OntarioTech UNIVERSITY

SOFE 4630U
Cloud Computing

Project Milestone
2022/03/29

Data Processing: Data flow - Apache Beam

Jerusha Macwan (100723319)

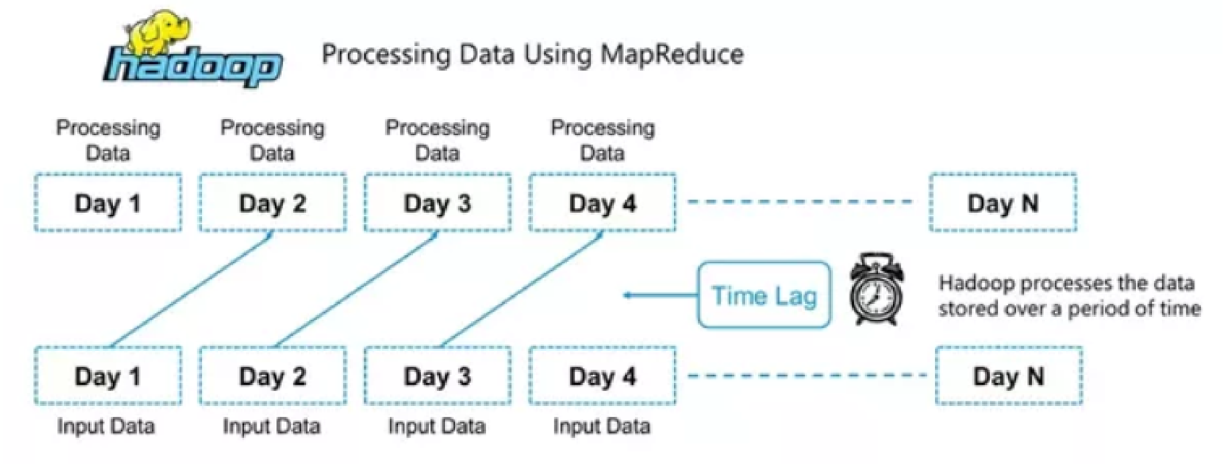
Google Cloud has another processing service called DataProc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and DataProc. Your comparison may include but is not limited to the major differences, advantages, disadvantages, and limitations.

Another processing service used in cloud environment is Dataprep.

Processing Services	Dataflow	Dataproc	Dataprep
Main use	Fully managed analytics solution Uses batch processing and autoscaling	Allows usage of open source tools for machine learning, batch processing and querying	Explores and prepares systematic and unsystematic data for machine learning and evaluation
Advantages	Reduces Latency and time taken for processing	Easy creation and management of clusters	User friendly Easy integration with other GCP services
Disadvantages	Scalability and performance bottleneck	No user interface for management of configuration specific to a cluster	Functions, datatypes and dictionaries defined by users are not supported

Suggest a practical application using both stream and batch processing that can be applied to a given dataset. It's expected to use the dataset uploaded in the third milestone but you can use any other dataset. If you decide to use another dataset, It should maintain both variety and huge volume. Your report should include but not limited to:
The application.
Its impact.
The used dataset (size, schema/structure).
A graph showing the proposed pipeline(s).
List of other tools (AI, clustering,…) needed to implement that application.

Batch processing refers to the processing of data blocks that have already been kept for a long length of time. For instance processing all of a big financial firm's transactions in a single week. This data contains records which might be in the range of millions for a single day. This data can be saved as a file, record or other type of data. This file will be processed at the end of the for several analysis that the firm wishes to conduct. The processing will take a long time. This will use Batch processing. The best framework which can be used here is Hadoop MapReduce. This figure will give a detailed description for the same.



Processing Data Using MapReduce

Figure 1 Processing data using MapReduce

Stream processing is significant for fraud detection and other applications. Anomalies can be detected that lead to fraud in real time and block transactions that are fraudulent before they are processed if stream processing is used in transaction data. The following figure shows the same.

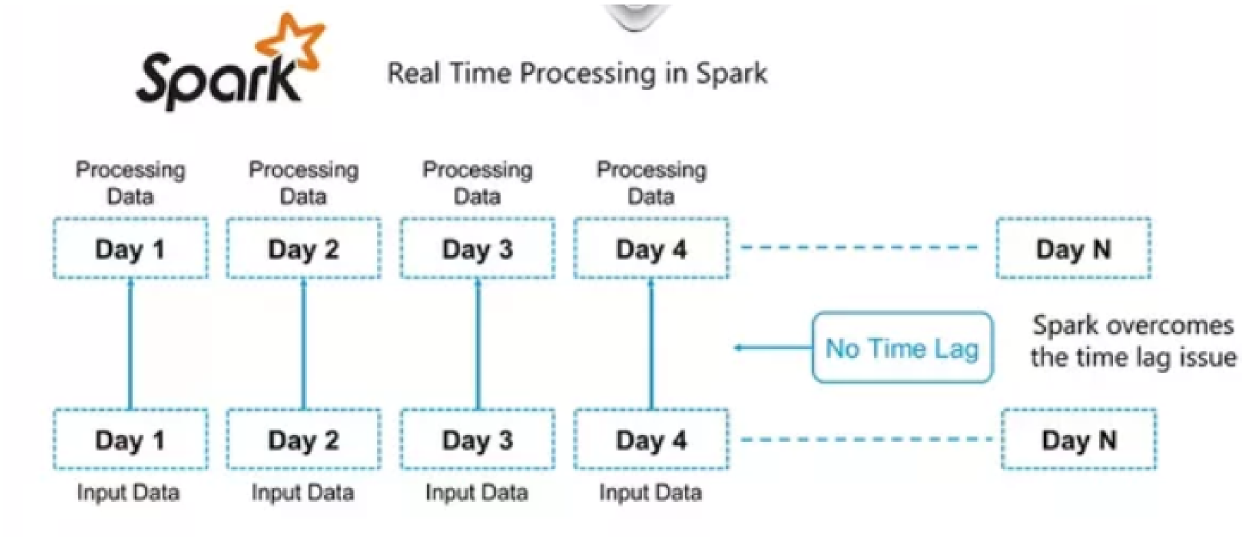


Figure 2

References

[1] Big Data Battle: Batch Processing vs Stream Processing[Online]. Available:
https://gowthamy.medium.com/big-data-battle-batch-processing-vs-stream-processing-5d94600d8103
[Accessed: 28-March-2022]

[2] Big Data Battle: Batch Processing vs Stream Processing[Online]. Available:
https://gowthamy.medium.com/big-data-battle-batch-processing-vs-stream-processing-5d94600d8103
[Accessed: 28-March-2022]