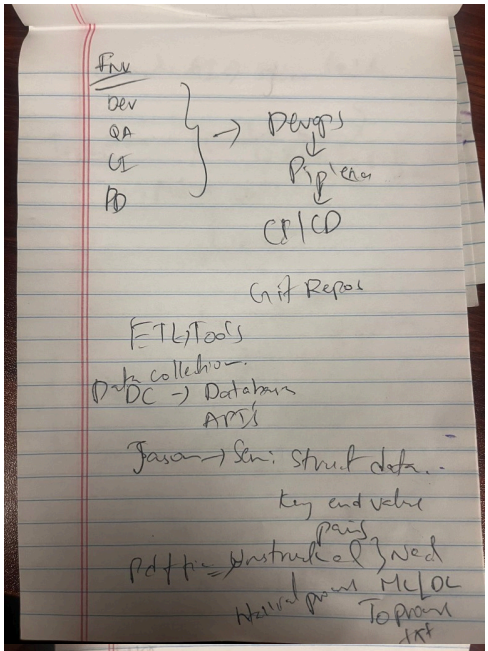
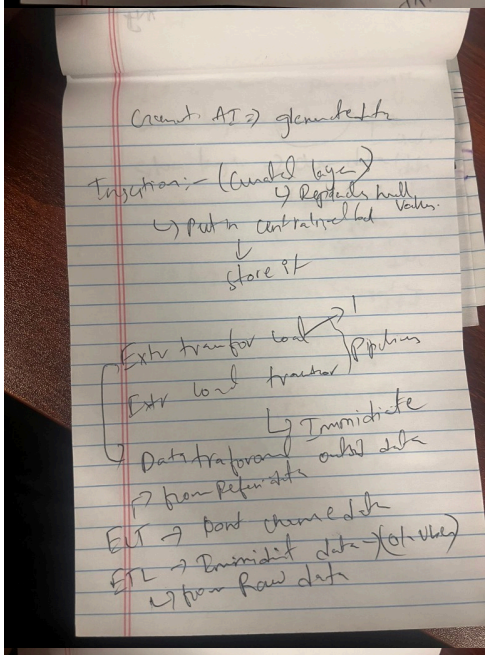
Day-1

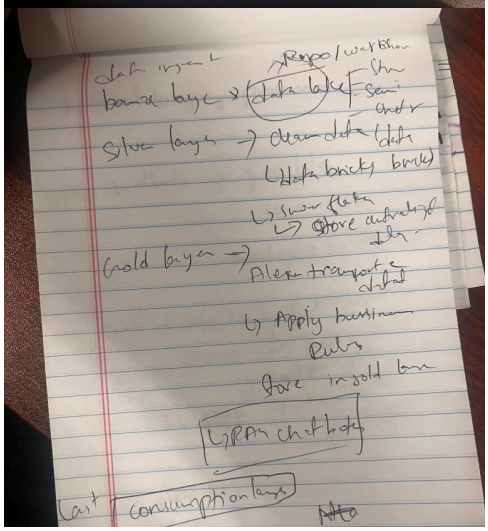
Wednesday, January 21, 2026 7:24 PM



- **“Every data has a story to tell.”**
- The role of:
- **Data Engineers** → build pipelines & infrastructure
- **Data Scientists / Analysts** → extract insights and meaning
- Your job is to **give voice to data** through processing and analysis.



- **Data Collection**
- Databases: MySQL, PostgreSQL, MongoDB
- APIs
- Files: CSV, JSON, XML
- Web scraping (BeautifulSoup, Scrapy)
- Event streams
- **Data Types**
- **Structured** → relational tables
- **Semi-structured** → JSON, XML
- **Unstructured** → PDFs, images, videos
- **Data Ingestion**
- Moving data into centralized storage
- Methods:
- **ETL** (Extract → Transform → Load)
- **ELT** (Extract → Load → Transform)

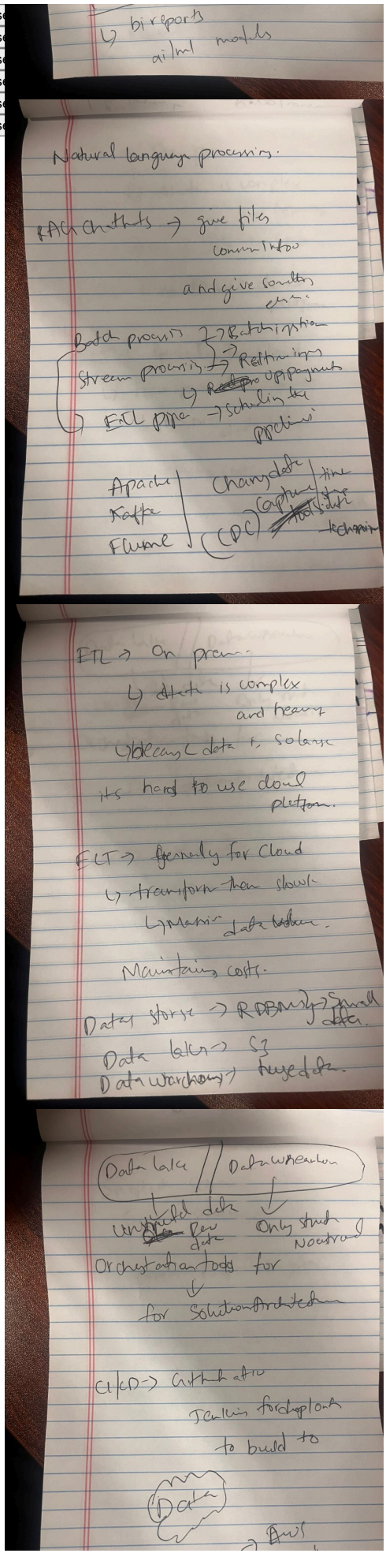


- **Batch Processing**
- Scheduled runs (daily, weekly, monthly)
- Used for reports, analytics, historical data
- **Stream / Real-Time Processing**
- Instant processing (Uber, UPI, food delivery)
- Tools mentioned: **Kafka, Apache NiFi**
- **CDC (Change Data Capture)**
- Captures only changed data using timestamps
- Efficient for real-time updates

- **Bronze Layer (Data Lake)**
- Raw, unprocessed data
- **Silver Layer**
- Cleaned and processed data
- **Gold Layer**
- Business-ready data with rules applied

- **Databases**
- Smaller, structured data
- **Data Warehouses**
- Large historical data
- Uses **fact & dimension tables**
- **Star schema / Snowflake schema**

Choose
Choose
Choose
Choose
Choose
Choose



- **Data Lakes**
- Raw, semi-structured & unstructured data

- **Power BI dashboards**
- Reporting & analytics
- Machine Learning models
- Predictive analytics
- NLP & Generative AI use cases

- **Data Governance**
- Security, privacy, compliance
- Data masking, NDAs, sensitive data handling
- **DevOps**
- CI/CD pipelines
- Managing dev → test → prod environments
- **Orchestration**
- Tools like **Airflow, Azure Data Factory (ADF)**
- Schedule & monitor pipelines

Data warehouses are designed to use parallel processing so that large volumes of data can
and processed faster by dividing the workload across multiple CPUs or nodes simultaneou

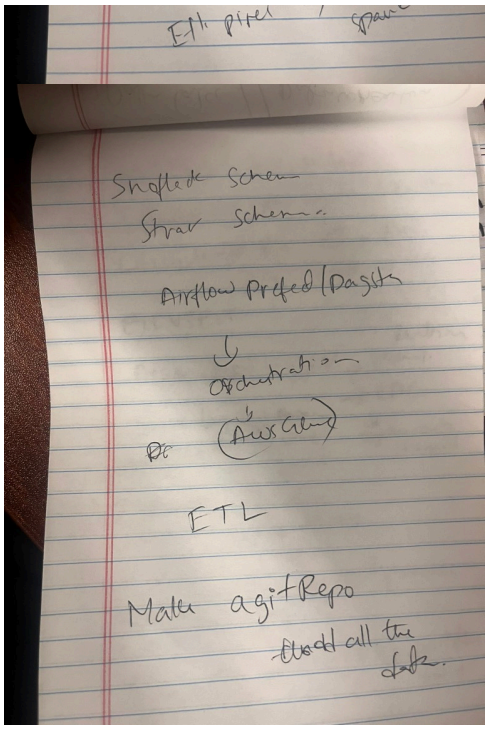
Parallel processing

Instead of:

- 1 CPU → slow

You get:

- Many CPUs / nodes → fast


ETL pipel
Snoflak Schem
Star Schem
Airflow Prefect Dagsta
Orchestrati
Aws Glue
ETL
Make a git Repo
that all the
data.