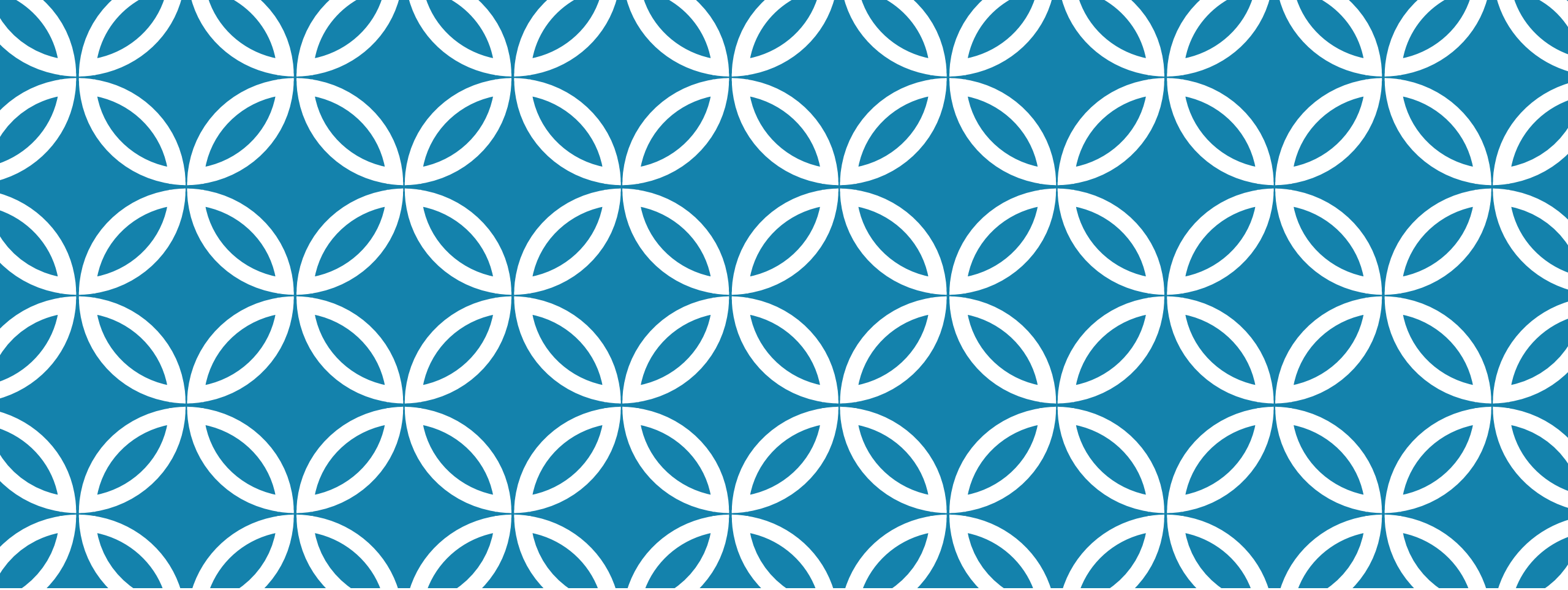


BIG DATA EXTRACTION

Dr. Salahuddin Shaikh

WHAT IS DATA EXTRACTION?

- Data extraction is the process of getting data from a source for further data processing, storage or analysis elsewhere. The term *data collection* is often used when talking about data extraction.
- Data is typically analyzed and then crawled through in order to get any relevant information from the sources (such as database or document). More data processing can also be done to add metadata.
- Data extraction refers to the process of procuring data from a given source and moving it to a new context, either on-site, cloud-based, or a hybrid of both.

DATA EXTRACTION AS EXTRACT/TRANSFORM/LOAD (ETL)

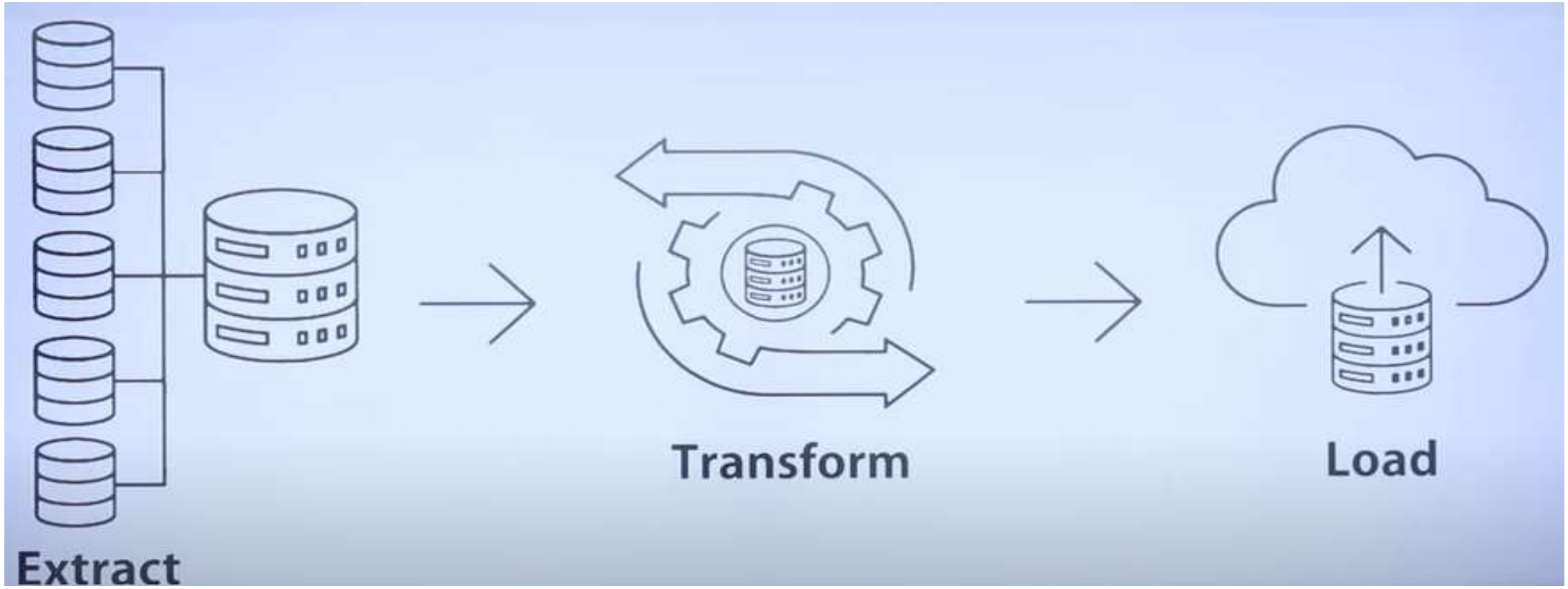
- It is perhaps the most important operation of the Extract/Transform/Load (ETL) process because it is the foundation for critical analyses and decision making processes that are vital to organizations.
- Data extraction is the backbone of ETL (Extract, Transform, Load), a process that drives the data and analytics workflows of many organizations. It’s the most demanding stage of any data-related project, requiring careful planning and execution to ensure a smooth data pipeline. Factors like data sources, extraction methods, and the accuracy of extracted data all play a role in determining the success of the data extraction process.

DATA EXTRACTION AS EXTRACT/TRANSFORM/LOAD (ETL)

The full ETL process lets organizations bring data from different sources into a single location.

- **Extraction** gathers data from one or more sources. The process of extracting data includes locating and identifying the relevant data, then preparing to be transformed and loaded.
- **Transformation** is where data is sorted and organized. Cleansing — such as removing missing values — also happens during this step. Depending on the destination you choose, data transformation could include data typing, JSON structures, object names, and time zones to ensure compatibility with the data destination.
- **Loading** is the last step, where the transformed data is delivered to a central repository for immediate or future analysis.

DATA EXTRACTION AS EXTRACT/TRANSFORM/LOAD (ETL)



WHY DATA EXTRACTION IS IMPORTANT

- To summarize studies in a common format to facilitate synthesis and coherent presentation of data
- To identify numerical data for meta-analyses
- To obtain information to assess more objectively the risk of bias in and applicability of studies
- To identify systematically missing or incorrectly assessed data, outcomes that are never studied, and underrepresented populations

HERE ARE JUST FIVE REASONS YOU SHOULD CONSIDER USING DATA EXTRACTION:

#1 It can improve accuracy & reduce human error

By automating data entry processes for repetitive tasks, data extraction tools can help improve the accuracy of your data inputs by reducing human errors.

If your staff is entering large amounts of data day in and day out, there is a high chance of errors and inaccuracies through human error. By automating the process, you can go some way to removing these errors and get more accurate data overall.

HERE ARE JUST FIVE REASONS YOU SHOULD CONSIDER USING DATA EXTRACTION:

#2 It can increase employee productivity

Removing the need for lots of manual data entry means your staff can spend more time on important tasks that only a human can do. Typically these types of tasks add more value to a business.

Because they are using their skills to complete more meaningful tasks (vs endless data entry) it means employees can feel more satisfied in their job, leading to an increase in productivity.

HERE ARE JUST FIVE REASONS YOU SHOULD CONSIDER USING DATA EXTRACTION:

#3 It can improve visibility

Using data extraction to stay on top of data processing allows your team to get their hands on data faster. This simple process of extracting and storing data means it is more visible to everyone in your business that needs to see it.

What's more, when employees have access to the information they need, there are no delays in waiting on the data being inputted into the system.

HERE ARE JUST FIVE REASONS YOU SHOULD CONSIDER USING DATA EXTRACTION:

#4 It can save you & your business time

Time is money. And wasted time is a big issue for businesses.

Any tool that can improve processes and save time should be explored. When used correctly, data extraction tools can save your business time, giving staff time to concentrate on more important tasks.

HERE ARE JUST FIVE REASONS YOU SHOULD CONSIDER USING DATA EXTRACTION:

#5 It can help reduce costs

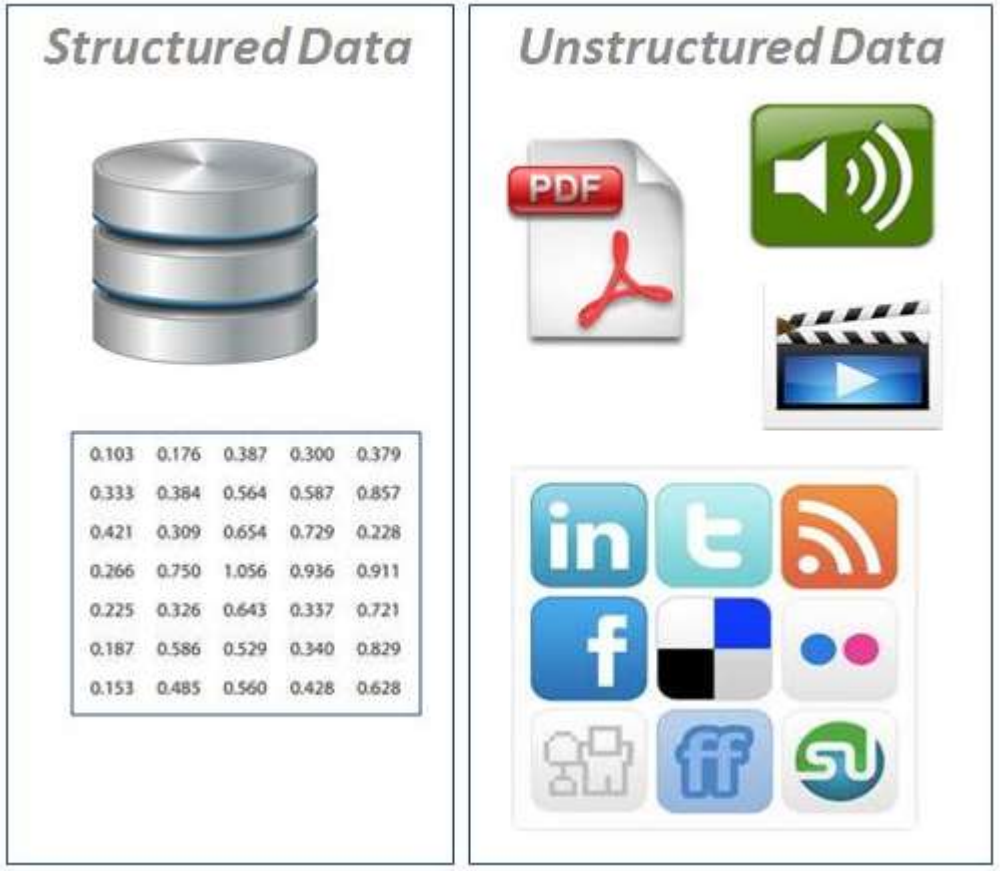
And finally, by automating long and repetitive tasks where possible, businesses can save money in both the short and long term.

In the day to day running of your businesses and as it grows, you don’t need to worry about scaling and investing in a large team to handle your data needs.

SO, HOW IS DATA EXTRACTED?

Now you know what data extraction is and why you should be using it in your business, how exactly does it work?

When it comes to data extractions, there two main types: structured and unstructured.



DATA TYPE

Here are what each type of data means:

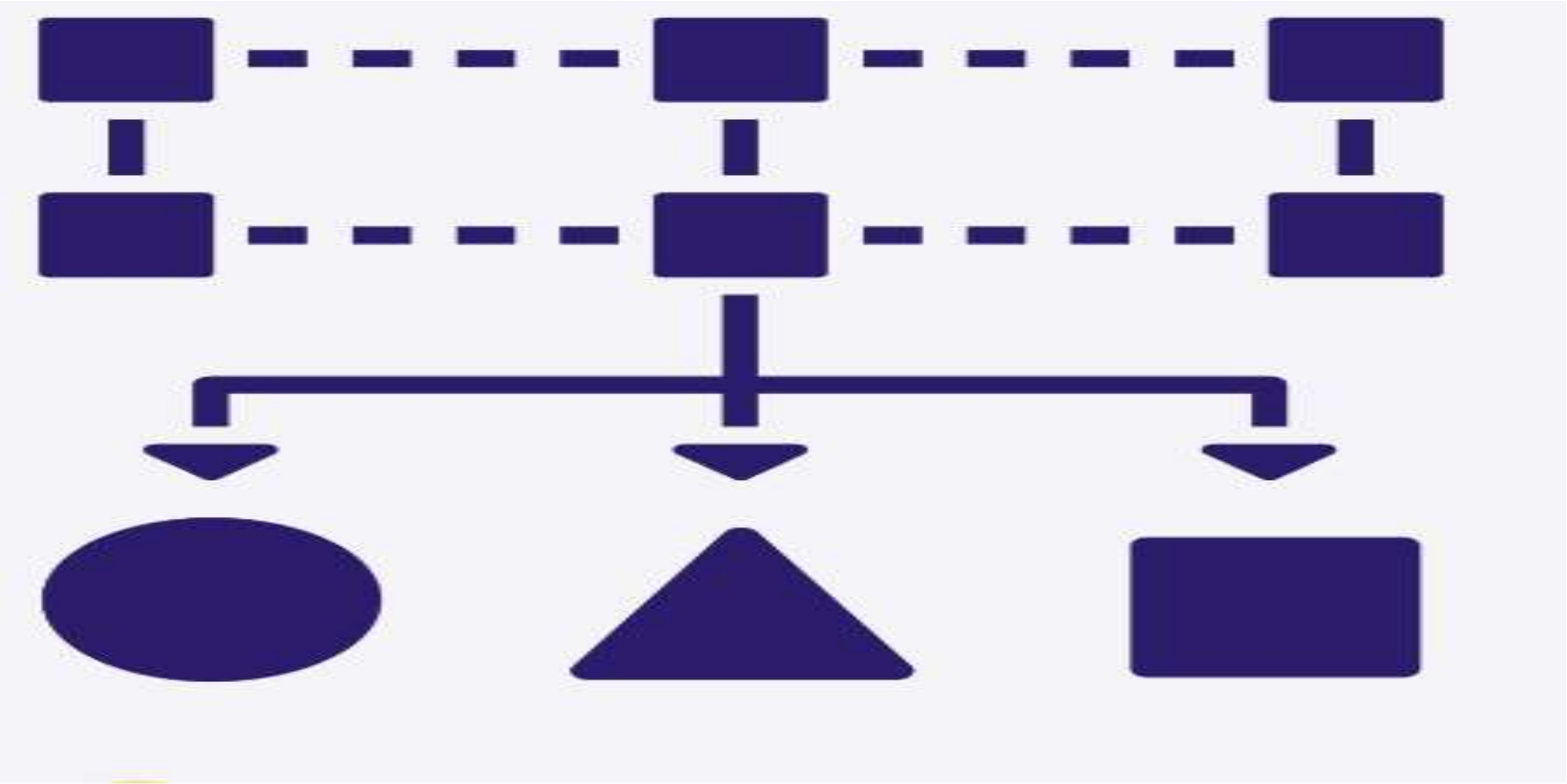
Structured data – when the process is typically performed within the source system. It’s common to utilize full or incremental extraction methods here.

Unstructured data – when you work with unstructured data, a large part of the job task preparing the data. Things like removing whitespace and symbols, removing duplicate results, and deciding how to handle any missing values.

HOW IS DATA EXTRACTED: STRUCTURED & UNSTRUCTURED DATA

- Virtually all data extraction is performed for one of three reasons:
- To archive the data for secure long-term storage.
- For use within a new context (during domain changes for example).
- In order to prepare it for later-stage analysis (the most common reason for extraction).
- Let’s start off by taking a look at how structured data is commonly derived.

HOW IS DATA EXTRACTED: STRUCTURED & UNSTRUCTURED DATA



Data extraction

STRUCTURED DATA EXTRACTION

- Structured data refers to data formatted according to standardized models, making it ready for analysis.
- It can be extracted via a relatively straightforward method known as logical data extraction. Structured data extraction is itself broken down into two subtypes, i.e., full and incremental extraction.

FULL EXTRACTION

As the name might already suggest, this method refers to a single-trip retrieval of data from a given source. It is extracted without any supplements in the form of additional logical information from the system. This is relatively uncomplicated when performed with the right data extraction tools.

That being said, if it is vital to know which changes to the data are continually being made within the source system, the second extraction method is required.

INCREMENTAL EXTRACTION

Extracting incrementally is an ongoing and more complex logical process, as it's not limited to the initial retrieval. Recurring visits to the source system are required in order to monitor for and extract any recent changes the source has made to the data. Determining which changes have occurred while avoiding repeated extraction of the entire data set is where additional logic is required. This is termed Change Data Capture (CDC) and is the preferred practice.

UNSTRUCTURED DATA EXTRACTION

- Without a doubt, extracting unstructured data is more complex than in the case of its structured counterpart. No wonder – the types of data that constitute this group are highly varied. Examples of data sources include web pages, emails, text documents, PDFs, scanned text, mainframe reports, or spool files. However, it's crucial to remember that the information contained within them is no less valuable than that found in structured forms!
- The capacity to extract and process unstructured data is equally as important despite the process's challenging nature. To render the data ready for analysis, further work is required, and it goes beyond mere extraction. Examples of this are removing whitespace, symbols, and duplicate results, or filling in missing values.

CHALLENGES OF DATA EXTRACTION

Even though data extraction is one of the most essential steps in the journey toward data analysis, it is not without its own challenges. Some of these include

Data Volume Management: Your data architecture is designed to handle a specific ingestion volume. If data extraction processes are created for small amounts of data, they may not function properly when dealing with larger quantities. When this happens, parallel extraction solutions may be necessary, but they can be challenging to engineer and maintain.

Data Source/API Constraints: Data sources vary and so do extractable fields. So it's important to consider the limitations of your data sources when extracting data. For instance, some sources like APIs and webhooks may have restrictions on how much data can be extracted at once.

CHALLENGES OF DATA EXTRACTION

Synchronous Extraction: Your extraction scripts must run with precision, taking into account factors such as data latency, volume, source limitations, and validation. The symphony of extraction becomes a complex masterpiece when multiple architectural designs are utilized to cater to different business needs.

Prior Data Validation: Data validation can happen at the extraction stage or the transformation stage. If done during extraction, one should check for any missing or corrupted data, such as empty fields or nonsensical values.

Intensive Data Monitoring: To ensure the proper functioning of your data extraction system, it is important to monitor it on several levels, including resource allocation (e.g. computational power and memory), error detection (e.g. missing or corrupted data), and reliability (e.g. proper execution of extraction scripts)