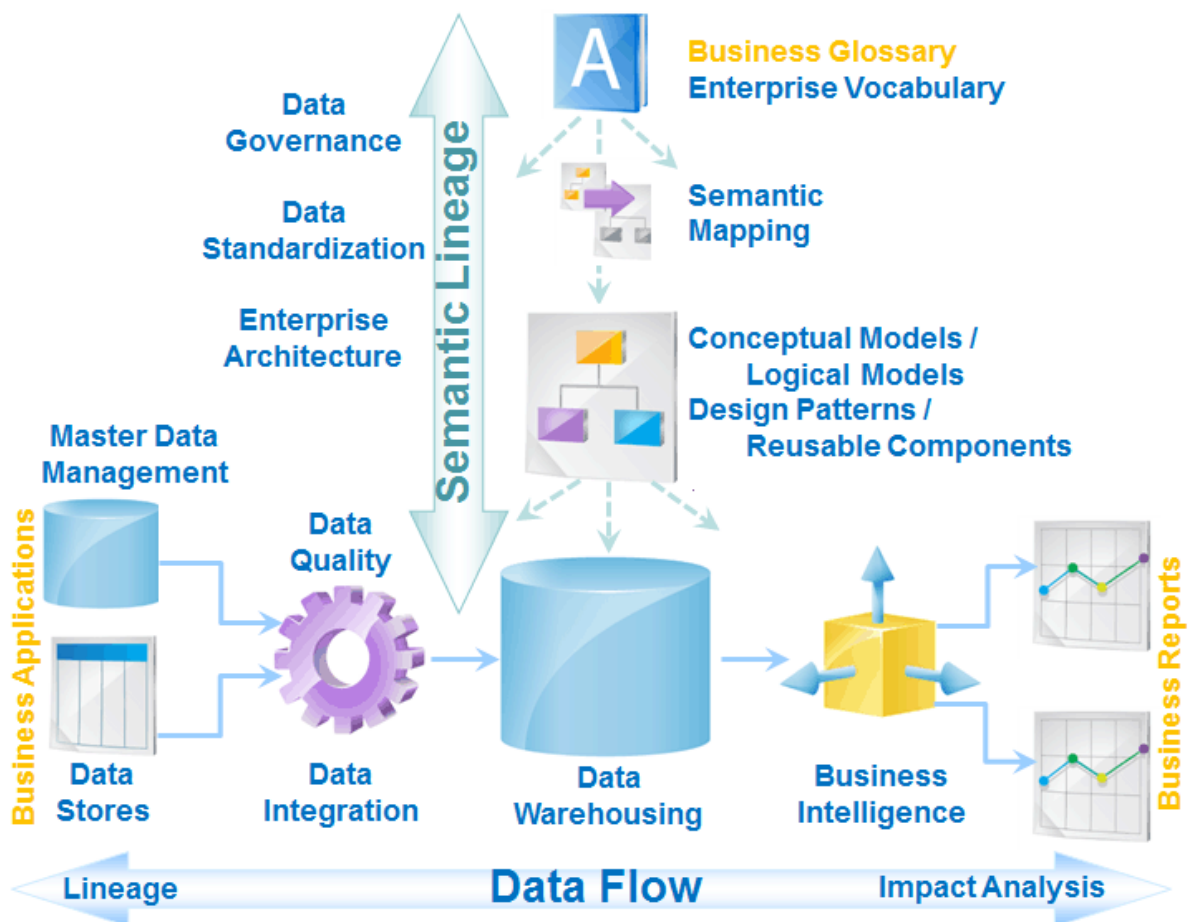


Qlik Talend Data Catalog Product Overview

Data Cataloging and Data Governance solutions must be built upon solid Metadata Management foundations: At the core of MM is a high scalability Metadata Version & Configuration Management system designed to support the continuous changes in the enterprise architecture with metadata harvesting (on prem & multi-cloud) and automatic stitching.

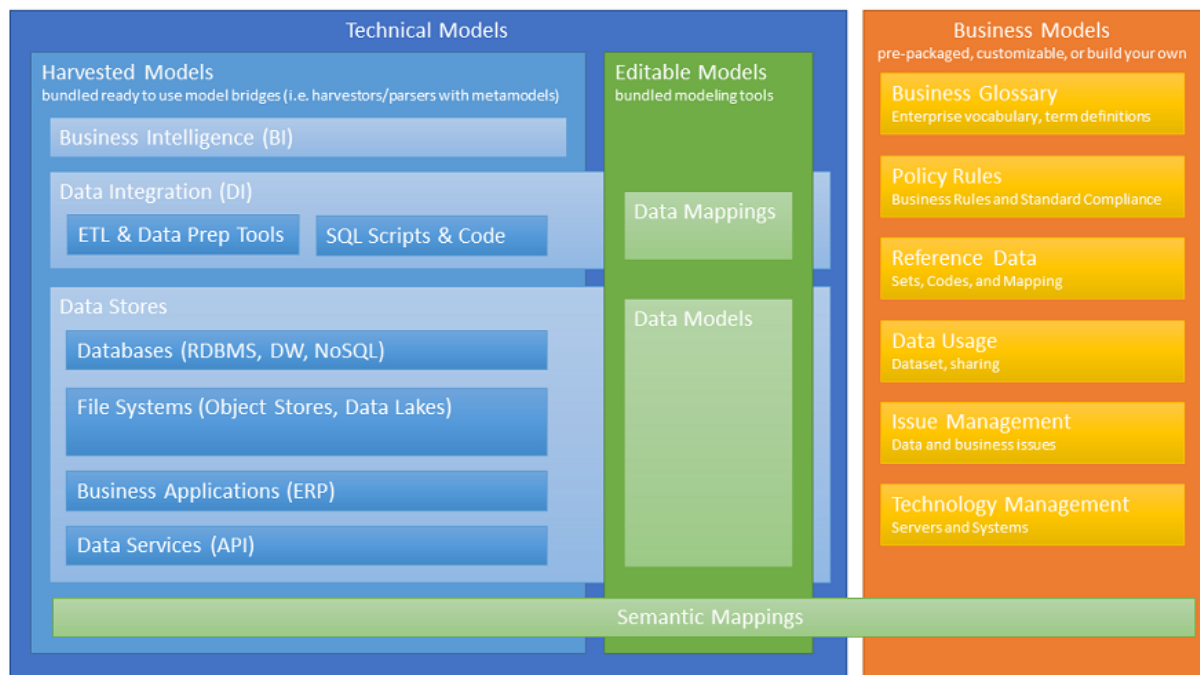
The Big Picture of Metadata Management for Data Governance

- with Business Glossary semantic traceability to the actual Enterprise Architecture,
- with data flow lineage & impact analysis, and support for DI & BI development Lifecycle Change Management.



Connecting Technical and Business Metadata for Data Governance

from DATA CATALOG with



Metadata Overview

Talend Data Catalog provides a comprehensive and well-integrated set of Metadata Management (MM), Data Cataloging (DC) and Data Governance (DG) solutions supporting on Considerations

premise, cloud based or hybrid Enterprise Architectures ranging from the classic Data Warehouse to the latest Data Lakes and Data Vaults:

- Integrated Solutions: Metadata Management, Data Cataloging, and Data Governance
- Multi-Deployments: On Premise, Cloud, or hybrid (Cloud with on premise harvesting)
- Multi-Architectures from the Data Warehouse to the new Data Lake / Data Vault
- Multi-Vendors from any Data Integration and Business Intelligence tools
- Multi-Technologies from files and SQL based RDBMS, to the new NoSQL, JSON, Avro, Parquet, XML and Hadoop big data technologies, REST API Data Services
- Multi-Storage File Systems: Data Cataloging by file crawling over Linux/Windows, Hadoop HDFS, Amazon S3, Azure Blob Storage, OpenStack Swift, Apache Kafka, etc.
- Multi-Configurations with Change Management and Incremental Metadata Harvesting, Comparison, Version and Configuration Management and automatic stitching

- Multiple integrated tools for Search, Data Flow and Semantic Lineage, Data Modeling, Data Mapping, Active Data Governance (generation of self-service DI and BI), Multi-Vendor BI Web Portal
- Data classification, both automatic and manual and both metadata and data detected.
- Fully customizable metamodel for custom modeling

Talend Data Catalog provides solutions for a full range of users:

- Most business end users in need of a multi-vendor Business Intelligence (BI) web Portal with quick access to the right report understand its content with proper business definitions from the enterprise glossary.
- Advanced business users and compliance officers looking for information traceability (data lineage) and data privacy for General Data Protection Regulation (GDPR), Sarbanes-Oxley (SOX) regulations, and more.
- Data stewards, data modelers, and data quality experts working on enterprise data standardization, common vocabulary, data modeling and business rules with powerful tools like the glossary, semantic mapper, data modeler.
- Data analysts in need of self-service data integration, preparation, and business intelligence using the data mapping design tool with active data governance (forward engineering) to their actual DI/BI tools.
- IT engineers, data scientists, data Integration and business intelligence developers looking complex multi-vendor and multi-architecture end to end lineage in great details down the design level information of each tool with support for change impact with full detailed version and configuration management.

Features

Metadata Harvesting

- Data Stores
 - Databases (Oracle, SQL Server, Teradata, IBM DB2, PostgreSQL, MySQL, AWS Redshift, GreenPlum, Netezza, SAP HANA, etc.)
 - Big Data (Hadoop Hive, HCatalog, Google Big query, etc.)
 - NoSQL (Cassandra, HBase, MarkLogic, MongoDB, etc.)
 - Flat Files (CSV, XLSX)
 - Hierarchical Files (JSON, Avro, Parquet, XML, XML XSD, etc.)
 - File Systems ((Linux/Windows)
 - Data Lake and Cloud Hadoop HDFS, Amazon S3, etc.)
 - Data Services (Open API, etc.)

- **Metadata Stores**
 - Data Modeling Tools (Erwin, ER/Studio, PowerDesigner, etc.)
 - Metadata Management (Atlas, Navigator, etc.)
 - Semantic Web Ontology (OWL/RDF)
- **Data Integration**
 - DI/ETL Scripts (Oracle PL/SQL, Teradata BTEQ/FastLoad/BulkLoad, Hadoop HiveQL, Sqoop, SAS code, etc.)
 - DI/ETL Tools (Informatica PowerCenter, IBM DataStage, Oracle ODI, Microsoft SSIS, SQP Data Services, SAS DI, Talend, etc.)
- **Business Intelligence** (SAP BusinessObjects, IBM Cognos, Microsoft SSAS/SSRS, Azure PowerBI, Oracle OBIEE, Microstrategy, Qlik, Tableau, ThoughtSpot, TIBCO Spotfire, etc.)
- **Business Applications** (SAP Business Suite, SAP Business Warehouse, Salesforce, etc.)

Metadata Management (MM)

- Configuration Manager (with automatic metadata stitching, and Enterprise Architecture diagramming)
- Metadata Search and Worksheets (metadata driven pre and post filters, semantic search language)
- Metadata Browser (hierarchical metadata browsers with custom metadata profiles per tool/technology)
- Metadata reporting capabilities where both search and browse end up to the same reporting page which is also directly available at WORKSHEETS > Manage. With a tabular and bulk editing capabilities
- Data Model Visualizer and modeler (fully editable ER Diagrams)
- Enterprise architecture diagramming and editing (fully editable architecture diagram)
- Data Flow Lineage and Impact Analyzer including data flow lineage and impact analysis down to the feature level, along with data vs control flow, data vs. semantic flow, highlight path and dynamic scoping
- Multi-Configuration Management (multi configurations for different enterprise architectures and groups)
- Multi-Version Management (efficient automatic incremental harvesting, with model history/SOX compliance)
- Metadata Comparator (comparison with previous versions for the impact of change)

Data Governance (DG)

- Glossary (with customizable workflow automation)

- Semantic Mapper (search driven, auto map, and multi-levels from glossaries to data stores via design models)
- Semantic Lineage Analyzer (term usage, and automatic glossary definition on data pass through)
- Local Documentation (quick in place editing of business names and descriptions while browsing harvested data stores)
- Glossary term classification (quick in place semantic linking while browsing harvested data stores, DI jobs, and BI reports)
- Data Tagging (applying reusable Labels available in search)
- Comments and Reviews (collecting business user feedback and managing reviews)

Data Cataloging (DC)

- File System Crawling (file type auto-detection, partitioning auto-detection)
- Data Profiling (from data sampling to full data profiling with statistical results)
- Semantic Discovery (data classes, patterns/lists machine learning)
- Relationship Discovery (data-detected matching data classes and metadata-detected inferred from usage in DI, BI, SQL, etc.)
- Social Curation (endorsement, warnings, certifications with impact on search)
- Data Modeling
- Data Store Documenter (automatic reverse engineering of naming standards with supervised machine learning)
- Data Store Modeler (with full data model diagram editing)
- Data Store Designer (new data store specifications and design)
- Data classification based on both data-detected and metadata-detected classification

Data Mapping

- Data Mapper (from business user data mapping specifications to design for bulk and feature/SQL with joins/filters/lookups)
- Metadata Applications
- BI Web Portal (Multi-vendor BI Web Portal with bi-directional integration, and glossary generation)

Active Data Governance (Forward Engineering)

- Data Modeling Tools (Erwin, ER/Studio, PowerDesigner, etc.)
- Data Integration Scripts (PL/SQL, BTEQ, HiveQL, etc.)

- Data Integration (to self-service / data prep tools)
- Business Intelligence (to self-service like Tableau or design layers like BO Universes.)

Administration, Customization, & Extensions

- Custom Attributes (metamodel extensions) (MyCompanyCertificationLevel, etc.)
- Customizable UI (menus, widget layout, etc.)
- REST API (glossary lookups, lineage trace, automatic harvesting, search, browse, update, etc.)

Object Watchers and Email Notification

- Provides watcher capabilities at the server level allowing user to effectively subscribe to objects and be notified of changes
- There is an adjustable frequency of notification from near real-time to daily.

Explore

Fundamentally, Talend Data Catalog is a metadata management tool. Thus, a key aspect of the feature of Talend Data Catalog centers around looking up, finding or discovering metadata that is harvested into the system.

In particular, you may use the search (both immediately as you type as well as a more advanced search page) by contained text. In addition, you may browse directly by metadata category and type. You may create worksheets to report on the metadata. Finally, you may locate elements by navigating the current configuration architecture.

Qlik Catalog

Qlik Catalog overview

Qlik Catalog is a modern enterprise data management solution that simplifies and speeds up how you catalog, manage, prepare, and deliver trustworthy, actionable data to business users across the enterprise.

Qlik Catalog stores technical and business metadata in a central catalog. From this catalog, data is made available to the user community through a user interface similar to a shopping experience on an e-commerce site. Users can search, browse, preview, understand, compare, and find appropriate datasets from the marketplace collection and place them in a shopping cart. From there users can publish data to an analytics tool or an external environment, on-premises or in the cloud, becoming available for other applications to leverage.

Architecture and data flow

Qlik Catalog empowers enterprises and users by instilling confidence in data. This is done by combining automated data ingestion, validation and profiling with enterprise-level transformation capabilities all from a centralized interface. Qlik Catalog maintains visibility throughout the data lifecycle—for example, tracking detailed ingest statistics and custom-built data transformations through lineage. The application has comprehensive and sophisticated data ingestion capabilities that support a variety of sources, including legacy structures such as Mainframe COBOL copybook and file formats such as XML and JSON. Datasets are stored in industry-standard formats, calculating profile statistics for every field and creating samples for each dataset. Data is also made available for querying through standard SQL engines. Depending on the underlying platform this can be Hive, Impala, or PostgreSQL.

The solution validates and profiles ingested records as good, bad, or ugly. An ugly record is one where some of the field data is problematic – it might be a string with control characters in it, mismatched data types, or non-numeric characters found in an integer field. A bad record does not conform to the specified record layout. For example, the record may have the incorrect number of columns, wrong or absent delimiters, or an embedded control character is causing a record break. Qlik Catalog ingest technology assures data validity and integrity by preventing bad and ugly records from presenting as good data.

Qlik Catalog enriches and stores technical and business metadata in a central catalog. From this logically ordered and secure environment, enterprise data is made available to the user community through a friendly interface that mimics shopping experience on an e-commerce site. Users can search, browse, preview, understand, compare, and find the most appropriate datasets from the overall marketplace collection and place them in a shopping cart. From there they can take action such as publishing data to their favorite analytic tool or delivering it to an external environment on-premises or in the cloud where it will be available for other applications to leverage.

Qlik Catalog offers two deployment options. Both are available on premises or in the cloud.

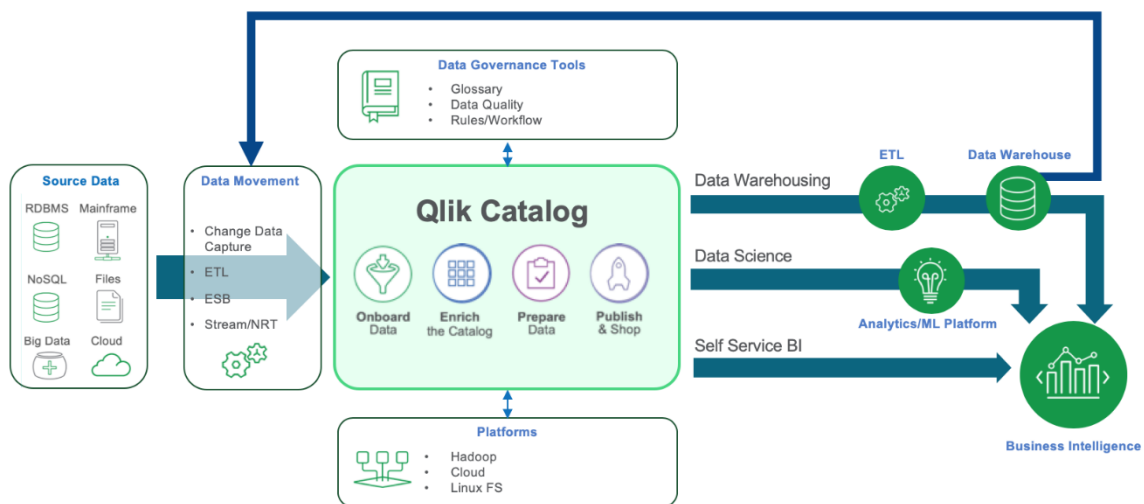
- On an edge node of a Hadoop cluster. In this configuration it stores data in HDFS (Hadoop Distributed File System) or if deployed in the cloud, in cloud-native storage such as AWS S3. Data is exposed for querying via Hive or Impala query engines and processed on the cluster using standard parallel processing frameworks such as MapReduce, Tez, and Spark.

This deployment option fully leverages the scalability of the underlying parallel infrastructure and is suitable for high volumes of data and heavy user workloads.

- On a single Linux server. In this configuration data is stored as files in the Linux file system and exposed for querying through PostgreSQL, leveraging the feature known as foreign tables.

This deployment option is suitable for more modest data volumes and workloads. This option can be especially attractive to organizations that do not have a need for higher data volumes or the expertise to manage a Hadoop environment. This deployment option provides the architectural foundation for the streamlined Qlik Catalog for QVDs offering focused on making QVDs easier to discover, understand, and consume.

Architecture and workflow



Module overview

The following functional modules provide a powerful and flexible range of data preparation, governance, and management functionality.

Catalog - Interactive marketplace dashboard

Catalog is an interactive marketplace dashboard providing immediate insight and actions for search and filterable entities and QVDs across the data ecosystem. It scores entities against operational, quality, and popularity or size KPIs providing immediate insight and actions for

filterable entities across the data ecosystem. Users access meaningful information such as lineage, joinable entities, sample data, creation date, source, connection, KPI scoring detail, and field profile statistics with one click. Shop for data functionality provides search, browse, preview, and explore features targeting specific entity collections.

Source - Data ingest, validation, quality, and profiling

Source enables administrator level users to deploy and manipulate enterprise data assets across clustered nodes with ease and efficiency through the powerful ingest framework of Qlik Catalog. Data analysts build custom environments to validate, profile, and register metadata in HCatalog, ensuring clean, executable datasets upon ingest of data records. Source provides guided wizards to build metadata-driven environments leveraging JDBC technology for relational databases, File Definition Language (FDL) for flat files, COBOL copybooks for mainframe data. An XML utility stages users through ingest and validation of XSDs for loading XML data, and JSON files validate with a hierarchical extraction utility. The data sourcing capabilities of Qlik Catalog can be configured to be scheduled for automatic execution in production environments through API scripting / integration.

Discover - Interactive query and metadata management

Discover enables analysts to view and explore all data sources and metadata to which they are granted access. Users can search, browse, create, delete, and edit properties at source, entity, and field levels, enabling collaborative data curation and governance. Datasets are logical collections of data assets in Qlik Catalog that can be created by any user to be transformed, published, or selected and retrieved through query language. Users can run most Hive or Impala commands from within Qlik Catalog or create custom views using a familiar shopping cart environment for visibility and control over data sources.

Prepare - Data transformation

Prepare is a simple and intuitive GUI that enables users to customize database tables into filtered and cleansed data sets through transform controllers: transform, filter, join, aggregate, router, sort, union, change data capture (CDC). The metadata and profiling statistics of Qlik Catalog guide users throughout the process, including end-to-end validation of the dataflow. Users can envision and graphically manipulate custom datasets to meet a wide variety of targeted data preparation and analytics requirements. Users can build and apply Pig functions to entities through the custom expression builder .

Publish - Deliver datasets

Publish enables one-time or recurring replication of datasets across enterprise cluster environments. Administrators define Publish targets, file format and execute or schedule on a one time or recurrent basis. Datasets can be published to various destination types including Hadoop (and Hive), local file, HDFS, FTP, Amazon Web Service (AWS) Simple Storage Service (S3), or any RDBMS supporting any protocol via Qlik Catalog Open Connector scripting. Publish provides the ability to define file type, field and record delimiters, header information, partition merge options, data obfuscation techniques, and environmental properties for openconnector.

Security - User and group administration

Security provides a console to administer role-based access permissions to users and assigns both users and data to groups. Administrators can create users and designate group access levels and permissions to entities. Qlik Catalog leverages enterprise security technologies such as Active Directory identity services and domain management and LDAP to query and dynamically synchronize active groups and personnel. Qlik Catalog can also integrate with access control policies defined in Hadoop tools such as Ranger, Sentry, HDFS, and HDFS with transparent data encryption. Support for impersonation and enforcement of HDFS access control methods enable incisive control and integration of existing security policies.

Ops

Ops provides a quick view of data quality and usage metrics for the most recent ingest activity. Operations display for recent data loads with standard associated metadata including Source, Entity, Job Status, Delivery Time, Start Time, End Time, Record Count, Good Records, Bad Records, Ugly Records, Filtered Records, and Last Update Time.