E T L

Introduction

Cyrus Lentin

E T L

- Is Short For **Extract Transform Load**,
- Functions (Activities) That Are Combined Into One Tool
- Used To Pull Data Out Of One Source And Place It Into Another Source

- Extract
  Is The Process Of Reading Data From A Source
- Transform
  Is The Process Of Converting The Extracted Data From Its Previous Form Into The Form It Needs To Be In The Target Source.
- Load
  Is The Process Of Writing The Data Into The Target Database

ETL Is Used To Migrate Data From One Database To Another, To Form Data Marts And Data Warehouses And Also To Convert Databases From One Format Or Type To Another.

ETL Process

Extract

- Data From Different Source (SAP, ERP, Other Operational Systems)
- Is Converted Into One Consolidated Format
- Which Is Ready For Transformation Processing

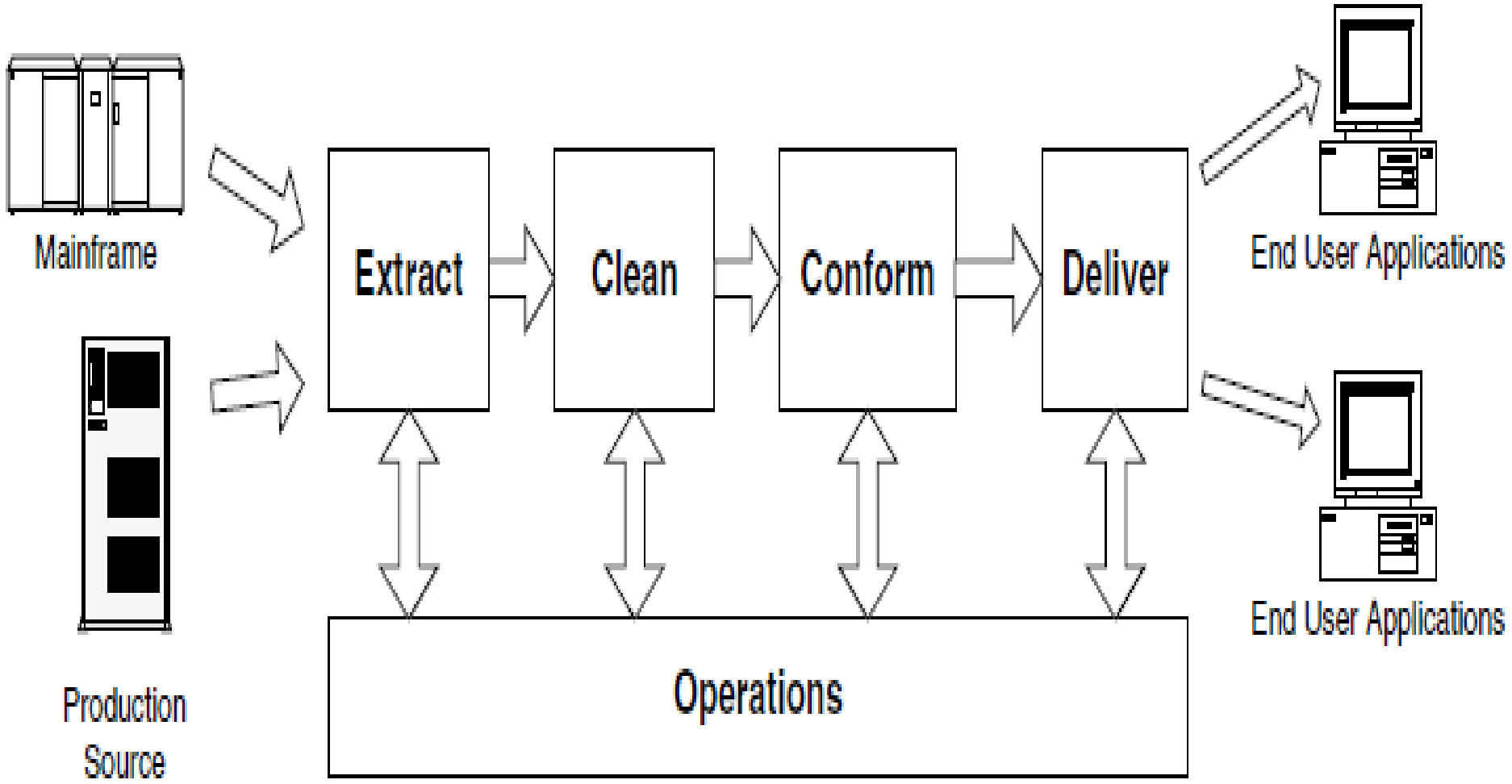
Transform

- Applying Business Rules (So-called Derivations, E.G., Calculating New Measures And Dimensions)
- Cleaning (E.G., Mapping NULL To 0 Or "Male" To "M" And "Female" To "F" Etc.)
- Filtering (E.G., Selecting Only Certain Columns To Load)
- Splitting A Column Into Multiple Columns And/Or Merging Multiple Columns To One
- Joining Together Data From Multiple Sources (E.G., Lookup, Merge)
- Transposing Rows And Columns
- Applying Any Kind Of Simple Or Complex Data Validation

Load

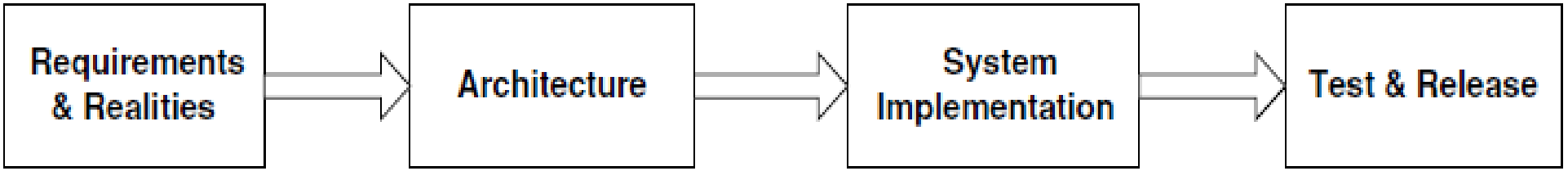
- Loading The Data Into A Data Warehouse Or Data Repository
- Making Data Ready For Reporting Applications

Data Flow Thread



Planning & Design Thread

- Requirements / Realities
- Architecture
- Implementation
- Test / Release



Planning & Design – Requirements

- Business Needs
- Compliance
- Data Profiling
- Security
- Data Integration
- Data Latency
- End User Delivery Interfaces
- Available Skills
- Legacy Licenses

Planning & Design – Architecture

- Hand-coded Versus ETL Vendor Tool
- Batch Versus Streaming Data Flow
- Horizontal Versus Vertical Task Dependency
- Scheduler Automation
- Exception Handling
- Quality Handling
- Recovery And Restart
- Metadata
- Security

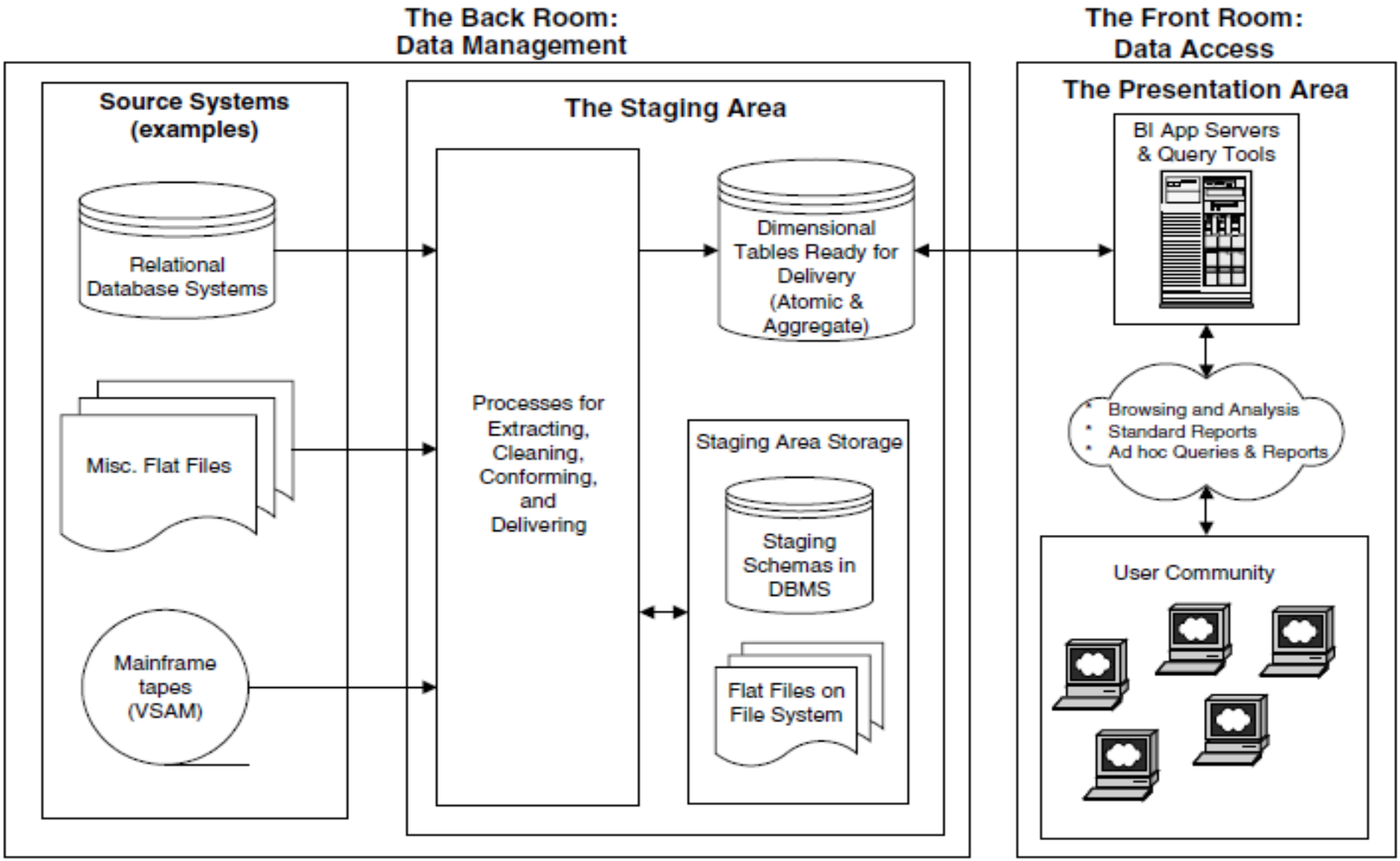
Planning & Design – System Implementation

- Hardware
- Software
- Coding practices
- Documentation Practices
- Quality Checks

Planning & Design – Test & Release

- Development Systems
- Test Systems
- Production Systems
- Handoff Procedures
- Update Propagation Approach
- System Snapshoting And Rollback Procedures
- Performance Tuning

Data Management / Data Access



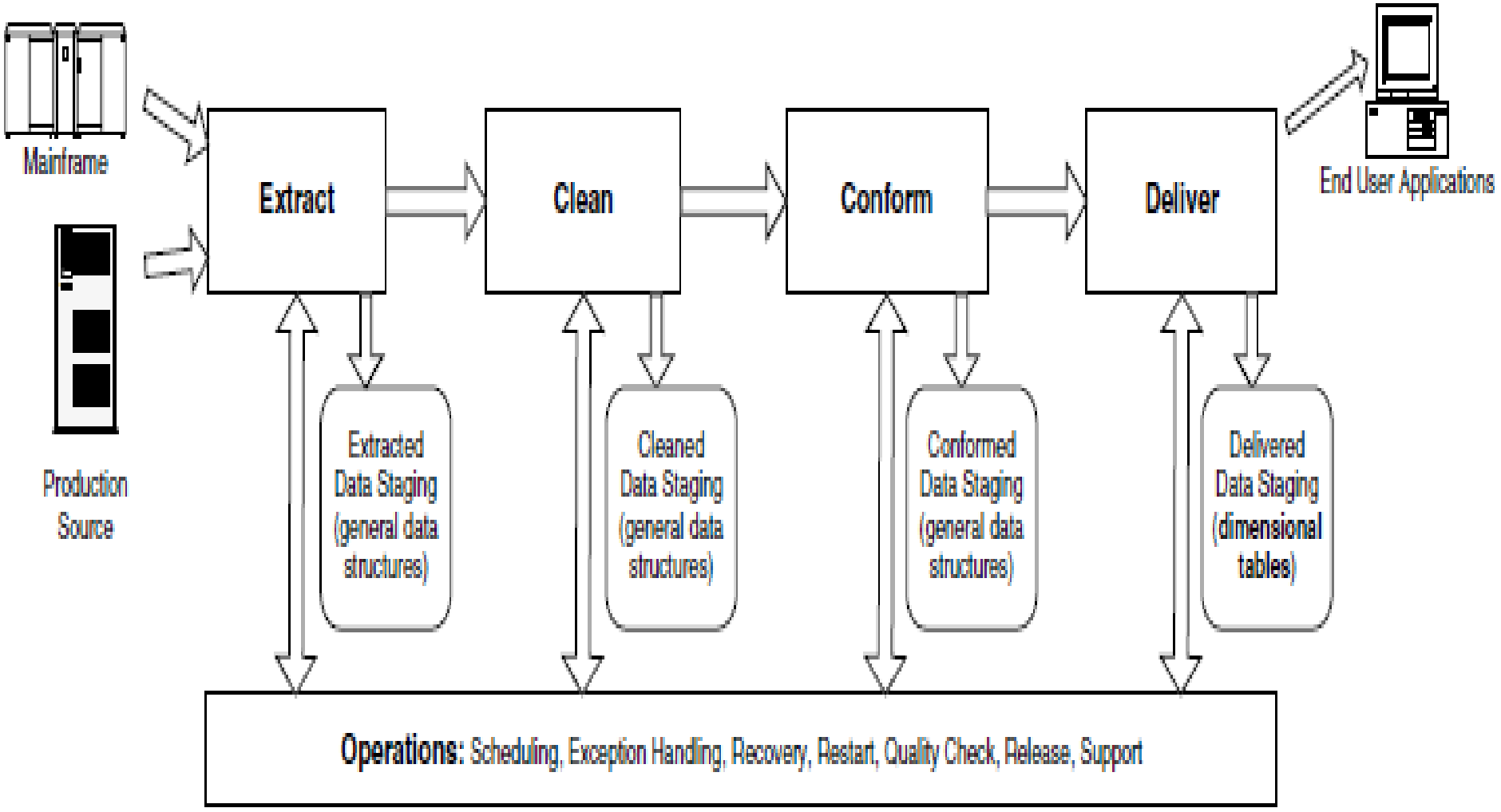
Why Segregation

- Providing Detailed Security At A Row, Column, Or Applications Level
- Building Query Performance-enhancing Indexes And Aggregations
- Providing Continuous Up-time Under Service-level Agreements
- Guaranteeing That All Data Sets Are Consistent With Each Other

Data Management / Staging (Back Room)

- Staging describes discrete steps in the back room.
- Staging almost always implies a temporary or permanent physical snapshot of data.
- Four staging steps found in almost every data warehouse,
  - Extract
  - Clean
  - Conform
  - Deliver

Staging – Block Diagram



Extract

- The Raw Data Coming From The Source Systems Is Usually Written Directly To Disk. Some Minimal Restructuring Happens Before Significant Content Transformation Takes Place
- Data From Structured Source Systems (IMS Databases, Or XML Data Sets) Often Is Written To Flat Files Or Relational Tables In This Step
- This Allows The Original Extract To Be As Simple And As Fast As Possible And Allows Greater Flexibility To Restart The Extract If There Is An Interruption
- Initially Captured Data Can Then Be Read Multiple Times As Necessary To Support The Succeeding Steps
- In Some Cases, Initially Captured Data Is Discarded After The Cleaning Step Is Completed, And In Other Cases Data Is Kept As A Long-term Archival Backup
- The Initially Captured Data May Also Be Saved For At Least One Capture Cycle So That The Differences Between Successive Extracts Can Be Computed

Clean

- In Most Cases, The Level Of Data Quality Acceptable For The Source Systems Is Different From The Quality Required By The Data Warehouse.
- Data Quality Processing May Involve Many Discrete Steps, Including
  - Checking For Valid Values Ensuring Data Consistency Across Values
  - Removing Duplicates
  - Checking Whether Complex Business Rules And Procedures Have Been Enforced
- The Results Of The Data-cleaning Step Are Often Saved Semi-Permanently Because The Transformations Required Are Difficult And Irreversible.
- Is The Cleaned Data Fed Back To The Sources Systems To Improve Their Data And Reduce The Need To Process The Same Data Problems Again
- Even If The Cleaned Data Cannot Be Physically Fed Back To The Source Systems, The Data Exceptions Should Be Reported To Build A Case For Improvements

Conform

- Data Conformation Is Required Whenever Two Or More Data Sources Are Merged In The Data Warehouse.
- Separate Data Sources Cannot Be Queried Together Unless Some Or All Of The Textual Labels In These Sources Have Been Made Identical And Unless Similar Numeric Measures Have Been Mathematically Rationalized So That Differences And Ratios Between These Measures Make Sense.
- Data Conformation Is A Significant Step That Is More Than Simple Data Cleaning.
- Data Conformation Requires An Enterprise-wide Agreement To Use Standardized Domains And Measures.

Deliver

- The Whole Point Of The Back Room Is To Make The Data Ready For Querying.
- The Final And Crucial Back-room Step Is Physically Structuring The Data Into A Set Of Simple, Symmetric Schemas Known As Dimensional Models, Or Equivalently, Star Schemas.
- These Schemas Significantly Reduce Query Times And Simplify Application Development.
- Dimensional Schemas Are Required By Many Query Tools, And These Schemas Are A Necessary Basis For Constructing OLAP Machines

Stage Or Not To Stage

- ETL Architect's Decision To Store Data In A Physical Staging Area Versus Processing It In Memory.
- Determine The Right Balance Between Physical Input And Output (I/O) And In-memory Processing.
- Determining Whether To Stage Your Data Or Not Depends On Two Conflicting Objectives:
  - Getting The Data From The Originating Source To The Ultimate Target As Fast As Possible
  - Having The Ability To Recover From Failure Without Restarting From The Beginning Of The Process
- The Decision To Stage Data Varies Depending On Your Environment And Business Requirements.
- Following Reasons Must Be Considered For Staging Data Before It Is Loaded Into The Data Warehouse:
  - Recoverability
  - Backup
  - Auditing
- Staging Area Types:
  - Persistent – History To Staging Files Are Maintained
  - Transient – Staging Files Are Deleted After Delivery

Stage Rules

- The Data-staging Area Must Be Owned By The ETL Team.
- Users Are Not Allowed In The Staging Area For Any Reason.
- Reports Cannot Access Data From The Staging Area.
- Only Etl Processes Can Write To And Read From The Staging Area.

ETL Team Owns The Data-staging Area.

- ETL Architect Designs The Tables & Decides Whether A Table Belongs In The Database Or Flatfile
- The ETL Architect Must Supply Processing & Data Storage Requirement to OS-Admin / DBA
- OS-Admin / DBA creates the required database / allocates required space and hands over to ETL Team

Data Access (Front Room)

- Indexing dimensional tables in the presentation area for query performance
- Choosing front-end tools, including query tools, report writers, and dashboards
- Writing SQL to solve end user queries
- Data-mining techniques
- Forecasting, behavior scoring, and calculating allocations
- Security on the tables and applications accessible by end users
- Metadata supporting end user tools
- End user training and documentation

Data Warehouse

- The aim of the data warehouse is to publish the organization's data assets to most effectively support decision making.
- The key word in this mission statement is publish. The success of a data warehouse begins and ends with its end users.
- Since the data warehouse is a decision support system, our main criterion of success is whether the data warehouse effectively contributes to the most important decision-making processes in the organization.

A data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making.

A Data Warehouse Is NOT
Product | Language | Project | Data Model | Transaction System

ETL Mission

All ETL System Must:

- Deliver Data Most Effectively To End User Tools
- Add Value To Data In The Cleaning And Conforming Steps
- Protect And Document The Lineage Of Data

Four Keys Steps:

- Extracting Data From The Original Sources
- Quality Assuring And Cleaning Data
- Conforming The Labels And Measures In The Data To Achieve Consistency Across The Original Sources
- Delivering Data In A Physical Format That Can Be Used By Query Tools, Report Writers, And Dashboards.

Thank you!

Contact:

Cyrus Lentin
cyrus@lentins.co.in
+91-98200-94236