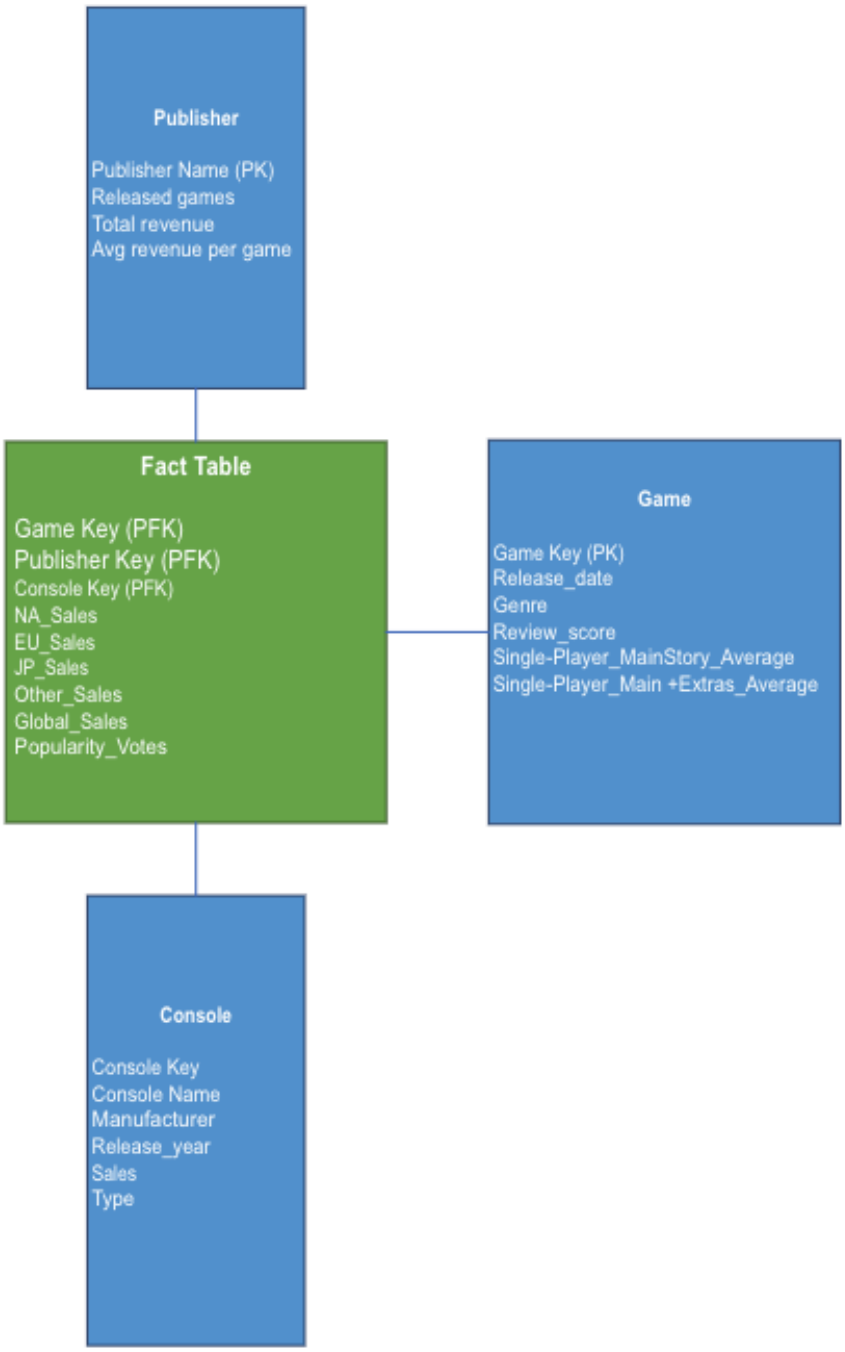
Project Phase 1: Conceptual Design
Group# 12

Students:
Manav Patel: 300074687
Othmane Ayoub: 300050124
Moumin Farah: 300026540

Project Description:

The project entails designing and implementing a data mart for video game sales data. The data mart will be created using EECS database or PostgreSQL. The data mart will include information such as the video game's name, platform on which it was released, year of release, genre, publisher, and sales figures for Europe, Japan, and other regions. Following the creation of the data mart, we will explore the data using online analytic processing (OLAP) techniques. This will allow us to analyze the data from various angles and gain insights into video game sales trend but also the reason that might influence the sales such as the year, the name of the console etc... In addition, we will use visualisation and mining techniques to uncover patterns and relationships in data.

1. Grain:
   The popularity and sales of a game made by a specific publisher for a specific console.



The diagram above shows the dimensions and attributes for each dimension for our conceptual model.

2. Details pertaining to our dimensions and dimensional attributes.

- Publisher dimension.

Publisher Name (PK): String, Sample value = ‘Rockstar games’.
Released games: Integer, minimum = 0 a and maximum = no limit, Sample value = ‘45’.
Total revenue: “$” + Float concatenated with k, m or bn (k=thousand, m=million, bn=billion), minimum = 0, Sample value = ‘$1.3bn’.
Avg revenue per game: “$” + Float concatenated with k, m or bn, minimum = 0, Sample value = ‘$20000’.

- Game dimension

Game Key (PK) : String, Sample value = ‘GTA V’
Release_date: [Release date of the game.] Date, Format: YYYY-MM-DD, Sample value = 1993-01-24
Genre: [Genre of the game.] String, One of the following 9 values: Action, Adventure, Comedy, Crime, Family, Fantasy, Mystery, Sci-Fi, Thriller.

Note: We plan to assimilate genres from our listed external sources and potentially add and remove some of these unique identifiers.
Review_score: [User's Review score of the game based on metacritics.] Integer, minimum = 0 and maximum = 100, Sample value = 87
Single-Player_MainStory_Average: [Fastest time it takes to complete the single-player main story] Date-Time, Format: HHh MMm, Sample value = 62h 42m. [***Note: The value of time cannot be negative]***.
Single-Player_Main +Extras_Average: [Average time it takes to complete the single-player main story + Extras] Date-Time, Format: HHh MMm, Sample value = 62h 42m. [***Note: The value of time cannot be negative]***.

- Console dimension

Console Key: String, [***Possible values: PS2, PSP, DS etc. Note: There are 44 such unique values. We may reduce this number based on the relevance of the gaming console as per current market. Since some consoles may have been discontinued***.] Sample Value = 'PS4'.
Console Name: String, [Similar to the keys above, there are 44 potential values.] Sample Value = 'Playstation 4'.
Manufacturer: [Manufacturer of the console] String, [***Currently there are more than 7 unique identifiers for this attribute. We will be pruning these classes based on current market relevance.]*** Sample Value = 'Sony'.
Release_year: Integer 4-digit, minimum = 1965 and maximum = 2023, Sample Value = '2022'
Sales: [Units sold in million] Float, minimum = 0, Sample Value = 155
Type: [Type of device] String keys : [Home, Handheld, Hybrid], Sample Value= 'Home'

3. **Details of facts and measures:**

Game Key (PFK): [Primary key for Game dimension] Integer, Sample Value = 1
Publisher Key (PFK): [Primary key for Publisher dimension] Integer, Sample Value = 2
Console Key (PFK): [Primary key for Console dimension] Integer, Sample Value = 3
Note: For all PFKs, the Integer number will be mapped to a unique class of that dimension. The number of such keys is dependent on the unique number of classes that exist for each dimension
NA_Sales: [Sales in North America (in millions)] Float, minimum = 0, Sample Value = 82.6
EU_Sales: [Sales in Europe (in millions)] Float, minimum = 0, Sample Value = 82.6
JP_Sales: [Sales in Japan (in millions)] Float, minimum = 0, Sample Value = 82.6
Other_Sales: [Sales in the rest of the world (in millions)] Float, minimum = 0, Sample Value = 82.6
Global_Sales: [Total worldwide sales(in millions)] Float, minimum = 0, Sample Value = 82.6
Popularity_Votes: [Number of popularity votes] Integer, minimum=0 , Sample Value = 15000

4. Checklist of the 10 design mistakes as discussed in class:

Criteria	Relevance
Place text attributes in the Fact table.	The foreign keys in the fact table are text attributes. A solution to avoid that would be the use of an integer surrogate key and include an index table that would help match the different text keys to the correct integer surrogate key.
Limit verbose descriptions to save space.	The concept design report details all the information about the project, the grain, the fact, the dimensions and the description of each of their attributes as well as the different constraint the data is subject to.
Normalize to save space (leads to slower queries!)	Space is not a constraint in our design. We prioritize query speed. Therefore, no normalization has been done to the data so that our model would remain a star.
Ignore the need to track changes.	Adding a time attribute called “Current_time” to save the initial load as well as the refresh time on change would help us keep track of changes.
Add new hardware to solve all query performance issues	We focus on the design of system and data integration as the main angle to increase the performance
Use operational key’s as the primary keys.	As mentioned in the first row, we plan on using surrogate key in order to avoid production keys and text keys.
Neglect to declare (and comply with) the grain.	In order to avoid this issue, we have a detailed data structure made from different dimensions (that have their own attributes) and a fact table. This will help us stick with the grain because it helps us visualize the grain.
Neglect a detailed design.	Setting clear goals/objectives, breaking the design down into many dimensions with each their foreign keys, getting help from the TA and validating the design to make sure that it represents the purpose of our research helped to deal with this issue.
Expect users to query normalized data.	The datasets came from various sources and various formats. We are expected to ensure the quality and conformity of the data.
Fail to conform Facts and Dimensions.	This can be avoided by clearly defining facts and dimensions, establishing naming conventions, using a consistent data model,

	using a data quality tool and continuously monitoring and improving.

5. Assumptions:

- As a game consulting firm, analyze data from the gaming industry to help your client in identifying the most bankable game genre, style, gaming platform.
- We assume that we have access to a subset of raw data for the gaming industry.
- Our datasets are taking a subset of what the actual game votes and sales might be, however the proportion indicates how popular or not a game is based on this subset.
- Our current set of data lists/classifies multiple such games, consoles and publishers that have been discontinued. For now, we assume that all the listed classes are relevant to our model. However, we plan to prune our dataset to make it more relevant to today's market.
- We assume that attributes related to game playtime are an average of the time that players take to finish a particular game in main story mode or story mode with extras. Essentially, we intend to use the metric as a standard time required to finish a game.
- We will be using our source datasets to create datasets as per our dimensional models.
- We assume that the sample used to get the review scores and measure the playtime of a game is representative of all its player base.

6. Possible Additions:
    a. Additionally, we feel adding a dimension called 'Marketing channel', can provide us with great insights on what marketing channels a certain game published utilized to promote their games. This could relate to how these channels affected the sales of a particular game for a game publisher. Unfortunately, we were unable to find an external source to support this hypothesis. However, if we find

an external source that provides us with such data, we do plan to incorporate such a dimension.

b. As mentioned in the 10 mistakes to avoid, an index table with an integer surrogate key that would uniquely refer to the game no matter how it is named in the different datasets will be added.

Team Work Summary:

Team Meeting 1	2023, February 5th - All members looked for potential dataset sources and listed them. - Decided that each of us would create a data model based on our understanding and take the best features of all three models created.
Team meeting 2	2023, February 7th - Discussed features of data models and Finalized grain + data model.
Team meeting 3 with TA	2023, February 10th - Discussed probable improvements in data model design. - Manav + Moumin : Model improvements

	- Othmane : Checklist supervision for 10 mistakes as discussed in class.
Team meeting 4	2023, February 12th - Manav + Moumin + Othmane : Report Finalization.

References – External sources that we intend to use for our dimensions (Subject to change):

- Dimension Game: https://www.kaggle.com/datasets/muhammadadiltalay/imdb-video-games?select=imdb-videogames.csv
- Dimension Game: https://www.kaggle.com/datasets/gregorut/videogamesales
- Dimension Game: https://www.kaggle.com/datasets/baraazaid/how-long-to-beat-video-games
- Dimension Console: https://www.kaggle.com/datasets/jaimepazlopes/game-console-manufactor-and-sales
- Dimension Publisher: https://vginsights.com/publishers-database