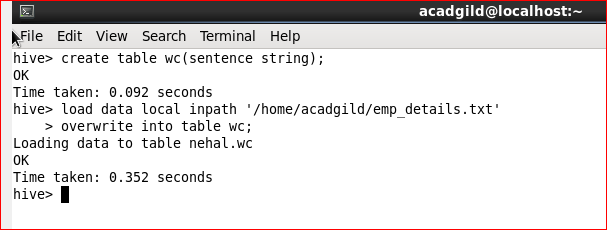
Problem Statement:

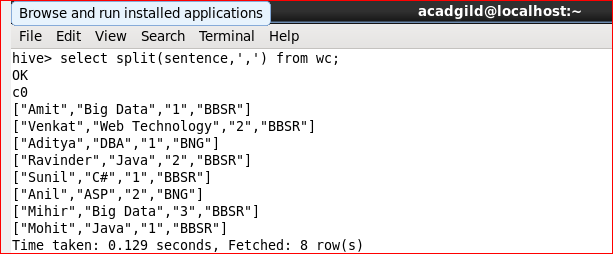
1. Perform word count in Hive for above given dataset

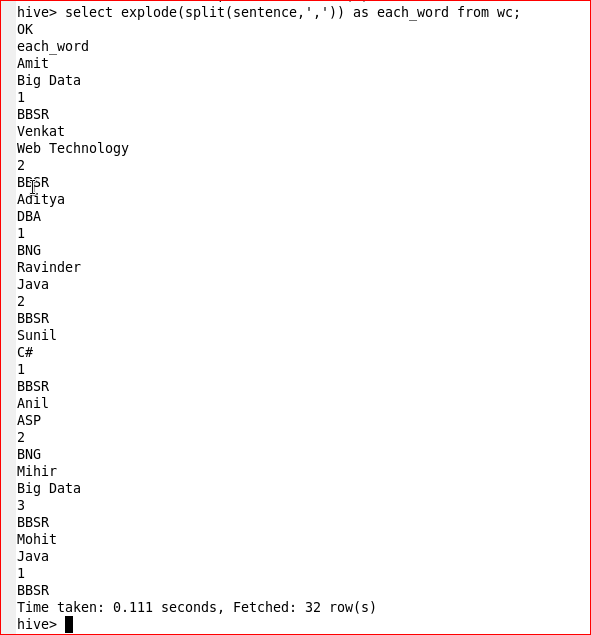
2. Explain the working of Partitioning in brief.

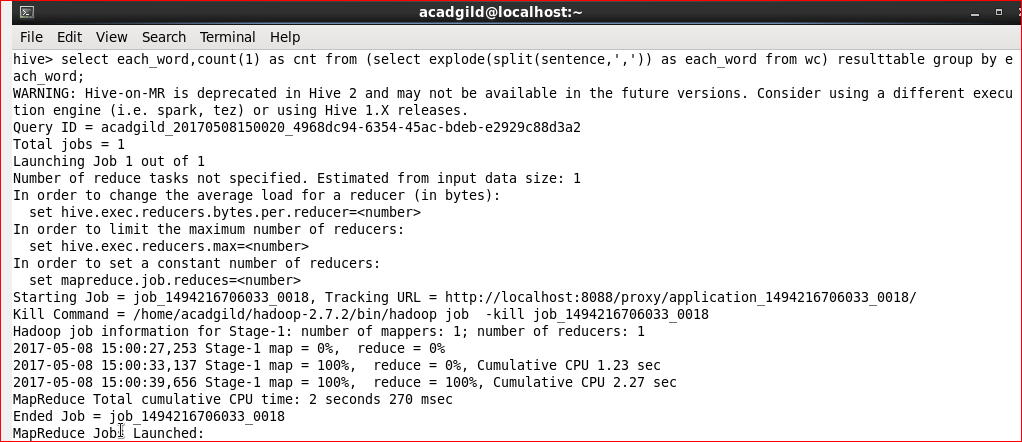
3. Explain the difference between Static and Dynamic Partitioning in Hive with an example.

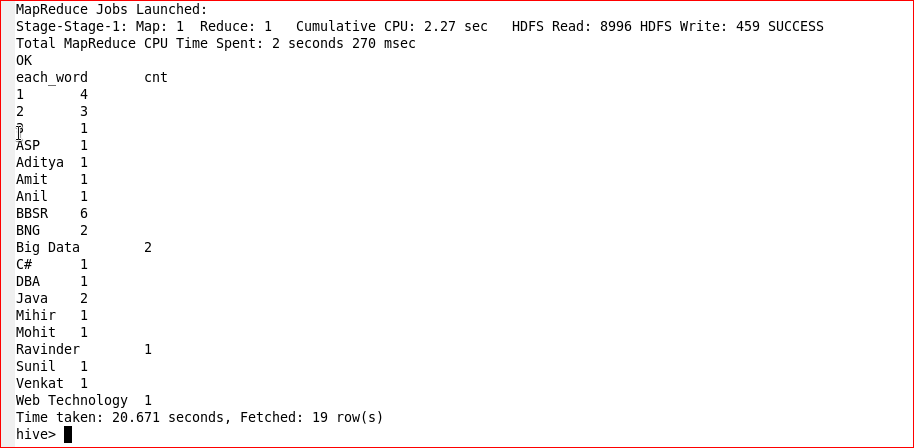
Q1. Perform word count in Hive for above given dataset.











Q2. Explain the working of Partitioning in brief.

Hive organizes tables into partitions. It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, and department. Using partition, it is easy to query a portion of the data.

Tables or partitions are sub-divided into **buckets,** to provide extra structure to the data that may be used for more efficient querying. Bucketing works based on the value of hash function of some column of a table.

For example, a table named **Tab1** contains employee data such as id, name, dept, and yoj (i.e., year of joining). Suppose we need to retrieve the details of all employees who joined in 2012. A query searches the whole table for the required information. However, if you partition the employee data with the year and store it in a separate file, it reduces the query processing time. The following example shows how to partition a file and its data:

The following file contains employee data table.

/tab1/employee data/file1

Id, name, dept, yoj

1, Gopal, TP, 2012

2, Kiran, HR, 2012

3, kaleel, SC, 2013

4, Prasanth, SC, 2013

The above data is partitioned into two files using year.

/tab1/employee data/2012/file2

1, Gopal, TP, 2012

2, kiran, HR, 2012

/tab1/employeedata/2013/file3

3, kaleel, SC, 2013

4, Prasanth, SC, 2013

Adding a Partition

We can add partitions to a table by altering the table. Let us assume we have a table called **employee** with fields such as Id, Name, Salary, Designation, Dept, and yoj.

Syntax:

ALTER TABLE table\_name ADD [IF NOT EXISTS] PARTITION partition\_spec

[LOCATION 'location1'] partition\_spec [LOCATION 'location2'] ...;

partition\_spec:

: (p\_column = p\_col\_value, p\_column = p\_col\_value, ...)

The following query is used to add a partition to the employee table.

hive> ALTER TABLE employee

> ADD PARTITION (year=’2013’)

> location '/2012/part2012';

Q3. Explain the difference between Static and Dynamic Partitioning in Hive with an example.

In Hive Partition concept there is two different type of partitions one is *Static Partition* and another one is Dynamic Partition. Here we will discussed about **Static partition vs dynamic partition in hive** differences.

In Hive There is two types of Partitions

i) Static Partition in Hive

ii) Dynamic Partition in Hive

Static Partition in Hive

* Insert input data files individually into a partition table is Static Partition
* Usually when loading files (big files) into Hivetables static partitions are preferred
* Static Partition saves your time in loading data compared to dynamic partition
* You “statically” add a partition in table and move the file into the partition of the table.
* We can alter the partition in static partition
* You can get the partition column value form the filename, day of date etc. without reading the whole big file.
* If you want to use Static partition in hive you should set property **set hive.mapred.mode = strict** This property set by default in hive-site.xml
* Static partition is in Strict Mode
* You should use where clause to use limit in static partition
* You can perform Static partition on Hive Manage table or external table

Example:

**CREATE** **TABLE** cityreport (cityid string, creport string, ctover string)  
partitioned **BY** (city string)  
row format delimited  
**FIELDS** terminated **BY** ‘|’  
stored **AS** textfile;

Loading data using static partitioning:

**LOAD DATA LOCAL inpath ‘/home/mahesh/hive-related/hyderabad.log’ INTO TABLE cityreport partition (city = ‘hyderabad’);**

Dynamic Partition in Hive

* single insert to partition table is known as dynamic partition
* Usually dynamic partition load the data from non-partitioned table
* Dynamic Partition takes more time in loading data compared to static partition
* When you have large data stored in a table then Dynamic partition is suitable.
* If you want to partition number of column but you don’t know how many columns then also dynamic partition is suitable
* Dynamic partition there is no required where clause to use limit.
* we can’t perform alter on Dynamic partition
* You can perform dynamic partition on hive external table and managed table
* If you want to use Dynamic partition in hive then mode is in nonstrict mode
* Here is hive dynamic partition properties you should allow

SET hive.exec.dynamic.partition = true;

SET hive.exec.dynamic.partition.mode = nonstrict;

Example:

hive> **INSERT** **INTO** **TABLE** t2 PARTITION(country) **SELECT** \* **FROM** T1;