Join Optimizations:

1. Map Side Join / Broadcast Join
2. Bucket Join
3. Sort merge Join

Loading the Data into SQL Table:

Customers Data:

```
[hduser@localhost Downloads]$ cat customers.txt;
3002,Nick Rimando,New York,100,5001
3007,Brad Davis,New York,200,5001
3005,Graham Zusi,California,200,5002
3008,Julian Green,London,300,5002
3004,Fabian Johnson,Paris,300,5006
3009,Geoff Cameron,Berlin,100,5003
3003,Jozy Altidor,Moscow,200,5007
3001,Brad Guzan,London,100,5005[hduser@localhost Downloads]$
```

Orders Data:

```
[hduser@localhost Downloads]$ cat orders.txt
70001,150.5,2012-10-05,3005,5002
70009,270.65,2012-09-10,3001,5005
70002,65.26,2012-10-05,3002,5001
70004,110.5,2012-08-17,3009,5003
70007,948.5,2012-09-10,3005,5002
70005,2400.6,2012-07-27,3007,5001
70008,5760,2012-09-10,3002,5001
70010,1983.43,2012-10-10,3004,5006
70003,2480.4,2012-10-10,3009,5003
70012,250.45,2012-06-27,3008,5002
70011,75.29,2012-08-17,3003,5007
70013,3045.6,2012-04-25,3002,5001[hduser@localhost Downloads]$
```

Now loading the data into HDFS:

```
[hduser@localhost Downloads]$ hdfs dfs -ls /user/hduser/practice_data
22/07/11 16:51:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup        283 2022-07-11 16:51 /user/hduser/practice_data/customers.txt
-rw-r--r--   1 hduser supergroup        412 2022-07-11 16:51 /user/hduser/practice_data/orders.txt
[hduser@localhost Downloads]$
```

Creating orders table and loading data into it:

```
hive> create table orders(order_no int, purchase_amount double, order_date date, customer_id int, salesman_id int) row format delimited fields terminated by '
,' lines terminated by '\n';
OK
Time taken: 4.517 seconds
hive> load data inpath "/user/hduser/practice_data/orders.txt" into table orders;
Loading data to table practice.orders
Table practice.orders stats: [numFiles=1, totalSize=412]
OK
Time taken: 2.488 seconds
hive> select * from orders limit 10;
OK
70001   150.5   2012-10-05      3005    5002
70009   270.65  2012-09-10      3001    5005
70002   65.26   2012-10-05      3002    5001
70004   110.5   2012-08-17      3009    5003
70007   948.5   2012-09-10      3005    5002
70005   2400.6  2012-07-27      3007    5001
70008   5760.0  2012-09-10      3002    5001
70010   1983.43 2012-10-10      3004    5006
70003   2480.4  2012-10-10      3009    5003
70012   250.45  2012-06-27      3008    5002
Time taken: 1.102 seconds, Fetched: 10 row(s)
hive>
```

Creating customers table and loading data into it:

```
hive> create table customers(customer_id int,customer_name string,city string,grade int,salesman_id int) row format delimited fields terminated by ',' lines t
erminated by '\n';
OK
Time taken: 0.317 seconds


hive> load data inpath '/user/hduser/practice_data/customers.txt' into table customers;
Loading data to table practice.customers
Table practice.customers stats: [numFiles=1, totalSize=283]
OK
Time taken: 0.651 seconds
hive> select * from customers;
OK
3002    Nick Rimando    New York        100     5001
3007    Brad Davis      New York        200     5001
3005    Graham Zusi     California      200     5002
3008    Julian Green    London  300     5002
3004    Fabian Johnson  Paris   300     5006
3009    Geoff Cameron   Berlin  100     5003
3003    Jozy Altidor    Moscow  200     5007
3001    Brad Guzan      London  100     5005
Time taken: 0.156 seconds, Fetched: 8 row(s)
hive>
```

Normal Join on orders and customers:

```
hive> select * from orders o inner join customers c on o.customer_id = c.customer_id;
Query ID = hduser_20220711191129_696e18f5-791c-4e40-9545-7cddcaf3b84e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
```

Below, we can see that the number of reducers is 1

```
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2022-07-11 19:11:46,856 Stage-1 map = 0%,  reduce = 0%
2022-07-11 19:12:03,198 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.28 sec
2022-07-11 19:12:13,701 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.89 sec
MapReduce Total cumulative CPU time: 4 seconds 890 msec
Ended Job = job_1657537881794_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 4.89 sec   HDFS Read: 17601 HDFS Write: 825 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 890 msec
OK
70009   270.65  2012-09-10      3001    5005    3001    Brad Guzan      London  100     5005
70013   3045.6  2012-04-25      3002    5001    3002    Nick Rimando    New York        100     5001
70008   5760.0  2012-09-10      3002    5001    3002    Nick Rimando    New York        100     5001
70002   65.26   2012-10-05      3002    5001    3002    Nick Rimando    New York        100     5001
70011   75.29   2012-08-17      3003    5007    3003    Jozy Altidor    Moscow  200     5007
70010   1983.43 2012-10-10      3004    5006    3004    Fabian Johnson  Paris   300     5006
70007   948.5   2012-09-10      3005    5002    3005    Graham Zusi     California      200     5002
70001   150.5   2012-10-05      3005    5002    3005    Graham Zusi     California      200     5002
70005   2400.6  2012-07-27      3007    5001    3007    Brad Davis      New York        200     5001
70012   250.45  2012-06-27      3008    5002    3008    Julian Green    London  300     5002
70003   2480.4  2012-10-10      3009    5003    3009    Geoff Cameron   Berlin  100     5003
70004   110.5   2012-08-17      3009    5003    3009    Geoff Cameron   Berlin  100     5003
Time taken: 45.653 seconds, Fetched: 12 row(s)
hive>
```

Map Side Join - Way 1: Using Auto Property

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


hive> select * from customers c inner join orders o on o.customer_id = c.customer_id;
Query ID = hduser_20220711191417_3491c816-b76a-4031-bf3b-cabe27cae1c7
Total jobs = 1
```

Below, we can see that there is no reducer job involved

```
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-07-11 19:14:38,385 Stage-3 map = 0%,  reduce = 0%
2022-07-11 19:14:45,791 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.22 sec
MapReduce Total cumulative CPU time: 1 seconds 220 msec
Ended Job = job_1657537881794_0009
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.22 sec   HDFS Read: 8479 HDFS Write: 825 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 220 msec
OK
3005    Graham Zusi     California      200     5002    70001   150.5   2012-10-05      3005    5002
3001    Brad Guzan      London  100     5005    70009   270.65  2012-09-10      3001    5005
3002    Nick Rimando    New York        100     5001    70002   65.26   2012-10-05      3002    5001
3009    Geoff Cameron   Berlin  100     5003    70004   110.5   2012-08-17      3009    5003
3005    Graham Zusi     California      200     5002    70007   948.5   2012-09-10      3005    5002
3007    Brad Davis      New York        200     5001    70005   2400.6  2012-07-27      3007    5001
3002    Nick Rimando    New York        100     5001    70008   5760.0  2012-09-10      3002    5001
3004    Fabian Johnson  Paris   300     5006    70010   1983.43 2012-10-10      3004    5006
3009    Geoff Cameron   Berlin  100     5003    70003   2480.4  2012-10-10      3009    5003
3008    Julian Green    London  300     5002    70012   250.45  2012-06-27      3008    5002
3003    Jozy Altidor    Moscow  200     5007    70011   75.29   2012-08-17      3003    5007
3002    Nick Rimando    New York        100     5001    70013   3045.6  2012-04-25      3002    5001
Time taken: 31.086 seconds, Fetched: 12 row(s)
hive>
```

Map Side Join – Way 2: Using hints

```
hive> set hive.ignore.mapjoin.hint = false;
hive> set hive.ignore.mapjoin.hint;
hive.ignore.mapjoin.hint=false
hive>
```

```
hive> select * from customers c inner join orders o on o.customer_id = c.customer_id;
Query ID = hduser_20220711191659_3794497e-25fb-4d88-8daf-4447563448da
Total jobs = 1

Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-07-11 19:17:21,112 Stage-3 map = 0%,  reduce = 0%
2022-07-11 19:17:27,345 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.22 sec
MapReduce Total cumulative CPU time: 1 seconds 220 msec
Ended Job = job_1657537881794_0010
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.22 sec   HDFS Read: 8479 HDFS Write: 825 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 220 msec
OK
3005    Graham Zusi     California      200     5002    70001   150.5   2012-10-05      3005    5002
3001    Brad Guzan      London  100     5005    70009   270.65  2012-09-10      3001    5005
3002    Nick Rimando    New York        100     5001    70002   65.26   2012-10-05      3002    5001
3009    Geoff Cameron   Berlin  100     5003    70004   110.5   2012-08-17      3009    5003
3005    Graham Zusi     California      200     5002    70007   948.5   2012-09-10      3005    5002
3007    Brad Davis      New York        200     5001    70005   2400.6  2012-07-27      3007    5001
3002    Nick Rimando    New York        100     5001    70008   5760.0  2012-09-10      3002    5001
3004    Fabian Johnson  Paris   300     5006    70010   1983.43 2012-10-10      3004    5006
3009    Geoff Cameron   Berlin  100     5003    70003   2480.4  2012-10-10      3009    5003
3008    Julian Green    London  300     5002    70012   250.45  2012-06-27      3008    5002
3003    Jozy Altidor    Moscow  200     5007    70011   75.29   2012-08-17      3003    5007
3002    Nick Rimando    New York        100     5001    70013   3045.6  2012-04-25      3002    5001
Time taken: 30.457 seconds, Fetched: 12 row(s)
hive>
```

Bucketed Join:

Here we use same data of orders and customers.

Creating Bucketed Orders table with 2 buckets and loading data from previous orders table:

```
hive> create table orders_bucketed(order_no int, purchase_amount double, order_date date, customer_id int, salesman_id int) clustered by (customer_id) into 2
buckets row format delimited fields terminated by ',' lines terminated by '\n';
OK
Time taken: 0.361 seconds
hive> insert into orders_bucketed select * from orders;
Query ID = hduser_20220711192238_bd479dfc-d14d-4eba-a6a5-aea5ff2d6b31
Total jobs = 1
```

```
hive> select * from orders_bucketed;
OK
70013   3045.6  2012-04-25      3002    5001
70012   250.45  2012-06-27      3008    5002
70010   1983.43 2012-10-10      3004    5006
70008   5760.0  2012-09-10      3002    5001
70002   65.26   2012-10-05      3002    5001
70011   75.29   2012-08-17      3003    5007
70003   2480.4  2012-10-10      3009    5003
70005   2400.6  2012-07-27      3007    5001
70007   948.5   2012-09-10      3005    5002
70004   110.5   2012-08-17      3009    5003
70009   270.65  2012-09-10      3001    5005
70001   150.5   2012-10-05      3005    5002
Time taken: 0.173 seconds, Fetched: 12 row(s)
```

```
[hduser@localhost Downloads]$ hdfs dfs -ls /user/hive/warehouse/practice.db/orders_bucketed
22/07/11 19:24:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup        170 2022-07-11 19:23 /user/hive/warehouse/practice.db/orders_bucketed/000000_0
-rw-r--r--   1 hduser supergroup        234 2022-07-11 19:23 /user/hive/warehouse/practice.db/orders_bucketed/000001_0
[hduser@localhost Downloads]$
```

Creating Bucketed Customers table with 4 buckets and loading data from previous customers' table:

```
hive> create table customers_bucketed(customer_id int,customer_name string,city string,grade int,salesman_id int) clustered by (customer_id) into 4 buckets ro
w format delimited fields terminated by ',' lines terminated by '\n';
OK
Time taken: 0.264 seconds
hive> insert into customers_bucketed select * from customers;
Query ID = hduser_20220711192626_e9ac8b56-21f0-438c-bd6c-d1bdebf31a02
Total jobs = 1
```

```
hive> select * from customers_bucketed;
OK
3004    Fabian Johnson  Paris   300     5006
3008    Julian Green    London  300     5002
3001    Brad Guzan      London  100     5005
3009    Geoff Cameron   Berlin  100     5003
3005    Graham Zusi     California      200     5002
3002    Nick Rimando    New York        100     5001
3003    Jozy Altidor    Moscow  200     5007
3007    Brad Davis      New York        200     5001
Time taken: 0.384 seconds, Fetched: 8 row(s)
hive>
```

```
[hduser@localhost Downloads]$ hdfs dfs -ls /user/hive/warehouse/practice.db/customers_bucketed
22/07/11 19:28:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   1 hduser supergroup         69 2022-07-11 19:27 /user/hive/warehouse/practice.db/customers_bucketed/000000_0
-rw-r--r--   1 hduser supergroup        104 2022-07-11 19:27 /user/hive/warehouse/practice.db/customers_bucketed/000001_0
-rw-r--r--   1 hduser supergroup         36 2022-07-11 19:27 /user/hive/warehouse/practice.db/customers_bucketed/000002_0
-rw-r--r--   1 hduser supergroup         68 2022-07-11 19:27 /user/hive/warehouse/practice.db/customers_bucketed/000003_0
[hduser@localhost Downloads]$
```

Now applying the Bucketed Join by using some properties:

```
hive> set hive.auto.convert.join = true;
hive> set hive.optimize.bucketmapjoin = true
    > ;
hive> set hive.auto.convert.join = true;
hive> set hive.optimize.bucketmapjoin = true;
hive> set hive.optimize.bucketmapjoin
    > ;
hive.optimize.bucketmapjoin=true
hive> set hive.auto.convert.join;
hive.auto.convert.join=true
hive>
```

```
hive> select * from customers_bucketed c inner join orders_bucketed o on c.customer_id = o.customer_id;
Query ID = hduser_20220711193232_218159c1-da94-41e4-80b7-b024967b9af7
Total jobs = 1
```

Below we can see that number of reducers are zero

```
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-07-11 19:32:56,519 Stage-3 map = 0%,  reduce = 0%
2022-07-11 19:33:05,304 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.46 sec
MapReduce Total cumulative CPU time: 1 seconds 460 msec
Ended Job = job_1657537881794_0013
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.46 sec   HDFS Read: 8613 HDFS Write: 825 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 460 msec
OK
3002    Nick Rimando    New York        100     5001    70013   3045.6  2012-04-25      3002    5001
3008    Julian Green    London  300     5002    70012   250.45  2012-06-27      3008    5002
3004    Fabian Johnson  Paris   300     5006    70010   1983.43 2012-10-10      3004    5006
3002    Nick Rimando    New York        100     5001    70008   5760.0  2012-09-10      3002    5001
3002    Nick Rimando    New York        100     5001    70002   65.26   2012-10-05      3002    5001
3003    Jozy Altidor    Moscow  200     5007    70011   75.29   2012-08-17      3003    5007
3009    Geoff Cameron   Berlin  100     5003    70003   2480.4  2012-10-10      3009    5003
3007    Brad Davis      New York        200     5001    70005   2400.6  2012-07-27      3007    5001
3005    Graham Zusi     California      200     5002    70007   948.5   2012-09-10      3005    5002
3009    Geoff Cameron   Berlin  100     5003    70004   110.5   2012-08-17      3009    5003
3001    Brad Guzan      London  100     5005    70009   270.65  2012-09-10      3001    5005
3005    Graham Zusi     California      200     5002    70001   150.5   2012-10-05      3005    5002
Time taken: 34.205 seconds, Fetched: 12 row(s)
hive>
```

Sort Merge Join:

Here we create orders_sort_merge table and customers_sort_merge table and load the data from the previous tables.

Creating customers_sort_merge table with 2 buckets with sort condition on Customer Id and loading the data:

```
hive> create table customers_sort_merge(customer_id int,customer_name string,city string,grade int,salesman_id int) clustered by (customer_id) sorted by (cust
omer_id) into 2 buckets row format delimited fields terminated by ',' lines terminated by '\n';
OK
Time taken: 0.426 seconds
hive> insert into customers_sort_merge select * from customers;
Query ID = hduser_20220711193749_8d25a937-5cf6-40cb-acaf-e2c4b2f36ab8
Total jobs = 1
Launching Job 1 out of 1
```

```
[hduser@localhost Downloads]$ hdfs dfs -ls /user/hive/warehouse/practice.db/customers_sort_merge
22/07/11 19:40:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup        105 2022-07-11 19:38 /user/hive/warehouse/practice.db/customers_sort_merge/000000_0
-rw-r--r--   1 hduser supergroup        172 2022-07-11 19:38 /user/hive/warehouse/practice.db/customers_sort_merge/000001_0
```

Creating orders_sort_merge table with 2 buckets with sort condition on Customer Id and loading the data:

```
hive> create table orders_sort_merge(order_no int, purchase_amount double, order_date date, customer_id int, salesman_id int) clustered by (customer_id) sorte
d by (customer_id) into 2 buckets row format delimited fields terminated by ',' lines terminated by '\n';
OK
Time taken: 0.277 seconds
hive> insert into orders_sort_merge select * from orders;
Query ID = hduser_20220711194544_bd2d5c5f-5417-4f60-adf1-e216c8a2270e
Total jobs = 1
```

```
[hduser@localhost Downloads]$ hdfs dfs -ls /user/hive/warehouse/practice.db/orders_sort_merge
22/07/11 19:48:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup        170 2022-07-11 19:46 /user/hive/warehouse/practice.db/orders_sort_merge/000000_0
-rw-r--r--   1 hduser supergroup        234 2022-07-11 19:46 /user/hive/warehouse/practice.db/orders_sort_merge/000001_0
[hduser@localhost Downloads]$
```

Join Operation: Sort Merge Join

```
set hive.auto.convert.sortmerge.join=true;
set hive.auto.convert.sortmerge.join.noconditionaltask=true;
set hive.optimize.bucketmapjoin=true;
set hive.optimize.bucketmapjoin.sortedmerge=true;
set hive.enforce.bucketing=true;
set hive.enforce.sorting=true;
set hive.auto.convert.join=true;
```

We can see that the number of Reducers here are zero.

```
hive> select * from orders_sort_merge o inner join customers_sort_merge c on c.customer_id = o.customer_id;
Query ID = hduser_20220711195255_da32d5fb-fdb6-4074-af3c-758596b43be4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1657537881794_0016, Tracking URL = http://Inceptez:8088/proxy/application_1657537881794_0016/
Kill Command = /usr/local/hadoop/bin/hadoop job  -kill job_1657537881794_0016

Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 0
2022-07-11 19:53:09,363 Stage-1 map = 0%,  reduce = 0%
2022-07-11 19:53:23,136 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.96 sec
MapReduce Total cumulative CPU time: 2 seconds 960 msec
Ended Job = job_1657537881794_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2   Cumulative CPU: 2.96 sec   HDFS Read: 17223 HDFS Write: 272 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 960 msec
OK
70012   250.45  2012-06-27      3008    5002    3008    Julian Green    London  300     5002
70011   75.29   2012-08-17      3003    5007    3003    Jozy Altidor    Moscow  200     5007
70003   2480.4  2012-10-10      3009    5003    3009    Geoff Cameron   Berlin  100     5003
70005   2400.6  2012-07-27      3007    5001    3007    Brad Davis      New York        200     5001
```