BIG DATA SYSTEMS – ASSIGNMENT 2

Title:

Implement an efficient data layout and retrieval strategy for a Hadoop Cluster

Overview & background:

A multinational financial services company has a large volume of financial transaction data generated from its branches and online services. The financial transaction data is generated in real-time and is too large to be processed and analyzed using traditional methods. The company needs a scalable and flexible big data solution that can handle the volume, velocity and variety of the data.

The company wants to use big data technologies to store, process and analyze the data to identify trends, detect fraud and make informed business decisions. The company has decided to use a Hadoop cluster with HDFS as its storage system and MapReduce for processing and analysis. The Hadoop cluster, with HDFS as its storage system, provides a cost-effective solution for storing and managing large amounts of data. MapReduce will provide powerful processing and analysis capabilities to extract valuable insights from the data.

Input: CSV data with flat schema with multiple records and features.Link is given in main page

Description:

1. STORAGE:

Each Storage Node will store the data based on below condition.

a. Mutually Exclusive feature data (column value) which is not common across records (rows): private node
b. Feature data common in two records: 2-way shared node
c. Feature data common in four records: 4-way shared node.
d. Feature data common in eight records: 8-way shared node.

Note: Private node, 2,4,8- way shared nodes are storage nodes which stores feature values which are common in 2, 4, 8 records respectively.

2. METADATA

Maintain record ID wise metadata about above storage deployments, which will explain how the feature values are stored across the storage nodes. The meta-data can be stored on a specific node.

3. RETRIEVAL:

For provided record ID, retrieval of record will refer step 2 to fetch all the required features (column values) from respective storage nodes to form the original record.

NOTE: You can apply different techniques to understand the similarity of feature values like normalization, standardization, vectorization etc.

Submission Requirements:

1. A Python / Java code which enables
   a. The given CSV data to be written, using the distributed storage layout strategy described, to optimize the performance for handling the large volume of data, and
   b. Counting the number of financial transactions per customer and per country.
   c. Filtering large volumes of data based on invoice generated by the online services to identify and resolve performance issues?

2. Generate daily, weekly and monthly reports on the financial transactions, including total value, number of transactions, and average transaction value?

3. Ensure data reliability and availability in the Hadoop cluster, and what strategies can you employ to recover from data loss or node failures?

4. Retrieval of any record given the record ID from the distributed storage.

5. **You can use a Hadoop cluster, a plain cluster of a set of nodes, or any BigData storage framework to demonstrate your data storage and retrievalcode. Describe your setup in detail.**

6. **You are allowed to use Hadoop ecosystem products like Pig, Hive, HBase etc.**

7. You should provide clear instructions to reproduce the submission on the Evaluator’s setup.

8. Your code and results should be reproducible

9. The implementation should be general purpose for any other CSV input file.

General Notes:

- Using Canvas, only the first member of the group has to upload the file. No submission over email will be considered.
- Submit the code and a document as a zip file. The document should be a Word or a PDF describing your setup, how to run your code and results as described in "Submission requirements".
- The document in the zip file should have full names of the group members along with the BITS Registration no. of each group member.
- Name the zip file in format like "Grp_<your_group_number>.zip" only. Don't add anything into the file names.

- Make sure that you upload the file well ahead of the deadline. At the last moment, we have seen several groups have faced issues while doing the submissions.
- The assignment report should have section on system configuration in which you need to describe Hadoop cluster configuration and configuration of products like Pig, Hive, HBase etc if you are making use of them.
- If used, PigLatin queries, Hive queries, HBase queries are also to be submitted. Describe in detail, how you moved the data to HDFS, loaded into Pig, Hive or HBase , if used in your solution.

- **Note - Since it is a group assignment, only one submission is expected from each group. Unnecessarily don’t upload the solution on individual basis. If it’s observed, then a penalty (25% reduction) will be applicable on it.**
- **Every group should record a mp4 video which should contain the executions with queries/answers.**
- **Name the video file in format like "Grp_<your_group_number>.mp4" only. Don't add anything into the file names.**
- **Plagiarism will be strictly dealt with and if found will result in cancellation of the Assignment and 0 marks being awarded to all the group members.**
- **The last date of submission will not be extended in any case.**