BY

SRIJHA REDDY GANGIDI
What is Hadoop?

- Open-source Software Framework
- Running Big Data Applications
- Cost-Effective at Handling unstructured data sets

- It includes distributed file system, a parallel processing framework and other components that support data ingestion, co-ordination of workflow, management of jobs and monitoring of the clusters.
Evolution of Hadoop

Created by dough cutting, a part of Apache project.
Hadoop Architecture
**Ambari**

- Ambari offers a Web-based GUI with wizard scripts for setting up clusters with most of the standard components.
- Ambari will help you provision, manage, and monitor a cluster of Hadoop jobs.

**Provision a Hadoop Cluster**
- Ambari provides a step-by-step wizard for installing Hadoop services across any number of hosts.
- Ambari handles configuration of Hadoop services for the cluster.

**Manage a Hadoop Cluster**
- Ambari provides central management for starting, stopping, and reconfiguring Hadoop services across the entire cluster.

**Monitor a Hadoop Cluster**
- Ambari provides a dashboard for monitoring health and status of the Hadoop cluster.
- Ambari leverages Ambari Metrics Systems for metrics collection.
- Ambari leverages Ambari Alert Framework for system alerting and will notify you when your attention is needed (e.g., a node goes down, remaining disk space is low, etc.).
HDFS (Hadoop Distributed File System)

- The Hadoop Distributed File System offers a basic framework for splitting up data collections between multiple nodes while using replication to recover from node failure.
- The large files are broken into blocks, and several nodes may hold all of the blocks from a file.
- HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a wide variety of concurrent data access applications.
HBase (Database)

When the data falls into a big table, HBase will store it, search it, and automatically share the table across multiple nodes so that MapReduce jobs can run locally. It runs top of HDFS.

HBase provides you with the following:

1. Low latency access to small amounts of data from within a large data set.
2. Flexible data model to work with and data is indexed by the row key.
3. Fast scans across tables.
4. Scale in terms of writes as well as total volume of data.
MapReduce

- It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster.
- The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
- The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples.
Hive is designed to regularize the process of extracting bits from all of the files in HBase.

- It offers an SQL-like language that will dive into the files and pull out the snippets your code needs.
- The data arrives in standard formats, and Hive turns it into a stash for querying.
Pig basically has 2 parts:
1. **PigLatin**: Write Pig script in PigLatin.
2. **Pig Interpreter**: Process the script using the interpreter.

Pig is recommended for people familiar with scripting languages like Python.
R(Statistics)

- Perform statistical analysis on data
- Provides elastic data analytics platform that will scale depending on the size of data set to be analyzed.
- Programmers can write MapReduce modules in R and run it using Hadoop parallel processing MapReduce mechanism to identify patterns in the datasets.

R + Hadoop = Rhadoop

Rhadoop Packages:
- ravro - read and write files in avro format
- plyrmr - higher level plyr-like data processing for structured data, powered by rmr
- rmr - functions providing Hadoop MapReduce functionality in R
- rhdfs - functions providing file management of the HDFS from within R
- rhbase - functions providing database management for the HBase distributed database from within R

R allows performing Data Analytics by various statistical and machine learning operations as follows:
- Regression
- Clustering
- Classification
- Recommendation
- Text Mining

R and Hadoop are a natural match in big data analytics and visualization.
Mahout (Machine Learning)

Apache Mahout is an open source project that is primarily used for creating scalable machine learning algorithms. Mahout is a project designed to bring implementations of algorithms for data analysis, classification, and filtering, to Hadoop clusters. It implements popular machine learning techniques such as:

- Recommendation
- Classification
- Clustering

- The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment.
- It includes several MapReduce enabled clustering implementations
Sqoop (Relational Database Collector)

- Efficiently transferring bulk data between Hadoop and Structured Relational Databases.
- Abbreviation for (Sq)L-to-Hado(op).

- Sqoop is a command-line tool that controls the mapping between the tables and the data storage layer, translating the tables into a configurable combination for HDFS, HBase, or Hive.
Flume/Chukwa (Log Data Collector)

Flume and Chukwa share similar goals and features.

Flume is distributed system for collecting log data from many sources, aggregating it, and writing it to HDFS. Flume maintains a central list of ongoing data flows, stored redundantly in Zookeeper. Flume adopts a “hop-by-hop” model.

Log processing with MapReduce has been difficult with Hadoop. Chukwa is a Hadoop subproject that bridges that gap between log handling and MapReduce. Chukwa distributes this information more broadly among its services. In Chukwa the agents on each machine are responsible for deciding what data to send.
ZooKeeper (Centralized co-ordination service)

It is a centralized service to maintain configuration information, naming, providing distributed synchronization and group services, which are useful for variety of distributed systems.

- ZooKeeper imposes a file system like hierarchy on the cluster and stores all of the metadata for the machines so that you can synchronize the work of the various machines.

- ZooKeeper allows distributed processes to coordinate with each other through a shared hierarchal namespace which is organized similarly to a standard file system. The name space consists of data registers - called znodes.

- A distributed HBase setup depends on a running ZooKeeper cluster. HBase by default manages a ZooKeeper cluster
Oozie (Workflow Scheduler System)

Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages – such as Map Reduce, Pig and Hive -- then intelligently link them to one another.

Oozie manages a workflow specified as a DAG (directed acyclic graph)
Conclusion

- Hadoop can handle large volumes of structured and unstructured data more efficiently than the traditional enterprise data warehouse.

- Hadoop has a robust Apache community behind it that continues to contribute to its advancement.

- All the modules in Hadoop are designed with a fundamental assumption that hardware failures are commonplace and thus should be automatically handled in software by the framework.
Reference

http://hortonworks.com/hadoop/hdfs/
http://www.edureka.co/blog/pig-programming-create-your-first-apache-pig-script/
https://cwiki.apache.org/confluence/display/Hive/Design
http://wikibon.org/wiki/v/HBase,_Sqoop,_Flume_and_More:_Apache_Hadoop_DEFINED