**Practical 8: Schema design using HBase**

**What is 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.

Unlike relational database systems, HBase does not support a structured query language like

SQL; in fact, HBase isn’t a relational data store at all. HBase applications are written in JavaTM

much like a typical Apache MapReduce application. HBase does support writing applications

in Apache Avro, REST and Thrift.

An HBase system is designed to scale linearly. It comprises a set of standard tables with

rows and columns, much like a traditional database. Each table must have an element

defined as a primary key, and all access attempts to HBase tables must use this primary key.

Avro, as a component, supports a rich set of primitive data types including: numeric, binary

data and strings; and a number of complex types including arrays, maps, enumerations and

records. A sort order can also be defined for the data.

HBase relies on ZooKeeper for high-performance coordination. ZooKeeper is built into

HBase, but if you’re running a production cluster, it’s suggested that you have a dedicated

ZooKeeper cluster that’s integrated with your HBase cluster.

HBase works well with Hive, a query engine for batch processing of big data, to enable fault

tolerant big data applications.

**An example of HBase**

An HBase column represents an attribute of an object; if the table is storing diagnostic logs

from servers in your environment, each row might be a log record, and a typical column

could be the timestamp of when the log record was written, or the server name where the

record originated.

HBase allows for many attributes to be grouped together into column families, such that the

elements of a column family are all stored together. This is different from a row-oriented

relational database, where all the columns of a given row are stored together. With HBase

you must predefine the table schema and specify the column families. However, new

columns can be added to families at any time, making the schema flexible and able to adapt

to changing application requirements.

Just as HDFS has a NameNode and slave nodes, and MapReduce has JobTracker and

TaskTracker slaves, HBase is built on similar concepts. In HBase a master node manages the

cluster and region servers store portions of the tables and perform the work on the data. In

the same way HDFS has some enterprise concerns due to the availability of the NameNode

HBase is also sensitive to the loss of its master node.

**HBase 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.

The master server manages these region servers and all these tasks take place on HDFS.

Given below are some of the commands supported by HBase Shell.

We can start the HBase interactive shell using “HBase shell” command as shown below.

1)Creating a Table using HBase Shell

We can create a table using the create command, here you must specify the table name and

the Column Family name. The syntax to create a table in HBase shell is shown below.

create ‘<table name>’,’<column family>’

**>create 'emp', 'office'**



Check the shell functioning before proceeding further. Use the list command for this

purpose. List is a command used to get the list of all the tables in HBase. It lists all the tables

in HBase.

**hbase(main):047:0> list**

‘

‘DESCRIBE’ command to describe the details and configuration of the HBase table. For example, version, compression, blocksize, replication etc. The syntax to describe the table is as follows.

Syntax: describe <‘namespace’:’table\_name’>

**hbase(main):047:0>describe ‘emp’**



‘Put’  command to insert data to rows and columns on a table. This would be similar to insert statement on RDBMS but, the syntax is completely different.

Syntax: put ‘<name\_space:table\_name>’, ‘<row\_key>’ ‘<cf:column\_name>’, ‘<value>’

In above examples, notice that we have added 2 rows; row key ‘1’ with one column ‘office:name’ and row key ‘2’ with three columns ‘office:name’, ‘office:gender’ and ‘office:age’. If you are coming from RDBMS world, you probably would confuse with this. Once you understand how column database works it’s not that difficult to get around it.

Also, note that last command from above example actually inserts a new column ‘office:age’ at row key ‘2’ with ’50’

Internally, HBase doesn’t do an update but it assigns a column with new timestamp and scan fetches the latest data from columns.





‘Scan’  command to get the data from the HBase table. By default, it fetches all data from the table.

Syntax: scan ‘<name\_space:table\_name>’



This scan’s the ’emp’ table to return name and age columns from starting row 1 and ending row 3.









Use ‘disable’ to disable a table. Prior to delete a table or change its setting, first, you need to disable the table. The syntax to disable the table is as follows.

Syntax: disable ‘<namespace>:<table\_name>’



Use is\_disabled to check if the table is disabled. When it disabled it returns ‘true‘



Let’s check if the table disabled by using describe



‘enable’ command is used to enable a disabled table. You need to enable a disabled table first to perform any regular commands., The syntax to enable the table is as follows.

Syntax: enable ‘<namespace>:<table\_name>’

Syntax: enable ‘<namespace>:<table\_name>’



Use drop command to delete a table. You should disable a table first before you drop it.

Syntax: drop ‘<table\_name>’



Use drop\_all command to delete many tables using a regular expression.

hbase(main):041:0> drop\_all ‘em.\*’

hbase(main):041:0> drop\_all ‘em.\*’





**1)Creating a Table using HBase Shell**

We can create a table using the create command, here you must specify the table name and

the Column Family name. The syntax to create a table in HBase shell is shown below.

create ‘<table name>’,’<column family>’

>create 'customer', 'address','order'



**2)Put: Inserts a new record into the table with row identified by 'row..'**

This command is used for following things

• It will put a cell 'value' at defined or specified table or row or column. •

It will optionally coordinate time stamp.

Syntax: put <'tablename'>,<'rowname'>,<'columnvalue'>,<'value'>

Example – with the help of put commands we have inserted new records in “customer”

table for address and order family. Here name “Nick” is Row key

put 'customer','Nick','address:city','Mumbai'

put 'customer','Nick','address:state','Maharashtra'

put 'customer','Nick','address:street','Steet1'

put 'customer','Nick','order:number','ORD-15'

put 'customer','Nick','order:amount','50'

Adding one more record of customer name “Justin”. Here name “Justin” is Row key

put 'customer','Justin','address:city','Pune'

put 'customer','Justin,'address:state','Maharashtra'

put 'customer','Justin','order:number','ORD-16'

put 'customer','Justin','order:amount','60'





3)Get: Returns the records matching the row identifier provided in the table

By using this command, you will get a row or cell contents present in the table. In addition

to that you can also add additional parameters to it like TIMESTAMP,

TIMERANGE,VERSIONS, FILTERS, etc. to get a particular row or cell content.

Syntax: get <'tablename'>, <'rowname'>, {< Additional parameters>}

**a) get 'customer', 'Nick'**

By executing above commands we are getting all the content of customer “Nick” as shown

in below screenshot.



**b) Additional parameters to get only address details get 'customer', 'Nick' , 'address'**

By executing above commands we are getting all the content of family address for row key

“Nick” from “customer” table as per screenshot below.



**c) Additional parameters to get only city details get 'customer', 'Nick' , 'address:city'**

By executing above commands we are getting all the content of family address city for row

key “Nick” from “customer” table as per screenshot below.



**4)Scan :The scan command is used to view the data in HTable. Using the scan command, you can get the table data.**

• This command scans entire table and displays the table contents.

• We can pass several optional specifications to this scan command to get more

information about the tables present in the system.

• Scanner specifications may include one or more of the following attributes. o

These are TIMERANGE, FILTER, TIMESTAMP, LIMIT, MAXLENGTH,

COLUMNS, CACHE, STARTROW and STOPROW.

Its syntax is as follows:

Syntax: scan <'tablename'>, {Optional parameters}

**scan 'customer'**

When we execute above commands in HBase then we will be getting all the table

“customer” contents along with additional parameters like timestamp as show in below

screenshot.



**5)Delete -Using the delete command, you can delete a specific cell in a table.**

• This command will delete cell value at defined table of row or column.

• Delete must and should match the deleted cells coordinates exactly.

• When scanning, delete cell suppresses older versions of values.

The syntax of delete command is as follows:

Syntax:delete <'tablename'>,<'row name'>,<'column name'>

**delete 'customer', 'Nick' ,'address:street'**

The above command below delete street from address family for row key “Nick” from

“customer” table.















**6)Alter** - This command alters the column family schema. To understand what exactly it does, we have explained it here with an example. Alter commands are useful for below cases -

• Altering single, multiple column family names

• Deleting column family names from table

• Several other operations using scope attributes with table

Syntax: alter <tablename>, NAME=><column familyname>, VERSIONS=>5

We can delete specific column family by using alter commands

alter 'customer', 'delete' => 'address'





After deleting “address” family from “customer” table. Let’s again check customer table

using “scan” commands as follow

>scan ‘customer’

As you can see from below screenshot we only now have order:amount and order:number.

















