IBM Industry Models

IBM Industry Model support for a data lake architecture

Version 1.0

Contents

1 Introduction

The data lake has emerged as the recognized mechanism to enable organizations to define, manage and govern the use of various big data technologies. This represents an evolution of big data towards the mainstream use in an enterprise and the associated focus on management of such assets.

IBM Industry Models have traditionally been a means for organizations to establish the necessary communications between the business and the respective technical initiatives in areas such as data warehousing, business process reengineering and services oriented architecture. Many of the same traditional imperatives for the use of IBM Industry Models also exist when organizations deploy a data lake:

- The need to establish a common cross enterprise set of assets for use by the business
- The need to ensure a common understanding of such assets by the business and technical users
- The need to enforce a common governance layer around the data lake

This document will provide the necessary guidelines and practices to organizations who want to use IBM Industry Models as a key part of their data lake initiative.

1.1 About this document

This document is intended to provide guidelines to organizations using IBM Industry Models to assist in building a data lake infrastructure. This document will cover the different considerations for using the various IBM Industry Model components (for example, Business Vocabulary, Data Models) in the context of a data lake. This document will also contain the initial set of approaches and development steps to be considered in building out different parts of the data lake by using IBM Industry Models.

This document was created with the assistance and feedback from a number of IBM customers who are beginning to build out data lake infrastructures using the IBM Industry Models as part of their development process.

1.2 What this document means as a “data lake”

While there are a number of different interpretations across the industry of the precise definition of a data lake, for the purposes of this document, the assumption is that a data lake is any collection of data repositories which an organization would like to govern and manage a single set of assets to be reused across the enterprise, including traditional information warehouses, operational hubs, landing zones (HDFS and Relational) and collections of deep data on HDFS clusters.

1.3 The use of tooling in this document

While the overall principles of this document would apply to the deployment of IBM Industry Models to a data lake using any suitable tooling, the examples in this document are using IBM tools. For example, IBM InfoSphere Information Governance Catalog (IGC) and IBM InfoSphere Data Architect (IDA).

1.4 Intended audience

This document is intended for anybody who is considering using IBM Industry Models to assist with the definition, governance and evolution of a data lake environment. Personnel such as data architects, business analysts, and modelers would find this document beneficial.

2 IBM Models and the overall data lake landscape

This chapter briefly describes the main component areas of the data lake and describes the most likely associated integration points that IBM Industry Models would have with the data lake.

2.1 Use cases and motivations for IBM Industry Model customers to consider a data lake.

In general, there are a number of well documented reasons as to why organizations are now considering the deployment of a data lake. These reasons usually fall into a number of areas:

- The need to define an infrastructure to manage analytical repositories over all data, whether structured, semi-structured or unstructured with lineage and the appropriate governance that will enable the organization to be compliant.
- Provide a means for storing document and their content for reuse by multiple users, including business users and data scientists and expanding the use cases to "Citizen Data Science Discovery work" (next generation BI)
- Provide the necessary data lineage back to source systems
- Enable the users to access the analytical contents in a self-service manner

Specifically, when looking at organizations who are using IBM Industry Models, an additional number of reasons have been mentioned in the context of data lake:

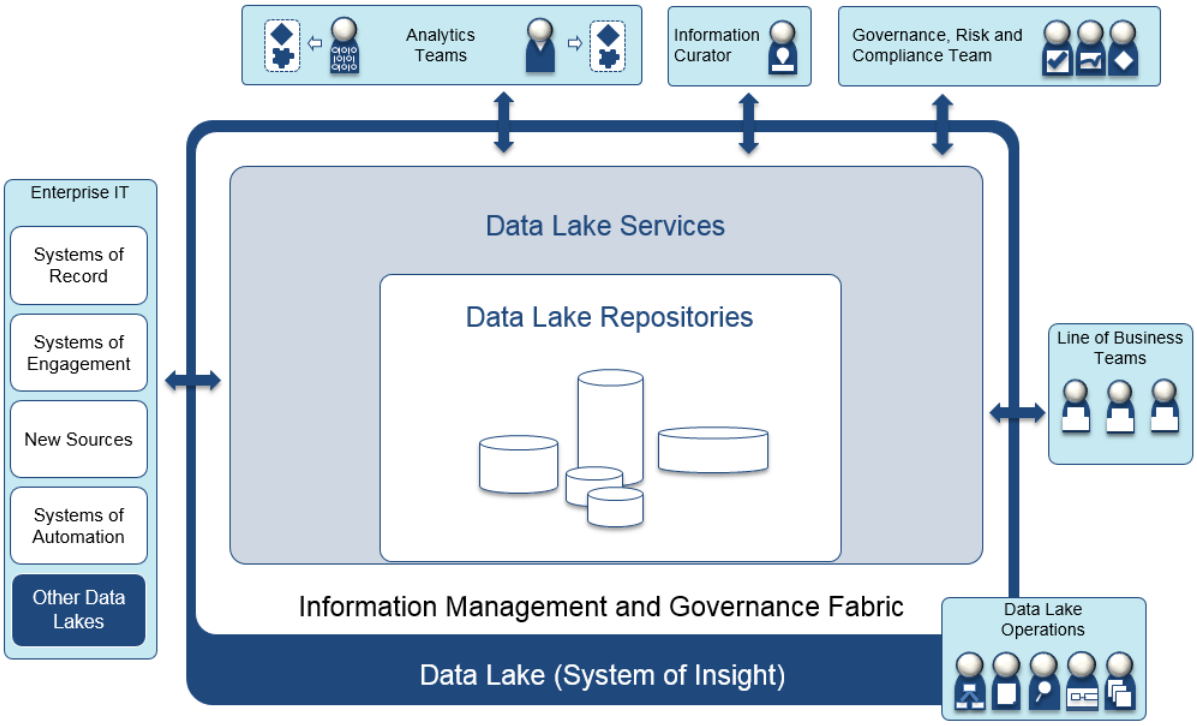
- Enable a mechanism to provide the reporting and lineage support required for certain regulatory/compliance needs
- Provide a basis for rationalization and reduction of IT effort that is needed to support a large number of current separate reporting environments.
- Create a set of canonical data structures to enforce consistency and standardization across the organization – enabling better reuse of data across the different LOBs.
- Use the data lake as the central clearing hub for all data outside of the core Systems of Record.
- Using the data lake as the basis for developing advanced analytic models, developing risk insights, operationalizing analytic/predictive models, and doing data integration
- Since the lake will form the underpinning of everything "to the right" of the transactional systems, IBM Industry Models are a foundational pillar, along with governance, in ensuring a data lake stays pristine.
- A highly evolved metadata management capability is critical in not only aiding the analysts in understanding the data that they are consuming, but also mandatory for the data governance processes and data management processes.

2.2 The data lake landscape

At the core of the data lake are the set of repositories that have been designated as fitting the criteria to be included within the data lake[1]. These repositories could range from traditional RDBMs information warehouses to operational data hubs to HDFS clusters.

The data lake services exist to ensure consistent and controllable access to the data lake as well as ensuring that the appropriate levels of integration/synchronisation are achieved between the data lake repositories and the broader enterprise IT systems.

Underpinning all of this is the necessary middleware called the Information Management and Governance Fabric which oversees all of the provisioning workflows into and across the data lake Repositories as well as providing access control, monitoring and audition capabilities.



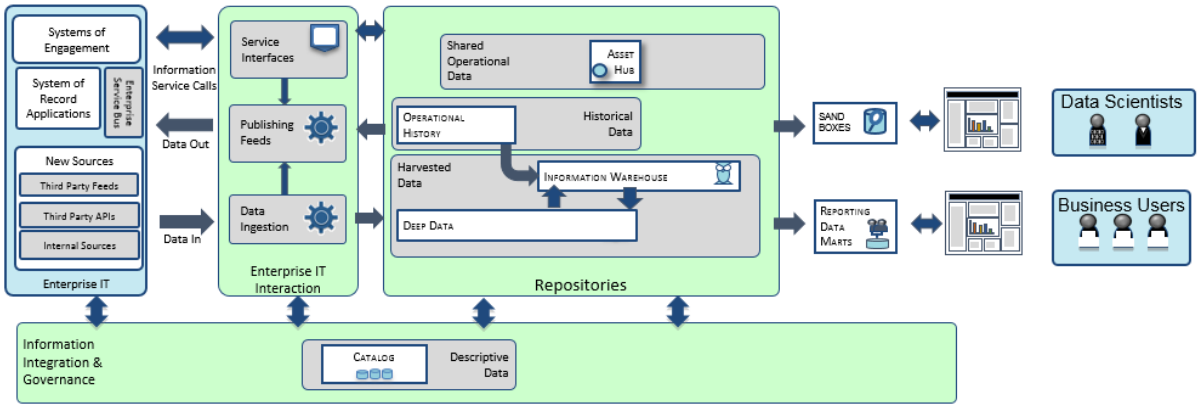
In terms of the various users of the data lake, this will range from the traditional line of business users, the teams responsible for the day to day operations of the data lake as well as some less traditional data management users:

- Analytics Teams – typically a group of users, including data scientists who are responsible for carrying out the advanced analytics across the data lake.
- Information Curator – a role which is responsible the management of the Catalog which will be used by users to find the relevant data elements within the data lake
- Governance Risk and Compliance team – this group is responsible for the definition of the overall governance program of the data lake and any associated reporting functions to demonstrate compliance.

[1] For more info on this landscape reference to the IBM Redbook: Designing and Operating a Data Reservoir http://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/sg248274.html?Open

2.3 IBM Industry Models and the data lake

There is in theory a broad range of components that could fit within a data Lake. For the purposes of this document only the subset of Data Lake components that are currently of typical interest to Industry Models customers deploying Data Lakes will be covered.



Some of the critical components in the diagram above in relation to IBM Industry Models are:

- Catalog - The set Business Vocabulary content, usually stored in IBM InfoSphere Governance Catalog are deployed into the *Catalog* component.
- Information Warehouse - The traditional data warehouse is located in the *Information Warehouse* component.
- Deep Data – This is data in a non-relational repository providing a historical record of the data from the systems of record.
- Reporting Data Marts - The Dimensional Warehouse Models would provide much of the design content for the *Reporting Data Marts*. In addition, where a client is looking to deploy a Kimball dimensional warehouse they can also use the Dimensional Warehouse Models to deploy a dimensional variant of the *Information Warehouse*.
- Asset Hub – The set of near-real-time operational data, typically grouped around data entities such as Customer, Product, Account, etc.
- Sandboxes – A (usually non-relational) store for data for experimentation purposes.

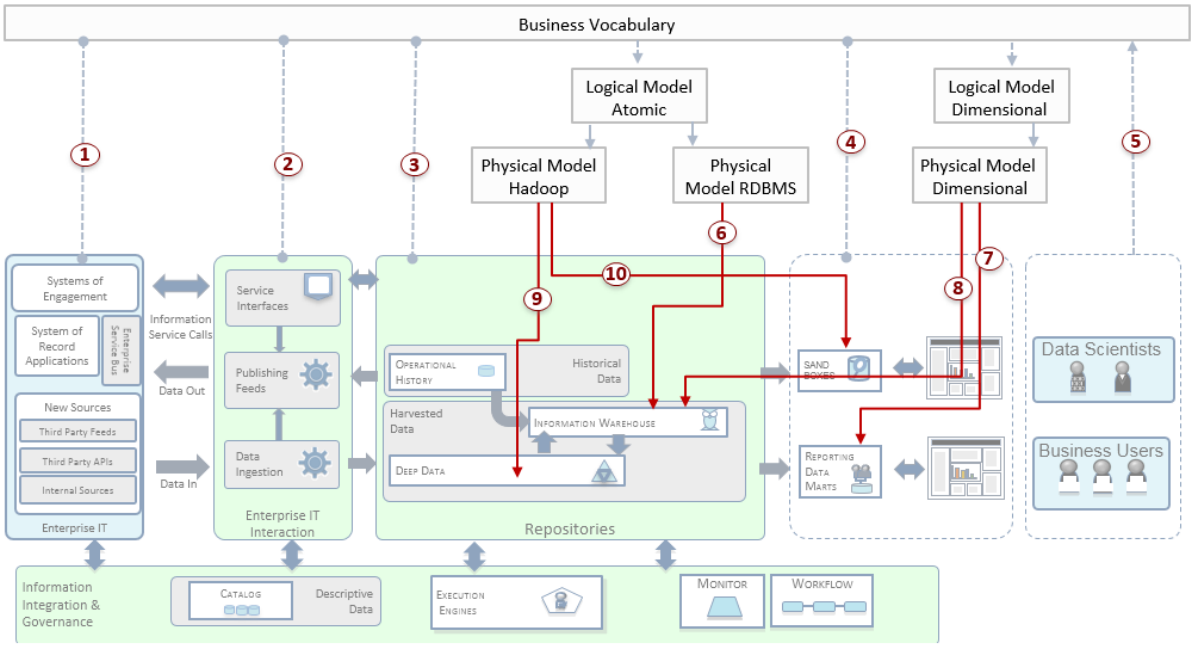
There are more publications that outline in more detail all of the potential components of the data lake reference architecture[2].

2.4 The role of IBM Industry Models and the big data Landscape

Typically, IBM Industry Models are design-time artifacts and are used to underpin the related development activities. In general, a number of IBM Industry Models components deploy into this physical landscape.

[2] For more info on this landscape reference to the IBM Redbook: Designing and Operating a Data Reservoir: http://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/sg248274.html?Open

The diagram below provides a summary of the main interaction of the different types of IBM Industry Model components with the typical areas of a big data landscape using Hadoop and more traditional RDBMS technologies.



In the above diagram, there are the following broad sets of activities:

Mapping of assets to the Business Vocabulary

1. Any relevant assets in the Systems of Record or Systems of Engagement can be mapped to the overall Business Vocabulary.
2. Any integration-related physical assets (for example, ETL jobs) are mapped to the Business Vocabulary
3. All repository assets should have mappings to the Business Vocabulary. This should include assets that were brought in as-is from upstream systems with no attempt to enforce a standard structure (for example, historical data)
4. Any downstream published data, both data marts as well as certain data scientist Sandboxes should be mapped to the Business Vocabulary. In the case of the sandboxes, a decision might have to be taken in terms of whether such assets have a broader value to the enterprise beyond the individual data scientist, in which case mapping to the Business Vocabulary. In some cases, the decision might be taken that such Sandboxes are transient and local to the individual data scientist, whereas in other cases organizations might wish to enforce some standardisation/reuse across some subset of the data scientist sandboxes.

Use of the Business Vocabulary for search and discovery

5. The Business Vocabulary should be the basis for any searching or discovery activities that are carried out by the Business Users and the Data Scientists.

Deployment of physical structures from the logical models

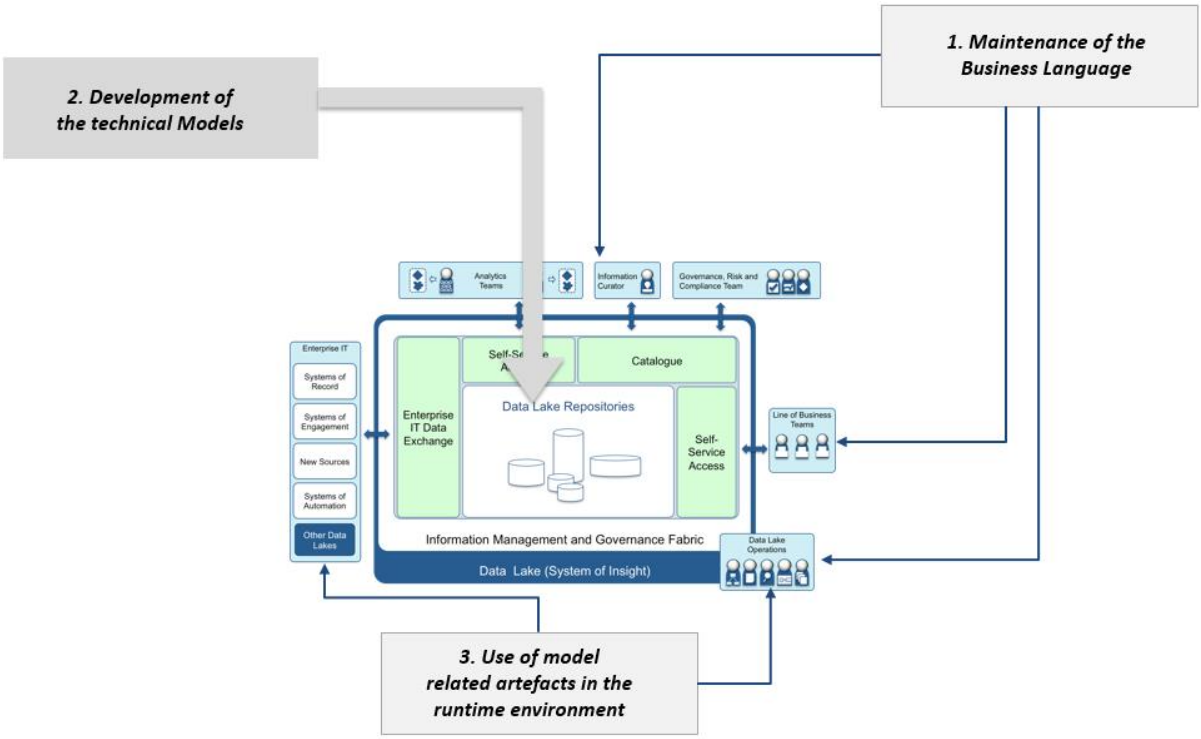
6. Use the Physical Atomic Models to deploy the necessary structures to build out the traditional Information Warehouse assets.
7. Use the Physical Dimensional Models to deploy the necessary structures to build out the Data Mart layer.
8. Use the Physical Dimensional Models to deploy the necessary structures to build out the Kimball-style Information Warehouse assets.
9. Deploy deep data assets on Hadoop (for example, typically HDFS structures either using HBase or Hive or other vendor-provided SQL interfaces) either for which there is a need for a specific structure[3].
10. As required, the Physical models might also be used to deploy the data scientist Sandboxes. Most likely in circumstances where an organization would like to enforce a common set of structures for their data scientists to use.

2.5 The different Model activities and the data lake

A key principal that is likely to be applied when considering the deployment of IBM Industry Models and the data lake is that of the separation of the definition of the business language and the definition of any associated structures to be used in the creation of the data lake repositories. This separation of what can be seen as the fundamentally different but related domains is central to this document and is introduced in this section.

In general, three separate areas of development activities have been identified in terms of the possible interaction of IBM Industry Models with the data lake infrastructure.

[3] For a more detailed description of the options for deploying IBM Industry Models to Hadoop, please refer to the separate document "*Guidelines for deploying an IBM Industry Model to Hadoop*"



Maintenance of the business language

The use of IBM Industry Model Business Vocabularies to enable a common business meaning of language by all data lake users. The heterogeneous nature of the technologies contained within the data lake and the fact that in many cases there is no requirement to enforce any kind of schema, means that there is increased importance placed on begin able to define a common business language. Specifically, such a language is required to act as a common point of reference to help all users have a common understanding of what is a diverse range of component. Therefore, there is the need for a separate dedicated lifecycle to focus on the creation and ongoing maintenance of this Business Vocabulary.

Development of the technical models

It is likely that there will be repositories within the data lake for which the organization would like to define a specific structure or schema, such as a traditional data warehouse. It is also likely that there will be certain repositories within the data lake for which there is no need for such an upfront schema definition, typically where there is a need to land data coming into the data lake in an “as-is” format and typically in technologies such as HDFS where there is not necessarily a need to define a schema up front. Where the organization has decided that there is need for an upfront definition of a schema, then the recommendation is that IBM Industry Data Models are used, to ensure consistency of structure across all repositories in the data lake

In addition, there is potentially a role for the use of UML-based IBM Industry Models to assist with the definition of the common services layer, although this is likely to be the focus of subsequent phases of activities on the data lake deployment guidelines.

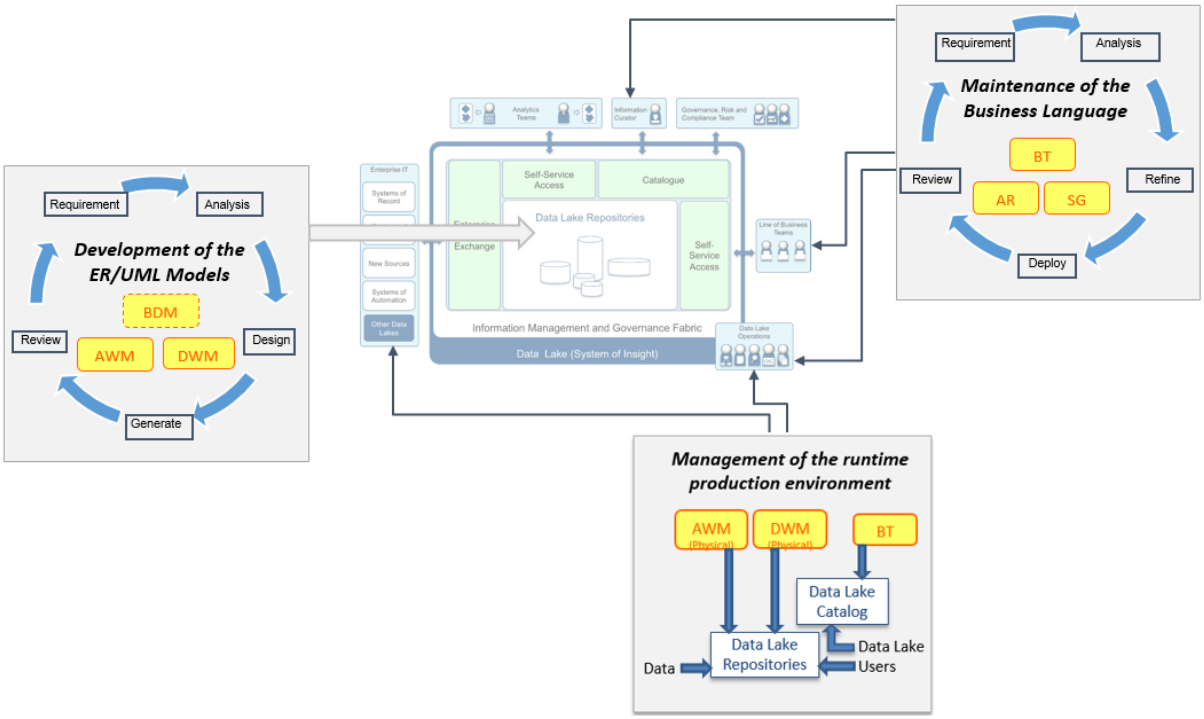
Use of IBM Industry Models in the runtime environment

This last lifecycle is concerned with the considerations for the use of IBM Industry Model Business Vocabularies and derived physical assets in the run time environment of the data lake. As well as the actual deployment of these artifacts for use in the runtime environment, there should also be considerations on defining the necessary feedback loops to the other activities.

There is also a significant difference in how the Business Vocabulary would be integrated with the runtime environment, given that there is a close connection between the metadata layer store used in the development of the business language and the metadata stored used in the actual runtime environment for these terms. Whereas for the technical models, there is likely to remain a significant and distinct separation between the ”Design Time” and “Run time” variants of these models with quite separate metadata stores.

3 The different lifecycles

The sets of activities mentioned in the previous chapter should be viewed as three separate but interrelated development lifecycles. These three lifecycles will each focus on the development and evolution of different IBM Industry Model components, with different participants and skillsets. It is also likely that these lifecycles will progress at different speeds, depending on the specific circumstances and business need.



The diagram above shows the three different lifecycles, the different IBM Industry Model components that they would typically be expected to evolve and manage and where these lifecycles would typically connect with the data lake Landscape.

Maintenance of the business language.

This lifecycle is concerned with the management and evolution of the Business Vocabulary set of content provided by IBM:

Business Terms (BT in the above Diagram) – The central set of terms or concepts that are used to describe the business. This set of terms is structured in a form that is independent of any business use. The primary objective is for these Business Terms to provide a central glossary or taxonomy for the data lake.

Analytical Requirements – The Analytical Requirements can supplement the core Business Terms used in the data lake Glossary by providing additional sample report-oriented measures and dimensions.

Supportive Content – The Supportive Content can supplement the core Business Terms used in the data lake Glossary by providing additional details relating to specific external or regulatory taxonomies.

In terms of the integration point with the data lake landscape, it can be seen that the business language defined by this lifecycle should be actively used as the common reference point by all of the users of the data lake, from Business Users to IT staff responsible for the data lake maintenance.

This will bring a set of challenges around the need, in some organizations, to ensure that these different users are able to see just the subset of the overall Business Vocabulary that pertains to their role, but to also ensure the ability for certain users to have a wide ranging view of the full Vocabulary and how it relates to the various associated data lake assets.

Development of the technical models

This lifecycle is concerned with the analysis and design of the necessary technical models (for example, ER or UML models) which should be guided by the initial requirements definition as done in the Business Vocabulary. The result of this lifecycle is the generation of the appropriate Physical models for deployment in the data lake runtime environment. In terms of data models that are managed by this lifecycle:

Business Data Model (BDM) – The Business Data Model (BDM) is intended to provide the initial design-independent ER representation of the relationships and constraints of the business requirements. The BDM is only provided with the IBM Industry Data Models for Insurance, Healthcare and Energy & Utilities. This model component is not supplied for Banking/FM, Telco and Retail where the review of the business relationships and constraints in the Business Terms is done via a deep taxonomy provided in the Business Vocabulary. However, in such cases, it might be likely that a high level “Conceptual Model” is employed to provide a very high level view of the main concepts and the primary high level relationships.

Atomic Warehouse Model (AWM) – The Atomic Warehouse Model (AWM) is a design-specific model that is intended to provide the basis for deployment of a central Inmon-style atomic data warehouse. In the context of the data lake this model can be used in the traditional role of deploying a conventional data warehouse, alternatively where there is a need to enforce a consistent schema on HDFS Deep Data structures, then this model can also be used. However, the level of normalization is likely to be less in such deployments due to the requirements in HDFS for flatter less normalized structures.

Dimensional Warehouse Model (DWM) – The Dimensional Warehouse Model (DWM) is a design-specific model that is intended to provide the basis for deployment of Kimball-style dimensional warehouse structures. In the context of the data lake, this model can be used to deploy either traditional data marts or in some cases can be used to provide conventional Kimball dimensional warehouses. There are also potential additional uses for this model in terms of:

- Deploying to dimensional structures on HDFS
- Deploying part of the infrastructure to support the data scientist activities.

The primary integration point between this lifecycle and the data lake landscape is that the physical models that are generated are used as the basis for the creation of the physical repositories within the data lake.

Management of the runtime data lake environment

This lifecycle is specifically concerned with the considerations around incorporating the various artifacts related to and derived from IBM Industry Models into the run time data lake environment. This lifecycle is not going to focus on the much broader set of tasks relating to the overall management of the complete data lake.

So this lifecycle will primarily be concerned with:

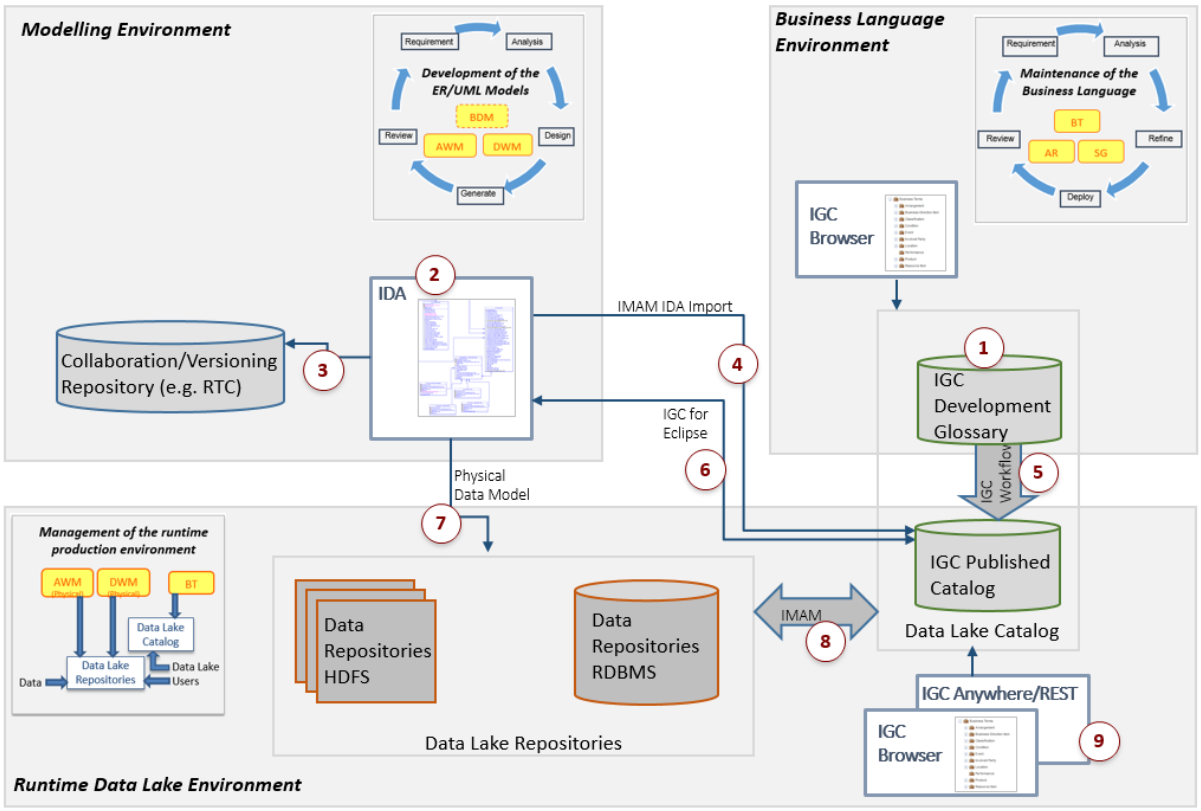
- The considerations for incorporating and managing the Business Terms that describe the business language into the production repository of the data lake catalog
- The considerations for the deployment of the Physical models derived from the technical models into the appropriate structures in the data lake repositories.

3.1 The underlying metadata repositories

In order to better understand the inter-relationship between the three lifecycles, it is important to understand the metadata repositories, which are used to underpin them. Essentially there are typically two different types of metadata repository used across these different lifecycles:

1. Data Lake Catalog - This is the environment that is used to store all of the metadata relating to the Business Terms along with the metadata relating to any associated physical data lake assets (data repositories, ETL flows, etc.) This is typically a metadata Repository that is very tightly integrated with the Information Governance Fabric of the data lake. This means that the processes and metadata environment used during the creation and evolution of the business language (Development Glossary) is the same as that used to underpin the use of that Vocabulary in the Runtime environment (Published Catalog)
2. Collaboration/Versioning repository – This is the environment which is used to store all of the technical models along with any associated relationships between these models. This Metadata Repository is typically a general purpose IT Development versioning and collaboration environment (for example, Rational Team Concert, CVS/Subversion, Git) and so is separate from the runtime data lake environment.

Finally, there are also the actual data repositories that are used to store the different instance data needed to be managed by the data lake.



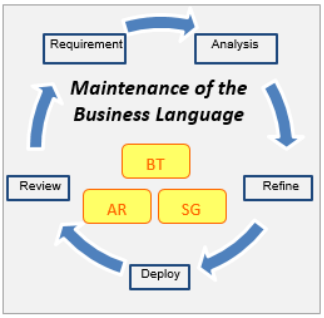
The above diagram shows the main flow of artifacts between the main repository components of the three different lifecycles.

1. In the business language environment, the necessary new and changed terms are created and reviewed using the IGC Browser tool with the specific metadata stored and managed in the IGC Development Glossary.
2. In the Modeling Environment, the modelers make the necessary data model enhancements needed to define the physical data structures for the data lake
3. The development of the models is likely to be managed and versioned as part of a standard source code management, collaboration and versioning repository.
4. There might be a need for the data models to also be imported into the data lake Catalog for reference purposes to provide further context to the business terms if needed. The master copy of the data models will still remain managed in the modeling environment.
5. When the new or changed terms have been appropriately reviewed in the Development Glossary they can then be promoted to the Published Catalog. Once in the Published Catalog, these terms can then be viewed, mapped to by the users of the runtime data lake environment.
6. It is possible to synchronize the read-only view of the Business Terms as seen by the Modelers with the master set of Production Business Terms. It is also possible to import any mappings created by the modeler between business terms and data model artifacts into the Published Catalog.
7. At the appropriate time the physical data model created by the models can be imported into the environment for managing the runtime data lake repositories.
8. The IGC Published Catalog is also used to store the metadata of the various other physical artifacts within the data lake.
9. Finally, the various users of the Runtime data lake can use the contents of the Published Catalog.

There are a number of additional publications available that describe in more detail the process regarding the IGC Development Glossary and Publication Catalog[4].

[4] Planning, designing and publishing the IGC Catalog : https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.5.0/com.ibm.swg.im.iis.bg.bestp.doc/topics/c_iadmgde_BuildingGlossary.html

4 Defining the business language of the data lake



4.1 Business language overview

The purpose of the business language is to provide a consistence set of terms that can be used by both business and technical teams to describe the information in the data lake. The business language is the initial contact point for most users of the data lake and so it is imperative that the language creates a positive first impression of the data lake's usefulness and quality. Therefore, the effort to create and maintain the language it with relevant content is a key success factor for the data lake.

As part of the data lake architecture, this language is stored in the Catalog, which contains the descriptions of the data the data lake and related rule that that define how this data is being managed. A business term captures the vocabulary used by the business and includes a textual description that defines the term itself. By associating assets in the Catalog with one or more business terms, a powerful semantic layer is created which indicates that a given information asset is classified by this particular business term. This classification can allow for searching by business terms so the users can find all the assets that are associated with that business concept. The business language is the created and maintained though an iterative lifecycle;

1. Requirements for business terms arise from the need to provide business meaningful descriptions of information in the data lake.
2. Analysis of existing business definitions and agreement with stakeholders that additions to the business language are correct and will be useful to all data lake users.
3. Refining the business language to ensure that it meets business language standards, is structured correctly.
4. Deployment of the language in the data lake catalog so that it is available to all data lake users.
5. Review of business terms based on feedback from users and the use of the business language.

Example 1 of Business Language – Customer A

Customer A Esperanto is the global business language, which describes data contained in the Customer A data lakes. The focus of the language in on the consistent and accurate communication between local federated organizations and global centralized functions. Customer A has used the Banking Data Warehouse Business Terms as the starting point for their development of Customer A Esperanto.

Customer A operates a federated data architecture with each country having its own data lake, which is used for local operations and reporting. Esperanto is the mandatory language to be used by all countries when transferring data between data lakes for global reporting. Countries are also encouraged to use Esperanto in the development of internal data repositories within their own data lakes.

4.2 Main roles in business language lifecycle

The creation of the business language requires individuals performing a range of roles. In the size of enterprises typically implementing a data lake architecture this will involve the creation of one or

more dedicated teams that are responsible for implementing and managing the language. The roles described below are the main actors in the five stages of the lifecycle.

Business Language Authors are the users who are involved in the creation and maintenance of the language.

- **Business Term Modelers/Analysts** understand the process of defining business terms, their relationships to other terms in the language and the holistic use of the business language in the data lake. They are responsible for the creation, definition and management of the individual terms in all or a subset of the business language. They have expertise in both IBM Industry Model Business Vocabulary and Information Governance Catalog.
    - **IGC workflow: Editor – create, modify terms etc.**
    - **IGC Role: Term Authors, <u>not</u> Information Asset Assignment, Catalog Users**
- **Business Subject Matter Experts** provide their business subject matter expertise both in the definition of business terms and identifying requirements for extending the business language. The business analysts work with business terms owners to establish terms that need to be represented in the language.
- **Business Language Stakeholders** work with the Business Term Modelers to ensure the definitions of the term in the language are consistent with the goals of the organization. The Stakeholders provide the executive sponsorship that empowers Business Term Modelers to create an enterprise language that is independent of specific lines of business.
    - **IGC workflow: Approver – just approve, not edit**
    - **IGC Role: Catalog Users**

Business Language Users are a diverse range of both business and technical roles who use the business language to describe, find and manage information in the data lake. The development of the business language needs to reflect the needs of all users

- **Line of Business Users** who use the business language to look up definitions of business concepts.
- **Data Modelers** who create mapping between the business language and data models. This activity is described in detail in Section 5.
- **Data Stewards** who define data quality rules using the terms in the business language.
- **Metadata Managers** who deploy the business language in the catalog and integrate it with information assets in the data lake.

Example of Business Language Organization

Customer A have created an organisational structure to manage and operate the lifecycle of their enterprise business language.

The Global Data Management (**Modelers, Business SMEs, Business Language Stakeholders**) is an enterprise team responsible for creating and governing the business meaning of the Esperanto business language. They analyze the new business terms and either reuse or add new terms inspired by definitions from the IBM Industry Model Business Terms. The team uses a formal IGC workflow to refine terms and obtain approval from stakeholders.

The Data Management Technical Authority (**Business Term Owners, Data Modelers**) is an enterprise team responsible for technical accuracy of the Esperanto business language. They refine the Esperanto Terms by validating the structure of language in IGC and by completing the mapping to data models.

The data lake Foundation (**Metadata Management**) is the enterprise team responsible for the distribution of Esperanto. They deploy the Esperanto Terms to the local and global data lakes.

4.3 IBM Industry Model artifacts involved

IBM Industry Model Business Vocabularies provide an accelerator for the creation of this language by providing a predefined language that describes that information that represented in the data models but is independent of the technical structure of the data models. IBM Industry Models have three components delivered in Information Governance Catalog (IGC) that can be used as the starting point of defining the business language of the data lake;

1. Business Terms which are organized by business categories and by hierarchies. This is the recommended starting point for the development of a data lake business language as they have been designed to enable users to create an enterprise-wide understanding of the data.
2. Analytical Requirements which comprise high-level reporting information and business measures along the axes of common dimensions. While the Analytical Requirements are primarily designed to allow business users to rapidly map reporting requirements to the data models, they can also be used as the part of a data lake business language that is focused on supporting calculated values or KPIs.
3. Supportive Glossary incorporate terminology that originates from external sources such as regulatory authorities and industry standard bodies. It is not recommended that Supportive Glossaries are used for the basis of a data lake business language. Instead the business language should be mapped to the relevant Supportive Glossary content designed to allow for the easy translation of information requirement expressed in these external languages.

The remainder of this chapter assumes that the IBM Industry Model Business Terms are being used as the basis of the the definition of a business language that is managed using IBM InfoSphere Information Govenenace Catalog.

One other possible artifact, the Business Data Model (BDM).

In industries where a BDM is provided, there might be some role for this model to be used as part of the definition of the business language. Some organizations might see this model as a core part of the set of artifacts that make up their business language. For the purposes of this document it is assumed that the BDM is part of the Life Cycle 2. However even in this case, it remains a critical asset in terms of assisting the technical/modeling teams to further understand the structure and constraints of the set of terms being defined as part of the language and what are the necessary downstream design level models needed to instantiate these business terms in the definition of the relevant repositories of the data lake.

Currently the BDM is available for the Insurance, Healthcare, and Energy & Utilities Models. In the other industries (Banking/FM, Retail, Telco) the role of the BDM is taken on by the extended taxonomy (for example, the Financial Services Data Model in Banking).

4.4 Inputs to the lifecycle

The inputs for this cycle are characterized by a number of design decision that will influence how the IBM Industry Model content is to be used to define the business language.

- Having a clear objective and scope of the business language
  - What data is being described?
  - What are the most common terms causing confusion required in the data lake?
  - Who is the intended audience of the business language?
  - Defining the department v's enterprise business language
  - Anti-pattern – spending time defining business language for data that might never be in the data lake
- Having a clear scope
  - Are we describing business concepts?
  - Are we describing the common information used by the business?
  - Are we defining lists of reference data?
  - Are we defining business concepts or reporting measures?
- Clear ownership of the business language
  - Enterprise level and not departmental or line of business
  - Business language not a technical language
  - Business language and not a data model
- Technical Environment
  - IGC set up with users
- Workflow on or off?

4.5 Setup considerations

This section describes the steps and considerations regarding the initiation of the various components associated with the initial definition and ongoing maintenance of the business language for the data lake.

4.5.1 Determine the needs of the different users

A key consideration is that there are likely to be a range of different types of users all attempting to avail of this business language. These uses are likely to have different requirements of the data lake and the language associated with it, they will have different levels of business and technical knowledge and will have different needs in terms of the breadth of the data lake they are interested in. There is also the important consideration of security/ownership considerations of the data in the data lake.

In general, the considerations pertaining to the different types of users of the data lake and how the business language will support them are:

The scope of business language that is suitable for the user.

In some cases - such as the data lake Operations personnel, the Data Curator and potentially Data Scientists - there might be a need for the user to have access to the full set of terms in the business language. In other cases, there might be a desire to restrict to certain users to only the subset of the business language that pertains to their area. This would have the benefit of presenting to users just the terms that they understand and are familiar with and avoiding overloading any search results with terms that are alien to the users. However, the converse is that in some cases there might be some lost opportunities when these users are not aware of other terms in other domains that might be of use to them in their role. [5]

Another scope-related consideration is to decide whether or to what extent to include links in the business language to the upstream Business Processes in the Systems of Record and of Engagement.

Of course with the question of scope come the security and ownership considerations.

Experimental or "Production" areas of the data lake

In most cases it is expected for the data lake to consist of Data Scientists and other analysts carrying out more experimental or investigative activities in addition to the Business users more interested in the "production" data and associated queries. The data and query needs and behaviors of these two sets of users are likely to be different and might influence the specific aspects of the business language being used by these users.

Whether there is a need for any local "dialects".

In many cases in larger organizations there might be cases where groups of users have their own vernacular of terms specific to them which are unknown or irrelevant outside of their group. Or perhaps there are synonyms for their terms that are called something different in other areas. It would be important to define the policy towards the support, or not, of such dialects. A lot would depend on the normal expectations and cultural norms in place in the organization. One approach might be to try to define a single set of terms for use by all users, alternatively to allow for the use of terms specific to a department, but in this case it would be recommended that such terms are linked as synonyms to the equivalent "enterprise version" of the term to ensure clarity of communication.

[5] There is a separate set of documentation from the IBM Industry Models team dealing with the question of setting up Enterprise and associated departmental Glossaries. For more information, contact IBM Industry Models team.

What is the level of structure required to be exposed.

There is likely to be varying degrees of hierarchical structure built into the business language from a simple flat glossary of business terms with little hierarchical structure to taxonomies with extensive and deep hierarchies of business terms. The latter might be the case in the Banking FM, Telco and Retail models where such deep taxonomies are provided.

The use of such hierarchies of business terms have the advantage of providing a degree of precision when it comes to understanding the semantic relationships between terms and the associated mappings those terms have to data lake physical assets. However, in general, such deep taxonomies can sometimes not resonate with the departmental business users, where they might prefer just to see a simple list of the relevant terms.

One option might be to retain the use of the taxonomies along with the deep hierarchies, but to ensure that these hierarchies are only visible to the users for whom such detail is welcome and relevant, with the mainstream business users just having visibility to the individual set of Business Terms that make sense in their specific context.

Business and technical terms.

There might be a need to have a collection of both business and technical terms in the language. In some cases, it might be necessary to ensure that detailed technical language is not exposed unnecessarily to business users. However, in other cases, for specific groups of users, the decision might be taken to provide additional more technical terms that to provide further context for the Business Term. An example of this would be the use of more generic or non-business terms to describe with more precision the location of business terms in an overall taxonomy.

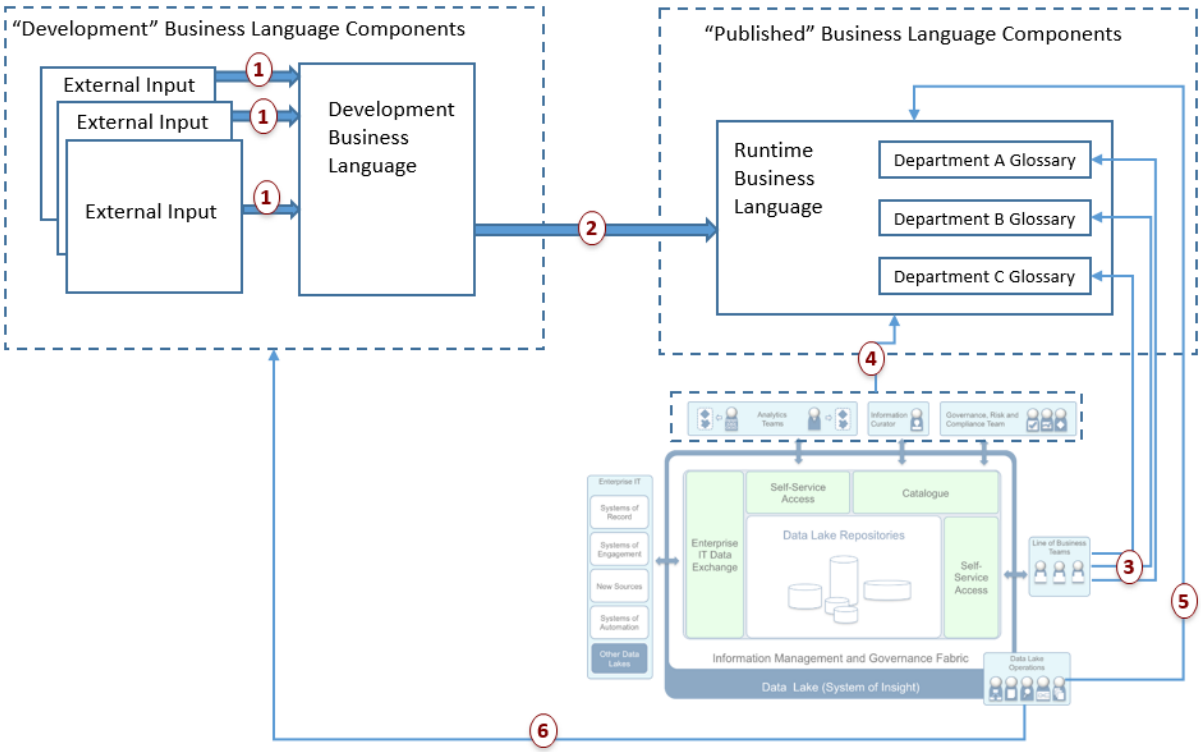
What artifacts to include in the business language

While the use of a set of Business Terms with appropriate definitions is likely to form the backbone of any business language, some consideration might also be given to what other artifacts to expose as part of the "business language". Specifically, it is a question of whether to include business metadata and technical metadata in the overall business language.

For example, in some cases it might be useful for the business language to also include parts of the supporting logical and physical models in order to give further background or context to a particular term or set of terms. The key question is whether the inclusion of such artifacts in the business language will genuinely assist various users of the data lake and when included that the artifacts are exposed only to the users for whom they will be useful. It would also be important to consider that if included in the business language such Logical and Physical models are only for reference purposes. Finally, there is also the need to ensure that when such models are included in the business language that there is a means to ensure that there is a process

4.5.2 Define the components of the business Language

In setting up the business language for the data lake, it is important to consider the required components to manage both the provisioning of terms and associated artifacts and mappings on a regular basis as well as the need to manage and control the access and usage of the contents of the business language.



The diagram above outlines what the possible components necessary for the setting up and maintenance of data lake business language might be. The main components are:

- Development business language – This is the set of business language components under development by the organization and have not yet been published for use by the users of the data lake.
- External Input – This relates to external models, documents, taxonomies that the organization uses as input to the development of the business language. Examples of this could be IBM Industry Model content, regulatory taxonomies, or DBpedia.
- Runtime business language – This is the set of business language components that have been certified by the relevant stewards/curators as suitable for use as part of the business language to describe the data lake.
- Departmental glossaries – These are specific subsets of the business language that have been identified as necessary to give groupings of business users the set of business terms that are specific to their business domain and using their vernacular.

The specific activities that are outlined in the above diagram are:

1 The relevant external inputs to the development business language are identified and from these inputs the appropriate subsets of business terms and other artifacts are copied into the development vocabulary. If it is expected that there will be periodic future updates from the source for a particular external input, then it is recommended to retain an area where such inputs can be landed and examined.
2 Promotion to the runtime business language
3 Access by the business users
4 Access by the Analysts, Curators and Governance
5 Access by the data lake operations personnel.

6 The ongoing input to the business language definition by the data lake operations personnel.

4.5.3 Defining a glossary suitable for business users

One other consideration in relation to the setting up of the business language is to determine the appropriate glossary of terms that is made available to the business users. In many cases the full content of the business language containing many technical terms and potentially artificial constructs in taxonomies means that there is a need to identify a subset of the overall business language to expose to the business users. Indeed, it would be critical to the success of the data lake overall that the subset of the business language that is used by the business users for their regular search and discovery activities is something they understand and are comfortable with.

Some of the likely characteristics that such subsets of the business language should have are:

- Meaningful to the Business Users, both the term itself and any associated definitions and related artifacts
- Any grouping of terms of logical/natural to Business Users
- Subject oriented
- The glossary of terms for any set of business users to be of limited size
- Designed with consumption by business user in mind
- Should be able to find the item in reality (should not be simply abstract) - anchored on business reality
- Might be grouped based on industry standard groupings

Bottom up and top down

As part of defining the glossary for Business Users, there is likely to be a need to achieve a balance between influencing the glossary with the set of terminology as used by the business users ("bottom up") with the definition of a standard set of terms and associated relationships to be propagated across the enterprise ("Top down"). If the glossary is overly influenced by the actual business terms used by each department, there might be a challenge with the establishment of a true cross enterprise business language. However, if the glossary is overly influenced by a top down canonical structure, it might alienate some of the business users with the potential use of overly generic terms or terms they don't use or recognize.

The ideal is likely to be a balance involving the use of synonyms where the conflict of various "local" business terms is unavoidable. There is also a key role for the business language stakeholders to ensure an appropriate management process is in place with provision to accommodating the requirements of the different groups of users of the data lake.

How to group the terms

A key design decision when defining a Business User focused glossary is to determine the most appropriate means of grouping the terms. There are a number of options and considerations to keep in mind when defining such structure of term groupings.

- Business groupings already used by the business – this would likely be the most obvious approach in many cases. This essentially is looking to reflect whatever natural groupings that are already in use by the business within the structure of the data lake business glossary. It

has the advantage of being something that is inherently meaningful to the business, but might result in reflecting any inconsistencies that are already inherent in these groupings.

- Data concept based – this approach might be considered where it is considered possible to expose the central data concept groupings used in the core business terms to the departmental business users. It has the advantage of resulting in a consistent and complete grouping scheme, however it would only be successful where these data concepts are already known to the business (or there is a willingness to educate the business users on these data concepts).
- Based on an external standard classification – in some cases there might be external standard classification schemes that could be used. Similar to the data concept approach, it would be necessary that such external classification schemes are already known to the regular business users. Examples of such external standards might be the APQC Process Classification Framework[6], or the BIAN (Banking Industry Architecture Framework) Service Landscape[7].

4.5.4 Using the IBM Industry Models to define the business language

The online IBM Industry Model Knowledge Centres contain a lot of the guidance in terms of how to create a glossary of terms.

Industry	Link to Business Terms in the Knowledge Center
Banking and FM	http://www.ibm.com/support/knowledgecenter/SSN364_8.8.0/com.ibm.ima.using/comp/vocab/terms.html
Insurance	http://www.ibm.com/support/knowledgecenter/SSRAR8_8.8.0/com.ibm.ima.product_overview/comp/vocab.html
Healthcare	https://www.ibm.com/support/knowledgecenter/beta/SS9NBR_9.1.0/com.ibm.ima.product_overview/comp/vocab.html
Energy and Utilities	http://www.ibm.com/support/knowledgecenter/SSAKTX_2.0.0/com.ibm.ima.product_overview/comp/vocab.html
Telco	http://www.ibm.com/support/knowledgecenter/en/SSAREY_8.5.1/com.ibm.ima.toc/welcome_version.html

4.6 Business language lifecycle

4.6.1 Initial creation of the business language

The potential set of steps likely in the initial creation of the business language are:

[6] https://www.apqc.org/pcf
[7] https://bian.org/servicelandscape/

1. Establish the overall set of principles that will apply to the data lake business language, based on the needs of the different users and the overall characteristics that would be required for the business language.
2. Set up the overall stewardship and curation processes
3. Identify a specific department or area for the pilot – ideally aligned with a specific business domain
4. Define the set of Business Terms that are necessary to adequately describe the selected business area and it component elements. Typically, this could initially be a flat glossary
    - Consider using the appropriate subset of IBM Industry Models to feed

4.6.2 Requirements

Data Analysts identify the business language or terms that describe the data that is to be transferred between the data lakes for a given business purpose.

Examples would be terms to describe the data related to customer behavior analysis or for a particular regulatory reporting requirement.

Only data that has to be shared as part of the data lake environment is to be included in the business language. The addition of unnecessary terms reduces the complexity of the language and limits the scope of downstream technical modeling tasks to data elements for which there is a real business need.

4.6.3 Analysis

The Data Analyst and Data Definition Owner analyze the requirements to identify where existing Business Terms can be reused, where new Business Terms are required, and where existing IBM Industry Model Business Terms can be leveraged.

While the Data Analyst has expert knowledge of the data being transferred it is likely that they might inadvertently add duplicate or inappropriate terms to the business language. The business language stakeholders provide expertise on the existing business language terms in use with the data lake and the IBM Industry Model terms and how they should be applied across multiple lines of business

Any new terms must follow the promotion process from draft to final published term.

Terms are not directly copied from IBM Industry Models Business Terms into the Development business language area. A new term is created and the description, labels etc. are manually recreated. A mapping back to the IBM Industry Model Business Terms is made using the ‘replaced by’ term association.

These tasks are carried out with IGC workflow switched on.

4.6.4 Refine

Once the initial draft of a new or changed Business Term is created, it is necessary for this term to then undergo a review process which involves a combination of the relevant Business Subject Matter Experts and the business language Stakeholders. This review process would consider:

- Does the Business Term being proposed make sense to the intended Business Users?
- Does the Business Term adhere to the company standards in terms of name and definition structure and completeness and that all of the relevant relationships to other terms and assets are correct?

4.6.5 Deploy

The new or updated Business Term is deployed to the Published Catalog for use by the various users of the data lake. One important consideration is to determine the access/viewing privileges that the various business users should have in relation to this term. For example, if a term is intended only for use by the users of a particular department, then it might be necessary to place the term in the appropriate IGC Category to ensure the required viewing privileges.

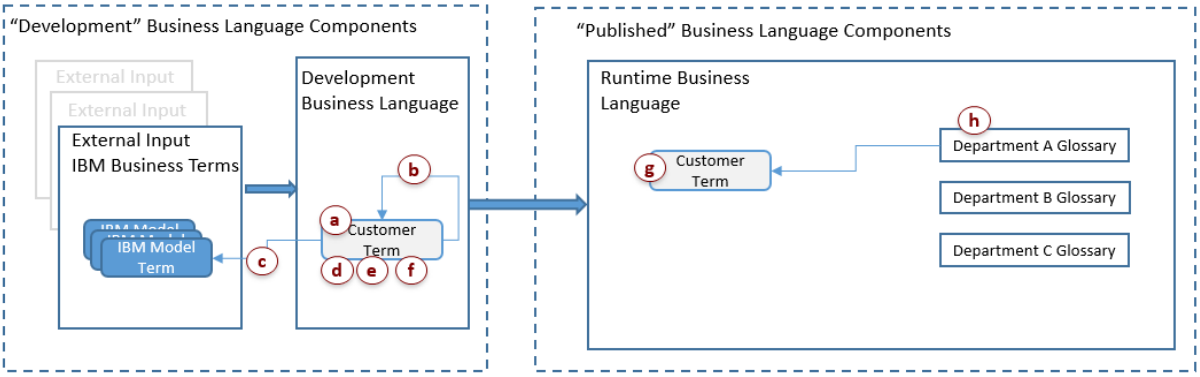
4.6.6 Review

It is important that the data lake business language remains active with, and relevant to, all the relevant data lake users of that term. So, to ensure that part or all of the business language does not go stale, it would be critical that an ongoing review process is in place to ensure that any requests for changes to existing Published terms are handled in an efficient and methodical manner.

All data lake users should have access to a straight forward mechanism to make suggestions regarding improvements or corrections to existing published terms or to make suggestions for completely new terms to be added. If such a review process is not in place, then it would almost certainly impact the credibility of the business language..

4.7 Example of a typical end to end flow of a term

This section shows an example of a typical flow concerning the evolution of a new or modified term across the different components of the business language landscape.



The diagram above shows a typical flow regarding the creation or modification of a new term, with the each of the steps described in the table below.

Step	Description	Role Involved	IGC Workflow Status
a	The Business Term is created or modified. Any additions or changes to the attributes (for example definition) are made. The Term is placed in the correct category.	Term Author	Draft
b	Make any associations between the New terms and existing Business Terms. This provides context with used of is-of or has-a associations.	Term Author	Draft

c	Make any associations between the new business term and any IBM Industry Model Terms (if they were inspired from these terms).	Term Author	Draft
d	Review new Business Term to ensure is structurally correct and make comments where appropriate. Note this role is not allowed to make changes to the terms – only make comments. Note that it is likely to provide users with Glossary Admin rights in order to see the Dev Glossary.	Data Modeller	Draft
e	Begin Approval Workflow after applying suggested structural corrections	Term Author	Draft > For Approval
f	Review Terms against checklist for example, are all fields completed, business approval obtained from Stakeholders?	QA team member	For Approval > Approved
g	Publish Terms to the Runtime business language in the published data lake Catalog.		Approved > Published
h	Make any necessary enhancements to Departmental Glossaries (for example, associate as Synonym)		Approved > Published

4.7.1 Ongoing maintenance of the business language

There are a number of considerations that might need to be taken into account relating to the ongoing maintenance of the business language. While these considerations will vary, what is definite is that there will be a need to accommodate such evolution of the business language. It is not realistic to expect that the initial definition of the business language will be complete. Factors that would necessitate such maintenance include:

- Ongoing evolution of the broader business and technical environments in which the business language resides.
- New versions of external models (for example, IBM Industry Models) and new versions of other relevant standards.
- Ongoing feedback from the users of the data lake. The business language has to be a living language and so there must be a process in place to manage what should be constant feedback from the various data lake users.
- The growth of the business language to accommodate the addition of new artifacts (for example, addition of new Rules and Policies).

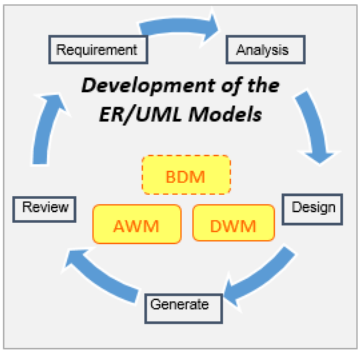
How to manage the governance of the business language - lack of stewardship makes it difficult to realize this as needed. IT functions as "custodian" and does a good job; however, without full business participation, success here will be limited.

4.8 Lifecycle output

The main output from this lifecycle will be the set of final published Terms and associated categories for use by the data lake users.

5 Modeling the data lake repositories.

5.1 Modeling lifecycle overview



The purpose of the modeling lifecycle is to provide a consistent set of structures for use when deploying the data lake repositories. IBM Industry Models can provide a basis for a single consistent canonical set of structures that would aid the usability and extensibility of the data lake.

The purpose of this chapter is to describe the set of considerations and steps that are related to the creation and deployment of data models and associated artifacts specifically in the context of a data lake deployment.

There is a much broader description of the main Data Modeling steps in the relevant sections of the IBM Industry Model Knowledge Center.

The Data Models associated with a data lake deployment are maintained through an iterative lifecycle:

1. The Requirements for the specific Data Modeling activity are received from the relevant business users. In the context of the data lake it is assumed that the provision of such requirements is done using the standard business language of the data lake.
2. The Analysis of the structural and content implications of the business requirement, especially as it might relate to any previous Data Modeling activities for data lake Repositories.
3. The Design of the platform-specific data models needed to provide the structures for the different data lake repositories. The design approach might potentially be quite different for these different repositories, for example the level of normalization needed for deployment to Deep Data HDFS structures, Data Scientist Sandboxes or RDBMS Information Warehouses.
4. The Generation of the required Physical Data Model and the provision of that Data Model to the data lake operations team
5. The Review of the feedback of the suitability of the data model both in terms of how it addressed the specific business users needs as well as if it needed any significant customizations due to performance or volumetrics challenges.

5.2 Inputs to the lifecycle

The input to this lifecycle is characterised by a combination of business and technical considerations that will influence the approach and output from the modeling activities in relation to the data lake.

- The range of business issues and sets of Business Users to be addressed. Fundamentally the need for a standard canonical set of structures across the data lake grows exponentially in line with the number and range of different parts of the business to be supported by the data lake.
- Approaches to the data lake in terms of the need for enforcing a schema. There is a wide range of opinion as to what is the appropriate collection of technology to be utilized across the data lake repositories and there is likely to be a general philosophy in terms of the

approach to the definition of a common schema. There is a spectrum ranging from data for which there is no need for any enforcement of schema (raw data that simply needs to be placed as-is in the data lake) to data structures where there is a high perceived need for structures (the typical Information Warehouse). There is also likely to be a number of artifacts in the middle of that spectrum where there are pros and cons that could be argued against the need for enforcing structure – for example data that is stored in HDFS that combines data from a number of different sources with different schema assumptions.

5.3 Main actors/roles involved

The main actors/roles involved in this lifecycle fall broadly into two different areas.

Business-related actors/roles

The first set of actors/roles are those who are responsible for ensuring that the business requirements for the area to be modeled are adequately and clearly communicated to the more technical oriented actors.

- **Business Subject Matter Experts** – who will provide the specific requirements and ensure that any questions or requests for clarification from the technical oriented users are addressed. They are also likely to be involved in reviews of the models as they are created prior to the generation of any final PDMs.

Technical oriented actors/roles

These are the people who are responsible for the actual translation of the business requirements to the set of physical artifacts needed to define the structures for the relevant repositories in the data lake.

- **Data modelers** - Translates the business needs defined as project scopes in IGC into scoped design data models.
- **Data architects** - Transforms the scoped LDM's into physical data model(s), targeting the latest production Physical Data Model(s). The data architect is responsible for setting enterprise data standards as well as for defining the structure and placement of solution data, including deciding the appropriate combination of Hadoop and RDBMS technologies to be used in the data lake. For transformation of the LDM to Physical Data Model(s) in the Hadoop environment, the Data Architect will be responsible to decide which denormalization patterns should be applied on specific areas and structures of the LDMs to support the Hadoop specific constraints
- **Database administrators** – Performs physical data modeling tasks to refine the physical data model(s) by adding specific physical properties to bring it one step closer to a functional, deployable model.

5.4 Artifacts involved

The main IBM Industry Model artifacts involved in this lifecycle are the following.

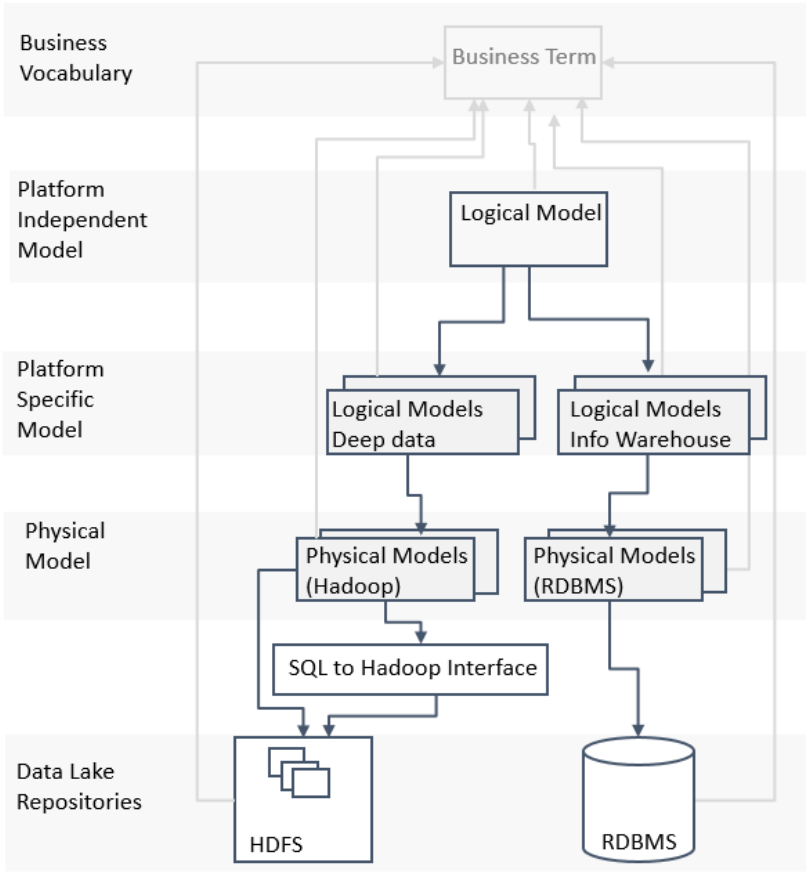
Business Data Model – primarily used as the high-level logical blueprint for subsequent design-level deployments. However, in some cases this model can also be the starting point for defining the data model structures needed for deployments to Hadoop/HDFS.

Atomic Warehouse Model – predominantly used as the basis for the deployment of the conventional Inmon-style atomic data warehouse. It can also be used as the basis for deployment of the HDFS Deep Data structures, where there is a need for a persistent schema in this repository.

Dimensional Warehouse Model – predominantly used in the deployment of either the data mart layer and/or the Kimball-style dimensional warehouse. Possible investigations required to establish the use of this model to also enable the deployment of certain aspects of the data structures needed to support the Data Scientists.

5.5 Main Steps

This section describes the main steps and associated considerations for the data modeling activities in relation to a data lake[8].



The diagram above shows a likely overall flow from an initial Business Term, as defined in Lifecycle 1, through the series of logical and physical models to the generation of the structures required for the relevant data lake repositories.

The main feature of this flow is the central role of the Platform Independent Logical Model which is used as the means of enforcing a consistency of structure and terminology in the downstream models, irrespective of whether they are intended for Hadoop or relational repositories.

Given the potentially large array of different technologies to be used in the data lake repositories, it might be necessary to then introduce series of Platform Specific Models, both logical and physical models to achieve the necessary set of structures needed across the data lake repositories.

[8] As part of the delivered IBM Industry Models content, there is accompanying documentation on the general steps to be followed when developing the data models.
For example :
http://www.ibm.com/support/knowledgecenter/SS9NBR_9.1.0/com.ibm.ima.using/usi_ibm_im/usi_ibm_im.dita

5.5.1 Requirement

This step relates to the review of the requirement regarding the particular data modeling activity to be followed to enable the update or creation of a structure for one of the data lake repositories.

During this step the data modeler is likely to communicate with the relevant business subject matter experts to ensure that they have an appropriate and complete understanding of the requirement as defined using the business language in IGC plus any accompanying documentation.

5.5.2 Analysis

This step relates to the data modeler and potentially the data architect carrying out the necessary analysis to determine the overall broad structure and composition of the area to be modeled.

The main influencing factors in this step include:

- The role of reference/input models such as the IBM Industry Models to provide a necessary starter set of structures for use when building out a new or enhanced data lake repository.
- Any models that have been used as part of the generation of previous data lake repositories. This is to ensure an appropriate level of consistency across the data lake repositories as well as avoiding unnecessary duplication/replication of repository structures.

5.5.3 Design

In terms of designing the models to be used in the creation or enhancement of the different data lake repositories, the key influencing factor is the need to manage the modeling process for potentially quite radically different target technologies.

For the design considerations when deploying to more traditional data warehouse technologies there are a range of documents and publications.

For the design consideration when deploying to Hadoop technologies, a separate accompanying document exists[9].

Another key consideration is the retention of any mappings to the IGC business terms that existed in the upstream logical platform independent models. With most target data lake technologies, it is possible to ensure that such business term mappings that might have been imbedded in the logical models are inherited through the different logical to physical transformations so that the resultant generated physical assets retain the mappings to the business terms. This inheritance of mappings to the business terms is important as it ensures that the generated physical artifacts are mapped to the key relevant business terms in the business language.

5.5.4 Generate

This step is concerned with the generation of the appropriate physical structures to be provided to the team responsible for the management of the data lake repositories.

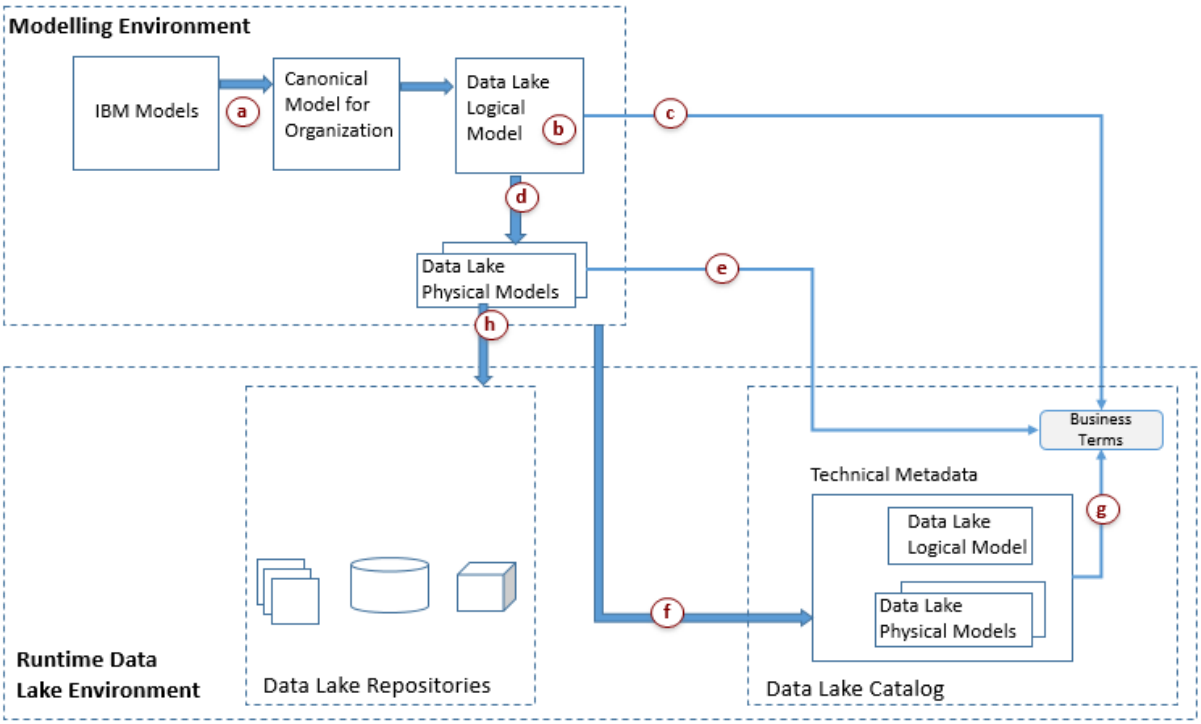
[9] "Guidelines for deploying IBM Models to Hadoop "

5.5.5 Review

This step is concerned with accommodating any feedback on the accuracy and suitability of the output from the Data Model process. Typically, this feedback might come from the DB administrators or other data lake operations personnel or from the Business Analysts in terms of the required business coverage of the data model artifacts.

5.6 Example of an end to end flow of a data model artifact

This section provides a description of a typical end-to-end flow of the data model artifacts from their initial update/creation to the subsequent uses both in supporting the business language and in deployment to the data lake repositories.



The diagram above shows a typical flow regarding the creation or modification of a new term, with the each of the steps described in the table below.

Step	Description	Role Involved
a	Copy entities from the IBM Industry Model LDM into a Canonical Model LDM. The copy will retain the mappings to the IBM Industry Model Business Terms. The persistence flag of all entities, attributes and relationships will be set to non-persistence as default. The IDA LDM and PDM files also be source controlled/versioned	Data Modeler
b	Changes are made to the data lake subset model for example, removing attributes, changing data types etc.	Data Modeler

	• Will all changes be made in this model with the IBM Industry Model LDM being used a read-only reference model? Will the IDA LDM and PDM files also be source controlled/versioned during this process? Will multiple data modelers be working on these files?	
c	Associations are made between data lake LDM and Business Terms using the IDA Eclipse plug-in.	Data Modeler
d	Data lake LDM is transformed to update the data lake Physical Data Model.	Data Modeler
e	Associations between data lake PDM and data lake Business Terms are confirmed using the IDA Eclipse plug-in. They will have been generated automatically by the transform.	Data Modeler
f	Current data lake LDM and PDM are imported into IGC using IMAM. Steps for this import to avoid bringing in any associations to the IBM Industry Model to be agreed and tested.	Data Modeler
g	Associations between data lake Data Models and data lake Business Terms are updated in IGC using the IDA Eclipse plug-in.	Data Modeler
h	The PDM is used as the basis for the creation/update of the data lake Repositories, as required	

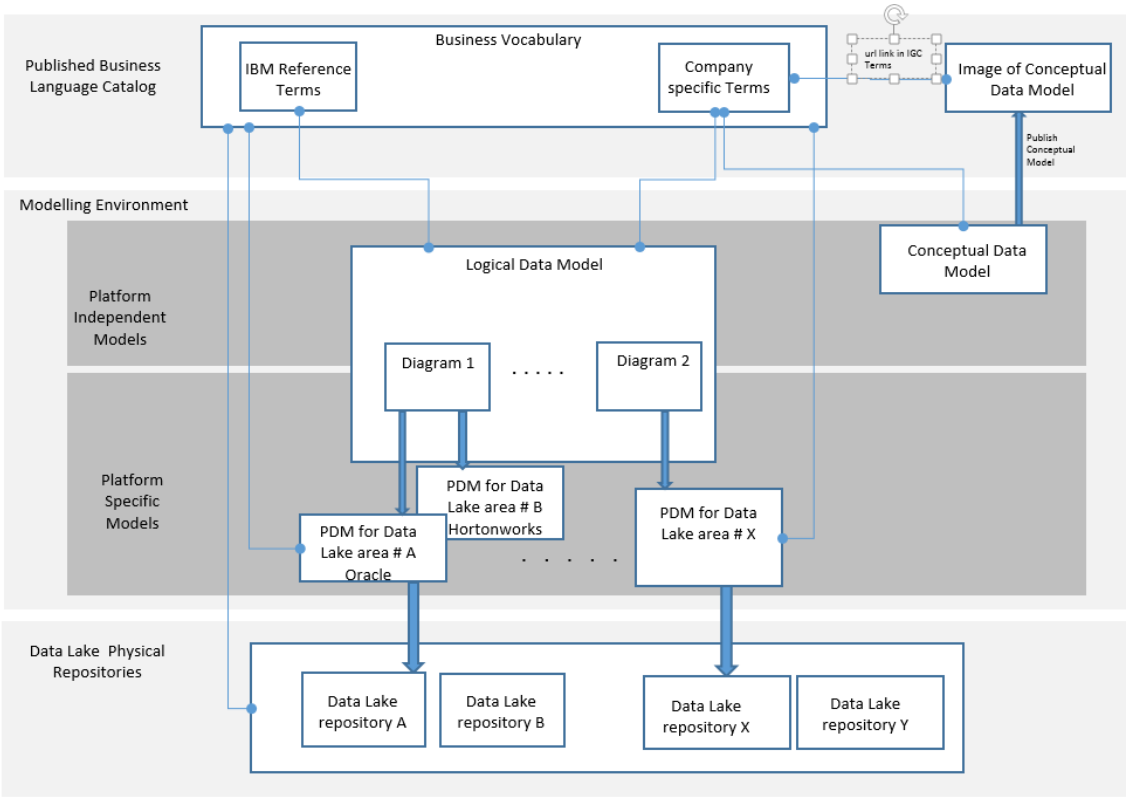
5.6.1 Detailed example of the end to end flow

The diagram below shows in more detail the possible flow and associated mappings from the IBM and company specific terms in the Business Vocabulary to the various models in the modeling environment and how these models in turn are used in the generation or maintenance of the different data lake repositories.

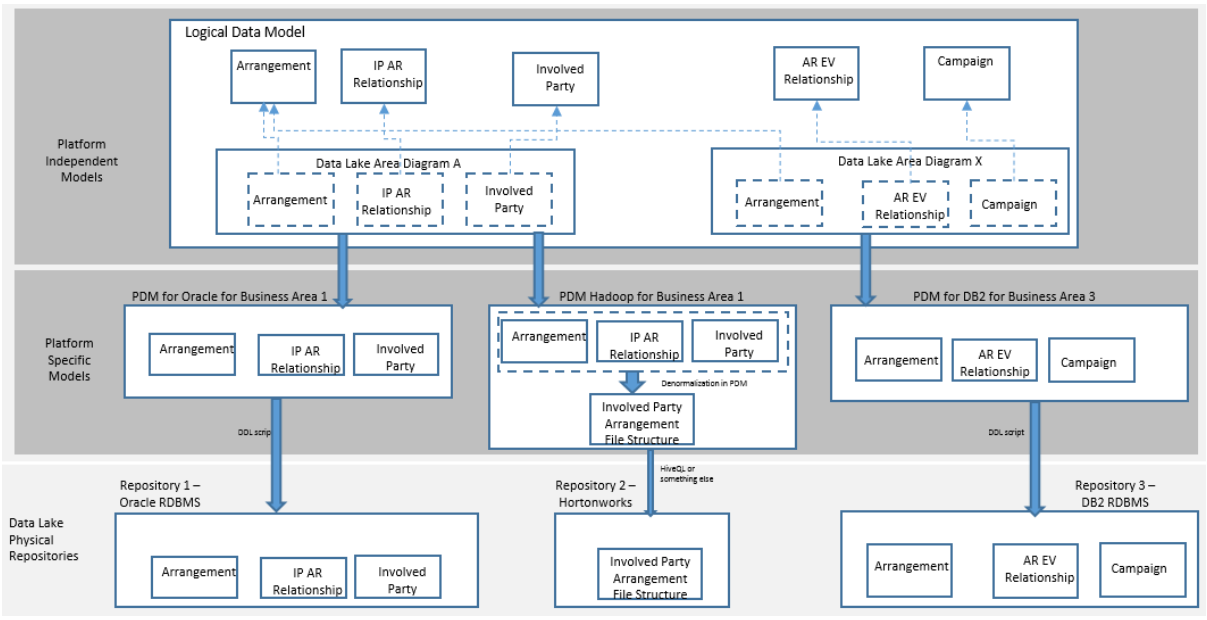
In this example the representation of the different areas of the data lake repositories to be generated are managed as diagrams in the ER modeling tool.

In this way a single overall platform independent model of the overall data lake can be represented with the diagrams representing the filters of the subsets needed for specific repositories.

The Physical Data Models (PDM) which are generated from the logical data models are the point at which any platform specific structural changes are made to align with the constraints or other factors of the specific physical repository.



This example can be expanded even more with some specific model examples as shown in the diagram below.



In the diagram above the Logical Data Model has a single set of actual entities (for example, Arrangement, Involved Party, etc), with each of the diagrams having read-only views of these entities. It is assumed that in each entity the specific subset of attributes that would be persisted are indicated using the IDA Persistence flag for that attribute. The setting of such a persistence flag is cumulative (all the attributes needed for all diagrams are set as “persisted” in the actual entity).

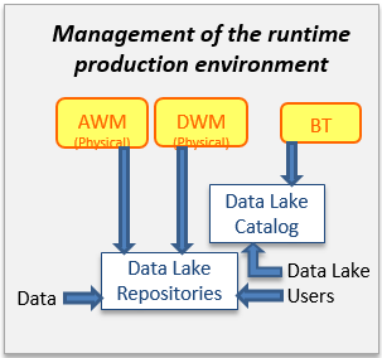
The above diagrams provides an example of the possible of some of the possible transformations of these entities as they are changed to be used to deploy to the different physical data lake repositories which are deployed using different technologies.

5.7 Lifecycle outputs

The output of this lifecycle are the physical models that are provided to the data base administrators and other data lake operations personnel for use to guide the deployment of new or changed data lake repositories.

6 Using the IBM Industry Models in the data lake runtime environment

6.1 Data lake deployment lifecycle overview



The purpose of this third lifecycle is the management of the incorporation of the models-related artifacts into the runtime data lake environment. The intention is not to describe the broader overall management of the data lake itself but the specific activities and considerations relating to the treatment of the interaction between the ongoing data lake operations and the parallel lifecycles related to the maintenance of the business language and the design/development of related model structures.

The main aspects of this lifecycle would include:

- The deployment of the published business language by the different data lake users.
- The deployment of the generated data model artifacts into the data lake repositories.
- The use of the business language by the Business Users to assist their search and discovery activities across the data lake.
- The use of the business language by the data lake operations team, specifically the ongoing mapping/tagging of any data lake components to the data lake catalog.
- The management of feedback from the various data lake users to the business language and data modelling Lifecycles.

6.2 Inputs to the lifecycle

The inputs to the portions of this lifecycle that are relevant to IBM Industry Models would mainly fall into two main categories:

- Business language terms and categories that have been promoted for use as part of the data lake catalog.
- Physical data models that have been generated from the data modeling Lifecycle, intended for use in the creation or maintenance of the data lake repositories.

6.3 Main actors/roles involved

The interaction of the various models artifacts and the ongoing operations of the data lake will fall into the responsibilities of a number of different actors:

Business Language Users responsible for the deployment and maintenance of the business language terms to the data lake catalog:

- Metadata managers

- Data Stewards
- Data Curators
- Data Governance Administrators

Data Repository Operations Users responsible for the deployment and maintenance of the necessary physical Data Models to the Data Repositories:

- Data Base Administrators
- Data lake Integration Administrators

Data lake users who avail of the IBM Industry Models elated artifacts in their regular use of the data lake

- Business Users – Who are mainly concerned with the use of the catalog to assist in their use of the BI tools that are connected to the data lake. Catalog is used to provide the basic Search and support for self service capabilities.
- Data Scientists- Who are mainly concerned with the creation and maintenance of sandboxes and who are using the catalog to help with the search and discovery of the assets across the data lake.
- Data lake Operations Staff – who are concerned with the daily running of the data lake. They will use the catalog to help them understand the relationships and meaning of various data lake assets they are maintaining and supporting.

6.4 Artifacts involved

The main IBM Industry Model related artifacts to be involved in this lifecycle include:

- The subset of the Business language that has been promoted for use in the Published Glossary.
- The generated physical versions of the Atomic Warehouse Models and Dimensional Warehouse Models.

6.5 Main data lake activities involving the IBM Industry Models.

The main activities involved in the usage of the IBM Industry Models can be broadly broken into the following main categories.

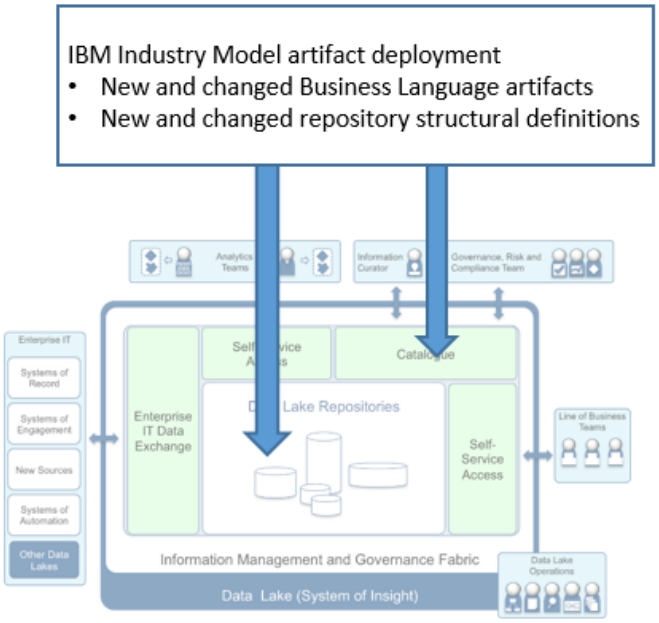
1. The deployment of the IBM Industry Models artefact to the data lake runtime environment.
2. The activities of the data lake users against the data lake repositories.
3. The addition of new or changed instance data into the data lake repositories.

6.5.1 IBM Industry Model deployment activities

This section is concerned with the specific activities regarding the deployment of the various IBM Industry Model artifacts to the data lake runtime environment.

This is mainly related to the following:

1. The deployment of new or changed model artifact to the data lake catalog.
2. The deployment of new or changed structural definitions (for example, Physical Data Models) to the data lake repositories.



In distinguishing between the principles behind the deployment of Terms to the data lake catalog and physical data models to data lake repositories, one key consideration should be given to the required relationships between the catalog terms and any data lake repositories. For example, it is likely that there could be deployment of terms to the catalog for which there is no equivalent data model structures in the data lake repositories, such as in the case of Terms needed to describe either unstructured data for which there is no need for a data model. However, any deployment of a data model for use in the data lake repositories for which there is no equivalent business term in the catalog should at least be investigated.

Deployment of new or changed model artifact to the data lake Catalog

This relates to the Business Terms and Categories that were created or updated as part of the overall "Defining the business language" as described in Chapter 4. This would include:

- Ensuring any new/changed terms are deployed in a way that is aligned with any category hierarchies.
- Informing the relevant users of the new or changed catalog artifact.
- Ensuring that any relationships/mappings from the new/changed catalog artifact to logical or physical assets are updated or created accordingly.
- Ensuring that any additional catalog assets such as Data Quality Rules and Policies are deployed and the appropriate relationships main or maintained with the correct business terms.

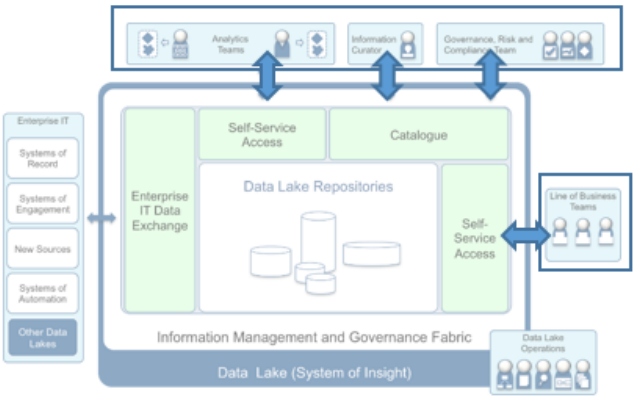
Deployment of new or changed structural definitions to the data lake Repositories

This relates to the Physical Data Models that were created or updated as part of the overall "Modeling the data lake Repositories" as described in Chapter 5. This would include:

- Ensuring that the appropriate Physical Data Models are handed off to the relevant data lake operations staff (for example, DBAs).
- Ensuring that any associated data lake catalog mappings are updated, in the case of the PDMs being supplied with mappings to catalog terms.
- There should be guidance in place to investigate situations where a PDM is being provided where there are no mapping to catalog terms. In such cases, unless these are pure technical artifacts, there should be a question about creating or updating data lake repositories for which there is essentially no business meaning.

6.5.2 Activities of data lake users

This section is specifically concerned with the relationship between the deployed IBM Industry Models artifacts and any new or changed instance data that impacts the data lake Repositories.

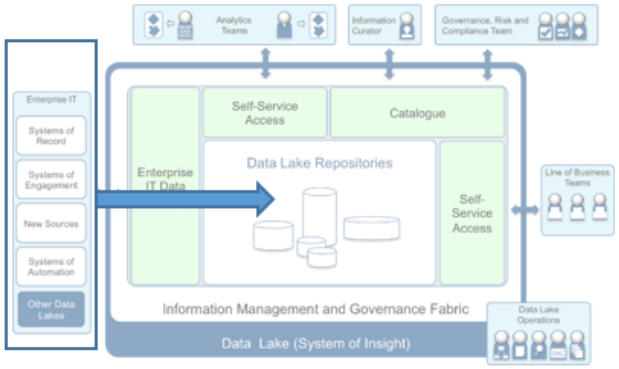


The typical activities would include:

1. Ensuring that all of the necessary attributes are available to the data lake users for a particular Data Catalog or Data Structure.
2. The data lake users providing feedback on the data lake catalog.
3. The data lake operations staff providing feedback on any changes needed to the model-driven data lake Repositories in areas such as additional fields/metrics required, additional data types needed, new structures needed to be created.
4. Monitoring the usage patterns of the data lake users, especially where there are any local customizations to suit specific departmental/local needs, or where there is a possibility to identify common questions/combinations/access paths which can be formalized to assist self service.
5. The data lake Governance personnel will provide feedback on the suitability and workability of the data lake catalog from a security, lineage and comprehension perspective.
6. The Data Scientists might provide input to new or changed Catalog Terms or data lake Repositories based on their investigation/discovery activities. There might also be requests from the Data Scientists to include specific Sandboxes as part of the overall set of governed data lake repositories. Finally there might be ongoing needs to ensure the coordination/alignment of their data sets with the production aspects of the data lake?

6.5.3 New or changed instance data in the data lake Repositories

This section is specifically concerned with the relationship between the deployed IBM Industry Models artifacts and the impact of any data instance additions or changes to the contents of the data lake repositories.



The typical activities might include:

1. Determining if any new data element needs to be mapped to the catalog.
2. Determine if any new data element should undergo further transformation to align with a standard structure in the data lake.
3. Any changes to the data require an associated change to the data lake Catalog.

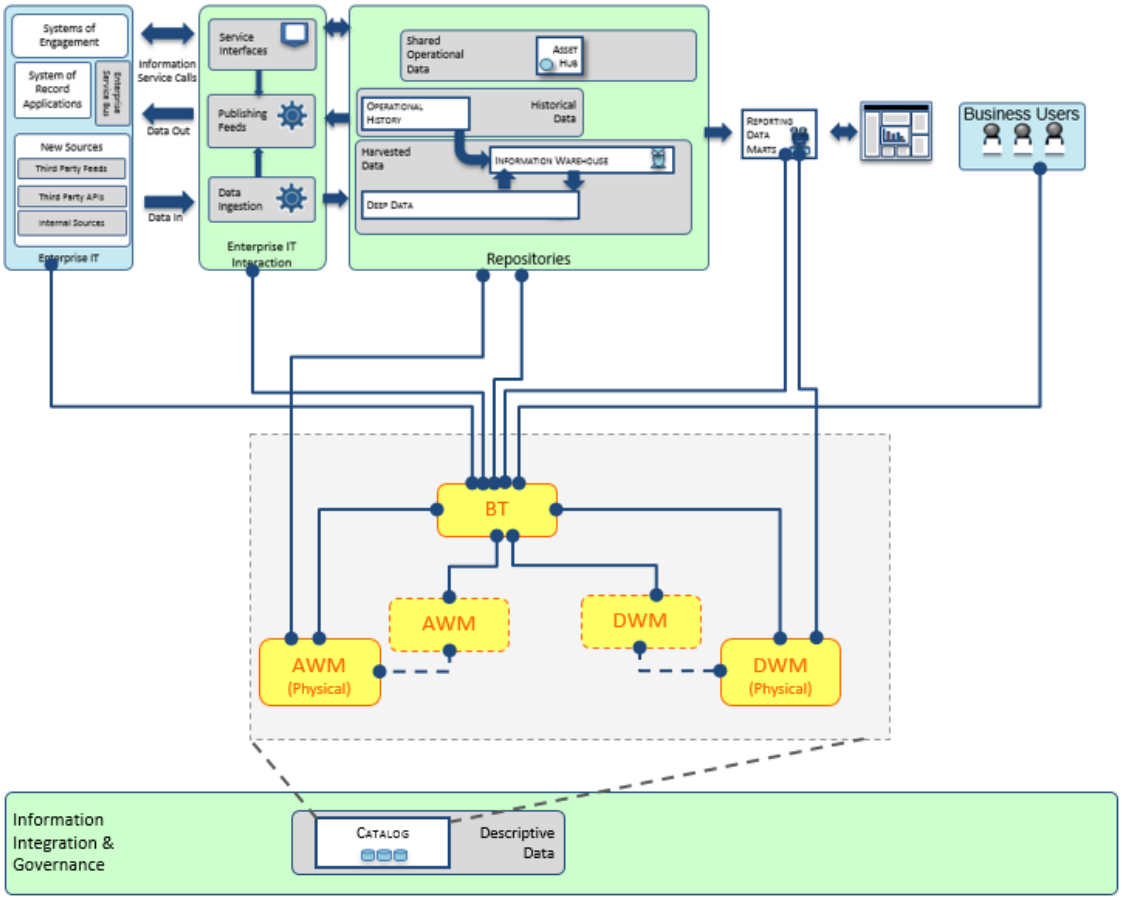
The most significant activity in relation to such new or changed data concerns the mapping of any new data lake Repositories to the data lake Catalog. In order for the underlying data lake governance activities to be efficient it is critical that the data lake catalog is kept aligned with any changes to the data lake repositories. A key question here relates to the most appropriate process for the “tagging” of any new or changed data lake repository structures.

Such mappings are rarely 1 to 1 but in many cases it is the case that 2-3 terms are needed to describe a column in a table

Central Role of the Business Terms in the data lake mapping.

The Business Terms in the data lake catalog are the central point for most of the mappings that might be carried out to against various data lake repositories. While it might be possible to also include copies of the associated Data Models, these are included in the data lake catalog only as supporting artifacts to provide more context to the business terms. It is assumed that the bulk of the mappings to the data lake Catalog would be done to the Business Terms.

The diagram below shows this central role of the business terms. In this example the various data lake repositories are mapped to the business terms. In addition, it is possible that certain upstream System of Record artifacts and downstream Data Mart artifacts might also be mapped to the business terms in the data lake catalog as necessary.



The above diagram also shows the potential relationships between the different data lake catalog artifacts. In addition to the capability to map between the data lake artifacts and the business terms, it is also possible for such mappings from the data lake artifacts and the Physical Models.

Approaches to the mapping of Business Terms and the data lake artifacts.

The best practices regarding the optimal approach to mapping the various data lake artifacts to the data lake catalog is still evolving. However, some of the possible considerations in this area are:

1. What should be mapped? This is essentially related to what the enterprise feels should be subject to a common governance fabric. While certain aspects are reasonably obvious choices (for example, any cross LOB data lake repositories), other artifacts might need to be considered (emerging Data Scientist sandboxes, local data marts, upstream Systems of Record). In other words, should it be mandatory for artifacts with a meaning/relevance to the business to be mapped to the catalog?
2. To what data lake Catalog elements should mapping be carried out? While it is possible to map to the Physical Models in the data lake catalog, as a starting point it would be suggested to ensure that all business meaningful data lake artifacts are mapped to at least one term.
3. At what level of granularity should artifacts be mapped? While there is a cost to making and maintaining the mappings, the ideal is to have artifacts mapped to the lowest level of granularity (for example, Column level).

4. When to do the mappings? Certainly the ideal is that the mapping to the data lake catalog is done as soon as possible, certainly as soon as that data lake repository is populated and available to the general data lake users – such users will be less likely to find/understand the data lake artifact without a mapping to the data lake catalog.
5. Who should do the mapping? Ideally the technical owner or steward of the artifact should be considered the owner of the mapping. It is likely that this person has an understanding of the technical aspects of the data lake artifact so that they can ensure it is mapped to the correct data lake Catalog item.
6. Who maintains the mapping?
7. What happens when there isn't anything suitable in the data lake Catalog to map to? There should be a process to enable the various data lake users to initiate a request for an addition or change to the data lake catalog.

6.6 Considerations for using IBM Industry Models in a data lake runtime environment.

This section describes the various possible considerations regarding the use of IBM Industry Model artifacts in the data lake runtime environment. Some or all of these considerations might be relevant to different data lake deployments.

6.6.1 Data lake Governance

A key role of the IBM Industry Models in the data lake runtime environment are their use to assist with the underpinning of the overall data governance of that lake.

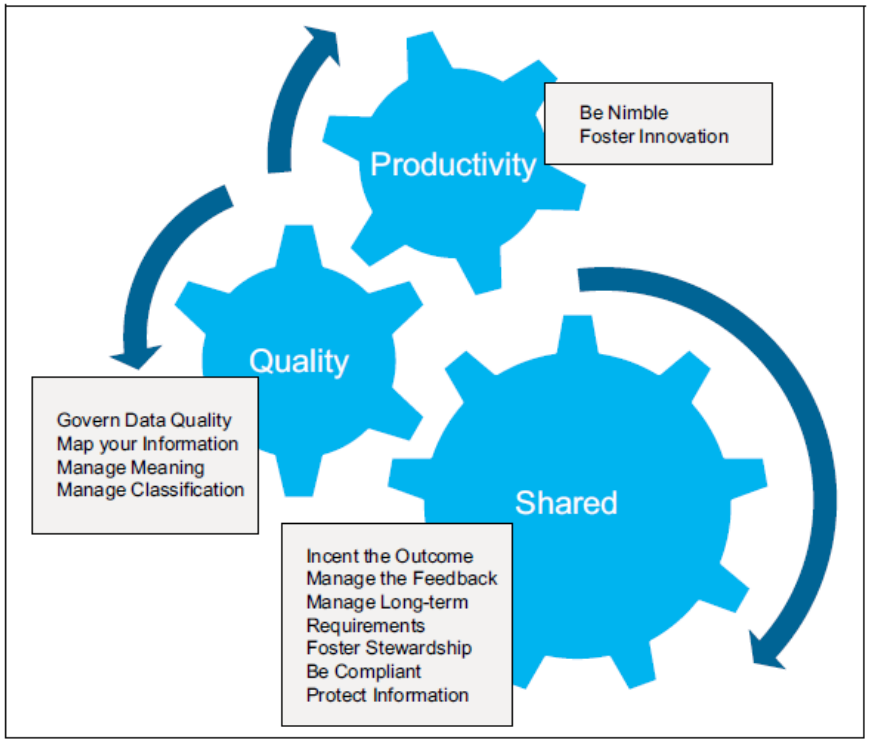
At the minimum there should be the tagging of all content in the data lake to a common Business Vocabulary so that at least a standard business meaning can be asserted across the data lake. The roles and steps to be considered as part of this setup are contained in the IGC Knowledge Center[10]

Role of IBM Industry Models in the overall set of data Governance principles

The IBM Redbook "Information Governance Principles and Practices for a big data Landscape"[11] describes a number of different principles to be taken into account when defining the appropriate level of data governance for a big data /data lake environment.

[10] Forming an Information Governance Team : https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.5.0/com.ibm.swg.im.iis.bg.bestp.doc/topics/c_Establish_gloss_governance.html

[11] http://www.redbooks.ibm.com/abstracts/sg248165.html?Open



The above diagram outlines the grouping of these 12 principles into the areas of

- Productivity
- Quality
- Shared

The overall goal of these principles is "*governing information for an organization is to move information as quickly as is practical while keeping the quality as high as is practical and as secure as is practical."* IBM Industry Models can play a distinct role in underpinning a number of these principles to assist in attaining that objective:

- Manage the long term requirements – The use of a standard business language and standard definitions can greatly assist in the consistent definition of requirements and the identification of reuse or overlaps. Especially the deep taxonomies of the Banking/FM, Telco and Retail models and the central Business Data Model of the Insurance, Healthcare and Energy & Utility models can provide the practical basis for the precise identification of and placement of requirements
- Foster Stewardship – The use of the IBM Industry Models to provide a consistent basis for the stewardship across the data lake assets.
- Manage Meaning – The provision of a common consistent business language, as defined in the IBM Industry Models, across lines of business is critical to ensuring common meaning.
- Manage Classification – The use of the various term and data model hierarchies can be used as an important input to the overall classification of the catalog elements and the associated data lake repository assets.

6.6.2 Levels of ownership of assets in the lake

Ownership considerations are related to which users or organizations will own the deployed data structures. With centralized or enterprise-level ownership, the need to enforce a common schema becomes critical, whereas with locally or personally owned artifacts or with artifacts owned by an external organization, such schema considerations are far less important. A high level of enterprise or central ownership of a particular set of artifacts means a stronger need for a commonly agreed upon structure or definition, hence the increased need to have such artifacts derived from a model[12].

What is the appropriate level of ownership for data in the lake? Data that is just for the cross enterprise use or should we also include local or even personal data. There would be a possible need to enable support for local overriding/augmenting of enterprise data.

It is important to separate the types of Business Users and their "ownership" of data from the Technical Users. The "ownership" will be different across these two groups of users.

- The Business Ownership will mainly focus on who is responsible for the correct definition, understanding and use of specific data lake repository elements from a business meaning perspective. Business users are concerned with different aspects (LOB data, SoR data, enterprise Insight data, specific department data, sandbox data)
- The Technical Ownership will mainly focus on the ownership of the inputs and outputs of the various data lake operations components that move the data through the data lake repositories.

It is also important to consider the other facets of "ownership",

- who has access?
- who has the entitlement to manage/govern the data?
- What about other data assets (for example, Data Quality Rules, Privacy Policies)?
- What is the appropriate level of metadata and governance for the data depending on the business context?
- When to incrementally curate and increase governance as data proves its value?
- Track usage, resolve anomalies

6.6.3 Classification of the data in the Lake

The overall potential set of classification schemes that might need to be considered in a data lake are described in Chapter 2.3.3 of the IBM Redbook “Designing and Operating a Data Reservoir”. While some of the various potential classification schemes mentioned (for example, Role Classifications, Activity classifications) can be found in the data models as enumerations or reference data, the main focus for the IBM Industry Models would be to underpin the Semantic Classifications – the schemes that classify the meaning of an element in the data lake.

[12] Text taken from IBM Journal or Research and Development Article “ Applying Data models to big data Architectures” http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6964825

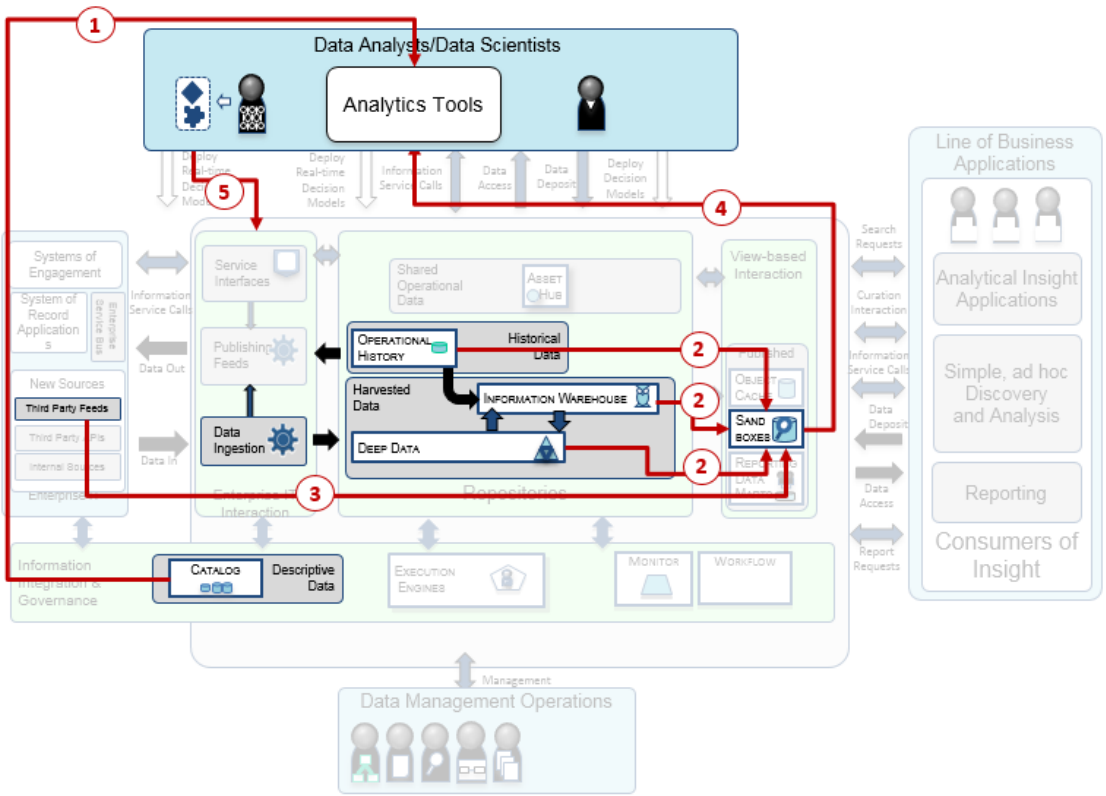
The IBM Industry Models provide the basis for the "Subject area classification" where the categories or groupings provided in the IBM Industry Models can be used to provide the basis for the classification of the Subject areas in the data lake. Similar to the set of considerations described section 4.5.3 of this document, the considerations for determining the ideal classification of Subject Areas could include:

- Business groupings already used by the business
- Data concept based
- Based on an external standard classification

.

6.6.4 Data Scientist Sandboxes

A key aspect of the data lake is the need to accommodate the needs of the activities of the Data Scientists. This is typically a group of users separate from the general business users who have a wider range of access to data and whose main role is the discovery of new insights to be used by the rest of the organisation.



The above diagram shows a typical flow of how the Data Scientist group might interact with the data lake assets. The typical steps outlined in this diagram are:

1. Query the catalog to see what data sources and data assets exists to address the particular business question. IBM Industry Models can be used here as the basis for the standardized catalog for use by the Data Scientists as well as the broader set of Business users.
2. Extract information from multiple stores in the Sandbox:
    - Ability to extract data from Relational and NoSQL data stores
    - Where there is a need to ensure reuse or consistency of the data being provided to Data Scientists, it might be possible to define a set of standardized structures, deployed from the models for use by the Data Scientists in creating their Sandboxes. This might help reduce some of the Data Scientist prep time.
3. In addition, discover and input to the sandbox any relevant sources of data not in the data lake (for example, 3rd party data sources). Given that this data is coming from outside the lake, it may or may not be recorded in the catalog. If it is, then using the IBM Models *Supportive Content* construct to indicate such external data sources might be beneficial.
4. Carry out Advanced Analytics on the info on the sandbox.
5. Ability to deploy decision models into the Data Management Area. If the creation of new or changed decision models means a change to the catalog, then this should be done via the business language lifecycle described in Chapter 4.

Other considerations - characteristics

Currently there is no explicit support in the IBM Industry Models for the types of data structures typically needed by Data Scientists. However, a number of potentially common patterns of use are emerging on some of the structures being used by organizations to support the Data Scientists, and it might be possible to derive some of these from the IBM Industry Models. Specifically, in the potential use of some of the aggregated/summary structures that can be used to provide the highly flattened structures used by Data Scientists. For example, some of the DWM dimensional structures or the “Summary Area” in the AWM of the Banking Data Warehouse might provide the basis for the creation of such flattened structures.

The concept of "pre-doing" some of the grouping of commonly used Data Scientist metrics for a particular area would be related to where in the lifecycle that data is. For example, for newly loaded data into the warehouse, there is little sense in trying to create such predefined groupings of Data Scientist metrics, as the typical usage patterns and frequently required collections of data fields have not yet been understood. However, such pre-defined structures might be possible for data that has been in the data lake for some time, as it would have been possible to observe and identify the most commonly used metrics,

6.6.5 Models and Security

Given the wide range of potentially sensitive information that is likely to be stored in an enterprise wide data lake, enforcing the appropriate level of security and access control across the data lake repositories is critical. So it would be important to ensure that the deployment of an IBM Industry Models artifacts into this runtime environment conforms with the overall data lake security layer.

This section outlines the main considerations for aligning the IBM Industry Models with a data lake security mechanism.

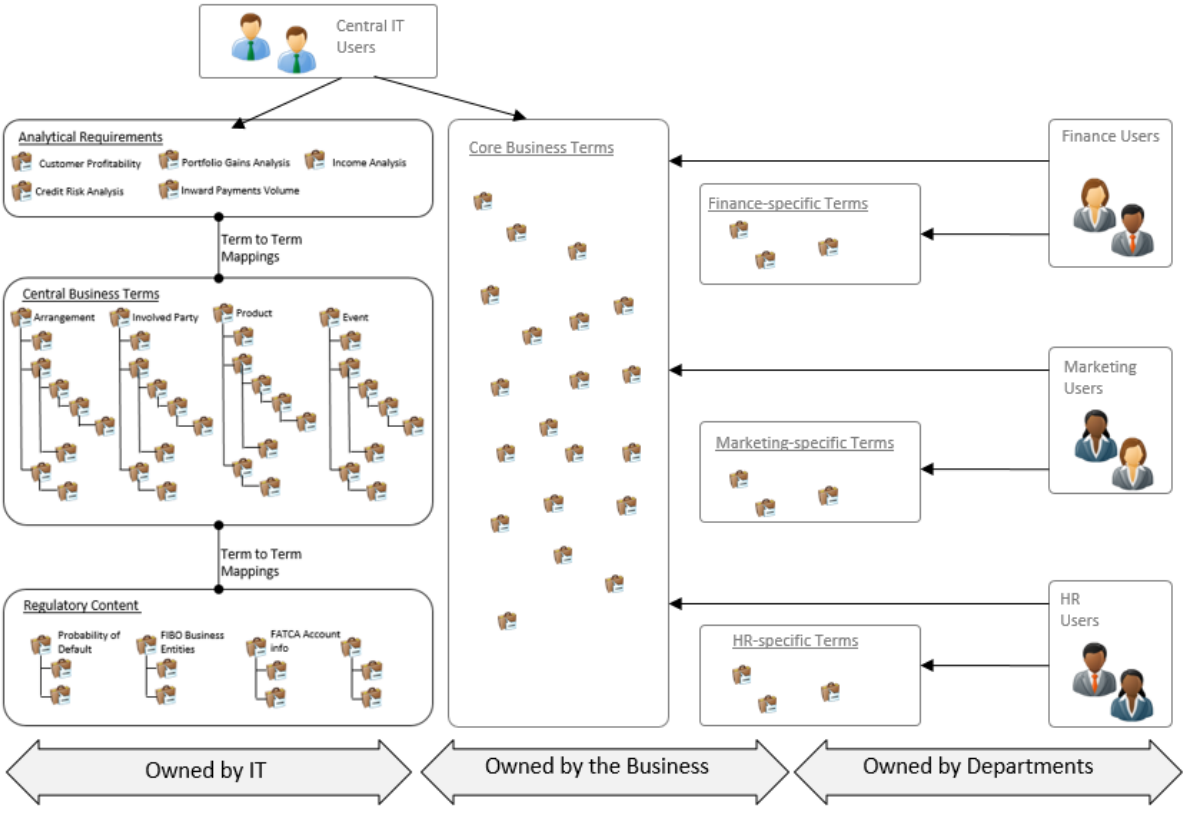
While the overall description of what is needed to enforce data lake security is not in the scope of this document, it might be helpful to describe some of the basic security approaches that the early adopter data lake implementers are using, and to identify how these can be aligned with the relevant IBM Industry Model artifacts.

1. Managing the Viewing privileges on the Catalog.
2. Using the Catalog to underpin the privileges of the data lake repositories.

The rest of this section will describe the use of the IBM Industry models with these different security approaches.

Managing Viewing privileges of the Catalog

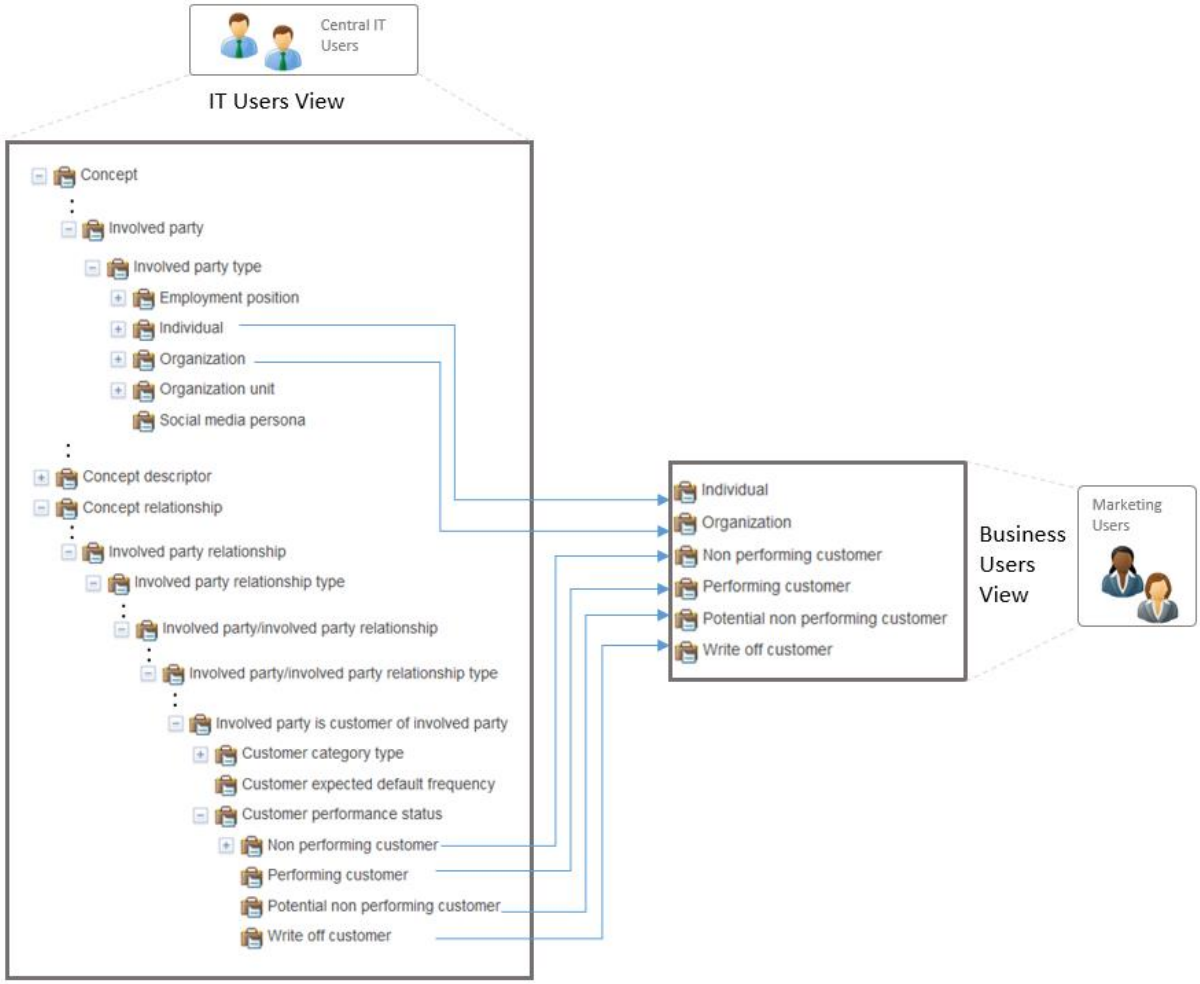
One potential objective is to set up a facility to enable different groups of the data lake to access the relevant subsets of the data lake catalog. For reasons of security and clarity there may often be a need for a hierarchy of viewing privileges where users can see terms pertaining only to their area supplemented with terms which are of general use to Users. There is also often a need to hide from Business Users more "technical" terms needed for the management and governance of the data lake, but which provide no obvious value to general business users.



The above diagram describes a potential example of how the IGC Catalog structures in the Published Catalog can be used to manage the different viewing privileges that the different data lake users might have.

Each of the different sets of Department users (for example Finance Users) can only see the terms that belong directly to their Department (for example, Finance Specific Terms) and the set of Core Business Terms that are visible to all Business Users

The Central IT Users of the data lake Ops team however have a full view of all of the details of the catalog, any central taxonomies used for classification as well as a view over the Business terms



The Diagram above shows a more detailed example:

The Central IT users have a full view of enterprise taxonomy, which could be an extensive deep hierarchy of terms. However, the business users in this example only see a small subset of the terms from that hierarchy, only the terms that make business sense to them and their context.

The main limitation of this approach is that it is concerned with just the viewing privileges on the data lake catalog (essentially the data lake metadata) and does not concern itself with the instance data in the data lake Repositories. There is a need for a separate security mechanism to be put in place for the data lake repository instance data.

Using the catalog to underpin the privileges of the data lake repositories

A more advanced approach to overcome the main limitation of the Catalog viewing privileges is to use the catalog metadata to underpin the actual access privileges to the data lake repository instance data. This means that the data lake catalog metadata and the data lake repository instance data are more tightly integrated from a security perspective.

6.6.6 Models and the Virtualization Layer across the data lake.

The current assumption is that there are no specific additional considerations in terms of the data lake Virtualization layer in the context of deploying Models related artifacts. One possible link might be the use of a subset of the Business Terms in the data lake catalog to provide the set of standardized tags to underpin the various virtualization activities

6.7 Lifecycle outputs

The main outputs of this Lifecycle would be the set of feedback to the first two lifecycles concerning the additions or changes to the Data Model artifacts needed to be included in future Model-related enhancements of the data lake Catalog or the data lake Repository structures.

IBM Corporation
IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
June 2016

IMW14878USEN-00

© Copyright IBM Corporation 2016