**ETL Project Proposal**

*Use this as a general guide for your project proposal. Please keep the headers below and populate with your proposal text. We have added general questions under each header for your group to address, but feel free to add other questions. It's okay if your final project deviates from your proposal. The purpose of the proposal is to help your team to formulate a plan, set clear goals and get to work ASAP.*

**Team Members:**

1. Zheng Qi
2. Oscar Hui
3. Chad Molenaar
4. Joe Quinn

**Project overview (200-300 words)**

● *What is your question of interest?*

How can both gamers and game publishers access data through **SQL/MongoDB etc** to better understand and/or for analyses for information such as:

○ Steam game sales?

○ Most successful publishers?

○ Most popular genres?

○ Size of online player base?

● *What is the expected outcome of your ETL project? (i.e. how will this data ETL project add value?) Think about how your ETL project will enrich the data and how this data might create value.*

This data will essentially be a summary of multiple different data sets to give an accurate picture of the game sales of Steam.

As a group, we are looking to allow both to have the information available to them that will help them analyse and extract data on a **Regional** and **GDP per capita basis.**

Game publishers can use this data to gain a better understanding of what the most popular genres are so they can choose to develop a game targeting a wider audience or aim at a more niche fanbase. The size of the online player base can help both gamers and publishers.

Gamers can use this data to look at the popularity of different publishers so they can get a better idea of how good the game will be before making a purchase. In addition to this, gamers can use this to help them decide on which game to buy. This is especially useful for users who want to be a part of an online community and play in larger online games.

Through the extraction, transformation and data storage processes used, anyone will be able to save resources such as time, money, and trusted information. An important value-add for both gamers and game developers.

**EXTRACT - Proposed data sources (minimum two data sources)**

* *Where is the data located?*
  + Data sources are Data.world, Kaggle and Steam.
* *What are the data set formats?*
  + Json and csv
* *How will you get this data?* 
  + Combination of API and downloaded data.
* *What are your assumptions about the data?* 
  + Potentially messy and may require pretty print. Utilising JSON can have its pitfalls as it comes out as a dictionary.
  + CSV files tend to hold a lot of duplicates, misspellings and null values as examples**.**
  + We are additionally holding the assumption that it will hold the Regional and GDP data we are looking for, however, this will be identified with further investigation of other potential databases/sets.
* *What does it cover?* 
  + We are expecting to locate common values or headers (such as geographical data or demographical data) that will enable us to answer the above questions through the cleansing and transformation process.
* *What is it missing?* ***TBA***

**TRANSFORM - Proposed clean-up and analysis**

* *What are the transformations you will apply to the data? (e.g. filtering, aggregation, derived columns)* 
  + Filtering, aggregation,
* *What steps will you take to clean the data and ensure its validity? (e.g. messy data, duplicated data, incorrectly formatted data)* 
  + *Drop the duplicated data. Convert the data to the right format.*
* *How will you identify potential issues with your data sources? (e.g. exploratory data analysis, simple statistics etc)* 
  + Exploratory data analysis
* *How will the data be integrated? (e.g. joins, merges)* 
  + Mergers
* *How will you apply these transformations (e.g. jupyter notebook, python script)* 
  + Pandas, Jupyter notebook, Json
* *IMPORTANT → Why did you apply these transformations? How did this enrich your data?* 
  + *The data requested through API is quite messy and non structured. To investigate the relationship between different factors, the data need to be converted from non-structured data to structured data.*

**LOAD - Data storage**

* *What type of database (relational, document) will you store the data?* 
  + *SQL database*
* *Why did you choose this database over another database?* 
  + *Structured data for easier comparison.*
* *What are your expected tables / documents and relationships between tables / documents in your database? TBA*

**Potential limitations**

* *What are the potential limitations of your above proposed steps?* 
  + *As the publishing time of different games is different, the aggregation results may be limited to a particular time.*
* *How can you control these potential issues?* 
  + *We will not be considering the recent published game and specify the selection to a certain period.*

#########################################################################

**ETL DIAGRAM**

It may be easier to explain the process in an ETL diagram (you don’t have to create a diagram but it may help your group to formulate a plan).

Most data engineers make sure to document the ETL processes for reference. One way to do this is to use ETL diagrams. This is usually drafted at the beginning of the project and finalised at the end of the project.

Here are a two examples of how an ETL diagram would look like:

