Oz: A Platform for Edge Data Collection

@burbabull, @heathenft

October 3, 2024

Abstract

The need for improved access to specialized data has never been greater than it is today. As organizations build software, deploy AI agents, and train ever-larger generative models, this need only continues to grow. The internet is the largest repository of publicly accessible information, but the novelty in this dataset has been exhausted. Organizations now need access to new sources of information. This whitepaper presents a mechanism to gather, distribute, and incentivize the collection of novel and specialized datasets required for the continued growth of the data-dependent economy.

The protocol provides a framework and infrastructure for collecting specialized data from any number of edge nodes. To generate data sets at scale, Oz uses Cascade, a modular architecture for incentivizing the collection of structured data from distributed network participants. By enabling the generation of specialized data sets, Oz enables organizations to completely replace expensive enterprise software with bespoke tooling without compromising the quality and integration of the complete solution.

1 Introduction

The production of commercially accessible datasets has come to rely almost entirely on a system of proprietary edge data collectors embedded in everyday products. This system is effective at creating large volumes of structured data for use in today's monolithic software organizations. But the landscape for the creation and use of software and data has changed substantially in the last two years due to advancements in generative AI and their effects on the global software industry.

Oz was founded based on seven theses about the role of data in the development of software:

1. **Software development has been commoditized**
   Two decades ago, developing software was considered a specialized skill. However, advancements in developer tooling have greatly simplified the process of designing and delivering new software tools. In recent years, generative AI has almost entirely eliminated the barrier to entry to developing new software. This specialization led to the growth of generic enterprise software products. These tools sufficiently but imperfectly matched the processes of many organizations, leading to widespread adoption.
2. **Simplified software development will lead to more specialized software tools**
   As software becomes easier to build, more software will be created. In the past, there was pressure on enterprise software organizations to add hundreds of small features for each client. In the future, software will be built to address the highly specific needs of each individual business.

The result will be uncompromised software tools suited to fewer and fewer companies, but matching their use cases precisely. [1]

3. **Access to high-quality data will inhibit the use of home-grown software tools**
   Enterprise software uses huge quantities of expensive and specialized data to deliver its core value, often paying in excess of $100,000 per year for individual data sources. This high upfront cost is mitigated by the large number of customers served by each enterprise software tool. For small organizations to effectively build software that exceeds the quality of off-the-shelf enterprise tools, they will need access to equivalent data at a much lower cost.
4. **The market for specialized data will exceed $50B annually**
   As bespoke software and agents become more common, the demand for data will increase. In total, the market for specialized data today exceeds $54B per year, led by organizations such as Mandiant, Epic, and Bloomberg [2]. As bespoke software becomes more common, the need for inexpensive specialized datasets will expand dramatically to satisfy the demand for analytics, reporting, and compliance requirements.
5. **Acquiring new data is more important than distributing existing data.**
   We've reached a point where every major company has scraped the majority of publicly available digital information. The internet served as a readily available trove of accessible information, carefully curated for years prior to its use in AI and analytics. Existing troves of data have been exhausted, and we've hit the "data wall" [3]. We now need more networks for originating new data instead of distributing, storing, and sharing what we already have. Much like software, tools for disseminating existing data will become commodified, whereas networks for originating novel data will retain their value long-term.
6. **Unlocking access to private data will enable continued advancements in AI**
   The proliferation of IoT devices in the early 2010s brought internet connectivity and sensor capabilities to billions of devices across the globe [4] [5]. The audio, video, and sensor data produced by these private devices can enable extended use cases of generative AI that are not possible with public internet data alone. Extensive integration infrastructure is required to collect, structure, and concentrate this data.
7. **The future of data collection is based on DePINs**
   Existing data brokers are structurally configured to sell data to large enterprises. By collecting data top-down with centralized edge and IoT networks, their products will remain too expensive, too complex, and too difficult to purchase for the emerging class of data customers. DePINs for data collection have the potential to exactly model the scale and price points required to meet the rising demand for data for the next generation of bespoke software and AI agents.

Oz protocol was designed to address these seven founding propositions. The protocol implements a mechanism for the creation and incentivization of modular data collection DePINs. Each data DePIN module can support any number of data providers and data consumers, enabling markets for data that are secure, decentralized, and fair to all participants.

The rest of this document outlines the technical architecture and applications of Oz protocol.

2 The Open Data Access Network

There are six key challenges to generating novel data sets:

Challenges

1. Integration and deployment: Connecting to existing data repositories and maintaining those connections over time as the underlying system is updated.
2. Incentivization. Rewarding valuable contributions and punishing disruptive or worthless contributions.
3. Flexibility. Accommodating any type of data (embeddings, text, video, audio, etc).
4. Privacy-preserving. Ensuring data is accessible only by its intended recipient and originator.
5. Scalability. Handling data producers and consumers of any scale, ranging from a single module of a single application to an enterprise data broker.
6. Censorship resistance. Ensuring no participant in the network can interfere with the transmission of data between its originator and recipient.

To address these issues, Oz has developed Cascade, a novel architecture for the development of modular data DePINs. Cascade is a protocol that enables the deployment of edge data collection devices by any network participant. Once deployed, these devices generate a secure stream of data from any user to any subscriber. To date, Oz has 5 active edge data collectors deployed by over 5,000 network participants, generating millions of data points per day.

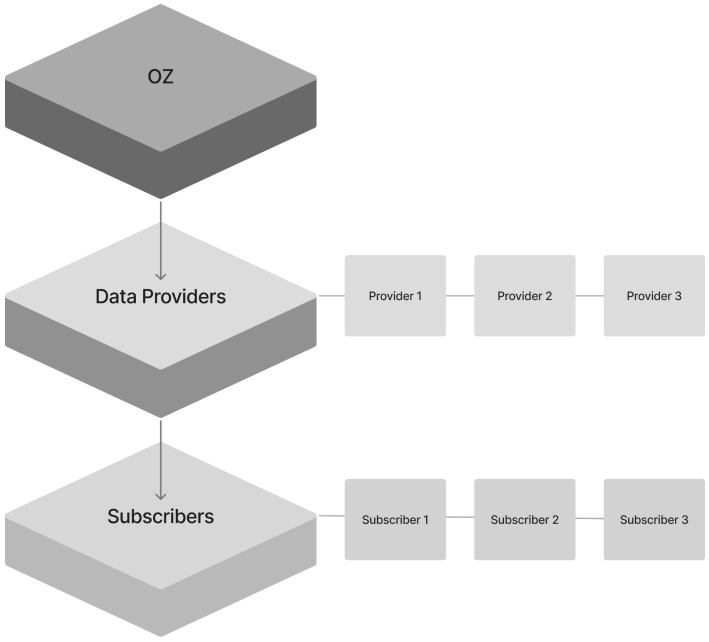

Figure 1: High-level architecture of Cascade, a mechanism for modular DePIN development.

2.1 Edge Data Collection

Sense Agents

Data enters the protocol via modular edge data collection devices called **Sense Agents**. Data providers can use Sense Agents to stream data directly to subscribers from the edge over secure channels. Each Sense Agent defines its own standalone DePIN with the following components:

1. **Single schema**. Each Sense Agent generates data in a pre-defined format for programmatic consumption by subscribers.
2. **Sensor suite**. Sense Agent must have an onboard sensor to convert either real-world or digital data into transmissible metadata.
3. **Decentralized subscription ledger**. Sense Agents must have a record of who is entitled to their data feeds. This information is tracked via a decentralized ledger.

Sense Agents can access sensor data from:

- Microphones
- Video equipment
- Temperature sensors
- Pressure sensors
- Physical presence sensors
- Actuators
- Existing software tools
- Networking equipment
- RF detectors (WiFi, cellular, Bluetooth)

Cascade provides standard interfaces that enable developers to connect existing sensor APIs with an incentive and distribution layer.

Sense Agents bring AI companies past the “data wall” by making data available that was previously either not digitized or kept behind private user-only APIs. The most readily accessible data are those generated by already deployed private IoT devices. By integrating with this vast existing sensor infrastructure, Sense Agents will dramatically expand access to an untapped repository of live data.

2.2 Data Distribution

Concord

Sense Agents share data with network participants based on records stored on a distributed ledger known as **Concord.** The ledger takes the following roles on the protocol:

1. Directing payloads from Sense Agents. Instructing Sense Agents to share data with the appropriate data buyer.
2. Distributing incentives. Transferring payment from the data buyer to the data provider.
3. Maintaining data provider reputation. Storing data provider reputation scores on the public ledger and making them available to future data buyers.

To ensure censorship resistance, Concord is deployed on a decentralized infrastructure as a set of smart contracts.

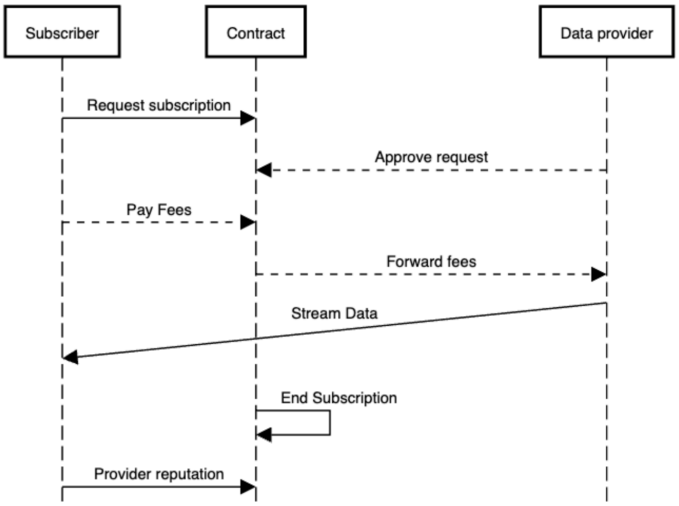

Figure 2: Lifecycle of a data subscription on Concord

2.2.1 Data Privacy

On Cascade, data providers send their payloads directly to subscribers via secure protocols without the involvement of a centralized intermediary. The ledger instructing providers with whom to share data is fully decentralized. This means data streams are completely private and censorship-resistant.

When a Sense Agent is created, a corresponding record on Concord is also created maintaining the publisher-subscriber relationship between data providers of a given Sense Agent type and the appropriate subscribers.

2.2.2 Data Security

Authorization is key to ensuring data feeds remain secure and accessible only to the intended audiences. Concord gives full control of data feeds to its originators, ensuring data is only sent directly from its originator to the intended recipient.

To prevent unauthorized access to data feeds, each data provider may choose to approve, reject, or terminate subscriptions to their data feeds.

2.3 Data Storage

Vana

User data is not stored directly within the Cascade architecture. Instead, Oz provides the tools to originate novel datasets, transmit them to buyers, and maintain data provider reputation.

To provide long-term data storage and durable incentivization, data providers may choose to subscribe their own data feed to the native Vana integration Sense Agent [6]. Once subscribed, each data provider's stream from a given Sense Agent is routed to the Oz tokenized Data Liquidity Pool (DLP) on Vana.

2.3 Fee Model

Concord incentivizes data providers to share novel data sets by providing fees in exchange for access to data feeds for a period of time. Under the subscription model, subscribers receive data that is live and directly provided by the source.

Subscription fees have the following components:

1. Sense Agent developer fee. Allocated to the original developer of the Sense Agent used to gather the data, and defined at the point the point of original deployment for each Sense Agent.
2. Protocol fee. Allocated to the Oz development team, set at zero at the time of writing.
3. Data provider fee. Allocated to the deployer of the Sense Agent data collection device.

Subscriptions are granted for a fixed number of days, as defined at the time of subscription.

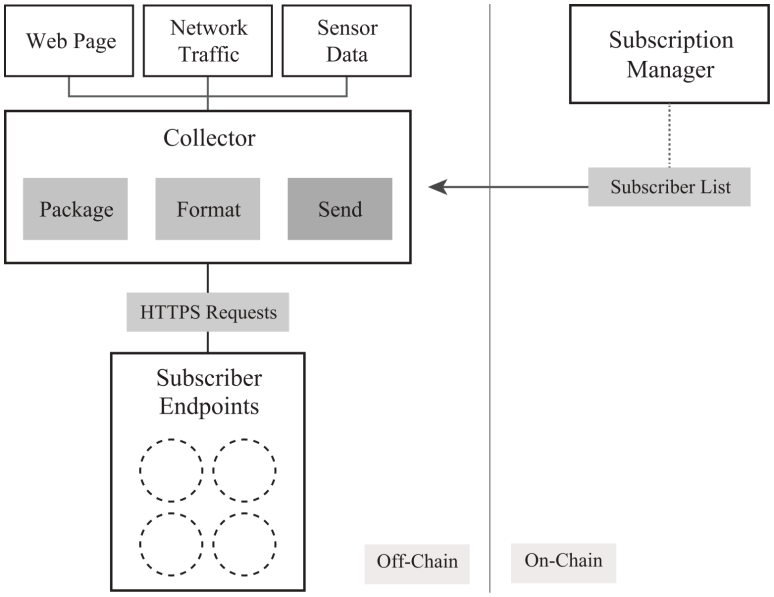

Figure 3: On- and off-chain components of Cascade's data subscription model

3 Applications of the Open Data Access Network

Today's data-dependent organizations purchase datasets from brokers that source raw data from proprietary sensor networks. The data that results from this process is consistent and high-volume, but extremely expensive.

The architecture of Oz protocol was designed to meet the business requirements of existing data-dependent organizations, but with a few additional benefits:

1. Elasticity. As the target audience for data expands, the scale of the datasets needed also expands. The cost of many datasets today exceeds $100,000 per year. Oz allows subscribers to consume a dynamic number of data feeds, from 1 to millions, allowing their data consumption to grow with their business.
2. Novelty. Existing data brokers source their data primarily from closed-source sensor networks. These include network devices, smartphones, and cellular networks. By incentivizing individuals across the globe to provide access to their local sensor network, Oz will grant access to data that was previously inaccessible to organizations.

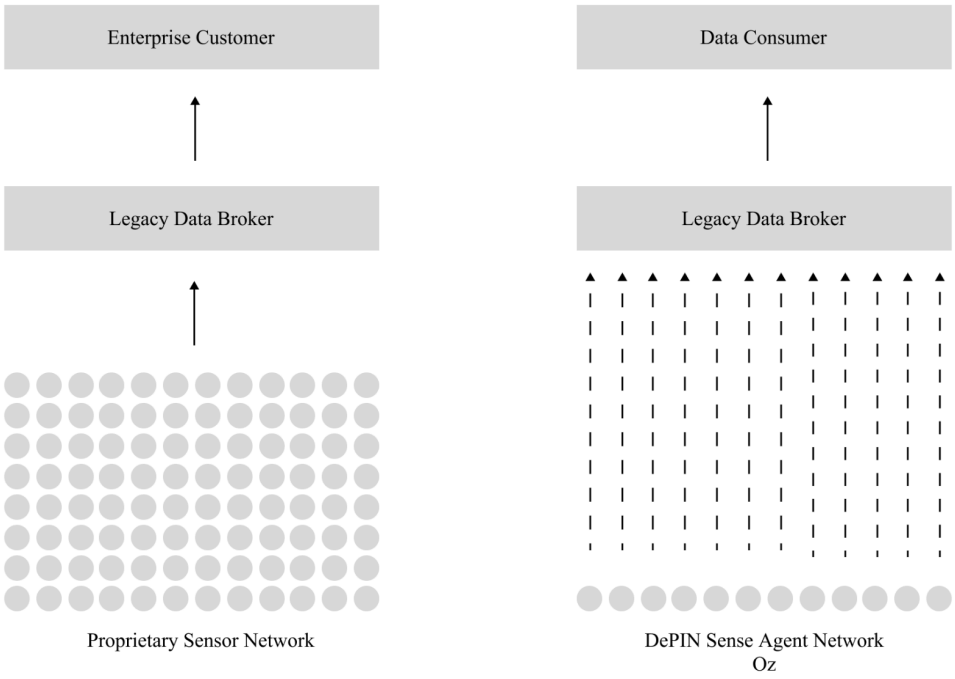

Figure 4: Legacy edge data collection architecture vs. modular Oz modular Sense Agent network

3.1 Data Walled Gardens

The current incentives for data collection place the interests of the public and the interests of the data provider at odds. The prevailing incentive structure encourages data providers to amass vast amounts of information and restrict access within proprietary systems in order to maximize profit.

In many critical sectors, such as healthcare, cybersecurity, public safety, and infrastructure, such restricted access can directly harm the public. Open data sharing has the potential to significantly improve outcomes in these areas. By effectively eradicating the middleman and encouraging the direct sharing of data between the origin and the end-user, Oz effectively eradicates this incentive structure and leads to more positive outcomes across key areas of society.

4 Conclusion

This whitepaper proposes a data collection mechanism that incentivizes the open sharing of specialized and novel datasets. The protocol is private, secure, and fair to all network participants.

The protocol is designed to attract developers with a balanced fee structure, and to incentivize data providers with simple deployment, transparency, and full control over their private data.

References

1. Far Reach, Inc. (2024). Bespoke software: Pros & cons. Far Reach. https://www.farreachinc.com/blog/bespoke-software-pros-cons/#:~:text=The%20bespoke%20software%20market%20was,21.5%25%20between%202023%20and%202032.
2. MarketsandMarkets. (2024). Threat intelligence market by solution, deployment mode, organization size, application, and region - Global forecast to 2026. MarketsandMarkets. https://www.marketsandmarkets.com/Market-Reports/threat-intelligence-security-market-150715995.html#:~:text=The%20Threat%20Intelligence%20Market%20is,)%20of%206.5%25%20by%202026.
3. Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv. https://doi.org/10.48550/arXiv.2211.04325
4. Maiti, P., Shukla, J., Sahoo, B., & Turuk, A. K. (2017). Efficient data collection for IoT services in edge computing environment. 2017 International Conference on Information Technology (ICIT), 101–106. https://doi.org/10.1109/ICIT.2017.40
5. Statista. (2024). Number of IoT connections worldwide 2022-2033. Statista. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/#:~:text=Number%20of%20IoT%20connections%20worldwide%202022%2D2033&text=The%20number%20of%20Internet%20of,over%20the%20next%20ten%20years.
6. Vana. (2024). DLP Spotlight: SYD - A new era in threat intelligence. Vana. https://www.vana.org/posts/dlp-spotlight-syd---a-new-era-in-threat-intelligence