

DataFrames and Advanced Techniques

A Spark DataSet is an extension of the RDD object. It has rows, can run queries, and has a schema (which leads to more efficient storage and optimization). A DataFrame is just a DataSet of Row Objects, and unlike a DataSet the schema is inferred at runtime rather than compile time. This also means DataSets can only be used in compiled languages (NOT Python).

There are some instances where an RDD might be preferred, but for the vast majority of operations **DataSets are king**. They are more efficient, simplify development and allow for interoperability with other libraries.

Revision #1

Created 8 April 2024 20:44:08 by Elkip

Updated 9 April 2024 15:31:35 by Elkip