The following is for Fortune 100 organizations
Based on **2010 industry data** verified by

4 analyst organizations

10+ Fortune 100 enterprises

Did you know…

**Information Systems?**
Information Systems in a Fortune 100 company
~10 thousand
100s per business area (e.g., billing)
30% mission critical
30% modified significantly annually

Did you know…

Databases?
Logical databases in a Fortune 100 ...

~10,000

100s per business area (e.g., billing)

~90% relational

100s new / year

Database size ...

20% large 1-10 TB

40% medium 10 GB-1 TB

40% small <10GB
Database growth…
- 2X every 18 mo  OLTP / transactional
- 2X every 12 mo  OLAP / decision support

Logical database complexity ...
- Tables/database: 100-200
- Attributes/table: 50-200
- Stored procedures/database: ~# tables
- Web services/database: < # tables but ↑↑
- Views: ~3X # tables
Did you know…

Database Design?

Database design level …

~100% table (SQL)
~0% Entity Relationship
Did you know…

Relational Database Integration?

Data integration accounts for …
~40% software project costs
Circa 1985
Tables: 2+
Dominant operation: join
Design: schema-based

Circa 2010
Tables: 10s-1000s
Dominant operations
- Extract, Translate, Load (ETL)
- File transfer
- Then … join
Design
- Evidence gathering (often manual)
- 60-75% schema-less
Did you know…

Information Ecosystems?

Information systems that interact to support domain-specific end-to-end processes, e.g., supply chain, billing, marketing, sales, engineering, operations, finance, …
Information Ecosystems involve …

100s-1,000s Information Systems
100s-1,000s Databases

Conclusion

Information systems & databases live in complex, deeply inter-connected worlds
Did you know…

**Database Integration @ Scale?**

Database integration in an **Information Ecosystem** …

- **100s-1,000s** source databases
- **1,000s-millions** function-specific views
Every organization with a database depends critically on...

Data + Data Integration
Boring huh?

Then how about …
Deeply sophisticated analysis [data integration] across ...

28 intelligence databases connected across 17 US intelligence agencies over 5 years totally missed …
Data Integration @ Scale

Data Integration @ Scale
What is Data Integration?

Combine 2+ data items to form a meaningful result

1980’s Data Integration

Account

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Acct#</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill</td>
<td>43 Main</td>
<td>123…</td>
<td></td>
</tr>
<tr>
<td>Sally</td>
<td>12 Peel</td>
<td>489…</td>
<td></td>
</tr>
</tbody>
</table>

Charge

<table>
<thead>
<tr>
<th>Acct#</th>
<th>Charge</th>
<th>Date</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>123…</td>
<td>$2.32</td>
<td>2/1/10</td>
<td></td>
</tr>
<tr>
<td>149</td>
<td>$1.17</td>
<td>1/7/10</td>
<td></td>
</tr>
</tbody>
</table>

TelcoSum (Charge) where Account.Acct# = Charge.Acct#

DI computation

Entity Mapping

Monthly Bill

Name  Address    Acct#  Amount
Bill   43 Main    123… $2.32
Sally  12 Peel    489… $0.00
What does DI mean?

- **Determined by application “semantics”**

- **Financial Close for a Large Enterprise**
  - **What:** Combine data from 100s-1,000s of systems
  - **Meaning:** Defined by financial accounting practices, jurisdictional regulations, industry standards, ...
  - **Speed:** 1-2 days after end of period
  - **Solutions:** e.g., SAP Financial, PeopleSoft Financial, Hyperion
  - **Governance:** laws and regulations
  - **Correctness:**
    - US = Generally Accepted Accounting Principles (US GAAP)
    - GB = Generally Accepted Accounting Practice (UK GAAP)
    - Etc.

- **Data Integration =**
  - Entity mapping
  - Data integration computation

Definitions (cont.)

- **Semantically homogeneous relational data descriptions**
  - **Verified by an authoritative expert**
  - **To describe the same entity (considering set inclusion)**
  - **For which there is one and only one simple, complete entity mapping between the relational schemas for the data descriptions**
  - **The entity mapping can be defined in relational expressions**

- **Semantically heterogeneous relational data descriptions**
  - **Verified by an authoritative expert**
  - **To describe the same entity (considering set inclusion)**
  - **For which there is not a complete entity mapping between the relational schemas for the data descriptions**
  - **Entity mappings cannot be defined in relational expressions**
Deeper Thoughts

- Schemas: sufficient but not necessary
- Simple map: 1:N, ... atomicity
- Degree of semantic heterogeneity

<table>
<thead>
<tr>
<th>Semantic homogeneity</th>
<th>Semantic heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete mapping</td>
<td>Partial mapping</td>
</tr>
<tr>
<td>No mapping</td>
<td></td>
</tr>
</tbody>
</table>

Consequences

- **Semantically homogeneous** relational data descriptions
  - Data independence + updatable => consistent views
  - Provide basis for
    - Correct
    - Efficient
  - Schema level design and integration

- **Semantically heterogeneous** relational data descriptions
  - No guarantee of: Data independence + updatable views
  - Provide no basis for
    - Correctness
    - Efficiency
  - Schema level design and integration not always possible

- **Data modelling and integration best practices**
  - Semantic homogeneity
  - Radical simplification
  - Standardization: ontologies, schemas, application solutions, ...
Semantic Heterogeneity...

- Is unavoidable @ scale
  - Autonomous design
  - Ontologically heterogeneous, e.g., Enterprise Customer
    - Legal entity-based ontology
    - Location-based ontology

- Dominates semantic homogeneity

Information Ecosystem Integration (3D)

Telco Service Provider
- VZW
- Worldcom
- MCI
- BA
- GTE

Enterprise Customer
- Wells Fargo
- Wachovia
- Bank of America
- Merrill Lynch

Business Domains
- Ordering
- Repair
- Marketing
- Sales
- Finance
Correctness & Efficiency

- Legally Binding Requirements
  - Law governing communications, financial transactions, privacy, …
    - E.g.: Customer Proprietary Network Information (CPNI)
  - Regulation: rules defining operational details
    - E.g.: FCC CPNI rules –customer views require manual verification
  - Customer agreements: SLAs for all services
  - Telco agreements: SLAs for all production systems

- Mission Critical: Operational & Information Correctness

- Complex Views
  - Customer Defined Views
  - Inconsistent & Conflicting Views
    - Business perspectives, manually defined, no ultimate truth or agreement
    - Conflicting telephone #: Long-distance vs. local

Data Modelling: Enterprise Customer

- Enterprise Customer data description
  - Enterprise customer identifier (ECID)
  - Enterprise customer data
    - Telco information, e.g., billing, must be mapped (integrated into) Enterprise Customer organizational units

- Enterprise Customer Ontologies = Conceptualizations
  - Legal entity
  - Locations
  - Telco contracts
  - ~40 others
Enterprise A

Legal entity structure (partial)
• ~400 legal entities
• 14 levels deep
• 20,000 accounts (~50/entity)
Location-based Ontology

Radically Simplified Schema

Contract-based Ontology

Radically Simplified Schema
Entity maps for semantically heterogeneous Enterprise Customer Entities are schema-less

**Location-based ECID schema**
- Location
  - Subclass of
  - Building
    - Subclass of
    - Office Building
    - Residential

**Legal entity-based ECID schema**
- Legal entity
  - Subclass of
  - Company
    - Part of
  - Child Company

**Contract-based ECID schema**
- Contract
  - In part of
  - Service or goods
  - Buyer
    - Subclass of
    - Enterprise

Enterprise Customer: 3 Departments in 2 locations
Location-Based Enterprise Customer IDs: 2 locations

Two Entities

Legal entity-based Enterprise Customer IDs: 3 organizations with 1, 2 and 3 departments

Four Entities
Multiple Versions of Truth

Data Integration Observations

- Each view is unique
  - Entity map
  - Data integration computation

- Views often conflict or mutually inconsistent

- Data integration for semantically heterogeneous relations
  - Schema-less
  - By evidence gathering
  - At instance level

- Views are
  - Dominant means of accessing data
  - Lenses into Information Ecosystems
Codd’s Rules Extended

- 0: System must be relational, a database, a management system
- 1: Information rule (tabular)
- 2: Guaranteed access rule (keys)
- 3: Systematic treatment of nulls
- 4: Active online catalog based on the relational model
- 5: Comprehensive data sublanguage rule
- 6: View updating rule
- 7: High-level insert, update, and delete
- 8: Physical data independence
- 9: Logical data independence
- 10: Integrity independence
- 11: Distribution independence
- 12: Nonsubversion rule
- 13: Relational power rule: The full power of relational technology applies to semantically homogeneous relational databases.

For Semantically Heterogeneous Databases

- Cannot be guaranteed
  - Relational model
  - Codd’s rules
  - Relational properties
    - Updateable views
  - Relational assumptions
    - Global schema
    - Single version of truth

- No basis for correctness or efficiency

- Unfortunately, like the real world
Shadow Approach

Database Approach: power and limits
- Data structures
- Schemas
- Data models
- Logic

Shadow Approach
- Meta-tagging
  - Equivalence (E-tags) for entity mapping after resolution
  - What (W-tags) for “meaning”
  - Projection (P-tags) for ETL over logical data structures

(Traditional) Data Integration

Total Semantic Heterogeneity
Shadow Theory-based Data Integration

<table>
<thead>
<tr>
<th>Billing view</th>
<th>Ordering view</th>
<th>Repair view</th>
<th>Sales view</th>
<th>Marketing view</th>
<th>Service view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence 1</td>
<td>Evidence 2</td>
<td>Evidence 3</td>
<td>Evidence 4</td>
<td>Evidence 5</td>
<td>Evidence N</td>
</tr>
</tbody>
</table>

Ontologies

Explicitly represented w/Multiple Versions of Truth.

Need a new language to implement rules by tags.

Business Rules

Enterprise portal

Ontology 1  Ontology 2  Ontology 3  Ontology 4

Schema

DB

Shadows

Metadata

data

Additional slides
Data Integration Pragmatics

- Codd’s Rules violated
  - Un-natural fit
  - Primary keys
    - Problems: technical, managerial
    - Autonomous design
    - Semantic heterogeneity
  - Few views updatable

- Schemas largely unavailable
  - 10% not RDBMSs
  - 30% legal
  - 60% autonomy

- Federation often prohibited
  - Legal: privacy, CPNI, ...

- Single Source of Truth
  - Makes sense in specific cases
  - Usually not!

- Views conflict
  - For good reason!
    - Sales vs. marketing vs. repair
  - Maintain conflicts or lose value

- Best practices rare
  - Semantic homogeneity
  - Radical simplification
  - Standardization

- Data Integration
  - Table-level (Not ER-level)
  - Instance-based
  - Seldom schema-based
  - Manual
  - Evidence gathering

Current Data Integration Technologies

- Relational DBMS
- Extract, Transform, & Load (ETL)
- Federation aka Enterprise information integration (EII)
- DI Platforms: 50+, e.g.,
  - IBM InfoSphere Information Server
  - Informatica Platform
  - Microsoft SQL Server Integration Services
  - Oracle Data Integrator, ...
  - SAP Data Integrator/Federator
- Instance-level integration
  - XML-based integration
  - Custom tools: Verizon
  - Dun & Bradstreet ECID maps

- Limitations
  - Schema-based
  - Federation issues
  - Limited (no!) support for
    - Schema-less integration
    - Instance-level integration
    - Evidence gathering
    - Conflicting views
    - Semantic heterogeneity
  - 50+ ad hoc tools
  - Expensive to migrate
  - Information Ecosystem onto data integration platform
Emerging Data Integration Technologies

- **DBMS Architecture**
  - In-Database Analytics
  - Database Appliances
- **Distributed DBMS Architecture**
  - Change Data Capture (CDC)
  - Data Access
  - Data Replication
  - Distributed Data Caching
- **Service Orientation**
  - Data Services platforms
  - Information as a Service
  - Enterprise Service Buses (ESBs)
- **Master Data Management (MDM)**

- **Data Discovery**
- **Data Quality Services**
- **Instance-level integration**
  - XML-based integration
  - Custom tools
  - Dun & Bradstreet ECID maps
- **Schema-less Tools**
  - Evidence gathering: Web, …
  - Entity/Identity Resolution
  - Social Networking:
    - Collaboration
    - Web 2.0
    - Social Search
- **Data Mining**
- **Schema-less Tools**
  - Data Mining
  - Knowledge Ontologies / Taxonomies
  - Classification & Analysis
  - Semantic Technologies
  - Content Analytics
- **Shadow Theory**
  - Plato’s Cave
  - Monadic 2nd Order Logic

**Heterogeneous meanings of a telephone number**

- **Telephone** attribute in customer database.
- **Telephone** entity = a physical phone.
- Owner identified uniquely with additional attribute NPA+NXX+XXX+CustomerID.
- Ported to cell or internet phone
- Domestic long distance services, charges determined by NPA.
- International long distance services, charges determined by country code.

- North America Numbering Plan (NANP) telephone number
  - (NPA) NXX-XXX
  - NPA - Number Plan Area
  - NXX - Central office code
  - XXX - telephone line #
## Limitations of Current Solutions

<table>
<thead>
<tr>
<th>Limitations of Current Solutions</th>
<th>Limitations of Current Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Model Principles Don’t Apply</td>
<td>Tools Don’t Meet Requirements</td>
</tr>
<tr>
<td>• Unique IDs (keys) are limiting</td>
<td>• No single model of data</td>
</tr>
<tr>
<td>• Modelling Choices: entities,</td>
<td>• Inconsistent / conflicting views</td>
</tr>
<tr>
<td>properties, or relationships</td>
<td></td>
</tr>
<tr>
<td>Schemas Are Seldom Available</td>
<td>Meta-data and Multi-database Scale</td>
</tr>
<tr>
<td>• 10% not RDBMSs</td>
<td>• information ecosystem (vs. database) management</td>
</tr>
<tr>
<td>• 30% legal</td>
<td>• Millions of views</td>
</tr>
<tr>
<td>• 60% owners maintain autonomy</td>
<td>• 10,000s of cross database dependencies</td>
</tr>
<tr>
<td>• Biz: impact operations</td>
<td></td>
</tr>
<tr>
<td>• Tech: impact schedule &amp; cost</td>
<td></td>
</tr>
<tr>
<td>Schema-based Integration Challenges</td>
<td>Migrating Information Ecosystems to Vendor Solutions is Prohibitive</td>
</tr>
<tr>
<td>• Inconsistent data models</td>
<td></td>
</tr>
<tr>
<td>• View Conflicts</td>
<td></td>
</tr>
<tr>
<td>• Cascaded Views</td>
<td></td>
</tr>
<tr>
<td>Federation Seldom Allowed (CPNI)</td>
<td></td>
</tr>
<tr>
<td>No Single Source of Truth</td>
<td></td>
</tr>
</tbody>
</table>

- **Pragmatics**
  - Bootstrap in a legacy environment

- **Conceptual Modelling Support**
  - Different models required for different workers
  - No round-trip engineering