**Health Monitoring System Report**

**1. Introduction**

The Health Monitoring System is designed to analyze health data for 10,000 patients tested in a diagnostic center. With the increasing need for efficient health data management, this system provides a structured approach to collecting, processing, and analyzing patient health information using big data technologies such as Apache Spark. The system aims to automate health monitoring and provide meaningful insights into patient health trends.

**2. Objectives**

The primary objectives of the Health Monitoring System are:

* **Data Generation**: Create a dataset of 10,000 patients with randomly assigned health parameters.
* **Data Processing**: Utilize Apache Spark for scalable and efficient data processing.
* **Statistical Analysis**: Compute meaningful health statistics to detect patterns and anomalies.
* **Visualization**: Display key findings through reports and dashboards for better decision-making.

**3. Data Generation and Structure**

The dataset consists of the following attributes:

* **Patient ID**: A unique identifier assigned to each patient.
* **Name**: Randomly generated patient names.
* **Age**: Random values ranging from 20 to 80.
* **Blood Pressure (BP)**: A systolic/diastolic value between (90-140) / (60-90).
* **Sugar Level**: Random floating value between 70 and 200 mg/dL.
* **Cholesterol**: Random floating value between 150 and 300 mg/dL.
* **Haemoglobin**: Random floating value between 11 and 18 g/dL.

**4. Data Processing Using Apache Spark**

**4.1 Loading and Cleaning Data**

* The generated dataset is stored in a CSV file and then loaded into a Spark DataFrame.
* Data types are adjusted to ensure numerical fields are correctly formatted.

**4.2 Data Analysis Using Spark SQL**

* SQL queries are executed on the DataFrame to compute average values of Sugar Level, Cholesterol, and Haemoglobin based on patient age groups.
* Results are stored and displayed for further insights.

**5. Statistical Analysis and Insights**

Using Apache Spark, the following statistical insights are derived:

* **Average Sugar Level per Age Group**: Helps in understanding diabetes trends among patients.
* **Average Cholesterol Level per Age Group**: Provides insights into cardiovascular health risks.
* **Average Haemoglobin Level per Age Group**: Assists in diagnosing anemia and related conditions.

**5.1 Key Findings**

* Patients in older age groups tend to have higher cholesterol levels.
* Blood sugar levels fluctuate across different age groups, requiring further investigation.
* Haemoglobin levels show variation, potentially indicating underlying health issues.

**6. Data Visualization and Reporting**

* The computed statistics are displayed using Spark's show() function for tabular insights.
* The results are stored in a processed CSV file for integration with visualization tools such as Power BI or Tableau.

**7. Conclusion and Future Enhancements**

This Health Monitoring System efficiently processes large-scale patient health data, providing valuable insights into health trends. Future enhancements include:

* **Real-Time Monitoring**: Implementing Kafka and streaming analytics for live data processing.
* **Machine Learning Integration**: Using predictive analytics to detect potential health risks.
* **Interactive Dashboards**: Creating visual reports for improved decision-making.

By leveraging big data technologies, this system offers a scalable and efficient solution for health data analysis, paving the way for smarter healthcare insights.

**Submitted By:** Kartikey  
**Course:** BTech in Computer Engineering (Big Data Analytics)  
**Institution:** Lovely Professional University