

УДК 519.6+519.7

М. В. Гладкий, асп.; П. П. Урбанович проф. (БГТУ, г. Минск)

РЕВОЛЮЦИЯ В ОБЛАСТИ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ (BIG DATA)

В настоящее время наблюдается значительный рост объемов информации, поступающей из самых различных источников (мобильные устройства, социальные сети, финансовые операции и многое другое). Мировой объем данных увеличивается более чем в два раза каждые два года, что приводит к открытию новых возможностей в области ИТ и изменяет ее роль в целом. Имеет смысл говорить не просто о модернизации привычных представлений о системах хранения и обработки информации, но и о новом витке информационной революции [1], основанной на методах и технологиях эффективной обработки, анализа и управления большими объемами данных, в том числе и в режиме реального времени.

Под термином «большие данные» (BigData) понимают совокупность технологий и методов производительной обработки динамически растущих объемов данных в распределенных информационных системах, обеспечивающих организацию качественно новой полезной информацией.

Для технологий BigData характерны пять признаков [2]:

- volume — объём данных считается большим, когда возникают затруднения при обработке этого объёма средствами традиционных СУБД (более 1 Петабайта);

- velocity — скорость обработки данных, в том числе и в режиме реального времени;

- variety — информация больших объемов очень редко бывает однородной. В подавляющем большинстве случаев общий массив данных включает как структурированные, так и неструктурированные данные;

- veracity — способность выделить и отсеять полезную составляющую данных от информационного «шума»;

- value — ценность информации определяет целесообразность её обработки. Собираемые данные должны давать ответы на предварительно сформулированные и вновь появляющиеся вопросы.

Выделяют два основных класса средств в области обработки и анализа больших данных: BusinessIntelligence (BI) и DataMining. К первому классу относят методы и инструменты для перевода необработанной информации в осмысленную, удобную форму и применяются в основном в тех случаях, когда необходимо выполнить:

- денормализацию баз данных;

- многомерную агрегацию данных;
- групповую консолидацию и бюджетирование;
- статистические выводы и вероятное моделирование;
- оптимизацию ключевых показателей эффективности;
- контроль версий и управление процессами.

Ко второму классу относят совокупность методов и средств обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Задачи, решаемые методами DataMining, принято разделять на описательные (наглядное описание имеющихся скрытых закономерностей) и предсказательные (предсказания в тех случаях, когда данных еще нет) [2]. При решении задач этих двух типов используют алгоритмы и методы классификации, кластеризации, машинного обучения и имитационного моделирования, основанные на применении корреляционного, сетевого и регрессивного анализа, деревьев решений, нейронных сетей, генетических алгоритмов, A/B-тестирования, краудсорсинга, поиска ассоциативных правил.

Технологии, используемые для сбора и хранения больших данных, можно разделить на 3 группы: программное обеспечение (ПО), оборудование и сервисные услуги. К ПО относят NoSQL решения, парадигму распределённых вычислений MapReduce, Фреймворк Apache Hadoop, NewSQL платформы SAP HANA и IBM Bluemix [1].

К технологическому оборудованию относят серверы и инфраструктурное оборудование. Серверы включают в себя хранилища данных, а к инфраструктурному оборудованию относят средства ускорения платформ, источники бесперебойного питания, комплекты серверных консолей и др.

Сервисные услуги включают в себя услуги по построению архитектуры системы базы данных, обустройству и оптимизации инфраструктуры, и обеспечению безопасности хранения данных.

ЛИТЕРАТУРА

1. Натан, М. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. Москва: Вильямс, 2015.

2. Сухобоков, А. А. Влияние инструментария BigData на развитие научных дисциплин, связанных с моделированием. Москва: МГТУ им. Н.Э. Баумана, 2015.