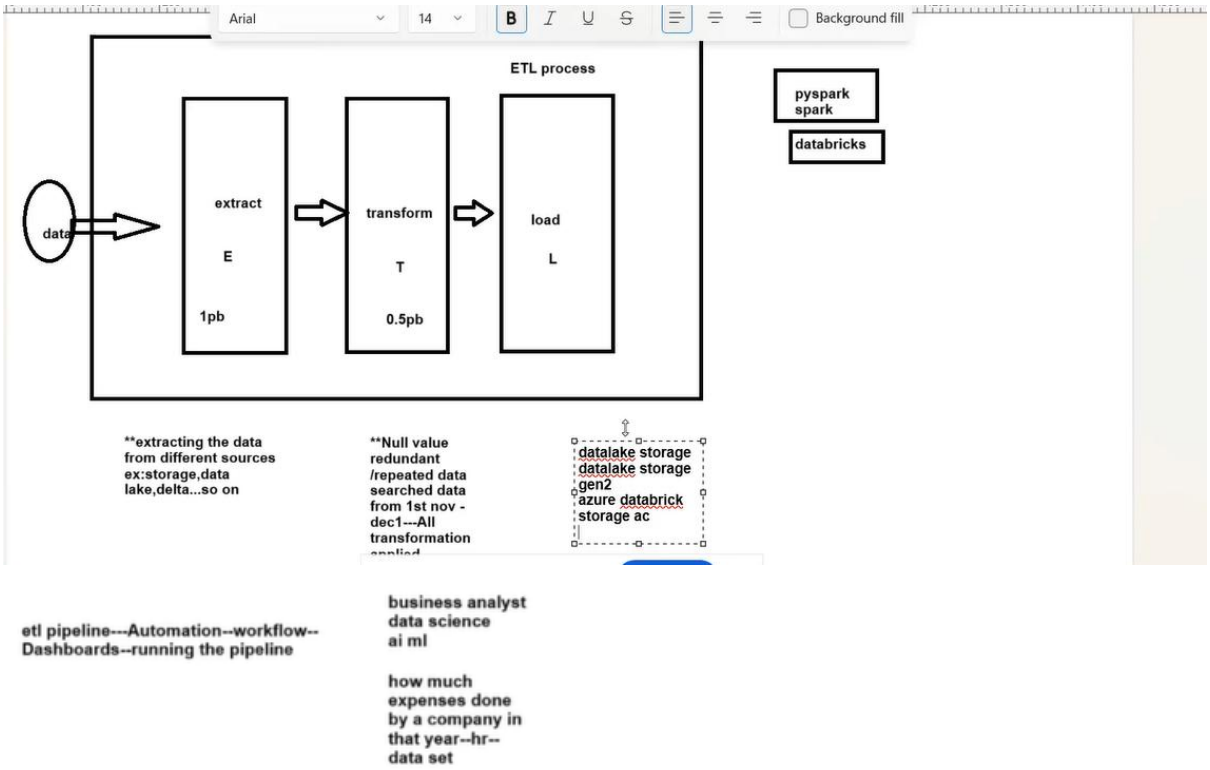
Day5 notes



Why PySpark for ETL?

1. Performance
2. Ease of Use
3. Scalability
4. Rich Ecosystem

- ETL pipeline automates the ETL process.
- In dashboards, the history, activity of data is stored
- When all queries for transformation are written in single notebook, if we run the pipeline, it automatically it extracts data and applies all the queries in transformation phase and output is received through proper validations.
- For data visualization, Azure synapse is used
- Real World Applications:
    - Perform Large scaledata cleansing and preparation
    - For Ai,ML applications
    - Analysing streaming data in real time
    - Handle structured and unstructured data effortlessly

- Scheduling is done to run the entire pipeline
- Testing: Testing incremental load data

Implementing incremental data load process, the data warehouse stays up to date with latest sales transactions without needing to reload entire dataset each time.

Step by step Implementation of ETL process for a Data Warehouse:

- Identify Incremental changes
- Capture changed data
- Transform Incremental Data
- Merge with existing data
- Update MetaData
-