ИСПОЛЬЗОВАНИЕ APACHE HADOOP ДЛЯ ОБРАБОТКИ БОЛЬШИХ НАБОРОВ ДАННЫХ

Бабич А.С., Черных Е.П.
Национальный технический университет «Харьковский политехнический институт», г. Харьков

Так как в последнее время тема больших данных приобретает все большую и большую популярность, появляется необходимость поиска решений для хранения и обработки этих данных. Apache Hadoop – является одним из решений.

Apache Hadoop представляет собой набор алгоритмов (фреймворк с открытым кодом, написанный на Java) для распределенного хранения и распределенной обработки очень больших наборов данных (Big Data) на вычислительных кластерах, построенных из стандартных аппаратных средств. Все модули в Hadoop разработаны с фундаментальным предположением, что аппаратные сбои – это обычное явление и, следовательно, должны обрабатываться автоматически в рамках программного обеспечения.

База фреймворка Apache Hadoop состоит из следующих модулей:

- Hadoop Common – содержит библиотеки и утилиты, необходимые для других модулей Hadoop;
- Hadoop Distributed File System (HDFS) – распределенная файловая система, которая хранит данные на машинах, обеспечивая очень высокую суммарную пропускную способность в пределах кластера;
- Hadoop YARN – отвечает за управление вычислительными ресурсами в кластерах и использование их для планирования заявок пользователей;
- Hadoop MapReduce – модель программирования для обработки больших данных. Подход локализации данных позволяет обрабатывать их быстрее и более эффективно с помощью распределенных вычислений, чем при использовании более мощной суперкомпьютерной архитектуры, которая опирается на параллельную файловую систему, где вычисления и данные связаны через высокоскоростные сети. Преимущество MapReduce заключается в том, что он позволяет распределено производить операции предварительной обработки и свертки. Операции предварительной обработки работают независимо друг от друга и могут производиться параллельно. Аналогично множество рабочих узлов могут осуществлять свертку. Для этого необходимо, чтобы все результаты предварительной обработки с одним конкретным значением ключа обрабатывались одним рабочим узлом в один момент времени.

Одной из основных целей Hadoop изначально было обеспечение горизонтальной масштабируемости кластера посредством добавления недорогих узлов без прибегания к мощным серверам и дорогим сетям хранения данных. Функционирующие кластеры размером в тысячи узлов подтверждают осуществимость и экономическую эффективность таких систем.

Таким образом, Apache Hadoop является одним из лучших решений для хранения и обработки больших данных.