Dell - Data Engineering Training

Day 1

Data Engineering Training

Who am I

Who am I

Udayakumar Mathivanan

Enterprise Architect | Technical Evangelist

Who am I

Solutions-oriented Enterprise Architect & Technical Evangelist with a proven track record of effective infrastructure design and software system implementation, through utilizing Best Practices and Agile Development Methodology to facilitate solution implementation in a medium or large enterprise environment across domains like **Healthcare, Automotive, Telecom, Retail, Banking and Finance. Subject Matter Expert on Data Science, Artificial Intelligence , Robotics , Gen AI and related technologies.**

Who am I

Nearly **two decades** of IT experience, 12 of which have been in the **Cloud & Data Engineering**

Completed Computing for **Data Science Graduate Program** at **Stanford University, CA**

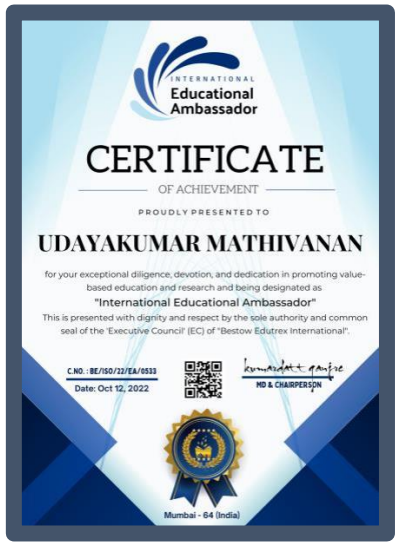
Global Business B.A., with Management Concentration, **University of South Florida, FL**

Served **Microsoft, Samsung and Amazon** in the past

Who am I

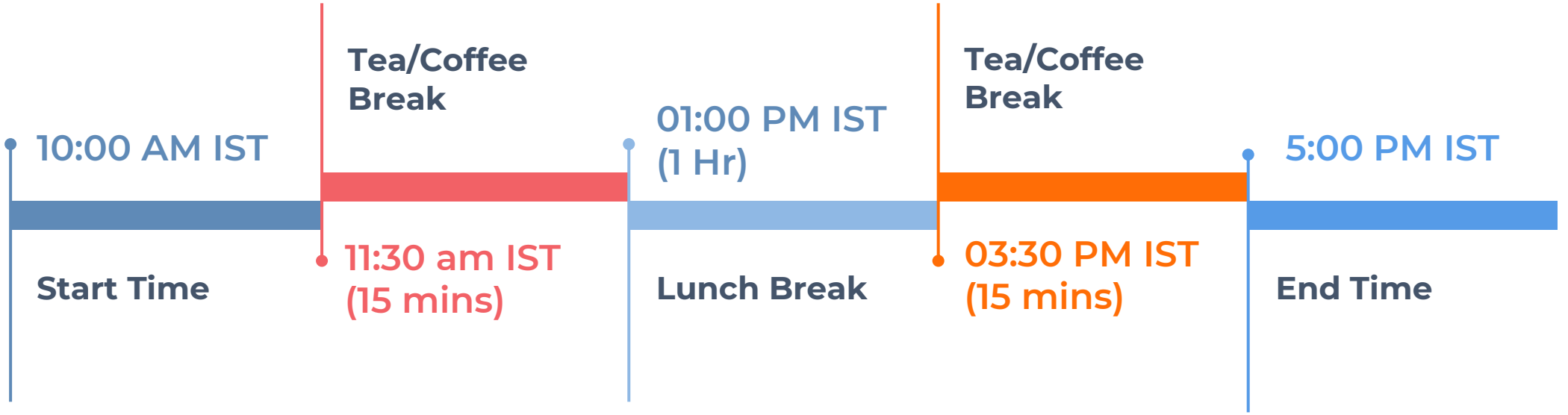
- Expertise on **AI Engineering, Data science & Statistics, Predictive Modeling, Deep Learning.**
- Member of **Microsoft, Amazon , Google International Technical Trainers Association.**
- Worked with Large Customers in **North America, Europe, Middle East and Asia.**
- Trained more than **50K+ Software/IT Professionals and Emerging Talents.**
- Active **Speaker/Blogger** on various User Groups/Forums across globe – Travelled More than **65+** Countries.

Who am I



Training & Break Timings

Training & Break Timings



Agenda

1. Introduction to Data Engineering
2. Stages in Data Engineering
3. Roles and Responsibilities of a Data Engineer
4. Types of data we have
5. Brief History of Python
6. Pandas in Data Engineering
7. What is Data Analysis
8. Linear Function
9. What is Exploratory Data Analysis

Introduction To Data Engineering

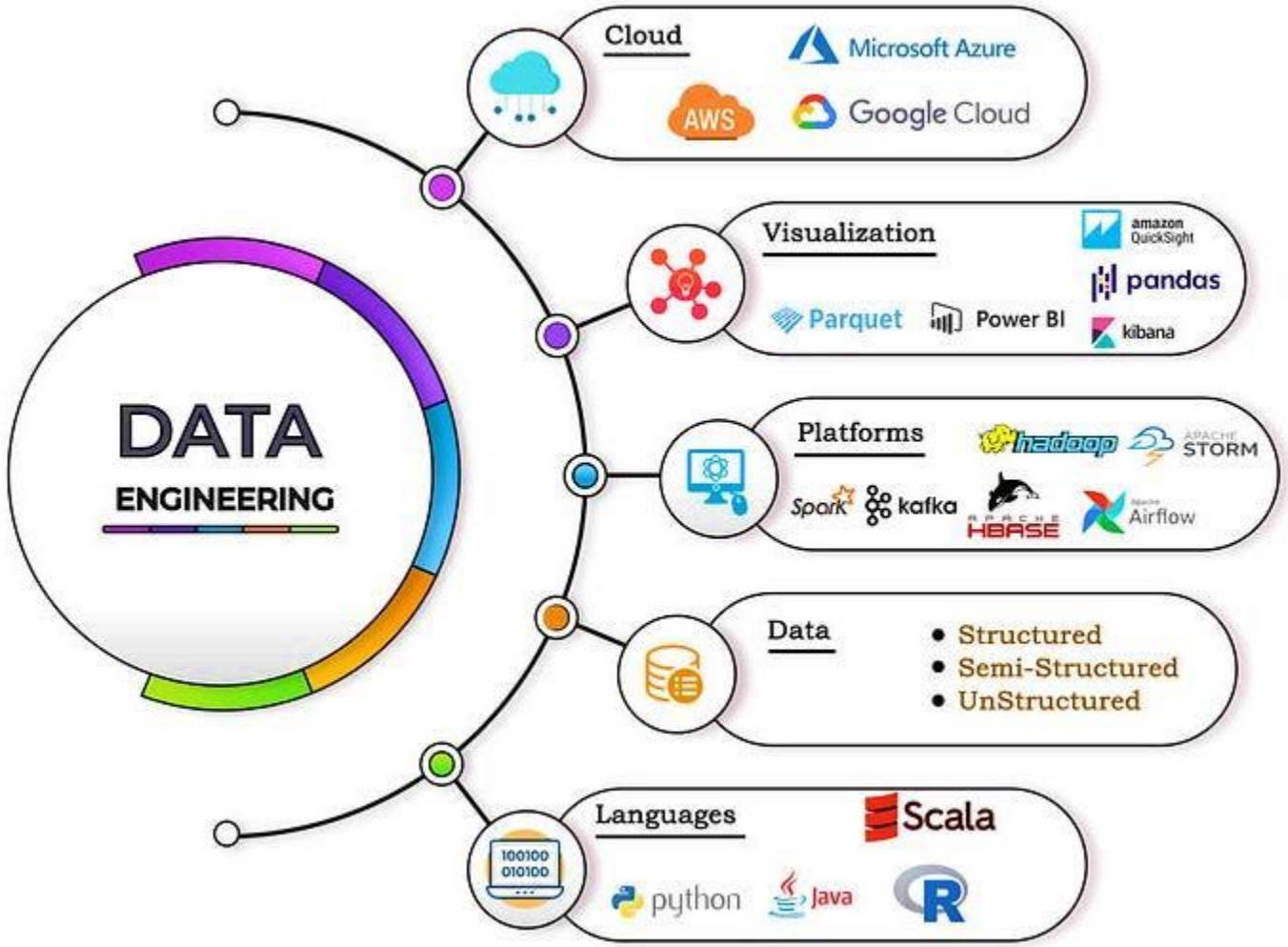
Introduction To Data Engineering

Data engineering is a set of operations to make data available and usable to data scientists, data analysts, business intelligence (BI) developers, and other specialists within an organization. It takes dedicated experts – data engineers – to design and build systems for gathering and storing data at scale as well as preparing it for further analysis.

What Is Data Engineering

Data engineering is the **process of designing, building, and maintaining** systems for **collecting, storing, processing, and analyzing large-scale data.** Data engineering **involves working with various data sources, such as databases, APIs, web scraping, streaming data,** etc., and transforming them into a unified and consistent format that can be used for further analysis or machine learning. Data engineering also involves **creating and managing data pipelines,** which are workflows that **automate the data flow from** the source to the destination, such as a **data warehouse, a data lake, or a dashboard.**

Data Engineering Overview

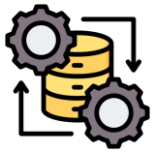


Stages in Data Engineering

Stages in Data Engineering

DATA ENGINEERING

Data sets



Pre-Processing

Classification

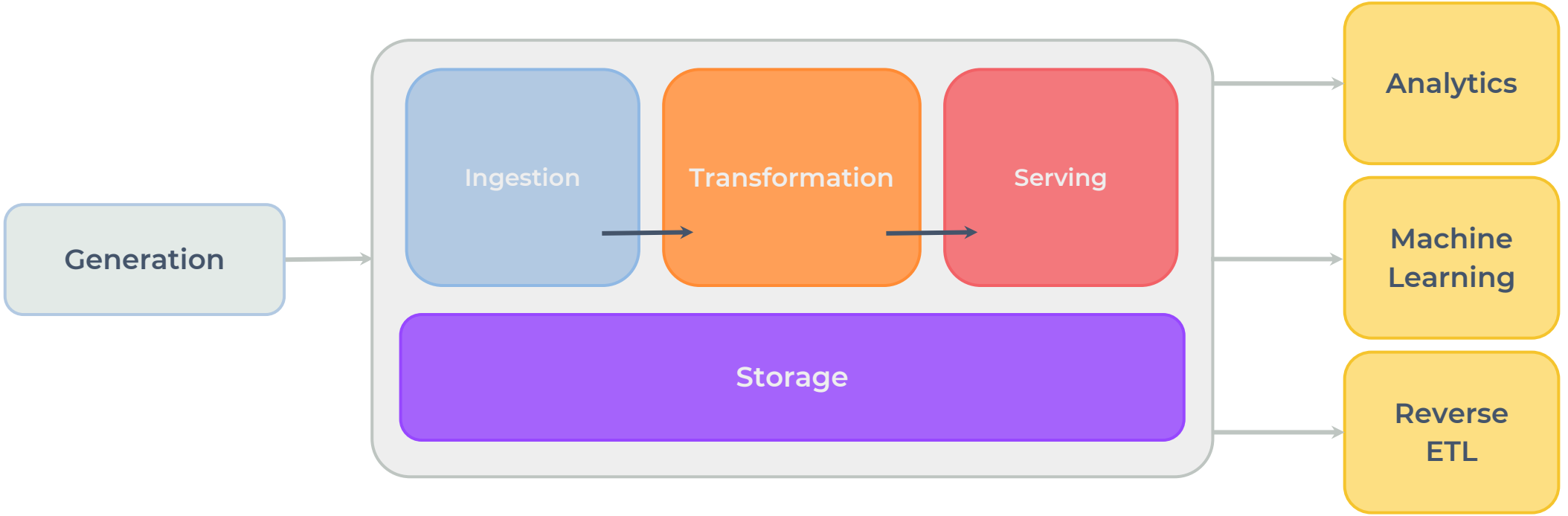
Database

Statistics

Analytics

Evaluation

Data Engineering Lifecycle

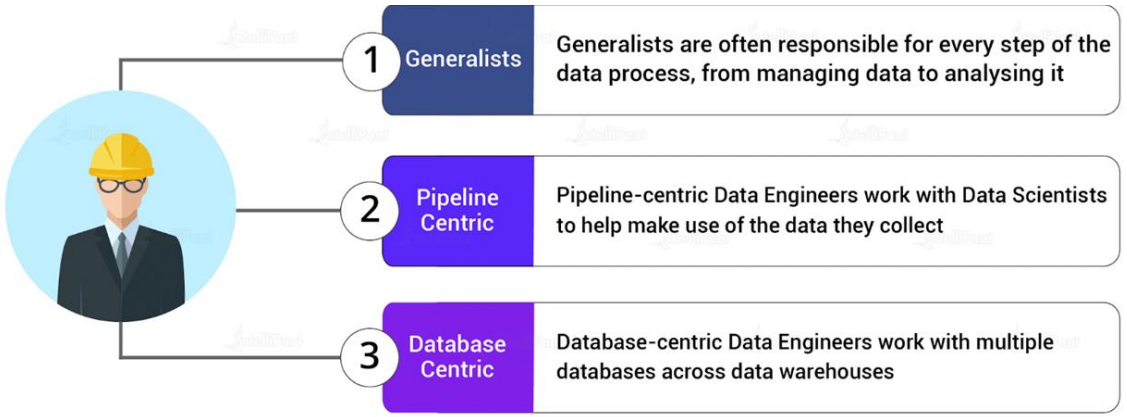


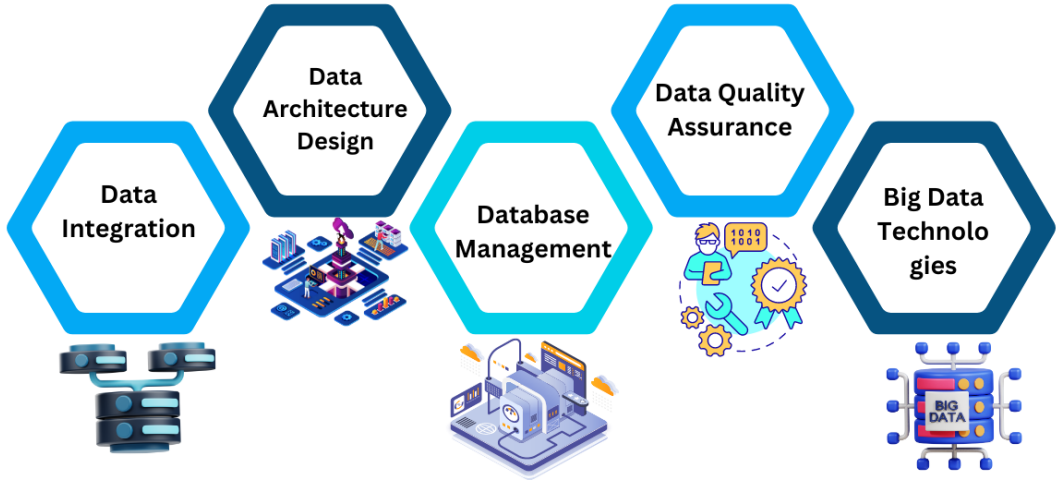
Undercurrents:

Security | Data Management | DataOps | Data Architecture | Orchestration | Software engineering

Roles and Responsibilities of a Data Engineer

Roles and Responsibilities of a Data Engineer

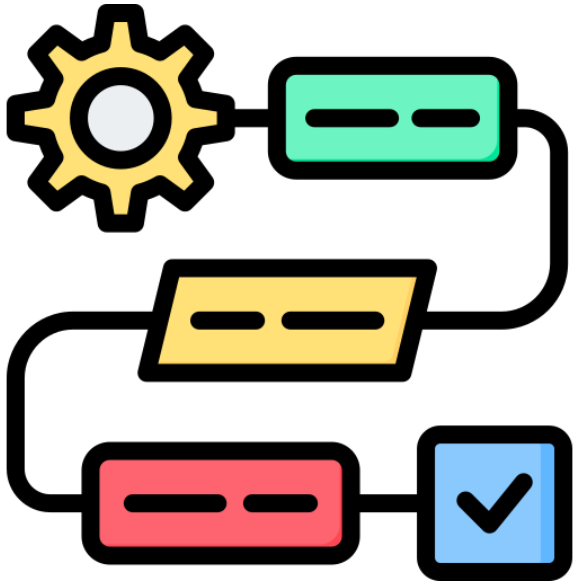




Data Engineering Process

A typical data engineering process includes:

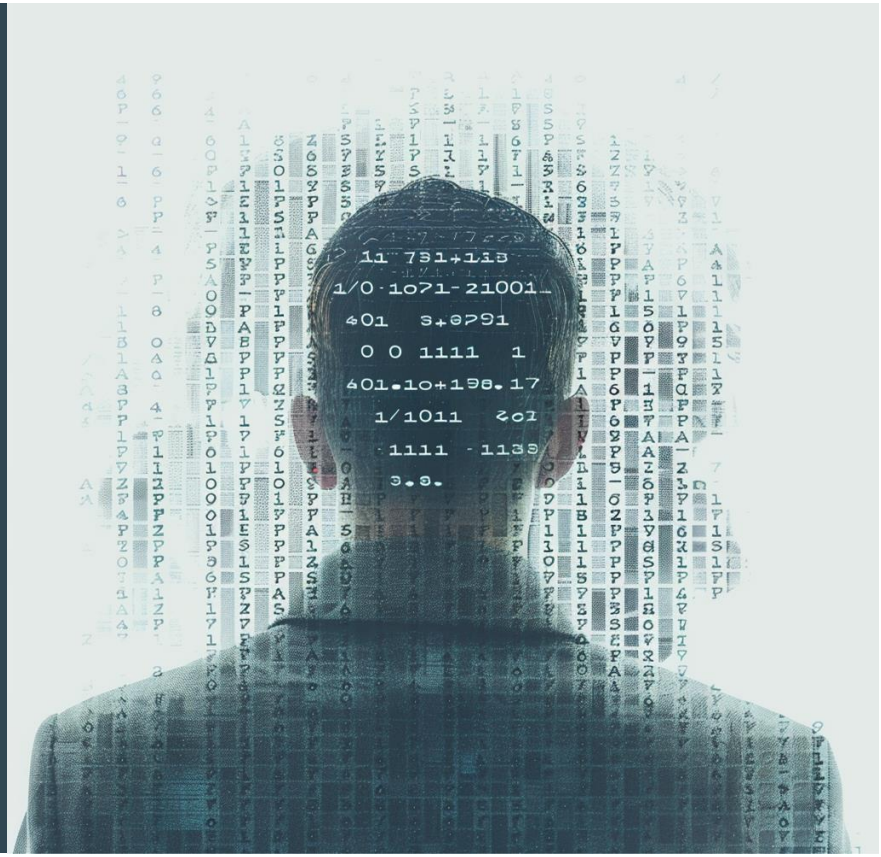
- **Data Flow:** This process enhances a standard data flow through a data pipeline to streamline data-driven models, such as ML models for real-time analysis.
- **Data Normalization and Modeling:** This process entails transforming data into easily accessible and usable formats to drive business decisions.
- **Data Cleaning:** Data cleaning eliminates incorrectly formatted, incomplete, or corrupted data from a data set when merging multiple data sources.
- **Data Accessibility:** This includes enhancing the experience of data access, as well as visualization using custom tools, charts, and illustrations.



Types Of Data We Have

Types Of Data We Have

A **massive amount of data** is produced every day because of the growth in the number of mobile users, rising internet penetration rates, and the accessibility of different eCommerce apps. Data science is a discipline that oversees gathering, processing, modeling, and analyzing data in order to acquire a better understanding of the data. Businesses use data science to improve decision-making, boost revenues, and accomplish growth.

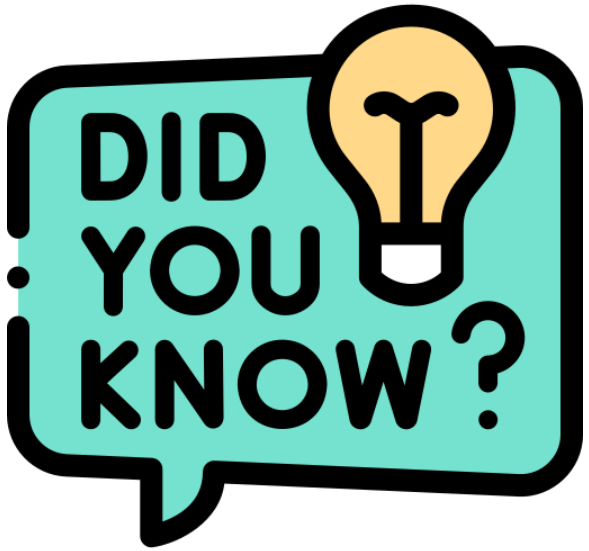


Data Facts

If we consider all the data that is currently available internationally, around 70% of it is user-generated according to a DM News report.

Data Facts

- According to one estimate, 1.145 trillion megabytes of data are produced daily.
- Statista estimates that in the previous year (2021), there were around 79 Zettabytes of data/information created, consumed, collected, and duplicated globally.
- According to forecasts made by CrowdFlower in its Data Scientist Report, text data makes up 91% of the data utilized in data science. According to the same survey, unstructured data consists of 33% images, 11% audio, 15% video, and 20% other types of data in addition to text.



Data Facts

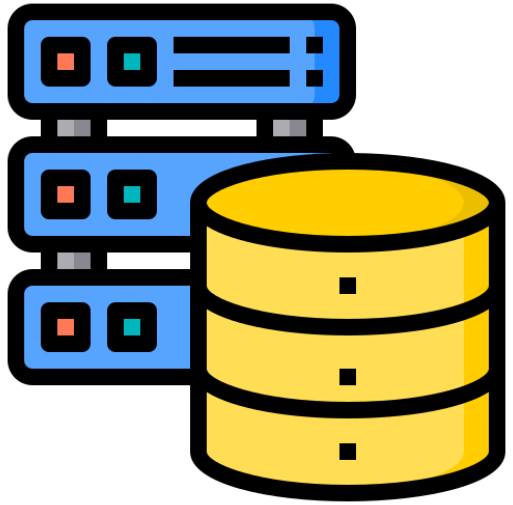
- In the worldwide digital universe, between 80 and 90% of the data is unstructured, according to one of the articles published on CIO.
- A user of the internet today would need 181 million years to download all the data from the internet.
- In 2023, about two professionals joined LinkedIn per second.
- The United States had 2670 data centers, making it the largest in the world in 2023.
- In 2023, according to Domo, every person on earth generated almost 2.5 quintillion bytes of data each day

Data Statistics For Future

- According to the same report from DOMO, in 2023, each person generated around 1.7 MB of data each second.
- By 2024, there will be 1.6 networked mobile devices and connections per person.
- By 2024, 149 zettabytes of data will have been copied, collected, and organized. Compared to the two zettabytes we produced in 2010, that is enormous.
- In 2026, it is anticipated that the market for data science platforms will be worth 322.9 USD billion.

Types Of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), ...
- Streaming Data



What Do We Do With This Data

Aggregation and Statistics

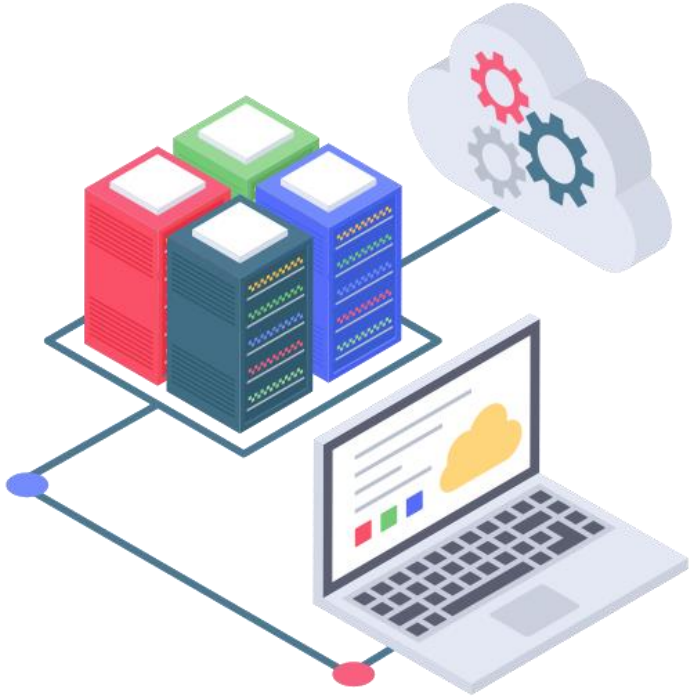
- Data warehousing and OLAP

Indexing, Searching, and Querying

- Keyword based search
- Pattern matching (XML/RDF)

Knowledge discovery

- Data Mining
- Statistical Modeling



Brief History Of Python

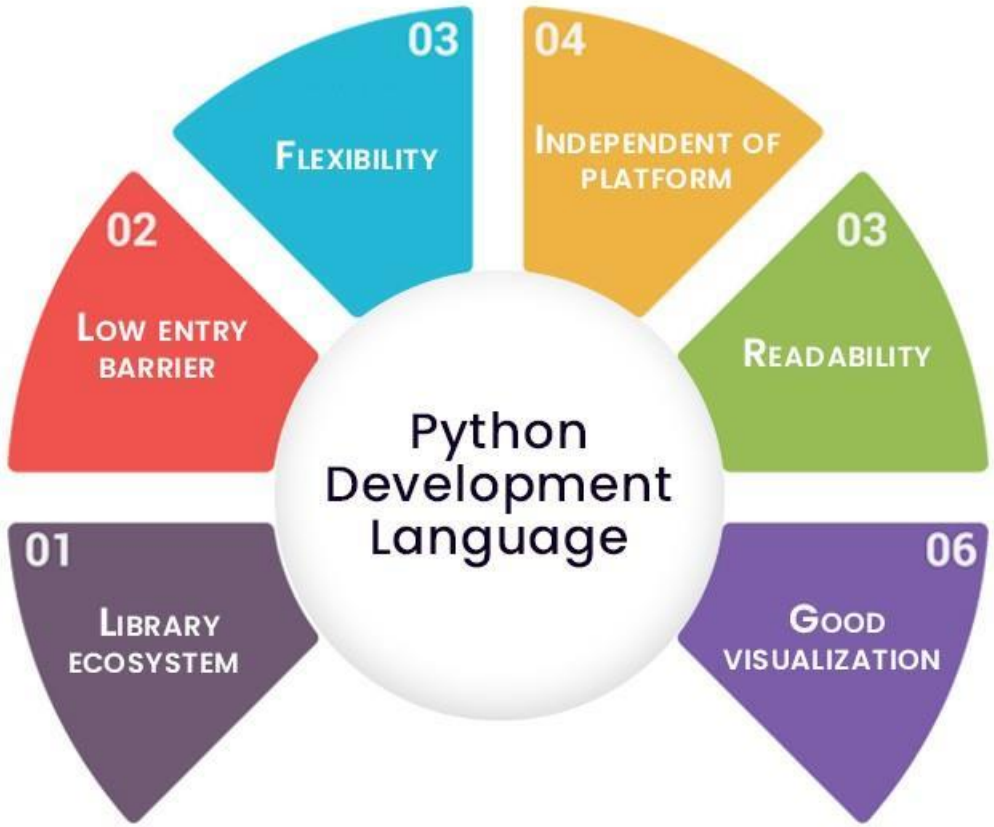
Brief History Of Python

- Invented in the **Netherlands**, early 90s by **Guido van Rossum**
- Named after **Monty Python**
- **Open sourced** from the beginning
- Considered as a **scripting language**, but is much more
- Scalable, **object oriented** and functional from the beginning
- Used by **Google** from the beginning

Guido van Rossum

Creator of Python

Benefits Of Python



Benefits of Python Development Language for AI and ML

Top Companies Using Python

In contemporary times, **126,424 websites** are made using the python programming language. Many top-notch companies have developed successful apps by using it. This is why it is considered the language of today and the future.

Pandas In Data Engineering

A library that simplifies task of data manipulation in Python.

- Abhijeet Dwivedi

Pandas In Data Engineering

Pandas are generally used for data science, but have you wondered why?

This is because pandas are used in conjunction with other libraries that are used for data science. It is built on the top of the **NumPy** library which means that a lot of structures of NumPy are used or replicated in Pandas.

The data produced by Pandas are often used as input for plotting functions of Matplotlib, statistical analysis in SciPy, and machine learning algorithms in Scikit-learn.

Why Pandas

- Data set **cleaning, merging, and joining.**
- Easy **handling of missing data** (represented as NaN) in floating point as well as non-floating point data.
- **Columns can be inserted and deleted** from Data Frame and higher dimensional objects.
- **Powerful group by functionality** for performing split-apply combine operations on data sets.
- **Data Visualization**

What is Data Analysis

What is Data Cleaning

It’s the process of **fixing errors, removing inconsistencies,** and generally whipping your data into shape for analysis.

Bad data = Bad decisions: Misspellings, missing values, and inconsistencies can seriously skew your results, leading to poor choices for your business or project.

Garbage in, garbage out: Even the fanciest analysis techniques won’t save you if your data is fundamentally flawed.

Time well spent: Data cleaning might seem tedious, but it saves you a world of headaches down the line, ensuring you’re making decisions based on rock-solid information.

What is Data Cleaning

let's map out exactly what needs to be **scrubbed**, **polished**, and **fixed** in a dataset.

Handle Missing Values: We need to find and address those pesky gaps in important columns.

Extract Numeric Values An important column holds the key to insights, but it's a jumbled mess right now. We'll extract just the numbers, ditch the currency symbols and extra text, and handle salary ranges.

Replace Inconsistent Values : Clean formatting in the column will make our analysis much smoother.

Parse 'Date' columns: We need to turn a date column into a true date format.

Standardize Columns : Removing random whitespace and fixing strange characters will make this text data easier to work with.

What is Data Cleaning

Handle missing values

Before we can fix missing data, we need to find out where it's hiding and how much we're dealing with.

```
# Identify missing values
missing_values = products.isnull().sum()
print("Missing Values:\n", missing_values)
```

Let's deal with the less problematic issues first;

```
# Fill missing product type values with a default value
products['product type'].fillna('Unknown', inplace=True)

# Fill missing Price values with a default value
products['Price'].fillna(0, inplace=True)
```

What is Data Cleaning

What is Data Analysis

Data Analysis is the technique to **collect, transform, and organize data** to make **future predictions**, and make informed **data-driven decisions**. It also helps to find possible solutions for a business problem.

There are six steps for Data Analysis.

- Ask or Specify Data Requirements
- Prepare or Collect Data
- Clean and Process
- Analyze
- Share
- Act or Report

What is Data Analysis

Analyzing Numerical Data with NumPy

NumPy is an array processing package in Python and provides a high-performance multidimensional array object and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

Arrays in NumPy

NumPy Array is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers. In Numpy, the number of dimensions of the array is called the rank of the array. A tuple of integers giving the size of the array along each dimension is known as the shape of the array.

Linear Function

Linear Function

A linear function has one independent variable (x) and one dependent variable (y), and has the following form:

$$y = f(x) = ax + b$$

This function is used to calculate a value for the dependent variable when we choose a value for the independent variable.

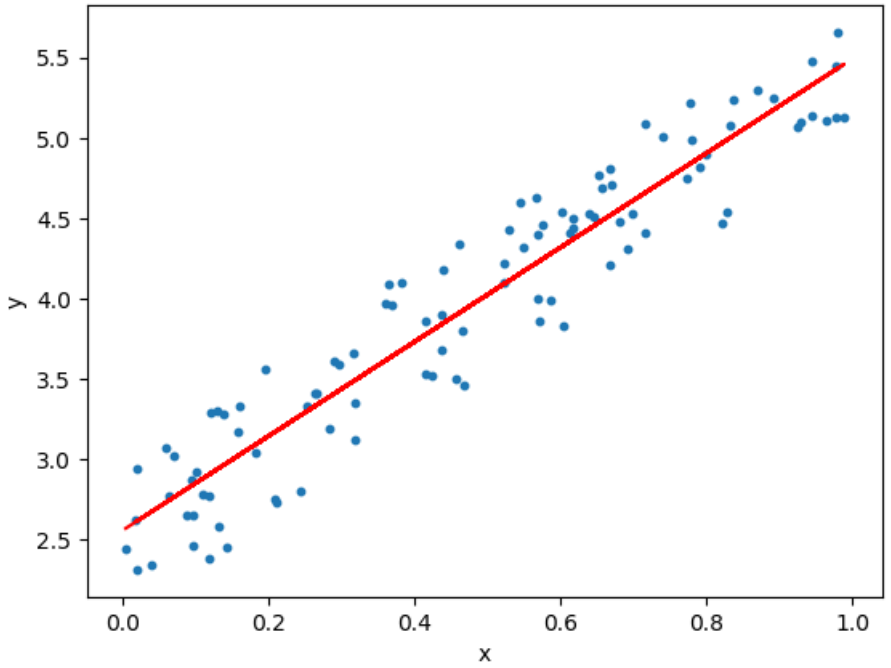
Explanation:

- f(x) = the output (the dependant variable)
- x = the input (the independent variable)
- a = slope = is the coefficient of the independent variable. It gives the rate of change of the dependent variable
- b = intercept = is the value of the dependent variable when x = 0. It is also the point where the diagonal line crosses the vertical axis.

Linear Function With One Explanatory Variable

A function with one explanatory variable means that we use one variable for prediction.

Linear Regression is usually the first machine learning algorithm that every data scientist comes across.



Where can Linear Regression be used?

It is a very powerful technique and can be used to understand the factors that influence profitability.

It can be used to forecast sales in the coming months by analyzing the sales data for previous months.

It can also be used to gain various insights about customer behaviour.

What Is Exploratory Data Analysis

What Is Exploratory Data Analysis?

Exploratory Data Analysis (EDA), also known as **Data Exploration**, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

'Understanding the dataset' can refer to a number of things including but not limited to...

Extracting important variables and leaving behind useless variables **Identifying outliers, missing values, or human error Understanding the relationship(s),** or lack of, between variables
Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process

What Is Exploratory Data Analysis?

Components of EDA

Main components of exploring data:

1. Understanding your variables
2. Cleaning your dataset
3. Analyzing relationships between variables

Contact Us

080-4524-9465

support@intellipaat.com