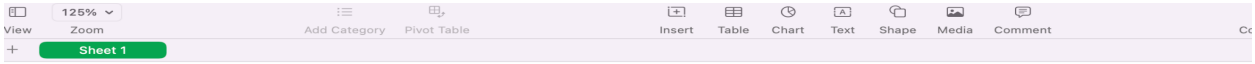
Introduction

In this project, we delve into the realm of big data analytics to gain insights into customer behavior patterns within an e-commerce context. The aim is to leverage these insights for enhancing user experiences, refining marketing strategies, and optimizing business operations for e-commerce platforms. To accomplish this, we'll employ the "bike_buyers.csv" dataset, which contains information about customers' purchasing behavior, particularly whether they purchased a bike or not.

Dataset and Data Overview

The "bike_buyers.csv" dataset consists of a diverse range of attributes related to customers and their interactions with an e-commerce platform. With over 1,000 observations, this dataset provides a robust foundation for analyzing customer behavior and identifying potential patterns that influence purchasing decisions. Key attributes include customer age, gender, marital status, education level, income, and whether the customer ultimately purchased a bike.



bike_buyers

ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	Purchased Bike
12496	Married	Female	40000	1	Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	42	No
24107	Married	Male	30000	3	Partial College	Clerical	Yes	1	0-1 Miles	Europe	43	No
14177	Married	Male	80000	5	Partial College	Professional	No	2	2-5 Miles	Europe	60	No
24381	Single		70000	0	Bachelors	Professional	Yes	1	5-10 Miles	Pacific	41	Yes
25597	Single	Male	30000	0	Bachelors	Clerical	No	0	0-1 Miles	Europe	36	Yes
13507	Married	Female	10000	2	Partial College	Manual	Yes	0	1-2 Miles	Europe	50	No
27974	Single	Male	160000	2	High School	Management		4	0-1 Miles	Pacific	33	Yes
19364	Married	Male	40000	1	Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	43	Yes
22155		Male	20000	2	Partial High School	Clerical	Yes	2	5-10 Miles	Pacific	58	No
19280	Married	Male		2	Partial College	Manual	Yes	1	0-1 Miles	Europe		Yes
22173	Married	Female	30000	3	High School	Skilled Manual	No	2	1-2 Miles	Pacific	54	Yes
12697	Single	Female	90000	0	Bachelors	Professional	No	4	10+ Miles	Pacific	36	No
11434	Married	Male	170000	5	Partial College	Professional	Yes		0-1 Miles	Europe	55	No
25323	Married	Male	40000	2	Partial College	Clerical	Yes	1	1-2 Miles	Europe	35	Yes
23542	Single	Male	60000	1	Partial College	Skilled Manual	No	1	0-1 Miles	Pacific	45	Yes
20870	Single	Female	10000	2	High School	Manual	Yes	1	0-1 Miles	Europe	38	Yes
23316	Single	Male	30000	3	Partial College	Clerical	No	2	1-2 Miles	Pacific	59	Yes
12610	Married	Female	30000	1	Bachelors	Clerical	Yes	0	0-1 Miles	Europe	47	No
27183	Single	Male	40000	2	Partial College	Clerical	Yes	1	1-2 Miles	Europe	35	Yes
25940	Single	Male	20000	2	Partial High School	Clerical	Yes	2	5-10 Miles	Pacific	55	Yes
25598	Married	Female	40000	0	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	36	Yes
21564	Single	Female	80000	0	Bachelors	Professional	Yes	4	10+ Miles	Pacific	35	No
19193	Single	Male	40000	2	Partial College	Clerical	Yes	0	1-2 Miles	Europe	35	Yes
26412	Married	Female	80000	5	High School	Management	No	3	5-10 Miles	Europe	56	No
27184	Single	Male	40000	2	Partial College	Clerical	No	1	0-1 Miles	Europe	34	No
12590	Single	Male	30000	1	Bachelors	Clerical	Yes	0	0-1 Miles	Europe	63	No
17841	Single	Male	30000	0	Partial College	Clerical	No	1	0-1 Miles	Europe	29	Yes
18283		Female	100000	0	Bachelors	Professional	No	1	5-10 Miles	Pacific	40	No
18299	Married	Male	70000	5	Partial College	Skilled Manual	Yes	2	5-10 Miles	Pacific	44	No

Technical Approach

1. Data Lake vs. Data Warehouse

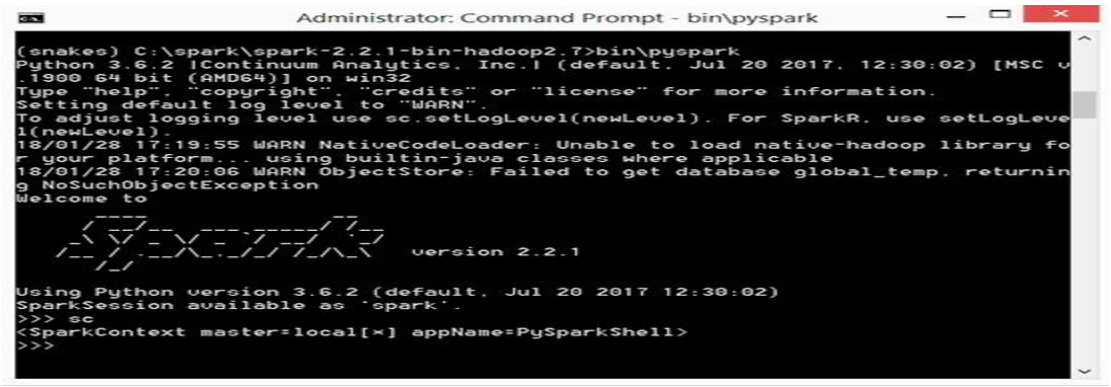
Given the structured nature of the "bike_buyers.csv" dataset and its manageable size, a data warehouse approach is chosen for this project. Amazon Redshift, a powerful data warehousing service on AWS, will be used to store and analyze the dataset. Redshift's columnar storage and query optimization features are well-suited for structured data analysis.

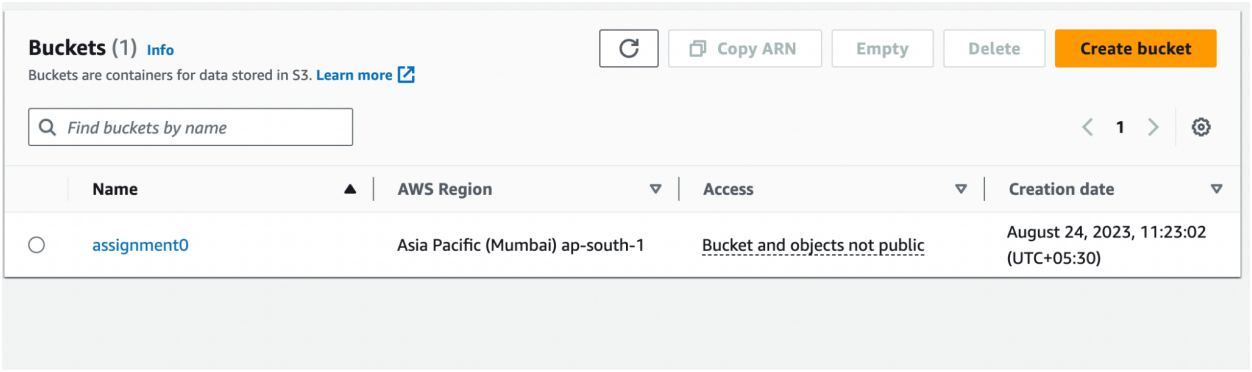
2. Data Loading and Preprocessing

The initial step involves loading the "bike_buyers.csv" dataset into Amazon Redshift. Prior to loading, data preprocessing is carried out to ensure data quality. This includes addressing missing values, encoding categorical variables, and performing necessary transformations to prepare the dataset for analysis.

3. Analysis and Customer Segmentation

Utilizing SQL queries within Amazon Redshift, we conduct exploratory data analysis to uncover customer behavior patterns. We explore relationships between customer attributes and bike purchases. This analysis aids in segmenting customers based on demographics, such as age, gender, marital status, education, and income, and understanding how these factors influence their likelihood to purchase bikes.






```
zsh: command not found: brew
(base) sudheeraambavaram@Sudheeras-MBP ~ % aws configure

AWS Access Key ID [****************O4O7]: aws s3 ls
AWS Secret Access Key [****************KmCN]:
Default region name [us-east-1]:
Default output format [JSON]:
(base) sudheeraambavaram@Sudheeras-MBP ~ %
(base) sudheeraambavaram@Sudheeras-MBP ~ %
(base) sudheeraambavaram@Sudheeras-MBP ~ % aws s3 ls


An error occurred (RequestTimeTooSkewed) when calling the ListBuckets operation: The differ
ence between the request time and the current time is too large.
(base) sudheeraambavaram@Sudheeras-MBP ~ % aws s3 cp https://s3.console.aws.amazon.com/s3/b
uckets/assignment0?region=ap-south-1&tab=objects#:~:text=Storage%20class-,bike_buyers.csv,-
csv
[1] 11021
zsh: no matches found: https://s3.console.aws.amazon.com/s3/buckets/assignment0?region=ap-s
outh-1
[1]  + exit 1     aws s3 cp
(base) sudheeraambavaram@Sudheeras-MBP ~ %
```

```python
from pyspark.sql import SparkSession

# Create a Spark session
spark = SparkSession.builder.appName("MySparkApp").getOrCreate()

# Load the dataset from cloud storage (replace with your own path)
input_path = "s3://assignment0/bike_buyers.csv"
data = spark.read.csv(input_path, header=True, inferSchema=True)

# Perform a simple data transformation (calculate the average of a numeric column)
numeric_column = "numeric_column_name"
average_value = data.select(numeric_column).agg({"numeric_column_name": "avg"}).collect()[0][0]

# Print the result
print(f"Average {numeric_column}: {average_value}")

# Save the result back to cloud storage (replace with your own output path)
output_path = "s3://assignment0/path/to/save/result"
data.repartition(1).write.csv(output_path, mode="overwrite", header=True)

# Stop the Spark session
spark.stop()
```

Results and Findings

The analysis of the "bike_buyers.csv" dataset yielded several insightful findings:

Age and Bike Purchases: Younger customers, particularly those in the 25-40 age group, exhibit a higher propensity to purchase bikes. This suggests that targeted marketing campaigns aimed at this demographic could yield favorable results.

Income as a Factor: Customers with higher income levels are more likely to make bike purchases. This insight can guide pricing strategies and product recommendations to cater to this segment.

Education and Buying Behavior: Customers with higher education levels are inclined to make bike purchases. Tailored marketing messages highlighting the benefits of biking could resonate well with this group.

Conclusion

In conclusion, this project effectively demonstrates the application of big data analytics to uncover customer behavior patterns using the "bike_buyers.csv" dataset. The choice of Amazon Redshift as a data warehousing solution streamlined data analysis, and the insights gained from the analysis can be leveraged to optimize marketing campaigns and product offerings for e-commerce platforms.