ТЕХНОЛОГИИ IBM ДЛЯ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Козлов А. В.

Лещёв А. Е. – ст. преп. каф. информатики

Термин “Big Data” определяет технологии и методы сбора и обработки структурированных и неструктурированных данных больших объёмов с целью получения воспринимаемых человеком результатов. Сюда, как правило, включают инструменты обработки данных в параллельном режиме, NoSQL-решения, MapReduce-алгоритмы, а также библиотеками проекта Hadoop.

Компания IBM предоставляет удобную платформу “big data platform” для обработки больших данных. Основными продуктами этой платформы являются продукты IBM InfoSphere BigInsights и IBM InfoSphere Streams.

Программное обеспечение IBM InfoSphere BigInsights предоставляет инструменты анализа больших данных на корпоративной платформе. Платформа поддерживает структурированные, полуструктурированные и неструктурированные данные и предоставляет широкий спектр инструментов для их обработки. ПО IBM InfoSphere BigInsights построено на базе популярных открытых продуктов от apache foundation для работы с большими данными, например, Hadoop, HBase, Zookeeper, Oozie и пр. Кроме того, в состав продукта включены и специализированные технологии от IBM, предоставляющие удобные средства для анализа больших данных – это такие технологии, как BigSQL, BigSheets и TextAnalytics.

Одной из задач, решаемой с помощью технологий Big Data, кроме обработки огромных массивов данных, является обработка данных в реальном времени. К таким задачам можно отнести, например, задачу оптимизации движения транспорта, когда в зависимости от количества автомобилей, ожидающих разрешающий сигнал для дальнейшего движения, система управляет работой светофоров. В задачах такого класса, очень важна скорость обработки данных, ведь такие данные имеют строго ограниченную область жизни, например, в случае со светофорами, такие данные уже через несколько минут станут бесполезными. Для обработки данных в real-time режиме, IBM предоставляет отдельный продукт. IBM InfoSphere Streams - это платформа для анализа больших данных в реалтайм-режиме, которая позволяет быстро принимать, анализировать информацию в режиме реального времени, как только она поступает из нескольких сотен разнообразных источников, например, датчиков, социальных сетей, интернет-служб в структурированном и неструктурированном виде. Важной отличительной характеристикой этой платформы является способность обрабатывать данные с очень высокой скоростью и большими объемами — порядка нескольких миллионов сообщений в секунду. InfoSphere Streams обладает следующими удобными возможностями:

- Анализ данных "в движении / на лету" — работает в режиме реального времени и позволяет добиться времени отклики менее миллисекунды, позволяя анализировать данные по мере возникновения.
- Легко интегрируется с корпоративными платформами, и позволяет анализировать широкий спектр данных, как структурированных, так и не структурированных
- Содержит большое число адаптеров, для приема данных в разнообразном виде, например, текста, аудиоданных, изображений, видео, электронной почты, Web-трафика, данных GPS, спутниковых данных, данных о финансовых транзакциях, показаний датчиков и пр.
- Содержит богатые наборы инструментов и акселераторы для выполнения расширенного анализа, например акселератор для работы с событиями телекоммуникаций, позволяющий анализировать большие объемы streaming-данных от систем телекоммуникаций практически в реальном времени, а также акселераторы получения данных для анализа социальных сетей.
- Занимается аспределением программы по обрабатывающим узлам кластера для анализа нескольких миллионов сообщений в секунду, со времением отклика менее миллисекунды.
- Содержит набор инструментов, позволяющих осуществлять фильтрацию для выделения только значимых данных из огромных объемов данных, что позволяет снизить затраты на хранение данных.
- Может быть расширение с одного сервера для нескольких сотен серверов в зависимости от требоний по данных и времени отклика, при этом позволяя добавлять или удалять узлы в графическом режиме, не обладая навыками администрирования Unix-систем.
- Предоставляет инструменты для обеспечения безопасности доступа к дынным и системе в целом
- Обладает возможностью адаптируемости к быстро меняющимся типам данных.

Список использованных источников:

1. Платформа IBM для работы с большими данными. URL: http://www-03.ibm.com/software/products/ru/category/bigdata
2. IBM InfoSphere BigInsights. URL: http://www-03.ibm.com/software/products/ru/infobigienteedit