The Data Ecosystem and Languages for Data Professionals

A Data Analyst's Ecosystem includes the infrastructure, software tools, frameworks, and processes used to Gather data, Clean Data, Mine Data, Visualize Data

Data

- Structured: Data follows a rigid format and can be organized into rows and colums. Ex: databases and spreadsheets.
- Semi-structured: mix of data that has consistent characteristics and data that does not conform to a rigid structure. Ex: email
- Unstructured: Data that is complex and mostly qualitative information that cannot be structured into rows and columns. Ex: photos, videos, pdf, text files, and social media content.

Data can come in a variety of file formats, such as: Relational Databases, Non-Relational Databases, APIs, Web Services, Data Streams, Social Platforms, Sensor Devices.

Data Repositories

- Databases
- Data Warehouse
- Data Marts
- Data Lakes
- Big Data Stores

Languages available in the data analysist ecosystems:

- Query Languages: SQL for querying and manipulating data
- Programming languages: Python for developing data applications
- Shell and Scripting languages: for repetitive operational tasks

Data Analysist Ecosystems

- Gathering, extracting, transforming, and loading data
- Data wrangling and cleaning
- Data analysis and mining
- Data visualization

Data

Structured data

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

Structured data is objective facts and numbers that can be collected, exported, stored, and organized in typical databases

Includes

- SQL Databases
- Online Transaction Processing
- Spreadsheets
- Online forms
- Sensor GPS and RFID tags
- Network and web server logs

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

Includes:

- E-mail
- XML and other markup languages
- Binary executables
- TCP/IP Packets
- Zipped files
- Integration of data

XML and JSON allow users to define tags and attributes to store data in a hierarchical form and are used widely to store and exchange semi-structured data

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

Unstructured data can deal with heterogeneity of sources and has a variety of business intelligence and analytics applications

Includes

- Web pages
- Social media feeds
- Images in varied file formats
- Video and audio files
- Documents and PDF files
- Power point presentations
- Media logs
- Surveys

Unstructured data can be stored in files and documents (such as a Word doc) for manual analysis or in NoSQL databases that have their own analysis tools for examining this type of data

To summarize

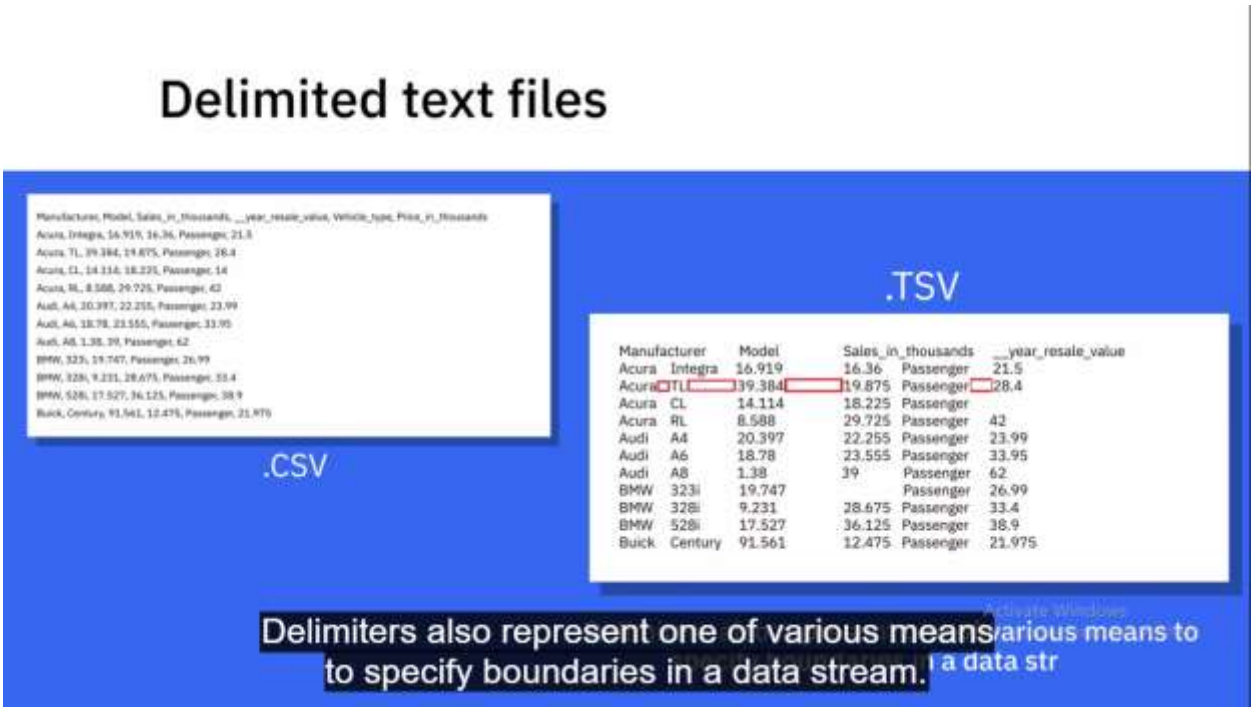
- Structured data is data that is well organized in formats that can be stored in databases and lends itself to standard data analysis methods and tools
- Semi-structured data is data that is somewhat organized and relies on meta tags for grouping and hierarchy
- Unstructured data is data that is not conventionally organized in the form of rows and columns in a particular format

Different types of file formats

Standard file formats:

1. Delimited text file formats, or. CLV

- Filed 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
- Most common delimiters are the Comma, Tab, Colon, Vertical Bar, and Space
- Comma-separated values (CSVs) and Tab-separated values (TSVs) are the most commonly used file types in this category



2. 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

3. Extensible Markup Language, or. XML

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
- Makes it simpler to share data between system

4. Portable Document Format, or .PDF

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

5. JavaScript Object Notation, or JSON

JSON is a text-based open standard designed for transmitting structured data over the web

- Languages-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
- That is a reason, many APIs and Web Services return data as JSON

Sources of Data

Common Sources of data

1. Relational Databases: the organizations have internal applications to support them in managing their day to day, such as Business activities, customer transactions, Human resources activities, workflows. These system uses relational databases such as SQL Server, Oracle, MySQL, and IBM DB2, to store data in a structured way.
    - For example, data from retail transactions system can be used to analyze sales in different regions
2. Flat file and XML Databases: external to the organization, there are other publicly and privately available datasets. Ex Government releasing demographic and economic datasets, Point of sale, financial, weather.
    - Which businesses can use to define strategy, predict demand, and make decisions related to distribution or marketing promotions.
    - Such data sets: Flat files, spreadsheets files, or xml documents

➢ 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

➢ Spreadsheets files

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

- XML Files
  - Contain data values that are identified and mark up using tags
  - Can support complex data structures
  - Common uses include online surveys, bank statements and other unstructured data set
- APIs (Application Program Interfaces) and Web Services
  - Which multiple users or applications can interpret with and obtain data for processing or analysis
  - Listen for incoming request, which can be in the form of web requests from users or network requests from applications and return data in plain text, XML, HTML, JSON, or Media files.
  - 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-relation data
    - APIs are also used for pulling data from database sources, within and external to organization.
- Web Scrapping
  - Extract relevant data from unstructured sources
  - Also known as Screen scrapping, Web harvesting, and Web data extraction
  - Download specific data based on defined parameters
  - Can extract text, contract information, images, videos, product items, and more from a website
  - Popular uses:

        - Providing prices 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
    - Popular web scrapping 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
        - 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
        - Popular technology:
            - Apache Kafka
            - Apache Spark
            - Apache STORM
        - RSS (Really Simple Syndication) feeds: capturing updated data from online forums and news sites where data is refreshed on an ongoing basis.
            - Online forums
            - New sites

Languages for Data Professionals

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

1. SQL

- SQL (Structured Query Languages) is a 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 produces-which means you can write a set of instructions and call them for later use
- Advantages 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 languages.
    - Its syntax allows developers to write programs with fewer lines of code using basic keywords such as select, insert, into and updated
    - Can retrieve large amounts of data quickly and efficiently
    - Runs on an interpreter system

2. Python

- Python is a widely-used open-source, general-purpose, high-level programming language.
- Its syntax allows programmers to express their concept 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
- Support multiple programming paradigms-object-oriented, imperative, functional, and procedural
- 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 scrapping
- Matplotlib and Seaborn to visually represent data in the form of the bar graphs, histogram, and pie-charts
- Opency for image processing

3. 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
- Benefits:
    - Open sources
    - Platform independent
    - Can be paired with many programming languages
    - Highly extensible
    - Facilities the handling of structured and unstructured data
    - Includes libraries such as Ggplot2 and Plotly that offer aesthetic graphical plots to its user
    - Allows data and scripts to be embedded in reports
    - Allows creation and interactive web apps
    - Can be used for developing statistical tools

4. 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, and data visualization

- Used in the development of big data frameworks and tools – Hadoop, Hive, Spark
- Well-suited for speed-critical projects

5. 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.
- It is most useful for repetitive tasks that may be time-consuming to execute by typing
- 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

6. 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, dashboard, and interactive reports

Summary and Highlights

In this lesson, you have learned the following information:

A data analyst ecosystem includes the infrastructure, software, tools, frameworks, and processes used to gather, clean, analyze, mine, and visualize data.

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 analyze 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.