Pyspark Optimization Part 1

Agenda

This is the tenth project in the Pyspark series. The ninth project mainly focuses on partitioning, which is nothing but dividing data structure into parts. In a distributed system like Apache Spark, it can be defined as a division of a dataset stored as multiple parts across the cluster. This 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.

Tech stack:

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

Key Takeaways:

- Understanding the project overview
- Introduction to Optimization
- Need of Optimization
- Understanding different Optimization techniques
- Introduction to Catalyst Optimizer
- Implementation of Catalyst Optimizer
- Implementation of File Formats Optimization
- Implementation of Cache and Persist
- Use of reduceByKey function
- Difference between reduceByKey and groupByKey function
- Understanding Arrow Optimization

Note:

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