Apache Hive

External vs Internal tables

Internal Tables: Tables whose data is managed and stored by the database system itself. The database system is responsible for managing the storage, metadata, and access to internal tables. Data in internal tables is typically stored in the native storage format of the database system, optimized for efficient storage and retrieval. Internal tables are tightly integrated with the database system and are subject to its performance optimizations and security mechanisms.

External Tables: External tables are tables that reference data stored outside the database system. Instead of managing the data directly, the database system provides a logical representation of the external data source. External tables are useful for querying data that resides in files or in other database systems without physically importing the data into the database.

Examples of external tables include tables that reference files stored in Hadoop Distributed File System (HDFS), files in Amazon S3, or tables in another database system accessed via a foreign data wrapper.

Task 1: Create an external table (user_ratings) in Hive where the data is stored in HDFS.

First create some txt data in HDFS.

1,12,5
2,15,3
3,12,3
4,15,4
5,14,4

```
Create external table user_ratings
(user_id int,
Film_id int,
Rating int)
Row format delimited
Fields terminated by ….
Location ‘path’;
```

Task 2: Create an Internal table (user_info) in Hive and pull the data from HDFS.

First create some txt data in HDFS.

1,25,'Teacher'
3,22,'Student'
5,25,'maker'
7,30,'Student'
9,35,'Teacher'

Create table user_info
(user_id int,
Age int,
designation string)
Row format delimited
Fields terminated by ….;

Joins

In HiveQL, the SQL-like query language used in Apache Hive, joins are used to combine rows from two or more tables based on a related column between them. Hive supports several types of joins, including inner joins, outer joins (left, right, and full outer joins), and cross joins.

An inner join returns rows when there is at least one match in both tables based on the join condition.

A left outer join returns all rows from the left table (the first table mentioned in the query) and matching rows from the right table. If there is no match, NULL values are returned for the columns of the right table.

A right outer join returns all rows from the right table and matching rows from the left table. If there is no match, NULL values are returned for the columns of the left table.

Task 3: Write inner, left and right join between the above internal and external tables and check the results.