Introduction to HBase

Agenda

- What is Hbase.
- Hbase Vs Hadoop.
- Conceptual View
- Column Family
- Column family Vs Qualifier
- HBase Processes
- Hbase Master
- Installation modes
- Pros of Hbase
- Cons of HBase

What is HBase

Hbase is a subproject of Hadoop & developed by Apache.

It is NoSQL database which can handle millions of rows in a table.

Data is logically organized into tables, rows and columns

HBase table contains data in key : value format.

HBase table contains all key in sorted manner.

Random read/write are very fast in Hbase.

HBase Vs Hadoop

HMaster can run on master machine of Hadoop or any other machine.

HBase uses Hadoop for storing its tables.

Datanodes of Hadoop works as HReigionServer of HBase.

Big tables are splitted into smaller parts and stored on HDFS(HReigionServer).

HBase can also run in standalone mode, in which it doesn't use Hadoop for storage.

Conceptual View

- A row has a sortable row key and an arbitrary number of columns.
- HBase stores the current timestamp while inserting the data.
- Hbase has concept of column family.
- <column family>:<qualifier> makes a column
- HBase can keep versioned data.
- You can configure how many version you want.

Column Family

Instead of columns, Hbase has column families.

Column families are part of table schema.

Combination of column family & qualifier makes a column.

Qualifiers are not part of table schema.

We can create as many qualifiers as required at runtime.

Different rows can have different no. Of columns.

Column family Vs Qualifier

Key	Value		
	cf1		cf2
	qf1	qf2	qf1
row1	value1	value2	value3
row2	value4		

```
ROW                 COLUMN+CELL
 row1               column=cf1:qf1, timestamp=1350804873477, value=value1
 row1               column=cf1:qf2, timestamp=1350804992484, value=value2
 row1               column=cf2:qf1, timestamp=1350805003788, value=value3
 row2               column=cf1:qf1, timestamp=1350812449169, value=value4
```

HBase Processes

- The HMaster (runs on master machine)
- The RegionServer (runs on slavemachine machine)
- The HBase client (interactive shell to run Hbase commands)

HMaster

- Master machine for HBase cluster.
- Does all the cordination activities.
- It keeos metadata about all HBase tables.
- ROOT region locates all the META regions.
- Assign user regions to the HRegionServers.
- Enable/Disable table and change table schema
- Monitor the health of each Server

Installation modes

Standalone

Does not use Hadoop for storage

Pseudo-distributed

Uses single node hadoop cluster for storing tables.

Fully distributed

Uses fully distributed Hadoop cluster for soring tables.

Pros of HBase

Distributed

Built on Hadoop HDFS

Fault tolerant, no data loss

Handles Big Data

High performance for write and read

Scalable (auto-sharding)

Cons of HBase

Does not support table Join.

Does not support Indexes.

Columns are not the part of table schema.

No security control.

Can lose it's data if Hadoop crashes.

...Thanks...