Spark and Scala :- 25 days – 25 hrs

Describe Features of Apache Spark

- How Spark fits in Big Data ecosystem
- Why Spark & Hadoop fit together

Define Spark Components

- Driver Program
  - Spark Context
- Cluster Manager
- Worker
  - Executor
  - Task
- Spark RDD
  - Spark Context
- Spark Libraries

Load data into Spark

- Different data sources and formats
  - HDFS
  - Amazon S3
  - Local File System
  - Text
  - JSON
  - CSV
  - Sequence File
- Create & Use RDD, Data Frames

Apply dataset operations to Resilient Distributed Datasets

- Transformation
- Actions
- Cache Intermediate RDD
  - Lineage Graph
  - Lazy Evaluation

Use Spark DataFrames for simple queries

- Create Data Frame
- Spark Interactive shell (Scala & Python)
- Spark SQL

Define different ways to run your application

Build and launch a standalone application

- Spark Program Life Cycle
- Function of Spark Context
- Different Way to Launch Spark Application
  - Local
  - Standalone
  - Hadoop YARN
  - Apache Mesos
- Launch Spark Application
  - Spark-Submit
  - Monitor the Spark Job

Describe & Create pair RDD

- Key-Value pair
- Apache Spark vs Apache Hadoop MapReduce
- Create RDD from existing non-pair RDD
- Create pair RDD by loading certain formats
- Create pair RDD from in-memory collection of pairs

Apply Operations on pair RDD

- Group ByKey
- Reduce ByKey
- Other Transformations
  - Joins

Control partitioning across nodes

- RDD Partition
- Types of Partition
  - Hash Partitioning
  - Range Partitioning

- Benefit of Partitioning
- Best Practices

More on Data Frames

- Explore Data in DataFrames
- Create UDFs (user define functions)
  - UDF with Scala DSL
  - UDF with SQL
- Repartition Data Frames.
- Infer Schema by Reflection
- DataFrame from database table
- DataFrame from JSON

Monitor Apache Spark Applications

- Spark Execution Model
- Debug and Tune Spark Applications

Identify Spark Unified Stack Components

- Spark SQL
- Spark Streaming
- Spark MLib
- Spark GraphX

Benefits of Apache Spark over Hadoop Ecosystem

Describe Spark Data pipeline Use Cases

- Spark Streaming Architecture
- Dstream and a spark streaming application
  - Define Use Case (Time Series Data)
  - Basic Steps
  - Save Data to HBase
- Operations on DStream
  - Transformations
  - Data Frame and SQL Operations
- Define Windowed Operation
  - Sliding Window

- Windowed Computation
- Window based Transformation
- Window Operations

• Fault tolerance of streaming applications

- Fault Tolerance in Spark Streaming
- Fault Tolerance in Spark RDD
- Check pointing

Describe Graph X

Define Regular, Directed, and property graphs

Create a Property Graph

Perform Operations on Graphs

Describe Apache Spark MLib

Describe the Machine Learning Techniques

• Classifications

• Clustering

• Collaborative Filtering

Use Collaborative filtering to predict user choice

Scala

• Introduction

• A first example

• Expressions and Simple Functions

• First Class function

• Classes and Objects

• Case classes and Pattern matching

• Generic types and methods

• Lists

• For- Comprehension

• Mutable State

• Computing with Streams

• Lazy Values

• Implicit Parameters and Conversions

- Handley / Milner type Interface
- Abstraction for concurrency