**Question**

*Simulate a data analysis pipeline for climate change's impact on agriculture. Modify the dataset on:*[*https://www.kaggle.com/datasets/prasad22/weather-data*](https://www.kaggle.com/datasets/prasad22/weather-data)*to support your needs. Ingest and analyze a simplified dataset and identify correlations between weather and crop yield. In addition, simulate the challenges of large datasets* and propose a conceptual data center model to manage them. Deploy your solution on the cloud server you will host under Google Cloud Free Tier (document processes or rationale to do with selecting a provider; registration and verification process; server configuration; deployment, monitoring, and maintenance of the server).

**The data pipeline summary conducted (ETL)**

|  |  |
| --- | --- |
| **Process** | **Description** |
| Extract | Collecting and ingesting the dataset from Kaggle, modifying it to include crop yield, crop type and region for agricultural relevance (using Kaggle api scrapper) |
| Transform | Cleaning, preprocessing and engineering the data. This also included the data augmentation process by normalizing the variables to produce another column names Crop\_Yield. Other cleaning processes included renaming column names and removing the time component from Date column |
| Load | Storing the processed dataset in a structured format (CSV) for analysis and modeling. |

**Dataset Augmentation**

* Adding crop\_yield and Season columns simulate analysis and trends.
* This was done by normalizing temperature, precipitation, humidity and windspeed variables.
* Simulating the season column was based on the Date column already given. Grouping method in python was implemented

**Data Ingestion and Preprocessing**

* Using python pandas and PySpark to load the dataset.
* Normalizing or standardize numerical features.
* Split the dataset into (X, Y) features to fit the LinearRegression model so as to calculate the coefficients of variables to understand feature importance on crop yield by location.

**Proposed Data Analysis**

* Temperature, humidity, precipitation and windspeed distributions across locations (using folium mapping)
* Location by total yield (bar graph /using plotly)
* Crop Yield Trend analysis by location (line graph, using plotly)
* Correlations of temperature, humidity, precipitation, windspeed against date
* Correlation of temperature, humidity, precipitation, windspeed to crop yield (at each location and using weighted averages of the climate conditions to the total yield by location)
* Variable importance to total yield (user should see how yield total adjusts based on the variables selected, a checkbox to tick variables and see yield behavior)

**Simulating Large Dataset Challenges**

* Slow data ingestion and processing, taking so time to execute a data analysis process
* Memory limitations on local machines.
* Difficulty in parallelizing computations.

**Proposed conceptual data center model**

* Using distributed storage (Hadoop HDFS, Google Cloud Storage).
* Implementing distributed computing (Apache Spark, Google Dataflow

**Cloud Deployment on Google Cloud Free Tier**

**Registration and Verification**

**Server Configuration**

* **Settting up a virtual machine (VM):**
* Go to Compute Engine > VM Instances.
* Create a new instance (e2-micro for Free Tier).
* Choose an OS and configure storage (30 GB).
* SSH into the VM and install Python, Pandas, Scikit-learn, Flask and other required libraries.
* Settting up a Jupiter Notebook server for analysis.

**Deploy the Solution**

* Google Cloud Storage to upload the modified dataset.
* Using VM to run the data pipeline and analysis scripts.
* Streamlit or FastAPI to create a simple web app for visualizing results.
* Deploy the app on the VM and expose it using an external IP.

**Monitoring and Maintenance**

* Google Cloud’s monitoring tools to track CPU, memory and disk usage.
* Alerts for high resource usage or downtime.
* Updating software and security patches.
* Scaling resources (upgrade to a larger VM).