**CIS4400**

**DataWarehouse for Analytics**

**Term Project**

**1 Introduction**

This project on DataWarehouse for Analytics provides an opportunity for you to actively integrate and apply all you have learned on class into a practical, real-world project. You are encouraged in your project to use any tools seen in class and tools you learn on your own. .

**2 Group forming**

Groups will consist of 4 members Max. Students are free to name their teams. I would encourage that you have a Captain. Your team is going to change the face of the company. Each of you will be working as a Data Engineer.

In the real world, you don’t get to choose who you work with. If you get a job at Microsoft, You don’t choose to work with Chuck or Jane. This is the reason why you will be randomly assign to group to promote collaboration between you and your peers. This will help you to deal with people from different background and this will also help you in the real world.

I know also things may change during the year and unexpected events may lead you to drop the class, if this is the case, be kind to let your teammate know about this.

To form the group, I have used the wheelofnames.com and pick each after the spin to form the group. The group are presented as followed.

|  |  |
| --- | --- |
| Group 1  Mohinur Abdullaeva  Leslie Aucapina  Gulnur Dogotari  Timur Naimov | Group 2  Jiang,Jacky Bahrawi,Nabeel Liu,Jeffery Alkaifi,Belal |
| Group 3  Huang,Juntian Rodriguez,Nicole Zhou,Pengyi Patel,Dhyey | Group 4  Amankwah,Christian Kataria,Siddhi Otgontamir,Yondonjamts Chowdhury,Nurul |
| Group 5  Arman,Siam Hou,Susanez Feng,Jiesen Sacoto,Louis | Group 6  Tenzin,Karma Yangchen JinWoo  Yinghua Wu  Fatimah Fatimah |
| Group 7  Greenbaum,Emma Shah,Nidhi  Sawda,Humaiya Alam,Ramisa | Group 8  Islam,Wasikul Noto,Nicholas Rahman,Janeeta Leon,Crystal |
| Group 9  Tanzim,Mahdi Samuels,Elvis Preeo,Thoukir Ahmed,Fatema | Group 10  Wexler,Sara Hoque,Kazi Souffront,Xavier Guallpa,Jason |

**3 Project Selection**

There will be 10 projects and students will be assigned randomly to students. If a group is not comfortable with a project, you will have until February 28th to submit another project which will require my written approval by email.

In the real word, you don’t get to choose the project you will work on, you will be assigning a project automatically. You are required to meet with me to discuss about the requirements for each project.

**4 Instructions for the project**

The project will have the following structure; however you are free to deviate from that

1. Business Problem

Give a description of the company.

Give the requirements and any missing requirements you have found.

1. Business Impact

 You will need to consider any risks, cost, and benefits that are related to this project.  How the persona, company will benefit from it?

[It is useful if you can provide an estimated impact of your target/moonshot analysis. If things go exactly as planned, how much does your company beneﬁt? If things go much better than planned, how much does your company beneﬁt?]

1. Business Persona

Who are going to be the people who use the system? Who are the actors of the system?

1. Data

The data sources that you are going to use and the metadata associated with them.

[Tell us (at least approximately) how large your dataset is. Highlight both the strengths and weaknesses of your dataset(s).]

1. Methods

In this section, we will highlight some of the key tools and elements of our tech stack that we will be using for this project. As we learn and build upon our skills during this program, we hope to add additional tools to this section and improve upon some of our initial assumptions.

1. Data Tools

Data Storage: What are the data storages that you are going to use.

Data Processing: How are you going to process/ingest the data?

Data Orchestration: Are you using any type of tools to automate your data pipeline

1. Interface

The final interface will be a web-based application where the Persona can visualize the results.  People should be able also to connect to your  data using Tableau/PowerBI and do their own analysis.

1. PowerPoint Presentation

Instructions for the powerpoint presentation will be given later in the course.

Part 1 to Part 7 must be the page that is shown when opening your github.

**5 Final Report**

You are expected to write a final report which will be your Github account and a PowerPoint presentation. You are expected to use all the tools we have used in class and if possible new tools as well. It is mandatory to have the diagrams in the document and PowerPoint as well as a screenshot of the tools you have used to manage the software. You are welcomed to innovate and introduce new concepts.

**6 Grading Rubrics**

|  |  |  |  |
| --- | --- | --- | --- |
| Demo skills | Up to 30 points | | Clear and comprehensive presentation of design and implementation details during demo. All group members contribute to demo. Instructions for the demo will be given a month before the presentation. |
| Written Report | Up to 50 points | | All sections and content described in Section 4 prompt are addressed. Experimental results characterizing system performance are presented. Clear and descriptive figures are included. Proper grammar/spelling and appropriate tone is used.  Efficient implementation, consistent structure, requirements are addressed, project management, justification of software methodology, diagrams are presents, lessons learned are clearly stated. Team Organization |
| Team Collaboration | Up to 10pt | | Efficient coordination with the team by using the different tools to build a team. Github Project/Microsoft Teams must shown students communicating with each other. **You must create a group chat on Microsoft Teams and add the Professor.** |
| Peer Evaluation | Up to 10pt | | Each member will grade the member of their team. You will receive a google form for that. **You must grade your team member otherwise you won’t get any grade for your peer evaluation.** |
| **Total:** | | **100 points** | |
| Extra functionality | Up to 10 points | | Extra functionality was shown which exceeds requirements. This extra functionality must involve considerable effort or creativity. |

**7 Project Status and Deadlines**

Groups will periodically submit status updates starting the last week of February.

If you are able to deliverer that project in time, it means you have the required skills to conduct and deliver a real-world project. Someday, you may not feel working on the project because it could be more complicated in your hands. This is fine and that’s ok, I have been there too as a student, but remember, any small and significant work you are doing, brings you closer to the end and to one your greatest accomplishment.

Best of Luck!

Group 1 🡪 Project 3

Group 2 🡪 Project 2

Group 3 🡪 Project 6

Group 4 🡪 Project 5

Group 5 🡪 Project 1

Group 6 🡪 Project 7

Group 7 🡪 Project 10

Group 8 🡪 Project 9

Group 9 🡪 Project 8

Group 10 🡪 Project 4

Project Milestones.

March 14th, 2024 Data Source, Dimensional Modeling

April 11th, 2024, ETL, ELT

May 11th, 2024 Visualization Tools

**8 Group Meeting**

You are required to have group meetings and those group meetings should be documented and add to project board. Furthermore, you are required to create team chat with your group on Microsoft Team and you must add your professor. I will provide a log sheet for that. You can use your preferred channel to communicate, however, I would recommend Microsoft Teams since you can reach each other quickly like in a work environment setting. You are expected to meet more than once and are welcome to make an appointment with the instructor for more clarification.

**9 Project Details**

This is the list of projects that you will be working on:

Project 1: Evaluate the correlation between TSA throughput and the weather July 1st 2023 to September 30th 2023

* I need a map (united states map) that will give me the concentration of passengers.
* A line charge that will give the number of passengers by airport per day. Make sure I can select the Airport.
* More requirements will come.

Keywords: Transportation, Decision Optimization, Algorithms, Geolocation

Project 2: Doctors ratings evaluation on ratemds. You will need to evaluate the ratings of those doctors on healthgrades as well. More requirements will come.

* Data will be provided by professor.
* Healthgrades data will be research alongside with the professor.

Keywork: Healthcare, Fuzz Matching, Geolocation

Project 3: Analysis of the inspection of the New York City Restaurants and their ratings on Yelp Reviews. Is there a correlation between the violation and the ratings?

* You must get the data from the NYC Open Data Set.
* You will working with the professor to minize the calls from Yelp to get the correct ratings for all restaurant.

Keywork: Consumer Discretion, Retail, Fuzz Matching, Geolocation

Project 4: Analysis of the For-Hire Vehicle Trip Records and High Volume For-Hire Vehicle Trip Records in New York City for the from January 1st 2023 to December 31th 2023 and the weather.

* You must get the data from the TLC website.
* You will have a dashboard that will allow me to filter for Green Cab or Yellow Cab. In this dashboard I must have the following
  + Filter by Borough
  + Total trips by taxi type
  + Total trips by taxi type by month, by day
  + Total trips by boroughs
  + Revenue by Boroughs
  + Congestion prices should be added to all trips below the 60th to evaluate the impact of the revenue per month, per quarter and per year.
* Weather from openmeteo

Keywork: Consumer Discretion, Retail, Fuzz Matching, Geolocation, Weather Data

Project 5: Analysis of the PPP Loans that were issue to Small Business during COVID-19.

* Total loan by companies
* Total loan by State
* Total Business by States
* Top 10 Lenders by State
* Total Loan by Lenders
* Classification by NAICS Code
* Aggreate data based on GDP for the past 4 years, or COVID-19 per zipcode

Keyword : Finance, Geolocation

Project 6: The Centers for Medicare and Medicaid Services wants to know the profession and the popularity of the all the doctors that are getting paid by State.

* Total Doctors by State
* Total Companies by State
* Ratings by County and Zip Code For each Specialty.

Keyword: Healthcare, Geolocation

Project 7: Analysis of the OTC Market Transaction.

* Data will be provided by the professor. You must study closely the OTC market.
* Requirements will be provided by the professor.
* You are required to do Technical Analysis
* You are required to do Deep Learning (Algorithm will be provided)

Keyword: Finance, Deep Learning, Technical Analysis, Time Series

Project 8: Walmart, Target and Amazon is looking to standardize they paycheck based on market data and cost of livings

* MIT living wages
* Minimum Wages per State
* Census Data
* DMA Data
* Levels FYI Data
* Job Board (EchoJobs, <https://topstartups.io/jobs/>, efinancialCareers, https://remoteworkjunkie.com/stack-overflow-alternatives/)

Your task will be to help HR to determine the salary of someone based on the location, the cost of living, and the market data. Other requirements will be provided by the Professor.

Keyword: Consumer Discretion, People Analytics, MacroEconomics

Project 9: You are looking to buy a new home and your criteria is looking for cost of living, the housing sale price by county, The housing listing price p and the crime information in the area, the school rating

* You will research the data with the professor and discuss the requirements

Keyword: Consumer Discretion, Macroeconomics, Real Estate

Project 10: Setting up a Datawarehouse that contains information about CME data. Make sure there is also reference data. Data will be provided by the professor. Data will be served through in a various way. The data to be retrieve will be provided by the professor and the data will be constantly updated.

Keyword: Finance, Commodities Market

Project 11: SEC data for all 4000 companies. The most important data are the accounting report.

Requirements are pending.

Grading Rubric

|  |  |  |
| --- | --- | --- |
| Section | Checklist | Grades |
| Collaboration | * Microsoft Teams Chat Creation * Github Project Board Created * Valid Task on Github Project (10 minimum). A valid task contains complete details about the task executed/assigned by a team member. It must be properly documented, and I can see the status of the Task at any point in time. |  |
| Presentation PowerPoint | * No Paragrah * Clear and Simple Tex   + San serif font   + No smaller than 20 pt   + No more than 1-2 Lines Per bullet   + No more than 3-5 bullets per slide   + Reveal Bullets one at a time   + If there is more than 6 bullets points, create another slide with the title of the previous slide follow by CONT’D   + Simple Color Scheme   + Perfect Visual Organization   + Don’t mix images with extended text   + Use High res Images   + One Image per Slide   + One graph per slide and provide context   + No 3D visualization   + Using only 2 colors (mostly Blue or Orange for Contrast)   + Give audience at least 30 seconds to look for the images |  |
| Presentation Speaking | * Eye Contact with Audience * No Staring * Vocal Variety * Appropriate hands Gesture * Appropriate Posture * No Leaning * No Slide Reading |  |
| Peer Grading | * Average of all grading made by your peer * If you have not grade anyone you will automatically receive 0 * if you didn’t grade someone in the group, you will receive a pro rated grade. For example, if you are in a team of 5 , you only grade 3 members out of the 5, you have receive the following grade for 5 members (10,10,10,10,10) You will only get the grade from the people that you evaluate it means (30/50) |  |
| Project Report | * Github Repo Created * Github Repo contains more than 20 commits * Github Repo contains SQL scripts that generated the datawarehouse * Github Repo contains script that extract data * Github Repo contains script that transform the data * Github Repo contains script that load the data * Github Repo contains your powerpoint * Github Repo contains diagram of your pipeline * GIthub Repo contains diagram of your dimensional modeling |  |