How Do You Monitor Data Quality and Why Should You (or Data Quality – What Actuaries Need to Know)

Mark S. Allaben, FCAS, MAAA

VP and Actuary

Information Delivery Services

CAS Annual Meeting November 2012

Anti-Trust Notification

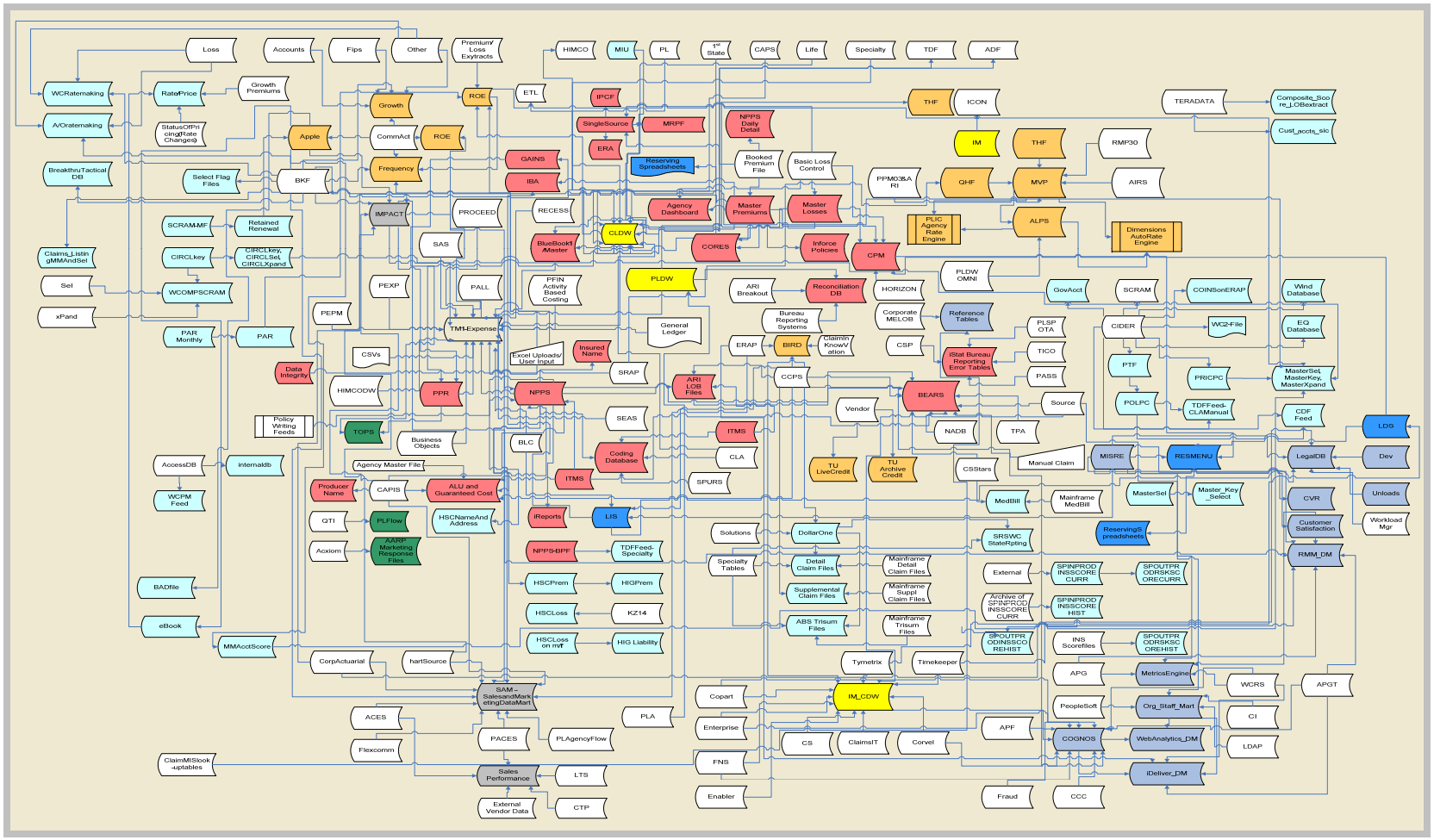
- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding-expressed or implied-that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy

Presentation Structure

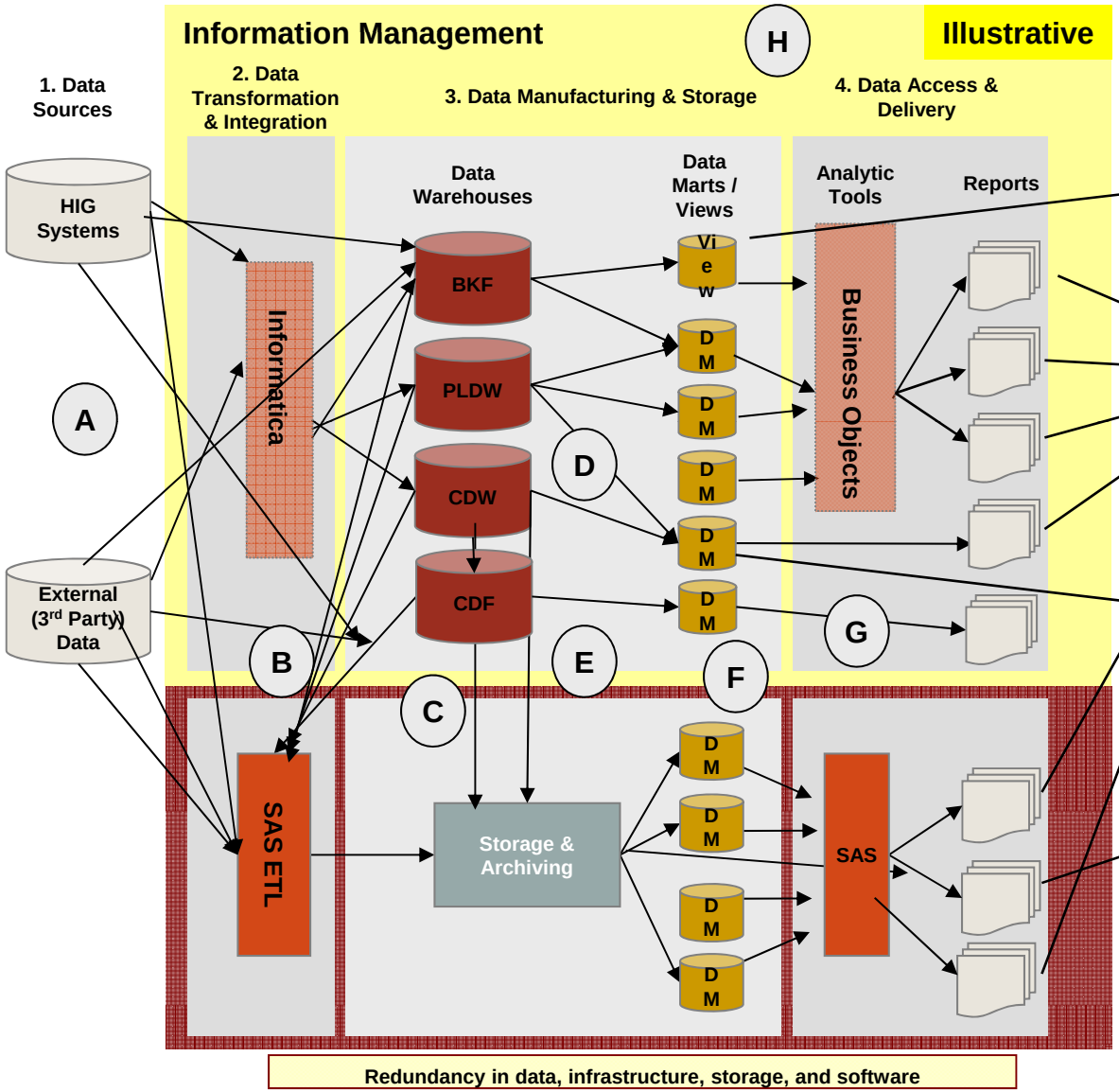
- **Foundational Concepts**
  - The Problem
  - Business Requirements for Data
  - Information Delivery Architecture
  - Metadata
- **Data Quality Program in Practice**
- **Appendix (Extra stuff)**

Information Architecture – The Problem

Typical Multi-line Insurer Current Data Architecture



Data Warehouse Environment Example of Issues



Key Observations

- **A.** Multiple Sources of Data
- **B.** Multiple Transformation & Integration (ETL) Tools
- **C.** Redundant storage of data
- **D.** Uncontrolled Access to data
- **E.** All data stored on the same tier / type of storage
- **F.** Data marts not always 'in sync' with data sources
- **G.** Multiple BI Tools
- **H.** No 'Single version of the truth' – No systemic reconciliation back to source systems

Metadata: Current State vs. Possible Future State Scenario

Current State:

Information Chaos

- **Multiple definitions for the same data element**
- **Multi-use data fields**
- **Excessive time & resources required to search for needed data**
- **Pockets of excellence**
- **Lack of enterprise data governance and stewardship**
- **One shot mapping efforts**
- **Not shared or reusable**
- **Use of incorrect sources**
- **Data redundancy**

Future State:

Data Quality Management

- **Agreed upon enterprise definitions**
- **Single-use data fields**
- **Increased efficiencies via short data searches**
- **Enterprise organizational effectiveness**
- **Centrally captured / reduced redundancy**
- **Shared and reusable**
- **Authoritative & certified sources**
- **Unlimited potential for creative use of data**
- **Provides competitive advantage**
- **Trusted data**
- **Provable, repeatable processes / results**

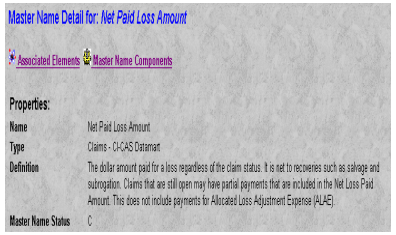
Future State Process Flow

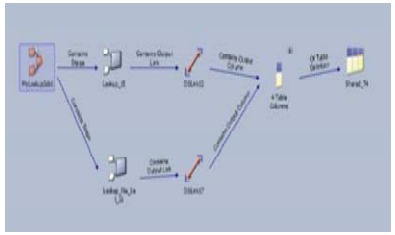
1. Analyst types the term "**Paid Loss Amount**" into the P&C Metadata Search System

2. He/she is quickly presented with a list of **exact name matches** and **synonyms**

3. He/she is determines "**Net Paid Loss Amount**" is the right field to use, it is an **"approved source"** and who the **Steward** is.

4. He/she is able to conduct an **impact analysis** and determine the **data lineage, where it was created, and the rules used to calculate it.**





Data Requirements

Solving for five data requirements is critical to the success of any initiative

Data Requirements	Description
Scalability	▪ Increased usage and appetite for additional data elements from other parts of the enterprise and from 3rd party sources will initiate a virtuous circle - increased use of data will lead to more **sophisticated questions** which **will lead to the need for more data** to make decisions, complete transactions, and conduct research. Increased capacity in people, process, and technology will enable capture of additional data **at decreasing marginal costs.** Scalability enables a shift from being extremely parsimonious in our data capture to capturing all potentially useful data
Trustworthy	▪ Knowledge of what data exists, where it is located, and confidence that the **quality level is sufficient for** conducting **analysis and making decisions**
Accessibility	▪ Easier and speedier access to existing data. All 2010 workstreams assume that data, 3rd party and internal, will be **available wherever and whenever needed** in the future processes
Granularity	▪ Data acquired by the customer interaction processes (New Business, Claims, etc.) and 3rd party providers are **detailed enough to meet research and transactional needs** of product, marketing, sales, and pricing
Connectivity	▪ Ability to **link data across** the **enterprise and** from **3rd parties** at a granular vs. summary level, to enable research, analysis and transactional processing

Achieving the five data requirements will make data available and useable across the enterprise.

Data Access & Delivery

Business Intelligence (BI)

An umbrella term that encompasses the processes, tools, and technologies required to turn data into information, and information into knowledge and plans that drive effective business activity. BI encompasses data warehousing technologies and processes on the back end, and query, reporting, analysis, and information delivery tools (that is, BI tools) and processes on the front end

Potential Issues

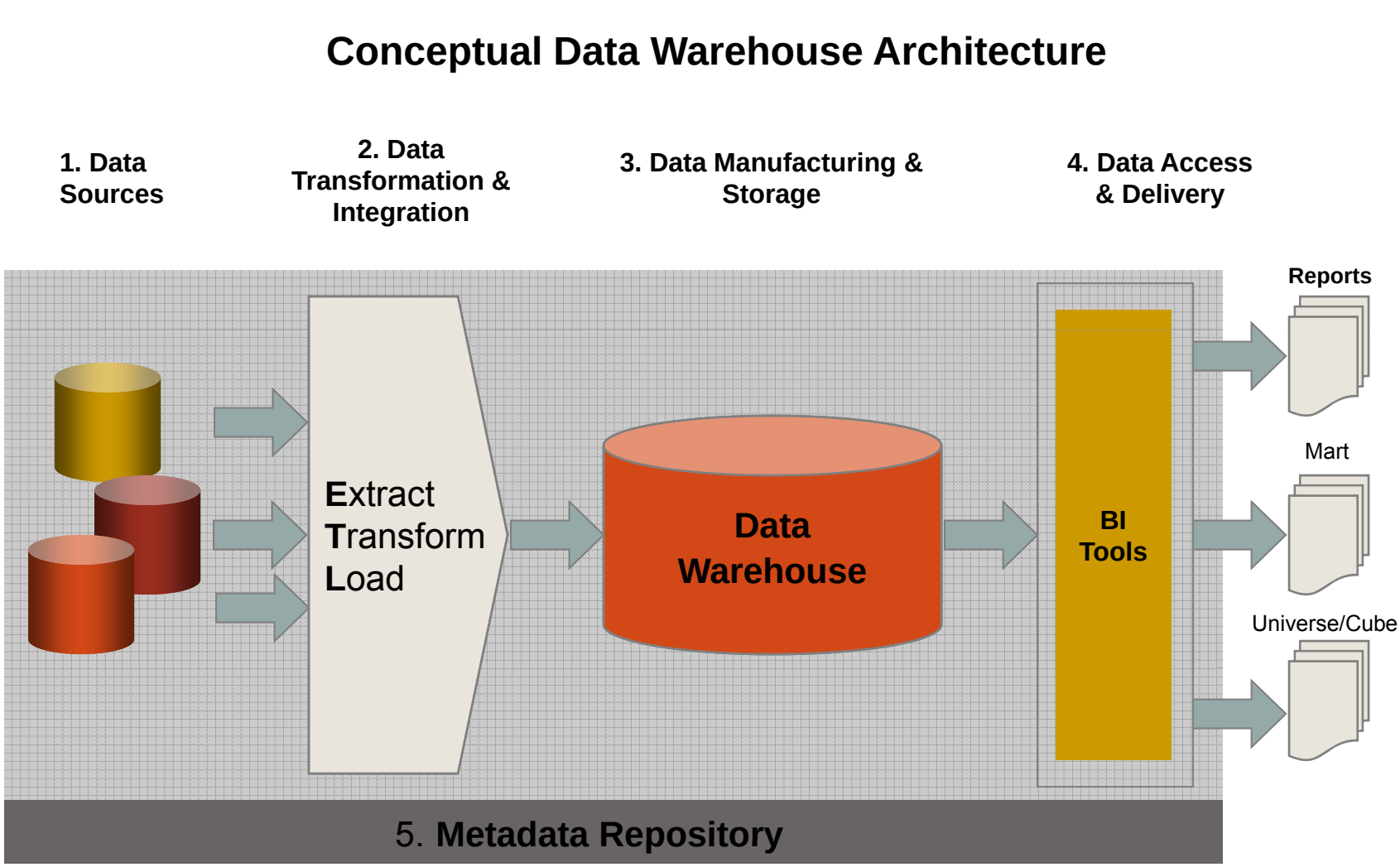
Multiple BI Tools

- Five Business Intelligence tools are in use
- Reports and Analytics cannot be easily reused
  - Dueling "Truths"
  - Reconciliation Efforts

	Purpose	Usage
Standard Reports	**Provides a pre-made document to provide information needed by user**	**Reports that require infrequent structural changes, and can be easily accessed electronically**
Queries	**Provides ability to data using a pre-defined query, or on an ad hoc basis**	**Research, analysis and reporting**
Analytical Applications	**Provides ability to easily access key performance indicators or metrics**	**Monitoring and accessing performance**
OLAP Analysis	**Alerts users to pre-defined conditions that occur**	**Research and Analysis**
Exception Based Reporting	**Provides ability to perform summary, detailed or trend analysis on requested data.**	**Notification without the need to perform detailed analysis**
Data Mining	**Ability to discover hidden trends with the data**	**Research and analysis of hidden trends with in the data**

Five Elements of Data Quality Management

Conceptual Data Warehouse Architecture



Data Transformation & Integration (ETL)

ETL (Extract, Transform and Load) is a common 3 step process designed for this purpose

- Extract data from multiple legacy sources
- Extract may be via
  - Intermediate files
  - Databases
  - Directly connecting to sources
- Multiple extract types
  - Full extract (refresh)
  - Incremental extract

- Works with the extracted data set
- Applies business rules to convert to desired state
- Cleanse and standardize data

- Inserts / updates the data warehouse database tables
- Intelligently add new data to the system

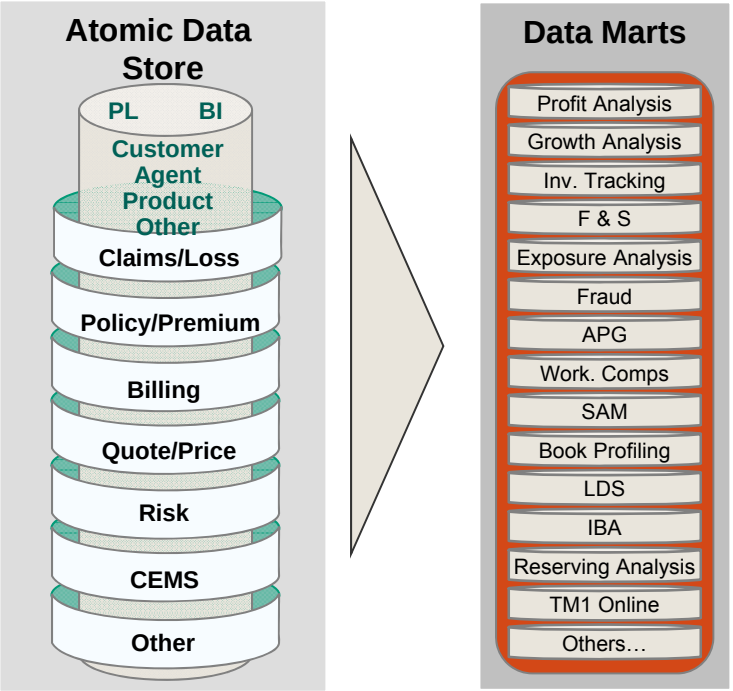
Data Manufacturing & Storage

Atomic Data Store

A shared, analytic data structure that supports multiple subjects, applications, or departments

Data Mart

A shared, analytic data structure that generally supports a single subject area, application, or department



Data Warehouse Architecture

There are different types of data warehouses and platforms, e.g.:

- centralized vs. federated
- Superdome v. Teradata v. Exadata

Potential Issues

Redundant Storage of Data

Uncontrolled Access to Data

All data stored on the same tier / type of storage

Data marts not always in-sync with data sources

Metadata

Metadata can provide a semantic layer between IT systems and business users—essentially translating the systems' technical terminology into business terms—making the system easier to use and understand, and helping users make sound business decisions based on the data (i.e. A Data Yellow Pages)

A *metadata repository* is: the logical place to uniformly retain and manage corporate knowledge (meta data) within or across different organizations in a company

Various types of meta data include:

Data Definitions

- List of common data elements and standard definitions

Business Rules

- Rules define data use, manipulation, transformation, calculation and summarization
- Business rules are mainly implemented by the ETL and reporting tools in a metadata dictionary

Data Standards

- Rules and processes on data quality

Data context

- Use of and dependencies on data within business units and processes

Technical Metadata

- Information on configuration and use of tools and programs

Operational metadata

- Information on change/update activity, archiving, backup, usage statistics

Potential Issues

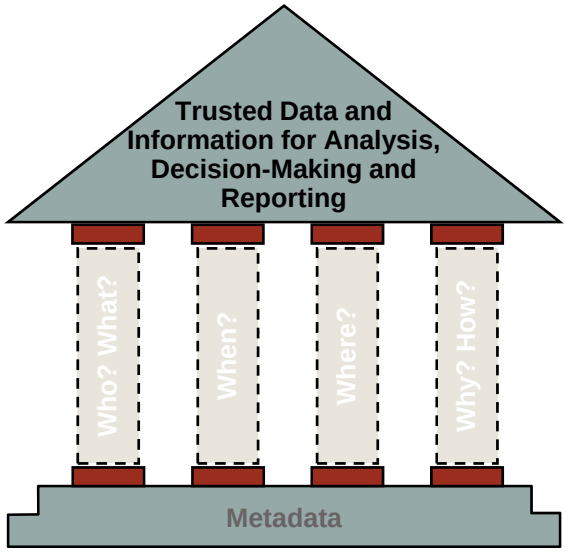
No Single Version of the Truth –
No systemic reconciliation back to source system

- Metadata is the crux of many of our data problems
  - Time would note be wasted
    - Less reconciliation
  - Not gathering useless / redundant data
    - Less storage

Metadata - What is Metadata?

Metadata is 'data about data'. **It tells us the meaning and context of a piece of data.**

- ***Who?***
  - Who owns this data?
  - Who's responsible for its quality?
  - Who has access to it?
- ***What?***
  - What's the definition of this data element?
  - What are the valid values?
- ***When?***
  - When was it last updated?
- ***Where?***
  - Where is this data stored?
  - Where does it originate from?
  - Where is it used?
- ***Why?***
  - Why is this piece of data important?
- ***How?***
  - How is it calculated?
  - How is it manipulated?



Example of Metadata:

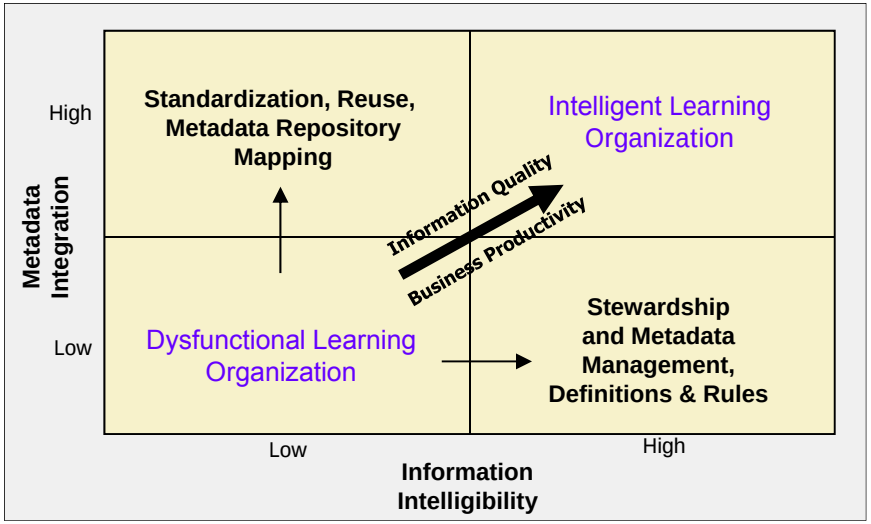
- **What does "Total Earned" mean?**
- **What is the definition and who is accountable?**
- **How is "Total Earned" formulated?**
- **Where does this data originate from?**
- **What software, hardware, and databases are involved?**

Often metadata is agreed-upon **definitions and business rules** stored in a centralized repository so that common terminology for business terms is used for all business users – even those across departments and systems. It can include information about **data's ownership, source system, derivation (e.g. profit = revenues minus costs), or usage rules**. It prevents data misinterpretation and poor decision making due to sketchy understanding of the true meaning and use of corporate data.

Metadata - What are the benefits of implementing a Metadata Strategy?

Benefits

- Common, embraced language between Business and IT
- Substantial opportunity to improve data quality through greater understanding of HIG data
- Improved business intelligence
- Reduced redundancy
- Consistency of data elements
- Reduced reconciliation efforts around data definition
- Alleviate loss of knowledge when staff transfers, retires or leaves the company
- Minimize the effort on learning new data sources
- Reduced development cycle times for new and existing systems
- Economies of scale
- Increased efficiencies via short data searches
- Improved efficiency of analysis



Imagine sending all of your most experienced employees away for a month.

- ***What would happen to your business?***
- ***Where would your employees go to get answers?***
- ***How long would it take and how many resources would have to be involved?***

The costs would be mitigated if you had a centralized metadata repository.

Data Quality in the Practice

- **Data Quality Program - Aspects**
  - Data Governance
  - Data Stewardship
  - Third Party Data Governance
  - Integration with Enterprise Risk Management
- **Solvency II - A Question of Translation**
- **Day Care, Feeding Babies and Data Quality**

15

Current Practice Complies with Solvency II

- **Data Governance Council**
  - Data Champion
  - Business Owned Data
  - Data Standards
  - Variance Requests (Exceptions)
  - Roles and Responsibilities
  - Enterprise Executive Representation
  - Staff Resources for the Data Governance Council

Current Practice Complies with Solvency II

- **Data Stewardship**
  - Data Stewardship Steering Committee
  - Data Naming Standards
    - Common (Certified) Definitions
  - Data Stewards
  - Variance Requests (Exceptions)
  - Roles and Responsibilities
  - Enterprise Executive Sponsorship

Current Practice Complies with Solvency II

Data Quality and Data Stewardship

- High Impact Data Elements
- Data Profiling Tools
- Scorecards
  - Validity
  - Completeness
  - Accuracy
  - Defaults
  - Defects allowed into production
- Remediation
- Actuarial Standard of Practice 23 – Data Quality

Current Practice Complies with Solvency II

Data Risk Management

- ERM - Operational Risk
- Audit Committee of the Board of Directors
- Internal Audit Annual Assessment
- Sarbanes-Oxley Compliance
- SEC Compliance

Current Practice Complies with Solvency II

Third Party Data (TPD)

- Third Party Data Governance Committee
  - Data
  - Enterprise-wide Licensing
  - SLA's with TPD Providers
  - Audit
- IT Creates Internal Service that can switch providers
- TPD Process Integrated with Procurement and Legal

Current Practice Complies with Solvency II

Documentation

- Metadata Repository
  - Data Definitions
  - Lineage
  - Data Mapping
  - Business and IT Metadata
- Software Development Cycle
- Architectural Review Committee
- Change Controls to Invest Projects

Current Practice Complies with Solvency II

Documentation

- Metadata Repository
  - Data Definitions
  - Lineage
  - Data Mapping
  - Business and IT Metadata
- Software Development Cycle
- Architectural Review Committee
- Change Controls to Invest Projects

Current Practice Complies with Solvency II

Translation from Practice to Solvency II

- **British English vs American English**

  – **Solvency II** – Accurate, Complete, Appropriate, Timely, Accessible, Comparable

  VS

  – **Company Practice** – Scalable, Trusted, Accessible, Granular, Connected

Day Care, Feeding of Babies and Data Quality

- **Story to Share with IT and Management**
  - Do we really need every data element right?
  - Is 95% right good enough?
  - Let's Pretend:
    - You run a day care center with 20 babies
    - You outsourced your formula preparation for the feeding of the babies
    - The outsourcing partner says, bad news, 1 out of the 20 bottles has rat poison in the formula
- **Do you Feed the babies?**

Appendix

Terms – Common Language

- Business Intelligence Tools
- Data Governance
- Data Warehouse
- Dimensional Data
- Master Data Management
- Metadata
- Metadata Repository
- Relational Data
- Staging

Glossary: Common Data Warehousing Terms & Definitions

1. Data Sources

- Source System: **Source System or Data Sources refers to any electronic repository of information that contains data of interest for management use or analytics**

2. Data Transformation & Integration (ETL)

- ETL: **The data transformation layer (aka Extract, transform, load - ETL or some variant) is the subsystem concerned with extraction of data from the data sources (source systems), transformation from the source format and structure into the target (data warehouse) format and structure, and loading into the data warehouse**

5. Metadata Management

- Metadata:
  - Metadata, or "data about data", is used not only to inform operators and users of the data warehouse about its status and the information held within the data warehouse, but also as a means of integration of incoming data and a tool to update and refine the underlying DW model.
  - Examples of data warehouse metadata include table and column names, their detailed descriptions, their connection to business meaningful names, the most recent data load date, the business meaning of a data item and the number of users that are logged in currently

Glossary: Common Data Warehousing Terms & Definitions

3. Data Manufacturing & Storage

- Data Warehouse: **A shared, analytic data structure that supports** multiple **subjects, applications, or departments. There are three types of data warehouses: centralized, hub-and-spoke, and operational data stores**
- Hub-and-Spoke Data Warehouse: **A data warehouse that stages and prepares data for delivery to downstream (i.e., dependent) data marts. Most users query the dependent data marts, not the data warehouse**
- Centralized Data Warehouse: **A data warehouse residing within a single database, which users query directly**
- Federated Marts or Environments: **An architecture that leaves existing analytic structures in place, but links them to some degree using shared keys, shared columns, global metadata, distributed queries, or some other method**
- Data Mart: **A shared, analytic data structure that generally supports a** single **subject area, application, or department. A data mart is commonly a cluster of star schemas supporting a single subject area**
- Dependent Data Mart: **A dependent data mart is a physical database (either on the same hardware as the data warehouse or on a separate hardware platform) that receives all its information from the data warehouse. The purpose of a Data Mart is to provide a sub-set of the data warehouse's data for a specific purpose or to a specific sub-group of the organization. A data mart is exactly like a data warehouse technically, but it serves a different business purpose: it either holds information for only part of a company (such as a division), or it holds a small selection of information for the entire company (to support extra analysis without slowing down the main system). In either case, however, it is not the organization's official repository, the way a data warehouse is**
- View: **Is a 'logical' provisioning of a subset of the data warehouse similar to a Data Mart**
- Tiered Storage: **Data is stored according to its intended use. For instance, data intended for restoration in the event of data loss or corruption is stored locally, for fast recovery. Data required to be kept for regulatory purposes is archived to lower cost disks**
- Operational Data Store (ODS): **A "data warehouse" with limited historical data (e.g. 30 to 60 days of information) that supports one or more operational applications with sub-second response time requirements. An ODS is also updated directly by operational applications**

Glossary: Common Data Warehousing Terms & Definitions

4. Data Access & Delivery

- Business Intelligence (BI): **is an umbrella term that encompasses the processes, tools, and technologies required to turn data into information, and information into knowledge and plans that drive effective business activity. BI encompasses data warehousing technologies and processes on the back end, and query, reporting, analysis, and information delivery tools (that is, BI tools) and processes on the front end**
- Business Intelligence Tools:
  - Business intelligence tools are a type of application software designed to help the business intelligence (BI) business processes. Specifically they are generally tools that aid in the analysis, and presentation of data. While some business intelligence tools include ETL functionality, ETL tools are generally not considered business intelligence tools
- Reporting:
  - The data in the data warehouse must be available to the organization's staff if the data warehouse is to be useful. There are a very large number of software applications that perform this function, or reporting can be custom-developed. Examples of types of reporting tools include:
    - Business intelligence tools: These are software applications that simplify the process of development and production of business reports based on data warehouse data
    - Executive information systems (known more widely as Dashboard (business)): These are software applications that are used to display complex business metrics and information in a graphical way to allow rapid understanding.
    - OLAP Tools: OLAP tools form data into logical multi-dimensional structures and allow users to select which dimensions to view data by.
    - Data Mining: Data mining tools are software that allow users to perform detailed mathematical and statistical calculations on detailed data warehouse data to detect trends, identify patterns and analyze data
- OLAP:
  - OLAP is an acronym for On Line Analytical Processing. It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. It is part of the broader category business intelligence, which also includes Extract transform load (ETL), relational reporting and data mining. The typical applications of OLAP are in business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas
- Spreadmart: **A spreadsheet or desktop database that functions as a personal or departmental data mart whose definitions and rules are not consistent with other analytic structures**

Metadata - Scope

A ***Metadata Management*** program enables our ability to find, understand, manage, govern, rationalize, share, reuse, and leverage information about data, business, applications, services, hardware and software.

Business Metadata

- Business name
- Business definition
- Standard abbreviations
- Valid values
- Formulas and calculations
- Derivation logic
- Business/Logical models

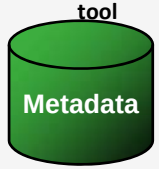
Technical Metadata

- Physical data models
- Application systems
- Program code
- File/Record layouts
- Tables/Fields/Rows
- ETL Transformation logic

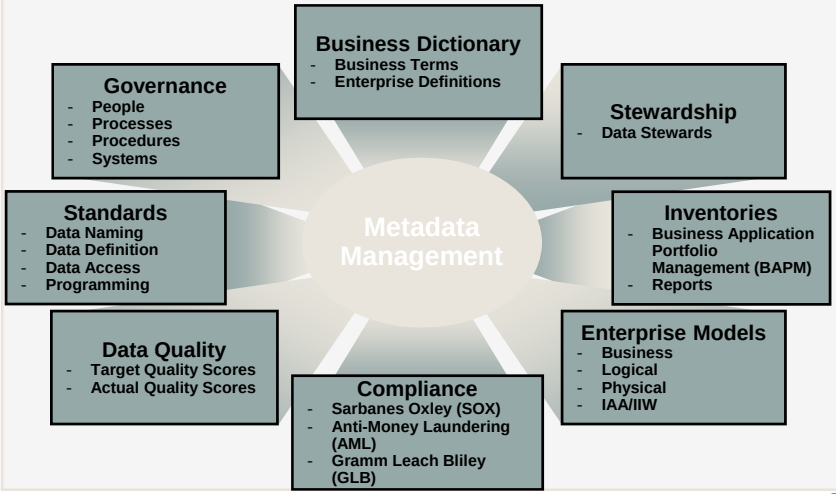
Operational Metadata

- Change logs
- Duration of load
- Number of rows added or changed
- Date entered
- Load dates
- Date last modified

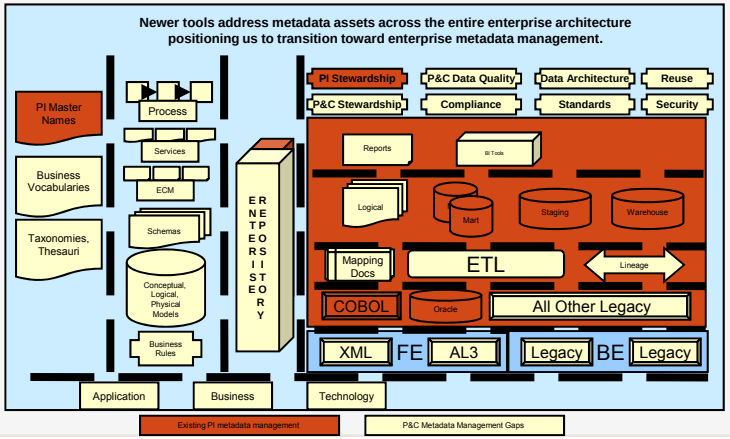
1 Centralized metadata repository / tool



2 Holistic metadata management approach



3 An expanded IT landscape which requires additional process and tool capabilities.



Metadata Implementation Program - The Five Deliverables

1. **Tool:** Acquire a metadata tool that will meet our business and IT requirements for Metadata Management
2. **Governance**: Implement the proper roles, responsibilities, policies, processes, procedures, and standards to most effectively manage our information assets
3. **Organization**: Consolidate various data management resources into a data asset management organization
4. **Communication Plan**: Establish an ongoing effort to educate and communicate to our employees all metadata strategy related initiatives
5. **Roadmap/Implementation:** Develop a preliminary roadmap with key implementation strategies for moving forward