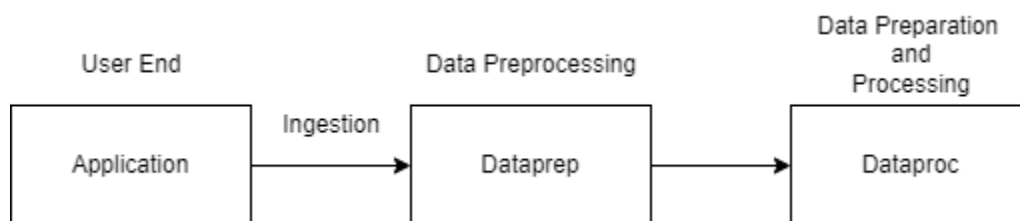


- 1. Google Cloud has another processing service called Dataproc. Name another processing service that is usually used in the cloud environment (not necessarily GCP). Compare between it and both Dataflow and Dataproc. Your comparison may include but is not limited to the major differences, advantages,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

Dataproc is a service offered by google cloud platform that allows users to manage different data processing and analytics services for their data, including Hadoop, HiveMQ, Apache spark, jupyter, and others. Users can create a cluster for their system to handle larger amounts of data processing, and still have it be fully managed by dataflow. Dataflow allows users to handle and manage this on a much larger scale. It can be used as a map reduce for processing of big data, and use the batch processing techniques to gain insights provided by machine learning, or real time data analytics. Essentially, it's a scalable data processing system that can easily be integrated with big data. Users also gain the benefit of containerizing their dataproc jobs using Kubernetes to fit with the configuration of their entire system. Dataproc also offers serverless deployment [1], and an easy to use API for developers to seamlessly integrate dataproc into their system architecture.

On the other hand, dataprep is a big data analytics service that leans more toward providing the customer with visualization of the data, and removal of redundant data. A similar process applies, where a cluster will be created and fully managed, but dataprep is seen more with the use case where data is first processed before being stored. Dataprep also deals with more visualization for the developer, whereas dataproc deals with actual real-time processing and collection of data. Unlike dataproc, dataprep is not as well integrated with Hadoop, whereas dataproc was initially designed for integration specifically with Hadoop. Google integrates machine learning to predict ideal data transformations, so that queries are optimized, and they utilize bigquery allowing users to process large amounts of data "dumps" (all types of schemas or non schemas) in a fast amount of time with a few clicks. According to google, data prep normally deals with the data processing after ingestion (before storage), while dataproc deals with processing the actual stored data, as well as the real time data in the next stage of the pipeline [2].



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

The practical application in my opinion would be a smart home control system that would control temperature of houses in an area. The application would contain sensors inside and outside the house to measure windspeed, temperature, humidity, weather conditions, and other environmental factors. The sensors inside would measure the temperature within the houses and there would be an actuator that regulates the humidity, and temperature within the house based on the conditions outside.

The application would function by creating a “subnetwork” of a city which would be selection of various houses in the same area, temperature data would be accumulated from outside each house to ensure a variety of datapoints for the application which would result in a more accurate determination of the weather conditions. As mentioned before, there would be two sources of data, so kafka would require two separate schemas for the data ingestion module:

Sensors within the house [home sensors] would follow the following schema:

- {temperature: float, lighting: int, humidity: float, number of people: int, area of house to regulate: float}

Sensors external to the house:

- {temperature: float, time of day: datetime, humidity: float, precipitation: float, windspeed: float, natural disaster: Boolean}

Data would be constantly streamed every second, for multiple houses, from both outdoor and indoor sensors. This would generate a large volume of data after a short while.

This application will follow the lambda architecture which uses a batch layer and speed layer, also known as a cold and hot path. The hot path follows quick analytics to make real time decisions, while the cold path usually caters to longer term analytics and storage. Before entered into the paths, and right after ingestion, preprocessing is needed to ensure the data is clean, there are no missing fields, etc. This application will use dataprep cluster for this since it follows the preprocessing mechanisms defined in the previous answer. The “removal of redundant data” feature that is offered by dataprep however, will not be utilized since we want all data to be stored in a data lake, and schema on read availability when the data is queried.

In this application, the cold path will constantly feed data to a machine learning model using HDInsight to update the model, and improve its efficiency at determining the best possible temperature based on current conditions. In this path, the data will also be fed into Azure DataLake that will store the data for a long period of time, allowing the user to benefit from the schema-on-read ability when the data is used for further processing.

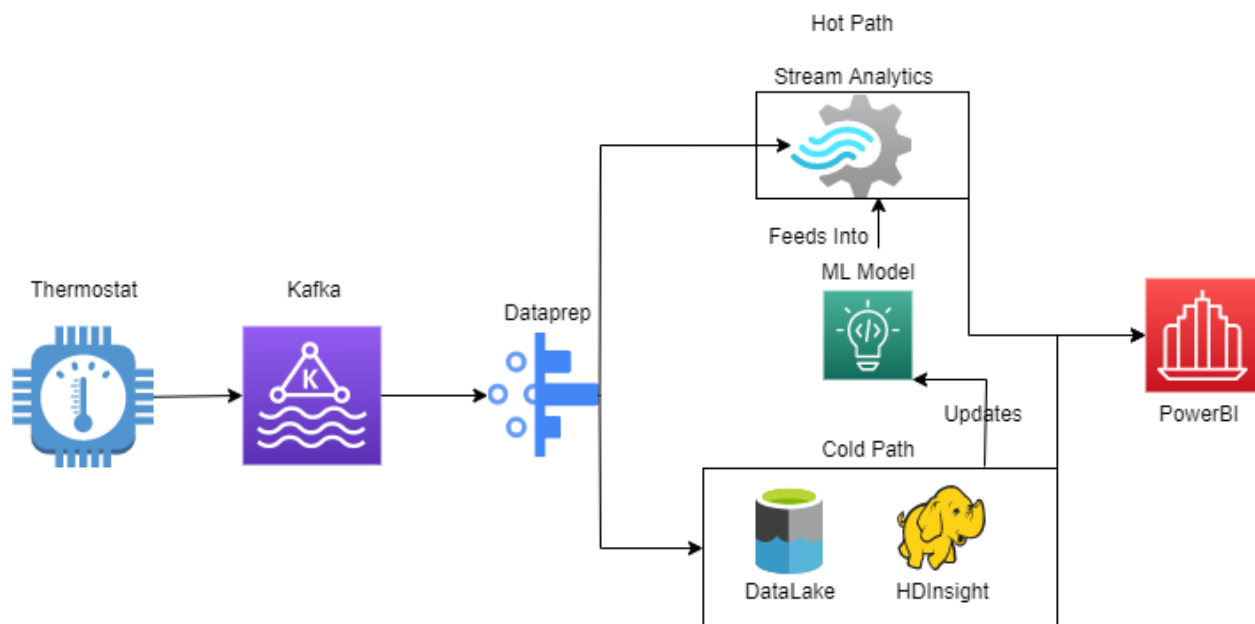
the hot path will use Azure stream analytics to make real time decisions based on the current conditions in the area. It will use the constantly updated machine learning model to get the best

possible temperature for a given house based on the conditions in that area from the given house, and its surrounding areas.

On the business end, a powerbi user interface will be used to display the distribution of data in given areas, and understand how the temperature is regulated over a given area. This will be useful in case the ML model contains any obvious visual bias that needs to be fixed, and it would cause the ML model to need to be retrained.

With the rising cost of energy, and the need to aim all users to be more energy efficient, this application will allow for smart controls of a users home to ensure that energy is not wasted. If a user leaves their home to go on vacation, or simply leaves for the workday, there will be no wastage of energy since the system will be able to determine if a user is home, and regulate the energy that way.

As explained, here is a graphical representation of the pipeline:



References

[1] “Dataproc | google cloud,” *Google*. [Online]. Available:

https://cloud.google.com/dataproc?utm_source=youtube&utm_medium=unpaidsoc&utm_campaign=CDR_ali_databases_jj6mp7sam10_CloudBytes_020421&utm_content=description.

[Accessed: 29-Mar-2022].

[2] “Dataprep by Trifacta | google cloud,” *Google*. [Online]. Available:

<https://cloud.google.com/dataprep>. [Accessed: 29-Mar-2022].