Galaxy Data Quality Program MIT IQ Industry Symposium

July 18-19, 2007

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

INGENIX®

Overview

- Ingenix and Galaxy
- Galaxy’s DQ program
- Evolving business needs and the pace of change
- Data quality in relation to evolving business needs

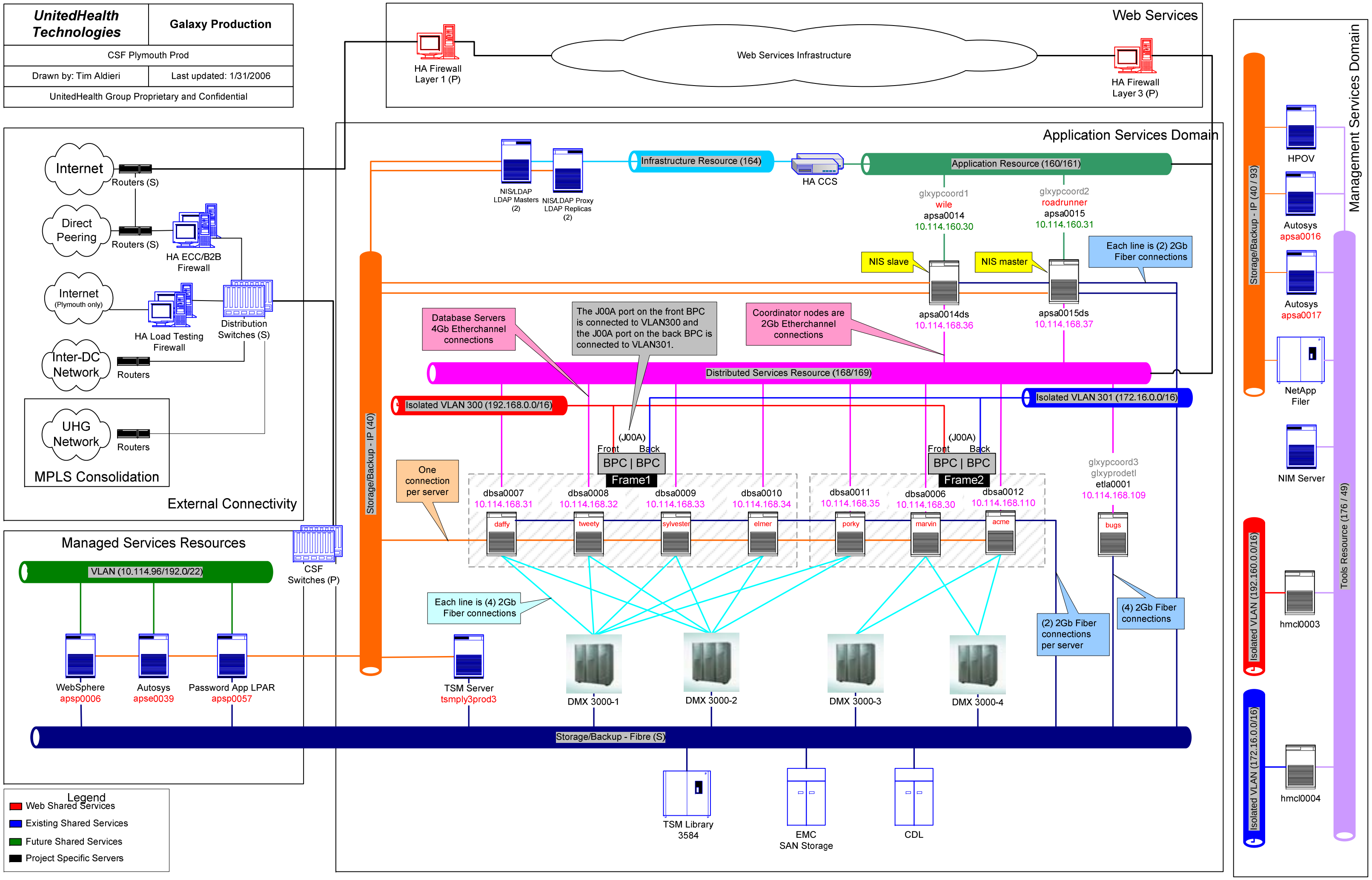
Ingenix Background

- 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).

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)
  - 18 TB of data; 62 TB footprint
  - 3,159 attributes across 12,632 columns in 600 tables (and counting)
  - Largest table: more than 1.5 billion rows
    - 1,704,717,031 on Claim Statistical Service as of 5/3/07
- Usage
  - Over 1,000 registered users
  - 7,888 queries per day / 256,656 per month, on average
  - Ad hoc, scheduled queries, production extracts to applications and marts
  - Direct access to Galaxy via user-selected tools – Sagent is administratively supported

Galaxy Physical Architecture



UnitedHealth Technologies	Galaxy Production
CSF Plymouth Prod	
Drawn by: Tim Aldieri	Last updated: 1/31/2006
UnitedHealth Group Proprietary and Confidential	

Database Servers 4Gb Etherchannel connections

The J00A port on the front BPC is connected to VLAN300 and the J00A port on the back BPC is connected to VLAN301.

Coordinator nodes are 2Gb Etherchannel connections

Each line is (2) 2Gb Fiber connections

One connection per server

Each line is (4) 2Gb Fiber connections

(2) 2Gb Fiber connections per server

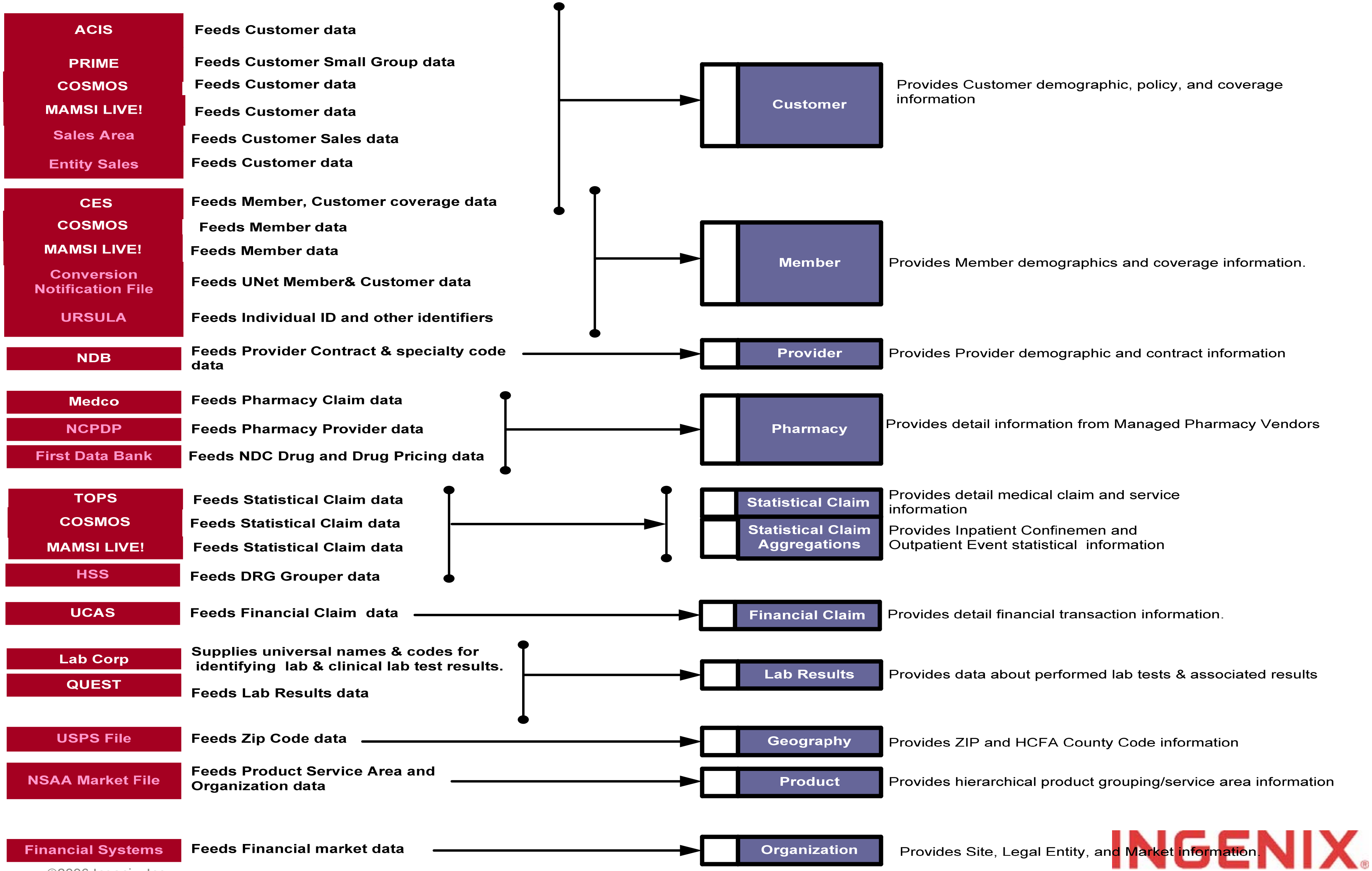
(4) 2Gb Fiber connections

Legend
- 🟥 Web Shared Services
- 🟦 Existing Shared Services
- 🟩 Future Shared Services
- ⬛ Project Specific Servers

System Components

- Hardware
  - 7 IBM P-series Servers P575
  - 2 IBM P-series Servers P510
  - 1 IBM P-series Server P570
  - 4 EMC DMX 3000 Storage Cabinets
  - Additional supporting servers for Sagent, Autosys, etc.
- Software
  - UDB with DPF v8.2
  - AIX 5.3.0
  - DataStage/PX 7.0.1
  - Optiload 3.1
  - CoSort 7.5.3
  - Autosys 4.5
  - Sagent 4.5i

Galaxy Source Systems & Subject Areas



ACIS — Feeds Customer data
PRIME — Feeds Customer Small Group data
COSMOS — Feeds Customer data
MAMSI LIVE! — Feeds Customer data
Sales Area — Feeds Customer Sales data
Entity Sales — Feeds Customer data
CES — Feeds Member, Customer coverage data

Customer — Provides Customer demographic, policy, and coverage information

COSMOS — Feeds Member data
MAMSI LIVE! — Feeds Member data
Conversion Notification File — Feeds UNet Member& Customer data
URSULA — Feeds Individual ID and other identifiers

Member — Provides Member demographics and coverage information.

NDB — Feeds Provider Contract & specialty code data

Provider — Provides Provider demographic and contract information

Medco — Feeds Pharmacy Claim data
NCPDP — Feeds Pharmacy Provider data
First Data Bank — Feeds NDC Drug and Drug Pricing data

Pharmacy — Provides detail information from Managed Pharmacy Vendors

TOPS — Feeds Statistical Claim data
COSMOS — Feeds Statistical Claim data
MAMSI LIVE! — Feeds Statistical Claim data
HSS — Feeds DRG Grouper data

Statistical Claim — Provides detail medical claim and service information
Statistical Claim Aggregations — Provides Inpatient Confinemen and Outpatient Event statistical information

UCAS — Feeds Financial Claim data

Financial Claim — Provides detail financial transaction information.

Lab Corp — Supplies universal names & codes for identifying lab & clinical lab test results.
QUEST — Feeds Lab Results data

Lab Results — Provides data about performed lab tests & associated results

USPS File — Feeds Zip Code data

Geography — Provides ZIP and HCFA County Code information

NSAA Market File — Feeds Product Service Area and Organization data

Product — Provides hierarchical product grouping/service area information

Financial Systems — Feeds Financial market data

Organization — Provides Site, Legal Entity, and Market information.

Functions of Galaxy Data

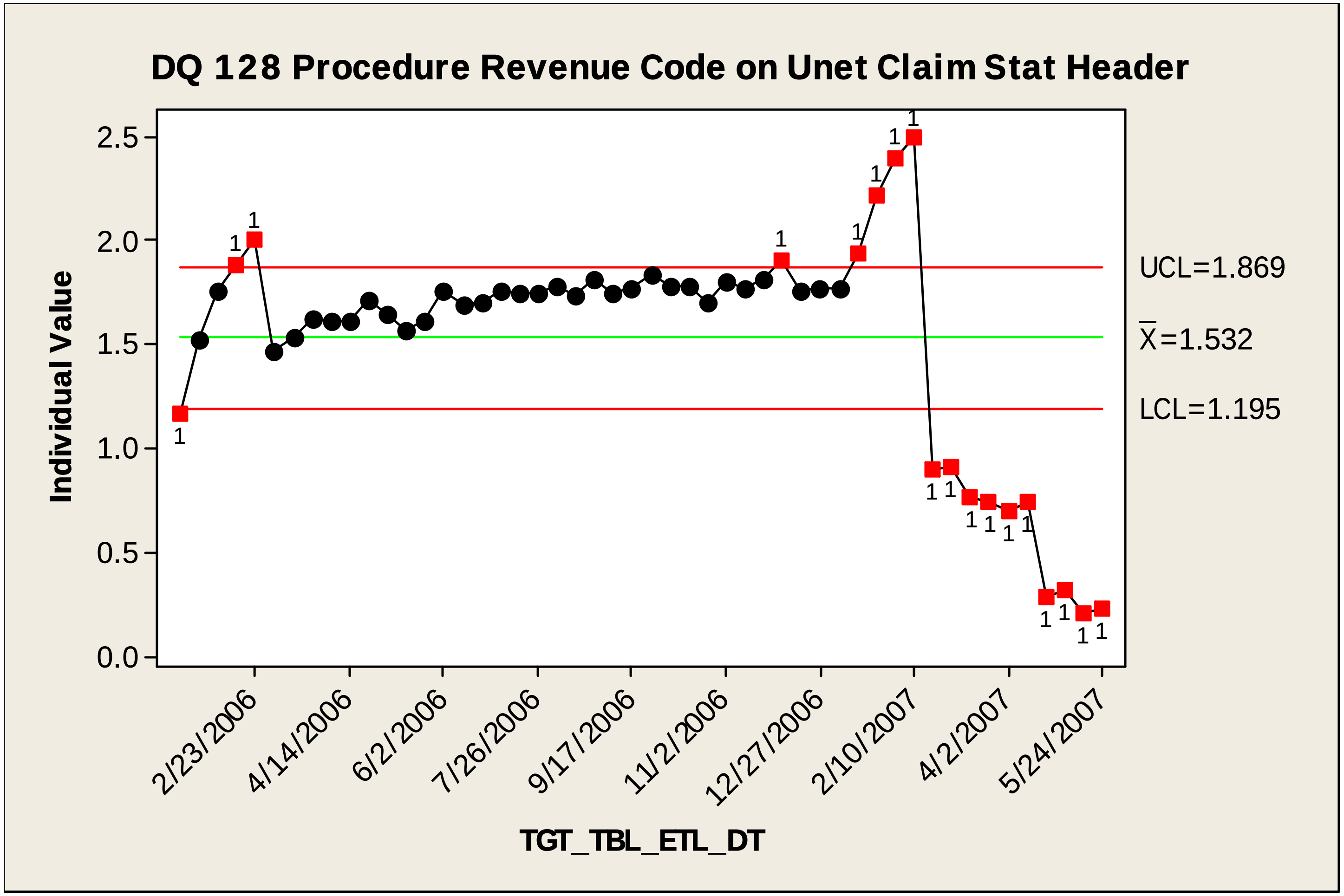
Galaxy is the single source of truth for key business functions

- Medical Trend Analytics
- Pricing
- Provider Utilization & Profiling
- Appropriateness of Care
- Network Adequacy
- Care Management / Pattern of Care / Preventive Care
- Fraud & Abuse
- Customer Reporting
- HEDIS Reporting
- Member Demographics
- Product Penetration

Galaxy’s Data Quality Program

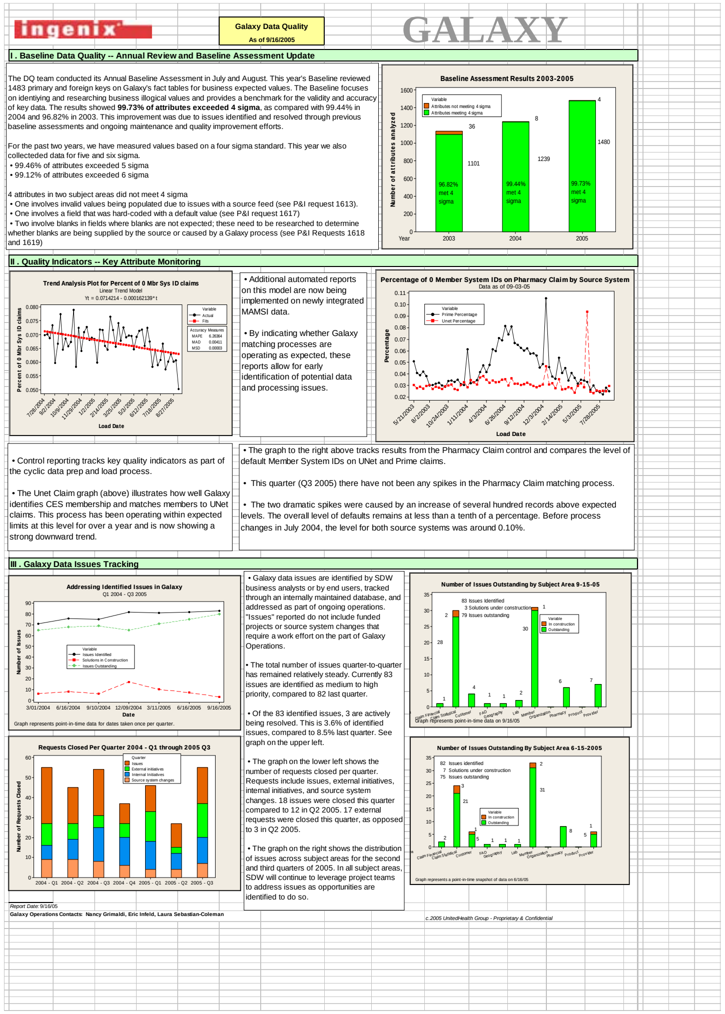
- Management recognized need for DQ when Galaxy was launched
- Theoretical / methodological foundations
  - Correct data problems at the source
  - Data as a product
  - Statistical process control
- Primary functions of DQ program
  - Monitor, measure, and report on Galaxy’s Data Quality
  - Recommend and implement actions based on findings
- Biggest initial challenge = establishing useful metrics
  - What to measure / how to measure
  - How to respond to the results of measurements
- 2003 Initiated metrics & reporting program
- 2004 Implemented first automated measures
- 2004-2007: Deliver weekly/cyclic, monthly, quarterly, semi- annual reporting through largely automated processes

Example of Weekly Measure



Quarterly Management Report

- Baseline Data Quality – annual review and baseline assessment
- Quality Indicators – key attribute monitoring
- Galaxy Data Issues Tracking



Galaxy Data Quality

As of 9/16/2005

GALAXY

I . Baseline Data Quality -- Annual Review and Baseline Assessment Update

The DQ team conducted its Annual Baseline Assessment in July and August. This year's Baseline reviewed 1483 primary and foreign keys on Galaxy's fact tables for business expected values. The Baseline focuses on identiying and researching business illogical values and provides a benchmark for the validity and accuracy of key data. The results showed **99.73% of attributes exceeded 4 sigma**, as compared with 99.44% in 2004 and 96.82% in 2003. This improvement was due to issues identified and resolved through previous baseline assessments and ongoing maintenance and quality improvement efforts.

For the past two years, we have measured values based on a four sigma standard. This year we also collecteded data for five and six sigma.

- 99.46% of attributes exceeded 5 sigma
- 99.12% of attributes exceeded 6 sigma

4 attributes in two subject areas did not meet 4 sigma

- One involves invalid values being populated due to issues with a source feed (see P&I request 1613).
- One involves a field that was hard-coded with a default value (see P&I request 1617)
- Two involve blanks in fields where blanks are not expected; these need to be researched to determine whether blanks are being supplied by the source or caused by a Galaxy process (see P&I Requests 1618 and 1619)

Baseline Assessment Results 2003-2005

II . Quality Indicators -- Key Attribute Monitoring

Trend Analysis Plot for Percent of 0 Mbr Sys ID claims

- Additional automated reports on this model are now being implemented on newly integrated MAMSI data.
- By indicating whether Galaxy matching processes are operating as expected, these reports allow for early identification of potential data and processing issues.

Percentage of 0 Member System IDs on Pharmacy Claim by Source System

Data as of 09-03-05

- Control reporting tracks key quality indicators as part of the cyclic data prep and load process.
- The Unet Claim graph (above) illustrates how well Galaxy identifies CES membership and matches members to UNet claims. This process has been operating within expected limits at this level for over a year and is now showing a strong downward trend.

- The graph to the right above tracks results from the Pharmacy Claim control and compares the level of default Member System IDs on UNet and Prime claims.
- This quarter (Q3 2005) there have not been any spikes in the Pharmacy Claim matching process.
- The two dramatic spikes were caused by an increase of several hundred records above expected levels. The overall level of defaults remains at less than a tenth of a percentage. Before process changes in July 2004, the level for both source systems was around 0.10%.

III . Galaxy Data Issues Tracking

Addressing Identified Issues in Galaxy

Q1 2004 - Q3 2005

Graph represents point-in-time data for dates taken once per quarter.

- Galaxy data issues are identified by SDW business analysts or by end users, tracked through an internally maintained database, and addressed as part of ongoing operations. "Issues" reported do not include funded projects or source system changes that require a work effort on the part of Galaxy Operations.
- The total number of issues quarter-to-quarter has remained relatively steady. Currently 83 issues are identified as medium to high priority, compared to 82 last quarter.
- Of the 83 identified issues, 3 are actively being resolved. This is 3.6% of identified issues, compared to 8.5% last quarter. See graph on the upper left.
- The graph on the lower left shows the number of requests closed per quarter. Requests include issues, external initiatives, internal initiatives, and source system changes. 18 issues were closed this quarter compared to 12 in Q2 2005. 17 external requests were closed this quarter, as opposed to 3 in Q2 2005.
- The graph on the right shows the distribution of issues across subject areas for the second and third quarters of 2005. In all subject areas, SDW will continue to leverage project teams to address issues as opportunities are identified to do so.

Number of Issues Outstanding by Subject Area 9-15-05

Graph represents point-in-time data on 9/16/05

Requests Closed Per Quarter 2004 - Q1 through 2005 Q3

Number of Issues Outstanding By Subject Area 6-15-2005

Graph represents a point-in-time snapshot of data on 6/16/05

Report Date: 9/16/05

Galaxy Operations Contacts: Nancy Grimaldi, Eric Infeld, Laura Sebastian-Coleman

c.2005 UnitedHealth Group - Proprietary & Confidential

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
  - Automated data collection
  - Regular reporting
  - DQ Community
- UHG growing, largely through acquisitions and partnerships
- Healthcare industry changing – relation of government to health care, new products, esp. consumer driven

Pace of Change for Galaxy

- 2004
    - Galaxy integrated data from MAMSI, a United Health Group acquisition
        - Used the existing structure
        - 1+ year to integrate
- 2006
    - Integrated data from three new source systems
    - Developed a new subject area, Revenue
    - Significantly expanded Customer subject area
    - Responded to healthcare industry changes
        - Part D data
        - HRA (Health Reimbursement Account) data
- 2007
    - Integrate data from additional acquisitions
    - Expand the Revenue subject area
    - Continue to support the use and enhancement of existing data.
- 2008
    - Two major integrations already scheduled
    - Potential for several others

Pace of Change for Galaxy DQ

- Biggest challenge
  - 2003 what to measure and how to measure
  - 2007 how to rapidly analyze and act on DQ data
- Baseline Assessment of Galaxy Data Quality
  - 2003
    - 800 person hours to pull and analyze data for first Baseline Assessment
    - Duration = more than 3 months
    - Measured 1137 attributes
  - 2006
    - Pulled 75% of data in less than 10 hours through an automated process
    - Measured 1506 attributes
    - Pull data quarterly
- Automated reports
  - 2004: 4 reports
  - 2007: 80 reports
  - Reports now implemented as part of standard development process.

2007 – 2008 Key UHG Business Needs

- UHG acquisitions and partnerships –
  - More data for Galaxy
  - More users need access
- Users need data sooner –
  - Time to integrate data into Galaxy must be shortened
- Legacy data critical for ensuring reporting continuity and analytics –
  - Continued support is necessary
- Data consistency across sources critical for reporting continuity and analytics –
  - Integration methodologies need to promote and enforce consistency

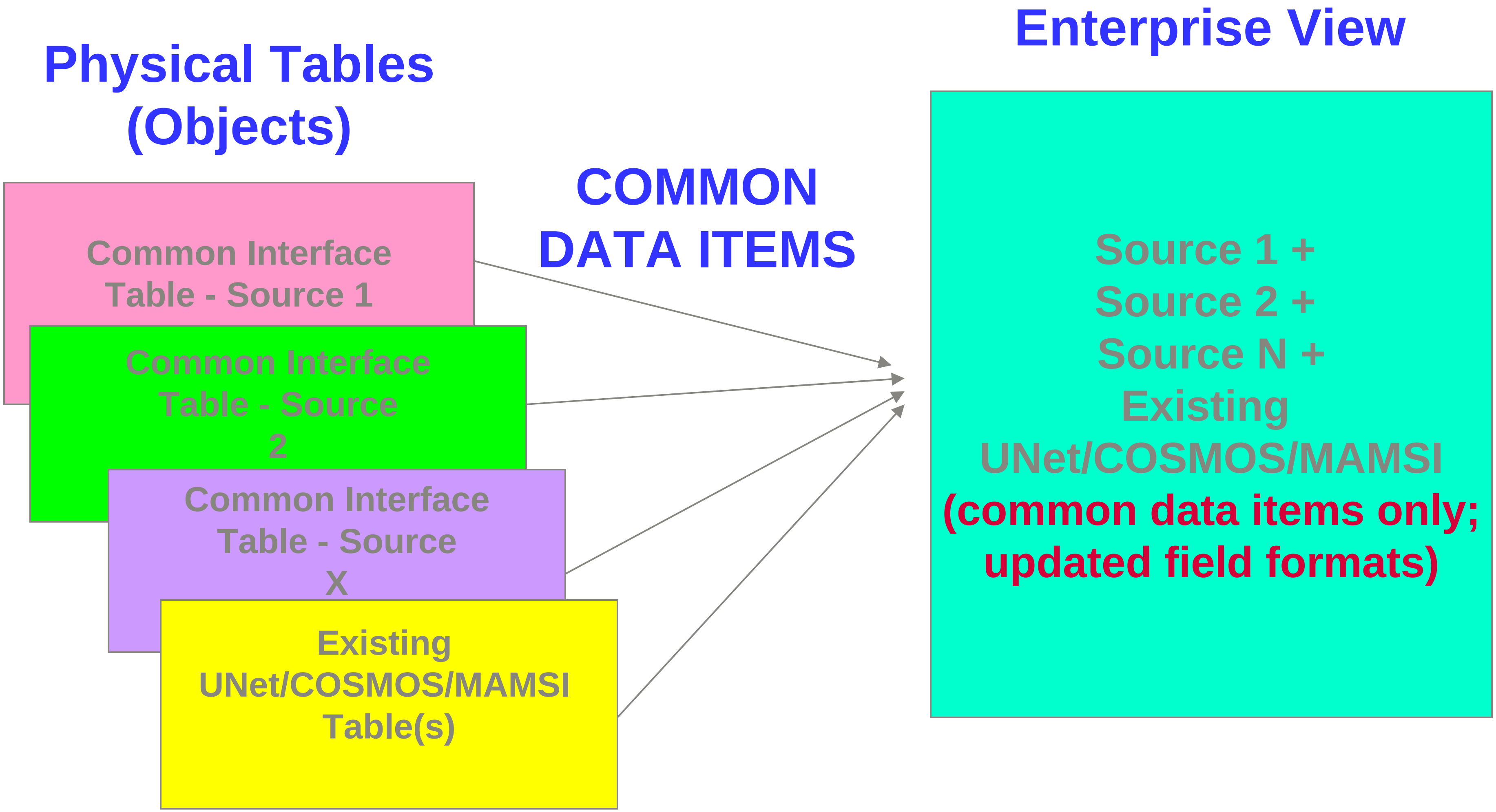
How to Respond?

- Data Quality included in set of changes to improve efficiency and agility
  - Common Interface – puts more responsibility on source systems for data quality
  - Gateway – changes how Galaxy prepares data.
- DQ measures
  - More comprehensive
  - Taken earlier in the process
  - More fully automated

Common Interface Approach

- Galaxy defines standard requirements and layouts for data
- Sources map to these requirements and feed to Galaxy
- Streamlined transformation/load into Galaxy
- Common model across the enterprise

Common Interface Architecture – Views



Physical Tables (Objects)

Common Interface Table - Source 1

Common Interface Table - Source 2

Common Interface Table - Source X

Existing UNet/COSMOS/MAMSI Table(s)

COMMON DATA ITEMS

Enterprise View

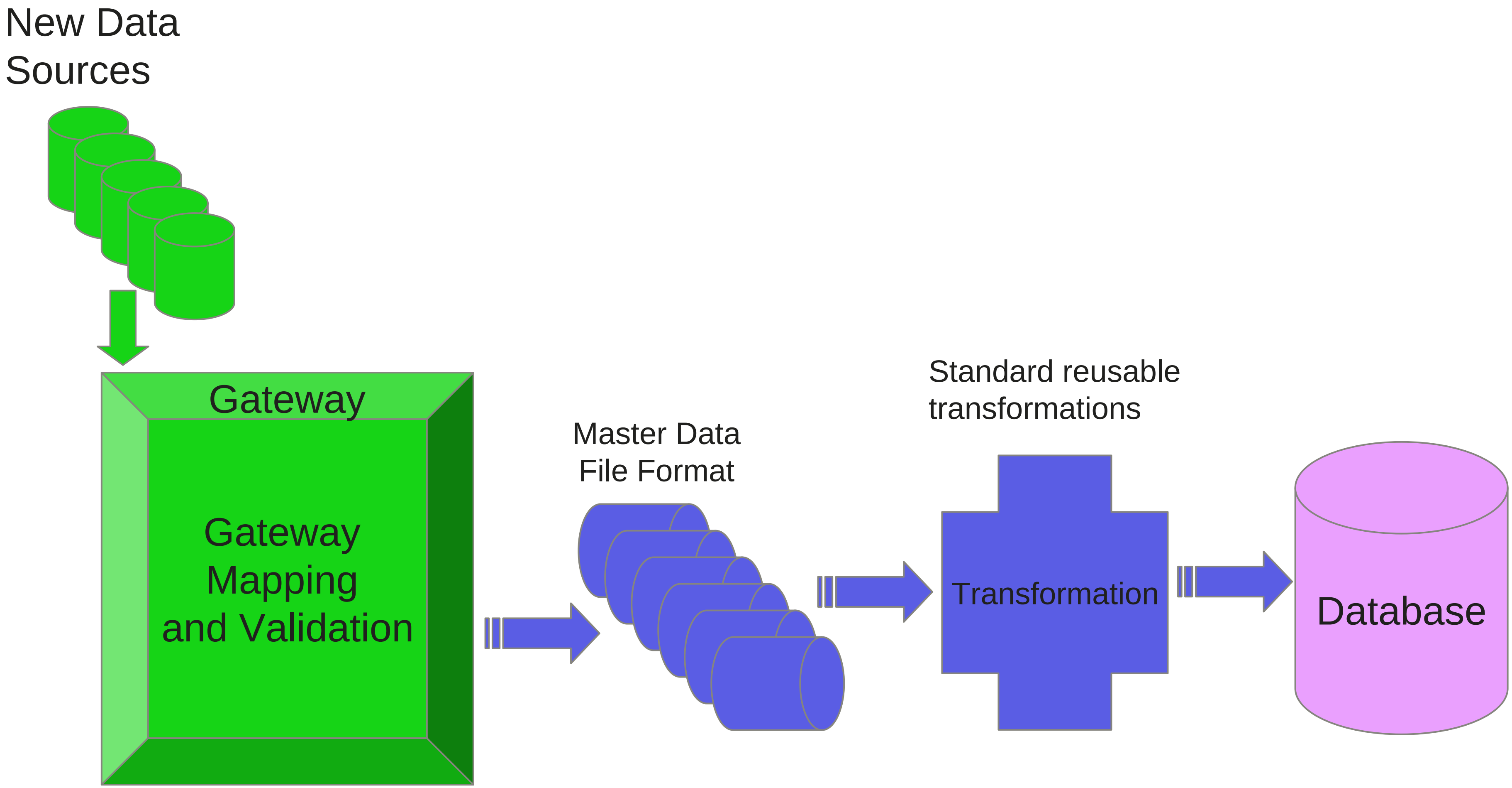
Source 1 +
Source 2 +
Source N +
Existing UNet/COSMOS/MAMSI
(common data items only; updated field formats)

Gateway Integration Tool

- Facilitates mapping disparate data sources into a Master Data Definition
- Applies generic transformation logic to the output
- Utilizes reusable transforms
- Performs automatic code generation
- Ensures consistency across source-to-target mappings
- Provides true-to-code documentation
- Incorporates data quality modules
- Increases speed and reduces complexity of data integrations

INGENIX®

Gateway Integration Tool



New Data Sources

Gateway

Gateway Mapping and Validation

Master Data File Format

Standard reusable transformations

Transformation

Database

©2006 Ingenix, Inc

Gateway – Data Quality Features

- DQ functions
  - Monitor and react to events in processing
  - Collect trend data
- Field validation
  - Data type checking
  - Value range checking
  - Valid value list checking
  - Assignment of default values
  - Informational, error and warning messages
- File validation
  - Format checking
  - Field counts / record length validation
  - Summary of field error and warning messages
  - Thresholds of summary counts of errors and warnings that allow job to be aborted if counts or percentages exceeded – generate alerts

tdwi
BEST PRACTICES
AWARD
2007

Back to Basic DQ

- Data in the warehouse is only as good as data in the source
  - Ensuring sources to supply better data through the Common Interface
- Manufacturing model: Data as a product produced through a process
  - Executing processes more consistently across the database through the Gateway
- Measure to improve
  - Gateway integrates and executes DQ measures consistently across the database.
  - Both tools measure ETL processes (timing of jobs, etc.) that affect other aspects of data quality from end-to-end

DQ: Chicken or Egg?

- After 4 years – back to the beginning
  - Applying theory/methodology more fully
  - Applying at the beginning of integrations
  - Applying more comprehensively across the warehouse
- Major re-thinking of all Galaxy processes
  - Interacting with customers
  - Writing specifications
  - Obtain source files
  - Mapping source-to-target
  - Implementing ETL
  - Building physical tables
  - Taking DQ measures
- DQ still requires championing
- New problem: How to analyze and respond to findings from the data gathered through new process.