УДК 004.82

В.С. Кавицкая, аспирант,
В.В. Любченко, д-р. техн. наук, проф.,
Одес. нац. политехн. ун-т

ОСНОВНЫЕ ИНСТРУМЕНТЫ ДЛЯ АНАЛИЗА BIG DATA

Розглядаються проблеми аналізу великих даних. Пропонується аналіз основних інструментів для аналізу великих даних. Визначаються переваги і недоліки основних інструментів.
Ключові слова: аналіз великих даних; інструменти аналізу великих даних; великі дані.

Рассматриваются проблемы анализа больших данных. Предлагается анализ основных инструментов для анализа больших данных. Определяются основные достоинства и недостатки инструментов.
Ключевые слова: анализ больших данных; инструменты анализа больших данных; большие данные.

The problems of analyzing big data are considered. Analysis of the main tools for analysis of big data is offered. The main strengths and weaknesses of the tools are identified.
Keywords: analysis of big data; big data analysis tools; big data.

Все больше организаций испытывают те или иные сложности при развитии проектов, связанных с большими данными (Big Data), а именно [1]**:**

— недостаток инфраструктурных мощностей;

— организационные сложности по внедрению новых подходов и сбору данных;

— обеспечение безопасности и соответствия стандартам;

— нехватка ресурсов.

Самой критичной проблемой является недостаток инфраструктурных мощностей, так как каждый инструмент обладает существенным рядом как достоинств так и недостатков для анализа Big Data. Предлагается анализ основных инструментов.

Рассмотрим основные инструменты для анализа Big Data, такие как:

— **Hadoop MapReduce** — открытая реализация на Java для Apache Hadoop.

— Реализации для NoSQL баз данных на примере **MongoDB.**

— Apache Spark — платформа для обработки больших данных.

— Онтологический подход.

Hadoop MapReduce — программная модель (framework) выполнения распределенных вычислений для больших объемов данных в рамках парадигмы map/reduce, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки заданий на параллельную обработку [2]**.**

Основные концепции Hadoop MapReduce можно сформулировать как:

— обработка/вычисление больших объемов данных;

— масштабируемость;

— автоматическое распараллеливание заданий; для анализа Big Data**:**

— применение MapReduce по производительности менее эффективно, чем специализированные решения;

— эффективность применение MapReduce снижается при малом количество машин в кластере;

— невозможно предсказать окончание стадии map;

— этап свертки не начинается до окончания стадии map.

MongoDB — это база данных документов с открытым исходным кодом, и ведущая база данных NoSQL. **Рассмотрим основные концепции MongoDB.**

— использование гибкой структуры данных;

— масштабируемость;

— размещение как на локальном оборудовании, так и в облаке;

— применение свойства ACID (atomicity, consistency, isolation, durability; атомарность, согласованность, изолированность, долговечность) на уровне документа.

MongoDB обладает следующими недостатками для анализа Big Data:

— неэффективная обработка больших объемов данных;

— низкая производительность при выполнении запросов map-reduce;

— низкая аналитическая мощность — необходима интеграция MongoDB и Hadoop.

Apache Spark использует модель для организации распределенных вычислений, основанную на понятии устойчивой к сбоям распределенной коллекции данных (RDD). Apache Spark обладает следующими преимуществами [3]:

— повышение отказоустойчивости и повышение скорости обработки информации в сравнении с MapReduce;

— эффективное выполнение итеративных алгоритмов за счет поддержки кэширования результатов в памяти.

— расширение технологий Hadoop его применения.

Необходимо отметить, что наличие огромного объема сырых данных еще не гарантирует высокие аналитические способности применяемых инструментов, так как при принятии решений необходимо еще строить гипотезы и теории, в соответствии с которыми могут развиваться события, т.е. сырые данные неразрывно должны рассматриваться с предметной областью. Вышеперечисленные инструменты не обладают данным свойством, с чем отлично справляется онтологический подход.

Онтология обладает следующими концепциями [4]:

— Систематичность — онтология представляет целостный взгляд на предметную область.

— Единообразность — материал, представленный в единой форме гораздо лучше воспринимается и воспроизводится.

— Научность — построение онтологии позволяет восстановить недостающие логические связи во всей их полноте.

Основными преимуществами применения онтологического подхода для анализа Big Data можно выделить:

— предоставление системного подхода в конкретной предметной области;

— четкое структурирование информации в конкретной предметной области;

— повышение эффективности информационного поиска за счет уменьшения времени обработки запроса;

— нахождение, определение и восстановление отсутствующих или скрытых логических связей.

Таким образом, были рассмотрены основные инструменты для анализа Big Data. Из проведенного анализа следует, что задачи, решаемые с помощью Hadoop MapReduce, должны отвечать одному основному требованию — они должны относиться к задачам, параллельным по данным. Реализации для NoSQL баз данных обладают низкой аналитической мощностью. Apache Spark является хорошим инструментом для анализа Big Data, однако он не учитывает все особенности предметной области. Применение онтологии для анализа позволяет решить проблему четкого структурирования и систематизации данных в конкретной предметной области, тем самым повышая эффективность выполнения запросов, своевременного предоставления информации при анализе Big Data.

Литература

1. Pettey C. Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data / C. Pettey, L. Goasduff // Gartner, 2011. — P. 11 — 21.
2. White, T. Hadoop: The Definitive Guide / T. White // O'Reilly Media, Inc, 2012. — P. 657.
3. Cluster Computing with Working Sets / M. Zaharia, M. Chowdhury, M. Franklin and other // HotCloud, 2010. — P. 1 — 7.
4. Gavrilova, T. Practical Design Of Business Enterprise Ontologies / T. Gavrilova, D. Laird // In Industrial Applications of Semantic Web. — Springer, 2005. — P. 61 — 81.