# Pre-requisites

Python

Basics of Spark

# Day 1

1. What is Bigdata?
2. Hadoop Yarn Architecture
3. An Introduction to Apache Spark
   1. Spark’s Basic Architecture
   2. Spark’s Language APIs
   3. Starting Spark
   4. The SparkSession
   5. Data Frames
      1. Partitions
   6. Transformations
      1. Lazy Evaluation
   7. Actions
   8. Spark UI
4. Spark’s Toolset
5. Structured API
   1. Data Frames and Datasets
   2. Schemas
   3. Overview of Structured Spark Types
      1. Data Frames Versus Datasets
      2. Columns
      3. Rows
      4. Spark Types
   4. Overview of Structured API Execution
      1. Logical Planning
      2. Physical Planning
      3. Execution
6. Basic Structured Operations
   1. Schemas
   2. Columns and Expressions
   3. Records and Rows
   4. Data Frame Transformations
7. Executing on Spark
   1. Run job from command line
   2. Go through execution plan / DAG on spark for existing model
8. Using logging

# Day 2

1. Details of the Case Study which we will use in our hands-on
2. Aggregations
   1. Aggregation Functions
      1. count
      2. min and max
      3. sum
      4. avg
   2. Grouping
   3. Window Functions
   4. User-Defined Aggregation Functions
3. Joins
   1. Join Expressions
   2. Join Types
   3. Inner Joins
   4. Outer Joins
   5. Left Outer Joins
4. Right Outer Joins
5. Data Sources
   1. The Structure of the Data Sources API
   2. CSV Files
   3. JSON Files
   4. ORC Files
   5. AVRO Files
   6. Parquet Files`
   7. Advanced I/O Concepts
      1. Splitable File Types and Compression
      2. Reading Data in Parallel
      3. Writing Data in Parallel
      4. Managing File Size
6. Integrating Spark with Hive
7. Spark SQL
   1. What Is SQL?
   2. How to Run Spark SQL Queries?
   3. Catalog
   4. Tables
   5. Views
   6. Select Statements
   7. Databases

# Day 3

1. Resilient Distributed Datasets (RDDs)
   1. What Are the Low-Level APIs?
   2. About RDDs
      1. Types of RDDs
      2. When to Use RDDs?
   3. Creating RDDs
   4. Manipulating RDDs
   5. Transformations
      1. distinct
      2. filter
      3. map
      4. sort
      5. Random Splits
   6. Actions
      1. reduce
      2. count
      3. first
      4. max and min
      5. take
   7. Saving Files
      1. saveAsTextFile
      2. SequenceFiles
      3. Hadoop Files
2. Caching
3. Persistence
   * 1. Different types of persistence
     2. When to use which kind of persistence.
4. Checkpointing
   * 1. What is checkpointing?
     2. localCheckpoint
     3. When and where to use?
5. Distributed Shared Variables
   1. Broadcast Variables
   2. Accumulators
      1. Basic Example
      2. Custom Accumulators

**Day 4**

1. Configuring Applications
2. PySpark UDF Introduction
   1. What is UDF?
   2. Why do we need it?
3. Create PySpark UDF (User Defined Function)
   1. Create a Data Frame
   2. Create a Python function
   3. Convert python function to UDF
4. Using UDF with Data Frame
   1. Using UDF with Data Frame select()
   2. Using UDF with Data Frame withColumn()
   3. Registering UDF & Using it on SQL query
5. Introducing Apache Parquet file format
   1. What is Apache Parquet?
   2. Parquet Format vs. CSV
   3. Advantages of Parquet Columnar Storage
   4. Primitive data types in Parquet format
   5. Apache Parquet Spark Example
      1. Spark Write Data Frame to Parquet file format
      2. Spark Read Parquet file into Data Frame
      3. Append to existing Parquet file
      4. Using SQL queries on Parquet
      5. Spark parquet partition – Improving performance
      6. Spark Read a specific Parquet partition
6. Apache Arrow in PySpark
   1. What is Apache Arrow?
   2. Apache PyArrow with Apache Spark
   3. What is the problem with existing Pandas/Spark conversion without PyArrow?
   4. How to use PyArrow in Spark to optimize?
   5. Enabling for Conversion to/from Pandas
   6. Pandas UDFs
   7. Compatibility Setting for PyArrow
   8. Converting Pandas Data frame to Apache Arrow Table
   9. PyArrow Table to Pandas Data Frame
   10. How does the PyArrow enabled conversion work internally?

**Day 5**

1. Performance Tuning
   1. Indirect Performance Enhancements
      1. Design Choices
      2. Object Serialization in RDDs
      3. Cluster Configurations
      4. Scheduling
      5. Data at Rest
      6. Shuffle Configurations
      7. Memory Pressure and Garbage Collection
   2. Direct Performance Enhancements
      1. Parallelism
      2. Improved Filtering
      3. Repartitioning and Coalescing
      4. User-Defined Functions (UDFs)
      5. Temporary Data Storage (Caching)
      6. Joins
      7. Aggregations
      8. Broadcast Variables

# Day 6

1. Machine Learning Basic Concepts
   1. Importing the Libraries
   2. Importing the Dataset
   3. Summary of Object-oriented programming: classes & objects
   4. Missing Data Treatment
   5. Categorical Data
   6. Splitting the Dataset into the Training set and Test set
   7. Feature Scaling
2. Analytics and Machine Learning
   1. What Is Spark’s MLlib?
   2. High-Level MLlib Concepts
   3. MLlib in Action
      1. Feature Engineering with Transformers
      2. Estimators
      3. Pipelining Our Workflow
      4. Training and Evaluation
      5. Persisting and Applying Models
3. Preprocessing and Feature Engineering
   1. Formatting Models According to Use Case
   2. Transformers
   3. Estimators for Preprocessing
      1. Transformer Properties
   4. High-Level Transformers
      1. SQL Transformers
      2. VectorAssembler
   5. Working with Continuous Features
      1. Bucketing
      2. Scaling and Normalization
      3. StandardScaler
   6. Working with Categorical Features
      1. StringIndexer
      2. Converting Indexed Values Back to Text
      3. Indexing in Vectors
      4. One-Hot Encoding
   7. Feature Manipulation
   8. Feature Selection

**Day 7**

1. Pipeline
   1. Building of pipeline
   2. Saving it
   3. Use on a different dataset
2. Classification
   1. Use Cases
   2. Types of Classification
   3. Classification Models in MLlib
   4. Logistic Regression
      1. Logistic Regression Intuition
      2. Sigmoid Function
      3. Model Hyperparameters
      4. Training Parameters
      5. Prediction Parameters
      6. Example
   5. Decision Trees
      1. Decision Tree Regression Intuition
      2. Pruning
      3. Overfitting in Decision Tree
      4. Entropy
      5. Information Gain
      6. Model Hyperparameters
      7. Training Parameters
      8. Prediction Parameters
3. Regression
   1. Use Cases
   2. Regression Models in MLlib
   3. Linear Regression
      1. Simple Linear Regression Intuition
      2. RMSE
      3. Model Hyperparameters
      4. Training Parameters
      5. Example
      6. Training Summary
   4. Decision Trees
      1. Model Hyperparameters
      2. Training Parameters
      3. Example