INDUSTRIAL TRAINING REPORT

Data Science Intern

AT

Incedo Technology Solutions Ltd

Submitted in partial fulfilment

of the

requirement for the award of the degree of

BACHELOR OF TECHNOLOGY (B. Tech)

in

Data Science & Engineering

Submitted by

Kanishk Mudgal

(209309047)

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

DEPARTMENT OF DATA SCIENCE AND ENGINEERING,
SCHOOL OF INFORMATION TECHNOLOGY
MANIPAL UNIVERSITY JAIPUR,
RAJASTHAN, INDIA-303007

November 2023

October 31, 2023

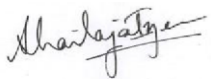
TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Kanishk Mudgal** has successfully completed his Summer Internship in **Banking** Department for period of **Aug 01, 2023,** to **October 28, 2023**.

We really appreciate **Kanishk's** efforts throughout his stay with us. He finished the training under the direction of **Mr. Praveen Kasana** and worked hard on all the tasks that were allocated to him**.**

We found him sincere, hardworking and result oriented. He worked well as part of a team during his tenure. We take this opportunity to thank him and wish him all the best for his future.

FOR INCEDO TECHNOLOGY SOLUTIONS LIMITED



AUTHORIZED SIGNATORY HUMAN RESOURCES

Incedo Technology Solutions limited
CIN: U72200DL2011PLC220071
Corporate Office : 248,Udyog Vihar, Phase -IV, Gurugram - 122015
Registered Office : A - 47, Lower Ground Floor, Hauz Khas, New Delhi - 110016
inquiries@incedoinc.com | www.incedoinc.com
Phone: +91 124 4345901

DECLARATION

I hereby declare that the Industrial Training Report on Datascience Internship is an authentic record of my own work as requirements of Industrial Training during the period from 09 Aug to 18 Oct 2023 for the award of the degree of B.Tech. (Data Science and Engineering), Manipal University Jaipur, Rajasthan under the guidance of **Mr. Praveen Kasana.**

.

(Kanishk Mudgal)
(209309047)

Date: 20-11-2023

Certified that the above statement made by the student is correct to the best of our knowledge and belief.

Examined by:
Dr. Aparna Tripathi

Head of Department:
Dr. Akhilesh Kumar Sharma

ACKNOWLEDGMENTS

First and foremost, I wish to express my sincere thanks and gratitude to my esteemed Mentor **Mr. Praveen Kasana** who has contributed so much for successful completion of my Industrial Training by his thoughtful reviews and valuable guidance. Next, I would like to sincerely thanks **Dr. Akhilesh Kumar Sharma** (Head of Data Science and Engineering Department), industrial training coordinator and faculty members whose assistance was sought during the training work for his co-operation and encouragement.

Kanishk Mudgal
209309047

ABSTRACT

During my internship at Incedo Technology Solutions Ltd, I undertook a comprehensive exploration of log files through exploratory data analysis (EDA) using Jupyter Notebook. This involved employing various statistical and visual techniques to derive meaningful insights from the data, facilitating a deeper understanding of patterns and anomalies within the log files.

Additionally, I successfully implemented S3 connectivity within Jupyter Notebook, streamlining the seamless transfer and retrieval of data between the notebook and Amazon S3. This not only enhanced data accessibility but also contributed to the efficiency of data-driven tasks.

Furthermore, I demonstrated proficiency in AWS bucket handling and Lambda API integration, enabling the company to leverage scalable and serverless computing for enhanced data processing. This involved managing and manipulating data stored in AWS buckets and seamlessly integrating Lambda functions to automate processes.

Moreover, I played a pivotal role in integrating Kibana, Logstash, and Elasticsearch, creating a robust ecosystem for log management and analysis. This integration facilitated real-time log monitoring, data parsing, and storage, enhancing the company's ability to extract valuable insights from large datasets.

In summary, my internship at Incedo Technology Solutions Ltd involved a multifaceted exploration of data science techniques, encompassing EDA, AWS connectivity, serverless computing, and advanced log management solutions. The skills acquired and projects completed during this period have significantly contributed to my expertise in data science and its practical applications.

CONTENTS

TOPICS

1: INTRODUCTION

1.1 Introduction

The advent of big data has revolutionized the way organizations extract value from their vast datasets, necessitating advanced data science techniques for efficient analysis and decision-making. During my internship at Incedo Technology Solutions Ltd, I focused on several key aspects of data science, contributing to the company's data-driven initiatives. This report provides an overview of my work, spanning exploratory data analysis (EDA), S3 connectivity, AWS bucket handling, Lambda API integration, and the integration of Kibana, Logstash, and Elasticsearch for comprehensive log management.

1.2 Scope of Work

The scope of my work during the internship at Incedo Technology Solutions Ltd encompassed a broad range of data science tasks, each contributing to the enhancement of the company's data infrastructure. The primary areas of focus included exploratory data analysis (EDA) of log files, establishing S3 connectivity within Jupyter Notebook, handling AWS buckets, integrating Lambda APIs for serverless computing, and implementing a comprehensive solution for log management using Kibana, Logstash, and Elasticsearch. The scope extended from understanding the intricacies of log data to optimizing data accessibility, processing, and real-time monitoring.

1.3 Problem Statement

The ever-increasing volume and complexity of log files pose a significant challenge for organizations, requiring effective solutions for data management and analysis. The problem addressed in this internship was the need for streamlined processes to extract meaningful insights from log data. This involved

tackling issues such as identifying patterns and anomalies within log files, improving data accessibility through cloud services like S3, optimizing data processing through AWS buckets and Lambda APIs, and implementing a robust log management system using Kibana, Logstash, and Elasticsearch. The overarching problem was to create a cohesive and efficient data science workflow that could handle and derive value from large-scale log data in a scalable and real-time manner, meeting the evolving needs of the organization.

2. Background Material

The proliferation of log files in contemporary IT environments underscores the importance of effective log management and analysis. With this backdrop, my internship aimed to address critical challenges associated with log data, ranging from understanding patterns and anomalies through EDA to optimizing data accessibility and processing through AWS services. The integration of S3 connectivity, AWS buckets, Lambda APIs, and Kibana, Logstash, and Elasticsearch reflects the diverse range of skills applied to enhance the organization's data infrastructure.

- ***ELK Stack:***

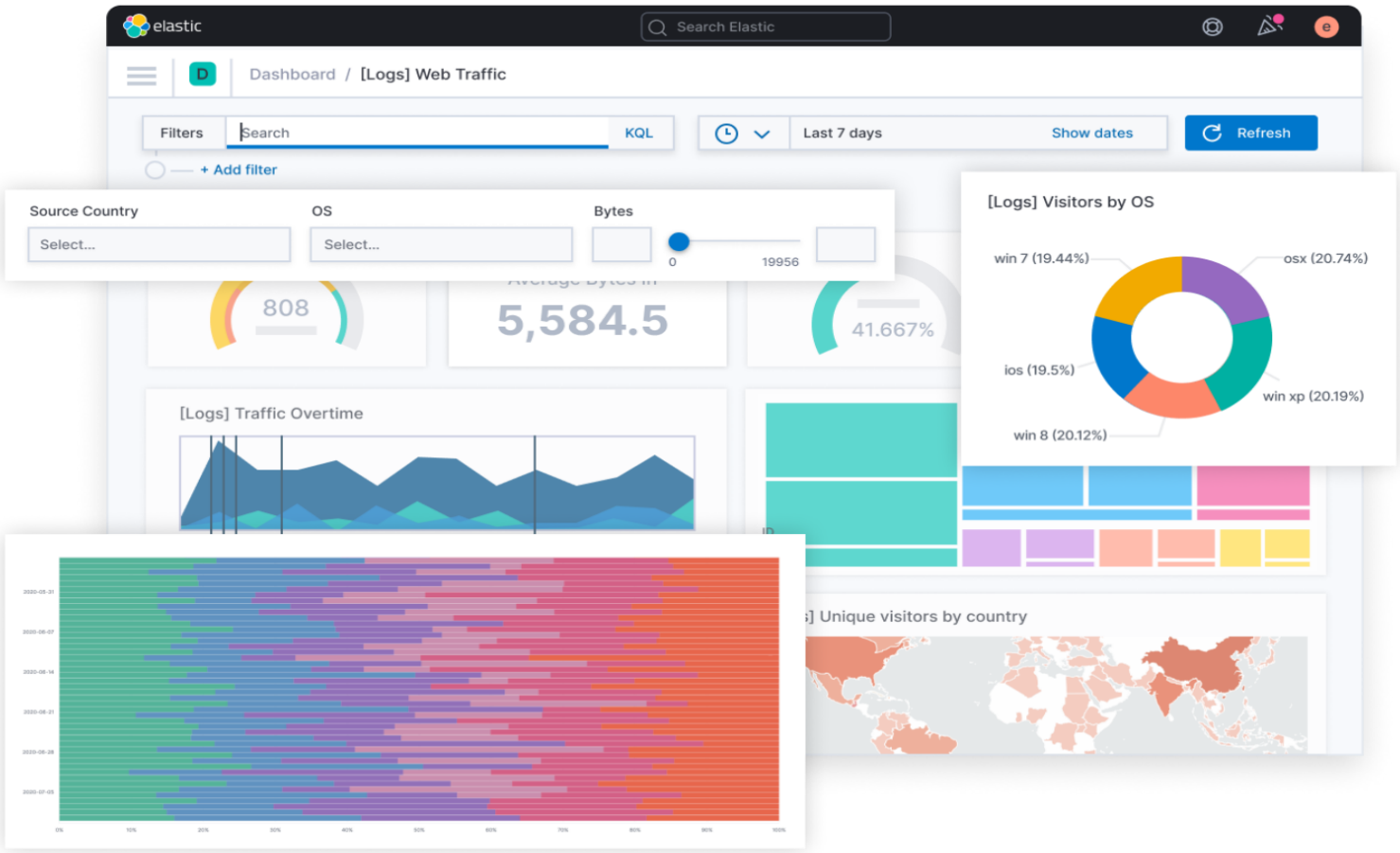
ELK Stack is a powerful and widely used open-source log and data analytics platform. The acronym "ELK" stands for Elasticsearch, Logstash, and Kibana, which are the three core components of the stack:

1. Elasticsearch: It is a distributed search and analytics engine designed for scalability and speed. Elasticsearch is used for indexing, searching, and analyzing large volumes of data in real-time. It provides a RESTful interface for querying the data.

2. Logstash: Logstash is a data processing pipeline that ingests, transforms, and enriches data from various sources. It supports a wide range of input sources, applies filters to parse and transform data, and outputs the processed data to various destinations, including Elasticsearch.

3. Kibana: Kibana is a web-based visualization tool that allows users to explore, analyse , and visualize data stored in Elasticsearch. It provides interactive dashboards, charts, and graphs, making it easier for users to understand complex data patterns and trends.

The ELK Stack is commonly used for log management, monitoring, and data analysis in various industries.



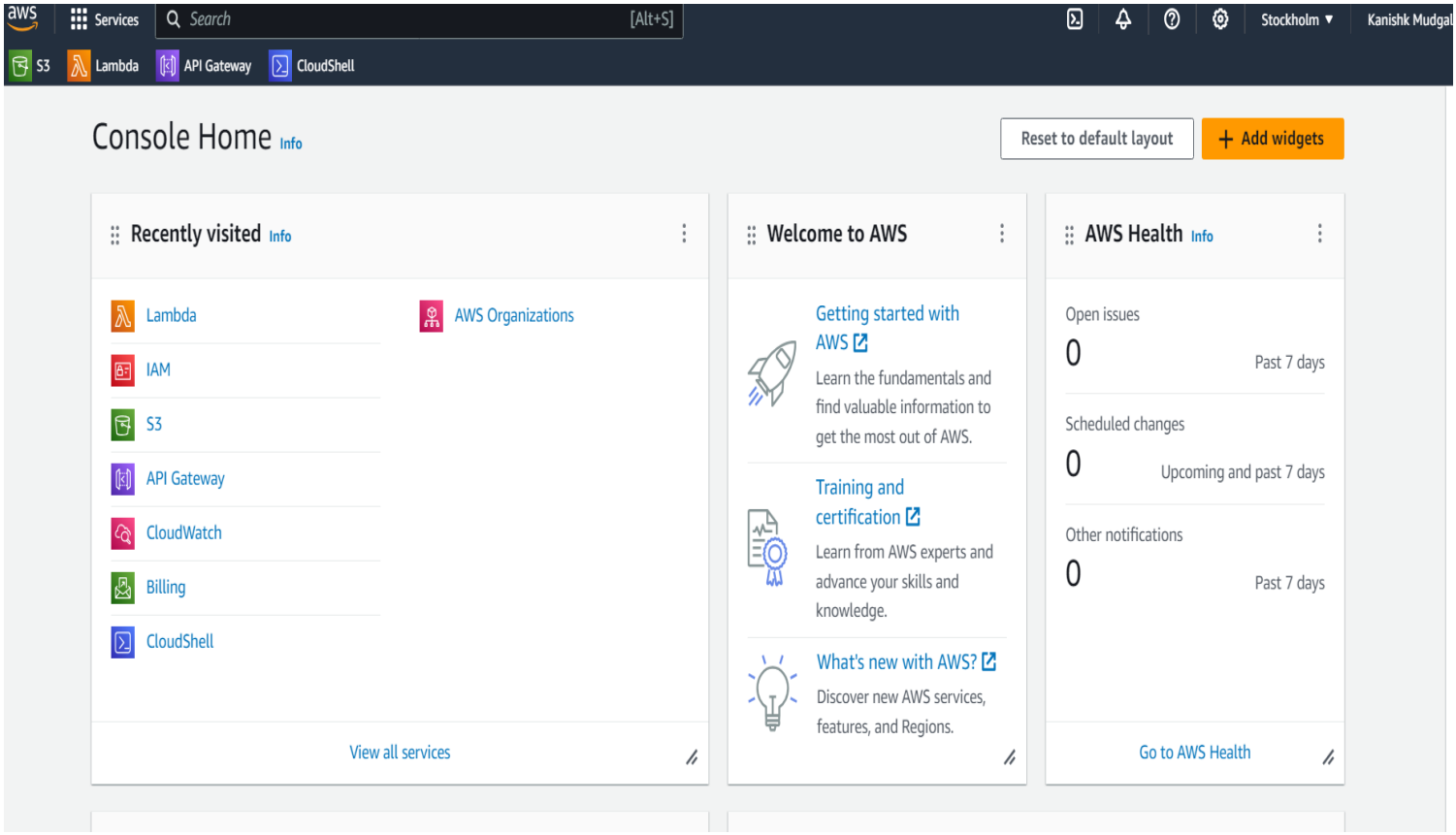
- ***Amazon Web Services (AWS):***

Amazon Web Services (AWS) is a comprehensive cloud computing platform provided by Amazon. It offers a wide range of services, including computing power, storage, databases, machine learning, analytics, and more. Some key AWS services relevant to data science and log management include:

1. Amazon S3 (Simple Storage Service):A scalable object storage service that allows users to store and retrieve any amount of data. It is often used for data storage and retrieval in data science projects.

2. AWS Lambda: A serverless computing service that lets you run code without provisioning or managing servers. It is commonly used to execute code in response to events, making it useful for automating tasks and integrating with other AWS services.

3. AWS CloudWatch: A monitoring service for AWS resources and the applications running on the platform. It provides real-time monitoring and actionable insights into AWS resources.



- ***Exploratory Data Analysis (EDA):***

Exploratory Data Analysis is a critical step in the data analysis process where analysts or data scientists visually and statistically explore datasets to understand their characteristics. Key components of EDA include:

1. Descriptive Statistics: Summarizing and describing the main features of a dataset, such as mean, median, standard deviation, and percentiles.

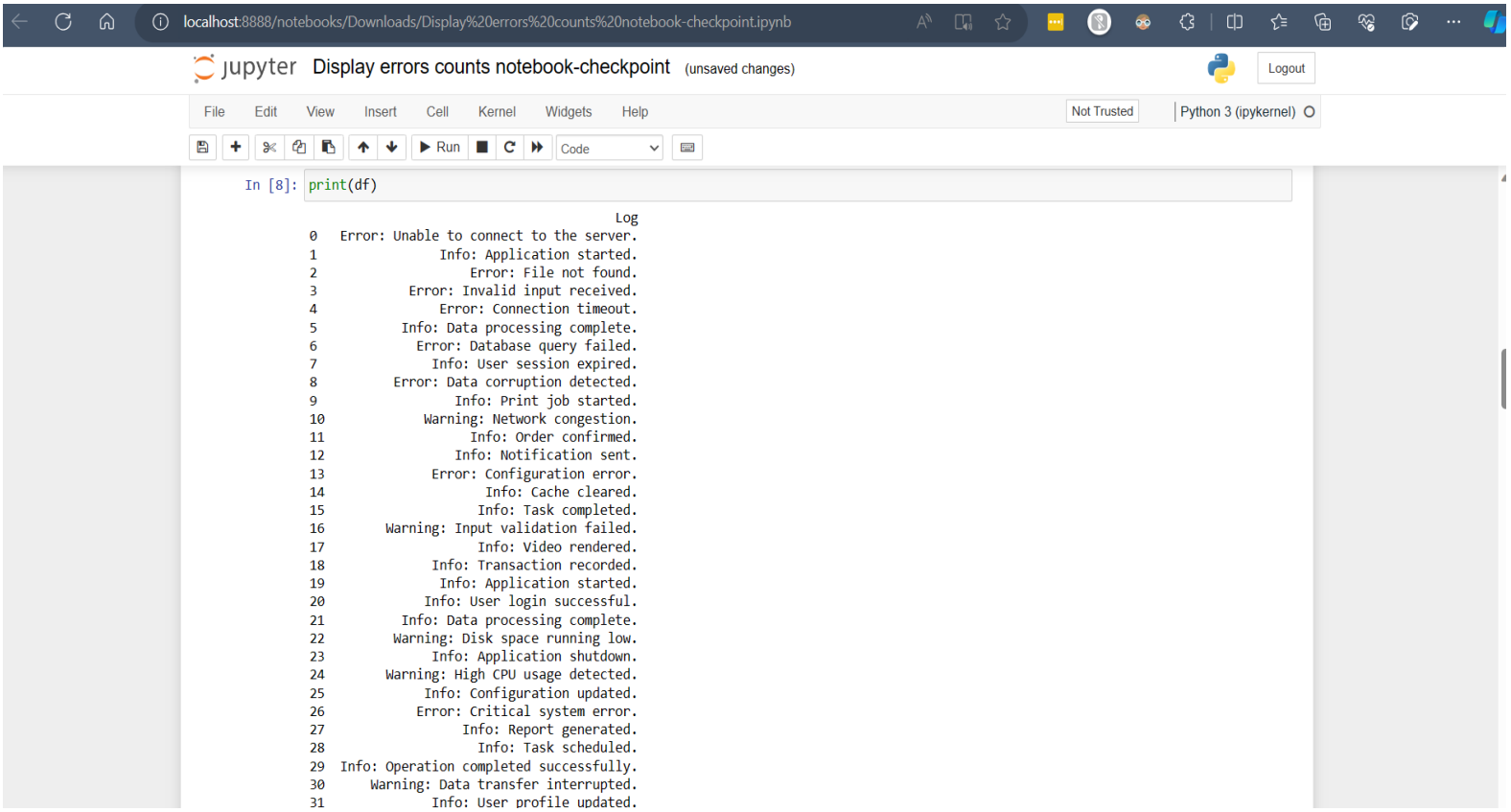
2. Data Visualization: Creating visual representations of the data, such as histograms, scatter plots, and box plots, to identify patterns, trends, and outliers.

3. Data Cleaning: Identifying and handling missing or inconsistent data to ensure the quality of the dataset.

4. Feature Engineering: Creating new features or transforming existing ones to better represent the underlying patterns in the data.

5. Correlation Analysis: Examining the relationships between variables to understand dependencies and potential insights.

EDA is crucial for gaining insights into the structure and nature of the data, informing subsequent steps in the data analysis pipeline, and guiding the development of models and hypotheses.



3. Methodology

1. Project Initiation:

- Define the scope and objectives of the project.
- Identify key stakeholders and gather initial requirements.
- Establish a clear understanding of the problem statement and project goals.

2. Literature Review:

- Review existing literature, frameworks, and methodologies relevant to the project.
- Identify best practices and insights from previous work in the field.

3. Data Collection:

- Identify and collect relevant data sources for analysis.
- Ensure data quality and address any issues related to missing or inconsistent data.
- Document the data sources, variables, and metadata.

4. Exploratory Data Analysis (EDA):

- Use statistical and visual techniques to explore and understand the characteristics of the data.
- Generate summary statistics, visualizations, and insights to inform further analysis.
- Identify patterns, trends, and potential outliers in the data.

5. Method Selection:

- Choose appropriate methods and algorithms based on the nature of the problem and the goals of the project.
- Consider factors such as scalability, interpretability, and computational efficiency.

6. Model Development:

- Develop and implement models based on the selected methods.
- Fine-tune model parameters and hyperparameters for optimal performance.
- Validate models using appropriate evaluation metrics.

7. AWS Services Integration:

- Integrate relevant AWS services such as S3, Lambda, or CloudWatch as needed.

- Configure connections and permissions to ensure seamless data flow and processing.

8. ELK Stack Integration:

- Set up and configure Elasticsearch, Logstash, and Kibana according to project requirements.
- Establish data pipelines for log management and real-time analysis.

9. Testing and Validation:

- Conduct thorough testing of the implemented solutions.
- Validate results against expected outcomes and business requirements.
- Address any issues or discrepancies through iterative improvements.

10. Documentation:

- Document the entire methodology, including data sources, preprocessing steps, analysis techniques, and results.
- Create user guides or documentation for future reference.

11. Communication and Reporting:

- Regularly communicate progress, challenges, and findings to stakeholders.
- Generate reports or presentations summarizing key insights and recommendations.

12. Iterative Refinement:

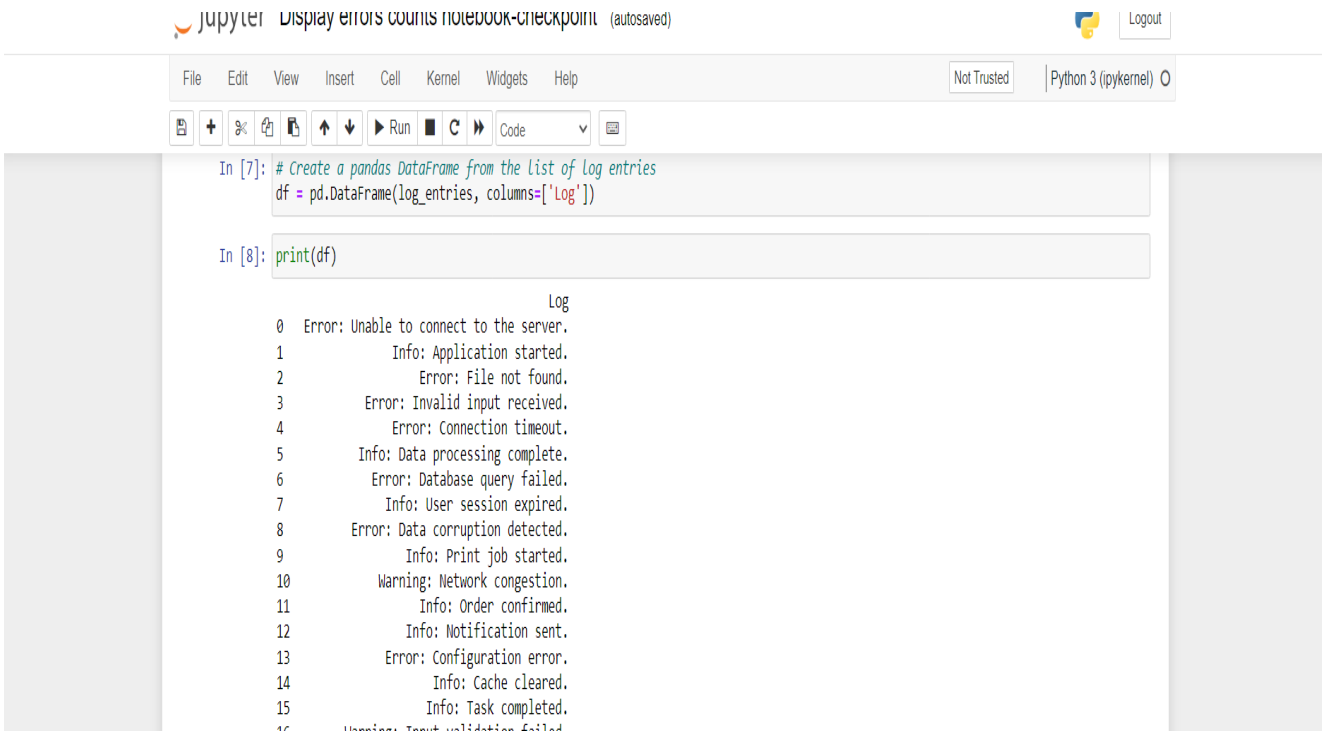
- Gather feedback from stakeholders and end-users.
- Iterate on the methodology based on feedback and evolving project requirements.
- Continuously refine and improve the approach throughout the project lifecycle.

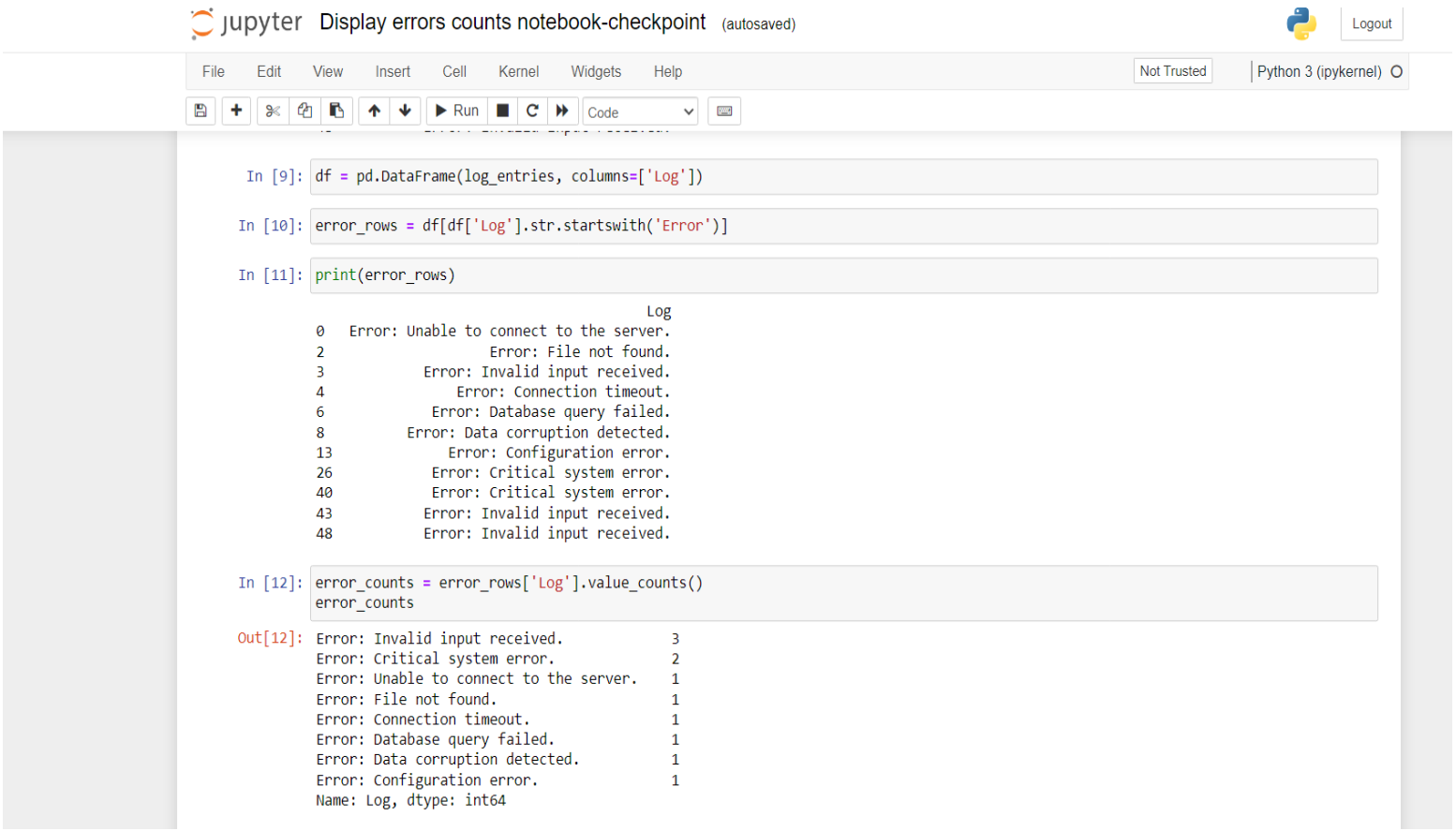
This methodology ensures a systematic and structured approach to problem-solving, integrating various components such as data analysis, model development, and technology integration in a cohesive manner.

4. Implementation

1. *Displaying errors using Python*

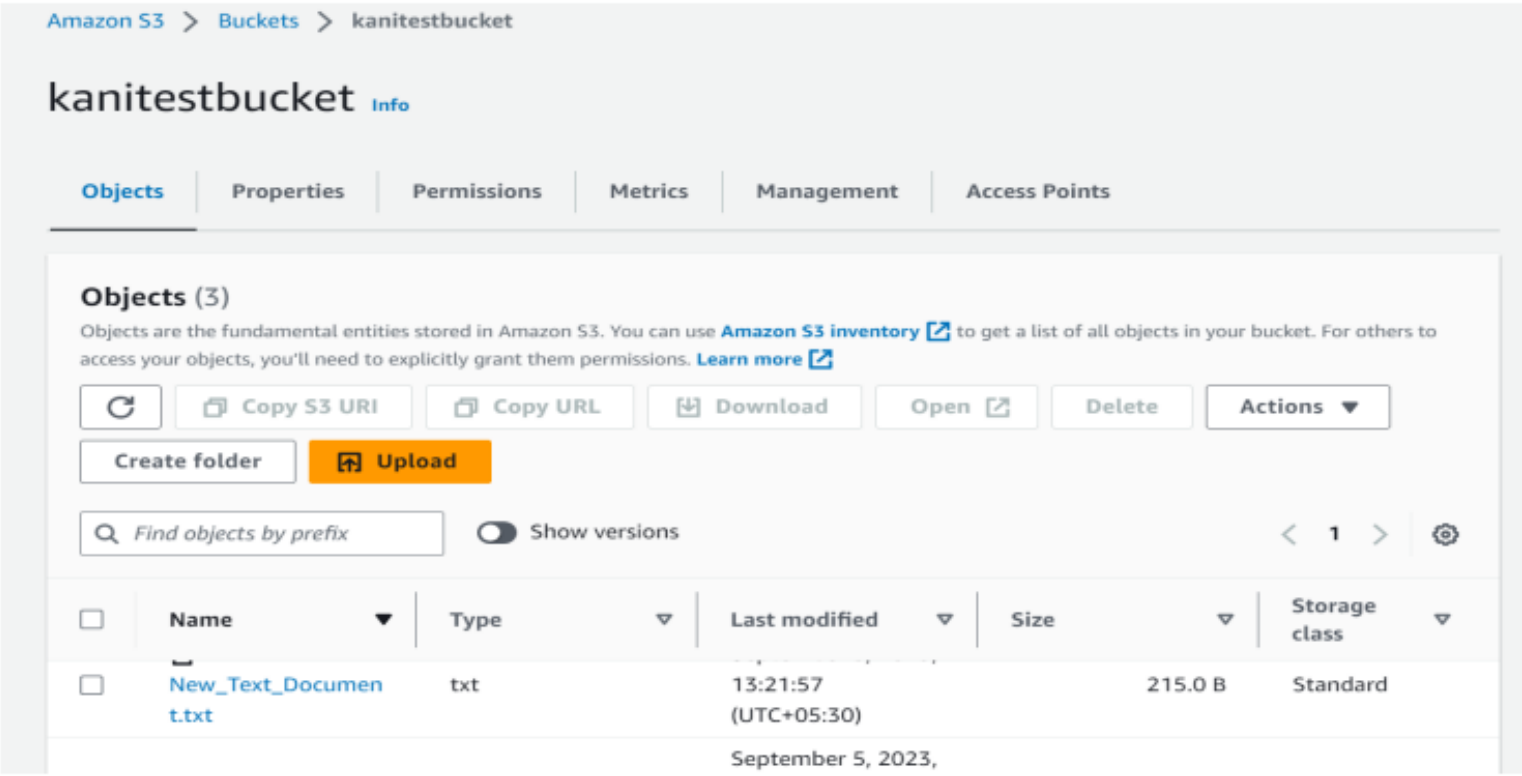
```
In [2]: import pandas as pd

In [3]: import mpld3

In [4]: import matplotlib.pyplot as plt

In [5]: import plotly.express as px

In [6]: log_entries = [
            "Error: Unable to connect to the server.",
            "Info: Application started.",
            "Error: File not found.",
            "Error: Invalid input received.",
            "Error: Connection timeout.",
            "Info: Data processing complete.",
            "Error: Database query failed.",
            "Info: User session expired.",
            "Error: Data corruption detected.",
            "Info: Print job started.",
            "Warning: Network congestion.",
            "Info: Order confirmed.",
            "Info: Notification sent.",
```



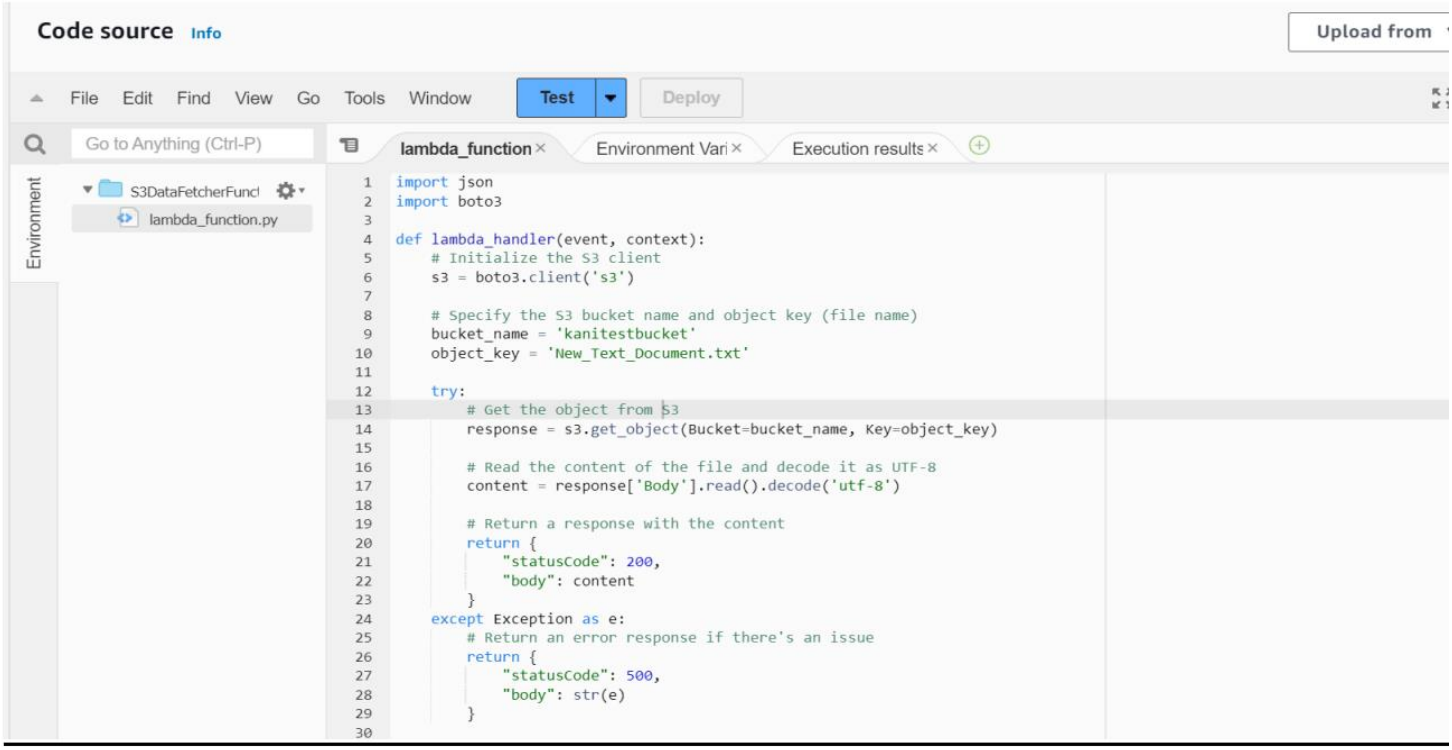


2. *S3 Bucket, Lambda Function , API gateway and jupyter notebook integration.*

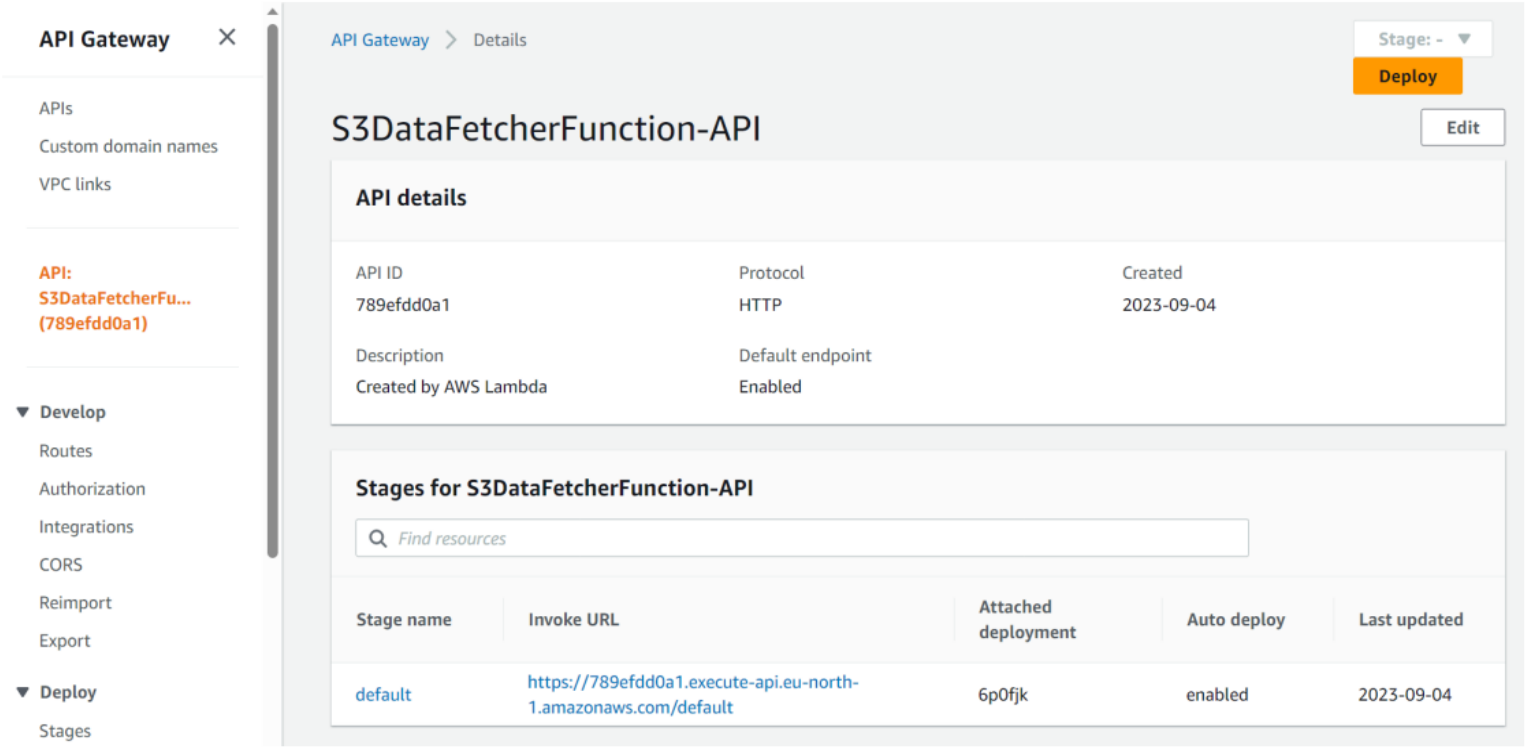
1. S3 Bucket and object creation:



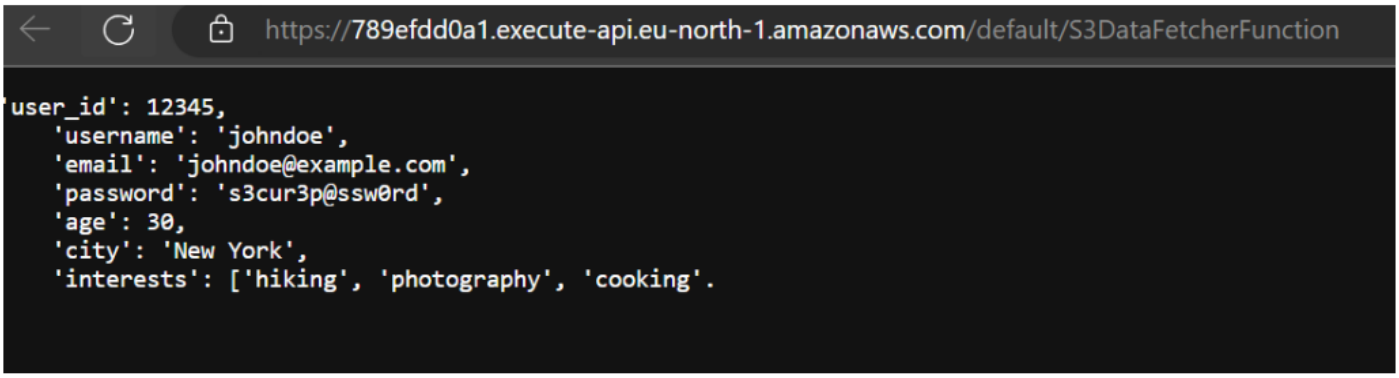
2. S3 data fetch Lambda function:



3. Added Trigger to lambda function – API Gateway



4. API direct link response when data feeded through Jupyter in S3 bucket object.



5. Feeding Data & calling Lamba API model in Jupyter notebook

KANISHK MUDGAL PROGRESS REPORT ON S3 , LAMBDA , API GATEWAY , JUPYTER INTEGRATION

In [58]:
```
import boto3
import requests
import json
import s3fs
service_name='s3',
```

In [ ]:
```
#----------------------------------DEALING WITH S3 BUCKET-----------------------------------------------#
```

In [136]:
```
# Configure AWS credentials
boto3.setup_default_session(
    aws_access_key_id='AKIA4NYFOBBAQ6EVBDV5',
    aws_secret_access_key='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',
    region_name='eu-north-1'
)
```

In [141]:
```
#----Giving input in bucket----#

# Specifying the S3 bucket name and object key for the text file
s3_bucket_name = 'kanitestbucket'
s3_key = 'New_Text_Document.txt'  # Choose a key name for your text file

# Define input data as a string
input_data = """'user_id': 12345,
    'username': 'johndoe',
    'email': 'johndoe@example.com',
    'password': 's3cur3p@ssw0rd',
    'age': 30,
    'city': 'New York',
    'interests': ['hiking', 'photography', 'cooking'.
"""

# Save the input data to a local text file
with open('input_data.txt', 'w') as file:
    file.write(input_data)

# Upload the local text file to S3
s3.upload_file('input_data.txt', s3_bucket_name, s3_key)

print(f"Input data has been saved as {s3_key} in the S3 bucket: {s3_bucket_name}")
```

```
Input data has been saved as New_Text_Document.txt in the S3 bucket: kanitestbucket
```

In [142]:
```
#----Print contents of txt file------#

# Specify the S3 bucket name and file key
s3_bucket_name = 'your-s3-bucket-name'
file_key = 'example.txt'  # Replace with the actual key of your text file

try:
    # Get the text file from S3
    response = s3.get_object(Bucket='kanitestbucket', Key='New_Text_Document.txt')

    # Read the content of the text file
    text_content = response['Body'].read().decode('utf-8')

    # Print the content
    print("Content of the text file:")
    print(text_content)

except Exception as e:
    print("Error:", str(e))
```

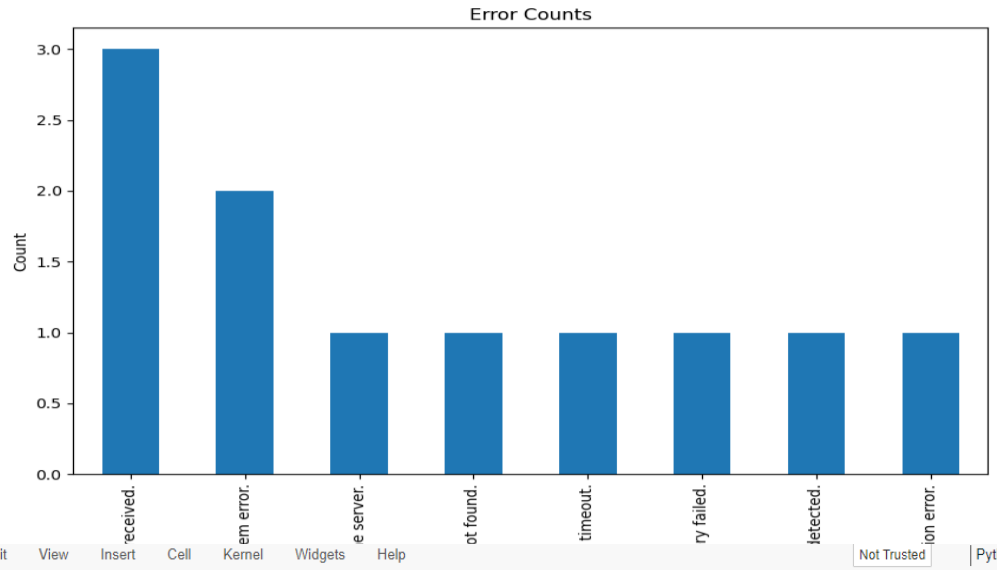
```
Content of the text file:
'user_id': 12345,
    'username': 'johndoe',
    'email': 'johndoe@example.com',
    'password': 's3cur3p@ssw0rd',
    'age': 30,
    'city': 'New York',
    'interests': ['hiking', 'photography', 'cooking'.
```

5. Results and Analysis

1.Results by applying EDA:

```
error_counts.plot(kind='bar')
plt.title('Error Counts')
plt.xlabel('Error Messages')
plt.ylabel('Count')
```

```
Out[13]: Text(0, 0.5, 'Count')
```

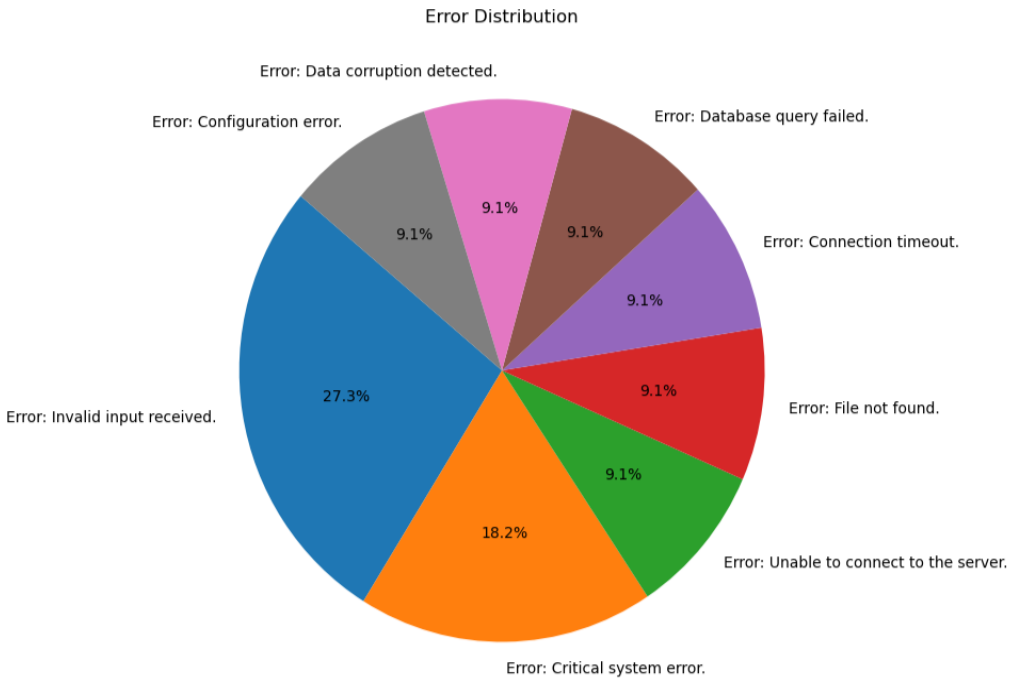


File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
In [27]: #Error display in pie chart format
plt.figure(figsize=(8, 8))
plt.pie(error_counts, labels=error_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Error Distribution')
```

```
Out[27]: Text(0.5, 1.0, 'Error Distribution')
```



2. Results by calling AWS API on Jupyter

In [127]:
```
#-------------------------------------------------------------------------------------------------------------------------
```

In [128]:
```
#----------------------------DEALING WITH LAMBDA function through API HERE--------------------#
```

In [149]:
```
# Set up the S3 and API Gateway clients
s3 = boto3.client('s3')
api_url = 'https://789efdd0a1.execute-api.eu-north-1.amazonaws.com/default/S3DataFetcherFunction'
```

In [151]:
```
import requests

# API Gateway URL
api_url = 'https://789efdd0a1.execute-api.eu-north-1.amazonaws.com/default/S3DataFetcherFunction'

try:
    response = requests.get(api_url)

    if response.status_code == 200:
        # Parse and print the response from the API Gateway
        api_response = response.text
        print("API Response:")
        print(api_response)
    else:
        print("Error:", response.status_code)
        print("Error Message:", response.text)

except Exception as e:
    print("Error:", str(e))
```

```
API Response:
'user_id': 12345,
    'username': 'johndoe',
    'email': 'johndoe@example.com',
    'password': 's3cur3p@ssw0rd',
    'age': 30,
    'city': 'New York',
    'interests': ['hiking', 'photography', 'cooking'.
```

In [ ]:
```
#-------------------------------------------------END-----------------------------------------------------------------
```

6. Conclusion:

6.1 Conclusion

The culmination of my work at Incedo Technology Solutions Ltd has significantly contributed to the advancement of the company's data science capabilities. The completion of exploratory data analysis (EDA) on log files using Jupyter Notebook provided valuable insights into data patterns and anomalies, laying the foundation for informed decision-making. Establishing S3 connectivity and handling AWS buckets enhanced data accessibility, streamlining the flow of information between Jupyter Notebook and Amazon S3.

The integration of Lambda APIs for serverless computing has not only optimized data processing but has also automated critical tasks, improving overall efficiency. Furthermore, the successful integration of Kibana, Logstash, and Elasticsearch has created a robust log management system, enabling real-time monitoring and analysis.

This work has not only showcased the application of data science techniques but has also demonstrated the practical implementation of cloud services and log management solutions, fostering a more data-driven and streamlined organizational environment.

6.2 Future Scope:

The completion of these projects opens avenues for future enhancements and expansions in several directions:

1. Advanced Analytics: Further exploration of advanced analytics techniques, machine learning models, and predictive analytics can be incorporated to derive deeper insights and facilitate proactive decision-making.

2. Scale and Performance Optimization: As data volumes grow, there is a scope for optimizing the current solutions to handle larger datasets and improve overall performance. This may involve revisiting configurations, adopting parallel processing, or exploring alternative cloud services.

3. Security and Compliance: Enhancements can be made to incorporate robust security measures and ensure compliance with industry regulations. This includes data encryption, access controls, and adherence to data governance policies.

4. User Interface Improvements: The development of user-friendly interfaces or dashboards within Kibana can enhance the accessibility of insights to a broader audience within the organization. This can empower stakeholders to interact with and derive value from the data.

5. Continuous Monitoring and Maintenance: Establishing a framework for continuous monitoring and maintenance of the implemented solutions will ensure their long-term sustainability. This includes monitoring data quality, updating models, and addressing any emerging issues.

6. Integration with Additional AWS Services: Exploring and integrating other AWS services relevant to the company's evolving needs can provide additional functionalities and benefits. For example, incorporating AWS Step Functions for orchestrating workflows or Amazon QuickSight for business intelligence.

In essence, the current work serves as a solid foundation for ongoing innovation and improvement, positioning the organization to adapt and thrive in the dynamic landscape of data science and cloud computing.