Data Quality Assessment Framework

ABSTRACT

Many efforts to measure data quality focus on abstract concepts and cannot find a practical way to apply them. Or they attach to specific issues and cannot imagine measurement beyond them. To avoid these traps, a team at Ingenix developed the Data Quality Assessment Framework (DQAF). Focusing on four objective data quality dimensions (completeness, timeliness, validity, and consistency), the DQAF defines 38 measurement types that can be applied to relational data (e.g., consistency across multiple columns) regardless of the specific content of the data This paper will discuss how the DQAF enables establishment of core data quality measures and a common model for analyzing and storing data quality results.

BIOGRAPHY

Laura Sebastian-Coleman
Ingenix, a subsidiary of UnitedHealth Group (UHG)

Laura Sebastian-Coleman leads a Data Quality Team at Ingenix, a subsidiary of UnitedHealth Group (UHG), where she has worked since 2003. During that time, she implemented a Data Quality Program for UHG's enterprise data warehouse. The program includes per load and quarterly collection and reporting on data quality metrics and a monthly Data Quality Community meeting. In 2009, she led the team that created the Data Quality Assessment Framework.

Before joining Ingenix, she worked the commercial insurance industry. She holds a Certificate in Information Quality from MIT. She is currently Director of Member Services for the International Association for Information and Data Quality (IAIDQ) and has delivered papers at MIT's International Conference on Information Quality and Industry Symposium, as well as at the IDQ Conference.

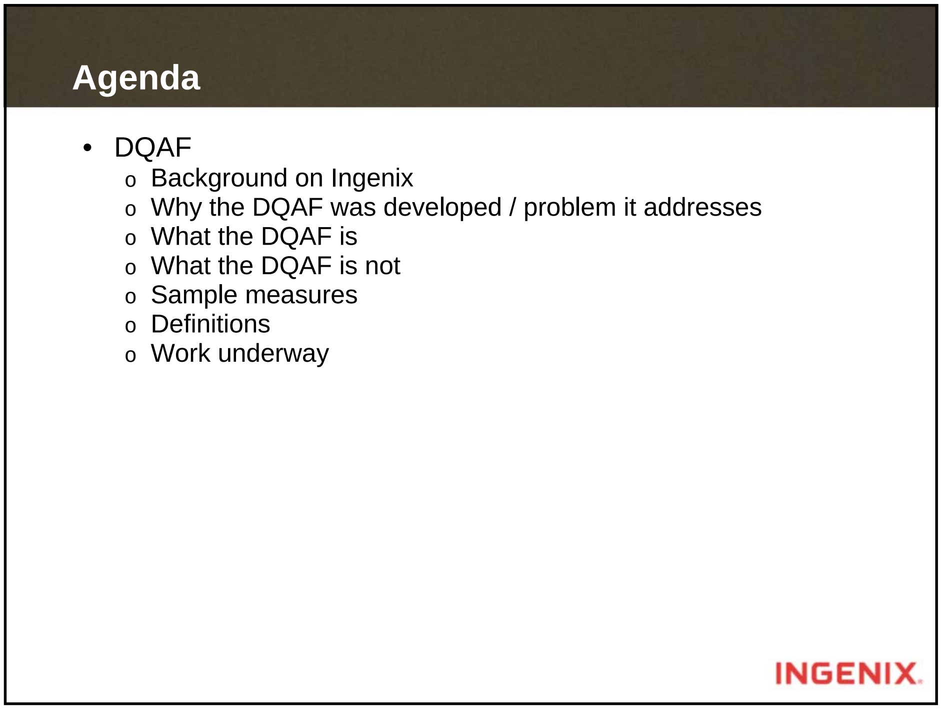
She holds a B.A. in English and History from Franklin & Marshall College, and Ph.D. in English Literature from the University of Rochester (NY).

DQAF
Data Quality Assessment Framework

MIT Information Quality Industry Symposium
July 14-16, 2010

Laura Sebastian-Coleman, Ph.D.
Ingenix – UnitedHealth Group

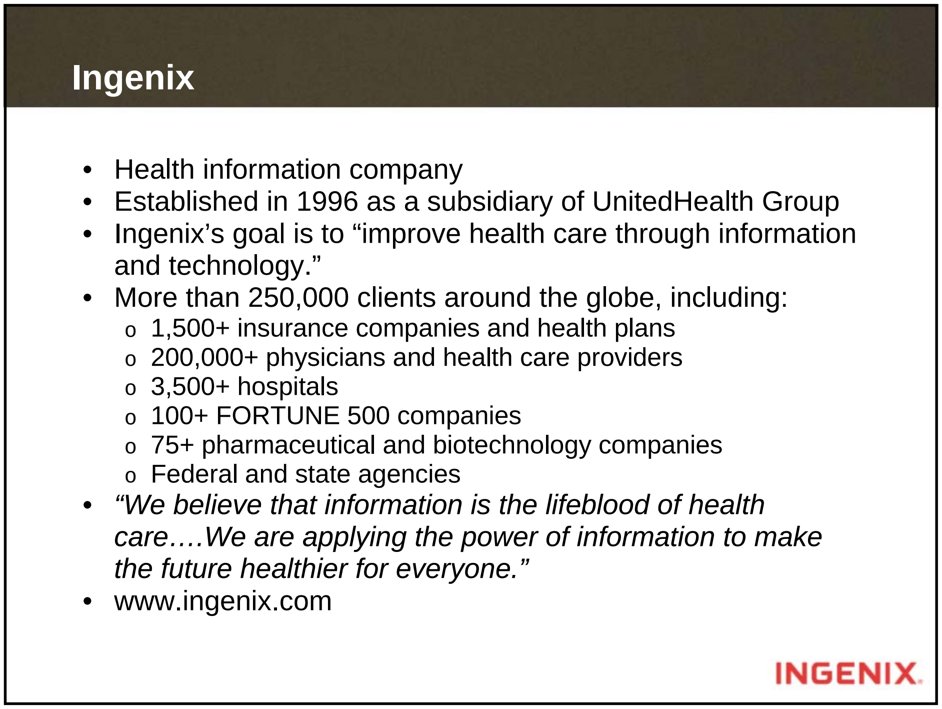
INGENIX.



Agenda

- DQAF
  - Background on Ingenix
  - Why the DQAF was developed / problem it addresses
  - What the DQAF is
  - What the DQAF is not
  - Sample measures
  - Definitions
  - Work underway

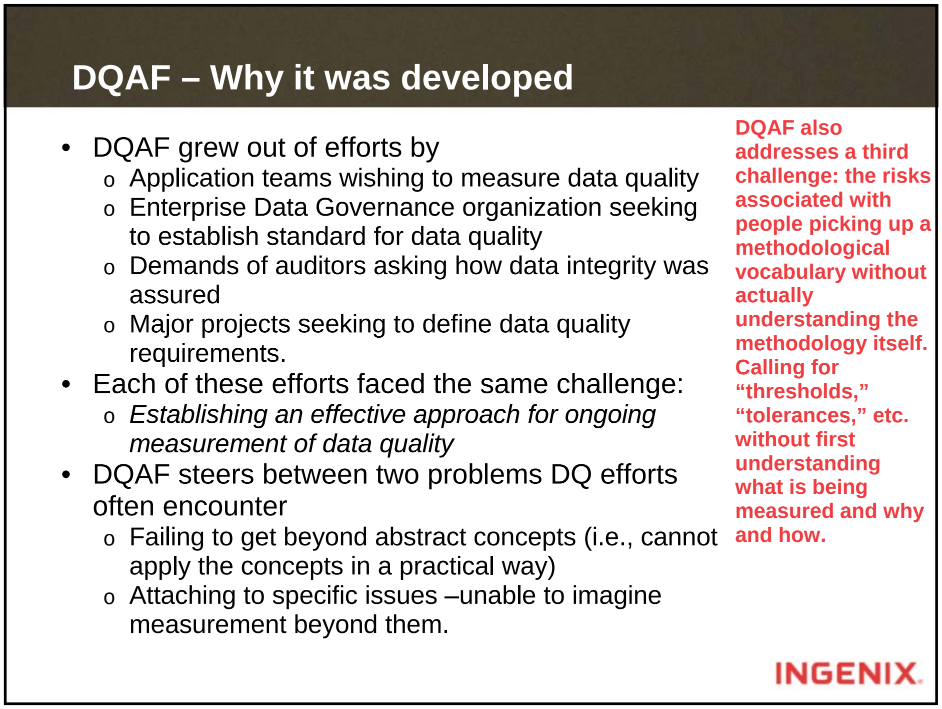
INGENIX.



Ingenix

- Health information company
- Established in 1996 as a subsidiary of UnitedHealth Group
- Ingenix's goal is to "improve health care through information and technology."
- More than 250,000 clients around the globe, including:
  - 1,500+ insurance companies and health plans
  - 200,000+ physicians and health care providers
  - 3,500+ hospitals
  - 100+ FORTUNE 500 companies
  - 75+ pharmaceutical and biotechnology companies
  - Federal and state agencies
- *"We believe that information is the lifeblood of health care….We are applying the power of information to make the future healthier for everyone."*
- www.ingenix.com

INGENIX.

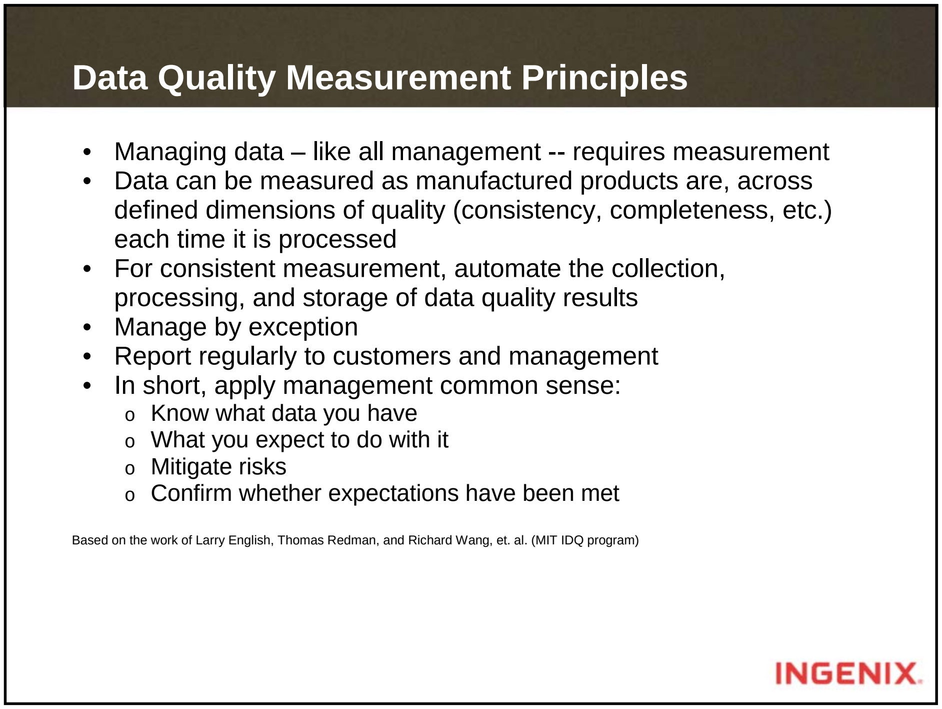


DQAF – Why it was developed

- DQAF grew out of efforts by
  - Application teams wishing to measure data quality
  - Enterprise Data Governance organization seeking to establish standard for data quality
  - Demands of auditors asking how data integrity was assured
  - Major projects seeking to define data quality requirements.
- Each of these efforts faced the same challenge:
  - *Establishing an effective approach for ongoing measurement of data quality*
- DQAF steers between two problems DQ efforts often encounter
  - Failing to get beyond abstract concepts (i.e., cannot apply the concepts in a practical way)
  - Attaching to specific issues –unable to imagine measurement beyond them.

DQAF also addresses a third challenge: the risks associated with people picking up a methodological vocabulary without actually understanding the methodology itself. Calling for "thresholds," "tolerances," etc. without first understanding what is being measured and why and how.

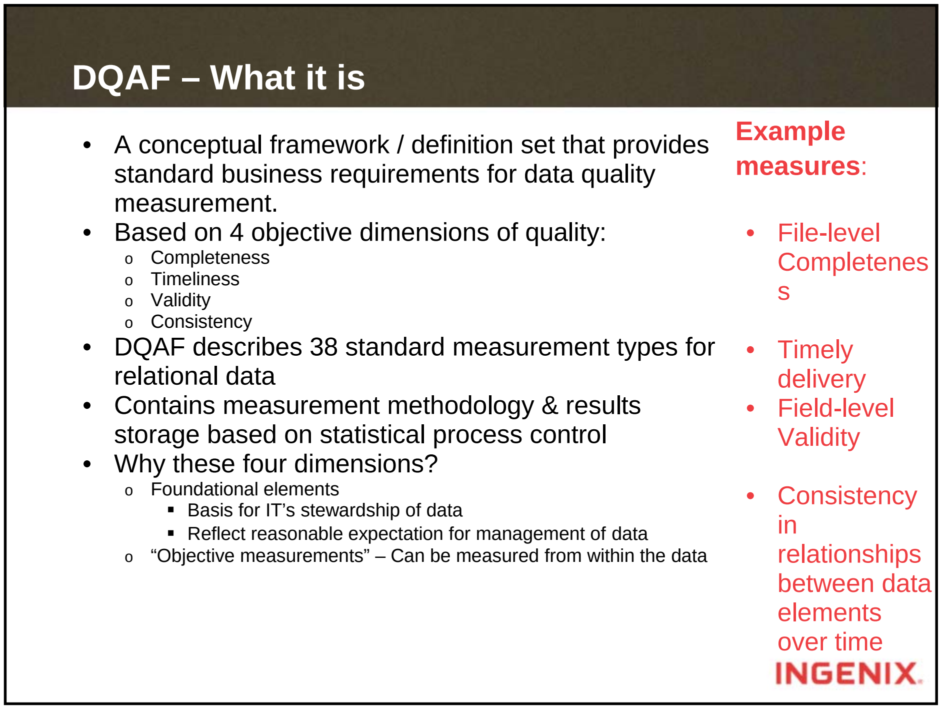
INGENIX.



Data Quality Measurement Principles

- Managing data – like all management -- requires measurement
- Data can be measured as manufactured products are, across defined dimensions of quality (consistency, completeness, etc.) each time it is processed
- For consistent measurement, automate the collection, processing, and storage of data quality results
- Manage by exception
- Report regularly to customers and management
- In short, apply management common sense:
  - Know what data you have
  - What you expect to do with it
  - Mitigate risks
  - Confirm whether expectations have been met

Based on the work of Larry English, Thomas Redman, and Richard Wang, et. al. (MIT IDQ program)

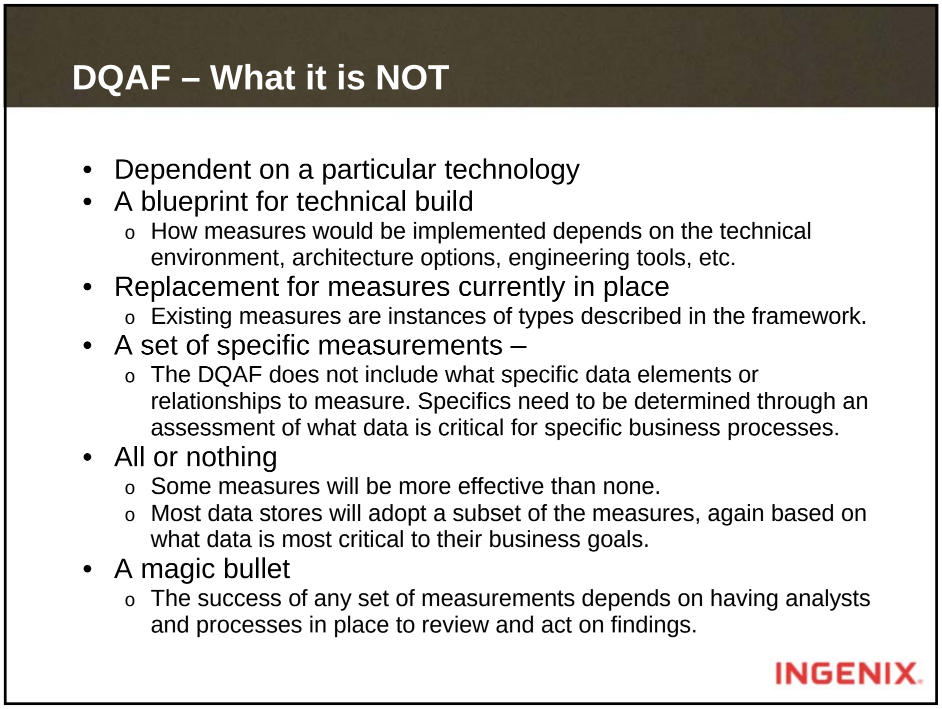


DQAF – What it is

- A conceptual framework / definition set that provides standard business requirements for data quality measurement.
- Based on 4 objective dimensions of quality:
  - Completeness
  - Timeliness
  - Validity
  - Consistency
- DQAF describes 38 standard measurement types for relational data
- Contains measurement methodology & results storage based on statistical process control
- Why these four dimensions?
  - Foundational elements
    - Basis for IT's stewardship of data
    - Reflect reasonable expectation for management of data
  - "Objective measurements" – Can be measured from within the data

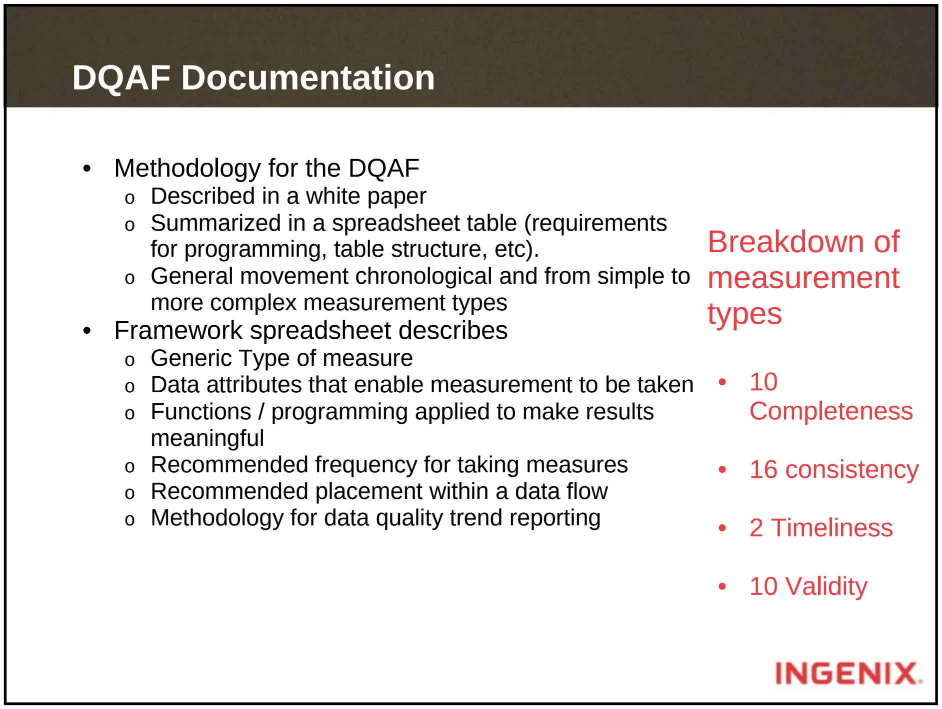
Example measures:

- File-level Completeness
- Timely delivery
- Field-level Validity
- Consistency in relationships between data elements over time



DQAF – What it is NOT

- Dependent on a particular technology
- A blueprint for technical build
  - How measures would be implemented depends on the technical environment, architecture options, engineering tools, etc.
- Replacement for measures currently in place
  - Existing measures are instances of types described in the framework.
- A set of specific measurements –
  - The DQAF does not include what specific data elements or relationships to measure. Specifics need to be determined through an assessment of what data is critical for specific business processes.
- All or nothing
  - Some measures will be more effective than none.
  - Most data stores will adopt a subset of the measures, again based on what data is most critical to their business goals.
- A magic bullet
  - The success of any set of measurements depends on having analysts and processes in place to review and act on findings.

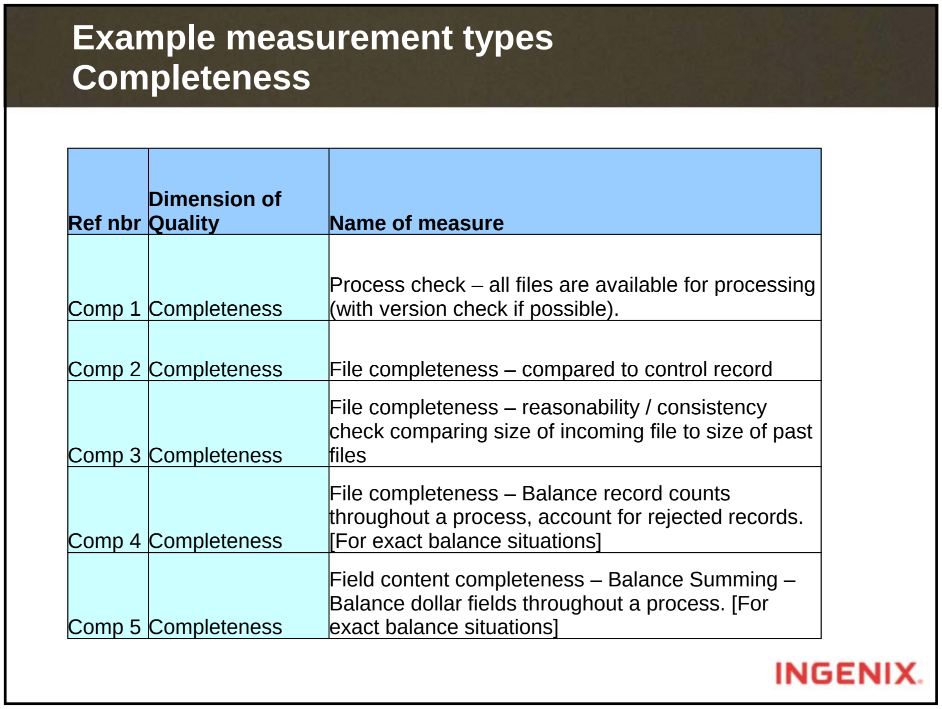


DQAF Documentation

- Methodology for the DQAF
  - Described in a white paper
  - Summarized in a spreadsheet table (requirements for programming, table structure, etc).
  - General movement chronological and from simple to more complex measurement types
- Framework spreadsheet describes
  - Generic Type of measure
  - Data attributes that enable measurement to be taken
  - Functions / programming applied to make results meaningful
  - Recommended frequency for taking measures
  - Recommended placement within a data flow
  - Methodology for data quality trend reporting

Breakdown of measurement types

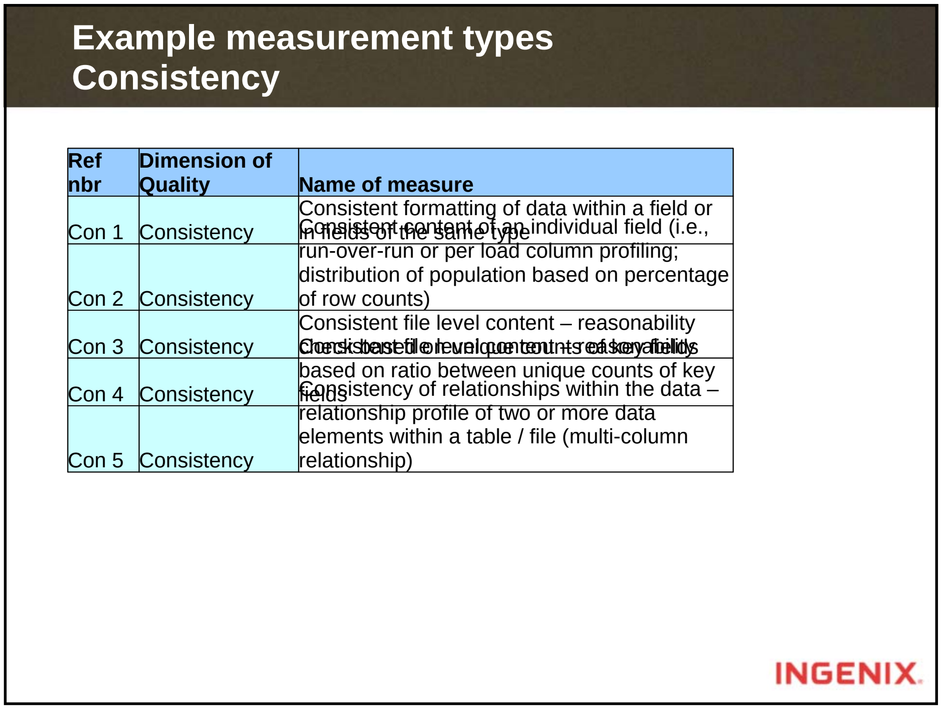
- 10 Completeness
- 16 consistency
- 2 Timeliness
- 10 Validity



Example measurement types Completeness

Ref nbr	Dimension of Quality	Name of measure
Comp 1	Completeness	Process check – all files are available for processing (with version check if possible).
Comp 2	Completeness	File completeness – compared to control record
Comp 3	Completeness	File completeness – reasonability / consistency check comparing size of incoming file to size of past files
Comp 4	Completeness	File completeness – Balance record counts throughout a process, account for rejected records. [For exact balance situations]
Comp 5	Completeness	Field content completeness – Balance Summing – Balance dollar fields throughout a process. [For exact balance situations]

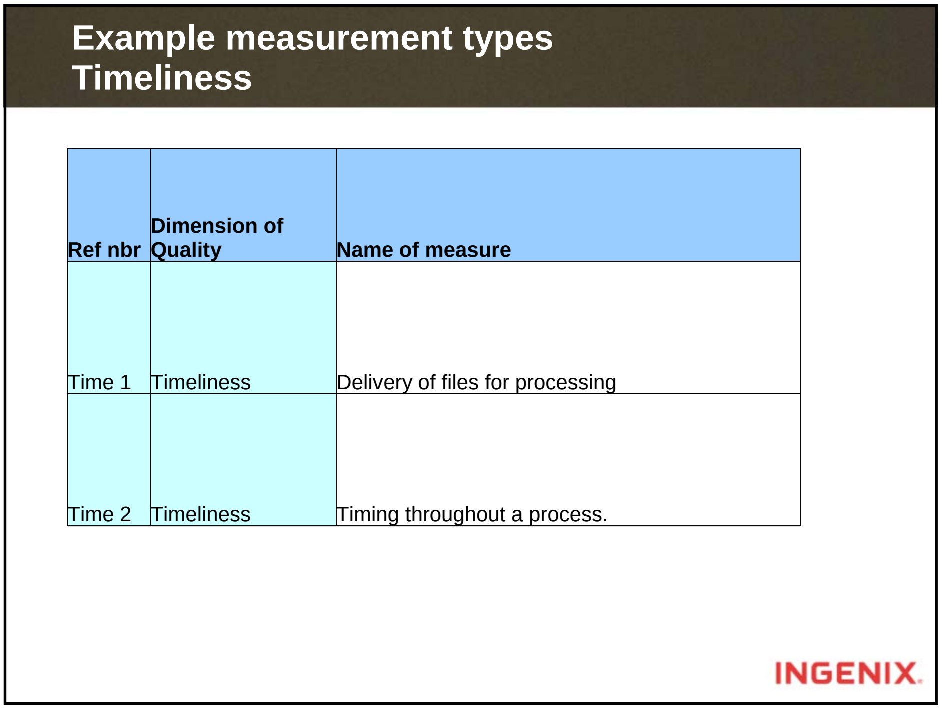
INGENIX.



Example measurement types Consistency

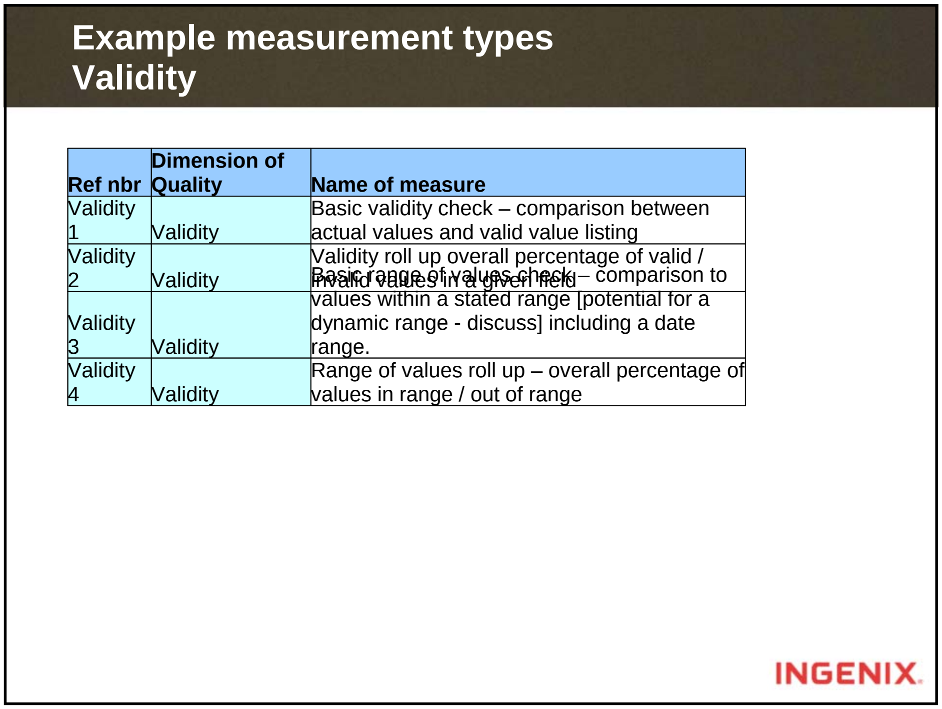
Ref nbr	Dimension of Quality	Name of measure
Con 1	Consistency	Consistent formatting of data within a field or in fields of the same type
Con 2	Consistency	Consistent content of an individual field (i.e., run-over-run or per load column profiling; distribution of population based on percentage of row counts)
Con 3	Consistency	Consistent file level content – reasonability check based on unique counts of key fields
Con 4	Consistency	Consistent file level content – reasonability check based on ratio between unique counts of key fields
Con 5	Consistency	Consistency of relationships within the data – relationship profile of two or more data elements within a table / file (multi-column relationship)

INGENIX.



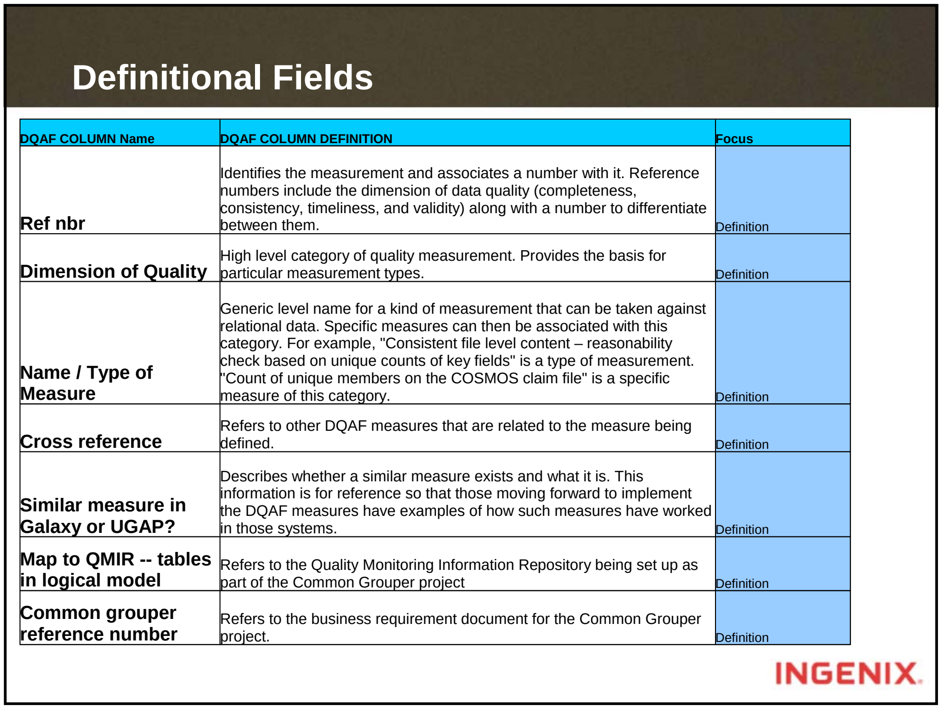
Example measurement types Timeliness

Ref nbr	Dimension of Quality	Name of measure
Time 1	Timeliness	Delivery of files for processing
Time 2	Timeliness	Timing throughout a process.



Example measurement types Validity

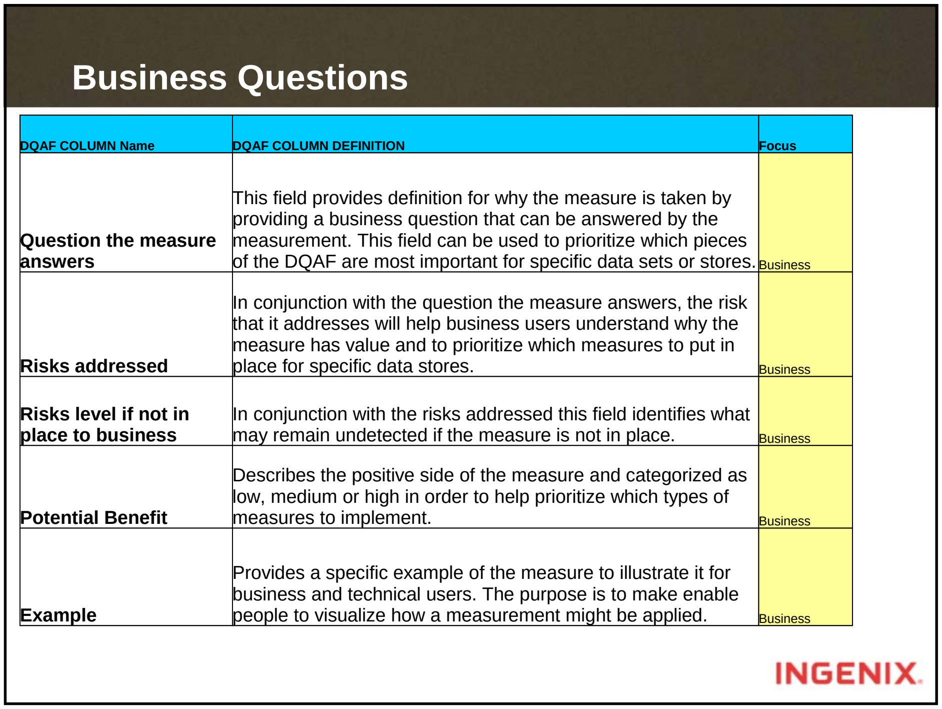
Ref nbr	Dimension of Quality	Name of measure
Validity 1	Validity	Basic validity check – comparison between actual values and valid value listing
Validity 2	Validity	Validity roll up overall percentage of valid / invalid values in a given field
Validity 3	Validity	Basic range of values check – comparison to values within a stated range [potential for a dynamic range - discuss] including a date range.
Validity 4	Validity	Range of values roll up – overall percentage of values in range / out of range



Definitional Fields

DQAF COLUMN Name	DQAF COLUMN DEFINITION	Focus
Ref nbr	Identifies the measurement and associates a number with it. Reference numbers include the dimension of data quality (completeness, consistency, timeliness, and validity) along with a number to differentiate between them.	Definition
Dimension of Quality	High level category of quality measurement. Provides the basis for particular measurement types.	Definition
Name / Type of Measure	Generic level name for a kind of measurement that can be taken against relational data. Specific measures can then be associated with this category. For example, "Consistent file level content – reasonability check based on unique counts of key fields" is a type of measurement. "Count of unique members on the COSMOS claim file" is a specific measure of this category.	Definition
Cross reference	Refers to other DQAF measures that are related to the measure being defined.	Definition
Similar measure in Galaxy or UGAP?	Describes whether a similar measure exists and what it is. This information is for reference so that those moving forward to implement the DQAF measures have examples of how such measures have worked in those systems.	Definition
Map to QMIR -- tables in logical model	Refers to the Quality Monitoring Information Repository being set up as part of the Common Grouper project	Definition
Common grouper reference number	Refers to the business requirement document for the Common Grouper project.	Definition

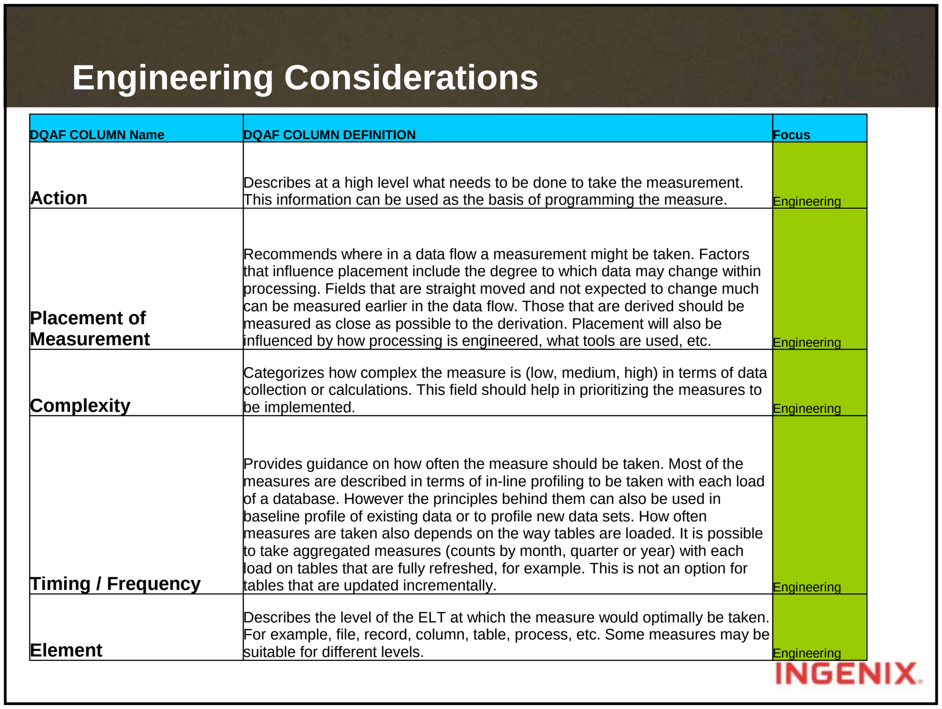
INGENIX.



Business Questions

DQAF COLUMN Name	DQAF COLUMN DEFINITION	Focus
Question the measure answers	This field provides definition for why the measure is taken by providing a business question that can be answered by the measurement. This field can be used to prioritize which pieces of the DQAF are most important for specific data sets or stores.	Business
Risks addressed	In conjunction with the question the measure answers, the risk that it addresses will help business users understand why the measure has value and to prioritize which measures to put in place for specific data stores.	Business
Risks level if not in place to business	In conjunction with the risks addressed this field identifies what may remain undetected if the measure is not in place.	Business
Potential Benefit	Describes the positive side of the measure and categorized as low, medium or high in order to help prioritize which types of measures to implement.	Business
Example	Provides a specific example of the measure to illustrate it for business and technical users. The purpose is to make enable people to visualize how a measurement might be applied.	Business

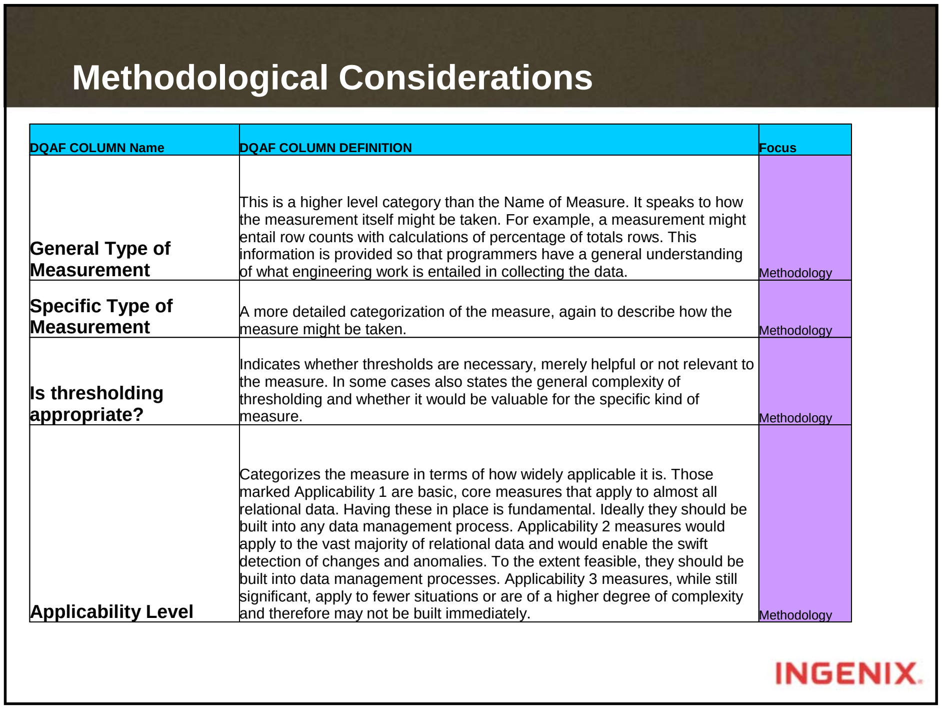
INGENIX.



Engineering Considerations

DQAF COLUMN Name	DQAF COLUMN DEFINITION	Focus
Action	Describes at a high level what needs to be done to take the measurement. This information can be used as the basis of programming the measure.	Engineering
Placement of Measurement	Recommends where in a data flow a measurement might be taken. Factors that influence placement include the degree to which data may change within processing. Fields that are straight moved and not expected to change much can be measured earlier in the data flow. Those that are derived should be measured as close as possible to the derivation. Placement will also be influenced by how processing is engineered, what tools are used, etc.	Engineering
Complexity	Categorizes how complex the measure is (low, medium, high) in terms of data collection or calculations. This field should help in prioritizing the measures to be implemented.	Engineering
Timing / Frequency	Provides guidance on how often the measure should be taken. Most of the measures are described in terms of in-line profiling to be taken with each load of a database. However the principles behind them can also be used in baseline profile of existing data or to profile new data sets. How often measures are taken also depends on the way tables are loaded. It is possible to take aggregated measures (counts by month, quarter or year) with each load on tables that are fully refreshed, for example. This is not an option for tables that are updated incrementally.	Engineering
Element	Describes the level of the ELT at which the measure would optimally be taken. For example, file, record, column, table, process, etc. Some measures may be suitable for different levels.	Engineering

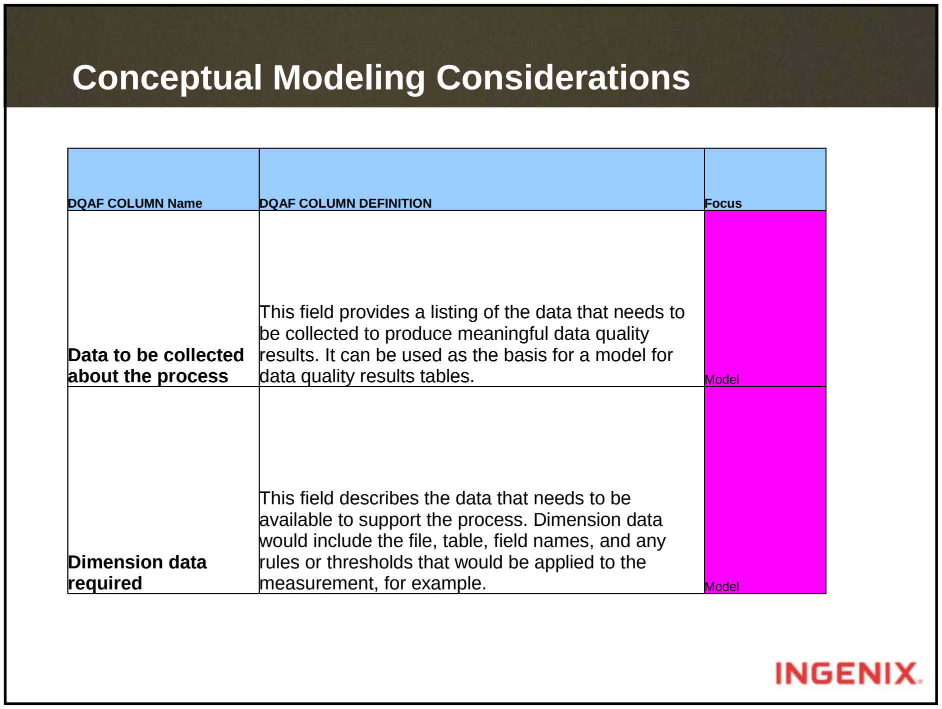
INGENIX.



Methodological Considerations

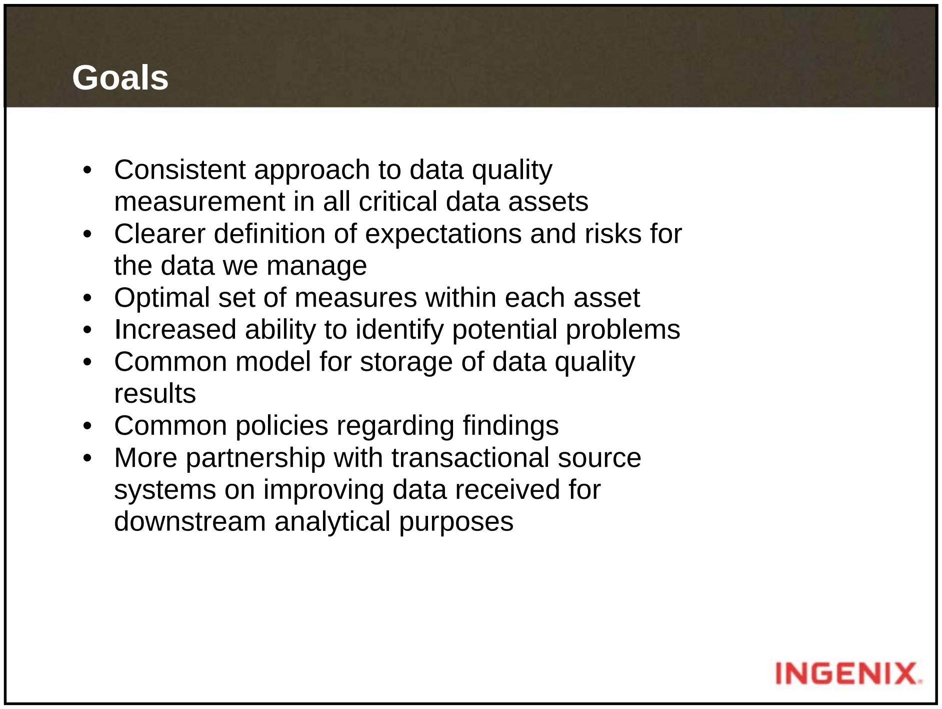
DQAF COLUMN Name	DQAF COLUMN DEFINITION	Focus
General Type of Measurement	This is a higher level category than the Name of Measure. It speaks to how the measurement itself might be taken. For example, a measurement might entail row counts with calculations of percentage of totals rows. This information is provided so that programmers have a general understanding of what engineering work is entailed in collecting the data.	Methodology
Specific Type of Measurement	A more detailed categorization of the measure, again to describe how the measure might be taken.	Methodology
Is thresholding appropriate?	Indicates whether thresholds are necessary, merely helpful or not relevant to the measure. In some cases also states the general complexity of thresholding and whether it would be valuable for the specific kind of measure.	Methodology
Applicability Level	Categorizes the measure in terms of how widely applicable it is. Those marked Applicability 1 are basic, core measures that apply to almost all relational data. Having these in place is fundamental. Ideally they should be built into any data management process. Applicability 2 measures would apply to the vast majority of relational data and would enable the swift detection of changes and anomalies. To the extent feasible, they should be built into data management processes. Applicability 3 measures, while still significant, apply to fewer situations or are of a higher degree of complexity and therefore may not be built immediately.	Methodology

INGENIX.



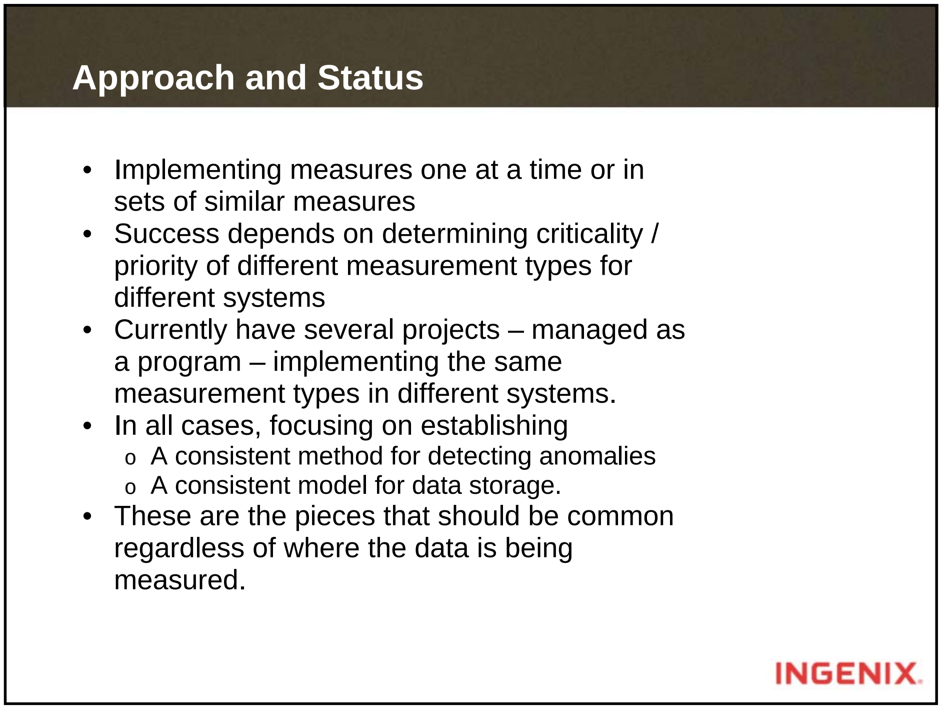
Conceptual Modeling Considerations

DQAF COLUMN Name	DQAF COLUMN DEFINITION	Focus
Data to be collected about the process	This field provides a listing of the data that needs to be collected to produce meaningful data quality results. It can be used as the basis for a model for data quality results tables.	Model
Dimension data required	This field describes the data that needs to be available to support the process. Dimension data would include the file, table, field names, and any rules or thresholds that would be applied to the measurement, for example.	Model



Goals

- Consistent approach to data quality measurement in all critical data assets
- Clearer definition of expectations and risks for the data we manage
- Optimal set of measures within each asset
- Increased ability to identify potential problems
- Common model for storage of data quality results
- Common policies regarding findings
- More partnership with transactional source systems on improving data received for downstream analytical purposes



Approach and Status

- Implementing measures one at a time or in sets of similar measures
- Success depends on determining criticality / priority of different measurement types for different systems
- Currently have several projects – managed as a program – implementing the same measurement types in different systems.
- In all cases, focusing on establishing
  - A consistent method for detecting anomalies
  - A consistent model for data storage.
- These are the pieces that should be common regardless of where the data is being measured.

INGENIX.