Summary:

ANLT-214 Data Engineering – Lab 2: Spark on Databricks

By Ali Taheri and Arda Ugur

October 2018

Our very first hands on experience with Apache Spark general-purpose cluster-computing frame work was quite enlightening. To begin with, the example provided by Professor Williamson was quite illustrative in demonstrating different aspects of Spark, using Databricks platform.

One of the limiting aspects of this lab assignment was the absence of massive datasets to select from, to have a genuine experience of “big data” Spark. Therefore, we would like to recommend making additional “big” datasets which wrangling and analyzing without utilizing Apache Spark would not be possible otherwise. Suffice it to state that, conceptual comprehension of transformations (mapping) and actions (reducing) is only possible when large data sets are employed in practice. Relatively small size of our selected data set, consequently the processing time of transformations and actions were short enough for us to experiment with our dataset.

For this lab assignment, we decided to work with the *Songs* dataset which has around 1 million rows of data points, although it is still a small dataset compared to the “airlines” dataset. The *Songs* dataset is consisting of 100 files, which allowed us to learn how to make a batch load of files in a Spark data frame. An important take away from us working with Spark data frames is that in comparison to conventional Pandas data frames, Spark data frames are more convenient for working with massive datasets and they offer a rich set of specialized functions. For example, functions such as ***agg*** allows users to perform aggregations on data frames and returns a new column with the calculated output and ***filter*** returns the rows of a data frame which matches the given condition. Another Spark function we find extremely useful for working with large datasets is ***Schema*** which returns the data frame schema in a tree format.

Our chosen *Songs* dataset is fairly clean, which almost did not require any extensive data clean up. Nonetheless, we decided to drop some of the columns with extensive amount of zero (0) values in order to have a more homogenous dataset. In addition to our exploratory analysis, we generated new columns for categorical attributes with ***StringIndexer***. Following these changes, our dataset became ready for performing predictive analytics by implementing machine learning algorithms. To do so, we utilized suitable packages such as ***pyspark.ml*** and ***sklearn***.

We would like to conclude our summary by emphasizing that overall, this assignment was a great exercise for learning and getting acquainted with a great new tool – Apache Spark.