ETL Concepts

By Kusuma Mam

26th March 2022

ETL

Source → transform → Target

why
- integration
- clean
- consolidate
- faster query, performance
- ↓ load on source data system
- aggregation
- homogenize data format/units
- queryable format

Data
- collection of facts about obj/person
- Quant/Qual

OLTP — Online TransProcess

ATM, POS (Point of Sale)

OLAP — Online Analytical processing

- classification/computation

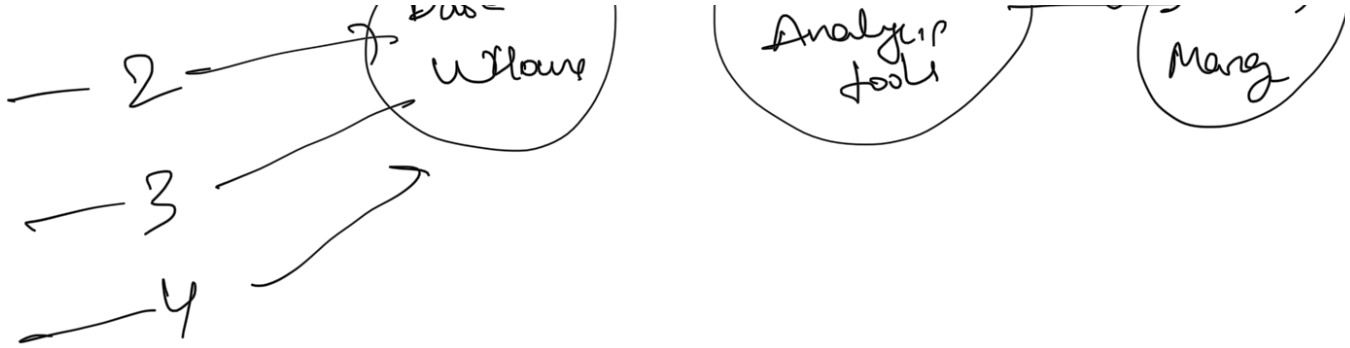
Ex: BI (Business Intelligence) Reports

Data Warehouse
- Large store of data accumulated from diff sources

Data Model

~~RDB~~ Pizza

Menu — OLAP



ETL → copy data from source dbs which
represents differently from same
→ based on consumers of data

Exa

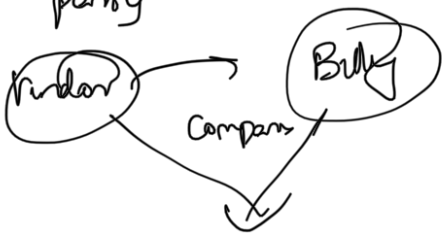
many dbs → Data warehouse

Migration

Unstructured data → store in BigData
Hadoop

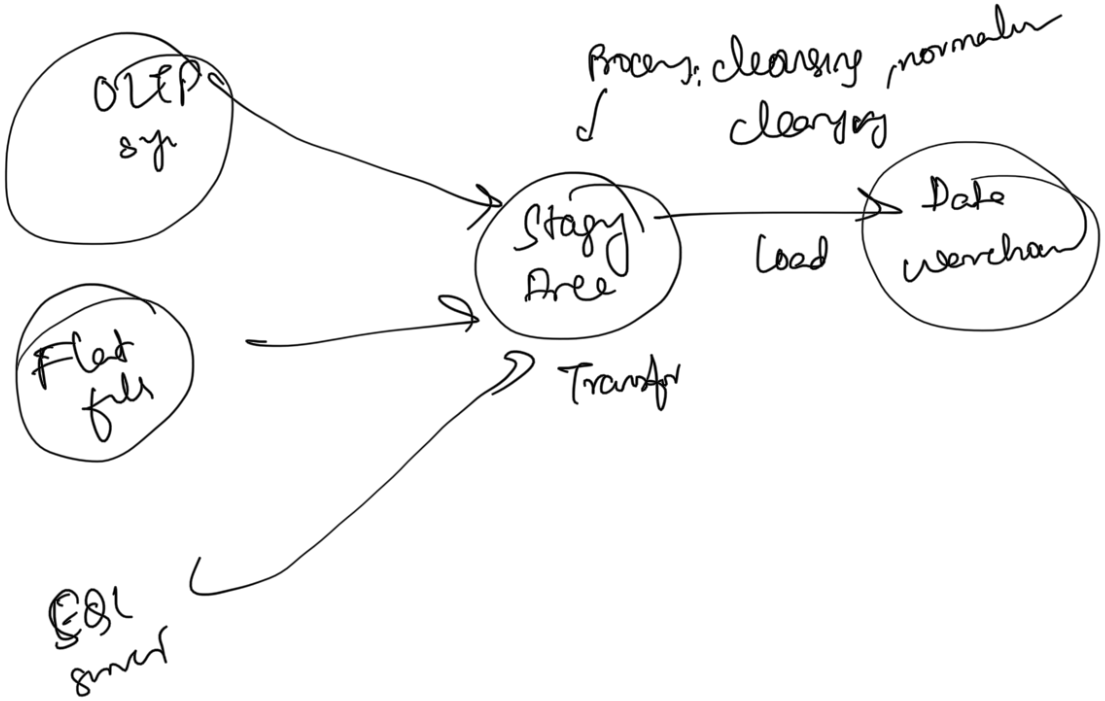
Exi: oracle to Hadoop
ETL

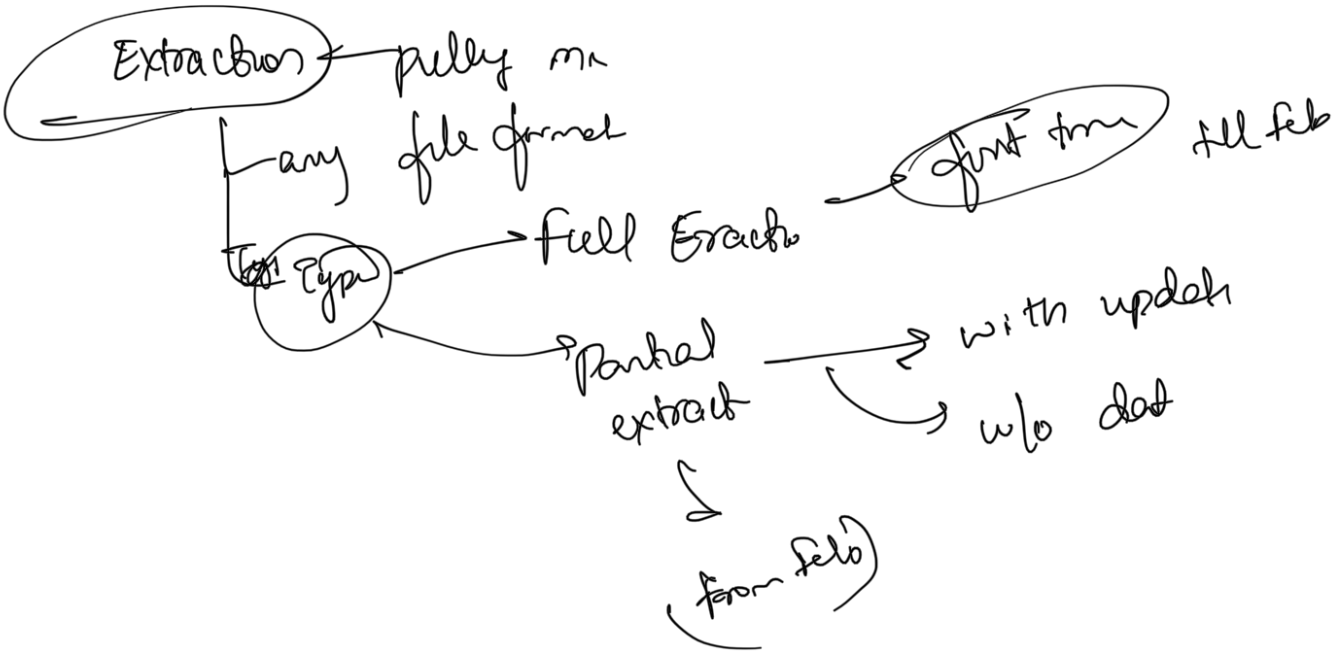
- Mergers and acquisitions
- 3rd party

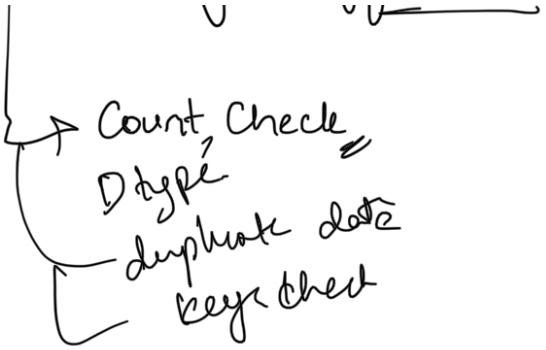


Data Integration
↳ single view of Truth.

- Nanny / Symbol. changes
- Abbreviation







Dirty Data

- unwanted
- has null → cost (null) → not counted
- diff abbrevs
- standards units
- Duplicate data
- Inconsistency.

Master Table entry should be same as OLTP table

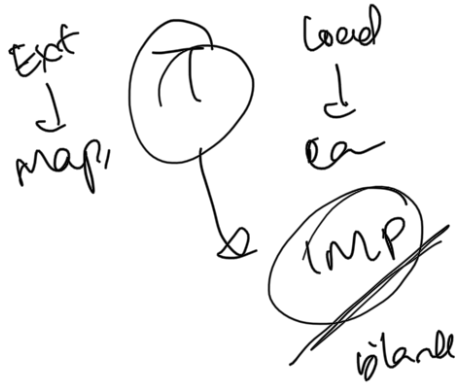
Data Transformation & Cleansing

Transformations

- ML → not always

Standardisation

Currency → changes everyday



blank S name to AAAA.

concatenate ETL learn

D Transform & D Cleansing Steps

OLTP → filter → D.wh

- Filtering
- set rules for for map
  - conversions
  - check | threshold value, date validations
  - exception handling
- Clean (Null → 0,
  male → M
- Look Up Table (Like Menu)

D Transform Validation

count key check

D Quality Check

→ Bef & After

- Null Value Check
- Referential Check
- Range Check

Process → Exception Tables

- List of value
  - Blanks
  - Duplicate

outliers

- Date len

Blank Vs Null

1 space

0

cont

Loading

↳ Load fails, Recover mechs

Types
- Initial
- Incremental
- Full Refresh

Load Techniques
- Batch load.
- Seq. load
- Parallel
- Incremental

Loading Verifications

↳ Key field

ETL tasks

↳ Understand Data to be used for report
- Review Data Mo (Target column needs
  ↳ J

Best Practices for ETL proc

- Never try to clean all data
- Clean something
- Determine cost of cleansy data
- Speed up query process

Example: delivery
holiday: done-in

ETL tools

Enterprise

Oracle Data Integrator
Informatica

Open Source

Jasper ETL
Apache Camel
Scriptella
Apache Nifi
&q

How to choose ETL

Volume
Flexible data acquisition

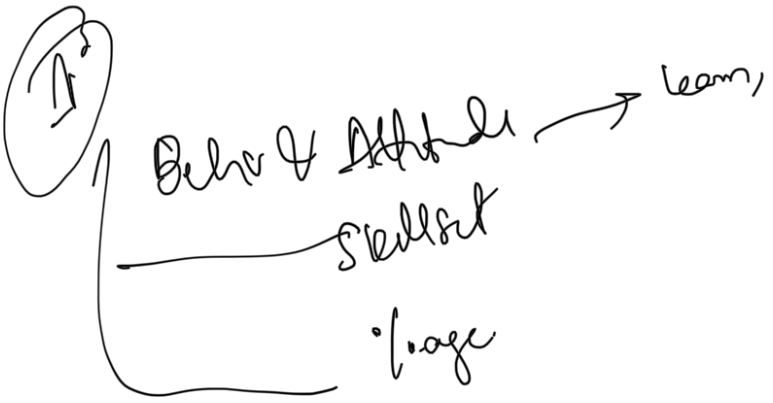
Transform

manage data in source & derived

Summary

Core study

worksheet



Team player

imp than self-centred

Appraisal: Team Minded

Team

Team completion vs whole proj completion

after doing,

help others

~~Start~~ Company
Salary
Learning opp

Big Org: Learn, Top roles