iTpreneur
Training | Placement | Project

11 Year's of Excellence

Exploring the Data Management (Step by Step Guide)

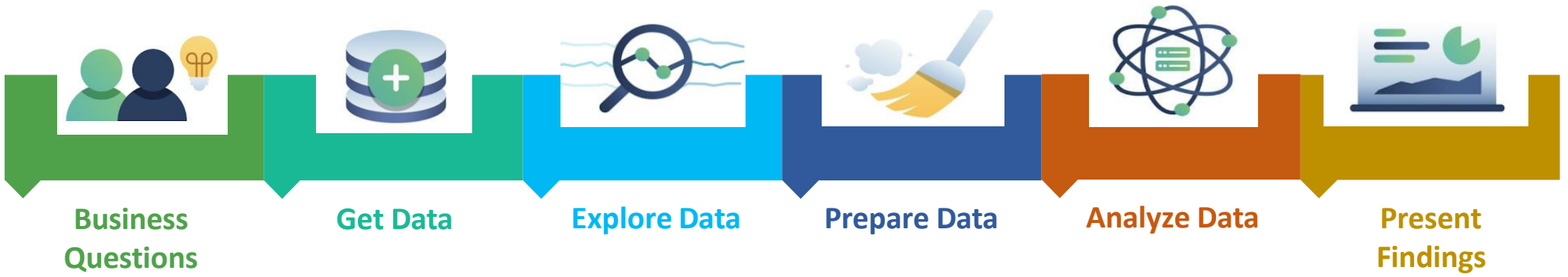
INDEX

TOPICS **PAGE NO**

DATA COLLECTION METHODS (CONVENTIONAL)	**8**
DATA COLLECTION METHODS	**9**
DATA COLLECTION CHALLENGES	**10**
DATA INTEGRATION APPROACHES	**20**
DATA INTEGRATION CONCLUSION & BENEFITS	**26**
DATA ANALYSIS – SOME IMPORTANT CONCEPTS: DATA PROFILING	**32**

Data Analysis – Step by Step Guide

The process of data analysis is a systematic approach that involves several stages, each crucial to ensuring the accuracy and usefulness of the results.



The data analysis process in a nutshell

Step 1: Defining objectives and questions

Define the objectives and formulate clear, specific questions that your analysis aims to answer. This step is crucial as it sets the direction for the entire process. It involves understanding the problem or situation at hand, identifying the data needed to address it, and defining the metrics or indicators to measure the outcomes.

Step 2: Data collection

Once the objectives and questions are defined, the next step is to collect the relevant data. This can be done through various methods such as surveys, interviews, observations, or extracting from existing databases. The data collected can be quantitative (numerical) or qualitative (non-numerical), depending on the nature of the problem and the questions being asked.

Step 3: Data cleaning

Define the objectives and Data cleaning, also known as data cleansing, is a critical step in the data analysis process. It involves checking the data for errors and inconsistencies, and correcting or removing them. This step ensures the quality and reliability of the data, which is crucial for obtaining accurate and meaningful results from the analysis.

Step 4: Data analysis

Once the data is cleaned, it's time for the actual analysis. This involves applying statistical or mathematical techniques to the data to discover patterns, relationships, or trends. There are various tools and software available for this purpose, such as Python, R, Excel, and specialized software like SPSS and SAS.

Step 5: Data interpretation and visualization

After the data is analyzed, the next step is to interpret the results and visualize them in a way that is easy to understand. This could involve creating charts, graphs, or other visual representations of the data. Data visualization helps to make complex data more understandable and provides a clear picture of the findings.

Step 6: Data storytelling

The final step in the data analysis process is data storytelling. This involves presenting the findings of the analysis in a narrative form that is engaging and easy to understand. Data storytelling is crucial for communicating the results to non-technical audiences and for making data-driven decisions.

Data Collection

Data collection is the process of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. It is an essential phase in all types of research, analysis, and decision-making, including that done in the social sciences, business, and healthcare.

Before an analyst begins collecting data, they must answer three questions first:

- What's the goal or purpose of this research?
- What kinds of data are they planning on gathering?
- What methods and procedures will be used to collect, store, and process the information?

Additionally, we can break up data into qualitative and quantitative types. Qualitative data covers descriptions such as color, size, quality, and appearance. Quantitative data, unsurprisingly, deals with numbers, such as statistics, poll numbers, percentages, etc.

Data Collection Methods (Conventional)

Primary Data Collection:

Primary data collection involves the collection of original data directly from the source or through direct interaction with the respondents.

- Surveys and Questionnaires (structured questionnaires or surveys to collect data from individuals or groups)
- Interviews (structured (with predefined questions), semi-structured (allowing flexibility), or unstructured (more conversational)
- Observations (observe and record behaviors, actions, or events in their natural setting)
- Experiments (control the conditions and collect data to draw conclusions about cause-and-effect relationships)
- Focus Groups (Focus groups bring together a small group of individuals who discuss specific topics in a moderated setting)

Secondary Data Collection:

Secondary data collection involves using existing data collected by someone else for a purpose different from the original intent. Researchers analyze and interpret this data to extract relevant information.

- Published Sources (books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data)
- Online Databases (research articles, statistical information, economic data, and social surveys)
- Government and Institutional Records (Database maintained by Government agencies, research institutions)
- Publicly Available Data (Data shared by individuals, organizations, or communities on public platforms, websites, or social media can be accessed and utilized for research)

Data Collection Methods

The methods used to collect data vary based on the type of application. Some involve the use of technology, while others are manual procedures. The following are some common data collection methods:

- Automated data collection functions built into business applications, websites and mobile apps
- Sensors that collect operational data from industrial equipment, vehicles and other machinery
- Collection of data from information services providers and other external data sources.
- Tracking social media, discussion forums, reviews sites, blogs and other online channels

Data Collection Challenges

Data quality issues. Raw data typically includes errors, inconsistencies and other issues. Ideally, data collection measures are designed to avoid or minimize such problems. That isn't foolproof in most cases, though.

Deciding what data to collect. This is a fundamental issue both for upfront collection of raw data and when users gather data for analytics applications. Collecting data that isn't needed adds time, cost and complexity to the process

Low response and other research issues. In research studies, a lack of responses or willing participants raises questions about the validity of the data that's collected.

Dealing with big data. Big data environments typically include a combination of structured, unstructured and semi structured data, in large volumes. That makes the initial data collection and processing stages more complex.

Finding relevant data. With a wide range of systems to navigate, gathering data to analyze can be a complicated task for data scientists and other users in an organization. The use of data curation techniques helps make it easier to find and access data. For example, that might include creating a data catalog and searchable indexes.

Data Cleaning

Data cleaning is the process of identifying and correcting errors and inconsistencies in data sets so that they can be used for analysis. In doing so, data professionals can get a clearer picture of what is happening within their businesses, deliver trustworthy analytics any user can leverage, and help their organizations operate more efficiently.

The more accurate your data set, the more accurate your insights will be. And as research from Harvard Business Review points out, when it comes to making business decisions, whether by executives or frontline decision makers, every insight matters. That's why data cleaning should be at the top of your list of priorities if you want to get the most out of your data. In this post, we will discuss the top five benefits of cleaning your data, real-life data cleaning examples, and seven steps to follow to clean your data properly.

Data Cleaning - Steps

Step 5: Check data integrity

Data professionals should then check for data integrity by ensuring that all data is accurate, valid, and up-to-date before proceeding to data analysis or data visualization. This is done by running data integrity checks or data validation tests on the data.

Step 6: Store data securely

Then, data professionals must store data securely in order to protect it from unauthorized access and data loss. This includes encrypting data at rest, using secure file transfer protocols for data transmissions, and regularly backing up data sets.

Step 7: Expose data to business experts

Finally, the last step is exposing data to business users. These domain experts have deep knowledge, and can quickly help identify data that's inaccurate or out of date. This mutual partnership between data and business teams requires the right self-service business intelligence solution, so business users can focus on exploring data to find data cleanliness issues.

Data Cleaning - Steps

Step 1: Identify data discrepancies using data observability tools

At the initial phase, data analysts should use data observability tools such as **Monte Carlo** or **Anomalo** to look for any data quality issues, such as data that is duplicated, missing data points, data entries with incorrect values, or mismatched data types.

Step 2: Remove data discrepancies

Once the data discrepancies have been identified and appropriately evaluated, data analysts can then go about removing them from the existing dataset. This may involve removing data entries or data points that are irrelevant, merging data sets together, and ensuring data accuracy.

Step 3: Standardize data formats

After data discrepancies have been removed, standardizing data formats is essential in order to ensure consistency throughout the dataset. For example, one data set may contain dates formatted differently than another data set. Data analysts should ensure that all data is stored in the same format, such as YYYY/MM/DD or MM/DD/YYYY, across all data sets.

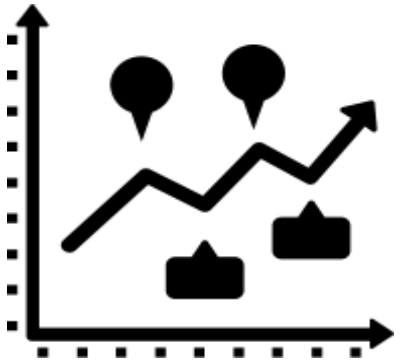
Step 4: Consolidate data sets

Then, different data sets can be consolidated into a single data set unless data privacy laws prevent them from doing so. Often, this requires breaking down silos between datasets and bringing them together. Many organizations rely on emerging data architectures, whether they're using or considering a data lake, data warehouse to do so. Consolidating data sets makes data analysis more efficient as it reduces data redundancy and streamlines the data processing process.

Data Analysis

Descriptive

What happened?



Summarizing and describing the primary properties of a dataset. It provides vital insights into the data's frequency distribution, central tendency, dispersion, and identifying position. It assists researchers and analysts in better understanding their data.

Diagnostic

Why did it happen?

It examines data to understand the root causes of events, behaviors, and outcomes. Uses diverse techniques and tools to identify patterns, trends, and connections to explain why certain events occurred.

Predictive

What is likely to happen?

Predictive analytics looks at current and historical data patterns to determine if those patterns are likely to emerge again. This allows businesses and investors to adjust where they use their resources to take advantage of possible future events.

Prescriptive

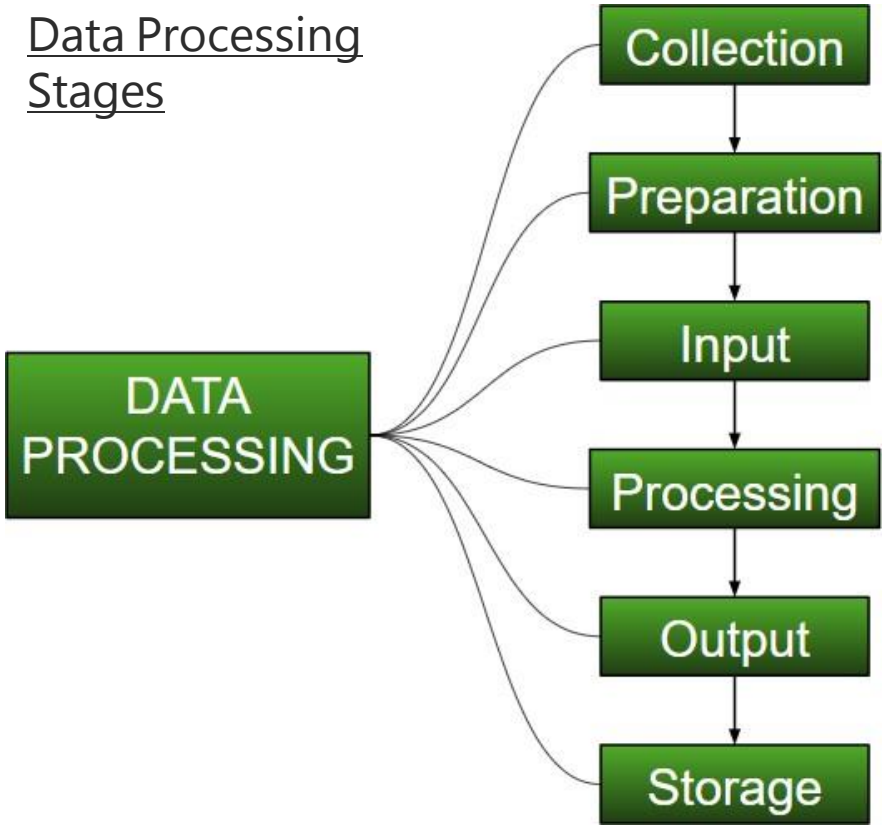
What is the best course of action?

Prescriptive analytics is a process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted outcomes. In essence, prescriptive analytics takes the "what we know" (data).

Data Processing

Data processing converts data into a valuable and acceptable format according to a predetermined sequence of actions. You can do this conversion either manually or mechanically. You can display the processed data in different formats and some examples include images, graphs, tables and charts. The program or data processing method you use can determine the layout of your data.

Data Processing Stages



The initial stage in data processing is the collection and compilation of data

Arranging data involves extracting and structuring the raw data to prepare it for the next step and eliminate redundant, incomplete or incorrect data

Transforming raw data into usable information and transfer the clean data to its intended destination

The computer processes the data entered during the previous stage for interpretation. Although the procedure may differ slightly based on the data source and the intended purpose, machine learning algorithms often complete this stage.

The data output stage is where non-data analysts can use and interpret the data. These professionals often transform data into understandable data that frequently comprise graphs, videos, photos, texts and other formats.

After analyzing and processing all the data, you can save it for future reference.

Data Processing - Forms of processed data output

Manual Processing

This method of data processing involves manually inputting information into a computer or physical document. The manual processing method also involves physically completing calculations and logical operations on the collected data. This method of data processing is useful for small or simple data sets.

Mechanical Processing

People perform this with the aid of a mechanical device or electronic equipment. Some mechanical processing tools include calculators and typewriters, which can enhance manual data processing techniques. This method of data processing is useful with straightforward datasets that require minimal adjustments.

Electronic Processing

This is a recent method of data processing. Electronic data processing is a quick and reliable method that improves data accuracy and efficiency. Many institutions and businesses use this method because of its ability to improve productivity. In electronic processing, the computer system follows a detailed set of instructions to automatically input and adjust data. This method can also generate data visualization results, which can expedite company decision-making processes.

Data Processing - Data Output File Types

Once your data processing software has sorted through raw data, processed it and made it into a format that is palatable for you, the next step is data output. There are several different data output file types that will help you better understand and be able to present your data in a usable way:

Text: Used to tell a story for data.

Chart: Used to show trends in data such as growth or decline.

Table: This is used to present mainly statistical data using rows and columns.

Image: Maps, vectors and other images can be used here to demonstrate particular insights from the data retrieved.

Data Processing - Methods

Batch processing: Batch processing is processing many data at once. This method aids in completing tasks such as payroll and month-end reconciliation.

Real-time processing: Real-time processing, also called stream processing, involves processing data in a short time to provide immediate and accurate results. Examples of real-time data processing systems include bank ATMs, traffic control systems and modern computer systems.

Online processing: Online processing continuously enters and processes data as long as the primary sources are available. An example of online processing of data is bar code scanning.

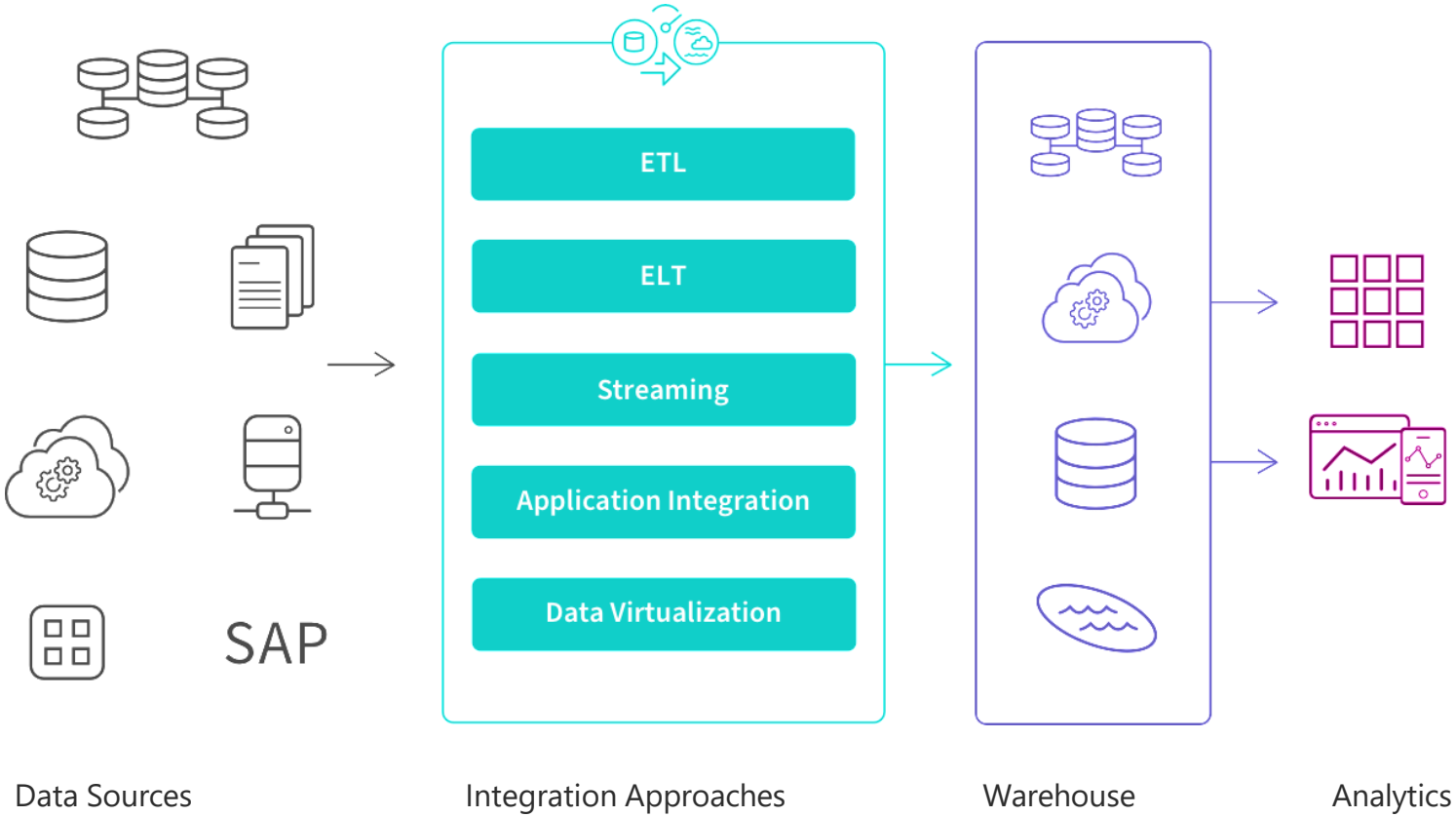
Multiprocessing: Multiprocessing is the ability of a computer system to support several processes and programs simultaneously. Multiprocessing operating systems allow multiple programs to run simultaneously.

Data Integration

Data integration refers to the process of bringing together data from multiple sources across an organization to provide a complete, accurate, and up-to-date dataset for BI, data analysis and other applications and business processes. It includes data replication, ingestion and transformation to combine different types of data into standardized formats to be stored in a target repository such as a data warehouse, data lake or data lakehouse.

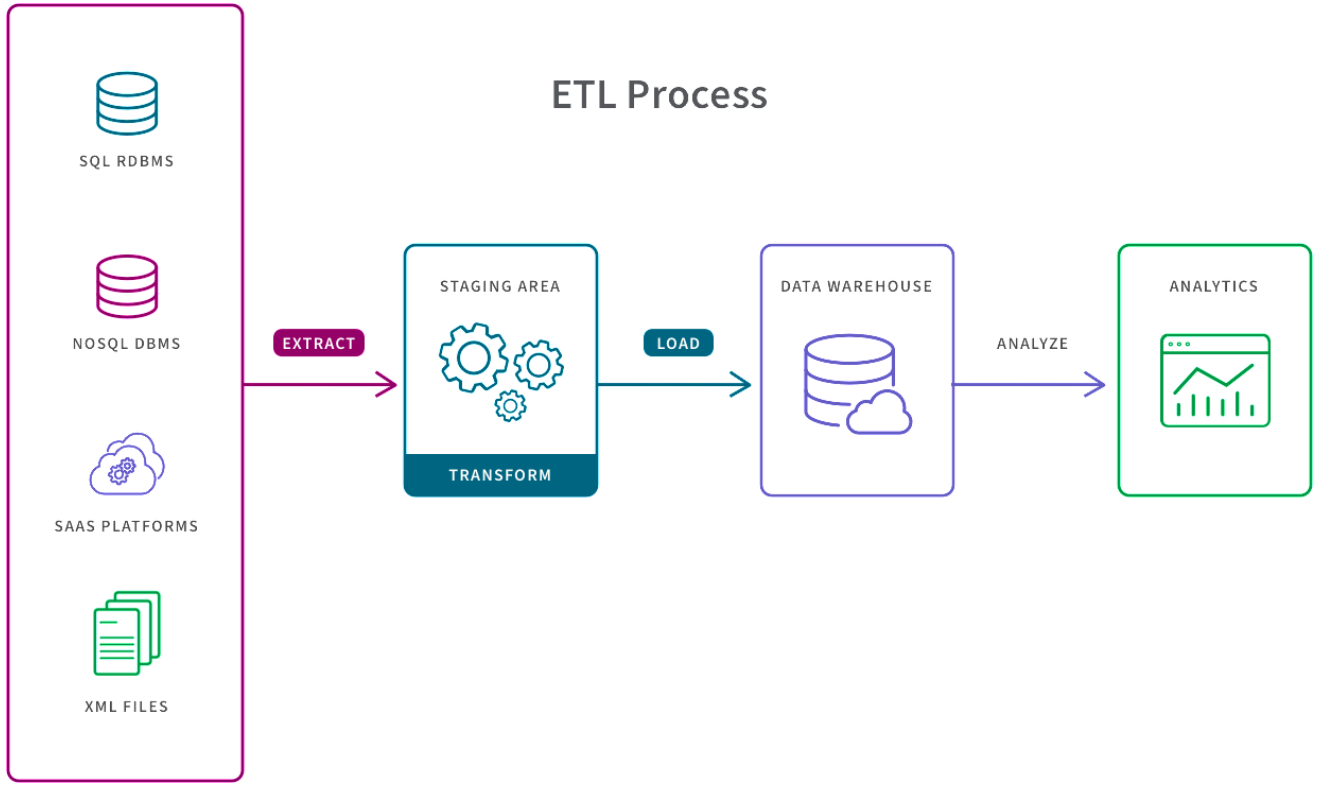
Data Integration Approaches

There are five different approaches, or patterns, to execute data integration: ETL, ELT, streaming, application integration (API) and data virtualization. To implement these processes, data engineers, architects and developers can either manually code an architecture using SQL or, more often, they set up and manage a data integration tool, which streamlines development and automates the system.



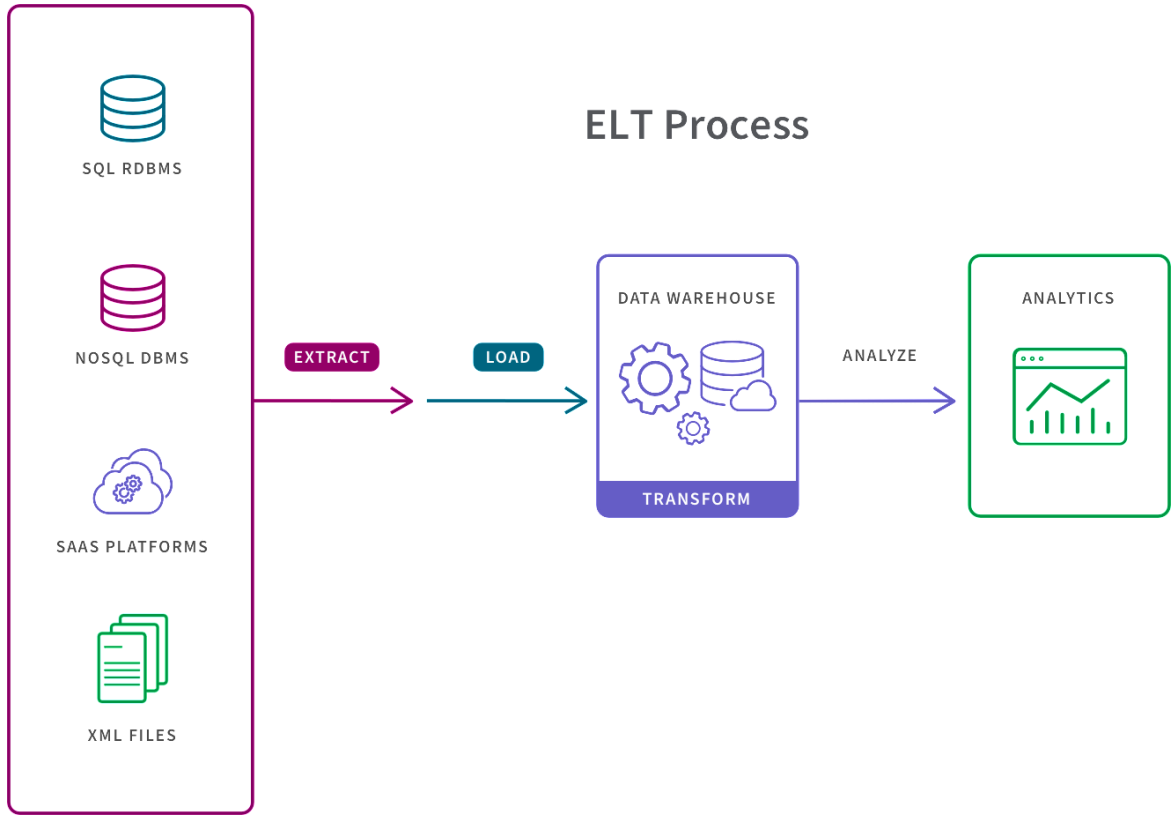
Data Integration Approach - ETL

1. ETL - An ETL pipeline is a traditional type of data pipeline which converts raw data to match the target system via three steps: extract, transform and load. Data is transformed in a staging area before it is loaded into the target repository (typically a data warehouse). This allows for fast and accurate data analysis in the target system and is most appropriate for small datasets which require complex transformations.



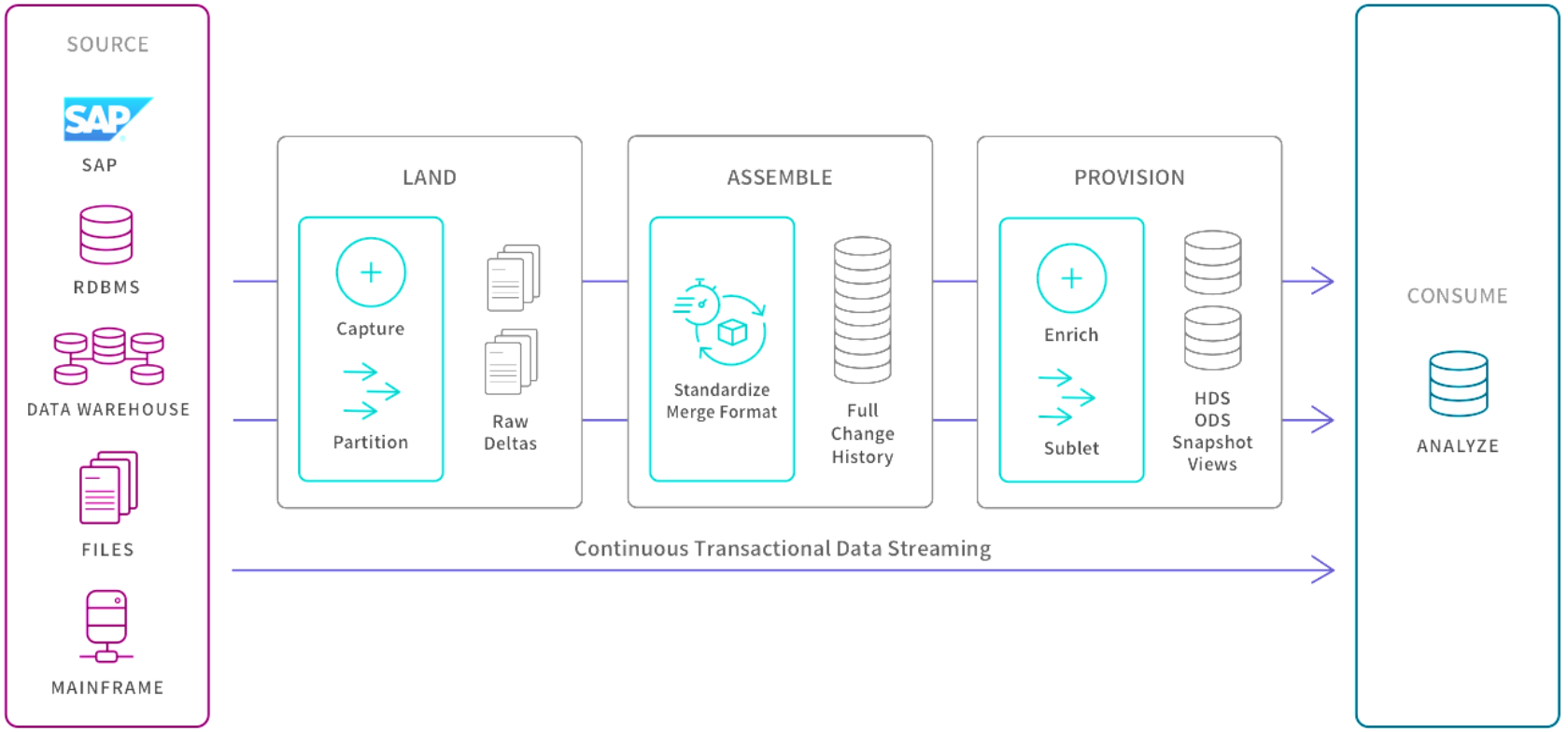
Data Integration Approach - ELT

2. ELT - In the more modern ELT pipeline, the data is immediately loaded and then transformed within the target system, typically a cloud-based data lake, data warehouse or data lakehouse. This approach is more appropriate when datasets are large and timeliness is important, since loading is often quicker. ELT operates either on a micro-batch or change data capture (CDC) timescale. Micro-batch, or "delta load", only loads the data modified since the last successful load. CDC on the other hand continually loads data as and when it changes on the source.



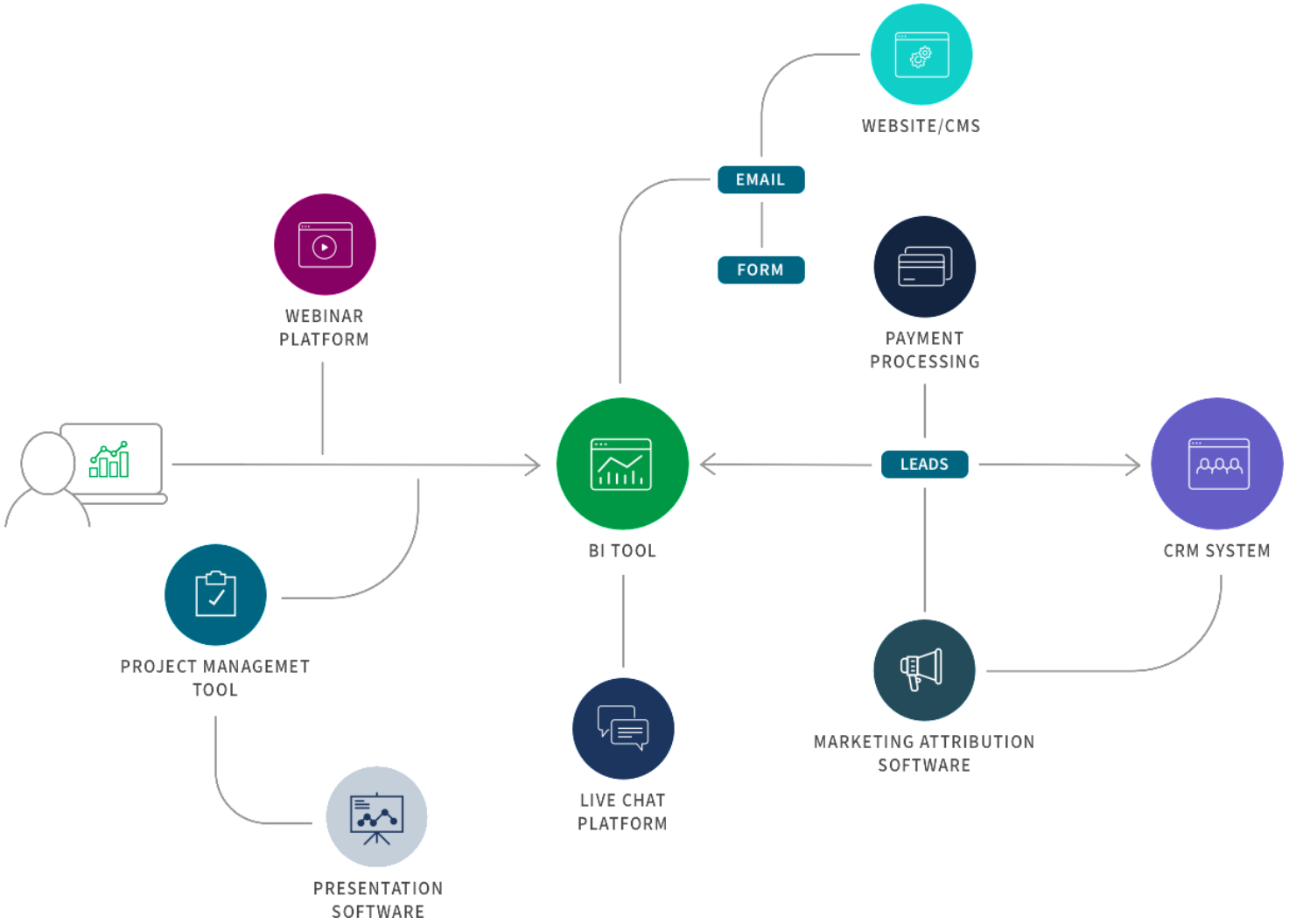
Data Integration Approach – Data Streaming

3. Data Streaming - Instead of loading data into a new repository in batches, streaming data integration moves data continuously in real-time from source to target. Modern data integration (DI) platforms can deliver analytics-ready data into streaming and cloud platforms, data warehouses, and data lakes.



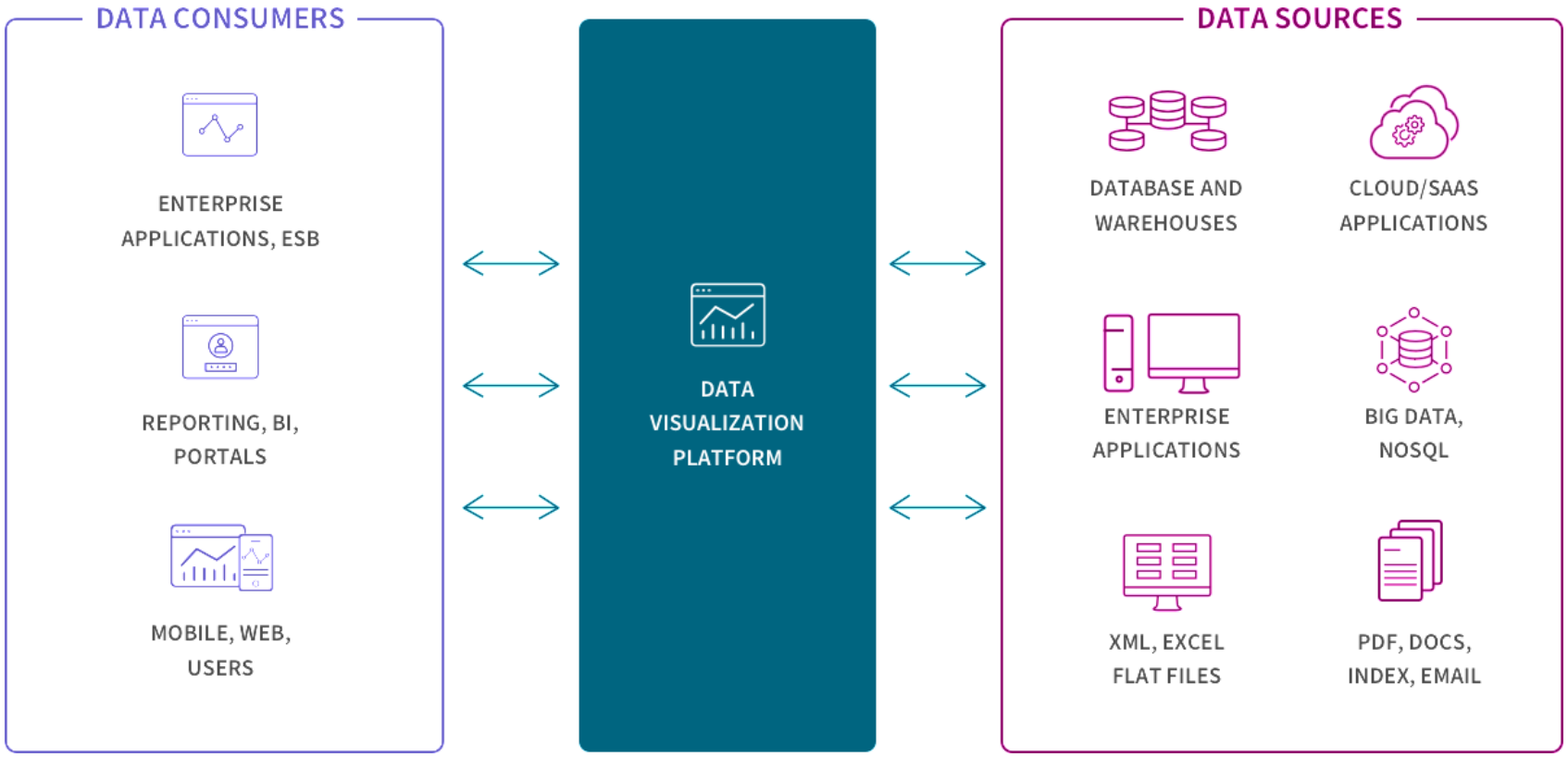
Data Integration Approach – Application Integration

4. Application integration - Application integration (API) allows separate applications to work together by moving and syncing data between them. The most typical use case is to support operational needs such as ensuring that your HR system has the same data as your finance system. Therefore, the application integration must provide consistency between the data sets. Also, these various applications usually have unique APIs for giving and taking data so SaaS application automation tools can help you create and maintain native API integrations efficiently and at scale. Here is an example of a B2B marketing integration flow:



Data Integration Approach – Data Virtualization

5. Data Virtualization - Like streaming, data virtualization also delivers data in real time, but only when it is requested by a user or application. Still, this can create a unified view of data and makes data available on demand by virtually combining data from different systems. Virtualization and streaming are well suited for transactional systems built for high performance queries.



Data Integration Conclusion & Benefits

Conclusion

Data Integration is Continually evolving

Each of these five approaches continue to evolve with the surrounding ecosystem. Historically, data warehouses were the target repositories and therefore data had to be transformed before loading. This is the classic ETL data pipeline (Extract > Transform > Load) and it's still appropriate for small datasets which require complex transformations.

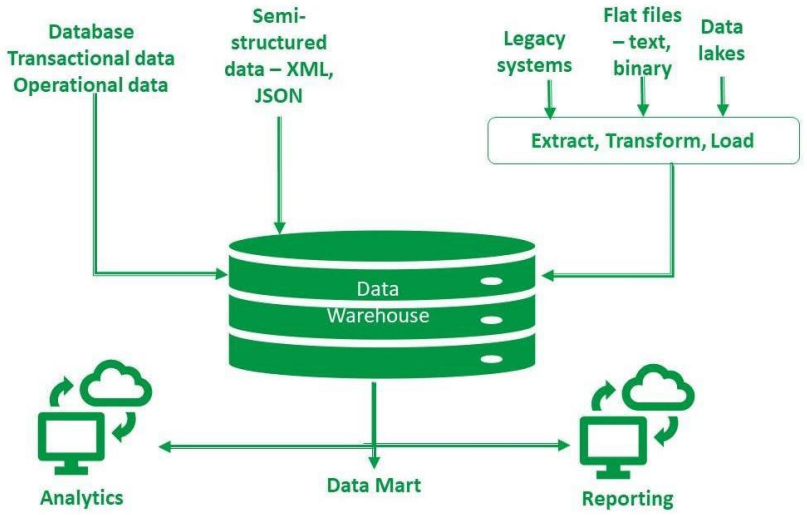
However, with the rise of Integration Platform as a Service (iPaaS) solutions, larger datasets, data fabric and data mesh architectures, and the need to support real-time analytics and machine learning projects, data integration is shifting from ETL to ELT, streaming and API.

Data Integration Benefits

1. Solves Legacy System Integration
2. Solves Unstructured Data Issues
3. Filters Duplicate Data
4. Handles Poor Quality Data
5. Provides Data Governance
6. Improves Performance Integration
7. Reduces Time of Integration
8. Reduce Dependent Resources
9. Increases Business Efficiency
10. Provides a Future Perspective

Data Integration & Storage Concepts – Data Warehouse

A data warehouse is a repository created for analytics and reporting purposes. It usually works on a structured storage (schema-on-write), unlike data lakes. Data warehouses primarily store past and current structured or semi-structured data, which is internal to the organization and available in standard format. Unstructured data (like that from the internet) should be processed and formatted with an ETL step before being ingested into a data warehouse. This makes the data consistent and of high quality—and, therefore, ready for analysis. You can say that a data warehouse is an analytical database used for business intelligence. The schema-based format makes data analysis easier.

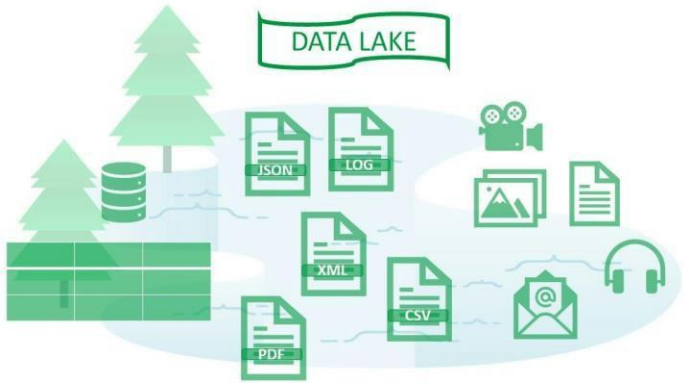


*Data warehouse can store both current and historical data in one place and is designed to give a long-range view of data over time, making it a primary component of business intelligence.

Data warehouses can be on-premise and cloud-based. Cloud data warehouses reduce the cost, deployment process, and infrastructure needs, and can automatically scale based on application needs.
A data mart is a subset of a data warehouse that stores operational data of a particular niche or line of business.

Data Integration & Storage Concepts – Data Lake

A data lake is a central storage repository that stores data in its native format. It uses flat architecture to store data, usually as object or file storage. Data lakes are vast and store any amount of unstructured, structured, or semi-structured big data. They work on the schema-on-read principle (i.e., do not have a predefined schema).



The data sources can be IoT devices, streaming data, web applications, and many others. Some of the data ingested might be filtered and ready to use as well — the kind of flexibility impossible with relational databases.

Since data lakes are configured on commodity hardware and clusters, they are highly scalable and inexpensive.
Data lakes can be configured on-premise or in the cloud. Again, on-premise data lakes are suitable for highly sensitive and secure data. However, having a cloud data lake reduces the cost of infrastructure and is easier to scale out.

Data Integration & Storage Concepts – Data Pipelines

A data pipeline is a method in which raw data is ingested from various data sources and then ported to data store, like a data lake or data warehouse, for analysis. Before data flows into a data repository, it usually undergoes some data processing.

There are two main types of data pipelines, which are batch processing and streaming data;

1. Batch Processing
2. Data Streaming

Data Pipeline Architecture

Data Ingestion: Data is collected from various data sources, which includes various data structures (i.e. structured and unstructured data). Within streaming data, these raw data sources are typically known as producers, publishers, or senders. While businesses can choose to extract data only when they are ready to process it, it's a better practice to land the raw data within a cloud data warehouse provider first. This way, the business can update any historical data if they need to make adjustments to data processing jobs.

Data Transformation: During this step, a series of jobs are executed to process data into the format required by the destination data repository. These jobs embed automation and governance for repetitive workstreams, like business reporting, ensuring that data is cleansed and transformed consistently. For example, a data stream may come in a nested JSON format, and the data transformation stage will aim to unroll that JSON to extract the key fields for analysis.

Data Storage: The transformed data is then stored within a data repository, where it can be exposed to various stakeholders. Within streaming data, this transformed data are typically known as consumers, subscribers, or recipients.

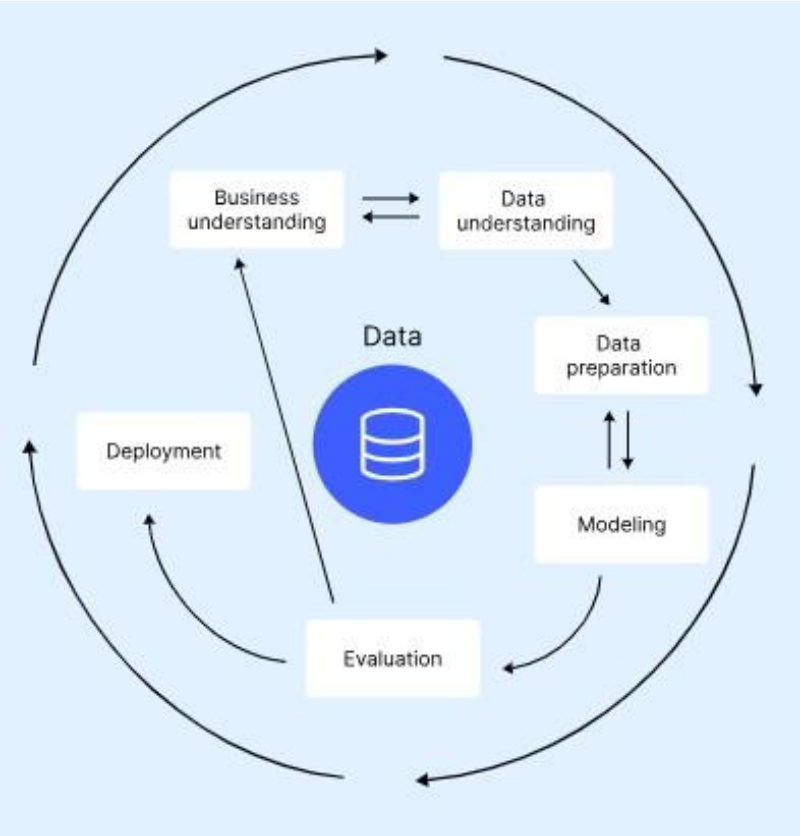
Data Analysis – Some Important Concepts: Data Mining

Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.

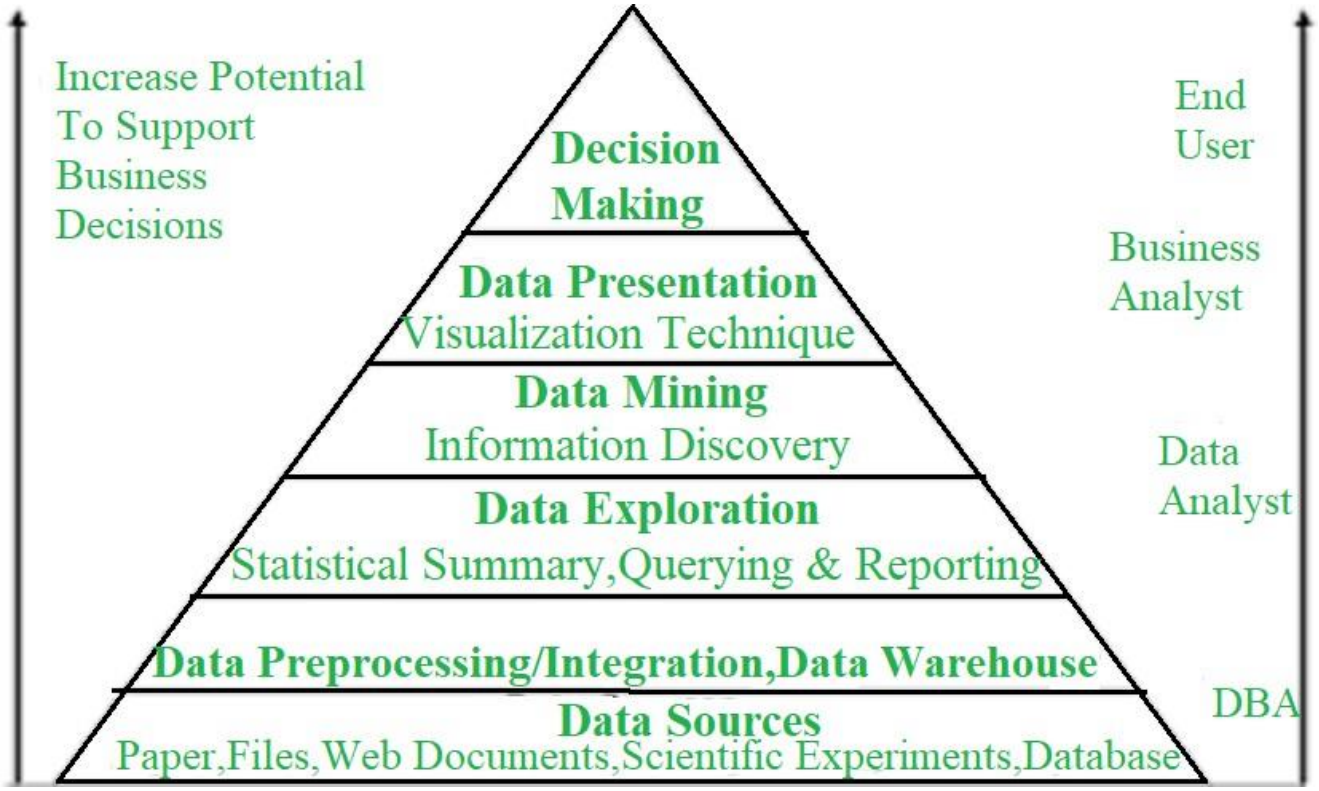
Companies use data mining software to learn more about their customers. It can help them to develop more effective marketing strategies, increase sales, and decrease costs. Data mining relies on effective data collection, warehousing, and computer processing.

Data mining is the process of analyzing a large batch of information to discern trends and patterns.
Data mining can be used by corporations for everything from learning about what customers are interested in or want to buy to fraud detection and spam filtering.

Data mining programs break down patterns and connections in data based on what information users request or provide.



Data Analysis – Some Important Concepts: Data Mining

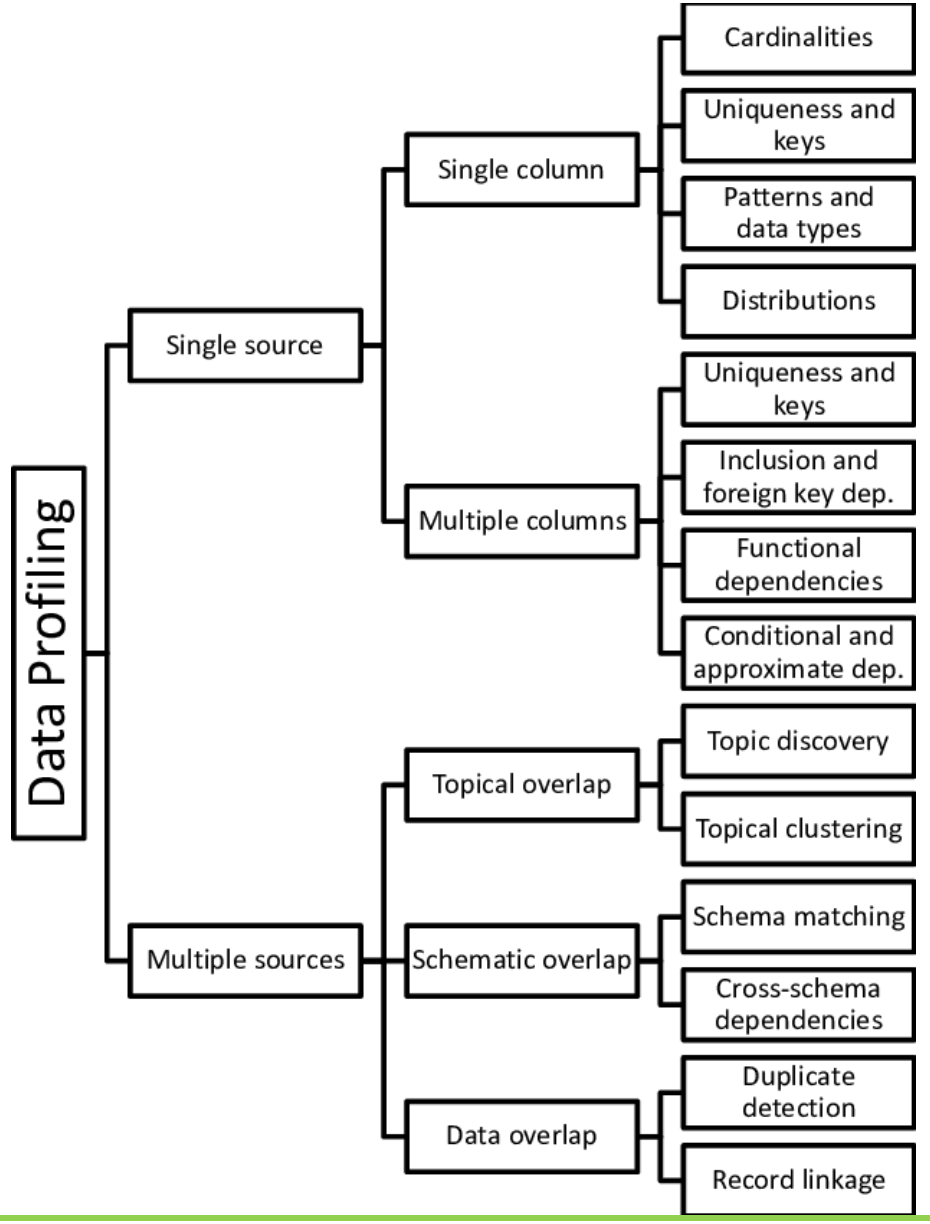


Data Analysis – Some Important Concepts: Data Profiling

Data profiling is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects.

Data profiling involves:

- Collecting descriptive statistics like min, max, count and sum.
- Collecting data types, length and recurring patterns.
- Tagging data with keywords, descriptions or categories.
- Performing data quality assessment, risk of performing joins on the data.
- Discovering metadata and assessing its accuracy.
- Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded value dependencies, and performing inter-table analysis.



Data Analysis – Some Important Concepts: Data Profiling

Types of Data Profiling

- **Structure discovery** — Structure discovery (or analysis) helps determine whether your data is consistent and formatted correctly. It uses basic statistics to provide information about the validity of data.

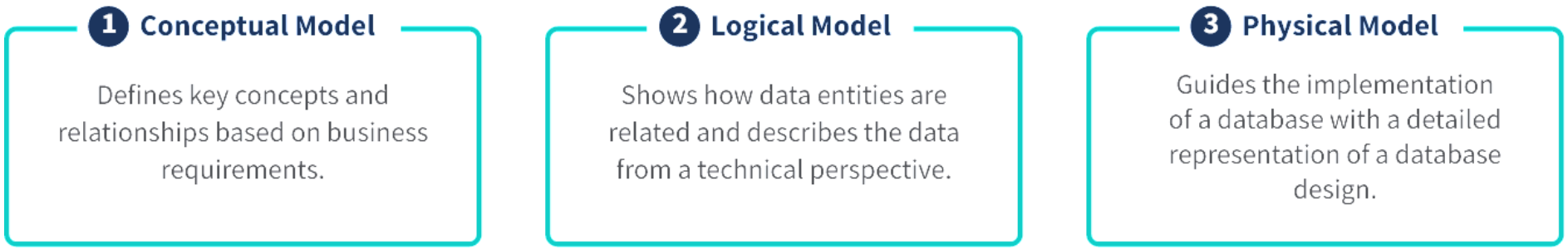
- **Content discovery** — Content discovery focuses on data quality. Data needs to be processed for formatting and standardization, and then properly integrated with existing data in a timely and efficient manner. For example, if a street address or phone number is incorrectly formatted it could mean that certain customers can't be reached, or a delivery is misplaced.

- **Relationship discovery** — Relationship discovery identifies connections between different datasets.

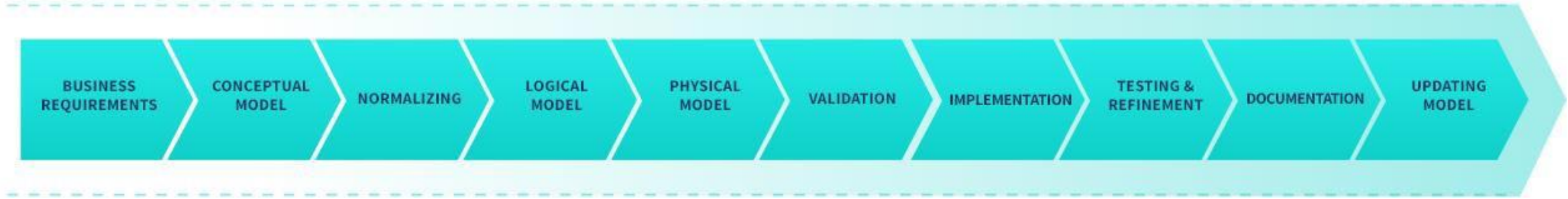
Data Analysis – Some Important Concepts: Data Modelling

Data modeling is the process of creating a diagram that represents your data system and defines the structure, attributes, and relationships of your data entities. Data modeling organizes and simplifies your data in a way that makes it easy to understand, manage, and query, while also ensuring data integrity and consistency. Data models inform your data architecture, database design, and restructuring legacy systems.

Types of Data Modelling -



Data Analysis – Some Important Concepts: Data Modelling



Data Analysis – Some Important Concepts: Data Wrangling

Data wrangling is the process of transforming and structuring data from one raw form into a desired format with the intent of improving data quality and making it more consumable and useful for analytics or machine learning.

The data wrangling process often includes transforming, cleansing, and enriching data from multiple sources. As a result of data wrangling, the data being analyzed is more accurate and meaningful, leading to better solutions, decisions, and outcomes.
Because of the increase in data collection and usage, especially diverse and unstructured data from multiple data sources, organizations are now dealing with larger amounts of raw data and preparing it for analysis can be time-consuming and costly.

Functioning –

Explore: Data exploration or discovery is a way to identify patterns, trends, and missing or incomplete information in a dataset. The bulk of exploration happens before creating reports, data visualizations, or training models, but it's common to uncover surprises and insights in a dataset during analysis too.

Cleanse: Data often contains errors as a result of manual entry, incomplete data, data automatically collected from sensors, or even malfunctioning equipment. Data cleansing corrects those entry errors, removes duplicates and outliers (if appropriate), eliminates missing data, and imputes null values based on statistical or conditional modeling to improve data quality.

Data Analysis – Some Important Concepts: Data Wrangling

Transform: Data transformation or data structuring is important; if not done early on, it can compromise the rest of the wrangling process. Data transformation involves putting the raw data in the right shape and format that will be useful for a report, data visualization, or analytic or modeling process. It may involve creating new variables (aka features) and performing mathematical functions on the data.

Enrich: Enrichment or blending makes a dataset more useful by integrating additional sources such as authoritative third-party census, firmographic, or demographic data. The enrichment process may also help uncover additional insights from the data within an organization or spark new ideas for capturing and storing additional customer information in the future. This is an opportunity to think strategically about what additional data might contribute to a report, model, or business process.

Validate: Validation rules are repetitive programming sequences that verify data consistency, quality, and security. Examples of validation include ensuring uniform distribution of attributes that should be distributed normally (e.g. birth dates) or confirming accuracy of fields through a check across data. This is a vital step in the data wrangling process.

Store: The last part of the wrangling process is to store or preserve the final product, along with all the steps and transformations that took place so it can be audited, understood, and repeated in the future.