Key concepts of Hadoop

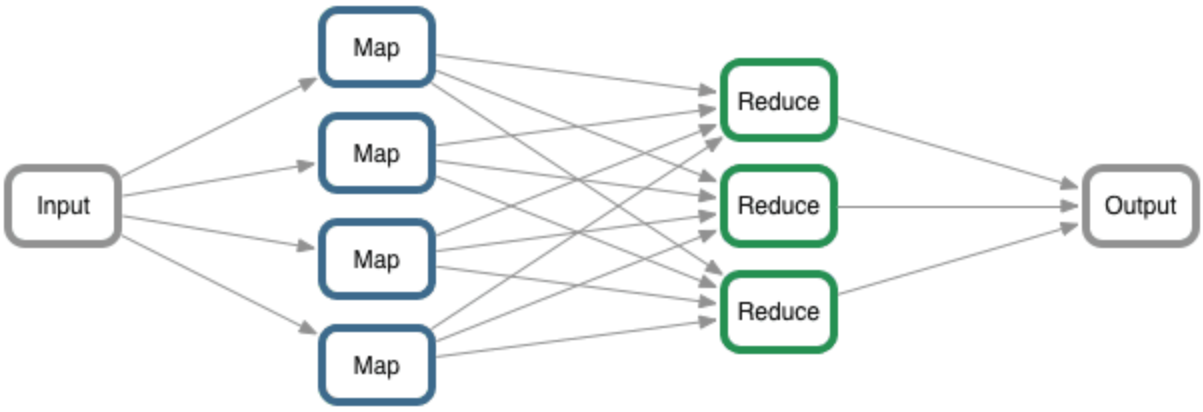
This section will define and discuss several terminologies used in Hadoop and its ecosystem. To facilitate comprehension, we will divide Hadoop into two major categories: the base module and the additional software packages and tools that can be installed separately or on top of Hadoop. Hadoop is the collective term for all of these elements.
To begin, let us examine the terms that comprise the base module.

1. Apache Hadoop

Apache Hadoop is an open-source platform for clustered data processing. It is based on the straightforward MapReduce programming style and enables dependable, scalable, and distributed computing. In this paradigm, both storage and computation are distributed.

2. MapReduce

MapReduce is a programming methodology for simultaneous data processing in a distributed setting. The MapReduce paradigm consists of two primary components. The first is the Map() method, which is responsible for filtering and sorting. The other is the Reduce() section, which is used to summarise the output from the Map section.



3. Hadoop Common

Apache Common includes utilities that are used to assist various Hadoop modules. It is essentially a collection of commonly used tools and utilities. Hadoop is mostly utilised by developers to aid in the building of applications.

4. Hadoop Distributed File System (HDFS)

HDFS is a distributed file system that runs on commodity hardware. It is extremely scalable and has a high throughput. In a clustered environment, data blocks are replicated and stored in a distributed fashion.

5. Yet Another Resource Negotiator (YARN)

YARN is in Hadoop 2 resource management. YARN's role in a clustered environment is to manage and schedule computing resources.

6. HBase

HBase is a free, scalable, distributed, and non-relational database management system. It is a Java application that is based on Google's Big Table. HDFS is the underlying storage file system.

7. Hive

Hive is a data warehouse management solution that enables the reading, writing, and management of massive amounts of data stored in a distributed storage system. For querying the dataset, it provides a SQL-like query language called HiveQL (HQL). Hive supports storage in HDFS as well as other suitable file systems such as Amazon S3.

8. Apache Pig

Apache Pig is a high-level framework for the study of huge data sets. Pig Latin is the language used to write Pig scripts. It abstracts away the underlying MapReduce routines, allowing developers to operate with the MapReduce concept without creating any code.

9. Apache Spark

Spark is a framework for cluster computing and a general-purpose compute engine for Hadoop data (large scale data-set). In memory, it is over 100 times faster than MapReduce. It is nearly ten times faster on disc. Spark may execute in a variety of environments/modes, including standalone mode, on Hadoop, and on Amazon EC2. It supports data access via HDFS, HBase, Hive, or any other Hadoop data source.

10. Sqoop

Sqoop is a command-line utility for data migration between relational databases and Hadoop databases. It is mostly used for data migration between relational and non-relational databases. Sqoop is derived from the start and final parts of two other terms: sql + hadoop.

11. Oozie

OOzie is a workflow engine for Hadoop. It manages Hadoop operations by scheduling workflows.

12. ZooKeeper

ZooKeeper is an open source platform for Hadoop distributed applications that provides a high-performance coordination service. It is a centralised service for configuration management, naming registry management, distributed synchronisation management, and group services.

13. Apache Flume

Apache Flume is a distributed data collecting, aggregation, and movement service. It is extremely efficient when dealing with big amounts of log and event data.

14. Hue

Hue is just a web interface for Hadoop data analysis. It is an open source project that facilitates the use of Hadoop and its ecosystem. Its primary objective is to improve the user experience. It has drag-and-drop functionality and editors for Spark, Hive, and HBase.

15. Mahout

Mahout is open source software that enables the rapid development of scalable machine learning and data mining applications.

16. Ambari

Ambari is a web-based monitoring and management solution for Hadoop clusters. It supports ecosystem services and tools including HDFS, MapReduce, HBase, ZooKeeper, Pig, and Sqoop, among others. Provisioning, maintaining, and monitoring Hadoop clusters are three of its primary functions.

Conclusion:

As the Hadoop ecosystem evolves, new applications, services, and solutions emerge. As a result, the world of big data will adopt new phrases and lingo. We must keep a tight eye on individuals in the know and comprehend them in real time.

We have attempted to highlight the most critical key phrases in the Hadoop ecosystem in this post. Additionally, we discussed a little bit about the ecosystem and why it is necessary for us to understand the terminologies. Hadoop is gaining traction as a mainstream technology, and as a result, more people are becoming involved.

Thus, the time has come to familiarise yourself with some of the fundamental concepts and words used in the Hadoop universe. There will be a plethora of new notions and phrases available in the future, and we must adapt properly.