CIS 4560 Term Project Tutorial

Authors: Kathlyne Alilain, Jennifer Mazas

Instructor: Jongwook Woo

Date: 05/05/2025

Lab Tutorial

Kathlyne Alilain (kalilai2@calstatela.edu)

Jennifer Mazas(jmazas@calstatela.edu)

05/05/2025

Evolution of Europe's Climate Monitoring Infrastructure

Objectives

In this hands-on lab, you will learn how to:

- Download and load the stations_info_tx_v31.0e.txt dataset (4.6 GB).
- Create Hive tables and perform queries using Hadoop Cluster.
- Conduct tempo-spatial analysis of maximum temperature patterns.
- Visualize regional trends in Excel.
- Understand the application of Big Data tools in climate research.

Platform Specs

- IBM Bluemix BigInsights
- CPU Speed: 2.4 GHz
- # of CPU cores: 4 cores per node
- # of nodes: 3 nodes
- Total Memory Size: 24 GB (8 GB per node x 3 nodes)

Step 1: Load Data to Hadoop Cluster

Log in to the remote cluster and download the data file:

```
$ ssh your_user@144.24.46.199
```

```
$ your_user@144.24.46.199 enter your password:
```

```
-bash-4.2$ wget -O stations_info_tx_v31.0e.txt https://knmi-ecad-assets-prd.s3.amazonaws.com/ensembles/data/stations_info_tx_v31.0e.txt
```

```
-bash-4.2$ wget -O stations_info_tx_v31.0e.txt https://knmi-ecad-assets-prd.s3.amazonaws.com/ensembles/data/stations_info_tx_v31.0e.txt
--2025-04-07 02:41:46--  https://knmi-ecad-assets-prd.s3.amazonaws.com/ensembles/data/stations_info_tx_v31.0e.txt
Resolving knmi-ecad-assets-prd.s3.amazonaws.com (knmi-ecad-assets-prd.s3.amazonaws.com)... 52.218.62.10, 52.218.40.26, 52.92.2.81, ...
Connecting to knmi-ecad-assets-prd.s3.amazonaws.com (knmi-ecad-assets-prd.s3.amazonaws.com)|52.218.62.10|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 433591 (423K) [text/plain]
Saving to: 'stations_info_tx_v31.0e.txt'

100%[======================================================>] 433,591      610KB/s   in 0.7s

2025-04-07 02:41:47 (610 KB/s) - 'stations_info_tx_v31.0e.txt' saved [433591/433591]
```

Upload the file to the Stations directory of HDFS:

```
hdfs dfs -mkdir Stations
```

```
hdfs dfs -mkdir Stations/stations_info/
```

```
-bash-4.2$ hdfs dfs -mkdir Stations
-bash-4.2$ hdfs dfs -mkdir Stations/stations_info/
```

```
hdfs dfs -put stations_info_tx_v31.0e.txt Stations/stations_info/
```

```
hdfs dfs -ls Stations/stations_info/
```

```
-bash-4.2$ hdfs dfs -put stations_info_tx_v31.0e.txt Stations/stations_info/
-bash-4.2$ hdfs dfs -ls Stations/stations_info/
Found 1 items
-rw-r--r--   3 kalilai2 hdfs     433591 2025-04-29 02:56 Stations/stations_info/stations_info_tx_v31.0e.txt
-bash-4.2$
```

Step 2: Create Hive Table

Start Beeline CLI and switch to your assigned database.

```
$ beeline
```

```
use [your_database];
```

If you do not have a database yet, create one:

```
CREATE DATABASE your_database_name;
```

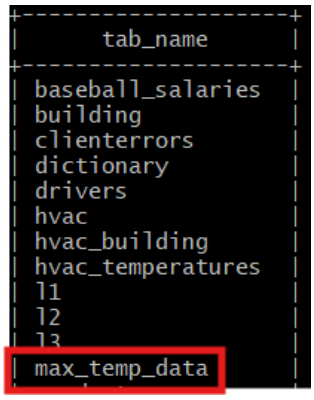
Create the external Hive table by running these commands:

```
DROP TABLE IF EXISTS max_temp_data;

CREATE EXTERNAL TABLE IF NOT EXISTS max_temp_data(station_id STRING,
      station_name STRING,
      country STRING,
      latitude DOUBLE,
      longitude DOUBLE,
      elevation DOUBLE,
      start_date STRING,
      end_date STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
STORED AS TEXTFILE LOCATION '/user/your_user/Stations/stations_info/'
TBLPROPERTIES ("skip.header.line.count"="1");
```

After creating the table, check if the table is created and confirm the structure of the columns using these commands:

```
SHOW tables;
```




```
DESCRIBE max_temp_data;
```

col_name	data_type	comment
station_id	string	
station_name	string	
country	string	
latitude	double	
longitude	double	
elevation	double	
start_date	string	
end_date	string	

To make sure the data is loaded correctly, run:

```
SELECT * FROM max_temp_data LIMIT 5;
```

max_temp_data.station_id	max_temp_data.station_name	max_temp_data.country	max_temp_data.latitude	max_temp_data.longitude	max_temp_data.elevation	max_temp_data.start_date	max_temp_data.end_date
1	Vaexjoe	SWEDEN	56.87	14.8	166.0	1950-01-01	2006-12-31
2	Falun	SWEDEN	60.62	15.62	160.0	1950-01-01	2024-12-31
3	Stensele	SWEDEN	65.07	17.15	325.0	1950-01-01	2004-12-31
4	Linkoeping	SWEDEN	58.4	15.53	93.0	1950-01-01	2024-12-31
5	Linkoeping-Malmslaett	SWEDEN	58.4	15.53	93.0	1950-01-01	2024-12-31

Step 3: Run Hive Queries

Run these queries for analysis:

Find out which countries had the highest average elevations of weather stations:

```
SELECT country, AVG(elevation) AS avg_elevation
FROM max_temp_data
GROUP BY country
ORDER BY avg_elevation DESC
LIMIT 10;
```

country	avg_elevation
ARMENIA	1955.075
KYRGYZSTAN	1398.5
IRAN, ISLAMIC REPUBLIC OF	1355.142857142857
BOSNIA AND HERZEGOVINA	1348.5
SWITZERLAND	1018.972972972973
TAJIKISTAN	934.0
NORTH MACEDONIA	779.5
AUSTRIA	758.0833333333334
SAUDI ARABIA	689.0
TÜRKIYE	568.6090909090908

10 rows selected (8.929 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

Which countries have the most stations (more spatial coverage)?

```
SELECT country, COUNT(*) AS num_stations

FROM max_temp_data

GROUP BY country

ORDER BY num_stations DESC

LIMIT 5;
```

country	num_stations
GERMANY	1013
RUSSIAN FEDERATION	318
SWEDEN	275
ITALY	204
NORWAY	183

5 rows selected (8.858 seconds)
0: jdbc:hive2://bigdaiun0.sub03291929060.trai>

What are the top 5 stations at the highest elevations?

```
SELECT station_id, country, elevation

FROM max_temp_data

ORDER BY elevation DESC

LIMIT 5;
```

station_id	country	elevation
2073	ITALY	3480.0
2941	ARMENIA	3223.0
15	AUSTRIA	3109.0
58	GERMANY	2964.0
878	ITALY	2600.0

Step 4: Downloading Data to Your Personal Computer

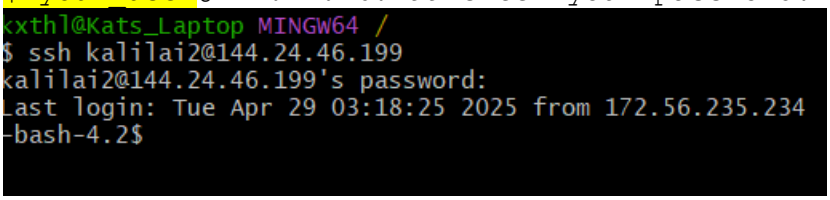
Run an INSERT OVERWRITE DIRECTORY query to save the table's content into a file in HDFS:

```
INSERT OVERWRITE DIRECTORY '/user/your_user/max_temp_export/'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
SELECT * FROM max_temp_data;
```

This will export all the data from your max_temp_data table into a new HDFS directory /user/your_user/max_temp_export/.

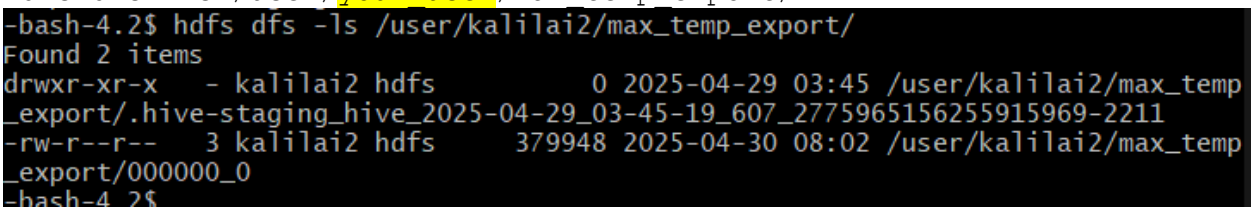
Open another terminal to download the output file at the HDFS path.

```
$ ssh your_user@144.24.46.199
$ your_user@144.24.46.199 enter your password:
```



Locate the file.

```
hdfs dfs -ls /user/your_user/max_temp_export/
```

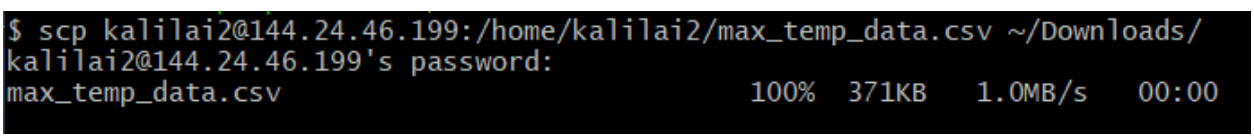


Export manually.

```
hdfs dfs -copyToLocal /user/your_user/max_temp_export/000000_0
max_temp_data.csv
```

Then transfer to local machine:

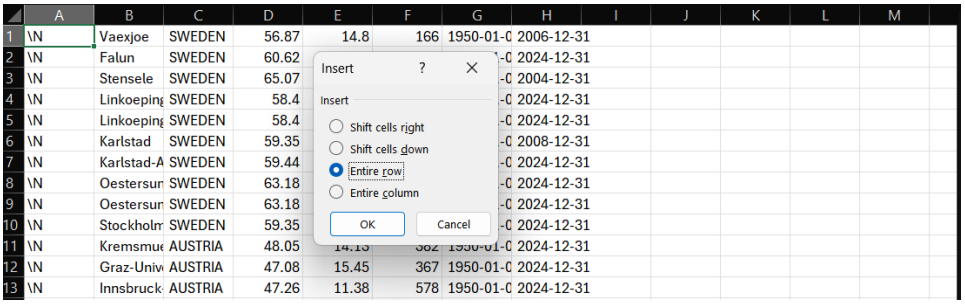
```
$ scp
your_user@your_ip_address:/home/your_local_machine/max_temp_data.csv
~/Downloads/
```



Step 5: Visualize Data

Open the downloaded file through Excel.

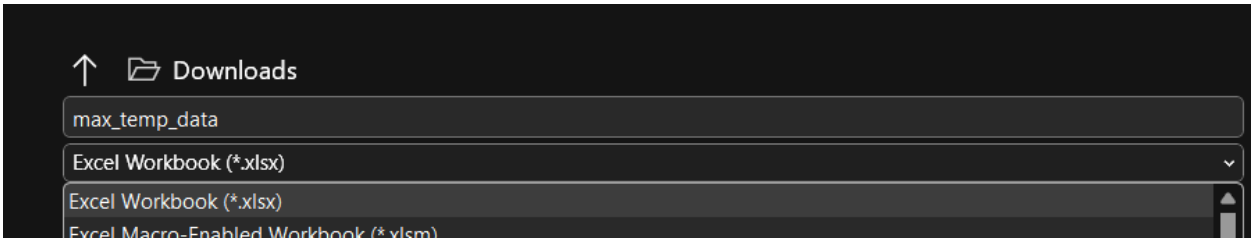
Right click on the first row and insert a new row above to create column headers.



Insert column headers: station_id, station_name, country, latitude, longitude, elevation, start_date, end_date

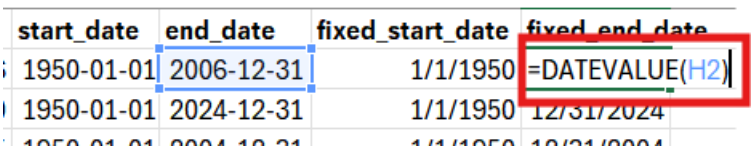
	A	B	C	D	E	F	G	H
1	station_id	station_name	country	latitude	longitude	elevation	start_date	end_date
2	\N	Vaexjoe	SWEDEN	56.87	14.8	166	1950-01-01	2006-12-31
3	\N	Falun	SWEDEN	60.62	15.62	160	1950-01-01	2024-12-31

Go to File > Save As, and select file type as 'Excel Workbook (*.xlsx).



Create two new columns called 'fixed_start_date' and 'fixed_end_date'. Use this formula, =DATEVALUE(cell), and select the cell containing the corresponding dates.

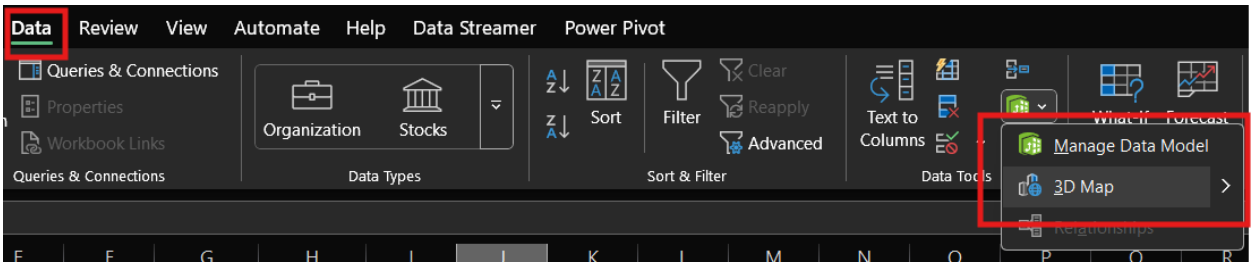
Drag the corner of the formulated cell to copy the formula to the rest of the column.



start_date	end_date	fixed_start_date	fixed_end_date
1950-01-01	2006-12-31	1/1/1950	12/31/2006
1950-01-01	2024-12-31	1/1/1950	12/31/2024
1950-01-01	2004-12-31	1/1/1950	12/31/2004
1950-01-01	2024-12-31	1/1/1950	12/31/2024
1950-01-01	2024-12-31	1/1/1950	12/31/2024

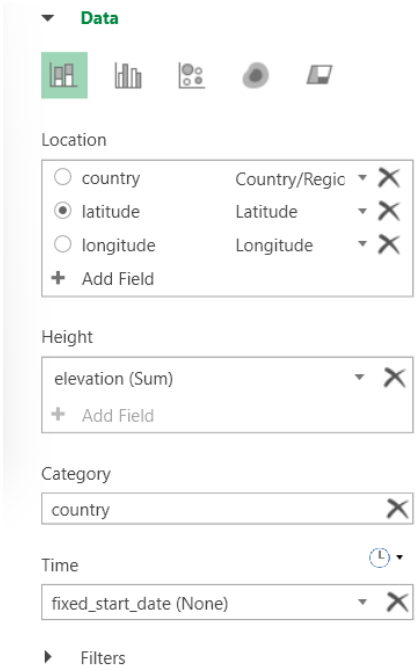
NOTE: You may also need to review the data to make sure everything is formatted correctly and fix any errors.

Next, go to the 'Data' tab, then 'Data Model' > 3D Map.



On the 3D Map screen, in the Data pane, select 'latitude' as the Location.

Drag 'elevation' to Height, 'country' to Category, and 'fixed_start_Date' to Time.



It should generate this map with the ability to see changes over time.

References

1. URL of Data Source,

   https://surfobs.climate.copernicus.eu/dataaccess/access_eobs.php#datafiles

2. URL of your Github, https://github.com/kalilai2/CIS-4560-01-Group-Project.git