# TREE*N*TEQ : A data liquidity platform.

## *Litepaper*

## *V.01*

### I. The state of data.

1. ***Data is everywhere***. It's the lifeblood of the global economy. The amount of data generated, consumed and ingested by individuals, institutions, communities, businesses is growing at an exponential rate. This data can be personal, machine generated, proprietary, or public data. We call these data entities to whom the data relates, the '*data principals*'. It is possible that data principals are not always individuals, communities or juridical entities, they can be machines as well. There is an emerging trend of Artificial Intelligence agents acting on behalf of traditional data principals. These agents may hold data through various technological means. Though these agents can act autonomously, and take complex decisions about data in their control, we assume that these agents would always have juristic managers, and thus the ultimate data principal would be a legally recognized entity.

2. ***Data is monetized***. Large scale businesses globally monetize data generated by data principals. Sophisticated business models ingest large reams of data to provide goods and services to public and private clients. Various governments, through multiple agencies, also collect public and private data to improve governance and meet sovereign objectives. An emergent trend is of small scall businesses, start-ups, individuals also collecting and seeking out data to build products, generate insights, and meet other similar objectives. This trend is expected to accelerate due to the availability of affordable GPUs tailor made for training and deploying AI systems, primarily based on large language models. We call all the above users of data '*data processors*'. Given recent technological advancements in AI systems and allied hardware, all data principals can also be data processors.

3. ***Data is siloed***. When data is created by data principals, it is most often stored independently by these data principals. Most data are held centrally, be it on the data principal's personal hardware devices, on cloud storage services, with data brokers, with data processors who have exclusively licensed their data, or with governmental/ quasi-governmental authorities. Often, data principals do not have meaningful custody or access to their own data. Over the past years, the trend of storing data on decentralized file sharing systems has also been on the rise.

4. ***Data is illiquid***. Owing to social, contractual and technological factors like the data being in siloes, restrictive access, lack of counterparty trust, licensing arrangements, and contractual requirements, data has been rendered an illiquid asset. Most data processors either collect data directly from their users, offering products and services for free- often commoditizing their users, or from businesses who specialize in purchasing data from third party data principals or processors. This purchased data creates a secondary market for data, which is subsequently aggregated, packaged and sold to potential data processors. These businesses are called '*data brokers'*. The sale and purchase of data by and through data brokers are by way of private contracts, and no information regarding such transactions are publicly available, except in certain cases. Data belonging to data principals who do not retain exclusive rights over their data, having sold, assigned or licensed their data to the data processors, may have their data sold multiple times in closed door secondary markets, to different data processors without their knowledge and often to their detriment. Additionally, if a data principle or a data processor seeks to sell their primary or secondary data, they do not have open access to robust public marketplaces.

5. ***Data is undervalued***. Even though data principals transact their data daily, there is no price discovery of the price of specific data. Data subjects often receive goods and services in exchange for their data. These transactions are technically barter transactions. These data barters constitute most of the data transacted. Where data is transaction for money, the price of data is not publicly made available, as these transactions are private. Due to the opacity of the current data markets, there exists information asymmetry as to the value of data, to the detriment of the data sellers. We posit that the combinatorial effect of current data marketplace is the gross undervaluation of data. The fact that most data principals, especially businesses, that own data do not reflect data as an asset in their books of accounts is testament to this undervaluation.

6. ***Data has network effects***. A single bit of data has limited value. As more data of the same type is aggregated, the value of each individual unit of data increases. Similarly, a single type of data has limited value and utility. As more data of different types are aggregated and connected, the value of both data types increases, exponentially.

7. ***Data is under-utilized***. From a productivity perspective, data is inexhaustible, as it can be used multiple times. This allows data to have potentially limitless use cases, mirroring limitless demand. This provides multiple monetization models for the same data. However, despite this inherent characteristic of data, it is often not used beyond its initial use case due to custodial, contractual and liquidity issues. Certain

data also becomes stale owing to its individual characteristics, thus making it perishable. If such perishable data is not processed during certain time periods, its value can diminish precipitously.

8. ***Data deserves to be liquid and listed***. Data, like fossilized oil, is sans value if trapped underground. It is only when data is aggregated, cleaned, and annotated, that it becomes fit for processing. We call this collection of pre-processing steps '*refining*'. Refining is the first step to liquidity. The second step would be the tokenization of data. Tokenization entails wrapping the refined data within a smart contract deployed on a blockchain. We argue this blockchain should be public for purposes of ensuring trust, transparency and security. Once tokenized, the third step requires the data to be stored. This storage can again be centralized or decentralized. We argue for storing the data in decentralized systems, using protocols like the Inter Planetary File System (IPFS) and other layers built on such protocols. This allows the data to be stored in a decentralized manner, attracting all benefits of decentralization, while ensuring legal custody, access and redundancy for the data principal. Decentralized data is also exigible to centralized or decentralized computation. The fourth and final step is listing the tokenized data on a public data marketplace. From this marketplace the listed, tokenized data can be exchanged, transferred, or transacted. Many transaction modules can be designed into the marketplace to facilitate nuanced, bespoke or templatized data transactions. The transactions can be global and automated, with human or machine counter parties. We propose such a public marketplace for listing tokenized data assets, which ensures liquidity to the tokenized data.

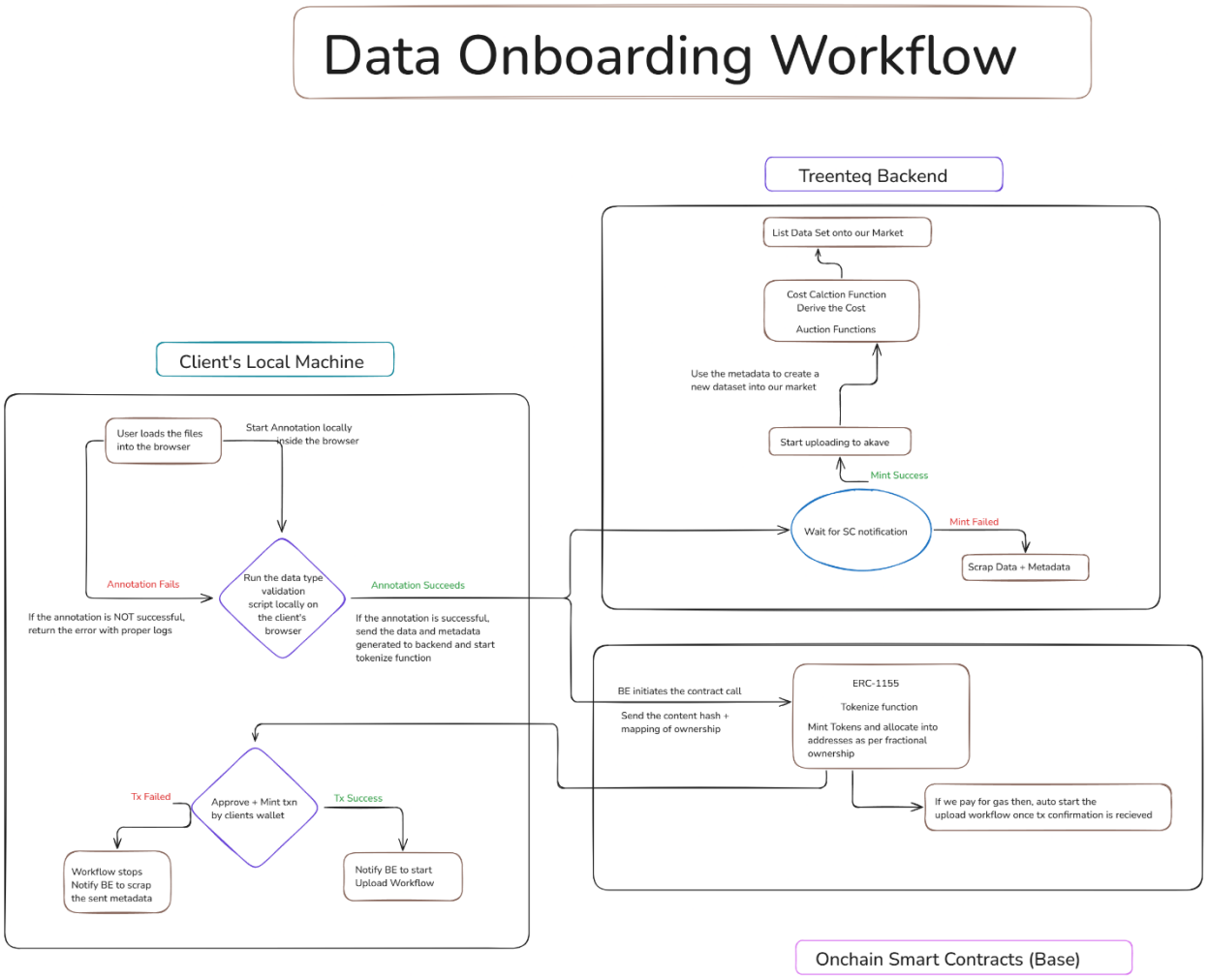
9. ***Other projects.*** There are early-stage projects that are currently attempting to solve the data liquidity problem as articulated by us above. They seek to do so by building public data marketplaces. These projects predominantly mediate data transactions using cryptographic tokens, also called digital assets or cryptocurrencies. We find that the leading example of a public tokenized data marketplace today are in the form of DataDAOs, which contain within them Data Liquidity Pools. However, regulatory uncertainty and restrictions applicable to digital asset projects pose significant hurdles in various jurisdictions and impact the potential for scaling a tokenized data marketplace at population scale. Lack of familiarity with web3 tools and systems also poses a non-negligible challenge in scaling potential web3 data marketplaces.

## II. Proposed solution architecture

1. ***Platform architecture.*** The marketplace for tokenized data built by us consists of four broad components. The first component is the '*data ingestion module'*, which

enables the data providers to upload and submit their datasets. The second component is the '*data refinement engine'* which allows for the annotation, organization and validation of datasets to ensure quality. The level of refinement can be modified and made more robust as the market place grows. Once the data is refined, we proceed to the third step of tokenizing the dataset. This is done through the '*tokenization framework*', which is a designed framework for converting refined datasets into tokens that can be traded or sold. Last component of the architecture is the '*client access portal'*. This portal is a an interface for clients to search, access, and purchase datasets using tokens or even fiat.

2. ***The data onboarding workflow.*** As represented diagrammatically below, the workflow is bifurcated into two parts, the clients local environment and treenteq's backend. The core innovation of our proposed architecture is the manner in which data is onboarded from a clients internal environment, to a distributed storage system, via our platform. The workflow is bifurcated into two parts, the client's internal environment and the treenteq backend. The detailed workflow is represented diagrammatically below:



## III. Anticipated benefit

1. Once a solution as described above is deployed, it stands to benefit all stakeholders in the data ecosystem, be they data principals, data processors or data marketplaces.

2. ***Monetization.*** Data principals who are individuals can monetize their personal and non-personal data through a multitude of means. They, armed with means of discovering the price of their data, can transact freely with their data in a near permissionless way. This allows independent data principals monetize their data and earn an income where there is none today. This income may be basic or substantial, recurring or one time. In all scenarios, if the data has real world applications, its price is almost guaranteed to be non-zero owing to market forces. This, we note, would be a very potent tool in the transitioning society from a pre-AI to a post-AI world. It is understood that AI systems would be responsible for large scale unemployment, placing many individuals in a rather precarious position. Monetizing data by individuals can be a mitigating factor against potential unemployment or diminished incomes.

3. Data principles who are judicial entities, like companies, societies, trusts, partnerships, and even municipalities and governmental agencies can supplement their existing revenues with little to no additional cost. This is relevant as such entities would presumably be in possession of large amounts of data generated as a byproduct of their day-to-day activities. This can be legacy or recurrent data. The potential to earn from such institutional data is unlimited, while the costs are limited, and in many cases, already incurred or sunk. Such additional revenue streams may have knock-on effects in all business and governance models, including reduction in the cost of goods and services offered, or taxes levied.

4. ***Availability.*** Data processors will have access to previously unavailable data, at fair and transparent prices. The provenance of data will be provable resulting in reduced risk of liability, especially when data is used for training AI models. We expect an exponential multiplication in the number of data processors, and new business models to be built and products offered. Competition is also likely to increase between data processors, resulting in positive externalities. In the context of computation, the most significant impact we anticipate would be felt by individual and small-scale enterprises, who require legal access to large data sets to train bespoke AI systems and models, leveraging personal AI supercomputers and retail next-gen robotics and edge solutions.

5. ***Price discovery***. A further benefit we see for data principals and processors is the ability to attain real market prices for their data, which is not available in the data industry today. Contemporaneous data valuation methods valuate data based on derivative methods, which leads to difficulty in ascribing 'fair market value' to any particular data set, especially for taxation and corporate compliance. With access to pricing for data, data processors are poised to classify data as 'assets'. This transformation is further made possible as reliance on services provided by third party data valuation service providers will no longer be mandatory or strictly required. With operational data markets, it will become significantly easier and cheaper to attain accurate, legally acceptable pricing for for data assets. This should usher in the era of data being considered a tradeable, reportable asset class. It may be noted that data assets have a large 'option value' owing to the perpetual development of new technologies which would require data, thus constantly growing the demand for existing data and potentially expanding its demand, and consequently, its price.

6. ***Data as digital oil.*** The analogy of data being digital oil is valid and appropriate. Today data, like oil over a century ago, is trapped underground in silos as noted above. A public data platform marketplaces stand to unlock significant value from mining data. Unlike oil, which generated private profits and public negative externalities, data would be the very opposite. Profits from the data market places can be decentralized and distributed equitably, all the while creating positive externalities. The positive externalities are caused due to the impact data based products and services stand to have in the economy.

7. ***New data marketplaces.*** We expect an exponential increase in computational abilities as well as advancements in AI system design, including LLMs. With the ability for smaller players to design, train and deploy their own foundational purpose-driven LLMs, we anticipate a corresponding demand for clean, liquid, legally available data. This would potentially lead to more market players entering the public data marketplaces to meet the growing demand for data.

## IV. Roadmap.

1. ***Querying 2.0***. In addition to expanding the types of datasets available and enhancing the tokenization process, we intend develop advanced technologies that enable users to perform computations directly on the data within our platform. This feature will allow users to submit analytical queries. Two very basic but instructive examples of this feature would be arming users with the ablity to initiate targeted queries about economic data about the percentage change in GDP for a specific country over a defined period; or request calculations related to carbon emissions

reductions based on various climate data sets, facilitating informed decision-making for sustainability initiatives, thereby empowering enabling economists and researchers to derive insights without needing to access raw data directly. This stands to exponentially accelerate research and development cycles across industries.

2. ***Integrated encryption.*** To achieve the ability to query raw data sets, we plan to implement a distributed node architecture that leverages cutting-edge encryption technologies. These would include the following

   ***(a)*** **Multi-Party Computation** (MPC). MPC allows multiple parties to jointly compute a function over their inputs while keeping those inputs private. For example, different nodes can collaborate to analyze sensitive financial data without revealing their own information;

   ***(b)*** **Fully Homomorphic Encryption** (FHE). FHE enables computations to be performed on encrypted data, ensuring that sensitive information remains secure throughout the analysis process. This is particularly useful for private and personal data, where privacy is paramount.

   ***(c)*** **Zero-Knowledge Succinct Non-Interactive Arguments of Knowledge** (ZK-SNARKs). This cryptographic method allows one party to prove to another that they possess certain information without revealing the information itself. For instance, a user could prove they have the necessary credentials to access specific datasets without disclosing their identity.

3. ***Incentivization.*** Upon completion of the computations, users will receive the results while compensating the platform with tokens. These tokens may be internal or external tokens issued by us on a public blockchain. The token distribution as contemplated currently is as under:

   ***(a)*** **Node Compensation**. Tokens will be allocated to the distributed nodes that performed the computations, incentivizing participation and ensuring fair compensation for their resources.

   ***(b)*** **Data Owner Royalties**. A portion of the tokens will be allocated as royalties to the original data owners whose datasets were utilized in the computation, promoting a fair revenue-sharing model.

   ***(c)*** **Network Fees**. Similar to traditional blockchain architectures, a small network fee will be applied to facilitate the transaction and maintain the platform's operational integrity.

## V. Conclusion

Though ambitious, we are convinced that the problems noted by us need solving. Our platform Treenteq will transform data into a liquid, accessible, and tradeable asset class by addressing the challenges of siloed, undervalued, and underutilized data. By introducing a platform that enables tokenization, decentralized storage, and transparent marketplaces, Treenteq will unlock the latent value of data for all stakeholders. Our aim is to ensure fair monetization for data principals, transparent access for data processors, and the creation of a robust, decentralized data economy. With advanced features like encrypted computation and equitable revenue-sharing models, Treenteq not only democratizes data access but also stands to be a pivotal infrastructure layer for the post-AI world, ensuring that data's economic and social potential is fully realized.

***