Learning plan1_gakyeong_bae

Learning plan1

Objective

1. Process and analyze large-scale datasets efficiently.
2. Write distributed data processing applications using PySpark.
3. Query and manipulate data using SparkSQL.
4. Utilize Databricks for collaboration, scaling, and optimization of Spark workflows.

Detail

- Study 1 hour on M, W

week 4 : Introduction to Apache Spark

Goal :

Understand Spark's architecture and components.

Topic :

- What is Apache Spark? Why is it needed?
- Spark Architecture and Components
- library: Spark Core / MLlib
- Spark Ecosystem : RDDs / DataFrames API

Resource :

- Apache Spark Official Documentation : https://spark.apache.org/documentation.html
- How Data Engineering Works: https://www.youtube.com/watch?v=qWru-b6m030&t=741s
- Apache Spark Essential Training :https://www.linkedin.com/learning/apache-spark-essential-training/welcome?u=2153100
- Spark by {Examples}: https://sparkbyexamples.com/pyspark-tutorial/

week 5 : Introduction to pySpark

Goal :

Learn to use Apache Spark with Python.

Topic :

- PySpark setup
- Working with RDDs and DataFrames.
- Common PySpark transformations and actions

Resource :

- W3schools pyspark tutorial : https://www.w3schools.com/python/pyspark_intro.php

-pyspark-examples: https://github.com/spark-examples/pyspark-examples

week 6 : Introduction to sparkSQL

Goal :

Query data using SQL within Spark.

Topic :

- Setting up and using SparkSQL.
- Converting DataFrames to SQL tables.
- Running SQL queries on Spark datasets. #### Resource :
- SparkSQL : https://spark.apache.org/docs/latest/sql-getting-started.html

-Introduction to Spark SQL and DataFrames : https://www.linkedin.com/learning/introduction-to-spark-sql-and-dataframes/install-pyspark-21042116?u=2153100

-Simplilearn SparkSQL