Наука и Образование
МГТУ им. Н.Э. Баумана
Сетевое научное издание
ISSN 1994-0408

Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2015. № 03. С. 207–240.

DOI: **10.7463/0315.0761354**

Представлена в редакцию: 18.02.2015

УДК [004.6+004.272.42+004.4]::001.5:[519.6+519.7+519.8]

Влияние инструментария Big Data на развитие научных дисциплин, связанных с моделированием

Сухобоков А. А.[1], **Лахвич Д. С.**[1,*]

*dmitry.lakhvich@optimalmngmnt.com

[1]МГТУ им. Н.Э. Баумана, Москва, Россия

Due to the expansion of use of Big Data solutions the conception of Big Data is considered and examples of the tasks demanding use of the Big Data tools are presented. The overview of Big Data tools is made, including Open Source products and proprietary tools developed by largest IT companies. Examples of using of Big Data tools in enterprise data processing and management systems are shown. Within overview, a classification of the Big Data tools is proposed that fills gaps of previously developed such classifications. Made overview allowed drawing conclusion that besides the solving of practical tasks the expansion of Big Data tools leads to uprise of new schools in scientific disciplines associated with simulation of technical, natural, socio-economic systems and the decision of practical problems basing on developed models. Made conclusion is illustrated by examples of following disciplines: numerical methods, imitational modeling, management of socio-economic systems and optimal control theory.

Ключевые слова: инструменты Big Data, кластеры Hadoop, научные дисциплины, модели с большим числом элементов

1. Понятие Big Data

В соответствии с современными представлениями Big Data представляет собой данные большого объёма, для которых характерны пять V – пять характеристик, начинающиеся на V: Volume - объём, Variety - разнообразие, Velocity – скорость, Veracity – достоверность и Value – ценность [1].

Volume – объём. Объём данных считается большим, когда возникают затруднения при обработке этого объёма средствами традиционных СУБД. При возникновении концепции Big Data таким объёмом считался 1 PB ($1*10^{15}$байт). С развитием процессорных технологий и технологий СУБД эта цифра может вырасти, однако рост не происходит быстро из-за отсутствия качественных изменений, обусловленных технологическими инновациями.

Внутренней причиной перехода к новым технологиям обработки данных является необходимость распараллелить обработку, распределить её на большое число независимых процессоров, каждый из которых обрабатывает свой фрагмент данных.

Распараллеливание – наиболее естественный способ преодоления сложностей вызванных большим объёмом данных. После первых проектов по обработке больших данных стало понятно, что объёмные дисковые стойки NAS-систем в состоянии обеспечить хранение большого объёма данных, но бутылочное горлышко интерфейса не в состоянии обеспечить пропускную способность, необходимую для обслуживания сотен параллельно работающих процессоров. Именно это понимание и дало толчок появлению инструментов Big Data.

Variety – разнообразие. Данные такого объёма очень редко бывают однородными. В подавляющем большинстве случаев общий массив данных включает как структурированные, так и неструктурированные данные. Под неструктурированными данными имеются в виду изображения, аудио-треки, фильмы и видео-ролики, данные социальных сетей. Пропорции структурированных и неструктурированных данных в разных массивах могут быть самыми разными, например от 1:9 до 9:1.

Velocity – скорость. Скорость трактуется не только как скорость прироста, но и как скорость обновления ранее полученных значений, что неизбежно влечёт за собой необходимость высокоскоростной обработки и получения результатов. В пределе – в реальном времени.

Veracity – достоверность. В условиях работы с большими объемами данных особое значение приобретает отделение достоверных данных от информационного «шума» и мусора, отсеивание этого шума и мусора.

Value – ценность. Именно ценность информации предопределяет целесообразность её обработки. Собираемые данные должны давать ответы на предварительно сформулированные и вновь появляющиеся вопросы. Эффекты, получаемые в результате сбора и обработки данных, должны оправдывать затраты на эти операции. Собираемые данные должны приносить пользу.

Перечисленный перечень ключевых характеристик Big Data появился не сразу. Сначала были сформулированы первые три – Volume, Variety и Velocity. Потом по одной последовательно добавились Veracity и Value.

2. Задачи Big Data

Приведём типичные области, в которых возникают данные, которые можно охарактеризовать как Big Data, и необходимо применять соответствующие инструменты:

- **Финансовые услуги.** Ожидается, что к 2020 году число ежедневных транзакций электронной коммерции и финансовых операций физических лиц увеличится до $450*10^9$ [2]. С учетом объёмов данных каждой транзакции и сроков их хранения для архивирования таких финансовых данных предпочтительно использовать инструменты Big Data. Помимо организации хранения больших объёмов данных они позволяют проводить сегментацию пользователей, подключать данные пользователей из социальных сетей и на основе выявленных корреляций формировать для пользователей индивидуальные предложения, лучше учитывающие их потребности. Кроме того, применение инструментов Big Data позволяет выявлять мошенников и предотвращать потери [3].

- **Производственные процессы в промышленности.** Данные поступают с сенсоров и используются для управления технологическими процессами, контроля объёма и качества выпускаемой продукции. Например, в компании Chevron с нефтяных скважин в сутки поступает 3000 PB данных [4]. Это в три раза больше, чем за сутки генерирует весь интернет. Использование этих данных позволило дополнительно ежегодно получать 15 млрд. долларов прибыли.
- **Здравоохранение.** The Institute for Health Technology Transformation показал, что человеческое тело представляет собой неиссякаемый источник больших данных [5]. Объём архивов изображений в медицине ежегодно возрастает на 20-40%:
  - объём данных одного снимка трёхмерной рентгеновской компьютерной томографии (3D CT Scan) составляет примерно 1GB;
  - объём данных одного снимка трёхмерной магнитно-резонансной компьютерной томографии (3D MRI) составляет примерно 150MB;
  - объём данных одного рентгеновского снимка составляет примерно 30MB;
  - объём данных одной маммограммы составляет примерно 120MB.

  Также наблюдается отчётливая тенденция быстрого роста числа носимых (wearable) устройств, которые находятся на теле пациентов и снимают информацию в реальном времени. Ожидается, что к 2018 году в мире будет использоваться 500 миллионов таких устройств [4].
- **Эксплуатация и обслуживание сложного оборудования.** Например, внутренние системы одного современного самолёта ежедневно порождают 1 TB данных [2]. Основное назначение собираемых данных – контроль состояния оборудования, планирование технического обслуживания и ремонтов для поддержания необходимого уровня его надёжности.
- **Сельское хозяйство.** Использование методов и инструментов Big Data для анализа цепочек ДНК отдельных растений и животных позволит радикально сократить время выведения новых сортов и пород с заданными свойствами, ранее занимавшее 10 лет и более. Выращивание сельхоз продукции на полях, оснащённых большим числом датчиков, в реальном времени передающих информацию об уровне влажности, освещённости, температуре, наличии питательных веществ в почве и пр. позволит управлять процессами выращивания урожая. Ожидается, что сочетание высокоэффективных сортов, оптимальной ирригации, правильной дозировки пестицидов, гербицидов и удобрений позволит к 2050 году повысить урожайность зерновых до 250% [6]. Объёмы данных, которые придётся при этом собирать и обрабатывать, характерны для Big Data.
- **Сложные логистические процессы.** Данные о размещении каждой упаковки товара на складах, данные об отгрузках и поступлении товаров имеют объёмы, измеряемые терабайтами, и в большинстве случаев могут быть обработаны SCM-системами, которые релевантны масштабам цепочки поставок. Необходимость использования инструментов Big Data в логистических сетях крупных компаний, военных и правительственных организаций возникла после перехода к современным технологиям, реализующим сбор и обработку данных с меток RFID,

установленных на каждой транспортной упаковке, а также сбор, хранение и обработка данных геолокации о каждом транспортном средстве.

- **Информационные технологии.** Более 10 миллиардов новых сообщений Facebook [7] и 500 миллионов новых твитов [8] появляется ежедневно. С ещё большей скоростью появляются новые записи в журналах соответствующих баз данных. Ещё один пример: Yahoo использует хранилище данных на базе Hadoop объёмом 255 PB, которое используется для поддержки более семисот миллионов пользователей [9].
- **Розничная торговля.** Обработка всей совокупности данных об истории продаж, объёмах запасов, ценах, а также других дополнительных данных, например, о постоянных клиентах, имеющих дисконтные карты, о конкурентах и т.д. позволяет понять факторы, влияющие на объёмы продаж, сформировать конкурентные цены и проводить эффективные маркетинговые компании [10].
- **Телекоммуникации.** В мире уже произведено больше мобильных телефонов, чем имеется людей на земле. 5 миллиардов мобильных телефонов находится в пользовании [2]. Накопление данных об оказанных клиентам услугах (звонках, SMS, передаче информации), а также последующая аналитическая обработка этих данных позволяет идентифицировать поведение пользователей и более точно определять их потребности. На основе этого можно оптимизировать инфраструктуру и сокращать затраты на развитие сети, с меньшими затратами и более полно удовлетворять потребности клиентов [11]. В случае требований гос. органов протоколировать и хранить определенное время все голосовые разговоры, SMS и трафик, телекоммуникационные компании не смогут решить эти задачи без применения инструментов Big Data.
- **Коммунальное хозяйство (электро-, водо-, тепло-, газоснабжение).** В каждом крупном городе в жилом секторе и на предприятиях установлены миллионы счётчиков, с которых регулярно собираются показания, счётчики подлежат учёту, периодически проводится их поверка и замена. Использование «умных» счётчиков, позволяющих регулярно регистрировать и передавать данные по сети, в сочетании с последующей обработкой собираемых данных позволяет улучшить качество обслуживания. В сервисный центр сразу поступают данные об отсутствии подачи электроэнергии, воды и т.п. [6]. Кроме того, расширение объёма передаваемых данных, например добавление сведений о поддерживаемой температуре, позволяет сэкономить 10% потребляемых ресурсов за счёт отказа от поставки излишнего тепла. Применение гибких тарифов позволяет влиять на поведение клиентов и сгладить пиковые нагрузки. Обработка накопленных данных за период позволяет учесть потребности клиентов и оптимизировать инфраструктуру [6].
- **Муниципальное управление.** Сбор и обработка данных об автомобильном трафике и загруженности магистралей позволяют гражданам оптимизировать маршруты перемещения, экономя время и автомобильное топливо. Использование этого же подхода для оценки использования общественного транспорта позволяет сократить затраты на него и улучшить качество обслуживания. Очень

востребованной является регистрация всех заездов автомобилей на парковки и выездов оттуда. Это позволяет водителям узнавать с помощью мобильного телефона наличие свободных мест на парковках и сокращает время поиска свободного места. Такой глобальный сервис уже имеется, и область его действия распространяется на 45 городов мира. [6].

- **Образование.** Применение инструментов Big Data позволяет сформировать и поддерживать индивидуальную модель для каждого обучаемого, в которой будут отражены его индивидуальные характеристики и предпочтения, сведения об уже изученных темах и предметах, отзывы и рекомендации, данные преподавателями и менторами. Соответствующий сервис может быть одновременно использован миллионами пользователей в режиме online обучения, но в тоже время предусматривать возможность расширения моделей обучаемых за счёт сведений поступающих из различных учреждений offline обучения (университетов, колледжей, курсов) [12], [13].

Обобщая возможные применения инструментов Big Data, перечислим типичные задачи, решаемые с их помощью:

- Аналитика по клиентам / объектам;
- Операционная и поведенческая аналитика;
- Построение хранилищ данных, экономически эффективных с точки зрения затрат на единицу объёма хранимых данных;
- Борьба с мошенничеством и контроль соблюдения норм.

3. Инструменты Big Data

3.1. Hadoop

За исходную точку, с которой началось развитие инструментов Big Data, можно принять создание Hadoop в начале 2000-х годов [14]. Хотя и до этого были работы по созданию распределённых файловых систем, именно в Hadoop распределённая файловая система HDFS (Hadoop Distributed File System) была объединена с фреймворком MapReduce. В результате появился инструмент, который стало возможно использовать для решения самых разных задач по сбору и обработке больших данных.

В используемом кластере выделяется главный узел, который организует процесс, и узлы данных, на которых располагаются фрагменты данных, и выполняется их обработка. Изначально предполагается, что узлы кластера – это низко-надёжные компьютеры эконом класса (commodity servers).

Основная идея HDFS состоит в том, чтобы разделить массив больших данных на блоки и распределить эти блоки между узлами данных вычислительного кластера. Все блоки имеют одинаковый размер. Поскольку узлы кластера не обладают высокой надёжностью, каждый блок размещается на нескольких узлах в соответствии с предварительно установленным коэффициентом репликации. В случае выхода одного из узлов из строя, все располагавшиеся на нём блоки копируются на другие узлы, чтобы поддерживать заданное число копий. Это обеспечивает устойчивость к отказам. HDFS

поддерживает традиционное иерархическое пространство имён: главным является корневой каталог, каталоги могут быть вложены друг в друга, в одном каталоге могут располагаться файлы и другие каталоги. Обновление существующих файлов не поддерживается, необходимо записывать изменившийся файл как новый.

MapReduce предназначен для того, чтобы организовать параллельный процесс решения задачи на кластере. Процесс выполнения строится из двух фаз: фазы отображения Map и фазы свёртки Reduce. Функции Map и Reduce определяются разработчиком в зависимости от решаемой задачи. Функция Map выполняет первичную обработку данных, лежащих в HDFS. Она запускается на узлах данных, где лежат блоки файла с исходными данными. Результаты работы функции Map передаются функции Reduce, которая объединяет результаты, полученные независимо выполнявшимися функциями Map. В связи с тем, что на первой фазе может выполняться большое число функций Map, для свёртки результатов может быть параллельно запущено много функций Reduce. Однако, в конечном счёте, все их результаты подаются на вход одной функции Reduce, которая формирует окончательный результат. Надёжность работы обеспечивается повторным выполнением задач. В случае если какая-то из задач не выполнилась из-сбоя и не выдала результат, она повторно запускается на другом узле.

В 2013 году появилась версия Hadoop 2.0 [15]. Два самых важных нововведения, которые были реализованы в этой версии:

- В составе Hadoop появился модуль YARN, отвечающий за управление ресурсами кластера и планирование заданий. MapReduce реализован поверх YARN, как один из вариантов обработки данных. С помощью YARN можно реализовать и другие схемы обработки, например, представленные графом сложной структуры, в узлах которого будут выполняться определённые функции. YARN обеспечивает возможность параллельного выполнения нескольких различных задач в рамках кластера и их изоляцию.
- Часть функций по мониторингу заданий была снята с центрального узла, распределяющего ресурсы, и перенесена на новый тип узлов ApplicationMaster. Распределение ресурсов реализовано более эффективно. За счёт этого повысилась производительность Hadoop на 10-15% при одной и той же конфигурации кластера. Ещё одним важным эффектом стал рост реального числа узлов, которые могут эффективно работать в кластере с 4-х до 30-40 тысяч.

3.2.Распределённые файловые системы для хранения больших объёмов данных

Когда Дуг Катинг искал файловую систему для создаваемого с нуля проекта Nutch, который, как часть, включал в себя Hadoop, он взял за основу идеи GFS (Google File System) и реализовал систему с открытым кодом NDFS (Nutch Distributed File Systems), которая в последствие стала называться HDFS. Это было хорошим решением с практической точки зрения - система была достаточно простой в реализации, идеально сочеталась с MapReduce и обеспечивала надёжное хранение файлов большого объёма, которые не могли поместиться в памяти одного узла. Однако, дальнейшее развитие Hadoop и теоретические исследования показали, что не всё идеально. HDFS:

- спроектирована в расчёте только на однопользовательский режим;
- имеет архитектуру Master-Slave, при которой все метаданные размещаются в головном узле, и поэтому система имеет ограничения по масштабированию;
- не полностью POSIX-совместима;
- хорошо справляется с большими файлами, но сильно теряет производительность при работе с большим числом мелких файлов.

Распределённые файловые системы на кластерах начали обсуждаться и проектироваться ещё в середине 90-х годов XX века после того, как стало понятно, что NAS-системы в состоянии обеспечить хранение большого объёма данных, но не в состоянии обеспечить пропускную способность, необходимую для параллельной работы большого числа процессоров, выполняющих обработку данных (независимо от того, кто управляет их работой - независимые пользователи или распределённая программа). С этого момента до настоящего времени спроектированы и разработаны десятки таких систем: Vesta, Galley, PVFS, Swift, GPFS, StorageTank, LegionFS, Google File System, Federated Array of Bricks (FAB), pNFS, Lustre, Panasas file system, zFS, Sorrento, Kybos, Ceph, Intel's Distributed Application Object Storage (DAOS), RADOS, Sirocco, Ursa Minor, SOS – этот список взят из разделов, посвящённых обзору близких работ всего двух публикаций, посвящённых созданию распределённых файловых систем [16], [17]. В начале 2000-х годов распределённые файловые системы стали рассматриваться как основной инструментарий для хранения данных объёмом несколько петабайтов.

Перечисленные выше системы разрабатывались как университетскими командами и стартапами, так и гигантами отрасли такими как Intel и IBM. Они имеют разный уровень готовности. В то время как одни из них представляют собой пилотные проекты, разработанные для иллюстрации некоторых новых концепций, другие уже являются зрелыми продуктами, прошедшими проверку и доводку во многих внедрениях. С ростом зрелости рынка в последние годы появились публикации, в которых приводятся результаты сравнительного тестирования нескольких распределённых файловых систем [18], [19], а также рассматриваются варианты замены HDFS в составе Hadoop [20].

3.3.Экосистема Hadoop

Широкий спектр задач, решаемых Hadoop, привёл к созданию большого числа дополнительных программных продуктов, которые органично расширяют возможности первоначальной системы. Быстрому появлению множества новых продуктов способствовал открытый характер лицензии Apache Hadoop. Нередко новые программные продукты начинали свое развитие в коммерческих компаниях, таких как: Twitter, Facebook, Amazon, но потом были переданы в сообщество под свободными лицензиями. К настоящему времени Hadoop является центром большой экосистемы. Число продуктов, входящих в экосистему, составляет несколько сотен. Существуют разные схемы классификации этих продуктов, различающиеся как по составу классов продуктов, так и по составу классифицируемых продуктов [21], [22], [23]. Ниже предлагается классификация входящих в экосистему продуктов, в которой сделана попытка закрыть белые пятна, присутствующие в каждом из перечисленных источников. В таблице 1

перечислены все выделенные классы, указывается назначение каждого класса, а также представлены примеры программных продуктов, отнесённых к этому классу.

Таблица 1. Состав экосистемы Hadoop

Класс продуктов	Назначение	Примеры продуктов
Распределенные файловые системы	Экосистема Hadoop сконфигурирована таким образом, что может использоваться целый спектр различных распределенных файловых систем. В большинстве прикладных задач применяется классическая HDFS, но некоторые вендоры развивают свои дополнительные решения в этой области.	HDFS, FTP File system, Amazon S3,Windows Azure Storage Blobs (WASB), IBM General Parallel File System, Parascale file system, Appistry CloudIQ Storage, Lustre, Ceph, Intel's Distributed Application Object Storage (DAOS)
NoSQL СУБД	Наибольшее распространение в экосистеме Hadoop получили NoSQL СУБД, которые весьма удобны для обработки и хранения разнообразной плохо структурированной информации. Все базы данных, задействованные в экосистеме, рассчитаны на большие массивы данных и реализуют повышенные требования по отказоустойчивости. Как правило, эти БД используют одну из следующих моделей данных: ключ-значение, документоориентированная, потокоориентированная, графоориентированная.	Apache HBase, Apache HCatalog, Hypertable, Apache Acuumulo, Apache Cassandra, Druid, Neo4j, InfoGrid
NewSQL Базы данных	Представляют собой новое поколение классических реляционных баз данных, но с улучшенной поддержкой горизонтального масштабирования и поддержкой шардирования.	MemSQL, VoltDB
Интерпретаторы SQL запросов	Реализуют интерпретацию SQL-like языка запроса, позволяя решать задачи выборки данных простым и знакомым индустрии способом. Каждая реализация отличается полнотой реализации SQL и производительностью.	Apache Hive, Apache Drill, Apache Spark SQL, Apache Phoenix, Cloudera Impala, HAWQ for Pivotal HD, Presto, Oracle Big Data SQL, IBM BigSQL
Интерпретаторы других языков запросов	Реализуют специализированные языки запросов, которые могут обладать преимуществами перед SQL на определенных классах задачах.	Apache Pig
Обработчики данных	Обеспечивают непосредственно способ обработки поступающих в экосистему данных. Обработчики отличаются типом решаемых прикладных задач, стеком применяемых технологий и способом реализации прикладных задач.	Apache MapReduce, Apache YARN, Apache Tez, Apache Storm, Apache Spark, Apache Hadoop Streaming, Apache Hama

Системы управления задачами кластера	Планирование и выполнение различных Hadoop и прочих задач. Управление очередью и временем исполнения.	Apache Oozie, Apache Fair Scheduler, Apache CapacityScheduler
Библиотеки безопасности и управления доступом	Реализуют пользовательские политики доступа к различным компонентам экосистемы, а также синхронизацию с пользовательскими службами различных операционных систем.	Apache Knox, Apache Ranger,
Сервисы аналитики и визуализации данных	Занимают центральное место в экосистеме Hadoop, как продукты, формирующие добавочную ценность обработанной информации. Реализуют непосредственную визуализацию данных, удобную для конечного пользователя.	Datameer, Pentaho, Pivotal HD, RapidMiner Radoop, SPARQLcity, Apache Giraph
Библиотеки машинного обучения	Решение задач кластеризации, фильтрации, категоризации данных.	Apache Mahout
Библиотеки сериализации данных	Хранение данных в удобном для прикладных приложений виде, а также упрощение обмена сложно структурированной информацией.	Apache Avro, Apache Thrift
Системы развёртывания	Установка дополнительных сервисов кластера. Управление конфигурациями кластера. Пуск, останов, реконфигурация Hadoop, сервисов всего кластера.	Apache Ambari
Сервисы сбора, модификации, перемещения данных, а также сервисы обмена сообщениями	Решают задачу опроса различных источников данных (получение логов с различных серверов, граббинг сайтов и т.д.), преобразования к общему формату, сохранение данных в используемой среде хранения.	Apache Flume, Apache Sqoop, Apache Kafka
Дистрибутивы Hadoop	Программные продукты различных производителей, включающие преднастроенные библиотеки и сервисы экосистемы Hadoop, как свободные, так и проприетарные.	Hortonworks HDP, Cloudera CDH, MapR Data Platform
Системы координации	Реализуют хранение конфигурации, именования, блокировки, и устранения “гонки” за ресурс для узлов кластера.	Apache ZooKeeper
Системы мониторинга и аудита	Мониторинг кластера при помощи досок объявлений (dashboards), ведение метрик, уведомление о событиях кластера (отсутствие места на диске, падение узла и т.д.)	Apache Ambari, Apache Knox, Apache Ranger, Apache ZooKeeper
Библиотеки тестирования	Тестирование различных модулей и компонентов экосистемы, тестовые наборы данных	Apache Bigtop
Прочие библиотеки	Библиотеки, упрощающие разработку прикладных приложений	Cascading Development Frame-work

Как видно из представленной таблицы, экосистема Hadoop представляет разработчику конструктор, который позволяет построить программно-аппаратное решение для широкого спектра задач с практически неограниченными объемами данных, которые нужно обработать или проанализировать. Многообразие представленных на рынке решений отображает колоссальные темпы развития данной экосистемы, а также всей совокупности решений, связанных с обработкой Больших данных.

3.4.Платформа IBM Bluemix

Компания IBM объединила все свои технологические решения по работе с Большими данными в составе платформы IBM Bluemix [24]. Платформа является инструментом для разработки веб- и мобильных приложений для работы с Big Data. Обеспечивается работа с данными в Hadoop, в SQL и NoSQL базах данных IBM, а также в базах данных других производителей, ориентированных на работу с большими данными, такими как MongoLab, ElephantSQL, Redis Cloud и др. В состав платформы входят:

- средства хранения и обработки данных: Hadoop, средство построения хранилищ данных, средства управления данными, средства управления контентом, средства поточной обработки данных;
- развитые средства аналитики для принятия решений, средства построения отчётов, средства контентной аналитики, аналитики по данным геолокации, средства предиктивной аналитики и датамайнинга.

Одной из ключевых особенностей платформы Bluemix является тесная интеграция с сервисами IBM Watson для взаимодействия с пользователями и обработки данных на естественном языке.

На базе платформы Bluemix IBM ведёт разработку и поставляет готовые решения для работы с большими данными:

- крос-функциональные для организации продаж, маркетинга, обработки финансовых данных, управления рисками, управления операциями, управления ИТ, обеспечения защиты от мошенничества, обработки данных персонала;
- отраслевые для различных отраслей: медицины, страхования и т.д.

3.5.Платформа SAP HANA

В 2011 году компания SAP вывела на рынок In-Memory платформу SAP HANA. В составе платформы объединены объектно-графическое, построчное и постолбцовое хранилища данных, а также несколько серверов приложений и их окружение [25], [26], в том числе:

- сервер выполнения SQL-запросов;
- планировщик и оптимизатор вычислений;

- сервер обработки текстовых данных;
- сервер работы с графами;
- сервер предиктивной аналитики;
- средства сжатия данных;
- серверные компоненты (среда исполнения) для языков программирования;
- библиотеки встроенных бизнес-функций

и т.д.

Ключевой особенностью SAP HANA является выполнение всех операций в оперативной памяти. Дисковая память используется только для протоколирования всех операций и поддержки актуальных резервных копий данных. В результате переноса всех процессов в оперативную память обеспечивается высокая скорость обработки данных. Объединение хранилищ данных и серверов приложений в рамках единой платформы позволяет перейти к двухуровневой архитектуре приложений вместо традиционной трёхуровневой (клиент – сервер приложений – сервер баз данных). Это также позволяет повысить скорость работы приложений за счёт устранения узкого бутылочного горлышка, которым являлся интерфейс между сервером приложений и сервером баз данных.

Высокая степень параллелизма обработки табличных данных при выполнении SQL-запросов обеспечивается в SAP HANA техническими решениями по распределению процессов обработки таблиц данных, их отдельных столбцов и фрагментов столбцов между разными серверами, процессорами и процессорными ядрами аппаратной среды.

Максимальные аппаратные конфигурации кластеров, которые используются для работы SAP HANA, на момент написания статьи насчитывали более 100 серверов, содержащих до 8 процессоров и до 12 ТВ оперативной памяти [27], [28]. Каждый процессор может содержать до 15 ядер. В результате, суммарно, максимальные конфигурации SAP HANA могут иметь более 12 000 процессорных ядер и более 1,2 PB оперативной памяти. В связи с высокой стоимостью таких конфигураций на практике в подавляющем большинстве случаев применяются существенно менее мощные аппаратные комплексы. По мере совершенствования технологий производства процессоров и памяти параметры инсталляций SAP HANA по объёму памяти и числу используемых процессорных ядер будут увеличиваться.

Проведенная оценка позволяет отнести платформу SAP HANA к нижнему сегменту инструментов Big Data. Выполнение всех процессов в оперативной памяти делает SAP HANA идеальным средством обработки данных в реальном времени.

4. Архитектурные решения с использованием инструментов Big Data

4.1. Использование инструментов Big Data в интернет-компаниях

Центральное звено хранения и обработки данных во всех крупных интернет компаниях реализуется с помощью инструментов Big Data. Потребности этих компаний, собственно, и были первой причиной создания этих инструментов. Так HDFS разрабатывалась как открытая альтернатива проприетарной GFS (Google File System) [14], [29]. Детальная архитектура созданных и используемых корпоративных систем, различающаяся из-за разнообразия решаемых задач, редко обсуждается в полном объёме, тем не менее, в блогах иногда представлены обобщённые схемы или обсуждаются отдельные технические подробности. Так в [30] и [31] показаны два различающихся варианта общей архитектура корпоративной системы Facebook по состоянию на 2012 г. Вариант из [30] приведен на рис.1.

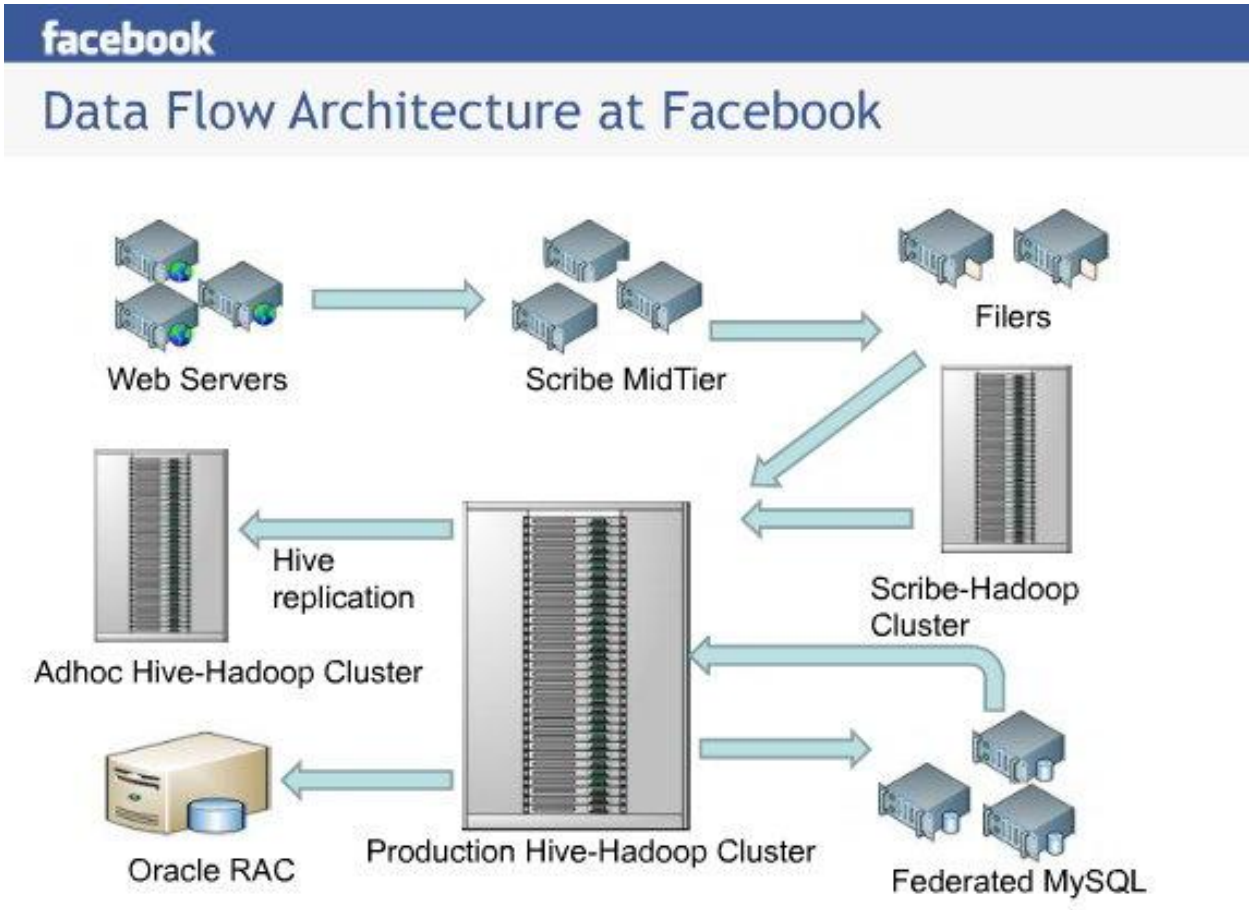

Рис. 1. Корпоративная архитектура, реализованная в Facebook

В отличие от Facebook для Yahoo не опубликована общая архитектура корпоративной системы за исключением эскиза [32], который представлен на рис.2, но

зато опубликован ряд блогов и интервью [9], [33], [34], [35], из которых можно понять, что в корпоративной системе используются такие продукты как Apache Hadoop, Apache Pig, Apache Oozie, Apache HBase, Apache Hive, HCatalog (сервер метаданных Hive), Apache Storm, YARN, Apache Falcon, Apache Spark, Apache Tez, Apache ZooKeeper, Tableau, MicroStrategy.

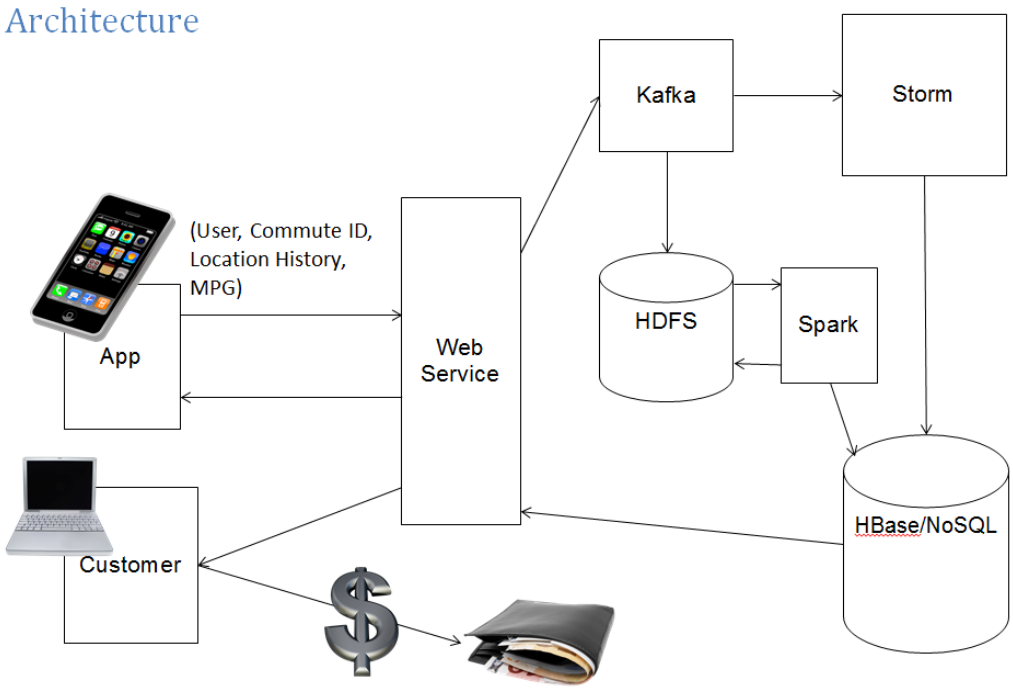

Рис. 2. Эскиз корпоративной архитектуры, реализованной в Yahoo

Представленные варианты архитектуры и различные комментарии позволяют сделать выводы, что основу корпоративной системы крупной интернет компании представляют кластеры Hadoop, причём используется один основной кластер и один или несколько дополнительных, решающих различные задачи. Поверх Hadoop используется SQL-подобная СУБД Hive и ряд дополнительных инструментов Big Data, среди которых можно выделить средства управления мастер-данными и средства бизнес-аналитики.

Такие же тенденции можно увидеть в корпоративной архитектуре крупных российских интернет-компаний. Подтверждением этого является описание опыта работы с Hadoop в Mail.Ru [36].

4.2.Подход и рекомендуемая архитектура SAP для использования инструментов Big Data в составе корпоративных системных ландшафтов

При проектировании корпоративных системных ландшафтов архитекторы компании SAP, лидирующей на мировом рынке корпоративных приложений, считают целесообразным рассматривать возможность применения Hadoop при следующих условиях [37], [22]:

- необходимо обрабатывать данные объёмом в петабайты или даже в перспективе экзабайты, в любом случае их объём намного больше 100 TB и превосходит возможности традиционных реляционных СУБД и SAP HANA;
- не требуется быстрого получения результатов или обработки данных в реальном времени;
- не предъявляется стандартных для транзакционной обработки данных требований обязательного выполнения транзакции или отката в исходное состояние.

Применение Hadoop в указанных случаях значительно увеличит сроки обработки данных, она будет занимать часы или даже дни, однако удельные затраты на единицу объёма данных (MB, GB) значительно сократятся.

Все случаи, когда целесообразно применять Hadoop, были классифицированы, и для каждого из них компания SAP предложила шаблон типовой архитектуры [37], [38]. Перечень разработанных шаблонов:

- Использование Hadoop в масштабах предприятия;
- Hadoop как гибкое хранилище данных;
- Hadoop как простая база данных;
- Hadoop как средство обработки данных;
- Hadoop для аналитики данных (простая аналитика);
- Hadoop для аналитики данных (двухфазная аналитика);
- Hadoop для аналитики данных (федеративные запросы / виртуализация данных).

Помимо архитектурных шаблонов компания SAP предложила также референсную архитектуру, показывающую, как Hadoop может быть встроен в ландшафт корпоративных приложений SAP [37], [39]. В рамках референсной архитектуры выделены компоненты технологии обработки данных, источники данных, аналитические приложения и бизнес-приложения. Интеграция Hadoop с аналитическими и бизнес-приложениями всегда реализуется через хранилище данных или базу данных (SAP HANA, SAP IQ, SAP ASE) с использованием специализированных компонентов обмена данными и управления качеством данных.

Для иллюстрации разработанной архитектуры компания SAP показала четыре примера её применения (без глубокой детализации, только на верхнем уровне):

- использование Hadoop для упреждающего обслуживания оборудования;
- использование Hadoop для выработки рекомендаций в реальном времени по розничным покупкам;
- использование Hadoop для выявления проблем идентификации у оператора телекоммуникационной сети.
- миграция в Hadoop хранилища данных объёмом более 1 PB.

Дополнительно, вместе с архитектурой компания SAP сформулировала общие принципы использования Hadoop в составе корпоративных системных ландшафтов и типовую последовательность действий по развёртыванию Hadoop в составе таких ландшафтов [37], [39]. Взаимодействие бизнес-приложений с Hadoop через SAP HANA

SAP рассматривает как одно из перспективных направлений развития корпоративных систем. В SAP HANA SPS 09 помимо возможностей обмена данными с Hadoop появилась возможность прямого вызова MapReduce из SAP HANA и обратного получения результатов[40].

Роль основного интегрирующего звена для работы с Big Data, которую SAP возлагает на SAP HANA, нашла подтверждение при создании комплекса для фиксации рекорда в книге рекордов Гиннесса. 5 марта 2014 года SAP построила связку SAP HANA–SAP IQ и продемонстрировала работу в режиме реального времени с базой данных объёмом 12,1 PB [41]. SAP HANA в этой связке обеспечивала взаимодействие с клиентами и играла роль гигантского кэша над СУБД, использующей жёсткие диски. 50% данных были структурированными и 50% - неструктурированными. Этим тестом SAP продемонстрировал возможность расширить область применения SAP HANA и других своих технологий до 12 PB без использования Hadoop. Естественно, что для обработки данных существенно больших объёмов будет требоваться применение Hadoop или других аналогичных решений.

5. Новые направления, возникающие в результате применения и дальнейшего развития инструментария Big Data, в научных дисциплинах, использующих моделирование

Основным, бросающимся в глаза, эффектом от развития инструментов для работы с Big Data и их последующего применения является возможность решения множества новых, прежде не решавшихся задач в самых разных областях деятельности, что показано в разделе 2 данной статьи. Однако есть и другой менее видимый эффект от распространения инструментов Big Data, который пока только формируется, – внутри дисциплин (областей знаний), связанных с моделированием окружающей действительности (технических, естественных или социально-экономических систем) и / или решением практических задач на базе использования построенных моделей, возникают новые направления. Эти новые направления позволяют решать свойственные каждой дисциплине задачи для систем, содержащих значительно большее количество отдельных элементов и связей между ними. В зависимости от сложности отдельных элементов новые направления могут решать задачи для систем с миллионами элементов, каждый из которых имеет сотни или тысячи параметров или для систем, содержащих миллиарды элементов, описываемых одним-двумя параметрами.

Первой такой частной дисциплиной, в которой возникло новое направление в связи с появлением инструментов Big Data, является теория алгоритмов сортировки и поиска данных, детально изложенная в [42]. В ходе создания распределённых файловых систем и ряда других программных продуктов Big Data вырабатывались новые подходы и методы теории сортировки и поиска, позволяющие работать с большими объёмами данных. Однако инструменты Big Data начинают проникать и в другие дисциплины, каждая из которых обладает свойственным ей специфическим аппаратом моделирования и решения

задач. В этом разделе мы более детально рассмотрим указанную тенденцию на примере следующих областей науки:

- численные методы,
- теория имитационного моделирования,
- теория управления, приложения которой можно разделить на модели и методы управления в технических системах, модели и методы управления в биологических системах, модели и методы управления в социальных и экономических системах.

Обсуждаемая тенденция появления новых направлений на основе применения инструментов Big Data не ограничивается перечисленными дисциплинами. Можно с уверенностью говорить о факте появления и начале развитии такого направления в теории графов, и в математической статистике, а также о прогнозе появления таких направлений в исследовании операций, в теории игр и в других дисциплинах. В теории графов развитие нового направления связано со спецификой графов социальных сетей, а в математической статистике – со средствами предиктивной аналитики на больших данных. Однако, анализ проявлений этой тенденции в перечисленных областях выходит за рамки данной статьи, как из-за ограниченности её объёма, так и из-за того, что дополнительно потребуется рассмотреть особенности наблюдаемой тенденции в дисциплинах, являющихся разделами математики. Это может стать предметом отдельной обзорной обобщающей публикации.

Фактически, формирующиеся новые направления возникают вследствие того, что появляется возможность решать проблему, привлекая десятки и сотни тысяч самостоятельных узлов обработки данных. Работу этих узлов нужно организовать, для этого нужно использовать новые подходы, новые алгоритмы, схемы распределения работ и консолидации полученных результатов.

Может показаться, что эти задачи не являются чем-то новым, – уже долгое время развивается направление суперкомпьютерных вычислений, в рамках которого решаются похожие задачи [43]. Кластеры Big Data и самые мощные суперкомпьютеры с распределённой памятью, тоже представляющие собой кластеры, на первый взгляд, очень похожи. В обоих случаях используется архитектура Master–Slave. Тем не менее, имеется много различий:

- в суперкомпьютерных кластерах используются значительно более высокоскоростные каналы для обмена данными между узлами;
- ветви программы, параллельно выполняющиеся в разных узлах суперкомпьютерного кластера, обмениваются между собой сообщениями или данными. В отличие от этого в кластере Hadoop возможности обмена данными очень ограничены. Map-задачи при запуске берут исходные данные из блоков HDFS, располагающихся на данном узле, а результат направляют на вход одной из Reduce-задач;
- при создании суперкомпьютеров используются более надёжные компоненты, чем commodity-сервера в кластерах Hadoop. Поэтому в алгоритмах для кластеров

Hadoop обязательно предусматривается ситуация, что в части узлов могут возникнуть сбои, и решаемые ими задачи надо решить повторно.

Предельно экономный подход, реализуемый в кластерах Hadoop, позволил снизить стоимость параллельной обработки данных примерно в 10 раз по сравнению с суперкомпьютерами. Это дало возможность массово применять инструменты Big Data в корпоративных системах и выйти далеко за пределы традиционных областей применения суперкомпьютеров:

- ядерная физика,
- моделирование климата,
- генная инженерия,
- проектирование интегральных схем,
- анализ загрязнения окружающей среды,
- создание лекарственных препаратов и новых материалов,
- проектирование эффективных форм с учётом гидро- и аэродинамики.

Использование инструментов Big Data в составе корпоративных систем с их большим числом параллельно обращающихся пользователей и режимом работы 24/7 сразу же заставило предъявлять к ним значительно более высокие требования по надёжности, доступности, скорости работы по сравнению с программами для суперкомпьютерных расчётов, используемыми, в основном, в научно-исследовательской и проектно-конструкторской деятельности. Эти более высокие требования, а также радикально отличающиеся от суперкомпьютерных систем принципы построения процессов обработки данных делают необходимым появление внутри каждой научной дисциплины нового направления, объединяющего разрабатываемые решения по использованию создаваемых инструментов в рамках экосистемы Hadoop.

Термин Big Data объединил под единый зонтик все теоретические результаты, технические решения, алгоритмы и инструменты для работы с большими объёмами данных, развиваемые в рамках экосистемы Hadoop. Эти решения алгоритмы и инструменты содержат элементы нового раздела теории алгоритмов сортировки и поиска. По аналогии, можно ввести похожие названия для новых направлений в других науках для того, чтобы отделить все наработки и инструменты от похожих по назначению аналогов в рамках суперкомпьютерных вычислений:

- Big Calculation - для нового направления в численных методах, связанного с вычислениями на кластерах Hadoop;
- Big Simulation - для нового направления в имитационном моделировании на базе кластеров Hadoop;
- Big Management - для нового направления в управлении социальными и экономическими системами на базе кластеров Hadoop;
- Big Optimal Control - для нового направления в теории оптимального управления на базе кластеров Hadoop.

Естественный вопрос, который может возникнуть: почему нельзя подвести все перечисленные направления под единый зонтичный термин Big Data? Дело в том, что для решения многих задач может потребоваться именно большой объём вычислений или большой объём моделирования, при этом объём используемых данных будет существенно меньше условного 1 PB, и данные не будут иметь некоторых других характеристик, свойственных Big Data согласно разделу 1 данной статьи. Иначе говоря, данные могут быть обработаны в обычной СУБД, но для реализации алгоритмов потребуется кластер Hadoop.

Ещё одним теоретическим обобщением, которое не позволяет отобразить характерные особенности процессов моделирования и обработки данных с использованием различных моделей, является подход 5W, описываемый во многих источниках (например, [44]). Согласно этому подходу, все инструменты, создаваемые в рамках экосистемы Hadoop и даже за её пределами, относят к средствам аналитики, отвечающим на один или несколько из следующих пяти вопросов:

- What is happening? – Что происходит?
- Why did it happen? – Что случилось?
- What could happen? – Что может случиться?
- What action should I take? – Какие действия я должен сделать?
- What did I learn, what's best? – Чему я научился, что является лучшим?

В этом случае инструменты, построенные на основе совершенно разных методов моделирования, математических и инженерных подходов, объединяются в одну группу. При таком объединении и обобщении нивелируется специфика используемых методов моделирования и тормозится развитие новых востребованных разделов отдельных дисциплин.

5.1. Big Calculation

Большинство алгоритмов численных методов изначально разрабатывались в последовательной парадигме. Мы рассмотрим только два примера, чтобы показать применение нового подхода, который возникает под влиянием использования инструментов Big Data. Первый пример - это симплекс метод для решения задач линейного программирования, который можно обобщённо представить следующим образом:

- Поиск одной из вершин выпуклого многогранника, представляющего собой область допустимых решений.
- Последующее перемещение по рёбрам этого многогранника от одной вершины к другой до тех пор, пока не будет найдена вершина, в которой целевая функция принимает максимальное значение.

Второй пример – метаэвристика поиска с чередующимися окрестностями для решения задач непрерывной и дискретной оптимизации VNS [45], которую в общем виде можно описать так:

- Определение последовательности размеров окрестностей и начальной точки.
- Циклический поиск локального оптимума, начиная с первого заданного размера окрестности и заданной начальной точки. Если в результате локального поиска на очередном шаге будет найдено новое лучшее значение оптимума, использовать на следующем шаге найденную точку оптимума в качестве начальной, в противном случае перейти к поиску на следующем размере окрестностей.

Появление вычислительных кластеров повлекло за собой разработку новых параллельных численных методов. Так компания SAP к 2004 г. испытывала потребность в переходе на параллельные алгоритмы решения задач целочисленного линейного программирования из-за имеющей место тенденции роста размерности при оптимизации цепочек поставок [46]. В компании проводились исследования по реализации параллельных алгоритмов с помощью декомпозиции исходных матриц на блоки и параллельного решения задачи оптимизации отдельных блоков на разных узлах кластера. Результаты проведенных исследований показали, что при увеличении числа параллельно обрабатываемых блоков основной матрицы более 25-30 дальнейшего увеличения скорости решения задачи не происходит.

Для повышения эффективности поиска глобального оптимума с помощью чередующихся окрестностей также было разработано несколько вариантов параллельного алгоритма VNS (PVNS). Наиболее эффективный из них заключался в наращивании числа решений, выбираемых в текущей окрестности, и параллельном выполнении локального поиска для каждого из них. Этот подход, как и в предыдущем примере, тоже предполагает использование нескольких или, как максимум, нескольких десятков параллельно работающих узлов кластера.

Инструменты Big Data в своём современном состоянии ориентированы на совершенно другие характеристики кластеров. Число параллельно работающих узлов может составлять десятки и сотни тысяч. Для эффективного использования таких ресурсов требуются принципиально иные алгоритмы. Например, та же задача линейного программирования может решаться следующим образом: параллельно будут найдены все вершины выпуклого многогранника и вычислены значения целевой функции в каждой из них, а потом из этих значений будет выбрано максимальное. При больших размерностях задачи каждый узел кластера будет последовательно решать несколько таких независимых подзадач. Однако для любой задачи конкретной размерности всегда будет существовать некоторое количество узлов в кластере, начиная с которого дальнейшее увеличение их числа при решении задачи описанным методом будет уменьшать общее время решения по сравнению с методами распараллеливания, ориентированными на несколько десятков узлов. Общий объём выполненных вычислений при использовании предложенного подхода существенно увеличится, однако общее время решения может значительно сократиться.

Точно также подход максимального распараллеливания может быть применён для быстрого решения задачи поиска вместо применения метаэвристики VNS, – область

допустимых решений может быть покрыта сеткой начальных точек, общее число которых может в несколько раз превысить число узлов в кластере. Из каждой начальной точки параллельно будет выполнен локальный поиск, а потом сопоставлены полученные результаты.

По мере расширения использования инструментов Big Data будут появляться всё новые и новые численные методы, ориентированные на возможности больших кластеров.

5.2.Big Simulation

Системы имитационного моделирования, работающие на отдельном компьютере или сервере, позволяют моделировать поведение максимум нескольких десятков тысяч объектов. Для преодоления этого ограничения был разработан новый подход к построению имитационных моделей – агентное моделирование и инструменты для построения агентно-ориентированных моделей с использованием грид-систем [47]. В рамках агентного моделирования имитационная модель представляет собой децентрализованное сообщество независимо действующих агентов. К настоящему моменту реализованы десятки инструментов агентного моделирования на базе суперкомпьютеров, с помощью которых можно строить имитационные модели, включающие сотни миллионов и миллиарды объектов. Это позволяет решать задачи, в которых необходимо моделировать большое число объектов, например:

- прогнозирование развития социально-экономических систем (стран, регионов, городов);
- моделирование миграционных процессов;
- имитация и оптимизация пешеходного движения;
- моделирование транспортных перевозок и транспортных систем;
- прогнозирование экологического состояния окружающей среды;
- моделирование работы систем сотовой связи

и др.

Появление в составе корпоративных систем кластеров Hadoop, которые могут содержать сведения о сотнях миллионов участников социальной сети, или о десятках миллионов активов (зданий, сооружений, единиц оборудования) естественным образом выдвигает вопрос об использовании этих данных для построения имитационных моделей. Например, имитационной модели, которая будет позволять прогнозировать надёжность работы оборудования в зависимости от использования различных стратегий технического обслуживания и ремонта. В качестве технических систем с большим числом элементов (десятки миллионов) могут выступать региональная или мультинациональная электрическая сеть, сеть трубопроводов, сеть железных дорог и т.д. Другой задачей, где необходимо использовать имитационную модель на базе кластера Hadoop, является прогноз развития социальной сети или прогноз поведения её участников при возникновении определённых обстоятельств.

Использование суперкомпьютеров не нашло широкого применения для решения этих задач. Во многом, на взгляд авторов, это обусловлено высокими затратами, возникающими при этом. Затраты в случае использования кластера Hadoop будут на порядок меньше, и сам кластер доступнее – часто он уже есть в корпоративном периметре или арендуется в общедоступном облаке. До появления Hadoop 2.0 сложно было ожидать реализации систем имитационного моделирования. Жёсткий двухстадийный однонаправленный процесс обработки данных MapReduce не соответствовал характеру многостадийной циклической обработки данных при имитационном моделировании с обязательным обменом данными между взаимосвязанными элементами модели после каждого шага. Появление YARN радикально изменило эту ситуацию. Стало возможным конструировать процессы обработки данных с любым числом стадий и реализовывать сложные схемы обмена данными между задачами, выполнявшимися на разных узлах.

Для упрощения разработки на базе YARN приложений, в которых необходимо обеспечить многостадийный процесс обработки данных и обмен информацией между узлами, была разработана платформа Apache Hama. Эта платформа реализует модель программирования BSP (Bulk Synchronous Parallelism). Согласно этой модели весь процесс вычислений состоит из последовательности супершагов [48]. Каждый супершаг выполняется параллельно каждым узлом, участвующим в BSP-вычислениях. Супершаг содержит три стадии: локальные вычисления, обмен информацией и синхронизационный барьер. Каждый узел имеет локальную память, которая доступна только этому узлу в течение всех супершагов. Кроме того во время локальных вычислений на данном супершаге каждый узел имеет доступ к сообщениям, посланным другими узлами во время предыдущего супершага. Он также может послать сообщения другим узлам во время стадии информационного обмена, чтобы они были прочитаны ими во время следующего супершага. Синхронизационный барьер позволяет синхронизировать работу всех узлов, чтобы обеспечить получение ими всех посланных им сообщений до начала следующего супершага. Предусматривается также обработка возможных сбоев. Каждый узел может использовать точки восстановления, чтобы эпизодически сохранять изменившуюся часть памяти в распределённую файловую систему. Это позволяет восстановить последнее запомненное состояние в случае сбоя.

В связи с созданием необходимого инструментария для разработки систем имитационного моделирования на базе кластеров Hadoop в ближайшем будущем можно ожидать появления сначала отдельных имитационных моделей, а потом систем имитационного моделирования класса Big Simulation.

5.3.Big Management

Под термином Management понимается управление в социальных и экономических системах. В настоящее время в автоматизированных системах управления используется всего два способа сведения множества частных показателей к укрупнённым показателям, отражающим соответствие принятой стратегии развития: использование системы сбалансированных показателей (Balanced Scorecard) и метод управления портфелями [49].

Система сбалансированных показателей строится как иерархическая система, основанием которой является множество показателей деятельности сотрудников низовых звеньев, а на верхнем уровне они консолидируются в небольшое число показателей, контролируемых топ-менеджментом и характеризующих деятельность компании в целом. Число уровней иерархии в системе сбалансированных показателей соответствует числу уровней иерархии в системе управления компанией. Система управления портфелями позволяет классифицировать и объединить в ограниченное число портфелей набор относительно однородных объектов или действий, которыми необходимо управлять. При этом число элементов в каждом портфеле может быть достаточно большим.

Использование программных инструментов для поддержки двух указанных способов менеджмента позволяет ставить чёткие количественно выраженные цели, контролировать их своевременное достижение или выявлять причины, почему они не достигнуты. В случае появления отклонений по результатам анализа их причин вносятся изменения в бизнес-процессы и/или корректировки в систему показателей.

Почему возникает потребность в инструментах Big Management?

1. Мы уже упоминали в предыдущем разделе, что крупная компания может владеть или использовать десятки миллионов активов (зданий, сооружений, единиц оборудования). Все активы нуждаются в профилактике (техническом осмотре, контроле состояния), ремонте, модернизации. Это всё работы или проекты, для которых необходимо выделение бюджета, специалистов, зачастую, временный вывод из эксплуатации и т.д. Если с каждым активом необходимо провести работы хотя бы один раз в квартал, мы сразу получим около 100 миллионов отдельных мелких или крупных проектов. До последнего времени системы управления проектами могли уверенно поддерживать одновременное планирование и учёт работы по нескольким десяткам тысяч проектов. После перевода SAP PPM (Project Portfolio Management) на платформу SAP HANA SAP заявил, что система сможет поддерживать неограниченное число проектов, однако сообщений об опыте внедрения для поддержки миллионов одновременно выполняемых проектов, не говоря уже о десятках или сотнях миллионов, пока не появлялось.

2. Системы сбалансированных показателей успешно функционируют в крупных компаниях, численность сотрудников в которых может составлять десятки тысяч человек. В масштабах целой страны, или такого объединения, как Евросоюз, общая численность управленцев может составлять несколько миллионов. Потребность в организации эффективной работы таких больших аппаратов управления очень велика, однако пока нет примеров внедрения системы сбалансированных показателей для таких больших структур.

5.4. Big Optimal Control

Основная задача теории оптимального управления – найти последовательность управляющих воздействий, которые обеспечат переход системы из имеющегося начального состояния в некоторое заданное конечное и при этом будет достигаться

максимум или минимум заданного критерия. Математическая модель, используемая для описания задачи, включает в себя: начальную точку, параметры управления, описание поведения системы, оптимизируемый критерий, существующие ограничения на ресурсы. Поведение детерминированных систем описываются дифференциальными уравнениями, дифференциальными уравнениями в частных производных и конечными автоматами. Вероятностные системы описываются стохастическими дифференциальными уравнениями и марковскими процессами.

Проблемы, связанные с решением практических задач оптимального управления привели к появлению отдельных групп численных методов для решения задач оптимального управления и специальных программных комплексов [50]. В случаях, когда формальное описание задачи не может быть сформулировано из-за его сложности, но может быть построена имитационная модель системы, оптимальное решение может быть найдено методами прямой оптимизации, работающими поверх имитационной модели.

На начальном этапе теория оптимального управления создавалась для оптимизации управления техническими системами. Примеры, приведенные в предыдущих разделах, показывают, что существует много технических систем с миллионами параметров управления (сеть газопроводов страны, региона, электрическая сеть аналогичных масштабов и т.д.). Для упрощения управления такими сложными объектами системы соответствующие автоматического управления строятся в виде многоуровневых иерархических систем. В большинстве случаев нижние уровни полностью управляются автоматически, на их долю приходятся рутинные операции и предотвращение аварийных ситуаций. На верхних уровнях таких больших систем из-за высокой сложности и недостаточной проработанности систем управления автоматическое управление в большинстве случаев подменяется автоматизированным - в контур управления включают людей-операторов.

По мере развития теория оптимального управления вышла за пределы чисто технических систем, и на стыке техники и технологий с естественнонаучными и экономическими дисциплинами появились и продолжают возникать всё новые и новые задачи большой размерности. Приведём по одному примеру:

- (на стыке с науками о земле): На давно разрабатываемом нефтяном месторождении пробурено несколько тысяч скважин, получены данные о характеристиках проницаемости пород в точках бурения, имеются данные сейсморазведки и данные истории добычи. По этим данным должна быть построена гидродинамическая модель месторождения, а затем решена задача оптимального управления, которая может быть поставлена в одном из двух вариантов:
  - в условиях заданного ограничения на бюджет найти совокупность геолого-технических мероприятий, которые позволят максимально повысить нефтеотдачу;
  - для заданного уровня нефтеотдачи найти совокупность геолого-технических мероприятий, для выполнения которых потребуется минимальный бюджет.

- (на стыке с сельским хозяйством): В рамках примера по использованию больших данных для увеличения производительности сельскохозяйственного производства, приведенного в разделе 2 данной статьи, должна решаться задача минимизации затрат на достижение заданного объёма урожая. Управляющими воздействиями в данном случае будут являться агротехнические мероприятия.
- (на стыке с экономикой): В рамках примера 1, приведенного в разделе 5.3 данной статьи, по планированию технического обслуживания и ремонта сложного технического комплекса, должна решаться задача минимизации затрат на достижение заданного уровня надёжности его работы. Управляющими воздействиями в данном случае будут являться работы по техническому обслуживанию и ремонту.

Для решения больших задач оптимального управления на базе низкобюджетных кластеров потребуется развивать всю совокупность решений и методов, перечисленных в предыдущих разделах 5.1, 5.2, 5.3: новые численные методы, приложения на основе модели программирования BSP, большие имитационные модели, методы построения иерархических систем управления и методы управления портфелями однородных процессов или объектов.

Список литературы

1. Demchenko Y. Defining the Big Data Architecture Framework (BDAF). Outcome of the Brainstorming Session at the University of Amsterdam. SNE Group. University of Amsterdam, Amsterdam, 2013. Режим доступа: http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf (дата обращения 03.02.2015).
2. Parker D. Changing the World with Big Data. Real-time with Real Results // IGEL (Initiative for Global Environmental Leadership) бизнес-школы Wharton School в университете штата Pennsylvania: сайт, ноябрь 2013. Режим доступа: http://igel.wpengine.netdna-cdn.com/wp-content/uploads/2013/11/David-Parker.pdf (дата обращения 03.02.2015).
3. Financial Services // Datameer: сайт компании. Режим доступа: http://www.datameer.com/solutions/industries/financial-services.html (дата обращения 03.02.2015).
4. Beloussov S. Cloud computing is now IT // Proc. of Parallels Summit 2014. Parallels: сайт компании. Режим доступа: http://sp.parallels.com/fileadmin/media/hcap/events/summit/2014/documents/Summit2014-keynote-SergueiBeloussov.pdf (дата обращения 03.02.2015).
5. The Body as a Source of Big Data // The Institute for Health Technology Transformation: сайт. Режим доступа: http://ihealthtran.com/images/Infographic-the-body-as-a-source-of-big-data-HealthIT-Infographic-NetApp-Infographic.pdf (дата обращения 03.02.2015).

6. Sustainability in the Age of Big Data. Special Report // IGEL (Initiative for Global Environmental Leadership) бизнес-школы Wharton School в университете штата Pennsylvania: сайт, сентябрь 2014. Режим доступа: http://d1c25a6gwz7q5e.cloudfront.net/reports/2014-09-12-Sustainability-in-the-Age-of-Big-Data.pdf (дата обращения 03.02.2015).
7. A Focus on Efficiency. A whitepaper from Facebook, Ericsson and Qualcomm. 16 сентября 2013. Режим доступа: https://fbcdn-dragon-a.akamaihd.net/hphotos-ak-prn1/851575_520797877991079_393255490_n.pdf (дата обращения 03.02.2015).
8. About Twitter, Inc. // Twitter, Inc. : сайт. Режим доступа: https://about.twitter.com/company (дата обращения 03.02.2015).
9. Asay M. Why the world's largest Hadoop installation may soon become the norm // Techrepublic: сайт, 12 сентября 2014. Режим доступа: http://www.techrepublic.com/article/why-the-worlds-largest-hadoop-installation-may-soon-become-the-norm/ (дата обращения 03.02.2015).
10. Retail // Datameer: сайт компании. Режим доступа: http://www.datameer.com/solutions/industries/retail.html (дата обращения 03.02.2015).
11. Telecommunications // Datameer: сайт компании. Режим доступа: http://www.datameer.com/solutions/industries/telecommunications.html (дата обращения 03.02.2015).
12. van Rijmenam M. Big Data Will Revolutionize Education // DATAFLOO: сайт, 29 апреля 2014. Режим доступа: https://datafloq.com/read/big-data-will-revolutionize-learning/206 (дата обращения 03.02.2015).
13. Dalisay T. Big Data in Education: Big Potential or Big Mistake // Socialnomics: сайт, 13 января 2014. Режим доступа: http://www.socialnomics.net/2014/03/05/big-data-in-education-big-potential-or-big-mistake/ (дата обращения 03.02.2015).
14. Уайт Т. Hadoop. Подробное руководство: пер. с англ. СПб.: Питер, 2013. 672 с. (Сер. Бестселлеры O'Reilly).
15. Емельянов И. Как обновление Hadoop 2.0 сделало «большие данные» доступнее для бизнеса // CIO: сайт, 17 октября 2013. Режим доступа: http://www.computerra.ru/cio/5598 (дата обращения 03.02.2015).
16. Weil S.A., Brandt S.A., Miller E.L, Long D.D.E., Maltzahn C. Ceph: A Scalable, High-Performance Distributed File System // OSDI '06: 7th USENIX Symposium on Operating Systems Design and Implementation, 2006. P. 307-320. Режим доступа: http://static.usenix.org/event/osdi06/tech/full_papers/weil/weil.pdf (дата обращения 03.02.2015).
17. Brim M.J., Dillow D.A, Oral S., Settlemyer B.W., Wang F. Asynchronous Object Storage with QoS for Scientific and Commercial Big Data // 8th Parallel Data Storage Workshop, November 18, 2013, Denver, CO. Режим доступа: http://www.pdsw.org/pdsw13/papers/p7-pdsw13-brim.pdf (дата обращения 03.02.2015).

18. Depardon B., Le Mahec G., Seguin C. Analysis of Six Distributed File Systems. Research Report HAL Id: hal-00789086, 2013. 44 p. Режим доступа: https://hal.inria.fr/hal-00789086/PDF/a_survey_of_dfs.pdf (дата обращения 03.02.2015).
19. Donvito G., Marzulli G., Diacono D. Testing of several distributed file-systems (HDFS, Ceph and GlusterFS) for supporting the HEP experiments analysis // Journal of Physics: Conference Series: 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013). IOP Publishing, 2014. Vol. 513. Art. no. 042014. DOI: 10.1088/1742-6596/513/4/042014
20. Harris D. Because Hadoop isn't perfect: 8 ways to replace HDFS // GIGAOM: сайт, 11 июля 2012. Режим доступа: https://gigaom.com/2012/07/11/because-hadoop-isnt-perfect-8-ways-to-replace-hdfs/ (дата обращения 03.02.2015).
21. Kerzner M., Maniyam S. Chapter 12. Big Data Ecosystem // Hadoop Illuminated, 2014. Режим доступа: http://hadoopilluminated.com/hadoop_illuminated/Bigdata_Ecosystem.html (дата обращения 03.02.2015).
22. Powlas T. Big Data in an SAP Landscape – ASUG Webcast Part 1 // SAP Community Network: сайт, 29 декабря 2013. Режим доступа: http://scn.sap.com/community/business-intelligence/blog/2013/12/29/big-data-in-an-sap-landscape-asug-webcast-part-1 (дата обращения 03.02.2015).
23. The Hadoop Ecosystem Table // GitHub: сайт. Режим доступа: http://hadoopecosystemtable.github.io/ (дата обращения 03.02.2015).
24. Big Data Overview: презентация IBM на семинаре Big Data & Analytics Day. Москва, 27 января 2015 года.
25. SAP HANA® Database for Next-Generation Business Applications and Real-Time Analytics. Explore and Analyze Vast Quantities of Data from Virtually Any Source at the Speed of Thought. SAP AG, 10 October 2013. 18 p. Режим доступа: http://www.slideshare.net/SAPMENA/hana-27077351 (дата обращения 03.02.2015).
26. SAP HANA Master Guide. SAP HANA Platform SPS 09. Document Version: 1.1 (2014-12-17). SAP AG. 84 p. Режим доступа: http://help.sap.com/hana/SAP_HANA_Master_Guide_en.pdf (дата обращения 03.02.2015).
27. Arnold A. In-memory Database Architecture & Landscape Options for a Real-time Business. SAP AG, October 2014. 33 p. Режим доступа: https://hcp.sap.com/content/dam/website/saphana/en_us/Technology%20Documents/SAP%20HANA%20Architecture%20and%20Landscape%20Options.pdf (дата обращения 03.02.2015).
28. Watts D., Krutov I. xREF: IBM x86 Server Reference. IBM, 2015. Режим доступа: http://www.redbooks.ibm.com/redpapers/pdfs/redpxref.pdf (дата обращения 03.02.2015).

29. Hauff C. Big Data Processing, 2014/15. Lecture 5: GFS & HDFS // Delft University of Technology: сайт. Нидерланды. Режим доступа: http://www.st.ewi.tudelft.nl/~hauff/BDP-Lectures/5_filesystem_gfs_hdfs.pdf (дата обращения 03.02.2015).
30. Lovett M. Apache Hadoop: The Open Source Elephant of Big Data // Trenton Systems: сайт компании, 6 июля 2012. Режим доступа: http://blog.trentonsystems.com/apache-hadoop-the-open-source-elephant-of-big-data/ (дата обращения 03.02.2015).
31. Popescu A., Bacalu A.-M. Life of Data at Facebook // myNoSQL : сайт, 3 сентября 2012. Режим доступа: http://nosql.mypopescu.com/post/30815314471/life-of-data-at-facebook (дата обращения 03.02.2015).
32. *Evans B., Graves T.* Storm and Spark at Yahoo: Why Chose One Over the Other // *yahoohadoop: сайт, 18 сентября 2014. Режим доступа:* http://yahoohadoop.tumblr.com/post/98213421641/storm-and-spark-at-yahoo-why-chose-one-over-the (дата обращения 03.02.2015).
33. Singh *S.* Apache HBase at Yahoo! – Multi-tenancy at the Helm Again // *Yahoo Developer Network: сайт, 7 июня 2013. Режим доступа:* https://developer.yahoo.com/blogs/hadoop/apache-hbase-yahoo-multi-tenancy-helm-again-171710422.html#more-id (дата обращения 03.02.2015).
34. Zicari R.V. Hadoop at Yahoo. Interview with Mithun Radhakrishnan: интервью // ODBMS Industry Watch: сайт, 21 сентября 2014. Режим доступа: http://www.odbms.org/blog/2014/09/interview-mithun-radhakrishnan/ (дата обращения 03.02.2015).
35. Feng A., Evans B., Dagit D., Patil K., Poulosky P., Kapur D., Lal A. The Evolution of Storm at Yahoo and Apache // *yahoohadoop : сайт, 29 сентября 2014. Режим доступа:* http://yahoohadoop.tumblr.com/post/98751512631/the-evolution-of-storm-at-yahoo-and-apache (дата обращения 03.02.2015).
36. Лапань М. Практическое слоноводство: презентация // Семинар Hadoop Kitchen в Mail.Ru Group, 27 сентября 2014. *Режим доступа:* http://www.youtube.com/watch?v=QAejaeTvj_M *начало с 1:20:40* (дата обращения 03.02.2015).
37. Burdett D., Tripathi R. CIO Guide. How to Use Hadoop with Your SAP® Software Landscape. SAP AG, Февраль 2013. 40 p. Режим доступа: http://hortonworks.com/wp-content/uploads/2013/09/CIO.Guide_.How_.to_.Use_.Hadoop.with_.Your_.SAP_.Software.Landscape.pdf (дата обращения 03.02.2015).
38. Powlas T. Using HANA and Hadoop, Key Scenarios - Part 2 ASUG Big Data Webcast // SAP Community Network: сайт, 29 декабря 2013. Режим доступа: http://scn.sap.com/community/business-intelligence/blog/2013/12/29/using-hana-and-hadoop-key-scenarios (дата обращения 03.02.2015).

39. Powlas T. Fitting Hadoop in an SAP Software Landscape – Part 3 ASUG Webcast // SAP Community Network: сайт, 29 декабря 2013. Режим доступа: http://scn.sap.com/community/business-intelligence/blog/2013/12/29/fitting-hadoop-in-an-sap-software-landscape-part-3-asug-webcast (дата обращения 03.02.2015).
40. Eacrett M. What is new in SAP HANA SPS 09 // SAP HANA: сайт, 21 октября 2014. Режим доступа: https://blogs.saphana.com/2014/10/21/what-is-new-in-sap-hana-sps-09/ (дата обращения 03.02.2015).
41. Hagman M. Guinness World Record – Largest Data Warehouse // SAP HANA: сайт, 5 марта 2014. Режим доступа: https://blogs.saphana.com/2014/03/05/guinness-world-record-largest-data-warehouse/ (дата обращения 03.02.2015).
42. Кнут Д.Э. Искусство программирования. Т. 3. Сортировка и поиск: пер. с англ. М.: Вильямс, 2012. 824 с.
43. Воеводин В. В., Воеводин Вл. В. Параллельные вычисления. СПб.: БХВ-Петербург, 2002. 608 с.
44. Analytics and Big Data and ROI…Oh My! : презентация учеб. материала // BlueSpire : сайт компании, 16 сентября 2014. Режим доступа: http://www.slideshare.net/bluespiremarketing/analytics-and-big-data-and-roioh-my-trendlab-webinar (дата обращения 03.02.2015).
45. Кочетов Ю. А., Младенович Н., Хансен П. Локальный поиск с чередующимися окрестностями // Дискретный анализ и исследование операций. 2003. Т. 10, № 1. С. 11-43.
46. Braun H. Optimization with Grid Computing // Workshop on Cyberinfrastructure (CI) in Chemical and Biological Process Systems: Impact and Directions, September 25-26, 2006, Arlington, VA. Режим доступа: https://smartmanufacturingcoalition.org/sites/default/files/optimization_with_grid_computing.pdf (дата обращения 03.02.2015).
47. Макаров В.Л., Бахтизин А.Р., Васенин В.А., Роганов В.А., Трифонов И.А. Средства суперкомпьютерных систем для работы с агент-ориентированными моделями // Программная инженерия. 2011. № 3. С. 2-14.
48. Fegaras L. Supporting Bulk Synchronous Parallelism in Map-Reduce Queries // University of Texas at Arlington: сайт, 2012. Режим доступа: http://lambda.uta.edu/mrql-bsp.pdf (дата обращения 03.02.2015).
49. Сухобоков А. А. Исследование и разработка моделей и архитектуры средств контроллинга для межрегиональных предприятий в составе систем класса ERP II: дис. … канд. техн. наук. М., МГТУ им. Баумана, 2009. 196 с.
50. Маджара Т. И. Интеллектуальная система для решения задач оптимального управления с вычислительными особенностями: дис. ... канд. техн. наук. Владивосток, Ин-т автоматики и процессов управления ДВО РАН, 2011. 149 с.

Science & Education of the Bauman MSTU
Electronic journal
ISSN 1994-0408

Science and Education of the Bauman MSTU, 2015, no. 03, pp. 207–240.

DOI: **10.7463/0315.0761354**

Received: 18.02.2015

The Big Data Tools Impact on Development of Simulation-Concerned Academic Disciplines

A.A. Sukhobokov[1,*], **D.S. Lakhvich**[1]

[*]artem.sukhobokov@yandex.ru

[1]Bauman Moscow State Technical University, Moscow, Russia

Keywords: Big Data tools, Hadoop clusters, scientific disciplines, models with big number of elements

The article gives a definition of Big Data on the basis of 5V (Volume, Variety, Velocity, Veracity, Value) as well as shows examples of tasks that require using Big Data tools in a diversity of areas, namely: health, education, financial services, industry, agriculture, logistics, retail, information technology, telecommunications and others. An overview of Big Data tools is delivered, including open source products, IBM Bluemix and SAP HANA platforms. Examples of architecture of corporate data processing and management systems using Big Data tools are shown for big Internet companies and for enterprises in traditional industries. Within the overview, a classification of Big Data tools is proposed that fills gaps of previously developed similar classifications. The new classification contains 19 classes and allows embracing several hundreds of existing and emerging products.

The uprise and use of Big Data tools, in addition to solving practical problems, affects the development of scientific disciplines concerning the simulation of technical, natural or socio-economic systems and the solution of practical problems based on developed models. New schools arise in these disciplines. These new schools decide peculiar to each discipline tasks, but for systems with a much bigger number of internal elements and connections between them. Characteristics of the problems to be solved under new schools, not always meet the criteria for Big Data. It is suggested to identify the Big Data as a part of the theory of sorting and searching algorithms. In other disciplines the new schools are called by analogy with Big Data: Big Calculation in numerical methods, Big Simulation in imitational modeling, Big Management in the management of socio-economic systems, Big Optimal Control in the optimal control theory. The paper shows examples of tasks and methods to be developed within new schools. The educed tendency is not limited to the considered disciplines: there are other ones such as graph theory, mathematical statistic, game theory, and operations research.

References

1. Demchenko Y. *Defining the Big Data Architecture Framework (BDAF). Outcome of the Brainstorming Session at the University of Amsterdam*. SNE Group. University of Amsterdam, Amsterdam, 2013. Available at: http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf , accessed 03.02.2015.
2. Parker D. *Changing the World with Big Data. Real-time with Real Results*. IGEL (Initiative for Global Environmental Leadership) Wharton School of Business of University of Pennsylvania : website, November 2013. Available at: http://igel.wpengine.netdna-cdn.com/wp-content/uploads/2013/11/David-Parker.pdf , accessed 03.02.2015.
3. Financial Services. Datameer: company website. Available at: http://www.datameer.com/solutions/industries/financial-services.html , accessed 03.02.2015.
4. Beloussov S. Cloud computing is now IT. *Proc. of Parallels Summit 2014*. Parallels: company website. Available at: http://sp.parallels.com/fileadmin/media/hcap/events/summit/2014/documents/Summit2014-keynote-SergueiBeloussov.pdf , accessed 03.02.2015.
5. The Body as a Source of Big Data. The Institute for Health Technology Transformation: website. Available at: http://ihealthtran.com/images/Infographic-the-body-as-a-source-of-big-data-HealthIT-Infographic-NetApp-Infographic.pdf , accessed 03.02.2015.
6. Sustainability in the Age of Big Data. Special Report. IGEL (Initiative for Global Environmental Leadership) Wharton School of Business of University of Pennsylvania : website, September 2014. Available at: http://d1c25a6gwz7q5e.cloudfront.net/reports/2014-09-12-Sustainability-in-the-Age-of-Big-Data.pdf , accessed 03.02.2015.
7. A Focus on Efficiency. A whitepaper from Facebook, Ericsson and Qualcomm, September 16, 2013. Available at: https://fbcdn-dragon-a.akamaihd.net/hphotos-ak-prn1/851575_520797877991079_393255490_n.pdf , accessed 03.02.2015.
8. About Twitter, Inc. Twitter, Inc. : website. Available at: https://about.twitter.com/company , accessed 03.02.2015.
9. Asay M. *Why the world's largest Hadoop installation may soon become the norm*. Techrepublic: website, September 12, 2014. Available at: http://www.techrepublic.com/article/why-the-worlds-largest-hadoop-installation-may-soon-become-the-norm/ , accessed 03.02.2015.
10. Retail. Datameer: company website. Available at: http://www.datameer.com/solutions/industries/retail.html , accessed 03.02.2015.
11. Telecommunications. Datameer: company website. Available at: http://www.datameer.com/solutions/industries/telecommunications.html , accessed 03.02.2015.

12. van Rijmenam M. *Big Data Will Revolutionize Education*. DATAFLOO: company website, April 29, 2014. Available at: https://datafloq.com/read/big-data-will-revolutionize-learning/206 , accessed 03.02.2015.
13. Dalisay T. *Big Data in Education: Big Potential or Big Mistake*. Socialnomics: website, January 13, 2014. Available at: http://www.socialnomics.net/2014/03/05/big-data-in-education-big-potential-or-big-mistake/ , accessed 03.02.2015.
14. White T. *Hadoop: The Definitive Guide*. Third ed. O'Reilly Media / Yahoo Press, 2012. 688 p. (Russ. ed.: White T. *Hadoop. Podrobnoe rukovodstvo*. St. Petersburg, Piter Publ., 2013. 672 p.).
15. Emel'yanov I. *How an update of Hadoop 2.0 made Big Data easier for business users*. CIO: website, October 17, 2013. Available at: http://www.computerra.ru/cio/5598 , accessed 03.02.2015.
16. Weil S.A., Brandt S.A., Miller E.L, Long D.D.E., Maltzahn C. Ceph: A Scalable, High-Performance Distributed File System. *OSDI '06: 7th USENIX Symposium on Operating Systems Design and Implementation*, 2006, pp. 307-320. Available at: http://static.usenix.org/event/osdi06/tech/full_papers/weil/weil.pdf , accessed 03.02.2015.
17. Brim M.J., Dillow D.A, Oral S., Settlemyer B.W., Wang F. Asynchronous Object Storage with QoS for Scientific and Commercial Big Data. *8th Parallel Data Storage Workshop*, November 18, 2013, Denver, CO. Available at: http://www.pdsw.org/pdsw13/papers/p7-pdsw13-brim.pdf , accessed 03.02.2015.
18. Depardon B., Le Mahec G., Seguin C. *Analysis of Six Distributed File Systems*. Research Report HAL Id: hal-00789086, 2013. 44 p. Available at: https://hal.inria.fr/hal-00789086/PDF/a_survey_of_dfs.pdf , accessed 03.02.2015.
19. Donvito G., Marzulli G., Diacono D. Testing of several distributed file-systems (HDFS, Ceph and GlusterFS) for supporting the HEP experiments analysis. *Journal of Physics: Conference Series: 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)*. IOP Publishing, 2014, vol. 513, art. no. 042014. DOI: 10.1088/1742-6596/513/4/042014
20. Harris D. *Because Hadoop isn't perfect: 8 ways to replace HDFS*. GIGAOM: website, July 11, 2012. Available at: https://gigaom.com/2012/07/11/because-hadoop-isnt-perfect-8-ways-to-replace-hdfs/ , accessed 03.02.2015.
21. Kerzner M., Maniyam S. Chapter 12. Big Data Ecosystem. In: *Hadoop Illuminated*, 2014. Available at: http://hadoopilluminated.com/hadoop_illuminated/Bigdata_Ecosystem.html , accessed 03.02.2015.
22. Powlas T. *Big Data in an SAP Landscape – ASUG Webcast Part 1*. SAP Community Network: website, December 29, 2013. Available at: http://scn.sap.com/community/business-intelligence/blog/2013/12/29/big-data-in-an-sap-landscape-asug-webcast-part-1 , accessed 03.02.2015.

23. The Hadoop Ecosystem Table. GitHub: website. Available at: http://hadoopecosystemtable.github.io/ , accessed 03.02.2015.
24. Big Data Overview: IBM presentation at Big Data & Analytics Day seminar. Moscow, January 27, 2015. (unpublished).
25. SAP HANA® Database for Next-Generation Business Applications and Real-Time Analytics. Explore and Analyze Vast Quantities of Data from Virtually Any Source at the Speed of Thought. SAP AG, October 10, 2013. 18 p. Available at: http://www.slideshare.net/SAPMENA/hana-27077351 , accessed 03.02.2015.
26. SAP HANA Master Guide. SAP HANA Platform SPS 09. Document Version: 1.1 (2014-12-17). SAP AG. 84 p. Available at: http://help.sap.com/hana/SAP_HANA_Master_Guide_en.pdf , accessed 03.02.2015.
27. Arnold A. *In-memory Database Architecture & Landscape Options for a Real-time Business*. SAP AG, October 2014. 33 p. Available at: https://hcp.sap.com/content/dam/website/saphana/en_us/Technology%20Documents/SAP%20HANA%20Architecture%20and%20Landscape%20Options.pdf , accessed 03.02.2015.
28. Watts D., Krutov I. *xREF: IBM x86 Server Reference*. IBM, 2015. Available at: http://www.redbooks.ibm.com/redpapers/pdfs/redpxref.pdf , accessed 03.02.2015.
29. Hauff C. *Big Data Processing, 2014/15. Lecture 5: GFS & HDFS*. Delft University of Technology: website. Netherlands. Available at: http://www.st.ewi.tudelft.nl/~hauff/BDP-Lectures/5_filesystem_gfs_hdfs.pdf , accessed 03.02.2015.
30. Lovett M. *Apache Hadoop: The Open Source Elephant of Big Data*. Trenton Systems: company website, July 6, 2012. Available at: http://blog.trentonsystems.com/apache-hadoop-the-open-source-elephant-of-big-data/ , accessed 03.02.2015.
31. Popescu A., Bacalu A.-M. *Life of Data at Facebook*. myNoSQL : website, September 3, 2012. Available at: http://nosql.mypopescu.com/post/30815314471/life-of-data-at-facebook , accessed 03.02.2015.
32. Evans B., Graves T. *Storm and Spark at Yahoo: Why Chose One Over the Other*. yahoohadoop: website, September 18, 2014. Available at: http://yahoohadoop.tumblr.com/post/98213421641/storm-and-spark-at-yahoo-why-chose-one-over-the , accessed 03.02.2015.
33. Singh S. *Apache HBase at Yahoo! – Multi-tenancy at the Helm Again*. Yahoo Developer Network: website, June 7, 2013. Available at: https://developer.yahoo.com/blogs/hadoop/apache-hbase-yahoo-multi-tenancy-helm-again-171710422.html#more-id , accessed 03.02.2015.
34. Zicari R.V. *Hadoop at Yahoo. Interview with Mithun Radhakrishnan.* ODBMS Industry Watch: website, September 21, 2014. Available at: http://www.odbms.org/blog/2014/09/interview-mithun-radhakrishnan/ , accessed 03.02.2015.
35. Feng A., Evans B., Dagit D., Patil K., Poulosky P., Kapur D., Lal A. *The Evolution of Storm at Yahoo and Apache*. yahoohadoop : website, September 29, 2014. Available at:

http://yahoohadoop.tumblr.com/post/98751512631/the-evolution-of-storm-at-yahoo-and-apache , accessed 03.02.2015.

36. Lapan' M. Practical elephant industry: presentation. *Hadoop Kitchen seminar in Mail.ru Group*, September 27, 2014. Available at: http://www.youtube.com/watch?v=QAejaeTvj_M beginning from 1:20:40, accessed 03.02.2015.
37. Burdett D., Tripathi R. *CIO Guide. How to Use Hadoop with Your SAP® Software Landscape*. SAP AG, February 2013. 40 p. Available at: http://hortonworks.com/wp-content/uploads/2013/09/CIO.Guide_.How_.to_.Use_.Hadoop.with_.Your_.SAP_.Software.Landscape.pdf , accessed 03.02.2015.
38. Powlas T. *Using HANA and Hadoop, Key Scenarios - Part 2 ASUG Big Data Webcast*. SAP Community Network: website, December 29, 2013. Available at: http://scn.sap.com/community/business-intelligence/blog/2013/12/29/using-hana-and-hadoop-key-scenarios , accessed 03.02.2015.
39. Powlas T. *Fitting Hadoop in an SAP Software Landscape – Part 3 ASUG Webcast*. SAP Community Network: website, December 29, 2013. Available at: http://scn.sap.com/community/business-intelligence/blog/2013/12/29/fitting-hadoop-in-an-sap-software-landscape-part-3-asug-webcast , accessed 03.02.2015.
40. Eacrett M. *What is new in SAP HANA SPS 09*. SAP HANA: website, October 21, 2014. Available at: https://blogs.saphana.com/2014/10/21/what-is-new-in-sap-hana-sps-09/ , accessed 03.02.2015.
41. Hagman M. *Guinness World Record – Largest Data Warehouse*. SAP HANA: website, March 5, 2014. Available at: https://blogs.saphana.com/2014/03/05/guinness-world-record-largest-data-warehouse/ , accessed 03.02.2015.
42. Knuth D.E. *The Art of Computer Programming. Vol. 3. Sorting and Searching*. Second edition. Addison-Wesley Professional, 1998. 800 p. (Russ. ed.: Knuth D.E. *Iskusstvo programmirovaniya. T. 3. Sortirovka i poisk*. Moscow, Vil'yams Publ., 2012. 824 p.).
43. Voevodin V.V., Voevodin Vl.V. *Parallel'nye vychisleniya* [Concurrent calculatios]. St. Petersburg, BHV-Petersburg Publ., 2002. 602 p. (in Russian).
44. Analytics and Big Data and ROI…Oh My! BlueSpire : company website, September 16, 2014. Available at: http://www.slideshare.net/bluespiremarketing/analytics-and-big-data-and-roioh-my-trendlab-webinar , accessed 03.02.2015.
45. Kochetov U.A., Mladenovich N., Hansen P. Local search with alternating neighborhoods. *Diskretnyi analiz i issledovanie operatsii*, 2003, vol. 10, no. 1, pp. 11-43. (in Russian).
46. Braun H. Optimization with Grid Computing. *Workshop on Cyberinfrastructure (CI) in Chemical and Biological Process Systems: Impact and Directions*, September 25-26, 2006, Arlington, VA. Available at: https://smartmanufacturingcoalition.org/sites/default/files/optimization_with_grid_computing.pdf , accessed 03.02.2015.

47. Makarov V.L, Bahazin A.R., Roganov V.A, Trifonov I.A. Capacities of Supercomputer System for Work with Agent-Based Models. *Programnaya injeneriya = Software engineering*, 2011, no. 3, pp. 2-14. (in Russian).
48. Fegaras L. *Supporting Bulk Synchronous Parallelism in Map-Reduce Queries*. University of Texas at Arlington: website, 2012. Available at: http://lambda.uta.edu/mrql-bsp.pdf , accessed 03.02.2015.
49. Sukhobokov A. A., *Issledovanie i razrabotka modeley i arhitekturi sredstv kontrolinga dlya mezhregionalnih predpriyatiy v sostave system klassa ERP II. Kand. diss.* [Research and development of models and architecture of controlling tools in ERP II systems for multiregional enterprises. PhD dissertation]. Moscow, Bauman MSTU, 2009. 196 p. (In Russian).
50. Madjara T.I., *Intellektualnaya systema dlya resheniya zadach optimalnogo upravleniya s vichislitelnimy osobennostyamy. Kand. diss.* [Smart system for solving of optimal control problems with computational peculiarities. PhD dissertation]. Vladivostok, Institute of Automation and Control Processes. Far Eastern Branch of RAS, 2011. 149 p. (in Russian).