**ASSIGNMENT – 27.6**

1. Using the below given two datasets you need to give a demo on the below joins in hive.

-Bucket Map join

- Sort-Merge Bucket join

- Sort-Merge Bucket Map join

- Left semi join

You can download the datasets from the below link <https://drive.google.com/open?id=0ByJLBTmJojjzSWhfXzI5NGh4RGM>

Dataset description of dataset1.csv is as follows: id,first\_name,last\_name,email,gender,ip\_address

Dataset description of tdataset2.csv is as follows: id,first\_name,last\_name

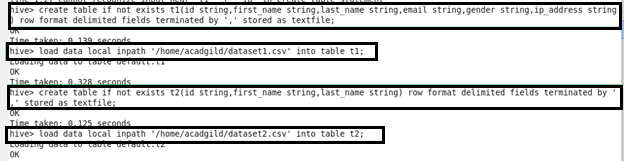
You can just join the two datasets using the above specified join types in hive.

**Bucket Map join:**

Tables being joined are bucketized on the join columns.

The number of buckets in one table is a multiple of the number of buckets in the other table, the buckets can be joined with each other.

Creating tables and loading the data into the created table.

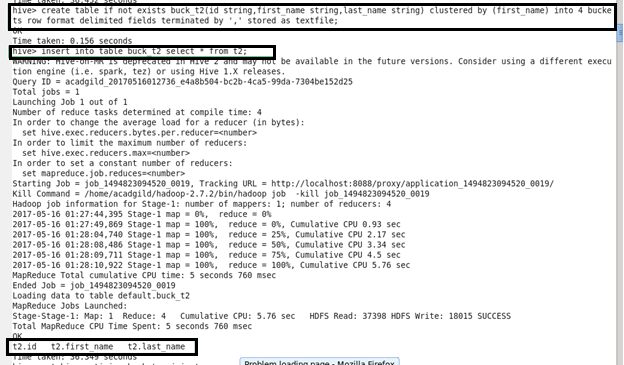


Creating bucketed tables to perform bucket map join and inserting data from the existing tables.

**Created table – buck\_t1 :**



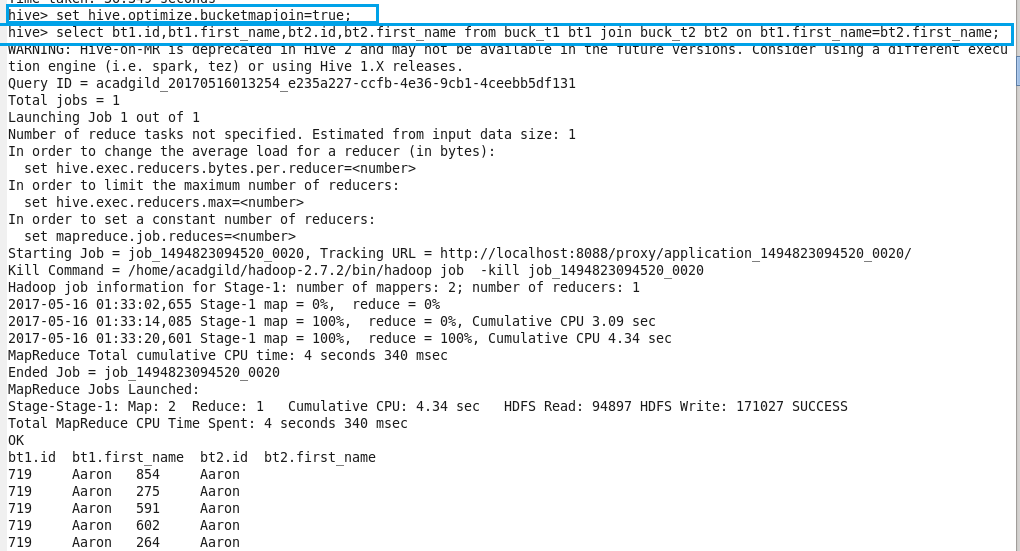
**Created table – buck\_t2 :**

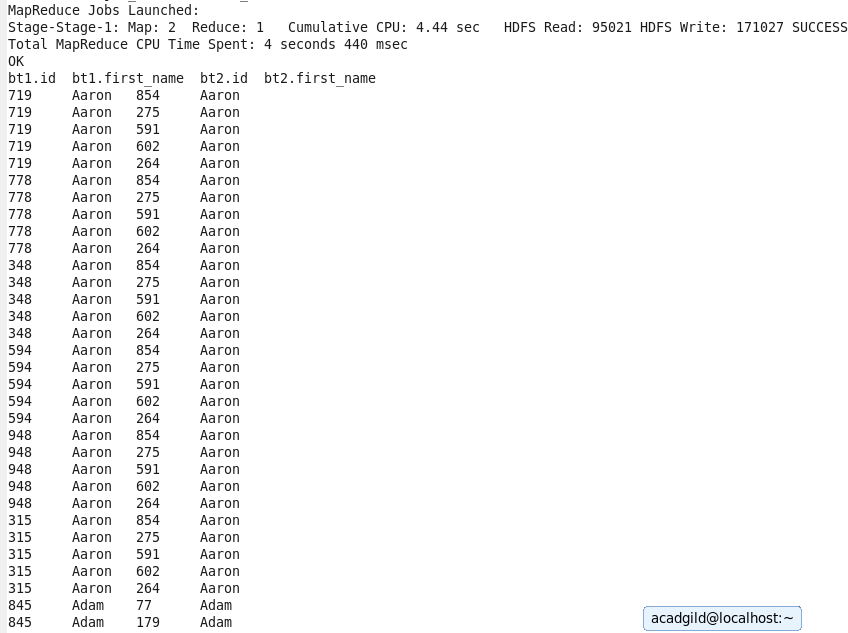


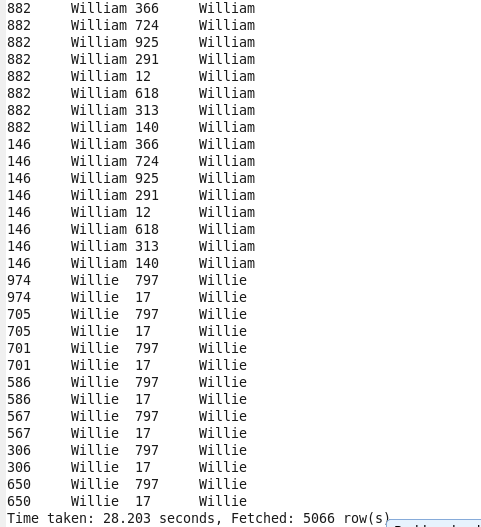
To perform bucket map join we need to set configuration i.e,

Set hive.optimize.bucketmapjoin=true;

With attached output :



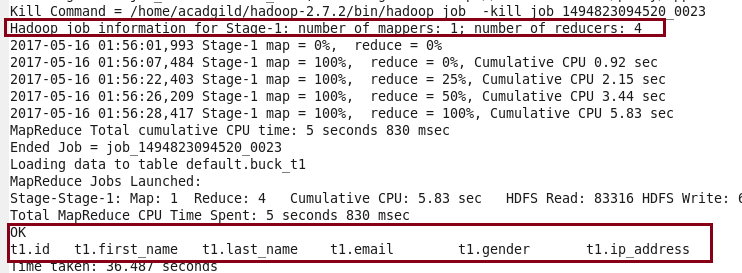
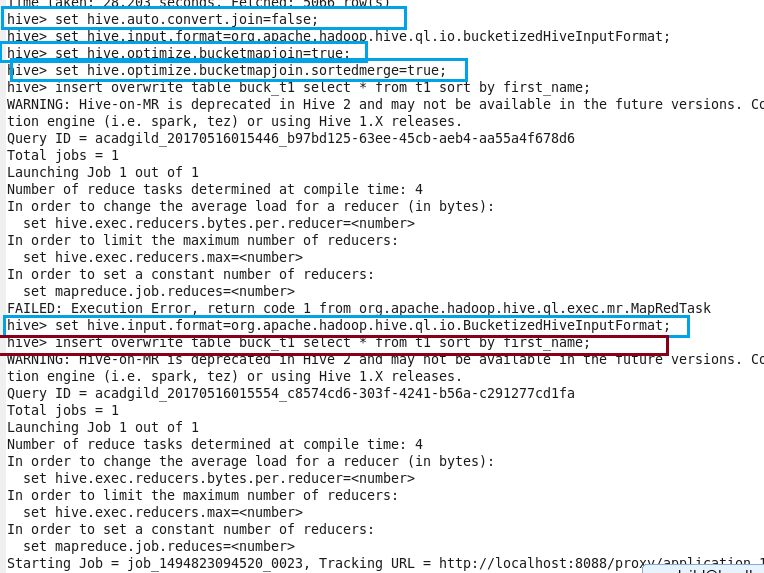


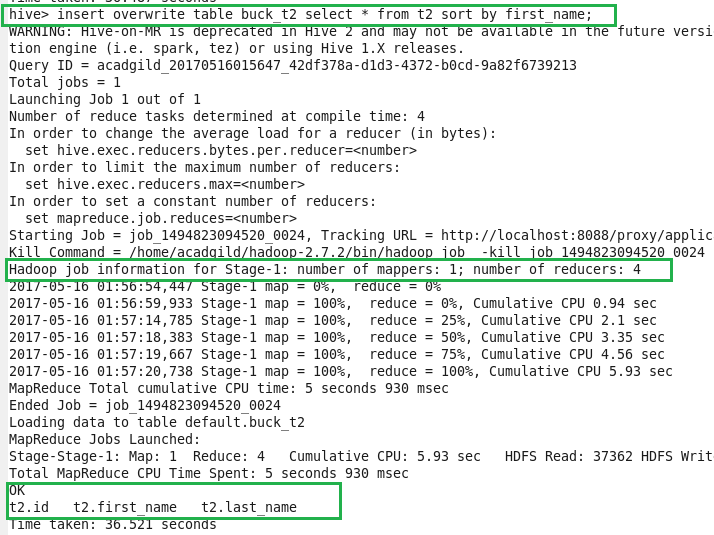


**Sort-Merge Bucket join :**

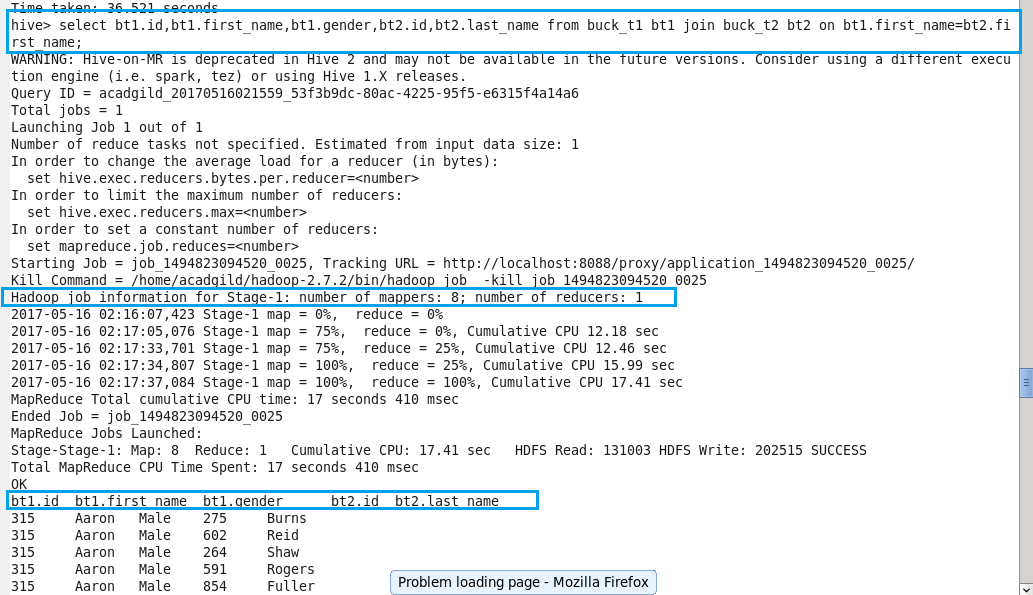
In SMB join in hive,each mapper needs a bucket from the first table and the corresponding bucket from the second table and then a merge sort join is performed. It is mainly used as there is no limit on file or partition or table join. It can best be used when the tables are large. In this join columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

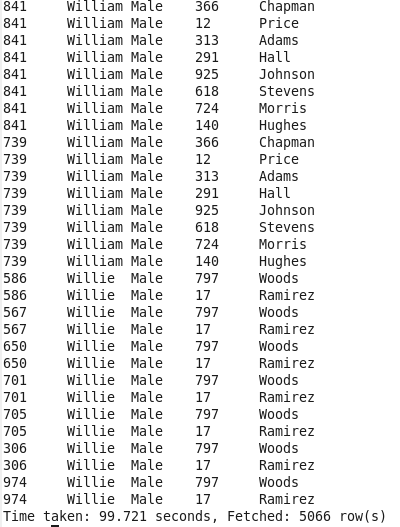
To perform SMB join we need to set following configuration showing in the attached screenshot then inserting the data in created bucketed tables.





**Performing SMB join attached with output :**

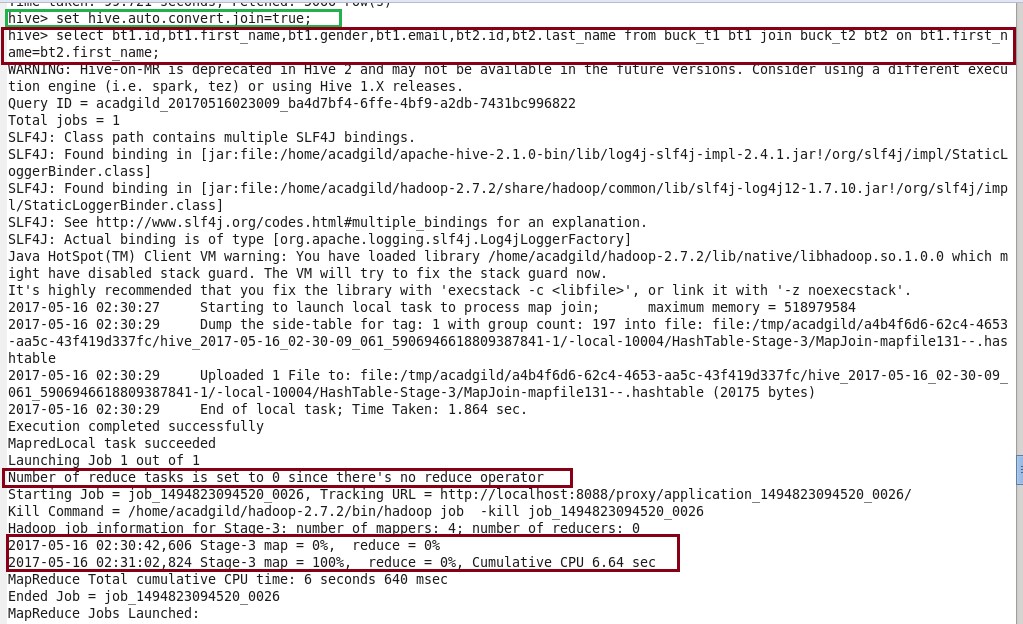
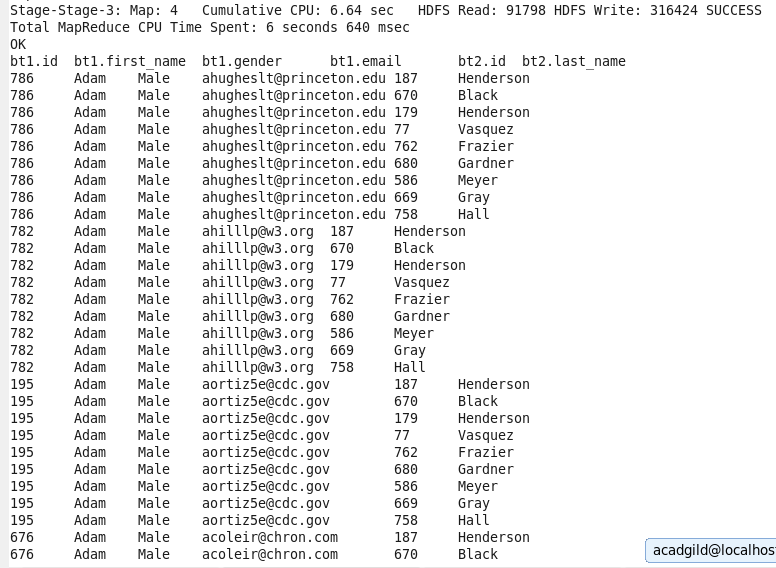


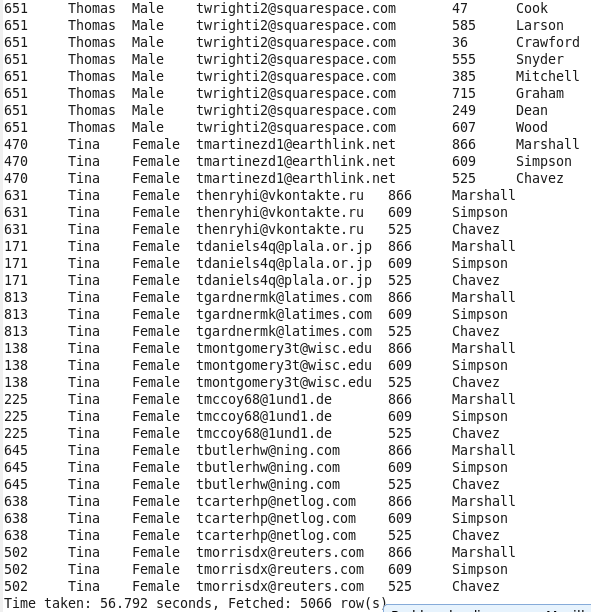


**SORT-MERGE BUCKET MAP JOIN:**

It is another Hive join optimization technique where all the tables need to be bucketed and sorted. In this case joins are very efficient because they require a simple merge of the presorted tables.

After setting configuration performing SMBM join on created bucketed tables attached with the output.

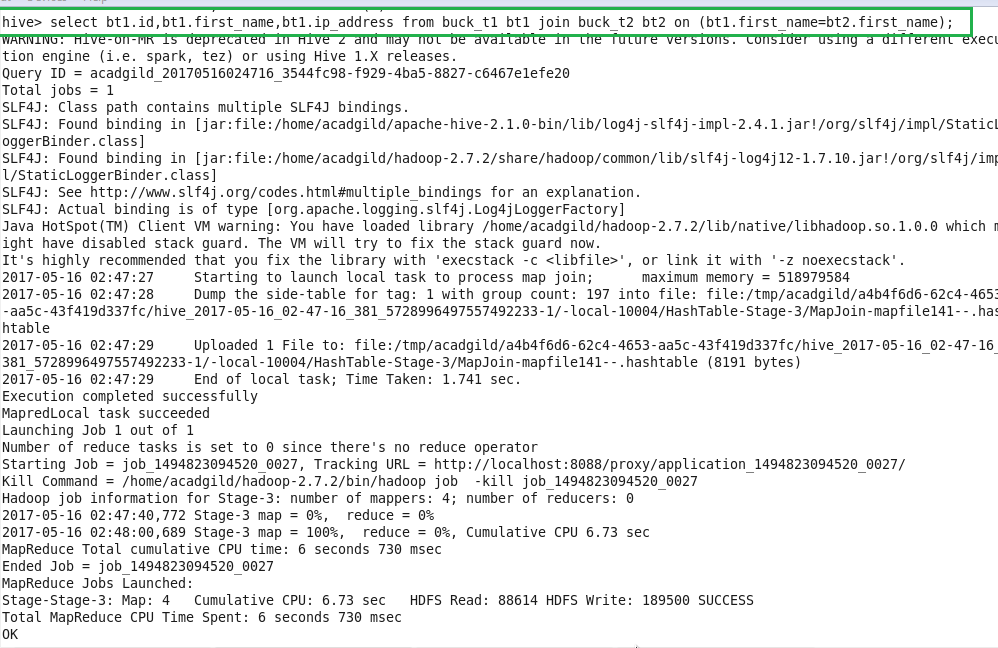
****

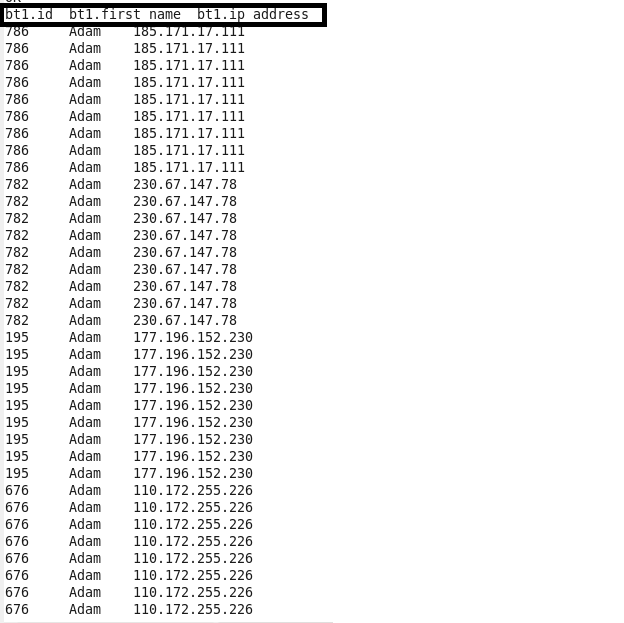


**LEFT SEMI JOIN:**

The left semi join is used in place of the IN/EXISTS sub-query in Hive. In a traditional RDBMS, the IN and EXISTS clauses are widely used whereas in Hive, the left semi join is used as a replacement of the same.

Left semi join eliminates the use of where and in keywords that we use in typical sub queries.

****

**OUTPUT :**

