Intro to Spark DataFrame API

What is a DataFrame?

- DataFrame is a structured data type in Spark that can lead better expressiviity, performance and space efficiency.
- DataSet is the other structuted data type in Spark that is also statically typed (so it is only available in Scala and Java).
- The focus is on *"what to do?"* instead of *"how to do?"*
- The data is structured as a table (similar to a table in a relational database). We can formally describe the schema either directly in the host language or via a Data Definition Language (DDL).
- Data types supported in DataFrame in Spark.
- A RDD can be easily converted into a `DataFrame` and vice versa but there is a cost in the transformation.

Examples

- Example 1: DataFrame Schemas
- Example 2: Working with Columns
- Example 3: Working with Rows
- Example 4: Wordcount with DataFrames
- Two complete examples of using data frames.
- Example 5: Distribution of M&Ms by State
- Example 6: Fire Call Analytics for San Francisco