Data Gathering and Warehousing

Final Project

Fictional Company/Organization

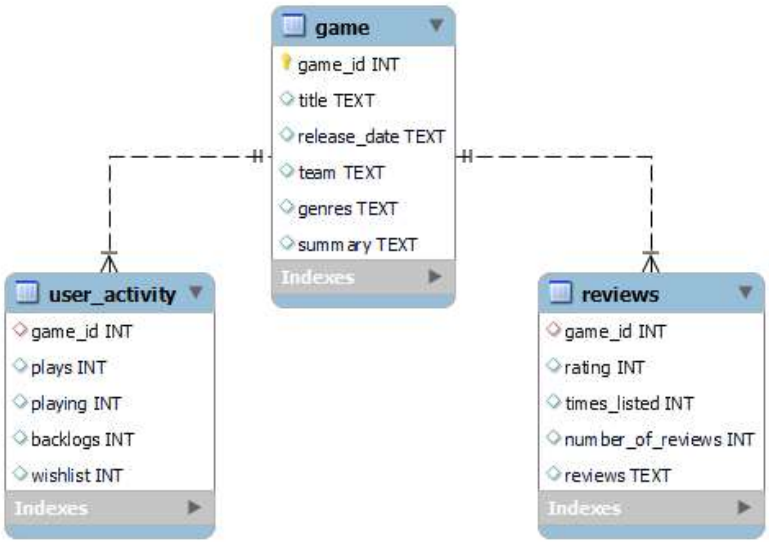
- Name:
    - Data-Solstice gaming
- Mission Statement:
    - We at Data-Solstice gaming make it our mission to create only games our players care about. Our goal is to only produce titles that our players care about and want to come back to play.
- Background information of the fictional company:
    - We founded our company because we noticed that most gamers were complaining about games being buggy or becoming repetitive and boring after a week. That is why we made it our mission to only deliver quality content for our users.

Data

- We will use data from a gaming site called Backloggd and various other companies that collect user reviews. The data we collect from the companies will include information regarding the game and its team. These sites also collect data from the player such as what they would rate the game, what was their opinion of the game and if the user is playing the game right now. The current dataset takes data from 1980 to 2023. The data was scrapped from the company and stored in a database.
- We ensured the quality of the data by checking its accuracy through the correct assignment of information to each game. We also looked at the data's consistency, we ensured the data did not have any inconsistent or contradictory information. Last, we checked the data for conformity to ensure that all units of measurements and data types are consistent. At the beginning of the data collection stage, we implement this.
- To standardize data, we clean the data after extracting it and before analyzing it. For items that appear to be duplicates, we will inspect if the information is completely identical. If the data is exactly the same, we will drop it. Otherwise, we will need to inspect and merge the items. First, we rename the columns in order to clean the data and make them easy to understand. To prevent missing records issues, we establish a rule to ensure that each column has the correct data type and fill missing data with NA values.
- The usage agreement that we have is we can use the data in any fashion we like since we have a subscription with each company to collect their data.

Database Design

- The database schema that works best for our data is the Star method. We will use the game information as the main table with the primary key and combine other tables with it. To assess information regarding the data.
- The database will comprise three tables ‘game’, ‘user_activity’, and ‘reviews’. The game_id will be the primary key in ‘game’ and a foreign key in the other tables.



Database Storage

- As most of our employees work remotely, we store the data on a server. This means we have to keep our database on a server so that it is accessible to the entire team.
- We will use the database management system SQLite, by using a .db file.
- The data file video_game_analysis.db and it is loaded in after importing the appropriate library.
    - with R
        - db <- dbConnect(dbDriver("SQLite"), dbname = "video_game_analysis.db")
    - with python
        - conn = sqlite3.connect('video_game_analysis.db')
- Our backup and recovery strategy entails leveraging cloud storage solutions, coupled with regularly scheduled backups. This approach guarantees enhanced data protection, preparing us for any unforeseen circumstances.

Database Access and Analysis

- The question we plan to find in our data is that teams developed the highest rated? What was the lowest? What did users say about the games?
- What genres had the highest rated titles, and what teams were associated with that. What is the user activity related to that genre?
- We plan to visualize the data by making a dashboard and creating a PowerPoint.