Spark SQL

# Apache Spark - Revision Notes

## 1. Spark SQL

Spark SQL is a module in Apache Spark for working with structured data. It allows querying data via SQL as well as the DataFrame and Dataset APIs.  
Key Points:  
- Enables querying of structured data using SQL and the Hive Query Language (HQL).  
- Can read data from various sources like JSON, Parquet, Hive, and JDBC.  
- Integrates with Spark’s core APIs and supports complex analytics.  
- Supports data warehousing capabilities and can be used to create temporary views.

## 2. Spark Streaming

Spark Streaming is a component of Spark that enables scalable and fault-tolerant stream processing of live data streams.  
Key Points:  
- Processes real-time data from sources like Kafka, Flume, and TCP sockets.  
- Data is processed in near real-time using mini-batches.  
- Outputs can be pushed to databases, file systems, or dashboards.  
- Useful for use cases like fraud detection, log processing, and real-time analytics.

## 3. Micro-batching

Micro-batching is the core concept behind Spark Streaming.  
Key Points:  
- Incoming data is divided into small batches (micro-batches).  
- Each micro-batch is processed as a small job in Spark.  
- Balances latency and throughput, making it suitable for many streaming applications.  
- Provides strong consistency and fault-tolerance by leveraging Spark’s RDD model.