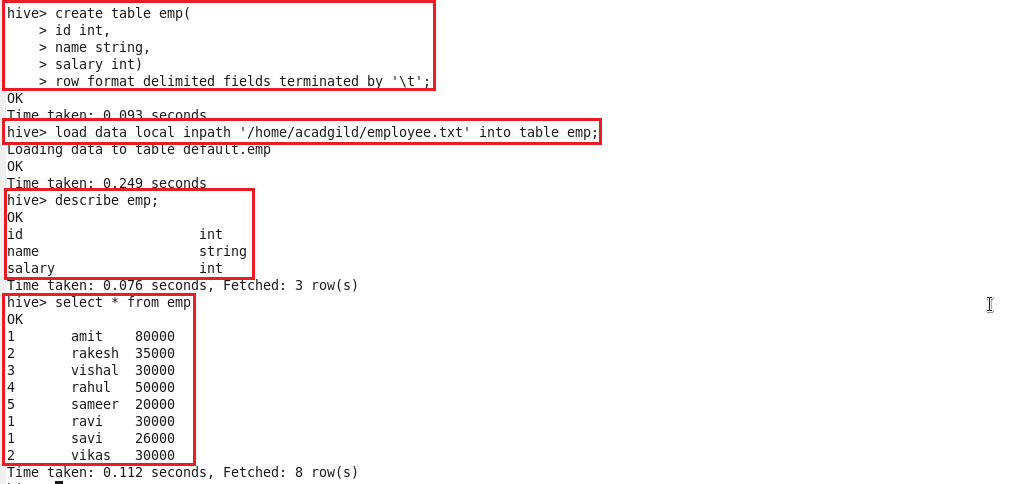
Assignment 27.5

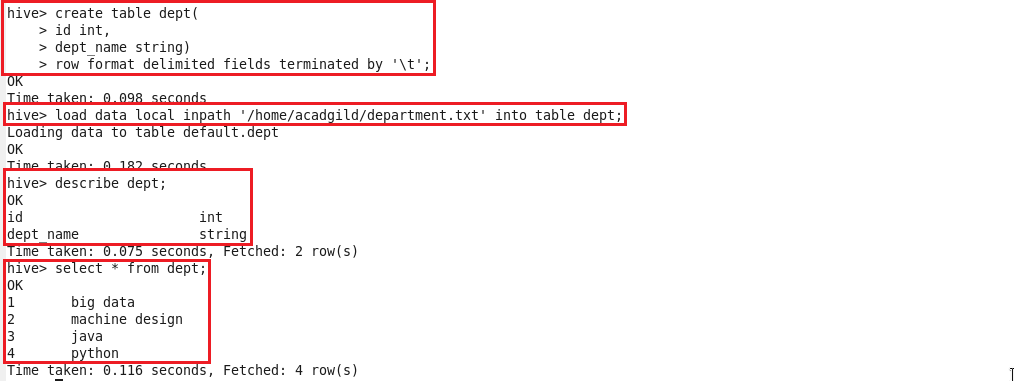
Problem Statement

Perform join optimizations in hive by following the steps in the below blog link

Employee table



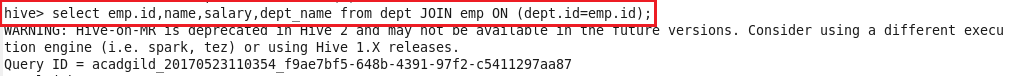
Department table

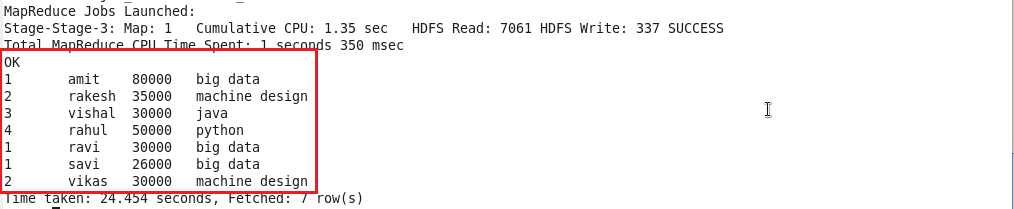


**Join table ordering (Largest table last)**.

As with any type of tuning, it is important to understand the internal working of a system. When Hive executes a join, it needs to select which table is streamed and which table is cached. Hive takes the last table in the JOIN statement for streaming, so we need to ensure that this streaming table is largest among the two.

The ‘emp’ table consists of department id, employee name, and employee salary. For any organization, this list can keep growing over the time. But, the ‘dept’ table will be static for most of the time.  Hence, when these two tables are joined it is important that the larger table comes last in the query. Let us see the optimized Hive query:





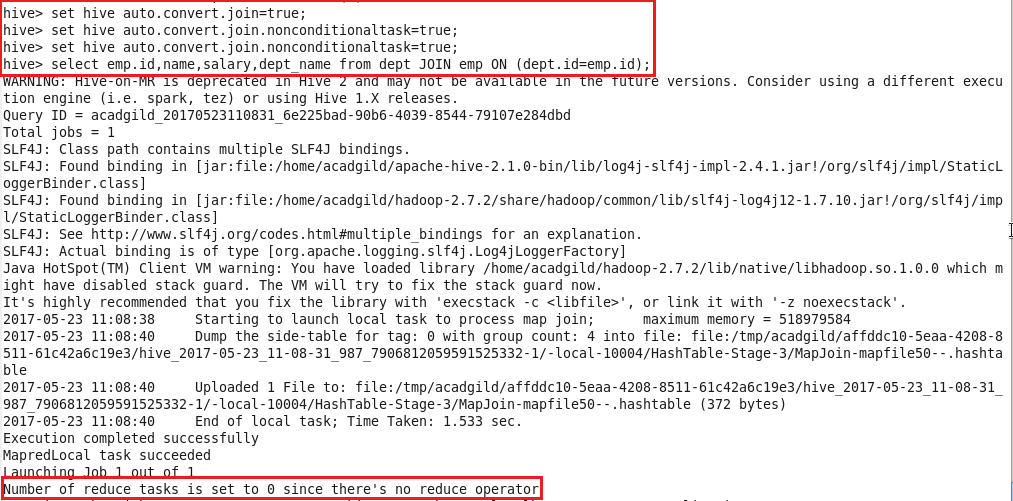
**Map Side Join:**

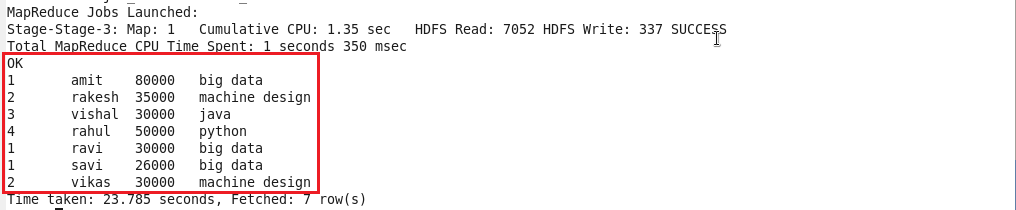
Also known as replicated join, a map-side join is a special type of join where a smaller table is loaded in memory and join is performed in map phase of MapReduce job. Since there is no reducer involved in the map-side join, it is much faster when compared to regular join.

To perform map-side join, set few configurations either into hive-site.xml OR directly from Hive shell. Below are the configurations which I have set from Hive shell.

hive> set hive.auto.convert.join=true;

hive> set hive.auto.convert.join.noconditionaltask=true;

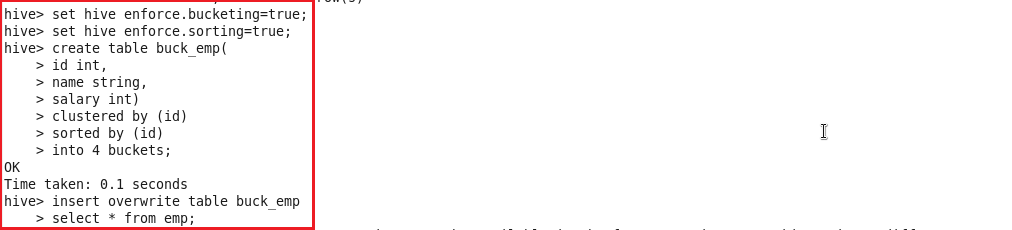




**Sort-Merge-Bucket (SMB) Map Join:**

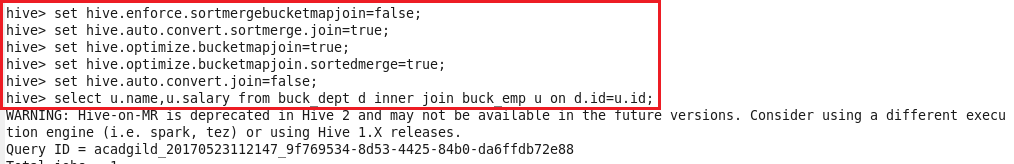
It is another Hive join optimization technique where all the tables need to be bucketed and sorted. In this case joins are very efficient because they require a simple merge of the presorted tables.

We have to create bucketed tables from our existing tables i.e.; emp and dept. Before creating bucketed table, we need to set below properties.





Now the stage is set to perform SMB Map Join to optimize Hive joining. Again, make some changes in properties to perform SMB Map join.



We will find that 4 mapper tasks are running (as we had 4 buckets). This helps in performing faster join operation when compared to regular joins.

