5-2022

# How Blockchain Solutions Enable Better Decision Making Through Blockchain Analytics

Sammy Ter Haar

*University of Arkansas, Fayetteville*

# How Blockchain Solutions Enable Better Decision Making Through Blockchain Analytics

**by**

**Sammy L. ter Haar**

**Advisor: Dr. Zachary Steelman**

**An Honors Thesis in partial fulfillment of the requirements for the degree Bachelor of Science in Business Administration in Information Systems – Business Analytics.**

**Sam M. Walton College of Business**
**University of Arkansas**
**Fayetteville, Arkansas**

**May 14, 2022**

## -Table of Contents-

**-Abstract-**

Since the founding of computers, data scientists have been able to engineer devices that increase individuals' opportunities to communicate with each other. In the 1990s, the internet took over with many people not understanding its utility. Flash forward 30 years, and we cannot live without our connection to the internet. The internet of information is what we called early adopters with individuals posting blogs for others to read, this was known as Web 1.0. As we progress, platforms became social allowing individuals in different areas to communicate and engage with each other, this was known as Web 2.0. As Dr. Mary Lacity put in her Blockchain for Business textbook, we are shifting towards the '*Internet of Value*' with blockchain's emergence (Lacity, 2020). Blockchain solution enabling true ownership of data is a major force in the push for Web 3.0. As individuals take back more power, it is important to ensure proper analysis of current and future situations. This is conducted through data analysis, but we now have an emerging technology that can record large amounts of information and provide quick insight. We can also ensure the data is clean through its native open-source architecture that prevents duplication of data points. The most complete and exact results can be found in a blockchain network where inputted data is automatically organized and scrubbed. Data grows exponentially, it is important to find ways to store this data while also ensuring it is complete.

**-Data-**

Understanding the unknown is done by consistently questioning and interpreting what is seen. All disciplines that people are taught are the result of discovering relationships as to why things happen. When an outcome is seen, the variables of cause and effect are key pieces that help in understanding. For example, when Sir Isaac Newton discovered the force of gravity when an apple hit him on the head. His discovery was not immediate when the apple hit his head, but after observing trials of different objects being dropped. The key to discovering what the observations mean is in the variables or *data.* (See Appendix for a glossary of key terms).

Data is information that can be processed and analyzed to reveal trends and causations for certain experiments. Data by itself means nothing, it's just recorded raw information that can come in different forms. *Qualitative data* is a type of data where no numerical values are included, examples of this type of data would be a list of all the countries in the world. *Quantitative data* is the opposite being data that are only numerical values, an example of this could be GDP per capita. Oftentimes, we see statisticians associate qualitative variables with numerical variables to find trends, for example, seeing the GDP per capita for each country. Behind each measure, understanding of the metrics can be scoped down further by recording data about the population of a country of interest. As you can tell by now, there is simply no limit to what can be decided as data. Companies, usually keep large databases filled with what is called *Big Data* to be used when needed. Employees are hired to navigate the database and find key data that might be of interest, this is known as data mining.

With computers, data is now being recorded at such a quick rate that people do not know what to consider important. As mentioned previously, knowing which data to use all depends on the overarching question at hand. A full understanding of the question will lead to finding the most useful data. Data can also be incomplete, meaning it must be 'cleaned' in a way where duplicates are removed, and missing pieces are filled into null spaces to obtain the most accurate results. In all markets, companies must be aware of their competitors and thus must have the correct data to compete. Real-time data is important to maintain relevance, it will not do a company any good if they are running analysis on outdated data, this combines the importance of recording data often as well as understanding it as quickly as possible. In the cyberspace, people want access to insights and could act nefariously in obtaining records. Data security is crucial to keep an advantage over competitors and prevent hackers from stealing information. According to Megan Leonhardt of CNBC, "Yahoo told the public in September 2016 it had experienced a breach in 2014 that affected at least 500 million accounts" (Leonhardt, 2019). Sensitive information of their users can be lost or compromised as a result of these hacks which can lead to poor public opinion for said company in question. Data variables are being created and recorded at exponential speeds leaving many analysts being flooded with millions of lines of data. How do you figure out what is important and how do you leverage your knowledge into a story? With the use of analytics.

## -Analytics-

Raw data has information that is uncovered via analysis. *Analytics* is the process by which raw data can be understood whether by metrics or through visual dashboards. To create an efficient analysis, the data must be cleaned and organized in a way that makes sense. The purpose of analytics is to reveal trends and information that was unknown previously. One technique that is widely used among businesses is the cross-industry process for data mining or "*CRISP*."

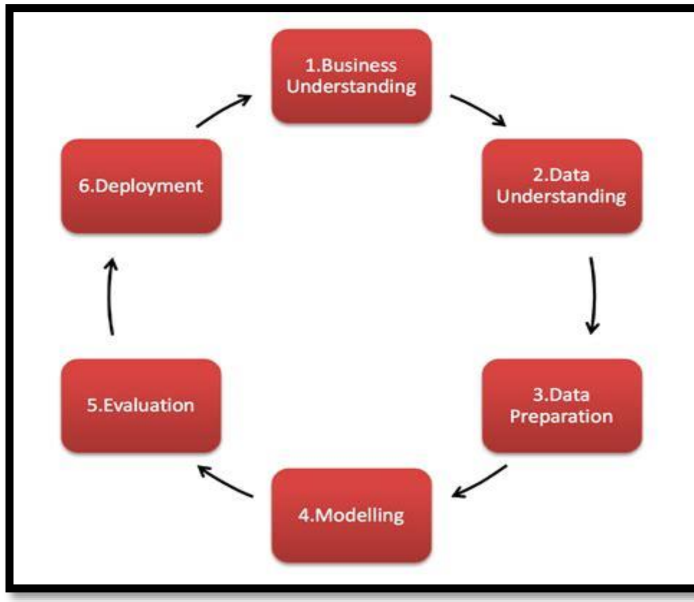


**FIGURE 2.**

**CRISP METHODOLOGY**

This method is a structured approach to understanding a dataset and running the correct models to yield the most useful results (Vorhies, 2016). As shown in *Figure 1*, an analyst would start with business understanding, then dive into the raw data for data understanding/preparation, in this stage the data is cleaned so that there are no missing values or duplicates that could alter true results. From there, a modeling approach is chosen. There are two types of approaches when it comes to modeling, supervised and unsupervised. *Supervised* approaches consist of regression analysis, decision trees, and neural networks. In this type of analysis, there is a known variable the analyst wants to know more about. An *unsupervised* approach is clustering, where raw information about the data is produced to highlight the prevalence of what is being seen in the dataset. After these models are run, the analyst can evaluate their outputs and produce a deliverable of their findings to be shared with higher-ups. It is important to note that multiple models can be used to uncover information for a dataset, in fact, it is encouraged. In addition to uncovering information and finding trends, the analyst must be able to communicate their findings in a comprehensible way. Brent Dykes, an expert in data storytelling uses Lego bricks, as seen in Figure 2, as a comparison for the importance of communicating what the data reveals (Dykes, 2021). He exhibits this process perfectly by using the individual Lego bricks as an example of

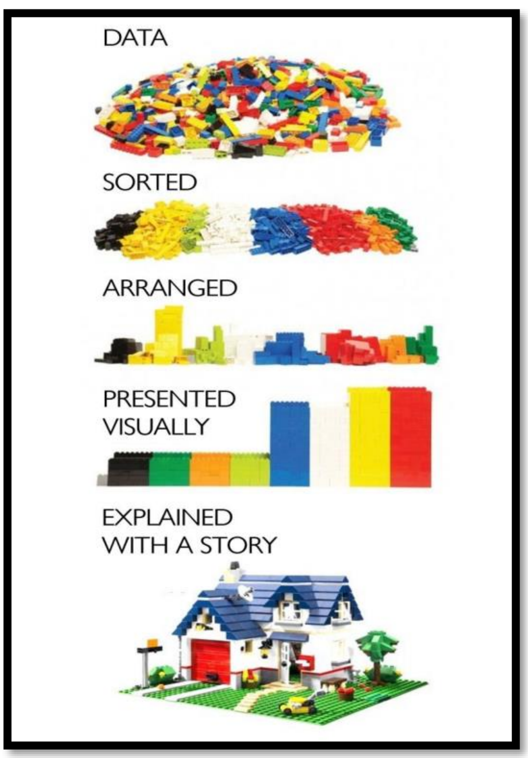


**FIGURE 1.**

**LEGO VISUAL EXPLAINING ANALYTICS**

raw data. Each brick is then sorted by color and arranged in a pleasing manner that can be presented visually. However, the key to analysis is understanding which is why we can see a beautiful Lego house built out of the pieces of raw data. Understanding is where the value is generated. The analysis is done on several platforms; visualizations can be created directly with raw data on software such as Alteryx or Tableau. SAS Viya is another software that can allow multiple forms of analytics that can be combined to create a dashboard with all the findings. The goal of completing this CRISP method is to engage in *predictive analytics*, where companies are trying to predict the future given their findings. This reiterates the importance of having real-time data to make accurate decisions that are relevant at said time. Analytics can directly reveal an opportunity where it is not seen. This means that data inputs are being done not only by the user manually but through the computers adding new types of data to their system. Devices can communicate with each other now and share data regarding their owner. New data variables are constantly being created at speeds that are impossible to keep up with for analysts.

**-Internet of Things-**

Machine learning allows a system to produce outputs that match a user's interest, this is done through system-to-system data sharing. The machine is conducting all the analytics in real-time and all in the backend that yields information that aligns with the user or operators' preference. This is done through the *internet of things* or *IoT*. This concept is the idea that any device with an on or off switch can communicate with other devices by connecting to the internet (Clark, 2016). IoT enables data sharing and the ability for these machines to produce outputs suited to a user's preferences. An example of objects communicating data to each other is an iPhone and iWatch. These two devices are connected via a user's Apple account that tracks health and fitness data about their user. The watch is also able to alert the user when they receive text messages and phone calls. The Amazon Alexa and Google home are two devices that are

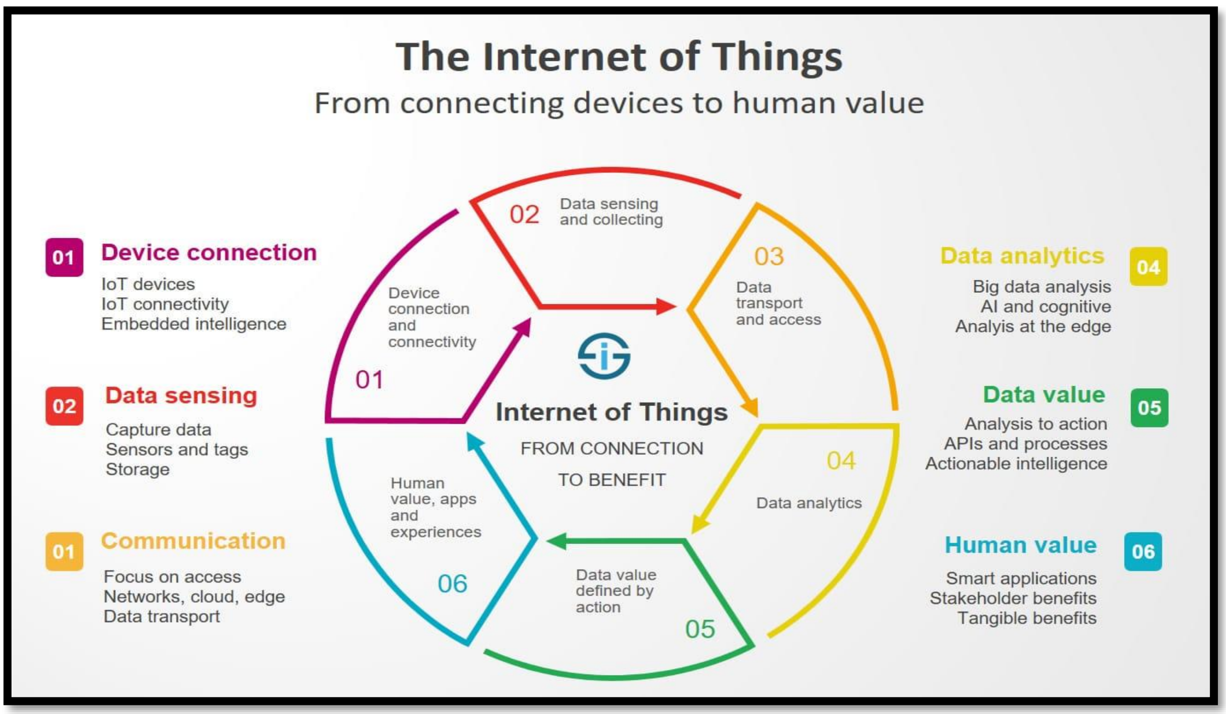


**FIGURE 3. INTERNET OF THINGS GRAPHIC**

voice-operated that can communicate to different devices to accomplish an end-user goal. Tasks such as “turn on the lights” are accomplished by the device recognizing the command and communicating to a smart light bulb to turn the light on (Rovai, 2021).

These are just two examples, there are multiple household essentials such as smart refrigerators that can communicate data to supported devices that are linked. These smart fridges use the power of artificial intelligence to constantly learn key aspects of their user (Rabbitte, 2020). This capability leverages the data already being stored per device and creates a medium for these devices to communicate and accomplish simple tasks. The smart device in this instance the fridge would use *machine learning* to recognize past habits so that it can provide more value for its user. For example, a smart fridge can connect to a shopping app like Amazon, and automatically add items to a user’s cart based on anticipated events that are consistent with past behavior. The automated machine learning algorithm will naturally mold to a user’s wants by recognizing trends in the data that is inputted. This is a revolutionary form of technology that is growing immensely. However, with each device possessing so much data, data scientists struggle to pick out which pieces should be analyzed further. Because we are flooded with data from multiple sources, having to aggregate the data across various locations, we end up with lots of data and specifically replicated data which may vary by location/device. Therefore, we struggle with fully trusting the data due to the multiple layers of aggregation and trust needed to gather from multiple databases. However, as we begin to move into this next “wave” of big data due to IoT devices constantly communicating on the internet, we need a new solution to not only aggregate the data but to ensure that the data is correct and valid, enter blockchain.

## -Blockchain-

*Blockchain* is a new technical solution that is a shared, immutable ledger that can record transactions, track assets, and allow trust in a world of distrust. Information on the blockchain

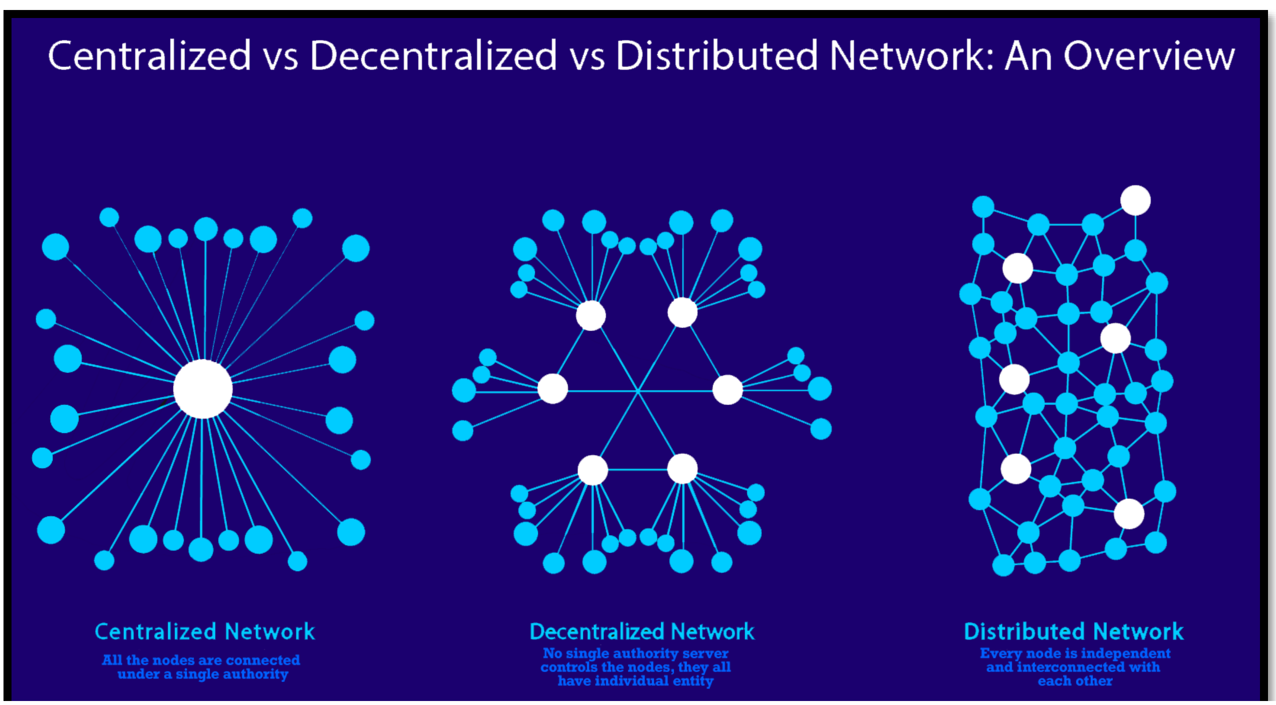


**FIGURE 4. CENTRALIZED NETWORK VS DECENTRALIZED NETWORK**

cannot be altered as it is cryptographically hashed by the previous block which is all recorded on its *distributed ledger*. Information on the blockchain can be seen due to the open-source nature of its code. The most widely recognized blockchain has to be Bitcoin. Bitcoins white paper was published by the anonymous Satoshi Nakamoto on October 31, 2008 (Bitcoin Whitepaper Explained, 2020). The white paper was formed by multiple developers connected through an email chain started by Nakamoto. It is worth mentioning that the other developers had no clue who Nakamoto was even after asking for his identity they would see their questions ignored. The white paper states that bitcoin would be a fully peer-to-peer electronic online cash payment system that enables funds to be transferred without the need for a third-party entity or financial institution.

Participants on the blockchain are known as *nodes*, which is just a device that can communicate directly to the blockchain. Users can download bitcoin clients that allow their device to operate as a node. Clients can be full, all data on-chain is rendered in your node or light, and only partial data on-chain is on your node. What makes a node different than owning bitcoin on a decentralized exchange is being able to fully possess your private key. Transactions on a blockchain are recorded via their decentralized distributed ledger. A node will send information to a public key address, the information cannot be accessed without the private key of the recipient. These transactions are validated via the consensus mechanism known as *proof-of-work*. Proof-of-work is a method of different nodes competing with each other to solve an equation that validates a transaction, the first node to complete the equation is rewarded in bitcoin, meaning there is an incentive to participate as a full client. There are other forms of consensus mechanisms as well, one popular version is known as Proof-of-stake where a node that possesses a high amount of a specific cryptocurrency, let's say Ethereum, is rewarded with the ability to validate transactions. *Proof-of-stake* reduces energy consumption and focuses on those who have high amounts of coins.

Blockchain has sparked the emergence of new markets and ideas in multiple different industries by promoting artwork through non-fungible tokens (NFT) and solving financial issues through *decentralized finance* or *DeFi*. Multiple altcoins, a coin that is an alternative to bitcoin, networks are being built to support a space where participants truly own their data and can select what to share and what to keep anonymous. During current events, developers interested in the blockchain space are trying to push for Web 3.0 solutions. According to Charles Silver of Forbes, "…Web 3.0 will bring us a fairer internet by enabling the individual to be a sovereign. True sovereignty implies owning and being able to control who profits from one's time and information" (Silver, 2020). With a technical innovation as massive as the blockchain, user data inputted, and trends will be of concern to any governing nation that supports this innovation. Thus, analytics is essential to maintaining not only security but display information that would influence a participant's decision-making.

**-Blockchain Analytics-**

The blockchain in itself creates multiple data points that are native to its technology, the process by which funds and assets are transferred is unique which means multiple agreements/transactions can occur simultaneously every day. With the whole Web 3.0 revolution taking place currently, various industries are looking at blockchain-based solutions to gain a competitive advantage. For example, in financial markets, key data points that are analyzed on the blockchain include transactions of cryptocurrency. Similar to trading stocks, there are multiple metrics that an investor keeps their eye on. In crypto it is the same way; investors pay close attention to the trading volume of any given crypto along with its current price action and compare that to previous trends they have seen so that an informed decision can be made.

With crypto and blockchain adoption gaining more traction there will be a huge demand for blockchain analytics which Bitquery defines as "Blockchain analytics or On-chain analytics is analyzing the blockchain data to understand various aspects of crypto networks. With *blockchain analytics*, you can monitor multiple economic indicators for cryptocurrencies and blockchain native protocols" (Bitquery, 2020). Information stored on-chain is not centrally located but recorded within a distributed decentralized ledger system that alleviates the need for a third-party institution. Therefore, the need for blockchain analytics is important to ensure its use is ethical.

Specific on-chain information that can be tracked includes specific smart contracts. A *smart contract* is a complete computer code-generated agreement between two parties that executes when stipulations of both sides are met (Herpy, 2022). Assuming these smart contracts become more common in Web 3.0, there will be huge deals with substantial amounts of money that are transacted within these smart contracts with the transfer of money occurring instantaneously. As these become more prevalent, on-chain analytics will be able to provide transparency to the transactions. In addition, blockchain analytics can analyze interactions and social connections among participants. Each participant's identity on-chain is protected as it is just a string of numbers and letters that make up an individual's public key. However, the blockchain can recognize these keys and see who and when a certain public key has been interacting with another.

Another use for blockchain analytics includes tracking off-chain data through IP or internet protocol, to catch bad actors. Inherently, the blockchain is meant to be a place to create trust in a world of distrust and this is a foundational belief for blockchain analytics company Chainalysis. Chainalysis is providing on-chain visibility through analytics ranging from "data, software, services, and research to government agencies, exchanges, financial institutions, and insurance and cybersecurity companies in over seventy countries. Their data platform powers investigation, compliance, and risk management tools that have been used to solve some of the world's most high-profile cyber-criminal cases and grow consumer access to cryptocurrency safely" (What We Do, n.d). Mt. Gox, a bitcoin exchange from Shibuya, Japan, was started in 2010 and was handling over 70% of all bitcoin transactions in 2014. Unfortunately, the exchange was hacked seeing the largest bitcoin heist at the time with losses of $1.7 Billion. Chainalysis utilized its platform to dive into the analytics of bitcoin and act as the official investigator for the Mt. Gox bankruptcy case (Redman, 2017). The location of the missing funds was able to be tracked, however, Chainalysis co-founder mentioned in a hearing that it does not mean the missing coins can be accessed. This example proves that blockchain analytics surveillance firms can be beneficial for tracking large asset movements and providing a location of where they are. Blockchain data cannot be duplicated which ensures that the data is clean for analysis, current forms of analysis include lots of wasted time examining raw data and scrubbing it to produce the most accurate results. Blockchain analytics only analyzes on-chain activity, and since we know the inputted data is an agreed-upon version from participants in the network, we can trust that the results will be accurate and actionable.

## -Future-

Data is vast and with the emergence of blockchain technology, it is only going to get larger. As adoption becomes more prevalent, the demand for the value being added by these projects will increase, and thus the importance of analytics to find that value. The blockchain industry is extremely healthy at this period in time. Many industries are researching the best way to implement the technology towards their needs. Blockchain technology is shaping the way we see the world by not only adding business benefits via Defi and asset tracking but also creating leisure time activities via gaming and artistic NFTs. One problem that arises within blockchain analytics is the ability to query multiple blockchains. Multiple blockchains being able to communicate means the data must be consistent so it does not break while executing. Analytics will play a significant role in that it's important to have analysis on all datasets rather than just one to enable better cross-functionality. There are blockchain projects such as Cosmos ($ATOM) and Chainlink ($LINK) that are working on a solution to link multiple blockchains together. These projects are working to break the barriers of blockchains by enabling them to transact with one another. On Chainlink, users have the capability to participate in any blockchain network (Reiff, 2022). This ability goes back to the issue of querying multiple datasets and the importance of having analysis on all datasets across all blockchains to ensure interoperability. Analytics is a fundamental of conducting business and maintaining customers. In addition to providing insight on economic indicators for cryptocurrencies and blockchain protocols, analyzing data on chain can also help uncover the locations of nefarious transactions to get rid of bad actors. The answer to all problems is within the raw big data that is not only inputted by users but now being created by the computers themselves. Discovery of these blockchains true capability will be found in the data.

**-References-**


*Bitcoin White Paper Explained | Paxful University*. (2020, October 20). Paxful Blog | Crypto Guides & Product Updates. https://paxful.com/university/bitcoin-whitepaper-explained/

*Blockchain Analytics: Everything you need to know*. (2020, August 12). Bitquery. https://bitquery.io/blog/blockchain-analytics-everything-you-need-to-know

*Building and Applying Predictive Models in IBM SPSS Modeler training webinar*. (2017). Smart Vision - Europe. https://www.sv-europe.com/crisp-dm-methodology/

Clark, J. (2016, November 17). *What is the Internet of Things, and how does it work?* IBM. https://www.ibm.com/blogs/internet-of-things/what-is-the-iot/

*Digital transformation and technologies through the human prism*. (n.d.). I-SCOOP. Retrieved April 19, 2022, from https://www.i-scoop.eu/digital-transformation/human-prism/

Dykes, B. (2021). *A Deeper Dive into LEGO Bricks and Data Stories* [Review of *A Deeper Dive into LEGO Bricks and Data Stories*]. https://www.effectivedatastorytelling.com/post/a-deeper-dive-into-lego-bricks-and-data-stories

Herpy, J. (n.d.). *Council Post: Smart Contracts And The Law: What You Need To Know*. Forbes. Retrieved April 19, 2022, from https://www.forbes.com/sites/forbesbusinesscouncil/2022/03/17/smart-contracts-and-the-law-what-you-need-to-know/?sh=496e85f83d03

Lacity, M. C., & Tapscott, D. (2020). *Blockchain Foundations: For the Internet of Value* (1st ed.). Epic Books.

Leonhardt, M. (2019, December 27). *The 10 biggest data hacks of the decade*. CNBC. https://www.cnbc.com/2019/12/23/the-10-biggest-data-hacks-of-the-decade.html

Marin, J., 2022. *Centralized vs Decentralized vs Distributed: a quick overview*. [online] Medium. Available at: <https://medium.com/@juliomacr/centralized-vs-decentralized-vs-distributed-a-quick-overview-1f3bd17b8468> [Accessed 21 April 2022].

Rabbitte, P. (2020, January 27). *Smart fridges set to get smarter with IoT & AI functionalities*. 150sec. https://150sec.com/smart-fridges-set-to-get-smarter-with-iot-ai-functionalities/13576/

Redman, J. (2017, June 22). *Chainalysis Says They've Found the Missing $1.7 Billion Dollar Mt Gox Bitcoins – Featured Bitcoin News*. Bitcoin News. https://news.bitcoin.com/chainalysis-says-theyve-found-the-missing-1-7-billion-dollar-mt-gox-bitcoins/

Reiff, N. (2022, April 21). *Chainlink (link) definition*. Investopedia. Retrieved April 26, 2022, from https://www.investopedia.com/chainlink-link-definition-5217559

Rovai, M. (2021, January 6). *Home Automation with Alexa*. Medium. https://towardsdatascience.com/home-automation-with-alexa-69f71524f164

Silver, C. (n.d.). *Council Post: What Is Web 3.0?* Forbes. Retrieved April 19, 2022, from https://www.forbes.com/sites/forbestechcouncil/2020/01/06/what-is-web-3-0/?sh=14a10ec458df

*The Blockchain Analysis Company*. (n.d.). Chainalysis. https://www.chainalysis.com/

Vorhies, W. (2016, July 26). *CRISP-DM – a Standard Methodology to Ensure a Good Outcome*. Data Science Central. https://www.datasciencecentral.com/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/

(2020). *An Ultimate Beginner's guide to Blockchain Technology* [Review of *An Ultimate Beginner's guide to Blockchain Technology*]. https://digiteer.in/an-ultimate-beginners-guide-to-blockchain-technology/

## -Appendix-

*Data* – Data is information that can be processed and analyzed to reveal trends and causations for certain experiments.

*Qualitative* – a type of data where no numerical values, strictly observational.

*Quantitative* – a type of data that are only numerical values.

*Big Data* – a term to describe millions of lines of data.

*Analytics* – The process by which raw data can be understood whether by metrics or through visual dashboards.

*CRISP* – The cross-industry process for data mining, is used to ensure accurate results when data mining.

*Supervised* – Consists of regression analysis, decision trees, and neural networks. In this type of analysis, there is a known variable the analyst wants to know more about.

*Unsupervised* – Clustering, where raw information about the data is produced to highlight the prevalence of what is being seen in the dataset.

*Internet of things* or *IoT* – Concept where individual devices can share data through the internet.

*Machine Learning* – Development of computer systems that can learn from past trends without the need for instructions.

*Blockchain* – Information kept on distributed ledger supported by multiple computers on a peer-to-peer network.

*Distributed Ledger* – A consensus of shared, replicated, and synchronized data spread across multiple locations.

*Node* – a device that can communicate directly to the blockchain.

*Proof-of-Work* – a method of different nodes competing with each other to solve an equation that validates a transaction, the first node to complete the equation is rewarded in said crypto.

*Proof-of-Stake* – a node that possesses a high amount of a specific cryptocurrency staked is rewarded with the ability to validate transactions.

*Decentralized Finance or Defi* – Blockchain concept that offers financial instruments without the need for a 3rd party entities like brokerages, exchanges, and banks.

*Blockchain Analytics* – Analyzing, identifying, and clustering data present on the blockchain to find trends and ween off bad actors.

*Smart Contract* – A transaction protocol that is a computer program used to automatically execute documentation when stipulations of all parties are met.

*Metaverse* – Emerging web 3 application that creates a decentralized digital world.

*Web 3.0* – Next form of internet enabling users to fully own their data.

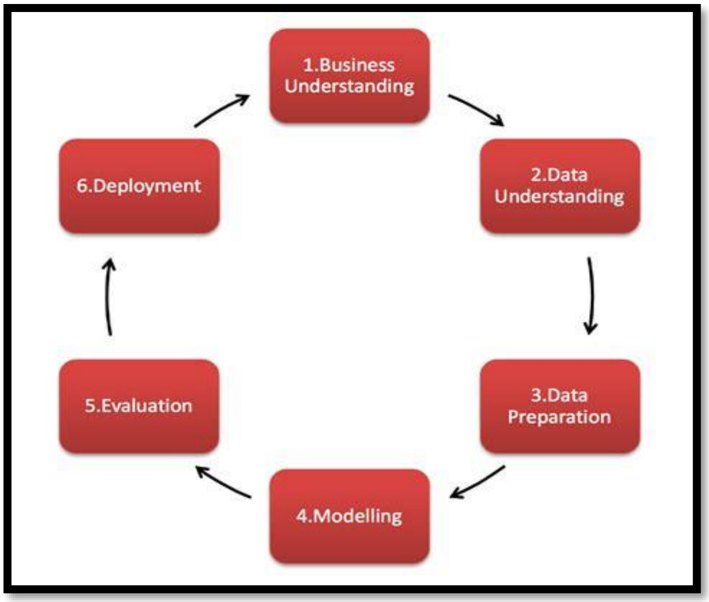


**FIGURE 3. CRISP METHODOLOGY**

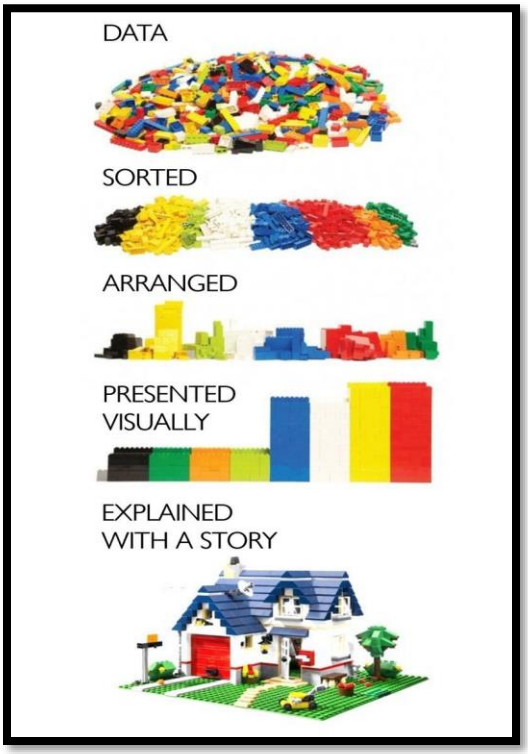


**FIGURE 4. LEGO VISUAL EXPLAINING ANALYTICS**

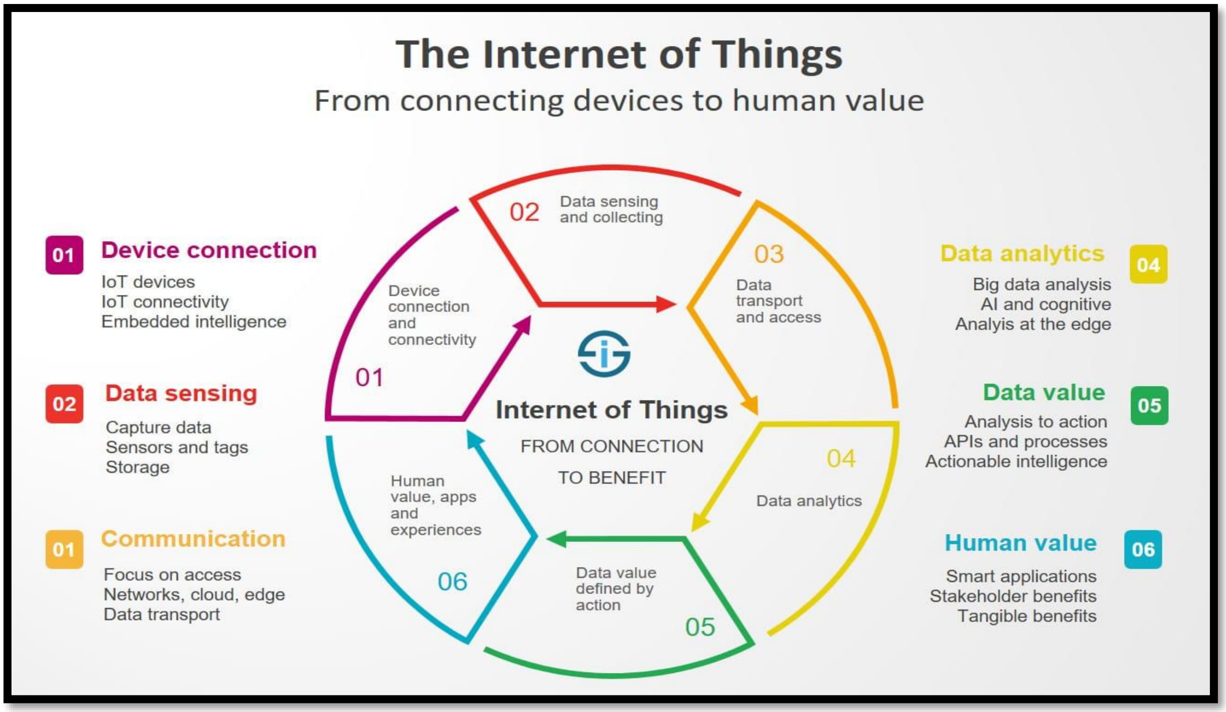


**FIGURE 3. INTERNET OF THINGS GRAPHIC**

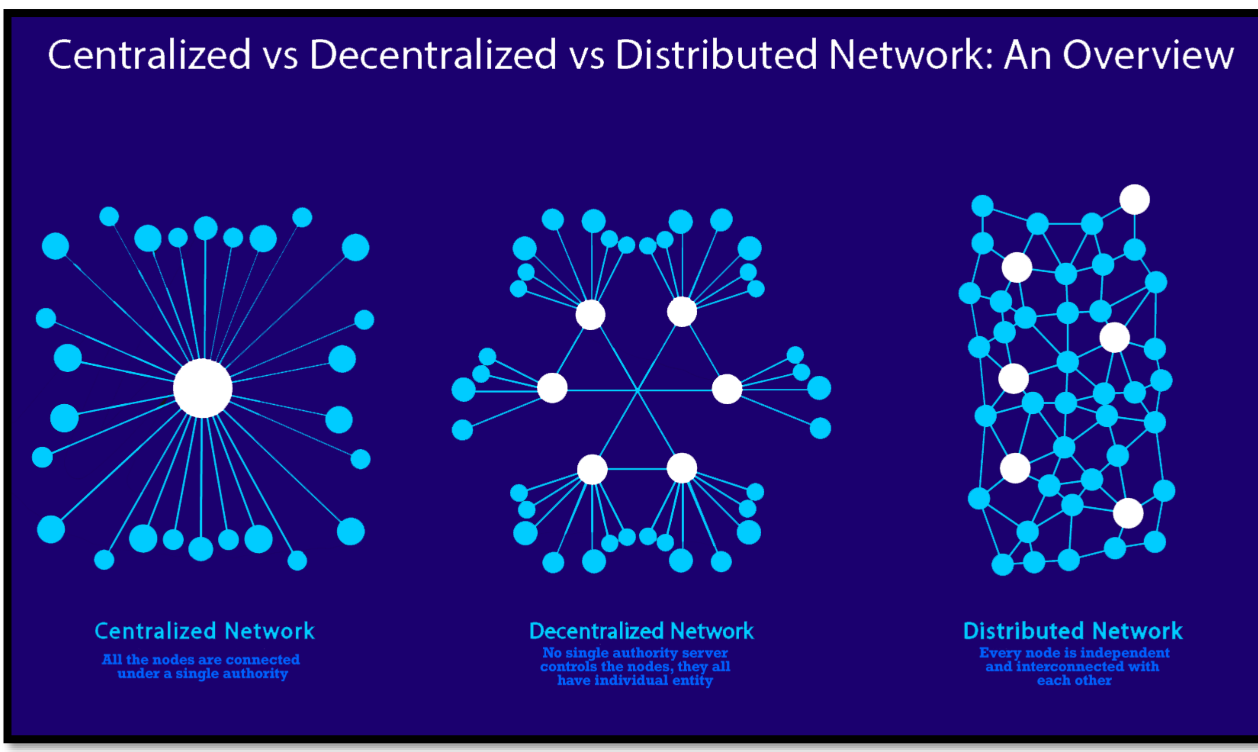


**FIGURE 4. CENTRALIZED NETWORK VS DECENTRALIZED NETWORK**