A picture containing drawing

Description automatically generated

**ITC 6460 Cloud Analytics**

**Final Project Report – Video Games Market Analysis**

**Group: Husky 2**

By

**- Katherine LaConte**

**- Likhitha Varakala**

**- Manmitha Pantangi**

**- Shivam Sinha**

**Introduction**

The dataset for this analysis encompasses two comprehensive files: game\_sales\_data.csv and vgsales.csv. These datasets provide an extensive overview of the video game industry, cataloging various aspects such as game titles, platforms, release years, genre classifications, and sales figures across different regions including North America, Europe, Japan, and other territories. This rich compilation of data offers a unique lens through which to view the dynamics of video game sales globally, capturing the trends, preferences, and shifts in the market over time. It serves as a critical foundation for the analytical tasks ahead, enabling a multifaceted exploration of what drives success in the video game industry.

A table with many different colored text

Description automatically generated with medium confidence**Fig 1:** Dataset 1

A table with many different colored text

Description automatically generated with medium confidence

**Fig 2:** Dataset 2

**Data Cleaning**

The data cleaning process for the video game sales datasets, `game\_sales\_data.csv` and `vgsales.csv`, was a meticulous task aimed at ensuring the data's accuracy, completeness, and consistency before analysis. This foundational step involved several critical actions to refine the datasets, making them suitable for in-depth examination and analysis. Initially, **duplicate entries** were identified and removed to prevent skewed results, ensuring that each record uniquely represented a video game. This step was crucial for maintaining the integrity of the dataset and avoiding inflation of sales figures or misrepresentation of market trends.

Handling **missing values** was another significant aspect of the data cleaning process, requiring careful consideration to decide whether to impute missing data based on available information or remove entries lacking critical details. This decision-making process was vital for preserving the dataset's quality without introducing bias or inaccuracies.

**Data type conversion** was carried out to align the data with the requirements of the analysis tools and techniques to be applied later. Sales figures were converted to numeric types to facilitate mathematical operations and comparisons, while release years were adjusted to date formats to enable temporal analyses. This step ensured that the datasets were in a state that accurately reflected the reality of the market, allowing for meaningful insights to be drawn.

Finally, **merging the datasets** was a complex but essential task, requiring alignment of similar columns, reconciliation of genre classifications, and ensuring consistent data granularity. This integration process created a unified dataset that offered a comprehensive view of the video game market, laying a solid foundation for the subsequent analytical phases.

Through these meticulous data cleaning and preparation steps, the datasets were transformed into a reliable source for analysis, setting the stage for uncovering valuable insights into video game sales trends and market dynamics.

**AWS Services**

In this analysis project, the utilization of Amazon Web Services (AWS) played a pivotal role in handling, analyzing, and visualizing large-scale video game sales data. Through the combined use of AWS S3, IAM, AWS Glue, Amazon SageMaker, and AWS QuickSight, a comprehensive and secure data analysis pipeline was established, enabling the extraction of valuable insights from the datasets `game\_sales\_data.csv` and `vgsales.csv`.

**AWS S3** served as the cornerstone for data storage, offering a highly durable, scalable, and secure solution for managing the extensive datasets required for this analysis. By leveraging S3, we ensured that our data was stored in a centralized location, accessible from anywhere, yet protected against unauthorized access and loss. This facilitated seamless integration with other AWS services for further processing and analysis.

**AWS Identity and Access Management (IAM)** was instrumental in safeguarding the analysis process. By meticulously managing permissions and roles, IAM enabled us to define who could access the datasets and analytical tools, thus ensuring that only authorized personnel had the ability to interact with the sensitive data. This level of security is critical in maintaining the integrity of the data analysis process and protecting against potential data breaches or unauthorized access.

**AWS Glue** played a critical role in the data preparation phase. As a fully managed extract, transform, and load (ETL) service, AWS Glue automated the time-consuming tasks of data cataloging, cleaning, enrichment, and transformation. It turned raw data from S3 into a structured format that was ready for analysis. This automation not only streamlined the data preparation process but also ensured consistency and reliability in the data being analyzed.

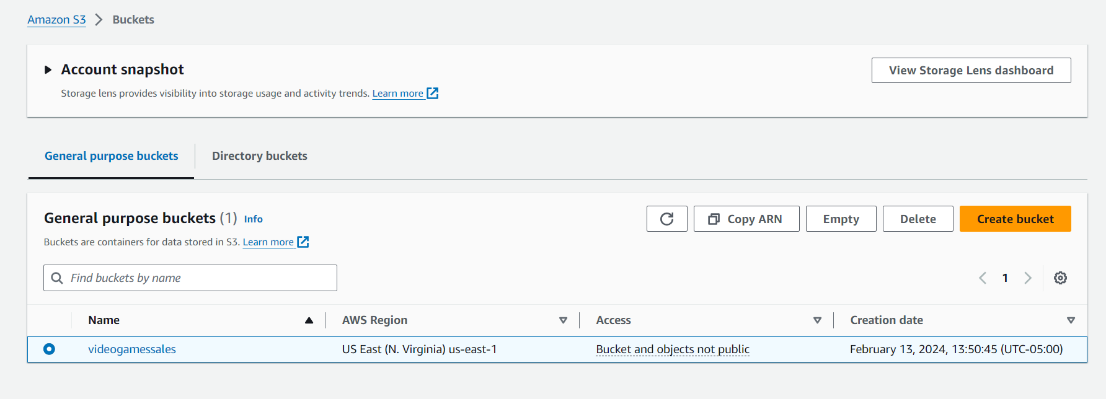
**Amazon SageMaker** was leveraged for its powerful machine learning capabilities. It enabled the building, training, and deployment of a regression model aimed at predicting global video game sales. SageMaker's comprehensive and user-friendly environment accelerated the development of the model by providing access to high-performance computing resources, pre-built algorithms, and the flexibility to experiment with different model configurations. This facilitated the creation of a predictive model that could accurately forecast sales trends based on historical data and other influencing factors.

**AWS QuickSight** was the final piece in the data analysis pipeline, bringing the insights and findings to life through interactive dashboards and visualizations. QuickSight's ability to connect directly to data stored in AWS services, such as S3 and AWS Glue, allowed for real-time analysis and visualization. The dashboards created provided a dynamic and intuitive interface for exploring the data, enabling stakeholders to quickly understand the analysis outcomes and make informed decisions. By leveraging QuickSight's advanced visualization features, we were able to highlight key trends, patterns, and anomalies in the video game sales data, making the insights accessible to a broad audience.

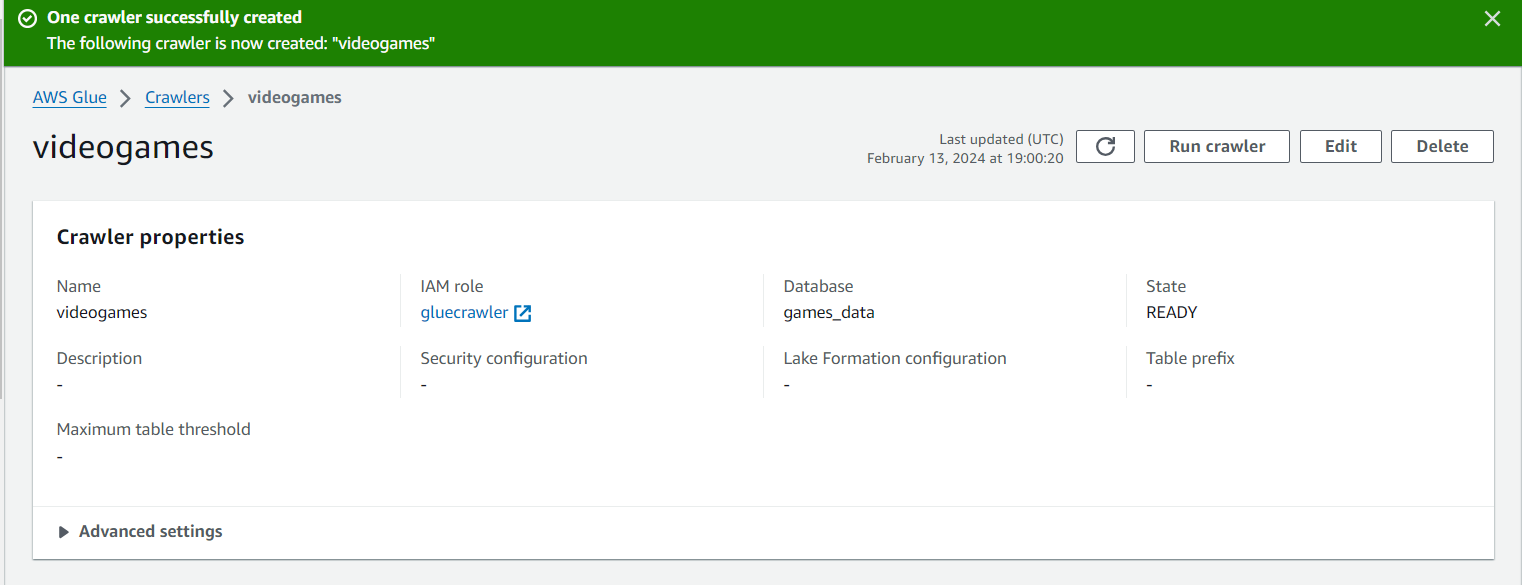
In summary, the synergy between AWS S3, IAM, AWS Glue, Amazon SageMaker, and AWS QuickSight formed the backbone of this comprehensive analysis project. This integration not only facilitated a secure and efficient data analysis workflow but also enabled the extraction of actionable insights from complex datasets, demonstrating the power of AWS in supporting data-driven decision-making in the video game industry.

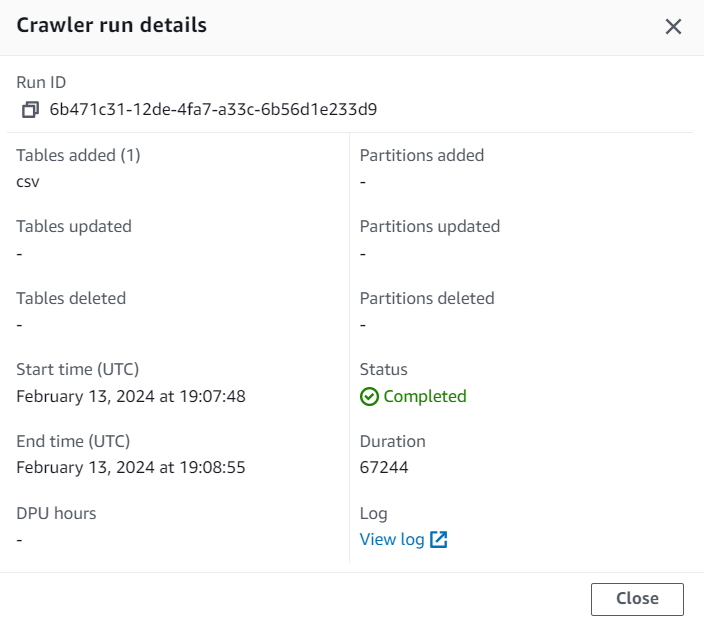
**Steps**

**Create an S3 Bucket:** Log into the AWS Management Console, navigate to S3, and create a new bucket to store your datasets (game\_sales\_data.csv and vgsales.csv). Ensure to configure the bucket settings according to your privacy and access requirements.

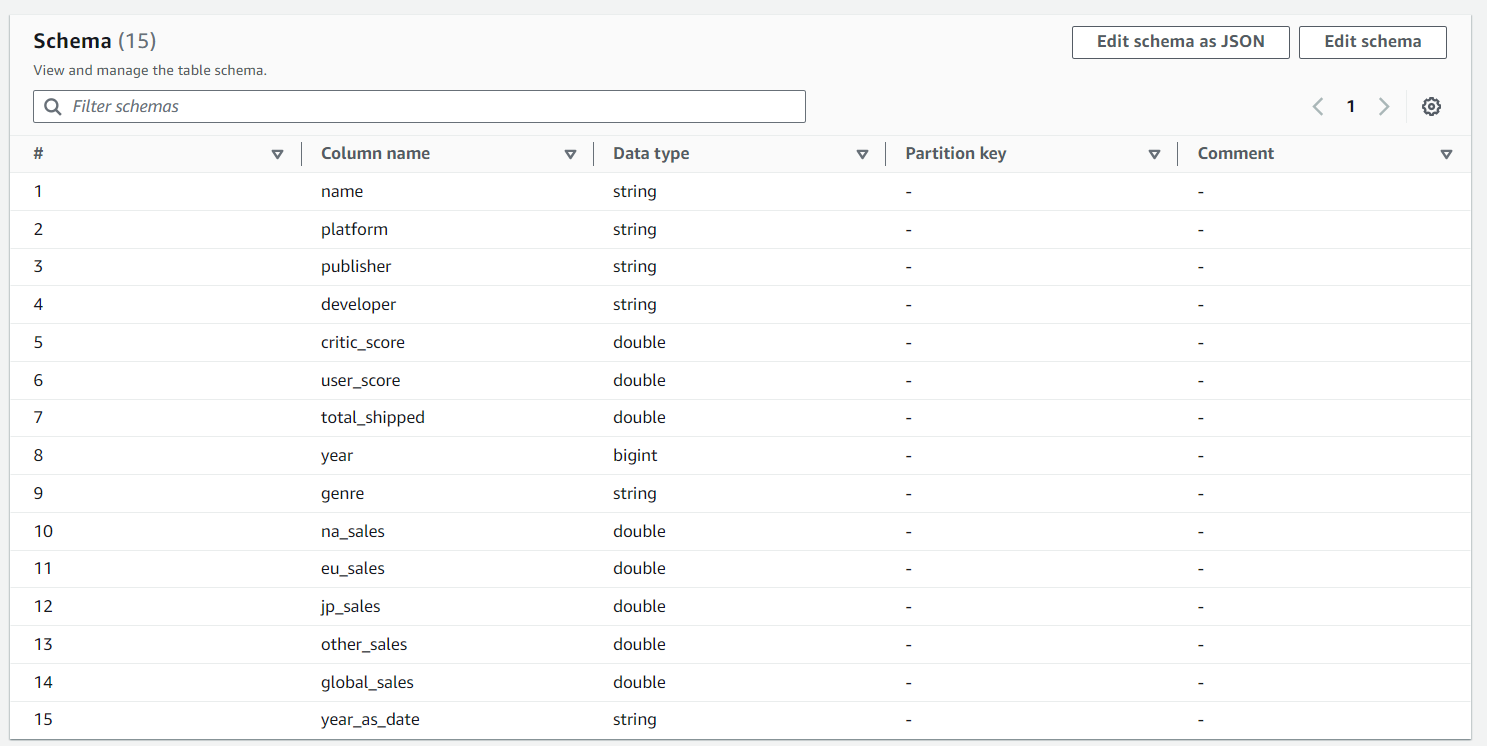
**Use AWS Glue for Data Cataloging:**

**Create a Crawler:** In AWS Glue, set up a new crawler pointing to your S3 bucket. The crawler scans your data and infers schemas.

**Run the Crawler:** Execute the crawler to populate the AWS Glue Data Catalog with tables representing your datasets.



**Table Schema:**

**Querying Data with AWS Athena:**

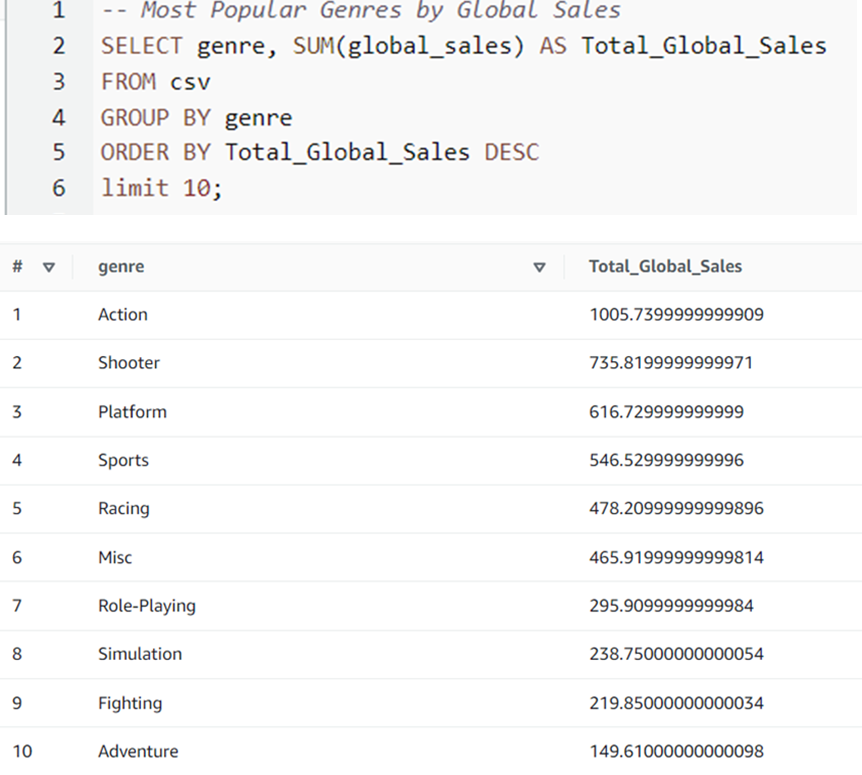
Navigate to AWS Athena in the AWS Management Console.

Ensure Athena is set to use the Data Catalog populated by Glue.

Write and execute SQL queries against your datasets to analyze video game sales trends, genre performance, and other insights.

**SQL Queries**

**Popular Genres by Global Sales:** Identifying the most successful genres worldwide, highlighting consumer preferences.

****

**Highest Critic Scores:** Filtering games with top reviews to understand the impact of critical acclaim on sales.

**A close-up of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Genre Performance Across Platforms and Over Time:** Examining how different genres fare on various gaming platforms and their popularity trends.

**A close-up of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Regional Genre Performance:** Comparing genre success in different geographic regions, offering insights into regional market tastes.

**A close-up of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Top Performing Games in Each Genre: S**howcasing standout titles within each genre for a focused look at market leaders

**A screenshot of a computer

Description automatically generated**

**Predictive Modeling**

**Data Preprocessing:**

The dataset used for training and testing the model consists of video game sales information, including details on platform, genre, release year, and publisher. Missing values in the 'Year' column were imputed with the mean value. Additionally, the 'Publisher' column was transformed, categorizing publishers with fewer than 50 occurrences as 'Small Publisher.'

Categorical variables such as 'Platform,' 'Genre,' and 'Publisher' were one-hot encoded to represent them numerically. Numerical features were standardized using the StandardScaler from scikit-learn.

**Model Architecture:**

The neural network model comprises an input layer with 91 neurons (after one-hot encoding), two hidden layers with 128 neurons each, and an output layer with 1 neuron for regression purposes. The model is compiled with the mean squared error (MSE) loss function and RMSprop optimizer.

**Model Training:**

The dataset was split into training and validation sets, with 80% of the data used for training. The model was trained for 18 epochs, and the training process was visualized using a plot displaying the training and validation loss over epochs.

**Model Evaluation:**

The trained model can now be used for predicting global video game sales. Predictions are made on a test set, and performance is evaluated using the mean squared error (MSE) metric. This metric provides insights into how well the model generalizes to unseen data.

A graph showing the growth of a company

Description automatically generated with medium confidence

X-axis (Epoch): This axis represents the number of training epochs, which are iterations over the entire dataset. Each point on the x-axis corresponds to one complete pass through your training data.

Y-axis (Loss): The y-axis represents the loss, specifically the mean squared error (MSE) in your case. The loss is a measure of how well the model is performing on the training and validation sets. It quantifies the difference between the predicted values and the actual target values.

Training Loss (Blue Line): The blue line shows the training loss at each epoch. This is the value of the loss function computed on the training set. The goal during training is to minimize this loss, indicating that the model is learning and improving its predictive capability.

Validation Loss (Orange Line): The orange line shows the validation loss at each epoch. This is the value of the loss function computed on a separate validation set, which the model has not seen during training. The validation loss helps to assess how well the model generalizes to new, unseen data.

Legend: The legend in the upper left corner helps identify which line corresponds to training loss and which corresponds to validation loss.

The purpose of monitoring these losses over epochs is to identify patterns that might indicate issues like overfitting (if training loss keeps decreasing but validation loss starts increasing) or underfitting (if both losses remain high). It helps you understand the training progress and the model's ability to generalize to new data.

**Results:**

A screenshot of a computer

Description automatically generated**Dashboard and Visualization**

The dashboard presents a comprehensive analysis of video game sales, highlighting key metrics such as the number of games, genres, platforms, publishers, and sales over the years. It features detailed charts and visualizations, including sales by platform, critic scores for top games, and global sales distribution. Notably, "**Super Mario Bros**." emerges as the top-selling game, showcasing its significant impact. This dashboard, powered by QuickSight, effectively synthesizes vast data into accessible insights, aiding stakeholders in understanding market trends and making informed decisions.

**A screenshot of a computer

Description automatically generated**

**Conclusion**

The comprehensive analysis, leveraging AWS services and advanced data processing techniques, provided deep insights into the video game industry's sales trends, consumer preferences, and market dynamics. By examining datasets, applying predictive modeling, and visualizing findings through QuickSight dashboards, we've uncovered valuable patterns that can inform strategic decisions. This project not only demonstrates the power of cloud computing in handling complex data analyses but also highlights the potential for data-driven approaches to anticipate market shifts and guide industry stakeholders towards informed, strategic decisions.