Data Engineering Day 18

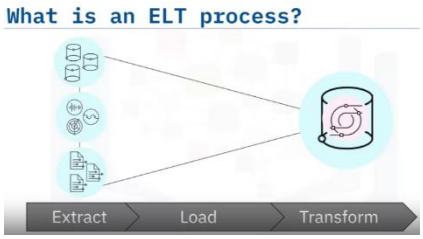
The credit for this course goes to Coursera. Click More

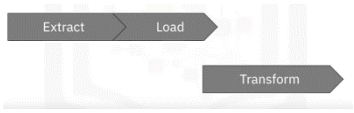
Another link : Azure data Engineer

ETL and Data Pipelines with Shell, Airflow and Kafka

 # Data Processing Techniques:

- **ETL and ELT Process:**
  - ELT can also be called as Extract Load and Transforms
  - It is also one of the methods to implement the Data pipeline methodology in which data will be extracted from various sources like websites, API, or any forms and loaded in databases.
  - The out goal of ETL and ELT are the same, however the order of implementation is different.

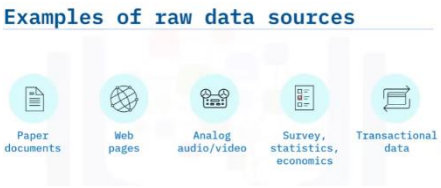




Use cases of ELT:

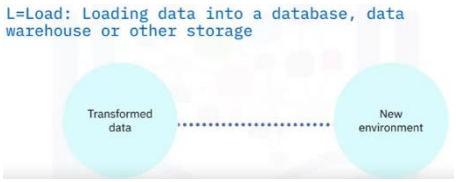
- Demanding the mass scalability of big data
- Computing real time analysis while streaming of big data
- Inter-networking with large, distributed data sources.

- **Why ELT is emerging:**
  - Cloud computing and Big Data are the main reasons for ELT being born.
  - Clear or distinct boundaries between transferring (moving data) and computing (processing data).
  - More flexibility

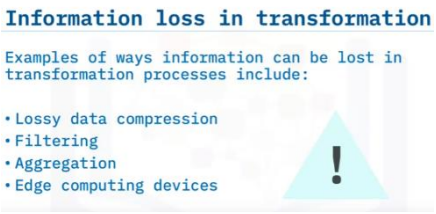


- More precise (No information / Data loss)

- **Extractions:**
  - Get access to data from various sources like IoT devices /sensors /camera and reading it in the applications for analysis.
  - Web scramming
  - Getting access to the data by using API's.
  - Data could both static and stream online.
  - Examples are weather stations data, social networking feeds, IoT devices.



- **Transformations:**
  - Processing data
  - Key responsibilities include cleaning, filtering, Joining, Feature Engineering, formatting, and data typing.
- **Loading:**
  - Moving or shifting data to new destinations such as databases, data warehouse, data lake or data mart.

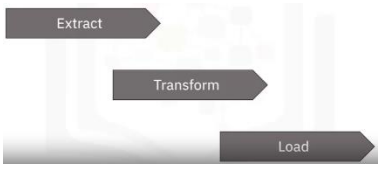


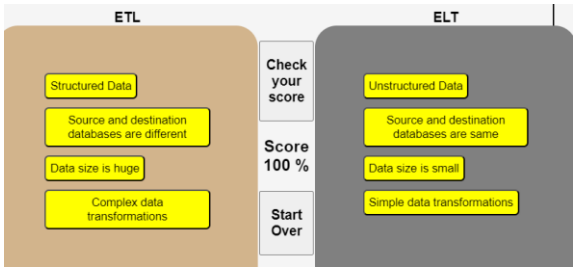
- **Difference between ETL and ELT:**

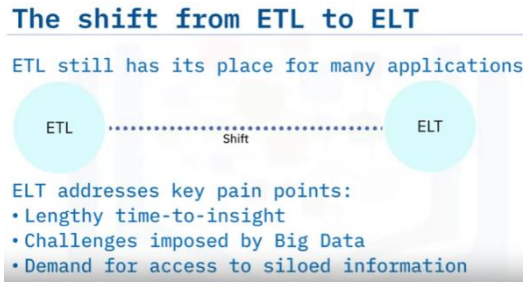
ETL	ELT
Transformations of data like filtering and stuffs happens within a pipeline.	Transformations of data like filtering and stuffs happens in destination environment.
It is a rigid method	Is flexible and allows end users to build their own transformation
It used structured or relational data and faces difficulties for scalability	Handles both structured and unstructured data where it resolves the scalability issues.
Takes time for versioning and development	Takes less time or shorter durations for versioning where it provides flexibility for user to create their own user dashboard and modify it as required.
Classical way of extracting, loading, and transferring data	Modern version of ETL where it handles more raw data for extract load and transfers.

 # Data Extracting Techniques:

- OCR: transferring papers docs to computer readable files.
- Analog to digital converters, CCD sampling
- Mail, phone, or in-person survey and polls.
- Web scraping
- API's
- Database querying
- Edge computing
- Biomedical devices



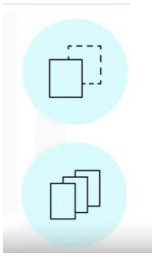




 # Data Transformations Techniques:

- Data transformations includes various operations such as:
  - Data typing
  - Data structuring
  - Anonymizing and encryptions
  - Cleaning for duplicate and missing values
  - Normalization where data is converted into same units for easy interpretations.
  - Filtering, sorting, aggregating, binning (small or least significant values are replaced by either mean or median for better analyzing)

 # Data Transformations Techniques:



- **full loading**
  - loading data into one large batch
  - used for tracking transactions in a new data warehouse.
  - Used for porting over transitions histories.
- **Incremental loading**
  - Data is appended to, not over written.
  - Used for accumulating transitions history.
- **Schedule**
  - Periodic loading like updating daily transitions or selling histories to the database.
- **On-demand**
- **Batch and stream.**
- **Push and pull.**
- **Parallel and serial**