Дуйсенов Н.Ж.. Даушеева Н.Н.

Big Data Management

Учебно-методическое пособие
для студентов специальности 5В070300- Информационные системы,
5В070400-Вычислительная техника и программного обеспечение

Шымкент 2019

УДК 681
ББК 87.256.6

Рекомендовано к печати решением учебно-методического совета университета «Мирас» (протокол № 10 от 28.05.2019 г.)

Рецензенты:
Кошкинбаева М.Ж., к.т..н, заведующая кафедройинформационных технологий и телекоммуникаций, университета «Мирас».
Бердалиева Г.А.,к.т.н., доцент, кафедры «Информационных систем»Южно-Казахстанский Государственный Университет им.М.Ауезова

Дуйсенов Н.Ж.
Даушеева Н.Н.
Big Data Management: Учебно-методическое пособие (для студентов специальности 5В070300- Информационные системы, 5В070400- Вычислительная техника и программного обеспечение).- Шымкент: университет «Мирас», 2019.- 104 с.

ISBN 978-99665-32-851-0

Учебное-методическое пособие предназначено для полготовки к занятиям, для выполнения самостоятельных работ по дисциплине «Big Data Management», а также для самостоятельного изучения дисциплины. Учебное пособие содержит необходимый теоретический материал по всему курсу, примеры решения типовых задач, а также задания для самостоятельного решения.
Для преподавателей и студентов ВУЗов.

УДК 001.891
ББК 87.256.6

СОДЕРЖАНИЕ

Введение

Целью изучения дисциплины «Big Data Managment» заключается в приобретении и закреплении навыков использования теоретических знаний для расчета переходных процессов и характеристик сигналов, проходящих через линейные электрические цепи, и построение качественных, и точных графиков переходных процессов. Основной задачей является углубленное изучение переходных процессов в линейных RLC – цепях и расчета таких процессов при помощи различных методов анализа.

В жизни всего человеческого общества в настоящее время, огромную роль играет разного рода информация, которая представляется людям в различной форме – от книг и газет до радиовещания и телевидения. Но в большинстве случаев, имеет место потоки информации в виде различных сигналов. Современному инженеру по сетям связи необходимо работать с огромными потоками информации и разрабатывать наиболее оптимальные способы её передачи и хранения, следовательно, ему приходиться сталкиваться с необходимость анализа переходных колебаний и прохождением сигналов через электрические цепи, поэтому данная тема является достаточно актуальной, а область её применения очень обширна. Знание методов анализа параметров переходных процессов играют важную роль в синтезе устройств, обслуживания оборудования, проектирования систем связи, и т.д. Все эти факты говорят о важности изучения данного направления.

Тема № 1. Big Data Managmentкак новый взгляд на проблемы управления и принятия решений

В этой главе вы познакомитесь с определениями и основными понятиями подхода к извлечению, анализу и трансформации данных, получившему название Big data - большие данные. Вы узнаете почему эта область деятельности заняла в планах правительства США отдельное место рядом с энергетикой, торговлей, транспортной инфраструктурой и образованием. Почему все больше высокотехнологичных стартапов выбирают тематику больших данных за основу своего бизнеса. Как появление этого феномена привело к необходимости обучения новых специалистов – Data Scientist – исследователь данных. Далее вас ждет экскурсия внутрь этого конгломерата научных направлений и технологий. Вы узнаете, что многие долгие и глубокие исследования прошлых лет органично породили техники больших данных. Успехи в суперкомпьютерных кластерах и параллельном программировании позволии создать технологические основы обработки больших данных. Как компании-гиганты информационных технологий начали использовать большие данные, и что представляет сегодняшний пейзаж распространения этого феномена.

Что такое Большие Данные

Термин Big Data появился как новый термин и логотип в редакционной статье Клиффорда Линча, редактора журнала Nature 3 сентября 2008 года, который посвятил целый специальный выпуск одного из самых знаменитых журналов теме “что могут значить для современной науки наборы больших данных”. И здесь использование слова “большие” было связано не столько с каким-то количеством, а с качественной оценкой, как например “Большая вода”. Время подтвердило справедливость выделения больших данных как отдельного феномена. Сегодня, согласно исследованиям агентства Gartner термин Big Data находится близко к пику знаменитого гартнеровского Hype Cycle – оценки относительной доли публикаций и обсуждений для различных технологических направлений. На рисунке 1.1 приведена опубликованная агентством кривая, определяющая гартнеровский цикл на 2013 год.

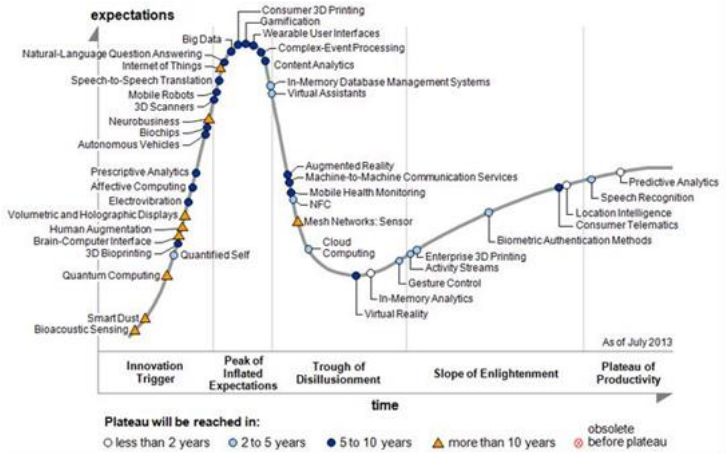


Рисунок 1.1 Гартнеровский цикл (Hype Cycle) новых технологий на 2013 год

В 2012 году в статье [1] Данаха Бойда и Кэт Крауфорд было сформулировано определение Big Data как культурного, технологического и научного феномена, включающего в себя : (1) Технология: максимизация вычислительно мощности и сложности алгоритмов для сбора, анализа, связывания и сравнения огромных наборов данных. (2) Анализ: изображение огромных наборов данных чтобы идентифицировать паттерны для того, чтобы делать экономические, социальные технические и юридические

утверждения. (3) Мифология: всеобщая уверенность, что огромные наборы данных представляют более высокую форму знаний и сведений, которые могут генерировать озарения, которые ранее были невозможны и с ореолом верности, объективности и точности.

Если заглянуть в Википедию, [http://en.wikipedia.org/wiki/Big_data] то можно найти определение, основанное на ключевых публикациях, развивающих употребление термина Big Data после упомянутого выше журнала Nature. Согласно этому определению, Big Data Managment– это термин, обозначающий множество наборов данных столь объемных и сложных, что делает невозможным применение имеющихся традиционных инструментов управления базами данных и приложений для их обработки. Проблему представляют сбор, очистка, хранение, поиск, доступ, передача, анализ и визуализация таких наборов как целостной сущности, а не локальных фрагментов. В качестве определяющих характеристик для больших данных отмечают «три V»: объём (англ. volume, в смысле величины физического объёма), скорость (англ. Velocity, означающее в данном контексте скорость прироста и необходимость высокоскоростной обработки и получения результатов), многообразие (англ. variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных). Ведущей характеристикой здесь является объем данных, который должен быть рассмотрен в аспекте приложений. В таблице 1.1 приведены все используемые сегодня производные единицы измерения объема данных.

Таблица 1.1

Название Размер по ГОСТ 8.417-2002 (приставки по СИ)		СимволПримечание:размер по стандартам МЭК		
байт	8 бит	B		
килобайт	10^3 B	KB	2^{10} =1024 байт	
мегабайт	10^6B	MB	$_2$20	байт
гигабайт	10^9B	GB	$_2$30	байт
терабайт	10^{12} B	ТБ	$_2$40	байт
петабайт	10^{15} B	PB	$_2$50	байт
эксабайт	10^{18} B	EB	$_2$60	байт
зеттабайт	10^{21} B	ZB	$_2$70	байт
йоттабайт	10^{24} B	YB	$_2$80	байт

Современность демонстрирует нам примеры чудовищных размеров генерируемых сегодня оцифрованных данных. Как утверждают гиганты ИТ индустрии (EMC, Cisco, IBM, Google) в 2012 году в мире было сгенерировано 2 зетабайта (2 * 1021) или 2 тысячи экзабайтов или 2 тысячи миллиардов гигабайтов информации, а в 2020 году эта величина достигнет 35 зетабайтов. За один день в 1012 году создано больше информации, чем было сгенерировано за весь 2002 год! Источниками этой лавины данных являются многочисленные цифровые устройства, концентрирующие и направляющие в бездонные просторы Интернета продукцию человеческого разума – твиты, посты в фэйсбуке и в контакте, запросы в поисковые системы, и т.п., а также данные от сенсоров и контроллеров

миллионов устройств, которые измеряют температуру и влажность, состояние дорог и кондиционеров и много другого, что сегодня объединяется термином “умные приборы”. Это видеопотоки с камер наблюдения, оцифрованные аудиосигналы, координаты GPS мобильных устройств и многое другое, порожденное машинами самостоятельно в процессе функционирования техники и существующее в виде битов данных. Все эти данные поступают на хранение в различные базы данных, хранилища и просто теряются. Часть этих данных доступна через Интернет, а часть имеет исключительно локальный доступ. В любом случае сегодня по оценкам экспертов используется не более 1-5 процентов от всех сгенерированных данных. Подход больших данных призван существенно увеличить использование имеющейся информации и позволить представить ее в подходящем для практического применения виде: принятия решений человеком или автоматического управления системами. Однако этому препятствует не только проблема количества – объем данных первая “v”. Для больших данных, как было уже отмечено важна вторая “v”- скорость. Результаты обработки больших данных должны быть получены за время, определяемое решаемой с их помощью проблемы. Это даст возможность превратить аналитику больших данных из инструмента, отвечающего на вопрос “кто виноват?”, характерного для традиционных систем аналитики, в инструмент для полученияответов “что делать?”. Аналитик в этом случае из врача патологоанатома превращается в терапевта. Скорость доступа к данным, скорость их процессинга является важным критерием качества технологий, входящих в большие данные.

Наконец третья “v” – разнообразие данных говорит о том, что Big Data Managmentдолжны эффективно обрабатываться независимо от их структурированности. Здесь принято выделять три основных вида данных по степени их структурированности.

Первый уровень – это привычные структурированные данные, которые могут быть представлены отделимыми и заранее определенными полями, в которых находятся биты, имеющие различную семантику. Например, все таблицы имеют в определенном поле заданной длины заголовки, в другом заранее заданном поле– один из фактов, в другом поле – другой из фактов, определяющих числовые или текстовые значения семантических переменных, содержащихся в заголовках. Структурированные данные хорошо хранить в реляционных базах данных и управлять такими данными удобно, используя специальный язык SQL – Structured Query Language. Несмотря на свою распространенность такие данные определяют только в 10 % от всего объема сгенерированных данных.

Второй уровень – это полуструктурированные (semistructured) данные. Данные такого типа имеют структурные разделители, но не могут быть представлены в виде таблицы из-за отсутствия части атрибутов у разных данных. Примером таких данных могут служить файлы в формате SGML – Standard Generalized Markup Language или BibTex в которых нет определенной схемы хранения данных, но семантический смысл различных элементов данных может быть определен по анализу самого файла. Иногда такие данные определяют как допускающие самоописание. Многие данные хранящиеся в Web относятся к полуструктурированным, данные библиографических описаний публикаций, научные данные.

Наконец, неструктурированные данные, которые по определению не могут подойти под ранее описанные виды. В них входят тексты, записанные символами различных языков, записи звуков, неподвижные изображения, видеофайлы, сообщения электронной почты, твиты, презентации и другая бизнес-информация вне выгрузок баз данных. Считается, что от 80 до 90 процентов всех данных в организациях относятся к неструктурированным данным. Нередко к неструктурированным относят и введенные выше полуструктурированные данные. Иногда шкалу разнообразия расширяют, используя целую шкалу от структурированных данных к полностью неструктурированным. Будем считать показатель вариативности данных нулевым для полностью неструктурированных данных и возрастающим до единицы для хорошо структурированных из реляционных баз.

Используя такие шкалы по всем трем “v” – составляющим больших данных приведем оценки объемов, скорости и вариативности данных от различных видов источников. На рисунке 1.2 приведено такое сравнение, сделанное на основе.

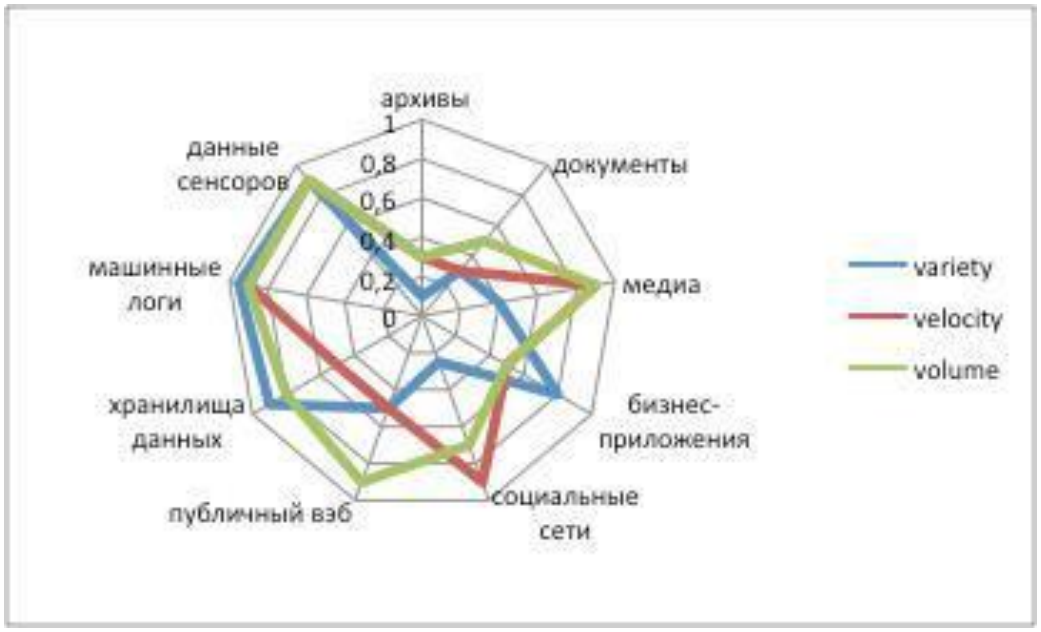


Рисунок 1.2 Big Data Managmentот различных источников имеют различные характеристики.

Тема № 2. Роль больших данных в технике, экономике и жизни

Big Data Managementкак феномен уже оказывают сильное влияние на бизнес и социальную жизнь многих людей. Это нередко происходит незаметно для средств массовой информации, но иногда, фокус публикаций переносится на какое-либо применение подхода больших данных, и непрофессиональные массы начинают осознавать всю грандиозность инструмента, попавшего в руки человека. И роль и возможности больших данных продолжают расти по мере усиления влияния компьютерных систем на всю деятельность людей на планете. Уже сегодня можно говорить о том, что киберпространство, где собственно и “живут” этиэксабайты данных оказалось тесно связанным с реальным миром, где живут люди, передвигается транспорт, работают заводы, цветут цветы. И не только активность реального мира отражается в киберпространстве, но и процессы в киберпространтве управляют многим, что происходит в мире реальности. Последнее время на рынке все больше становится “умных” вещей, вещей, выполняющих привычные и нужные для нас материальные функции, но содержащих в себе управляющие компьютеры, позволяющие чрезвычайно разнообразить их поведение, улучшить возможности взаимодействия с ними. И эти вещи немедленно строят свое отражение в киберпространстве. Интернет вещей – IOT (Internet of Things) растет экспоненциально, и, взаимодействуя с “Интернетом людей”, образует невиданный симбиоз киберпространства и реального мира. Big Data Managementпозволяют увидеть имеющиеся в киберпространстве данные в прагматическом ракурсе решаемой задачи, поставленной человеком (по крайней мере пока человеком). И если это маркетолог из Амазона – крупнейшего в мире Интернет-

магазина, то его задача наблюдая за следами в киберпространстве оставляемыми каждым человеком, придумать какой товар предложить ему для покупки. Извлекая из океана данных все, что относится к этому человеку он анализирует где он живет по координатам мобильного телефона, что искал в Гугле или Яндексе, что покупал прошлые годы, сколько платил за товары и т.п. Обработка этих данных позволяет в подходящее время сделать предложение о покупке от которой будет трудно отказаться. Ну а если это аналитик из Агентства национальной безопасности, то анализируя данные о миллионах людей, которые они сами оставили в киберпространстве, он может обнаружить персоны с аномальным поведением и предотвратить террористический акт, который мог унести сотни и тысячи жизней. В этой книге мы не будем касаться столь специфических использований подхода больших данных. Отметим только, что постоянный анализ потока сообщений в социальных сетях – а это типичная задача больших данных- может дать весьма достоверную информацию о происходящих в обществе процессах.

Джон Клейнберг, профессор Корнелльского университета сказал: «Big Data Managmentпозволяют мне определить горячие точки, в которых начинаются процессы, которые станут господствующими в будущем. Если бы интернета с социальными сетями не было, если бы не существовало подхода больших данных, я бы никогда не смог инструментально определить эти горячие точки». Одним из наиболее известных молодых политологов является Джастин Гример, 28-летний исследователь из Стэнфорда, который объединил математику и политическую науку. Его исследования базируются на обработке больших данных, включая структурированную и неструктурированную информацию из соцсетей, блогов, форумов, выступлений в Конгрессе, новостных порталов. Суть его работы состоит в том, чтобы определить, как действуют прямые и обратные связи, выражающиеся в тех или иных политических решениях, между людьми в Конгрессе и их избирателями.

Гарвардский университет открыл институт количественных социальных наук. Его директор Гарри Кинг говорил в своем выступлении о новых возможностях аналитики в социальных науках: « Это революция, и она только началась. Эта революция стала реальностью благодаря возможности обработки огромного количества источников данных самого различного формата, как структурированных, так и неструктурированных, как вычислимых, так иневычислимых». Эндрю Гельман, один из наиболее авторитетных статистиков и политологов Америки говорит: «Методы не изменились, но Big Data Managmentсделали их эффективными. Теперь математика и статистика – это интересно и весело. Это просто круто»

Big Data Managment– это действительно круто. Вот другой пример. Анализируя Big Data Managmentинтернет-запросов, исследователи обнаружили странный феномен. Уже несколько лет всплеск поисковых запросов Google по таким терминам, как лечение гриппа, симптомы гриппа и т.п. на несколько недель предваряет начало стремительного нарастания эпидемии гриппа. Эта закономерность уже сегодня используется для проведения превентивных мер по предотвращению во многих штатах эпидемии гриппа, подготовке врачей, освобождению лечебных коек и т.п.

Следует отметить, что используемая до этого информация, поступающая от участковых врачей и пунктов неотложной помощи, как правило, отставала от реальной картины. Специалисты Федеральной резервной системы выяснили, что статистика поисковых запросов Google относительно покупки домов является более надежным источником для определения тенденций в увеличении или уменьшении объемов продаж недвижимости и динамики жилищного строительства, чем прогнозы наиболее известных экономистов. По мнению участников Всемирного экономического форума 2012 года в Давосе, те, кто оседлает тему интеллектуального анализа больших данных, станут хозяевами информационного пространства. Этой теме был посвящен специальный доклад на Форуме «Big Data Managment– большое влияние». Ключевой вывод доклада – цифровые активы становятся не менее значимым экономическим активом, чем золото или валюта.

Исследования, проведенные профессором Бринйолфсоном (E.Brynjolfsson) и двумя его коллегами в 2012 году, показали, что анализ и прогнозирование на основе больших данных берется на вооружение корпоративной Америкой.

Они изучили 179 крупных компаний и обнаружили, что те из них, кто взял в последние год-полтора на вооружение интеллектуальный анализ больших данных получил немедленное улучшение экономических показателей на 5-6%. С учетом оборота и размеров этих компаний это очень и очень много и показывает сумасшедшую рентабельность вложений в интеллектуальный анализ больших данных. И если сегодня, пока, как показывают исследования агентства Wikibon research, компании не получают должной отдачи от инвестиций в технологии Big Data и от каждого вложенного доллара пока возвращается половина, то в следующие три – пять лет ситуация кардинально изменится, и ROI составит не менее 3,5 долларов на один доллар инвестиций.

Если попытаться оценить роль больших данных в экономике и развитии бизнесагосударственном масштабе, то непременно следует изучить обширные доклады Института глобального развития корпорации МакКинси (McKinsey Global Institute)под названием Bigdata: Thenextfrontierforinnovation, competition, andproductivity(Большие данные: очередной рубеж для инноваций, конкуренции ипродуктивности), а также Gamechangers: FiveopportunitiesforUSgrowthandrenewal (Меняющие правила игры: пять возможностей для роста обновленияСША). Один из основных выводов, которые сделаны авторами состоит в том, что Big Data Managmentстановятся заметным двигателем роста валового национального продукта (GDP). Анализ, выполненный McKinsey, показал, что Big Data добавят $325 млрд к ВВП к 2020 году. И этот вклад вполне сравним с другими факторами, которые окажут позитивное влияние на экономику. Это энергетика (шельфовая добыча нефти и газа), торговля, инфраструктура (инвестирование в транспортную сеть, строительство дорог и железнодорожных путей), а также образование и трудоустройство. Авторы доклада иллюстрируют этот вывод диаграммой, приведенной на рисунке 1.3.

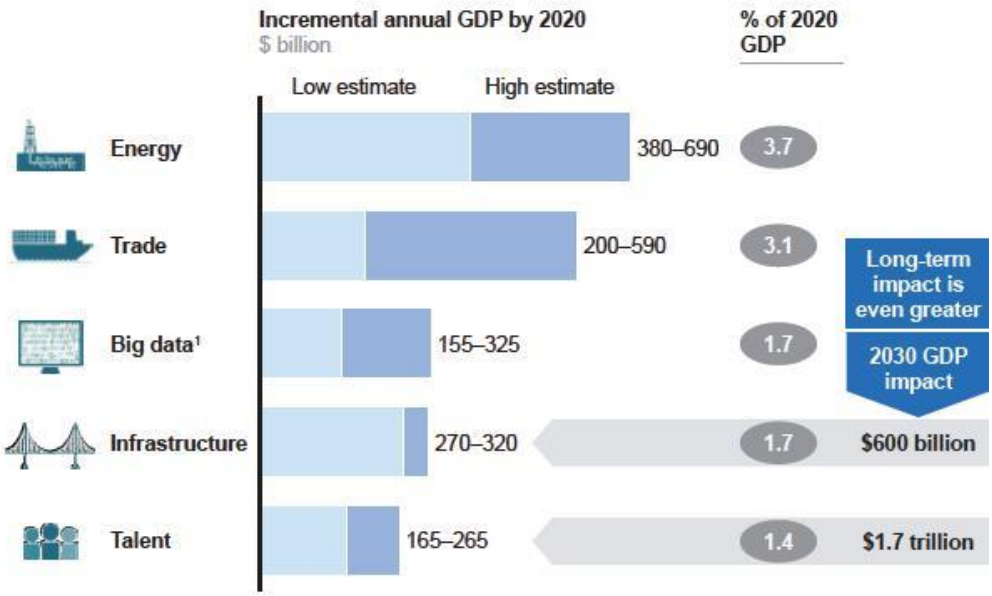


Рисунок 1.3. Пять главных источников прироста ВВП в США.

Большие данные, разумеется не представляют собой отдельную отрасль экономики и выделение этого направления человеческой деятельности и соответствующих бизнесов лишь призвано показать какая доля прироста ВВП зависит от того, будут ли предприятия и организации различного профиля деятельности внедрять в свои бизнес-процессы технологии и сервисы индустрии больших данных. Области, на которые Big Data

оказывают наибольшее воздействие - это продуктивность, предпринимательство и инновации. Именно эти области являются болевыми точками экономики. Один из серьезных вызовов сегодня является старение рабочей силы вслед за населением. То есть средний возраст сотрудника американской компании превышает допустимую норму. По прогнозам, ситуация в ближайшее время будет только ухудшаться — и, чтобы поддерживать рост ВВП на прежнем уровне, необходимо увеличить производительность на 30%. По прогнозам McKinsey, Big Data Managmentмогут дать прирост производительности и сгенерировать дополнительные доходы в различных секторах экономики.

В докладе выделены здравоохранение (0,7% рост производительности, 300 млрд. долларов ежегодного прироста объемов), торговля (0,5-1,0 %, +60% прироста), производство (7% производительность,+50% - объемы). Кроме этого в докладе также назван такой специфический сектор как глобальные персональные данные о местоположении (700 млрд долларов дохода для конечного пользователя и 100 млрд долларов для сервис-провайдеров). В докладе показано, что для администрации Евросоюза использование технологий больших данных может приносить ежегодно доход в 250 миллиардов евро и рост производительности на 0,5%. Мы рассмотрим здесь более детально три основных сектора.

Тема № 3. Область использование Big DataManagement

Здравоохранение

В экономике США здравоохранение представляет собойодин из самых больших секторов, создающий 17% ВВП и использующий 11% всей рабочей силы. Ежегодно расходы на здравоохранение возрастают почти на 5%. И это неизбежно в условиях старения населения. Анализ, сделанный в докладе, показывает как Big Data Managmentмогут не только создать дополнительный источник компенсации затрат, но и повысить качество медобслуживания. В основе больших данных может быть объединена информация, хранящаяся в четырех главных пулах данных, которые сегодня невзаимосвязаны. Это (1) фармацевтические данные полученные в ходе исследований и испытаний, (2) данные из клиник по историям болезни и диагностике, (3) данные о поведении пациентов, их покупки, отзывы, данные от домашних медицинских приборов и даже от одежды и обуви, такие как кроссовок с сенсорами, (4) данные от медицинских учреждений об оказании услуг, аптек об отпуске препаратов, сведения о ценах на рынке здравоохранения. На основе анализа всех этих данных предполагается развивать пять направлений использования больших данных.

1. Операционная деятельность медицинских учреждений. Ожидается что здесь будет получено ежегодное сокращение расходов не менее чем на 165 миллиардов ежегодно. Здесь станут возможными исследованияэффективности лечения благодаря обработке всей доступной информациио практике лечения. На основе анализа всех известных историй болезни и диагностики в практику врачей войдет широкое использование систем поддержки принятия решений, позволяющих предоставить клиницисту невиданный ранее доступ к опыту тысяч коллег по всей стране. Существенного сокращения затрат и повышения качества жизни дают методы персональной и превентивной медицины, основанные на удаленном мониторинге пациентов. Распространение различных сенсоров активностей человеческого организма, подключаемых к смартфонам, позволяет сократить необходимость проведения лабораторных исследований, предотвратит неожиданные осложнения, а автоматическое напоминание о необходимости проведения самостоятельных лечебно-профилактических манипуляций повысит качество назначенного лечения.

2. Система ценообразования и оплаты. Ожидается генерирование около 50 миллиардов долларов ежегодно, которые также сократят расходы на здравоохранение. Анализ счетов и поступлений с помощью автоматических процедур, основанных на машинном обучении и нейронных сетях, позволит сократить число ошибок и хищений при оплате. Сегодня от 2 до 4 процентов ежегодных платежей оказываются невнесенными. Формирование ценовых планов, учитывающих реальные возможности населения и потребность в услугах, также увеличивает общие поступления от пациентов. Только системы работающие с большими данными позволяют перейти к оплате, основанной на производительности и совместно регулировать расходы на медикаменты и труд медперсонала.

3. Исследования и разработки. Big Data Managmentв этом направлении здравоохранения позволят получить доходы в 100 млрд долларов ежегодно и сократить расходы бюджета на 25 миллиардов в год. Наибольший эффект здесь ожидается от новых возможностей предиктивного моделирования при разработке лекарственных препаратов. Не меньшее влияние статистические алгоритмы и инструменты больших данных производят на планирование клинических испытаний и привлечение пациентов к таким испытаниям. Обработка результатов таких испытаний еще одно важное приложение больших данных. Особое место в исследованиях и разработках в здравоохранении сейчас занимают инновации в персонализированной медицине. Основываясь на обработки гигантских объемов генетической информации, доступной сегодня каждому, врачи смогут назначать абсолютно уникальные лекарственные средства и методы лечения. Наконец, разработки по выделению паттернов заболеваний позволят получить хорошие прогностические оценки развития различных видов болезней, выделить профили рисков и не только провести профилактические мероприятия, но и спрогнозировать необходимость разработок методов лечения, эффективных для будущих видов заболеваний.

4. Новые бизнес-модели. Основанные на цифровых данных в здравоохранении эти модели могут дополнять существующие или дажеконкурировать с некоторыми из них. Это агрегаторы данных, которые поставляют очищенные и скомпонованные блоки данных, удовлетворяющих заданным условиям, третьим лицам. Например все истории болезней пациентов, применявших тот или иной фармакологический препарат важны для фармпредприятий и они готовы покупать такие данные. Другим потенциалом новых бизнес-моделей являются он-лайн платформы для пациентов и врачей, медицинских исследователей и фармакологов.

5. Массовый скрининг и предупреждение эпидемий. Это направление опирается на Big Data Managmentи развитие технологий позволит строить как географические и социальные модели здоровья населения, так и предиктивные модели развития эпидемических вспышек.

Производство

Этот сектор экономики всегда приводился в движениеинформацией, содержащейся в различных данных. Начиная с разработок, основанных на маркетинговых исследованиях, моделях будущих продуктов воплощаемых в системах автоматизированного проектирования (CAD-Computer Aided Design), системах автоматизированного производства(CAM- Computer Aided Manufacturing), управляющих технологическими процессами производства и заканчивая системами отношений с клиентами(CRM – Customer Relationship Management), производственная цепочка полностью зависит от качества и актуальности доступных данных. Big Data Managmentпозволят расширить возможности производственных систем и обеспечить как сокращение расходов на производство, так и невиданное ранее ускорение освоения новых изделий и оптимальную организацию цепочек поставок и управления трудовыми ресурсами. Глобализация требует перманентного контроля процессов на разбросанных по всей планете производственных и торговых

площадках. Только радиочастотных меток, которые сегодня вовлечены в процессы поставок будет продано не менее 209 миллиардов штук в 2021. Производственные системы уже в 2010 году манипулировали почти двумя экзабайтами данных. Ведь например один только Боинг-737 при своем производстве генерирует около 240 терабайт данных. Основные направления использования больших данных в производственной цепи ценностей могут быть представлены нижеследующими.

1. Исследования, разработка и дизайн продуктов. Использование больших данных здесь позволяет ускорить производственный процесс, учесть пожелания потребителя, сократить стоимость разработки благодаря подходу открытых инноваций. Менеджмент получает жизненный цикл продукта, который учитывает создание и развертывание платформ коллаборативного труда. Учет потребностей рынка постепенно заменяется на учет потребностей отдельных потребителей.

2. Изготовление. Процессы изготовления развиваясь вступили в фазу называемую “цифровая фабрика”. В них разработанная цифровая модель продукта трансформируется в технологическую модель, пригодную для материализации на установках обработки и синтеза материалов, механической обработки деталей, 3D принтинга. В изготовлениипринципиальной основой новых цехов является “интернет вещей” (IOT). Информация от сенсорных распределенных сетей позволяет оптимизировать процессы изготовления включая наносборки. IOT подход в нефтепереработке позволяет достигнут экономии затрат в 60% благодаря мониторингу, автоматизации, синхронизации всех процессов, технических и организационных систем.

3. Маркетинг, процесс продаж и послепродажной поддержки. Поступающие от потребителей данные, отзывы и технические отчеты позволяют влиять на будущие направления разработок, определять профиль потенциальных покупателей, обеспечивать своевременное обслуживание и модернизациюуже проданного и установленного оборудования.

Особое место в производстве занимает производство электрической энергии. Этой части индустриального сектора также свойственно широкое использование информационных технологий и Big Data Managmentиграют там важную роль. Мы посвятили применению больших данных в электроэнергетике всю вторую главу книги.

Торговля (Retail)

Этот сектор экономики также получаетвозможность увеличить доходы и повысить продуктивность за счет внедрения технологий больших данных в информационные платформы поддержки и организации бизнеса. Собственно в инновационных компаниях этого сектора и были разработаны и развернуты первые системы, использующие Big Data Managmentдля увеличения эффективности продаж. Возможности Интернет-магазинов по мониторингу поведения клиентов оказались ключевыми для появления рекомендательных систем, одного из главных двигателей технологий больших данных. Именно в торговле ожидается не менее чем полупроцентный прирост производительности только за счет использования больших данных и рост доходов почти на 60%. Пример такого ритэйлера как пионер в области обмена электронными данными Wal-Mart показывает возможности больших данных для учета и управления товарами и товарными цепочками от поставщика до покупателя. Применив инновационные технические средства – радиочастотные метки (RFID) Wal-Mart получил возможность трассировки индивидуальных товаров что не только повысило качество операций, но и позволило развить новые стратегии построения отношений с поставщикамии клиентами. Вообще в торговле могут быть выделены 16 точек приложения больших данных, где это создает существенный эффект. Рассмотрим эти возможности сгруппировав по типовым направлениям в торговле.

1. Маркетинг. Здесь Big Data Managmentв первую очередь создают опору для кросс-продаж – продаж по принципу “вы возможно также хотите…”. Рекомендательные системы

позволяют создавать до 30% продаж такому магазину как Amazon. Другая точка приложения больших данных – это использование информации о местоположении для корректных рекламных сообщений и рекомендаций. Смартфоны и другие мобильные устройства делают сервисы, основанные на информации о местоположении (Location Based Services), мощным инструментом продаж. Еще одна точка приложения – анализ поведения покупателя в торговых залах. Анализ типовых паттернов поведения позволяет реализовать эффективную раскладку товаров, позиционирование витрин и т.п. Микросегментацияклиентов с использованием изощренных методов анализа потока кликов индивидуального клиента позволяет строить программы бонусов или скидок, учитывающих весьма тонкие особенности, что приводит к стабильному повышению лояльности клиентов. Наконец, практически недостижимой без использования больших данных, является управление маркетинговыми операциями на основе анализа мнений (sentiment analysis). Для этого собираются высказывания клиентов в социальных сетях, блогах, форумах и извлекается информация об их отношении к тем или иным товарам, услугам или поставщикам. Многоканальность источников информации о клиентах также рассматривается как одна из точек приложения больших данных. Совместная обработка данных о покупках вместе с данными о домах, в которых проживают покупатели, данных об их доходах, детях, работе позволяет усилить акции по продвижению товаров, отыскать точки заинтересованности в услугах.

2. Продажи. В этом функциональном направлении торговли Big Data Managmentпозволяют увеличить прозрачность выполнения операций. Мониторинг всех действий по продаже и проблем при их исполнении дает ключ к операционным усовершенствованиям и новым механизмам управления операциями. Другой точкой приложения больших данных здесь является оптимизация трудовых затрат. Это контроль за выполнением сотрудниками трудовых обязанностей, мониторинг расписаний и распорядка и механизмов поощрения и наказания.

3. Цепь поставок. Важнейшими точками приложения больших данных к повышению эффективности цепи поставок является прежде всего возможность оперативной инвентаризации и управления учетом в целом. Второй такой точкой служит оптимизация логистики и распределения. Возможности технологий геолокации порождая потоки данных, тем самым дают возможности наблюдать и управлять логистикой. Наконец, Big Data Managmentоткрывают возможность наглядно информировать поставщиков товаров о предпочтениях покупателей и добиваться успеха в переговорах о закупках как в отношении их объемов, так и по ценам.

4. Новые бизнес-модели. Доступ к обширным данным опокупкам и предпочтениям покупателей открывает возможности появления новых бизнес-моделей в торговле. Одними из последних здесь были сервисы по сравнению цен у различных продавцов и рынки в Интернете (Web based markets). Особенностью таких рынков является предоставление покупателям огромных объемов информации о товарах и услугах,различные сравнения, доступа к отзывам других покупателей и т.д. Перечисленными выше секторами экономики роль больших данных в росте ВВП не исчерпывается и практически вся хозяйственная и социальная жизнь человека уже испытывает их влияние. Особую роль, которая будет обсуждаться ниже, Big Data Managmentиграют в образовании. Авторы видят здесь удивительную возможность ведения учета образовательной траектории каждого человека в стране, начиная с его глубокого детства, с первых данных, получаемых в детских садах, через школьные документы, описывающие текущую успеваемость, особенности поведения и склонности, итоговые документы аттестации, учебу в колледжах и вузах и далее в институтах повышения квалификации и профессиональных курсах. Здесь видится возможность рассмотрения родившихсястране людей как важный национальный ресурс, который подлежит мониторингу и анализу как в среднестатистическом смысле, так и в целях прогностических выводов о будущем состоянии трудовых ресурсов. Надеясь, что этому проекту удастся посвятить другие публикации, здесь мы рассматриваем только

одинаспект больших данных в образовании – необходимость подготовки кадров, умеющих эффективно использовать возможности больших данных в любой области человеческой деятельности. Возможно, когда то эти умения станут частью любой профессии и, тогда не придется говорить об этой специфике отдельно. Но в настоящее время потребности общества делают необходимым появление специалистов по большим данным в форме отдельной профессии. Название этой профессии Data Scientist – исследователь данных. В США произвели оценку потребностей в специалистах такой профессии и пришли к выводу, что уже в 2018 году в США будет нехватка исследователей данных в количестве 190 000 человек! Известный журнал Harvard Business Review так озаглавил один из своих выпусков: “Data Scientist: The Sexiest Job Of the 21st Century” – исследователь данных – самая привлекательная работа 21 столетия. Кто такой этот исследователь данных - обобщается сегодня в целом ряде публикаций [4]

- Любит данные
- Исследовательский склад ума
- Цель работы – нахождение закономерностей в данных
- Практик, не теоретик
- Умеет и любит работать руками
- Эксперт в какой-либо прикладной области (обычно, но не обязательно)
- Работает в команде
- Препочтительное образование:
- Computer Science
- Статистика, математика
- Точные науки- физика, инженерия и т.п.
- Магистры и кандидаты наук

Исследователь данных не является математиком, но обязательно владеет техниками больших данных.

Исследователь данных не является программистом, но обязательно умеет работать с программными средствами и владеет технологиями больших данных.

А что включают в себя техники и технологии больших данных является предметом содержания следующей главы.

Тема 4. Техники и технологии больших данных

Мы будем использовать следующие определения, которые различают техники и технологии.

- Техника (чего-либо) – способ или процедура выполнения какой-либо задачи
- Технология – приложение результатов науки, чаще всего к промышленным или коммерческим целям

Большие данные, было сказано выше, как феномен, включают в себя не только собственно данные как объект операций, но и комплекс особых операций над этим объектом, поскольку объект оказался специфическим и неподвластным известным ранее методам. Операции над объектом всегда включают в себя ответ на вопрос “Как?” – и это и есть собственно выбор той или иной техники, и ответ на вопрос “Чем?” - а это выбор технологии, инструмента и методики его применения.

Техники больших данных

Сначала перечислим те функциональные операции над данными, методы их хранения и обработки, некоторые из которых рассматриваются в этой книге. Разумеется, этот список не исчерпывает всего многообразия динамично развивающихся техник, однако позволяет увидеть, что можно делать с большими данными для достижения целей, стоящих перед исследователем.

- Извлечение данных
- Краудсорсинг – сбор данных от большого числа источников
- Консолидация данных
- Визуализация
- Машинное обучение (с учителем и без учителя)
- Генетические алгоритмы
- NLP (Natural Language Processing) – Обработка естественного языка
- Нейронные сети
- Анализ сетей
- Оптимизация
- Распознавание образов
- Предиктивное моделирование
- Регрессионный анализ
- Обработка сигналов и анализ временных рядов
- Сентиментный анализ – извлечение «чувств»
- Пространственный анализ
- A/B тестирование
- Анализ правил ассоциации
- Классификация
- Кластерный анализ

Использование техник и методов в практически любом проекте с использованием больших данных происходит в функциональном потоке, основные составляющие которого приведены на рисунке 2.1

Рисунок 2.1 Функциональный поток работы с данными

Консолидация данных

Этот целый набор техник, направленных на извлечение данных из разных источников, обеспечение их качества, преобразования в единый формат и загрузку в хранилище данных – “аналитическую песочницу” (analytic sandbox) или «озеро данных» (data lake). Техники консолидации данных различаются по виду аналитики выполняемой системой :

- Пакетная аналитика (batch oriented)
- Аналитика реального времени (real time oriented)
- Гибридная аналитика (hybrid)

При пакетной аналитике периодически производится выгрузка данных из различных источников, данные анализируются на наличие сбойных фрагментов, шума и производится их фильтрация. При выполнении аналитики реального времени данные производятся источниками непрерывно и образуют набор потоков данных.

Анализ этих потоков и своевременное получение результатов в заданном темпе требуют обеспечить асинхронное получение данных в виде некоторых сообщений и маршрутизировать эти сообщения в нужные процессинговые узлы для обработки. Для гибридной аналитики как правило сообщения данных должны быть не только маршрутизированы на процессинг, ноинтегрированы в аналитическую песочницу для дальнейшей обработки по результатам накопления данных за значительные интервалы времени.

Данные, полученные в результате консолидации, должны соответствовать определенным критериям качества. Качество данных - это критерий, определяющий полноту, точность, актуальность и возможность интерпретации данных. Данные могут быть высокого и низкого качества. Данные высокого качества - это полные, точные, актуальные данные, которые поддаютсяинтерпретации. Такие данные обеспечивают получение качественного результата: знаний, которые смогут поддерживать процесс принятия решений. Данные низкого качества - это так называемые грязные данные. Грязные данные могут появиться по разным причинам, таким как ошибка при вводе данных, использование иных форматов представления или единиц измерения, несоответствие стандартам, отсутствие своевременного обновления, неудачное обновление всех копий данных, неудачное удаление записей-дубликатов и т.д. Наиболее распространенные виды грязных данных:

1. пропущенные значения ;
2. дубликаты данных;
3. шумы и выбросы.

Указанные виды грязных данных могут быть очищены с помощью различных методов, зависящих от типа данных (текст, числа.изображения).Интеграция данных базируется на понятии стандарта представленияинтегрированных данных. Стандарт описывает формат хранения информации. Ондолжен быть удобным для использования при просмотре данных и построениямоделей интеллектуального анализа данных.

Существует целый рядспецификаций, претендующих на место стандарта представленияинтегрированных данных. Большое распространение в качестве контейнера дляхранения данных получил XML. Изначально этот язык разметки информационныхстраниц не предназначался для интеллектуального анализа данных. Но развитиенереляционных баз данных привело к тому, что XML оказался удобен в качествеосновы унификации данных из разных источников. Для хранения интегрированныхданных используются и международные стандарты, специально разработанныесовсем для иных целей. Например, широко известен стандарт ISO 15926. Когдаавтоматизированное проектирование и трехмерное моделирование были впервыекомпьютеризированы, возникла потребность в стандарте, который мог бырегистрировать изменения для производственного предприятия на протяжениивсех стадий его жизненного цикла. Стандарт должен был описывать хранилищеданных, содержащее требования к промышленному объекту, проект предприятия,физические объекты, из которых состоит промышленный объект, и изменения,всех объектов хранения. Как результат был предложен стандарт ISO 15926 -«жизненный цикл промышленного объекта».

Он весьма устойчиво развивался, и сэтим развитием идея внедрения этого стандарта стала получать все большеераспространение в промышленности. Выходя за пределы передачи данных, ISO15926 быстро стал стандартом для совместного функционирования и совместнойработы сложных моделей данных всех типов, включая 2D и 3D данные. Выборстандарта интегрированных данных во многом обуславливается характеромобрабатываемой информации, инструментарием и окружением, доступнымиинтегратору, популярностью стандарта в предметной области. Единого для всехсистем анализа данных стандарта не существует.

Совокупность процессов, определяющих консолидацию, называют ETL – Extraction-Transformation-Loading (Извлечение-Преобразование-Загрузка).

Вприложения бизнес-аналитики (BI- Business Intelligence) в процессы ETL включались весьма сложные преобразования данных, такие как квантование, позволяющее снизить объем обрабатываемых данных, нормализация – процесс приведения реляционных таблиц к каноническому виду или числовых данных к единому масштабу, кодирование данных – введение уникальных кодов для сжатия данных.

В техниках больших данных обычно полагают, что необходимо работать непосредственно с грязными данными, поскольку нередко именно характер сбоев может стать предметом анализа, а сжатие данных представляет собой функцию собственно аналитических алгоритмов.

Возможность же хранения данных в исходном виде должны предоставлять технические средства аналитической системы. Качество больших данных нередко трудно оценить методами формальных алгоритмов, и тогда прибегают к визуализации на раннем этапе исследования. Кроме оценки качества и выбора метода препроцессинга, визуализация может помочь перейти к важному этапу аналитики - выбору моделей, гипотез для достижения конечной цели - принятия решений.

Визуализация

Техника визуализации является мощным методом интеллектуального анализа данных. Как правило, ее используют для просмотра и верификации данных перед созданием модели, а также после генерации прогнозов. Визуализация - это преобразование численных данных в некоторый визуальный образ, в целях упрощения восприятия больших массивов информации.

Для осуществления визуализации служат визуализаторы. Визуализаторы могут являться либо отдельным приложением, либо плагином или частью другого приложения. Возможности визуализаторов очень широки.

В настоящее время они могут представлять информацию практически во всех мыслимых видах, лишь бы аналитик мог сформулировать, что он хочет видеть.

Простейшие привычные визуализации базируются на двумерных диаграммах и гистограммах.

Характеристики объектов данных откладываются по осям. Каждая точка на графике соответствует какому-либо объекту из базы данных. На рисунке 2.2 и рисунке 2.3 приведены привычные точечная и столбцовая диаграммы.

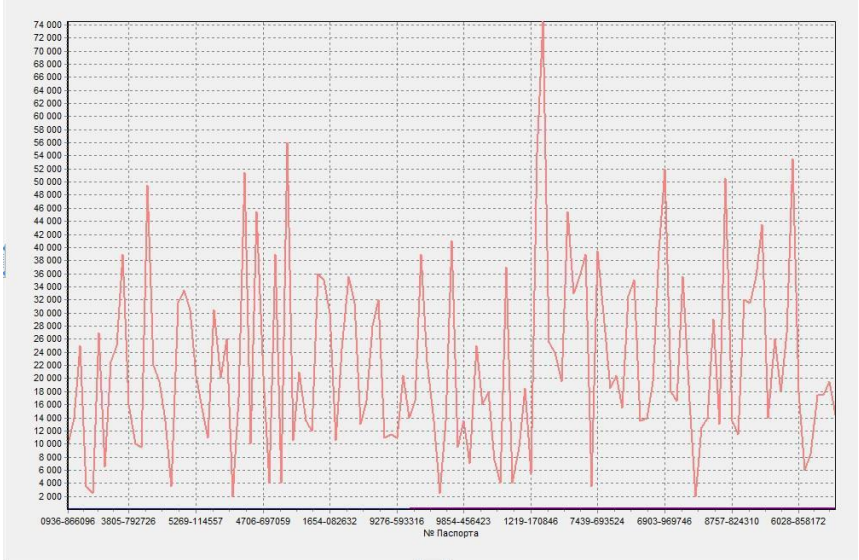


Рисунок 2.2 График зависимости одной действительной переменной от значений другой переменной

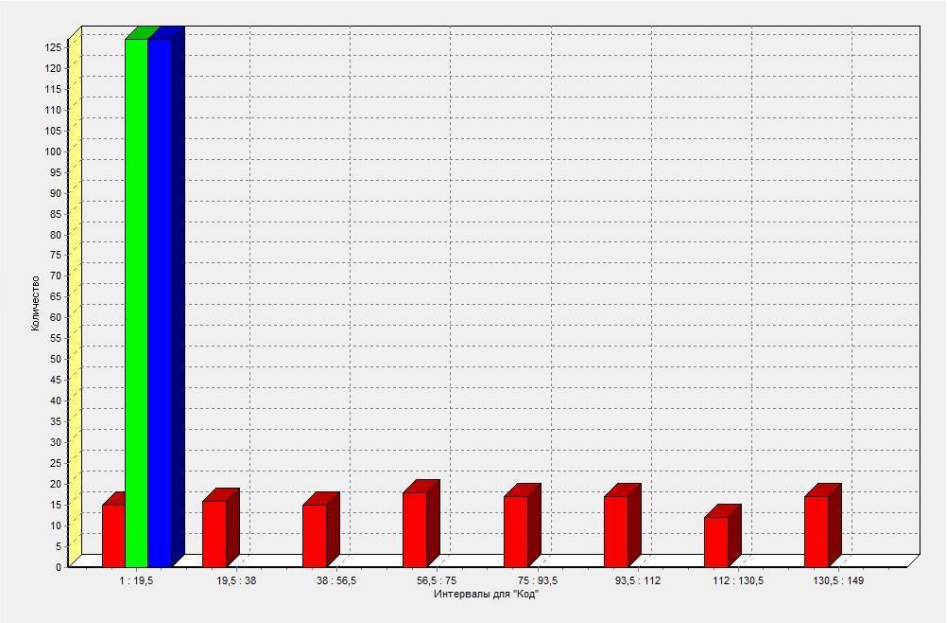


Рисунок 2.3 Столбцовая диаграмма зависимости целочисленной переменной от другой целочисленной переменной (кода измерения)

Более специфические виды визуализации можно отнести к одному из следующих типов.

Визуализация текстов

Если данные представляют собой тексты на естественном языке, то первичную помощь в анализе может оказать визуализация с помощью размеченного текста. Визуализатор подсчитывает частоту упоминаний того или иного слова, и присваивает словам условный вес, зависящий от этой частоты. Слова разного веса при визуализации имеют различную разметку, а значит разное представление на экране. Одни слова выглядят больше других.

Рисунок 2.4 Рейтинговая диаграмма текста

Этот тип визуализации помогает исследователю очень быстро ухватить основные мысли текста.

Визуализация кластеров

Одной из часто используемых визуализаций является визуализация кластеров. Кластерами называют группы в чем-то схожих или близких по свойствам объектов. Алгоритмы кластеризации, т.е. разбиение множества объектов на группы, мы рассмотрим ниже, а здесь покажем только как может быть визуализирована их работа. Большинство визуализаторов поддерживает алгоритмы кластеризации и способно разделять данные на кластеры. Обычно для визуального представления кластеров для объектов из разных кластеров используются контрастные цвета.

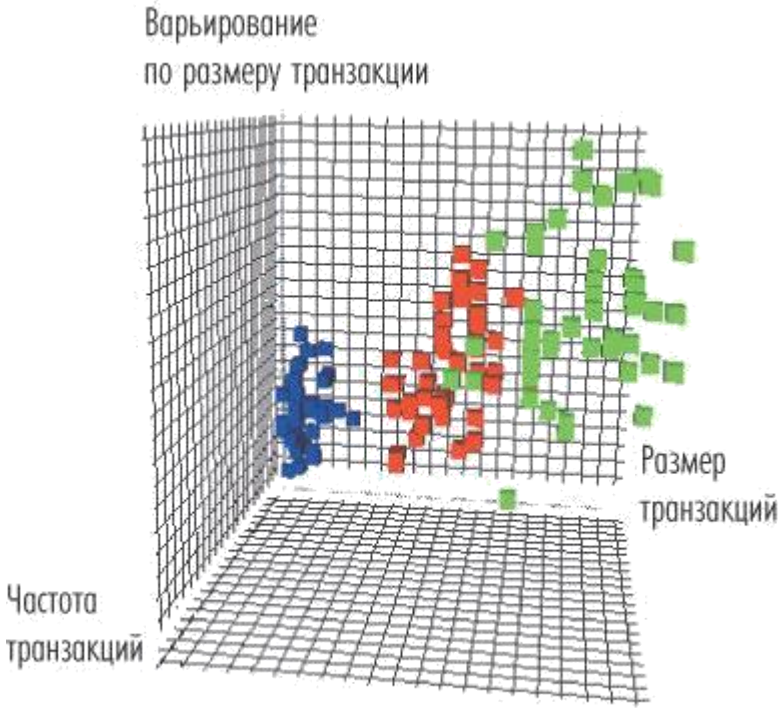


Рисунок 2.5 Диаграмма, иллюстрирующая кластеризацию в трехмерном пространстве измерений

Основная проблема при кластеризации данных – определение оптимального количества кластеров. Для того, чтобы упростить аналитику оценку и сравнение различных вариантов, некоторые визуализаторы предлагают построение кластерограм.

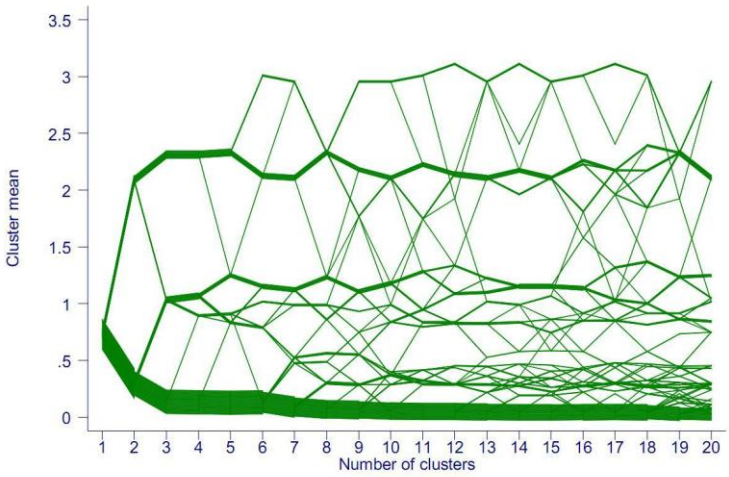


Рисунок 2.6 Кластерограмма – зависимость размеров кластеров от их числа

Визуализация ассоциаций

Визуализация ассоциаций демонстрирует частоту, с которой те или иные элементы появляются вместе в наборе данных, за счет чего определяется структура организации данных (например, речь может идти о том, какие продукты часто продаются вместе). Также возможна визуализация информации о силе ассоциации данных.

В различных визуализаторах картинка будет различной, в зависимости от выбранных обозначений. Если правила представляются на сетке, а виды и сила ассоциации изображаются в точках сочленения в узлах сетки в виде столбцов, дисков и меток, то визуализация ассоциаций станет выглядеть так, как показано, например, на рисунке 2.7.

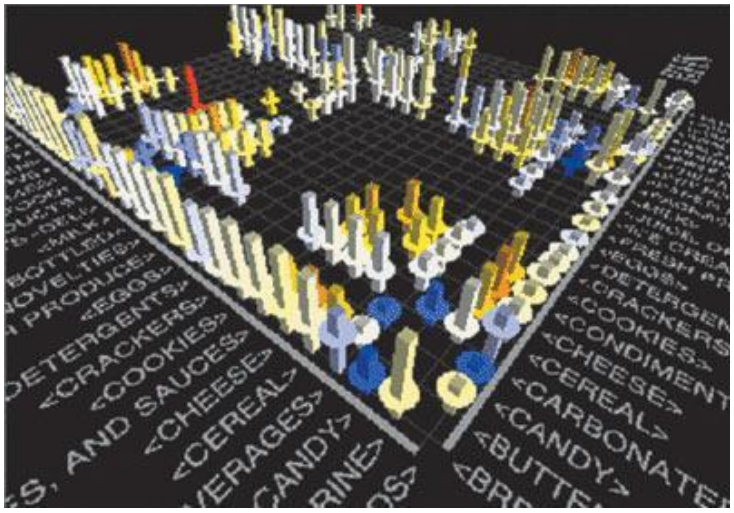


Рисунок 2.7 Диаграмма ассоциативных связей между словами текста

В качественных визуализаторах сетку с 3D-правилами можно масштабировать, вращать и панорамировать. В них можно менять цвета и шрифты, экспортировать картинку в различных форматах.

Частным случаем визуализации ассоциаций является визуализация информационных потоков. В этом случае некоторым образом изображается массив источников и приёмников информации, а затем они соединяются с помощью условных обозначений. На рисунке 2.8

показана визуализация связей точек, имеющих определенную привязку к географическим координатам.

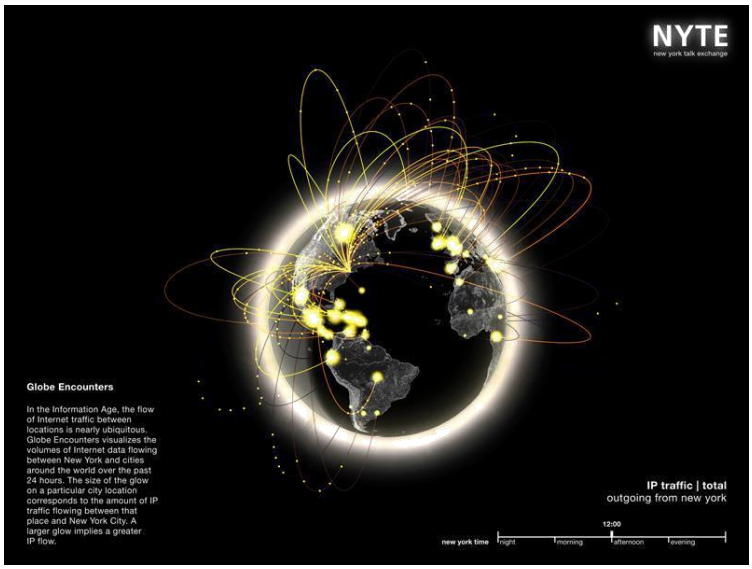


Рисунок 2.8 Графовая модель бинарного отношения с численными параметрами связи и геоинформацией

В приведенном на рисунке 2.8 примере визуализируется IP трафик города Нью-Йорк в США. Дуги соединяют его с другими городами мира. Размер вспышек показывает относительный объем IP трафика, так что можно сразу сказать, с каким городом идет более активный обмен информацией.

Ландшафтная визуализация

Ландшафтная визуализация заключается в представлении данных в виде трехмерного ландшафта — столбчатых диаграмм, с индивидуальными высотой и цветом, что позволяет показывать количественные и реляционные характеристики данных и быстро идентифицировать в данных как тенденции и взаимосвязи, так и аномалии. Типичный ландшафтный визуализатор позволяет аналитикам наблюдать за развертыванием данных, фильтровать их на лету, а также генерировать анимационные презентационные ролики. Такой визуализаторделает пространственно-ориентированную информацию доступной для быстрого понимания и осмысливания, как, например, показано на рисунке 2.9.

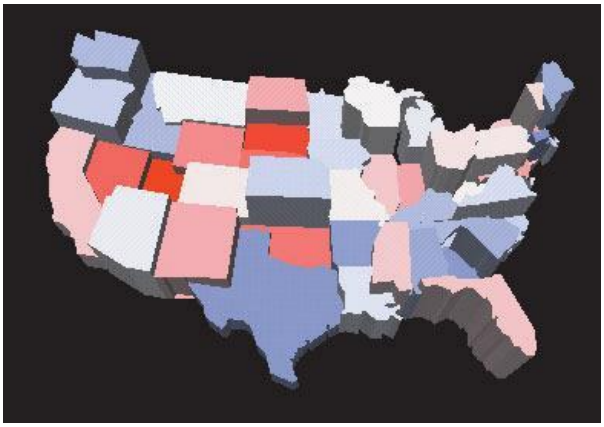


Рисунок 2.9 Ландшафтная диаграмма в геокоординатах

При работе ландшафтного визуализатора, как правило, доступны все операции навигации через ландшафт: панорамирование, вращение в 3D-пространстве, масштабирование с укрупнением интересующей области и т.п. Пользователь может выполнять как операцию приближения, с целью получения детализированной информации о специфических данных, так и операцию подъема (уменьшения масштаба ландшафта), чтобы посмотреть, как данные встроены в окружающую среду.

Визуализация гипотез

Визуализация гипотез позволяет показывать выявленные закономерности, подтверждающие выдвигаемые гипотезы. Представление информации в различных визуализаторах отличается. Например, если строки круговых 3D-диаграмм отображают признаки, использованные классификатором, то каждая круговая диаграмма отражает вероятность того, что величина признака или диапазон значений подходят для классификации. На рисунке 2.10, представленном ниже, анализируется зарплата работающего населения США. Визуализатор отражает атрибуты, которые могут влиять на классификацию по зарплате. Атрибуты представлены рядами круговых трехмерных диаграмм. Высота круговой диаграммы (цилиндра) показывает количество записей в данной категории; цвет показывает, что зарплата больше или меньше 50 тыс. долл. На каждый атрибут может быть несколько круговых диаграмм, например, для обозначения пола (мужской/женский) имеется две диаграммы, а для возраста — восемь диаграмм. Их количество зависит от количества закономерностей,выявленных визуализатором.

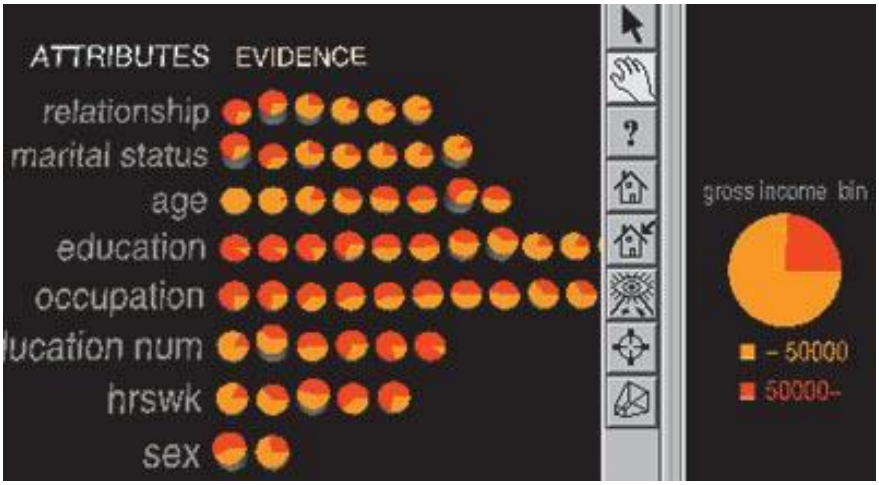


Рисунок 2.10 Диаграмма множественных оценок для поддержки оценивания гипотез

Визуализация деревьев решений

Визуализация деревьев решений позволяет представить иерархически организованную информацию в виде ландшафта и обозревать все множество данных или их часть в виде узлов и ветвей. Ландшафт может быть как двумерным, так и трехмерным. Количественные и реляционные характеристики данных делаются видимыми с помощью иерархически соединенных узлов. В каждом узле находятся числа (двумерный вариант) или гистограммы, высота и цвет которых соответствуют значениям данных (трехмерный вариант). Линии, соединяющие узлы, показывают взаимосвязи. (См. рисунок 2-11).

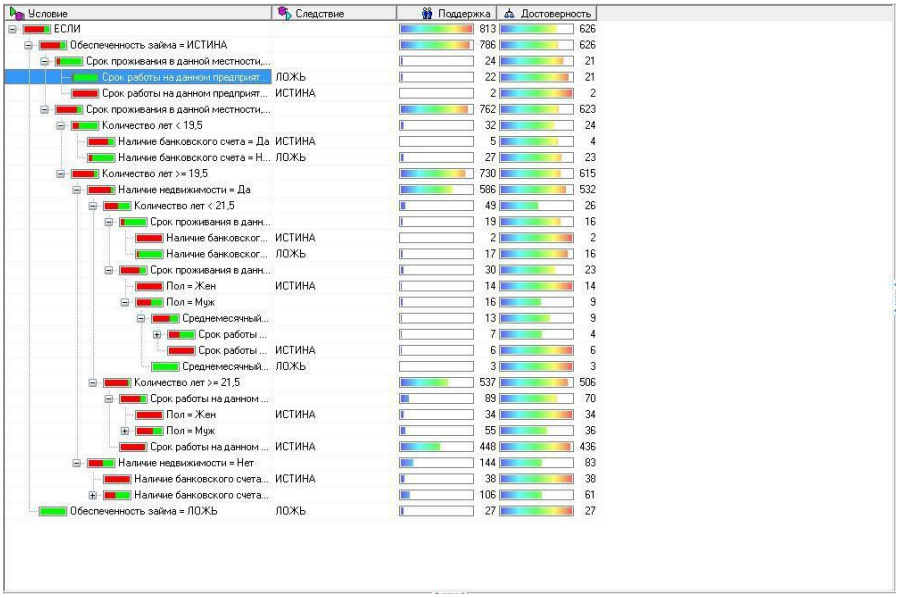


Рисунок 2.11 Диаграмма дерева решений

Визуализатор позволяет анализировать весьма сложные деревья решений. Каждый узел в дереве отражает точку принятия решения. В зависимости от того, как модель оценивает данные по отношению к решению, выбирается ответвление. На рисунке 2.12 показана работа такого визуализатора при обработке зависимостей возраст-доход.

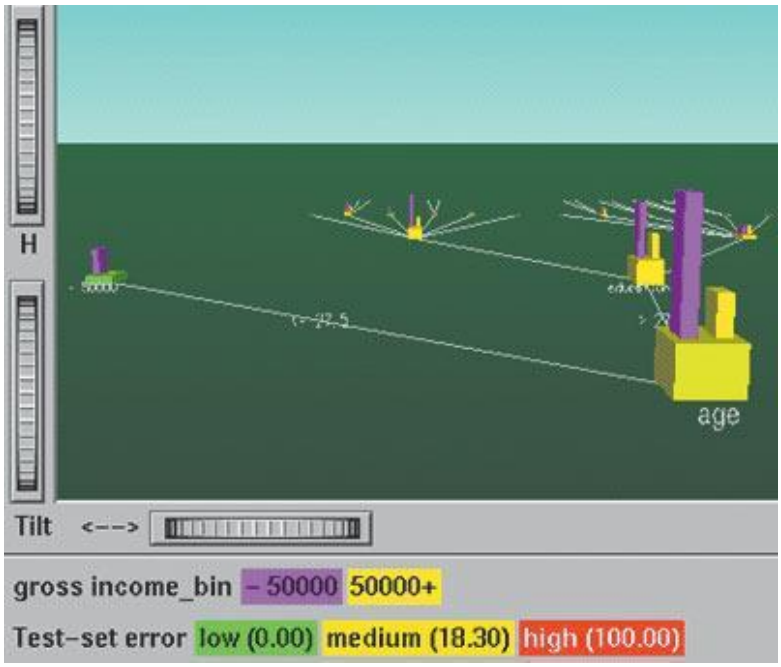


Рисунок 2.12 Диаграмма графа решений

Многомерная визуализация

Набор данных иногда слишком сложен для представления его в двумерном или даже трехмерном представлении. В этом случае удобно обратиться к системам многомерной визуализации данных. К обычной 3D-системе координат добавляются дополнительные

измерения, ассоциированные, например, с размером и цветом элементов данных (рисунок 2.13).

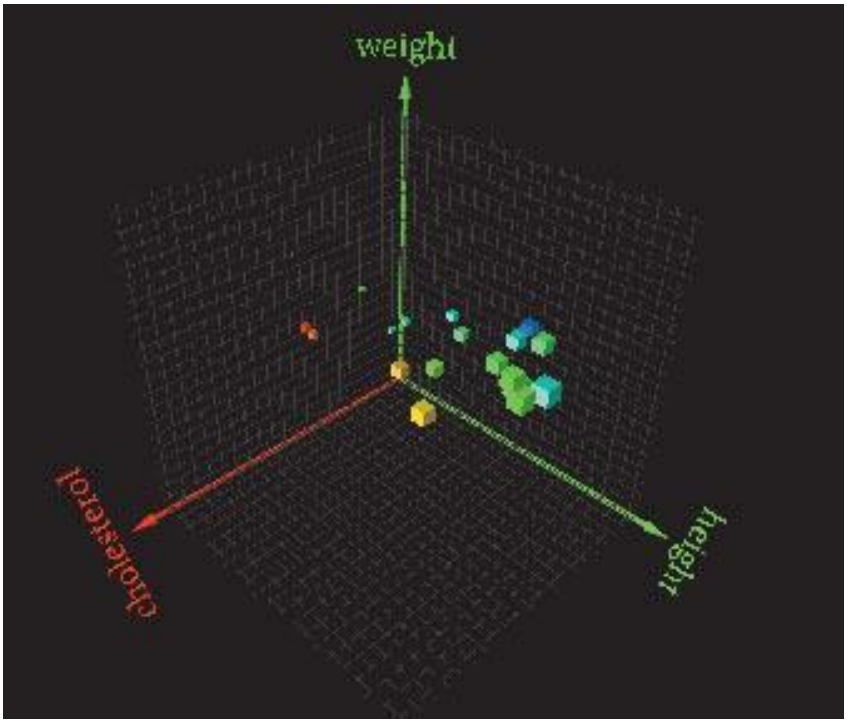


Рисунок 2.13 Диаграмма представления данных в пятимерном пространстве (четвертое измерение задается цветом, а значения – размером элемента)

Следующая группа техник больших данных – самая многочисленная и разнообразная – относится к извлечению данных (Data Mining). Здесь используются разнообразные математические методы и алгоритмы. Часть из них опирается на изощренные математические исследования, а часть использует методы «грубой силы» с некоторыми эвристиками, позволяющими сократить число переборов при определении результата.

Тема 5. Классификация

Техника классификации является одной из базовых методик интеллектуального анализа больших данных. Ее нередко используют при построении модели аналитических систем наряду с еще одной техникой - кластеризацией.

Классификация - это распределение объектов (наблюдений, событий) исследования по *заранее известным* классам на основании сходства признаков.

В отличие от классификации кластеризация производит распределение объектов (наблюдений, событий) по *неизвестным заранее* классам.

Классификация производится в соответствии с принципами машинного обучения с учителем (Supervised Machine Learning). Для проведения классификации с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат классификации. Каждый объект (запись базы данных) должен содержать информацию о некоторых признаках объекта. Процесс классификации, как правило, сводится к следующим шагам.

1. Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое. Обучающее множество - множество, которое включает данные, использующиеся для конструирования модели. Множество содержит входные и выходные (целевые) значения примеров. Выходные значения предназначены для обучения модели. Тестовое множество также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки модели.

2. Каждый объект набора данных относится к одному предопределенному классу. На этом этапе используется обучающее множество, на нем происходит конструирование модели. Полученная модель представляется классификационными правилами, деревом решений или математическими формулами.

3. Производится оценка правильности модели. Известные значения из тестового множества сравниваются с результатами использованияполученной модели. Вычисляется уровень точности - процент правильно классифицированных объектов в тестовом множестве.

Задачей классификации часто является предсказание категориальной зависимой пременной на основе выборки непрерывных и/или категориальных переменных. Например, можно предсказать, кто из клиентов энергетической компании является потенциальным покупателем определенного вида услуг, а кто - нет, кто перейдет на новый тарифный план, а кто - нет, и т.д. Это задачи бинарной классификации, в них зависимая переменная может принимать только два значения (например, да или нет, 0 или 1).

Если зависимая переменная может принимать значения из некоторого множества предопределенных классов, то такая классификация является мультивариантной. Такая классификация используется, например, когда желательно определить, какой тарифный план захочет выбрать клиент. В этом случае рассматривается множество классов для зависимой переменной. Количество признаков объекта, характеризующих категориальную переменную, может быть различным. Одномерные данные содержат только один признак для каждого объекта. Эти данные позволяют сравнивать признаки, устанавливать насколько значения отличаются друг от друга, обращать особое внимание на объекты с нетипичными значениями. Примером одномерных данных является информация о среднем расходе электроэнергии в регионе по районам.

Она позволяет назвать район с самым высоким уровнем энергопотребления, понять, насколько отличаются уровни средних расходов в различных районах друг от друга, обратить внимание на районы, где уровень расхода самый высокий. Двумерные данные содержат информацию о двух признаках для каждого из объектов. Кроме того, что они дают возможность получить два набора одномерных данных, двумерные данные позволяют установить, существует ли связь между двумя признаками, насколько сильно связаны признаки, можно ли предсказать значение одного признака по значению другого и если да, то с какой надежностью.

Например, данные опроса предприятий о том, удовлетворены ли они уровнем энергоснабжения и уровнем энергопотребления, позволяют установить, есть ли связь между уровнями энергоснабжения и энергопотребления на предприятии.

Многомерные данные содержат информацию о трех или более признаках для каждого объекта. В дополнение к той информации, которую можно извлечь из одномерных и двумерных наборов, многомерные данные можно использовать для получения информации о том, существует ли зависимость между признаками, насколько они взаимосвязаны (речь идет не только о попарной взаимосвязи признаков, но и о зависимости в совокупности), можно ли предсказать значение одной переменной на основании значений остальных.

Примером многомерных данных является список клиентов энергетической компании, в котором перечислены уровни энергопотребления, численность сотрудников, годовой оборот средств и профиль.

Таблица 2.1

№	Клиент	Численность сотрудников	Годовой оборот (млн. руб.)	Энергопотребление (тыс. кВт/ч)	Профиль
1	ООО “ВторРесурс”	57	200	100	производство
2	АО “Семь”	32	200	100	производство
3	Адм. Локотского р-на	89	0	8	гос. учреждение
4	ООО “Роснабор”	49	20	12	учебное зав-е
5	ТСЖ “Высотка”	4	5	30	услуги
6	ОАО “Механик”	27	20	400	производство
7	ООО “Настрой-ММ”	36	1	300	производство
8	ООО “Сухпай”	37	5	40	производство
9	ОАО “Рыба”	83	7	400	производство
10	ООО “Водосток”	13	5	50	торговля
11	ООО “Кривошип”	14	3	9	производство
12	ИП “Иванов”	9	2	100	производство
13	ООО “Локуст”	1134	50	1000	производство
14	ООО “Микротех”	21	22	100	производство

При анализе таблицы несложно установить связь между профилем иэнергопотреблением. Если профиль “производство”, то энергопотреблениевысокое. Значения признаков, которые выражаются с помощью чисел, имеющихсодержательный смысл, называют количественными данными. Данные,приведенные в столбцах “Численность сотрудников”, “Годовой оборот” и“”Энергопотребление” таблицы, являются количественными. В то же время числа,приведенные в первом столбце этой таблицы, не являются количественнымиданными: они только указывают на номер клиента в реестре. Эти числа не имеютсодержательной интерпретации. С количественными данными можно выполнятьвсе обычные операции над числами, такие, как вычисление среднего и оценкуизменчивости. Если данные регистрируют определенное качество, которымобладает объект, их называют качественными. Даже если значениям этогокачества можно поставить в соответствие числа, то обрабатывать эти числа какколичественные данные нельзя. Пример качественных данных –столбец“Профиль” в вышеприведенной таблице. Возможные значения включают“производство”, “торговля”, “услуги”, “гос. учреждение”. Для классификациииспользуются различные математические методы, называемые «машинноеобучение». Основные из них, которые используются при анализе большихданных:

- классификация с помощью деревьев решений; байесовская (наивная) классификация;
- классификация при помощи искусственных нейронных сетей; классификация методом опорных векторов; статистические методы, в частности, линейная регрессия;
- классификация при помощи метода ближайшего соседа; классификация CBR-методом;

Тема 6.Классификация при помощи генетических алгоритмов

Методы различаются приспособленностью и удобностью для работы с различными типами данных. Одни хорошо подходят для работы с количественными признаками, другие – качественными. Алгоритмы, реализующие быстрые методы, нередко уступают по качеству более медленным. Оценку качества конкретной программной реализации метода классификации принято производить с помощью методики, называемой ROC – кривой (receiver operating characteristic). Эта характеристика называется иначе кривой ошибок и показывает связь доли верных положительных классификаций от общего числа положительных классификаций (TPR - True positive rate) с долей ошибочных положительных классификаций от общего числа отрицательных классификаций (FPR – False positive rate). На рисунке 2.14 приведена ROC для некоторого «хорошего» классификатора, а для сравнения пунктирной прямой линией показана ROC для чисто случайного угадывания.

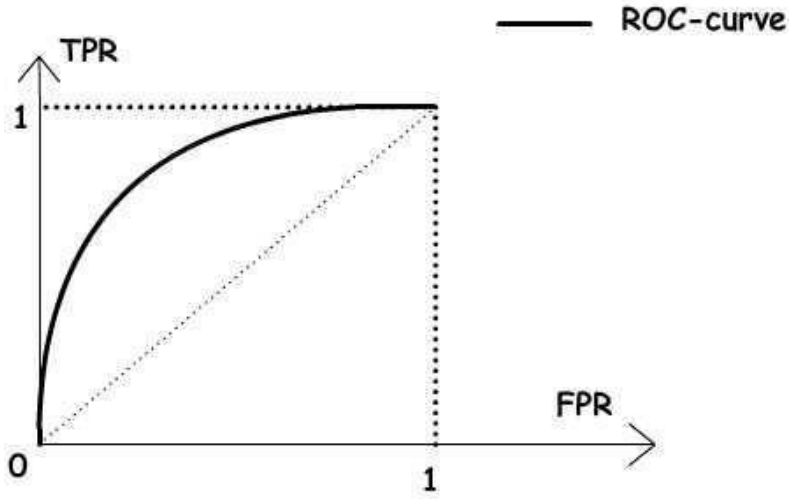


Рисунок 2.14 Кривая ошибок для «хорошего» классификатора (сплошная линия) и чисто случайного угадывания (пунктирная линия).

Агрегированной характеристикой качества классификатора служит площадь под ROC- кривой, называемая AUC (area under curve). Чем больше это значение, тем лучше модель классификации. AUC обычно используется для сравнения моделей классификации.

Рекомендуем заинтересованному читателю неплохой современный обзор методов и алгоритмов классификации [5].

Кластеризация

Техника кластеризации является подходом к классификации данных в случае, когда заранее неизвестно, к какому классу должен быть отнесен любой из имеющихся объектов. Кластеризация осуществляется автоматическим нахождением групп, на которые должны быть разбиты анализируемые объекты. Такой процесс может рассматриваться как машинное обучение без учителя(Unsupervised Machine Learning). Известно более 100 разных алгоритмов кластеризации, мы обратим внимание на наиболее типичных: иерархический кластерный анализ и кластеризация методом k-средних.

Иерархический кластерный анализ включает два вида методов: агломеративные и дивизимные. Самыми распространенными являются ***иерархическиеагломеративные методы.***Сущность этих методов заключается в том,чтовначале каждый объект рассматривается как отдельный кластер. Объединение кластеров происходит пошагово. На

основании матрицы расстояний или матрицы сходства признаков объединяются наиболее близкие объекты. Если матрица сходства первоначально имеет размерность m×m, то через m -1 шагов все объекты будут объединены в один кластер. Множество методов иерархического кластерного анализа различается используемыми мерами близости и алгоритмами объединения. Разница между алгоритмами заключается в способе вычисления близости. Вообще, понятие близости для объектов данных играет ключевую роль для всех методов кластеризации. Критерий близости двух объектов данных как правило оказывается связан с семантическими аспектами «похожести» в анализируемой предметной области. Для читателей, которые интересуются способами количественного вычисления похожести, иногда выражаемых с помощью более формальных понятий расстояния, мы рекомендуем прекрасную и, пожалуй, самую полную на сегодняшний день «Энциклопедию расстояний» [6].

В общем виде алгоритм иерархического кластерного анализа можно представить в виде последовательности процедур:

1. Значения исходных признаков нормируются.
2. Составляется матрица расстояний или матрица мер близости.
3. Находится пара самых близких кластеров.
4. По выбранному алгоритму объединяются эти два кластера.

Пункты 2, 3 и 4 повторяются до тех пор, пока все объекты не будут объединены в один кластер или до достижения заданного "порога" близости. Алгоритмы делятся на две группы: алгоритмы добавления объекта в кластер и алгоритмы объединения двух кластеров. Для включения нового объекта в существующий кластер применяются различные алгоритмы агрегации.

Метод одиночной связи. На основании матрицы расстояний определяютсядва наиболее близких объекта, они и образуют первый кластер. Далее выбирается объект, который будет включен в этот кластер. Таким объектом будет тот, который имеет наименьшее расстояние хотя бы с одним из объектов, уже включенных в кластер. На следующем шаге аналогично включается в кластер следующий объект и так далее до образования единственного кластера.

Метод полных связей.Включение нового объекта в кластер происходиттолько в том случае, если расстояние между объектами не меньше некоторого заданного уровня.

Метод средней связи.Для решения вопроса о включении нового объекта вуже существующий кластер вычисляется среднее значение меры близости, которое затем сравнивается с заданным пороговым уровнем (как в предыдущем методе).

Метод Уорда.Данный метод предполагает,что первоначально каждыйкластер состоит из одного объекта. Сначала объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений:

$$= \sum \sum (\ \ -)^2$$

для каждого номера кластера и всех объектов с номерами $\in (1,)$, $\in (1, \ )$, где i - номер объекта, nl - количество объектов в l - том кластере, j - номер признака, k - количество признаков, характеризующих каждый объект. В дальнейшем объединяются те объекты или кластеры, которые дают наименьшее приращение величины vl .

Для объединения двух кластеров применяются следующие алгоритмы:

Метод дальнего соседа.Степень близости оценивается по степениблизости между наиболее отдаленными объектами кластеров.

Метод ближайшего соседа.Степень близости оценивается междунаиболее близкими объектами этих кластеров.

Метод средней связи.Степень близости оценивается как средняявеличина степеней близости между объектами кластеров.

Метод медианной связи.Расстояние между любым кластеромSи новымкластером, который получился в результате объединения кластеров P и Q, определяется как расстояние от центра кластера S до середины отрезка, соединяющего центры кластеров P и Q.

Противоположны агломеративным по логическому построению процедур классификации ***иерархические дивизимные методы***. Начальным условием дивизимных методов является то, что первоначально все объекты объединены в один кластер. В процессе классификации по определенным правилам постепенно от этого кластера отделяются группы схожих между собой объектов. Таким образом, на каждом шаге количество кластеров возрастает, а мера расстояния между кластерами уменьшается.

Другая группа методов кластеризации – ***кластеризация методомk-средних***. Методы этого вида находят широкое применение из-за простоты реализации на алгоритмических языках и большого быстродействия. Суть кластеризации k-средних в том, что метод стремится минимизировать суммарное квадратичное отклонение объектов кластеров от центров этих кластеров. Говоря более простым языком, это итеративный алгоритм, который делит данное множество объектов на k кластеров, элементы, которых являются максимально приближенными к их центрам, а сама кластеризация происходит за счет смещения этих же центров. Следует оговорить то, что метод k-средних очень чувствительный к выбросам ишуму, которые могут существенно исказить результаты кластеризации. Так что перед кластеризацией, часто необходимо пропустить обрабатываемые данные через фильтры, предназначенные для чистки данных.

Процесс простейшей кластеризации методом k-средних состоит из следующих шагов:

1. Выбрать в пространстве объектов точки, которые будут центроидами (точками в центре кластера) соответствующих k кластеров. Выборка начальных центроидов может быть как случайной так и по определенному алгоритму.
2. В цикле, который продолжается до тех пор, пока центроиды кластеров не перестанут изменять свое положение, оцениваем каждый объект и смотрим, к какому центроиду какого кластера он является близлежащим.
3. Если найден близлежащий центроид, привязываем объект к кластеру этого центроида.
4. Когда все объекты перебраны, высчитываем новые координаты центроидов k кластеров.
5. Проверяем координаты новых центроидов. Если они соответственно равны предыдущим центроидам — выходим из цикла, кластеризация завершена, если нет, возвращаемся к пункту 2.

На рисунке 2.15 представлена последовательность преобразований методом k-средних в визуализированной форме

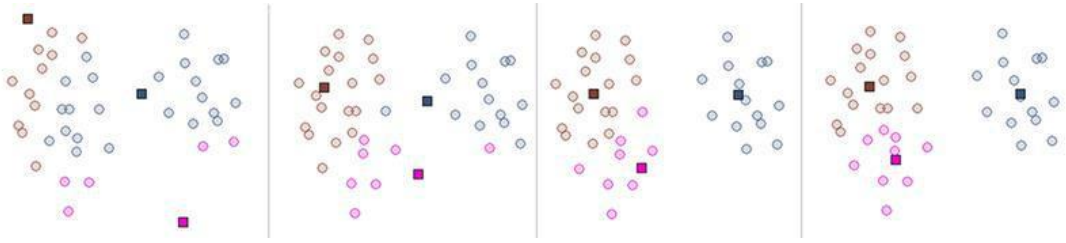


Рисунок 2.15 Графическое представление шагов преобразования данных при кластеризации

При работе с большими данными получил распространение алгоритм кластеризации, названный ***Affinity propogation*** (AP). В отличие от описанных выше он основан на концепции передачи сообщений (message passing) между точками, ассоциированными с

анализируемыми объектами данных. Алгоритм AP не требует предварительного знания числа кластеров, как требует метод k-средних. Отметим его главные особенности и опишем основные шаги.

Пусть кластеризуемый набор объектов данных может быть ассоциирован с множеством точек, произвольной внутренней структуры { } , для любой пары которых определена функция (. , .) похожести, близости (similarity), так , что если (,) > (,) ,то это означает,что точка более похожа на точку,чемточка .выполнение алгоритма сводится к итеративному выполнению двух шагов передачи сообщений между точками для обновления двух формируемых алгоритмом матриц:

– матрица ответственности с элементами (,) значения которых показывают насколько хорошо подходит для того, чтобы служить как экземпляр по сравнению с другими кандидатами экземпляров для .

– матрица доступности с элементами (,) значения которых представляют насколько подходяще было бы для взять как его экземпляр, принимая во внимание предпочтительность других точек как такой экземпляр.

Начальное значение этих матриц берется нулевым и матрицы могут быть представлены как таблицы логарифмов вероятностей, для удобства операций.

Шаги обновления матриц следующие:

(,) ← (,) − max{ (, ′) + (, ′)}
′≠

(,) ← min(0, (,) + ∑ max(0, (′,))) ≠
′∉{ , }

(,) ← ∑ max(0, (′,))
′≠

Агоритм AP является относительно новым, но представляяется весьма перспективным. Его авторы показали необычайную эффективность его работы в системе распределенных вычислений и хранения данных [7]

Выбор метода кластеризации зависит от набора данных, к которому он применяется, и тех требований, которые предъявляются к точности и скорости работы алгоритма. В разных ситуациях предпочтения отдаются различным подходам. Более подробно с особенностями методов рекомендуется знакомиться в специальной литературе. Можем рекомендовать для входа в предметную область неплохой обзор [8].

Регрессионный анализ

Регрессионный анализ — техника моделирования данных, направленная на исследование их взаимосвязи. В простейшем случае регрессионный анализ используют для построения моделей прогнозирования новых числовых значений на основе набора известных значений. Регрессионные модели позволяют, в частности, выявить особенности функционирования конкретного объекта и на их основе предсказывать будущее поведение объекта. Данные в регрессионном анализе делятся на пары: зависимая переменная - независимая переменная. Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Основная задача построения модели – настроить параметры таким образом, чтобы модель наилучшим образом приближала данные. Критерием качества приближения (целевой функцией)обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента. Если модель описывается линейным уравнением – то она называется линейной регрессией. В зависимости от количества независимых переменных, регрессия может быть простой или множественной.

Простая линейная регрессия записывается уравнением:

y=a+bx,

визуально представляемым графиком прямой, как на рисунке 2.16.

x называется независимой переменной или предиктором;

y – зависимая переменная или переменная отклика,это значение,которое мыожидаем для *y*, если мы знаем величину *x*;

a –свободный член линии оценки;это значение *y*,когда *x=0;*

b –угловой коэффициент или градиент оценённой линии;она представляет собойвеличину, на которую *y* увеличивается в среднем, если мы увеличиваем *x* на одну единицу.

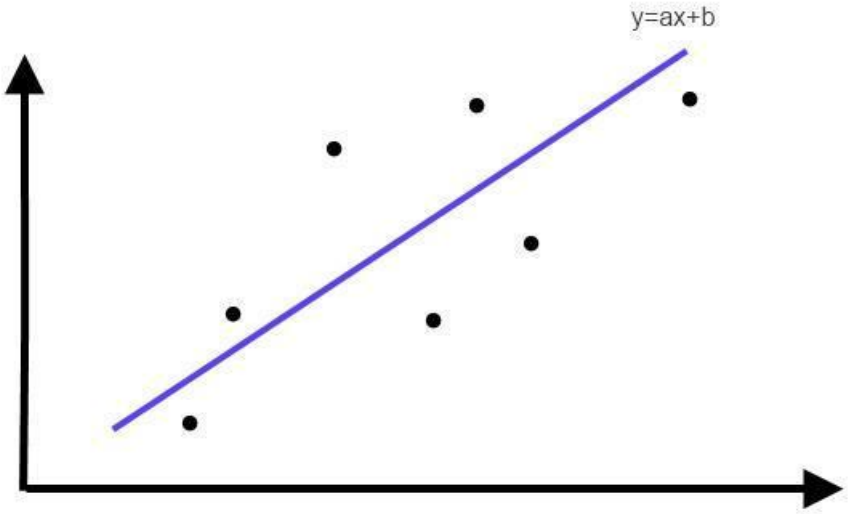


Рисунок 2.16 График линейной регрессии

Величина *x* называется независимой переменной или предиктором;y– зависимая переменная или переменная отклика, это значение, которое мы ожидаем для y, если мы знаем величину *x*;*a*– свободный член линии оценки; это значение *Y*, когда *x=0;b*– угловой коэффициент или градиент оценённой линии; она представляет собой величину, на которую *y* увеличивается в среднем, если мы увеличиваем *x* на одну единицу.

Если зависимая переменная зависит отчем более одной независимой переменной, то регрессионную модель называют *множественной линейнойрегрессией* -

$$Y = b0 + b1X1 + b2X2 + \ldots + bnXn$$

Построение модели заключается в подборе коэффициентов *b0,b1,b2* и т.д . Этот процесс называется «обучением» регрессионной модели. Для обучения можно в простейшем случае найти решение системы линейных уравнений с неизвестными *b0,b1,b2*:,образующейся при объединении соотношений$Y = b0+b1X1+ b2X2$длям«обучающих» наборов данных , взятых из реальных экспериментов или компьютерного моделирования

$\{ {}^{1}, {}_{1}{}^{1}, {}_{2}{}^{1} \ldots \}; \{ {}^{2}, {}_{1}{}^{2}, {}_{2}{}^{2} \ldots \}; \ldots\ldots\{ \ , {}_{1} , {}_{2} \ldots \};$

Решение системы возможно различными способами. Наиболее часто применяют метод минимизации среднего квадрата разностей между имеющимися данными и прогнозом значения, получаемого из модели. На рисунке 2.17 проиллюстрирован этот подход.

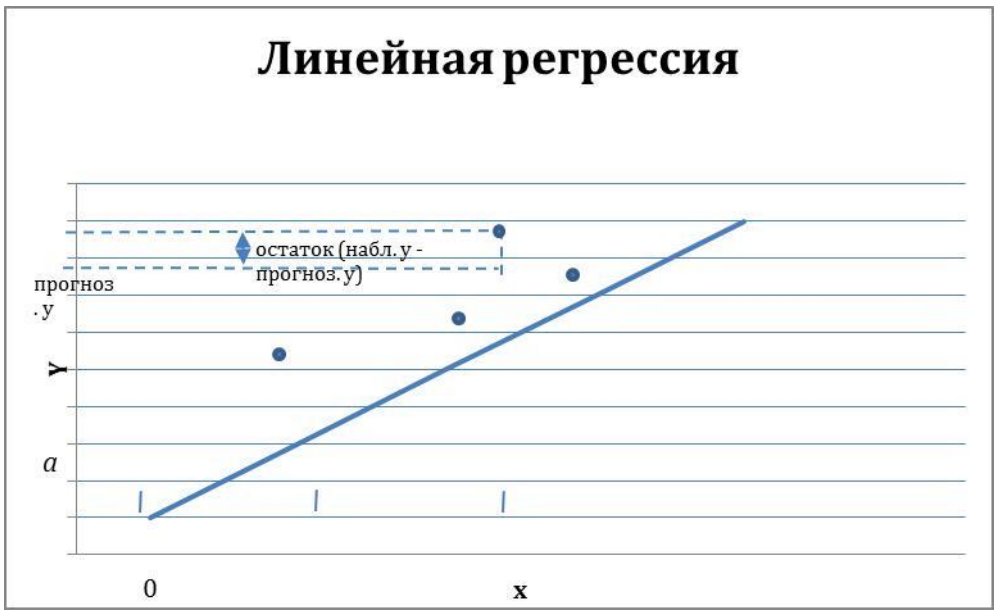


Рисунок 2.17 Подбор коэффициентов регрессии по минимуму суммы квадратов отклонений прогноза и имеющихся данных.

При построении модели важным фактором является избавление или, наоборот, включение выбросов. Выброс – это точка, координаты которой значительно противоречат большинству значений координат в наборе данных. Очистка данных от выбросов может кардинально изменить один или больше параметров модели (т.е. угловой коэффициент или свободный член). Если нет уверенности в незначимости имеющихся в данных выбросов, исследуют модели, как с их учетом, так и по очищенным данным. Нужно ли учитывать выброс решают после рассмотрения изменения коэффициентов регрессии. При построении линейной регрессии всегда проверяется гипотеза о том, что угловой коэффициент линиирегрессии равен нулю. То есть имеет ли смысл регрессионная модель. Для проверки гипотезы может быть использован любой из алгоритмов проверки гипотез из статистического анализа.

Области применения регрессионного анализа весьма разнообразны. Отметим сначала одну из важных – прогнозирование значений некоторого временного ряда. Типичными примерами таких рядов являются температура воздуха в данном месте, биржевые курсы акций, показания интеллектуального измерителя в энергосети, средства на счете и т.п.

В этой задаче в качестве зависимой переменной рассматривается прогнозируемое значение величины в еще не наступившие моменты времени, а в качестве предикторов нередко рассматриваются значения этой величины в уже прошедшие моменты. Таким образом модель линейной регрессии для прогнозирования временного ряда, то есть значения переменной y в еще не наступивший момент времени $t+1$ на основании знания значений этой переменной в прошлые моменты времени t, будет выглядеть следующим образом

$$+1 = 0 + 1 \quad -1 + 2 \quad -2 + 3 \quad -3 + \cdots$$

Очевидно, что прогнозирование сводится к правильному подбору коэффициентов . Модель такого вида называется авторегрессионной моделью. Применяют и более сложные модели для прогноза, учитывающие кроме прошлых значений анализируемой величины и внешние предикторы. Так, если в регрессию включить еще набор значений внешнего фактора , то модель, называемая в этом случае моделью авторегресии-скользящего среднего (ARIMA), будет выглядеть таким образом

$$+1 = 0 + 1 \quad -1 + 2 \quad -2 + 3 \quad -3 + \cdots + 0 \quad + 1 \quad -1 + \quad 2 \quad -2 + \cdots$$

Для использования такой модели также необходимо иметь эффективную процедуру нахождения коэффициентов.

Известны и другие более сложные модели для прогнозирования значений временных рядов. Алгоритмы, решающие эту задачу называют предсказательными или предиктивными (predictive algorithms). Чрезвычайно высокая их распространенность в индустриальных системах аналитики и управления привела к необходимости стандартизации описаний алгоритмов и портируемости их программных реализаций из одной системы в другую. Для этих целей был разработан специальный язык PMML – predictive model markup language. PMML обеспечивает простое использование предсказательных аналитических моделей несколькими приложениями одновременно. Так, можно обучить модель в системе, применяемой для планирования выхода годного продукта с линии сборки, выразить ее в PMML, настроить, оттестировать в среде разработки, а затем быстро перенести в другую систему, где использовать, например, для предсказания выработки продукта в другом производственном процессе. Кроме того, PMML — исключительно гибкий стандарт, рассчитанный на удовлетворение потребностей современных специалистов — бизнес-аналитиков,применяющих передовые методы моделирования. Например, поскольку предсказательные модели рассчитаны на решение конкретных проблем и их преимущества проявляются только при применении по прямому назначению, в сложных случаях одной модели для поиска решения будет недостаточно. PMML позволяет строить приложения с несколькими моделями, включая ансамбли моделей. Каждую модель можно экспортировать в PMML и выполнить в рабочей среде так, как запланировано в приложении. В настоящее время многие поставщики программного продукта для предиктивной аналитики предоставляют PMML описание своих алгоритмов. Чтобы показать читателям насколько велико уже множество таких поставщиков и какие алгоритмы уже реализованы и описаны, мы решили привести в книге таблицу, содержащую такую полезную информацию.

Таблица. 2.1

Company / Project	**Software**	**PMML Producer**	**PMML Consumer**	**Supported Model Type**
	KnowledgeSTUDIO	PMML 3.2		Decision Trees Regression Models (Linear and Logistic) Neural Networks Clustering Models Rule Set Models (Scorecards)
	KnowledgeSEEKER	PMML 3.2		Decision Trees
	StrategyBUILDER	PMML 3.2		Decision Trees (Strategy Trees)

	Augustus 0.4x	PMML 4.0	PMML 4.0	Decision Trees Regression Naïve Bayes Baseline (With Segmentation on all of the above)
	Augustus 0.5x	PMML 4.1	PMML 4.1	Baseline Clusters Naïve Bayes Regression RuleSet Trees (With Segmentation on all of the above)
	Augustus 0.6x	PMML 4.1	PMML 4.1	Baseline Clusters Trees (With Segmentation on all of the above)
	BigML Public API	PMML 4.1		Decision Trees (classification and regression)
	Strategy Tree Optimization	PMML 3.0, 3.1	PMML 3.0, 3.1	Decision Tree
	PowerCurve™ Strategy Management	PMML 3.0, 3.1, 4.0, 4.1	PMML 3.0, 3.1, 4.0, 4.1 PMML 3.0, 3.1, 4.0, 4.1	Decision Tree Regression Model
	IBM SPSS Statistics 21	PMML4.0	PMML2.0 through 4.0	Clustering Models Decision Trees

Кроме линейных моделей регресии применяют и нелинейные. В ряде случаев они позволяют получить более компактные описания, с мпньшим числом параметров регрессии. Существует несколько классов нелинейных регрессий:

1) регрессии, нелинейные по предикторам, но линейные по коэффициентам, например, полиномы $=_1+_2{}^2+_3{}^3+\cdots$ или гиперболы $= +$

2) регрессии, нелинейные по коэффициентам

В качестве примера первого типа нелинейной регрессии можно привести привести связь между временем доставки груза в транспортной компании и оценкой пробок на улицах города компанией Яндекс. Нелинейную регрессию второго класса рассмотрим более подробно на примере весьма важной *логистическойрегрессии* или логит-регрессии.

Эта регрессия определяет значение зависимой переменной *Y* с помощью нелинейной функции особого вида, показанной на рисунке 2.18 и называемой логистической

$$= \frac{1}{1 + \quad -(\ _0+ \ _1 1+ \ _2 2+ _3 3+\cdots)}$$

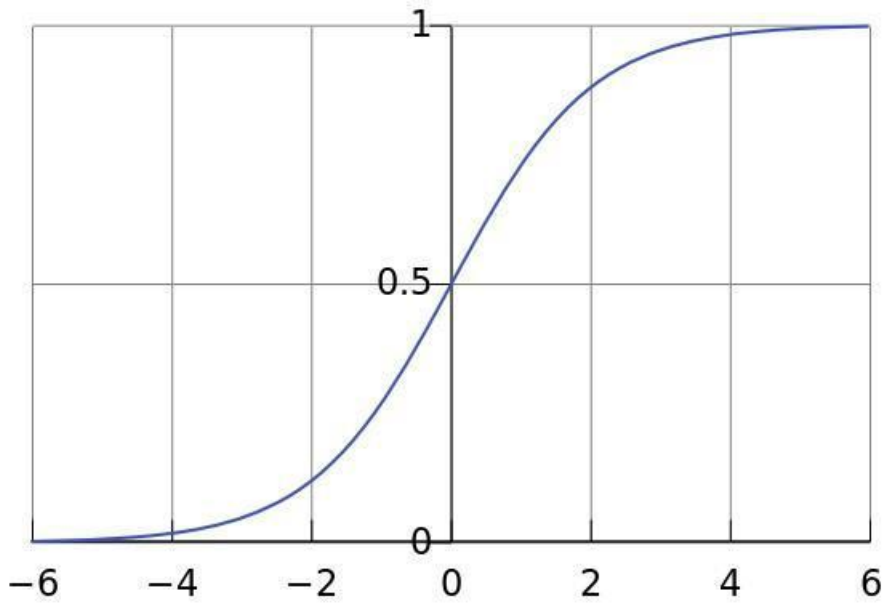


Рисунок 2.18. Логистическая кривая, определяющая логистическую регрессионную модель по одной переменной

Как видно, переменная *Y* принимает значения из интервала (0,1) и обычно используется как оценка вероятности некоторого события, например, принадлежности набора переменных (предикторов) некоторому классу. В этом случае логистическая регрессия может использоваться как алгоритм для классификации данных. Обучение модели сводится к определению таких коэффициентов , для которых на обучающей выборке принадлежащие анализируемому классу наборы данных порождают существенно большиезначения Y по сравнению с наборами данных не относящихся к этому классу. Какв других алгоритмах классификации для подбора коэффициентов модели используют метод ROC- кривых.

Приведем пример применения логистической регрессионной модели в мониторинге режимов энергосети. Пусть предикторы определяют значения мощности в некоторых фидерах сети и их значения измеряются интеллектуальными измерителями и поступают в аналитический центр. Некоторые распределения (наборы) значений мощностей с высокой вероятностью приводят к аварийным отключениям в сети, и задача предупреждения аварий такого типа может быть сведена к обнаружению таких сочетаний мощностей в выбранных фидерах. С помощью компьютерного моделирования сети или на основании изучения реальных данных о процессах в сети, приведших к аварийным отключениям, можно обучить модель, то есть найти такую совокупность коэффициентов , которая обеспечивает максимизацию значений *Y* для предаварийных ситуаций по сравнению со значениями,соответствующимипроцессам не приводящим к аварии. Технике

регрессионных моделей посвящено большое число работ, обзор наиболее интересных из них приведен в [9]

Тема № 7 Анализ ассоциативных правил

Техника анализа ассоциативных правил является одной из распространенных методик интеллектуального анализа данных. Ее используют для построения модели системы, знания о которой могут быть извлечены путем обнаружения закономерностей между связанными объектами.

Ассоциативные правила - это правила для количественного описания взаимной связи между двумя или более событиями.

Примерами приложения ассоциативных правил могут быть следующие задачи:

–*выявление наборов услуг*,которые часто заказываются вместе или никогда незаказываются вместе;

–*определение доли клиентов*,положительно относящихся к изменениям втарифных планах;

–*определение профиля потребителей электроэнергии*;определение долислучаев, в которых задолженность клиента приводит к убыткам кредитора.

Для описания их решения разработана специальная терминология. Важнейшими понятиями являются транзакция и предметный набор. Транзакция – это некоторое множество событий, происходящих совместно. Предметный набор - непустое множество предметов, появившихся в одной транзакции. В примере с формированием и приобретением клиентом пакета услуг, транзакцией будут события формирования и приобретения, а предметным набором – список услуг, входящих в приобретенный пакет.

Ассоциативное правило состоит из двух предметных наборов, называемых условие и следствие, записываемых в виде $X \rightarrow Y$, что читается следующим образом: «Из X следует Y». Таким образом, ассоциативное правило формулируется в виде: «Если <условие>, то <следствие>». Правила обычно отображаются с помощью стрелок, направленных от условия к следствию, например, услуга1 → услуга3.Условие и следствие иногда называются соответственно левосторонним и правосторонним компонентами ассоциативного правила.

Ассоциативные правила имеют характеристики, отображающие особенности связи между условием и следствием. В основном оперируют двумя характеристиками — поддержкой (записывается как supp или S) и достоверностью (записывается как conf или C).

Поддержка – это отношение количества транзакций, содержащих как условие, так и следствие, к общему числу транзакций.

$$S(A \rightarrow B) = P(A \cap B) = \frac{\text{количество транзакций, содержащих A и B}}{\text{общее количество транзакций}}$$

Достоверность – это мера точности ассоциативного правила, она определяется как отношение количества транзакций, содержащих условие и следствие, к количеству транзакций, содержащих только условие

$$C(A \to B) = P(B \mid A) = P(B \cap A) / P(A) = \frac{\text{количество транзакций, содержащих A и B}}{\text{количество транзакций, содержащих только A}}.$$

Вычислив характеристики связи, можно делать аналитические заключения. Если поддержка и достоверность достаточно высоки, можно утверждать, что любая будущая транзакция, которая включает условие, будет также содержать и следствие.

Рассмотрим пример.

Таблица 1– Пример набора транзакций

№	**Транзакция**
1	Передача данных, электроснабжение, мобильная нотификация
2	Телевидение, интернет
3	Интернет
4	Безакцептная оплата, визуализация расходов, мобильная нотификация, услуги курьера, интернет
5	Безакцептная оплата, электроснабжение, интернет, мобильная нотификация
6	Телевидение, мобильная нотификация
7	Электроснабжение, мобильная нотификация
8	Электроснабжение, визуализация расходов, мобильная нотификация, интернет
9	Безакцептная оплата, передача данных, визуализация расходов, мобильная нотификация, интернет

Возьмем ассоциацию электроснабжение → мобильная нотификация. Поскольку количество транзакций, содержащих как электроснабжение, так и мобильную нотификацию, равно 4, а общее число транзакций — 9, то поддержка данной ассоциации будет:

S(электроснабжение→мобильная нотификация) = 4 / 9 = 0,44.

Поскольку количество транзакций, содержащих только электроснабжение (условие), равно 4, то достоверность данной ассоциации будет:

C(электроснабжение→мобильная нотификация) = 4 / 4 = 1.

Иными словами, все транзакции, содержащие электроснабжение, также содержат и мобильную нотификацию, из чего делаем вывод о том, что данная ассоциация может рассматриваться как правило. Все клиенты, заказывающие электроснабжение, также пользуются услугой мобильной нотификации.

Методики поиска ассоциативных правил конструируются так, чтобы обнаруживать все ассоциации, которые удовлетворяют ограничениям на поддержку и достоверность, введенным аналитиком. При анализе больших данных это приводит к необходимости рассматривать десятки и сотни тысяч ассоциаций. Обработка такого количества правил вручную нереальна. Число правил необходимо уменьшить. Однако характеристик поддержки и достоверности недостаточно, чтобы сузить круг важных для системы ассоциаций. Требуется предусмотреть фильтрацию правил по значимости.

Значимость тем выше, чем выше зависимость предметных наборов. Если условие и следствие независимы, то поддержка правила примерно соответствует произведению

поддержек условия и следствия, то есть *S(A→B)* ≈ S*(A)*S(B)*. Это значит, что хотя условие и следствие часто встречаются вместе, не менее часто они встречаются и по отдельности. Например, если услуга *1* встречался в 90 транзакциях из 100, а услуга *2* — в 60 и в 50 транзакциях из 100 они встречаются вместе, то несмотря на высокую поддержку (*S(услуга1→услуга2) = 0,5*) это не обязательно правило. Просто эти услуги покупаются независимо друг от друга, но в силу их популярности часто встречаются в одной транзакции. Поскольку произведение поддержек условия и следствия *S(услуга1)*S(услуга2)* = 0,9 · 0,6 = 0,54 отличается от *S(услуга1→ услуга2)* = 0,5 всего на 0,04, предположение о независимости услуг *1* и *2* достаточно обоснованно. Если условие и следствие независимы, то правило вряд ли представляет интерес, даже если его поддержка и достоверность высоки. Ассоциация малозначима для модели системы. Для измерения значимости применяют субъективные характеристики связи, которые называются интерес (lift) и уровень (leverage).

Интерес — это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом. Значения интереса большие, чем единица, показывают, что условие чаще появляется втранзакциях, содержащих следствие, чем в остальных. Можно сказать, что интерес является обобщенной мерой связи двух предметных наборов: при значениях лифта > 1 связь положительная, при 1 она отсутствует, а при значениях< 1 — отрицательная. Интерес вычисляется следующим образом:

L(*A* →B) = C(A→B) / *S*(*B*).

Рассмотрим ассоциацию мобильная нотификация → электроснабжение из таблицы 1.

S(*электроснабжение*) = 4/9 = 0,44;*C*(*мобильная нотификация → электроснабжение*) = 4/7 = 0,57.

Следовательно, *L*(*мобильная нотификация→электроснабжение*) = 0,57/0,44 = 1,295.

Теперь рассмотрим ассоциацию мобильная нотификация → интернет.

S(*интернет*) = 6/9= 0,67;*C*(*мобильная нотификация → интернет*) = 4/7 =0,57.

Тогда *L*(*мобильная нотификация→интернет*) = 0,57/0,67 = 0,85.

Большее значение интереса для первого правила показывает, что подключение мобильной нотификации больше влияет на покупку электроснабжения, чем интернета. Хотя интерес используется широко, он не всегда оказывается удачной мерой значимости правила. Правило с меньшей поддержкой и большим интересом может быть менее значимым, чем альтернативное правило с большей поддержкой и меньшим интересом, потому что последнее применяется для большего числа транзакций. Увеличение числа транзакций приводит к возрастанию значимости связи между условием и следствием. Чтобы учитывать эту значимость введено понятие уровня. Уровень — это разность между наблюдаемой частотой, с которой условие и следствие появляются совместно (то есть поддержкой ассоциации), и произведением частот появления (поддержек) условия и следствия по отдельности:

T(*A* → *B*) = *S*(*A* → *B*)– *S*(*A*)*S*(*B*).

Рассмотрим ассоциации визуализация расходов → мобильная нотификация и электроснабжение → мобильная нотификация, которые имеют одинаковую поддержку C = 1, поскольку визуализация расходов и электроснабжение всегда продаются вместе с мобильной нотификацией (см. таблицу 1). Интересы для данных ассоциаций также будут одинаковыми, поскольку в обеих ассоциациях поддержка следствия S(мобильная нотификация) = 7/9 = 0,77.

Тогда L(визуализация расходов → мобильная нотификация) = L(электроснабжение → мобильная нотификация) = 1/0,77 = 1,3.

S(визуализация расходов → мобильная нотификация) = 3/9 = 0,33; S(визуализация расходов) = 0,33; S(мобильная нотификация) = 0,77.

Таким образом, T(визуализация расходов → мобильная нотификация) = 0,33 – 0,33 · 0,77 =0,08.

S(электроснабжение → мобильная нотификация) = 0,44; S(электроснабжение) = 0,44; S(мобильная нотификация) = 0,77.

Следовательно, T(электроснабжение → мобильная нотификация) = 0,44 – 0,44 · 0,77 =0,1.

Итак, значимость второй ассоциации больше, чем первой. Правило электроснабжение → мобильная нотификация важнее для модели, так как оно встречается чаще, то есть применяется для большего числа транзакций. Достоверность, поддержка, интерес и уровень используются для последующего ограничения набора рассматриваемых ассоциаций путем установки порога значимости, ниже которого ассоциации отбрасываются. Существует целый ряд алгоритмов, позволяющих делать это во время обнаружения правил ассоциации. Простейший алгоритм поиска ассоциативных правил перебирает все возможные комбинации условий и следствий, оценивает для них поддержку и достоверность,

а затем исключает все ассоциации, которые не удовлетворяют заданным ограничениям. Число возможных ассоциаций с увеличением числа предметов растет экспоненциально. Если в базе данных транзакций присутствует k предметов и все ассоциации являются бинарными (то есть содержат по одному предмету в условии и следствии), то потребуется проанализировать

$k2^{-1}$ассоциаций. Поскольку реальные базы данных транзакций, рассматриваемые при интеллектуальном анализе данных, обычно содержат тысячи записей, вычислительные затраты при поиске ассоциативных правил огромны.

Чтобы обеспечить приемлемую эффективность, в процессе генерации ассоциативных правил широко используются алгоритмы, позволяющие уменьшить количество ассоциаций, которое требуется проанализировать. Одной из наиболее распространенных является методика, основанная на обнаружении так называемых частых наборов, когда анализируются только те ассоциации, которые встречаются достаточно часто. Эта методика носит название Apriori.

Алгоритм Apriori предусматривает несколько шагов:

Предобработка данных:все данные приводятся к бинарному виду;типданных изменяется.

Выявление часто встречающихся наборов элементов:подсчитываются1-элементные часто встречающиеся наборы.

Генерация наборов элементов: генерируются потенциально часто встречающиеся наборы элементов (их называют кандидатами) и подсчитываются поддержки для кандидатов.

Проверка кандидатов:удовлетворяют ли значения поддержки кандидатовминимальному порогу.

Извлечение правил:чтобы извлечь правило из часто встречающегосянабора F, следует найти все его непустые подмножества; для каждого подмножества M мы сможем сформулировать правило $M \Rightarrow (F - M)$, если достоверность правила $C(M \Rightarrow (F - M)) = S(F)/S(M)$ не меньше установленного порога достоверности.

Более подробно с реализацией алгоритма можно познакомиться в специальной литературе [10] .

Алгоритм Apriori замедляется процессом генерации кандидатов в популярные предметные наборы. Кроме этого, он требует многократного сканирования базы данных транзакций, а именно столько раз, сколько предметов содержит самый длинный предметный набор. Поэтому был предложен ряд алгоритмов, которые позволяют избежать генерации кандидатов и сократить требуемое число сканирований набора данных.

Одним лучших методов поиска ассоциативных правил является алгоритм, получивший название *Frequent Pattern-Growth* (**FPG**), что можно перевести как

«выращивание часто встречающихся предметных наборов». Он позволяет не только избежать затратной процедуры генерации кандидатов, но уменьшить необходимое число проходов базы данных до двух. FPG состоит из следующих шагов:

Сжатие базы данных транзакций в компактную структуру, что обеспечивает очень эффективное и полное извлечение часто встречающихся предметных наборов;

Построение FP-дерева с использованием технологии разделения и захвата, которая позволяет выполнить декомпозицию одной сложной задачи на множество более простых;

Извлечение из FP-дерева часто встречающихся предметных наборов с помощью условных FP-деревьев.

Извлечение правил.

На рисунке 2.19, заимствованнном из [11] показано как за десять транзакций (TID=1,2,3…10) строится FP – дерево предметных наборов из элементов a,b,c,d.

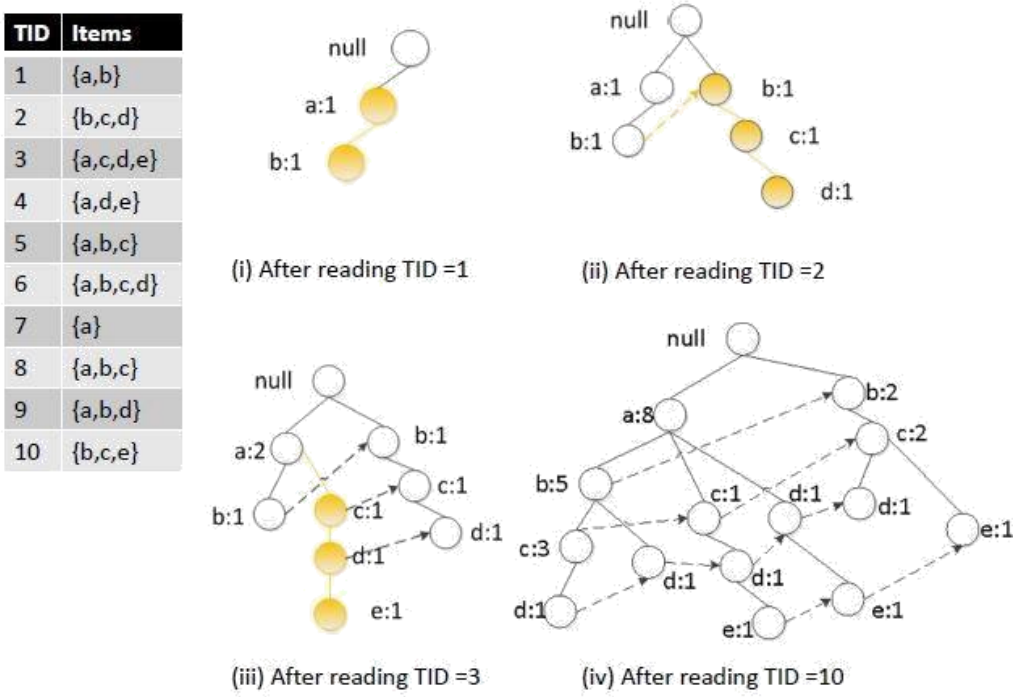


Рисунок 2.19. Иллюстрация работы алгоритма Frequent Pattern-Growth.

Более подробно с работой алгоритма можно познакомиться на ресурсе [12].

Анализ ассоциативных правил постоянно развивается. Появляются новые методики и алгоритмы. Неизменной остается важность этой техники анализа данных для решения задач, имеющих дело с предметными наборами.

Тема № 8 Нейронные сети

По этим термином в извлечении данных (Data Mining) понимается целый набор техник, основанных на моделях обработки, имитирующих деятельность человеческого мозга. Так же, как и мозг, они состоят из большого числа связанных между собой однотипных элементов – *нейронов*, которые моделируют нейроны головного мозга. На рис. 2.20 показана функциональная схема нейрона.

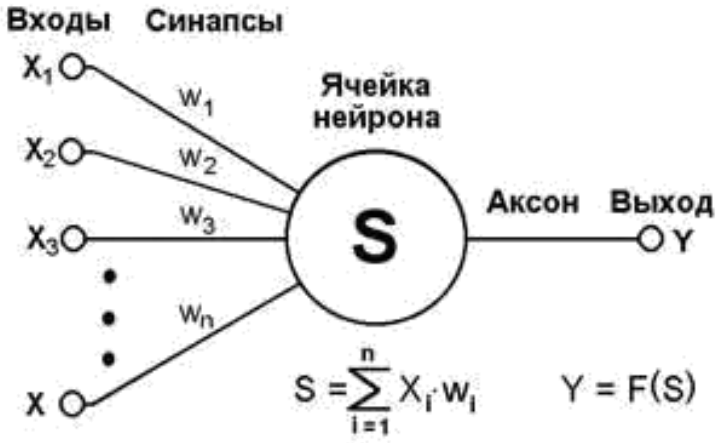


Рисунок 2.20. Функциональная схема одиночного нейрона, содержащая набор входов, соединенных с помощью синапсов с ядром, имеющим выходной аксон.

Искусственный нейрон, так же, как и живой, имеет ряд входов. Они обозначенykак x1… xn. На них поступает сигнал, который по синапсам, связывающим входы сядром, передается ядру. Ядро нейрона (S) осуществляет обработку входных

сигналов. Аксон Y связывает нейрон с нейронами другого слоя, если таковые

есть. Иначе представляет собой выходной сигнал. Каждый синапс имеет вес

(w1…wn), который определяет, насколько соответствующий вход нейрона влияет

на его состояние. Величину = ∑ называют состоянием нейрона,где =1 , , – число входов нейрона,значение-*i*–го входа и вес-*i*–го синапсасоответственно. Значение аксона нейрона в общем случае записывается как= ( ).Нелинейная функция*F*называется активационной.Активационнаяфункция должна обладать свойством резко возрастать на коротком интервале аргумента в окрестностях порогового значения, принимать приблизительно одно значение до этого интервала и приблизительно одно (большее) значение – после этого интервала. Опыт показал, что многие алгоритмы нейронных сетей плохо работают или не работают с линейными функциями. Поэтому наиболее часто в качестве активационной функции используется так называемый*сигмоид*

1= 1+ − . (кстати,эта функция уже была введена,когда рассматриваласьлогистическая регрессия). Основное достоинство этой функции в том, что она дифференцируема на всей оси абсцисс и имеет очень простую производную. При уменьшении параметра α сигмоид становится более пологим, вырождаясь в горизонтальную линию на уровне 0,5 при α=0. При увеличении а сигмоид все больше приближается к функции единичного скачка. Нейронная сеть представляет собой соединение нескольких нейронов, представляющее собой обычно структуру в виде нескольких слоев одиночных нейронов одинаковыми свойствами, которое позволяет организовать параллельную обработку данных, преобразуя входной набор в выходной. На рисунке 2.21 показан пример многослойной нейронной сети.

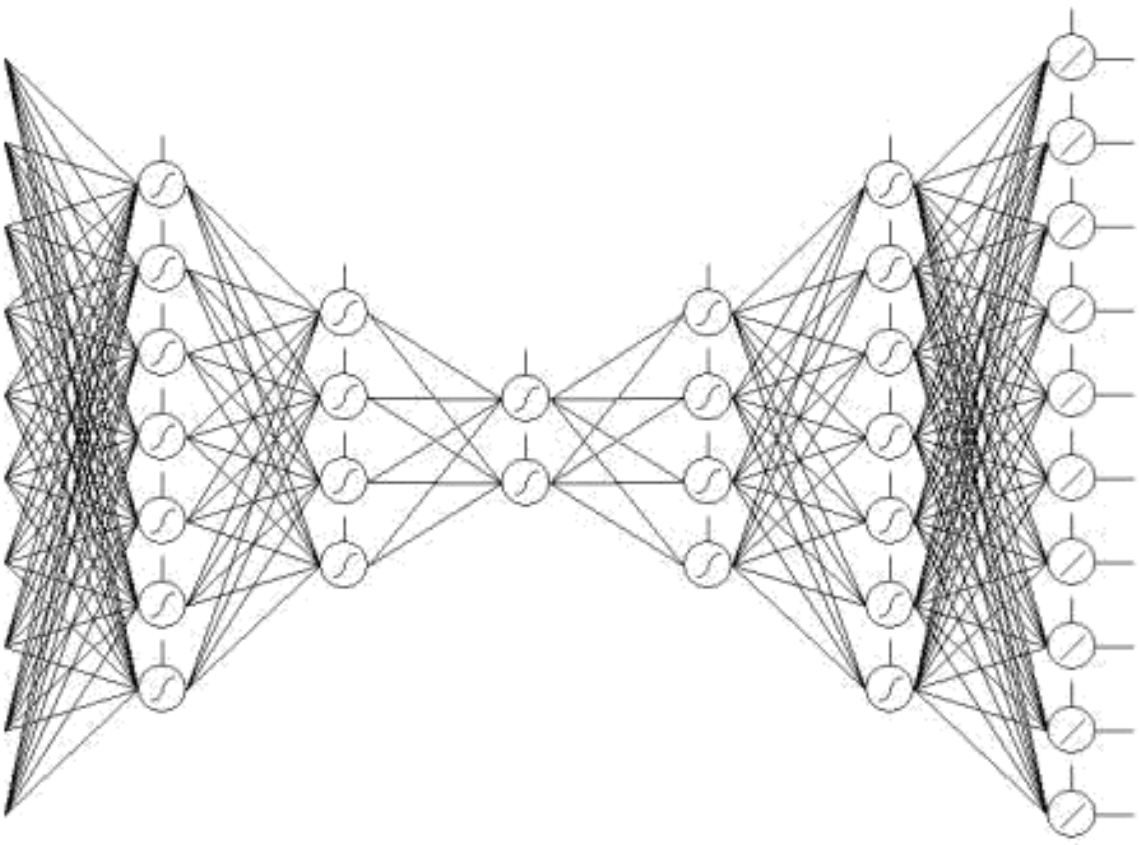


Рисунок 2.21. Пример шестислойной нейронной сети.

Нейронные сети используют для создания, как описательных моделей данных, такмоделей для прогнозирования. С помощью нейронных сетей могут решаться задачи кластеризации, классификации, предсказания значений и другие. Нейронные сети могут рассматриваться как сложные регрессионные модели, но их популярность, в первую очередь определяется хорошей проработанностью алгоритмов настройки коэффициентов модели, или, как обычно называют этот процесс, алгоритмов обучения сети. В связи с этим в литературе, чаще всего нейронные сети рассматриваются как обучающиеся системы, основная разновидность Machine Learning (систем машинного обучения). Для использования нейронной сети она должна пройти обучение. Под обучением сети понимается процесс настройки весов синапсов, так чтобы выход сети был адекватен задаче. Следует помнить, что сеть «понимает» не то, что от нее требуют, а то, что проще всего обобщить. Широко известны примеры обучения сети не классификации объектов на различном фоне, а классификации этих фоновых изображений. В процессе обучения сеть в определенном порядке просматривает обучающую выборку. Порядок просмотра может быть последовательным, случайным и т. д. Некоторые сети, обучающиеся без учителя (например, сети Хопфилда), просматривают выборку только один раз. Другие (например, сети Кохонена), а также сети, обучающиеся с учителем, просматривают выборку множество раз, при этом один полный проход по выборке называется эпохой обучения. При обучении с учителем набор исходных данных делят на две части — собственно обучающую выборку и тестовые данные; принцип разделения может быть произвольным. Обучающие данные подаются сети для обучения, а проверочные используются для расчета ошибки сети (проверочные данные никогда для обучения сети не применяются). Такимобразом, если на проверочных данных ошибка уменьшается, то сеть действительно выполняет обобщение. Если ошибка на обучающих данных продолжает уменьшаться, а ошибка на тестовых данных увеличивается, значит, сеть перестала выполнять обобщение и просто «запоминает» обучающие данные. Это явление называется переобучением сети или

оверфиттингом. В таких случаях обучение обычно прекращают. В процессе обучения могут проявиться другие проблемы, такие как паралич или попадание сети в локальный минимум поверхности ошибок. Невозможно заранее предсказать проявление той или иной проблемы, равно как и дать однозначные рекомендации к их разрешению.

Все выше сказанное относится только к итерационным алгоритмам поиска нейросетевых решений. Для них действительно нельзя ничего гарантировать и нельзя полностью автоматизировать обучение нейронных сетей. Однако, наряду с итерационными алгоритмами обучения, существуют не итерационные алгоритмы, обладающие очень высокой устойчивостью и позволяющие полностью автоматизировать процесс обучения.

Если в сети только один слой, процесс ее обучения может быть представлен достаточно простым процессом.

Слой Кохонена состоит из некоторого количества *n* параллельно действующих линейных нейронов. Все они имеют одинаковое число входов *m* и получают на свои входы один и тот же набор входных данных = (1,2, …). На выходе *j*-го линейного элемента получается значение = 0 + ∑ =1,где величины есть весовые коэффициенты*i*-входа*j*–гонейрона. Первое слагаемое является заданным порогом для каждого нейрона. Выходной нелинейный элемент выбирает тот нейрон, выход которого оказывается максимальным и тем происходит классификация наборов входных данных. Как строится алгоритм обучения. Приведем здесь широко применяемый алгоритм Кохонена.

1. Весовые коэффициенты , устанавливаются в некоторые малые значения. Устанавливаются входы = (1,2, …), соответствующие распознаваемому классу.

2. Вычисляются значения вспомогательного вектора для каждого входа по формуле = ∑ =1(−)2

3. Определяется *m*, такое что = min

4. Пересчитываются все для нейрона с номером *m* по формуле= (−),где величина коэффициента0.5≤μ≤1.

5. Если решение не было достигнуто, возвращаемся к шагу 2

Способности к распознаванию у многослойных сетей значительно превосходят те же способности у однослойной сети. Зато несколько усложняется процесс обучения этой сети. Еще более сложен процесс обучения сетей с обратной связью. Однако именно такие сети позволяют проводить наиболее качественное построение модели, получать наиболее точные прогнозы. Название *сети собратной связью* они получили из-за используемого алгоритма обучения,вкотором рассматривается сигнал, распространяющийся от выходного слоя к входному, т. е. в направлении, противоположном направлению распространениясигнала при нормальном функционировании сети. Нейронная сеть обратного распространения состоит из нескольких слоев нейронов, причем каждый нейрон слоя *i* связан с каждым нейроном слоя *i+1*, т. е. речь идет о *полносвязной* нейронной сети.Как уже отмечалось, в общем случае задача обучения нейронной сети сводится к нахождению некой функциональной зависимости *Y=F(X)* где *X* – входной, а *Y* – выходной векторы. В общем случае такая задача, при ограниченном наборе входных данных, имеет бесконечное множество решений. Для ограничения пространства поиска при обучении ставится задача минимизации целевой функции ошибки нейронной сети, которая находится по методу наименьших квадратов:

$$( ) = \frac{1}{2} \sum_{=1} (-)^2$$

где, , – значения – го выхода, целевое значение - го выхода и число нейронов в выходном слое соответственно.

Простейший метод градиентного спуска, рассмотренный выше, очень неэффективен в случае, когда производные по различным весам сильно отличаются. Это соответствует

ситуации, когда значение F(S) для некоторых нейронов близка по модулю к 1 или, когда модуль некоторых весов много больше.

В этом случае для плавного уменьшения ошибки надо выбирать очень маленькую скорость обучения, но при этом обучение может занять непозволительно много времени. Простейшим методом усовершенствования градиентного спуска является введение *моментат*, когда влияние градиента на изменение весов изменяется со временем. Дополнительным преимуществом от введения момента является способность алгоритма преодолевать мелкие локальные минимумы.

Детальнее с нейронными сетями с обратной связью можно ознакомиться в обширной литературе, посвященной их теории, список можно найти, например, здесь [13].

Важным аспектом применения нейронных сетей в анализе данных является то, что данные могут содержать даты, порядковые номера, символьные строки и т.п.

А в нейронных сетях все входные и выходные параметры представлены в виде чисел. То есть данные для анализа могут быть как количественными, так и качественными (категориальные). В этом случае необходимо преобразование качественных данных в числовые.

Качественные данные мы можем разделить на две группы: упорядоченные и неупорядоченные. Для рассмотрения способов кодирования этих данных мы рассмотрим задачу о прогнозировании своевременной оплаты электроэнергии.Возможным примером упорядоченных данных может быть, например, просрочка платежа клиентом:

Таблица 2.2

нет	**неделя**	Месяц	**квартал**	год и больше

Примером неупорядоченных данных могут, например, являться данные, например, о факторах риска, характерных для клиентов.

Таблица 2.3

нет	**судимость**	плохая кредитная история	**просрочки платежей**	низкий доход

В первом случае правильным будет установка в соответствие каждому значению своего веса, отличающегося на 1 от веса соседнего значения. Так, число 2 будет соответствовать просрочке в месяц. Таким образом, просрочка будет закодирована числами в диапазоне [0..4].

В принципе аналогично можно поступать и для неупорядоченных данных, поставив в соответствие каждому значению какое-либо число. Однако это вводит нежелательную упорядоченность, которая может исказить данные, и сильно затруднить процесс обучения. В качестве одного из способов решения этой проблемы можно предложить поставить в

соответствие каждому значению один из входов нейронной сети. В этом случае при наличии этого значения соответствующий ему вход устанавливается в 1 или в 0 при противном случае. К сожалению, данный способ не эффективен при большом количестве вариантов входного значения. В этом случае число входов нейронной сети разрастается до огромного количества. Это резко увеличивает затраты времени на обучение. В качестве варианта можно использовать несколько другой подход. В соответствие каждому значению входного параметра ставится бинарный вектор, каждый разряд которого соответствует отдельному входу нейронной сети.

Основным достоинством применения нейронных сетей является возможность решать различные неформализованные задачи. При этом можно очень просто моделировать различные ситуации, подавая на вход сети различные данные и оценивая выдаваемый сетью результат

Список техник больших данных весьма обширен и авторы не претендуют на освещение всех таких математических методов аналитики. Однако, ключевые из них мы постарались представить, чтобы заинтересованный читатель обратился к поисковым системам и нашел нужное для его задачи.

А мы переходим к обзору того компьютерного инструментария, который принято называть технологиями.

Тема № 9. Технологии и инструменты больших данных

Мы рассмотрим базовые технологии и инструменты, которые сегодня получили наибольшее распространение в известных проектах. Этот список не исчерпывает всех уже апробированных технологий и тем более находящихся в разработке, однако он позволяет получить достаточно целостное представление о том “чем” пользуются сегодня исследователи данных и какими инструментами необходимо владеть, чтобы развернуть проект с использованием больших данных.

- Hadoop – открытая программный каркас (framework) для работы с гигантскими объемами данных включая имплементацию MapReduce
- Hbase – открытая распределенная нереляционная СУБД, входящая вHadoop
- MapReduce – модель параллельной обработки для гигантских наборов данных в распределенных системах, имплементированная в Hadoop
- Mushup – приложение использующее и комбинирующее представление данных или функциональности от двух и более источников
- Metadata – данные дляописание данных
- Нереляционные СУБД
- R – язык программирования для статистической обработки и графики
- Stream Processing - обработка потоков данных
- Визуализация – приложения для графического представления данных и их взаимосвязей
- Big Table - СУБД Hbase Google File System
- BI – Business Intelligence – приложения для анализа и представления данных
- Cassandra – открытая СУБД для распределенного хранения данных
- Облачный компьютинг – парадигма использования компьютеров как предоставления компьютерных услуг
- Хранилища данных
- Распределенные компьютерные системы
- Dynamo – система хранения данных от Amazon

- ETL – extract-transform-load компьютерные приложения работы с БД

Технологии больших данных должны обеспечивать решениями и инструментами, позволяющими реализовывать описанные выше техники на значительных объемах разнородных данных с необходимой скоростью. Достигается это высокой параллелизацией вычислений и распределенным хранением данных. Несмотря на потребность значительной вычислительной мощности и памяти, как правило,развертывание программных продуктов больших данных производится на кластерах из компьютеров среднего или даже низкого класса (commodity computers). Это позволяет масштабировать системы больших данных без привлечения существенных затрат. В последнее время для развертывания систем больших данных все шире применяются облачные сервисы (cloud computing services). В случае имплементации системы в облаке узлы вычислительного кластера реализуются на виртуальных машинах облачной инфраструктуры и гибко адаптируются к задаче, снижая затраты на использование. Это служит дополнительным фактором, привлекающим многих разработчиков строить системы больших данных на облачных платформах.

В настоящей книге мы рассмотрим технологии, поставляемые в виде открытого кода (Open Source Projects). Коммерческие закрытые технологии охраняются правом интеллектуальной собственности, отчего не могут быть описаны нами с должной детализацией, а их использование требует приобретения законченного программного продукта, имеющего весьма высокую стоимость, что практически исключает знакомство с ним с малыми затратами.

Наиболее популярной технологией больших данных, считающейся де-факто стандартом для построения систем аналитики, работающих в пакетном режиме, является совокупность решений и программных библиотек, объединенных под названием Hadoop. Если Big Data Managmentпоступают в виде высокоскоростных потоков и реагирование системы должно происходить с малой задержкой, то вместо пакетной аналитики применяется аналитика реального времени. Здесь пока не возникло де-факто стандартных подходов и из наиболее популярных мы рассмотрим технологию под названием Storm.

Apache Hadoop

Под названием Hadoop сообщество Apache продвигает технологию, основан*ную* на использовании специальной инфраструктуры для параллельной обработки больших объемов данных. Hadoop обеспечивает среду для функционального программирования задач, автоматического распараллеливания работ, смещения вычислительной нагрузки к данным.

История Hadoop непосредственно связана с разработкой Google File System (2003 г) и затем реализацией технологии MapReduce (2004 г). На основе этих компонент в 2005 г появилось приложение поиска информации Apache Nutch, которое на следующий год дало дорогу проекту Apache Hadoop.

Hadoop состоит из четырех функциональных частей:

Hadoop Common Hadoop HDFS

Hadoop MapReduce Hadoop YARN

Hadoop Common – это набор библиотек и утилит, необходимых для нормального функционирования технологии. В его состав входит специализированный упрощённый интерпретатор командной строки.

HDFS (Hadoop Distributed File System) – это распределенная файловая система для хранения данных на множестве машин в больших объемах. Проектировалась так, чтобы обеспечивать:

- Надежное хранение данных на дешевом
- В ненадежном оборудовании;
- Высокую пропускную способность чтения-записи;
- Потоковый доступ к данным;

- Упрощенную модель согласованности;
- Архитектуру аналогичную Google File System.

В основе архитектуры HDFS лежат узлы хранения – серверы стандартной архитектуры, на внутренних дисках которых хранятся данные. Для всех данных используется единое адресное пространство. При этом обеспечивается параллельный ввод-вывод информации с разных узлов. Таким образом, гарантируется высокая пропускная способность системы.

HDFS оперирует на двух уровнях: пространства имён (Namespace) и хранения блоков данных (Block Storage Service) (рисунок 2.21). Пространство имён поддерживается центральным узлом имён (Namenode), хранящим метаданные файловой системы и метаинформацию о распределении блоков файлов. Многочисленные узлы данных (Datanode) непосредственно хранят файлы. Узел имён отвечает за обработку операций файловой системы— открытие и закрытие файлов, манипуляция с каталогами и т.п. Узлы данных отрабатывают операции по записи и чтению данных. Узел имён и узлы данных снабжаются веб-серверами, отображающими текущий статус и позволяющими просматривать содержимое файловой системы.

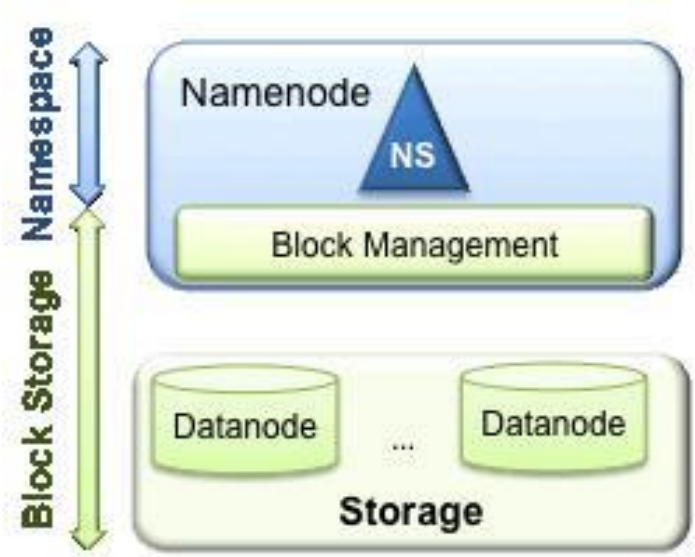


Рисунок 2.22. Структура файловой системы HDFS

У HDFS нет POSIX-совместимости. Не работают Unix-команды ls, cp и т.п. Для монтирования HDFS в Linux ОС необходимы специальные инструменты, например, HDFS-Fuse. Файлы поблочно распределяются между узлами. Все блоки в HDFS (кроме последнего блока файла) имеют одинаковый размер – от 64 до 256 Мб. Для обеспечения устойчивости к отказам серверов, каждый блок может быть продублирован на нескольких узлах. Коэффициент репликации (количество узлов, на которых должен быть размещён каждый блок) определяется в настройках файла. Файлы в HDFS могут быть записаны лишь однажды (модификация не поддерживается), а запись в файл в одно время может вести только один процесс. Таким простым образом реализуется согласованность данных.

Hadoop MapReduce – это наиболее популярная программная реализация модели параллельной обработки больших объемов данных путем разделения на независимые задачи, решаемые функциями Map и Reduce.

Алгоритм MapReduce получает на вход 3 аргумента: исходную коллекцию данных, Map функцию, Reduce функцию, и возвращает результирующую коллекцию данных.

Collection MapReduce(Collection Source, Function Map, Function Reduce)

Исходными коллекциями данных являются наборы записей специального вида, Это структура данных типа *Ключ,Значение* (KEY, VALUE). Пользователю необходимо задать

функции обработки Map и Reduce. Алгоритм сам заботится о сортировке данных, запуске функций обработки, повторном исполнении упавших транзакций и много чем еще. Результирующая коллекция состоит из результатов анализа в легко интерпретируемом виде. Работа алгоритма MapReduce состоит из трех основных этапов: Map, Group и Reduce. В качестве первого этапа над каждым элементом исходной коллекции выполняется Map функция. Как правило, она принимает на вход одну запись вида (KEY, VALUE), и возвращает по ней некоторое количество новых записей (KEY1, VALUE1), (KEY2, VALUE2), ..., т.е. преобразует входную пару {ключ: значение} в набор промежуточных пар. Также эта функция играет роль фильтра — если для данной пары никаких промежуточных значений возвращать не нужно, функция возвращает пустой список.

KeyValueArray Map(object itemFromSourceCollection)

Можно сказать, что обязанность Map функции конвертировать элементы исходной коллекции в ноль или несколько экземпляров объектов {ключ: значение}. Это продемонстрировано ниже на рисунке 2.23:

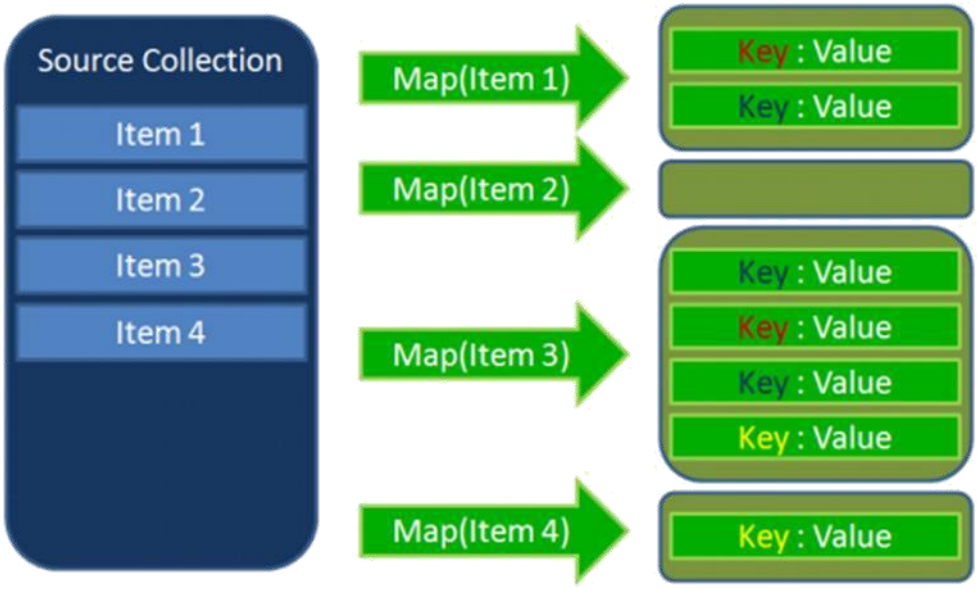


Рисунок 2.23. Формирование структуры данных типа «ключ-значение» функцией Map

На втором этапе (Group) алгоритм сортирует все пары {ключ: значение}и создает новые экземпляры объектов, сгруппированные по ключу. Операция группирования выполняется внутри алгоритма MapReduce и пользователем не задается. В результате происходит объединение всех значений для одного и того же ключа и результатом является пара {ключ: список значений} как показано на рисунке 2.24.

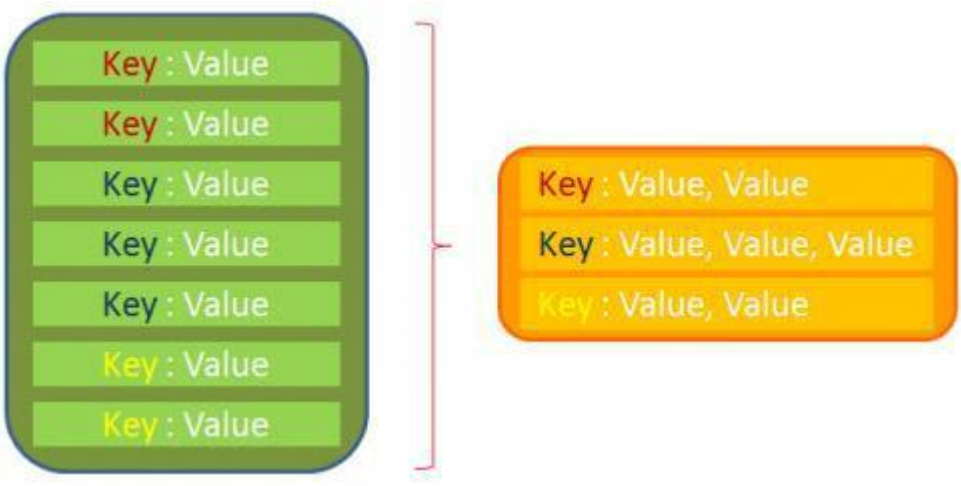


Рисунок 2.24. Группировка данных по ключу

Последним шагом выполняется функция Reduce для каждого сгруппированного экземпляра объекта {ключ: список значений}:

ItemResult Reduce(KeyWithArrayOfValues item)

Формально функция работает по принципу {(KEY, VALUE1), (KEY, VALUE2), ...} → (KEY1, VAL1), (KEY2, VAL2), ...

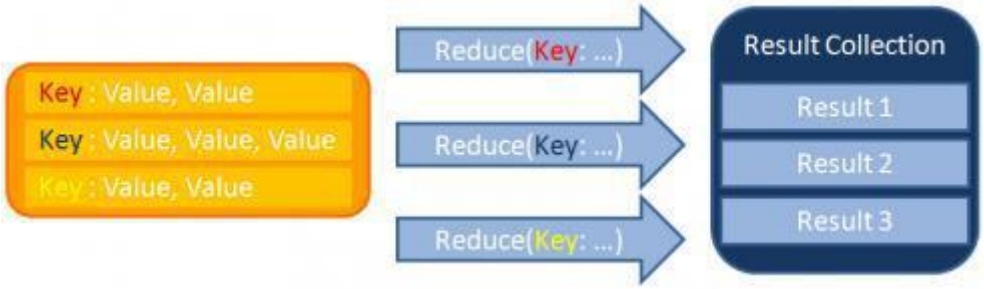


Рисунок 2.25. Исполнение функции Reduce, возвращающей коллекцию всего массива данных по запрошенному ключу.

Функция Reduce возвращает экземпляры объекта {ключ: свернутое значение}, которые включается в результирующую коллекцию.

Для примера, рассмотрим упрощенный вариант задачи, стоящей перед поисковыми системами. Допустим, у нас есть база данных страниц в Интернете, и мы хотим, сколько раз ссылаются на каждую страницу. Пусть есть страница first.com со ссылками на first.com, second.com, third.com, страница second.com с двумя ссылками на first.com и страница third.com, на которой нет ссылок вообще. Чтобы иметь единый формат исходной коллекции данных, определим вид каждой сохраненной страницы как (KEY = URL, VALUE = TEXT).

Запустим на этой базе данных операцию Map. Получив на вход текст страницы, она выделит все исходящие ссылки:

```
void Map(String key, String value) {
for key in GetUrls(value) {
OutputRecord(key, 1)
}
}
```

Мы получим список записей вида (KEY = URL, VALUE = 1).

(first.com, 1)
(second.com, 1)
(third.com, 1)

(first.com, 1)
(first.com, 1)

Единичка обозначает одну обнаруженную ссылку. На шаге Group таблица сортируется, чтобы можно было объединить записи с одинаковым ключом.

(first.com, 1)
(first.com, 1)
(first.com, 1)
(second.com, 1)
(third.com, 1)

Записи группируются по ключу.

(first.com, 1, 1, 1)
(second.com, 1)
(third.com, 1)

На шаге Reduce учитываются все записи с одинаковым ключом (то есть одинаковым URL). Псевдокодфункции:

```
void Reduce(String key, Iterator it) {
int count = 0;
while (it.HasMoreRecords()) {
count += it.GetValue();
}
OutputRecord(key, count);
}
```

После выполнения Reduce выходная таблица будет иметь вид:

(first.com, 3)
(second.com, 1)
(third.com, 1)

Результаты легко интерпретируются.

В качестве базового языка написания функций используется Java. Для программирования существует популярный Hadoop плагин в Eclipse. Но можно обойтись и без него: утилиты *Hadoop streaming* позволяют использовать в качестве Map и Reduce любой исполняемый файл, работающий со стандартным вводом-выводом операционной системы (например, утилиты командной оболочкиUNIX, скрипты Python, Ruby и т.д.), есть также SWIG-совместимый прикладнойинтерфейс программирования *Hadooppipes* на C++. Кроме того, в состав дистрибутивов Hadoop входят реализации различных обработчиков, наиболее часто используемых в распределённой обработке.

Особенностью Hadoop является перемещение вычислений как можно ближе к данным. Поэтому пользовательские задачи запускаются на том узле, который содержит данные для обработки. По окончанию фазы Map происходит перемещение промежуточных списков данных для обработки функцией Reduce. Их объем, как правило, мал, что обеспечивает высокое быстродействие системы. На рисунке 2.26 показаны основные шаги выполнения вычислений в распределенной среде Hadoop.

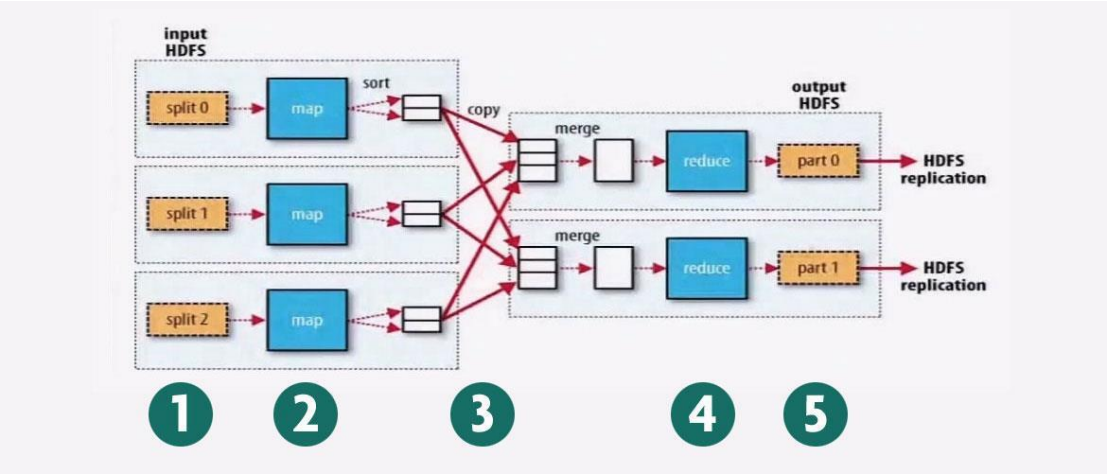


Рисунок 2.26. Пять основных шагов исполнения задачи в Hadoop.

Заметим здесь, что кроме Hadoop существуют разные имплементации MapReduce. Изначально MapReduce был реализован компанией Google. Позднее появились другие реализации алгоритма. Развитием MapReduce от Google стал проект с открытым исходным кодом - MySpace Qizmt - MySpace's Open Source Mapreduce Framework. Другой известной версией алгоритма является та, что реализована в системе MongoDB

Hadoop YARN (*Yet Another Resource Negotiator*) – платформа управления ресурсами системы, ответственная за распределение вычислительных ресурсов серверов и расписание выполнения пользовательских задач.

В первых версиях Hadoop MapReduce включал планировщик заданий *JobTracker*, начиная с версии 2.0 (2013 г.) эта функция перенесена в YARN. В ней модуль Hadoop MapReduce реализован поверх YARN. Программные интерфейсы по большей части сохранены, однако полной обратной совместимости нет.

YARN иногда называют кластерной операционной системой. Это обусловлено тем, что платформа ведает интерфейсом между аппаратными ресурсами и различными приложениями, использующими вычислительные мощности.

Основой YARN является логически самостоятельный демон — планировщик ресурсов (ResourceManager), абстрагирующий все вычислительные ресурсы кластера и управляющий их предоставлением приложениям распределённой обработки. Ему подотчетны многочисленные менеджеры узлов (Node Manager), ответственные за отслеживание текущего статуса и нагрузки отдельных серверов.

Работать под управлением YARN могут как MapReduce-программы, так и любые другие распределённые приложения, поддерживающие соответствующие программные интерфейсы. YARN обеспечивает возможность параллельного выполнения нескольких различных задач в рамках системы серверов. Разработчику распределённого приложения необходимо реализовать специальный класс управления приложением (AppMaster), который отвечает за координацию заданий в рамках тех ресурсов, которые предоставит планировщик ресурсов. Планировщик ресурсов отвечает за создание экземпляров класса управления приложением и взаимодействия с ними через сетевой протокол.

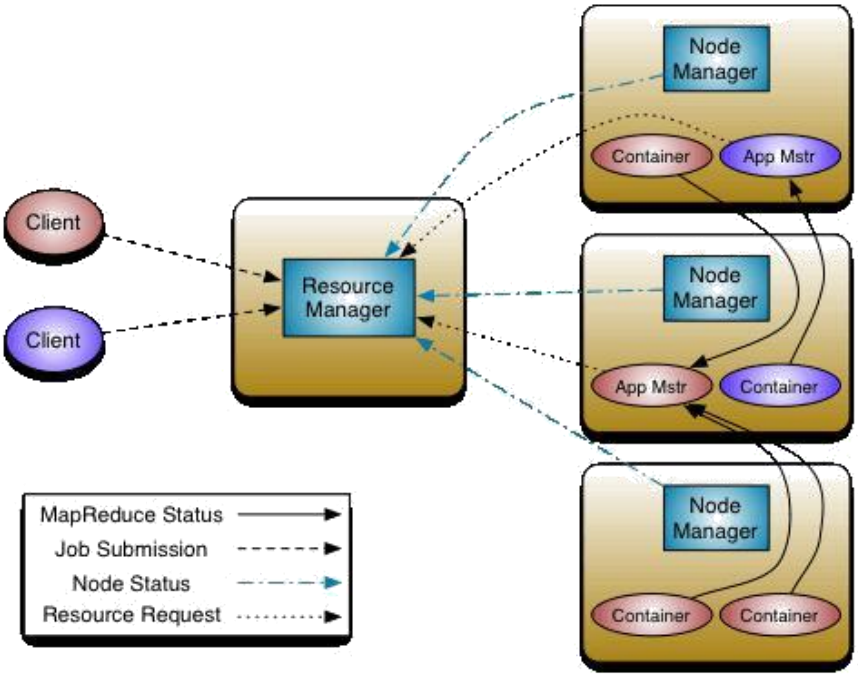


Рисунок 2.27. Управление кластером в среде Hadoop YARN

На основе Hadoop создан целый ряд продуктов для обработки данных. Вот список лишь наиболее популярных из них:

- Pig – высокоуровневый язык потоков данных для параллельного программирования;
- HBase – распределенная база данных, которая обеспечивает хранение больших таблиц;
- Cassandra – устойчивая к ошибкам, децентрализованная база данных;
- Hive – хранилище данных с функциями объединения данных и быстрого поиска;
- Mahout – библиотека методов машинного обучения и извлечения знаний

Hadoop является очень динамично развивающейся технологией. Поэтому наиболее свежую информацию рекомендуется получать в Интернете на сайте http://hadoop.apache.org/. А неплохой учебник по Hadoop был переведен на русский язык и издан в 2013 году издательством Питер [14].

Тема № 10. Storm – система потоковой обработки

Storm является бесплатной технологией и программной реализациейраспределенной вычислительной системы реального времени [15]. Эта системапозволяет строить надежную обработку неограниченных потоков данных подобнотому как Hadoop делает это с пакетной обработкой. Storm применяется дляаналитики реального времени, онлайнового машинного обучения, непрерывныхвычислений, распределенных ETL и других операций с потоками больших данных.Storm может интегрироваться с технологиями очередей и баз данных, которыеуже используются и не зависит от языка программирования. Основой Stormявляются Storm топологии и Storm кластер. Кластер является объектом,подобным Hadoop кластеру, а вместо запуска MapReduce job здесь запускаютсяStorm topologies. Jobs и Topologies имеют ключевое различие – первые внормальном режиме завершают работу, а вторые обрабатывают сообщениявсегда. В Storm кластере имеется два типа узлов master

node и worker nodes(рисунок 2.28). На master node запускается демон называемый Nimbus, которыйподобен JobTracker в Hadoop. Nimbus ответственнен за распределение кода порабочим узлам кластера, распределение задач по машинам и запуск и остановкурабочих процессов. Каждый рабочий процесс выполняет подмножествотопологии. Работающая топология состоит из многих рабочих процессов,распределенных по многим машинам. Каждый рабочий узел (worker node) имеетдемон под названием Supervisor. Этот модуль слушает все процессы на своеймашине и запускает и останавливает их по инициативе Nimbus. Координациямежду Nimbus и всеми Supervisor производится через специальный кластер,называемый Zookeeper. Этот кластер также хранит на своем дисковомпространстве состояние всех процессов, что позволяет восстанавливать послесбоя отдельно любую машину рабочего кластера.

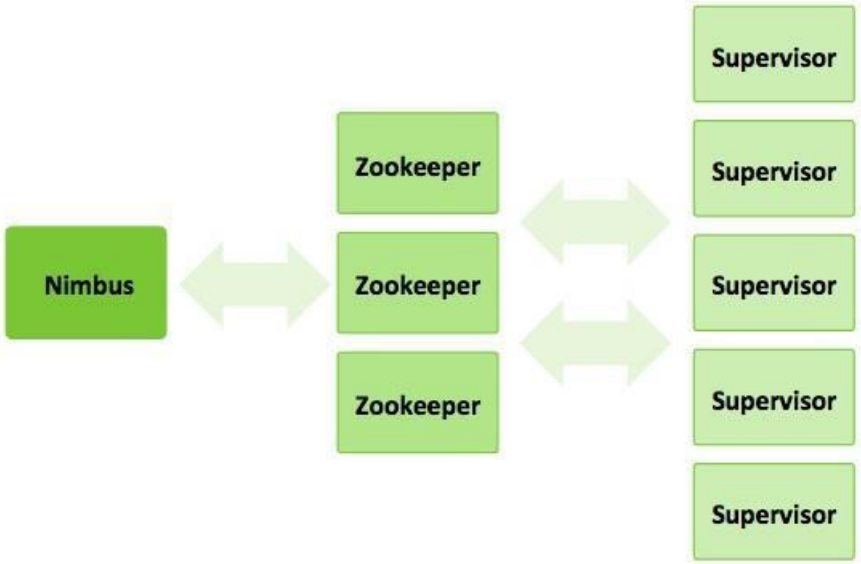


Рисунок 2.28 Структура Storm – кластера и основные потоки управления.

Чтобы выполнить вычисления в реальном времени на Storm нужно создать топологию (topologies) – граф вычислений. Каждый узел в топологии содержит логику процессинга и линк между узлами, показывающий как данные должны быть переданы между узлами.

Основной абстракцией в Storm является поток (stream). Потоком называется неограниченная последовательность кортежей (tuples). Источники потоков данных для обработки представляются в топологии абстракцией, называемой spout, а обработчики потоков, которые могут выполнять функции, фильтровать потоки, агрегировать или объединять потоки данных, взаимодействовать с базами данных называются bolt. На рисунке 2.28 приведен пример топологии, в которой обрабатываются два потока данных с помощью четырех обработчиков

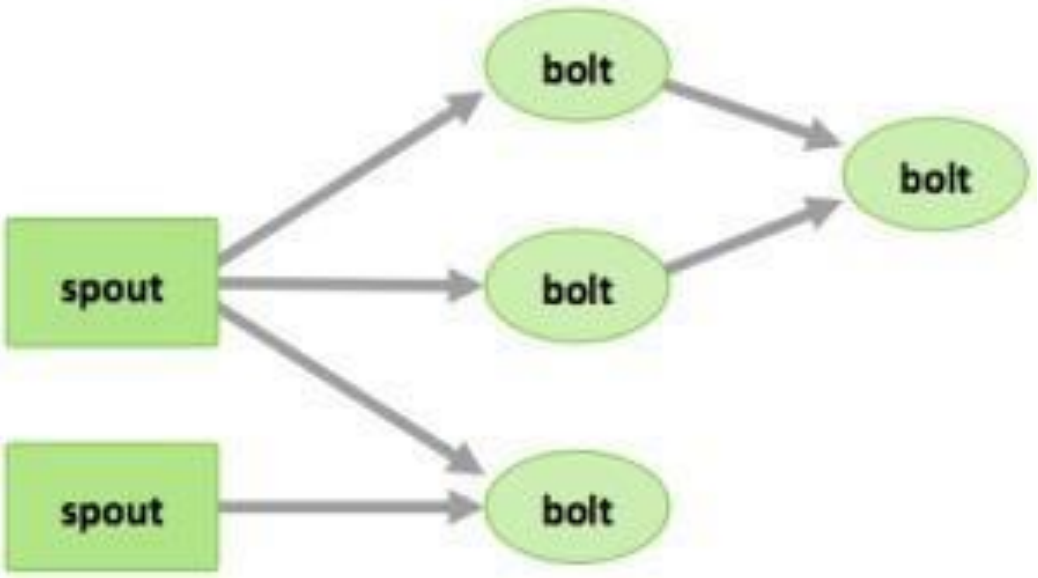


Рисунок 2.28. Диаграмма топологии для исполнения некоторой задачи.

Назовем в качестве примера имплементации Storm проект Predictive Analytics jn H2J and Hortonworks Data Platform [16]. Эта учебная имплементация позволяет получить реальные результаты для многих практически интересных случаев, таких как обработка потока кликов на Web – страницах.

Еще одна технология, получившая новую жизнь благодаря приложениям больших данных - это функциональный язык программирования R.

Язык программирования R

В последнее время R де-факто стал основным языком для Data Science получив широкое распространение среди аналитиков больших данных. Прежде всего R_язык программирования для статистической обработки данных и работы сграфикой, но в тоже время, это свободная программная среда с открытым исходным кодом, развиваемая в рамках проекта GNU.

Поскольку R – Open Source software, то его можно скачать и установить совер-шенно свободно [17]. Работает R в режиме командной строки и для удобства его обычно используют установив графическую оболочку R-Studio [18]. Работа с R напоминает работу с научным калькулятором, поскольку выполнение строк кода программы происходит по мере завершения каждой исполнимой конструкции. Код программы на R состоит из выражений. Они могут быть следующих видов:

- присваивание;
- условные выражения;
- арифметические выражения. Примеры выражений в R:

> *y<-100*
> *if (1==1) 1 else 0 [1] 1*
> *100/5[1] 20*

Компонентами выражений в R являются объекты или функции. Обычно каждое выражение пишется на отдельной строке, но допускается записывать несколько выражений в одной строке, разделяя их точкой с запятой, как показано ниже.

> *"LinuxCareer.com";sin(pi);5^7 [1] "LinuxCareer.com"*
[1] 1.224647e-16
[1] 78125

Объект можно рассматривать как элементарный компонент ("thing") языка программирования R. Например, объектами являются:

- числовой вектор;

- символьный вектор; - список; - функция.

Примеры объектов в R:

```
>       c(1,5,2,7,9,0)
[1] 1 5 2 7 9 0
>       c("GNU R programming tutorial","LinuxCareer.com") [1] "GNU R programming tutorial" "LinuxCareer.com"
>       list("GNU R programming tutoial",c(1:5),"this is also an object in R")
[[1]]
[1] "GNU R programming tutoial"
[[2]]
[1] 1 2 3 4 5
[[3]]
[1] "this is also an object in R"
>       function(a,b) {a/b} function(a,b) {a/b}
```

Символ в R - это имя переменной. Поэтому, если вы присваиваете объект переменной, тем самым вы присваиваете объект символу. Окружение в R представляет собой набор таких символов, созданных для определенной цели.

Пример символов в R:

```
>       x<-3
>       y<-"R tutorial"
```

В примере выше x и y являются символами.

В R функция - это объект, который получает в качестве аргументов другие объекты, и в качестве результата тоже возвращает объект. Знали ли вы, что в R оператор присваивания "<-" является функцией? Вместо выражения:

```
> a<-1
```

вы можете вызвать функцию "<-" с аргументами "a" и "1", как показано ниже:

```
> '<-'(a,1)
```

Несколько примеров функций в R:

"<-" - присваивание;

"+" - суммирование;

"if" - оператор;

"[" ссылка на вектор.

Примеры использования функций в R:

```
>       '+'(1,1) [1] 2
>       'if'(1>3,"one greater than three", "one less than three") [1] "one less than three"
>       '['(a,1)
[1] 1
```

В R объекты не изменяются. Это значит, что R будет копировать объект, а не только ссылку на объект. Рассмотрим следующий объект. Мы определяем функцию, которая присваивает i-му элементу вектора x значение 4, как показано ниже:

```
> f<-function(x,i){x[i]<-4}
```

Давайте посмотрим, что произойдет, если мы определим вектор w и передадим его в качестве аргумента в функцию f.

```
>       w<-c(1,3,6,7)
>       f(w,1)
>       w
[1] 1 3 6 7
```

Мы видим, что вектор w, когда он передается функции, просто копируется, поэтому функция не модифицирует сам вектор.

Все в R является объектом. Они используются не только для хранения данных в случае векторов, списков или других структур данных. Другими примерами объектов в R являются функции, символы или выражения. Например, именафункций в R являются символьными объектами, которые указывают на функции (которые тоже являются объектами), как показано ниже:

```
>       functionname<-function(x,y) x+y
>       functionname
function(x,y) x+y
```

В R имеются некоторые специальные обозначения. Это: NA - используется для представления отсутствующих значений, означает "not available";

Inf или -Inf - результат вычислений, когда полученное число слишком велико, слишком мало, или имеет место деление на нуль;

NaN - результат вычислений, который не может существовать, например при делении нуля на нуль, означает "not a number";

NULL - часто используется в качестве аргумента функций, означает, что этому аргументу не присвоено никакого значения.

R часто приводит данные одного типа к другому. Например, когда вы вызываете функцию с аргументом неправильного типа, R пытается конвертировать этот аргумент, чтобы функция могла работать. Другим примером является случай, когда мы определяем вектор с числовыми значениями. В этом случае R присваивает вектору тип "integer":

```
>       x<-c(1:10)
>       typeof(x)
[1] "integer"
```

Теперь, если мы изменим значение четвертого элемента вектора x, R автоматически изменит тип вектора на "double":

```
>       x[4]<-4.1
>       typeof(x)
[1] "double"
```

Интерпретатор - это программа, которая выполняет написанный код. Нет необходимости компилировать код R, как в случае C, C++ или Java. Это означает, что R является интерпретируемым языком.

Интерпретатор R обрабатывает выражения R в несколько этапов. Во-первых, он анализирует выражения и переводит их в соответствующую функциональную форму. Давайте вызовем функцию quote(), чтобы посмотреть, как это происходит.

```
>       typeof(quote(if(1>3) "one is greater than three" else "one is less than three"))
[1] "language"
```

Вншеприведенное выражение возвращает объект "language". Чтобы увидеть, как R обрабатывает выражение, мы создаем дерево синтаксического разбора.

```
>       as(quote(if(1>3) "one is greater than three" else "one is less than three"),"list")
[[1]]
'if'
[[2]]
1 > 3
[[3]]
[1] "one is greater than three"
[[4]]
[1] "one is less than three"
```

Теперь давайте применим функцию typeof() сэлементам такого списка, который покажет, как в R интерпретируется выражение.

> lapply(quote(if(1>3) "one is greater than three" else "one is less than three"),typeof)

[[1]]
[1] "symbol"
[[2]]
[1] "language"
[[3]]
[1] "character"
[[4]]
[1] "character"

Как вы можете видеть, некоторые части оператора if не включаются в анализируемое выражение. Это элемент else. Кроме того, интересно отметить, что первый элемент списка - это символ, который указывает на функцию if(). Даже несмотря на то, что синтаксис оператора if отличается от вызова функции, интерпретатор R транслирует выражение в вызов функции с именем функции в качестве первого аргумента и другими аргументами, как в списке выше.

Эта краткая иллюстрация конструкций языка R позволяет видеть ряд удобных свойств для его применения в задачах обработки данных. Однако, самое главное преимущество R в наличие гигантских по широте и возможностям готовых функций обработки. Их принято объединять в пакеты (packages). Наиболее известны и популярны пакеты функции, входящие в так называемую CRAN - Comprehensive R Archive Network [19] . Приведем здесь только названия основных групп пакетов функций на R, поскольку полное описание пакетов могло бы занять сотни страниц.

- Байесовский интерфейс
- Хемометрия и вычислительная физика
- Анализ, мониторинг и проектирование клинического исслеования
- Кластерный анализ и конечноэлементные модели
- Дифференциальные уравнения
- Вероятностные распределения
- Вычислительная эконометрика
- Анализ экологических и данных и данных об окружающей среде
- Планирование экспериментов и анализ экспериментальных данных
- Эмпирические финансы
- Статистическая генетика
- Графическое отображение, динамическая графика и визуализация
- Высокопроизводительны е и параллельные вычисления на R
- Машинное обучение и статистическое обучение (Machine learning&Statistical Learning)
- Анализ медицинских изображений
- Мета-анализ
- Мультивариантная статистика
- Обработка естественного языка (Natural language Processing)
- Численная математика
- Официальная статистика и методология опросов
- Оптимизация и математическое программирование
- Анализ данных фармокинетики
- Филогенетика и особые сравнительные методы (Phylogenetics, Espeially Comprative Methods)

- Психометрические модели и методы
- Исследования репродуцирования (Reproducible Research)
- Робастные статистические методы
- Статистика социальных исследований
- Анализ пространственных данных
- Управление и анализ пространственно-временных данных (Handling and Analyzing Spatio-Temporal Data)
- Анализ наблюдений (Survival Analysis)
- Анализ временных рядов
- Web технологии и сервисы
- Графические модели в R (gRaphical Models in R)

Практически любая задача обработки данных сегодня может быть решена правильным применением функций из перечисленных выше пакетов. Поэтому, прежде чем взяться за написание программы для вашей задачи, настоятельно рекомендуем потратить время на изучение готовых функций обработки. Как правило, потраченное на это время с лихвой компенсируется экономией на кодирование и отладку собственного кода. Еще один весьма полезный ресурс для вхождения и освоения языка R, который мы рекомендуем - [20]

Тема № 11. Аналитика больших данных как корпоративный проект

В этом разделе мы рассмотрим как выглядит типовой жизненный цикл проекта по аналитике больших данных. За основу мы приняли подход одного из ведущих игроков на рынке больших данных – корпорации EMC [21]. Большинство проектов по аналитике больших данных по существу решаемых проблем ищут решение одной или комбинации следующих задач:

- Поиск нового: редких фактов, один из миллионов или миллиардов объектов и событий

Поиск классов: нахождение новых типов объектов и поведений

- Поиск ассоциаций: нахождение необычных невероятных совместно случающихся ассоциаций идентификация связей между различными вещами, людьми или событиями, которые много ближе чем шесть ступеней разделения тесного мира [22]

Эти задачи не всегда явно могут быть сформулированы, а их результаты правильно интерпретированы, и исследователям данных приходится взаимодействовать с целой командой специалистов, которые являются важной и непременной частью команды проекта.

Жизненный цикл проекта аналитики больших данных

Многочисленные проекты по использованию больших данных для получения аналитических результатов имеют, как правило, много общего в составе и порядке выполнения. Эксперты выделяют типовой жизненный цикл таких проектов, отмечая цикличность их развития.

На рисунке 2.30 приведена диаграмма жизненного цикла проекта аналитики больших данных, ставящего целью получение информации для принятия решения, основанного на имеющихся данных.

Рисунок 2.30 Диаграмма жизненного цикла проекта аналитики больших данных.

Этот цикл подтверждается многими проектами, которые были выполнены различными компаниями. Приведем здесь основные направления проектов по аналитике больших данных.

Рекомендательные системы: Онлайн магазины иWeb–порталы используютHadoop, чтобы сравнивать пользовательские запросы и покупки, чтобы выделить поведенческие профили и характеристики и рекомендовать клиентам подходящие товары и услуги. LinkedIn использует этот подход для своих предложений “Люди, которых Вы возможно знаете”. Amazon прямо предлагает купить что-либо на основании данных рекомендательной машины.

Анализ “чувств” (sentiment analysis): использован расширенная аналитикатекста на основе Hadoop на основе анализа неструктурированных текстов в социальных сетях и СМИ включая посты в Twitter и Facebook, чтобы определить какие чувства испытывает пользователь к выбранной компании, бренду или продукту. Анализ может фокусироваться как на статистически средние типы людей, так и индивидуальных пользователей.

Моделирование рисков:финансовые компании,банки и некоторые другиеиспользуют Hadoop и аналитические песочницы чтобы анализировать большие объемы данных о транзакциях, чтобы определить риск и опасность финансовых активов, чтобы подготовить сценарии “что-если” основанные на имитационном моделировании поведения рынков, а также оценить риск от потенциальных клиентов.

Детектирование хищений:использование техник больших данных дляобъединения поведения клиентов, исторических данных и транзакций, чтобы детектировать активности, подозрительные на хищения. Компании кредитных карт используют технологии больших данных для определения по потоку транзакций подозрительных на операции с похищенной картой.

Анализ маркетинговых кампаний:отделы маркетинга во всех видах индустриивсегда занимались мониторингом и определением эффективности маркетинговых кампаний, но только возможности больших данных позволили получить анализ высокогранулированных потоков кликов на сайтах и звонков на телефонах с учетом местоположения клиентов.

Анализ оттока клиентов:предприятия используютHadoopи методы машинногообучения для выделения поведенческих паттернов по анализу данных о клиентах для выявления вероятных отказов от услуг компании.

Анализ социальных графов:благодаря новым видам хранилищ данных дляэффективного храненения больших графов и Hadoop из структур связей между людьми

извлекается информация о степени влияния отдельных персон на другие. Это помогает выделить “наиболее важных” клиентов, которыми часто являются вовсе не те, кто покупает больше всего продуктов, а те, чье мнение влияет на большинство других клиентов компании.

Аналитика пользовательского опыта:постоянно ориентированные на клиентапредприятия используют Hadoop и другие технологии больших данных для интеграции данных от всех каналов взаимодействия с клиентами, такими как чат на сайте компании, звонки в колл-центр и т.п. для понимания влияния каждого из видов взаимодействия и оптимизации политик компании по работе с клиентами.

Мониторинг сетей:сбор и обработка всех логов и сообщений,генерируемыхоборудованием информационной сети предприятия обеспечивает администраторов диагностикой узких мест и проблем безопасности. Используя Hadoop удается анализировать исторически значимые отрезки времени для того, чтобы видеть эволюцию сети и влияние модернизаций на поведение. Эти же методы анализа оказываются пригодными и не только для информационных сетей, но также и транспортных, где может решаться задача повышения эффективности и сетей в электроэнергетике.

Рассмотрим более детально все этапы жизненного цикла аналитического проекта. Отметим, работа над проектом может производиться одновременно в нескольких фазах и переход от фазы к фазе нередко происходит не только в прямом, но и в обратном направлении, если оказывается, что работа на текущем этапе не может быть выполнена на основе имеющихся результатов.

Исследование

На этом этапе изучается бизнес-область задачи, включая относящуюся к вопросу историю, включая подобные проекты в этой или других организациях. Оцениваются ресурсы проекта в терминах кадров, технологий времени и доступных данных. Вычленяется задача, которая должна быть решена в проекте и декомпозируется на составляющие подзадачи. Формулируются начальные гипотезы для тестирования и начала изучения данных. Возможность перехода к следующему этапу определяется положительным ответом на вопрос: “У нас достаточно информации, чтобы набросать аналитический план и показать его другим для обзора?”

Подготовка данных

На этой фазе развертывается аналитическая песочница, в которой будет производиться работа над проектом. Выполняются ETL процессы и производится первичное ознакомление с данными и оценивается их качество для проекта. Переход к следующей фазе проекта может быть осуществлен при положительном ответе на вопрос: “У нас есть достаточно хорошие данные, чтобы начать построение модели?”

Планирование модели

На этой фазе производится выбор методов, техник и потоков работ для оценивания моделей. Изучаются данные для выявления отношений между переменными и выделения ключевых переменных и моделей, которые представляются предпочтительными для использования. Ключевой вопрос перехода к следующему этапу: “У нас есть хорошая идея о типе модели, которую следует попробовать? Можем мы теперь конкретизировать аналитический план?”

Построение модели

На этой фазе выполняются работы по выделению данных для тестирования, обучения модели и данных для получения рабочих результатов. Далее выбираются эффективные технологии и инструменты для построения и обучения моделей и

выполняется весь поток работ по построению модели и получения результатов анализа с помощью этих моделей. Работа с моделью может быть перенесена в следующую фазу при положительном ответе на вопрос: “У нас есть достаточно робастная (малочувствительная к значениям параметров) модель, которая дает интерпретируемые результаты? Есть ли достаточная уверенность в ее корректности?”

Обсуждение результатов

Эта фаза проекта весьма важная, поскольку только в процессе обсуждения можно убедиться, что цели, поставленные на фазе исследования, достигнуты. В обсуждении принимают участие все заинтересованные лица и поэтому результаты оказываются сформулированными на языке различных потребителей результата: других аналитиков, бизнес-менеджеров, инвесторов и т.д. На этом этапе выделяются ключевые результаты и как правило продолжаются работы по использованию и совершенствованию или даже переработке моделей.

Использование

Это фаза аналитического проекта представляет собой совокупность процессов по подготовке отчетов как технического, так и бизнес содержания. Запускается пилотный проект по анализу вновь поступающих данных и модель вводится вбизнес процессы организации. Big Data Managmentпредставляют собой неисчерпаемое поле для аналитических проектов. В этом смысле они могут рассматриваться как дальнейшее развитие бизнес-аналитики (BI) с возможностями обработки несоизмеримо большего объема самых разнообразных данных, получаемых из различных источников как внутри, так и вне организации. Однако сила больших данных заключается в возможностях построения предиктивных (предсказательных) систем для очень сложных взаимосвязанных естественных и порожденных человеческой деятельностью процессов. Возможности инструментов для больших данных сегодня позволяют строить такие предсказательные системы, работающие в реальном масштабе времени. Это позволяет встраивать их в сложные технико-организационные системы для поддержки механизмов автоматического управления. В качестве весьма важного примера мы выбрали применение больших данных в построении сложных систем в электроэнергетике.

Тема № 12. Big Data Managmentв электроэнергетике

В этой главе вы найдете ряд важных и интересных приложений подхода, техники и технологий больших данных к задачам современной электроэнергетики. Все рассматриваемые приложения изложены на примерах конкретных примеров проектов, выполненных или продолжающихся в США и других странах. Авторы старались использовать в изложении русскоязычную терминологию, которую предлагают авторы объемистого коллективного труда институтов РАН и предприятий российской электроэнергетики, [24] однако в ряде случаев, мы будем пользоваться терминами, принятыми в англоязычной литературе без перевода, если перевод выглядит неочевидным или неуместным.

Интелеллектуальная электроэнергетика

В настоящее время электроэнергетические системы в большинстве стран модернизируются и развиваются на основе концепции глубокой интеграции электроэнергетических сетей (Power Grid) и сетей компьютерных или как их называют инфокоммуникационных (Network). При этом оба вида сетей не просто развиваются и обогащаются новыми функциональными элементами и протоколами взаимодействия, а порождают глубокий синергетический эффект, связанный с невиданными ранее возможностями анализа состояния целой огромной энергосистемы в реальном времени, прогнозирования процессов в ней, интерактивного взаимодействия с клиентами и управления оборудованием. Такая концепция получила название Smart Grid – интеллектуальная энергосеть. Общую функционально–технологическую идеологию этой концепции, отражает сформулированное IEEE определение SmartGrid как концепции полностью интегрированной, саморегулирующейся и самовосстанавливающейся электроэнергетической системы, имеющей сетевую топологию и включающей в себя все генерирующие источники, магистральные и распределительные сети и все виды потребителей электрической энергии, управляемые единой сетью информационно-управляющих устройств и систем в режиме реального времени.

Концептуально, Smart Grid обычно иллюстрируют картиной, подобной изображенной на рисунке 3.1.

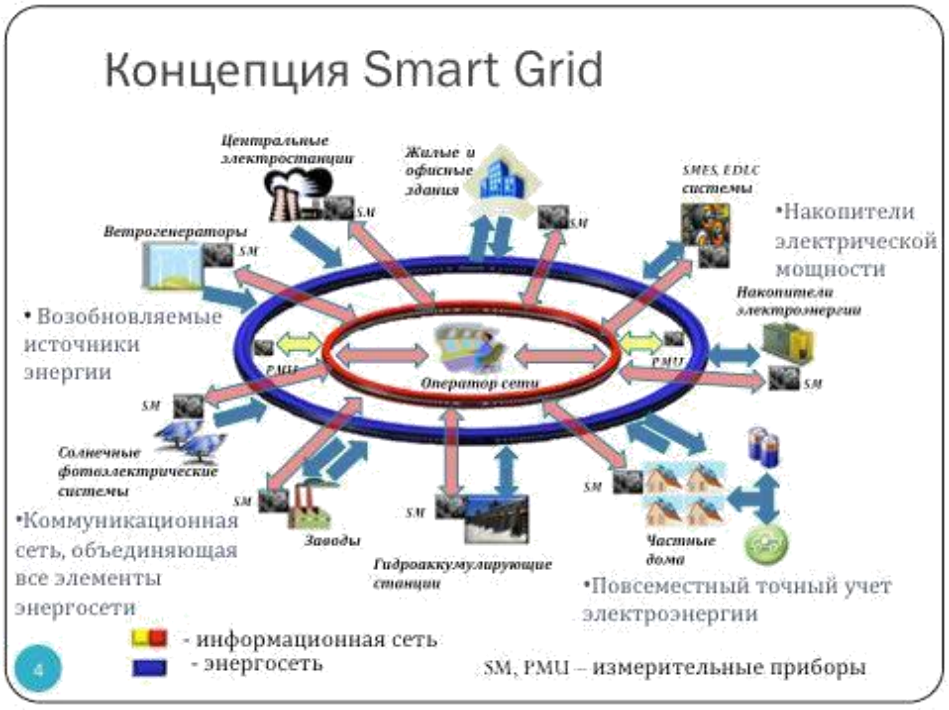


Рисунок 3.1 Smart Grid как объединяющая инфраструктура электроэнергетики

Преодолев за последние годы период исследований и начального развития, эта концепция стала де-факто стандартным подходом как для игроков электроэнергетического рынка, так и для правительственных регуляторов. В РФ этот подход находится в начальной фазе развития и реализуется в концепции Интеллектуальной электроэнергетической системы с активно-адаптивной сетью (ИЭС ААС), под которой понимается система, в которой все субъекты электроэнергетического рынка (генерация, сеть, потребители) принимают активное участие в процессах передачи и распределения электроэнергии. Всоставе ИЭС электрическая сеть из пассивного устройства транспорта и распределения электроэнергии превращается в активный элемент, параметры и характеристики которой изменяются в зависимости от режимов работы энергосистемы. Практическое воплощение данных направлений осуществляется во взаимосвязи и развитии положений стратегических

 документов ОАО «ФСК ЕЭС» – Программы инновационного развития и Политики инновационного развития и модернизации. Согласно документам программы можно выделить особенности новой концепции энергосетей.

Основные новые качества ИЭС ААС:

 Обеспечение равного доступа любых производителей и потребителей электрической энергии к услугам инфраструктуры. Создание специальных интерфейсов для унифицированного и надежного подключения к сетям возобновляемых и нетрадиционных источников энергии на условиях параллельной работы в составе энергосистемы.

 Участие в управлении режимом работы ИЭС генерации, управляемых элементов сетевой инфраструктуры, потребителей электроэнергии.

Обеспечение «активности» потребителей электроэнергии за счет их оснащения интеллектуальными системами учета с возможностью ситуативного управления спросом. Обеспечение за счет применения этих систем рационального использования энергии в нормальных режимах и управления потреблением электроэнергии с целью поддержания требуемых параметров функционирования ИЭС.

 Наличие достаточных объемов информации о текущем состоянии электрическойсети и ее элементов (включая векторные измерения), и о внешней среде(освещенность, осадки, гололед, ветровые нагрузки и другие метеофакторы), атакже современной системы управления, позволяющей в реальном временииобрабатывать указанную информацию.Обеспечение максимальной самодиагностики элементов ИЭС, использование ее результатов в алгоритмах функционирования автоматических систем режимного и противоаварийного управления.Наличие распределенных и иерархических централизованных систем режимного и противоаварийного управления, основанных на адаптивных алгоритмах реального времени.

 Применение быстродействующих программ и вычислительных ресурсов, обеспечивающих как выработку автоматических управляющих воздействий, так и предоставление рекомендаций (с помощью экспертных и других систем) диспетчерскому, оперативно- технологическому и ремонтному персоналу для реализации управляющих воздействий и проведения необходимых работ.

Целевые функции новой концепции развития сетей электроэнергетики Smart Grid в более общем виде отражены в документах Министерства энергетики США (DOE - Department of Energy) [25]. Согласно этим документам Smart Grid нацелена на удовлетворение ключевых ценностей (keygoals) :

Доступность -обеспечение потребителей энергией без ограничений взависимости от того, когда и где она им необходима, и в зависимости от оплачиваемого качества.

Надежность -возможность противостояния физическим иинформационным негативным воздействиям без тотальных отключений или высоких затрат на восстановительные работы, максимально быстрое восстановление (самовосстановление) работоспособности;

Экономичность -оптимизация тарифов на электрическую энергию дляпотребителей и снижение общесистемных затрат.

Эффективность -максимизация эффективности использования всехвидов ресурсов, технологий и оборудования при производстве, передаче, распределении и потреблении электроэнергии.

Органичность взаимодействия с окружающей средой -максимальновозможное снижение негативных экологических воздействий;

Безопасность –не допущение ситуаций в электроэнергетике,опасных длялюдей и окружающей среды.

На рисунке 3.2 представлено взаимодействие составляющих частей в Smart Grid.

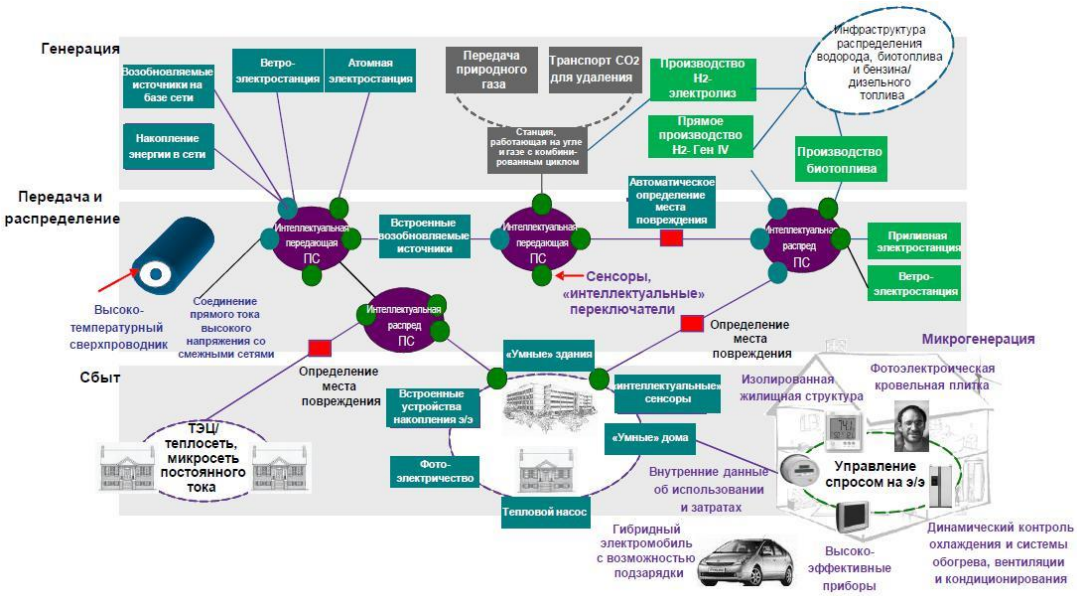


Рис.3.2 Взаимодействие участников интеллектуальной электроэнергетической системы на основе концепции Smart Grid

Столь сложное и разнообразное взаимодействие регулируется протоколами, охватывающими от физического до прикладного уровня, обеспечивая эффективное взаимодействие и поддерживая технический прогресс в совершенствовании всех компонент энергосистемы. Отметим, что несмотря на существование общей концепции и ее поддержку во всех странах, существует национальная специфика в требованиях на применение различных протоколов на сетях того или иного государства. Особо выделяется здесь требование регулятора в Германии, где необходимо обеспечить открытые протоколы взаимодействия, так, чтобы оборудование и программное обеспечение любых поставщиков могло работать совместно, позволяя потребителю выбирать поставщиков на конкурентном принципе, исходя из собственных требований и потребностей. В последующем изложении мы остановим внимание только на тех сторонах интеллектуальных энергосистем, которые касаются обработки информации с целью придания этим системам новых отличных от традиционных энергосистем функций. Другие важные аспекты, такие как архитектуры и протоколы дляпостроение инфокоммуникационных сетей, управления генерирующим оборудованием и звеньями сети энергопередачи останутся в стороне нашего изложения.

Системы обработки данных в интеллектуальных энергосетях

Чтобы понять, почему современные системы обработки данных в интеллектуальных энергосетях развиваются в направлениях, оперирующими с большими данными, нужно, прежде всего, исходить из базовых трендов развития современной электроэнергетики. Выделим здесь несколько из них, которые оказывают наибольшее влияние на информационную составляющую.

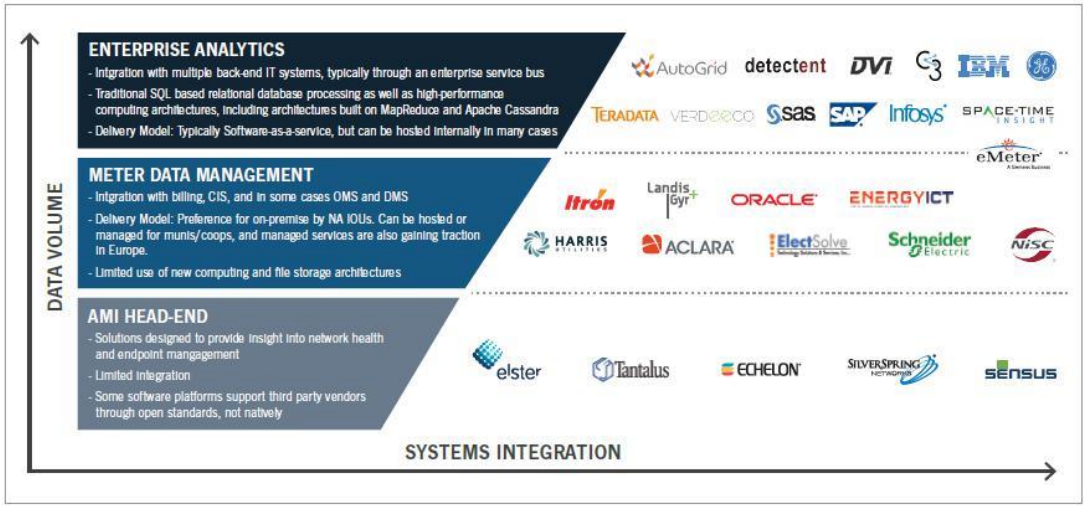
Во-первых, это постоянно растущие требования как к экологичности так и к повышению энергетической эффективности энергосистем и сетей. Это привело к, использованию широкого спектра источников возобновляемой энергии, таких как ветрогенераторов и солнечных электрогенераторов. Однако, в силу нестабильного характера их характеристик генерации появились требования к эффективному быстрому управлению нагрузками в энергосистемах. Наряду с динамичностью, здесь особенно нужно отметить и бурный рост распределенности узлов генерации по сети. Практически каждый включенный в сеть потребитель может оказаться источником энергии, если его

потребности окажутся на каком-то интервале времени меньшими, чем возможности принадлежащих ему локальных электрогенераторов. Характерно, что существенная распределенность и высокая динамичность появляются даже в тех случаях, когда в работе сети участвуют не только возобновляемые источники, но более традиционные, но современные, такие как набирающие популярность микрогазовые турбины, биогенераторы. Во-вторых, аспект связан с появлением и внедрением на сетях накопителей электроэнергии различного уровня присоединения. Сегодня появление аккумуляторных батарей поддержки группы домохозяйств становится весьма популярным решением для выравнивания энергопотребления и переноса пиков потребления с часов повышенной оплаты на другие интервалы времени менее дорогого тарифа. Таким образом главными катализаторами развития систем больших данных в электроэнергетике являются:

Активное управление нагрузочными характеристиками потребителей Эффективное управление распределенными системами генерации с большим числом источников Эффективное управление и мониторинг для многочисленных динамичныхнестабильных системам генерации.Но, разумеется, более традиционные задачи энергосистем и сетей также не остаются вне сферы обслуживания новых систем обработки данных. Прогнозирование нагрузки и эффективное управление элементами сети с целью энергосбережения и предотвращение перегрузок

Гибкая тарификация и детектирование утечек и хищений Управление имуществом и техническое обслуживание Функционирование в режиме устранения аварии. Все системы, предназначенные для обработки данных в энергосетях, как правило основаны на сборе информации от многочисленных (миллионов) измерителей в том числе интеллектуальных приборов учета потребителей (Smart Meters) и промышленных приборов и систем учета (АСКУЭ). Также теперь это всевозможные сенсоры на оборудовании энергосети, отражающие работу и состояние оборудования, сенсоры и информационные системы погодных условий и климата. Ввиду исключительно больших объемов обрабатываемых данных, их разнообразия по форме и семантике, требований к оперативной и целостной их обработки, компьютерные системы, дающие возможность имплементировать новые функции в энергосети и энергосистемы, относятся к категории Big data (больших данных). Для выделения систем и технологий, относящихся к обработке данных в Smart Grid используется термин Soft Grid. Анализ текущего состояния и наиболее успешных примеров из Soft Grid изложен в последующих параграфах.

Рыночный срез поставщиков решений по секторам, определяемым объемами обрабатываемых данных представлен на рисунке 3.3. Диаграмма не претендует на исчерпывающее представление всех игроков, однако достаточно репрезентативна.



SOURCE: GTM RESEARCH

Рисунок. 3.3 Таксономия поставщиков решений для информационных систем в электроэнергетике.

Тема № 13. Обзор проектов Soft Grid, ориентированных на аналитику в распределенных системах

Решение задач в этой области относят к проблемам, обозначаемым термином“аналитика на сетях” – Grid Analytics и “аналитика для клиентов” – Customer DataAnalytics. На конференции Greenmedia 2013 [27] отмечалось, что аналитика на сетях принесет в проекты интеллектуальных сетей большую долю инвестиций, которая составит не менее 8,7 млрд. долларов к 2020 году. Главной задачей здесь ставят оптимизацию распределенных сетей с помощью программного управленияаналитики. Решение задач построения эффективных алгоритмов опирается на использование весьма точных имитационных моделей сети, таких как GridLab-D разработанной по заказу министерства энергетики США [28] и многочисленных данных, собираемых как с существующих смарт счетчиков, так и устройств цифровой защиты и управления, скомбинированных с программными решениями и коммуникациями, чтобы управлять ими. В то же время предприятия поворачиваются к новым классам сетевого оборудования, которые не могут гладко войти в существующие режимы работы. Это - солнечные батареи, ветровые генераторы и другие распределенные источники энергии, накопители энергии, которые балансируют локальное потребление на основе изменяющихся условий и запросов от мириад источников. Интеграция, оптимизация и менеджмент данных для управления распределенными энергоресурсами является достойной задачей для изощренных алгоритмов и технологий больших данных.

Аналитика для клиентов оценивается инвестиционным потенциалом не менее чем в 7.1 млрд. долларов. Эксперты выделяют три главных направления аналитических систем, ориентированных на клиентов:

защита доходов,тонкое предсказание нагрузкидетальная сегментация пользователей.

Проекты в этой сфере позволяют генерировать гибкие ценовые программы, привлекательные потребителям, проводить хорошо таргетированные маркетинговые мероприятия, строить модели точного поведения клиентов для дальнейшей адаптации ценовой политики.На этом рынке идет постоянный рост проектов как от стартапов, так и от известных игроков в электроэнергетике. Например, General Electric, в январе 2013 года запустила свою Big Data платформу Grid IQ Insight [29] как инструмент для консолидации всех данных от существующих систем управления сетями, интеллектуальных приборов учета и сенсоров сети вместе с неструктурированными данными о погоде и лентами постов из Facebook и Twitter.Корпорация GE работает со многими предприятиями энергетики в разных странах, доказывая ценность своей платформы путем быстрого развертывания и прототипирования для различных применений от станций зарядки электромобилей и мониторинга распределенных солнечных solar “hot spot” до анализа роста растений, чтобы помочь предприятиям держать линии электропередачи свободными от веток деревьев. В начале 2014 года GE запустила свою Инновационную программу в энергопредприятиях дляобслуживания предприятия включая AEP и Indianapolis Power & Light, которым это помогает стать узловыми партнерами путем обменов данными и идеями. В ближайшее время GE выводит несколько продуктовых аналитических пакетов, поддерживаемых платформой Grid IQ. В первую очередь эта платформа будет основой для понимания

измерений “meter insight”, построенной для получения преимуществ от доступа к массе источников, в большей части еще неиспользуемых данных, от десятков миллионов интеллектуальных приборов учета и счетчиков, развернутых по всему миру.

Платформа Grid IQ Insight бурно развивается и в ближайшие месяцы она возможно будет первым аналитическим инструментом, дающем GE новую экспертизу, простирающуюся от широкого диапазона сетевых продуктов от трансформаторов и подстанций до программных платформ мониторинга и управления сетями. Объединяя данные от интеллектуальных измерителей с помощью специального программного обеспечения, платформа позволяет реализовать разнообразные функции, основанные на весьма тонком анализе с использованием Data Mining и других технологий и техник больших данных .

Среди аналитических функций, основанных на измерениях во многих точках, GE предлагает весьма эффективный способ детектирования хищений энергии – сравнивая данные приборов учета, на подстанциях и сенсоров на трансформаторах определяя точки, где энергия может быть потеряна или украдена. И здесь особое значение имеет работа в пятнадцатиминутных интервалах для лучшего предсказания нагрузки.

Важно заметить, что развернутые GE миллионы интеллектуальных измерителей, могут использовать оборудование других производителей и успешно работают в сетях AMI (Advanced Metering Infrastructure), установленных другими поставщиками, такими как Silver Spring Networks в США, или Grid Net в Австралии. Решения GE полностью агностичны к тому какое оборудование будет использоваться для сбора данных от интеллектуальных измерителей.

В тоже время GE запустила и свой собственный Smart Metering Operations Suite (SMOS) [30] , который описывается как завершенная измерительная и управляющая система данными и транзакциями. SMOS включает в себя средства мониторинга, управления, поддержания работыGE усиливает свое сотрудничество с компаниями по партнерству C3 Energy [31] . Ниже мы подробнее рассмотрим продукты этой компании. А в апреле 2013 года корпорация инвестировала $105 млн в компанию Pivotal [32] , дочку VMware и EMC разработавшую облачную платформу аналитики данных. На первый взгляд непонятен смысл таких инвестиций в конкурирующий продукт, поскольку это партнерство может повлиять на работы по собственной разработке Grid IQ Insight. Но по-видимому, это объясняется тем, что все big data проекты являются пилотными, и трудно оценить заранее их успешность, Поэтому следует делать этот рынок наиболее разнообразным, чтобы как можно быстрее идти к большимконтрактам с энергетиками в будущем, в том числе в направлении строительства ими собственных хранилищ данных и формирования типовых промышленных применений. Тогда возглавив или принимая участие в нескольких проектах компания имеет больше шансов стать успешной на новом рынке. Один из больших вызовов, который в ближайшее время ожидается в аналитике данных в энергетике - это кто построит самую большую аналитическую библиотеку и работающую быстрее всех. Путь к победе лежит через развертывание разнообразных подходов и скорейшее их применение в реальных проектах.

Провайдеры рынка аналитики данных , такие как Oracle, EMC, SAP, IBM, Teradataподобные, нацелились на интеллектуальные сети используя тот факт, что предприятия в энергетике еще не погрузились достаточно глубоко в поток больших данных. Судя по известному отчету Nucleus Research 2011, на каждый вложенный в аналитику доллар получается ROI (Return of Investment) возврат почти 11 долларов (10.66) при рассмотрении более 60 развертываний. Видимо предприятия энергетики могут ожидать такого же ROI. Однако в докладе GTM Research’s Soft Grid 2013-2020 показано, что к 2020 году для этой отрасли уровень возврата инвестиций вряд ли превысит 7 долларов на один вложенный. Причины в сильном регулировании в отрасли и невозможности рассчитывать на рост потребления, как например, в торговле. Вместо этого в энергетике следует ожидать львиной доли возврата от вложений в аналитику от сокращения затрат, улучшения

эффективности, и менеджмента в возникающих бизнесах как использование зеленых технологий и охраны окружающей среды, а также все большего появления пользователей с собственными источниками энергии, как солнечные батареи на крышах.

Рассмотрим еще некоторые проекты в области аналитики энергосистем, ведущиеся крупными компаниями в области платформ для интеллектуальных измерителей.

Подразделение Smart Grid компании Siemens [33] и крупный игрок на рынке больших данных - корпорация Teradata [34] объявили о глобальном стратегическом сотрудничестве в области больших данных. Благодаря этому тандему Siemens Smart Grid сможет оптимизировать свой портфель решений, нацеленных на повышение прозрачности операционной деятельности сетевых и сбытовых энергокомпаний. Для заказчиков Siemens Smart Grid это даст возможность повысить надежность своей инфраструктуры и более эффективно управлять своими энергосетями Применяемая унифицированная архитектура данных Unified Data Architecture, являющаяся одной из главных идей корпорации Teradata, является основой для реализации управления, обработки и анализа данных, позволяющей энергокомпаниям эффективно использовать свои большие данные. Данные такого объема неизбежно возникают, когда энергокомпания начинает управлять обновлённой инфраструктурой на основе решений Siemens Smart Grid, сочетающей в себе расширенные средства автоматизации, датчики нового поколения, коммуникационные системы и программные приложения.

Такие заказчики Teradata, как Southern California Edison и Oklahoma Gas and Electric, уже обрабатывают и анализируют огромные объемы данных, что позволяет им предоставлять потребителям современные услуги. Энергокомпании теперь могут быстро оценивать затраты и время, необходимое для восстановления энергоснабжения в случае аварии, и держать в курсе событий своих клиентов. По информации Teradata, запросы о потерях при передаче энергии с разбиением по типу производителя, географическому местоположениюпогодным условиям позволяют лучше планировать структуру и распределение нагрузки энергосети. Техническое обслуживание теперь можно проводить на основании данных о фактическом состоянии и износе оборудования, а не в соответствии с утверждённым графиком. Использование географических данных позволяет более эффективно использовать ремонтные бригады.

Еще один гигант – компания Эриксон (Ericsson) [35] становится оператором более 600 тысяч точек интеллектуальных счетчиков, что увеличит доставку данных для E.ON – одной из крупнейших энергетической компаний в мире на 3000%. Smart Grids буду обеспечивать более эффективное управление энергосетью и опрелять интеграцию с возобновляемыми источниками энергии. Еще один шаг в направлении Сетевого Сообщества, где Big Data Managmentи облачный компьютинг определяют контуры будущего для индустрии энергетики, транспортировки и ИТ.

Решения Эриксон будут собирать данные с интеллектуальных счетчиков и поставлять их в производственную ИТ сеть E.ON [36]. Данные будут экспортироваться ежедневно, что по сравнению с ежемесячными измерениями увеличивает их объем на 3000 процентов. Решение Эриксон по интеллектуальным измерениям развертывается в концепции как сервис и позволяет E.ON стать более эффективной. Решение комбинирует управляемые сервисы, консалтинг и и системы интеграционных сервисов, включающих считывание с интеллектуальных счетчиков и управление ими, мониторинг элементов сети измерений. Соглашение об уровне сервиса SLA (Service Level Agreement) определяет менеджмент рабочей силы, имущества, бизнес процессов как в отношении сервисов на местах, так и в центре управления. Это важный шаг к гладкой интеграции локально получаемой энергии от возобновляемых источниковинфраструктуру существующих и перспективных сетей. Принятое законодательство в Евросоюзе требует от энергопроизводителей или распределенных системных операторов развернуть интеллектуальные системы сбора данных о потреблении энергии для того чтобы делать шаги к увеличению эффективности использования энергии. Президент Эриксон по

северной Европе ицентральной Азии Robert Puscaric сказал: «По мере того как Сетевое Сообщество приходит в жизнь, и мы на пути к соединению наших 50 миллиардов девайсов, мы уверены, что большой частью M2M (Machine-to-Machine) коммуникацийбудут Smart Grids и Smart Meetering». Ericsson также развертывает сети и сервисы для Smart Grid в Эстонии и Италии.

E.ON – одна из крупнейших частных энергетических компаний в мире выбрала Эриксон для консалтинга, решений системы интеграции сервисовинтеллектуальных энергосетей для своих клиентов не случайно. Эриксон будетсобирать данные от интеллектуальных счетчиков и обрабатывать в своейвнутренней сети на основе использования вычислительных ресурсов компании.Одновременно E.ON исследует пути, исключающие прямой аутсорсингвычислительных процессов в сторонней компании. Строить собственные дата-центры? Какой наилучший способ для энергетических компаний Германиискомбинировать неясное будущее сетей сбора данных от интеллектуальныхизмерителей и непрекращающегося давления управлять самой большойразделенной мощностью солнечных электростанций? Совместно с известнымлидером в области облачных решений компанией IBM корпорация E.ONанонсировали потенциальное решение, которое могло бы преодолеть границы, ипоказать, как облачная инфраструктура может быть применена для решениязадач в области интеллектуальных энергосетей в интересах предприятийэнергетики для постоянно нарастающих вызовов со стороны больших данных.

E.ON ,будет базировать инфраструктуру интеллектуальных счетчиков надатацентре IBM в Германии, используя ее Smart Cloud и Intelligent Energy ServiceEnablement Platform (IESEP). Целью является улучшение развертывания именеджмента интеллектуальных счетчиков, упрощение интеграции свозобновляемыми источниками энергии и другими инновационными сервисамиодновременно с улучшением уровня услуг для потребителей энергии.

E.ONвладеет 68 гигаваттными источниками энергии в Европе и обеспечивает ею 26миллионов потребителей. Несмотря на гигантские масштабы корпорация несобирается развертывать множество компьютерных центров, а начинаетиспользование частного облака для клиентов в Германии, Австрии и Швейцарии.Первыми шагами для ориентированных на клиентов приложениями являются«профили использования» для информации о «time-of-use-rates”» т.е. типичныеизменения нагрузки во времени, характерные для каждого отдельногопользователя и представленные с помощью специального кодирования. Иногдаих называют паттерны потребителя, которые могут быть сравнены систорическими данными. Эти свойства являются весьма общими приразвертывании ителлектуальных счетчиков в настоящее время. Предприятия,кроме того, предложили включить в патерны специфические детали кромеиспользуемой динамики нагрузки, которые реализуются и в ориентированных наклиента (customer-facing) приложениях, так и во внутри корпоративных бизнес-кейсах для платформ. Big Data Managmentи аналитика продолжают бытьдрайверами новых инноваций и эффективного использования каждый новый проект вводит инновации в информационное окружение энергосистем. Siemens и Teradata предлагают сквозную интеграцию операционных данных с данными интеллектуальных счетчиков для их последующего анализа на единой платформе, реализуя тем самым новый подходэксплуатации и развитию энергосетей. Компании совместно разрабатывают модели данных на основе логической модели данных Teradata Utilities Logical Data Model [37] , являющейся фундаментом бизнес-аналитики для энергетической отрасли. Эта модель предоставляет структуру и стандарты данных для решенияважных вопросов бизнеса, делает данные доступными, обеспечивает их повторное использование в других приложениях, а также позволяет быстро и надежно передавать информацию потребителям энергии и надзорным органам.

IBM разработала технологию Hybrid Renewable Energy Forecasting (HyRef) [38] которая предназначена для эффективного внедрения возобновляемых источников энергии в

сети энергоснабжения. Инструмент использует комбинацию продвинутой системы предсказания погоды с использованием анализа изображений облаков и камеры смотрящие в небо. Приложение может быть сфокусировано на гранулярность до единственной ветровой турбины. HyRef использует Big Data Managmentи аналитику чтобы смоделировать и предсказать когда и как будут работать турбины. Точность предсказания по выработке мощности достигает 10%. Энергетическая компания в Китае State Grid Corporation of China (SGCC), сообщила о том, что уже развернула для применения систему HyRef для эффективной интеграции с сетью источников возобновляемой энергии. На рисунке 3.4 приведен фрагмент рекламного постера, раскрывающего преимущества HyRef. Пилотный проект на 670 Мегаватт в городе Zhangbei является самым крупным в мире, который комбинирует ветровую и солнечную энергетику, накопители энергии и инновационные передающие системы. Развертывание уникальной системы больших данных осуществили исследователи из IBM Lab в Китае и TJ Watson Research Lab in Yorktown, NY.

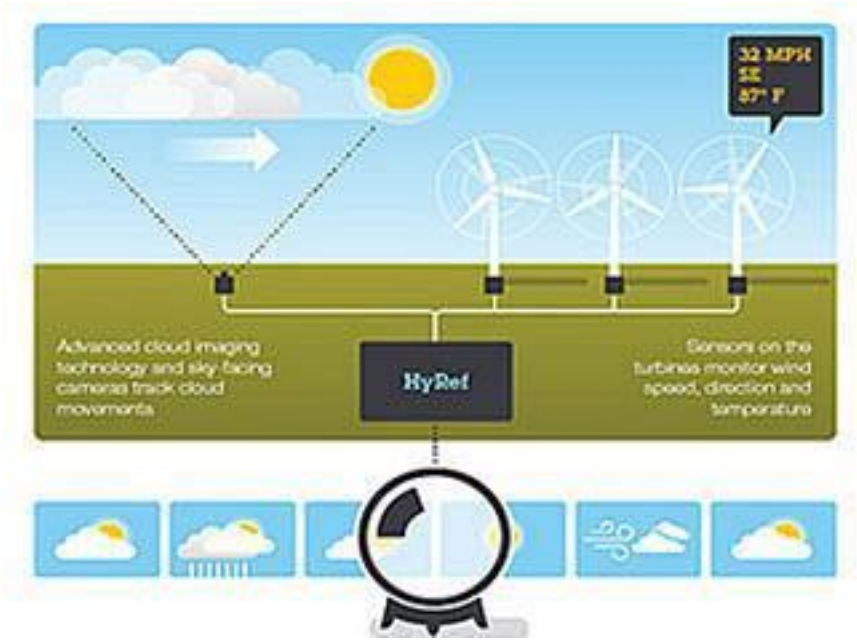


Рисунок 3.4 Рекламный постер системы HyRef поясняющий возможности весьма точного учета погодных условий для стабилизации сетей с зелеными источниками.

На рынке компаний, специализирующихся на программных продуктах для энергетики, заметное место занимает компания C3Energy [31] .

В последнее время ею проведена большая работа по применению своего программного решения агрегирования больших данных и аналитики к интеллектуальным сетям. Последние четыре приложения, анонсированные в этой сфере, реализованные над ядром обрабатывающей системы (data engine), уже используются большим числом корпоративных и институциональных клиентов для планирования энергии и имущества.

Первыедваприложениеназываются C3 Asset and System Risk и C3 Grid Investment Planning . Первое из них позволяет постоянно оценивать состояние работоспособности для имущества для снижения операционных расходов, предотвращения отказов и минимизации непредвиденных капитальных расходов.

Второе приложение на основании данных всех видов (денежные потоки, лояльность клиентов, влияния на сеть) имущества оценивает ранг различных видов инвестиций в сеть. Позволяет информировать о необходимости затрат на замену элементов сети.

Еще одно приложение - C3 AMI Operations предназначено помочь персоналу установке, вводу в эксплуатацию и обслуживанию интеллектуальных измерителей и связывающей их сети.

Бурно растущий рынок привлекает все новых игроков. Многим из них сопутствует успех. Компания AutoGrid Systems [40] – калифорнийский стартап, основанный в 2011 году, построила свою платформу для работы с большими данными для интеллектуальных сетей. Недавно японский гигант компания NTT Data [41] приобрела лицензию на эту платформу. Собственный продукт на основе этой платформы будет интегрировать интеллектуальные измерители и управление demand response для японских предприятий и зданий. Этобудетблизкоктому, чтоназывают demand response optimization and management system (DROMS). Однако продукт будет иметь специфические черты.

Платформа AutoGrid использует интеграцию данных от различных источников используя unstructured data engine –инструмент интеграции данных более высокого уровня. Корневой технологией платформы является Energy Data Platform

– облачное ядро (Cloud Engine) аналитики и менеджмента неструктурированных данных. Оно берет данные от множества источников и применяет предиктивный алгоритм реального времени к потоку. Компания является заметным генератором идей и терминов. Так, например, в понимании компании Big Data = Device Control, Business Intelligence, More.

Уже работающее приложение на основе платформы Energy Data Platform было недавно продемонстрировано для работы в составе DR системы demand response повышенной эластичности.

Партнеры планируют добавить новую аналитику и инструменты бизнес-менеджмента которые могли бы ранжировать прогнозирование потребление энергии клиентам и поведенческую аналитику (behavioral analytics) чтобы интегрировать ее с управлением потреблением (DR), управлениемраспределенной генерацией и другими неизвестными ранее функциями. Существует несколько сценариев как будет идти развитие рынка электроэнергетики, но в любом случае, по мнению многих экспертов, нужно идти по пути открытия целого набора стратегических вопросов обо всех компонентах цепочки ценностей (value chain) в германском дерегулированном энергетическом ландшафте. E.ON уже вовлечена в похожий проект по cloud-based smart metering с Ericsson, в рамках которого управляется в реальном времени более 600 тысяч точек измерений. Но такие проекты сфокусированы на страны с почти завершенным покрытием всей национальной энергосети системами интеллектуальных измерителей (smart metering).

В отличие от Франции, Италии и Великобритании германское правительство отклонило принятие рекомендации Евросоюза “продавить” интеллектуальные счетчики в массы до 80 процентов резидентов к 2020 году. Вместо этого в Германии ищут более оптимальную модель smart metering, основанную на потребностях клиентов и игроков дерегулированного рынка энергии. Выбор лежит в технологиях подключения клиентов к сети и строгим требованиям к безопасности данных и основанной на стандартах требований интероперабельности. Регулятор в Германии в настоящее время требует установки интеллектуальных приборов учета во всех новых зданиях и у потребителей более 6 мегаватт, а также для площадок, продуцирующих более 7 мегаватт-часов в сеть распределенной генерации. Учитывая, что пока развертывание интеллектуальных приборов учета не было эффективным по стоимости, планируется установка таких приборов только у 23 процентов хозяйств к 2022 году. Однако немецкий план включает в себя другие компоненты, которые отсутствуют в других национальных планах. Это в частности идея центрального хаба или домашнего шлюза (in-home gateway) который связывает водяные, газовые, и электрические измерители и другие домашние приборы с магистральными сетями поставщиков. Так называемый BSI gateway , названный в честь немецкого министерства безопасности должен обеспечивать стандартное безопасное подключение приборов многих производителей [42]. Это обусловлено сильно дерегулированным энергетическим рынком Германии, где необходимо обеспечить присутствие многих продавцов энергии с их разнообразными приборами и услугами на основе единой сети. Правительство Германии пытается быть драйвером инноваций требуя развертывания

открытой платформы, на основе которой потребители с различными требованиями могут найти подходящих поставщиков.

В то же время Германия является страной с наиболее развитой сетью распределенных солнечных станций с общей мощностью в 34.5 гигаватт летом, обеспечивающей около 7 процентов национальной мощности в первой половине 2013. Так что проблемой является постоянное непредсказуемое изменение входящих в сеть производителей энергии, что требует выполнения требований по стабилизации сетевых элементов. Здесь имеется набор требований по Умным инверторам (Smart Inverter) которые могут помочь балансировать распределенные солнечные электростанции при включении их в общую сеть. Всеэто связано также с развивающимся рынком накопителей энергии которые также помогают снижать влияние большого числа солнечных батарей на распределенные сети. В целом требуется балансировать работу сети с изменчивыми солнечными батареями и потребительскими паттернами использования энергии не затрачивая много капитальных ресурсов. По мере того как потребители становятся «энергетически сознательными» (energy-conscios) и технологии обеспечивают больше информации, становится возможным получить бенефиты и для потребителей и для поставщиков. Столь высокая сложность алгоритмов обработки и гигантские объемы данных, требуют организации эффективных вычислительных процессов для их реализации.

Последние несколько лет наблюдается медленный, но постоянно нарастающий переход энергопредприятий на облачные компьютерные архитектуры. Это происходит весьма медленно, потому что энергопредприятия все еще опасаются переводить свои ИТ потребности в облако, особенно такие как управление сетью.

В то же время масштабируемость, гибкость, и высокая адаптивность к новым задачам, характерные для облачных решений, являются весьма притягательными для энергопредприятий, нуждающихся в менеджменте двухсторонних потоков как информации, так и энергии между ними и технологически оснащенными потребителями. На фронте интеграции данных и энергии в сети дом-энергопредприятие, корпорация IBM является сегодня ключевым игроком в известном проекте Model City Mannheim , где отрабатываются и тестируются технологии будущего для интеллектуальных сетей для всей Европы. Главный вопрос, который исследуется, это как много функций сети можно передать в облако, которое хотя и является частным, но сосредотачивает в себе весьма много ключевой информации обо всех игроках.

Специфические задачи, стоящие перед предприятиями отрасли рассматривались на конференции SOFTGRID 2013 1-2 октября 2013 года под названием :

“Предприятия энергопроизводства: Общая картина через большие данные” [43].

Судя по докладам экспертов, до 2020 года прогнозируется инвестиционный объем до 4,2 млрд долларов в интеграцию всех данных внутри предприятий для полноценной аналитики. Секция “Разработка бизнес-кейсов для больших данных и аналитики для современных сетей” включала доклады EMC , Oracle и стартап EcoFactor и AutoGrid в которых фокус был на перводе операций на предприятиях в облачные вычисления и как продать аналитические возможности в департаменты производства энергии.

Индустрия smart grid растет и имеет много конкурентов от стартапов, таких как AutoGrid, C3 Energy and Verdeeco, до гигантских корпораций как IBM, Oracle, Siemens и General Electric – которые стремятся завладеть рынком “big data” уверенно приходя к предприятиям энергетики по всему миру. Все они глубоко вовлечены в информационные обмены и тестирование процессов, основанных на больших данных, в бизнес-процессы энергоотрасли, и многие уже двигаются вперед к big-data-ready продуктам коммерческого масштаба для интеллектуальных энергосетей. Вот еще несколько примеров.

Компания стартап Silver Spring Networks [44] предлагает продукт под названиемUtilityIQ suite of analytics, которым заинтересовались крупные поставщики smart meter такие как Itron и Elster. В состав этого продукта входит целый набор программного обеспечения для решения различных задач энергопредприятий.

UtilityIQ Advanced Metering Manager – автоматизирует дорогостоящие и времяемкие процессы сбора данных от измерителей. Поддерживая постоянное соединение с приборами измерения продукт позволяет увеличить точность сбора данных и производить удаленное подключение и отключение отдельных приборов для обслуживания. Также поддерживается детектирование отказов и нарушения изоляции. С помощью UtilityIQ Metering данные интегрируются в систему MDM энергопредприятия.

UtilityIQ Power Monitor обеспечивает проактивный режим предупреждений об изменениях уровней напряжения за пределы установленного предприятием порога для каждого индивидуального измерителя с фиксацией его местоположения. Это программное обеспечение усиливает интеллектуальность распределенной системы, поскольку позволяет следить только за измерителями с данными, отклоняющимися от нормальных.

UtilityIQ Outage Detection System фиксирует время, тип, местоположение отказа в сети, коррелируя с другими точками отказа, что позволяет существенно ускорить восстановление. Данные об отказах, коррелированные по множествам местоположения сетевых элементов позволяют предприятию быстрее идентифицировать и выявлять проблемы в сети, автоматизировать процессы составления спецификаций отказов и отчетов об отказах и восстановлении. Этот продукт также может быть легко интегрирован в OMS back-office системы энергопредприятия подобно

UtilityIQ Advanced Metering Manager

UtilityIQ Demand Response Manager предназначен для поддержки HAN device control, DR program management и аналитики, нужной для имплементации программ управления нагрузкой как по цене энергии, так и по прямым значениям нагрузки. Этот продукт работает совместно с IQ Advanced Metering Manager и Silver Spring CustomerIQ energy web portal для реализации алгоритмов управления в соответствие с общей ситуацией на сети и соглашениями с потребителями.

UtilityIQ Network Element Manager UtilityIQ Firmware Upgrader программные модули, которые обеспечивают различные возможности реконфигурирования системы и апгрейда. Еще один модуль – GridScape for DA application обеспечивает функции конфигурирования, менеджмента и функций безопасности для сетей распределенного автоматического управления (Distribution Automation networks).

Интересные продукты компании eMeter [45] , демонстрируют возможности весьма гибкого построения информационных систем для энергопредприятий. Компания принадлежит концерну Siemens, который в свою очередь имеет

партнерские отношения с известным поставщиком решений в области Big Data компанией Teradata. Самый знаменитый продукт компании – Energy IP – гибкая масштабируемая платформа для Smart Grid приложений. Корневые функциональности системы разработчики определяют как «центральную нервную систему» для энергопредприятий. Платформа состоит из машины синхронизации данных, репозитория данных, построена на основе SOA – Service Oriented Architecture, содержит персистентное приложение управления потоком работ - Workflow Engine, важной частью платформы, гарантирующей надежность использования является Audit Tracker, в состав платформы входит также генератор отчетов – Reports&Reporting Framework и подсистема перманентной валидации VVV – Validation, Estimation&Editing. Одной из главных проблемно-ориентированных компонент платформы является аналитическое ядро Analytics Foundation.

Нельзя не отметить еще одну важную особенность платформы – естественная интеграция с производственными системами SAP. Через веб-сервис MDUS – Meter Data, Unification and Synchronization Energy IP интегрируется с SAP IS-U системой SAP для энергопредприятий.

На рынке аналитических систем для энергетики появилась также Toshiba, которая владеет собственным гигантом в области измерителей Landis+Gyr и известным

поставщиком MDM (Meter data Management) решений компанией Ecologic Analytics, которая также позиционирует свои новые разработки как аналитические системы.

Здесь безусловно имеется достаточно места для совершенствования применений интеллектуальных измерителей. Успешные стартапы создают новые подходы и технологии и быстро развиваются, поскольку большинство североамериканских энергопредприятий идут по пути интегрирования данных измерений в управление сетями и бизнесом. Часть успешных стартапов поглощается крупными игроками на рынке хранения и обработки данных. В качестве примера можно привести компанию Data Raker, которая успешно выйдя на рынок облачных аналитических платформ, была приобретена гигантом Oracle [46] .

Oracle купил компанию DataRaker – поставщика программных решений SaaS (Software-as-a-Service) для энергопредприятий, в первую очередь для обработки данных с интеллектуальных счетчиков. Компания обслуживает по контракту 24 миллиона счетчиков (реально установлено 18 млн) и работает с крупными предприятиями энергетики. Пользователи услуг от DataRaker имеют возможность более глубокой аналитики чем все развернутые системы. Oracle планирует комбинировать аналитику DataRaker со своим продуктом Oracle Utilities в одну интегрированную платформу. Продукты компании охватывают пользовательские информационные системы, управление трудовыми ресурсами, менеджмент ресурсами предприятий и платформы менеджмента интеллектуальных измерителей и распределительных сетей. Выход на рынок аналитикиинтеллектуальных сетей встречает конкурентов таких как IBM, SAS, Teradata, EMC, SAP. Все они используют потоки данных от интеллектуальных измерителей, сенсоров на линиях, систем управления подстанциями. Oracle развертывает комплексную систему Advanced Analytics Cloud Service. Это позволяет обнаруживать хищения, предотвращать отказы, обеспечивать мобильных сотрудников необходимой информацией. Кроме того, это позволяет нацелиться на управление балансом нагрузки, потерь в линиях, менеджментом напряжения и предиктивным моделированием.в интересах операторов сети. Внимание уделяется интеграции новых средств с старыми данными предприятий.

С этой целью используется специфическая для энергетики форма Common Information Model (CIM) и преобразуется в стандартные форматы данных. Общее представление о структуре этого продукта можно получить, посмотрев состав входящих в него приложений и их направленность на рисунке 3.5.

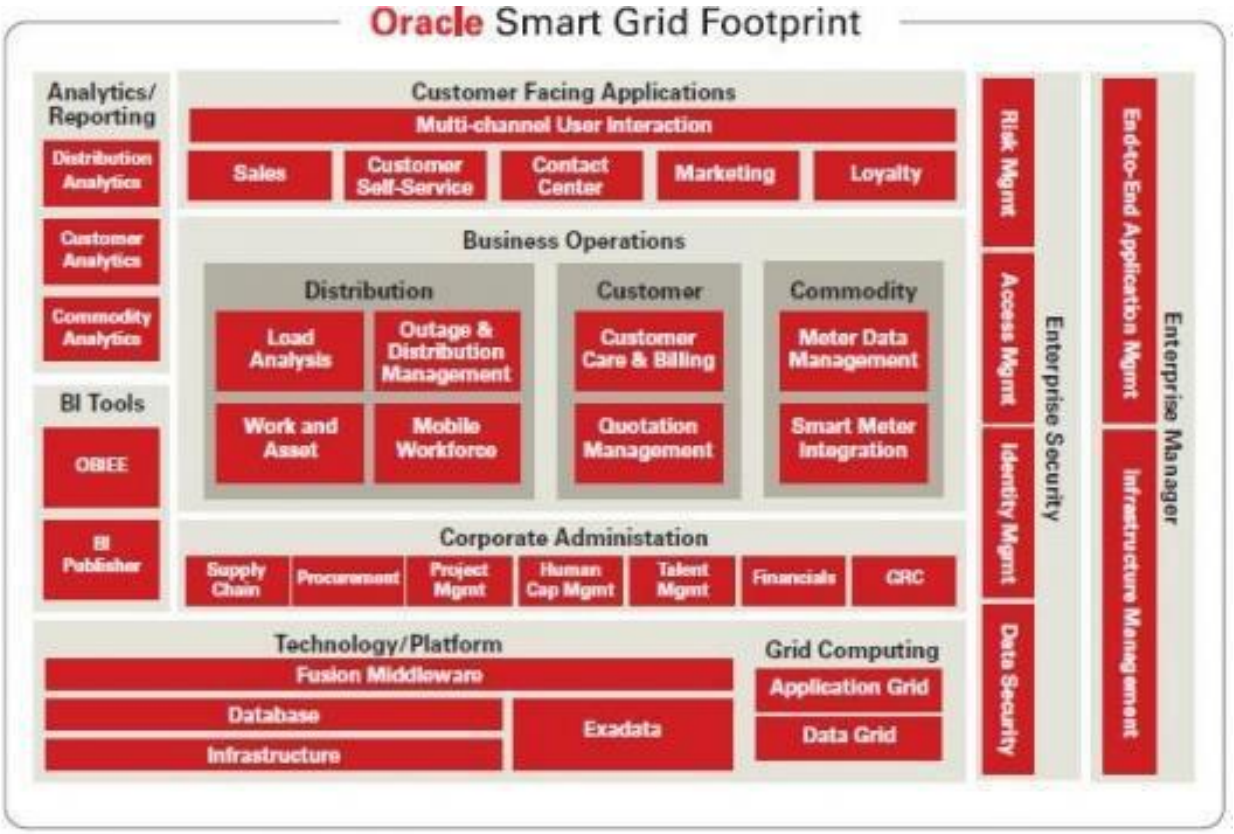


Рисунок 3.5 Общая структура программной системы Oracle Utilities.

Oracle ведет большие работы над технологиями in-memory analytics (аналитика в памяти) для получения преимуществ перед конкурентами за счет более быстрых процессов обработки. Уже сегодня в работу включены последние технологии, среди которых Exadata database appliance, Exalytics in-memory machine, что делает реальным главный план компании - построить эффективный мост между традиционными департаментами энергетиков, такими как отдел расчетов с потребителями, управления распределения рабочих ресурсов, менеджмента имущества и т.п. Планируется каждые шесть - девять месяцев вводить в строй новую дэшборд для использования персоналом предприятия.

Edison Foundation провел исследования, которые показали, что к октябрю 2013года в США было развернуто 46 миллионов интеллектуальных приборов учета , которые создают более миллиарда точек данных в день. Это весьма значительный информационный источник имеющий большую потенциальную ценность. На конференции Soft Grid 2013 [43] в ходе секции Last-Mile Analytics: Insights Into AMI, DG, EV and Beyond, аналитик из GTM Research заявил, что первоначально AMI сети были предназначены исключительно для удаленного интервального съема данных со счетчиков с целью биллинга потребителей, однако очевидно, что эти данные могут быть извлечены для других целей и в том числе для distribution automation (DA) work.

На рисунке 3.6 показано развитие архитектуры инфраструктуры сбора данных с распределенной сети интеллектуальных измерителей.

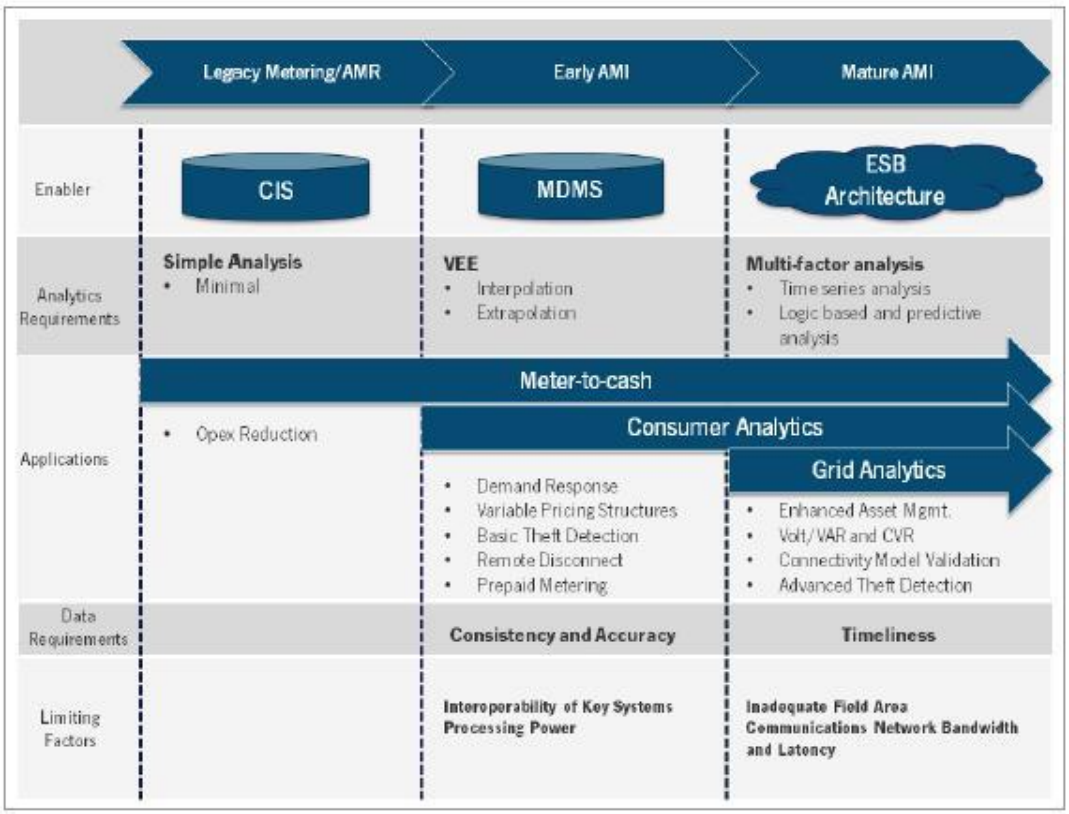


Рисунок 3.6 Эволюция архитектуры AMI (Источник: GTM Research).

Многие компании строят специальные сетевые решения, чтобы в полной мере решить такие задачи. Энергетики смотрят как пользовательская информация может быть проанализирована и превращена в полезную. Этот процесс будет большой волной и возможно не одной, идущей вслед за волной инноваций. Это начало процесса, который будет продолжаться долго и за пределами нашеговремени жизни.

Еще один активный игрок на рынке аналитический систем в энергетике компания 3TIER [49] помогает организациям просчитывать риски для проектов, связанных с возобновляемыми источниками энергии. Решения компании позволяют предсказывать нестабильность погодных условий и прогнозировать на этой основе риски в ветровой и солнечной энергетике. В основе лежит глубокий анализ актуальных метеонаблюдений и исторических закономерностей. Сервис имеет две составляющих Оценку и Прогнозирование. Он доступен через Интернет, и требует быстрого и бесперебойного доступа к большим данным. Обрабатывается большое количество погодных и климатических данных и создается огромный массив моделирования и здесь удалось добиться успеха благодаря применению для хранения данных горизонтально масштабируемой NAS системы EMC Isilon.

Компания eMeter запустила бизнес-модель, основанную на облачном сервисе. Почти сразу же, Siemens анонсировал возможность использования в облачном режиме своего EnergyIP meter data management suite . Оплата теперь может осуществлять на основе помесячных расчетов на один измеритель. Вслед за этим eMeter сообщила о запуске аналитических инструментов в облачном сервисе. Теперь собранные данные могут быть проанализированы для оценки состояния AMI, менежмента загрузкой оборудования, анализа отказов и событий, защиты доходов и мониторинга нагрузки. Облачный сервис позволил объединить данные не только от интеллектуальных измерителей, но и от AMR и даже старых электромеханических счетчиков. Все это превращает компанию в одну из самых амбициозных на рынке использования интеллектуальных измерителей.

Виден серьезный тренд к переносу всей инфраструктуры Smart Grid в облако. IBM ведет работы с E.ON по развертыванию cloud-based smart metering ITinfrastructureчтобы обслуживать Германию, Австрию и Швейцарию.

Подобно тому как это делают алгоритмы продвинутых поисковых машин или прогнозаторов погоды, система уже рассмотренной выше компании AutoGrid [40] Energy Data Platform добывает структурированные и неструктурированные данные, генерируемые энергосетью и устройствами, соединенными с ней чтобы раскопать и извлечь паттерны использования, устанавливает корреляцию между ценами и использованием или анализирует взаимосвязи десятков тысяч переменных. С помощью этой платформы энергопредприятия могут прогнозировать как много мощности будет требоваться на следующей неделе или даже минуте или секунде. Крупные промышленные потребители могут оптимизировать свою работу, чтобы избежать штрафов за перегрузки.

Itron – компания производитель счетчиков электроэнергии [11] . В последнее время главной целью было обеспечить их работу в составе AMI – продвинутой измерительной инфраструктуры. Это уже стало зрелой отраслью в Северной Америке и все энергопредприятия развернули такую инфраструктуру. Однако использование всех возможностей такой инфраструктуры далеко от полноты. Вотчетах GTM Research говорится об использовании данных в AMI для менеджмента отказов и оптимизации напряжения в двух лидирующих вторичных приложениях в Северной Америке. Один случай, который породил большой интерес называется CVR – conservation voltage reduction. Это метод стратегического понижения напряжения на выбранных распределительных фидерах, чтобы усилить эффективность, улучшить фактор мощности, и сократить вредное воздействие на оборудовании как распределительные трансформаторы. Исторически CVR использовался на основе модельного подхода для компенсации потерь на линиях. Но сейчас более трети энергопредприятий используют возможность измерения напряжения на окончаниях линий с помощью интеллектуальных измерителей и применять CVR. Компания Itron для разработок в этой области выбрала опытного партнера – компанию Beckwith

Electric – известную своими продуктами как в области CVR так и еще одной технологии , называемой Volt/VAr Optimization [50].

Было неожиданным обнаружить в числе многочисленных продуктов стартапов и гигантов ИТ рынка разработку российского происхождения. Облачная платформа DMMessenger, разработана входящей в Группу IBS компанией Luxoft [51]. Она помогает снижать потребление электроэнергии в условиях пиковых нагрузок, анализируя терабайты данных, полученных с электросчетчиков.

Платформа DMMessenger представляет собой инструмент анализа данных о потреблении электроэнергии, обрабатывающий практически любые объемы данных, полученные со счетчиков потребителей. DMMessenger предназначен для разработчиков решений Smart Grid, которые хотят добавить в свои продукты средства коммуникации с потребителями, обработки данных счетчиков, распределенное хранилище данных, а также интеграцию с социальными сетями.

Архитектура DMMessenger решает проблему анализа «больших данных», используя распределенное хранилище, которое масштабируется по мере необходимости, а также параллельные вычисления в облаке. Распределенный аналитический инструмент используется для обработки миллионов записей в реальном времени и в параллельном режиме.

Использование DMMessenger помогает энергосбытовым компаниям классифицировать потребителей по их профилям нагрузки, местоположению и другим параметрам, а затем выбрать тех, кто сможет обеспечить наибольшую пользу, например, ограничивая потребление энергии во время пиковых нагрузок. DMMessenger тесно интегрирован с Customer Information System (CIS) и использует открытые программные интерфейсы (APIs) для отправки запросов в базу данных этой системы. В облаке не

хранится никакой конфиденциальной информации, что позволяет защитить персональные данные клиента.

После сегментирования потребителей и выделения целевой аудитории DMMessenger предоставляет энергосбытовым компаниям инструмент с широкими возможностями для взаимодействия и двусторонней связи.

Запрос на снижение потребления отправляется клиентам в виде электронных сообщений, IVR-звонков, Facebook-оповещений или с помощью приложения для смартфона. Независимо от способа контакта с потребителем участие в программах энергосбережения поощряется социальными элементами – играми, участием в конкурсах и соревнованиях.

Пилотные проекты в системах обработки больших данных в электроэнергетике нередко затрагивают весьма глубокие основы построения информационных систем. Приведем здесь пример аналитической системы с одноуровневой распределенной архитектурой, основанной на отсутствии центров обработки данных.

Распределенная по территории, самообслуживаемая интегрированная машина аналитики данных может быть построена на основе объединения интеллектуальных измерителей, солнечных панелей, аккумуляторных батарей и сетевых устройств. Это удалось сделать в ходе проекта в г.Шарлотте Северная Каролина, на энергопредприятии Duke Energy [52] . Подключив к интеллектуальной сети 3000 штук 35 долларовых микрокомпьютеров Raspberry Pi. удалось реализовать полностью распределенные вычисления для выполнения аналитики и обслуживания сети. Говоря о возможностях аналитики на такой сети, для сравнения можно привести тот факт, что такую же вычислительную мощность имел суперкомпьютер Deep Blue, который обыграл Гарри Каспарова на шахматном поединке в 1997 году. Развитие таких архитектур, которые можно называть гетерархическими, продолжается. Планируется развертывание 150 000 компьютеров на сети в Цинциннати, и такая производительность будет представлять собой уже половину от суперкомпьютера IBM’s Watson. Основу программной реализации системы аналитики составляет распределенное хранилище и система поддержки распределенных вычислений. В целом для организации распределенной сети и поддержки систем аналитики в энергосетях разработана платформа Distributed Grid Computing Platform From the Node Up. Эта система представляет собой развитую сеть с многими физическими каналами связи между узлами. Она объединила через сеть 3G и LTE новые интеллектуальные измерители вместе с 14 000 интеллектуальными счетчиками, передающими данные через PLC модемы. В сети также работают различные внутридомовые системы мониторинга нагрузки, управления приборами и сенсорные системы. Для обмена сообщениями используется легкий протокол MQTT и для более робастного трафика протокол AMQP. В то же время транслируются промышленные протоколы для линейных сенсоров DNP-3 для интеллектуальных измерителей Modbus,для мониторинга трансформаторов – SNMP.

Для развертывания проекта был создан консорциум производителей для сети включающий таких как Ambient, известного поставщика интеллектуальных счетчиков Echelon, DMS провайдер Alstom, IT системный интегратор Accenture, поставщик решений по сетевой интеграции накопителей и распределенной автоматизации S&C Electric и мобильный оператор Verizon. В итоге была разработана архитектурараспределенной шины обмена сообщениями ODMBA (open-source, distributedmessagebusarchitecture). Не менее важной задачей было построить на такой распределенной сети производительные методы аналитики.Данные от телекоммуникационной сети, данные о погоде, данные от пользователей, данные от распределенных энергоресурсов нормализуются, валидируются и хранятся хранилище на базе Hadoop.

Эксперты говорят, что компьютерная мощность подобных сетей позволяет уже сегодня не только служить целям повышения надежности сетей, но и превращать пользовательские данные в полезную для них информацию. Каждый проект в этом

направлении оказывается не только реализацией алгоритмов энергосбережения, но и проектом пользовательской аналитики, в котором пользователи получают возможность анализировать свои данные по потреблению, чтобы найти потери и какие их приборы работают неэффективно. Можно сказать, что целью реализуемых в проекте систем является превратить энергопредприятия в “доверенных советников” (trusted energy advisors) для своих потребителей. Компания Landis+Gyr Americas [53], которая была приобретена Тошибой за 2.3 млрд долларов в 2011 г. берет курс на существенные инвестиции в аналитику данных. Все данные с 3 млн измерителей в штате Техас и др. уже поступают для анализа вызовов на обслуживание, предотвращения перегрузок и технического состояния измерителей. Эти данные также используются в местном муниципальном энергопредприятии для работы с отдельными потребителями по оборудованию их хозяйств технологиями DR (demand response) или установки им солнечных батарей. Этот подход представляет собой ви новой услуги: предоставление данных третьей стороне для подключения продавцов соответствующих товаров и услуг.

Интересный опыт по побочному, но весьма важному для практических энергетиков приложению сети сбора и обработки данных от интеллектуальных измерителей есть у GE Digital Energy, которая сравнивает показания счетчиков и состояние погоды с тем, чтобы обнаружить установленные снаружи и не дающие правильных показаний из-за нарушения температурных условий. В целом по мере приобретения доверия к результатам анализа данных, энергетики начинают включать такие системы в управление производственными задачами с целью автоматизации. Как показывает опыт многих проектов, когда энергетики начинают сотрудничать с партнерами по аналитике больших данных требуется время для оценки наиболее значимых и эффективных по затратам применений аналитики в текущих и перспективных бизнес-процессах. Происходит неоднократный возврат к уже решенным, казалось бы задачам. В этомитеративном процессе удается найти правильную “смесь технологий” для получения ценного конечного результата. Причем технологии приходится рассматривать для решения всего круга задач, связанных не только с собственно аналитикой и энергетикой, но и физическими каналами связи и вычислительными платформами. Сюда включаются и различные беспроводные технологии коммуникаций, оптические сети, солнечные батареи и накопители энергии и конечно интеллектуальные измерители. Но самой трудной задачей оказалось найти кто и как будет платить за все это. Уверенность в успешном решении и этой задачи придает тот факт, что уже выполненные проекты показывают, что подход глубокой интеграции энергетических и компьютерных технологий является экономически выгодным. Так, что продолжение работ в настоящее время, должно быть расширено и направлено на поиск эффективных бизнес-моделей.

Системы управления энергопотреблением

Использование технологий управления потреблением Demand Response (DR) [54] (в российской литературе используют также термин «ценозависимое потребление») состоит в организации процессов физического уменьшения потребления мощности в интервалы времени повышенной цены на электроэнергию. Процессы инициируются оператором сети, который используя специальный протокол, отправляет своим потребителям сообщения о необходимости уменьшить потребление или перейти на дополнительные генераторы через указанный интервал времени. В ответ на такое сообщение оборудование потребителя, допускающее кратковременные отключения без существенной потери функций, производят отключение. Такими нагрузками являются, например, водонагреватели, многочисленные в США кондиционеры офисных помещений и домашних хозяйств. Потребитель также отправляет оператору сети сообщения о процессе снижения нагрузки. На рисунках 3.7 и 3.8 показано как имплементируются процессы DR в общую структуру интеллектуальной сети. Сигналы запросов и ответа используют настоящее время две версии протокола взаимодействия, более ранняя из которых применяется для работы с генераторами и

простыми потребителями, а более поздняя – для работы с потребителями, позволяющими более тонкое реагирование.

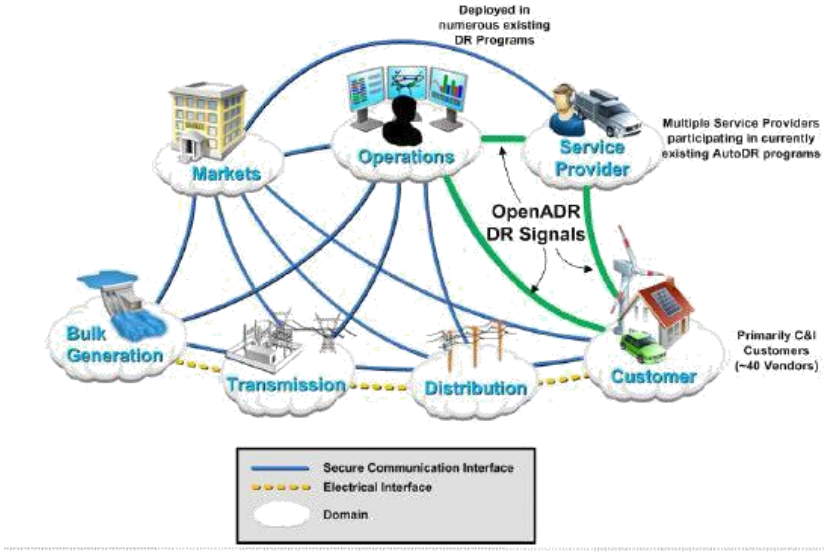


Рисунок 3.7 Место процессов Demand Response в архитектуре Smart Grid

Для обмена сообщениями в ходе процессов DR в настоящее время используется открытый протокол ADR 2.0 [http://www.openadr.org],

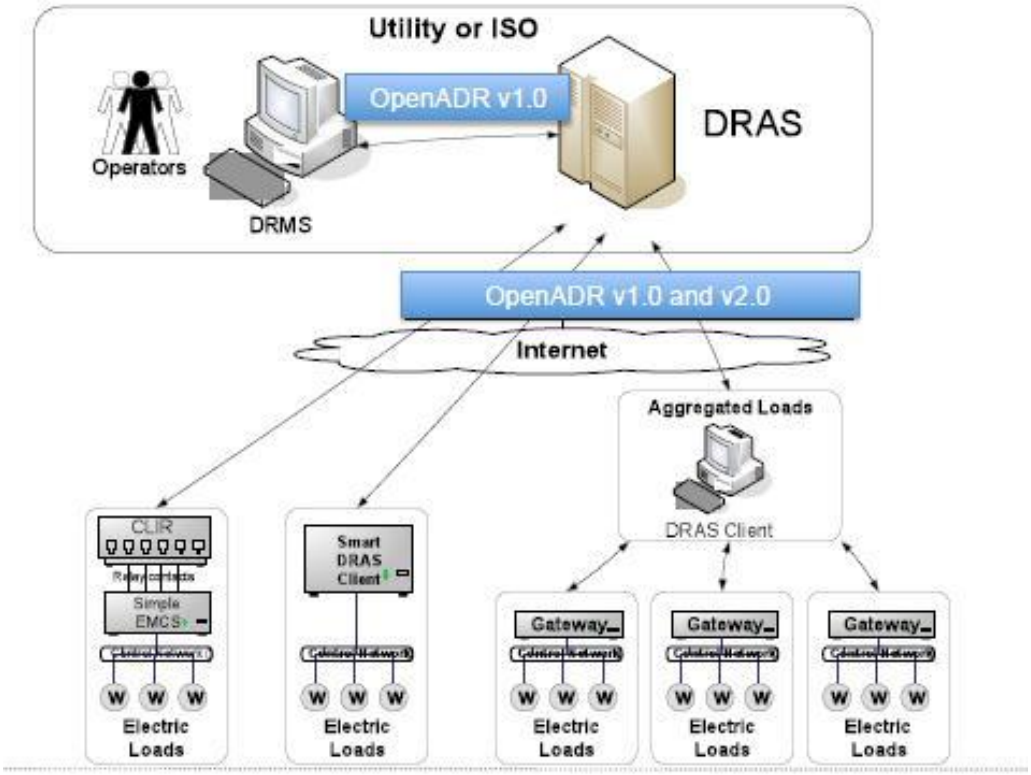


Рисунок 3.8 Сравнение места версий протокола OpenADR v1.0 и OpenADR v2.0

Приведем здесь краткое описание этого протокола, для того, чтобы процессы DR были более понятны. В спецификации OpenAutomatedDemandResponseCommunicationsSpecification (Version 1.0), Национальная лаборатория в Беркли(LawrenceBerkeleyNationalLaboratory) описала OpenADR [55]

как«коммуникационную модель данных, сконструированную чтобы управлять посылкой и приемом DR сигналов от электростанции или независимого системного оператора потребителям электроэнергии. Назначением такой модели данных является взаимодействовать с управляющими системами офисных зданий и объектов промышленности, которые запрограммированы чтобы реагировать на DR сигнал, полностью автоматически, без какого-либо ручного вмешательства. Спецификация OpenADR является очень гибкой, весьма общей и позволяет обеспечивать обмен информацией между весьма разными электростанциями или сетевыми операторами - Independent System Operator (ISO) и его конечными пользователями.

Концепция открытых систем позволяет любому имплементировать сигнальную систему обеспечивая автоматизацию серверов и клиентов. Эта спецификация задает также стандарт, определяющий интерфейс к функциям и свойствам сервера автоматических DR запросов - Demand Response Automation Server (DRAS) , который используется для обслуживания различных Demand Response programs и динамического ценообразования для клиентов. Во время действий события Demand Response event, электростанция или оператор ISO/RTO передает в DRAS информацию, что изменилось и что надо спланировать, в частности, задается время начала и окончания. Типовое сообщение к DRAS содержит строки такого типа:

PRICE_ABSOLUTE – The price per kilowatt-hour
PRICE_RELATIVE – A change in the price per kilowatt-hour
PRICE_MULTIPLE – A multiple of a basic rate per kilowatt-hour

LOAD_AMOUNT – A fixed amount of load to shed or shift

LOAD_PERCENTAGE – The percentage of load to shed or shift

Заметьте, что в первых трех случаях самому клиенту предоставляется возможность наилучшим образом отреагировать на изменения. Например, коммерческие потребители могут быть извещены об изменении в цене в интервале пикового периода и самостоятельнопринять решение об сокращении нагрузки или принятия оплаты по повышенному тарифу.

Система Energy Management System (EMS) офисного или индустриального здания может быть запрограммирована так, чтобы временно изменить температуру в здании на несколько градусов или притушить, или совсем выключить некоторые несущественные осветительные приборы. В последних двух случаях нормальный сброс нагрузки в автоматическом режиме основан на существующих соглашениях. Если цены продолжают расти еще выше, то EMS может эскалировать DR программу до уровня уменьшения или отключения кондиционеров в течение такого пикового периода. Используемые сейчас стандарты позволяют обмениваться и записывать всю необходимую информацию о DR событиях, включая присвоение имен и идентификаторов событий, текущего состояния, режима исполнения, различных видов нумерации, сигналов тревоги и состояния перехода к источнику возобновляемой энергии и наоборот, данные участника рынка, тестовые сигналы еще масса всевозможных полезных данных.вЕвропе нет такой всеобъемлющей сети кондиционеров, создающих основную нагрузку в летнее время и имеются в наличии водяные резервуары для накопления энергии. Однако, здесь нестабильность порождается большим числом ветрогенераторов и солнечных батарей.

В некоторых городах доля такой энергии достигает 50%, а в среднем равна 10% в отличие от 2% в США.

Так появился подход агрегирования нагрузок многих пользователей, чтобы вырабатывать отклик на команды сети в секунды или доли минуты. В качестве примера можно рассмотреть бельгийскую компанию Restore [56] которая построил платформу для

агрегирования нагрузок и реагирования на требования оператора сети. Компания заключила контракт с оператором энергосистемы - компанией Elia, являющейся transmission system operator (TSO) - оператором передачи энергии (сетевой компанией) в Бельгии, для обеспечения отклика со стороны сети, который можно считать соответствующим так называемым требованиям Primary Reserve requirements – подобному по эффекту используемому в США сервисам сверх быстрого регулирования (super-fast frequency regulation services), однако не требующего специальных накопителей. Технология состоит в том, что анализируются, прогнозируются и выравниваются на ежеминутной и часовой основе флуктуации потребления мощности от отдельных приборов путем отключения нагрузок на короткие интервалы времени. Это позволило уже продолжительный период времени избежать блэкаутов для сетей с такими мощными потребителями как холодильные камеры крупнейшего логистического узла в Бенелюксе изарядной станции на 175 электроавтомобилей. Используя необходимые данные в системе строится хорошее предсказание, чтобы помочь промышленным потребителям уменьшить их потребление в самые лучшие моменты, чтобы избежать повышенных штрафов за перерасход.

В Калифорнии на крышах домов становится все больше солнечных батарей и защитники окружающей среды радуются вместе с домовладельцами появившейся энергетической независимости. Однако для энергокомпаний и операторов сетей это новые проблемы. И это потому, что солнечные системы генерации практически полностью лишены инструментов для того чтобы говорить энергетикам сколько энергии они могут вернуть в сеть, и когда они это сделают. И если бы их было немного, но в Калифорнии более 170 000 распределенных солнечных систем, подключенных сегодня к сети, и не учитывать их влияние на стабильность сети невозможно. И солнечные батареи только один из вызовов, с которыми сегодня встречаются лицом к лицу энергетики. И здесь могут помочь инновации в интеллектуальных измерителях, интеллектуальные инверторы, системы мониторинга. Компания Clean Power Research разработала систему менеджмента солнечных генераторов с помощью виртуальных измерений. Используя спутниковые данные о состоянии погоды удается выполнять точные предсказания поминутное состояние генерации на каждой из солнечных панелей.

Программное обеспечение для этого разработала компания SolarAnywhere FlletView, которая сотрудничает с сетевым оператором штата и получила грант от департамента энергетики. Построение модели использовало «тонны» данных от индивидуальных солнечных систем, таких как объем генерации и характеристики, отображение их на геокарте, информацию о взаимных соединениях в сети, к которой подключены источники. Эти данные сравнивались с результатами моделирования и данными от спутников в реальном времени. Сегодня введена вэксплуатацию рабочая версия программы для моделирования выходов всех 170 000 солнечных станций каждые полчаса и весьма точного предсказания на интервал до 30 минут вперед.

Используя следующий грант компания планирует произвести интеграцию этих данных cAutomatic Load Forecasting System (ALFS) которая прогнозирует на часовой интервал и день вперед. Такая интеграция позволит вырабатывать требования к балансу нагрузки на каждый час автоматически и с большой точностью.

В конечном итоге будет построена система автоматического реконфигурирования сети путем отключения отдельных фидеров. И чем ниже гранулярность модели сети, тем больше может получатьсяэффект от такого управления. Калифорнийский университет в Сан-Диего получил грант в 499 900 долларов чтобы продемонстрировать программу, которая может обеспечить лучшее предсказание в реальном времени эффектов, возникающих от различного поведения и подключения солнечных станций, с помощью кластерного анализа.

Разрабатывается также система для прямого использования этих данных для управления линиями распределительной сети. Не все такие линии сегодня управляемы, но в

работе определяется, какие из них будут эффективно влиять на нагрузку и требуют первостепенной модернизации.

В г.Остин штат Техас энергокомпания Austin Energy завершила развертывание системы Demand Response Optimization and Management System (DROMS) разработанной компанией AutoGrid [40] .

Размер пилотной фазы невелик и включает в себя 60 термостатов и 15 зарядных систем для электромобилей. Интеграция DR системы с зарядниками для электромобилей является новой задачей и первые результаты дают много материала для понимания и проектирования будущих алгоритмов.

В системах DR сигналы вкл/выкл должны за короткое время поступать на все потребляющие приборы на основе их классификации по типичным профилям нагрузки, таким как частота максимальной нагрузки, т.е. сколько раз в день они должны потреблять больше всего энергии.

Значение нагрузки, которое может варьироваться очень быстро, также имеет важное значение для осуществления интеграции различных приборов. Существенную роль играет оснащение всех потребителей-приборов сенсорами и телеметрией для создания обратной связи. Так обеспечивается агрегирование нагрузки с большой гибкостью и быстротой, чтобы следовать командам от центральной управляющей платформы. При большом числе управляемых нагрузок задача требует оперирования с большими объемами данных с малой латентностью, что требует новых алгоритмов и программных решений. Но есть задачи, требующие еще большей скорости обработки информации.

Компания Space Time Insight, известная своими разработками в области ситуационной аналитики, сообщила, что работает над проблемами управления сверхбыстрыми данными энергосетей, такими как измерение с блоков фазоров PMU (Phasor measurement unit), которые собираются 60 раз в секунду. Результаты представлены в проекте с Electric Power Research institute .

Тема № 14. Big Data Managmentпревращаются в энергию: виртуальные электростанции

Выработка энергии источниками возобновляемой энергии зависит от погоды и часто достигает максимума, когда ее не требуется.

Сегодняшние газовые генераторы, которые могут гибко реагировать на потребность увеличения выработки путем добавления горючего, являются дорогим и неэкологичным способом решения этой проблемы. Однако, практически всегда достаточно специальным образом изменить график потребления и перераспределить нагрузку по питающим фидерам, как потребность в дополнительных генераторах может исчезнуть.

Поэтому появился подход, названный “виртуальной электростанцией” (virtual power plant - VPP) [59] .

Это технология, относящаяся к smart grid, и заключающаяся в том, чтобы координированно управлять потреблением электроэнергии в большом количестве домов и офисов, частично отключая и включая потребителей в соответствии с требованиями на энергопотребление, так, чтобы синхронизировать потребление с подъемами или спадами многих электростанций, включая все возобновляемые источники. Но для этого необходимо иметь доступ к тысячам устройств для достаточно быстрого отклика с достаточной координацией, и это представляет собой серьезную задачу.

Последние четыре года для решения такой задачи реализовывался проект тройкой компаний из Канады: Maritime Provinces of New Brunswick, Nova Scotia and Prince Edward Island.

Проект называется PowerShift Atlantic, и реализовывает совместное управление комбинации на 11.5 мегаватт соединенных нагрузок, от промышленных кондиционеров и водонасосных станций до тысяч дистанционно управляемых водонагревателей в домах жителей.

В отличие от классических систем Demand Response эти нагрузки не просто способны отключаться на время, чтобы уменьшить волнения в сети, они способны поглощать избыточную мощность – критическое свойство чтобы избежать необходимости снижать генерацию от прибрежных и континентальных ветрогенераторов в 675 мегаватт. Это соответствует целям достичь 40% возобновляемой энергии в 2020 году.

Итак, вместо запроса многим потребителям отключиться в манере все-или-ничего, была построена структура с оборудованием индивидуальных пользователей агрегированная в пять индивидуальных Demand Management Service Providers чтобы мониторить и управлять ими на региональной основе.

Они в свою очередь соединяются с централизованной “виртуальной электростанцией” (агрегатором агрегаторов) - системой, которая работает как интерпертатор и координатор сетевого оператора по отношению каждой из пяти групп, выдавая поминутные инструкции для обеспечения долговременной оптимизации.

В основе лежит идея распределенной иерархически управляемой системы сетевого балансирования.

На рисунке 3.9 показана агрегация потребителей и генераторов в координированные кластеры, что собственно и образует четыре виртуальных электростанции.

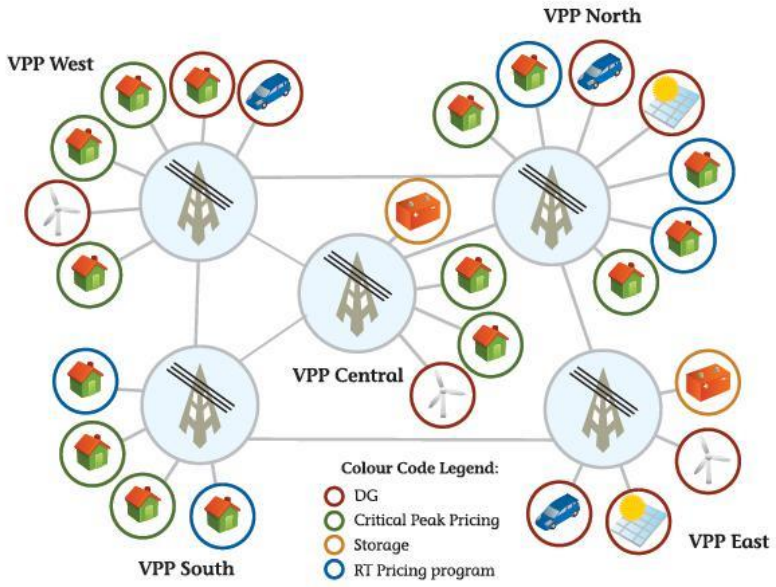


Рисунок 3.9 Логическая структура энергосистемы с четырьмя координированными виртуальными электростанциями

Обозначения DG – распределенный источник, Critical Peak Pricing – ценообразование

В пиковый период, Storage - накопители, RT Pricing Program - участники программы оплаты в реальном времени.

Основные блоки системы – это индивидуальные нагрузки, которые позволяют сдвигать во времени и обрезать потребляемую мощность. Они включают в себя множество термонагревателей, как водяные, холодильники и системы кондиционирования зданий, которые медленно изменяют температурные установки в течение дня. Они также включают водяные насосы, которые могут выбирать, когда на дне они могут поднимать воду и могут это делать в подходящий интервал времени.

Офисные здания могут осуществлять работу в режиме Demand Response, не нарушая комфорт и производительность своих пользователей. Наконец, поставщики электроэнергии могут использовать этот инструмент для подключения возобновляемыхресурсов таких как ветроагрегаты и солнечные батареи к сети , так чтобы минимизировать прерывистую природу этих ресурсов.

В проекте, который описан выше, участвовали производители промышленных систем нагревания, систем температурного поддержания и стартап занимающийся менеджментом нагрузки. Один из стартапов – Intgral Analytics имеет программный продукт для предсказания и управления распределенными в доме или офисе нагрузками, так чтобы соответствовать потребностям сети, другой стартап Enbala Networks, который занимается технологиями быстрого управления водяными насосами и накопителями энергии в системах быстрого реагирования в системах отклика сети. В результате удалось построить систему такого управления потребителями, чтобы снизить вариабельность профиля нагрузки для группы в целом. Возможно, кого-то заинтересует более детальное описание идеи агрегации нагрузки. На рисунке 3.10 показано, как формируется профиль нагрузки при агрегировании.

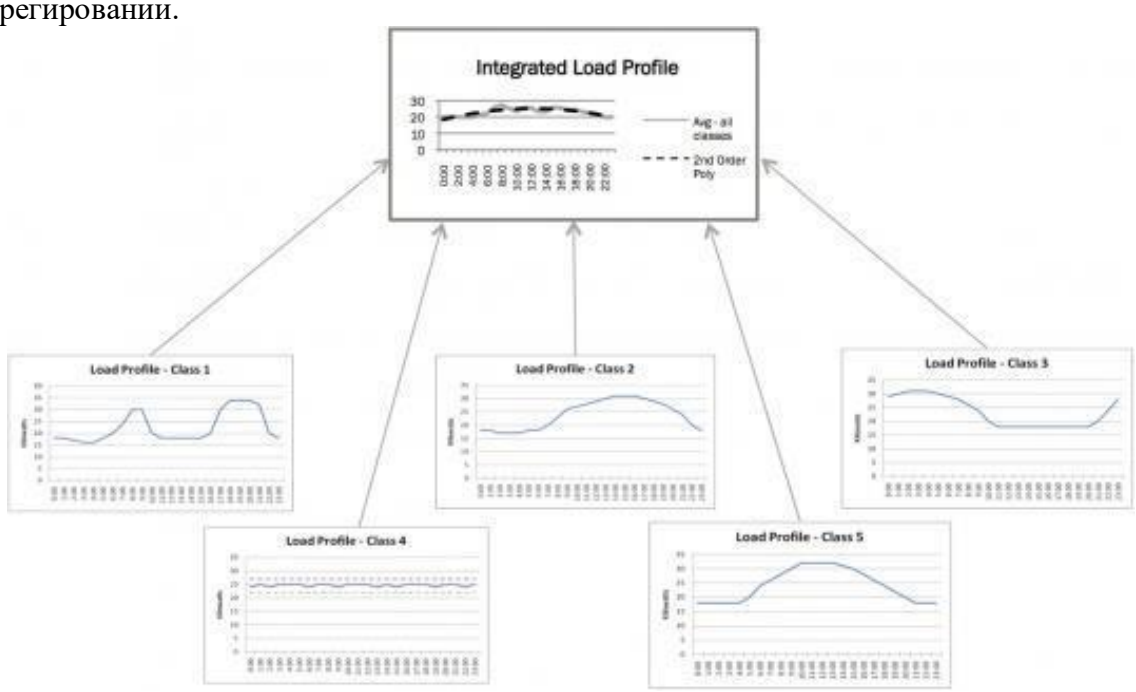


Рисунок 3.10 Профили нагрузки и их интеграция

Виртуальная электростанция решает целый ряд взаимосвязанных задач, диктуемых широким применением источников возобновляемой энергии. Исходя из стоимости выбросов CO2 в атмосферу, стоимость надежности ветрогенераторов и солнечных батарей, стоимость точного прогноза потребления, путем анализа возможного снижения пиков, прерывистости генерации, предсказания генерируемой мощности, горячего резерва, вырабатываются все необходимые команды на уменьшение потребления, участия источников возобновляемой энергии, прогнозирование оплаты и степень работы систем DR. На рисунке 3.11 показано такая функциональная структура.

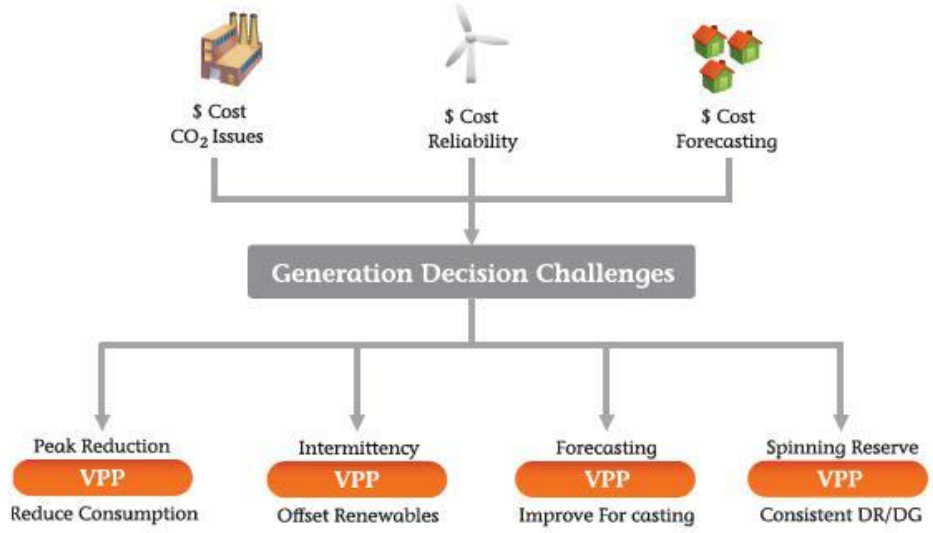


Рисунок 3.11 Взаимосвязанные задачи интеграции распределенных возобновляемых источников в энергосистему

Кластеризация распределенных источников и потребителей в виртуальные подстанции, на основе эффективной агрегации является важной частью развертывания виртуальной электростанции. Различные кластеры могут использовать различные технологии управления нагрузкой и генерацией. На рисунке 3.12 показано как объединенные и спомощью microgrid, так и обычными фидерными подключениями участники энергосистемы могут использовать локальные подсистемы Demand Response и локальные системы управления распределенными генераторами для предотвращения чрезмерных закупок.

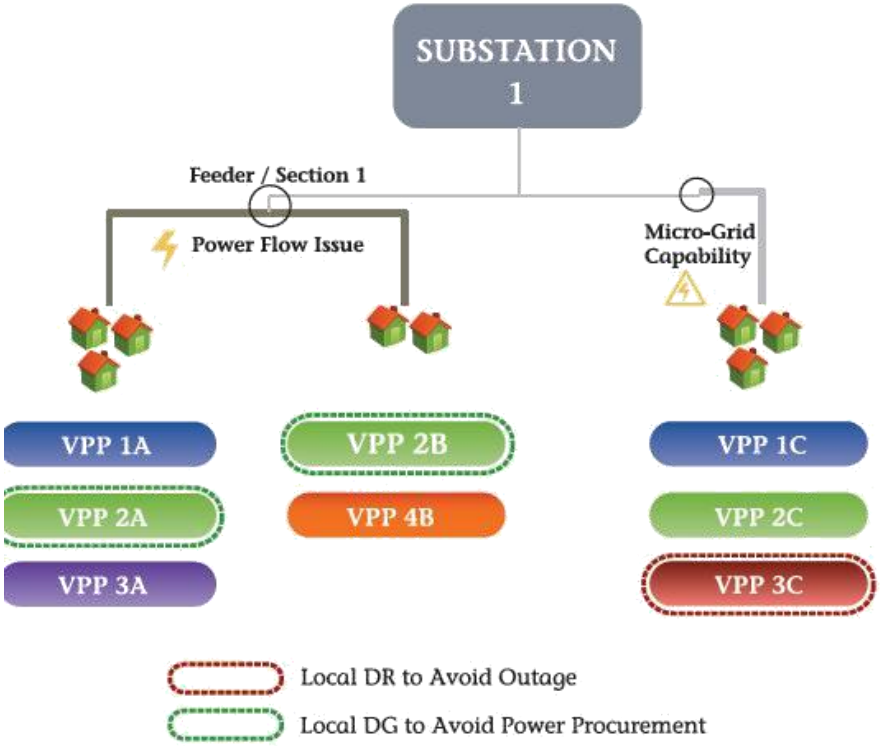


Рисунок 3.12 Создание виртуальных подстанций в кластеризуемых энергосистемах.

Термин «виртуальная электростанция» был поддержан игроками рынка. Упоминаемая ранее компания AutoGrid [40] выпустила продукт, который называется Software Defined Power Plants (™) - программно-реализованная электростанция. И это не просто еще одно новое маркетинговое словосочетание. Эта система позволяет производителям электроэнергии вместо добавочных расходов на горючее использовать более тонкое его использование. Энергоэффективность сети возрастает и при тех же затратах пользователи могут получить дополнительные объемы электроэнергии, эквивалентные установке новых генерирующих мощностей. Используя предиктивную аналитику, поведенческие алгоритмы и основанное на физике процессов понимание сети, система дает предприятиям видеть и управлять распределением мощности и расходом на всей обслуживаемой территории в реальном времени. Совместно с Energy Data Platform (EDP) осуществляется управление всеми доступными ресурсами для обеспечения унифицированного менеджмента и непрерывной оптимизации баланса энергопотоков. На рисунке 3.13 показаны основные взаимодействующие компоненты платформы. Система поддерживает работу Demand Response режима без дополнительных программных компонент. Также этот комплекс позволяет интегрировать в сеть возобновляемые источники энергии и вводить в состав сети накопители, обеспечивая управление всеми этими компонентами с наибольшей экономической эффективностью. В конечном результате сокращаются капитальные затраты, минимизируются выбросы вредных газов и соблюдаются требования регуляторов при удовлетворении потребителей.

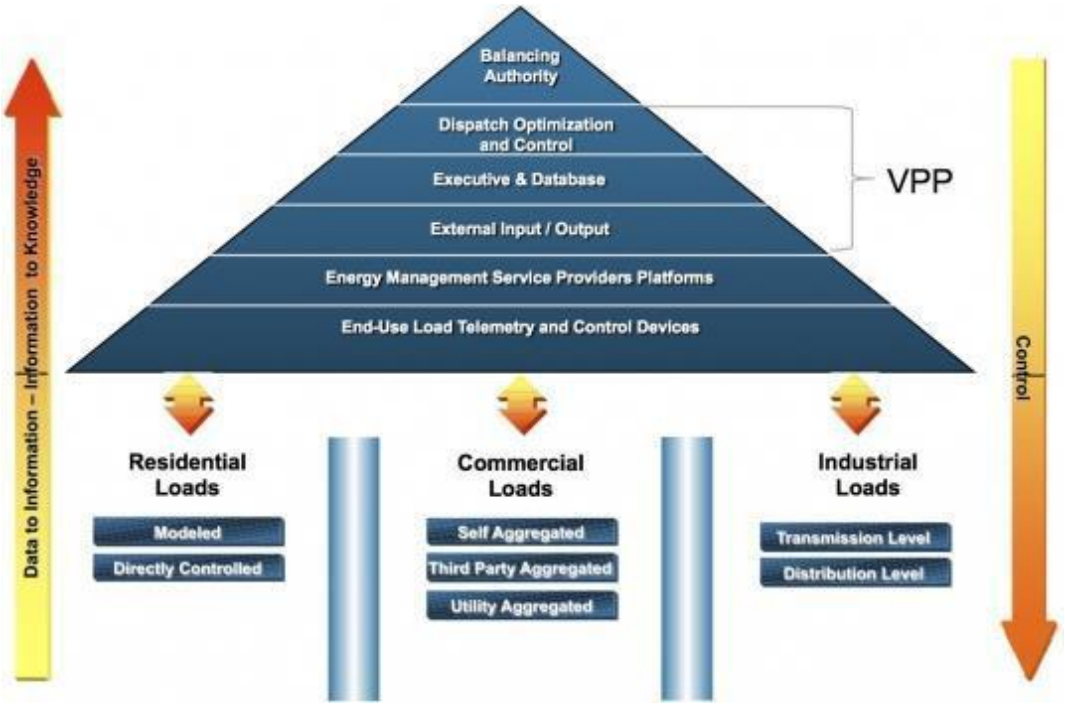


Рисунок 3.13 Замкнутый информационный поток в платформе виртуальной электростанции.

Заметим здесь, что все продукты AutoGrid используют две определяющие компоненты рынка электроэнергии: физика и человеческое поведение. Основой платформы EDP также является жидкостная, основанная на физике, модель сети, учитывающая поведение всех устройств, подключенных к ней- зданий, электроизмерителей, промышленных площадок, зарядных устройств для электроавтомобилей, оборудование квартир, передающих линий и тп. Разработчики отмечают, что детали, точность моделирования и предсказания здесь имеют первостепенное значение. Больницы имеют совершенно иные запросы на чем отели или торговые представительства. Потребление

энергии офисными зданиями колеблется в течение дня в зависимости от использования лифтов и компьютеров. Сегмент распределения центра города совсем отличается от пригородного. В модели могут быть запущены сценарии, основанные на накопленных данных или на ожидаемом поведении. Что если к полудню температура изменится на 12 градусов? Как изменятся запросы от кондиционеров в центре города? Что будет если цена за киловатт-час возрастет на 2 цента, сколько потребителей уменьшат свое потребление? А если возрастет на 2.25 цента? Короче говоря, система позволяет рассматривать сеть не как физическую абстракцию, а как реальную жизнь.

Анализ показывает, что на 3400 мегаваттной системе можно развернуть виртуальную электростанцию до 750 мегаватт за счет экономичного расхода энергии. В результате будет получено 3.5 млн долларов годовой экономии. Так виртуальные электростанции занимают место рядом с микрогридами в трендах развития электроэнергетических систем.

Рассмотрим еще один аспект применения больших данных в трендах развития энергосистем. Их эволюция затронула в настоящее время рост использования новых элементов, таких, как накопители энергии различного уровня.

Современные энергосети включают в свой состав немало элементов, которые еще недавно были редкими и требовали особого, уникального подхода как при проектировании сети, так и при эксплуатации. К таким элементам прежде всего стоит отнести накопители электрической энергии. Существует немало различных видов накопителейэлектрической энергии, которые могут использоваться сегодняшними потребителями. Это и большие аккумуляторные батареи, и системы сжатого воздуха, маховики, воздушные охладители, промышленные моторы, водяные помпы т.п. Но вопрос, как измерить их ценность и окупаемость при использовании на энергосетях, является весьма сложным. И дело не только в том, что это новые технологии и существует немало проблем просто для их подключения к сети, но и проблем экономических и регуляторных. Конечно в первую очередь интересуются тем как будут возвращаться инвестиции в эти весьма недешевые устройства для их владельцев. Но не менее важно понимание их ценности для сети в целом, где наличие накопителей влияет на сглаживание пиков нагрузки, существенно уменьшает нестабильность ветровых и солнечных источников энергии и способствует решению других подобных проблем. Моделирование системного влияния накопителей на работу сети является самостоятельной сложной задачей и ее решение стало предметом исследователей из Electric Power Research Institute (EPRI). Одним из результатов их исследований стал программный инструмент Energy Storage Valuation Tool (ESVT) [60]. С его помощью можно превратить данные о сети с накопителями в измерения ценные для энергопредприятий, сетевых операторов, производителей накопителей, проектировщиков новых сетей.

Рассмотрим пользовательский интерфейс этого инструмента, который позволяет в весьма простой форме задавать исходные параметры сети и получать как технические, так и финансовые результаты использования накопителей. На рисунке 3.14 показан графический пользовательский интерфейс в явном виде задающий четыре шага анализа.

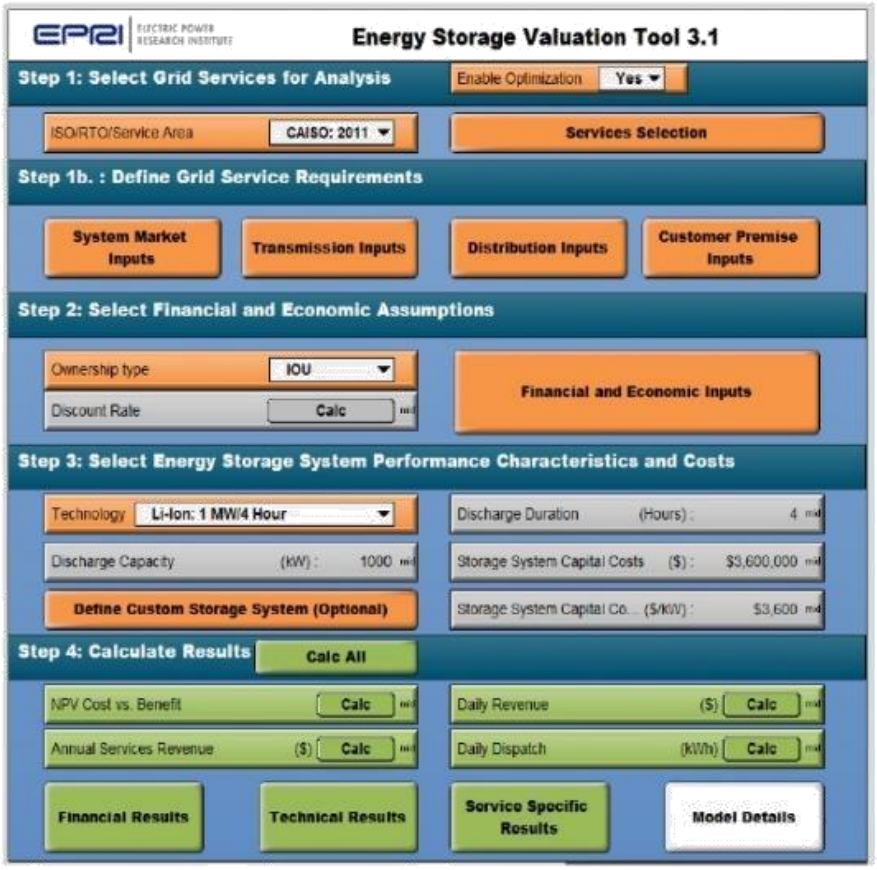


Рисунок 3.14 Пользовательский интерфейс аналитического инструмента ESVT для оценивания эффективности развертывания накопителей в сети

На первом шаге выбирается вид сетевого сервиса, который требуется подвергнуть анализу. Инструмент позволяет анализировать системы передачи, распределения энергии, пользовательские сервисы, рыночное позиционирование. На втором шаге задаются финансовые и экономические параметры. На шаге три инструмент позволяет задать технологии используемых накопителей из широкого спектра: аккумуляторные батареи различного типа, накопители на сжатом воздухе (CAES), насосные гидроаккумуляторы и др. Здесь задаются и количественные характеристики применяемых накопителей. Четвертый шаг запускает процесс анализа и позволяет получить результаты в различной удобной для дальнейшего использования форме.

Проведенные исследования позволили получить множество данных о затратах на киловатт и стоимости киловатт часа для различных технологий накопителей. С помощью этого инструмента можно анализировать как миллисекундный отклик балансировки сети так и долгосрочные колебания нагрузки. Свои исследования ERPI проводил по контракту с Калифорнийским регуляторным органом - California Public Utilities Commission. Интересными результатами использования являются также предложения по появлению новых рыночных структур, отсутствующих в сегодняшней энергетике, но которые могут эффективно существовать при широком использовании накопителей. Поскольку в настоящее время имеется масса непонимания в этой области, моделирование позволяет увидеть как влияет появление технологий накопителей на рынок и технические характеристики, и в этом состояло выполненное исследование. Но это можно считать только первыми двумя шагами в направлении, определенным исследователями. В общем случае обработка данных в ESVT может быть представлена основными этапами, показанными на рисунке 3.15, где ключевым является хорошая модель.

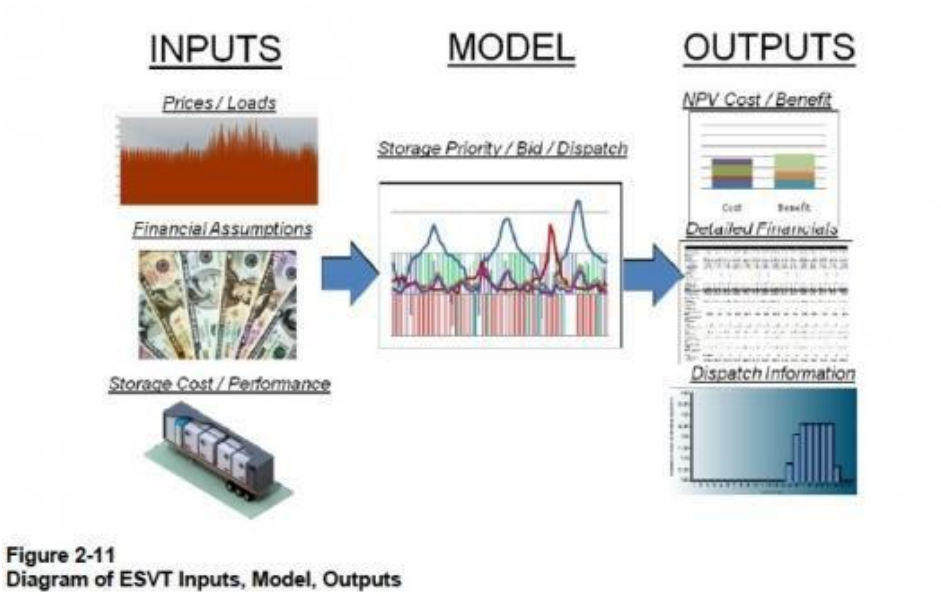


Figure 2-11
Diagram of ESVT Inputs, Model, Outputs

Рисунок 3.15 Базовый поток обработки в ESVT. Ключевым элементом системы является модель, учитывающая множество аспектов.

Следующий шаг развития этого инструмента будет направлен на анализ непрямого влияния развертывания накопителей на сетях и в том числе на окружающую среду. А затем предполагается получить описание бизнес-кейсов, описывающих монетизируемую ценность длявладельцев накопителей в условиях реального технического, экономического и регуляторного окружения.

Накопители как постоянные элементы в сети позволяют не просто сглаживать нестабильность возобновляемых источников энергии, но и существенно увеличивать надежность при распределении энергии потребителям. Для решения таких задач требуется использование весьма точных моделей сетей, например, таких, как разработанных компанией AutoGrid. Основой модели является сочетание двух составляющих рынка электроэнергии – физическое и человеческое поведение. Специально разработанная программная система Energy Data Platform использует жидкостную физическую модель сети, принимая во внимание все компоненты, соединенные друг с другом: зданий, измерителей, промышленных построек, зарядников электромобилей, квартирных комплексов, линий передачи и т.п.

Washington State University вовремяпиковыхпериодов.При использовании пятиминутных выборок значений напряжения через платформу сбора данных OpenWay компания поставщик энергии идентифицировала проблемный фидер и применила развертывание батарейного накопителя для ликвидации проблем напряжения.

В более сложных случаях системы сбора распределенных данных включают в себя геоинформационные системы и соответствующие модели интегрируются с моделями AMI (Advance Metering Infrastructure)соответствующими системами SCADA энергопредприятий для автоматического управления volt/VAR на основе моделирования потоков мощности и имитационного моделирования сети и потребителей.

Особое место в системах такого типа занимают модели дляиспользование данных AMI для сетей с возобновляемыми источниками и микрогридами, содержащими накопители энергии. Бурное внедрение таких сетей в структуру традиционной экономики энергетики привело к возникновению нового взгляда на возможности управления потреблением, отличающегося от того, что типичен для стран, где он был изобретен. В США, где впервые был придуман подход DR, описанный выше, энергопредприятие или оператор сети за день вперед или за час вперед предупреждают о необходимости уменьшить потребление отключая кондиционеры или водонагреватели или включить

запасной генератор. В Европе нет такой всеобъемлющей сети кондиционеров в летнее время и имеются в наличие водяныерезервуары для накопления энергии. Однако здесь нестабильность порождается большим числом ветрогенераторов и солнечных батарей.

В некоторых городах доля такой энергии достигает 50%, а в среднем равна 10% в отличие от 2% в США. И солнечные батареи и ветрогенераторы - только один из главных вызовов, с которыми сегодня встречаются лицом к лицу энергетики. И здесь постоянно идут разработки, предлагающие инновации, объединяющие интеллектуальные измерители, новые интеллектуальные инверторы, системы мониторинга и управления оборудованием. Компания Clean Power Research разработала систему менеджмента солнечных генераторов с помощью виртуальных измерений. Используя спутниковые данные о состоянии погоды удается выполнять точные предсказания поминутное состояние генерации на каждой из солнечных панелей. Программное обеспечение для этого разработала компания SolarAnywhere FlletView, которая сотрудничает с сетевым оператором штата и получила грант от департамента энергетики. Построение модели использовало тонны данных от индивидуальных солнечных систем, таких как объем генерации и характеристики, отображение их на геокарте, информацию о взаимных соединениях в сети, к которой подключены источники. Эти данные сравниваются с результатами моделирования и данными от спутников в реальном времени. Сегодня является рабочей версия программы для моделирования выходов всех 170000 солнечных станций, развернутых в Калифорнии, каждые полчаса и предсказания на каждые последующие получасовые интервалы. Используя следующий грант компания планирует произвести интеграцию этих данных с Automatic Load Forecasting System (ALFS) которая прогнозирует на часовой интервал и день вперед. Такая интеграция позволит вырабатывать требования к балансу нагрузки на каждый час автоматически и с большой точностью.

В конечном итоге будет построена система автоматического реконфигурирования сети путем отключения отдельных фидеров. И чем выше гранулярность сети, тем больше может получаться эффект от такого управления. Калифорнийский университет в Сан-Диего получил грант в 499900 долларов чтобы продемонстрировать программу, которая может обеспечить лучшее предсказание в реальном времени эффектов от различного уровня проникновения солнечных станций с помощью кластерного анализа. Разрабатывается также система для прямого использования этих данных для управления линиями распределительной сети. Не все такие линии сегодня управляемы, но в работе определяется какие из них будут эффективно влиять на нагрузку и требуют первоочередной модернизации.

Рассмотрим еще один проект энергосистемы с накопителями, установленной в Пенсильвании, США на объединительной сети 10 самых крупных операторов передачи электроэнаргии – PJM Interconnection [62]. Компания Ecoult инсталлировала накопительную систему для поддержки сервисов регулирования общей мощностью до 3MW. На рисунке 3.16 показана контейнерная конструкция батарейного накопителя.

Рисунок 3.16 Накопитель 3MW UBer TM и система управления Ecoult на PJM Interconnection

Проиллюстрируем работу системы управления, которая на основе анализа сигнала от сети весьма точно осуществляет предсказание и компенсацию мощности путем точного регулирования циклов частичного разряда и заряда рядов батареи.

Качество предсказания в системе может иллюстрировать рисунок 3.17

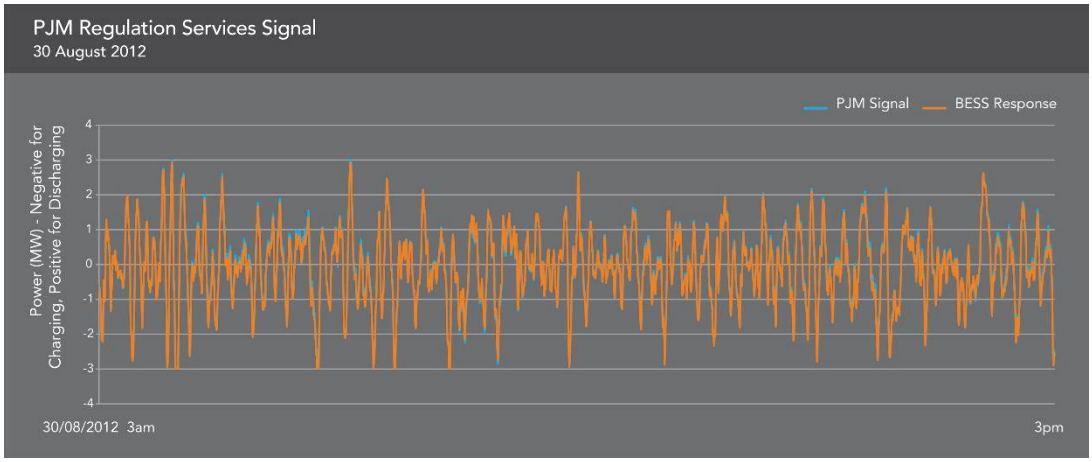


Рисунок 3.17 График сигнала от PJM и отклик системы регулирования BESS (Battery Energy Storage Service)

Процессы, характеризующие состояние заряда (SOC- State of Charge) для отдельных рядов батареи показывают (см. рисунок 3.18) качество балансирования развернутой системы накопителя.

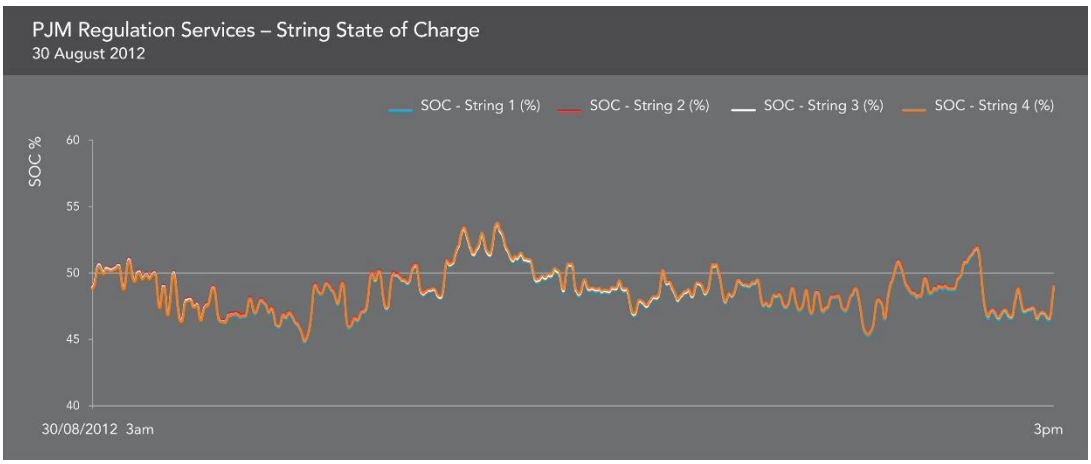


Рисунок 3.18 Процессы заряда батареи накопителя по рядам элементов.

Как видно из изложенного, появление управляемых накопителей энергии на сетях позволяет весьма эффективно компенсировать нестабильности нагрузки и генерации, что оказывается существенным фактором развития энергосистем с распределенными возобновляемыми источниками и сложными топологиями взаимных связей. Однако, использование таких накопителей требует сбора и высокоскоростной обработки большого объема данных.

Тема № 15. Системы мониторинга и управления ресурсами в норме и в аварийной ситуации

В этом разделе мы рассмотрим проекты более традиционной направленности, однако реализуемые на основе концепции и технологий больших данных.

Недавно компания Space-Time Insight – стартап, работающий в области объединения данных реального времени и предиктивного анализа, вступил на рынок североамериканских производителей электроэнергии таких как Southern California Edison и Hydro One, Florida power&Light San Diego Gas& Electric , вместе с калифорнийским сетевым оператором California ISO (Independet System Operator). Стартап поднял 20 млн раунда С и работает с ведущими компаниями в энергетике еще и в Азии и Европе. Основная направленность развития компании - это разработки в направлении аналитики имущества (asset analytics) и сетевой разведки (grid intelligence). Реально работающие сегодня, развернутые компанией фрагменты будущей гигантской системы позволяют не просто смотреть на карту и видеть, где сейчас ваши транспортные средства и самолеты, но и сразу получать информацию, что находится на этих транспортных средствах. Полностью ли они загружены? Опаздывают они или идут по графику. Правильно ли выбрано транспортное средство для доставки? Идет ли по наиболее правильному и эффективному маршруту? Такие решения трудно принять, если обладаешь недостаточной информацией. Компания интегрировала свою разработанную ранее производственную платформу чтобы обслуживать целый ряд подобных задач. Например, геопространственный инструмент, разработанный два года назад для производственного муниципального района Сакраменто, обслуживает сейчас как пожарных, так и территориальный водоканал. Большой рыночный потенциал таких разработок не останавливает исследования и развивает движение в направлении логистических задач. Развитие в сторону индустрии управления ресурсами

открыло новые возможности и вывело на конкуренцию с целым множеством технологических провайдеров в электроэнергетике.

Общие инвестиции в компанию составили 45 млн. И к концу года она попала в топ 10 поставщиков мониторинга для Smart Grid систем. Эксперты GTM Research отметили успешную интеграцию традиционных средств ГИС и средств обработки данных следующего поколения и возможностей аналитики. Компания конкурирует с традиционными поставщиками ГИС как Intergraph и Esri и строит партнерские отношения с General Electric и Google Maps. В своих решениях компания использует интеграцию данных реального времени с системами, построенными на основе платформ, таких как EMC

Greenplum , SAP HANA и отдельно с подобными инструментами больших данных от Oracle, IBM, OSI soft.

Постоянно анализируя потребности рынка и тренды, компания Space-Time Insight от привычных сервисов геоинформационных систем для энергетиков перешла к встраиванию функций аналитики реального времени и предиктивной аналитики. Первые результаты были получены в транспортировке газа и нефти для компаний в прибрежной зоне и компаний глобальных сервисов доставки. Логистические операции требуют значительной аналитики реального времени. Смысл дальнейшей работы как превратить разработанные специализированные системы в повторяемые и сократить время выхода на рынок как самостоятельных продуктов. Задачи, решаемые здесь - это узнать где находятся автомобили или самолеты, что они везут, вовремя или опаздывают, какие маршруты наиболее эффективны, каким образом следует доставлять тот или иной груз. Главной проблемой было эффективно решить задачу интеграция геосервисов с новым поколением сбора и обработки данных. Задачи интеграции информационных систем такого масштаба требуют от участников проекта опыта масштабного системного проектирования.

Сама по себе задача мониторинга, анализа и эффективного управления ресурсами всегда была актуальна для индустрии вообще и, в частности, для энергетики. Однако возможности оперирования с большими данными открыли здесь новые возможности. И на этом рынке постоянно появляются новые игроки.

Компания Schnider Electric [63] выпустила программную платформу Orbit для интеграции существующей ГИС с мобильными системами для ремонтных бригад со спецификой мобильных устройств разного уровня. Orbit построена как облачный сервис с мобильными клиентами и использует в качестве ядра Microsoft Windows Azure cloud platform .

Это делает более простым и эффективным разработку интерфейсов между множеством старых систем и новыми приложениями.

Система показала себя весьма эффективной в ходе тестирования на задаче плановой инспекции энергокомпаний. Были обнаружены многие недостатки в использовании методов разработки известных мобильных приложений при построении индустриальных мобильных приложений. Потребовалось время для поиска более надежных решений и сейчас проблемы устранены. Подробнее вы можете познакомиться с тем как работает Orbit ниже. Система относится к мобильным приложениям, работающим совместно с облачной инфраструктурой. В настоящее время существуют версии для планшетов на Windows 7, 8 и iPad.

На рисунке 3.19 показана типичный интерфейс при работе полевого персонала.

Рисунок 3.19 Интерфейс системы Orbit в процессе выполнения работ по обслуживанию

Система может рассматриваться как специализированная геопространственная система, но она использует корпоративную ГИС и не хранит геоданные в себе, что позволяет считать ее применение безопасным. Вы только задаете виды работ, которые должен выполнять полевой персонал, быстро корректировать и дополнять их список. В процессе работы пользователи создают и подтверждаютсписки имущества, получают доступ к документации и заданиям, отмечают факт исполнения и фиксируют состояние обслуженного объекта. Внутренняя структура системы может быть описана диаграммой, показанной на рисунке 3.20.

Рисунок 3.20 Пять основных модулей системы Orbit, взаимодействующих с корпоративным геопространственным сервисом через ArcFM сервер системы

Применение систем, подобных Orbit, позволяет резко повысить производительность труда полевого персонала, снизить уровень ошибок и сильно увеличить сроки эксплуатации оборудования и имущества.

Весьма важные применения больших данных связаны с системами, от работы которых во многом зависит работа энергосистем в период аварий и отключений. Компания ComEd (Commonwealth Edison Company, [65] являющейся частью базирующейся в Чикаго компании Exelon Corporation), летом 2011 года встретилась с самым худшим сочетание штормов и отказов в истории северного Иллинойса. Отключение коснулось 2.6 миллиона человек что более чем вдвое превышает трехлетнее среднее число отключений. В одном только июле 18 000 отказов освещения, 600 трансформаторов, 700 столбов, 850 000 потребителей остались без энергии. За пять дней более миллиона звонков поступило на колл-центр. Среди жалоб на работу энергетиков в период шторма одной из главных была плохаякоммуникация с энергетиками. Обращение к веб-сайту было бесполезным, указывают клиенты, так ка не содержало нужной информации. Текстовые СМС сообщения по запросу не соответствовали действительности. После этого были сделаны существенные изменения в системе коммуникации с клиентами во время аварий. Была развернута специальная система коммуникаций при авариях - OCS (outage communication system), которая позволяет видеть интегрированную визуализацию информации об аварии, улучшенный извещатель о шторме- визуализатор, более точную информацию о местонахождении персонала, улучшенную систему репортинга персонала. Главная задача систем такого типа всегда ставится так: иметь “одну версию правды” для персонала и внешних пользователей. Окружение OCS должно основываться на интегрированной платформе для информационного менеджмента и проактивной коммуникации. Кроме того, платформа должна обеспечивать будущий анализ и получение отчетов для других типов оперативных данных. Новая система обслуживает внутреннюю аудиторию через модель бизнес-аналитики (BI) предоставление знаний об аварии через дэшборд, интегрированные таблицы и систему репортинга, в том числе через мобильные устройства (например мобильные карты аварий). Внешняя аудитория, такие как пользователи, средства массовой информации, также имеют интерфейсы доступа к данным системы. Местное правительство и органы регулирования, могут получать информацию через различные каналы, такие как вэб, электронную почту6 приложения для смартфонов, текстовые сообщения, интерактивные голосовые системы и даже факс. С точки зрения бизнес-процессов OCS пересекается с многими областями имеющихся сегодня систем коммуникации при авариях. И, чтобы обеспечить развертывание новой системы нужно было вовлечь в процесс ее появления и развертывания многих различных подразделений и отдельных людей. Были организованы семинары, на которых обсуждались функциональности системы, и без этого вряд ли внедрение системы было бы успешным, если бы решение было принято исключительно директоратом.

Главным партнером по поставке нового ПО, в том числе и аналитической системы , работающей в режиме, близко к реальному времени была компания Ventyx. На вершине корневой платформы находятся средства аналитики, средства представления персонала и потребителя, Второй уровень платформы занимают средства двунаправленной коммуникации между приложениями расширенной аналитики Ventyx и фронт-энд приложениями для коммуникаций с пользователями (такие как колл центр, почтовая система, карта аварии и т.п.), а также с приложениями для внутренней аудитории –дэшборд принятия решений, производственные карты аварий, и др. Сезон штормов 2012 года вызвал существенно менее острые нарекания на информированность о ходе аварии. И персонал, и клиенты и другие внешние аудитории получали оперативную и непротиворечащую информацию. В этот сезон было 123 000 подписок на приложения для смартфона, 730 000 пользователей подписались на двусторонние СМС рассылки. Для внутренней аудитории работало 40 досок принятия решений, интерактивная карта работала на сайте и имелись возможности читать сообщения энергетиков на Facebook.

Сегодня на сайте компании http:// www.comed.com вы можете всегда видеть (см. рисунок 3.21) карту отключений и отказов в реальном времени, видеть процесс борьбы с аварией, получать прогноз по ситуации в конкретном выбранном адресе.

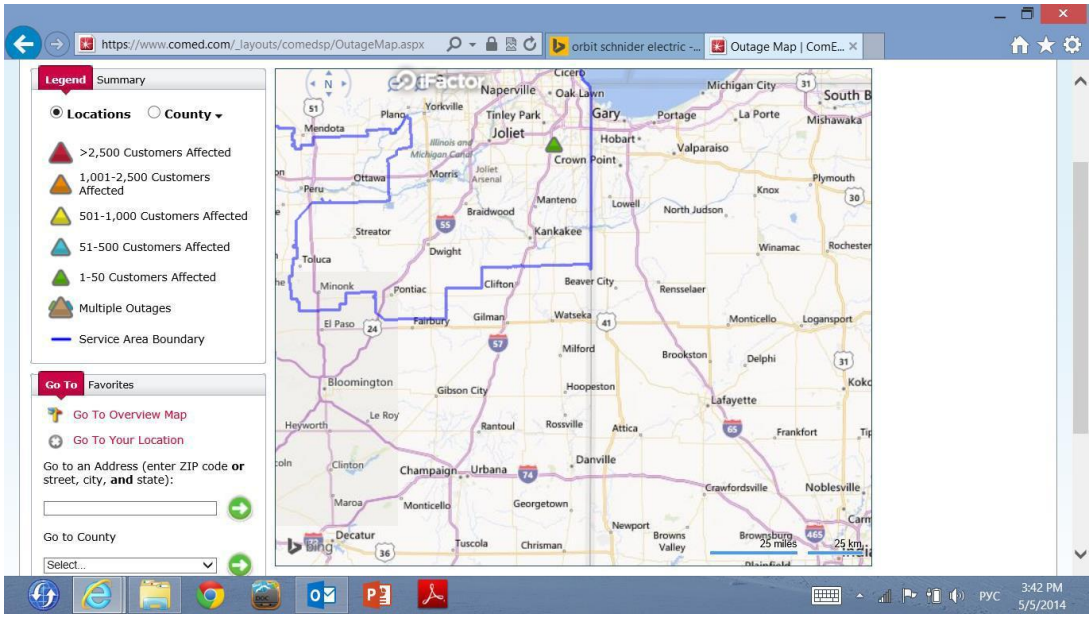


Рисунок 3.21 Сайт мониторинга аварий и блэкаутов компании Commonwealth Edison Company.

Приведем еще один пример применения больших данных к задачам мониторинга и противодействия авариям и блэкаутам. Пример касается энергосистемы национального масштаба Индии.

В Индии почти треть национальной электроэнергии теряется из-за хищений и неэффективности и при этом еще половина населения вообще не имеют доступа к электрической сети. При этом постоянно случаются блэкауты и во время одного из последний из них остались в темноте почти 600 миллионов человек. Все эти проблемы предоставляют широкие возможности для решения в рамках индустрииинтеллектуальных сетей. ЦентральноеправительствоИндииинициировалопрограмму Restructured Accelerated Power Development and Reform Programme, или R-APDRP .

Эта программа предусматривает выделение более 10 миллиардов долларов на модернизацию сетей в ближайшие годы. Главными игроками ожидается здесь участие Infosys, Wipro and Tata Consultancy Services (TCS) - гигантов индийской ИТ индустрии. Одной из главных проблем которая встретилась на пути выполнения программы, является недостаточное инвестирование в распределительном секторе электроэнергетики. И это как известная проблема курицы и яйца – без ясной картины, где происходят потери, непонятно как насытить сеть мощностями. Поэтому 2 млрд было решено инвестировать в фазу проекта для эмпирического нахождения технических потерь для каждого распределительного фидера и трансформатора в каждом городе.

Решение этой грандиозной для Индии задачи включает развертывание измерительной сети, построения ГИС для энергетики, сети автоматической регистрации измерений на каждом трансформаторе и линии, а также индексирование каждого потребителя, обслуживаемого этим трансформатором. Когда будут выделены все технические проблемы, можно оценить также коммерческие, которые окажется возможно решить.

Это детектирование нелегальных подключений, обнаружение ошибочных биллинговых операций и т.п. Для охвата самых малых фрагментов сети требуется

развертывать дополнительные коммуникации к существующим трансформаторам и развертывать для них системы мониторинга через сотовые сети связи и WiMAX. Не менее важным является развертывание бэк-энд ИТ систем для биллинга, сбора данных, точных измерений массы новых данных. Эти проекты безусловно относятся к большим данным и требуют привлечения новых больших инвестиций. Неизбежные финансовые ограничения диктуют поэтапный характер модернизации национальной энергетической сети. На рисунке 3.22 приведена схема для опытной зоны, охватывающей все основные типы сетей и потребителей, на которой производится апробация новых разрабатываемых решений и технологий.

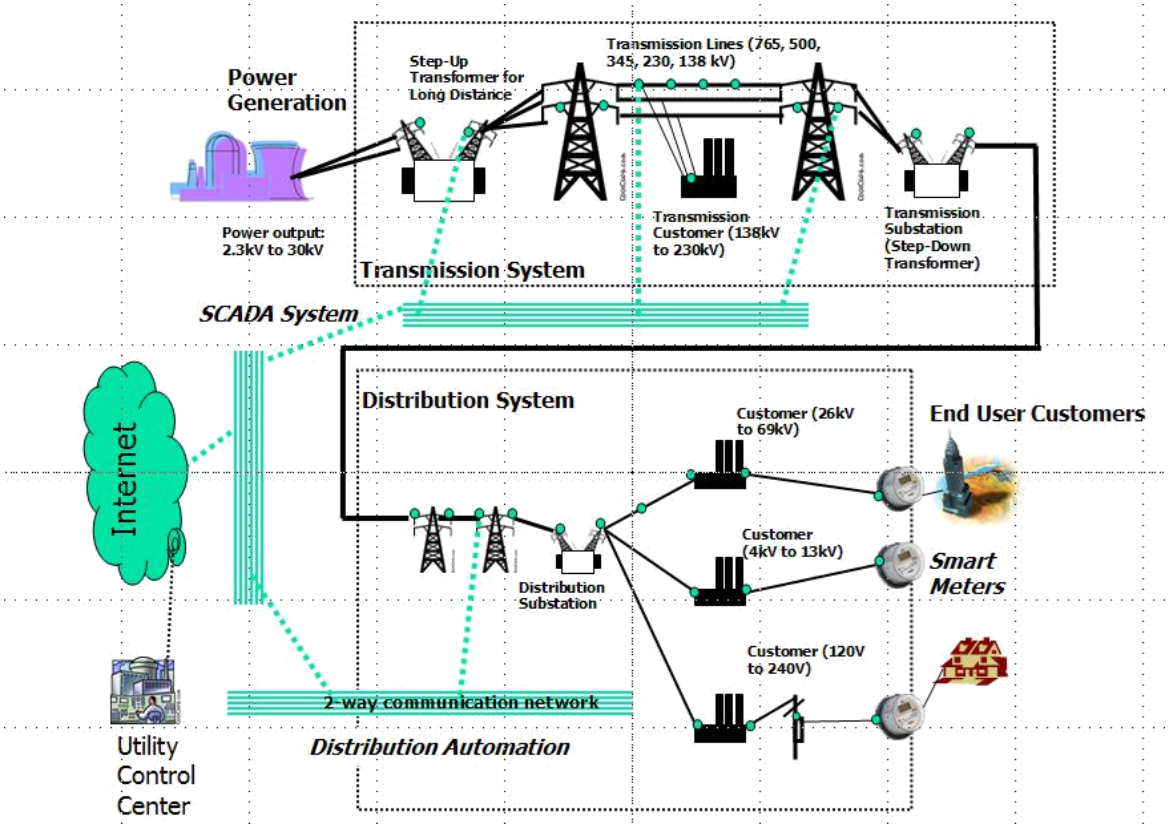


Рисунок 3.22 Концепция интеллектуальной энергосети

Индии согласно одному из проектов R-APDRP.

В январе 2014 года Министерство энергетики Индии назвало 14 проектов по интеллектуальным сетям, сочетающих развертывание интеллектуальных измерителей и имплементацию систем менеджмента отказов и пиковых нагрузок, таких как, например, развертывание проекта по внедрению demand-response со стоимостью 100 млн долларов.

Для снижения затрат на развертывание таких проектов Министерством энергетики Индии был объявлен конкурс на низкостоимостный интеллектуальный измеритель с повышенной секретностью, который мог бы быть основой для построения интеллектуальных сетей национального масштаба.

Учитывая природу блэкаутов в энергосети Индии, для многих больших индустриальных коммерческих и критических правительственных учреждений оказывается притягательным переходить на генерацию своей собственной энергии.

Даже удаленные от цивилизации сельские сообщества нацеливаются на солнечные генераторы и системы генерации, использующие биомассу в составе микрогридов. Более 1000 микрогридов должно быть введено в строй в Индии до 2017 года как для деревень, так и крупных предприятий.

Координация их работы на основе сбора и обработки больших данных будет достойнойзадачей для совместной работы предприятий и специалистов в области энергетики и информационных технологий.

Списоклитературы

1. K. C. danah boyd, «CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholary phenomenon,»*Information, Communication & Society,* т. 15, № 5, pp. 660- 679, 2012.

2. Kapow Software, «Kapow Software,» A Kofax Company, 2014. [В Интернете]. Available: http://www.kapowsoftware.com. [Дата обращения: 6 May 2014].

3. А. Васильев, «Wikibon прогнозирует рост возвратавложений в технологии Big Data,» 20 Сентябрь 2013. [В Интернете]. Available: http://www.computerra.ru/83558/wikibon-big-data-forecast/. [Дата обращения: 6 May 2014].

4. Л. Жуков, «Профессия Data Scientist,» в *Конференция"Big Data Managmentв национальной экономике", http://www.ospcon.ru/files/media/Zhukov.pdf*, Москва, 2013.

5. R. V. R. Kumar, «Classification Algorithms for Data Mining: A Survey,» *International Journal of Innovations in Engineering and Technology (IJIET),* т. 1, № 2, pp. 7- 14, August 2012.

6. D. E. Deza M.M., Encyclopedia of Distances, London - New York: Springer-Verlag Berlin Heidelberg, 2009.

7. D. Brendan J. Frey, «Clustering by passing messages between data points,»*Scince,*т. 315, pp. 972-976, 2007.

8. P. Berkhin, «Survey of Clustering Data Mining Techniques,» 2009. [В Интернете]. Available: http://www.cs.iastate.edu/~honavar/clustering-survey.pdf. [Дата обращения: 7 May 2014].

9. L. X. H. H. H. K. L.-B. Frank Hutter, «Algoritm Runtime Prediction: Methods&Evaluation,» 5 November 2013. [В Интернете]. Available: http://arxiv.org/abs/1211.0906. [Дата обращения: 14 May 2014].

10. Wikipedia, «Small-world experiment,» Wikipedia Foundation, Inc., [В 2] Интернете]. Available: http://en.wikipedia.org/wiki/Small-world_experiment. [Датаобращения: 21 May 2014].

11. D. Dietrich, «The Dirt Little Secret of Big Data Projects,» 18 April 2013. 3] [В Интернете].Available: https://infocus.emc.com/david_dietrich/the-dirty-little-secret-of-big-data-projects/. [Дата обращения: 21 May 2014].

12. «Концепция интеллектуальной электроэнергетической системы сактивноадаптивной сетью,» ОАО "НТЦ электроэнергетики", Москва,2012.

13. Prepared for the US Department of Energy by Litos SategicCommunication under Contract No DE-Ac26-04NT41817, «The SmartGrid: an Introduction,»http://energy.gov/sites/prod/files/oeprod/documentsandMedia/DOE,2010.

14. Z. Pollock, «UTILITY AMI ANALYTICS FOR THE SMART GRID 2013-6] 2020: Applications, Markets and Strategies,» GTM Research, 2013. «Smart Grid Analytics M&A: Sensus Buys Verdeeco,» 7] www.greentechmedia.com.«GridLab-D Welcome,» www.gridlabd.org.

15. GE Digital Energy, «Grid IQ (TM) Insight. Delivering Operational Insights 9] for Utilities Globally,» Digital Energy, Atlanta, US, 2014.

16. GE Digital Energy, «SMOS Smart Metering Operations Suite,» 0] www.gedigitalenergy.com, 2014.

17. C3 Energy, «Smart Grid Analytics Software,» www.c3energy.com, 2014. 1]

18. Pivotal, «Pivotal Big Data Suite. Changing the Economics of Big Data/ 2] Forever,»www.gopivotal.com.Siemens Smart Grid,w3.siemens.com/smartgrid/global/en/pages/default.aspx.Teradata, www.teradata.com.

19. «Energy and Utilities Industry,» wwwericsson.com.

20. Ericsson - E.ON, «Press Release,»http://www.engerati.com/blogs/ericsson-eon-deal-announced-smart-utilities-scandinavia.

21. «Teradata Uilities Data Model,» http://www.teradata.com/logical-data-7] models/Teradata-Utilities-Logical-Data-Model/.

22. J. Montgomery, «Renewable Energy World,»в*IBM's HyRef Seeks to SolveWind's Intermittency Problem http://www.teradata.com/logical-data-models/Teradata-Utilities-Logical-Data-Model/*, Orlando, FL, 2015,August, 15.

23. M. Arancibia, «IBM desaarrola una technologiaaa paraimpulsar lasenergias limpias,» 10 December 2013. [В Интернете]. Available: http://www.capital.cl/negocios/ibm-desarrolla-una-tecnologia-para-impulsar-las-energias-limpias/. [Дата обращения: 21 May 2014].

24. «AutoGrid,» Autogrig, 2014. [ВИнтернете]. Available: http://www.auto

25. grid.com/. [Дата обращения: 24 April 2014].

26. «AutoGrid Lands NTT Data as Big Data Energy Partner (NTS AdvancedTechnology),» Autogrid, 2014. [В Интернете]. Available: http://www.auto-grid.com/gallery/autogrid-lands-ntt-data-as-big-data-energy-partner-nts-advanced-technology/. [Датаобращения: 24 April 2014].

27. «BSI Protection Profile (Germany),» Elster Group GmbH, 2014. [ВИнтернете]. Available: http://www.elster.com/en/bsi-protection-profile-germany. [Датаобращения: 24 April 2014].greentechgrid, «Soft Grid 2013: From Big Data Potential to Real-WorldValue,» Greentech Media Inc., 2014. [В Интернете]. Available: http://www.greentechmedia.com/articles/read/soft-grid-2013-from-big-data-potential-to-real-world-value. [Датаобращения: 24 April 2014].

28. Silver Spring Networks, «Silver Spring Networks,» Silver SpringNetworks, 2014. [В Интернете]. Available:http://www.silverspringnet.com/. [Дата обращения: 24 April 2014].

29. Siemens, «Grid Application Platform,» Siemens AG, 1996-2014. [ВИнтернете]. Available:http://w3.siemens.com/smartgrid/global/en/products-systems-solutions/smart-metering/emeter/Pages/overview.aspx. [Датаобращения: 24 April 2014].

30. Oracle, «Oracle Press Release,» Oracle, 13 December 2012. [ВИнтернете]. Available:http://www.oracle.com/us/corporate/press/1885589. [Датаобращения:24 April 2014].

31. Silicon.Fr, «Oracle s'offre data Raker et le Big Data analyique del'energie,» 2014. [В Интернете]. Available: http://www.silicon.fr/oracle-acquisition-dataraker-big-data-analytique-cloud-energie-81975.html. [Дата обращения: 21 May 2014].

32. Z. Pollock, «Utility AMI Analytics for the Smart Grid 2013-2020:Applications, Markets and Strategies2013,» GTM Research a Greentech Media, Inc., SF, 2013.

33. TIER by Vaisala, «Renewable Energy. Assestment and Forecasting,» 9] 3Tier, Inc., 2014. [В Интернете]. Available: http://www.3tier.com/en/. [Дата обращения: 25 April 2014].

34. Cooper Power Systems, «Volt/VAR Management,» 2014. [В Интернете]. Available:http://www.cooperindustries.com/content/public/en/power_systems/solutions/ivvc.html. [Дата обращения: 28 April 2014].

35. Luxoft, «Smart Energy management,» Luxoft, 2014. [ВИнтернете]. 1] Available: http://www.luxoft.com/energy/dmmessenger-energy-consumption-regulation-platform/. [Датаобращения: 25 April 2014].

36. J. S. John, «Duke Energy: From Smart Grid Devices to Grid ComputingPlatform,» 2 October 2013. [В Интернете]. Available: http://www.greentechmedia.com/articles/read/duke-energy-from-smart-grid-devices-to-grid-computing-platform. [Датаобращения: 5 May 2014].

37. Landis+Gyr, «Landis+Gyr: Manage energy better,» Landis+Gyr, 2014. [ВИнтернете]. Available: http://www.landisgyr.com/. [Датаобращения: 25 Apil 2014].Wkipedia, «Demand responce,» 16 April 2014. [ВИнтернете]. Available:http://en.wikipedia.org/wiki/Demand_response. [Датаобращения: 28April 2014].

38. Berkley Laboratory, «PIER Demand Response Research Center,» 5] Berkley Lab, 2014. [В Интернете]. Available: http://openadr.lbl.gov/.[Дата обращения: 28 April 2014].

39. REstore, «Services for utilities,» 2012. [ВИнтернете]. Available: http://www.restore.eu/. [Дата обращения: 5 May 2014].

40. Space Time Insight, «Situational Intellegence for a Smarter World,»Space Time Insight, 2014. [В Интернете]. Available:http://www.spacetimeinsight.com/. [Дата обращения: 28 April 2014].

41. Electric Power Research institute, «Electric Power Research institute,»Electric Power Research institute, 2014. [В Интернете]. Available: http://www.epri.com/Pages/Default.aspx. [Дата обращения: 28 April 2014].

A. Zurborg, «Unlocking Customer Value: The Virtual Power Plant,»2010. [В Интернете]. Available:http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/ABB_Attachment.pdf. [Дата обращения: 29 April 2014]. http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId= 0] 000000003002000312».

42. J. S. john, «EPRI's Cool New Grid - Scale Energy Storage Tool,» 19 June 2013. [В Интернете]. Available: http://www.greentechmedia.com/articles/read/EPRIs-cool-new-grid-scale-energy-storage-tool. [Датаобращения: 21 May 2014].

43. Ecoult - Energy Storage Solutions, «Frequency Regulation Services,» 2] Ecoult, 2014. [ВИнтернете]. Available: http://www.ecoult.com/case-studies/pjm-pa-usa-frequency-regulation-services/. [Датаобращения: 5 May 2014].

44. Schnider Electric, «Welcome to Orbit,» Schnider Electric, 2013. [В 3] Интернете]. Available: https://orbit.schneider-electric.com/. [Датаобращения: 5 May 2014].

45. Schnider Electric, «GIS - Geospatial Services - ArcFM Server,» Schnider 4] Electric, 2014. [ВИнтернете]. Available: http://www.schneider-electric.us/sites/us/en/solutions/smart-infrastructure/smartinfra-gis-home.page. [Датаобращения: 5 May 2014].ComEd, «ComEd powering lives: An Exelon Company,» Commonwealth

46. EdisonCompany, 2014. [В Интернете]. Available:http://www.comed.com. [Дата обращения: 5 May 2014].

47. Ministry of POWER Government of India, «R-ARDRP - Restructured Accelerated Power Development&Reforms Programme,» Powwr

48. Finance Corporation Ltd., 17 June 2011. [ВИнтернете]. Available:http://www.apdrp.gov.in/. [Дата обращения: 5 May 2014].