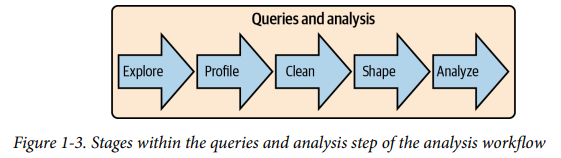


1. First, data is generated by **source systems**, a term that includes any human or machine process that generates data of interest. Data can be generated by people by hand, such as when someone fills out a form or takes notes during a doctor’s visit. Data can also be machine generated, such as when an application database records a purchase, an event-streaming system records a website click, or a marketing management tool records an email open
2. The second step is moving the data and storing it in a database for analysis. I will use the terms **data warehouse**, which is a database that consolidates data from across an organization into a central repository, and data store, which refers to any type of data storage system that can be queried. Other terms you might come across are data mart, which is typically a subset of a data warehouse, or a more narrowly focused data warehouse; and data lake, a term that can mean either that data resides in a file storage system or that it is stored in a database but without the degree of data trans‐ formation that is common in data warehouses. Data warehouses range from small and simple to huge and expensive
   1. Usually a person or team is responsible for getting data into the data warehouse. This process is called **ETL**, or **extract, transform, load**. Extract pulls the data from the source system. Transform optionally changes the structure of the data, performs data quality cleaning, or aggregates the data. Load puts the data into the database. This process can also be called **ELT, for extract, load, transform**—the difference being that, rather than transformations being done before data is loaded, all the data is loaded and then transfor‐ mations are performed, usually using SQL. You might also hear the terms source and target in the context of ETL. The source is where the data comes from, and the target is the destination, i.e., the data‐ base and the tables within it. Even when SQL is used to do the transforming, another language such as Python or Java is used to glue the steps together, coordinate scheduling, and raise alerts when something goes wrong. There are a number of commercial products as well as open source tools available, so teams don’t have to create an ETL system entirely from scratch.
3. Once the data is in a database, the next step is **performing queries and analysis**. In this step, SQL is applied to explore, profile, clean, shape, and analyze the data. Figure 1-3 shows the general flow of the process



* 1. **Exploring** the data involves becoming familiar with the topic, where the data was generated, and the database tables in which it is stored.
  2. **Profiling** involves checking the unique values and distribution of records in the data set.
  3. **Cleaning** involves fixing incorrect or incomplete data, adding categorization and flags, and handling null values.
  4. **Shaping** is the process of arranging the data into the rows and columns needed in the result set.
  5. **Analyzing** the data involves reviewing the output for trends, conclusions, and insights.

Although this process is shown as linear, in practice it is often cyclical—for example, when shaping or analysis reveals data that should be cleaned.

1. **Presentation of the data** into a final output form is the last step in the overall workflow. Businesspeople won’t appreciate receiving a file of SQL code; they expect you to present graphs, charts, and insights. Communication is key to having an impact with analysis, and for that we need a way to share the results with other people. At other times, you may need to apply more sophisticated statistical analysis than is possible in SQL, or you may want to feed the data into a machine learning (ML) algorithm. Fortunately, most reporting and visualization tools have SQL connectors that allow you to pull in data from entire tables or prewritten SQL queries. Statistical software and languages commonly used for ML also usually have SQL connectors.