Project Milestone-- Data Processing: Dataflow- Apache Beam

Shanjay K

1. Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc.

An alternative to Google Cloud DataProc is Amazon EMR. Amazon EMR is another big processing service that is usually used in the cloud environment and is a managed cluster platform that simplifies running big data frameworks.

- AWS EMR vs GCP DataProc vs GCP Dataflow
    - Differences
        - Amazon EMR has a greater market share with about 11x more customers and is used in about 2x more countries.
        - Google Dataflow can automatically run jobs on a cluster such as balancing work or Scaling the number of workers for a job which can be very time consuming on other systems.
    - Similarities
        - Resizing
        - Autoscaling
        - Flexible virtual machine types

2. 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.

A practical application using both stream and batch processing that can be applied to a given dataset could be sensor data from some type of social media platform. Let's take Facebook for example where content coming in will be stream processed and displayed to other users while simultaneously being stored for later processing using an HDFS. Later, the stored immutable data can be batch processed and analyzed to view certain trends such as times of most user activity, if world events relate to increased user activity, etc.

The dataset that'd be used would be things such as a user's info (geographical location, age, gender, etc) to view trends amongst other to see if there are any similarities. There would need to be AI tools that would automatically search through all tops and analyze trends to see if an action x leads to y being discussed and so on.