Data Engineer

Data Cleansing, DataVault, Hash

Link

https://www.tableau.com/learn/articles/what-is-data-cleaning

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*=“42”, *Birthday*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction ⬅
- Data Transformation and Data Discretization
- Summary

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

Data Preprocessing

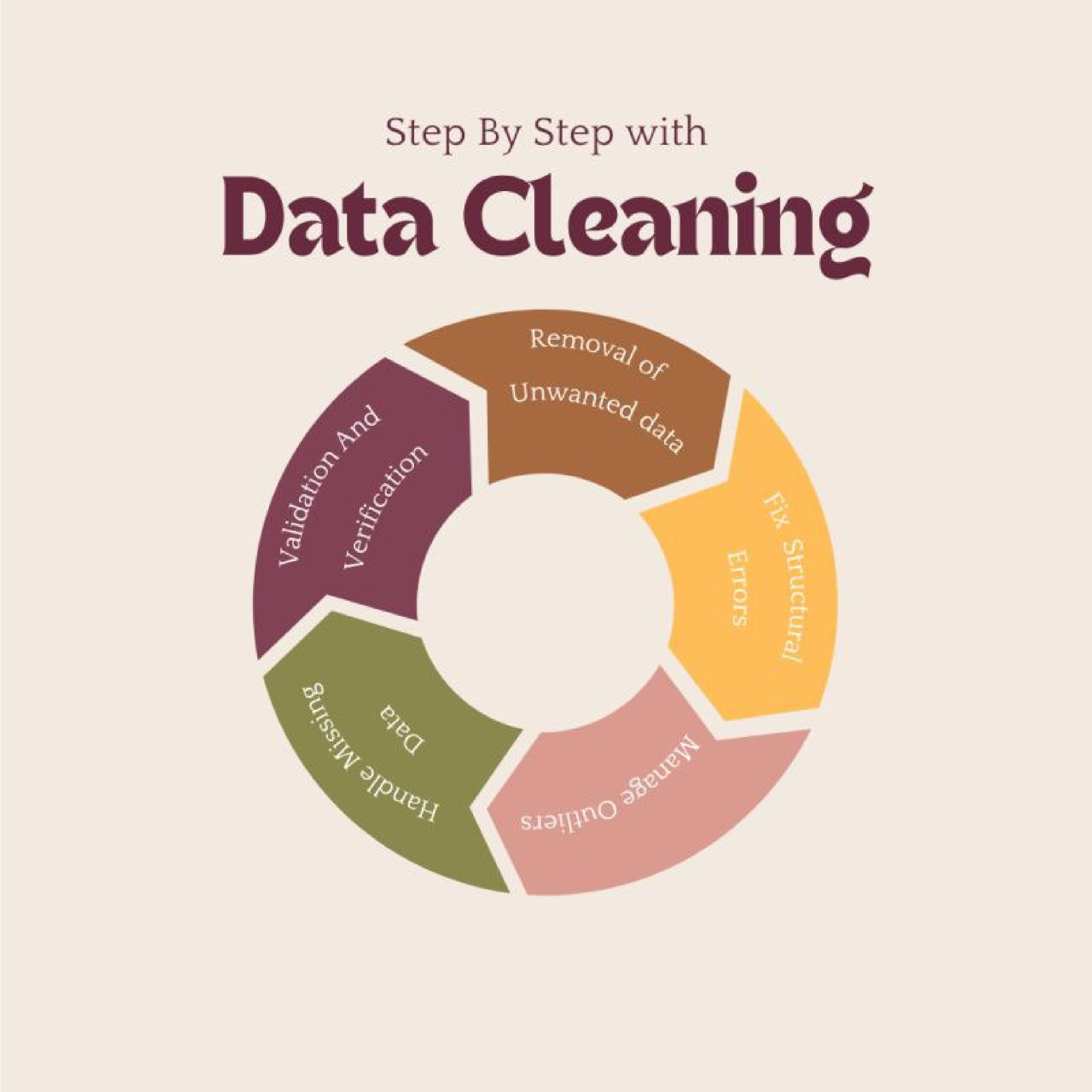
- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary



Step By Step with

Data Cleaning

Removal of Unwanted data

Fix Structural Errors

Manage Outliers

Handle Missing Data

Validation And Verification

Data Cleaning Isn’t Glamorous, but It’s important

No fancy dashboards.

No colorful charts.

Just rows, columns... and hidden chaos.

90% of the time, when a report “looks off”, it’s not the formula but messy data underneath.

In one client project, the actual game-changer wasn’t visuals.

It was fixing:

- Dates in weird formats
- Duplicates messing up totals
- Typos causing broken relationships
- Irrelevant rows bloating file sizes

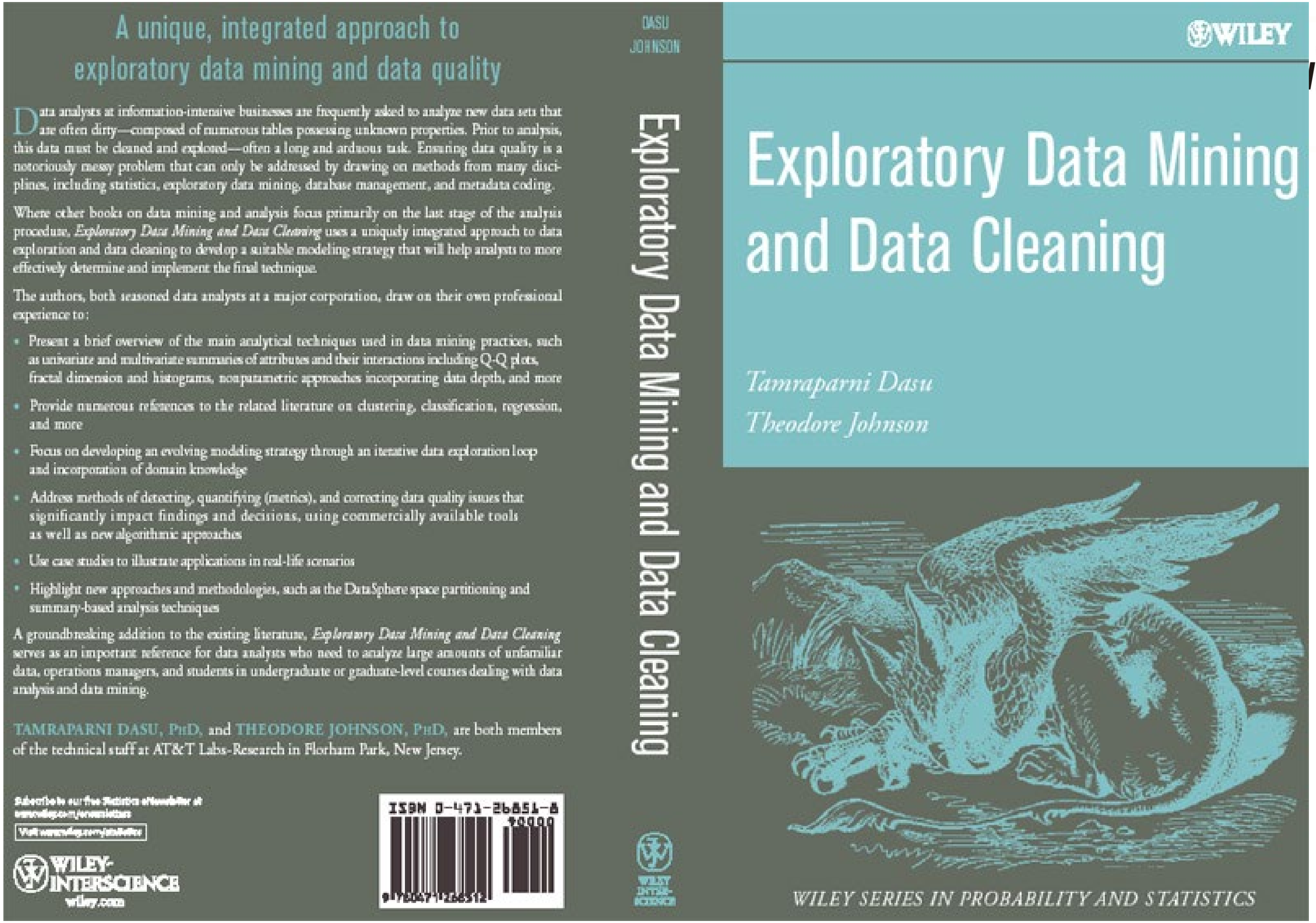
After cleanup, this is what happened:
✨ Reports refreshed 3x faster
✨ Accuracy shot up
✨ Teams made decisions with confidence, not confusion

you can’t build a castle on a shaky foundation.

In analytics, clean data = solid ground.
Might not be glamorous.
But it pays. Every single time.

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation



A unique, integrated approach to exploratory data mining and data quality

Data analysts at information-intensive businesses are frequently asked to analyze new data sets that are often dirty—composed of numerous tables possessing unknown properties. Prior to analysis, this data must be cleaned and explored—often a long and arduous task. Ensuring data quality is a notoriously messy problem that can only be addressed by drawing on methods from many disciplines, including statistics, exploratory data mining, database management, and metadata coding.

Where other books on data mining and analysis focus primarily on the last stage of the analysis procedure, *Exploratory Data Mining and Data Cleaning* uses a uniquely integrated approach to data exploration and data cleaning to develop a suitable modeling strategy that will help analysts to more effectively determine and implement the final technique.

The authors, both seasoned data analysts at a major corporation, draw on their own professional experience to:

- Present a brief overview of the main analytical techniques used in data mining practices, such as univariate and multivariate summaries of attributes and their interactions including Q-Q plots, fractal dimension and histograms, nonparametric approaches incorporating data depth, and more
- Provide numerous references to the related literature on clustering, classification, regression, and more
- Focus on developing an evolving modeling strategy through an iterative data exploration loop and incorporation of domain knowledge
- Address methods of detecting, quantifying (metrics), and correcting data quality issues that significantly impact findings and decisions, using commercially available tools as well as new algorithmic approaches
- Use case studies to illustrate applications in real-life scenarios
- Highlight new approaches and methodologies, such as the DataSphere space partitioning and summary-based analysis techniques

A groundbreaking addition to the existing literature, *Exploratory Data Mining and Data Cleaning* serves as an important reference for data analysts who need to analyze large amounts of unfamiliar data, operations managers, and students in undergraduate or graduate-level courses dealing with data analysis and data mining.

TAMRAPARNI DASU, PhD, and THEODORE JOHNSON, PhD, are both members of the technical staff at AT&T Labs-Research in Florham Park, New Jersey.

WILEY-
INTERSCIENCE
wiley.com

ISBN 0-471-26851-8
90000
9 780471 268512

DASU
JOHNSON

Exploratory Data Mining and Data Cleaning

WILEY
INTER-
SCIENCE

WILEY

Exploratory Data Mining and Data Cleaning

Tamraparni Dasu
Theodore Johnson

WILEY SERIES IN PROBABILITY AND STATISTICS

Data Vault

- Brief History and Revisit Some Definitions
- Three Basic Building Blocks of the Data Vault
- Advanced Features
- Questions

Data Vault –
Brief History and Revisit Some Definitions

- 1970 – Dr. E.F. Codd of IBM
- 1979 – First Working Relational Database by Relational Software Incorporated
    - Oracle v2
- 1991 – William H. Inmon published ‘Building the Data Warehouse’

Data Vault –
Brief History and Revisit Some Definitions

- Legacy System –
  - '… any system that has been put into production.'

    (para-phrased W.H. Inmon)
- Operational Data Store –
  - '… a subject-oriented, integrated, volatile, current or near current collection of operational data.'

    W.H. Inmon

Data Vault –
Brief History and Revisit Some Definitions

- Data Warehouse –
  - ‘… a subject-oriented, integrated, time-variant, non-volatile collection of data designed for support of business decisions’

    W.H. Inmon
- Data Vault –
  - ‘… a detail-oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business.’

    Dan Linstedt

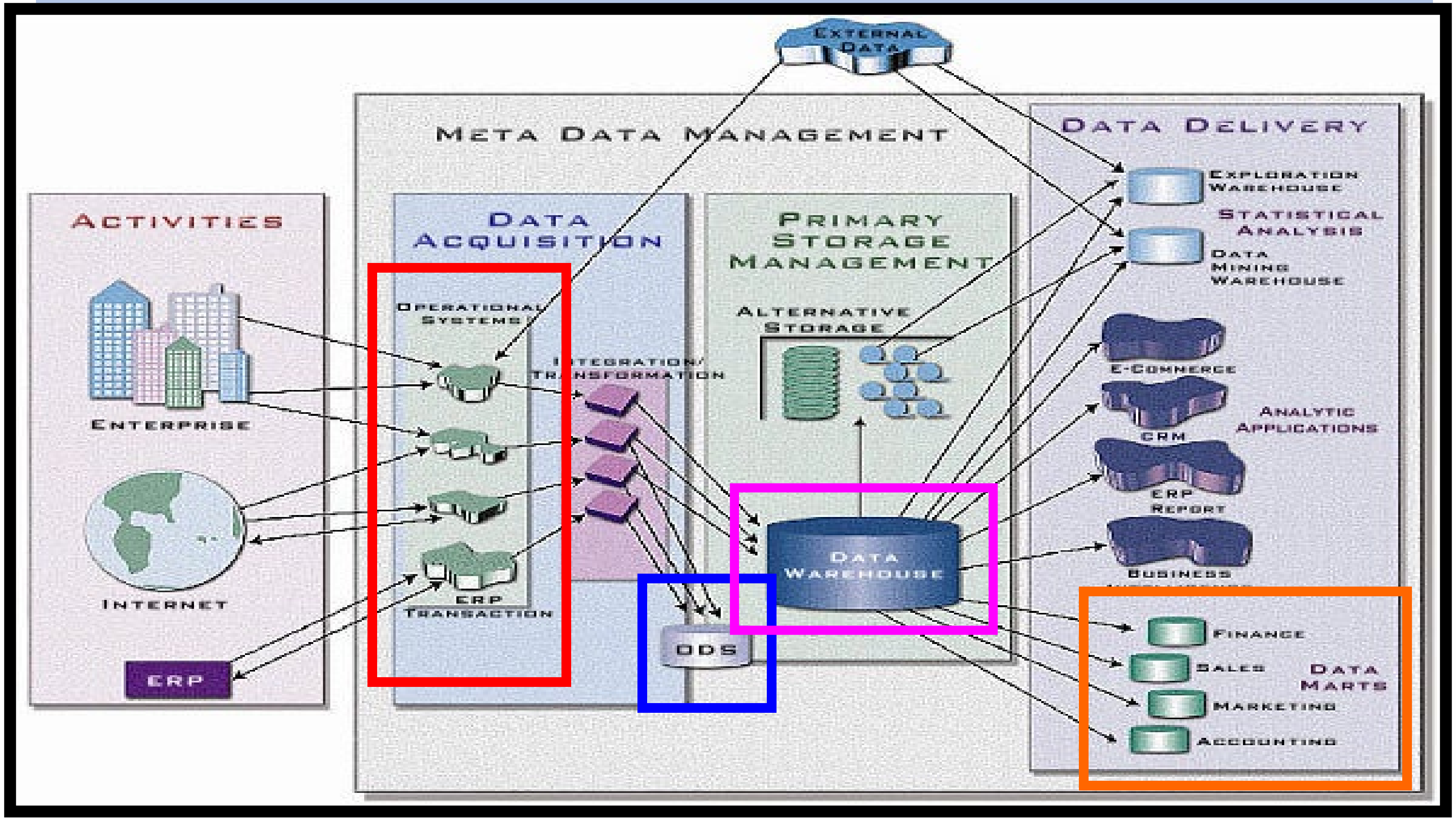
Data Vault –
Brief History and Revisit Some Definitions

- Data Mart –
  - '… a subset of a data warehouse, for use by a single department or function.'

    www.e-formation.co.nz/glossary.asp
- Corporate Information Factory –
  - '… the framework that exists that surrounds the data warehouse; typically contains an ODS, a data warehouse, data marts, DSS applications, exploration warehouses, and so forth.'

    W.H. Inmon

Data Vault –
Brief History and Revisit Some Definitions



* Source: Bill Inmon and Claudia Imhoff

Data Vault – Why?

- Why do we need it?
  - We finally have a Data Model that will work for small, medium, or large business
    - Anyone building a Data Warehouse can use these techniques.
  - We’ve got issues in constructing the data warehouse from 3rd normal form, or star schema form.
    - There are inherent road blocks to each method that we must solve technically through our Data Model.

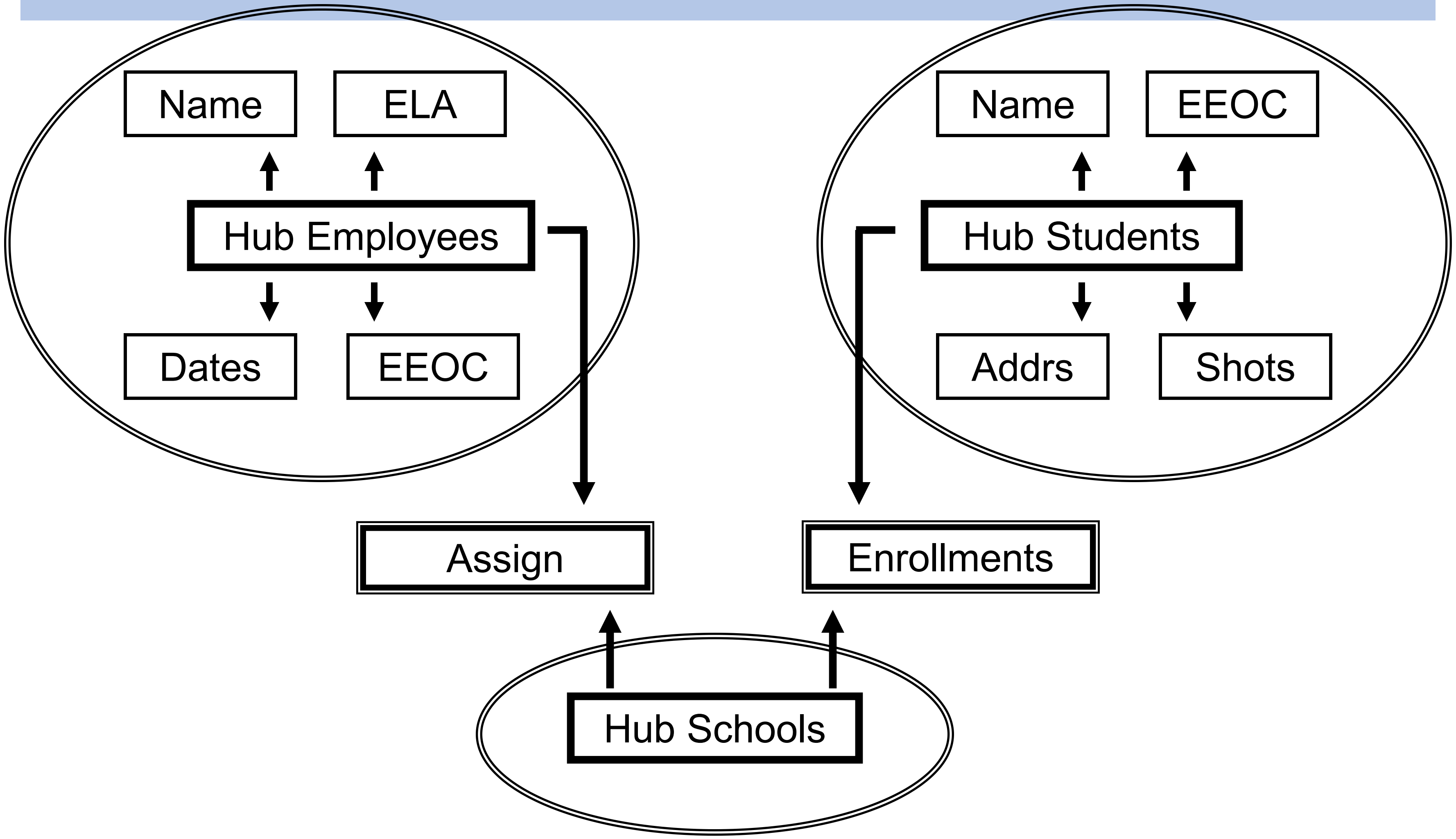
Data Vault

- Brief History and Revisit Some Definitions
- Three Basic Building Blocks of the Data Vault
- Advanced Features
- Questions

Data Vault –
Three Basic Building Blocks

- Hub – stand alone table; list of unique business keys; used for business identification
- Satellite – descriptive data; historical data; used for descriptive information for the HUB or LINK
- Link – associative table; list of unique relationships between keys; used for relationships between HUBs and LINKs

Data Vault – Three Basic Building Blocks
Preview



Data Vault – Three Basic Building Blocks
HUB

A Hub is a list of unique business keys.

Primary Key

<Business Key>

Load DTS

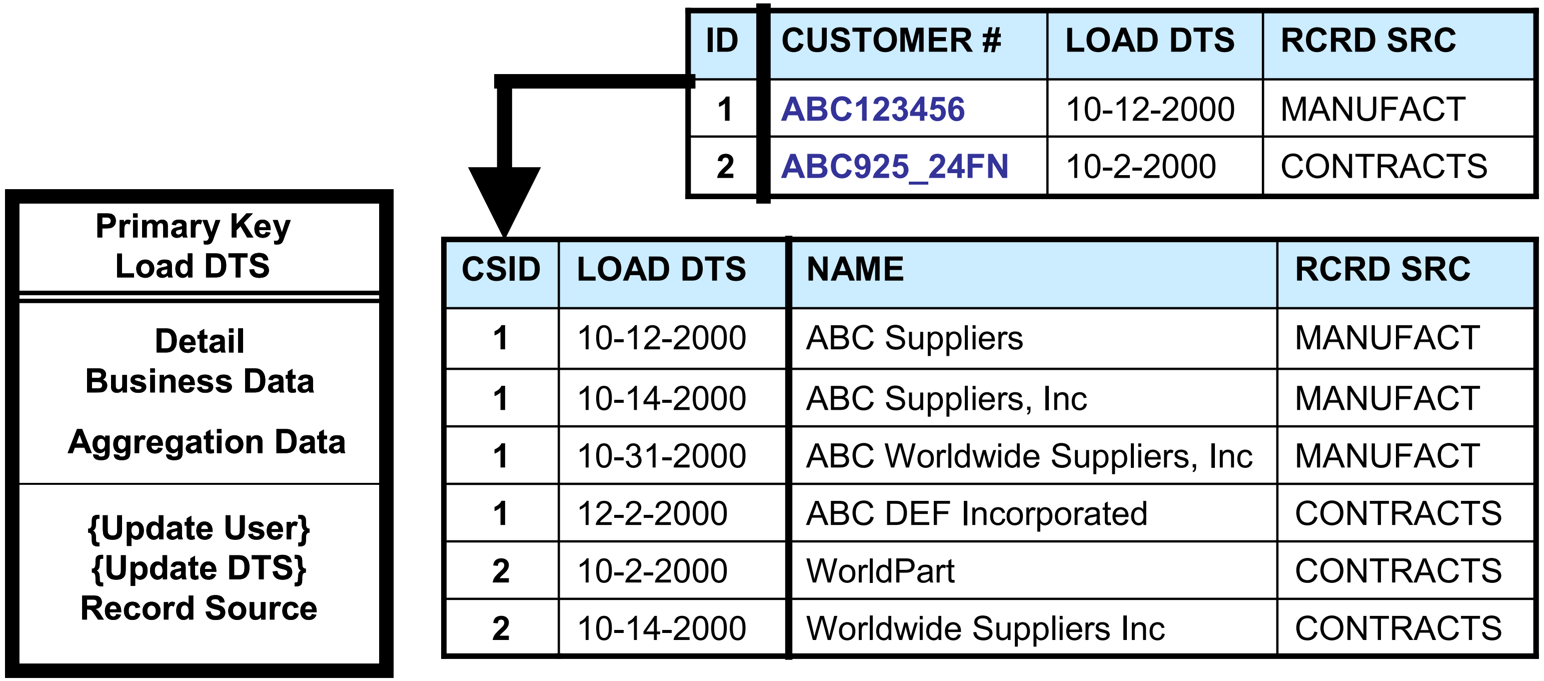
Record Source

Sample Data Set “CUSTOMER”

ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT
2	ABC925_24FN	10-2-2000	CONTRACTS
3	DKEF	1-25-2000	CONTRACTS
4	KKO92854_dd	3-7-2000	CONTRACTS
5	LLOA_82J5J	6-4-2001	SALES
6	HUJI_BFIOQ	8-3-2001	SALES
7	PPRU_3259	2-2-2000	FINANCE
8	PAFJG2895	2-2-2000	CONTRACTS
9	929ABC2985	2-2-2000	CONTRACTS
10	93KFLLA	2-2-2000	CONTRACTS

Data Vault – Three Basic Building Blocks SATELLITE

A Satellite is a time-dimensional table housing detailed information about the hub's business keys.



Primary Key
Load DTS

Detail
Business Data

Aggregation Data

{Update User}
{Update DTS}
Record Source

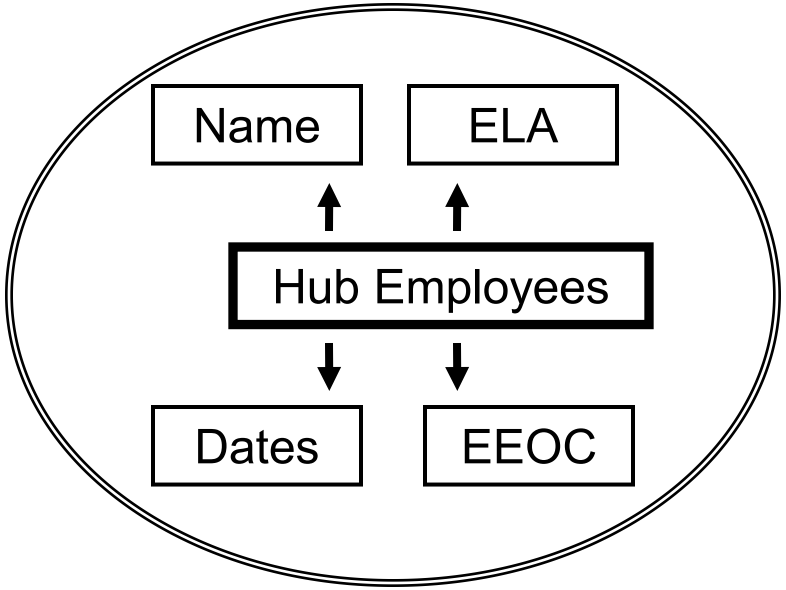
ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT
2	ABC925_24FN	10-2-2000	CONTRACTS

CSID	LOAD DTS	NAME	RCRD SRC
1	10-12-2000	ABC Suppliers	MANUFACT
1	10-14-2000	ABC Suppliers, Inc	MANUFACT
1	10-31-2000	ABC Worldwide Suppliers, Inc	MANUFACT
1	12-2-2000	ABC DEF Incorporated	CONTRACTS
2	10-2-2000	WorldPart	CONTRACTS
2	10-14-2000	Worldwide Suppliers Inc	CONTRACTS

CUSTOMER NAME SATELLITE

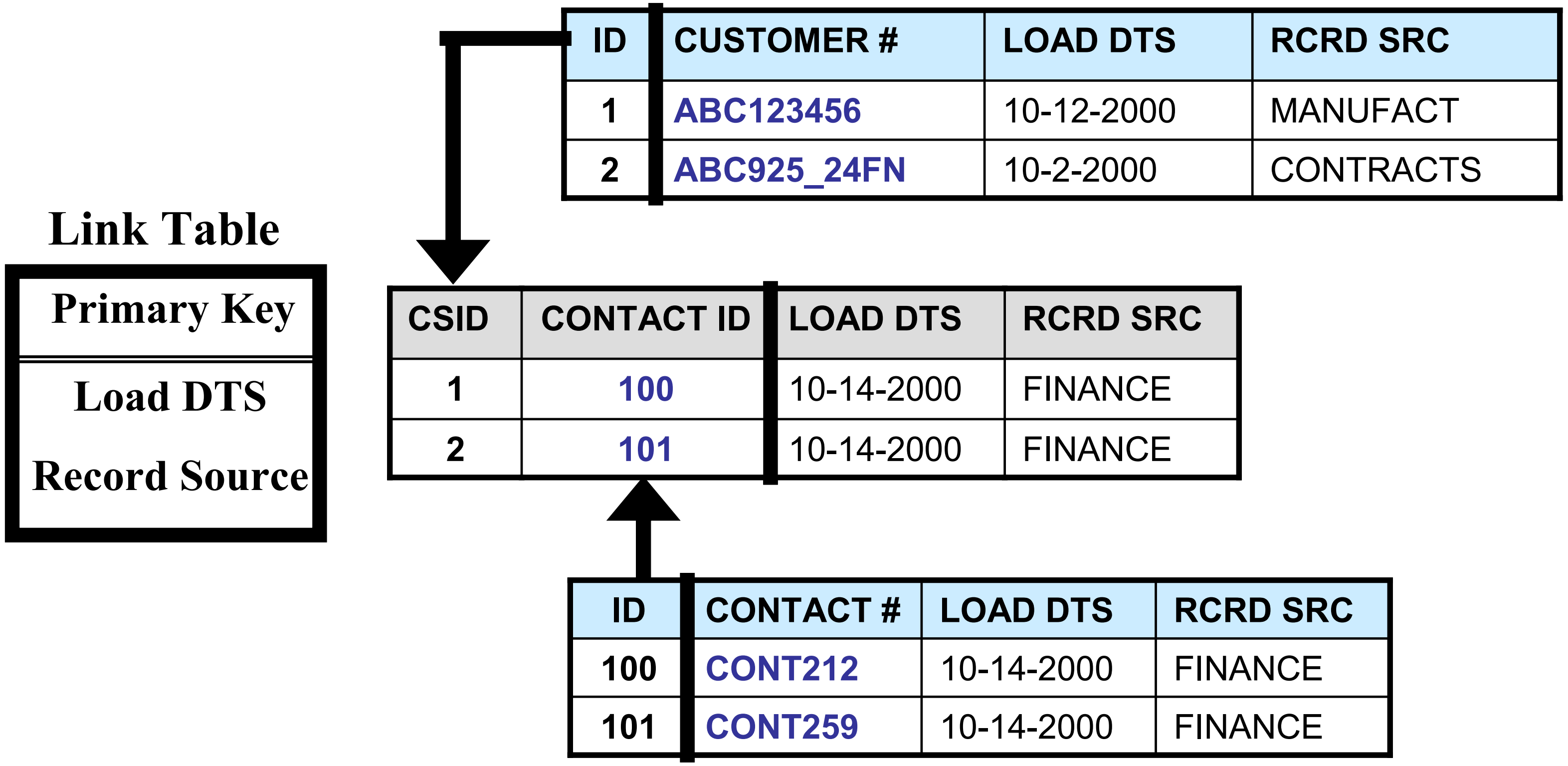
Data Vault – Three Basic Building Blocks

Employees HUB and some of its Satellites



Data Vault – Three Basic Building Blocks
LINK

A Link is an associative or intersection table, representing the connection between information between business elements.



Link Table

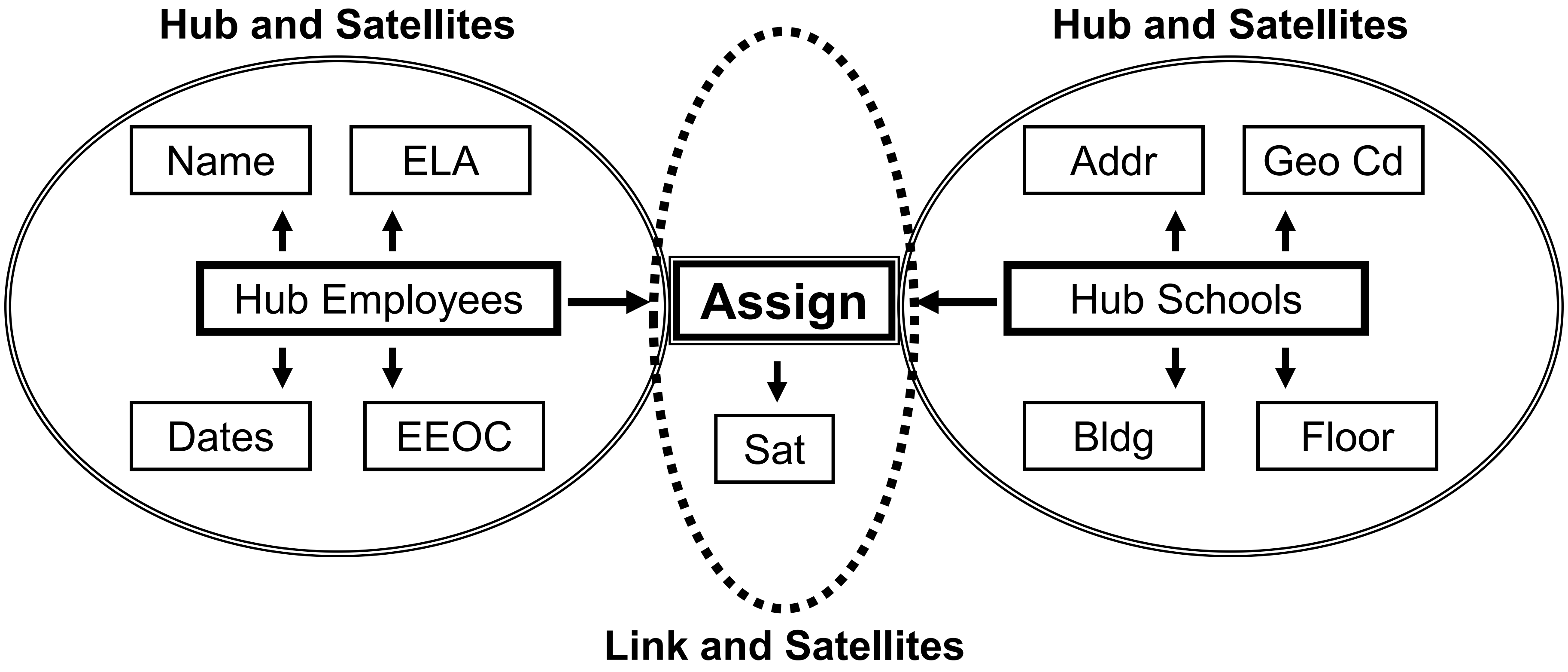
Primary Key
Load DTS Record Source

ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT
2	ABC925_24FN	10-2-2000	CONTRACTS

CSID	CONTACT ID	LOAD DTS	RCRD SRC
1	100	10-14-2000	FINANCE
2	101	10-14-2000	FINANCE

ID	CONTACT #	LOAD DTS	RCRD SRC
100	CONT212	10-14-2000	FINANCE
101	CONT259	10-14-2000	FINANCE

Data Vault – Three Basic Building Blocks



Data Vault

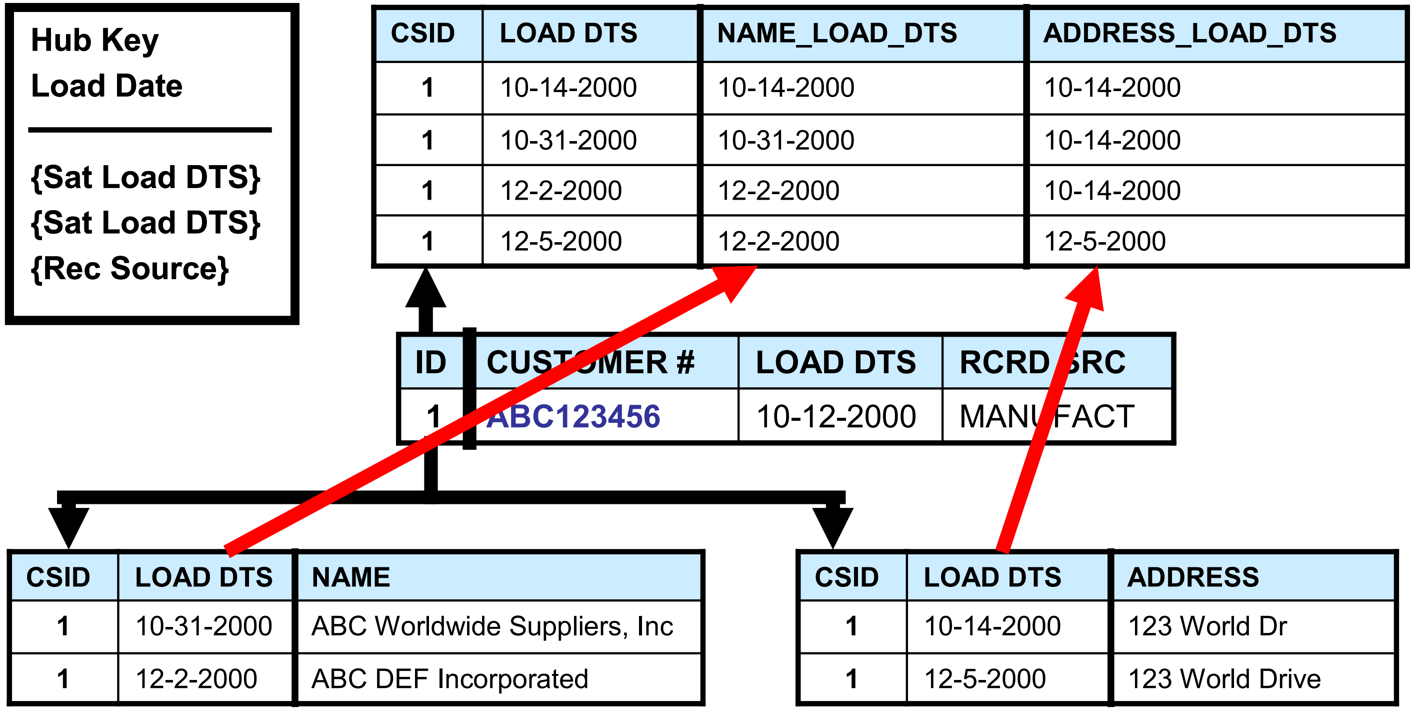
- Brief History and Revisit Some Definitions
- Three Basic Building Blocks of the Data Vault
- Advanced Features
- Questions

Data Vault – Advanced Features

- Point-In-Time –
  - A structure which sustains integrity of joins across time to all the SATELLITES that are connected to the HUB or LINK.
- Bridge –
  - A single row table that contains the latest Load Date Time Stamp (DTS). Similar to Point-In-Time except it spans a subject-area or a schema.
- User Grouping Link –
  - The information provides the user with a customized view from a reporting standpoint and does not affect the underlying information.

Data Vault – Advanced Features
Point-In-Time (PIT)

A structure which sustains integrity of joins across time to all the satellites that are connected to the hub.



Hub Key
Load Date

{Sat Load DTS}
{Sat Load DTS}
{Rec Source}

CSID	LOAD DTS	NAME_LOAD_DTS	ADDRESS_LOAD_DTS
1	10-14-2000	10-14-2000	10-14-2000
1	10-31-2000	10-31-2000	10-14-2000
1	12-2-2000	12-2-2000	10-14-2000
1	12-5-2000	12-2-2000	12-5-2000

ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT

CSID	LOAD DTS	NAME
1	10-31-2000	ABC Worldwide Suppliers, Inc
1	12-2-2000	ABC DEF Incorporated

Customer Name Satellite

CSID	LOAD DTS	ADDRESS
1	10-14-2000	123 World Dr
1	12-5-2000	123 World Drive

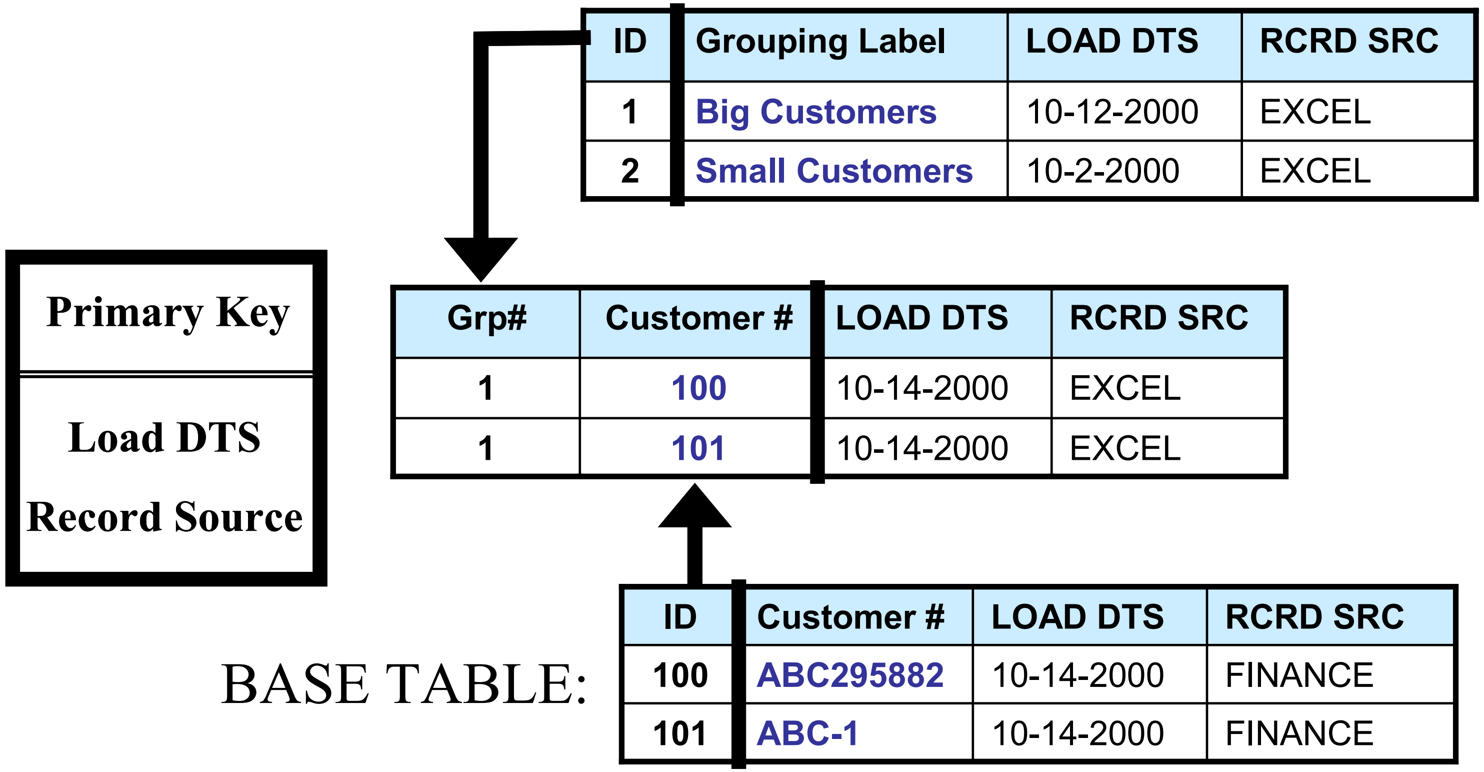
Customer Address Satellite

Data Vault – Advanced Features
Bridge

- A single row table that contains the latest Load DTS with multiple columns. A Bridge is not a helper table.
- Similar to a PIT Table except it spans or applies to a subject-area or schema. A PIT Table is HUB (LINK) and SATELLITE specific.

Data Vault – Advanced Features
User Grouping Link

The User Grouping Link, allows users to "state" how they want roll-ups to occur – in situations where source data doesn't exist.



ID	Grouping Label	LOAD DTS	RCRD SRC
1	Big Customers	10-12-2000	EXCEL
2	Small Customers	10-2-2000	EXCEL

Grp#	Customer #	LOAD DTS	RCRD SRC
1	100	10-14-2000	EXCEL
1	101	10-14-2000	EXCEL

Primary Key

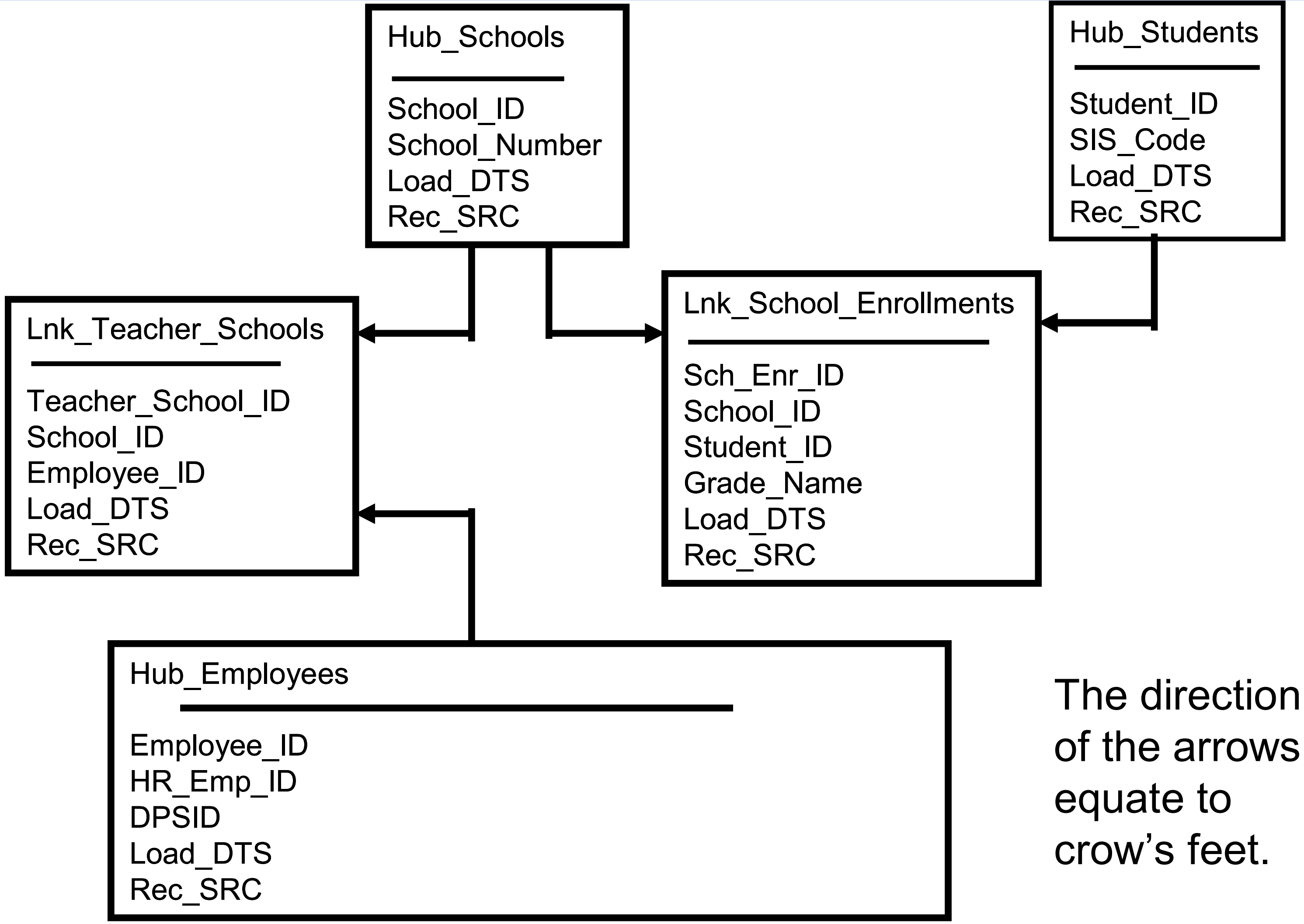
Load DTS

Record Source

BASE TABLE:

ID	Customer #	LOAD DTS	RCRD SRC
100	ABC295882	10-14-2000	FINANCE
101	ABC-1	10-14-2000	FINANCE

Data Vault – How is DPS using DV



The direction of the arrows equate to crow's feet.

Data Vault – Why using DV

- Storage considerations.
- Vertical partitioning of data (rate of change).
- All the FACTS all the TIME.
- Scalability and Extensibility.

Data Vault - Links

- https://bimanu.de/blog/data-vault-modellierungsmodell-von-dan-linstedt/
- https://www.databricks.com/de/glossary/data-vault
- https://de.wikipedia.org/wiki/Data_Vault

Data Vault Hash

- Hashing keys in Data Vault allows integration keys to be loaded in a deterministic way from multiple sources in parallel. This also removes the need for key lookups between related entities.

Data Vault Hash - Links

- https://www.scalefree.com/blog/architecture/hash-keys-in-the-data-vault/
- https://www.tpximpact.com/knowledge-hub/blogs/tech/hash-keys-data-warehousing-1
- https://www.scalefree.com/knowledge/webinars/data-vault-friday/data-vault-hashing-or-not/