

# Building a Quality-Oriented Data Warehouse

Alison Torres, Director  
Teradata Warehouse Consulting

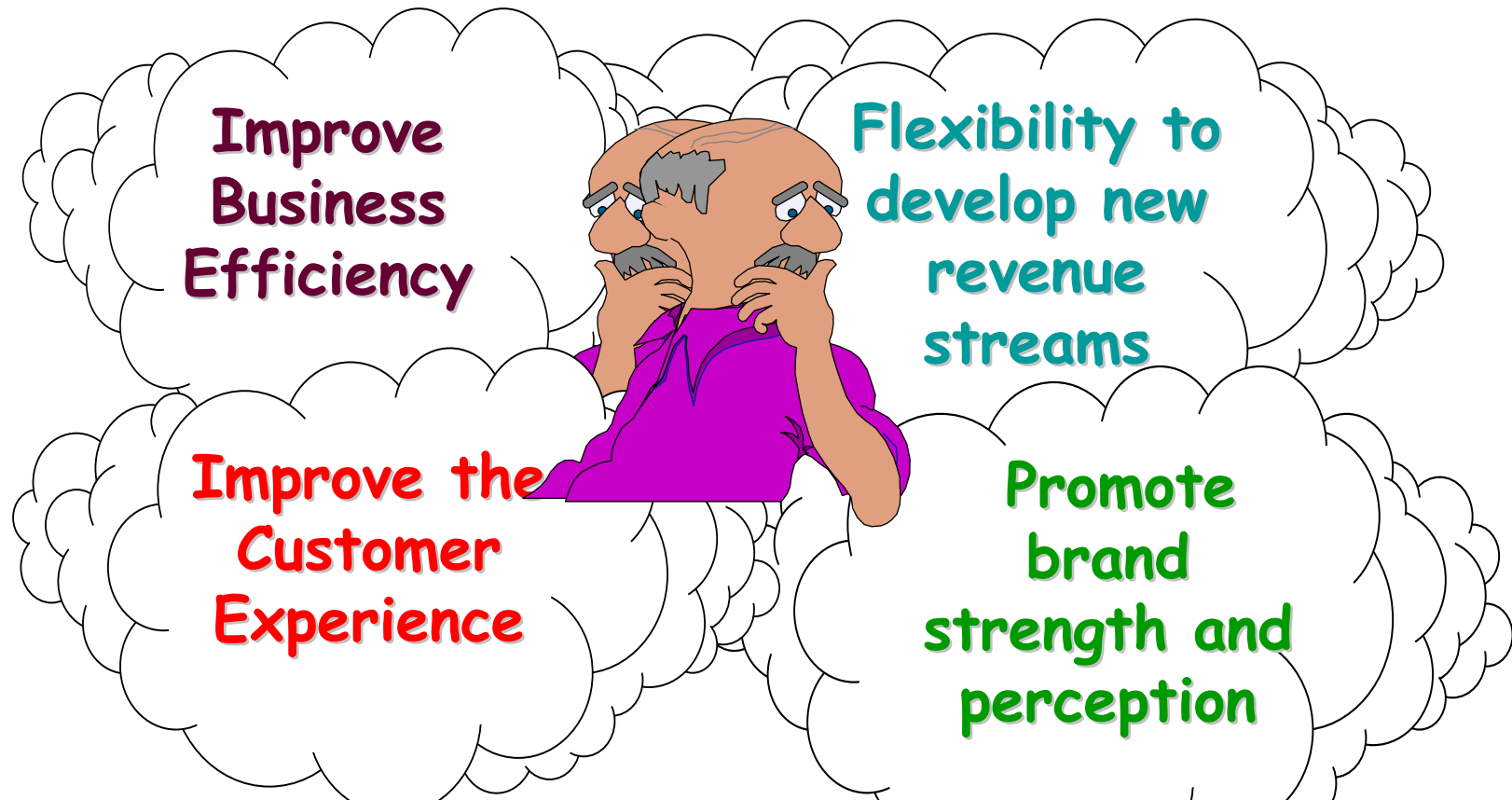


# Overview

---

- Key Business Objectives
- The Promise of Data Warehousing
- Business Rules
- Data Quality Defined
- Data Quality Misconceptions
- Dimensions of Data Quality
- Business Impact of Data Quality Issues
- Root causes of Data Quality problems
- Solutions for Data Quality Problems
- Building a Quality Oriented Data Warehouse

# Key Business Drivers



*The Data Warehouse is a great place to manage business decisions, building it right from the start is a big help to enable that management.*

*-Alison Torres*

# Business Imperatives

---

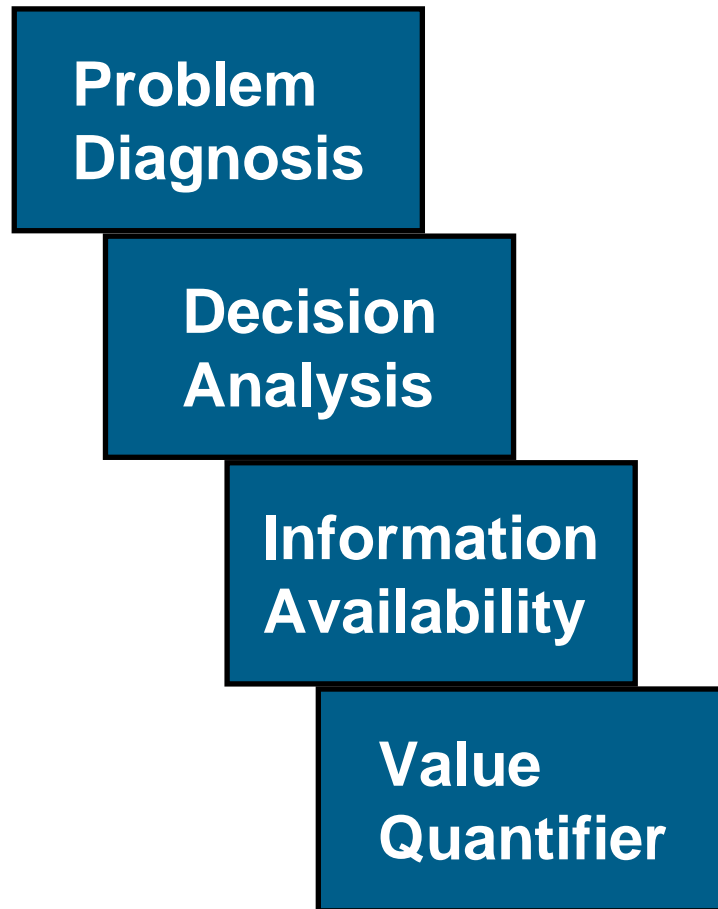
- **Must get closer to the *customer* !**
- **Must *improve productivity* of knowledge workers !**
- **Must be able to *integrate* new technologies *quickly* !**
- **Must become *flexible* to facilitate rapid '*market*' changes !**

**If you are not achieving your data warehouse objectives, there could be a problem with your business rules...**

***BUSINESS RULES are a foundational element of achieving your objectives, they are the 'must haves'.***

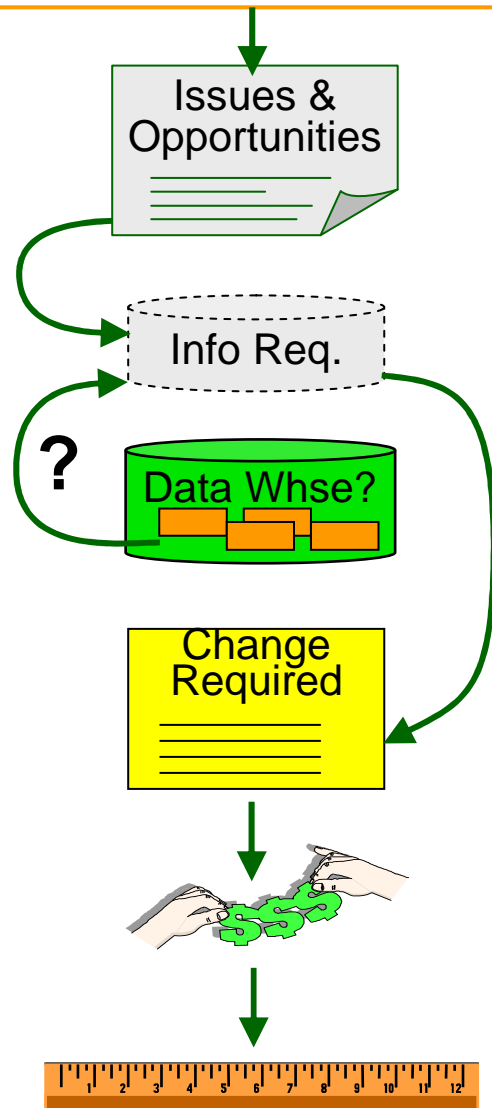
***BUSINESS RULES in a governmental environment would be determined by laws and the ramifications of not following them.***

# How Do You Discover Business Needs?



- Executive interviews by industry experts
- Draw out real business problems
- Uncover hidden agendas, laws
- Identify business champions
- Protect company and individual interests
- Reveal bottom-line impact
- Produce and present business value and priority report
- Build solid foundation for next steps

# Six Key Business Questions



*What are your key business issues and opportunities?*



*What information do you need to support these issues and opportunities?*

*Which information needs are being satisfied today by the data warehouse?*

*What process change is/would be associated with the issue and opportunity?*

*What is the value of that process change?*

*How will that value be measured?*

# Data Warehouse Quality

---

- **One Single Version of the Truth**
  - **Speaking one language**
  - **No dueling numbers**
- **Single Source**
  - **One stop shopping**
  - **Saves user time by eliminating need to obtain multiple interfaces and files**
- **Accessibility and timeliness of information**
  - **Allows End-User access without IT dependence**
  - **Reporting takes minutes or hours not weeks or months**

\* Often there are laws which prohibit having all the data in a single data warehouse. In that case, think of this on a departmental level where each new application would be integrated into the existing data warehouse.

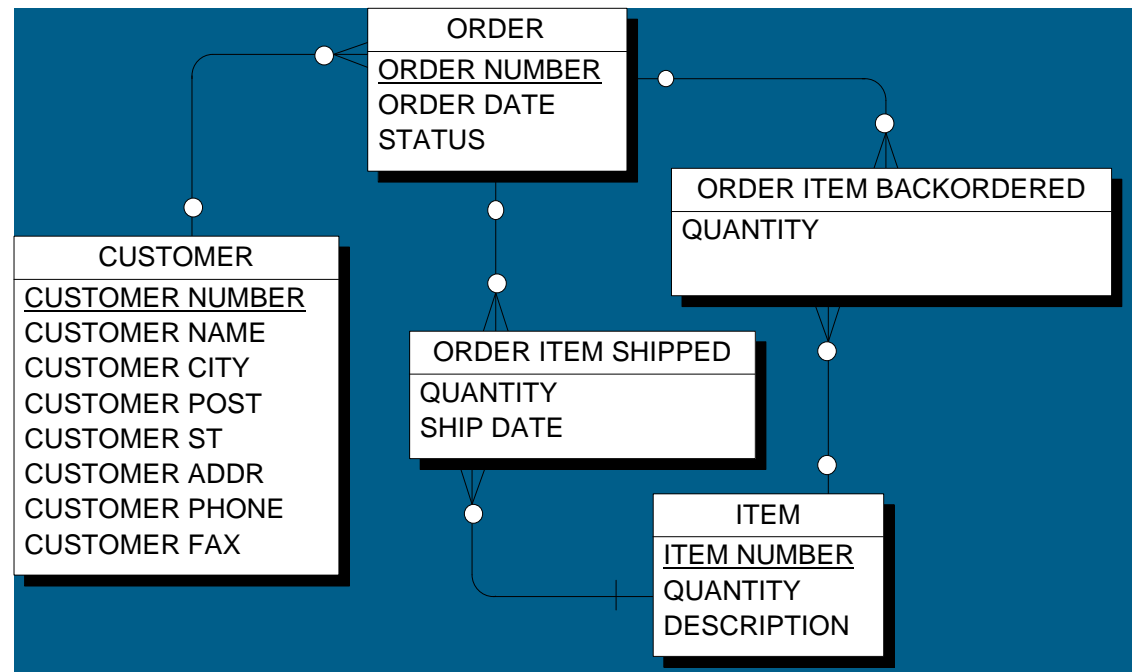
# Data Warehouse Quality

---

- **Cost avoidance**
  - **Multiple applications access the same data**
  - **Offload Legacy systems for DSS reporting**
  - **Avoid development of standalone applications to address specific data needs**
  - **Eliminate redundant systems**
- **Infrastructure to integrate future corporate acquisitions and data growth**
  - **Platform proven, highly scalable**
- **Ability to anticipate and respond to changes in the competitive marketplace**

# Model the Business

- Business information MODELING
  - > *Determined by BUSINESS RULES & VISION*
  - > The LOGICAL MODEL NEVER changes UNLESS
    - Underlying way of doing business changes, or
    - Adding NEW subject areas to MODEL will not impact existing model



# Model the Data (LDM)

- A Logical Data Model (LDM) is the result of information modeling
- A LDM is a diagram which shows:
  - > Entities (data of importance to the organization)
  - > Relationships between entities
  - > Attributes (properties of the data)
- LDMs are completely technology independent of any particular database or hardware platform

***A schematic view of the environment and a mock-up representation of something in the real world.***



# What is a Business Rule?

---

- **Set of conditions that govern a ‘business event’ so that it occurs in a way that is acceptable to the business.**
- **Business people identify rules that define all possible and permissible/not permissible conditions for the business**
- **Business rules should be written for and understood by business people in natural language and independent of technology**
- **Business rules are meant to be challenged by business people and implemented in technology that allows for controlled, but spontaneous business change**

**“An exciting new technology called business rules is beginning to have a major impact on the IT industry, more precisely, on the way we develop and maintain computer applications. Business rules can be seen in some respects as the next (and giant) evolutionary step in implementing the original relational vision.”**

**...C. J. Date, The Business Rules Approach to Application Development**

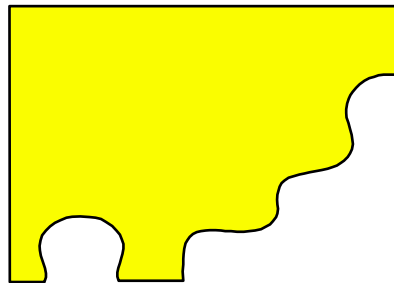
# Business Rules Challenges

**Operational Systems are usually designed to solve a specific business problem and are rarely developed to a coordinated corporate plan.**

**“And get it done quickly; we don’t have time to worry about corporate standards ...”**

**Taxation**

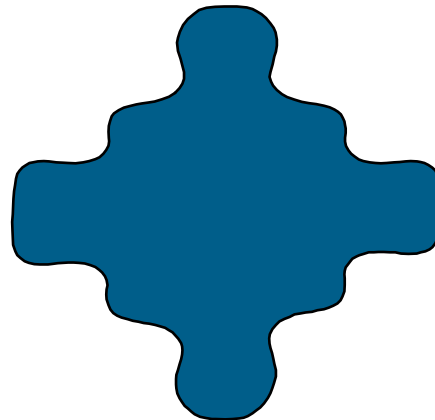
**Compliance**



**Different keys,  
same data**

**Welfare**

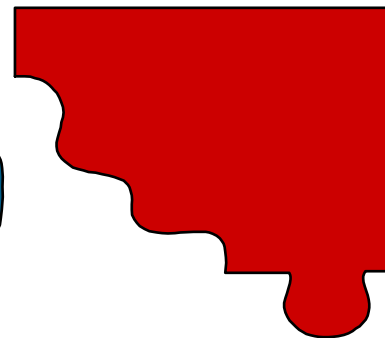
**Collections**



**Same data as other  
apps, but uses  
different names  
for it**

**Education**

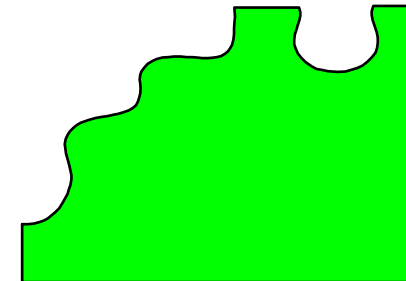
**Adjudication**



**Different data,  
but uses same  
names as data  
in other apps with  
different meaning**

**Food Stamps**

**Litigation**



**Unique data found  
only here, nowhere  
else**

# Data Quality, How Serious Is This Issue?



“Poor data quality is the norm rather than the exception, but most organizations are in a state of denial about this issue.”

-- 1998

*from “Gartner Group”*

‘Our **Data Warehouse** matches our **Operational Systems**.’

# Data Quality Defined

**Data quality** is defined as the *measure of the suitability* of data for its *intended purpose*.

- In the context of customer relationship management the intended use depends on your company's strategy and how you want to relate to your customers.
  - > Am I sending a marketing brochure?
    - ☞ I want to find a good customer.
  - > Am I a tax collector?
    - ☞ I want to find a bad customer.

**Data quality** is a *business problem*, not an IT problem. It is a way of doing business. All business processes need to be designed from the ground-up to insure data quality.

**Perspectives will vary depending upon usage.  
Who makes the decision?**

# Common Data Quality Misconceptions

---

- It is the IT department's job to assure data quality.
  - > Everyone in the organization must work together.
- Data quality problems can be fixed with data cleansing tools and techniques.
  - > Save your money.
- If the data in the warehouse matches the source system IT has done its job well.
  - > Too narrow a vision.
- Data quality problems can be fixed by modifying front-end systems to perform the required edits and validation.
  - > What about process issues?
- Bad data should be dropped and not loaded into the data warehouse
  - > Creates a data quality problem of another type

# Twelve Dimensions of Data Quality

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	A measure of the amount of information captured about an object or event.	All information about a specific call is captured including duration, start and stop time, origination and termination information, billing information, network information, etc.	None of the network related information for a specific call is captured. Nothing is known about how the call was handled by the network.
Completeness	A measure of information gaps within a specific entity occurrence.	Name, age, and occupation are known for all customers.	Name and age are known for all customers but occupation is known for only 50% of the customers.
Uniqueness	A measure of unnecessary information replication.	Customer information is stored once for each customer.	Certain customer's records are duplicated due to variations in the spelling of the name, alternate address, etc. The records are not linked in any way.
Interpretability	A measure of semantic standards being applied	A date is stored as 11 June 2010.	A date stored as 11062010 is interpreted as November, 06, 2010.
Timeliness	A measure of how current a record is.	All customer addresses represent the current place of dwelling.	Many customers have changed their address without informing the company.
Precision	A measure of exactness	The amount of tax due for this specific transaction is \$0.104.	The amount of tax due for this specific transaction is stored as \$0.10.
Depth	A measure of the amount of entity or event history that is retained.	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

# Dimensions of Data Quality - Real World Accuracy

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of agreement that exist in situations with data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities in the real world.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	A measure of the amount of information.	All information about a specific call is captured including duration, start and stop.	None of the network related information for a specific call is captured. Nothing is known.
Completeness		transaction is \$0.104.	transaction is stored as \$0.10.
Uniqueness			
Integrity			
Timeliness			
Presentability			
Depth	A measure of the amount of entity or event history that is retained.	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

## Accuracy - matches the real world

Description:

A measure of information correctness

Conformance:

A balance of \$10,000 is stored as \$10,000

Non-Conformance:

A balance of \$10,000 is stored as \$12,500

# Dimensions of Data Quality - Multiple Databases

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of events recorded versus the real world quantities of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	A measure of the information captured about an object.	All information about a specific call is captured including duration, start and stop time, origination and termination information, billing information, network information.	None of the network related information for a specific call is captured. Nothing is known about how the call was handled by the network.

## Consistency - multiple databases

Description:

A measure of the degree of conflicts that exist in situations with redundant data.

Conformance:

A balance of \$10,000 in ABC system is also stored as \$10,000 in the XYZ system

Non-Conformance:

A balance of \$10,000 in ABC system is stored as \$12,500 in the XYZ system

Integrity	A measure of the relationship between data to another item of related information.	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

# Dimensions of Data Quality - Unrecorded Data

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	the amount of about an	All information about a specific call is captured including duration, start and stop time, origination and termination information, billing information, network information, etc.	None of the network related information for a specific call is captured. Nothing is known about how the call was handled by the network.

## Entirety - unusable or unavailable

Description:

A measure of the quantities of entities created vs. the real world or the number of actual events.

Conformance:

All phone calls made were recorded and stored for billing.

Non-Conformance:

Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.

# Dimensions of Data Quality - Recorded Data, Missing Details

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	A measure of the amount of information captured about an	All information about a specific call is captured including duration, start and stop time, origination and termination information, billing information, network	None of the network related information for a specific call is captured. Nothing is known about how the call was handled by the network.

## Breadth - dropped information

**Description:**

A measure of the amount of information captured about an object or event.

**Conformance:**

All information about a specific call is captured including duration, start and stop time, origination and termination information, billing information, network information, etc.

**Non-Conformance:**

None of the network related information for a specific call is captured. Nothing is known about how the call was handled by the network.

# Dimensions of Data Quality - Voluntary Data

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	A measure of the amount of information captured about an object or event.	All information about a specific call is captured including duration, start and stop time, origination and termination information, billing information, network information, etc.	None of the network related information for a specific call is captured. Nothing is known about how the call was handled by the network.
Completeness	A measure of information gaps within a specific entity occurrence.	Name, age, and occupation are known for all customers.	Name and age are known for all customers but occupation is known for only 50% of the customers.
Uniqueness		Customer information is stored once for	Certain customer's records are duplicated

## Completeness - optional collection

**Description:** A measure of information gaps within a specific entity occurrence.

**Conformance:** Name, age, and occupation are known for all customers.

**Non-Conformance:** Name and age are known for all customers, but occupation is known for only 50% of customers.

## Dimensions of Data Quality - Duplicate Data

### Uniqueness - slight variations

**Description:** A measure of unnecessary information replication.

**Conformance:** Customer information is stored once for each customer.

**Non-Conformance:** Certain customer's records are duplicated due to variations in the spelling of the name, alternate address, etc. The records are not linked in any way.

Completeness	A measure of how much information is known for all customers.	Name and age are known for all customers but occupation is known for only 50% of the customers.	
Uniqueness	A measure of unnecessary information replication.	Customer information is stored once for each customer.	Certain customer's records are duplicated due to variations in the spelling of the name, alternate address, etc. The records are not linked in any way.
Interpretability	A measure of semantic standards being applied	A date is stored as 11 June 2010.	A date stored as 11062010 is interpreted as November, 06, 2010.
Timeliness	A measure of how current a record is.	All customer addresses represent the current place of dwelling.	Many customers have changed their address without informing the company.
Precision	A measure of exactness	The amount of tax due for this specific transaction is \$0.104.	The amount of tax due for this specific transaction is stored as \$0.10.
Depth	A measure of the amount of entity or event history that is retained.	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

# Dimensions of Data Quality - Data Formats

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of	A balance of \$10,000 in the ABC system is	A balance of \$10,000 in the ABC system is
<div style="background-color: #0056b3; color: white; padding: 10px; border: 2px solid #0056b3;"> <h2 style="margin: 0;">Interpretability - misinterpretation</h2> <p><b>Description:</b> A measure of semantic standards being applied.</p> <p><b>Conformance:</b> A date is stored as 11 June 2010 or "11062010".</p> <p><b>Non-Conformance:</b> A date stored as "11062010" is interpreted as Nov 06, 2010.</p> </div>			
Uniqueness	A m inf	Customer information is stored once for each customer.	Certain customer's records are duplicated due to variations in the spelling of the name, alternate address, etc. The records are not linked in any way.
Interpretability	A measure of semantic standards being applied	A date is stored as 11 June 2010.	A date stored as 11062010 is interpreted as November, 06, 2010.
Timeliness	A measure of how current a record is.	All customer addresses represent the current place of dwelling.	Many customers have changed their address without informing the company.
Precision	A measure of exactness	The amount of tax due for this specific transaction is \$0.104.	The amount of tax due for this specific transaction is stored as \$0.10.
Depth	A measure of the amount of entity or event history that is retained.	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

# Dimensions of Data Quality - Concurrent Data

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real	All phone calls that were made were recorded and stored for billing	Calls to a particular NPA-NNX were not recorded due to a switch profile problem
Timeliness	A measure of how current a record is.	All customer addresses represent the current place of dwelling.	Many customers have changed their address without informing the company.
Precision	A measure of exactness	The amount of tax due for this specific transaction is \$0.104.	The amount of tax due for this specific transaction is stored as \$0.10.
Depth	A measure of the amount of entity or event history that is retained.	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

## Timeliness - missed updates

**Description:** A measure of how current a record is.

**Conformance:** All customer addresses represent the current place of dwelling.

**Non-Conformance:** Many customers have changed their address without informing the company.

# Dimensions of Data Quality - Physical Data Types

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.

## Precision - inappropriate rounding

**Description:** A measure of exactness.

**Conformance:** The amount of tax due for this specific transaction is \$0.104.

**Non-Conformance:** The amount of tax due for this specific transaction is stored as \$0.10.

Timeliness	A measure of information correctness	Home addresses represent the current place of dwelling.	Many customers have changed their address without informing the company.
Precision	A measure of exactness	The amount of tax due for this specific transaction is \$0.104.	The amount of tax due for this specific transaction is stored as \$0.10.
Depth	A measure of the amount of entity or event history that is retained.	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

# Dimensions of Data Quality - Adequate Data

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Completeness			
Entity			
Breadth			
Consistency			
Uniqueness			
Integrity			
Timeliness	A measure of the amount of entity or event history that is retained.	All customer addresses represent the current place of dwelling.	Many customers have changed their address without informing the company.
Precision	A measure of validity with respect to another item of related information.	The amount of tax due for this specific transaction is \$0.104.	The amount of tax due for this specific transaction is stored as \$0.10.
Depth	A pool of valid values	A complete history of orders, bills, and payments is retained for all customers.	Orders, bills, and payment information is only retained for one year. Each month, the prior year records are deleted for that month to make room for the new information.
Integrity		A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain		Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

## Depth - missing history

**Description:** A measure of the amount of entity or event history that is retained.

**Conformance:** A complete history of orders, bills, and payments, is retained for all customers.

**Non-Conformance:** Orders, bills, and payments, are retained for one year. Each month the prior year records are deleted for that month to make room for the new information.

# Dimensions of Data Quality - Referential Integrity

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Integrity	A measure of validity with respect to another item of related information.	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

## Integrity - relationship inconsistency

**Description:** A measure of validity with respect to another item of related information.

**Conformance:** A call detail record contains a from number of (802) 888-9999.

**Non-Conformance:** The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.

# Dimensions of Data Quality - Confusing Data

Dimension	Description	Conformance	Non-Conformance
Accuracy	A measure of information correctness	A balance of \$10,000 is stored as a balance \$10,000.	A balance of \$10,000 is stored as a balance \$12,500.
Consistency	A measure of the degree of conflicts that exist in situations with redundant data.	A balance of \$10,000 in the ABC system is also stored as \$10,000 in the XYZ system.	A balance of \$10,000 in the ABC system is stored as \$12,500 in the XYZ system.
Entirety	A measure of the quantities of entities created, versus the real world or the number of actual events.	All phone calls that were made were recorded and stored for billing.	Calls to a particular NPA-NNX were not recorded due to a switch profile problem. Revenue for these calls will be lost.
Breadth	A measure of the amount of information captured about an	All information about a specific call is captured including duration, start and stop	None of the network related information for a specific call is captured. Nothing is known

## Domain - inconsistency

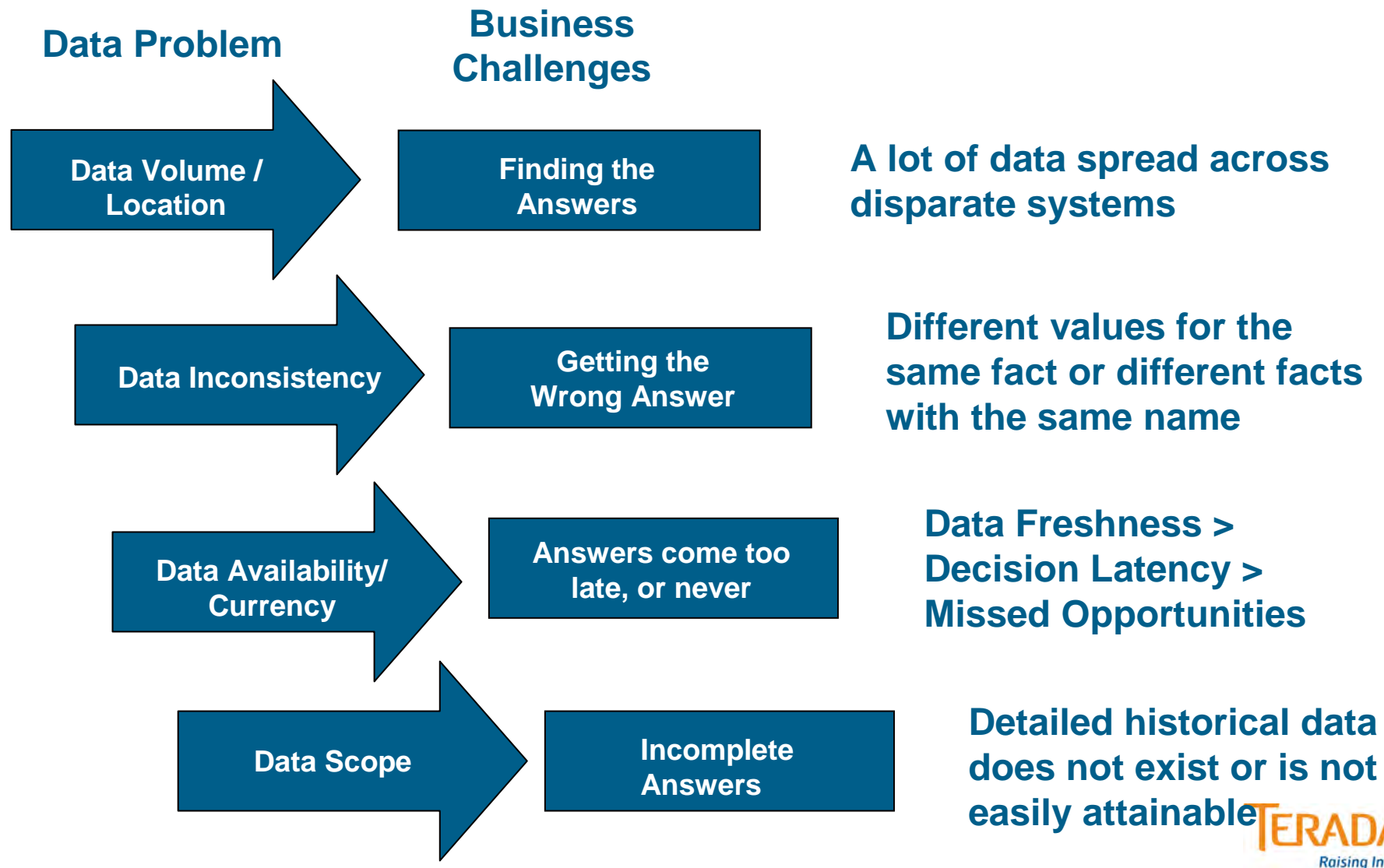
Description: A pool of valid values.

Conformance: A call number of (802) 888-5773.

Non-Conformance: Confusion of Domains, Invalid Data for a given Domain.  
Erroneous NPA values, confusion between NPA and NNX values.

Integrity	A call detail record contains a from number of (802) 888-9999.	The Terminating Point Master table indicates that due to an area code split the 271 NNX is now in the 732 NPA.	
Domain	A pool of valid values	Area codes 201, 908, 732.	Confusion between NPA and NXX codes.

# What's Standing in the Way of Meeting Data Quality Objectives?



# Possible Root Causes of Data Quality Problems

---

- Processes are not designed to capture data that is needed (customer is not asked to provide data that is needed)
- Customer response issues (customer not motivated to provide data or data capture questions are ambiguous or confusing)
- Processes do not include functions to capture data that has been provided
- Data not validated by data capture functions (data input errors)
- External data sources not reliable
- Inconsistent or poorly defined metadata (business data definitions, code sets, etc.)
- Redundant data capture functions (same data element captured in different processes or at different times)

# Possible Root Causes of Data Quality Problem

---

- Data aging problems (no process to maintain data that has been captured)
- Data replication errors (physical copies of data mismanaged)  
Note: A data warehouse is a physical copy of operational data.
- Data derivation errors (inconsistent business rules for creation and maintenance of derived data elements)
- Inconsistent application of business rules or poorly defined business rules for the storage and maintenance of data
- Fraud and security issues (data content not properly protected)
- Database definition problems (data normalization, data definition, and referential integrity errors)

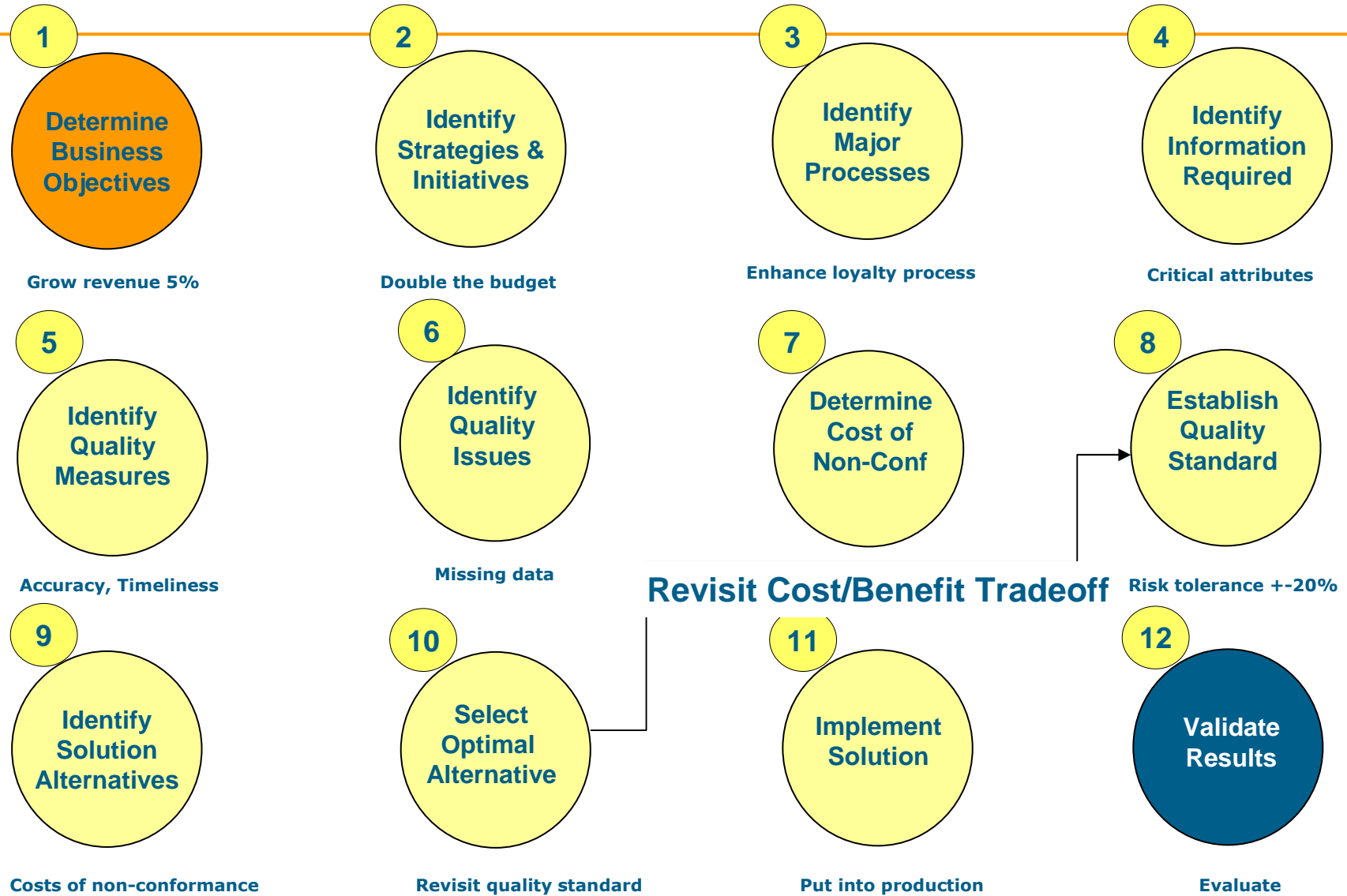
## What Is Needed

---

Put a process in place to allow **IT** and **business** users to work together to identify and correct the root causes of specific data quality issues that are negatively impacting critical operational and analytical processes.

The assumption is that the proper operation of these processes is key to achieving critical CRM and business objectives. These processes are directly linked to specific revenue growth and expense reduction objectives.

# Business Driven Quality Improvement Process



# Building a Quality Oriented Data Warehouse

---

- Make sure that data sources come from the data base of record.
- Some operational systems store data attributes redundantly - make sure you obtain the attribute of record.
- All data attributes are not of equal importance. Determine the critical data elements and focus your initial efforts on making sure they are correct.
- Design ETL/ELT processes to stage the incoming data and perform validations before loading into target tables if possible.

# Building a Quality Oriented Data Warehouse

---

- Allow time during implementation to discover, analyze, and document valid domain values and business rules for each attribute. Involve business SMEs and source system experts as much as possible.
- Get samples of source data loaded as soon as possible during the design phase. Use the RDBMS to assess the quality of the data and determine what you are up against.
- Don't assume that documentation is correct. Use the RDBMS to validate that attributes contain what they are supposed to.

# Building a Quality Oriented Data Warehouse

---

- Build quality control processes into the ETL/ELT to make sure that data is not dropped and that all data elements are accurately replicated into the data warehouse.
- Leverage the data warehouse as a quality measurement platform to support all defined KPIs and quality measures.
- Leverage the data warehouse as the platform to support root cause analysis on identified data quality issues. The ability to drill down to detailed data and source attributes will be key to diagnosing problems with derived attributes.

Alison Torres

alison.torres@teradata-ncr.com

---

**TERADATA**®  
*Raising Intelligence*