Qlik Data Catalyst

Technical Overview

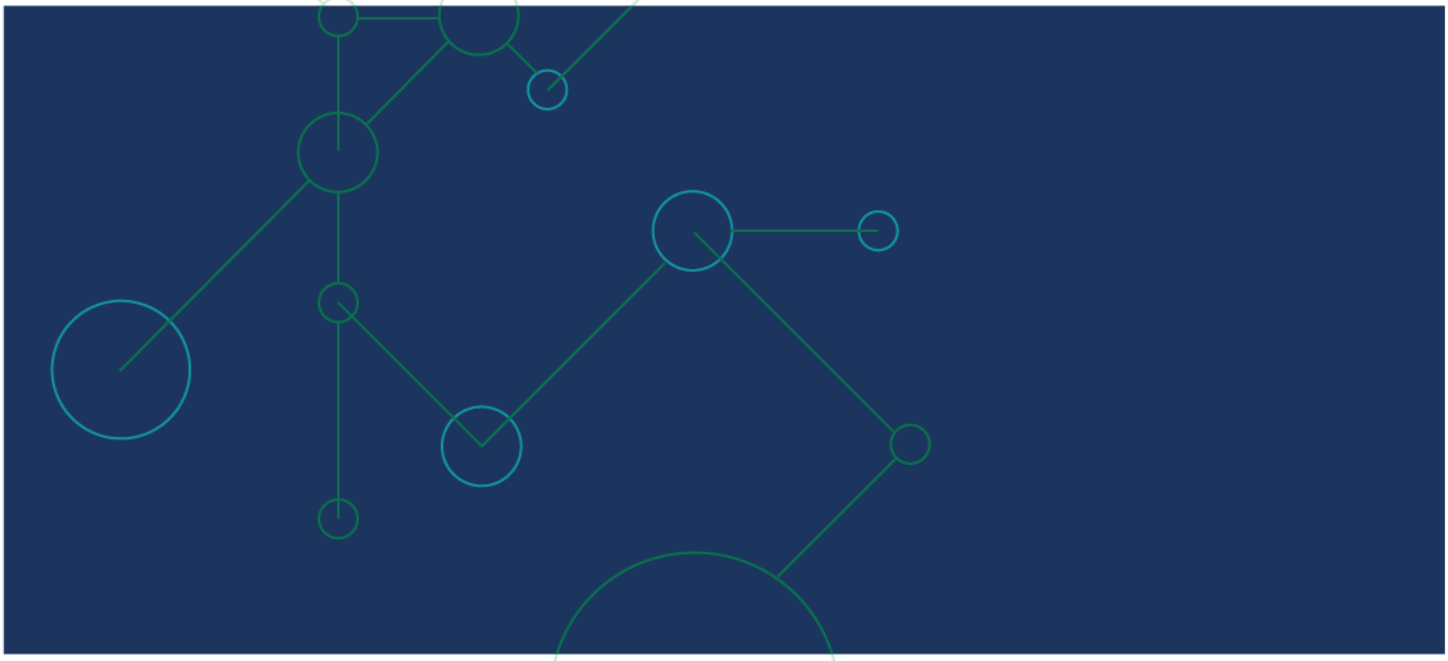


TABLE OF CONTENTS

Introduction to Qlik Data Catalyst™ 3
- Highlights 3

A Technical Architectural View **5**

Qlik Data Catalyst User Interface and Modules **5**
- Qlik Data Catalyst's Integrated Modules Share a Common Framework 5
- Source Module: Data Ingest, Validation, Quality, and Profiling 7
- Catalog and Discover Modules: Explore the Data Collection: Preview and Shop for Data 7
- Prepare Module: Data Transformation 7
- Security Module: User and Group Administration 8
- Publish Module: Deliver Datasets 8
- Reports/Dashboard 8

The Fundamental Role of Metadata **8**
- Metadata: A First-Class Citizen in the Marketplace 10
- Qlik Data Catalyst Captures Business Insights from Data Analysts and Through Automation 10
- Qlik Data Catalyst's Metadata Management Process Is Automatic and Open 10

Qlik Data Catalyst Builds a Data Layer in Hadoop That Is Persistent, Managed, and Secure — The Data Marketplace **11**

Data Conductor: Control Across the Enterprise Data Ecosystem **13**

Qlik Data Catalyst Security and Data Governance **15**
- Qlik Data Catalyst Takes a Platform Approach to Data Marketplace Security and Data Governance 15
- Authentication 16
- Authorization 16
- Accounting 16
- Qlik Data Catalyst, Hive, and HDFS: Managing User Permissions Across the Hadoop Stack 16
- Impersonation 17
- Encryption 17
- Data Masking 17

Qlik Data Catalyst is a Native Hadoop Application **18**

How Qlik Data Catalyst Interacts with the Hadoop Cluster **18**

Qlik Data Catalyst Deployment Options Support Enterprise Ready **19**

Integrating Qlik Data Catalyst with Other Applications in Your Enterprise IT Environment **19**

Conclusion **20**

INTRODUCTION TO DATA CATALYST

Qlik Data Catalyst is a modern enterprise data management solution that simplifies and speeds up how you catalog, manage, prepare, and deliver your trustworthy, actionable data to business users across your enterprise. Qlik Data Catalyst builds a secure, enterprise-scale repository of all the data your business has available for analytics, giving your data consumers a single, go-to catalog to find, understand, and gain insights from any underlying enterprise data source. The solution's data preparation and metadata tools streamline the transformation of raw data into analytics-ready assets, while the product's Smart Data Catalog and graphical user interface (GUI) help your users easily discover and select whatever data they need. Built on a platform of hardened data security and featuring governance capabilities, you can easily integrate Qlik Data Catalyst with any of your other data management tools to gain enterprise-grade scalability, reliability, and performance.

This document provides a technical overview of the Qlik Data Catalyst solution. It is designed to help teams considering deploying a data marketplace — including your users in IT, data governance, analytics, and business communities — to understand the Qlik Data Catalyst architecture, functionality, and performance characteristics.

HIGHLIGHTS

Qlik Data Catalyst is a fully integrated solution to onboard, catalog, prepare, and deliver enterprise data for a wide range of applications, including the following:

- Agile Analytics
- Enterprise Data On-Demand
- Migration from ETL to Modern Data Preparation and Delivery
- Data Governance Collaboration
- Mainframe Data for Modern Analytics

The solution gives your users access to all the functionality they need to deploy and maintain enterprise data, and the consolidated, scalable management platform can replace or augment a costly mix of scattered, redundant databases and data flows.

By giving your users self-service, on-demand access to data, Qlik Data Catalyst dramatically expands users' ability to use and share information and analytics to drive business decisions.

Qlik Data Catalyst powers the new era of governance. As data consumers increasingly adopt self-service models, they become experts. Qlik Data Catalyst engages your data analysts, business users, and other data consumers with interactive metadata at the point of use, encouraging the crowdsourced curation of data as an enterprise asset.

In Qlik Data Catalyst, metadata drives the entire data marketplace process. The Qlik Data Catalyst Smart Data Catalog integrates metadata from source systems, data validation and profiling, data preparation, business user tags and comments, security and user access tracking, and data load and publishing logs. The resulting robust layer of technical, operational, and business metadata enables self-service access to the data and stronger governance while exposing a rich new set of actionable data insights.

Qlik Data Catalyst can access data on any platform, giving your users full control over how much profiling is applied to data assets across an enterprise data ecosystem.

The solution is a fully native, modern data management solution leveraging the performance, scalability, and economic advantages of data storage and compute platforms such as Hadoop, Spark, Amazon Web Services (AWS), Google Cloud Platform, and Microsoft Azure. Qlik Data Catalyst makes the promise of increased agility and scale associated with these platforms a reality by combining their core functionality with the full suite of capabilities required to securely deploy a production-grade, enterprise-wide data marketplace solution.

Qlik Data Catalyst is built for the enterprise. It includes the features essential for your organization to securely manage data at scale — including high availability, intelligent data ingestion of legacy sources, strong data encryption and data masking, access controls, and technical and business metadata that give your business users context and meaning to understand and leverage data assets more effectively.

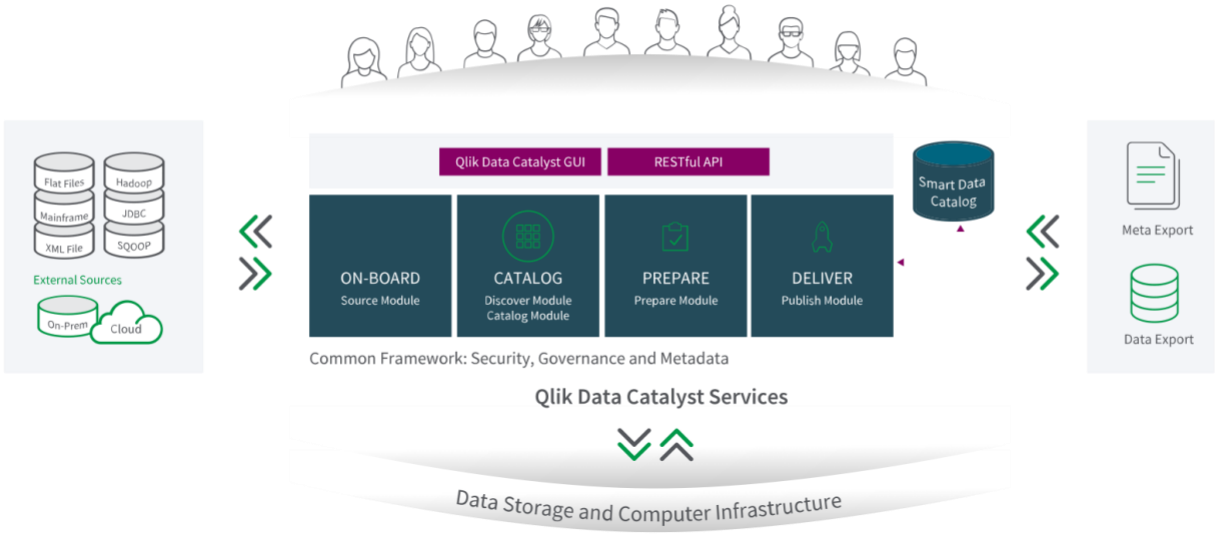
A Technical Architectural View

Qlik Data Catalyst runs on top of modern data storage and computing platforms including Hadoop, Spark, AWS, Google Cloud Platform, and Microsoft Azure. While leveraging these platforms for their raw power, Qlik Data Catalyst adds layers of functionality required to on-board, discover, prepare, and deliver business-ready enterprise data to your data consumers. Qlik Data Catalyst includes a common framework of services to support security, data governance, and metadata management. These services enable a robust data-as-a-service platform.

1. The Qlik Data Catalyst GUI is a browser-based and intuitive application that allows your IT staff and business users to drive the data marketplace process and access all of the platform's features and functions. GUI actions direct activity in the other two parts of Qlik Data Catalyst: the Smart Data Catalog and Qlik Data Catalyst Services.

2. The Qlik Data Catalyst Smart Data Catalog is a relational database that manages and maintains all of the metadata collected and generated along every step of the data delivery process. It can securely exchange metadata with other applications and data catalogs.

3. Qlik Data Catalyst Services take action on data, such as ingestion (which includes automated validation, profiling, and history management), metadata creation and management, and data preparation. Actions taken on data are executed on the underlying data storage and compute layer, making the Qlik Data Catalyst Services extremely lightweight.



Qlik Data Catalyst User Interface and Modules

Your users interact with Qlik Data Catalyst modules through the product's intuitive GUI. Multiple modules (for example, Source, Catalog, Discover, Prepare, and Publish) provide capabilities to manage each process from on-boarding and cataloging to preparing and delivering business-ready enterprise data. Other modules, specifically Report and Security, provide services that help users administer and monitor activity in the marketplace.

Qlik Data Catalyst Integrated Modules Share a Common Framework

The Qlik Data Catalyst solution is made up of the set of integrated modules previously referenced that work in a common framework with data management, metadata, security, and governance capabilities. All modules together leverage and follow consistent policies, such as user access privileges, data encryption, and file-naming conventions. This ensures your users have an optimal experience and your administrators can easily maintain the environment. It also means the data marketplace is more secure because a single, consistently applied set of security measures is implemented seamlessly across your entire environment. No fault lines or potential failure points exist in security between different applications or point products.

SOURCE	CATALOG / DISCOVER	PREPARE	PUBLISH
• Onboard data and metadata from all sources and formats • Standardize record formats, data types and character sets • Convert mainframe, JSON and XML data to a queryable format • Validate records against expected format/types/values • Profile new fields of data, generate a statistical profile • Apply post-processing rules, Find/act on insights in data	• Browse the data collection • Define custom search criteria, save custom views/result sets • Save and share custom views • Review data scores - quality, operational, and popularity • Look at technical and business metadata for each data item • Preview sample data • Explore particular data items • Save data for subsequent use	• Create custom data preparation flow to generate new datasets • Drag and drop canvas; no coding needed • Join, route, filter, aggregate, sort, union and more • Create custom data transformation • One time or repeating. Immediate or scheduled. • Combine data in various states of refinement, from raw to ready	• Publish data for consumption • One time or repeating, Immediate or scheduled • Secure and protect data for downstream use **REPORT** • Track user activity, data volumes, and cluster health • Pre-defined reports: operational, governance, and security • Build new reports, leveraging metadata repository
SECURITY			
• Synchronize users and groups with Active Directory or LDAP • Leverage Kerberos security	• Honor cluster permissions • Obfuscate/mask confidential data	• Audit access and lineage • Support impersonation	• Control users' access to data

Source Module: Data Ingest, Validation, Quality, and Profiling

Source lets your users onboard, convert, load, validate, and profile data quickly and easily through a Qlik Data Catalyst powerful ingestion framework. The process produces clean datasets, with tracked history, stored in the underlying data storage platform, managed by Qlik Data Catalyst, and ready to query.

Source also provides guided wizards to build automated processes for ingesting relational data, mainframe data sources, JSON and XML files, and flat files. Metadata defining the source format—such as database schemas, COBOL copybooks, or XSD files—can be imported directly from the Source when available. When such metadata is not available in the source format, it can be created with automated assistance from Qlik Data Catalyst. This metadata is used to convert source data to standardized character sets and formats optimized for analytics as it is loaded. Hierarchical data (such as mainframe files, XML, and JSON) is automatically turned into tables that can subsequently be queried in SQL or any business intelligence (BI) tool.

As part of the onboarding process, the Qlik Data Catalyst validation process sorts incoming records into good, bad, and ugly groups, based on the extent to which each record complies with the expected record format and data values and types. The record count for each group and messages documenting the anomalies of ugly and bad records are available to users following each data onboarding job. The solution also generates a statistical profile for every field of data entering the marketplace. This statistical data is added into the Qlik Data Catalyst Smart Data Catalog, providing your users with detailed information describing the exact content and character of each field of data.

Qlik Data Catalyst allows your users to define and apply custom business rules against new data as it is onboarded. Leveraging validation and profiling metadata, Qlik Data Catalyst post-processing rules provide your users with a powerful, flexible tool to automatically find, expose, and take action on new insights hidden in the data itself. For example, rules can be created to identify and protect personally identifiable information (PII), flag potential duplicate data, or highlight significant changes in data quality or load job outcomes.

Qlik Data Catalyst manages history of ingested data, automatically creating partitions and incremental snapshots. You can synchronize history and update information with other metadata repositories, such as HCatalog, Atlas, or Navigator. The Qlik Data Catalyst data sourcing processes can be scheduled for automatic execution for lights-out operation. They run natively on the underlying high-performance, parallel big data platforms, allowing for faster data on-boarding and scalability even as data volumes grow.

Catalog and Discover Modules: Explore the Data Collection. Preview and Shop for Data

Qlik Data Catalyst builds a Smart Data Catalog that documents every aspect of the data and data management process. This information is presented through the Qlik Data Catalyst catalog and discover modules.

Catalog gives your users an Amazon-like shopping experience where they can search, browse, preview, understand, compare, and find the most appropriate entities from the overall marketplace collection. Search tools allow your users to apply filters, to zero in on the entities that match their criteria. KPIs grade each entity, so the user can understand the quality of the data, its operational status, and popularity among users. Preview options allow your users to see sample data or quickly review more detailed metadata regarding individual entities. Data items added into the shopping cart (entities or entire sources) can go directly into data preparation or publishing jobs, be saved as datasets for later use and shared with other users. Collaboration tools allow your users to share insights through user reviews or tags and share data collections and results.

The Qlik Data Catalyst Discover module provides a hierarchical view of data sources, entities, and fields. Discover enables your users to view and manage associated metadata to all data sources and entities to which they have been granted access, including custom tags. Users comfortable writing their own SQL can run queries directly from within Qlik Data Catalyst and create custom views of the data that can be shared with other users.

Prepare Module: Data Transformation

The Qlik Data Catalyst Prepare module offers a simple and intuitive environment that lets your users create powerful transformations to turn raw data sources into business-ready data. Prepare allows your users to create a dataflow by connecting operators from a palette, including Transform, Filter, Join, Aggregate, Sort, Union, Change Data Capture, and Route. Your users can also create custom operators and include them in the dataflow. Graphical Prepare dataflows are translated into native jobs that execute on the cluster. Qlik Data Catalyst metadata and profiling statistics guide your users throughout the process, including end-to-end validation of the dataflow. The dataflow can be tested interactively as it is developed and saved for automated, scheduled execution.

Security Module: User and Group Administration

Qlik Data Catalyst provides a console to administer role-based access permissions to your users, and assigns both users and data to groups. This ensures data protection on a shared platform; for example, if your marketing and HR data items are assigned to different groups, members of one group cannot access the other's. The Security module allows an administrator to create users and designate group

access levels/permissions to data sources, entities, and Qlik Data Catalyst modules. The platform can synchronize user and group information with Active Directory or LDAP. It can also integrate with access control policies defined in tools such as Ranger, Sentry, and HDFS.

Publish Module: Deliver Datasets

The Qlik Data Catalyst self-service Publish module enables one-time and scheduled export of data in configurable formats for consumption in other environments in compliance with user permissions and security requirements. Datasets are logical collections of data objects that can be replicated to other destinations and formats including Hive, HDFS (including ORC and Parqet), FTP/FTPS, AWS/S3, Azure ADLS/WASB, local files, or any other protocol via Qlik Data Catalyst Open Connector scripting. Publish gives you the ability to define the file type, field delimiters, record delimiters, header information, partition merge options, data obfuscation techniques, and environmental properties for Open Connector.

Reports/Dashboard

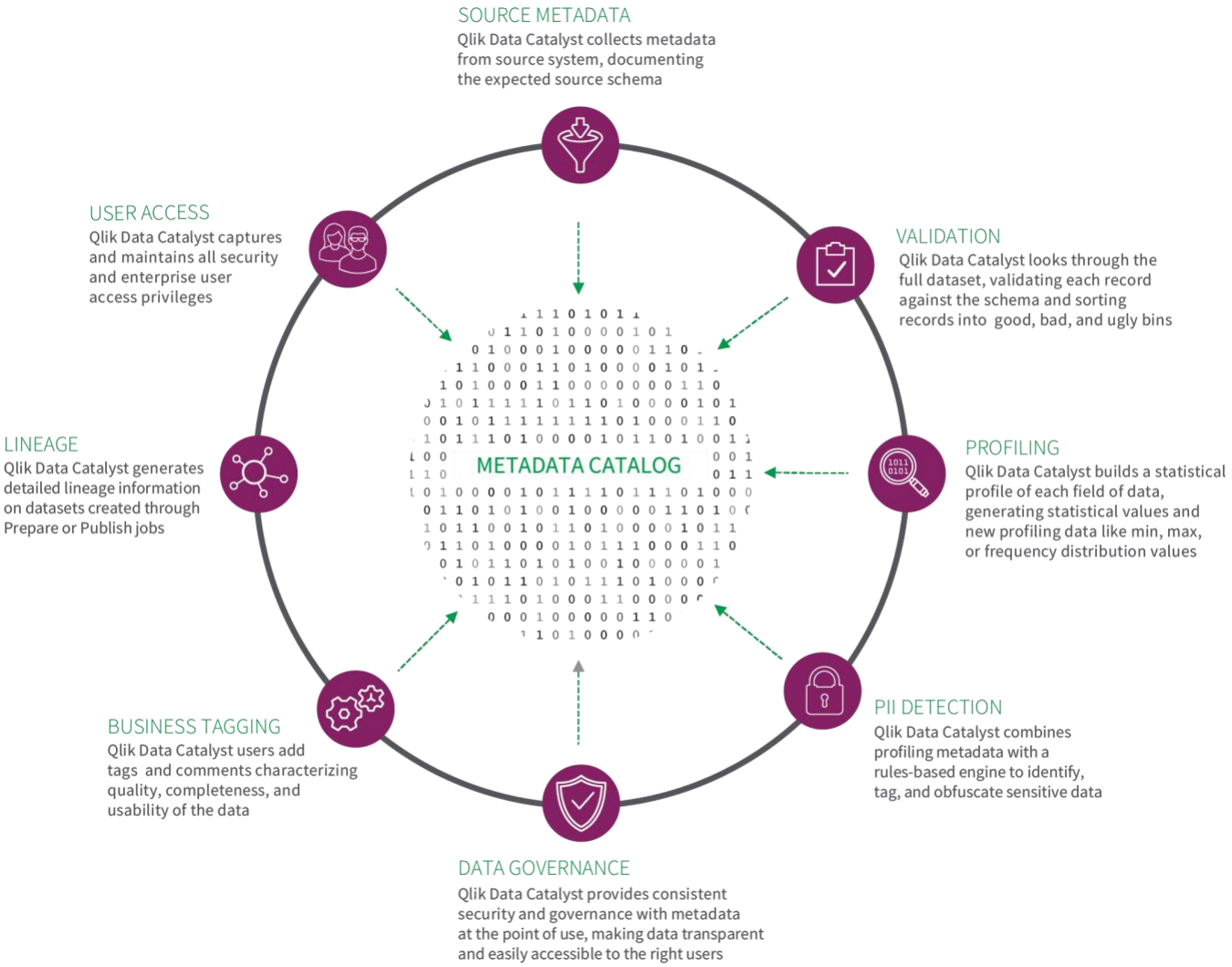
Qlik Data Catalyst includes reporting capabilities that enable your users to track data volumes and monitor cluster health and user activity. Examples include Trended Record Volumes (Operational), Timeliness and Quality (Governance), and Access by User and Group (Security). Reports are generated as bar, line, or grid charts that can be exported.

In addition to its out-of-the-box reports, Qlik Data Catalyst exposes all of its metadata by providing open access to its relational repository through the use of reporting views. Your users can leverage them to create custom reports using their BI tool of choice. These views will persist even though the underlying metadata schemas may change with product releases, ensuring long-term compatibility.

The Fundamental Role of Metadata

Metadata plays a central role in Qlik Data Catalyst, far beyond metadata's traditional role of documenting data about data with simple cataloging or classification. In Qlik Data Catalyst, the Smart Data Catalog drives the data marketplace management process and enables many of the Qlik Data Catalyst key capabilities. Starting with data sourcing, metadata is collected to document the expected source schema. Next, Qlik Data Catalyst validates every incoming record against the expected format. Build out of the Smart Data Catalog continues with the generation of profile statistics of each field. This continuous process sets the foundation for your users to generate valuable insights from the secure, accurate, and understood data across the marketplace.

As your data stewards, analysts, and business users work with the data, new metadata is created and added into the Smart Data Catalog. Tags, definitions, and metrics generated both manually and automatically help characterize the quality, completeness, business meaning, and usability of each data item. Data generated through Prepare or Publish jobs is linked to ancestors and descendants, providing detailed lineage information. Security and user access privilege details are also captured and kept in the catalog.



Metadata: A First-Class Citizen in the Data Marketplace

In the Qlik Data Catalyst platform, metadata and the data it describes are managed together as a single unit. Every action or new piece of information associated with a data entity is immediately recorded and

linked with that data in the Smart Data Catalog. When data is processed, the output inherits all relevant metadata associated with the input.

Data and metadata are always active, always synchronized, and always right. In this sense, metadata in Qlik Data Catalyst is elevated to a status as an equal partner to data itself in the marketplace.

This tight coupling of data with metadata throughout each step of the process is what allows Qlik Data Catalyst to provide your organization with a data marketplace that is structured, well organized, and ready to use. Based on this foundation, the data marketplace can be built and maintained by data analysts rather than deeply technical IT staff. The tight coupling also enables your users to directly access data in the data marketplace on a self-service and on-demand basis without help from IT.

Another important benefit of an open, metadata-driven environment is the ease of automation and integration into other enterprise tools and databases. Because all Qlik Data Catalyst objects and activities are represented as metadata, schedulers, applications, and repositories can synchronize with and execute Qlik Data Catalyst jobs via a standard API. This enables a Qlik Data Catalyst-managed data marketplace to be well integrated with existing enterprise tools and operational flows.

Qlik Data Catalyst Captures Business Insights from Data Analysts and Through Automation

Qlik Data Catalyst also collects a rich layer of business metadata by allowing your users working in the Qlik Data Catalyst GUI to crowdsource and share business names, and definitions, blogs, and tags associated with different data entities. The more analysts work with the data, the better documented it becomes, replacing error-prone handoffs between the business and IT with a self-service, collaborative process for handling shared data.

The solution leverages metadata from the robust and rich data profiling and validation to create additional business value. Using pattern matching technology, Qlik Data Catalyst can automatically identify and take necessary actions to appropriately describe, tag, and secure the information. Specific data categories such as PII and other sensitive data classifications seen in the Payment Card Industry Data Security Standard (PCI DSS) are easily identifiable through the Qlik Data Catalyst continuously-enriched data profiles and Smart Data Catalog.

Qlik Data Catalyst Metadata Management Process is Automatic and Open

The Qlik Data Catalyst Smart Data Catalog resides in a standard relational database. It can be easily shared with other enterprise metadata management platforms and catalogs. The data model for the Smart Data Catalog is documented, and can be accessed via the Qlik Data Catalyst open RESTful APIs as well as via the metadata import/export function.

Qlik Data Catalyst Builds a Data Layer That Is Persistent, Managed, and Secure — The Data Marketplace

With Qlik Data Catalyst, your organization builds a persistent data layer—a collection of data in the underlying data storage layer that is maintained, managed, and enhanced over time. This serves as an interactive data marketplace in your enterprise and provides a vital link between data providers (e.g., applications, third-party data feeds) on the one hand and data consumers (e.g., analytic applications, data warehouses, or data analysts) on the other. This is fundamentally different from extraction, transformation, and load (ETL) tools, whose goal is to move data from a source to a target while cleaning and transforming data en route, without creating a persistent set of data along the way.

When you are using Hadoop, AWS, Google Cloud Platform, or Microsoft Azure to store data in the marketplace, Qlik Data Catalyst adds a robust set of data management, security, governance, and metadata as required by any enterprise-ready data management platform. This ensures that the data stored in your persistent data layer forms a data marketplace, not a data swamp.

The inclusion of these enterprise data management capabilities is central to the ability of Qlik Data Catalyst to give your users self-service, on-demand access to data in the marketplace; all the data is always ready and understood. As illustrated in the diagram, by maintaining data in the data marketplace at each stage of maturity, Qlik Data Catalyst gives your users more flexibility and power in selecting data that matches their analytic and information needs.

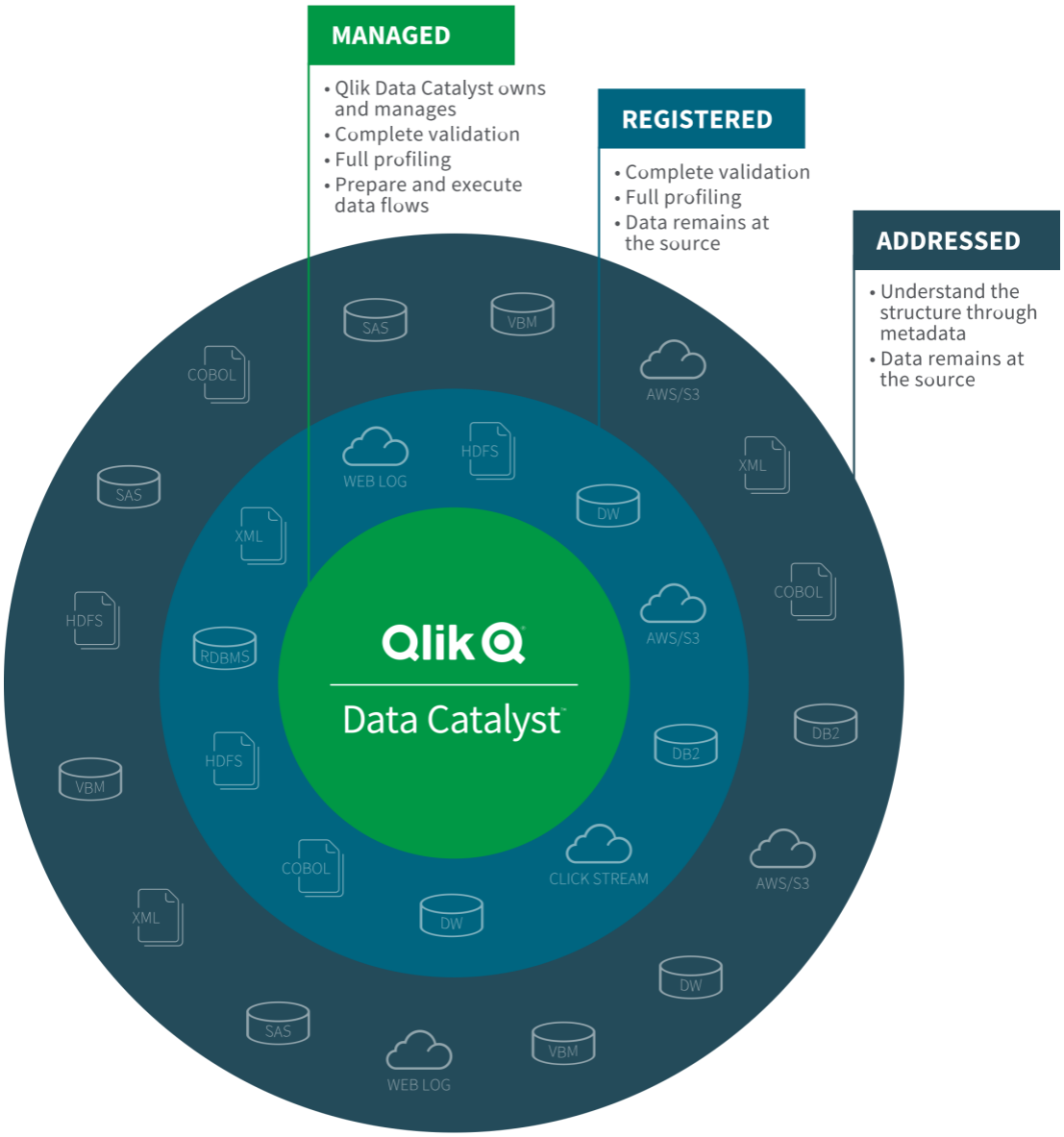
Qlik Data Catalyst sits as an abstraction layer on top of a Hadoop cluster or storage system like AWS, and manages the set of files that make up the data marketplace. These files are created whenever Qlik

Data Catalyst onboards new data into the marketplace (initially and during regular updates), when data is cleansed or enhanced, and when new datasets are prepared for use by your end users.

Qlik Data Catalyst maintains this persistent layer of data over time and throughout every stage of managing data in the data marketplace process—from raw to ready. There are three conceptual levels of data maturation in this raw-to-ready life cycle. Bronze data is data in its rawest form when it has just been brought into the data marketplace. Silver data has been somewhat cleansed and enhanced and is ready for access by data analysts. Gold data is even more enhanced and is appropriate for access by business users.

Data Conductor: Control Across the Enterprise Data Ecosystem

Data management with Qlik Data Catalyst provides deep understanding of data that is secure, trusted, and accessible from anywhere in the landscape of platforms, processes, and standards. Qlik Data Catalyst Data Conductor with level control is a powerful feature allowing your users to control how much information and profiling is applied to data assets across an enterprise data ecosystem. Three levels are available within Qlik Data Catalyst including Managed, Registered, and Addressed, giving your enterprise flexibility to determine where data should reside (whether within or outside a data marketplace) without sacrificing visibility of any data.



Each level defines a layer of visibility provided to the user:

1. **Managed:** Users have complete on-demand access to all of the data and metadata through the marketplace for managed data sources. When adding a new managed data source into the marketplace, Qlik Data Catalyst ingests all of the data from the source into the Qlik Data Catalyst data storage layer. Metadata from the source environment, along with metadata created through the Qlik Data Catalyst validation and profiling process, is added into the Qlik Data Catalyst Smart Data Catalog.

2. **Registered:** Metadata and sample data from registered data sources are fully integrated into the Qlik Data Catalyst Smart Data Catalog, allowing your users to easily search, find, explore, and preview data on demand. Actual data from registered data sources is not by

default stored in the marketplace. By promoting a Registered data source to Managed status, your users can launch the process that fully onboards the data into the Qlik Data Catalyst marketplace.

3. **Addressed:** With Addressed data sources, only metadata from the source environment is integrated into the Qlik Data Catalyst Smart Data Catalog. The data itself is not onboarded into Qlik Data Catalyst. Validation and profiling metadata are not created or available. Addressed data sources can be promoted to Registered or Managed status as needed to provide more complete access to metadata or data.

Your users have the ability to control the default level and to promote data from the Addressed to the Registered or Managed levels or from the Registered to the Managed level. Furthermore, your users may demote data from the Managed to the Registered or Addressed levels or from the Registered to the Addressed level.

As an example of active data management, an ingested data source may no longer require its data values to be stored in the Qlik Data Catalyst managed data marketplace, although its metadata may still be relevant for discovery purposes. The user within Qlik Data Catalyst can demote the entity containing the record from Managed to Addressed and refresh the metadata as needed. Similarly, the data can be promoted back to the Managed level if active management of the data in the marketplace is once again necessary.

Qlik Data Catalyst Security and Data Governance

Qlik Data Catalyst includes a robust set of enterprise-scale security and data governance capabilities to ensure that the data in a Qlik Data Catalyst data marketplace is completely protected and secure. These include authentication, authorization, accounting, encryption, and data masking.

Qlik Data Catalyst Takes a Platform Approach to Data Marketplace Security and Data Governance

The integrated Qlik Data Catalyst solution includes a tightly coupled set of capabilities to address all aspects of implementing and maintaining an enterprise-scale data marketplace. From a security perspective, this platform approach means that all of the Qlik Data Catalyst security and data governance measures are implemented across the data marketplace processes—consistently and without exception.

For example, when a user is granted new privileges to access a new data source, those actions are logged in the Qlik Data Catalyst metadata layer and implemented throughout the Qlik Data Catalyst application. Likewise, data encrypted or masked upon being ingested into the marketplace remains obscured as long as it is in the data marketplace, and even optionally as the data is exported to other systems. Consistent use of impersonation means that actions by authorized, authenticated users are always logged, regardless of whether the user is ingesting, enhancing, cleansing, preparing, or exporting data in the marketplace.

By incorporating consistent security and governance capabilities throughout the data marketplace process and making those measures transparent and accessible to your Qlik Data Catalyst administrators through an intuitive point-and-click GUI, Qlik Data Catalyst eliminates potential security failure points in the data marketplace and consolidates administration and maintenance of security measures.

Authentication

Qlik Data Catalyst integrates directly with enterprise identity management technologies such as Active Directory and LDAP to synchronize active groups and users. Qlik Data Catalyst will then authenticate users against corporate Active Directory/LDAP/Kerberos infrastructure.

Authorization

The Qlik Data Catalyst solution implements authorization and manages each user's access to data in the marketplace through a system of permissions, roles, users, and groups. Each Qlik Data Catalyst group represents a set of data assets from within the data marketplace as defined by an administrator. Your users are assigned to one or more groups and given different permissions to interact with data assets within each group.

Authorization may also be granted by Sentry/Ranger or HDFS ACLs when Qlik Data Catalyst is run in impersonation mode. When this is the case, executed jobs and queries are associated with that specific user and any policies set via these external systems will be honored.

This system gives Qlik Data Catalyst administrators very granular control to efficiently establish each user's ability to create, read, or write data at the level of a specific source or entity.

Accounting

Qlik Data Catalyst logging provides audit trail information on all consequential actions taken by users in the Qlik Data Catalyst Service, either through the Qlik Data Catalyst GUI or Qlik Data Catalyst RESTful API. Log data is recorded in the Qlik Data Catalyst Smart Data Catalog and covers all actions that impact data, metadata, or both.

Qlik Data Catalyst Hive, and HDFS: Managing User Permissions Across the Hadoop Stack

Qlik Data Catalyst runs natively on top of a Hadoop cluster and executes MapReduce/Hive/Pig code to store, access, and manipulate data in the Qlik Data Catalyst data marketplace. Throughout this exchange, Qlik Data Catalyst complies with and enforces the data and file access control methods provided by HDFS and Hive.

Specifically, HDFS (the Hadoop Distributed File System) honors the Unix/POSIX permission model with user(u), group(g), and others(o) rights associated with file and directory entries. Additionally, HDFS provides extended Access Control List (ACL) entries to accommodate user/ group permissions that do not fit within a strict hierarchy.

Qlik Data Catalyst works with Ranger and Sentry, when available, to coordinate users' database-level access controls.

Impersonation

Impersonation becomes a mandate once Hadoop clusters are securely accessible by authenticated Kerberized users. Qlik Data Catalyst uses impersonation when accessing files in HDFS to ensure data security while also creating transparency and an audit trail of any time an individual user's access to the data resulted in any change. Qlik Data Catalyst uses impersonation so that that every request by a Qlik Data Catalyst user to create, read, or update a file in HDFS results in a persistent record of that action, documented with that user's ID and a time stamp of the action in HDFS.

Impersonation support enables Qlik Data Catalyst, functioning as a Service User with access to all commands and directories, to execute tasks/run jobs on the cluster on behalf of a user in a secure way. Therefore, any permissions or privileges that are applicable for that user (via Sentry or similar tools) will be honored.

Encryption

Qlik Data Catalyst provides encryption capabilities to help clients meet stringent regulatory requirements (HIPAA, PCI DSS, FISMA, Sarbanes-Oxley) regarding data privacy and protection of valuable data assets. Qlik Data Catalyst offers two methods to encrypt data in the marketplace. First, Qlik Data Catalyst works with Hadoop HDFS transparent encryption. Qlik Data Catalyst reads and writes data to and from HDFS blocks, seamlessly propagating the encryption in place within Hadoop directories.

Second, Qlik Data Catalyst provides field-level encryption through the Qlik Data Catalyst application interface, allowing your users to mark data to be encrypted as it is ingested into the marketplace or later, when it is already in the marketplace. Qlik Data Catalyst also allows your users to flag data that arrived in the data marketplace encrypted from its source system.

Data Masking

With Qlik Data Catalyst, your users can flag individual fields of data in the marketplace as sensitive and the solution will mask data in those fields using a variety of obfuscation methods. Qlik Data Catalyst supports a variety of masking techniques, including substitution, shuffling, number and date variance, and encryption. These techniques can be flexibly configured to secure data in term-specific, targeted ways.

Qlik Data Catalyst Is a Native Hadoop Application

Qlik Data Catalyst was purpose-built to fully exploit the performance and economic advantages of Hadoop and its associated open source projects. As a fully native Hadoop application, Qlik Data Catalyst stores data and executes all data access or data processing activities on the Hadoop cluster and leverages the Hadoop massively parallel processing architecture to drive faster performance. This allows the platform's users to leverage the full power of the Hadoop parallel processing model while loading, bringing large volumes of data into the marketplace.

Through its edge-node architecture, Qlik Data Catalyst provides the most flexible, open data marketplace solution on the market. Qlik Data Catalyst can be deployed on-premises behind the firewall or in the cloud including elastic data processing, as well as in a hybrid of the two.

Because Qlik Data Catalyst follows Hadoop standards, it's tested and certified with all major Hadoop distributions, including Cloudera, Hortonworks, MapR, and Amazon EMR, and is available through the Microsoft Azure Marketplace. By ensuring certification with all major Hadoop distributions and

compliance with Hadoop open software standards, Qlik Data Catalyst helps your organization use and benefit from Hadoop while avoiding getting locked into any proprietary technology.

Qlik Data Catalyst also leverages new Apache Hadoop-related projects as they evolve to deliver the functionality, maturity, and stability required for use in an enterprise-scale data marketplace solution. This includes integration with Hive, Pig, Tez, and Spark as well as Cloudera Sentry, Hortonworks Ranger, and MapR POSIX Client projects. By constantly tracking and selectively integrating with new Hadoop-related technology as it matures, Qlik Data Catalyst gives your organization a low-risk, low-effort way to automatically benefit from ongoing innovation in the open-source Hadoop ecosystem.

How Qlik Data Catalyst Interacts with the Hadoop Cluster

Qlik Data Catalyst interacts with data in HDFS by building on underlying Hadoop technologies. MapReduce jobs, Spark jobs, Hive queries, and Pig scripts are generated by the Qlik Data Catalyst server in response to actions taken by users in the Qlik Data Catalyst GUI or initiated by Qlik Data Catalyst API calls. The actions are translated into standard Hadoop API calls and executed on the Hadoop cluster, often generating new files, which are registered in HCatalog. Results of any actions taken at the HDFS level are shared with your end user and recorded in the Qlik Data Catalyst Smart Data Catalog.

Data ingested by Qlik Data Catalyst is stored in standard file formats established by the Apache Hadoop open source software framework, including Delimited, Parquet, ORC, and Avro. All files created by Qlik Data Catalyst in HDFS are registered in HCatalog so they can be viewed by any Hadoop-enabled application. Likewise, any Hive scripts or Pig code generated by Qlik Data Catalyst fully complies with the open source standard.

Qlik Data Catalyst Deployment Options Support Enterprise Ready

Qlik Data Catalyst can be deployed on-premises, in the cloud, or in a hybrid model. The Qlik Data Catalyst server and Smart Data Catalog can be hosted onsite or off-site or parts in each place, regardless of where the associated data storage and processing layer is hosted. Qlik Data Catalyst can also be migrated from one deployment model to another with nominal re-configuration of parameters.

This provides maximum flexibility, allowing your organization to evolve its Qlik Data Catalyst implementation incrementally in step with your overall cloud strategy.

Qlik Data Catalyst is installed on the edge node of the cluster; Qlik Data Catalyst is not installed on each node in the cluster. This makes the initial install and any subsequent upgrades of Qlik Data Catalyst fast and easy, thereby minimizing downtime during typical service-affecting tasks. The focus on enterprise ready enables your organization to leverage Qlik Data Catalyst as the consistent end-to-end data marketplace platform.

Integrating Qlik Data Catalyst with Other Applications in Your Enterprise IT Environment

Qlik Data Catalyst can be easily integrated with other applications or enterprise data management platforms to share data or metadata. Data in the Qlik Data Catalyst marketplace can be used as a landing zone to receive data updates from operational systems and to execute data validation, cleansing, and transformation operations before loading that data into a data warehouse. Other organizations integrate Qlik Data Catalyst marketplace with downstream analytic systems, databases, or applications across the enterprise. Qlik Data Catalyst can also be integrated with enterprise data governance tools, catalogs, or metadata repositories to ensure that information collected in Qlik Data Catalyst is part of the overall enterprise data management process.

Conclusion

Qlik Data Catalyst provides your organization with end-to-end functionality to collect, prepare, deliver, and manage a data marketplace, enabling you to maximize the value of your enterprise data assets while minimizing development and operations costs.

By creating a persistent layer of secure, well-organized, and accessible enterprise data beginning at ingest and making each asset easily accessible to your business users through an intuitive GUI, Qlik Data Catalyst helps your company deliver a next-generation model of data-as-a-service across the organization.

Our platform enabling data-as-a-service is a data marketplace, bringing together your data providers and data consumers to securely and collaboratively share information. Through it, you drive expanded

use of data across business groups and in a diverse set of reporting, analytic, and visualization applications.

Built on a shared platform of robust data security and data governance services, Qlik Data Catalyst allows your organization to confidently expand access to enterprise data while reducing risk and exposure to inappropriate data access or loss.

One Go-To Source for All Data

Qlik Data Catalyst brings all of your enterprise data together into one, centralized repository where it can be instantly accessed for a wide range of uses. All data is cataloged, validated, and profiled so it's easy to find and understand.

Data You Can Trust

Qlik Data Catalyst is fueled by rich metadata, which provides visibility and context for all data at a granular level. Your users know exactly what the data is, where it came from, and what's been done to it so they can act on it with greater confidence than ever before.

Self-Service; Shop for Data

Your analysts using Qlik Data Catalyst simply shop for data—accessing, exploring, preparing, and sharing data on their own, without any help from IT. Preparing data takes only minutes, not months.

Complete Governance and Control

Qlik Data Catalyst takes a platform approach to data security that centralizes control and applies governance at the point of use. Enterprise-grade encryption and authentication, along with granular-level permissions and tracking, help ensure data security and compliance.

Learn more at www.qlik.com.

About Qlik

Qlik® is on a mission to create a data-literate world, where everyone can use data to solve their most challenging problems. Only Qlik's end-to-end data management and analytics platform brings together all of an organization's data from any source, enabling people at any skill level to use their curiosity to uncover new insights. Companies use Qlik products to see more deeply into customer behavior, reinvent business processes, discover new revenue streams, and balance risk and reward. Qlik does business in more than 100 countries and serves over 48,000 customers around the world.

qlik.com