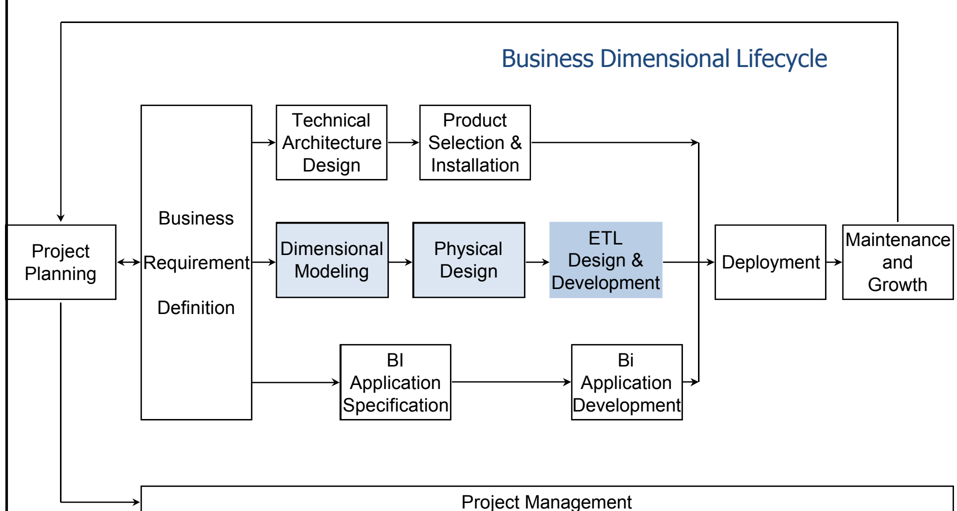
CSI4142 Data Science

Data staging

(Notes by Hl Viktor © Refer to Kimball et. al. Chapters 9 and 10, Han et.al. Chapter 3)

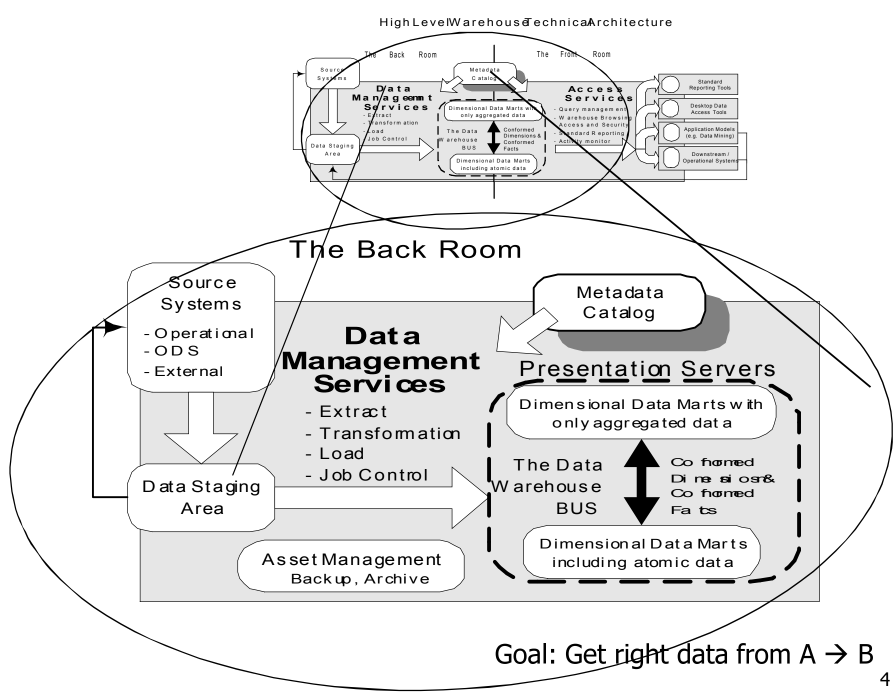
Data Staging (ETL)



The goal of data staging

Getting **the right data**

from A (Sources) to B (Data Marts)



Goal: Get right data from A → B

So, what is the best way to do data staging? Considerations

- Round up the requirements
- Consider the Business Needs
- Study the Sources
- Look out for data limitations
- Decide on scripting languages
- Look at the staff skills
- Remember legacy licences (!!!)

5

The Data staging steps

- A: Planning
  1. High level plan
  2. Choose a tool
  3. Detailed planning: dimension management, error handling
  4. Detailed planning by target table
- B: Develop One-time Historic Load
- C: Develop Incremental Load

6

Step A1: High-level Planning

- Create a very high-level, **one-page schematic** of the source-to-target flow
- Identify starting and ending points
- Label known data sources
- Include placeholders for sources yet to be determined
- Label targets
- Include notes about known problems

7

Step A2: Choose data staging tools

- Do it yourself, in source system code
- Use a data staging tool
  - All major data warehouse vendors now offer one
  - SQL Server Integration Services (SSIS) part of Microsoft SQL Server
  - IBM Cognos DecisionStream
  - IBM WebSphere DataStage
  - Oracle Warehouse Builder
  - Talend Open Studio
  - Syncsort by Syncsort for sorting and summarizing

8

Step A3: General planning

- Extraction from multiple sources (timing, information fusion)
- Archiving (when?)
- Data quality management
- Change management (when? how?)

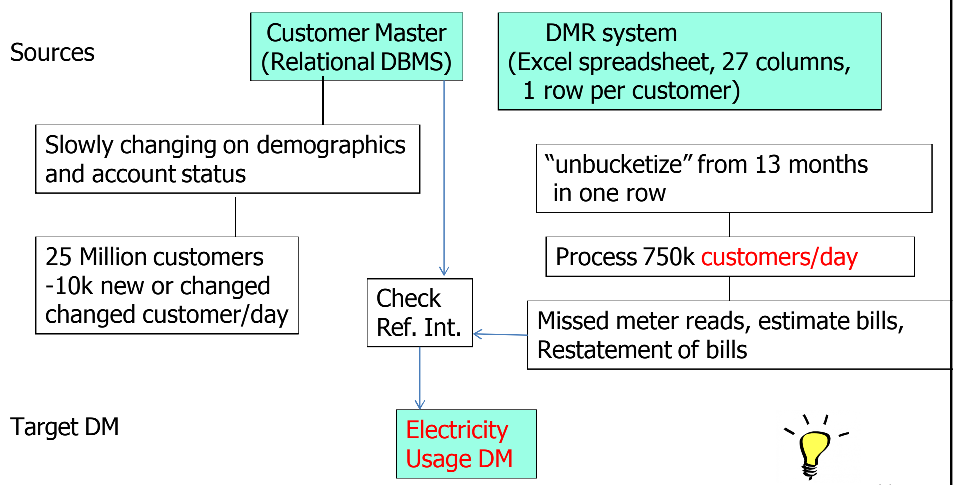
9

Step A4: Detailed planning by table

- Drill down by target table, graphically sketching any complex data restructuring or transformations
- Identify attribute hierarchies (normalize the source)
- Graphically illustrate the surrogate-key generation process
- Develop a preliminary job sequencing

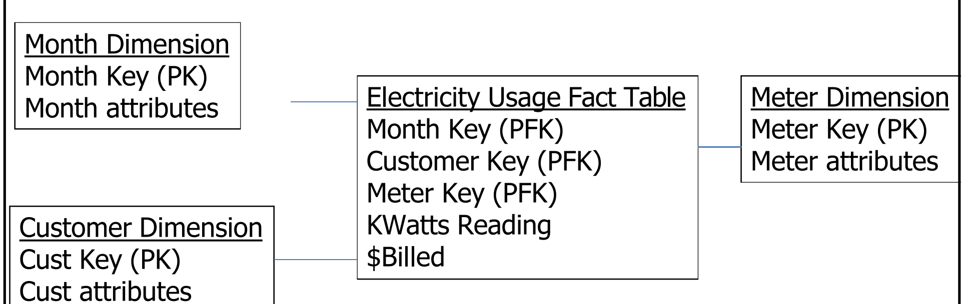
10

Basic high-level data staging plan schematic: An electricity usage example



Electricity usage star schema

- We would get the details about the meter types, etc. from the suppliers
- Customer would include demographic data based on postal code (e.g. from Stats Canada)



Source data

CustomerID	kWh	Bill	KWh	Bill	Kwh	Bill	…	Kwh	Bill
AWD1001	389	$54.59	750	$88.77	500	$76.00		600	$74.01
DWAS4522	900	$203.44	750	$127.61	400	$70.23		700	$101.23
…									

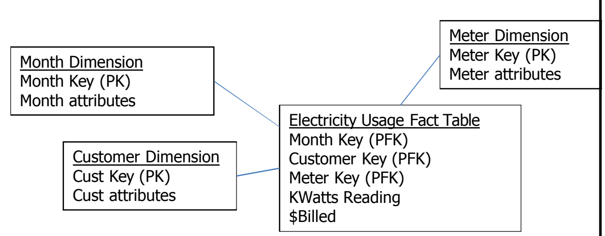
- Readings data from Jan 2016 to Jan 2017 (Excel): 27 columns
- RDMS records for Customers

Customer-ID	Name	City	Meter-type	other attributes
AWD1001	Jane	Calgary	LCD100	
DWAS4522	Joe	Ottawa	RCD203	

13

Unbucketize…

CustomerID	kWh	Bill	Month
AWD1001	389	$54.59	Jan-16
AWD1001	750	$88.77	Feb-16
AWD1001	500	$76.00	Mar-16
…	…	…	…
AWD1001	600	$74.01	Jan-17
DWAS4522	900	$203.44	Jan-16
DWAS4522	750	$127.61	Feb-16
DWAS4522	400	$70.23	Mar-16
…	…	…	…
DWAS4522	700	$101.23	Jan-17

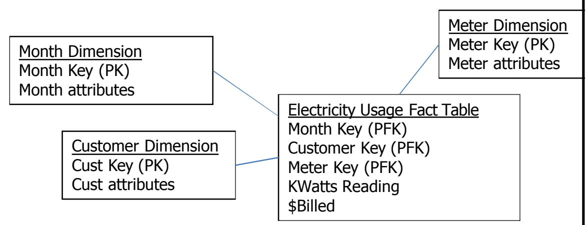


Customer-ID	Name	City	Meter-type	other attributes
AWD1001	Jane	Calgary	LCD100	
DWAS4522	Joe	Ottawa	RCD203	

14

Next steps…

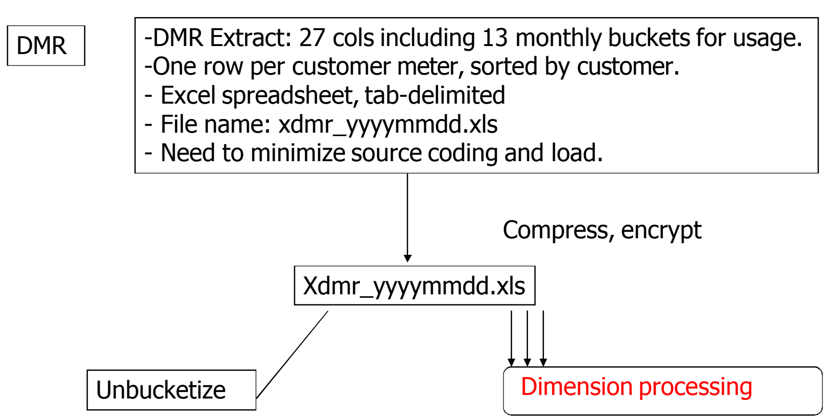
CustomerID	kWh	Bill	Month
AWD1001	389	$54.59	Jan-16
AWD1001	750	$88.77	Feb-16
AWD1001	500	$76.00	Mar-16
…	…	…	…
AWD1001	600	$74.01	Jan-17
DWAS4522	900	$203.44	Jan-16
DWAS4522	750	$127.61	Feb-16
DWAS4522	400	$70.23	Mar-16
…	…	…	…
DWAS4522	700	$101.23	Jan-17



Customer-ID	Name	City	Meter-type	other attributes
AWD1001	Jane	Calgary	LCD100	
DWAS4522	Joe	Ottawa	RCD203	

- Dimension processing
- Historic fact load

First draft of historic load schematic



Dimension Processing…. Customer

- We use surrogate keys (auto-numbers)

Cust-key	Meter-type
1	LCD100
2	RCD203
3	LCD100

Cust-key	Name	City	Province	Other demographic attributes
1	Jane	Calgary	Alberta	
2	Joe	Ottawa	Ontario	
3	Ann	Montreal	Quebec	

Customer-ID	Name	City	Meter-type	other attributes
AWD1001	Jane	Calgary	LCD100	
DWAS4522	Joe	Ottawa	RCD203	
WOW354	Ann	Montreal	LCD100	

Cust-key	*CustomerID*	kWh	Bill	Month
1	*AWD1001*	389	$54.59	Jan-16
1	*AW 10*	750	$88.77	Feb-16
1	*AW 1*	500	$76.00	Mar-16
…	…	…	…	…
1	*AWD 01*	600	$74.01	Jan-17
2	*DW 22*	900	$203.44	Jan-16
2	*DW S 22*	750	$127.61	Feb-16
2	*DWAS4 22*	400	$70.23	Mar-16
…	…	…	…	…
2	*DWAS4522*	700	$101.23	Jan-17
3	…			

17

Dimension Processing…. Meter

- We use surrogate keys (autonumbers)

Cust-key	Meter-kry
1	200
2	201
3	200

Meter-key	Meter-type	Manufacturer	Warranty	Other specs
200	LCD 0	LG	1 year	
201	RC 3	Bell	5 years	

Meter-key	Cust-key	CustomerID	kWh	Bill	Month
200	1	AWD1001	389	$54.59	Jan-16
200	1	AWD1001	750	$88.77	Feb-16
200	1	AWD1001	500	$76.00	Mar-16
	…	…	…	…	…
200	1	AW 1	600	$74.01	Jan-17
201	2	DW 522	900	$203.44	Jan-16
201	2	DW 522	750	$127.61	Feb-16
201	2	DW 22	400	$70.23	Mar-16
	…	…	…	…	…
201	2	DWAS4 22	700	$101.23	Jan-17
200	3	WOW354	910	$123.44	Jan-16
200	3	WOW354	730	$126.61	Feb-16
200	3	WOW354	410	$73.23	Mar-16
	…	…	…	…	…
200	3	WOW354	720	$121.23	Jan-17

18

Dimension Processing…. Date

- We use surrogate keys (autonumbers)
- We may also include a Read-Date (if available)

Meter-key	Cust-key	Month	kWh	Bill	Month
200	1	300	389	$54.59	Jan-16
200	1	301	750	$88.77	Feb-16
200	1	302	500	$76.00	Mar-16
	…	…	…	…	…
200	1	313	600	$74.01	[illegible]
201	2	300	900	$203.44	[illegible]
201	2	301	750	$127.61	[illegible]
201	2	302	400	$70.23	[illegible]
	…	…	…	…	…
201	2	313	700	$101.23	[illegible]
200	3	300	910	$123.44	Jan-16
200	3	301	730	$126.61	Feb-16
200	3	302	410	$73.23	Mar-16
	…	…	…	…	…
200	3	313	720	$121.23	Jan-17

Month-key	Month	Year	Season	Event
300	January	2016	Winter	
301	February	2016	Winter	
302	March	2016	Spring	
…				
313	January	2017	Winter	Canada 150
314	February	2017		Canada 150

19

The Final Data: Fact and Dimensions

- Link via surrogate keys

Month-key	Month	Year	Season	Event
300	January	2016	Winter	
301	February	2016	Winter	
302	March	2016	Spring	
…				
313	January	2017	Winter	Canada 150
314	February	2017		Canada 150

Meter-key	Cust-key	Month	kWh	Bill
200	1	300	389	$54.59
200	1	301	750	$88.77
200	1	302	500	$76.00
	…	…	…	…
200	1	313	600	$74.01
201	2	300	900	$203.44
201	2	301	750	$127.61
201	2	302	400	$70.23
	…	…	…	…
201	2	313	700	$101.23
200	3	300	910	$123.44
200	3	301	730	$126.61
200	3	302	410	$73.23
	…	…	…	…
200	3	313	720	$121.23

Meter-key	Manufacturer	Warranty	Other specs
200	LG	1 year	
201	Bell	5 years	

Cust-key	Name	City	Province	Other demographic attributes
1	Jane	Calgary	Alberta	
2	Joe	Ottawa	Ontario	
3	Ann	Montreal	Quebec	

20

Step A4: First draft of historic load schematic for the fact table (recap)

Unbucketize:
Usage 1 keyed to yyyymm-1….
When yyyymm < subscribe_date, stop processing row.
-Will have 13x rows.
-Data presorted by customer, so can perform cust_key lookup on same pass

↓

Fact_stage1 (only fields relevant to fact table), includes customer and date surrogate keys

↓ Sort by meter type, lookup meter_key

Fact_stage2, include meter surrogate key

↓ Sort by date, lookup month_key

Fact_stage3, includes month surrogate key

21

Planning Step A4: First draft of historic load schematic for the fact table (cont)

Fact_stage3, includes month surrogate key

↓ Optional (Sort by date, lookup read-date key)

Fact_stage4, includes read-date surrogate key

↓ Bulk load into electric_usage_info

Electricity Usage Fact

22

The Data staging steps

- A: Planning
  1. High level plan
  2. Choose a tool
  3. Detailed planning: dimension management, error handling
  4. Detailed planning by target table
- B: Develop One-time Historic Load
  1. Populate dimension tables
  2. Populate fact table (and create data mart)
  3. Consider data preprocessing for analytics
- C: Develop Incremental Load

23

B: Develop one-time historic load

1. Build and test the historic dimension table loads
2. Build and test the historic fact table loads, including surrogate key lookup and substitution

24

Step B1: Populate dimension tables

- Static dimension extract
- Creating and moving the result set
  - Data compression
  - Data encryption
- Static dimension transformation
- Simple data transformations
- Surrogate key assignment
- Combining from separate sources
- Validating one-to-one and one-to-many relationships

25

Surrogate key assignment (another example)

- Use integer "autonumbers", increasing by 1
- Maintain a table with the production_key → surrogate_key matches

SKU	Product	Brand	Supplier
12319319	Milk	Natrel	Saputo
12319336	Milk	Quebon	Saputo
12319353	Milk	Grand Pre	Lactantia
12319370	Cream	Quebon	Saputo
12319387	Cream	Natrel	Saputo
12319404	Brie	Yellow	Metro
12319421	Brie	French	Metro
12319438	Cheddar	Trappe	Fromage
12319455	Gouda	Trappe	Fromage

SKU	Product-key
12319319	
12319336	
12319353	
12319370	
12319387	
12319404	
12319421	
12319438	
12319455	

Product--key	Product	Brand	Supplier
	Milk	Natrel	Saputo
	Milk	Quebon	Saputo
	Milk	Grand Pre	Lactantia
	Cream	Quebon	Saputo
	Cream	Natrel	Saputo
	Brie	Yellow	Metro
	Brie	French	Metro
	Cheddar	Trappe	Fromage
	Gouda	Trappe	Fromage

26

Step B1:
Populate a simple dimension table (cont).

- Load
  - Bulk loader
  - Turn off logging
  - Pre-sort the file
  - Transform with caution
  - Aggregations
  - Use the bulk loader to perform "within-database" inserts
- Index management
  - Drop and re-index
  - Keep indexes in place

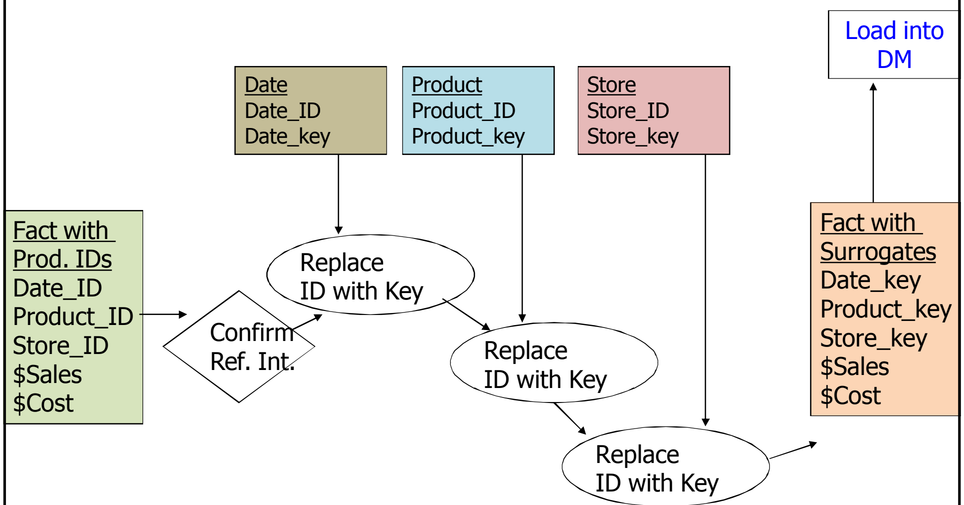
27

Step B3:
Historic load of Atomic-level DM

- Fact table processing
  - Fact table surrogate key lookup
- Ensure Referential Integrity!!!

28

Surrogate key pipeline for Store Example



29

Transformations for Analytics and Data Mining

- Flag normal, abnormal, out of bounds, or impossible facts
- Recognize random or noise values from context and mask out
- Apply a uniform treatment to null values
- Flag fact records with changed status
- Classify an individual record by one of its aggregates
- Add computed fields as inputs or targets
- Map continuous values into ranges
- Normalize values between 0 and 1
- Convert from textual to numeric or numeral category
- Emphasize the unusual case abnormally to drive recognition

30

Steps to transform the data

(Chapter 3 of Han et. al.)

1. Data cleaning
2. Data integration and transformation
3. Data reduction

Design decision: done during data staging or by user applications (or at both ends)

- Depends on domain, organization culture, end user needs and skills

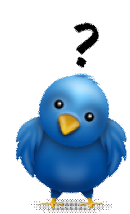
31

Why clean the data?

- Incomplete; e.g. age missing
- Noisy; e.g. age = 130 (human)
- Inconsistent; e.g. province = "BC" and postal code = "K1N"

Others:

- Redundant duplicates (referential integrity: "John Smith")
- Incorrect formats (inches versus meters)
- Etc.



32

Data Cleaning

- Importance
  - "Data cleaning is the number one problem in data science"—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

33

Missing values

Product--key	Product	Brand	Supplier
100	Milk	Natrel	Saputo
101	Milk		
102	Milk	Grand Pre	Lactantia
103	Cream	Quebon	Saputo
104	?	Natrel	Saputo
105	Brie	Yellow	Metro
106	?	?	?
107	Cheddar	Trappe	Fromage
108	Gouda	Trappe	Fromage

- Ignore
- Fill manual
- Use default value (e.g. unknown)
- Use mean value (e.g. average income of all clients)
- Use mean value of class or grouping (e.g. average income of all clients from Orleans suburb in 30-35 age group)
- Use most probable value (e.g. use a decision tree to predict age of a client)

- May introduce BIAS into data
- May not be correct!

34

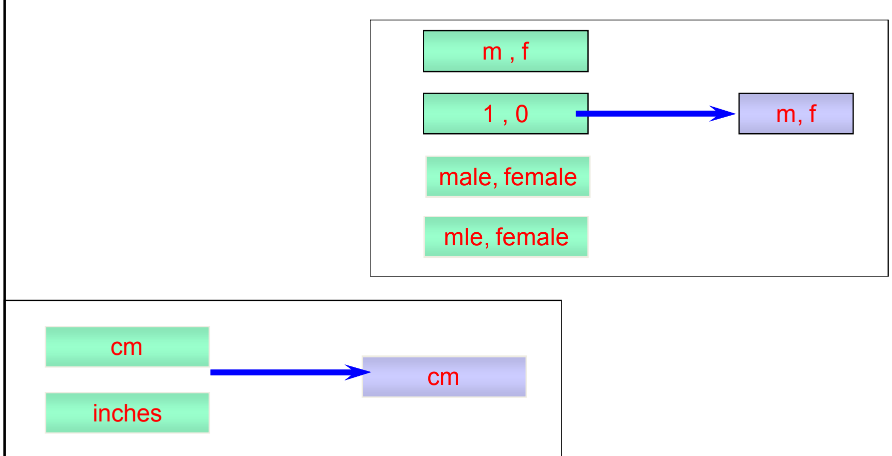
Data Cleaning: Transform data

- Eliminate anomalies:
  - No unique key
  - Data naming, coding anomalies
  - Data meaning anomalies
  - Spelling and text inconsistencies

CUSNUM	NAME	ADDRESS
90328575	Oracle Corp	100 NE 1st Street, Tampa
90328575	Oracle	100 NE. First St., Tampa
90238475	Oracle Services	100 North East 1st St., FLA
90233479	Oracle Limited	100 N.E. 1st St.
90233489	Oracle Computing	15 Main Road, Ft. Lauderdale
90234889	Oracle Corp. UK	15 Main Road, Ft. Lauderdale, FLA
90345672	Oracle Corp UK Ltd	181 North Street, Key West, FLA

35

Data Cleaning: Multiple standards



36

Data Cleaning: Noisy Data

- Noise: **random error or variance in a measured variable**
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - inconsistency in naming convention

SKU	Product	Brand	Supplier	Price
12319319	Milk	Natrel	Saputo	$5.99
12319336	Milk	Quebon	Saputo	$5.49
12319353	Milk	Grand Pre	Lactantia	$4.99
12319370	Cream	Quebon	Saputo	$3.49
12319387	Cream	Natrel	Saputo	$449.00
12319404	Brie	Yellow	Metro	$7.99
12319421	Brie	French	Metro	$6.49
12319438	Cheddar	Trappe	Fromage	$6.99
12319455	Gouda	Trappe	Fromage	$0.19

37

Data Cleaning: Noise

Idea: Smooth out the noise from the data

- Binning: place data in buckets or bins of neighbors
  - 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: fit the data to a function using linear or multiple linear regression (more later)
- Clustering: useful for finding outliers (more later)
- Should always involve human inspection

38

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

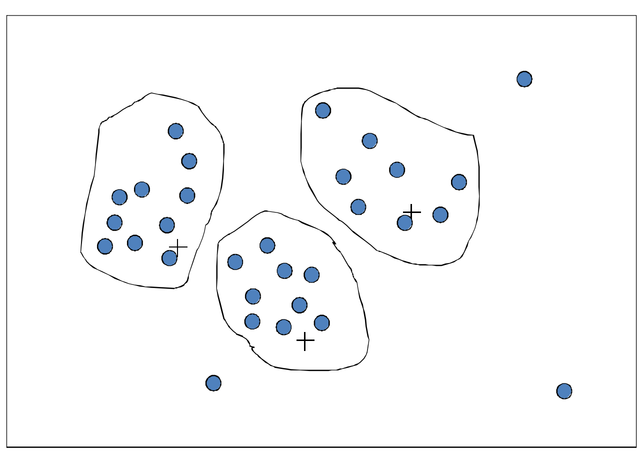
* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

39

Cluster Analysis: See the outliers



40

Steps to transform the data

(Chapter 3 of Han et. al.)

1. Data cleaning
2. Data integration and transformation
3. Data reduction

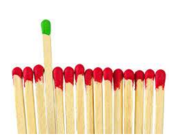
41

Data integration and Information fusion

- Top-down (schema level) versus bottom-up (data driven) versus hybrid approaches

Goal: Keep original information as far as possible

- Schema integration
- Object matching
- Useful when multiple sources



42

Data Transformation: Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  – Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  – Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|ν'|) < 1

43

Data transformation: Attribute construction

cust-key	name	date-of-birth	income
200	Joe	10-Dec-00	120,000
201	Jane	12-Dec-99	23,000
202	James	12-Dec-70	23,300
203	Joey	13-Dec-71	89,000
204	Joel	02-Jan-70	18,000
205	Jamie	02-Feb-85	88,000
206	Joanna	04-Mar-89	121,000
207	Joaxim	03-Dec-99	22,900
208	Jo	04-Jan-98	90,000

44

Data transformation: Attribute construction

cust-key	name	Age	income
200	Joe	16	120,000
201	Jane	17	23,000
202	James	46	23,300
203	Joey	45	89,000
204	Joel	47	18,000
205	Jamie	32	88,000
206	Joanna	28	121,000
207	Joaxim	17	22,900
208	Jo	19	90,000

cust-key	name	Age-range	income range
200	Joe	12 to 17	120K to 130K
201	Jane	12 to 17	10K to 25K
202	James	45 to 50	10K to 25K
203	Joey	45 to 50	80K to 90K
204	Joel	45 to 50	10K to 25K
205	Jamie	30 to 40	80K to 90K
206	Joanna	26 to 30	120K to 130K
207	Joaxim	12 to 17	10K to 25K
208	Jo	18 to 25	80K to 90K

- Domain dependent

45

Steps to transform the data

(Chapter 3 of Han et. al.)

Tasks

1. Data cleaning
2. Data integration and transformation
3. Data reduction
   1. Attribute (Feature) Selection
   2. Sampling

46

Detecting redundancies: Attribute selection

Given two attributes *a* and *b*, measure how strongly

a → *b*

- E.g. there is (should be) a direct correlation between Date-of-birth and Age ☺
- We only need to keep one!
- Detecting correlations is actually a HUGE general problem in data mining and Big Data applications
- Consider the combinatorics!!!!
- Built-in in Machine Learning environments (later)

47

Correlation Analysis (Categorical Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: **population**

48

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group (different from expected!)

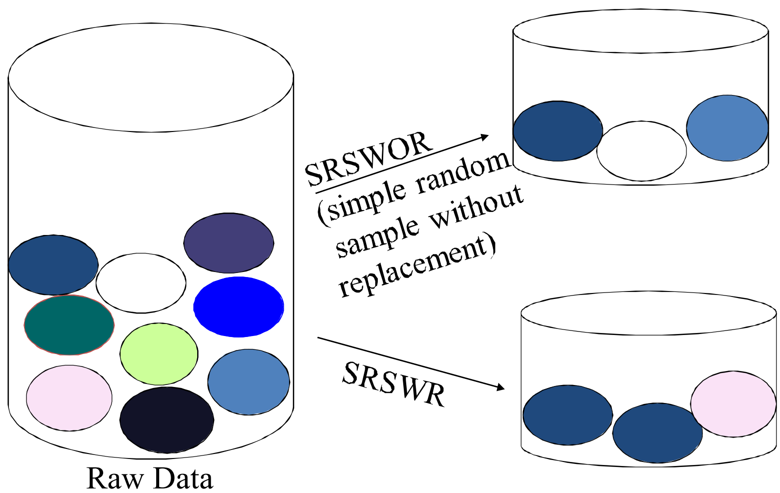
49

Data Reduction Method: Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Choose a representative subset of the data
  – Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  – Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Beware: If not carefully designed, then sampling may not reduce database I/Os (page at a time)

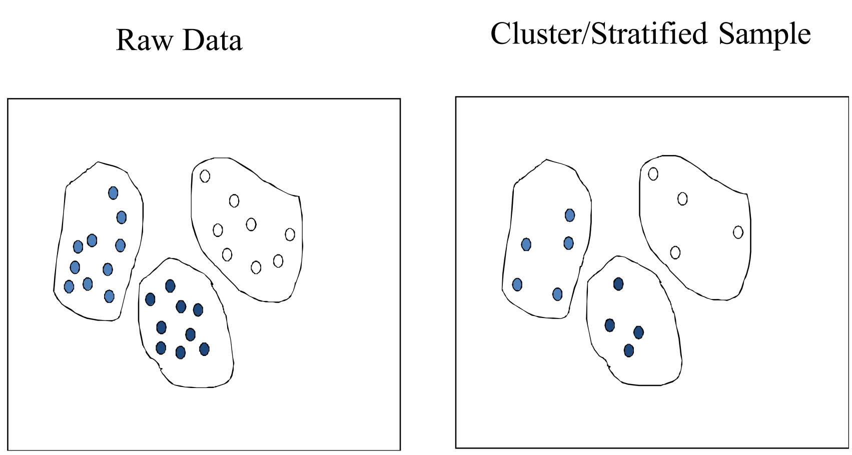
50

Sampling: with or without Replacement



51

Sampling: Cluster or Stratified Sampling



52

Reflection

Data preprocessing is a crucial, time consuming step of any data science effort

- Remember: "junk in, junk out"

Tasks

1. Data cleaning
2. Data integration and transformation
3. Data reduction

53

So, you have designed your data mart, loaded the historic data....

What is next, in terms of data staging?

54

The Next Step…

C: Incremental table staging

Step C1: Determining whether dimension record has been changed

Today's complete set
of dimensional records
Customer_id
Customer_type
Name
Address
Service_begin_date
….

Yesterday's complete set
of dimensional records
Customer_id
Customer_type
Name
Address
Service_begin_date
….

Compare difference &
Apply policy decisions

Known changed set
of dimensional records
Customer_id
Customer_type
Name
Address
Service_begin_date
….

56

Handling change

Slowly changing dimensions

- Type 0: No change (e.g. Date-of-Birth)
- Type 1: Overwrite
- Type 2: Add new row
- Type 3: Keep history (add new attribute)
- Type 4: Add history table/dimension

57

Handling Change: Type 1 overwrite

- Often caused by data capturing errors

Cust-key	Name	Age	City	Marital-Status
122	Ann	20	Ottawa	Single

⇓

Cust-key	Name	Age	City	Marital-Status
122	Anne	20	Ottawa	Single

58

Handling Change: Type 2a

- Add new row

Cust-key	Name	Age	City	Marital-Status
122	ANN	20	Ottawa	Single

- Suppose we currently have 2345 cust-keys in our mart

Cust-key	Name	Age	City	Marital-Status
2346	ANN	20	Montreal	Single

- From today, Ann is linked to the FACT using cust-key 2346

59

Handling Change: Type 2b

- Add new row

Cust-key	Name	Age	City	Marital-Status
122	ANN	20	Ottawa	Single

- Suppose we currently have 2345 cust-keys in our mart

Cust-key	Name	Age	City	Marital-Status	Current?
122	ANN	20	Ottawa	Single	No
2346	ANN	20	Montreal	Single	Yes

- From today, Ann is linked to the FACT using cust-key 2346

60

Handling Change: Type 2c

- Add new row

Cust-key	Name	Age	City	Marital-Status
122	ANN	20	Ottawa	Single

- Suppose we currently have 2345 cust-keys in our mart

Cust-key	Name	Age	City	Marital-Status	Effective-date
122	ANN	20	Ottawa	Single	13/2/2002
2346	ANN	20	Montreal	Single	1/1/2018

- From today, Ann's record is linked to the FACT with cust-key 2346

61

Handling Change: Type 3

- A new attribute is used to keep history

Cust-key	Name	Age	City	Marital-Status
122	ANN	20	Ottawa	Single

Cust-key	Name	Age	City	Old-Marital-Status	Effective date	New-Marital-Status
122	ANN	20	Ottawa	Single	14/02/2018	Married

62

Handling Change: Type 4

- Add another, new, separate "history" dimension
- Customer dimension has current data:

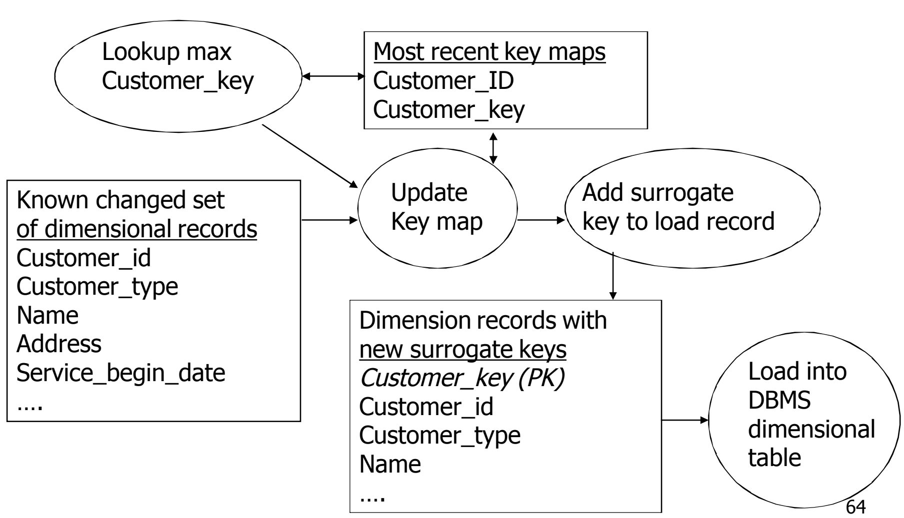
Cust-key	Name	Age	City	Marital-Status
2347	ANN	20	Montreal	Married

- "Customer-History" dimension keeps history:

Cust-key	Name	Age	City	Marital-Status	Effective-date
122	ANN	20	Ottawa	Single	13/2/2002
2346	ANN	20	Montreal	Single	1/1/2018
2347	ANN	20	Montreal	Married	14/2/2018

63

Step C1: Handling changed dimensions



64

Step C2: Incremental Fact Table Staging

- Incremental fact table extracts
  - New transactions
  - Updated transactions (correcting info)
  - Database logs
  - Replication
- Incremental fact table load
- Speeding up the load cycle
  - More frequent loading
  - Partitioned files and indexes
  - Parallel processing

65

Step D: Aggregate Table and OLAP Loads

- Build aggregates
- Maintain aggregates
- Prepare OLAP loads (if any)
  - Cube-like structure based on dimensional model
  - MOLAP engines build own optimized aggregates
    - Oracle Essbase
    - Microsoft Analysis Services
    - DB2 UDB OLAP

66

The last data staging step: Automation

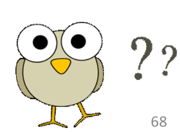
- Typical operational functions
  - Job definition: flow and dependency
  - Job scheduling: time and event based
  - Monitoring
  - Logging
  - Exception handling
  - Error handling
  - Notification
- Determine job control approach
- Record extract metadata
- Record operations metadata
- Ensure data quality
- Set up archiving in the data staging area
- Develop disk space management procedures

67

Summary…

- Designing and building the data mart
  - Dimensional modeling
  - Aggregates and Indexes
  - Data staging
- Online Analytic Processing (OLAP)

Next: Machine Learning



68