Accenture Labs

accenture

Bringing data together:

A MODERN DATA ECOSYSTEM

From managing data to mining information

Today's enterprise data management enables companies to "know what they know," but is not able to widen the aperture to reveal insights that are outside an initial query's scope. It is difficult to find and access relevant contextual data.

Companies need an integrated view across the entire organization providing a common way to govern and manage data. This view combines critical pieces of metadata: technical, business, and operational across the data supply chain—from sources, through transformations and refiners, to staged and derived data, their models and their composites, to reports and analytics applications.

Tracking data to provide a holistic view is still largely a manual activity today. Organizations conduct interviews with subject matter experts and record the details in systems and spreadsheets that may or may not get updated. And that can only ever provide a static snapshot of metadata at a given point in time.

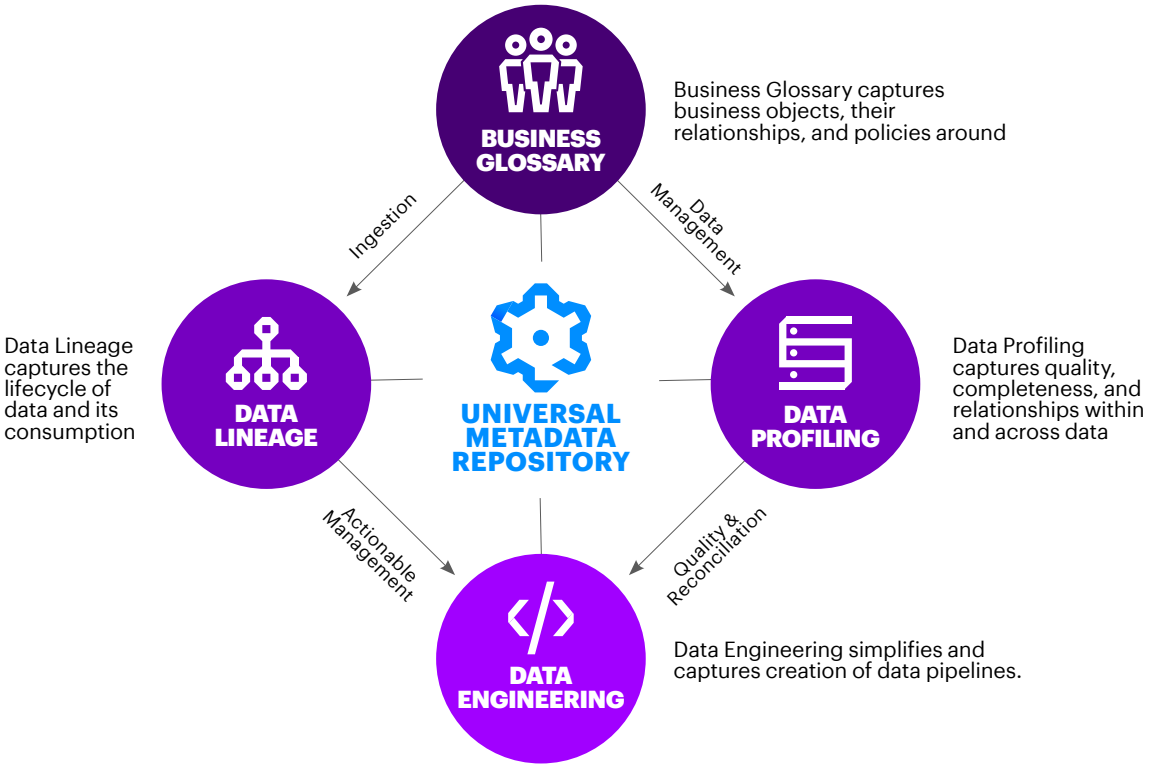
Rarely are these systems connected to the data storage and pipeline mechanisms. The emerging solutions in this space are niche: Some focus on the data lake, some track the lineage of pipelines, and others examine content based on semantic models and profiling. But each tool is restricted to a niche capability and operational environment. There is a gap between where the business rules are defined and documented and how they are actually enforced.

Introducing the Universal Metadata Repository

Our Universal Metadata Repository (UMR) uses a data supply chain specific knowledge graph—the same technology that underpins today's Internet—to provide a single logical view of the data supply chain to support the following capabilities:

A single federated system: A single logical view of an organization's data incorporates the views of many tools and allows for them to interoperate and exchange metadata. UMR augments existing tools, providing a way to exchange metadata across business and technical domains. In this way we can stitch together the results of data engineering, profiling, lineage and glossary tools (See Figure 1).

FIGURE 1: The Universal Metadata Repository (UMR) sits at the center of the data ecosystem bringing together technical, operational and business metadata from underlying tools such as data engineering, profiling, lineage and business glossary.



A living system: Machine learning augments and accelerates the manual process. Data rules and metadata are discovered automatically and can be dynamically updated as data changes.

Data rules are enforced: Rules are enforced through dynamic queries automatically detecting discrepancies, and in many cases generating implementation updates.

Where an organization has already invested in a number of underlying tools, UMR unlocks their value by combining the metadata into a single federated view. An organization building a metadata management capability from the ground up can unlock value by using the UMR with a subset of complementary tools providing data quality, lineage, and catalog.

The UMR augments the following tools:

Data Catalog & Discovery

The UMR augments the enterprise data catalog by providing advanced internet search-like functionality including keyword search, autocomplete options and filtering the catalog based on other metrics such as timestamps, data quality metrics and domains.

Data Quality

By tracking all the metrics and how they relate to one another, the UMR monitors data quality programmatically. Using the metadata across the supply chain, the system computes aspects such as accuracy, completeness, timeliness and trustworthiness. The UMR allows each data user to add their own measure of quality that is automatically computed and tracked over time.

Data Lineage

Users can trace the data from its system of origin all the way to where it is being consumed, tracking changes along the way with the UMR. The UMR captures parent-child relationships, transformations and operations performed on the data, the type of technology or transformation script and timestamp.

Policy & User Management

The UMR uses Role Based Access Control (RBAC) to create user roles and allocate permissions accordingly. For example, an admin user group will have enhanced privileges as opposed to the Regular User whose default access will be limited.

Business Glossary

The business glossary defines the business terms and definitions related to the physical assets so that users can understand the semantics behind the data and what/how it is being used. Using the UMR we create the mappings between the physical data and the business terms to help stakeholders across the organization collaboratively agree on the definitions, rules, and policies that define their data.

Connectors

Custom connectors plug into any metadata or pipeline tool and extract the metadata into the UMR. These connectors essentially parse the technical, business metadata and data mappings from the underlying systems (data sources, ETL tools, metadata generators and recorders) and hydrate the knowledge graph (UMR) to provide one cohesive place to search and view metadata from across the entire organization.

FIGURE 2: Screenshot of the UMR front end application shows data lineage for a given dashboard

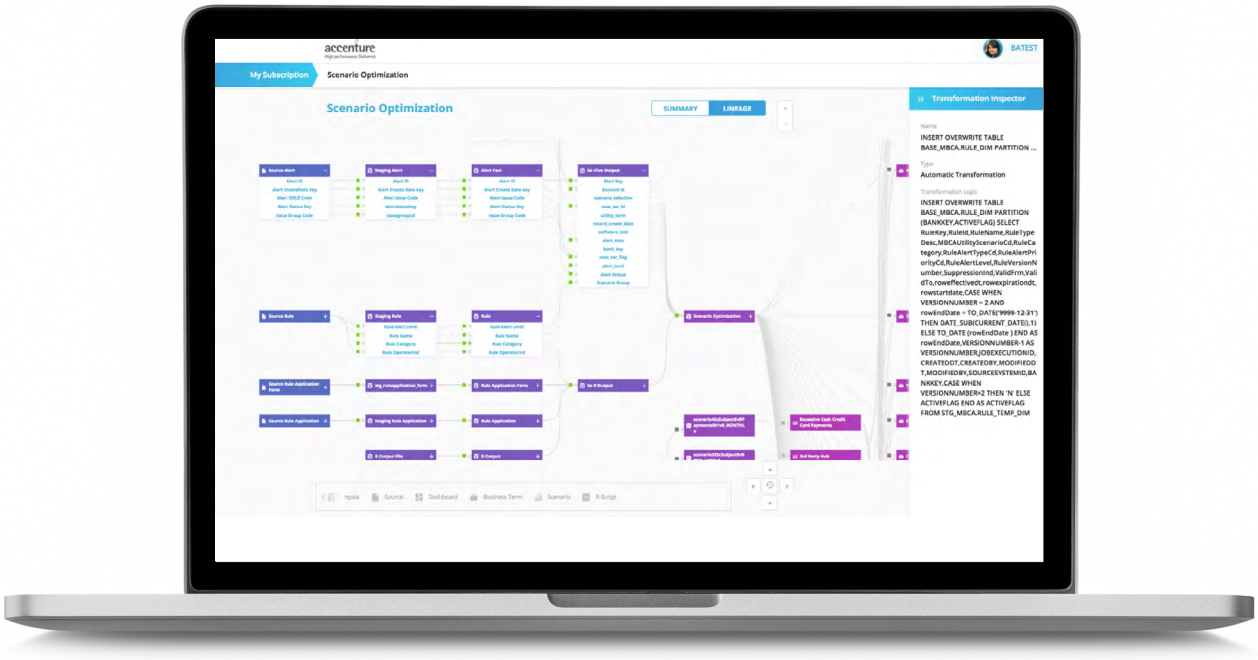
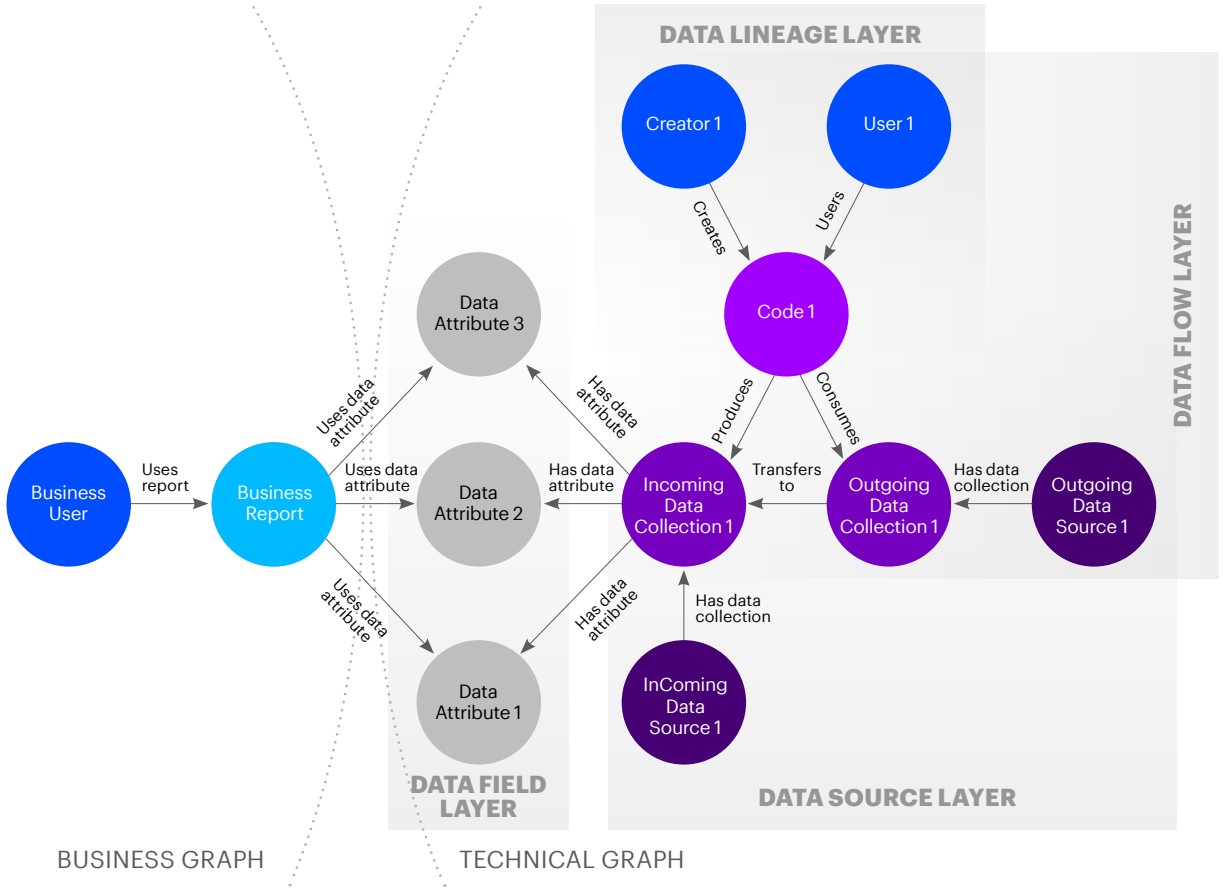


FIGURE 3: UMR Instance Graph – It shows how the entities are modeled and related to each other.



UMR Instance Graph

Let's make the UMR more concrete. Figure 3 shows an example of a UMR instance graph. The UMR defines the class or schema with definitions of what represents a node, what represents an edge and how the nodes are related to each other. An instance graph is a view of the schema populated with data at a given point in time. For example: the UMR Schema might define a node for a "User". In the instance graph, "Business User, "Creator 1" and "User 1" are instances of that node. In this graph, the circles represent the node or entities, and the directed edges map the relationships. Reading this graph left to right, we can start with a "Business User" who uses a "Business Report" that uses_data_attribute "Data Attribute 2" from an "Incoming Data Collection 1" that was produced by "Code 1" created by "Creator 1."

A simple traversal of the knowledge graph reveals a lot of answers:

"What Data was used to generate Business Report?"

"What is the most connected piece of data in this ecosystem",

"Who created a certain dataset"

"Which user is allowed to use a certain data or report"

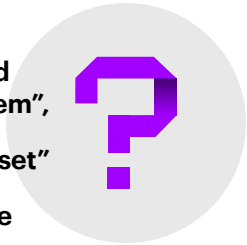
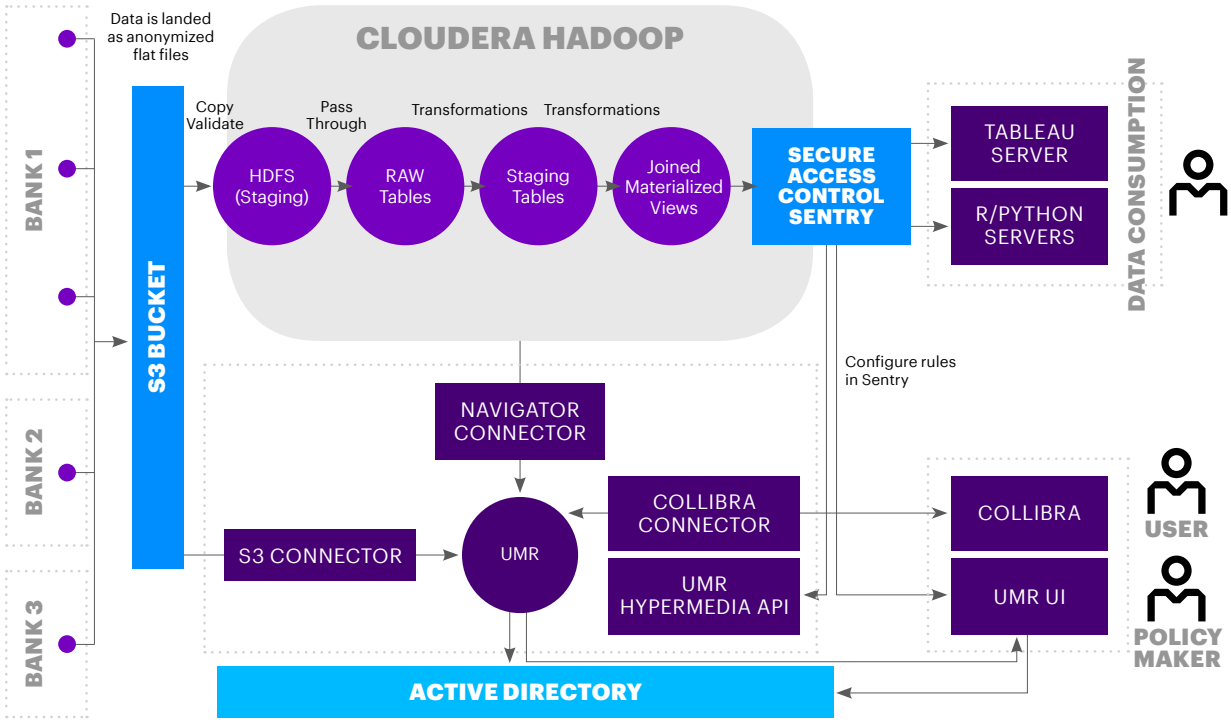


FIGURE 4: Reference implementation of UMR for an Anti-Money Laundering Ecosystem. The top flow with yellow arrows shows the data moving through the data supply chain process while the bottom grey arrows show how metadata is being captured and used along the way to provide a 360-degree view of the data.

Architecture Description



UMR in action: Anti-money Laundering

We have implemented the UMR for an anti-money laundering (AML) multi-tenant utility that allows member banks to get the most value from their data and analytics initiatives while sharing the associated cost. The UMR helps the consortium track their data across multiple systems and transactions in a programmatic and transparent way (See Figure 4).

The UMR for AML automatically stores technical metadata on data related to customers, accounts, transactions, transfers and alerts generated to signify if a given transaction or transfer is likely to be a money laundering activity.

It provides a data catalog with discovery capabilities to search for specific entities. You can search on specific data domains(tables), business terms (fields) or dashboards (reports).

It maps the technical metadata to the business metadata – business terms, their definitions, processes, owners etc. to provide business context to the physical data entities.

UMR shows end to end lineage: Starting with a dashboard, the business terms that are represented in the dashboard to the scenario/models powering this dashboard, the underlying data tables feeding these models all the way back to the original source files providing the data.

Bringing Data Together

The assertion that "every company is a data company" seemed radical when it was first made a decade ago. Today it seems obvious as companies compete on their ability to access the best data.

But there's a disconnect between what companies want from their data and their ability to deliver. Traditional data stores, data warehouses, and data lakes don't solve the underpinning problem that data is often held in separate tools, systems and locations. It is managed by separate groups. And that lack of connectedness makes it difficult to understand, trace and access the data throughout the data supply chain and across multiple systems. The UMR addresses these issues and much more. The Universal Metadata Repository reduces the friction in your data supply chain to unlock value in your data.

Contributors

SONALI PARTHASARATHY
Technology Research Principal,
Data Management Lead,
Accenture Labs
sonali.parthasarathy@accenture.com

TERESA TUNG
Managing Director, Global
Systems & Platforms R&D Lead,
Accenture Labs
teresa.tung@accenture.com

SABINA MUNNELLY
Senior Managing Director,
Accenture Innovation Center
for Data & Analytics Lead
sabina.munnelly@accenture.com

SANJAY KUMAR JOSHI
Senior Managing Director,
ATC Lead for Intelligent Data Group
sanjaykumar.joshi@accenture.com

About Accenture

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions — underpinned by the world's largest delivery network — Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With 477,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at www.accenture.com

About Accenture Labs

Accenture Labs incubates and prototypes new concepts through applied R&D projects that are expected to have a significant strategic impact on Accenture and its clients. Our dedicated team of technologists and researchers work with leaders across the company and business partners to invest in, incubate and deliver breakthrough ideas and solutions that help our clients create new sources of business advantage.

Accenture Labs is located in seven key research hubs around the world: San Francisco, CA; Washington, D.C.; Dublin, Ireland; Sophia Antipolis, France; Herzliya, Israel; Bangalore, India; and Shenzhen, China; and 25 Nano Labs. The Labs collaborates extensively with Accenture's network of nearly 400 innovation centers, studios and centers of excellence located in 92 cities and 35 countries globally to deliver cutting-edge research, insights and solutions to clients where they operate and live. For more information, please visit www.accenture.com/labs.