Medical Appointment No-Show Analysis Using Hadoop/MapReduce

Ze Li, 24203409

Abstract—This project implements a comprehensive medical appointment behavior analysis system using Hadoop/MapReduce to identify patterns and risk factors associated with patient no-shows. The solution combines distributed data processing with interactive visualization to deliver actionable insights for healthcare optimization.

Dataset [1]: https://www.kaggle.com/datasets/joniarroba/noshowappointments
Source code: https://github.com/zeli8888/COMP47780-CLOUD-COMPUTING-PROJECT.git
Online dashboard: https://zeli8888-cc-medical-dashboard.streamlit.app/

Index Terms—Hadoop, MapReduce, Medical Analytics, Cloud Computing, Data Analysis

1 Application Overview

This project implements a comprehensive medical appointment behavior analysis system using Hadoop/MapReduce framework. The system is designed to process large-scale medical appointment data and identify critical patterns and risk factors associated with patient no-shows in healthcare systems. By leveraging cloud computing capabilities, the solution enables healthcare providers to optimize appointment scheduling, improve resource allocation, and enhance patient communication strategies through data-driven insights.

The architecture integrates Hadoop/MapReduce for batch processing with an interactive Streamlit dashboard for result visualization. The system successfully processes 110,527 medical appointment records, providing multi-dimensional analysis across six key factors that influence appointment adherence.

2 Objectives

2.1 Primary Objectives

- **Large-scale Data Processing**: *Completed* - Successfully implemented Hadoop/MapReduce to process 110,527 medical appointment records using efficient cloud processing with capability for processing even larger datasets
- **Multi-dimensional Analysis**: *Completed* - Analyzed no-show patterns across 6 key dimensions: gender, age, health conditions, neighborhood, SMS interventions, lead time
- **Risk Factor Identification**: *Completed* - Identified high-risk patient segments and environmental factors contributing to no-shows through comprehensive data analysis
- **Interactive Visualization**: *Completed* - Developed a comprehensive Streamlit dashboard for exploring analysis results and generating actionable insights
- **Cloud Deployment**: *Completed* - Deployed solution on AWS EMR demonstrating scalability and cloud-native implementation capabilities

2.2 Secondary Objectives

- **Data Quality Management**: *Completed* - Implemented robust data validation and data cleaning procedures in MapReduce jobs
- **Performance Optimization**: *Completed* - Utilized combiner classes and efficient data structures for optimal processing performance
- **Cross-platform Compatibility**: *Completed* - Ensured solution functionality across both local Hadoop clusters using Docker and cloud environments with AWS

3 Problem Description and Dataset Collection

3.1 Problem Significance

Patient no-shows represent a significant challenge in healthcare systems worldwide, leading to substantial operational and financial impacts:

- **Resource Wastage**: Unused medical staff time and equipment capacity resulting in inefficient resource utilization
- **Financial Losses**: Estimated annual losses of $150 billion in the US healthcare system due to missed appointments [2]
- **Reduced Access**: Missed appointments that could be allocated to other patients, decreasing overall healthcare accessibility
- **Health Inequality**: Disproportionate impact on underserved communities and vulnerable populations

3.2 Dataset Characteristics

The analysis utilizes the Kaggle Medical Appointment No-Shows dataset with the following characteristics:

- **Source**: Kaggle Medical Appointment No Shows Dataset [1]
- **Volume**: 110,527 medical appointments
- **Time Period**: Records collected between April 29 and June 8, 2016
- **Key Variables**:
  - Patient demographics (Gender, Age, Neighborhood)
  - Appointment details (ScheduledDay, AppointmentDay)
  - Health conditions (Hypertension, Diabetes, Alcoholism, Handicap)
  - Intervention data (SMS reminders)
  - Target variable (No-show status)

4 Methodology and Implementation

4.1 Technical Architecture

The system employs a multi-layer architecture designed for scalability and performance:

- **Storage Layer**: HDFS for distributed data storage and management
- **Data Processing Layer**: Hadoop/MapReduce with Java-based implementation
- **Visualization Layer**: Streamlit dashboard with interactive Plotly visualizations

4.2 MapReduce Implementation

The core analysis is implemented through specialized MapReduce components:

4.2.1 PatientDemographicsMapper

Key features of the mapper implementation:

- **Multi-dimensional Analysis**: Analyzes patient attendance patterns across gender, age, neighborhood, and health conditions
- **Appointment Lead Time Analysis**: Calculates and categorizes scheduling lead time to study its impact on attendance rates
- **Health Status Assessment**: Examines attendance patterns for patients with single diseases and multiple comorbidities
- **SMS Reminder Effectiveness**: Evaluates the impact of SMS reminders on patient attendance rates
- **Data Quality Control**: Includes age validation, data completeness checks, and error record tracking

4.2.2 PatientDemographicsReducer

Key features of the reducer implementation:

- **Data Aggregation**: Performs count aggregation for key-value pairs across all analysis dimensions
- **Statistical Summary**: Computes total counts for attended and missed appointments in each analysis category
- **Result Output**: Generates final summarized datasets for further analysis and reporting

4.3 Analysis Dimensions

The system performs simultaneous analysis across six critical dimensions:

1) **Gender Analysis**: Comparative no-show rates between male and female patients
2) **Age Group Analysis**: Categorization into children/youth, young adults, middle-aged, and seniors
3) **Health Conditions**: Impact analysis of hypertension, diabetes, alcoholism, and handicaps
4) **Geographical Analysis**: Neighborhood-level pattern identification and risk mapping
5) **SMS Intervention**: Effectiveness assessment of reminder message systems
6) **Lead Time Analysis**: Relationship between scheduling delay and appointment adherence

5 Suitability of Hadoop/MapReduce

Hadoop/MapReduce demonstrated exceptional suitability for medical data analysis requirements:

- **Scalability**: Efficiently processed 110K+ records with demonstrated potential for million-record datasets through linear scalability
- **Fault Tolerance**: Built-in data replication and recovery mechanisms ensured data reliability during processing
- **Cost Effectiveness**: Cloud deployment on AWS EMR provided pay-per-use economic model with optimal resource utilization
- **Parallel Processing**: Enabled simultaneous multi-dimensional analysis through cloud computing capabilities

6 Software Features

6.1 Core Analytical Features

6.1.1 Comprehensive Patient Profiling

- Demographic analysis across age groups and gender distributions
- Geographical distribution pattern identification and mapping
- Health condition correlation assessment with no-show probabilities

6.1.2 Operational Intelligence

- Neighborhood-level risk mapping and geographic hot-spot detection
- Appointment lead time impact quantification and optimization insights
- SMS reminder effectiveness evaluation and intervention assessment
- Resource allocation recommendations based on risk patterns

6.2 Technical Features

6.2.1 Data Quality Management

- Data completeness and consistency verification
- Automated data cleaning procedures
- Solid error handling and exception management

6.2.2 Visualization Capabilities

- Interactive dashboard with intuitive navigation and exploration
- Multiple chart types including bar, pie, line, and gauge visualizations
- Comparative analysis tools for multi-dimensional insights

6.2.3 Deployment Flexibility

- Local Hadoop cluster compatibility for on-premises deployment
- Docker containerization support for environment consistency
- AWS EMR cloud deployment for scalable cloud infrastructure

7 Worked Example

7.1 Data Processing Pipeline

The system processes individual patient records through a comprehensive analytical pipeline. Detailed demo example can be found at readme.md.

1) **Input Validation**: Raw appointment data undergoes validation and parsing

```
if (gender.isEmpty() ||
↪  ageStr.isEmpty() ||
↪  noShow.isEmpty()) {
    return;}

int age;
try {
    age =
    ↪  Integer.parseInt(ageStr);
    if (age < 0 || age > 120)
        return; // Filter invalid
        ↪  ages
} catch (NumberFormatException e)
↪  {return;}

boolean valid = analyzeScheduling
↪  LeadTime(context,
↪  scheduledDay, appointmentDay,
↪  attendanceStatus);
if (!valid) {return;}
```

2) **Multi-dimensional Mapper Analysis**: Simultaneous evaluation across all analytical dimensions

```
analyzeSchedulingLeadTime(context
↪  , scheduledDay,
↪  appointmentDay,
↪  attendanceStatus);

analyzeByGender(context, gender,
↪  attendanceStatus);

analyzeByAge(context, age,
↪  attendanceStatus);

analyzeByHealthCondition(context,
↪  hypertension, diabetes,
↪  alcoholism, handicap,
↪  attendanceStatus);

analyzeByNeighbourhood(context,
↪  neighbourhood,
↪  attendanceStatus);

analyzeSMSReminder(context,
↪  smsReceived,
↪  attendanceStatus);
```

3) **Simple Reducer Aggregation**: Simple aggregation for structured output

```
public void reduce(Text key,
↪  Iterable<IntWritable> values,
↪  Context context)
        throws IOException,
        ↪  InterruptedException {
    int sum = 0;
    for (IntWritable val :
    ↪  values) {
        sum += val.get();
    }
    result.set(sum);
    context.write(key, result);
}
```

4) **Result Generation**: Structured output for dashboard visualization

```
GENDER_F_Attended        57245
GENDER_F_NoShow          14591
GENDER_M_Attended        30962
GENDER_M_NoShow          7723
...
```

7.2 Key Analytical Findings

7.2.1 Demographic Patterns

- **Young adults (19-35 years)**: Highest no-show rate at 23.8%, indicating need for targeted engagement strategies
- **Senior patients (55+ years)**: Lowest no-show rate at 15.6%, demonstrating better appointment adherence
- **Gender analysis**: Comparable rates between male (20.0%) and female (20.3%) patients

7.2.2 Scheduling Impact Analysis

- **Same-day appointments**: Achieved lowest no-show rate at 4.6%, supporting urgent care scheduling models
- **Extended lead times**: Appointments between 30 and 90 days showed 33.2% no-show rate
- **Progressive impact**: Strong inverse relationship between scheduling delay and appointment adherence

7.2.3 Intervention Effectiveness

- **SMS recipients**: Unexpectedly higher no-show rates (27.6%) suggesting potential issues with message

content or timing. Higher rates may also be due to the fact that patients with longer lead times tend to receive SMS.
- **Non-recipients**: Lower no-show rates (16.7%) indicating the need for communication strategy optimization
- **Key insight**: SMS intervention requires content personalization and timing optimization

7.3 Dashboard Interaction Example

Healthcare administrators can interact with the system through multiple workflows:

- **Dimension Selection**: Choose specific analytical dimensions through sidebar navigation
- **Data Analysis**: Gain insights from visualized charts and detailed aggregated data through interactive tables

8 Conclusion

8.1 Project Achievements

This project successfully demonstrates the practical application of Hadoop/MapReduce for medical data analysis through several key achievements:

- **Technical Implementation**: Developed robust MapReduce jobs capable of processing real healthcare data with comprehensive validation and error handling
- **Analytical Depth**: Identified significant patterns and correlations across multiple clinical and operational dimensions
- **Operational Relevance**: Generated actionable insights for healthcare service optimization and resource allocation decisions
- **Architectural Scalability**: Created cloud-ready solution with proven deployment methodology and linear scaling characteristics

8.2 Healthcare Implications

The analysis revealed several critical insights for healthcare operations:

- SMS communication strategies require optimization in content, timing, and personalization
- Young adult populations need focused engagement approaches and tailored communication
- Reduced scheduling lead times significantly improve appointment adherence rates
- Geographical variations indicate need for localized strategies and community-specific interventions

8.3 Future Enhancements

Several opportunities for extension and improvement were identified:

- **Real-time Processing**: Integration with streaming data for immediate insights and interventions
- **Predictive Modeling**: Machine learning integration for individual no-show risk prediction
- **Expanded Data Sources**: Incorporation of socioeconomic, environmental, and behavioral factors
- **Advanced Visualization**: Geospatial mapping and temporal trend forecasting capabilities

The project establishes a solid foundation for data-driven healthcare optimization using cloud computing technologies, demonstrating significant potential for improving operational efficiency and patient care delivery through scalable analytical capabilities.

References

[1] Kaggle, "Medical Appointment No Shows Dataset," 2016. [Online]. Available: https://www.kaggle.com/joniarroba/noshowappointments
[2] A. M. Chen, "Socioeconomic and demographic factors predictive of missed appointments in outpatient radiation oncology," Frontiers in Health Services, 2023.
[3] Amazon Web Services, "Amazon EMR Management Guide," 2025. [Online]. Available: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html
[4] Streamlit, "Streamlit Documentation," 2025. [Online]. Available: https://docs.streamlit.io/