What is the Hive Metastore?

The Hive Metastore is the central repository of Apache Hive metadata. It stores metadata for Hive tables (schema and location) and partitions in a relational database at Visa this is typically MySQL. It provides client access to this information by using metastore service API.

What is the difference between Managed and External Tables in Hive?

Managed tables are Hive owned tables where the entire lifecycle of the tables' data is managed and controlled by Hive. When you drop a Hive Managed table it removes the schema from the metastore and all data stored associated with the table.

External tables are tables where Hive is referencing the data typically stored in Hadoop HDFS or flat files (csv). When you drop the external table referenced data remains, only the metastore schema is removed.

Are indexes used with Hive?

Indexes are a pointer or reference to a record in a table as in relational databases. Indexing is a relatively new feature in Hive. In Hive, the index table is different than the main table. Indexes facilitate in making query execution or search operation faster.

Which is preferred at Visa Hive indexing or partitioning?

Indexes are used to speed the search of data within tables. Partitions provide segregation of the data at the HDFS level, creating sub-directories for each partition. Partitioning allows the number of files read and amount of data searched in a query to be limited, which is generally faster and preferred use case at Visa.

What does a hive explain plan do?

Hive provides an EXPLAIN command to return a query execution plan without running the query. We can use an EXPLAIN command for queries if we have a doubt or a concern about performance. The EXPLAIN command will help to see the difference between two or more queries for the same purpose. Used in this manner: Explain select count(\*) from transactions