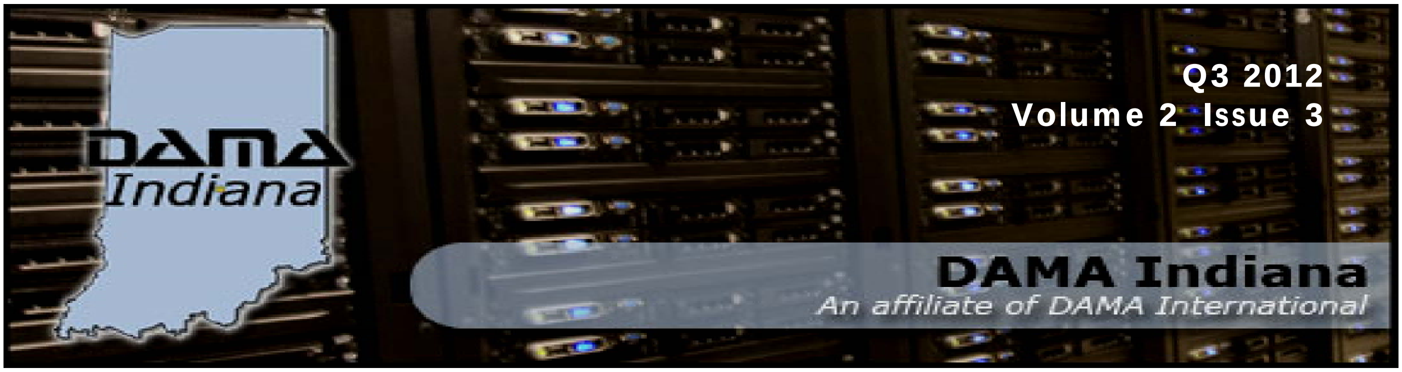


Q3 2012
Volume 2 Issue 3

DAMA Indiana
An affiliate of DAMA International

Web site address www.DAMAIndiana.org
LinkedIn Group DAMA Indiana
Facebook DAMA Indiana
Twitter @DAMAIndiana

Inside This Issue

Welcome to DAMA Indiana News!

Welcome to the fall edition of the DAMA Indiana newsletter! In this issue, member Colleen Delahanty shares her Big Data experience at the TDWI conference and President Sue Peoni helps us recognize the signs of DQO.

Preparations are well under way for the October chapter meeting. More details are included in this newsletter and will be sent via e-mail. At the October meeting, we will be electing the Board. Please let us know if you are interested in a position. The VP of Administration position will be open. Please check out page 9 for more information on board positions.

Do you have a question or an idea for a newsletter article? The top of the newsletter contains the web site and social media contact information for the chapter. In addition, the contact information for the Board is on the last page. We would love to hear from you! Also, please let us know at newsletter@damaindiana.org if you have any changes to your contact information.

From the President's Pen...

By Sue Peoni

Considering DQO

As I write this, my first two REAL, LIVE, MDM projects since arriving at ADESA have FINALLY been approved (Hooray!). This is truly good news, but it also means I am facing serious DQO with my business partners. What is DQO, you might ask? DQO stands for Data Quality Overwhelm – a clever, and very appropriate, term I picked up from the IAIDQ (Information/Data Quality Professional Open Community) group on LinkedIn. DQO is that look in the eyes of the business when you tell them you want them to REALLY manage their customer data; that you are about to redesign all of their nice, comfortable business processes and remove their ability to write over attributes at will; that you will be asking them to do this without adding any more resources; you are asking them to (ack!), CHANGE!!! All of my business partners who have been enthusiastically pushing me to get MDM moving now have a completely different look about them: Uh-oh, now what do we do?

Please see *President's Pen* on page 2

President's Pen from page 1

The person on IAIDQ posing the question received lots of good advice in return. Unfortunately, most of it involved expensive impact analyses that I probably don't have the budget or time to do. Despite my constraints, I do acknowledge the need to think about DQO, plan for it, and work with it. A word mentioned in the IAIDQ thread is "journey" – and I find myself using that word frequently in presentations and conversations with my business partners. We talk a lot about this being one step at a time toward better data quality, and I promise not to force more change on them than they can deal with. I also try my best to see the situation from their perspective without giving into their fears. If I can find the DQ issue that drives them crazy and fix that in the first implementation, I find they are much more willing to work with me on process changes. Unfortunately, some of those issues don't get resolved until a later phase -- then all of my best selling skills come into play!

I also find many of my business partners want a copy of the cookbook – "Just follow these six easy steps, and Data Quality is yours!" Unfortunately, after many years of doing this, I've concluded that, while based on a common set of principles, good DQ is largely situational. It must be crafted to fit a unique set of circumstances, business practices, and constraints. Even within the same organization, the DQ practices necessary to maintain good customer data are much different than those used to maintain DQ for product or vendor data.

So off we go on our adventure! I promise to let you know at exactly what point the look of panic fades and confidence in MDM starts to appear. Maybe they'll let me take pictures!

-- Sue

> *"DQO is that look in the eyes… when you tell them you want them to REALLY manage their customer data."*

Save the Date!

The next DAMA Indiana meeting date is:

Thursday, October 18th

NOTE: We will be back at the Eli Lilly Corporate Center.

DAMA BOARD POSITION

The Vice President of Administration position will be available for the October elections.
Please let us know if you are interested!

⬅ Check this out!

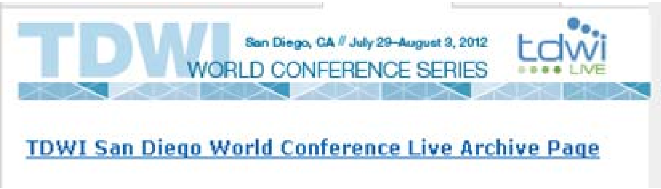
TDWI World Conference 2012: Big Data – The Tipping Point

By Colleen Delahanty

I recently had the opportunity to attend the TDWI World Conference in San Diego. This is the 3rd TDWI conference that I've attended since 2001. I have always been impressed with the quality of the events and this one did not disappoint. TDWI prides themselves on their educational sessions being a sales-free zone. They do also offer an exhibit area for those that choose to visit the vendors to discuss and view demonstrations of their products. There were multiple sessions to choose from each day on a variety of topics. There were also 2 keynote speakers who addressed the entire group. Traditional topics pertaining to data warehousing and data analytics were offered as well as the most recent trend in data and the theme of this year's conference…….'Big Data'. I attended a few of the traditional sessions, but tried to focus most of my time on the concept of 'Big Data'.

Big Data Defined

So, what exactly is this thing called 'Big Data'? It isn't solely about volume. It is about the drive and determination to always want to know more. The expectation of having the data and knowing what to do with it is growing. The consequences of not knowing are growing. The way that we are thinking is changing. We now have a 'Google it' mentality, where having access to an immediate answer is an expectation, and having data to arrive at the answer is a necessity Data is coming from everywhere! We have sensors connected to things that we never would have imagined a short time ago………'smart' vending machines collect data, workout clothes monitor your workout and give you feedback on suggested areas of improvement, sensors in your mouth monitor plaque build-up, just to name a few. IBM estimates that there are 2.5 quintillion bytes created daily, with 90% of existing data created in the last 2 years and the average rate of growth will be 40 – 60% per year!



TDWI San Diego World Conference Live Archive Page

Although the definition of Big Data is evolving, there are some characteristics that constitute it:

- Volume – Large quantities
- Velocity – Sometimes having data that is 2 minutes old is too late.
- Variety – Structured, unstructured (text, video, streams, etc…), combination of structured and unstructured
- Viscosity – Variable rate of data being sent
- Complexity
- Ambiguity – not easily interpreted without context

Capturing the data isn't the biggest challenge. Conditioning the data to get it to an acceptable quality in a timely manner and figuring out how to present it in an understandable form for human consumption is the biggest challenge. The value of an individual piece is high immediately but begins to deteriorate shortly after that. This isn't to say that the data isn't useful over time, just that as time passes, the value shifts to the aggregation of the data and away from the individual piece of data.

> *"Capturing the data isn't the biggest challenge."*

Continued on page 4

TDWI continued from page 3

So what technologies are available to deal with the challenges of Big Data? There are many to choose from and they are rapidly changing. It is difficult to keep up with the latest trends. Categories of DBMS available include: Relational, Non-relational, Analytic, Operational, NoSQL, NewSQL, Key Values, Big Tables, and many more.

Hadoop (which falls into the categories of Non-relational and Analytic) is one of the latest crazes in unstructured data processing transformation and analysis. It is a distributed file-based system, written by engineers, for use by engineers. It is not intended for use by the casual SQL user. Its primary purpose is to provide a place to store data where it can be read and analyzed. Given that purpose, once the data is written, it cannot be updated. Hadoop works in conjunction with tools named MapReduce, Pig and Hive. MapReduce is the framework used for writing applications that rapidly process vast amounts of data. It functions as the job scheduling/execution component. Pig is a procedural language used to analyze large sets of data. Hive is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries (using HiveQL), and the analysis of large datasets stored in Hadoop compatible file systems.

Hadoop is not intended to replace the traditional RDBMS, but work in conjunction with and compliment it. Many architectures continue to use a RDBMS for their traditional structured data and Hadoop for their unstructured data. Visualization tools and techniques are recommended to allow users to digest the vast amount of data that is being presented to them.

Advice on Big Data
Regardless of the tool(s) chosen, here are some words of advice:

- Don't choose your technology based on the latest hype. Choose your technology based on the problem that you are trying to solve.
- Once external data is brought in-house, you don't want to continue moving it around. Bring the analytics to the data. This is especially true for large quantities of data and/or unstructured data.

Keynote Highlights
One of the keynote speakers was from eBay. His presentation highlighted their use of 'extreme' analytics. eBay has no tangible product to sell. They provide the means by which others sell and buy products. Analytics has to be a part of their DNA in order for them to be successful. Some of their data statistics that stood out were:

- 50 TB of new data added every day
- Millions of queries executed every day
- More than 150,000 data elements recorded
- 2.6 trillion rows in the largest table

The form of their data varies from traditional structured data, to semi-structured data (data that has a generic form, but big chunks of text as part of that form), to totally unstructured data. Their platform architecture includes a traditional

Continued on page 5

TDWI continued fron page 4

data warehouse accessed by SQL, as well as the unstructured piece which is housed by Hadoop. Virtual data marts are built on top of the multiple forms of data.

Their extensive use of analytics was evident in their decision to implement the eBay Buyer Protection Plan. The decision to implement this came with a huge amount of risk, so they needed to be confident, given that it is based solely on trust of the seller.

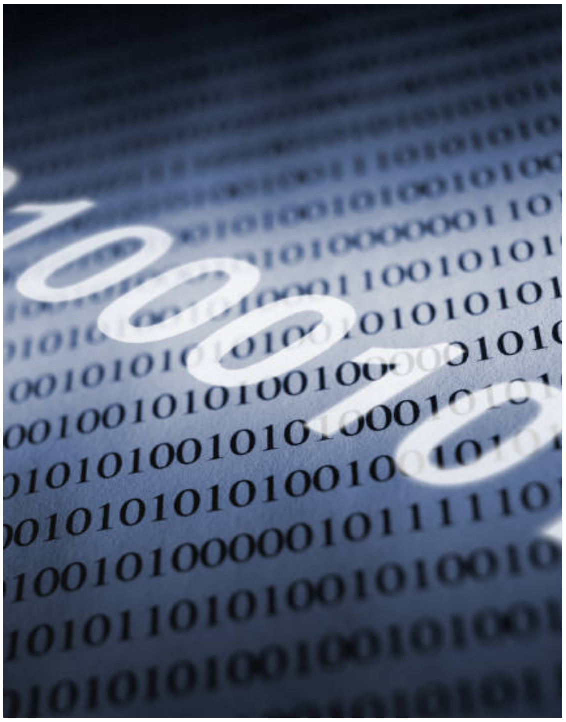
The other keynote address was set up as a mock debate about whether Big Data is really new and whether or not it will change the world. On one end of the spectrum, the argument is that Big Data is nothing new. We've been down this road before with file based platforms, mainframe-like platforms and languages that are SQL-like. The opposing view is that Big Data, with the volume and velocity at which it is coming, is like nothing we've ever seen before. Standard methods of access and presentation soon will be inadequate and better methods of visualization are needed. It was an entertaining way to make the point that like in most debates, the truth lies somewhere in the middle.

In conclusion, like in the past, this TDWI was both enjoyable and educational. I learned some new concepts and reinforced some old beliefs. I would recommend the TDWI conferences to anyone who has the chance to attend.

Check out these Big Data Resources:

Hadoop Web Site
http://hadoop.apache.org/

TDWI Keynote
http://tdwi.org/live/sandiego2012



Member Question: Are you big on BIG DATA?

We would like to hear YOUR thoughts on Big Data and the value to you. Responses (which remain anonymous, unless you specifically request that your name be shared) will be included in an upcoming issue.

Are you looking at Big Data programs or projects?

What is the biggest driver for your interest in Big Data?

What resources have been the most helpful to you?

What do you think are the biggest challenges with Big Data?

Please send your thoughts to newsletter@DAMAIndiana.org.

Big Data: Hadoop

By Mershard Frierson

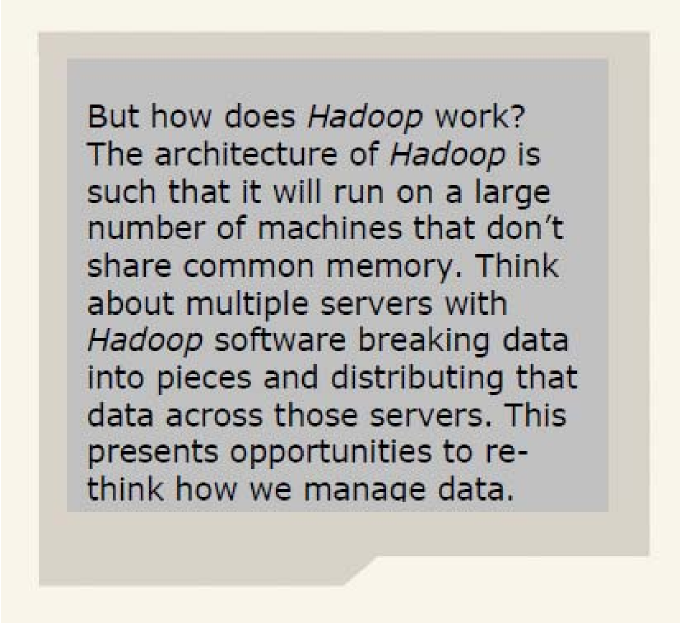
INTRODUCTION

“Big Data” analytics is the growing topic in the arena of scalable information technology. The attractiveness of Hadoop is that it adapts easily to scalable deployment and data utilization requirements. Who would have ever imagined a world where the concept of data outgrowing a *relational*-database would ever dominate our conversation? The team that developed *Hadoop* visualized today’s informational landscape.

IS THE ENTERPRISE READY?

To continue this exploration, let’s first determine if the enterprise is ready for Hadoop. And in making that determination, we will start where most managers begin their analysis. We know it as the Cost\Benefit Analysis- usually associated with adaptation of emerging technology. If cost is a primary consideration for *readiness*, then it must be a relief to know that *Hadoop* is essentially *open-source*!

For questions, contact Mershard at:
Mershard Frierson
317-210-DATA (3282)
BIG Data Consultant
www.mershard.com

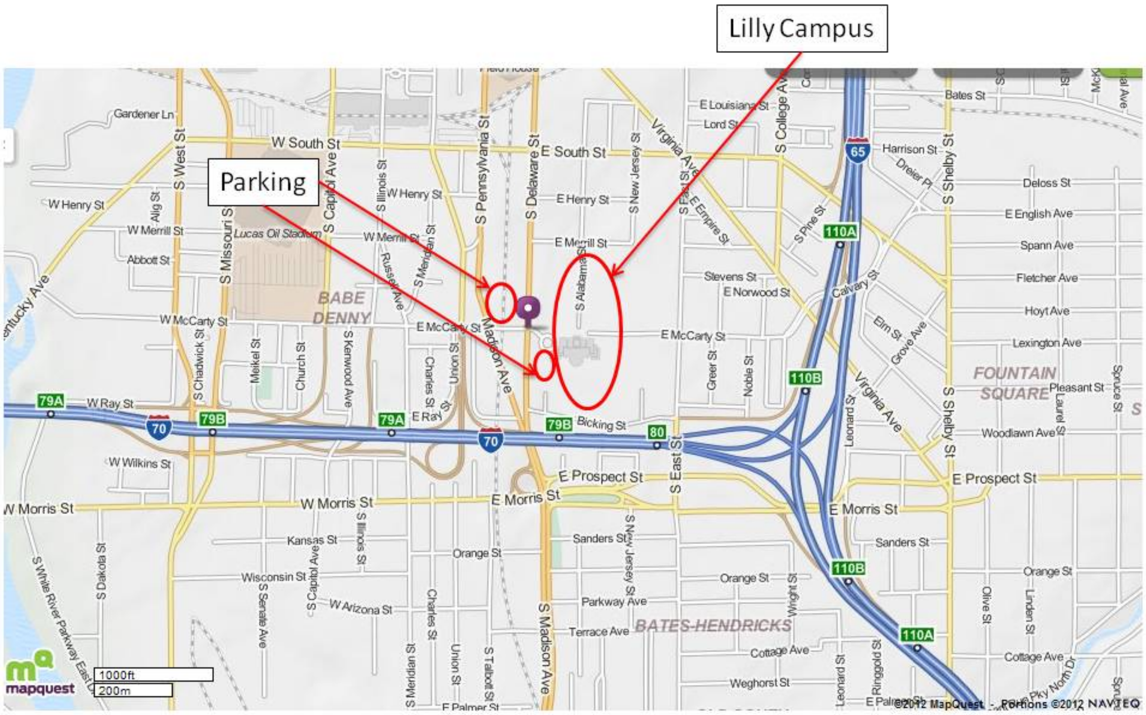


But how does *Hadoop* work? The architecture of *Hadoop* is such that it will run on a large number of machines that don’t share common memory. Think about multiple servers with *Hadoop* software breaking data into pieces and distributing that data across those servers. This presents opportunities to re-think how we manage data.

PROS AND CONS TO HADOOP

There is more to preserving data integrity than just breaking it into pieces and re-assembling data. In fact, data integrity and regulatory compliance are still present challenges to the mainstream integration of Hadoop. This regulatory compliance speed-bump may balance out the elevated expectation of delivery that is often associated with new technologies. Finally, there are many types of data: medical, financial, and payment card, where multiple state and federal laws are applicable. These are a few considerations, as we prepare for the age of “Big-Data”.

October Chapter Meeting



Join us on **Thursday, October 18th** for an informative day you won't want to miss!

At our session, we are big on Big Data and getting in the loop on Hadoop. We will be featuring speakers as well as breakout sessions on a variety of topics.

Our featured speaker of the day will be Michael Covert – Founder and CEO of Analytics Inside. He will be speaking on the topic of Big Data Solutions. We will also hear from Informatica, break into roundtable discussion groups, and get a preview of the upcoming data modeling conference.

Lunch will be provided courtesy of Informatica.

Please note: We are back at the Eli Lilly Corporate Center, but our room has changed!

We will be meeting in the Leadership Development Center, near the main entrance of Lilly Corporate Center. (McCarty/Delaware entrance with the fountain) Please go to the main entrance and let security know that you are there for the DAMA Indiana meeting. You will be directed upstairs to the conference facility. Above is a map of the Lilly location. Visitor parking is located both in the attached garage (beside by the fountain) or in the garage at the northwest corner of Delaware and McCarty (enter from McCarty street).

Please register by sending an e-mail to info@damaindiana.org. We hope to see you there!

Professional Development Opportunities

The Data Warehousing Institute (TDWI) hosts several conferences during the year. Check out http://tdwi.org for more information.

tdwi
The Premier Source for Business Intelligence and Data Warehousing
TDWI International
Follow us

http://www.datamodelingzone.com/

DATA MODELING ZONE 2012

November 12th-15th 2012 in Baltimore, Maryland

***Sharpen* your data modeling skills and *Connect* with the community!**

http://www.information-management.com/conferences/mdmnewyork/

MDM & DATA GOVERNANCE SUMMIT
NEW YORK
2012
October 14-16, 2012 | New York Marriott Marquis

For additional conferences and webinars, check out Dataversity at: http://www.dataversity.net/education/events/

DATAVERSITY™

Enterprise Data World 2013
Call for Presentations
http://edw2013.dataversity.net

Proposals due by October 1, 2012!!!

Around Town

Here are other area events that may be of interest to data professionals:

IndyPASS (Professional Association for SqlServer)
Check the website at:
http://indiana.sqlpass.org/

INOU**G (Indiana Oracle Users Group)**
October 25th – User Group Meeting
www.inoug.org

Robert Seiner has articles and resources at: http://www.tdan.com/

T D TDAN.com A N
The Data Administration Newsletter
SINCE 1997

Check out articles, DM radio, and webinars at: http://www.information-management.com/

Click here to find out more!
information management

Reminder

Attending conferences and professional meetings counts toward CBIP and CDMP recertification credits. Visit the ICCP site today: http://www.iccp.org/cgi-bin/pdform.php

Spotlight: Board Elections

It is already time to start thinking about the slate for October elections!

In the Q3 2011 issue of the DAMA Indiana newsletter, each board member described his/her position responsibilities and favorite things. Check it out at: **http://www.damaindiana.org/Newsletters/DAMA_Indiana_News_Q3_2011_FINAL.pdf**

Board positions are 1-year commitments, with annual elections at the Fall (typically October) meeting.

Our chapter by-laws guide the operation of the chapter and detail the board position responsibilities. An overview of the by-laws was also in the Q3 2011 issue. The board also welcomes volunteers who are willing to help with meetings and the newsletter.

If you have any questions, please contact a current board member (see contacts on this page).

Board Elections

Elections are just around the corner! We will be voting for the slate of officers during the October meeting. If you are interested in a board position, please contact Dan Heffern at VPAdministration@damaindiana.org

DAMA Indiana Board

President: Sue Peoni
President@damaindiana.org

VP Administration: Dan Heffern
VPAdministration@damaindiana.org

VP Communications: Tom Morris and Christi Denney
VPcommunications@damaindiana.org
newsletter@damaindiana.org

VP Finance: Gene Boomer
VPFinance@damaindiana.org

VP Online Content: Christina Knotts
VPOnlineContent@damaindiana.org

VP Programs: Michael Irick and Ravi Chittaranjan
VPPrograms@damaindiana.org