Papers :-

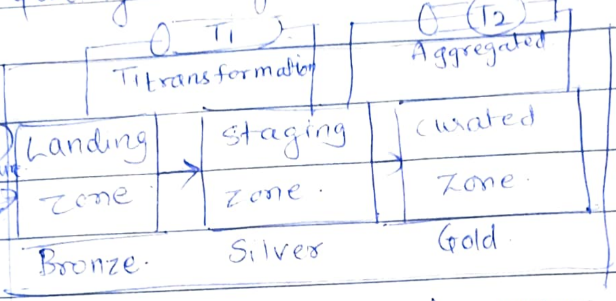
Date:

Data lake design

→ Objective — leverage data insights
how quickly u go to insights.

⇒ horizontal split (partitioned / buck
z-ordered

External (EL) source
1) (Row oriented data)
2) Rawdata / unstructure
ii) schemaless data
Batch
streaming



(Hive)
Metastore.
→ mapping betn logical name space to physical assets.

5) Normalized (No redundancy).
6) compress data before ingestion.

Datalake onpremise - HDFS.
cloud - storage S3/ADLS.

Landing Zone - specific location on memory. → for particular subject areas.
eg. 1) Amazon apparels landing zone.
2) Retail business/subject zones are landing zone.
3) saggrigated by time zone.

T1 transforⁿ

Staging zone - T1 tranⁿ ← scheduled
triggered.

(more widely used) Triggered — message queue → publisher / subscriber
→ queueing system.
→ because of variable injestion triggered is used.
→ staging zone is called source of all truth. all model u build, all dashboard u monitored, this is the zone. It has to be versioned (immutable)
→ Data is denormalized (join table and quickly get it)
→ have notion of downstream applⁿ.
→ It is done in non linear fashion.
→ columnar data. (parquette format) OCR
→ horizontal split (partitioned / bucketed).
→ data here stays for longer duration
→ during transformation, data cleaning is done. (missing value not in perspective of ML model)
→ Keep marker to imputed data.
→ in healthcare missing data should not be imputed.
→ More imputation more noise.

Note

→ compressed file can not be split for ~~spark~~ spark parallel processing

→ Splittable compression is used.

→ Understand ingestion time, type of data, CPU uncompressing time and compressdata if needed.

day end → operationaldata ingestion → Trans – aggregate – available for next Busi Day.
token / churning / coupons. prediction.

Benchmarking – find optimum parameters.

Benchmarking – 1TB of data – how much time it takes.

2 TB – in same time → scale out → clouds are popular.

→ Access controllist →. Landing zone access is subjective → file level read access

→ Staging zone → who has access to particular column, partition.

→ Landing zone access → Data engineers / Staging zone – data analyst.

→ Data types conversion.

Star / Snowflake Schema

→ schemabased datahere., schema is enforced here.

→ Two types of tables, facts. and dimensions.

facts – Business operations are captured by facts.

Date, cid, $\overset{tx}{\text{tid}}$, $\overset{duct}{\text{Pid}}$, $\overset{store}{\text{sid}}$, Amt. → T^x data. facts.

information about customer
" Product.
" Store.
Referencedata. → dimensions

Curated Zone – precomputed statistics.

→ management wants to know KPIS. monthly (aggregation).

→ freq of customer, regionwise

→ which store has – how many pitfalls.

→ moving average of revenue of last 4 weeks.

Two core objective → ① Enable the lead time.
② sources of features of ML model / statistical inference.

→ Some organization may push this data to datawareh / featurestores
↳ source of all truth ML

→ data here is in small size

→ data is stored mostly row oriented fashion.

→ This is called referce architecture. not standard. architecture

small organization may follow only two zones.

Large organization may have more than three zone.

→ metastores usually used in staging zone.