Contents

[Cloud Data Pipeline Design Report for Video Game Sales Analysis 1](#_Toc132198790)

[1. Introduction 1](#_Toc132198791)

[2. Overall Concept Design 1](#_Toc132198792)

[3. Cloud Service Provider Selection 2](#_Toc132198793)

[4. Platforms and Components 2](#_Toc132198794)

[Data Ingestion: 2](#_Toc132198795)

[ETL Transformation: 2](#_Toc132198796)

[Storage of Data for Analytics: 3](#_Toc132198797)

[Analysis and Visualization: 3](#_Toc132198798)

[5. Pipeline Diagram and Component Overview 3](#_Toc132198799)

[6. Conclusion 4](#_Toc132198800)

# 

# Cloud Data Pipeline Design Report for Video Game Sales Analysis

## 1. Introduction

The dataset used in this project contains historical video game sales data, including game titles, platforms, release years, genres, publishers, regional sales, and other relevant information. The purpose of the project is to analyze the dataset, identify trends and patterns in video game sales, and generate insightful visualizations for decision-making in the gaming industry. The motivation for creating a cloud data pipeline is to enable a scalable and efficient solution that can handle large amounts of data, provide robust analytical capabilities, and offer seamless integration with various cloud services.

## 2. Overall Concept Design

The original format of the dataset is CSV, with columns such as Name, Platform, Year\_of\_Release, Genre, Publisher, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Global\_Sales, Critic\_Score, Critic\_Count, User\_Score, User\_Count, Developer, and Rating. The data schema can be represented as a table with these fields.

During the ETL process, the following transformations will be applied:

* Filter out records with missing or incomplete data
* Normalize and standardize text fields (e.g., case conversion, whitespace trimming)
* Convert year of release to a date format for time-based analysis

The potential analyses and visualizations to be performed include:

* Top-selling games by region and globally
* Sales trends over time
* Market share analysis by platform, genre, and publisher
* Correlation between critic/user scores and sales

## 3. Cloud Service Provider Selection

For this project, we have chosen Google Cloud Platform (GCP) as the cloud service provider. GCP offers a wide range of services for data ingestion, storage, processing, and analytics. Additionally, GCP provides a strong ecosystem for integrating various components of the pipeline and offers comprehensive documentation and support for implementing solutions.

## 4. Platforms and Components

### Data Ingestion:

**Google Cloud Storage (GCS):** Used for storing raw CSV files and processed data in a scalable and secure storage solution.

For data ingestion and storage, we chose Google Cloud Storage (GCS) due to its durability, availability, and performance. GCS is a highly scalable and cost-effective storage solution, allowing us to store large volumes of data in various formats, including CSV and JSON. The choice of GCS ensures that our solution can handle the increasing volume of data as the gaming industry continues to grow.

### ETL Transformation:

**Cloud Dataflow**: A fully managed service for processing and transforming the data in the pipeline. We will use Apache Beam with Python SDK for defining the ETL logic.

The justification for using Cloud Dataflow with Apache Beam and Python SDK in our pipeline is based on several key factors. First, Cloud Dataflow is a fully managed service that can scale automatically to handle the data processing workload, making it ideal for handling large-scale datasets. This ensures that the data pipeline will be able to handle varying data volumes without manual intervention.

Second, Apache Beam provides a unified programming model that allows us to define and execute data processing pipelines, irrespective of the underlying execution engine. This ensures that the ETL logic can be easily adapted to other processing engines or cloud platforms if needed.

Third, Python is a widely-used programming language with extensive support for data processing libraries, making it a popular choice for developing ETL pipelines. Using the Python SDK for Apache Beam enables us to leverage existing Python skills and libraries, streamlining development and maintenance.

### Storage of Data for Analytics:

**BigQuery**: A serverless, highly scalable, and cost-effective data warehouse designed for analytics. We will store the transformed data in BigQuery tables for efficient querying and analysis.

Google Cloud Dataflow was selected for the ETL process due to its serverless, fully managed nature, and its ability to handle both batch and streaming data processing. Cloud Dataflow provides an auto-scaling feature, which allows the pipeline to adjust its resources according to the data volume and processing requirements. This ensures that the pipeline remains cost-effective while delivering high performance, even when dealing with large datasets or complex transformations.

### Analysis and Visualization:

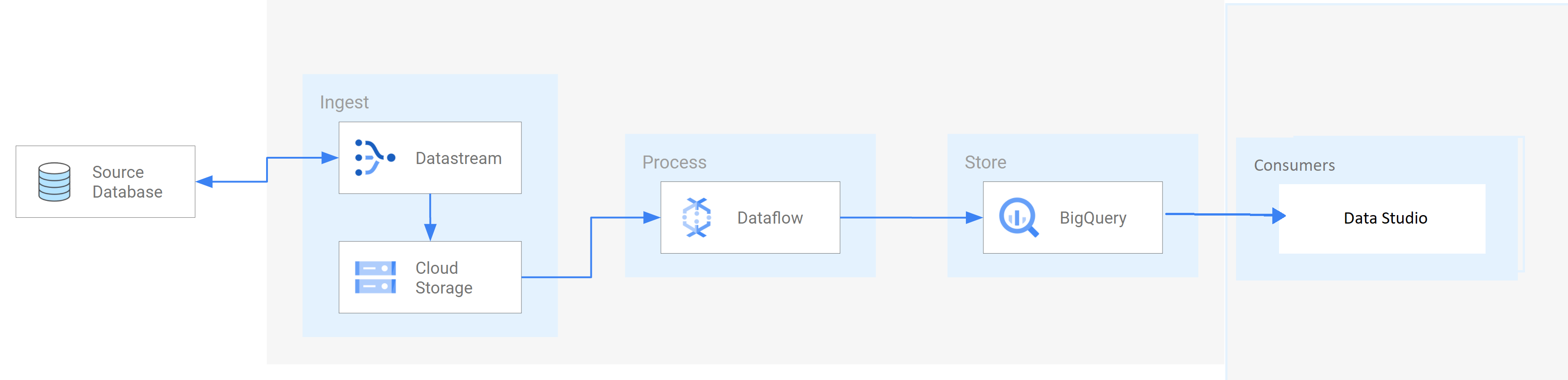
**BigQuery**: We will use SQL queries to perform various analyses, such as aggregations, filtering, and projections on the stored data.

BigQuery was chosen as the storage solution for analytics due to its serverless, highly scalable, and cost-effective architecture. BigQuery's columnar storage format and massive parallel processing capabilities enable it to perform complex analyses and aggregations on large datasets with ease. Furthermore, BigQuery's integration with other GCP services, such as Data Studio and Cloud Dataflow, ensures a seamless and efficient data analysis process.

**Data Studio**: A visualization tool that connects to BigQuery, allowing us to create interactive and shareable dashboards and reports based on the analytical results.

For data visualization, we chose Google Data Studio, a powerful and user-friendly tool that enables us to create interactive and insightful visualizations. Data Studio integrates seamlessly with BigQuery, allowing for efficient and real-time data exploration. The choice of Data Studio ensures that our pipeline delivers high-quality visualizations that are easily accessible and shareable, enabling stakeholders to make data-driven decisions.

## 5. Pipeline Diagram and Component Overview



**Google Cloud Storage**: Stores the raw data and acts as a source and sink for Dataflow during the ETL process.

**Cloud Dataflow**: Reads data from GCS, applies the necessary transformations, and writes the transformed data back to GCS or directly into BigQuery.

**BigQuery**: Stores the transformed data in an optimized format for analytics and serves as the data source for Data Studio.

**Data Studio**: Connects to BigQuery to generate visualizations and dashboards based on the analyzed data.

By implementing this cloud data pipeline on GCP, we can efficiently store, process, analyze, and visualize our video game sales dataset to gain valuable insights and make informed decisions in the gaming industry.

## 6. Conclusion

In conclusion, the cloud data pipeline design proposed in this report leverages the power of Google Cloud Platform (GCP) to create an efficient, scalable, and robust solution for analyzing video game sales data. By utilizing GCP's comprehensive suite of services, such as Google Cloud Storage, Cloud Dataflow, BigQuery, and Data Studio, we have designed a pipeline that addresses the challenges of data ingestion, storage, processing, and visualization.

The ETL process is built using Cloud Dataflow, which provides the flexibility to apply necessary transformations and ensures data quality. The transformed data is then stored in BigQuery, a highly scalable and cost-effective data warehouse designed for analytics. BigQuery enables us to perform complex analyses and aggregations on large datasets with ease, providing a solid foundation for data-driven decision-making.

Furthermore, Data Studio is used to create insightful and interactive visualizations, allowing stakeholders to explore the data and identify trends, patterns, and opportunities within the gaming industry. The seamless integration between BigQuery and Data Studio ensures an efficient and user-friendly experience for data analysis and reporting.

The proposed cloud data pipeline not only builds upon the prototype implemented in Databricks but also extends its capabilities by leveraging the power of GCP's managed services. This design ensures that the pipeline is both flexible and adaptable, accommodating future changes in data volume, variety, or analytical requirements. Additionally, the pipeline's modularity allows for easy integration with other GCP services or third-party tools, further enhancing its capabilities and utility.

In summary, the cloud data pipeline design for video game sales analysis provides a comprehensive solution that harnesses the power of GCP to deliver valuable insights and empower data-driven decision-making in the gaming industry.