УДК 004.67

Засухина Ольга Александровна,
доцент кафедры «Электроснабжение промышленных предприятий»,
ФГБОУ ВО «Ангарский государственный технический университет»,
e-mail: olga_a_z @mail.ru
Ершов Егор Витальевич,
обучающийся группы ЭЭ-20-1,
ФГБОУ ВО «Ангарский государственный технический университет»,
e-mail: egormp3@mail.ru
Головатюков Леонид Константинович,
обучающийся группы ЭЭ-22-1,
ФГБОУ ВО «Ангарский государственный технический университет»,
e-mail: leonid.golovatiukov@mail.ru
Шитенков Григорий Александрович,
обучающийся группы ЭЭ-22-1,
ФГБОУ ВО «Ангарский государственный технический университет»,
e-mail: gregoryshitenkov@yandex.ru»

ТЕХНОЛОГИЯ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ (BIG DATA)

Zasukhina O.A, Ershov E.V., Golovatiukov L.K., Shitenkov G.A.

BIG DATA PROCESSING TECHNOLOGY

Аннотация. *Рассмотрены описание больших данных, технологии обработки больших данных (big data), системы хранения больших данных Hadoop.*

Ключевые слова: *большие данные, объем, разнообразие, скорость, noSQL, система хранения данных, бизнес-операции.*

Abstract. *The description of big data, big data processing technologies (big data), Hadoop big data storage systems are considered.*

Keywords: *big data, volume, variety, speed, NoSQL, data storage system, business operations.*

Одно из стремительно развивающихся направлений IT-технологий — это большие данные — Big Data. Основным способом описания больших данных является аббревиатура 3 V — volume, velocity, variety (рисунок 1) [1].

Объем (Volume). В самом простом определении большие данные — это данные, которые слишком велики для работы на компьютере. Однако данное определение — относительное, поскольку то, что является большим для одной системы, может быть пустяком для другой системы в другое время.

Это закон Мура — хорошо известное наблюдение в области информатики о том, что физическая емкость и производительность компьютеров удваиваются каждые два года. То, что занимало весь диск компьютера 10 лет назад легко помещается на флэшку сейчас. С другой стороны, размеры файла с современной кинокамеры достигают до 18 гигабайт в минуту и эти объемы проблема для обычного компьютера.

Рисунок 1 - Три преимущества больших данных

Скорость (Velocity). Скорость обработки тоже относительное понятие. Есть научные исследования, которые получают в течении продолжительных исследований, а затем заносятся и не меняются годами, но есть и другие данные, например, сообщения в социальных сетях — это десятки и сотни миллио-

нов строк в день. Даже съем данных температуры с прибора учета каждую миллисекунду приведет к большому потоку изменяемых данных, которые нужно успевать обработать в реальном времени.

Разнообразие (Variety). Первых два V это по сути обычные современные большие базы данных или Data Warehouse. И обработка таких данных уже традиционные и привычные технологии. Проблема возникает при добавлении третьей V — разнообразия данных. Здесь речь идет не только о строках и столбцах хорошо отформатированных данных. У вас может быть неструктурированный текст, например, книги и сообщения в блогах, а также комментарии к новостям и твитам. Исследования оценили, что 80 процентов корпоративных данных могут быть неструктурированными. Сюда также могут входить фотографии, видео и аудио. Исследование компании Forrester Research показывает, что разнообразие является важнейшим фактором для создания технологии Big Data. Фактически, при разговоре о Big Data, разнообразие упоминается в четыре раза чаще, чем объем данных или скорость.

Необходимость обработки больших данных привело к развитию новых технологий [2]. Хранение и обработка данных происходит в огромных кластерах объединенных вместе компьютеров. Такие кластеры могут насчитывать тысячи и даже десятки тысяч узлов. На сегодняшний день существуют множество Big Data-инструментов для анализа данных. Анализ данных представляет собой процесс проверки, очистки, преобразования и моделирования данных с целью получения полезной информации, выводов и обоснований для принятия решений.

Такие кластеры легко наращиваются (масштабирование вширь), позволяя решить проблему объемов хранения и увеличения вычислительных мощностей. Проблема хранения неструктурированных данных решается при помощи хранения первичных данных в виде файлов в специальной распределенной файловой системе (например, HDFS) или не реляционных базах данных (например, древовидных или сетевых). Такое хранение данных еще называют noSQL базы данных. Для запросов к таким данным разработаны языки запросов доступа и поиска — noSQL языки запросов. Это быстро развивающееся направление обработки данных. Здесь постоянно идет исследование и доработка математического аппарата и моделей. Пока еще нет общих стандартов, они находятся на стадии наработок и обсуждений. Слишком разнообразны способы хранения и виды хранимой информации. Кроме того, здесь могут быть и реляционные данные. И если изначально noSQL расшифровывалось как — не SQL, то сейчас под этим термином подразумевают — не только SQL. Еще одна проблема Big Data — преобразование данных. Технология ETL — это процесс транспортировки данных, при котором информацию из разных мест преобразуют и кладут в новое место. ETL расшифровывается как extract, transform, load, то есть «извлечь, трансформировать, загрузить». Один из основных процессов в управлении хранилищами данных, который включает в себя: извлечение данных из внешних источников; трансформация и проверка данных, чтобы они соответствовали потребностям бизнес-модели баз данных; загрузка их в хранилище данных. В Big Data изначально невозможно очистить, проверить и преобразовать данные, поэтому здесь применяется технология ELT. Данные извлекаются и загружаются все, а процесс трансформации и проверки на соответствие происходит при запросе к ним. Еще одним большим пластом науки и технологии Big Data, является развитие семантических анализаторов (СА). СА пытается вытянуть информацию по запросу из различных текстов. Этот раздел науки находится в непрерывном развитии. В настоящее время такие анализаторы есть только для самых распространенных языков в мире. Для английского языка анализаторы наиболее отработаны. Достоверность их распознавания достигает 80–90 %, для русского 60–70 %. Ученые говорят, что необходимо достижение рубежа распознавания в 1–2 % ошибочной информации. Еще Big Data активно использует самообучающиеся автоматы — программы, которые в ходе своей работы на основе множественных данных учатся составлять оптимальные алгоритмы поиска и нахождения решения. После определенного времени работы такой программы, даже ее разработчику почти невозможно разобрать как достигнуто программой то или иное конечное решение.

Одна из самых распространённых в настоящее время технологий — фреймворк Hadoop — проект фонда Apache Software Foundation [3]. Apache Hadoop занимает первое место в списке. Большие данные будет

сложно обрабатывать без Hadoop, и специалисты по данным хорошо это знают. Hadoop — это не только полностью открытая и бесплатная система хранения больших данных, но и сопутствующий набор утилит, библиотек, фреймворков, дистрибутивов для разработки. Эта основополагающая технология хранения и обработки больших данных является проектом верхнего уровня Apache Software Foundation. Hadoop состоит из четырех частей:

- HDFS - распределенная файловая система, предназначенная для работы на стандартном оборудовании;
- MapReduce - модель распределённых вычислений, представленная компанией Google, используемая для параллельных вычислений;
- YARN - технология, предназначенная для управления кластерами;
- библиотеки для работы остальных модулей с HDFS.

KNIME Analytics Platform ведущий open source фреймворк для инноваций, зависящих от данных. Он помогает раскрыть скрытый потенциал, найти новые свежие идеи, или предсказать будущие тенденции. KNIME Analytics Platform содержит в себе более 1000 модулей, сотни готовых к запуску примеров, широкий спектр интегрированных инструментов и широкий выбор современных доступных алгоритмов, определённо, это идеальный набор инструментов для любого специалиста в data science. OpenRefine (ранее Google Refine) мощный инструмент для работы с сырыми данными: их очистки, преобразования из одного формата в другой и расширения с помощью веб-сервисов и внешних данных. OpenRefine помогает с легкостью исследовать большие наборы данных; R-Programming открытая среда программирования для статистических вычислений и графики. Язык R широко используется среди майнеров данных для разработки статистического программного обеспечения и анализа данных. Простота его использования и расширяемость значительно повысили популярность R в последние годы. Помимо интеллектуального анализа данных, он предоставляет статистические и графические методы анализа, включая линейное и нелинейное моделирование, классические статистические тесты, анализ временных рядов, классификацию, кластеризацию. Существует широкий спектр инструментов для работы с большими данными, которые помогают хранить, анализировать, составлять отчеты и делать с данными намного больше. Это программное обеспечение превращает скудные биты данных в мощное топливо, которое стимулирует глобальные бизнес-процессы и способствует принятию решений, основанных на знаниях. Когда-то использование больших данных произвело революцию в области информационных технологий. Сегодня компании используют ценные данные и внедряют инструменты больших данных, чтобы превзойти своих конкурентов. На конкурентном рынке как устоявшиеся компании, так и новички применяют стратегии, опираясь на обработанные данные, чтобы зафиксировать сигнал, отследить пожар и получить прибыль. Большие данные позволяют организациям определять новые возможности и создавать новые типы компаний, которые могут комбинировать и анализировать отраслевые данные.

Таким образом, чистые, актуальные и наглядные данные предоставляют полезную информацию о продуктах, оптимизируют бизнес-операции и влекут за собой значительные экономические преимущества.

СПИСОК ЛИТЕРАТУРЫ

1. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. META group Inc., 2001.

2. Фрэнк Б. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики: Альпина Паблишер — 2017, 320 с.

3. Самойлова, И. А. Технологии обработки больших данных / И. А. Самойлова. — Текст : непосредственный // Молодой ученый. — 2017. — № 49 (183). — С. 26-28. — URL: https://moluch.ru/archive/183/46957/ (дата обращения: 26.10.2022).