CSCI 5040 Programming Assignment 2: Walmart

Learning Outcomes

- Open a CSV file in Spark
- Convert a DataFrame to a DataSet
- Create a dataset as an SQL table.
- Discover properties of a dataset using Spark's API

Required Reading

Chapter 2 and 3 in the textbook

Motivation!

I found a dataset on Walmart's sales data for 45 different stores. Let's see what we can learn, shall we?

Here's what we learned in the September 11th class:

- DataFrames are good for working with columns of data. (grouping and summary statistics)
- DataSets are good for working with rows (i.e. records) of data. (filtering and maping)

The purpose of using a DataSet rather than a DataFrame is that DataSets compile much faster than DataFrames, they are type-save, and they allow for manipulations on objects.

Instructions

The CSV file provided has five columns. Those columns are presented along with their Scala data types.

- Store: BigInt
- Dept: BigInt
- Date: String
- Weekly Sales: Double
- IsHoliday: Boolean

Try to answer this questions without resorting to SQL. Answer each question along with this document with the code you used to solve this. For the last question, it may take multiple lines of code and some scratch pad work.

1. What is the largest weekly sales in the entire file? For this, I used a DataFrame and the code on page 26.
2. What was the store and department number of the store with the highest weekly sales? For this, I used a DataFrame and the code on page 27.

3. What was the average weekly sales for each of the 45 stores? For this, you'll need to display all 45 stores in order of store number (so, 1 to 45) and their store average. I did this with a data frame and the "avg" function, which is found in "org.apache.spark.functions.avg".
4. How many records have negative sales? I did this using a DataFrame and the "where" clause.
5. How many records occur in the year 2011? I did this using a DataFrame using the "where" function similar to page 40 in the book. I came up with a different answer using a dataset and the "filter" method. In either case, the "less than or greater to" operator is "<=" and the "greater than or equal to" operator is ">=".
6. Manipulate the dataset so that each weekly sales value is represented in thousands instead of single dollars. Display the first 20 results. I did this with the map method.
7. Filter and manipulate the walmart table to remove all records with positive sales, then convert those negative sales figures into positive values. Display the first 20 results.
8. This will take multiple queries to answer. Are down days more likely or less likely to happen on a holiday? We know more down days will happen on non-holidays simply because there are more non-holiday days. Divide the number of down days (days with negative sales) on holidays by all holidays. Divide the number of down days on non-holidays by all non-holidays. Which percentage was higher? There's your answer.

Notes and Comments

Build a text file of the commands that you used to produce the results. You may have to use a combination of DataFrames and Datasets to accomplish your task.