Assignment 2
Bigdata Analytics

Date of Submission: 17th September'24

1. Describe different data cleaning techniques and discuss how these can improve the quality of big data.
2. Define missing data and explain how it can affect big data analysis. What are some common methods for handling missing data in large datasets?
3. What are the techniques to identify and handle outliers in big data? Provide examples of when removing outliers may lead to loss of important information.
4. Describe a real-world scenario where missing data significantly influenced the results of a big data analysis. How was the problem resolved?
5. Given a high-dimensional dataset, explain how you would apply feature selection methods to improve the performance of a predictive model.
6. Describe the advantages of using Apache Pig and Apache Hive for scalable data preprocessing in big data environments.
7. Explain the differences between Pig Latin and HiveQL. In what scenarios would you prefer one over the other for data preprocessing tasks?
8. Discuss how Apache Pig and Hive help in performing ETL (Extract, Transform, Load) operations in big data workflows. Provide an example of a data preprocessing task that can be efficiently managed using these tools.