Assignment 27.6

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>

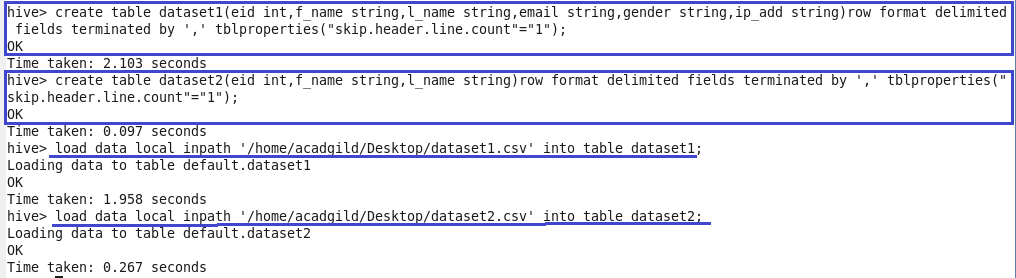
Dataset description of dataset1.csv is as follows: id,first\_name,last\_name,email,gender,ip\_address Dataset description of tdataset2.csv is as follows: id,first\_name,last\_name Note : You can just join the two datasets using the above specified join types in hive.

Bucket Map join

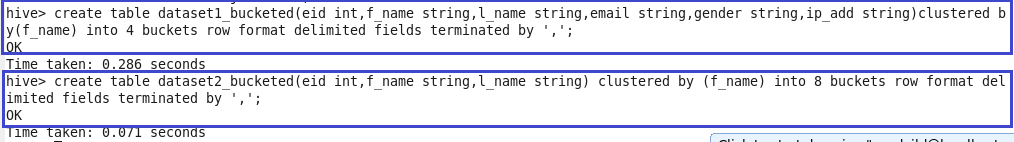
The constraint for performing Bucket-Map join is:

If tables being joined are bucketed on the join columns, and 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.

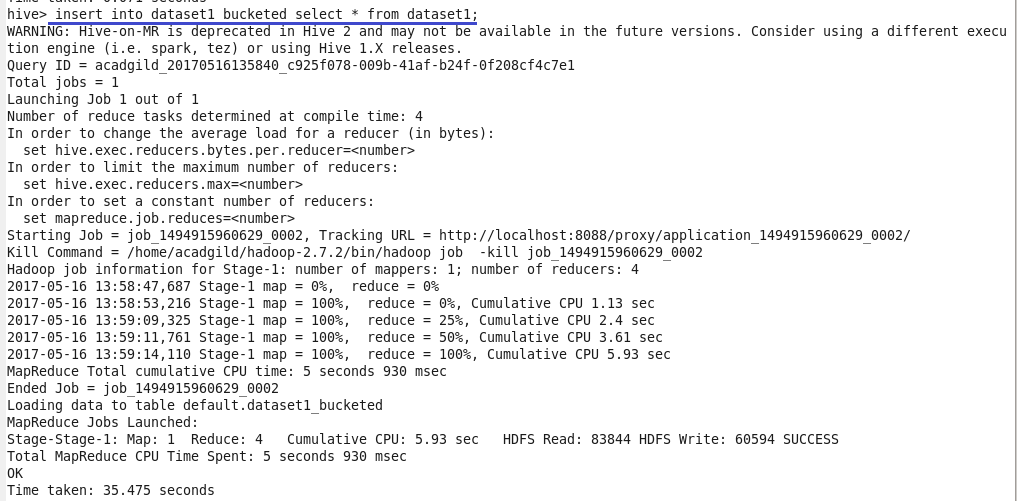
Creating tables and loading both datasets in them.

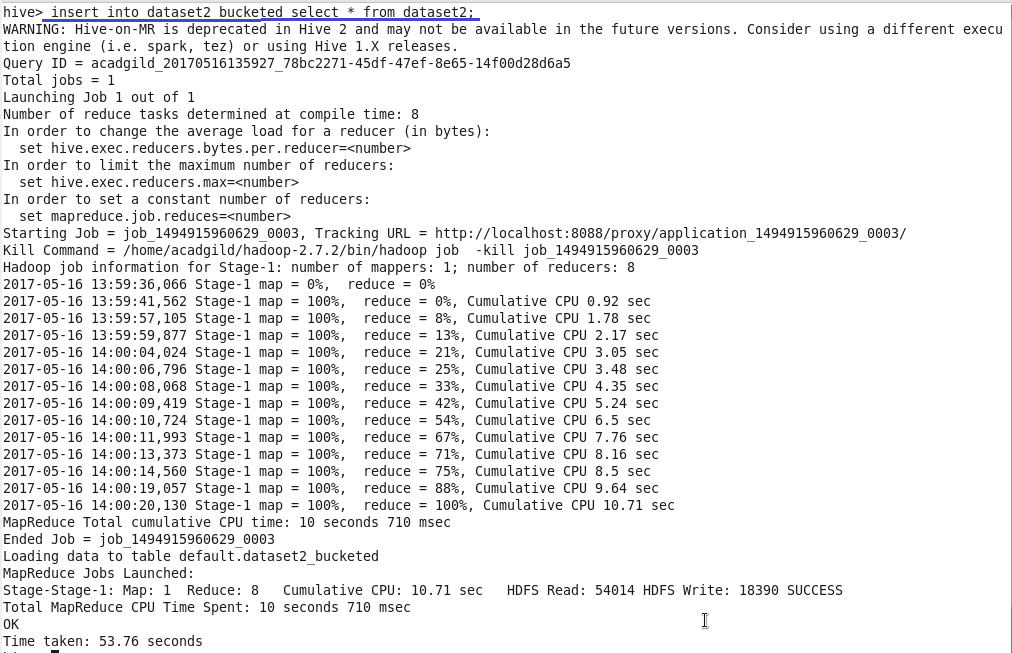


Bucketing table for dataset1 into 4 buckets and for dataset2 into 8 buckets.

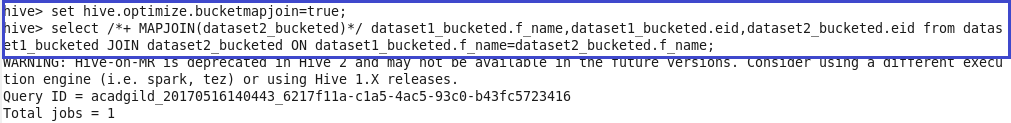


Loading data into bucketed tables from original tables.





Query for Bucket Map join



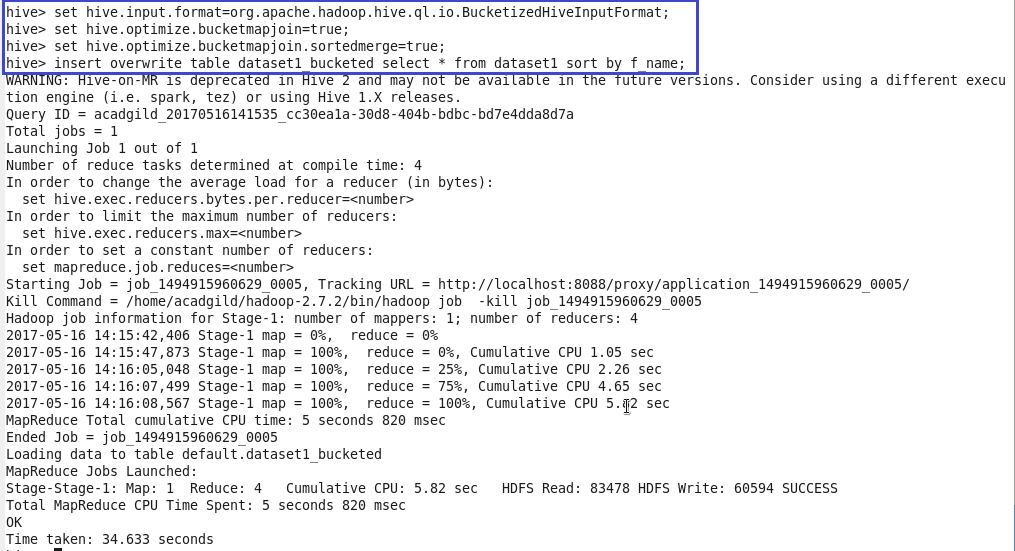
Output

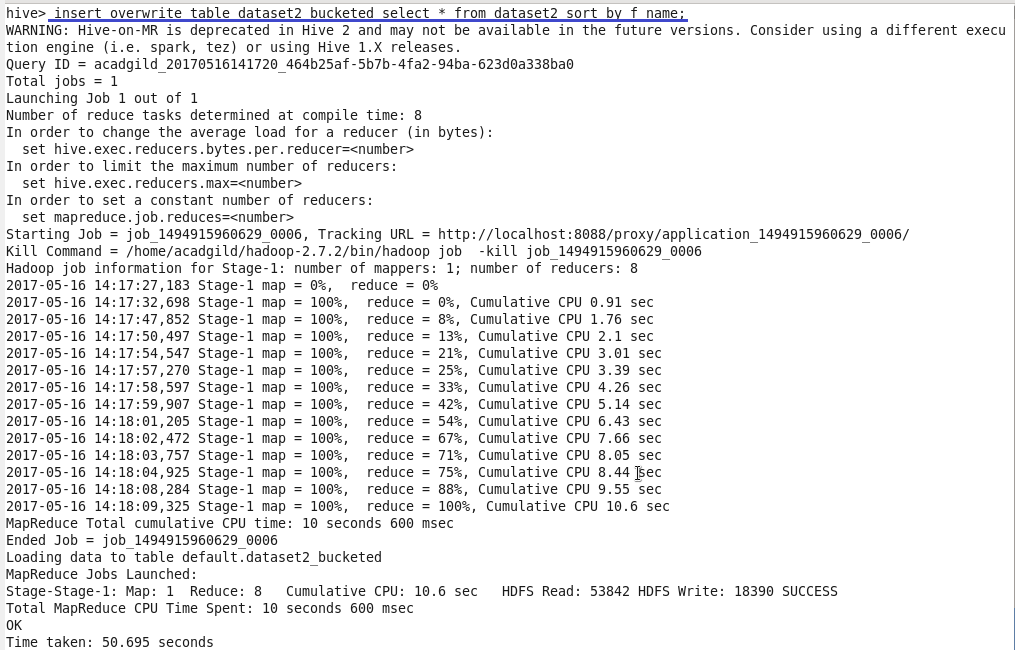


Sort Merge Bucket Map join

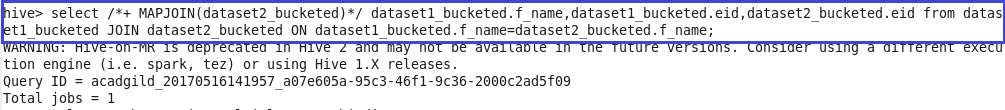
If the tables being joined are sorted and bucketized on the join columns and have the same number of buckets, a sort-merge join can be performed. The corresponding buckets are joined with each other at the mapper.

Here we have 4 buckets for dataset1 and 8 buckets for dataset2.



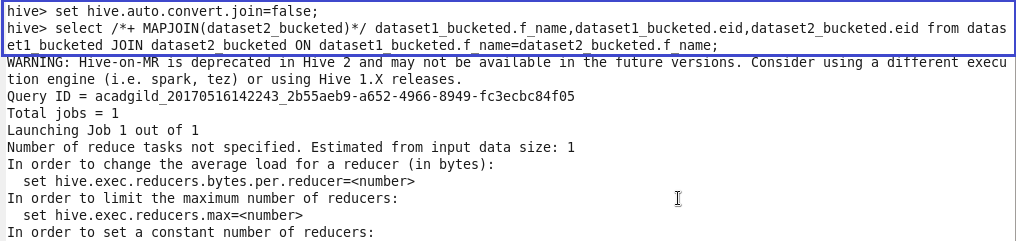


Query and output for SMBM join





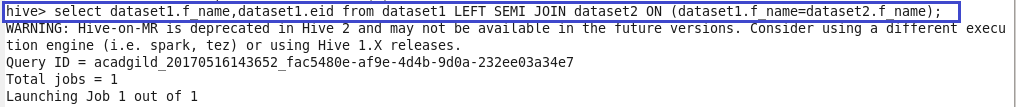
Sort merge bucket join





Left Semi join

The restrictions of using LEFT SEMI JOIN is that the right-hand-side table should only be referenced in the join condition (ON-clause), but not in WHERE- or SELECT-clauses etc.



Output

