Correlations and aggreations between stocks using Spark

2IMD15 Data Engineering - Milestone 2

Group 1: Daniel Teixeira Militao (1486314), Manon Wientjes (1398903), Pawel Budzynski (1511734),
Tong Zhao (1416790), Ugne Laima Ciziute (1495186)

Dataset and correlation functions

The dataset we used is stock data from January until mid April 2020[1]. We use them to find interesting correlations between different stocks. We ended up picking the opening price of stocks for the months of February and March on an hourly basis for working hours and days which allowed us to have 666 vectors with 369 dimensions each. We used Pearson correlation with average as an aggregation function and Total correlation with identity function as aggregation.

Pre-processing

In order to be able to compute correlations, the vectors need to be of the same length. Hence, we needed to filter out some stocks and handle missing values by interpolating and extrapolating. Pre-calculations are done to speed up computation of correlations. The pipeline in Figure 1 depicts the steps taken during pre-processing.

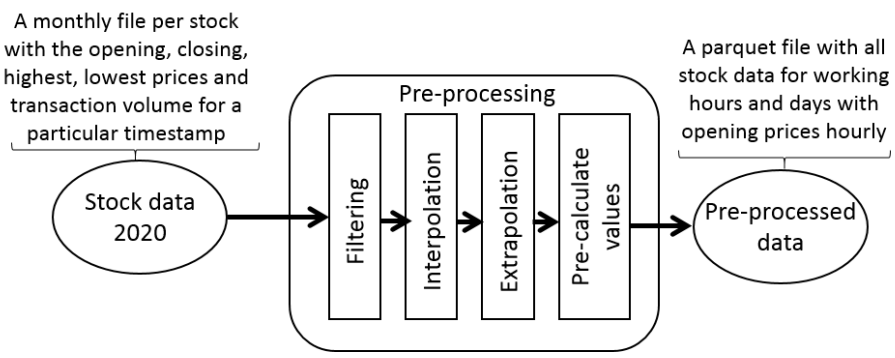

Figure 1: Pre-processing pipeline.

Correlations calculation

Correlations are calculated according to Figure 2. First groups of stocks are created to ensure that each worker gets a similar workload. To prevent redundant comparisons between group combinations, duplicate combinations are filtered out. In the end, $flatMap$ is used to calculate the correlations between every combination.

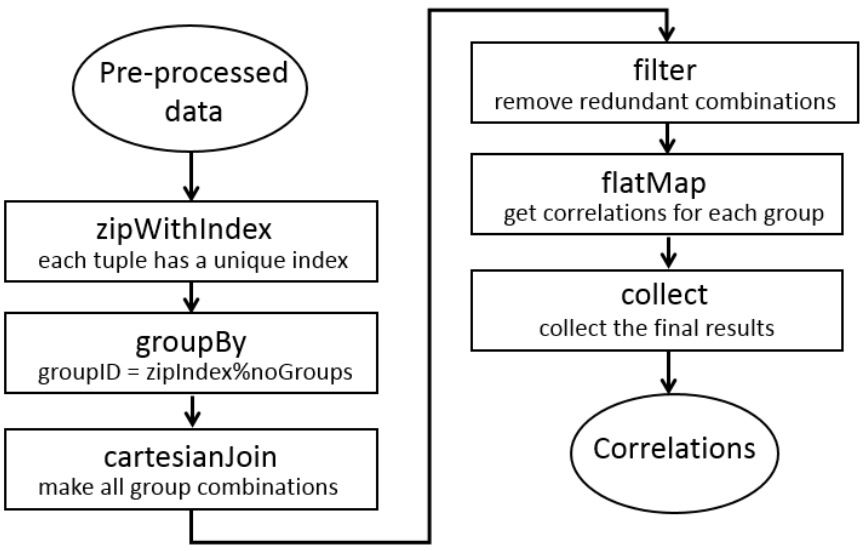

Figure 2: Correlations calculation pipeline.

Experimental performance

We ran experiments for $p = 3$ on the number of vectors used to see how the running time would behave, the results of which can be seen in Table 1. Furthermore, for $p = 4$ using 111 vectors with 198 dimensions Pearson took 1,653 seconds and using 185 vectors with 198 dimensions TC took 1,778 seconds.

Table 1: Running time (sec) for increasing no. of vectors for $p = 3$.

Number of vectors	222	444	666
Pearson Correlation (369 dim.)	95	590	1,928
Pearson Correlation (198 dim.)	79	371	1,230
Total Correlation (198 dim.)	69	547	1,844

After running experiments, it could be concluded that for the same data complexity Pearson is faster than TC, which is inline with our expectation. Further experiments on a more powerful cluster on a Microsoft Azure server with 11 executors with a total of 40 cores and 28GB RAM were performed. The reason for these extra experiments was to see how our application would scale and if it still efficiently used all cores and we are happy to report that it does. Table 2 shows the results.

Table 2: Running time (sec) for increasing no. of vectors on the large cluster for $p = 3$.

Number of vectors	222	444	666
Pearson Correlation (369 dim.)	81	375	1,163
Total Correlation (369 dim.)	57	421	1,481

Insights

All the results of the top ten Pearson correlations are above 0.999. The top one refers to currency exchange markets. Next 3 positions seem to look quite interesting as they show strong correlation between value of a company located in Geneva and different exchange rates of Swiss franc. Total correlation gives way more interesting results. Amsterdam_MT and Madrid_MTS refer to the same company but its high correlation with Xetra_750000 is an interesting finding. The second and third rows show stocks that behaved very similar during this period.

Table 3: Ten highest Pearson correlations for March+February opening price

Pair		Value
Forex_AUD	(CME_6J, CME_6A)	0.99997
Mailand_STM	(Paris_STM, Forex_ZARCHF)	0.99997
Mailand_STM	(Paris_STM, Forex_TRYCHF)	0.99997
Mailand_STM	(Paris_STM, Forex_MXNCHF)	0.99997
Paris_STM	(CME_6J, Mailand_STM)	0.99997
Mailand_STM	(Paris_STM, CME_6M)	0.99997
Mailand_STM	(Paris_STM, CME_6J)	0.99997
Paris_STM	(CME_6M, Mailand_STM)	0.99997
Paris_STM	(Forex_MXNCHF, Mailand_STM)	0.99997
Mailand_STM	(Paris_STM, Forex_HUFCHF)	0.99997

Table 4: Ten highest Total Correlation correlations for March opening price

Pair			Value
Amsterdam_MT	Xetra_750000	Madrid_MTS	4.0257
Paris_CNP	Madrid_ACS	London_CCL	4.0158
London_ICP	Madrid_ACS	London_CCL	4.0066
Amsterdam_MT	Paris_CNP	Madrid_MTS	4.0039
Paris_CNP	London_ICP	London_CCL	4.0023
Amsterdam_MT	Madrid_ACS	Madrid_MTS	3.9748
Paris_CNP	London_ICP	Madrid_ACS	3.9733
Amsterdam_MT	Viena-Exchange_AT0000743059	Madrid_MTS	3.9523
Paris_STM	London_SMIN	Mailand_STM	3.9442
Amsterdam_MT	Paris_CS	Madrid_MTS	3.9441

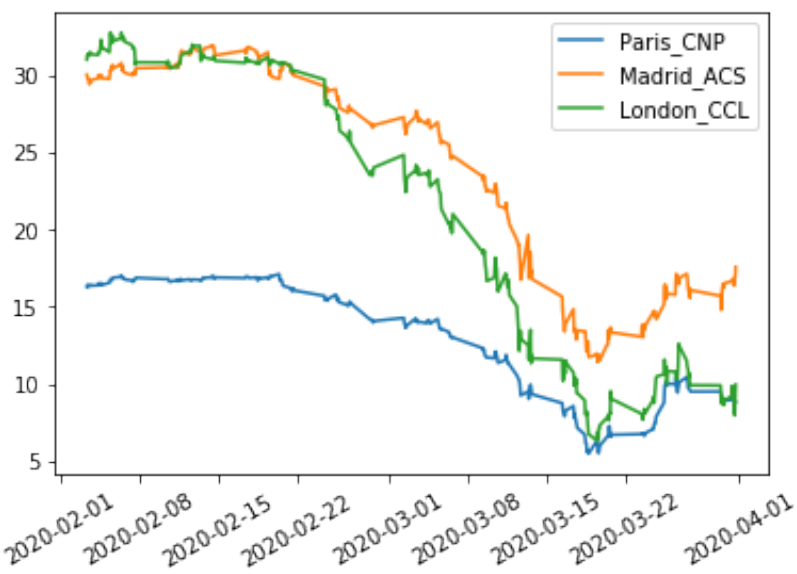

Figure 3: Example of interesting correlation found by Total Correlation.

[1] https://canvas.tue.nl/courses/10287/files/2383551/download?download_frd=1
Video presentation: https://drive.google.com/drive/folders/1UdWW7MyOGa_bWTe7GlDN_zmjLIgE_br-