Storage and Retrieval System for Bitcoin Traffic and Price Analysis

Rohan Thakur, Tarun Chopra and Sashi Gandavarapu

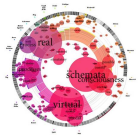
Objectives

Real time analysis Bitcoin twitter traffic vs Bitcoin transaction traffic (from Blockchain)

Key Questions: Is amount of twitter traffic correlated with number of bitcoin transactions?
Is the sentiment on twitter correlated with number of bitcoin transactions?

Correlation with commodities such as Oil, Gold, Silver and Copper

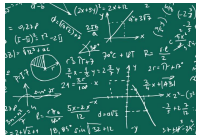
Key Questions: Has bitcoin become synonymous with trading of commodities?
Which commodities have most correlation with historic bitcoin prices?



Model and visualize data in a dashboard, automatic refresh for real-time updates.

What is Bitcoin?

- Virtual Currency
- De-centralized.
- Online traded.

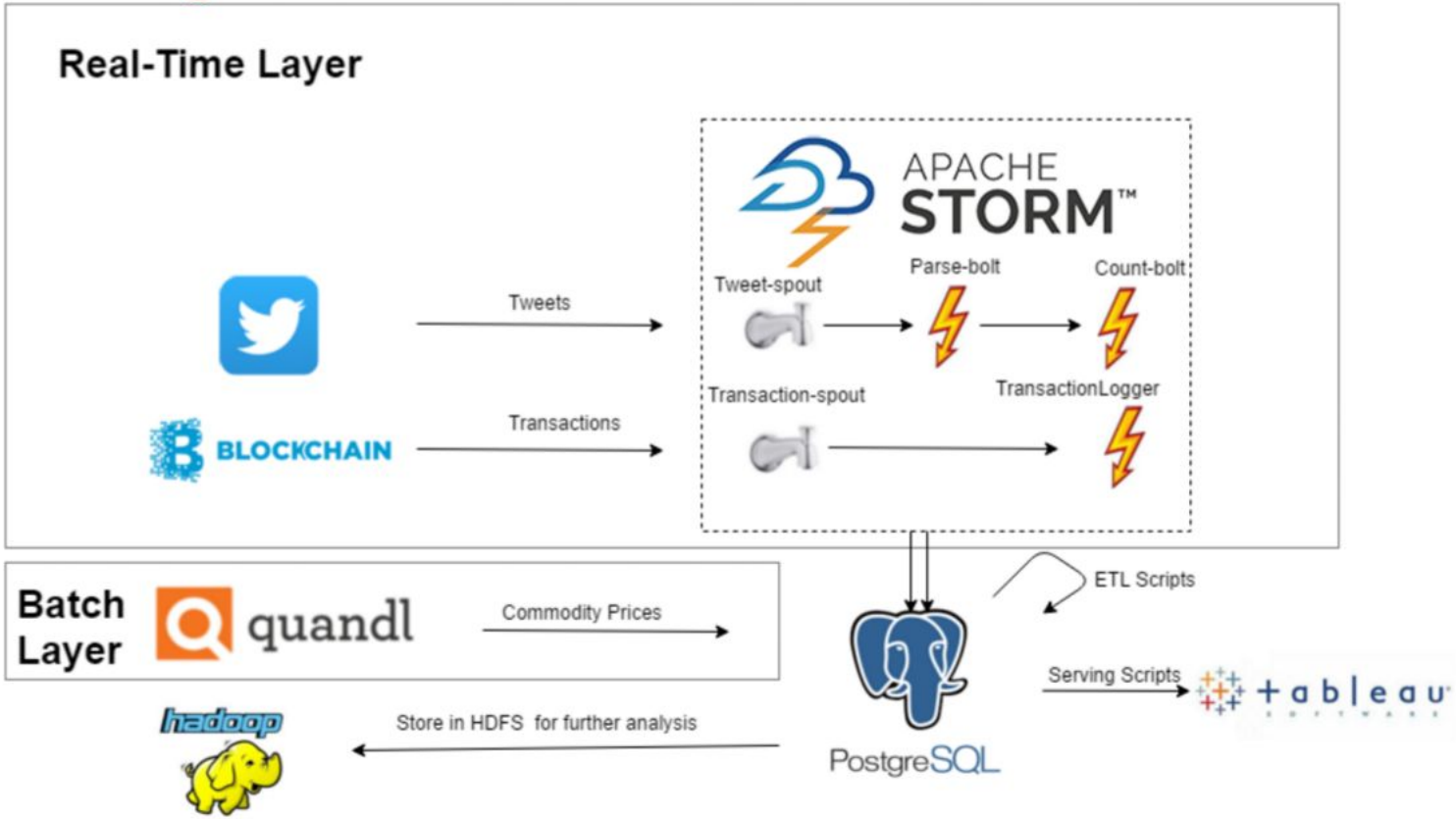


- Complex Math and Crypto Algorithms

- 15 Millions Bitcoin in market with each valued at $420

- Public verifiable ledger called Blockchain.
- Prevents fraudulent acts.

Our Project: Lambda Architecture



Lambda Architecture (Cont'd)

- Real-Time Data: Apache Storm
  - We had two spouts to stream our data: One monitoring the twitter stream, and one monitoring the latest block released on the blockchain
  - For tweets we logged a timestamp and the actual tweet. For transactions we logged the timestamp for when the transaction was verified, transaction id, and hash key of the block
  - We conducted hourly sentiment analysis and traffic comparison of the tweets versus bitcoin
- Batch Data: Python
  - Daily price data for chosen commodities and bitcoin were queried and downloaded as dataframes, exported to csv, and inserted into our database periodically.
  - Batch scripts (python and SQL) will be run every day through a shell script to compute aggregate stats
- Database: Postgres
- Dashboard and Final Data Model: Tableau
  - We were able to connect our postgres tables to Tableau using Cloudera and a Postgres Server
  - Data analysis and visualization was conducted in Tableau

Volume, Velocity and Variety

- Volume
  - We were able to log over 500k blockchain transactions and over 2k bitcoin tweets over 3-4 days
  - For the chosen commodities, we pulled daily price data from 1/1/2010 until current date from Quandl, much smaller in size than our streaming data
- Velocity
  - Twitter stream has high variance and is completely dependent on interest in bitcoin at any given time. Varies from several tweets per second to less than a tweet per hour.
- Variety
  - Apart from blockchain data which was obtained by parsing JSONs, we mostly dealt with simple strings and dates, which required little transformation

Data Collected - Snapshots

timestamp	tweet
4/14/2016 3:15:24 AM	I just "deposited" #bitcoin into my @circlepay for the first time. Feels weird. :-/. like losing your virginity. #CyptoCurrency
4/14/2016 3:18:01 AM	#Bitcoin conference #russia results: the ethereum boom is here #crypto #cryptocurrency https://t.co/T8lyXzcvF0 https://t.co/WKuMRMrbHk
4/14/2016 3:18:20 AM	Current #Bitcoin Price 1 BTC = $424 USD: This may be the luckiest day of your life https://t.co/lX6heyMDGQ https://t.co/sBUfaq2U2N
4/14/2016 3:18:32 AM	RT @TheBitCoinPRO: This Former Regulator Is Now a Bitcoin Backer - Fortune - https://t.co/jm9muo7I0R
4/14/2016 3:18:54 AM	My monster has 1,000 hp left! Ive earned a total of 165,760 satoshi https://t.co/JVEMkLP4jj #monstercoingame #Bitcoin
4/14/2016 3:20:03 AM	RT @McEnroeDenny: Sending bitcoin without a QR code - paper wallet https://t.co/PxtGHT8436 #bitcoin#crypto#btc
4/14/2016 3:22:49 AM	RT @RedditBTC: Hashpower is the Ends, Bitcoin is the Means https://t.co/IAfpT0vjLp
4/14/2016 3:30:06 AM	RT @wildanimal402: #Airbnb just acquired a team of #bitcoin and #blockchain experts https://t.co/icrMrSnhC1 https://t.co/6UV9qqv3ck
4/14/2016 3:31:41 AM	Current #Bitcoin Price 1 BTC = $424 USD: What you can learn from 4BC https://t.co/oPw0TmEb7s https://t.co/8VI8MTztsn
4/14/2016 3:32:15 AM	Re: Welcome To The Blockchain: Quote from: Icathia on April 13, 2016, 07:38:48 AMSo nice to hear that... https://t.co/BQBXVXQDjf #bitcoin

Recent Tweets Containing Bitcoin

10,000 rows

hashkey	timestamp	transaction_id
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091602
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091348
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091435
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091084
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091063
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091306
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091175
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091361
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091178
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091214
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141090820
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141090770
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091076
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141090822
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091246
0000000000000000000f2edbcfdf1fa13c4a0dd7c665b02163f8c4d5c69e2001d	4/14/2016 3:02:50 AM	141091068

Bitcoin Transaction logs

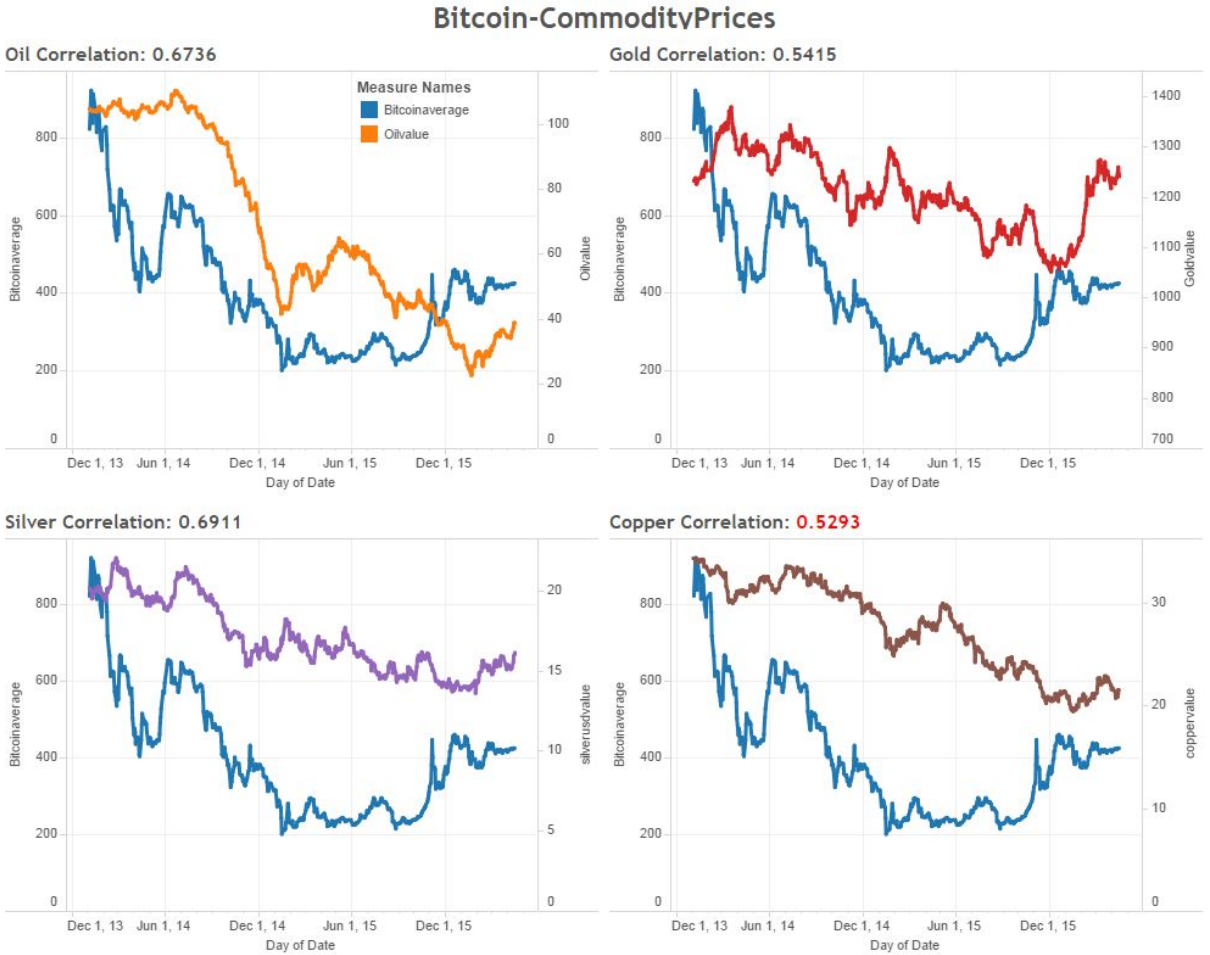
date	bitcoinaverage	goldvalue	oilvalue
12/16/2014	337.16	1,199.250	55.910
12/17/2014	324.07	1,199.000	55.640
12/18/2014	313.60	1,210.750	56.300
12/19/2014	313.11	1,197.500	55.520
12/22/2014	328.27	1,195.250	56.900
12/23/2014	334.22	1,179.500	55.590
12/24/2014	331.53	1,177.000	56.030
12/29/2014	317.14	1,194.000	54.440
12/30/2014	315.59	1,186.500	52.390
12/31/2014	314.44	1,199.250	52.000
1/2/2015	315.45	1,184.250	51.780
1/5/2015	271.78	1,192.000	48.870
1/6/2015	277.04	1,211.000	46.570
1/8/2015	289.61	1,206.500	45.680
1/9/2015	285.04	1,211.250	45.190
1/12/2015	269.63	1,222.000	43.550
1/13/2015	242.22	1,239.000	41.500
1/14/2015	200.28	1,228.750	41.650

Commodity Closing Prices

Dashboard 1:

Dashboard 1 shows historic price correlation between commodities and bitcoin closing price.

This visualization updates daily after getting a feed from quandl API.



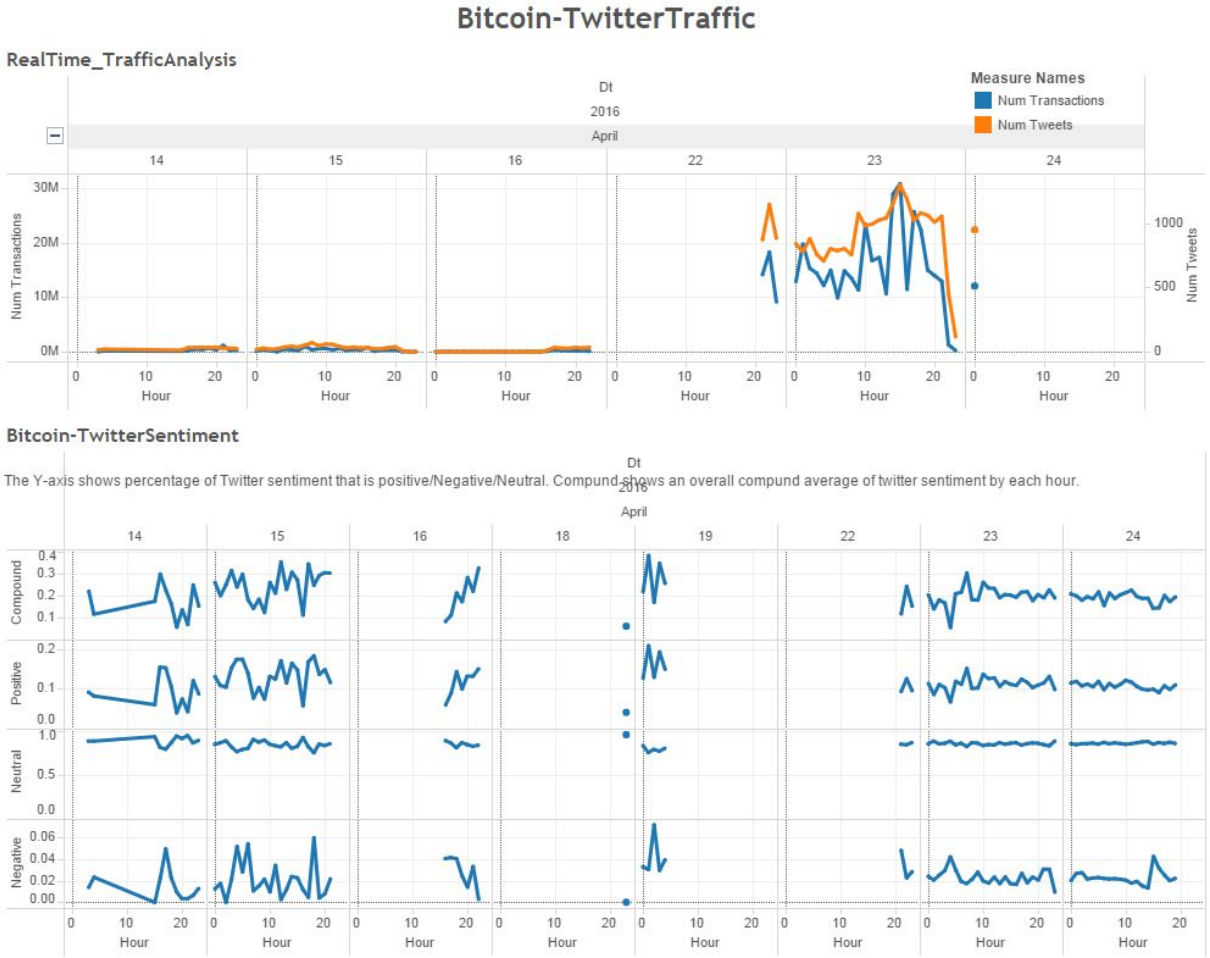
Dashboard available at https://public.tableau.com/profile/publish/Bitcoin-Commodity-TwitterTrafficCorrelation/Bitcoin-TwitterTraffic#!/publish-confirm

Dashboard 2:

Top portion of dashboard 2 shows hourly correlation of volume of twitter traffic to volume of bitcoin transactions.

Bottom portion of dashboard 2 shows hourly correlation of sentiment of twitter traffic to volume of bitcoin transactions.

This visualization updates hourly after getting a feed from twitter API and blockchain API.



Dashboard available at https://public.tableau.com/profile/publish/Bitcoin-Commodity-TwitterTrafficCorrelation/Bitcoin-TwitterTraffic#!/publish-confirm

Conclusions

1. We found that bitcoin prices have a moderate to strong correlation to some of the commodities we explored. Bitcoin price correlation with
    a. Gold: 0.54
    b. Oil: 0.67
    c. Silver: 0.69
    d. Copper: 0.53
2. In the few days of twitter data collected, the dashboards show a reasonable correlation between the volumes of traffic. But, to draw a definite conclusion, the system needs to be running for a few more days.
3. Most of the twitter sentiment during the data collection period is neutral (about 85%).

Limitations, Challenges, Extensions

- Challenges:
  - Twitter data collection seems to be missing some tweets that have the keywords in them: SOLVED !
  - Blockchain API is sometimes prone to failure, when it loses track of the structure of the blockchain.

    This can lead to some pauses in our data, which leads to some gaps in time for data collected. In real-world application, this would be monitored closely and brought up and running ASAP
  - Compare_traffic sql needs to be optimized in order to be run quickly. Due to the huge amounts of tweet and bitcoin transaction data, currently compare_sql query is taking a long time to complete.
- Limitations:
  - Archive the postgres database into hive/hadoop architecture for later analysis purposes.
- Extensions:
  - Time series analysis on bitcoin prices and commodity prices(as opposed to a simple correlation metric)