Real-time Data Quality
The Right Answers at the Right Time

Pat Herbert
R&D Director
DataFlux – A SAS Company
DataFlux – A SAS Company

- DataFlux ([www.dataflux.com](http://www.dataflux.com)) is a wholly owned subsidiary of SAS.
- DataFlux is a leading provider of data quality integration solutions.
- DataFlux enables organizations to analyze, improve and control their data through an integrated technology platform. Through its enterprise data quality integration solutions, companies can build a solid information foundation that delivers a unified view of customer, product or supplier data.
Data Integration – A Historical Perspective

- Data Management has been an issue since the first files were created.
- SAS has been a leader in the Data Management space for over 30 years.
- ETL (Extract-Transform-Load) was a term that became a definition for a range of applications and processes that were used to create data warehouses and data marts.
- SAS produced ETL Studio which became Data Integration Studio as it took on more of the capabilities needed in the broader Data Integration space.
Data Integration – A Historical Perspective

- Data Quality issues arose and it became clear that the data that was stored in data warehouses and data marts needed to be of high quality in order to deliver useful, reliable results.

- SAS acquired DataFlux for its Data Quality expertise and technologies.

- ETL Studio/Data Integration Studio and the SAS Data Quality Server provide Data Quality capabilities as a result – embedding and leveraging technologies from DataFlux.
Data Integration – A Historical Perspective

DataFlux (A SAS Company) was positioned by Gartner, Inc. in the Leaders Quadrant – the highest position a software provider can achieve - in the “Magic Quadrant for Data Quality Tools, 2006” report.
Data Integration – A Historical Perspective

The Data Quality Methodology from DataFlux is fundamental to most Data Quality initiatives – Batch or Real-Time.
Data Integration – A Historical Perspective

- Data Integration as a term was coined to cover all aspects of data acquisition and delivery within an enterprise – from the real-time/transactional systems on the front-end, to the batch-oriented ETL processes, all the way through to the decision support systems that analyze and present data on the back-end.
Data Integration – A Historical Perspective

- SAS has a complete product set that covers the broad Data Integration landscape.
Data Integration – A Historical Perspective

- Data Quality initiatives are being pushed from the back-end and batch systems into the front-end systems.
- Real-Time Data Quality was born…
Data Integration
What is Real-Time?

- It’s really all about acceptable latency or lag between an event and a reaction.
- All things have some degree of latency.
- Extremely low latency measured in terms of the speed of light across a short distance (nanoseconds).
- High latency measured in seconds, minutes, hours, days, or longer.
- We are interested in the things between.
What is Real-Time?

- Physical or Hard real-time systems require tasks to be completed in a specified timeframe (usually 100th of a second or faster)
  - Pace maker, EKG, medical monitoring systems
  - Aircraft flight controls, missile guidance systems
  - Vehicle dashboard, ABS, telephone switches,…
What is Real-Time?

- Operational or Soft real-time systems require tasks to be completed as fast as possible but typically, the timeframe requirements are much longer (sub-second up to a second or more) or non-critical
  - POS (UPC scanner, price lookup), Air carrier ticketing systems, GPS systems
  - Decision support dashboard (investment trading)
  - Call center applications
  - ERP systems, Web-based order-entry, Billing, Inventory, …
What is Real-Time?

- Non real-time systems – sooner is better but measured in minutes, hours, days, weeks or even quarterly or more.
  - POS (manual, prices stored nightly)
  - Decision support dashboard (corporate management)
  - Change data capture parts of an ETL process
  - ETL batch processes – they still have time constraints but they don’t typically have someone waiting on them
Goal of Real-Time Data Quality: Better Answers, Faster

- Better answers are more accurate, more consistent, more reliable, more complete.
- “Better” is not necessarily “best” or “perfect”.
- Perfect answers may require more interaction that may interrupt the process.
Goal of Real-Time Data Quality: Better Answers, Faster

- Faster answers allow for information to be delivered more quickly.
- “Faster” does not necessarily mean “fastest” or “immediate”.
- How fast is fast enough? You have to weigh cost vs. need.
- Do you need the fastest computer money can buy or one that is “reasonably” fast?
- We almost always settle on “reasonably” fast and let cost help determine our definition of “reasonable”.
Goal of Real-Time Data Quality:
Better Answers, Faster

- We need to identify the sweet-spot and aim for that.
- We want both (Better and Faster) and often to an extreme.
- As we look at real-time Data Quality initiatives, we must to weigh cost (of implementation and cost reductions) against time to get answers while working to get better results.
Goal of Real-Time Data Quality: Better Answers, Faster
Why Real-Time Data Quality Now?

- Technology has improved in several areas.
  - Service Oriented Architecture (SOA)
  - Networks
  - CPU speeds
  - Data storage/retrieval speeds
Why Real-Time Data Quality Now?

- Data management trends moving toward shorter times to answers. Correcting data issues earlier in the process helps (better answers faster).

- It is becoming more “reasonable” (less costly) to get better answers, faster. “Faster” can now be real-time at a reasonable cost.

- We are all looking for competitive advantages. If you can provide better answers, faster than your competitors, you will have a real advantage over them.
What is the SAS® DataFlux® Real Time Data Integration Server?

- **Service Oriented Architecture (SOA) based server that executes Data Quality web services in real-time.**
  - SOA design is an open, standards based implementation.
  - Can be called from SAS, SAP, Java, C, C++, C#, or any other web service enabled “client”.
  - Write once, deploy once, run from anywhere.
  - Data Quality business rules can be shared by all applications that are web service enabled.
  - Data Quality business rules can also be shared by batch jobs and other processes – allowing for improved performance in large batch processing environments.
What is the SAS® DataFlux® Real Time Data Integration Server?

- Very efficient at providing small/simple services as well as large/complex services and anything in-between.
- Supports multiple locales.
- Inherently scalable and fault tolerant.
- Runs on many OS platforms (Windows, Unix, Linux).
- Can be secured to prevent unauthorized use.
What is the SAS® DataFlux® Real Time Data Integration Server?

- Data Quality web services are defined in easy to use graphical user interface design environment (dfPower Architect).

- Data Quality web services are deployed, managed and tested via management application with graphical user interface (Integration Server Manager).
What is the SAS® DataFlux® Real Time Data Integration Server?
What is the SAS® DataFlux® Real Time Data Integration Server?
What is the SAS® DataFlux® Real Time Data Integration Server?
Advantages of Real-Time Data Quality

- Advantages of Real-Time Data Quality as provided by the SAS® DataFlux® Real Time Data Integration Server
  - Reduced costs
  - Better overall system performance
  - More accurate, consistent and reliable results
  - Simplified data entry
  - Improved customer satisfaction
  - Less time to better decisions
  - Competitive edge
Advantage: Reduced costs

- Business rules can be defined and implemented in one place rather than needing several implementations that all must be maintained thus reducing the cost of system maintenance and management.

- Example: SAP system (used for 2 different purposes and customized by 2 different departments), Siebel system, three internally developed systems, and a system from a vendor that requires outside consulting services to maintain.

- Reduced burden on the IT infrastructure because data quality rules are executed in one place rather than multiple times and/or in multiple locations after erroneous data has been replicated throughout your systems.
Advantage: Reduced costs

- Only new data has to be cleansed and that reduces IT costs to an even greater degree because typical ETL systems perform the same data cleansing tasks on all data in the system, not just the new data.

- System management is simplified with the Integration Server Manager – a graphical user interface application that allows for centralized management and deployment of Data Quality services.
Advantage: Better overall system performance

- Nightly ETL jobs can finish in less time. This allows for the data to be cleansed in a distributed fashion before the ETL processes run.

- Data storage systems are not accessed as heavily because data is cleansed prior to storage. Reduced access to data storage systems means, better transaction rates, more time for more transactions or less new equipment to support more transactions.

- ERP systems see performance improvements because some of the work needed can be offloaded to other systems.
Advantage: More accurate, consistent and reliable results

- More accurate results when data quality is addressed in real-time, as it enters the system because issues can be resolved immediately while there is context and/or while originator of the bad data is available to correct it.
  - Context example: data is being entered from a batch of records, all for a given family and the family name is spelled incorrectly in one of the records. If the error is flagged immediately, the person entering the data can correct it because the other records provide the hint about what the proper spelling is.
  - Originator of data example: a web form that prompts for a phone number can reject updates if the phone number does not have enough or has too many digits or if it has letters, ...

- More consistent results: Fixing data quality issues in real-time, as it enters the system, means that it can't be replicated in a bad form.

- More reliable results because standards applied more readily with less impact on the other systems. Data quality rules can be improved more aggressively.
Advantage: Simplified data entry

- Only information that is necessary has to be entered while superfluous values can be omitted – allowing the data quality rules to complete missing values. This not only simplifies data entry, it also speeds up the process of entering the data and improves the overall quality of the data at the same time.

- Doing this at point of data entry provides feedback about if enough info was provided rather than having partial information delivered to back-end systems that must be remediated later and/or multiple times before use.
Advantage: Improved customer satisfaction

- Data can be checked as it is being provided for accuracy and correctness and the customer can be prompted if there are errors. This means fewer errors on the back end; fewer delays in order processing; fewer calls back to the customer.

- Customers can be more readily identified with less specific information due to robust identity management and matching capabilities.
  - Customers feel more like they are “remembered” or “recognized” by the company they are doing business with.
  - Customers don’t have to provide exact spellings and data does not have to be entered perfectly in order to locate information about the customer in question.
Advantage: Improved customer satisfaction

- Existing customer information can be used to fill in fields so that the customer does not have to provide the same information multiple times.
- Faster transactions result from reduced data entry and fewer required fields which makes the provider seem more responsive to the customer.
- Note: same improvements can be applied to non-customer information but customer specific issues are often the driving force in the desire for more real-time capabilities. This is becoming especially true as more customer self-service applications and systems are created.
Advantage: Improved customer satisfaction
Advantage: Improved customer satisfaction
Advantage: Improved customer satisfaction
Advantage: Improved customer satisfaction
Advantage: Improved customer satisfaction
Advantage: Less time to better decisions

- Just another way to say “better answers, faster” with a focus on the Business Intelligence and/or decision support.
  - Because the quality of data is improved as it enters the system, decision support applications can get to better data sooner.
  - Because the ETL processes can run more quickly, the Business Intelligence/decision support applications can start accessing the data in the data warehouse sooner which means reports with better data can be delivered faster.
Advantage: Competitive edge

- If you can provide better answers, faster than your competitors, you will have an advantage over them to...
  - Win the hearts of your users and your customers.
  - Win the praise of your employers and your investors.
  - Win the battle with your competition.

- Application of Real-Time Data Quality capabilities will give you a competitive edge.
Can you have too much of a good thing?

- Non-repeatable results [expand](this can be overcome to some extent with good practices [expand])
- Larger burden on the operational systems [expand](this can be overcome with good choices and better operational systems [expand])
- Higher costs [expand](can be overcome by improving the right parts of systems rather than trying to modify the entire system to be 100% real-time [expand]).
Are you ready for Real-Time Data Quality?

- Don’t introduce Real-Time Data Quality just because you can.
- Build Real-Time Data Quality plans around business objectives but pay attention to technical limitations/restrictions.
- Do what is reasonable.
- Balance is the key – balance between speed, quality and cost.
- The SAS® DataFlux® Real Time Data Integration Server is uniquely positioned to deliver on the promise of Real-Time Data Quality now. If you are ready, so are we!
Questions?

- As time permits.

- Contact info:
  - Pat.Herbert@DataFlux.com
  - +1.919.677.8000
see you in Stockholm 22-24 May 2007!
www.sas.com/sasforuminternational

SAS® FORUM INTERNATIONAL
THE PREMIER INTERNATIONAL EVENT FOR ENTERPRISE BUSINESS INTELLIGENCE