Summary & Highlights

In this module, you learned that:

* Hadoop is an open-source framework for Big Data that faced challenges when encountering dependencies and low-level latency.
* MapReduce, a parallel computing framework used in parallel computing​, is flexible for all data types, addresses parallel processing needs for multiple industries and contains two major tasks, “map” and “reduce​.”
* The four main stages of the Hadoop Ecosystem are Ingest, Store, Process and Analyze, and Access​.
* Key HDFS benefits include its cost efficiency, scalability, data storage expansion and data replication capabilities. Rack awareness helps reduce the network traffic and improve cluster performance. HDFS enables “write once, read many”operations.
* Suited for static data analysis and built to handle petabytes of data, Hive is a data warehouse software for reading, writing, and managing datasets. Hive is based on the “write once, read many” methodology, doesn’t enforce the schema to verify loading data and has built-in partitioning support.
* Linearly scalable and highly efficient, HBase is a column-oriented non-relational database management system that runs on HDFS and provides an easy-to-use Java API for client access. HBase architecture consists of HMaster, Region servers, Region, Zookeeper and HDFS. A key difference between HDFS and HBase is that HBase allows dynamic changes compared to the rigid architecture of HDFS.