УДК 004.62

СОВРЕМЕННЫЕ МЕТОДЫ И ТЕХНОЛОГИИ ПРЕОБРАЗОВАНИЯ БОЛЬШИХ ДАННЫХ

Н.Б. Толпинская, А. Сычев

Появление феномена больших данных стало закономерным последствием роста компьютерных технологий. На основе определения данных, информации и знаний показан процесс получения последних. Рассмотрена цикличность получения знаний, начиная со сбора и хранения данных, их преобразования в информацию и заканчивая аналитической обработкой. Большие данные объединяют техники и технологии, которые экстремально практично извлекают смысл в рамках поставленной задачи. Представлен хронологический обзор соответствующих технологий, показаны основные тенденции развития больших данных, инструменты для реализации и основные сложности прикладного применения. Раскрывается понятие больших данных как работы с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках в целях увеличения эффективности деятельности, создания новых продуктов и повышения конкурентоспособности. Подчеркнута актуальность развития технологий больших данных.

Ключевые слова: Big Data; знания; информация; Hadoop; Data Mining; Business Intelligence; анализ; обработка; информационные технологии; систематизация.

ЧОҢ МААЛЫМАТТАРДЫ КАЙРА ТҮЗҮҮНҮН ЗАМАНБАП МЕТОДДОРУ ЖАНА ТЕХНОЛОГИЯЛАРЫ

Н.Б. Толпинская, А. Сычев

Чоң маалыматтар феномени компьютердик технологиялардын өнүгүүсүнүн мыйзам ченемдүү көрсөткүчү болуп калды. Маалыматтарды, информацияларды жана билимдерди аныктоонун негизинде чоң маалыматтарды алуу процесси көрсөтүлдү. Макалада билим алуунун циклдүүлүгү маалыматтарды чогултуу жана сактоо, аларды информацияга айлантуудан тартып аналитикалык жактан иштеп чыгууга чейин каралды. Чоң маалыматтар техникаларды жана технологияларды бириктирет, алар коюлган милдеттердин алкагында маалыматтын маңызын экстремалдуу пайдалуу алып чыгат. Тиешелүү технологияларга хронологиялык сереп салуу жүргүзүлдү, чоң маалыматтардын өнүгүү тенденциялары, ишке ашыруу үчүн аспаптар жана аларды колдонуудагы негизги кыйынчылыктар көрсөтүлдү. Чоң маалыматтар түшүнүгү чоң көлөмдөгү жана ар кыл курамдагы маалыматтар менен иштөө катары ачылып берилди, ишмердүүлүктүн натыйжалуулугун жогорулатуу, жаңы продукттарды түзүү жана атаандаштыкка туруктуулукту жогорулатуу максатында маалыматтар бат-бат жаңыртылып турат жана ар түрдүү булактардан алынат. Макалада чоң маалыматтар технологиясын өнүктүрүүнүн актуалдуулугу баса белгиленген.

Түйүндүү сөздөр: Big Data; билимдер; маалыматтар; Hadoop; Data Mining; Business Intelligence; талдоо жүргүзүү; иштеп чыгуу; маалымат технологиялары; системага салуу.

MODERN METHODS AND TECHNOLOGIES FOR BIG DATA CONVERSION

N.B. Tolpinskaya, A. Sychev

The emergence of the phenomenon of big data has become a logical consequence of the growth of computer technology. Based on the definition of data, information and knowledge shows the process of obtaining the latter. The article discusses the cyclical nature of obtaining knowledge, starting with the collection and storage of data, their conversion into information and ending with analytical processing. Big data combines technology and technology that is extremely practical to extract meaning within the framework of the task. A chronological review of relevant technologies is presented, the main trends in the development of big data, tools for implementation and the main difficulties of applied applications are shown. The concept of big data is revealed as work with information of huge volume and diverse composition, which is very often updated and located in different sources in order to increase business efficiency, create new products and increase competitiveness. The relevance of the development of big data technologies is underlined.

Keywords: Big Data; knowledge; information; Hadoop; Data Mining; Business Intelligence; analysis; processing; information technology; systematization.

Введение. Почти за семидесятилетнюю историю компьютерных технологий так и не появилось однозначно четкого определения данных, информации, знаний и того, как первые связаны с результатами обработки. Зачастую их используют как синонимы, однако между ними существует принципиальная разница. Это связано с невероятными темпами развития технологии работы с данными, с одной стороны, и практически не развивающейся теории информации, не изменившейся с 50-х годов прошлого века, с другой.

В приведенном обзоре отражены технологии и инструменты преобразования данных в знания.

Данные несут в себе сведения о событиях, произошедших в материальном мире, они являются регистрацией сигналов, возникших в результате этих событий [1]. Это формализованный способ фиксирования в дискретной форме объективных фактов и характеристик. Когда данные структурированы, упорядочены, сгруппированы и категоризированы, они становятся содержательной информацией.

Информация – это совокупность данных, упорядоченная с определенной целью, придающей им смысл.

Существуют и другие формулировки информации, однако необходимо отметить, что универсальных определений этих понятий нет, и в зависимости от области деятельности, они объясняются по-разному.

Знание трактуется как информация, снабженная смыслом, готовая к продуктивному применению. Знание представляет собой совокупность оформленного опыта, ценностей, контекстуальной информации, экспертного понимания, составляющих основу для оценки и интеграции, нового опыта и информации [2].

На основании новой информации, после аналитической обработки можно получить знания, необходимые для принятия управленческих решений.

Принятие решений – это выбор наилучшего в некотором смысле варианта решения из множества допустимых на основании имеющейся информации. На рисунке 1 представлен полный цикл получения знаний.

Для решения поставленной задачи выбираются необходимые данные, которые обрабатываются и преобразуются в информацию. Полученная информация анализируется, и в результате анализа получаются новые знания о предметной области задачи. Эти знания преобразуются в допустимые варианты решения задачи, а в результате принятия решения обычно принимается одно наилучшее в рамках поставленной задачи.

Таким образом, процесс извлечения знаний из данных состоит из трех этапов. Первый этап – сбор и хранение данных. Второй – их обработка и преобразование в содержательную информацию. Третий – получение знаний, необходимых для принятия управленческих решений.

Сбор и хранение данных. Сбор данных может производиться человеком или автоматически. Среди аналитиков существует твердо устоявшееся мнение, что лишних данных не бывает, нужно собирать все доступное, т. к. неизвестно, что понадобится для моделирования поведения различных показателей. С этим сложно не согласиться, однако стоит отметить, что сведения должны быть так или иначе привязаны к тематике организации. Исходя из области деятельности и поставленных задач, перед началом сбора следует определить набор интересующих критериев и источники интересующей информации. Технология сбора определяется на основании необходимого набора показателей, применимых

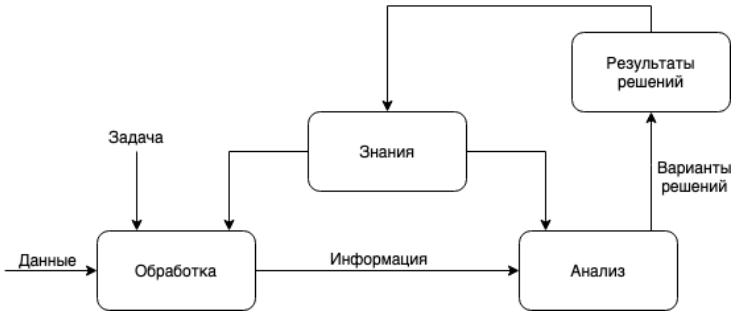

Рисунок 1 – Цикл получения знаний

к ним методам обработки и техническим средствам. Собранные данные приводятся к формализованному виду, пригодному для последующего хранения и обработки. Как правило, они сохраняются в виде таблиц баз данных, XML-файлов, Excel- таблиц и т. п.

В настоящее время стремительно развиваются интернет-технологии, количество данных растет в геометрической прогрессии, что вызывает проблемы их хранения и последующей обработки. Данные, собранные из соцсетей, геосистем и прочих интернет-источников, представляют практический интерес для компаний различных сфер деятельности (торговля, кредитование, здравоохранение), поскольку позволяют извлечь персонализированную информацию и сделать обслуживание клиентов более индивидуальным.

Последнее десятилетие речь идет не просто о данных, а о Больших данных (Big Data) и далее речь пойдет именно о них, поскольку эта проблема стала первоочередной на текущий момент.

Итак, после того как данные собраны и сохранены в различных файлах, начинается извлечение из них содержательной информации.

От данных к информации. Информация – объект динамический, который меняется при изменении первоначальных сведений и актуален только на момент их взаимодействия с методами обработки. При отсутствии источника данных и непосредственного получателя, процесс их перевода в определенную информацию не имеет смысла. Можно выделить следующие методы преобразования: статистический анализ, контекстуализация, разбиение на категории, отбор. После преобразования данных мы получаем информационный массив с определенной степенью снижения их первоначального объема. Причем этот процесс можно повторить несколько раз и получить более содержательную информацию.

Само понятие Big Data включает в себя не просто огромные массивы данных, а совокупность технологий их обработки для получения информации. Результаты обработки данных необходимо сохранить для проведения последующего анализа. Отобранные данные сохраняются в специализированных хранилищах данных.

Хранилище данных – предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки бизнес-анализа и формирования отчётов для выбора оптимальных управленческих и стратегических решений организации. Они строятся на базе систем управления базами данных и систем поддержки принятия решений [3].

Для извлечения данных из различных источников, их преобразования и очистки с последующей записью в хранилище данных, используют процессы ETL (Extract Transform Load). В настоящее время ETL-системы все более широко применяются для консолидации данных с целью их дальнейшего анализа. Наиболее распространенными разработчиками ETL-инструментов можно назвать Oracle, Informatica и IBM [4].

Так IBM предлагает два решения: Data Stage и Data Manager, в который встроено OLAP, что упрощает структуру хранилища данных. Пользователи Unix и Linux могут использовать решения PowerCenter и PowerMart от Informatica и Oracle Data Integrator от Oracle, которым характерна возможность масштабирования. Для пользователей Microsoft предлагается SSIS решение, позволяющее создавать пользовательские компоненты.

Технология хранилищ данных возникла в начале 80-х годов прошлого века. В основе концепции хранилища данных лежат две основные идеи – интеграция разъединенных детализированных данных (описывают некоторые конкретные факты, свойства, события и т. д.) в едином хранилище и разделение наборов данных и приложений, используемых для оперативной обработки и применяемых для решения задач анализа. Выделяют 5 основных архитектур хранилищ данных, которые можно расположить в порядке наиболее часто применяемых следующим образом: доминирующей архитектурой является звезда, за ней следует шина витрин и централизованная архитектура, и лишь небольшой процент проектов основан на независимых витринах и федеративной архитектуре.

На сегодняшний день существуют два основных подхода к архитектуре хранилищ данных. Это так называемая корпоративная информационная фабрика (Corporate Information Factory, сокр. CIF) Билла Инмона и хранилище данных с архитектурой шины (Data Warehouse Bus, сокр. BUS) Ральфа Кимболла (Ralph Kimball) [5].

Наибольшее распространение получил подход Инмона, в его модели атомарные данные организованы в реляционные базы и находятся в нормализованном хранилище данных, причем суммарные данные доступны для использования через специализированные хранилища, средства data mining и OLAP; что же касается зависимых витрин данных, то только они организованы с помощью пространственных моделей, как и у Р. Кимболла [6]. Однако модель BUS более проста в понимании и более эффективна в доступе к данным (архитектура хранилища реализована по схеме “звезда” или “снежинка”, а связь таблиц данных и измерений по схеме “шина”), особенно при сложных анализах, но требует организации сложных процедур подготовки, загрузки данных и управления изменениями данных.

С развитием технологий Big Data, взгляд на организацию архитектуры хранилищ данных изменился. Уже в начале века начались разработки специализированных баз данных, так как универсальные системы управления, построенные на безразмерной архитектуре, не учитывающие аппаратные изменения платформ, не могли реализовать оптимальные решения для некоторых классов приложений, требующих определенной специфики. Для хранения используют объектные хранилища, позволяющие работать с растущими объемами данных. Использование плоского адресного пространства дает возможность быстро найти нужные данные, как на локальном, так и облачном сервере. Помимо этого, они оснащены внутренним механизмом проверки корректности файлов и другими функциями, обеспечивающими доступность данных.

С ростом влияния интернета на жизнь людей встала новая проблема – обработка быстрорастущих данных в реальном времени. Эта информация представляет интерес для научных исследований в области развития web-технологий, которые все больше развиваются в направлении полной индивидуализации под конкретных пользователей и их паттерны поведения. Это, в свою очередь, требует обработки, хранения, анализа и доступности огромного количества информации в режиме реального времени. Не следует забывать о бизнесе и государственном секторе, где получение своевременных данных увеличивает доходы бизнеса и оптимизирует управленческие решения. Разработчикам пришлось по-новому взглянуть на потоковую передачу данных.

Так в 2004 г. была предложена модель распределенной обработки MapReduce [7] для больших объемов данных и распределенная файловая система Google File System. На этой основе была разработана платформа Hadoop [8], архитектура которой состоит из трех уровней. Первый использовался для хранения данных – это распределенная файловая система Hadoop Distributed File System (HDFS). Второй слой представляет собой фреймворк обработки данных MapReduce, реализующий одноименную модель группировки информации. На практике оказалось сложным писать нужные запросы, поэтому стала развиваться целая отрасль прикладных программ для решения конкретных задач, это третий уровень. К таким инструментам относятся: Hive – для работы с данными, расположенными в первом слое при помощи SQL-подобного языка; Pig – позволяющий описывать более сложные процессы трансформации данных при помощи функционального языка PigLatin; Mahout, Cascading, Scalding и др. поддерживающие пакетную обработку данных.

Но мир не стоит на месте, и обрабатывать просто большие данные оказалось недостаточным, их следовало обрабатывать быстро, так, например, препроцессинг веб-логов перед помещением их в кластер. В связи с этим стали появляться новые инструменты: Apache Storm [9], Apache Samza [10], Apache Flink [11], работающие с последовательно поступающими данными (естественная потоковая обработка) и Apache Storm/Trident и Apache Spark [12], поддерживающими пакетную обработку. Каждый из них прекрасно решает свой круг задач, однако ввиду отсутствия единого менеджера ресурсов, не было возможности создавать кластеры, предназначенные для различных рабочих нагрузок.

Такой менеджер появился в концепции Hadoop 2.0. Им стал фреймворк Yet Another Resource Negotiator (YARN) [13], работа которого заключалась в построении очереди выполняемых в кластере задач, выделении ресурсов выполняемым задачам, мониторинге процесса и перераспределении ресурсов в случае ошибок при их выполнении. Модель MapReduce позволяет выполнять обработку больших массивов

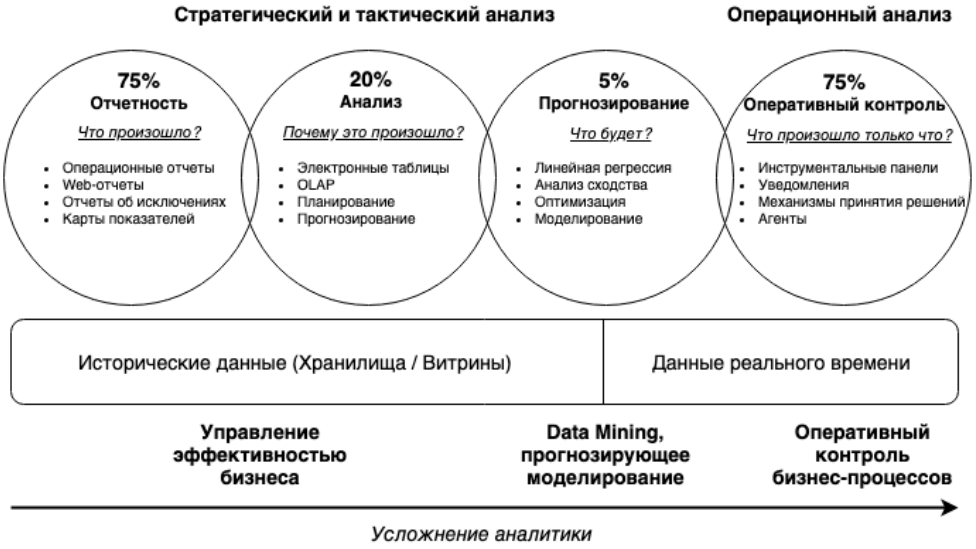

Рисунок 2 – Многообразие аналитических средств

данных на кластерах, собранных из доступных серверов. Однако еще полвека назад, в 1963 г. Мелвин Конвей предложил модель Fork/Join, которая, в отличие от MapReduce, лучше приспособлена не к кластерам, а к многоядерным процессорам. Так, уже в Hadoop 2.0. реализованы две новые модели параллельных вычислений Apache Tez и Apache Spark. Последний имеет собственную надстройку SparkSQL для работы с данными при помощи SQL-запросов.

Таким образом, процесс извлечения информации представляет собой последовательное выполнение определенных шагов с применением подходящих для выбранных данных, методов и технологий обработки.

От информации к знанию. Получение знаний – завершающий шаг цикла. Здесь речь пойдет об аналитическом анализе данных, помещенных в хранилище.

Для превращения информации в знание человеку необходимо творчески её переработать. В переработке информации задействованы логические процессы мозга, а также правила и общественные связи. Если процесс извлечения знаний переложить на компьютер, то можно говорить об аналитической обработке специализированной информации с применением соответствующих методов (сюда относят методы статистического анализа) и моделей обработки (потоковая модель, модель обработки в реальном времени, модель MapReduce и т. д.).

За последние 10 лет в сфере аналитического анализа данных произошли большие изменения: от реализации отдельных методов анализа, до встроенных платформ быстрой разработки специализированных аналитических приложений. На рисунке 2 показаны основные инструменты, используемые для проведения как стратегического, так и операционного анализа.

Технологии бизнес-анализа оперируют историческими данными, а Big Data используют данные реального времени. Бизнес-анализ является описательным процессом результатов, достигнутых в определенный период времени, между тем как скорость обработки больших данных предоставляет возможность сделать анализ предсказательным, способным предлагать достоверные рекомендации на будущее. Технологии Big Data позволяют анализировать большее количество их типов, по сравнению с инструментами бизнес-аналитики, что дает возможность фокусироваться не только на структурированных хранилищах [14].

Открывшиеся возможности анализа в реальном времени применительно к Big Data вывели технологию больших данных на новый уровень развития. Это стало возможным за счет использования технологий аналитики в памяти (in-memory), таких как: DSSD, Apache Spark, GemFire. Уже разработаны готовые решения in-memory системы управления – данные изначально хранятся в оперативной памяти,

и доступ к ним практически не занимает времени, пример такого решения – SAP HANA. Эта система, являясь системой управления базами данных, предоставляет доступ к памяти любого BI-инструмента: можно загрузить данные из таблиц Excel, систем Cognos, Oracle BI и других.

Заключение. На сегодняшний день информация является наиболее ценным продуктом деятельности человечества. На службе производства информации находятся множество современных научных разработок в области информационных технологий. При этом существует реальный дефицит знаний. В эпоху динамичности экономических процессов владение знаниями, умение управлять потоками информации, является серьезным конкурентным преимуществом организаций, повышает их эффективность, увеличивает скорость реагирования на внешние изменения. Это делает процесс получения знаний крайне актуальным.

В последнее десятилетие наблюдается катастрофический рост объема данных и существующие способы их хранения и обработки не справляются с этой задачей. В связи с этим, подход к преобразованию стал меняться, что повлекло за собой очередной виток развития информационных технологий, а точнее, технологий работы с большими данными.

Для многих организаций, имеющих огромные запасы данных, процесс получения из них новых знаний стоит на первом месте, поскольку это может принести экономический рост. А если рассматривать государственные организации, то наличие знаний – это процветание государства, причем по всем направлениям.

Современное сообщество имеет колоссальные технические возможности работы с данными, которые, к сожалению, опережают уровень развития их прикладного использования. Big Data открывает огромные возможности планирования как в отдельных отраслях экономики, здравоохранения, образовании, так и для государства, как для организации в целом. Развивая технологии Big Data можно поднять значение информации, как фактора производства, на совершенно новый качественный уровень. Информация станет не только равноценна труду и капиталу, но и возможно станет наиважнейшим ресурсом современной экономики [15].

Литература

1. *Хорошилов А.В.* и др. Информационные системы в экономике / А.В. Хорошилов и др. М.: МЭСИ, 1998.
2. Чем отличаются данные от информации. Понятие информации. Информация и данные. Отличия понятий информации и данных. Информация – структурированные данные. URL: http://cdd-evo.ru/chem-otlichayutsya-dannye-ot-informacii-ponyatie-informacii-informaciya-i-dannye/ Joerg Reinschmidt, Allison Francoise. Business Intelligence Certification Guide. IBM Red books;
3. Хранилище данных. URL: https://ru.wikipedia.org/wiki/ Хранилище_данных
4. *Коновалов М.В.* ETL: обзор и роль в развитии компаний [Текст] // Технические науки в России и за рубежом: матер. VII межд. науч. конф. (г. Москва, ноябрь 2017 г.) / М.В. Коновалов. М.: Буки-Веди, 2017. С. 31–34. URL B/ (дата обращения: 09.11.2018).
5. Основные подходы к архитектуре Хранилищ данных. URL: http://www.interface.ru/home.asp?artId=537
6. OLAP и многомерные базы данных. URL: https://www.monographies.ru/ru/book/section?id=4638
7. MapReduce. URL: https://ru.wikipedia.org/wiki/MapReduce
8. Hadoop. URL: https://hadoop.apache.org
9. Apache Storm. URL: https://storm.apache.org/.
10. Apache Samza. URL: http://samza.apache.org/.
11. Apache Flink: Scalable Batch and Stream Data Processing. URL: https://flink.apache.org/.
12. Apache Spark – Lightning-Fast Cluster Computing. URL: https://spark.apache.org/.
13. Использование Hadoop YARN. URL: https://www.ibm.com/developerworks/ru/library/bd-hadoopyarn/index.html
14. Большие данные vs бизнес-аналитика. URL: http://datareview.info/article/bolshie-dannyie-vs-biznes-analitika/
15. *Веретенников А.В.* BigData: анализ больших данных сегодня / А.В. Веретенников // Молодой ученый. 2017. № 32. С. 9–12. URL: https://moluch.ru/archive/166/45354/ (дата обращения: 06.04.2019).