Spark and Kafka Development – 75 Hours

1. **Introduction to Big Data**
2. **Big Data Components**
3. **Introduction to Apache Spark**
    a. What is Apache Spark?
    b. Starting the Spark Shell
    c. Using the Spark Shell
    d. Getting Started with RDD
    e. Getting Started with DataFrames
    f. DataFrame Operations
4. **DataBricks**
    a. Introduction to DataBricks
    b. Create your own DataBricks workspace
    c. Create a notebook inside your home folder in DataBricks – Hands-on
    d. Understand the fundamentals of Apache Spark notebook
    e. Create and attach to a Spark cluster – Hands-on
    f. Adding Libraries in DataBricks – Hands-on
    g. How Data Bricks is different than tradition Apache Spark clusters
    h. Benefits of using Data Bricks
5. **Introduction to RDD**
    a. RDD Overview
    b. RDD Data Sources
    c. Creating and Saving RDDs
    d. RDD Operations
    e. LAB: Loading and writing Unstructured file using RDD with RDD Operations
6. **Transforming Data with RDDs**
    a. Writing and Passing Transformation Functions
    b. Transformation Execution
    c. RDD lazy Evaluation
    d. RDD Partitions and Coalesce
    e. LAB: Transformation with RDD and Repartition it.
7. **Working with DataFrames**
    a. Creating DataFrames from Data Sources
    b. Saving DataFrames to Data Sources
    c. DataFrame Schemas
    d. Eager and Lazy Execution
    e. Analyzing Data with DataFrame Queries
    f. Querying DataFrames Using
    g. Column Expressions

h. Grouping and Aggregation Queries
i. Joining DataFrames
j. Catalyst Execution Plan
k. LAB: Data Frame Actions and Transformation
l. LAB: ETL Using Data Frame m. LAB: Conversion of RDD to DataFrame

8. Querying Tables and Views with Apache Spark SQL

a. Querying Tables in Spark Using SQL
b. Querying Files and Views
c. The Catalog API
d. Comparing Spark SQL, Apache Impala, and Apache Hive-on-Spark
e. LAB: Querying Tables using SparkSQL

9. Writing, Configuring, and Running on Notebook

a. Apache Spark Applications
b. Writing a Spark Application in Cell
c. Running an Application
d. DataBricks Spark Application Web UI
e. Configuring Application Properties
f. Log aggregations in Spark

10. Batch ETL Using Spark

a. Connecting RDBMS to Spark
b. Ingesting Data to Spark DataFrames
c. Business flow using a use case
d. LAB: Batch ETL use case – Passport Analysis and e-commerce sales analysis

11. Kafka Introduction

a. Architecture
b. Overview of key concepts
c. Overview of ZooKeeper
d. Cluster, Nodes, Kafka Brokers
e. Consumers, Producers, Logs, Partitions, Records, Keys
f. Partitions for write throughput
g. Partitions for Consumer parallelism (multi-threaded consumers)
h. Replicas, Followers, Leaders
i. How to scale writes
j. Disaster recovery
k. Performance profile of Kafka
l. Consumer Groups, “High Water Mark”, what do consumers see
m. Consumer load balancing and fail-over
n. Working with Partitions for parallel processing and resiliency

12. Writing Kafka Producers

a. Introduction to Producer Java API and basic configuration

13. Writing Kafka Consumers Basics

a. Introduction to Consumer Java API and basic configuration

14. **Low-level Kafka Architecture**
    a. a. Motivation Focus on high-throughput
    b. Embrace file system / OS caches and how this impacts OS setup and usage
    c. File structure on disk and how data is written
    d. Kafka Producer load balancing details
    e. Producer Record batching by size and time
    f. Producer async commit and commit (flush, close)
    g. Pull vs poll and backpressure
    h. Compressions via message batches (unified compression to server, disk and consumer)
    i. Consumer poll batching, long poll
    j. Consumer Trade-offs of requesting larger batches
    k. Consumer Liveness
    l. Managing consumer position (auto-commit, async commit and sync commit)
    m. Messaging At most once, At least once, Exactly once
    n. Performance trade-offs message delivery semantics
    o. Performance trade-offs of poll size
    p. Replication, Quorums, ISRs, committed records
    q. Failover and leadership election
15. **Writing Advanced Kafka Producers**
    a. Using batching (time/size)
    b. Using compression
    c. Async producers and sync producers
    d. Commit and async commit
    e. Default partitioning (round robin no key, partition on key if key)
    f. Controlling which partition records are written to (custom partitioning)
    g. Message routing to a particular partition (use cases for this)
    h. Advanced Producer configuration
    i. Lab 1: Use message batching and compression
    j. Lab 2: Use round-robin partition
    k. Lab 3: Use a custom message routing scheme
16. **Writing Advanced Kafka Consumers**
    a. Adjusting poll read size
    b. Implementing at most once message semantics using Java API
    c. Implementing at least once message semantics using Java API
    d. Implementing as close as we can get to exactly once Java API
    e. Re-consume messages that are already consumed
    f. Using ConsumerRebalanceListener to start consuming from a certain offset (consumer.seek*)
    g. Assigning a consumer a specific partition (use cases for this)
    h. Lab 1 Write Java Advanced Consumer
    i. Lab 2 Adjusting poll read size
    j. Lab 3 Implementing at most once message semantics using Java API
    k. Lab 4 Implementing at least once message semantics using Java API
    l. Lab 5 Implementing as close as we can get to exactly once Java API

17. **Schema Management in Kafka**
    a. Avro overview
    b. Avro Schemas
    c. Flexible Schemas with JSON and defensive programming
    d. Using Kafka's Schema Registry
    e. Topic Schema management
    f. Validation of schema
    g. Lab1 Topic Schema management
    h. Lab 2 Validation of schema
    i. Lab 3 Prevent Consumer from accepting unexpected schema / defensive programming
18. **Kafka REST Proxy**
    a. Using the REST API to write a Producer
    b. Using the REST API to write a Consumer
    c. Lab Writing REST Producer
    d. Lab Writing REST Consumer
19. **Kafka Connect**
    a. Kafka Connect Basics
    b. Modes of Working: Standalone and Distributed
    c. Configuring Connectors
    d. Tracking Kafka Connector Offsets
    e. Lab using Kafka Connect – Sync with Mysql and writing data to mysql
20. **Introduction to KSQL – Confluent Based**
    a. Introduction to KSQL
    b. Using KSQL
    c. KSQL - Data Manipulation
    d. KSQL - Aggregations
    e. Lab using KSQL
21. **RealTime ETL and Event partitions**
    a. Connecting Spark with Kafka
    b. Writing Spark Streaming Application
    c. Spark Structured Streaming and DStreams
    d. Aggregations on Spark Streaming
    e. LAB: Real-Time Event partitions using Spark Streaming