

Information Integration



SDSC SAN DIEGO
SUPERCOMPUTER CENTER

After this video you will be able to

- Explain the data integration problem
- Define integrated views and schema mapping
- Describe the impact of increasing the number of data sources
- Appreciate the need to use data compression
- Describe Record Linking, Data Exchange and Data Fusion Tasks

A Business Case (from IBM)

ex

*“Mergers and acquisitions in the past decade have increased our customer base by 200 percent.
Having a single view of the customer, we’re more accurately able to target and cross-sell across our brands.”*

Suncorp is a diversified financial services group that offers general insurance, banking, life insurance and wealth management services. With operations in Australia and New Zealand, Suncorp has over AU\$95 billion in assets, more than 16,000 employees and relationships with over nine million customers. The financial services organization maintains five operating divisions, managing 14 market brands, and is supported by corporate and shared services divisions.

Suncorp-Metway Ltd wanted a single, integrated view of its customers to ensure its marketing campaigns didn't encourage internal conflict between the brands and duplication of efforts, both of which had a negative effect on the bottom line.

solution
↓
as single database
information integration

Deconstructing the Case – Hypothetically

↓ ↓
↓ ↓

Insurance Company's Partial Schema

Policies(PolicyKey, PolicyTypeKey, Agent, Conditions)

PolicySales(PolicyKey, PolicyholderKey, StartDate,
TransactKey, Premium, CoveragePeriod,
CoverageLimit)

Transactions(TransactKey, Date, Time, Amount,
Balance)

Policyholders(PolicyHolderKey, Name, Address,
City, State, ZIP)

Claims(PolicyKey, ClaimKey, TransactKey,
ClaimAmount)

ClaimDescription(ClaimKey, TypeKey, ClaimantKey,
ProcCode, Description)

Claimants(ClaimantKey, Name, Address, City, State,
ZIP)

ClaimTypes(TypeKey, Description)

PolicyTypes(PolicyTypeKey, Name, Description)

Bank's Partial Schema

Accounts(AcctNumber, AcctType, MemberID,
MemberType, TypeID, StartDate, EndDate,
InterestRate, CreditLimit)

Individuals(MemberID, FName, MI, LName, SSN,
Nationality, DoB, LegalStatus,
FullAddress, Phone, PhoneType, Email)

Corporations(MemberID, Name, RegisteredAddress,
CorporationType, Signatory1,
Signatory2, DNBNumber, Phone, Email)

Transactions(TrID, AcctNum, Date, Time,
TransactionType,
Description, TransactionAmount,
Debit/Credit, Balance, Payoff)

AccountType(TypeID, Name, Description)

TransactionTypes(Ttype, Name, Description)

Disputes(AccntNumber, DisputeID, TrID, Date,
DisputeAmt, Explanation, Valid, ValidatorID)

Integrated Views

PolicySales(PolicyKey, PolicyholderKey,
StartDate, TransactKey, Premium,
CoveragePeriod, CoverageLimit)

Policyholders(PolicyHolderKey, Name,
Address, City, State, ZIP)

Accounts(AcctNumber, AcctType, MemberID,
MemberType, TypeID, StartDate,
EndDate, InterestRate, CreditLimit)

Individuals(MemberID, FName, MI, LName,
SSN, Nationality, DoB,
LegalStatus, FullAddress, Phone,
PhoneType, Email)

- Find current customers of the insurance company who are also customers of the bank, and create this integrated view

discountCandidates(custID,
address, policyKey, AcctNumber)

join

- from different data sources build
- view: relation from others

Schema Mapping

Policyholders(PolicyHolderKey, Name,
Address, City, State, ZIP)

Individuals(MemberID, FName, MI, LName, SSN,
Nationality, DoB, LegalStatus, FullAddress, Phone,
PhoneType, Email)

discountCandidates(custID, address, policyKey, AcctNumber)

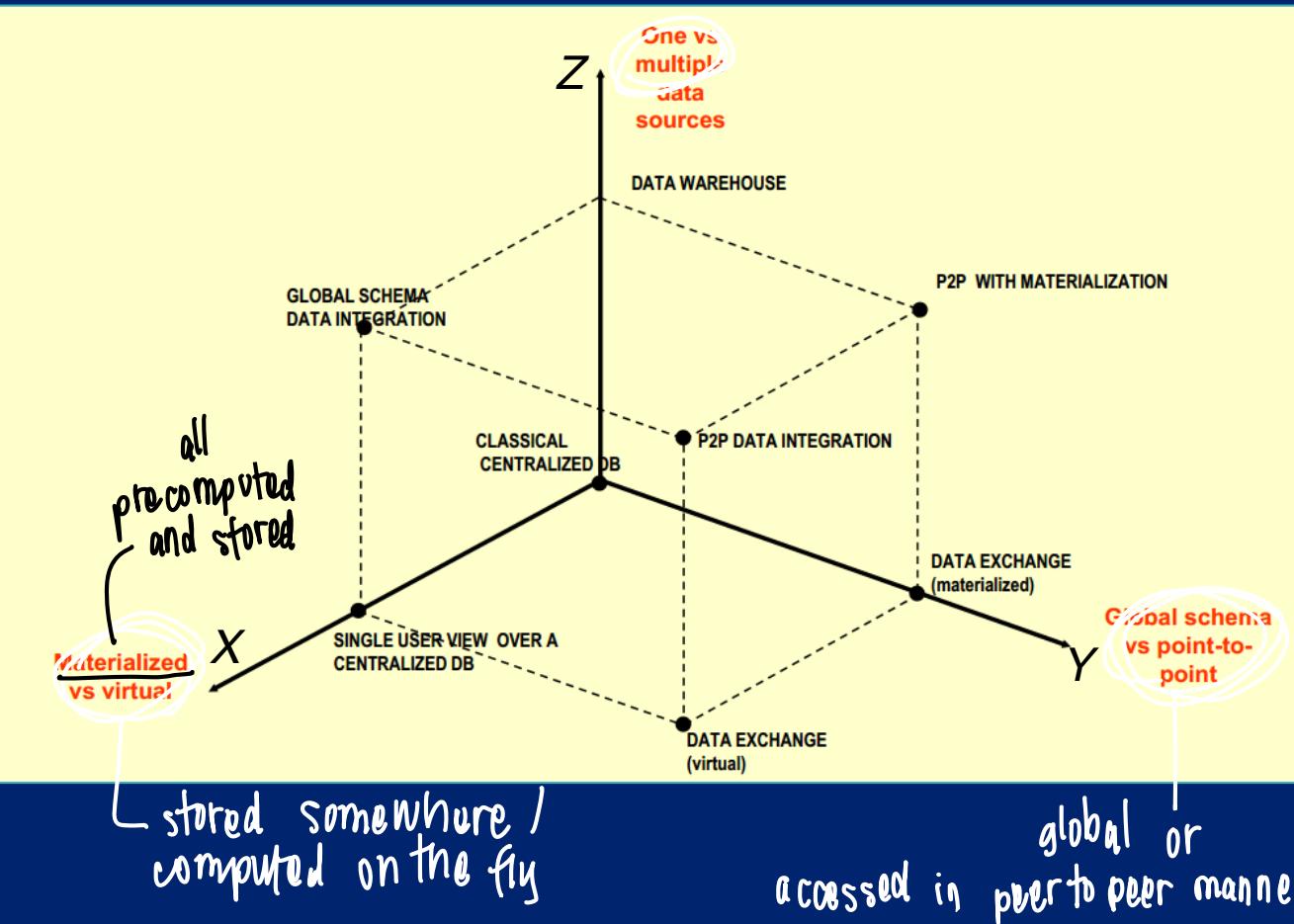
PolicySales(PolicyKey, PolicyholderKey,
StartDate, TransactKey, Premium,
CoveragePeriod, CoverageLimit)

Accounts(AcctNumber, AcctType, MemberID,
MemberType,TypeID, StartDate,
EndDate, InterestRate, CreditLimit)

• Step
establish correspondence between
the attributes of the view (target
relation)

and of source
relations

Querying Integrated Data



- Find the bank account number of a person whose policyKey is known
 - SELECT AcctNumber FROM discountCandidates where policyKey = '4-937528734'

query architecture of
The data integration
system

Record Linkage

Policyholders(PolicyHolderKey, Name,
Address, City, State, ZIP)

Individuals(101, Stephen, C., Jones, 123-45-6789, US,
10/02/1983, citizen, "231 Cedar St. LA, CA 90005", 661-266-9374,
landline, scjones@gmail.com)

Individuals(102, Elizabeth, , McFarlane, 123-54-6789, US,
06/18/1978, citizen, "4157 Elm St. LA, CA 90005", 213-266-9374,
mobile, emlane@gmail.com)

Individuals(103, Liz, P., McFarlane-Gray, 123-92-2318, US,
06/18/1978, citizen, "231 Cedar St. LA, CA 90005", 213-702-4343,
landline, emlane@gmail.com)

Individuals(104, Lisa, M., Brady, 423-45-6209, US, 08/09/1975,
foreign-student, "231 Cedar St. LA, CA 90005", 302-266-9374,
landline, scjones@gmail.com)

Policyholders(3-764528104, Liz, P., McFarlane-Gray, 4157 Elm
St. LA, CA, 90005) → which address should be used?
→ which bank account?

problem → ensure that set of records that belong to single entity are recognized
— clustering, matching rules
Individuals(MemberID, FName, MI, LName, SSN,
Nationality, DoB, LegalStatus, FullAddress, Phone,
PhoneType, Email)

complete : no eligible record from the source should be absent

and accurate: all the entries in the integrated relation should be correct

matching person
• is in both databases
• has the same name & address in the two databases

The “Big Data” Problem

- **Many sources**
 - Hundreds of tables
 - Schema mapping problem is a combinatorial challenge
- **Pay-as-you-go model**
 - Only integrate sources that are needed when needed
- **Probabilistic Schema Mapping**

how elements from
~~two~~ sources would relate +
to each other

determine how
they relate to target
schema

Designing Mediated Schema

- **Customers – an integrated table**
- **Candidate designs**
 - Create the customer table to include individuals and corporations – use a flag called customer_type to distinguish. In the mediated schema
 - Individual.(FN+MI+LN), PolicyHolder.Name, Corporations.Name map to Customer_Name
 - Names of two types of customers become two different attributes
 - Ind.FullAddress, Corp.RegisteredAddress, PH.(Address+City+State+Zip) map to Customer_Address
- **Issues**
 - Should the DOB, Nationality, Legal Status be included in this table?
+ associating probability values with each of these options

Attribute Grouping

- How to evaluate if two attributes should go together?

- How similar are the attributes —
 - Individual names vs. policyholder names?
 - Individual names vs. Corporation names?
 - How likely is it that two attributes would co-occur?
 - Should the DOB put in the same schema as the individual name?
 - How about DOB and the corporation name?
- how closely the two attrs represent the same real world concept
- _ when 2 attrs are likely to be different

Customer Transactions

BankTransactions(TransactionID (TID),
TransactionBeginTime(TBT), TransactionEndTime(TET),
TransactionAmount(TA), Credit-Debit(CD),
TransactionParty(TP), Transaction Description(TD), Balance(B),
Payoff(P))

InsuranceTransactions(TransactionID (TID), TransactionDateTime
(TDT), TransactionType(TT), Amount(A), TransactionDetails(TDT))

Med1($\{TID\}$, $\{TBT, TET, TDT\}$ $\{TA+CD, A\}$, $\{TP, TD, TDT\}$, $\{TT\}$,
 $\{B\}$, $\{P\}$)

Med2($\{TID\}$, $\{TBT\}$, $\{TET\}$, $\{TDT\}$ $\{TA+CD, A\}$, $\{TP\}$, $\{TD\}$, $\{TDT\}$,
 $\{TT\}$, $\{B\}$, $\{P\}$)

Med3($\{TID\}$, $\{TBT, TDT\}$, $\{TET, TDT\}$ $\{TA+CD, A\}$, $\{TP\}$, $\{TD, TDT\}$,
 $\{TT\}$, $\{B\}$, $\{P\}$)

...

Compute pairwise attribute similarity and using a threshold plus/minus an error, put similar attributes in the same cluster

For every subset of uncertain pairs create a mediated schema

which one to choose?

Probabilistic Mediated Schema

qualitative account of the method

- Find source schemas that are consistent with a mediated schema

- A source schema is consistent with a mediated schema if two different attributes of the source schema do not occur in a cluster in two different clusters

- candidate
mediate
schemas {
- $\text{Med3}(\{\text{TID}\}, \{\text{TBT}\}, \{\text{TDT}\}, \{\text{TET}\}, \{\text{TDT}\})$ {TA+CD, A}, {TP}, {TD, TDT}, {TT}, {B}, {P}) is better than * more consistent with bank transactions
 - $\text{Med1}(\{\text{TID}\}, \{\text{TBT}, \text{TET}, \text{TDT}\})$ {TA+CD, A}, {TP, TD, TDT}, {TT}, {B}, {P}) with respect to BankTransactions

- Choose the k best mediated schema — when top capability estimates are close to each other
 - count number of consistent sources for each candidate mediated schema
 - use this count to come up with a probability estimate

Pause

A Data Integration Scenario

- 4 data sources each with one relation

- S₁: Treats(Doctor, ChronicDisease)
- S₂: Discharges(Doctor, Patient, Clinic)
- S₃: Treats(Doctor, ChronicDisease)
- S₄: Surgeons(SurgeonName)

same schema — possible,
independent from each
other

- Target schema (manually designed)

- TreatsPatient(Doctor, Patient)
- HasChronicDisease(Patient, ChronicDisease)
- DischargesPatientsFromClinic(Doctor, Patient, Clinic)
- Doctors(DoctorName)
- Surgeons(SurgeonName)

* possibility to
add more sources
(more clinics)
as the system
grows

Example Schema Mapping

- Local-as-View (LAV) mapping — write the relations in each source as a view over the target schema
 - Mapping source schemas to target schema
 - Easier to add sources

maps to
/

```
SELECT doctor, chronicDisease
FROM TreatsPatient T, HasChronicDisease H
WHERE T.Patient = H.Patient
```

S1. $Treats(d, s) \rightarrow TreatsPatient(d, p) \text{ AND } HasChronicDisease(p, s)$

S2. $Discharges(d, p, c) \rightarrow DischargesPatientFromClinic(d, p, c)$

S3. $Treats(d, s) \rightarrow TreatsPatient(d, p) \text{ AND } HasChronicDisease(p, s) \text{ AND } Doctors(d)$

S4. $Surgeons(d) \rightarrow Surgeons(d)$

Query Answering

- **Query**

- Which doctors are responsible for discharging patients?
- SELECT DoctorName
- FROM Doctors D₁, DischargesPatientsFromClinic D₂
- WHERE D₁.DoctorName = D₂.DoctorName



how to translate
this query to a
query that can
be sent to the
sources

- **Query reformulation**

- Automatically transform query against the target schema to
the simplest query against source schemas

- **Ideal Answer**

- SELECT Doctor
- FROM S₃.Treats T, S₂.Discharges D
- WHERE T.Doctor = D.Doctor

becomes a scalability problem in a big data integration scenario

becomes worse as
a number of sources
increases

Integration of Public Health Infrastructure

— connects local hospitals
in the Washington DC area
designed to exchange disease information

Washington DC Disease Surveillance System (WADSS)

CATEGORY	SOLUTION	DESCRIPTION
Data Exchange	Integration hub with HL7 messaging	All internal and external data moves through commercial integration hub that transforms HL7 V2 data into a consistent HL7 V3 representation.
Terminology	SNOMED LOINC	Implemented standard concept terminologies SNOMED and LOINC for coding of clinical and lab data.
Conceptual	RIM-based integrated data repository	A centralized, commercial data repository was natively designed on the HL7 RIM to normalize clinical data from disparate sources. Implemented a data quality algorithm to manage patient matching and identify duplicate records.
Architecture	PHIN architecture	Developed architecture consistent with the CDC's Public Health Information Network requirements.

*Used to enable interoperability between existing hospital and lab systems and WADSS.

Reference Information Model — global schema that this industry has developed and expects to use as a standard

Data Exchange

Health Level-7 or HL7 refers to a set of international standards for transfer of clinical and administrative data between software applications used by various healthcare providers.

```
<!DOCTYPE ADT_A03 SYSTEM "hl7_v231.dtd">
<ADT_A03>
<MSH>
  <MSH.1></MSH.1>
  <MSH.2>^&lt;&gt;</MSH.2>
  <MSH.3><HD.1>LAB</HD.1></MSH.3>
  <MSH.4><HD.1>767543</HD.1></MSH.4>
  <MSH.5><HD.1>ADT</HD.1></MSH.5>
  <MSH.6><HD.1>767543</HD.1></MSH.6>
  <MSH.7>19900314130405</MSH.7>
  <MSH.9>
    <CM_MSG_TYPE.1>ADT</CM_MSG_TYPE.1>
    <CM_MSG_TYPE.2>A04</CM_MSG_TYPE.2>
  </MSH.9>
  <MSH.10>XX3657</MSH.10>
  <MSH.11><PT.1>P</PT.1></MSH.11>
  <MSH.12><VID.1>2.3.1</VID.1></MSH.12>
</MSH>
<EVN>
  <EVN.1>A01</EVN.1>
  <EVN.2>19980327101314</EVN.2>
  <EVN.3>19980327095000</EVN.3>
  <EVN.4>I</EVN.4>
  <EVN.6>19980327095000</EVN.6>
</EVN>
<PID>
  <PID.1>1</PID.1>
  <PID.3.LST>
    <PID.3><CX.1>123456789ABCDEF</CX.1></PID.3>
```

HL-7

```
MSH|^~\&|LAB|767543|ADT|767543|19900314130405||ADT^A04|XX3657|P|2.3.1<CR>
EVN|A01|19980327101314|19980327095000|||19980327095000<CR>
PID|1||123456789ABCDEF|123456789ABCDEF|PATIENT^BOB^S||19590520|M||
  612345 MAIN STREET^^ANYTOWN^CA^91234||714-555-1212|
  714-555-1212|||123456789ABCDEF||U<CR>
PD1|||WELBY<CR>
PV1||0||NEW|||SPOCK<CR>
```

↳ data variety
problem

- Find all prescriptions and lab reports of patient #19590520 containing serum protein, along with age-specific normal values between 1/1/2012 and 9/1/2015
 - The patient went to three different clinics and four different labs in this period
 - The doctor's own office uses a relational database for EHR

data exchange problem
transform data from
source schema to the
schema of the receiving
system (RIM)

Data Exchange

- Given a source database with a finite number of relations, a set of schema mappings, and a set of constraints that the target schema must satisfy, the data exchange problem is to find a finite target database such that both the schema mappings and the target constraints are satisfied.

how to populate
the tuples in the
target
database

*Using Codebooks

- Logical Observation Identifiers Names and Codes (LOINC) is a database and universal standard for identifying medical laboratory observations.

2. COMPONENT	Text	255	First major axis-component or analyte
3. PROPERTY	Text	30	Second major axis-property observed (e.g., mass vs. substance)
4. TIME_ASPECT	Text	15	Third major axis-timing of the measurement (e.g., point in time vs 24 hours)
5. SYSTEM	Text	100	Fourth major axis-type of specimen or system (e.g., serum vs urine)
6. SCALE_TYP	Text	30	Fifth major axis-scale of measurement (e.g., qualitative vs. quantitative)
7. METHOD_TYP	Text	50	Sixth major axis-method of measurement
8. CLASS	Text	20	An arbitrary classification of the terms for grouping related observations together. The current classifications are listed in Table 32. We present the database sorted by the class field within class type (see field 23). Users of the database should feel free to re-sort the database in any way they find useful, and/or to add their own classifying fields to the database. The content of the laboratory test subclasses should be obvious from the subclass name.

BP.ATOM	Blood pressure atomic
BP.CENT.MOLEC	Blood pressure central molecular
BP.MOLEC	Blood pressure molecular
BP.PSTN.MOLEC	Blood pressure positional molecular
BP.TIMED.MOLEC	Blood pressure timed molecular
BP.VENOUS.MOLEC	Blood pressure venous molecular
CARD.RISK	Cardiac Risk Scales Framingham
CARD.US	Cardiac ultrasound (was US.ECHO)
CARDIO-PULM	Cardiopulmonary
CLIN	Clinical NEC (not elsewhere classified)

MOLPATH.DELDUP	Gene deletions or duplications
MOLPATH.INV	Gene inversion
MOLPATH.MISC	Gene miscellaneous
MOLPATH.MUT	Gene mutation
MOLPATH.REARRANGE	Gene rearrangement
MOLPATH.TRINUC	Gene trinucleotide repeats
MOLPATH.TRISOMY	Gene chromosome trisomy
MOLPATH.TRNLOC	Gene translocation

Using Compressed Data

- reducing data size through compression
so that this encoder form is smaller than the original representation
- **Compression** : creating an encoded representation of data ,

- Encoded representation of data so that it uses less space
- Dictionary encoding — column store relational database

Record #	Patient ID	Date	Test Code	Test Result
1	100	1/1/2012	SE-AC	14.5
2	502	1/1/2012	BP-S	123
3	301	1/2/2012	HAC	5.8
4	502	1/1/2012	BP-D	91
...
...
10M	1274	7/20/2016	SE-AC	13.8

Using Compressed Data

• Compression

- Encoded representation of data so that it uses less space
- Dictionary encoding

replace with a number

Record #	Patient ID	Date	Test Code	Test Result
1	100	1/1/2012	32	14.5
2	502	1/1/2012	125	123
3	301	1/2/2012	174	5.8
4	502	1/1/2012	126	91
...
...
10M	1274	7/20/2016	32	13.8

Orig. Test Code	Encoded Test Code
SE-AC	32
BP-S	125
HAC	174
BP-D	126
...	...
...	...
SE-AC	32

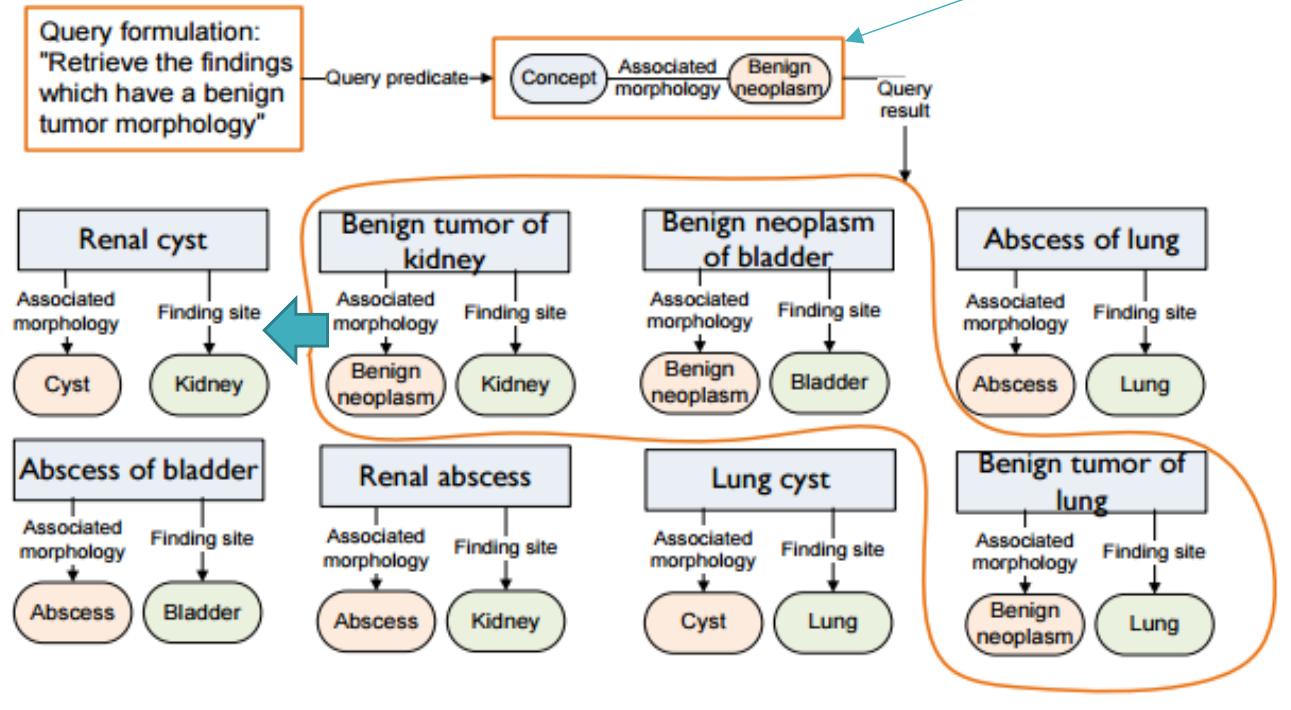
Data compression is an important technology for big data.

reduce the amount of space required by storing data in a column oriented data store and by using compression

Ontological Data

Ontology queries are graph queries

Example: Result of retrieving concepts with */associated morphology/* specified as */benign neoplasm/*



Ontology

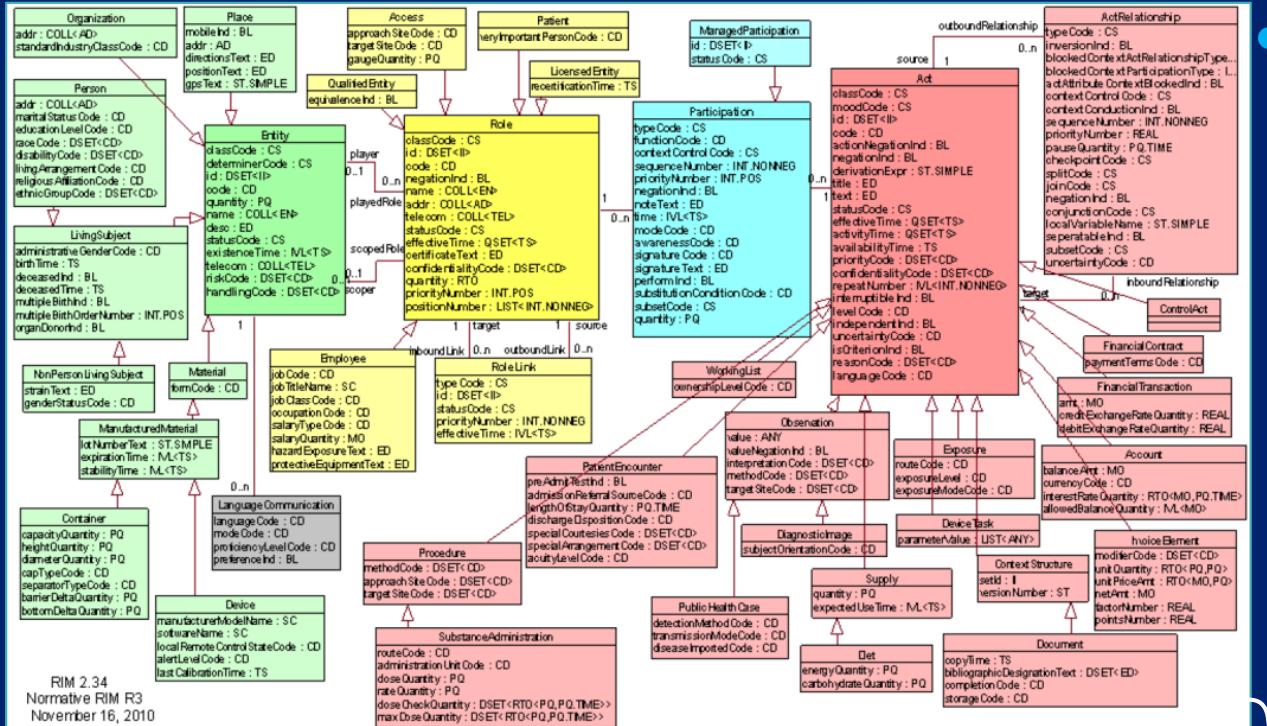
- A set of terms of a domain
- Relationships between the terms

systematized nomenclature of medicine

SNOMED

- A medical ontology used for clinical data

The Takeaway Points



an integration system in a public health domain (and others)
must be able to handle variety

process of taking data represented in one model in one source system and converting it to equivalent data in another model (the target system)

Integration across multiple data models

- Global Schema – RIM
- Data Exchange

- Format conversions
- Constraints
- Compressed Data — store and query efficiency
- Model transformation

Query transformation:
process of taking a query on the target schema
converting it to a query against a different data model

queries and analysis performed by a data analysts should be against this global schema

Pause

Integration for Multichannel Customer Analytics

forums
surveys

- **Customer analytics**

- processes and technologies that give organizations the customer insight necessary to deliver offers that are anticipated, relevant and timely

- **Questions one would like to ask**

- Is our product launch going well?
- Is there an emerging product issue?
- Where should the product team focus its development dollars?
- Are there more effective methods for positioning current products?
- Which services have the best chance of surviving a turbulent market?
- Is there a product defect in the market?



* for customer
focused products
relates to
→ identify info that
the companies product, its features and its
utility

Data Fusion

task

goal: find the values of data items from a source

S: set of data sources

- **Data sources**

- **Data Items**

D: set of data items
: aspect of real world entity

- A product
- A part of a product
- A feature of a product
- The utility of a product feature
- ...

value

— for each data item a source can (but not necessarily will) provide a value

- Using data from a subset of sources find the true value or a true value distribution of a data item

because not all data sources

will have relevant information about the data item

- Assemble all such values for the real-world entity represented by the data items

Too Many Sources

- Too many sources = too many values — < differ
conflict
- ~~Voting~~^{mechanism} to select the “right” value
 - Simple voting can be problematic – Veracity problem
 - Source Reliability
 - Copy Detection
 - Statistical techniques to estimate
 - Trustworthiness of sources — evaluate whether it's reporting basic or known facts correctly
 - Bias introduced by copies (copy detection)
 - True distribution of values for data items

Source Selection

- **The problem**

- Choose only useful sources
- Adding sources first improves integration accuracy then reduces it

- **The solution**

- Order candidate source based on a measure of “goodness”
- Add sources until the marginal benefit is less than the marginal cost
- Current techniques scale well

