ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ SAP HANA И MAPREDUCE ДЛЯ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ

Ачкасов И.В.

Харьковский национальный университет радиоэлектроники

Термин BigData начал активно использоваться приблизительно в 2011 году. BigData это не только большой объем хранения данных, но так же методы для оптимизации и работы с этими данными. Методы, которые используются для небольших наборов данных, уже перестают работать корректно в этой отрасли. Они должны уметь обрабатывать огромные наборы данных, например содержание всех страниц в интернете.

Основными принципами работы с BigData являются:

1. Горизонтальная масштабируемость. Если данные увеличились в несколько раз, увеличили на столько же количество железа, и все продолжает работать с той же мощностью.
2. Отказоустойчивость. Свойство работать без сбоев при выходе из строя одного из узлов системы.
3. Локальность данных. По возможности, данные обрабатываются на том же сервере, на котором и хранятся, для ускорения работы за счет исключения времени, потраченного на передачу файлов между машинами.

Самое популярное программное обеспечение для работы с большими данными:

1. NoSQL – «не только SQL». Набор подходов для реализации и работы с нереляционными базами данных.
2. MapReduce – модель распределения вычислений, применяющаяся для обработки данных размерами в ПБ и более. Представлена компанией «Google».
3. Hadoop – фреймворк, применяется для поиска и управления данными на высоконагруженных сайтах.
4. SAP HANA – высокопроизводительная NewSQL платформа для хранения и обработки данных.

Результаты опроса компанией T-Systems показали, что 30% опрошенных компаний выбрали технологии in-memory (SAP HANA), 18% выбрали NoSQL, 15% выбрали аналитические платформы, например компаний Splunk и Dell и менее популярными оказались Hadoop/MapReduce.

Рассмотрим систему SAP HANA. В ней находится индексный сервер, который управляет авторизацией, сеансами, транзакциями и обрабатывает команды. Также он хранит соответствия между файлами в ROM и их кэшированными образами в RAM. HANA поддерживает как строчное, так и поколоночное хранение данных, которое предоставляет больше возможностей.

Клиентские приложения получают доступ к базе данных HANA непосредственно с использованием JDBC, либо через подсистему Extended Services (XS) с использованием HTTP. JDBC— соединение с базами данных

на Java. Для доступа через XS используются сервлеты Java или приложения на JavaScript, находящиеся на стороне сервера.

Для преодоления ограничений, связанных с вводом-выводом SAP HANA была построена на основе сервера для вычислений по технологии in-memory. Это значит, что первый доступ к таблице вызывает необходимость чтения и поддержки всей таблицы в памяти. Поддержку файлов журналов и долговременное хранение данных на диске обеспечивают процессы фонового режима. Хранение данных по столбцам сокращает объем требуемых операций считывания и устраняет необходимость индексирования данных.

Приложения могут действовать в обход процессора SQL, получая непосредственный доступ к подсистеме вычислений с помощью запросов на основе XML. Существует три типа non-SQL объектов: Attribute Views, Calculation Views и Analytic Views. Во многих случаях использование этих объектов вместо запросов SQL позволяет улучшить характеристики производительности приложений.

Для работы с данными, которые могут быть разложены на пары ключ - значение, без риска при этом потерять контекст или какие-либо неявные взаимосвязи, можно использовать представленный компанией Google фреймворк MapReduce. Неявные взаимосвязи есть в графах (ребра, поддеревья, дочерние и родительские отношения, веса и т.п.), причем далеко не все такие взаимосвязи могут существовать на конкретном узле. Поэтому большинство алгоритмов для работы с графами требуют полной или частичной обработки графа при каждой итерации. В MapReduce это зачастую невозможно или очень сложно сделать.

MapReduce предполагает, что данные организованы в виде некоторых записей. Обработка данных происходит в 3 стадии:

1. Стадия Map. Происходит фильтрация данных. На выходе получаются множество пар в виде ключ - значение. Функции map проходят в разных потоках, то есть могут выполняться независимо и параллельно.
2. Стадия Shuffle. Данные, полученные на выходе функции map, раскладываются по категориям (корзинам). Так же как и map, выполняются в разных потоках.
3. Стадия Reduce. Каждая корзина соединяется с другой корзиной такого же типа. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce [Электронный ресурс],-

https://habrahabr.ru/company/dca/blog/267361/

SAP HANA [Электронный ресурс],-

https://ru.wikipedia.org/wiki/SAP_HANA