Big Data Analytics Project: Hospital Stay & Volume Analysis

Student: D.Jeshwitha

Roll No. / Academic Year: 2211CS010669 (S2-87) / 2025 - 2026

1. Executive Summary

This report summarizes the progress and initial findings of the Big Data Analytics (BDA) project focused on hospital admission data. The primary goal is to leverage Apache Spark (PySpark) to analyze 55,500 patient records to identify key operational trends. The project has successfully completed the Data Loading, Preprocessing, and **Temporal Analysis** phases, revealing a stable average Length of Stay (LOS) despite fluctuations in monthly patient volume. Future work is focused on completing categorical and bivariate analyses to provide actionable insights for resource planning and patient care optimization.

2. Methodology and Technical Stack

2.1. Technical Environment

Feature	Description	Value/Type
Platform	Distributed processing framework	**Apache Spark** (PySpark)
Data Volume	Total records analyzed	**55,500** rows
Tooling	PySpark DataFrames, pyspark.sql.functions	

2.2. Data Preprocessing

The raw data was transformed into a usable structure for large-scale aggregation:

1. **Temporal Transformation:** A new field, Admission_Year_Month, was created from the Date of Admission to facilitate time-series analysis.
2. **Categorical Binning:** The Age column was binned into an **Age Group** categorical variable (e.g., 21-40, 61-80) to segment the patient population.

2.3. Sample Data Structure

Column Name	Data Type	Role in Analysis
Date of Admission	Date	Time-series base
Age Group	String	Key categorical variable
Length of Stay (Days)	Numeric	Target variable (LOS)

3. Completed Analysis: Temporal Trends

3.1. Analytical Approach

The Temporal Analysis utilized PySpark's aggregation capabilities by grouping the entire dataset by Admission_Year_Month. The following two key operational metrics were computed:

- **Monthly_Patients**: Total patient volume (count) per month.
- **Avg_Length_of_Stay**: Mean value of the LOS (in days) per month.

3.2. Key Findings

The results reveal distinct patterns that are crucial for hospital management:

- **Stable LOS:** The average Length of Stay is remarkably consistent across the observed months, consistently falling within a narrow range of **15 to 16 days**. This establishes a critical operational baseline.
- **Fluctuating Volume:** Patient admissions show significant month-to-month variation (e.g., sampled range from **686 to 1001** patients), suggesting the need for flexible staffing and resource deployment to handle peak periods.

Admission_Year_Month	Monthly_Patients	Avg_Length_of_Stay
2019-05	686	15.126
2019-06	907	15.422
2019-07	957	15.928
2019-08	1001	15.695

4. Planned Future Work

The project is moving into the deeper analytical phases to explore causal factors and predict outcomes.

Analysis Phase	Goal	Impact
1. Categorical Analysis (LOS by Age)	Quantify LOS variations across different Age Group demographics.	Inform resource allocation and specialized care planning for high-LOS age segments.
2. Bivariate Analysis (LOS vs. Billing)	Investigate the correlation between Length of Stay and associated billing data.	Identify drivers of operational cost and efficiency within the hospital system.
3. Predictive Modeling	Develop a Machine Learning Regression Model to predict patient LOS upon admission.	Enhance hospital forecasting, capacity management, and bed allocation strategies.

5. Conclusion

The successful processing of 55,500 records using Apache Spark validates the technical pipeline established for this project. The initial Temporal Analysis provides foundational insights into patient volume and LOS stability. The remaining analytical phases will transform this raw data into sophisticated, actionable intelligence to optimize hospital operations.