Introducing ETL

Round up Requirements

- Business Needs
- -information required by business users to make decisions
- -business needs drive choice of data sources and their transformation
- -maintain dialog amoung ETL team, data architect and modelers,business analyst and business users.

- Compliance
- -ensure reported numbers are accurate, complete
- -saving archived copies of data sources and subsequent data staging
- -providing proof of complete transaction flow that changed data
- -supplying proof of security of data copies over time.

- Data Quality
- -Three important demands
- -First Data is important for users to function
- -Second data is distributed
- -Third increase in demands for compliance
- Security
- -which data to publish and to whom
- -seek advice from security manager

- Data Integration
- -it makes all systems work together seamlessly
- Serious data integration must take place amoung transaction systems
- -takes form of conforming dimensions and conforming facts in DW
- -CDs means setting common dimensional attributes across business processes
- -CAs means making agreements on common business metrics

- Data Latency
- -how quickly source data must be delivered to business users
- -it affect ETL architecture
- -good algorithm, parallelization and good hardware can speed up batch oriented data
- -ETL must convert batch processing to stream processing if latency is most urgent

- Archiving
- -staging data after each major activity in ETL process
- -archive all the staged data on permanent storage device
- -each staged /archived data should have metadata describing origin and processing steps that produced data
- User delivery Interfaces
- -should work closely with modeling team to deliver data in a format which is easy to understand, well structured and easy to retrieve

- Available skills
- -ETL design should be based on resources and skills available
- -decide to buy vendors package or hand code
- -technical issues and license cost

Steps In ETL

- Extracting
- Cleaning and conforming
- Delivering
- Managing

Extracting Data

SS1-Data Profiling

- -use tools
- -plays two roles strategic and tactical

SS2- Change Data capture System

- -goals are
- -Isolate changed source data to allow selective processing
- -Capture all changes
- -tag changed data with reason codes
- -support compliance tracking with additional meta data
- -perform CDC early before bulk data transfer to DW

Audit columns

- -source is appended with audit columns to store data and data and time a record was modified
- -columns are populated via database triggers
- -sometimes columns populated with source applications

Timed extracts

- -select all rows where data in create or modified date fields equal SYSDATE-1(yesterdays)
- -unreliable
- -duplicate rows when restarted from mid process failures

Full diff compare

- -full snapshot of yesterdays data and compare with today's data for what has changed
- -it is detailed process
- -time consuming
- -do on source machine

Database log scraping

- -snapshot of redo log at scheduled time for transactions identifications which changes table required for ETL
- -what if log is full

SS3- Extract System

- -use file or stream
- Cleaning and conforming data

SS4 Data cleansing system

- -fix impure data
- -architecture for cleansing data
- -goals are
- -early diagnosis of data quality issues
- -requirements for source systems and integration efforts to supply better data

- -ability to provide specific description of data errors expected to be encountered in ETL
- -Framework for capturing all data quality errors

Quality screens

- -Each quality screen is a test
- -if test fails drop error record into error event schema
- -if test fails chose to halt process, send offending data into suspension or tag the data

- -column screen test data within single column
- -whether column contains unexpected null values, value falls outside range, or value fails to adhere to required format
- -structure screen test relationship of data across columns

Responding to quality event

- -halt process
- -sending offending record to suspense file for later processing
- -Tagging data and passing it through next step in pipeline

Subsystem 5 Error Event Schema

- -centralized dimensional schema to record every error event thrown by quality screen

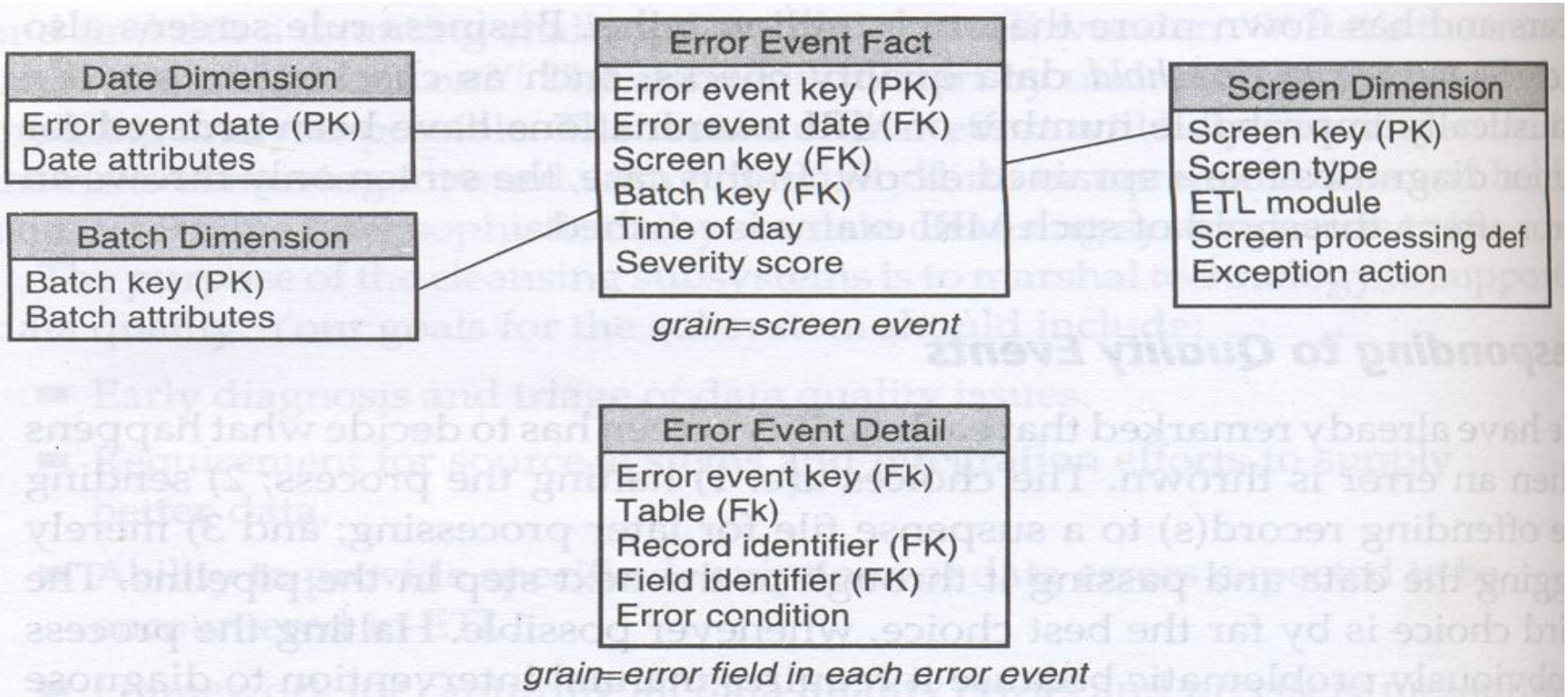

Figure 9-1 Error event schema.

Subsystem 6 Audit Dimension Assembler

- -special dimension assembled by ETL for each fact table
- -contains metadata context at moment when specific fact table record is created

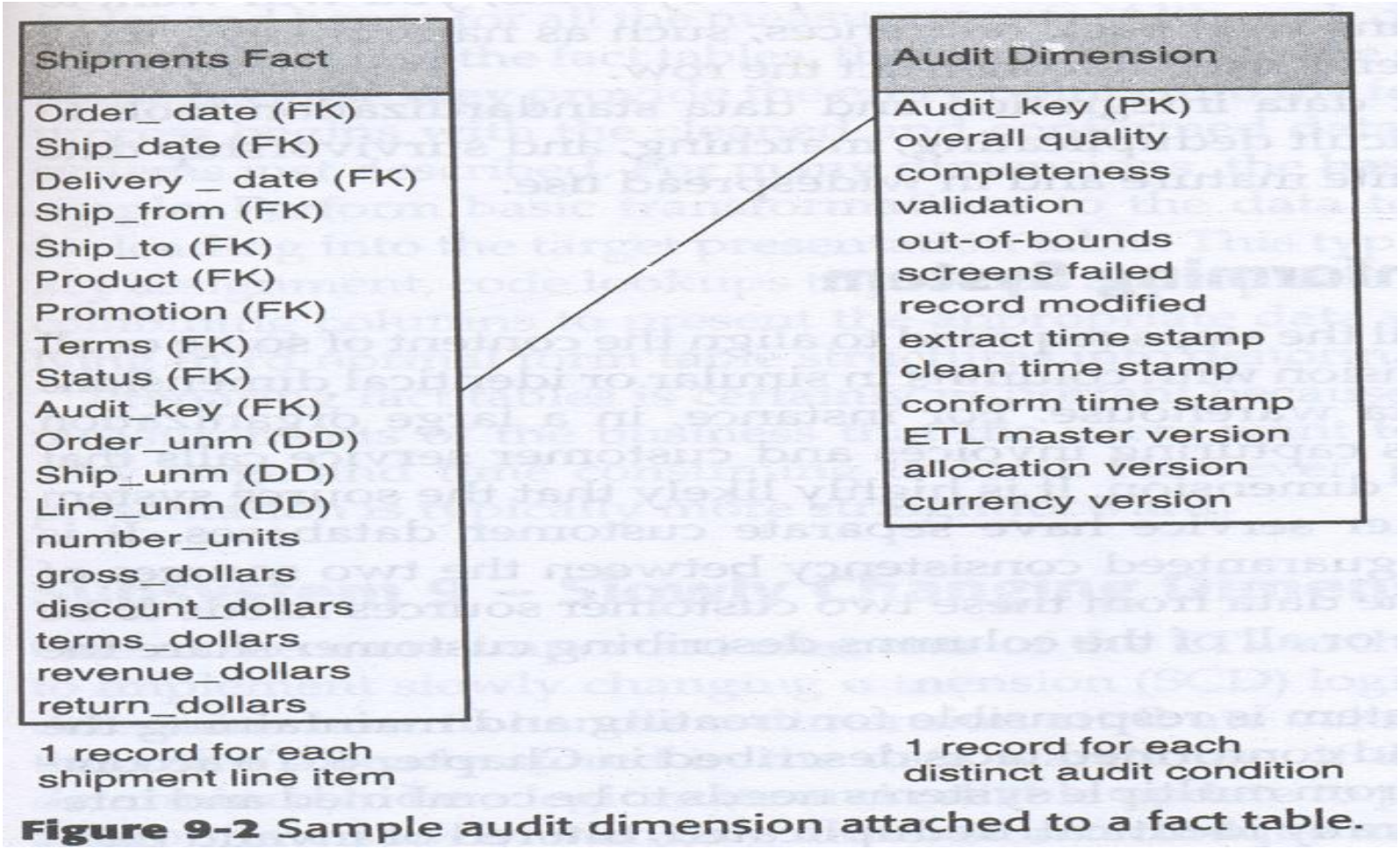

Figure 9-2 Sample audit dimension attached to a fact table.

- Subsystem 7 -Deduplication System
- Subsystem 8 –Conforming System
- -steps required to align contents of some or all columns across dimensions
- -fact tables for invoices and customer service calls that use customer dimensions
- -responsible for maintaining conformed dimensions and conformed facts

Delivering Data for Presentation

- **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.
- when the data warehouse receives notification that an existing row in a dimension has changed, there are three basic SCD responses —
- Type 1 overwrite, type 2 add a new row, and type 3 add a new column.