Jana

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. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Advantages

- Serverless simplicity-it doesn't require software installation,additional licensing costs and overhead
- Fast exploration and anomaly detection-data is explored instantly
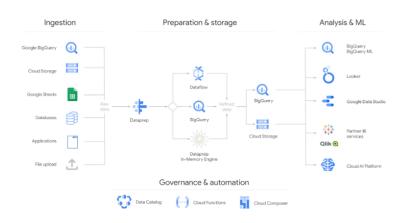- Easy and powerful data prep-predicts next data transformation

Disadvantages/Limitations

- API access limited
- Don't have access to data quality rules
- No scheduling/plan management
- Support is limited

Dataflow Vs DataProc Vs DataPrep

- Dataprep-transformation/preparation tasks are ETL processes
  - Preps and cleans the data for its future use
- Dataflow- batch/stream processes of data
  - Supports automatic provision to clusters
- Dataproc-clustering computer framework
  - Supports manual provision to clusters

Dataprep ELT pipeline architecture

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. Your report should include but not limited to

- Many companies require both stream and batch processing. Where you can process current data and large batches of data at different points
- This way businesses can operate more efficiently and faster
- Example of applications for this is social media platforms, more specifically instagram
- Instagram requires processing of large amounts of data due to its millions of users and requires processes current data like when someone posts something instantly
- The impact that batch/stream has is huge as if these applications can't do this on time people won't be using the platforms
- Instagram requires AI as it uses the users patterns to suggested different pages and have personalized ads to the person's interests