KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện Thoại: (08) 38.354.266 - Fax:(08) 38.350.096

CSC12107 – INFORMATION SYSTEM FOR BUSINESS INTELLIGENCE
PROJECT
BUILDING AND MINING DATA WAREHOUSE

1. General Information

Assignment ID	PROJECT
Estimated duration:	10 weeks
Submission deadline:	22/12/2024
Assignment type:	Student Group
Submission chanel:	Moodle
Teachers:	Hồ Thị Hoàng Vy, Tiết Gia Hồng, Nguyễn Ngọc Minh Châu
Contacts:	hthvy@fit.hcmus.edu.vn, tghong@fit.hcmus.edu.vn nnmchau@fit.hcmus.edu.vn

2. Learning outcomes

This assignment is to gain the following outcomes:

- G3.3 Design a Star or Snowflake data model diagram through the Multidimensional Design from analytical business requirements and OLTP system
- G5.1 Deploy the ETL procedure to extracting data from disparate databases and data sources, and then transforming the data for effective integration into a data warehouse using SSIS tool
- G5.2 Operate the basic OLAP technologies using SSAS tools.
- G5.3 Create a dashboard and other visualizations to analyze and communicate the data from DW using SSRS or excel...
- G5.4 Applying the data mining algorithms in Analysis Services to your data.

3. Requirements and submission rules

The objective of this project is to create a data warehouse utilizing air quality data from the Environmental Protection Agency's (EPA) Daily Summary AQI by County across 10 States. The aim is to analyze this data warehouse in order to identify trends and patterns in U.S. air quality over 3 years (2021-2023)

3.1 Data description

Describe meaning of the properties of the following data sources (only describe the properties necessary for the project):

- US Daily AQI report by County(2021-2023):
  - Data: **3 .csv files under "Air_Quality_Data" folder**
  - Description: AirData Download Files Documentation (epa.gov)
  - Data Source: Download Files | AirData | US EPA
- Geography data: (2B) uscounties.csv
- AQI Category Definition: Table 1.

AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Table 1. AQI basics. Ref: AQI Basics | AirNow.gov

3.2 Design data warehouse (DW), synthesize, load data from the sources into DW, then design and build Cube

Suggestions:

- Map the above data sources to get the values for building Geography dimension with dimensional hierarchy as follows: State > County
- Transform the datetime data to create the Date dimension with dimensional hierarchv: Year > Quarter > Month > Dav

- Define and design other dimensional hierarchies to meet OLAP and Report requirements

3.3 OLAP and Report

Analytical requirements:

- The report should include a visual representation of the data and an analysis based on the results of the questionnaire.
- The analysis requirements are open-ended questions. It is therefore necessary to answer these questions based on both the data and the domain knowledge available, i.e. your understanding about the data and real world.
- The analysis should be **concise** and to the point, avoiding lengthy and verbose writing.
- Tips: You could present your analysis based on the hints in each question.
- Please do not hesitate to cite your references used in the report.

Data analysis questions:

1. Report the **min and max** of AQI value for each **State** during **each quarter of years**. *Analysis hints:* How do the AQI values fluctuate during the year? Pay attention to the values (max, min). Are any unusually large or small?
2. Report the **mean and the standard deviation** of AQI value for each **State** during **each quarter of years**. *Analysis hints:* How do the AQI values fluctuate during the year? Pay attention to the values (mean, std, max, min). Are any unusually large or small?
3. Report the number of days, and the mean AQI value where the air quality is rated as "very unhealthy" or worse for each State and County. *Analysis hint:* What is the AQI limit above which air quality is "very unhealthy" or worse?
4. For the four following states: Hawaii, Alaska, Illinois and Delaware, **count the number of days** in each air quality **Category** (Good, Moderate,etc.) by **County**. *Analysis hints:* Comparing the data of the states and counties, focus on the distribution of the harmful air condition. What could you conclude about the differences?)
5. For the four following states: Hawaii, Alaska, Illinois and Delaware, compute the **mean AQI** value by **quarters**. *Analysis hints:* Comparing the data of the states over the year. What could you conclude about the fluctuations?
6. Design a report to demonstrate the AQI fluctuation trends over the year for the four following states: Hawaii, Alaska, Illinois and California. *Analysis hint:* Give your opinion about the fluctuations of AQI value.
7. Build graphs/charts for the above reports.
8. Use a regional map to visually represent (by color) the mean AQI value in regions during a year. *Example:*

US mean AQI of four states: Alaska, Delaware, Hawaii, Illinois over the year 2023

Month	Alaska	Delaware	Hawaii	Illinois
2023-01	40.339	43.151	38.581	44.517
2023-02	28.032	47.893	34.405	40.057
2023-03	29.077	46.570	29.600	45.179
2023-04	24.994	53.278	25.500	48.929
2023-05	25.632	50.699	28.364	65.065
2023-06	20.050	82.956	24.435	90.900
2023-07	22.762	56.355	24.462	55.857
2023-08	30.117	53.000	26.544	50.501
2023-09	17.956	47.822	25.256	48.688
2023-10	29.303	41.032	18.419	40.935
2023-11	30.620	46.012	24.156	41.243
2023-12	36.558	43.284	23.593	35.737

Question for bonus points:

9. Report the **mean, the standard deviation, min and max** of AQI value group by **State** and **County** during **each quarter of the year**. *Analysis hints:* Pay attention to the values (mean, std, max, min). Are any unusually large or small? Compare the standard deviation values between question 1 and 2, explain.

10. Create a new attribute, **DayLightSaving**, in a suitable table. **DayLightSaving** may have two values:

    True: Between March 12, 2023, and November 5, 2023

    False: Otherwise

    Report the mean AQI value by State, Category, DayLightSaving over years. *Analysis hint:* Is there any notable difference on the air quality during the Daylight Saving period compared to the other?

11. Count the number of days by State, Category in each month.

    Be caution: The Category in the data set is calculated for each County, not State.

12. Report the number of days by Category and Defining Parameter. *Analysis hints*: What is your opinion on the pollution situation in the United States as a whole? Additionally, please identify the primary factors that the country should consider in order to enhance air quality

3.4 Data Mining:

- Suggestion: Using models to predict the air quality in the next periods such as next quarter (Q1-2024), next month (01-2024), etc.

- Students propose applications of any case, explain the algorithm used, why, how the results are, etc.

3.5 Conclusion:

A brief summary of the group's project outcomes including the following elements:

- Based on the data and reports, please give an overview of air quality in US counties in 2023 and explain the aforementioned arguments.
- This section should also include a summary of the project's achievements and suggestions for potential improvement areas.

4. Assessment

- Midterm Q&A: ETL process (data flow, data cleaning, ETL data from source to DW)
- Final Q&A: Completed project (mining DW with reports, OLAP, mining, periodical automatic job creation to perform ETL)
- You can refer to any documents or ask for help during the assignment, but cheating or plagiarism will always result in a 0 for the project. No exceptions!!!
- The teacher evaluates the total score for each group, the group determines the percentage of each member's score depending on the level of contribution to the project.

5. References

Air quality index - WikipediaAir quality index - Wikipedia

US Counties Database | Simplemaps.com

6. Other rules

- Students work in groups and post the source code on Github
- Project includes:
    - The report file includes:
        - Members Information
        - Details of work assignment, % tasks completed
        - Export report from github
        - Analytical sections on data, statistics, visualizations, etc.
    - Main content:
        - Analysis and design of databases (NDS, DDS)
        - Data ETL process analysis (cleaning, transformation, data integration,...)
        - Data mining (OLAP, Report, Mining)

- Source:
  - Script to create database NDS, DDS
  - Project ETL, mining…
- Video:
  - Presentation and demo video (max 10 min)
  - All the members should contribute in the presentation

- Project submission plan:
  - Midterm: around week 5-6
  - Final: around week 11-12