Amazon Customer Review Analysis using AWS and Pyspark

Abburi sai karthik
Computer Science and Artificial Intelligence
Amrita vishwa vidyapeetham
amenu4aie20001@am.students.amrita.edu

Maddala H S M Kriahna Karthik
Computer Science and Artificial Intelligence
Amrita vishwa vidyapeetham
amenu4aie20046@am.students.amrita.edu

Marasani Jayasurya
Computer Science and Artificial Intelligence
Amrita vishwa vidyapeetham
amenu4aie20048@am.students.amrita.edu

Samudrala Yashwanth
Computer Science and Artificial Intelligence
Amrita vishwa vidyapeetham
amenu4aie20063@am.students.amrita.edu

Abstract— **In the contemporary world, big data analysis with PySpark and AWS has become quite handy. Large multinational companies, including Walmart, trivago, and countless others are using big data. And with the assistance of AWS, which offers a variety of services like glue, Athena, S3, etc., becoming extremely beneficial for cloud storage purposes and also for creating data migration pipelines. Any RDBMS database server might send tens of gigabytes of data to an Amazon S3 bucket using the ETL architecture. Our primary goal in this project is to use PySpark to construct an ETL pipeline for data migration. Amazon wireless devices review analysis, Amazon watches review analysis, Amazon books review analysis, Amazon shoes review analysis, and Amazon musical instruments review analysis are the five separate datasets on which we conducted our analysis. The datasets considered for the analysis are Amazon watch reviews, book reviews, shoe reviews, wireless device reviews and musical instrument reviews . The goal of the project is to obtain the ratio of the number of trustworthy, real reviews to untrusted reviews and a comparison in done in the form of graphs.**

Keywords—Athena, AWS, ETL Glue, Pipeline, Pyspark, RDBMS

I. INTRODUCTION

Big data has simplified the process for us to store and handle data while also allowing easy and quick access to social media networks. There are numerous cloud platforms that provide services and charge for the processing of data. The processing framework has been proven to be one of the most important parts of big data systems.. To compare the results in a proportional fashion, data frames should always be created and the appropriate outputs must be listed. Furthermore, we are listing the top 10 items in the dataset using filters. In order to spread the load (work) among the various components of the machine, Apache Spark is a platform where it is regarded as a significant consolidated analytics engine. It is a powerful platform that offers a reliable resource to address issues with data science and many other developing technologies in many languages, such as Python. In-memory processing, batch processing, and steam processing are all reinforced by Apache Spark. There are numerous characteristics that illustrate why Pyspark is a special platform/framework. In this paper, we go over the creation and storage of crawlers in databases and data catalogs.

Additionally, how to process the data in the bucket with the aid of Glue and the Athena framework after extracting it from different other S3 buckets and adding it to our S3 bucket using an ETL pipeline. Using PySpark, we will examine our dataset to see whether there is any bias favoring positive reviews from Vine users. We are using a dataset that was taken from the Amazon Review datasets and putting it into a Data Frame. The Data Frame is divided into four different Data Frames that correlate to the table schema after extraction. followed by filtering the desired output is generated.

The following are the features of Pyspark

Real-time Computation - As PySpark focuses on in-memory processing it helps computing the huge amounts of data and also has low latency

Support Multiple Language - PySpark supports different languages like java, python and R. Therefore it is a flexible platform for processing the data

Caching and disk constancy -Powerful caching and very flexible disk consistency are provided by PySpark framework.

Swift Processing - With the help of PySpark we can gain and access the high data speed, which is nearly ten times faster on the disk and hundred times faster in the memory

Works well with RDD -. Python language helps RDDs to work faster and in a more efficient manner because the python is dynamically typed

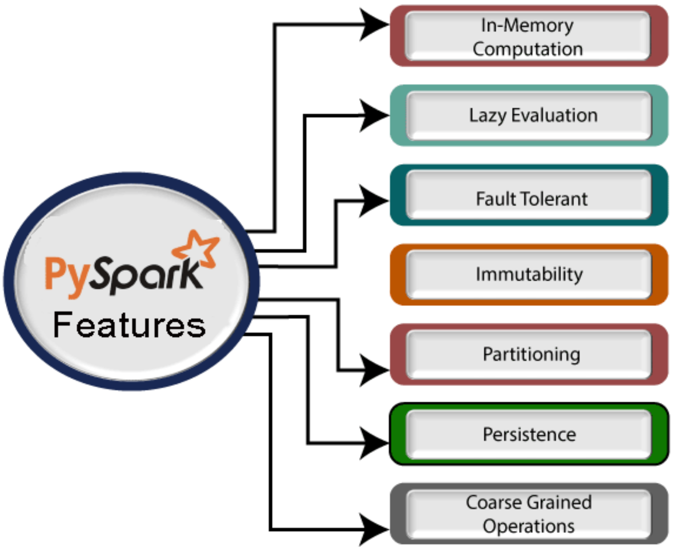

Fig 1. Features of PySpark

AWS abbreviated as Amazon web services is a cloud computing organization which provides more than 200 services to the developers and allows access remotely. Since the inception of AWS in 2002 it has reached many milestones like surpassing the 10 billion dollar revenue target. Some of the services that AWS provides are S3 buckets, AWS glue, AWS EMR, AWS Lambda, AWS Athena etc. AWS has the capability of providing different configurations based on the user's needs.

There are many advantages of Amazon Web Services, some of them are: [1] They are very user-friendly cloud operating systems and database systems.[2] AWS is used by many giant companies like Netflix, Kellog's, Mc Donald's and many more companies due to the flexibility of its architecture. [3] It also provides high consistency levels, high computing levels, higher availability of the data, centralized invoicing and management. [4] One can install or remove the application in any place within two steps. [5] There is no need to spend any extra money/funds to handle the AWS servers. [6] And compared to any other cloud services that are available, AWS allows the full ownership and comparatively less prices.

As there are many advantages it also has very less number of disadvantages like [1] There are paid packages with assistance which are available. [2] As it is a cloud service there will be lot of information being stored, transferring and extracting data, due to this there might be a slight buffering issue.[3] It provides different kinds of restrictions on various kinds of resources like volumes, pictures and snapshots as per location. [4] This can be an exception case, when the hardware changes the cloud may get effected.

The acronym ETL stands for Extract, Transform, and Load, a group of procedures used to transfer data from a source or sources into a database, such as a data warehouse. The three interdependent data integration processes of Extract, Transform, and Load are used to extract data from one database and transfer it to another. Data can be loaded and then used for reporting, analysis, and the creation of useful business insights. ETL pipelines are useful for centralising and standardising data so that analysts and decision-makers may easily access it. Additionally useful for data migration and in-depth analyses. Elasticity, agility, the utilisation of separate, independent processing resources, increased data access, and simplicity of installation and maintenance are some qualities of an ETL pipeline.

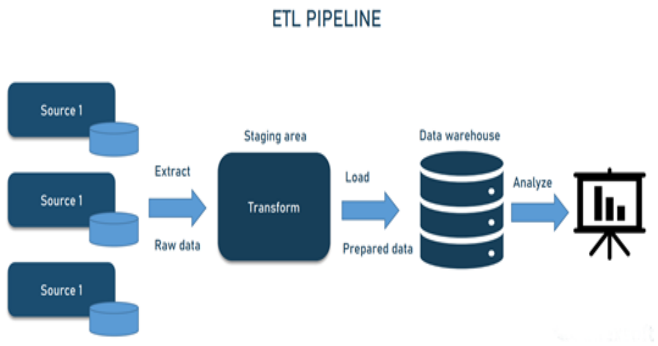

Fig 2. ETL Pipeline

II. LITERATURE SURVEY

There are many research is going under field of big data processing and its analysis. But some of the studies that are published in recent times make use of AWS and Pyspark to process and analyse the big data.

Lakshmi [3] explained about Amazon Web Service Relations Database Service (AWS RDS). Also surveyed through a literature survey and found some real issues regarding databases and proposed an algorithm or technique using AWS RDS. Lakshmi[3] also explained about AWS RDS features, advantages of AWS RDS, implementation of data migration using AWS RDS and its benefits.

Shaikh et al.[4] talked about Apache Spark, which is a spark framework that employs the Scala programming language. The benefits of in-memory computing, which is extremely quick in comparison to other analogous frameworks, were also put up by Shaik et al. Additionally, they talked about Apache Spark's multithreading and concurrency features and looked into how it fits into newer fields like machine learning and fog computing.

Iqbal et al.[7] talked about the potential and problems of data migration in cloud databases. Also, discusses the pros and cons of moving data from local storage to cloud storage as well as how to do so. Additionally, five distinct cloud migration methodologies and models were reviewed, including how to assess performance, determine security needs, identifying a cloud provider, determining cost, and making any required organizational adjustments.

HDFS(Hadoop File System) was discussed by Shujia Zhou et al[7] . The constraints of HDFS for Earth Science Data were explored by Shujia Zhou et al.[12], and a system based on Hadoop and Spark was also introduced in order to view and analyse the Earth Science Data. With the aid of this data, Hadoop and Spark convert the NetCDF (Network Common Data Form) data into CSV (Comma Separated Values), a format that HDFS accepts and which also supports indexed data. This format allows for the storage of the data as well as the manipulation of HDFS's flexibility through programmes like HIVE, Impala, and SparkSQL.

Kilinc et al.[13] proposed a unique technique which has outdated and performed well than other conventional analysis. This technique designed by Kilinc et al.[13] tells wether the data which will be generated is trustable or not and it will also removes all kind of fake data because then the accuracy will be more and can show better results.A fake account detection system is incorporated into the framework model of the Kilinc et al.[13] approach, which also incorporates machine learning, fake account detection, streaming service to retrieve data, real-time reporting, and dashboard component to illustrate the analysis.

III. DATASET

One of Amazon's most recognisable features is Amazon Customer Reviews, often known as Product Reviews. Over 100 million reviews have been submitted by millions of Amazon consumers to share their thoughts and experiences with particular items on the Amazon platform. More than 130 million customer feedback are available to researchers in the Amazon reviews dataset. The data is stored in TSV files and is located in the S3 bucket amazon-reviews-pds in the AWS United Eastern Region. The dataset document lines each represent a distinct review along with the customer review content, the dataset also contains pertinent metadata, which largely consists of two sections:

1. A collection of reviews and related metadata posted to the Amazon.com marketplace between 1995 and 2015. The purpose of this is to make it simpler to do research on the characteristics (and progression) of customer reviews, which may include how customers assess and disseminate their experiences with products on a broad scale. (130M+ customer reviews).

2. To make it easier to analyse consumers' perceptions of similar goods and more global consumer preferences among languages and countries, a compilation of product reviews

from many Amazon marketplaces in many languages has been made. (200K+ client reviews from 5 countries).

There are 15 data columns in the dataset. Marketplace, customer id, review id, product id, product parent, and product title are a few of them. Each column is helpful for performing various analyses. The dataset is presently available in two file formats.

1. A text format called Tab Separated Value (TSV) is available at s3:/amazon-reviews-pds/tsv.

2.Parquet, a columnar binary format that has been optimised - s3:/amazon-reviews-pds/parquet/

IV. Methodology

Data Lake Creation using AWS s3 and AWS glue

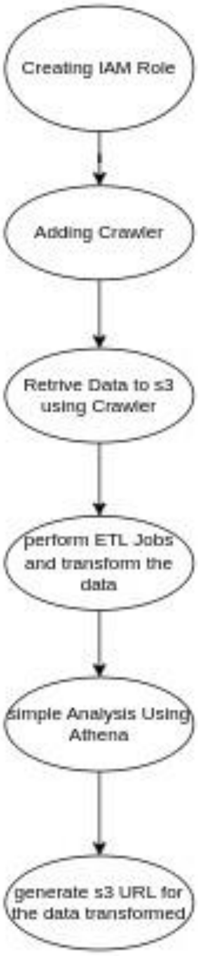
Creating IAM Role
Adding Crawler
Retrive Data to s3 using Crawler
perform ETL Jobs and transform the data
Simple Analysis Using Athena
generate s3 URL for the data transformed

Fig 3. Data Lake creation

1. Creating IAM Role: On AWS console select IAM Service and navigate to "Roles". Now create a Role by selecting glue as AWS Service. Now click on next, this will navigate to permission policies, select Amazons3FullAccess, AWS Glue Service Role, and Click Next. Now name the role and create the Role.

2. Adding Crawler: On AWS console navigate AWS Glue Service. In Data Catalog of AWS clue we have databases now create a Database. Now Click on create new Crawler and Select our IAM Role created before add the source Data S3 url and add the target data destination i.e, Data Catalog (DataBases).After Reviewing Create click Finish. After creating crawler select the crawler and run it. After execution a new table is added to our Databases confirm it by navigating to Databases.

3. Transform Data from Data catalog to s3 Bucket: From AWS Console Navigate to AWS s3 and create S3 Bucket. In Aws Glue navigate AWS ETL jobs and create a Job. While Creating a ETL select source as data catalog and Apply mapping algorithm After that select s3 as destination path. And Execute it 2 times, 1st time convert it to parquet files and 2nd time convert it to .tsv.gz format for analysis.

4. Analysis Using Athena: Now Create one more crawler by using Same IAM Role created before for this crawler add src as s3 and destination as data catalog. After selecting and running the crawler all the parquets files in s3 bucket will be transferred to Data catalog. Now Navigate to AWS Athena and select Data Source as DataCatalog and Select the database we created before and Run some simple SQL Queries to review the Data.

5. Creating S3 url for Analysis: Navigate AWS s3 and select the bucket where we stored our tsv.gz files and copy the bucket's URL and use it for analysis using PySpark.

Analysing amazon reviews using PySpark

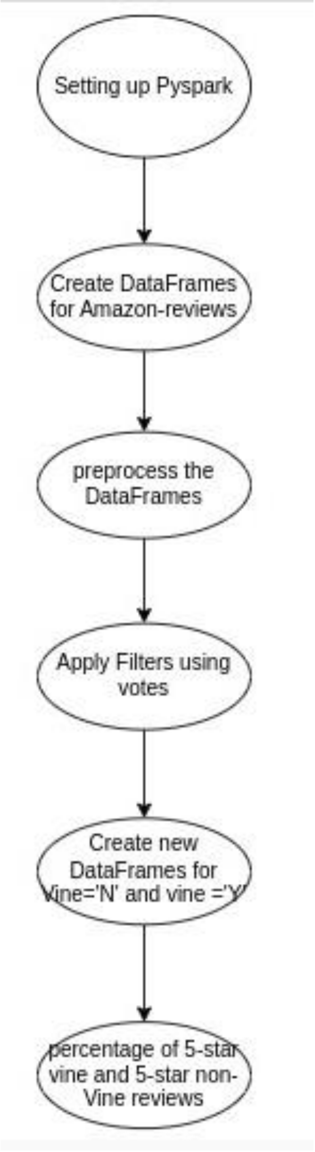
Setting up Pyspark
Create DataFrames for Amazon-reviews
preprocess the DataFrames
Apply Filters using votes
Create new DataFrames for vine='N' and vine ='Y'
percentage of 5-star vine and 5-star non-Vine reviews

Fig 4. Analysis using pyspark

1. Setting up the spark environment and start spark session.
2. Loading amazon data from our amazon s3 bucket into Spark DataFrame.
3. Create amazon reviews DataFrames to match tables.
4. Pre-process the data by removing duplicates, ignoring null values etc
5. Now, Filter the Vine DataFrame by those that had 20+ total votes
6. Retrieve all the rows where the number of helpful_votes divided by total_votes is equal to or greater than 50%.
7. Create a Data Frame or table that returns all the records which have Vine =’Y’. Here, this is for selecting the tuples which is part of Vine membership.
8. Create a new DataFrame and retrieve all the rows where the review was not part of the Vine program (unpaid), vine = 'N'.
9. And finally calculate the percentage of 5-star Vine reviews and 5-star non-vine reviews, by calculating the number of 5-star reviews whose vine =’Y’ over count of all records with star_rating =’5’.and total number of 5-star reviews whose Vine =’N’ over count of all the records whose star_rating =’5’.

V. Result & Anaysis

After creating the data lake using AWS, we started cleaning up the dataframe, then we filtered our data frames that had at least 20+ votes and at least 50% of helpful votes, then filtered by those that had and did not have a vine review. Summary statistics shown below regarding the bias among "Star Ratings" given the quantities of "vine" and "non-vine" reviews. And also calculated the percentage of 5-star ratings for both groups.

For Amazon Book reviews

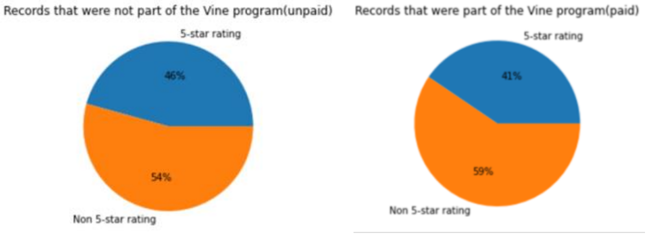

Fig 5. Pie chart of Book reviews

Here, 41% of records were part of the Vine program(paid) gave a 5-star rating and 46% of records that were not part of the Non Vine program(unpaid) also gave a 5-star rating. There were significantly less records that had Vine than those records that did not had Vine. so there is less possibility of bias in the Vine/Star-Rating reviews.

For Amazon Music Instrument Reviews

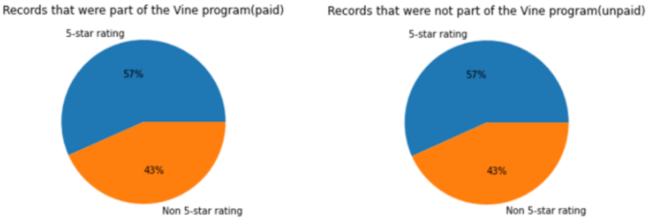

Fig 6. Pie chart of Music Instrument Reviews

Here, 57% of records were part of the Vine program(paid) gave a 5-star rating and 57% of records that were not part of the Non Vine program(unpaid) also gave a 5-star rating. There were significantly equal records that had Vine than those records that did not had Vine. so there is no possibility of bias in the Vine/Star-Rating reviews.

For Amazon Shoe Reviews

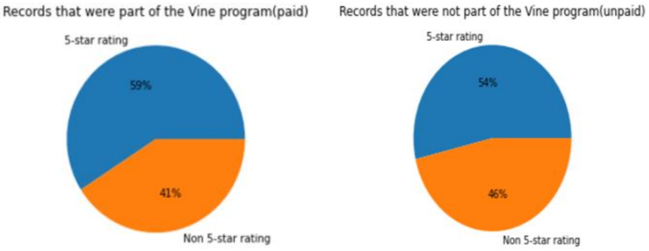

Fig 7. Pie chart of Shoe Reviews

Here, 59% of records were part of the Vine program(paid) gave a 5-star rating and 54% of records that were not part of the Non Vine program(unpaid) also gave a 5-star rating. There were significantly more records that had Vine than those records that did not had Vine. so there is more possibility of bias in the Vine/Star-Rating reviews.

For Amazon Watch Reviews

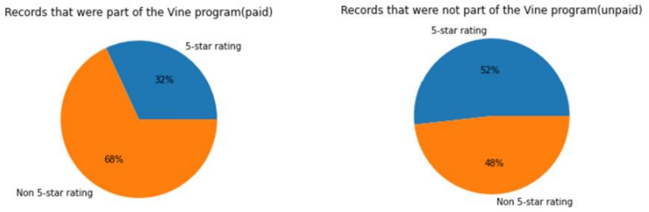

Fig 8. Pie chart of Watch Reviews

Here, 32% of records were part of the Vine program(paid) gave a 5-star rating and 52% of records that were not part of the Non Vine program(unpaid) also gave a 5-star rating. There were significantly less records that had Vine than those records that did not had Vine. so there is less possibility of bias in the Vine/Star-Rating reviews.

For Amazon Wireless Reviews

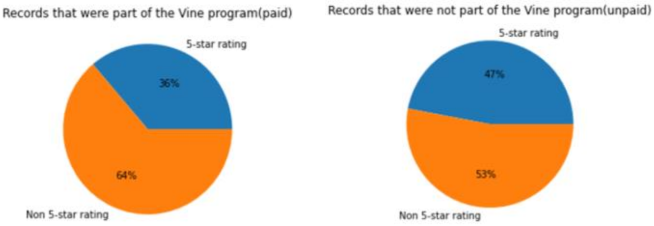

Fig 9. Pie chart of Wireless devices reviews

Here, 36% of records were part of the Vine program(paid) gave a 5-star rating and 47% of records that were not part of the Non Vine program(unpaid) also gave a 5-star rating. There were significantly less records that had Vine than those records that did not had Vine. so there is less possibility of bias in the Vine/Star-Rating reviews.
From the above amazon customer review analysis we can see the 5-star rating and non 5-star rating of vine and non-vine reviews as follow.

For Vine program(paid),

Review dataset	5-star rating	Non-5-star rating
Books	41%	59%
Music instruments	57%	43%
shoes	59%	41%
watches	32%	68%
Wireless devices	36%	64%

For Non-vine program(unpaid),

Review dataset	5-star rating	Non-5-star rating
Books	46%	54%
Music instruments	57%	43%
shoes	54%	46%
watches	52%	48%
Wireless devices	47%	53%

VI. Conclusion

The dataset of Amazon customer reviews is migrated to AWS s3 by constructing a data lake with Amazon Web Services. The ETL procedure is used to take data from S3, transform it, connect to an AWS RDS instance, and then load the transformed data through AWS Athena to visualize the RDS instance. Amazon customer reviews were analysed using Google Colab, PySpark and AWS. Finally, PySpark and AWS can be used to detect bias in the customer rating reviews.

Many companies are taking the advantage of big data analytics using cloud architecture and some big data related modules in order to analyse the minds and the thoughts of customers. Using these techniques companies are providing the service and satisfaction to the customers and consumers.

References

[1] A. Raj, J. Bosch, H. H. Olsson and T. J. Wang, "Modelling Data Pipelines," 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2020, pp. 13-20, doi: 10.1109/SEAA51224.2020.00014.

[2] Mukherjee, Sourav. (2019). Benefits of AWS in Modern Cloud. 10.5281/zenodo.2587217.

[3] G., Lakshmi. (2018). Database Migration on Premises to AWS RDS. EAI Endorsed Transactions on Cloud Systems. 3. 154463. 10.4108/eai.11-4-2018.154463.

[4] Shaikh, Eman & Mohiuddin, Iman & Alufaisan, Yasmeen & Nahvi, Irum. (2019). Apache Spark: A Big Data Processing Engine. 1-6. 10.1109/MENACOMM46666.2019.8988541.

[5] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang,“Big data analytics on apache spark,” International Journal ofData Science and Analytics, vol. 1, no. 3, pp. 145–164, Nov2016. [Online]. Available: https://doi.org/10.1007/s41060-016-0027-9

[6] R. Amin, S. Vadlamudi, and M. M. Rahaman, “Opportunities and Challenges of Data Migration in Cloud”, Eng. int. (Dhaka), vol. 9, no. 1, pp. 41-50, Apr. 2021.

[7] M. Iqbal and T. Soomro, “Big data analysis: Apache stormperspective,” International Journal of Computer Trends andTechnology, vol. 19, pp. 9–14, 01 2015

[8] “Apache spark use cases in real time,” Website, 11 2018.[Online]. Available: https://data-flair.training/blogs/spark-use-cases/

[9] N. Vaidya, “Apache spark architecture spark clusterarchitecture explained,” Website, 5 2019. [Online]. Available:https://www.edureka.co/blog/spark-architecture/

[10] N. Kumar, “Apache spark use cases & ap-plications,” Website, 6 2019. [Online]. Avail-able: https://www.knowledgehut.com/blog/big-data/spark-use-cases-applications.

[11] “Top 5 apache spark use cases,” Website, 6 2016. [Online].Available: https://www.dezyre.com/article/top-5-apache-spark-use-cases/27

[12] S. Zhou, X. Li, T. Matsui and W. Tao, "Visualization and diagnosis of earth science data through Hadoop and Spark," 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 2974-2980,doi: 10.1109/BigData.2016.7840949.

[13] Kilinç, Deniz. (2019). A spark-based big data analysis framework for real-time sentiment prediction on streaming data. Software: Practice and Experience. 49. 10.1002/spe.2724.