HADOOP ANALYTICAL TOOLS

Top 15 Hadoop Analytical Tools

1. Apache Spark

- Apache Spark enables batch, real-time, and advanced analytics over the Hadoop platform. Spark provides in-memory data processing for the developers and the data scientists
- It has become the default execution engine for workloads such as batch processing, interactive queries, and streaming, etc.
- Companies, including Netflix, Yahoo, eBay, and many more, have deployed Spark at a massive scale.

2. MapReduce

- MapReduce is the main component of Hadoop.
- It is a software framework for writing applications that process large datasets in parallel across hundreds or thousands of nodes on the Hadoop cluster.
- Hadoop divides the client's MapReduce job into several independent tasks that run in parallel to give throughput.
- The MapReduce job is divided into map task and reduce task.
- Programmers generally write the entire business logic in the map task and specify light-weight processing like aggregation or summation on the reduce task.
- The MapReduce framework works in two phases- Map phase and the Reduce phase.
- The input to both the phases is the key-value pair.

3. Apache Impala

- Apache Impala is an open-source tool that overcomes the slowness of Apache Hive. It is a native analytic database for Apache Hadoop.

4. Apache Hive

- Apache Hive is a java-based data warehousing tool designed by Facebook for analyzing and processing large data.
- Hive uses HQL (Hive Query Language) like SQL that is transformed into MapReduce jobs for processing huge amounts of data.
- It provides support for developers and analytics to query and analyze big data with SQL like queries (HQL) without writing the complex MapReduce jobs.

5. Apache Mahout

- Apache Mahout is an open-source framework that normally runs coupled with the Hadoop infrastructure at its background to manage large volumes of data.
- We can use Apache Mahout for implementing scalable machine learning algorithms on the top of Hadoop using the MapReduce paradigm.
- It is a library of the scalable machine learning algorithm.
- Previously, it uses the Apache Hadoop platform, but now it focuses more on Apache Spark.
- Apache Mahout is not restricted to the Hadoop based implementation; it can run algorithms in the standalone mode as well.
- Apache Mahout implements popular machine learning algorithms such as Classification, Clustering, Recommendation, Collaborative filtering, etc.

6. Pig

- Pig is an alternative approach to make MapReduce job easier.
- Yahoo developed Pig to provide ease in writing the MapReduce.
- Pig enables developers to use Pig Latin, which is a scripting language designed for pig framework that runs on Pig runtime.
- Pig Latin is SQL like commands that are converted to MapReduce program in the background by the compiler.
- It translates the Pig Latin into MapReduce program for performing large scale data processing in YARN.

7. HBase

- HBase is an open-source distributed NoSQL database that stores sparse data in tables consisting of billions of rows and columns.
- It is written in Java and modeled after Google's big table.
- HBase provides support for all kinds of data and built on top of Hadoop.
- HBase is used when we need to search or retrieve a small amount of data from large data sets.

8. Apache Storm

- A storm is an open source distributed real-time computational framework written in Clojure and Java.
- With Apache Storm, we can reliably process unbounded streams of data (ever-growing data that has a beginning but no defined end). Apache Storm is simple and can be used with any programming language.
- Apache Storm is used for real-time analytics, continuous computation, online machine learning, ETL, and more.
- Among many, Yahoo, Alibaba, Groupon, Twitter, Spotify uses Apache Storm.

9. Tableau

- Tableau is a powerful data visualization and software solution tool in the Business Intelligence and analytics industry.
- It is the best tool for transforming the raw data into an easily understandable format with zero technical skill and coding knowledge.
- Tableau allows users to work on the live datasets and to spend more time on data analysis and offers real-time analysis.
- Tableau turns the raw data into valuable insights and enhances the decision-making process.
- It offers a rapid data analysis process, which results in visualizations that are in the form of interactive dashboards and worksheets.
- It works in synchronization with the other Big Data tools.

10. R

- R is an open-source programming language written in C and Fortran.
- It facilitates Statistical computing and graphical libraries.
- We can use R for performing statistical analysis, data analysis, and machine learning.
- It is platform-independent and can be used across multiple operating systems.

11. Talend

- Talend simplifies ETL and ELT for Big Data.
- It accomplishes the speed and scale of Spark.
- It handles data from multiple sources.

12. Lumify

- Lumify is open-source, big data fusion, analysis, and visualization platform that supports the development of actionable intelligence.
- With Lumify, users can discover complex connections and explore relationships in their data through a suite of analytic options, including full-text faceted search, 2D and 3D graph visualizations, interactive geospatial views, dynamic histograms, and collaborative workspaces shared in real-time.

13. KNIME

- KNIME offers simple ETL operations.
- It is an open-source, scalable data-analytics platform for analyzing big data, data mining, enterprise reporting, text mining, research, and business intelligence.

14. Apache Drill

- It is a low latency distributed query engine inspired by Google Dremel.
- Apache Drill allows users to explore, visualize, and query large datasets using MapReduce or ETL without having to fix to a schema.
- It is designed to scale to thousands of nodes and query petabytes of data.

15. Pentaho

- Pentaho offers real-time data processing tools for boosting digital insights.

- It is data integration, orchestration, and a business analytics platform that provides support ranging from big data aggregation, preparation, integration, analysis, prediction, to interactive visualization.