**Data Analytics Basics for Everyone**

**MODULE 1: 8th August 2021**

Data is available in a variety of structured and unstructured datasets residing in text, images, videos, clickstreams, user conversations, social media platforms, the Internet of things.

A modern data ecosystem includes a network of interconnected and continually evolving entities that include:

* + Data that is available in a host of different formats, structure, and sources.
  + Enterprise Data Environment in which raw data is staged so it can be organized, cleaned, and optimized for use by end-users.
  + End-users such as business stakeholders, analysts, and programmers who consume data for various purposes.

Emerging technologies such as Cloud Computing, Machine Learning, and Big Data, are continually reshaping the data ecosystem and the possibilities it offers.

Data Engineers, Data Analysts, Data Scientists, Business Analysts, and Business Intelligence Analysts, all play a vital role in the ecosystem for deriving insights and business results from data.

• **Data Engineering** convert raw data into usable data

• **Data Analytics** use this data to generate insights

• **Data Scientists** use Data Analytics and Data Engineering to predict the future using data from the past

• **Business Analysts and Business Intelligence Analysts** use these insights and predictions to drive decisions that benefit and grow their business

Based on the goals and outcomes that need to be achieved, there are **four primary types of Data Analysis:**

* + **Descriptive** Analytics, that helps decode “**What happened**.”
  + **Diagnostic** Analytics, that helps us understand “**Why it happened**.”
  + **Predictive** Analytics, that analyzes historical data and trends to suggest “**What will happen next**.”
  + **Prescriptive** Analytics, that prescribes “**What should be done next**.”

The **Data Analysis process** involves:

* + **Developing an understanding of the problem and the desired outcome:** Data analysis begins with understanding the problem that needs to be solved and the desired outcome that needs to be achieved. Where you are and where you want to be needs to be clearly defined before the analysis process can begin.
  + **Setting a clear metric for evaluating outcomes**: This stage of the process includes deciding what will be measured. For example, number of product X sold in a region and how it will be measured.
  + **Gathering, cleaning, analyzing, and mining data to interpret results:** **Gathering data** once you know what you're going to measure and how you're going to measure it, you identify the data you require, the data sources you need to pull this data from, and the best tools for the job.

**Cleaning data**: Having gathered the data, the next step is to fix quality issues in the data that could affect the accuracy of the analysis.

**Analyzing and mining data**: Once the data is clean, you will extract and analyze the data from different perspectives. You may need to manipulate your data in several different ways to understand the trends, identify correlations and find patterns and variations.

**Interpreting Results:** After analyzing your data and possibly conducting further research, which can be an iterative loop, it's time to interpret your results. As you interpret your results, you need to evaluate if your analysis is defendable against objections, and if there are any limitations or circumstances under which your analysis may not hold true. Presenting your findings: Ultimately, the goal of any analysis is to impact decision making. The ability to communicate and present your findings in clear and impactful ways is as important a part of the data analysis process as is the analysis itself. Reports, dashboards, charts, graphs, maps, case studies are just some of the ways in which you can present your data.

**VIEWS:**

* I define data analytics as the process of collecting information and then analyzing that information to confirm various hypothesis.
* To me, data analytics also means storytelling with data. Using data to clearly and concisely convey the state of the world to the people around you.
* Data analysis is the use of information around you to make decisions. Just like you get up every morning, you watch the news. The weather report will tell you the temperature for the day, whether it's going to rain. That may dictate what you're going to wear or what activities you can do. Data analysis isn't an abstract concept, it's something that we do naturally, but it has a technical name and now people are being paid to do it in a much larger or grander experience. But really, it's not that complicated.
* Data analytics is really any sets of data that you can use to review information, anything that's going to help you to understand what is going on. In my case as a CPA, I am always looking at financial state. I'm always analyzing data to predict where someone's been, where they are right now, and where they're headed. That data helps me to see further and almost predict the future of any company that I'm working with.
* Data analytics is the collecting, cleansing, analyzing, presenting, and ultimately sharing of data and your analysis to be able to help communicate exactly what's going on with your business, what's going on in the data so that you can help make better decisions.

**Data Analytics vs. Data Analysis**

**Analysis** - detailed **examination of the elements** or structure **of something**

**Analytics** - the systematic **computational analysis of data** or statistics

**Analysis can be done without numbers or data**, such as business analysis psycho analysis, etc. **Whereas Analytics, even when used without the prefix "Data", almost invariably implies use of data for performing numerical manipulation and inference**. Some experts even say that Data Analysis is based on inferences based on historical data whereas Data Analytics is for predicting future performance. The design team of this course does not subscribe to this view, and you will see why later in the course as you become familiar with the terms like predictive analytics, prescriptive analytics, etc.

**MODULE 2: 9TH AUGUST 2021**

The role of a Data Analyst spans across:

* + Acquiring data that best serves the use case.
  + Preparing and analyzing data to understand what it represents.
  + Interpreting and effectively communicating the message to stakeholders who need to act on the findings.
  + Ensuring that the process is documented for future reference and repeatability.

**A day in the life of a Data Analyst can include a number of possibilities —**

* from acquiring data from varied data sources to creating queries for pulling data from data repositories,
* foraging through rows of data to look for insights,
* creating reports and dashboards, and interacting with stakeholders for gathering information and presenting the findings, it’s a spectrum.
* And yes, the big one — cleaning and preparing the data so that the findings have a credible basis — which, by the way, is a large part of what any Data Analyst may find themselves doing in their jobs.

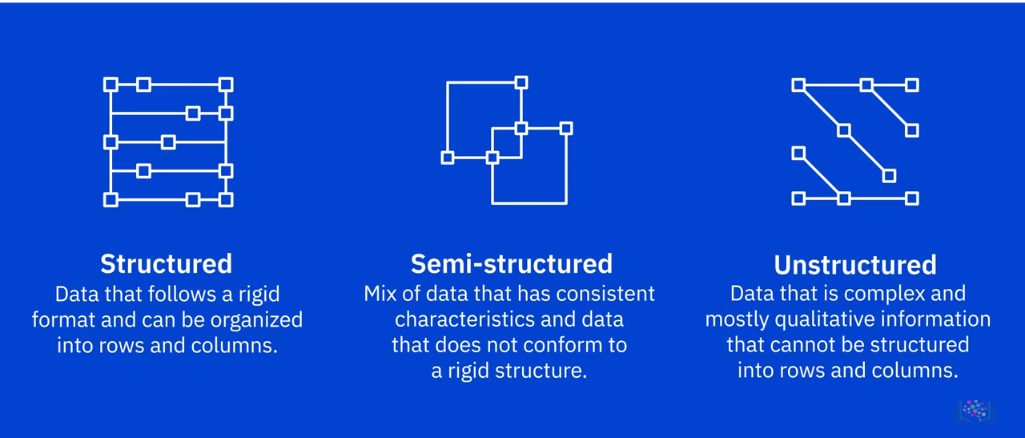
In order to play this role successfully, **Data Analysts need a mix of technical, functional, and soft skills:**

* + **Technical Skills** include varying levels of proficiency in using spreadsheets, statistical tools, visualization tools, programming and querying languages, and the ability to work with different types of data repositories and big data platforms.
  + An understanding of Statistics, Analytical techniques, problem-solving, the ability to probe a situation from multiple perspectives, data visualization, and project management skills – all of which come under Functional Skills a Data Analyst needs in order to play an effective role.
  + **Soft Skills** include the ability to work collaboratively, communicate effectively, tell a compelling story with data, and garner support and buy-in from stakeholders. Curiosity to explore different pathways and intuition that helps to give a sense of the future based on past experiences are also essential skills for being a good Data Analyst.

**MODULE 3 - THE DATA ECOSYSTEM AND LANGUAGES FOR DATA PROFESSIONALS (12 August 2021)**

**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. Data that follows a rigid format and can be organized neatly into rows and columns is structured data. Eg: **excel spreadsheets**, database, Online forms, Sensors such as Global Positioning Systems (or GPS) and Radio Frequency Identification(or RFID) tags; and Network and Web server logs.
  + **Semi-Structured Data**, that is data which is partially organized and partially free form. Semi-structured data is a mix of data that has consistent characteristics and data that doesn’t conform to a rigid structure. **It contains tags (like genre of movies) and elements, or metadata, which is used to group data and organize it in a hierarchy.** For example, **emails**. An email has a mix of structured data, such as the name of the sender and recipient, but also has the contents of the email, which is unstructured data.
  + **Unstructured Data,** that is data which cannot be organized conventionally into rows and columns. It does not follow any particular format, sequence, semantics, or rules. For example, **photos, videos, text files, PDFs, and social media content**. Some of the sources of unstructured data could include: Web pages, social media feeds, Images in varied file formats (such as JPEG, GIF, and PNG) Video and Audio files, Documents and PDF files, PowerPoint presentations, Media logs.

Data is unorganized information that is processed to make it meaningful. Generally, data comprises of facts, observations, perceptions, numbers, characters, symbols and images that can be interpreted to derive meaning. 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 (a term that includes databases, data warehouses, data marts, data lakes, and big data stores) 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.

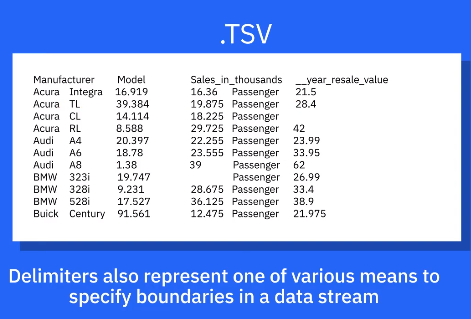
Data professionals need a host of languages that can help them extract, prepare, and analyze data. These can be classified as:

* + **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.

**Different Types of File Formats:**

1. **Delimited text file formats or .CSV**

## A delimiter is a sequence of one or more characters for specifying the boundary between independent entities or values.eg: comma, tab, colon, vertical bar, and space. **Comma-separated values (or CSV**s) **and tab-separated values (or TSVs)** are the most commonly used file types in this category.

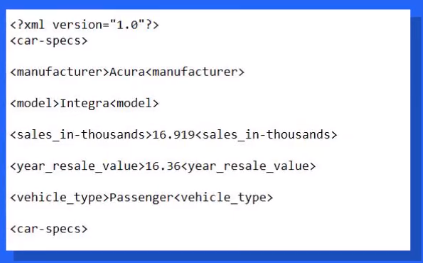


1. **Microsoft Excel Open .XML Spreadsheet, or .XLSX**

It falls under the spreadsheet file format. It is an XML-based file format created by Microsoft.

1. **Extensible Markup Language, or .XML**

Extensible Markup Language, or XML, is a markup language with set rules for encoding data.The XML file format is both readable by humans and machines. It is a self-descriptive language designed for sending information over the internet. XML is similar to HTML in some respects, but also has differences. For example, an .XML does not use predefined tags like .HTML does.



1. **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, which means it can be viewed the same way on any device.

1. **JavaScript Object Notation, or. JSON**

JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web. The file format is a language-independent data format that can be read in any programming language. JSON is easy to use, is compatible with a wide range of browsers, and is considered as one of the best tools for sharing data of any size and type, even audio and video. That is one reason, many APIs and Web Services return data as JSON.

**Module 4 - Understanding Data Repositories and Big Data Platforms( August 21, 2021 )**

In this lesson, you have learned the following information:

A Data Repository is a general term that refers to data that has been collected, organized, and isolated so that it can be used for reporting, analytics, and also for archival purposes.

The different types of Data Repositories include:

**Databases**, which can be relational or non-relational, each following a set of organizational principles, the types of data they can store, and the tools that can be used to query, organize, and retrieve data.

**Data Warehouses**, that consolidate incoming data into one comprehensive storehouse.

**Data Marts**, that are essentially sub-sections of a data warehouse, built to isolate data for a particular business function or use case.

**Data Lakes**, that serve as storage repositories for large amounts of structured, semi-structured, and unstructured data in their native format.

**Big Data Stores**, that provide distributed computational and storage infrastructure to store, scale, and process very large data sets.

**ETL, or Extract Transform and Load**, Process is an automated process that converts raw data into analysis-ready data by:

**Extracting data from source locations.**

**Transforming raw data by cleaning, enriching, standardizing, and validating it.**

**Loading the processed data into a destination system or data repository.**

**Data Pipeline**, sometimes used interchangeably with ETL, encompasses the **entire journey** of moving data from the source to a destination data lake or application, using the ETL process.

**Big Data** refers to the vast amounts of data that is being produced each moment of every day, by people, tools, and machines. The sheer velocity, volume, and variety of data challenge the tools and systems used for conventional data. These challenges led to the emergence of processing tools and platforms designed specifically for Big Data, such as Apache Hadoop, Apache Hive, and Apache Spark.

**MODULE 5: GATHERING DATA August 21, 2021**

* The process of identifying data begins by determining the information that needs to be collected, which in turn is determined by the goal you seek to achieve.
* Having identified the data, your next step is to identify the sources from which you will extract the required data and define a plan for data collection. Decisions regarding the timeframe over which you need your data set, and how much data would suffice for arriving at a credible analysis also weigh in at this stage.
* Data Sources can be **internal or external** to the organization, and they can be **primary, secondary, or third-party**, depending on whether you are obtaining the data **directly from the original source, retrieving it from externally available data sources, or purchasing it from data aggregators.**
* Some of the data sources from which you could be gathering data include **databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys and observation studies.**
* Data that has been identified and gathered from the various data sources is combined using a variety of tools and methods to provide a single interface using which data can be queried and manipulated.
* The data you identify, the source of that data, and the practices you employ for gathering the data have implications for quality, security, and privacy, which need to be considered at this stage.

**MODULE 6 – WRANGLING DATA (August 22, 2021)**

In this lesson, you have learned the following information:

Once the data you identified is gathered and imported, your next step is to make it analysis-ready. This is where the process of Data Wrangling, or Data Munging, comes in. Data Wrangling is an iterative process that involves data exploration, transformation, and validation.

Transformation of raw data includes the tasks you undertake to:

* Structurally manipulate and combine the data using Joins and Unions.
* Normalize data, that is, clean the database of unused and redundant data.
* Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.
* Clean data, which involves profiling data to uncover quality issues, visualizing data to spot outliers, and fixing issues such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.
* Enrich data, which involves considering additional data points that could add value to the existing data set and lead to a more meaningful analysis.

A variety of software and tools are available for the Data Wrangling process. Some of the popularly used ones include Excel Power Query, Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python, and R, each with their own set of characteristics, strengths, limitations, and applications.

**MODULE 7- ANALYING AND MINING DATA (August 23, 2021)**

**Statistics** is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data. Statistical Analysis involves the use of statistical methods in order to develop an understanding of what the data represents.

**Statistical Analysis can be:**

**Descriptive**; that which provides a summary of what the data represents. Common measures include Central Tendency, Dispersion, and Skewness.

**Inferential**; that which involves making inferences, or generalizations, about data. Common measures include Hypothesis Testing, Confidence Intervals, and Regression Analysis.

**Data Mining**, simply put, is the process of extracting knowledge from data. It involves the use of pattern recognition technologies, statistical analysis, and mathematical techniques, in order to identify correlations, patterns, variations, and trends in data.

There are several techniques that can help mine data, such as, classifying attributes of data, clustering data into groups, establishing relationships between events, variables, and input and output.

A variety of software and tools are available for analyzing and mining data. Some of the popularly used ones include Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio, and SAS, each with their own set of characteristics, strengths, limitations, and applications.

**MODULE 8 - COMMUNICATING DATA ANALYSIS FINDINGS (August 23, 2021)**

Data has value through the stories that it tells. In order to communicate your findings impactfully, you need to:

* Ensure that your audience is able to trust you, understand you, and relate to your findings and insights.
* Establish the credibility of your findings.
* Present the data within a structured narrative.
* Support your communication with strong visualizations so that the message is clear and concise, and drives your audience to take action.

**Data visualization** is the discipline of communicating information through the use of visual elements such as graphs, charts, and maps. The goal of visualizing data is to make information easy to comprehend, interpret, and retain.

For data visualization to be of value, you need to:

* Think about the key takeaway for your audience.
* Anticipate their information needs and questions, and then plan the visualization that delivers your message clearly and impactfully.

There are several types of graphs and charts available for you to be able to plot any kind of data, such as **bar charts, column charts, pie charts, and line charts.**

You can also use data visualization to build dashboards. **Dashboards** organize and display reports and visualizations coming from multiple data sources into a single graphical interface. They are easy to comprehend and allow you to generate reports on the go.

When deciding which tools to use for data visualization, you need to consider the ease-of-use and purpose of the visualization. **Some of the popularly used tools include Spreadsheets, Jupyter Notebook, Python libraries, R-Studio and R-Shiny, IBM Cognos Analytics, Tableau, and Power BI.**

**MODULE 9 - OPPORTUNITIES AND LEARNING PATHS (August 23, 2021)**

Data Analyst roles are sought after in every industry, be it **Banking and Finance, Insurance, Healthcare, Retail, or Information Technology.**

Currently, the demand for skilled data analysts far outweighs the supply, which means companies are willing to pay a premium to hire skilled data analysts.

Data Analyst job roles can be **broadly classified** as follows:

* Data Analyst Specialist roles - On this path, you start as a Junior Data Analyst and move up to the level of a Principal Analyst by continually advancing your technical, statistical, and analytical skills from a foundational level to an expert level.
* Domain Specialist roles - These roles are for you if you have acquired specialization in a specific domain and want to work your way up to be seen as an authority in your domain.
* Analytics-enabled job roles - These roles include jobs where having analytic skills can up-level your performance and differentiate you from your peers.
* Other Data Professions - There are several other roles in a modern data ecosystem, such as Data Engineer, Big Data Engineer, Data Scientist, Business Analyst, or Business Intelligence Analyst. If you upskill yourself based on the required skills, you can transition into these roles.

There are several paths you can consider in order to gain entry into the Data Analyst field. These include:

* An academic degree in Data Analytics or disciplines such as Statistics and Computer Science.
* Online multi-course specializations offered by learning platforms such as Coursera, edX, and Udacity.
* Mid-career transition into Data Analysis by upskilling yourself. If you have a technical background, for example, you can focus on developing the technical skills specific to Data Analysis. If you do not have a technical background, you can plan to skill yourself in some basic technologies and then work your way up from an entry-level position.

THE END