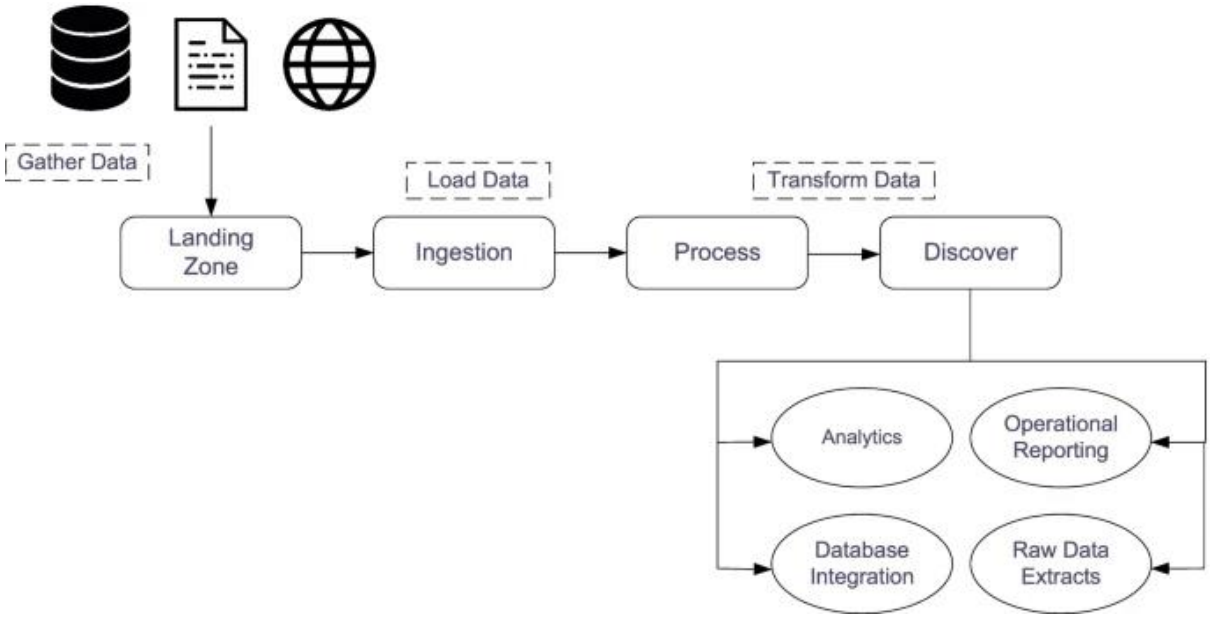
Big Data Analysis with IBM Cloud Database

Phase-3 Development Part 1

Project 5 : Big Data Analysis



Problem Statement:

Dive into the world of big data analysis with IBM Cloud Databases. Uncover hidden insights from vast datasets, from climate trends to social patterns. Visualize your findings and derive valuable business intelligence. Embark on data-driven adventures, exploring the endless possibilities of big data!

To begin building a big data analysis solution using IBM Cloud Databases, follow these steps:

Create an IBM Cloud account. You can create a free account to get started.

- Choose the appropriate database service. IBM Cloud offers a variety of database services, including Db2 and MongoDB. Choose the service that is best suited for your needs.

- Set up a database instance. Once you have chosen a database service, you need to set up a database instance. This will involve choosing a region and a plan.

- Develop queries or scripts to explore and analyze the selected dataset. Once you have set up a database instance, you can start to develop queries or scripts to explore and analyze the selected dataset. You can use the database console to develop and execute queries and scripts.

- Perform basic data cleaning and transformation as needed. Before you can analyze your data, you may need to perform some basic data cleaning and transformation. This may involve removing duplicate records, correcting errors, and transforming the data into a format that is compatible with your chosen analysis tools.

IBM DB2:

Description: IBM Db2 is a family of data management products, including database servers, developed by IBM.

Role in the Project: Used for storing structured data, providing a reliable and scalable database solution

To set up a database instance after choosing the database, you need to follow these steps:

- **Create a database instance.** This can be done using the database management tool that you are using. For example, to create a database instance in Db2, you would use the CREATE DB command.
- **Configure the database instance.** This includes setting things like the database name, the database user accounts, and the database parameters.
- **Start the database instance.** This can be done using the database management tool that you are using. For example, to start a database instance in Db2, you would use the START DB command.

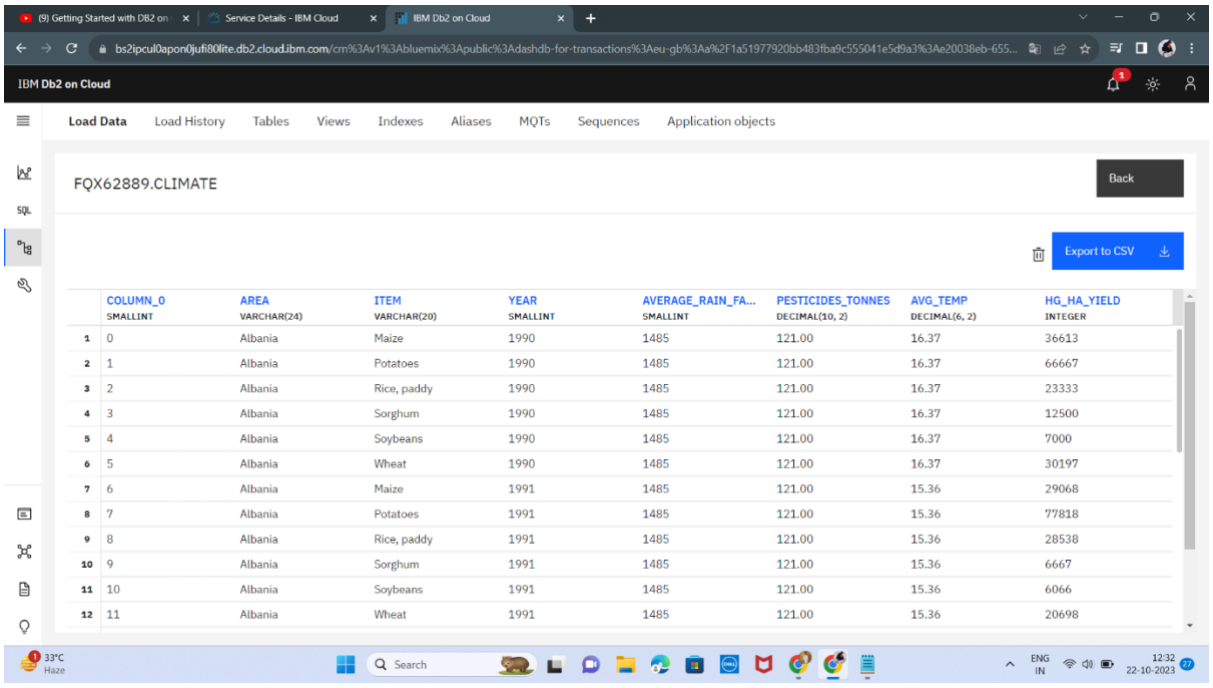
Once the database instance is created, configured, and started, you can start using it to store and manage your data.

Loading the dataset:

The first step is to load the dataset into IBM Cloud Databases.

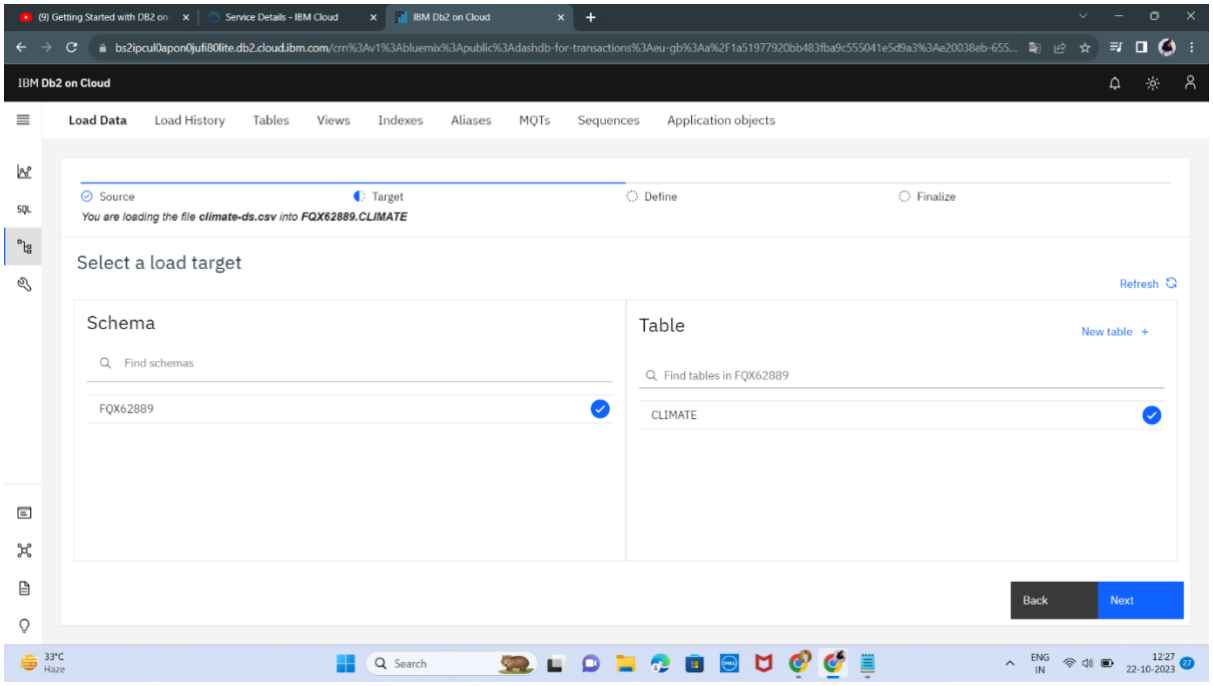
To load the dataset into IBM cloud DB2, we can use the following steps:

By using DB2 database, load the Dataset into IBM cloud database

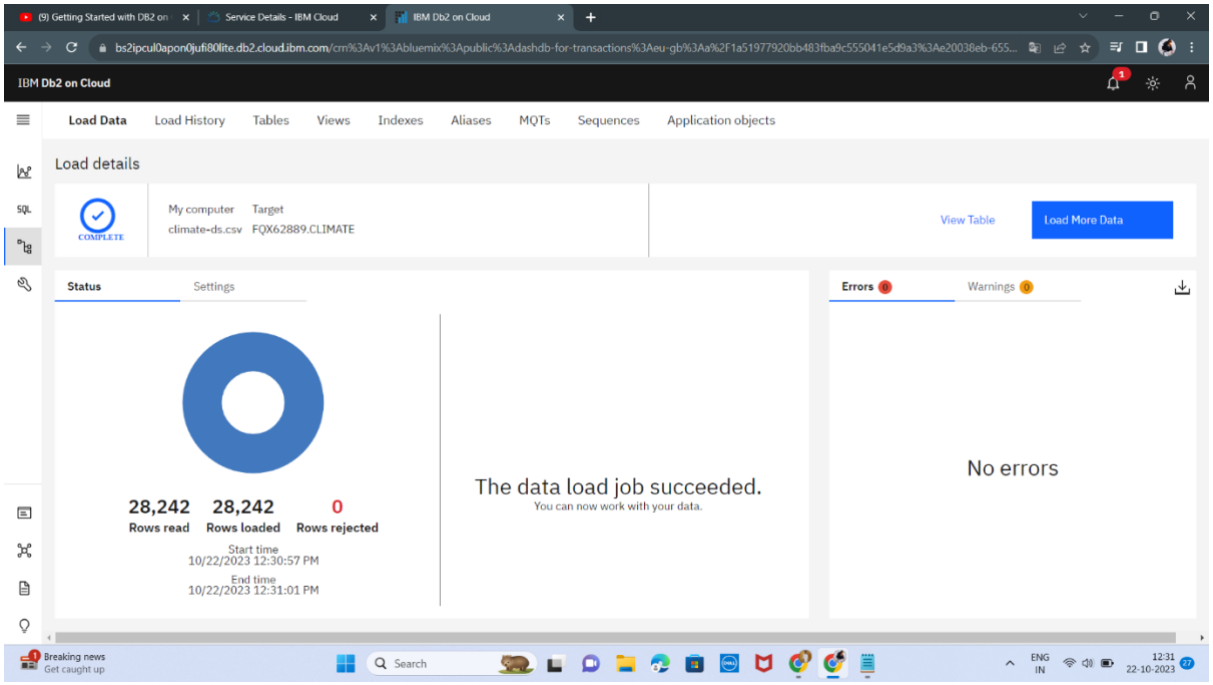


	COLUMN_0 SMALLINT	AREA VARCHAR(24)	ITEM VARCHAR(20)	YEAR SMALLINT	AVERAGE_RAIN_FA... SMALLINT	PESTICIDES_TONNES DECIMAL(10, 2)	AVG_TEMP DECIMAL(6, 2)	HG_HA_YIELD INTEGER
1	0	Albania	Maize	1990	1485	121.00	16.37	36613
2	1	Albania	Potatoes	1990	1485	121.00	16.37	66667
3	2	Albania	Rice, paddy	1990	1485	121.00	16.37	23333
4	3	Albania	Sorghum	1990	1485	121.00	16.37	12500
5	4	Albania	Soybeans	1990	1485	121.00	16.37	7000
6	5	Albania	Wheat	1990	1485	121.00	16.37	30197
7	6	Albania	Maize	1991	1485	121.00	15.36	29068
8	7	Albania	Potatoes	1991	1485	121.00	15.36	77818
9	8	Albania	Rice, paddy	1991	1485	121.00	15.36	28538
10	9	Albania	Sorghum	1991	1485	121.00	15.36	6667
11	10	Albania	Soybeans	1991	1485	121.00	15.36	6066
12	11	Albania	Wheat	1991	1485	121.00	15.36	20698

Before loading the dataset to the database we want to create the TABLE NAME ,here we have created the table name as “**Climate**”



By clicking the next icon the dataset will start to load.



The Dataset is successfully loaded into the database without any Rejection.

To develop a simple query to explore and analyze a dataset using the IBM Cloud Db2 console:

- Log in to the IBM Cloud console and navigate to the Db2 service.
- Click on the database instance that you want to use.
- Click on the SQL tab.

In the SQL editor, enter the following query:

-- Perform basic data cleaning

-- Drop any rows with missing values.

DELETE FROM climate

WHERE column_0 IS NULL OR area IS NULL OR item IS NULL OR year IS NULL OR average_rain_fall_mm_per_year IS NULL OR pesticides_tonnes IS NULL OR avg_temp IS NULL OR hg_ha_yield IS NULL;

-- Convert all of the columns to numeric values.

ALTER TABLE climate

ALTER COLUMN column_0 SET DATA TYPE DECIMAL(10,2);

ALTER TABLE climate

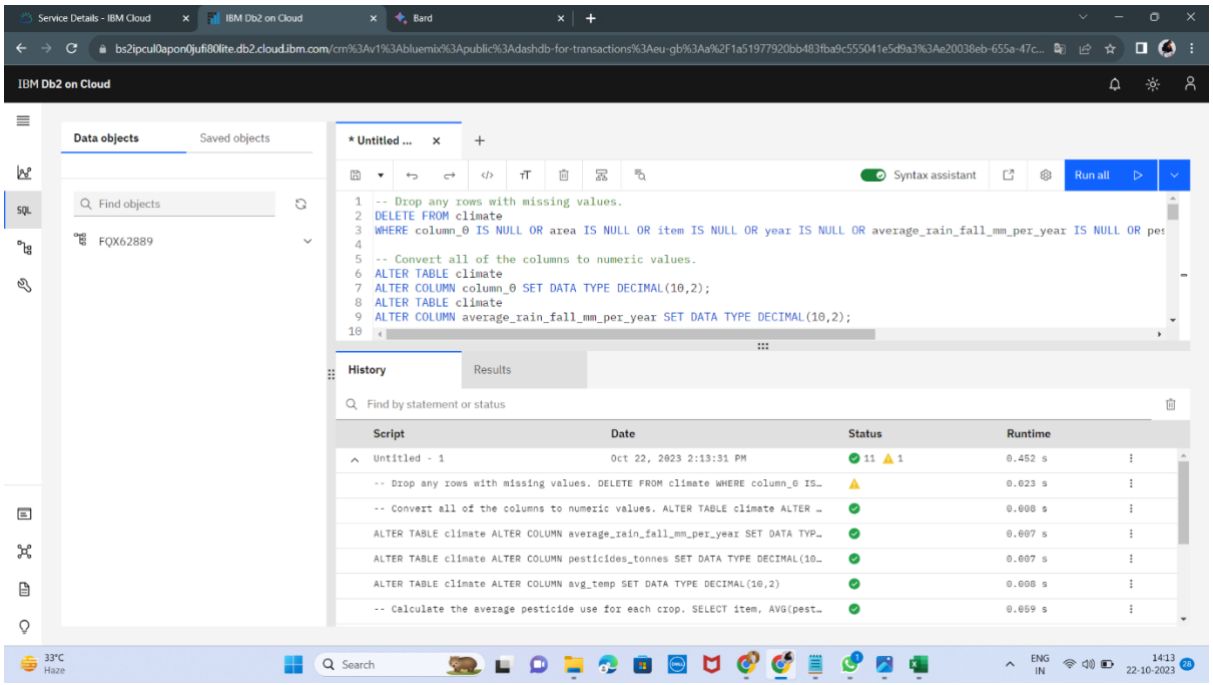
ALTER COLUMN average_rain_fall_mm_per_year SET DATA TYPE DECIMAL(10,2);

ALTER TABLE climate

ALTER COLUMN pesticides_tonnes SET DATA TYPE DECIMAL(10,2);

ALTER TABLE climate

ALTER COLUMN avg_temp SET DATA TYPE DECIMAL(10,2);



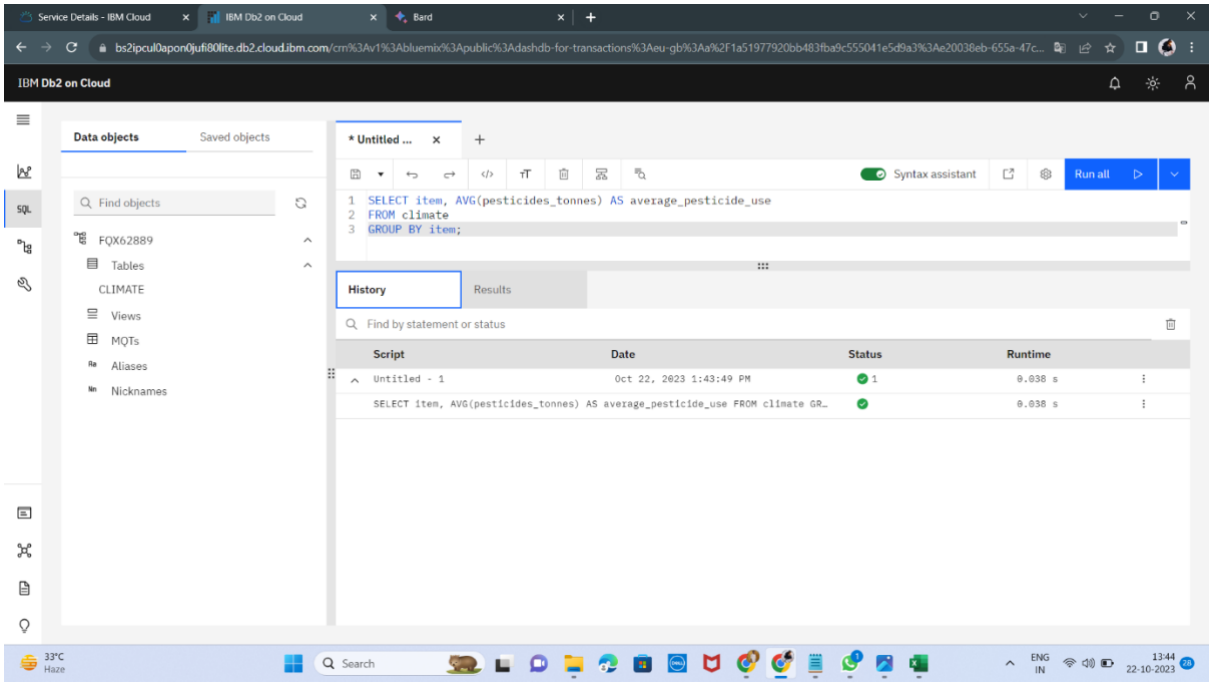
-- Perform basic data transformation

-- Calculate the average pesticide use for each crop.

SELECT item, AVG(pesticides_tonnes) AS average_pesticide_use

FROM climate

GROUP BY item;

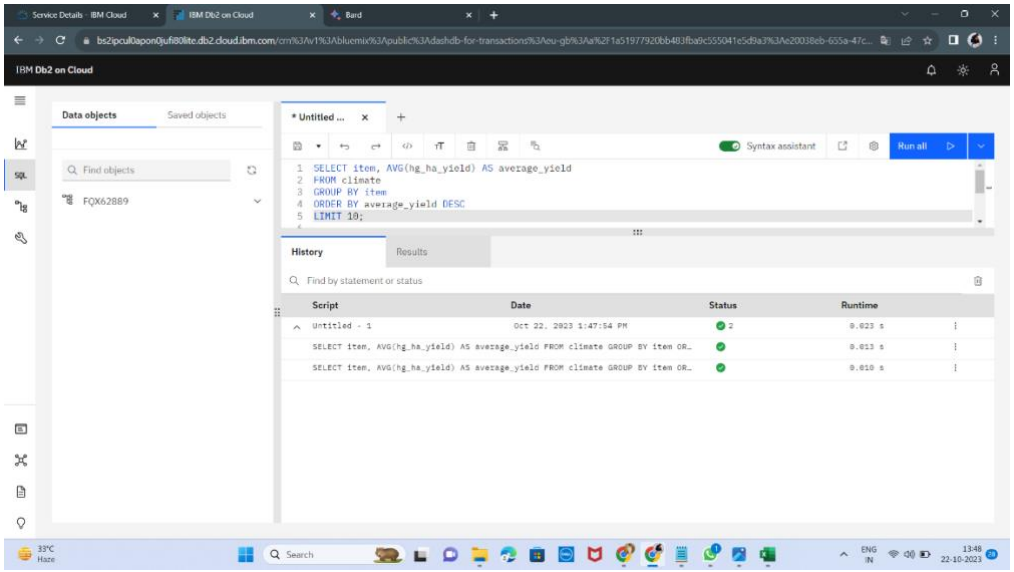


-- Identify the crops with the highest and lowest average yields.

SELECT item, AVG(hg_ha_yield) AS average_yield From climate GROUP BY item ORDER BY average_yield DESC LIMIT 10;

SELECT item, AVG(hg_ha_yield) AS average_yield FROM climate GROUP BY item ORDER BY average_yield

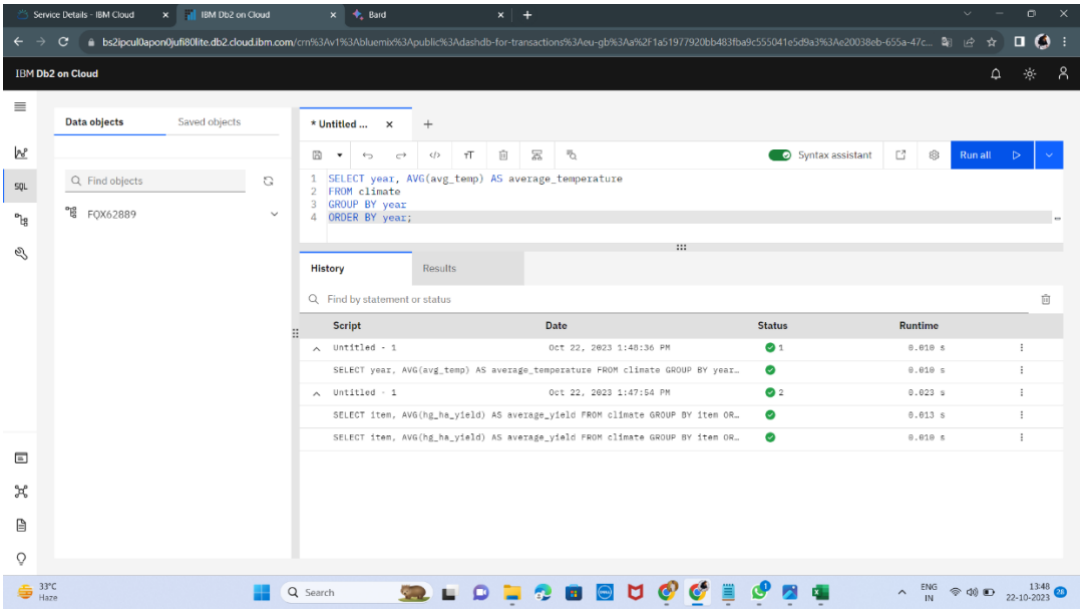
ASCLIMIT 10;



-- Calculate the trend in average temperature over time.

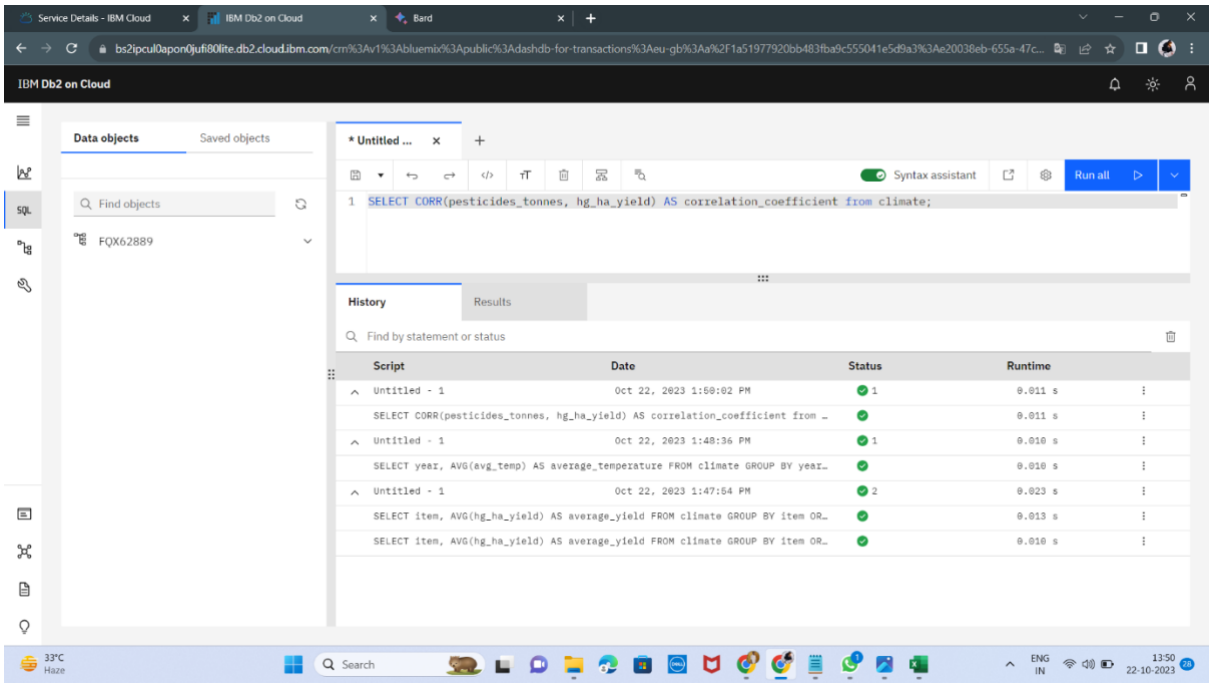
SELECT year, AVG(avg_temp) AS average_temperature

FROM climate GROUP BY year ORDER BY year;



-- Calculate the correlation between pesticide use and average yield.

SELECT CORR(pesticides_tonnes, hg_ha_yield) AS correlation_coefficient from climate;



-- Identify the crops that are most and least sensitive to changes in pesticide use.

Calculate the percentage change in average yield for each crop for every 10% increase in pesticide use

delta_yield = (hg_ha_yield - LAG(hg_ha_yield, 1) OVER (PARTITION BY item ORDER BY year)) / LAG(hg_ha_yield, 1) OVER (PARTITION BY item ORDER BY year) * 100;

Calculate the average percentage change in yield for each crop

average_delta_yield = delta_yield.groupby('item').mean();

Sort the crops by average percentage change in yield

average_delta_yield.sort_values(ascending=False, inplace=True);

Print the crops that are most and least sensitive to changes in pesticide use

print('Crops that are most sensitive to changes in pesticide use:')

print(average_delta_yield.head(10))

print('Crops that are least sensitive to changes in pesticide use:')

print(average_delta_yield.tail(10))

SELECT item, AVG(delta_yield) AS average_delta_yieldFROM (SELECT item, (hg_ha_yield - LAG(hg_ha_yield, 1) OVER (PARTITION BY item ORDER BY year)) / LAG(hg_ha_yield, 1)

OVER (PARTITION BY item ORDER BY year) * 100 AS delta_yield FROM climate) AS delta_yield_table

GROUP BY item ORDER BY average_delta_yield DESC

LIMIT 10;

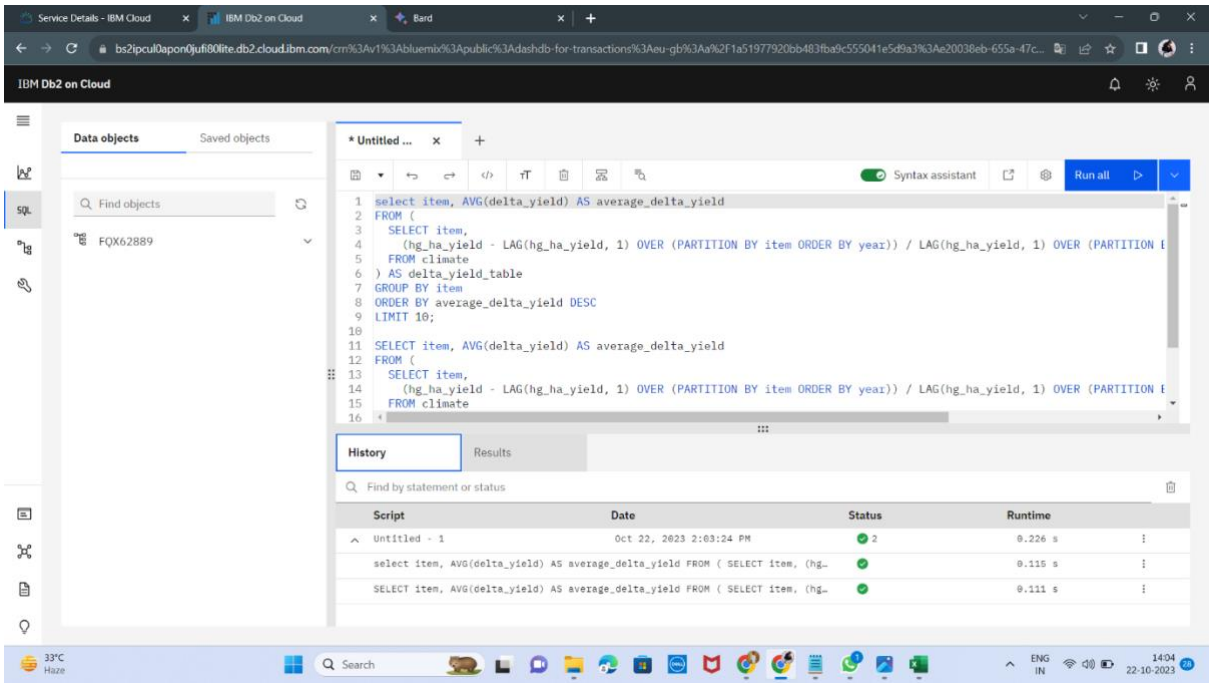
SELECT item, AVG(delta_yield) AS average_delta_yield

FROM (SELECT item,(hg_ha_yield - LAG(hg_ha_yield, 1) OVER (PARTITION BY item ORDER BY year)) / LAG(hg_ha_yield, 1) OVER (PARTITION BY item ORDER BY year) * 100 AS delta_yield

FROM climate) AS delta_yield_table

GROUP BY item ORDER BY average_delta_yield ASC

LIMIT 10;



We can use this query to identify the crops that are most and least sensitive to changes in pesticide use. This information can be used to develop strategies to reduce pesticide use and improve crop yields.

Team Leader name :

Praveen Kumar A (Reg no :111421104091)

Team Members :

Vijayakumar C (Reg no :111421104118)

Praveen joel (Reg no :111421104090)

Yashwanth Kumar S (Reg no :111421104123)

Nighil Ananth V (Reg no :111421104072)