December 2020

FEDERATED DATA ACCESS

HP INC

Federated Data Access

1 Abstract

Data has become critical to drive decisions, manage devices, support engineering teams, support service support teams, and many other uses on data driven organizations.

Data is coming from multiple sources, collected and stored in multiple systems and not easy to share avoiding duplications and inefficiencies; assuring data quality and consistency.

Current approaches are creating data layers which internally are managing copies of the data for operational purposes.

Due to the amount of data systems available it is also difficult to be aware of the existence of this data or difficulties to share it so usually it ends with teams creating their own data sets (with different quality governance, transformations and logic). This can (usually) lead into inconsistencies on data being reported depending on the system being used; inconsistencies between same data on different systems. This is usually tried to be solved with Data Cataloguing tools which provides visibility of the available data sets/systems and their description context, tables, attributes.

Data is also governed by different owners but, as this data is shared through synchronizations/copies, it is difficult to keep this ownership (and governance) of their transformations; including data quality, consistency and correctness.

The Federated Data Access (FDA) is a system which will allow:

- Publish and subscribe to data sets
- Governance data sets

FDA will allow the above securely and avoiding creating data copies.

2 Problems solved

Single access point to all domains within a company or organization which are maintained, accessible with required access control.

Avoid duplicated data sets. When we are talking about Big Data we are talking about very large amounts of data and the involved cost to process and store is expensive; specially when double or triple paid.

Governance, data belongs to a domain (or a joint of different domains). These domains are the ones having better knowledge of their data in terms of quality, consistency and reliability. If data is not governed it can end (usually does) in different teams manipulating same data in different ways. The result of this (besides the associated cost of duplicated work, processing and storage cost) is that on different touch points the business the same data is showing different values. When talking about business reports to take decisions, customer facing data or data to be used by machine learning algorithms this can lead to lack of credibility and incorrect decisions.

3 Prior solutions and limitations

Some of previous solutions were generating one duplication level in order to manage data access.

Some others were managing the direct connection to the source data systems (DB connection, REST API…) but were not managing the subscription of data sets, managing lineage and data governance.

4 New solution description

Federated Data Access (FDA) system makes a role to manage data traffic between data systems (or sources) and data consumers. It enables the existence of one single data set for the same domain through lineage and governance, enables consumers to discover and access multiple domains from the FDA itself. It also makes consumers agnostic of the exact data system (server/database/table) being accessed through the FDA subscription which acts as a proxy.

Novelty is on putting together single data access system for multiple data sets, data sets publication and subscription (to have this single data access system) and the complete process being governed by the corresponding data owners.

Subscription is not based on consumer being notified when change in data set but allowing the consumer to query FDA as if they were querying the original data system. The query type (SQL, REST API…) will be determined by how Data System has published the data set (DB connection, REST API…). FDA is not defining a homogeneous query language which consumers use and FDA translates into original Data System language. It will require complete query language implementation, query optimization and manage multiple languages according to the Data Systems heterogeneity.

Data System has to be onboarded stablishing its data domain and data administrators/curators so, they are registered as authorized to perform some administrative operations. This will stablish the Domain Governance team for their published data sets; when they are published.

Data System will call FDA to publish new data set. Each data set will require information about involved domains, its data source, transformations and data quality/consistency status. FDA will call all domains governance members to review new data set. Each domain governance will approve/deny the new data set; it is sufficient one denial to not accept new data set. If denied domain governance will provide reason (duplicated data set, incorrect transformation, bad data quality…) and Data System will have to work with domain to solve the issue. If accepted data set will be registered, store its metadata and creates the lineage relationships with the original data sets (if applies).

For simplicity and clarity, it is assumed data systems have been onboarded (stablishing their domains and user roles being assigned), users are logged in the system (authentication) and are authorized to work on some domain governance (authorization based on the roles assigned as per the domain they belong).

How the system will work:

1. Data set publication

   New data set publications can be due to 2 reasons: non-derived data and derived data. But in both cases the process will be the same and are described with steps 1 to 4 in Figure 1.
   There will a data developer, from the data consumer area, which creates in data system a new data set (derived or non-derived).
   Step 1: data system request to Publication service permission to publish new data set. Information included in the request: involved domains, data sources, transformations done to create the data set, quality and consistency status).

Step 2: FDA will use provided domains to call affected governance domain teams. If data set is from a new domain then this new domain will be created, and data system will become the governance team.
Step 3: Each governance team (related with the affected domains) will emit their vote.
Step 4: If all governance teams have voted positive (accepted the new data set) the data system will provide connection details so that FDA can stablish connection (through a service account on the data system) and FDA will include domain, domain lineage with other data sets, sources, transformations, quality and consistency data in data catalogue and make it available for subscription. Connection details by data system is what will be used when consumer wants to query that data set.

2. Data consumption

Data consumption will be performed by data developers (which can involve data engineers, analytic teams…) which will explore FDA (data catalogue) to identify data sets to subscribe. Once subscribed (subscription will be supervised by the related domains governance teams to assure data developer/consumer can be authorized to access it). If subscription succeeds the consumer will receive the domain token so can be used for future accesses.
The process can be described with steps 5 to 9 on Figure 1.
Step 5: data developer in a data consumer environment is exploring catalogue through FDA subscription service. The data catalogue informs about available domains, available data sets and lineage between data sets.
Step 6: once data developer identifies the data set which is interested it requests it subscription.
Step 7: Each governance team (related with the requested data set) will emit their vote. If all votes are positive (accepted) then consumer will also receive domain token and a temporal access token. This temporal access token is used to query FDA for a given domain which is required to be renewed to assure security when accessing the FDA.
Step 8: Data consumer knows from the data catalogue the type of query/access it can request FDA for the data set (into the data system). Consumer will use the temporal token to query FDA for the domain (it won't use final data system connection string or data tables only FDA domain).
Step 9: FDA validates the temporal token received and translates into which data system to query. Replaces the FDA domain by the corresponding end data system connection and tables/columns. When data system returns results, they will be returned to consumer.

If consumer is already subscribed to a data set then in step 5, consumer, will request new temporal token for the subscribed domain. Step 6 FDA will review authorization, subscription and generate a new temporal access token. Step 7 new temporal toke is returned to consumer. Step 8 and 9 will be exactly the same as above.

Steps 4 provides a connection string so that FDA knows how to connect with data system (DB, REST API or other) so that FDA can translate FDA queries into end data system queries (redirect queries). This redirection should be through a DB service account or REST API or mechanism that allow FDA to access Data System.

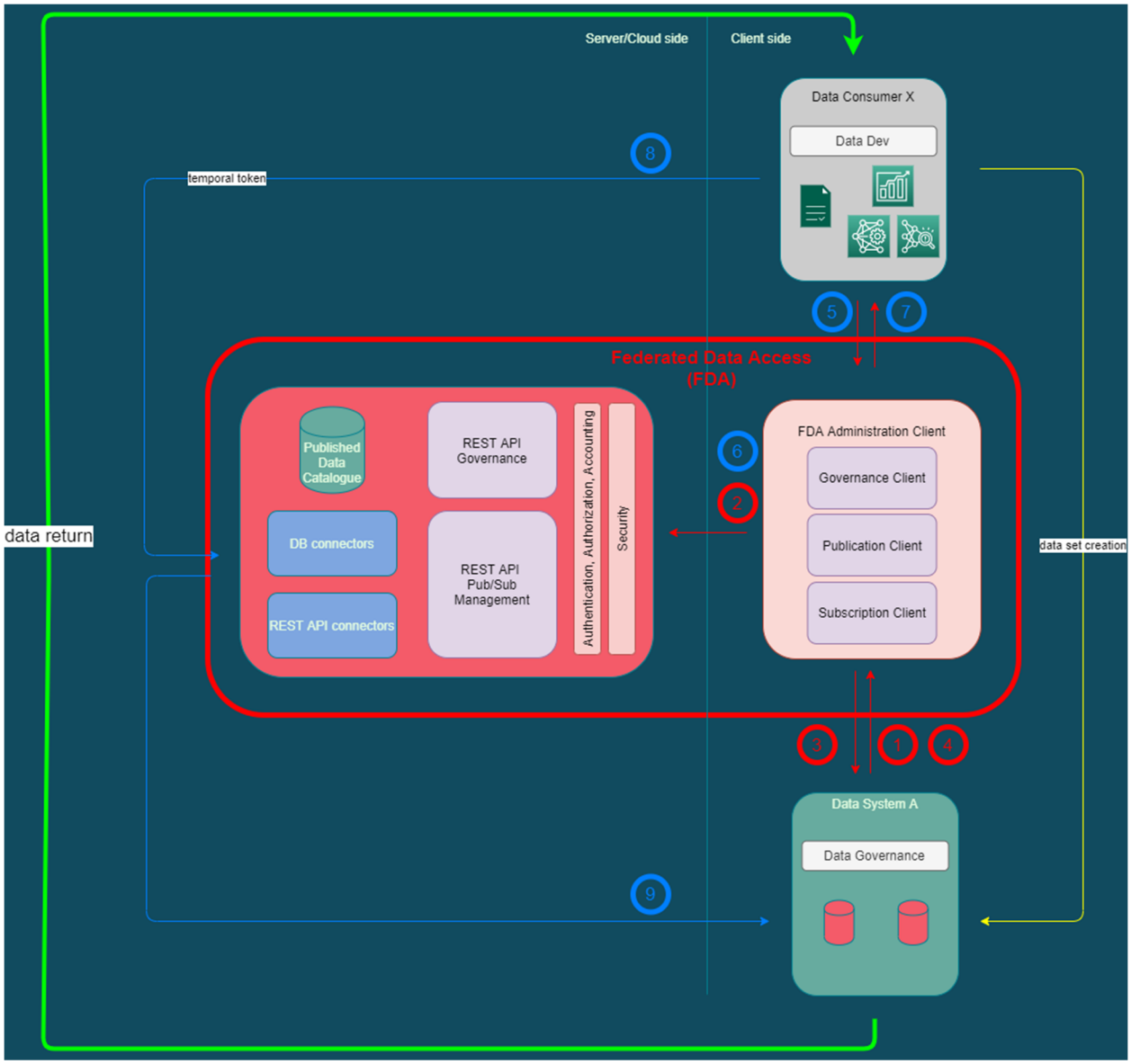

Figure 1 Federated Data Access (FDA) functional block diagram

5 Advantages of our invention

The advantages are:

- To provide a single access to curated data sets (Golden Data) across company or involved organizations.
- To publish curated data sets and stablish its ownership and governance.
- To divide data sets into Governance groups.
- To enforce a process which assures no new datasets overlaps or duplicate already existing data sets
- To stablish data access authorization.
- To allow authorized consumers access the data through subscriptions.

- To make any change on data system (system/DB replacement or data set improvement) transparent to data consumers as they access through the FDA subscription.
- To publish transformations/enrichment of new data sets with the right owner's acceptance according to their data lineage.
- To avoid duplicating data in the access layer.

Disclosed by Roque Bonilla Lucas, HP Inc.