EMR CLUSTER
&
DATA PROCESSING

Mentored By :

MD Sarfaraz Ahmad (Sr. Manager)

Presented By :

Mandar Pimparkar (CSD- Intern (2261459)

Project Group P1

2023
Cognizant

2023
Cognizant

PROBLEM STATEMENT

Our Client wants to get a set up of an EMR cluster to process and analyze large datasets using big data frameworks like Spark, and also needs clear instructions on how to launch a sample cluster using Spark, and how to run a simple PySpark script that will be stored in an Amazon S3 bucket.
The instructions should cover the essential tasks in three main workflow categories: Plan and Configure, Manage, and Clean Up. This will allow the company to focus on data analysis and insights rather than spending hours setting up the infrastructure for data processing.

MEET OUR TEAM!

LIKHITESH A L	**PRODUCT OWNER**
K V SAI ROHITH	**SCRUM MASTER**
NEHA WAIKAR	**DELIVERY TEAM**
YASH BUTY	**DELIVERY TEAM**
MANDAR PIMPARKAR	**DELIVERY TEAM**
YASH GHARDE	**DELIVERY TEAM**

ROLES PERFORMED:

- **PRODUCT OWNER**
  - PERFORMED CEREMONIES OF THE PRODUCT OWNER.

2. **DELIVERY TEAM**
   - WORKED AS A BACKEND DEVELOPER.
     - WORKED ON AWS SERVICES
       - AWS SNS
       - AWS S3
     - WORKED ON THE CREATION OF A SIMPLE USER INTERFACE
     - WORKED ON DATA PROCESSING USING DATABRICKS

USER STORIES*

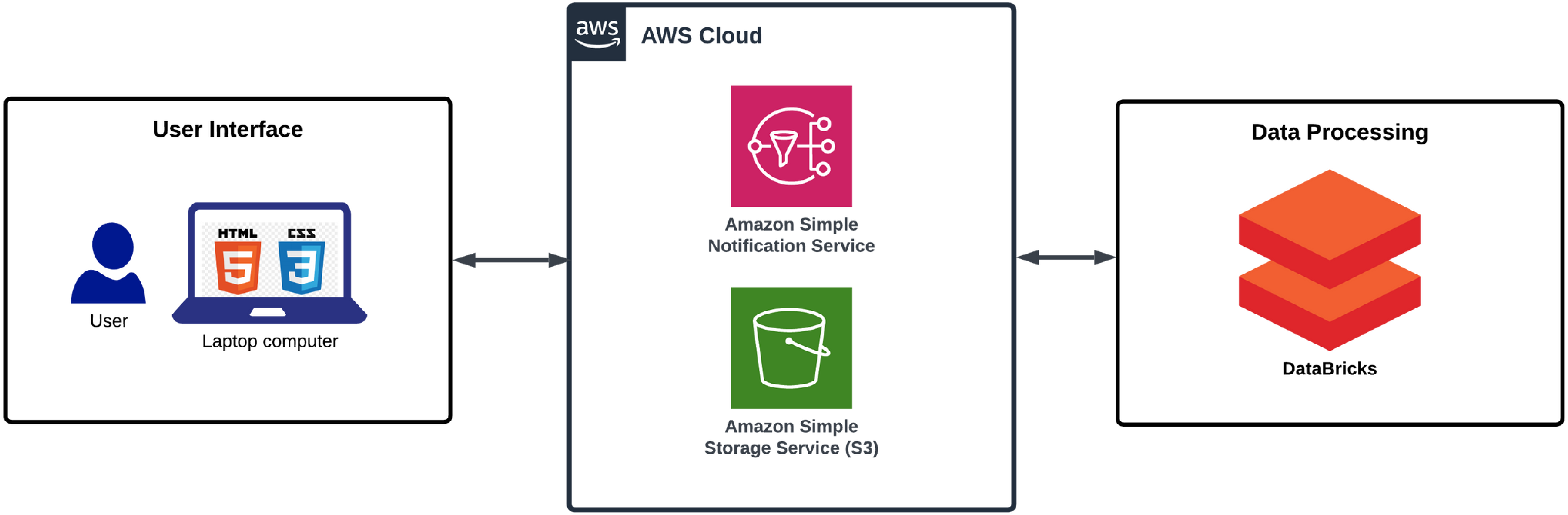
- As a client, I should be able to upload and retrieve the data using UI so that the team can analyze and give processed data for making good business decisions.
- As a team, we need to create two S3 Buckets so that the data can be uploaded and retrieved by the client.
- As a team, we need to create a simple user interface and connect it to both S3 Buckets so that client data is directly stored in the S3 bucket.
- As a team, we should create a notification service using Amazon SNS, so that we get notified once the data is uploaded by the client inside the bucket.

* I WORKED ON

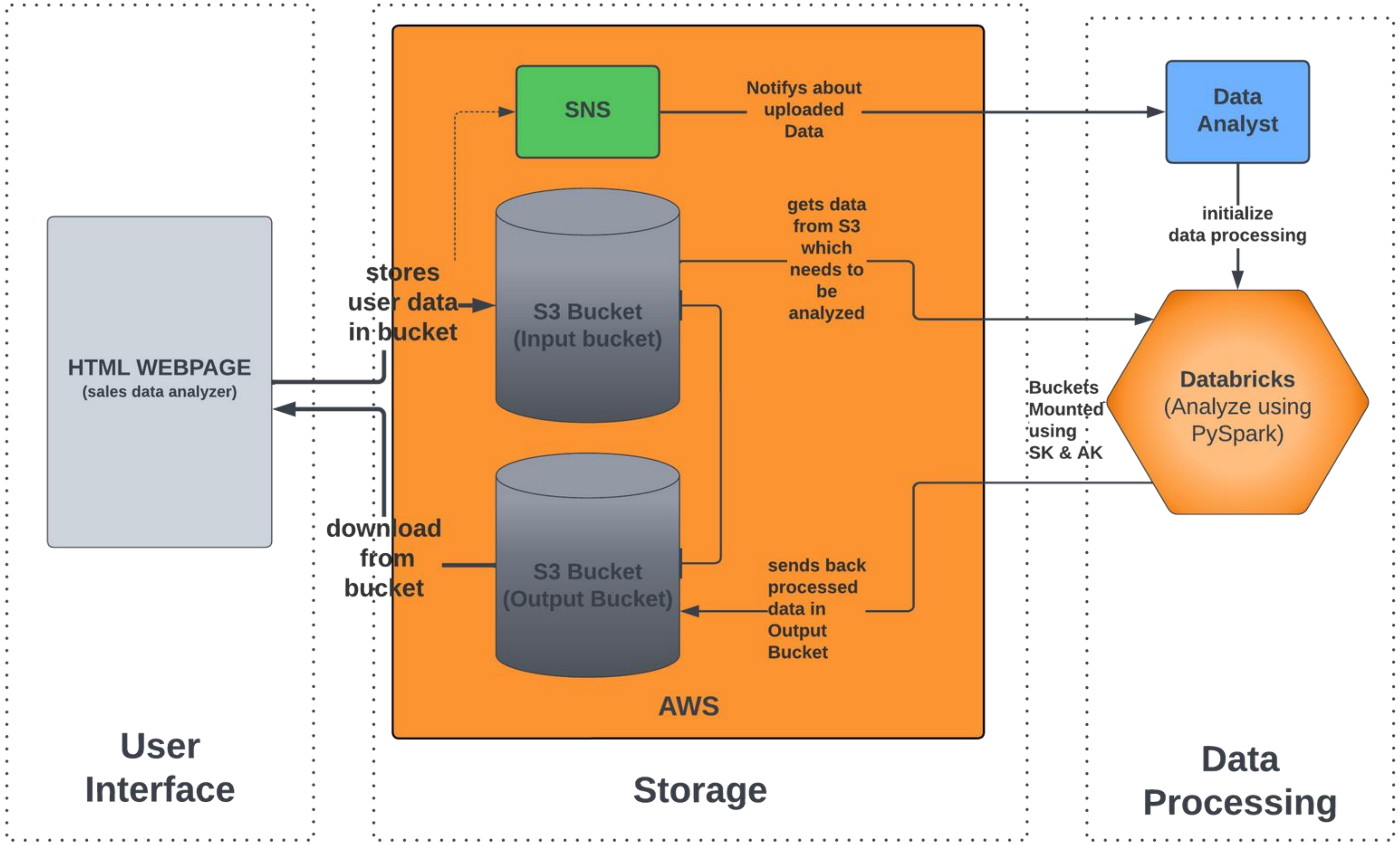
TECH STACK

USER STORY	TECHNOLOGY USED	CHALLENGES(if any)	ACCEPTANCE CRITERIA	STORY POINTS
As a client, I should be able to upload and retrieve the data using UI so that the team can analyze and give processed data for making good business decisions	• HTML & CSS	• NA	• The system should provide a user-friendly interface that allows the client to easily navigate and interact with the data. • Data should be in .csv	5
As a team, we need to create two S3 Buckets so that the data can be uploaded and retrieved by the client	• AWS S3	• NA	• Configure S3 according to client requirement.	3
As a team, we need to create a simple user interface and connect it to both S3 Buckets so that client data is directly stored in the S3 bucket.	• HTML & CSS • AWS S3	• using FLASK to connect webpage with S3 buckets	• It should be simple to use and can hold only .csv files	5
As a team , we should create a notification service using Amazon SNS, so that we get notified once the data is uploaded by client inside the bucket.	• AWS S3 • AWS SNS	• NA	• we should get a notification as soon as the data is uploaded by the user	5

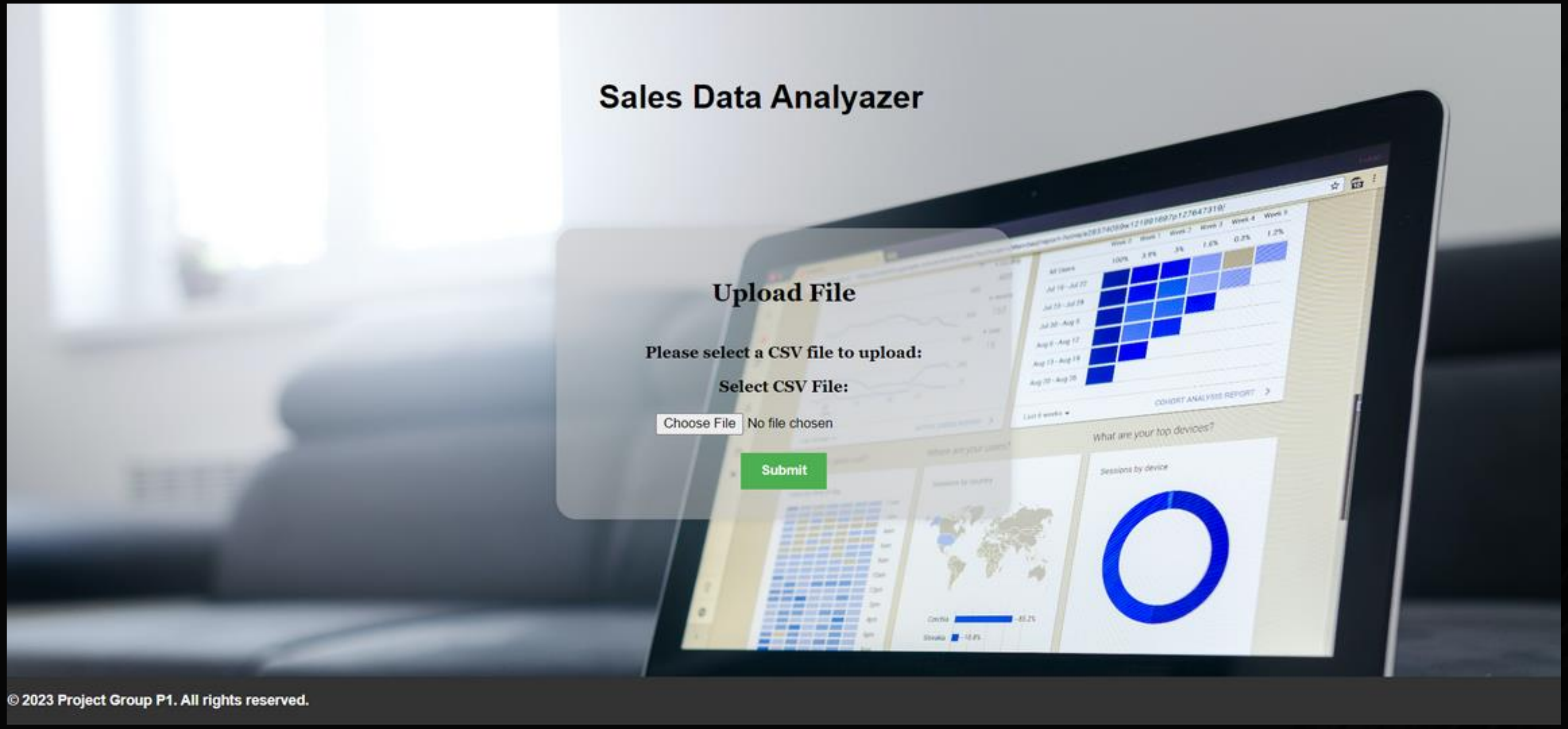
METHODOLOGY



PROJECT IMPLEMENTATION


User Interface
HTML WEBPAGE
(sales data analyzer)
stores user data in bucket
download from bucket
SNS
S3 Bucket (Input bucket)
S3 Bucket (Output Bucket)
AWS
Storage
Notifys about uploaded Data
gets data from S3 which needs to be analyzed
sends back processed data in Output Bucket
Buckets Mounted using SK & AK
Data Analyst
initialize data processing
Databricks (Analyze using PySpark)
Data Processing

USER INTERFACE



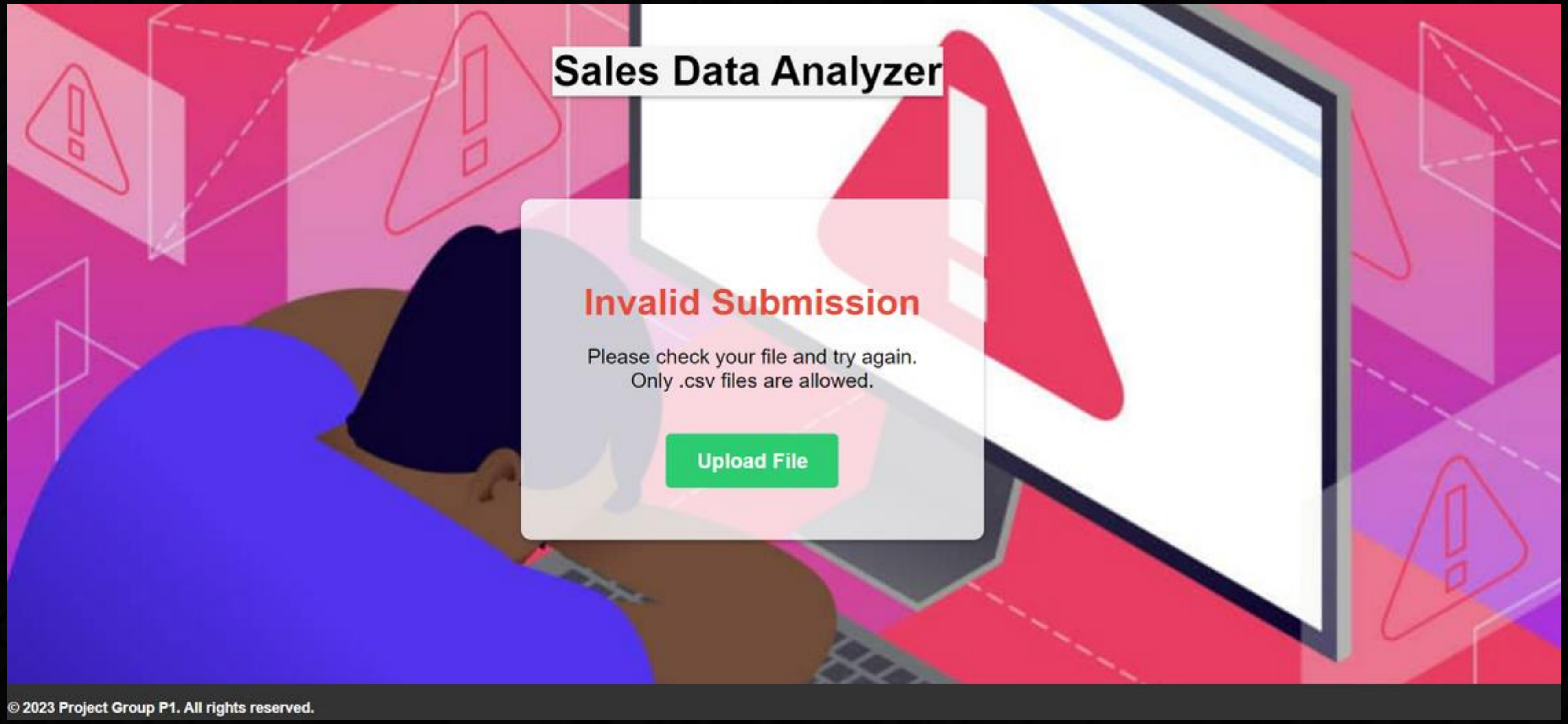
Sales Data Analyazer

Upload File

Please select a CSV file to upload:

Select CSV File:

© 2023 Project Group P1. All rights reserved.



Sales Data Analyzer

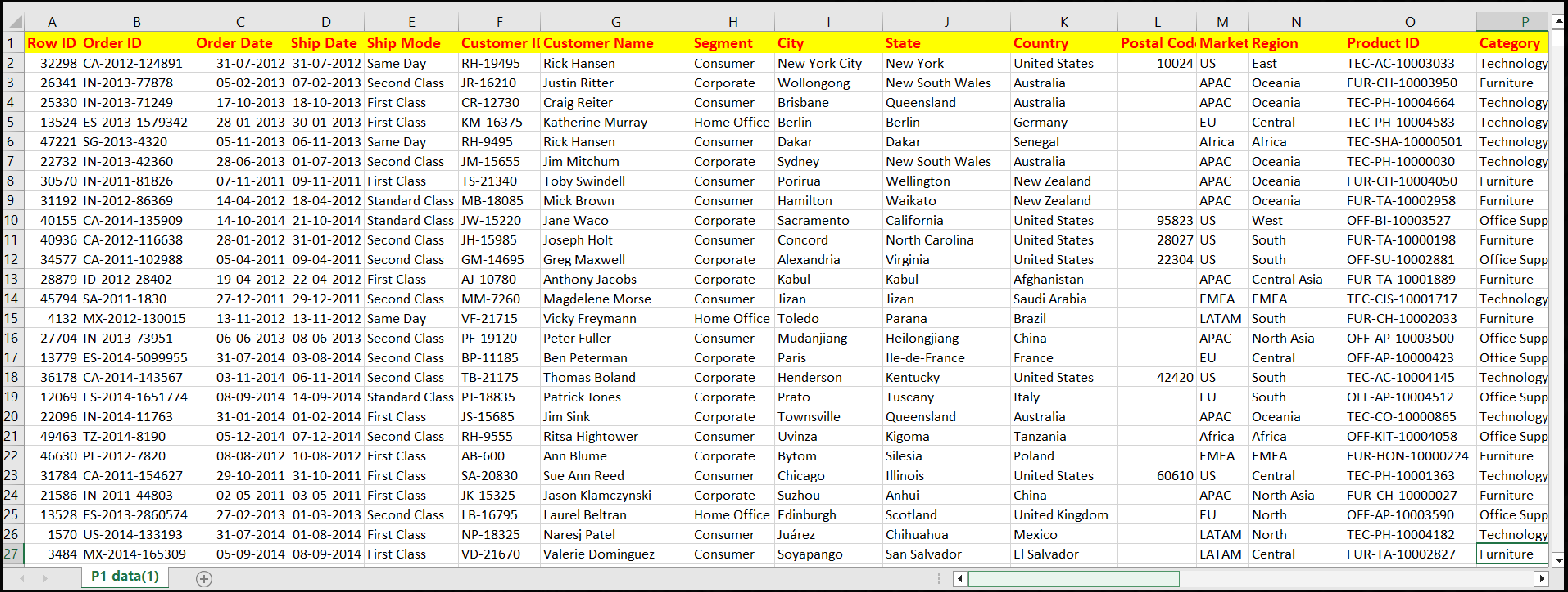
Invalid Submission

Please check your file and try again.
Only .csv files are allowed.

© 2023 Project Group P1. All rights reserved.

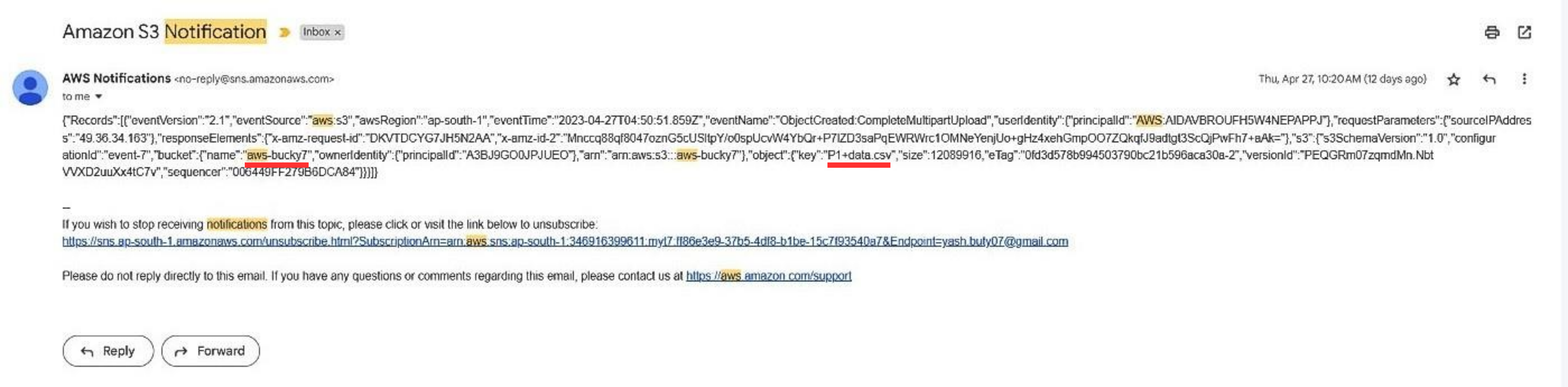
DATASET USED

P1 Data.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer II	Customer Name	Segment	City	State	Country	Postal Cod	Market	Region	Product ID	Category
2	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	United States	10024	US	East	TEC-AC-10003033	Technology
3	26341	IN-2013-77878	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales	Australia		APAC	Oceania	FUR-CH-10003950	Furniture
4	25330	IN-2013-71249	17-10-2013	18-10-2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland	Australia		APAC	Oceania	TEC-PH-10004664	Technology
5	13524	ES-2013-1579342	28-01-2013	30-01-2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin	Germany		EU	Central	TEC-PH-10004583	Technology
6	47221	SG-2013-4320	05-11-2013	06-11-2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar	Senegal		Africa	Africa	TEC-SHA-10000501	Technology
7	22732	IN-2013-42360	28-06-2013	01-07-2013	Second Class	JM-15655	Jim Mitchum	Corporate	Sydney	New South Wales	Australia		APAC	Oceania	TEC-PH-10000030	Technology
8	30570	IN-2011-81826	07-11-2011	09-11-2011	First Class	TS-21340	Toby Swindell	Consumer	Porirua	Wellington	New Zealand		APAC	Oceania	FUR-CH-10004050	Furniture
9	31192	IN-2012-86369	14-04-2012	18-04-2012	Standard Class	MB-18085	Mick Brown	Consumer	Hamilton	Waikato	New Zealand		APAC	Oceania	FUR-TA-10002958	Furniture
10	40155	CA-2014-135909	14-10-2014	21-10-2014	Standard Class	JW-15220	Jane Waco	Corporate	Sacramento	California	United States	95823	US	West	OFF-BI-10003527	Office Supp
11	40936	CA-2012-116638	28-01-2012	31-01-2012	Second Class	JH-15985	Joseph Holt	Consumer	Concord	North Carolina	United States	28027	US	South	FUR-TA-10000198	Furniture
12	34577	CA-2011-102988	05-04-2011	09-04-2011	Second Class	GM-14695	Greg Maxwell	Corporate	Alexandria	Virginia	United States	22304	US	South	OFF-SU-10002881	Office Supp
13	28879	ID-2012-28402	19-04-2012	22-04-2012	First Class	AJ-10780	Anthony Jacobs	Corporate	Kabul	Kabul	Afghanistan		APAC	Central Asia	FUR-TA-10001889	Furniture
14	45794	SA-2011-1830	27-12-2011	29-12-2011	Second Class	MM-7260	Magdelene Morse	Consumer	Jizan	Jizan	Saudi Arabia		EMEA	EMEA	TEC-CIS-10001717	Technology
15	4132	MX-2012-130015	13-11-2012	13-11-2012	Same Day	VF-21715	Vicky Freymann	Home Office	Toledo	Parana	Brazil		LATAM	South	FUR-CH-10002033	Furniture
16	27704	IN-2013-73951	06-06-2013	08-06-2013	Second Class	PF-19120	Peter Fuller	Consumer	Mudanjiang	Heilongjiang	China		APAC	North Asia	OFF-AP-10003500	Office Supp
17	13779	ES-2014-5099955	31-07-2014	03-08-2014	Second Class	BP-11185	Ben Peterman	Corporate	Paris	Ile-de-France	France		EU	Central	OFF-AP-10000423	Office Supp
18	36178	CA-2014-143567	03-11-2014	06-11-2014	Second Class	TB-21175	Thomas Boland	Corporate	Henderson	Kentucky	United States	42420	US	South	TEC-AC-10004145	Technology
19	12069	ES-2014-1651774	08-09-2014	14-09-2014	Standard Class	PJ-18835	Patrick Jones	Corporate	Prato	Tuscany	Italy		EU	South	OFF-AP-10004512	Office Supp
20	22096	IN-2014-11763	31-01-2014	01-02-2014	First Class	JS-15685	Jim Sink	Corporate	Townsville	Queensland	Australia		APAC	Oceania	TEC-CO-10000865	Technology
21	49463	TZ-2014-8190	05-12-2014	07-12-2014	Second Class	RH-9555	Ritsa Hightower	Consumer	Uvinza	Kigoma	Tanzania		Africa	Africa	OFF-KIT-10004058	Office Supp
22	46630	PL-2012-7820	08-08-2012	10-08-2012	First Class	AB-600	Ann Blume	Corporate	Bytom	Silesia	Poland		EMEA	EMEA	FUR-HON-10000224	Furniture
23	31784	CA-2011-154627	29-10-2011	31-10-2011	First Class	SA-20830	Sue Ann Reed	Consumer	Chicago	Illinois	United States	60610	US	Central	TEC-PH-10001363	Technology
24	21586	IN-2011-44803	02-05-2011	03-05-2011	First Class	JK-15325	Jason Klamczynski	Corporate	Suzhou	Anhui	China		APAC	North Asia	FUR-CH-10000027	Furniture
25	13528	ES-2013-2860574	27-02-2013	01-03-2013	Second Class	LB-16795	Laurel Beltran	Home Office	Edinburgh	Scotland	United Kingdom		EU	North	OFF-AP-10003590	Office Supp
26	1570	US-2014-133193	31-07-2014	01-08-2014	First Class	NP-18325	Naresj Patel	Consumer	Juárez	Chihuahua	Mexico		LATAM	North	TEC-PH-10004182	Technology
27	3484	MX-2014-165309	05-09-2014	08-09-2014	First Class	VD-21670	Valerie Dominguez	Consumer	Soyapango	San Salvador	El Salvador		LATAM	Central	FUR-TA-10002827	Furniture

SNS TRIGGERD



Amazon S3 Notification Inbox ×

AWS Notifications <no-reply@sns.amazonaws.com>
to me

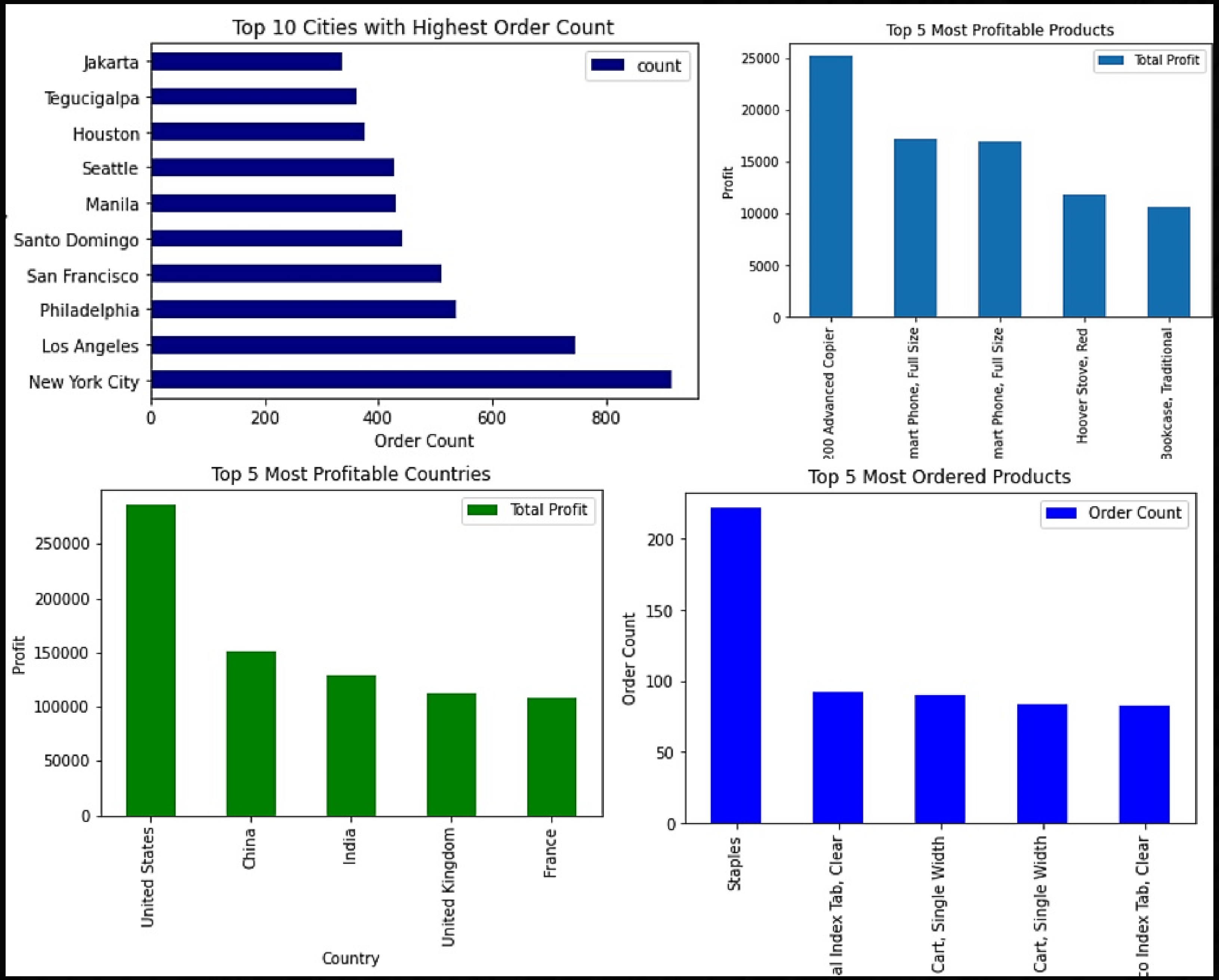
Thu, Apr 27, 10:20 AM (12 days ago)

```
{"Records":[{"eventVersion":"2.1","eventSource":"aws:s3","awsRegion":"ap-south-1","eventTime":"2023-04-27T04:50:51.859Z","eventName":"ObjectCreated:CompleteMultipartUpload","userIdentity":{"principalId":"AWS:AIDAVBROUFH5W4NEPAPPJ"},"requestParameters":{"sourceIPAddres
s":"49.36.34.163"},"responseElements":{"x-amz-request-id":"DKVTDCYG7JH5N2AA","x-amz-id-2":"Mnccq88qf8047oznG5cUSltpY/o0spUcvW4YbQr+P7lZD3saPqEWRWrc1OMNeYenjUo+gHz4xehGmpOO7ZQkqfJ9adtgt3ScQjPwFh7+aAk="},"s3":{"s3SchemaVersion":"1.0","configur
ationId":"event-7","bucket":{"name":"aws-bucky7","ownerIdentity":{"principalId":"A3BJ9GO0JPJUEO"},"arn":"arn:aws:s3:::aws-bucky7"},"object":{"key":"P1+data.csv","size":12089916,"eTag":"0fd3d578b994503790bc21b596aca30a-2","versionId":"PEQGRm07zqmdMn.Nbt
VVXD2uuXx4tC7v","sequencer":"006449FF279B6DCA84"}}}]}
```

--
If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
https://sns.ap-south-1.amazonaws.com/unsubscribe.html?SubscriptionArn=arn:aws:sns:ap-south-1:346916399611:myt7:ff86e3e9-37b5-4df8-b1be-15c7f93540a7&Endpoint=yash.buty07@gmail.com

Please do not reply directly to this email. If you have any questions or comments regarding this email, please contact us at https://aws.amazon.com/support

ANALYZED DATA



2023
Cognizant

PROJECT GROUP P1

APPENDIX

USER STORIES

- As a client, I should be able to upload and retrieve the data using UI so that the team can analyze and give processed data for making good business decisions. (3)*

- As a team, we need to create two S3 Buckets so that the data can be uploaded and retrieved by the client. (3)*

- As a team, we need to create a simple user interface and connect it to both S3 Buckets so that client data is directly stored in the S3 bucket. (3)*

- As a team, we should create a notification service using Amazon SNS, so that we get notified once the data is uploaded by the client inside the bucket. (3)*

USER STORIES

- As a team, we should be able to create and launch a cluster on Databricks so that we can process and analyze the user data. (3)*
- As a team, we should be able to mount both the S3 buckets on DataBricks, so that we can access user data. (5)*
- As a team, we should be able to write a PySpark script for analyzing the data according to user requirements. (3)*
- As a team, we should be able to upload analyzed data from Databricks to the S3 bucket and display it on UI for client usage. (3)*

PRODUCT BACKLOG

1 EPIC

Setting up S3 Buckets

USER STORY:

- Setting up two S3 buckets
- Creating a notification service for new data uploaded in Bucket using Amazon SNS.

2 EPIC

Data Manipulation using DataBricks

USER STORY:

- Creating and launching cluster on DataBricks
- Mounting Both S3 buckets on DataBricks
- Creating PySpark Script to analyze data according to user Requirements
- Uploading analyzed data from DataBricks to S3 Bucket

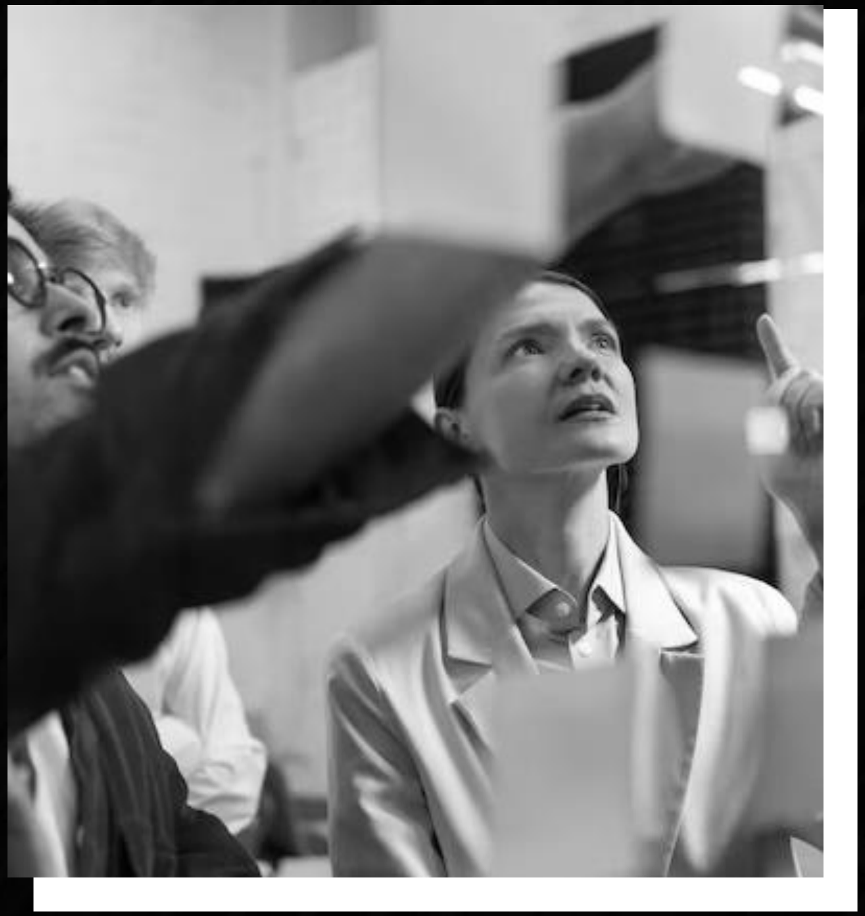
PRODUCT BACKLOG

❸ EPIC

Data Transfer using UI

USER STORY:

- Creating a User Interface
- Connecting S3 bucket with UI for Data Transfer



❹ EPIC

Product Testing & Documentation

USER STORY:

- Performing Unit Testing
- Performing Performance Testing
- Performing Integration Testing
- Create Closure documents
- Create SDD Documents

TECH STACK

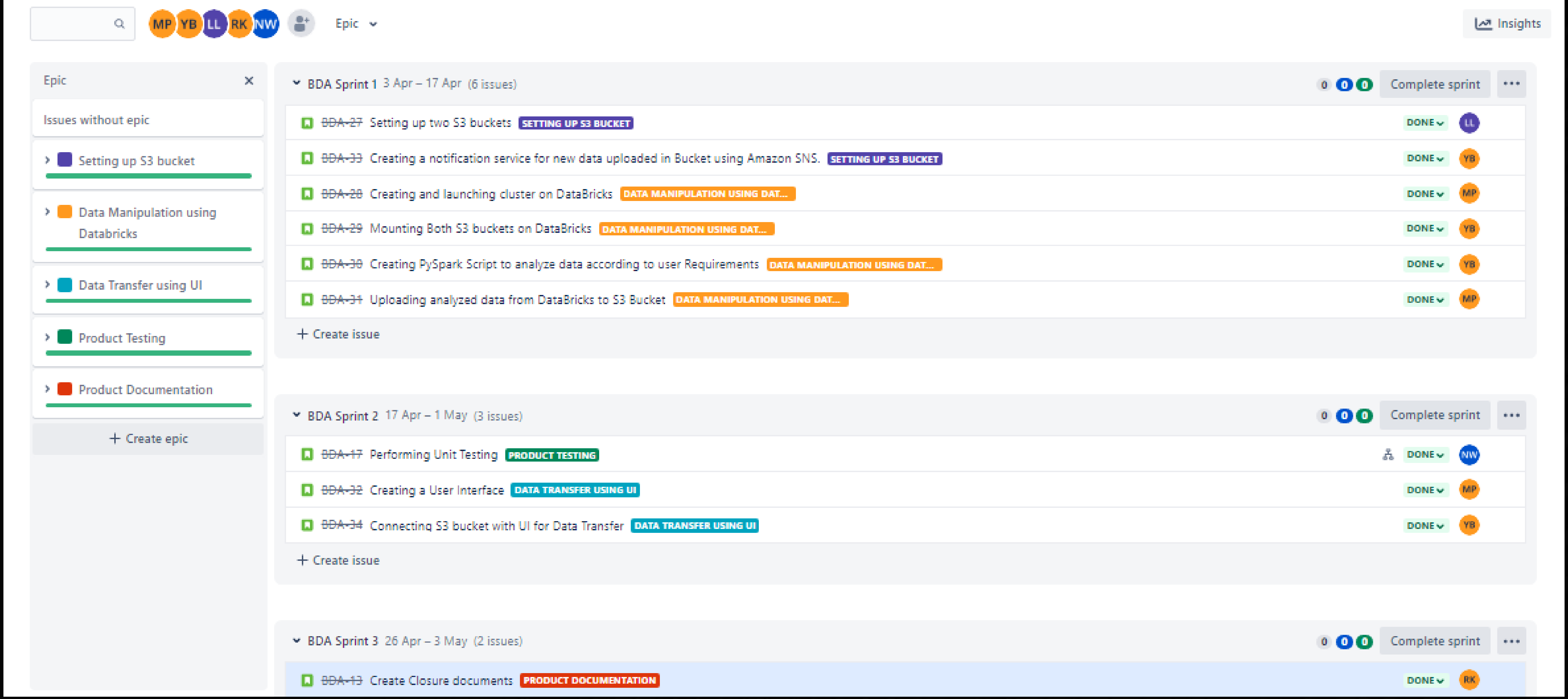
USER STORY	TECHNOLOGY USED	CHALLENGES(if any)	ACCEPTANCE CRITERIA	STORY POINTS
As a client, I should be able to upload and retrieve the data using UI so that the team can analyze and give processed data for making good business decisions	• HTML & CSS	• NA	• The system should provide a user-friendly interface that allows the client to easily navigate and interact with the data. • Data should be in .csv	3
As a team, we need to create two S3 Buckets so that the data can be uploaded and retrieved by the client	• AWS S3	• NA	• Configure S3 according to client requirement.	3
As a team, we need to create a simple user interface and connect it to both S3 Buckets so that client data is directly stored in the S3 bucket.	• HTML & CSS • AWS S3	• NA	• It should be simple to use and can hold only .csv files	5
As a team , we should create a notification service using Amazon SNS, so that we get notified once the data is uploaded by client inside the bucket.	• AWS S3 • AWS SNS	• NA	• we should get a notification as soon as the data is uploaded by the user	5

TECH STACK

USER STORY	TECHNOLOGY USED	CHALLENGES (if any)	ACCEPTANCE CRITERIA	STORY POINTS
As a team, we should be able to create and launch a cluster on Databricks so that we can process and analyze the user data.	• DATABRICKS	• using a free version dont allow us to keep our cluster active all the time • as emr access was not provided to us we need to find an alternative	• Cluster needs to be active all the time	3
As a team, we should be able to mount both the S3 buckets on DataBricks, so that we can access user data	• AWS S3 • DATABRICKS	• NA	• user data should be directly fetched from S3 buckets	3
As a team, we should be able to write a PySpark script for analyzing the data according to user requirement.	• DATABRICKS	• NA	• should be able to code and give results accoerding to user needs.	5
As a team, we should be able to upload analyzed data from Databricks to the S3 bucket and display it on UI for client usage.	• AWS S3 • DATABRICKS • HTML & CSS	• NA	• result should be easily accessiblke by the client	5

JIRA DASHBOARD

click here →



Issues without epic

- Setting up S3 bucket
- Data Manipulation using Databricks
- Data Transfer using UI
- Product Testing
- Product Documentation

BDA Sprint 1 3 Apr – 17 Apr (6 issues) 0 0 0 Complete sprint

- ~~BDA-27~~ Setting up two S3 buckets SETTING UP S3 BUCKET DONE LL
- ~~BDA-33~~ Creating a notification service for new data uploaded in Bucket using Amazon SNS. SETTING UP S3 BUCKET DONE YB
- ~~BDA-28~~ Creating and launching cluster on DataBricks DATA MANIPULATION USING DAT... DONE MP
- ~~BDA-29~~ Mounting Both S3 buckets on DataBricks DATA MANIPULATION USING DAT... DONE YB
- ~~BDA-30~~ Creating PySpark Script to analyze data according to user Requirements DATA MANIPULATION USING DAT... DONE YB
- ~~BDA-31~~ Uploading analyzed data from DataBricks to S3 Bucket DATA MANIPULATION USING DAT... DONE MP

BDA Sprint 2 17 Apr – 1 May (3 issues) 0 0 0 Complete sprint

- ~~BDA-17~~ Performing Unit Testing PRODUCT TESTING DONE NW
- ~~BDA-32~~ Creating a User Interface DATA TRANSFER USING UI DONE MP
- ~~BDA-34~~ Connecting S3 bucket with UI for Data Transfer DATA TRANSFER USING UI DONE YB

BDA Sprint 3 26 Apr – 3 May (2 issues) 0 0 0 Complete sprint

- ~~BDA-13~~ Create Closure documents PRODUCT DOCUMENTATION DONE RK

CEREMONIES OF PRODUCT OWNER

Sprint Planning

- Main role play product owner and scrum master
- Planning for each sprint and sprint estimation is done

Creating backlog

- The main role is played by the product owner
- The user story and story point estimation take place

Product Grooming

- The main role is played by the product owner with other members of the team before each sprint starts

Reviewing Sprint

- This is done together with Scrum so as to see where the project requirements are met

Serving as Primary Contact

- The product owner works as the main contact between the client and the team members

CEREMONIES OF SCRUM MASTER

Sprint Planning

- Main role play product owner and scrum master
- Planning done by product owner and work assign to team by scrum master.

Daily Stand-up

- Daily meeting arrange by scrum master 15 min for taking updates.

Sprint Review

- Meeting lead by scrum master and taking review from deployment team.

Sprint Retrospective

- Meeting lead by both scrum master and product owner reviewing what is being implemented in sprint and is there room for improvement.

Product Backlog Grooming

PRODUCT BACKLOG GROOMING

- **This is a meeting held during sprint about the coming backlog.**
- **Main people**
  - Scrum master
  - Product Owner
- **Lead by Product Owner**
- **Points to be discussed:**
  - What is coming in the next sprint?
  - Discussion with the development team.
  - Breaking down broad user stories into smaller items.
  - Identifying roadblocks and minimizing risks related to backlog items.

BEST CODE PRACTICES

Variable, Class and Function Naming convention

Clear and Concise comments

Code Indentation

Reusability and Scalability

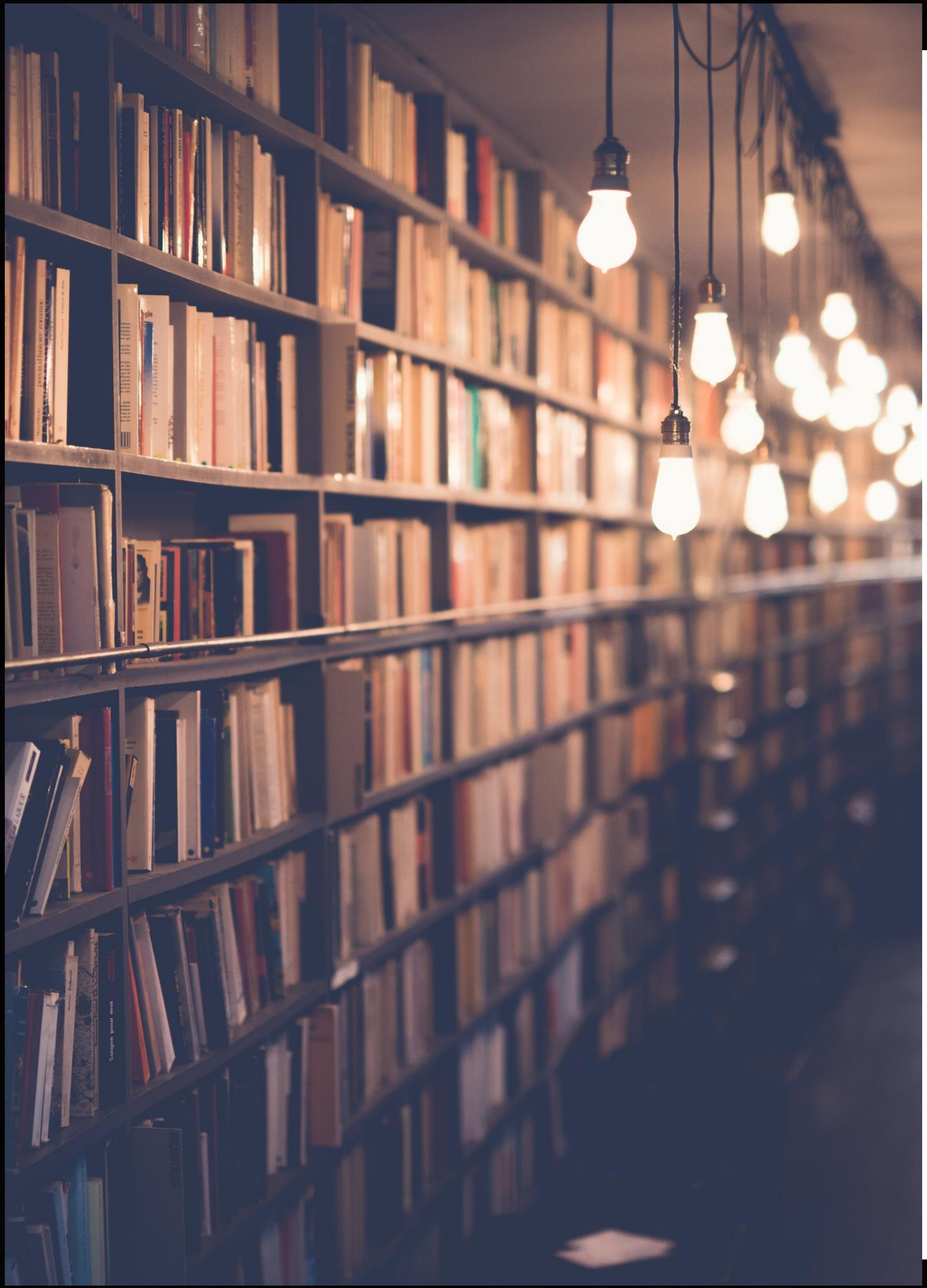
DRY Principle

PROJECT DOCUMENTATION:

GitHub Repository

Agile Dashboard

Onedrive



THANK YOU