2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

Big Data Analytics

Sachchidanand Singh
Business Analytics Division
IBM India Software Lab (ISL)
Pune, India
sachcsin@in.ibm.com

Nirmala Singh
Data Warehouse Division
Mahindra Satyam
Pune, India
nirmala_singh@mahindrasatyam.com

Abstract—In this paper, we explain the concept, characteristics & need of Big Data & different offerings available in the market to explore unstructured large data. This paper covers Big Data adoption trends, entry & exit criteria for the vendor and product selection, best practices, customer success story, benefits of Big Data analytics, summary and conclusion. Our analysis illustrates that the Big Data analytics is a fast-growing, influential practice and a key enabler for the social business. The insights gained from the user generated online contents and collaboration with customers is critical for success in the age of social media.

Keywords: InfoSphere BigInsights, WX2, Massively Parallel Processing (MPP), Data-warehousing-as-a Service (DaaS), ParAccel Analytic Database (PADB), Omne optimizer & SAND Analytic Platform

I. INTRODUCTION

IBM estimates that every day 2.5 quintillion bytes of data are created – so much that 90% of the data in the world today has been created in the last two years. [1] It is a mind-boggling figure and the irony is that we feel less informed in spite of having more information available today.

The surprising growth in volumes of data has badly affected today's business. The online users create content like blog posts, tweets, social networking site interactions and photos. And the servers continuously log messages about what online users are doing.

The online data comes from the posts on the social media sites like facebook and twitter, YouTube video, cell phone conversation records etc. This data is called Big Data.

II. WHAT IS BIG DATA?

Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools. The difficulty can be related to data capture, storage, search, sharing, analytics and visualization etc.

The Big Data spans across three dimensions: Volume, Velocity and Variety.

- Volume – The size of data is very large and in terabytes and petabytes.
- Velocity – It should be used when streaming in to the enterprise in order to maximize its value to the business. The role of time is very critical here.
- Variety – It extends beyond the structured data, including unstructured data of all varieties: text, audio, video, posts, log files etc.

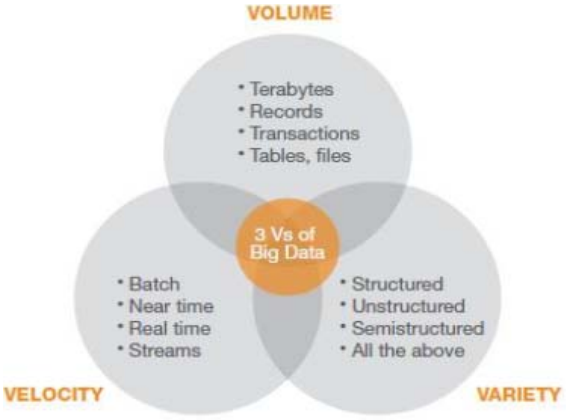

Figure 1. The three Vs of Big Data

III. WHY BIG DATA?

When an enterprise can leverage all the information available with large data rather than just a subset of its data then it has a powerful advantage over the market competitors. Big Data can help to gain insights and make better decisions.

Big Data presents an opportunity to create unprecedented business advantage and better service delivery. It also requires new infrastructure and a new way of thinking about the way business and IT industry works. The concept of Big Data is going to change the way we do things today.

The International Data Corporation (IDC) study predicts that overall data will grow by 50 times by 2020, driven in large part by more embedded systems such as sensors in clothing, medical devices and structures like buildings and bridges. The study also determined that unstructured information - such as files, email and video - will account for 90% of all data created over the next decade. But the number of IT professionals available to manage all that data will only grow by 1.5 times today's levels. [2]

The digital universe is 1.8 trillion gigabytes in size and stored in 500 quadrillion files. And its size gets more than double in every two years time frame. If we compare the digital universe with our physical universe then it's nearly as many bits of information in the digital universe as stars in our physical universe. [3]

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

IV. CHARACTERISTICS OF BIG DATA

A Big Data platform should give a solution which is designed specifically with the needs of the enterprise in the mind. The following are the basic features of a Big Data offering-

- Comprehensive - It should offer a broad platform and address all three dimensions of the Big Data challenge -Volume, Variety and Velocity.
- Enterprise-ready - It should include the performance, security, usability and reliability features.
- Integrated - It should simplify and accelerates the introduction of Big Data technology to enterprise. It should enable integration with information supply chain including databases, data warehouses and business intelligence applications.
- Open source based - It should be open source technology with the enterprise-class functionality and integration.
- Low latency reads and updates
- Robust and fault-tolerant
- Scalability
- Extensible
- Allows adhoc queries
- Minimal maintenance

V. BIG DATA OFFERINGS

There are many vendors offering Big Data analytics solutions like IBM, Kognitio, ParAccel & SAND etc. The following are the key Big Data offerings available in the market for the enterprise use-

A. IBM InfoSphere BigInsights

It is an Apache Hadoop based solution to manage and analyze massive volumes of the structured and unstructured data. It is built on an open source Apache Hadoop with IBM Big Sheet and has a variety of performance, reliability, security and administrative features. The IBM Big Sheet is a sophisticated text analytics module used for data exploration. The BigInsights is able to analyze data in its native format without imposing any schema/structure and enables fast ad-hoc analysis.

The BigInsights on the Cloud are available in both basic and enterprise editions with the options of public, private and hybrid cloud deployments. The basic edition is an entry-level offering available at no charge and it helps organizations to learn about Big Data analytics. The clients can seamlessly move to the enterprise edition when they are ready and can set up a Hadoop cluster in approximately 30 minutes to start data analysis with low usage rate of $0.60 per cluster/hour. The basic and enterprise edition both versions include a developer sandbox where clients can develop a new generation of business analytics applications.

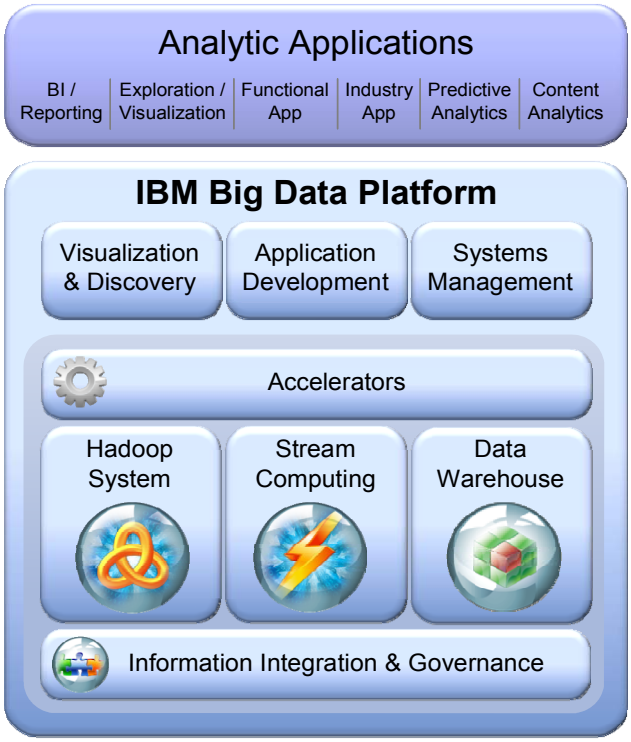

Figure 2. IBM Big Data Platform

The IT professionals and students looking to build a career and skills in Big Data & Apache Hadoop can take advantage of IBM's BigDataUniversity.com website where users can learn the basics of Hadoop, stream computing and open-source software development.

B. WX2 Kognitio Analytical Platform

It is a fast and scalable in-memory analytic database platform. It can be deployed in any of the 03 ways: as a software-only license, as a fully configured data warehouse appliance running on industry-standard hardware, or on-demand via Kognitio's cloud-based data-warehousing-as-a service (DaaS) solution.

The biggest (measured by data) WX2 customer has bought a license for 9 ½ terabytes of user data. The typical Kognitio WX2 configurations have a couple of hundred gigs of user data per CPU core. The biggest current system measured by nodes has 300 servers. [4]

C. ParAccel Analytic Platform

It offers ParAccel Analytic Database (PADB) which is a columnar, massively parallel processing (MPP) analytic database platform. It has strong features for the query optimization and compilation, compression and network interconnect.

ParAccel Analytic Database (PADB) is a schema-neutral database and its users employ agile load-and-go analytic methodology. It comes with Omne optimizer which can optimize SQL code of any length and complexity irrespective

of its poor structure. A user can develop routines for parallelized in-database execution using the ParAccel's Extensibility Framework.

D. IBM InfoSphere Streams

It enables continuous analysis of massive volumes of streaming data with sub-millisecond response time. This IBM offering provides a highly scalable and agile infrastructure which can support a wide variety of structured and unstructured data types.

E. SAND Analytic Platform

It is a columnar analytic database platform that achieves linear data scalability through massively parallel processing (MPP). It breaks the constraints of the shared-nothing architectures with fully distributed processing and dynamic allocation of resources. It has been designed to supports thousands of concurrent users with mixed workloads and infinite query optimization. Also it supports in-memory analytics, full text search and SAND box for the immediate data testing.

The SAND Analytic Platform focuses on the complex analytic tasks including customer loyalty marketing, churn analytics and financial analytics.

VI. BIG DATA ADOPTION TRENDS

Big Data analytics is a fast growing and influential practice. International Data Corporation (IDC) forecast shows that Big Data technology and services market is expected to grow up to $16.9 billion in 2015. It represents a compound annual growth rate (CAGR) of 40% or about 7 times that of the overall information and communications technology (ICT) market. [5]

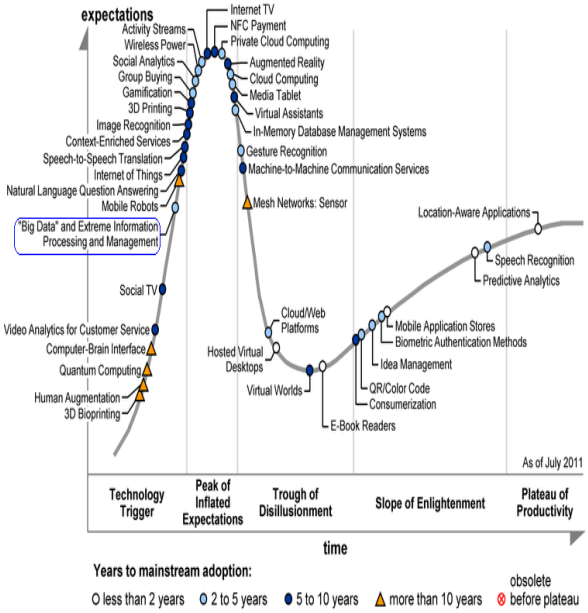

Figure 3. Gartner Hype Cycle [6] [7]

VII. VENDOR AND PRODUCT SELECTION

In this section we will define the entry and exit criteria for vendor and product selection. The following are the main criteria for the selection of vendor and product-

- Scalability supports
- Integration with existing systems
- Performance
- Complexity
- Security & Reliability
- Pricing
- Adherence to regulatory and legal compliance
- Vendor's financial stability
- Vendor comparison based on the selected criteria

VIII. BIG DATA BEST PRACTICES

In this section will focus on the best practices to be followed while working with Big Data offerings. The important points are listed below-

- Build understanding of Big Data & its offerings may be by self study, consult or involve expert
- Start small & take baby steps, build the confidence gradually
- Plan optimal security
- Identify the right strategy - Do risk management
- Do proof of concept (PoC) & pilot projects
- Provision for the surprises & changes

IX. CUSTOMER SUCCESS STORY

A. Vestas Wind Systems

Vestas Wind Systems is a Danish energy company using IBM Big Data analytics software and powerful IBM systems to improve wind turbine placement for optimal energy output. Turbine placement is a major challenge for the renewable energy industry and Vestas expects to accelerate the adoption of wind energy internationally and expand its business into the new markets.

B. El Corte Inglés (ECI)

El Corte Inglés (ECI) is a Spanish company using Kognitio WX2 to substantially increase accuracy in all aspects of the marketing campaign planning. ECI has over 60 stores across Spain. And it offers everything from food to furniture.

C. TRA

TRA is a media and marketing research company. They have chosen Kognitio's WX2 as the underlying database solution for TRA's Media TRAnalytics(TM) ROI Measurement Solution. The TRA selected WX2 because based on the database growth projection and its current commercial database platform was not able to perform the rapid analysis.

D. Merkle

Merkle is a leading database marketing agency using ParAccel Analytic Database (PADB) for its newest generation of KnowledgeLink applications. It is a combination of technology, process and service required to create and maintain an accurate and comprehensive representation of a customer across multiple channels and business lines.

E. Sypherlink

Sypherlink is a leading technology firm using the SAND Analytic Server as a high-performance analytic platform for the customers whose applications require a physical centralized database.

X. BENEFITS OF BIG DATA

The McKinsey Global Institute (MGI) conducted study on Big Data in the five domain areas - healthcare in the United States, the public sectors in Europe, retail in the United States and manufacturing and personal-location data globally. And their study confirms that Big Data can generate value in each of the listed domain areas. For example a retailer using Big Data can increase its operating margin by more than 60 percent. [8]

If US healthcare plans to use Big Data creatively and effectively to drive efficiency and quality then the sector could create more than $300 billion in value every year. Two-thirds of that would be in the form of reducing US healthcare expenditure by about 8 percent. [8]

The government administrators in the developed European economy could save more than €100 billion ($149 billion) in the operational efficiency improvements using Big Data solutions. [8]

XI. BIG DATA CHALLENGES

The main challenges of Big Data are data variety, volume, analytical workload complexity and agility. Many organizations are struggling to deal with the increasing volumes of data. In order to solve this problem, the organizations need to reduce the amount of data being stored and exploit new storage techniques which can further improve performance and storage utilization.

XII. SUMMARY AND CONCLUSION

Big Data is a new gold rush & key enabler for the social business. A large or medium sized company can neither make sense of all the user generated content online nor can collaborate with customers, suppliers and partners effectively on social media channels without using Big Data analytics. The collaboration with customers and insights from user generated online contents are critical for the success in the age of social media.

In a study by McKinsey's Business Technology Office and McKinsey Global Institute (MGI) firm calculated that the U.S. faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of Big Data.[9]

The biggest gap is the lack of the skilled managers to make decisions based on analysis by a factor of 10x.Growing talent and building teams to make analytic-based decisions is the key to realize the value of Big Data.

ACKNOWLEDGMENT

Our thanks to AIEEE for allowing us to modify manuscript templates they had developed for conference proceedings.

REFERENCES

[1] Improving Decision Making in the World of Big Data http://www.forbes.com/sites/christopherfrank/2012/03/25/improving-decision-making-in-the-world-of-big-data/

[2] World's data will grow by 50X in next decade, IDC study predicts http://www.computerworld.com/s/article/9217988/World_s_data_will_grow_by_50X_in_next_decade_IDC_study_predicts

[3] The 2011 Digital Universe Study: Extracting Value from Chaos http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

[4] Kognitio WX2 overview
http://www.dbms2.com/2008/01/26/kognitio-wx2/

[5] IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy http://outsourcing.ulitzer.com/node/2195534

[6] Gartner Hype Cycle 2012 for Emerging Technologies http://sembassy.com/wp-content/uploads/2011/10/gartner-hype-cycle-2012.gif

[7] Gartner Hype Cycle 2012
http://www.gartner.com/id=2065716

[8] Big data: The next frontier for innovation, competition and productivity http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

[9] Big data: The next frontier for competition http://www.mckinsey.com/features/big_data