Datameer
stay curious

Datameer Big Data Governance

Bringing open-architected and forward-compatible governance controls to Hadoop analytics

As big data moves toward greater mainstream adoption, its compliance with long-standing enterprise standards and industry regulations is becoming increasingly important.

As the concept of the data lake – a central repository for storage and self-service access to any data – begins to take hold, these concerns become even more acute. Strong data governance capabilities, like the ability to audit additions and changes to data, trace your data's lineage and sharing, assign role-based access to data and perform impact analysis are vital if big data technology is to become a standard part of the enterprise technology toolkit.

Governing Hadoop

Unfortunately, Hadoop, the dominant big data technology, does not natively offer such features. The notion of tracking changes in data, and of controlling access to data in a granular fashion (where certain users have access to certain subsets of it) is assumed functionality in the world of the enterprise data warehouse. Not so in the Hadoop world.

Another facet of working with data in Hadoop is that analysts tend to spread their work between different tools in its ecosystem, including Apache Hive, Apache Pig, MapReduce and higher-level platforms built atop Hadoop. So even as some governance features are added to individual components, the need for an overarching governance system is clear.

These facets of working with big data raise new challenges and risks, such as:

- Ensuring users only have access to data for which they are authorized
- Maintaining centralized policies (such as ACLs) that are enforced across different tools in the ecosystem
- Gaining visibility into which data is used downstream for which analytics, and by whom
- Ensuring analytical results are based on valid and high-quality data
- Ascertaining how data flows through the system and how it is transformed
- Determining how changes to data and analytics will affect other assets in the workflow
- Meeting internal and external compliance or regulatory requirements
- Identifying who may have made changes to the data where analytics processes are failing or producing unexpected results

Datameer and **Good Governance**

Datameer provides just such a solution, ensuring that customers don't have to choose between self-service big data analytics and a robust, governable data architecture. Since Datameer runs natively on Hadoop, it can track any and all changes to the data made in the Datameer environment, regardless of which Hadoop component might be used to execute the processing tasks.

Perhaps equally important, as new Hadoop components and execution engines are introduced and support for them is added to Datameer, work dispatched to those components will inherit the same governance features provided by Datameer. Each of Datameer's governance functionalities addresses one or more of the following five pillars of strong data governance, enabling businesses to finally democratize data access with confidence.

- Quality and consistency
- Data policies and standards
- Security and privacy
- Regulatory compliance
- Retention and archiving

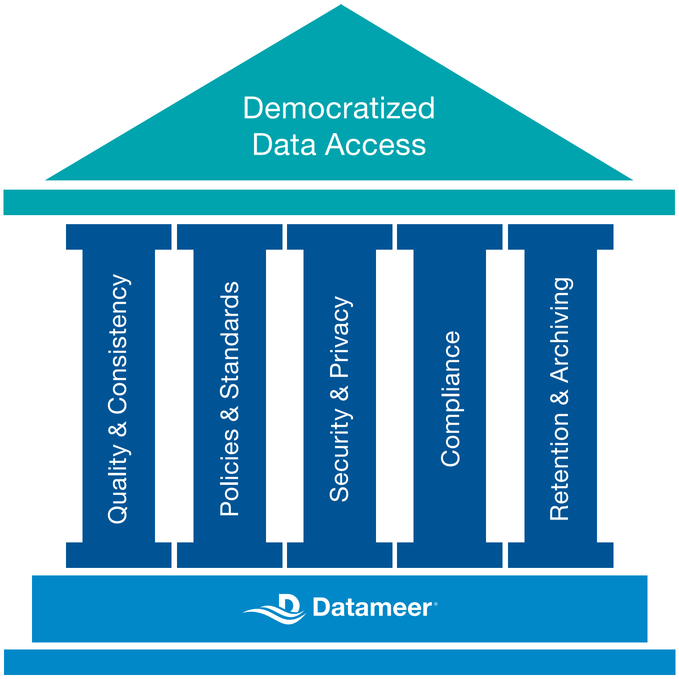


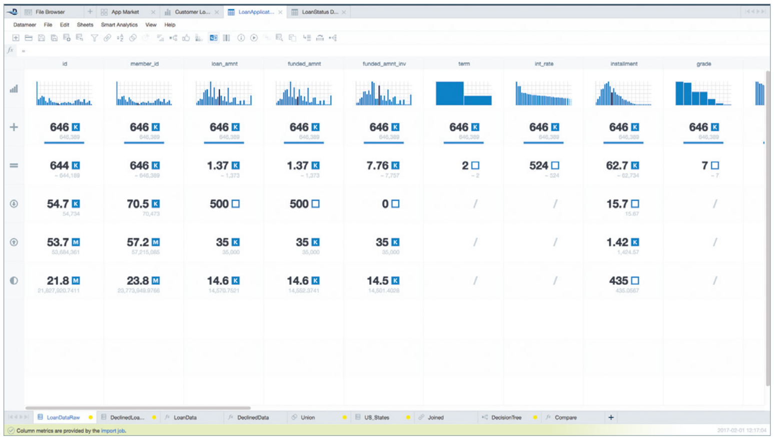
FIGURE 1

Comprehensive big data governance functionality addresses each of these pillars, ensuring fully governed yet democratized data access in a data lake

1. Quality and Consistency

Data quality and consistency are imperative when it comes to ultimately extracting value from big data.

If at any point in the data pipeline there is a question about data validity, the overall value of the resulting insights is in question. Datameer's data profiling tools enable you to check and remediate issues like dirty, inconsistent or invalid data at any stage in the analytics pipeline, and provides transparency into every change, from the original data set all the way through to the final visualization. Additionally, derived fields and metrics can be shared across projects to ensure consistency of calculations and thus results.



	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade
+	646 K 646,389	646 K 646,389	646 K 646,389	646 K 646,389	646 K 646,389	646 K 646,389	646 K 646,389	646 K 646,389	646 K 646,389
=	644 K ~ 644,189	646 K ~ 646,389	1.37 K ~ 1,373	1.37 K ~ 1,373	7.76 K ~ 7,757	2 □ ~ 2	524 □ ~ 524	62.7 K ~ 62,734	7 □ ~ 7
	54.7 K 54,734	70.5 K 70,473	500 □	500 □	0 □	/	/	15.7 □ 15.67	/
	53.7 M 53,684,361	57.2 M 57,215,085	35 K 35,000	35 K 35,000	35 K 35,000	/	/	1.42 K 1,424.57	/
◐	21.8 M 21,827,920.7411	23.8 M 23,773,949.9766	14.6 K 14,570.7521	14.6 K 14,552.3741	14.5 K 14,501.4028	/	/	435 □ 435.0567	/

Column metrics are provided by the import job.

FIGURE 2

With Datameer's Flipside, analysts can view the data type, count, max, min, uniqueness, mean and average,
to understand the shape and quality of their data at every step of the analytics process

Specifically, Datameer offers these Quality and Consistency capabilities:

- **Data Profiling** with Flipside provides simple, highly accessible, visual data profiling that lets users easily spot outliers in data, quickly and early in the analytics process. As data quality issues are remediated, those efforts are themselves logged, so they may be audited later. Meanwhile, downstream analyses are further safeguarded from dirty data and erroneous results are readily avoided.
- **Data Statistics** monitoring can detect dirty, corrupt or invalid data early, auto-detect and quantify calculation errors (like divide-by-zero) that might affect analytic results, ensure consistent data volume and throughput and ensure completeness of records and data sets throughout the pipeline.
- **Metadata Management** catalogs all data and analytics artifacts, provides a REST API for programmatic access and audit, and integrates with external data governance tools and frameworks that are filtering into the broader Hadoop ecosystem. This empowers IT managers and compliance officers to get a comprehensive view of their assets across multiple vendor platforms and technologies.

2. Data Policies and Standards

Data access policies are the first line of defense against risk for businesses. For IT, the goal is to implement policies that allow them to manage risk appropriately, while still meeting business needs.

Specifically, Datameer supports the following capabilities to aid in enforcing data access policies:

- Secure data views enable administrators and privileged users to expose a subset of fields to specific groups or users, and apply masking, anonymization or aggregation to sensitive data fields, while leveraging the column-level security and metadata security of both Datameer and external systems like Apache Sentry.
- Multi-stage analytics pipelines enable end users to build data preparation or analytics workbooks on top of secure data views, and apply further policies and role-based security at every stage of the pipeline, from ingest to export.

3. Security and Privacy

True Big Data security needs to exceed that of the Hadoop Distributed File System's built-in capabilities.

Fine-grained access control is important, both at the row and column level, and any added metadata needs to carry with it the same level of security. Integration with enterprise identity management systems like Active Directory/LDAP should be a given. And role-based controls on downloading or exporting data and accessing administrative functions are mission-critical. Datameer provides all these and more across your end-to-end big data analytics pipeline.

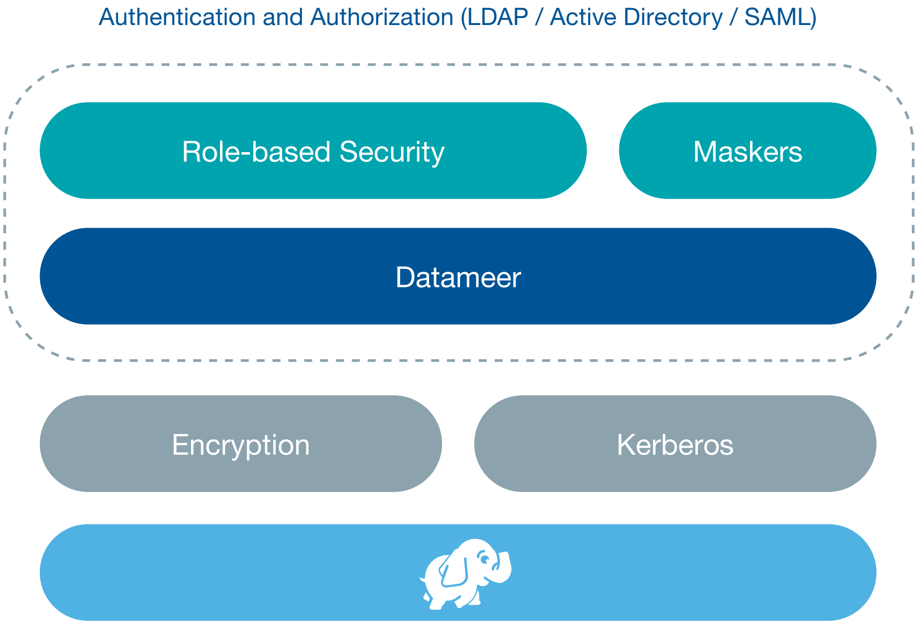

Authentication and Authorization (LDAP / Active Directory / SAML)
Role-based Security
Maskers
Datameer
Encryption
Kerberos

FIGURE 4

Datameer authentication/ authorization and access control features

Datameer's security & privacy facilities include:

- **LDAP / Active Directory / SAML** support allowing enterprise identity management standards to be retained and leveraged out-of-the-box.
- **Encryption-at-rest and Encryption-in-transit:** Datameer works seamlessly with the built-in capabilities of HDFS and YARN to encrypt all data in Hadoop, and adds wire-level SSL encryption of all data transmitted to the user's browser.
- **Secured Impersonation** ensures jobs run as, and created data belonging to, the authorized Datameer user or group, and that these permissions and audit trail are captured in all Hadoop ecosystem components like HDFS and YARN.

- **Role-based access control** allows IT to control which users can perform which tasks throughout the Datameer application. For example, you can give bulk ingest abilities to IT staff only, while still allowing analysts to upload their own files on an ad hoc basis.
- **Permissions and Sharing** means all Datameer artifacts, including imported data, export jobs, workbooks and infographics can be shared with an individual, a group, multiple group or everyone, and even allows individual visualization widgets to be published to secured or unsecured URLs.
- **Bi-directional, point-and-click integration with Apache Sentry 1.4**, delivers centralized security policy management for data in Hive,Impala, HDFS and Datameer.
- **Column security and anonymization functions,** letting users transform data, including removing columns, filtering rows or anonymizing data with secure hashing functions.

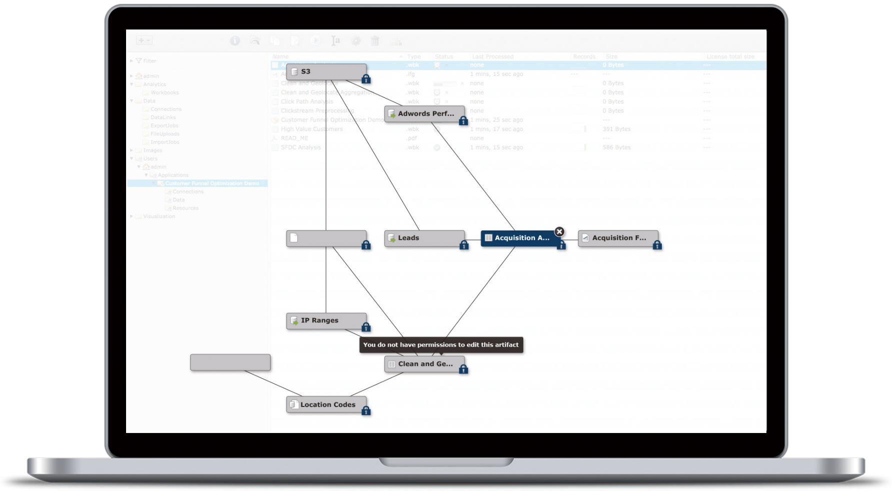
4. Regulatory Compliance

Across several industries, there are legal imperatives for Big Data governance. From Sarbanes Oxley, to Basel, to HIPAA and PCI compliance, without strict data governance functionality, big data technologies can't be deployed in some environments.

(Big) Data Lineage

In addition to legal requirements, there are also numerous functional requirements and productivity needs which make big data lineage extremely useful.

For example, looking at the final output of a sophisticated analysis visually can provide a lot of useful information very quickly. But it can also be a bit opaque, requiring a leap of faith in order to trust the efficacy of the analysis. The number of steps and transformations that data can undergo can make it much more difficult to understand the genesis of the analytical results than the insights they provide. In some scenarios, the implicit trust may be enough, but in most enterprises, such analyses must be verifiable.



You do not have permissions to edit this artifact

FIGURE 5

Datameer's big data lineage functionality

Datameer's new cross-artifact lineage features facilitate this, making such verification almost as easy as consumption of the results themselves. Through an easily understood graphical representation, users can understand the source and journey of all data in the analysis, respective of the permissions they've been granted by the owners of each step in the process, or by administrators. This brings peace of mind that generates confidence in analytical results, and with such confidence comes buy-in, deeper trust and broader adoption of the solutions. Data lineage tools also enable rapid discovery, easily answering the questions of who to turn to about various data or analytics tasks and facilitating reuse. This tight feedback loop between analysis, insight and discovery creates the virtuous cycle the big data world so urgently needs.

Specifically, Datameer's data lineage capabilities include:

- Cross-artifact dependency graphs, which permits tracing upstream back to the source, or seeing downstream dependencies, and allowing users to ensure that valid data has been used and follow what transformations and analytics functions were applied.
- Dependencies REST API, help synchronize Datameer's lineage with that of external metadata management systems and applications, and helps package-related artifacts for deployment.
- Worksheet lineage, providing lineage information at the worksheet and field level.

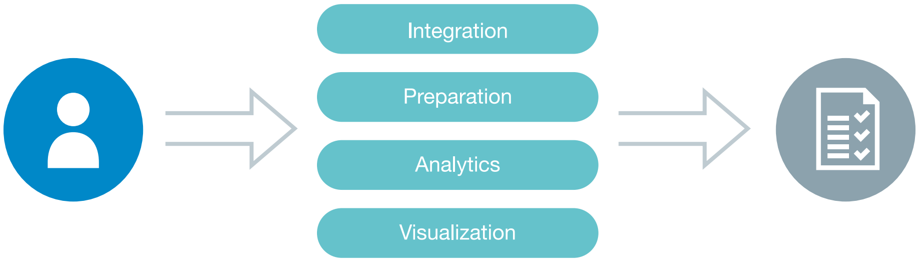
Audit

In Datameer, all relevant user and system events, including data creation and modification, job executions, authentication and authorization actions, and data downloads are automatically and transparently logged. These logs can be analyzed in Datameer itself, or by an external system.

The data in these logs can also be used for periodic security audits. In addition to the types of events already mentioned, the logs also contain important data about users and their interaction with the system (not just the data). This includes information about groups and roles, their assignments, artifact sharing, logins and failed login attempts, password updates, enabling and disabling of specific users and more.

FIGURE 6

Audit logs are created to track every relevant user and system event



Specifically, Datameer offers these audit capabilities:

- **User Action Log:** Datameer maintains a log file with all relevant user and systems events and information, such as CRUD events on Datameer entities and other system entities, job executions, data downloads, volumes ingested and many more.
- **Security Audit Log:** Datameer maintains a dedicated security audit log that captures relevant actions for security investigations and audits, including all authentication attempts, logouts and changes to permissions for data and other artifacts.
- **A software development kit (SDK)** that allows external systems to be apprised of user and system audit events as they happen.
- **Audit Reports (HUM application):** Datameer provides pre-built reports that aggregate, analyze and visualize log data in the form of a Datameer application called “HUM” (Health, Usage and Monitoring).

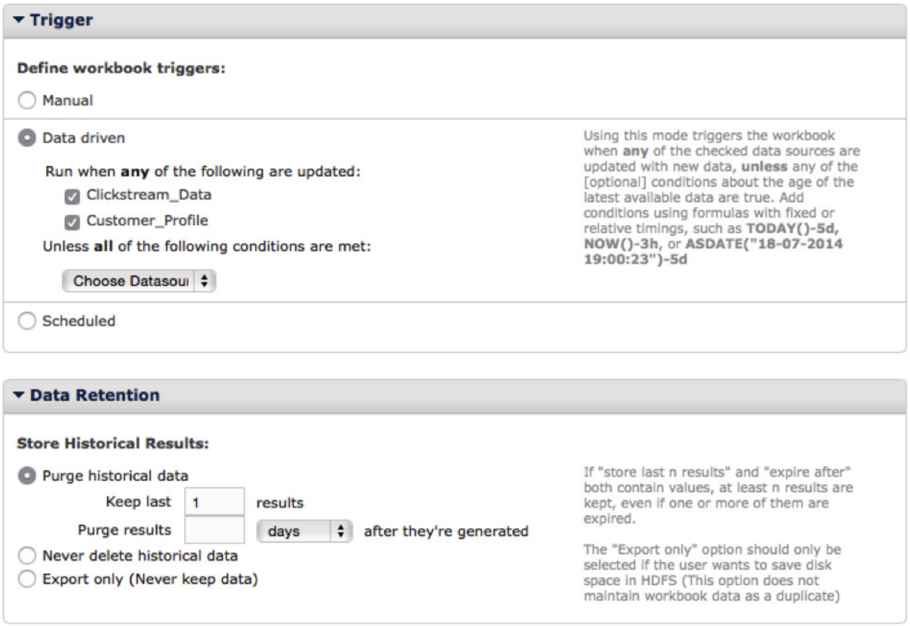
5. Retention and Archiving

In Datameer, flexible retention rules allow each imported data set's retention policy to be configured by an individual set of rules.

It is easy to configure Datameer to keep data permanently, or to purge records that are older than a specific time window. Independent of time, retention rules can also be configured based on the number of runs of ELT ingests or analytics workbook executions. Security rules allow retired data to be either instantly removed, retained until a specified time or manually removed after system administrator approval.

FIGURE 7

Datameer's data retention functionality



Trigger

Define workbook triggers:

○ Manual

● Data driven

Run when **any** of the following are updated:

☑ Clickstream_Data

☑ Customer_Profile

Unless **all** of the following conditions are met:

Choose Datasou

Using this mode triggers the workbook when **any** of the checked data sources are updated with new data, **unless** any of the [optional] conditions about the age of the latest available data are true. Add conditions using formulas with fixed or relative timings, such as **TODAY()-5d, NOW()-3h**, or **ASDATE("18-07-2014 19:00:23")-5d**

○ Scheduled

Data Retention

Store Historical Results:

● Purge historical data

Keep last 1 results

Purge results days after they're generated

○ Never delete historical data

○ Export only (Never keep data)

If "store last n results" and "expire after" both contain values, at least n results are kept, even if one or more of them are expired.

The "Export only" option should only be selected if the user wants to save disk space in HDFS (This option does not maintain workbook data as a duplicate)

Future-Proof

One more dimension to the big data governance story is the emerging governance tools and frameworks built specifically for the Hadoop ecosystem. While critical mass for these and other tools and standards is still gathering, this creates a crucial requirement that any governance features used today in platforms like Datameer be forward-compatible with frameworks that will be important in the future.

Datameer's open, API-based approach to governance was conceived with such forward-compatibility in mind. Event-driven metadata can be transmitted to other systems in real time, allowing for the same audit information to be available in the Datameer environment and to external governance systems.

Additionally, these same APIs can be used to integrate with Software Configuration Management (SCM) systems like Git and Subversion, to add governance to the deployment of operational analytics workflows into production.

Get Moving, Responsibly

With comprehensive big data governance in place, openly-architected and forward-compatible, Datameer removes a serious barrier to big data initiatives in the enterprise, even where highly sensitive data is used.

While the Hadoop ecosystem evolves and data governance standards and frameworks emerge, Datameer gives you the rigorous data governance capabilities you need right now, with an architectural approach that protects your investment as new systems and standards are introduced.