SOFE 4630U: Cloud Computing
March 29, 2022
Group 11 - Group Report
Project Milestone - Data Processing: Dataflow - apache beam
Fajer Zayed (100672347), Ireni Ruthirakuhan (100657302), Raveenth Maheswaran (100704540), Yale Wang (100673933)

GitHub Link: https://github.com/fzayed/Project-Milestone-Group-11.git

---

Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP).

- Amazon EMR
- DataPrep

Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Comparing DataFlow, DataProc and Amazon EMR

- Similarities between **DataFlow**, and **DataProc** and **Amazon EMR**
    - All give you ETL data warehousing
    - Amazon EMR runs a simplified version of DataProc big data frameworks

- Major Differences
    - Data Proc gives you access to Hadoop
    - Dataflow gives a platform to use Apache Beam
    - Amazon EMR allows you to choose which file systems you want to work off of
    - DataFlow and DataProc are both serverless while Amazon EMR has a server platform
    - DataFlow offers both batch and stream processing while DataProc only offers batch processing

- Advantages of Data Proc
    - Simplifies operations
    - Easy install/resize
    - Very secure as it provides multiple security integrations such as Apache Ranger, Kerberos, and Personal Authentication
    - It is flexible to use, as it is serverless and uses Kubernetes to manage the clusters
    - It is cost-effective compared to other options, as it uses 57% less than the average total cost of ownership of on-premise data lakes
- Disadvantages of Data Proc
    - No choice of version choice for Hadoop/hive/spark stack.
    - Cannot pause/stop a cluster.
    - DataProc doesn’t offer stream processing (hot path)
    - DataProc needs manual provisioning of clusters
        - Takes time to configure clusters
        - Requires more resources to do so

- Advantages of Dataflow
    - Don’t need to manually balance
    - Automatically scales
    - Quality of support

    - Product direction
    - Horizontally scalable to ensure all resources are utilized efficiently
    - It is very reliable and consistent
- Disadvantages of Dataflow
    - Not as flexible
    - Can't customize implementation
    - Not preferred if Hadoop dependencies are required
    - Not as fast compared to other data processing software
        - But not as significant

- Advantages of Amazon EMR
    - Scalable
    - Flexible
    - Secure
    - Fully managed service
- Disadvantages of Amazon EMR
    - RAM is fixed
    - Harder to manage

- Limitations of Amazon EMR
    - LImit to APIs burst capacity (rate limit)
    - Limit to number of apis that can be called at a single time (bust limit)

Comparing DataFlow, DataProc and DataPrep

- Similarities between **DataFlow** and **DataProc** and **DataPrep** include
    - These are GCP products
    - Used for big data processing
- Major Differences between dataflow and dataproc and dataprep include
    - Dataprep = ui driven
    - Dataflow batch and stream processing of data
    - dataProc has machine learning and data science as a service

- Advantages of Data Proc
    - Easy to use
    - Hands on approach
- Disadvantages of Data Proc
    - Cannot choose which version to use of the particular stack
    - Inability to stop the cluster or pause it

- Advantages of Dataflow
    - Uses batch and stream processing of data
    - Creates new pipelines for data processing and resources produced
    - Fully managed

    - Used for batch and stream processing
    - Fast
    - Serverless
    - Cost effective
    - Provides portability with processing jobs
    - Removes operational overhead
    - Engine separates computation and storage improving data latency and autoscaling
- Disadvantages of Dataflow
    - Cannot be scalable without violating API contract
        - Cannot scale to 0 workers

- Advantages of DataPrep
    - Quickly explore new datasets
    - Flexible
    - Support data transmission needs
    - Easy to use
- Disadvantages of DataPrep
    - Only use as a medium of processing data further use → BigQuery
    - Data quality rules are not available

- Limitations of DataProc
    - Cannot stop or pause a DataProc cluster
    - UI for managing the cluster specific configuration is not available
- Limitations of Dataflow
    - Can only run 25 concurrent jobs at most
    - Limitation of 1000 compute engine instances
- Limitations of DataPrep
    - Number of workspaces limit to 1000
    - Limited access to APIs

Comparing DataFlow, DataProc and Data Pipeline

	Google Cloud **DataFlow**	Google Cloud **DataProc**	Amazon Web Services **Data Pipeline**
Batch Processing	Yes	Yes	Yes
Stream Processing	Yes	No	Yes
Provisioning	Automatic	Manual	Automatic
Data Writes	PubSub, GCS, BigQuery, SQL, etc.	Vertex AI, BigQuery, DataPlex	DynamoDB, SQL, RedShift tables
Data Sharing	Yes	Yes	Yes

Suggest a practical application using both stream and batch processing that can be applied to a given dataset.

- There are many practical applications that can be used towards stream and batch processing through the incorporation of DataFlow
    - In fact, DataFlow can be applied to situations where stream analytics (business insights), real-time AI (fraud detection), and log processing (identify system health) is required.
- We will focus on a social media application that uses (Stream analytics)
    - Batch processing to track logins
    - Stream processing to track social media interactions

Its impact.

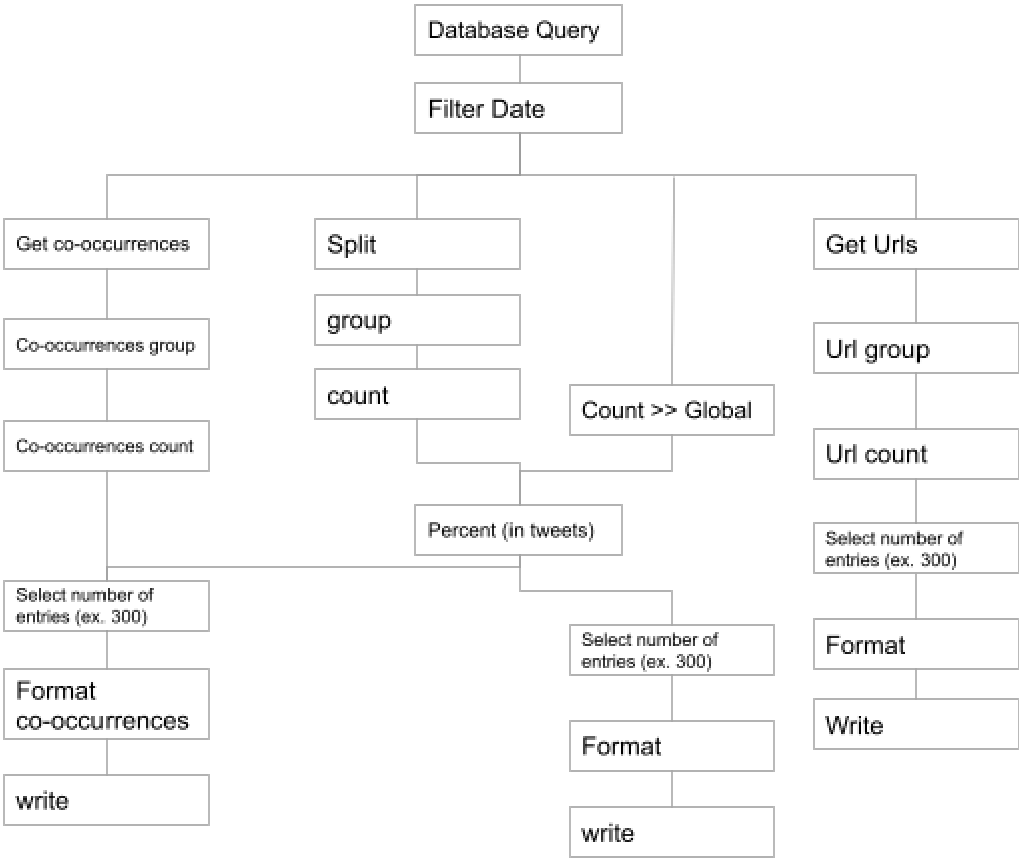
- This application can have an incredible scalability to number of users
- Social media applications are good for maintaining many connections
- Implementing lambda architecture can greatly improve performance and data handling in various ways
    - Can reduce complex queries by implementing stream processing for real time requests
    - Storing data in batches rather than storing data individually

The used dataset (size, schema/structure).

- The used dataset was the same images and csv files from Lab 3
- The data can be organized based on the user's activity in their social media account
    - CSV file can store multiple types of data such as user's friends list, posts, their media files, etc

Dataflow pipeline graph/diagram

The graph of the pipeline presented below represents the dataflow of when data is taken from a social media platform and then stored within GCP. Within this pipeline diagram we can see that the values branches/forks into 3 different sub branches. One branch identifies the popular words in percent, popular urls from count and the last one identifies the fitting word co-occurrences. With these results from each branch, all are then written to BigQuery Tables.



List of other tools (AI, clustering,…) needed to implement that application

- From edge devices using a pub/sub can then create a topic and therefore have a subscription. This can then be deployed as batch and stream processing through the use of dataflow.
  - To analyze the received data, tools such as BigQuery, BigTable, and VertexAI can be used to create tables, and datasets
  - This can be used to help with the production build and call of external APIs

- The other tools used in the videos include:
  - Analyzing data with BigQuery
    - To ingest user activity
  - Vision API
    - To analyze content of images using pre-trained neural networks
  - Cloud translation
    - To provide the service in many languages

References

https://cloud.google.com/dataflow

https://cloud.google.com/blog/products/gcp/analyzing-tweets-using-cloud-dataflow-pipeline-templates
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html