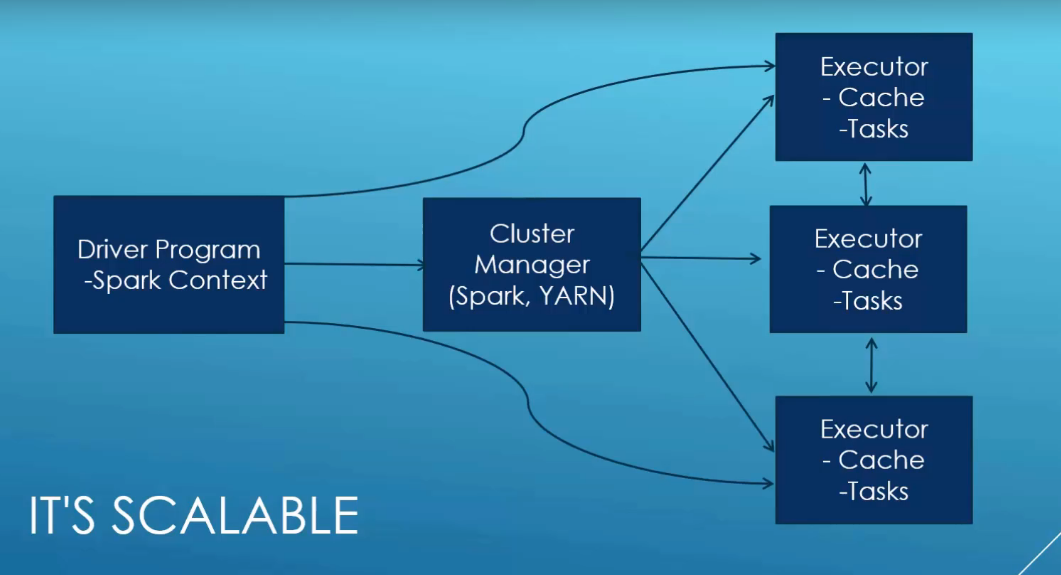
PySpark

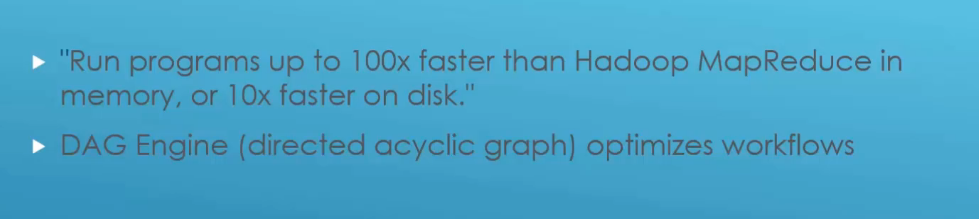
What is Spark: -

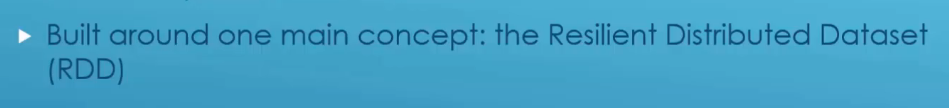


Spark Architecture: -

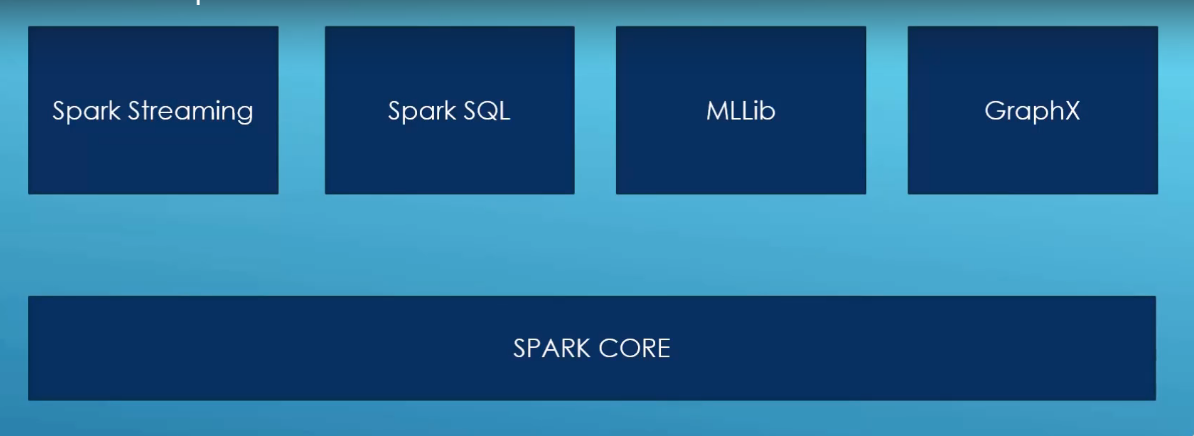


Why spark is fast: -

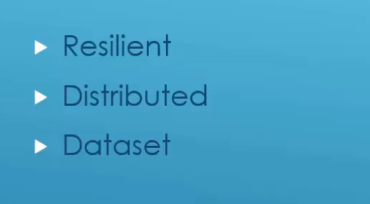




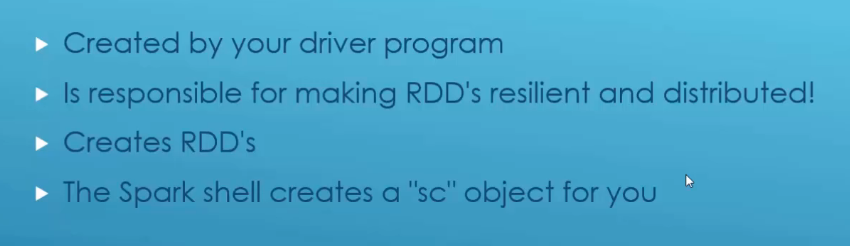
Spark Components: -



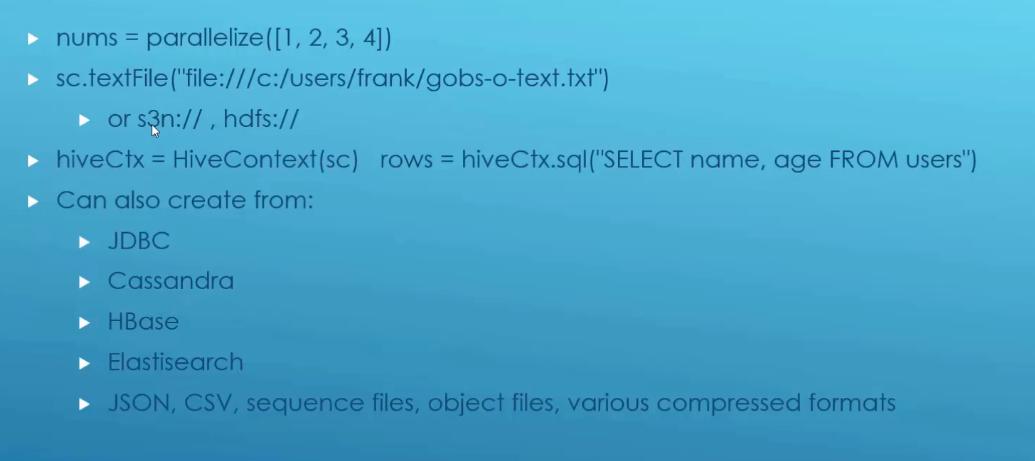
Introduction to RDD: -



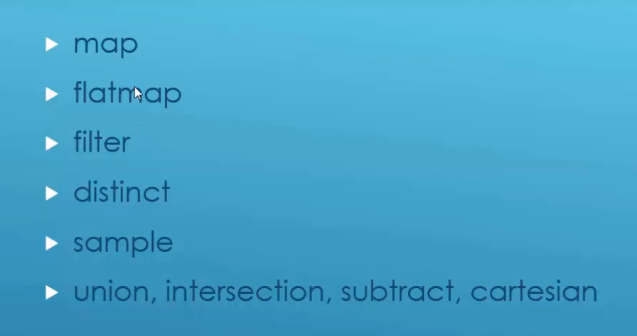
Spark Context: - Starting point of the spark framework



Creating RDDs: -

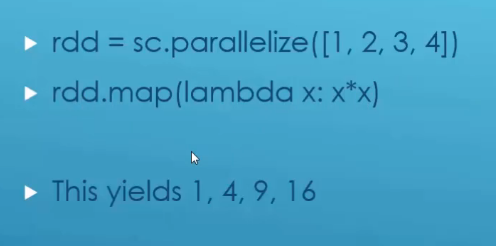


Common Transformation operations: -

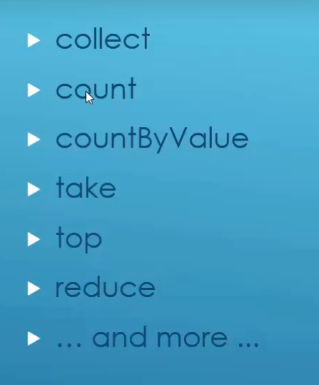


* Map – One to One relationship
* Flatmap – One to Many relationship

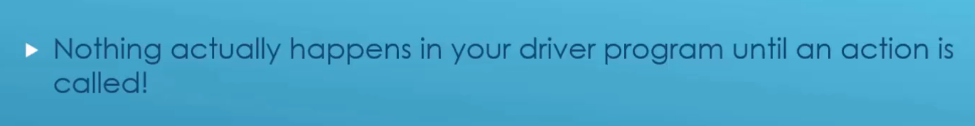
Map example: -



Some Actions: - Getting a result out of an RDD



Lazy Evaluation: -



Ratings Histogram example: -

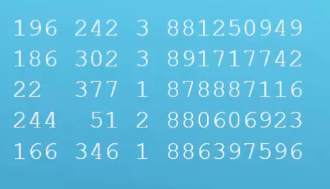
Import what we need: -



Set up our context: -



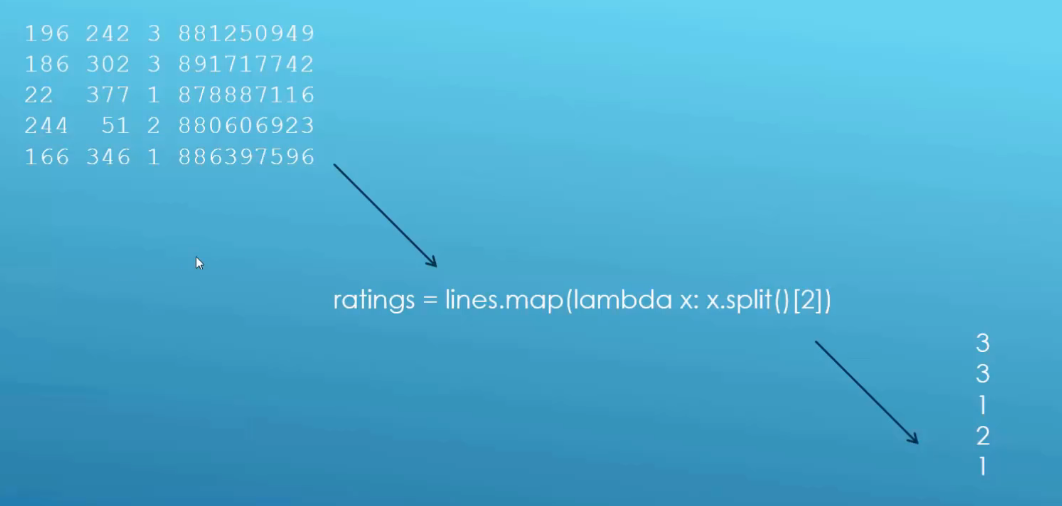
Our data: -



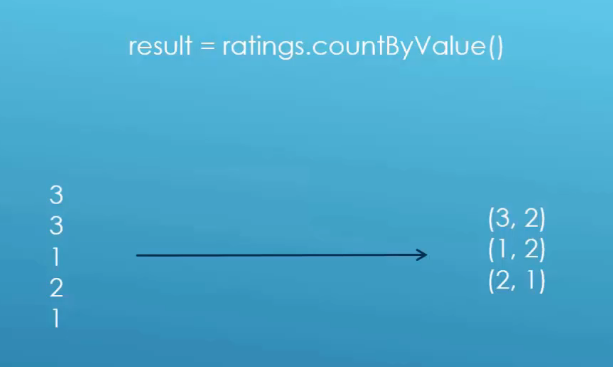
Loading the data: -



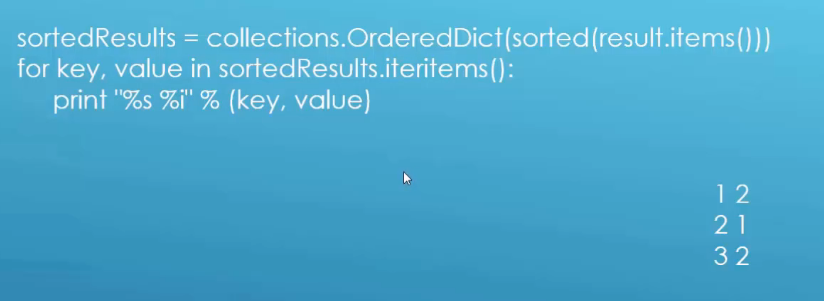
Extract the data we care about: -



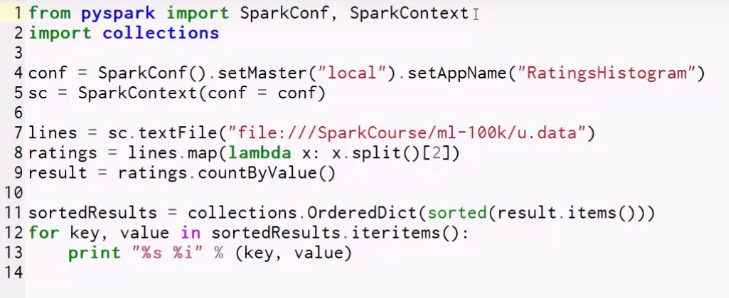
Performing an action: -



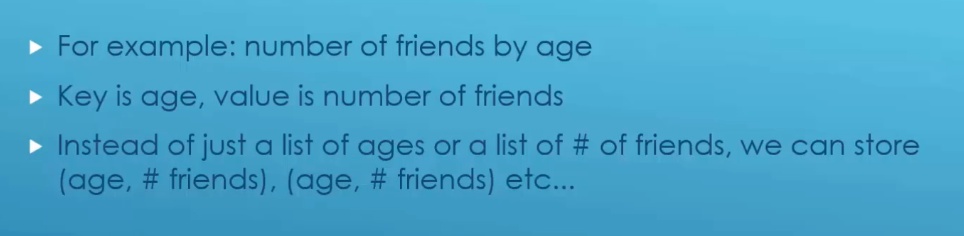
Sort and display the results: -



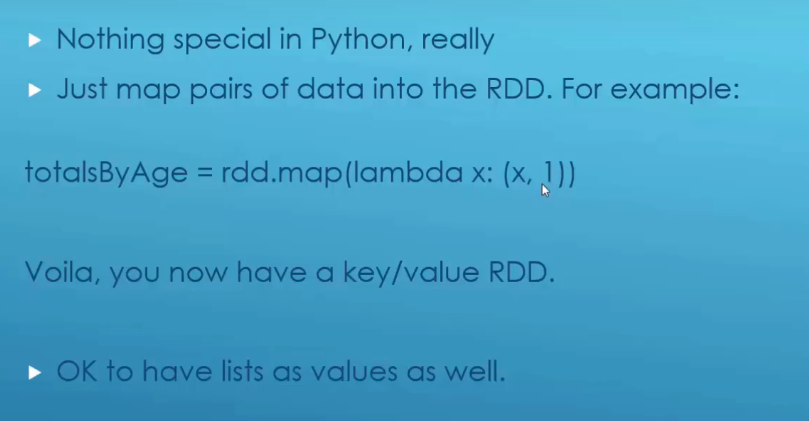
Ratings-counter: -



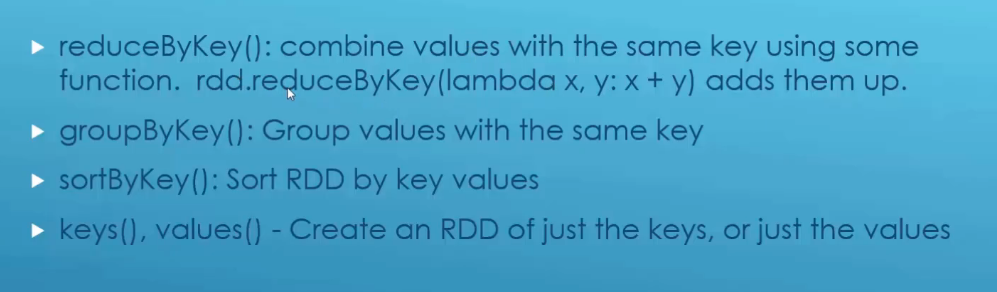
RDD’s can hold key/value pairs: -



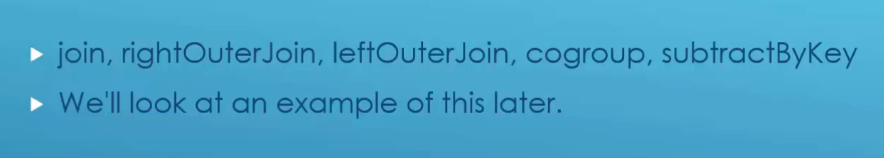
Creating a key/value RDD: -



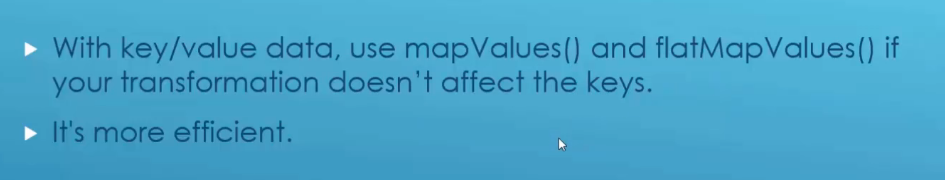
Spark can do special stuff with key/value data: -



We can do SQL-Style joins on two key/value RDD’s: -

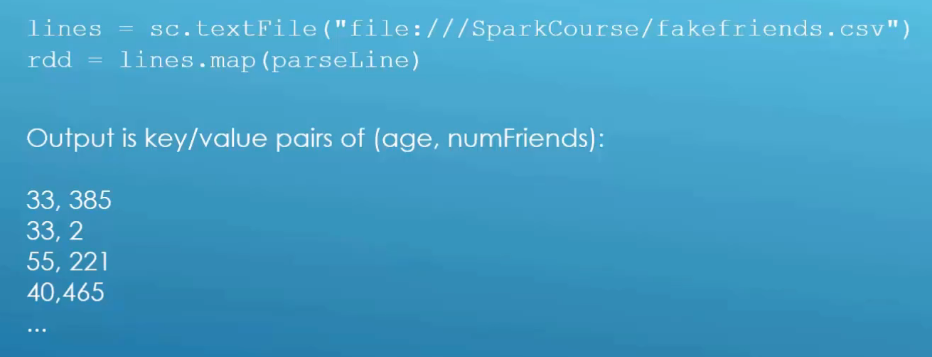


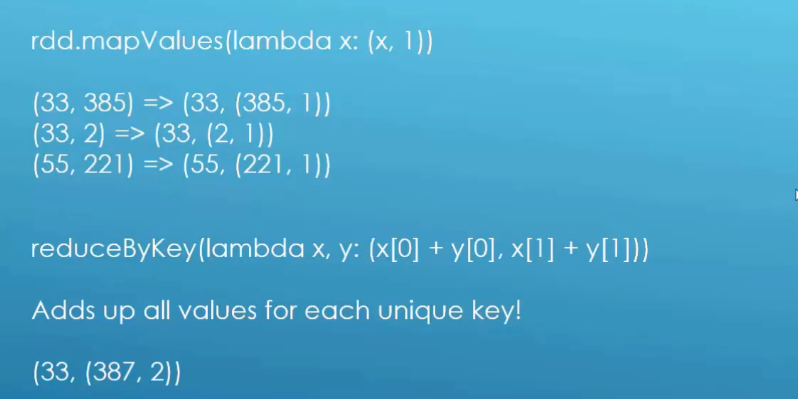
Mapping just the values of a key/value RDD’s: -

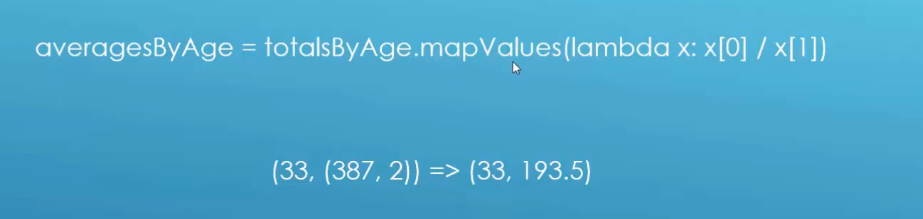


Average friends by age: -

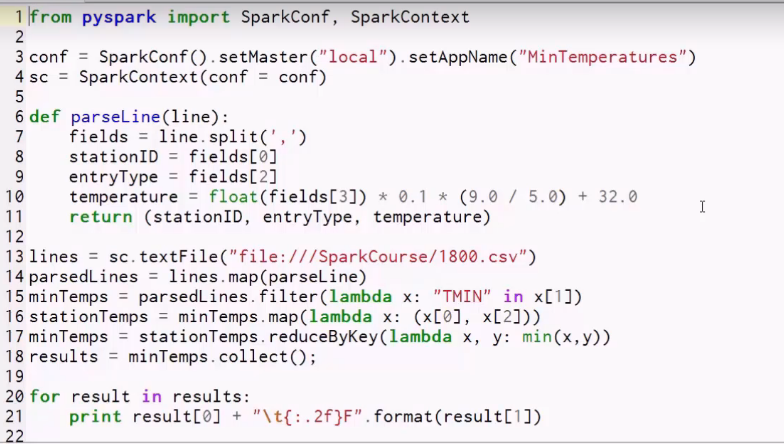






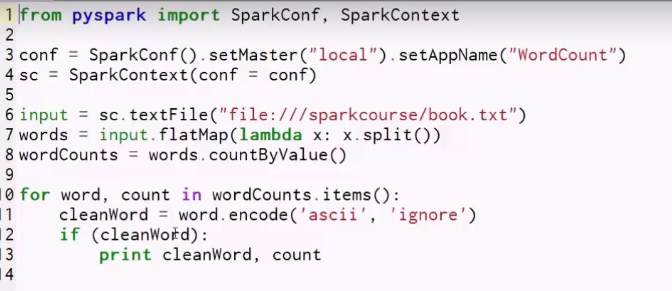


Filtering RDD’s: - Minimum Temperature Example



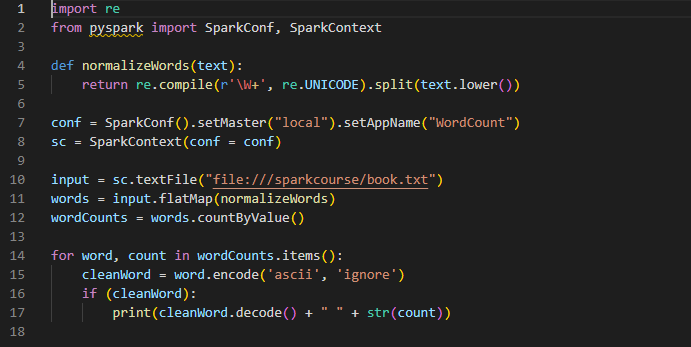
* Filter function accepts a Boolean function (returns True or False)
* If the output is True, it will keep the value otherwise discard it

Word count example: -

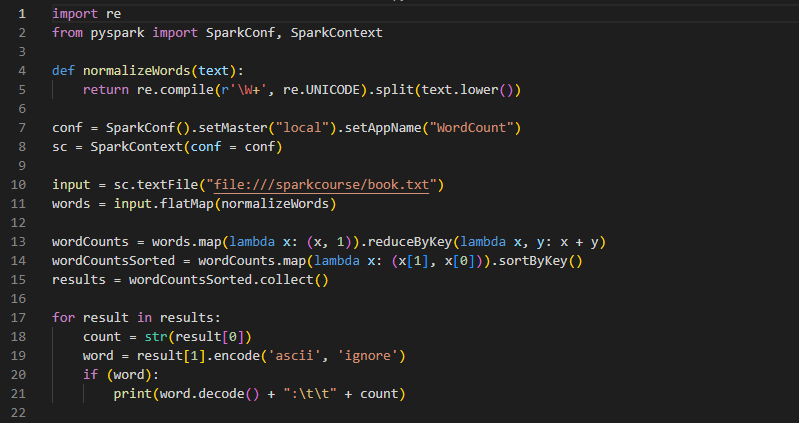


* Results are not really great
* Capitalizatin issues and special char issues etc

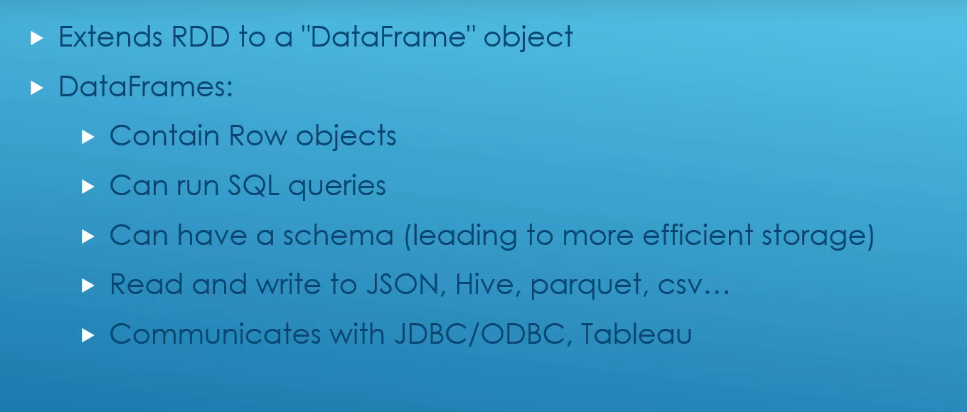
Word count better: -



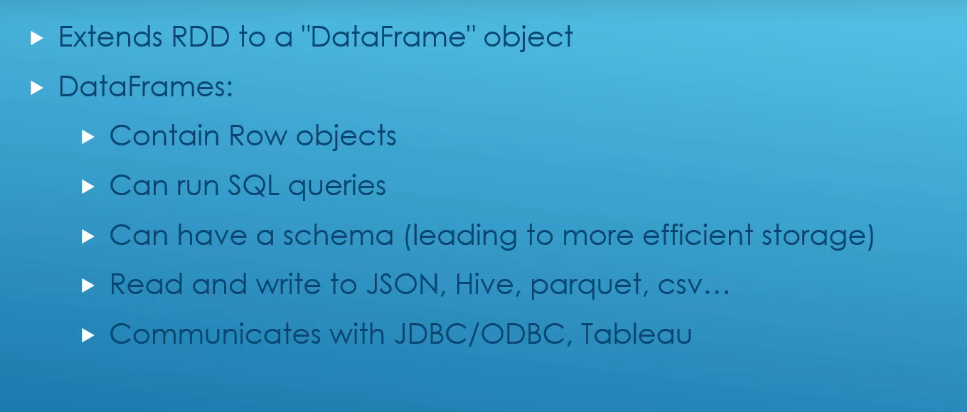
Word count better sorted: -



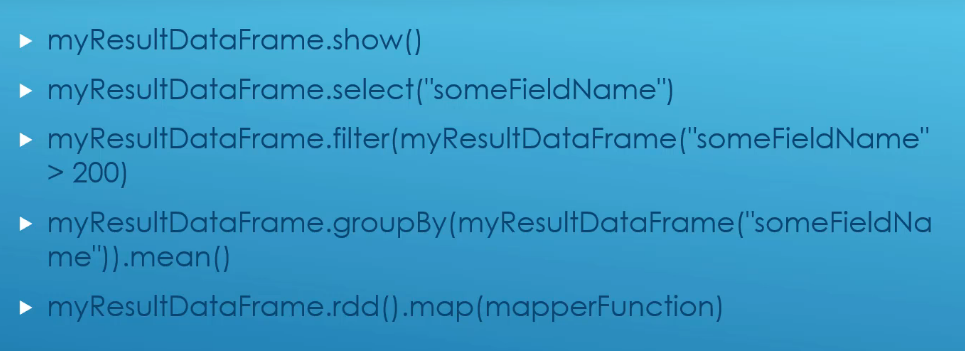
Dataframes: -



Using spark SQL in Python: -



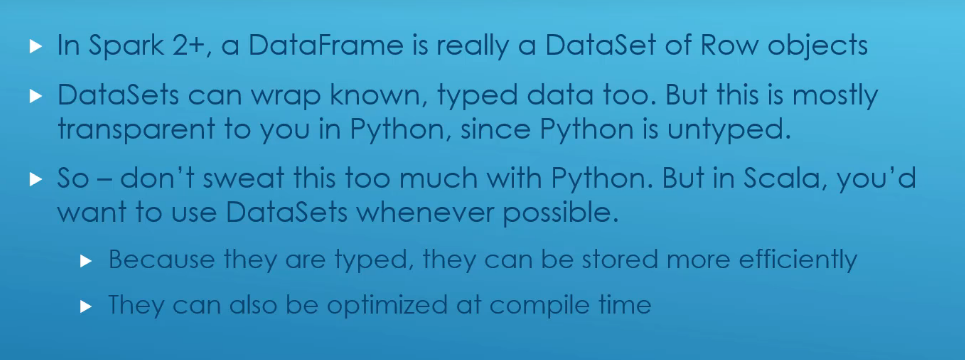
Other stuff you can do with dataframes: -



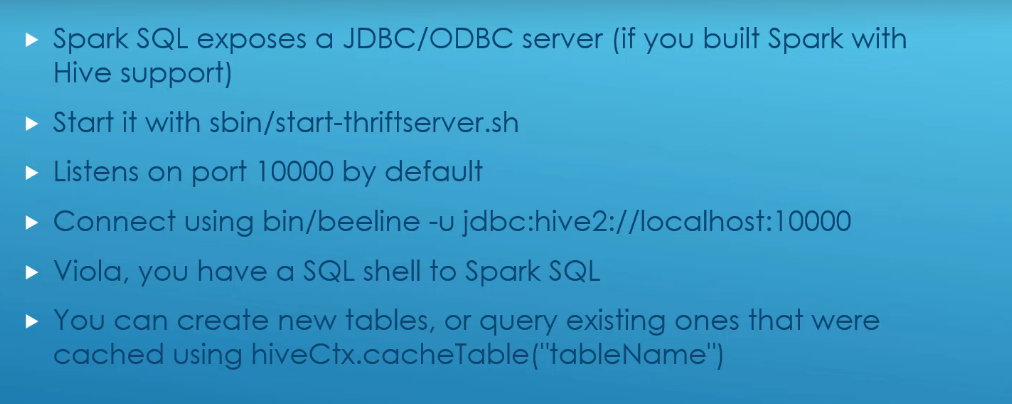
Dataframe are the new hotness: -



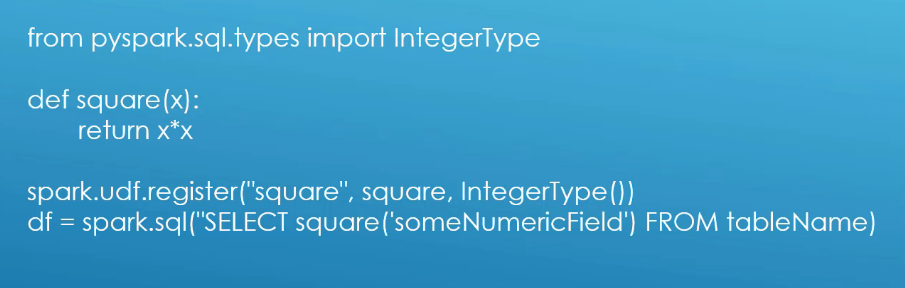
Dataframes vs Datasets: -



Shell access: -



UDF’s in spark: -



* You need to pass the data type the function will take as input