IE494 BIG DATA PROCESSING

Stage – 1 Report

Group Members :

- Jayswal Krishil – 202203040

Tenzing: A SQL Implementation on the MapReduce Framework

2. Objective :

- **Tenzing's core architecture :**
  Exploring the Tenzing's core architecture by focusing on it's main components.
- **Basic SQL features :**
  Exploring basic SQL features like projection, filtering , aggregation and join and exploring different startegies for achieveing this.
- **Statistical Analysis of performance :**
  We will analyze the performance by highlighting how Tenzing's SQL features are designed for parallel execution, making them scalable and efficient in large-scale data environments.
- **Advanced SQL Features :**
  We will explore some advance SQL features for Analytics purposes and OLAP extensions.
- **Benchmarking and Comparison :**
  Finally doing some benchmark results and side by side comparison of Tenzing with other Implementations like Hive, etc.

3. Plan :

- **Introduction and Overview :**
  Giving a basic overview of what Tenzing does and why there is a need of such system.
- **History and Motivation :**
  Getting an overview of the challenges faced in past and motivation behind building Tenzing.
- **Implementation Architecture Overview :**
  Describing Tenzing's core architecture, focusing on its four major components: worker pool, query server, client interfaces, and metadata server and lifetime of a query.

- **SQL Implementation on MapReduce :**
  Dive into Tenzing's SQL implementation over MapReduce, covering key features like projection, joins, and OLAP operations. Focus on how Tenzing efficiently handles distributed structured data while optimizing query execution for large-scale analysis.
- **Performance analysis and Benchmarking :**
  Evaluate Tenzing's performance through benchmarking against traditional systems, focusing on query latency, scalability, and efficiency.
- **Comparison with other such systems :**
  Highlight how MapReduce-based optimizations enhance distributed query execution, and compare results with other SQL-on-MapReduce implementations such as Hive or HadoopDB.
- **Conclusion and References :**
  Highlight key takeaways from the performance analysis and its implications for future developments in distributed data management.
- **Further Enhancement :**
  Explore the possibility of expanding its analytical capabilities with machine learning and AI features, as well as improving user interfaces to further simplify interactions for non-technical users.

4. Work done so far :

- In preparation for the term paper, I have thoroughly read the research paper on Tenzing and explored various online articles to gain a comprehensive understanding of Tenzing's SQL implementation on the MapReduce framework.

- Additionally, I conducted further research through various online sources to gather insights into performance metrics, comparison with other SQL-on-MapReduce systems, and the challenges associated with implementing SQL in a distributed environment.

5. References :

- **Chattopadhyay, Biswapesh, et al. "Tenzing a sql implementation on the MapReduce framework." Proceedings of the VLDB Endowment 4.12 (2011): 1318-1327.**
  https://www.vldb.org/pvldb/vol4/p1318-chattopadhyay.pdf

- **https://stephenholiday.com/notes/tenzing/**