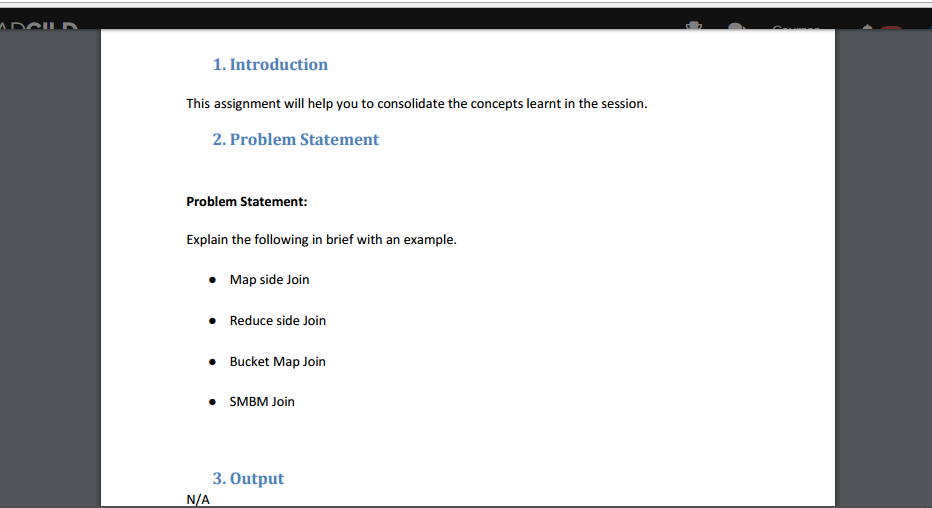
****

**Map-side join:**

* In order to perform Map-side joins, and there should be two files
* One - larger size
* the other - smaller size.

*The file size can be set small file by making use of the following property:*

* Hive.mapjoin.smalltable.filesize=(default it will be 25MB)
* Assume that there are two types of dataset
* one of larger in size
* other of smaller in size
* both having common id in column.
* Assume performing map side join by joining the ID on the datasets.
* Select /\*+MAPJOIN(dataset2)\*/ dataset1.first\_name,dataset1.eid,dataset2.eid from dataset1 join dataset2 on datset1.first\_name=datset2.first\_name;
* By using map-side join the reducer count will be set to 0 automatically.
* **Reduce-Side Joins:**  
    
  Reduce-Side joins are more simple when compared to Map-Side joins.
* Because the datasets inside need not to be structured.
* When compared efficiency will be less since both datasets will undergo MapReduce shuffle phase.
* Records having same key will brought together in reducer phase.
* One can use the Secondary Sort technique in order to control the order of the records.

**Bucket-Map join:**

* If in case the tables need to be joined will be bucketed up to join columns.
* The number of buckets present in a table will get multiple t other numbers of the buckets in other table.
* The buckets is capable to be joined with each other.
* Consider there are 2 tables which are bucketed.
* we can perform bucket-map join between these two datasets.
* For performing bucket-map join we need to set this property.

Set hive.optimize.bucketmapjoin=true;

***Sort merge bucket map join:***

* One must define the tables which need to be CLUSTERED BY same column as well SORTED BY in the same column and order INTO same amount of buckets.
* SMB join which is in Hive, each mapper will read a bucket from first table as well the bucket from the second table .
* Later a merge sort join is performed.
* In HIVE Sort Merge Bucket (SMB) join is mainly used since there is no file limit or partition or table join.
* SMB join will be best suited when the tables are large.
* In SMB join the columns are bucketed and will be sorted by making use of the join columns.
* All tables should maintain same number of buckets as of in SMB join.

*We have to set following options:*

Set hive.input.format=org.apache.hadoop.hive.ql.io.BucketizedHiveInputFormat;

Set hive.optimize.bucketmapjoin=true;

Set hive.optimize.bucketmapjoin.sortedmerge=true;