KOENIG
step forward

Data Processing with PySpark

Module 1 : Introduction to Apache Spark

- Introduction to Big Data
- What is Apache Spark?
- Evaluation of Apache Spark
- Features
- Spark Architecture
- Spark Vs Hadoop Map Reduce
- Spark SQL Vs HIVE

Module 2 : Installation

- Installation on MAC
- Installation on Windows
- With Scala and Intellij
- Creating DataBricks Account
- DataBricks Compute, Notebook, tables

Module 3 : PySpark Introduction

- Why Pyspark?
- Need for Pyspark
- Spark Python Vs Scala
- Pyspark features
- Real-life usage of PySpark
- Web/Application

Module 4 : PySpark Basics

- SparkSession
- SparkContext
- Stage
- Executor
- RDD
- Parallelize

Module 5 : RDDs (Resilient Distributed Datasets)

- Parallelize

- Read Text File
- Read CSV
- Create RDD
- RDD Persistence and Caching Mechanism
- RDD Features
- RDD Limitations
- RDD Lineage
- Action
- Pair Functions- Paired RDD
- Repartition and Coalesce
- Shuffle Partitions
- Cache vs Persist

5.1 PERSISTENCE Options:

- MEMORY_ONLY
- MEMORY_SER_ONLY
- DISK_ONLY
- DISK_SER_ONLY
- MEMORY_AND_DISK_ONLY

5.2 CORE COMPUTING

- Fault Tolerance model in spark
- Different ways of creating a RDD
- Word Count Example
- Creating spark objects(RDDs) from Scala Objects(lists).
- Increasing the no of partitions
- Aggregations Over Structured Data:
- reduceByKey()

5.3 GROUPINGS AND AGGREGATIONS

- Single Grouping and Single Aggregation
- Single Grouping and multiple Aggregation
- Multi Grouping and Single Aggregation
- Multi Grouping and Multi Aggregation

5.4 Various Actions and Transformations

- countByKey()
- countByValue()
- sortByKey()

- zip()
- Union()
- Distinct()

Module 6 :- SQL - DataFrame

- Introduction
- Making data Structured
- Case Classes
- ways to extract case class objects
- using function
- using map with multiple exressions
- using map with single expression
- Sql Context
- Data Frames API
- DataSet API
- RDD vs DataFrame vs DataSet
- Create a DataFrame
- Create an empty DataFrame
- Convert RDD to DataFrame
- Convert DataFrame to Pandas
- union() & unionAll()
- unionByName()
- UDF (User Defined Function)
- map()

Module 7 : SQL Functions

- Aggregate Functions
- Window Functions
- Date and Timestamp Functions
- JSON Functions
- Read & Write JSON file

Module 8 :- Built-In Functions

- when()
- expr()
- lit()
- split()
- concat_ws()
- substring()
- translate()
- regexp_replace()

- overlay()
- to_timestamp()
- to_date()

Module 9 :- External Sources

- Working with sql statements
- Spark and Hive Integration
- Spark and mysql Integration
- Working with CSV
- Working with JSON
- Transformations and actions on dataframes
- Narrow, wide transformations
- Addition of new columns, dropping of columns ,renaming columns
- Addition of new rows, dropping rows
- Handling nulls
- Joins

Module 10 : DEPLOYMENT MODES

- Local Mode
- Cluster Modes(Standalone , YARN