

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.