Introduction to HBase

For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com

Agenda

Disadvantage of SQL
What is NoSQL ?
Advantages of NoSQL
What is HBase
HBase Vs Hadoop.
Conceptual view
Column family
Column family Vs Qualifier
Installation modes
Pros of HBase
Cons of HBase

Disadvantage of SQL Database

Can’t process large amount of data.

Not easily scalable.

No partial failure.

No fault tolerant.

Need a well defined schema .

Not an open source.

What is NoSQL ?

Stands for Not only SQL (NoSQL).

NoSQL can handle structured/unstructured data.

NoSQL can use Hadoop for its storage.

NoSQL does not require a fixed table schema.

HBase is an example of NoSQL.

Advantages of NoSQL

NoSQL can handle any size of data.

NoSQL is schema less.

Partial failure of system.

NoSQL uses HDFS for storage.

Fault tolerant is taken care by Hadoop.

Easy to implement (open source).

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 can handle unstructured data.

HBase Vs Hadoop

HBase uses HDFS for storing its tables data.

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

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

HMaster is the master process of HBase which controls all HBase activities.

HRegionServer is the slave process of HBase.

Conceptual View

A row has a sortable row key and an arbitrary number of columns.

HBase can maintain versioned data.

You can configure how many versions you want.

HBase stores the current timestamp while inserting the data.

HBase has concept of column family.

<column family>:<qualifier> makes a column

Column Family

Instead of columns, HBase has column families.

Column families are part of table schema.

Qualifiers are not part of table schema.

We can create as many qualifiers as required at runtime.

Combination of **column family** & **qualifier** makes a column.

Different rows can have different no. of columns.

Column Family Vs Qualifier

Key	Value		
	official		personal
	department	designation	name
Emp001	IT	Hadoop Admin	Naveen
Emp002			Mahesh

```
hbase(main):017:0> scan 'employee'
ROW                     COLUMN+CELL
 Emp001                 column=official:department, timestamp=1366042568512, value=IT
 Emp001                 column=official:designation, timestamp=1366042654784, value=Hadoop Admin
 Emp001                 column=personal:name, timestamp=1366042592684, value=Naveen
 Emp002                 column=personal:name, timestamp=1366042679251, value=Mahesh
2 row(s) in 0.0370 seconds

hbase(main):018:0>
```

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 storing tables.

Pros of HBase

Distributed

Built on Hadoop HDFS

Handles Big Data

High performance for write and read

Scalable (auto-sharding)

Fault tolerant, no data loss

Cons of HBase

Does not support table join.

Does not support group by.

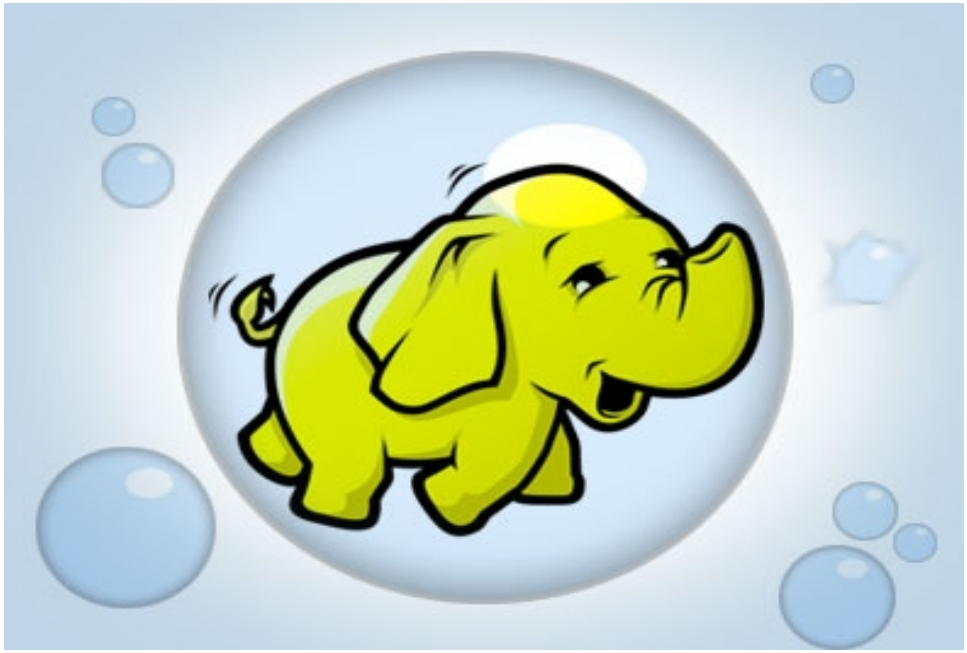
Does not support Indexes.

Columns are not the part of table schema.

Can lose it's data if Hadoop crashes.

Not suitable for transactions

...Thanks...



For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com