**Data Analytics Laboratory**

**Task 10**

**No-SQL database –Apache HBase**

## Aim

To set HBase shell environment and create tables, insert rows, display its contents.

## Introduction to HBase

* HBase is a column-oriented non-relational database management system that runs on top of Hadoop Distributed File System (HDFS).
* HBase provides a fault-tolerant way of storing sparse data sets, which are common in many big data use cases. It is well suited for real-time data processing or random read/write access to large volumes of data.
* HBase does not support a structured query language like SQL. Each table must have an element defined as a primary key, and all access attempts to HBase tables must use this primary key.
* Supports primitive data types including numeric, binary data and strings; and a number of complex types including arrays, maps, enumerations and records.

|  |  |
| --- | --- |
| HBase | RDBMS |
| Column oriented | Row-oriented (mostly) |
| Flexible schema, columns can be added on the fly | Fixed schema |
| Designed to store Denormalized data | Designed to store Normalized data |
| Good with sparse tables | Not optimized for sparse tables |
| Joins using MapReduce which is not optimized | Optimized for joins |
| Tight integration with MapReduce | No integration with MapReduce |
| Horizontal scalability – just add hardware | Hard to shard and scale |
| Good for semi-structured data as well as structured data | Good for structured data |

## Prerequisites

**1. What is Hbase?**

Hbase is a column-oriented database management system which runs on top of HDFS (Hadoop Distribute File System). Hbase is not a relational data store, and it does not support structured query language like SQL.

**2. What is column families?**

Column families comprise the basic unit of physical storage in Hbase to which features like compressions are applied.

**3. Explain what is the row key?**

Row key is defined by the application. As the combined key is pre-fixed by the rowkey, it enables the application to define the desired sort order. It also allows logical grouping of cells and make sure that all cells with the same rowkey are co-located on the same server.

**4. What is the difference between getting and Scan?**

Get will return only single row from Hbase table based on row key . Scan command returns set of rows depending upon given search condition.

**5. What happens when deleting a row?**

At the time of deletion, data is not physically deleted from the file system instead make tombstone marker. Physical deletion happens during compaction.

## In-Lab Tasks

To start the HBase environment type hbase shell in the terminal

$hbase shell

hbase(main):001:0>

Use the create command to create a new table. You must specify the table name and the ColumnFamily name.

hbase(main):001:0> create 'student', 'name', 'subject'

0 row(s) in 0.4170 seconds

=> Hbase::Table - student

Use the list command to confirm your table exists

hbase(main):002:0> list 'student'

TABLE

student

1 row(s) in 0.0180 seconds

=> ["test"]

To put data into your table, use the put command.

We are storing data for student 1

hbase(main):003:0> put 'student', '001', 'name:firstname', 'Arunkumar'

0 row(s) in 0.0850 seconds

hbase(main):004:0> put 'student', '001', 'name:lastname', 'Gopu'

0 row(s) in 0.0110 seconds

hbase(main):005:0> put 'student', '001', 'name:fullname', 'Arunkumar Gopu'

0 row(s) in 0.0100 seconds

hbase(main):003:0> put 'student', '001', 'subject:s1', 'Data Analytics'

0 row(s) in 0.0650 seconds

hbase(main):003:0> put 'student', '001', 'subject:s2', 'Data mining'

0 row(s) in 0.0450 seconds

hbase(main):003:0> put 'student', '001', 'subject:s3', 'Operating system'

0 row(s) in 0.01050 seconds

Inserting data for second student

hbase(main):003:0> put 'student', '002', 'name:firstname', 'David Jackson'

0 row(s) in 0.0850 seconds

hbase(main):004:0> put 'student', '002', 'name:lastname', 'Samuel'

0 row(s) in 0.0110 seconds

hbase(main):005:0> put 'student', '002', 'name:fullname', 'David Jackson Samuel'

0 row(s) in 0.0100 seconds

hbase(main):003:0> put 'student', '002', 'subject:s1', 'Computer Networks'

0 row(s) in 0.0650 seconds

hbase(main):003:0> put 'student', '002', 'subject:s2', 'R programming'

0 row(s) in 0.0450 seconds

hbase(main):003:0> put 'student', '002', 'subject:s3', 'Data Analytics'

0 row(s) in 0.01050 seconds

Scan the table for all data at once. One of the ways to get data from HBase is to scan. Use the scan command to scan the table for data.

hbase(main):006:0> scan 'student'

ROW COLUMN+CELL

001 column=name:firstname, timestamp=1421762485768, value=Arunkumar

001 column=name:lastname, timestamp=1421762494563, value=Gopu

001 column=name:fullname, timestamp=1421762492345, value=Arunkumar Gopu

001 column=subject:s1, timestamp=1421762487389, value=Data Analytics

001 column=subject:s2, timestamp=1421762492834, value=Data mining

001 column=subject:s3, timestamp=1421762437483, value=Operating system

002 column=name:firstname, timestamp=1421762485768, value=David Jackson

002 column=name:lastname, timestamp=1421762494563, value=Samuel

002 column=name:fullname, timestamp=1421762492345, value=David Jackson Samuel

002 column=subject:s1, timestamp=1421762487389, value=Computer Networks

002 column=subject:s2, timestamp=1421762492834, value=R programming

002 column=subject:s3, timestamp=1421762437483, value=Data Analytics

3 row(s) in 0.0230 seconds

To get a single row of data at a time, use the get command.

hbase(main):007:0> get 'student', '001'

COLUMN CELL

name:firstname timestamp=1421762485768, value=Arunkumar

name:lastname timestamp=1421762482362, value=Gopu

name:fullname timestamp=1421762487287, value=Arunkumar Gopu

subject:s1 timestamp=1421762485345, value=Data Analytics

subject:s2 timestamp=1421762482367, value=Data mining

subject:s3 timestamp=1421762488736, value=Operating system

6 row(s) in 0.0350 seconds

If you want to delete a table or change its settings, as well as in some other situations, you need to disable the table first, using the disable command. You can re-enable it using the enable command.

hbase(main):008:0> disable 'student'

0 row(s) in 1.1820 seconds

hbase(main):009:0> enable 'student'

0 row(s) in 0.1770 seconds

To drop (delete) a table, use the drop command.

hbase(main):011:0> drop 'student'

0 row(s) in 0.1370 seconds

To exit the HBase Shell and disconnect from your cluster, use the quit command. HBase is still running in the background.

hbase(main):008:0>quit

$stop-hbase.sh

stopping hbase....................

$stopCDH.sh

**Results**

The program is implemented in python and the output is observed.

**Faculty Signature**