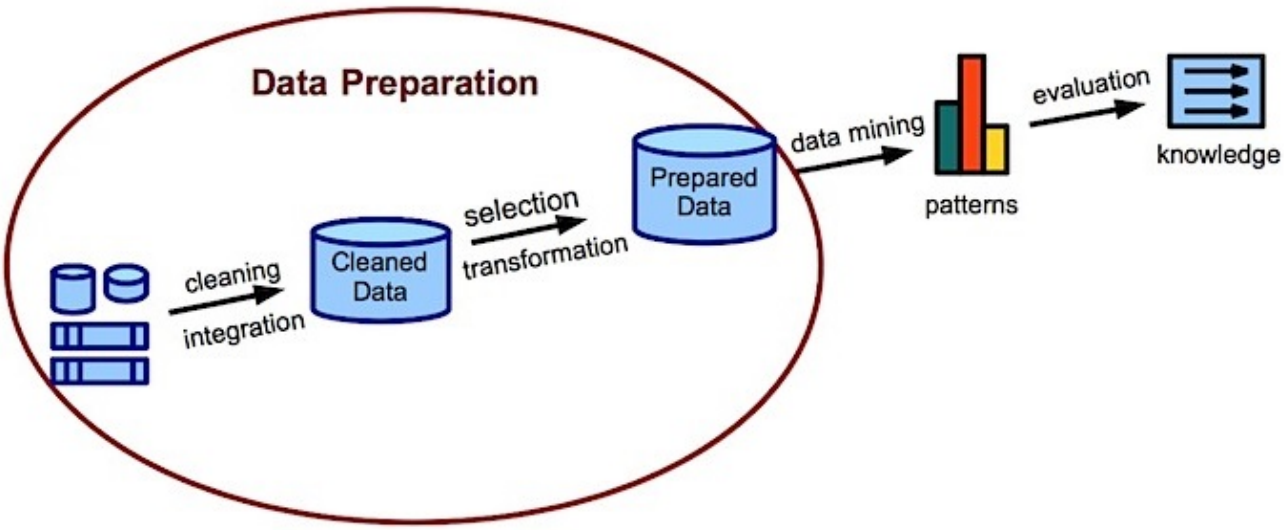
Data Preparation

Good **data preparation** allows for efficient analysis, limits errors and inaccuracies that can occur to data during processing, and makes all processed data more accessible to users. It's also gotten easier with new tools that enable any user to cleanse and qualify data on their own.

1. What is Data Preparation?

Data preparation is the process of **cleaning** and **transforming** raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.



Data preparation is often a lengthy undertaking for data professionals or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.

For example, the data preparation process usually includes standardizing data formats, enriching source data, and / or removing outliers.

2. Benefits of Data Preparation

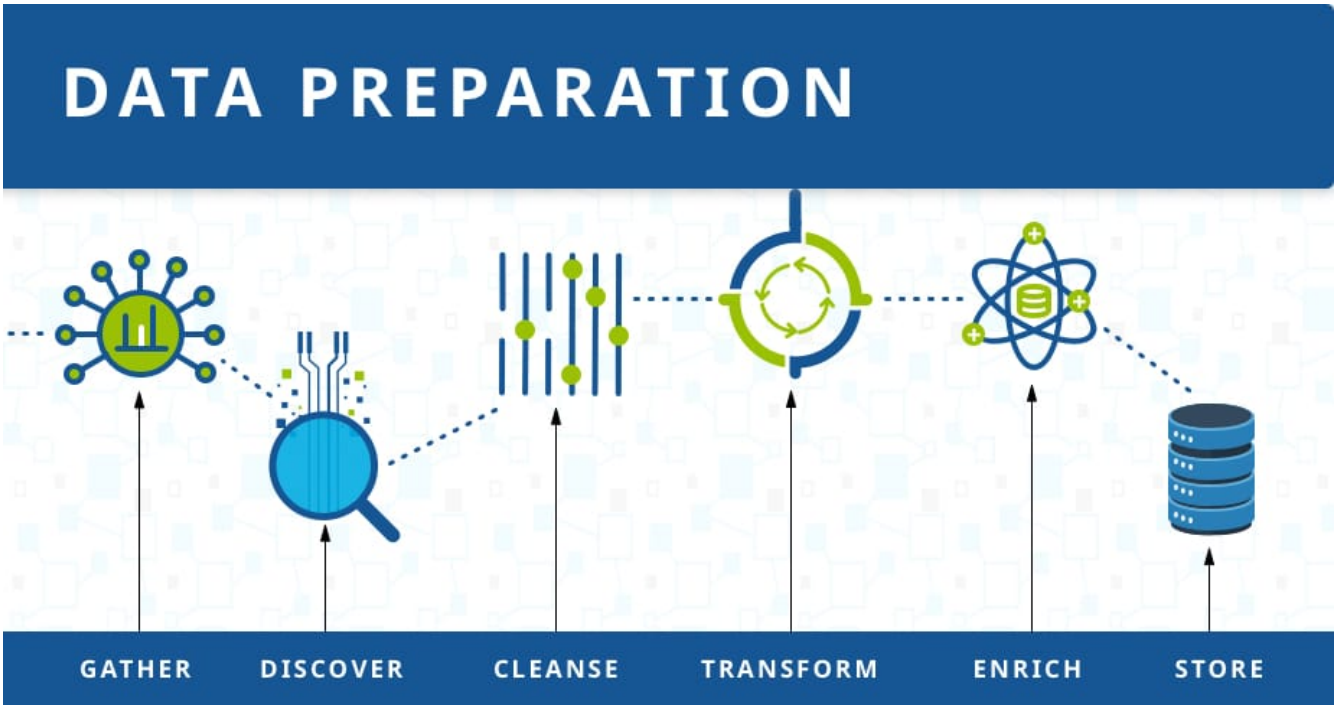
76% of data scientists say that data preparation is the worst part of their job, but the efficient, accurate business decisions can only be made with clean data.

Data preparation helps in:

- **Fix errors quickly** — Data preparation helps catch errors before processing. After data has been removed from its original source, these errors become more difficult to understand and correct.
- **Produce top-quality data** — Cleaning and reformatting datasets ensures that all data used in analysis will be high quality.
- **Make better business decisions** — Higher quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient and high-quality business decisions.

3. Data Preparation Steps

The specifics of the data preparation process vary by industry, organization and need, but the framework remains largely the same.



3.1 Gather Data

The data preparation process begins with **finding the right data**. This can come from an existing data catalog or can be added ad-hoc.

3.2 Discover and Assess Data

After collecting the data, it is important to discover each dataset. This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context.

3.3 Cleanse and Validate Data

Cleaning up the data is traditionally the most time consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps. Important tasks here include:

- **Removing extraneous data and outliers.**
- **Filling in missing values.**
- **Conforming data to a standardized pattern.**
- **Masking private or sensitive data entries.**

Once data has been cleansed, it must be validated by testing for errors in the data preparation process up to this point. Often times, an error in the system will become apparent during this step and will need to be resolved before moving forward.

3.4 Transform and Enrich Data

Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience.

***Enriching* data** refers to adding and connecting data with other related information to provide deeper insights.

3.5 Store Data

Once prepared, the data can be **stored** or channeled into a third party application such as a business intelligence tool clearing the way for processing and analysis to take place.

4. The Future of Data Preparation

Initially focused on analytics, data preparation has evolved to address a much broader set of use cases and can be used by a larger range of users. Although it improves the personal productivity of whoever uses it, it has evolved into an enterprise tool that fosters collaboration between IT professionals, data experts, and business users.

5. Conclusion

Data preparation creates higher quality data for analysis and other data management related tasks by eradicating errors and normalizing raw data before it is processed. It is critical, but takes a lot of time and might require specific skills. The well known saying **"garbage-in garbage-out"** is very relevant to this domain.