****

**COURSE: BIG DATA ARCHITECTURE**

**COURSE DESCRIPTION: DSA 3030A**

**LECTRURE: PROF.EDWARD OMBUI**

**MEMBERS: NEEMA NDANU- 666457**

**MITCHELLE KANURI-666565**

**ROSE MASINDE- 666750**

**NATASHA KOKI- 665602**

**REPORT: PREDICTING DIABETES RISK IN POPULATION HEALTH MANAGEMENT**

**INTRODUCTION:**

Given the importance of diabetes as a worldwide health concern, the project's goal was to create an end-to-end Big Data analytic application that uses the Apache Hadoop environment to predict diabetes risk in population health management. Effective management and intervention stressed early detection through risk prediction as a critical approach.

**PROBLEM DOMAIN AND RELEVANCE**

Given the prevalence of diabetes, its influence on global health, and the abundance of data available, our group decided to focus on forecasting diabetes risk. Through the methodical examination of numerous risk factors, such as age, weight, smoking history, blood glucose level, hypertension, and other factors, the application sought to provide medical experts with useful information for early intervention.

**ARCHITECTURAL OVERVIEW**

The application's architecture leverages multiple components of the Apache Hadoop ecosystem, according to a distributed computing paradigm: -

Data Ingestion: Apache Nifi is utilized to smoothly collect health data from various sources and feed it into the Hadoop environment for analysis.

Data Preprocessing: To ensure that the ingested data is formatted correctly for analysis, Apache Pig is used for preprocessing tasks like data cleaning and transformation.

Data Analysis: Using Apache Hadoop, Apache Pig and programming language Python, we can efficiently process and analyze the data to extract insights.

Data Visualization: Using Power BI, we presented our data analysis findings through interactive dashboard and reports, enhancing understanding and interpretation.

Scalability and fault tolerance are prioritized in this architectural design since they are essential for effectively managing large-scale databases.

**FUNCTIONALITY OR USABILITY**

A wide range of features specifically designed for diabetes risk prediction are available in the application: -

Ingestion of Data: Apache Nifi is used to gather a broad range of health information from many sources, such as age, gender, blood pressure, heart disease, smoking history, BMI, HbA1c level, blood sugar level, and diabetes status. The information is then streamed into the Hadoop ecosystem's storage.

Data Preprocessing: To prepare the imported data for analysis, Apache Pig handles missing values, standardizes data formats, and does feature engineering.

Data analysis: Leveraging Apache Pig and Python scripts, Apache Hadoop analyzes preprocessed data to find patterns in diabetes risk variables and estimates the probability of developing diabetes based on feature availability.

Data Visualization: Power Bi facilitates the presentation of analysis findings through interactive visuals, enhancing comprehension and decision.

**IMPLEMENTATION DETAILS**

Project Management: The implementation process is planned into discrete phases, including data ingestion, processing, analysis, visualization, testing and deployment. Team members are assigned tasks based on their skills sets, ensuring effective coordination and timely completion.

**Implementation Technically:**

Apache Nifi as central hub for data collection, providing a unified interface for connecting to various sources. The integration of Apache Nifi into ecosystem entails meticulous setup and configuration procedure outlined in comprehensive guides. These guides meticulously detail the steps required to established Apache Nifi as primary node within the Hadoop Distributed File Systesm (HDFS).

Apache Pig scripts were created to carry out data cleaning and preprocessing operations to prepare the incoming data for analysis. The dataset was cleaned and structured correctly for further analysis thanks to the important procedures like data transformation, filtering, and feature extraction that these scripts took care of. Pig's capabilities were leveraged during the preprocessing phase to speed data preparation and set the foundation for relevant and accurate analysis.

During the data analysis phase, the application leveraged a combination of Apache Pig and custom Python script (UDF) to tackle complex analytics tasks. By invoking Python script within Pig, specialized algorithms were seamlessly integrated into the analysis pipeline. This approach enabled the development and application of tailored algorithms for diabetes risk prediction, capitalizing on Pig’s advanced data processing capabilities. Pig's parallel processing capabilities facilitated scalability, allowing the program to handle the demands of large-scale data analysis with ease. The constructive collaboration of Python Script and Pig empowered the extraction of insightful information crucial for information decision-making in diabetes risk assessment and community health management.

The application's capabilities were enhanced through integration with potent visualization tools like Power BI, which made it possible to create interactive visualizations. Using clear visual aids, stakeholders may efficiently examine and comprehend the data analysis findings, leading to more profound understanding and well-informed choices. The application fully used the examined data's potential by enhancing user engagement and comprehension through visualization.

The project achieved its goals of strong data analysis and efficient visualization in population health management and diabetes risk prediction by carefully planning the implementation process and following best practices.

**LESSON LEARNT**

Effective Big Data Tool Utilization: The research made clear how important it is to choose the appropriate tools for working with large-scale datasets. Efficient processing and analysis were made possible by utilizing Apache Hadoop, Nifi, and Pig, highlighting the significance of matching tools to project objectives.

Difficulties with Data Quality: Missing or inconsistent data discovered during preprocessing brought to light how important it is to use strong data transformation and cleaning methods. Early resolution of these issues highlighted the importance of paying close attention to data quality.

Algorithm Selection: A careful trade-off between computing complexity and accuracy had to be made while selecting appropriate algorithms for diabetes risk prediction. The research showed how crucial it is to conduct extensive analysis and testing to determine which algorithms best meet the assignment's demands.

Joint Development:

Successful teamwork has been identified as a critical component of project success. The smooth completion of duties was made possible by open communication and seamless teamwork, highlighting the importance of creating a collaborative workplace.

Best Practices and Documentation: Throughout the development lifecycle, the project emphasized how important it is to follow best practices. The process, decision, and configuration documentation were crucial to the project's comprehension, upkeep, and expansion. The team ensured uniformity and clarity in project execution by documenting architectural choices and development processes.

Technical Difficulties: Handling a variety of data sources and putting complicated algorithms into practice caused technical difficulties that called for creative solutions. Overcoming these obstacles demonstrated how important it is to be flexible, creative, and open to trying new things.

By combining these insights, the initiative not only met its short-term goals but also cleared the path for ongoing development and success in subsequent undertakings.

**FINAL REMARKS AND COMPREHENSIVE EVALUATION**

The project has shown a strong potential to improve healthcare outcomes through the creation of a Big Data analytic application for diabetes risk prediction in community health management. The application is ready to assist medical professionals in early detection and intervention by utilizing the Apache Hadoop environment.

A thorough grasp of the application's design, functionality, implementation specifics, and lessons learned was made possible by the thorough summary. It demonstrated a thorough understanding of the technical nuances of the project and the difficulties experienced during development.

Through the efficient application of Big Data techniques, resolution of data quality issues, and promotion of cooperative development, the project accomplished its goals and laid the groundwork for further improvements. The project's sustainability and scalability are ensured by the focus on best practices and documentation, which also provides a strong platform for future innovation in population health management.

In summary, the research not only provides diabetes risk prediction in the short term, but it also paves the way for future advancements and improvements in healthcare analytics.