CA675 CLOUD TECHNOLOGIES- Assignment 2

Student's Name(s):	Alen George Shibu, Sarath Mukundan Adavakkat, Suhara Fathima Shahul Hameed
Student Number(s):	23271013, 23267908, 23266257
Student Mail ID(s):	alen.georgeshibu7@mail.dcu.ie, sarath.mukundanadavakkat2@mail.dcu.ie , suhara.shahulhameed2@mail.dcu.ie
Group Number:	34
Program:	Master of Science in Computing (Data Analytics)
Project Title:	Loan Defaulters' Demographic Visualizer
Module Code:	CA675
Lecturer:	Dr. Michael Scriney
Assignment Submission Date:	19th December 2023
GitLab Repository:	https://gitlab.computing.dcu.ie/georgea7/loan-defaulters-demographic-visualiser
Video Link:	https://drive.google.com/file/d/1TxjfonAQfwKhxW9sZhYix8H_NiM3RFzr/view?usp=drive_link

Table of Contents

INTRODUCTION 4

Choice of Technology..4

RELATED WORK 5

SYSTEM ARCHITECTURE..6

Source of the data 7

Data pre-processing..7

DATA PROCESSING..8

A.Distribution of Age 9

B.Distribution of Income 9

C.Distribution of Occupation Type..10

VISUALISATION..12

CHALLENGES..13

RESPONSIBILITY SECTION..13

RELEVANT SCREENSHOTS 14

List Of Figures

Figure 1:EMR-cloud2 7
Figure 2: Outliers in data 8
Figure 3: S3 bucket 8
Figure 4:Age distribution 9
Figure 5:Income Distribution 10
Figure 6: Occupation type distribution 11
Figure 7: Dashboard 12
Figure 8: S3 bucket cloud-assignment-S3 14
Figure 9: Code and dataset stored 14
Figure 10: Cluster EMR cloud-2 15
Figure 11: Executed PySpark code 15
Figure 12: Output stored in S3 bucket 16

INTRODUCTION

In the constantly evolving financial services industry, understanding the intricacies of loan defaulters is critical for risk assessment and proactive decision-making. It provides an extensive collection of information, uncovering significant patterns and weaknesses among those who default on loans.

Visualizing age, income, occupation type and gender data specifically for loan defaulters is an important analytical tool in identifying the risk factors linked with loan repayment. By focusing on this specific subset of persons who have defaulted on loans, we can acquire useful insights into the demographic and socioeconomic aspects that may contribute to defaulting behaviour. Understanding the demographic profile of defaulters by age, income, and gender enables financial institutions to estimate risk more properly. It aids in the refinement of lending strategies and risk assessment models by identifying trends or groups that may be more prone to loan defaults. We hope that by visualizing this, we can improve knowledge and guide solutions for mitigating default risk while increasing financial inclusion and stability.

Choice of Technology

The technology used for visualizing demographic data of loan defaulters - Amazon Web Services (AWS), PySpark, and Tableau—provides a strong and complete approach.
Amazon EMR (Elastic MapReduce) is a big data cloud platform provided by Amazon Web Services (AWS). It is a fantastic solution for efficiently processing large-scale datasets. It makes use of Apache Spark, among other frameworks, to enable distributed data processing. PySpark, a Python API for Spark, makes it easier to manipulate and analyze data. Using Amazon EMR with PySpark allows for the handling of huge volumes of demographic data, computations, and the extraction of valuable insights.
PySpark is a strong data processing tool, particularly when dealing with enormous datasets. Its interface with Amazon EMR enables distributed computing, making data translation, cleaning, and analysis more efficient.
Tableau, on the other hand, specializes in data visualization and interactive dashboard creation. It can connect to data sources such as Amazon EMR and display processed and analyzed demographic insights. Tableau has an easy-to-use interface for making visually appealing charts, graphs, and dashboards. It enables stakeholders to study and comprehend complicated demographic patterns among loan defaulters interactively.

RELATED WORK

1.Demographic Analysis 2022 [link]

The Central Bank of Ireland's "Demographics of the Financial Sector" report is likely to provide a detailed analysis of the composition and diversity of the financial industry workforce. This study sheds light on the demographic makeup, including age, gender, ethnicity, and maybe other diversity-related criteria. It examines how different demographic groups are represented at various career levels, noting potential trends or changes over time. The report concentrates on gender diversity, stressing the presence of women in positions of leadership, and it may also discuss ethnic or racial diversity within the sector. It could also highlight problems in reaching diversity targets, actions to encourage inclusion and the positive effects of a diverse workforce on corporate performance.

2.German Loan Default Prediction Project[link]

The focus of a German loan default prediction project is anticipated to be on using machine learning techniques to forecast the risk of loan defaults. Analysing a dataset containing borrower information, loan attributes, and previous default records is required. To prepare the dataset for modelling, the project may include data pre-treatment activities such as cleaning and feature engineering. To develop prediction models, many machine learning techniques, such as logistic regression or decision trees, may be used. The accuracy and efficacy of these algorithms in forecasting loan defaults are assessed during their evaluation. The project's findings frequently provide insights into major elements impacting default probabilities and aid in the improvement of credit risk assessment methods.

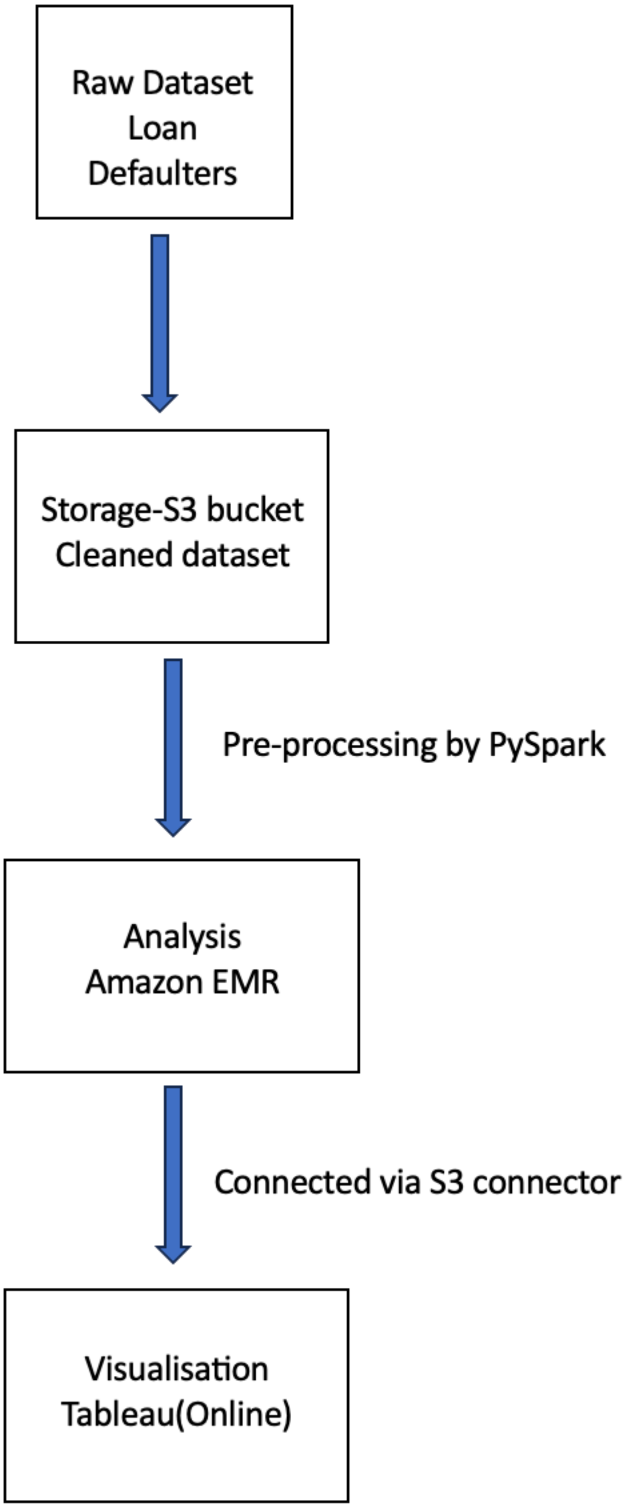
3.The Effect of Demographic Characteristics on Loan Performance of Commercial Banks In Kenya[link]

This study investigates how demographic characteristics influence commercial bank loan performance in Kenya. The study explores the impact of demographic variables such as age, income, education, and work status on borrowers' repayment behavior using data from various demographic variables. Statistical analysis results show substantial relationships between specific demographic variables and loan performance, providing insights that could guide more focused lending strategies or risk assessment processes in Kenya's commercial banking sector.

Our work focuses on demographic factors such as age, gender, income, and occupation of loan defaulters. We visualize these factors in a dashboard to understand these trends based on the gender of loan defaulters.

SYSTEM ARCHITECTURE

The system architecture of our visualization is given below:



DATASET

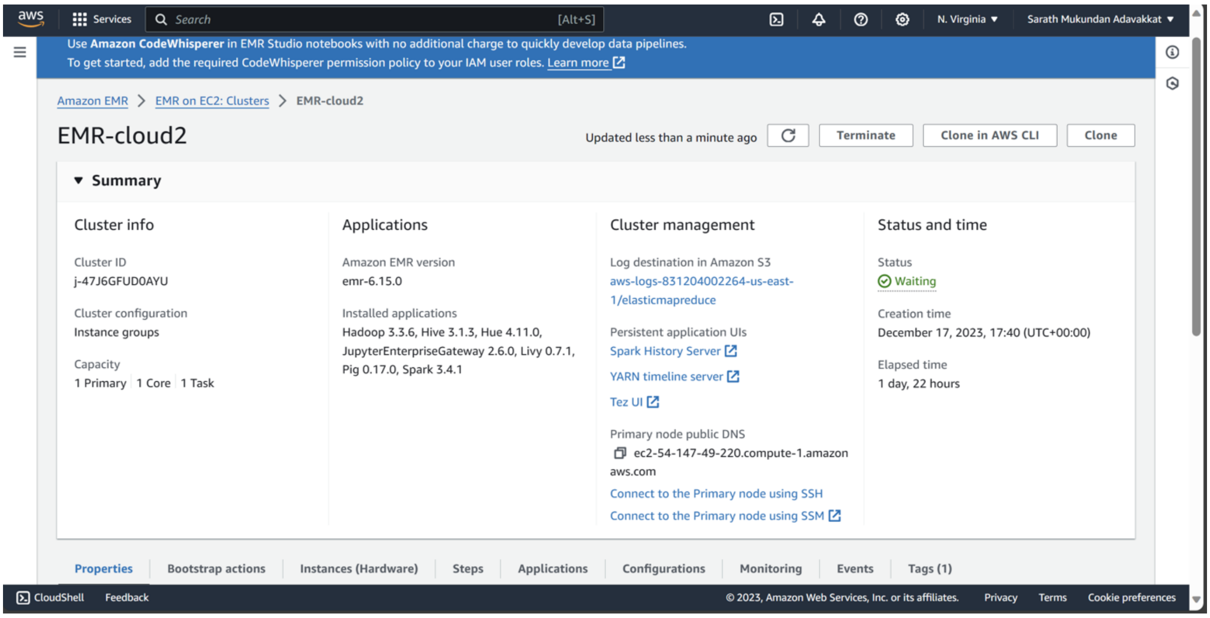
Source of the data

The data was collected from the Kaggle website.
Link:https://www.kaggle.com/datasets/gauravduttakiit/loandefaulter/data?select=application_data.csv
The original dataset consists of 122 columns and 307512 rows which included loan default status as 0 and 1. Target variable (1 - client with payment difficulties: he/she had a late payment, 0 – no payment difficulties). Out of 122 columns we required only 11 columns for our aim.

Data pre-processing

Link to source code repository: https://gitlab.computing.dcu.ie/georgea7/loan-defaulters-demographic-visualiser

The dataset preparation and cleaning are done in PySpark in the following steps :



Use **Amazon CodeWhisperer** in EMR Studio notebooks with no additional charge to quickly develop data pipelines.
To get started, add the required CodeWhisperer permission policy to your IAM user roles. Learn more

EMR-cloud2

Updated less than a minute ago

▾ **Summary**

Cluster info

Cluster ID
j-47J6GFUD0AYU

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version
emr-6.15.0

Installed applications
Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Pig 0.17.0, Spark 3.4.1

Cluster management

Log destination in Amazon S3
aws-logs-831204002264-us-east-1/elasticmapreduce

Persistent application UIs
Spark History Server
YARN timeline server
Tez UI

Primary node public DNS
ec2-54-147-49-220.compute-1.amazonaws.com

Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

Status
✓ Waiting

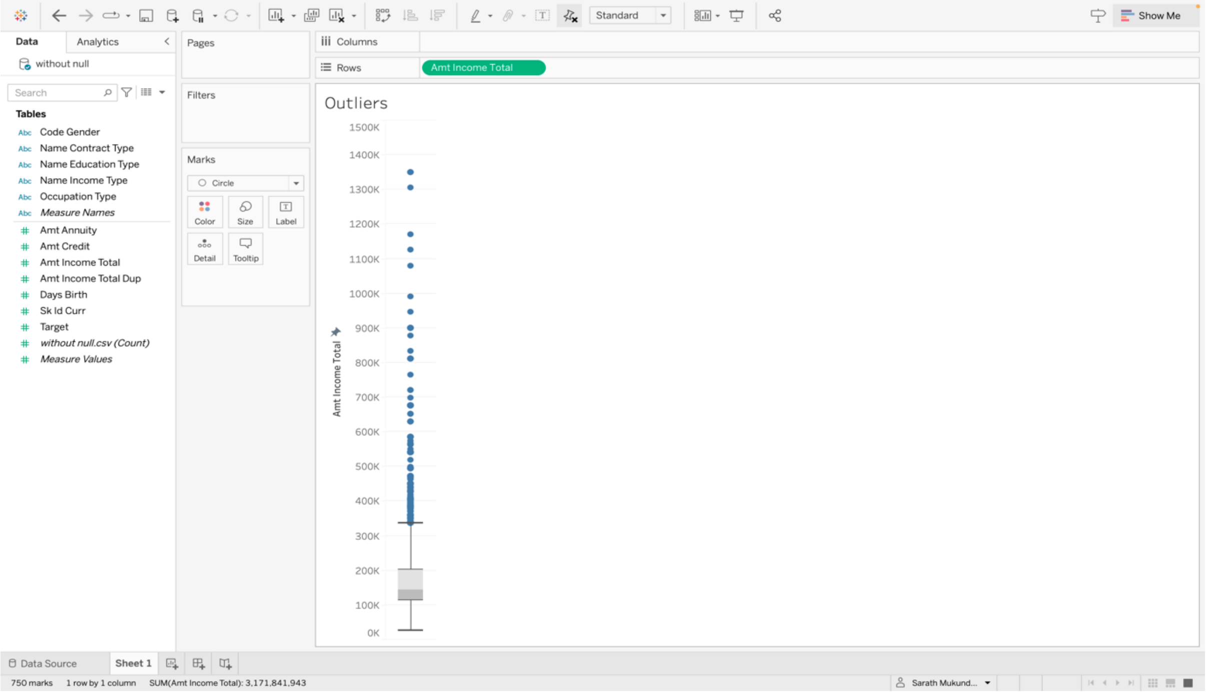
Creation time
December 17, 2023, 17:40 (UTC+00:00)

Elapsed time
1 day, 22 hours

© 2023, Amazon Web Services, Inc. or its affiliates.

Figure 1:EMR-cloud2

Initially, we selected rows with target = 1 i.e., loan defaulters as we are focussing on these data closely. Selection of required columns is performed which are SK_ID_CURR", "TARGET", "NAME_CONTRACT_TYPE","CODE_GENDER","AMT_INCOME_TOTAL","AMT_CREDIT","AMT_ANNUITY","NAME_INCOME_TYPE","NAME_EDUCATION_TYPE","DAYS_BIRTH", "OCCUPATION_TYPE".
Next, the DAYS_BIRTH column consisted of negative values which were converted to age by dividing the negative value by 365 and taking the absolute floor value as the age. Row with null values were eliminated. Checked for outliers in the tableau by box plotting and removed irrelevant data outliers.

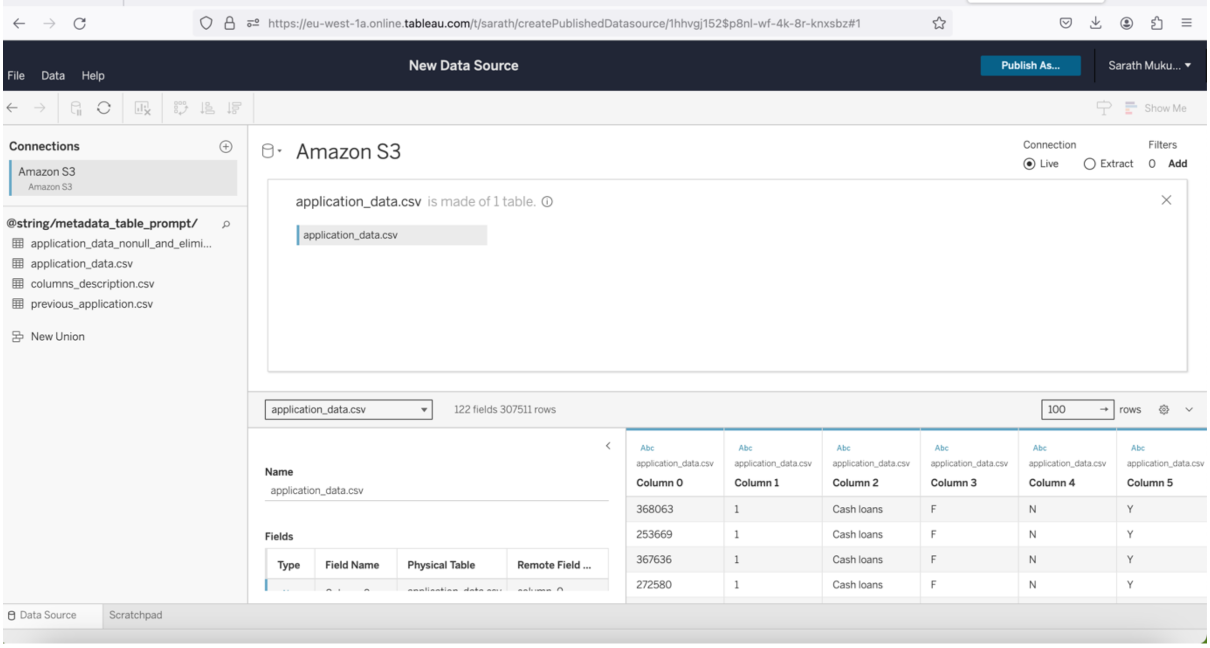


750 marks 1 row by 1 column SUM(Amt Income Total): 3,171,841,943

Figure 2: Outliers in data

DATA PROCESSING

The cleaned dataset from S3bucket is imported to Tableau online by using S3 connector.



application_data.csv is made of 1 table.

application_data.csv

application_data.csv 122 fields 307511 rows 100 rows

Name
application_data.csv

Fields

Type	Field Name	Physical Table	Remote Field ...

Abc application_data.csv Column 0	Abc application_data.csv Column 1	Abc application_data.csv Column 2	Abc application_data.csv Column 3	Abc application_data.csv Column 4	Abc application_data.csv Column 5
368063	1	Cash loans	F	N	Y
253669	1	Cash loans	F	N	Y
367636	1	Cash loans	F	N	Y
272580	1	Cash loans	F	N	Y

Figure 3: S3 bucket

The distribution of age, income and occupation of loan defaulters based on gender is visualised.

A.Distribution of Age

We have visualised the distribution for the age of loan defaulters based on gender which can be seen below.

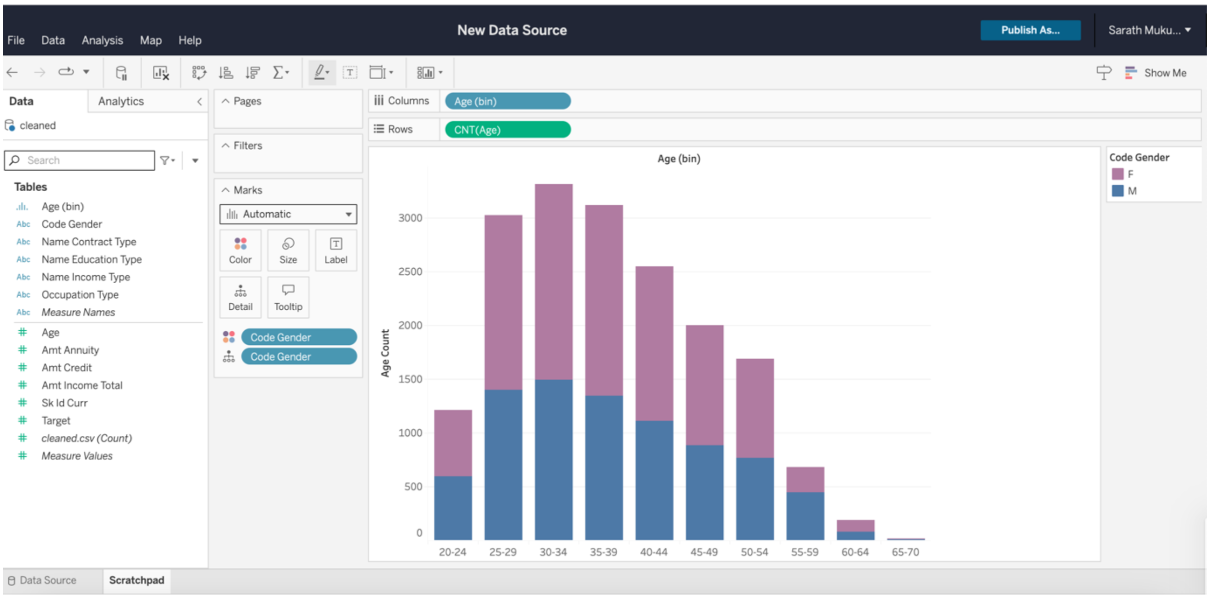


Figure 4:Age distribution

Insights from graph:

- The graph shows the distribution of age for males and females, with males in blue and females in purple colour. The x-axis shows the age interval, and the y-axis shows the number of people in each interval.
- The graph depicts the count of people in different age categories, from 20 to 70.
- Higher number of women in each age category compared to men.
- A slight peak in the 30-34 age category can be seen which says more people fall under this age category for loan defaulters.
- There is a gradual decline after the age category 30-34 i.e., fewer loan defaulters after this age group.

B.Distribution of Income

The below graph shows the distribution of loan defaulters income based on gender:

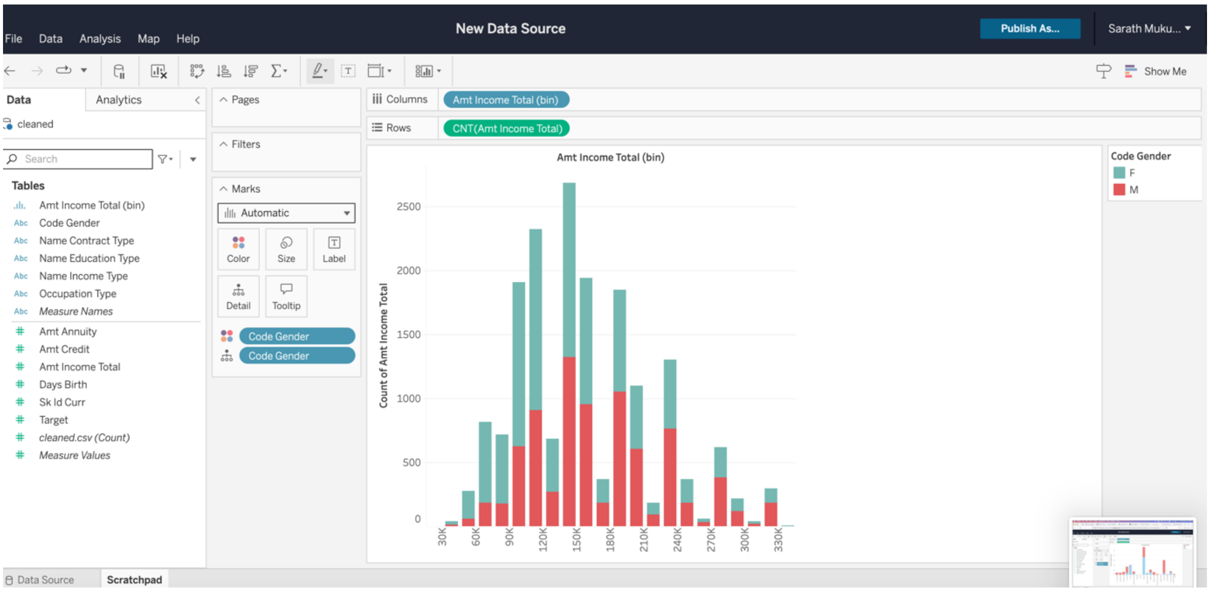


Figure 5:Income Distribution

Insights from the graph:

- The graph shows the distribution of income for males and females, with males in red and females in green. The x-axis shows the income section, and the y-axis shows the number of people in each section.
- More number of people reside in the range of €150k with almost the same for males and females.
- In the high-income range, we can see comparatively fewer females than male loan defaulters.
- There can be seen high no. of defaulters in the mid-income range.

C.Distribution of Occupation Type

The below graph represents the distribution of occupation type based on the gender of loan defaulters:

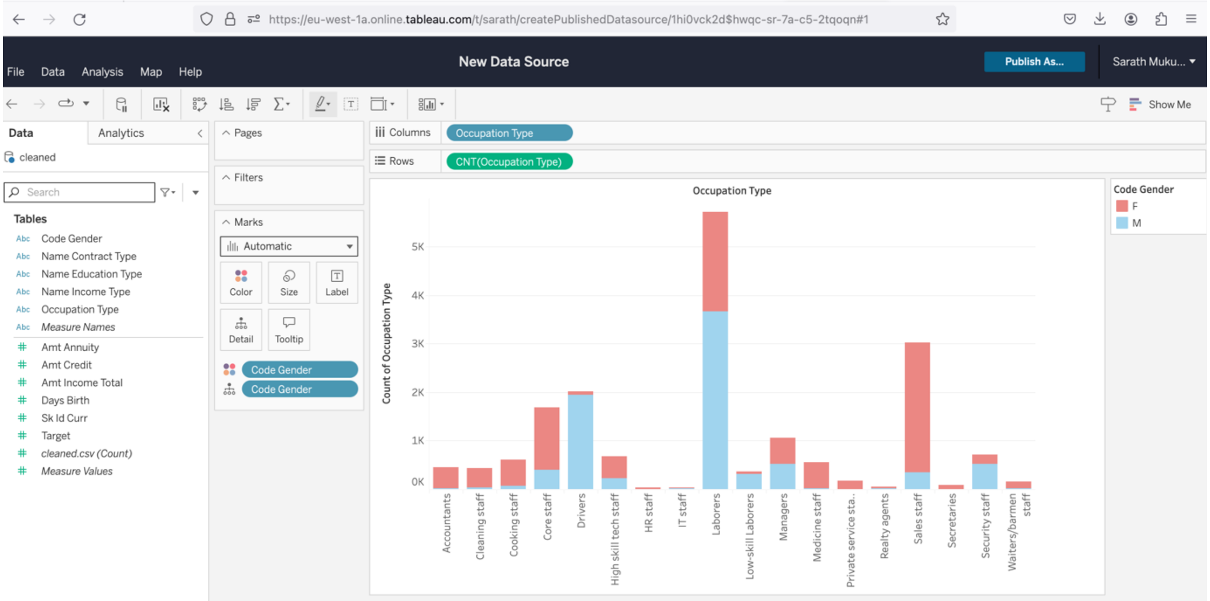


Figure 6: Occupation type distribution

Insights from the graph:

- The graph shows the distribution of occupation type for males and females, with males in blue and females in peach colour. The x-axis shows the occupation type, and the y-axis shows the number of people in each occupation.
- It can be seen more loan defaulters fall under the labourers category and the least in HR staff, realty agents and IT staff.
- More male loan defaulters in drivers and labourers occupations. In the case of female loan defaulters, more no. in labourers and sales staff.

VISUALISATION

Tableau (Online) was employed to visualise a dashboard with the distribution of age, income and occupation type based on gender for loan defaulters which can seen below:

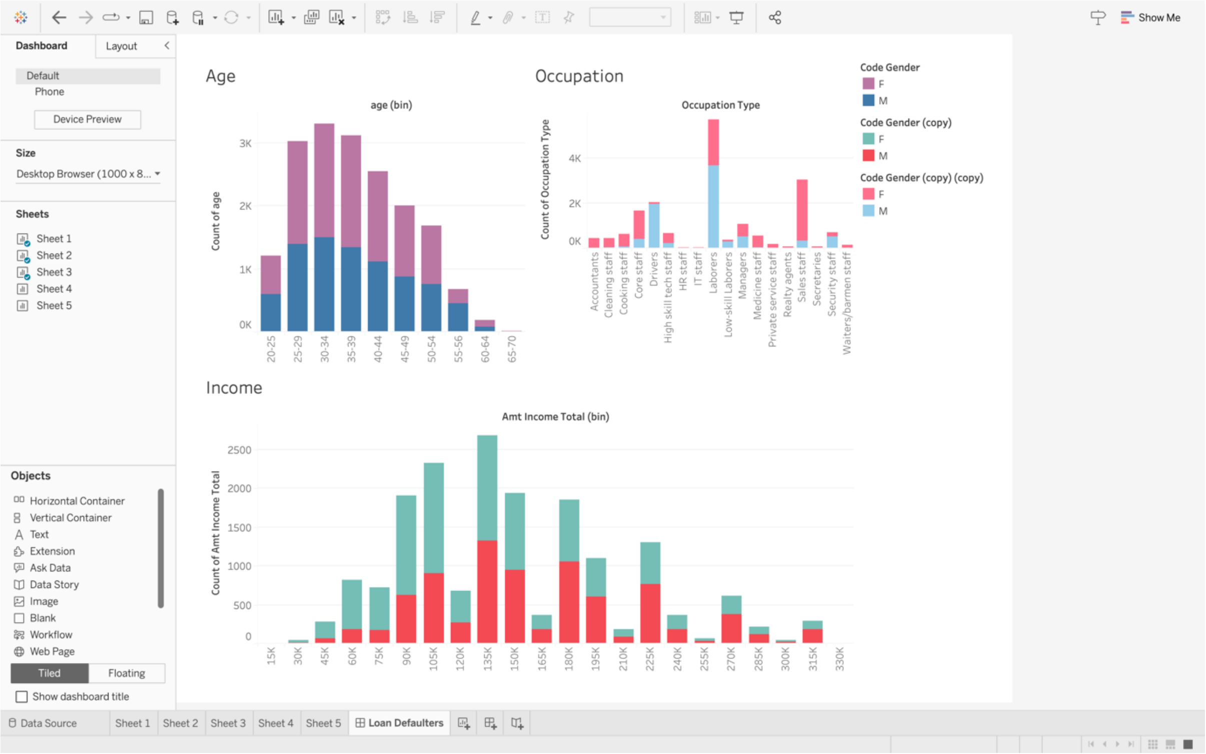


Figure 7: Dashboard

- The visualization shows a constant trend of more female loan defaulters throughout all age groups, with a noteworthy peak in the 30-34 age group. This knowledge could be critical for targeted financial initiatives or risk assessment aimed at reducing loan defaults, particularly among younger groups.

- A remarkable income distribution among male and female loan defaulters can be seen, emphasizing similarities in the middle and discrepancies in the upper-income groups. This knowledge can be useful in establishing targeted financial strategies or risk assessment models, particularly for high-income male defaulters and middle-income people of both genders.

- Loan defaulters are prevalent in the "Labourers" category, showing that this occupation type has the highest number of defaulters, regardless of gender.

- The graph, on the other hand, shows significantly lower counts of defaulters in the occupation categories of HR workers, real estate agents, and IT staff, implying that these occupations had fewer cases of loan defaults.

CHALLENGES

- There were no challenges or difficulties in data processing and cleaning in PySpark.
- The configurations of different AWS were quite difficult to grasp and keep track of, it was a time-consuming process. Later, after understanding the concept we found it to be really simple.
- There was some issue while connecting the AWS S3 bucket to Tableau online but later after some research, it was resolved.

RESPONSIBILITY SECTION

Regular communication and collaboration among all team members were done. To track work, share code snippets, and record any issues discovered, we used shared documentation like Google Docs. Search for the dataset was performed by everyone.

Alen:

- In charge of data input into AWS via PySpark.
- Created Python scripts to clean and pre-process raw data.
- Handled difficulties with missing values, outliers, and data formatting.
- Worked with Sarath to ensure data quality and consistency.

Sarath:

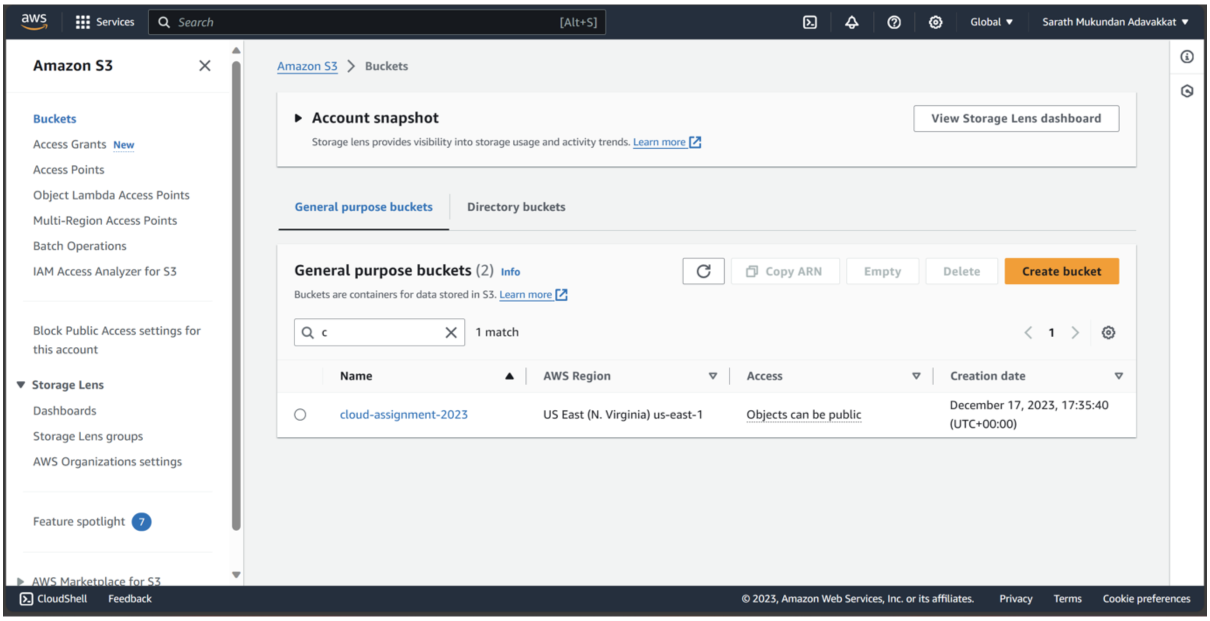
- Collaborated with Alen on data quality checks and pre-processing.
- Used Tableau to perform exploratory data analysis (EDA) to gain insights about age, income, employment, and loan defaulters based on gender.
- Performs feature engineering to prepare data for modelling and visualization.
- Documented the analysis procedure and results.

Suhara:

- Based on the preprocessed data, used Tableau to build insightful visualizations.
- Created dashboards and visual representations of age, income, occupation, and loan defaulters based on gender.
- Collaborated seriously with Alen and Sarath to fully comprehend the data and insights obtained.
- Wrote a detailed report outlining the important facts and insights.

RELEVANT SCREENSHOTS

1.Created S3 bucket named cloud-assignment-2023



Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

General purpose buckets (2) Info

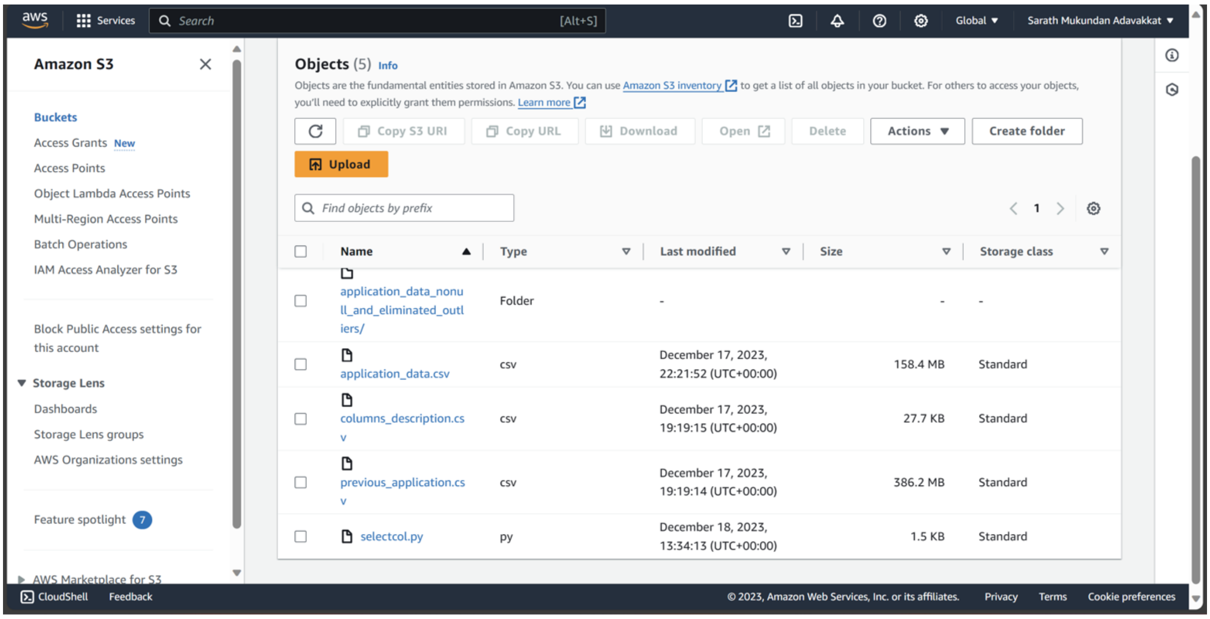
Buckets are containers for data stored in S3. Learn more

c — 1 match

	Name	AWS Region	Access	Creation date
○	cloud-assignment-2023	US East (N. Virginia) us-east-1	Objects can be public	December 17, 2023, 17:35:40 (UTC+00:00)

Figure 8: S3 bucket cloud-assignment-S3

2. Stored our raw dataset and PySpark code.



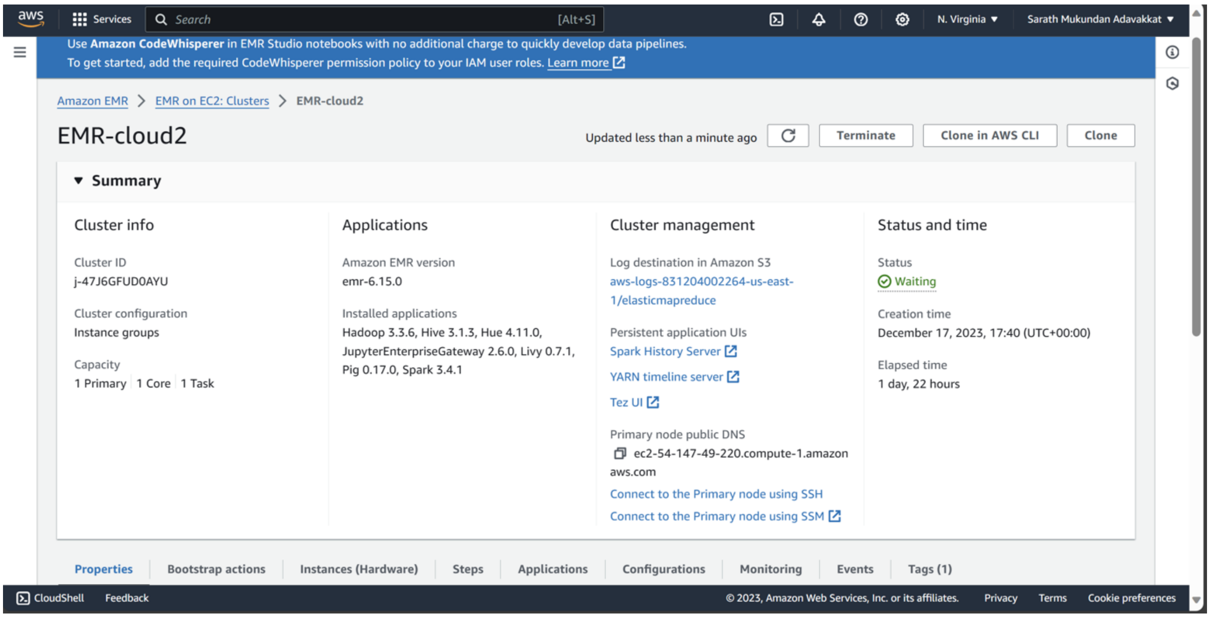
Objects (5) Info

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	application_data_nonull_and_eliminated_outliers/	Folder	-	-	-
□	application_data.csv	csv	December 17, 2023, 22:21:52 (UTC+00:00)	158.4 MB	Standard
□	columns_description.csv	csv	December 17, 2023, 19:19:15 (UTC+00:00)	27.7 KB	Standard
□	previous_application.csv	csv	December 17, 2023, 19:19:14 (UTC+00:00)	386.2 MB	Standard
□	selectcol.py	py	December 18, 2023, 13:34:13 (UTC+00:00)	1.5 KB	Standard

Figure 9: Code and dataset stored

3.Configuration of cluster created EMR-cloud2 using AWS EMR



Use **Amazon CodeWhisperer** in EMR Studio notebooks with no additional charge to quickly develop data pipelines.
To get started, add the required CodeWhisperer permission policy to your IAM user roles. Learn more

Amazon EMR > EMR on EC2: Clusters > EMR-cloud2

EMR-cloud2

Updated less than a minute ago

Summary

Cluster info

Cluster ID
j-47J6GFUD0AYU

Cluster configuration
Instance groups

Capacity
1 Primary | 1 Core | 1 Task

Applications

Amazon EMR version
emr-6.15.0

Installed applications
Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, Livy 0.7.1, Pig 0.17.0, Spark 3.4.1

Cluster management

Log destination in Amazon S3
aws-logs-831204002264-us-east-1/elasticmapreduce

Persistent application UIs
Spark History Server
YARN timeline server
Tez UI

Primary node public DNS
ec2-54-147-49-220.compute-1.amazonaws.com
Connect to the Primary node using SSH
Connect to the Primary node using SSM

Status and time

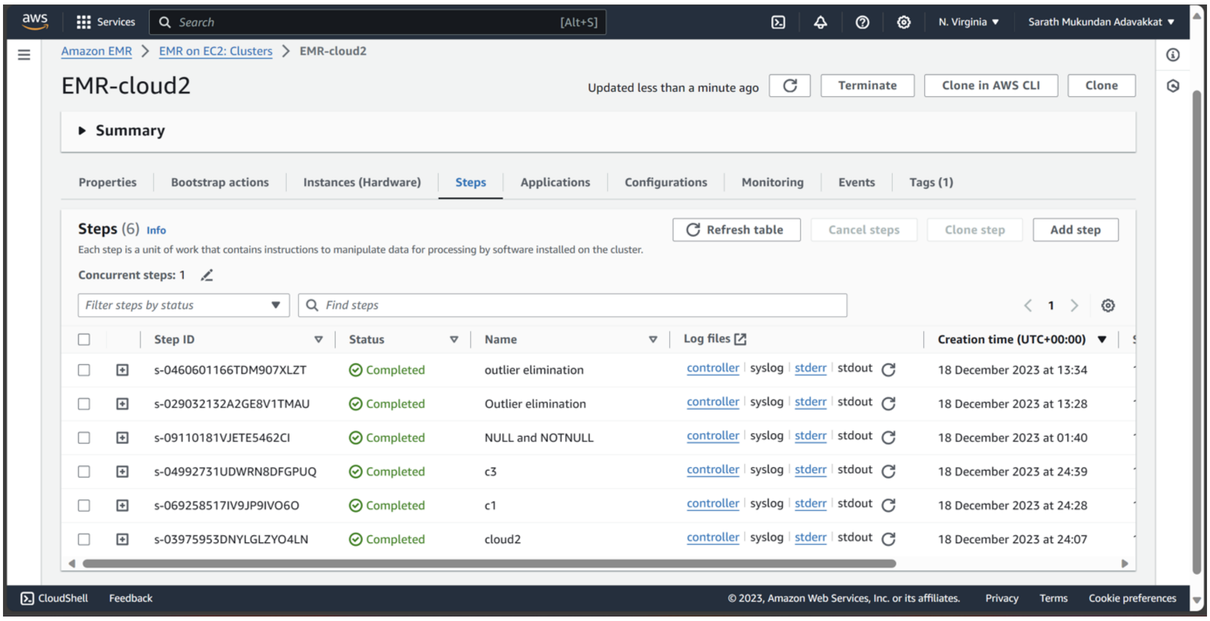
Status
Waiting

Creation time
December 17, 2023, 17:40 (UTC+00:00)

Elapsed time
1 day, 22 hours

Figure 10: Cluster EMR cloud-2

4.Executed PySpark codes using steps in EMR



Amazon EMR > EMR on EC2: Clusters > EMR-cloud2

EMR-cloud2

Updated less than a minute ago

Steps (6) Info

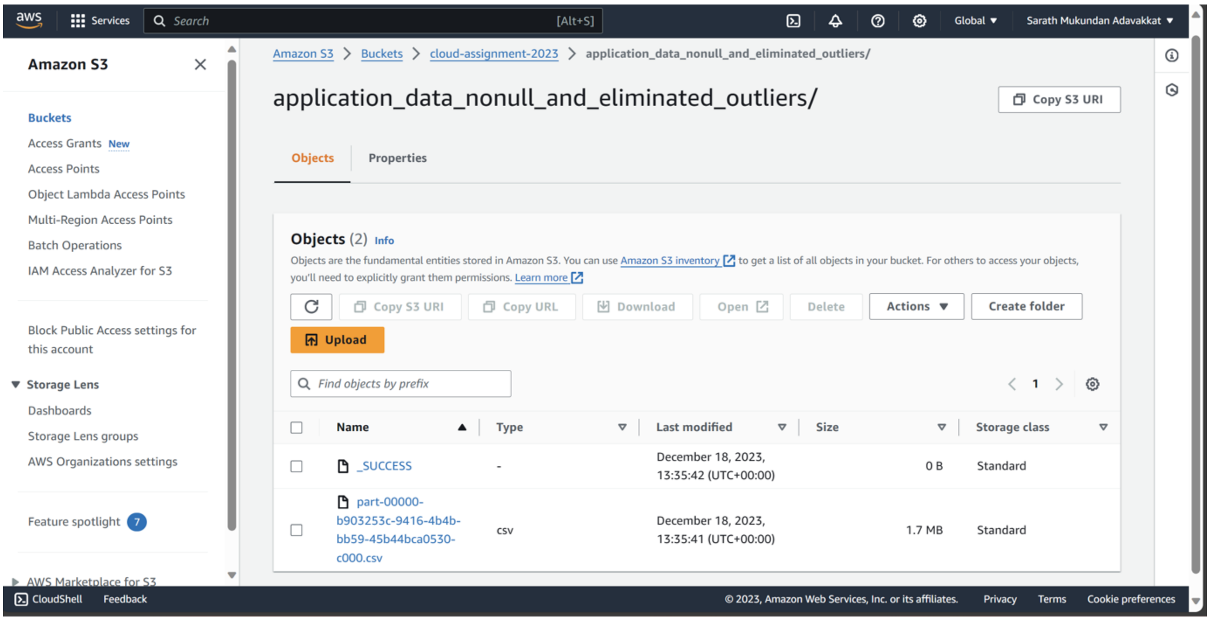
Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

Concurrent steps: 1

Step ID	Status	Name	Log files	Creation time (UTC+00:00)
s-0460601166TDM907XLZT	Completed	outlier elimination	controller | syslog | stderr | stdout	18 December 2023 at 13:34
s-029032132A2GE8V1TMAU	Completed	Outlier elimination	controller | syslog | stderr | stdout	18 December 2023 at 13:28
s-09110181VJETE5462CI	Completed	NULL and NOTNULL	controller | syslog | stderr | stdout	18 December 2023 at 01:40
s-04992731UDWRN8DFGPUQ	Completed	c3	controller | syslog | stderr | stdout	18 December 2023 at 24:39
s-069258517IV9JP9IVO6O	Completed	c1	controller | syslog | stderr | stdout	18 December 2023 at 24:28
s-03975953DNYLGLZYO4LN	Completed	cloud2	controller | syslog | stderr | stdout	18 December 2023 at 24:07

Figure 11: Executed PySpark code

5. The output of PySpark code executed is stored in the same S3 bucket under the folder application_data_nonull_and_eliminated_outliers


aws
Services
Search
[Alt+S]
Global
Sarath Mukundan Adavakkat
Amazon S3
Buckets
Access Grants New
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets > cloud-assignment-2023 > application_data_nonull_and_eliminated_outliers/
application_data_nonull_and_eliminated_outliers/
Copy S3 URI
Objects
Properties
Objects (2) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
_SUCCESS	-	December 18, 2023, 13:35:42 (UTC+00:00)	0 B	Standard
part-00000-b903253c-9416-4b4b-bb59-45b44bca0530-c000.csv	csv	December 18, 2023, 13:35:41 (UTC+00:00)	1.7 MB	Standard

Figure 12: Output stored in S3 bucket