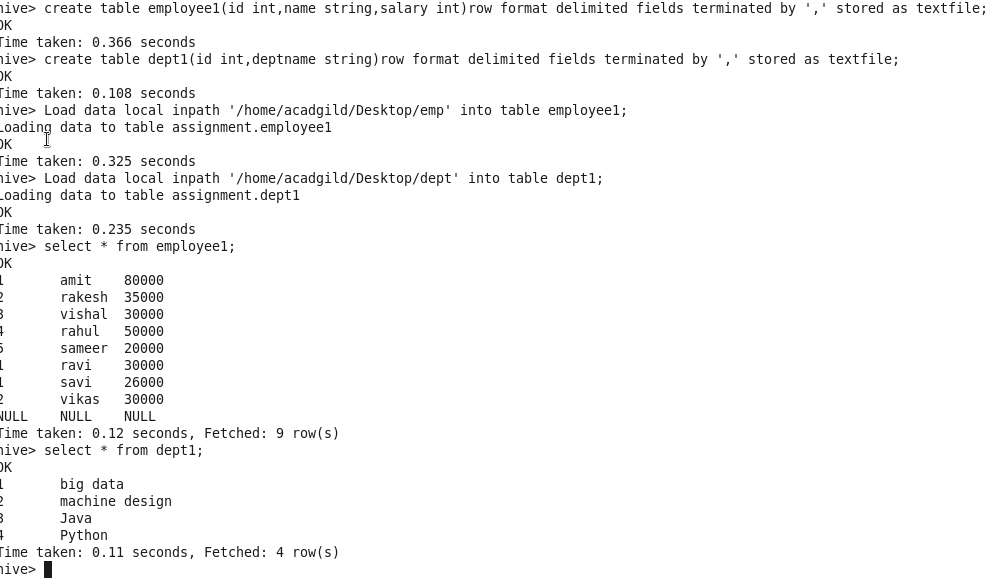
**ASSIGNMENT 27.5**

* **Perform join optimizations in hive**

Hive, like other SQL databases, allows users to join various tables. However, Joins can be computationally expensive, especially on big tables.

For join optimization in Hive, we can use repartition joins, replication joins and semi joins.

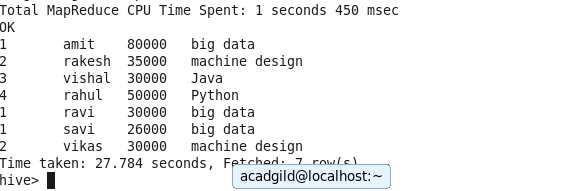
First we create tables and load the dataset



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.

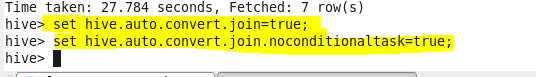
Hence, when these two tables are joined it is important that the larger table comes last in the query.





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.



Once you are done with the configuration, execute the same join operation as we performed above.

