HADOOP VS HBASE: WHAT'S THE DIFFERENCE?

Hadoop is a group of open source software projects that are overseen by The Apache Software Foundation (ASF). These projects make big data a lot easier to manage by providing a framework for

storing and processing massive amounts of data across networked clusters of low-cost hardware without compromising reliability. Instead of relying on expensive, high-end hardware for resilience, the software is able to detect and handle failures at the application layer with reliability strategies and data replication. Hadoop can expand to accommodate more data by adding nodes, and parallel processing capabilities mean quick results, even when processing huge amounts of data.

The core components of the Hadoop stack include:

- **Hadoop Distributed Filesystem (HDFS)**

Storage component - infinitely scalable and flexible to enable storage and analysis of unlimited amounts and types of data, across clusters of industry-standard commodity hardware in an easily accessible format.

- **MapReduce**

Processing component - unique approach moves the processing software to the data instead of moving huge data sets over the network to be processed by software. Processing occurs in two phases: the map phase and the reduce phase. In the map phase, input is divided into smaller sub-problems and processed. In the reduce phase, the answers from the map phase are collected and combined to form the output of the original, bigger problem.

- **YARN**

Short for *Yet Another Resource Negotiator*, YARN manages computing resources such as CPU and memory, including the scheduling of resource requests.

The latest releases can be downloaded directly from hadoop.apache.org, but most enterprise users purchase commercial distributions from vendors. The commercial offerings bundle the software with maintenance, support, and other enhancements.

HBase is a supporting component in the Hadoop group of open source projects, a non-relational database that is the de facto standard for working with Hadoop and is frequently included in commercial distribution bundles. Modeled after Google's BigTable, HBase is capable of hosting extremely large tables (billions of rows, millions of columns) atop clusters of commodity hardware and provides real-time, programmatic and query access to HDFS.

Unlike *columnar* relational databases, which store data in columns, HBase is a *column-oriented,* NoSQ, database that uses column families to group similar or frequently accessed data together. Built on top of HDFS, HBase enables low-latency queries and updates for large tables, so that single rows can be accessed quickly from a billion-row table. HBase achieves this by storing data in indexed StoreFiles on HDFS. Additional benefits include a flexible data model, fast table scans, scale in terms of writes, and the ability to handle sparse data sets that are common to many big data scenarios.