**1. Can you explain the difference between supervised and unsupervised learning?**

Supervised Learning Technique - as the name suggests, it has the presence of a supervisor. When we teach or train the machine using data that is well labelled.

Example: A basket filled with fruits. All fruits are labelled with respect to the shape, color etc. After training when the machine is asked to identify a fruit. It can easily do it by the previously learned information.

Different types of supervised learning algorithms are, Classification, Regression, Support Vector Machine, KNN, Decision Tree, Random Forest etc.

Unsupervised Learning Technique - Training of a machine using information i.e neither classified nor labelled.

Different types are, K means clustering, Agglomerative, Probabilistic etc.

**2.What is overfitting in machine learning, and how can it be prevented?**

Overfitting means the model fits too much to the noise from the data. This often results in low error on training sets but high error on test or validation sets. When overfit on the training data, it means performs poorly on new test data.

Overfitting can be prevented by using cross validation techniques. It is used to check the performance of a model to solve the issues like overfitting by evaluating the model on multiple subset or divisions of the data. Like K Fold Cross Validation, the data is split into K groups and evaluated.

**3.Explain the bias-variance tradeoff in model training.**

To evaluate the model’s performance, Bias variance trade off is the point where we are adding just noise by adding model complexity or flexibility. Training error goes less as it has to, but the test error starts to goes up. The model after the bias trade off begins to overfit.

**4.What is cross-validation, and why is it important in machine learning?**

Cross Validation - It is used to check the performance of a model to solve the issues like overfitting and underfitting by evaluating the model on multiple subset or divisions of the data. Like K Fold Cross Validation, the data is split into K groups and evaluated. Major steps are partitioning the dataset into a number of subset, holding out a set for evaluation then training the model on the other sets. This gives a more reliable estimate of how the model performs across different training set because it provides an average score across different training samples used.

**5.Describe the concept of feature engineering and its significance in model building.**

Feature Engineering is the process of creating new features or transforming existing features to improve the performance of a machine-learning model. It involves selecting relevant information from raw data and transforming it into a format that can be easily understood by a model. The goal is to improve model accuracy by providing more meaningful and relevant information.

**6.What is PySpark, and how does it relate to Apache Spark?**

PySpark is the Python API for Apache Spark, PySpark allows developers to write Spark applications. Spark supports both batch processing and real time processing. It performs in memory processing that means it stores the data in the RAM of server which allows quick access and in turn accelerates the speed of analytics. DAG (Directed Acyclic Graph) allows for more complex and flexible data.

**7. Explain the advantages of using PySpark for big data processing.**

* Large-Scale Data Processing
* Fault Tolerance
* Integration with Python

**8. Describe the main components of PySpark's architecture.**

* Driver Program - Main application of spark. Entry point of a spark application. It runs the main function and creates Spark context to coordinate the execution of tasks on the cluster. It manages the execution flow of the spark application. Divides the application into tasks and schedules them for execution. Spark context serves as a client to connect to the spark cluster manager.
* Cluster Manager - It is responsible for managing resources across the cluster and scheduling tasks.
* Worker Nodes - Responsible for executing tasks on the cluster. Runs tasks assigned by the driver program.
* Task - It represents units of work that are sent to the executor nodes for execution.It performs computation.

**9.What is an RDD, and how does it differ from traditional data structures in Python?**

RDD - Resilient Distributed Dataset, is a fundamental data structure in Spark. Transformations and Actions are the two fundamental types of operations that can be applied to RDD.

**10.Explain the concept of lineage in RDDs.**

Lineage represents the sequence of transformations that were applied to create an RDD. It is a directed acyclic graph (DAG). DAG is used to represent the logical execution plan of spark application. It represents the sequence of transformation and actions on RDDs. Optimises execution plan for better performance.

**11.How can you create an RDD in PySpark?**

**12. Difference btw Transformation and Actions. Explain with example**

Transformation and actions are the two fundamental types of operations that can be applied to RDD.

Transformations - are operations that transform the data such as filtering, mapping, aggregating etc. They are performed lazily, that means they are not executed immediately when called. Spark keeps track of the transformation operations in a DAG. They are not executed until an action is called. This allows spark to optimize the execution plan.

Actions - are operations that trigger the execution of transformations and return a value. Actions are operations that initiate the computation and return a result to the driver program or write data to an external storage system.

Example: collect, count, save as txt file, reduce

**Example**, we have a dataset and we need to read, map and filter the save as txt file. Then transformation steps are read, map and filter and action is to save as txt file. In Spark jobs, when we check the output, the output will only be shown until we call for action. Just by running the code with transformation lines compiler doesn’t shows the output. Once we run the code with action only we would be able to check the output in spark jobs.