**Dataflow**

- A service that formats data generated through streaming and batch processing
- The pipeline has 3 steps
    - Read data from a source, transform data, write data back into a sink
    - Data retrieved to read is obtained from a source and put into a parallel collection
        - Where parallel means that it can be distributed across different machines
    - Transforms the data within the P collection by using operations forming new P collections
    - Final transform then send the final P collection to a data sink
- In the case where you need to easily share the pipelines with team members and organizations

- Dataflow is used to deploy and execute pipeline
    - VMs do the data processing

- Compute and storage are handled separately

- Use cases
    - Stream analytics
    - Real time AI
        - Fraud detection
    - Log processing
        - System health

**Apache_beam**

- Library that is used to describe the pipeline
- Defining both batch and streaming data parallel-processing pipelines

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

- DataProc
    - Contains data tools to help with streaming, batch processing, querying
    - Provides a way to help with the creation of clusters and its management
- Other processing services used within the cloud environment include:
    - Dataprep
    - Databricks lakehouse platform
    - Amazon kinesis
    - Snowflake
    - Microsoft SQL server
    - Apache Kafka
    - Amazon EMR
    - Spark Streaming
    - Confluent

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

- 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

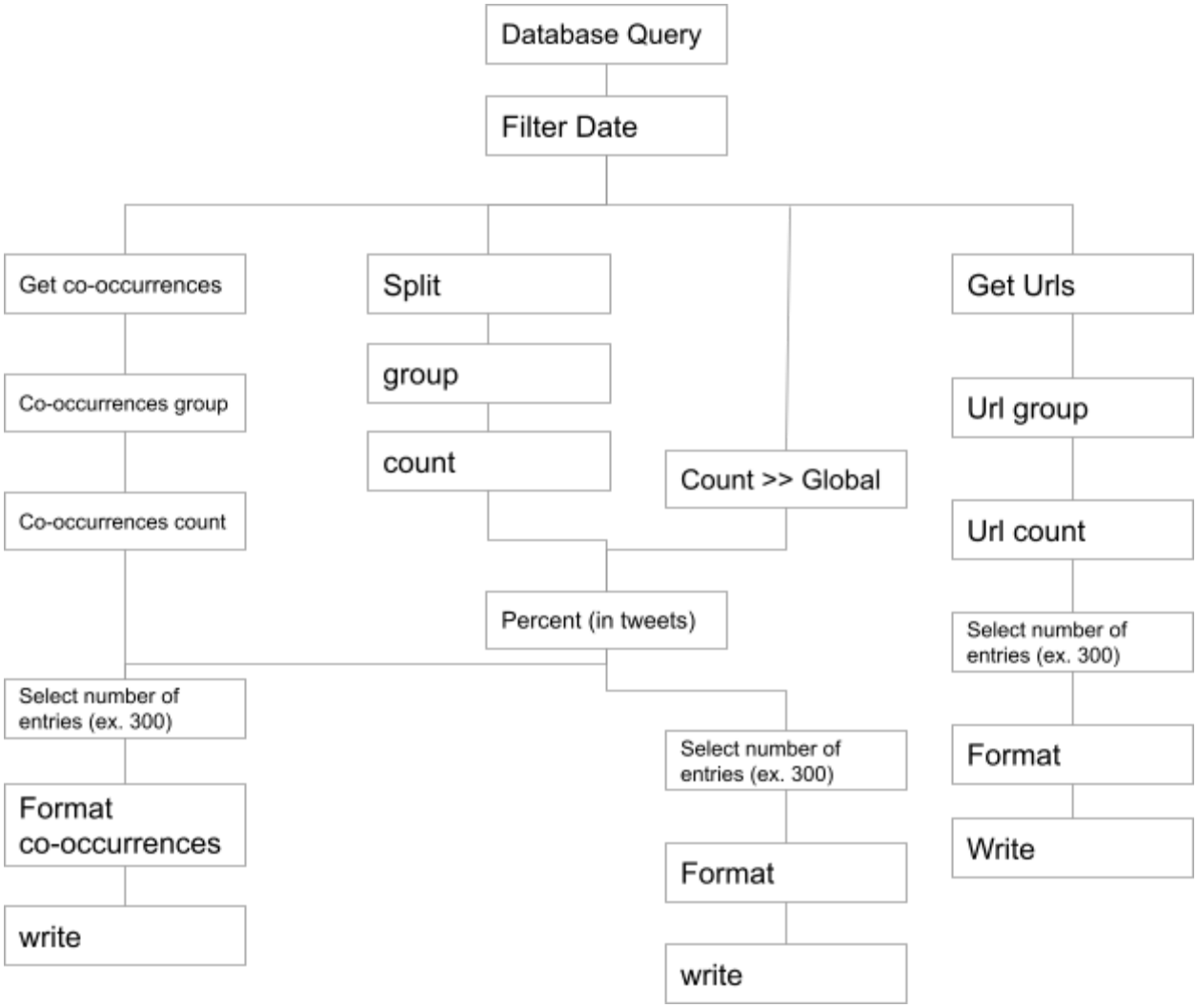
        - UI for managing the cluster specific configuration is not available
- Limitations of Dataflow
    - Can only run 25 concurrent jobs at most
    - Limitation of 1000 compute engine instances
- Limitations of DataPrep
    - Number of workspaces limit to 1000
    - Limited access to APIs

According to the dataset given in the previous project section

- Practical applications that use stream and batch processing include
    - Stream analytics
        - Business insights
    - Real time AI
        - Fraud detection
        - Enabling predictive analytics
    - Log processing
        - System health can be identified through the logs

**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.




**Other tools needed to implement the 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
- It can also be used with tools listed below:
    - Cloud machine learning
    - Apache cassandra
    - Mongodb
    - redis

**References**

https://wisdomplexus.com/blogs/dataproc-vs-dataflow-vs-dataprep/#:~:text=Dataproc%20is%20a%20Google%20Cloud,on%2Ddemand%20and%20fully%20automated.

https://cloud.google.com/dataflow
https://cloud.google.com/blog/products/gcp/analyzing-tweets-using-cloud-dataflow-pipeline-templates