GoodLuck | Page No.
Date

4. Introducing ETL

1] Round up requirements
① business needs
② compliance
③ data quality
④ security
⑤ data integration
⑥ data latency
⑦ archiving
⑧ available skills

2] business needs
- It is the info required by business users to make decisions

3] compliance
- It ensures reported numbers are accurate & complete
- used for supplying proof of security of data copies over time

4] Data quality
- There are 3 imp demands -
① Data is imp for users to function
② Data is distributed -
③ increase in demand for compliance

5] security
- which data to publish & to whom
- seek advice from security manager.

6] Data integration
- It makes all systems work together seamlessly
- CD - means setting common dimensional attributes across business processes

ROYAL

- CAs - means making agreements on common business metrics

7] Data latency
- how quickly source data must be delivered to business users
- It affects ETL architecture

8] Archiving
- Archiving means staging data after each major activity in ETL process.
- Staged data is archieved on permanent storage device
- Each archieved data should have metadata describing origin & processing steps

9] User delivery interfaces
- It should work closely with modelling team to deliver data in a format which is easy to understand, well structured & easy to retrieve

10] Steps in ETL
- (1) Extracting
  (2) cleaning & conforming
  (3) Delivering
  (4) Managing

11] Extracting data
(1) subsystem 1 → data profiling
- Plays two roles - strategic & tactical

(2) subsystem 2 → Change data capture system
- Goals are -
  1. capture all changes
  2. tag changed data with reason codes

GoodLuck

3. support compliance tracking.

③ subsystem 3 - Extract system
- use file or stream
- cleaning & conforming data

④ subsystem 4 - Data cleansing system
- fix impure data
- goals are
1. early diagnosis of data quality issues
2. requirements for source systems & integration efforts to supply better data.

⑤ subsystem 5 - Error event schema
- It is the centralised dimensional schema to record every error event thrown by quality screen

⑥ subsystem 6 - Audit Dimension assembler
- It contains metadata context at moment when specific fact table record is created.

⑦ subsystem 7 - Duplication system

⑧ subsystem 8 - conforming system
It is responsible for maintaining conformed dimensions & conformed facts.

⑨ subsystem 9 - Slowly changing dimension manager
- The ETL system must determine how to handle a dimension attribute value that has changed from the value already stored in the data warehouse.