**Data Literacy for Data Science**

# **Understanding Data**

## **Lesson Overview: Understanding Data**

In this lesson, "Understanding Data," you'll explore data basics through videos on data comprehension and sources. The reading covers metadata's role, and quizzes reinforce your understanding. The lesson ends with a summary video, ensuring you grasp essential data concepts.

|  |  |
| --- | --- |
| **Asset name and type** | **Description** |
| “Understanding Data:” video | Gain foundational insights into data comprehension. The video will also help you explore data types, characteristics, and their importance in various fields of study. |
| “Data Sources” video | Explore the diverse origins of data in this video, uncovering where and how data is generated, collected, and utilized in different contexts. |
| “Working on Varied Data Sources and Types” video | This video equips you with the skills to effectively manage and analyze data from a wide range of sources and in various formats, ensuring adaptability in data handling. |
| “Metadata” reading | Read an excerpt on “Metadata” and discover the significance of metadata in data analysis through this reading, which encompasses three primary metadata types: technical, process, and business metadata. |
| Practice quiz | Test your knowledge of metadata concepts based on the previous reading. |
| “Lesson Summary” video | Recap the key points from this lesson. |
| Practice quiz | Take a practice quiz to assess your comprehension of data fundamentals. |

## **Understanding data**

Understanding Data Types: Structured, Semi-Structured, and Unstructured

In our recent lesson, we explored the different types of data: structured, semi-structured, and unstructured. Think of structured data as a neatly organized library where every book has a specific place on the shelf, making it easy to find what you need. This type of data is well-defined and can be stored in databases, like numbers in a spreadsheet or records in a SQL database. It’s straightforward and can be analyzed using standard methods.

On the other hand, semi-structured data is like a collection of books that are organized by genre but not by author or title. It has some organization, using tags or metadata to group information, but it doesn’t fit neatly into rows and columns. Examples include emails or XML files. Finally, unstructured data is like a messy attic filled with various items—photos, videos, and documents—where there’s no clear organization. This type of data is more challenging to analyze because it doesn’t follow a specific format.

summary of the types of data:

* **Structured Data**:
  + Well-organized and follows a specific format.
  + Stored in databases (e.g., SQL databases, spreadsheets).
  + Easy to analyze using standard methods.
* **Semi-Structured Data**:
  + Some organization but lacks a fixed schema.
  + Uses tags or metadata for grouping (e.g., emails, XML files).
  + Cannot be stored in traditional rows and columns.
* **Unstructured Data**:
  + No identifiable structure or organization.
  + Includes various formats like images, videos, and documents.
  + More challenging to analyze due to its diverse nature.

## **Data Sources**

concept of **Data Sources** in a simple way.

Data sources are like different places where we can find information that we can use for analysis. Imagine you are a detective trying to solve a mystery. You would gather clues from various places: a library (books), a witness (people), and even the internet (websites). Similarly, in the world of data science, we have various sources to collect data, such as relational databases (like SQL Server), flat files (like CSV files), APIs (which are like bridges to get data from other applications), and even web scraping (which is like gathering information from websites).

For example, think of a restaurant that wants to know which dishes are popular. They might look at their sales data stored in a database, check customer reviews on social media using APIs, and even scrape data from food blogs to see what people are saying about their menu. By combining all this information, they can make better decisions about what to serve!

summary of **Data Sources**:

* **Data Sources** are various places where information can be collected for analysis.
* Common types include:
  + **Relational Databases**: Structured data storage (e.g., SQL Server, MySQL).
  + **Flat Files**: Simple text files like CSV, where data is organized in rows and columns.
  + **APIs**: Interfaces that allow applications to communicate and share data (e.g., social media data).
  + **Web Scraping**: Extracting data from websites based on specific criteria.
  + **Data Streams**: Continuous flows of data from sources like IoT devices or social media.

These sources help organizations gather insights to make informed decisions, such as analyzing sales trends or customer preferences.

## **Viewpoints: Working with Varied Data Sources and Types**

**working with varied data sources and types** into simpler terms:

1. **What are Data Sources?**
   * Data sources are places where data comes from. This can include:
     + **Relational Databases**: These store data in tables (like a spreadsheet).
     + **NoSQL Databases**: These are more flexible and can store data in different formats, like documents or key-value pairs.
     + **Files**: Such as CSV (comma-separated values), JSON (JavaScript Object Notation), and XML (eXtensible Markup Language).
2. **Types of Data:**
   * Data can be structured (organized in a specific way, like tables) or unstructured (not organized, like text or images).
   * Examples of unstructured data include logs (records of events), documents, and social media posts.
3. **Challenges:**
   * Different data sources and types can be tricky to work with because:
     + They may have different formats (like how dates are written).
     + They might require different tools or methods to access and analyze them.
     + You may need to clean or transform the data to make it usable.
4. **Flexibility and Learning:**
   * As a data professional, you need to be flexible and open to learning new skills to handle various data types and sources effectively.

In summary, working with varied data sources and types means understanding where data comes from, recognizing the different formats it can take, and being prepared to tackle the challenges that arise when using this data.

summary of **working with varied data sources and types**:

* **Data Sources**: Places where data originates, including relational databases, NoSQL databases, and various file formats (CSV, JSON, XML).
* **Types of Data**: Can be structured (organized in tables) or unstructured (like text or images).
* **Challenges**: Different formats and requirements can complicate data handling, necessitating cleaning and transformation.
* **Flexibility**: Data professionals must be adaptable and willing to learn new skills to effectively manage diverse data types and sources.

## **Metadata**

**Metadata and Metadata Management**

**Objectives**

After completing this reading, you will be able to:

* Define what metadata is
* Describe what metadata management is
* Explain the importance of metadata management
* List popular tools for metadata management

**What is metadata?**

Metadata is data that provides information about other data.

This is a very broad definintion. Here we will consider the concept of metadata within the context of databases, data warehousing, business intelligence systems, and all kinds of data repositories and platforms.

We'll consider the following three main types of metadata:

* Technical metadata
* Process metadata, and
* Business 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 that record information about the tables stored in a database, like:
  + each table's name
  + the number of columns and rows each table has
* A data catalog, which is an inventory of tables that contain information, like:
  + the name of each database in the enteprise data warehouse
  + the name of each column present in each database
  + the names of every table that each column is contained in
  + the type of data that each column contains

The technical metadata for relational databases is typically stored in specialized tables in the database called the System Catalog.

**Process metadata**

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

Many important enterprise systems are responsible for collecting and processing data from various sources. Such critical systems need to be monitored for failures and any performance anomalies that arise. Process metadata for such sytems includes tracking things like:

* process start and end times
* disk usage
* where data was moved from and to, and
* how many users access the system at any given time

This sort of data is invaluable for troubleshooting and optimizing workflows and ad hoc queries.

**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. These business-minded users are thus interested in business metadata, which is information about the data described in readily interpretable ways, such as:

* how the data is acquired
* what the data is measuring or describing
* the connection between the data and other data sources

Business metadata also serves as documentation for the entire data warehouse system.

**Managing metadata**

Managing metadata includes developing and administering policies and processes to ensure information can be accessed and integrated from various sources and appropriately shared across the entire enterprise.

Creation of a reliable, user-friendly data catalog is a primary objective of a metadata management model. The data catalog is a core component of a modern metadata management system, serving as the main asset around which metadata management is administered. It serves as the basis by which companies can inventory and efficiently organize their data systems. A modern metadata managment model will include a web-based user interface that enables engineers and business users to easily search for and find information on key attributes such as CustomerName or ProductType. This kind of model is central to any Data Governance initiative.

**Why is metadata management important?**

Good metadata management has many valuable benefits. Having access to a well implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data.

Well managed metadata helps you to understand both the business context associated with the enterprise data and the data lineage, which helps to improve data governance. Data lineage provides information about the origin of the data and how it gets transformed and moved, and thus it facilitates tracing of data errors back to their root cause. Data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives.

The key focus areas of data governance include availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organization.

**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
* Unifi Data Catalog
* data.world
* Informatica Enterprise Data Catalog

**Summary**

In this reading, you learned that:

* Metadata is data that provides information about other data, and includes three main types: technical, process, and business metadata
* The technical metadata for relational databases is typically stored in specialized tables in the database called the system catalog
* A primary objective of business metadata management modelling is the creation and maintenance of a reliable, user-friendly data catalog
* Having access to a well-implemented data catalog greatly enhances data discovery, repeatability, governance, and can also facilitate access to data
* Metadata management tools from IBM include InfoSphere Information Server and Watson Knowledge Catalog

## **Lesson Summary: Understanding Data**

importance of understanding data in data science. Here’s a simplified breakdown of the key points:

* **Data Types**: There are three main types of data:
  + **Structured Data**: Organized in a clear format, like tables in a database.
  + **Semi-Structured Data**: Has some organization but doesn’t fit neatly into tables, like emails or XML files.
  + **Unstructured Data**: Chaotic and varied, such as social media posts or videos, requiring advanced tools to analyze.
* **Data Sources**: Data can come from various places, including:
  + Internal applications used by organizations.
  + Publicly available datasets or proprietary datasets that can be purchased.
* **Data Formats**: Common formats for data include:
  + **CSV**: A flat file format for storing data in a table.
  + **JSON**: A flexible format for data exchange that doesn’t require a predefined structure.
* **APIs**: Application Programming Interfaces (APIs) allow access to data from different sources, like Twitter or Facebook, for analysis.
* **Role of Data Scientists**: Data scientists need to understand these data types and sources to analyze data effectively, often working with large datasets.

summary of the lesson on understanding data:

* **Data Types**:
  + **Structured Data**: Organized in tables with rows and columns (e.g., databases).
  + **Semi-Structured Data**: Some organization but not in strict formats (e.g., XML, emails).
  + **Unstructured Data**: Chaotic and varied data (e.g., social media posts, videos) that often requires advanced analysis tools.
* **Data Sources**: Data can be sourced from internal applications, publicly available datasets, or purchased proprietary datasets.
* **Data Formats**: Common formats include CSV for flat files and JSON for flexible data exchange.
* **APIs**: APIs enable access to data from various platforms (e.g., Twitter, Facebook) for analysis.
* **Role of Data Scientists**: They must understand these data types and sources to effectively analyze large datasets.

## **Glossary: Understanding Data**

Welcome! This alphabetized glossary contains many of the terms in this course. These terms are important for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Video where the term is introduced** |
| Comma-separated values (CSVs) | Delimited text files where the delimiter is a comma. Used to store structured data. | Understanding Different Types of File Formats |
| Delimited text file formats | Text files are used to store data where each line or row has values separated by a delimiter. A delimiter is a sequence of one or more characters specifying the boundary between values. Common delimiters include comma, tab, colon, vertical bar, and space. | Understanding Different Types of File Formats |
| NoSQL databases | Databases are designed to store and manage unstructured data and provide analysis tools for examining this type of data. | Types of Data |
| Online Transaction Processing (OLTP) Systems | Systems that focus on handling business transactions and storing structured data. | Types of Data |
| Relational databases | Databases are designed to store structured data with well-defined schemas and support standard data analysis methods and tools. | Types of Data |
| Sensors | Devices such as Global Positioning Systems (GPS) and Radio Frequency Identification (RFID) tags generate structured data. | Types of Data |
| Spreadsheets | Software applications like Excel and Google Spreadsheets are used for organizing and analyzing structured data. | Types of Data |
| SQL Databases | Databases that use Structured Query Language (SQL) for defining, manipulating, and querying data in structured formats. | Types of Data |
| Tab-separated values (TSVs) | Delimited text files where the delimiter is a tab. Used as an alternative to CSV when literal commas are present in text data. | Understanding Different Types of File Formats |

# **Data Literacy**

## **Lesson Overview: Data Literacy**

In the "Data Literacy" lesson, you'll dive into the fundamental aspects of data through a series of instructional videos. You'll gain insights into data collection, organization, and the distinctions between Relational Database Management Systems (RDBMS) and NoSQL databases. Additionally, you'll navigate the world of Data Marts, Data Lakes, ETL processes, and Data Pipelines. The lesson ends with a summary video, ensuring you grasp essential data concepts.

|  |  |
| --- | --- |
| **Asset name and type** | **Description** |
| “Data Collection and Organization” video | Explore the fundamentals of collecting and organizing data in this instructional video. |
| “Relational Database Management System” video | Gain insights into the workings of Relational Database Management Systems (RDBMS) through this informative video. |
| “NoSQL” video | Delve into NoSQL databases and their unique characteristics in this video. |
| “Data Marts, Data Lakes, ETL, and Data Pipelines” video | Learn about Data Marts, Data Lakes, ETL processes, and Data Pipelines in this comprehensive video. |
| “Considerations for Choice of Data Repository” video | Discover key factors to consider when selecting a data repository in this video. |
| “Data Integration Platforms” video | Explore the role and significance of Data Integration Platforms through this enlightening video. |
| Practice quiz | Test your knowledge of Data Integration Platforms with this practice quiz. |
| “Lesson Summary” video | Recap the key points from this lesson. |
| Practice quiz | Take a practice quiz to assess your comprehension of data literacy. |
| Glossary | Access a comprehensive glossary that defines and clarifies key data literacy terms relevant to the field of data science. |
| “Summary” reading | Summarize the essential concepts and insights enhancing your understanding of data literacy within the context of data science. |

## **Data Collection and Organization**

concept of **data repositories** in simple terms.

A data repository is like a big storage box where all kinds of information are kept. Imagine you have a collection of books, each containing different stories and facts. Just like you organize your books on shelves, a data repository organizes data so that businesses can easily find and use it for their operations or to analyze trends. This can include everything from customer information to sales data, all neatly stored in databases, data warehouses, or big data stores.

To make it relatable, think of a data warehouse as a library. In a library, books from various genres are collected and organized so that readers can find what they need quickly. Similarly, a data warehouse gathers information from different sources, cleans it up, and stores it in one place for easy access and analysis. This helps businesses make informed decisions based on the data they have.

summary of the topic on **data repositories**:

* **Definition**: A data repository is a storage system for collected and organized data, used for business operations and analysis.
* **Types**:
  + **Databases**: Collections of data that allow for input, storage, and retrieval using a Database Management System (DBMS).
  + **Data Warehouses**: Central repositories that consolidate data from various sources for analysis, using the ETL (Extract, Transform, Load) process.
  + **Big Data Stores**: Systems designed to handle and process very large datasets, often used for big data analytics.
* **Importance**: Data repositories help businesses efficiently manage data, making reporting and analytics more credible and streamlined.

## **Relational Database Management System**

relational databases in simple terms.

A relational database is like a well-organized library where information is stored in tables, similar to how books are arranged on shelves. Each table has rows and columns: rows represent individual records (like a single book), and columns represent different attributes of that record (like the title, author, and genre of the book). For example, in a customer table, each row would contain information about a different customer, such as their name and address, while the columns would define what kind of information is stored, like "Customer Name" or "Customer Address."

Now, imagine you have a table for customers and another for their transactions, like purchases they made. These tables can be linked together using a common piece of information, such as a "Customer ID." This allows you to easily find all transactions related to a specific customer, just like finding all the books written by a particular author in a library. This linking capability helps you gather insights and make better decisions based on the data you have.

summary of relational databases:

* **Definition**: A relational database organizes data into tables, which consist of rows (records) and columns (attributes).
* **Linking Tables**: Tables can be related through common data, like a "Customer ID," allowing for easy retrieval of related information.
* **Data Integrity**: Relational databases minimize redundancy and ensure data consistency by storing information in a structured way.
* **Querying**: SQL (Structured Query Language) is used to query and manipulate the data efficiently.
* **Advantages**: They offer flexibility, ease of backup, and ACID compliance, ensuring reliable data transactions.
* **Use Cases**: Commonly used for online transaction processing (OLTP), data warehousing, and IoT solutions.

## **NoSQL**

NoSQL databases in simple terms.

NoSQL databases are like flexible storage boxes for data. Unlike traditional databases that have strict rules about how data should be organized (like putting all your toys in specific labeled boxes), NoSQL allows you to store data in a more relaxed way. You can mix different types of data together, whether it's structured (like a list of names), semi-structured (like a JSON file), or unstructured (like a collection of photos). This flexibility makes NoSQL databases great for handling large amounts of data, especially in today's world where we have lots of information coming from various sources like websites and mobile apps.

Imagine you have a big toy chest where you can throw in all your toys without worrying about how they are arranged. You can easily find your favorite toy later, even if it’s mixed in with others. Similarly, NoSQL databases let you store and retrieve data quickly without needing to follow strict rules. This makes them perfect for businesses that need to adapt quickly to changing data needs.

summary of NoSQL databases:

* **Definition**: NoSQL stands for "not only SQL" and refers to non-relational databases that allow flexible data storage.
* **Flexibility**: Unlike traditional databases with fixed schemas, NoSQL databases can store structured, semi-structured, and unstructured data in a more relaxed manner.
* **Types**: There are four common types of NoSQL databases:
  + **Key-value stores**: Store data as key-value pairs (e.g., Redis).
  + **Document-based**: Store data in documents (e.g., MongoDB).
  + **Column-based**: Store data in columns rather than rows (e.g., Cassandra).
  + **Graph-based**: Use graphs to represent data and relationships (e.g., Neo4J).
* **Advantages**: NoSQL databases can handle large volumes of data, are cost-effective, and provide better scalability and performance for modern applications.

## **Data Marts, Data Lakes, ETL, and Data Pipelines**

concepts in simple terms:

1. **Data Mart**:

* A **data mart** is like a small section of a data warehouse, focused on a specific area of a business, such as sales or finance.
* It provides relevant data to specific users or teams, making it easier for them to access the information they need for their reports and analysis.
* Think of it as a specialized library section that only contains books related to a particular subject.

2. **Data Lake**:

* A **data lake** is a large storage space that holds all types of data in its original form, whether it's structured (like tables), semi-structured (like JSON files), or unstructured (like videos or text).
* It allows organizations to keep all their data without having to sort it first, making it useful for future analysis.
* Imagine a big pool where you can throw in all kinds of items without worrying about organizing them right away.

3. **ETL (Extract, Transform, Load)**:

* **ETL** is a process used to prepare data for analysis. It involves three steps:
  + **Extract**: Collecting raw data from different sources.
  + **Transform**: Cleaning and organizing the data to make it usable (like fixing errors or changing formats).
  + **Load**: Putting the cleaned data into a data warehouse or another storage system.
* Think of it as a cooking process where you gather ingredients (extract), prepare them (transform), and then serve the dish (load).

4. **Data Pipeline**:

* A **data pipeline** is a system that moves data from one place to another, often in real-time.
* It can include ETL processes but is broader, covering the entire journey of data from its source to its destination.
* Imagine a conveyor belt that continuously moves items from one station to another, ensuring everything is processed and delivered efficiently.

summary of the concepts:

* **Data Mart**:
  + A small section of a data warehouse focused on a specific business area (e.g., sales or finance).
  + Provides relevant data for specific users or teams.
* **Data Lake**:
  + A large storage space for all types of data in its original form (structured, semi-structured, unstructured).
  + Useful for keeping data for future analysis without immediate organization.
* **ETL (Extract, Transform, Load)**:
  + A process to prepare data for analysis involving:
    - **Extract**: Collecting raw data.
    - **Transform**: Cleaning and organizing the data.
    - **Load**: Storing the cleaned data in a data warehouse.
* **Data Pipeline**:
  + A system that moves data from one place to another, often in real-time.
  + Encompasses the entire journey of data, including ETL processes.

## **[Viewpoints: Considerations for Choice of Data Repository](https://www.coursera.org/learn/what-is-datascience/lecture/McOUm/viewpoints-considerations-for-choice-of-data-repository)**

The topic we are discussing is about choosing the right data repository, which is like selecting the best storage space for your important items, but in this case, it's for data.

When data professionals decide on a data repository, they consider several factors, such as:

* **Type of Data**: Is the data structured (like a neatly organized spreadsheet), semi-structured (like an email), or unstructured (like a video)?
* **Performance Needs**: How fast do you need to access the data? Is it for quick transactions or for analyzing large amounts of information?
* **Volume of Data**: How much data are you dealing with? If it's a lot, you might need a special system designed for big data.

Imagine you have a collection of books, toys, and clothes. You wouldn't store them all in the same box, right? You'd want to keep your books on a shelf, your toys in a toy chest, and your clothes in a wardrobe. Similarly, data professionals choose different types of databases based on the kind of data they have and how they plan to use it.

summary of the topic:

When selecting a data repository, data professionals consider several key factors:

* **Type of Data**: Understanding whether the data is structured, semi-structured, or unstructured.
* **Performance Requirements**: Determining how quickly the data needs to be accessed.
* **Volume of Data**: Assessing the amount of data to decide if a standard or big data system is needed.
* **Access Type**: Considering whether the data will be accessed frequently or for long queries.
* **Compatibility**: Ensuring the repository works well with existing tools and systems.
* **Security Features**: Evaluating the security measures provided by the repository.
* **Scalability**: Ensuring the repository can grow with the organization’s needs.

Choosing the right data repository is crucial for effective data management and analysis.

## **Data Integration Platforms**

**data integration** in simple terms.

Data integration is like putting together pieces of a puzzle. Imagine you have different pieces of information scattered across various places, like customer details in sales, marketing, and finance. Data integration helps you gather all these pieces and combine them into one complete picture. This way, you can see everything about your customers in one place, making it easier to analyze and understand their behavior.

For example, think of a library where books are stored in different sections. If you want to find all the books by a specific author, you would need to check each section. Data integration acts like a librarian who knows where every book is located and can quickly bring all the books by that author to you. This unified view allows businesses to make better decisions based on comprehensive data.

summary of **data integration**:

* **Definition**: Data integration is the process of combining data from different sources to create a unified view.
* **Purpose**: It helps organizations access, transform, and manage data effectively for analysis and decision-making.
* **Usage Scenarios**: Common uses include ensuring data consistency, master data management, data sharing, and data migration.
* **Relation to ETL and Data Pipelines**: Data integration involves using data pipelines to move data, while ETL (Extract, Transform, Load) is a specific process within data integration.
* **Modern Solutions**: Current data integration tools often include pre-built connectors, support for big data, and capabilities for cloud environments.

## **Lesson Summary: Welcome to Data Literacy**

simple summary of the lesson on **data literacy**:

1. **Understanding Data**: Data literacy means being able to read, understand, create, and communicate data. It’s important for making sense of the information we collect.
2. **Data Storage**: There are different ways to store data, depending on what type it is:
   * **Relational Databases**: Good for structured data (like tables with rows and columns).
   * **NoSQL Databases**: Better for semi-structured or unstructured data (like documents or social media posts).
3. **Data Warehouses and Lakes**:
   * **Data Warehouse**: A place to store organized data for reporting and analysis.
   * **Data Lake**: A storage area for large amounts of raw data in its original format.
4. **Data Pipelines**: These are processes that help collect, transform, and move data so it can be analyzed. One common method is called **ETL** (Extract, Transform, Load), which prepares raw data for analysis.
5. **Importance for Data Scientists**: As a future data scientist, understanding these concepts helps you manage and analyze data effectively, leading to better insights and decisions.

## **Data Literacy for Data Science Lesson Glossary**

Welcome! This alphabetized glossary contains many of the terms in this course. These terms are important for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Video where the term is introduced** |
| ACID-compliance | Ensuring data accuracy and consistency through Atomicity, Consistency, Isolation, and Durability (ACID) in database transactions. | Relational Database Management System |
| Cloud-based Integration Platform as a Service (iPaaS) | Cloud-hosted integration platforms that offer integration services through virtual private clouds or hybrid cloud models, providing scalability and flexibility. | Data Integration Platforms |
| Column-based Database | A type of NoSQL database that organizes data in cells grouped as columns, often used for systems requiring high write request volume and storage of time-series or IoT data. | NoSQL |
| Data at rest | Data that is stored and not actively in motion, typically residing in a database or storage system for various purposes, including backup. | Considerations for Choice of Data Repository |
| Data integration | A discipline involving practices, architectural techniques, and tools that enable organizations to ingest, transform, combine, and provision data across various data types, used for purposes such as data consistency, master data management, data sharing, and data migration. | Data Integration Platforms |
| Data Lake | A data repository for storing large volumes of structured, semi-structured, and unstructured data in its native format, facilitating agile data exploration and analysis. | Data Marts, Data Lakes, ETL, and Data Pipelines |
| Data mart | A subset of a data warehouse designed for specific business functions or user communities, providing isolated security and performance for focused analytics. | Data Marts, Data Lakes, ETL, and Data Pipelines |
| Data pipeline | A comprehensive data movement process that covers the entire journey of data from source systems to destination systems, which includes data integration as a key component. | Data Integration Platforms |
| Data repository | A general term referring to data that has been collected, organized, and isolated for business operations or data analysis. It can include databases, data warehouses, and big data stores. | Data Collection and Organization |
| Data warehouse | A central repository that consolidates data from various sources through the Extract, Transform, and Load (ETL) process, making it accessible for analytics and business intelligence. | Data Collection and Organization |
| Document-based Database | A type of NoSQL database that stores each record and its associated data within a single document, allowing flexible indexing, ad hoc queries, and analytics over collections of documents. | NoSQL |
| ETL process | The Extract, Transform, and Load process for data integration involves extracting data from various sources, transforming it into a usable format, and loading it into a repository. | Data Marts, Data Lakes, ETL, and Data Pipelines |
| Graph-based Database | A type of NoSQL database that uses a graphical model to represent and store data, ideal for visualizing, analyzing, and discovering connections between interconnected data points. | NoSQL |
| Key-value store | A type of NoSQL database where data is stored as key-value pairs, with the key serving as a unique identifier and the value containing data, which can be simple or complex. | NoSQL |
| Portability | The capability of data integration tools to be used in various environments, including single-cloud, multi-cloud, or hybrid-cloud scenarios, provides flexibility in deployment options. | Data Integration Platforms |
| Pre-built connectors | Cataloged connectors and adapters that simplify connecting and building integration flows with diverse data sources like databases, flat files, social media, APIs, CRM, and ERP applications. | Data Integration Platforms |
| Relational databases (RDBMSes) | Databases that organize data into a tabular format with rows and columns, following a well-defined structure and schema. | Data Collection and Organization |
| Scalability | The ability of a data repository to grow and expand its capacity to handle increasing data volumes and workload demands over time. | Considerations for Choice of Data Repository |
| Schema | The predefined structure that describes the organization and format of data within a database, indicating the types of data allowed and their relationships. | Considerations for Choice of Data Repository |
| Streaming data | Data that is continuously generated and transmitted in real-time requires specialized handling and processing to capture and analyze. | Considerations for Choice of Data Repository |
| Use cases for relational databases | Applications such as Online Transaction Processing (OLTP), Data Warehouses (OLAP), and IoT solutions where relational databases excel. | Relational Database Management System |
| Vendor lock-in | A situation where a user becomes dependent on a specific vendor’s technologies and solutions, making it challenging to switch to other platforms. | Data Integration Platforms |

## **Summary: Data Literacy for Data Science**

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

* The basics of data collection and organization methods.
* What RDBMS is and its significance.
* NoSQL databases and their flexible schema.
* Types of data storage and the ways to process data.
* The factors influencing data repository selection.
* The various data integration tools and the solutions they provide.