Generating Income from Your Data

Whitepaper Draft 1

DataNation.Guru

November 2024

Table of Contents

1 Motivation

Today, we stand at the edge of something truly remarkable. In just over two short years, generative AI has given us a way to transform the way we interact with technology, with breakthroughs like ChatGPT showing us what's possible. We are in the midst of the next great revolution “The fourth industrial revolution” and generative AI is at the heart of it. This isn't a passing trend. It’s going to change everything we do, and it's here to stay.

Every few weeks, we see an incredible number of companies and individuals pushing the boundaries, sharing their AI models, and driving innovation at a pace we’ve never seen before. However, here’s the question: can it go faster? Absolutely, data is the driving key factor and there just isn’t enough of the right high quality data to train these models. Data will help build these models, whether its specialised or not, data needs to be gathered and presented in a format that it can be utilised by these AI models.

Right now, many of the Large Language Models(LLMs) you see on platforms like Hugging Face are just static tools, artifacts built for inference. But here's the problem: there’s no real community around the data. No place to contribute, no way to improve or enrich the datasets. Everything about data needed for LLMs is left to the data owner, this slows innovation down.

LLMs of any size have the potential to assist professionals across various industries, the key to their success is high quality data. Companies large and small have large amounts of domain specific data that could be used to generate additional incomes. This data sits in silos within these companies, the companies are unsure how their data can be utilised to generate an extra income for their businesses, they do not know how to address privacy or other regulatory issues. Unlike tangible assets, which can be protected primarily through physical means, intangible assets such as data require additional considerations.

How can we bring the spirit of Data collaboration to AI? How do we build communities where we share knowledge, income, contribute to datasets, and ultimately create AI models that are more accurate, more powerful, and therefore more innovative? The need is clear, it’s time to bring it all together.

Data Nation is developing a framework enabling "programmable" Data IP on the blockchain. This will also enable fractionalised sharing of income from Dataset assets. Our goal is to establish a specialised set of smart contracts, empowering companies or individuals to monetise their dataset IP based on transparent, creator-defined terms as well as ensuring the privacy of data is at the centre of the platform along with addressing regulatory issues.

Data Nation democratises data and expertise where data and knowledge is shared via the platform. Where custom datasets can be created either upon request or placed on the platforms marketplace for sale and fractional ownership.

Data Nation utilises the Cardano blockchain technology, enabling contributors and stakeholders of Data to be rewarded fairly, fostering continuous innovation and growth for Data analysts and AI Engineers on a decentralised, secure and reliable platform.

This is combined with a seamless user experience with our partners Maestro https://www.gomaestro.org for payment processes and https://iagon.com/ for data storage. Our marketplace users do not need to be aware that they are interacting with a blockchain.

Data Nation provides a private, decentralised, reliable and cost effective platform enabling sustainable income anywhere across the globe; business that would otherwise be taken by the centralised companies in the AI platform ecosystem such as Amazon, Google, Microsoft and others.

2 Introduction

This document (the “Data Nation paper”) provides a high-level specification of the DataNation.Guru Platform. The DataNation ecosystem is dedicated to support, educate along with social responsibility; striving to contribute positively to a global DataNation community and its future development.

DataNation’s platform challenges the way that Data is improved, valued, monetised, stored and sold. DataNation does this by using the unique characteristics of blockchain, decentralised control with a focus on dApps. Using the capabilities of the Carano ecosystem to allow for greater privacy for end-users and their data, at a cost effective price that also delivers an income for contributors and stakeholders.

3 DataNation Assets

DataNation Usage via Ada or fiat currency

Participation in the DataNation ecosystem is facilitated through purchasing usage credits via Ada or a fiat currency, there is no requirement for a native token.

DataSet Stakeholders publish their Dataset assets on the platform allowing data to be cleaned and transformed thus improving the quality of the data. This is carried out using DataNation platform tools and ensuring any privacy issues are addressed. The data quality can be improved by the Stakeholders or requests can be made for Data Analysts or AI Engineers on the platform to make improvements to datasets. This can be either for a cost of time or a fractional share of the income generated by the dataset. Once the datasets quality is improved it can be quality audit checked and valued via the platform’s algorithms and then published on the Data Nations MarketPlace.

Potential dataset customers have their requirements matched with datasets, these customers will then have access to the high quality datasets fitting their requirements with a valuation and detailed explanation of the data within the dataset.

Dataset Stakeholders will be able to sell fractional ownership income rights to their datasets if they want to do so on the MarketPlace. This will enable any income generated from the dataset to be shared equally between the fractional owners.

Dataset Stakeholders and Fractional owners will be paid any income generated in Ada and will therefore need to connect their Cardano Wallets to the platform.

Fractionlised NFT Tokenisation of a Dataset

A fractionalised NFT is an NFT divided into multiple pieces, allowing for fractional ownership of the original token. This process lowers the barrier of entry, as the NFT fractions cost less than the full token.

CIP-68 enables the creation of tokens with programmable metadata. This means that the metadata associated with a token can be updated and modified without the need to mint new tokens. This flexibility allows for the fractionalised ownership of NFTs on Cardano.

Each tokenised Dataset will include a Data Nation reference NFT and 1,000 user tokens which will contain metadata that allows new Stakeholders to own a fraction of a Dataset and access the treasury contract to withdraw income. The NFT has an asset certificate, version hash, KYC barcode, fractional sale and lock in time period which can be set via input options from the Dataset Owner, the longer the lock in period the more income is generated for the fractionalised token owners. The user tokens linked to the NFT can be sold or given by the original owner to contributors that have made quality improvements to the Dataset.

Each Dataset will contain an embedded hash of the NFT asset certificate so that it can be traced back to its owners.

Each time a new Dataset is uploaded onto the Data Nation platform, it will have a hash of the asset certificate for identification purposes stored within it, this will be stored under a hidden hash key.

Who will use DataNation.Guru (URL placeholder)

Individuals or companies will all carry out KYC, the types of users will be: -

1) Data Analysts and AI Engineers – these can be individuals, a group of individuals or a company.
2) Purchasers of Dataset usage – these are the customers that use the Datasets on the platform
    a. A customer can pay to use the Dataset which creates an income for that Dataset
    b. A customer can purchase a Dataset outright and prevent future customer usage having sole rights to that Dataset
    c. A customer can purchase (a share) fractional tokens of a Dataset and benefit from any future income generated by the Dataset
3) Dataset owners can also request enhancements to their Datasets for specific quality improvements.

4 Roadmap

2024-2025 (Stage 1)

1) Integrate Maestro https://www.gomaestro.org for payment processes.
2) Integrate with https://iagon.com/ a DePin (Decentralised Physical Infrastructure Network) for distributed data storage.
3) Implement the monetisation of Datasets via smart contracts.
4) The Data Nation platform will enable individuals/companies to register and carry out KYC
5) Enable companies to deploy Datasets onto the platform along with the ability to enhance them.
6) Develop dataset quality improvement tools i.e. Data cleansing and transformation
7) The platform's algorithms check the quality of the dataset and value the contents
8) Publish the dataset onto the MarketPlace (this is optional as some companies may use the platform to just improve their datasets only)
9) Allow income usage analysis of the Datasets.
10) Enable users to test the first 1000 records from any Dataset after which they can -
    a. Purchase a fractional share of the Dataset
    b. Purchase the Dataset outright taking it off the marketplace.
    c. Owner requests and pays or gives fractional income share of their Dataset to a Data Analyst/AI Engineer on the Marketplace who enhance their Dataset

Late 2025 (Stage 2)

1) Store Treasury in Stablecoin
2) Think about a possible ICO, TGE
3) Future Design enhancements - how LLMs can be placed onto platform with their training Datasets and digitised

5 Income Generation

To ensure that Owner and Data Analysts/AI Engineers receive fair compensation for their work, blockchain technology will be integrated into the payment and licensing processes. Blockchain's decentralised and transparent ledger system provides a secure and tamper-proof method for recording transactions and agreements.

Smart contracts, which are self-executing contracts with the terms of the agreement directly written into code, can automate payments to developers based on usage metrics, purchase or payment terms. This system not only guarantees timely and accurate payments but also enhances trust between customers and Stakeholders by providing an immutable record of transactions. Furthermore, blockchain can facilitate microtransactions and fractional ownership, allowing owners to be compensated proportionally to the usage of their Datasets globally in Ada.

Income and sales from Datasets

Income and sales from Datasets can be generated in the following ways:

1) Subscription payments to Data Nation platform for Dataset tools usage and deployment
2) Selling fractional ownership of Datasets, this price is set by the owner of the dataset
3) Buying a Dataset outright. Again price set by the original Owner.

Who is in Charge of Setting Up the Contracts on Data-Nation

Creators and Dataset Rights Holders:

- **Smart Contract Deployment:** The original Creators / owners of the Datasets are responsible for initiating and specifying smart contracts parameters that represent their ownership of their Dataset on the Data Nation platform.
- **Defining Terms:** They set the terms on the Data Nation platform for sale or fractionalised revenue sharing within these contracts as well as the lock in time period.
- **Content Registration:** By registering their Dataset on Data Nation, they establish a verifiable record of ownership and terms of use or sale.

Data Nation Development Team:

- **Platform Development:** The Data Nation development team maintains the platform infrastructure and tools.
- **Standard Contract Templates:** The Data Nation development team also provide parametrised smart contract templates to automate the process for creators or Dataset rights holders to deploy their income fractional criteria onto Data-Nation within a smart contract.

Dataset purchase

The Dataset Marketplace is where Datasets can be sold. Datasets can be sold on the marketplace where a user will be able to buy fractional ownership in Datasets and share in the income they generate.

Datasets are initially valued by the original Owner of the Dataset before putting up for sale, each fractional token is split according to an initial value given. i.e. 1,000 asset tokens at $20 each = $20,000 value of a Dataset.
The Dataset subsequently generates an income from usage or a sale of the Dataset.
The sale price of any Dataset is set by the creator/owner. The Dataset can be sold outright if the owner sets a bid price.

Proceeds from sales are paid into the Sellers Wallet. Periodically owners request income amounts owed, these are deposited into the fractional owner's wallet via the Treasury Contract

When a Dataset is added to the platform a unique encrypted asset identifier is embedded into the Dataset, this Hash links a reference to its NFT assets within the Dataset.

Datasets will be available for evaluation and a customer can purchase a Datasets usage or an outright purchase then it would not be available for future downloads to others.

6 Benefits of Data Nation Assets on Blockchain

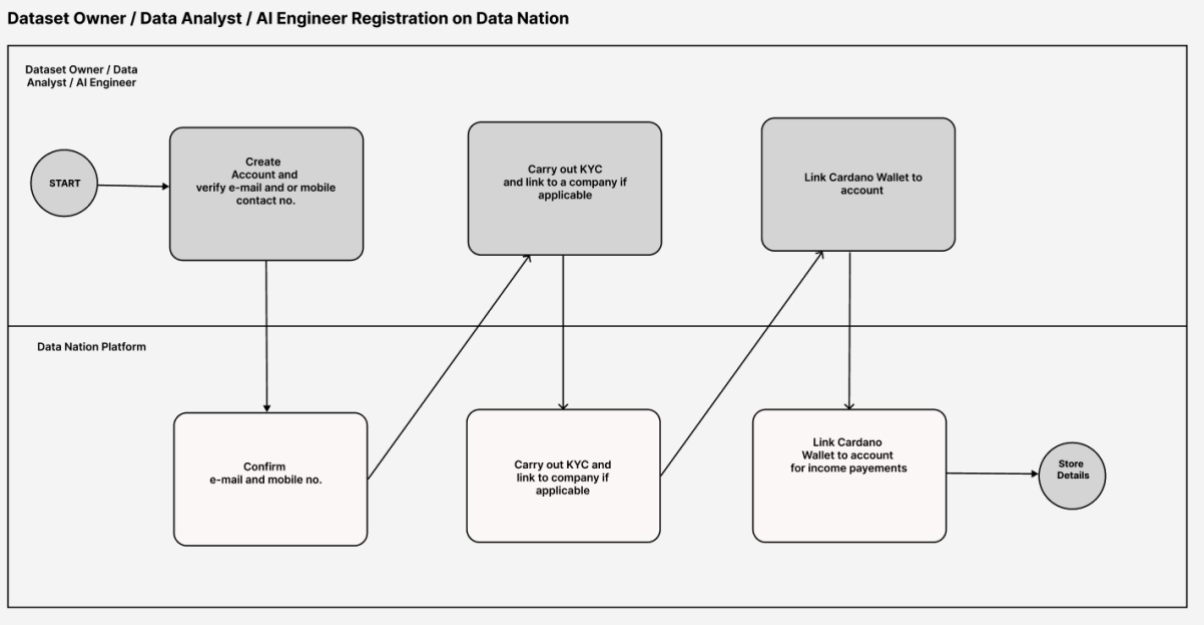
What does the Platform provide regarding on chain data

1) The ability to track Datasets. Creators and owners of all Datasets are all stored on chain proving who created which Dataset and when.
2) Payments made for Dataset usage can be displayed and broken down by fractional ownership
3) Users can view popular Dataset usage and buy into these Datasets with fractionised ownership.
4) Users can view how Datasets have been enhanced with their quality improved and by whom.
5) Data Nation is a totally decentralised private platform
6) Data Nation is a low entry to barrier platform as costs can be micro payments, and a user pays only for what they use.

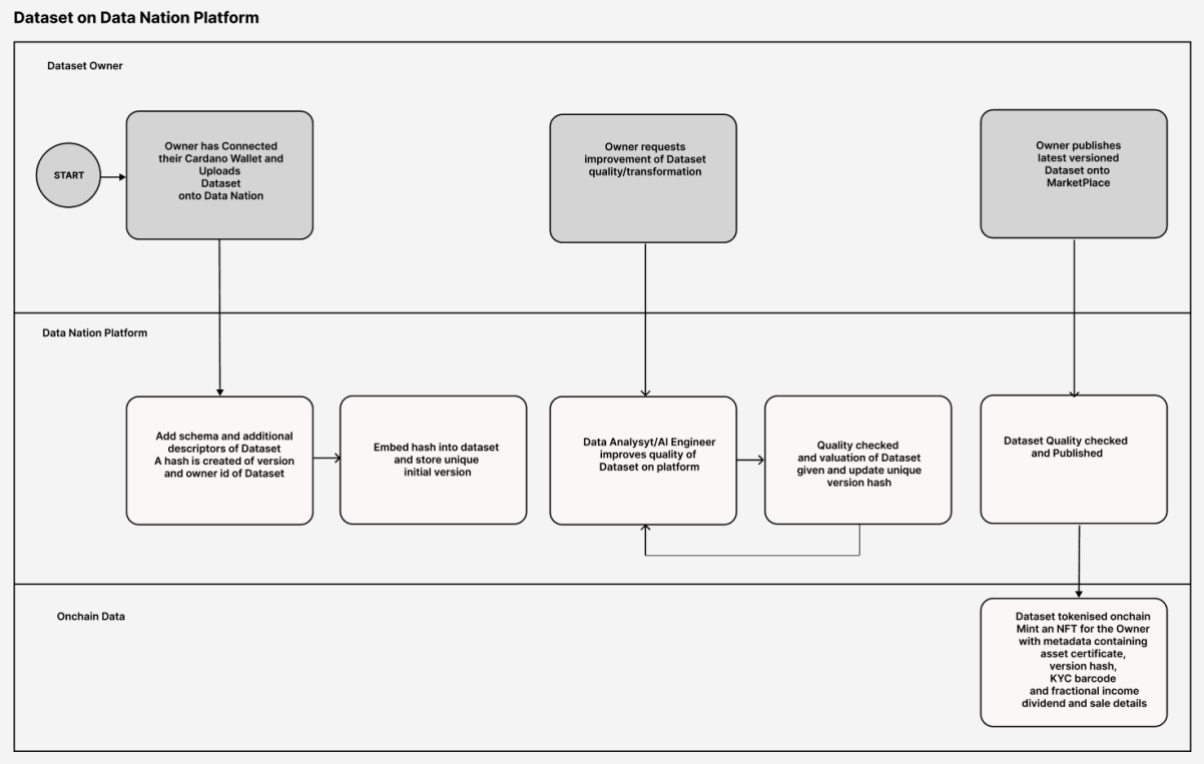
The ability to view the above data on chain gives all Stakeholders and Users the ability to prove ownership rights and usage payments.

7 System Design and Technical Specifications

Registration



Dataset Actions on Data Nation



The following are the essential aspects that make the DatNation platform better than others.

The quality checking is to be carried out by opensource tools such as Great Expectations - https://greatexpectations.io/ others to TBD or built long term.

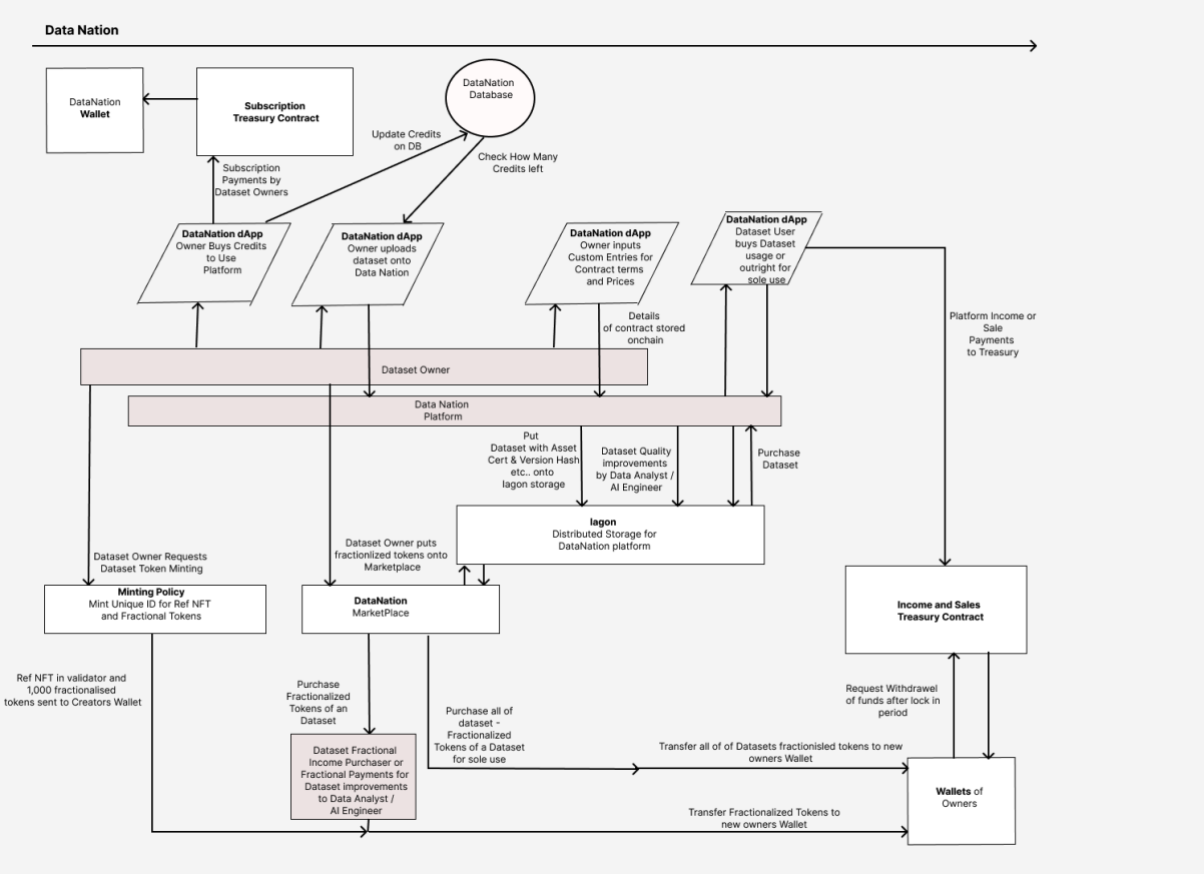
The valuation of the dataset is to be carried out by graph theory algorithms (TBD).
This will quality check the dataset and conclude on what the data can be used for and assign it a value. It may recommend further improvements. This will be an area based on new bespoke algorithms.

This is a major aspect of the Platform as the matching of datasets with potential requirements from customers on the Marketplace (TBD).

These are all key aspects to the platforms success and are not blockchain dependant.

Smart Contract Interactions

LLM Marketplace and API Usage



The above high level diagram shows the interaction between the different actors using the Data Nation platform, and Treasury Smart Contracts, they are in the beige box or rectangles apart, the Dataset user is only shown on the dApp.

These actors being –

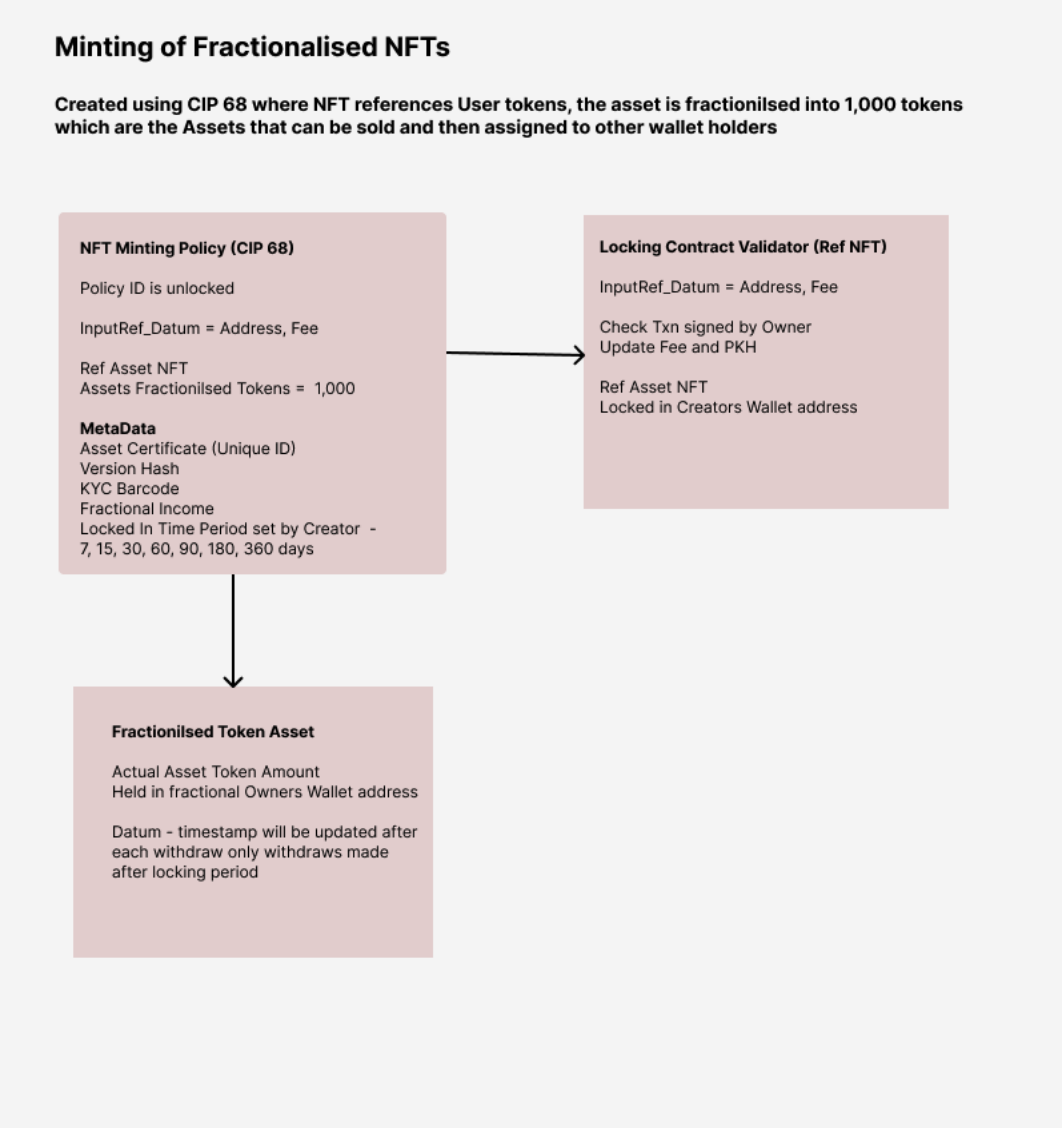
Dataset Owner – the owner and creator of the Dataset. The Owner defines the parameters for the fractionalised CIP 68 asset i.e. at what price can these fractionalised tokens be purchased and can the datset be sold outright if so at what price.

Dataset Fractional Income Purchaser – customer who wants to buy fractional ownership of an LLM so that they can earn income from it over a staking period of 90 days, after which it is re-staked for another 90 days.

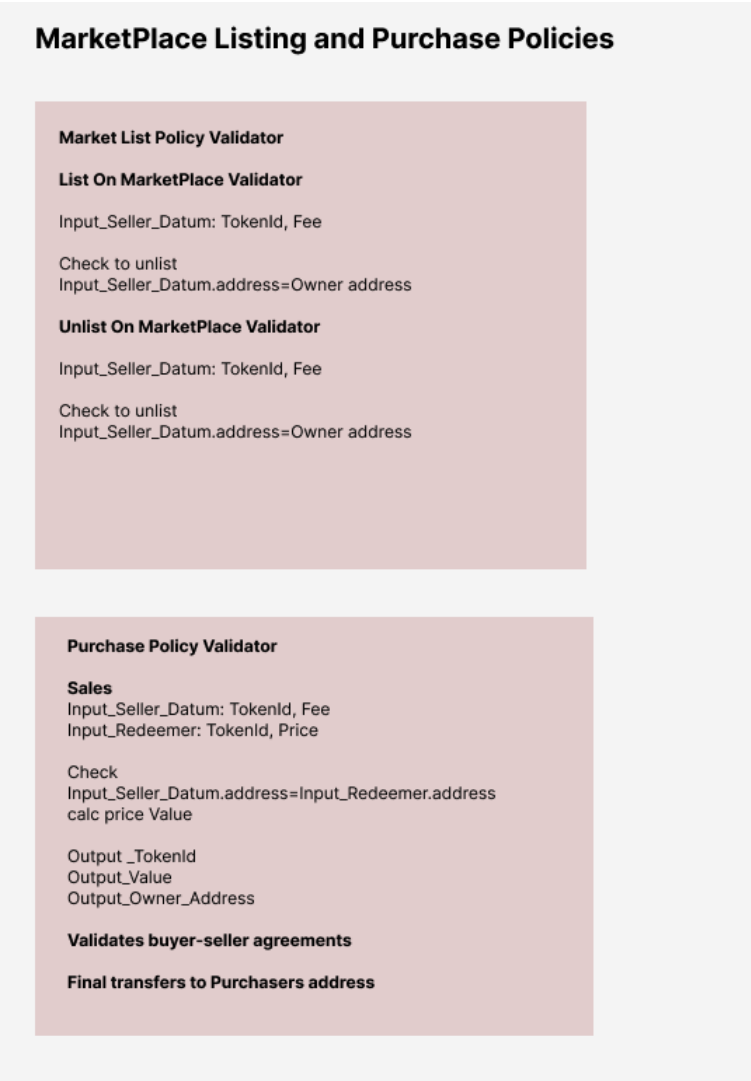
Data Analyst/ AI Engineer– an AI Engineer or company who want make improvments to the Datset on the DataNation platform, they can be paid with fractional ownership of the Dataset or on a cost basis.

Dataset User – a person or company that are using the Dataset.

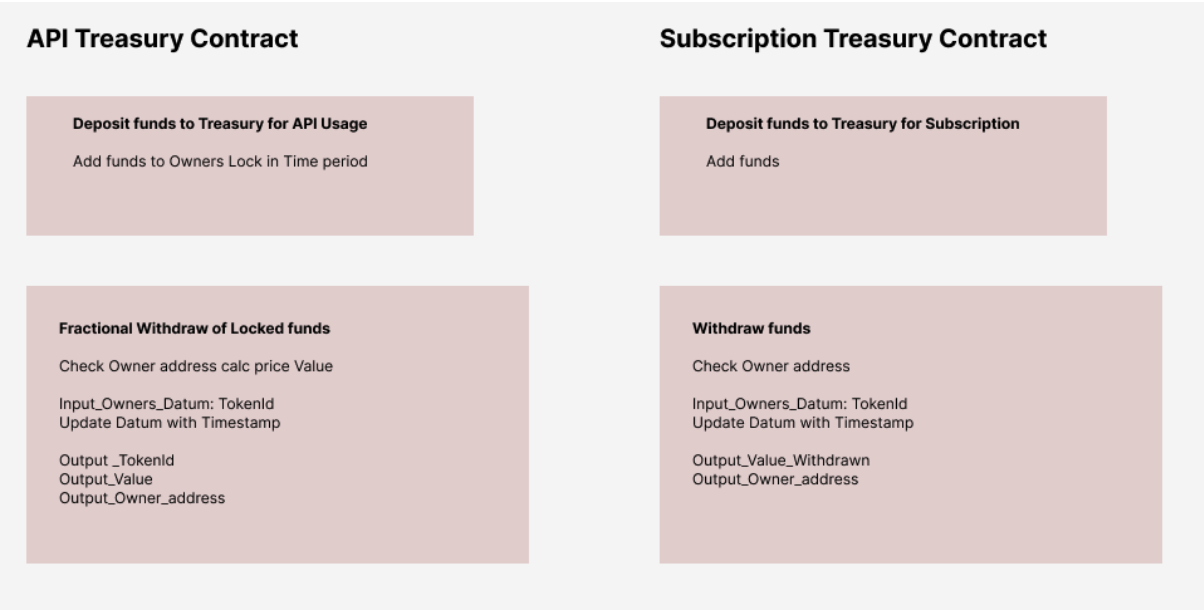
The Minting Contract



Marketplace Contracts for selling CIP-68 FT token assets



The Treasury Contracts



Creating and storing the Fractionalised Asset

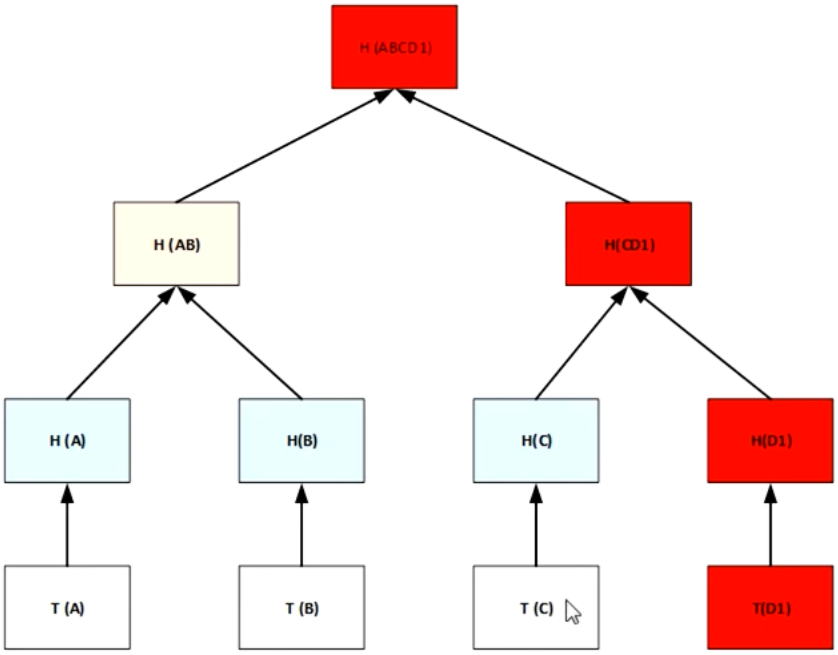
Combining CIP-68 with Fractionalisation

By merging the updatable nature of CIP-68 NFTs with fractional ownership, you create a dynamic asset that multiple parties can own and influence. It works as follows:

1. Minting the Updatable NFT:
    - An ref NFT and associated tokens are minted following the CIP-68 standard.
    - The NFT is governed by a smart contract that allows updates to its metadata and associated tokens.
2. Locking the FT in a Smart Contract:
    - The NFT is transferred to a smart contract that holds it securely.
    - The contract outlines the rules for updating the NFT and managing ownership.
3. Associated Fractional Tokens:
    - Fungible tokens representing fractions of the NFT can be sold on the Marketplace.
    - Each token signifies a percentage of ownership of income from the use of the LLM.

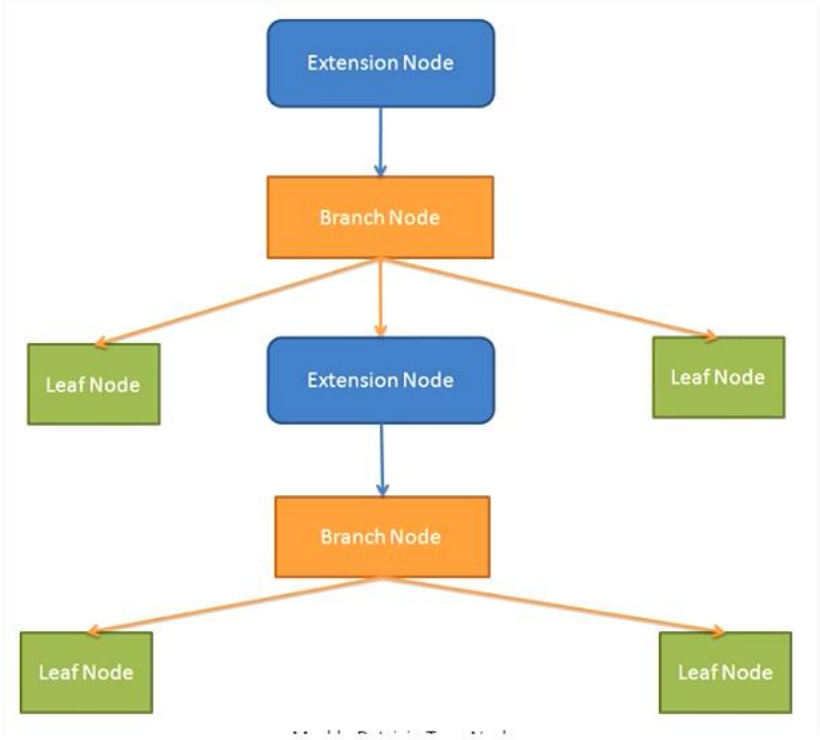
How can we update the fractionalised NFT's tokens

The fractionalised tokens would be stored in a **Merkle Tree** as in the following diagram. The Merkle Tree will enable meta data to be updated on the fractionised tokens. Each node has a Hash which is grouped via the tree branches all the way up to the root node Hash, this ensures that nothing is changed within the tree without rebuilding the Hash's.



Merkle Patricia Forestry Tree

A Merkle Patricia Forestry (MPF) will be used to implement this project. This is an enhanced version of the Merkle Tree where extensions to branch's can be implemented. This is a data structure combining Merkle Trees and Patricia Tries, used for efficient and secure data storage and verification.



A contract will check if an owner is within the Merkle Tree then process the contract accordingly by updating the appropriate tokens(nodes) metadata(datum).

How will this be coded in Aiken

The approach to code this will be based on the Merkle Patricia Forestry implementation as that will enable nodes metadata to be updated, which in turn will represent a fractional token for a LLM.
A Merkle Patricia Forestry (MPF) is a key:value structure which stores values.

The library for this is on Github and linked to below -
https://aiken-lang.github.io/merkle-patricia-forestry/aiken/merkle_patricia_forestry.html

Functional Requirements
Update Functionality

- **Objective**: Implement a function to update a value associated with a specific key within the Merkle Patricia Tree.
- **Requirements**:
    - Accept a **key** and a **new value** as inputs.
    - Traverse the tree to locate the node corresponding to the key.
    - Update the node's datum with the new value.
    - Recalculate hashes from the updated node up to the root to maintain tree integrity.
    - Return the updated root hash of the tree.

Access Control

- Only authorized entities (e.g., users with valid signatures) can perform updates.
- Implement signature verification to authenticate users.

Data Validation

- Ensure that the key and value inputs are valid and conform to expected formats.
- Validate that the update does not violate any business rules or constraints.

Event Logging

- Log events after successful updates for auditing purposes.
- Include details such as the key, old value, new value, and updater's identity.

Updating MetaData on a fractionalised token

As all of the Fractionalised tokens will already have been created, we will only need to update the tokens.

The ‘value’ old and new entries will be updated on the node to reflect the last date a withdrawel was made from the Treasury Contract.
The proof will be obtained off chain it is set up, note proofs are only valid for a precise trie root hash and state. So inserting (resp. removing) any item into (resp. from) the trie will invalidate any previously generated proof.

The update of the Merkle tree will be based on the following Aiken code :

The Update function within the MPF GitHub library will be utilised -

```
update (
  self: MerklePatriciaForestry,
  key: ByteArray,
  proof: Proof,
  old_value: ByteArray,
  new_value: ByteArray,
) -> MerklePatriciaForestry
```

The Aiken coded update function will be extended as follows –

```
fn update(tree: MerklePatriciaForestry, key: ByteArray, new_value: ByteArray) ->
MerklePatriciaForestry {
  // Locate the node corresponding to the key
  match tree.get_node(key) {
    Some(node) => {
      // Update the node's value
      let updated_node = Node.update_value(node, new_value)

      // Recalculate hashes up to the root
      let updated_tree = tree.update_node(key, updated_node)

      // Return the updated tree
      updated_tree
    },
    None => {
      // Handle the case where the key does not exist
      error("Key not found in the Merkle Patricia Tree")
    }
  }
}
```

This would be unit tested as follows –

```
test fn test_update_existing_key() {
  let initial_tree = MerklePatriciaForestry.new()
  let key = "sample_key".to_byte_array()
  let value = "initial_value".to_byte_array()
  let tree_with_value = initial_tree.insert(key, value)

  let new_value = "updated_value".to_byte_array()
  let updated_tree = update(tree_with_value, key, new_value)

  assert(updated_tree.get_value(key) == Some(new_value))
}
```

8 Database Schema Design

In this DApp, will need to keep track of subscriptions usyage of the platform and will also need the ability to search and filter on Dataset income and sale data so that payments can be made to owners of Datasets and viewed and analysed in detail by Users. These payments will be grouped to a time period basis(TBD) onchain. This can be done efficiently if we store the these and user details in a SQL database with indexes for quick access. The source of truth of payments, however, will always be the blockchain, and the database is a cached breakdown of the blockchain data payment. By introducing data replication, we must ensure that the database is updated accordingly when a transaction is submitted to the blockchain. The schema for the SQL DB TBD.

For ref :-

Notes and Links made on Notion to refer to and add to in the future

The following is a Notion document in which I have listed any links I have used and my notes on Graph Theory using ChatGPT.

https://www.notion.so/12fcd0828571809d8cb8fa36c0be2b1c?v=dca0e2e2213642b98dd6cfec64da98eb&pvs=4