Caitlin Post

ENGL 182

Topic: Big Data – A Public Asset

Abstract: The paper starts with describing how data is a valuable asset for the society through some examples from the recent past. The data asset is classified into 3 categories: raw data, insights, and transactions. Insights and transactions are resourceful data which carry a lot of value in various applications which makes it expensive; whereas, raw data is just a bunch of characters with no value to it. This is followed by examples of how big data could improve quality of life in future, specifically in health and education. Further, the paper stresses on the issues with the current big data ecosystem and how they slowdown innovation. Ownership is the major issue as majority of the data assets are owned by only a couple big tech companies. Other issues are constrained access, disjoint data, and privacy. The paper stresses on the need for a new big data ecosystem which solves the mentioned issues and boosts innovation in the society. Further into the paper, a decentralized data ecosystem model is introduced with very brief technical details. The paper intends to describe the various parties involved in the system and how they interact to create a fair space for everyone. Apps and services collecting user data and users are responsible for the creation of raw data in the system. The ecosystem allows data scientists from all over the world to access raw data and convert them into transaction and insight assets. Companies with B2B or B2C services can use those transaction and insight assets to provide quality products. In return, these companies pay money for those assets which rewards data scientists, users, and data collecting apps appropriately.

Key Words: Big Data, raw data, insights, transactions, privacy, ownership, decentralization, digital assets

The value of Big Data

"The amount of data stored by businesses nearly doubles every 12 to 18 months" (Babcock). Data has emerged as an asset in the last few years. 7 of the top 10 companies in the world are tech giants which rely on data for important insights and applications. It is an asset which has already generated huge wealth. As people use more web-connected services, more and more information is collected through activity on the services. As data-driven insights get more accurate, companies will use them to optimise their business strategies to maximise profits. Let's take the example of the auto industry: vehicles now feature GPS and telematic systems in their cars, using which car manufacturers have been able to collect and monetize from data on customer driving habits. Particularly, General Motor's. telematics system collects telemetry data, which they have used to offer lower insurance premiums to customers who drive fewer miles. This boosted GM's customer satisfaction. In another example, "American Express Co. launched an analytics and consulting business that draws on the purchasing behaviour of its 90 million credit card holders across 127 countries. This organization, American Express Business Insights, hopes to attract direct marketers by using proprietary data to enhance customer acquisition and retention programs" (Banerjee et al., 2).

Raw data itself is just a sequence of characters, which unless seen in the context of usage, has no value. Data is used to derive useful information which can be monetised. Data itself doesn't have any value. Look at the Information value pyramid developed by Accenture in Figure 1. The pyramid has 3 levels: raw data, insights, and transactions. **Raw data** is on the bottom and carries very less value itself. Moving up the pyramid creates larger value opportunities, but they are difficult to execute. **Insights** are on the second level of the pyramid, and find various applications in all sorts of things. Some examples of insights were given in the previous paragraph. Another example is Google keeping a track of searches related to "flu" and predicting outbreaks. Google uses search patterns related to different

symptoms and diseases, and informs local health authorities about risks of any outbreaks. Also, Amazon detects products with high demand and releases them in its own brand called "Amazon Basics". **Transactions** lie on the top of the pyramid and carry the most value. Data at the "transactions" level enables companies to execute their end-to-end processes better, and could help improve point-of-sale retail transactions, marketing campaign rollouts or fraud detection (Banerjee et al., 4). Data at Transaction level is more of personal level data. Example would be Google and Amazon using individual user's records to recommend new products or to provide relevant search results. This kind of data carries the most value and has proved to be most profitable. Remember that insights and transaction level data is derived from raw data only. They are just stages of making the data usable. Raw data itself does not have any usable or productive value and only poses risks to user privacy.

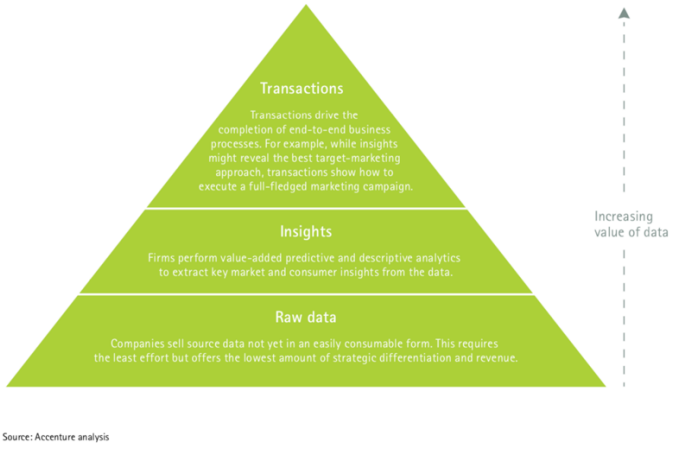

Figure 1 (Source: Accenture)

How will big data benefit the mankind?

The examples of big data benefitting the mankind are endless. I am going to lay down few of them:

1. **Health:** In order to tackle big data challenges and perform smoother analytics, various companies have implemented AI to analyse published results, textual data,

and image data to obtain meaningful outcomes. IBM Corporation is one of the biggest and experienced players in this sector to provide healthcare analytics services commercially. IBM's Watson Health is an AI platform to share and analyse health data among hospitals, providers and researchers (Dash, S). As the society innovate, healthcare will become much cheaper and more technology oriented. AI systems could be trained to ask users for symptoms and prescribe drugs on its own. As image processing becomes more advanced, AI could detect skin diseases and warn us of any health risks. With the use of wearable technology and IOT all around us, AI could warn us of unhealthy patterns in our lifestyles and could also warn of us potential health risks such as cancer.

2. **Education:** Online learning is the future of education. Companies have been building AI's to educate students right from the primary level. As these AI's get more powerful, a sustainable education model would be developed which could be accessed by anyone. It would make schools irrelevant and turn universities into research hubs. It would be able to teach each student at its own pace, keeping track of their progress. And it would also allow students to go in depth into topics of their own choice. It would also be able to assess strengths and weaknesses of individual students and hence provide a customized curriculum to students.

Issues with the ecosystem

1. **Ownership:** The major thing I see wrong with the current data model is its ownership structure. Right now, the data is considered proprietary asset of the company that generates it. But users play a major role in the generation of data, as their inputs on the numerous apps and web services is the source of data. Hence, users deserve some ownership of the data and the rights over it. Meanwhile, companies which collect

data through apps and services, should also be incentivized for the collection. Hence, data should be partly owned by both the parties.

Big data and your privacy (by John Dillard), talks about big data in agriculture. Farmers have been using services which collect various data on their farmland about weather, irrigation, sunlight, yield, etc. and gives them insights. These systems could also recommend them optimal amounts of irrigation, fertilizers, or pesticides based on the collected information. Although these services are made and marketed to help farmers, this data flows to large databases storing terabytes of data about farms around the world. And the collected data is sold to generate more profit. Recently, farmers have been concerned about the selling of their data and wonder how their data is used. Moreover, they have voiced concerns about the ownership of the data and demanded profits from the selling of data. Such examples can be seen in some other industries equipped with big data infrastructure. Hence, it is important for users to have ownership in their data for the data ecosystem to sustain.

European Union identified this issue some time ago and passed a law "article 20" to protect users' rights over their data. Europeans now have the right to "receive the personal data concerning him or her… in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance…". Also, users can now ask companies to delete their personal data. This was a revolutionary step by European Union.

2. **Constrained Access:** Data can only be accessed by very large companies. Either the ones that collect the data, or the ones with very high capital which buy rights to data from data brokers. Smaller companies and young data scientists have no way to get access to the large amounts of the proprietary data which limits the innovation possible. Moreover, smaller companies have major disadvantages due to inferior

insights on their users. This gives larger companies an edge over their user-base, demotivating young entrepreneurs. Also, data scientists and young students have potential to generate insights from raw data, which bigger companies have not been able to. If data is openly accessible to generate insights, innovation would be boosted, and mankind would benefit more.

3. **Disjoint Data:** Most of the user data is stored in several centralized servers in broken parts. If those data are combined, the value of that resulting data would increase exponentially. For example: Amazon collects data about users' shopping preferences. Google collects data about your search history and web browsing. And, Facebook collects data about your social interests and connections. If the data from these three companies were combined, the resulting data would be way more powerful than those independent sets. It could provide way more insights on users.
4. **Data Security:** Although data breaches are rare, but whenever they take place, massive risks originate since sensitive data such as passwords and payment details are revealed. Centralized servers are very vulnerable to such attacks because all the data is stored in closely connected systems.

Therefore, data is an asset. Bing Song cites "Data has become and will continue to be a foundational basis for our society. It is much like the air we breathe, water we drink and electricity we depend on. Moreover, each of us is a source of data, and we all constantly contribute to data flows". As the amount of big data is increasing by the day, the future possibilities of innovation keep increasing, and so does the data's value. But the ecosystem which has developed around big data is not optimal. It should be an asset owned and controlled by the society and not corporates. The system lacks transparency and hence there exists no confidence between users and the big companies. A lot of users are concerned about

privacy of their data. Moreover, they are curious what is happening with their data. This has made people concerned.

I believe that there is a need for an alternate ecosystem for data which solves these issues and accelerates innovation in the society. A system that is fair to everyone. Further in this paper, I present a proposal for a decentralised big data ecosystem which could solve this purpose.

Model: Decentralized Big Data Ecosystem

With the release of bitcoin, the leading cryptocurrency, the tech world was introduced to the technology of blockchain which stores ledgers on a peer to peer network of nodes. With further innovation into blockchain, Ethereum introduced smart contracts which are self-executing contracts between 2 or more parties mentioning the terms of agreement between them. These contracts permit trusted agreements and transactions to be processed without any external mechanisms. We build our model using the two technologies.

The decentralized model is built over a large number of peer-to-peer network of nodes (hosts) which can store data, execute smart contracts, and perform all kinds of computing operations which current cloud computing systems offer. It is setup to be trustless which means that the users don't have to trust a central authority for the integrity of the system. The network as a whole is owned by everyone who is a member of the network. The network considers data as an immutable asset whose access is only given to entities who have been granted the access according to the smart contracts. It can store 3 types of data assets which are raw data, insights, and transactions. These assets specify ownership and access rules. There are 5 kinds of parties in this network:

1) **Party 1:** Apps or Services which collects user data
2) **Party 2:** Users or individual members on the network

3) **Party 3:** Individuals or companies which use data to create new AI algorithms or develop some kind of insights from the data. They don't create applications and hence don't have any profits
4) **Party 4:** Apps or companies which either use algorithms from Party 3 or use data directly to provide services to customers or other businesses
5) **Party 5:** Individuals who invest into the nodes used in the network and are paid for the amount of work done by the node

   Note: A single entity might be participating as more than one of these parties in the network.

Input of Data into the Network – Creation of raw data asset

Apps which collect user data (Party 1), and the users (Party 2) using them are responsible for the input of the data into the network. These apps would be programmed in a very similar way to how they are programmed right now. The only difference would be that their background logic and database would be hosted on the decentralized network. They would be able to save any acquired data on the network and would be able to access that data at any time in their application code to perform operations (Only at application level – explained later). The network would register the data as a **raw data** asset which is owned by the Party 1 and Party 2 (whose data it is) at agreeable terms. For this paper let's assume that Party 1 owns 30% of the asset, whereas Party 2 owns 70% of it.

Let me explain this through an example. Suppose Google app collects data about a user's basic personal details and stores that on the database. This input of information will create a raw data asset in the system. The asset contract will specify a unique ID for the asset along with brief detail about the data. It will also mention the owners along with their respective

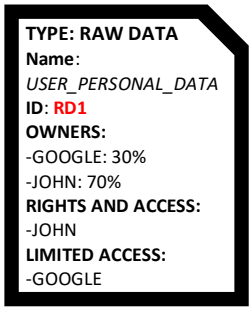

Figure 2 a)

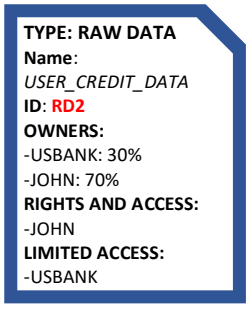

Figure 2 b)

ownership ratios. Moreover, it will clearly specify that what party in the contract gets what rights or access over the asset. Figure 2 represents an asset contract for such information stored by google for a user named John.

Pool of Data

As multiple apps would generate user data over time, a large pool of data would be created. This pool of data would keep getting larger with time. The network would be able to combine data of a user from multiple apps which would be way more valuable than those disjoint datasets. Figure 2 b) shows another data asset provided by US Bank about John's credit info. Using both Google and US Bank data on John could develop very strong insights. This solves the issue of **Disjoint Data** mentioned earlier.

Role of Party 3 – Turning raw data assets into insights and transactions

These would be companies or programmers, particularly data scientists whose goal would be to make use of data in a way which hasn't been done before. They would be training new AI algorithms, creating useful derived data from the already existing datasets or generating resourceful insights. They will essentially be converting raw data in the system into **insights** and **transactions**. They won't be providing any services using these innovations and hence, would be able to access data at application level free of cost on the network. They would indeed be creating another asset on the network, of type insights or transactions, which will be partly owned by them and the owners of the data assets used. For this paper let us assume that Party 3 would share the ownership equally at 50% with the owners of the data asset. Furthermore, each data asset would further be given a relevance score based on how important the data was to the resultant asset created. The owners of each data asset would be assigned ownership of the newly created assets in the proportion of the relevance of the data asset. For example, a data scientist in Seattle has developed insights relating users' ability to pay off debt. It uses data from Google and US Bank from the example above. It uses data of

hundreds of thousands of users available from both the companies. For simpler understanding let us assume that data of only four users is used. Each kind of data asset is assigned a relevance value. These relevance scores are such that they sum up to 1, for ease of calculation. In this case, credit history is given a relevance score of 0.6 as it is very critical to the process. Consequently, personal information is rated at 0.4. Let us refer to these 4 users as User1, User2, User3, and User 4 and to the developer as DEV. Figure 3 a) shows the contracts of the data assets used. And Figure 3 b) shows how ownership ratios for the newly created asset is evaluated.

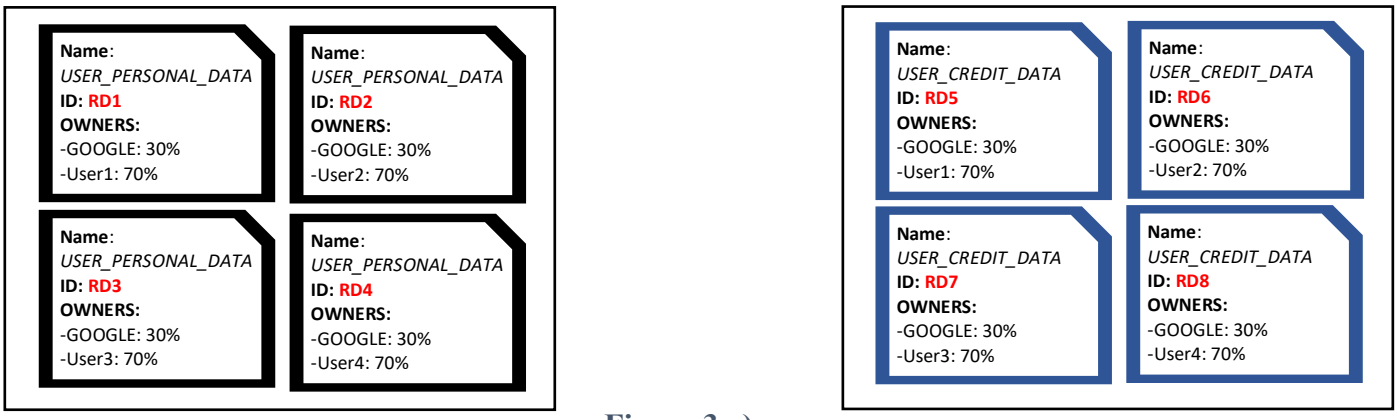

Figure 3 a)

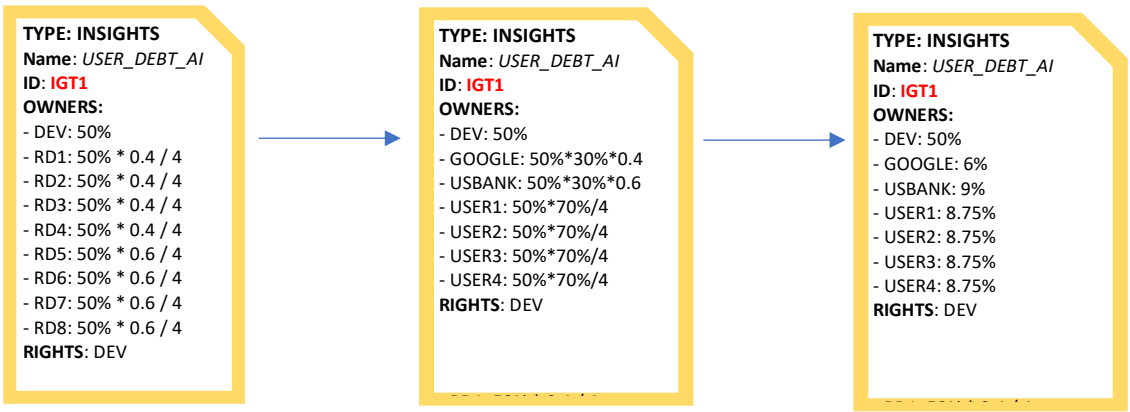

Figure 3 b)

Hence, assets created by party 3 are under shared ownership of the data collectors (party 1), users (party 2), and innovators (party 3). Rights for this asset are given to party 3 as they are the creators of the asset.

Role of Party 4 - Profit Generators using insight and transaction assets

These are the companies which provide B2B or B2C services. They earn profits from their

services. They use data assets of the type **insights** and **transactions** to provide useful services and user experiences to their customers. As they use these assets, they make payments to the owners of the assets they used. They generate the revenues which are used to pay all the other parties on the network. The price of the assets they used would be decided based on various market factors. In some cases, companies might be negotiating with Party 3 for access to assets created by them.

Access to Data

Earlier in the paper it was established that raw data is itself of no use and only poses threats to user's identity. Hence, raw data relating to a user can only be accessed by the user itself. If any other party has access to an asset, they will be able to use the asset through data kits. Data kits provide a layer of abstraction between data and programmer which lets the programmer use data in a meaningful way in its code without making the data vulnerable.

Conclusion

The proposed model solves the issue of ownership of data by maintaining contracts for all the assets. Further, it also fixes the problem of constrained access by letting anyone around the world create insights and transactions out of raw data. It also addresses privacy threats by never displaying actual raw data. Moreover, a decentralized system is very secure in itself. Additionally, it also combines disjoint data sets obtained from different services.

Such a big change in the ecosystem can only be made when there is widespread awareness and enough people are voicing their concerns. Governments have consistently been behind technology due to rapid development in the field in the recent past. Even most of the internet laws were passed many years after the issues were first realized. As this issue becomes more heated, more people would come up with new solutions tackling the issues. But implementing any such framework is going to come with a lot of challenges, particularly in its approval and adaption.

BIBLIOGRAPHY

Banerjee, S., Bolze, J.D., McNamara, J.M. and O'Reilly, K.T. (2011)

How Big Data Can Fuel Bigger Growth. Accenture.

https://na.eventscloud.com/file_uploads/006832b470b7dbe3ff0b3fc885fb4b7b_bigdata.pdf

Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. J Big Data 6, 54 (2019). https://doi.org/10.1186/s40537-019-0217-0

John Dillard. Big data and your privacy.

Farm Journal Media (Vol. 138, Issue 7)

https://go-gale-com.offcampus.lib.washington.edu/ps/retrieve.do?tabID=T003&resultListType=RESULT_LIST&searchResultsType=SingleTab&searchType=BasicSearchForm&currentPosition=17&docId=GALE%7CA379197414&docType=Article&sort=Relevance&contentSegment=ZCUM&prodId=PPES&contentSet=GALE%7CA379197414&searchId=R2&userGroupName=wash_main&inPS=true

Charles Babcock

Data, Data, Everywhere

https://www.informationweek.com/data-data-everywhere/d/d-id/1039328?

Bing Song

Big data as the next public good – Washington Post

https://www.washingtonpost.com/news/theworldpost/wp/2018/05/02/big-data/