**Assignment 5:- Hive and HBASE**

**Associated Data Files**

This Data set is about Olympics. You can download the data set from the below link:

https://drive.google.com/open?id=0ByJLBTmJojjzV1czX3Nha0R3bTQ

**DATE SET DESCRIPTION**

The data set consists of the following fields.

Athlete: This field consists of the athlete name

Age: This field consists of athlete ages

Country: This fields consists of the country names which participated in Olympics

Year: This field consists of the year

Closing Date: This field consists of the closing date of ceremony

Sport: Consists of the sports name

Gold Medals: No. of Gold medals

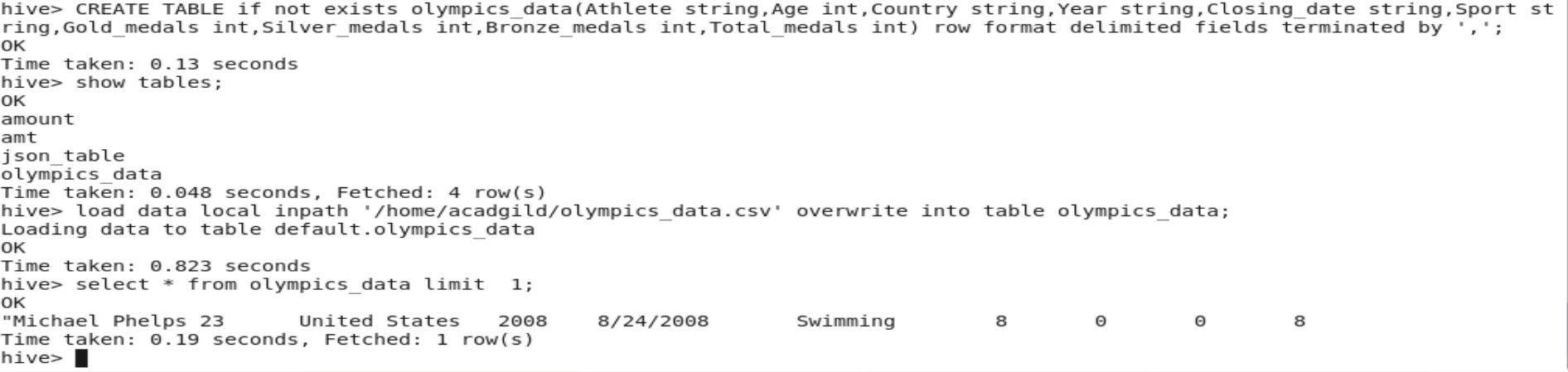
Silver Medals: No. of Silver medals

Bronze Medals: No. of Bronze medals

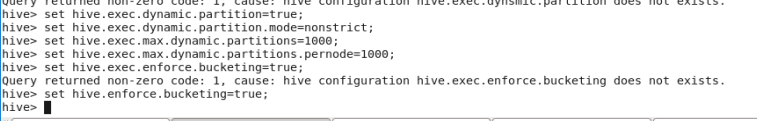
Total Medals: Consists of total no. of medals

First we will prepare the data set ready to perform all the actions asked in the problem statement below.

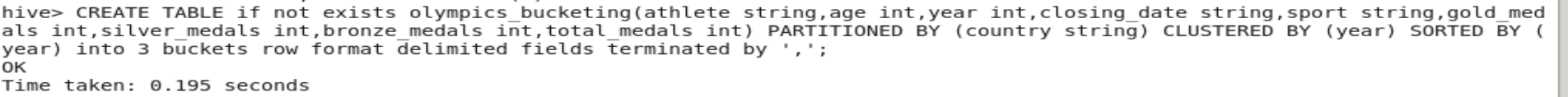
* Hive commands to create database, use the database just created and create a table
* Load the Olympics.csv file into Olympics\_data table



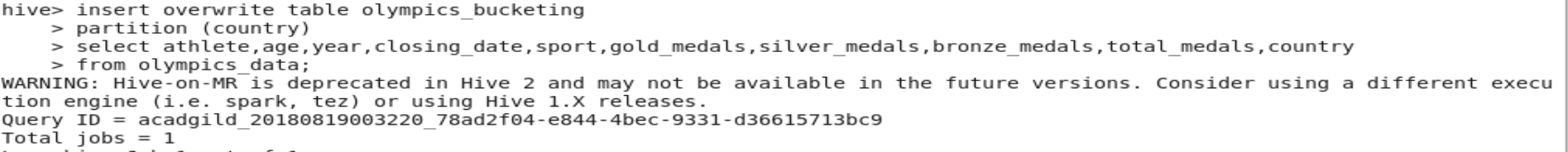
* **Now we will create PARTITION AND BUCKETING operation:**
  1. Hive partition divides table into number of **partitions** and these partitions can be further subdivided into more manageable parts known as **Buckets** or **Clusters**.
  2. In Hive Partition, each partition will be created as directory. But in Hive Buckets, each bucket will be created as file.
* **Setting the Below Properties in Hive Command Line Before Proceeding Further for Bucketing Scripts:**



* **Then create the table with these optimization technics as shown in the below screen shot:**



* **Now insert the contents of the Olympic\_table into the Olympic\_bucket table as shown in the below screen shot**



* **List the top 1 record from Olympic\_bucket table as shown below**

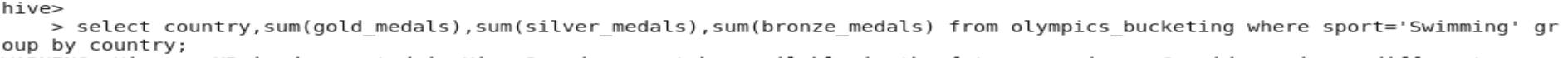


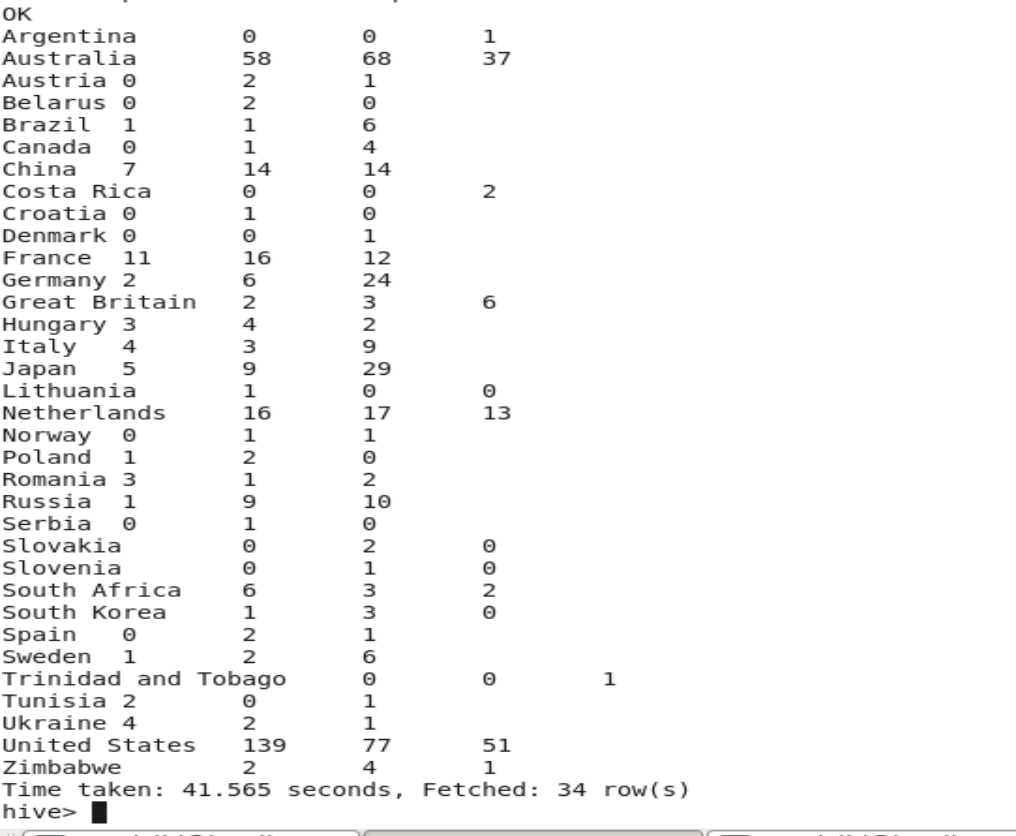
**Problem Statement**

**Task 1.1**

1. Write a Hive program to find the number of medals won by each country in swimming.

**Query and output:-**





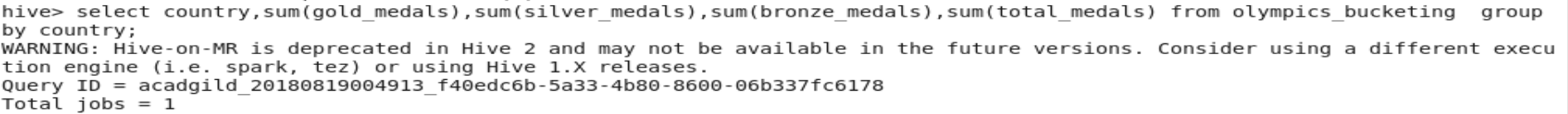
1. Write a Hive program to find the number of medals that India won year wise.

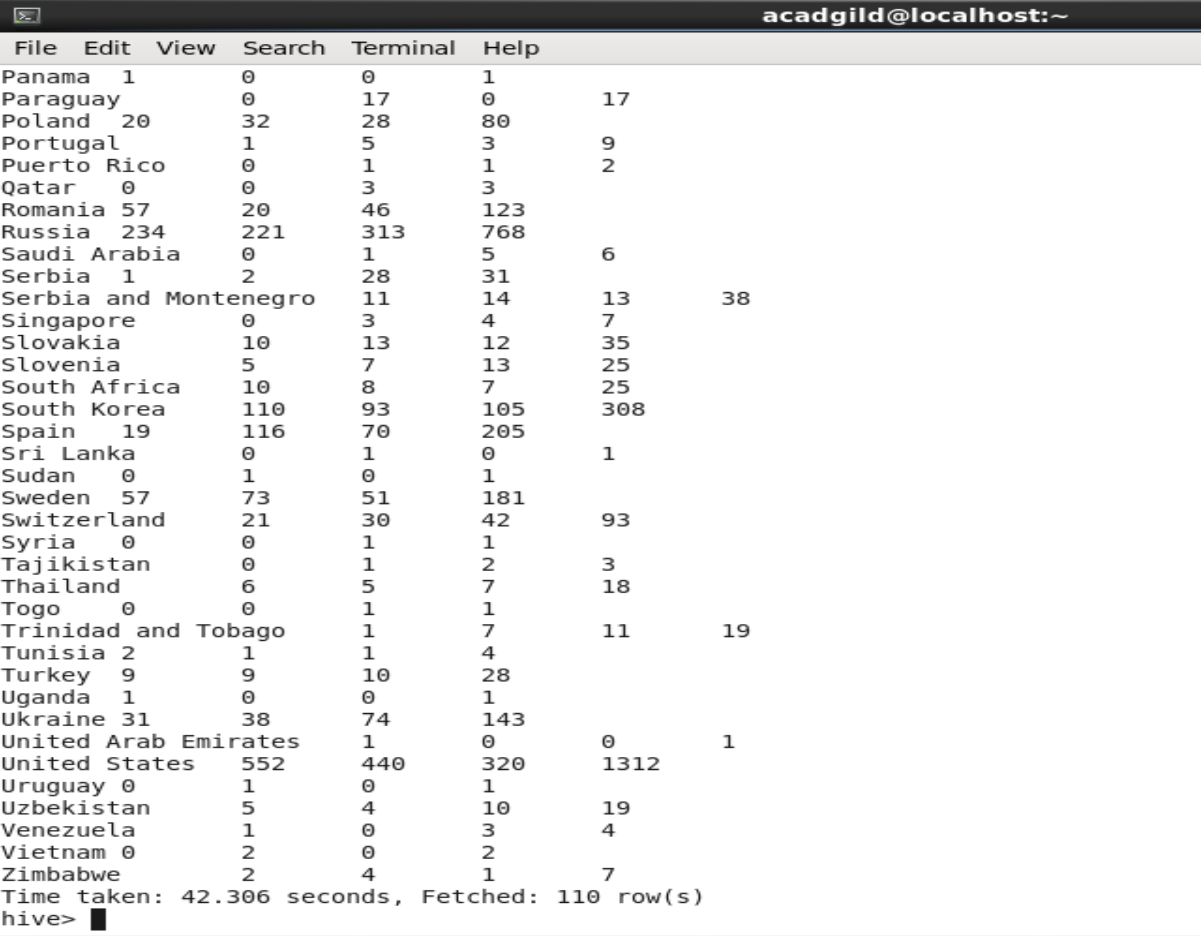
**Query and Output:-**



1. Write a Hive Program to find the total number of medals each country won.

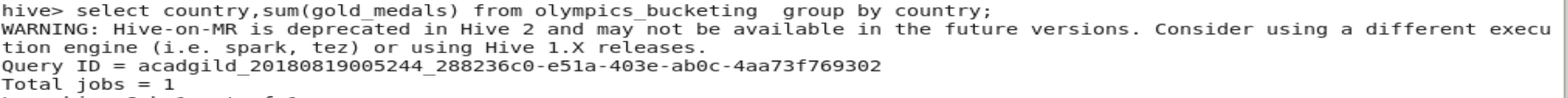
**Query and Output:-**

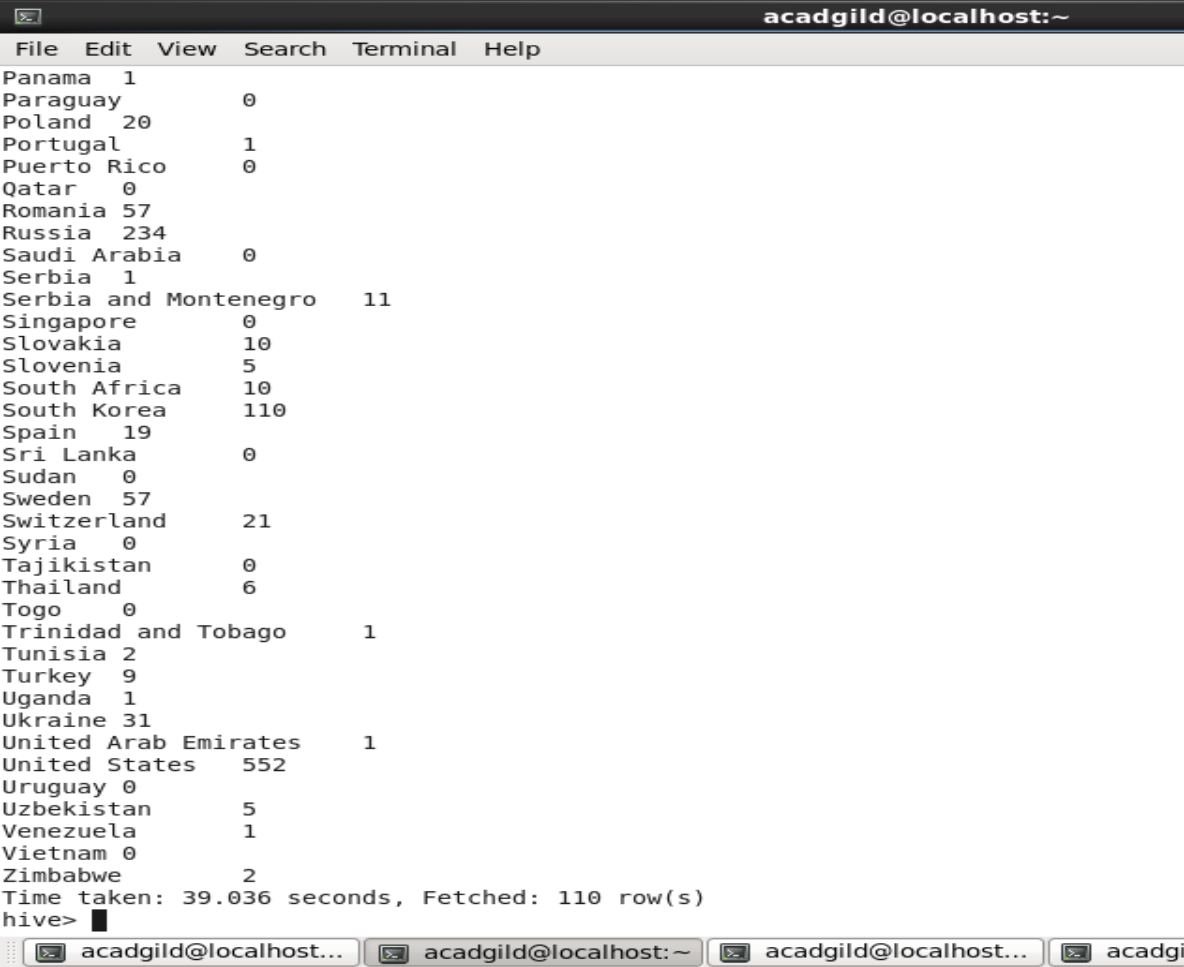




1. Write a Hive program to find the number of gold medals each country won.

**Query and Output:-**





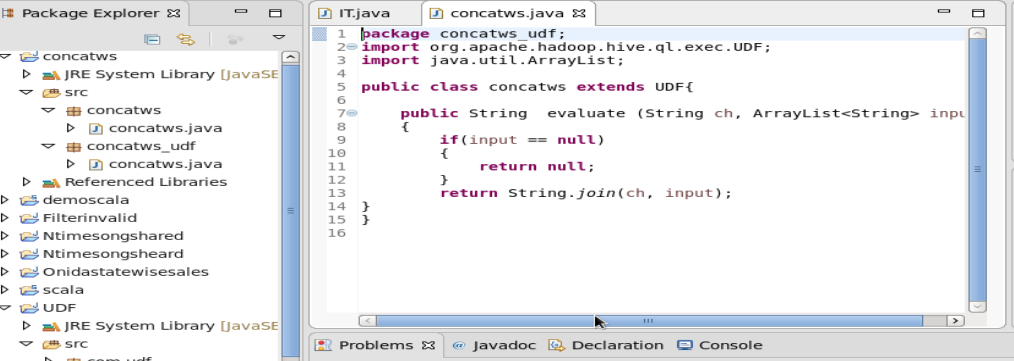
**Task 1.2**

Write a hive UDF that implements functionality of string concat\_ws(string SEP, array<string>). This UDF will accept two arguments, one string and one array of string.

It will return a single string where all the elements of the array are separated by the SEP.

**Java Program:**

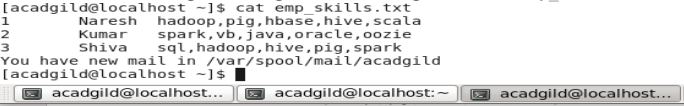
Following is the java program that will take two arguments: separator string as ch and array list of string as input and returns the joined string to the calling command:



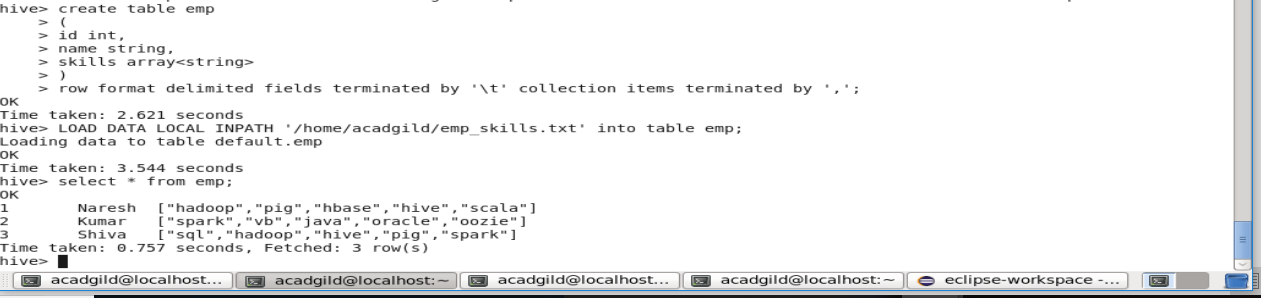
Then I have created a jar file from the above program with the name **concatws.jar as shown below**



Now I have created the emp\_skills.txt file which contains the below information as shown below.



Created the emp table and Load the emp\_skills data into emp table as shown below.



Then I have registered this jar to existing hive session with the following command:

* ADD JAR concatws.jar;

Then I have created a temporary function with the following command:

* CREATE TEMPORARY FUNCTION con AS ‘concatws\_udf.concatws’;



Then I have executed the following command to work with udf function:

* Select con(“|”, skills) from emp;
* Select con(“:”, skills) from emp;
* Select skills from emp;



**Task 1.3**

Link: https://acadgild.com/blog/transactions-in-hive/

Refer the above given link for transactions in Hive and implement the operations given in the blog using your own sample data set and send us the screenshot.

**Settings:**



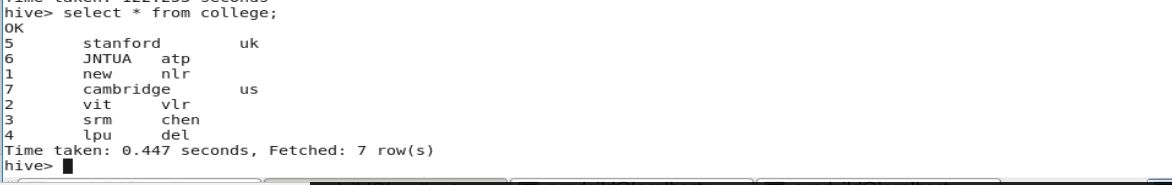
**Create Table in Hive Table**



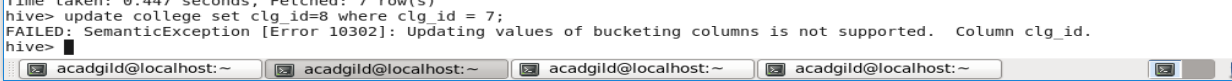
**Inserting the Data in Hive Table**

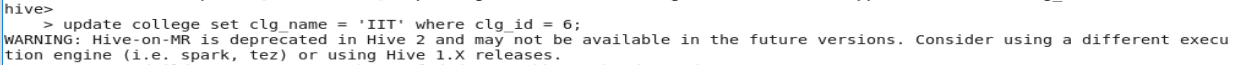


**Selecting the Data in Hive Table**



**Updating the Data in Hive Table**





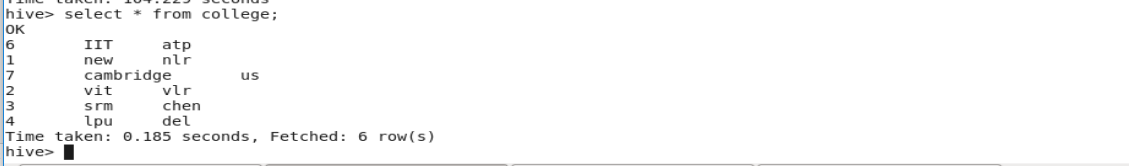
**Output:-**



**Deleting a Row from Hive Table**



**Output:-**



**Task 2.1**

Answer in your own words with example.

1. What is NoSQL data base?

NoSQL is an approach to database design that can accommodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.

1. How does data get stored in NoSQL database?

Data Storage is not based on a single data model. Most outstanding ones are key-value pair, graph, document, and columnar.

1. What is a column family in HBase?

A column family defines shared features to all columns that are created within them (think of it almost as a sub-table within your larger table). You will notice that HBase columns are composed of a combination of the column family and column qualifier (or column key): 'family:qualifier'  
  
Where the qualifier can be an arbitrary array of bytes, the column family has to be composed of printable characters. Also, on HDFS, the column family is what is stored in human-readable format, example: '/hbase/table/region/<colfamX>'  
  
Remember - all columns families must be created up-front whereas columns can be added on the fly. Hence understanding the design of your data access pattern is crucial when first building your HBase instance.

1. How many maximum number of columns can be added to HBase table?

It has been recommended to keep the number of column families under three. But there is no magic number like this. Why not two? Why not four? Technically, HBase can manage more than three of four column families. However, you need to understand how column families work to make the best use of them.

1. Why columns are not defined at the time of table creation in HBase?

An HBase table is made of column families which are the logical and physical grouping of columns. The columns in one family are stored separately from the columns in another family.

A single column family contains one or more columns, Column families must be defined at table creation time but columns can be added dynamically after table creation (if an insert statement states a column that does not exist for a column family it will create it).

1. How does data get managed in HBase?

The data in the HBASE is managed in such a way that,

* **WAL:** Write Ahead Log is a file on the distributed file system. The WAL is used to store new data that hasn't yet been persisted to permanent storage; it is used for recovery in the case of failure.
* **BlockCache:** is the read cache. It stores frequently read data in memory. Least Recently Used data is evicted when full.
* **MemStore:** is the write cache. It stores new data which has not yet been written to disk. It is sorted before writing to disk. There is one MemStore per column family per region.
* **Hfiles** store the rows as sorted KeyValues on disk.

All the above are sub components of Region server which manages the data in the HBASE.

1. What happens internally when new data gets inserted into HBase table?

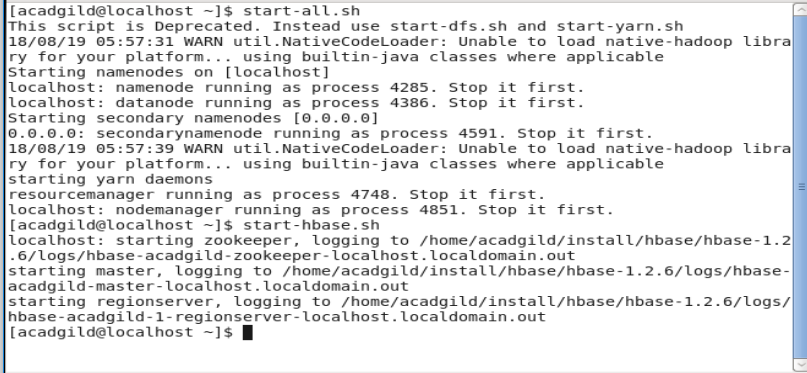
* The client gets the Region server that hosts the META table from ZooKeeper.
* The client will query the .META. Server to get the region server corresponding to the row key it wants to access. The client caches this information along with the META table location.
* It will get the Row from the corresponding Region Server.

For future reads, the client uses the cache to retrieve the META location and previously read row keys. Over time, it does not need to query the META table, unless there is a miss because a region has moved; then it will re-query and update the cache.

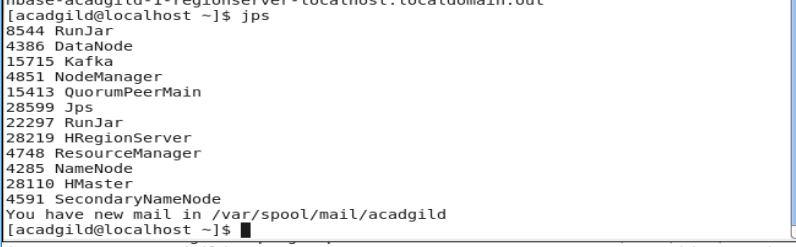
**Task 2.2**

To start working with HBase actual commands we need to start the Hadoop and HBase services with the following commands:

* Start-all.sh
* Start-hbase.sh

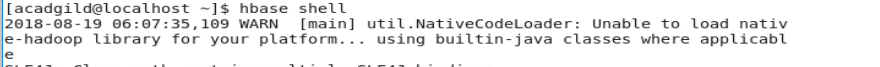


* Jps command to check whether the services started or not



1. Create an HBase table named 'clicks' with a column family 'hits' such that it should be able to store last 5 values of qualifiers inside 'hits' column family.

Below command to open the hbase.



Following is the command to create a table ‘clicks’ with column family ‘hits’ and having 5 versions:

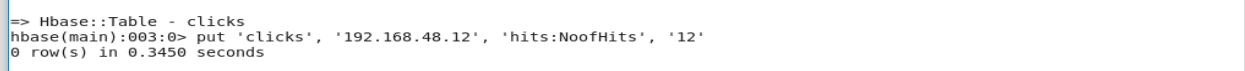
* Create ‘clicks’, NAME=>‘hits’, VERSIONS=>5



2. Add few records in the table and update some of them. Use IP Address as row-key. Scan the table to view if all the previous versions are getting displayed.

To add records in the clicks table we use below put command and we use row ip address 192.168.48.12 as row key:

* Put ‘clicks’, ‘192.168.48.12’, ‘hits:noOfHits’, ‘12’

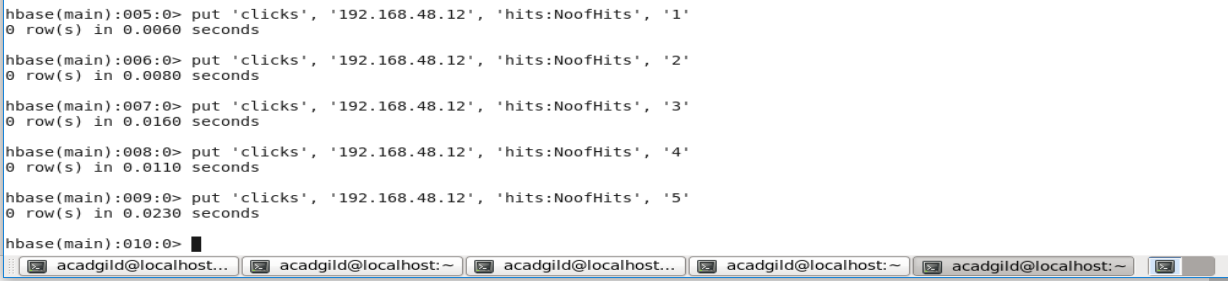


To view the contents of the table use scan command as show below:

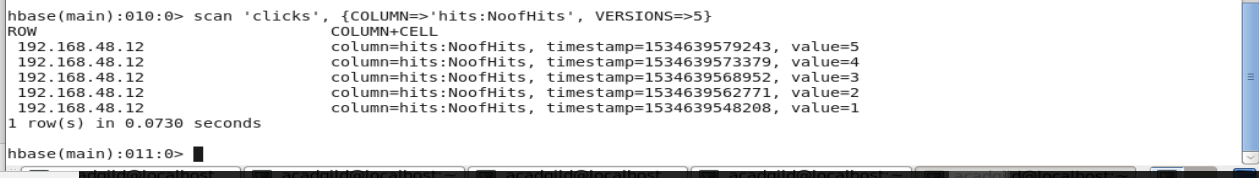
* Scan ‘clicks’



Now we will try to update the same row key for 5 times and check whether we are getting all the version values or not with the help of following commands used in the below screen shot:



Now we are going to see the last 5 qualifiers inside hits column family with the help of the following command used in the below screen shot:



Thanks,

Naresh Kumar