WHITE PAPER

INTELLIGENT BUSINESS STRATEGIES

Organising The Data Lake

The Central Role Of An Information Catalog

By Mike Ferguson
Intelligent Business Strategies
April 2017

Prepared for: Waterline Data

Table Of Contents

DATA DELUGE - THE EVER INCREASING DEMAND FOR DATA

The fight for business survival is increasingly dependent on data and insights

In today’s enterprise, the fight for business survival and business growth has never been so intense. Companies are doing everything in their power to compete by producing more comprehensive insights about their customers, their business operations, their risks and their finances in this rapidly moving digital world. Of course insights can only be produced with data and in this increasingly self-service world, the thirst for new data to analyse has skyrocketed in just about every corner of the enterprise.

The demand for new data has skyrocketed as companies seek to produce new insights from analysis

The number of data sources is exploding

Data of all varieties is being consumed and analysed

To quench this thirst and produce the insights needed, new sources of internal data and external data are being ingested to add to what we already know. Popular new data sources include clickstream data from web server logs, in-bound customer email, external data from open government and weather data. Sensors are being deployed in production lines, supply chains, and in assets and products, all in the name of producing insight to optimise operations and understand product or asset usage. The demand is so intense that the amount of data and the number of data sources is exploding. It is without doubt a data deluge and many companies are struggling to cope with it. Data of all varieties is being consumed and analysed, making it harder to track what is going on. Data volumes are on the rise and in some cases have already reached into hundreds of terabytes or petabytes. Also the rate at which data is arriving (velocity) has accelerated to millions of events per second in some cases. All of this is underway right now.

There is a hive of activity underway

But how is all this being managed and who is tracking what is going on across all data stores and locations?

But if you take a step back, you could rightly ask several questions and struggle to find answers. For example: How can you track all this activity? What data is being collected? Who is collecting it? Where is it being collected? How do we know what to govern? With so many data sources and data being ingested everywhere, how can you profile all this data manually? Who owns what? Who is producing what data and what insights? How do we know what is going on so that we can bring order to chaos in the rapidly moving digital world?

Information catalog technology holds the key to staying in control of an ever increasing data landscape

This paper looks at this problem and explains why we need to invest in newly emerging information catalog technology to stay in control of the ever increasing data landscape. It looks at how information catalogs enable us to organise and rapidly discover new data, track what data and insights are being produced and publish these as services so they are easy for others to find and consume to help deliver business value across the enterprise. We will look at the key requirements for this technology, the approaches as to how information catalogs work and then look at how one vendor—Waterline Data—is stepping up to the challenge of helping to manage and maximise the value produced from this newly emerging distributed data landscape.

THE CHANGING DATA LANDSCAPE – MULTIPLE DATA STORES IN A DISTRIBUTED DATA LAKE

The practices and architecture needed to build data warehouses to query and analyse data are now very well understood and widely deployed

It is now more than twenty-five years since the emergence of the enterprise data warehouse, and for most companies today, the architecture required and best practice approaches needed to build one have become very well understood. Master data and transaction data are extracted from source systems, after which data is cleaned, transformed and integrated into a target data warehouse system. Subsets of that trusted and integrated data are then moved into dependent physical data marts or made available through virtual data marts, both of which are specifically designed for analysis and reporting.

However, new data is now in demand that is not well suited to data warehouses

New analytical workloads have emerged across multiple analytical data stores beyond the data warehouse

Over recent years however, the demand to analyse the aforementioned new types of data has caused significant changes to the analytical landscape. The requirement now is to deal with high volumes of structured, semi-structured and unstructured data from internal and external sources together with new types of analytical workloads needed to derive insight from analysing these types of data. As a result, new data stores and analytical platforms have now emerged beyond the analytical RDBMSs that underpin our data warehouses. These include cloud storage, NoSQL column family databases and key value stores capable of ingesting data rapidly, NoSQL graph DBMSs for graph analysis, Hadoop, and streaming data analytics platforms. All of these are now in play. The result is that companies are now creating *analytical ecosystems* consisting of multiple data stores that still include the traditional data warehouse. This is shown in Figure 1.

Companies are now creating analytical ecosystems consisting of multiple data stores in addition to the data warehouse

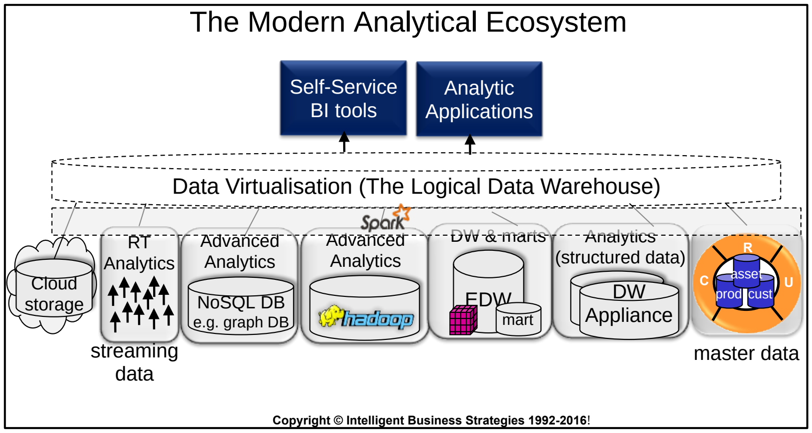


Figure 1

Setting aside the data virtualisation layer to hide complexity and simplify data access, the fact of the matter is that data is now being ingested into multiple platforms for preparation and analysis by analytical applications and tools.

Complexity has increased with multiple analytical data stores on–premises and in the cloud and data of multiple types being ingested into all of them

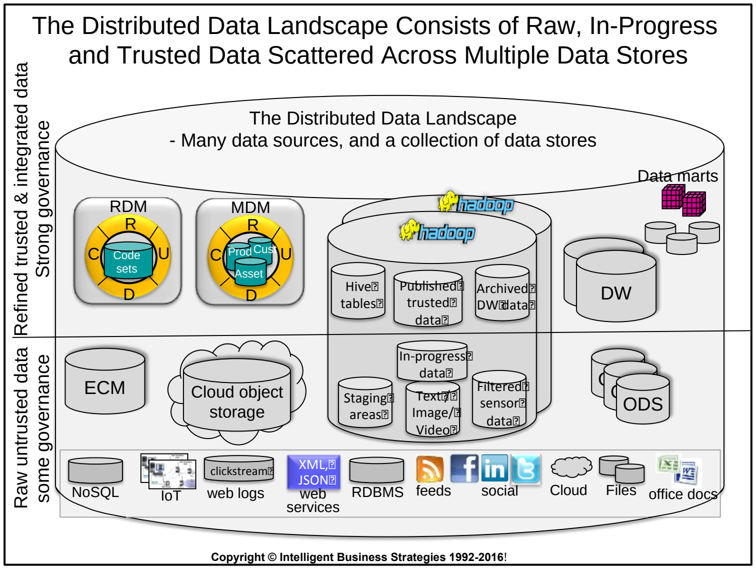
The point here is complexity has increased. There are many more data sources. Multiple analytical data stores exist. Data from these sources is being ingested into analytical platforms both on-premises and on the cloud. Multiple types of data exist and multiple types of analytical workload are going on in real-time on streaming data, on data held in cloud storage, on Hadoop and on traditional data warehouse systems.

Problems With The Current Approach

Despite the hype about a single Hadoop system becoming a centralised data lake, many companies now have a distributed data lake with multiple data stores

This may include multiple Hadoop systems, relational DBMSs, NoSQL data stores and cloud storage

Looking at this, it doesn't take long to realise that despite the hype about a single Hadoop system becoming a centralised data lake, the data landscape has and is becoming increasingly more distributed as shown in Figure 2. In some companies it already includes multiple MDM systems (e.g., for different domains), multiple operational data stores, multiple data warehouses, multiple data marts and *multiple* Hadoop systems.


The Distributed Data Landscape Consists of Raw, In-Progress and Trusted Data Scattered Across Multiple Data Stores

Copyright © Intelligent Business Strategies 1992-2016!

Figure 2

With data in multiple data stores it is hard to know what data is located where

Data relationships across multiple data stores are unknown

We also often don't know what kind of processing is going on and what jobs and analytical models exist to prepare and analyse it

In this kind of set-up, data resides in multiple data stores, in various states of quality and it is very hard to know what data is located where. Also, subsets of data may be duplicated within or across data stores. For example, multiple versions of a file may exist in the Hadoop Distributed File System (HDFS), multiple copies of master data in data warehouses, data marts and in Hadoop, etc. We don't know all the relationships that exist in the data. We don't know what data is being prepared and analysed let alone who is doing it and where it is happening. We also don't know what scripts, ETL jobs and/or self-service data preparation 'recipes'[1] are being used to work on data across these systems and what analytical models exist to analyse it.

Some would say the solution here is straightforward: create a centralised data lake with all data in one Hadoop system. However, in reality, it is not always that simple. For example, many companies already have several Hadoop systems. Also, once you have captured certain kinds of data, it may be too big to move to a single central location, for example, sensor data captured in the cloud. Also, even if that is not the case, some data, once it has been captured, may not be able to be moved for legal reasons because doing so would result in violating compliance with legislation in the specific jurisdiction in which it was captured.

[1] The term 'recipe' is often used by self-service data preparation vendors to describe the steps taken by a data scientist or business user to prepare data for analysis

Siloed Approaches To Managing And Governing Data

Our attempts to manage and govern data over the years have resulted in silos

Data warehouses, MDM systems, RDM systems, streaming platforms and Hadoop

So what have we done about this problem so far to help manage and govern data in this increasingly distributed data landscape? Several advances have occurred over the years. We have built data warehouses. We have built master data management (MDM) and reference data management (RDM) systems to clean, integrate and standardise master and reference data. Now we are also preparing data for analysis on big data platforms and streaming data platforms. We are also using Hadoop as a staging area and offloading ETL processing from data warehouse. However the problem is that to date we have taken a stand-alone project-oriented approach to these initiatives and have inadvertently created silos (see Figure 3) with different data cleansing and integration tools in use in each silo.

Different tools for data management have often been used in each of these silos

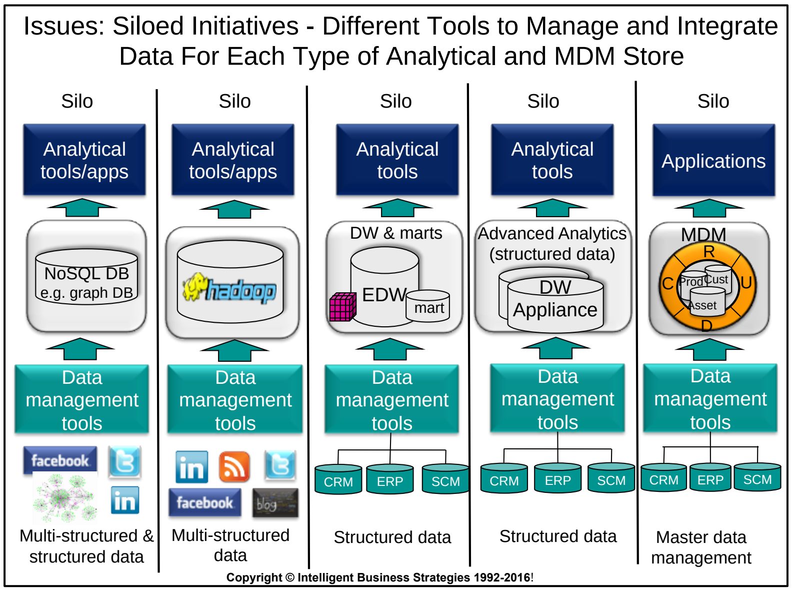


Figure 3

The impact is that costs are too high, agility is poor, skills are spread thinly, metadata is fractured and maintenance is complex and time consuming

The impact of this siloed approach is that:

- Siloed data sets diverge, introducing inconsistencies into the analytics and subsequent decision making process. It is quite common to have boardroom debates about which numbers are more trustworthy – sales, marketing or finance's; HQ or business units; and so on.
- The cost of data cleansing, integration and governance is much higher than it should be
- Multiple IT and self-service data integration and quality (DI/DQ) technologies and techniques are being used that are not integrated in any way
- Multiple approaches may be in use on a single platform like Hadoop
- Skill sets are fractured across different projects and technologies
- Speed of development is slow because so many different technologies and approaches are being used
- Maintenance is complex, time consuming and costly
- Fractured metadata exists across tools or no metadata at all in some cases

Lots of re-invention is occuring rather than reuse

There is no catalog to document the existence of data across systems and to classify it to allow it to be governed

Data is not organised to make it easy to understand and there is nowhere that documents the existence of artefacts in use across the landscape

- Metadata lineage is unavailable when data is being transformed by multiple different tools and custom analytical applications
- Lots of re-invention is occurring rather than re-use
- Inconsistent DI/DQ rules may be in use for the same data
- There is no central repository or catalog to document the existence of data in each of these systems and no tagging mechanism to classify the data to allow it to be governed consistently anywhere it is stored
- There is no place where you can go to see the data quality profile of data across these data stores
- There is no place where you can go to see the existence of ETL jobs, self-service recipes, scripts, analytical models, interactive notebooks, custom applications, etc. in use across this landscape
- Data is not organised across this landscape to make it easy to understand what stage, in terms of data preparation, the data is at. Is it raw data? Is it data ‘in-progress’ going through preparation? Is it trusted data or trusted insight ready for consumption?

THE STRUGGLE TO FIND DATA FOR ANALYSIS

There is no common place to tell you what data is available, what data preparation jobs and what analytical models exist as a service that you can reuse

The problem here is that there is no common place for anyone to go to tell them what data, what ETL jobs, what data preparation recipes, what analytical models, etc. are available for use and where they are. As a result, data scientists and business analysts often struggle to find the data they need for analysis. They could therefore potentially ingest and pay for the same data multiple times even if it has already been brought in somewhere else in the enterprise. They have no idea if the data they need for analysis has already been prepared by someone else and so could spend a lot of time and effort repeating the exercise for no reason. They often have nowhere to go to help them know what data already exists and what they can reuse.

In fact, the security infrastructure around data siloes prevents them from being able to find data they need even if they tried looking. Usually analysts only have access to ‘their silo’ and have no way of finding or leveraging data that exists in other silos. It is a catch 22: they can’t find data without access to a silo, but they can’t access a silo without specifying what data they need and requesting access to it.

THE ADDED CHAOS OF UNMANAGED SELF-SERVICE DATA PREPARATION

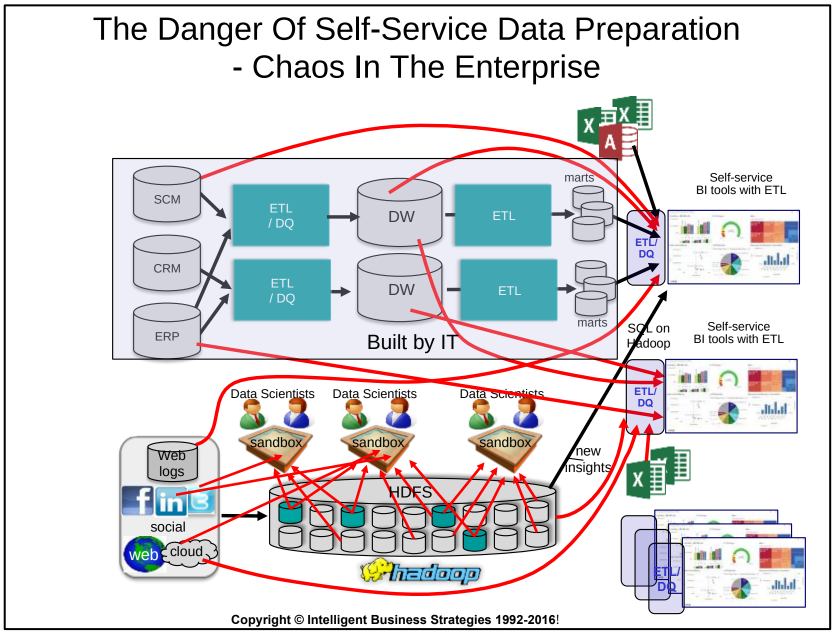
Also IT is no longer able to keep pace with the demand for data

Nowadays both IT and self-service business users are preparing and integrating data

Furthermore, the impact of the data deluge described earlier is that it is unrealistic to assume that IT has sufficient resources available to keep pace with business demand for new data and do all the ETL processing on all data on behalf of the business. Subsequently, business users are demanding access to self-service data preparation tools to do their own data cleansing and integration. This is not to replace IT efforts but to augment them. Self-service data preparation is already available from existing enterprise data management tool suite vendors and new self-service data preparation vendors also exist. In addition, BI tool vendors have recently added self-service capability.

The problem with self-service data preparation is that it can very quickly degenerate into chaos when there is no governance (Figure 4). Here, users of stand-alone self service data preparation tools or self-service BI tools with built-in data preparation capability can connect to many data sources, access data and start transforming and trying to integrate data.

Self-service data preparation environments can very easily become chaotic if not governed


The Danger Of Self-Service Data Preparation - Chaos In The Enterprise

Copyright © Intelligent Business Strategies 1992-2016!

Figure 4

UNCERTAINTY AROUND TRUST AND DATA CONFIDENCE

If everyone is doing their own self-service data preparation then are the metadata specifications each user defines shared?

A key question here is that if each user is doing their own self-service data preparation, then where are all the metadata specifications stored that they define to transform, clean and integrate data? Are they held separately for each user? Can metadata be shared? If there is no way see what data has already been prepared and to see the metadata lineage to show how it was prepared, then the chances are very high that users will not trust the data. They will most probably chose to re-invent rather than re-use data preparation jobs simply because there would be uncertainty around trust in the prepared data and the confidence to use it.

Business users have no place to go to find out what trusted, prepared and integrated data already exists

IT also don't know what data business users have prepared

The issue here is that the business user has no place to go to find out what trusted, prepared and integrated data already exists, nor to tell for a specific data set where it came from and what was done to prepare it. Equally there is no place that IT can go to see what data business users have already prepared. The bottom line is that there is a real need for some kind of centralised service like an information catalog that publishes what data is available across an organisation as well as its current level of data quality. It also needs to publish the existence of ETL and self-service jobs and services used to produce trusted data.

BRINGING ORDER TO CHAOS - THE NEED FOR AN INFORMATION CATALOG

The answer to managing all these issues is to establish an enterprise information catalog

Having identified the issues with current approaches to managing and governing data, the question is: how do you bring order to the chaos? If the data is distributed as in Figure 2, could such a data lake be managed, governed and function as if all the data was centralised?

Information catalog technology plays a central role in organising and governing a data lake

You can see what data and artefacts exist across multiple data stores both on-premises and in the cloud

The answer is that it can be done if an enterprise information catalog is established. This technology plays a central role in organising and governing a data lake because its job is to know about all data and data relationships across the distributed multiple data store environment. It should know about registered data sources and ingested data. It should be able to tag data to know what it is, and what it means, and offer multiple data classification schemes[2], each with different levels of classification, so that data can be labelled in a way that makes it obvious as to how to govern it. In addition, the enterprise information catalog also needs to know about the policies and rules that must be enforced to govern data classified and tagged with specific levels of retention, confidentiality, quality (trustworthiness) and value. All of this metadata is stored and accessible in the information catalog.

A key capability of an information catalog is its ability to automatically crawl, discover and tag data in one or more data stores

Information catalogs should provide a comprehensive set of capabilities to manage and govern the distributed data landscape

Learning from the failure of static metadata repositories that were outdated almost as soon as they were populated, a powerful and necessary capability of the catalog is its ability to automatically crawl, discover and tag data in one or more data stores and dynamically keep the catalog up to date. (figure 5).

The Information Catalog Crawls The Data Lake To Discover, Tag And Inventory Data In RDBMSs, Hadoop And Cloud Storage

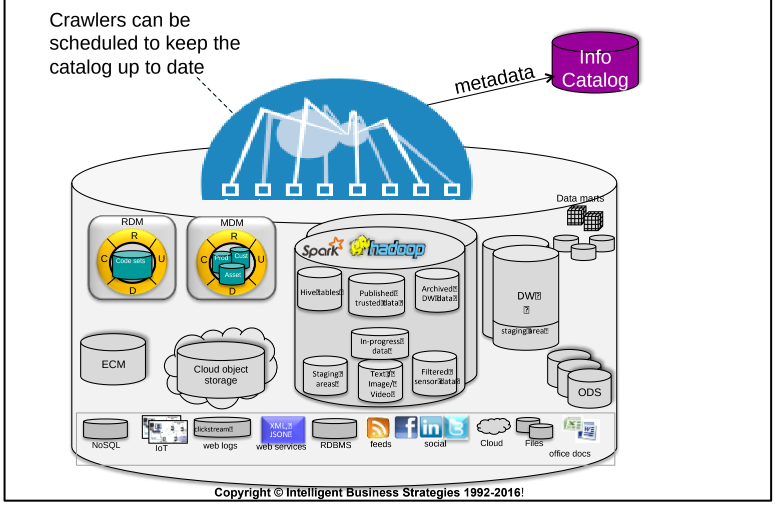


Figure 5

However this is just one critical capability. What other capabilities should an information catalog provide? Let's take a look in the next section.

[2] For example data confidentiality, data retention, data trustworthiness, and business value classification schemes, each of which would have different levels, e.g., levels of confidentiality ranging from public data to top secret data.

KEY REQUIREMENTS FOR AN INFORMATION CATALOG

The enterprise information catalog has a number of key requirements. Below is a very comprehensive list of capabilities that you should evaluate for relative importance for your organization's situation and needs.

Data Discovery

Authorised users should be able to nominate and register new data sources in an information catalog

- Allow authorised users to nominate/bookmark and register data sources containing data of interest in the catalog so that it can be brought into a data store included in the distributed data lake

Automated data discovery, data profiling, tagging and creation of data lineage is a core requirement because there is just too much data to do all this manually

- Automatically discover ingested data in multiple different data stores within a data lake as seen in Figure 5. This would include RDBMSs, Hadoop, cloud storage and NoSQL databases. During discovery it should be possible to:
  - Automate data profiling to understand the quality of every item
  - Automate discovery of the same, similar and related data across multiple data stores regardless of whether the data names for these data are different
  - Automate field level tagging (naming) and annotation
  - Automate derivation of data lineage

Manual as well as automated tagging

- Allow users to manually tag data to introduce it into the catalog

Collaboration

Support virtual communites and roles to enable collaboration to flourish

- Allow virtual communities to be created and maintained to allow people to collaborate over data and other artefacts, either yet to be published or already published in the catalog.
- Create roles within the catalog that users can be assigned to, e.g., data owners, data experts, data curators/producers, data stewards, approvers, catalog editors, consumers

Collaborative governance and stewardship

- Curate, collaborate over, and manually override tags automatically generated by the software
- Allow nominated users to collaborate over, vote on and approve data names either created by automatic software tagging or created manually by authorised users

Data Governance

- Support import of business glossary terms and ontology vocabularies to be used as semantic classification tags

Automated determination of the meaning of data items

- Support automatic semantic tagging at field, data set, folder, database and collection level.

Automated classification of data to know how to govern it

- Support multiple pre-defined data classification (tagging) schemes to classify data in order to help organise and govern it. Examples of such schemes would be those that indicate levels of data confidentiality, data retention, and data trustworthiness (quality)

- Add user defined data classification (tagging) schemes to classify data in order to help organise and govern it.

Extensibility with user defined and 3rd party classification schemes

- Add 3rd party classifiers to the catalog to extend it to enable support for new data types and sources
- Automate data classification by making use of pre-defined patterns, user defined patterns (e.g., regular expressions or reference lists) to automatically identify and classify specific kinds of data in a data lake

Pre-defined and user defined patterns to enable automated data identification

- Automate data classification by making use of analytics to observe, learn and predict the classification of data in a data lake

Crowd sourcing

- Allow manual tagging to crowd source classification of data and other artefacts in the catalog to organise them in order to govern and manage the data lake. It is critical to allow users to tag data with the terms they are used to, so they and their peers can find it again using those terms

Automatically detect and tag sensitive data

- Automatically detect and classify or manually classify sensitive data (e.g., personally identifiable information - PII) in the catalog to enable governance of that data via tag-based data privacy and security policies
- Allow multiple governance and use tags to be placed on data including:
  - A level of confidentiality tag, e.g., to classify it as PII
  - A level of quality tag
  - A level of data retention tag
  - A business use tag, e.g., Customer engagement, risk management, etc.
  - Tagging a file to indicate its retention level or which processing zone it resides in within a data lake, e.g., ingestion zone, approved raw data zone, data refinery zone, trusted data zone

Multiple tagging schemes

Tagging of files and individual data items

- Automatically propagate tags by using machine learning to recognise similar data across multiple data stores

Attach policies to tags and tags to data to know how to govern it

- Define, manage and attach policies and rules to tags to govern any data published in the catalog that has been classified using those tags

Establish a business glossary for common data names

- Nominate data items within the catalog to be used in the creation of a business glossary to establish a common vocabulary for the enterprise

Generate schema for discovered data to make it easily and quickly accessible

- Derive and generate schemas from discovered data to enable that data to be easily and quickly accessed by authorised users via 3rd party tools. Examples here include:
  - The generation of Hive tables on discovered data in Hadoop
- The generation of virtual tables in data virtualization servers on data automatically discovered across multiple data stores in a distributed data lake

Governance of Artefacts

- Import the metadata from 3rd party tools to automatically discover, classify and publish
  - IT-developed ETL jobs, self-service data preparation 'recipes' in the catalog

- BI tool artefacts (queries, reports, dashboards), analytical models and data science notebooks in the catalog

to understand what is available across the distributed data lake to prepare, query, report and analyse data held within it

Automatically discover metadata in other tools to catalog existing artefacts across the landscape

- Manually classify (tag) and publish IT developed ETL jobs, self-service data preparation 'recipes', BI queries, reports, dashboards, analytical models and data science notebooks to the catalog

Findability

Create easy to use marketplaces in the catalog to publish data and insight as a service

Faceted search across the data landscape to make it easy to find data and artefacts

- Create 'marketplaces' within the catalog to offer up data and insights as a service to subscribing consumers
- Support faceted search to zoom in on and find data and other artefacts published in the catalog that a user is authorised to see
- Allow search facets to include tags and properties, lineage (e.g., only data from DW or derived from DW data), data set location and format
- Allow users to easily find data in the catalog, select it and then launch third party tools such as data visualization or self-service data preparation tools to work on that data if authorised to do so
- Allow users to easily find:
  - IT-developed ETL jobs and business-developed self-service data preparation 'recipes' in the catalog and to launch or schedule them to provision data
  - BI queries, reports and dashboards in the catalog and to launch tor schedule them to provision insights
  - Interactive data science notebooks and analytical models in the catalog and launch or schedule them to refine and analyse data to provision insights
- Allow users to search metadata to find data sets that they are not authorized to use and request access to those data sets
- Understand relationships across data and artefacts in the catalog to make recommendations on related data
- Allow consumers to collaborate over and rate data and artefacts within the catalog in terms of quality, sensitivity and business value

Trust

See end-to-end lineage and data profiles for trusted data

- Allow users to easily see lineage from end-to-end in technical and business terms and navigate the lineage graph to explore related data
- Provide data profiles to help users decide whether it meets their needs
- Be able to easily distinguish between disparate and common business vocabulary data names in the catalog

Scalability

- Scale automatic discovery processing and profiling using scalable technologies like Apache Spark to deal with the volume of data in multiple underlying data stores

Integrate with other tools to make it easy to publish into the catalog and find data in the catalog

Integration

- Integrate the catalog with other tools and applications via REST APIs

Catalog Approaches – What They Offer And How To Use Them

While this is a very comprehensive set of requirements, no product yet supports all of this today. Instead we are seeing information catalog products adopt different approaches, which makes them more suitable to some uses and capabilities than others. Key approaches being used include:

Different information catalog technologies use different approaches to catalog data and artefacts

- Automated discovery of data in files, databases and other data stores
- Automated discovery of metadata in databases and files with crowdsourcing
- Automated discovery of artifacts across multiple data management and BI / Analytical tools
- Automated discovery of query activity
- Manual cataloging by data stewards

Some products use a combination of these techniques

Some products use a combination of these techniques. Let's look at these in more detail.

Automated Data Discovery Across Multiple Data Stores

Automated data discovery is fundamental to understanding the meaning of data in a data lake and knowing where it is stored

Data can be automatically profiled to unserstand its quality

Automation is a fuldamental part of remaining agile as the data landscape increases in complexity

Automated discovery and search indexing of data in multiple data stores is the core of any information catalog. This looks at each and every data item in an attempt to automatically work out what the data is and what it means. Pre-defined patterns, user-defined patterns and analytics help automatically identify what the data is, how to name it, how to classify it and if it relates to other data already catalogued. Automated indexing also allows facets to be created to search for data that is published in the catalog. This approach is fast and ideal for a data lake where data is often ingested from data sources that are unknown or not well understood. This approach also means that mapping disparate data to common business vocabularies is automated and not dependent on crowdsourcing to understand that several data items discovered all in fact are the same data despite having different data names. It also allows semi-structured data to be discovered and potentially even structured data to be automatically derived from unstructured data (e.g., using text analysis entity recognition) during the discovery process. Once discovered, teams can of course collaborate over tagged data and override or approve what has been done automatically. A major advantage of this approach is that the information catalog has utility almost immediately if the automated tagging is reasonably accurate because the catalog has a critical mass of tagged fields and files without requiring any manual effort.

Automated Metadata Discovery Across Multiple Data Stores

Automation metadata discovery does not look at data itself and relies on people to resolve data ambiguities

This approach differs from automated data discovery because there is no attempt to look at the data, just the metadata (assuming it exists). The problem here is that the data may have no metadata (e.g., an unlabelled CSV file in HDFS) or the same piece of data exists in multiple data stores with different data names. If this occurs, then it relies on people to indicate the meaning of the data or that multiple data items have the same meaning. It is not automatic as in the case of automated data discovery. It relies on crowd sourcing. In addition, automated data profiling is not possible using the catalog but is doable via separately purchasable stand-alone data profiling tools. The issue with this approach is, of course, how good is the metadata? Metadata is often cryptic or misleading. Metadata ingested from legacy systems sometimes has to conform to a 4, 6 or 8 character format and is virtually impossible to decipher. Fields are frequently overloaded to store data not reflected in their name, e.g., taxid field may be used to store passport numbers for non-citizens. Even if there is metadata and it reflects the data, often the terms used are either ambiguous (e.g., id or name or amount) or simply not what the business users would use (e.g., customer_id instead of account number). The more data is unlabelled or mislabelled in a data lake, the more crowd sourcing is needed to gain an understanding of what the data means. And the more manual crowdsourcing required, the longer it takes to get your catalog to a critical mass of usefulness. However if unlabelled files in systems like Hadoop have Hive tables projected on them, then it would be possible to understand the meaning of the data from the Hive tables defined, since this is metadata that can be accessed. Automated discovery would not be possible without the Hive metadata. Also manual initiation of 3rd party data quality profiling may be needed to understand data quality. This may slow down the ability to get data under control in a data lake.

AUTOMATED DISCOVERY OF ARTEFACTS ACROSS MULTIPLE TOOLS

Automated discovery of ETL jobs, self-service data preparation jobs, BI reports, dashboards, predictive models and data science notebooks helps identify, document, govern and organise what is available for reuse in a distributed data environment

This approach is not about data. It is about automatically cataloguing ETL jobs, self-service data preparation recipes, BI reports, queries, dashboards, predictive models and data science notebooks that exist across the data landscape so that they can be tagged and published in an information catalog. It is typically done by crawling the metadata in other tools to understand what already exists. It also means that people will not re-invent them and will more likely reuse them. Again a lot will depend on the software product here as to what metadata connectors it has and what level of automated artefact discovery is possible. This will work well across data warehouses and data marts and is also needed in a distributed data lake environment. This approach is needed in addition to automated data discovery to reap the full benefits of a catalog.

AUTOMATED DISCOVERY OF QUERY ACTIVITY

Automated discovery of queries can help optimise design of physical databases as it also provide evidence of where main query processing activity is taking place

Again this approach is not about data but about the use of data in queries. It is dependent on reading the log files of database servers and other query engines. It can tell you about frequency of queries and their performance. This is very useful for mature data marts, for example where interactive query processing from self-service BI tools is heavily used. Here, it could provide evidence to indicate that changes in physical database design are needed to optimise query processing at different times of the day. It would also help understand how the tables are being joined by various queries.

However, in a logical data lake environment there is no common log across multiple platforms. Even in a physical, single Hadoop cluster based data lake, the data is ingested from different systems and in many cases has never been queried, so this approach will not help find data nor join it. Even when it comes to optimization, a lot of data access via Hive or other SQL on Hadoop engines is by a limited number of data scientists who may only run SQL queries on this data a few times. In that sense it is not huge numbers of interactive query users albeit that interactive self-service BI tools can be used to access Hadoop data via Hive and other SQL on Hadoop engines. However, it is of more use in optimising the design of physical data models in systems where data is known and modelled, e.g., on-line transaction processing (OLTP) systems, data marts and data warehouses.

Manual Data Cataloging By Data Stewards

Collaborative governance and maintenance of the catalog

This approach allows data stewards to manually link business metadata to technical metadata and publish it in the calalog. This will always be needed but preferably in conjunction with other approaches like automated data discovery that significantly accelerate the process of populating the catalog with tagged data. The disadvantage of this approach of course is that it is slow. The advantage is that it enables collaborative governance and also that manual cataloging is likely to be very accurate as there is precise control on where in the catalog the data appears. This capability also allows for collaborative maintenance of the catalog itself.

Leveraging Waterline Data To Govern, Organise And Provision Your Data

Waterline Data is an Information Catalog vendor whose Smart Data Catalog product runs on-premises or in the cloud

Having understood the key requirements and approaches to cataloguing data and other artefacts to organise the data lake, this section looks at how one vendor's product helps companies achieve this. That vendor is Waterline Data.

Waterline Data was founded in 2013. They currently have customers in multiple countries worldwide. Their flagship product is the Waterline Data Smart Data Catalog which runs on-premises and/or in the Amazon Web Services (AWS) cloud.

Automated Data Discovery Using Smart Data Catalog

Waterline Smart Data Catalog can automatically discover and inventory data across multiple data stores including Hadoop, RDBMS and NoSQL databases

Waterline Data Smart Data Catalog uses a combination of automated data discovery with human review and some manual curation and tagging to enable inventory and governance of data in an enterprise data lake. Initially the scope of the product was automated discovery and cataloguing of data on various Hadoop systems. However in version 4, Waterline Data have now widened that scope to include NoSQL and relational DBMSs with cloud storage in development (expected Q2 2017). The idea is to discover data in all data stores as depicted earlier in Figure 5.

Waterline Data Smart Data Catalog crawls all or part of HDFS, relational data sources and soon cloud storage, automatically cataloguing all the data it finds. Using Hadoop or Spark as its underlying scalable engine, it inspects the data values of each and every field to automatically discover the meaning of data, at which point it can automatically tag the data to give it a name. Figure 6 gives a very high level picture of what is going on during this process with tab delimited data files and files containing JSON data.

Smart Data Catalog can crawl structured and semi-structured data

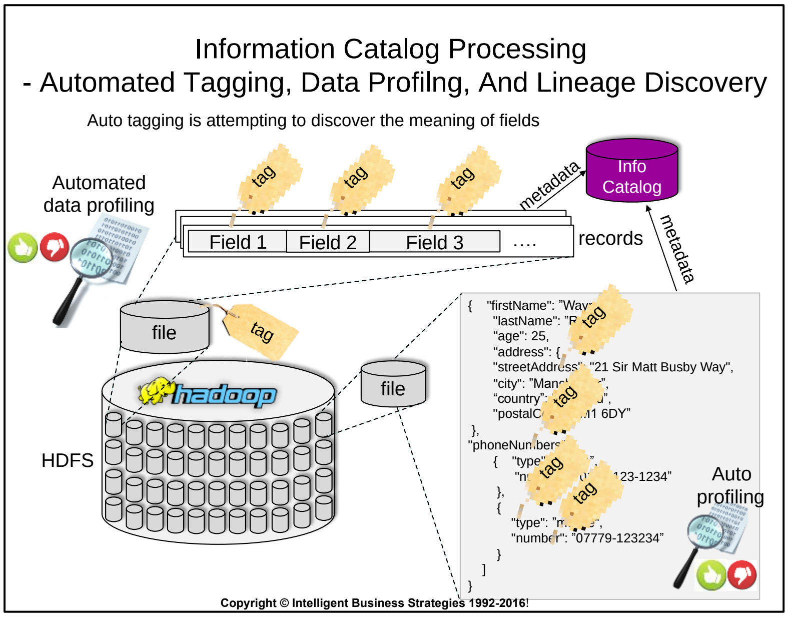


Figure 6

Automated Data Profiling At Scale

Smart Data Catalog can automatically profile data in many different file formats on Hadoop to automatically assess data quality

During this process, data is also automatically profiled at scale to assess its quality. This is done by first discovering the file format. Waterline Data Smart Data Catalog can recognise JSON, XML, RC, ORC, PARQUET, Sequence, AVRO, and delimited data files. Any schema within the files is determined before reading the data within it to create a data quality profile of it. Algorithms are used during automated profiling to calculate metrics and statistics about the data. Again this is done at scale. This speeds up the process of identifying anomalies in huge amounts of data on Hadoop and also helps improve data scientist productivity since all they have to do is look at the catalog to quickly determine data quality and ascertain what they need to do to prepare the data for analysis.

Smart Data Catalog can also automatically detect sensitive data and changes to schema in Hadoop files

Waterline Data Smart Data Catalog can also detect sensitive data during profiling so people know where that data is across all files in HDFS. This is a major help in governing at risk sensitive data. In addition, changes to schema in Hadoop files can also be detected and data scientists made aware so they can determine the impact on any existing data preparation jobs and analytic applications.

Automated Discovery Of Partitioned Data Sets

Unlike relational databases where new data is inserted into existing tables, in Hadoop, new data is usually kept in new separate files. These are called partitioned data sets. However, for both cataloguing purposes and analytical purposes, all the files in a partitioned data set should be treated as a single concatenated file. Waterline Data Smart Data Catalog automatically detects partitioned data sets and treats all the files as a single logical file during search. It provides tagging and profiling statistics both at the data set as well as individual partition level. This is probably the most common data pattern in Hadoop and if not treated correctly, would result in the catalog returning thousands or tens of thousands of partial results (partitions) instead of a single complete result.

Automated Discovery Of Lineage

Lineage is also automatically generated to help determine where data came from

In addition to automated data profiling of Hadoop data, Waterline Data Smart Data Catalog also provides a powerful file lineage discovery capability. This is extremely important when it comes to data governance on Hadoop since it makes it possible to automatically determine where data in a Hadoop file has come from. The metadata generated from this process is then available for access so that users can determine the trustworthiness of the data.

Automated Tagging

Automatic tagging is also possible to determine meaning of data being ingested into and stored in the data lake

Tags can be automatically propogated by using machine learning to recognise similar data across multiple data stores

In addition to automated data profiling and automated data lineage discovery, Smart Data Catalog can also automatically tag data to help data scientists and business analysts quickly understand its meaning. In addition, users can override automatically created tags if they disagree with what has been generated. Automatic tagging is especially important when bringing external data into the data lake. A really powerful feature here is Waterline Data Smart Data Catalog's ability to propagate tags during automatic tagging. The software does this by using machine learning to learn by observing how users tag data so that it can recognise similar content, format and other characteristics across files during the automated tagging process. This capability allows it to automatically propagate tags to similar data. This result is the ability to see similar data across the data lake simply by searching or drilling in on a tag.

Users with the right privileges can then curate these suggested tags and either accept or reject them, thereby helping machine learning algorithms to improve the accuracy of tag propagation.

Furthermore, tag management in Smart Data Catalog allows you to specify who can create tags and create user-defined rules to determine exactly when data should be labelled with a specific tag. Users can also manually tag data.

Advanced Scalable and Fault Tolerant Search

Smart Data Catalog also embeds Apache SOLR to enable near real-time indexing, scalable, fault tolerant, advanced full-text search across the entire data lake

One of the challenges with data lake management and operation is the ability to find data across hundreds of thousands of files and tables and the ability to be able to do this at scale. In order to address this problem, Waterline Smart Data Catalog embeds Apache SOLR. Built on Apache Zookeeper, SOLR also has replication, distribution, rebalancing and fault tolerance built in which makes it easy to scale up to accommodate high volumes of data. This capability enables Smart Data Catalog to support near real-time indexing to ensure that data is searchable as soon as it is ingested. It also enables advanced full-text search queries to be supported to match on phrases, wildcards, joins, grouping, etc., across any data type.

Discovered data is also automatically indexed and facets generated to make it easy to find data in the catalog nomatter where it resides in the data lake

In addition, Smart Data Catalog can also automatically create facets from the metadata generated in automatic discovery to enable faceted search based navigation to data. Faceted search makes it easy to find data in the catalog simply by clicking on a term (a facet). However, Waterline Smart Data Catalog goes further by supporting multi-faceted search to allow users to find data within the data lake more quickly. Facets are dynamically created based on the search results and include many data set and field properties including tags, origin (established by stitching together lineage), data set type, format, modification time, size and others.

Once found, users can immediately see a profile of data quality (generated by automatic profiling) and bookmark files to create a direct access shortcut to that data. Users can also request notification of changes to schema or tagging. Also all usage activity is logged for auditing purposes.

Schema Generation On Hadoop Data

Automatic Hive table schems generation on Hadoop speeds up access to new data in the data lake

In addition to faceted search based access, Waterline Data Smart Data Catalog makes it possible for business analysts to quickly find Hadoop files and access them with BI tools. In a Hadoop system with hundreds of thousands of files, this is a very powerful capability. Users simply use search-based navigation to find the data file in the catalog and then automatically generate a Hive table on top of the file at the click of a mouse. In this way, schema can be projected onto files in HDFS to allow business analysts with SQL generating BI tools like Qlik Sense, Tableau and even Excel to quickly access and query the data. Furthermore, users have an option of using tags or business terms for field names alleviating the need for business analysts to remember physical and often cryptic physical field names. Of course, not all users will be authorised to access all files. Nevertheless the combination of automated discovery, automated tagging and automated Hive table generation really can help reduce time to value.

Business Glossary Creation And Import

Automatic business glossary generation from discovered data as well as business glossary import

On top of all this, it is possible to create a business glossary within the catalog and also import business glossary metadata. This way, consumers of data in the data lake have a common understanding of data in Hadoop that they are authorised to see.

CONCLUSIONS

The data landscape in many enterprises is increasing in complexity

There is no question that the data landscape within many enterprises is increasing in complexity with data from many more data sources now being ingested and data being analysed across multiple analytical data stores. Furthermore, data is being captured in cloud storage, Hadoop and on relational databases. As the data lake becomes more complex the idea that you are going to know what exists and what is happening across this landscape without the aid of an information catalog is near on impossible to imagine. This technology has not only risen in importance, it is practically mandatory if you are going to organise, operate and govern data in a distributed or centralised data lake.

Information catalog technology is fast becoming 'must have' technology to stay on top of what exists across this landscape

The information catalog is central to the success of any data lake implementation. With hundreds of new data sources, hundreds of thousands of files, huge data volumes, high velocity streaming data and IT and business users now active in preparing and analysing data in a data lake, user need software that can see across it to stay on top of what we have. Users need to know what's new and what has changed and still make it easy to find and analyse data to deliver business value.

Automation is required as data variety is too great and more is arriving too fast for everything to be done manually

However, the catalog of technical metadata is not enough. Business analysts would not be able to find the data sets that they need by searching using cryptic field names. It is paramount that data is tagged with corresponding business metadata to make it findable, understandable and usable for self-service analytics as well as governable by the tag-driven governance tools and policies. The idea that you can do this manually is nothing short of fantasy. Automation is now a must and with all of this needing to be managed, organising, managing and operating a data lake is just not going to be possible without an Information Catalog.

Waterline Data's Smart Data Catalog can see across Hadoop, relational and NoSQL data stores to help organise and govern data in a distributed data lake

Waterline Data's Smart Data Catalog offering is a welcome tool in helping companies get on top of this problem by making it possible to discover, inventory, find and understand your data in an increasingly complex data landscape. With many companies offloading staging areas and ETL processing from data warehouses to Hadoop and a lot of the aforementioned new data also being ingested onto this low-cost scalable platform, creating an information catalog of Hadoop data is not a bad place to start in organising and governing the data in your data lake. The need to know what is on NoSQL databases and relational DBMS based data warehouses is also part of the requirement. Then if you start capturing data on cloud storage you can broaden the scope of what is catalogued to accommodate this environment.

Whether you have a centralised Hadoop only data lake or a more complex multi-platform distributed data lake, Waterline Data is in a good position to help you come to grips with how to manage and govern your data while also enabling you to quickly find, understand and provision trusted data so that people can to deliver business value.

About Intelligent Business Strategies

Intelligent Business Strategies is a research and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, analytical processing, data management and enterprise business integration. Together, these technologies help an organisation become an *intelligent business*.

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant he specialises in Big Data, BI/Analytics and Data Management. With over 35 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, big data, data governance, master data management and enterprise architecture. He has spoken at events all over the world and written numerous articles and blogs providing insights on the industry. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS and European Managing Director of Database Associates, an independent IT industry analyst organisation. He teaches popular master classes in Big Data Analytics, New Technologies for Business Intelligence and Data Warehousing, Data Virtualisation Enterprise Data Governance, Master Data Management, and Enterprise Business Integration.

Water Lane, Wilmslow
Cheshire, SK9 5BG
England
Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-mail: info@intelligentbusiness.biz

Organising The Data Lake – The Central Role Of The Information Catalog