ETL

Day-4 – Cleaning & Conforming

Cyrus Lentin

Cleaning & Conforming

- Cleaning And Conforming Are The Main Steps Where The ETL System Adds Value
- The Other Steps Of Extracting And Delivering Only Move And Reformat Data
- Cleaning And Conforming Actually Changes Data
- This Provides Guidance Whether Data Can Be Used For Its Intended Purposes
- Cleaning & Conforming Process:
  - Design Objectives
  - Cleaning Deliverables
  - Checks and Their Measurements
  - Conforming Deliverables
- Three Deliverables:
  - Data Profiling Report
  - Error Report
  - Audit Report
- A Powerful Cleaning And Conforming System Is Built Around These Three Tangible Deliverables

Design Objectives

- Understand Stake-Holders
  - Data Warehouse Manager
  - Business Owner
  - Technical Owner
  - Dimension Manager (Master Record Department)
  - Fact Table Provider (Transaction Record Department)
- Design Objectives
  - Thorough
  - Fast
  - Corrective
  - Transparent
- Balancing Conflicting Priorities
  - Completeness v/s Speed
  - Corrective v/s Transparent
  - Data Quality

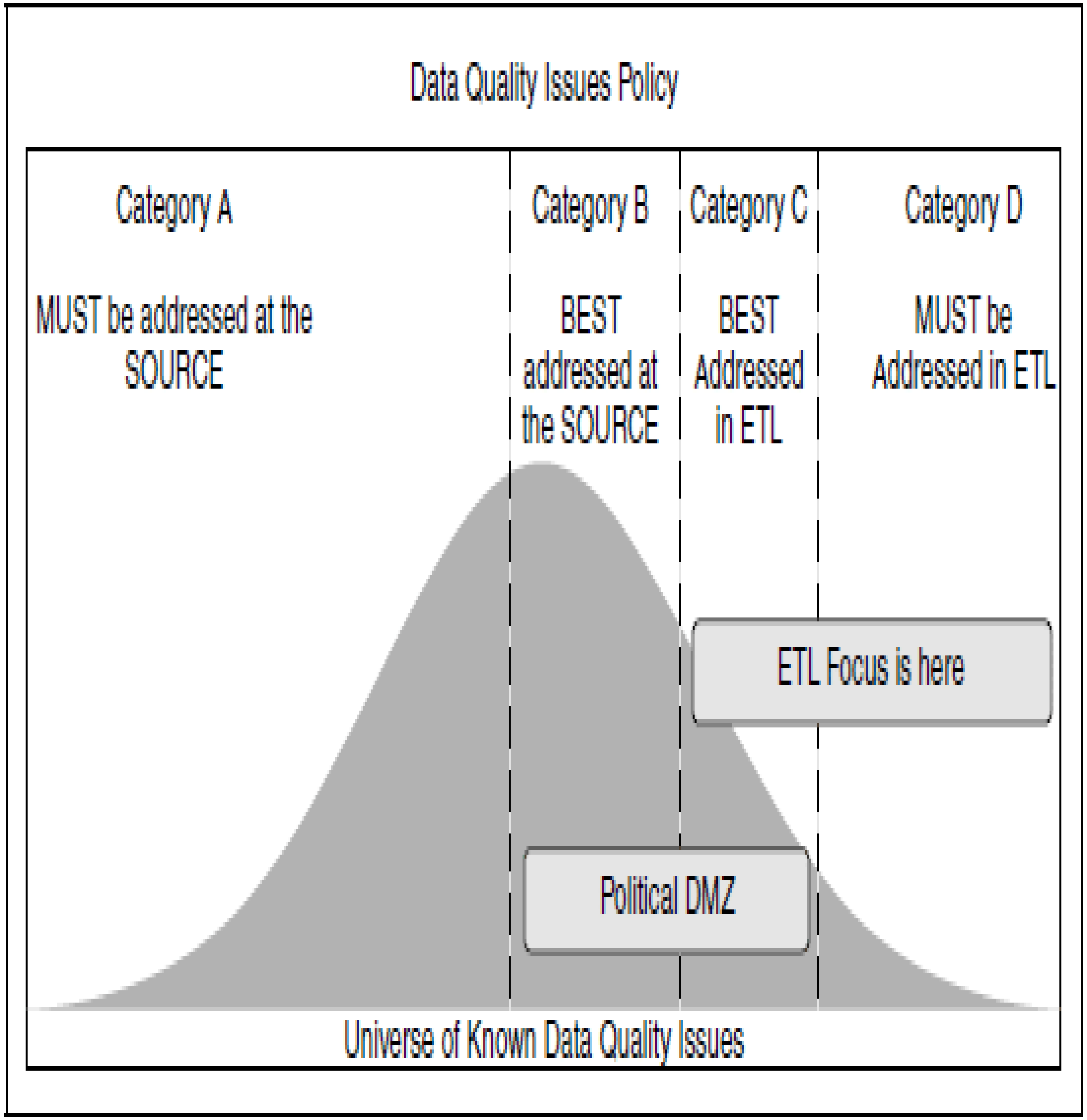
Defining Data Quality

Data Must Be Accurate

- Correct
- Unambiguous
- Consistent
- Complete
  - All Fields Defined
  - All Records Present

Data Quality Policy

- Category A
  Missing Info Of Customer / Vendor
- Category B
  Missing Info Of Customer / Vendor
- Category C
  Missing Or Incomplete Information From Independent Third-party Data Suppliers
- Category D
  Missing Or Incomplete Information From Independent Third-party Data Suppliers


Data Quality Issues Policy
Category A
MUST be addressed at the SOURCE
Category B
BEST addressed at the SOURCE
Category C
BEST Addressed in ETL
Category D
MUST be Addressed in ETL
ETL Focus is here
Political DMZ
Universe of Known Data Quality Issues

Cleaning Deliverables

Cleaning Deliverables Sub-system Should Offer The Following Data-quality Insights:

- Is Data Quality Getting Better Or Worse?
- Which Source Systems Generate The Most/Least Data-quality Issues?
- Are There Interesting Patterns Or Trends Revealed In Scrutinizing The Data-quality Issues Over Time?
- Is There Any Correlation Observable Between Data-quality Levels And The Performance Of The Organization As A Whole?

Also Should Be Able To Answer:

- Which Of My Data-Quality Checks Consume The Most/Least Time In My ETL Window?
- Are There Data Quality Checks That Can Be Retired Because The Types Of Issues That They Uncover No Longer Appear In Our Data?

Deliverables

- Data Profiling Report
- Error Report
- Audit Report

Checks and Their Measurements

- Detailed Design Stage
- Contains A Set Of Fundamental Checks And Tests At The Core Of Most Data-cleaning Engines
- It Describes What These Functions Do, How They Do It
- It Describes How They Build Upon One Another To Deliver Cleaned Data

Anomaly Detection

- A Data Anomaly Is A Piece Of Data That Does Not Pass The Data Quality Test
- Finding Data Anomalies May Be Perceived By Some Outside The ETL Scope
- Detecting Data Anomalies Will Be The Responsibility Of The ETL Team

Anomaly Detection

- All Records
- Data Sampling

Checks and Their Measurements

Types Of Checks

- Column Property Checks
- Structure Checks
- Data Checks
- Value Checks

Based On The Findings Of These Checks, The ETL Job Stream Can Choose To:

- Pass The Record With No Errors
- Pass The Record, Flagging Offending Column Values
- Reject The Record
- Stop The ETL Job Stream

Conforming Deliverables

Incoming Data Needs To

- Be Made Structurally Identical
- Filtered Of Invalid Records
- Standardized In Terms Of Its Content
- De-duplicated
- Converted Into The New Conformed Image

Process For Building Conformed Data

- Standardizing
- Deduplication
- Surviving

Standardizing

- Descriptive Attributes Vary Across Multiple Data Sources
- These Are Not Errors But Variations Of Good Data
- Standardizing Is Capturing & Correcting These Variations
- The Corrections Should Be Based On The Requirements Of The Target System

Data Validation & Correction As Require By The Target System

Matching & Deduplication

- Matching, or deduplication, involves the elimination of duplicate standardized records
- Duplicate can be easily detected through the appearance of identical values in some key column—like social security number, telephone number, or charge card number
- In other cases, no such definitive match is found, and the only clues available for deduplication are the similarity of several columns that almost match
- Specialized data integration matching tools are now mature and in widespread use and deal with these very specialized data-cleansing issues
- The matching software must compare the set of records in the data stream to the universe of conformed dimension records and return:
  - A numeric score that quantifies the likelihood of a match
  - A set of match keys that link the input records to conformed dimension instances
- Organizations with a need for very robust deduplication capabilities can choose also to maintain a persistent library of previously matched data & use this consolidated library to improve their results

Surviving

- Survivorship Refers To The Process Of Creating / Filtering A Set Of Standardized & Deduplicated Records
- Surviving Data Is Filtered Into A Separate Table That Combines The Column Values From Each Of The Records To Build Conformed Target Records For Fact Or Dimension Tables
- This Entails Establishing Business Rules That Must Applied When Writing Out For Survived Records
  - Source-To-Target Mapping (eg Validation Of State Names, Categories, Class, Etc)
  - Survivorship Block Of Records (for master-detail type of transaction, ensure record(s) present)

Data Profiling Report

Good Data Profiling Analysis Takes The Form Of A Specific Metadata Repository Describing:

- Schema Definitions
- Business Objects
- Domains
- Data Sources
- Table Definitions
- Synonyms
- Data Rules
- Value Rules
- Issues That Need To Be Addressed

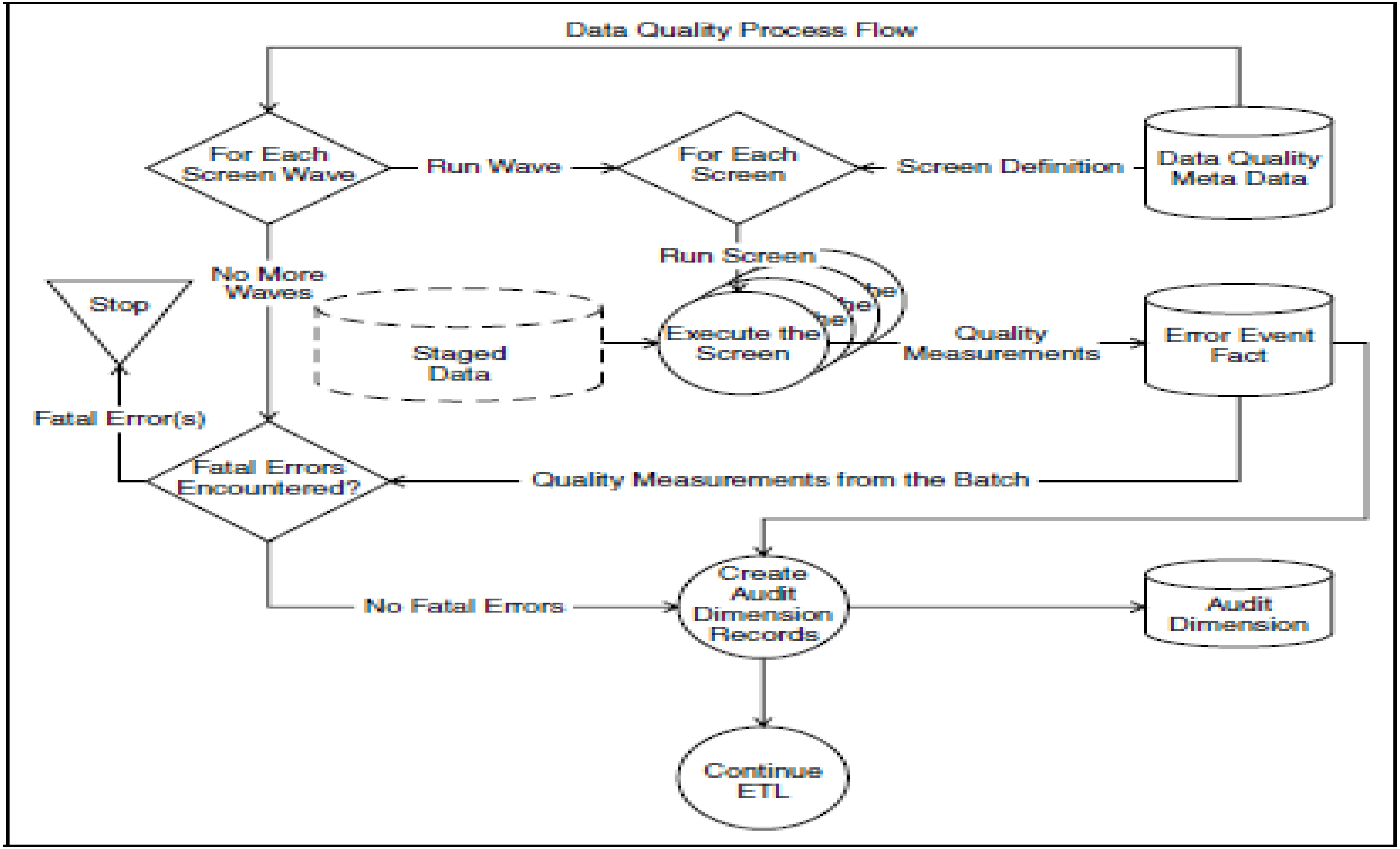
Error Report

- Each Data-quality Error Or Issue Surfaced By The Data-cleaning Subsystem Is Captured As A Row In The Error Report
- The Attributes Of The Error Report Are As Follows:
  - Error Date / Time
  - ETL Stage
  - Processing Order Number
  - Severity Score
  - Exception Action
  - Error
  - SQL Statement

Audit Report

- To Associate Data-quality Indicators With The Final Tables As Per The Target System, We Need To Build A Table Wise Audit Report.
- The Audit Report Is Prepared For Each Table In The Target System
- The Audit Report Captures Important ETL Processing Milestone Like
  - Timestamps
  - Outcomes
  - Significant Errors
  - Correction
  - Frequency Of Error Occurrence
  - Overall Data-quality Score
- Audit Reports Are Created As The Final Step Of The Processing For Cleaned And Conformed Tables And Must Contain A Description Of The Fixes And Changes That Have Been Applied
- Audit Report Must Be Prepared For Each ETL Job
- Audit Report Must Be Circulated To Business Owners & Technical Owners Of Source & Target System

Data Quality Process Flow



Thank you!

Contact:

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