2025

CS 4407
Data Mining & Machine Learning

PROGRAMMING ASSIGNMENT UNIT 2

INSTRUCTOR: NIRMAL ADHIKARI

MINING AND MACHINE LEARNING:

DATABASE AND ANALYTICS TOOLS

INTRODUCTION

Effective data analysis depends on the choice of database systems, statistical tools, and application programming interfaces (APIs). Each component supports a different stage of the analytics pipeline, from data storage to advanced modeling. This paper compares a traditional database, an analytical database, and a NoSQL database, and then examines how MySQL, R, and Hadoop work together to support data-driven decision making.

COMPARING DATABASE TYPES

A traditional database, often a relational database management system (RDBMS), organizes data into tables with predefined schemas. Systems like MySQL enforce structured relationships and use Structured Query Language (SQL) for data management. They are well suited for transactional operations where consistency and reliability are essential (Elmasri & Navathe, 2020).

Analytical databases, by contrast, are optimized for large-scale queries and complex analyses. They often employ columnar storage to speed up aggregations and reduce input/output demands. Systems such as Amazon Redshift or Google BigQuery allow analysts to process massive datasets efficiently, focusing on read-heavy workloads rather than frequent updates (Pavlo et al., 2017).

NoSQL databases break away from fixed schemas and relational tables. They store data as documents, key-value pairs, or graphs, providing flexibility for unstructured or semi-structured data. Examples include MongoDB and Cassandra, which scale horizontally and handle high-velocity data streams common in modern applications. While they sacrifice some consistency guarantees, they excel in handling diverse data types and real-time analytics.

MYSQL: A TRADITIONAL DATABASE

MySQL is a widely used open-source RDBMS known for its stability and performance in transactional systems. It enforces ACID properties (atomicity, consistency, isolation, durability), ensuring that critical operations like banking transactions remain reliable. MySQL supports complex joins and indexing, making it suitable for structured business data. In an analytics context, MySQL often serves as the initial repository where raw data is stored and validated before being moved to analytical or NoSQL systems for further exploration.

R: A STATISTICS PACKAGE

R is a programming language and environment designed for statistical computing and graphics. It offers extensive libraries for data cleaning, visualization, and machine learning. Analysts use R to run statistical tests, build predictive models, and create interactive plots. It complements databases by providing the analytical capabilities needed after data extraction. When paired with MySQL, R can query the database directly, pull relevant datasets, and apply advanced models to uncover trends and patterns.

HADOOP: AN API AND DEVELOPMENT ENVIRONMENT

Hadoop is an open-source framework that enables distributed storage and processing of large datasets across clusters of computers. It uses the Hadoop Distributed File System (HDFS) and the MapReduce programming model to process vast amounts of data efficiently. As an API and development environment, Hadoop allows developers to build scalable analytics applications that can integrate with both R and MySQL. Data can be ingested from MySQL, processed in Hadoop for large-scale computations, and then analyzed in R for statistical insights.

INTEGRATION IN AN ANALYTICS SYSTEM

MySQL, R, and Hadoop illustrate how different tools support the full analytics workflow. MySQL stores and organizes structured data, ensuring data integrity at the source. Hadoop provides the infrastructure to process and transform both structured and unstructured data at scale. R then performs the statistical analysis and modeling required for machine learning tasks. Together, these tools create a robust system where reliable storage, scalable processing, and advanced analytics are interconnected.

CONCLUSION

Traditional, analytical, and NoSQL databases each serve distinct purposes in modern analytics. Traditional databases like MySQL offer reliable structured storage, analytical databases handle large-scale queries, and NoSQL databases provide flexibility for unstructured data. In a complete analytics system, MySQL ensures consistent data management, Hadoop supports distributed processing, and R delivers statistical and machine learning capabilities. Integrating these tools enables organizations to manage data effectively and extract actionable insights.

REFERENCES

Elmasri, R., & Navathe, S. B. (2020). *Fundamentals of database systems* (7th ed.). Pearson.

Pavlo, A., Anguelov, D., Fetterly, D., Luo, L., & Ellner, S. (2017). Self-driving database management systems. *Communications of the ACM, 60*(5), 33–37. https://db.cs.cmu.edu/papers/2017/p42-pavlo-cidr17.pdf

Word count: 618