**Installation of Hortonworks Virtual Machine:**

1-Go to the this url and download for macos x version:

<https://www.virtualbox.org/wiki/Downloads>

2-Install the virtualbox tool.

3-Go to the this url and download hortonworks sandbox for virtual box:

<https://hortonworks.com/downloads/#sandbox>

Create a new virtual machine on virtualbox, and setup the Hortonworks sandbox.

**Usage of Hortonworks Sandbox:**

1.Via virtualbox sandbox:

Terminal can be used via sandbox after installation.

2.Through the browser:

This is possible to connect dashboard from browser through <http://127.0.0.1:8888>

To login information:

Username: maria\_dev

Password: maria\_dev

3.Connect with your terminal:

You can connect your virtual machine by using ssh command: ‘-ssh [root@127.0.0.1](mailto:root@127.0.0.1) -p 2222’

**Hbase Commands and Examples**

**Status**: General view of Hbase (number of servers).

->status

1 active master, 0 backup masters, 1 servers, 0 dead, 11.0000 average load

**Version**: The version of the Hbase.

->version

1.1.2.2.6.0.3-8, r3307790b5a22cf93100cad0951760718dee5dec7, Sat Apr 1 21:41:47 UTC 2017

**Create**: “create <table name> , <column family>” : Create a table in Hbase.

->Create 'Saltuk\_Tutorial\_Table','Column1','Column2'

0 row(s) in 1.3110 seconds

**List**: List all tables that created before.

->list

TABLE

ATLAS\_ENTITY\_AUDIT\_EVENTS

Gdelt

atlas\_titan

driver\_dangerous\_event

driver\_dangerous\_events

iemployee

testTable

7 row(s) in 0.3980 seconds

=> ["ATLAS\_ENTITY\_AUDIT\_EVENTS", "Gdelt", "atlas\_titan", "driver\_dangerous\_event", "driver\_dangerous\_events", "iemployee", "testTable"]

**Disable**: “disable ‘table\_name’: To delete a table or change table settings first thing we need to do disable that table.

--disable 'Saltuk\_Tutorial\_Table'

0 row(s) in 2.4520 seconds

**Enable**: “enable ‘table\_name’: To enable a table that disabled before.

->Enable 'Saltuk\_Tutorial\_Table'

0 row(s) in 1.2820 seconds

**Describe**: “describe ‘table\_name’: Description of the table.

->Describe 'Saltuk\_Tutorial\_Table'

Table Saltuk\_Tutorial\_Table is ENABLED

Saltuk\_Tutorial\_Table

COLUMN FAMILIES DESCRIPTION

{NAME => 'Column1', BLOOMFILTER => 'ROW', VERSIONS => '1', IN\_MEMORY => 'false', KEEP\_DELETED\_CELLS => 'FALSE', DATA\_BLOCK\_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN\_VERSIONS => '0',

BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION\_SCOPE => '0'}

{NAME => 'Column2', BLOOMFILTER => 'ROW', VERSIONS => '1', IN\_MEMORY => 'false', KEEP\_DELETED\_CELLS => 'FALSE', DATA\_BLOCK\_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN\_VERSIONS => '0',

BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION\_SCOPE => '0'}

2 row(s) in 0.0270 seconds

**Exists**: “exists ‘table\_name’:To check table is exists or not.

->exists 'table\_test'

Table table\_test does not exist

0 row(s) in 0.0110 seconds

**Drop**: “drop ‘table\_name”:To delete a table.But you need to first disable that table.

->disable 'Saltuk\_Tutorial\_Table'

0 row(s) in 2.2630 seconds

->drop 'Saltuk\_Tutorial\_Table'

0 row(s) in 1.2680 seconds

->list

TABLE

ATLAS\_ENTITY\_AUDIT\_EVENTS

Gdelt

atlas\_titan

driver\_dangerous\_event

driver\_dangerous\_events

iemployee

testTable

7 row(s) in 0.0100 seconds

**Exit**:To exit from Hbase Shell.

**Insert/Update Data**: “put ‘table\_name’ , ‘row’ , ‘<columnFamily:columnName>’,’<value>’ “: Insert data to given column, or update data.

->scan 'iemployee'

ROW COLUMN+CELL

1 column=insurance:dental, timestamp=1493989899812, value=metlife

1 column=insurance:health, timestamp=1493989899801, value=anthem

1 column=insurance:life, timestamp=1493989899842, value=metlife

1 column=insurance:vision, timestamp=1493989899828, value=visionOne

1 column=payroll:grade, timestamp=1493989899756, value=G16

1 column=payroll:salary, timestamp=1493989899768, value=250000.00

1 column=personal:city, timestamp=1493989899696, value=San Fransisco

1 column=personal:fname, timestamp=1493989899628, value=Mike

1 column=personal:lname, timestamp=1493989899673, value=Young

1 column=personal:zip, timestamp=1493989899737, value=12345

1 column=skills:interpersonal-rating, timestamp=1493989899792, value=medium

1 column=skills:management, timestamp=1493989899780, value=executive,creator,innovative

1 row(s) in 0.2980 seconds

->put 'iemployee','1','personal:phone\_num','5055150044'

0 row(s) in 0.0950 seconds

->scan 'iemployee'

ROW COLUMN+CELL

1 column=insurance:dental, timestamp=1493989899812, value=metlife

1 column=insurance:health, timestamp=1493989899801, value=anthem

1 column=insurance:life, timestamp=1493989899842, value=metlife

1 column=insurance:vision, timestamp=1493989899828, value=visionOne

1 column=payroll:grade, timestamp=1493989899756, value=G16

1 column=payroll:salary, timestamp=1493989899768, value=250000.00

1 column=personal:city, timestamp=1493989899696, value=San Fransisco

1 column=personal:fname, timestamp=1493989899628, value=Mike

1 column=personal:lname, timestamp=1493989899673, value=Young

1 column=personal:phone\_num, timestamp=1502177673670, value=5055150044

1 column=personal:zip, timestamp=1493989899737, value=12345

1 column=skills:interpersonal-rating, timestamp=1493989899792, value=medium

1 column=skills:management, timestamp=1493989899780, value=executive,creator,innovative

1 row(s) in 0.0460 seconds

**Read Data**: “get ‘table\_name’ , ’row’ “ : To read row data in table .

->get 'iemployee','1'

COLUMN CELL

insurance:dental timestamp=1493989899812, value=metlife

insurance:health timestamp=1493989899801, value=anthem

insurance:life timestamp=1493989899842, value=metlife

insurance:vision timestamp=1493989899828, value=visionOne

payroll:grade timestamp=1493989899756, value=G16

payroll:salary timestamp=1493989899768, value=250000.00

personal:city timestamp=1493989899696, value=San Fransisco

personal:fname timestamp=1493989899628, value=Mike

personal:lname timestamp=1493989899673, value=Young

personal:phone\_num timestamp=1502177673670, value=5055150044

personal:zip timestamp=1493989899737, value=12345

skills:interpersonal-rating timestamp=1493989899792, value=medium

skills:management timestamp=1493989899780, value=executive,creator,innovative

13 row(s) in 0.0700 seconds

**Delete Data**: “delete ‘table\_name’ , ‘row’ ,’columnFamily:columnName’ , ‘timeStamp’ ”:Delete a specific data from table.

->delete 'iemployee','1','personal:lname'

0 row(s) in 0.0400 seconds

->get 'iemployee','1'

COLUMN CELL

insurance:dental timestamp=1493989899812, value=metlife

insurance:health timestamp=1493989899801, value=anthem

insurance:life timestamp=1493989899842, value=metlife

insurance:vision timestamp=1493989899828, value=visionOne

payroll:grade timestamp=1493989899756, value=G16

payroll:salary timestamp=1493989899768, value=250000.00

personal:city timestamp=1493989899696, value=San Fransisco

personal:fname timestamp=1493989899628, value=Mike

personal:phone\_num timestamp=1502177673670, value=5055150044

personal:zip timestamp=1493989899737, value=12345

skills:interpersonal-rating timestamp=1493989899792, value=medium

skills:management timestamp=1493989899780, value=executive,creator,innovative

12 row(s) in 0.0630 seconds

**Scan**:”scan ‘table\_name’ ” : To use view Htable in Hbase.

->scan 'iemployee'

ROW COLUMN+CELL

1 column=insurance:dental, timestamp=1493989899812, value=metlife

1 column=insurance:health, timestamp=1493989899801, value=anthem

1 column=insurance:life, timestamp=1493989899842, value=metlife

1 column=insurance:vision, timestamp=1493989899828, value=visionOne

1 column=payroll:grade, timestamp=1493989899756, value=G16

1 column=payroll:salary, timestamp=1493989899768, value=250000.00

1 column=personal:city, timestamp=1493989899696, value=San Fransisco

1 column=personal:fname, timestamp=1493989899628, value=Mike

1 column=personal:phone\_num, timestamp=1502177673670, value=5055150044

1 column=personal:zip, timestamp=1493989899737, value=12345

1 column=skills:interpersonal-rating, timestamp=1493989899792, value=medium

1 column=skills:management, timestamp=1493989899780, value=executive,creator,innovative

1 row(s) in 0.0210 seconds

**Count**: ”count ‘table\_name’ “: Counts the row in the Htable.

->count 'driver\_dangerous\_event'

4 row(s) in 0.0450 seconds

=> 4

**Pig Commands and Examples**

1.Go to: 127.0.0.1:8888

2.Login with maria\_dev, username:maria\_dev, password:maria\_dev

3.Download the dataset sample:

<https://drive.google.com/open?id=0B1vC0K2TfbWteklkYXBTWFI1OFE>

4. Upload all data sets to FilesView->usr/Maria\_Dev

5.Open the PigView.

**Load**: ”veriableName = LOAD ‘data\_path’ USING PigStorage( ‘,’);” : To assign data set as veriable.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

**Dump**:”dump relation\_name;” : To display result on the screen.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

Dump student;

**Describe**:”describe relation\_name;” : Scheme of data. Describing the data.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

Describe student;

**Group**:”group\_data\_name = Group relation\_name By relation ;” : To use group data that in one relation or more.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

group\_data = GROUP student by age;

**Cogroup**: ”cogroup\_data\_name = Cogroup group1 By relation1 , group2 By relation2 ;“ : To make group 2 or more .different dataset or relation.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

employee = LOAD 'employee.txt' USING PigStorage(',')

as (id:int, name:chararray, age:int, city:chararray);

cogroup\_data = COGROUP student by age, employee by age;

**Join**: “relation\_name\_join = Join relation1 By key1, relation2 By key2 ;” : To combine records from 2 or more relations.

customers = LOAD 'customers.txt' USING PigStorage(',')

as (id:int, name:chararray, age:int, address:chararray, salary:int);

orders = LOAD 'orders.txt' USING PigStorage(',')

as (oid:int, date:chararray, customer\_id:int, amount:int);

test = JOIN customers BY id, orders BY customer\_id;

**Union**:”union\_data = Union relation1 , relation2 ;” : To merge the content of 2 relations.Their columns must be identical.

student 1= LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

student2 = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

union\_student = Union student1,student2;

**Split**: Split relation\_name Into split\_relation1 If (condition) , split\_relation2 If (condition);” : Split relation into two or more relation.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

Split student Into student\_detais1 if age<23, student\_details2 if (age>22 and age<25);

**Filter**:”filtered\_data = FILTER data\_name By (condition) ;” : Select from data based on a condition.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

filter\_data = FILTER student BY city == 'Chennai';

**Foreach**:”foreach\_data = Foreach data\_name Generate (required data);” : Generates specified data transformations.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

foreach\_data = FOREACH student GENERATE id,age,city;

**Order** **By**: “ordered\_data = Order relation\_name By parameter ASC/DESC ;” : Sort data based on parameter.

student = LOAD 'data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:int, phone:chararray,

city:chararray );

order\_data = Order student By firstname ASC;

**Resources:**

* <https://www.tutorialspoint.com/hbase/>
* <https://www.tutorialspoint.com/apache_pig/>
* <https://hortonworks.com/hadoop-tutorial/introduction-apache-hbase-concepts-apache-phoenix-new-backup-restore-utility-hbase/>
* <https://hortonworks.com/tutorial/how-to-process-data-with-apache-pig/>
* <http://hbase.apache.org/book.html#quickstart>
* <https://hortonworks.com/tutorial/hadoop-tutorial-getting-started-with-hdp/>
* <http://pig.apache.org/docs/r0.17.0/start.html>