BAHRIA UNIVERSITY (KARACHI CAMPUS)

Assignment-01

(Big Data Analytics)

Class: **BSE [4]-7 (B)** **(Morning)**

Course Instructor: **Dr. Salahuddin Shaikh** Submission Date: **31/10/2024**

Date: **(22/10/2024)** Max Marks**:** **5 M (CLO1-2)**

Student's Name: ____________________ Reg. No: _________

Assignment Title:

"Advanced Use Cases and Performance Analysis of Hadoop and GFS"

Tasks:

1. **Case Study Analysis:**

In today's digital world, companies generate and process massive amounts of data daily. Two leading tech giants, **Facebook** and **Google**, have built their data infrastructures on robust distributed file systems to handle this data efficiently.

- **Facebook** uses **Hadoop** to manage its vast amounts of user data. Every interaction on Facebook—from likes and shares to comments and posts—generates data that must be stored and analyzed. With billions of users worldwide, Facebook processes petabytes of data daily to provide personalized content, optimize ads, and improve overall user experience. Hadoop's scalability and batch processing capabilities allow Facebook to aggregate and analyze data efficiently, supporting its recommendation systems, log analysis, and user behavior insights.
- On the other hand, **Google** initially developed **Google File System (GFS)**, later evolving into **Colossus**, to manage its enormous search index and store user data across its services like Gmail, Google Drive, and Google Photos. Colossus ensures data is replicated across multiple servers, providing high reliability and fault tolerance for real-time applications. Google uses this system to support cloud storage and search functionalities, enabling users to access and retrieve their data seamlessly, while also powering its vast ad network and analytics tools.

Both companies, while using different systems, tackle the challenge of managing Big Data in their unique ways—Facebook focusing on batch processing and user interaction analysis with Hadoop, and Google focusing on real-time data management and retrieval with Colossus.

  - **Analyze how each company utilizes these systems for their specific Big Data needs.**

2. **Performance Optimization in Hadoop:**

   - Explore strategies for optimizing data storage and processing in Hadoop (e.g., tuning replication factor, block size).
   - Propose ways to improve performance for large-scale data processing tasks.
3. **Fault Tolerance in GFS:**
   - Discuss how GFS handles node failures and data recovery in a distributed system.
   - Suggest potential improvements or alternatives based on newer technologies.
4. **Comparison of Performance Metrics:**
   - Compare how Hadoop and GFS handle large file storage, fault tolerance, and scalability based on your research and experiments.
   - Create a performance metrics table summarizing key findings.
5. **Conclusion:**
   - Reflect on the advantages and challenges of using Hadoop and GFS in today's Big Data landscape.

Submission Deadline: 31st October 2024