Project Milestone-- Data Processing: Dataflow- apache beam

Dataprep, Dataproc, and Dataflow are three distinct parts of the new age of data processing tools in the cloud. They perform different tasks yet are related to each other.

	Dataprep	DataProc	Dataflow
Differences	- A service for visually exploring, cleaning, and preparing data for analysis - Easy to use - Fully automated provisioning of clusters - BigTable and BigQuery - UI driven data processing - Scales on-demand	- A managed service that allows you take advantage of open source data tools for batch processing, querying, streaming, and machine learning - Simple, easy to use - Provisioning clusters is done manually - Apache Spark and Hadoop - Data Science/ML Ecosystem	- A serverless data processing service - Relatively difficult - Serverless, automatic provisioning of clusters - Apache Beam - Batch and Stream Processing of data - Resources produced or removed on-demand
Advantages	- No limit to the file size - It will be easy to use and integrate other services provided by Google Cloud Platform - People can access multiple data sources from Cloud Storage and BigQuery for further analysis - No need for any VM - Prepared data can be used by services like Google Cloud Machine Learning Engine to train ML models and analysis	- Low cost - Super fast - Integrated - Managed - Simple and familiar	- Less operational costs - Stream data analytics with speed - Simplify operations and management
Disadvantages /Limitations	- Sort transform is not supported - It doesn’t support user-defined functions - It doesn’t support custom dictionaries and data types - User access to administrative	- You cannot change the machine type of an existing cluster, Dataproc does not support this - Therefore, you need to consider the size of the individual machines in advance	- Bound to Google technologies - Not suited for experimental data processing jobs

	functions is not supported - There may also be some limitations to file formats	- Dataproc does not limit the number of nodes in a cluster, but other software may have limitations	