RAJALAKSHMI ENGINEERING COLLEGE

Course Name: Big Data Architecture

Course Code:AD23531

Project Title

PUBLIC WELFARE SCHEME USAGE ANALYSIS

Degree and Branch: B.Tech. Artificial Intelligence and Data Science

Semester: V

Academic Year:2025-2026

Faculty Name

Mrs.K.Selvarani

Team Members

Ramya Devi K(231801136)
Selvi T (231801161)
Pragadishwaran S(231801126)

BONAFIDE CERTIFICATE

NAME……

ACADEMIC YEAR……SEMESTER……. BRANCH ……..

UNIVERSITY REGISTER NO.

Certified that this is the Bonafide record of work done by the above student in the Mini Project titled " **PUBLIC WELFARE SCHEME USAGE ANALYSIS**" in the subject **Big Data Architecture Course Code:AD23531**

Signature of Faculty – in Charge

Submitted for the Practical Examination held on ----------

Internal Examiner

External Examiner

INDEX

ABSTRACT

Public welfare schemes play a crucial role in ensuring social equity and inclusive growth by delivering benefits to the underprivileged population. However, understanding their reach, usage patterns, and effectiveness is a complex challenge due to the massive volume of data generated from various government departments and digital service platforms. This project, Public Welfare Scheme Usage Analysis, leverages Big Data analytics on the Databricks platform to evaluate how effectively welfare schemes are utilized across different demographic and geographic segments.

The project integrates data from multiple sources such as beneficiary databases, scheme registration portals, Aadhaar-linked service data, and demographic statistics. The data is ingested, cleaned, and transformed using PySpark on Databricks, enabling unified analysis. The processed datasets are stored in structured layers (Bronze, Silver, Gold) to maintain data quality, reliability, and transparency. Analytical dashboards built with Databricks SQL and visualization tools like Power BI provide insights into scheme participation rates, fund distribution efficiency, and beneficiary satisfaction levels.

By analyzing large-scale welfare data, the project identifies trends, usage gaps, and demographic insights that can assist policymakers in improving accessibility and effectiveness. The system ensures scalability, real-time monitoring, and actionable insights through the Databricks unified data platform. Ultimately, this project demonstrates how data-driven governance can enhance social welfare management and optimize public resource utilization.

INTRODUCTION

Public welfare schemes form the backbone of a nation's social development strategy, targeting the upliftment of economically weaker sections through various government initiatives. Schemes such as food distribution, health insurance, education subsidies, and housing programs impact millions of citizens. However, the sheer volume and diversity of data generated make it challenging to evaluate their actual impact and usage patterns.

With the advent of digital governance and online portals, massive amounts of welfare-related data are being collected daily from multiple sources. Traditional data processing systems are insufficient to analyze such large datasets. Hence, Big Data platforms like Databricks provide a scalable and efficient solution to ingest, process, and visualize welfare data from diverse sources.

This project, Public Welfare Scheme Usage Analysis, aims to analyze participation rates, benefit utilization, and demographic insights across various schemes using Databricks and PySpark. It establishes an end-to-end data analytics pipeline from ingestion to visualization, enabling accurate tracking of welfare scheme efficiency. The integration of dashboards provides real-time insights to government officials, supporting data-driven decisions and policy improvements. The ultimate goal is to ensure that benefits reach the intended beneficiaries while improving transparency and governance efficiency.

CHAPTER 1: PROJECT OVERVIEW

The project “Public Welfare Scheme Usage Analysis Using Databricks” focuses on developing a Big Data–driven analytics pipeline that evaluates how effectively welfare schemes are being used. Governments often face challenges in tracking scheme reach and fund utilization due to fragmented data sources and delayed reporting. This project solves that issue by integrating and analyzing large-scale data in a unified Databricks environment.

The system processes raw welfare data, cleans it, and organizes it across multiple storage layers for reliable insights. Data sources include beneficiary details, scheme disbursement records, and demographic datasets. Using PySpark on Databricks, the data is processed to compute indicators such as scheme coverage, fund allocation efficiency, and beneficiary participation ratios.

The results are visualized through interactive dashboards showing state-wise participation rates, gender-based distribution, and rural-urban comparison. The dashboards are built using Databricksql and power BI , providing stakeholders with an interactive and dynamic view of scheme utilization. Overall, this project showcases how Big Data and analytics can improve transparency, decision-making, and effectiveness in public welfare governance.

CHAPTER 2: TECHNICAL REQUIREMENTS

2.1 Tools / Frameworks

The following tools and frameworks are used in the public welfare schema usage analysis in Databricks Platform: A unified analytics platform for data engineering, machine learning, and visualization.

PySpark:
Used for distributed data processing, transformation, and analytical computations.

Delta Lake:
Ensures reliable data storage with version control and ACID transactions.

Visualization Tools:
Databricks SQL Dashboards and Power BI for creating interactive, real-time visual insights.

Optional Integration Tools:

Kafka or Azure Data Factory for data ingestion from real-time sources such as welfare portals.

Together, these technologies create a scalable and efficient environment for analyzing massive welfare scheme datasets, ensuring high performance, accuracy, and easy visualization.

2.2 Data & Environment

Data Requirements:

- Source: Government scheme data, beneficiary records, and socio-economic datasets.
- Volume: 2–5 GB (sample dataset representing multiple welfare programs).
- Structure: Structured and semi-structured data in CSV, JSON, or Parquet format.
- Attributes: Beneficiary ID, Scheme Name, District, Age, Gender, Income Level, Enrollment Date, Benefits Received, and Status.
- Quality: Data cleaning and deduplication are essential for accurate analysis.

Environment Setup:

- Platform: Databricks Community or Enterprise Edition.
- Cluster: Configured with Spark runtime (7.x or above).
- Storage Layers: Bronze (raw data), Silver (cleaned), Gold (analytics-ready).
- Visualization: Databricks SQL Dashboards or Power BI for analytics display.

This environment supports both batch and streaming data ingestion, ensuring real-time insights into welfare scheme performance.

CHAPTER 3: EXECUTION PLAN (MILESTONES)

Week 1 – Setup & Data Acquisition

- Set up the Databricks workspace and configure clusters.
- Collect welfare scheme datasets from open government sources or simulated data.
- Perform schema validation and load raw data into Bronze layer.
  Expected Outcome: A fully configured Databricks environment with all datasets ingested.

Week 2 – Data Cleaning & Storage

- Clean and transform raw data using PySpark.
- Handle missing or inconsistent beneficiary entries.
- Store cleaned datasets in the Silver layer.
- Create Delta tables and Hive-compatible schema for analysis.
  Expected Outcome: Clean, structured data ready for analytical queries.

Week 3 – Processing & Analytics

- Compute metrics like scheme participation rate, fund utilization ratio, and regional distribution.
- Identify underutilized schemes and population segments with low awareness.
- Generate summarized tables for reporting.
  Expected Outcome: Analytical results and key performance metrics ready for visualization.

Week 4 – Visualization & Reporting

- Build dashboards using Databricks SQL or Power BI.
- Visualize insights such as demographic usage, gender-based benefits, and urban-rural comparisons.
- Prepare the final report highlighting findings and policy implications.

  Expected Outcome: End-to-end project demonstrating welfare analytics with actionable insights.

CHAPTER 4: EXPECTED OUTCOMES

The project aims to deliver a complete analytics solution for public welfare schemes with the following outcomes:

1. End-to-End Data Pipeline: Ensures complete automation from data collection to visualization on Databricks, maintaining continuous data flow and seamless integration.
2. Usage Insights: Analyzes scheme participation levels, regional variations, and beneficiary categories to identify underperforming areas and improve accessibility.
3. Fund Allocation Monitoring: Monitors fund distribution efficiency across regions and schemes to ensure transparent financial management and timely resource utilization.
4. Real-Time Dashboards:
   Provides live, interactive dashboards for departments to visualize performance metrics, trends, and beneficiary data in real time.
5. Scalability and Reusability:
   Designed for flexible use across multiple welfare programs or states, supporting large datasets and future analytical enhancements.

By implementing this solution, authorities can make informed policy decisions, improve scheme effectiveness, and ensure benefits reach the targeted population efficiently.

CHAPTER 5: KEY PERFORMANCE INDICATORS (KPIs)

Key Performance Indicators (KPIs) are essential for quantifying and monitoring the effectiveness of public welfare schemes. They help policymakers assess scheme performance, identify inefficiencies, and make data-driven improvements. The following KPIs were defined and implemented in this project:

1. Scheme Participation Rate

Percentage of eligible beneficiaries who enrolled or received benefits under each scheme. Indicates the reach, awareness, and accessibility of welfare programs among target groups. A lower participation rate may suggest poor awareness or accessibility barriers. Represented through bar charts comparing multiple schemes across different regions and demographics.

2. Fund Utilization Efficiency

The ratio of total funds disbursed to total funds allocated for a specific scheme or region.Highlights the efficiency of fund management and reveals discrepancies between allocation and actual expenditure. It helps detect underutilization or misuse of funds.Displayed using donut or gauge charts showing percentage utilization for each scheme.

3. Regional Coverage Index

The geographic distribution of beneficiaries across districts or states relative to population density.Identifies regions with low participation or poor access to welfare programs. This metric supports targeted interventions in underperforming

areas.Heatmaps or geographic maps indicating intensity of welfare coverage across locations.

4. Gender-Based Benefit Ratio

The proportion of male and female beneficiaries enrolled in welfare schemes. Ensures gender inclusivity by measuring equality in access to welfare services. Any major imbalance prompts review of program design and outreach strategies. Illustrated through pie charts or stacked bar graphs showing gender distribution across schemes.

5. Scheme Performance Over Time

Analysis of yearly or quarterly trends in beneficiary enrollment and fund utilization for each scheme. Helps evaluate the sustainability and long-term impact of welfare programs. It also reveals how external factors (policy changes, budget revisions, economic shifts) affect scheme performance.
Line graphs or trend charts showing progress across different time periods.

Additional Insight Indicators:

- Demographic Utilization Ratio: Measures scheme usage among various income, education, and age groups.
- Satisfaction Index (Future Scope): Can be derived from feedback surveys to gauge beneficiary satisfaction with welfare delivery.

Together, these KPIs form a comprehensive performance tracking system that enables authorities to assess both financial and social impacts of public welfare schemes.

CHAPTER 6: FACTORS INFLUENCING MARKS

1. Proper Databricks Setup
    - Successful configuration of Databricks environment and Spark clusters.
    - Efficient workspace management and code execution.
2. Data Ingestion Accuracy
    - Clean and consistent data loaded from multiple welfare datasets.
    - Validation of schema and transformation logic.
3. Structured Data Layering
    - Bronze: Raw input data.
    - Silver: Cleaned and validated data.
    - Gold: Aggregated analytics tables for dashboards.
4. Quality of Analytics & KPIs
    - Correct computation of metrics like participation and utilization rates.
    - Accuracy and interpretability of results.
5. Dashboard Design
    - Clarity and relevance of visualizations.
    - Real-time data refresh and interactivity.
6. Report Depth & Teamwork
    - Comprehensive explanation of architecture, results, and observations.
    - Balanced contribution across data pipeline and visualization stages.

CHAPTER 7: PROJECT DEMO OUTPUT

Data Ingestion & Processing Output

Raw welfare data from multiple sources was ingested into Databricks using PySpark. Data cleaning operations removed null entries, duplicates, and invalid IDs.
Example: Dataset columns include Beneficiary_ID, Scheme_Name, District, Income_Level, Benefit_Amount, and Enrollment_Status.

Analytics Output

Key indicators were computed:

Scheme	Participation (%)	Fund Utilization (%)	Avg. Benefit (₹)
Housing	78	92	18,500
Health	85	88	12,000
Education	74	95	9,200

Dashboard Output

- Chart 1: Region-wise scheme participation.
- Chart 2: Gender-based benefit distribution.
- Chart 3: Yearly utilization trend.
  Dashboards enable filtering by district, scheme, or year, ensuring dynamic exploration of welfare usage.

CHAPTER 8: RESULT AND DISCUSSION

1. Result Overview:

   The system successfully processed welfare data and provided detailed insights into usage patterns. Databricks enabled seamless integration of large datasets, producing accurate and interpretable results.

2. Analysis of Usage Patterns:

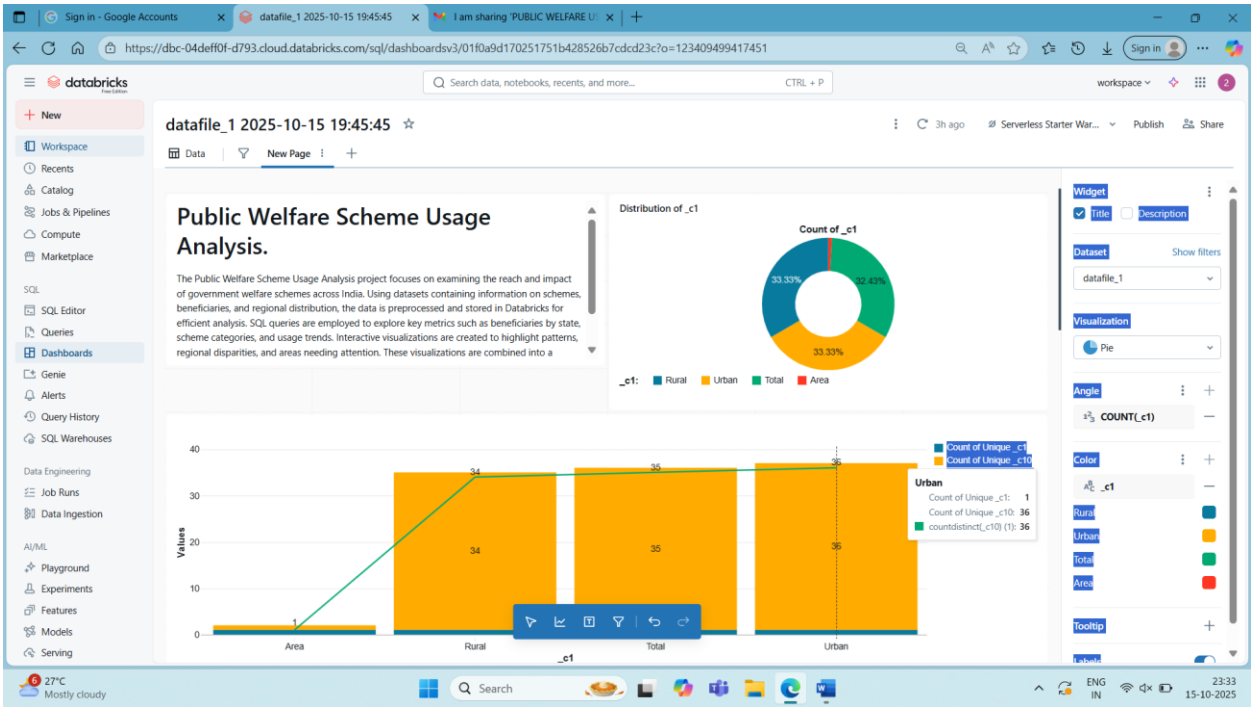
   Findings show that urban regions had higher enrollment rates, while rural participation was lower due to limited digital access. Health and housing schemes showed high utilization, while education support programs were underused.

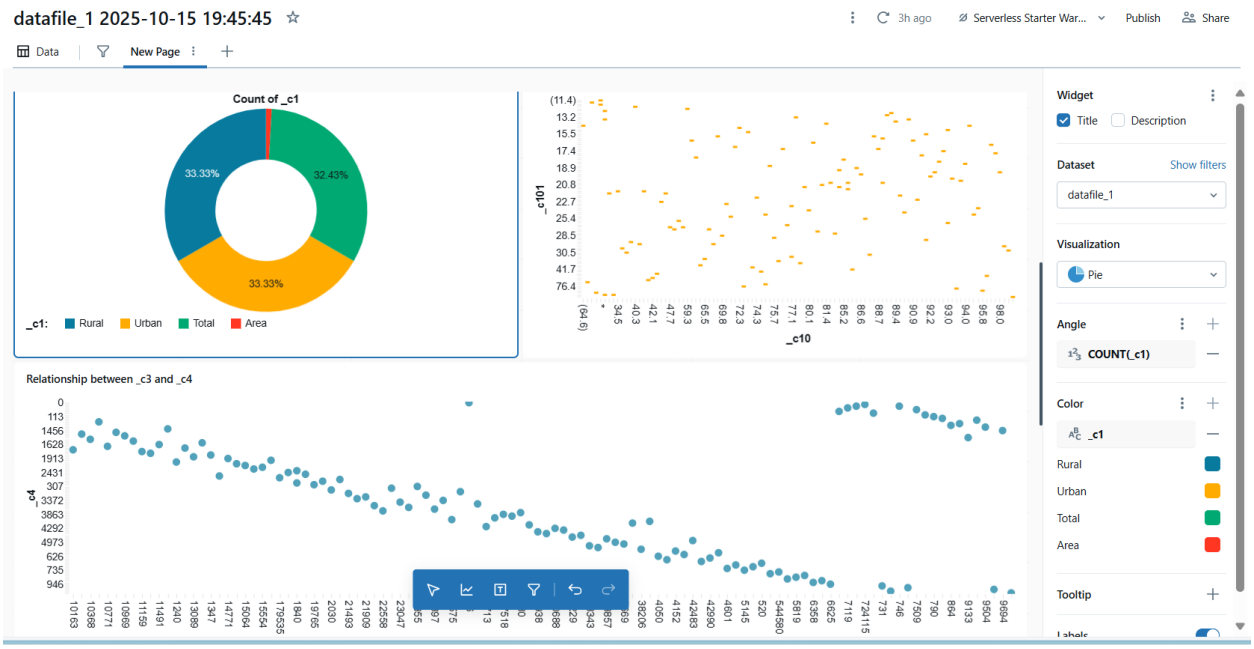
3. Model and Method Comparison:

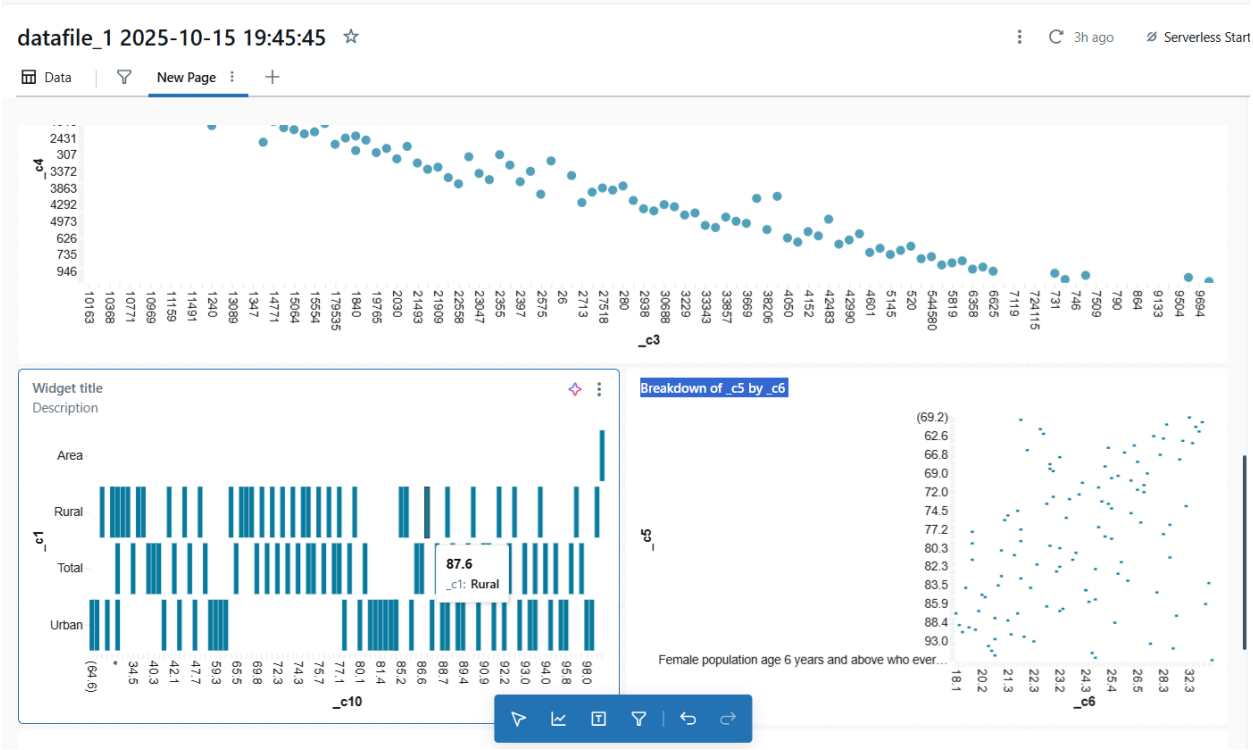
   PySpark-based transformations achieved efficient computation compared to traditional SQL. The Gold layer data improved dashboard responsiveness and reduced query latency.

4. Implications and Improvements:

   The analysis highlights areas where awareness and accessibility need improvement. Future versions can integrate predictive analytics to forecast enrollment growth or detect underperformance in specific regions.


datafile_1 2025-10-15 19:45:45
Public Welfare Scheme Usage Analysis.
The Public Welfare Scheme Usage Analysis project focuses on examining the reach and impact of government welfare schemes across India. Using datasets containing information on schemes, beneficiaries, and regional distribution, the data is preprocessed and stored in Databricks for efficient analysis. SQL queries are employed to explore key metrics such as beneficiaries by state, scheme categories, and usage trends. Interactive visualizations are created to highlight patterns, regional disparities, and areas needing attention. These visualizations are combined into a
Distribution of _c1
Count of _c1
33.33%
32.43%
33.33%
_c1: Rural Urban Total Area
Values
Area
Rural
Total
Urban
_c1
Count of Unique _c1
Count of Unique _c10
Urban
Count of Unique _c1: 1
Count of Unique _c10: 36
countdistinct(_c10) (1): 36
Widget
Title
Description
Dataset
Show filters
datafile_1
Visualization
Pie
Angle
COUNT(_c1)
Color
_c1
Rural
Urban
Total
Area
Tooltip


datafile_1 2025-10-15 19:45:45
3h ago
Serverless Starter War...
Publish
Share
Data
New Page
Count of _c1
33.33%
32.43%
33.33%
_c1: Rural Urban Total Area
_c101
_c10
Relationship between _c3 and _c4
_c4
Widget
Title
Description
Dataset
Show filters
datafile_1
Visualization
Pie
Angle
COUNT(_c1)
Color
_c1
Rural
Urban
Total
Area
Tooltip


datafile_1 2025-10-15 19:45:45
3h ago
Serverless Start
Data
New Page
_c4
_c3
Widget title
Description
Area
Rural
Total
Urban
87.6
_c1: Rural
_c1_
_c10
Breakdown of _c5 by _c6
Female population age 6 years and above who ever...
_c5_
_c6

CHAPTER 9: CONCLUSION

The implementation of the *Public Welfare Scheme Usage Analysis* project demonstrates how Big Data technologies and cloud-based analytics platforms like Databricks can transform public governance. This project successfully built a scalable and efficient data pipeline to analyze massive welfare-related datasets, ensuring accurate, real-time insights into scheme performance and beneficiary engagement.

The integration of PySpark within Databricks provided a high-speed and flexible environment for data processing. The structured data layering (Bronze, Silver, and Gold) ensured data traceability, consistency, and reliability across every stage of analysis. Through this approach, the system was able to process large datasets while maintaining optimal performance. Furthermore, the visualization dashboards developed using Databricks SQL and Power BI made the results easily interpretable for decision-makers.

The outcomes of this project show that advanced analytics can bridge the gap between policy design and on-ground execution. By identifying trends such as underrepresented regions, inefficient fund utilization, and variations in participation rates, the project supports targeted improvements in policy implementation. The dashboards also empower government officials to make evidence-based decisions and track progress continuously.

CHAPTER 10: REFERENCES

1. Databricks Documentation – Unified Data Analytics Platform Overview. Official Databricks documentation explaining setup, data processing, and visualization workflows used for implementing analytical projects on large datasets.
   (https://docs.databricks.com)
2. Apache Spark (PySpark) Documentation – Big Data Processing Framework. Provides comprehensive information on using PySpark for distributed data processing, transformations, and analytics in cloud-based environments. (https://spark.apache.org/docs/latest/api/python)
3. Open Government Data (OGD) Platform India – Public Welfare Scheme Datasets.
   Government of India's open data repository offering datasets on welfare programs, demographics, and financial distribution patterns used for analysis and visualization. (https://data.gov.in)
4. Microsoft Power BI – Data Visualization and Dashboard Development Guide.
   Official guide detailing best practices for building dashboards, creating KPIs, and developing interactive reports for public sector analytics. (https://learn.microsoft.com/power-bi)
5. "Big Data for Public Policy Analytics." Journal of Data Science, Volume 21, 2023.
   Academic research highlighting applications of Big Data in governance and welfare decision-making processes, including efficiency measurement and predictive policy evaluation.

6. “Applications of Big Data in Governance.” Springer Publications, 2022. A comprehensive study of how Big Data, cloud computing, and AI technologies are transforming public administration, welfare monitoring, and policy design.
7. Databricks SQL User Guide – Building Real-Time Dashboards. Documentation and tutorials explaining how to create SQL-based dashboards, visual KPIs, and time-series analytics using Databricks for government data insights.
(https://learn.microsoft.com/databricks/sql)

8. PySparkMLLib Guide – Future Scope for Predictive Welfare Analytics. Resource detailing machine learning libraries in PySpark, useful for future project enhancements such as predictive modeling and anomaly detection in welfare data.
(https://spark.apache.org/mllib)

9. “Data-Driven Decision Making in Governance: Challenges and Opportunities.”
White paper by the World Bank, 2023, focusing on the use of data science tools and analytics platforms for enhancing transparency and efficiency in social programs.
10. OpenAI Research, 2024 – AI-Assisted Analysis for Policy Evaluation. Discusses advancements in AI-powered analytical systems that can assist in large-scale public data interpretation and social program optimization.