---------------------------------------------------------------------------------------------------------------------------------------------------------------------

**BI : Introduction :** 2030 More than 180 Zettabytes of data. How to harness the power of data? Make informed decisions.

|  |  |
| --- | --- |
| A diagram of a person typing on a keyboard  Description automatically generated | A diagram of data ecosystem  Description automatically generated |
| A diagram of a company's landscape  Description automatically generated | A diagram of a diagram  Description automatically generated |
| A close-up of a chart  Description automatically generated | A diagram of a triangle with points and a flag  Description automatically generated |

Data is unorganized information that is processed to make it meaningful. Generally data comprises of facts, observations, perception, numbers, characters, symbols, and images that can be interpreted to derive meaning.

Data :

1. Structured : Data that follows a rigid format and can be organized neatly into rows and columns. Typically in databases and spreadsheets.

* Data can be stored in well-defined schemas such as databases.
* Can be represented in a tabular manner with rows and columns.
* For example : SQL Databases, Online Transactions Processing (OLTP), Spreadsheets, Online forms, Sensor GPS/RFID, Network and sensor logs.

1. Semi-structured : Mix of data that has consistent characteristic and data that doesn’t conform to a rigid structure..

* Has some organizational properties but lacks a fixed or rigid schema.
* Can’t be stored in form of rows and columns as in databases.
* Contains tag, and elements or metadata which is used to group data and organize it in a hierarchy.
* For example : Emails, XML/JSON(other markup languages), binary executables, TCP/IP Packets, Zipped files, Integration of data.

1. Unstructured : Data that is complex, and mostly qualitative information that is impossible to reduce to rows and columns. For example Photos(JPEG, GIF, PNG), videos, PDF filesand social media content. Unstructured data can be stored in files and documents such as Word doc for manual analysis or in NoQL databases.

Data can come in variety of file formats, such as :

1. Relational Database
2. Non-Relational Database
3. APIs ; API can return data in a wide variety of formats such as plain text, XML, HTML, or JSON among others.
4. Web ervices
5. Data Stream
6. Social media platforms
7. Sensor devices.

Data repositories : (a term that includes databases, data warehouses, data marts, data lakes and big data stores)

Type, format and sources of data influence the type of data repositories that you can use to collect, store, clean, analyze, and mine the data for analysis.

Data Gathering, ETL => Data Wrangling and Cleaning => Data Analysis and Mining => Data Visualization

A close-up of a number

Description automatically generated

Delimited text Files :

* Text files used to store data as text in which each line, or row has values separated by a delimiter;
* A screenshot of a computer

  Description automatically generatedwhere delimiter is a sequence of one or more characters for specifying the boundary between independent entites or values.
* Most common delimiters are the comma, Tab, colon, vertical bar and space.
* CSVs(Comma) or TSVs(Tab) are the most commonly used file types in this category.
* Delimiters also represent one of various mean to specify boundaries in a data stream.

Microsoft Excel Open XML Spreadsheet or XLSX (Workbook) :

* It is an XML based file format created by Microsoft.
* In an XLSX, there can be multiple worksheets, and each worksheets is organized into rows and columns at the intersection of which is the cell . Each cell contains data.
* XLSX uses the open file format, which means it’s generally accessible to most other applications.

A screenshot of a computer code

Description automatically generated

Extensible Markup Languages (XML) :

* Markup language with set rules for encoding data.
* XML file format is both readable by humans and machines.
* It’s self descriptive language designed for sending information over the internet.
* XML is similar to HTML in some respects, but also has differences.
* XML is platform independent and programming language independent and therefore simplifies data sharing between various systems.

Portable Document Format (PDF) :

* A file format developed by adobe to present documents independent of application software, hardware and operating system which means it can be viewed the same way on any device.
* This format is frequently used in legal and financial documents and can also be used to fill in data such as for forms.

JavaScript Object Notation (JSON) :

* Text based open standard designed for transmitting structured data over the web.
* The file format is a language independent data format that can be read in any programming language.
* JSON is eay to use, is compatible with a wide range of browsers, and is considered as one of the best tools for sharing data of any size and type even audio and video.
* Many APIs and web services return data as JSON.

Common sources of data :

* Relational databases, ie. SQL Server, Oracle, MySQL, IBM DB2, to store data in a structured way.
* Flat files(CSV) and XML datasets, ie. Government organizations releasing demographic and economic datasets on an ongoing basis. Also companies that sell specific data such as Weather data, Financial data. Flat files store data in plain text format with one record or row per line, and each values separated by delimiters such as commas, or tabs. XML contain data values that are identified or marked up using tags. Flat files maps to a single table, but XML can support more complex data structures such as hierarchical.
* APIs (Application Program Interfaces) and web services, typically listen for incoming request which can be in the form of web requests from users or network requests from applications and return data in plain text, XML , HTML, JSON or media files. For example : Twitter and Facebooks API to source data from tweets and posts for performing tasks such as opinion mining or sentiment analysis. Also Stock Market APIS used for pulling data for trading and analysis.
* Web Scraping, used to extract relevant data (images, videos, text) from unstructured sources. Also known as screen scraping, web harvesting and web data extraction. Web scraping makes it possible to download specific data from web pages baed on defined parameters. For example, BeautifulSoup, Scrapy, pandas and selenium.
* Data Streams and Feeds , for aggregating constant streams of data flowing from sources such as instruments, IoT devices and applications, GPS data from cars, computer program, webites and social media posts. This data is generally timestamped and also geo tagged for geographical identification. For example, KAFKA, Apache Spark, and Apache Storm.
* RSS (Really simple syndication) for capturing update data from online forums and news sites where data is refreshed on an ongoing basis. Using feed reader, which is an interface that converts RS text files into a stream of updated data, updates are streamed to user devices.

Using SQL, you can :

* Insert, update, and delete records in a database.
* Create new databases, tables and views
* Write stored procedures.

Advantages of SQL :

* SQL is portable and platform independent
* Can be used for querying data in variety of DB and data repositories
* Has a simple syntax that is similar to English languages
* Syntax allows developers to write programs with fewer line of code using basic keywords (SELECT, INSERT)
* Can retrieve large amounts of data quickly and efficiently
* Runs on an interpreter system.

Powershell is a cross platform automation tools and configuration framework by Microsoft that is optimized for working with structured data format, such as JSON, CSV, XML, and REST APIs, websites and office applications. It consists of command line shell and scripting language. Shell and scripting languages are commonly used for automating repetitive operational tasks.

Metadata is data that provides information about other data. We'll consider the following three main types of metadata:

* Technical metadata

Technical metadata is metadata which defines the data structures in data repositories or platforms, primarily from a technical perspective. For example, technical metadata in a data warehouse includes assets such as: Tables, a data catalog.

* Process metadata

Process metadata describes the processes that operate behind business systems such as data warehouses, accounting systems, or customer relationship management tools.

* Business metadata

Users who want to explore and analyze data within and outside the enterprise are typically interested in *data discovery*. They need to be able to find data which is meaningful and valuable to them and know where that data can be accessed from.

A diagram of data warehouse

Description automatically generated**Popular tools for metadata management**

Popular metadata management tools include:

* IBM InfoSphere Information Server
* CA Erwin Data Modeler
* Oracle Warehouse Builder
* SAS Data Integration Server
* Talend Data Fabric
* Alation Data Catalog
* SAP Information Steward
* Microsoft Azure Data Catalog
* IBM Watson Knowledge Catalog
* Oracle Enterprise Metadata Management (OEMM)
* Adaptive Metadata Manager

A data repository is a general term used to refer to data that has been olleted, organized, and isolated so that it can be used for business operations or mined for reporting and data analysis. It can be a small or large database infrastructure with one or more databases that collect, manage, and store data sets. Data repositories help to isolate data and make reporting and analytics more efficient and credible while also serving as a data archive.Types of data repositories, include :

1. Databases

Is collection of data or information, designed for the input, storage , search and retrieval and modification of data. DBMS is a set of programs that creates and maintains the database. It allows you to store, modify, and extract information from the database using a function called querying. RELATIONAL (SQL) and Non RELATIONAL (No-SQL).

1. Data Warehouses

Works as a central repository that merges information coming from disparate sources and consolidates it through the ETL process into one comprehensive database for analytics and business intelligence. ETL process helps to extract data from different data sources, transform the data into a clean and usable state, and load the data into the enterprises data repository. Related to data warehouses are the concepts of Data marts and Data Lakes. Both have historically been relational since much of the traditional enterprise data has resided in RDBMS.

1. Big data stores

That include distributed computational and storage infrastructure to store, scale, and process very large datasets.

A Relational DB is a collection of data organized into a table structure, where the tables can be linked, or related based on data common to each. Tables are made of rows and columns, where rows are the ‘records’ and the columns is the ‘attributes’. Relational DB range from small desktop systems to massive cloud-based systems. Cloud based relational DB also referred to as Database as a service, are gaining wide use as they have access to the limitless compute and storage capabilities offered by the cloud. For example Amazon RDS, Google Colud SQL, IBM DB2 on Cloud and SQL Azure.

RDBMS doesn’t work well with semi-structured and unstructured data and is therefore, not suitable for extensive analytics on such data.

**NoSQL(Not only SQL)** is a non relational database design that provides flexible schemas for the storage and retrieval of data. Nowadays become more popular in the era of cloud, big data and high volume web and mobile applications. No is an abbreviation for ‘not only’. NoSQL have flexible schemas that allows programmers to create and manage modern applications. They don’t use a traditional row/column/table database degin with fixed schemas, and typically not use the structured query language to query data, although some may support SQL.

NoSQL allows data to be stored in a schema-less or free-form fashion. There are 4 common types of NoSQL based on the model being used for storing data :

1. **Key value store** : Stored as a collection of key value pairs. The key represents an attribute of the data and is a unique identifier. Both keys-values can be anything from simple integers or strings to complex JSON documents. Key-Value stores are great for storing user serssion data and user preferences, making real time recommendations and targeted advertising and in-memory data caching. However, if you want to be able to query the data on specific data value, need relationships between data values, or need to have multiple unique key, a key-value store may not be the best fit. For Example : Redis, Memcached, and DynamoDB.
2. **Document Based :** store each record and its associated data within a single document. Enable flexible indexing, powerful ad hoc queries and analytic over collections of documents. Document databases are preferable for eCommerce platforms, medical record storage, CRM platforms, and analytics platform. However if you are looking to run complex search queries and multi operation transactions, a document based database may not be the best option for you. For example MongoDB, documentDB, CouchDB and Cloudant.
3. **Column Based :** store data in cell grouped as columns of data instead of rows. A logical grouping of columns, that is column that are usually accessed together, is called a column family. Example,a customer name and profile information will most likely be accessed together but not their purchase history. So, customer name and profile information data can be grouped into a column family. Since column databases store all cells corresponding to a column as a continuous disk entry ,accessing and searching the data becomes very fast. Column databases can be great for systems that require heavy write request, storing time series data, weather data, and IOT data. But if you need to use complex queries or change your querying patterns frequently, thi may not be the best option for you. FOR example is Apache HBase and Cassandra.
4. A black and white chart with white text

   Description automatically generated**Graph Based :** use a graphical model to represent and store data. They are useful for visualizing ,analyzing, and finding connections between different pieces of data. Circles are node, and the arrows repreent relationships. Graph DB are an excellent choice for working with connected data, which is data that contains lots of interconnected relationships. Graph DB are great for social media networks, real time product recommendations, network diagram, fraud detection and access management. But if you want to process high volumes of transaction, it may not be the best choice, because graph DB are not optimized for large volume analytics queries. For example, Neo4J and cosmosDB.

**NoSQL** was created to the limitations of tradional relational database technology. Which advantages:

1. Ability to handle large volumes of structured, semi-structured and unstructured data.
2. Ability to run as distributed systems scaled across multiple data centers, which enables them to take advantages of cloud computing infrastructure.
3. An efficient and cost-effective scale out architecture that provides additional capacity and performance with the addition of new nodes.
4. A diagram of data warehouse

   Description automatically generatedSimpler design, better control over availability, and improved scalability that enables you to be more agile, more flexible and to iterate more quickly.

**Data Warehouse** works like a multi-purpose storage for different use cases. By the time data comes into the warehouse, it has already been modeled and structured for specific purposes, meaning it is analysis ready. Data warehouses serve as the single source of truth – storing current and historical that has been cleaned, conformed and categorized. A data warehouse is a multi purpose enable of operational and performance analytics.

A diagram of data warehouse

Description automatically generated

A data mart is a subsections of the data warehouse, built specifically for a particular business, function, purposes or community of users. The idea is to provide stakeholders data that is most relevant to them when they need it. For example sales or finance team accessing data for their quarterly reporting and projections. Data marts offers analytical capabilits for a restricted area of the data warehouse, and offer isolated security and isolated performance. The most important role is business specific reporting and analytics.

**A diagram of data processing

Description automatically generatedA data lake** is storage repository that can storage large amounts of structured, semi structured, unstructured in their native format, classified and tagged with metadata. While Data WH stores data processed for a specific need:

* a data lake is a pool of raw data where each data elements is given a unique identifier and is tagged with metatags for further use.
* Data from a data lake is selected and organized based on the use case you need it for.
* Unlike data warehouse, a data lake would retain all source data without any exclusions. And data could include all types of data soures and types.
* Data lakes are sometimes also used as staging area of a data warehouse.
* The most important role of a data lake is in predictive and advanced analytics.

A diagram of a cylinder

Description automatically generatedETL is how raw data is converted into analysis ready data. It is an automated process in which you :

* gather raw data from identified sources,
* extracting information needed for reporting and analysis.
* Cleaning, standardizing, and transforming data into usable format
* Loading data into a data repository.

1. Extract => Data from source locations is collected for transformation. Can be Batch processing, stream processing,
2. Transformation => executions of rules and functions that converts raw data into data that can be used for analysis. For example making date formats, making unit of measurement across all sourced data, removing duplicate data, filtering out data that you do not need
3. Load => Processed data is transported to a destination system or data repository. It could be initial loading (populating all data in the repo), incremental loading (applying ongoing updates and modifications a needed periodically), full refresh (erasing contents of one or more tables and reloading with fresh data).

ETL and data pipelines used interchangeably. And although both move data from source to destination, data pipeline is a broader term that :

1. Encompasses the entire journey of moving data from one system to another, of which ETL is subset.
2. Can be used for architected for batch processing, for streaming data and a combination of batch and streaming data.
3. High performing systems that support both long running batch queries and smaller interactive queries.
4. The destination for a data pipeline is typically a data lake, althotugh the data may also be loaded to different target destinations such as another application or a visualization tools.

Most popular Data pipelines : Apache Beam, DataFlow

VIDEO ANALOGY Data Lakes, Data Warehouse dengan Konsep Restaurant :

https://www.coursera.org/learn/business-intelligence-essentials/lecture/Y59c1/optional-data-lakehouses-explained

**DATA MODELLING**

Build a house :

1. You know your requirements, but don’t know how much it would cost
2. to resolve this (unknown cost), discuss your requirements with architect who provides you with a blueprint (blueprint makes easy).

Data model similar to architect’s blueprint.

Data modeling is the process of creating a blueprint.

Data modeling creates a representation of data for efficient storage, retrieval, and analysis to meet specific requirements. This helps to define how data is organized and the relationship between different entities; and provides a blueprint for the database to manage and use data.

Data modeling can be :

* defined as the process of visually representing of visually representing all the complex relationships between different data that will form your database.
* It represents data and information in the form of text or symbols.
* To create a database structure, a data analysts establish a robust data framework with the help of data models, which helps to understand the data better and provide an efficient way to update and delete data.

Data model is :

* like an outline that data architects use to build a solid foundation for a system.
* It helps create primary models and
* Establishes link between different data elements.

3 Types of data models based on levels of abstractions :

1. Conceptual data model
   1. To convey business structure and concepts at a high level.
   2. Business stakeholders and data architects collaborate to create data models that define and organize information to build a system.
2. Logical data model
   1. Defines the features and connection of entities in a system independently of any specific database management systems (DBMS).
   2. It serves as a foundation for creating a physical data model by adding more details.
3. Physical data model.
   1. Explains the implementation of logical data models using a specific DBMS.
   2. It converts abstract concepts related to the system’s entities into schemas, customized for a particular database platform.

A diagram of types of data models

Description automatically generated

How data models are useful .

1. Collaborate with stakeholders
2. Defines entites and their attributes
3. Establish connections between data objects
4. Create accurate data model for use by applications developers, DB admin and others takeholders.

4 Types of data models based on the types of data and industry :

1. Entity Relationshipd (E-R model)
   1. Visual representation of entities and thie relationships
   2. Establishes an entity attributes, connection and constraints .
2. Hierarchical model
   1. Arranges data in a tree like structure.
   2. Data get linked to a central root from which the hierarchy expands like a tree,
   3. Each level has one parent node and multiple child nodes.
3. Network model
   1. Extension of the hierarchical model.
   2. Allows multiple relationships between the connected nodes and has many parent nodes.
   3. Additionally multiple relationships are linked between two record types
4. Relational model
   1. Organizes data into tables using column and rows.
   2. Uses key to identify each record in the database.

**Contoh ER Model : Contoh Hierarchical model**

A diagram of a patient

Description automatically generatedA diagram of a shoe store

Description automatically generated

**Contoh network Model :**

A diagram of a company

Description automatically generated with medium confidenceA diagram of sales and sales

Description automatically generated

Benefits of data modelling :

* Provides insight to spot missing or incorrect data.
* Make smarter decisions immediately.
* Help to clean data by standardizing it , developing hierarchies, and ensuring consistency and usability.
* Improves communication between technical and non-technical stakeholders for organizational growths
* Save resources by designing a database in advance.
* Data modelling compliance within the organization by understanding its data protection responsibilities to ensure proper data governance and compliance.

Increasing number of internet connected devices that we interact with on a daily basis record vast amounts of data about us..

BIG DATA is ‘’the dynamic, large and disparate volumes of data being created by people, tools and machines. It requires new, innovative, and scalable technology to collect, host and analytically process the vast amount of data gathered in order to drive real time business inights that relate to consumers, risk, profit, performance, productivity, management, and enhanced shareholder value.’’ There is no one definition of big data but there are certain elements that are common across the different definitions, such as velocity. Volume, variety, veracity and value. 5 V of Big data.

The ‘V’ of Big DATA :

1. Velocity is the speed at which data accumulates. Data is being generated extremely fast in a process that never stops.
2. Volume is the scale of the data or increase in the amount of data stored. Increase in data sources, higher resolution sensors/images, and also scalable infrastructure are the factor that increase Volume of BIG DATA.
3. Variety is the diversity of the data. Structured data fits neatly into rows and columns in relational databases, while unstructured data is not organized in a predefined way like tweets, blog posts, pictures and videos. Variety also reflet that data comes from different sources, machines, people and processes, both internal and external to organizations.
4. Veracity is the quality and origin of data and its conformity to facts and accuracy. Attributes include consistency, integrity, completeness and ambiguity. Drivers include cost and the need for traceability.
5. Value is our ability and need to turn data into value. Value isn’t just profit. It may have medical or social benefits, as well as customer, employee, or personal satisfaction.

BIG DATA Processing tools :

1. Apache Hadoop : collection of tools that provides distributed storage and processing of big data.
2. Apache Hive : Data warehouse **for data query and analysis** built on top of Hadoop.
3. Apache Spark : Distributed data analytic framework designed to perform complex data analytics in real time.

Consideration for choice Data Repository :

1. A Number of different factors influence the selection of the right data repo
2. Type of data (structured, semi, unstructured)
3. Schema of data
4. Performance requirements
5. Whether you are working with data at rest or streaming data
6. Data encryption needs (security features)
7. Volume of data and whether you need a BIG Data system.
8. Storage requirements (Scalability)
9. Frequency of data access (Frequen updates, keep in vault for long time)
10. Database repo that you need.
11. Purposes of data repo (Transactional, Analytical, Archival, Data Warehousing)
12. Compatibility with existing ecosystem
13. Cost of various solutions.

Data integration is a discipline compriing the practices, architectural, techniques and tools that allow organizations to ingest, transform, combine, and provision data across various data types.

Data Integration includes :

1. Accessing, queueing, or extracting data from operational systems
2. Transforming, and merging extracted data either logically or physically.
3. A group of logos and symbols

   Description automatically generatedData quality and governance.
4. Delivering data through an integrated approach for analytic purposes.

Capabilities of modern data integration platform :

1. Pre-built connectors and adapters
2. Open Source architecture
3. Optimization for both batch processing of large scale data and continuous data stream or both
4. Integration with Big Data sources
5. Additional functionalities for data quality and governance, compliance and security
6. Portability between on premise and different types of cloud environments (single cloud, multi cloud, hybrid).

TOOLS :

1. RDBMS databases such as MySQL, Postgre SQL
2. NoSQL databases such as MongoDB or Cassandra
3. Graph databases such as Neo4J
4. Python, with Pandas
5. Apache airflow for building data pipelines/data orchestration
6. Spark or Hadoop for BIG Data Processing
7. SQL Server Data Repository
8. Okada for data integration
9. AWS Redshift Data Warehouse
10. AWS S3 Data Lake
11. Internal tool for ETL pipeline
12. Kafka or Websphere MQ for handling streaming data
13. Talend for ETL
14. Beautiful Soup and Scrappy for web scraping
15. Variety of cloud storages
16. Lifelong learner

Congratulations! You have completed this lesson. At this point in the course, you know:

* Databases, which can be relational or non-relational, each following a set of organizational principles, the types of data they can store, and the tools used to query, organize, and retrieve data.
* Data warehouses consolidate incoming data into a comprehensive storehouse.
* Data marts are essentially subsections of a data warehouse and are built to isolate data for a particular business function or use case.
* Data lakes serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.
* Big data stores provide distributed computational and storage infrastructure to store, scale, and process large data sets.
* The extract, transform, and load (ETL) process is an automated process that converts raw data into analysis-ready data by:
  + Extracting data from source locations
  + Transforming raw data by cleaning, enriching, standardizing, and validating it
  + Loading the processed data into a destination system or data repository
* The data pipeline often used interchangeably with the ETL process, encompasses the entire journey of moving data from the source to a destination data lake or application using the ETL process.
* Data modeling is used to visually represent the data using text and symbols. On the other hand, a data model is a structure used to organize data to ensure that all the data objects required by the database or data warehouses are accurately presented.
* Data models are classified based on their levels of abstraction, such as conceptual, logical, and physical.
* The four types of data models based on the types of data used are: Entity-relationship or E-R model, hierarchical model, network model, and relational model.
* The data modeling provides insights into the data, helps clean data, improves communication between stakeholders, saves resources, and supports compliance.
* Big data refers to the huge amounts of real-time data that people, tools, and machines produce. The sheer velocity, volume, and variety of data challenge the tools and systems used for conventional data. These challenges led to the emergence of processing tools and platforms designed specifically for big data, such as Apache Hadoop, Apache Hive, and Apache Spark.

1. SQL is the standard querying language for RDBMSs.
2. This is one of the abilities of RDBMSs that make them very well suited for OLTP applications.
3. Document-based NoSQL databases store each record and its associated data within a single document and work well with Analytics platforms.
4. A data mart is a sub-section of the data warehouse used to isolate a subset of data for a particular business function, purpose, or community of users.
5. Velocity, in the context of Big Data, is the speed at which data accumulates.
6. Hadoop, a java-based open-source framework, allows distributed storage and processing of large datasets across clusters of computers.
7. The emergence of NoSQL technology has made it possible for data marts and data warehouses to be used for both relational and non-relational data.
8. Graph-based NoSQL databases use a graphical model to represent and store data and are used for visualizing, analyzing, and finding connections between different pieces of data.
9. A Data Lake can store large amounts of structured, semi-structured, and unstructured data in their native format, classified and tagged with metadata.
10. Veracity, in the context of Big Data, refers to the accuracy and conformity of data to facts.
11. Spark is a general-purpose data processing engine used for performing complex data analytics in real-time.