Name: Mohammad Khan
Student Number: 150952987

Bitcoin analysis using big data jobs

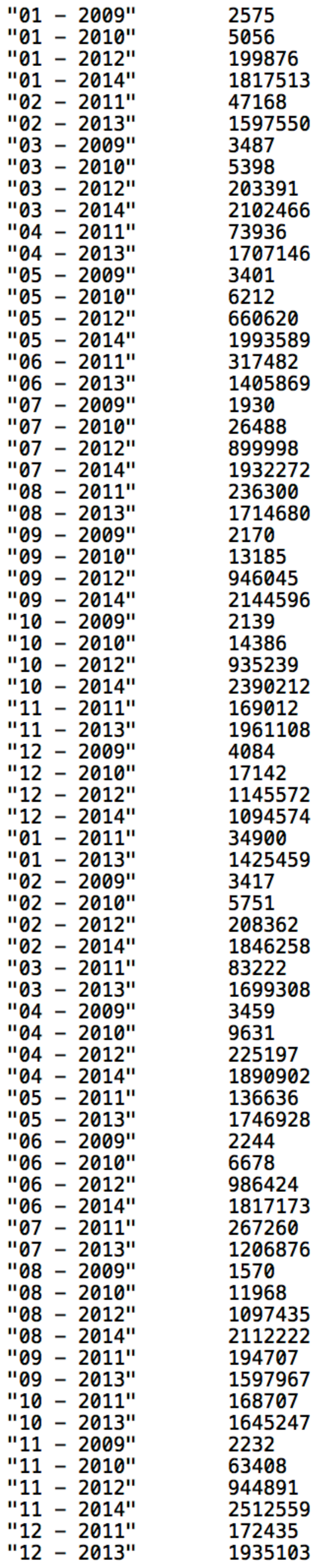
Part A

For this part I need to create a bar chart that shows the all the transactions that occurred from every month in the dataset. For this I used the MapReduce technique because it goes through every individual transactions committed and also by getting the year and the month of each transactions.

The pair that I will emit will consist of the key, representing month and year, and value, being one. A combiner was added to make things quicker, which does the same thing as the reducer, which is adding all the values emitted for each key in each mapper. An example of this is if you had 7 “01-2009” produced by a mapper, then the combiner will produce “01-2009”, 7, instead of 7 1’s to the reducer.

A try and except within the mapper was used to make sure there was not any failures in any lines of code. The main idea of the mapper is to check any lines passed, converting the time into a month and year, in string format. The following screenshot is the output from running the job.

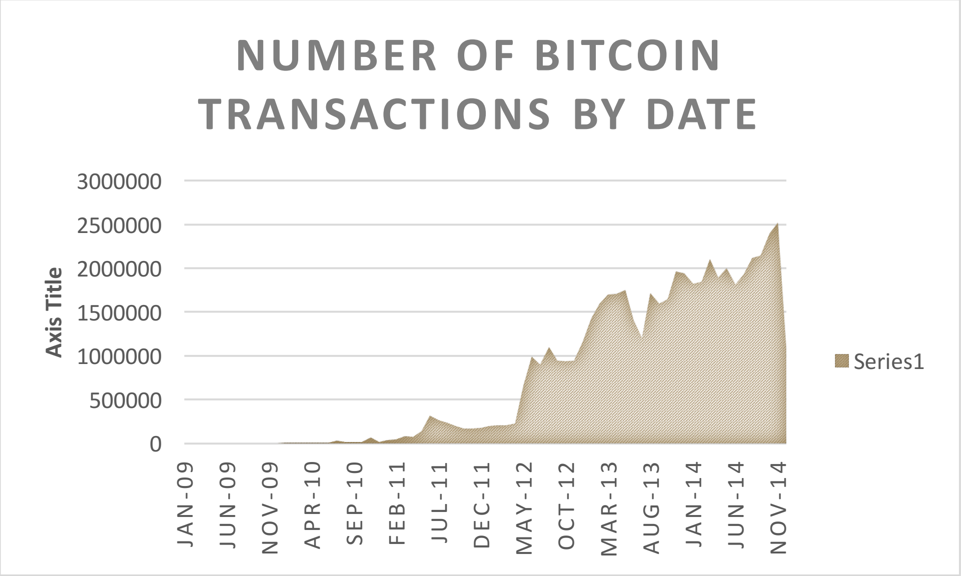
Mohammad M Khan
150952987



```
"01 - 2009"        2575
"01 - 2010"        5056
"01 - 2012"        199876
"01 - 2014"        1817513
"02 - 2011"        47168
"02 - 2013"        1597550
"03 - 2009"        3487
"03 - 2010"        5398
"03 - 2012"        203391
"03 - 2014"        2102466
"04 - 2011"        73936
"04 - 2013"        1707146
"05 - 2009"        3401
"05 - 2010"        6212
"05 - 2012"        660620
"05 - 2014"        1993589
"06 - 2011"        317482
"06 - 2013"        1405869
"07 - 2009"        1930
"07 - 2010"        26488
"07 - 2012"        899998
"07 - 2014"        1932272
"08 - 2011"        236300
"08 - 2013"        1714680
"09 - 2009"        2170
"09 - 2010"        13185
"09 - 2012"        946045
"09 - 2014"        2144596
"10 - 2009"        2139
"10 - 2010"        14386
"10 - 2012"        935239
"10 - 2014"        2390212
"11 - 2011"        169012
"11 - 2013"        1961108
"12 - 2009"        4084
"12 - 2010"        17142
"12 - 2012"        1145572
"12 - 2014"        1094574
"01 - 2011"        34900
"01 - 2013"        1425459
"02 - 2009"        3417
"02 - 2010"        5751
"02 - 2012"        208362
"02 - 2014"        1846258
"03 - 2011"        83222
"03 - 2013"        1699308
"04 - 2009"        3459
"04 - 2010"        9631
"04 - 2012"        225197
"04 - 2014"        1890902
"05 - 2011"        136636
"05 - 2013"        1746928
"06 - 2009"        2244
"06 - 2010"        6678
"06 - 2012"        986424
"06 - 2014"        1817173
"07 - 2011"        267260
"07 - 2013"        1206876
"08 - 2009"        1570
"08 - 2010"        11968
"08 - 2012"        1097435
"08 - 2014"        2112222
"09 - 2011"        194707
"09 - 2013"        1597967
"10 - 2011"        168707
"10 - 2013"        1645247
"11 - 2009"        2232
"11 - 2010"        63408
"11 - 2012"        944891
"11 - 2014"        2512559
"12 - 2011"        172435
"12 - 2013"        1935103
```

Mohammad M Khan
150952987

A bar graph was created from the data above, its formed from the number of transactions from the start to the end of the dataset.



From the bar chart above we can see that: from January 2009 - February 2011 there was a very small number of transactions made using bitcoin. Conversely, from July 2011 - May 2012 there was a steady increase of the number of transactions. After May 2012 there was a substantial increase in transaction number and since then a steady increase from 1 million to 2.5 million in November 2014. This shows an overall correlation of people using bitcoins much more over time.

Mohammad M Khan
150952987

Part B

For the second part I need to filter out the top 10 donors of the WikiLeaks bitcoin wallet. For this part I used Spark as it is easy to use joins.

Firstly, the vout data was filtered so that transactions with only WikiLeaks wallet as the receiver are shown. This was done by splitting the line and selecting the 4th element (wallet ID) in the line and checking it with the WikiLeaks wallet ID using the filter transformation.

After that, the map transformation was used to restructure the data into the form where you have the key (hash of the transaction) and the value (null as you don't need anything else from the vout line). The vin data was prepared by splitting each line and then restricting the vin data to a key (being txid which will be used to join with the hash on vout) and value (txhash, vout) pair. This is going to be important to get information of the donors, as the value gives data on previous transactions.

Step 4, a join was used between WikiLeaks vout data and the whole vin data, this will grab the hash of the previous transaction. Next, a key, value structure is made but this time it will be the key being represented by the tuple of (txhash/hash, vout/n).

Next, a second join is used to get the vout data and again it will reconstruct the data so that the key is the same as the tuple above and the value is the wallet ID and number of bitcoins sent by a specific donor.

A map transformation is done to get the id of the wallet and the number of bit coins donated. To do this the type is changed to float so that I can manipulate it in to sorting. Number of donors are counted and partial counts and a list is made of donors once with total donations. Once the data is sorted, the top 10 are printed. Which is shown below in GBP:

ID of Wallet	Amount (in BTC)	GBP in millions
17B6mtZr14VnCKaHkvzqpkuxMYKTvezDcp	46515.1894803	143.3
19TCgtx62HQmaaGy8WNhLvoLXLr7LvaDYn	5770.0	17.8
14dQGpcUhejZ6QhAQ9UGVh7an78xoDnfap	1931.482	5.9
1LNWw6yCxkUmkhArb2Nf2MPw6vG7u5WG7q	1894.37418624	5.8
1L8MdMLrgkCQJ1htiGRAcP11eJs662pYSS	806.13402728	2.5
1ECHwzKtRebkymjSnRKLqhQPkHCdDn6NeK	648.5199788	2.0
18pcznb96bbVE1mR7Di3hK7oWKsA1fDqhJ	637.04365574	2.0
19eXS2pE5f1yBggdwhPjauqCjS8YQCmnXa	576.835	1.8
1B9q5KG69tzjhqq3WSz3H7PAxDVTAwNdbV	556.7	1.7
1AUGSxE5e8yPPLGd7BM2aUxfzbokT6ZYSq	500.0	1.5