

Data Analytics for Accounting, 1e (Richardson)
Chapter 2 Data Preparation and Cleaning

- 1) Mastering the data requires a firm understanding of what data is available to you and where it is stored, as well as being skilled in the process of extracting, transforming, and loading (ETL).
- 2) A flat file is a means of storing data in one place, such as in an Excel spreadsheet, as opposed to storing the data in multiple tables, such as in a relational database.
- 3) A foreign key is an attribute that is required to exist in each table of a relational database and serves as the unique identifier for each record in a table.
- 4) A primary key is an attribute that is required to exist in each table of a relational database and serves as the unique identifier for each record in a table.
- 5) A foreign key is an attribute that exists in relational databases in order to carry out the relationship between two tables.
- 6) A composite primary key is made up of the three or more primary keys in the tables that it is linking.
- 7) Descriptive attributes are attributes that exist in relational databases that are neither primary nor foreign keys.
- 8) Once you have extracted the data of interest, it will need to be validated for completeness and existence.
- 9) The E in IMPACT Cycle represents existence.
- 10) The T in IMPACT Cycle represents transfer.
- 11) The L in IMPACT Cycle represents loading.
- 12) In order to obtain the right data, it is important to have a firm grasp of what data is available and how it is stored.
- 13) Data normalization can reduce data redundancy and improve data integrity.
- 14) Much like the IMPACT cycle, requesting data is often an iterative process.
- 15) Unlike the IMPACT cycle, requesting data is not an iterative process.
- 16) If the extraction and transformation steps have been done correctly, the loading part of the ETL process should be the simplest step.
- 17) After obtaining the data and determining the purpose and scope of the data request, the next

step is to validate the data.

18) Comparing the number of records that were extracted to the number of records in the source database is an example of validating the data for integrity.

19) Formatting negative numbers is an example of cleaning the data.

20) A template can make communication easier between data requestor and provider.

21) Mastering the data can also be described via the ETL process. The ETL process stands for:

- A) Extract, total, and load data.
- B) Extract, transform, and load data.
- C) Enter, transform, and load data.
- D) Enter, total, and load data.

22) All of the following are Audit Data Standards (ADS) developed by the American Institute of Certified Accountants *except*:

- A) Investments subledger standards
- B) General Ledger standards
- C) Procure-to-Pay subledger standards
- D) Order-to-Cash subledger standards

23) When using [EmployeeID] as the unique identifier of the Employee table, [EmployeeID] is an example of which of the following:

- A) Foreign key
- B) Composite key
- C) Primary key
- D) Key attribute

24) The purpose of extracting data is:

- A) To validate the data for completeness and integrity
- B) To load the data into the appropriate tool for analysis
- C) To identify and obtain the data from the appropriate source
- D) To identify which approach to data analytics should be used

25) The purpose of transforming data is:

- A) To validate the data for completeness and integrity
- B) To load the data into the appropriate tool for analysis
- C) To identify and obtain the data from the appropriate source
- D) To identify which approach to data analytics should be used

- 26) The purpose of loading data is:
- A) To validate the data for completeness and integrity
 - B) To put the data into the appropriate tool for analysis
 - C) To identify and obtain the data from the appropriate source
 - D) To identify which approach to data analytics should be used
- 27) All of the following are included in the five steps of the ETL process *except*:
- A) Determine the purpose and scope of the data request
 - B) Obtain the data
 - C) Validate the data for completeness and integrity
 - D) Scrub the data
- 28) Which of the following best exemplifies a way that data will need to be cleaned after extraction and validation?
- A) Remove headings and subtotals
 - B) Validate date/time fields
 - C) Remove trailing zeroes
 - D) Compare string limits for text fields
- 29) _____ is the metadata that describes each attribute in a database.
- A) Relational database
 - B) Data dictionary
 - C) Descriptive attributes
 - D) Flat file
- 30) Removing headings or subtotals from data is an example of which of the following?
- A) Validating the data for completeness
 - B) Validating the data for integrity
 - C) Cleaning the data
 - D) Obtaining the data
- 31) Correcting inconsistencies across data is an example of which of the following?
- A) Validating the data for completeness
 - B) Validating the data for integrity
 - C) Cleaning the data
 - D) Obtaining the data
- 32) Formatting negative numbers in the data is an example of which of the following?
- A) Validating the Data for Completeness
 - B) Validating the Data for Integrity
 - C) Cleaning the Data
 - D) Obtaining the Data

33) Removing leading zeroes and non-printable characters from the data is an example of which of the following?

- A) Validating the data for completeness
- B) Validating the data for integrity
- C) Cleaning the data
- D) Obtaining the data

34) Comparing descriptive statistics for numeric fields within the data is an example of which of the following?

- A) Validating the data for completeness
- B) Validating the data for integrity
- C) Cleaning the data
- D) Obtaining the data

35) Comparing the number of records within the data is an example of which of the following?

- A) Validating the data for completeness
- B) Validating the data for integrity
- C) Cleaning the data
- D) Obtaining the data

36) Validating date/time fields within the data is an example of which of the following?

- A) Validating the data for completeness
- B) Validating the data for integrity
- C) Cleaning the data
- D) Obtaining the data

37) Which of the following best describes the purpose of a primary key?

- A) To ensure that each row in the table is unique
- B) To create the relationship between two tables
- C) To provide business information, but are not required to build a database
- D) To support business processes across the organization

38) Which of the following best describes the purpose of a non-key attribute?

- A) To ensure that each row in the table is unique
- B) To create the relationship between two tables
- C) To provide business information
- D) To support business processes across the organization

39) Which of the following best describes the purpose of relational databases?

- A) To ensure that business rules are enforced
- B) To increase information redundancy in the organization
- C) To provide business information to data analysts
- D) To support business processes across the organization

- 40) Which of the following best describes the purpose of a foreign key?
- A) To ensure that each row in the table is unique
 - B) To create the relationship between two tables
 - C) To provide business information
 - D) To support business processes across the organization
- 41) A data dictionary is paramount in helping data analysts do which of the following?
- A) Maintain databases.
 - B) Identify the data they need to use.
 - C) Communicating insights.
 - D) Track outcomes.
- 42) A data dictionary is paramount in helping database administrators do which of the following?
- A) Maintain databases.
 - B) Identify the data they need to use.
 - C) Communicating insights.
 - D) Track outcomes.
- 43) At which step of the ETL process should you try to answer the question "What tools will be used to perform data analytic tests or procedures and why?"
- A) Step 1: Determine the purpose and scope of the data request.
 - B) Step 2: Obtain the data.
 - C) Step 3 or 4: Transformation.
 - D) Step 5: Loading the data for data analysis.
- 44) At which step of the ETL process should you try to answer the question "What other information will impact the nature, timing and extent of the data analysis?"
- A) Step 1: Determine the purpose and scope of the data request.
 - B) Step 2: Obtain the data.
 - C) Step 3 or 4: Transformation.
 - D) Step 5: Loading the data for data analysis.
- 45) At which step of the ETL process should you try to answer the question "What business problem will the data address?"
- A) Step 1: Determine the purpose and scope of the data request.
 - B) Step 2: Obtain the data.
 - C) Step 3 or 4: Transformation.
 - D) Step 5: Loading the data for data analysis.
- 46) At which step of the ETL process should you try to answer the question "Where are the data located in the financial or other related systems?"
- A) Step 1: Determine the purpose and scope of the data request.
 - B) Step 2: Obtain the data.
 - C) Step 3 or 4: Transformation.
 - D) Step 5: Loading the data for data analysis.

47) When obtaining the data yourself, you should do all of the following before you begin *except*:

- A) Identify the tables that contain the information you need.
- B) Identify which attributes specifically hold the information you need in each table.
- C) Identify how those tables are related to each other.
- D) Identify any errors or issues from the extraction.

48) There are a variety of methods that you could take to retrieve the data, including SQL. What does SQL stand for?

- A) Systems Query Language.
- B) Systems Question Language.
- C) Structured Question Language.
- D) Structured Query Language.

49) All of the following are benefits of using a normalized relational database *except*:

- A) Completeness.
- B) No redundancy.
- C) Business rules are enforced.
- D) Data is stored in one place.

50) Which of the following is most likely to be the primary key in an Employee table?

- A) Employee ID
- B) Employee Social Security Number
- C) Employee Name
- D) Employee Type

51) Chapter 2 describes, the various ways in which data can be stored for differing purposes. Describe the two ways data can be organized and the purpose for each organizational structure.

52) Taylor is a new staff accountant for a fortune 100 company. After hearing that she just successfully completed an Accounting Data Analytics course, her boss said, "Get me a listing of all our deadbeat customer, so I can cut off their credit." After asking clarifying questions, Taylor was able to determine that root request was "Which customer, with a credit limit over \$10,000, have more than \$5,000 outstanding for more than 90 days at the prior quarter's end?" Use the ETL Techniques briefly describe the process Taylor with have to complete to answer her boss' question. (Assume that Taylor does not have direct access the data, data will export into an Excel file, and she will complete the analysis with an Excel Pivot Table.)

53) Define and compare a primary key, a foreign key, and non-key attribute.

54) Assume that you will be up for a promotion next month and you'd like to analyze a recently acquired database to show off your data analytic skills. After identifying the goal of your data analysis, using the first step of the IMPACT cycle, what steps would you take if you have direct access to the database?

55) Assume that you have just completed the extraction process on a data set. As you begin validating the data for completeness and integrity, you notice an error. Describe the steps you might take to determine the source of an error.