happiest minds
The Mindful IT Company
Born **Digital** . Born **Agile**

HAPPIEST MINDS TECHNOLOGIES

Big Data 101

Creating Real Value from the Data Lifecycle

Happiest People . Happiest Customers

Contents

Abstract.....3

Introduction.....3

Data lifecycle – phases and their challenges.....4
- Data Acquisition & Data Warehousing
- Data Extraction & Structuring
- Data Modeling & Data Analysis
- Data Interpretation

5Vs of Big Data – Problems & Challenges.....6
- Volume & Scalability
- Heterogeneous & unstructured nature of Big Data
- Data governance and security
- Infrastructure and system architecture

Securing Success in Big Data Projects.....7

Businesses in the Era of Big Data.....8

Conclusion.....8

References.....8

Copyright Information.....9

Abstract

The big impact of Big Data in the post-modern world is unquestionable, un-ignorable and unstoppable today. While there are certain discussions around Big Data being really big, here to stay or just an over hyped fad; there are facts as shared in the following sections of this whitepaper that validate one thing - there is no knowing of the limits and dimensions that data in the digital world can assume. As of now, there are only predictions, forecasts that form the basis for businesses to decide their own course of action by taking advantage of this new window of opportunity with high-end intricate technology based resources and tools. These are nothing but ways and means to harness the mammoth and draw necessary inferences for making smart, intelligent business decisions.

The other outstanding feature of this century is the mushrooming of more billion-dollar enterprises around the world than ever before. This phenomenon is making the world a more competitive, knowledge-rich, mercurial and dynamic place. Survival of the fittest has never been more pertinent in the business sector. This backdrop serves as the most important reason for businesses to be data-driven. It is the only premise that can yield confident, sure-shot and actionable insights to make future-ready, fool-proof decisions. Even though, intuition and acumen – the human factors will always have a large role to play in deciding course of success, a strong data foundation can go a long way in minimizing risks of observational understanding or guess.

While this sounds like a good game to get on, the challenge however, lies in data itself. As the name suggests – it is big, heterogeneous, unstructured and scattered. In that case, how to mine diamonds off rocks? From acquisition of data that comes in an ad hoc manner, to storing the right metadata and data integration – are all different facets with both opportunities and challenges. Data analysis, structuring, retrieval and modeling are other key aspects with their own set of challenges. This paper looks at deriving real value from the big data giant through a look at the data lifecycle, its dimensions and challenges, best practices and application benefits.

Introduction

Eric Schmidt of Google had stated in 2010 that every two days we create as much information as we did from the dawn of civilization up until 2003. That was 2010.

In 2014, an IDC and EMC report stated that the digital universe is doubling every two years, and will reach 40,000 exabytes (40 trillion gigabytes) by 2020. Here are some more interesting and latest statistics on data:

- 90% of the world's data has been created in last two years
- Big Data market growth projection is $53.4 billion by 2017, from $10.2 billion in 2013
- 70% of the digital universe, approx. 900 exabytes are generated by users
- 98% of global information is now digital, that was 25% in 2000
- 10% increase in data visibility means additional $65.7 million for a typical Fortune 1000 company

These facts contribute to the imperative that enterprises must develop effective methodologies and upscale capabilities to gather, process, and harvest data, now before it gets too late. The world spent a large part of 2013 in data collection. It is now essential to create capabilities that can narrow down big data into relevance and importance, and carve out only the information that matters most to the business. The Big Data Revolution is not in the volume explosion of data, but in the capability of actually doing something with the data, making more sense out of it.

In order to build a capability that can achieve beneficial data targets, enterprises need to understand the data lifecycle and challenges at different stages.

Data lifecycle – phases and their challenges

The end-to-end lifecycle of data involves multiple phases as represented below, with challenges at each phase. Big Data needs contextual management from a haphazard, heterogeneous mix. There is also a question of credibility, uncertainty and error in understanding the relevance of data. This management of data requires smart systems and also better human collaboration for user interaction. A lot of debate on Big Data is focused on the technology aspect. However, there is also a lot more than technology required to set up the fundamental basis of managing data analysis. It doesn't reckon throwing away of existing structures, warehouses and analytics. Instead, needs to build up on existing capabilities in data quality, Master Data Management and data protection frameworks. Data Management needs to be seen from a business perspective, by prioritizing business needs and taking realistic actions.

Data Acquisition & Data Warehousing

Data always has a source. It doesn't come out of nowhere. And just as big as data is, so are the multifarious sources that can produce up to 1 million terabytes of raw data every day. This enormity and dispersion in data is not of much use, unless it is filtered and compressed on the basis of several criteria. The foremost challenge in this aspect is to define these criteria for filters, so as to not lose out any valuable information. For instance, customer preference data can sourced from the information they share on key social media channels. But then, how to tap the non-social media users who might also be an important customer segment. What are the data sources for them?

Data reduction is a science that needs substantial research to establish an intelligent process that brings down raw data to a user-friendly size without missing out the minute information pieces of relevance. And this is required in real-time, as it would be an expensive and arduous affair to store the data first and reduce later.

An important part of building a robust Data Warehousing platform is the consolidation of data across various sources to create a good repository of master data, which will help in providing consistent information across the organization.

Data Extraction & Structuring

Data that has been collected, even after filtering, is not in a format ready for analysis. It is has multiple modes of content, such as text, pictures, videos, multiple sources of data with different file formats. This mandates for a Data Extraction Strategy that integrates data from diverse enterprise information repositories and transforms it into a consumable format.

Data is basically of two categories – structured and unstructured. Structured data is that which is available in a pre-set format such as row and column based databases. These are easy to enter, store and analyze. This type of data is mostly actual and transactional.

Unstructured data on the other hand is free form, attitudinal and behavioral. This does not come in traditional formats. It is heterogeneous, variable and comes in multiple formats, such as text, document, image, video and so on. Unstructured data is growing at a super-fast speed. In 2011, IDC held a study that stated that 90 percent of all data in the next decade will be unstructured. However, from a business benefit perspective, true value and insights reside in this massive volume of unstructured data that is rather difficult to tame and channelize.

Extract-Transform-Load (ETL) is the process that covers the entire stage of getting data loaded in the proper, cleaned format from the source to the target data warehouse. There are several ETL tools in available, principles of making the right selection are same as that of deciding the right course of big data implementation as explained later in the paper.

Data Modeling & Data Analysis

Once the proper mechanism of creating a data repository is established, then sets in the rather complex procedure of Data Analysis. Big Data Analytics is one of the most crucial aspects and room for development in the data industry. Data analysis is not only about locating, identifying, understanding, and presenting data. Industries demand for large-scale analysis that is entirely automated which requires processing of different data structures and semantics in an understandable and computer intelligent format.

Technological advancements in this direction are making this kind of analytics of unstructured possible and cost-effective. A distributed grid of computing resources utilizing easily scalable architecture, processing framework and non-relational, parallel-relational databases is redefining data management and governance. Databases today have shifted to non-relational to meet the complexity of unstructured data. NoSQL database solutions are capable of working without fixed table schemas, avoid join operations, and scale horizontally.

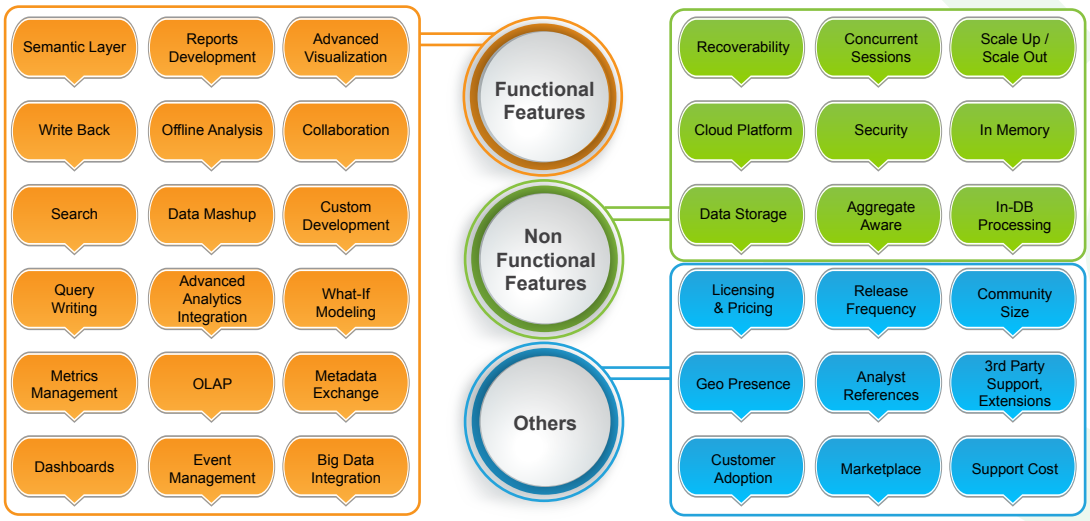
Data sciences has emerged as a sophisticated discipline that draws from various elements across statistical techniques, mathematical modelling and visualization. It encapsulates:

- Data manipulation & analytic applications addressing automation, application development and testing
- Data modelling covering key areas like experimental design, graphical models and path analysis
- Statistics and machine learning through classical and spatial statistics, simulation and optimization techniques Text data analysis through pattern analysis, text mining
- and NLP by developing and integrating solutions or deploying packaged solutions

Data Interpretation

The most important aspect of success in Big Data Analytics is the presentation of analyzed data in a user-friendly, re-usable and intelligible format. And the complexity of data is adding to the complexity of its presentation as well. Sometimes, simple tabular representations may not be sufficient to represent data in certain cases, requiring further explanations, historical incidences, etc. Sometimes, predictive or statistical analysis from the data is also expected from the analytics tool to support decision-making. In other words, the final phase or culmination of the entire Big Data exercise is Data Interpretation or Data Visualization.

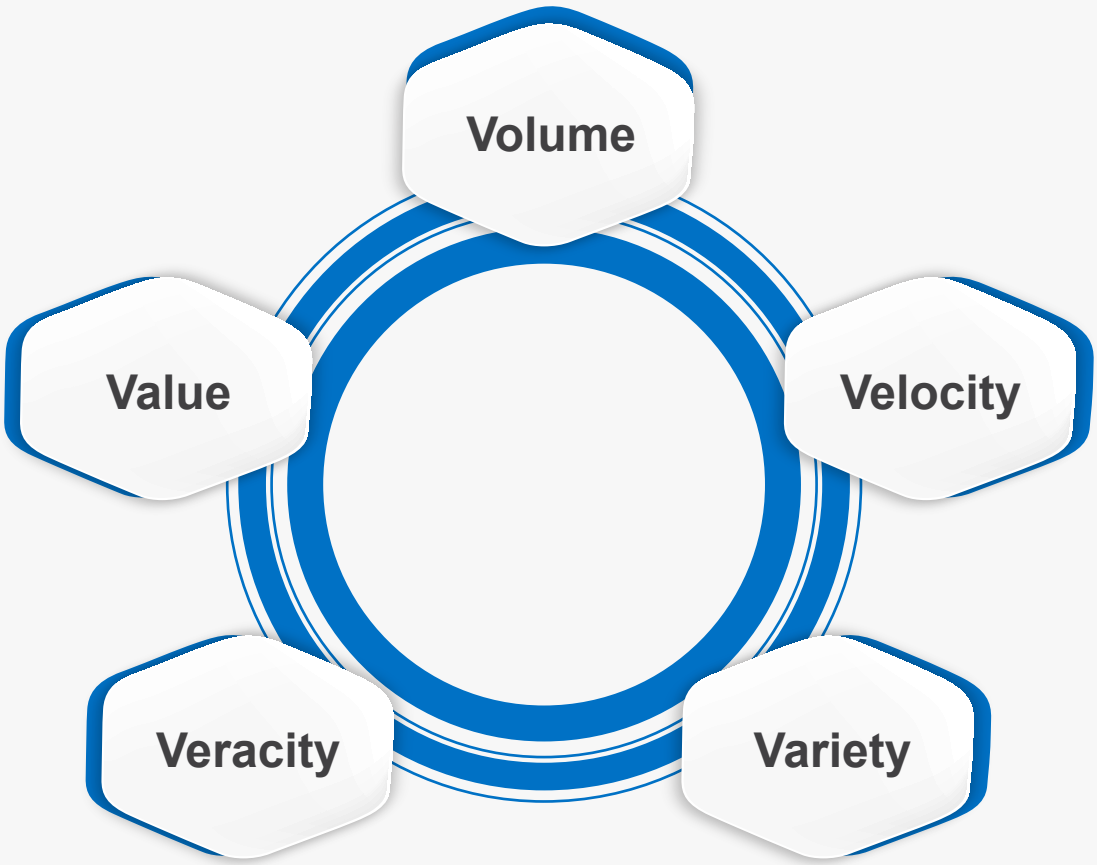
Visualization of data is a key component of Business Intelligence. Here's a snapshot of the Visualization framework that assists in Business Intelligence.



Interactive Data Visualization is the next big thing that the industry is moving into. From static graphs and spreadsheets to using mobile devices and interacting with data in real time – the future of data interpretation is becoming more agile and responsive.

5Vs of Big Data – Problems & Challenges

In 2001, Gartner had introduced the 3V's definition of data growth that was then in the inception stages. 3V signified the three dimensions of – Volume, Velocity and Variety. Now, the industry recognizes 5V, adding Veracity and Value as the additional aspects of data. While volume, velocity and variety are spices that the world has already tasted and still continue to be a major consideration in the data domain, veracity and value are aspects of the modernized data that throw up major challenges. Increasing channels or sources of data such as social media has made users a major part of data contribution and consumption. This is a boon and a bane at the same time. While it opens up a huge window into understanding consumers, there is a massive amount of junk that is created at various levels. Credibility and trustworthiness of data has never been of more concern before. Value, of course, is the big thing out of big data that is the driver of the entire revolution. Businesses no longer look only for access to Big Data, but generating true value out of it is the real objective.



This five-dimensional emergence of big data has given rise to related problems that cut across all the phases of the data lifecycle.

- **Volume & Scalability**

  This is the basic problem that every system or tool grapples with when dealing with Big Data – it is big and there is no knowing of the limits of its scale. Therefore, Big Data tools and infrastructure need to ensure enough scalability and elasticity to be able to handle the sonic speed of data growth.

- **Heterogeneous & unstructured nature of Big Data**

  As explained earlier, most data is unstructured and therefore, heterogeneous is nature. In terms of sources, formats, modes and feeds – data influx happens in all shapes and sizes. Analytic tools therefore need to be smart enough to decipher all the diverse natures of data, assimilate them with advanced algorithm development, optimization and automation to bring it on a uniform, consumable format.

- **Data governance and security**

  Increase in mobility and access to information has led to massive discussions around data governance, protection and security. Industries such as banking, healthcare, pharma, and defense are under strict compliance and regulatory mandates that make it a tough job to create a proper data protection framework. It is not enough to have an IT infrastructure and security in place. Data governance has taken primary importance in these sectors where opportunity is big in Big Data, but risks can be huge.

- **Infrastructure and system architecture**

  While the advanced technologies of Hadoop and MapReduce are scaled to meet the 5Vs of big data, they assert significant demands on infrastructure in terms of scale, storage capacities that are efficient and cost-effective. Intelligent storage capacities can leverage through data compression, automatic data tiering and data deduplication. The question is how much is needed to implement Big Data and how much is enough.

Securing Success in Big Data Projects

With all its opportunities and challenges, there are certain guiding principles in the implementation of Big Data that can push the envelope for success. The secret lies in a robust data strategy and data management program that is aligned to the business goals and strategies.

Points to be noted and considered before jumping into big data are:

- Discern business requirements before beginning to gather data – what is the actual business need
- Big Data implementation is a business decision, not an IT or technological function
- Take small steps to Big Data – an agile and iterative implementation approach can go a long way in yielding results giving the business space to respond, adapt and realize the true value in Big Data
- Standardize Big Data efforts into IT governance program in order to make up for skill shortages
- Align Big Data strategy with Cloud Strategy to tackle several issues around storage, security and scalability
- Embed Data Analytics into the system DNA to see real value by including big data into organizational data and breaking silos of teams

Businesses in the Era of Big Data

Big Data, being big enough, is merely just warming up. Soon, every aspect of the human society is going to be affected by big data. There is scope for every sector in every field to make great use of Big Data in enabling better lives for users. Here are some classic examples of Big Data in play in the field of business:

- Deeper levels of understanding and targeting customers: A large US retailer has been able to accurately predict when a customer of theirs is expecting a baby. Churn management has become easily predictable for telecom companies and car insurance companies are able to understand how well their customers are driving.
- Optimizing Business Process: Big Data is not only giving a peek into the external audience, but also a great way for introspection into business processes. Stock optimization in retail through predictive analysis from social media, web trends and weather forecasts is leading huge cost benefits. Supply chain management is particularly benefitting from data analytics. Geographic positioning and radio frequency identification sensors can now track goods or delivery vehicles and optimize routes by integrating live traffic data.
- Driving smarter machines and devices: The recently launched and widely talked about Google's self-driven car is majorly using Big Data tools. The energy sector is also taking advantage by optimizing energy grids using data from smart meters. Big Data tools are also being used to improve performance of computers and data warehouses.
- Smarter Financial Trading: High-Frequency Trading (HFT) is finding huge application of Big Data today. Big Data algorithms used to make trading decisions has led to a majority of equity trading data algorithms taking into account data feeds from social media networks and news websites to make decisions in split seconds.

These are some of the existing illustrations where Big Data is in application in the business sector. There are several other avenues, with newer ones opening by the day, where Big Data can drive organizations into being smarter, secured and connected.

Conclusion

Big Data is the biggest opportunity in the modern world. There is a whole set of advantages that businesses can yield from big data and evolve as smart and connected enterprises. Also, it has now become evident that Big Data is permeating all aspects of life, making it an imperative for businesses, whether they are ready or not. However, there are also unique and new challenges thrown up by the data revolution that enterprises need to be careful about. With proper caution and preparedness, a business can thrive into excellence with big data, without being exposed to a risk situation. Knowing how big data works and being aware of the challenges is one of the first steps towards preparing for taking on the big elephant.

References:

- http://www.ibmbigdatahub.com/blog/10-big-data-implementation-best-practices
- http://www.intel.in/content/dam/www/public/us/en/documents/solution-briefs/big-data-101-brief.pdf
- http://hbr.org/special-collections/insight/big-data
- http://www.experian.co.uk/assets/business-strategies/white--papers/wp-big-data-evolution-not-revolution.pdf
- http://tdwi.org/whitepapers/list/advanced-analytics/big-data-analytics.aspx

About the Author

Anand Veeramani
Practice Director - Big Data & Database Services

Anand Veeramani has 15 years of IT experience, including 3 years in Big Data technologies. He has worked in various roles including Chief Architect, SME, Practice Lead, Account Manager, Transition Manager, Senior DBA for different database & Big data services. Key past projects include: - Architecting Enterprise Data warehouse solutions for the largest Bank in USA - Managing Complex database Consolidation & Migration Initiative for largest Bank in USA - Managing end to end database service transition for leading Airlines in North America

Happiest Minds

Happiest Minds, the Mindful IT Company, applies agile methodologies to enable digital transformation for enterprises and technology providers by delivering seamless customer experience, business efficiency and actionable insights. We leverage a spectrum of disruptive technologies such as: Big Data Analytics, AI & Cognitive Computing, Internet of Things, Cloud, Security, SDN-NFV, RPA, Blockchain, etc. Positioned as “Born Digital . Born Agile”, our capabilities spans across product engineering, digital business solutions, infrastructure management and security services. We deliver these services across industry sectors such as retail, consumer packaged goods, edutech, e-commerce, banking, insurance, hi-tech, engineering R&D, manufacturing, automotive and travel/transportation/hospitality.

Headquartered in Bangalore, India; Happiest Minds has operations in USA, UK, The Netherlands, Australia and Middle East.

Business Contact: business@happiestminds.com **Media Contact:** media@happiestminds.com

Copyright Information

E-mail: business@happiestminds.com
Visit us: www.happiestminds.com

Follow us on