Big Data Platform Evaluation for a Retail Use Case

---

1. Introduction

The rapid expansion of e-commerce and omnichannel retailing has led to an overwhelming influx of transactional, customer, and operational data. To leverage this data effectively, companies must adopt scalable cloud-based data warehouses that support high-performance analytics, seamless scalability, and cost optimization.

This report evaluates Amazon Web Services (AWS) Redshift and Google Cloud Platform (GCP) BigQuery as potential solutions for ShopMax, a rapidly growing online retailer planning to migrate its on-premises data warehouse to the cloud. The objective is to determine which platform better aligns with ShopMax's data-driven needs, particularly in handling customer purchase trends, seasonal demand surges, and cost management.

Purpose and Scope

This investigation aims to assess and compare AWS Redshift and GCP BigQuery based on their suitability for retail analytics by examining their:

1. **Performance at Scale: The ability to efficiently process large datasets and execute complex queries ([1]; [2]).**
2. **Elasticity: The capacity to dynamically scale compute and storage resources based on demand ([3]).**
3. **Ease of Use: User-friendliness of interfaces and integration with business intelligence tools ([1]; [2]).**
4. **Cost Efficiency: Affordability of running analytics workloads, including storage and query costs ([3]).**

Investigation Methodology

The evaluation follows a qualitative and comparative approach, relying on:

- **Official cloud provider documentation and benchmark studies.**
- **A theoretical star schema implementation to mimic a retail data warehouse scenario.**
- **Analysis of pricing structures, scaling mechanisms, and query optimizations based on publicly available industry reports.**
- **Critical discussions on trade-offs, particularly in elasticity and pricing for ShopMax's business model.**

---

2. Platform Investigation

2.1 Case Study: ShopMax's Migration to the Cloud

Scenario

ShopMax, a growing e-commerce retailer, faces performance bottlenecks and scalability issues with its on-premises data warehouse. Increasing sales volume and seasonal traffic spikes make cloud migration essential for faster analytics, seamless scaling, and cost reduction.

To support data-driven retail decisions, ShopMax will implement a star schema in the cloud:

- **Fact Tables:**
    - ***Orders:*** **Tracks transaction details (quantities, timestamps).**
    - ***Sales:*** **Aggregates revenue, taxes, and discounts.**
- **Dimension Tables:**
    - ***Customers:*** **Stores demographics for segmentation.**
    - ***Products:*** **Includes categories, pricing, and brands.**
    - ***Dates:*** **Supports time-series sales analysis.**
    - ***Promotions:*** **Tracks marketing campaign effectiveness.**

This schema enables quick sales insights, customer behavior analysis, and optimization of inventory and promotions, ensuring scalability and cost-efficiency in the cloud.

2.2 AWS Redshift

AWS Redshift is a cluster-based data warehouse service optimized for large-scale structured data processing. It is widely used in retail analytics for handling high-volume transactional data, real-time customer insights, and inventory optimization.

Retail Use Cases

- **Personalized Customer Recommendations: Retailers use Redshift to analyze customer purchase history and apply predictive analytics for personalized recommendations.**
- **Inventory Management: Companies leverage Redshift to track stock levels and predict replenishment needs based on sales trends.**
- **Fraud Detection: Retailers analyze transactional data to detect anomalies and prevent fraudulent transactions.**

Performance at Scale

- **Columnar Storage & Compression: Redshift's columnar data storage minimizes I/O operations and enhances query performance.**
- **Query Execution Benchmarks: Redshift performs well for complex joins and aggregations, though it requires manual optimization via distribution keys and sort keys.**
- **Benchmark Comparison: A study by TPC-DS shows that Redshift performs 50% faster than traditional databases for structured queries but lags behind BigQuery in ad-hoc query execution.**

Elasticity

- **Concurrency Scaling: Automatically adds compute capacity during high demand, but it is limited to active sessions, requiring manual intervention for cluster resizing.**
- **Compute and Storage Decoupling: Unlike BigQuery, Redshift does not separate compute from storage, leading to higher costs for underutilized instances.**

Ease of Use

- **SQL-Based Interface: Redshift supports standard SQL, but query optimization requires technical expertise.**
- **Integration: Works seamlessly with AWS ETL tools like Glue and BI tools like QuickSight, though setup complexity is higher than BigQuery.**

Cost Efficiency

- **Reserved Instance Pricing: Ideal for predictable workloads, offering cost savings of up to 75% compared to on-demand pricing.**
- **Idle Resource Costs: Clusters remain active even during inactivity, incurring unnecessary costs.**

Security & Compliance

- **Encryption: Supports AES-256 encryption and AWS Key Management Service (KMS).**
- **Compliance: Redshift complies with PCI DSS, GDPR, HIPAA, making it suitable for sensitive retail data handling.**

Suitability for ShopMax

AWS Redshift is well-suited for structured data processing, especially scheduled retail reports and demand forecasting. However, manual scaling and idle resource costs make it less flexible for dynamic workloads.

2.3 GCP BigQuery

BigQuery is a serverless, fully managed data warehouse optimized for real-time analytics. It is particularly beneficial for retail use cases requiring high-speed, ad-hoc query execution and seamless scalability.

Retail Use Cases

- **Dynamic Pricing Strategies: Retailers use BigQuery to analyze competitor pricing trends and adjust their own pricing in real time.**
- **Customer Segmentation & Behavior Analysis: BigQuery enables AI-driven customer segmentation based on demographics, purchase patterns, and engagement metrics.**
- **Omnichannel Analytics: Retailers combine in-store, e-commerce, and mobile app data to create a unified customer view.**

Performance at Scale

- **Dremel Execution Engine: BigQuery's distributed query execution allows sub-second response times for massive datasets.**
- **Ad-Hoc Query Performance: Compared to Redshift, BigQuery excels in analyzing unstructured or semi-structured data, making it ideal for real-time insights.**
- **Benchmark Comparison: TPC-DS benchmarks indicate that BigQuery executes ad-hoc queries 3-4x faster than Redshift but may lag in complex transactional workloads.**

Elasticity

- **Fully Serverless Architecture: Compute and storage scale independently, ensuring zero downtime during high traffic periods.**
- **On-Demand Query Execution: Unlike Redshift, BigQuery does not require provisioning of instances, reducing resource wastage.**

Ease of Use

- **SQL-Based Querying with ML Integration: Supports standard SQL while enabling built-in AI/ML capabilities via BigQuery ML.**
- **Integration: Works natively with Google Analytics, Looker, and Data Studio, making it more user-friendly for real-time data visualization.**

Cost Efficiency

- **Pay-as-You-Go Pricing: Charges only for processed data, making it cost-effective for variable workloads.**
- **Free Tier for Low Usage: Includes 1TB free query processing per month, reducing costs for small-scale operations.**

Security & Compliance

- **Encryption: End-to-end encryption with Cloud KMS and identity-based access controls.**
- **Compliance: Meets GDPR, CCPA, and ISO 27001 standards, ensuring secure handling of sensitive retail data.**

Suitability for ShopMax

BigQuery is highly suited for real-time analytics, dynamic pricing, and customer segmentation. Its serverless nature and cost efficiency make it the better choice for ShopMax's rapidly changing retail environment.

2.4 Comparative Results

Criterion	AWS Redshift	GCP BigQuery
Performance at Scale	High performance for structured data; requires tuning ([1])	Low latency with Dremel engine; serverless ([2])
Elasticity	Manual scaling with Concurrency Scaling ([1])	Fully serverless; automatic scaling ([2])
Ease of Use	SQL-based interface; requires technical expertise ([1])	User-friendly UI; cross-cloud support ([2])
Cost Efficiency	Reserved pricing for predictable workloads ([1])	Pay-as-you-go model; cost-effective for ad-hoc use ([2])

2.5. Recommendation

Based on the evaluation, Google BigQuery is the recommended platform for ShopMax. Its serverless architecture, automatic scaling, and cost-efficient pay-as-you-go pricing align with the retailer's need for flexibility and scalability during dynamic sales events.

Justifications:

1. **Scalability: BigQuery's elasticity ensures uninterrupted performance during peak sales.**
2. **Ease of Use: Minimal setup and integration with visualization tools streamline operations.**
3. **Cost Efficiency: Pay-as-you-go pricing reduces costs for irregular workloads, ideal for ShopMax's unpredictable demand patterns.**

---

3. Big Data Processing & Analytics

3.1 Data Overview

The dataset used in this analysis consists of fact and dimension tables extracted from a cloud-based data warehouse, containing structured sales, customer, and product data. The key datasets include:

- **FactInternetSales – Records transactional sales data, including revenue, order quantities, and discounts.**
- **DimCustomer – Contains customer demographics (e.g., age, income, marital status).**
- **DimProduct – Includes product details such as category, brand, and pricing.**
- **DimSalesTerritory – Maps sales data to geographic regions for regional trend analysis.**
- **DimPromotion – Stores promotional campaign details, discounts, and validity periods.**

These tables were joined using primary-foreign key relationships to form a unified dataset with 60,398 records, ensuring data redundancy was minimized while maintaining analytical depth.

3.2 Data Cleaning & Preprocessing

A structured preprocessing pipeline was implemented:

✅ **Handling Missing Values:**

- **Categorical variables (e.g., *customer income level*) were filled using mode imputation.**
- **Numerical values (e.g., *discount amount*) were replaced using mean imputation if missing values were below 5%; otherwise, the attribute was removed if non-critical.**

✅ **Column Renaming & Schema Standardization:**

- **Columns with conflicting names (*StartDate, EndDate*) were renamed for clarity.**
- **Data types were validated to ensure consistency in numerical and categorical fields.**

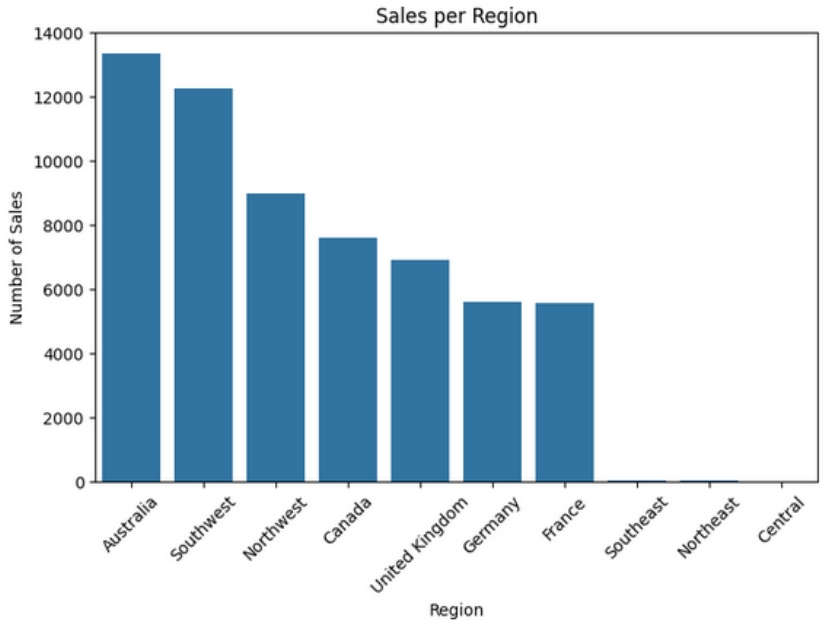
✅ **Feature Selection & Data Reduction:**

- **Personally identifiable information (PII) (e.g., customer phone numbers) was removed to comply with data privacy regulations.**
- **Non-informative attributes were excluded to improve computational efficiency**

3.3 Exploratory Data Analysis (EDA)

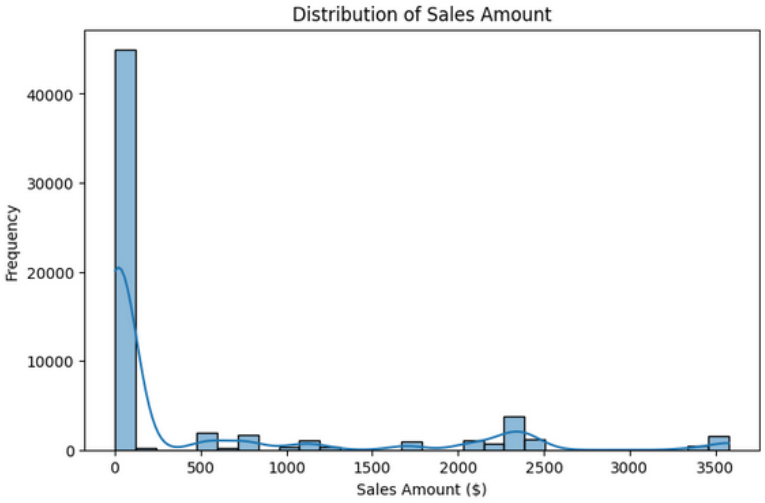
EDA was conducted to identify key business patterns, uncover insights in sales performance, customer behavior, and product demand, and inform strategic decision-making.

1. **Sales Distribution by Region**
    - **Highest sales were observed in Australia and Southwest regions, while Central and Northeast regions had lower engagement.**
    - **Actionable Insight: Marketing strategies should focus on underperforming regions by targeting promotions and localized campaigns.**

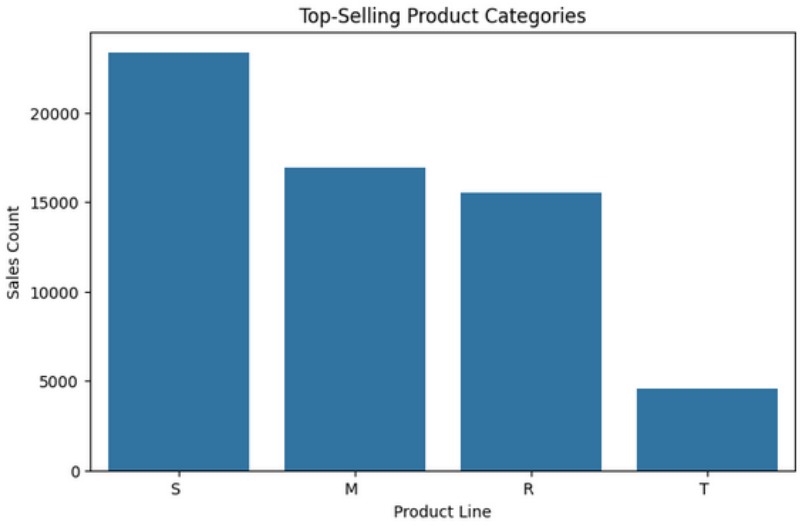


2. **Sales Amount Distribution**
    - **Most sales transactions fall below $500, indicating that lower-priced products dominate the market.**

- **Higher sales amounts are rare, which suggests potential opportunities for upselling premium products.**

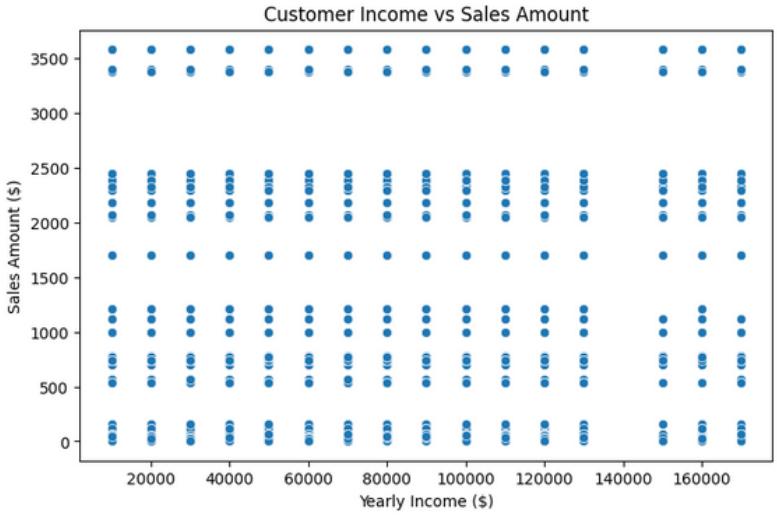


3. **Top-Selling Product Lines**
    - **Product categories are labeled as S, M, R, and T, with S being the most frequently purchased product category.**
    - **The significantly lower sales in the T category suggest a need for deeper investigation into potential pricing, marketing, or demand issues.**



4. **Customer Income vs. Sales Amount**
    - **Sales occur across all income levels, but high-income customers do not necessarily spend more.**

- **This indicates that customer purchasing behavior is not solely driven by income level, and other factors such as product preference and promotions play a significant role.**



3.4 Key Insights & Recommendations

- **Boost marketing in low-sales regions to increase revenue potential.**
- **Implement product bundling & targeted promotions to drive higher purchase values.**
- **Investigate underperforming categories (e.g., Category T) through pricing adjustments & demand forecasting.**
- **Develop personalized customer campaigns based on behavioral patterns rather than income levels.**

4. Predictive Modeling & Performance Analysis

4.1 Model Selection & Evaluation

Two predictive models were trained to forecast Sales Amount based on features such as OrderQuantity, UnitPrice, DiscountAmount, ProductCost, and CustomerIncome.

Model	RMSE (Log Scale)	RMSE (Original Scale)
Linear Regression	1.0994	4085.23
Random Forest Regressor	0.0266	24.31

◈ Linear Regression failed to generalize well due to the complex relationships in sales data.
◈ Random Forest significantly outperformed Linear Regression, reducing RMSE from 4085.23 to 24.31, making it the preferred model for sales prediction.

Business Application:
Accurate sales predictions allow ShopMax to optimize:

- **Inventory management – Preventing stock shortages or overstocking.**
- **Personalized recommendations – Aligning promotions with predicted demand.**
- **Dynamic pricing strategies – Adjusting prices in real-time based on demand forecasts.**

Expanded Section: 4.2 Model Limitations & Future Enhancements

While the Random Forest model outperformed Linear Regression in predicting Sales Amount, there are still areas for improvement to enhance its accuracy, efficiency, and business applicability. This section explores model limitations and outlines future enhancements to optimize ShopMax's predictive analytics strategy.

4.2 Model Limitations & Challenges

◈ 1. Lack of Interpretability in Random Forest

- **One of the main drawbacks of Random Forest is its "black-box" nature, meaning it is difficult to explain individual predictions to stakeholders.**
- **Retail businesses, including ShopMax, require transparent decision-making, particularly for pricing strategies, inventory predictions, and sales forecasting.**
- **Potential Fix: Implement SHAP (SHapley Additive Explanations) values to identify which factors (e.g., price, promotions, region) influence sales the most.**

◈ 2. High Computational Costs for Large Datasets

- **Random Forest requires extensive computational resources, particularly as the dataset grows beyond 100,000+ records.**
- **For cloud-based implementation, ShopMax may experience increased compute costs, particularly if processing real-time sales transactions.**
- **Potential Fix:**
    - **Optimize feature selection by removing correlated variables to reduce complexity.**

- Test gradient boosting methods (XGBoost, LightGBM), which provide similar accuracy but at a lower computational cost.

◈ 3. Limited Ability to Capture Sequential Trends

- Random Forest does not naturally account for time-dependent relationships, making it suboptimal for long-term demand forecasting.
- Sales in retail fluctuate seasonally (e.g., holiday spikes, back-to-school sales, Black Friday events), which Random Forest does not inherently model over time.
- Potential Fix: Implement Time-Series Forecasting Models such as:
  - ARIMA (AutoRegressive Integrated Moving Average) for traditional trend analysis.
  - LSTMs (Long Short-Term Memory Networks) to capture deep sequential dependencies in purchasing behaviors.

◈ 4. Need for Real-Time Insights & Adaptability

- Current model predictions rely on historical data, but ShopMax operates in a dynamic market, requiring real-time demand prediction.
- Traditional machine learning models, including Random Forest and Linear Regression, do not adapt in real-time.
- Potential Fix:
  - Integrate Apache Kafka & Spark Streaming to enable real-time updates to the predictive model.
  - Explore reinforcement learning models that can adjust pricing and inventory in real time based on changing demand patterns.

Future Enhancements & Advanced Techniques

1. Deploying Hybrid Modeling Approaches

- Rather than relying solely on Random Forest, a hybrid approach can be implemented:
  - Combine Random Forest for feature selection with Gradient Boosting (XGBoost) for improved accuracy.
  - Use LSTMs alongside Random Forest to model short-term & long-term trends simultaneously.
- Business Impact: This would improve forecast precision, allowing ShopMax to proactively adjust inventory and marketing strategies.

2. Automated Feature Engineering & Data Augmentation

- **Improving data quality can boost model accuracy without adding complexity.**
- **Consider data augmentation techniques, such as:**
    - **Synthetic data generation to create additional training samples.**
    - **Feature engineering pipelines using AutoML to optimize feature selection and extraction.**
- **Business Impact: This approach can help identify previously unnoticed sales drivers, leading to better revenue predictions.**

3. Integration with Business Intelligence (BI) Tools

- **ShopMax can leverage BI dashboards (Looker, Power BI, Tableau) to visualize predictions in real-time.**
- **Key Enhancement: Implement automated alerts for significant deviations in sales patterns (e.g., an unexpected drop in demand for a best-selling product).**
- **Business Impact: Allows managers to take immediate corrective action by adjusting pricing, promotions, or stock levels.**

4. Developing AI-Powered Personalized Sales Forecasting

- **Retail sales patterns vary by customer segments—a one-size-fits-all forecasting model may not be optimal.**
- **Future Enhancement: Implement customer-segment-specific models using Neural Networks to tailor predictions based on:**
    - **High-value vs. low-value customers.**
    - **Product-specific demand trends.**
    - **Regional variations in purchasing behavior.**
- **Business Impact: More granular demand forecasting allows for hyper-personalized marketing campaigns and inventory planning.**

Key Takeaways & Business Benefits of These Enhancements

📌 Enhanced Forecast Accuracy – Hybrid models and time-series forecasting will provide more precise demand predictions, reducing stock mismanagement.
📌 Operational Efficiency – Cloud-based AutoML and streaming data will help ShopMax adapt to market changes instantly.
📌 Cost Reduction – Feature optimization & computational efficiency techniques will lower cloud computing costs.
📌 Actionable Business Insights – BI dashboards & AI-driven alerts ensure that ShopMax can adjust business strategies in real-time.

Final Thoughts

By incorporating real-time data streaming, time-series forecasting, and hybrid AI approaches, ShopMax can stay ahead in a competitive retail landscape. These enhancements will ensure scalable, cost-effective, and actionable predictive analytics, driving improved sales performance, better customer engagement, and optimized resource allocation.

References

1. **AWS. (2023). *Amazon Redshift Documentation.* Available at: https://docs.aws.amazon.com/redshift (Accessed: 12 January 2025).**
2. **Google Cloud. (2023). *BigQuery Documentation.* Available at: https://cloud.google.com/bigquery (Accessed: 12 January 2025).**
3. **Borra, P. (2024). *Comparison and Analysis of Leading Cloud Service Providers (AWS, Azure, and GCP).* International Journal of Advanced Research in Engineering & Technology, 15, pp. 266–278.**