Comprehensive PySpark Interview Questions and Answers

1. What is PySpark and what are its main components?

PySpark is the Python API for Apache Spark, an open-source unified analytics engine for large-scale data processing. Main components include Spark Core (the foundational engine), Spark SQL (for querying structured data), Spark Streaming (for real-time data processing), MLlib (for machine learning), and GraphX (for graph processing).

2. How do you set up a PySpark environment?

Setting up PySpark involves installing Spark, setting up Java (as Spark runs on the JVM), and configuring environment variables like `SPARK_HOME` and `PYSPARK_PYTHON`. You can install PySpark via pip with `pip install pyspark`. Configuration may also involve setting up a cluster manager such as YARN or Kubernetes.

3. What is the role of a SparkSession?

The SparkSession is the entry point for programming Spark with the DataFrame and SQL API. It allows you to create DataFrames, execute SQL queries, and manage Spark jobs.

24. What are some real-world use cases of PySpark?

Real-world use cases include real-time analytics, fraud detection, recommendation systems, large-scale data transformations, and machine learning model training.

25. How do you handle schema evolution in PySpark?

Schema evolution involves handling changes in data structure over time. PySpark can handle schema evolution by using options like `mergeSchema` when reading from sources like Parquet or

Avro.