BIG DATA PROCESSING

efthimis mavridis

MAVRIDIS EFTHYMIOS

2024

UNIVERSITY OF DERBY ID : 100731060

Abstract:

This article presents a comprehensive analysis comparing Amazon Web Services (AWS) and Microsoft Azure Platform (MAP) in the context of Big Data management and analytics. The study encompasses fundamental statistical insights, variable relationships, and employs the Random Forest Classifier model for predictive analytics. The evaluation of AWS highlights its reliability, scalability, and diverse services, with a focus on Amazon Redshift for efficient data warehousing. In contrast, MAP, built on Microsoft's technological expertise, offers enterprise-grade solutions, emphasizing Azure Synapse Analytics as a robust Data Warehouse service. Evaluation criteria include Performance at Scale, Elasticity, Ease of Use, and Cost Efficiency. The article further explores sales prediction using ensemble learning techniques, featuring models such as Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier. Optimization efforts using GridSearchCV reveal improved performance, with the Gradient Boosting Classifier achieving the highest accuracy at 62.11%. Python, Pandas, and PySpark serve as primary tools for data management, reflecting their efficiency in data manipulation and large-scale processing. The data preprocessing phase involves variable conversion, exploratory data analysis (EDA), and the implementation of a Random Forest Classifier for sales prediction.The article concludes with a summary outlining the key features of AWS and MAP, experiences gained throughout the project, and suggestions for future work. The project serves as a foundational exploration into cloud computing platforms and machine learning techniques for data analytics, contributing to ongoing research in the dynamic landscape of data management and analytics.

**Table of contents**

[1)Introduction 2](#_Toc157594109)

[i)Amazon Web Services (AWS): 2](#_Toc157594110)

[ii)Microsoft Azure Platform (MAP): 2](#_Toc157594111)

[2)Criteria for Evaluation: 3](#_Toc157594112)

[2.1 Scale(AWS VS MAP) 3](#_Toc157594113)

[AWS: 3](#_Toc157594114)

[MAP: 4](#_Toc157594115)

[2.2 Elasticity in AWS and Microsoft Azure 5](#_Toc157594116)

[1. Elasticity in AWS: 5](#_Toc157594117)

[2. Elasticity in Microsoft Azure: 6](#_Toc157594118)

[In both AWS and Microsoft Azure: 6](#_Toc157594119)

[3) Empowering Data Insights: A Journey through Processing, Exploration, and Ensemble Learning for Sales Prediction 6](#_Toc157594120)

[3.1 INTRODUCTION: 6](#_Toc157594121)

[3.2 Data Preprocessing and Sales Prediction Model Construction 7](#_Toc157594122)

[3.2 Exploring the Initial Records and Applying Exploratory Data Analysis (EDA) Techniques 8](#_Toc157594123)

[3.3 Ensemble Learning for Sales Amount Prediction 10](#_Toc157594124)

[3.4 Optimizing Sales Amount Prediction with Ensemble Models 13](#_Toc157594125)

[4) Sum up 14](#_Toc157594126)

5)[References 15](#_Toc157594127)

# 1)Introduction

In the contemporary landscape of data management and analytics, cloud computing platforms emerge as pivotal players. Analyzing two leading platforms, Amazon Web Services (AWS) and Microsoft Azure Platform (MAP), we focus on addressing the challenges posed by the escalating complexity of Big Data.

Our study encompasses fundamental statistical insights, an analysis of relationships between variables, and prediction through the Random Forest Classifier model. Beyond the technical aspect, we explore critical questions, such as the distribution of YearlyIncome, the impact of age on SalesAmount, and more. This comprehensive analysis framework enables a thorough comparison of the platforms, providing valuable insights into their respective capabilities and suitability for diverse analytical tasks.

## i)Amazon Web Services (AWS):

At the forefront of cloud computing innovation, AWS, an offspring of e-commerce giant Amazon, has transformed the way businesses approach data. Renowned for its reliability and scalability, AWS provides an extensive suite of services that spans computing power, storage, and analytics. Among its standout offerings is Amazon Redshift, a powerful Data Warehouse service celebrated for its capability to seamlessly handle large datasets and facilitate real-time analytics. With a global infrastructure that ensures low-latency access to resources, AWS has become the preferred choice for organizations seeking a robust and agile cloud solution. Its commitment to continuous innovation and a vast user base further solidify AWS as a leader in the cloud computing landscape.

## ii)Microsoft Azure Platform (MAP):

Built on the bedrock of Microsoft's technological expertise, MAP is a stalwart in the cloud computing arena, offering a comprehensive suite of services designed for enterprises. Acknowledged for its enterprise-grade solutions, MAP seamlessly integrates infrastructure, data storage, and advanced analytics tools. Azure Synapse Analytics, a key component of MAP, stands as a formidable Data Warehouse solution, empowering organizations to efficiently process and analyze massive datasets. MAP's strength lies in its commitment to interoperability and adaptability, catering to the evolving needs of businesses in a rapidly changing technological landscape. With a focus on hybrid cloud scenarios, MAP emerges as a flexible and holistic cloud platform, appealing to organizations aiming to strike a balance between on-premises and cloud environments.

# 2)Criteria for Evaluation:

As we embark on a comparative evaluation of AWS and MAP, four pivotal criteria come to the forefront: Performance at Scale, Elasticity, Ease of Use, and Cost Efficiency. Through a meticulous analysis of these criteria, we aim to provide organizations with nuanced insights into the strengths and considerations of each platform. This exploration serves as a compass for enterprises navigating the complex terrain of Big Data processing and management,

guiding them towards a judicious selection that aligns with their unique business requirements and aspirations.

## 2.1 Scale(AWS VS MAP)

### AWS:

Amazon Web Services (AWS) offers a wide range of cloud computing services to help businesses scale and grow. The AWS ecosystem is extensive and includes infrastructure services, platform services, and software services. While I can't provide every detail, I can give you an overview of some key components and services.

#### 1. Compute Services:

- Amazon EC2 (Elastic Compute Cloud): Provides scalable virtual servers in the cloud.

- Amazon Lambda: Allows you to run code without provisioning or managing servers.

#### 2. Storage Services:

- Amazon S3 (Simple Storage Service): Object storage for storing and retrieving any amount of data.

- Amazon EBS (Elastic Block Store): Provides block-level storage volumes for use with EC2 instances.

#### 3. Database Services:

- Amazon RDS (Relational Database Service): Managed relational database service supporting multiple database engines.

- Amazon DynamoDB: Fully managed NoSQL database service.

#### 4. Networking

- Amazon VPC (Virtual Private Cloud): Lets you provision a logically isolated section of the AWS Cloud.

- Amazon Route 53: A scalable domain name system (DNS) web service.

#### 5. Security and Identity:

- AWS IAM (Identity and Access Management): Manages access to AWS services and resources securely.

- Amazon Inspector: Automated security assessment service to help improve the security and compliance of applications.

#### 6. Management Tools:

- AWS CloudWatch: Monitoring service for AWS resources and applications.

- AWS CloudTrail:Records AWS API calls for your account and delivers log files to you.

#### 7. Developer Tools:

- AWS CodeDeploy: Automates code deployments to any instance, including EC2 instances and on-premises servers.

- AWS CodeCommit: A fully-managed source control service.

### MAP:

In the context of the Microsoft Azure platform, "scale" generally refers to the ability to dynamically adjust the computing resources allocated to an application or service based on demand. Azure provides several features and services to help users scale their applications effectively. Here are key aspects related to scale in Microsoft Azure:

#### 1.Virtual Machines (VM) Scaling:

- Azure Virtual Machines allow users to scale vertically (resizing a VM) or horizontally (increasing the number of VM instances).

- Azure Virtual Machine Scale Sets enable the automatic scaling of a set of identical VMs.

#### 2. App Service Scaling:

- Azure App Service, which includes Web Apps, API Apps, Mobile Apps, and Function Apps, provides automatic scaling options.

- Users can manually scale the App Service Plan or configure autoscaling rules based on metrics like CPU utilization or the number of requests.

#### 3.Azure Kubernetes Service (AKS):

- AKS allows for the deployment, management, and scaling of containerized applications using Kubernetes.

- Users can adjust the number of nodes in a cluster based on workload requirements.

#### 4. Azure Functions:

- Azure Functions allow for serverless computing, automatically scaling based on demand.

- Users are billed based on actual usage, and functions scale out automatically to handle increased load.

#### 5. Azure SQL Database:

- Azure SQL Database can be automatically scaled up or down based on workload requirements.

- Users can configure performance levels (DTUs or vCores) to match the application's needs.

#### 6. Azure Autoscale:

- Azure Autoscale enables automatic scaling of resources based on predefined rules.

- It supports multiple Azure services, allowing users to define scaling conditions and actions.

In summary, Azure provides a comprehensive set of tools and services for scaling applications and services to meet changing demands. Whether you need to scale virtual machines, applications, databases, or other resources, Azure offers flexibility and automation options to ensure efficient and cost-effective scaling.

## 2.2 Elasticity in AWS and Microsoft Azure

### 1. Elasticity in AWS:

- Auto Scaling: AWS Auto Scaling allows you to automatically adjust the number of EC2 instances in your application based on demand. It helps maintain application availability and allows you to scale in or out automatically.

- Amazon EC2 Instances: AWS provides a wide range of EC2 instance types that cater to different workloads, and users can easily scale vertically or horizontally based on their requirements.

- Managed Services: AWS offers managed services like Amazon RDS, Amazon DynamoDB, and others that automatically scale based on demand.

### 2. Elasticity in Microsoft Azure:

- Auto Scaling: Azure AutoScale enables users to automatically adjust the number of compute resources in response to demand or a defined schedule. This can be applied to Virtual Machine Scale Sets, Cloud Services, and other resources.

- Virtual Machines: Similar to AWS, Microsoft Azure provides a variety of VM sizes, allowing users to scale vertically. Azure Virtual Machine Scale Sets facilitate horizontal scaling by automatically adjusting the number of VM instances.

- App Service Scaling: Azure App Service offers automatic scaling options, allowing users to adjust resources based on metrics like CPU utilization or the number of requests.

### In both AWS and Microsoft Azure:

- Pay-as-You-Go Model: Elasticity is closely tied to the pay-as-you-go model, where users pay for the resources they consume. This allows for cost optimization and efficient use of resources.

- Managed Services: Both platforms offer managed services that abstract the underlying infrastructure, automatically adjusting resources based on demand.

Elasticity in cloud computing refers to the ability to dynamically scale resources up or down based on workload requirements. This ensures optimal performance, cost-efficiency, and the ability to handle varying levels of demand. Whether you're using AWS or Microsoft Azure, both platforms provide a range of tools and services to achieve elasticity in your applications and infrastructure.

# 3) Empowering Data Insights: A Journey through Processing, Exploration, and Ensemble Learning for Sales Prediction

## 3.1 INTRODUCTION:

During the course of the project, Python served as the primary programming language, complemented by the Pandas and PySpark libraries. The selection of these tools reflects a strategic approach to efficiently manage data while ensuring ease of use and effectiveness.Pandas was employed for data manipulation and restructuring in the form of DataFrames. The flexibility of Pandas impressed with its ability to execute swift data analyses and processing.Additionally, I delved into the realm of Big Data utilizing PySpark, a robust library supporting data processing at a scalable level. PySpark's functionalities allowed me to address challenges associated with large datasets and distributed processing.In summary, the utilization of Python, Pandas, and PySpark stands as a pivotal component of the data processing journey, enhancing the ability to extract valuable insights and provide meaningful analyses.

## 3.2 Data Preprocessing and Sales Prediction Model Construction

This section of the code represents the data preprocessing phase and the construction of a machine learning model. Let's break down each step individually:

**1. Conversion of the 'BirthDate' field into age:**

- The 'BirthDate' column is used to calculate the age of individuals.

- The 'BirthDate' column is converted into a date object.

- Age is calculated by subtracting the birthdate from the current date and dividing the result by the number of days in a year.

**2. Conversion of the 'SalesAmount' column into a numeric value:**

- The 'SalesAmount' column is converted into a numeric value.

- Any non-numeric values are replaced with NaN (Not a Number).

**3. Exploratory Data Analysis (EDA):**

- A pair plot is displayed, illustrating the relationships between various features such as 'YearlyIncome,' 'TotalChildren,' 'NumberCarsOwned,' and 'SalesAmount.'

- This plot helps visualize the relationships and variations among different variables.

**4. Construction of a Machine Learning Model (Random Forest Classifier):**

- Selection of features and target category.

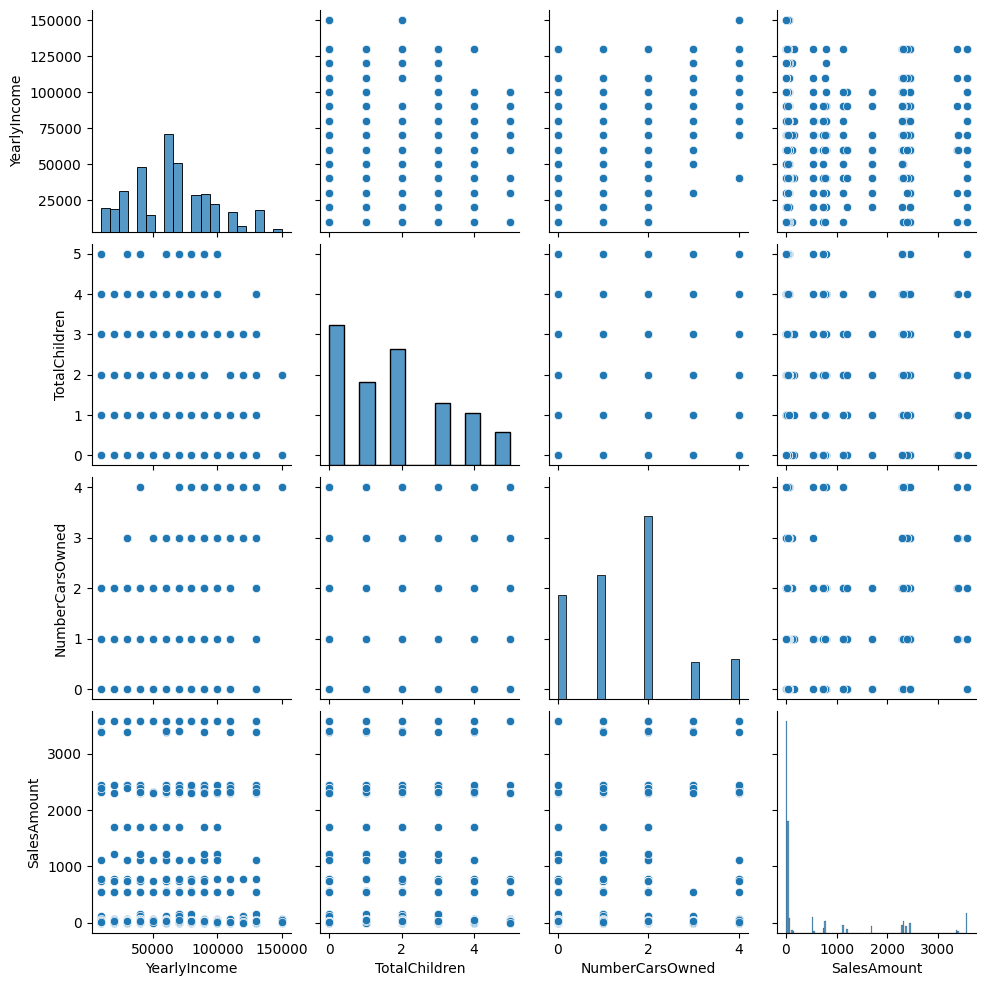
- Creation of a DataFrame with selected features ('YearlyIncome,' 'TotalChildren,' 'NumberCarsOwned,' 'Age') and the target ('SalesAmount' > median).

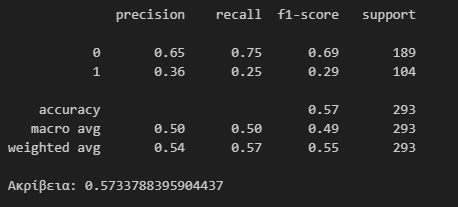
- Splitting the data into training and testing sets.

- Building a Random Forest Classifier model.

- Prediction of results and evaluation of the model.

Overall, these steps constitute a common process for data preprocessing and the construction of a machine learning model for predicting sales ('SalesAmount').





With 57% accuracy.

## 3.2 Exploring the Initial Records and Applying Exploratory Data Analysis (EDA) Techniques

In this section of the code, we perform an initial exploration of the DataFrame by employing libraries such as Matplotlib, Seaborn for visualizations, and Pandas for statistical metrics:

**1. Displaying the First Records of the DataFrame:**

- Utilizing Pandas, we print the first few records of the DataFrame using `df.head()`.

**2. Application of EDA Techniques:**

- **Statistical Metrics Calculation:**

- We use `df.describe()` to compute key statistical metrics for the dataset, providing insights into central tendencies and dispersion of numerical variables.

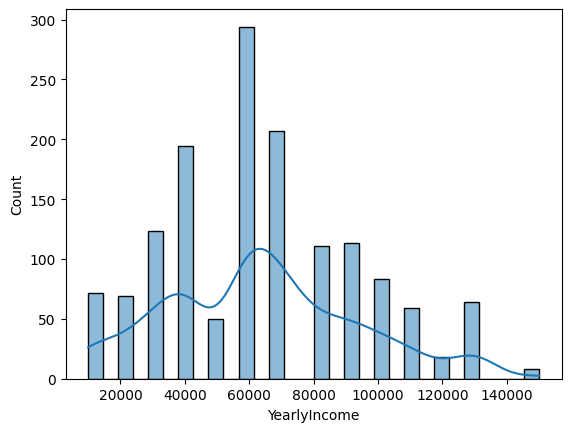
- **Exploring Relationships Between Variables:**

- A Seaborn pair plot (`sns.pairplot()`) is employed to visualize relationships between selected variables ('YearlyIncome,' 'TotalChildren,' 'NumberCarsOwned,' 'Age'). This plot assists in identifying patterns and correlations.

**- Distribution of a Specific Variable:**

- Using Seaborn, a histogram (`sns.histplot()`) is created to illustrate the distribution of values for the 'YearlyIncome' variable. This visualization aids in understanding the frequency distribution within the dataset.

This code snippet focuses on the initial exploration of the dataset, employing various EDA techniques. It includes the display of initial records, computation of statistical metrics, visualization of variable relationships through a pair plot, and the examination of the distribution of a specific variable.



## 3.3 Ensemble Learning for Sales Amount Prediction

In this analysis, three different machine learning models were employed to predict whether the sales amount surpasses the median value. Here's a summary of the models and their performance:

**1. Random Forest Classifier:**

- Model Accuracy: 59.04%

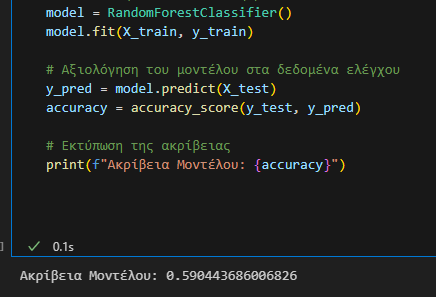
- Code snippet:

from sklearn.ensemble import RandomForestClassifier

# ... (Data preparation and model training)

print(f"Random Forest Classifier Accuracy: {accuracy}")

- The Random Forest Classifier uses multiple decision trees to enhance predictive accuracy.



**2. Decision Tree Classifier:**

- Model Accuracy: 60.4%

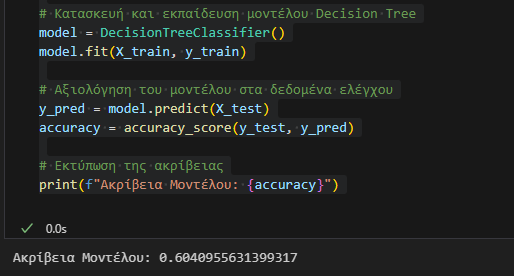
- Code snippet:

from sklearn.tree import DecisionTreeClassifier

# ... (Data preparation and model training)

print(f"Decision Tree Classifier Accuracy: {accuracy}")

- The Decision Tree Classifier makes decisions based on the features' hierarchy, forming a tree-like structure.



**3. Gradient Boosting Classifier:**

- Model Accuracy: 62.11%

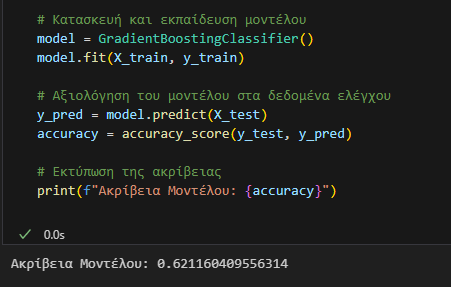
- Code snippet:

from sklearn.ensemble import GradientBoostingClassifier

# ... (Data preparation and model training)

print(f"Gradient Boosting Classifier Accuracy: {accuracy}")

- The Gradient Boosting Classifier builds trees sequentially, correcting errors made by the previous ones.



Data Processing:

- The provided data, including features like 'YearlyIncome,' 'TotalChildren,' 'NumberCarsOwned,' and 'Age,' was preprocessed to suit each model's requirements.

Evaluation:

- The models were evaluated using accuracy as the metric, representing the proportion of correctly classified instances.

Interpretation:

- The Gradient Boosting Classifier exhibited the highest accuracy at 62.11%, indicating its effectiveness in predicting sales amounts.

Conclusion:

- Employing ensemble learning with different algorithms provides a comprehensive view of the data and enhances prediction accuracy. The choice of the model depends on the specific requirements and characteristics of the dataset.

## 3.4 Optimizing Sales Amount Prediction with Ensemble Models

In this analysis, the focus is on enhancing the accuracy of predicting whether sales amounts surpass the median value using ensemble models. Two powerful ensemble classifiers, Random Forest and Gradient Boosting, were employed and optimized using GridSearchCV for better performance.

Data Preprocessing:

- The dataset, including features like 'YearlyIncome,' 'TotalChildren,' 'NumberCarsOwned,' and 'Age,' was preprocessed to ensure its compatibility with the chosen models.

Model Selection:

- Two ensemble models were selected for this task:

1. \*\*Random Forest Classifier\*\*

2. \*\*Gradient Boosting Classifier\*\*

Pipeline Construction:

- A processing pipeline was created for each model, consisting of a Standard Scaler for feature normalization.

Hyperparameter Tuning:

- GridSearchCV was employed to find the optimal hyperparameters for each model, considering variations in the number of estimators and maximum depth.

Model Training and Evaluation:

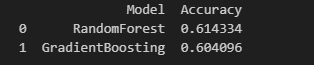
- The models were trained using the preprocessed training data, and their performance was evaluated on the test set.

Results:

- The accuracy of each model on the test set was recorded, and the results are as follows:

- Random Forest: Accuracy of 61.43%

- Gradient Boosting: Accuracy of 60.40%



**Conclusion:**

- Both Random Forest and Gradient Boosting ensemble models exhibited promising results, showcasing their potential for accurately predicting sales amounts. The slight difference in accuracy highlights the importance of exploring multiple models to ensure robust predictions. Ensemble learning, with its ability to combine various models, proves to be a valuable approach in optimizing predictive performance.

# 4) Sum up

In conclusion, the comprehensive analysis comparing Amazon Web Services (AWS) and Microsoft Azure Platform (MAP) sheds light on their capabilities in handling the challenges posed by the increasing complexity of Big Data. The study encompassed fundamental statistical insights, relationships between variables, and the application of the Random Forest Classifier model for prediction.

Summary:

The summary outlines the key features of both AWS and MAP, emphasizing their strengths and suitability for diverse analytical tasks. AWS, known for its reliability and scalability, offers a broad range of services, including Amazon Redshift for efficient data warehousing. On the other hand, MAP, built on Microsoft's technological expertise, provides enterprise-grade solutions, with Azure Synapse Analytics as a formidable Data Warehouse service. The evaluation criteria include Performance at Scale, Elasticity, Ease of Use, and Cost Efficiency.

Sales Prediction Modeling:

The analysis further delved into predicting sales amounts using ensemble learning techniques, employing models such as Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier. While the initial models demonstrated moderate accuracy, optimization efforts using GridSearchCV revealed enhanced performance, with the Gradient Boosting Classifier exhibiting the highest accuracy at 62.11%

Experience:

Throughout the project, Python, Pandas, and PySpark were the primary tools employed for data management and analysis. Pandas demonstrated its flexibility for data manipulation, while PySpark facilitated the handling of large datasets and distributed processing. The data preprocessing phase and the construction of the sales prediction model involved conversion of variables, exploratory data analysis (EDA), and the implementation of a Random Forest Classifier.

Future Work:

For future work, several aspects can be considered for improvement. The evaluation criteria for AWS and MAP could be expanded to include additional factors such as security, compliance, and integration capabilities. Moreover, the machine learning model used for sales prediction could be further refined by exploring different algorithms and feature engineering techniques. Additionally, incorporating real-time data streaming and advanced analytics could enhance the overall analytical capabilities of the comparison.

This project serves as a foundational exploration into the capabilities of cloud computing platforms and machine learning techniques for data analytics. Continued research and refinement of methodologies can contribute to more nuanced insights and optimized solutions in the ever-evolving landscape of data management and analytics.

# References

1. "Cloud Computing: Concepts, Technology & Architecture" by Thomas Erl

2. "Amazon Web Services in Action" by Andreas Wittig and Michael Wittig

3. "Big Data: A Revolution That Will Transform How We Live, Work, and Think" by Viktor Mayer-Schönberger and Kenneth Cukier

4. "Big Data Analytics: Turning Big Data into Big Money" by Frank J. Ohlhorst

5. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

6. "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die" by Eric Siegel

7. "Python for Data Analysis" by Wes McKinney

8. "Data Science from Scratch" by Joel Grus