Journal of Information Technology Management

ISSN #1042-1319

A Publication of the Association of Management

CHALLENGES AND BEST PRACTICES FOR ENTERPRISE ADOPTION OF BIG DATA TECHNOLOGIES

SUBHANKAR DHAR
SAN JOSE STATE UNIVERSITY
subhankar.dhar@sjsu.edu

SOURAV MAZUMDER
IBM SOFTWARE GROUP
smazumder@us.ibm.com

ABSTRACT

In this paper, we discuss the emerging use of Big Data technologies in Enterprise Data Warehouses and Business Intelligence for the purpose of fostering innovation and producing better business insights and decisions. We present best practices and techniques of Big Data implementation using a 3-legged Big Data environment strategy along with the challenges of enterprise adoption of Big Data technologies. Hence this paper is relevant from an academic as well as a practitioner's perspective.

Keywords: Business Analytics, Big Data, Data Mining, Map Reduce, Hadoop, NoSQL, Data Warehouse

INTRODUCTION

Organizations typically store their critical data arising from various transactions in relational databases. However, there is a large amount of unstructured data generated from blogs, images, email, social media, scientific experiments, various surveys, etc., that contain useful information. Mining these kind of data is becoming extremely useful for making intelligent business decisions.

Owing to the fact that this non-traditional data accounts for 80% of enterprise data and growing in leaps and bounds, in last few years, enterprises have already started putting enough focus and effort to figure out how to bring in the traditional and non-traditional data together under one common umbrella of Enterprise Data Management strategy. In that attempt, enterprises have started trying out a new genre of data management technologies, popularly known as Big Data technologies. However, not much of success has been seen so far due to the still evolving nature of the Big Data technologies which poses challenges in adhering to the Quality of Service aspects any enterprise needs to cater for in Enterprise Data Management.

In this paper we discuss some of those key challenges to bring in traditional and non-traditional data under the same umbrella of Enterprise Data Management. We also aim to cover the best practices so far found to be helpful in alleviating those challenges. In overall course of this paper we also cover an overview of Enterprise Data Landscape and key characteristics of Big Data technologies to set the required context.

OVERVIEW OF BIG DATA LANDSCAPE

Various analyst reports predict that volume of data is growing at an annual rate of 40%. IDC forecasted 50 folds growth of digital universe from the beginning of 2010 to the end of 2020 [7]. In overall this data volume can be classified in three major categories.

Transactional Data: This type of data is typically generated by controlled interactions between enterprise and its customers and other internal/external stakeholders through a defined set of enterprise applications and interfaces. The structure of this data is decided in design time and generally relational in nature. For example, data generated by ecommerce, ERP (Enterprise Resource Planning), SCM (Supply Chain Management), CRM (Customer Relationship Management), HR (Human Resource) systems, AIS (Accounting Information Systems) including general ledger are all part of transactional data.

Observational data

This type of data is typically generated by machines/sensors as ancillary to the main application data while business processes get executed. The structure of this data though typically decided in design time but they are non-relational in nature. The examples include data arising from blogs, sensors that monitor specific events, customer call logs in call centers, etc.

Social data or Interaction data

This type of data is typically gathered from voluntary participation of the stakeholders of an enterprise through a defined process or casual interactions. The structure of this data is typically open/free flowing and not decided in design time of the enterprise processes or applications. This includes feedback from customers, information gathered from social media like Twitter, Google Plus, LinkedIn and Facebook etc.

As expected, due to the volume and disparate nature of data within it, the Enterprise Data Landscape in today's world, badly suffers from challenges with respect to data storage, consumption and quality. The key challenges are noted below:

1. Inability to store and process ever growing (with sudden jump) owing to the high cost traditional EDW technology solutions.
2. In ability to ingest data in quick turnaround time given design, cleansing and distillation required to align to semantics and syntax of enterprise canonical data model.
3. High latency between the time data is generated to the time when data is available for consumption.
4. Difficulty in verification of authenticity and precision of data
5. Inability to demonstrate overall value of storing and processing high volume of data with flexibility and quick turnaround time.
6. Complexity in integrating the non-traditional kinds of data with traditional data from other enterprise systems for data mining and analysis.

Characteristics of Big Data Technology

Big Data technologies typically address all of the challenges of Enterprise Data Landscape noted in the above section based on their basic offerings in support for Volume, Velocity, Variety, Value [1] and also Veracity. The key technical features of Big Data technologies which help in addressing these 5 Vs are summarized in Table 1.

BIG DATA ADOPTION CHALLENGES

As described in the previous section, though the Big Data technologies bring required features to handle complex requirements of today's Enterprise Data Landscape, there are still many challenges in adoption of the same at enterprise level [8]. Here we present the key challenges in adoption of Big Data technologies for addressing Enterprise Data Management requirements as seen across the industry verticals.

1. Interoperability: Enterprises have already invested in business intelligence to develop solutions. Seamlessly integrating Big Data solutions with existing Enterprise systems and BI is quite difficult and additional investment is often necessary. There are specific issues related to BI visualization tools, which needs further improvement, providing greater flexibility and support for various forms of data. Data ingestion from upstream systems in a fast and predictable manner is also challenge sometimes.

   Vendors in Hadoop and NoSQL solution landscape are coming out with specific adapters to solve these problems. For example: Cloudera's Tera Data adapter, Tableau Adaptor, IBM's integrated Infospehere BigInsight platform [3].

Table 1: Summary of Technical Features of Big Data

Flexible Schema and Data Storage	♠ Supports frequent need to change data model in transparent way ♠ Supports varied data model requirement of multiple applications with same solution ♠ Support for Data Quality
Multi-Level Key Value Object Store	♠ Supports complex object hierarchy of real life data (instead of force fitting into a tuple model) ♠ Ensures collocation of related information for faster performance
Auto Sharding/ Partitioning	♠ Support horizontal scalability for high volume of data through CAP [2] model ♠ Ensures elastic behaviour of hardware infrastructure based on load
Distributed Data Access and Processing	♠ Supports scalability in processing large volume of data across shards – high read throughput and high write throughput ♠ Ensures bringing in 'processing near the data' ♠ Supports different workloads with priority
Efficient Indexing	♠ Supports secondary index on any fields for search across unrelated and unstructured data ♠ Ensures vertically partitioned index supporting shared nothing data shards
Reliability	♠ Supports high availability and failover in an optimal way without incurring much latency ♠ Ensures durability of data
Access	♠ Supports standard access protocols like JDBC/ODBC, JSON, Rest, etc. ♠ Supports data ownership and isolation
Low Total Cost Of Ownership	♠ All of these above features can be implemented in low cost commodity hardware and can support required SLAs ♠ Supports in premise infrastructure or external cloud model for better cost model ♠ Many of the Big Data technologies are Open Source with Open Core model with low cost for support compared to traditional data warehousing solutions ♠ Even in case of licensed Big Data Technology solution the cost of ownership with growth of data volume does not increase linearly and price per TB is cheaper than traditional data warehouse solution

2. Manageability: Managing a big cluster of hundreds of nodes poses problem related to infrastructure management and initial shocker to organization. Though different vendors provide different support for monitoring, management and recovery of big clusters, complete solution is typically missing which can help regular administration folks to embrace Big Data rapidly.

   In Hadoop ecosystem some of the vendors (e.g. IBM [3], Cloudera [4], HortonWorks [5] etc.) are trying to bring in integrated offerings in the area of manageability. From opens source offerings Ganglia is also used widely for the same purpose.
3. Security: Data, while getting generated and being accessed, need to be controlled properly in enterprise context. Otherwise that can lead to compliance issues, unintentional data loss, and exposure of data to non-legitimate users, accumulation of data without the right quality.

Several architectural patterns are emerging in securing the data from unsolicited and unintentional access. Among them using proxy server to protect regular users from data access is the popular one. Also the technology features available (or soon to be available) like Snapshots, are useful in protecting data.

4. Maturity: Complexity of the Big Data technology space poses challenges in selecting right solution and support vendors. Given multiple technology solutions and vendors mushrooming every month, it is difficult to settle on a technology choice and the partner. Hadoop ecosystem today has support offering from 5 to 6 key vendors including big players like IBM and Microsoft. NoSQL solutions are converging to the extent where most of the solutions provide the majority of the features.

   Selecting a vendor complying with the Open Core model is healthy practice observed by most of the organizations. Also checking for existing credentials, size and ability to provide end to end solution offering is very important while selecting the vendor. By now several reports are available from key analysts' organizations on Big Data technology and vendor trends.
5. Development Scalability and Maintainability: Lack of IDEs, Testing, Deployment and Administration tools (which suit the Big Data scale) make the development phase of Big Data slow and also pose challenges in maintenance. Big Data needs mixed skills in developers ranging from application logic, data modeling as well as infrastructure administration proficiency. Availability of proper set of tools in these areas could help enterprises developing Big Data applications faster and also maintaining the same.

   Recently, IDEs, Test suits and Administration tools for Big Data are more and more emerging from different technology and software service vendors. Integration of existing popular tools (like Eclipse) with Big Data technology perspective is also happening [6].
6. Reusability: Big Data adoption needs proper data modeling and unified Big Data Architecture across structured and unstructured data elements with huge volumes. Otherwise the reusability of the solution at enterprise level cannot be achieved. The sheer volume, velocity and variety of the Big Data poses challenges to the cognitive human mind as they mainly owe to the principles of multi-processing in distributed computing world. People struggle to visualize the Big Data solution comprising all of these aspects and tend to land up with non-reusable solutions.

   Nevertheless the disciplines of Data Science and Big Data architecture are emerging strongly giving rise to more number of patterns and models in these areas with blend of concepts from different part of software technologies and software engineering. Typical two in a box (Data Science merged with Data Architecture) approach works better to define and solve Big Data problems in Enterprises.

BEST PRACTICES IN ADOPTING BIG DATA TECHNOLOGIES

As of today many organizations have actually started their journey in adopting Big Data technologies in last few years. In this section we delve down into details of three key best practices that are already prevalent in industry and proven to be helpful in alleviating challenges in enterprise adoption of Big Data technologies discussed in previous section.

1. Adopt 3-legged Big Data environment strategy seamlessly integrated with each other. The first one is for the Developer community which can be a small scale cluster with very limited data volume. The second environment is for the analyst community which is also a smaller cluster but big enough to prove the use cases with reasonable volume and variety of data. The third one is for the business community which is large scale cluster and will potentially grow over period. The developer builds the use case (with a reasonable set of flexibilities) for research. The researchers try out various options of the use case and decide on the best analytics model and release the same to the business users in the same environment. The business users try the use case/models in research environment and either accepts it or rejects the same. In case of acceptance the use case gets deployed to the business environment. In case of rejection the researchers further tune it either himself with the given flexibility or reaches out to the developer for more flexibilities.

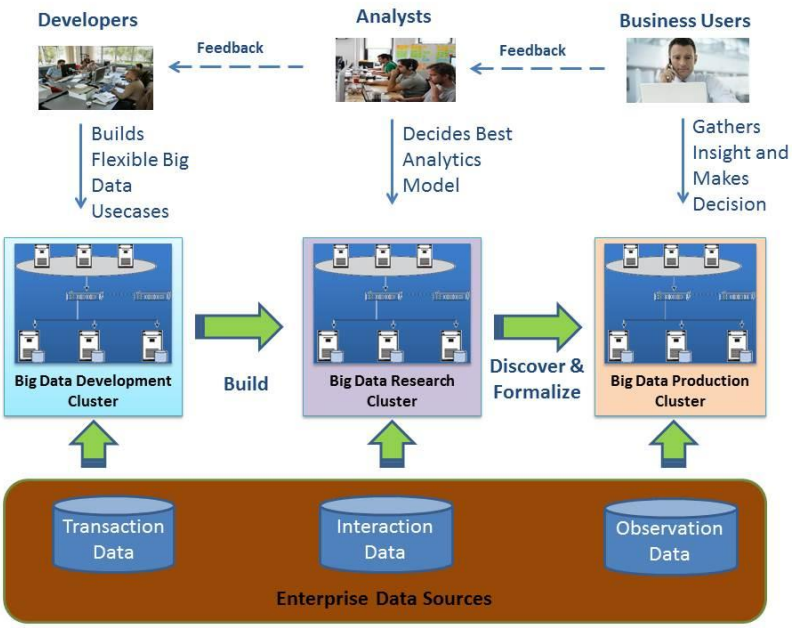

Figure 1: 3-Leg Environment Strategy for Big Data

This strategy not only fosters an iterative model where all key contributors can work in a collaborative way in proving business case for Big Data use cases but also can serve the purpose of sustaining the business case (and maintaining the use cases) in long term.

Agile development is way to go to adopt such a strategy. The typical cycle is PoC (Proof of Concept), Pilot, Refine, Establish, Extend and Standardize. Learning from PoC and Pilot phase has to be refined and then establishing the core process using few use cases/applications. The paradigm of agile development creates an environment of fast learning and aids effective solving of people's skill related issue. Also this can take care of the management and maintainability challenges by making administrator exposed to the solution in early PoC and Pilot cycle of the use case deployment where they get exposed to the deployment nuances.

Overall, this strategy is very helpful in addressing challenges like Reusability, Manageability, Development Scalability and Maintainability.

2. To run with a 3-legged environments with Big Data next best practice is to build standard abstraction layer for data modeling, data integration, data visualization and data management to hide the complexity of data diversity, isolate the security concerns and also facilitate common conduit of integration with rest of the enterprise systems. This will help solving the key challenges like Security, Technology Maturity and Interoperability as well as it will solve developers and modelers' skill issues leading to Development Scalability and Maintainability challenges. Partnering with key stakeholders on enterprise IT, who will use and contribute the Big Data platform is key to define the common abstraction layer.

Again iterative approach is required here to build such abstractions as they cannot be a dependency for starting the Big Data journey.

3. The next step is to create an integrated Big Data workbench and the process around the same to manage the 3- legged environments and abstractions. Building the use case in development environment, then passing the same to researcher and eventually deploying the same to production for business users should be seamless and one click affair. This step can help in addressing the challenges like Manageability, Development, Scalability and Maintainability. Same set of tools used in 3 environments to data ingest, process, visualize and manage helps in establishing the process of 'what you see what you get'. This can take care of the management and monitoring challenges effectively as use case can be easily deployed back to Research and Development environment for troubleshooting and solution.

The integrated Big Data workbench also handles the security concerns transparently as it becomes the only interface to access the data and environment. Similarly modeling challenges can be also addressed through this as this relives the researchers/business analysts from environment and tools related technical complexities.

The integrated Big Data workbench can be either custom developed or any of the off-the-shelf products (like Eclipse with IBM BigInsight Perspective, Talend, Pentaho, Karmasphere) can be adopted.

The integrated Big Data workbench can be either custom developed or any of the off-the-shelf products (like Eclipse with IBM BigInsight Perspective, Talend, Pentaho, Karmasphere) can be adopted.

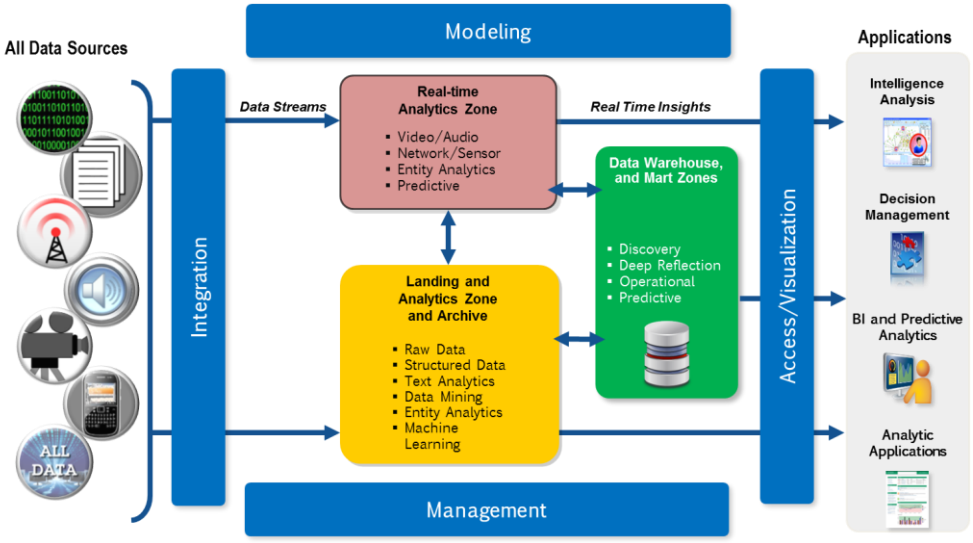

Figure 2: Key Big Data Abstraction to Run in a 3-Leg Environment

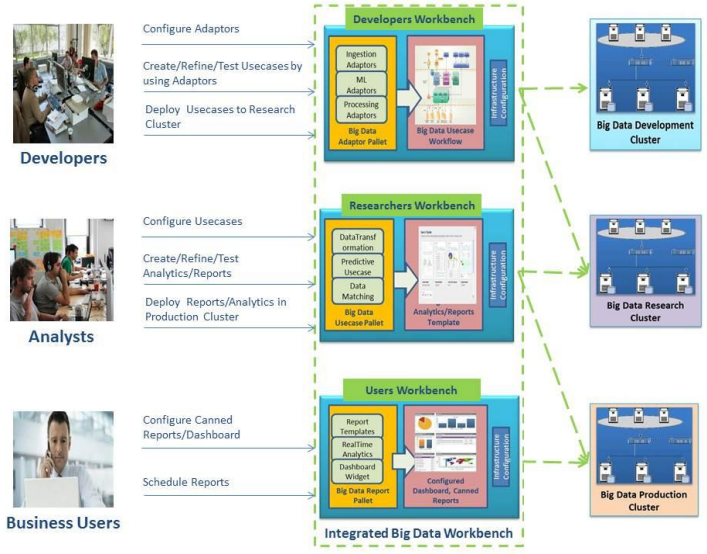

Figure 3: Integrated Big Data Workbench

CONCLUSIONS

In the last few years, Big Data technologies have gained considerable attention due to its potential to transform data mining and business analytics practices and the possibility for a wide range of highly effective decision making tools and services. With new tools, technologies and infrastructure available at our disposal, it has become much easier to capture, store and analyze unstructured data in the enterprise.

The future of Big Data technologies looks very promising as it analyzes all kinds of unstructured data, with a goal to make better decision making. However, enterprise adoption of Big Data is still at its infancy. There are several challenges that need to be addressed before Big Data technologies will attain its full potential.

In this research, we discuss the key characteristics of Big Data technologies along with several challenges, emerging trends and best practices of Big Data implementation using a 3-legged Big Data environment strategy. We believe that our proposed methodology for Big Data implementation will be guiding principles for researchers, practitioners and various organizations that are looking for solutions in this area. We hope that this work will provide key insights to the design and implementation challenges of Big Data Technology solutions.

REFERENCES

[1] Christian, B., Boncz, P., Brodie, M.L., and Erling, O. 2011. "The Meaningful Use of Big Data: Four Perspectives – Four Challenges," in Proceedings of the Twenty-five Semantic Web and Database researchers met at the 2011 STI Semantic Summit, Riga, Latvia.

[2] Dr. Eric A. Brewer, "Towards Robust Distributed Systems", http://www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf

[3] IBM BigInsights. http://www-03.ibm.com/software/products/en/infobigienteedit

[4] Cloudera. 2013. Cloudera Manager Enterprise Edition 4.5.x Release Notes http://www.cloudera.com/content/cloudera-content/cloudera-docs/CM4Ent/4.5.1/Cloudera-Manager-Enterprise-Edition-4.5.x-Release-Notes/Cloudera-Manager-Enterprise-Edition-4.html

[5] Foley, M. and Shah, H. 2012. "Deploying and Managing Hadoop Clusters with Ambari," Hadoop Summit.

[6] Cynthia M. Saracco, Daniel Kikuchi, and Thomas Friedrich, "Developing, publishing, and deploying your first Big Data application with InfoSphere BigInsights" http://www.ibm.com/developerworks/data/library/techarticle/dm-1209bigdatabiginsights/index.html?ca=dat

[7] John Gantz and David Reinsel, "THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East" http://idcdocserv.com/1414

[8] Stonebraker, M., "Big Data, Big Problems," 2011. Communications of the ACM, Volume 55 Issue 2.

AUTHOR BIOGRAPHIES

Subhankar Dhar is a Professor of Management Information Systems at San José State University, San Jose, CA, USA. He is also an affiliate faculty member of the Silicon Valley Center for Entrepreneurship. Dr. Dhar's research interests are in the areas of Big Data, mobile and cloud computing, wireless ad hoc and sensor networks. He teaches a variety of courses including computer networks, distributed systems and web-based computing. His publications have appeared in reputed international journals and gave presentations to various international conferences. He serves as a member of the editorial board of International Journal of Business Data Communications and Networking. He is a reviewer of papers for various international journals, conferences and scholarly publications. He also served as a member of the organizing committee of various international conferences like International Conference on Broadband Networks (BroadNets) and International Workshop on Distributed Computing (IWDC). Dr. Dhar has several years of industrial experience in software development, consulting for Fortune 500 and high-tech industries including product planning, design, and information systems management. Dr. Dhar received his Ph.D. in Mathematics from the University of South Florida.

Sourav Mazumder is a Big Data Architect and Evangelist in IBM Software Group, has over 18 years of IT experience and 6 years in Big Data. He has vast IT experience in industries like Manufacturing, Insurance, Telecom, Banking, Retail, Logistics in USA, Europe, Australia, Japan and India. Influenced key decision makers in fortune 500 companies to embark into Big Data journey as early as in 2010. TOGAF and iCMG certified Software Architect, Sourav is a regular speaker in Big Data conferences with his latest paper published in IEEE-ITMC.