PROJECT REPORT

ON

"BIG-DATA Dataset Analysis Using Databricks, PySpark and Microsoft Azure"

FOR

"DINGIR,CIT"

BY

ALOK KUMAR
AVINASH KUMAR SONI
ABHISHEK KUMAR SINGH
AVINASH KUMAR KUSHWAHA
AMAN KUMAR SINGH

University/College Name:- Cambridge Institute of Technology, Tatisilwai, Ranchi, Jharkhand(835103)

Intership Period :- 20-05-2024 -07-07-2024

Internship Organization :- DINGIR

SUBMITTED TO - Deshbandhu Mishra Sir

Big Data Internship Project Assignment: Dataset Analysis Using Databricks, PySpark and Microsoft Azure

- **Objective:**
  - *The primary objective of this project is to analyze State/UT wise Accidents Classified according to Age of impacting Vehicles during 2017.*
  - *By leveraging Databricks powerful data processing and analysis capabilities, we aim to identify trends, patterns, and significant factors contributing to road accidents.*
  - *This analysis will provide insights that can help in formulating strategies to reduce road accidents in the future.*

- **Dataset Used:**

https://www.data.gov.in/resource/stateut-wise-accidents-classified-according-age-impacting-vehicles-during-2017

- **Source:**

https://www.data.gov.in/

- **Team Members:**

Sl.No.	Name	College Roll. No.	Registration No.	Semester
1	ALOK KUMAR	363/CSE/2022	22050440011	4th
2	AVINASH KUMAR SONI	342/CSE/2022	22050440030	4th
3	ABHISHEK KUMAR SINGH	381/CSE/2022	22050440005	4th
4	AVINASH KUMAR KUSHWAHA	378/CSE/2022	22050440029	4th
4	AMAN KUMAR SINGH	03/DIP/CSE/2022	22050445003	4th

- **Project Responsibilities and Roles Assigned**

1. **ALOK KUMAR** - Data Collection, Editing, Azure Databricks & Presentation
2. **AVINASH KUMAR SONI** - Data Processing and Transformation & Created Microsoft Azure Blob
3. **ABHISHEK KUMAR SINGH** - Statistical Analysis
4. **AVNASH KUMAR KUSHWAHA** - Data Visualization
5. **AMAN KUMAR SINGH** - Data Ingestion, Preprocessing, Report Writing & GitHub Linking

- **Tools and Technologies:**
  - - Databricks
  - - Databricks File System (DBFS)
  - - PySpark
  - - Spark SQL
  - - Python
  - - Visualization libraries (e.g., Matplotlib, Plotly)
  - - Azure Databricks
  - - Git and GitHub
  - - Microsoft Excel

- **Tools and Technologies Details:**

1. Databricks: Databricks is a unified analytics platform that accelerates innovation by unifying data science, engineering, and business. It is used for data processing, machine learning, and collaborative data analysis.

2. Databricks File System (DBFS):DBFS is a distributed file system integrated with Databricks, allowing easy storage and management of large datasets.

3. PySpark: PySpark is the Python API for Apache Spark, enabling Python developers to write Spark applications.

4. Spark SQL: Spark SQL is a module for structured data processing with Apache Spark. It allows querying of structured data using SQL.

5. Python: Python is a versatile programming language widely used in data science and machine learning.

6.Visualization Libraries: Visualization is crucial for understanding data patterns and communicating insights.(e.g., Matplotlib, Plotly)

7. Azure Databricks: Azure Databricks is a cloud-based analytics platform that combines the capabilities of Apache Spark and Databricks, provided as a first-party service on Microsoft Azure. It is designed to help users perform large-scale data processing and analytics, machine learning, and data engineering tasks.

8. Git and GitHub: Git is a version control system used for tracking changes in source code. GitHub is a platform for hosting and collaborating on Git repositories.

9. Microsoft Excel: Microsoft Excel is used for initial data inspection and basic analysis.

- **Detailed Responsibilities & Roles Assigned:**

1. **ALOK KUMAR - Data Collection, Transformation, Azure Databricks Linking & Presentation**

   I. *Data Collection*:
   - Finding the Dataset
   - Verified the upload by listing the contents of the DBFS directory

   II. *Data Cleaning and Transformation*:
   - Cleaned the dataset by handling any missing or inconsistent data.
   - Perform necessary transformation to prepare the data for analysis

   III. *Azure Databricks Linking*:
   - Connected Azure Blob Storage with Azure Databricks
   - Connected Azure Databricks with Community Databricks

   IV. *Created Presentation*:
   -Gave a detailed overview of the projects
   -Explained Various Steps using appropriate Screenshots for better Understanding of the Project

2. AVINASH KUMAR SONI - Data Processing & Created Microsoft Azure Blob Storage

I. *Load the Dataset into a Spark DataFrame*:

- Used PySpark to read the dataset from DBFS into a Spark DataFrame.

II. *Created a Temporary View*:

- Created a temporary view of the DataFrame for SQL querying.

III. *Created Microsoft Azure Blob Storage*:

- Created an Azure Storage Account
- Created Blob Storage Folders in Azure
- Create a Container in the Storage Account

3. ABHISHEK KUMAR SINGH - Statistical Analysis

I. *Analyze Demographic Impact using Spark SQL*:

- Calculated the total number of people injured or died in each state.
- Determine the total share of each state.
- Analyze the impact on different age groups.

II. *Find Patterns and Trends*:

- Identify any patterns or trends in the data (e.g., which age group is most affected in each state).

4. AVNASH KUMAR KUSHWAHA - Data Visualization

I. *Created visualizations to show*:

- The total number of people died/injured by vehicles in each state.
- The distribution of injured individuals across different age groups.
- Any trends or patterns identified in the analysis.

5. AMAN KUMAR SINGH - Data Ingestion, Preprocessing, Report Writing, & GitHub

I. *Setting up Databricks Environment*:

- Created a Databricks workspace.
- Set up a cluster with appropriate configurations.

II. *Upload the Dataset to DBFS*:

- Uploaded the dataset to Databricks File System (DBFS).
- Verified the upload by listing the contents of the DBFS directory

III. *Created a Report*:

- Compiled the findings into a comprehensive report.
- Included analysis, visualizations, and recommendations based on the data.
- Discussed any limitations or challenges faced during the analysis.

IV. *GitHub Linking*:

- Linked the Code and the databricks notebook with the GitHub
- GitHub Repository Link :
  https://github.com/aksgithub250502/Big-Data-Project-Repository.git

● Challenges Faced During the Analysis and It's Solutions:

I. Data Quality and Completeness:

Challenges:
Missing Data: Some records had missing values for critical variables.
Data Inconsistencies: Inconsistent data entries and formats across different sources.
Solutions:
Data Imputation: Used statistical methods or machine learning models to fill in missing values.
Data Cleaning: Cleaned inconsistencies using automated scripts and manual checks.

II. Data Integration:

Challenges:
Combining Multiple Sources: Integrating data from various sources with different formats.
Data Merge Issues: Merging datasets with different levels of detail and quality.
Solutions:
ETL Processes: Implement Extract, Transform, Load (ETL) processes to standardize and integrate data.

III. Technical Challenges:

Challenges:
Processing Large Datasets: Efficient handling and processing of large datasets.
Performance Optimization: Avoiding performance bottlenecks during complex queries and transformations.
Solutions:
Optimization Techniques: Use Spark's optimization features like caching and partitioning.

IV. Visualization Challenges:

Challenges:
Interactive Visualizations: Creating interactive visualizations that effectively communicate findings.
Data Representation: Ensuring accurate data representation without bias.
Solutions:
Advanced Visualization Tools: Used libraries like Plotly for interactive visualizations.
Best Practices: Followed data visualization best practices to ensure clarity and accuracy.

V. Collaboration and Coordination:

Challenges:
Team Coordination: Ensuring effective communication and coordination among team members.
Version Control: Managing changes to the codebase and data.
Solutions:
Collaboration Tools: Used tools like Slack and GitHub for communication and version control.
Regular Meetings: Hold regular team meetings to discuss progress and address issues.

VI. Time Constraints:

Challenges:
Project Timeline: Balancing thorough analysis with project deadlines.
Resource Allocation: Efficiently managing time and resources across different project phases.
Solutions:
Prioritization: Prioritize tasks based on importance and deadlines.

VII.Connecting Azure Databricks with Databricks Community Edition

Challenges:
Authentication and Authorization: Setting up correct authentication.
Resource Configuration: Properly configuring resources for seamless connectivity.
Solutions:
Configuration Guides: Followed detailed configuration guides and best practices.
Support Channels: Utilized support channels and documentation for troubleshooting.

- **Future Perspectives:**
  - **Extended Analysis:** Incorporate more recent data and additional variables.
  - **Collaborative Efforts:** Work with government agencies and stakeholders to implement and monitor interventions.
  - **Technology Integration:** Use IoT and telematics for real-time monitoring and predictive analytics.

- **Conclusion:**

  - The analysis of State/UT wise accidents classified according to the age of impacting vehicles during 2017 highlights critical insights into road safety in India.
  - The project provided significant insights into the impact of vehicle age on road accidents.
  - Data-driven approaches are crucial in addressing road safety issues.
  - Continuous efforts are needed to improve road safety standards and reduce accidents.

- **References:**

Databricks Documentation:
- Databricks Overview: https://docs.databricks.com/en/index.html
- Databricks File System (DBFS): https://docs.databricks.com/en/dbfs/index.html

PySpark Documentation:
- PySpark API Reference: https://spark.apache.org/docs/latest/api/python/
- Spark SQL Guide: https://spark.apache.org/docs/latest/sql-programming-guide.html

Python Libraries:
- Pandas Documentation: https://pandas.pydata.org/docs/
- NumPy Documentation: https://numpy.org/doc/

Visualization Libraries:
- Matplotlib Documentation: https://matplotlib.org/stable/index.html
- Plotly Documentation: https://plotly.com/python/

Azure Databricks Documentation:
- Azure Databricks : https://learn.microsoft.com/en-us/azure/databricks/

Road Accident Data Sources:
- Open Government Data (OGD) Platform India: https://www.data.gov.in/
- Road Accidents Dataset : https://www.data.gov.in/catalog/road-accidents-india-2017

General References:
- YouTube: https://www.youtube.com/
- Google: https://www.google.com/

❖ These references provide a comprehensive foundation for understanding the tools, technologies, and methodologies used in the analysis of state/UT wise road accidents classified according to the age of impacting vehicles during 2017. They also offer additional resources for further study and exploration in data analysis, statistical analysis.

***** END *****