OCTOPAI

The Big Question in BI:
Do I Need a
Data Dictionary,
Business Glossary or
Catalog?

White Paper

- In recent years data volumes have increased greatly in enterprises, in parallel with business-driven demand to extract value from data.
- This has impacted Business Intelligence (BI) teams, who must meet this demand while dealing with increased data volumes.
- BI teams struggle because they need to find and understand the data required for BI delivery. In the past, manual effort was undertaken to do this analysis.
- Today, manual effort cannot scale and an automated approach to BI Intelligence is needed. BI Intelligence is what supplies BI developers with the knowledge they need about the data to successfully meet BI delivery expectations.
- BI Intelligence is achieved via several capabilities, all of which must be automated. These are:
  - **Business Glossary:** the inventory of business terms, their meanings, and all facts of business significance about them.
  - **BI Catalog:** a subcomponent of the Business Glossary which is the inventory of reports, their content described in business terms, and the capture of the database columns from which the report content is taken.
  - **Data Dictionary:** the inventory of databases, tables, and columns and their physical characteristics from across the enterprise production data landscape.
  - **Data Discovery:** search capabilities in the environment provided by the Data Dictionary integrated with the Business Glossary.
  - **Data Lineage:** the most complex and difficult of all metadata to harvest, Data Lineage shows how data flows via data supply chains across the production data landscape including all transformations. Visualization is critical for Data Lineage, and it must be presented as a comprehensible map.
- **All of these BI Intelligence capabilities must be present in a single platform, must be automated, and must be integrated. Anything less will fail to provide BI developers with the BI Intelligence they require, and will severely impact BI delivery.**

Introduction

By now nearly everyone has heard that "data is the new oil" – the most valuable resource of the 21st Century. This is a good analogy in more ways than one. Raw data is like crude petroleum that comes out of the ground. It needs to be refined to produce something useful. Just like refined gasoline powers automobiles, so information powers business decisions and insights. How is information refined from the raw data? That is the job of Business Intelligence (BI).

Just as it takes effort and technology to extract gasoline from crude petroleum, so it takes effort and technology for BI to extract information from raw data. The major part of this challenge is not the data visualization, which is done in a reporting tool in the final step of what is often a complex end-to-end process. That is like the service station that delivers gasoline to the car. The biggest part of the challenge for BI comes before this final step, in discovering and understanding the data needed for reporting requirements, and figuring out the details of the end-to-end data flow through production environments. This is BI Intelligence - the sum of all the knowledge about the data needed for successful BI delivery.

BI managers are increasingly recognizing that they are impacted by a lack of understanding of this overall environment, and cannot easily contribute to its management.

Drivers for BI Intelligence

Traditionally, BI has been thought of in terms of the final step of data visualization, or, worse yet, as just a reporting tool. Yet this is very far from what is needed to do BI. The end-to-end understanding that is required includes ETL, analysis services, data stores (such as data warehouses), reporting tools, and more. All of these components involve different technologies from different vendors, and in most enterprises they will be implemented in varying levels of maturity.

For many years BI met these challenges simply by manual effort. As enterprises realized that they could unlock much greater value from their data resource,

the demands on BI increased sharply, and today achieving BI Intelligence using manual effort alone is no longer a viable option. At the same time, the volume of data that enterprises manage has grown beyond anything imagined even just a few years ago. Much more data is being produced by enterprises, and data is increasingly being brought in from the outside. Enterprises expect value to be extracted from these enormously increased data volumes, further increasing the pressure on BI.
All this has happened over a period of decades, accelerating in recent years, as shown in Figure 1.
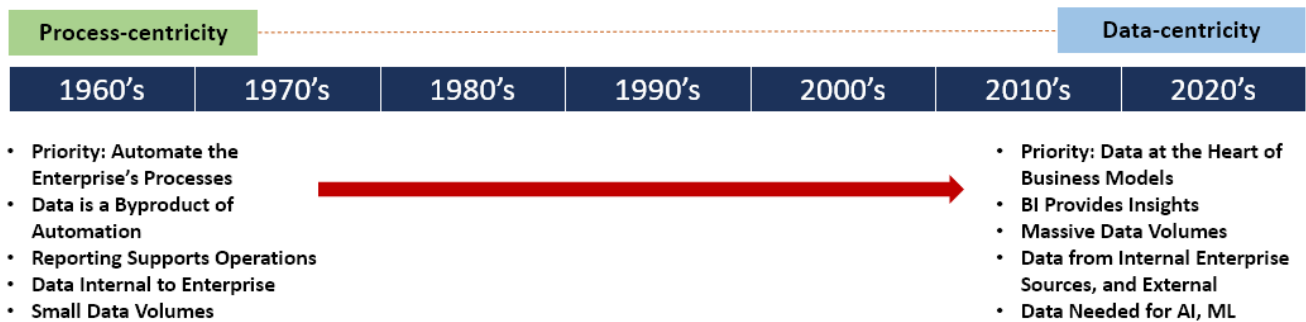

Process-centricity						Data-centricity
1960's	1970's	1980's	1990's	2000's	2010's	2020's

- Priority: Automate the Enterprise's Processes
- Data is a Byproduct of Automation
- Reporting Supports Operations
- Data Internal to Enterprise
- Small Data Volumes

- Priority: Data at the Heart of Business Models
- BI Provides Insights
- Massive Data Volumes
- Data from Internal Enterprise Sources, and External
- Data Needed for AI, ML

Figure 1: *A Brief History of IT, and The Shift to Data-centricity*

The upshot is that there is more business-driven demand for BI, which has increased the demand for BI Intelligence - to discover data, understand it, and find out how it flows. As a result, BI managers are increasingly recognizing that they are impacted by a lack of BI Intelligence, which degrades their ability to respond to the increased demand, and also to contribute meaningfully to the management of the environments that BI relies on. Manual approaches cannot scale. BI Intelligence needs automation to meet the challenge.

Understanding BI Intelligence

A major obstacle that has to be overcome before this challenge can be met is the attitudes towards BI Intelligence. Since BI developers traditionally used manual effort for gaining BI Intelligence, IT managers and business users are often unaware of what BI Intelligence is and why it is needed.

When the need for BI Intelligence is pointed out, the reaction is often that it is "just analysis" and "something we have always done". BI Intelligence is not "just analysis". Analysis can be done in lots of different ways for lots of different reasons. For instance, documenting a manual workflow prior to automating it requires analysis, but this kind of analysis has nothing to do with BI Intelligence. For BI Intelligence we need to:

- Thoroughly understand the specific items of information being asked for in the report requirements
- Determine if the reporting requirements have already been met in an existing report
- Discover the data needed for the reporting requirements
- Confirm that the data that has been discovered truly supports the information being asked for in the report requirements
- Find out where the data is coming from to understand the scope of what the data covers, and how fresh the data is.
- Determine the best location to obtain the data from

All this was done manually by BI developers in the past - every time for every report - and as we have pointed out, these manual efforts cannot scale. Where manual efforts have continued to be tried, the result has been that speed and quality of BI delivery suffer. IT managers and business users become frustrated with BI teams, even when these teams are performing as well as they possibly can.

There is an even more important point here. Old school analysis in BI was simply conceived of as BI developers going out and capturing the knowledge needed for their work. The knowledge does not pre-exist – it has to be gathered from scratch for each development exercise. BI Intelligence, however, has the central idea that the hard work of capturing knowledge has already been undertaken through automation, and a central platform is available for BI developers to access any time they need it, as shown in Figure 2.
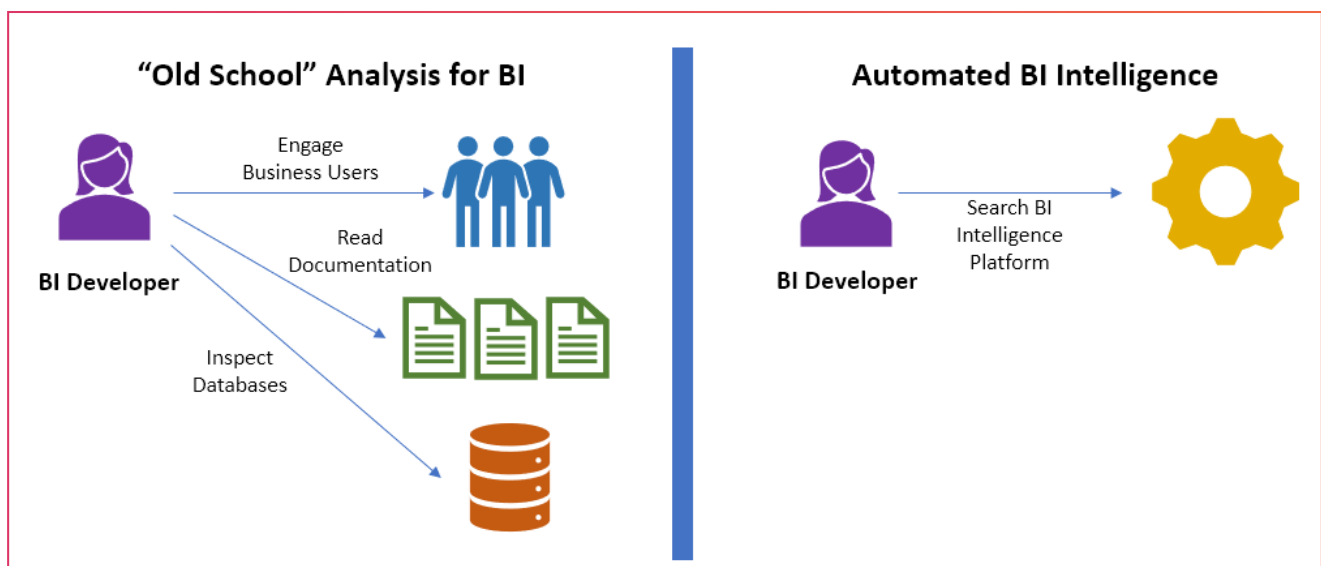

Figure 2: *Old School BI Analysis vs. Automated BI Intelligence*

What Does Automation of BI Intelligence Look Like?

All stakeholders in the information delivery process must come to understand what BI Intelligence is, what it means for BI, and why and how it must be automated. The best way to understand this is in concrete terms based on what automation of BI Intelligence has to offer.

The technologies that automate BI Intelligence have a range of capabilities. These capabilities are generally agreed to be:

- Business Glossary
- BI Catalog
- Data Dictionary
- Data Discovery
- Data Lineage

Each of these capabilities does something different, and in the past they have often been implemented by different technologies. Today, however, enterprises are increasingly demanding that single tools incorporate all of them. There is a single "whole" story behind an enterprise's data, and these capabilities collectively are what tells the whole story.

In theory, each of these capabilities could be provisioned by information supplied by humans. That would be an advance on Old School BI Analysis, where information was collected from scratch for each development effort, and never maintained afterwards. However, the scale and complexity of today's production data landscapes is so vast that this is simply not an option. So we need to consider not only what these capabilities must offer, but also how they use automation. As can easily be imagined, these capabilities are useful to roles other than BI developers, and Figure 3 shows some of them.

Capability	BI Developer	Business User	DBA	Business Analyst	Data Governance Analyst
Business Glossary	●	●		●	●
BI Catalog	●	●		●	●
Data Dictionary	●		●		●
Data Discovery	●				
Data Lineage	●		●	●	●

Figure 3: *BI Intelligence Capabilities and Usage by Selected Roles*

These roles extend to stakeholders who govern, manage, or use data. This is not really surprising given the central role of data as the "new oil" in our economy, but it does highlight the distinctions between the BI Intelligence capabilities and different constituencies in the enterprise that benefit from these capabilities. Unfortunately, these different constituencies may advocate independently for capabilities that interest them. This can be problematical as technologies and methodologies may be chosen which do not integrate to provide the solutions needed for BI Intelligence. A better approach is to drive everything from the end-to-end vision that BI Intelligence has, and ensure the requirements of the other constituencies are addressed.

That said, let us now look at these different capabilities in detail.

Business Glossary

When data becomes information it is expressed in a business context, meaning it is described using business terms. A Business Glossary is a tool that tracks these business terms, their definitions, and other useful information about them.

Some people in IT and the business think that a Business Glossary is of relatively little help. Surely the business users are already familiar with the business terms they use and their meanings. While this is true in part, it is far from the whole picture.

BI developers use a Business Glossary to check their work and understand the concepts they are dealing with. Requirements often need to be captured, rather than gathered, meaning that requirements provided by the business may not be fully precise and have to be tightened up by the BI developers themselves. Rather than pepper the business users with questions, starting only from the received "requirements", the BI developers can consult the Business Glossary to get exact definitions of what the information requested seems to be, and confirm these with the business. This greatly streamlines the BI development process and means developers do not have to rely on users as the sole source of all their information.

A more mundane use case involves report labeling. BI developers need to know how to label the data that appears in the reports and dashboards they develop. This is a problem of terminology – using the right words. With a Business Glossary, a BI developer can find the exact right term to use. This in turn prevents confusion on the part of those reading the report, who may be different to the individuals who asked for the report to be created. BI developers do not want to be in the position of having to guess what a label should be, and also want to confirm that any labels they have been asked to use by the business are standard. The Business Glossary helps with this.

Similarly, users who read a report may need to understand what a label means on the report. The Business Glossary provides the definition for each label. Why would users need this if they are already supposed to be familiar with the terms they use? The reality is that users who are new to their job, or who are dealing with reports they are less familiar with, or who spot something in a report that seems odd to them, want to check their understanding of what they are looking at. They can easily do this with a Business Glossary.

Very often people think of a Business Glossary like a dictionary – there is a term and a short definition, and it all seems pretty simple. But Business Glossaries are much more than this. They can contain a definition that is as long as it needs to be. Many different stakeholders can contribute to the definition, making it more and more precise over time. Calculations are also held in Business Glossaries too. For certain business terms, particularly metrics, the methodologies used to create them are also stored. Beyond this, Business Glossaries allow an enterprise to store all facts of business significance for a business term. This could include changes resulting from new business practices, or from changes in technology.

An increasingly important aspect of Business Glossaries is their social engagement. Business Glossaries can show who is a Subject Matter Expert (SME) for data elements. They can also permit conversations and alerts about data. Such conversations add to the store of knowledge about the data.

These are just some of the advantages of a Business Glossary. But how is any of this automated? The answer lies in another capability – **the BI Catalog**.

For years, enterprises have struggled to inventory the reports they have, what they contain, and what data is used in them. A central location where all this information is stored is a BI Catalog, but the vast majority of enterprises have never been able to implement one because, once again, no alternative to a massive, expensive, unreliable, top-down manual exercise existed.

Without a BI Catalog, enterprises have no reliable way to prevent the same report, or close variants of it, from being implemented over and over again. Nor is there any place where the information content of reports can be inspected - to find how data is labeled and grouped within specific business contexts. A BI Catalog provides for all of these needs.

A BI Catalog closely related to a Business Glossary. The reason for this is that the labels that appear in reports are actually business terms, which need to be defined, be deduplicated into synonyms, and have acronyms and abbreviations related to full business terms.

Automation is a critical need for a BI Catalog, and tools are available today that can literally inventory all the reports for an entire enterprise in just a few hours or less. They show what information appears in the reports, using report labels. Such tools show captured data transformations, particularly calculations, that occur in the report layer. Another feature of the tools that automate the BI Catalog is the ability to connect the data point on the report, described by a label, with a column in a physical database. This is an achievement of incredible importance, not just for BI Intelligence, but for Data Governance as we shall see a bit later on in this paper. Figure 4 summarizes how BI Catalog functionality contributes to a Business Glossary.
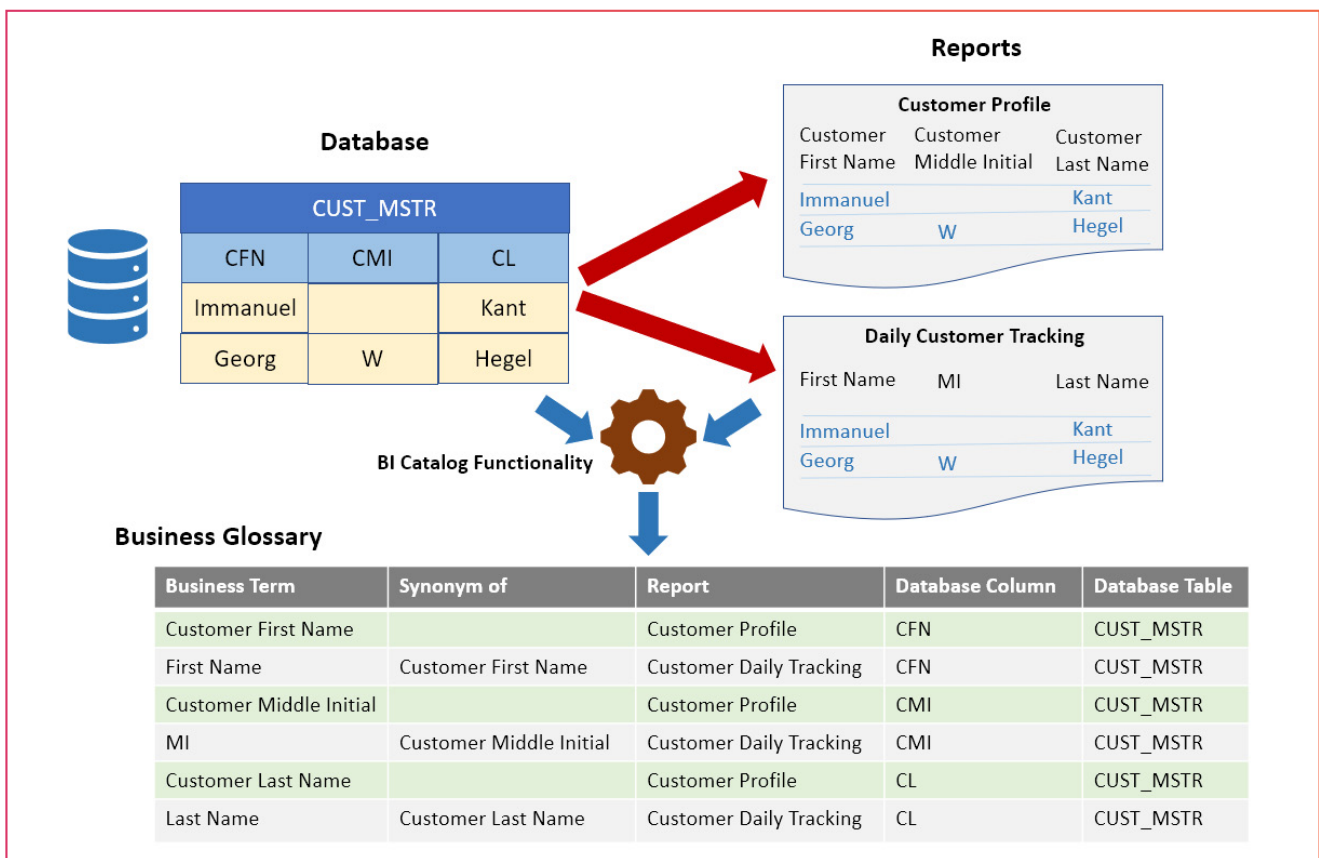

Figure 4: *Business Glossary Built Out Using Automated BI Catalog Functionality*

Business Term	Synonym of	Report	Database Column	Database Table
Customer First Name		Customer Profile	CFN	CUST_MSTR
First Name	Customer First Name	Customer Daily Tracking	CFN	CUST_MSTR
Customer Middle Initial		Customer Profile	CMI	CUST_MSTR
MI	Customer Middle Initial	Customer Daily Tracking	CMI	CUST_MSTR
Customer Last Name		Customer Profile	CL	CUST_MSTR
Last Name	Customer Last Name	Customer Daily Tracking	CL	CUST_MSTR

Given that a BI Catalog can automate harvesting of all the information about an enterprise's reports in a very short time, it is the best option to create a Business Glossary. In other words, a BI Catalog gives us an automated Business Glossary. All the labels in all the reports are Business Terms and are used to populate the Business Glossary. This would take an enormous amount of time and resources using manual effort.
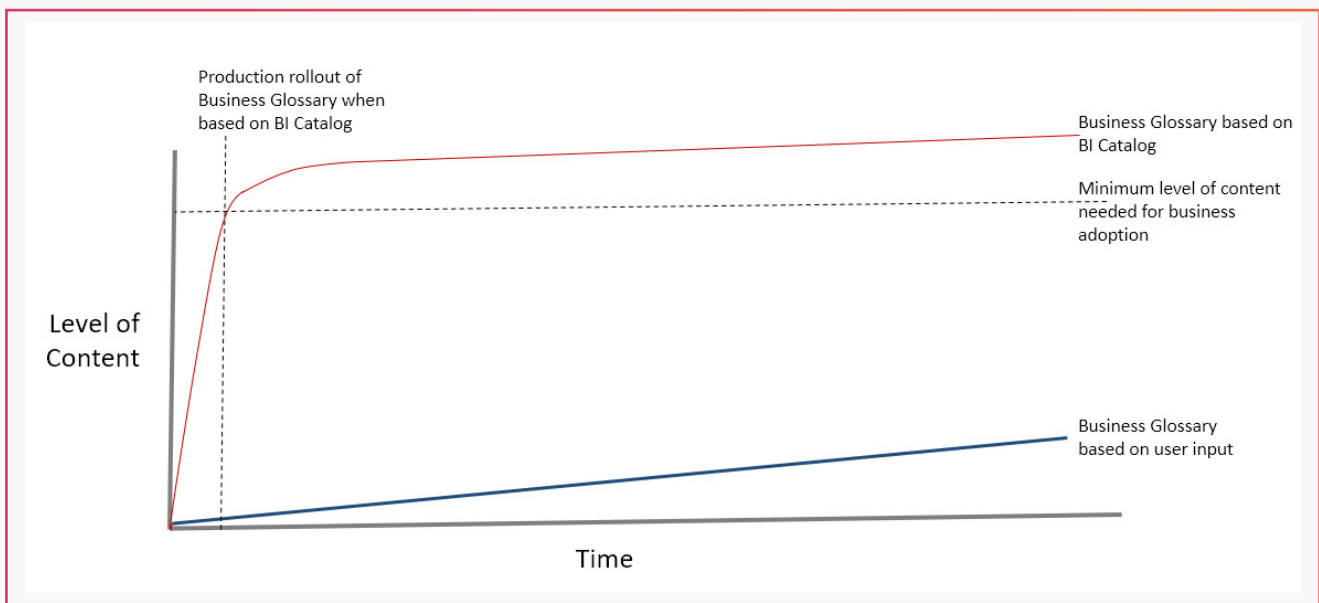

Figure 5: *Content Requirements for Deployment of a Business Glossary*

Figure 5 illustrates a common problem with Business Glossaries that a BI Catalog solves. Typically, a standalone Business Glossary (lacking a BI Catalog) is deployed with no content. Business users are then instructed to create the content. But at this point there is no content for the business users to get value from, so they are reluctant to take the considerable amount of time needed to add any. As a result, adoption usually fails. Going live with an empty Business Glossary is not a good idea.

By contrast, a Business Glossary based on an automated BI Catalog can be put into production immediately. A huge amount of useful content is created by the BI Catalog. Just knowing what reports exist and what business concepts appear in these reports is incredibly valuable for BI developers and business users alike. Adoption is therefore much more likely.

There is one other aspect of the automated BI Catalog that must not go unnoticed. It is essentially functioning in a cross-platform manner. There are different report tools consuming data from different database platforms. Having cross-platform capabilities is essential for an automated BI Catalog. Without such capabilities, or with too few of them, the only alternative is population through manual effort.

A Data Dictionary is an inventory of the physical database schemas, tables, and columns that exist in the relational databases of an enterprise. It can sometimes be extended to files and non-relational databases, but we will consider relational databases here as they are by far the most common content of Data Dictionaries.

Data Dictionaries work by reading the system catalog of a database and pulling in the specification of the objects in the database. For a column this might include:

- Column Name
- Column Datatype
- Column Null Rule
- Whether the Column is a Primary Key
- Whether the Column is a Foreign Key
- User-entered descriptive information present in the System Catalog

Since it is relatively easy in technical terms to harvest this information from relational databases, Data Dictionaries have been around for quite a while.

The information in a Data Dictionary helps BI developers, and also business users who are able to undertake self-service analytics. It acts as a quick guide to see what tables exist and what columns they contain.

Again, cross-platform capabilities are essential for a Data Dictionary. Its automated functionality must reach across the different database platforms found in the enterprise.

The Chasm between Data Dictionaries and Business Glossaries

There is one huge problem with Data Dictionaries. In their most basic form, they only describe the physical structure of a database, and that is usually not sufficient for a BI developer to understand what a particular column contains. How can a BI developer know if a given column contains the exact data they need for a new report?

This is a great problem that has plagued Data Dictionaries and Business Glossaries from the start. A Business Glossary contains business terms, but which database columns correspond to these business terms? A Data Dictionary contains database columns, but which business terms correspond to these database columns?
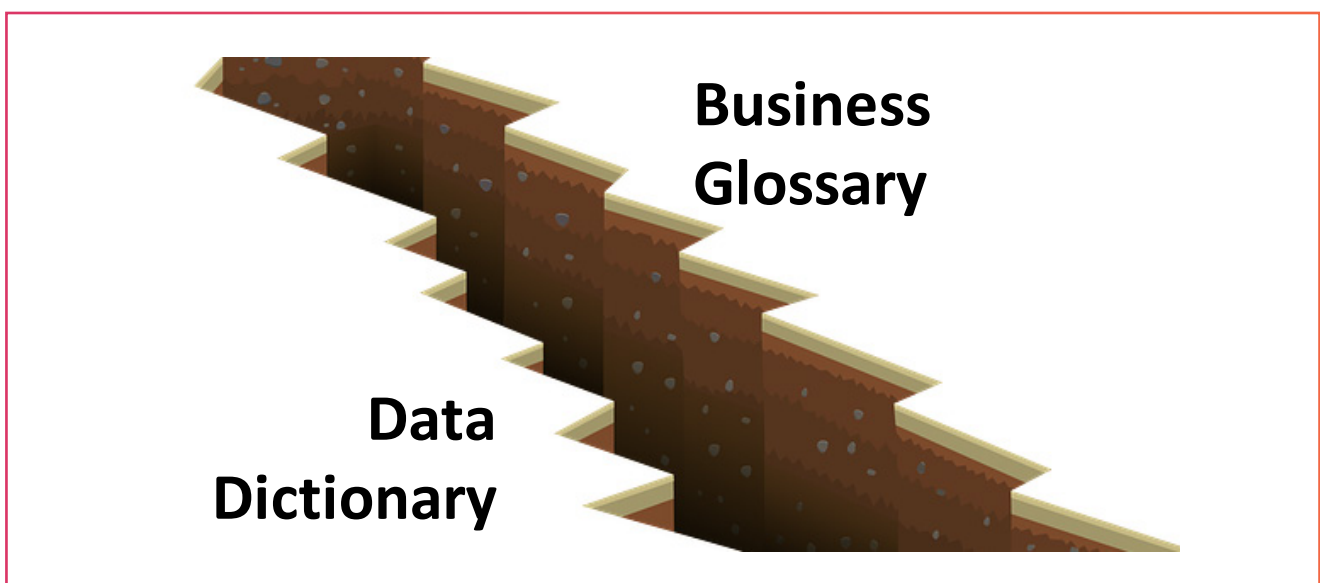

Figure 6: *The Chasm between Data Dictionaries and Business Glossaries*

How can this gap between the Data Dictionary and the Business Glossary be closed? One way is, of course, pure manual effort, which some enterprises have tried. This is enormously expensive and takes a long time. The results cannot be guaranteed to be free from error. It is usually done by looking at the data values in the physical columns. Sometimes programming or ETL logic may be inspected, but this requires even more technical resources, who are usually tied up on more important projects. As a result, manual inspection to try to figure out what a column means in business terms can also be unreliable. In reality, the expense relative to useful results produced means that enterprises simply do not undertake these exercises.

Another approach that is not uncommon is data profiling. "Data profiling" can mean a few different things, but in the sense we are using it here it means automatically examining the content of a column and classifying that content. But this too has its limitations. For instance, a tool may see that every value in a column has an "@" sign followed by an Internet domain, and so it can be inferred it is an email address. But what kind of email address is it – Customer Email, Employee Email, Vendor Email, or something else? We still cannot get to the business context of what the column represents. The only way forward from this point is manual effort once again, and, as we have seen, in today's world that will not work.

This brings us back to the BI Catalog as part of the Business Glossary. The BI Catalog bridges the chasm between Data Dictionary and Business Glossary by linking a report label to a physical column from which the data described by that label comes. As we saw earlier, and as illustrated in Figure 4, the BI Catalog does this in an automated and exact fashion. Knowing that a specific column is used in a specific report and has been given a specific label (business term) tells us all the basics of what we need to know about the column.

Once again, we need to consider the cross-platform capabilities involved here. Data Dictionaries and BI Catalogs gather information from a variety of different platforms in the data visualization layer and the database layer and render them into an integrated, business-understandable set of information. This is a remarkable achievement.

Yet there is still more that we need if we are to automate and support BI Intelligence at a level needed by BI developers and BI self-service business users.

Data Discovery

Data discovery is something that is also carried out within a Data Dictionary environment. By virtue of creating an inventory of all schemas, tables, and columns, the Data Dictionary provides a central list of all data assets from which a BI developer might draw data for a particular report. Advanced search capabilities in the Data Dictionary help the BI developer find columns that are candidates for usage in a report – when coupled with the business understanding provided by the Business Glossary.

The amount of time taken to search for data needed for BI and other kinds of analytics is known to be very great in relation to the time taken to develop a report once the data has been discovered. With a Data Dictionary that incorporates automated Data Discovery this is greatly sped up and the whole process becomes more efficient.

Negative results are also important when it comes to Data Discovery. Suppose a new report has to incorporate weather data, and it is unclear if the enterprise has any weather data or not. A manual search would take a very long time to confirm that the enterprise has no weather data. By contrast, the search could be performed in a few minutes in a Data Dictionary. However, it must be understood that these advantages only come with a Data Dictionary that is linked to a Business Glossary with a BI Catalog. To pick up on our example, it would seem to be a simple matter of searching for "weather" in the column names captured by a Data Dictionary to get meaningful results. But no table or column name might exist that includes the text "Weather". However, "Weather" would be likely to appear on a report if such data exists. Thus, a Data

Dictionary linked to a Business Glossary enabled by a BI Catalog functionality makes Data Discovery much more valuable.

Data Discovery without the link to the Business Glossary can create a very frustrating situation. Even medium-sized enterprises have millions of columns in their data landscapes, and when these are brought into a Data Dictionary they create a huge volume of metadata. Simply deploying the tool and presenting the content to BI developers, or anyone else working with data, and telling them "here's all the information about our physical data" is a poor strategy. How do you find what you need in a sea of many millions of columns, when there is nothing more than the technical specification available for each column? It is the link to the Business Glossary that makes all the difference, and makes Data Discovery possible by searching on business terms.

However, there is one more supremely important facet of knowledge about data that is needed to complete the picture for BI Intelligence, and this is Data Lineage.

Automated Data Lineage

If you ask a data professional to name the area of BI Intelligence that is the most valuable but also the most difficult to capture, they will almost always say it is Data Lineage. "Data Lineage" can be defined as understanding how data flows across a data landscape, forming a data supply chain where it is eventually consumed by reports in the BI layer. However, this is not very precise, and there are several different levels of Data Lineage as shown in Figure 7.

1	Database A Feeds Data to Database B (What Data Is Not Specified)	**High Level**
2	Database A Feeds a Type of Data to Database B (e.g. Customer Data – No Further Specification)	
3	Database A Feeds a Named Dataset to Database B (e.g. An Extract File whose specification is known)	
4	Specific Columns in Database A are Fed into Specific Columns in Database B (The Columns in A are "Mapped" to the Columns in B	**Low Level**

Note: This list may be decomposed into even more granular levels of Data Lineage

Figure 7: *Levels of Data Lineage*

This is important because some tools claim that they can automate "Data Lineage" but it is not clear what level of Data Lineage they are claiming to automate. Knowing about high level Data Lineage has some uses, but is of little value to BI developers.

A further problem that exists with tools that claim to automate Data Lineage is that some only automate the visualization of Data Lineage, but do not have the capabilities to harvest the Data Lineage. Other tools both harvest the Data Lineage and visualize it. This is an important distinction for two reasons:

1. Tools that only visualize Data Lineage must be fed with the metadata about Data Lineage. How that is collected is not specified, and often implies manual effort, which is not possible.
2. Visualization of Data Lineage is very important. Business Glossaries, BI Catalogs, and Data Dictionaries are essentially text-based. By contrast, Data Lineage needs to build a map of data flows. This requires a diagrammatic presentation, but one in which drill-down and drill-across capabilities exist, and in which textual information can also be presented. This is why the ability to harvest Data Lineage must be coupled with a very well-designed visual interface.

So we have two capabilities that we must consider within Data Lineage: harvesting the metadata about Data Lineage; and visualizing the metadata about Data Lineage.

Why Does BI Intelligence Need Data Lineage?

Given that Data Lineage is difficult to obtain and present, why is it so necessary for BI Intelligence?

One reason is that a Business Glossary can give us the business definition that applies to a column in a database, but it cannot tell us what populations of data are present in that column. Figure 8 illustrates this requirement. In this example a company operates in the USA and Canada and a BI developer is asked to produce a report covering all employees in both countries.

With Data Lineage, the BI developer gets the picture we see in Figure 8 and knows that there is a Global Combined Employee Database fed by data from the USA and Canada. This is the logical place to get the data for the new report. Without Data Lineage, the BI developer might only discover the Canadian Employee database and US Employee database and try to extract data from both for the report. A "Data Universe" is the set of populations of data in a database. A Business Glossary will provide a useable definition of Employee, but it does not tell us what the Data Universe is

in a given database. A Data Dictionary will tell us all the tables and columns that exist in the production data landscape, but not what Data Universes exist in them. Data Lineage helps the BI developer to see what data is going into a database, and so what its Data Universe is, which is a critical success factor for BI delivery.

The use case we have just discussed that is illustrated by Figure 8 also highlights that BI developers need to see Data Lineage at two different levels:

- Table-to-Table
- Column-to-Column

Automated data lineage capabilities must be able to capture and visualize lineage at these two levels. The BI developer can see a bigger picture at the Table-to-Table level and then "zoom in" to see the Column-to-Columns level.

Beyond this, the addition of automated Data Lineage capabilities to the Business Glossary, BI Catalog, and Data Dictionary capabilities provides the end-to-end data flow understanding that is so vital.
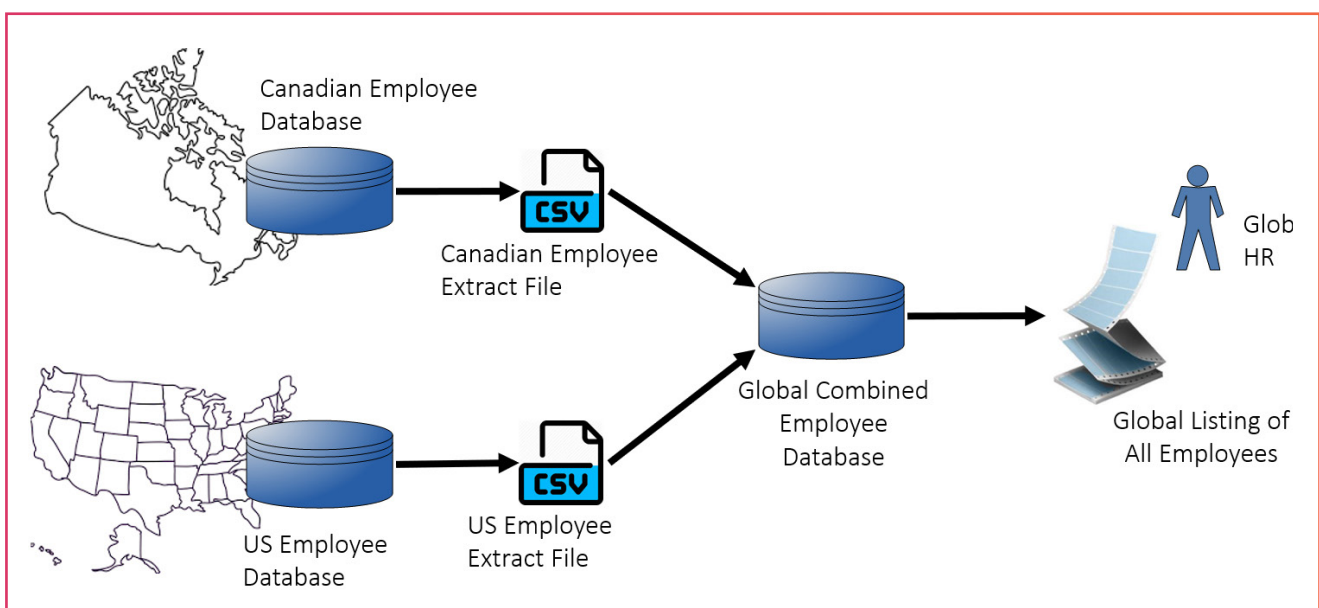

Figure 8: *Data Lineage Provides Understanding of Data Universe*

Data Lineage Removes Fear, Uncertainty, and Doubt

While Data Lineage is very important for new report development, it is absolutely essential for the support tasks BI developers are involved in. Many BI developers live in constant fear of being asked by business users to show where the data that appears in the reports they have developed comes from. Many business users are plagued by worry about the reliability of the information they see in the reports they use to make business decisions. When a business user spots something that seems to be an anomaly, they worry that this is just the tip of the iceberg and everything in the BI environment is unreliable.

These fears arise because, at least prior to BI Intelligence, neither BI developers nor business users had any idea of the end-to-end flows of data across the production data landscape. Even worse, both communities knew that they had no way of getting to an end-to-end view other than manual effort.

Hoping that the business users will not question the reports is not a strategy, and it is just a matter of time before a business user does see something they question. Then the issue becomes urgent as the BI developer is asked to

stop what they are doing and figure out an answer. The longer it takes to provide the answer, the greater the doubts grow in the mind of the business user.

This is where BI Intelligence really shines. Without it, BI developers are doomed to manual effort that will require starting with the report that is in doubt and manually inspecting database structures, ETL logic, stored procedures, and getting guidance from subject matter experts and upstream application support teams to interpret what they find. The time involved, the chance of missing something or misinterpreting something, and the factor of imposing on a lot of people who are already working full time on other things, make this a very unsatisfactory process.

By contrast, with an automated BI Intelligence suite, BI developers can literally press a button and see everything that they need right away in order to satisfy the doubts raised by the business user. The great spider web of the data lineage, with all its complexity is revealed and the path from the data point in doubt can be traced back to its source in minutes. Figure 9 shows an example of what this might look like.
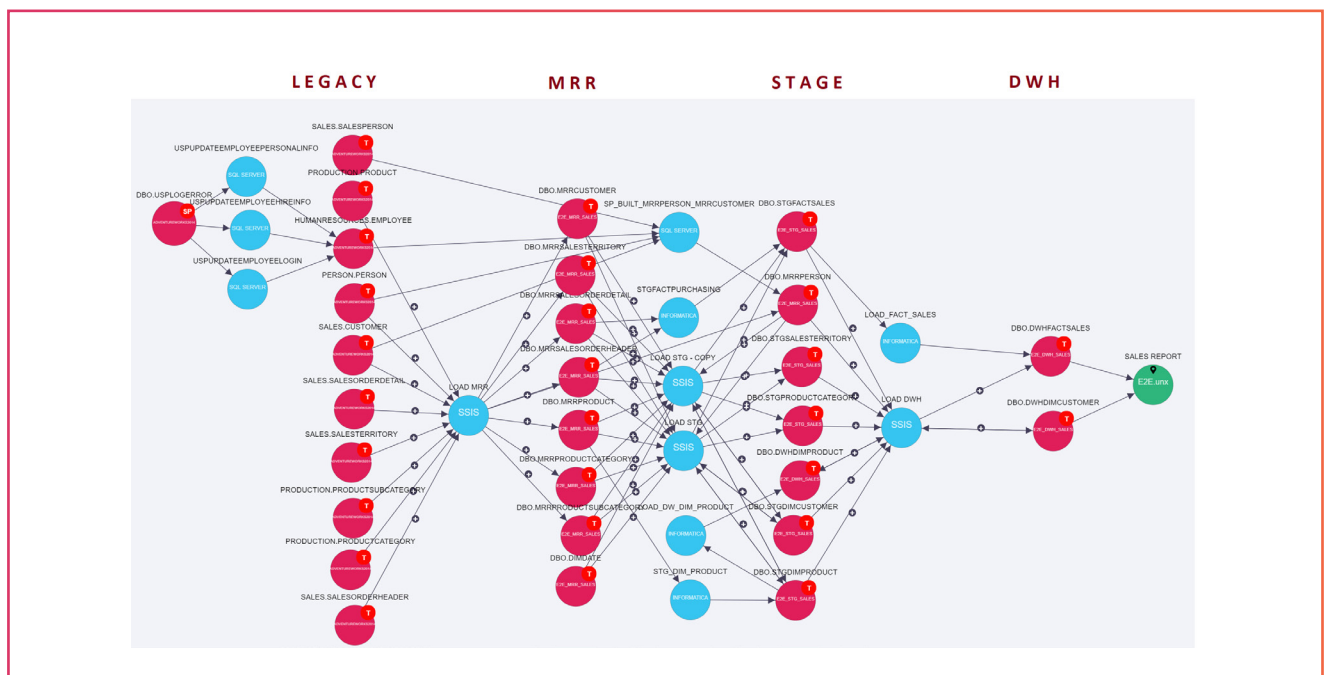

Figure 9: *Example of Data Lineage Map*

Maybe the business user simply did not understand something about the data, and there really is not a problem. Or maybe some uncommunicated change happened upstream and has affected the report. Either way, the Data Lineage Map will quickly provide the answer.

Harvesting Data Lineage Metadata

We have looked at just a couple of benefits of Data Lineage, and will consider some more shortly. But before that it is very important to understand what needs to happen for true Data Lineage to be achieved. As we noted earlier, there are a lot of vague definitions of "Data Lineage", and claims that it is provided need careful investigation.

Harvesting the metadata needed for Data Lineage is a very complex task because so many different technical components are involved. Figure 10 illustrates the complexity, even though it is simplified.
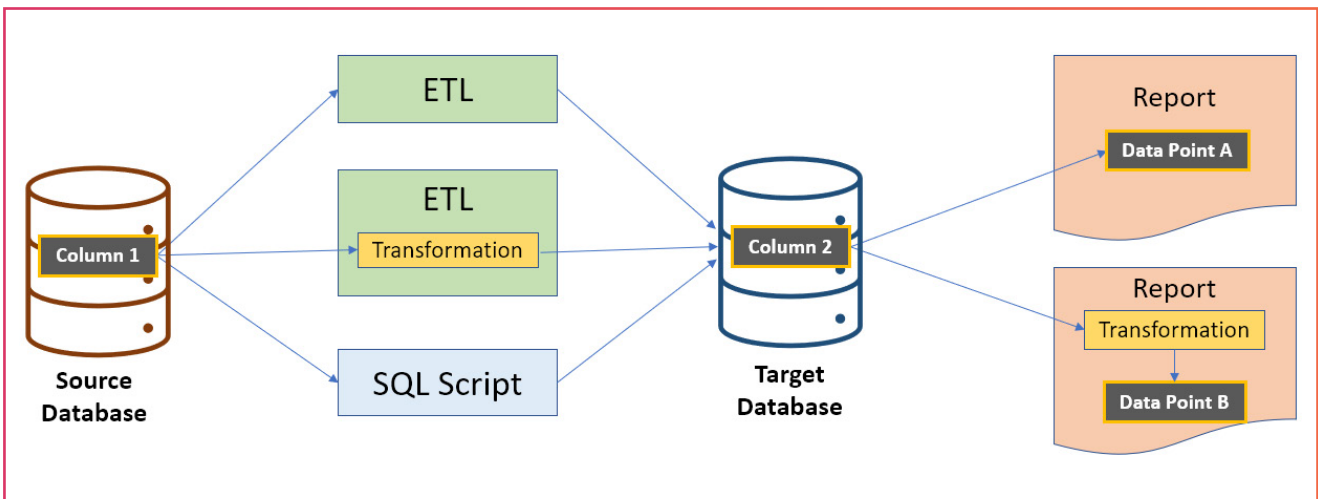

Figure 10: *Scope of Harvesting Data Lineage Metadata*

There are many different database platforms, ETL tools, scripting languages for SQL, and report packages. They exist in different layers, with sublayers, and there are different commercial products that can be found across the data supply chain. What is critically important is to harvest the data lineage end-to-end, not just in some of the layers, and to be able to extract the needed metadata from the different commercial products that are implemented in these layers. Wherever transformations occur these must be understood too.

Once again we see the need for the functionality involved to be cross-platform, and to be able to function at the table-to-table and column-to-column level in the database layer. And no end-to-end picture can be achieved without also understanding what is happening in the report layer.

It is no wonder that Data Lineage is so difficult to solve in a successful way. It is far harder than what we have seen for Business Glossary, BI Catalog, Data Dictionary, and Data Discovery. Yet it is easy to see how Data Lineage integrates with all of these capabilities. Indeed, it has to as this complete, integrated set of capabilities is what BI Intelligence requires to show the entire end-to-end picture. If one of them is missing, or any of them is incomplete, then BI Intelligence will not provide what is needed to support BI delivery and any investment will be wasted.

Now that we have presented the components of BI Intelligence in detail, let's summarize what each one gives an enterprise, and how an enterprise can know when it needs each of them.

Business Glossary

- Manages information at the business level, based on terms and definitions
- A foundational capability for BI Intelligence, since all data objects have to ultimately be understood in business contexts.
- Useful for managing knowledge about concepts that are not widely understood, helping new staff become acquainted with the data they work with, interpretation of reports, conforming terminology across reports.
- A social platform that centralizes knowledge about information, which anyone in the enterprise can contribute to and consult.
- Needed at the start of any program of BI Intelligence and, frankly, any metadata initiative. The Business Glossary is essentially table stakes.

BI Catalog

- An automated inventory of reports, the information they present, the transformations they contain, and the data they source from.
- An essential capability for BI Intelligence. The BI Catalog provides understanding of the data visualization layer and enables governance of this layer.
- A very useful part of BI Intelligence. It supports BI developers by showing what reports exist and how these are using data.
- The BI Catalog can automate the population of a Business Glossary.
- Needed as a first step in BI Intelligence, and to broaden the scope and content of a Business Glossary.

Data Dictionary

- An automated inventory of databases, schemas, tables, columns, and their technical characteristics.
- The basis for Data Discovery.
- Limited in usefulness without a Business Glossary, as even an average-sized enterprise may have millions of columns in its data landscape. The columns have to be related to business terms in the Business Glossary.
- The BI Catalog is needed to link the column level information to the business terms in the Business Glossary.
- Needed to expand the scope of BI Intelligence to include database components. Essential for governing technical aspects of data environments (which can be partially accomplished using a Data Dictionary without a Business Glossary or BI Catalog, but is limited without these capabilities).

Data Discovery

- Search capabilities that go on top of a Data Dictionary that is linked to a Business Glossary, and thus can be searched using business concepts.
- Supports BI Intelligence by helping BI developers quickly find sources of data and assess which is the best one to use.
- Very useful throughout the enterprise for other functions that need to find data, e.g. data scientists, data engineers building environments like data warehouses.
- Requires that there is a Data Dictionary linked to a Business Glossary that has been populated by a BI Catalog. It can be implemented when these other capabilities have been established.

Data Lineage

- The ultimate capability needed to support BI Intelligence and master all data environments. It provides an end-to-end view of how data flows across the data landscape, across different platforms, and any transformations along the way.
- Integrates the knowledge managed in the Business Glossary, BI Catalog, and Data Dictionary to provide a comprehensive set of information for BI Intelligence.
- Diagrammatic visualization of information is just as important as textual information in this capability, as is drill-down functionality to get to successive levels of detail.
- Automation of the harvesting of Data Lineage information from the disparate technical environments is the most important part of this capability.
- Data Lineage is useful to a wide range of stakeholders across the enterprise.
- Absolutely necessary for BI Intelligence, but requires the other components to be in place too.

From this brief survey we can see that there are different capabilities that build on each other, culminating in Data Lineage. Yet Data Lineage is what is so necessary for BI Intelligence. It is no wonder that enterprises are looking for technologies that bundle all of these capabilities. Such technologies will provide a single solution for BI Intelligence, although the people and process part of the implementation may still have to be built out in an iterative manner for each included capability. Nevertheless, this is the obvious answer to what is needed to support BI Intelligence.

Conclusion

The need for BI Intelligence has only been understood in the past few years, and then in only very general terms. The fact that BI Intelligence requires a set of distinct but integrated automated capabilities that coexist in a single platform has only been appreciated much more recently. We are now beginning to see the emergence of BI Intelligence tools that truly support BI delivery. This response cannot come fast enough. BI delivery is under enormous strain because of the growth in data volume and business demand. At the same time self-service BI tools are becoming more prevalent and are spreading the BI footprint across the entire enterprise. The insatiable demand for BI delivery must be met, and BI Intelligence which harnesses automation is an essential component of how this demand will be satisfied, and satisfied successfully.