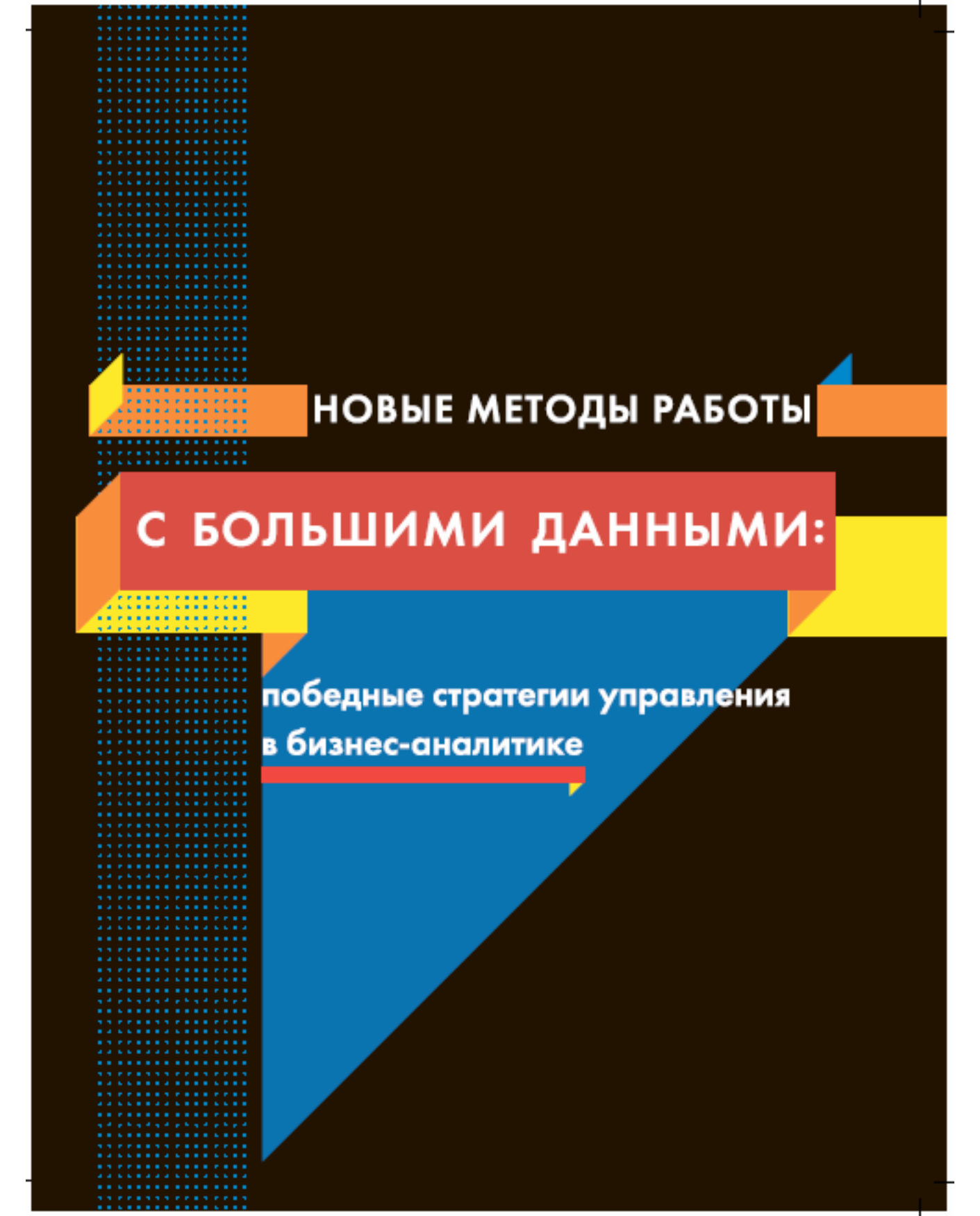




НОВЫЕ МЕТОДЫ РАБОТЫ

С БОЛЬШИМИ ДАННЫМИ:

**победные стратегии управления
в бизнес-аналитике**

Новые методы работы

С БОЛЬШИМИ ДАННЫМИ:

победные стратегии управления
в бизнес-аналитике

Premier
Business
Partner



TM



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Новые методы работы С БОЛЬШИМИ ДАННЫМИ: победные стратегии управления в бизнес-аналитике

Научно-практический сборник
под редакцией профессора А. В. Шмида

Москва
ПАЛЬМИР
2016

ББК 32.973

Н59

Научно-практический сборник

Авторы: **А. В. Шмид** (ответственный редактор, составитель, автор вступительной статьи),
Б. А. Позин, И. В. Галахов, М. А. Агейкин, Д. О. Александров, М. Р. Касимов, Н. И. Клема-
шев, Г. А. Ежов

Дизайн и верстка: Юлия Потресова

Н59 **Новые методы работы с большими данными: победные стратегии управления в бизнес-аналитике** : Научно-практический сборник. Под редакцией доктора технических наук, профессора А. В. Шмида. — М.: ПАЛЬМИР, 2016. — 528 с.: илл.

ISBN 978-5-93811-026-7

Книга знакомит с новейшими технологиями обработки больших данных (Big Data) на примере платформы BIG DATA IBM (базовые технологии экспертной системы Watson IBM).

Переломным в информатике был 2007 год: согласно отчету International Data Corporation (IDC), весь мир впервые за тот год произвел объем данных, превышающий возможности всех существующих сред хранения информации: человечество вступило в эпоху больших данных. Мы оказались свидетелями рождения новой экономики, создания систем поддержки принятия решений, невиданных технологических операций с «информационным сырьем» — своего рода «заводов» по его переработке. Это направление определяется в мире термином Predictive analytics — Предсказательная аналитика

В книге обобщен опыт ЗАО «ЕС-лизинг» в проектировании систем по технологии BIG DATA IBM, содержатся практические лабораторные работы — своего рода «букварь» по обучению действиям с большими данными, — созданные корпорацией IBM для MIT (Массачусетского технологического института), однако адаптированные преподавателями и магистрантами базовой кафедры ЗАО «ЕС-лизинг» «Информационно-аналитические системы» в МИЭМ НИУ ВШЭ для учебных курсов. Книга полезна также широкому кругу читателей, интересующихся проблемами современных информационных технологий.

ББК 32.973

ISBN 978-5-93811-026-7

© EC-leasing, 2016

© Ю. В. Потресова, оформление, 2016

© ПАЛЬМИР, издание, 2016

Часть I

Новейшие технологии обработки Больших Данных



Глава 1

Революция в области философии и технологий принятия корпоративных решений: Big Data

В ближайшие пять лет все компании на рынке разделятся на победителей и побежденных в зависимости от качества их аналитики.

Ginni Rometty
(IBM CEO. 2 марта 2012 года)

1.0 Введение

1.0.1 От информационно-аналитических к информационно-решающим системам

В настоящее время мы являемся свидетелями эволюции коммерческих предложений в области инструментария создания аналитических систем предприятий: от частных решений, например по визуализации аналитических результатов, к представлению интегрированных платформ с полной функциональностью для проведения аналитики по любым типам данных в ранее не доступных объемах. Например, платформы: IBM Big Data, IDOL10 от HP, не остались в стороне ORACLE и SAP. Обзор возможностей платформ от основных поставщиков приведен в таблице 2.4 (стр. 64).

1.0.2 Эволюция от ИТ-станков к ИТ-заводам

Конкурентные преимущества корпораций в конечном счете определяются качеством принимаемых их руководством решений. В этом смысле большинство рыночных предложений однобоко: предлагается только анализ, а синтез решения по умолчанию остается за человеком.

В то же время никого уже не удивляет, что до 80% операций на торгах проходит без участия человека: решения о покупке-продаже принимаются ИТ-роботами.

Суть действительно происходящей сейчас на наших глазах цифровой революции состоит не только в том, что для принятия конкурентных решений анализируются ранее невиданные объемы ранее недоступных для анализа типов данных, но также и в том,

что во все большей степени качество решений машин начинает превосходить качество решений людей. Это надо осознать и к этому надо готовиться, чтобы не оказаться в числе проигравших. Совокупность технологий, стоящих за подготавливаемой всем ходом технического прогресса цифровой революцией, сейчас условно и называется: Big Data.

1.1 Эволюция синтеза управляющих решений

1.1.1 Традиционная схема принятия решений

Человеку разумному свойственно принимать решения. В зависимости от статуса человека последствия принимаемых решений могут коснуться только его одного, а могут повлиять на судьбы партий, корпораций, армий, государств, на судьбы миллионов людей.

Человеку разумному свойственно желать успеха, поэтому он стремится принимать решения, наилучшие из возможных.

Технологии выбора наилучших из возможных решений отработаны и не менялись на протяжении веков. Они подразумевали и подразумевают использование двух групп профессионально подготовленных людей:

- одни отвечают за сбор информации, влияющей на принятие решения (разведчики, аудиторы, бухгалтеры, следователи, диагносты, социологи);
- другие отвечают за подготовку и принятие вариантов решения на основе собранной информации и исторического опыта принятия аналогичных решений (советники, аналитики, эксперты, оракулы, астрологи, провидцы).

Названные профессиональные группы, по сути, обеспечивали реализацию трех (в современной интерпретации) бизнес-процессов:

П1 — поиск и первичную фиксацию искомой информации (добычу информации);

П2 — доставку информации до места ее использования;

П3 — обработку информации и подготовку на ее основе вариантов принятия решения.

Само же решение принимается руководителем — лицом или органом, с точки зрения нашего рассмотрения реализующим две основные функции:

- определение своей информационной потребности для принятия решения, то есть выдача задания остальным участникам процесса на предоставление необходимой для принятия решения информации;

- собственно принятие решения на основе полученной информации.

На рис. 1.1 в самом общем виде представлена схема взаимодействия руководителя с двумя названными профессиональными группами и выделены основные бизнес-процессы, реализуемые людьми (а не компьютерами в составе автоматизированных систем — АС).

То есть, по сути, в упрощенном виде представлен макет оргструктуры взаимодействия трех участников процесса подготовки принятия решения, состоящего из двух бизнес-процессов, обозначенных на рис. 1.1: детализации и реализации.

Оргструктуре, приведенной на рис. 1.1, соответствует ролевая модель в следующем виде:

1. Руководитель.

В процессе детализации: дает задание аналитику на поиск и предоставление необходимой для принятия решения информации (определяет информационную потребность).

В процессе реализации: получает от аналитика искомую информацию; на ее основании принимает решение.

2. Аналитик.

В процессе детализации: детализирует информационную потребность руководителя, определяя состав и местонахождение источников первичной информации, а также состав подлежащей фиксации информации от каждого из источников. Дает задание на поиск и фиксацию искомой информации разведчику (поисковику).

В процессе реализации: получает первичную информацию от разведчика, анализирует, верифицирует и обобщает ее. Приводит информацию к виду, заданному информационной потребностью, и передает руководителю.

3. Разведчик (поисковик).

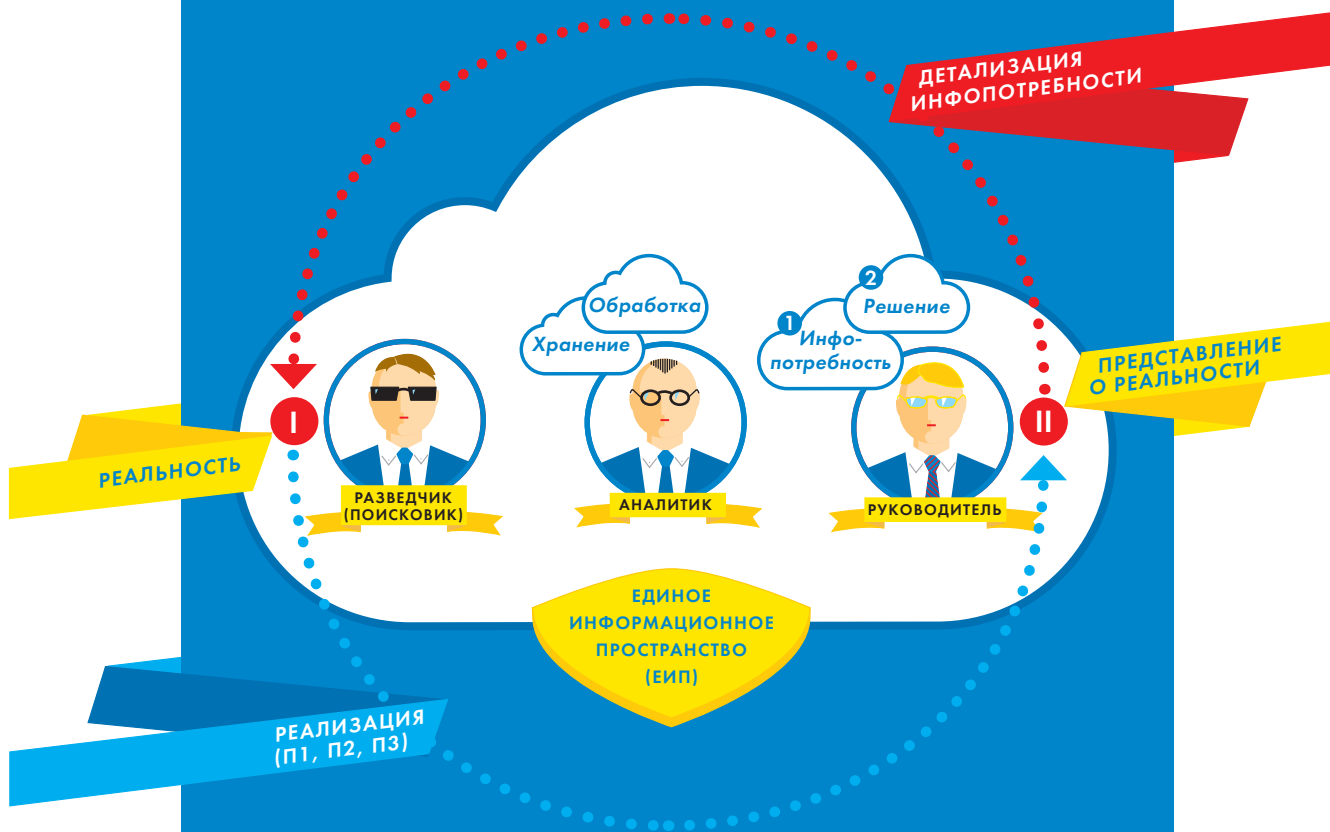
В процессе детализации: налаживает связь с источником (подключает источник).

В процессе реализации: фиксирует информацию источника и передает ее аналитику.

Именно такая и подобные ей схемы подготовки и принятия решений существовали веками.

При реализации таких схем руководители интуитивно стремились придерживаться еще не сформулированного до середины XIX века Чарльзом Дарвином принципа: «Выживает не самый сильный и не самый умный, а тот, кто лучше всех приспосабливается к изменениям».

ТРАДИЦИОННАЯ ПОСТАНОВКА ЗАДАЧИ
ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ
ПРИНЯТИЯ РЕШЕНИЙ
(УДОВЛЕТВОРЕНИЕ ИНФОРМАЦИОННОЙ ПОТРЕБНОСТИ
РУКОВОДИТЕЛЯ)



ИНФОПОТРЕБНОСТЬ ДОЛЖНА УДОВЛЕТВОРЯТЬСЯ:

БЫСТРО

информация
об изменениях
обстановки должна
поступать к моменту
принятия решения
(быстрее,
чем у конкурентов)

ПОЛНО

необходимая
и достаточная
для принятия
решений

ДОСТОВЕРНО

исключение
фальсификации

Борьба за выживание ведется и поныне в мире реальности, только в наше время Реальность стала цифровой, стала киберпространством или цифровым миром [1]. Приведенные на рис. 1.1 вековые принципы создания традиционных схем принятия решений актуальны и поныне.

Именно на основе традиционной схемы работали и в большинстве случаев в России продолжают работать информационно-аналитические системы (ИАС), для которых цифровая реальность ограничена структурированными данными.

Традиционная схема принятия решений так подробно рассматривается здесь с единственной целью: установить, что же изменяется в ней в ходе цифровой революции.

1.1.2 Эволюция базовых технологий в традиционной схеме принятия решений: предпосылки цифровой революции

Появление в середине XX века компьютеров (нового элемента в многовековой истории принятия решений), последующее появление сетей, обеспечивающих практически мгновенную доставку зафиксированной информации к месту обработки, а затем и цифровых средств первичной фиксации информации в реальном времени (видеокамеры, радиометки, цифровые телефоны, мобильные устройства, персональные компьютеры) завершило формирование необходимых условий для комплексной цифровой автоматизации в совокупности всех трех бизнес-процессов подготовки и принятия решений (П1, П2, П3).

На наших глазах происходят быстрые количественные изменения во всех трех бизнес-процессах традиционной схемы принятия решений:

П1 — возрастают объемы и множатся типы доступной для анализа информации. Это происходит за счет как расширения географического охвата и числа контролируемых процессов природы, общества, материального производства и потребления в связи с ростом числа и состава фиксирующих устройств, так и взрывного роста объема фиксируемой информации по каждому из контролируемых процессов;

П2 — возрастают возможности цифровых сетей по доставке зафиксированной информации к местам обработки, как по географическому охвату контролируемых процессов, так и по пропускной способности;

П3 — возрастают вычислительные возможности компьютеров по анализу доступной информации, а в последнее время во все

Рис. 1.1
Традиционная
постановка задачи
информационной
поддержки принятия
решений

возрастающей степени уже и по синтезу окончательных решений. Причем это возрастание в соответствии с правилом Мура происходит в геометрической прогрессии.

Названные количественные факторы в совокупности приводят к выводу о неизбежности грядущих качественных изменений взрывного порядка в области принятия решений — цифровой революции.

Цифровая революция подразумевает получение для компаний — участников технологического прорыва подавляющих конкурентных преимуществ, обусловленных качественно новыми технологиями подготовки, и принятия управленческих решений, перед компаниями, не применяющими подобных технологий.

Как отмечается в [1], цифровая революция, происходящая на наших глазах в нашем цифровом мире, несет в себе как огромные возможности, так и огромные угрозы.

Возможности — для тех, кто впишется в этот новый, формирующийся на наших глазах цифровой мир. Угрозы — для тех, кто упустит такую возможность.

Сам этот мир огромен: согласно данной в [1] оценке, всю информацию, содержащуюся во всех средах хранения цифрового мира на 2011 год, можно было бы оценить астрономической цифрой: около 1800 эксабайт!

То есть:

$1800 \text{ эксабайт} = 1800 \times 1024 \text{ петабайт} = 1800 \times 1024^2 \text{ терабайт} = 1800 \times 1024^3 \text{ гигабайт} = 1800 \times 1024^4 \text{ мегабайт} = 1800 \times 1024^5 \text{ килобайт} = 1800 \times 1024^6 \text{ байт.}$

Остановимся на рассмотрении новых технологий, предопределяющих приход цифровой революции.

1.2 Новые технологии цифровой революции в области принятия управленческих решений (Big Data)

Big Data сейчас, видимо, наиболее популярная и обсуждаемая тема в области ИТ.

Рынок завален огромным количеством рекламных по своей сути материалов, и их количество таково, что за этим частотолом важных, но деталей достаточно трудно рассмотреть основные инновационные идеи, в совокупности и определяющие технологический прорыв.

Попытаемся провести выделение и классификацию инновационных технологических идей, лежащих в основе цифровой

революции, в наиболее общей форме, без привязки к деталям: то есть с точки зрения здравого смысла.

В основу классификации положим оценку влияния технологических идей на качество управленческого решения, то есть на то, что и является искомым конкурентным преимуществом, достигаемым в итоге цифровой революции.

Именно качество принимаемых управленческих решений и разделяет противоборствующие на рынке компании на победителей и побежденных.

Интуитивно ясно, что качество управленческого решения определяется двумя основными факторами:

- информированностью решающего центра (объемом доступной информации, ее своевременностью и достоверностью);
- интеллектуальностью решающего центра (умением адекватно воспользоваться доступной информацией).

Попробуем провести выделение определяющих технологий Big Data (BD) отдельно по этим двум базовым факторам.

Как отмечается во всех доступных материалах, новые технологии не заменяют и не отменяют уже существующие технологии, а дополняют, развивают и интегрируют их, совместно обеспечивая выход на новый виток технического прогресса.

1.2.1 Технологии BD, влияющие на интеллектуальность решающего центра.

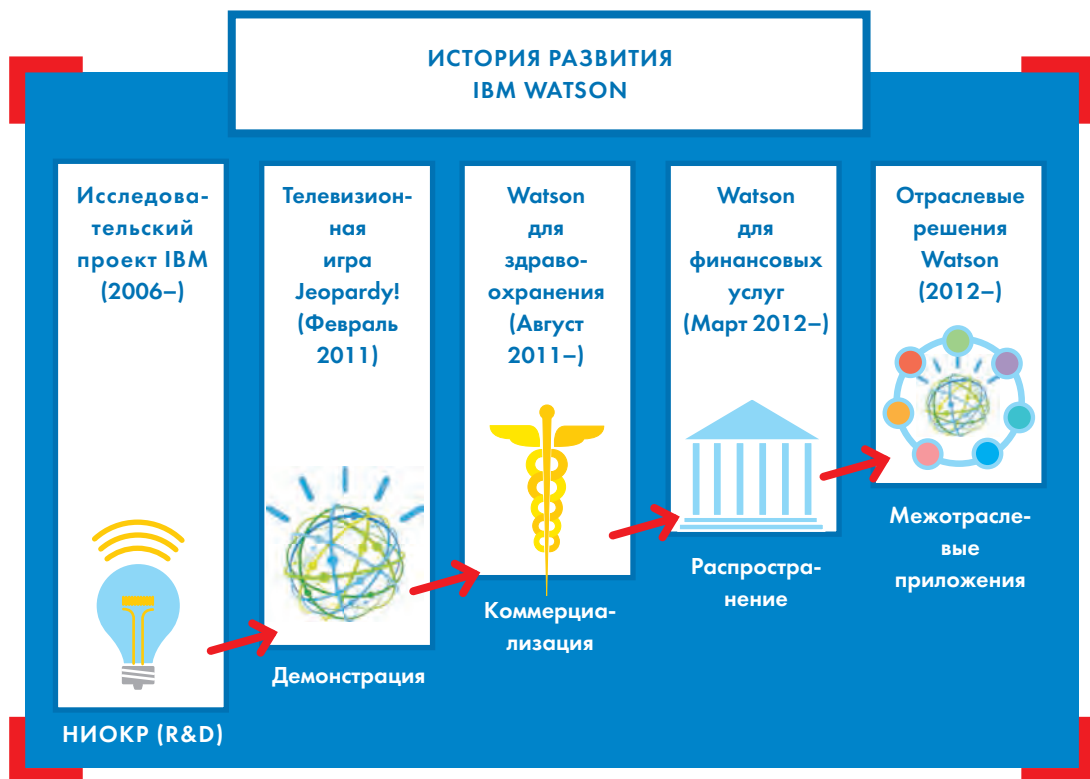
1.2.1.1 Кто будет принимать решения — человек или машина?

В 2005 году делегация наших специалистов (в которую вошел и автор) посетила лабораторию IBM в Цюрихе и была озадачена со стратегическим прогнозом развития ИТ на период 2005–2015 годов, составленным нобелевскими лауреатами. Суть прогноза состояла в том, что к 2015 году компьютеры будут принимать лучшие, нежели люди, решения во многих областях человеческой деятельности: в банковском деле, медицине, экономике, биржевых операциях и др.

Прогноз основывался на общеизвестной тенденции возрастания производительности компьютеров в геометрической прогрессии (в 2 раза каждые 2 года — правило Мура), с соответствующей тенденцией возрастания возможных объемов хранения.

А чем компьютер производительнее, тем более интеллектуальными оказываются принимаемые им решения.

В качестве исходной точки прогноза выбран 1998 год: в этом году компьютер IBM обыграл в шахматы чемпиона мира



Г. Каспарова. С этого момента в шахматы играть с компьютером бесполезно — он будет выигрывать у человека всегда.

Превосходство машины над человеком в области принятия управленческих решений в большинстве областей человеческой деятельности прогнозировалось на 2015 год и в ряде других областей.

Как же IBM выполняет свой собственный прогноз? И выполняет ли? 2 марта 2012 года в Новом Орлеане (США) на конференции основных бизнес-партнеров IBM первая за 105-летнюю историю IBM назначенная CEO IBM женщина, Джинни Рометти, сделала следующие заявления:

- первоочередной задачей IBM на ближайшие 5 лет является создание нового поколения аналитических систем;
- в ближайшие 5 лет все компании на мировом рынке разделятся на победителей и побежденных в зависимости от качества их аналитических систем;
- 12 апреля 2012 года IBM объявит о коммерческой доступности экспертно-решающей системы Watson.

Это объявление IBM Watson состоялось.

То, как IBM представляет историю развития Watson, показано на рис. 1.2.

Интересно, что старт исследовательскому проекту был дан в 2006 году — через год после выхода стратегического прогноза по развитию ИТ (R&D на рис. 1.2).

Состоявшаяся в 2011 году демонстрация IBM Watson была как нельзя более широкой: на всю телевизионную аудиторию США. В качестве демонстрационного примера было выбрано участие Watson в телевизионной игре Jeopardy (рис. 1.3) — прототипе передачи «Своя игра» в РФ.

Система Watson, как и участники-люди, общалась с ведущим, отвечала на вопросы и в итоге победила людей с подавляющим счетом.

Впечатление от первой публичной демонстрации Watson было потрясающим. Фантастика публично становилась реальностью. Но, как видно из рис. 1.2, это было только началом, незначительным эпизодом на карьерном пути Watson.

Приведем краткую информацию о Watson, собранную нами из доступных источников.

В основу Watson положены три революционные технологии, обозначенные на рис. 1.4.

Watson названа в честь президента IBM Томаса Вотсона (Thomas J. Watson), создавшего в 1964 году прорывную для тех времен

Рис. 1.2
История развития
IBM Watson

Рис. 1.3
2011 год — Watson
победила в первой
в истории телеигре-
викторине Jeopardy
с участием компьютера

IBM WATSON – СОЧЕТАНИЕ РЕВОЛЮЦИОННЫХ ТЕХНОЛОГИЙ



технологии: машины серии IBM 360, на несколько последующих десятилетий обеспечивших IBM абсолютное лидерство в области ИТ на мировом рынке.

В СССР архитектурные аналоги этих машин независимо от IBM также производились с 1971 по 1992 год под названием Единая система электронных вычислительных машин (ЕС ЭВМ), а с 1990 по 1996 год под руководством автора совместно с IBM был реализован проект по проектированию и производству в РФ трех моделей ЕС ЭВМ на элементной базе IBM.

Названием Watson IBM подчеркивает прорывной характер и этой технологии.

Watson построена по технологии DeepQA (Deep Question Answering), включающей генерацию гипотез, массовый сбор фактов, анализ и ранжирование гипотез на основе фактов. Взаимодействие основных ее элементов схематично представлено на рис. 1.5.

Команда IBM обеспечила и ежедневно обеспечивает Watson миллионами документов, включая энциклопедии, словари, тезаурусы, новостные статьи, литературные работы. Watson также использует базы данных, таксономии и онтологии, особенно DBPedia, WordNet, и Yago. Watson может обрабатывать 500 гигабайт в секунду, что эквивалентно чтению одного миллиона книг в секунду.

Следует отметить, что для человека существует фундаментальный предел: каждый из нас за всю жизнь в состоянии прочесть не более 12–14 тысяч книг.

Сопоставление этих фактов приводит к очевидному выводу: с точки зрения информированности Watson уже сейчас подавляюще более осведомлена о реальности, нежели любой человек. И этот разрыв в осведомленности будет только возрастать.

Более 100 различных методов используются для анализа естественной речи, определения источников данных, генерации гипотез, поиска и ранжирования фактов, объединения и подтверждения гипотез.

В медицине Watson помогает анализировать специфические симптомы, историю болезни, историю наследственных заболеваний и выполнять синтез этих данных с доступной неструктурированной и структурированной медицинской информацией, включая медицинские книги и статьи. IBM поясняет, что Watson не пытается заменить врачей, но помогает им избежать ошибок и уточнить диагноз. Благодаря своим свойствам информированности и интеллектуальности Watson снижает вероятность ошибки по первичному

Рис. 1.4
IBM Watson —
сочетание
революционных
технологий

Рис. 1.5
Алгоритм работы
экспертно-решающей
системы Watson

ПОМОЩЬ ОНКОЛОГАМ В ДИАГНОСТИРОВАНИИ И ВЫБОРЕ СХЕМЫ ЛЕЧЕНИЯ РАКА



ПРОБЛЕМА

Требуется улучшенная индивидуальная диагностика рака и разработка плана лечения с ускоренной адаптацией последних достижений медицины

РЕШЕНИЕ

Объективное рассмотрение процедур на основе клинических, медицинских данных и сведений о пациенте

Анализ и рекомендации, основанные на оценке фактов в соответствии с данными, знаниями и стандартами

Необходимость дополнительной информации и обучение на опыте

Задача снизит вероятность ошибки при первоначальной диагностике рака, достигающей в среднем 20–44%



диагнозу, поставленному человеком, с 20–40% до 2–3% (рис. 1.6).

В итоге сегодня Watson первая и пока единственная система такого класса, реально претендующая на замещение человека при принятии решений в цифровом мире — наивысшее достижение в области ИТ.

В юбилейном выпуске журнала «Популярная механика» (англ. Popular Mechanics) [2] приведены 10 интервью с ведущими учеными мира, обрисовывающими пределы нашего знания в биологии, физике, медицине, информационных технологиях. Характерно, что в качестве наивысшего в мире достижения в области информационных технологий избрана экспертно-решающая система Watson.

В журнале приведено интервью с Дэвидом Ферруччи — руководителем разработки Watson, в котором он старается объяснить интервьюирующим его экспертам, что Watson пока не содержит искусственного интеллекта по той простой причине, что ученым пока непонятно, что такое интеллект и как работает мозг, это

Рис. 1.6
← Применение WATSON
в онкологии

всего лишь синтез новейших достижений в различных областях математики, лингвистики, ИТ.

Хотя внешне Watson выглядит как система, обладающая искусственным интеллектом, да еще каким!

В феврале 2013 года Watson сдала в медицинском университете выпускные экзамены, приобретая тем самым юридическое право лечить людей. Watson осуществляет первичный прием пациентов в 30 госпиталях США, проводя, кроме опроса пациентов, анализ любых видов медицинских данных, она является на настоящее время ведущим консультантом в области онкологии в США. Известно также, что Watson проходит обучение в трех крупнейших американских банках и некоторых других областях.

Итак, цитата от Дэвида Ферруччи [2]: *«Компьютер решает отдельные задачи гораздо эффективнее человека, но мыслить, как человек, машина не может».*

Это про Watson.

Для практики гораздо важнее выделенная курсивом часть цитаты. Истина в том, что круг задач, которые машина уже решает лучше, чем человек, стремительно расширяется (от шахмат в 1998 году).

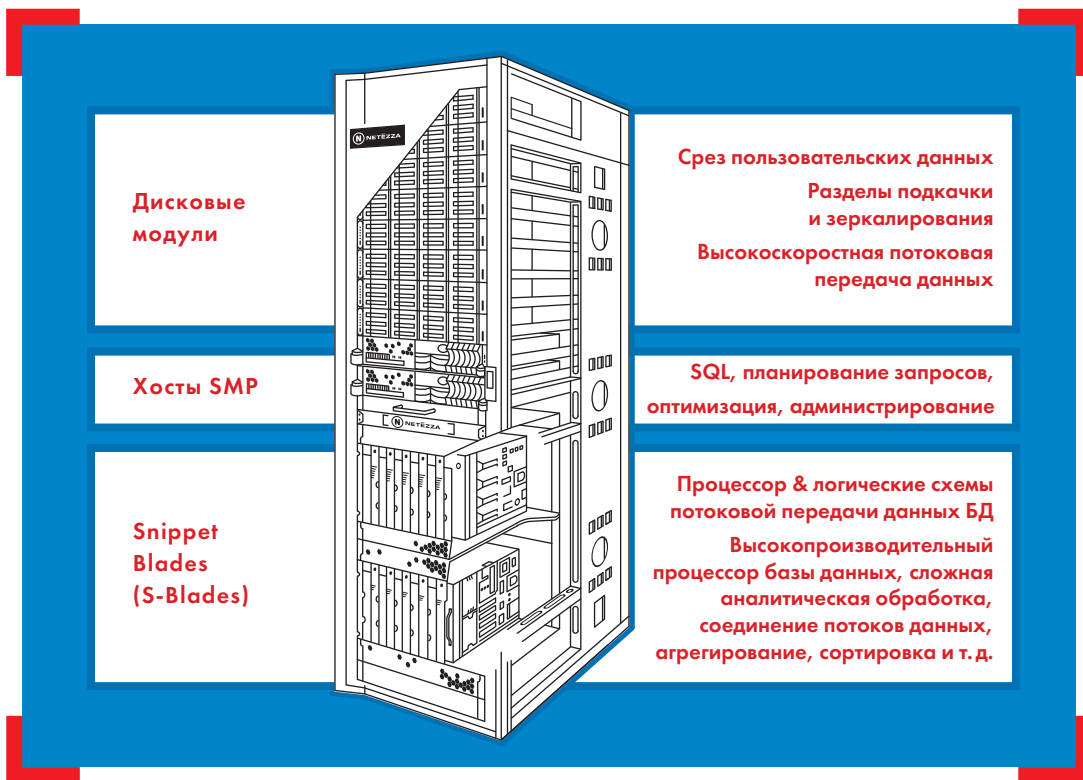
И если от качественных решений в медицине выигрывают все, то, например, в играх в экономику и финансы будут и проигравшие — те, для кого качество решений Watson окажется недоступным.

Видимо, именно это и имела в виду IBM CEO, когда предвещала, что в ближайшие пять лет все фирмы разделятся на победителей и побежденных в зависимости от качества их аналитики.

Исходя из изложенного, мы можем сделать вывод о том, что в традиционной схеме принятия решений (см. рис. 1.1) появляется *качественное* изменение: машина по всевозрастающему кругу задач отбирает у человека роль руководителя (в части функции «принятие решений»), правда, пока оставляя человеку целеполагание и предоставляя людям счастливую возможность обучать себя.

До самообучения машин в проекте Watson дело пока не дошло. Тем не менее появление *самообучающихся* компьютеров также запланировано IBM, но пока отложено (ориентировочно) на 30 лет [3].

В ходе цифровой революции борьба за повышение интеллектуальности решающего центра приводит и, согласно прогнозам, будет во все возрастающей степени приводить к замещению человека машиной при принятии решений, обеспечивая переход от аналитических к экспертно-решающим системам.



1.2.1.2 Специализированные ускорители обработки аналитических запросов к реляционным базам данных: скорая помощь аналитикам (людям и роботам)

Цифровая реальность, в которой и ведется интеллектуальная борьба за выживание, быстро меняется во времени.

По Дарвину следует, что чем быстрее участники борьбы будут принимать адекватные изменению реальности решения, тем выше их шансы на победу в конкурентной борьбе.

Если обратиться к традиционной схеме принятия решений (см. рис. 1.1), то очевидно, что ускорять следует исполнение всех трех бизнес-процессов, предшествующих принятию решения: П1, П2, П3.

Ускорители обработки запросов предназначены для ускорения процесса П3 — аналитической обработки накопленной информации с целью подготовки и оценки последствий различных вариантов решений.

← Рис. 1.7
Общий вид
и основная
функциональность
ускорителя IBM PureData
System for Analytics

Суть проблемы, на решение которой ориентированы ускорители, состоит в следующем: традиционные ИАС быстро обрабатывают запросы аналитиков к базе в том случае, если структура запросов и объекты поиска в базе заранее известны. То есть заранее определены: модель данных, логическая и физическая структуры реляционной базы. Однако аналитики высшего уровня заранее не могут определить для себя, какие именно запросы базе они будут осуществлять. Более того, известно, что последующие запросы будут наверняка изменяться в зависимости от ответов на предыдущие запросы. В такой ситуации (нерегламентированные заранее запросы) ответы от базы будут поступать недопустимо медленно. Известный выход — для ускорения получения ответов перепроектировать и переустановить базу — абсолютно неприемлем по времени. Тупик.

Специализированные ускорители обработки нерегламентных (непредвиденных при проектировании базы) запросов и предназначены для выхода из этого тупика в интересах аналитиков высшего уровня, хотя являются скорой помощью для аналитиков любой квалификации: и людей, и роботов.

Доступными на рынке ускорителями являются, например: IBM PureData System for Analytics (ранее Netezza), EMC Greenplum, HP Vertica, Oracle Exadata, SAP Hana.

Идеи, на которых они базируются, в общем, очевидны: массовый параллелизм и адаптивная динамическая перестройка структуры базы под запрос.

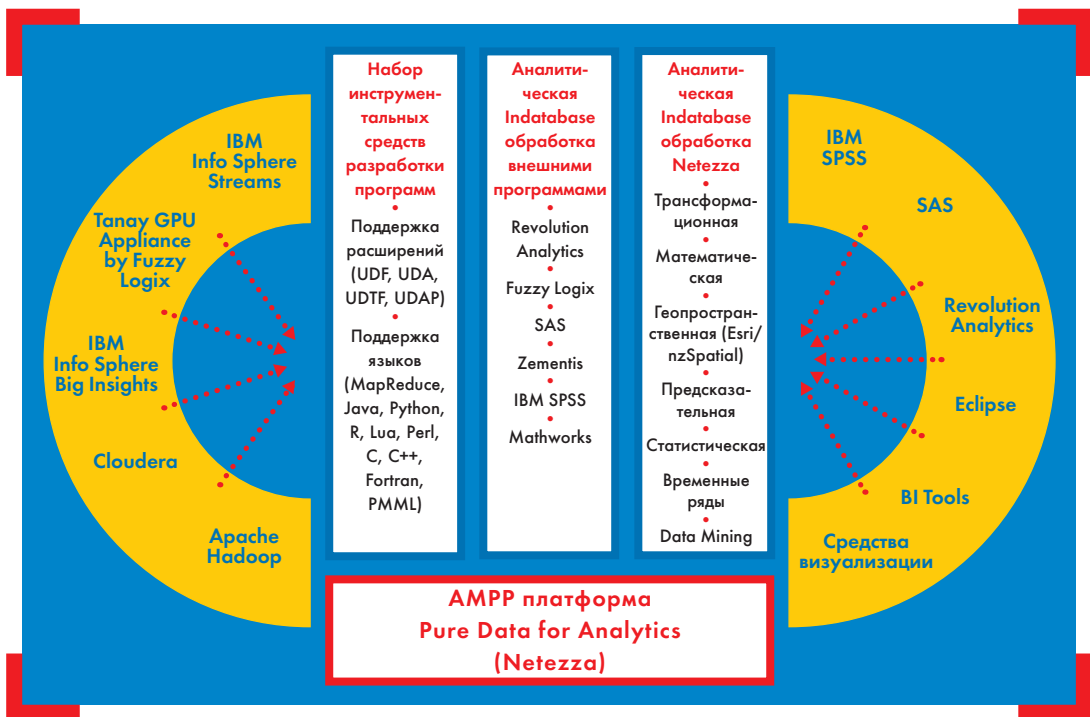
Общий вид и основная функциональность ускорителя IBM PureData System for Analytics приведены на рис. 1.7 из [3].

Выполнение запросов в лучших примерах ускоряется почти в 200 раз (!!!) для анализа данных с объемом до 1 петабайта в одном, представленном на рис. 1.7, ящике. А ящиков может быть и много. И это ведь без переделки приложения вообще! Просто ускоритель подключается как сервер базы приложения для SQL-запросов любых видов.

Один-два дня на подключение — и скорая помощь пришла.

1.2.1.3 Интеграция новых и старых технологий повышения интеллектуальности решающего центра: «назад в будущее»

Примером интеграции абсолютно новых и традиционных технологий повышения качества решений может служить сконструированная IBM вокруг PureData System for Analytics функционально



полная платформа для аналитики и прогнозирования развития ситуаций, представленная на рис. 1.8 из [3].

Левая часть рисунка представляет состав технологий сбора данных, где наряду с давно применяемыми технологиями (Apache Hadoop для работы с неструктурированными данными всех видов и его усовершенствованного корпорацией IBM варианта IBM InfoSphere BigInsight), показаны и новейшие технологии потоковой обработки, рассматриваемые в следующем разделе (IBM InfoSphere Streams).

Центральная часть рисунка — ядро платформы, аналог корпоративного хранилища данных (КХД), но на основе PureData System for Analytics (скорость, скорость!).

А правая часть рисунка представляет давно освоенный аналитиками состав продуктов анализа и прогнозирования: Cognos BI, IBM SPSS для построения статистических моделей и т.д.

Прагматичный синтез старого и нового выводит бизнес-процесс принятия решений на основе структурированных данных на качественно новый уровень. Это происходит за счет включения в бизнес-процесс новых типов данных: неструктурированных и потоковых.

← **Рис. 1.8**
Платформа
для аналитики
и прогнозирования
развития ситуаций,
сконструированная
вокруг PureData System
for Analytics

1.2.2. Технологии Big Data, влияющие на информированность решающего центра

1.2.2.1 Объем доступной информации цифрового мира и проблема проклятия размерности

Любой решающий центр максимально информирован, если в своих решениях он, так или иначе, учитывает всю необходимую и достаточную для принятия решения информацию нашего цифрового мира [1].

Как же велик объем этой всей информации нашего мира? Другими словами, как на самом деле велика эта Big Data? Каковы тенденции изменения этого объема во времени?

Интересующие нас оценки содержатся, например, в материале IDC, разработанном при спонсорской поддержке EMC [1].

На рис. 1.9 из этого материала приводится график ежегодного роста объема возникающей во всем мире информации с 2007 по 2011 год.

Первое, что бросается в глаза, это то, что этот объем действительно огромен: он измеряется в эксабайтах!

А второе — тенденция роста! При всей его огромности он еще (по мнению IDC) и возрастает не менее чем в 10 раз каждые 5 лет.

← **Рис. 1.9**
Ежегодный
рост объема
возникающей во всем
мире информации

СОЗДАВАЕМАЯ ИНФОРМАЦИЯ И ДОСТУПНЫЕ СРЕДСТВА ХРАНЕНИЯ



Source: IDC, 2008

Так, по мнению IBM CEO V.Rometty, пять дней в 2013 году по объему возникающей в цифровом мире информации эквивалентны всему 2007 году [4].

Возникает закономерный общий вопрос: а какую часть из подобного объема постоянно возникающей информации мы можем в принципе успеть запомнить, с тем чтобы на досуге разобраться, нет ли там чего-либо полезного для принятия наших решений? Ведь в случае невозможности сохранения эта информация как неотвратимо возникнет, так и безвозвратно пропадет!

Что может повлиять на информированность решающего центра? Очевиден и ответ: сохранить можно объем возникающих данных, не превышающий имеющегося свободного объема всех сред хранения данных нашего мира текущего года!

Соотношения объемов ежегодно возникающих данных и объема данных, который можно в принципе ежегодно запоминать во всех средах хранения нашего мира, представлены на рис. 1.10 из [1]. Синяя кривая — объем ежегодно возникающих данных в эксабайтах по годам.

Красная кривая — объем данных, который можно в принципе запомнить в текущем году из возникших.

Круговая диаграмма отражает процентные отношения запоминаемых объемов данных по средам хранения.

← Рис. 1.10
Количество
создаваемой
информации и объем
устройств хранения
данных:
синяя кривая — объем
ежегодно возникающих
данных в эксабайтах
по годам;
красная кривая — объем
данных, который можно
в принципе запомнить
в текущем году из вновь
возникших. Круговая
диаграмма отражает
процентные отношения
запоминаемых объемов
данных по средам
хранения

Важнейшим выводом из рис. 1.10 является тот факт, что начиная с 2007 года объем возникающих ежегодно данных начинает во всевозрастающей степени превышать объем данных, который мы могли бы в принципе запомнить!

И в 2011 году более половины возникающих в цифровом мире данных уже пропадает навсегда бесследно! Безвозвратные потери мировых данных нарастают!

Возникает *проблема проклятия размерности*.

Мы во всевозрастающей степени теряем представление о нашей реальности в связи с ограниченными возможностями хранения ее цифровых описаний.

Это явление можно назвать, например, амнезией (потерей памяти — а это название болезни) цифрового мира.

Можно долго рассуждать на тему о том, важная или неважная информация была утрачена, но поскольку мы ее не анализировали, то все эти рассуждения пустые. Потери могут повлиять на информированность решающего центра.

Как же лечить эту болезнь (амнезию цифрового мира), в принципе снижающую информированность решающего центра за счет безвозвратной утраты всевозрастающей части мировых данных с сопутствующей утратой содержащегося в этих данных (возможно) полезного для решения контента?

1.2.2.2 Технологии Big Data, влияющие на информированность решающего центра

Идея Big Data № 1 — потоковая обработка (streaming): преодоление проклятия размерности при хранении данных.

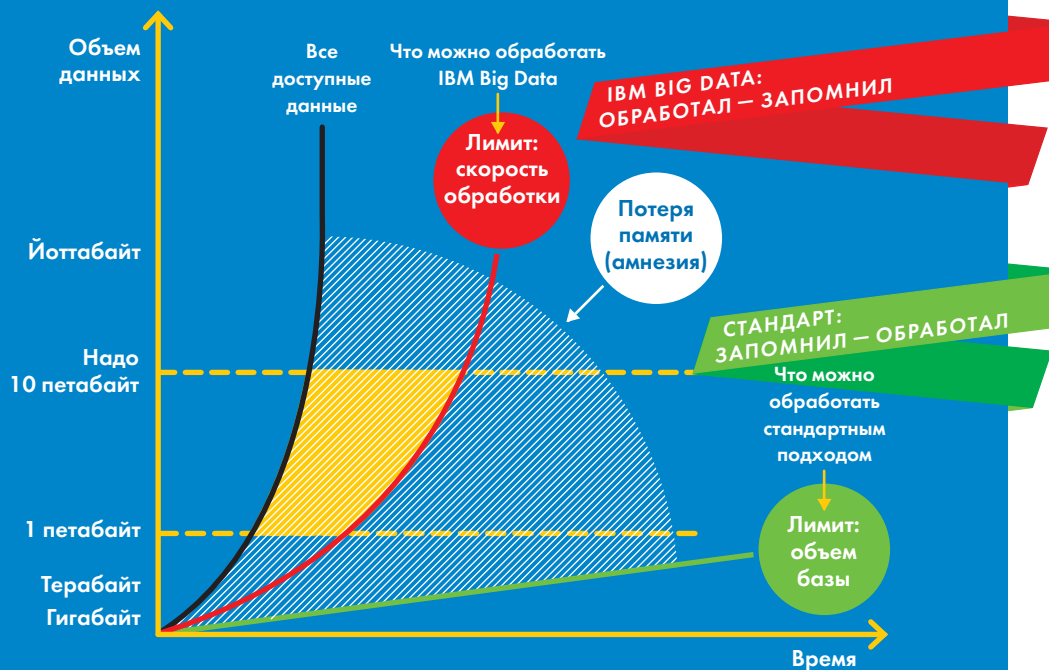
Как нас учат психологи, давайте ответим на вечный вопрос: в чем ваша проблема, что вас беспокоит? Ну не можем мы запомнить все мировые данные (памяти не хватает — а это непреодолимо), а поэтому, возможно, теряем контент.

Очевидно, что нам нужен именно контент. Если можно обойтись без запоминания первичных данных, то почему нет?

Откуда же берется необходимость запоминания *всей информации*, порождаемой источником?

Да из исторически сложившихся и привычных способов анализа данных: мы сначала их копируем к себе и запоминаем, а потом начинаем в них копаться с целью выделения контента. Так работают сейчас практически все значимые аналитические системы. При подобном подходе (запомнил — проанализировал) запоминается как сам контент, так и весь возможный шум, который мог

«АМНЕЗИЯ КОРПОРАЦИИ» ИЛИ ОБЪЕМ ВОСПРИНИМАЕМЫХ ИАС ПЕРВИЧНЫХ ДАННЫХ



КАЧЕСТВО ИАС

1. Объем первичных данных
2. Глубина и скорость анализа (площадь анализа)
3. Кто готовит и кто принимает решение

**ПОТОКОВАЯ ОБРАБОТКА
IBM BIG DATA
КАРДИНАЛЬНО
ПОВЫШАЕТ ОБЪЕМ
ИСПОЛЬЗУЕМЫХ ДАННЫХ
ДЛЯ АНАЛИТИЧЕСКОЙ
РАЗВЕДКИ**

Центральный банк
Российской
Федерации
640 ТБ

Российская
государственная
библиотека
им. Ленина
7 ТБ

бы быть выделен и уничтожен, но только в результате анализа данных, который при традиционном подходе *следует* за запоминанием (extraction — transformation — loading: ETL).

То есть нашей проблемой является наш же подход к анализу: запомнил — проанализировал.

Но ведь давным-давно в смежных областях (радиолокации, акустике, теории систем управления) существуют и принципиально иные по последовательности операторов подходы к анализу именно совершенно неструктурированных данных: например, алгоритмы фильтрации с целью выделения сигнала на фоне шума. Сначала выделение полезного сигнала на фоне шума, а затем анализ полезного сигнала и принятие решения.

Принципиальным при подобных подходах является тот факт, что не требуется предварительного запоминания всего объема данных, порождаемого источником, а отделение полезного контента от шума производится ДО принятия решения и/или запоминания контента.

Streaming (потокковая обработка) — это одна из новых технологий Big Data. Она меняет порядок обработки данных от источника: вместо «запомнил — обработал» появляется последовательность «обработал — запомнил».

Описание подобной технологии, включенной, например, в состав платформы IBM Big Data, содержится в [4].

Очевидно, чем выше доля устраняемого шума, тем меньше выделенного полезного контента придется запомнить для последующих принятий решений по историческим данным и тем ниже требования к объемам хранения.

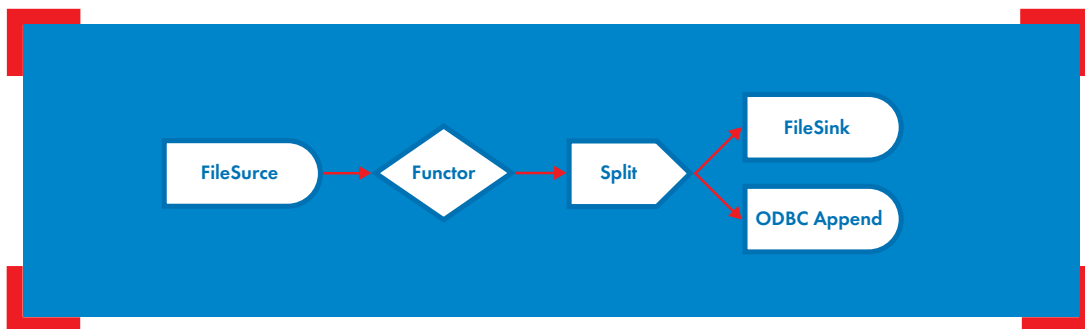
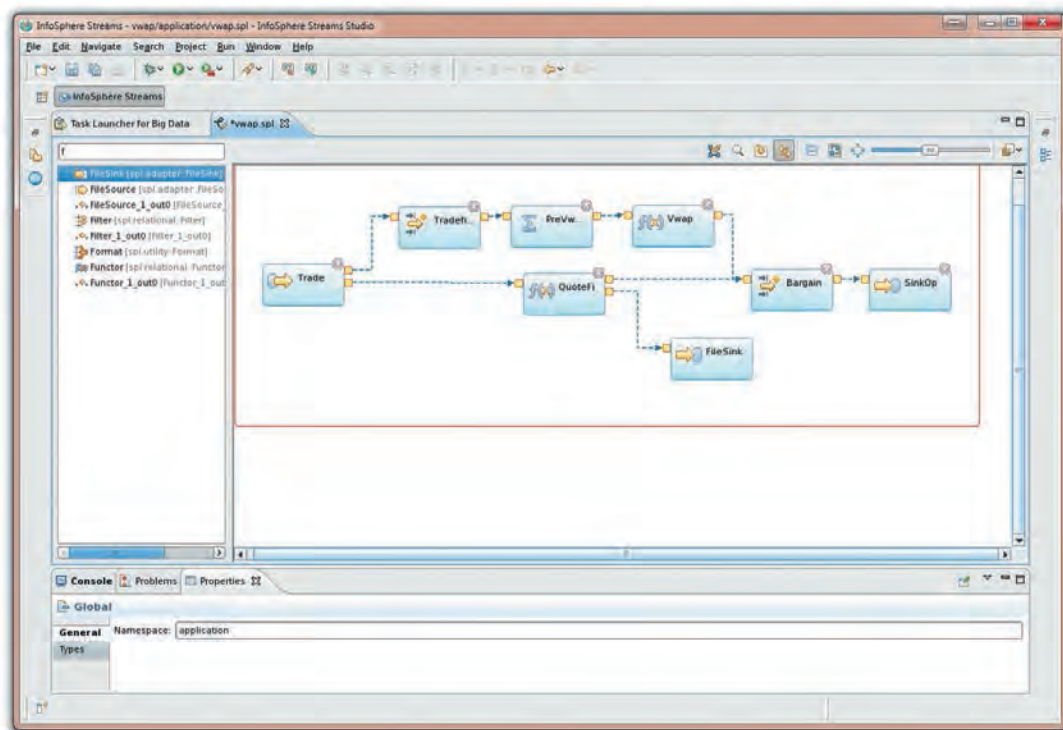
Поскольку большинство источников работают в реальном времени, то и фильтрацию приходится производить в реальном времени, то есть быстро. В итоге streaming, снижая требования к памяти, повышает требования к производительности при фильтрации.

Фундаментальным пределом при потоковой обработке является не объем памяти, а производительность при фильтрации.

Ничего не дается даром. Утешением является тот факт, что поскольку источники работают независимо, то при потоковой обработке возможен массовый параллелизм — то есть применение больших ферм параллельно работающих компьютеров с возможностью их масштабирования.

На рис. 1.11 в самом общем виде демонстрируются взаимосвязь основных соотношений информированности ИАС: доступных для анализа объемов первичных данных при традиционном

← **Рис. 1.11**
Амнезия корпорации:
зеленая линия —
объем данных, доступных
для традиционного
подхода: «запомнил —
обработал»;
красная кривая —
объем данных,
доступных для потоковой
обработки: «обработал —
запомнил»;
черная кривая —
все доступные
для обработки данные



подходе «запомнил — обработал» и потоковой обработке «обработал — запомнил».

Разница в площадях под красной и зеленой кривыми по отношению к площади под черной кривой и характеризует уровень амнезии — количество потерянных во времени первичных данных.

В качестве примеров приведены объемы информации, хранимой в самой большой библиотеке страны и в Центральном банке РФ.

← Рис. 1.12
Пример расчета
средней взвешенной
на объеме (VWAP)
и запись значений
из QuoteFilter в файл

← Рис. 1.13
Поток данных,
к которому применяется
преобразование
и разделение на два
возможных выхода
по предустановленной
логике

Рекордсменом по объему хранения в РФ является Яндекс: емкость его информационных хранилищ приближается к 10 петабайтам. Но и такие возможности хранения не спасают от амнезии.

Очевидно, что наиболее перспективной технологией с точки зрения повышения информированности ИАС является технология потоковой обработки, отодвигающей время наступления амнезии конкретной ИАС в ее жизненном цикле.

Поток (stream) в этой технологии представляет собой непрерывную последовательность элементов данных любых видов. Поток-овое приложение может быть интерпретировано в виде графа, содержащего узлы, связанные направленными ветвями. В узлах графа располагаются операторы или адаптеры, обрабатывающие данные из потоков. Сами потоковые приложения могут разрабатываться как в графической, так и в текстовой форме.

На рис. 1.12 и 1.13 приведены фрагменты программы на языке потоковой обработки SPL (Streams Processing Language) от IBM [4] в графической и структурной формах.

Рекомендации по применению потоковой обработки связаны на бизнес-уровне с тремя бизнес-требованиями:

1. Огромный (не помещающийся в память) объем данных для анализа, поступающих с предельной скоростью.
2. Время реакции (анализа данных и синтеза решения) системы: миллисекунды.
3. Любые типы источников.

Наличие любого из первых двух означенных факторов уже является рекомендацией для применения потоковой обработки. По сути, это рекомендации по работе с очень быстрыми источниками, производящими очень большие объемы данных.

Для описания подобных реалий цифрового мира вводится и новый подход к классификации источников и данных [4].

Источники классифицируются по четырем измерениям вариативности (V4):

- скорость выдачи информации (Velocity),
- объем выдаваемой информации (Volume),
- достоверность выдаваемой информации (Veracity),
- способ выдачи информации (Variety).

И для некоторых типов источников применима только потоковая обработка: запомнить не успеете.

Данные же делятся на две категории:

- данные в покое (уже запомненные);
- данные в движении (поступающие от источника, но еще не запомненные).

**BIG DATA:
ИЗМЕНЕНИЕ ПАРАДИГМЫ
ПРИ РАБОТЕ С ДАННЫМИ**

ТРАДИЦИОННЫЙ ПОДХОД
Структурный и повторяемый
анализ

ЗАПОМНИЛ — ОБРАБОТАЛ

БИЗНЕС

Определяет,
что спросить



ИТ

Структурирует данные
для ответа на вопрос

Месячная
отчетность

Анализ
прибыльности

Анализ
анкет

**Ограничение:
ПАМЯТЬ**

БОЛЬШИЕ ДАННЫЕ
Итеративность
и исследование

ОБРАБОТАЛ — ЗАПОМНИЛ

ИТ

Обеспечивает платформу
для креативного анализа



БИЗНЕС

Исследует,
что можно спросить

Отношение
к бренду

Стратегия
продуктов

Оптимизация
ресурсов

**Ограничение:
ПРОИЗВОДИТЕЛЬ-
НОСТЬ**

Данным в покое [4] ставится в соответствие море, а данным в движении — Ниагарский водопад.

Очевидно, что потоковая обработка как метод анализа данных влияет как на философию, так и на технологию в традиционной схеме принятия решений.

Например, в функции «Описание инфопотребности» (см. рис. 1.1) вместо традиционного: «предоставьте среднюю зарплату по моей корпорации» — появится что-то вроде: «сформируйте мне психологический портрет инсайдера в моей корпорации и соберите по этому портрету всю информацию по внутрикорпоративным и внешним источникам с целью его выявления».

В переводе на русский язык: «пойди туда — не знаю куда, принеси то — не знаю что»!

Эта фраза из сказки как нельзя лучше описывает изменения в философии принятия решений, вносимые потоковой обработкой:

- априорно состав, форма и географическое расположение источников, в отличие от традиционной схемы, неизвестны;
- если где-то и имеется полезный для решения контент, то он, скорее всего, сильно замусорен посторонней информацией — шумом.

Выручает тот факт, что проанализированы могут быть несравнимые с доступными традиционными технологиями объемы данных и любые типы источников. Поле поиска резко расширяется вплоть до всего киберпространства, и нужны креативные идеи, сужающие его до условий задачи.

Возникающее принципиальное изменение подхода к работе с данными отражено на рис. 1.14.

Главное, что следует из рис. 1.14, — это изменение традиционных взаимоотношений и ролей бизнеса и ИТ: *людей из бизнеса и ИТ надо переучивать на принципиально иные бизнес-процессы и учить работать вместе.*

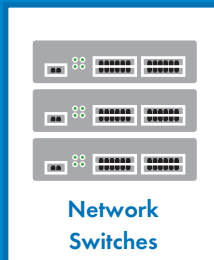
Садиться вместе и думать, куда пойти («пойди туда — не знаю куда») из вариантов, предоставляемых платформой креативного анализа (см. рис. 1.14).

В составе *креативной команды управления креативной платформой IBM Big Data* изменяет архитектуру бизнес-процессов принятия решений, оргструктуру и роли персонала *по традиционной схеме принятия решений.*

На наш взгляд, именно в этом и состоит коренное отличие современной схемы принятия решений от традиционной, порождаемое невиданными ранее технологическими возможностями по «копанию в данных» во всем киберпространстве.

← Рис. 1.14
Изменение парадигмы
при работе с данными

БОЛЬШИЕ ДАННЫЕ — ГОРЯЧАЯ ТЕМА,
потому что технологии сделали возможным
анализ **ВСЕХ** доступных данных



ЭФФЕКТИВНО С ТОЧКИ ЗРЕНИЯ ЗАТРАТ
УПРАВЛЯТЬ И АНАЛИЗИРОВАТЬ
ВСЕ ДОСТУПНЫЕ ДАННЫЕ
В ИХ ПЕРВОЗДАННОМ ВИДЕ —
СТРУКТУРИРОВАННЫЕ, НЕСТРУКТУРИРОВАННЫЕ,
ПОТОКОВЫЕ

С технологической точки зрения очевидно, что современная платформа для реализации ИАС должна обладать технологиями для работы как с данными в покое (структурированными и неструктурированными), так и с данными в движении (мощными потоками данных от любых типов источников).

Поэтому очевидно, что неотъемлемым свойством платформы должен быть инструмент интеграции всех видов данных для последующего комплексного анализа и принятия решения, в общем виде такой, как показано на рис. 1.15.

Идея Big Data № 2: обучаемые системы реального времени

Целью обучения ИТ-систем несомненно является улучшение их характеристик по двум базовым факторам, влияющим на качество решения: информированности и интеллектуальности ИТ-системы.

Любые идеи обучения людей или систем базируются на применении накопленных в прошлом знаний для принятия решений в настоящем либо прогнозирования будущего, на идее обратной связи из прошлого в настоящее.

В популярной форме можно объяснить идею обучения ИТ-систем по аналогии со сборкой мозаики [4].

Действительно, гораздо легче найти место для очередного доставаемого из коробки элемента мозаики, если воссоздаваемый при сборке мозаики образ уже в значительной степени определен и начинает проясняться, что же нужно искать в коробке с не вставленными еще элементами мозаики.

То же самое можно сказать и по поводу очередного приходящего от источника элемента информации: уже накопленный контент оказывается полезным для отделения контента, вновь пришедшего в составе новой порции данных, от шума.

Здесь речь идет о системах с обратной связью с архитектурой, представленной на рис. 1.16 из [3].

Специально подчеркивается, что (как следует из рис. 1.15) данная технология применима ко всем известным источникам.

Можно опять провести аналогию с мозаикой, а можно сказать, что по мере накопления исторического опыта принятия решения система с обратной связью как бы учится: становится все умнее и умнее и качество принимаемых ею решений повышается.

Например, обучение может состоять в накоплении метаданных: состава шаблонов для выделения полезного контента из входного потока.

Поскольку книга [4] написана ведущими сотрудниками IBM, а экспертно-решающая система Watson также является изделием

Рис. 1.15

← Анализ
всех доступных данных

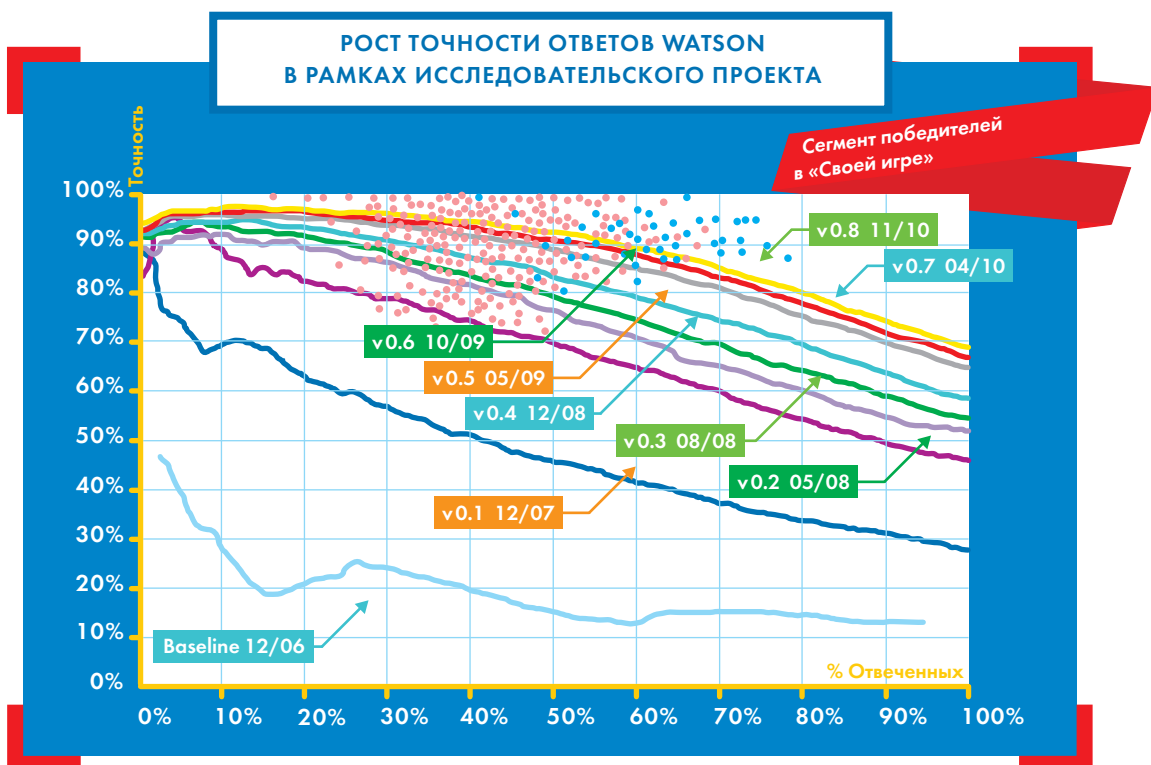


Рис. 1.16
Экосистема
аналитической
обработки
в реальном времени

IBM, то можно предположить, что когда говорится о том, что Watson непрерывно учится, то речь идет именно об этом. В подтверждение сказанного на рис. 1.17 приводятся кривые роста точности ответов Watson во времени в процессе его обучения для выработки победных стратегий в игре Jeopardy. Из рис. 1.17 следует, что сам процесс обучения может длиться годами.

Но гораздо важнее другое. Мы сталкиваемся с новой реальностью: возникновением обучаемых компьютеров и возникновением новой специальности — учителей обучаемых компьютеров — членов креативной команды обучаемого компьютера.

В целом можно сказать, что повышение эффективности выделения полезного сигнала на фоне шума при решении задач ИТ на основе адаптивных алгоритмов радикальным образом сказывается на показателях назначения системы в целом. Снижаются требования к необходимому для принятия решения объему вычислений, к объемам памяти. А в ряде случаев (по аналогии с адаптивными системами управления) принятие выигрышных решений вообще невозможно без применения адаптивных систем.

Идея Big Data № 3: открыть для себя цифровой мир

В настоящее время большинство предприятий, во всяком случае, в нашей стране, используют при принятии решений аналитические системы, базирующиеся на внутренней информации предприятия.

Поучительной для них может явиться информация, содержащаяся в рис. 1.18 из [1].

На рисунке классифицируется и оценивается вся информация цифрового мира (1,773 эксабайт в 2011 году) по трем категориям:

1. контент, генерируемый пользователями предприятий — 1,234 эксабайта (желтый круг).
2. контент, интересующий предприятия — 1,530 эксабайта (синий круг).
3. доля контента в генерируемом пользователями объеме, интересная для предприятий, — 1000 эксабайт (пересечение желтого и синего кругов — зеленая область).

То есть из всего полезного контента, обеспечивавшего информированность решающего центра предприятий (1530 эксабайт), 1000 эксабайт находится вне предприятий, и только 530 эксабайт — внутри предприятий.

Не забывая о внутреннем контенте предприятия, надо сконцентрироваться на изучении внешнего контента — открыть для себя цифровой мир.

Рис. 1.17
Рост точности ответов
Watson в рамках
исследовательского
проекта



Идея Big Data № 4: свободный адаптивный поиск и добыча информации — на войне как на войне

В предыдущих разделах мы остановились на ключевых технологических идеях IBM Big Data по работе с информацией цифрового мира: как переварить огромные, ранее недоступные объемы информации в реальном времени, как отфильтровать контент от шума и куда следует обратиться за информацией.

Мы теперь понимаем, что большая часть нужной для выживания предприятия информации находится вне предприятия, и знаем, как ее обработать. Но как ее обнаружить и получить в интересах своего предприятия?

То есть как эффективно реализовать функцию П1 (добычу информации) традиционной схемы принятия решений, но уже в цифровом мире?

Ведь если добыча внешней для предприятия информации не организована, то и обрабатывать нечего. От качества оружия — поисковой платформы для обнаружения и рассмотрения всей необходимой информации как внутри предприятия, так и вне его — зависит информированность решающего центра.

Способы добычи информации на протяжении веков делились на два класса: законные и незаконные.

Рис. 1.18
Пользовательская
и корпоративная
информация

Ничего не изменилось сегодня и в цифровом мире.

Будем говорить, что законную добычу информации в киберпространстве осуществляют поисковики, а незаконную (с точки зрения атакуемой стороны) осуществляют разведчики (см. рис. 1.1). Реалиями киберпространства являются и те и другие.

Незаконные способы добычи информации реализуются хакерами и спецслужбами (современный пример: разоблачения Сноудена) через вирусы, черви и т. д.

Законные же способы добычи информации технологически реализуются через инструментарий поисковых платформ, предоставляющих для команд аналитиков (ИТ и бизнес) возможности свободного креативного поиска во всем киберпространстве (см. рис. 1.14).

На рис. 1.19 приводится архитектура поисковой платформы Watson Explorer — одной из поисковых платформ IBM Big Data для работы с текстовыми данными в покое и движении от любых видов источников.

Этот инструмент позволяет вести во всем киберпространстве поиск, индексацию и визуализацию контента в интересах единой креативной команды предприятия [4].

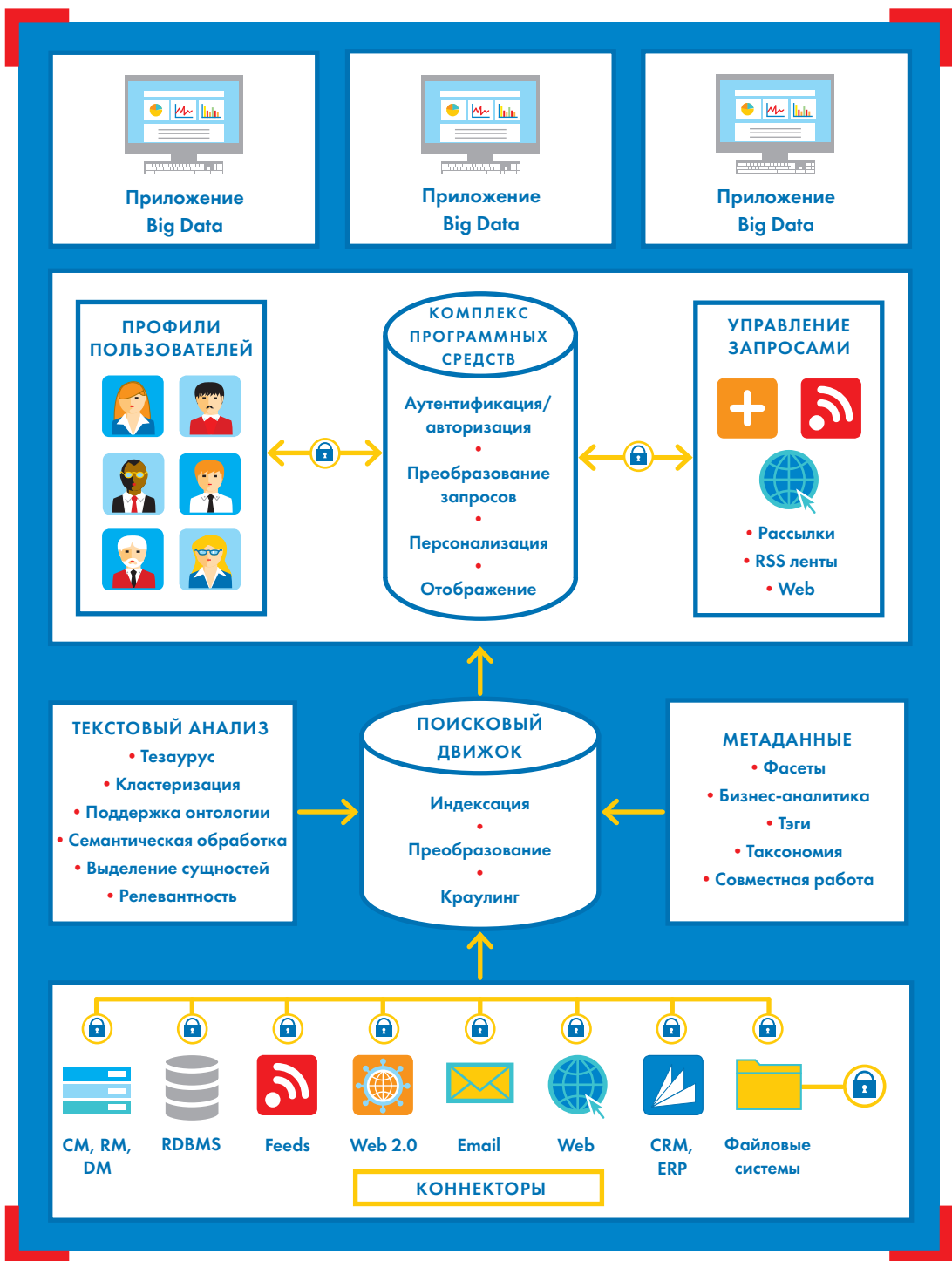
IBM Watson Explorer архитектурно разделен на четыре слоя:

- слой коннекторов предоставляет возможности стандартного доступа к множеству разнотипных внешних и внутренних для предприятия источников данных;
- процессинговый слой организует поиск источников в интересах и по заданиям креативной команды, нанесение найденных источников на карту поиска команды (навигацию), передачу информации для визуализации;
- пользовательский слой организует ситуационный центр поиска и визуализации найденной информации для креативной команды;
- в слое приложений присутствует специфичная для данной креативной команды аналитика найденной информации, как раз и реализующая конкурентные преимущества.

В этом примере наглядно проявляются концептуальные отличия современной схемы принятия решений от традиционной.

Watson Explorer реализует концепцию свободного поиска по всему киберпространству под управлением креативной команды, с обратной связью на эту команду.

Платформа предоставляет команде заранее неизвестные ВОЗМОЖНОСТИ, которые в ходе поиска расширяются, что кардинально изменяет бизнес-процесс доступа к искомой информации



и вместе с ним оргструктуру и ролевую модель реализации бизнес-процесса поиска, доступа к информации и последующего принятия решения.

Оба бизнес-процесса (П1 и П3 на рис. 1.1) из статических и априорно predetermined становятся динамическими и с обратными связями: уже полученная информация позволяет расширять или сужать круг поиска и повышать качество решения по сравнению с традиционной схемой.

В принципе архитектурное решение и функциональный состав этого примера могут служить прототипом для работы с любыми видами данных, а не только с текстовыми.

1.3 Реализация современных схем принятия решений на основе платформ Big Data в общем случае: сборка инструмента из конструктора платформы

1.3.1 Общие соображения: философия сборки

Каковы же необходимые и достаточные условия получения конкурентных решений? Любая платформа Big Data может быть уподоблена детскому конструктору: она содержит множество сопрягаемых между собой различных элементов, но оставляет ребенку возможности для творчества: сборки модели чего-либо. В инструкциях к конструктору на моей памяти содержались описания сборки моделей паровоза, самолета и т. д. Но можно было проявить фантазию и собрать что-то совершенно удивительное — такое, чего нет ни у кого.

Целью сборки из конструктора платформы IBM Big Data является получение на основе стандартных элементов современной схемы принятия решений, обеспечивающей компании конкурентные преимущества. Но поскольку конструкторы (платформы) доступны всем, то все и могут начать собирать одинаковые «паровозы» по инструкции.

Откуда же взяться конкурентным преимуществам? Где взять такое, чего нет ни у кого?

Если обратиться к рис. 1.14, то ответ напрашивается сам собой. Конструктор дает возможность собрать креативную платформу для принятия решений — инструмент команды.

Но креативность платформы обеспечивается совокупным интеллектуальным потенциалом работающих с платформой людей: бизнес и ИТ вместе — квалификацией команды.

Рис. 1.19
Архитектура
Watson Explorer



Да взять хотя бы Watson с его очевидно подавляющими преимуществами при принятии решений как в телеиграх, так теперь и в медицине и далее везде.

IBM не скрывает, что конструктором, из которого собирался Watson, является платформа IBM Big Data. И платформа от Watson теперь стала коммерчески доступной. Но это совсем не означает, что даже получив в свои руки сконфигурированный под Watson инструмент, кто угодно получит те же самые результаты. Ведь с 2006 года на креативной платформе работал и работает огромный совместный коллектив медиков (и не только) и информатизаторов, обеспечивавший конкурентные преимущества проекта Watson.

Наличие качественного инструмента работы с цифровой реальностью является только *необходимым* условием успеха, *достаточным* условием является наличие креативной команды, в совершенстве владеющей инструментом.

1.3.2 Что мы будем собирать из конструктора? И какое изделие мы собираемся предъявлять на испытания?

По документации (общедоступной, во всяком случае, для бизнес-партнеров IBM) конструктор IBM Big Data раскрывается более чем в 900 различных продуктах, шаблонах, готовых решениях и т. д. А уж число сочетаний этих элементов просто астрономическое. В силу этого должны быть сформулированы путеводные принципы сборки изделия под задачу (схемы принятия решения, оргструктуры, ролевой модели).

Задачей же является построение из элементов конструктора и технологического обеспечения современной схемы принятия решений, как минимум превосходящей уже существующую традиционную схему принятия решений вашего предприятия. А как максимум — обладающую конкурентными преимуществами в вашей отрасли бизнеса.

Мы теперь понимаем, что конкурентное преимущество схемы принятия решений определяется *креативной командой*, владеющей современной платформой для автоматизации подготовки и принятия решений.

Рис. 1.20
Минимальный
функциональный состав
конструктора ИАС

Остановимся на определении технических требований к функциональности подобной современной платформы в общем виде.

Список функций, которые должны быть реализованы в составе платформы, приведен на рис. 1.20, в левой части которого представлены искомые функции, а в правой части им поставлены в соответствие возможные варианты решений из состава продуктов IBM Big Data.

В качестве упражнения можно поставить в соответствие набору базовых функций возможные варианты решений из других платформ Big Data.

Все означенные на рисунке элементы платформы в общем случае необходимы, однако эффективность работы *креативной команды* определяется удобствами пользования *инструментом*. Поэтому наличие хорошего *ситуационного центра* для управления всеми стадиями работы с данными при принятии решения (по аналогии с Watson Explorer, см. рис. 1.19) обязательно.

Очевидно, что все обратные связи между стратегиями поиска и добычи данных, анализа данных, подготовки и передачи вариантов решения руководителю существуют и замыкаются через *креативную команду*, работающую в *ситуационном центре*. Причем, в отличие от традиционной схемы, обратные связи не

WATSON

предоставляет три класса сервисов



ОТВЕТЫ НА ВОПРОСЫ

- Большие объемы данных и фактов
- Уточняющие вопросы
- Запросы на естественном языке
- Генерация гипотез



ИССЛЕДОВАНИЕ

- Найти подтверждающие факты для заданных гипотез
- Определить нехватку фактов для оценки гипотезы
- Уточнить гипотезу



ПРИНЯТИЕ РЕШЕНИЙ

- Использование и анализ отраслевых баз знаний и моделей принятия решений
- Генерация решения на основе подтверждающих фактов
- Обучение и обратная связь

являются априорно статичными, а *обязаны* быстро изменяться во времени:

- по результатам уже произведенного поиска;
- по реакциям реальности на уже принятые решения.

И конкурентные преимущества определяются качеством работы *креативной команды*: качеством работы *людей*.

То есть со звездных высот надо опуститься на грешную землю и сказать: как и ранее, мы имеем дело с человеко-машинной системой.

Большие данные — большими данными, а внедрять-то АС — проза жизни! — придется по ГОСТ Р 34 (Автоматизированные системы), включающему покупные изделия — детали из конструктора и инфраструктурные компоненты.

Пример Watson не является исключением: решения принимает собственно экспертная система, а ее обучение, постановка отраслевых задач, сбор и анализ информации под задачу проводятся креативной командой.

1.4 Заключение: что происходит с традиционной схемой решений (рис. 1.1)? Что делать нам? Чему учиться?

1.4.1 Революция: решения принимают роботы, а не люди

Пришло время выполнить данное в разделе 1 обещание и рассмотреть основные изменения в традиционной схеме принятия решений, порожденные как всевозрастающей оцифровкой Реальности, так и появлением совокупности технологий IBM Big Data. Рассмотреть на уровне оргструктуры и ролевой модели — действий людей.

Человек покидает мировую сцену принятия решений и садится в партер в роли наблюдателя.

На наш взгляд, наиболее революционным из изменений в традиционной схеме принятия решений является переход функции руководителя (в части принятия решений) от человека к экспертно-решающей системе. Функция руководителя «задание информации» хотя и видоизменяется, но пока остается за человеком. Фирмами (пока еще) юридически командуют люди.

Уместно показать набор из трех сервисов, реализуемых экспертно-решающей системой Watson (рис. 1.21) при обсуждении решаемых проблем с людьми, и последующего САМОСТОЯТЕЛЬНОГО принятия решения.

Процесс принятия решения состоит из:

- выбора гипотезы, в наибольшей степени отвечающей имеющимся фактам;
- запроса недостающих фактов, подтверждающих гипотезу;
- признания гипотезы решением.

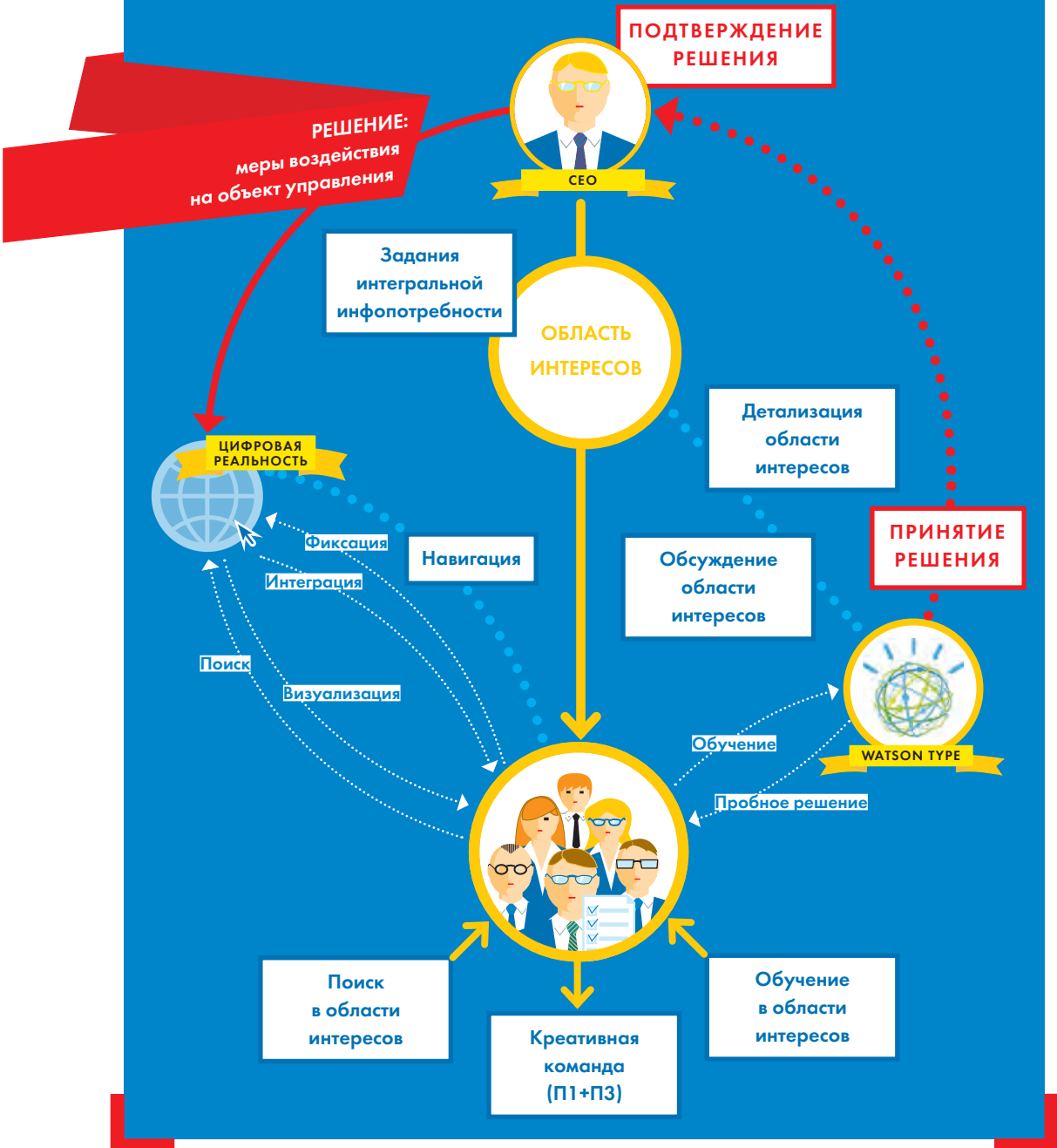
В случае с медициной состав гипотез конечен и определен заранее: это все известные болезни из справочника. Стратегией принятия решения может быть план проведения исследований (анализов) пациента для исключения гипотез (диагнозов), не соответствующих результатам анализов, а затем предоставление принятого решения (диагноза) на подтверждение врачу (CEO, если это не медицина).

Рис. 1.21
Три класса сервисов
IBM Watson [4]

СОВРЕМЕННАЯ СХЕМА ПРИНЯТИЯ РЕШЕНИЙ:

ЧЕМУ УЧИТЬ?

Работа в креативной команде



С учетом всего изложенного можно предположить, что эскиз современной схемы принятия решений с применением экспертно-решающей системы мог бы выглядеть так, как показано на рис. 1.22.

1.4.2 Эволюция: люди пока остаются, но их роли изменяются

Итак, отметим различия в ролях людей в традиционной (см. рис. 1.1) и современной (см. рис. 1.22) схемах принятия решений. Чего уж точно не было в роли пока еще *руководителя* (см. рис. 1.1), а теперь *СЕО* (см. рис. 1.22), так это общения на естественном языке с роботом при обсуждении вариантов развития событий (обсуждения гипотез), отрицания роботом предложений *руководителя* по решениям (отрицания гипотез на основе фактов). Чистый Голливуд!

Но таков вектор развития.

О чем *никогда* не думали *аналитики* (сравните рис. 1.1 и рис. 1.22) — это то, что им придется обучать *робота*, который будет все более и более замещать их при общении с юридическим *руководителем*.

О чем не имели представления *поисковики* и *разведчики* (см. рис. 1.1) — это то, что роль *людей* в количественном отношении по этим функциям сходит на нет и в законной коммерческой деятельности ее совсем не видно.

Роль переместилась в состав ролей *креативной команды ситуационного центра*.

Для поиска и добычи информации не придется покидать помещение *ситуационного центра*, тем более ездить в другие города и страны.

Что же остается?

1. Целеполагание (объявление области интересов — старому заданию инфо-потребности) — пока остается за человеком: *руководителем*.

На какую тему будем копать в киберпространстве?

2. Разработка и реализация стратегии поиска и получения информации о реальности (бизнес-процессы П1 и П3 на рис. 1.1) остаются за людьми, интегрируются в одном процессе и концентрируются в *креативной команде*.

Поиск становится свободным (во всем киберпространстве) и «ИТ-еративным» (см. рис. 1.22).

То есть, как и следовало ожидать, вековые функции *руководителя*, *аналитика*, *разведчика* никуда не делись, просто во все

← Рис. 1.22
Современная схема
принятия решения

ВИРДЖИНИЯ РОМЕТТИ – IBM CEO

11 марта 2013 года

«В ближайшие 5 лет все компании на рынке
разделятся на победителей и побежденных
в зависимости от качества
корпоративных решений.

И опыт, и интуиция не помогут.

Людей надо переучивать.

Нигде этому сейчас не учат, но мы будем!»

ЧЕМУ
УЧИТЬ?

НОВОЙ КУЛЬТУРЕ ПРИНЯТИЯ РЕШЕНИЙ
(PREDICTIVE ANALYTICS),
обеспечивающей конкурентные преимущества



большей степени доля работы средств автоматизации по отношению к доле работ людей возрастает.

А традиционная схема принятия решений — это современная схема принятия решений минус компьютеры.

Частный случай, но все функции и роли традиционной схемы сохранены, только перекомпонованы в оргструктуре: назад в будущее.

1.4.3 Где то звено, за которое можно вытащить всю цепь?

Золотым фондом предприятия в эпоху цифровой революции является коллективный интеллектуальный потенциал креативной команды, реализующей совокупность функций (см. рис. 1.22).

Именно он в конечном счете определяет две базовые составляющие качества принимаемых решений: информированность и интеллектуальность решающего центра (см. Раздел 1) и в конечном счете конкурентоспособность предприятия.

Ничего нового: качество команды управленцев — проблема жизни и смерти фирмы.

Проблема в том, что эта команда управленцев перемещается в ситуационный центр и должна одинаково хорошо владеть как ИТ-технологиями, так и бизнес-технологиями.

Для CEO проблема: где брать таких разносторонних людей, а для CIO — чему и как их учить.

1.4.4 Стратегические угрозы XXI века в области ИТ и проблема подготовки кадров для работы в эпоху цифровой революции: что делать нам, чему учить, как учить? (Рис. 1.23)

С технологической точки зрения цифровая революция привела к возникновению в области ИТ новой реальности: обучаемых компьютеров (экспертных систем — ЭС) и новой специальности — учителей обучаемых компьютеров.

Обучаемые компьютеры становятся технологической основой поддержки принятия управленцами решений на уровне компаний, стран, министерств и ведомств и прочая и прочая.

А специальность учителя обучаемых компьютеров, игроков креативной команды (см. рис. 1.22) становится массовой и крайне востребованной специальностью: они (учителя) будут присутствовать в составе любых команд успешных управленцев.

Главным итогом возникновения обучаемых компьютеров является тот факт, что в ближайшем будущем качество корпоративных

Рис. 1.23

← Что делать нам,
чему учить,
как учить

решений не сможет и не будет опираться только на опыт и интуицию управленцев (как это было ранее в условиях реализации традиционной схемы принятия решений [5]).

Конкурентные преимущества в условиях цифровой революции будут достигаться с учетом прогнозирования последствий принимаемых управленцами решений (predictive analytics) [3].

А технологическая «гонка вооружений» в этой области (predictive analytics) будет идти и уже идет за достижение превосходства по основным характеристикам применяемых ЭС: информированности и интеллектуальности [1]

Согласно прогнозам McKinsey [7], predictive analytics (PA) как новая область ИТ-технологий в недалеком будущем станет и новой областью экономики, превосходящей по своей значимости нефтегазовый сектор. Нефтегазовому сектору уготовано 7–8-е место в объеме мировой экономики, в то время как сектору «Predictive Analytics» McKinsey сулит второе место. При том что, по оценке той же McKinsey, объем всего сектора Big Data, к которому относится и predictive analytics, в 2012 году был ничтожно мал.

Очевидно, что сектор PA в настоящее время является наиболее быстро растущим сектором мировой экономики — драйвером роста. С тем отличием, что «сырьем» для переработки здесь будут не нефть и газ, а огромный и быстро растущий объем мировых данных с необходимостью создания «заводов» по переработке «сырья» — ЭС.

Действительно, согласно [3] большие данные являются для каждой страны таким же национальным ресурсом, как нефть или газ. Но обладание ресурсом еще не делает страну сильной. Надо научиться этот ресурс перерабатывать и обращать в товар, востребованный на мировом и национальном рынках.

По аналогии с нефтегазовой отраслью речь идет о создании технологий по добыче этого ресурса, заводов по его переработке в финальную продукцию, логистике, наконец, (что заменит трубопроводы?) по доставке финальной продукции к точкам ее потребления.

В отличие от нефтегазовой отрасли обе технологии (добыча и переработка) в секторе экономики PA совмещаются в едином «заводе» — обучаемом компьютере (см. рис. 1.22) — экспертной системе (ЭС).

Очевидно, что товаром может быть как сам «завод» в целом (ЭС), так и его отдельные «станки» — модели и методы добычи, доставки и переработки «сырья» — больших данных. Товаром может

быть и продукция, выпускаемая «заводом», — информационные продукты (ИП), удовлетворяющие информационную потребность руководителей в схемах принятия решений (см. рис. 1.1 и рис. 1.22). А кадры, необходимые для создания и развития сектора РА, как и ранее (по аналогии с нефтегазовой отраслью), — это ученые и инженеры, обеспечивающие создание и развитие науки и инженерной практики «новой индустриализации»: организации производств нового поколения (информационных «заводов») и на их основе (заводов): информационных продуктов (ИП). Причем продукция (ИП), выпускаемая на «заводах», будет доставляться до точек потребления при помощи новой «логистики» — облачного доступа.

Налицо две угрозы любой национальной экономике.

Во-первых, потеря конкурентоспособности предприятий, для которых будут недоступны современные и будущие технологии принятия конкурентоспособных решений.

Проблема в том, что современные ЭС уже относятся к классу обучаемых систем, в которых персонал постоянно совершенствует приданные ему средства автоматизации в процессе эксплуатации (дообучает ЭС). И конкурентные свойства ЭС определяются совокупностью качеств персонала и ЭС.

Таким образом, остается открытым вопрос доступности персонала, способного дообучать ЭС в ходе эксплуатации для достижения конкурентных преимуществ.

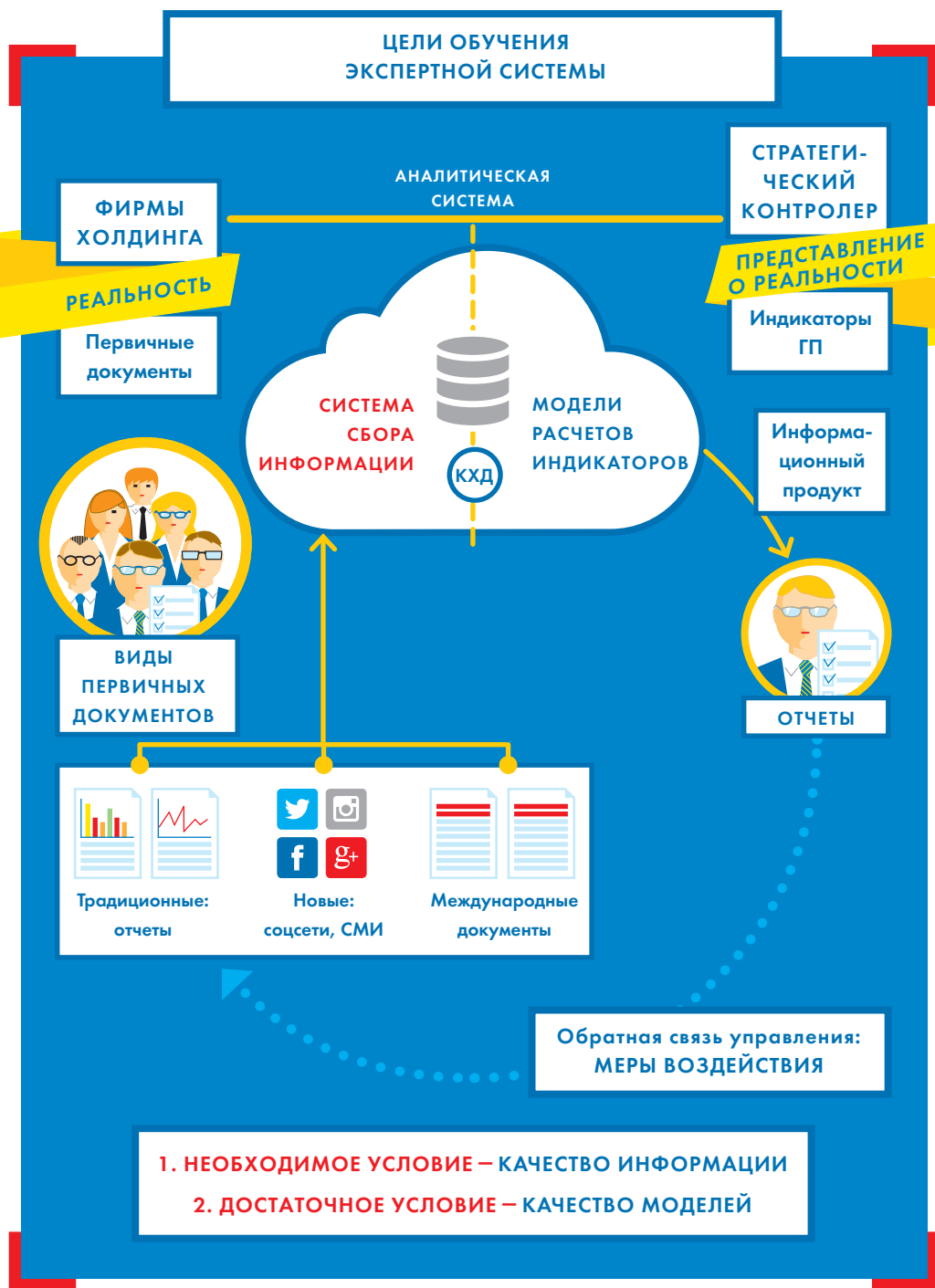
Вместе с тем, по мнению V. Rometty [3], в США (в 2013 году!) этому пока не учат, но собираются учить (predictive analytics).

Проблема обучения персонала для создания, эксплуатации и развития ЭС является международной.

Во-вторых, снижение конкурентоспособности экономики в целом в случае выпадения из мирового разделения труда в новом секторе инновационной экономики — «машиностроения» ЭС («заводов» по переработке «сырья» — больших данных).

Несмотря на новизну и необычность сектора создания нового класса нематериальных активов (обучаемых ЭС), сектора услуг по их обучению и, наконец, услуг по производству информационных продуктов, базовые законы экономики никто не отменял. Объемы производства и в этом секторе будут определяться как числом работающих, так и производительностью их труда.

Производительность же труда при создании ЭС (как и ранее в программировании) будет определяться наличием средств автоматизации разработки заключительного продукта из готовых крупных строительных блоков — программ (ранее



в программировании — языки высокого уровня, СУБД, мониторы транзакций — и т. д.).

При создании ЭС на основе коммерчески доступной платформы, например IBM Big Data, основные компоненты платформы позиционируются фирмой — изготовителем платформы именно как акселераторы (ускорители) разработки ЭС, то есть как средства радикального повышения производительности труда в новом «машиностроении» — производстве ЭС. И специалисты в области нового «машиностроения» должны обладать знаниями и навыками в области платформ Big Data, гарантирующими достижение высокой производительности труда при создании ЭС.

Итак, для удовлетворения потребностей инновационного развития экономики (например, в РФ), по экспертным оценкам, в ближайшие годы потребуются десятки тысяч специалистов в области создания и развития ЭС — нового поколения обучаемых компьютеров.

Необходимым условием (но не достаточным!) самой возможности организации обучения такого рода специалистов является наличие доступа при обучении к современным средствам проектирования ЭС — платформам IBM Big Data — букварю новой компьютерной грамотности, конструктору для первоначальной сборки ЭС.

Краткое описание состава конструктора платформы IBM Big Data содержится в главе 2.

Однако для ответа на вопрос, чему же еще надо учить специалистов в области создания и развития ЭС после первоначальной сборки ЭС из конструктора платформы, *следует вспомнить*, что коммерчески значимым информационным продуктом, производимым ЭС, является *решение* в той или иной области человеческой деятельности. И специалисты этих специальностей должны уметь работать с ЭС в целом, оценивать качество принимаемых ЭС решений и уметь повышать качество этих принимаемых решений в ходе эксплуатации и развития ЭС.

Иными словами — владеть знаниями и навыками по доводке функционала, реализуемого ЭС, до желаемого состояния: уметь обучать ЭС как по составу учитываемых аргументов предметной области (развитие информированности ЭС), так и по моделям принятия решений на основе собираемой в предметной области информации (развитие интеллектуальности).

Для конкретизации понимания деталей процесса обучения ЭС (чему учить) обратимся к простейшей модели принятия решений из [2], представленной на рис. 1.24. Обратим внимание

Рис. 1.24
Основные процессы
подготовки
РЕШЕНИЯ

на основные процессы, реализуемые в этой схеме: сбор данных и выделение информации, сохраняемой в корпоративном хранилище данных (КХД), методы и модели выработки решения и мер воздействия на ситуацию с последующим анализом реакции объекта управления на меры воздействия.

Из рис. 1.24 со всей очевидностью следует, что главной целью обучения ЭС является максимальное сближение реальности (содержащейся в источниках первичной информации об объекте анализа или управления) и представления руководителя о реальности (формируемого на основе ознакомления с информационными продуктами ЭС).

Только при достижении максимальной близости реальности и представления о ней у руководителя меры воздействия на объект управления (решение) приведут к желаемому результату.

Из соображений здравого смысла следует, что повлиять на качество решения в принципе можно двумя путями: расширяя состав и объем накапливаемой информации по предметной области всех видов и видоизменяя модели выработки решения по имеющейся информации.

По аналогии с адаптивными системами управления с обратными связями можно предположить, что наличие обратных связей как при поиске данных и выделении из данных информации, так и при принятии решений может положительно повлиять на качество работы ЭС.

На рис. 1.22 как раз и приводится одна из возможных современных схем принятия решений [2]. Она отличается наличием множественных обратных связей на всех этапах работы ЭС, в значительной степени изменяющих ролевые модели персонала АС, с возникновением новых требований к квалификации персонала, работающего в креативных командах обучаемых компьютеров. Очевидно, что требования к квалификации персонала креативной команды обучаемого компьютера определяются позициями специалистов в ролевой модели современной схемы принятия решений (см. рис. 1.22). Поэтому главными требованиями к знаниям и квалификации персонала креативных команд обучаемых компьютеров, сформулированных по-крупному, являются: умение работать со всей совокупностью обратных связей; повышение качества решений обучаемого компьютера за счет умения и знания учителя обучаемого компьютера.

По сути дела, старая специальность учителя, но в новой реальности: на основе новых технологий обучения и необычных учеников — диалектика.

Очевидно, что необходимой материальной основой обучения является доступ в процессе обучения как к технологиям обучения, так и к ученикам: к макетам обучаемых компьютеров, к тренажерам обучения.

С целью удовлетворения потребности доступа к такого рода технологиям фирмой «ЕС-лизинг» совместно с IBM в конце 2012 года создан первый в РФ Центр компетенции по технологии IBM Big Data — базовым технологиям создания ЭС Watson.

Технологической основой Центра компетенции является централизованная информационно-аналитическая система (ЦИАС) — платформа для одновременного исполнения многих учебных задач, тренажеров, прикладных программных пакетов ИАС, — созданная по проекту фирмы «ЕС-лизинг».

А назначением ЦИАС с точки зрения реализации процесса обучения является создание для будущих учителей обучаемых компьютеров класса их учеников: группировки виртуальных обучаемых компьютеров в различных областях бизнеса, практику по обучению которых и проходят будущие участники креативных команд.

Благодаря наличию ЦИАС, с точки зрения внешних пользователей, Центр компетенции выглядит так, как показано на рис. 1.25, — в виде мощного «пылесоса» разнообразных данных и их последующей обработки в интересах множества виртуальных обучаемых компьютеров: прикладных задач внешних пользователей, макетирующих их исполнение и обучающих свой персонал креативных команд.

По сути дела, на основе Центра компетенции IBM — «ЕС-лизинг» реализуется доказавшая свою эффективность модель инновационного развития: концентрации усилий вокруг решения поставленной задачи академической науки (ИПИ РАН), отраслевой науки («ЕС-лизинг»), вузовской науки (НИУ ВШЭ). И обучения студентов и аспирантов на основе практики реализации живых проектов. Само же обучение специалистов в настоящее время организуется на основе базовой кафедры «ЕС-лизинг» — «Информационно-аналитические системы», созданной в МИЭМ НИУ ВШЭ в 2013 году с предоставлением облачного доступа преподавателям и студентам к возможностям Центра.

Завершившие обучение по магистерской программе кафедры получают сертификат Академической программы IBM для работы в области IBM Big Data.

Первоначальное обучение основным технологиям платформы IBM Big Data реализуется на основе лабораторных работ,

СОЗДАНЫ «ПЫЛЕСОС» ДАННЫХ
И ИНФРАСТРУКТУРА ОБРАБОТКИ
В ИНТЕРЕСАХ МНОЖЕСТВА ЗАДАЧ
(«МОТОР» ЦЕНТРА КОМПЕТЕНЦИИ BIG DATA)

ИСТОЧНИКИ ИНФОРМАЦИИ:

- международные;
- федеральные;
- региональные;
- муниципальные;
- хозяйствующие субъекты

Макет — стартовая точка
создания любого бизнес-проекта

ЕС-ЛИЗИНГ

МАКЕТ

Группировка
бизнес-приложений,
тренажеров,
учебных программ



ИАП

ЦОД



Кафедра ИАС
МИЭМ НИУ ВШЭ



ИПИ РАН

ОБЛАЧНЫЙ
ДОСТУП



КЛИЕНТЫ:
юридические
и физические
лица

разработанных IBM для MIT и адаптированных магистрантами кафедры под возможности Центра компетенции. Состав лабораторных работ и инструкции по их использованию в процессе обучения содержатся в части II.

Особенностью обучения на базовой кафедре ИАС является непрерывное участие студентов под руководством преподавателей в практическом проектировании ИАС, реализуемых для заказчиков «ЕС-лизинг».

Уже решенные задачи и возникшие в результате их решения ИАС становятся все расширяющейся основой для обучения новых поколений студентов. Расширяющимся составом учеников в классе обучаемых компьютеров.

Примеры уже решенных задач, доступных для обучения в ЦИАС, представлены в главе 3.

Центр компетенции IBM Big Data — «ЕС-лизинг» является материально-технической основой для проведения НИР и ОКР в области создания современных информационно-аналитических систем, в том числе ЭС.

Для заинтересованных организаций и физических лиц Центром компетенции IBM — «ЕС-лизинг» оказывается целый спектр услуг по оказанию помощи в освоении технологий Big Data и организации обучения персонала по макетированию аналитических систем для выбора направлений их развития.

В части II содержится описание правил доступа к Центру компетенций со стороны внешних пользователей.

Очевидны выводы из изложенного.

1. В конкурентной борьбе как за качество принимаемых корпоративных решений, так и за долю рынка в новой экономике ключевую роль играет производительность труда проектировщиков ЭС. Целевая производительность труда может быть достигнута только с применением платформ типа IBM Big Data — крупноблочных конструкторов для строительства ЭС.
2. Применение технологий IBM Big Data приводит к появлению обучаемых ЭС, отличающихся от традиционных ЭС схемой принятия решений и наличием возможности обучения ЭС в ходе эксплуатации, с соответствующим изменением требований к ролям и квалификации персонала.
3. Подготовка специалистов для крупноблочного проектирования и последующего обучения ЭС требует доступа будущих учителей обучаемых компьютеров к своим ученикам:

Рис. 1.25
Инфраструктура
обработки в интересах
множества задач

тренажерам обучения ЭС, виртуальным обучаемым компьютерам.

4. Группировка виртуальных обучаемых компьютеров (класс учеников) для различных областей бизнеса предоставляется ЦИАС, использующей технологические возможности Центра компетенции Big Data (IBM – ЗАО «ЕС-лизинг»).

Научно-практический сборник

А. В. Шмид
Б. А. Позин
И. В. Галахов
М. А. Агейкин
Д. О. Александров
М. Р. Касимов
Н. И. Клемашев
Г. А. Ежов

**Новые методы работы
с Большими Данными:
победные стратегии управления
в бизнес-аналитике**

Ответственный редактор

А. В. Шмид

Редактор

Т. И. Якушкина

Дизайн

Ю. В. Потресова

Компьютерная верстка

Ю. В. Потресова

Корректор

Т. И. Якушкина



Подписано в печать 15.12.15. Формат 70 × 90/16

Усл.-печ. л. 38,61.

Издательство ПАЛЬМИР

Лицензия ИД № 00795 от 20.01.2000

potressov@gmail.com

Отпечатано в типографии

Книга знакомит с новейшими технологиями обработки Больших данных (Big Data) на примере платформы BIG DATA IBM (базовые технологии экспертной системы Watson IBM).

Переломным в информатике был 2007 год: согласно отчету International Data Corporation (IDC), весь мир впервые за тот год произвел объем данных, превышающий возможности всех существующих сред хранения информации: человечество вступило в эпоху Больших данных. Мы оказались свидетелями рождения новой экономики, создания невиданных технологических операций с «информационным сырьем», систем поддержки принятия решений — своего рода «заводов» по его переработке. Это направление определяется в мире термином Predictive analytics — Предсказательная аналитика.

В книге обобщен опыт ЗАО ЕС-лизинг в проектировании систем по технологии BIG DATA IBM, содержатся практические лабораторные работы — своего рода «букварь» по обучению действиям с Большими данными, — созданные корпорацией IBM для MIT (Массачусетского технологического института), однако адаптированные преподавателями и магистрантами базовой кафедры ЗАО ЕС-лизинг «Информационно-аналитические системы» в МИЭМ НИУ ВШЭ для учебных курсов. Книга полезна также широкому кругу читателей, интересующихся проблемами современных информационных технологий.



9 785938 110267