Pyspark Optimization Part 2

Agenda

This is the eleventh project in the Pyspark series. The tenth project mainly focuses on Spark Optimization techniques that are used to modify the settings and properties of Spark to ensure that the resources are utilized properly and the jobs are executed quickly. It also includes the execution of different optimization techniques such as Catalyst optimization, File formats optimization etc. This project mainly focuses on how to optimize PySpark using Shared variables, Serialization, Parallelism and built-in functions of Spark SQL.

Tech stack:

➔Language: Python
➔Package: Pyspark
➔Services: Spark

Key Takeaways:

- Understanding the project overview
- Introduction to Optimization
- Need of Optimization
- Understanding different Optimization techniques
- Difference between take and collect function
- Implementation of File Formats Optimization
- Implementation of Cache and Persist
- Use of reduceByKey function
- Difference between reduceByKey and groupByKey function
- Understanding Spark SQL built-in function
- Different types of Shared variables
- Implementation of Broadcast variables
- Implementation of Accumulator variables
- Serialization in PySpark
- Parallelism in PySpark

Note:

- S3 link for dataset - s3://airlines123/airline/data.zip