Paper 079-2013

In-Database Data Quality – Performance for Big Data

Charlotte Crain, Mike Frost, and Scott Gidley, SAS Institute Inc., Cary, NC

ABSTRACT

Data quality and high performance have joined forces. Today is an era of big data, extremely large data warehouses and potential security issues for moving data. Traditional data quality is performed with an ETL-like operation of extracting, processing, and publishing back to the source. Performance, or a potential for security issues associated with moving data, requires a call for a new approach. This paper explains how SAS® data quality functions can be invoked in-database, eliminating the need to move data and thus delivering data quality that meets the need for near real-time performance for today's business. Graphic results comparing performance metrics of traditional data quality operations against in-database data quality will be presented along with details illustrating how these results scale with database resources.

BUSINESS EVOLUTION – BIG DATA

Did you know in 2011 humans created 1.8 zettabytes of data? That would be equivalent to 200 billion high-definition movies that are at least 120 minutes long. It would take one person 47 million years to watch all those movies. All of this data that humans create doubles every two years? [1] "Big Data is the frontier of a firm's ability to store, process, and access … all the data it needs to operate effectively, make decisions, reduce risks, and serve customers. There are varying definitions of what "Big Data" is relative to a given context. From a pragmatic or practical viewpoint, a definition of big data must be actionable for *both* IT and business professionals**.** A couple questions of big data are the following:

Store: Can you capture and store the data?

Process: Can you cleanse, enrich, and analyze the data? "[2]

Historically, data has been collected and managed at the level of individual departments for their own needs. Each department developed procedures, data formats, and terminology that fit its unique situation and preferences. As long as there was no need to integrate or exchange the data, inconsistencies in the data from department to department were harmless. Today, however, mission goals and legal mandates require large organizations to report on their activities at the enterprise level, driving a need for consistency and accuracy across the organizational structure.

In addition to legacy systems being integrated or migrated, organizations are realizing the game-changing potential of analyzing the vast amounts of data that they generate or collect. This is driving a new desire to continuously organize and archive extremely large data sets of information - an approach and mechanism both referred to as big data.

Today, organizations use a variety of so-called big data solutions to collect data from sources as diverse as social media, RFID sources, SCADA, electronic metering devices, and machine and software logs. Applying traditional data management processes and methods to these vast amounts of information managed by a big information store can be impractical if the time that is needed to apply them results in latency that limits the usefulness of the information. For example, if it takes several hours to run a process that can be used to trace the source of a series of negative commentary on social media back to the particular customer sales transaction or service call that initiated it, the damage to the organization's brand or reputation is difficult to recover from.

DATA QUALITY – ESSENTIAL FOR DATA MANAGEMENT

Within the realm of data management, data quality is generally defined by attributes such as state of completeness, validity, consistency, timeliness, accuracy, and auditability that makes data appropriate for a specific use.

Each attribute is defined with some examples. None of the data quality dimensions is complete by itself, and many times dimensions are overlapping. [3]

- **COMPLETENESS**: The expected data attributes are provided. It is possible to have data not available for a given data element, yet it is still is considered complete because it is not needed by the organization or cannot be collected during a certain time period. [3]

- **VALIDITY:** Are the data correct? This is critical for the "data trust factor" for any organization. [3]

- **CONSISTENCY:** Values, formats, and definitions are consistent across data warehouses or the enterprise or that the data systems across the enterprise are synchronized

  Example of data in-consistency:
  A credit card is canceled and inactive, but the card billing status shows 'due'. The data is inconsistent, when it is in sync in the narrow domain of an organization, but not in sync across the organization. *Data can be complete, but inconsistent* [3]

- **TIMELINESS**: The extent to which information is readily available for a business process, report, or analysis is extremely important because ***"data delayed is data denied"*** [3]**.**

  Example of Timeliness:
  Customers service providing up-to date information to the customers. The timeliness depends on user expectation. An online availability of data could be required for a room allocation system in hospitality, but an overnight data is fine for a billing system. [3]

- **ACCURACY**: Does the data represent "real-world" values, objects, or events expected to fulfill business requirements?

  Example of Accuracy:
  Correct spellings of contact information such as person names, addresses, phone. However; this does not imply that the information is accurate. [3]

- **AUDITABILITY**
  Auditability means that any transaction, report, accounting entry, bank statement, and so on, , can be tracked to its originating transaction. This would need a common identifier, which should stay with a transaction as it undergoes transformation, aggregation and reporting.

  Example of non-auditable data:
  A surgery report cannot be linked to the Doctor ID of preliminary diagnosis OR the pathologist ID. [3]

Data quality problems across industries are not just an IT problem or a business department problem. It is an organization's problem. Examples of the impact of these problems include but are not limited to the following:

- Missed service level agreements for certain operations which specify the roles and responsibilities pertaining to management and delivery of data quality expectations
- Avoiding the use of data quality processes for situations where missing service level agreements is not an option
- Not applying data quality processes as frequently as ideal
- Legacy applications that continue to produce more of the same poor quality data
- Data quality applied on an ad hoc basis, leading to inconsistent application from department to department, person to person, or even run to run.
- Manual and redundant effort to perform repetitive data quality

All of these examples result in data quality issues that compromise the results of reporting, analysis, forecasting, and other key business requirements or they impede business processes thereby reducing the timeliness and accuracy of decision making. The lack of prompt and accurate data quality also impedes opportunities for analytics modernization or acceleration of building new models and knowing the customer better and faster. Despite the obvious benefits, one of the reasons for not applying data quality reliably can be related directly to performance.

BIG DATA QUALITY IN TODAY'S BUSINESS AND TECHNOLOGY ENVIRONMENT

"Did you know that bad data or poor data quality costs U.S. businesses $600 billion annually?"[4]

Because of this fact, what is the best way to perform data quality and have extremely large amounts of data that is sitting in a data warehouse such as Teradata or on a Hadoop file system? Driving this need is a critical business requirement for a new approach to knowing our customers and making effective decisions quickly and correctly. What technologies for data quality and big data can help me to stay competitive, retain and increase customer satisfaction by literally demonstrating that we actually "know our customer"? What technologies for data quality and big data can help me respond to regulatory requirements, meet service level agreements for internal and/or external customers, create efficiencies to free up skilled individuals who can then use their expertise for better or more sophisticated business analysis or automated processes?

The attribute or dimension of data quality at stake for big data is one of time or timeliness where we know that "data delayed is data denied". In today's fiercely competitive and regulated environments, this is not acceptable and a need to change how the business operates is in order. The speed of data quality enables IT to meet time sensitive, data quality-related service level agreements and open up other areas where traditional data quality solutions have not been used or used often enough because the compute and process time was prohibitively slow.

TECHNOLOGY ENVIRONMENT

Today, hardware platforms allow for up to 16 exabytes of memory (16GB to 512TB is typical usage) and inexpensive storage systems are used in big data environments. As more data are collected and stored, the hardware supports ever demanding business needs to process and analyze tremendously large amounts of data. "Modern database systems often contain a large number of powerful processors, optimized disks, and fast network connections, along with highly developed software that optimizes query processing and manages security, users, queues, and priorities. Because organizations have invested large sums of money in the acquisition and maintenance of these systems, the organizations need to leverage their resources as much as possible".[5]

THE RIGHT DATA AT THE RIGHT TIME

Decision support is having the just enough information to make the right decision …real-time decision support means "getting the right information, to the right people, just in time." The data must be good enough with the right level of data quality and it must be on-time, complete, and factual. Just as the pendulum swung in the late 1980s from mainframes to personal computers, these real-time messages must be tempered … with what is right for our organization…[6] A corollary follows that we need ways of managing data quality for big data in order to believe it to be trusted for right time decision making.

If we use big data in a timely fashion, then how is it possible to perform data quality on big data in an efficient, low network traffic, high-performance environment when moving vast amounts of data over a network and then performing traditional data quality on a compute server and moving results again back over the network is not an option? One way to help achieve this is to push the work of data quality down to the data, so limitations associated with performance does not become a roadblock or bottleneck for data quality being applied to big data.

In 2013, SAS will release the first-ever product for delivering on this promise - data quality functions that can be invoked within the database and called in-database without the need for extraction of data. SAS will initially provide this in-database data quality with Teradata.

IN-DATABASE PROCESSING

SAS in-database processing delivers better answers faster by moving relevant data processing closer to the data and improving integration between SAS and database management systems.

This is accomplished by exposing the data quality functions as stored procedures inside the database. This can add value to your use of SAS by reducing data movement, improving performance runtimes, and speeding up analytic development and deployment. Performing data quality inside the database can remove the data quality efforts being done downstream by report analysts and reduce the data quality efforts for statistical/mathematical modelers. Manual and redundant effort to perform repetitive data quality would be alleviated.

Performance is crucial from an IT perspective and performing data quality tasks inside Teradata will significantly reduce bottlenecks that result from moving data over a network. Service level agreements for delivery of data quality expectations have a much greater chance of being continually met or exceeded and data quality can be applied where missed service level agreement is not an option. This also alleviates downstream ad hoc data quality efforts by end users that are time consuming, redundant, and inconsistent in application.

In-database processing addresses these challenges by moving data quality tasks closer to the data and improving the integration between SAS®9 and the enterprise data warehouse (EDW) from Teradata. Performing these tasks inside Teradata will significantly reduce bottlenecks that result from moving data over a network. Customers can also tap the respective strengths of SAS and Teradata technologies. In addition to reduced data movement, which results in improved performance run times and reduction in redundant data, analytic models can be developed and deployed faster, transforming your data into precious and usable insights or revelations.

Business Requirements Change

Customer dynamics are changing more frequently because of the *way* we can complete monetary transactions whether we are trading stock, buying a jacket online, going to the amusement park and paying by credit card, or using our smart phone to purchase a piece of artwork at a local art fair. These dynamics accelerate the need to develop and refresh analytic models that can be managed securely. It also accelerates the need to develop data quality processes that can keep pace with the demand to develop and refresh analytic models to continually “know your customer” or “know our customer better and faster as their behavior changes”. By implementing in-database data quality procedures, increased model development efforts to keep pace with business demands can be realized. This occurs because some of the data quality handled by analysts before they could begin building models has been removed, which means more or better models can be developed and deployed faster.

“A major financial services organization plans to leverage the efficiencies of SAS in-database processing with their Teradata system to increase SAS analytic processing speed ten-fold. This organization also estimates a reduction in model deployment time (from development to delivery) in terms of hours versus weeks. The reduced total cost of ownership (TCO) these efficiencies generate results in higher return on investment (ROI). This initiative will enable joint SAS and Teradata customers to focus more on deriving analytic insights from their Teradata system than managing their detail data stores.” [6] Adding data quality capabilities inside the database such as creating a match code for customers, patients, or business can allow better identification and analysis of individual customers, patients, households, and businesses.

Another example that comes from the financial sector involves banks that are challenged to grow revenue and control costs in the environment of a slow economic recovery that is compounded by dramatically increased regulatory compliance demands. Customers are more likely than ever to move their business to competition, which could be other large banks or even community banks and credit unions. To retain customers, acquire new ones, and cross-sell to existing customers, banks must leverage every byte of internal and external data so that they can understand customers’ needs better. This also enables them to act on those needs as quickly as possible with informed decisions to improve and keep customer retention, cross-sell, and recommendations to others that leads to additional customers. With big data, we need to ensure data quality and we need to do it sufficiently fast to meet business demands in a fast paced world**.**

SAS EMBEDDED PROCESS

The SAS® Embedded Process places a SAS process beside database query processes so that SAS components can be leveraged to collaboratively process data flowing directly out of and back into an SQL query. It features a multi-threaded design for query parallelism.

The SAS Embedded Process is an integral part in MPP database implementations to enable a broad range of embedded processing inside MPP databases.

With SAS In-database data quality, you can manage data quality in-place by embedding data quality functions, a Teradata framework or environment, processes, and applications inside the database. There is no large scale copying of data from the database to a file system and server to perform data quality routines such as standardization, parsing, and matching thereby increasing performance and reducing network traffic.

DATA QUALITY FUNCTIONS IN-DATABASE FOR TERADATA

Table 1 shows examples of parsing and extraction functions.

Function	Description	Example	
Parsing	Segments a string into semantically atomic “tokens”	Mr. Roy G Biv Jr	
		Prefix	Mr
		Given Name	Roy
		Middle Name	G
		Family Name	Biv
		Suffix	Jr
Extraction	Extracts context-specific entities or attributes from text string	Blue men's long-sleeved button-down collar denim shirt	
		Color	Blue
		Material	denim
		Item	shirt

Table 1. Parsing and Extraction Functions

Table 2 has examples of pattern analysis, identification analysis, and gender analysis functions. Pattern analysis shows simple representation of a text string’s character pattern. It is used for pattern frequency analysis in profiling jobs.

Function	Description	Example	
Pattern Analysis	Shows simple representation of a text string’s character pattern. Used for pattern frequency analysis in profiling jobs.	Input	Pattern
		919-677-8000	999-999-9999
		NC	AA
Identification Analysis	Determines the type of data represented by a text string	Input	Identity
		John Smith	NAME
		SAS Institute	ORGANIZATION
Gender	Determines the	Input	Gender

Analysis	gender of a name	Jane Smith	F
		Sam Adams	M
		P. Jones	U

Table 2. Pattern Analysis, Identification Analysis, and Gender Analysis Functions

Table 3 has examples of standardization and casing functions. Standardization outputs a preferred standard representation of a string.

Function	Description	Example	
Standardization	Outputs a preferred standard representation of a string	Input	Output
		No Car	NC
		919.6778000	(919) 677-8000
		Smith, Mister James	Mr James Smith
Casing	Applies context-specific casing rules	Input	Output
		DATAFLUX CORP	DataFlux Corp
		ronald mcdonald	Ronald McDonald

Table 3. Standardization and Casing Functions

Table 4 shows the matching function or matchcode generation. A matchcode is a fuzzy representation of a string, similar to a checksum for a file.

Function	Description	Example	
Matching	Generates a “matchcode” for a text string. A matchcode is a fuzzy representation of a string, similar to a checksum for a file. Records can be clustered by sorting by matchcodes. Fuzzy lookups can be performed via matchcode searches.	Input	Matchcode
		Gidley, Scott A	XYZ$$$
		Scotty Gidleigh	XYZ$$$
		Mr Scot Gidlee Jr	XYZ$$$
		Mr Robert J Brauer	ABC$$$
		Bob Brauer	ABC$$$

Table 4. Matching Function

The in-database functions will generate matchcodes and manipulation such as clustering, entity resolution, and fuzzy lookups on matchcodes is done using SQL programs.

The data quality functions use the Quality Knowledge Base (QKB), which contains domain-specific rules and reference data used to analyze and transform strings and out-of-box support for the Contact Information domain.

SCALABILITY AND PERFORMANCE

In-database processing for data quality scales to any volume and the scalability is directly related to the available hardware resources. By using this technology you can:

- have fewer points of failure during execution
- leverage fully an existing database investment
- keep data in the database (no network data movement)
- appreciate markedly improved performance for data quality
- perform data quality not possible before because of time constraints

Figure 1 describes the time-related issue associated with a need for data quality while data volumes are ever increasing.

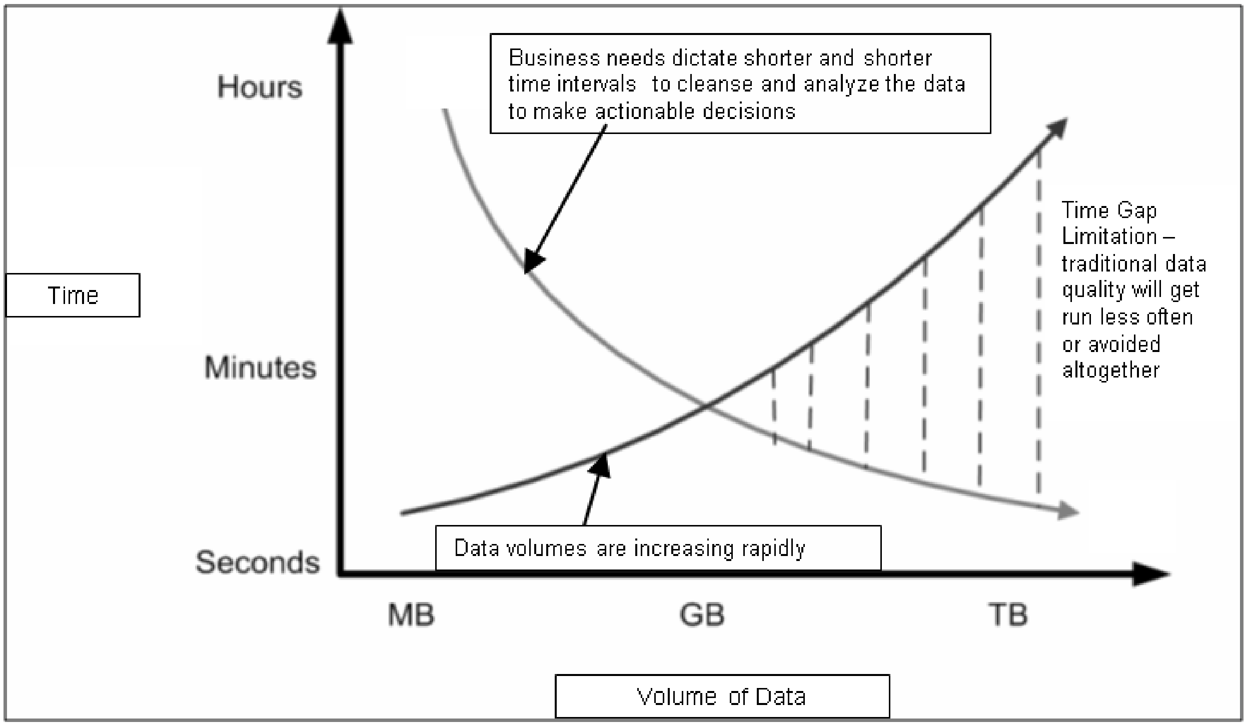
Hours
Business needs dictate shorter and shorter time intervals to cleanse and analyze the data to make actionable decisions
Time
Minutes
Time Gap Limitation – traditional data quality will get run less often or avoided altogether
Seconds
Data volumes are increasing rapidly
MB
GB
TB
Volume of Data

Figure 1. A time gap when data quality processes cannot adequately handle big data.

Figure 2 shows performance of generating matchcodes for names on a PC (the traditional data quality approach) versus in-database on Teradata (one node) and in-database on Teradata (16 nodes). In this early result 50 million records can be processed in a little over two minutes or about 1.4 billion records per hour. Note that both axes in this chart are plotted on a logarithmic scale.

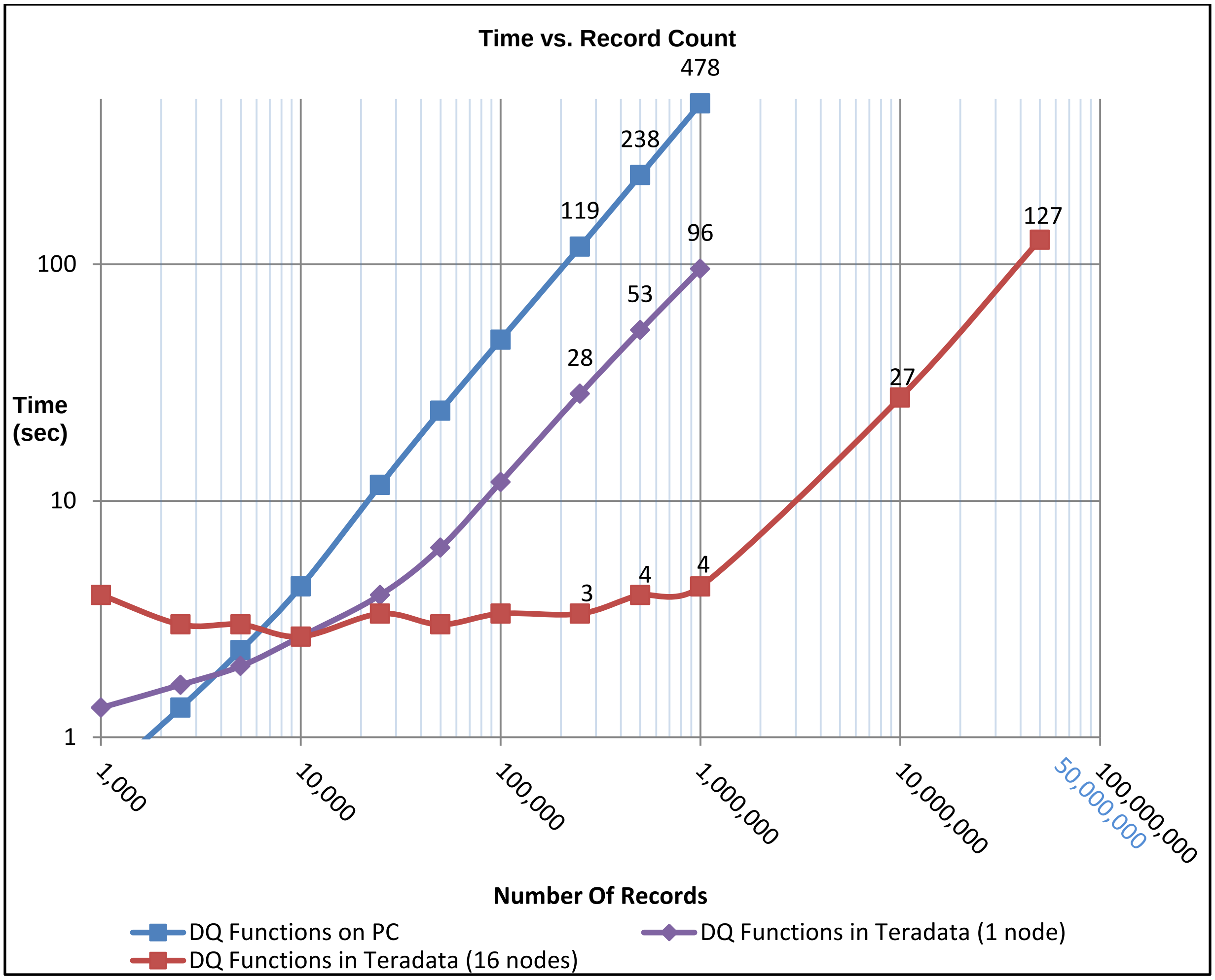

Figure 2. Initial Comparison of Performance

Figure 3 shows throughput versus record count. If processing under 10,000 records, there is little value for in-database data quality. As the number of records increases above 10,000, the 16 node Teradata instance continues to show linear scalability out to a million records, until it reaches a point of stability at 200 times more throughput than traditional data quality.

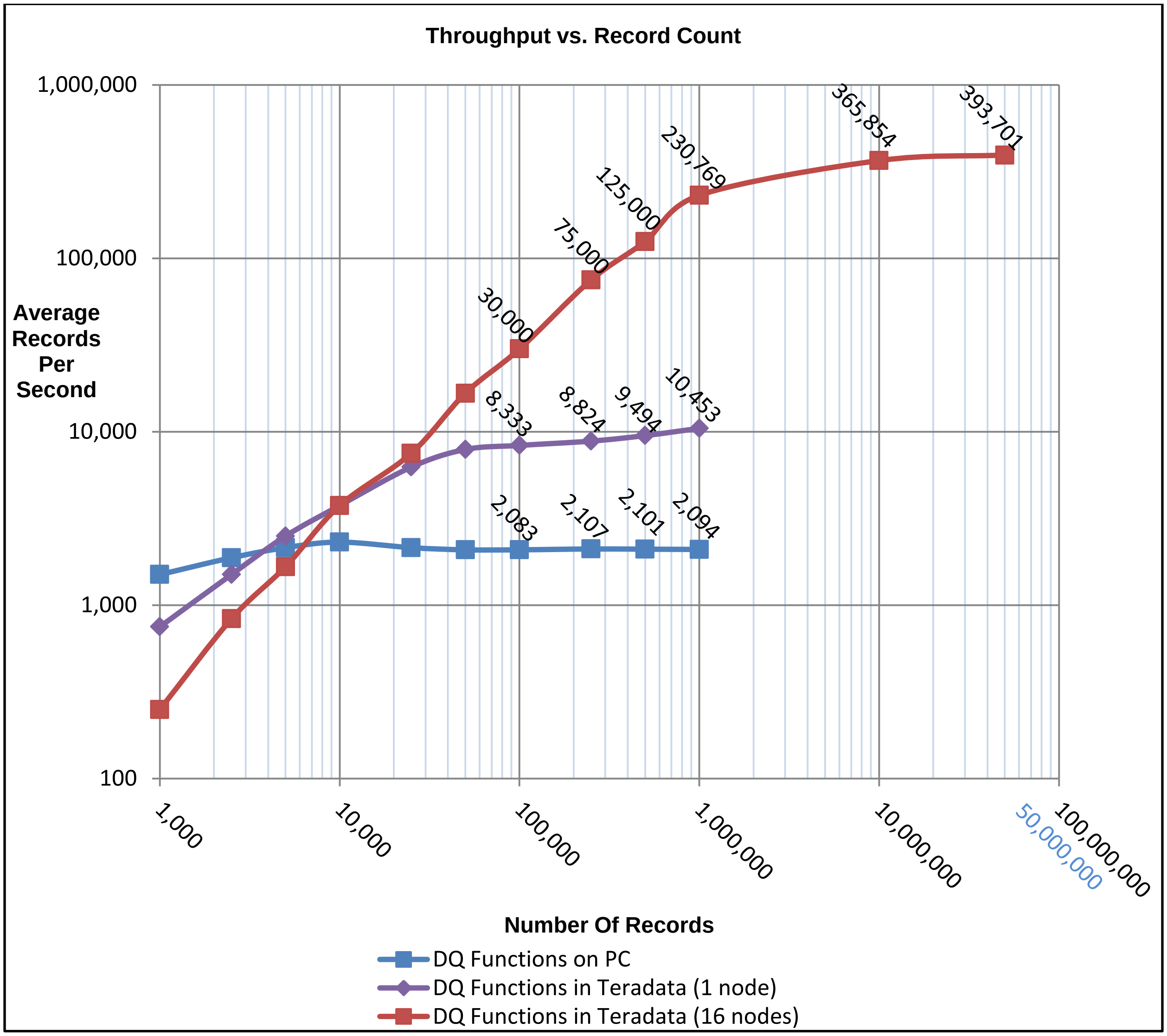

Figure 3. Throughput versus Record Count

SAS AND TERADATA: STRATEGIC PARTNERSHIP

There is corporate commitment from the highest levels and process alignment in sales, pre-sales, support, and program management. Research and Development have a joint product roadmap with continuous product release. Multiple joint resources include the following:

- Center of Excellence
- Business Analytics Innovation Center
- Product Advisory Council
- Dedicated R&D teams to integrate and optimize SAS for Teradata

The SAS and Teradata partnership mission is an integrated and robust business analytics and data warehousing environment from two industry leaders. Solutions that allow companies to focus on higher value business opportunities and a reduction in the complexity and cost for decision making will provide:

- reduced data movement, redundancy, and latency issues
- improved data quality and data consistency
- lower total cost of ownership and investment protection

CONCLUSION

This paper introduced data quality for big data and the in-database data quality procedure concepts for performance.

Data management and data quality processes must be sufficiently fast and accurate in supporting the continuum of the myriad of downstream big data consumption. Big data can unlock significant value by making information transparent and usable at much higher frequency and sophisticated analytics can substantially improve decision-making thereby improving development of next generation products and services.[7]

In-database processing for data quality scales to any volume and the scalability is directly related to the available hardware resources. The value of "time to data quality":

- significantly reduces the time required for data quality processes to be completed.
- allows IT to now meet once unattainable data quality service level agreements.
- enhances business intelligence and analytics programs. By proving accurate data faster than ever before, organizations can run against more types of clean data and have it run more often, leading to a greater confidence in their analytical and Business Intelligence results.

REFERENCES

[1] Luckie, C. 2013. *Big Data" Facts and Statistics That Will Shock You.* Fathom. Available at http://www.fathomdelivers.com/big-data-facts-and-statistics-that-will-shock-you/

[2] Gualtieri, M. 2012. *The Pragmatic Definition of Big Data*. Available at http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data

[3] Execution-MiH. Data Quality Definition- What is Data Quality? Available at http://www.executionmih.com/data-quality/accuracy-consistency-audit.php

[4] Luckie, C. 2013. "*Big Data" Facts and Statistics That Will Shock You,* 2013. Available at http://www.fathomdelivers.com/big-data-facts-and-statistics-that-will-shock-you/

[5] Shamlin, D., Duling, D. 2009. *In-Database Procedures with Teradata: How They Work and What They Buy You***.** SAS Institute, Cary, NC. Available at http://support.sas.com/resources/papers/proceedings09/337-2009.pdf.

[6] Nelson, B. G. 2004. Real-Time Decision *Support: Creating a Flexible, Architecture for Real-Time Analytics*, ThotWave Technologies, Cary, North Carolina

[7] Manyika, J., Chui, wet al**.** May 2011**.** *Big data: The next frontier for innovation, competition, and productivity,* McKinsey Global Institute

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Charlotte Crain
SAS Campus Drive
SAS Institute Inc.
Charlotte.Crain@sas.com