SOFE 4630U
Cloud Computing

Project Milestone
2022/03/29

Data Processing: Data flow - Apache Beam

https://github.com/PranjalS1/CloudPM-ApacheBeam.git

Sabesan Sivakumar (100701928)
Jerusha Macwan (100723319)
Matthew English (100704553)
Pranjal Saloni (100653360)

Q1

Group Answer:

Another processing service used in cloud environments is Google Cloud Dataprep.

	Dataflow	Dataproc	Dataprep
Main use	Exeutes Apache Beam pipelines within the GCP environment.	Allows usage of open source tools for machine learning, batch processing and querying	Visually explores, cleans and transforms structured and unstructured data for analytics and reporting by converting the file into JSON,CSV or a table format.
Major differences	Follows batch and stream processing of data. Can only observe and stream processing data.	Dataproc is a Google Cloud product with Data Science/ML service for Spark and Hadoop.	Dataprep is UI-driven, scales on-demand and fully automated.
When to use?	Dataflow provides a clear separation between processing logic and the underlying execution engine.	Dataproc should be used if the processing has any dependencies to tools in the Hadoop ecosystem.	Dataprep is used only as a medium of processing data for further use, such as in BigQuery.
Advantages	Helps understand the functioning and the limits of a system.	Utilizes Spark and Hadoop; help create and manage clusters quickly.	Allows quick exploration of new datasets.
Disadvantages/Limitations	Limited job computation. Reduces Latency and time taken for processing	No choice of selecting a specific version of Hadoop/hive/spark stack. Size of clusters must be known in advance	Limited access to APIs.

		as it cannot be changed	

Q2

Group Answer:

- Batch processing is when the data is being collected first then being processed while stream processing the data and processing is happening in real time. Batch processing is used when there is a large amount of data to be processed.
- An example of batch processing will be payroll. In a payroll application the large amount of data would be each employee name and their account information. Once the data is collected then it is processed all together in a batch as every employee gets paid on the same day. Shown below is a diagram of a batch processing for payroll

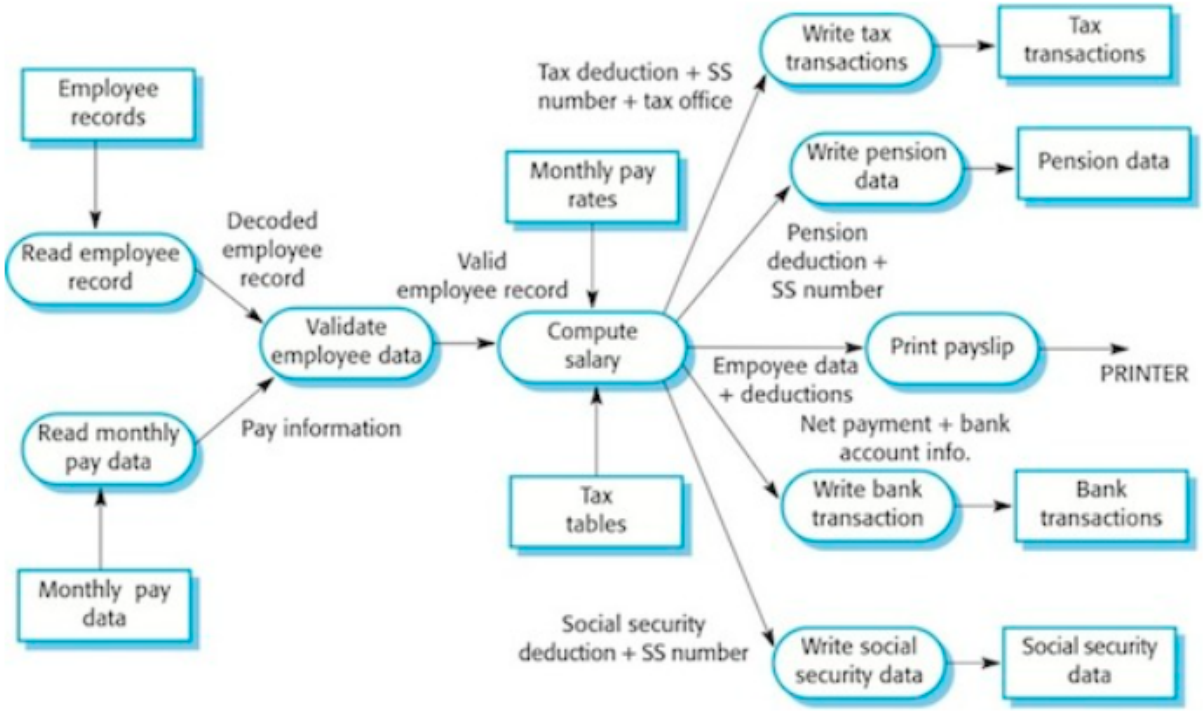

Figure 1: Batch Processing Payroll
Source: Adapted from[1]

- An example of stream processing would be fraud detection as you want to collect and process the data in real time in order to detect fraudulent activity. Both of these examples can be put together as we can detect fraud from payroll.
- An example of a fraudulent activity that may occur in payroll is pay rate altercation fraud.

- This application will have a huge impact as it helps ensure both employees and payees get treated equally and quickly identify and remove any fraudulent activities.
- Tools that can be implemented for this application is BigQuery. BigQuery is a GCP service that can help manage and analyze all sorts of data.
- In order to be able to successfully stream data we need to ensure that we have some IAM permission such as bigquery.datasets.get and bigquery.tables.get .

References

[1] Batch Data Processing Systems. [Online]. Available: https://ifs.host.cs.st-andrews.ac.uk/Books/SE9/Web/Architecture/AppArch/BatchDP.html. [Accessed: 28-Mar-2022].