ebook

# Data Fabric from Equifax

## A technology, security, and privacy overview

Published June 2023

# Contents

## Who should use this document

This document is intended for Equifax customers and partners to gain a deeper knowledge about the **technology used to ingest, store, and ultimately deliver products and solutions to customers and partners** from the Data Fabric. This is a part of the overall cloud-native transformation at Equifax.

**Business leaders** are encouraged to learn more about the business transformation underway at Equifax while **technology leaders** can reference this document to learn about the stages, components, and data privacy inherent in the **Data Fabric from Equifax**.

## Data Fabric migration overview

As part of the Equifax cloud-native transformation, we are **migrating data, products, applications, and networks to the cloud** for greater flexibility, scale, reliability and system insights.

As the final component of a multi-year journey, the data migration is designed with testing and validation at the core. Once in the cloud, our vast array of differentiated data sources available will be housed on a **single platform known as the Data Fabric, logically separated with strict compliance and data governance controls**. Some of these data sources, including **employment and income and digital identity, are only available from Equifax**.

The Data Fabric is a **cloud-native enterprise data management platform** that aggregates all data received by Equifax into a single environment, deployed regionally on the Google Cloud Platform (GCP). This data is subsequently made available to customers in the form of products solutions, such as scores, models, and attributes. Our Data Fabric generally consists of two integrated capabilities — **data pipelines** and **analytical services.** Underlying governance capabilities known as catalog services enable Data Privacy and Protection (DPP) to be an integral component of the Data Fabric.

## Data pipelines

The Data Fabric onboards and manages data assets ingested by Equifax by providing a set of API-based services through the data lifecycle stages of **Preparation, Ingestion, Keying and Linking, Journaling, and Purposing**.

- **Data Preparation and Ingestion** are the services for receiving and pre-processing raw data assets by applying data quality rules and converting data into a format that can be further processed and analyzed.

- **Keying and Linking** are the operations that identify the entity (e.g., person, business, or other) to which specific data should be associated. Each entity has a unique key assigned to it and expressed as a numeric identifier. This key is added to the specific data during this stage, which then enables disparate data to be linked to the same entity.

- **Journaling** is the service that receives and stores the prepped and ingested data. This stage is called Journaling because the specific technique employed to store data is to essentially record sequential observations (i.e., comparable to journal entries). Journaling is responsible for persisting new observations and for combining new observations with existing master observations. For example, an observation could be a new address associated with a person.

- **Purposing** is the service that receives journaled data and applies rules relating to a specific use case. Once the rules are applied, the resulting data is available to be viewed or extracted. The Purposing processes are typically executed shortly after data is journaled, with the goal of combining any new observations with other existing, historical observations, typically at a person or entity level.

## Analytical services

The Data Fabric includes a set of analytical services that enables business users, such as data stewards and data scientists, to analyze and create insights using data from the analytical environment (as described in more detail below), which is essentially a collection of snapshots taken during the data lifecycle stages.

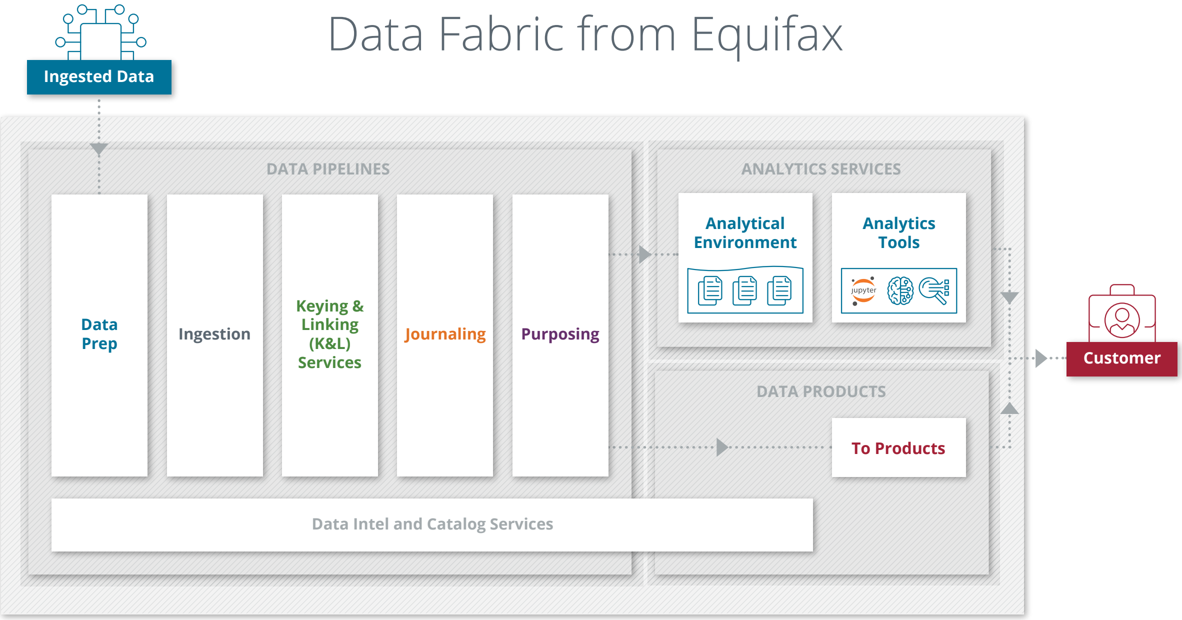
## Catalog services

The Data Fabric includes a set of catalog services to record information about the data assets in Data Fabric, which is also referred to as metadata or "data about data." Catalog services are designed to record, track, and manage information about the data assets, the transformation operations performed against the data, and the destinations of any copies or extracts made of the data.

The catalog services consists of two components:

1. a GCP-hosted component called the Catalog that is an integrated part of the data fabric
2. a licensed third-party solution called Collibra.

The Catalog is focused on supporting the operational needs of the Data Fabric, while Collibra is focused on the broader, corporate-wide metadata requirements. The combination of these components provides a comprehensive view on Data Fabric data assets. In the future, these components will become integrated. This diagram illustrates the previously described Data Fabric capabilities:



As a frame of reference, the following chart compares the components of a physical manufacturing warehouse with the Data Fabric. The warehouse receives shipments of parts from various sources, records, and categorizes them. The parts are then unpacked, sorted, and properly stored in the warehouse to facilitate the manufacturing process. In some cases, the parts may be picked, partially assembled, and then returned to the warehouse until needed. Ultimately, the parts are moved to the manufacturing floor and assembled into a finished product, which is then packaged and shipped to the customer's destination.

| Physical manufacturing warehouse | Digital warehouse (Data Fabric) |
|---|---|
| **Parts** | Ingestion of contributed data |
| **Partial assembly** | Data Preparation and Keying and Linking (e.g., cleansing, deduplication, keying and linking specific data) |
| **Recording/Cataloging** | Catalog services |
| **Warehouse** | Data Fabric data repositories |
| **Final assembly** | Purposing (e.g., data extraction, filtering, purposing) |
| **Packaging and shipping** | Transformation to a target format, including encryption and data/file transfers |

## Data privacy and protection

Four key call-outs related to Data Privacy are covered in this section:

1) Data isolation (or segregation)
2) Keying and Linking capabilities
3) 'Least privilege' applied to analytical services
4) Data retention

### Data isolation (or segregation)

Historically, our applications were built for specific purposes and were embedded with hard-coded business logic. This strategy created a complex legacy infrastructure that was resource-intensive to maintain. The Data Fabric solves these problems by storing the data assets into a single, connected platform designed with the technology allowing for data isolation — as required by our legal, contractual or regulatory obligations and other business requirements. Additionally, it offers the flexibility to easily implement and modify business logic on the platform.

GCP, like aspects of our legacy infrastructure, is a multi-tenant environment with data kept logically separated from users based on approved access levels. For data residency and performance purposes, data is stored in regional repositories based on the geographical use case (e.g., United States, United Kingdom, or Australia). The Data Fabric is hosted on the GCP and is also designed to support multi-tenancy, meaning that it can process and house data assets from different sources or customers while maintaining logical separation between these assets unless an intentional decision to combine the data assets is made.

#### *Domains and subdomains*

Multi-tenancy has been implemented in the Data Fabric on multiple levels as required by our business by utilizing *domains and subdomains*. A *domain* is a broad category of data that we ingest (e.g., Credit, Employment, Telecommunications, Wealth), and a *subdomain* is a subcategory of the *domain*. In other words, a domain is a collection of *subdomains* and a *subdomain* may only belong to one *domain*. An example of a *subdomain* of the *employment domain* would be Payroll. Each domain stores data in its own storage repository. (Please see the retention section later in this document for more details on data storage.)

In the context of *domains and subdomains*, the Data Fabric helps solve problems with the legacy infrastructure because our business is able to define and validate rules for their *domains and subdomains*. These teams can easily implement and modify business logic at the *domain* or *subdomain* level. In the end, our Data Fabric will consist of minimal hard-coded rules, relying on the business logic rules applied to the *domain or subdomain* level.

### *Data pipelines*

Conceptually, one can consider a *domain* as a dedicated pipe that maintains data separate from other *domains* until business logic is applied to combine data from different *domains*. The *subdomains* allow for further categorization within *domains*.

- Data remains in its assigned *domain* through the data pipelines and will only be allowed to mix with data in other *domains* using approved purposing rules.
- Logical separation relies on access controls and devaluation to prevent co-mingling of data assets that could potentially violate our legal, contractual or regulatory obligations and other business requirements. The following rules are consistent with our security framework:

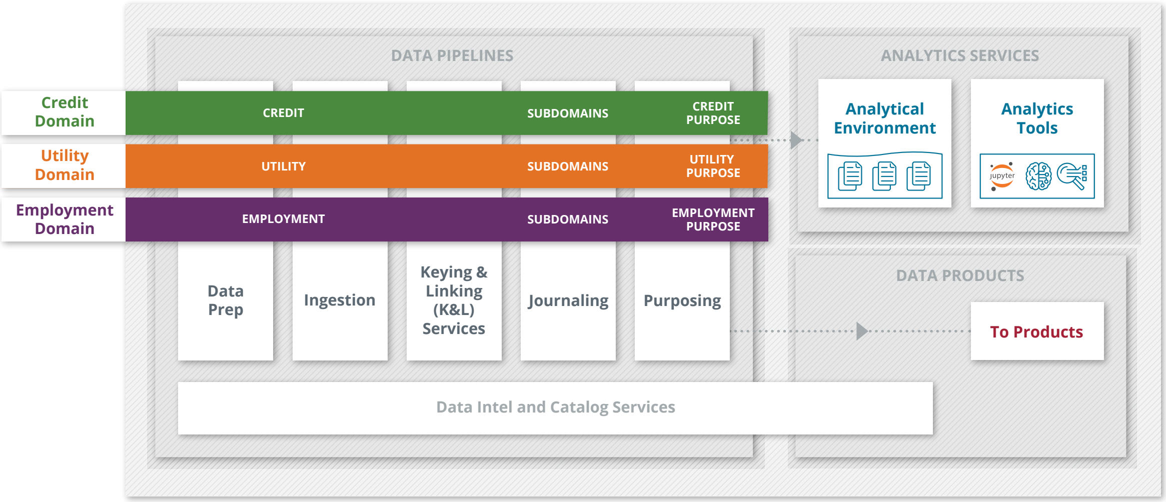
  - We require data contributors to encrypt their data using the Data Fabric’s public key. The data is then ingested and decrypted using the Data Fabric’s private key. (Similarly, data that is transferred from the Data Fabric is encrypted with the recipient’s public key.)
    - All data reposed in Data Fabric is encrypted using GCP native encryption.
    - The most sensitive contributed data is also encrypted at the field level using our Barricade library, a tool which allows us to manage the encryption and decryption of data directly, using GCP supported cryptographic libraries.
    - Finally, the most sensitive contributed data is also encrypted at the field level.

  - Data within *subdomains* of a particular domain is encrypted with the same key. Data in a particular *subdomain* is encrypted with a key that is different from data in a *subdomain* in another *domain*. In other words, for data segregation, data within each subdomain of a particular domain is considered “friendly” data, Access to the key needed to decrypt the data in each domain is controlled through access groups.

  - All of the Data Fabric and GCP Services are secured using internal Identity and Access Management (IAM) roles that are specific to the type of account being used and the role that account is intended to serve.
    - Service accounts are not usable by end users and have access tokens protected and automatically managed by Google Key Management Service (KMS).
    - Service account permissions apply to each stage in the data lifecycle and are provisioned by the Equifax Security’s IAM team.
    - User account access is centrally managed with roles assigned by Equifax and utilizes multi-factor authentication.
    - Data in the pipelines is generally accessible only through service accounts, meaning there is no human equivalent with access to the data.[1]
    - While access to data environments by system operators or administrators is allowed, these accounts do not have access to the encryption keys and do not have direct access to unencrypted data.

The following diagram illustrates the logical separation of *domains* within the data pipelines. In this illustration, data for each business *domain* (e.g., Credit, Utility, Employment) are isolated from other *domains* using logical data pipelines.

Data Fabric from Equifax



*Analytical Services (Analytical Environment)*

Data that is copied over to Analytical Services will reflect its original logical separation. Each application utilizing the data will be approved to access data sets based on a governance review. In some cases, where review has approved the use, data from different data sets may be accessed by the same application. As explained below, the principle of least privilege is followed for all data access implementations.

**Keying and Linking capabilities**

Keying and Linking is a critical service in the data pipelines that facilitates journaling, purposing, and ultimately the delivery of products and solutions to Equifax customers and partners. Keying and Linking is the process of identifying the entity that information is associated with and assigning a unique key to it. Keys are assigned at the entity level, and an entity can be identified across domains using a "combined key" when allowed by business rules.

Here is an example of Keying and Linking:

| | |
|---|---|
| John Smith, 123 main st, NY, SSN 123456789 | **Key: 75629** |

And

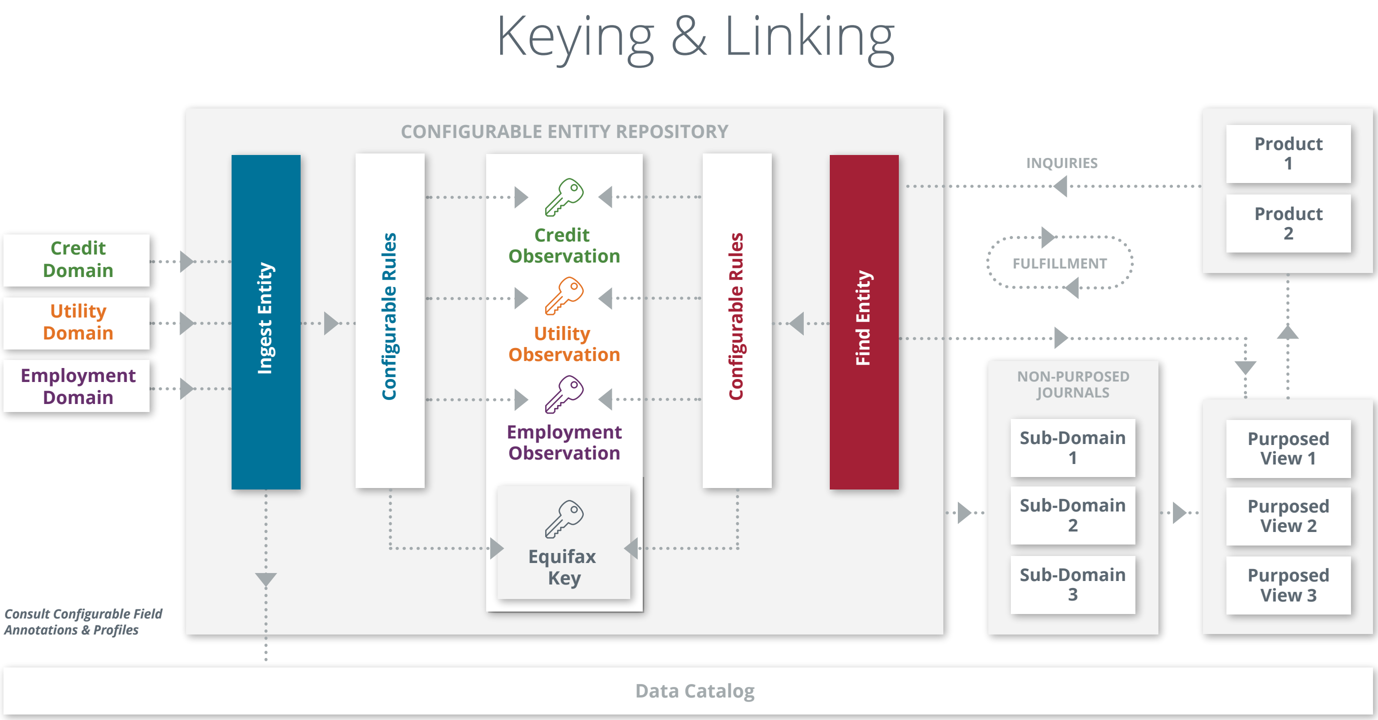
| | |
|---|---|
| John Smith, SSN 123456789, Acct 8765432, Payment 100.00 | **Key: 75629** |

By using the key "75629," these two data sets can be linked together and associated with a single entity (in this case a person). These two data sets could be from the same data source or different data sources within one or more *domains*. They are keyed and linked based on which rules are implemented for the Data Fabric at the business unit or regional level.

*Keying and Linking functional design*

Each data source ingested into the Data Fabric requires defined rules to be orchestrated at the field level to specify how annotated fields can participate in the Keying process. This is defined in a Keying and Linking profile which is a configuration to prescribe which Keying and Linking operations are available for the data received. In practical terms, this means that the data elements used for a key are defined specifically for the use case, not generically across all data within the Data Fabric. This allows our business to utilize a common capability with the flexibility to define specific rules for each respective use case.

The following diagram illustrates the functional design of the Keying and Linking process:



The configurable annotations are combined with the profile information for each source. This drives what each source will contribute to the Keying and Linking process. The profile information contains a set of configuration options that prescribe what each source can do in the keying process. Additionally, it points to the rule sets to be applied when each source is processed (i.e., which fields within the data source can be used).

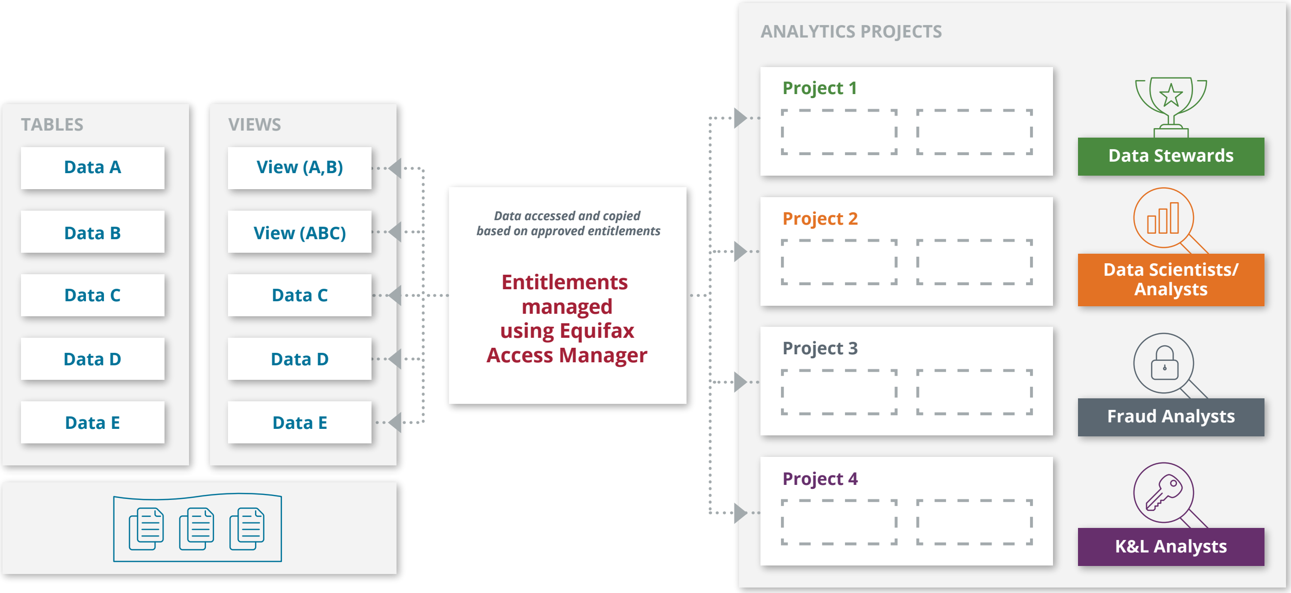
The Keying and Linking process is important and enables us to offer a unique capability because it allows for entity resolution across data sources when data elements are inconsistent. In other words, the Data Fabric can identify an entity based on many combinations of data elements instead of a predetermined list of required fields, which is adaptable to meet our unique business and regulatory environments.

### Least privilege applied to analytical services

As referenced above, analytical services provided by the Data Fabric enable our data analysts to carry out comprehensive analysis across large amounts of data to gain insights that benefit our customers. Unlike the data pipelines that are only accessed with service accounts, users like data stewards and data scientists naturally access the data in analytical services and, therefore, require us to implement the concept of least privilege.[2]

There are two concepts that come together in the Data Fabric which enable granular access to data: **tables and views**. In general, tables represent the physical storage of data, whereas views are access rights that enable users to view only certain portions of the physical data stored in tables. Because portions of data do not need to be duplicated for different uses, views enable us to better adhere to our data retention governance objectives. Views can also be defined at the most granular level: the data element. Users can create as many views as necessary to support their business objective, which are each subject to their own access entitlement. Views, therefore, also fulfill the least privilege requirements.

The following diagram illustrates the governed data access in the analytics environment:

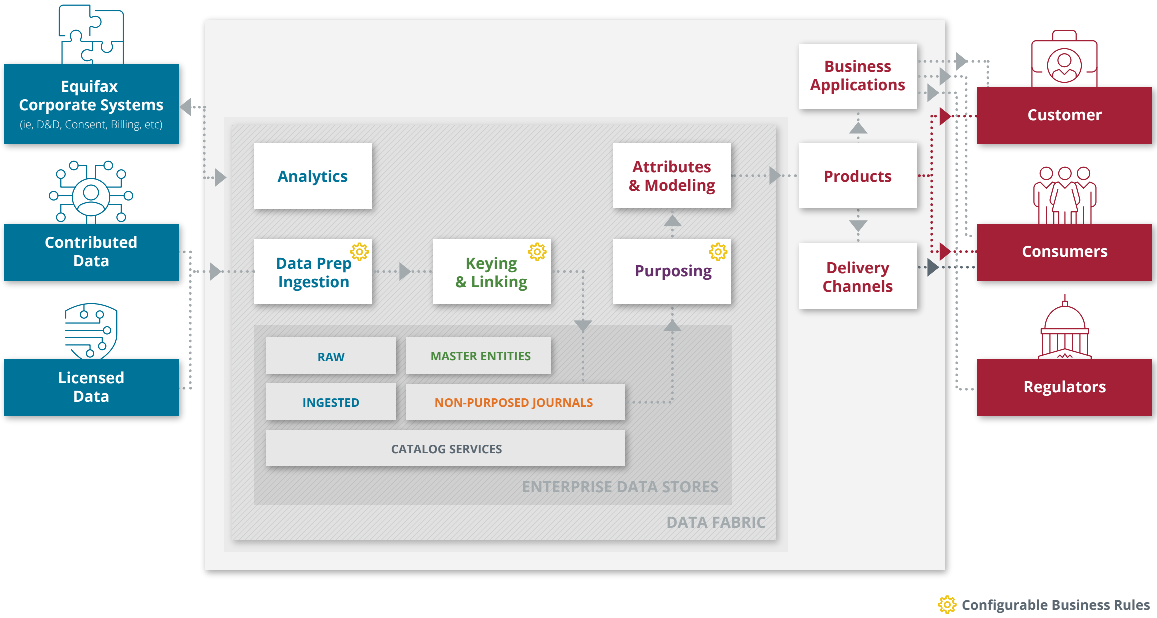


The Data Fabric ultimately provides more granular access control than the legacy on-premise environment. The chart below illustrates how data access works with the legacy and new Data Fabric environments:

| | Legacy | Data Fabric analytics | Remarks |
|---|---|---|---|
| Project level entitlements | Managed at Application level | Managed through Access Manager | The Data Fabric enables standardized access management. |
| Data entitlement | Role-based access control (RBAC) | Functional equivalent of RBAC using 'views' | The same user experience is provided in both environments. |
| Analytics tools entitlements | Open access across projects allowing data movement across projects | Data movement only within a project | The Data Fabric provides better control over data movement. |
| Purpose-based governance | Leverage tool (Privacera) and manual audits; data movement is monitored | At the time of publication, this is the same as the legacy method. | A new single solution is currently under development. |

### Data retention

The Data Fabric reposes data in different locations as represented by this view:



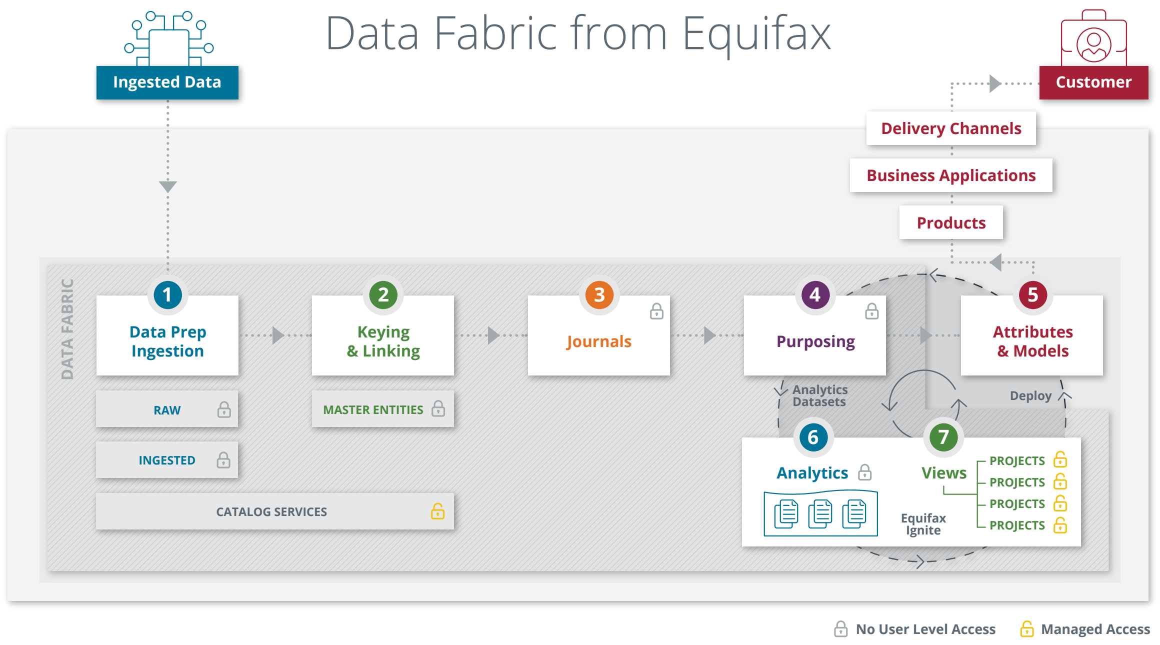
Each storage repository is a GCS bucket, a BigTable instance, or a BigQuery data set. A GCS bucket can be considered roughly equivalent to a file folder, and a BigTable instance or a BigQuery data set can be considered equivalent to a database.

All data assets in the Data Fabric can be broadly classified into (i) a "system of record," which is in the data pipelines and (ii) snapshots and copies of the system of record that are carried over to analytical services. (With respect to the views mentioned above, they are views of the snapshots or copies carried over to analytical services.) The Data Fabric can implement retention policies in each of the repositories based on the two classifications (e.g., system of record with one retention period and a snapshot or copy another) in compliance with the Equifax Global Retention Policy.

Though retention periods can be implemented at the system level in each repository instance, the Data Fabric is designed to expect retention policies and periods to be managed by data stewards using meta data services within the data catalog.

## Visual example

The diagram below shows a high-level flow of data in the Data Fabric. The steps provide an example of how data contributed from a credit data furnisher flows through the Data Fabric platform.



**1** Data is received from a data furnisher and routed to Data Fabric for prep and ingestion. At this stage the data is put into the “Credit” domain and “Consumer Credit” subdomain based on the data source (i.e., the credit data furnisher).

**2** Once ingested, the data is keyed based on the configuration for the furnisher, credit domain, and consumer credit subdomain.

**3** The keyed data is then passed to the journaling stage and the keys are used to update the credit domain journals for the entities present in the furnished data.

**4** One purpose for the credit domain is consumer credit reports; data goes through the credit reporting purposing process to aggregate data necessary to deliver a credit report. Data is also used for analytic projects, so the data is passed to the analytical environment.

**5** Data is then used by Attributes, Modeling and Product/ Business Application to prepare and deliver the credit report to the customer.

**6** Purposed data gets transferred to the analytical environment for analytics projects based on approved use cases.

**7** Data in the analytical environment is accessible by users based on job role using the tables and views concept.

## Learn more about our **cloud-native strategy and find additional resources** to address your questions.

[1] Limited human accounts are utilized to maintain data integrity and handle error-correction. These accounts are monitored and limited.
[2] In general, ‘least privilege’” means that a user is given access only to the data needed for that user to perform their job.