IMAAD IMRAN HAJWANE
202101132
TOPIC: JOINNG IN HIVE

Perform various Join Operations in Hive

- Creating Tables:

```
hive> CREATE TABLE customers (
    >     customer_id INT,
    >     customer_name STRING,
    >     city STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.111 seconds
hive> CREATE TABLE orders (
    >     order_id INT,
    >     customer_id INT,
    >     order_date STRING,
    >     total_amount DOUBLE
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.082 seconds
```

- Inserting Values in tables:

```
hive> INSERT INTO TABLE customers VALUES
    > (1, 'Alice', 'New York'),
    > (2, 'Bob', 'Los Angeles'),
    > (3, 'Charlie', 'Chicago'),
    > (4, 'David', 'Houston');
Query ID = vishwa_20240926153636_0462cf01-6676-40af-a883-0f2ee18d8259
Total jobs = 3
```

```
hive> INSERT INTO TABLE orders VALUES
    > (101, 1, '2024-09-01', 200.50),
    > (102, 2, '2024-09-02', 150.75),
    > (103, 1, '2024-09-05', 300.00),
    > (104, 3, '2024-09-03', 500.00),
    > (105, 4, '2024-09-07', 100.00);
Query ID = vishwa_20240926153725_5f176d90-04fb-4d7c-ac70-75193ae457e9
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
```

1. Inner Join

- **Definition**: An inner join returns only the rows that have matching values in both tables. If a row in either table does not have a corresponding match, it will not appear in the result set.
- **Use Case**: Useful when you want to combine related data from two or more tables and only include the records that share a common value in the specified columns.

- Inner Join:

```
hive> SELECT c.customer_id, c.customer_name, o.order_id, o.total_amount
    > FROM customers c
    > INNER JOIN orders o ON c.customer_id = o.customer_id;
Query ID = vishwa_20240926153816_2e564819-1435-476d-b25c-ac2a9c7d31fd
Total jobs = 1
```

```
Total MapReduce CPU Time Spent: 2 seconds 30 msec
OK
1       Alice   101     200.5
2       Bob     102     150.75
1       Alice   103     300.0
3       Charlie 104     500.0
4       David   105     100.0
Time taken: 22.691 seconds, Fetched: 5 row(s)
```

2. Left Join (Left Outer Join)

- **Definition**: A left join returns all the rows from the left table and the matched rows from the right table. If there is no match, NULL values will be returned for columns from the right table.
- **Use Case**: Useful when you want to keep all records from the left table regardless of whether there is a match in the right table.

- Left Join:

```
hive> SELECT c.customer_id, c.customer_name, o.order_id, o.total_amount
    > FROM customers c
    > LEFT JOIN orders o ON c.customer_id = o.customer_id where c.customer_id=1;
Warning: Map Join MAPJOIN[13][bigTable=?] in task 'Stage-3:MAPRED' is a cross product
Query ID = vishwa_20240926154259_8922039e-38e9-425b-9f7f-70875723eb05
Total jobs = 1
```

```
Total MapReduce CPU Time Spent: 2 seconds 150 msec
OK
1       Alice   101     200.5
1       Alice   103     300.0
Time taken: 21.677 seconds, Fetched: 2 row(s)
```

3. Right Join (Right Outer Join)

- **Definition**: A right join returns all the rows from the right table and the matched rows from the left table. If there is no match, NULL values will be returned for columns from the left table.
- **Use Case**: Useful when you want to keep all records from the right table regardless of whether there is a match in the left table.

  - Right Join:

```
hive> SELECT c.customer_id, c.customer_name, o.order_id, o.total_amount
    > FROM customers c
    > RIGHT JOIN orders o ON c.customer_id = o.customer_id where total_amount>=200;
Query ID = vishwa_20240926154620_be773de1-678e-4009-bce9-53cb2b422238
Total jobs = 1
```

```
Total MapReduce CPU Time Spent: 2 seconds 290 msec
OK
1       Alice   101     200.5
1       Alice   103     300.0
3       Charlie 104     500.0
Time taken: 21.986 seconds, Fetched: 3 row(s)
```

4. Full Join (Full Outer Join)

- **Definition**: A full join returns all the rows from both tables. If there is a match between the tables, the corresponding rows will be merged. If there is no match, NULL values will be included for columns where there is no corresponding record.
- **Use Case**: Useful when you want to retrieve all records from both tables, regardless of whether they have a match

  - Full outer Join:

```
hive> SELECT c.customer_id, c.customer_name, o.order_id, o.total_amount
    > FROM customers c
    > FULL OUTER JOIN orders o ON c.customer_id = o.customer_id  where c.customer_name="Alice";
Query ID = vishwa_20240926154942_d5f78b74-8b16-4e5d-ac75-34fc89f75c44
Total jobs = 1
```

```
Total MapReduce CPU Time Spent: 2 seconds 240 msec
OK
1       Alice   101     200.5
1       Alice   103     300.0
Time taken: 20.937 seconds, Fetched: 2 row(s)
```

Summary of Differences

- **Inner Join**: Only includes matched records.
- **Left Join**: Includes all records from the left table and matched records from the right.
- **Right Join**: Includes all records from the right table and matched records from the left.
- **Full Join**: Includes all records from both tables, matched and unmatched.