4. How to monitor and maintain this solution so we can proactively warn our users of any issues?

Ans.

Monitoring the Data Processing Pipeline

- Create a dashboard to analyze the processed data. We can create many metrics which we can monitor on a regular basis. If there is any deviation in metrics we can trigger alerts. For example, number of rows in the data for a given day or number of missing values in the data etc.
- We can create expectations for the data using python packages like '**Great Expectations**'. If the expectations are not satisfied then an alarm can be raised.
- If the data pipeline fails for any reason, we can set an alarm

Monitoring the Model

- As we will be collecting the actual weather data and the model predictions, we can monitor how the models are performing in production. If the model quality decreases for any reason we can trigger an alarm.
- We can monitor for data distribution shifts. If there is a difference in data distribution which was used for modeling and production, then the model needs to be retrained.

Maintain

- We can track the usage patterns for the data tables, dashboards etc. If any table is not being used we can stop processing data for that particular table. We can establish touch points with analysts and data scientists. Understand the change in their requirements and make changes to the data processing. Understand any new features they are expecting and provide them on the platform.
- Finding out the bottlenecks in the process and increasing the efficiency to reduce cost and improve the impact on the business

5. Share three best practices for data engineers when designing pipelines to implement your recommendations.

Ans

Top 3 Recommendations for the data engineers:-

- Provide a data dictionary, definitions for the features created and data lineage for the datasets.
- Partition the data for different tables after understanding the requirements and access patterns from the data scientist and data analysts. **Setting up a process to analyze what is the value per byte for the data stored?**
- **Build a data platform like a product. Prioritize long term growth and sustainability over short term gains**

Bonus Recommendations:-

- Avoid using PII data, in case it is not avoidable then the data should be anonymized and processed for downstream purposes. Setting up a process for data access (Role based access control). Setting up the process to take care of data privacy / data deletion requests.
- Parameterize the pipelines / creating API's for common operations
- Incrementally ingest data by using change capture systems
- Store the data in Parquet format and optimize the file size as per the requirement of the compute engine.
- Setting baseline expectations for data reliability.
- What is the value per byte for the data stored
- Queries Per Second (QPS), Query latency and data latency