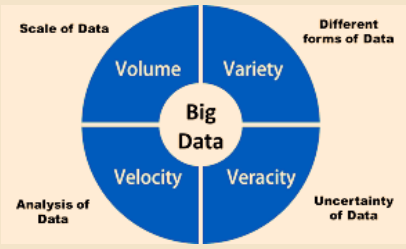
BigData Analytics on Cancer dataset

GITAM
DEEMED TO BE UNIVERSITY

Presenter Names: B. Harsha Vardhan, MD. Naushad Ali GVS Niharika, L Jayanth
Guide Name: Pragyaban Mishra

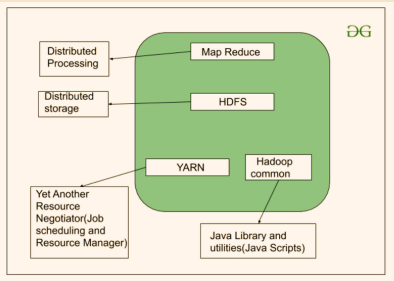
Introduction

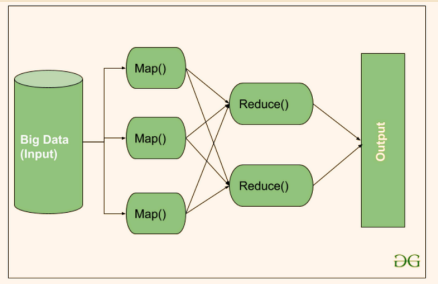
Introduction:
Big Data refers to the vast and continuously growing datasets characterized by the 5V's: Volume, Variety, Velocity, Variability, and Veracity. In healthcare, big data analytics plays a crucial role in storing, processing, and analyzing medical data such as mammographic images. It enables predictive insights and enhances clinical decision-making, leading to early detection and better treatment of diseases like cancer.



Problem Statement:
Traditional data management techniques struggle to handle large and complex healthcare datasets, especially medical imaging data. The need for an efficient big data analytics framework is crucial for processing, analyzing, and predicting cancerous tumors accurately. Our project aims to improve diagnostic efficiency using big data techniques and machine learning models.

Results





Data Storage & Management:
- HDFS is used for distributed cancer data storage
- MapReduce processes and categorizes tumor data

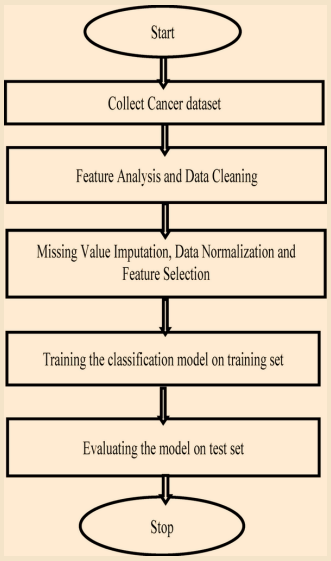
Data Processing Flow:
1. Data Ingestion – Loading cancer dataset into HDFS
2. Mapper Phase – Categorizes tumor characteristics into severity levels
3. Reducer Phase – Aggregates data and calculates survival impact
4. Final Output – Processed data stored back in HDFS for further analysis

Methodologies

Feature-Based Categorization:
- Tumor Size & Shape: Large tumors linked with lower survival rates
- Concavity & Marginal Adhesion: Used for cancer severity classification

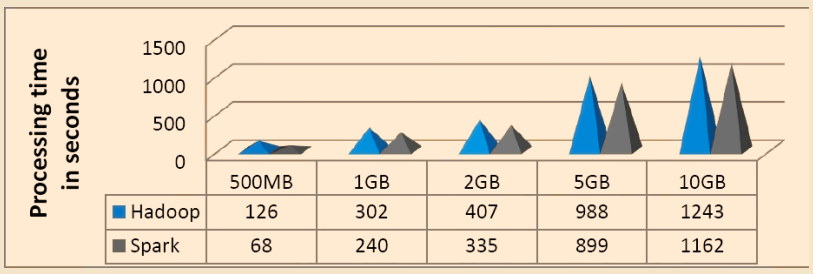
Hadoop-Based Processing:
- MapReduce Implementation in Java
- Survival Rate Calculation based on tumor characteristics

Error Handling:
- Missing values replaced with mean values
- Invalid entries handled during data preprocessing



Conclusions

This project demonstrates how big data analytics can enhance cancer data processing. By leveraging Hadoop and MapReduce, tumor features were analyzed efficiently, providing insights into cancer severity based on numeric attributes. Future enhancements will include integration with real-world hospital databases and improving computational efficiency.



System Features & Analysis

Dataset Used: Wisconsin Breast Cancer Dataset (Numeric Data)
Key Features Analyzed:
1. **Perimeter Mean:** Measures tumor size
2. **Area Mean:** Evaluates tumor spread
3. **Concavity Mean:** Determines tumor shape irregularities
4. **Bare Nuclei:** Counts cell nuclei lacking cytoplasm

Processing Approach:
- Hadoop Distributed File System (HDFS): For large-scale data storage
- MapReduce: For parallel data processing

System Features:
- Hadoop Integration: Scalable big data processing for medical datasets
- Automated Feature Analysis: Identifies key cancer indicators
- Efficient Computation: Parallel processing for fast and reliable results

References

1. K. Shailaja et al., "Applications of Big Data Analytics: A Systematic Review", International Journal of Engineering Research in Computer Science and Engineering, volume 5, 2018.

2. American Cancer Society. Breast Cancer Facts & Figures 2005 2006. Atlanta: American Cancer Society, Inc. http://www.cancer.org/.

3 Ms Shweta Srivastava et al., "A Review Paper on Feature Selection Methodologies and Their Applications", International Journal of Engineering Research and Development, Volume 7, PP. 57-61, 2013.

4. Abdur Rahman Onik et al., "An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier, International Journal of Computer Applications, volume 13, 2015.