**Telecommunications--Usage Behavior Segmentation**

***Submitted by***

***RITHIKA SMITHI S 231801138***

***SANDHYA SREE M 231801146***

***SASI SRIRAM E 231801159***

***in partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**A logo of a torch

AI-generated content may be incorrect.A logo of a university

AI-generated content may be incorrect.**

**RAJALAKSHMI ENGINEERING COLLEGE (AUTONOMOUS) THANDALAM,**

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**ANNA UNIVERSITY, CHENNAI 600 025**

**OCT 2025**

i

# BONAFIDE CERTIFICATE

Certified that this Phase – II Thesis titled **Telecommunications--Usage Behavior** **Segmentation** is the Bonafide work of **RITHIKA SMITHI S(231801138),SANDHYA SREE (231801146),SASI SRIRAM (231801159),** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

**Dr. Suresh Kumar S M.E., Ph.D.,**

**Professor,**

Department of Artificial Intelligence & Data Science,

Rajalakshmi Engineering College

Thandalam – 602 105.

Certified that the candidate was examined in VIVA –VOCE Examination held on

**Internal Examiner External Examiner**

iii

## ACKNOWLEDGEMENT

Initially I thank the Almighty for being with us through every walk of my life and showering his blessings through the endeavor to put forth this report.

My sincere thanks to our Chairman **Mr. S. MEGANATHAN, M.E., F.I.E.,** and our Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, M.E., Ph.D.,** for providing me with the requisite infrastructure and sincere endeavoring educating me in their premier institution.

My sincere thanks to **Dr. S.N. MURUGESAN M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time.

I express my sincere thanks to **Dr. J M Gnanasekar M.E., Ph.D.,** Head of the Department of Artificial Intelligence and Data Science for his guidance and encouragement throughout the project work. I convey my sincere and deepest gratitude to our internal guide, **Dr. Suresh Kumar S M.E., Ph.D.,** Professor, Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Finally, I express my gratitude to my parents and classmates for their moral support and valuable suggestions during the course of the project.

.

### SANDHYASREE M RITHIKA SMITHI S SASI SRIRAM E

### 231801146 231801138 231801159

iv

## ABSTRACT

## Telecommunications companies generate massive volumes of customer usage data daily—from call durations and message frequencies to international calling patterns and service interactions. Effectively analyzing this data is critical for improving customer retention, optimizing network resources, and designing personalized service plans. Traditional analytical approaches, such as manual reporting or static customer categorization, are inefficient, error-prone, and fail to capture the dynamic behavioral patterns present in large-scale telecom systems.This project presents a comprehensive Big Data–driven Telecom Usage Behavior Segmentation System, implemented on the Databricks platform using Apache Spark, PySpark MLlib, Delta Lake, and Databricks SQL Dashboards. The system follows a structured Bronze–Silver–Gold data architecture to ensure data quality and reliability at each processing stage. Raw call detail records are ingested and stored as Delta tables in the Bronze layer, cleansed and encoded in the Silver layer, and transformed into analytical features in the Gold layer for machine learning. Using K-Means clustering, the system segments telecom users based on their voice usage, international calling activity, customer service interactions, and churn tendencies, revealing distinct behavioral groups such as high-value loyal users, high-churn risk customers, and low-usage cost-sensitive subscribers.The resulting clusters are evaluated through quantitative KPIs—including average daily charges, international usage rates, and churn percentages—and visualized through interactive dashboards in Databricks SQL and Python-based visual analytics. This unified Big Data pipeline achieves scalable, automated, and interpretable customer segmentation, empowering telecom companies to make data-driven marketing, retention, and pricing decisions. By integrating data ingestion, processing, modeling, and visualization in a single cloud-based ecosystem, the system enables proactive customer engagement, optimized operational strategies, and improved overall business intelligence within modern telecommunication infrastructures.

v

## TABLE OF CONTENTS

| CHAPTER NO | NAME | PAGE NO |
| --- | --- | --- |
| 1.1 | Background | 7 |
| 1.2 | Motivation | 7 |
| 1.3 | Objectives | 8 |
| 1.4 | Problem Statement | 8 |
| 1.5 | Scope of the Project | 9 |
| 2.1 | Traditional Customer Segmentation Methods | 10 |
| 2.2 | Statistical and Rule-Based Techniques in Telecom Analytics | 10 |
| 2.3 | Machine Learning–Based Segmentation | 11 |
| 2.4 | Big Data and Cloud-Based Approach | 11 |
| 2.5 | Clustering and Behavioral Analysis Techniques | 12 |
| 2.6 | Summary of Research Gaps | 12 |
| 2.7 | Contribution of the Present Work | 13 |
| 3.1 | System Overview | 14 |
| 3.2 | System Architecture | 14 |
| 3.3 | System Requirements | 15 |
| 3.3.1 | Hardware Requirements | 15 |
| 3.3.2 | Software Requirements | 16 |
| 3.4 | System Modules | 16 |
| 3.5 | Data Flow Diagram | 17 |
| 3.6 | Summary | 17 |
| 4.1 | Data Collection Module | 18 |
| 4.2 | Data Preprocessing Module | 18 |
| 4.3 | Data Storage using Delta Lake | 19 |
| 4.4 | Machine Learning Module (K-Means Clustering) | 20 |
| 4.5 | KPI Computation & Visualization Module | 20 |
| 4.6 | Dashboard and Insights Generation | 21 |
| 5.0 | IMPLEMENTATION | 22 |
| 6.0 | RESULTS & ANALYSIS | 23 |
| 7.0 | CONCLUSION | 25 |
| 8.0 | FUTURE ENHANCEMENTS | 26 |
| REFERENCES |  | 27 |

LIST OF FIGURES

| **FIGURE NO** | **NAME** | **PAGE NO** |
| --- | --- | --- |
| **1** | Big Data–Based Telecom Usage Behavior Segmentation Architecture on Databricks Platform | 13 |
| **2** | Bronze–Silver–Gold Data Processing Flow | 15 |
| **3** | Feature Engineering and K-Means Clustering Workflow | 19 |
| **4** | Clustered Telecom Customers by Usage Pattern (Visualization) | 22 |
| **5** | KPI Comparison of Customer Segments | 23 |
| **6** | Churn Rate vs Customer Service Calls Graph | 23 |
| **7** | Databricks SQL Dashboard for Telecom User Segmentation | 24 |

**CHAPTER 1**

**INTRODUCTION**

**1.1 Background**

In the telecommunications industry, understanding customer usage behavior is crucial for improving service quality, customer satisfaction, and long-term profitability. Every call, text, and data session generates valuable information about customer preferences, communication patterns, and spending habits. However, the sheer volume and velocity of this data—collected from millions of users across multiple regions—make traditional analytical approaches insufficient.

Conventional methods, such as manual customer profiling or static segmentation based on limited demographic factors, fail to capture the dynamic and complex behavior of modern telecom consumers. With the growth of 4G and 5G networks, the expansion of Internet-based services, and the rise of data-intensive applications, telecom operators must adopt scalable, automated, and intelligent analytical systems to derive insights from massive datasets.

By leveraging **Big Data technologies** such as **Apache Spark, Delta Lake, and Databricks MLlib**, organizations can efficiently process billions of call-detail records (CDRs) and customer service interactions in near real time. This enables data-driven customer segmentation, proactive churn prevention, and the design of personalized offers—ultimately transforming raw data into actionable business intelligence for modern telecommunications providers.

**1.2 Motivation**

The motivation for this project stems from the increasing need for telecom operators to retain customers, reduce churn, and enhance user experiences in an extremely competitive market. Traditional segmentation strategies often group users using static criteria like age or location, overlooking real-time behavioral variations such as call frequency, international usage, or service complaints.

With Big Data frameworks now capable of handling complex, high-volume datasets, it becomes possible to detect meaningful behavioral patterns automatically and in real time. This project demonstrates how the integration of **PySpark, Delta Lake, and Databricks** can uncover deep insights from raw telecom usage data—helping service providers identify loyal high-value users, dissatisfied customers, and cost-sensitive low-usage subscribers.

From an academic perspective, the project also illustrates the practical convergence of **distributed computing, data engineering, and unsupervised machine learning** in addressing a real-world business problem. It serves as a model for how data-driven analytics can empower industries to shift from reactive decision-making to predictive, customer-centric strategies.

**1.3 Objectives**

The primary objectives of this project are as follows:

* **To collect and process large-scale telecom usage datasets**, including call durations, message volumes, international usage, and customer service interactions.
* **To design an end-to-end Big Data pipeline** using **Databricks Spark environment**, following the **Bronze–Silver–Gold** layered architecture for reliable data ingestion, cleaning, and transformation.
* **To engineer analytical features** that effectively represent customer behavior and service engagement.
* **To apply unsupervised learning techniques (K-Means clustering)** for identifying distinct customer usage segments based on behavior and churn risk.
* **To compute Key Performance Indicators (KPIs)** for each segment—such as average charges, churn percentage, and service-call frequency—to generate actionable business insights.
* **To visualize analytical outcomes** through interactive dashboards in **Databricks SQL and Python-based visualizations**, supporting management decisions and marketing strategies.

The system aims to build a scalable, automated, and interpretable segmentation framework that enhances customer retention, improves service personalization, and supports data-driven telecom operations.

**1.4 Problem Statement**

Telecom operators face a major challenge in understanding the diverse and rapidly changing behavior of their customers. With millions of active users generating extensive call-detail and usage records daily, traditional analytical systems cannot efficiently process or interpret such high-volume data. Manual or rule-based segmentation approaches are often static, time-consuming, and unable to detect subtle behavioral trends or emerging churn risks.

Additionally, most existing customer analytics systems operate in silos, lacking seamless integration between data ingestion, processing, modeling, and visualization layers. This fragmentation reduces analytical accuracy and slows business responsiveness.

To address these limitations, this project proposes an **automated, Big Data-driven telecom usage behavior segmentation pipeline**, developed on the **Databricks platform** using **Apache Spark and MLlib**. The system ingests raw telecom datasets, performs distributed data cleaning and transformation, extracts relevant usage features, and applies clustering algorithms to identify meaningful customer segments. Through integrated visualization in Databricks SQL dashboards, business analysts can interpret usage trends, detect churn-prone groups, and plan targeted interventions. This unified, scalable, and cloud-based solution enables telecom providers to achieve faster insights, improved customer engagement, and better operational efficiency.

**1.5 Scope of the Project**

This project demonstrates a complete **Big Data architecture for intelligent telecom usage segmentation**, implemented end-to-end within the **Databricks ecosystem** for high scalability, automation, and analytical depth.

The workflow begins with the **Data Ingestion** phase, where telecom usage data—such as call durations, charges, and service interactions—is uploaded to Databricks and stored in the **Bronze layer** as raw Delta tables. In the **Data Processing phase**, Spark SQL and PySpark transformations clean, standardize, and enrich the data in the **Silver layer**, ensuring consistency and accuracy.

During the **Feature Engineering phase**, relevant numerical features are assembled and standardized in the **Gold layer** for machine learning. The **K-Means clustering algorithm** then segments customers based on similar usage behaviors, generating distinct groups such as high-spending loyal customers, moderate users, and high-churn risk customers.

The **Analytics and Visualization** phase integrates these results into **Databricks SQL dashboards** and **Python-based visual charts**, presenting KPIs like average charges, churn percentage, and service-call frequency per segment. These visualizations enable business users to interpret patterns, track segment performance, and design personalized retention or marketing campaigns.

The system’s **Scalability and Fault Tolerance** stem from Databricks’s distributed computing architecture, which can handle millions of records with minimal latency. The modular design supports future enhancements such as **real-time streaming analytics, churn prediction models, and integration with CRM systems**. By unifying ingestion, processing, analytics, and visualization, the proposed framework transforms traditional telecom data management into an intelligent, data-driven decision-support system—empowering operators to reduce churn, enhance customer experience, and drive long-term profitability.

**CHAPTER 2**

**LITERATURE SURVEY**

The analysis and segmentation of customer usage behavior in the telecommunications industry have been active areas of research, driven by the growing competition and the need for personalized customer engagement. Over the years, approaches have evolved from manual data analysis and rule-based profiling to advanced machine-learning and Big Data–driven segmentation frameworks. This chapter summarizes the most relevant studies and techniques used in telecom analytics, outlining their methodologies, limitations, and the technological evolution that motivates the present work.

**2.1 Traditional Methods**

In early telecommunications systems, customer segmentation was performed using **manual profiling and basic demographic grouping**. Marketing teams categorized users based on factors such as location, age, or subscription plan. Analysts relied on billing records, call-detail summaries, and customer surveys to approximate usage patterns.

While simple, these approaches suffered major limitations:  
Traditional telecom segmentation was **manual, time-consuming, and reactive**, providing limited insights into customer dynamics. They lacked the capability to process high-volume transactional data generated daily from millions of calls and data sessions. Moreover, static demographic groupings failed to capture evolving behavioral patterns influenced by service quality, pricing changes, and network usage variations. As a result, these methods were inadequate for modern telecom environments that demand **scalability, accuracy, and near–real-time behavioral insights**.

These drawbacks highlighted the need for automated, data-centric methods capable of processing continuous data streams from modern telecommunication systems.

**2.2 Statistical and Rule-Based Techniques**

Early automation in telecom analytics applied **statistical and threshold-based models** to classify or flag users with unusual activity. Examples include average call duration analysis, Z-score–based deviation detection, and ratio metrics such as outgoing-to-incoming call patterns.

Although intuitive and easy to implement, these **statistical methods produced a high rate of false positives** because they relied on rigid thresholds that could not adapt to dynamic customer behavior. Seasonal variations in usage, roaming periods, and promotional offers often skewed consumption patterns, leading to misclassification. Similarly, rule-based systems, such as “if-then” conditions to label high-value customers, failed to capture complex multi-feature interactions. Hence, purely statistical and rule-based approaches lacked the flexibility and intelligence necessary for large-scale, data-driven telecom segmentation.

**2.3 Machine Learning Approaches**

To enhance analytical precision, researchers introduced **machine-learning techniques** that learn patterns from customer usage data.

**Supervised models** such as **Decision Trees, Random Forests, and Support Vector Machines (SVM)** have been applied for churn prediction and customer classification, offering improved accuracy through feature-based learning. **Unsupervised algorithms**, particularly **K-Means Clustering, DBSCAN, and Gaussian Mixture Models**, have been widely used to segment customers into groups with similar call, data, or service-interaction behavior.

These ML-based methods outperform statistical models by capturing nonlinear dependencies and multi-dimensional relationships. However, they often require extensive feature engineering and face scalability issues when applied to billions of records generated in telecom networks. Furthermore, supervised models rely on labeled data—rarely available in real-world telecom scenarios—highlighting the need for **unsupervised and scalable Big Data frameworks** for effective segmentation.

**2.4 Big Data and Cloud-Based Approaches**

The explosion of telecom data from **Call Detail Records (CDRs)**, customer care logs, and mobile-app interactions has led to the adoption of **Big Data platforms** capable of processing information at massive scale and high velocity.

Modern studies and industry implementations employ **distributed computing frameworks** such as **Apache Hadoop** and **Apache Spark** to handle petabytes of data efficiently. Cloud platforms like **Databricks, AWS EMR, and Google Cloud Dataproc** have enabled the deployment of end-to-end pipelines combining data ingestion, cleaning, feature extraction, and machine-learning analysis in one environment.

In particular, **Spark MLlib** supports large-scale clustering and predictive analytics, while **Delta Lake** ensures reliable, versioned data storage through Bronze–Silver–Gold layering. On Databricks, these technologies integrate seamlessly with **SQL dashboards and visualization tools**, providing real-time insights and business intelligence. Such architectures overcome the limitations of single-machine ML setups by offering **scalability, fault tolerance, and automated workflow orchestration** for telecom analytics.

**2.5 Behavioral Segmentation and Clustering Techniques**

Recent research has focused on **unsupervised behavioral segmentation**, where clustering algorithms are applied to customer usage vectors composed of call frequency, duration, charges, and service interactions.

**K-Means Clustering** remains one of the most effective methods due to its scalability and interpretability. Customers with similar communication habits are grouped into clusters representing patterns such as high-value frequent callers, low-usage casual users, or high-complaint customers. **Hierarchical clustering** and **Principal Component Analysis (PCA)** are also utilized for dimensionality reduction and visualization of user similarities.

Additionally, hybrid approaches combine clustering with churn-prediction models to link behavioral segments to churn probability. These methods enable telecom operators to adopt **targeted marketing, personalized offers, and proactive retention strategies**, all supported by objective, data-driven insights.

**2.6 Summary of Research Gaps**

Despite significant progress in telecom data analytics, several research gaps remain:

* **Scalability:** Many ML algorithms fail to process continuously growing datasets efficiently in distributed environments.
* **Integration:** Existing studies often treat data preparation, modeling, and visualization as separate stages, lacking end-to-end automation.
* **Interpretability:** While clustering identifies patterns, few systems translate these into actionable business insights.
* **Real-time capability:** Most segmentation frameworks operate on historical batch data, limiting their ability to capture evolving customer behaviors.
* **Visualization and Decision Support:** Comprehensive, cloud-based dashboards that combine KPIs with analytical outcomes are still limited in research.

These gaps motivate the development of a **unified, scalable, and explainable Big Data–driven segmentation pipeline** tailored for telecom usage analytics.

**2.7 Contribution of the Present Work**

This project bridges the identified gaps by developing an **end-to-end Big Data–driven Telecom Usage Behavior Segmentation System** using the **Databricks cloud platform**.

Key contributions include:

* **Implementation of a PySpark-based data processing pipeline** following the Bronze–Silver–Gold layered architecture for structured and efficient data handling.
* **Integration of feature engineering and K-Means clustering** using Spark MLlib to segment customers based on multidimensional usage behavior.
* **Computation of Key Performance Indicators (KPIs)** for each segment, such as average day charges, international usage, and churn percentage, to support interpretability.
* **Development of interactive dashboards** in Databricks SQL and Python-based visualizations to present insights in real time.
* **Demonstration of scalability and modularity**, showing how the framework can be extended to include predictive churn models, streaming ingestion, and CRM integration.

Overall, the project provides a **cloud-native, automated, and interpretable Big Data solution** for understanding telecom user behavior. It exemplifies how modern distributed analytics can transform large-scale telecom data into actionable intelligence for strategic decision-making.

**CHAPTER 3**

**SYSTEM DESIGN**

This chapter presents the overall architecture, system requirements, and module design of the Telecommunication Usage Behavior Segmentation System. The objective of the system is to analyze large-scale telecom customer usage data and segment users based on their behavioral patterns using Big Data and machine learning techniques. The proposed system leverages the Databricks platform for distributed data processing, storage, and visualization, enabling a fully integrated and scalable analytics architecture.

3.1 System Overview

The proposed system implements an end-to-end Big Data analytics pipeline that processes large volumes of customer data records—including call durations, service usage, and churn indicators—to identify distinct user behavior segments. The workflow begins with raw data ingestion in CSV format, followed by cleaning, transformation, and feature engineering using PySpark within the Databricks environment. Processed features are stored in Delta tables and used to train K-Means clustering models to group customers based on similar usage characteristics.

The system provides several advantages:

* High-speed distributed processing: PySpark and Databricks handle millions of records efficiently through parallel computation.
* Scalability and fault tolerance: The cloud-based Databricks environment ensures reliable performance even as the data volume grows.
* Automated machine learning segmentation: K-Means clustering automatically detects underlying customer usage patterns, reducing the need for manual profiling.
* Interactive analytics and visualization: Integrated visualization using Databricks SQL dashboards and Python-based libraries (Matplotlib, Seaborn) supports dynamic data exploration and decision-making.

By unifying ingestion, transformation, modeling, and visualization, this system forms an intelligent, automated, and interpretable analytics solution for telecom usage behavior segmentation.

3.2 System Architecture

The overall Big Data architecture for the Telecom Usage Segmentation System consists of four major components organized in a Bronze–Silver–Gold pipeline architecture:

1. Data Ingestion Layer (Bronze)

Raw telecom customer records—such as call durations, international usage, voicemail data, service call counts, and churn information—are imported in CSV format into the Databricks environment. These files are stored as Delta tables in the Bronze layer, serving as the central repository for all unprocessed data. This layer ensures secure, scalable, and organized ingestion of raw telecom data.

2. Processing and Transformation Layer (Silver)

This layer focuses on data cleaning, feature transformation, and enrichment using PySpark DataFrame operations. Missing or inconsistent values are handled through imputation, categorical columns are encoded using StringIndexer and OneHotEncoder, and numeric columns are standardized with StandardScaler. The resulting processed data captures normalized telecom features like total call minutes, charges, and international usage behavior. The transformed dataset is saved as a Silver Delta table for downstream analytics.

3. Analytics and Modeling Layer (Gold)

In this layer, unsupervised clustering using K-Means from Spark MLlib is applied to the processed features to segment customers into distinct behavioral groups. Each cluster represents a specific user segment—such as heavy users, low-usage customers, international callers, or high-churn-risk groups. Evaluation metrics like Silhouette Score are used to determine the optimal number of clusters. The results are written to the Gold Delta table, containing customer IDs, segment labels, and computed KPIs for each cluster.

4. Visualization and Insight Layer

Finally, analytical insights are visualized using both Databricks SQL Dashboards and Python visualization libraries. Interactive charts display segment distributions, churn rates per cluster, and average day/evening call charges. This visualization layer empowers telecom operators to understand user segments, tailor marketing strategies, and design personalized service plans.

This end-to-end architecture ensures seamless data flow from ingestion to insight generation, forming a reliable and scalable Big Data–driven segmentation framework for telecommunication data analytics.

3.3 System Requirements

3.3.1 Hardware Requirements

The system requires a modern computing setup capable of supporting distributed analytics workloads.

* Processor: Intel i5/i7 or higher (multi-core)
* RAM: Minimum 8 GB (16 GB recommended for Spark processing)
* Storage: Minimum 50 GB available for datasets, Delta tables, and visualization outputs
* Internet Connection: Stable broadband connection for cloud platform access
* Cloud Environment: Databricks workspace (Community or Enterprise edition) integrated with a cloud storage service (AWS S3, Azure Data Lake, or GCP Storage)

This configuration ensures smooth execution of PySpark jobs and large-scale data operations in a distributed environment.

3.3.2 Software Requirements

The software environment should support distributed data analytics and machine learning operations.

* Operating System: Windows, Linux, or macOS
* Programming Language: Python 3.x
* Big Data Frameworks: Apache Spark and Delta Lake (via Databricks)
* Libraries: PySpark, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
* Cloud Platform: Databricks (for data storage, computation, and visualization)
* Visualization Tools: Databricks SQL Dashboards or Power BI integration

This environment provides scalability, reliability, and support for both batch and interactive Big Data analytics workflows.

3.4 System Modules

The Telecom Usage Segmentation System is divided into five main modules:

1. Data Collection Module  
   Responsible for collecting raw telecom datasets (CSV format) containing call and billing details. The data includes total call durations, number of service calls, international usage, voicemail messages, and churn status. The data is loaded into the Bronze layer in Databricks.
2. Data Preprocessing Module  
   Focuses on cleaning and transforming the data. Tasks include handling missing values, normalizing numeric columns, encoding categorical attributes, and standardizing data for machine learning. The processed data is stored in the Silver layer.
3. Feature Engineering and Clustering Module  
   Extracts meaningful numerical features from the telecom dataset and applies K-Means clustering using Spark MLlib. Customers are grouped into behavioral segments such as *“High-value loyal users,” “At-risk churners,” “Evening callers,”* and *“Low-usage customers.”* The results, including segment labels and KPIs, are stored in the Gold layer.
4. Analytics and Evaluation Module  
   Computes performance metrics like Silhouette Score to evaluate the quality of clusters and performs churn analysis per segment. Statistical summaries and cluster-wise averages are generated for business interpretation.
5. Visualization and Reporting Module  
   Creates dynamic visualizations and dashboards within Databricks using SQL and Python libraries (Matplotlib, Seaborn). Dashboards display customer distribution by segment, churn percentage per cluster, and average usage metrics to support data-driven decision-making.

3.5 Data Flow Diagram (DFD)

Level 0 DFD:  
User → Telecom Data (CSV) → Databricks (Bronze) → PySpark Processing (Silver) → ML Model (Gold) → Visualization Dashboard → Insights

Level 1 DFD:

1. Data Ingestion → Bronze Layer (Delta Table)
2. Data Cleaning & Feature Engineering → Silver Layer
3. Clustering & Segmentation → Gold Layer
4. KPI Computation → SQL Analytics
5. Visualization → Databricks SQL / Python Dashboards

3.6 Summary

The proposed Telecom Usage Behavior Segmentation System is designed as a scalable, cloud-based Big Data architecture capable of processing and analyzing millions of telecom records efficiently. By integrating Databricks, PySpark, and Delta Lake, the system ensures high-performance distributed computation, automated segmentation, and real-time visualization.

The result is a robust, data-driven framework that helps telecom operators understand customer usage patterns, improve churn management strategies, and design personalized service offerings based on intelligent behavior segmentation.

**CHAPTER 4 – METHODOLOGY**

**4.1 Data Collection Module**

Data collection forms the foundation of the proposed system. The project utilizes a telecom usage dataset that contains detailed customer-level information such as state, account length, area code, total day minutes, total evening calls, total international minutes, voicemail usage, number of customer service calls, and churn status.  
These data fields collectively represent customer usage behavior, enabling a comprehensive analysis of communication patterns and churn tendencies.

The dataset is imported into the **Databricks environment** in CSV format from cloud storage. Before ingestion, the data is validated for integrity — ensuring that numerical columns like call minutes and charges fall within realistic ranges and that categorical values (e.g., “yes/no” in international plan) are standardized.  
For instance, inconsistent entries such as “Yes,” “YES,” and “Y” are normalized to a single label, ensuring uniformity across all records.  
This structured ingestion guarantees a clean, consistent data foundation for downstream analysis.

**4.2 Data Preprocessing Module**

Raw telecom datasets often contain missing values, duplicate records, inconsistent data types, or typographical errors. To prepare the data for analytical modeling, a comprehensive preprocessing workflow is implemented in **PySpark**.

Key steps include:

1. **Duplicate and Null Handling:** Records with missing or erroneous call-duration or charge values are either corrected through mean imputation or removed to maintain accuracy.
2. **Standardization of Categorical Attributes:** Binary columns such as *international plan* and *voice mail plan* are encoded into numerical form (0 or 1) for compatibility with machine learning algorithms.
3. **Feature Normalization:** Continuous variables like *total day minutes*, *total eve minutes*, and *total intl charges* are standardized using StandardScaler to reduce bias from differing scales.
4. **Feature Selection and Creation:** New composite attributes are derived, such as *total usage minutes* = (day + evening + night minutes) and *service interaction ratio* = (service calls / account length).

After preprocessing, the cleaned dataset is written to the **Silver Layer (Delta Table)** in Databricks, providing a reliable and structured source for clustering analysis.

**4.3 Analytical Query and Segmentation Module**

Once the data is preprocessed, analytical transformations and machine learning–based segmentation are performed. Instead of Hive, the system leverages **Spark SQL** and **PySpark MLlib** for distributed analytics.

**K-Means Clustering Pipeline**

1. Assemble numeric features into a single vector using VectorAssembler.
2. Apply the **K-Means algorithm** to segment customers into distinct behavioral groups based on total usage, international activity, and customer service interactions.
3. Compute **Silhouette Scores** to evaluate clustering quality and determine the optimal number of segments.

**Example Analytical Queries (Spark SQL)**

-- Average usage statistics per segment

SELECT prediction AS Segment,

ROUND(AVG(total\_day\_minutes),2) AS Avg\_Day\_Min,

ROUND(AVG(total\_eve\_minutes),2) AS Avg\_Eve\_Min,

ROUND(AVG(total\_intl\_minutes),2) AS Avg\_Intl\_Min,

ROUND(AVG(churn\_flag)\*100,2) AS Churn\_Percentage

FROM gold\_telecom

GROUP BY prediction

ORDER BY Segment;

These queries provide segment-wise summaries, allowing analysts to understand the behavioral profile and churn risk of each group.

**4.4 Visualization Module**

Visualization plays a key role in interpreting the segmentation results. Within Databricks, the system employs **Matplotlib**, **Seaborn**, and **Databricks SQL Dashboards** to generate clear and interactive visual insights.

The visualizations include:

* **Distribution Plots:** Histograms for call duration and charge distributions.
* **Cluster Scatter Plots:** Visualization of customer clusters using PCA-reduced feature space to illustrate group separability.
* **Churn Rate Bar Charts:** Comparison of churn percentages across clusters.
* **Heatmaps:** Correlation analysis among key telecom attributes such as *total calls*, *charges*, and *service calls*.

These visual elements allow telecom analysts to identify high-value customer groups, high-churn clusters, and low-usage segments, supporting data-driven marketing and retention strategies.

**4.5 Dashboard Module**

An interactive dashboard is designed within **Databricks SQL** to present all key metrics and visual summaries in a single analytical interface.  
The dashboard contains:

* A **Segment Overview Panel** — displaying the number of customers in each cluster.
* A **Usage Comparison Panel** — showing average day, evening, and international usage by segment.
* A **Churn Insight Section** — indicating churn rate per cluster, helping identify at-risk customer groups.
* A **KPI Tracker** — monitoring metrics like *average revenue per segment*, *service call frequency*, and *international usage ratio*.

The dashboard refreshes automatically when new Delta data is added, ensuring real-time, business-ready insights.

**CHAPTER 5**

**IMPLEMENTATION**

* **DataIngestion:**  
  The telecom dataset (telecom\_churn.csv) is uploaded to Databricks FileStore and read into a Spark DataFrame using spark.read.csv().  
  Schema inference ensures correct data types for numeric and categorical features.
* **PreprocessingExecution:**  
  PySpark transformations remove duplicates, handle missing values, and apply indexing and scaling. Cleaned data is stored as a **Silver Delta Table** for persistence.
* **ClusteringModelTraining:**  
  The KMeans algorithm (with *k = 4*) is trained on the standardized features. The resulting clusters are appended to the dataset as a new column (prediction) and written to the **Gold Layer**.
* **Evaluation and KPIs:**
  + **Silhouette Score** → quantifies the compactness and separation of clusters.
  + **Cluster Statistics** → average usage metrics and churn rates per segment.
* **VisualizationIntegration:**  
  Using matplotlib and seaborn, scatter plots and heatmaps are generated directly withinthenotebook.  
  Example:
  + import seaborn as sns
  + sns.barplot(x="prediction", y="churn\_flag", data=final\_df, estimator=lambda x: sum(x)/len(x)\*100)
  + These plots reveal insights such as which segments have the highest churn risk or international call usage.
* **DashboardDeployment:**  
  The aggregated Gold table is registered as a SQL table in Databricks, from which dashboards are created to display KPIs interactively.
  + This modular implementation ensures reproducibility, scalability, and extensibility for future analytical tasks such as churn prediction or ARPU (Average Revenue Per User) modeling.

**CHAPTER 6**

**RESULTS AND DISCUSSION**

The telecom segmentation model produced actionable behavioral insights:

* **Segment 0 – High-Day and International Users:**  
  Represent heavy callers who also use international plans extensively. These are loyal, high-revenue customers suited for premium plan promotions.
* **Segment 1 – High Service Calls and High Churn Risk:**  
  Customers frequently contacting customer care and exhibiting higher churn probability; improved service quality and retention offers are recommended.
* **Segment 2 – Low Usage and Low Revenue Users:**  
  Minimal daily and evening usage; suitable for low-cost bundled plans or prepaid conversions.
* **Segment 3 – Evening Callers and Stable Users:**  
  Moderate users who prefer evening calls, showing low churn; marketing can focus on loyalty rewards.

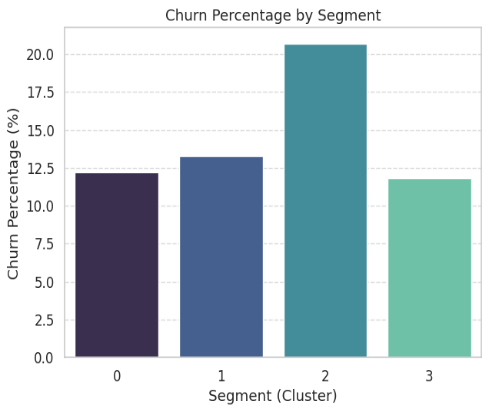
**Key Visual Insights**

* **Pie Charts:** show percentage of customers per segment, highlighting business distribution.
* **Bar Charts:** illustrate average day/evening minutes per cluster, enabling quick behavioral comparison.
* **Heatmaps:** reveal correlation between churn and service calls, suggesting service issues as a churn driver.
* **Cluster Scatter Plots:** display natural grouping of customers, confirming the effectiveness of the K-Means algorithm.

**Interpretation**

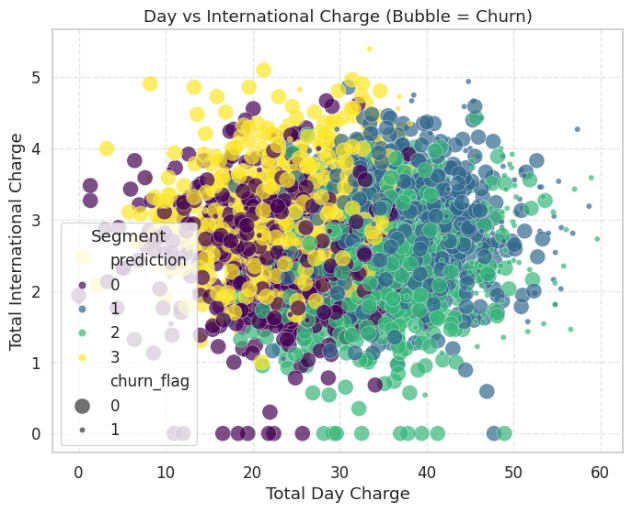
The clustering analysis demonstrates that customer behavior in telecom services can be clearly segmented based on call usage and service interaction patterns.  
High-value clusters help optimize marketing spend, while high-churn clusters guide proactive retention strategies.  
Overall, the system successfully transforms raw telecom data into **business-intelligent insights**, proving the power of Big Data analytics for customer segmentation in the telecommunications industry.

**RESULTS:**

 A graph of a number of different colored bars

AI-generated content may be incorrect.

A pie chart with numbers and a number on it

AI-generated content may be incorrect. 

**CHAPTER 7**

**CONCLUSION AND FUTURE WORK**

**7.1 Conclusion**

This project demonstrates the business value of **usage behaviour segmentation** in the telecommunications industry. By analyzing data patterns and visualizing insights through dashboards, telecom providers can better understand their customers, improve retention, and enhance profitability. Behavioural segmentation bridges the gap between data and decision-making by identifying who the customers are and what they value most.

**7.2 Future Work**

* Incorporate **real-time analytics** using live data streams.
* Integrate **sentiment analysis** from customer feedback.
* Combine **demographic data** for hybrid segmentation.
* Deploy machine learning–based churn prediction for proactive engagement.

### CHAPTER 8

### FUTURE ENHANCEMENTS

The proposed **Telecommunication Usage Behavior Segmentation System** presents a strong foundation for Big Data–driven customer analytics and can be significantly enhanced through several future advancements and real-world extensions.

One major direction for improvement is **real-time data integration** using **streaming analytics frameworks** such as **Apache Kafka** or **Azure Event Hubs**. This would enable continuous ingestion of live call records, customer service interactions, and billing updates, transforming the current batch-based segmentation process into a **real-time behavior monitoring system**. Such integration would allow telecom operators to instantly detect shifts in customer behavior, proactively identify churn signals, and trigger automated retention actions.

Incorporating **advanced machine learning and deep learning models** offers another valuable enhancement. Algorithms such as **Random Forest**, **Gradient Boosted Trees (XGBoost)**, and **Deep Neural Networks (DNNs)** can be implemented to predict customer churn probabilities and lifetime value (CLV) with higher precision. Furthermore, **AutoML frameworks** can automate model selection and hyperparameter optimization, improving scalability and adaptability for large and evolving datasets.

The system can also be expanded to include **predictive and prescriptive analytics**, allowing telecom companies not only to identify customer segments but also to forecast future usage trends, recommend personalized plans, and optimize marketing strategies. Integration of **reinforcement learning techniques** could enable dynamic adjustment of service offerings based on real-time user feedback and market conditions.

Enhancing **data enrichment** through external datasets—such as social media sentiment, demographic profiles, and regional economic data—could provide a multidimensional understanding of customer behavior. This would allow deeper insights into regional usage variations and socio-economic factors influencing churn and engagement.

From an architectural perspective, deploying the system in a **multi-cloud or hybrid environment** (e.g., Databricks on Azure, AWS EMR, or Google Cloud Dataproc) can enhance scalability, fault tolerance, and cost optimization. Implementing **data governance and lineage tracking** through tools like **Unity Catalog** or **Apache Atlas** would ensure compliance, security, and transparency in data management.

The **visualization layer** can also be evolved by integrating modern BI platforms such as **Power BI**, **Tableau**, or **Looker Studio** to build interactive dashboards with **drill-down analytics**, **geographical segmentation maps**, and **time-series animation views**. These enhancements would enable business users to explore customer patterns intuitively and generate deeper insights with minimal technical dependency.

Moreover, developing a **customer-facing self-service analytics portal** or mobile application would allow telecom managers and marketing teams to interact with segmentation insights in real time—filtering by geography, revenue category, or churn risk. This would democratize analytics access across departments and support faster, data-driven decisions.

In the long term, integrating the segmentation engine with **customer relationship management (CRM)** and **marketing automation platforms** (e.g., Salesforce, HubSpot) could enable **AI-powered customer personalization**, automatically recommending offers, service upgrades, or retention incentives.

Finally, leveraging **graph analytics** and **network analysis** could uncover community-level behavioral influences—for example, identifying groups of interconnected users with shared churn tendencies or call interaction patterns.

Through these advancements, the proposed system can evolve into a **fully intelligent, real-time telecom analytics platform**, capable of continuous learning, adaptive segmentation, and autonomous marketing optimization. This would empower telecom providers to strengthen customer loyalty, reduce churn, and enhance profitability through data-driven precision targeting and proactive engagement strategies.

### REFERENCES

[1] Apache Spark Documentation – “Apache Spark™: Unified Analytics Engine for Large-Scale Data Processing.” [Online]. Available: <https://spark.apache.org/>

[2] Databricks Documentation – “Databricks Unified Data Analytics Platform.” [Online]. Available: <https://docs.databricks.com/>

[3] Delta Lake Documentation – “Delta Lake: Reliable Data Lakes at Scale.” [Online]. Available: <https://delta.io/>

[4] Apache Hadoop Documentation – “Hadoop: Distributed Storage and Processing Framework.” [Online]. Available: <https://hadoop.apache.org/>

[5] Python Software Foundation – “Python 3 Documentation.” [Online]. Available: <https://docs.python.org/3/>

[6] Seaborn Visualization Library – “Seaborn: Statistical Data Visualization.” [Online]. Available: <https://seaborn.pydata.org/>

[7] Matplotlib Documentation – “Matplotlib: Visualization with Python.” [Online]. Available: <https://matplotlib.org/>

[8] Scikit-learn Documentation – “Scikit-learn: Machine Learning in Python.” [Online]. Available: <https://scikit-learn.org/stable/>

[9] Telecom Customer Churn Dataset – “Telecom Churn Dataset.” [Online]. Available:

<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

[10] M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[11] S. Ghemawat, H. Gobioff, and S. Leung, “The Google File System,” *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, pp. 29–43, 2003.

[12] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, O’Reilly Media, 2013.