PROJECT CLOSURE DOCUMENT

Project Group P1

1. Likhitesh A L
2. Sai Rohith
3. Mandar Pimparkar
4. Yash Buty
5. Neha Waikar
6. Yash Gharde

User Document

Table of Content

URL

The "URL" section in the project closure document outlines the URL or web address that the end-users can use to access the application. This section should provide the details of the URL.

User Manual

We have provided a step-by-step guide to help users get started with the application. Below are instructions on how to use its different features.

This section will give you the minimum system requirements required to access the webpage.

Minimum System requirements

- Windows 8 or above
- Microsoft Edge browser or Google Chrome or Mozilla Firefox
- Memory (RAM): at least 2Gb available

Features:

1. How to sign in to the User interface
2. How to upload a file in User interface
3. How to download a file in User interface

We hope this user manual has provided you with all the information you need to get started and make the most of our application.

Trouble Shooting:

This troubleshooting guide is designed to help you identify and resolve issues that may arise when uploading a CSV file to process and analyze it using a DataBricks cluster.

1. **User Interface Issues:**

If you encounter issues with the user interface where you upload the CSV file, try the following:

- Ensure that the file format is CSV.
- Verify that the file size is within the maximum limit set by the interface.
- Make sure that the file is not corrupted or damaged.
- Clear your browser cache and try again.
- Try using a different browser or device to access the user interface.

2. **DataBricks Cluster Issues:**

If you encounter issues with the Databricks cluster, try the following:

- Check that your cluster is running and has not been terminated.

- Verify that the cluster configuration is correct, including the Spark version and number of nodes.
- Check that the required libraries and packages are installed on the cluster.
- Ensure that the PySpark script is properly uploaded and imported into the notebook.
- Verify that the PySpark script is configured correctly to read the CSV file uploaded to the user interface.

3. **Processing and Analysis Issues:**

If you encounter issues with the processing and analysis of the CSV file, try the following:

- Verify that the PySpark script is correctly written to handle the CSV file format.
- Check that the necessary transformations and analysis functions are included in the script.
- Verify that the output file is saved to the correct location and in the correct format.
- Check that the resulting file is not empty or contains incorrect data.
- Monitor the cluster resources and logs for any errors or warnings during the processing and analysis.
- If none of these troubleshooting steps resolve your issue, contact the support team or the redressal contact for further assistance. Be sure to provide detailed information about the issue and steps taken so far to help expedite the resolution.

Redressal Contact:

In case of any issues or complaints, users and clients can reach out to the following contacts:

Support Team

Name	Email	Phone Number	Department
Yash Buty	yashbuty123@gmail.com	xxxxxxxxxx	Technical Support Team
Mandar Pimparkar	manadr98@gmail.com	xxxxxxxxxx	Maintenance Team
Likhitesh A L	likhitesh@gmail.com	xxxxxxxxxxx	Customer Service

Users and clients can also raise their concerns through the company's support ticket system or contact form available on the website. The support team will respond to the issues and complaints within the defined SLA (Service Level Agreement).

Client End Document

Table of Content

Flow of Project (SOP)

Development Methodologies

The development methodology for this project will be Agile. Agile methodology emphasises frequent communication and collaboration between the development team and the stakeholders, which helps to ensure that the final product meets the requirements and expectations of the client.

The key principles of Agile methodology that will be followed in this project are:

1. **Iterative development:** The project will be broken down into smaller, manageable iterations or sprints. Each sprint will have a specific set of deliverables that must be completed within a set time frame.
2. **Continuous feedback:** Feedback will be solicited from the client and other stakeholders throughout the development process to ensure that the project is on track and meeting expectations.
3. **Prioritizations:** Prioritizations of tasks will be done in collaboration with the client to ensure that the most critical features and functionalities are completed first.
4. **Flexibility:** The Agile methodology allows for changes to be made throughout the development process, which means that the project can be adapted to changing requirements and circumstances.
5. **Collaboration:** Frequent communication and collaboration between the development team and the stakeholders will be a key component of this methodology.

Overall, the Agile methodology will be used to ensure that the project is completed on time, within budget, and meets the expectations of the client.

Tools and Technologies:

1. **AWS S3 Bucket:** for storing the customer product details
2. **Data bricks:** for analysis the data
3. **AWS SNS:** get the notification

4. **Python & HTML:** for creating the front end UI

These technologies were chosen for their ability to efficiently and effectively handle the requirements of the project. The use of cloud-based services also allowed for easy scalability and maintenance of the system.

Roles and Responsibilities

Our Team:

Name	Role	Responsibilty
K V Sai Rohith	Scrum Master	1: Facilitate Scrum Events 2: Coach and Mentor the Team 3: Remove Impediments 4: Protect the Team
Likhitesh A L	Product Owner	1: Define and Prioritize Product Backlog 2: Collaborate with the Team 3: Maximize the Value of the Product 4: Ensure the Product Vision 5: Make Decisions
Yash Gharde	Developer	1: Writing and testing code 2: Collaborating with the team 3: Continuous improvement
Neha Waikar	Developer	1: Writing and testing code 2: Collaborating with the team 3: Continuous improvement

Yash Buty	UI Designer	1: Designing the UI 2: Collaborate with the development team 3: Continuously improve the user experience
Mandar Pimparkar	UI Designer	1: Designing the UI 2: Collaborate with the development team 3: Continuously improve the user experience

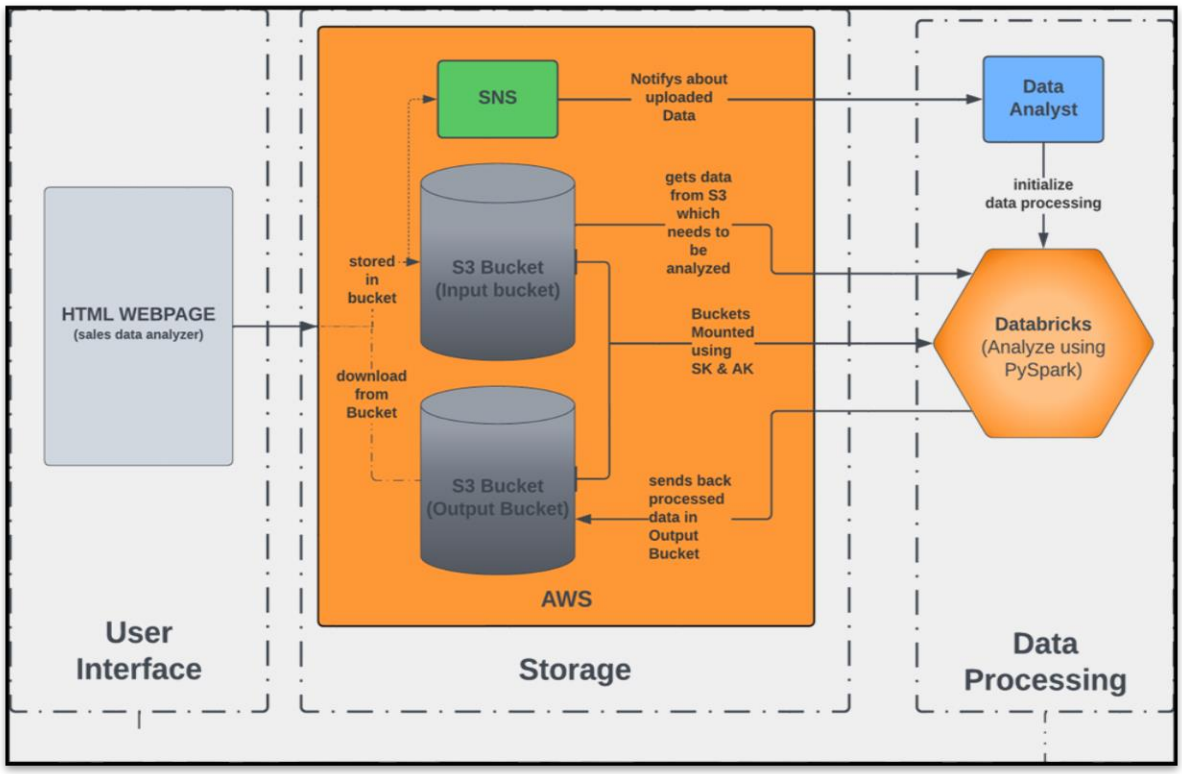
Resources (Code Documentation):

Overview

This project involves setting up a Databricks cluster to process and analyze large datasets using big data frameworks like Spark. The aim is to provide clear instructions on how to launch a sample cluster using Spark and run a simple PySpark script that will be stored in an Amazon S3 bucket. The instructions cover the essential tasks in three main workflow categories: Plan and Configure, Manage, and Clean Up. This allows the company to focus on data analysis and insights rather than spending hours setting up the infrastructure for data processing.

The project also involves creating a user interface where users can upload a CSV file containing raw data. The file is automatically processed in the Databricks cluster, and the resulting file is sent back to the user interface where the user can download and view the output.

Architecture and Design



1. **User Interface:** The user interface is built using Python and HTML and allows users to interact with the system by adding the data.
2. **AWS S3:** S3 is used to store files uploaded by the user in the cloud.
3. **AWS SNS:** SNS is used for notification when user upload the data.
4. **Databricks:** It is used for the analysis the data.

Key Members

Development Team

Name	Email	Phone Number
Neha Waikar	waikarneha21@gmail.com	xxxxxxxxxx
Yash Buty	yash.buty07@gmail.com	xxxxxxxxxx
Mandar Pimparkar	mandarmpimparkar@gmail.com	xxxxxxxxxx
Likhitesh A L	likhilikhitesh@gmail.com	xxxxxxxxxx
K V Sai Rohith	kvsrohith23@gmail.com	xxxxxxxxxx
Yash Gharde		xxxxxxxxxx

Support Team Document

Table of Content

Trouble Shoot

The Big Data analysis is a web-based project that allows users to add or upload the data through an UI. Although the system is designed to work seamlessly, some issues may arise while using it. This troubleshooting document will help you identify and resolve any issues that you may encounter.

The following are common issues that may arise while using the application and their solutions.

1. **Connectivity Issues**

- Check if the user has the necessary permissions to access the Databricks workspace, S3 bucket, and other related resources.

- Verify that the VPC and subnets are configured correctly for communication with the Databricks cluster.
- Check the network connectivity and firewall settings to ensure there are no blocking issues.

2. **Databricks Cluster Issues**

- Check the status of the Databricks cluster to ensure it is running and all nodes are functioning correctly.
- Verify that the necessary libraries and packages are installed on the cluster.
- Check the resource allocation and usage to ensure the cluster has enough resources to handle the workload.

3. **PySpark Script Issues**

- Verify that the PySpark script is uploaded to the correct location in the S3 bucket and that the user has the necessary permissions to access it.
- Check the syntax and code of the PySpark script for any errors or issues.
- Ensure that the PySpark script is compatible with the Spark version and configuration used in the Databricks cluster.

4. **User Interface Issues**

- Verify that the user has the necessary permissions to access the user interface and upload the CSV file for processing.

- Check the user interface for any errors or issues that may be preventing the user from completing the task.
- Review the logs and error messages generated by the user interface to identify the source of the problem.

5. **Other Issues**

- Review any other issues reported by the user and attempt to replicate the problem.
- Check the system and application logs for any errors or issues that may be impacting the Databricks cluster or related resources.
- Escalate the issue to the relevant team or individual if unable to resolve it.
- It is important to communicate clearly and effectively with the user throughout the troubleshooting process, providing regular updates and guidance to help them resolve the issue. It is also important to document the issue and the steps taken to resolve it for future reference.

Common Problems

Here are some common problems that users may encounter while using the Movie Distribution Centre:

1. **Lack of user training:** Users who are not familiar with the system may face difficulties navigating through the interface or using specific features.

2. **Data Compatibility**: The data uploaded by the user may not be compatible with the processing tools used in the Databricks cluster. This could cause errors or incorrect output.
3. **Connectivity Issues**: Connectivity issues between the Databricks cluster and the S3 bucket could cause delays or prevent the data from being processed.
4. **Resource Allocation**: Insufficient resources allocated to the Databricks cluster could cause performance issues or even failure of the data processing.
5. **AWS configuration issues:** Users may encounter issues with the AWS configuration, such as insufficient permissions or incorrect settings.
6. **UI usability issues:** Users may find the UI confusing or difficult to use, leading to frustration and errors.
7. **Data update and deletion issues:** Users may encounter issues when updating or deleting movie data, leading to inconsistencies in the data.
8. **Integration issues:** Users may face challenges integrating the system with other applications or platforms, such as payment gateways or social media channels.
9. **Performance issues:** As the system deals with a large amount of data, users may experience performance issues such as slow response times or system crashes.

Maintenance

Maintenance on the Movie Distribution Centre involves regular updates, fixes, and improvements to ensure the system is running efficiently and effectively.

Here are some examples of maintenance activities that can be performed on the product:

1. **Regular backups:** It is important to regularly back up the system's data to ensure that data loss does not occur in case of any unforeseen events. Backups can be done daily, weekly or monthly depending on the system usage.
2. **Software updates:** Regular software updates are essential to keep the system secure and bug-free. This includes updating the operating system, web server, and database software.
3. **Security updates**: As the system contains sensitive data, regular security updates are necessary to protect against data breaches and cyber attacks.
4. **User support:** Providing user support is essential to ensure the system is being used effectively. This includes providing user documentation, training, and troubleshooting support to users who encounter issues.
5. **Disaster recovery planning:** Disaster recovery planning is important to ensure that the system can be restored quickly in case of any disasters such as system crashes or natural disasters.

6. **User feedback analysis:** Analyzing user feedback is important to identify areas of improvement and to understand user needs and preferences.

Hardware and Software Required

Here are some hardware and software requirements for the this application:

Hardware Requirement:

1. Processor: Intel Core i5 or higher
2. RAM: 8 GB or higher
3. Storage: 500 GB or higher
4. Network Interface: Gigabit Ethernet
5. Display: 1920 x 1080 resolution or higher

Software Requirement:

1. Operating System: Windows 10 or higher, or Ubuntu 20.04 or higher
2. Programming Language: Pyspark
3. Framework: Python 3.9, HTML
4. Cloud Services: AWS S3 for storing user data.

Outcomes:

Implementing the security considerations and recommendations will help to ensure the following outcomes:

- Improve security
- Compliance
- Reliability
- Scalability

Overall, implementing these security considerations and recommendations can help you build a secure, reliable, and scalable solution that meets your requirements and protects your data.

Recommendations for Future:

To maintain the security of the system, the following recommendations are made:

1. Regular security audits: Conduct regular security audits to identify new risks and vulnerabilities.
2. On-going security training: Provide ongoing security training for employees to ensure that they are aware of security risks and how to avoid them.
3. Keep software up-to-date: Ensure that all software used in the system is up-to-date to prevent vulnerabilities caused by outdated software.

Decommission Of Product/Function

Table of Content

Basic SOP:

When it is time to decommission the Databricks cluster and associated resources, the following steps should be taken:

1. **Data Backup**

- Ensure that all important data is backed up before decommissioning the cluster and associated resources.
- Store the backup data in a secure location that is easily accessible for future reference.

2. **Termination of Databricks Cluster**

- Terminate the Databricks cluster to release the associated resources and stop incurring additional costs.
- Verify that the cluster has been successfully terminated and that no associated resources are running.

3. **S3 Bucket Cleanup**

- Remove any data and files stored in the S3 bucket that are no longer needed or relevant.
- Verify that the S3 bucket is empty and that no additional costs will be incurred for storage.

4. **Security Cleanup**

- Ensure that all security-related resources are properly cleaned up, including IAM roles, security groups, and VPC configurations.
- Verify that no security-related resources are left running and that no additional costs will be incurred.

5. **Documentation**

- Update the project documentation to reflect the decommissioning of the Databricks cluster and associated resources.
- Store the updated documentation in a secure location for future reference.

Security:

Security is a crucial aspect of this project, as it involves the processing and storage of sensitive data. The following are some security measures that should be implemented:

1. **Review Access Control:** Review access controls to ensure that only authorized personnel have access to the product or function that is being decommissioned. This includes reviewing access logs and disabling or revoking access for anyone who no longer needs it.

2. **Use Encryption:** All data should be encrypted during transfer and at rest. This can be done using SSL/TLS encryption for data in transit and using encryption tools like AWS S3 server-side encryption or client-side encryption for data at rest.
3. **Implement Access Controls:** Access to the Databricks cluster and the S3 bucket should be restricted to authorized personnel only. This can be done by implementing proper authentication and authorization mechanisms like AWS IAM policies or Azure Active Directory.
4. **Back Up Data:** Back up all data associated with the product or function being decommissioned. Ensure that the data is securely stored and that it can be restored if needed.
5. **Disable the product or function:** Disable the product or function being decommissioned to prevent unauthorized access or use. This may include revoking any licenses or disabling any software or hardware associated with the product or function.
6. **Monitor for security threats:** Monitor for security threats during the decommissioning process. This includes monitoring for any attempts to access the product or function being decommissioned, as well as monitoring for any unusual network traffic or activity.
7. **Document of the decommissioning process:** Document the decommissioning process, including any steps taken to ensure security. This Includes documenting any access controls that were reviewed or changed, any data that was backed up, and any steps taken to disable the product or function.

8. **Monitoring:** The project should be monitored to detect and prevent security breaches. This includes monitoring access logs, network traffic, and other relevant security indicators.
9. **Communication:** Stakeholders, including users and management, would be informed of the decommissioning process and the security measures being taken to ensure that sensitive data is protected. This process would be done through mail.

   By considering these security-based aspects, organizations would ensure that decommissioning the software product is carried out in a secure and controlled manner.

Responsibilities of our support team during decommissioning of a software product:

1. **Communicate with stakeholders:** The support team would communicate with all stakeholders, including users, management, and other relevant teams, to inform them of the decommissioning process and answer any questions or concerns they may have.

2. **Plan the decommissioning process:** The support team would work with other teams to plan the decommissioning process, including identifying the systems and applications that need to be decommissioned, determining the timeline, and identifying any potential risks or challenges.

3. **Provide technical support:** The support team should provide technical support to other teams during the decommissioning process, including assisting with the backup and retention of data, removal of the software from systems, and ensuring that all configurations and settings are properly removed.
4. **Document the process:** The support team should maintain accurate documentation of the decommissioning process, including any steps taken, decisions made, and any issues or challenges encountered. This documentation can be used to evaluate the process after it is complete and for future reference.
5. **Test the decommissioning:** The support team should work with other teams to test the decommissioning process to ensure that it is effective and that all data and systems are properly secured.
6. **Provide post-decommissioning support:** The support team should continue to provide support to other teams after the decommissioning process is complete, including answering questions or addressing any issues that may arise. By taking on these responsibilities, the support team can ensure that the decommissioning process is carried out smoothly, with minimal disruption to the organization, and that all sensitive data and systems are properly secured.

Project Closure Document (Project Group P1)