Interview Questions

1. What is Apache Spark MLlib?
    a. MLlib is Apache Spark's scalable machine learning library, which runs on a distributed environment using Spark as the processing engine.
2. Which library provides hypothesis testing in PySpark MLlib?
    a. Pearson Correlation
3. What is ML PipeLines?
    a. It is a means to combine the different operations of Spark MLlib, i.e., imputer, transformer and model. The imputer modifies data samples and removes null values. The transformer transforms data points, for example, the TF-IDF vectorizer, which performs TF-IDF vectorization on the data set. A model is a pyspark.mllib model, for example, Logistic Regression.
4. List some algorithms that are present in PySpark MLlib.
    a. mllib.classification
    b. mllib.clustering
    c. mllib.classification
    d. mllib.regression
    e. mllib.recommendation
5. What are the different machine learning tools available in Spark mllib?
    a. Algorithm
    b. Transformation
    c. Pipeline
    d. Persistence
    e. Utilities.
6. Does Spark support SVM with SGD?
    a. Yes, Spark supports SVM with SGD. It is a stochastic gradient descent optimiser that is used to optimise a model for a given data set. It is an iterative method.
7. What are the SGD methods that are supported by mllib?
    a. Logistic Regression with SGD
    b. SVM with SGD
8. What are the distribution generating functions that are supported by PySpark mllib?
    a. Normal distribution, uniform distribution and gamma distribution are some of the distribution generating functions that are supported by PySpark mllib.
9. What is the difference between K-means and bisecting-k means?
    a. K-means clustering clusters around the centroid, that is, it splits the data points from the start into k clusters. On the contrary, the bisecting-k means algorithm splits the data points into sub-clusters.
10. What are the advantages of Spark-mllib?

a. Easy-to0use library
b. Can be used with huge amounts of data points
c. Suitable for working with SQL queries

11. What is Spark-Conf, which is used in spark-mllib?
    a. It is used to set configuration and the parameters while submitting a Spark job. These parameters include variables such as the Spark cluster's IP address, the Spark executor's memory and the number of cores to be used.
12. What is a sparse vector?
    a. A local vector contains both integer-type and 0-based indices. It also contains double-typed values, which are stored on a single machine. In MLlib, two types of local vectors are supported, namely, Dense and Sparse vectors. A sparse vector is one in which most of the entries are zero.
13. When do you perform regression or use a regression-type model in spark-mllib?
    a. Regression is performed when we are predicting a value. For example, consider a scenario wherein you want to predict the number of jumps given and the number of steps a person has to follow. In this case, we will use regression. We will use particular steps as a feature and the number of jumps will be used as output.
14. When do you use StringIndexer?
    a. StringIndexer is used when a label does not have an integer value. Models generally prefer prediction columns to be of integer type rather than string type. It is similar to the LabelEncoder of sklearn.