ETL Report, Working-Draft V1:

Regarding Marketing Analytics for Video Game Sales (Name Pending)

Compiled by Group 6
Members:
Hugh, Jason, Ryan and Todd

Introduction

The goal of this project was to determine what factors were the main drivers behind video game sales on a global and regional basis. Since marketing data is usually private, our analysis was based on trends observed across publicly available sales data sourced from a web-scraped Kaggle-dataset [1]. The dataset spans over 30 years of sales records from the 80s' to 2019 and features over 2,000 rows of critical reviews and categorical data (developer, publisher, ESRB-rating and genre) with regards to the video games featured.
To get around the roadblock induced by the private nature of streamed marketing data, we simulated a data-stream using Azure Databricks and Kafka. We automated the process using an Azure Datafactory, triggered at 15-minute intervals and saved the data consumed to a SQL database. The retrieved data was then used to train a machine-learning model for predictive analysis w.r.t global and regional sales. The flowchart below summarizes our setup:

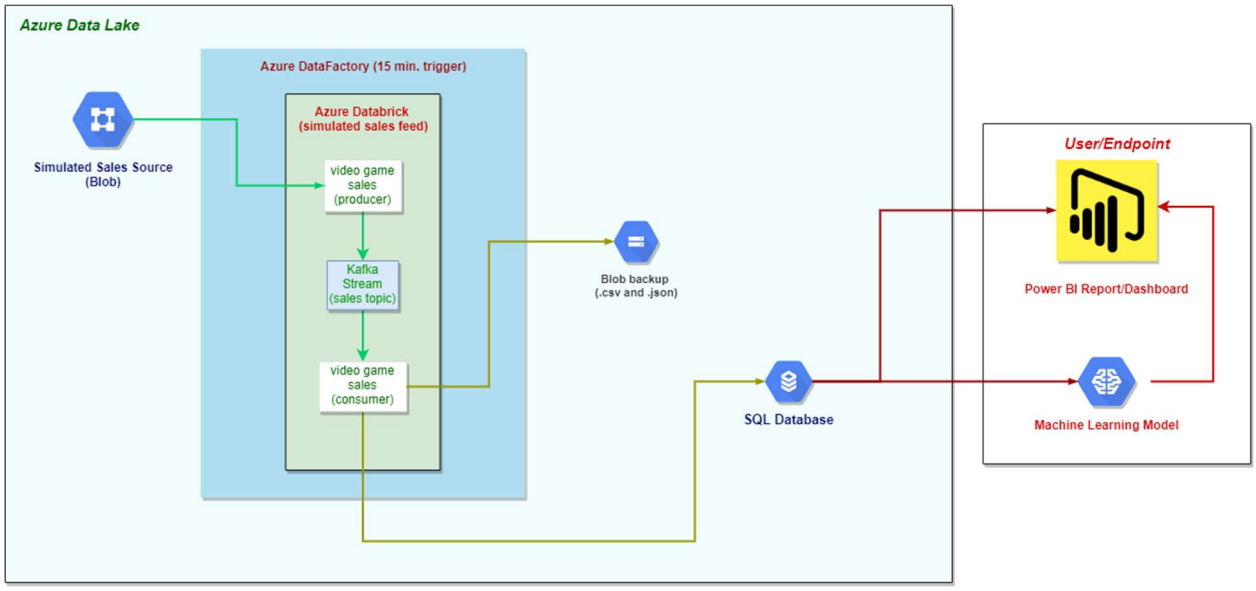


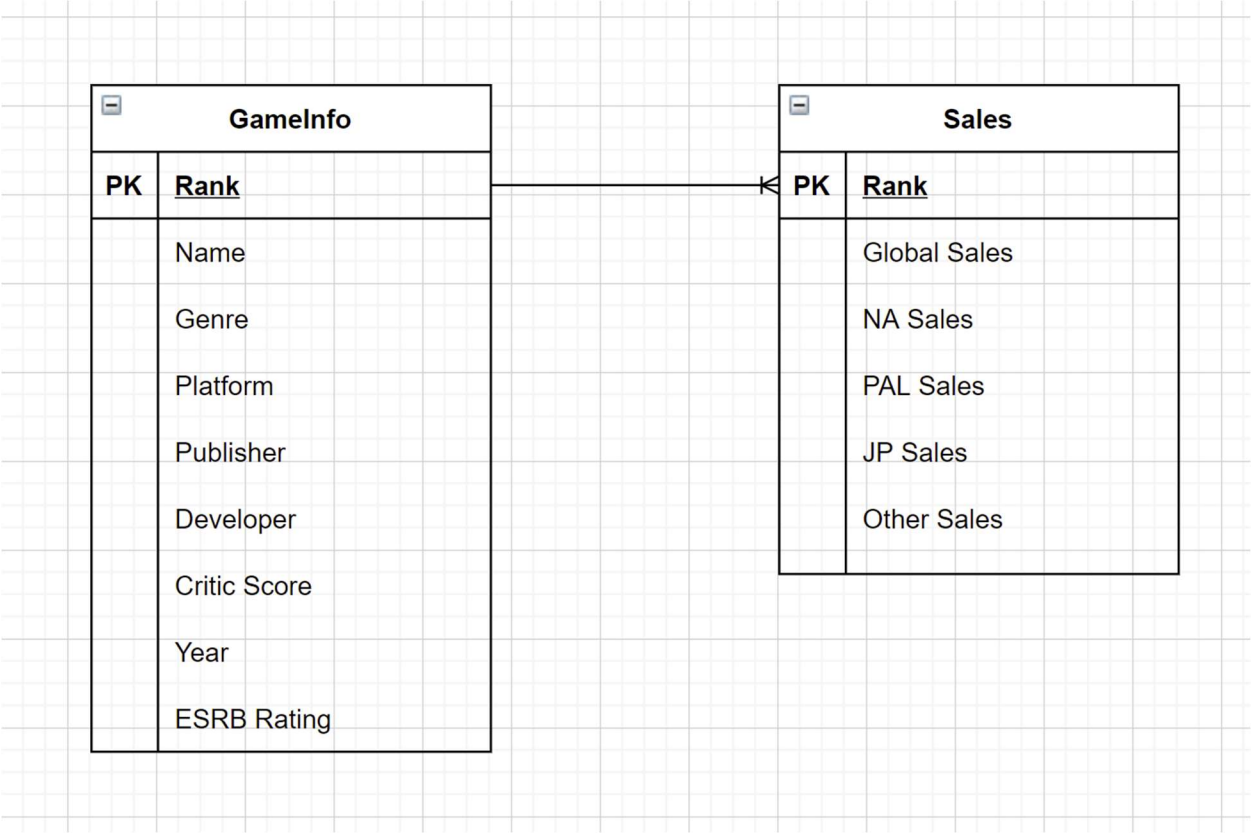
Figure 1: Data-Flow representation mediated by Kafka, from source to endpoint

Regarding the Setup:

Our main data source [1] was saved as a .csv to a blob in an Azure Data Lake. An Azure Databrick was then setup to create the Kafka Producer and Consumer for the simulated data-stream.
The Kafka Producer was made so that it pulled data refined from the source blob (the data was cleaned and converted to a list of dictionaries), and pushed it to a Kafka topic, one row at a time, at 5-second intervals.
The Kafka Consumer then pulled from the Kafka topic, the rows that were pushed from the Kafka producer, and accumulated them in a list of dictionaries. After the Consumer had retrieved all of the data from the Producer, it saved it to an SQL Database (and a blob as a backup) where it could be retrieved and used for data-analysis and machine-learning.
The Databrick was encapsulated in an Azure Datafactory, to act as a pipeline for the flow from the source blob (simulating live marketing data) to the SQL Database (and blob backup). The Datafactory was made to trigger flows every 15 minutes to simulate a real-world automated process.

SQL Database Setup

The SQL database that was setup for this project was setup according to the following ERD.



As you can see in the ERD there are two two tables that are in the database that use Rank as the primary key and the foreign key to join the two tables together. To setup the database and load the data from the CSV into the SQL database I created a schema that would create two tables that were empty, which could then be filled in with data from a temporary table. You can see an example of the schema and the queries used to fill the tables below.

```
Drop Table if EXISTS GameInfo
Drop Table if EXISTS Sales
CREATE TABLE Sales(
    [Rank] int primary key,
    [Total_Shipped] [float] NULL,
    [Global_Sales] [float] NULL,
    [NA_Sales] [float] NULL,
    [PAL_Sales] [float] NULL,
    [JP_Sales] [float] NULL,
    [Other_Sales] [float] NULL
)

CREATE TABLE GameInfo (
    [Rank] int primary key,
```

```
    [Name] [nvarchar](max) NULL,
    [Genre] [nvarchar](max) NULL,
    [ESRB_Rating] [nvarchar](max) NULL,
    [Platform] [nvarchar](max) NULL,
    [Publisher] [nvarchar](max) NULL,
    [Developer] [nvarchar](max) NULL,
    [Critic_Score] [float] NULL,
    [User_Score] [float] NULL,
    [Year] [int] NULL
    CONSTRAINT FK_GameInfo_Sales
        FOREIGN KEY (Rank) -- field on this Table
        REFERENCES Sales(Rank) -- what table do we refer to and what field do we
refer to
)
```

```
INSERT into Sales ([Rank], [Total_Shipped], [Global_Sales], [NA_Sales],
[PAL_Sales], [JP_Sales], [Other_Sales])
    SELECT [Rank], [Total_Shipped], [Global_Sales], [NA_Sales], [PAL_Sales],
[JP_Sales], [Other_Sales]
    FROM TotalShipped_Test

INSERT into GameInfo ([Rank], [Name], [Genre], [ESRB_Rating], [Platform],
[Publisher], [Developer], [Critic_Score], [User_Score], [Year])
    SELECT [Rank], [Name], [Genre], [ESRB_Rating], [Platform], [Publisher],
[Developer], [Critic_Score], [User_Score], [Year]
    FROM TotalShipped_Test
```

Regarding the Data:

Using the data retrieved and with some machine-learning refinement, we will attempt to answer the following questions with regards to our marketing analytics, viz.:

1. How much does a game's rating affect its sales?
2. How much does a game's review (critical/user) affect its sales?
3. Do different consoles sell more games than others?
4. How do different regions affect game sales?
5. Do certain genres sell more than others?
6. How are genres trending over time?
7. How are global sales trending over time?
8. What factors have the most influence on games sales?

More potential questions if we can find the relevant datasets:

1.) How does an emerging platform affect sales?
2.) How did the pandemic affect sales?
3.) (census related) What percentage of retail sales accounts for video games in the US, and potentially Japan?

The end-goal is to paint a picture of what makes a best-selling game.

References:

1.) ***Video Games Sales 2019***, "*Sales and Scores for more than 55,000 games*"**.** Retrieved from Kaggle

2.) ***Video Game Dataset***, "*474417 Game with Metacritic Score, Ratings, Genres, Publishers, Platforms, ...*" Retrieved from Kaggle

3.) VGChartz (For web-scraping)

4.) US Census Bureau, Retail Trade: Summary Statistics for the U.S., States, and Selected Geographies: 2017. Survey/Program: Economic Census, TableID: EC1744BASIC, Dataset: ECNBASIC2017. (directlink)