EQUINIX

White Paper
AI Data Marketplace

DATA FUELS DIGITAL AGENCIES

THE AI DATA MARKETPLACE SOLUTION AT EQUINIX®

Enabling trusted, effective data sharing in a secure environment to support AI innovation and meet mission needs

Contents

One Platform. One Solution.

Simplify Data Exchange and Monetization

Easy to share. Easy to consume. Easy to sell. Easy to buy. All in one platform.

Ensure Data Traceability and Integrity

Fully secured. AI models and data lineage tracking.

Bring Algorithms to the Data*

Run analytics wherever data resides. No need to move data.

Enjoy Proven Trust and Neutrality

Turnkey data center and technology capabilities in ONE solution. No need to integrate point solutions.

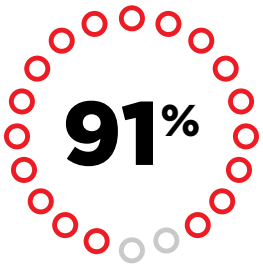
Global Distributed Solution

Support for data sharing in different regions/markets for data residency/compliance.

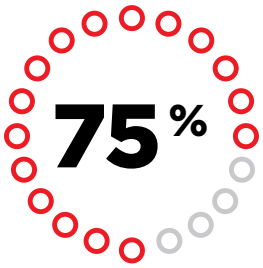
* Depending upon the trust model, bring data to the algorithm or algorithm to the data.

Introduction

Government agencies increasingly depend on data to support and automate decisions and actions. AI-based approaches—where data is needed to learn how to reason—are being leveraged to predict trends, optimize processes, speed up decision-making, enhance accuracy, improve pattern recognition and much more. The amount and diversity of data accessible to AI-based algorithms determine their functionality and accuracy. The demand for data from within and outside of government agencies is increasing exponentially.



91% of geospatial intelligence (GEOINT) stakeholders believe AI has the potential to greatly improve GEOINT effectiveness—with the largest impacts on national security, emergency response and urban planning.[1]



75% of enterprise applications use 10 external data sources, on average.[2]

Government agencies want to use and share their data and algorithms effectively, but they may be hindered in doing so because they fear losing control of it. As well, in many large agencies data remains siloed within groups who are reluctant to share it with each other. However, if agencies are unable to address these data and algorithm sharing challenges, they face the danger of an AI winter—starving algorithms with too little data for accurate and meaningful results.

Data marketplaces—also known as digital data marketplaces, or DDMs—allow data providers and data consumers to share, buy or sell data and algorithms privately and securely (and without violating any government regulations such as GDPR) using a programmable, community-owned, safe and secure infrastructure that organizes trust. Governance models regulate how the members of the marketplace interact. The marketplace facilitates legal contracts and asset registration that enable third-party services (data anonymization, conflict arbitration, analytics tools, etc.) and payment management. Sharing data assets via a secure, trusted and neutral data marketplace driven by a consortium established on the basis of a common benefit is a promising opportunity for government agencies meeting their missions in the 21st century.

The ability to merge physical models with digital content and conduct deep analysis at extraordinary speeds presents unprecedented opportunity to transform the productivity, capacity, and capability of the geospatial intelligence workforce.

MeriTalk, in collaboration with USGIF[1]

Data Marketplace Overview

The high-level architecture depicted in Fig. 1 shows the data marketplace as an entity owned and operated by a membership organization through which data providers and data consumers can interact and transact based on community rules and individual agreements.

A data marketplace must have a process for creating membership rules, and its process for admission must require a prospective member to agree to comply with all of them. Members create agreements regarding how they want to collaborate and execute transactions for data science workflows, enabling algorithms to train on one or multiple data sources. Agreements between parties are digitized as smart contracts that orchestrate and authorize all necessary steps needed to access and use the data. The contract also controls whether the results—the trained model—of the data science workflow can be moved out of the shared infrastructure.

The basic role of the data marketplace is to organize and facilitate interactions between data suppliers and algorithm developers to explore, select and agree to create, execute and complete data science transactions. A data marketplace is a global structure enabling sovereign organizations, which require absolute control of their data assets, to offer and under strict conditions make assets available to achieve mutual benefits that no single organization could achieve on its own.

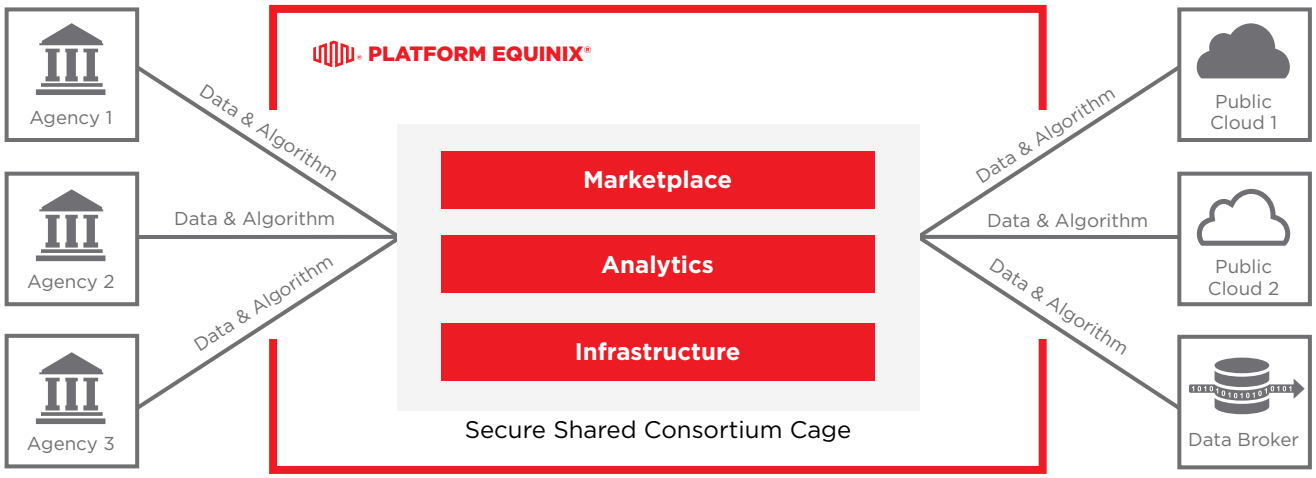


Fig. 1. A data marketplace enables sovereign organizations to make assets available to others for mutual benefit.

Use Cases and Requirements

There are many data brokers and data aggregators who buy and sell data and specialize in various sectors. The use cases and benefits are echoed in several government agencies as well, which show a rich diversity in applications across agencies, governance tasks and policy areas.[3]

Sector	Agency	Use Case	Consumer and Provider Benefit
Law Enforcement	Homeland Security	Biometric Identification and Predictive Reasoning	Prevent terrorist attacks, assess potential security risks, and enforce immigration and customs laws.[3]
Healthcare	Health and Human Services	Emerging Safety Concerns	Post-market surveillance and risk assessment of drugs and medical devices based on analysis of adverse events and medication error reports.[3]
Financial Services	U.S. Securities and Exchange Commission	Regulatory Enforcement	Identify violators of federal securities laws governing accounting fraud, trading misconduct, and unlawful investment advisors and asset managers.[3]
Intelligence	Defense	National Security	Using computer vision to aid video analysis in intelligence, surveillance and reconnaissance activities.[4]
Civilian	Housing and Urban Development	Citizen Assistance	Chatbot provides citizens with information about rental assistance, agency programs and civil rights complaints procedures.[3]

From a mission standpoint, stakeholders believe GEOINT-related AI will have the greatest impact on:[1]

National security

Urban planning and development

Emergency response/ natural disaster aid

There are many data marketplace solutions already out there, but they aren't fully addressing government agencies' data sharing challenges and concerns. In many cases, the data marketplace solutions available in the industry today do not adequately satisfy the demands of data providers or data consumers. A solution that gives agencies the confidence to share data and algorithms with each other and with partner organizations and service providers* as part of data marketplaces is needed.

Data Provider Requirements	What This Means
Full Control and Auditability	Gain full transparency into copies of data maintained in the infrastructure. Get control over which data can be taken out of the marketplace and which algorithms can be run on the data. Delete or make inaccessible data no longer being shared.
No Cloud Lock-in	Prevent data from being stored by a cloud provider for confidentiality and cost reasons (egress fees for moving data out).
Support for Different Security Trust Models	Supports different sharing models for data with different security requirements: bring the data consumer algorithm to the data; send data to the public cloud for use with data consumer algorithms; share data and algorithms in a neutral third-party location.
Distributed Solution	Allows certain data to be traded without leaving a region. Additionally, allows access to and processing of data at the edge to decrease latency and costs.
Data Licensing and Governance Options	Supports different licensing models for different types of data, and different governance models with respect to the operation of the marketplace.

Data Consumer Requirements	What This Means
Usage Privacy	Privacy with respect to how data is being used to create AI model assets.
Data Lineage	Know the source of the data to ensure people are not using incorrect data that leads to biased or inaccurate models.
Provider Reputation and Data Quality	Confidence in the data provider, data quality and data certification to ensure the quality of the AI models created.
Choice of Analytics	Access to different analytics vendors with expertise in different vertical domains. A choice of AI/machine learning (ML) frameworks and tools. The ability to bring in analytics frameworks in docker containers.
Support for Data Scientists and Production Workflows	The flexibility to support data scientists' production workflows: the ability to experiment with sample data on their own computers, build models with real data in a secure location, and take the built model out of the marketplace for use in a production environment.

3 Reasons the AI Data Marketplace Solution at Equinix Is Different

1. Multiple Data Sharing and Trust Archetypes are Supported

Different datasets warrant different data sharing and trust models. The AI Data Marketplace solution at Equinix makes it easy to bring data and algorithms into a secure, software-definable and geo-distributed data exchange sandbox. AI algorithms can be trained on data from different owners at different locations via the data marketplace, making the AI Data Marketplace solution at Equinix right for most government agency use cases. The solution supports all three of these data sharing mechanisms:

Distributed Model

Bring the algorithm to the data
Data providers who are unwilling to let data leave their premises due to confidentiality or intellectual property concerns can utilize a private cage at Equinix to host an AI training stack and run federated, privacy-preserving algorithms that require higher power density requirements.

Federated Model

Bring the data and algorithms to a neutral exchange location
Data and algorithm providers who prefer that their assets remain inaccessible in each other's locations can utilize secure, neutral exchange infrastructure cages inside Equinix data centers for data trading and algorithm use. Governance is negotiable. Raw data and algorithms are never taken outside the shared cage.

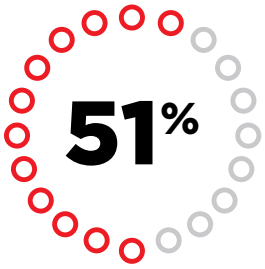
Centralized Model

Bring the data to the algorithm
Data providers who are comfortable sharing low-risk or non-confidential assets can utilize a public cloud marketplace or a hybrid model where the data to be shared is stored in a persistent manner in a private cage at Equinix, then moved into the public cloud infrastructures used by data science organizations for sharing or model training purposes.

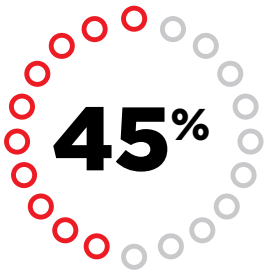
2. A Multi-Zone Architecture Ensures Data Is Always Secure

The AI Data Marketplace solution at Equinix deploys an architecture with three separate security zones (domains) for the control plane software, the provider/consumer data exchange, and the data provider's permanent data storage location (secure repository). Among the security advantages:

- **Protection against hacking**—Cybercriminals cannot access data being exchanged between parties in the data exchange zone.
- **Inability to take raw data out**—AI/analytics pipelines are executed in a sandbox run on a Kubernetes cluster where ingress/egress is strictly controlled.
- **Restricted access pattern monitoring**—Providers do not have access or visibility into how data consumers are using purchased data, and the IP of data consumer algorithms is protected.
- **Flexibility in security hardware**—Providers have several encryption options to ensure their data can never be accessed in the clear in the data exchange zone.
- **Time-bound access to data**—dProxy provides a layer of indirection that ensures a provider's real data storage location is never divulged to a data consumer.
- **Auditability and lineage tracking**—Lineage tracking can be done for any AI model created in the secure sandbox (data sources, AI frameworks, who did the model training).



51% of agencies surveyed identified security concerns as one of their biggest challenges as they look to expand AI over the next decade.[1]



45% of the largest 142 U.S. federal agencies have expressly manifested interest in AI/ML by planning, piloting or implementing such techniques.[3]

3. Federated Analytics Enable Privacy and Efficient Data Handling

The AI Data Marketplace solution at Equinix is distributed, meaning that a marketplace can simultaneously manage multiple geo-distributed data exchange locations at any given point in time. This allows it to support federated learning frameworks where local AI models on private infrastructure stacks (as shown in Fig. 2 below) can be built, then aggregated into a global AI model at a mutually trusted neutral location like Equinix. The AI Data Marketplace solution at Equinix allows data scientists to invoke third-party federated learning frameworks via Kubeflow. A federated learning approach is primarily useful for two reasons:

1. Privacy-preserving AI – When data providers want algorithm providers to ship their algorithm to the data location because they do not want to let raw data out of their security domain, federated learning can be leveraged to build a model locally, then share the anonymized model with the data consumer.
2. Efficient handling of large datasets at the edge – Federated learning is also useful when the size of the dataset being generated at the edge is large. Rather than sending it to a far-off central core location, one can build a local AI model and ship the model (typically kilobytes) to the central location rather than the raw data, which can reach into the terabytes. Since traffic doesn't have to be backhauled from the edge to a core location, it costs less.

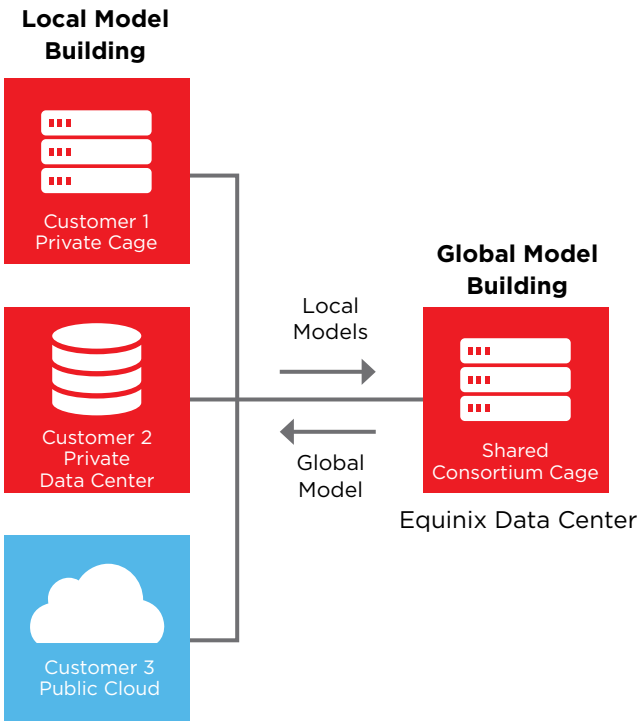


Fig. 2. A federated learning framework.

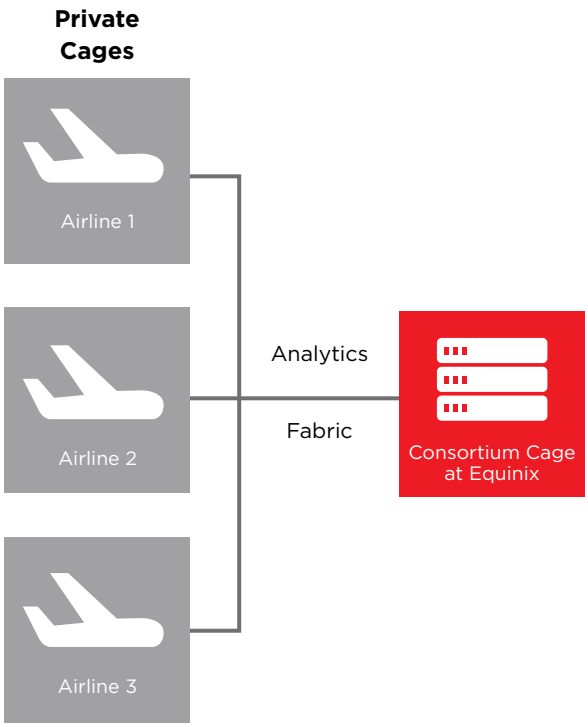


Fig. 3. Data and algorithm exchange between three airlines.

Data Marketplace Proof of Concept: Airlines

This proof of concept is indicative of multiple entities sharing data and algorithms in a marketplace via a distributed model (as described on page 7). We simulated three separate airline operators generating data in three different data zones. The goal was to: 1) show data sharing in a consortium-governed data marketplace via smart contracts while each airline retained complete administrative control of its raw data; and 2) determine whether federated/distributed analytics is as accurate in building an AI model as when data is brought to a central location.

Airlines that want to share data can store data in their own data centers or in private cages at Equinix at any metro location they choose. If agencies and enterprises are unable to host AI training hardware in their own private data centers due to the high power requirements of AI training hardware, they can have private AI stacks in private cages at Equinix, or they can get AI as a managed service at an Equinix data center of their choice.

Fig. 3 on the prior page shows the three-site configuration used to conduct this experiment. The table below demonstrates that the federated and centralized model training approaches provide a very similar level of AI model accuracy. This proves that an algorithm can be moved to the data to preserve data privacy and avoid the expense of moving large datasets, and an anonymized local model can then be moved to a central location to build a global model without any significant loss in accuracy.

Model 1.0: Random Forest: Top 5 Features*	
Predictive Performance	
Metric 1: AP **Metric 2:** AUROC	Federated = Centralized[1] Federated 1.7% higher than Centralized[2]
Bandwidth Performance	
Volume of data transferred	Federated 99.96% lower than Centralized[3]

1 Centralized AP = 0.19; 2 Centralized AUROC = 0.59; 3 Data volume ~ 7.5GB

Model 2.0: Random Forest: Top 20 Features*	
Predictive Performance	
Metric 1: AP **Metric 2:** AUROC	Federated 4.8% lower than Centralized[4] Federated 3.2% lower than Centralized[5]
Bandwidth Performance	
Volume of data transferred	Federated 99.96% lower than Centralized[6]

4 Centralized AP = 0.21; 5 Centralized AUROC = 0.62; 6 Data volume ~ 7.5GB

* In terms of the feature importance metric from Random Forests

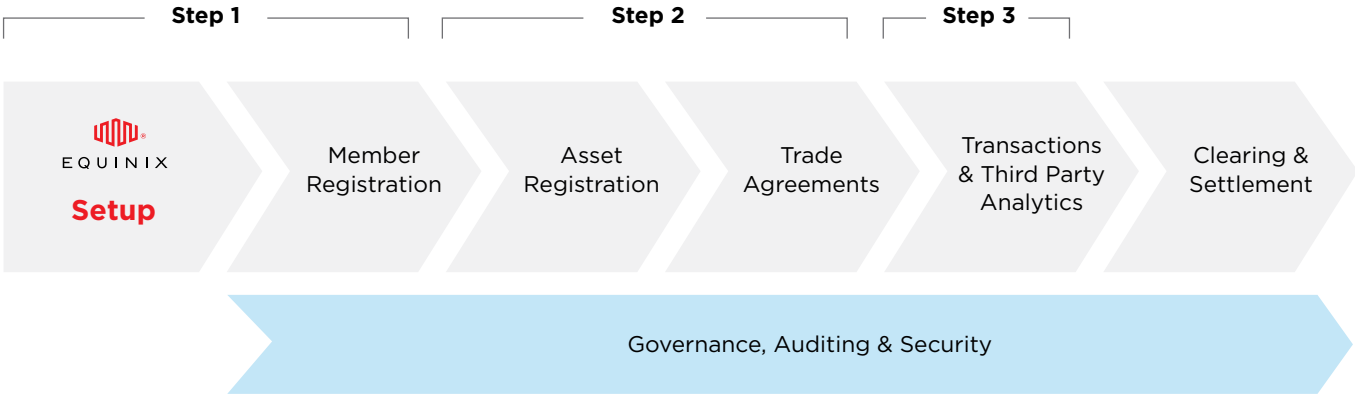


Fig. 4. Data marketplace usage workflow.

Steps to Standing Up a Data Marketplace

Step 1
Consortium Setup and Membership Registration

The first step is for a consortium to set up a data marketplace. This consortium may elect to create and offer a shared data processing infrastructure governed and administered by the membership organization. The consortium specifies the location of consortium-shared infrastructure, most likely close to where member data is located. Members may choose to store their data permanently in a private section of the membership infrastructure, or in separate, private data cages offered by a neutral data center provider such as Equinix.

Alternatively, a high-speed connection from a private data center, located near the data center hosting the consortium infrastructure, can be used to transport data for processing. Algorithm developers often deploy initially in public cloud infrastructures. Therefore, the consortium infrastructure also needs to be in proximity to cloud infrastructure services. Many large data centers offer cloud exchange facilities, enabling the consortium infrastructure to be located close to public clouds (e.g., AWS, Google, Microsoft Azure). Equinix data centers are interconnection hubs that are close to public clouds and end user devices.

Step 2
Asset Registration and Trade Agreements

After setup, a consortium provides access credentials to its members so they can register themselves and be admitted to the data marketplace. Once admitted, members can complete their registration and start registering their assets. Members decide which information they want to publish publicly in the marketplace to attract interactions to do business with other members. Once members agree to explore sharing data with each other and establish a contract that arranges visibility of available assets, information describing the available data (meta-data) becomes visible to prospective members. This allows members to negotiate access and usage of specific datasets and/or algorithms.

Once members agree on which assets can be accessed and used, they create a data trade agreement. This agreement authorizes subsequent data science transaction execution that accesses and uses data. All agreements are stored in an immutable, distributed ledger, which provides auditability for compliance or dispute resolution. The AI Data Marketplace solution at Equinix offers a blockchain ledger (Hyperledger) that allows the data asset trade agreements between members to be specified via smart contracts. It also logs all transactions on the Hyperledger for auditing and lineage tracking purposes.

Step 3
Data Scientist and Production Workflows

Data scientists can experiment in the data marketplace. They can upload their data science workflows from their laptops or from public clouds, then establish contracts with providers and examine the quality of their datasets on a test basis. After they are convinced of the quality of the datasets, they can purchase them and train their models in the marketplace data exchange location. They can also bring their own private datasets into the marketplace. Once data scientists have successfully trained a model, they can do model inferencing in the data marketplace or take the trained model out, with permission, and use it for inferencing in their private clouds.

The AI Data Marketplace Solution at Equinix

Deploy it for:

- Bilateral Data Exchanges Between Companies
- Single Entity-Driven Data Marketplaces
- Consortium-Based Data Marketplaces
- Data Sharing Between Agencies

Ready to get started?

Learn how the Artificial Intelligence (AI) Data Marketplace deployed on Platform Equinix® can help you meet your mission.

eqix.it/DESBfederal

References

1. Mapping AI to the GEOINT Workforce," Meritalk and the United State Geospatial Intelligence Foundation, April 2020.
2. "100 Data and Analytics Predications Through 2021," Gartner, 2017.
3. David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuellar, "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies," February 2020.
4. "Project Maven to Deploy Computer Algorithms to War Zone by Year's End," U.S. Department of Defense, July 21, 2017.

AI Data Marketplace_White Paper_US-EN | 0720