60_Subrato_D16AD

BDA - Experiment 2

Aim: To implement word count using MapReduce and RDBMS operations such as selection, union, projection, joins etc. using Pyspark.

Theory: • **MapReduce:** programming model and an associated implementation for processing and generating big datasets with a parallel distributed algorithms on a cluster.

Map → Process the input data into small chunks of data
Reduce ⇒ Combⁿ of shuffle and reduce stage to produce output

• **Spark:** Apache spark is a cluster computer technology designed for fast computation and is based on Hadoop MapReduce. It is designed to cover wide range of workloads such as batch applications, interactive queries and stream processing.

• **Pyspark:** Allows us to work Resilient Databases (RDD) in python. Pyspark combines python's learnability and ease of use with the power of Apache spark to enable processing and analysis of data at any size.

Conclusion: Word count program and RDBMS operations have been performed.