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**

cognizant

# 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

# 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.<br>• 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<br>• 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<br>• AWS SNS | ▪ NA | • we should get a notification as soon as the data is uploaded by the user | 5 |

# TECH STACK

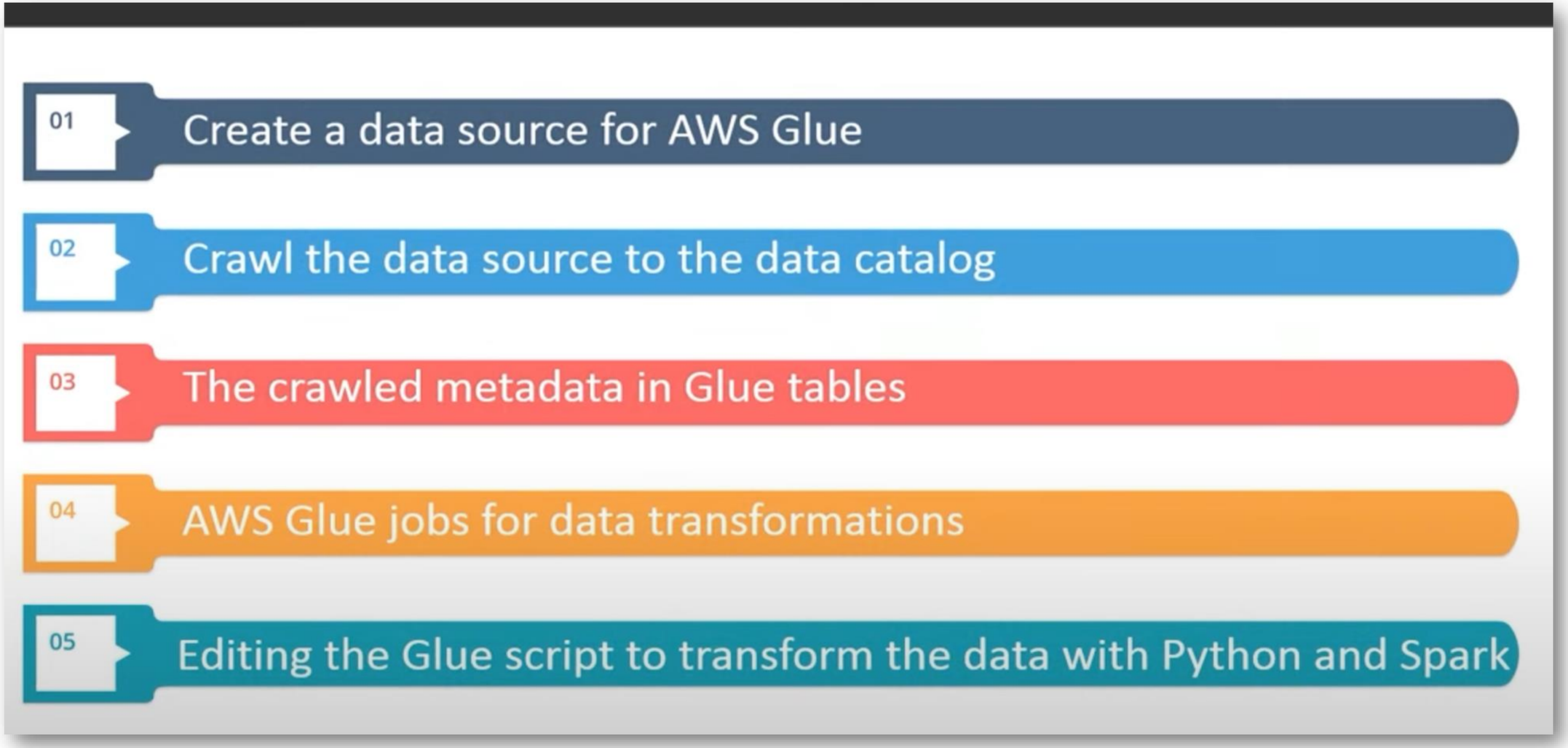
| USER STORY | TECHNOLOGY USED | CHALLENGES (if any) | ACCEPTANCE CRITERIA | STORY POINTS |
|---|---|---|---|---|
| As a developer, I should be able to create and launch a cluster on Databricks so that we can process and analyze the user data.* | • DATABRICKS | • Community version didn't allow us to keep cluster active all the time<br>• 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<br>• 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<br>• DATABRICKS<br>• HTML & CSS | ▪ NA | • Resultant file should be easily accessablle by the client | 5 |

# Proposed Solution

## A. Using Amazon Glue

Amazon Glue

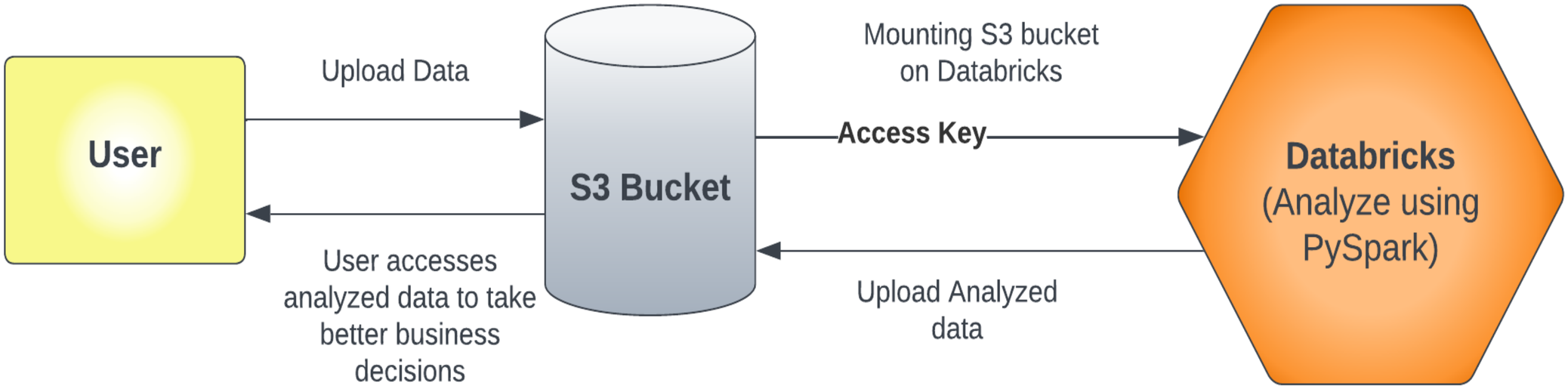
**Amazon Glue**: It's a fully managed ETL service that makes it simple and cost effective to categorize your data, clean it and move it reliable between various datastores.



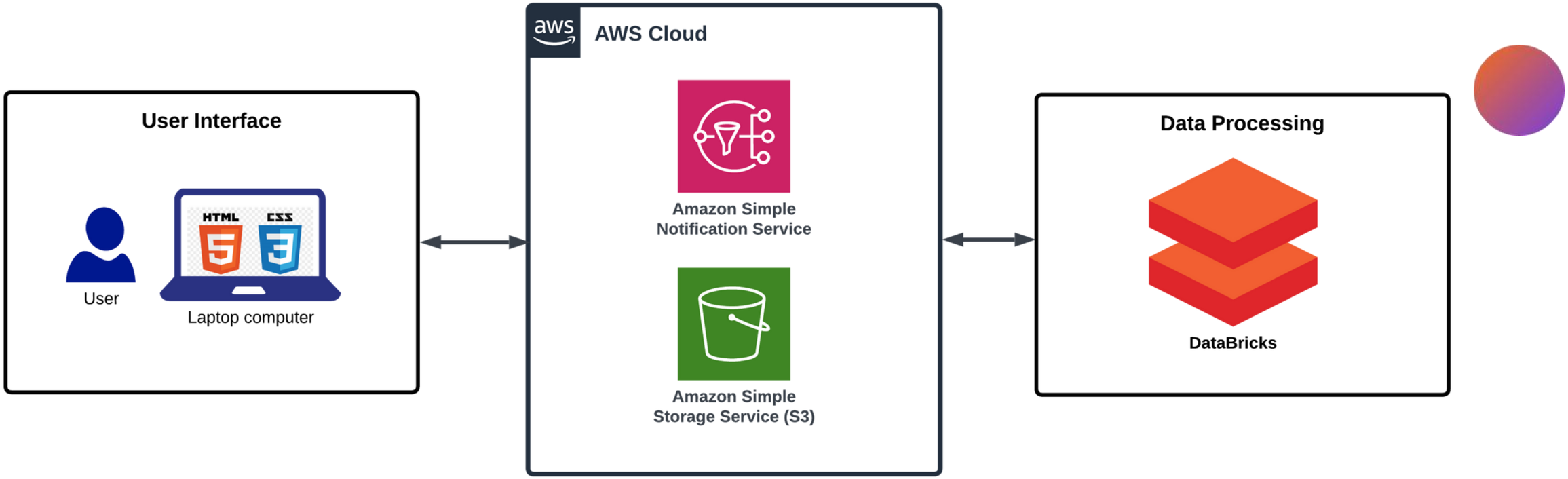
# Proposed Solution

databricks

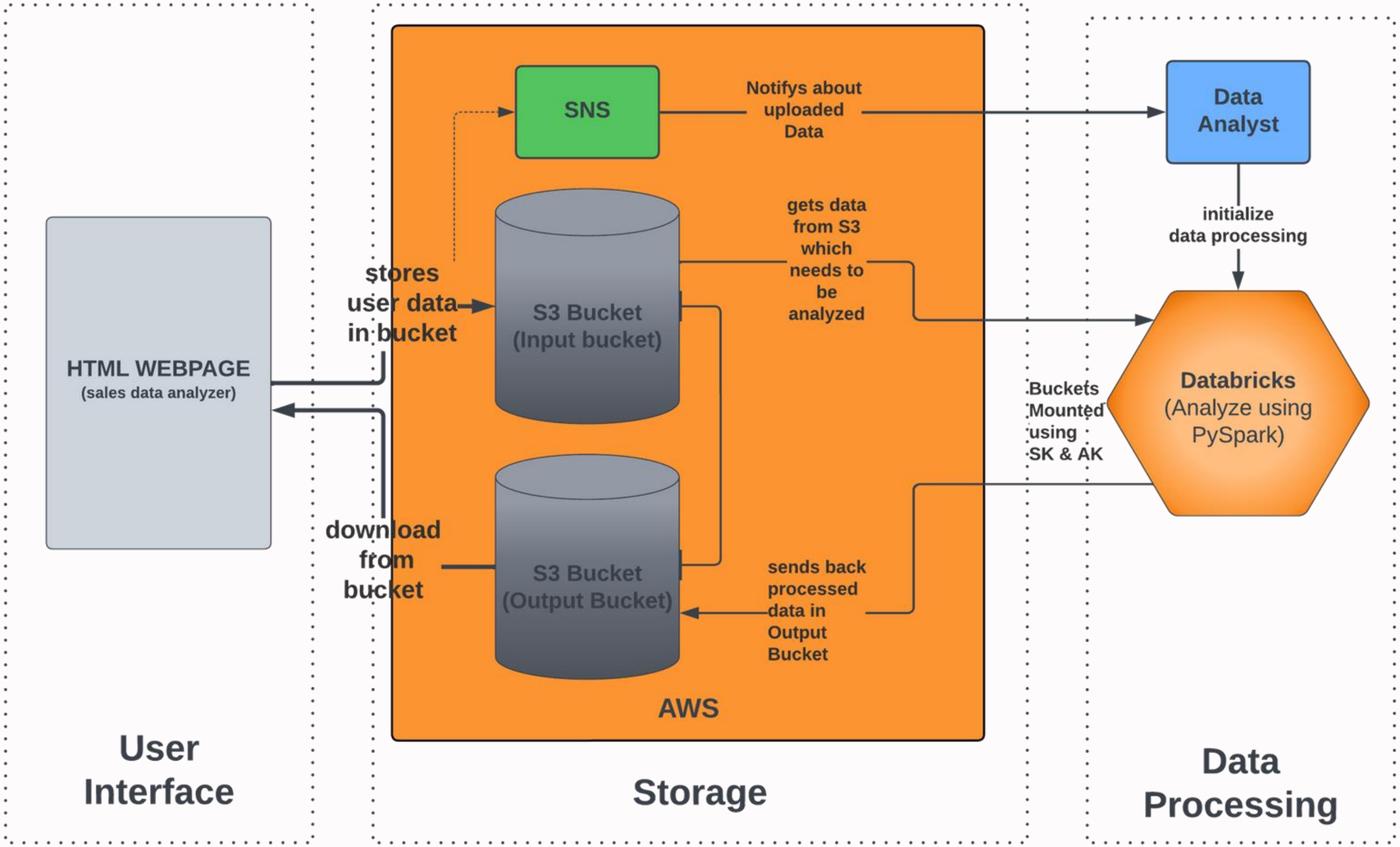
## B. Using databricks + S3 Bucket



# METHODOLOGY

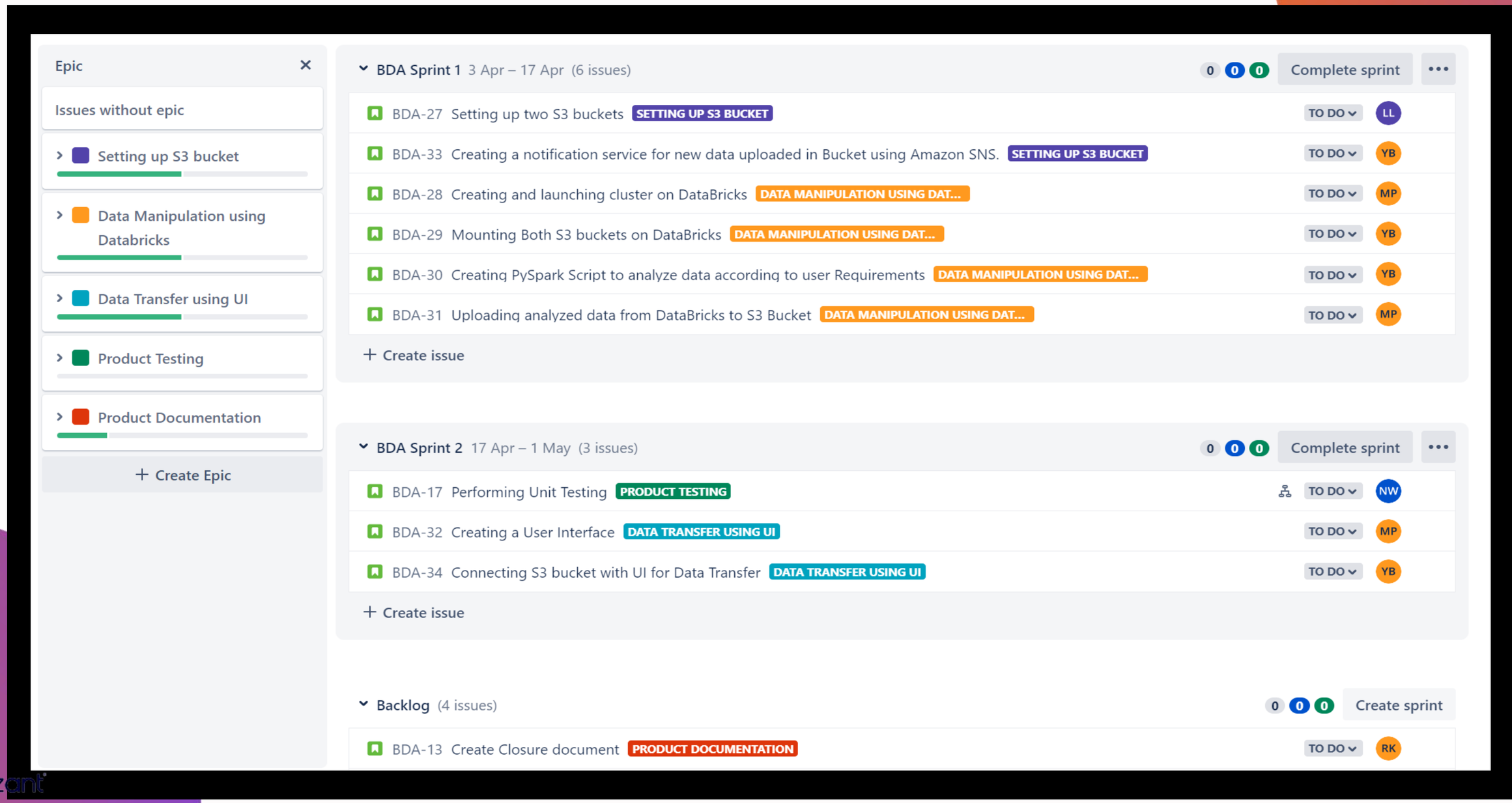


# PROJECT IMPLEMENTATION



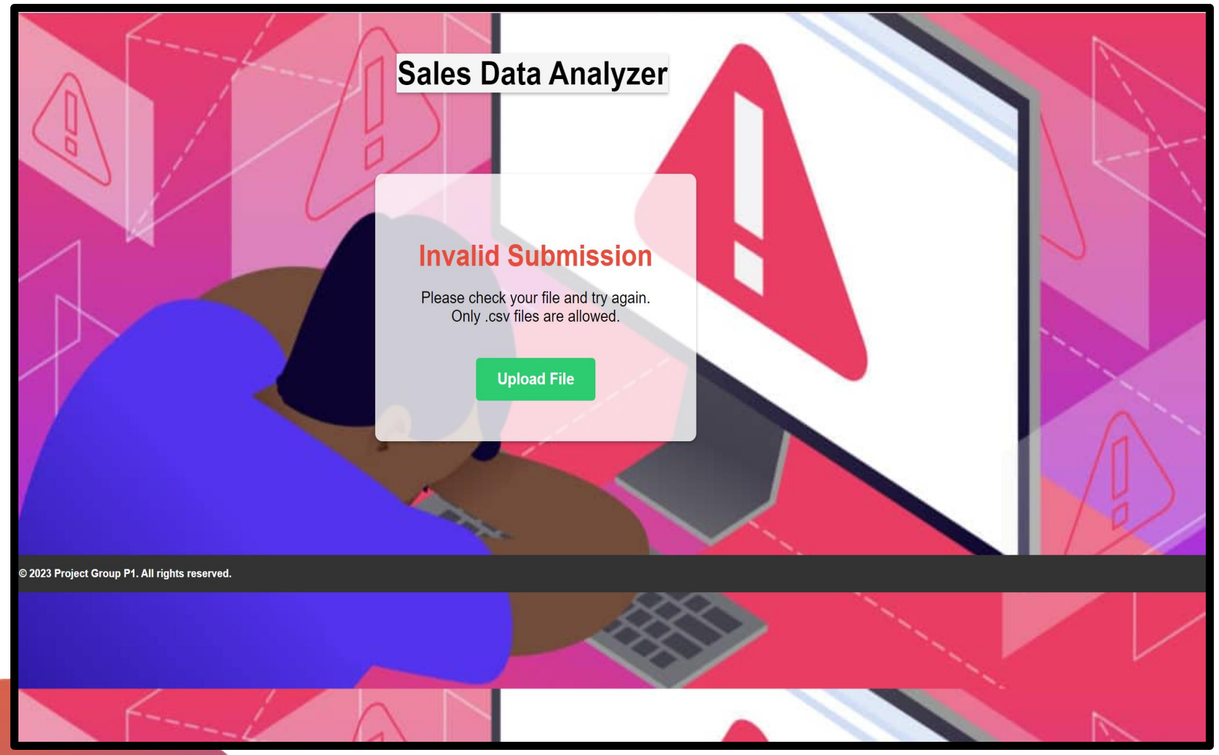
# JIRA DASHBOARD

click here →

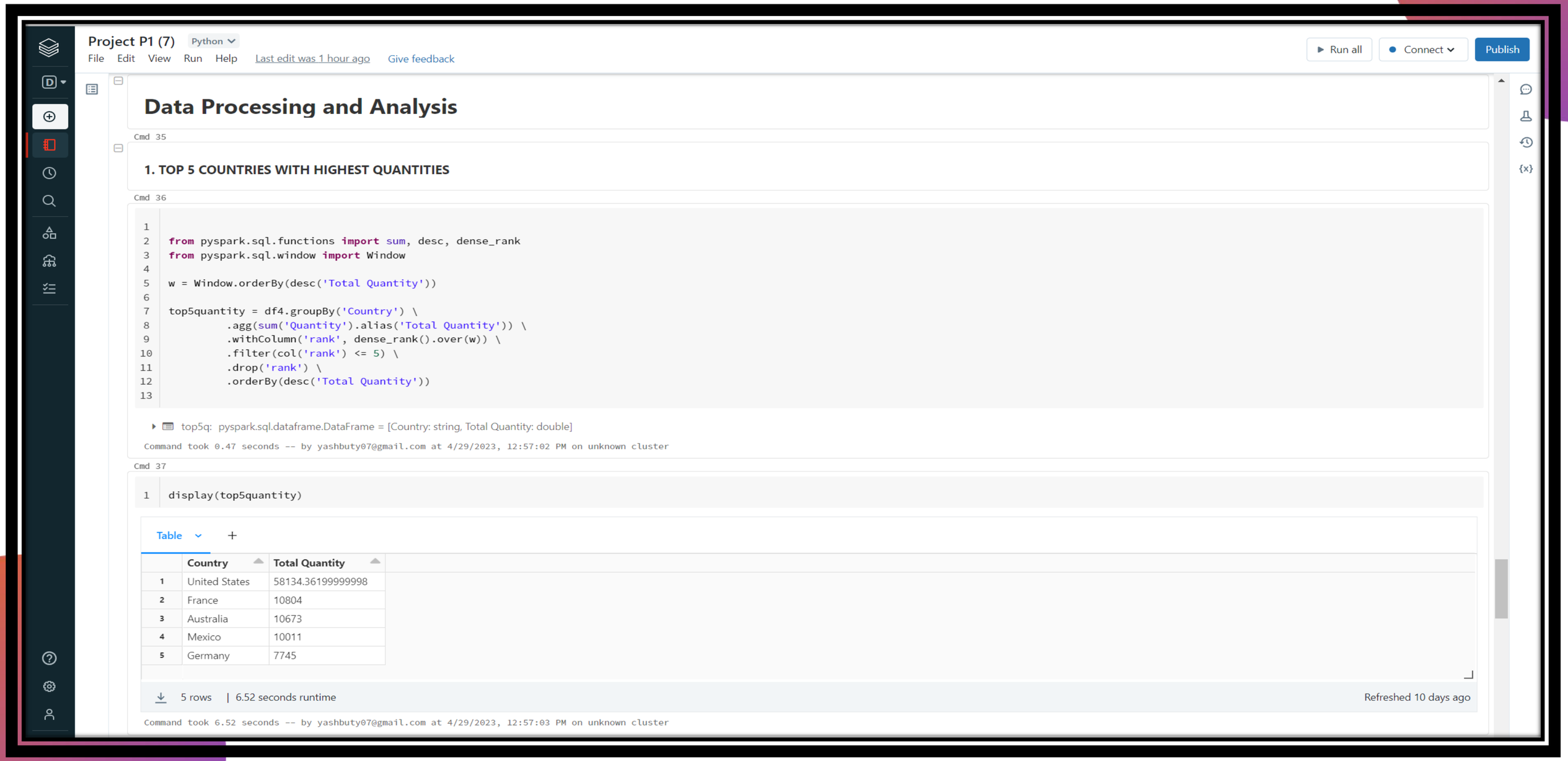



cognizant

# USER INTERFACE




# DATABRICKS


Project P1 (7)
Python
File Edit View Run Help
Last edit was 1 hour ago
Give feedback
Run all
Connect
Publish
Data Processing and Analysis
Cmd 35
1. TOP 5 COUNTRIES WITH HIGHEST QUANTITIES
Cmd 36
from pyspark.sql.functions import sum, desc, dense_rank
from pyspark.sql.window import Window
w = Window.orderBy(desc('Total Quantity'))
top5quantity = df4.groupBy('Country') \
.agg(sum('Quantity').alias('Total Quantity')) \
.withColumn('rank', dense_rank().over(w)) \
.filter(col('rank') <= 5) \
.drop('rank') \
.orderBy(desc('Total Quantity'))
top5q: pyspark.sql.dataframe.DataFrame = [Country: string, Total Quantity: double]
Command took 0.47 seconds -- by yashbuty07@gmail.com at 4/29/2023, 12:57:02 PM on unknown cluster
Cmd 37
display(top5quantity)
Table
Country
Total Quantity
1 United States 58134.36199999998
2 France 10804
3 Australia 10673
4 Mexico 10011
5 Germany 7745
5 rows | 6.52 seconds runtime
Refreshed 10 days ago
Command took 6.52 seconds -- by yashbuty07@gmail.com at 4/29/2023, 12:57:03 PM on unknown cluster

# CLIENT SALES DATA

| Row ID | Order ID | Order Date | Ship Date | Ship Mode | Customer | Customer Name | Segment | City | State | Country | Postal Co | Market | Region | Product | Category | Sub-Cate | Product | Sales | Quantity | Discount | Profit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 32298 | CA-2012-124891 | 31-07-2012 | 31-07-2012 | Same Day | RH-19495 | Rick Hansen | Consumer | New York | New York | United Sta | 10024 | US | East | TEC-AC-10 | Technolog | Accessorie | Plantronic | 2309.65 | 7 | 0 | 762.1845 |
| 26341 | IN-2013-77878 | 5/2/2013 | 7/2/2013 | Second Class | JR-16210 | Justin Ritter | Corporate | Wollongor | New Soutl | Australia | | APAC | Oceania | FUR-CH-10 | Furniture | Chairs | Novimex E | 3709.395 | 9 | 0.1 | -288.765 |
| 25330 | IN-2013-71249 | 17-10-2013 | 18-10-2013 | First Class | CR-12730 | Craig Reiter | Consumer | Brisbane | Queenslar | Australia | | APAC | Oceania | TEC-PH-10 | Technolog | Phones | Nokia Sma | 5175.171 | 9 | 0.1 | 919.971 |
| 13524 | ES-2013-1579342 | 28-01-2013 | 30-01-2013 | First Class | KM-16375 | Katherine Murray | Home Offi | Berlin | Berlin | Germany | | EU | Central | TEC-PH-10 | Technolog | Phones | Motorola S | 2892.51 | 5 | 0.1 | -96.54 |
| 47221 | SG-2013-4320 | 5/11/2013 | 6/11/2013 | Same Day | RH-9495 | Rick Hansen | Consumer | Dakar | Dakar | Senegal | | Africa | Africa | TEC-SHA-1 | Technolog | Copiers | Sharp Wire | 2832.96 | 8 | 0 | 311.52 |
| 22732 | IN-2013-42360 | 28-06-2013 | 1/7/2013 | Second Class | JM-15655 | Jim Mitchum | Corporate | Sydney | New Soutl | Australia | | APAC | Oceania | TEC-PH-10 | Technolog | Phones | Samsung S | 2862.675 | 5 | 0.1 | 763.275 |
| 30570 | IN-2011-81826 | 7/11/2011 | 9/11/2011 | First Class | TS-21340 | Toby Swindell | Consumer | Porirua | Wellingtor | New Zealand | | APAC | Oceania | FUR-CH-10 | Furniture | Chairs | Novimex E | 1822.08 | 4 | 0 | 564.84 |
| 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-10 | Furniture | Tables | Chromcraf | 5244.84 | 6 | 0 | 996.48 |
| 40155 | CA-2014-135909 | 14-10-2014 | 21-10-2014 | Standard Class | JW-15220 | Jane Waco | Corporate | Sacrament | California | United Sta | 95823 | US | West | OFF-BI-10 | Office Sup | Binders | Fellowes P | 5083.96 | 5 | 0.2 | 1906.485 |
| 40936 | CA-2012-116638 | 28-01-2012 | 31-01-2012 | Second Class | JH-15985 | Joseph Holt | Consumer | Concord | North Car | United Sta | 28027 | US | South | FUR-TA-10 | Furniture | Tables | Chromcraf | 4297.644 | 13 | 0.4 | -1862.31 |
| 34577 | CA-2011-102988 | 5/4/2011 | 9/4/2011 | Second Class | GM-14695 | Greg Maxwell | Corporate | Alexandria | Virginia | United Sta | 22304 | US | South | OFF-SU-10 | Office Sup | Supplies | Martin Yal | 4164.05 | 5 | 0 | 83.281 |
| 28879 | ID-2012-28402 | 19-04-2012 | 22-04-2012 | First Class | AJ-10780 | Anthony Jacobs | Corporate | Kabul | Kabul | Afghanistan | | APAC | Central As | FUR-TA-10 | Furniture | Tables | Bevis Conf | 4626.15 | 5 | 0 | 647.55 |
| 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-10 | Technolog | Phones | Cisco Sma | 2616.96 | 4 | 0 | 1151.4 |
| 4132 | MX-2012-130015 | 13-11-2012 | 13-11-2012 | Same Day | VF-21715 | Vicky Freymann | Home Offi | Toledo | Parana | Brazil | | LATAM | South | FUR-CH-10 | Furniture | Chairs | Harbour C | 2221.8 | 7 | 0 | 622.02 |
| 27704 | IN-2013-73951 | 6/6/2013 | 8/6/2013 | Second Class | PF-19120 | Peter Fuller | Consumer | Mudanjiar | Heilongjia | China | | APAC | North Asia | OFF-AP-10 | Office Sup | Appliance | KitchenAic | 3701.52 | 12 | 0 | 1036.08 |
| 13779 | ES-2014-5099955 | 31-07-2014 | 3/8/2014 | Second Class | BP-11185 | Ben Peterman | Corporate | Paris | Ile-de-Frar | France | | EU | Central | OFF-AP-10 | Office Sup | Appliance | Breville Re | 1869.588 | 4 | 0.1 | 186.948 |
| 36178 | CA-2014-143567 | 3/11/2014 | 6/11/2014 | Second Class | TB-21175 | Thomas Boland | Corporate | Henderson | Kentucky | United Sta | 42420 | US | South | TEC-AC-10 | Technolog | Accessorie | Logitech d | 2249.91 | 9 | 0 | 517.4793 |
| 12069 | ES-2014-1651774 | 8/9/2014 | 14-09-2014 | Standard Class | PJ-18835 | Patrick Jones | Corporate | Prato | Tuscany | Italy | | EU | South | OFF-AP-10 | Office Sup | Appliance | Hoover Sto | 7958.58 | 14 | 0 | 3979.08 |
| 22096 | IN-2014-11763 | 31-01-2014 | 1/2/2014 | First Class | JS-15685 | Jim Sink | Corporate | Townsville | Queenslar | Australia | | APAC | Oceania | TEC-CO-10 | Technolog | Copiers | Brother Fa | 2565.594 | 9 | 0.1 | 28.404 |
| 49463 | TZ-2014-8190 | 5/12/2014 | 7/12/2014 | Second Class | RH-9555 | Ritsa Hightower | Consumer | Uvinza | Kigoma | Tanzania | | Africa | Africa | OFF-KIT-10 | Office Sup | Appliance | KitchenAic | 3409.74 | 6 | 0 | 818.28 |
| 46630 | PL-2012-7820 | 8/8/2012 | 10/8/2012 | First Class | AB-600 | Ann Blume | Corporate | Bytom | Silesia | Poland | | EMEA | EMEA | FUR-HON- | Furniture | Tables | Hon Comp | 1977.72 | 4 | 0 | 276.84 |
| 31784 | CA-2011-154627 | 29-10-2011 | 31-10-2011 | First Class | SA-20830 | Sue Ann Reed | Consumer | Chicago | Illinois | United Sta | 60610 | US | Central | TEC-PH-10 | Technolog | Phones | Apple iPho | 2735.952 | 6 | 0.2 | 341.994 |
| 21586 | IN-2011-44803 | 2/5/2011 | 3/5/2011 | First Class | JK-15325 | Jason Klamczynski | Corporate | Suzhou | Anhui | China | | APAC | North Asia | FUR-CH-10 | Furniture | Chairs | SAFCO Exe | 2754 | 6 | 0 | 358.02 |
| 13528 | ES-2013-2860574 | 27-02-2013 | 1/3/2013 | Second Class | LB-16795 | Laurel Beltran | Home Offi | Edinburgh | Scotland | United Kingdom | | EU | North | OFF-AP-10 | Office Sup | Appliance | KitchenAic | 5273.7 | 10 | 0 | 1898.4 |
| 1570 | US-2014-133193 | 31-07-2014 | 1/8/2014 | First Class | NP-18325 | Naresj Patel | Consumer | Juárez | Chihuahua | Mexico | | LATAM | North | TEC-PH-10 | Technolog | Phones | Motorola S | 1713.84 | 4 | 0 | 445.52 |
| 3484 | MX-2014-165309 | 5/9/2014 | 8/9/2014 | First Class | VD-21670 | Valerie Dominguez | Consumer | Soyapango | San Salvad | El Salvador | | LATAM | Central | FUR-TA-10 | Furniture | Tables | Hon Comp | 2106.496 | 8 | 0.2 | 526.496 |
| 30191 | IN-2011-10286 | 17-12-2011 | 20-12-2011 | First Class | PB-19210 | Phillip Breyer | Corporate | Taipei | Taipei City | Taiwan | | APAC | North Asia | FUR-TA-10 | Furniture | Tables | Lesro Conf | 1715.16 | 2 | 0 | 720.36 |

# 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**

# 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

cognizant

# 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

## 3 EPIC

### Data Transfer using UI

**USER STORY:**

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

## 4 EPIC

### Product Testing & Documentation

**USER STORY:**

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