Federated Sharing Appliance Product Specification

Product Name: Federated Data Sharing Appliance (FDSA)
Version: 2.0
Product Owner: Prabhu Ganeshan
Date: Aug 2024

Overview

The Federated Data Sharing Appliance (FDSA) is a robust platform designed to enable secure, collaborative data sharing across organizations while ensuring each entity maintains governance over its own data sources. This appliance facilitates seamless integration with existing infrastructures, provides advanced data access management, and ensures compliance with data privacy standards. It is designed with a user-centric approach, offering a streamlined system for managing data access requests, query processing, and data result auditing.

Two Levels of Data Sharing

FDSA supports two distinct levels of data sharing, with data contributors deciding which level meets their requirements:

- Distributed Sharing (Level 1): At this level, researchers can analyze record-level data that is made available in an airlocked workspace, allowing for more detailed analysis.
- Federated Sharing (Level 2): At this level, researchers see only derived data and results in the workspace. The analysis of record-level data is done at the data source and, after a rigorous quarantine process, is aggregated and anonymized to ensure privacy.

Versatile Deployment Options

FDSA is optimized for flexibility, offering seamless deployment in both on-premise and cloud environments, intranet or internet accessibility, and easy integration into various infrastructure setups.

Cloud or on-premise Deployment: FDSA is designed to be cloud-agnostic, providing organizations the freedom to choose the cloud provider that best suits their needs. However, for organizations that require direct oversight of their infrastructure due to regulatory compliance, security policies, or operational preferences, FDSA can also be deployed on-premise.

Intranet or internet accessibility: FDSA can be deployed within an intranet environment, restricting access to internal networks for heightened security. It can also be configured with a fully qualified domain and set up as a public IP, making it accessible over the internet.

Tailored to your infrastructure: Regardless of your existing infrastructure, FDSA can be tailored to fit. It can be deployed and managed according to your specific operational needs based on your setup, without tying you down to any particular provider or environment.

Key Features and Functionality

Secure Data Sharing

FDSA ensures that all shared data is governed by the originating organization, maintaining compliance with data privacy and security regulations. Organizations have the ability to set and enforce governance policies over their data. Granular access control mechanisms allow organizations to define who can access specific data sets. All data shared between entities is encrypted and transmitted securely. The appliance includes a built-in mobile authenticator that provides two-factor authentication (2FA) to enhance the security of data access.

Data Access Approvals

FDSA offers a simplified, streamlined system for reviewing and managing data access requests. This system is integrated with the AD Workbench FAIR data access request framework, enabling efficient approval or denial of requests. Fine-tuned controls allow administrators to base access on detailed parameters.

FDSA Job/Task Query Management

FDSA manages and tracks the status of queries submitted by researchers, providing visibility into the entire query lifecycle. The system tracks queries through various states: Queued, Initializing, Running, Quarantined, Approved/Rejected, Complete/Rejected. It also logs and tracks the history of each query, enabling auditing and review.

User-Friendly Design

FDSA is designed with a easy-to-use UI that makes it easy for organizations to upload and manage datasets, operations, access, and maintenance. A centralized, easy-to-navigate dashboard simplifies all user operations and actions, including data access approvals and auditing of processed queries audit. There are also features for adding new users, admins, auto end-user creation upon data access requests, and user action management (e.g., role changes, disabling accounts, resetting MFA).

Easy Integration

FDSA supports seamless integration with the data contributor's infrastructure and internal systems, including data access request decision tools and applications. It enables extendibility and integration with external systems and internal applications through webhooks for managing data access decisions. It is also compatible with existing server environments, data management tools, and traffic restrictions through whitelisting process and limited IP traffic.

Data Quarantine and Audit Process

FDSA provides a quarantine process for data results, granting administrators the authority to review and audit processed data before release. They can either approve or reject it for release and provide feedback on their decision.

Docker Registry - Model Read

During data processing, FDSA can access and read published data researcher models from a secure Azure Docker registry and process inside a container.

Data Connectivity

FDSA includes database connectors that enable easy connections to remote data sources served as federated data. It features built-in connectors for PostgreSQL, supporting structured data sets and facilitating complex data queries across multiple databases.

Tech Stack

The Federated Data Sharing Appliance (FDSA) is built on a robust and scalable tech stack, incorporating a range of modern technologies and third-party components to ensure high performance, security, and seamless integration.

Core Technologies

- **Docker:**
  Utilized for containerization, allowing FDSA to deploy and manage applications in a consistent environment across multiple systems. Docker ensures the portability and scalability of the appliance.
- **Docker Compose:**
  Facilitates the orchestration of multi-container Docker applications. Used for defining and running the multi-container FDSA services, ensuring they work together seamlessly.
- **nginx:**
  Employed as a high-performance reverse proxy and web server, nginx handles load balancing, security, and routing within the FDSA environment.
- **Python:**
  The core programming language used for developing FDSA's backend services and automation scripts. Python's extensive libraries and frameworks contribute to rapid development and integration.
- **Flask:**
  A lightweight Python web framework, Flask is used for building FDSA's RESTful APIs and web services. It allows for quick and flexible development of secure and scalable applications.

- **PostgreSQL:**
  The primary database management system used in FDSA, PostgreSQL provides robust, scalable, and secure data storage, with support for complex queries and transactions.
- **Hasura GraphQL Engine:**
  Integrates seamlessly with PostgreSQL to provide a powerful GraphQL API for real-time data access. Hasura automates much of the data access process, making it easier to manage and query federated data.
- **Keycloak:**
  A comprehensive open-source identity and access management solution, Keycloak is used in FDSA for managing authentication & authorization, and 2FA.

Third-Party Software Notices and Information

FDSA incorporates the following third-party components, each integral to the appliance's functionality:

- **Docker** - For containerization and deployment.
- **nginx** - For reverse proxy, load balancing, and web server functionalities.
- **Python** - The core programming language for backend services.
- **PostgreSQL** - The primary relational database management system.
- **Flask** - A web framework for building APIs and web services.
- **Docker Compose** - For orchestrating multi-container Docker applications.
- **Hasura GraphQL Engine** - For providing a real-time GraphQL API.
- **Keycloak** - For managing authentication and authorization processes.