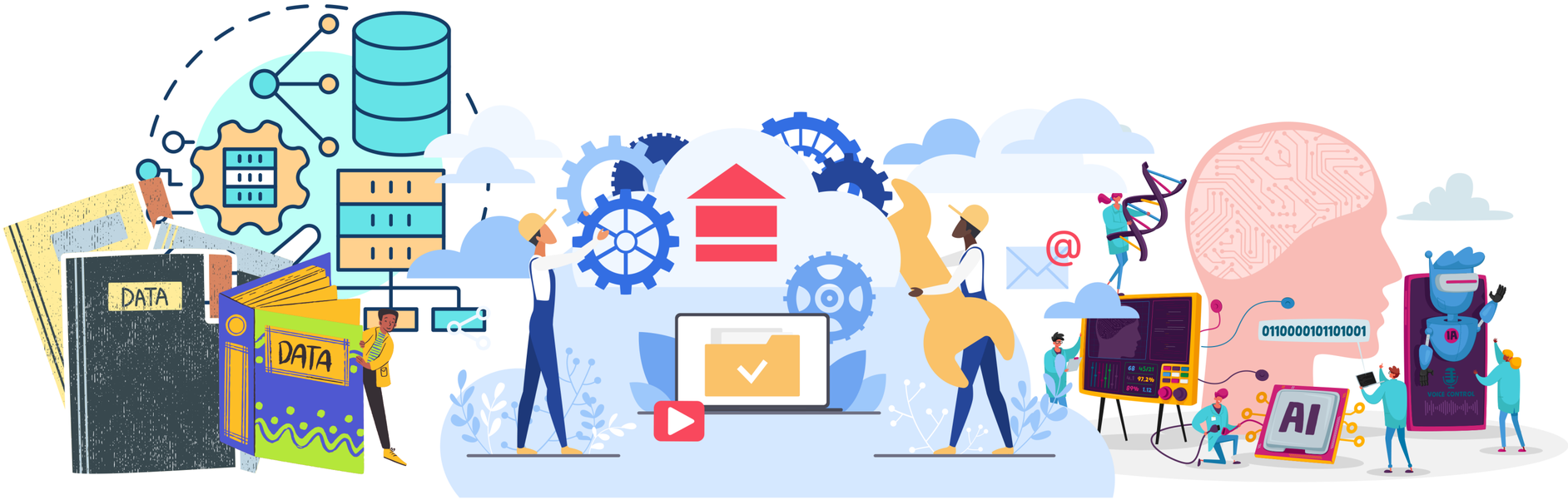
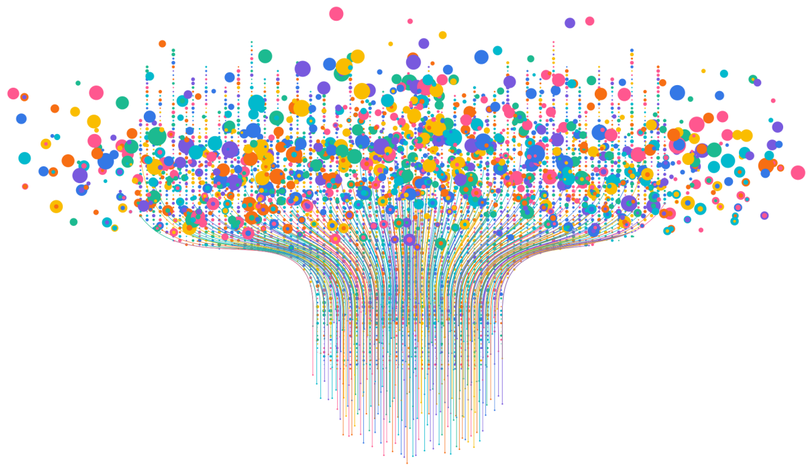
THE DATA ECOSYSTEM



EXPLAINED

SHIVANG KAINTHOLA

THE DATA ECOSYSTEM



Any development or decision making in the economy, industry, academics, governance, business etc. is **done the basis of** the existing **data**.

There is an entire ecosystem existing for **utilizing data** in such major use cases, **which encompasses many processes**, and the careers that help make it happen.

THE FUNDAMENTALS OF DATA

Data can existentially be defined by its :

Source
Where is the data coming from ?

License
What rules apply to the use of this data

Content
What does this data tell you ?

Format
What format is the data in ?

Schema
Does the data have any structure or hierarchy ?

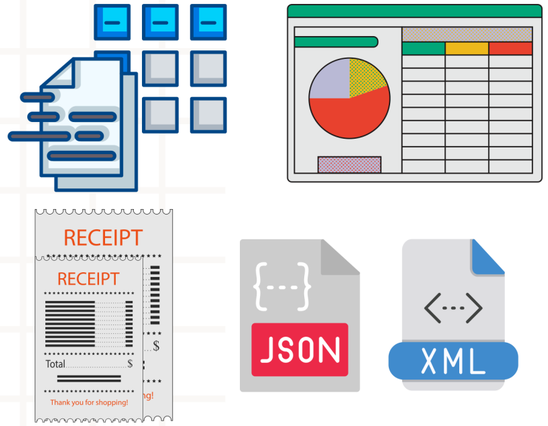
Usable
Is this data useful ?

Validity
Is this data accurate and reliable ?

Size and Storage
How much data is there, and how is it stored ?

Covering most of these factors, we can separate data into two broad categories :

Structured Data



Example : Spreadsheets, MySQLdatabases, forms,GPS data, binary files, XML/JSON files etc.

- Has a well defined **schema** and formats
- Can be represented in tabular format.
- Can be easily indexed.

Unstructured Data

Example : web pages, social media posts, documents, presentations, media logs, audio files etc.

- Has no defined schema or structure
- Can be in different formats.
- Cannot be easily indexed.

BIG DATA

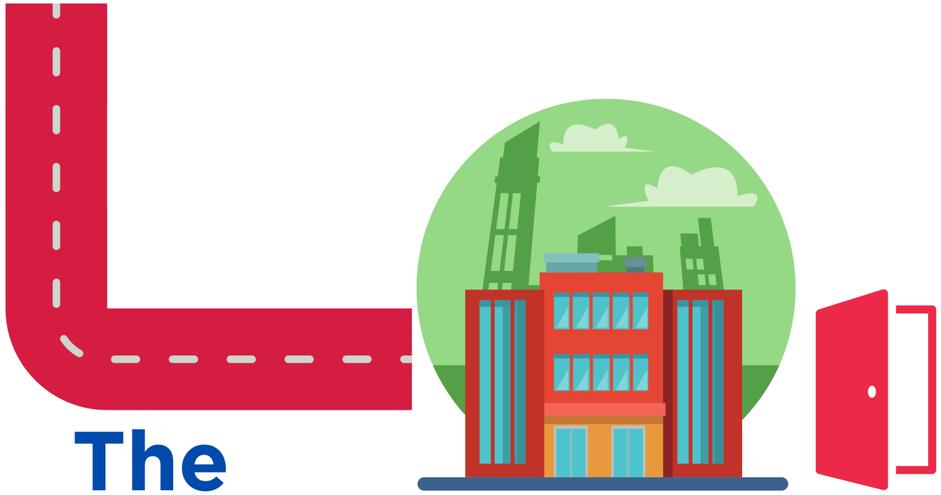
Today, there are **billions** of phones, computers, servers, sensors, cameras, and devices etc. constantly **generating** or **collecting raw data**, about all types of use cases ,at such a high volume and size,
it is called **Big Data**

This raw data is not obligated to be structured, have the correct format, size, or even be usable.

But all development relies on understanding this data,

and that begins with

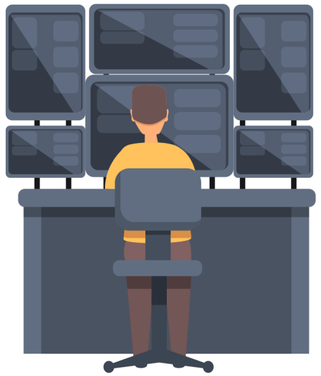
KEY PLAYERS IN THE DATA ECOSYSTEM



The organisation

has a definite goals to achieve from the data

Data Engineer



Researcher

Business Analyst

Database Engineer

Data Analyst

Data Scientist

THE DATA ENGINEER

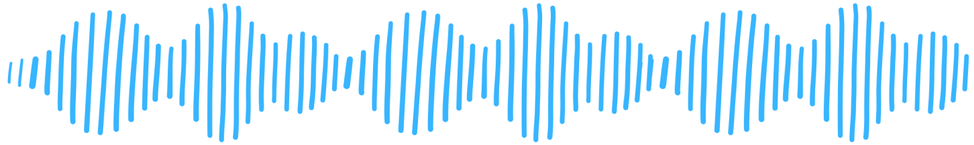
The data engineer is responsible for :

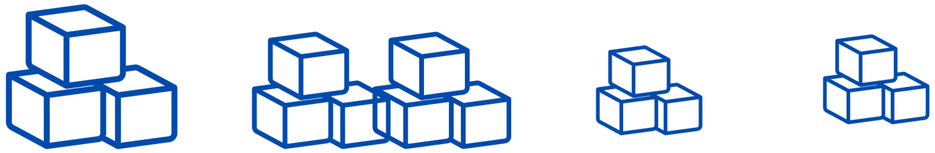
- Collection of data from the plethora of sources
- Transforming raw data to usable formats
- Designing structures to store the various types of data
- Maintaining repositories to store data
- Ensure data governance (data complies to quality and security standards)

(Thank them for your CSVs ;-)

Sources

Continuous stream of data





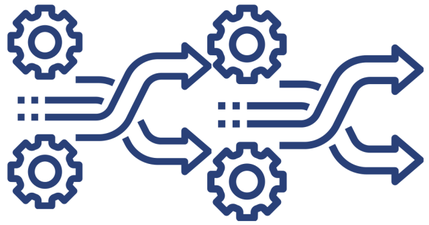
Separate blocks of data

Extract and Collect Data

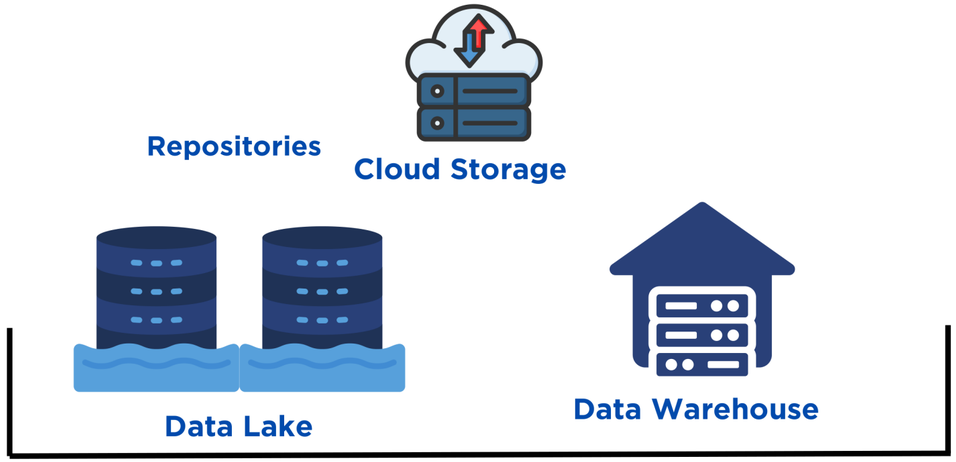
The data engineer organizes the whole process as a big **pipeline**.

The entire raw data from the plethora of sources can be **extracted** and collected as a constant **stream** of data, or in chunks or **blocks**.

Extracted Raw Data



Transform



Store and Load

The extracted and collected data is **transformed**, i.e., it is converted to a decent format, it is measured, labelled (metadata) and any preprocessing required is carried out.

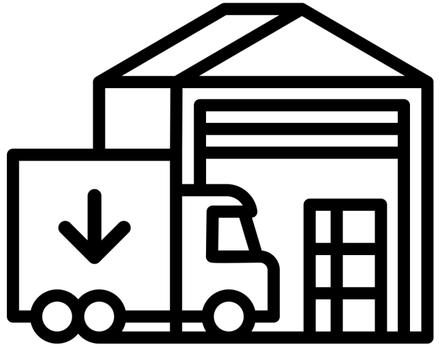
And then, it is stored in repositories like ***Data Lakes***, in combination with a ***Data warehouses***, all of which can be augmented with cloud computing.

TRANSFORM

The transformation of data involves :

- **Generating metadata** (data about the data) like table names, columns, etc.
- Making or **changing the structure** of data and data types within
- Detecting **irregularities or errors** within values and correcting them
- Removing **redundancies** and fixing **missing values**

LOAD



The ‘load’ step includes :

- Identifying what type of storage of is required
- Storing the transformed data in the prepared storage infrastructure
- Ensure data is under governance while in storage.
- Storing unstructured and structured data in appropriate repositories
- Unstructured data is often stored in NoSQL databases, while structured data is stored in relational databases

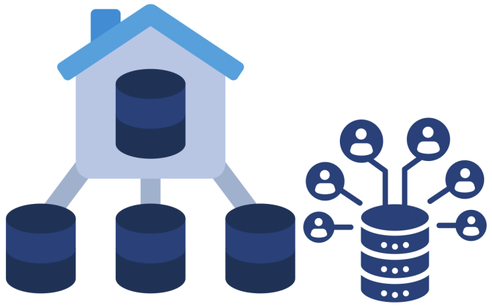
REPOSITORIES

Data Lakes

- Stores LARGE amount of raw collected data cheaply
- Has basic processing capabilities
- Scalable and adaptive
- Can be deployed by combining physical and cloud storage

Data Warehouses

- Stores extracted raw data
- Executes better processing, transformation on the data
- Provides backup options
- Avails the processed data to end users



Data Marts

- Hosts data independently or from data lakes and warehouses
- Avails data to end users or industrial functions
- Hybrid marts can host data from variety of sources

DATA WAREHOUSE TOOLS

teradata.

Google Big Query

snowflake®

Amazon Redshift

IBM DB2

ORACLE EXADATA

THE PIPELINE :
ETL / ELT

- This entire process of the pipeline from collection, transformation and storage / loading of data is popularly called **Extract-Transform-Load (ETL)**.
- In some cases, the data is loaded before pre-processing/transformation, where the process is called **Extract-Transform-Load (ETL)**.
- The data afterwards is availed to **destinations / end users**.

ETL TOOLS

Informatica PowerCenter

HEVO

IBM Infosphere Information Server

AWS Glue

Skyvia

Improvado

END USERS / DESTINATIONS

The Data Analyst

The Data Scientist

The Business Analyst /
Business Intelligence Analyst

Data Analysis Tools / Platforms

Power BI

tableau

QlikSense

Google Cloud

Looker

aws

IBM COGNOS

hadoop

THE DATA ANALYST

is typically going to :

- Have ***strong domain knowledge*** of whatever scientific, economic or social field the data will be used in
- **Inspect and clean data** to derive insights
- Identify correlations, **patterns**, and **apply statistical methods**
- **Visualize** data and ***report findings***, especially on critical aspects of the business/operation ; make **presentations** of reported findings
- Aid the data engineer in finding the best data sources

DATA SCIENCE

Azure Machine Learning

IBM Watson

Google's AutoML

Amazon SageMaker

Data **science** involves the use of scientific methods, algorithms, processes, and systems to extract insights and knowledge from structured and unstructured data.

Data **analysis** is an essential part of data science, a stepping stone towards identifying prediction targets and building of predictive machine learning models.

LightGBM

scikit learn

Keras

CatBoost

TensorFlow

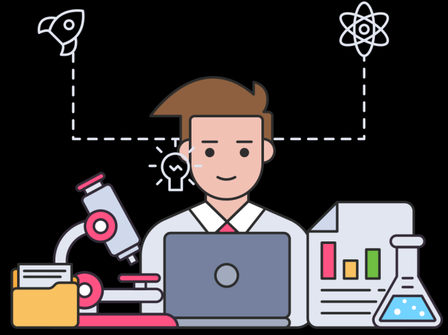
PyTorch

XGBoost

NVIDIA CUDA

Data science includes **deep** learning, **reinforcement** learning, **natural language processing**, **image processing**, **generative model** development etc.

THE DATA SCIENTIST



The data scientist is responsible for :

- Analyzing data for actionable insights, often collaborating with the data analyst
- Design models using machine learning and deep learning techniques and train on available data to forecast/predict business outcomes
- Deploying and maintaining various predictive models, essentially conducting the lifecycle of machine learning models

BUSINESS ANALYSTS

Business analysts primarily focus on **understanding business needs**, **identifying opportunities** for improvement, and **translating requirements into actionable insights** and recommendations for decision-making.

They **collaborate with data analysts** to gather and interpret data, while working alongside data scientists to **leverage advanced analytics** and **predictive modeling techniques in solving business problems** and driving strategic initiatives.

Jira

Agile

Lucidchart

Google Analytics

BUSINESS INTELLIGENCE ANALYSTS

Business intelligence analysts **focus on market forces and external influences** that shape their business

They collect, process, and analyze data to provide actionable insights that help organizations make informed decisions, optimize operations, and drive strategic initiatives within the data ecosystem.

RETENTION / DELETION

After the data is accessed for analysis, processing etc, there can be a many possibilities for the data depending on the goals and policies of the organization :

1. **SHARING** : The data and the results of its analysis may be shared with other organizations.

2. **RETENTION** : The data may be **archived** and retained by the organization.

3. **DELETION** : The data may be deleted.

THANK YOU !

WHILE I'M WORKING ON THE NEXT ARTICLE, I'M OVER HERE ON