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)

1.

Sabesan Sivakumar Answer:

Processing Service	Dataflow	DataProc	Databricks
Main use	-An area where Apache Beam jobs can be run - Can only watch and stream processing data	-Utilizes Spark and Hadoop services for open-source data that can be used for batch processing, streaming, etc.	-Utilizes Spark services -Uses the “lakehouse” platform. -Allows queries to be run on data that is not fully structured in a database schema
Advantages	- Fast - Simplicity: Since it is technically “Serverless”	- Low cost - Can utilize both Spark and Hadoop services	- Versatile(Python, Java, SQL) - Can be utilized for small jobs
Limitations/Disadvantages	- Individual jobs can have a maximum of 1000 compute engine instances	- Machine type cannot be changed of an already existing cluster	- Most of its services are available separate Azure cloud services so there really is no need.

2.

Group Answer:

- Batch processing is when the data is being collected first then being processed while stream processing the data and processing are 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's name and 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 batch processing for payroll

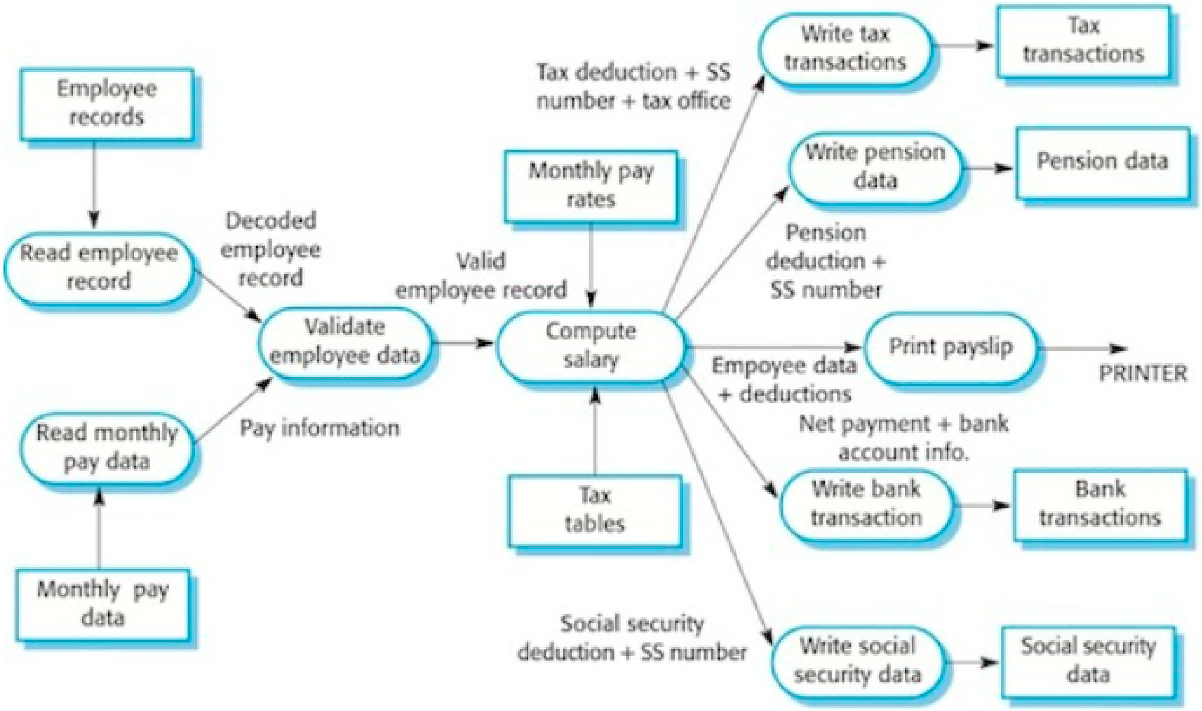

Figure 1: Batch Processing Payroll
Source: Adapted from[1]

- 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 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.
- Tool that can be implemented for this application is BigQuery. BigQuery is a GCP service that can help manage and analyze all sorts of data. I
- In order to be able to successfully stream data we need to ensure that we have some IAM permission such as big query.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].