ocean

A Decentralized Data Exchange Protocol

Reference Marketplace Framework

Ocean Protocol Foundation Ltd

October 2017

A joint project of

BIGCHAIN DB DEX

Abstract

This technical primer presents a summary of the core marketplace attributes and components required to facilitate the successful deployment of the decentralized data exchange protocol and network called Ocean Protocol.

Modern society runs on data. Modern artificial intelligence extracts value from that data. However, the power of both data*AI is siloed. The goal of Ocean Protocol is to liberate data, and open it up to AI, thereby distributing the power of data*AI. This liberation will be driven by asset tokenization propelled by blockchain. Like streams to an ocean, Ocean Protocol is the confluence of Blockchain with AI.

Table of Contents

1. Introduction

Ocean Protocol ("Ocean") is a decentralized data exchange protocol and network that incentivizes the publishing of data for use in the training of artificial intelligence (AI) models.

The network leverages blockchain technology to facilitate the distribution and consumption of data in a safe, secure, and transparent manner. Ocean provides the mechanism for storing every asset's metadata including links to the data itself, data ownership, and associated data IP licensing information.

On top of the protocol sit data marketplaces that access and serve the underlying data assets. Each marketplace acts as the last mile in connecting data providers with consumers. Ocean incentivizes uploading of high-quality data, including data intended for use in public data commons. Control of assets within the Ocean Protocol network is provided to the respective rights holder, with first-class privacy measures baked in. It also provides programmable market mechanics, making fair, yet flexible pricing easy. Additionally, Ocean is designed for industrial-scale usage.

Look no further than the government of the United Kingdom for the rationale driving Ocean. According to the report on *Growing the Artificial Intelligence Industry in the UK*, released jointly by the Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy on October 15th, 2017, the UK "...could add an additional USD $814 billion (£630bn) to the UK economy by 2035, increasing the annual growth rate of GVA from 2.5 to 3.9%."

[However,] to continue developing and applying AI, the UK will need to increase ease of access to data in a wider range of sectors. This Review recommends: - Development of data trusts, to improve trust and ease around sharing data -

Making more research data machine readable - Supporting text and data mining as a standard and essential tool for research. [UK AI]

There is discernible motivation for adopting AI, as highlighted above. It is also apparent that impeding AI growth will have adverse effects on economies, and likely on society as well. We find ourselves at an inflection point, and it is our strong belief that Ocean Protocol provides a clear path forward.

This paper briefly introduces Ocean Protocol and the associated marketplace requirements. Please note that a more detailed explanation of Ocean Protocol will be available upon release of both the technical and business white papers.

1.1. Mission

Ocean Protocol's mission is to create more equitable outcomes for the owners and consumers of data by unlocking data through a thoughtful application of both technology and governance. Ultimately, Ocean Protocol aspires to untangle the world's data in a safe and secure manner for the benefit of all.

1.2. Key Drivers

The primary goal of the Ocean network is to create a global supply chain of data for consumption by AI's. This data will be of two principle types: "commons", or free data; and priced data. The data itself can be provided in raw form, or "cleansed" and modelled. Ocean Protocol will facilitate access to the data through marketplaces that cater to the specific needs of their consumer base.

Also, critical to the network is ensuring data provenance and security. With Ocean, data providers will be able to control who (or what) accesses their data assets, as well as how and where the data assets are being used. This virtual paper trail is immutable and inherent to the network.

2. Ocean Protocol Overview

The following is a summary of the key technical attributes required for a fully functioning marketplace ecosystem within the Ocean data exchange protocol.

Further details of Ocean's underlying technology stack will be covered in the full technical white paper release.

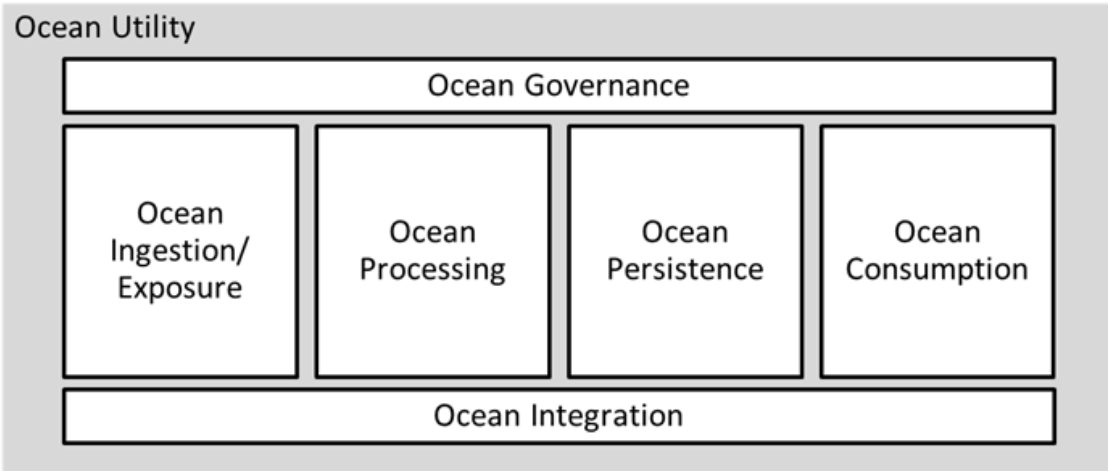
2.1. Core Ocean Capabilities

Ocean Protocol will support the following core capabilities:

1. **Data Exposure/Ingestion** - Data assets are exposed by data providers. These assets could be raw data with little to no modelling, or fully transformed data models, similar to what would be found in enterprise data warehouses. These assets can reside within the network, as is the case with free public data, or outside the network behind firewalls.
2. **Data Processing** - Data processing provides the compute mechanisms required to cleanse, transform, and analyze exposed data. Ocean Protocol's processing functionality will provide data curators with the ability to combine and normalize exposed data in order to create new assets. This capability will also provide the means for deploying AI algorithms. Processing can be provided on-premises behind firewalls when required, by data marketplaces or by registered data processors within the network.
3. **Data Persistence** - Data persistence provides the mechanisms for storing post-processing result sets. These mechanisms could provide simple blob or file storage, similar to HDFS, MOLAP or ROLAP data stores for analytic consumption, in-memory persistence for low-latency data

access, tuple or document stores for scalable operational data storage, highly indexed data storage optimized for search, IPFS, Storj, Swarm, etc.

4. **Data Consumption** - Data consumption provides the means for end-users/consumers to leverage the underlying data assets. This mechanism will generally be provided by marketplaces providing an interface to the data, whether B2C, B2B, M2M, etc.
5. **Data Integration** - The integration mechanism provides secure, end-to-end access to the network's data while enforcing authorization and entitlement protocols. Ocean Protocol's integration capabilities will manifest as API's and microservices deployed and maintained by registered integration providers.
6. **Data Governance** - Data governance is a first-class citizen of Ocean Protocol. In fact, the core mechanisms are baked in, from data provenance established through the immutable recording of all transactions within the network via blockchain, to the quantifiable establishment of golden sources via tokenization metrics, to the creation of data dictionaries via curated MDM registries (i.e. the entity that stakes the most defines the Master Data Management standards).
7. **Utility** - Ocean Protocol itself is a utility as it provides the basic infrastructure substrates for a public service. These substrates include (but are not be limited to) the tokenization mechanisms, or the means for transacting within the network, as well as the marketplace protocols that orchestrate Ocean's capabilities. All the capabilities listed above are accessible through the transmission of tokens within the utility network. [Utility]



2.2. Ocean Tokenization

Tokenization is the means of transacting within the ecosystem, and because the tokens can be exchanged to procure network services, they are treated as utility tokens. [Utility] Additionally, Ocean's blockchain technology removes the possibility of infinite reproducibility in digital asset like data and algorithms. As such, it can be confirmed that each unit of value was transferred only once, solving the long-standing problem of double spending. [Double Spend]

In fact, Ocean allows any service or resource within the network, from data to storage and compute to API access to governance mechanisms, to be tokenized. Network users must acquire tokens to leverage the network resources/services on offer. Tokens can be acquired through purchase (via exchanges), or by offering a value-added service within the ecosystem that nets the provider tokens.

Take, for example, a data provider. As a network asset contributor, a data provider can make their data available in exchange for tokens. To access this data asset, a data consumer only needs to provide the provider with the requisite number of tokens. The handshake between counterparties is a simple mechanism handled by Ocean's underpinning blockchain substrate.

But Ocean goes even further than simple Peer-to-Peer data asset transfer. Any service within Ocean can be tokenized. An example of this would be compute. Ocean can provide secure transient compute capabilities through trusted compute providers that facilitate the offloading of computational stress from the data provider. In this way, an AI developer could request a data asset from a provider, facilitating the transaction through a token transfer. The data provider could then push the dataset to a trusted compute provider, facilitating the processing with a percentage of the tokens received for the asset from the AI developer. The data provider would then provide the AI developer with the location of the compute cluster to push the algorithm for processing alongside the asset. Once processing is complete, the trusted compute provider would transfer the resultset to an agreed upon persistence location for use by the AI developer, and drop both dataset and algorithm from its memory banks. (Please note, this is a simplified process, as does not take into account

capabilities like Homomorphic Encryption that could potentially remove the need for dropping assets.)

With tokenization, Ocean offers a common mechanism of exchange to reduce the friction generally associated with asset sharing.

2.3. Key Network Contributors

There are five key network contributors: Data Providers, Data Consumers, Data Referrers, Network Keepers, and Regulators. Each contributor plays a unique and critical role in the operation of Ocean described below:

Data Providers

Data Providers are the core actors to the Ocean Protocol and network. Without Data Providers, Ocean Protocol does not exist as they provide the network data assets in exchange for tokens. The assets provided may be raw data files, blobs, structured, semi-structured, unstructured, etc. The data may be heavily modelled and available for usage as MOLAP or ROLAP data, or completely un-modelled and available via distributed file stores like HDFS or IPFS. Data Providers can be broken down into the following subsets:

- **Data Owner** - Data Owners are the original proprietor and purveyor of the data asset. They legally own the data IP and can facilitate usage of their data assets when compliant with regulations
- **Data Custodian** - Data Custodians holds data on behalf of their consumers, as well as maintain the value of the data assets in compliance with regulations. They do this by validating assets against benchmarks for usability, accuracy, and relevance. They are also responsible for creating and maintaining the metadata mappings for any data asset.

Data Consumers

Data Consumers are the primary users and beneficiaries of Ocean's data assets. Ocean consumption is open to all, and will be made up of individuals,

start-ups, small to medium sized companies, and large-scale multinational enterprises and governments. As stated previously, it is Ocean Protocol's goal to open up access to an extensive array of varied data for use by AI. As such, Ocean is perfectly suited to meet the needs of AI specialists, Data Scientists, Big Data Engineers, and Business Intelligence professionals.

Data Mashers

Data Mashers sit at the cross section of Data Providers and Data Consumers. Mashers provide a value-added service to the network by performing data cleansing, transformation, and normalization across multiple sets of data, effectively "mashing" data together. The resultsets of the data mash-up function will be treated as unique data assets for use within Ocean.

Data Referrers

Data Referrers will promote the use of Ocean to Data Providers, and facilitate the linking of Data Consumers to data assets. Consequently, the responsibility of identifying valuable data assets and their corresponding purveyors will be that of Ocean's Data Referrers. This key role will manifest itself through the development of marketplaces, from which data assets will be procured from providers, and exposed to consumers.

Network Keepers

Network Keepers provide and manage the orchestration of Ocean's critical substrate functionality. Keepers run as nodes within the network and provide one or more of the functionalities listed below. These nodes earn mining tokens for exposing the functional components to network users. They are also penalized in the event that service fails to meet established network governance thresholds. Core Keeper functionality includes, but may not be limited to, the following functionality:

- Replication and Consensus Transaction Validation for Unspent Tokens
    - Smart assets
    - Custom validation
- Market Mechanics

    - Matchmaking via API's and microservices
- Marketplace Protocols
    - Compute
    - Storage/Persistence
    - Gateways
        - Permissioning
        - Resolution
        - Integrity
        - Certification
    - Pricing
    - Registry Curation

It is envisioned that Data Providers and Referrers will make up a large proportion of Network Keepers. However, these services could also be provided by third party contributors with expertise in specific functional areas, like Data Integration or Data and Platform Audit.

Regulators

While this may be contrary to popular opinion, regulators are critical to provide guidance for the protocol and network. Ocean's use of blockchain does not absolve contributors of their requirement to protect data assets to the utmost. Inclusion of all vested parties is critical to Ocean's success. In fact, working with regulators and auditors to satisfy compliance will reduce overall friction within the network, as it will remove contributory reluctance, as well as impediments to consumption. The added benefit is that these safeguards can be intrinsically tokenized within the protocol, adding even more impetus to play by the rules.

Most contributors will access Ocean Protocol via data marketplaces that will be built on top of the Ocean Protocol and others like regulators and keepers will interface with the Ocean Protocol Foundation directly. Any new services in the ecosystem will have access points on the Ocean Protocol and the data marketplaces.

2.4. Data Governance

Data Governance is critical to the successful operation of any data platform. As such, Data Governance is provided first-class citizenship within the Ocean Protocol network. The function manifests itself through the immutable nature of transactions on the blockchain. Any transaction that ever occurs within Ocean Protocol is recorded. To ensure compliance by network contributors, Ocean incentives transactional recording in two ways:

1. By providing actors with an indelible record of request-response communications for adjudication purposes; and,
2. By taking advantage of Ocean's network effects and offering opportunities like incremental discounting for multiple sequential spends.

Please note: These token functions will be elaborated on further in the technical whitepaper.

Creating a virtual breadcrumb trail of all transactions through the entire network stack makes establishing provenance and auditability relatively easy. With this in mind, Data Provenance has been a core focus of BigchainDB ("BDB"), a scalable blockchain database provider, since its inception, and remains so after solution adoption by 40+ corporates. By providing connectors/API plug-ins to BDB nodes for all integration points, Ocean Protocol will be able to track data usage throughout the network. In the case of regulatory compliance, this capability is especially beneficial as compliance issues often subside so long as verifiable audit and provenance can be established.

Furthering Ocean Protocol's Data Governance capabilities, trusted curated registries work in tandem with Staking (more on this later) to facilitate the deployment of best-of-breed governance policies and standards. These registries will provide an adoption mechanism for standardized Master Data Management (MDM) policies, and associated Data Dictionaries, potentially across entire domains (or even across domains). Applying these policies could be as simple as subscribing to the top registry entry, and enforcing the associated policies and framework to an existing data asset.

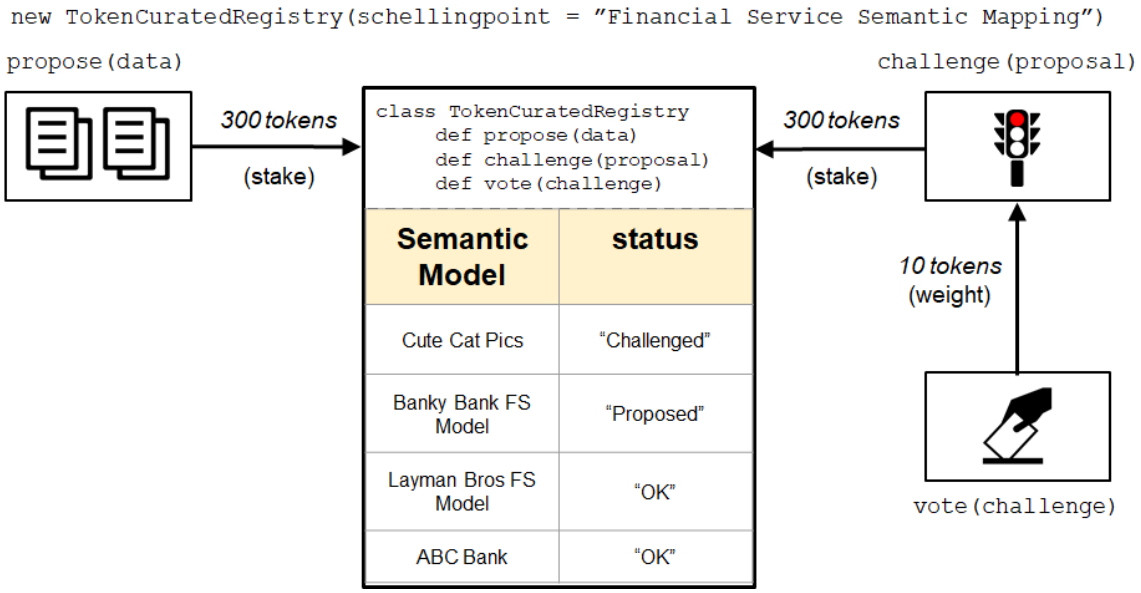
2.5. Curated Registries

Ocean Protocol is fundamentally a Delegated Proof of Stake (DPOS) network. This means that in order to transact, and actor must stake tokens. This crypto-economic mechanism gives rise to the concept of Curated Registries for any type of asset or service within the network. For example, Semantic abstractions of underlying complex data models have been valuable implementation instruments for decades. However, this paradigm has failed to garner widespread adoption because of two fundamental issues:

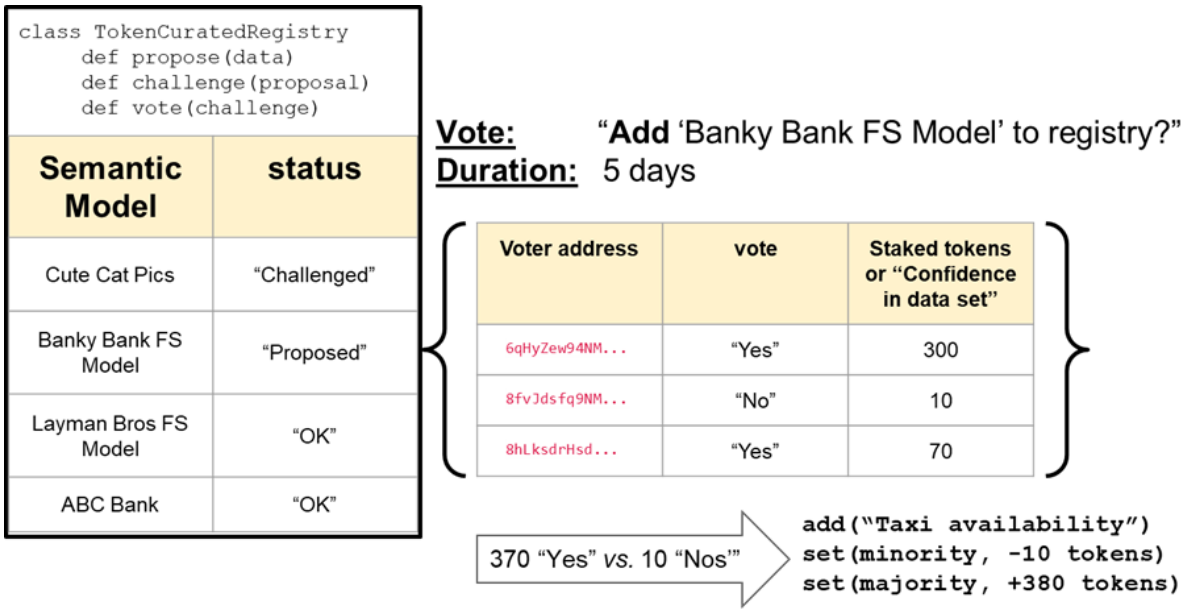
1. Lack of Semantic Layer standardization leads to competing Semantic models; and,
2. Poorly formed Semantic models create implementation issues.

Staking resolves these two problems. In the case of the first, were there a trusted curated semantic registry, then semantic model developers could increase stake to increase their position within the registry. The higher in the registry, the more likely your model is to be adopted, and thus become the semantic standard. When multiple models offering significantly similar implementations exist, there is rationale for merging the models and combining each independent parties' stake to raise the consortium's position within the registry. Lastly, in the case of a poorly formed semantic models within a registry, a challenge to that model could be invoked. If the model is poor in relation to other models, then it is in the best interests of the registry actors to remove the model in question. As such, the challenge would be upheld, and the challenged model would be removed.

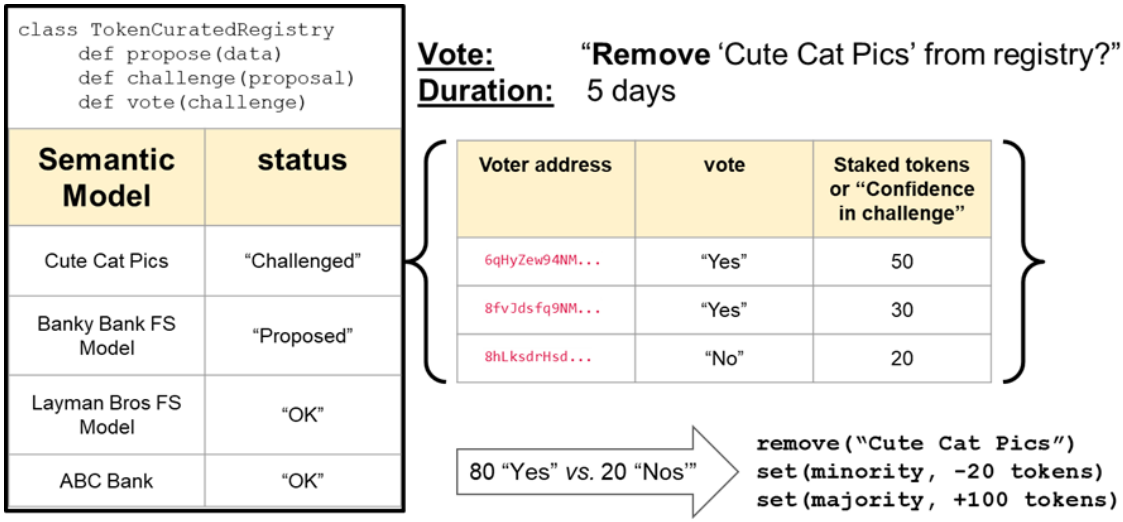
A generalized registry on and off-boarding process is described below:



Activity 1: Proposing and Challenging a Registered Semantic Model



Activity 2: Proposing a New Semantic Model Entry



Activity 3: Challenging an Existing Semantic Model Entry

Please note that these registries could exist for anything from data and models to fully formed applications. Please also note that the specifics of DPOS will be discussed in greater detail in the Ocean technical white paper.

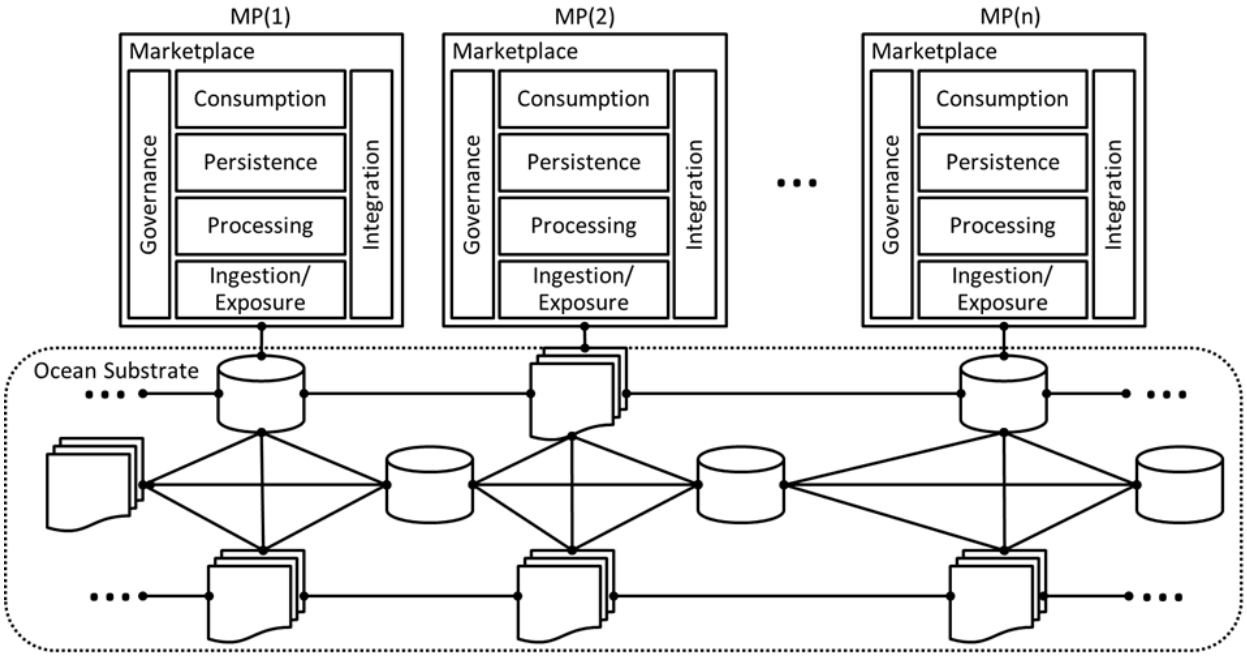
2.6. Ocean Marketplace Deployment Strategy

Ocean's marketplaces will act as Grand Bazaars for data, enticing consumers with their alluring data assets, while at the same time attracting data providers to the network because of access to a broad consumer base. Each marketplace can cater to specific domains by providing data relevant to that domain only, or appeal to broader consumer base by providing cross-domain data assets and mash-ups. Initial marketplace development will lean heavily on the experience of DEX. DEX is ingrained within the data marketplace community, helping to drive the paradigm since inception, and will bring a wealth of knowledge and best practice to developing the first marketplaces on top of Ocean protocol together with industry and government.

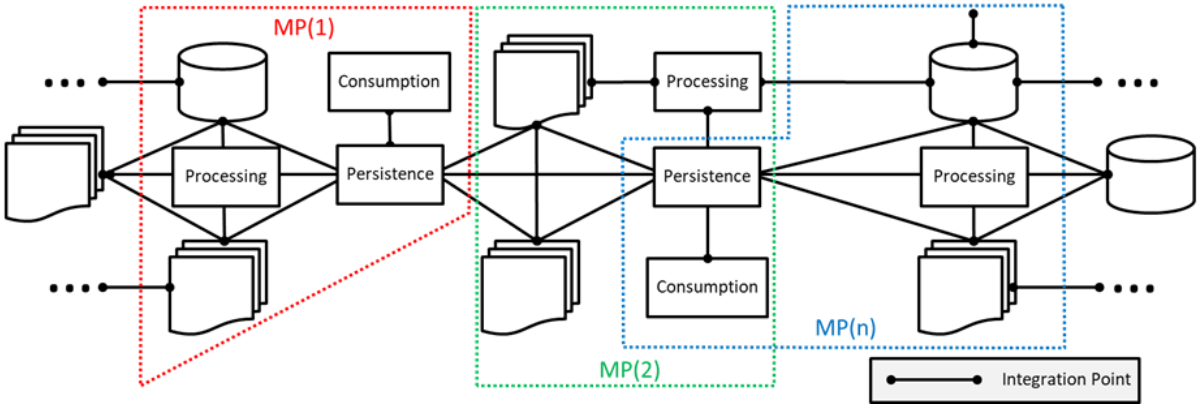
Initially, the inclination will be for marketplaces to manifest as holistic, end-to-end solutions that include the means to ingest, process, persist, consume, and govern data. Gradually, however, network effects will take over, and more suitable providers with distinct capabilities and expertise will emerge. This means that over time, providers of specific capabilities (e.g. in-memory persistence) will materialize to compete with generic marketplace offerings. As a result, marketplaces will naturally begin to dismantle in favor of a modular

approach. By doing so, best-in-class marketplaces will be in constant flux as they select capabilities from the evolving best-in-class providers within the network.

This ethos may lead to a modularized, evolutionary design approach for marketplace solutions. Consequently, it could be in the marketplaces' best interests to link together the best end-to-end capabilities available from the network at that moment, in order to attract both providers and consumers to their offering. Eventually, the primary function of marketplace providers could be to act as referrers of both data and component capabilities.



Marketplace Deployment Evolution - Anticipated Initial Structure



Marketplace Deployment Evolution - Future State

2.7. Types of Data

Ocean will expose three primary types of data:

1. **Proprietary Data** - This is data that is controlled by a data provider/owner, and is generally unique to that provider/owner.
    - Example: Proprietary autonomous vehicle data.
2. **Regulated Data** - This is data that is controlled jurisdictionally through regulation or other means. While the data may not be unique, its accessibility is limited generally due to privacy constraints.
    - Example: Personal medical history data.
3. **Free or "Commons" Data** - This is data that is generally free or open for use. This type of data generally has limited restriction on its usage.
    - Example: National Census data.

2.8. Pricing

It is envisioned that, at least initially, Ocean will enlist three pricing schemes for data, depending on the data type and its fungibility or uniqueness:

1. **Free Data** - This data is open to all consumers with no restrictions. We want to encourage a growing data commons for the world. The token design elaborates on the incentive structure.
2. **Proprietary/Regulated Non-Free Fungible Data** - With data that is common but controlled, the pricing is easy(ish): just use an exchange. Exchanges are low friction and let the market determine the price. We plan to support data exchange functionality in the Ocean Protocol.
3. **Proprietary/Regulated Non-Free, Non-Fungible Data** - For data that is unique, pricing becomes more difficult. The price could simply be fixed. However, if priced too low, it's a lost revenue opportunity. And if priced too high, no one will buy it. To address these concerns, we explored several pricing schemes and distilled them into three options: fixed price, auction, and royalties. Each has unique mechanics and requirements from a crypto standpoint and is explained in the table below.

3. Engagement Model

This following describes how Ocean will engage with both Data Providers and Data Consumers.

3.1. Data Providers

For any data network to succeed, the right players must be activated at the right time. This is no different for Ocean. First and foremost, this initially means rigorous engagement with Data Providers in order to plum the network. Without data, there is no Ocean. Thankfully, through DEX and BigchainDB, Ocean currently has 50+ data providers lined up for the network's initial data seeding phase.

The onboarding of data providers will be relatively straightforward. Initially, all data will reside in-situ, and be exposed to the network protocol via light-touch API's. To expose their assets, data providers will navigate to their marketplace portal of choice, and select the option to provide data. Next, they will register with the network, providing information about the data owner. Once this is complete, a daemon script will be pushed to be deployed to the provider's data repository, upon which access will be granted to the network via an administration portal. This portal will allow the data provider to designate which assets should be exposed to the network, plus any consumption parameters. All of the gathered information will then be deployed to keeper nodes for provenance, and the assets will be exposed to the consumer base.

4.2. Data Consumers

Without the demand-side of the network, Ocean is unviable. Thus, proper engagement of Data Consumers is critical. However, the timing of this activity is also essential. Too early, and consumers won't see the value of the network. Too late, and providers won't realize the return on their contribution. This *Goldilocks Dilemma* is exacerbated by Ocean's core target base of AI

researchers and startups because AI's need massive amounts of data. Auspiciously, we seem to have hit an inflection point, as AI adoption, along with the understanding of AI's inherent need for data, becomes more prevalent. [UK AI] This understanding will assist in placating data providers who traditionally look for immediate return on investment. In Ocean's case, the potential upside for further AI advancement are too great to leave the ecosystem.

Onboarding data consumers will be relatively simple. Like data providers, consumers will engage with their marketplace of choice, uploading information about user(s). Once these users are validated - a process that could require rigorous KYC for certain marketplaces like those catering to Financial Services or sensitive healthcare data - they will be able to start consuming data from providers. Depending on the marketplace and its associated providers, the consumption mechanisms could include embedded dashboards and mash-up windows, to full access to data assets via download.

For AI's, the process will include additional steps, such as providing access to distributed sandbox environments, or the potential to encrypt AI algorithms and push them to homomorphically encrypted, containerized data assets to run either in-situ or via trusted registered data processors (e.g. through a potential combination of OpenMined and Amethix). This type of transaction will be explained further in the technical whitepaper.

3.3. Delivery Strategy - <Hello World>

All solutions have an inception point. The foundations of the Ocean Protocol will be laid by DEX and BigchainDB in partnership with a consortium of industry and government contributors, centered in Singapore. This Genesis Program will run for 18 months divided into six unique project sprints, each in a regulated industry vertical that bring together all vested interests in compliance with local and national laws.

4. Conclusion

This technical brief presented Ocean Protocol - a protocol and network that incentivizes the proliferation of vast supplies of data for use in training artificial intelligence (AI) models.

5. Acknowledgements

The lead authors (Trent McConaghy, Dimi de Jonghe, Tim Daubenschütz, Chirdeep Singh Chhabra, and Don Gossen) would like to thank everyone who gave feedback, comments or other contributions to this paper, particularly Bruce Pon and Adam Drake, as well as the rest of the BigchainDB & DEX teams.

6. References

[UK AI] https://www.uk.gov

[Filecoin] https://filecoin.io/

[IPDB] https://ipdb.foundation/

[IPFS] https://ipfs.io/

[Storj] https://storj.io/

[Swarm] http://swarm-gateways.net/bzz:/theswarm.eth/

[Utility] https://news.21.co/thoughts-on-tokens-436109aabcbe

[Double Spend] https://en.wikipedia.org/wiki/Double-spending

ocean

Ocean Protocol Foundation. A Non-Profit Foundation
www.oceanprotocol.com