PySpark Integration

Business Overview

Apache Spark is a distributed processing engine that is open source and used for large data applications. It uses in-memory caching and efficient query execution for quick analytic queries against any quantity of data. It offers code reuse across many workloads such as batch processing, interactive queries, real-time analytics, machine learning, and graph processing. It provides development APIs in Java, Scala, Python, and R.

Data Pipeline

A data pipeline is a technique for transferring data from one system to another. The data may or may not be updated, and it may be handled in real-time (or streaming) rather than in batches. The data pipeline encompasses everything from harvesting or acquiring data using various methods to storing raw data, cleaning, validating, and transforming data into a query-worthy format, displaying KPIs, and managing the above process.

Agenda

This is the sixth project in the Pyspark series. The fifth project involved the introduction to PySpark SQL, SQL function, and various joins available in PySpark SQL with the help of business case study. This project mainly focuses on the integration of PySpark with Amazon S3 and MySQL database to perform ETL(Extract-Transform-Load) and ELT(Extract-Load-Transform) operations.

Tech stack:
➔Language: Python
➔Package: Pyspark
➔Services: AWS S3, MySQL

PySpark:

PySpark is a Python interface for Apache Spark. It not only lets you develop Spark applications using Python APIs, but it also includes the PySpark shell for interactively examining data in a distributed context. PySpark supports most of Spark's capabilities, including Spark SQL, DataFrame, Streaming, MLlib, and Spark Core. In this project, you

will learn about core Spark architecture, Spark Sessions, Transformation, Actions, and Optimization Techniques using PySpark.

MySQL

MySQL is a SQL (Structured Query Language) based relational database management system. The platform can be used for data warehousing, e-commerce, logging applications, etc.

Amazon S3

Amazon S3 is a simple web service interface for object storage that allows you to store and retrieve unlimited amounts of data from anywhere on the internet.

It is used for Backup and archive for on-premises or cloud data Content, media, and software storage and distribution

Key Takeaways:

- Understanding the project overview
- Introduction to PySpark
- Introduction to Amazon S3
- Create bucket in Amazon S3
- Store data in S3 buckets
- Introduction to MySQL database
- Need of PySpark integration
- Understanding the concept of ETL
- Difference between ETL and ELT
- PySpark integration with Amazon S3
- PySpark integration with MySQL database