PROJECT SUMMARY

About project.

Project objectives

- ETL Pipeline Design: Participants must design and implement an ETL pipeline to process the raw sensor data efficiently.
- Data Warehouse Setup: Create fact and dimension tables within the data warehouse to store structured data.
- Data Quality Assurance: Ensure data integrity and quality during the ETL process. Real-time Integration, consider real-time data ingestion to enable timely insights for farmers.
- Scalability, develop a scalable solution to accommodate large volumes of sensor data.

Tools used

SQL within snowflake, python and Lucidspark to design the entity relationship diagram.

Data extraction process

I extracted the raw data from the RAWDATA schema, creating staging tables in the HIT schema and copied data into it using SQL queries.

To make my cleaning process easier, I used the "Try_Cast" function when loading data into staging tables. This replaces values which cannot be parsed into the appropriate datatypes be replaced with null.

Data transformation

Null Value Check:

Null values were examined in each table to identify missing data points.

Categorical Value Standardization:

Wrongly spelled categorical values were addressed to ensure consistency within the dataset.

Mean and Median Calculation:

To handle missing values, mean and median calculations were performed for the relevant columns. Python was employed for this task since SQL lacks a built-in median method.

Imputation with SQL:

After determining that the mean and median values were closely aligned, SQL was used to fill missing values with the calculated means.

Hourly Averages:

Hourly averages were computed for specific columns, such as humidity, light intensity, and soil moisture, to impute missing data more accurately. These averages were then used as reference values to fill in null entries in the respective columns.

Spelling Correction in PEST_TYPE:

I identified and corrected spelling errors in the PEST_TYPE column. For instance, 'Slogs' was changed to 'Slugs', and 'Aphiods' was corrected to 'Aphids'.

Spelling Correction in PEST_SEVERITY:

Similarly, I corrected misspellings in the PEST_SEVERITY column. For example, 'Hihg' was changed to 'High'.

Description Spelling Corrections:

I reviewed and corrected the descriptions in the PEST_DESCRIPTION column for spelling errors. For instance, 'infestation deteced' was changed to 'infestation detected', and 'high infestation riskkk' was corrected to 'high infestation risk'.

Spelling Correction in WEATHER_CONDITION:

In this step, I focused on the WEATHER_CONDITION column. Spelling errors, such as 'Claar' (corrected to 'Clear') and 'Party Cloudy' (corrected to 'Partly Cloudy'), were fixed to ensure consistent weather condition data.

Handling Null Values:

When the TIMESTAMP column had null values, I checked if any other weather-related columns had non-null values. If they did not, meaning all other rows were null, I deleted those rows.

Handling Missing Values in WIND_SPEED and PRECIPITATION:

I used Python to compare the mean and median values for the WIND_SPEED and PRECIPITATION columns. After confirming that they were similar, I used SQL UPDATE statements to fill null values in these columns with the calculated means. This imputation enhanced data completeness.

These data cleaning steps improved the quality and consistency of both tables, ensuring that misspelled values were corrected, null values were appropriately handled, and missing data was imputed with relevant statistics, resulting in cleaner and more accurate datasets for analysis.

Creating dimension and fact tables also assigning primary & foreign keys and clustering or indexing columns

Location Dimension Table:

I created a table to store location-related data, like names, latitude, longitude, and region.

I used the location name as the unique identifier (primary key).

I filled this table with data from my existing LOCATION_DATA_TABLE.

Crop Dimension Table:

I made a table for crop information, including types, yields, and growth stages.

Each crop type became a unique identifier (primary key).

I loaded this table with data from the CROP_DATA_TABLE.

Pest Dimension Table:

I set up a table for pest details, such as types and descriptions.

I gave each pest type a unique ID (primary key) and loaded data from my PEST_DATA_TABLE.

Sensor Dimension Table:

I created a table to hold sensor data, including readings like temperature, humidity, and battery level.

Each sensor's ID was used as its unique identifier (primary key).

I filled this table with data from the SENSOR_DATA_TABLE.

Irrigation Dimension Table:

I established a table to manage irrigation data, including methods and duration.

The combination of timestamp and sensor ID became a unique identifier (primary key).

I populated this table with data from my IRRIGATION_DATA_TABLE.

Soil Dimension Table:

I made a table to handle soil-related information, like composition and pH levels.

The timestamp served as the unique identifier (primary key).

I loaded this table with data from the SOIL_DATA_TABLE.

Time Dimension Table:

I created a table for timestamps, with details like years, months, and weekdays.

The timestamp was the unique identifier (primary key).

I put data from my existing SENSOR_DATA_TABLE into this table.

Weather Dimension Table:

I set up a table for weather data, covering conditions, wind speed, and precipitation.

The timestamp was the unique identifier (primary key), and it linked to the time dimension table.

I filled this table with data from the WEATHER_DATA_TABLE.

Sensor Fact Table:

I created a table for sensor-related facts.

This table contains sensor IDs and timestamps.

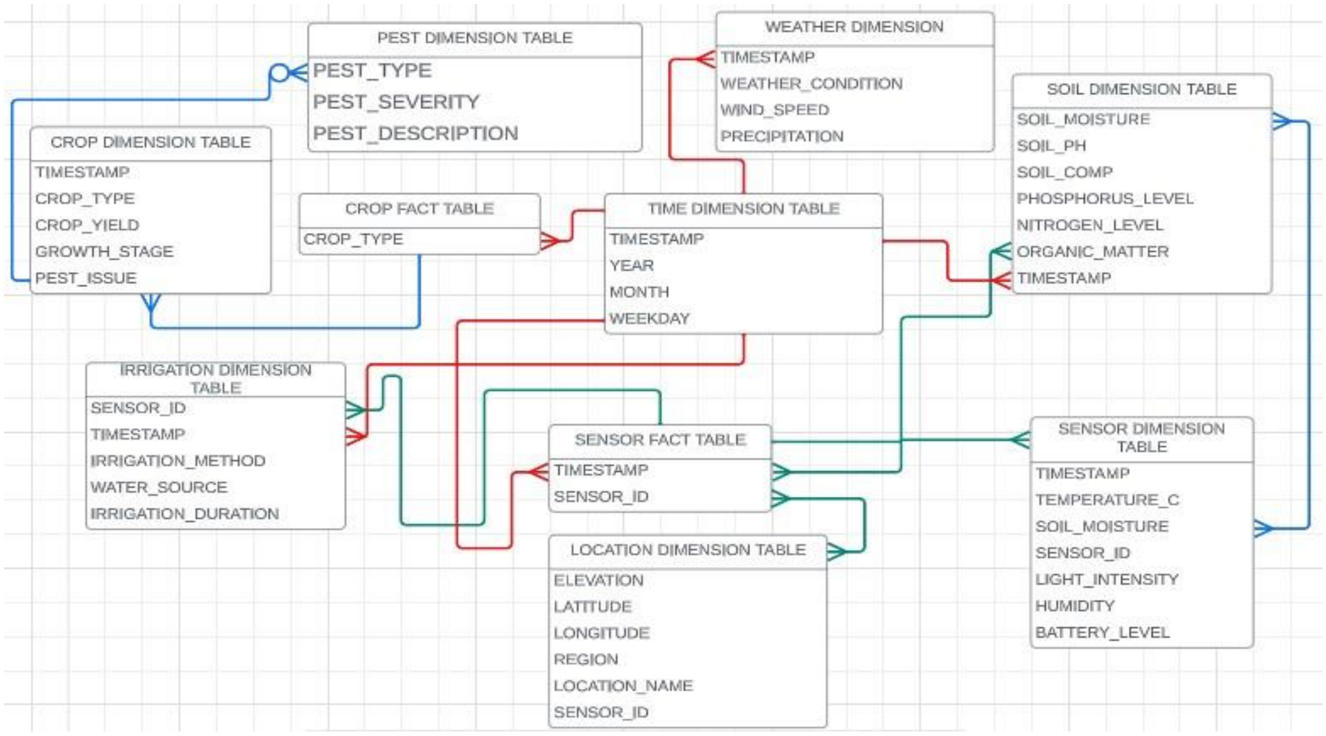
Crop Fact Table:

Another fact table was made for crop-related data.

It holds crop types.

These processes organized and structured the data, ensuring it is easily accessible and useful for analysis. Essentially creating different folders to neatly store and categorize your files for better understanding and insights.

ENTITY RELATIONSHIP DIAGRAM


PEST DIMENSION TABLE
PEST_TYPE
PEST_SEVERITY
PEST_DESCRIPTION
WEATHER DIMENSION
TIMESTAMP
WEATHER_CONDITION
WIND_SPEED
PRECIPITATION
SOIL DIMENSION TABLE
SOIL_MOISTURE
SOIL_PH
SOIL_COMP
PHOSPHORUS_LEVEL
NITROGEN_LEVEL
ORGANIC_MATTER
TIMESTAMP
CROP DIMENSION TABLE
TIMESTAMP
CROP_TYPE
CROP_YIELD
GROWTH_STAGE
PEST_ISSUE
CROP FACT TABLE
CROP_TYPE
TIME DIMENSION TABLE
TIMESTAMP
YEAR
MONTH
WEEKDAY
IRRIGATION DIMENSION TABLE
SENSOR_ID
TIMESTAMP
IRRIGATION_METHOD
WATER_SOURCE
IRRIGATION_DURATION
SENSOR FACT TABLE
TIMESTAMP
SENSOR_ID
SENSOR DIMENSION TABLE
TIMESTAMP
TEMPERATURE_C
SOIL_MOISTURE
SENSOR_ID
LIGHT_INTENSITY
HUMIDITY
BATTERY_LEVEL
LOCATION DIMENSION TABLE
ELEVATION
LATITUDE
LONGITUDE
REGION
LOCATION_NAME
SENSOR_ID