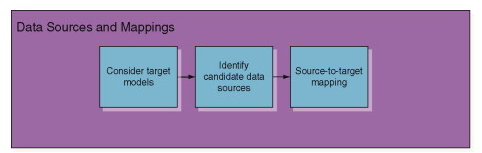
**Establishing Usability of Candidate Data Sources:**

In The Data Requirements Analysis Process, identified the data requirements, candidate data sources are determined and their quality is assessed using the data quality assessment process

Those responsible for accessing, extracting, and preparing data for loading into an analytical platform must understand the particulars about the data in the source systems prior to its extraction and integration. This involves more than trusting the supplied documentation, data dictionaries, or printouts of data models

**Identify candidate data sources:** Consult the data management teams to review the candidate data sources containing the identified data elements, and review the collection of data facts needed by the consuming applications. For each fact, determine whether it corresponds to a defined data concept or data element, exists in any data sets in the organization, or is a computed value (and if so, what are the data elements that are used to compute that value), and then document each potential data source.

**Develop source-to-target mappings:** Because this analysis should provide enough input to specify which candidate data sources can be extracted, the next step is to consider how that data is to be transformed into a common representation that is then normalized in preparation for consolidation. The consolidation processes collect the sets of objects and prepare them for populating the consuming applications.



### What is data profiling?

Data profiling refers to the process of examining, analyzing, reviewing and summarizing data sets to gain insight into the quality of data. [Data quality](https://searchdatamanagement.techtarget.com/definition/data-quality) is a measure of the condition of data based on factors such as its accuracy, completeness, consistency, timeliness and accessibility.

Additionally, data profiling involves a review of source data to understand the data's structure, content and interrelationships.

This review process delivers two high-level values to the organization: It provides a high-level view of the quality of its data sets; and two, it helps the organization identify potential data projects.

Given those benefits, data profiling is an important component of [data preparation](https://searchbusinessanalytics.techtarget.com/definition/data-preparation) programs. Its assistance helping organizations to identify quality data makes it an important precursor to data processing and data analytics activities.

Moreover, an organization can use data profiling and the insights it produces to continuously improve the quality of its data and measure the results of that effort.

Data profiling may also be known as data archeology, data assessment, data discovery or data quality analysis.

Organizations use data profiling at the beginning of a project to determine if enough data has been gathered, if any data can be reused or if the project is worth pursuing. The process of data profiling itself can be based on specific business rules that will uncover how the data set aligns with business standards and goals.

**Data profiling in business intelligence :**

Data profiling is **the process of examining, analyzing, and creating useful summaries of data**. The process yields a high-level overview which aids in the discovery of data quality issues, risks, and overall trends. Data profiling produces critical insights into data that companies can then leverage to their advantage.

### Types of data profiling

There are three types of data profiling.

* **Structure discovery.** This focuses on the formatting of the data, making sure everything is uniform and consistent. It uses basic statistical analysis to return information about the validity of the data.
* **Content discovery.** This process assesses the quality of individual pieces of data. For example, ambiguous, incomplete and null values are identified.
* **Relationship discovery.** This detects connections, similarities, differences and associations among data sources.

# Techniques of Data Profiling

## 1. Column profiling

Column profiling scans through a table and counts the number of times each value shows up within each column. This method can be useful to find frequency distribution and patterns within a column of data.

## 2. Cross-column profiling

Cross-column profiling is made up of two processes: key analysis and dependency analysis. Key analysis examines collections of attribute values by scouting for a possible primary key. Dependency analysis is a more complex process that determines whether there are relationships or structures embedded in a data set. Both techniques help analyze dependencies among data attributes within the same table.

## 3. Cross-table profiling

Cross-table profiling uses foreign key analysis, which is the identification of orphaned records and determination of semantic and syntactic differences, to examine the relationships of column sets in different tables. This can help cut down on redundancy but also identify data value sets that could be mapped together.

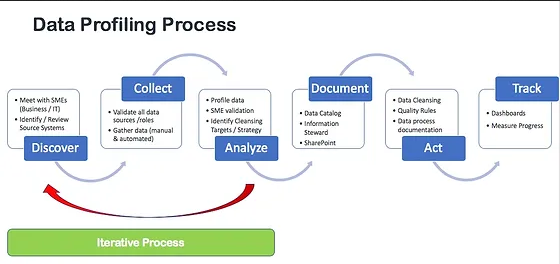
## 4. Data rule validation

Data rule validation uses data profiling in a proactive manner to verify that data instances and data sets conform to predefined rules. This process can be achieved either through batch validation or ongoing validation service.

### What are the steps in the data profiling process?

Data profiling helps organizations identify and fix data quality problems before the data is analyzed, so data professionals aren't dealing with inconsistencies, null values or incoherent schema designs as they process data to make decisions.

Data profiling statistically examines and analyzes data at its source and when loaded. It also analyzes the metadata to check for accuracy and completeness.



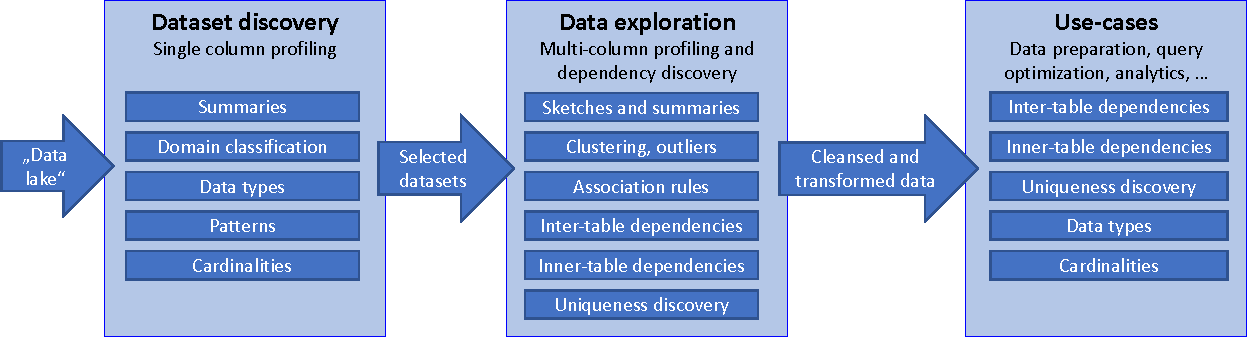
These are the **main steps** for profiling data sets.

It typically involves either writing queries or using data profiling tools.

A high-level breakdown of the process is as follows:

1. The first step of data profiling is gathering one or multiple data sources and the associated metadata for analysis.
2. The [data is then cleaned](https://searchdatamanagement.techtarget.com/definition/data-scrubbing) to unify structure, eliminate duplications, identify interrelationships and find anomalies.
3. Once the data is cleaned, data profiling tools will return various statistics to describe the data set. This could include the [mean](https://searchdatacenter.techtarget.com/definition/statistical-mean-median-mode-and-range), minimum/maximum value, frequency, recurring patterns, dependencies or data quality risks.

For example, by examining the frequency distribution of different values for each column in a table, a data analyst could gain insight into the type and use of each column. Cross-column analysis can be used to expose embedded value dependencies; inter-table analysis allows the analyst to discover overlapping value sets that represent [foreign key](https://searchoracle.techtarget.com/definition/foreign-key) relationships between entities.



### Benefits of data profiling

Data profiling returns a high-level overview of data that can result in the following benefits:

* leads to higher-quality, more credible data;
* helps with more accurate [predictive analytics](https://searchbusinessanalytics.techtarget.com/definition/predictive-analytics) and decision-making;
* makes better sense of the relationships between different data sets and sources;
* keeps company information centralized and organized;
* eliminates errors, such as missing values or outliers, that add costs to data-driven projects;
* highlights areas within a system that experience the most data quality issues, such as [data corruption](https://searchsqlserver.techtarget.com/definition/data-corruption) or user input errors; and
* produces insights surrounding risks, opportunities and trends.

**===================================================================**

**Attribute Anlysis:**

An **attribute analysis** is a powerful tool used to identify competitors' positions within a market and a critical **analysis** when positioning a product, service, or brand. It both helps to identify key differentiators and shows how crowded or competitive any particular positio4

**Example:**

An **attribute** is defined as a quality or characteristic of a person, place, or thing. Real life individuals and fictional characters possess various **attributes**. For **example**, someone might be labeled beautiful, charming, funny, or intelligent.

1. Range Analysis:

The range analysis process has six steps, each of which is discussed briefly in the following sections:

1. Understand the risks
2. Construct a quantitative model and identify the inputs it requires
3. Gather inputs and run the model
4. Check and validate the model
5. Reconcile the outcomes with reality
6. Revise the model and its parameters as necessary

The steps for cost and schedule range analysis are similar. Where there are important differences, they are noted in the discussion.

**Step 1: Understand the risks**

The purpose of this step is to gain a broad recognition of the uncertainties we need to take into account. The approach usually follows standard company practices, using workshops and generating risk registers.

**Step 2: Construct a quantitative model**

The purpose of this step is to build a quantitative model that is appropriate for the analysis and to identify the information that will be needed for it

Building useful and valid models is not easy. It requires a sound understanding of the main risks, and the underlying sources of uncertainty and their relationships.This understanding leads to a set of estimating parameters, on the right, that combine to represent the combined effects of the uncertainties, on the cost in this case.

No model will be absolutely comprehensive and precise. Judgment is required to determine which aspects of the uncertainty in a project should be incorporated in a model and how this should be done.

**Step 3: Gather inputs and run the model**

The purpose of this step is to gather data to populate the model and produce initial outputs (Figure 7).

Inputs are commonly generated in a range analysis workshop, covering the structure established in Step 2. We usually use a template which is structured to avoid systematic bias as far as possible.

**Step 4: Check and validate the model**

The purpose of validation is to pick up errors in the model, unrealistic input ranges or unrealistic prior expectations that the model will challenge, and so build confidence that the model is realistic (Figure 9).

Several approaches may be used, often in combination:

* Compare outcome with expectations, benchmarks and experience, and investigate if the outcomes ‘don’t smell right’
* Discuss and diagnose apparent anomalies with estimators, planners, managers and specialists
* Conduct sensitivity analyses
* Undertake detailed model audits if anomalous outcomes persist.

Robust model validation should always be undertaken, to help avoid common pitfalls, which include:

* Adopting an uncritical approach to models, amounting to blind faith in their veracity and the outcomes they generate
* Having unrealistic expectations of what a model can do and the accuracy of its outcomes
* Misunderstanding the effects of nodal bias in schedule networks, and the associated non-linear effects on overall durations.

**Step 5: Reconcile the model outcomes with reality**

The purpose of this step is to gain acceptance for unexpected or unwelcome findings, by demonstrating why the outcomes align with reality and are logical consequences of the model (Figure 10).

Discussion and diagnosis of anomalies and examination of sensitivity analysis are useful approaches. They often yield valuable insights into the challenges a project will face.

It is very important not to leave unresolved differences between the analysis and key stakeholders’ beliefs. If key stakeholders do not trust the outcomes, then the modelling effort is likely to have been largely wasted.

This step can bring about a revision of an estimate or a plan if the analysis has exposed unrealistic assumptions or other issues.

**Step 6: Revise the model and its parameters as necessary**

The estimate or plan may have to be revised to achieve an outcome required for the project to be a success. In this case, it is good practice to run the model based on the revised plan and estimate to confirm that the changes will deliver the desired outcome and not introduce any new risks.

Project teams will often use these models to explore options and carry out ‘what-if’ analyses, to illustrate to senior managers what challenges a project faces and possibly argue for variations to key performance measures or for additional resources to be allocated to the work.

**Summary**

Range analysis cannot be a mechanical process, but it should be systematic; Figure 11 shows the six main steps. The diagram highlights what is important at each step – use these to plan the range analysis.

Remember that models are not decision makers. They only help to make sense of a project, and their outcomes must be interpreted for the circumstances and the business requirement in which they are being implemented.

**2.Sparness**

A variable with **sparse(thin,light) data** is one in which a relatively high percentage of the variable's cells do not contain actual **data**. ... Random **sparsity** occurs when NA values are scattered throughout the **data** variable, usually because some combinations of dimension values never have any **data**.

Typically, sparse data means that there are many gaps present in the data being recorded. For example, in the case of the sensor mentioned above, the sensor may send a signal only when the state changes, like when there is a movement of the door in a room. This data will be obtained intermittently because the door is not always moving. Hence, this is sparse data.

However, if the sensor records, say, wind speed, the values change constantly. Thus, the data set that is obtained is dense.

**3.Format Evaluation**

Evaluation is a process that critically examines a program. It involves collecting and analyzing information about a program's activities, characteristics, and outcomes. Its purpose is to make judgments about a program, to improve its effectiveness, and/or to inform programming decisions

We often collect data so that we can find patterns in the data, like numbers trending upwards or correlations between two sets of numbers. Depending on the data and the patterns, sometimes we can see that pattern in a simple tabular presentation of the data.

Example

a customer accounts database containing a data field called ACCOUNT\_NUMBER, which always turned out to be composed of a two-character prefix followed by a nine-digit number.

There was existing code that automatically generated a new account number when a new customer was added.

It turned out that embedded in the data and the code were rules indicating how an account number was generated. Evidently, the two-character code represented a sales region, determined by the customer’s address, whereas the numeric value was assigned as an increasing number per customer in each sales region.

Because this attribute’s value carried multiple pieces of information, it was a classic example of an overloaded attribute.

Consider the following symbol classifications.

n Letter

n Digit

n Punctuation

n White space

Symbol pattern assignment is the first pass at pattern analysis (Figure 10.1). In

each string, we assign one of the foregoing classifications to each character

appearing in a data value. When all the characters in a string have been classified, the

string will have an associated pattern string as well. For each value string, we prepare

and record its pattern string. When all value strings have been analyzed, there is

a column of associated pattern strings ready to be collated and counted.

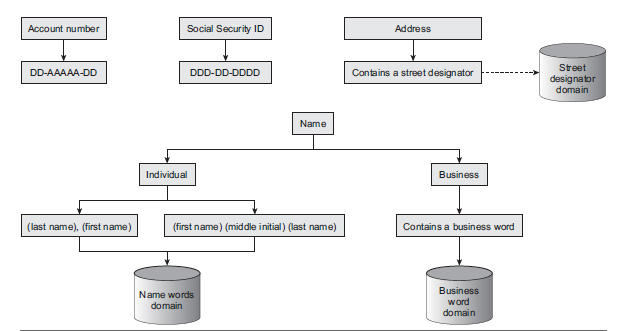
At this point, there are two tasks to be accomplished. The first is to look for

recurring patterns within the set of generated pattern strings; the second is to

check the generated pattern strings against the known sets of patterns. Either way,

the goal is to present candidate patterns representing rule-based domains to the

user.



Our next method for pattern analysis takes a more macro

view of the data by categorizing strings instead of symbols. At this point, all strings

can be classified as:

n Alphabetic

n Alphanumeric

n Numeric

n First name

n Last name

n Business word

n Address words

n One of any other categorized word class

In each attribute value, we now assign to each white space–separated string

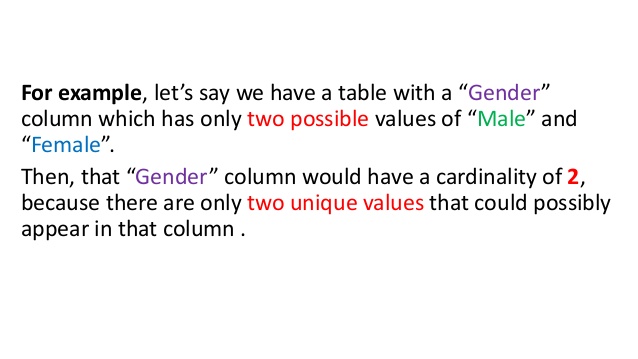
one of the word categories, forming a new pattern string. After all the strings

have had patterns assigned, again, these patterns can be collated and counted, and we check both for recurring patterns and for matches to previously known patterns.

**4. Cardinality and uniqueness**, where cardinality refers to the number of discrete values that appear within a column, and uniqueness tests to see that each row in a table has a unique value for any particular set of attributes.

A table's **cardinality** is the number of **rows** or **records** in the table. For an index, **cardinality** is considered the number of **unique** values in the index. A **unique** index would have **cardinality** equal to the number of **rows** in the table.

**High**-**cardinality** column values are typically identification numbers, email addresses, or user names. An example of a data table column with **high**-**cardinality** would be a USERS table with a column named USER\_ID. This column would contain unique values of 1-n. ... **Low**-**cardinality** refers to columns with few unique values.



**5. Frequency Distribution**

A data profiling tool will scan all the values in a column and provide a frequency distribution of that column's values, suggesting the type and potential uses of each column.

Frequency distribution, in statistics, a graph or data set organized to show the frequency of occurrence of each possible outcome of a repeatable event observed many times.

The **frequency** of a particular data **value** is the number of times the data **value** occurs. For example, if four students have a score of 80 in mathematics, and then the score of 80 is said to have a **frequency** of 4. The **frequency** of a data **value** is often represented by f.

A frequency table tabulates the number of times values from a data set appear within a configured range. As an example, you might have a list of employee scores and want to display the frequency of scores within certain ranges. ... You can then display this data on a bar chart to more readily compare the frequencies.

**6. Value Absence**

* **There are actually two problems associated with the absence of values that can be explored through data profiling.**
* **The first involves looking for values that are not there, and the second is to look for non-values that are there**

Data profiling can help you perform essential analysis such as:

* Provide a reality check for the perceptions and assumptions you may have about the quality of your data
* Verify your data matches the metadata that describes it
* Identify different representations for the absence of data (i.e., NULL and other missing values)
* Identify potential default values
* Identify potential invalid values
* Check data formats for inconsistencies
* Prepare meaningful questions to ask subject matter experts

**Abstract Type Analysis**

An abstract type is a more semantically descriptive qualification of a type definition that conveys business meaning. For example, “people names,” “telephone numbers,” and “ZIP codes” are all abstract data types that qualify as character strings.

==========

A relationship analysis identifies pairs of foreign keys and primary keys. Use this information to determine how data sets can be joined together.

A relationship analysis identifies pairs of foreign keys and primary keys. Use this information to determine how data sets can be joined together.

**1. Domain Analysis**

Domain Analysis is the **process** that identifies the relevant objects of an application domain. The goal of Domain Analysis is Software Reuse. The higher is the level of the life-cycle object to reuse, the larger are the benefits coming from its reuse, the harder is the definition of a workable **process**.

What are the inputs of domain analysis?

Raw domain knowledge from any relevant source is taken as input. Participants in the process can be, among others, domain experts and analysts. Outputs are (semi)formalized concepts, domain processes, standards, logical architectures, etc. Subsequent activities produce generic design fragments, frameworks, etc.

*Domain analysis* is the process by which a software engineer learns background information. He or she has to learn sufficient information so as to be able to understand the problem and make good decisions during requirements analysis and other stages of the software engineering process. The word ‘domain’ in this case means the general field of business or technology in which the customers expect to be using the software.

Some domains might be very broad, such as ‘airline reservations’, ‘medical diagnosis’, and ‘financial analysis’. Others are narrower, such as ‘the manufacturing of paint’ or ‘scheduling meetings’. People who work in a domain and who have a deep knowledge of it (or part of it), are called *domain experts*. Many of these people may become customers or users.

To perform domain analysis, you gather information from whatever sources of information are available: these include the domain experts; any books about the domain; any existing software and its documentation, and any other documents he or she can find. The interviewing, brainstorming and use case analysis techniques discussed later in this chapter can help with domain analysis. Object oriented modelling, discussed in the next chapter, can also be of assistance.

The following benefits will make this work worthwhile:

• **Faster development:** You will be able to communicate with the stakeholders more effectively, hence you will be able to establish requirements more rapidly. Having performed domain analysis will help you to focus on the most important issues.

• **Better system:** Knowing the subtleties of the domain will help ensure that the solutions you adopt will more effectively solve the customer’s problem. You will make fewer mistakes, and will know which procedures and standards to follow. The analysis will give you a global picture of the domain of application; this will lead to better abstractions and hence improved designs.

• **Anticipation of extensions:** Armed with domain knowledge, you will obtain insights into emerging trends and you will notice opportunities for future development. This will allow you to build a more adaptable system.