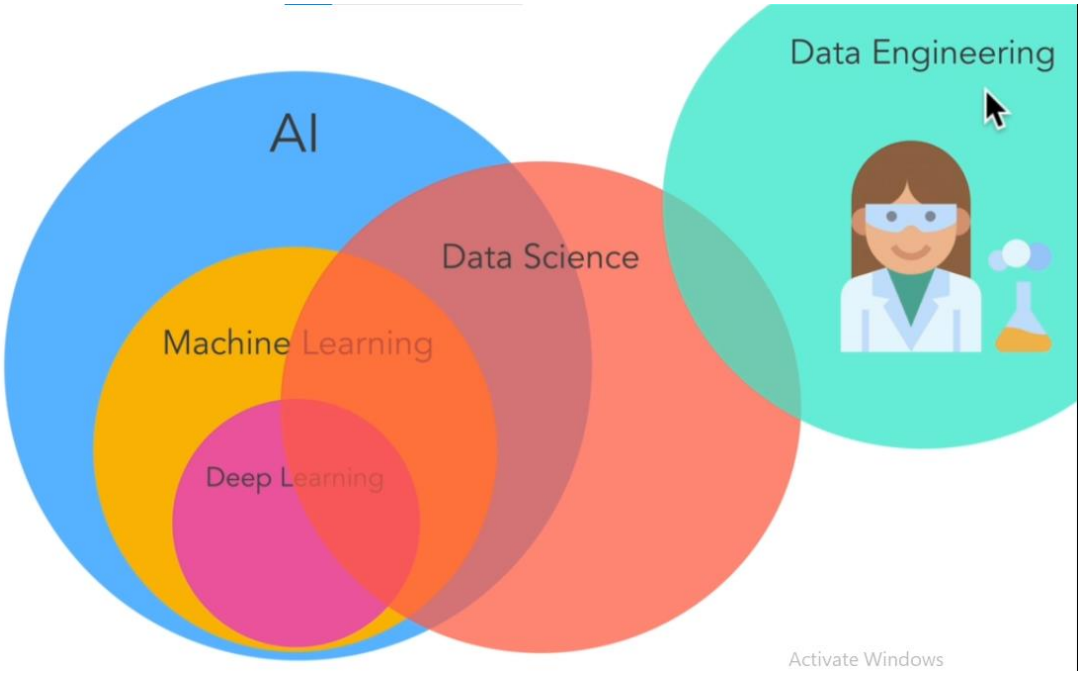
ML-9

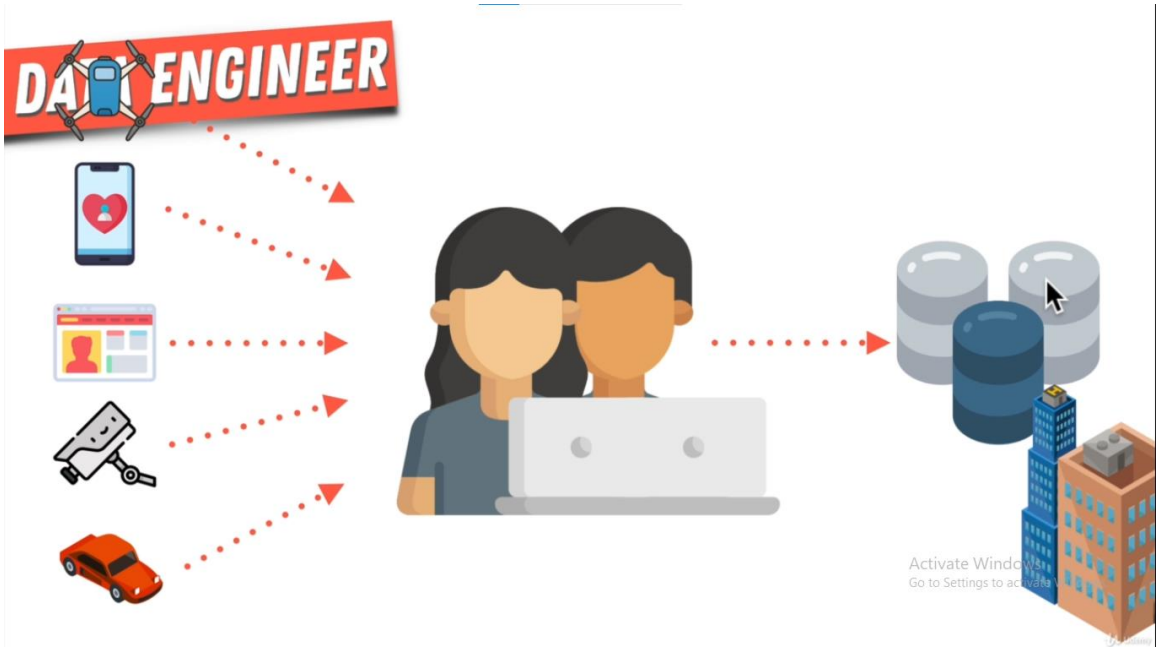
DATA ENGINEERING

Data Engineering



Well a data engineer would take all of the data points that are incoming, let's say for a company.

Well let's say this company has all these products and all these data are coming from their users from their security cameras, from their website, from IOT devices & a data engineer takes all this information and then produces it and maintains it in databases or a certain type of computers so that the business has access to this data in an organized fashion. So, they're kind of like the librarians where they collect all this information and they organize it for us. So, that people like machine learning or data science experts can use this.



Data

Types of data

Structured Data

This is data that is actually organized well for us to understand. Data gets collected in different ways from different sources.

This data is usually in a table or what we call a matrix a table that makes it easy to read.

Attributes are usually **columns** and **rows** are usually **instances** and sometimes we have the outputs in there in right most column like we saw with machine learning where we predict an output based on the inputs and structure

data usually comes from things such as relational databases like **mySQL** and other relational databases that allow us to do something called SQL queries.

Semi Structured data

There's still structure data but often in something like an **Excel, CSP** which we've seen or **Json** form.

This type of data is mostly available in **Kaggle.com**

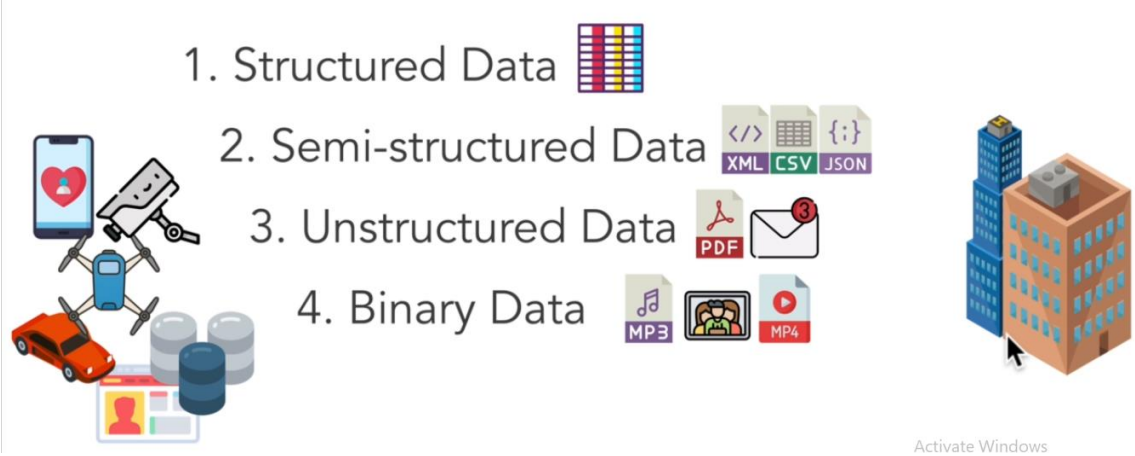
Unstructured Data

They're not usually in simple formats that we can really manipulate and analyze easily.

These are often things such as **emails** or PDFs or a sort of documents, where perhaps actually understanding them is a little bit more difficult.

Binary Data

These are things such as audio files, image files video files they're in binary, that is ones and zeros that computers can understand.



Data Mining

Data mining simply means preprocessing and extracting some knowledge from the data

So, we use some sort of data to extract knowledge from data is data mining.

Big Data

It means that we have a lot of data, a lot of variables so much data in fact that you can't run on your laptop and there's just too much data for your laptop or computer to hold. Usually when we talk about big data, it's data that's so big that you need to have it running on cloud computing or multiple computers such as A.W.S., Google cloud etc. because they have a lot of computers and a lot of storage to store this data.

Big Data is data that usually can't have just on one computer.

Data Pipeline

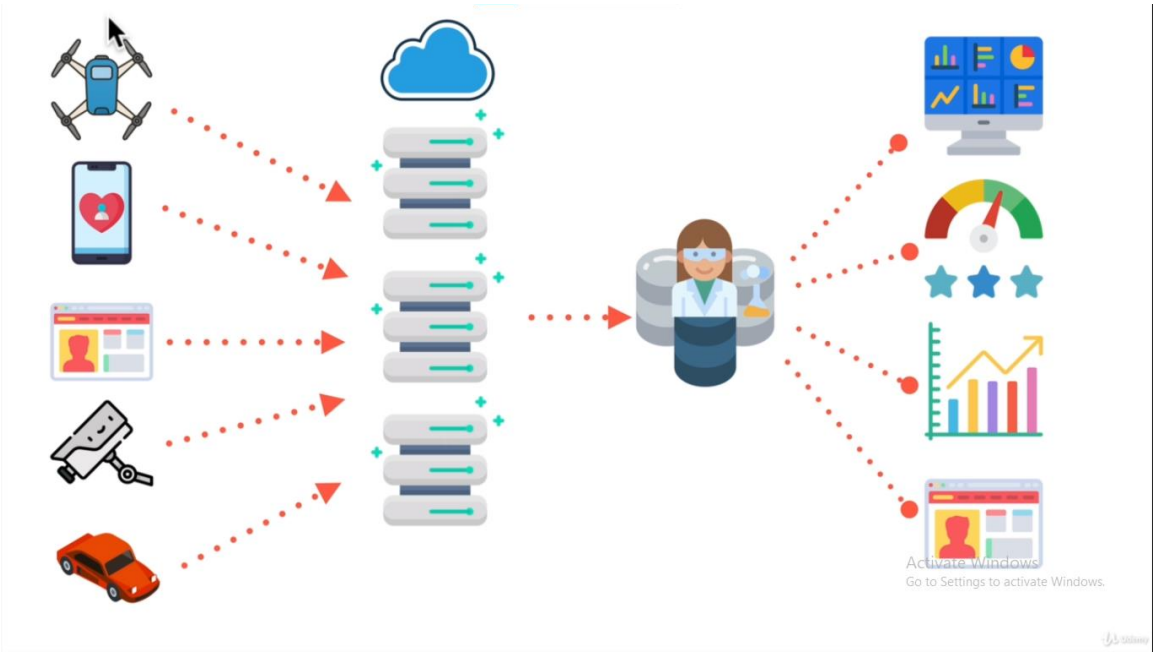
Data Pipeline is essentially a pipeline that a data engineer built to essentially use the fact that we had this big amount of data, we need to extract information from this data using data mining.

So, we need to bring or build a pipeline that allows us to flow from that unknown large amount of data to a pipeline that extracts data to a more useful form.

Data Engineering Work

So, a data engineer creates a data pipeline where all the information that different devices like IOT devices, mobile applications, web apps, cameras, cars and pretty much anything that collects data and stores information or logs data into servers or to the cloud. A data engineer essentially accumulates all this information into nicely packed databases and stores engines, so that different parts of the company can create visualizations. They can monitor the performance of their product. They can get business insights and make business decision from this data and even use this data on their apps for example for user profiles

Before a data scientist or a machine learning expert or even a business intelligence or data analyst gets hired for a big company, the thing that they need to do before all of that can be done is to hire a data engineer they build the pipeline that allows us to work as data scientists.



Tools of Data Engineering

Work of different experts in workflow

Software Engineer (red)

A software developer app developer mobile developer they build programs and apps that users and customers use and that releases data

Data Engineer (Blue)

Data engineer would build this piping and pipeline for us to ingest data and store it in different services like Hadoop like Google big query so that that data can be accessed by the rest of the business.

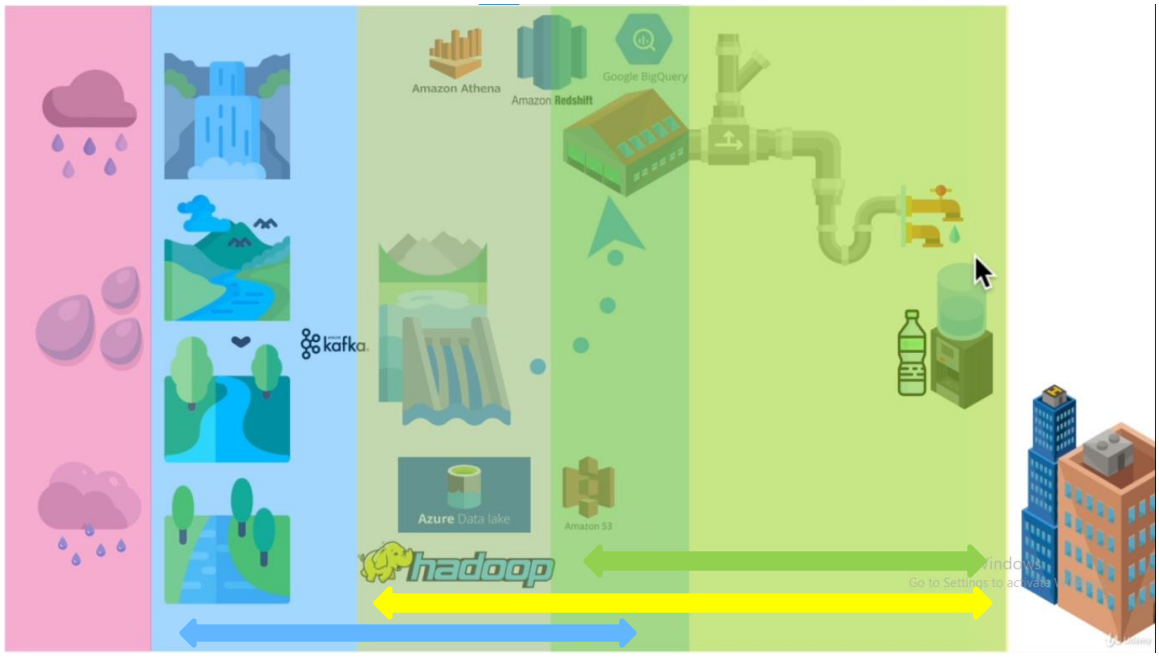
Data Scientists (Yellow)

Data scientists that use the data lake as well as the data scientists to extract information

and deliver some sort of business value.

Data analysts (Green)

Finally, we have data analysts or business intelligence to use something like a data warehouse or structured data to again derive business value.



What data engineering do

Data engineer three main tasks:

1)

They build an E.T.L. pipeline or an extract transform load pipeline

E.T.L. pipeline is the idea that data engineer extracts data that has been generated by all of systems.

They extract that data and then when they extract that data they transform the data into a useful form, that is into a form that can be loaded into a data warehouse.

So, they extract the data, they transform the data and they loaded into something like a data warehouse. So, that the data can be used by the rest of the company.

So that the data can be used by the rest of the company and they used programming languages like Python, go, and Java to accomplish these EDL jobs or extract transform load so that's the first task.

2)

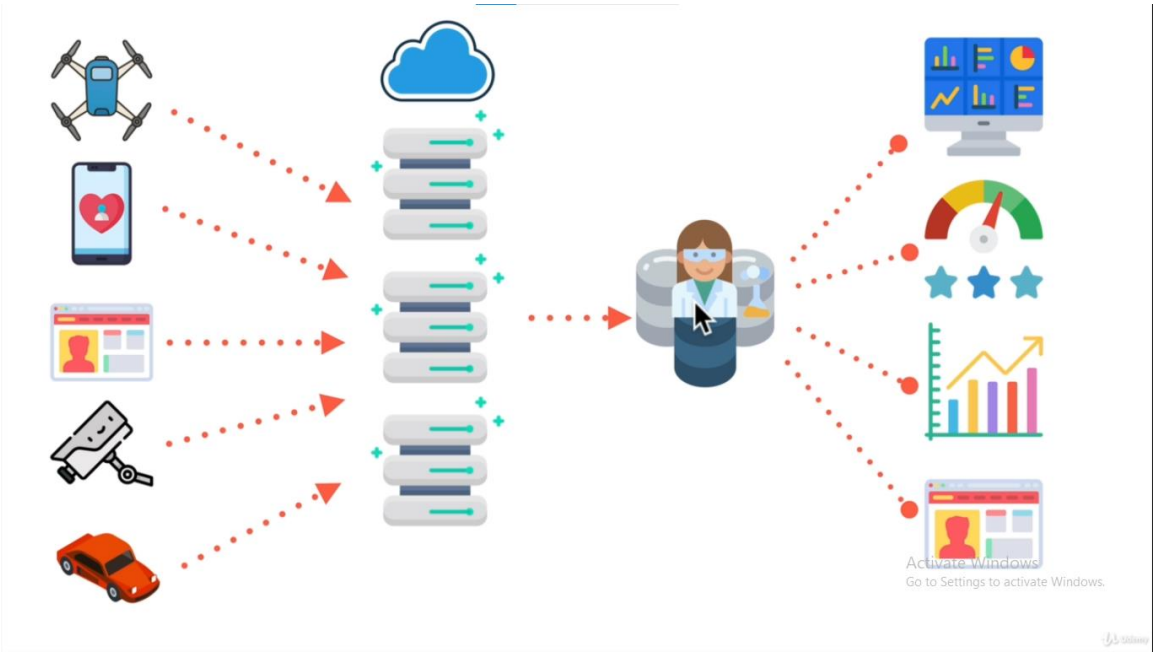
The second one is building analysis tools.

A data engineer needs to make sure that when any part of the system breaks, we are notified.

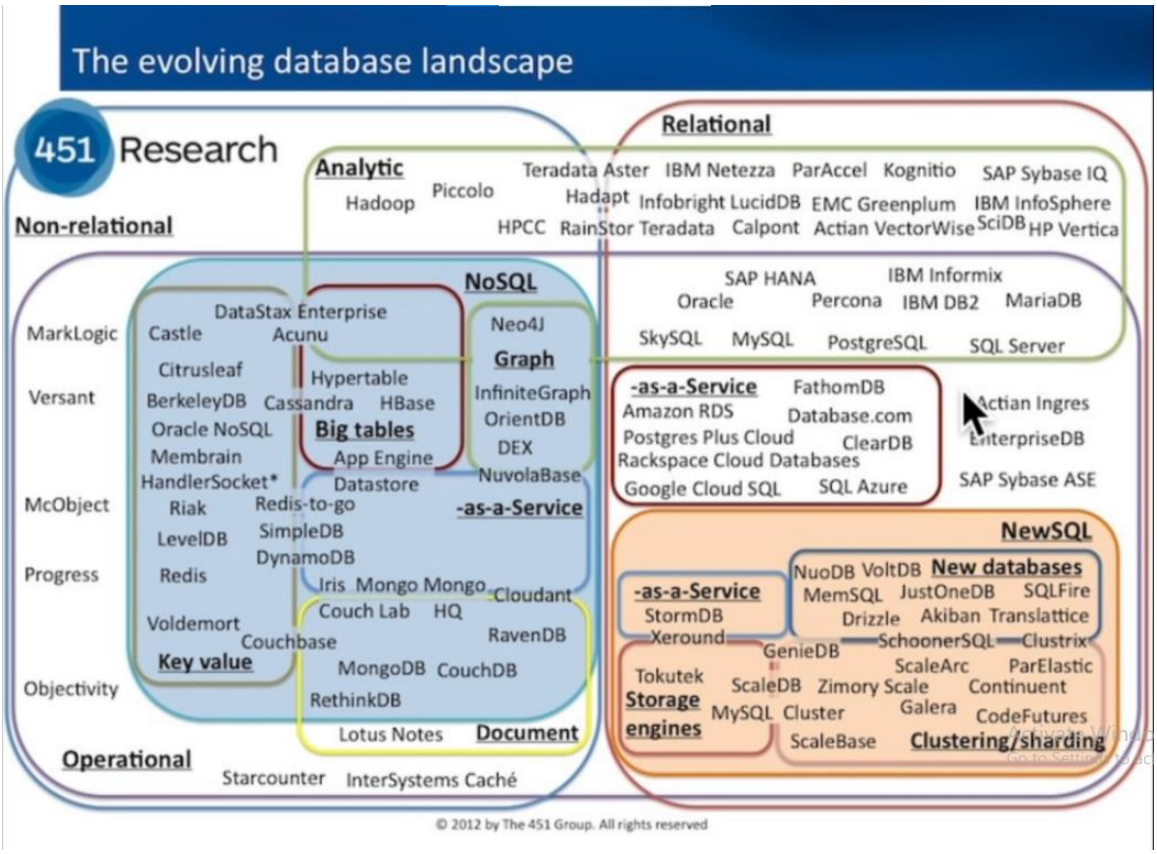
So a data engineer allows data scientists, data analysts, business intelligence people to use tools to analyze the data but also to make sure that the system that they've put in place is running correctly.

3)

Finally, their third main task is obviously to maintain the data warehouse and data lakes that is making sure that everything in there is accessible for other parts of the companies to use.



Type of databases



Databases are just computers, the computers with a hard drive essentially just like your computers and is able to store files and have files available anytime.

DBMS

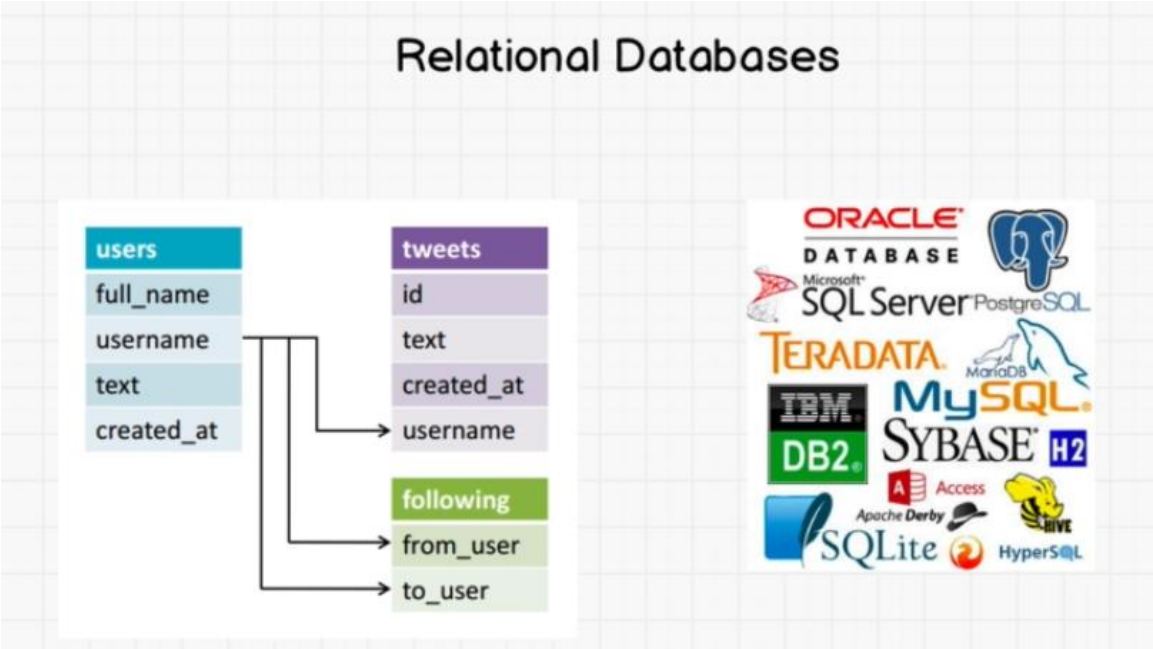
It's a collection of programs which allows us to access databases and work with data.

And it also allows control access to database users

Now there are two types of DBMS that are really popular right now:

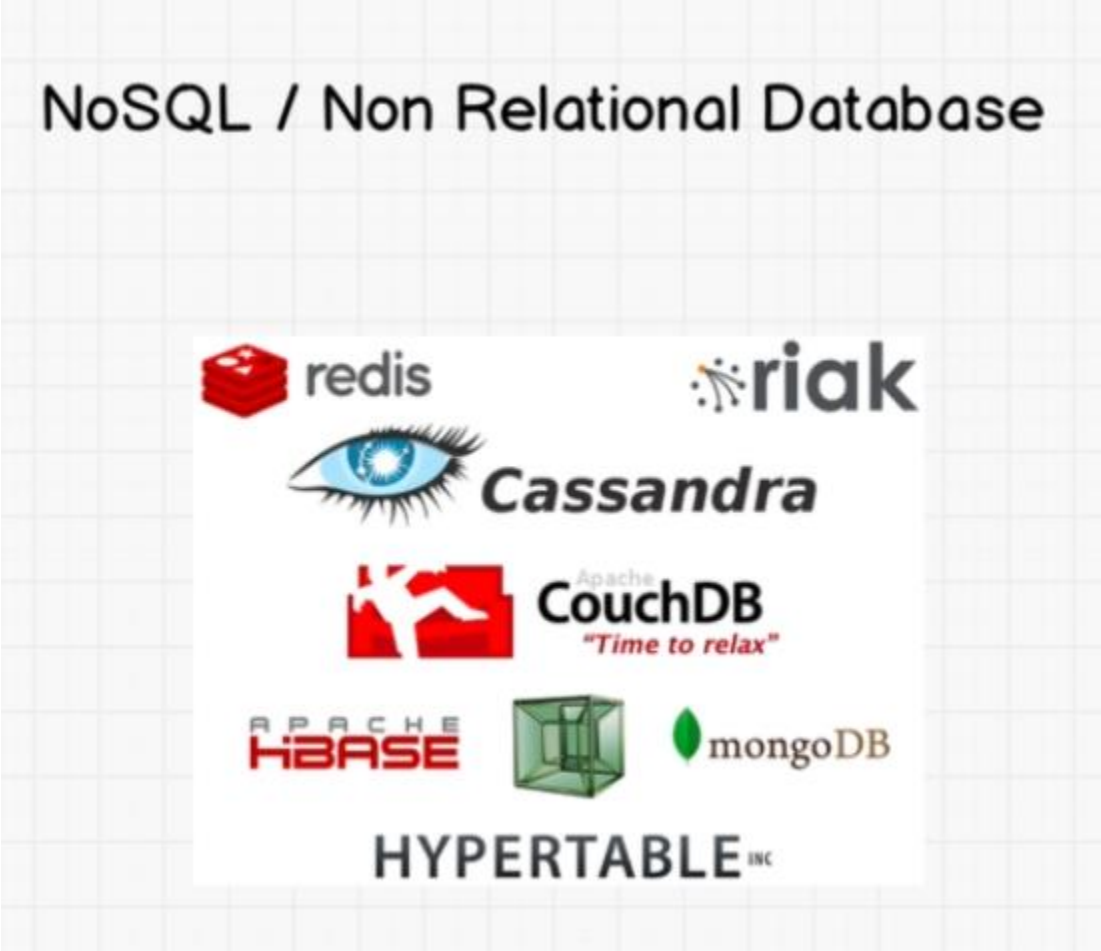
1. PostgreSQL
2. MongoDB

Relational Databases

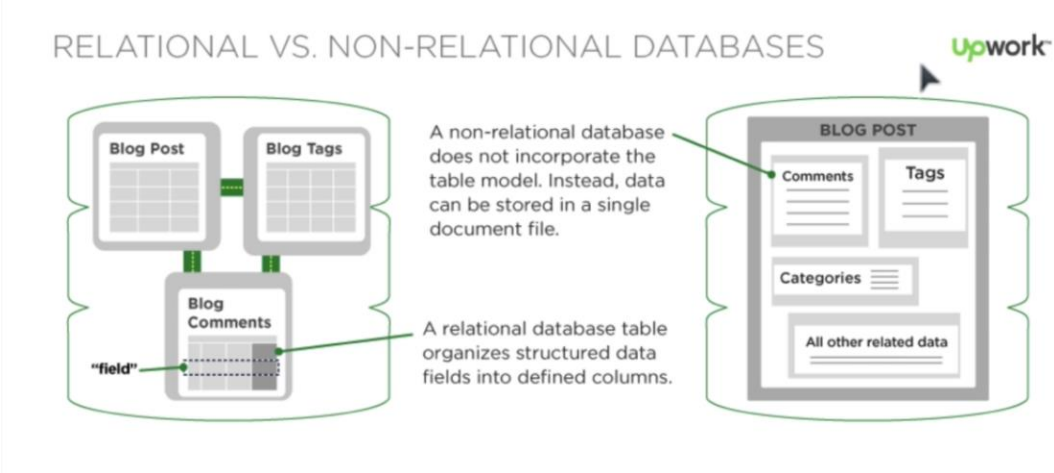


They all follow a same standard format. Relational databases consist of two or more tables with columns and rows. So, in this case(figure) **users** is a table and a **full name, user name**, **text created_at** our columns and whatever values they have here are rows. Each row represents an entry, in each column sorts a very specific type of information like name address or phone numbers and then the relation between tables and field is called a schema in a relational database the schema must be clearly defined before any information can be added.

NoSQL / Non-Relational Database



A non-relational database are more like folders just assembling related information of all types. Now Mongo D.B. is something called document oriented. It stores information as documents.



Hadoop

Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving **massive amounts of data** and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

HDFS

HDFS is a distributed file system that handles large data sets running on commodity hardware.

HDFS a file system just like you have on your computer that allowed it to store files on multiple computers.

That is we can use Hadoop on multiple computers to store as much data as we want.

MapReduce

Once we store data we need to perform some jobs, some processing on that data and map produce in Hadoop allowed us to perform jobs against this data that we had in a data lake using languages like Java or Python.

HIVE

Hive makes your Hadoop cluster feel like it's a relational database even though it isn't. It allows you to use hive and write SQL queries against your HDFS file system.

Apache Spark and Apache Flink

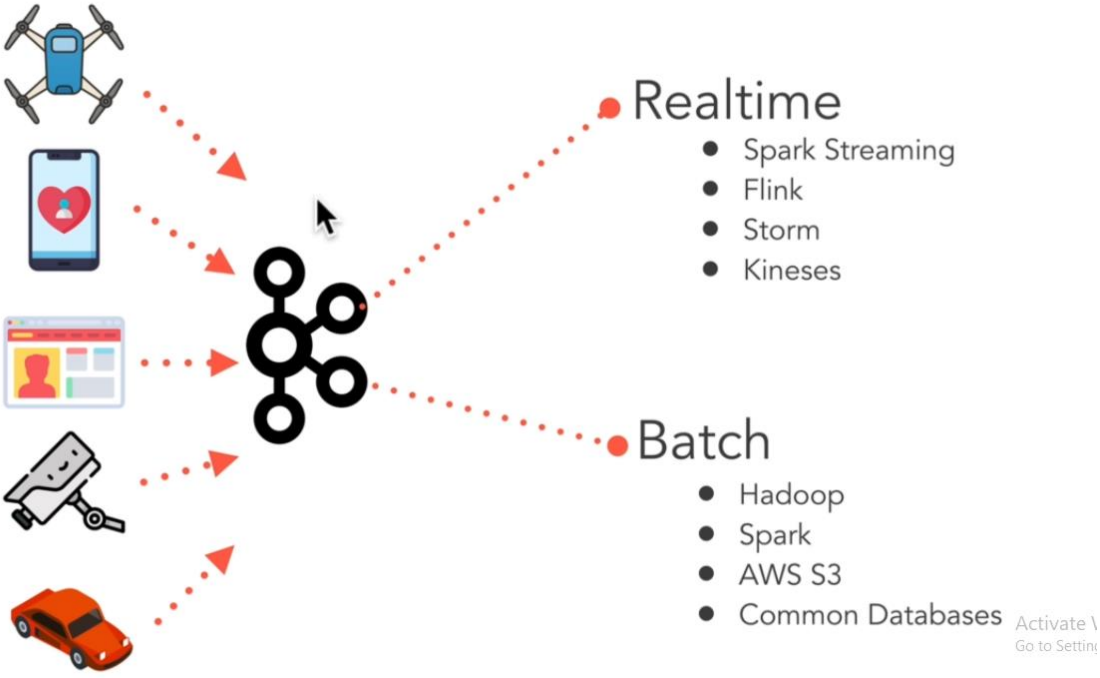
Kafka and Stream Processing

Idea of stream processing is now becoming more and more popular instead of processing data

using just batch jobs let's say every hour or every day.

There's now a movement towards this idea of real time stream processing, that means when data is received processing it right away instead of waiting a bit until we can do batches at a time this allows us to have faster reactions to data.

One of the main tools that's being used, right now for stream processing is something called **Kafka**


Realtime
Spark Streaming
Flink
Storm
Kineses
Batch
Hadoop
Spark
AWS S3
Common Databases