Introduction to Spark Programming

Course Outline

1. Scala Ramp Up
    - Scala Introduction, Variables, Data Types, Control Flow
    - The Scala Interpreter
    - Collections and their Standard Methods (e.g. map())
    - Functions, Methods, Function Literals
    - Class, Object, Trait, case Class
2. Introduction to Spark
    - Overview, Motivations, Spark Systems
    - Spark Ecosystem
    - Spark vs. Hadoop
    - Acquiring and Installing Spark
    - The Spark Shell, SparkContext
3. RDDs and Spark Architecture
    - RDD Concepts, Lifecycle, Lazy Evaluation
    - RDD Partitioning and Transformations
    - Working with RDDs - Creating and Transforming (map, filter, etc.)
4. Spark SQL, DataFrames, and DataSets
    - Overview
    - SparkSession, Loading/Saving Data, Data Formats (JSON, CSV, Parquet, text ...)
    - Introducing DataFrames and DataSets (Creation and Schema Inference)
    - Supported Data Formats (JSON, Text, CSV, Parquet)
    - Working with the DataFrame (untyped) Query DSL (Column, Filtering, Grouping, Aggregation)
    - SQL-based Queries
    - Working with the DataSet (typed) API
    - Mapping and Splitting (flatMap(), explode(), and split())

- DataSets vs. DataFrames vs. RDDs

5. Shuffling Transformations and Performance
    - Grouping, Reducing, Joining
    - Shuffling, Narrow vs. Wide Dependencies, and Performance Implications
    - Exploring the Catalyst Query Optimizer (explain(), Query Plans, Issues with lambdas)
    - The Tungsten Optimizer (Binary Format, Cache Awareness, Whole-Stage Code Gen)
6. Performance Tuning
    - Caching - Concepts, Storage Type, Guidelines
    - Minimizing Shuffling for Increased Performance
    - Using Broadcast Variables and Accumulators
    - General Performance Guidelines
7. Creating Standalone Applications
    - Core API, SparkSession.Builder
    - Configuring and Creating a SparkSession
    - Building and Running Applications - sbt/build.sbt and spark-submit
    - Application Lifecycle (Driver, Executors, and Tasks)
    - Cluster Managers (Standalone, YARN, Mesos)
    - Logging and Debugging
8. Spark Streaming
    - Introduction and Streaming Basics
    - Spark Streaming (Spark 1.0+)
    - DStreams, Receivers, Batching
    - Stateless Transformation
    - Windowed Transformation
    - Stateful Transformation
    - Structured Streaming (Spark 2+)
    - Continuous Applications
    - Table Paradigm, Result Table
    - Steps for Structured Streaming

- Sources and Sinks
- Consuming Kafka Data
- Kafka Overview
- Structured Streaming - 'kafka' format
- Processing the Stream