**Introduction to data analysis**

**Module 1**

**1.1 Modern data ecosystem and the role of data analytics.**

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.

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.
* Setting a clear metric for evaluating outcomes.
* Gathering, cleaning, analyzing, and mining data to interpret results.
* Communicating the findings in ways that impact decision-making.

**1.2 The data analytics role**

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.

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 2**

**2.1 The data ecosystem and languages for data professionals**

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.

**2.2 Understanding data repository and big data platform**

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 3**

**3.1 Gathering data**

* 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.

**3.2 Wrangling data**

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 4**

**4.1 Analyzing and Mining data**

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.

**4.2 Communicating data analysis findings**

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 5**

**5.1 Opportunities and learning paths**

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 your self in some basic technologies and then work your way up from an entry-level position.

**5.2**

**Using Data Analysis for Detecting Credit Card Fraud**

Companies today are employing analytical techniques for the early detection of credit card frauds, a key factor in mitigating fraud damage. The most common type of credit card fraud does not involve the physical stealing of the card, but that of credit card credentials, which are then used for online purchases.

Imagine that you have been hired as a Data Analyst to work in the Credit Card Division of a bank. And your first assignment is to join your team in using data analysis for the early detection and mitigation of credit card fraud.

In order to prescribe a way forward, that is, suggest what should be done in order for fraud to get detected early on, you need to understand what a fraudulent transaction looks like. And for that you need to start by looking at historical data.

**Here is a sample data set that captures the credit card transaction details for a few users.**

A screenshot of a computer

Description automatically generated

Descriptive techniques of analysis, that is, techniques that help you gain an understanding of what happened, include the identification of patterns and anomalies in data. Anomalies signify a variation in a pattern that seems uncharacteristic, or, out of the ordinary. Anomalies may occur for perfectly valid and genuine reasons, but they do warrant an evaluation because they can be a sign of fraudulent activity.

**Past studies have suggested that some of the common events that you may need to watch out for include:**

* A change in frequency of orders placed, for example, a customer who typically places a couple of orders a month, suddenly makes numerous transactions within a short span of time, sometimes within minutes of the previous order.
* Orders that are significantly higher than a user’s average transaction.
* Bulk orders of the same item with slight variations such as color or size—especially if this is atypical of the user’s transaction history.
* A sudden change in delivery preference, for example, a change from home or office delivery address to in-store, warehouse, or PO Box delivery.
* A mismatched IP Address, or an IP Address that is not from the general location or area of the billing address.

**Before you can analyze the data for patterns and anomalies, you need to:**

* **Identify and gather all data points that can be of relevance to your use case.** For example, the card holder’s details, transaction details, delivery details, location, and network are some of the data points that could be explored.
* **Clean the data.** You need to identify and fix issues in the data that can lead to false or incomplete findings, such as missing data values and incorrect data. You may also need to standardize data formats in some cases, for example, the date fields.

Finally, when you arrive at the findings, you will create appropriate visualizations that communicate your findings to your audience. The graph below samples one such visualization that you would use to capture a trend hidden in the sample data set shared earlier on in the case study.

A graph of a transaction value

Description automatically generated

**In the next section you will be asked to answer the following 5 (five) questions based on this case study:**

1. List at least 5 (five) data points that are required for the analysis and detection of a credit card fraud. (3 marks)
2. Identify 3 (three) errors/issues that could impact the accuracy of your findings, based on a data table provided. (3 marks)
3. Identify 2 (two) anomalies, or unexpected behaviors, that would lead you to believe the transaction may be suspect, based on a data table provided. (2 marks)
4. Briefly explain your key take-away from the provided data visualization chart. (1 mark)
5. Identify the type of analysis that you are performing when you are analyzing historical credit card data to understand what a fraudulent transaction looks like. [Hint: The four types of Analytics include: Descriptive, Diagnostic, Predictive, Prescriptive] (1 mark)