SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

INT300
YEAR 2(TERM-3)- FINAL REVIEW
VENDOR RECOMMENDATION PROJECT

Project Guide: Mr. Sivasankar
(L&T TC-3 DATA ANALYTICS TL)
PRESENTED BY: RISHITHA THOKA(E0322026)

Problem Statement/Objective

- The vendor recommendation project aims to address the challenges of manual vendor selection processes by leveraging PySpark for data processing and analysis. The objective is to develop an efficient system that automates vendor evaluation, shortlisting, and recommendation, while also creating a user-friendly web interface for stakeholders to access and review recommended vendor details easily.

INTRODUCTION

Our vendor recommendation project leverages the power of PySpark to streamline the vendor selection process. By utilizing PySpark's data processing capabilities, we aim to enhance efficiency and accuracy in identifying vendors who meet our specific requirements. The project also includes a web interface for displaying the recommended vendor details, ensuring transparency and ease of access for stakeholders.

Literature Survey

YEAR	PRODUCT	AUTHOR/DEVELOPER	TITLE	REMARKS
2015	Paper	Chih-Hsuan Wang	Using quality function deployment to conduct vendor assessment and supplier recommendation for business-intelligence systems	This paper presents an integrated framework to help business planners conduct vendor assessment, supplier selection and product (software) recommendation.
2023	Paper	Jane Cheng, Peng Zhao	Sustainable Big Data Analytics Process Pipeline Using Apache Ecosystem	The article explores cutting-edge big data workflow technologies like Hadoop, Spark, Airflow, etc., proposing an industrial data workflow pipeline tailored for large-scale processing in data-driven industrial analytics applications. It addresses challenges in real-world big data analytics, offering insights for researchers and professionals while fostering interdisciplinary studies in the field.
2022	Paper	Yasmine Sabri, Guido J. L. Micheli, Enrico Cagno	Supplier selection and supply chain configuration in the projects environment	To facilitate an inclusive evaluation of a wide range of capabilities of the suppliers, this research examines the application of a broader perspective for supplier selection in the projects environment.

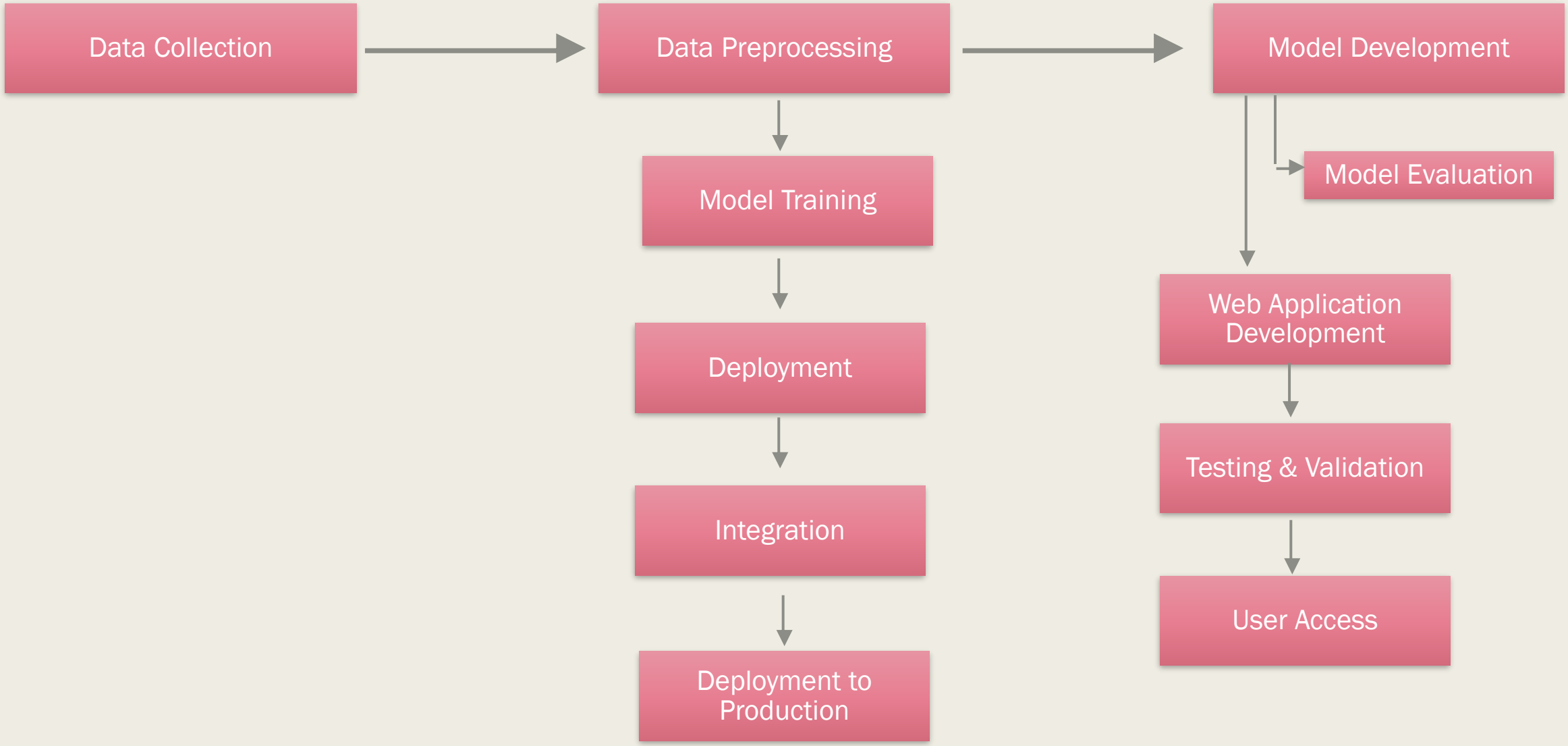
Literature Survey

YEAR	PRODUCT	AUTHOR/DEVELOPER	TITLE	REMARKS
2021	Paper	Sarah S. Alrumiah, Mohammed Hadwan	Implementing Big Data Analytics in E-Commerce: Vendor and Customer View	The paper delves into how Big Data Analytics (BDA) in e-commerce boosts decision-making, understanding consumer behavior, and revenue for vendors, despite challenges like shopping addiction and high costs, necessitating efficient data management strategies
2022	Paper	Ziyuan Xia, Anchen Sun, Jingyi Xu, Yuanzhe Peng, Rui Ma, Minghui Cheng	Contemporary Recommendation Systems on Big Data and Their Applications: A Survey	The survey paper provides a comprehensive analysis of recommendation systems' evolution, categorizing methodologies into content-based, collaborative filtering, knowledge-based, and hybrid approaches. It highlights challenges like data sparsity and scalability while emphasizing real-life applications and the potential for significant enhancements in user experiences through big data-driven advancements.

Research/Product survey

Our project involves conducting comprehensive research and surveys to identify the key criteria for vendor selection across various industries. Additionally, we will explore existing vendor recommendation systems and tools to understand best practices and potential areas for improvement. Through this research, we aim to design an effective vendor recommendation algorithm tailored to our specific needs, leveraging PySpark's capabilities, and create a user-friendly web interface for displaying recommended vendor details.

Workflow


Data Collection
Data Preprocessing
Model Development
Model Training
Model Evaluation
Deployment
Web Application Development
Integration
Testing & Validation
Deployment to Production
User Access

Methodology

- **Data Collection:**
  - Obtain data from various sources such as databases, APIs, or CSV files.
  - Store the data in a format that can be easily accessed by PySpark, such as Parquet or CSV.
- **Data Preprocessing:**
  - Clean the data by handling missing values, outliers, and formatting issues.
  - Perform data transformation tasks such as feature engineering, normalization, or encoding categorical variables.
- **Model Development:**
  - Use PySpark to develop machine learning models for vendor recommendation. Common models include collaborative filtering, content-based filtering, or hybrid approaches.
  - Split the data into training and testing sets for model evaluation.
- **Model Training:**
  - Train the recommendation model using the training dataset.
- **Model Evaluation:**
  - Evaluate the trained model using metrics such as accuracy.
  - Use the testing dataset to assess how well the model generalizes to new data.
- **Deployment:**
  - Deploy the trained model using PySpark's deployment capabilities, such as Apache Spark.
  - Expose the model as an API endpoint that can receive input data and provide recommendations.

Methodology

- **Web Application Development:**
  - Develop a web application using frameworks like Flask or streamline.
  - Create a user interface where users can input their preferences or requirements for vendor recommendations.
- **Integration:**
  - Integrate the deployed model API into the web application backend.
  - Handle requests from the frontend, send them to the model API, and display the recommendations on the webpage.
- **Testing and Validation:**
  - Test the end-to-end workflow to ensure all components are functioning correctly.
  - Validate the recommendations provided by the web application against expected results.
- **Deployment to Production:**
  - Deploy the web application to a production environment where it can be accessed by users.
  - Monitor the application for performance, scalability, and any potential issues.

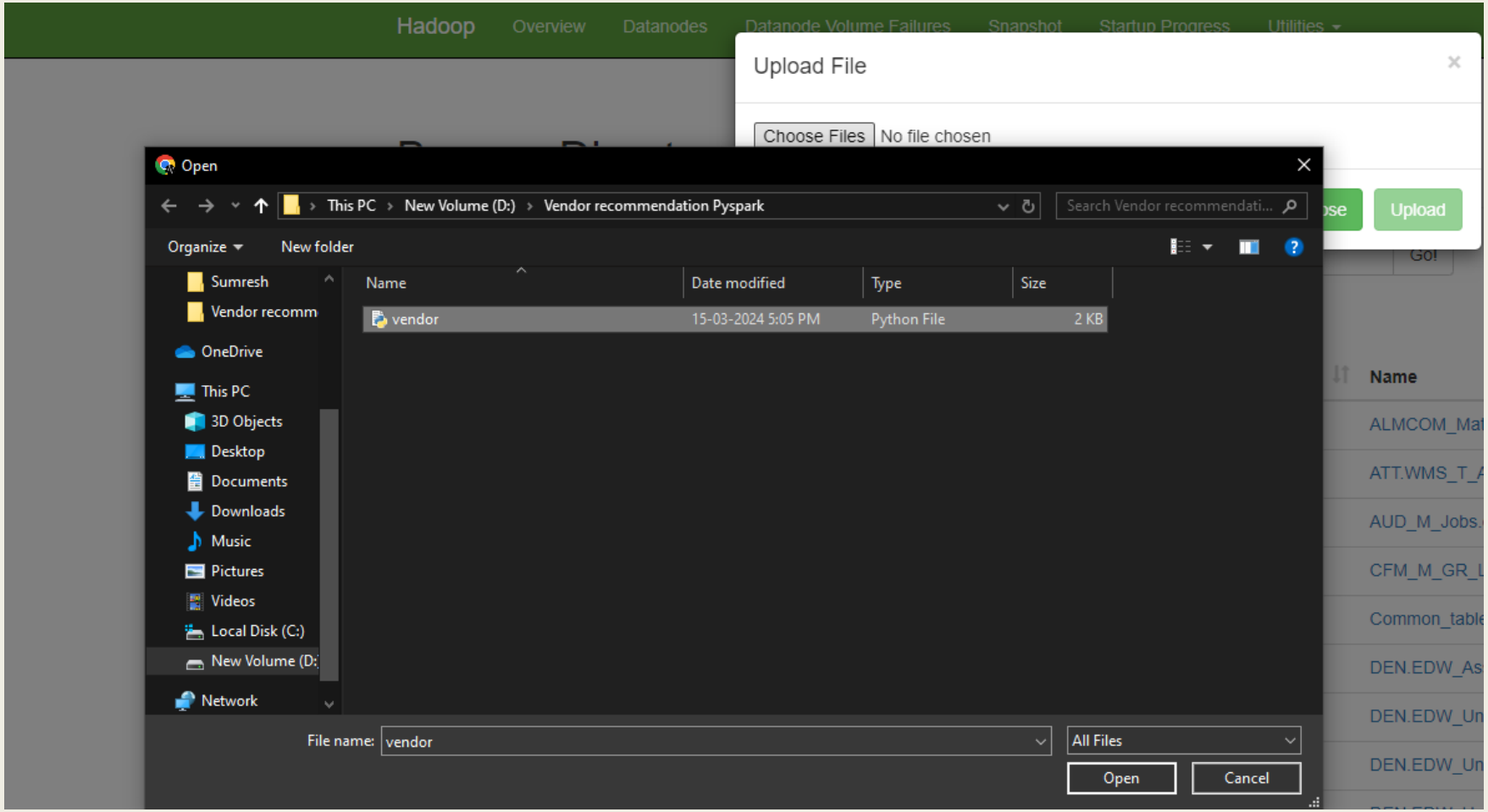
System Requirements

- **PySpark and Its Libraries:**
  - **PySpark:** PySpark is the core technology used for distributed data processing and machine learning tasks. It allows you to leverage Apache Spark's capabilities for handling large-scale datasets and running computations in parallel.
- **HDFS (Hadoop Distributed File System):**
  - HDFS is used for distributed storage of large volumes of data. In the context of a PySpark project, HDFS can be used to store input data, intermediate results, and output data generated during the data processing and modeling stages.
- **Apache Airflow:**
  - Apache Airflow is a workflow management platform that allows you to schedule, monitor, and manage complex workflows or pipelines. It can be used to orchestrate the different stages of your vendor recommendation project, such as data preprocessing, model training, and deployment.
- **Visual Studio:**
  - Visual Studio (assuming you mean Visual Studio Code or Visual Studio IDE) can be used as an integrated development environment (IDE) for writing and debugging code, particularly Python code for PySpark tasks. It provides features like code highlighting, debugging tools, and extensions for Python development.

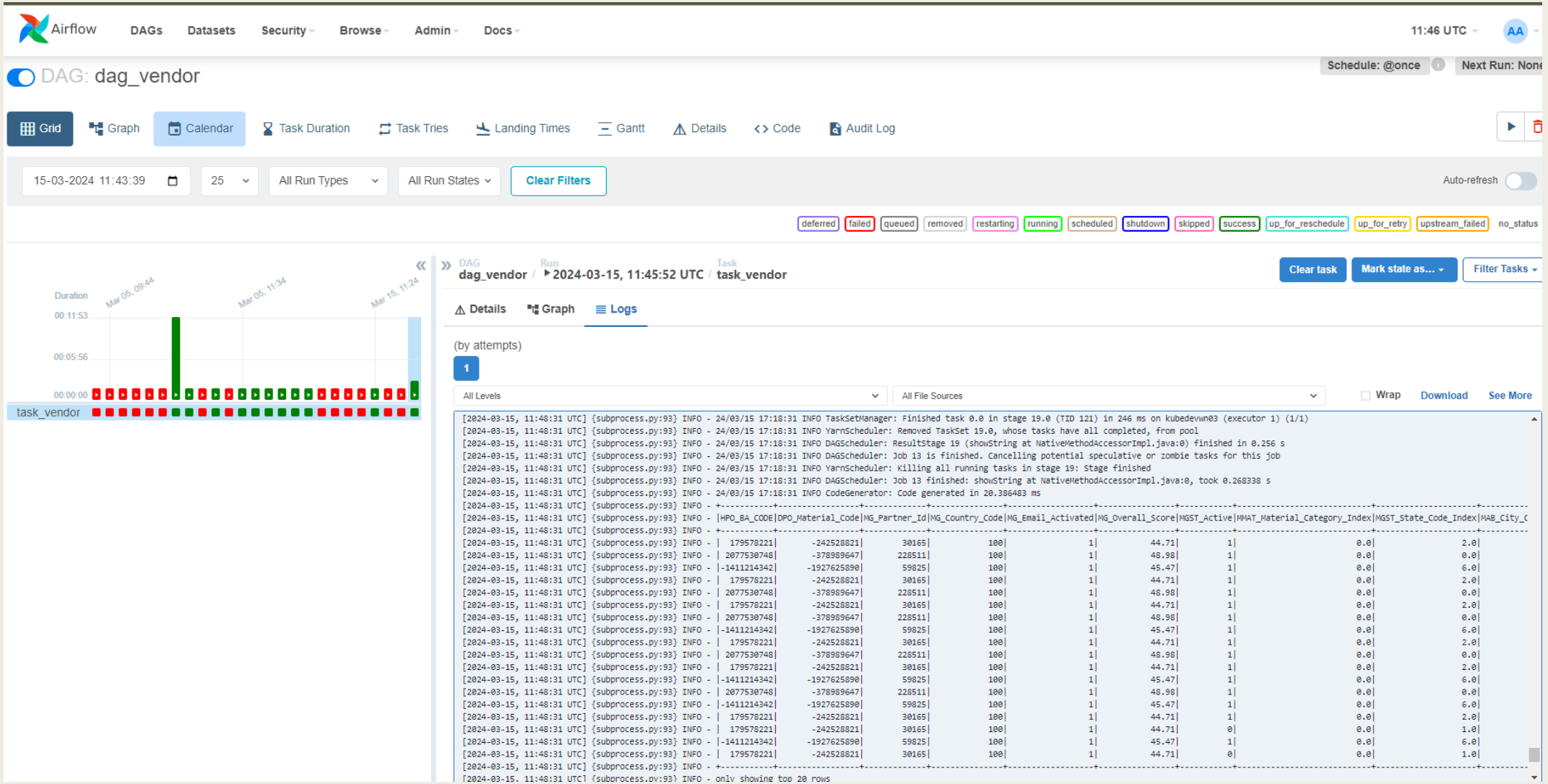
Work Done

- Data Collection: Gathered relevant data such as vendor profile id, product/service details, and Company feedback(.
- Data Preprocessing: Utilized PySpark for data cleaning, transformation, and feature engineering to prepare the data for analysis.
- Feature Selection: Identified key features and criteria for vendor evaluation, including experience, reputation, quality, and compliance
- Vendor Evaluation: Implemented PySpark algorithms to evaluate vendors based on the selected features and criteria.

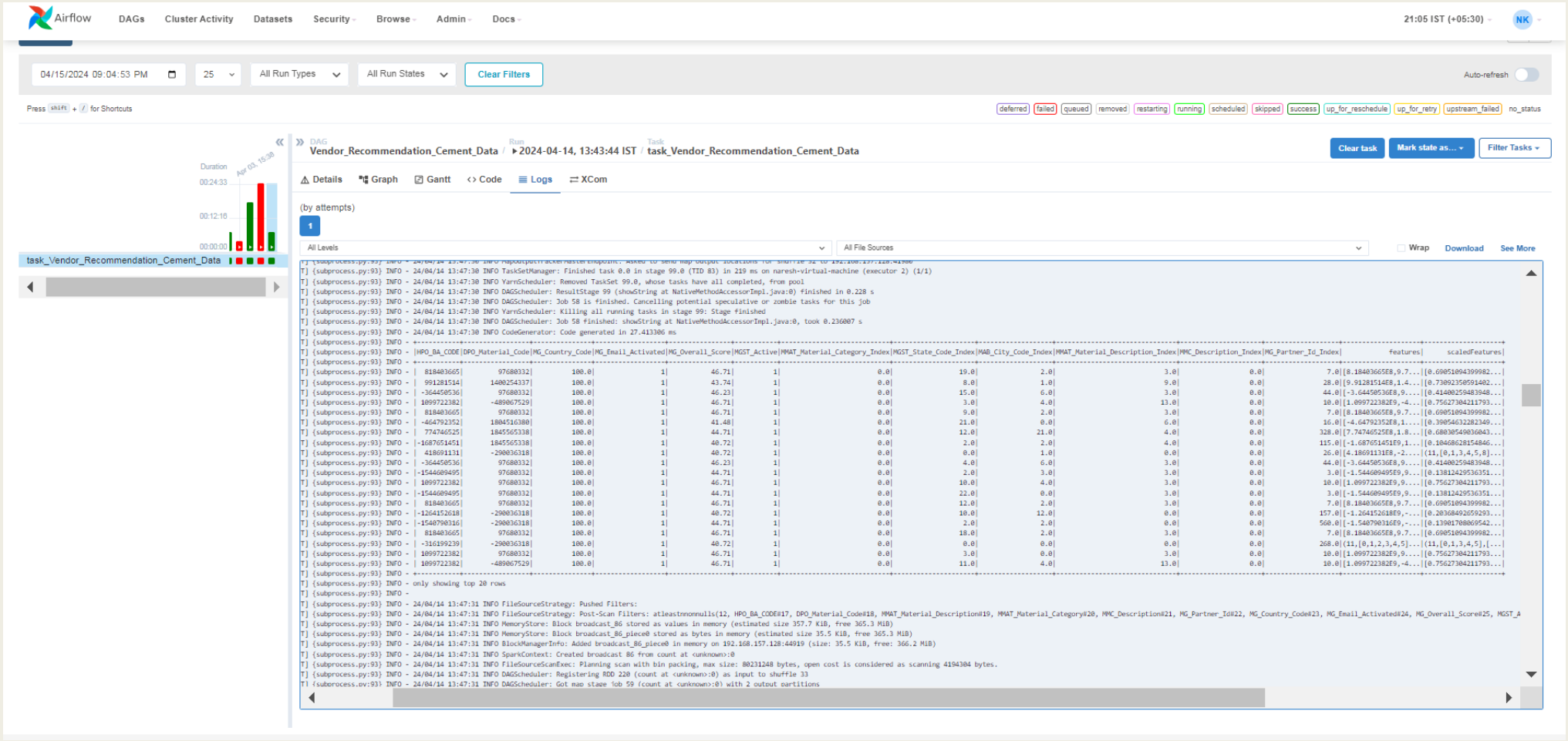
Work Done


Hadoop
Overview
Datanodes
Upload File
Choose Files
No file chosen
Upload
Open
This PC
New Volume (D:)
Vendor recommendation Pyspark
Search Vendor recommendati...
Organize
New folder
Sumresh
Vendor recomm
OneDrive
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
New Volume (D:
Network
Name
Date modified
Type
Size
vendor
15-03-2024 5:05 PM
Python File
2 KB
File name:
vendor
All Files
Open
Cancel
Name

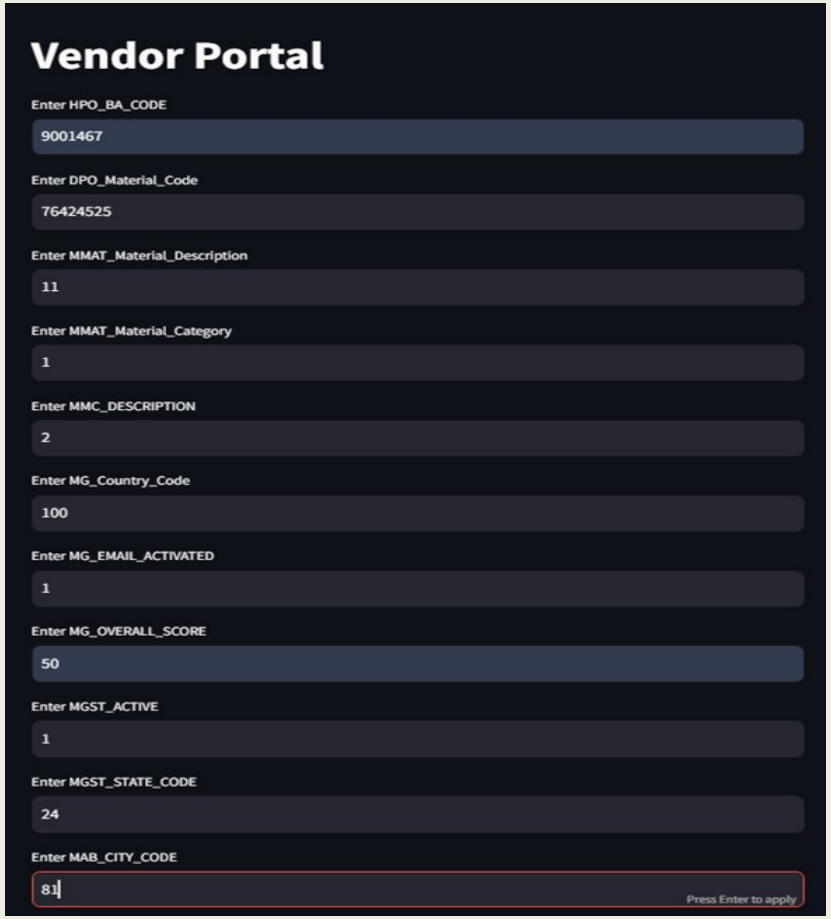
Work Done


Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
11:46 UTC
AA
DAG: dag_vendor
Schedule: @once
Next Run: None
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
15-03-2024 11:43:39
25
All Run Types
All Run States
Clear Filters
Auto-refresh
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Mar 05, 09:44
Mar 05, 11:34
Mar 15, 11:24
00:11:53
00:05:56
00:00:00
task_vendor
DAG
dag_vendor
Run
2024-03-15, 11:45:52 UTC
Task
task_vendor
Clear task
Mark state as...
Filter Tasks
Details
Graph
Logs
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO TaskSetManager: Finished task 0.0 in stage 19.0 (TID 121) in 246 ms on kubedevwn03 (executor 1) (1/1)
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO YarnScheduler: Removed TaskSet 19.0, whose tasks have all completed, from pool
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO DAGScheduler: ResultStage 19 (showString at NativeMethodAccessorImpl.java:0) finished in 0.256 s
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO DAGScheduler: Job 13 is finished. Cancelling potential speculative or zombie tasks for this job
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO YarnScheduler: Killing all running tasks in stage 19: Stage finished
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO DAGScheduler: Job 13 finished: showString at NativeMethodAccessorImpl.java:0, took 0.268338 s
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - 24/03/15 17:18:31 INFO CodeGenerator: Code generated in 20.386483 ms
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - +----------+-----------------+------------+--------------+-----------------+---------------+----------+----------------------------+----------------------+----------
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - |HPO_BA_CODE|DPO_Material_Code|MG_Partner_Id|MG_Country_Code|MG_Email_Activated|MG_Overall_Score|MGST_Active|MMAT_Material_Category_Index|MGST_State_Code_Index|MAB_City_C
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - +----------+-----------------+------------+--------------+-----------------+---------------+----------+----------------------------+----------------------+----------
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 1| 0.0| 2.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 2077530748| -378989647| 228511| 100| 1| 48.98| 1| 0.0| 0.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - |-1411214342| -1927625890| 59825| 100| 1| 45.47| 1| 0.0| 6.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 1| 0.0| 2.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 2077530748| -378989647| 228511| 100| 1| 48.98| 1| 0.0| 0.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 1| 0.0| 2.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 2077530748| -378989647| 228511| 100| 1| 48.98| 1| 0.0| 0.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - |-1411214342| -1927625890| 59825| 100| 1| 45.47| 1| 0.0| 6.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 1| 0.0| 2.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 2077530748| -378989647| 228511| 100| 1| 48.98| 1| 0.0| 0.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 1| 0.0| 2.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - |-1411214342| -1927625890| 59825| 100| 1| 45.47| 1| 0.0| 6.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 2077530748| -378989647| 228511| 100| 1| 48.98| 1| 0.0| 0.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - |-1411214342| -1927625890| 59825| 100| 1| 45.47| 1| 0.0| 6.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 1| 0.0| 2.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 0| 0.0| 1.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - |-1411214342| -1927625890| 59825| 100| 1| 45.47| 1| 0.0| 6.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - | 179578221| -242528821| 30165| 100| 1| 44.71| 0| 0.0| 1.0|
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - +----------+-----------------+------------+--------------+-----------------+---------------+----------+----------------------------+----------------------+----------
[2024-03-15, 11:48:31 UTC] {subprocess.py:93} INFO - only showing top 20 rows

Work Done


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
21:05 IST (+05:30)
NK
04/15/2024 09:04:53 PM
25
All Run Types
All Run States
Clear Filters
Auto-refresh
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
00:24:33
00:12:16
00:00:00
task_Vendor_Recommendation_Cement_Data
DAG
Vendor_Recommendation_Cement_Data
Run
2024-04-14, 13:43:44 IST
Task
task_Vendor_Recommendation_Cement_Data
Clear task
Mark state as...
Filter Tasks
Details
Graph
Gantt
Code
Logs
XCom
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
only showing top 20 rows

Work Done


Vendor Portal
Enter HPO_BA_CODE
9001467
Enter DPO_Material_Code
76424525
Enter MMAT_Material_Description
11
Enter MMAT_Material_Category
1
Enter MMC_DESCRIPTION
2
Enter MG_Country_Code
100
Enter MG_EMAIL_ACTIVATED
1
Enter MG_OVERALL_SCORE
50
Enter MGST_ACTIVE
1
Enter MGST_STATE_CODE
24
Enter MAB_CITY_CODE
81
Press Enter to apply

Work Plan

Day1(29-02-2024)	Day of Joining(Finishing Formalities), Introduction to Pyspasrk
Day2(01-03-2024)	Discussion of Vendor Recommendation project
Day3(04-03-2024)	Installation of PySpark
Day4(05-03-2024)	Task 1 and Task 2 done (Given by Company)
Day5(06-03-2024)	Learning Pyspark Data Preparation and executing Project
Day6(07-03-2024)	Learning Pyspark Data Preprocessing and executing Project
Day7(08-03-2024)	Data Preprocessing
Day8(11-03-2024)	Attended Webinar
Day9(12-03-2024)	Data Normalisation
Day10(13-03-2024)	Airflow Presentation (In Company)
Day11(14-03-2024)	Learning about Flask Api
Day12(15-03-2024)	Installation of Hadoop
Day13(16-03-2024)	First Review

Work Plan

Day14(18-03-2024)	Feature scaling data
Day15(19-03-2024)	Leave
Day16(20-03-2024)	Correcting errors in data normalization & Data Standardization
Day17(21-03-2024)	Learning about ml algorithms in detail
Day18(22-03-2024)	Applying ml algorithm to the data frame(logistic reg)
Day19(25-03-2024)	Researching about pyspark data analysis projects
Day20(26-03-2024)	Correcting errors in ml algorithm code
Day21(27-03-2024)	Coding the ml algorithms
Day22(01-04-2024)	Learning about ANN
Day23(02-04-2024)	Completed Task Given by company
Day24(03-04-2024)	Coding ML Algorithms
Day25(04-04-2024)	Leave
Day26(05-04-2024)	Second Review

Work Plan

Day14(18-03-2024)	Learning about Fast API
Day16(20-03-2024)	Learning about Streamlit
Day17(21-03-2024)	Selection of the Algorithm
Day18(22-03-2024)	Creating webpage using Streamlit
Day20(26-03-2024)	Integrating Steel and Cement files with webpage
Day21(27-03-2024)	Preparing Report
Day23(02-04-2024)	Completion of Project
Day26(05-04-2024)	Final Review

Tentative Results/Expected Outcome

- **Improved Vendor Recommendations:**
  - The trained recommendation model is expected to provide improved vendor recommendations based on user preferences and historical data analysis.
- **Personalized Recommendations:**
  - Users should receive personalized recommendations tailored to their specific needs and preferences, enhancing their overall experience.
- **Efficient Data Processing:**
  - Utilizing PySpark and distributed processing techniques ensures efficient handling of large-scale data, leading to faster processing times and improved scalability.
- **Model Performance Metrics:**
  - The model's performance metrics, such as accuracy, precision, recall, or F1-score, are expected to meet or exceed predefined thresholds, indicating a reliable recommendation system.
- **User Interface Interaction:**
  - The web application's user interface should provide an intuitive and user-friendly experience for inputting preferences and viewing recommendations.
- **Scalable Deployment:**
  - The deployed model should be scalable to handle varying loads of user requests, ensuring consistent performance under different traffic conditions.

Tentative Results/Expected Outcome

- **Automated Workflow:**
  - Utilizing Apache Airflow for workflow management enables automation of data preprocessing, model training, and deployment tasks, reducing manual intervention and potential errors.
- **Monitoring and Logging:**
  - Implementing monitoring and logging mechanisms ensures real-time tracking of system performance, errors, and anomalies, facilitating proactive maintenance and troubleshooting.
- **Integration and Compatibility:**
  - The web application should seamlessly integrate with the deployed model's API, allowing for smooth data exchange and recommendation delivery.
- **Business Impact:**
  - Ultimately, the expected outcome is a vendor recommendation system that adds tangible business value, such as increased sales, customer satisfaction, and retention through personalized and relevant recommendations.

References

JOURNAL REFERENCES:

- Wang, C.H., 2015. Using quality function deployment to conduct vendor assessment and supplier recommendation for business-intelligence systems. *Computers & Industrial Engineering*, *84*, pp.24-31.
- Sabri, Y., Micheli, G.J. and Cagno, E., 2022. Supplier selection and supply chain configuration in the projects environment. *Production planning & control*, *33*(12), pp.1155-1172.
- Alrumiah, S.S. and Hadwan, M., 2021. Implementing big data analytics in e-commerce: Vendor and customer view. Ieee Access, 9, pp.37281-37286.
- Peng, Y., 2022. A survey on modern recommendation system based on big data. arXiv preprint arXiv:2206.02631.
- Cheng, J. and Zhao, P., 2023. Sustainable Big Data Analytics Process Pipeline Using Apache Ecosystem. In Encyclopedia of Data Science and Machine Learning (pp. 1247-1259). IGI Global.

WEB REFERENCES:

- PySpark Tutorial – javatpoint
- https://www.youtube.com/watch?v=5peQThvQmQk
- https://www.youtube.com/watch?v=6kEGUCrBEU0&t=257s

THANK YOU