The Data Ecosystem and Language for Data Professionals

Overview of the Data Analyst Ecosystem:

- A Data Analyst's ecosystem includes the infrastructure, software, tools, frameworks and processes to gather, clean, mine and visualize the data.
- Data is unorganized information that is processed to make meaningful. Generally, data comprises of facts, observations, perceptions, numbers, characters, symbols and images that can be interpreted to derive meaning.

Types of Data:

- Based on how well defined the structure of the data is, it can be categorized as ,-
  (i) **Structured-** data that follows rigid format and can be organized into rows and columns.
  (ii) **Semi-structured**- mix of data that has consistent characteristics and data that doesn't conform to rigid structure. For e.g., Emails- it has mix of structured (name of sender and recipient) and unstructured (contents of the email) data.
  (iii) **Unstructured-** data that is complex and mostly qualitative information that can't be stored as rows and columns. For e.g., photos, videos, pdf etc.

Understanding Different Types of File Formats:

- **Delimited Text Files-** files used to store data as text. Each value is separated by a delimiter (comma, tab etc.)
- **XLSX Files-** It falls under spreadsheet file format.
- **Extensible Markup Language (XML)-** It is a markup language that sets rules for encoding data. It is readable by both humans and machines. It is platform independent and programming language independent, making it simpler to share data between various systems.
- **Portable Document Format (PDF)-** a file format to present documents independent of application software, hardware and operating systems.
- **Java Script Object Notation (JSON)-** It is a text based open standard designed for transmitting structured data over the web and it is language independent.

Sources of Data:

- **Relational Databases-** Organizations have internal applications to support them in managing their day to day business activities, customer transactions, workflows. These systems use relational databases such as MySQL server to store data in structured way which can be a source of analysis.
- **Flat Files-** There are companies that sell specific data that businesses can use to define strategy, predict demand etc. Such data are typically made available as flat files, spreadsheet files or XML documents. One of the most common flat file formats is CSV.
- **Application Program Interface-** Many data providers and websites provide API's and web services. API and web services typically listen for incoming requests, which can be in the form of web requests from users or computer networks from applications and return data in plain text, XML, HTML, JSON or media files.

Language for Data Professionals:

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

  Power shell 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.