Day 1: Spark Architecture, APIs, and Resource Management

1. **Recap of Spark Architecture**
    - Overview of Spark's core components
    - Execution model: Jobs, Stages, and Tasks
2. **Structured APIs and Low-Level APIs**
    - Understanding DataFrames and Datasets
    - Differences between Spark 2 and Spark 3
    - Low-level RDD APIs
3. **Resource Management and Scheduling on EMR**
    - Overview of Amazon EMR and its capabilities
    - Configuring YARN on EMR
    - Using Instance Groups and Instance Fleets for resource management
    - Job scheduling and priority management in EMR
4. **Spark Internals**
    - Catalyst optimizer: Logical and physical plan
    - Rule-based and cost-based optimization
    - Tungsten execution engine: In-memory computing, binary processing, and cache-aware computation

Day 2: Advanced Spark Techniques and Performance Optimization

1. **Advanced DataFrames and Datasets**
    - Optimizing DataFrame transformations
    - Type-safe processing using Datasets
    - Custom aggregations and UDFs/UDAFs in EMR
    - Using Glue Data Catalog with Spark SQL
2. **Language-Specific Spark**
    - Case Classes in Scala
    - Pandas API for Python
    - Running SQL queries on DataFrames & optimizing them
3. **Performance Tuning and Optimization Techniques**
    - Various joins in Spark and their performance impact
    - Adaptive Query Execution (AQE)
    - Best practices for writing efficient Spark code
    - Serialization strategies: Using Kyro
    - Data partitioning, bucketing, and custom partitioning
    - Cache, persist, and checkpoint strategies
4. **Memory Management and Tuning**
    - Spark's memory model
    - Tuning memory parameters and garbage collection
    - Off-heap storage and its benefits

Day 3: Advanced Topics in Streaming, ETL, and Fault Tolerance

1. **Optimized ETL Pipelines with EMR**
    - Building ETL pipelines using Spark on EMR
    - Handling schema evolution with AWS Glue
    - Best practices for data partitioning and bucketing
    - Techniques to prevent shuffling
2. **Fault Tolerance in Spark**
    - Reliable Spark Streaming: WALs, RDDs, and Availability
    - Advanced Spark Streaming: Structured Streaming vs. DStreams
    - Stateful transformations and exactly-once semantics
    - Watermarking and window operations
3. **Efficient Data Storage and Access**
    - Using S3 as a data source and sink in Spark
    - Optimizing S3 access patterns
    - Introduction to Magic Write for efficient data writes to S3
    - Comparing file formats: Parquet, ORC, Avro
4. **Exercise: Monitoring and Debugging Spark Jobs**
    - Monitoring using CloudWatch
    - Debugging with Spark UI