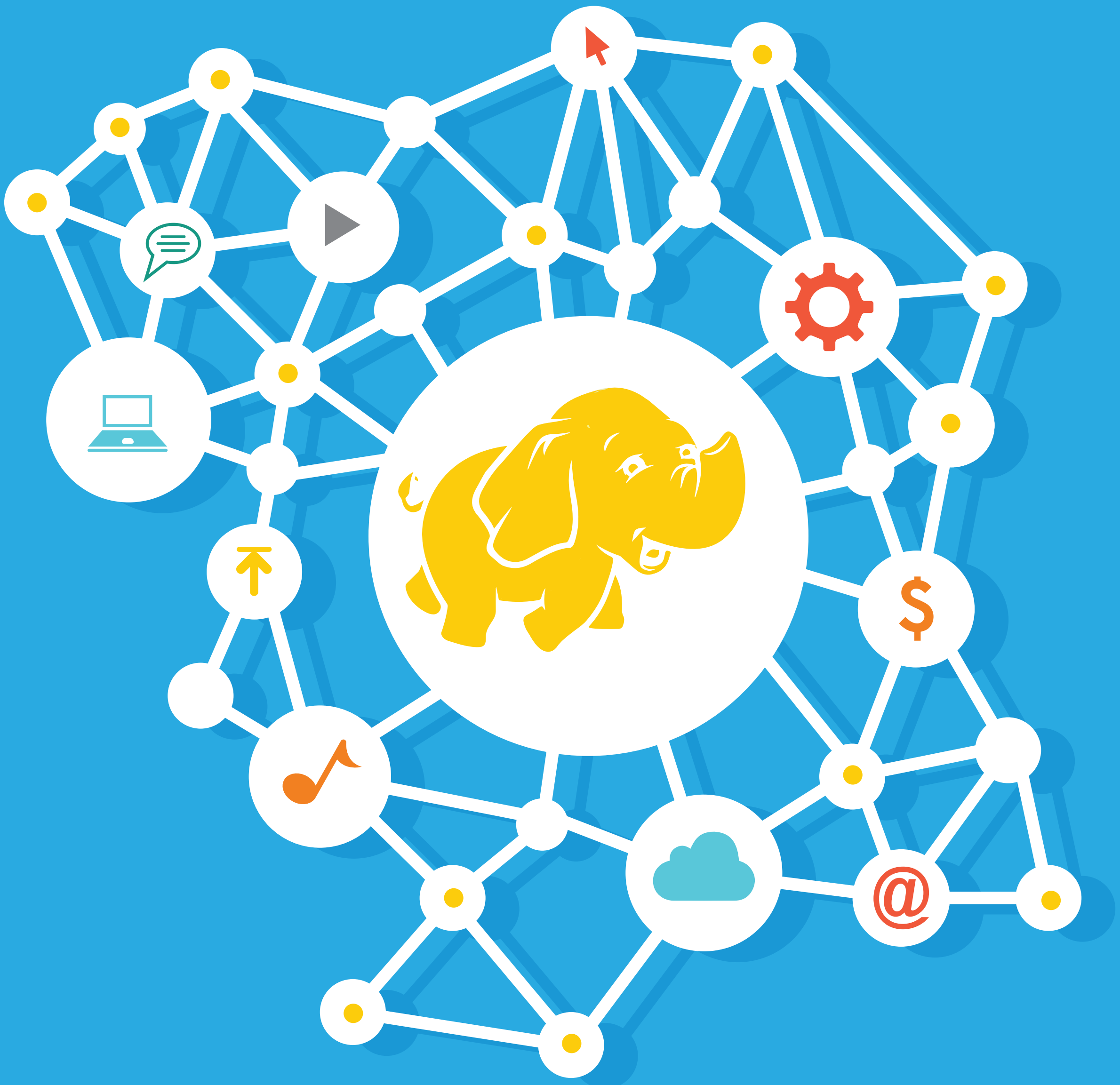
simplilearn

8 ESSENTIAL CONCEPTS OF BIG DATA AND HADOOP

A QUICK REFERENCE GUIDE



What is Big Data and Hadoop?

Big Data refers to large sets of data that cannot be analyzed with traditional tools. It stands for data related to large-scale processing architectures.

Hadoop is the software framework that is developed by Apache to support distributed processing of data. Initially, Java™ language was used to develop Hadoop, but today many other languages are used for scripting Hadoop. Hadoop is used as the core platform to structure Big Data and helps in performing data analytics.

Table of Contents

Chapter: 1	Important Definitions
Chapter: 2	MapReduce
Chapter: 3	HDFS
Chapter: 4	Pig vs. SQL
Chapter: 5	HBase componenets
Chapter: 6	Cloudera
Chapter: 7	Zookeeper and Sqoop
Chapter: 8	Hadoop Ecosystem

Chapter: 1 — Important Definitions

TERM	DEFINITION
Big data	Big Data refers to the data sets whose size makes it difficult for commonly used data capturing software tools to interpret, manage, and process them within a reasonable time frame.
Hadoop	Hadoop is an open-source framework built on the Java environment. It assists in the processing of large data sets in a distributed computing environment.
VMware Player	VMware Player is a free software package offered by VMware, Inc., which is used to create and manage virtual machines.
Hadoop Architecture	Hadoop is a master and slave architecture that includes the NameNode as the master and the DataNode as the slave.
HDFS	The Hadoop Distributed File System (HDFS) is a distributed file system that shares some of the features of other distributed file systems. It is used for storing and retrieving unstructured data.
MapReduce	The MapReduce is a core component of Hadoop, and is responsible for processing jobs in distributed mode.
Apache Hadoop	One of the primary technologies, which rules the field of Big Data technology, is Apache Hadoop™.
Ubuntu Server	Ubuntu is a leading open-source platform for scale out. Ubuntu helps in utilizing the infrastructure at its optimum level irrespective of whether users want to deploy a cloud, a web farm, or a Hadoop cluster.
Pig	The Apache Pig is a platform which helps to analyze large datasets that includes high-level language required to express data analysis programs. Pig is one of the components of the Hadoop eco-system.
Hive	Hive is an open-source data warehousing system used to analyze a large amount of dataset that is stored in Hadoop files. It has three key functions like summarization of data, query, and analysis.
SQL	It is a query language used to interact with SQL databases.
Metastore	It is the component that stores the system catalog and metadata about tables, columns, partitions, etc. It is stored in a traditional RDBMS format.
Driver	Driver is the component that Manages the lifecycle of a HiveQL statement.
Query compiler	A query compiler is one of the driver components. It is responsible for compiling the Hive script for errors.

Query optimizer	A query optimizer optimizes Hive scripts for faster execution of the same. It consists of a chain of transformations.
Execution engine:	The role of the execution engine is to execute the tasks produced by the compiler in proper dependency order.
Hive server	The Hive Server is the main component which is responsible for providing an interface to the user. It also maintains connectivity in modules.
Client components	The developer uses client components to perform development in Hive. The client components include Command Line Interface (CLI), web UI, and the JDBC/ODBC driver.
Apache HBase	It is a distributed, column-oriented database built on top of HDFS (Hadoop Distributed Filesystem). HBase can scale horizontally to thousands of commodity servers and petabytes by indexing the storage.
ZooKeeper	It is used for performing region assignment. ZooKeeper is a centralized management service for maintaining and configuring information, naming, providing distributed synchronization, and group services.
Cloudera	It is a commercial tool for deploying Hadoop in an enterprise setup.
Sqoop	It is a tool that extracts data derived from non-Hadoop sources and formats them such that the data can be used by Hadoop later.

Chapter: 2 — **MapReduce**

The MapReduce component of Hadoop is responsible for processing jobs in distributed mode. The features of MapReduce are as follows:

Distributed data processing

The first feature of MapReduce component is that it performs distributed data processing using the MapReduce programming paradigm.

User-defined map phase

The second feature of MapReduce is that you can possess a user-defined map phase, which is a parallel, share-nothing processing of input.

Aggregation of output

The third feature of MapReduce is aggregation of the output of the map phase, which is a user-defined reduce phase after a map process.

Chapter: 3 — **HDFS**

HDFS is used for storing and retrieving unstructured data. The features of Hadoop HDFS are as follows:

Provides access to data blocks

HDFS provides a high-throughput access to data blocks. When an unstructured data is uploaded on HDFS, it is converted into data blocks of fixed size. The data is chunked into blocks so that it is compatible with commodity hardware's storage.

Helps to manage file system

HDFS provides a limited interface for managing the file system to allow it to scale. This feature ensures that you can perform a scale up or scale down of resources in the Hadoop cluster.

Creates multiple replicas of data blocks

The third feature of MapReduce is aggregation of the output of the map phase, which is a user-defined reduce phase after a map process.

Chapter: 4 — **Pig vs. SQL**

The table below includes the differences between Pig and SQL:

Difference	Pig	SQL
Definition	HDFS provides a limited interface for managing the file system to allow it to scale. This feature ensures that you can perform a scale up or scale down of resources in the Hadoop cluster.	It is a query language used to interact with SQL databases.

Example	customer = **LOAD** '/data/customer.dat' **AS** (c_id,name,city);
	sales = **LOAD** '/data/sales.dat' **AS** (s_id,c_id,-date,amount);
	salesBLR = **FILTER** customer **BY** city == 'Bangalore';
	joined= **JOIN** customer **BY** c_id, salesBLR BY c_id;
	grouped = **GROUP** joined BY c_id;
	summed= **FOREACH** grouped **GENERATE GROUP**, **SUM**(joined.salesBLR::amount);
	spenders= **FILTER** summed **BY** $1 > 100000;
	sorted = **ORDER** spenders **BY** $1 **DESC**;
	DUMP sorted;
		SELECT c_id , **SUM**(amount) AS CTotal
		FROM customers c
		JOIN sales s ON c.c_id = s.c_id
		WHERE c.city = 'Bangalore'
		GROUP BY c_id
		HAVING SUM(amount) > 100000
		ORDER BY CTotal **DESC**

Chapter: 5 — **HBase components**

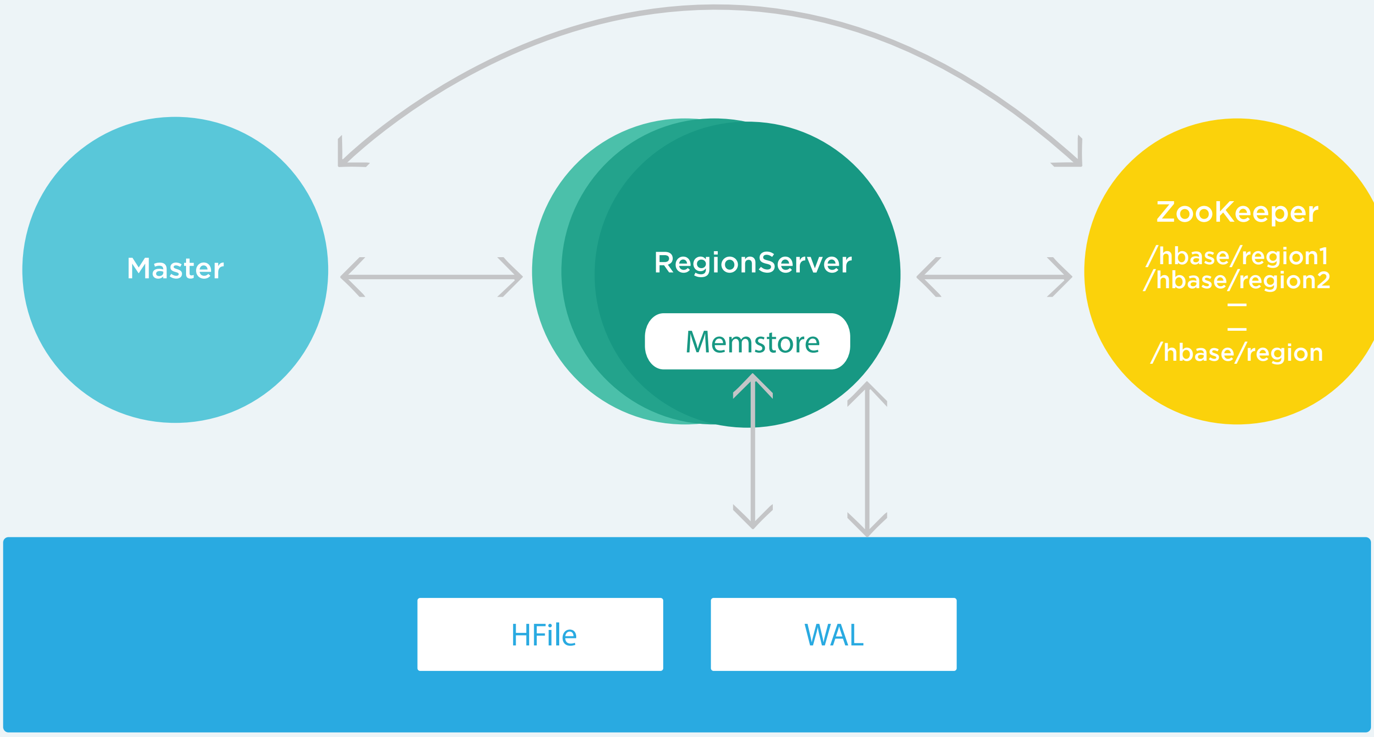
Introduction

Apache HBase is a distributed, column-oriented database built on top of HDFS (Hadoop Distributed Filesystem). HBase can scale horizontally to thousands of commodity servers and petabytes by indexing the storage. HBase supports random real-time CRUD operations. HBase also has linear and modular scalability. It supports an easy-to-use Java API for programmatic access.

HBase is integrated with the MapReduce framework in Hadoop. It is an open-source framework that has been modeled after Google's BigTable. Further, HBase is a type of NoSQL.

HBase Components

The components of HBase are HBase Master and Multiple RegionServers



An explanation of the components of HBase is given below:

HBase Master

It is responsible for managing the schema that is stored in Hadoop Distributed File System (HDFS).

Multiple RegionServers

They act like availability servers that help in maintaining a part of the complete data, which is stored in HDFS according to the requirement of the user. They do this using the HFile and WAL (Write Ahead Log) service. The RegionServers always stay in sync with the HBase Master. The responsibility of ZooKeeper is to ensure that the RegionServers are in sync with the HBase Master.

Chapter: 6 — **Cloudera**

Cloudera is a commercial tool for deploying Hadoop in an enterprise setup. The salient features of Cloudera are as follows:

It uses 100% open-source distribution of Apache Hadoop and related projects like Apache Pig, Apache Hive, Apache HBase, Apache Sqoop, etc.

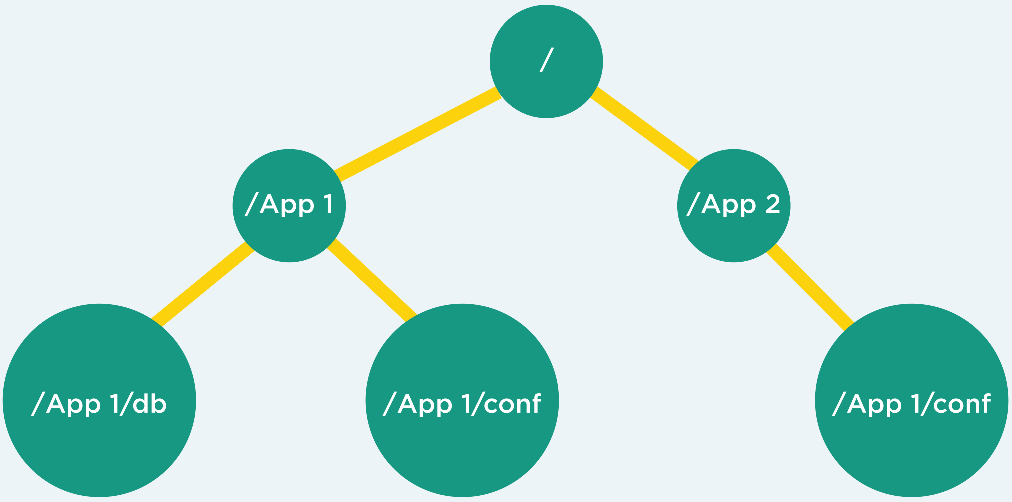
It has its own user-friendly Cloudera Manager for system management, Cloudera Navigator for data management, dedicated technical support, etc.

Chapter: 7 — **Zookeeper and Sqoop**

ZooKeeper is an open-source and high performance co-ordination service for distributed applications. It offers services such as Naming, Locks and synchronization, Configuration management, and Group services.

ZooKeeper Data Model

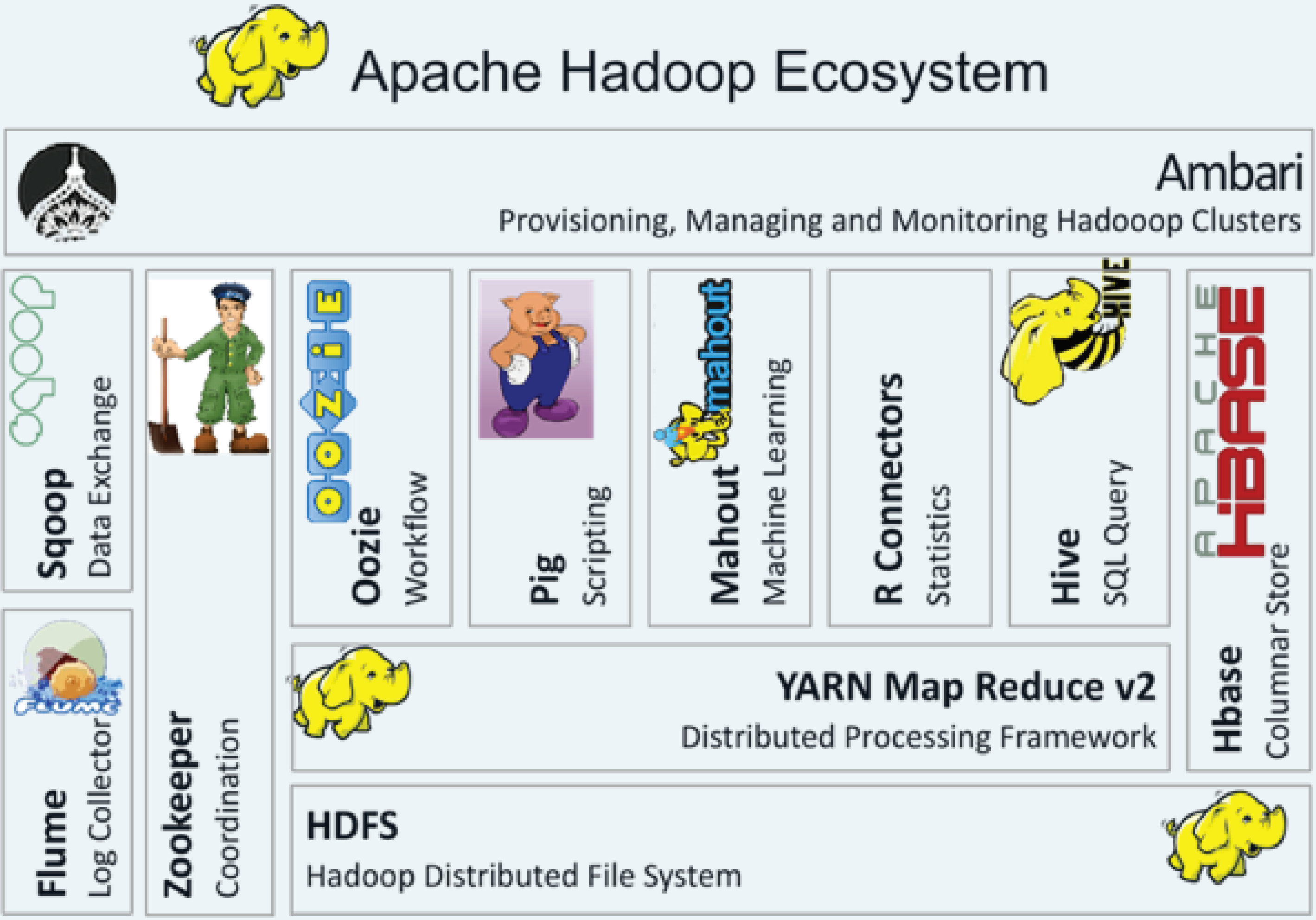
ZooKeeper has a hierarchical namespace. Each node in the namespace is called Znode. The example given here shows the tree diagram used to represent the namespace. The tree follows a top-down approach where '/' is the root and App1 and App2 resides in the root. The path to access db is /App1/db. This path is called the hierarchical path.



Sqoop is an Apache Hadoop ecosystem project whose responsibility is to import or export operations across relational databases like MySQL, MSSQL, Oracle, etc. to HDFS. Following are the reasons for using Sqoop:

- SQL servers are deployed worldwide and are the primary ways to accept the data from a user. Nightly processing is done on SQL server for years.
- It is essential to have a mechanism to move the data from traditional SQL DB to Hadoop HDFS.
- Transferring the data using some automated script is inefficient and time-consuming.
- Traditional DB has reporting, data visualization, and other applications built in enterprises but to handle large data, we need an ecosystem.
- The need to bring the processed data from Hadoop HDFS to the applications like database engine or web services is satisfied by Sqoop.

Chapter: 8 — Hadoop Ecosystem



The image given here depicts the various Hadoop ecosystem components. The base of all the components is Hadoop Distributed File System (HDFS). Above this component is YARN MapReduce v2. This framework component is used for the distributed processing in a Hadoop cluster.

The next component is Flume. Flume is used for collecting logs across a cluster. Sqoop is used for data exchange between a relational database and Hadoop HDFS.

The ZooKeeper component is used for coordinating the nodes in a cluster. The next ecosystem component is Oozie. This component is used for creating, executing, and modifying the workflow of a MapReduce job. The Pig component is used for performing scripting for MapReduce applications.

The next component is Mahout. This component is used for machine learning based on machine inputs. R Connectors are used for generating statistics of the nodes in a cluster. Hive is used for interacting with Hadoop using SQL like query. The next component is HBase. This component is used for slicing of large data.

The last component is Ambari. This component is used for provisioning, managing, and monitoring Hadoop clusters.