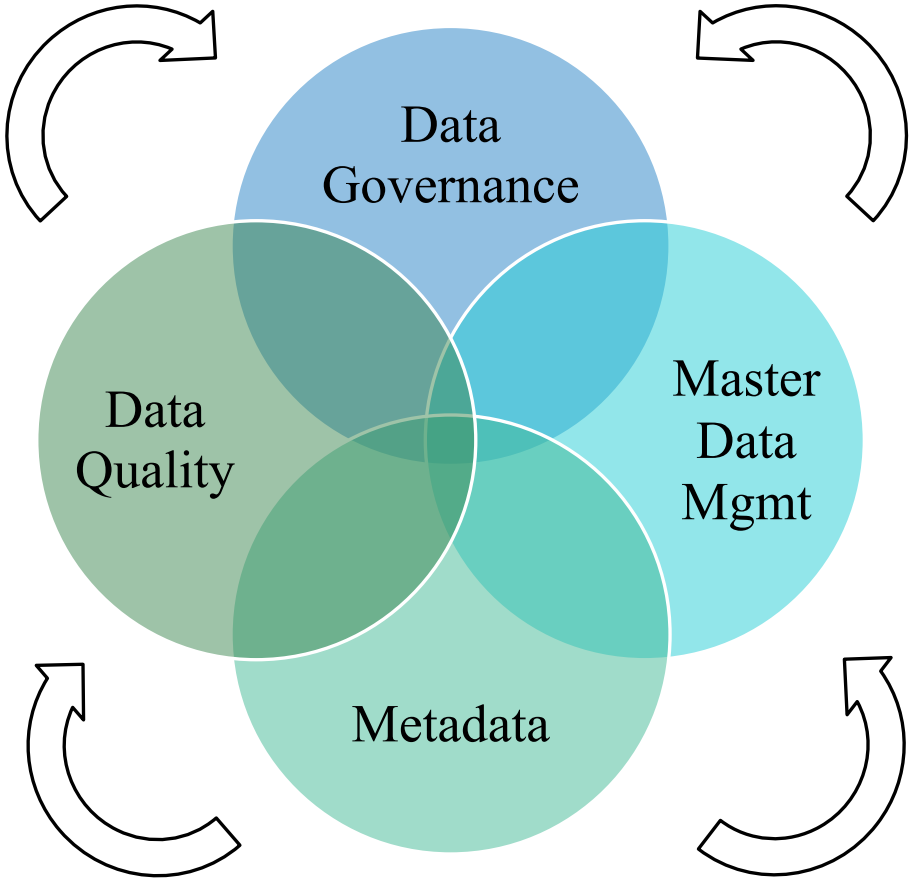
Introduction

School of Information Studies
Syracuse University

Agenda:

- Understand the importance of data governance when it comes to a data warehousing initiative
- Explain master data management
- Define data quality, and discuss its importance in the data warehouse
- Explain the role of metadata in the data warehouse
- Discuss ways to secure data in the data warehouse

They Are All Related



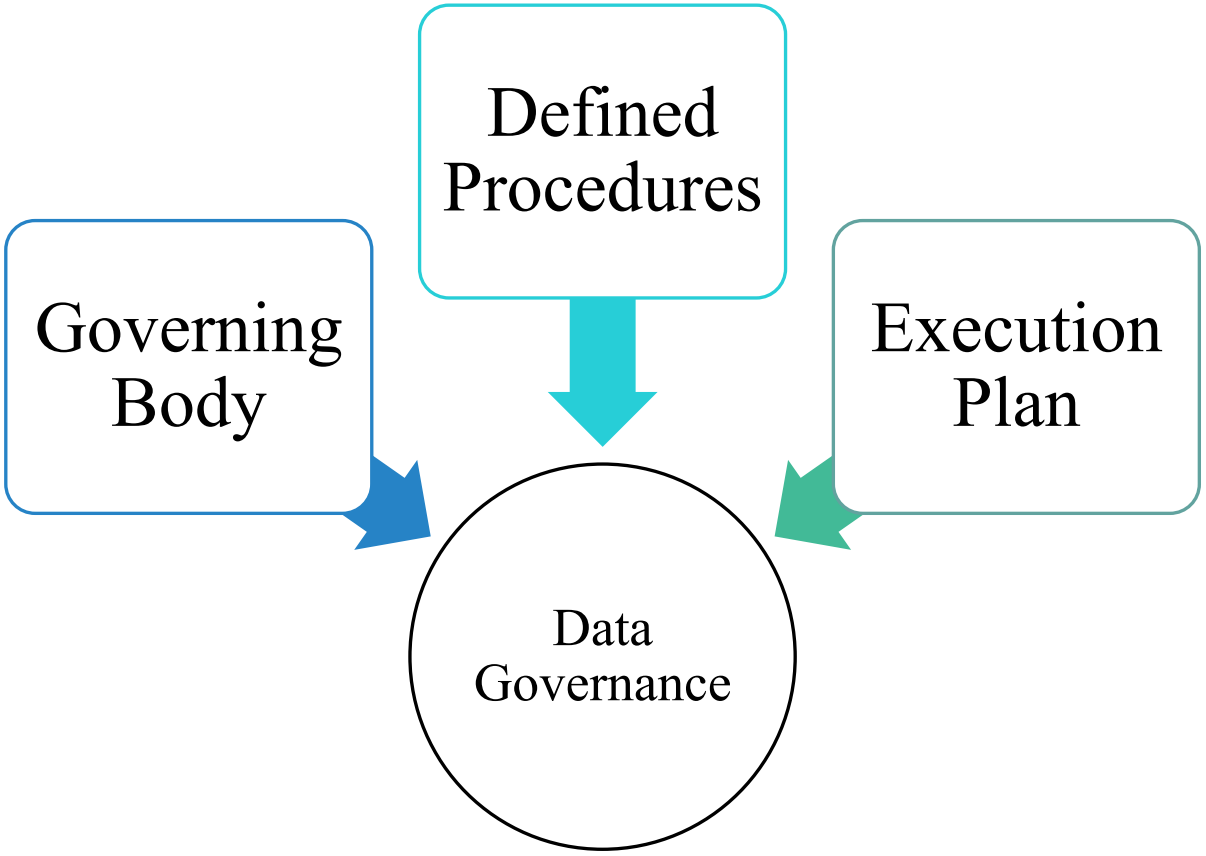
What Is Data Governance?

School of Information Studies
Syracuse University

What Is Data Governance?

- Data governance refers to the overall management of an organization's data. It addresses data
  - Availability,
  - Usability,
  - Integrity, and
  - Security.

Doing Data Governance



Why Do It?

An organization's most important asset is its data!

- Protect the data, and maintain regulatory compliance.
- Devise a plan to share data to improve decision-making or improve customer relations.
- Maximize actionable insights from the data you have.
- Ensure data quality.

Data Governance in Practice

- Be Noninvasive
  - It's not an us-vs.-them issue its for the benefit of the entire organization.
  - No data silos or data hoarding..
- Goals
  - Be as open and transparent as possible in your policies and procedures.
  - Support other groups who require use of your data.
  - Collaborate institutionally to break down information silos. It's *everyone's* data.

Why Do We Need Data Governance?

School of Information Studies
Syracuse University

Example: Why We Need Data Governance

Cust ID	City	State	Zip	Phone
1	Buffalo	NY	13244	315-443-1212
2	Syracuse	NY	13244	443-0092

- Does Customer 1 live in Buffalo?
  - If we assume the Zip is correct, it's Syracuse, New York.
- What is the area code of Customer 2's phone number?
  - If we assume the Zip is correct, it's 315.
- We don't want to assume! As an organization we need consensus as to what to do!

Typical Data Governance Questions

- Which attributes from each source are the authority?
  - Customer data contains e-mails from three systems—which one do we use?
- How should the data be cleaned?
  - If the Zip code and city/state don't match, how do we remedy this?
- What are the rules we use to establish hierarchies/relationships among our data?
  - We can categorize products by vendor/supplier, but is there a more fruitful/valuable way to do this?
- What formula/criteria should be used for weighing factors?
  - We want to charge back IT labor to departments. How is this accomplished on projects impacting multiple departments?

Roles in Data Governance

School of Information Studies
Syracuse University

Typical Roles in DG

- CIO or CDO
- Risk manager
- Data administrator/data steward
- Data custodian

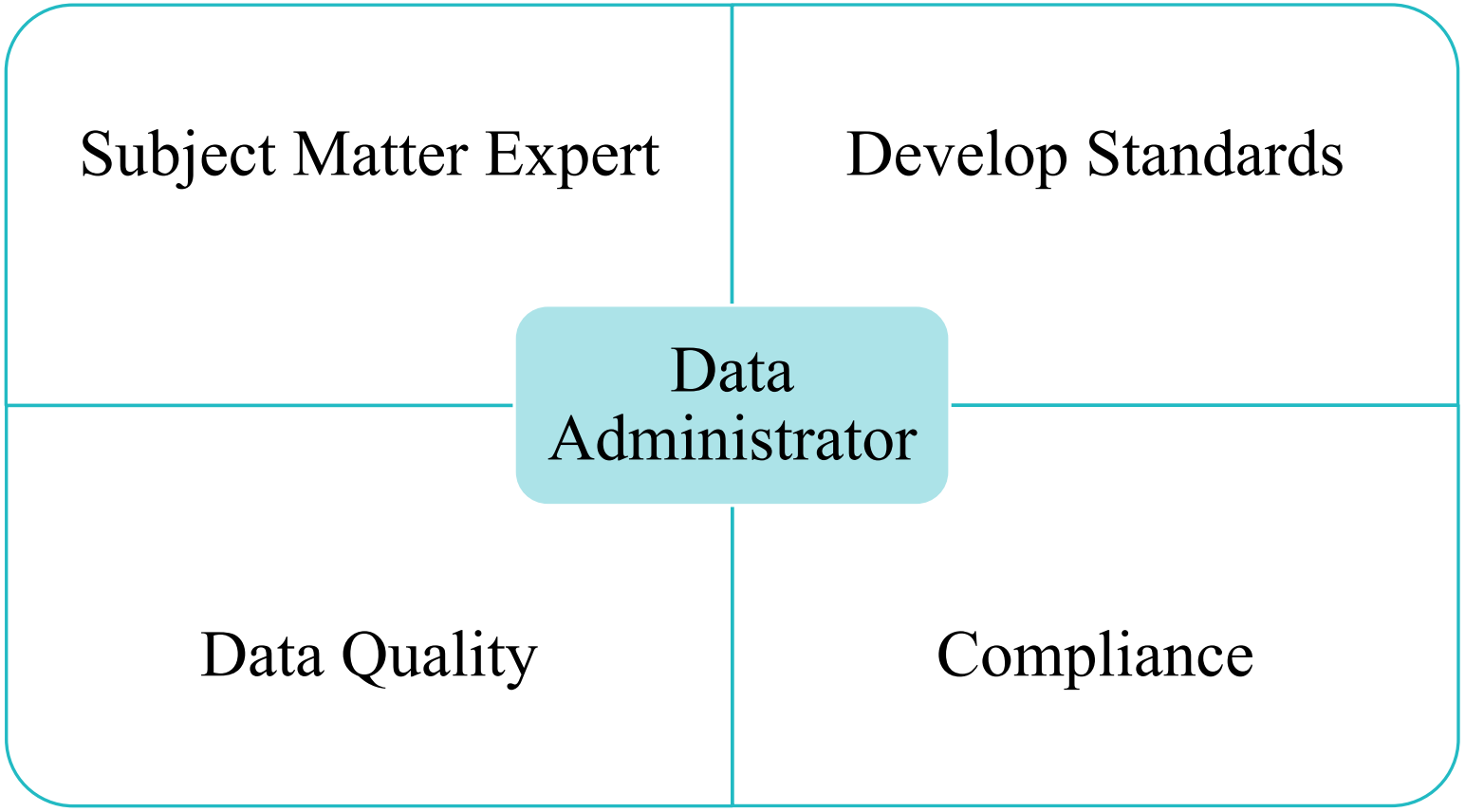
Data Governance Leadership

- Executive-level team
- Sets goals for the organization and expectation of culture around DG
- Assures cooperation throughout the organization
- Risk management
- Approval of policies
- Appointment of other roles

Data Administrator

- Also known as a **data steward**; has administrative control over an information asset and is ultimately accountable for it.
- This person is usually the organization's business expert in the domain (sales, inventory, financials, customer records, etc.).
- It is not a job title! It defines a relationship/responsibility to data.

Responsibilities of the DA



Data Administration Committees

- It is typical of information assets to span multiple functional domains.
  - Examples: customers or students in a university
- In these cases a committee of experts performs data administration.
  - Example: customer data administration team might include representatives from marketing, sales, and billing.

Data Custodian

- Responsible for the technical aspects of the data.
  - Controlling access
  - Maintaining documentation
  - Auditing data quality rules
- Works with the data administrator.
- Liaison to the IT team.
- Again, it's not a job title, it's a responsibility.

What Is Master Data Management?

School of Information Studies
Syracuse University

What Is Master Data Management?

- Creating a single “reference copy” of key business entities.
- Examples: customers, vendors, products, employees.
- Offers an organization a single version of what “customer” means.
- Helps reduce inconsistencies in data distributed throughout the enterprise.
- Aims to provide clean, reliable data.
- MDM systems can provide automated rules and utilities to maintain “golden records” for business entities.
- "Single version of the truth."

Master Data Are Entities

...Not Business Rules or Relationships!

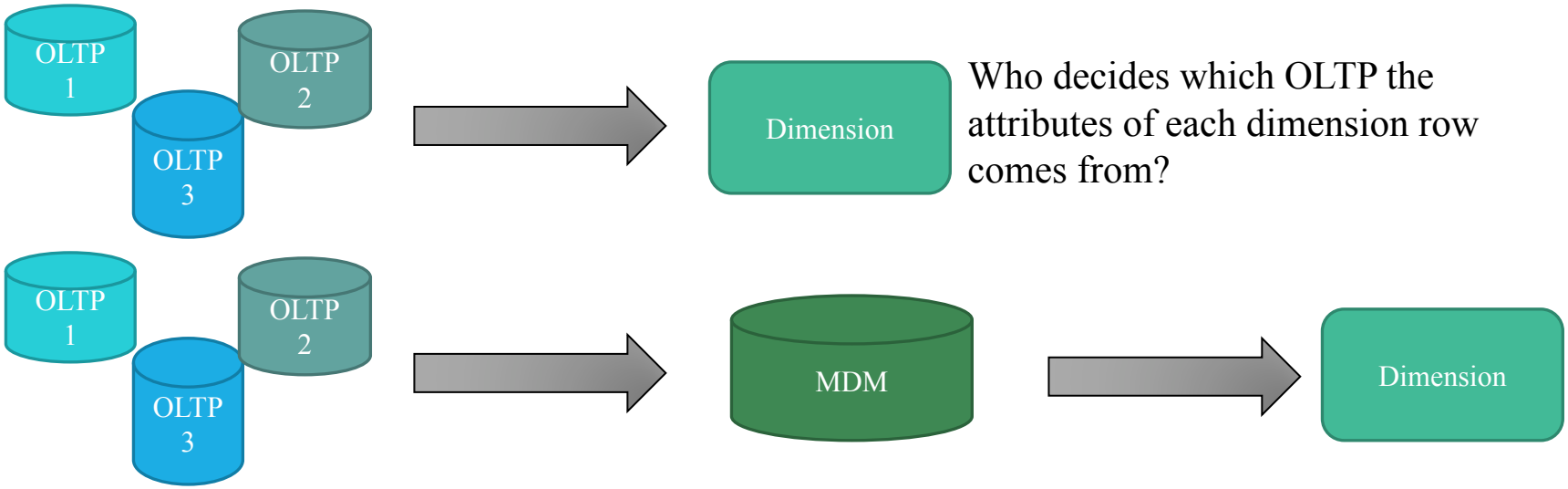
Master Data	Not Master Data
Customers	Orders/Sales
Employees	Payroll
Products	Subscriptions
Courses	Applications
Students	Enrollments

Why Do We Need MDM?

School of Information Studies
Syracuse University

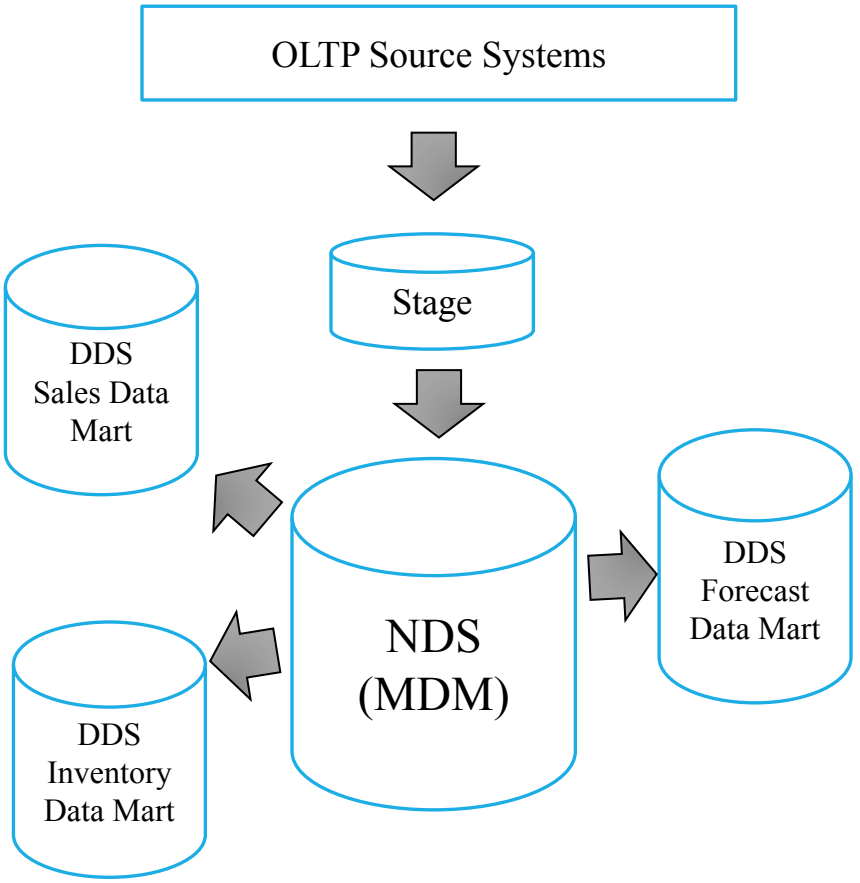
How Does This Relate to Data Warehousing?

MDM initiatives will help us create conformed dimensions in our data warehouse, specifically when the dimensional data is sourced from multiple systems.



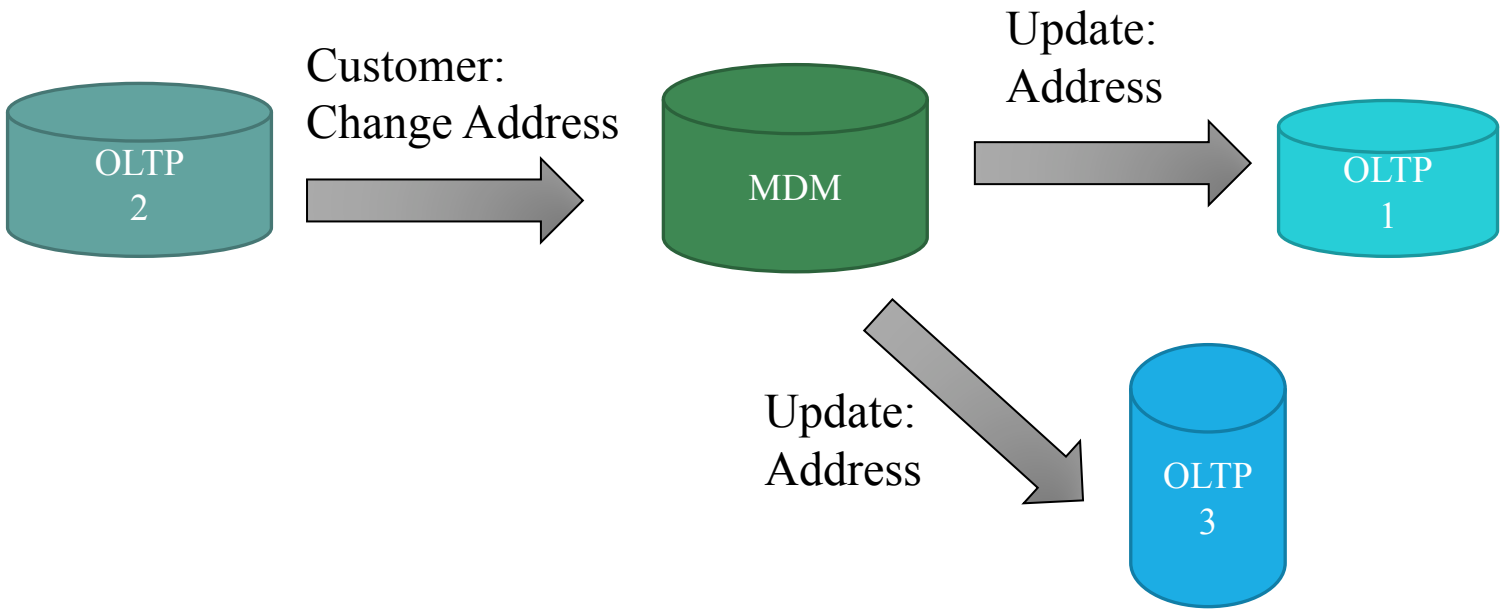
The Data-Centric DW Is Made for MDM

- The Inmon technical architecture of:
  - Hub and spoke
  - Normalized data stores (NDS)
  - data-centric approach
- Makes the process of MDM much easier.
- The tables in the NDS are the organizations' "single version of the truth!"
- The data are there; all that is left are the processes to manage them!



Using Master Data to Update OLTP

- Helps organizations maintain consistent records throughout the enterprise.



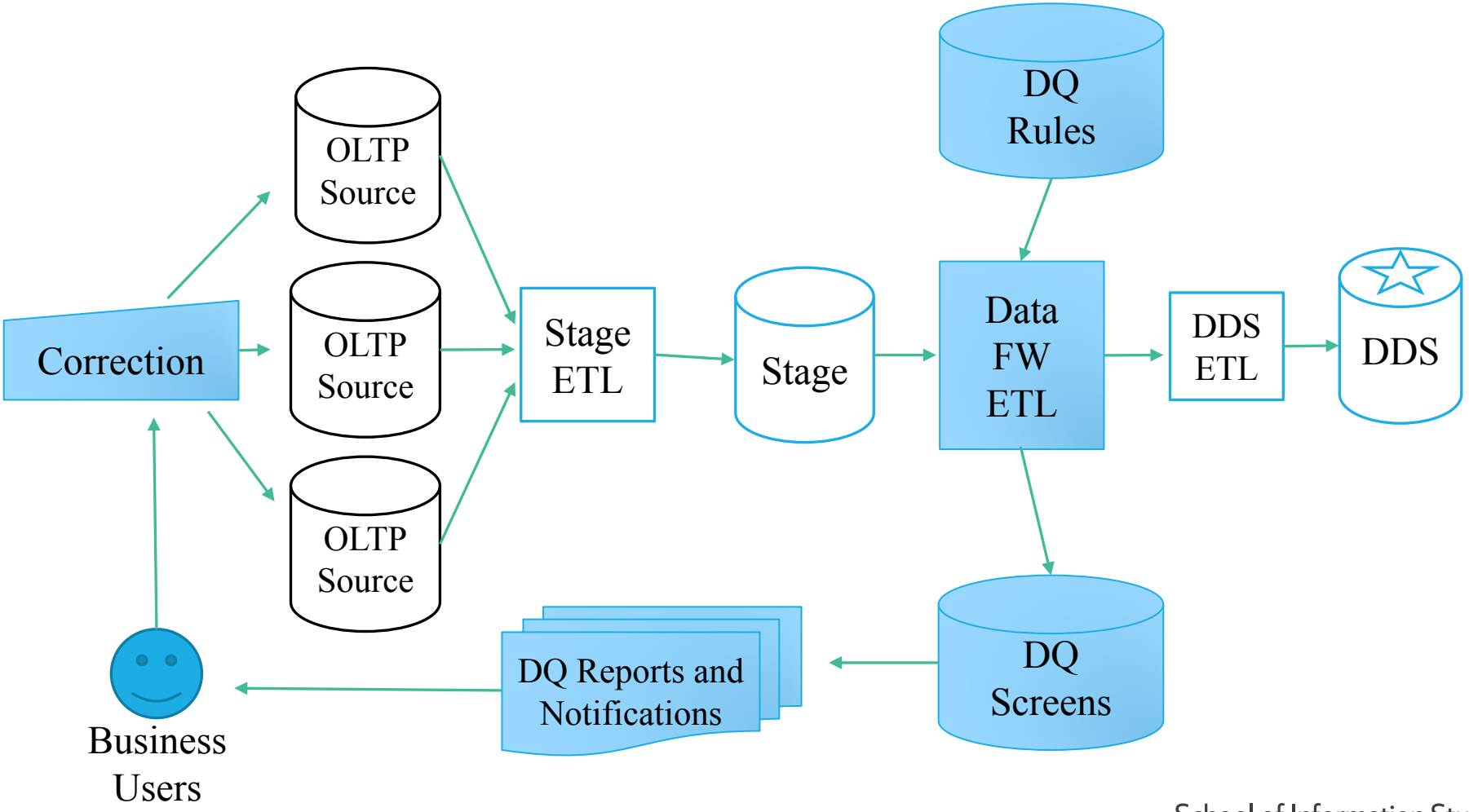
What Is Data Quality?

School of Information Studies
Syracuse University

What Is Data Quality?

- Activities to ensure the data in the data warehouse are correct and complete.
- There's no point to a data warehouse unless we can trust the data in it!

The Data Quality Process



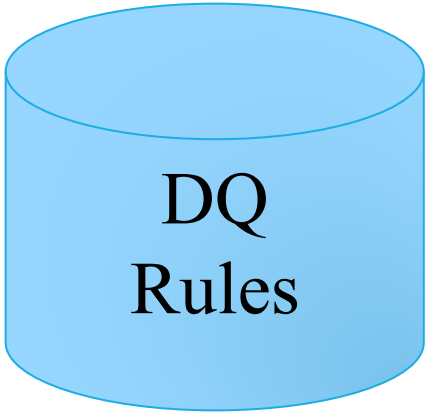
Why Not Automate?

- **Q: Why not fix the data automatically?**
- A: Data governance. It is not your decision to make—it's the data administrator's or DG team. If we auto-correct, we will never understand that it was wrong at the source.
- **Q: Why not notify, then fix automatically?**
- A: Creates an organizational-wide inconsistency because data are never corrected at the source. We know of the inconsistency but never address it.

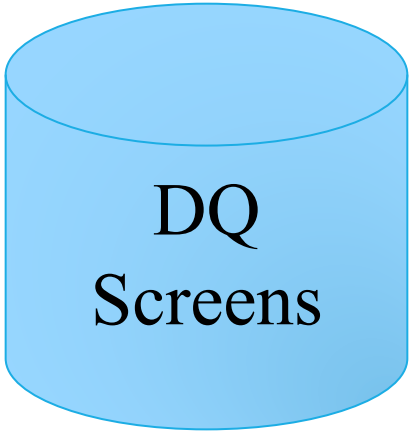
Data Quality Rules and Screens

School of Information Studies
Syracuse University

Data Quality Databases



- A database of rules and actions for each rule (accept, allow, warn, or fix).
- Keep rules simple.



- A database of actual data that failed a DQ screen.
- Includes information about which rule triggered the screen.

Five Criteria of Data Quality

- **Five criteria:** accuracy, precision, timeliness, completeness, consistency.
- For any set of data we determine a **data reliability** score based on one or more of the five factors.
- We write this score as a formula representing the percentage reliability of the data. Score should be between 0 and 1.
- This information is for the business user.
- Who determines this score? It's why you need data governance!

Example Incoming Validation DQ Rules: GPA

- DQ rule 1: accurate GPA
  - GPA must be a between 0.000 and 4.000.
  - Actions: REJECT.
- DQ rule 2: precise GPA
  - GPA must be three decimal places.
  - Actions: WARN.
- DQ rule 3: complete GPA
  - GPA must not be null.
  - Actions: FIX—Replace null with 0.000.
- Whenever a rule is triggered, it is written to the DQ screens database.

Three Types of DQ Rules

1. **Incoming data validation**
   - Rules are checked as staged data enters the DW (NDS, ODS, or DDS). Most common.
2. **Cross-reference validation**
   - Rules to check incoming data against the data already in the DW.
   - E.g., Monday's website visitors against a running average of the last three Mondays +/- 15%
3. **Data warehouse internal validation**
   - Check DW data against itself, typically for aggregates.
   - E.g., yearly sales in summary table matches actual number of sales by year.

The Data Quality Process in Action

School of Information Studies
Syracuse University

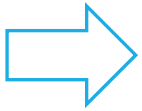
DQ Process in Action

DQ Rules

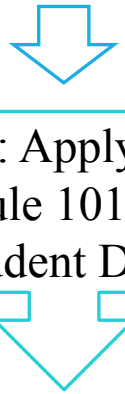
Rule Key	Name	Type	Action	…
101	GPA between 0 and 4	Incoming	Reject	…
102	Replace null with 0 in GPA	Incoming	Fix	…

Staged Student Data

SID	Name	GPA	…
992	Banks, Robyn	4.592	…
994	Gator, Allie	NULL	…
997	Tenz, Curt	-2.35	…



ETL: Apply DQ Rule 101 to Student Data



DQ Key	Rule Key	SID	Name	GPA	DQ Action	DQ Timestamp	…
1	101	992	Banks, Robyn	4.592	Reject	2017-05-12 9:00	…
2	101	997	Tenz, Curt	-2.35	Reject	2017-05-12 9:00	…

Data Cleansing

School of Information Studies
Syracuse University

Data Cleansing/Scrubbing

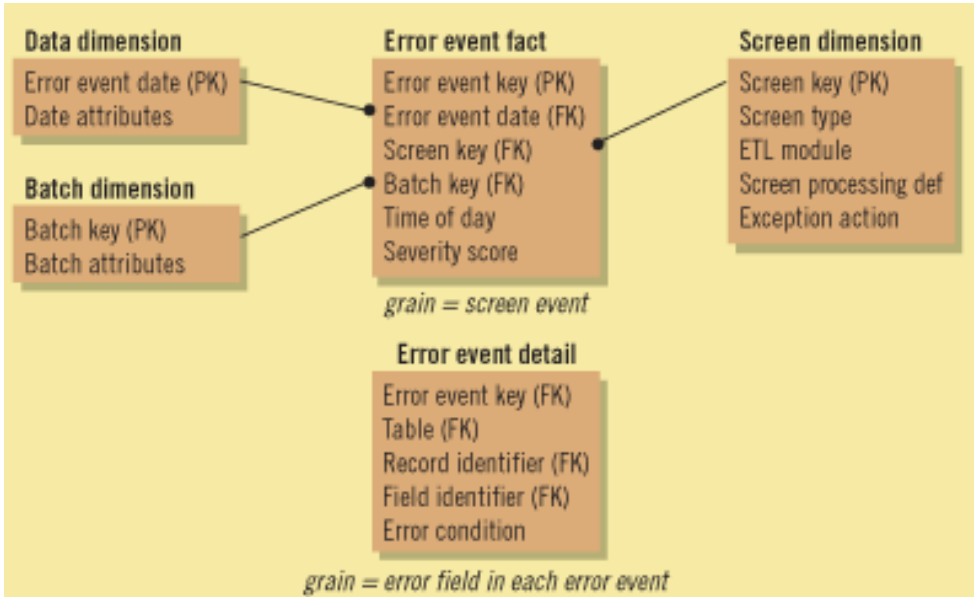
- The process of identifying and correcting bad data
- Trivial cases:
  - Replacing nulls with defaults
  - Fixing case: MIKE → Mike
  - Formatting: 3154432911 → 315-443-2911
- Advanced cases (matching)
  - Regular expressions: e-mails, IP addresses
  - Lookups on a business key
  - Fuzzy matching: Do not → Don't
  - Rule-based: [Bill, Will, William, Billy] → Bill

External Sources

- Use external datasets or web APIs to perform validation of common data.
- Examples:
  - E-mail address validation
  - Address validation
  - Postal/Zip codes
  - Country names to codes
  - GeoIP lookup/IP address validation
  - Credit card validation/Luhn check
  - Phone number validation

Error Event Schema

- A centralized dimensional model for logging failed data quality screens.
- Fact table grain is an error event.
- Dimensions are date, ETL job, quality screen source.
- A row added whenever there is a quality screening event that results in an error.
- Schema can also be used for warnings or fixes.



Types of Metadata

School of Information Studies
Syracuse University

Metadata

- Simply defined, metadata is “data about data."
- It exists to describe other data.
- In the data warehouse, metadata is internal facing. It supports the operations of the data warehouse.

All Kinds of Metadata

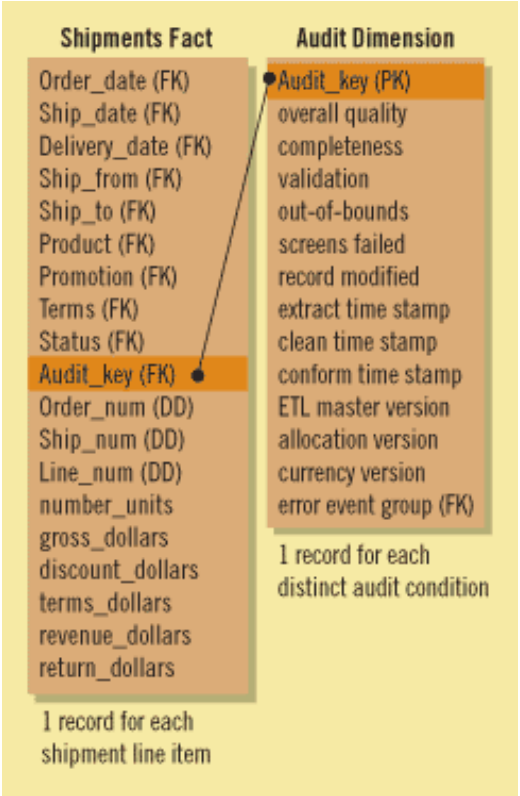
- Data definition: defines each attribute in the DW
- Data structure: defines structure of tables—keys, indexes, constraints
- Source system: defines source system data—attributes and their definitions
- ETL process: data for managing the ETL process
- Data quality: DQ rules, actions
- Audit: explains how data get into the DW
- Usage: tracks usage of DW data

Auditing Dimension, Lineage, and Dependency

School of Information Studies
Syracuse University

Audit Dimension

- ETL process metadata.
- A special dimension, assembled in the back room by the ETL system.
- Useful for tracking down how data in your schema “got there” or “was “changed.”
- Each fact and dimension table uses the audit dimension for recording results of the ETL process.
- There are two keys in the audit dimension for **original insert** and **most recent update.**



Data Lineage and Dependency

- **Lineage:** the ability to look at a data element and see how it was populated.
  - Audit tables help here.
- **Dependency:** the opposite direction. Look at a source table, and identify the cubes and star schemas that use it.
  - Custom metadata tables

Introduction

School of Information Studies
Syracuse University

Case Study Exercise

- You are the new CIO or CDO of a major university. One of your primary challenges is the current management practices around student data.
- Some of the issues are:
  - Admissions, the bursar, and the registrar all have different definitions of what is and isn't a student.
  - Academic departments keep records for their own students. Updates are compiled at inconsistent times throughout the semester.
  - When a student changes his or her name or address, the change must propagate through several systems. Right now it does not, and there are inconsistencies depending on whom you ask for the information.
  - Some students are alumni. Some applicants are students. Some alumni are applicants.

Solution to Part 1

School of Information Studies
Syracuse University

Part 1: Data Governance

- Devise a high-level plan for data governance.
- At this point you should not make any decisions but think about:
  - Who are the stakeholders?
  - What are your goals?
  - Which types of decisions must you make?
  - How will you organize to address the issues?

Part 1: Talking Points (1/2)

- You should assemble a DG team. Your team should include key stakeholders from the appropriate departments.
- The stakeholders do not need technological expertise—only knowledge in their subject areas and of the business processes.
- Data governance team
  - Data warehousing representative
  - Project manager
  - Data administrators from admissions, bursar, and registrar

Part 1: Talking Points (2/2)

- DG team decides on
  - Agreeable definitions of "student"
  - Sources of quality student data
  - A procedure to maintain consistency throughout organization
  - When student transitions to another role like alumni
  - Establishment of priorities and oversee implementation plan

Solution to Part 2

School of Information Studies
Syracuse University

Part 2: Your Plan

- With the data governance out of the way, it is time to think up solutions.
- Explain how you will solve these problems technologically, but keep your response at high level.
- Use techniques and approaches we discussed so far in this lesson.

Part 2: Talking Points

- Screen student data for anomalies and inconsistencies. Alert the appropriate parties of any issues.
- Maintain master data/golden record of student data. This master data, agreed upon by the DG committee, would then serve as a source to update shadow copies of student data.
- Notify the appropriate data administrators when the status of a student changes.
  - Student -> alumni
  - Alumni -> student
- None of this could be done without agreeable definitions for "what is a student."