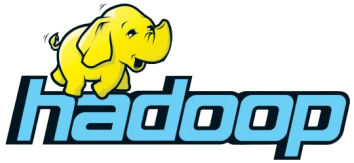
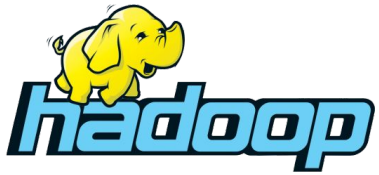


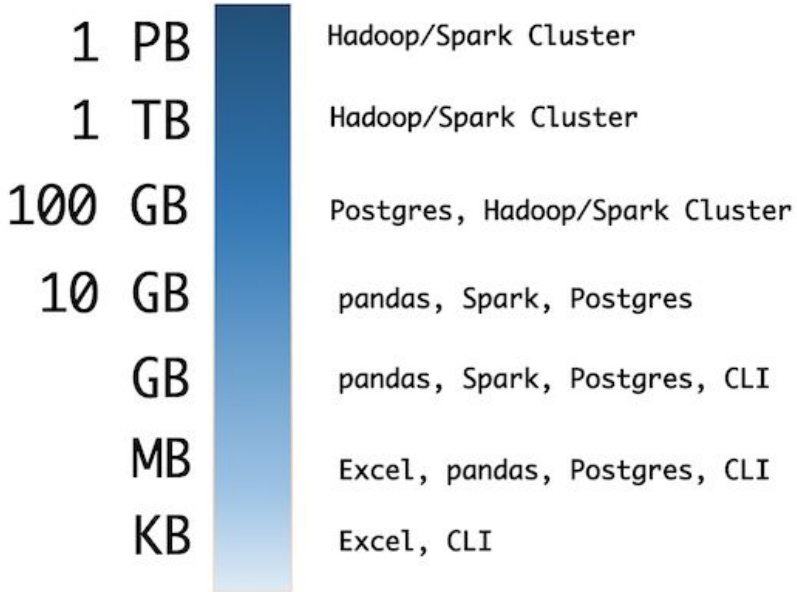
Introduction to Hadoop and Hive

March 18th, 2024

Shashank

RESEARCH, INNOVATION & IMPACT
Data Science Institute





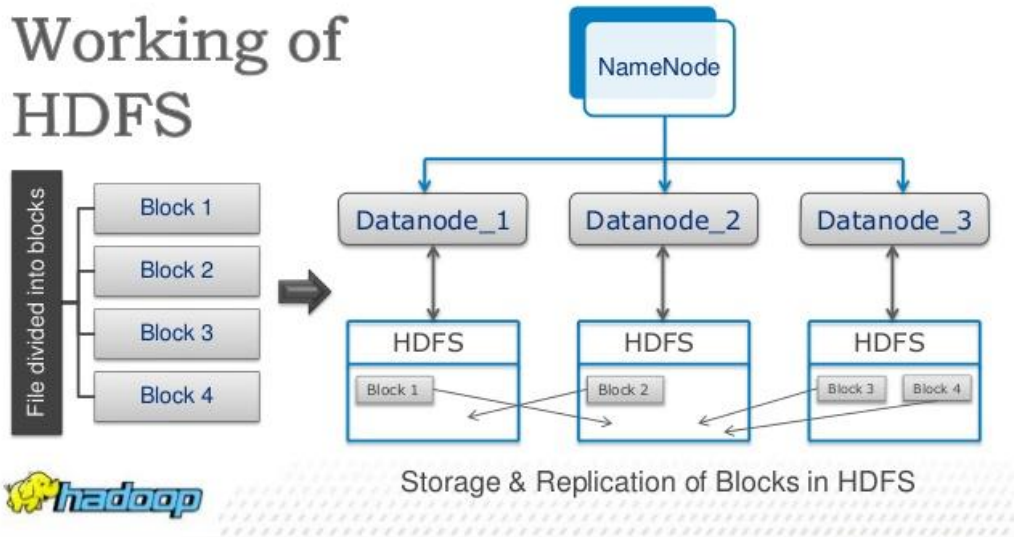
- Hadoop is an open-source framework designed for distributed storage and processing of big data sets.
- The core of Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce.
- Hadoop splits files into large blocks and distributes them across nodes in a cluster, processing data in parallel.

Core Components of Hadoop

Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data. It's designed to run on commodity hardware and has high fault tolerance.

MapReduce: A programming model for large scale data processing. It splits the data into smaller blocks and processes them in parallel to speed up the processing.

YARN (Yet Another Resource Negotiator): Manages and allocates system resources, allowing multiple data processing engines to handle data stored in HDFS.



Introduction to Hive

- Hive is a data warehousing tool built on top of Hadoop, designed to make querying and analyzing large datasets easier.
- Hive provides a SQL-like interface (HiveQL) to query data stored in various databases and file systems that integrate with Hadoop
- This allows users familiar with SQL to easily run queries on big data.

Key Features and Use Cases of Hive

- Offers a mechanism for structuring data and querying it with HiveQL, a SQL-like language.
- Allows extension of its capabilities through custom mappers and reducers.
- Simplifies querying and analysis for SQL users, eliminating the need to learn Java or MapReduce.

- Ideal for data warehousing applications where large datasets are analyzed for business intelligence.
- Facilitates log processing and analysis.
- Enhances machine learning and predictive modeling, especially when used with other Hadoop ecosystem tools like Spark.