Galaxy Data Quality Program
MIT IQ Industry Symposium

16-17 July 2008

Ingenix
United Health Analytics
Galaxy – Shared Data Warehouse
Laura Sebastian-Coleman
IS Manager – Data Quality & End User Support

INGENIX.

Overview

- Background / context
  - Ingenix and Galaxy
  - Galaxy's DQ program
- Value Measurement Initiative – or why we need to measure our SDLC to improve Galaxy's data quality
- Some things never change – or How Galaxy's experience applies to other situations

INGENIX.

Ingenix Background

- A global healthcare information company
- Founded in 1996 to develop, acquire, and integrate some of the nation's best-in-class healthcare information capabilities
- Significant and rapidly evolving portfolio of tools and services now transform data into actionable, fact-based, technology-enabled decision support
- Ranked among the top 10 providers of informatics by *Healthcare Informatics* magazine in June 2006
- Today there is an Ingenix product at work in nearly every U.S. healthcare organization.
- Ingenix is a wholly owned subsidiary of UnitedHealth Group (UHG).

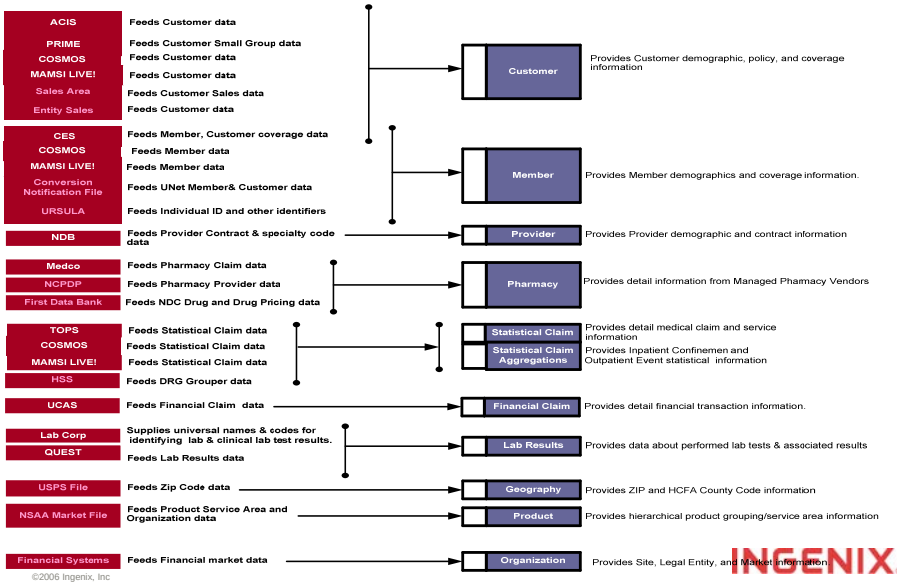
INGENIX.

Galaxy Overview

- Atomic Data Warehouse with transformations
- Integrates data from more than a dozen subject areas (claim, membership, customer, provider, etc.) across multiple sources
- Size
  - 350 source input files from more than 25 distinct internal and external sources (and counting)
  - 15 TB of data; 62 TB footprint
  - 3,438 attributes across 15,069 columns in more than 500 user-facing tables
  - Largest table: more than 1.5 billion rows
    - 1,782,687,382 on Claim Statistical Service as of 3/31/08
- Usage
  - Over 1,000 registered users
  - So far in 2008, averaging more than 450,000 queries / month
  - Ad hoc, scheduled queries, production extracts to applications and marts
  - Direct access to Galaxy via user-selected tools – Sagent is administratively supported

INGENIX.

Galaxy Source Systems & Subject Areas



Last year, I described our current situation

- Galaxy = a mature, enterprise data warehouse
- High demand for data and for organizational services
- Galaxy's DQ program also relatively mature
  - Defined metrics – baseline, semi-annual data profile
  - Automated data collection – complete with alerts, statistically established thresholds on key attributes
  - Regular reporting – post load, quarterly
  - DQ Community – user group
- UHG growing, largely through acquisitions and partnerships
- Healthcare industry changing – relation of government to health care, new products, esp. consumer driven

INGENIX.

And how we would meet new challenges

- Demand for more data from acquisitions
- Demand for faster integrations
- "Common Interface" approach – same structure for incoming data, regardless of source
- "Gateway" to drive consistency across sources
- DQ built into the process

- I was anticipating smooth sailing, since the pieces were all falling into place….

INGENIX.

What we did not count on was

- New, new challenges: evolution of the user community at the same time that demand is increasing for Galaxy data.
- The down side of success
- Revenue model
- Users new to Galaxy
- New employees
- Desire for faster integrations
- New business relationships

INGENIX.

Changes within user community

- Different users of data
- Different uses for data
- Different assumptions about the data
- Different questions about the data
- Different perceptions of the data

- These things throw open the flood gates to problems with the foundational necessity of “fitness for use” as a standard for quality.

INGENIX.

Effort to meet demand

- More projects
- Larger projects
- More complex projects
- New expectations about projects
- New tools, each with a learning curve
- New employees
- Competition for resources, especially “knowledge workers”
- These factors put stress on the organization and especially on the software development life cycle.

INGENIX.

Result

- A negative impact on data quality
- Actual – as defined by DQ metrics
- Perceived – as defined by end user perceptions

INGENIX.

How to Respond? Metrics

- Launched program for new metrics in January 2008
  - Measure where the pain is
  - Project work
    - On time, on budget
  - Project quality
    - Defect tracking
  - Data delivery
    - Are new sources delivering as promised?
- Prevent new pain from emerging
- Continue standard DQ metrics – conformance to business rules, expected populations, etc.
- Continue production database metrics – availability

INGENIX.

Sample metrics

See handouts

- Project delivery
- Project quality
- Data delivery

INGENIX.

What's the take-away? Sticking to Basics

- Data in the warehouse is only as good as data in the source – needs constant vigilance
- Manufacturing model: Data as a product produced through a process – SDLC = a key part of that process
- Measure to improve (not just to measure…)
  - Baseline key SDLC processes (budget, schedule, spec)
  - Keep measures simple –
    - on time, or not?
    - How early, how late?

INGENIX.