Cloud Project 4

Clarissa Branje
100716458

Describe how to apply MapReduce to count the words within a certain document.
In the MapReduce word exclude model, we figure out the recurrence of each word. Here, the job of the Mapper is to plan the keys to the current qualities and the job of the Reducer is to total the keys of normal qualities. Along these lines, everything is addressed as Key-esteem pair.

Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.
Apache Hadoop

Differences
Dataproc is compatible with Hadoop and the Hadoop Distributed File System (HDFS) (HDFS). When choosing computation and data storage solutions for Dataproc clusters and jobs, the following features and factors should be taken into account: Cloud Storage and HDFS: The Hadoop Distributed File System (HDFS) is used by Dataproc for storage.

Advantages
- Availibilibility
- Scaling
- Ease of use
- Performance
- Opensource
- Compatibility
- Cost-effective
- Varied data sources

Disadvantages
- Security
- Processing overhead
- Interactive processing

Limitations
- An issue with Small Files
- Slow Processing Speed
- Support for Batch Processing only
- No Real-time Data Processing

Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decided to use another dataset, It should maintain both variety and huge volume. Your report should include but not be limited toThe application. Its impact.The used dataset (size, schema/structure).A graph showing the proposed pipeline(s). List of other tools (AI, clustering,…) needed to implement that application.

Batch processing: large amounts of data in a single batch over a set period of time
Stream processing: immediate processing of a continuous stream of data.

An application that utilizes both of these processes could be a movie streaming service like Netflix. This movie application could use a data set like: The Movie Data set: 239Mb
This dataset includes Metadata on over 45,000 movies. 26 million ratings from over 270,000 users. Inside files include movies_metadata.csv: The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries, and companies.

Batch processing could be processing an update for adding or removing movies in season (for example Holiday movies in December) while Stream processing is processing real-time transactions like creating new accounts or processing media requests.

https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset