Page No.
Date

BDA - Experiment 5

Aim: Create HIVE database and descriptive based statistics, visualization using HIVE/PIG

Theory:

- Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarise big data.
- Provides a high level interface for querying and managing huge datasets stored in HDFS.
- Features:
  - Stores schema in a database and processes data into HDFS.
  - Designed for OLAP
  - Provides SQL type language for querying, called HiveQL or HQL
- Unlike Traditional Relational Databases, Hive follows a "schema-on-rea approach. Data is stored in HDFS without a predefined schema, & the schema is applied at time of query execution.
- Metastore is used to store metadata about the tables, columns, partitions and respective locations in HDFS.

Conclusion: Hive databases have been created and analysis along with visualisation were peuformed on this database.