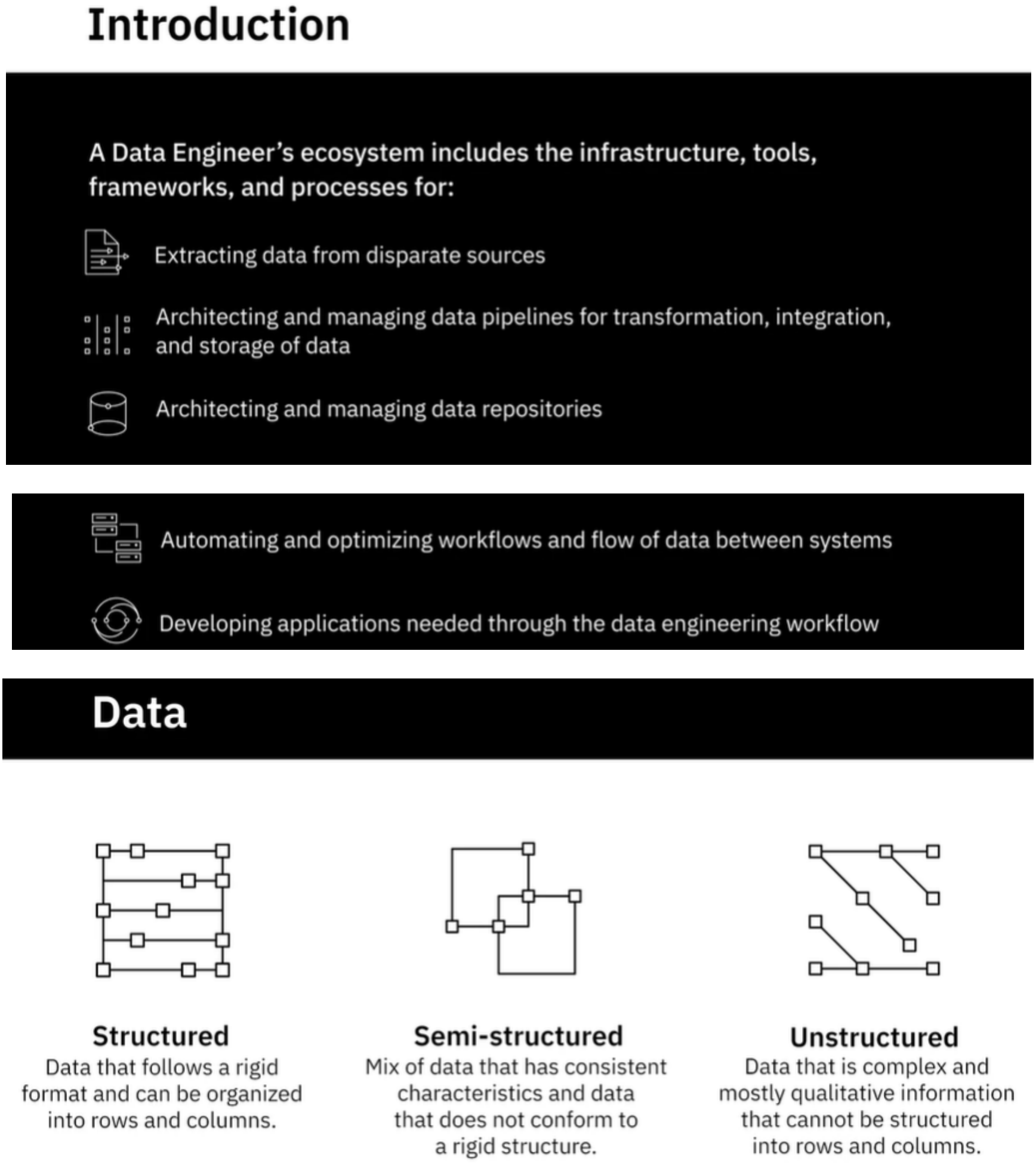
Week 2 - The Data Engineering Ecosystem

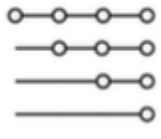
Overview of the Data Engineering Ecosystem



Data

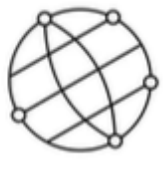
Data also comes in a wide-ranging variety of file formats being collected from a variety of data sources,

Relational Database



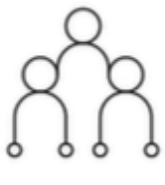
Non-Relational Database

APIs

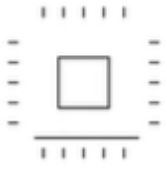


Web Services

Data Streams

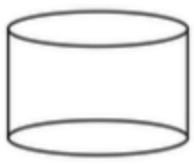


Social Platforms



Sensor Devices

Data Repositories

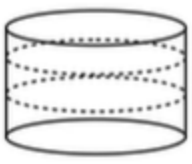


Transactional
or
Online Transaction Processing (OLTP) System

Designed to store high volume day-to-day operational data

Typically relational, but can also be non-relational

Data Repositories

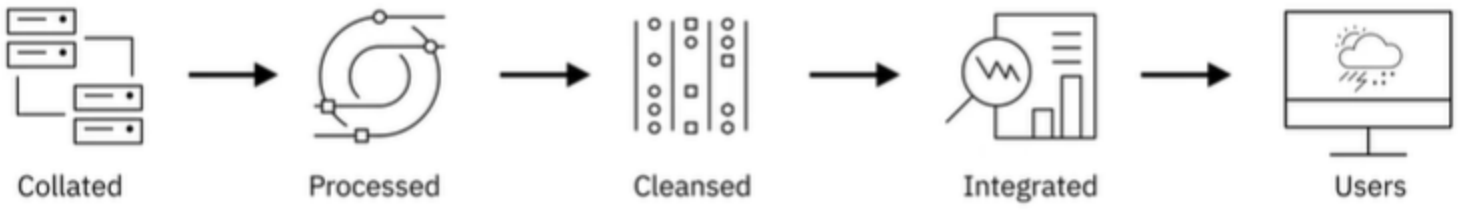


Analytical
or
Online Analytical Processing (OLAP) Systems

Optimized for conducting complex data analytics

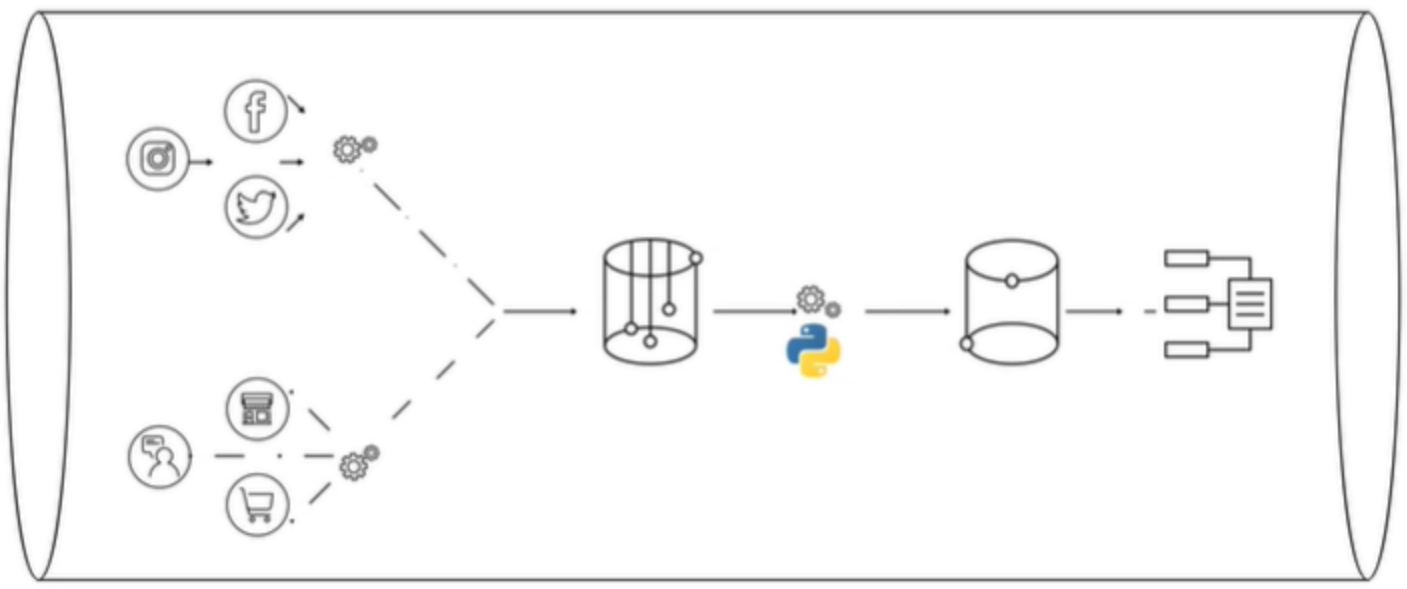
Include relational and non-relational databases, data warehouses, data marts, data lakes, and big data stores

Data Integration



Combine data from disparate sources into a unified view, accessed by users to query and manipulate the data.

Data Pipeline

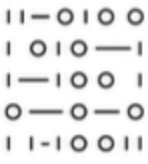


A set of tools and processes that cover the entire journey of data from source to destination systems.

Languages

Languages available in the Data Analyst Ecosystem:

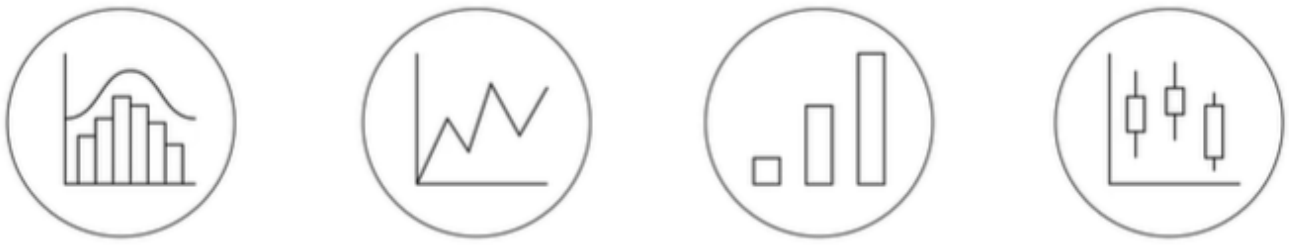
Query languages
For example, SQL for querying and manipulating data



Programming languages
For example, Python for developing data applications

Shell and Scripting languages
For repetitive operational tasks

Business Intelligence (BI) and Reporting Tools



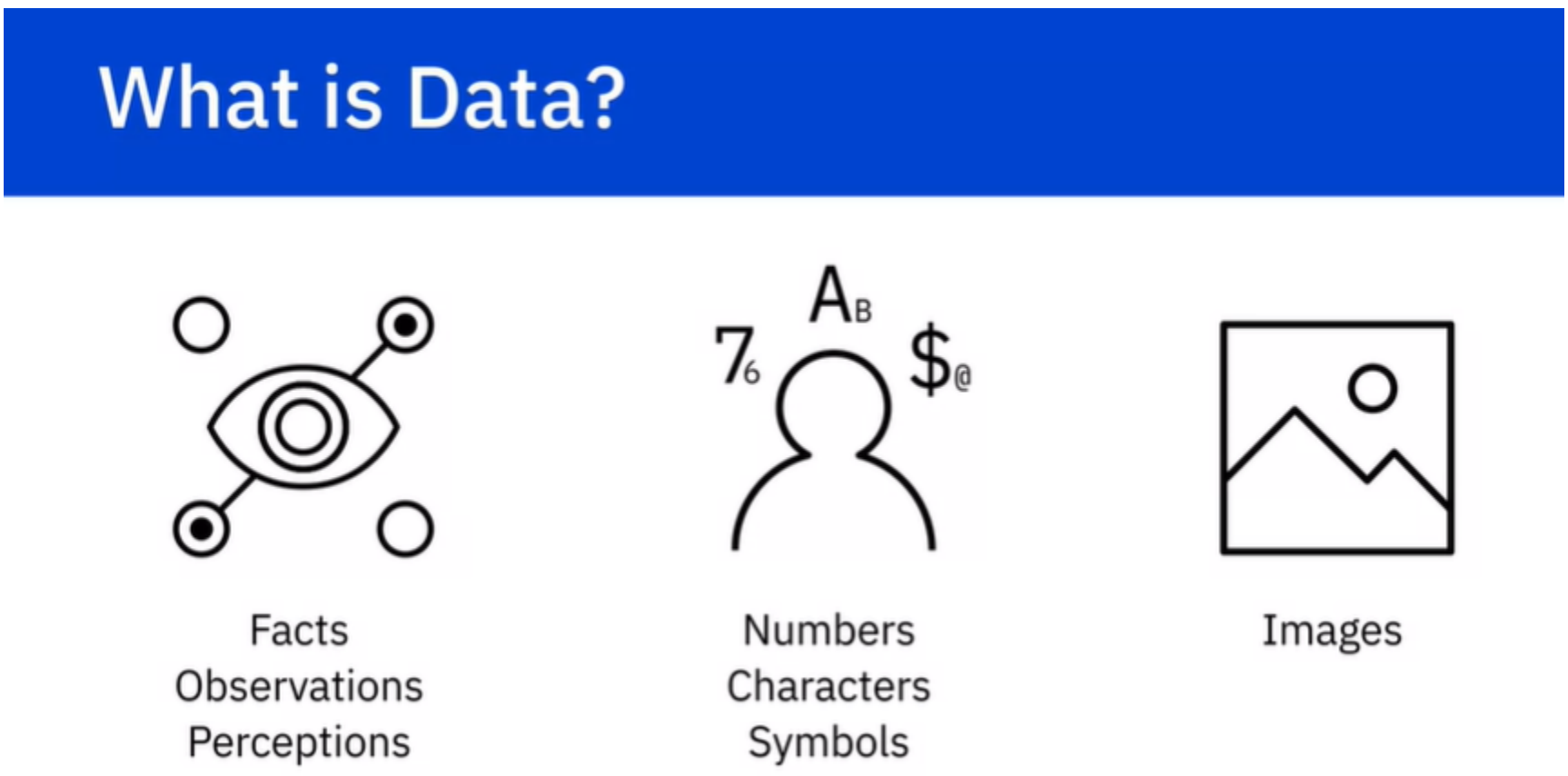
- Collect data from multiple data sources and present them in a visual format, such as interactive dashboards
- Visualize data in real-time and pre-defined schedule
- Drag and drop products that do not require knowledge of programming

Conclusion

Automated tools, frameworks, and processes for all stages of the data analytics process are part of the Data Engineer’s ecosystem.

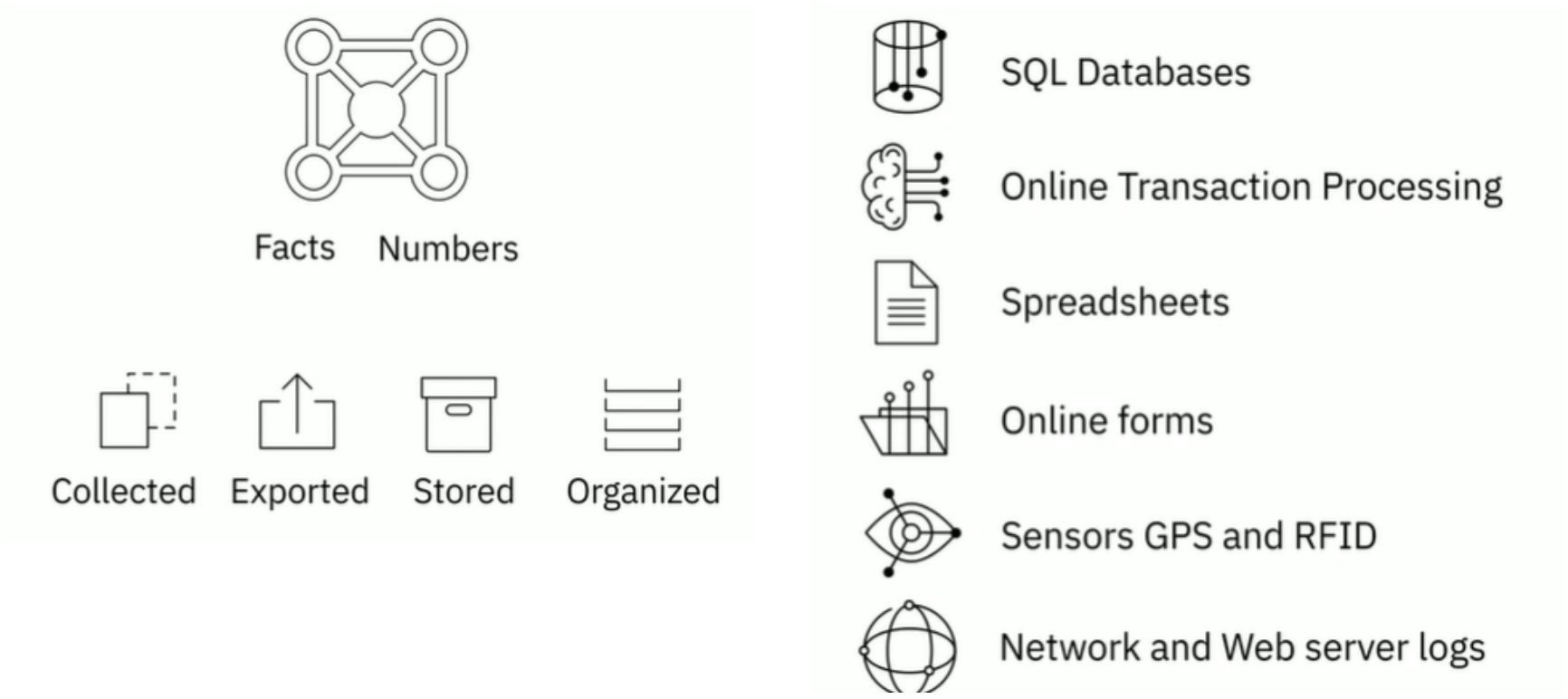
It‘s a diverse, rich, and challenging ecosystem.

Types of Data



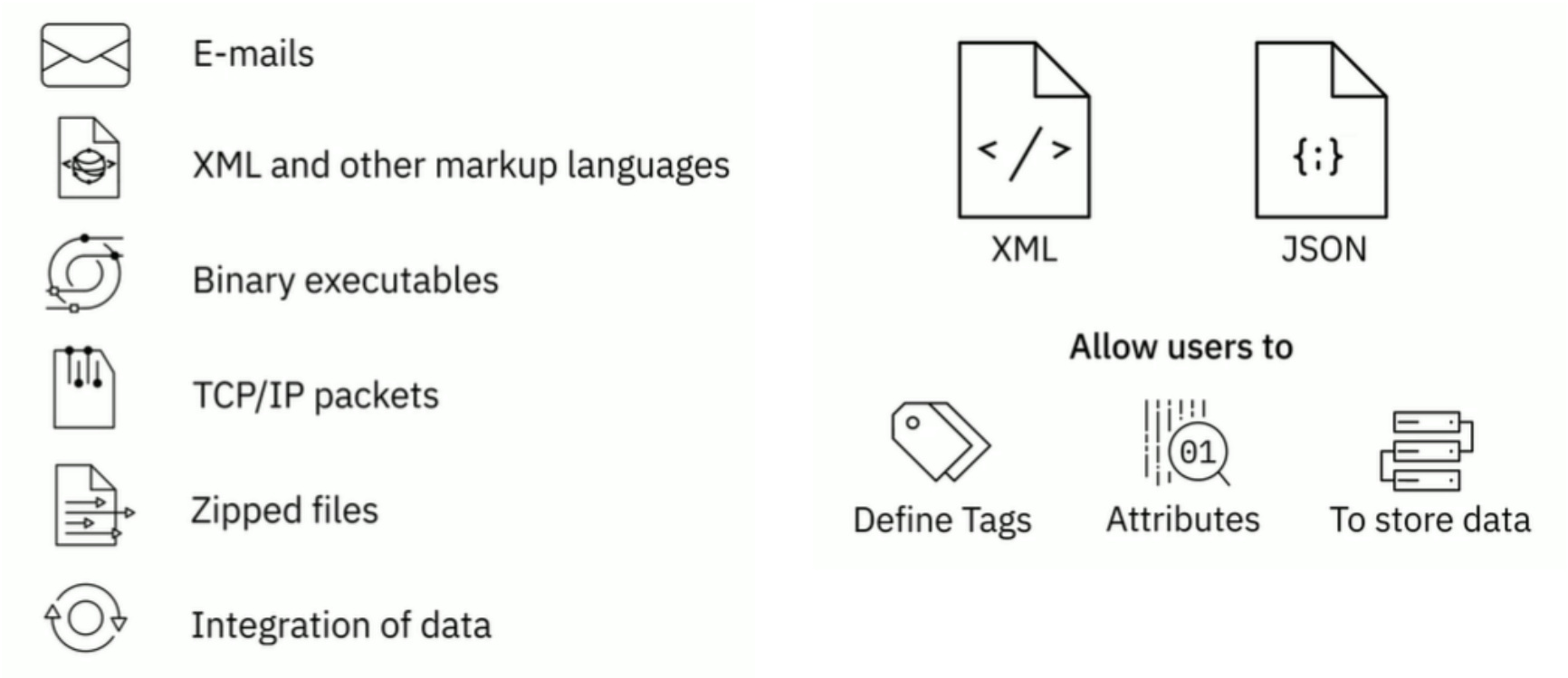
Structured data

- Has a well-defined structure
- Can be stored in well-defined schemas
- Can be represented in a tabular manner with rows and columns



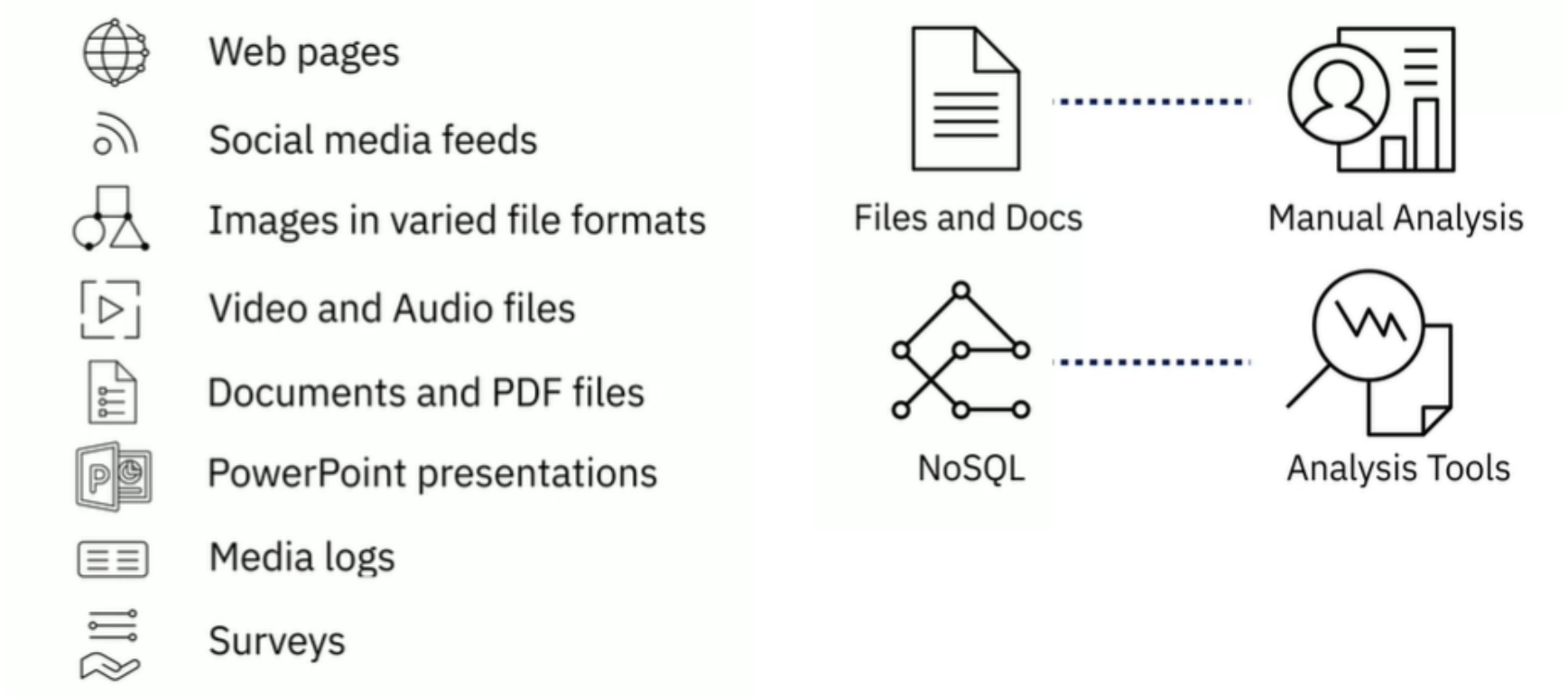
Semi-Structured data

- Has some organizational properties but lacks a fixed or rigid schema
- Cannot be stored in the forms of rows and columns as in databases
- Contains tags and elements, or metadata, which is used to griup data and organize it in a hierarchy



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 particular format, sequence, semantics, or rules



Conclusion

Structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools;

Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy;

Unstructured data is data that is not conventionally organized in the form of rows and columns in a particular format. In the next video, we will learn about the different types of file structures.

Understanding Different Types of File Formats

Standard file formats:

1. Delimited text file formats, or .CSV
2. Microsoft Excel Open .XML Spreadsheet, or .XLSX
3. Extensible Markup Language, or .XML
4. Portable Document Format, or .PDF
5. JavaScript Object Notation, or .JSON

Delimited 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

CSV

Comma-separated values

TSV

Tab-separated values

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

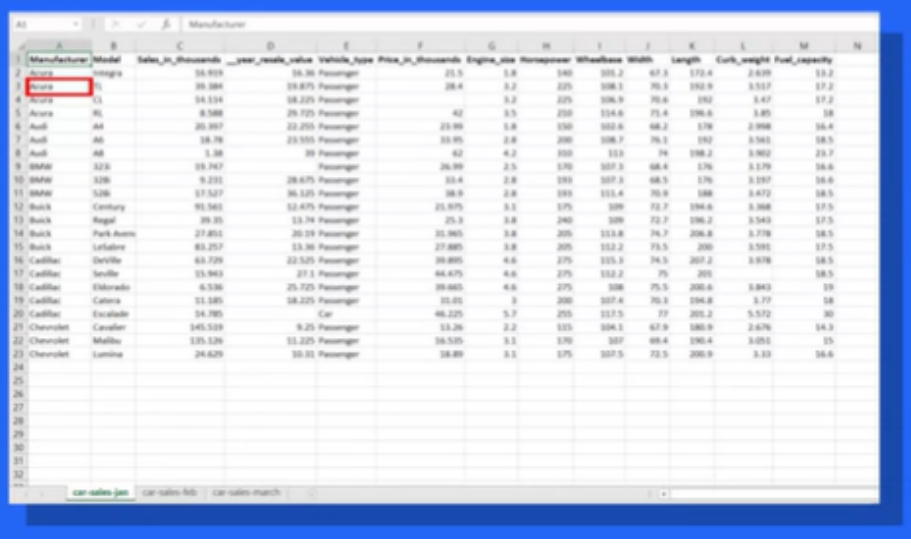
Manufacturer	Model	Sales_in_thousands	__year_resale_value
Acura Integra	16.919	16.36 Passenger	21.5
Acura TL	39.384	19.875 Passenger	28.4
Acura CL	14.114	18.225 Passenger	
Acura RL	8.588	29.725 Passenger	42
Audi A4	20.397	22.255 Passenger	23.99
Audi A6	18.78	23.555 Passenger	33.95
Audi A8	1.38	39 Passenger	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

Delimiters also represent one of various means to specify boundaries in a data stream

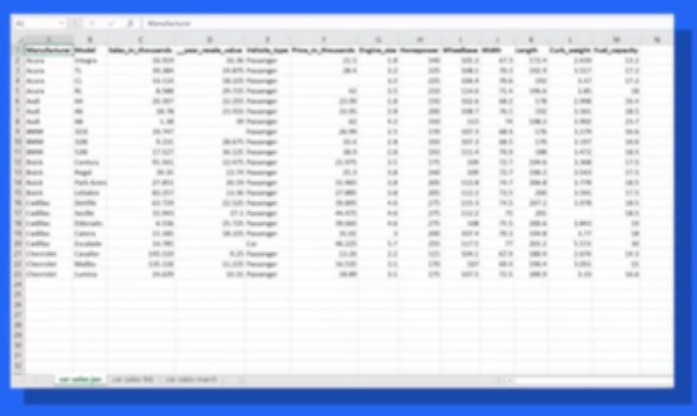
Microsoft Excel Open XML Spreadsheet, or .XLSX

Is a Microsoft Excel Open XML file format that falls under the spreadsheet file format. It is an XML-based file format created by Microsoft.

Cell



Microsoft Excel Open XML Spreadsheet, or .XLSX



- Open file format, accessible to most other applications
- Can use and save all functions available in excel
- Is a secure file format as it cannot save malicious code

Extensible Markup Language or .XML

```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura<manufacturer>

<model>Integra<model>

<sales_in-thousands>16.919<sales_in-thousands>

<year_resale_value>16.36<year_resale_value>

<vehicle_type>Passenger<vehicle_type>

<car-specs>
```

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

- Readable by both humans and machines
- Self-descriptive language
- Similar to .HTML in some respects
- Does not use predefined tags like .HTML does
- Platform independent
- Programming language independent

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

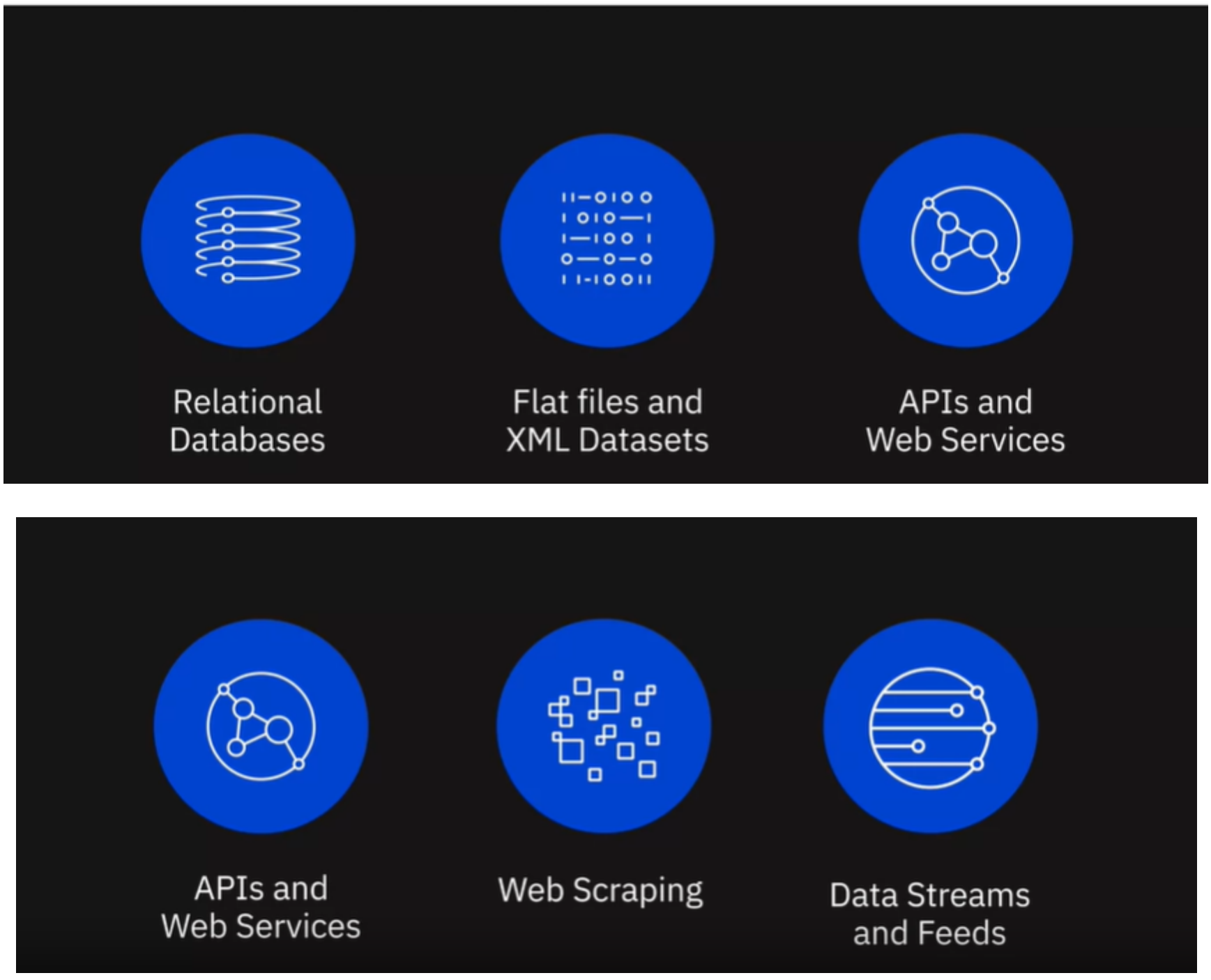
```
{
    "Employee": [
        {
          "id":"1",
          "Manufacturer": "Audi",
          "Model": "Integra",
        },
        {
          "id":"2",
          "Manufacturer": "Buick",
          "Model": "LeSabre",
        },
        {
          "id":"3",
          "Manufacturer": "Cadillac",
          "Model": "Escalade",
        }
    ]
}
```

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

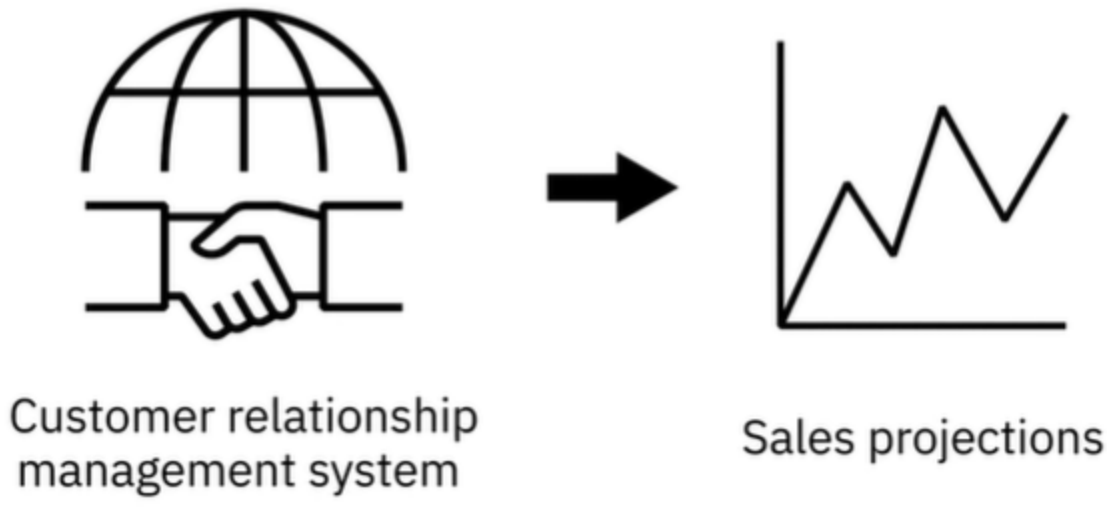
Sources of Data

Common sources of data:



Relational Databases

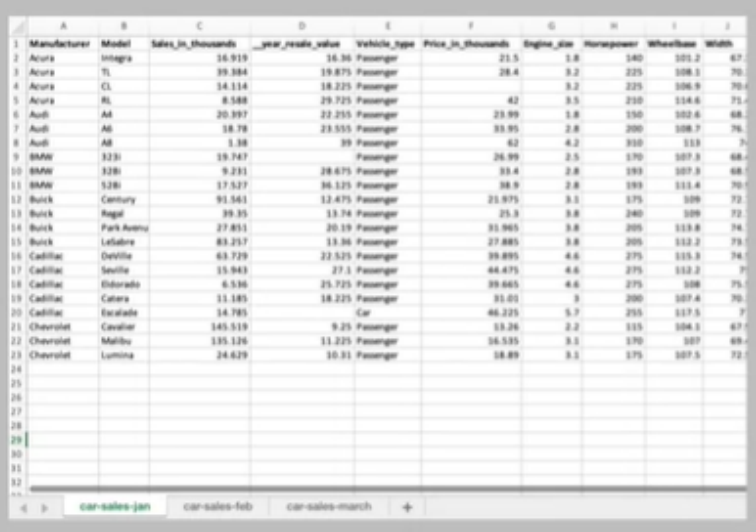
Store structured data that can be leveraged for analysis



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

Spreadsheet files



	A	B	C	D	E	F	G	H	I	J
1	Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands	Engine_size	Horsepower	Wheelbase	Width
2	Acura	Integra	16.919	16.36	Passenger	21.5	1.8	140	101.2	67
3	Acura	TL	39.384	19.875	Passenger	28.4	3.2	225	108.1	70
4	Acura	CL	14.114	18.225	Passenger		3.2	225	106.9	70
5	Acura	RL	8.588	29.725	Passenger	42	3.5	210	114.6	71
6	Audi	A4	20.397	22.255	Passenger	23.99	1.8	150	102.6	68
7	Audi	A6	18.78	23.555	Passenger	33.95	2.8	200	108.7	76
8	Audi	A8	1.38	39	Passenger	62	4.2	310	113	74
9	BMW	323i	19.747		Passenger	26.99	2.5	170	107.3	68
10	BMW	328i	9.231	28.675	Passenger	33.4	2.8	193	107.3	68
11	BMW	528i	17.527	36.125	Passenger	38.9	2.8	193	111.4	70
12	Buick	Century	91.561	12.475	Passenger	21.975	3.1	175	109	72
13	Buick	Regal	39.35	13.74	Passenger	25.3	3.8	240	109	72
14	Buick	Park Avenu	27.851	20.19	Passenger	31.965	3.8	205	113.8	74
15	Buick	LeSabre	83.257	13.36	Passenger	27.885	3.8	205	112.2	73
16	Cadillac	DeVille	63.729	22.525	Passenger	39.895	4.6	275	115.3	74
17	Cadillac	Seville	15.943	27.1	Passenger	44.475	4.6	275	112.2	7
18	Cadillac	Eldorado	6.536	25.725	Passenger	39.665	4.6	275	108	75
19	Cadillac	Catera	11.185	18.225	Passenger	31.01	3	200	107.4	70
20	Cadillac	Escalade	14.785		Car	46.225	5.7	255	117.5	7
21	Chevrolet	Cavalier	145.519	9.25	Passenger	13.26	2.2	115	104.1	67
22	Chevrolet	Malibu	135.126	11.225	Passenger	16.535	3.1	170	107	69
23	Chevrolet	Lumina	24.629	10.31	Passenger	18.89	3.1	175	107.5	72

car-sales-jan | car-sales-feb | car-sales-march

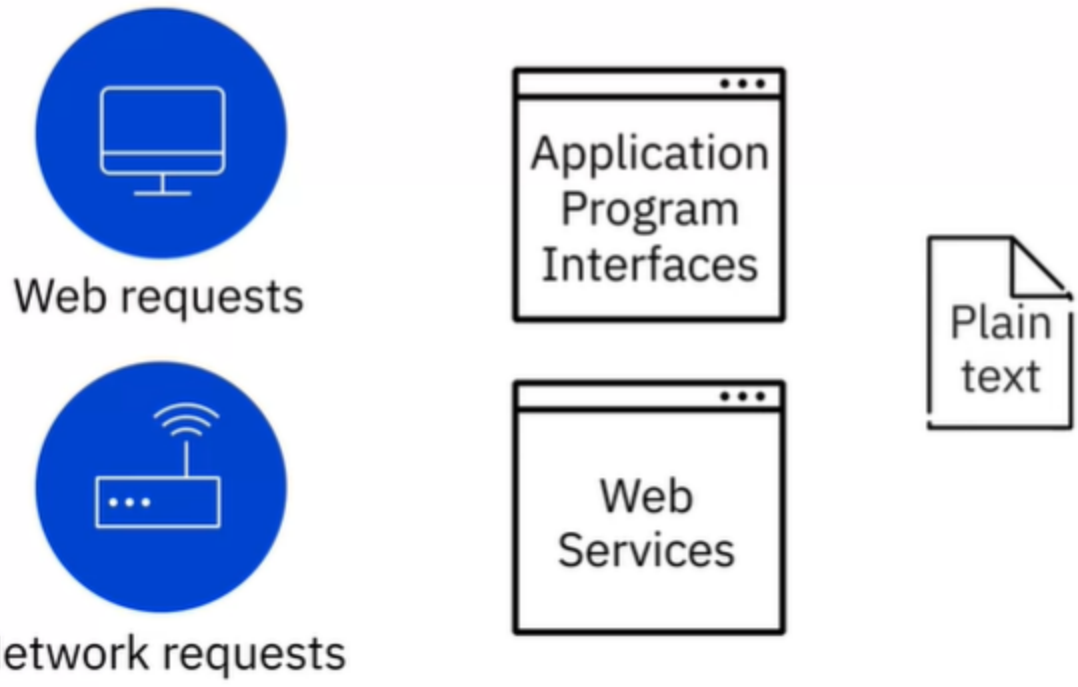
- Special type of flat files
- Organize data in a tabular format
- Can contain multiple worksheets
- .XLS or .XLSX are common spreadsheet formats

XML files

- Contain data values that are identified or marked up using tags
- Can support complex data structures
- Common uses include online surveys, bank statements, and other unstructured data sets

```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura<manufacturer>

<model>Integra<model>

<sales_in-thousands>16.919<sales_in-thousands>

<year_resale_value>16.36<year_resale_value>

<vehicle_type>Passenger<vehicle_type>

<car-specs>
```

APIs and Web Services



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 uses:

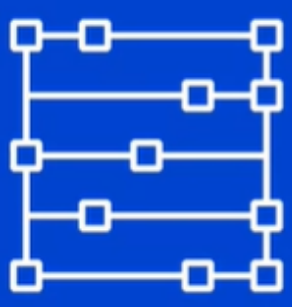
- Providing price comparisons by collecting product details from retailer, manufacturers, and eCommerce websites
- Generating sales leads through public data sources
- Extracting data from posts and authors on various forums and communities
- Collecting training and testing datasets for machine learning models

Web scraping

Popular web scraping tools:

- BeautifulSoup
- Scrapy
- Pandas
- Selenium

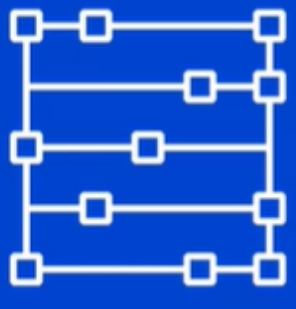
Data Streams and feeds



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

Languages for Data Professionals

Introduction

- Query languages are designed for accessing and manipulating data in a database (SQL)
- Programming languages are designed for developing applications and controlling application behavior (Python, R, Java)
- Shell and Scripting languages are ideal for repetitive and time-consuming operational tasks (Unix/Linux Shell, PowerShell)

SQL

SQL, or Structured Query Language, is a querying language designed for accessing and manipulating information from, mostly, though not exclusively, relational databases.

SQL

Using SQL, you can:

- Insert, update, and delete records in a database
- Create new databases, tables, and views
- Write stored procedures

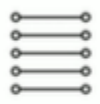
SQL

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 similar to 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

Python

Python is a widely-used open-source, general-purpose, high-level programming language.

Has in-built functions for frequently used concepts

Supports multiple programming paradigms – object-oriented, imperative, functional, and procedural

Python

Python is one of the fastest-growing programming languages in the world.

- Easy to learn
- Open-source
- Can be ported to multiple platforms
- Has widespread community support
- Provides open-source libraries for data manipulation, data visualization, statistics, mathematics

Python

Its vast array of libraries and functionalities also include:

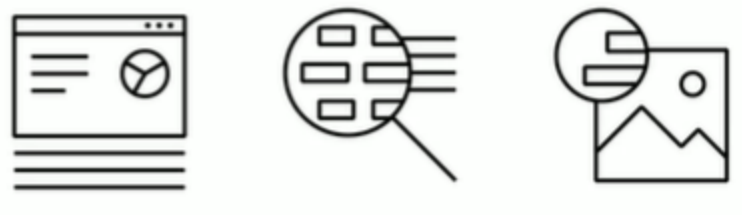
- Pandas for data cleaning and analysis
- Numpy and Scipy, for statistical analysis
- Beautifulsoup and Scrapy for web scraping
- Matplotlib and Seaborn to visually represent data in the form of bar graphs, histogram, and pie-charts
- Opency for image processing

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

Key benefits:

- Open-source
- Platform-independent
- Can be paired with many programming languages
- Highly extensible
- Facilitates the handling of structured and unstructured data

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

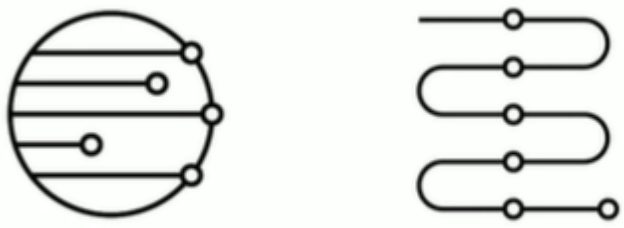
Java

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

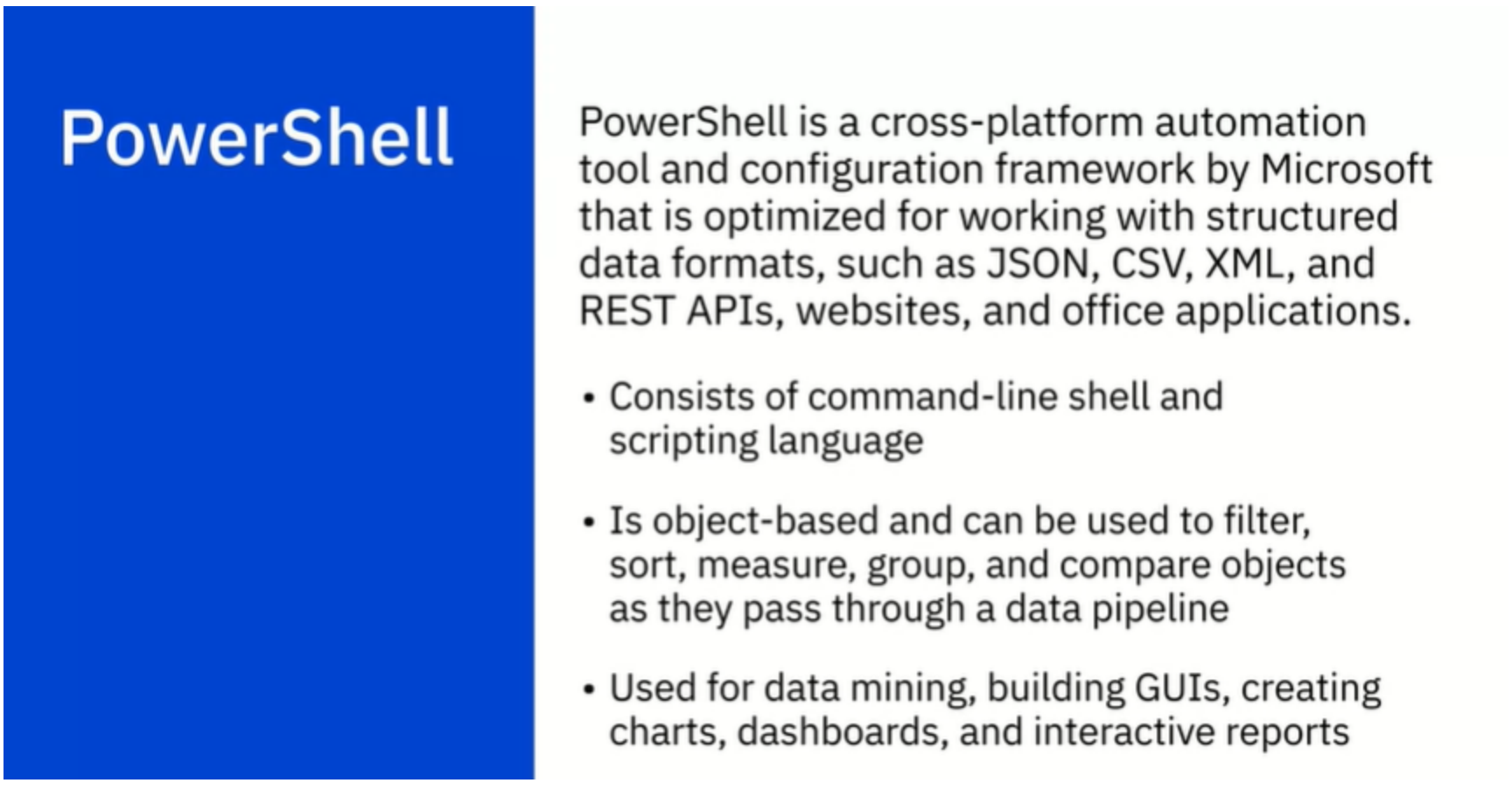
A Unix/Linux Shell is a computer program written for the UNIX shell. It is a series of UNIX commands written in a plain text file to accomplish a specific task.



Unix/ Linux Shell

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



Reading: Metadata and Metadata Management

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0100EN-SkillsNetwork/readings/Reading_Metadata_and_Metadata_Management.md.html?origin=www.coursera.org

Summary and Highlights

A Data Engineer’s ecosystem includes the infrastructure, tools, frameworks, and processes for extracting data, architecting and managing data pipelines and data repositories, managing workflows, developing applications, and managing BI and Reporting tools.

Based on how well-defined the structure of the data is, data can be categorized as

- Structured data, that is data which is well organized in formats that can be stored in databases.
- Semi-structured data, that is data which is partially organized and partially free-form.
- Unstructured data, that is data which can not be organized conventionally into rows and columns.

Data comes in a wide-ranging variety of file formats, such as, delimited text files, spreadsheets, XML, PDF, and JSON, each with its own list of benefits and limitations of use.

Data is extracted from multiple data sources, ranging from relational and non-relational databases, to APIs, web services, data streams, social platforms, and sensor devices.

Once the data is identified and gathered from different sources, it needs to be staged in a data repository so that it can be prepared for analysis. The type, format, and sources of data influence the type of data repository that can be used.

Data professionals need a host of languages that can help them extract, prepare, and analyse data. These can be classified as:

- Querying languages, such as SQL, used for accessing and manipulating data from databases.
- Programming languages such as Python, R, and Java, for developing applications and controlling application behavior.
- Shell and Scripting languages, such as Unix/Linux Shell, and PowerShell, for automating repetitive operational tasks.

Quiz

Practice Quiz

Question 1

Automated tools, frameworks, and processes for all stages of the data analytics process are part of the Data Engineer's ecosystem. What role do data integration tools play in this ecosystem?

- Store high-volume day-to-day operational data in data repositories
- Cover the entire journey of data from source to destination
- **Combine data from multiple sources into a unified view that is accessed by data consumers to query and manipulate data**
- Conduct complex data analytics

Question 2

Which of these data sources is an example of semi-structured data?

- Documents
- Social media feeds
- **Emails**
- Network and web logs

Question 3

Which one of the provided file formats is commonly used by APIs and Web Services to return data?

- XML
- Delimited file
- **JSON**
- XLS

Question 4

What is one example of the relational databases
discussed in the video?

- Spreadsheet
- XML
- Flat files
- **SQL Server**

Question 5

Which of the following languages is one of the most popular querying languages in use today?

- R
- **SQL**
- Java
- Python

Graded Quiz

Question 1

There are two main types of data repositories – Transactional and Analytical. For high-volume day-to-day operational data such as banking transactions, Transactional, or OLTP, systems are the ideal choice.

- **True**
- False

Transactional, or OLTP, systems are designed and optimized for handling high-volume transactions.

Question 2

Which of the following is an example of unstructured data?

- Zipped files

- **Video and Audio files**
- XML
- Spreadsheets

Question 3

Which one of these file formats is independent of software, hardware, and operating systems, and can be viewed the same way on any device?

- XML
- XLSX
- **PDF**
- Delimited text file

PDF format is independent of software, hardware, and operating systems, and can be viewed the same way on any device.

Question 4

Which data source can return data in plain text, XML, HTML, or JSON among others?

- **APIs**
- Delimited text file
- XML
- PDF

APIs can return data in a wide variety of formats such as plain text, XML, HTML, or JSON among others.

Question 5

In the data engineer’s ecosystem, languages are classified by type. What are shell and scripting languages most commonly used for?

- Manipulating data
- Building apps
- **Automating repetitive operational tasks**
- Querying data