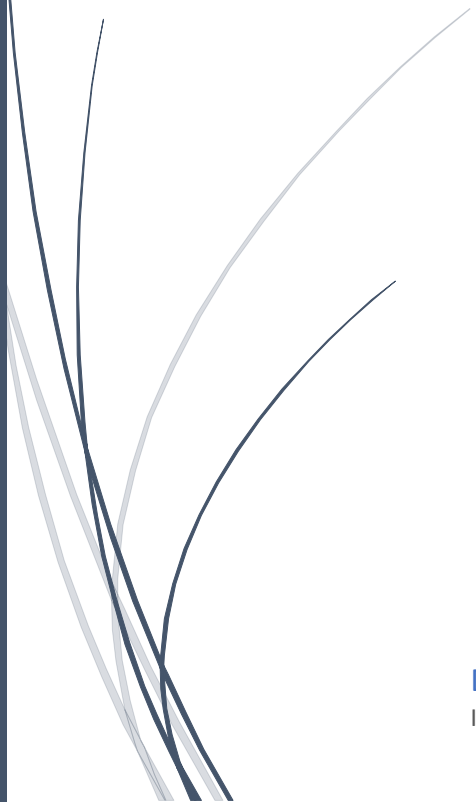
9/1/2019

Requirement Gathering V0.3



Dr Avinash kumar singh & Arvind narayanan
INTAIN PVT. LTD.

AI for Blockchain

Problem Definition (Storage)

Given the preferences, and the historical input variables, we need to rank the storage token. We can set the preference values between [1-10] while the input variables are available in terms of price, processing fee, transection time, etc. for each storage token.

The details about the input variables are given in table 1. In total we have 10 variables, these variables are further mapped with respect to the preference. For example, fees and price are mapped w.r.t to cost/budget while speed is expressed in transection time or block time.

Table 1. Input variable details

Preference	Attribute Name	Attribute symbol	Unit
Cost/budget	Price	a_1	USD ($)
	Processing Fees	a_2	USD ($)
Speed	Transection/Block Time	a_3	Minutes
Security	Security (FCAS Score))	a_4	No Unit
Health	Transection Volume	a_5	Number
	Exchange Volume	a_6	Number
Component Specific variables	Bandwidth-up	a_7	MBPS
	Bandwidth-down	a_8	MBPS
	Storage-total	a_9	TB
	Storage-left	a_{10}	TB

If the rating of the preference is high, it shows the high importance on the other hand if the rating is low, it shows lower importance for that preference (choice) while optimizing it. The output variable is denoted in below table. Given the preference a choice will be made for each output variable. We assume that the output variables are independent, and their choice do not impact to each other.

Table 2. Output variable details

Storage	Domain
Genaro (GNX)	d_1
Storj (STORJ)	d_2
Filecoin (FIL)	d_3
Sia (SC)	d_4

Dataset Representation

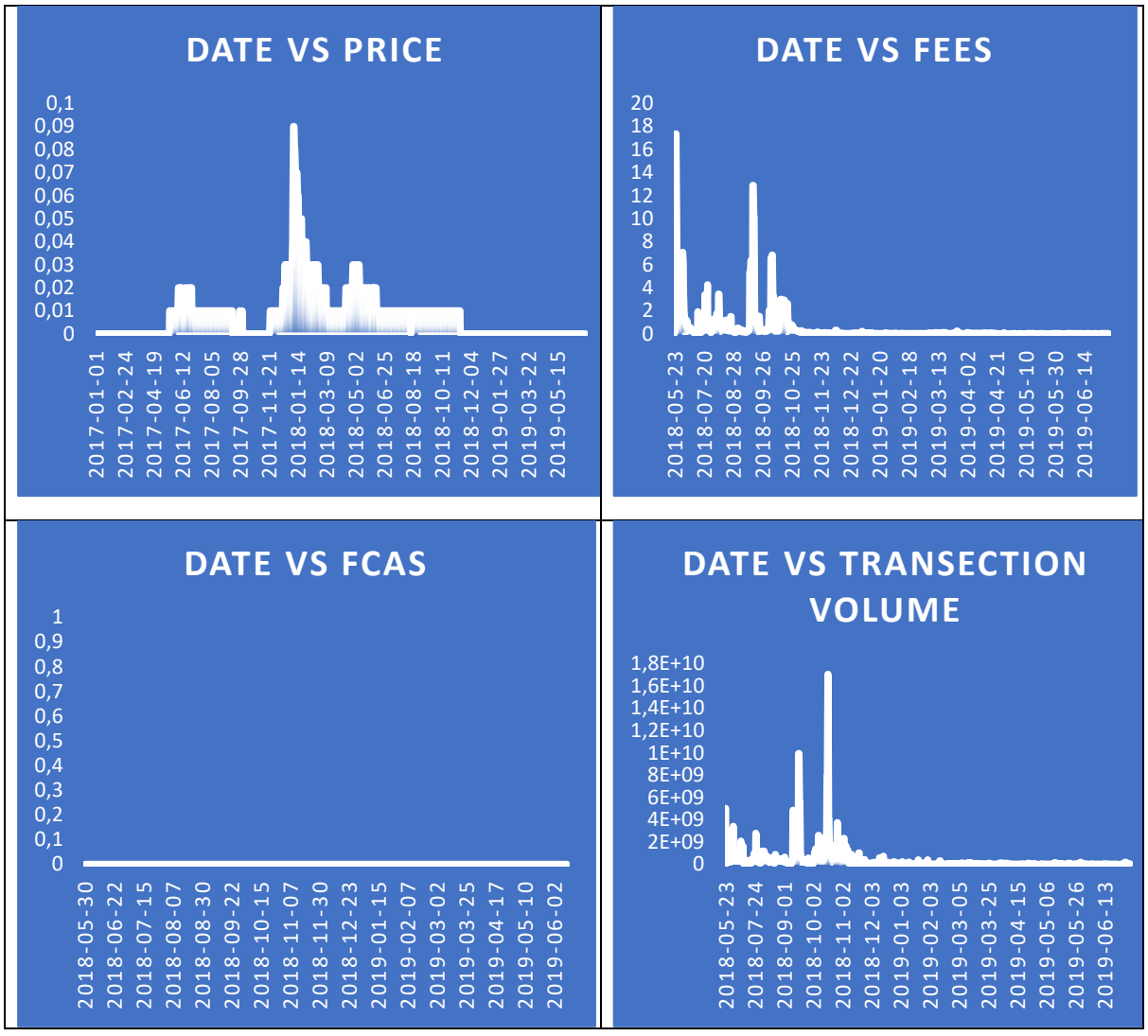
The complete set of input variables and their respective values are represented in table-2. Table-2 represents the values for SIA, the other storage tokens have the same data appearance. The unknown values are represented as “MV” while the values who are not available are represented as NA. Most of the variables are discrete in nature and captured against the date. The distribution of the data and how it changes over the time is depicted in Table-4, Fig-1, Fig-2, Fig-3,…, Fig-8 respectively for each input variable.

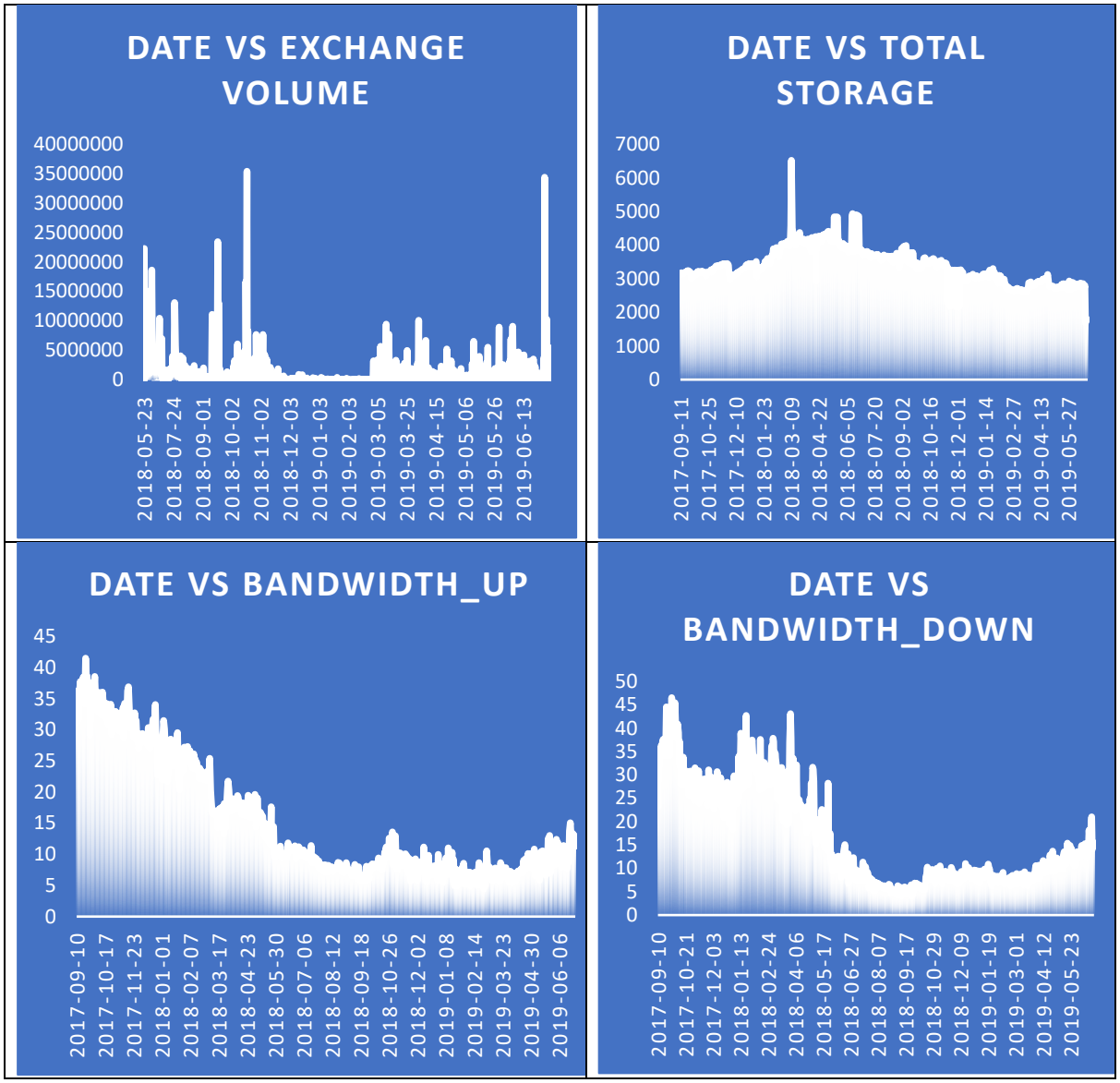
Table 3. Historical data for SIA

Date	Price	Fees	Blocktime	FCAS	Trans_vol	Exch_vol	Band_up	Band_down	Stor_tot	Stor_left
2018-06-01	0,02	2,73356		NA	311449142	1485207,5	10,5	12,3	3888	
2018-06-02	0,02	2,95644		NA	418182067	2276746,2	10,2	11,5	3950,6	
2018-06-03	0,02	1,55572		NA	223518252	1165243,1	10,6	11,6	3982,6	
2018-06-04	0,02	1,73695		NA	214125735	1095537,9	9,98	11,1	3983,6	
2018-06-05	0,01	3,69859		NA	465117361	2211251,7	11,3	12	3900,4	
2018-06-06	0,02	5,36227		NA	685252282	3448963,7	10,2	9,59	3918,2	
2018-06-07	0,02	4,83004		NA	605238381	3018698,9	9,34	10,9	3780,4	
2018-06-08	0,02	7,06507		NA	3422832284	18655976	10,1	12,6	3832,2	
2018-06-09	0,02	6,09654		NA	1133853781	5606328,5	9,91	12,1	3789,1	
2018-06-10	0,02	3,5038		NA	772831504	3898757,2	9,82	12,1	3824	
2018-06-11	0,01	1,86457		NA	561930013	2440782,5	10,1	10,2	3857,7	
2018-06-12	0,01	1,22175		NA	437111472	2060368,6	9,75	10,6	3888	

*The value for the block time and storage left are given against the time, we didn't understand the data so didn't put it here.

Table 4. Variables distribution graph





Problem Formulation

For the simplification of the variables, we now denoted the preference variable by some variables as represented in below table. Each variable can have the value between 1 to 10.

Table 5. Preference variable representation

Preference	Symbol
Cost/budget	w_1
Speed	w_2
Security	w_3
Health	w_4
Component Specific variable	w_5

COST Function:

We can define the cost function in terms of linear combination of preference and input variable and then we can minimize/maximize the cost. The expression is given below.

$$W = w_1(a_1 + a_2) + w_2 a_3 + w_3 a_4 + w_4(a_5 + a_6) + w_5(a_7 + a_8 + a_9 + a_{10})$$

Hypothesis-1

Approximate each input variable and then find the minimum cost of each, the assumption is "the local minima would lead to the global minima".

Hypothesis-2

Minimize the overall cost function to achieve the global minima

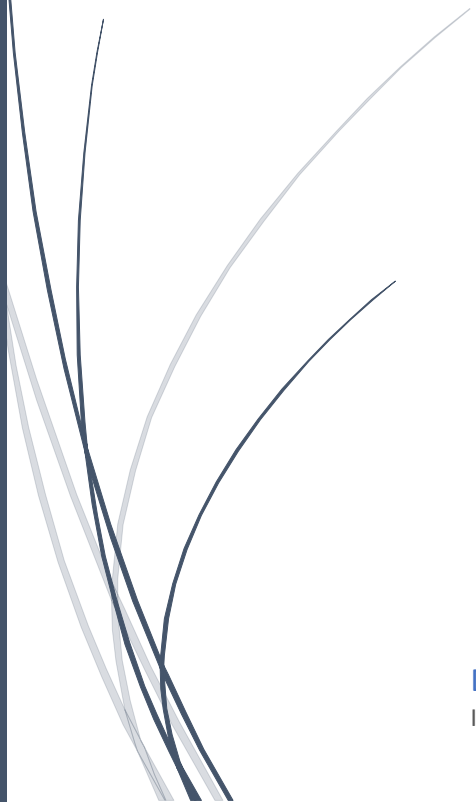
Evaluation Criteria

We would take continuity as a performance criterion. For example, we have a month data to perform the testing, we predicted the optimal choice at 1st day, keeping the fact that it would stay optimal for the next one month. Later we do prediction for every day, the continuity of the decision can be evaluated by how many times it appeared in that month divided by number of days in month.

$$contuity = \frac{number\ of\ time\ same\ token\ apperared\ in\ the\ month}{number\ of\ days\ of\ the\ month}$$

8/22/2019

Requirement Gathering V0.2



Dr Avinash kumar singh & Arvind narayanan
INTAIN PVT. LTD.

AI for Blockchain

Problem Definition (Storage)

Given the input attributes price, fees, block time, security etc. and the preference of the optimization levers find the optimal value of storage, computing and database.

In order to precisely define the problem, we used the below tabular representation. In total we have 8 attributes. We added one additional variable to capture more insight about the data. Unit is used to capture the measurement of the variable. A variable can have the discrete value, or it can have the categorical value. All the variables have discrete values.

Table 2. Input variable details

Attribute Name	Attribute symbol	Unit
Price	a_1	USD ($)
Fees	a_2	USD ($)
Transection/Block Time	a_3	Minutes
Security (FCAS Score))	a_4	No Unit
Transection Volume	a_5	Number
Exchange Volume	a_6	Number
Bandwidth	a_7	MBPS
Storage	a_8	GB

We have excluded the following variables due to their non applicability for solving the storage issue. They are listed below.

- Size
- CPU
- RAM
- Location

The input variables are specified in the specification or a configuration sheet, while the preference variables are expressed in terms of rating. If the rating is high, it shows the high importance on the other hand if the rating is low, it shows lower importance for that preference (choice) while optimizing the it. The preference table is given below. Rating can be defined as low, high, medium or on the scale of 1-10.

Table 2. Preference Table

Optimization Lever	Rating
Cost/Budget	[a-b]
Speed	[a-b]
Security	[a-b]
Health	[a-b]
Component Specific Variables	[a-b]

The output variable is denoted in below table. Given the attribute values a choice will be made for each output variable. We assume that the output variables are independent, and their choice do not impact to each other.

Table 3. Output variable details

Storage	Domain
Genaro	d_1
Storj	d_2
Filecoin (FIL)	d_3
Sia (SC)	d_4

If we keep the above convention, our dataset would be represented as given below.

Table 4. Dataset Representation

Sample	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	Storage
s_1	-	-	-	-	-	-	-	-	$[d_1 - d_4]$
s_2	-	-	-	-	-	-	-	-	$[d_1 - d_4]$
s_3	-	-	-	-	-	-	-	-	$[d_1 - d_4]$
s_4	-	-	-	-	-	-	-	-	$[d_1 - d_4]$

Since, we would be using supervised learning to solve the problem, we would be needing the decision variable $[d_1 - d_4]$ for every sample. Every sample is represented by the collection of 8 attributes or a 12-dimensional vector ($S \in \mathbb{R}^8$).

Data Understanding

For the storage we have four choices it can be either Genaro, Storj, Filecoin or Sia. The relation between the input variables and the actual used variables (in the shared files) are summarized below.

Table 5. Data Representation for Genaro

S.No	Attribute Name	Where to find
1	Price	NAStorj.csv and NAGenaro-network.csv Use 'AVG' value
2	Fees	Flipside folder, final_daily.csv. Use 'ALL'.
3	Block Time	Blocktime.csv (coin folder)
4	FCAS Score	Flipside folder, fcas.csv
5	Transection Volume	Flipside folder, final_daily.csv. Use Token volume
6	Exchange Volume	Flipside folder, final_daily.csv. Use usd volume.
7	Bandwidth	up.csv and down.csv (coin folder)
8	Storage	Total.csv (coin folder)

Storj and Genaro has the same reference so made a combine table. The reference mapping for other two are given below.

Table 6. Data Representation for Filecoin

S.No	Attribute Name	Where to find
1	Price	NAFilecoin.csv. Use 'AVG' value.
2	Fees	**Data not found**
3	Block Time	Blocktime.csv (coin folder)
4	FCAS Score	**Data not found**
5	Transection Volume	**Data not found**
6	Exchange Volume	**Data not found**
7	Bandwidth	up.csv and down.csv (coin folder)
8	Storage	Total.csv (coin folder)

Reason for Data not Found

- 'Fee' is given in the 'Flipside' folder in the file 'final_daily.csv'. No data for 'Filecoin (FIL)' could be found in this file.
- 'FCAS' score can be found in 'flipside folder' in 'fcas.csv' file. Again, no data could be found for 'Filecoin (FIL)'.
- Transaction volume and exchange volume can be found as 'token_volume' and 'usd_volume' respectively in the 'final_daily.csv' file in the flipside folder. No data could be found for 'Filecoin (FIL)'.

Table 7. Data Representation for Sia

S.No	Attribute Name	Where to find
1	Price	NASiacoin.csv. Use 'AVG' value..
2	Fees	Flipside folder, final_daily.csv. Use 'ALL'.
3	Block Time	Miningdb.json (Siacoin folder)
4	FCAS Score	Flipside folder, fcas.csv
5	Transection Volume	Flipside folder, final_daily.csv. Use Token volume
6	Exchange Volume	Flipside folder, final_daily.csv. Use usd volume.
7	Bandwidth	Bandwidthpricesdb.json file (Siacoin folder)
8	Storage	Storage.json (Siacoin folder)

We have few doubts about the data, for example

- We have to use the 'average' value for 'price'. This value is zero for some samples?
- Can you explain some cases like, given the input values and the preference this should be the choice.

8/13/2019

Requirement Gathering V0.1

Dr Avinash kumar singh & Arvind narayanan
INTAIN PVT. LTD.

AI for Blockchain

Problem Definition

Given the input attributes (cost/budget, speed, security, health, etc.) find the optimal value of storage, computing and database.

In order to precisely define the problem, we used the below tabular representation. In total we have 12 attributes. We added two additional variables to capture more insight about the data. Unit is used to capture the measurement and domain is used to define the range of the variable. A variable can have the discrete value, or it can have the categorical value. For example, we can define the budget in $ or you can also set it between a domain [1-10] where 10 represents high importance and 1 denotes low importance. Both ways are fine for handling the data.

Table 3. Input variable details

Optimization Lever	Attribute Name	Attribute symbol	Unit	Domain
Cost/Budget	Price	a_1	$	[a-b]
	Fees	a_2	$	[a-b]
Speed	Transection/Block Time	a_3	S/M/H/D	[a-b]
Security	Security (FCAS Score))	a_4	?	
Health	Transection Volume	a_5	number	[a-b]
	Exchange Volume	a_6	number	[a-b]
Component Specific Variables	Size	a_7	?	[a-b]
	Bandwidth	a_8	MBPS/GBPS	[a-b]
	CPU	a_9	number	[a-b]
	RAM	a_{10}	GB	[a-b]
	Storage	a_{11}	GB	[a-b]
Other	Location	a_{12}	City/country	[a-b]

If we keep this representation to define our dataset, we would be needing information about

- Whether the user will provide data in discrete or domain(categorical) or in combination of both.
- We would be needing unit/domain of every attribute

The output variable is denoted in below table. Given the attribute values a choice will be made for each output variable. We assume that the output variables are independent, and their choice do not impact to each other.

Table 4. Output variable details

Storage	Computing	Database	Domain
Genaro	iExec	Bluezelle (BLZ)	d_1
Storj	Sonm (SNM)	OrbitalDB	d_2
Filecoin (FIL)	Golem (GNT)	BigCHainDB	d_3
Sia (SC)	Kingsd		d_4

If we keep the above convention, our dataset would be represented as given below.

Table 5. Dataset Representation

Sample	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	Storage	Computing	Database
s_1	-	-	-	-	-	-	-	-	-	-	-	-	$[d_1 - d_4]$	$[d_1 - d_4]$	$[d_1 - d_3]$
s_2	-	-	-	-	-	-	-	-	-	-	-	-	$[d_1 - d_4]$	$[d_1 - d_4]$	$[d_1 - d_3]$
s_3	-	-	-	-	-	-	-	-	-	-	-	-	$[d_1 - d_4]$	$[d_1 - d_4]$	$[d_1 - d_3]$
s_4	-	-	-	-	-	-	-	-	-	-	-	-	$[d_1 - d_4]$	$[d_1 - d_4]$	$[d_1 - d_3]$

Since, we would be using supervised learning to solve the problem, we would be needing the decision variable $[d_1 - d_4]$ for every sample. Every sample is represented by the collection of 12 attributes or a 12-dimensional vector ($S \in \mathbb{R}^{12}$).

Data Understanding

Let's take the storage problem first. For the storage we have four choices it can be either Genaro, Storj, Filecoin or Sia. The data for each is summarized below.

Table 6. Data Representation for Genaro

S.No	Genaro	Fields	Field Type
1	blocktime_X	Time, blocktime	integer (14), float (2)
2	contractformation_X	Time, contact information	integer (14), float (2)
3	difficulty_X	Time, difficulty	integer (14), float (4)
4	down_X	Time, down	integer (14), float (2)
5	downusd_X	Time, down usd	integer (14), float (0)
6	facebooklikes_X	Time facebook likes	integer (14), float (3)
7	free_X	Time, free	integer (14), float (2)
8	githubcontributors_X	Time, github contributors	integer (14), float (2)
9	githubstars_X	Time, github stars	integer (14), float (4)
10	githubwatchers_X	Time, github watches	integer (14), float (3)
11	hashrate_X	Time, hash rate	integer (14), float (4)
12	hosts_X	Time, hosts	integer (14), float (3)
13	hostsonline_X	Time, host online	integer (14), float (3)
14	NAGenaro-Network	Date, open, high, low, close, volume, market, cap, average	date (yyyy-mm-dd), float (0), float (0), float (0), float (0), integer (8), float (0), float (0)
15	newcontractformation_X	Time, new contact information	integer (14), float (3)
16	price_X	Time, price	integer (14), float (2)
17	redditors_X	Time, redditors	integer (14), float (5)
18	sfperfees_X	Time, sfperfees	integer (14), float (3)
19	total_X	Time, total	integer (14), float (4)
20	twitterfollowers_X	Time, twitter follower	integer (14), float (5)
21	up_X	Time, up	integer (14), float (2)
22	upusd_X	Time, upusd	integer (14), float (0)
23	usd_X	Time, usd	integer (14), float (0)
24	used_X	Time, used	integer (14), float (3)
25	used_X	Time, used	integer (14), float (1)

There are 25 xl files inside the Storage->Genaro, in most of the files except the “NAGenaro-Network”, it has two columns. The first column of all these files are fixed “time” and the second column is different as shown in the above table. Let’s look at another storage “Storj”

Table 7. Data Representation for Storj

S.No	Storj	Fields	Field Type
1	blocktime_X	Time, blocktime	integer (14), float (2)
2	contractformation_X	Time, contact information	integer (14), float (2)
3	difficulty_X	Time, difficulty	integer (14), float (4)
4	down_X	Time, down	integer (14), float (2)
5	downusd_X	Time, down usd	integer (14), float (0)
6	facebooklikes_X	Time facebook likes	integer (14), float (3)
7	free_X	Time, free	integer (14), float (2)
8	githubcontributors_X	Time, github contributors	integer (14), float (2)
9	githubstars_X	Time, github stars	integer (14), float (4)
10	githubwatchers_X	Time, github watches	integer (14), float (3)
11	hashrate_X	Time, hash rate	integer (14), float (4)
12	hosts_X	Time, hosts	integer (14), float (3)
13	hostsonline_X	Time, host online	integer (14), float (3)
14	NAStorj	Date, open, high, low, close, volume, market, cap, average	date (yyyy-mm-dd), float (0), float (0), float (0), float (0), integer (8), float (0), float (0)
15	newcontractformation_X	Time, new contact information	integer (14), float (3)
16	price_X	Time, price	integer (14), float (2)
17	redditors_X	Time, redditors	integer (14), float (5)
18	sfperfees_X	Time, sfperfees	integer (14), float (3)
19	total_X	Time, total	integer (14), float (4)
20	twitterfollowers_X	Time, twitter follower	integer (14), float (5)
21	up_X	Time, up	integer (14), float (2)
22	upusd_X	Time, upusd	integer (14), float (0)
23	usd_X	Time, usd	integer (14), float (0)
24	used_X	Time, used	integer (14), float (3)
25	used_X	Time, used	integer (14), float (1)

The attribute behaviors are similar as of Genaro. The Filecoin is identical so, we didn’t describe that. The Last one is Sia, represented below.

Table 8. Data Representation for Sia

S.No	Sia	Fields	Field Type
1	activehosts	Date, host, host online	integer (14), integer (3), integer (3),
2	bandwidthpricesdb	Date, up, down, upusd, downusd	integer (14), float (2), float (2), float (0), float (0)
3	miningdb	Time, hashrate, difficulty, blocktime	integer (14), float (0), float (0), float (1)
4	socialimpact	Time, githubwatchers, githubstars, githubcontributors, redditors, twitterfollowers, facebooklikes,	integer (14), integer (3), integer (4), integer (2), integer (5), integer (5), integer (5)
5	storage	Date, total, used	integer (14), float (4), float (2)
6	storagepricesdb	Date, price, sfperfees	integer (14), float (3), float (2)
7	ussage	Date, used, free	integer (14), float (1), float (2)
8	NASiacoin	Date, open, high, low, close, volume, market, cap, average	date (yyyy-mm-dd), float (0), float (0), float (0), float (0), integer (8), float (0), float (0)

We have few doubts about the data, for example

- If the input variables are price, fee, etc. (please refer to table-1 column1 and 2”), what are these variables/files?
- How they are relevant or related to the input variables?
- How they are helpful in solving the problem.
- The data of Filecoin and the Genaro folder and looking same (are the copied mistakenly?).
- The data is collected with respect to either time or date, in other words the data is time dependent or in other words it is dynamic (changing with time) ?, do we expect more data in the same format
-