Big Data

Big Data Course

Eng.Mostafa Nabieh

Mostafa Nabieh

وزارة الاتصالات وتكنولوجيا المعلومات

MANNING

وزارة الاتصالات وتقنية المعلومات
MINISTRY OF COMMUNICATIONS AND INFORMATION TECHNOLOGY

AI TALENTS HUB

UDACITY

PLURALSIGHT

Learning Space
YOUR SPACE TO LEARN FUTURE SKILLS

BIG DATA

Mostafa Nabieh

HUAWEI

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

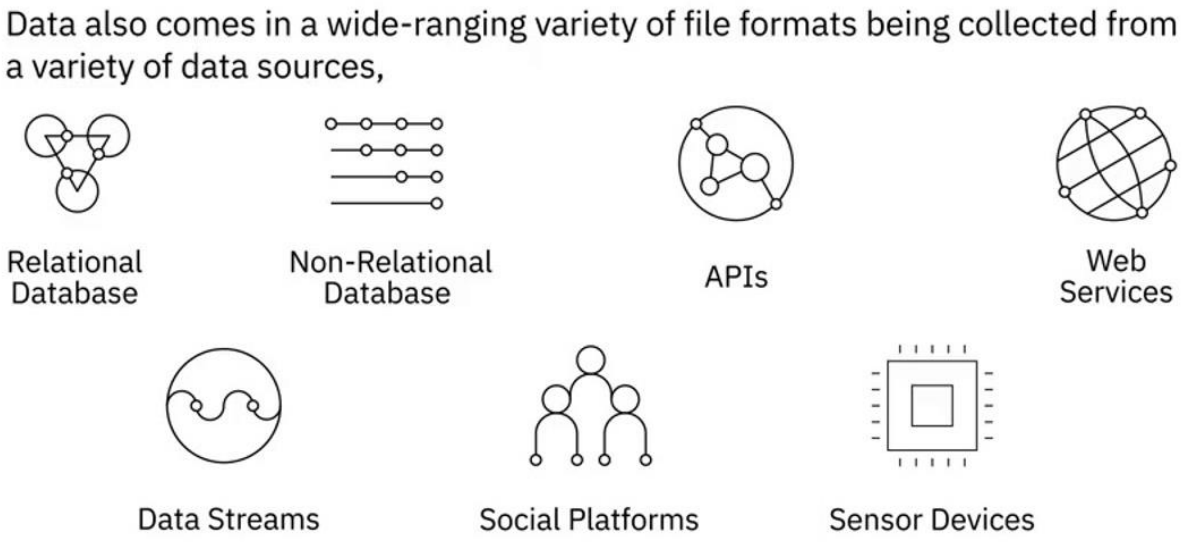
IBM

Mostafa Nabieh

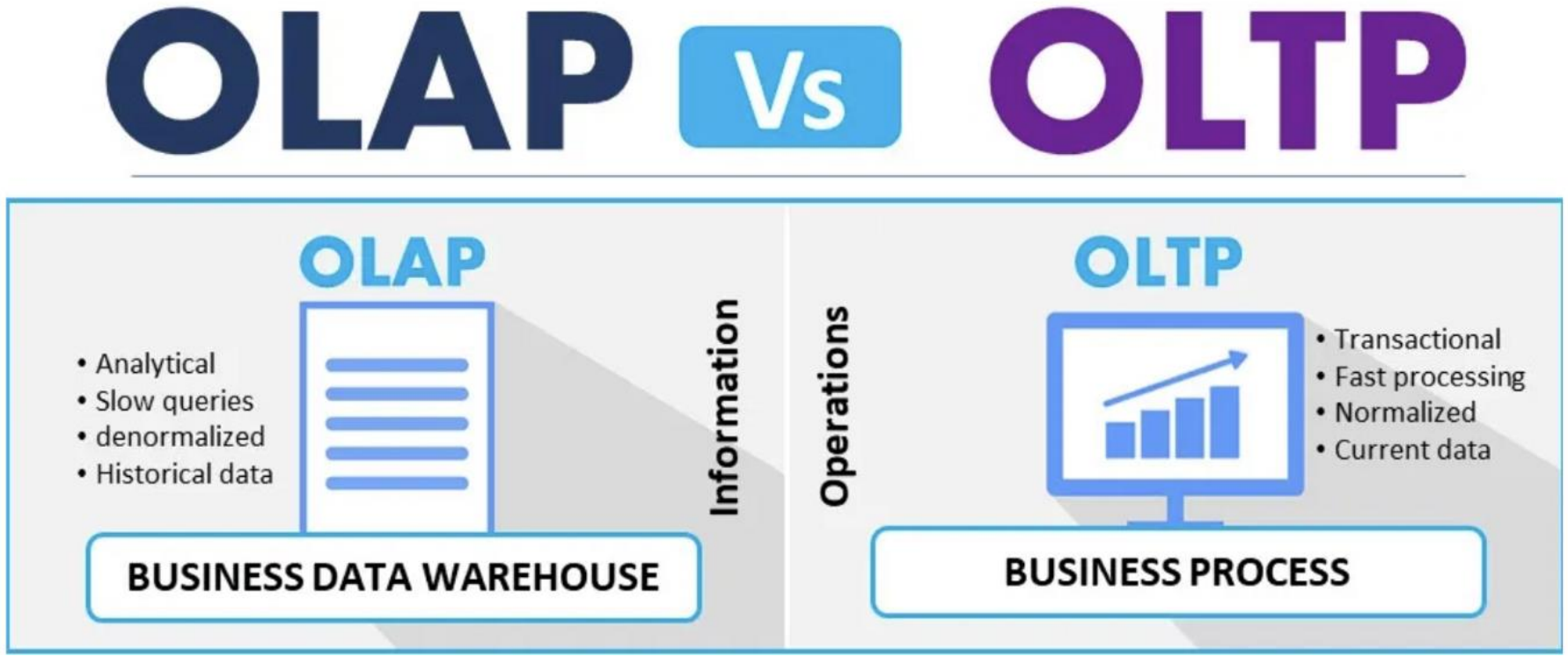
CONTENTS

Data

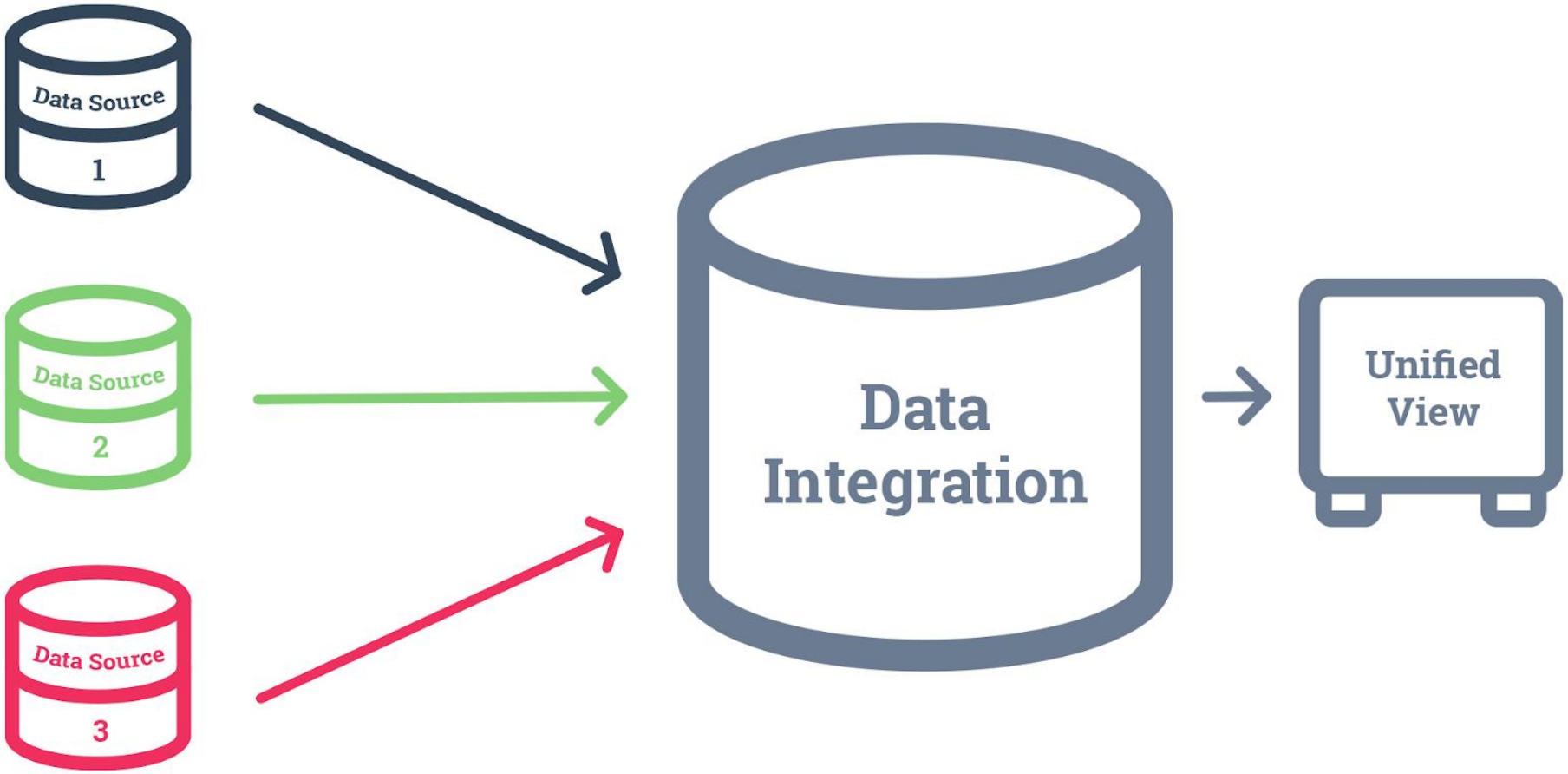
- Data also comes in a wide-ranging variety of file formats being collected from a variety of data sources



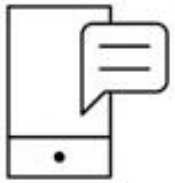
Data Repositories



Data Integration



Language



Query languages

For example, SQL for querying and manipulating data

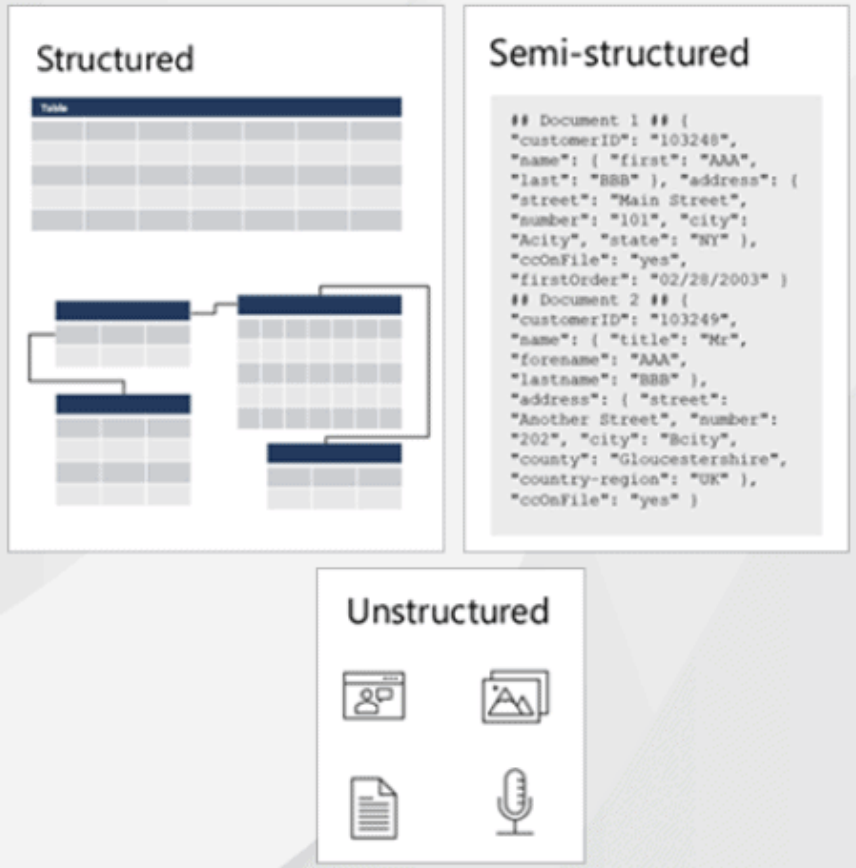
Programming languages

For example, Python for developing data applications

Shell and Scripting languages

For repetitive operational tasks

Types of data



Structured Data

VS

Semi-Structured Data

VS

Unstructured Data

Structured Data vs Semi-Structured Data vs Unstructured Data

Structured data	Semi-structured data	Unstructured data
Databases	XML / JSON data Email Web pages	Audio Video Image data Natural language Documents

Structured Data

Has a well-defined structure

Can be stored in well-defined schemas

Can be represented in a tabular manner with rows and columns

Structured Data

SAMPLE_BASIC_TABLE

Columns | Data | Model | Constraints | Grants | Statistics | Triggers | Flashback | Dependencies | Details | Partition

Actions...

	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	PRODUCT	VARCHAR2(256 BYTE)	Yes	(null)	1	(null)
2	MARKET	VARCHAR2(256 BYTE)	Yes	(null)	2	(null)
3	YEAR	VARCHAR2(256 BYTE)	Yes	(null)	3	(null)
4	SCENARIO	VARCHAR2(256 BYTE)	Yes	(null)	4	(null)
5	SALES	NUMBER(25,0)	Yes	(null)	5	(null)
6	STATENAME	VARCHAR2(256 BYTE)	Yes	(null)	6	(null)
7	COGS	NUMBER(25,0)	Yes	(null)	7	(null)
8	MARKETING	NUMBER(25,0)	Yes	(null)	8	(null)
9	PAYROLL	NUMBER(24,0)	Yes	(null)	9	(null)
10	MISC	NUMBER(23,0)	Yes	(null)	10	(null)
11	BEGINV	NUMBER(25,0)	Yes	(null)	11	(null)
12	ADDITIONS	NUMBER(25,0)	Yes	(null)	12	(null)

2011.01.27-ML.xlsx - Microsoft Excel

	IC Code	Filename	Rend Type ID	Width	Height	Nbr Colors (1,2,4)	Resolution (dpi)	Color Space (rgb, cmyk)	Run Length	Vid FPS	Vid Codec ID	Aud Samp Rate	Aud Res	Image Map Euid	Create System Rendition s	Create Online Thumb	Vid Loop		Asset ID	Asset Rend ID	Fil ID
2	sun000017	sun000017v4.flv	69	853	480	4			15	30		22	16						138945	646682	
3	sun000019	sun000019v4.flv	69	853	480	4			10	30		22	16						138947	646683	
4	sun000020	sun000020v4.flv	69	853	480	1			14	30		22	16						138948	646684	
5	sun000021	sun000021v4.flv	69	853	480	4			9	30		22	16						138949	646685	
6	sun000022	sun000022v4.flv	47	426	240	4			17	30		22	16						139431	646678	
7	sun000023	sun000023v4.flv	69	853	480	4			81	30		22	16						139432	646686	
8	sun000024	sun000024v1.flv	69	853	480	1			11	30		22	16						139433	646687	
9	sun000025	sun000025v4.flv	69	853	480	1			17	30		22	16						139434	646688	
10	<END>																				

Semi-Structured Data

Has some organizational properties but lacks a fixed or rigid schema

Cannot be stored in the form of rows and columns as in databases

Contains tags and elements, or metadata, which is used to group data and organize it in a hierarchy

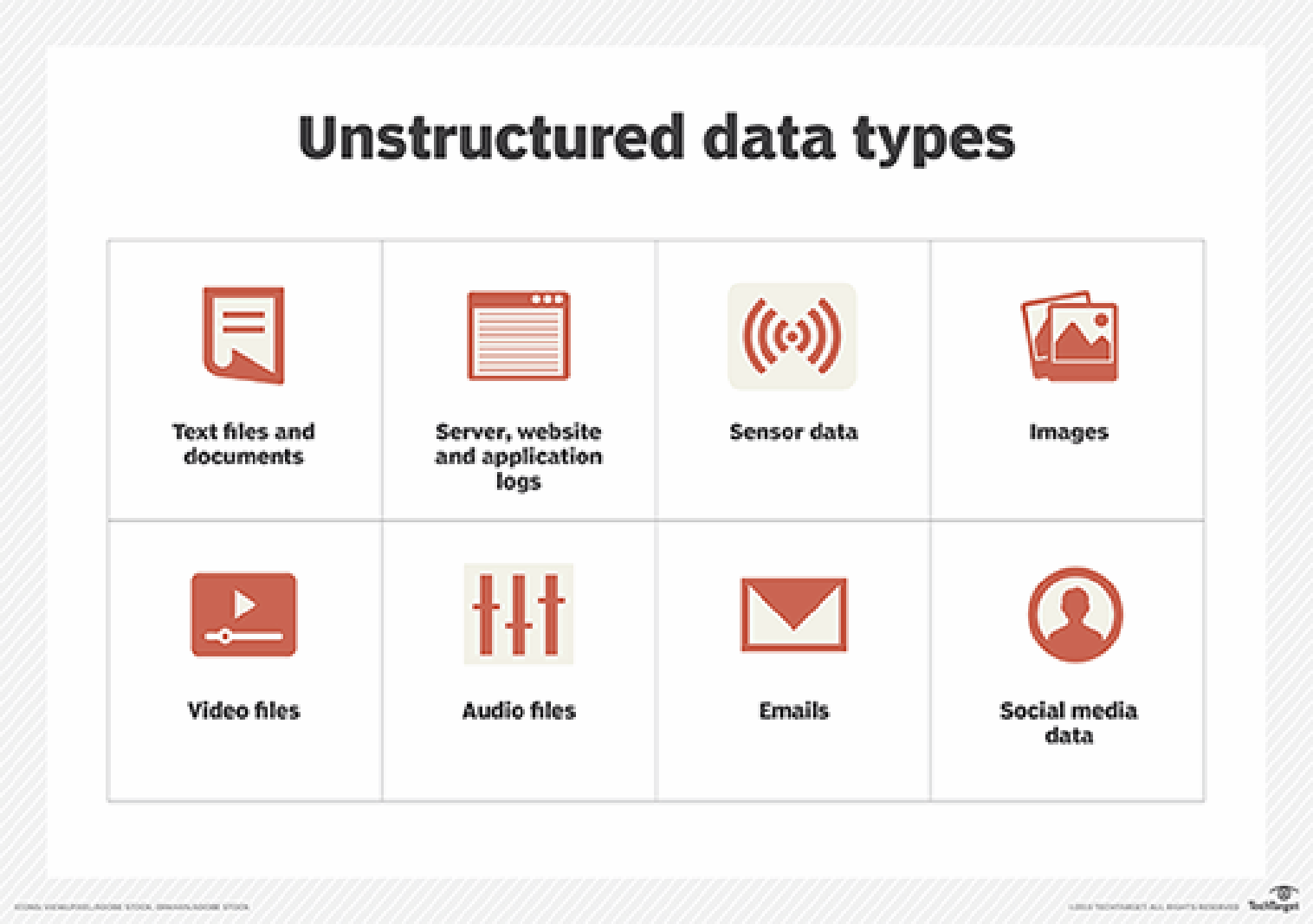
Semi-Structured Data

```
- <menu>
  - <area text="Welcome" file="index.html">
      <submenuitem text="New in Scribus 1.5" file="readme.html"/>
      <submenuitem text="Specification" file="specs.html"/>
    </area>
  - <area text="Documentation" file="intro.html">
    - <submenuitem text="Introduction" file="documentation.html">
        <submenuitem text="Editorial Notes" file="editorial.html"/>
        <submenuitem text="About the Team" file="about1.html"/>
      </submenuitem>
    - <submenuitem text="Setup" file="config.html">
        <submenuitem text="Configuring Scribus" file="settings1.html"/>
        <submenuitem text="Hyphenation and Spellchecking" file="hyphenator.html"/>
        <submenuitem text="Font Setup" file="fonts1.html"/>
        <submenuitem text="Fonts in Depth" file="fonts2.html"/>
      </submenuitem>
    - <submenuitem text="Scribus Basics" file="about2.html">
        <submenuitem text="Quick Start Guide" file="qsg.html"/>
        <submenuitem text="Command Line Reference" file="cli.html"/>
        <submenuitem text="Keyboard Shortcuts" file="keys.html"/>
        <submenuitem text="Mouse Shortcuts" file="mouse.html"/>
        <submenuitem text="Document Information" file="docinfo.html"/>
        <submenuitem text="Working with Frames" file="WwFrames.html"/>
        <submenuitem text="Working with Text" file="WwText.html"/>
        <submenuitem text="Text Properties" file="TextProp.html"/>
        <submenuitem text="Search and Replace" file="SearchReplace.html"/>
        <submenuitem text="Working with Styles" file="WwStyles.html"/>
        <submenuitem text="Working with Images" file="WwImages.html"/>
```

Unstructured Data

- Does not have an easily identifiable structure
- Cannot be organized in a mainstream relational database in the form of rows and columns
- Does not follow any format, sequence, semantics, or rules

Unstructured Data



Standard file formats

DELIMITED TEXT FILE FORMATS, OR .CSV

MICROSOFT EXCEL OPEN .XML SPREADSHEET, OR .XLSX

EXTENSIBLE MARKUP LANGUAGE, OR .XML

PORTABLE DOCUMENT FORMAT, OR .PDF

JAVASCRIPT OBJECT NOTATION, OR .JSON

Delimiter text files

- Files used to store data as text Each value is separated by a delimiter
- Delimiter - A sequence of one or more characters for specifying the boundary between independent entities or values.
- Comma, Tab, Colon, Vertical Bar, Space
- Comma-separated values Tab-separated values

CSV and TSV

Delimited text files

```
Manufacturer, Model, Sales_in_thousands, __year_resale_value, Vehicle_type, Price_in_thousands
Acura, Integra, 16.919, 16.36, Passenger, 21.5
Acura, TL, 39.384, 19.875, Passenger, 28.4
Acura, CL, 14.114, 18.225, Passenger, 14
Acura, RL, 8.588, 29.725, Passenger, 42
Audi, A4, 20.397, 22.255, Passenger, 23.99
Audi, A6, 18.78, 23.555, Passenger, 33.95
Audi, A8, 1.38, 39, Passenger, 62
BMW, 323i, 19.747, Passenger, 26.99
BMW, 328i, 9.231, 28.675, Passenger, 33.4
BMW, 528i, 17.527, 36.125, Passenger, 38.9
Buick, Century, 91.561, 12.475, Passenger, 21.975
```

.CSV

.TSV

Manufacturer	Model	Sales_in_thousands	__year_resale_value
Acura Integra	16.919	16.36 Passenger	21.5
Acura TL	39.384	19.875 Passenger	28.4
Acura CL	14.114	18.225 Passenger	
Acura RL	8.588	29.725 Passenger	42
Audi A4	20.397	22.255 Passenger	23.99
Audi A6	18.78	23.555 Passenger	33.95
Audi A8	1.38	39 Passenger	62
BMW 323i	19.747	Passenger	26.99
BMW 328i	9.231	28.675 Passenger	33.4
BMW 528i	17.527	36.125 Passenger	38.9
Buick Century	91.561	12.475 Passenger	21.975

Extensible Markup Language or .XML

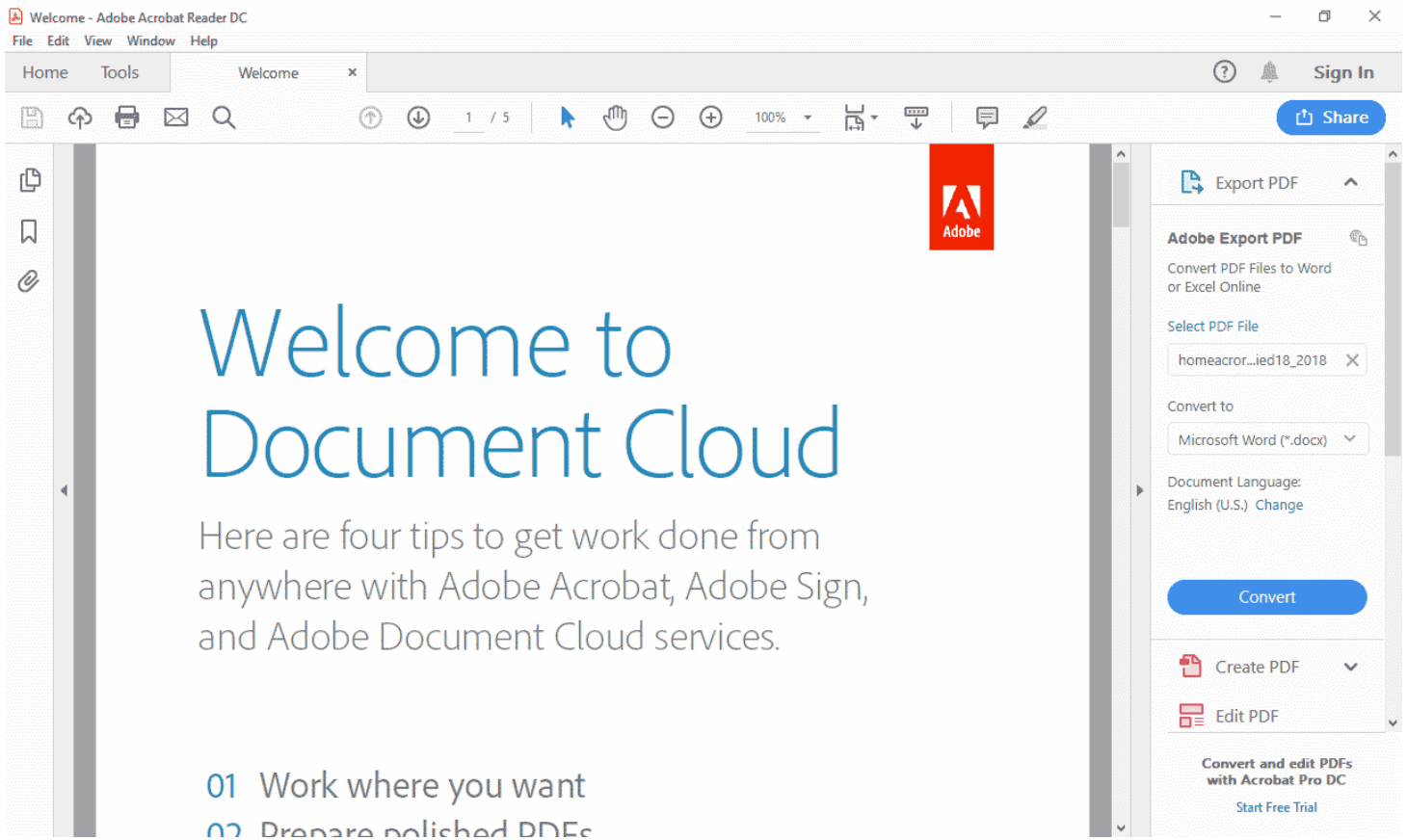
Extensible Markup Language, or XML, is a markup language with set rules for encoding data.

- Readable by both humans and machines
- Self-descriptive language
- Like HTML in some respects
- Does not use predefined tags like .HTML does
- Platform independent
- Programming language independent
- Makes it simpler to share data between systems

```
<?xml version="1.0" encoding="UTF-8"?>
- <EmployeeData>
  - <employee id="34594">
      <firstName>Heather</firstName>
      <lastName>Banks</lastName>
      <hireDate>1/19/1998</hireDate>
      <deptCode>BB001</deptCode>
      <salary>72000</salary>
    </employee>
  - <employee id="34593">
      <firstName>Tina</firstName>
      <lastName>Young</lastName>
      <hireDate>4/1/2010</hireDate>
      <deptCode>BB001</deptCode>
      <salary>65000</salary>
    </employee>
</EmployeeData>
```

Portable Document Format or PDF

- Portable Document Format, or PDF,Is a file format developed by Adobe to present documents independent of application software, hardware, and operating systems.
- Can be viewed the same way on any device
- Is frequently used in legal and financial documents
- Can also be used to fill in data for forms

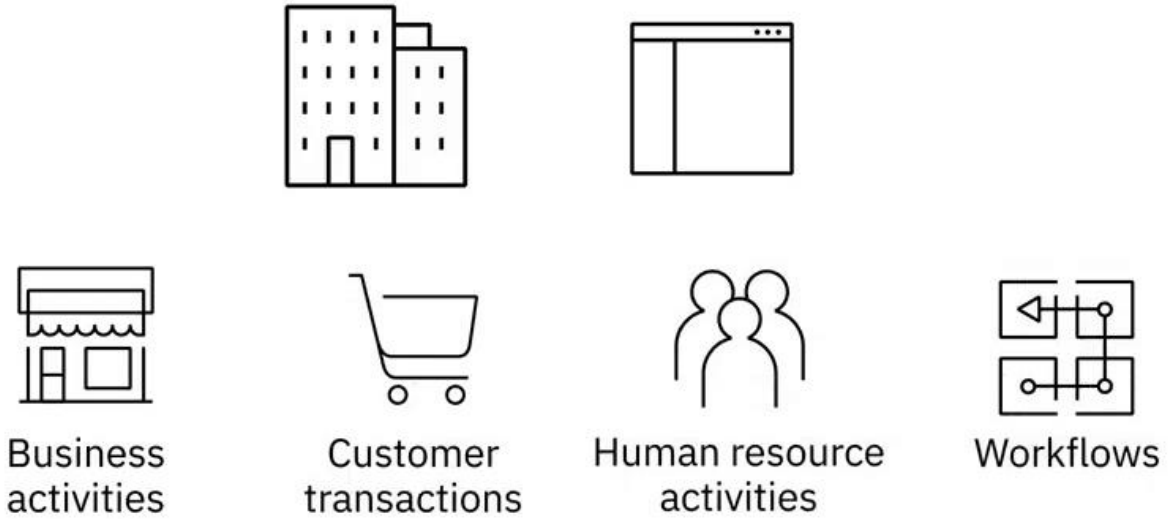


JavaScript Object Notation or JSON

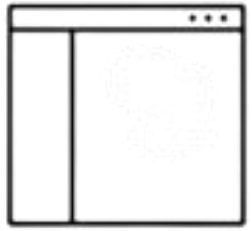
- JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web.
- Language-independent data format
- Can be read in any programming language
- Easy to use
- Compatible with a wide range of browsers
- Considered as one of the best tools for sharing data

```
{
    hey: "guy",
    anumber: 243,
  - anobject: {
        whoa: "nuts",
      - anarray: [
            1,
            2,
            "thr<h1>ee"
        ],
        more: "stuff"
    },
    awesome: true,
    bogus: false,
    meaning: null,
    japanese: "明日がある。",
    link: http://jsonview.com,
    notLink: "http://jsonview.com is great"
}
```

Common sources of data



Relational Database



Microsoft SQL Server

ORACLE

MySQL

IBM DB2

Flat files

- Store data in plain text format
- Each line, or row, is one record
- Each value is separated by a delimiter
- All of the data in a flat file maps to a single table
- Most common flat file format is .CSV

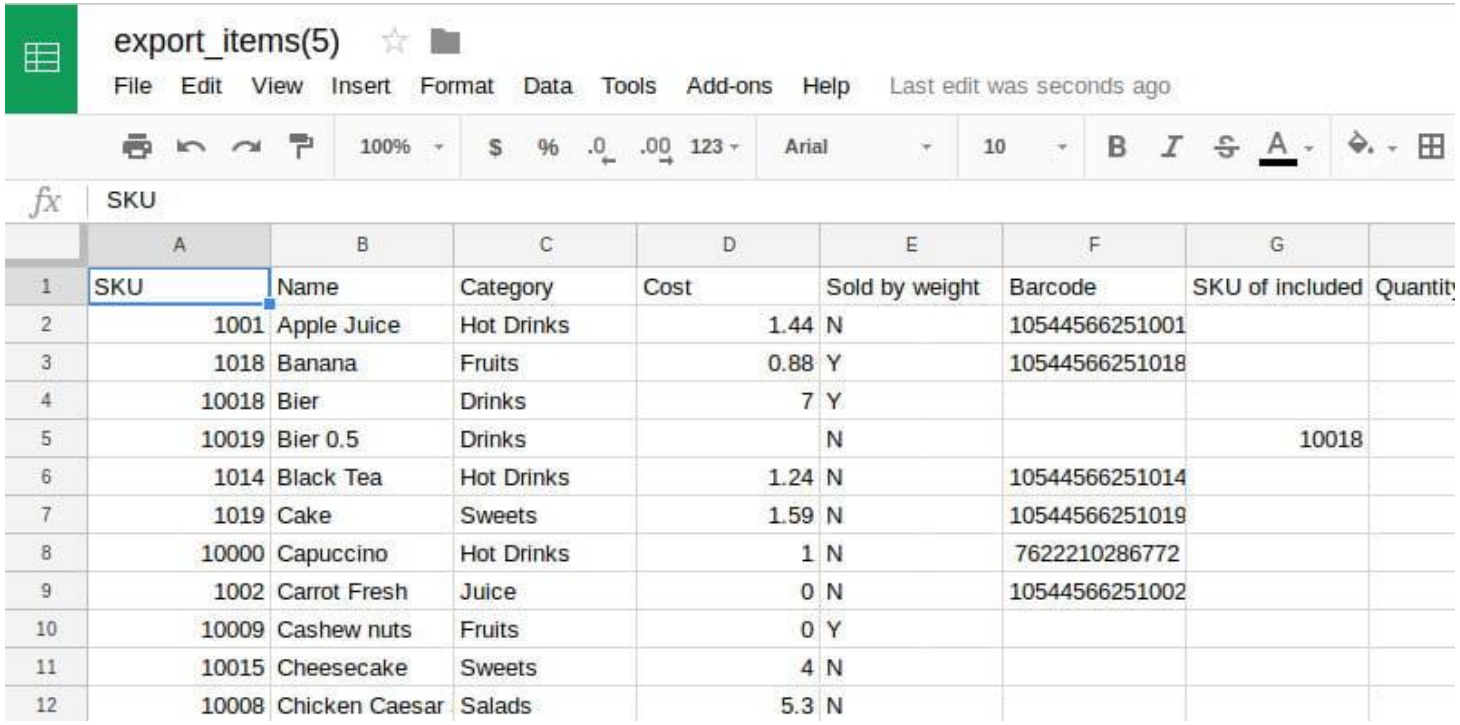
*Untitled - Notepad

File Edit Format View Help

```
"OrderID", "CustomerID", "OrderDate"
"01","001", "06/06/2021"
"02","369", "06/06/2021"
"03","151", "06/06/2021"
"04","014", "06/06/2021"
"05","061", "06/06/2021"
"06","220", "06/06/2021"
```

Spreadsheet files

- Special type of flat files
- Organize data in a tabular format
- Can contain multiple worksheets
- .XLS or .XLSX are common spreadsheet formats
- Other formats include Google Sheets, Apple Numbers, and LibreOffice Calc.



	A	B	C	D	E	F	G	
1	SKU	Name	Category	Cost	Sold by weight	Barcode	SKU of included	Quantit
2	1001	Apple Juice	Hot Drinks	1.44	N	10544566251001		
3	1018	Banana	Fruits	0.88	Y	10544566251018		
4	10018	Bier	Drinks	7	Y			
5	10019	Bier 0.5	Drinks		N		10018	
6	1014	Black Tea	Hot Drinks	1.24	N	10544566251014		
7	1019	Cake	Sweets	1.59	N	10544566251019		
8	10000	Capuccino	Hot Drinks	1	N	7622210286772		
9	1002	Carrot Fresh	Juice	0	N	10544566251002		
10	10009	Cashew nuts	Fruits	0	Y			
11	10015	Cheesecake	Sweets	4	N			
12	10008	Chicken Caesar	Salads	5.3	N			

Popular examples of APIs

- Twitter and Facebook APIs for customer sentiment analysis
- Stock Market APIs for trading and analysis
- Data Lookup and Validation APIs for cleaning and co-relating data

Web Scraping

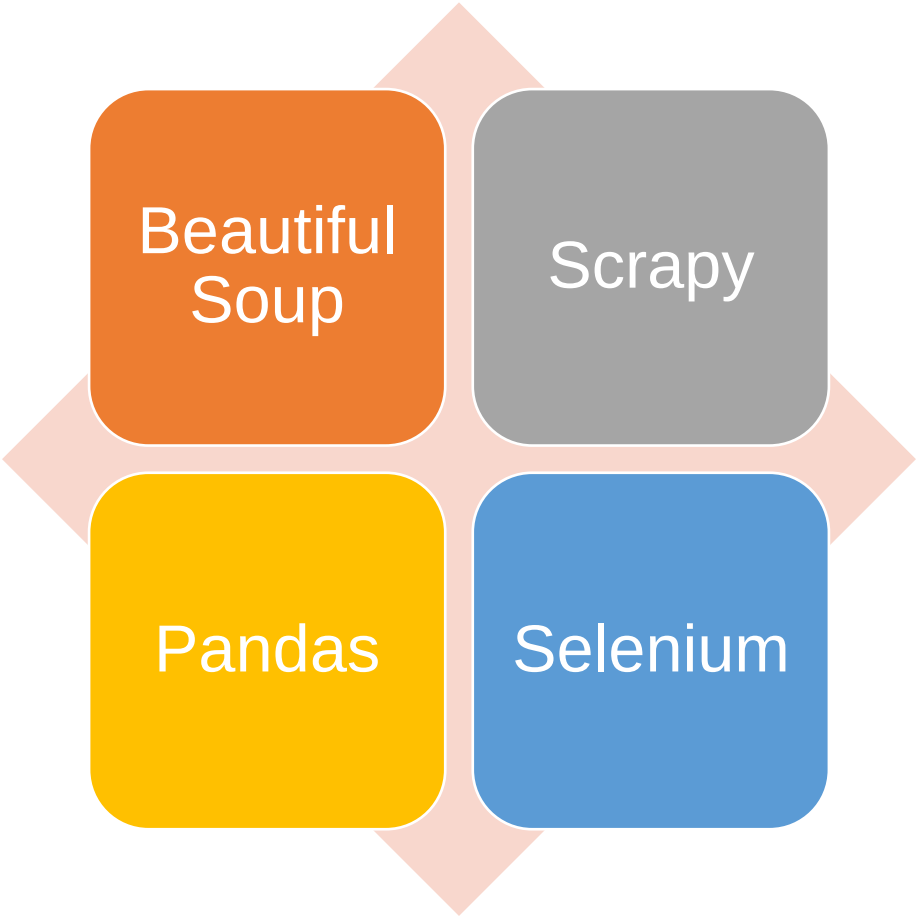
- Extract relevant data from unstructured sources

- Also known as Screen Scraping, Web harvesting, and Web data extraction

- Downloads specific data based on defined parameters

- Can extract text, contact information, images, videos, product items, and more...

Web scraping Popular web scraping tools



Data Streams and feeds

Aggregating streams of data flowing from instruments, IoT devices and applications, GPS data from cars, computer programs, websites, and social media posts

- Stock and market tickers for financial trading
- Retail transaction streams for predicting demand and supply chain management
- Surveillance and video feeds for threat detection

Data Streams and feeds

SOCIAL MEDIA FEEDS FOR SENTIMENT ANALYSIS

SENSOR DATA FEEDS FOR MONITORING INDUSTRIAL OR FARMING MACHINERY

WEB CLICK FEEDS FOR MONITORING WEB PERFORMANCE AND IMPROVING DESIGN

REAL-TIME FLIGHT EVENTS FOR REBOOKING AND RESCHEDULING

Data Streams and feeds

kafka

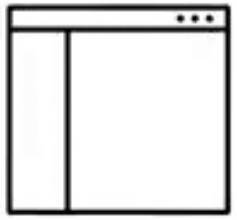
APACHE Spark™

APACHE STORM™

Data Streams and feeds

- RSS (or Really Simple Syndication) feeds Capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.

Online forums



News sites

SQL

- Querying Language designed for accessing and manipulating information from, mostly, though not exclusively, relational databases.
- Using SQL, you can:
  - Insert, update, and delete records in a database
  - Create new databases, tables, and views
  - Write stored procedures

Advantages of using SQL

- SQL is portable and platform independent
- Can be used for querying data in a wide variety of databases and data repositories
- Has a simple syntax that is like the English language
- Its syntax allows developers to write programs with fewer lines of code using basic keywords
- Can retrieve large amounts of data quickly and efficiently
- Runs on an interpreter system

Python

- Python is a widely-used open-source, general-purpose, high-level programming language.
- Its syntax allows programmers to express their concepts in fewer lines of code
- An ideal tool for beginning programmers because of its focus on simplicity and readability
- Great for performing high-computational tasks in large volumes of data
- Has in-built functions for frequently used concepts
- Supports multiple programming paradigms object-oriented, imperative, functional, and procedural

Python

- Its vast array of libraries and functionalities also include:
- Pandas for data cleaning and analysis
- NumPy and SciPy, for statistical analysis
- Beautiful soup and Scrapy for web scraping
- Matplotlib and Seaborn to visually represent data in the form of bar graphs, histogram, and pie-charts
- Opency for image processing

R-Programming

- R is an open-source programming language and environment for data analysis, data visualization, machine learning, and statistics.
- Widely used for:
  - Developing statistical software
  - Performing data analytics
  - Creating compelling visualizations

R Programming

R-Programming

Key benefits:

- Includes libraries such as Ggplot2 and Plotly that offer aesthetic graphical plots to its users
- Allows data and scripts to be embedded in reports
- Allows creation of interactive web apps
- Can be used for developing statistical tools

R Programming

JAVA

- Java is an object-oriented, class-based, and platform-independent programming language originally developed by Sun Microsystems.
- One of the top-ranked programming languages used today
- Used in a number of data analytics processes —cleaning data, importing and exporting data, statistical analysis, data visualization
- Used in the development of big data frameworks and tools — Hadoop, Hive, Spark
- Well-suited for speed-critical projects

Unix/ Linux Shell

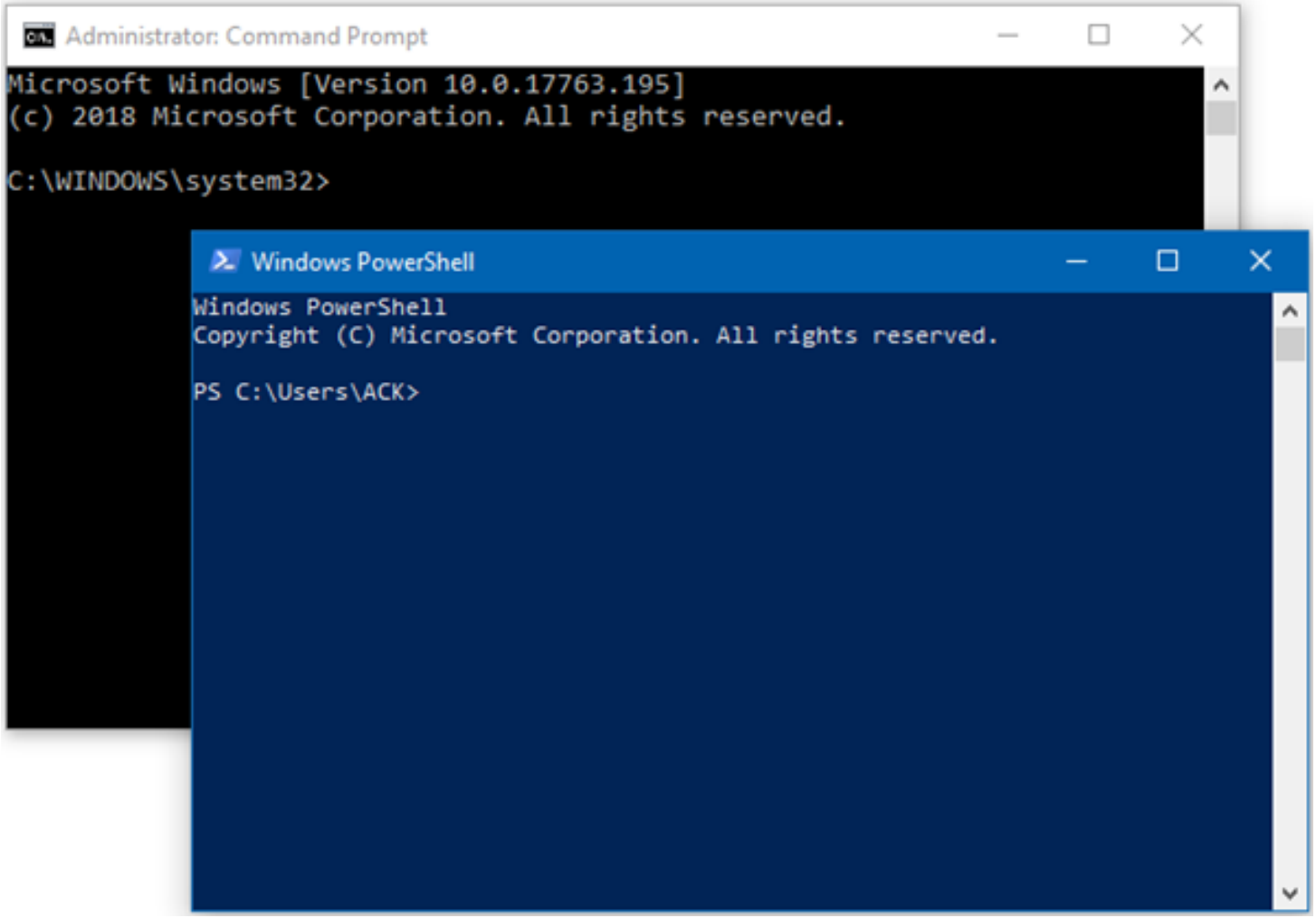
- A Unix/Linux Shell is a computer program written for the UNIX shell. It is a series of UNIX commands written in a plain text file to accomplish a specific task.

Typical operations performed by shell scripts include:

- File manipulation
- Program execution System administration tasks such as disk backups and evaluating system logs
- Installation scripts for complex programs
- Executing routine backups
- Running batches

PowerShell

- PowerShell is a cross-platform automation tool and configuration framework by Microsoft that is optimized for working with structured data formats, such as JSON, CSV, XML, and REST APIs, websites, and office applications.
    - Consists of command-line shell and scripting language
    - Is object-based and can be used to filter, sort, measure, group, and compare objects as they pass through a data pipeline
    - Used for data mining, building GUIs, creating charts, dashboards, and interactive reports



Thank you