World of BigData

"Problems cannot be solved with the same mind set that created them." – Albert Einstein

Data Governance -What, When, Where, Why, Who and How of Data

July 19, 2020July 19, 2020 veejayendraa Data Governance Data Catalog, Data Governance, Data Lineage

i
7 Votes

Today's digitally connected world is producing massive amount of data. This data could provide lots of information and answer many questions. It has become vital for organizations, for countries and for each of us.

Some of the examples where data is helping us and our society are –

- Usage of data in the medical field to predict the chances of developing any particular diseases and taking preventive measures based on this information.
- Predicting traffic conditions in a city at a particular time and developing infrastructure based on this information
- Identifying potential customers by analyzing their search history and shopping behavior. Providing targeted ads to such identified customers.
- Detecting credit card fraud by analyzing the location of usage of the card and cardholder, frequency of purchase, etc.
- Forecasting the sale during holiday season etc.

There are many such examples that can prove the importance of data and the importance of information that is locked within the data. In today's world data must be seen as an important tool that can shape up our behavior, our thinking, our future.

> **Data analytics or Data Intelligence is now no longer a tool to get a competitive edge within a business but it has now become essential to remain in business.**

To reap full benefits of data we need data which is –

- Good in quality
- Timely
- Accurate
- Properly managed and secured

Along with this we need information like –

- What is the origin of data?
- Where it is stored?
- What information does it store?
- Who owns it? How can I access it?
- What is the freshness of data? When last time it was ingested?

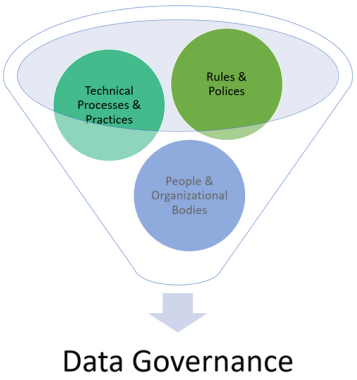
Organizations can see an improvement in business performance if they successfully consider the who – what – how – when – where and why of data to not only ensure security and compliance but to extract value from all the information collected and stored across the business.

But for this, organizations need to have tools and skills to manage the data properly to increase their value by many folds and make it much more useful. And this is where Data Governance comes in the overall picture of Data Engineering & Analytics.

What is Data Governance?

Data governance refers to the overall management of data in terms of the availability, usability, integrity, and security of the data in an organization. Majorly data governance depends upon business policies and hence there should be agreed –

- Enforced rules & Policies
- Organizational Roles, Ownership, Accountabilities
- Technical processes, tools & practices



Some people get confused between Data Management & Data Governance but there are distinctive differences between Data Management and Data Governance. Data management is concerned with the usage of data in making good business decisions whereas Data governance is concerned with how disciplined we are in managing the data across the organization.

Data Governance – Whose responsibility?

Data Governance involves the whole organization but some are involved more than others. But there should be defined responsibilities and based on which defined roles should be created within an organization.

Data Owners: Data owners (or data sponsors) are people who have the authority to make decisions for the data they own and enforce these decisions throughout the organization. They can be appointed at the entity / domain/department level. Data owners are ultimately accountable for the state of the data as an asset.

Data Stewards: Data stewards (or data champions) make sure that the data policies and data standards are adhered to in daily business. These people will often be the subject matter experts for a data entity. Data stewards are responsible for taking care of the data as an asset.

Data Governance Committee: Typically, a data governance committee establishes the main forum for approving data policies and data standards and handle escalated issues. Depending on the size and structure of each organization there may be sub fora for each data domain (eg finance, risk – (subdivided into credit risk, market risk and operational risk), trades, employee).

These roles should optionally be supported by a Data Governance Office with a Data Governance Team comprising of **Solution and Data Architect**, **Master Data Governance Lead**, **Compliance specialist, Data Engineers and Data Analyst**.

One of the most important aspects of assigning and fulfilling the roles is having a well-documented description of the roles, the expectations and how the roles interact. This will typically be outlined in a RACI matrix describing who is **R**esponsible, **A**ccountable, to be **C**onsulted and to be **I**nformed within certain enforcement, process or for a certain artifact as a policy or standard.

Why Data Governance is MORE important to a Data lake?

Answer to this questions lies in definition of data lake itself.

> *Data lake allows for the ingestion of large amounts of raw structured, semi-structured, and unstructured data that can be used for analysis as and when needed*.

Most of the organizations when they start implementing data lake solution, they start putting a vast amount of data, dumping everything into it hoping that they will use this data down the line. But after some time they lose track of what's there and turning data lake into a dump yard.

In the words of research firm Gartner

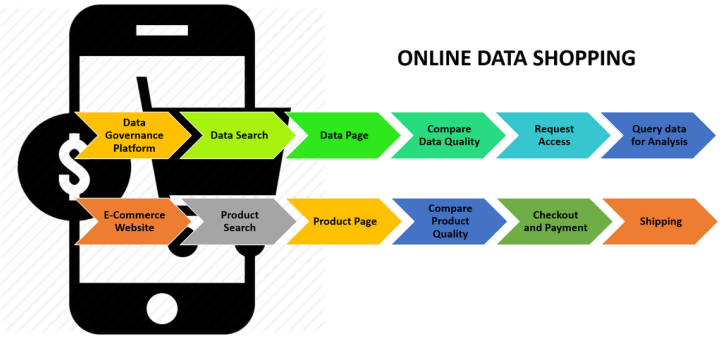
> "without at least some semblance of information governance, the data lake will end up being a collection of disconnected data pools or information silos all in one place…Without descriptive metadata and a mechanism to maintain it, the data lake risks turning into a data swamp."
>
> *https://www.gartner.com/en/newsroom/press-releases/2014-07-28-gartner-says-beware-of-the-data-lake-fallacy (https://www.gartner.com/en/newsroom/press-releases/2014-07-28-gartner-says-beware-of-the-data-lake-fallacy)*

Data stored within the data lake could be very sensitive and may contain personal identifiable information (PII) data. However, this sensitive data must be protected, compliant with privacy laws and regulations. This makes data governance a critical pillar in designing a data lake. Hence governance should be incorporated as part of the design in the beginning or at least minimum standards should be incorporated since inception.

For the success of any enterprise data lake, it is essential that data citizens should be able to do the following tasks with ease –

- Search, understand, and use that data to derive useful information from it.
- Should be able to compare the trustworthiness of data.
- Should be easily able to know how to get access to any data and from whom.

In simple words, getting the data from the data lake should be like a shopping experience.



For e.g when we know what item we want to purchase, we go to online shopping app, search for the item, browse through searched items, compare their values, compare their quality by going through reviews and rating provided by others, purchase few items by adding them to our basket and finally do the payment for the items. Once the vendor receives the order and payment, they ship the items to our address.

Similarly, once you know your use case you should be able to search the data, compare the data quality, and finally get access to the data by sending an access request to the data owner. Once the data owner approves your request, you should be able to access the data.

All this can happen if and only if Data within your data lake is governed properly. Data Governance majorly deal with following areas –

Business Glossary / Data Dictionary

We all need a common language to communicate with each other in our day to day life. Similarly in an organization, the same term could be used in different ways. For e.g. if I go to the Finance department of a bank and ask them for PnLs, they will give me actual PnLs. While if I talk to people in the Risk department of the same bank, they may give me Projected PnLs.

The other example could be Party information stored as Party information in some system and its unique identifier as party_id whereas it could be stored as Counterparty information in some systems within the bank and its identifier could be termed as counterparty_id.

This disparity results in lots of confusion and results in wastage of precious time while understanding the data.

A well maintained Business Glossary becomes vital in such situations and becomes very handy in removing such confusions. It becomes the organization's official semantic translator. People can talk in their own language as long as they are communicated to each other through that semantic translator. So all they need to learn is how to communicate with each other using that semantic translator.

Tools like Lumada (https://www.hitachivantara.com/en-us/products/data-management-analytics/lumada-data-services/lumada-data-catalog.html) (earlier known as Waterline, before acquisition by Hitachi), Collibra (https://www.collibra.com/) and Alation (https://www.alation.com/) help us in creating and maintaining the Business Glossary. These tools allow us to map Business Glossary to underlying tables and columns. Generally, organizations use crowdsourcing within the organization to build enterprise Business Glossary. But ultimately its onus on Data Owners and Data Stewards to manage it properly.

Data Domains

Every organization has some specific business/businesses. And every business line has some data domains.

> Data Domains are high-level categories of Organizational Data for the purpose of assigning accountability and responsibility for the data.

For example within an Investment Bank, there are domains like Trade Data – Equity Trade, FX Trade, etc as subdomains, Counterparty Data, Risk data – Market Risk, Credit Risk as subdomains, Reference data – curve, surface, interest rates data, Journals & Balances data, etc.

Similarly, in the Telecom organization, examples of data domains would be CDR (call detail record) data, Customer data, etc.

These data domains basically drive any organization's business and become the focal point of any analysis that an organization wishes to consider. That's why its very crucial to identify data domains for the success of data governance strategy.

Each domain/ subdomain within an organization should have –

- Data Owners & Stewards
- Policy and Standards to maintain that data
- Business Glossary / Data Dictionary
- Data Catalogs
- Data Quality Scores
- Business Processes
- Systems & Applications to manage this data.

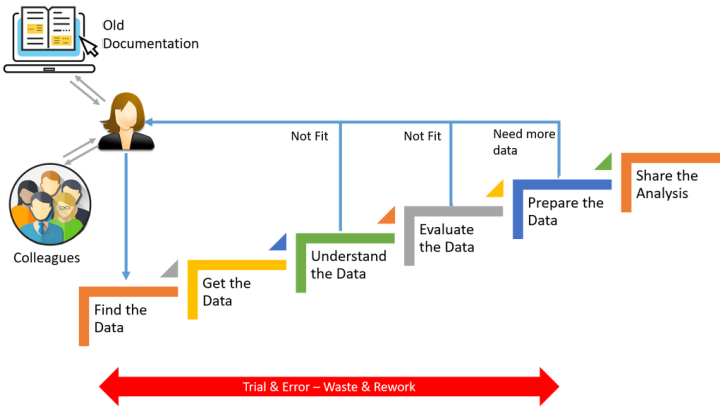
Cataloging

With the vast amount of data stored in a data lake, it becomes hard to keep track of what data is already available and thus becomes very difficult to search for a dataset. A solution to this is the Data catalog.

> A Data Catalog is a collection of metadata consisting of context, data quality score, sensitivity score and information about artifacts that use this data such as BI Reports, API, Machine Learning models, etc. combined with search tools that help others to find the data that they need and then get access to it.

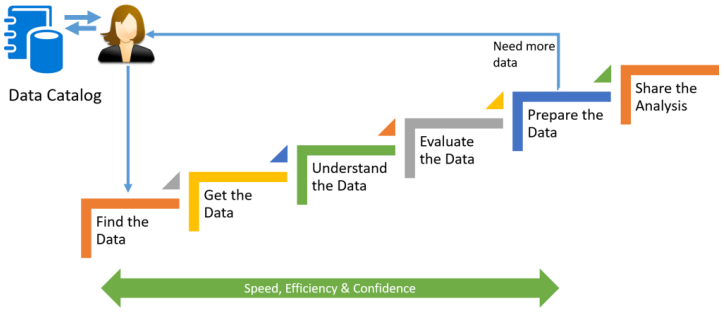
Properly maintained Data Catalog helps data citizens to work in self-service mode by empowering them to shop and find trustworthy data.

Without a catalog, analysts look for data by going through the documentation, talking to colleagues, relying on tribal knowledge, or simply working with familiar datasets because they know about them. The process is filled with trial and error, waste and rework, and repeated dataset searching that often leads to working with "close enough" data as time is running out.



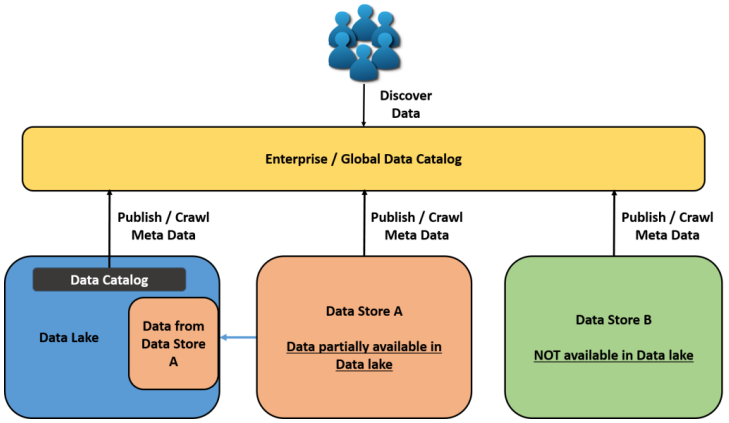
Analysis preparation without Data Catalog

With a data catalog the analyst is able to search and find data quickly, see all of the available datasets, evaluate and make informed choices for which data to use, and perform data preparation and analysis efficiently and with confidence.



Analysis preparation with the help of Data Catalog

Within a data lake environment where multiple frameworks, tools, and technologies are used for Data Ingestion, Transformation, and Data Access the most efficient way is to maintain a central data catalog and use it across various frameworks like Apache Hadoop, Apache Spark, Hive, Impala, Presto, EMR, Athena and Glue on AWS, similarly DataProc and BigQuery on GCP and various other available tools. This ensures metadata integrity and applying easy data governance rules. This central data catalog should be further integrated with Global / Enterprise Data Catalog so that we can get one common picture across all the data sources within the organization.

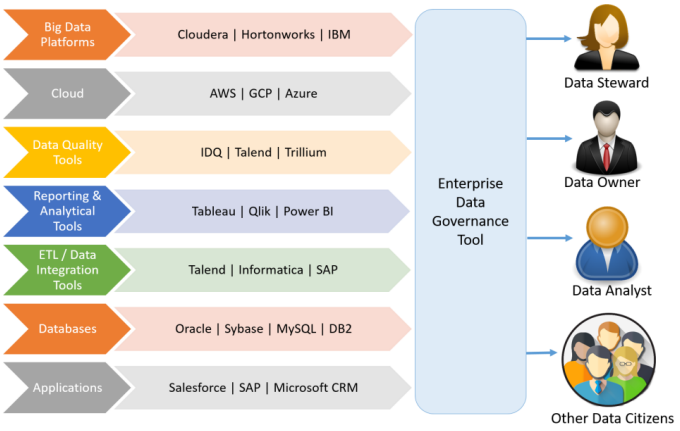


Global Data Catalog

The choice of data catalog belonging to the data lake depends upon its implementation. If an organization is using Cloudera based implementation than Cloudera Navigator (https://www.cloudera.com/products/product-components/cloudera-navigator.html) is a natural choice up till version CDH 6 but in Cloudera Data Platform (CDP) (https://www.cloudera.com/products/cloudera-data-platform.html), Apache Atlas (https://www.cloudera.com/products/open-source/apache-hadoop/apache-atlas.html) would take its place.

If data lake is implemented on AWS using its native services than Glue Data Catalog (https://docs.aws.amazon.com/glue/latest/dg/components-overview.html) serves the purpose of data catalog which is integrated with all the other services like AWS EMR, Glue ETL, RedShift, and Athena.

Similarly, on GCP, Google Data Catalog (https://cloud.google.com/data-catalog) is there which is integrated with DataProc, PubSub, GCS, and Big Query.



Enterprise Data Governance Integration

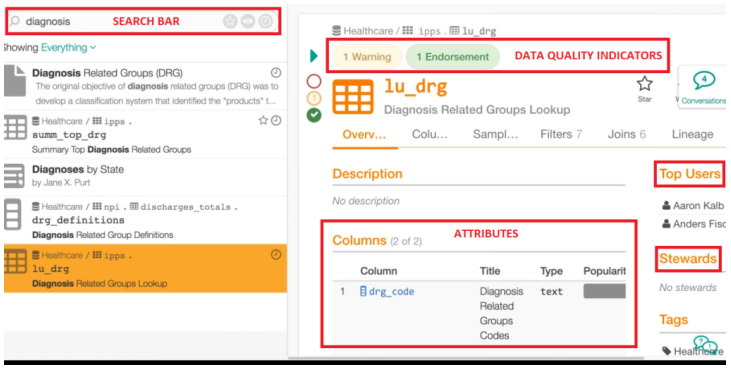
Tools like Collibra, Alation, etc also provide the data cataloging capability and they can be used as Enterprise Data Catalog service. Collibra provides a nice integration with Cloudera Navigator and Atlas. It has provided a connector to integrate with Cloudera Navigator (https://marketplace.collibra.com/listings/cloudera-navigator-to-collibra-integration-m4/) and connector to integrate with Atlas (https://marketplace.collibra.com/listings/apache-atlas-hortonworks-to-collibra-integration-m4/) as well.

Collibra Connectors

Collibra also provides connectors to integrate with AWS cloud services like AWS Glue Data Catalog, S3, Athena, Dynamo DB and RedShift. Similarly, it provides connectors for GCP services like Big Query, GCS & Big Table either through API integration or through crawling metadata store / hive metadata store. Similar connectors are available for Azure services like Azure Data Catalog.

Alation is capable of crawling Hive metastore, Databases, Data Warehouses, etc, and build a common data catalog.

Below is the sample of how a typical data catalog looks like. It allows data citizens to search over various datasets and allow them to see relevant information like Data Quality ratings, Endorsements, information about various attributes belonging to data, Data Owners, Data Stewards and Data Users.



Sample Data Catalog on Alation

Data Lineage

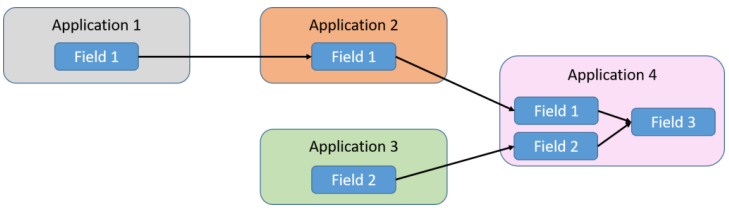
Data lineage provides the answers to questions like *where did this data come* from and *who else uses it*.

There are various ways to populate data lineage information, such as from SQL / Stored Procedure, ETL tools, and frameworks for e.g. Apache Spark creates a DAG which is sort of lineage information. Lineage information at its very basic core has data elements/nodes connected to each other through logic/transformation. These nodes can then be linked back with the cataloged nodes.

Lineage helps organizations connect different systems and processes to provide a complete picture of how data flows across the enterprise.

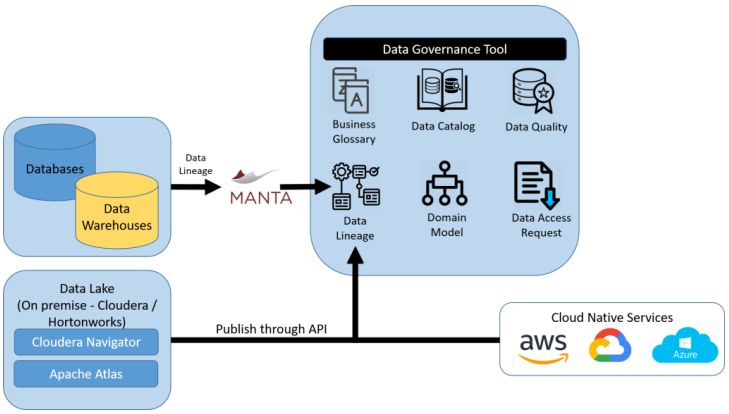
In short, cataloging enables discovery at rest, while lineage shows how it got there and where it goes from there.

Data lineage reveals how data is transformed through its life cycle across systems, applications, APIs and reports. It automatically maps relationships between data to show how data sets are built, aggregated, sourced and used, providing complete, end-to-end relationship visualization. This helps in developing a better understanding of data, enhanced trust, and providing sharper business insights. It even enables impact analysis at a granular level — columnar, table, or business report — of any changes to downstream systems.



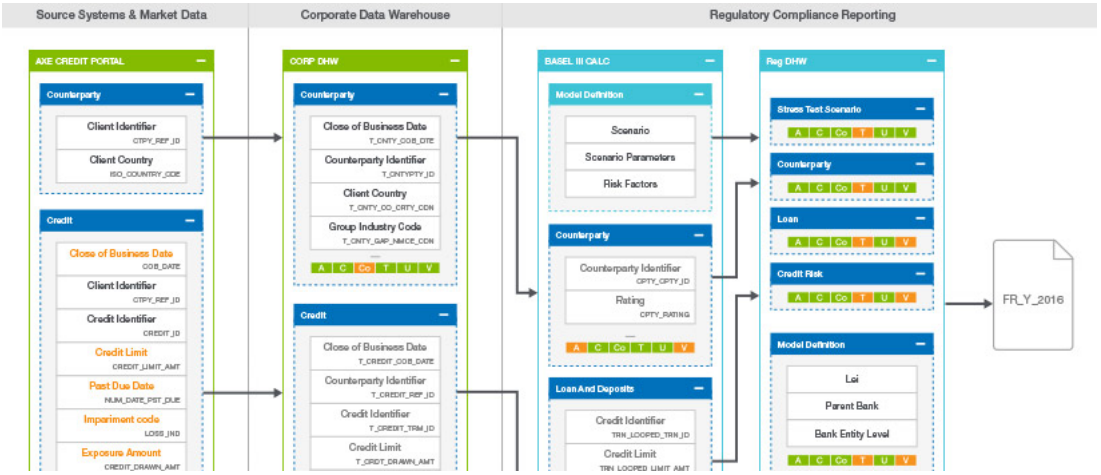
Similar to Data Cataloging, here also we need a service/tool to build Global Data Lineage which can get lineage information from different data stores including data lake as well.

A tool like Manta (https://getmanta.com/) is capable of fetching lineage information from Databases, Data Warehouses and it can further send this lineage information to Enterprise Data Governance tools like Collibra or Alation or Ataccama (https://www.ataccama.com/product/metadata-management-and-data-catalog) so that lineage information can be linked with Business Glossary, Data Catalog, Data Quality and profiling reports.



Most of us explore things visually. And having a simple means of exploring the data and its relationships using visual exploration allows a natural path for users to get insight into data, and leads data citizens quickly to actionable analysis.

By exporting the lineage information from across the data sources / ETL & BI tools helps us in building quick insights



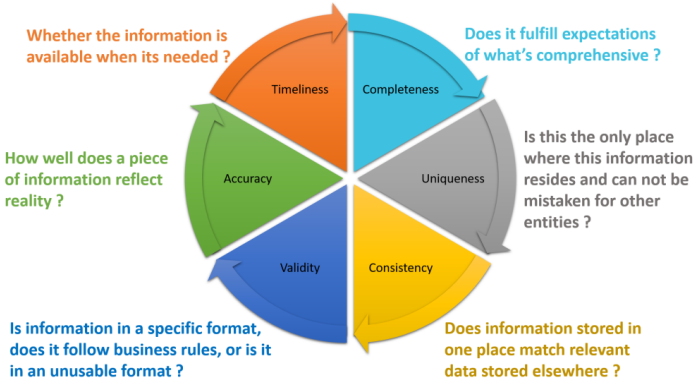
Data Lineage visualization in Collibra

Data Quality & Profiling

By now we have seen how valuable data can be, how it can provide useful insights to an organization and help them in taking better decisions. But a bad quality is not that valuable/useful.

> Data Quality relates to data completeness, accuracy, consistency, timeliness, uniqueness, data masking, and standardization. It ensures all these attributes are applied and data is correctly classified before making it available for usage.

There are mainly 6 attributes that need to be addressed while deciding the data quality score for any dataset.



Attributes of Data Quality

For data lakes, it becomes more important as we may get an exception during ETL or during querying the data because of bad quality data which could result in wastage of computing power and time.

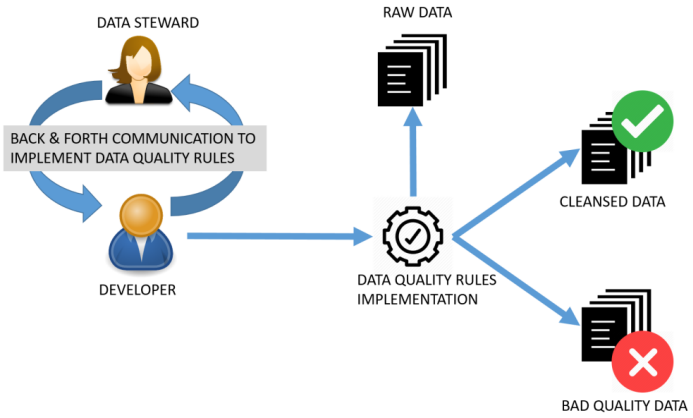
Consider the scenario where you have implemented an ETL job to process a huge volume of data and on a bad day your job fails because of a single bad quality record. It becomes really difficult sometime to pinpoint the bad record which is causing this failure.

Similarly, imagine the situation where issues could come while business users are running their huge queries and trying to prepare some report/analysis which needs to be shared with regulators before the end of the day. This would put everybody on the burner.

Data Quality should be part of ETL / ELT data pipelines so that these challenges could be addressed in a timely manner. Data Quality checks should run once the data has been acquired/landed in the landing directory of the data lake.

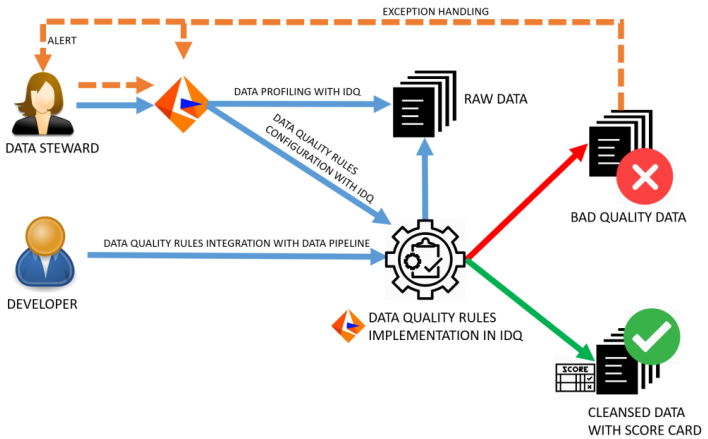
For data lake kind of big data solutions, depending upon skill set available with the Organizations, they could design & develop their own Rule-Based Custom Validation/DQ Framework. They could leverage the scalable distributed computing power of Apache Spark / Flink etc.

Mostly data quality rules are business-specific. Its the responsibility of data stewards or Business Analysts to specify business quality rules. IT developers take the requirements from Business Analysts and implement those business quality rules based on their interpretation. This process is prone to errors and also causes delays.



DATA QUALITY PROCESS WITHOUT IDQ

There are Data Quality Tools like IDQ (Informatica Data Quality) which provide UI interface for data profiling and configuring data quality rules. Data Stewards could easily use that UI interface to configure data quality rules. Developers can integrate those quality rules with the data pipeline by calling some API. IDQ could run these rules over its own compute engines or it could run these rules as Spark job on the same data lake environment/cluster where ETL / ELT data pipelines would be running.

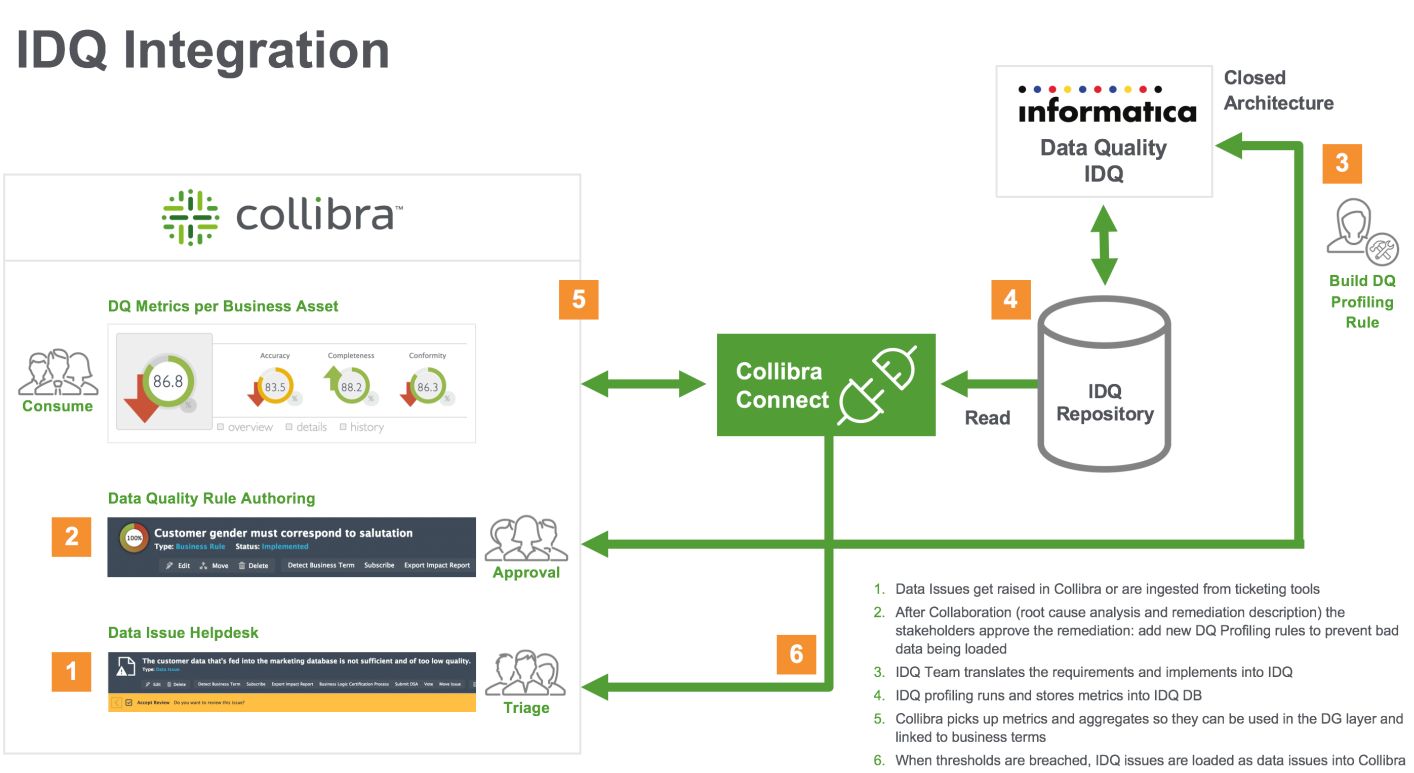


Data Quality Process With IDQ

IDQ has the capability of Exception / Case management and raises alerts if the number of data quality errors are above some configured threshold value.

Data Quality metadata like Data Quality Score could easily be imported into Data Cataloging / Data Governance tools like Collibra. Alation exposes some API through which data quality status of a particular dataset could easily be set.

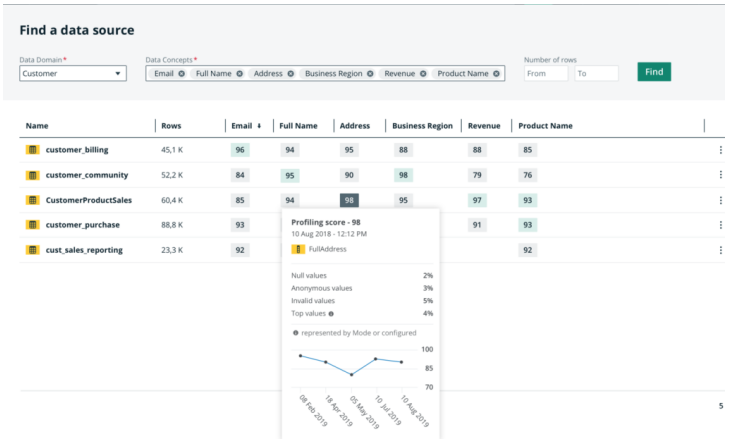
Other than that Collibra also provides direct integration with IDQ by the means of a connector (https://marketplace.collibra.com/listings/informatica-data-quality-integration-m4/).



Collibra-IDQ Integration

Once the data quality scores are made available to Collibra or Alation these tools could show data quality status at the time of finding datasets through the data catalog. Based on their requirements analysts could decide which data set they want to use for their analysis/reporting work.

For e.g. in the below data sets if an analyst wants to do some analysis around customer address than “CustomerPurchaseSales” seems to have better data quality and should be used for analysis purposes.



Data Quality Scores in Collibra

Access Management

Till now we have seen how properly governed data could help business people in finding relevant datasets by means of a business glossary, cataloging, lineage and comparing data quality scores. But to carry out their analysis they also need to get access to this data. This is like checking out your shopping basket and doing payment.

At this point, the Data Governance tool should enforce identity and access management rules. The purpose of Access Management in the context is to filter, remove, mask or in some other way enforce the access policies described with Governance, Data Use Agreement and Privacy & Risk policies established within the organization.

We can see how properly governed data allow data citizens to access an absolutely pristine data set for their analysis work. No coding. No begging for help. No artificial limits or obstacles. They get the best data the organization has to offer that they have the authorization to use. That is what we call **true data democratization**.

Disclaimer — Views presented in the article are author's own and does not reflect views of any organization author is associated with.