TEAM TRIUMPH BRIGADE

EMR CLUSTER
&
DATA PROCESSING

Presented by-
Yash Buty
Developer (Delivery Team)

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 Apache Spark and 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

CEREMONIES OF BACKEND TEAM

1. EXECUTE THE TASK ASSIGNED BY SCRUM MASTER WITHIN DEADLINE

2. PROVIDE UPDATES TO SCRUM MASTER IN DAILY STANDUPS

BACKEND TEAM:

- WORKED ON AWS SERVICES
  - AWS S3
  - AWS SNS

- WORKED ON DATA PROCESSING AND ANALYSIS USING DATABRICKS

cognizant

USER STORIES

- As a developer, we need to create two S3 Buckets so that the data can be uploaded and retrieved by the client.

- As a developer, 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 developer, 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.

USER STORIES

- As a developer, we should be able to create and launch a cluster on Databricks so that we can process and analyze the user data.
- As a developer, we should be able to mount both the S3 buckets on DataBricks, so that we can access user data.
- As a developer, we should be able to write a PySpark script for analyzing the data according to user requirements.

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. • Client Data should be in .csv	3
As a developer, I 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 developer, I 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 easy to use and should be able to hold csv file format	5
As a developer , I 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

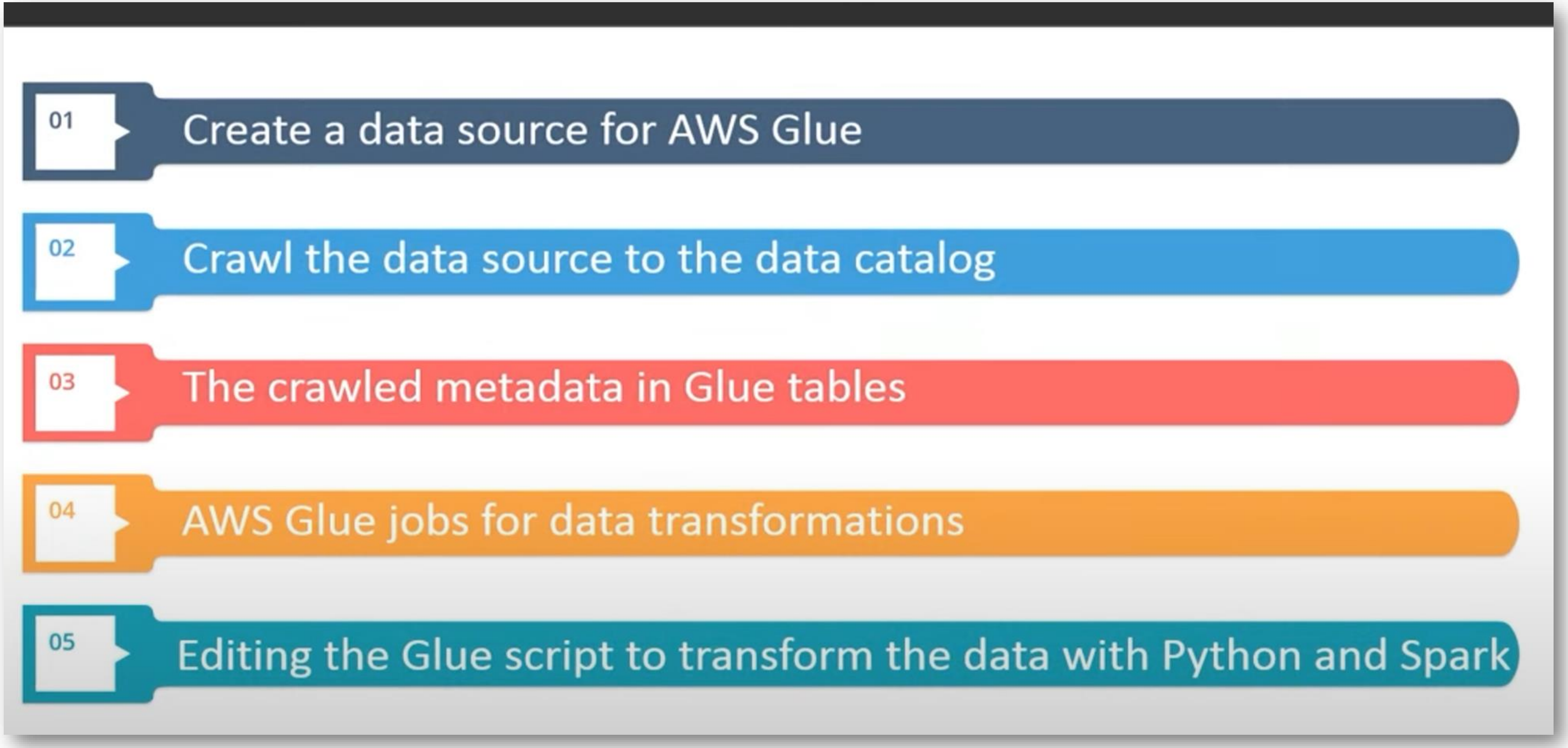
USER STORY	TECHNOLOGY USED	CHALLENGES (if any)	ACCEPTANCE CRITERIA	STORY POINTS
As a developer, I should be able to create and launch a cluster on Databricks so that we can process and analyze the user data.*	• DATABRICKS	• Community version didn't allow us to keep cluster active all the time • as EMR access was not provided to us we need to find an alternative	• Cluster needs to be active all the time	3
As a developer, I 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 developer, I should be able to write a PySpark script for analyzing the data according to user requirement.*	• DATABRICKS	▪ NA	• Output of the code should be according to client requirement.	5
As a developer, I 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	• Resultant file should be easily accessablle by the client	5

Proposed Solution

A. Using Amazon Glue

Amazon Glue

Amazon Glue: It’s a fully managed ETL service that makes it simple and cost effective to categorize your data, clean it and move it reliable between various datastores.



01 Create a data source for AWS Glue

02 Crawl the data source to the data catalog

03 The crawled metadata in Glue tables

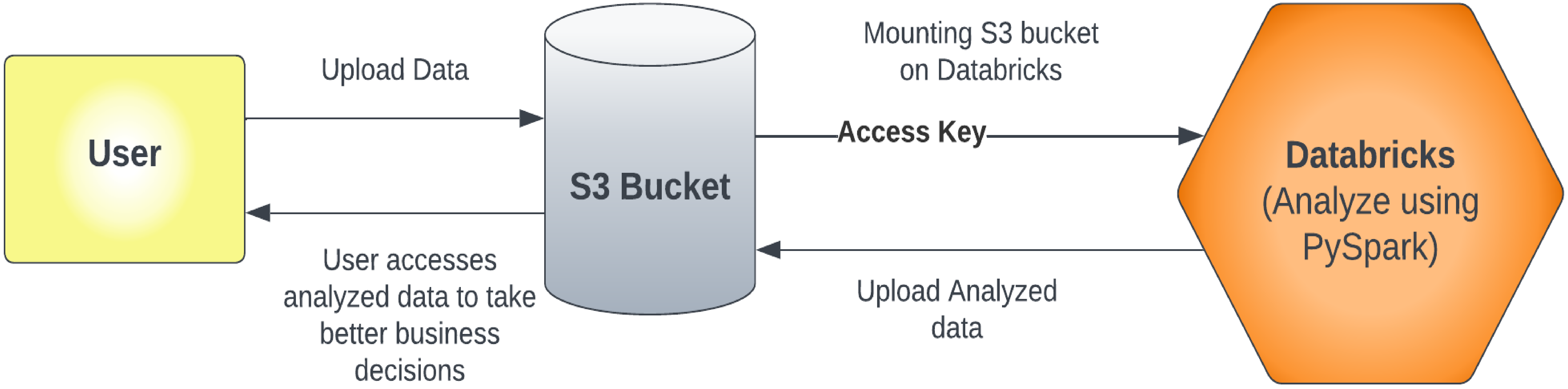
04 AWS Glue jobs for data transformations

05 Editing the Glue script to transform the data with Python and Spark

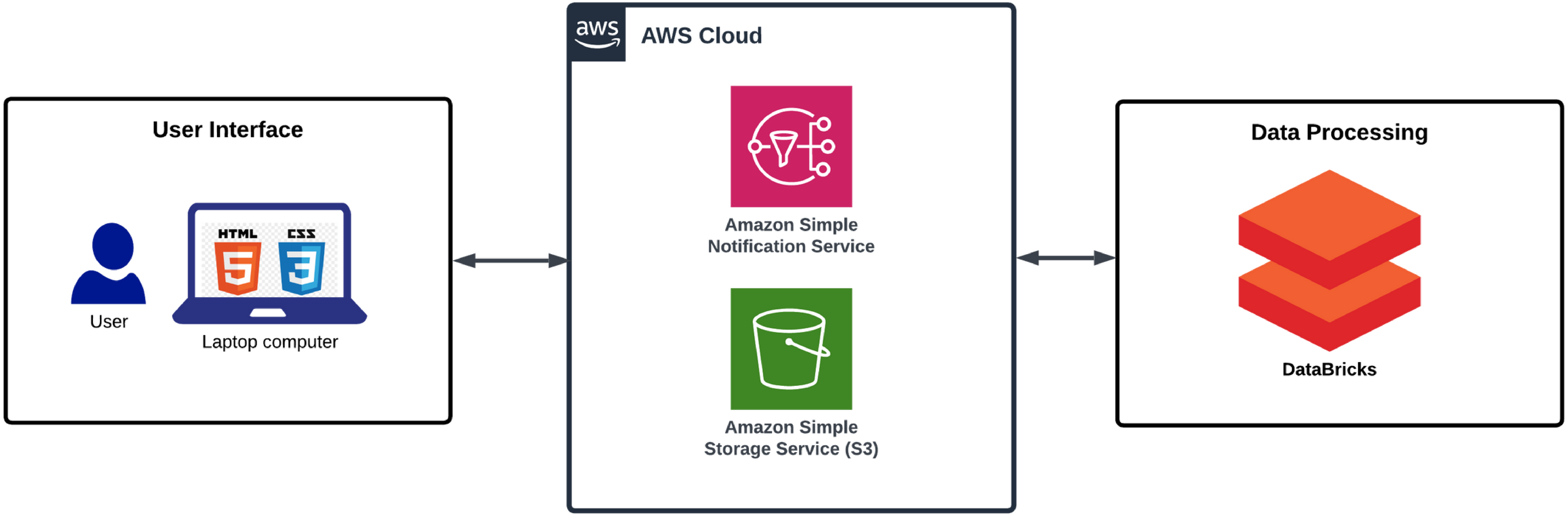
Proposed Solution

databricks

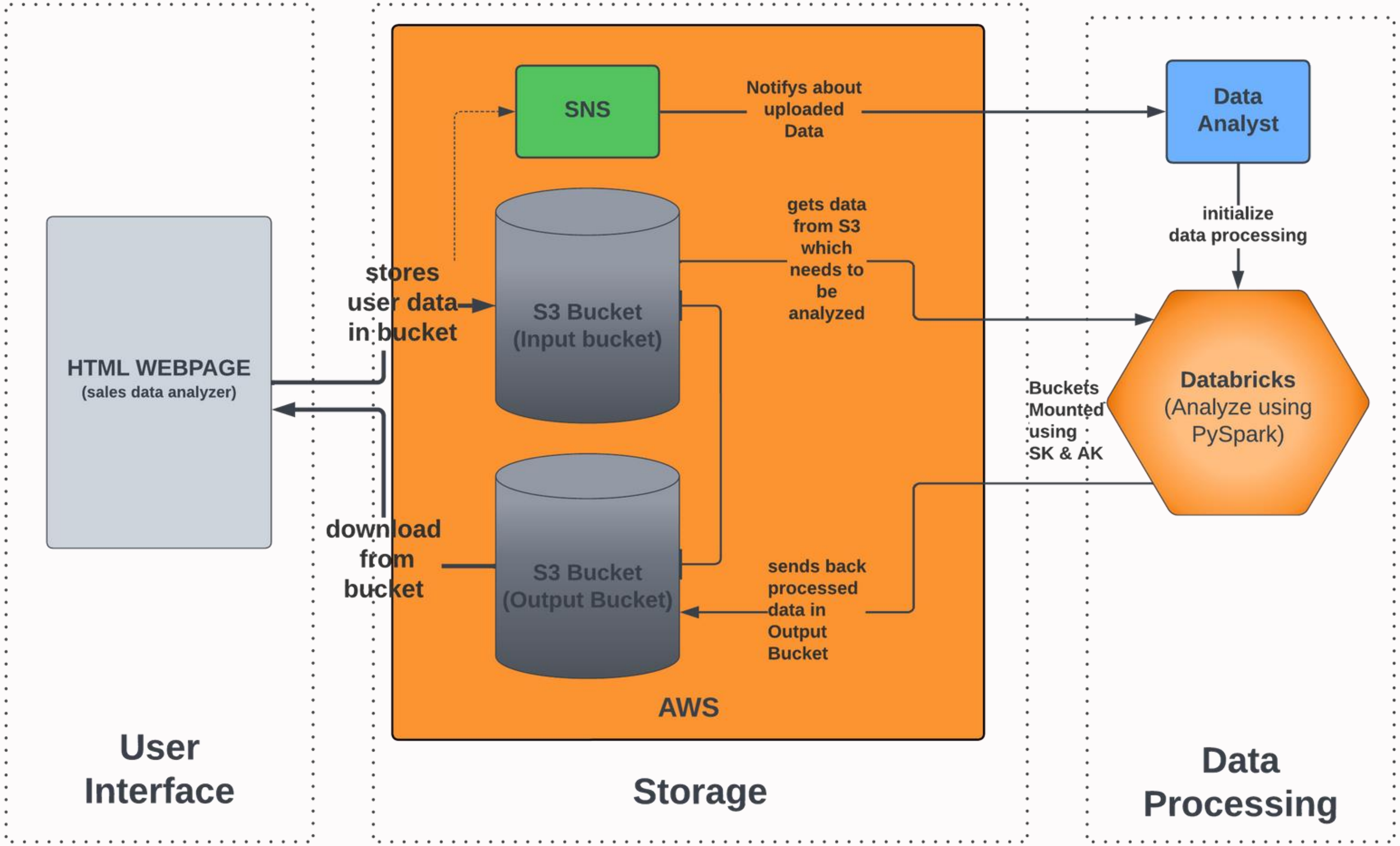
B. Using databricks + S3 Bucket


User
Upload Data
User accesses analyzed data to take better business decisions
S3 Bucket
Mounting S3 bucket on Databricks
Access Key
Upload Analyzed data
Databricks (Analyze using PySpark)

METHODOLOGY

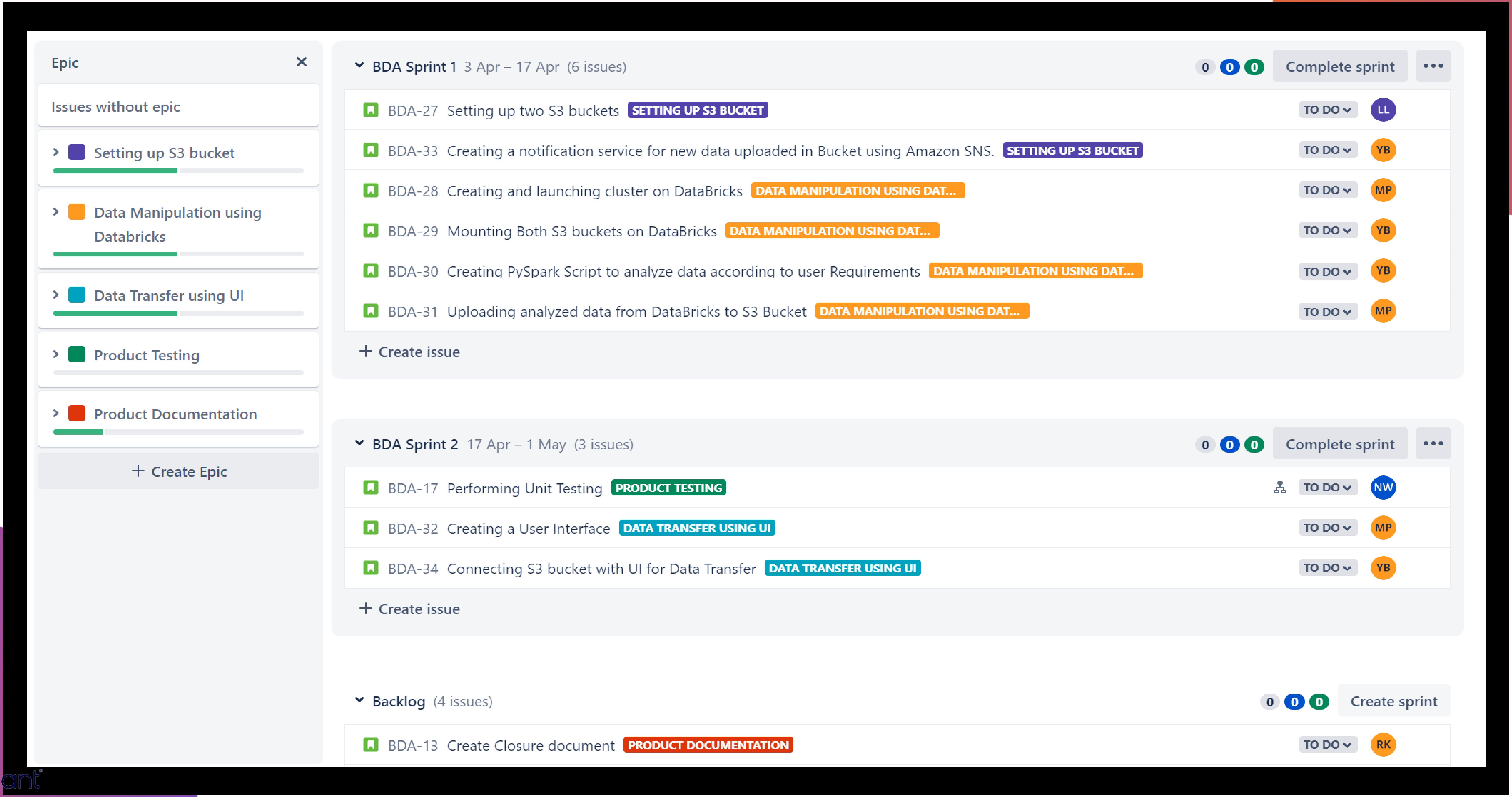


PROJECT IMPLEMENTATION



JIRA DASHBOARD

click here →



Epic ✕

Issues without epic

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

+ Create Epic

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

+ Create issue

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

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

+ Create issue

Backlog (4 issues) 0 0 0 Create sprint

BDA-13	Create Closure document	PRODUCT DOCUMENTATION	TO DO	RK

Thank You

cognizant

APPENDIX

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 takes 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

cognizant

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

Product Documentation

To access the GitHub Repository,
click here

PRODUCT BACKLOG

❶ 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.

❷ 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

cognizant

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