Hbase

Overview

HBase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable.

One can store the data in HDFS either directly or through HBase. Data consumer reads/accesses the data in HDFS randomly using HBase. HBase sits on top of the Hadoop File System and provides read and write access.

HBase is a column-oriented database and the tables in it are sorted by row. The table schema defines only column families, which are the key value pairs. A table have multiple column families and each column family can have any number of columns. Subsequent column values are stored contiguously on the disk.

Architecture

MasterServer

The master server:

- Assigns regions to the region servers and takes the help of Apache ZooKeeper for this task.
- Handles load balancing of the regions across region servers. It unloads the busy servers and shifts the regions to less occupied servers.
- Maintains the state of the cluster by negotiating the load balancing.

Is responsible for schema changes and other metadata operations such as creation of tables and column families.

Regions

Regions are nothing but tables that are split up and spread across the region servers.

Zookeeper

- Zookeeper is an open-source project that provides services like maintaining configuration information, naming, providing distributed synchronization, etc.
- Zookeeper has ephemeral nodes representing different region servers. Master servers use these nodes to discover available servers.
- In addition to availability, the nodes are also used to track server failures or network partitions.
- Clients communicate with region servers via zookeeper.
- In pseudo and standalone modes, HBase itself will take care of zookeeper

Installation

Before installing Hadoop into Linux environment, we need to set up Linux using ssh (Secure Shell). Java is the main prerequisite for Hadoop and Hbase. After installing java, you have to install Hadoop.

Install HBase in any of the three modes: Standalone mode, Pseudo Distributed mode, and Fully Distributed mode.

Shell

HBase contains a shell using which you can communicate with HBase. HBase uses the Hadoop File System to store its data. It will have a master server and region servers. The data storage will be in the form of regions (tables). These regions will be split up and stored in region servers.

General Commands:

- status - Provides the status of HBase, for example, the number of servers.
- version - Provides the version of HBase being used.
- table_help - Provides help for table-reference commands.
- whoami - Provides information about the user.

Data Definition Language:

- create - Creates a table.
- list - Lists all the tables in HBase.

- disable - Disables a table.
- is_disabled - Verifies whether a table is disabled.
- enable - Enables a table.
- is_enabled - Verifies whether a table is enabled.
- describe - Provides the description of a table.
- alter - Alters a table.
- exists - Verifies whether a table exists.
- drop - Drops a table from HBase.
- drop_all - Drops the tables matching the ‘regex’ given in the command.
- Java Admin API - Java provides an Admin API to achieve DDL functionalities through programming.

Data Manipulation Language:

- put - Puts a cell value at a specified column in a specified row in a particular table.
- get - Fetches the contents of row or a cell.
- delete - Deletes a cell value in a table.
- deleteall - Deletes all the cells in a given row.
- scan - Scans and returns the table data.
- count - Counts and returns the number of rows in a table.
- truncate - Disables, drops, and recreates a specified table.
- Java client API - Java provides a client API to achieve DML functionalities, CRUD (Create Retrieve Update Delete) operations and more through programming.
