Big Data:

Powering the Next Industrial Revolution

Author: Abhishek Mehta

April 2011

Executive Summary

Data is a key 'raw material' for a variety of socio-economic business systems. Unfortunately, the ability to maximize the use of and value from data has been under leveraged. Till now. A massive change in the technology ecosystem, explosion in sensors emitting new data and sprouting of new business models is revolutionizing the data landscape. Forever changing the way we look at, solve and commercialize business problems.

This is the world of Big Data, where 'big' is merely a reference to the size of the opportunity, not the inability to address it.

It is ushering us into the next industrial revolution, which will be powered by data factories, fed by the most valuable raw material of all, data.

The total value at stake - $5.000,000,000,000[i]. Five trillion dollars.

Data - The New Normal

There is a new normal emerging for Big Data, shattering hitherto taboo and unimaginable standards and beliefs that really were not that logical. So everything from what data was considered worthy of storage to how best to analyze and consume the data has been re-defined. The new ground rules may seem like a common sense approach to data, but have never before been implemented universally.

Data Trash - An Oxymoron

There is nothing, note nothing, called data trash. Irrespective of the source, structure, format, frequency and identity of the data, all data is valuable. In the new normal, data trash is an oxymoronic phrase. If a certain kind of data seems to have no value today, it is simply because we have not mined it enough or cannot yet comprehend what its use could be. As an example, when Google started harnessing (the satellite imagery), capturing (street views) and then sharing (for free) geographical data 5 years ago, not many people got what the value of it really was. 5 years hence, think again.

Store all the data, all the time

Not only is there no such thing as data trash, we need to store all of the data being produced in the world today. Thanks to Moore's law[ii], the cost of storing a terabyte of data has fallen from $1 million in the 1970's to $50 today. Moreover, this rapid decline in storage cost, coupled with the increase of storage capacity in increasingly smaller footprint has rendered tape as a storage mechanism obsolete. An organization can finally store all of the data it creates, buys and consumes, for every individual customer, with all associate history and for every single interaction.

Size does not matter

If the data in all of the books in the Library of Congress was sized, it would total roughly 15 terabytes[iii]. Twitter generates about 15 terabytes of data a day. Walmart processes 1 million transactions per hour and stores around 3-5 petabytes of data (1 petabyte equals 1,000 terabytes). Google can process 1 petabyte of data each hour and has the ability to store the entire web on its servers (around 25 petabytes of data), many times over.

By 2010, all the digital data in existence was around 1,200 exabytes (1 exabyte equals 1 million terabytes). It has been estimated that if we were to put in place and size all words ever spoken by mankind, it would be around 5 exabytes.

And this gargantuan mound of data continues to grow at a rapid pace. With the explosion in sensors (there are 4 billion cell-phones in the world) and rise of ever more ambitious projects (the Large Hadron Collider[iv] generates 40 terabytes of data per second), data will continue to occupy the space it is given.

Beyond these massive numbers is a silver lining. The world has the capability to make Big Data small. Size of the data is no longer a limiting factor, but an opportunity to build a unique capability and competitive differentiator if you have the power to harness it all.

Structure does not matter either

5% of all data produced in 2010 was structured. The rest, using the much abused industry term, was 'unstructured' data. This term usually refers to data streams like email, voice, GPS, video and others we cannot comprehend. The value of it is unquestionable. How we uncover it the challenge.

While I do not like the term 'unstructured data', I do agree in principle that irrespective of data sitting in existing data warehouses where a format has been force fitted to it (making it structured) or data stored as flat files waiting for the structure to be found beneath (hence 'unstructured' or better said complex), the information it can give us is valuable.

Data is personal again

The ability to take data that exists in multiple silos, whether they are within one organization (multiple Systems of Record taking in data) or across many (commerce data, healthcare data, financial data, telecom data, online data, social data etc), and tie it together at the individual level is game changing in nature. This has been very successfully done by Google for web search. Every single IP address has its own customized 'internet', in essence enabling mass personalization of a very wide swath of data. The fact that tools can allow for this massive democratization of data and enable individualization to its nth degree is fascinating as it renders sampling useless. And analyzing data opens up possibilities to address critical socio-economic problems that simply could not be solved before.

The Technology Mega-Shift

The driving force defining this new normal is a tectonic shift in technology that would not have been possible were it not for the dot com boom; which led to the creation of a world wide platform that was infinitely powerful, scalable and cheap. Those developments have gathered pace in the last 5 years as what was considered nascent projects then, have now become the building blocks of the architecture shaping the next generation Big Data capabilities.

Commodity hardware

Making $50 terabyte disks stand up to the rigors of more than a billion people simultaneously researching a massive online 'database' is not easy. Yet this is exactly what Google does every second. What they had to write the book on is how to take commodity hardware and with the right fault tolerance, processor power and networking, mesh it into a single unit as powerful as supercomputers, at $2,000 a box.

"Store everything" file systems

Storage systems are getting simpler and more powerful. Distributed File Systems, such as Hadoop, now allow users and companies to store all kinds of data in one place. So whether you need to handle transactional data, emails, voice logs, GPS triangulations, video feeds, chat transcripts, photos, machine data or data not even in existence today, the good news is it can all be stored, indexed and processed on one platform. Structured and unstructured data can finally co-exist, even be processed and tagged to one ID, all on the same platform.

Faster is better - Let's game the network

It's a rule of nature, when things get big, they get difficult to carry around. Newer technologies are rendering the necessity to move data or process it before you could use it as unnecessary. Two interesting tools, Tableau, a visual inteligence software tool and Hadoop, an open source Data Operating system, are good examples of this. Both Tableau and Hadoop have leveraged a unique advance as well as change in discipline.

Rather than move the data to your code, which is time consuming, messy, and sometimes downright impossible - because one should not try to move one petabyte of data through a one gigabyte network; they process the data where it is stored by taking the code to the data (megabits moved through a gigabit line (a lot faster) and then bringing the results back to the user (once again, much smaller size of data to be moved).

“ Size of data is not longer a limiting factor, but an opportunity to build a unique capability and competitive differentiator if you have the power to harness it all.”

This opens up a range of possibilities, from the ability to model an entire population, rather than samples, to reducing the time taken to experiment with new data. It also allows the user to model with all instances, including tail events, hence building inherently superior algorithms with much better predictive ability.

Massively-parallel processing (MPP) analytics

While massively parallel systems have existed for a while, MPP analytics takes the parallelism on offer and supercharges analytical processing as well. The ability to write algorithms that can now leverage the benefits of parallelism is the next big leap in computing. It finally gives the ability to mankind to solve problems that were either too expensive, too time consuming or just computationally impossible to solve.

The work being done on the Human Genome is a good example. It was first decoded in 2003. It took 10 years to do it and as you may know, it wasn't an easy thing to do. Decoding the human genome involves taking three billion base pairs and then processing various combinations and analyzing all of it at scale. With new technologies and the advancements, what took ten years in 2003 is now down to a week. And at a tenth the cost.

Just imagine the challenging problems that can now be turned on their head. With these rapid advancements, we can start questioning the assumptions that we were forced to take due to the technological limitations imposed on us and solve problems we thought were not solvable.

Factories of the future

We are truly fortunate to be witnessing the birth of the second industrial revolution. This revolution is being fueled by the factories of the future - data factories.

So what is a data factory? Think Google, Facebook, ComScore and Zynga. They are what I refer to as our data factories. A lot more are in the works with the explosion in sensors, data and networks. What defines them is the fact that data is at the core of what they build, deliver, commercialize and serve to their customers. The other thing that they have in common is that they are very profitable entities that are redefining established business paradigms on growth, value and margins.

But here is the interesting bit. Even companies like Walmart, HSBC, Aetna, Pepsi, McDonalds, Cleveland Clinic and NSA are data companies. Walmart collects, processes and delivers to its customers the world's leanest and fittest supply chain, a data business. Industries outside the web properties, saddled as they are with legacy technology platforms, just have not enjoyed the same level of innovation and disruption that data brings to their business models…yet. The same technology principles that have completely obliterated web property business models (Hadoop being the major force for change) will be tailored to other industry verticals and cause similar disruption.

As always, startups are leading that charge.

Tableaus is redefining the Business Intelligence space, Tresata is verticalizing the same open source platform that powers Google, Facebook, Yahoo for financial data. Cloudera is aiming to be the 'Red Hat' of Hadoop. And there are many others.

This data powered industrial revolution will be bigger, faster and more disruptive than the first one because data is a core asset for many industries and when enabled, can help solve problems that could not be solved before. Everything from fraud, disease management to smart grids and dynamic traffic management is now within grasp of being solved. The ability to automate the data pipeline and then rapidly find information in the massive amounts of data will be critical to success.

The path has already been laid by the web properties.

Data science and scientists are key

The ability to manage, clean, process and model massive amounts of data is a sill that has historically been undervalued. That is about to change. Data experts, just like algorithm experts, are a key resource for enterprises. Infact, they are more valuable than any other resource - except data itself. The skills that define this talent have a technical component (stats, java, C, hardware management, etc) as well as a strong domain component (financial data experts, healthcare data experts, etc). While the web properties realized this and corrected the imbalance between algorithmic and data management skills, the other industries will be forced to be adapt if they too want to leverage the power of Big Data.

Step 1: Understand how data drives your industry and how data can disrupt it

Where does one start?

Step 1 involves identifying and understanding what are the key decisions or key problems your business needs to solve. Most businesses know the key decisions they are measured on and how they impact the larger enterprise. What do you need to deliver value to your customers? What decisions do you need to make to efficiently manage your enterprise? What decisions do you need to make to increase the value you deliver to your customers?

Once you have identified your key business problems, you need to think about how can data - data you have today or may be able to get tomorrow - improve your ability to make those business decisions?

Here is an example. Say you're a retailer and the key decision you have to make is making sure that stores shelves are stocked with firewood when the winter storm arrives. The questions you need to ask may include:

1. **Geographical Data.** Which areas of the country are worse affected in the winter?
2. **Weather Data.** What are the historical weather patterns? What are they forecasting?
3. **Logistics Data.** What are current stock levels in stores? How much inventory do they have? What will not be stocked to accommodate this need?
4. **Sales Data.** What are the historical sales figures? At what price was I selling them at? What are my margins on the product? Am I pricing it fairly? How do I ensure I am not engaging in price gouging?
5. **Economic Data.** What are current inflation trends? How are they different from last season? What are current market prices of lumber forecast to be? Is emerging market demand impacting process?
6. **Performance Data.** What are my sales projections? What margin do I want to make? What inventory levels should I keep? What should my restocking cycle be? Have I got the supply agreements done? Do I need to hedge my price contracts?
7. **Competitive Data.** Who are my competitors in the area? What are they charging for the same?
8. **Customer Data.** Who are my customers? What do they buy? Are they brand loyal? What was their feedback to us last year? Have they called my call centers with complaints about certain brands? What are they tweeting about? What are their friends buying? And who are they recommending? How do I send them coupons? What promotions do they typically act on?

As you can see, understanding your key business decisions, and the data that can support that decision making, is a very important first step.

Step 2: Automate the data assembly line

Once you have identified what your key decisions are and how data impacts them or can improve the performance of it, step 2 requires the automation of the data assembly line. This is where the biggest opportunity lies. Ford completely changed the landscape of the world of car manufacturing with the first automobile assembly line when they started building the Model T, and the

"With these rapid advancements, we can start questioning the assumptions that we were forced to take due to the technological limitations imposed on us and solve problems we thought were not solvable."

same needs to be done with Big Data.

Google was the first to prove that data assembly lines can be treated the same way as manufacturing. And automation is the only way to scale and manage large quantities of data in an efficient manner.

While this has been done with great success in the web properties, it has only addressed web data (indexing links, building ad optimization networks). A similar step change is needed for other kinds of data, like financial, or health or telecom.

The concept of the assembly line is similar to manufacturing. What was an outstanding question for many was the applicability of what had been done in the web properties, with Google automating the data assembly line for web data, to other industries. That too is being addressed.

Some new startups, like Tresata, are proving it can be applied to data other than web data (in their case financial services). Raw materials have to come in - in this case it would be data. It needs to be processed - cleaned, parsed, de-duped, merged - like steel plates are for cars and then shaped into the final product (in this case data products) tailored to addressing specific business problems.

This ability to automate this assembly line at scale is a key requirement, one that will determine who will be the key players - people, companies, tools, and technologies - that will drive this next industrial revolution.

Companies that can deliver better data to decision makers at the right time and faster pace and ever cheaper costs to make decisions they could not make before are going to be the winners.

It is important to remember that big data and simple models (where tail events and population dynamics fed by each individual are part of the process to help build models, not just stress test them) are more powerful than small data and more complex models. The tools now exist to help you do that.

Step 3: Enhance your data assets

Step 3 is the leap. Once the assembly line has started empowering the key business decisions, it is time to change the rules once again. Look for ways to enhance and build upon your existing data assets.

What data, that you may not have access to today but could in the future, would dramatically improve the performance of your business :

1. How could it deliver incremental value to customers?
2. What would it deliver to customers?
3. What problems could it solve that can't be solved today?
4. What new products and services could be offered to meet unmet needs or create new markets?

For instance, ask how do you take what you have in your existing internal data, be it transactional or behavioral or financial - and add, enhance or improve it with external data or new data that you haven't even captured or new sources that you may be capturing but aren't leveraging. And if the two were able to be combined together, what will it change. I can assure it will change things.

A great example is transactional fraud. Transactional fraud in the United States accounted for $100 billion in lost value[v]. All typical modeling tools as well as data elements that could have been applied to addressing this problem haven been tried. And the problem remains.

Now let's 'leap it'. In experiments performed by a company we mentioned before (Tresata), by combining financial data streams with location information, so a payments provider now has access to not just the financial transactional data at time of sale, but also the location of the user of that payment instrument, one can think about eliminating fraud emanating from stolen payment and identity information. What it means for the consumer is an end to the embarrassment of having your

credit card declined when you are on vacation in Bora Bora, as well as the added protection that if your card was stolen, it could not be used because the thief cannot fake your location data, which can be verified real time. Game changed in fraud…forever. What it required is the capability to manage structured data with unstructured data stored, modeled and delivered at massive scale. Big Data is here.

Whether you look at structured data or unstructured data, ask yourself how do you differentiate your data asset from your competition or from other similar companies?

A peek into the future

So what will these data factories look like?

A great example of one such company is also a personal favorite, Zynga. Zynga is the biggest gaming company on Facebook. Zynga is also redefining gaming as we know it – taking it from the multi-million dollar 2 year development cycles for blockbuster games, and transforming it to real-time massively multi-player simple games with a huge social component (because games are meant to be played with friends). At the heart of it though, Zynga, like Google, is a data and analytics company. And they have architected an organization structure that's equally dynamic. They have hundreds of games that they offer, and have embedded data scientists with every one of them. These data scientists use Tableau to actively look at the data coming from the games the people are playing. They have the ability in real time to change features of the game or the tools they offer in it, be it the weapons in Mafia Wars™ or fruits and vegetables you can buy (yes buy with real dollars) and plant in Farmville™.

Rather than centralizing analytics, they completely democratized it. In the future, organizations will have to do the same. They will have to democratize the access to data at all levels of an organization, forever changing the way we make decisions, deliver products and services and how we charge for it.

Welcome to the future.

About Tableau

Tableau Software helps people see and understand data. Ranked by Gartner and IDC in 2011 as the world's fastest growing business intelligence company, Tableau helps anyone quickly and easily analyze, visualize and share information. More than 7,000 companies get rapid results with Tableau in the office and on-the-go. And tens of thousands of people use Tableau Public to share data in their blogs and websites. See how Tableau can help you by downloading the free trial at *http://www.tableausoftware.com/solutions/big-data.*

About the Author

Abhishek Mehta

Abhishek, former Strategy and Development Executive at Bank of America, is an expert in big data and consumer payments. A featured speaker and leading edge thinker on these topics, Abhishek is a die-hard supporter of all things open source and is recognized in the industry as a visionary on how to create value by building, transforming or disrupting business eco-systems. He has over a decade of experience in various strategic and operational leadership roles in banking, tech and consulting, and has lived and worked in Asia, Europe and the US. Abhishek is also the Founder and President of Foundation Ten10, a one-of-a-kind network driven non-profit focused on training, educating and nurturing children with learning disabilities, and a member of the Faculty at one of the premier Retail Banking Management Programs in the US.

End notes

[i]Estimated based on Net Global Corporate Revenue and Profit pools in 2010 as reported by IMF, World Bank, McKInsey and Others

[ii]Moore's Law states that the number of transistors on a chip will double about every two years - http://www.intel.com/technology/mooreslaw/

[iii]1 terabyte can store around 300 hours of video shot on your iPhone

[iv]Large Hadron Collider project (http://lhc.web.cern.ch/lhc/) is the worlds world's largest and highest-energy particle accelerator. It is expected to address some of the most fundamental questions of physics, advancing the understanding of the deepest laws of nature.

[v]LexisNexis report on fraud in the transactional systems