Chapter 2: Fundamentals of Data Analysis

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 2 - Sections & Objectives

- 2.1 What is Data Analysis
  - Explain how data is used to create knowledge.
- 2.2 Using Big Data
  - Use software tools to visualize a data analysis following the Data Analysis Lifecycle process.
- 2.3 Data Acquisition and Preparation
  - Configure data for analysis.
- 2.4 Big Data Ethics
  - Explain why ethics are important when using Big Data.
- 2.5 Preparation for Chapter 2 Internet Meter Labs
  - Analyze data by using an external application and SQLite.
- 2.6 Summary
  - Summarize the concepts presented in this chapter.

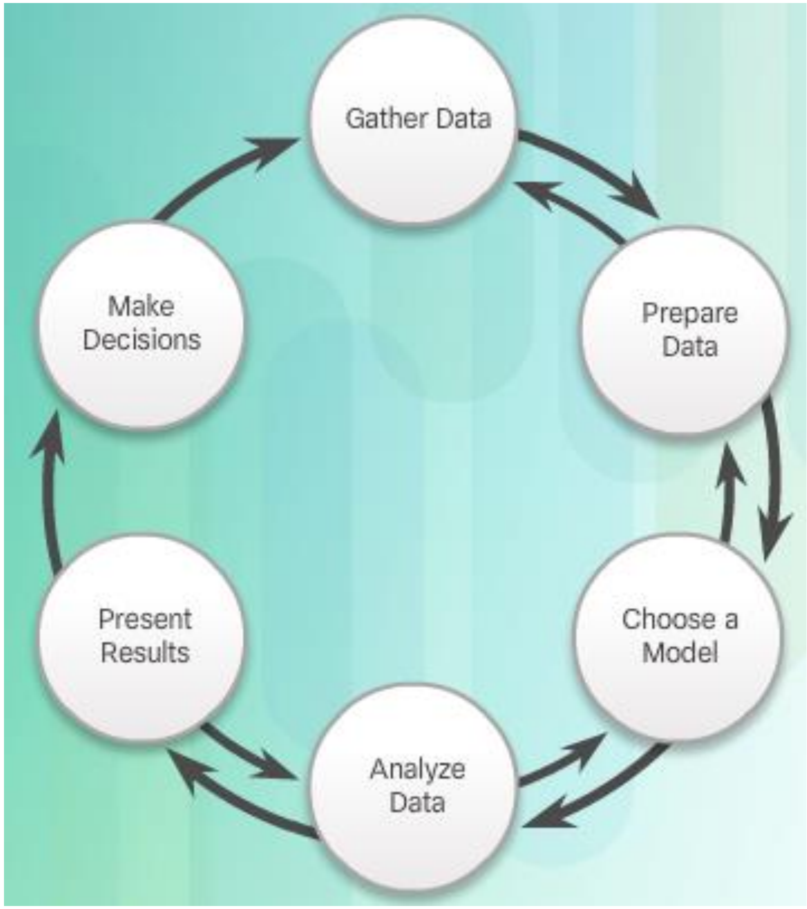
2.1 What is Data Analysis

Cisco | Networking Academy®
Mind Wide Open™

What is Data Analysis?

Analytics Models

- The six-step Data Analysis Lifecycle
- Data Analytics tools should provide:
  - Ease of use
  - Data manipulation
  - Sharing
  - Interactive exploration



What Are Analytics?

Analytics Models cont…

- The Python programming language has become a commonly used tool for handling and manipulating data.
- Python will be used in this course to perform data cleaning, analysis, and manipulation.
- Jupyter Notebooks will be used as both a document for written instructions as well as a Python command interface for running code.

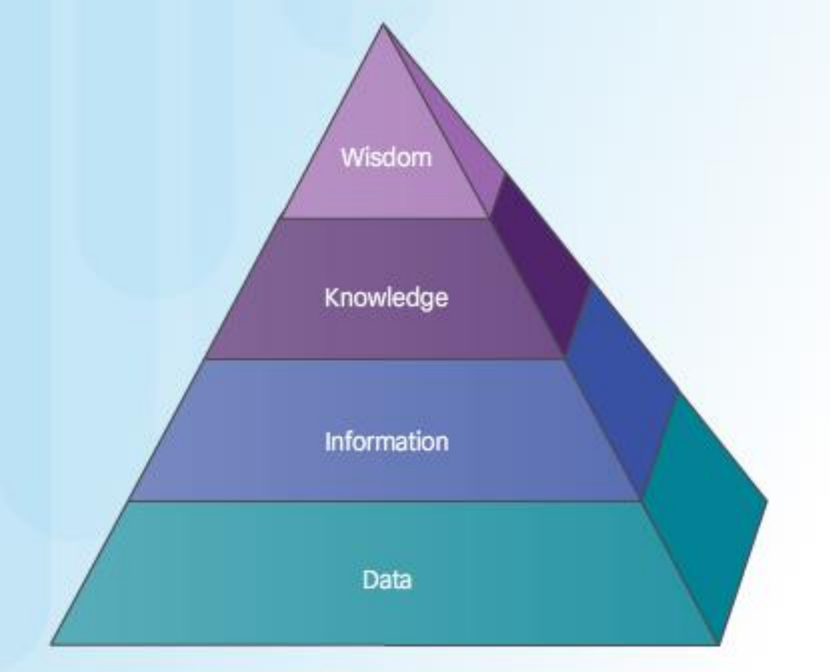
- The libraries that will be used in this course:
  - **NumPy** – This library adds support for arrays and matrices. It also has many built-in mathematical functions for use on data sets.
  - **Pandas** – This library adds support for tables and time series. Pandas is used to manipulate and clean data, among other uses.
  - **Matplotlib** – This library adds support for data visualization. Matplotlib is a plotting library capable of creating simple line plots to complicated 3D and contour plots.

2.2 Using Big Data

Cisco | Networking Academy®
Mind Wide Open™

Types of Data Analysis

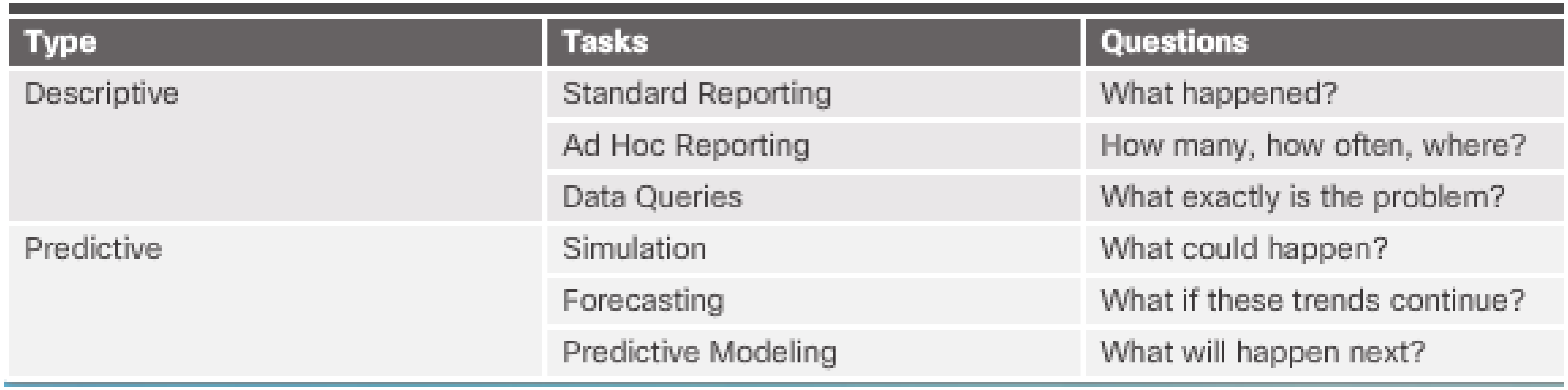
- Scalable technologies are enabling data center administrators to manage the top three aspects of Big Data:
  - **Volume**
  - **Velocity**
  - **Variety**
- The **Data**, **Information**, **Knowledge**, and **Wisdom** (DIKW) model shows the transitions that data undergoes until it gains enough value to inform wise decisions. This is called **Business Intelligence**



Using Big Data

Why Analyze Big Data?

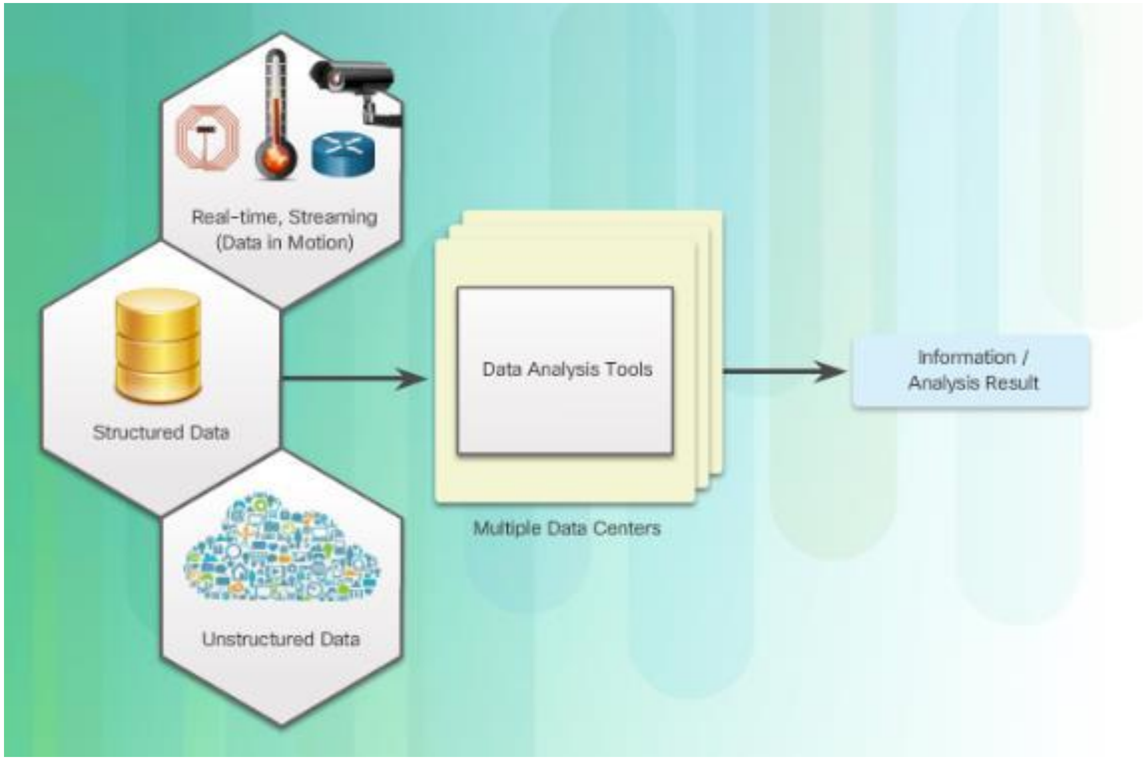
- Multiple types of analytics provide organizations and people with information that can drive innovation, improve efficiency and mitigate risk.
  - **Descriptive analytics** - Relies solely on historical data to provide regular reports on events that have already happened.
  - **Predictive analytics** - Can infer missing data and establish a future trend line based on past data. It uses simulation models and forecasting to suggest what could happen.
  - **Prescriptive analytics** - Recommends actions or decisions based on a complex set of targets, constraints, and choices.



Type	Tasks	Questions
Descriptive	Standard Reporting	What happened?
	Ad Hoc Reporting	How many, how often, where?
	Data Queries	What exactly is the problem?
Predictive	Simulation	What could happen?
	Forecasting	What if these trends continue?
	Predictive Modeling	What will happen next?

Timely Analysis of Big Data

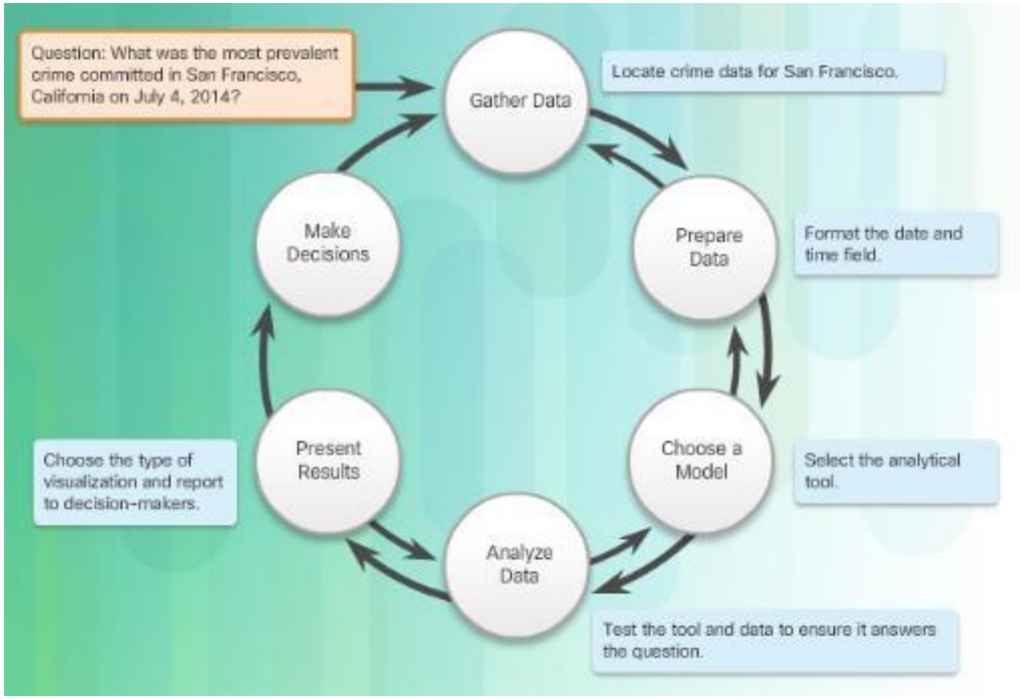
- With Big Data, much of the value of data is derived from creating opportunities to take action immediately.
- Data-driven decisions can have the following benefits:
  - Increased time to research and develop products and services
  - Increased efficiency and faster manufacturing
  - Faster time to market
  - More effective marketing and advertising



Using Big Data

Data Analysis Lifecycle

- **Gathering the data** - The process of locating data and then determining if there is enough data to complete the analysis.
- **Preparing the data** - This step can involve many tasks to transform the data into a format appropriate for the tool that will be used.
- **Choosing a model** - This step includes choosing an analysis technique that will best answer the question with the data available.
- **Analyzing the data** - The process of testing the model against the data and determining if the model and the analyzed data are reliable. Were you able to answer the question with the selected tool?
- **Presenting the results** - The process of communicating the results to decision-makers.
- **Making decisions** - Organizational leaders incorporate the new knowledge as part of the overall strategy. The process begins anew with gathering data.


Question: What was the most prevalent crime committed in San Francisco, California on July 4, 2014?
Gather Data
Locate crime data for San Francisco.
Prepare Data
Format the date and time field.
Choose a Model
Select the analytical tool.
Analyze Data
Test the tool and data to ensure it answers the question.
Present Results
Choose the type of visualization and report to decision-makers.
Make Decisions

2.3 Data Acquisition and Preparation

Cisco | Networking Academy®
Mind Wide Open™

Data Acquisition and Preparation

Sources of Data

There are many different sources of data.

- A vast amount of historical data can be found in files such as:
  - MS Word documents
  - Emails
  - Spreadsheets
  - MS PowerPoints
  - PDFs
  - HTML
  - and plaintext files
- Public and Private Archives
- CSV, JSON, and XML files use plaintext, a common format, and are compatible with a wide range of applications
- The Web can be mined for data using a web scraping application

Sources of Data cont…

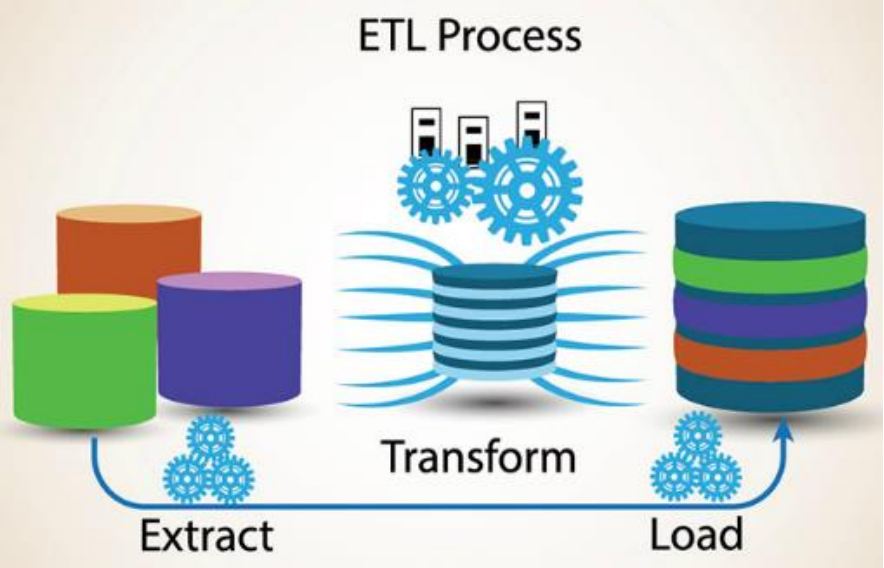
- The IoT uses sensors create data
  - Sensors in smartphones, cars, airplanes, street lamps, and home appliances capture raw data
- The list of things with sensors grows every year
  - The IoT contributes to the growth of Big Data

Data Acquisition and Preparation

Data Preparation

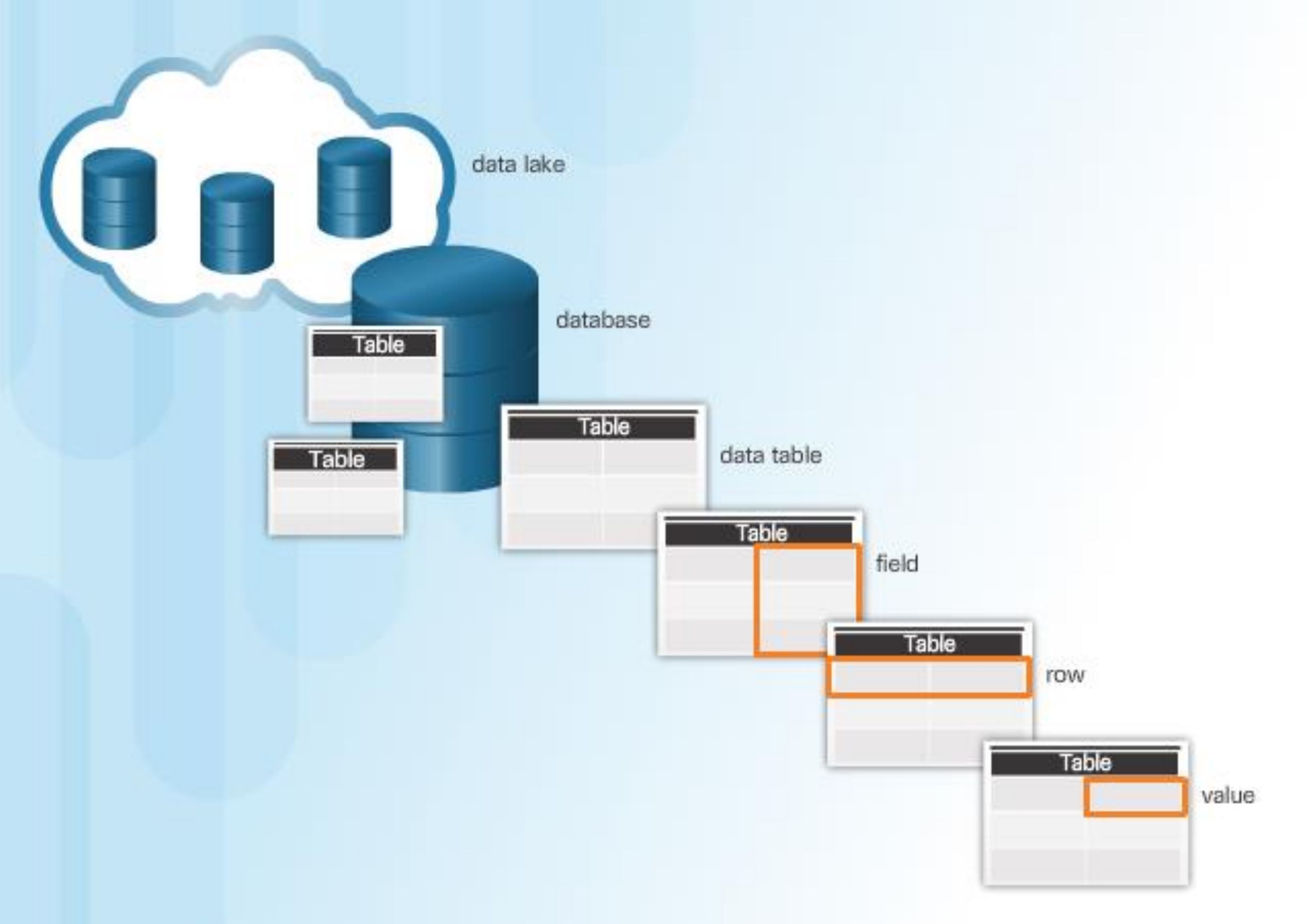
- Collected data may not be compatible or formatted correctly
  - Data must be prepared before it can be added to a data set

- Extract, Transform and Load (ETL)
  - process for collecting data from a variety of sources, transforming the data, and then loading the data into a database



Data Structures

- Relational Database Tables
  - Fields (Column)
  - Rows (Rows)
  - Values (Cells)
- Python
  - Strings
  - Lists
  - Tuples
  - Sets
  - Dictionaries



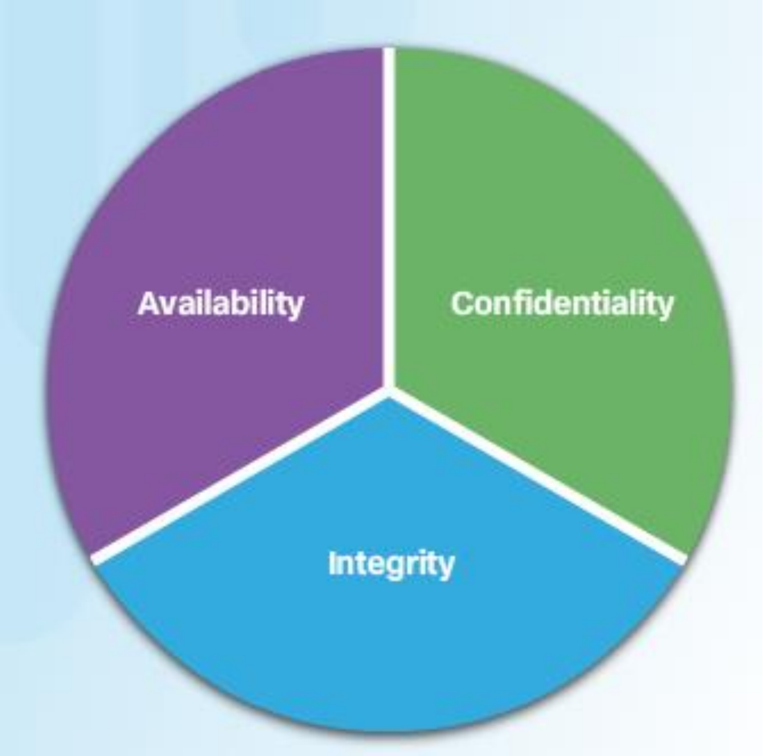
2.4 Big Data Ethics

Cisco | Networking Academy®
Mind Wide Open™

Big Data Ethics

What are the Ethical Concerns?

- Data protection regulations varies from country to country
- Confidentiality, integrity and availability, known as the CIA triad is a guideline for data security in an organization
- Four general cloud security controls:
  - Deterrent
  - Preventive
  - Detective
  - Corrective



2.5 Preparation for Ch2 Internet Meter Labs

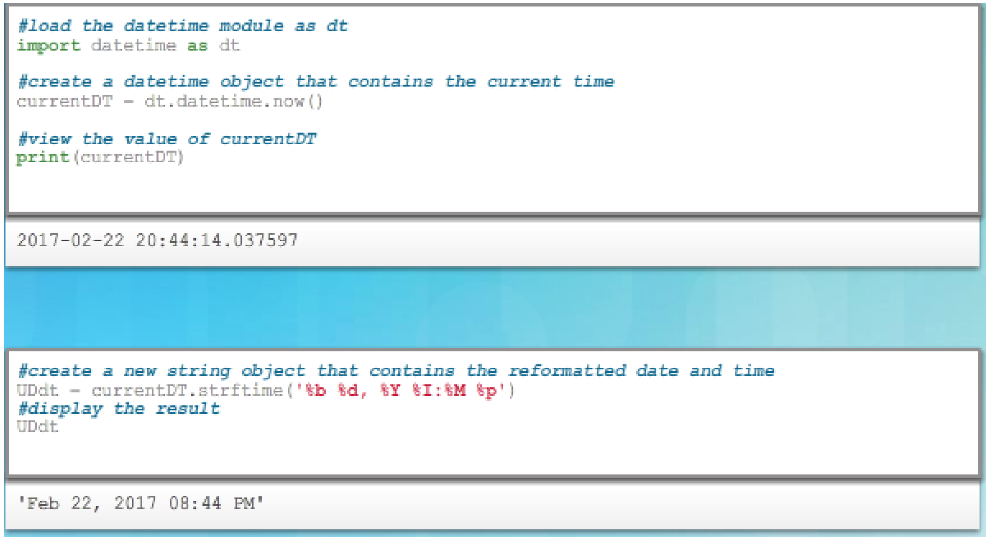
Cisco | Networking Academy®
Mind Wide Open™

Preparation for Chapter 2 Internet Meter Labs

Part 1

- The **datetime** module is included in most Python distributions as a standard library; however, it must be imported to be used in your code.

- The csv module allows reading and writing to .csv files.

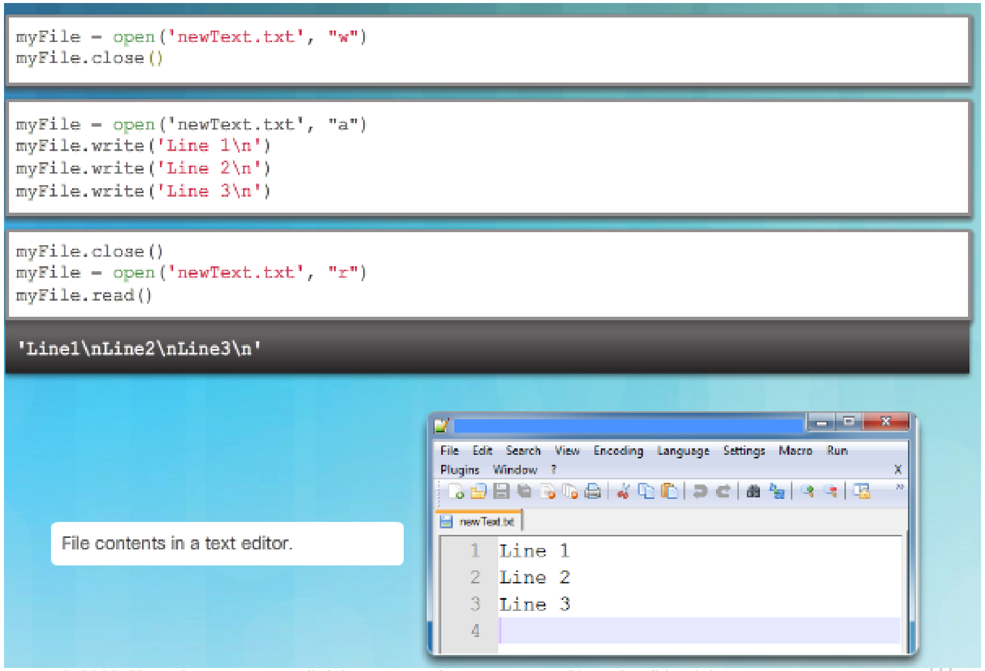


```
#load the datetime module as dt
import datetime as dt

#create a datetime object that contains the current time
currentDT = dt.datetime.now()

#view the value of currentDT
print(currentDT)
```

```
2017-02-22 20:44:14.037597
```

```
#create a new string object that contains the reformatted date and time
UDdt = currentDT.strftime('%b %d, %Y %I:%M %p')
#display the result
UDdt
```

```
'Feb 22, 2017 08:44 PM'
```



```
myFile = open('newText.txt', "w")
myFile.close()
```

```
myFile = open('newText.txt', "a")
myFile.write('Line 1\n')
myFile.write('Line 2\n')
myFile.write('Line 3\n')
```

```
myFile.close()
myFile = open('newText.txt', "r")
myFile.read()
```

```
'Line1\nLine2\nLine3\n'
```

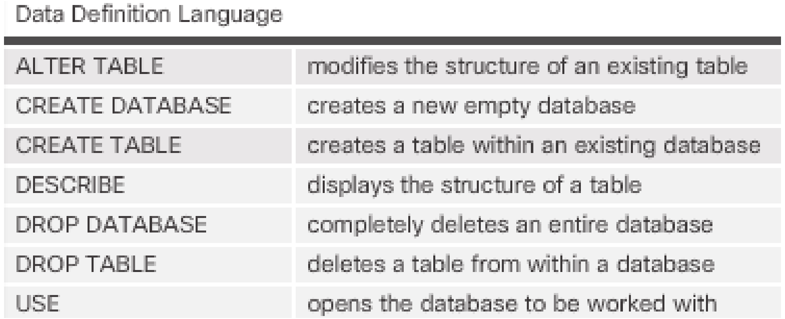
File contents in a text editor.

```
Line 1
Line 2
Line 3

```

Preparation for Chapter 2 Internet Meter Labs

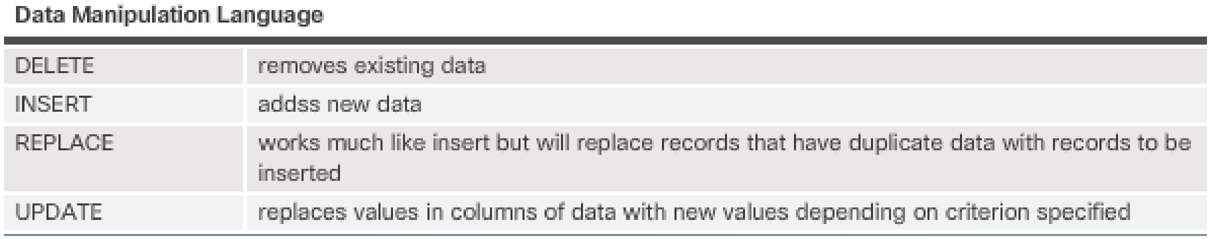
Part 2

- SQLite is an SQL implementation using a client server method of operation
  - Uses connections established between Python and an SQL database by creating an SQL connection object
- SQL can be said to be a language composed of three special purpose languages
  - Data Definition Language
  - Data Manipulation Language
  - Data Query Language



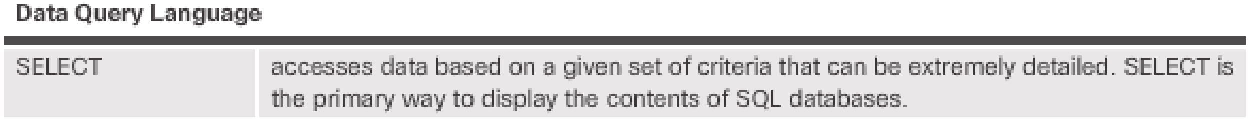
Data Definition Language

ALTER TABLE	modifies the structure of an existing table
CREATE DATABASE	creates a new empty database
CREATE TABLE	creates a table within an existing database
DESCRIBE	displays the structure of a table
DROP DATABASE	completely deletes an entire database
DROP TABLE	deletes a table from within a database
USE	opens the database to be worked with



Data Manipulation Language

DELETE	removes existing data
INSERT	addss new data
REPLACE	works much like insert but will replace records that have duplicate data with records to be inserted
UPDATE	replaces values in columns of data with new values depending on criterion specified



Data Query Language

SELECT	accesses data based on a given set of criteria that can be extremely detailed. SELECT is the primary way to display the contents of SQL databases.

2.6 Summary

Cisco | Networking Academy®
Mind Wide Open™

Chapter Summary

Summary

- Data can no longer be stored on a few machines or processed with just one tool
- Decision makers will increasingly rely on data analytics to extract the required information at the right time, in the right place, to make the right decision
- Descriptive analytics relies solely on historical data
- Predictive analytics attempts to predict what may happen
- Prescriptive analytics predicts outcomes and suggests courses of actions that will hold the greatest benefit for an organization
- Files, the Internet, sensors, and databases are all good sources of data.
- Extract, Transform and Load (ETL) is a process for collecting data from a variety of sources, transforming the data, and then loading the data into a database
- The CIA triad is a guideline for data security for an organization