Data Engineering is a field in which we have to deal with huge amounts of data which is stored. And make a pipeline for the business analyst and machine learning engineers.

So basically we are given raw data and we have to make it useful for the people who are going to derive business insight from that data.

we take that raw data and build the pipeline and do transformations along the journey of the data lifecycle where

we might have to perform multiple tasks along the way according to the situation such as ingestion, transformation and serving.

multiple technology stacks will be used to make it happen.

but before we do anything we have to understand how data is generated. at what frequency , what kind of data , and in what variety.

We have to get the full picture of how it's created and at what rate. Before moving forwards , it's essential for us to get to know how data is generated.

The data engineering lifecycle is a small part of the data lifecycle because it's long and tedious and goes through multiple transformations along the way.

source where data is generated is beyond our reach something that we can not control.and it can be anything iot devices , an application message queue , a sensor in an airplane etc.

Anything that generates huge amounts of data can be sourced and that data generated can be used for insight of that machine.

The problem where we can get is the schema of the source where data is generated.and also schema can change over time.

After that we come to the storage part which is also important in data journey.we have to choose a storage solution which is fast in retrieval and puts less computational strain.

This storage is done multiple times in the pipeline. This stage impacts how we are going to use it in later stages.

we have to keep in mind the architecture for read and write speed, can it suffer the process, what kind of storage is it.

frequently accessed data is called hot data ,lukewarm data is in range of week or month, less frequently is cold data in case of failure it is retrieved.

ingestion - so the process from taking the data from source

into another location where we can do data preprocessing is called data ingestion.and the source is out of our control so it might create some problems such as it might stop working,

it might give data of poor quality etc.

transformation - so changing data from one form to another is called transformation of the data.and we have to do these transformations because the data is not in proper format to be used by B analyst , ml team so transformation of data becomes necessary.

converting data columns into numerical form, removing bad values,and applying normalization.

Data featurization is also part of the transformation of data.

We have to understand the business logic to understand the featurization.

serving data -

so after all the ingestion transformation storage we come to the end part of the life cycle this is where we see the report .

machine learning- so machine learning is a sub field of data science where with the help of historical data we try to predict the future.

Or with the help of ml we try to find a function that can map input values to output values.

In data engineering the whole process is done so we can solve business problems and increase revenue, giving some business insight to shareholders.

reverse etl - opposite of the etl that we performed extract transform and load is called reverse etl.

the output we get at end , and is feed back to the source system