ETL PIPELINE OUTLINED STEPS

The detailed summary of the steps taken to carry out the ETL process for **Agricultural Monitoring Systems.**

The primary objective of this project is to extract, transform, and load (ETL) the raw sensor data into structured fact and dimension tables within our snowflake data warehouse.

The steps taken below are.

STEP 1:

The raw data was Loaded from all source tables (SensorDataRaw, WeatherDataRaw, SoilDataRaw, CropDataRaw, PestDataRaw, IrrigationDataRaw, LocationDataRaw) into staging tables (STG_CROPDATARAW, STG_IRRIGATIONDATARAW, STG_LOCATIONDATARAW, STG_PESTDATARAW, STG_SENSORDATARAW, STG_SOILDATARAW, STG_WEATHERDATARAW) within the Data Warehouse AVONMAET Schema.

STEP 2: DATA PREPROCESSING AND DATA CLEANING

The data in the staging table was inspected for issues with quality and structure. The anomalies found in the data and the cleaning action taken is listed below

FOR TABLE STG_CROPDATARAW

Anomalies discovered

- For Crop type column

1) Cron instead of Corn

2) Wheaat instead of Wheat

3) Presence of null values (0.0002% of the column)

4) Wrong data type

- For Growth stage

1) Flowerring instead of Flowering

2) Vegatation instead of Vegetation

3) Presence of NA and Null values (10.02% of the column)

4) Wrong Data type

- For Pest Type

1) Aphidds instead of Aphids

2) Presence of null values and NA (10% of the column)

3) Wrong data type

- Crop Yield

1) Presence of NA and Null values (10.02%)

2) No non numeric values

3) Convert to SI units (Kgm-2)

4) Wrong DataType

- Timestamp

1) Wrong data type

FOR TABLE STG_PESTDATARAW

Anomlies detected

- For Pest_type

1) Aphiods instead of Aphids

2) Slogs instead of Slugs

3) Wrong Data Type

- Pest Description

1) Poorly spelt words

2) Wrong Data Type

- Pest_Severity

1) Hihg instead of High

2) Replace NA with Null

3) Wrong Data Type

- Timestamp

1) Wrong Data Type

FOR THE WEATHERDATARAW TABLE

Anomalies found in this table are

- The weather condition

1) Poorly spelt words

2) Presence of NA

3) Wrong Data Type

- The Wind Speed

1) Presence of NA and Null

2) Conversion of Unit

3) Wrong Data Type

- The Precipitation

1) Presence of NA and Null

2) Conversion of Unit

3) Wrong Data type

- The Timestamp

1) Wrong Data Type

FOR THE IRRIGATIONDATARAW TABLE

Anomalies detected

- For Sensor_id column

1) Extracting the sensor_id from the string

2) Wrong DataType

- Irrigation Method

1) Wrongly Spelt words

2) Wrong Data Type

- Water Source

1) Wrongly Spelt words

2) Presence of Null Values

3) Wrong Data Type

- Irrigation Duration

1) Presence of NA

2) Converting units to seconds

3) Wrong Data Type

- Timestamp

1) Wrong Data Type

FOR SOILDATARAW TABLE

Anomalies spotted in the table are

- Presence of NA/Null Values in the columns

- Wrong Data Types for all the columns

FOR SENSORDATARAW TABLE

Anomalies observed in the table are

1) Presence of string in the Timestamp and Sensor ID Column

2) Presence of NA/NULL Values

3) Wrong Datatype for columns

FOR LOCATIONDATARAW TABLE

Anomaly found in this table are

1) Invalid characters from sensor ID

2) Wrong Data Type for the columns

The anomalies have been effectively rectified through the application of SQL scripts. In cases where the data contained NA or Null values, our approach primarily involved substitution with the statistical mean for numerical columns and mode for categorical columns. The comprehensive set of SQL scripts utilized for this purpose is enclosed within the accompanying file for reference.

CREATION OF DIMENSIONS AND FACT TABLES

After transforming and cleaning the data appropriately, the next step was to create dimensions and facts table.

To design a database for the business using the provided tables and transform them into dimension (dim) and fact tables, we followed a typical data warehousing approach. Below, we will outline a possible schema design and provide SQL scripts for creating the tables.

Dimension Tables:

- Dim_Crop_Type (From CropData Table):

This table represents information related to crop types.

It include columns like Crop_ID (Primary Key), Crop_Type

- DIM_GROWTH_STAGE (From CropData Table):

This table represents information related to growth stage of the crops.

It may include columns like growth_stage_ID (Primary Key), growth_stage.

- DIM_PEST_TYPE (from cropdata and pestdata)

This table contains information about the various kinds of Pests and its ID.

- DIM_IRRIGATION_METHOD (From Irrigation Data)

This table contains information about the various irrigation methods and its ID.

- DIM_WATER_SOURCE (From Irrigation Data)

This table contains information about the various water sources for irrigation and its ID

- DIM_DATE (for Timestamp)

This table represents date-related information for time-based analysis.

It may include columns like DateID (Primary Key), Timestamp, Year, Month, Day, Hour, Minute, Second, DayOfWeek, DayOfYear, WeekOfYear, and other attributes.

- Dim_Weather_Condition (Weather data Table):

This table represents information related to weather conditions.

It may include columns like Weather_Condition_ID (Primary Key), Weather_Condition.

- DimSensor (Dimension Table):

This table represents information related to sensors.

- DimLocation (Dimension Table):

This table represents information related to sensor locations.

It may include columns like LocationID (Primary Key), SensorID, LocationName, Longitude, Latitude, Elevation, Region, and other attributes.

Fact Tables:

- Fact_Crop (Fact Table):

This table stores quantitative data related to crops.

It include columns like Timestamp (Foreign Key to Dim_Date), CropTypeID (Foreign Key to DimCropType), CropYield, GrowthStage, and other measures.

- Fact_Irrigation (Fact Table):

This table stores quantitative data related to irrigation.

It may include columns like SensorID (Foreign Key to DimSensor), Timestamp (Foreign Key to DimDate), IrrigationMethodID (Foreign Key to DimIrrigationMethod), WaterSource, IrrigationMin, and other measures.

- Fact_Pest (Fact Table):

This table stores quantitative data related to pests.

It may include columns like Timestamp (Foreign Key to DimDate), PestID (Foreign Key to DimPest), and other measures.

- Fact_Sensor (Fact Table):

This table stores quantitative data related to sensors.

It may include columns like SensorID (Foreign Key to DimSensor), Timestamp (Foreign Key to DimDate), Temperature, Humidity, SoilMoisture, LightIntensity, BatteryLevel, and other measures.

- Fact_Soil (Fact Table):

This table stores quantitative data related to soil.

It may include columns like Timestamp (Foreign Key to DimDate), SoilComp, SoilMoisture, SoilpH, NitrogenLevel, PhosphorousLevel, OrganicMatter, and other measures.

- Fact_Weather (Fact Table):

This table stores quantitative data related to weather.

It may include columns like Timestamp (Foreign Key to DimDate), WeatherConditionID (Foreign Key to DimWeatherCondition), WindSpeed, Precipitation, and other measures.

These tables provide a structured foundation for your data warehousing solution, allowing the data analysts and scientists to perform analytics and generate insights based on your data.

Query Optimization, Final Quality Check and Documentation

After creating the Dim and Fact tables, the next step is to optimize the queries for a better performance. One way to go about this is to use proper Index.

Indexing is not supported in snowflake so another way is to use clustered column.

A clustering key in Snowflake is important because it determines the physical storage order of data, optimizing query performance by reducing data movement during query execution and improving locality for frequently accessed data.

This step was carried out to improve query performance. Refer to the SQL script to preview.

Quality Check

Quality check was done to ensure that the data was clean and no further inconsistencies and from the check, the data seem to be in good shape.