PYSPARK

ASSIGNMENT 26-12-2023

ETL WITH SPARK

ETL (Extract, Transform, Load) operations can be performed efficiently using PySpark, a Python library for Apache Spark. PySpark provides an interface for working with large-scale data processing. Here's a general outline of how you can perform ETL tasks using PySpark:

1. Setting up PySpark

pip install pyspark

1. Initializing PySpark

from pyspark.sql import SparkSession

# Initialize SparkSession

spark = SparkSession.builder.appName("ETL Example").getOrCreate()

1. Extract Data

# Read data from a CSV file

df = spark.read.csv("path\_to\_your\_file.csv", header=True, inferSchema=True)

1. Transform Data

# Example transformation: Selecting specific columns

transformed\_df = df.select("col1", "col2", "col3")

5.Load Data

# Write transformed data to a new CSV file

transformed\_df.write.csv("path\_to\_output.csv", header=True)

1. Execute ETL process

# Read data

data = spark.read.csv("path\_to\_your\_file.csv", header=True, inferSchema=True)

# Perform transformations

transformed\_data = data.select("col1", "col2", "col3")

# Write transformed data

transformed\_data.write.csv("path\_to\_output.csv", header=True)

