BI4SME

boosting business intelligence skills for SME growth

BI4SME - R2 – Training materials Unit 0: Introduction

GRANT AGREEMENT 2021-1-ES01-KA220-VET-000033132

Co-funded by the European Union

Contents

UNIT 0: Introduction

0.1. Training objectives

The goal of this training unit is, in the first place, to **teach the foundations of Business Intelligence (BI)** and other related areas in the context of Information Technologies. In the second place, the goal is to teach a series of open-source, free tools and techniques with the purpose of providing knowledge on BI and analytics. This way, you will attain a high-level overview of the different concepts and work fields.

This course is **designed for non-IT professionals**; therefore, it is not expected from the students to have advanced skills like programming languages, cloud services, etc. However, if you have some knowledge of mathematics, statistics, or tools like Excel, it could be helpful to this course.

The practical laboratories included in the training courses are focused on **providing hands-on examples of the use of BI technologies and tools in the small business contexts** given that our primary students could be SME managers and employees. This way small and medium-sized enterprises could increase the opportunities of improving their business and achieve growth in the enterprise market.

0.2. Introduction to BI and disciplines

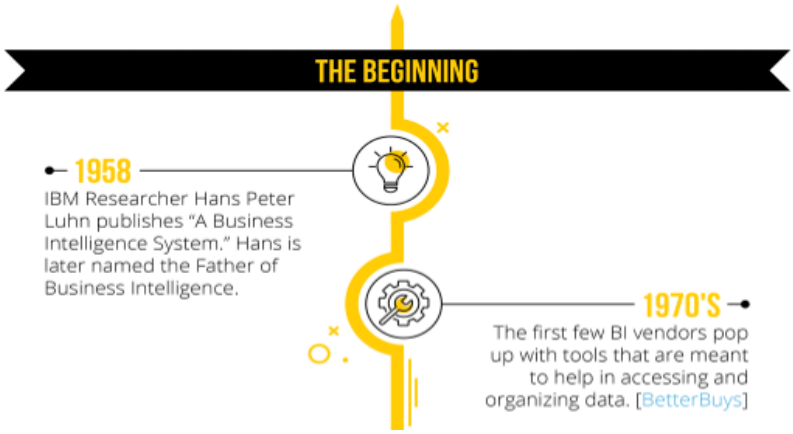
0.2.1. Basic concepts and brief history

First of all, it is important to clarify some of the concepts. The terms ***"Business Intelligence", "Business Analytics", "Big Data", and "Data Analytics"*** (often used as synonyms) are in fact different areas of study with a common goal: to enable data analysis to extract as much information as possible.

"Business Intelligence" and *"Business Analytics"* may seem similar since both share the same principle: make the best use of information to make better decisions. In the same way, the terms "Data analytics" and "Business analytics" are often used interchangeably, yet the two are quite distinct.

Despite this, they have subtle differences in terms of 4 key concepts: what data they analyse, where they are stored, what they do with the information, and what variable each one studies.

Before proceeding in going deeper into all 4 disciplines, let's have a look on the **brief history of Business Intelligence** given that our training will be mainly focused on this area.



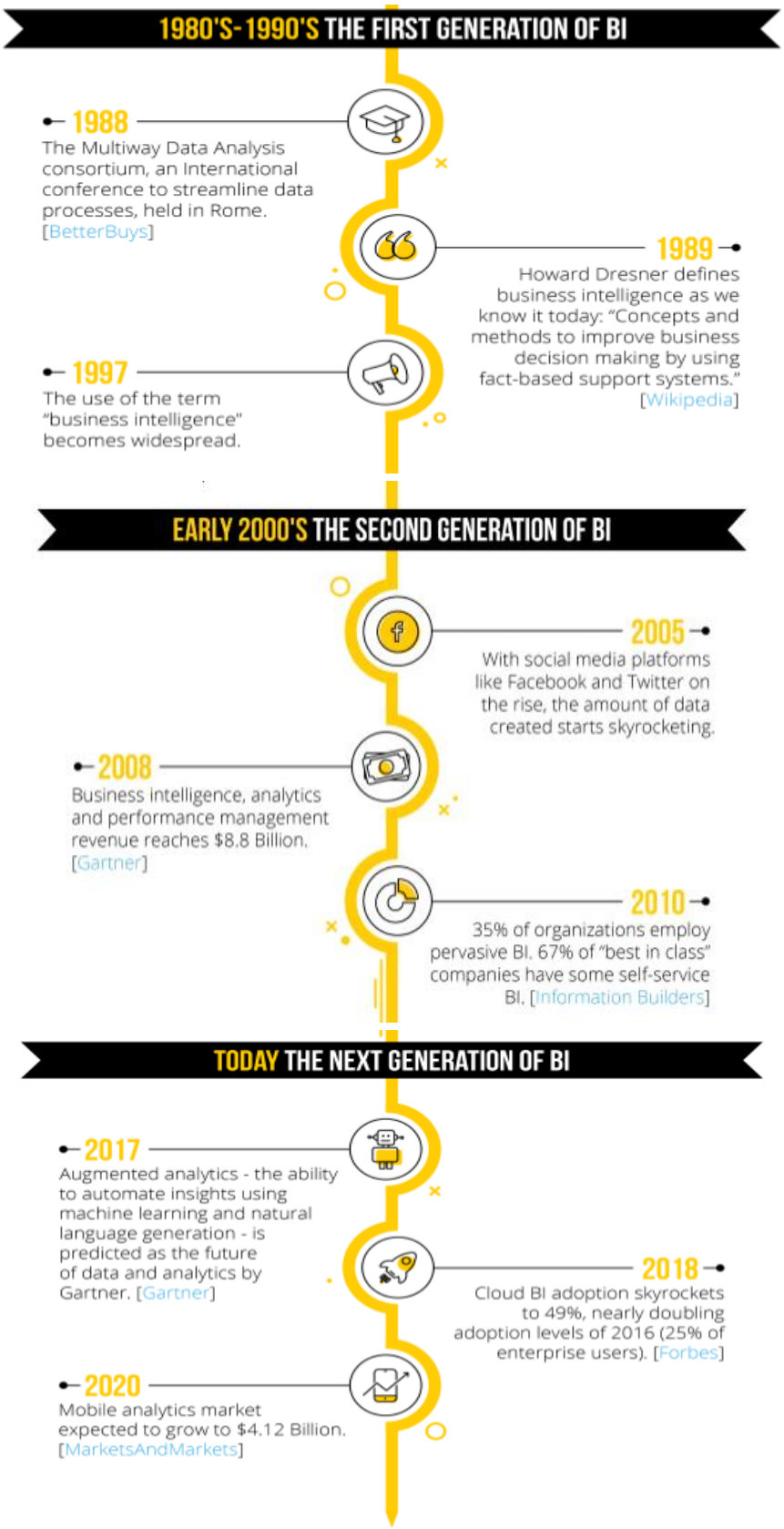


Figure 1. Resumed diagram explaining the history of BI from DZone blog.

0.2.2. Business Intelligence

Business Intelligence (BI) can be described as a set of **methodologies, processes, architectures and technologies that transform raw data into meaningful and actionable business solutions**. The useful information is used to enable strategies, tactics, operational insights and decision-making.

Business Intelligence aids in data-driven decisions, but the benefits are incomparable and extended beyond the driven business solutions. Essentially, Business Intelligence systems are data-driven Decision Support Systems (DSS). Business Intelligence is sometimes used interchangeably with briefing books, report and query tools, and executive information systems.

The main benefits are:

- Data-driven business decisions
- Faster analysis and intuitive dashboards
- Increased organizational efficiency
- Improved customer experience
- Improved employee satisfaction
- Trusted and governed data
- Increased competitive advantage

Business intelligence (BI) teams run queries on data, which are eventually presented to end users, or to individuals responsible for making business decisions, or used as input for machine learning algorithms or other Data Science projects. One common problem encountered here is that if the summaries from analytics databases cannot support the type of analysis the BI team wants to do, then the whole process needs to run again, this time with different transformations.

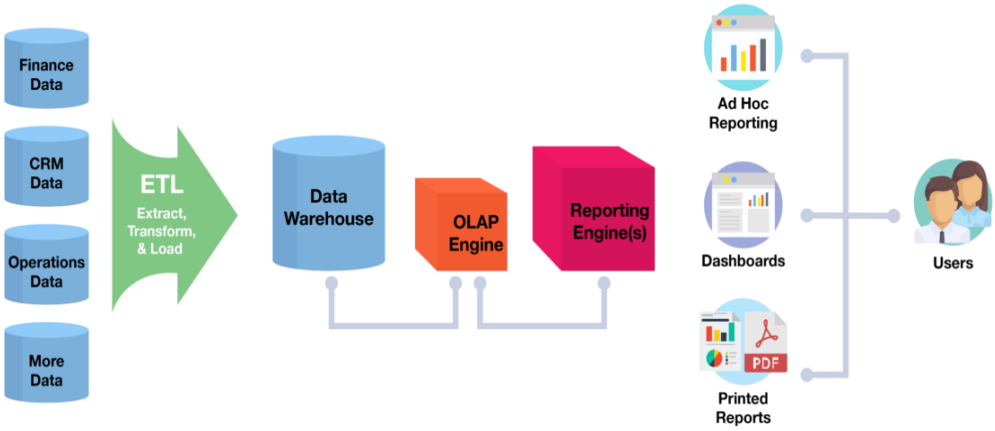


Figure 2. BI system design at a high-level

Business Intelligence is oriented to the past; through reports the history of the company is examined to understand its development. The main goal is **to allow professionals to use data with maximum productivity**.

Some of the tools we find in Business Intelligence are:

- Real-time monitoring
- Dashboard development and reporting
- Benchmarking
- Implementation BI software
- Data visualization
- Performance management
- Data and text mining

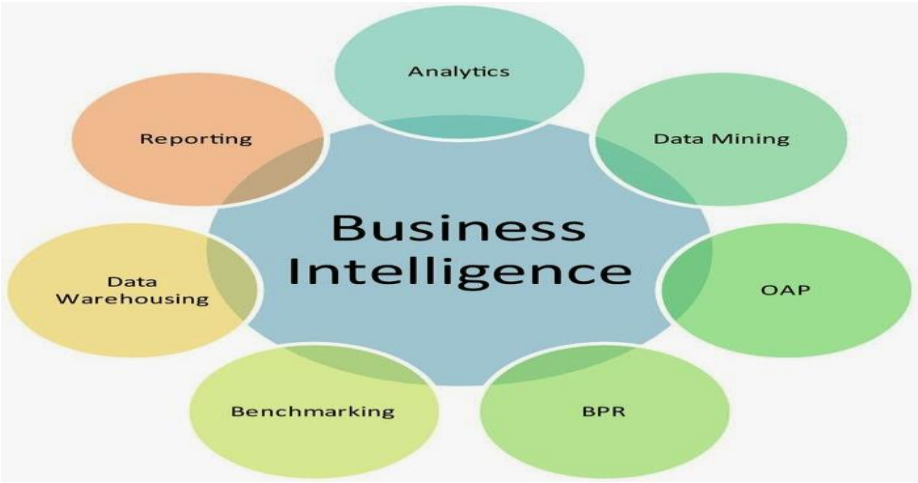


Figure 3. Diagram of BI tools

Co-funded by the European Union

GRANT AGREEMENT 2021-1-ES01-KA220-VET-000033132

0.2.3. Business Analytics

Business Analytics (BA) is an **approach to data analysis within a company** and is commonly identified as a complement to **Business Intelligence (BI), which focuses on historical and current data to understand the performance the company is experiencing** so far and aid in planning and identifying business patterns or issues.

Business Analytics focuses on the future, that is, it facilitates the creation of a future vision based on predictive models that influence the development of new paths and strategies.

While BI studies the internal statistics of our company, BA uses different external sources to show us the best paths, for example, it can use trends in our sector and macroeconomic data. Both BI and BA technologies focus on optimizing our company's processes and decision-making, the main difference is that the first corrects operational errors (current and past), and the second focuses on not committing them in the future.

As we have just seen, both BI and BA are key to understanding the facts of the past and the trends of the future. Together they will help us to improve our business strategy more wisely.

0.2.4. Big Data

Big Data is an area of knowledge that studies ways to **capture, process and manage large amount of data**. In recent years, Big Data was defined by the "3 Vs": *volume, velocity, variety*. Nowadays Big Data is defined by the "6Vs", also including *veracity, value and variability*.

0.2.4.1. Volume

The name 'Big Data' itself is related to a size of data that is enormous. Volume is a huge amount of data. To determine the value of data, the size of data plays a very crucial role. If the volume of data is very large, then it is considered as 'Big Data'. This means that a particular data can be considered as Big Data depending on the volume of data.

If we had this volume of data in a traditional system, it would be so huge that we would not be able to persist or process this amount of information with the current technologies. For instance, if we have a traditional database in a single machine, we shouldn't have enough space to persist all the information. Even if we could have it, the operations like queries for consulting some data would take hours or days, or more.

0.2.4.2. Velocity

Velocity refers to the high speed of accumulation of data. In Big Data, velocity data flows in from sources like machines, networks, social media, mobile phones etc. There is a massive and continuous flow of data. This determines the potential of data and how fast the data is generated and processed to meet the demands.

The purpose of Big Data architectures is to reach a high speed of data processing considering that we are dealing with this high volume of data. Consequently, this involves new ways of storing information that requires organizing information into several distributed files, instead of a single central server. Hence the capacity of splitting large data sets into small chunks is also necessary to process the information in parallel.

0.2.4.3. Variety

It refers to the nature of data which might be structured, semi-structured and unstructured. It also refers to heterogeneous sources. On Big Data architectures we can have data from various sources with different data types, while on Business Intelligence and Analytics, we can work with data that Big Data systems have previously structured.
Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

- **Structured data**: this data is basically organized data. It generally refers to data that has defined the length and format.
- **Semi- Structured data**: This data is basically semi-organized data. It is generally a form of data that does not conform to the formal structure of data. Log files are examples of this type of data.
- **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are examples of unstructured data which can't be stored in the form of rows and columns.

0.2.4.4. Veracity

It refers to inconsistencies and uncertainty in data, which means that data can sometimes get messy and, therefore, quality and accuracy are difficult to control. Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources. The data used for further analysis should be reliable and complete.

0.2.4.5. Value

The data itself has no value for the company unless you turn it into something useful. Data itself is of no use or importance, but it needs to be converted into something valuable to extract information.

If you have a large amount of data but this data is not relevant to be used in analytics to extract trends or statistics, this information has a low or null value. Hence, the value is a very important attribute to increase the production of useful results during the analysis.

0.2.4.6. Variability

Data variability, also known as spread or dispersion, refers to how spread out a set of data is. Variability gives users a way to describe how much data sets vary and allows users to use statistics to compare their data to other sets of data.

In the context of Big Data, variability refers to the number of inconsistencies in the data. Variability can also refer to the inconsistent speed at which big data is loaded into your database. Lastly, big data itself can be classified as a variable, because of the multitude of data dimensions that result from the multiple disparate data types and sources available.

0.2.5. Data Analytics

Analytics has become a driving force for business development and transformation, providing organizations with the capabilities needed to create and implement new, creative strategies that improve customer experiences, enhance growth opportunities, and provide new revenue streams.

When a business is planning its sales strategies for an upcoming season or holiday, it might use business analytics to predict product demand so it can optimize stock and ensure they're able to meet a specific business goal.

However, with data analytics, that same hypothetical business might use data to discover that women between the ages of 18 and 24 are the most likely to buy those products—and then personalize their marketing campaign accordingly.

Data analysis is the process of collecting, cleaning, inspecting, transforming, storing, modelling and querying data (along with various other related tasks). Their goal is to identify trends and patterns that reveal important insights and increase efficiency to support decision-making. This can be applied in business, but also in other domains, such as science, government, or education.

The classification of Data Analytics, according to its purpose, can be divided into four categories: descriptive, diagnostic, predictive and prescriptive:

- **Descriptive analysis** provides an objective description of what has happened in the past.
- **Diagnostic analysis** seeks to understand the reasons behind what has happened in the past.
- **Predictive analytics** uses past data to make predictions about future trends.
- **Prescriptive analytics** provides actionable steps to reach a specific goal.

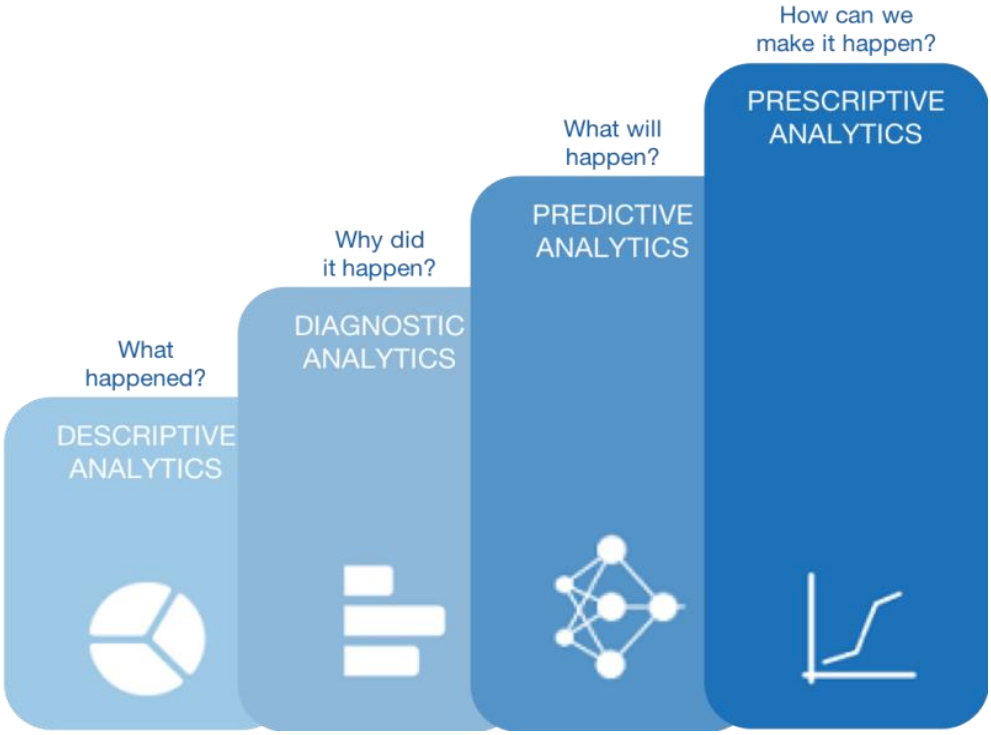


Figure 4. Data analytics categories.

0.2.6. Comparative

Let's wrap it up! In the following table, you can see the summary of the similarities and the differences of all 4 concepts: ***"Business Intelligence", "Business Analytics", "Big Data", and "Data Analytics".***

	Business Intelligence	**Business Analytics**	**Big Data**	**Data Analytics**
What data do they analyze?	Structured data	Structured data	Structured and non-structured data	Structured and non-structured data
Which is the focus of the analysis?	Oriented to the past	Oriented to the future	-	Oriented to the past and the future

Where is the information persisted?	Stores data on a central server	-	Stores data on a distributed server	-
What variable does each one study?	Studies internal statistics	Studies trends or macroeconomic indicators	-	Identifies trends and patterns not only for business
What do they do with the information?	Corrects operational errors	Works to avoid making mistakes from the past	Captures and processes information	1. Describes what happened in the past 2. Understands the reasons of it 3. Makes predictions about future trends 4. Provides actionable steps to reach a specific goal

0.3. BI and Big Data Architectures

0.3.1. Business Intelligence Architecture

It is interesting to have a notion of what is a Business Intelligence Architecture and how it works at a high level because during the course we are going to use some tools that are based on this same structure.

The Business Intelligence systems are designed with an architecture divided into several layers. This design can be adapted to the context of the business needs, the organization which it is part of, etc. One of the most extended is based on 5 layers, which are:

1. Data sources
2. Data integration
3. Data management services (or analytical stores)
4. Reporting and Analytical services (or BI tools)
5. Information delivery and consumption services

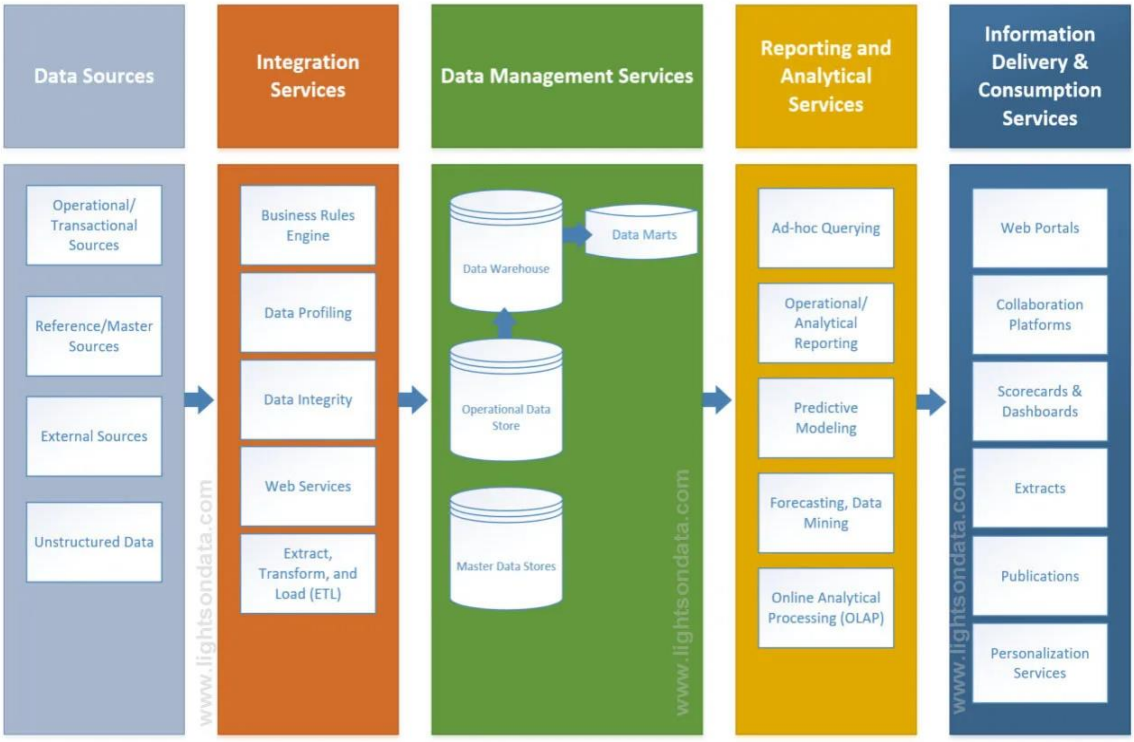


Figure 5. BI architecture model from lightsondata.com

0.3.1.1. Data sources

This layer oversees connecting the different sources from which your data enters your ecosystem. The type of sources can be:

- Operational/ transactional sources – Ex: online purchase/donation forms
- Reference/ master sources – Ex: HR information system, reference master data (such as fund types)
- External sources – Databases, APIs, Webservices, message systems, etc.
- Unstructured data – Ex: files with documentation, emails, phone conversation recordings, etc.

You need to be aware of all these sources and how data is captured from each one. This way we can evaluate the data quality level and ensure the proper transformation. Besides, we can enforce the use of business rules before it can be stored within the systems you manage.

One of the most common systems is databases. In module 2 we'll introduce SQL, a common language for most types of databases.

0.3.1.2. Data integration

This layer oversees collecting data as it evaluates, standardizes, updates, and transforms the data through:

- **Business rules & procedures** – create and optimize them and ensure the data follows them.
- **Data profiling** – a step often missed, but of great value in understanding the data better and gauging its quality level.
- **Data integrity** – ensuring data is recorded or captured as intended.
- **Data manipulation** – the actual process of importing and transforming data according to our target systems data model.

0.3.1.3. Data management (Analytical Stores)

This refers to the storage systems where the processed data is going to be persisted. We have different types of storing options:

- Data warehouse
- Data marts
- Operational data store(s)
- Master data store(s)

Data warehouse

The data warehouse is the centrepiece of the BI system built for data analysis and reporting. Data in an enterprise exists in different formats in various sources and is not necessarily consistent from one source to another. To resolve differences and potential conflicts, a data warehouse consolidates data from different sources and makes the data available in one unified, harmonized form.

The process of getting to this “one version of the truth” for an enterprise or organization is divided into three main steps: extract, load, and transform. We used to call this the ETL process. We’ll go deeper into this topic in Unit 3.

Data mart

A data mart offers the analytical capability for a restricted area of data, for example, for just one functional domain or department in an enterprise. Data marts can help avoid one department from interfering with another department’s data. They can also simplify data analytics or meet a smaller, more specific requirement.

Operational data store

An operational data store (ODS) is a type of database that's often used as an interim logical area for a data warehouse. ODSes are designed to integrate data from multiple sources for lightweight data processing activities such as operational reporting and real-time analysis.

While in the ODS, data can be scrubbed, resolved for redundancy and checked for compliance with the corresponding business rules. An ODS can be used for integrating disparate data from multiple sources so that business operations, analysis and reporting can be carried out while business operations are occurring. This is where most of the data used in current operations is housed before it's transferred to the data warehouse for longer-term storage or archiving.

Master data store

In the master data storage, we are going to persist and store the critical business data of the company shared by a lot of sub-systems. Master data generally falls into 4 groups: people, things, places, and concepts. Other categorizations within those groupings are called subject areas, domain areas, or entity types.

0.3.1.4. Reporting and Analytical services

In this layer we will use the Business Intelligence tools in order to process the information we have previously centralised in the Data warehouse. We can retrieve data or transform it into more valuable information through:

- Ad-hoc queries
- Operational/ analytical reporting
- Predictive modelling
- Forecasting & data mining
- Online analytical processing (OLAP)
- Others

We will go deeper into these tools in Units 4 and 5.

0.3.1.5. Information Delivery and consumption services

The layer oversees showing the information we have prepared and refined in the analytical service layer. With the information created, we can reach our intended audience. This is usually done through:

- Web portals – Ex: an intranet through which user access is maintained
- Collaboration platforms – Ex: a SharePoint site
- Dashboards/ scorecards – Ex: PowerBI or Tableau
- Extracts – Ex: Excel data sheets
- Publications – Ex: Executive summaries, board reports, organization or department newsletters
- Personalization services – Ex: a report automatically delivered through an email, or a dashboard dynamically tailored to its user

0.3.2. Big Data Architecture

The main difference between Big Data Architectures and traditional BI architectures is the requirements explained in section 0.3.3. A Big Data Architecture must be able to manage huge volumes of data with high speed of processing.

Therefore, in the centre of this architecture, we'll have a new type of storage called Data Lake, which is a repository that stores large amounts of incoming data. In a Data Lake, we have 2 different zones: the Raw Data zone and the Master Data zone.

In the Row Data zone, the system ingests all the information extracted from the sources with minimal validation. Here it doesn't matter if the information is incomplete, inconsistent, or has not been filtered. The important fact is to keep the original information without being transformed because this way in the future we could access this data and reprocess it again in case we need it.

In the Master Data zone, we can use the original information from the Raw zone and transform it to clean it, filter it, enforce all the validations and

Co-funded by the European Union

GRANT AGREEMENT 2021-1-ES01-KA220-VET-000033132

constraints to standardize it, etc. This way in the Master Data zone we'll have a lot of refined and valuable data, which could be provided for data scientists, feed machine learning process, etc.

A Data Lake is built on top of a distributed file system where we can split the data on different machines (or nodes). The most used technologies are HDFS (Hadoop Distributed File System) or Amazon S3 (Simple Storage Service).

Besides, we'll need technologies with the capacity to manage these huge, distributed files, to process them in parallel on different machines, and orchestrate all the processing. The most used technologies for this purpose are Hadoop Map Reduce, Apache Spark, and Apache Flink.

Big Data is a broad and complex field of work and requires certain advanced knowledge. It is not the objective of this course to delve into this topic but to show general fundamentals. Big data architectures are the layout that allows data to be optimally ingested, processed and analysed. If you are interested in learning more about this topic you can find a lot of useful information on the Internet.

0.4. References

- Sisense (2018, May 10). A Brief History of Business Intelligence [Infographic]. https://www.sisense.com/blog/infographic-brief-history-business-intelligence/?0
- Talend (n.d.). *Business Analytics vs. Data Analytics: Which is better for your Business.* https://www.talend.com/resources/business-analytics-vs-data-analytics/
- IBM (n.d.) *Business Analytics conozca sus datos, descubra sus insights.* https://www.ibm.com/mx-es/analytics/business-analytics
- Van Loon, R. (2023, Jannuary 27). *What's the difference between data analytics and Business Analytics* Simplilearn. https://www.simplilearn.com/business-analytics-vs-data-analytics-article
- Entreprise Big Data Framework (2019, April 27). *The difference between Analytics, Business Intelligence and Big data*. https://www.bigdataframework.org/analytics-business-intelligence-and-bi-whats-the-difference/
- GeeksforGeeks (n.d.). *6V's of Big Data.* https://www.geeksforgeeks.org/5-vs-of-big-data/
- Indicative n.d.) *What is Data Variability?.* https://www.indicative.com/resource/data-variability/
- Smart Panel (2023, June 1). Diferencias *entre Big Data, Business Intelligence y Business Analytics.* https://www.smartpanel.com/diferencias-entre-big-data-business-intelligence-y-business-analytics/
- Firican, G. (n.d.). *10 components of the Business Intelligence landscape.* LightsonData. https://www.lightsondata.com/business-intelligence-landscape-components/
- Sydle (n.d). *Big Data, Business Intelligence y Analytics: Cuàl es la diferencia?.*

https://www.sydle.com/es/blog/big-data-business-intelligence-analytics-612d5993f797755bfcfd57a6/

- Lutkevich, B. (n.d.). *Operational data Store. Tech Target.* https://www.techtarget.com/searchoracle/definition/operational-data-store
- Sharma, R. (n.d.). *What is Big Data Architecture? Definition, Layers, Process & Best Practices.* Upgrad. What is Big Data Architecture? Definition, Layers, Process & Best Practices | upGrad blog
- Hajdarbegovic, M. (2020, January 28). *Data Lake Architecture: A Comprehensive Guide.* Virtasant. https://www.virtasant.com/blog/data-lake-architecture