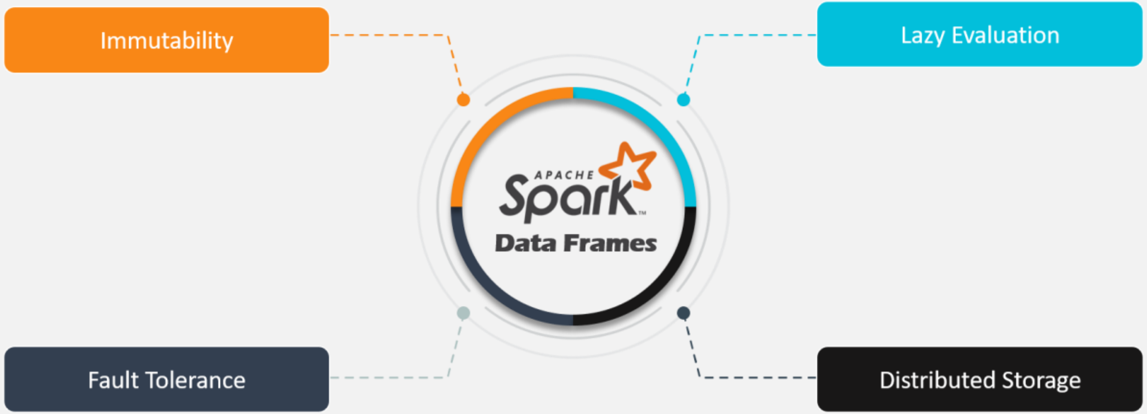
Spark DataFrames



Python and Spark

- In this course the main way we will be working with Python and Spark is through the DataFrame Syntax.
- If you’ve worked with pandas in Python, R, SQL or even Excel, a DataFrame will feel very familiar!

Python and Spark

- Spark DataFrames hold data in a column and row format.
- Each column represents some feature or variable.
- Each row represents an individual data point.

Python and Spark

- Spark began with something known as the “RDD” syntax which was a little ugly and tricky to learn.

- Now Spark 2.0 and higher has shifted towards a DataFrame syntax which is much cleaner and easier to work with!

Python and Spark – Why DataFrame

There is always a question which strikes in mind that if we already have RDD, why do we need DataFrame than?

DataFrame is one step ahead of RDD. Since it provides memory management and optimized execution plan.

- Custom Memory Management
- Optimized Execution plan

Python and Spark

- Spark DataFrames are able to input and output data from a wide variety of sources.
- We can then use these DataFrames to apply various transformations on the data.

Python and Spark

- At the end of the transformation calls, we can either show or collect the results to display or for some final processing.
- In this section we’ll cover all the main features of working with DataFrames that you need to know.

Python and Spark

- Once we have a solid understanding of Spark DataFrames, we can move on to utilizing the DataFrame MLlib API for Machine Learning.

Python and Spark

- After this section you will have a section for a “DataFrame Project”.
- This Project will be an analysis of some historical stock data information using all the Spark knowledge from this section of the course.

Python and Spark

- It will serve as a quick exercise review to test all the skills learned in this section.
- Let’s get started with learning the basics of Spark DataFrames!

Spark DataFrames Project Exercise