SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

AN AUTONOMOUS INSTITUTION

Question Bank

PART – A:

1. How does Stats models differ from Scikit-learn?
2. Where scikit learn were commonly used?
3. Mention the functions of Job Tracker.
4. State the need for Map and Reduce function.
5. Write the features of Hadoop.
6. What is the primary purpose of Scikit-learn?
7. What is the primary use of the Statsmodels library in Python?
8. Differentiate RDBMS and Hadoop.
9. What is HBase? List the major components of HBase.
10. Draw the architecture of YARN.

PART – B:

1. In a classification task, you observe that your model is over fitting the training data. What techniques from SCIKIT –learn can you use to address this issue.
2. Explain the procedure involved in installation of Stats models.
3. An IT manager is interested in adopting Map Reduce for data processing. Highlight three key features contribute to its effectiveness in large scale data analysis.
4. Illustrate the architecture of YARN and its core components in detail.
5. Describe in detail the processing model of Hadoop.
6. Describe the role and interaction of the three primary HDFS daemons Namenode, Data node and secondary name node in a Hadoop cluster.Provide examples of scenarios where these daemons play a crucial role in maintaindata integrity and availability.
7. Build a sckitlearn model using house prices dataset and use estimators to analyze and predict accuracy using KNN. Also demonstrate the model using cross validation and feature extraction.
8. Enumerate the architecture of HBase with suitable diagram.

PART – C:

1. There is a number of documents where each document is a set of terms. It is required to calculate a total number of occurrences of each term in all documents. Illustrate various stages involved in above scenario and write mapper and reducer code for the same.
2. A data engineer is tasked with explaining HDFS to a team of developers. Describe the components of HDFS and discuss how block replication contribute to fault tolerance in Hadoop.
3. A large manufacturing company wants to implement a predictive maintenance system to minimize unexpected machinery breakdowns and reduce downtime. The company has chosen Scikit-learn for building the predictive model.

i)Discuss the steps the company should follow to implement the predictive maintenance system using Scikit-learn, starting from data collection to model deployment.

ii) Explain how Scikit-learn's algorithms, like Random Forest and Support Vector Machines (SVM), can be applied in this scenario. Discuss their advantages and potential drawbacks.

iii) What metrics should the company use to evaluate the performance of the predictive maintenance model, and why?

4. A database administrator is considering whether to use a traditional RDBMS or Hadoop for a new data – intensive project. Compare and contrast the strengths and weaknesses of each system, considering factors like scalability ,and data types.