Chapter 9
Big Data Techniques, Tools, and Applications

Yushi Shen
Microsoft Corporation, USA

Ling Wu
EMC² Corporation, USA

Yale Li
Microsoft Corporation, USA

Shaofeng Liu
Microsoft Corporation, USA

Qian Wen
Endronic Corp, USA

ABSTRACT

This chapter covers big data technologies and tools, including the NoSQL database, HDFS, MapReduce, SMAQ stack, and the Hadoop Ecosystem. It also introduces the appliance products that help the customer for their big data analytics.

INTRODUCTION

In the era of big data, data are counted by GB-TB-PB. Data must be distributed among many machines. Building a useful big data database is an extremely complex process. How do you select only data that is appropriate from available data sources? How can you remove duplicated, corrupted or meaningless data, and convert the remainder to useful information for business insights? How do you store the lot, so that the user can handle the volume and variety with limited resources?

New approaches to processing and analyzing big data have some common characteristics. They take advantage of commodity hardware, to enable scaled-out and parallel-processing techniques, use non-relational data storage capacity to process

DOI: 10.4018/978-1-4666-4801-2.ch009

unstructured data, and apply advanced analytics and data visualization technologies, to convey insights to businesses.

BIG DATA TECHNOLOGIES

Hadoop is an open source framework for processing, storing and analyzing huge amounts of unstructured data. The fundamental concept is to break Big Data into multiple smaller data sets, so each data set can be processed and analyzed in parallel. Hadoop is best for large, but relatively simple database filtering, sorting, converting and analysis. (Wikipedia on Apache Hadoop)

The Hadoop ecosystem is made up of a number of complimentary sub-projects. Here is a list of Hadoop components (Apache Software Foundation, 2013):

- Hadoop Distributed Filesystem (HDFS), which creates replicas of data blocks, and distributes data on computer nodes over the cluster (Borthakur, 2013);
- MapReduce - MapReduce divides jobs into two parts. The "Map" function divides a query into multiple jobs, and the "Reduce" function combines the results to form the output (Hadoop – MapReduce Tutorial, 2013);
- HBase is a Hadoop database that provides random, real-time read and write access to HDFS;
- Hive is an analysis tool: it uses a SQL like syntax to rapidly develop queries. Mostly used for offline batch processing, ad-hoc querying and statistical analysis of large data warehouse systems;
- Mahout is a framework for deploying many machine learning algorithms on large datasets, mostly used in clustering, classification and text mining.
- Pig is the platform that analyzes large data sets. The Pig structure is amenable to substantial parallelization, so as to effectively handle very large volumes of data sets. Pig uses a language called Pig Latin, and has the characteristics of easy programming, auto optimization and extensibility;
- OOzie is an open source workflow scheduler system to manage Apache Hadoop data processing jobs. Oozie workflow consists of actions and dependencies. Users create Directed Acyclical Graphs (DAG) to model workflow. Oozie manages the dependencies at runtime, and executes the actions when the dependencies identified in the DAG are satisfied. Yahoo!'s workflow engine uses OoZie to manage jobs running on Hadoop (Yahoo!, 2010);
- ZooKeeper is a centralized service, which enables highly reliable distributed coordination. It maintains configuration information, provides distributed synchronization and group services for distributed applications;
- Flume is a distributed system that brings data into HDFS. The Apache Flume website describes Flume as "a distributed, reliable and available service for efficiently collecting, aggregating and moving large amounts of log data. It enables applications to collect data from its origin and send it to the HDFS;"
- HCatalog provides table management and storage management for data created using Hadoop. HCatalog provides a shared schema and data type mechanism, can interoperate across data processing tools such as Pig, Hive and MapReduce.
- BigTop is a project for packaging and testing the Hadoop ecosystem. It puts 100% open source apache Hadoop big data stack together, including Hadoop, Hbase, Hive, Mahout, flume and etc. This full stack of components provide the user a complete data collection and analytics pipeline (Apache Incubator PMC).