Top 20 Spark Concepts to Learn

Foundations (Basics – Must Know First)

1. **Spark Architecture** – Driver, Executors, Cluster Manager, Jobs, Stages, Tasks.
2. **RDD (Resilient Distributed Dataset)** – transformations, actions, persistence.
3. **DataFrame API** – schema, columns, operations.
4. **Spark SQL** – querying structured/semi-structured data, SQL vs DataFrame.
5. **Lazy Evaluation** – DAG creation, execution plans.
6. **Spark Configurations & Deployment Modes** – local, standalone, YARN, Kubernetes.

Intermediate (For Real Projects & ETL)

7. **Data Sources & Formats** – Parquet, ORC, JSON, CSV, Avro, Delta.
8. **Joins in Spark** – broadcast joins, shuffle joins, skew handling.
9. **Window Functions** – ranking, aggregations, time-based windows.
10. **Partitioning & Shuffling** – partition strategies, coalesce, repartition, bucketing.
11. **Caching & Persistence** – memory/disk strategies, storage levels.
12. **Spark SQL Functions** – built-in functions, UDFs, UDAFs, Pandas UDFs.
13. **ETL with Spark** – extract (Kafka, DBs, files), transform (cleansing, enrichment), load (DWH, data lake).

Advanced (Optimization & Scalability)

14. **Catalyst Optimizer & Tungsten Engine** – query optimization internals.
15. **Serialization & Kryo** – efficient data handling.
16. **Spark Streaming / Structured Streaming** – micro-batch, continuous, watermarking, event-time processing.
17. **Checkpointing & Fault Tolerance** – recovery in streaming/batch.
18. **Performance Tuning** – Spark UI, job/stage monitoring, skew fixes, partition tuning.
19. **Cluster Resource Management** – executor/driver memory, cores, parallelism.
20. **Integration with Ecosystem** – Kafka, Hive, HDFS, Delta Lake, MLlib, Airflow.

Learning Strategy:

- **Step 1**: Understand Spark architecture + RDDs + DataFrames.
- **Step 2**: Move into SQL, joins, window functions, ETL workflows.

- **Step 3**: Learn performance tuning, streaming, and integration with real-world tools.