PySpark Development

Course Contents

Module 01: A Brief Primer on PySpark

- A Brief Primer on PySpark
- Brief Introduction to Spark
- Apache Spark Stack
- Spark Execution Process
- Newest Capabilities of PySpark
- Cloning GitHub Repository

Module 02: Resilient Distributed Datasets

- Creating RDDs
- Schema of an RDD
- Understanding Lazy Execution
- Introducing Transformations – .map(…)
- Introducing Transformations – .filter(…)
- Introducing Transformations – .flatMap(…)
- Introducing Transformations – .distinct(…)
- Introducing Transformations – .sample(…)
- Introducing Transformations – .join(…)
- Introducing Transformations – .repartition(…)

Module 03: Resilient Distributed Datasets and Actions

- Introducing Actions – .collect(…)
- Introducing Actions – .reduce(…) and .reduceByKey(…)
- Introducing Actions – .count()
- Introducing Actions – .foreach(…)
- Introducing Actions – .aggregate(…) and .aggregateByKey(…)
- Introducing Actions – .coalesce(…)
- Introducing Actions – .combineByKey(…)
- Introducing Actions – .histogram(…)
- Introducing Actions – .sortBy(…)
- Introducing Actions – Saving Data
- Introducing Actions – Descriptive Statistics

Module 04: DataFrames and Transformations

- Creating DataFrames
- Specifying Schema of a DataFrame
- Interacting with DataFrames
- The .agg(…) Transformation
- The .sql(…) Transformation
- Creating Temporary Tables
- Joining Two DataFrames
- Performing Statistical Transformations
- The .distinct(…) Transformation

Module 05: Data Processing with Spark DataFrames

- Filtering Data
- Aggregating Data
- Selecting Data
- Transforming Data
- Presenting Data
- Sorting DataFrames
- Saving DataFrames
- Pitfalls of UDFs
- Repartitioning Data