What is Hadoop?

Apache Hadoop is a 100 percent open source framework that pioneered a new way for the distributed processing of large, enterprise data sets. Instead of relying on expensive, and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data. With Hadoop, no data is too big data.

Hadoop Architecture

A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode and DataNode. Though it is possible to have data-only worker nodes and compute-only worker nodes, a slave or worker node acts as both a DataNode and TaskTracker. In a larger cluster, the Hadoop Distributed File System (HDFS) is managed through a dedicated NameNode server to host the file system index, and a secondary NameNode that can generate snapshots of the NameNode's memory structures, thus preventing file-system corruption and reducing loss of data.



**Apache Hadoop Architecture**

**The Apache Hadoop framework comprises:**

**• Hadoop Common –** Contains libraries and utilities needed by other Hadoop modules

**• Hadoop Distributed File System (HDFS) –** A distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster

**• Hadoop YARN –** A resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications

**• Hadoop MapReduce–**A programming model for large-scale data processing

# Why Big Data Hadoop

In a fast-paced and hyper-connected world where more and more data is being created, Hadoop’s breakthrough advantages mean that businesses and organizations can now find value in data that was considered useless.

Organizations are realizing that categorizing and analyzing Big Data can help make major business predictions. Hadoop allows enterprises to store as much data, in whatever form, simply by adding more servers to a Hadoop cluster. Each new server adds more storage and processing power to the cluster. This makes data storage with Hadoop less expensive than earlier data storage methods.

# Hadoop and Big Data

With 90 percent of data being unstructured and growing rapidly, Hadoop is required to put the right Big Data workloads in the right systems and optimize data management structure in an organization. The cost-effectiveness, scalability and systematic architecture of Hadoop make it more necessary for organizations to process and manage Big Data.