

PySpark with Kafka and Databricks Content

Table of Content:

1. What is Python?
2. Python Setup
 - Python3 Installation
 - PyCharm Installation
 - Jupyter-lab Installation
3. Data Types
 - Numbers
 - Variable Assignment
 - Strings
 - Strings Slicing and Indexing
 - String Properties and methods
 - Print Formatting with Strings
 - Lists
 - Dictionaries
 - Tuples
 - Sets
 - Booleans
4. Variables
5. Comparison operators
 - Comparison operators(<,>,<=,>=,==,!=)
 - Chaining comparison operators
6. Python Statements
 - if, elif and else statements
 - for loops
 - while loops
 - List Comprehension
7. Methods and Functions
 - Introduction to function
 - Basics of Functions
 - Logic with python function
 - Tuples unpacking
 - Interactions b/w python functions
 - Lambda Expressions(map, flatmap, filter functions)
8. Object Oriented Programming
 - Introduction
 - Attributes and Class Keyword
 - Class object, Attributes, Methods

- Inheritance and Polymorphism
- Magic/Dunder Methods
- 9. Modules and Packages
- 10. Error and Exception Handlings
- 11. Advanced Python
 - Python I/O
 - Reading and Writing to file and folder
 - Collections Module
 - DateTime Module
 - Math and Random Module
 - Logger Module
 - Regular Expression Module
 - Zipping and Un-zipping Module
- 12. Internals of Python
- 13. Pandas Module
 - Core components of Pandas, Series and Data frames
 - Processing data from CSV, Json, XML, Parquet, Database.
- 14. Work-Shop
 - Mutli Processing Module (Optional/Workshop).
 - Web Components (Optional/Workshop)
 - Flask
 - Flask Socket-IO

PySpark:

- 1. SparkCore
 - Why Spark?
 - Bird View of Spark Architecture
- Spark Core:
 - Abstractions in Spark.
- 2. RDD
 - What is RDD?
 - What are the different ways to create an RDD
 - parallelize
 - textfile
 - wholetextfile.
 - What are RDD Partitions and there importance
 - About RDD Parallelism
- 3. DAG
 - Jobs
 - Stages

- Tasks
- 4. Transformations and Actions
 - What are Narrow and Wide Transformations
 - Understanding and working on different transformations and Actions
- 5. In-detail Understanding about Py-spark Architecture
 - Overview of Pyspark Architecture
 - Py4j Module
 - Py4j Gateway Server
 - Python Runner and Python Worker
 - Compute method
 - Understanding Pyspark Serializations and De-serializations
 - Marshall
 - Pickle
- 7. Variables
 - Closure
 - Broadcast
 - Accumulator
- 8. Discussing Spark-Core optimizations techniques

PySpark-SQL:

1. Disadvantages of Pandas Dataframe
 - What is Spark Dataframe
 - Different ways of creating Dataframes.
 - RDD to DF and DF to RDD
 - Working with different data sources like CSV, XML, Excel, JSON, JDBC, Parquet, Hudi (Optional/Workshop) by using Different Spark SQL API's
 - Select, where, groupby, case, otherwise, etc.
2. Join
 - Hints
 - Broadcast

- Merge-sort
- Shuffle hash Join

3.Dataframe Persistence/Memory Management Techniques

- cache
- persist
 - MEMORY_ONLY, MEMORY_AND_DISK, MEMORY_ONLY_SER, MEMORY_AND_DISK_SER, DISK_ONLY, MEMORY_ONLY_2, MEMORY_AND_DISK_2
- unpersist

4.Windowing operations in Spark

- What is window and different types of windows
- Time-based
- Offset-based
- Analytics functions: rank, dense rank, row number, lead, lag , ect
- Spark Catalyst Optimizer/ Spark Query Engine
- Parsed logical plan, Analysed logical plan, Optimized logical plan, Physical plan
- Explain method
- Adaptive Query Executions
- Optimizing Skew joins

5.Understanding concepts of YARN

- Deploying pyspark Applications in YARN in client and cluster modes
- Discussing spark deployment strategies
 - Static deployment
 - Dynamic deployment

6. Spark Streaming

- Understanding Kafka Concepts
 - Creating PyKafka producers and consumers

7.Understanding the concept of spark structed streaming and integrating kafka - spark

8.AWS

S3(Datalake)

EMR

Lambda

Athena

9. Databricks

9.Final Project