Data Warehouse: The data warehouse is the place where the processed data is stored for immediate retrieval and usage.

Data Lakes: The data lakes is the place where the raw data is stored, the data is then processed and sent to the data warehouses. The data lakes consists of all types of data such as structured, semi-structured and unstructured data.

These all follow the ETL flow:

Extract,Transform and Load:

The Extraction can take place in batch processing, where the large chunks of data are taken as batches and moved from source to the destination.

Tools for batch processing are: Blendo and Stich

Stream Processing:The stream processing is done for applications which are time senstive about thev data. The data updation  in real time is critical, in such cases we use stream line processing.

Load:Populating the entire data into the repository, data checks and check if there was any data missing.

Data Pipeline:

The ETL process is the subset of the data pipeline where all batch and stream processings takes place

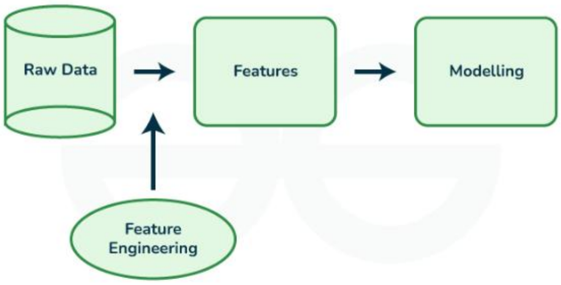
The popular data pipeline tools are : Apache beam, Data flow and the kafta

Big Data:refers to the dynamic ,large and disparate volumes of data created by people tools and machine.It requires new,innovative, and scalable technology to collect, host ,and analytically process the vast amount of data gathered in order to derive real time business insights that relate to consumers,risk,profit,performance,productivity management, and enhanced share holder value.(Formal Definiton)

Feature Engineering:

Feature engineering is the process of transforming raw data into features are suitable or machine learning models

The feature engineering heavily helps in in the success of the machine learning modelsThis feature engineering technique helps in understanding the patterns and relationships of the data making it easy for the machine learning models to learn more effectively.

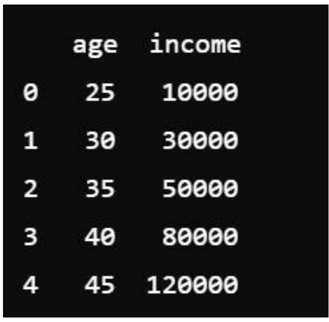


Processes Involved in Feature Engineering are:

1) Feature Creation: By observing the relationships and patterns , we tend to create the pattern by data intuition.

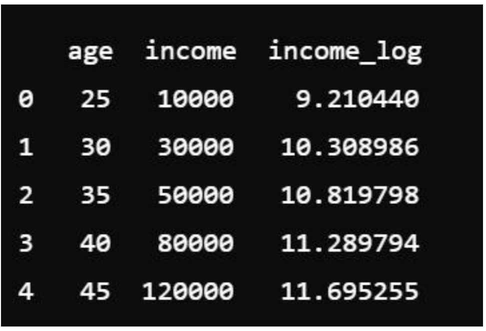
Types of feature creation:
1) Domain-specific: Creating the features based on the business rules or industry standards
2) Data driven:Creating new features such as calculating aggregations,
3) Synthetic: Generating the new features by combining the pre-existing features.

2) Feature Transformation:
   1) Normalization: rescaling the features to have a similar range, such as between 0 and 1, to prevent some features from dominating others.
   2) Scaling: basically standardizing the values to a particular level so that the values can easily be compared. For example 1 dollar equal to 80 rupees here the rupees would be getting the more value so that we would be standardizing it
   3) Encoding: Transforming categorical features into a numerical representation
   Example: one-hot encoding and label coding.The encoding also helps in reducing the dimesinality of the data.
   4) Trasfromation: Transforming the features using mathematical operations to change the distribution or scale of the features.



```
   age  income
0   25   10000
1   30   30000
2   35   50000
3   40   80000
4   45  120000
```

---------------------->



```
   age  income  income_log
0   25   10000    9.210440
1   30   30000   10.308986
2   35   50000   10.819798
3   40   80000   11.289794
4   45  120000   11.695255
```

Here we can observe that as the age increase the income is increasing this is called postive skewness. The postive skewness can be eradicated by applying the logarithmic operations to normalize the features.

3) Feature Extraction:
   1) Dimensionality Reduction--by using the principal component analysis()
   2) Feature Combination:
   3) Feature Aggregation
   4) Feature Transformation

Encoding examples:

```
import pandas as pd
import numpy as np
import seaborn as sns
df = pd.read_csv("Encoding Data.csv")
df.head(10)

#mapping the true or false and yes or no with 0 and 1
df['bin_1'] = df['bin_1'].apply(
    lambda x: 1 if x == 'T' else (0 if x == 'F' else None))
df['bin_2'] = df['bin_2'].apply(
    lambda x: 1 if x == 'Y' else (0 if x == 'N' else None))
sns.countplot(df['bin_1'])
sns.countplot(df['bin_2'])
```

```
#Label Encoding

from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['ord_2'] = le.fit_transform(df['ord_2'])
sns.set(style ="darkgrid")
sns.countplot(df['ord_2'])


#One-Hot Encoding
from sklearn.preprocessing import OneHotEncoder
enc = OneHotEncoder()
enc = enc.fit_transform(df[['nom_0']]).toarray()
encoded_colm = pd.DataFrame(enc)
df = pd.concat([df, encoded_colm], axis=1)
df = df.drop(['nom_0'], axis=1)
df.head(10)
```

Feature Selection: