International Journal of Advanced Science and Technology
Vol.115 (2018), pp.37-44
http://dx.doi.org/10.14257/ijast.2018.115.04

A Study of Consideration Factors for the Implementation of Big Data System

Jeong-Beom Kim

Head Professor of BigData Specialist Master Degree Dept. Namseoul Univ.
jbkim@nsu.ac.kr

Abstract

Since big data system implementation are conducted by many organization and commercial companies largely these days, some important factors should be considered during big data related project. The focal point of this paper is to do the case study of consideration factors about big data system implementation focused on large companies or organizations. Before starting the big data system roll-out, important factors should be considered by implementation team with full project scope. Those consideration factors can be whether the tasks are batched oriented, real time processing, large volume of unstructured data, log data oriented, high skilled analysis technologies, Spark platform for in-memory processing , scale out/up infra structure. This study can provide decision makers involved in big data system implementation with perspective and significant views in order to get the better TCO(Total Cost of Ownership). The main purpose of this paper is to define considerable factors which are required by implementation team to do the effective and successful project roll-out.

Keywords: *Big Data, Hadoop, Spark, Implementation Strategy, Performance, Availability, Batch Processing, Real Time Processing*

1. Introduction

1.1. The Needs for this Study

Since big data system implementation by many organization and commercial companies is growing rapidly, effective project roll-out is becoming critical issue in big data related project cases. The main purpose of this paper is to define scopes of considerable factors during the implementation of big data system required by data scientist correctly, specifying the role and guidelines for the purpose of successful project. And this study can help project team with smooth implementation and effective development as useful road map. Below figure shows the trend of big data system [1].

Received (November 26, 2017), Review Result (January 26, 2018), Accepted (February 1, 2018)

ISSN: 2005-4238 IJAST

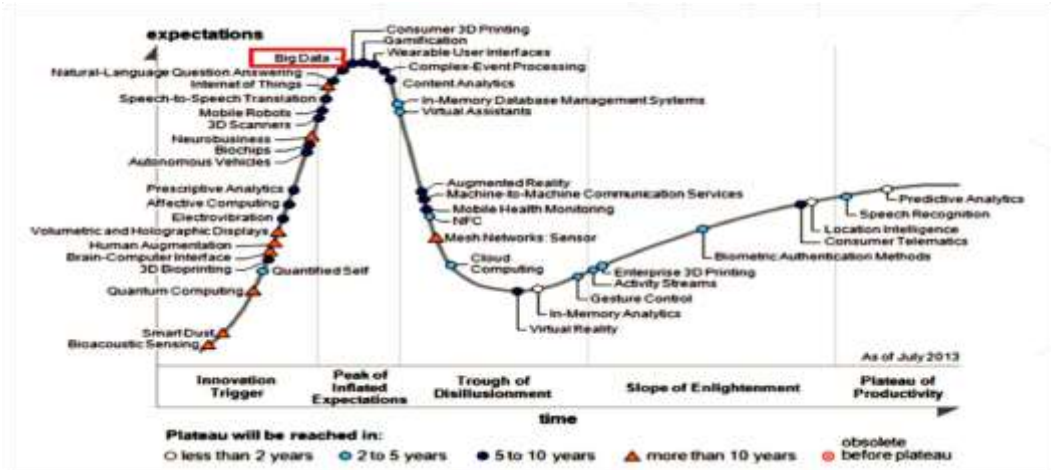

Figure 1. Trend of Big Data System(source by Gartner Report 2014)

1.2. Characteristics of Big Data System Platform

As the 4th industrial revolution emerging in all industry, many companies and government organizations have been implementing big data system project with high expectation. Also the technology level of AI(Artificial Intelligence) is getting improving, similar projects are being implemented. For example IOT and machine learning study have been conducted by many global companies like Google, Apple, and Amazon. In Korea, Samsung Electronics Company and SK Hynix have big data project implemented, and are running big data system at production sites to increase productivity and competiveness. Also many of global companies are aggressively involved in the development of Big Data related technology. For example IBM has invested a lot in machine learning development named WATSON AI with long term vision. As the market size is growing fast, the needs for big data specialist is demanding quickly. There are remarkable characteristics in doing the implementation of big data project compared traditional project as below. Firstly, the hard ware infra structure should have high availability based on open source. The H/W system should be high end level system structure with 3 layer data and dual meta information containers. This can cost more 10% or 20% compared to existing high end system. For the roll-out of SPARK big data system, it is recommended to establish Hadoop system to store the data from the view points of large volume data with longer containing data. GPU(General Processing Unit) is used together with traditional CPU(Central Processing Unit) in case of AI(Artificial Intelligence) project implementation to secure high performance. Secondly, data collection and store method is totally different from traditional system since big data system should handle unstructured data which have various type as well as structured/semi-structured type data. For the collection of unstructured data, crawling agents or tools are being used together with data transfer tools like ETL(Extract Transformation and Loading) or MQ(Message Queue). Thirdly, big data system should be expandable with scale up functions to process big data. And cloud environment is positively recommended to secure high availability as well as performance. Fourthly, to process the big volume data should be processed under distributed environment based on memory processing function. Since SPARK system is I/O(input and out) intensive and CPU intensive processing, it has memory based processing structure while Hadoop system is DISK intensive processing. Big data system refers to the process of big data collection, processing, storing, analysis and visualization or presentation [2,3,4,5,6].

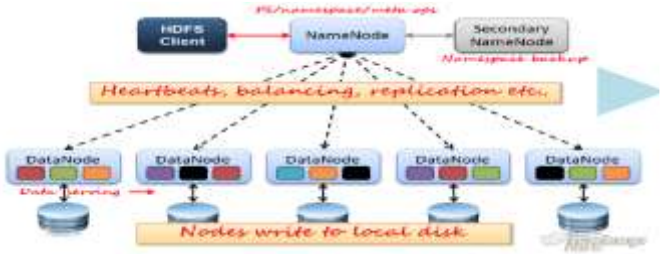

Figure 2. Characteristic of Hadoop Big Data System

2. Consideration Factors for Big Data Project Roll-out Compared with Traditional System Project

2.1. Difference between Big Data System and Traditional System from the Perspective View Points of Implementation

This study is addressing critical factors which are results from many big data project, not from traditional project or survey project data. This study have collected data from many SNS like Facebook, Twitter, and Blogs using R-Studio open source by selecting related key words.

As successful big data implementation, followings strategy should be executed through the study. Firstly, the implementation should be user oriented and integrated system. In details, the project scope needs to be user centered enterprise modeling and collection of user requirements with integrated structure. Secondly, the system should be equipped and deployed with the latest big data technology. Hadoop big data eco system is essential as system infra-structure, and Splunk or Hana system should be implemented as real time analysis solution. Thirdly, Through implementing user oriented big data system, the system needs to be easy to use and provide helpful results. Fourthly, the implementation should be followed by improvement of big data analysis process. As for this, new type process of collection and distribution unstructured data should be adapted, which is the major difference from traditional project. Fifthly, big data specialists group is organized by top management decision. If there are enough big data specialists, additional specialists needed to be scouted from outside. Without proper big data resource power, the successful implementation cannot be guaranteed. One of the key success factors of big data system implementation is keeping specialist called Data Scientists who have the skills of big data system and technology[7,8,9,10,11].

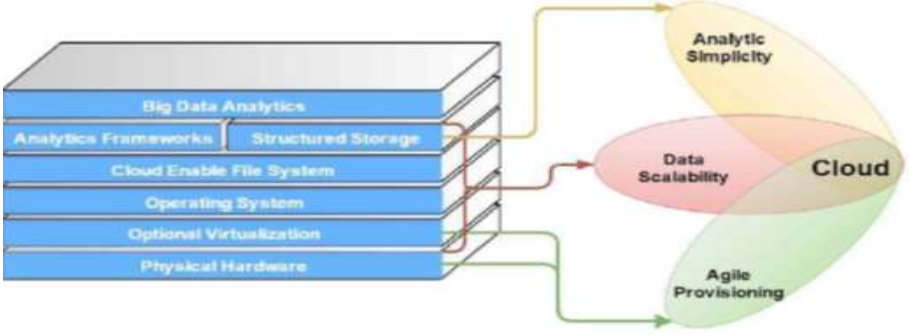

Figure 3. Big Data System Implementation Strategy

2.2. Major Role of Big Data System Implementation Team

To do the successful big data implementation, the role of project team members is important.

There can be four key roles of big data project implementation as following. Firstly, the role of project head is the most important. He should coordinate with project team members by representing user requirements to enable the project to be successful by managing big data project efficiently. If there are problems during the project, he should solve these using mutual communication among project members and report to top management about project progress on regular basis. He can do the decision about the allocation of internal project resource with top priority. Secondly, the role of project team members should be clarified with documents. During the development and implement stage, these members should support user group and development team, and also understand the requirements from users clearly and develop these needs for the system implementation. Thirdly, the role of user support team is also important. They should explain current job process and also address ways of improvement, and coordinate development team about new process planning, as users representing position. They should discuss with develop team about process planning and implementation scope for the purpose of successful project within project time period. They play the role of verifying user requirements by the implemented system and also check whether system is operating well, or not. They should prepare the test data in advance and make test scenario to finish test job. Once the test is finished, they should educate the users how to use the new system for the smooth operation. Fourthly, role of the change manage team should not be ignored. One of the crucial success factors of project is the role of change management team who are checking whether user are using the system without any problem or not, and receiving additional requirements from end user sides as well as upgrading new system to deliver the values to users [12,13,14,15].

Figure 4. Big Data System Trend by Netcraft

2.3. Role and Responsibility of Change Management during Big Data Project Roll-out

According to the project surveys which are related with big data implementation, 65% of unsuccessful cases are related with the absence of change management team or implementation by unskilled team members as major vulnerability. It is forecasted that 70% of enterprise and big organization have the problems with project change management. Below table conveys meaning how to improve change management with perspective point of view.

To solve the vulnerable points of big data project, change manage team should play important role of stream lined process standardization with strict guide line , and easy to use as well as upgrade through deploying smart maintenance activities. One of the best ways for these is to clarify and enhance the role and responsibilities of change management by deploying effective methods which are containing to current situation and organization. For example, change management team should verify the final test whether the system is running without any problems before actual roll-out. Usually, there are four category in change management team organization. The first one is initiator of change management who is doing the role of submitting change requirement by end user side. The second one is change manage manager who is doing the key role of change management, reviewing the requirements and relay them to final approval manager. The

third one is final approval who is doing the role of final approval with proper methods. The last one is change manage team who is doing the role of change management on site[16,17,18].

3. Consideration Factors about Big Data Project Roll-Out

3.1. Consideration Factors about Business Process Design

During the stage of business process design, data process and service process should be considered. Data process includes the specification about data handling process, knowledge visualization process, and data collection structure along with containing policy. Service process includes the specification about service target, service hurdle policy, and maintenance target.

3.2. Consideration factors about Data Collection

During the stage of data collection process, data collection tools should be considered. Data crawling process includes the specification about network separation and connection in terms of data security, data preprocessing by ETL(Extract Transformation, and Loading) or ESB(Enterprise Service Bus) or Crawling Agent, data structure which are structured or semi-structured or unstructured, information of data source which are DBMS(Data Base Management System) or NoSQL(Not only Structured Query Language) or HDFS(Hadoop Distributed File System), and process for identification unable for the purpose of security. Data collection tools can be Scribe, Flume, Chukwa. Data Integration tool can be sqoop. Queue tools can be Kafka, Bookeeper. Web Crawler tools can be Heritrix, Nutch, Scrapy, PHP-crawler, awesome-crawler.

3.3. Consideration factors about Data Store

During the stage of data store, data keep place should be considered. Data keep process includes the specification about DFS(Distributed File System) which is needed for congregated large volume data mainly used in AI learning system, Grid File System, Hadoop DFS, AWS(Amazon Web Service), in-memory DFS, and retain of superior data store place. Calculation of data store volume is as blow.

Data volume = Data to be store per day * keeping days * compression rate

3.4. Consideration factors about Data Analysis

During the stage of data analysis, analysis method. Analysis method includes the specification about analysis for KPI(Key Performance Index) which is statistics analysis environment and AI analysis environment, AI analysis service which is real time or batch analysis, and AI analysis for decision making.

3.5. Consideration Factors about Data Visualization

During the stage of data visualization, knowledge presentation and visualization should be considered. Knowledge presentation includes the specification about knowledge presentation based on rule, meaning network, frame along complex knowledge presentation, and ontology knowledge presentation. Visualization tool can be D3, Highcharts, and Google.

3.6. Consideration Factors about Human Resource Calculation

During the stage of human resource calculation, resource calculation method for service development and analysis should be included. Resource allocation by function point includes definition of service function, function cost and time, function level. Also

resource calculation by each unit should be executed, which involves EA(Enterprise Architect), SA(Solution Architect), TA(Technical Architect), AA(Application Architect), DA(Data base Architect), and Unit developer. EA is in charge of total design based on business and strategy. SA is in charge of specified design based on software characteristics. TA is in charge of infra architecture and environment maintenance. AA is in charge of service development architecture and development standard maintenance. DA is in charge of data and DBMS(Data Base Management System) maintenance and design. Resource allocation by solution unit is needed from the perspective views from solution cost and duration.

3.7. Consideration Factors about Requirements of Big Data System

The most important requirement of big data system architecture should be data driven architecture. In terms of data driven, data analysis operation is quite dependant on big data system environment. Hadoop platform is good for batch analysis, while Spark platform is good for real time analysis. To secure data store with huge volume, Spark system with Hadoop platform is more recommendable[19,20].

Following table shows about the requirements big data system implementation.

Table 1. Big Data System Recommended Requirements

node	Specification	Remarks)
Name Node	. Usage : Big Data system(Hadooop) Name node . Requirements − CPU : 2 CPU * 8 core − Memory : 256 GB − Internal Disk : 1TB HDD * 2 (OS mirroring, RAID controller) − External Disk : 4 TB * 4 (RAID 5) − NIC : 10 GB * 2(Dual N/W)	Master node
Data Node	. Usage : data node . Requirements − CPU : 2 CPU * 8 core − Memory : 128 GB − Internal Disk : 1TB HDD * 2 (OS mirroring, RAID controller) − External Disk : 4 TB HDD * 12 NIC : 10 GB * 2(Dual N/W)	Slave Node

5. Conclusion

The goal of this study is to provide responsible users with data governance in the age of Big Data. This study can help big data system team members with successful roll out by providing important guides as considerable factors. As a summary, important factors such as system implementation which is aligned with business purpose, effective system compared to invested cost, system which has considered expansion, successful project and change management, communication with end user, should be considered. To do the effective big data project implement, the total diagnosis and guide are needed from the beginning stage of big data system project to real implementation stage.

The next step of this paper will be to find out the success factors of implementation in real cases with enough survey data focusing on business result and TCO(Total Cost of Ownership).

Acknowledgement

This paper is a revised and expanded version of a paper entitled "An Empirical Study of Effective Ways for Improving Big Data Project" presented at the workshop in Daejeon university on December 2017."

"Funding for this paper was provided by Namseoul University."

References

[1] Gartner, Gartner's 2014 Hype Cycle for Emerging Technologies, **(2013)**.
[2] Big Data Computing Technology, Hanbit Academy, **(2016)**, pp. 14-31.
[3] D. Gollmann, "Computer Security", John Wiley & Sons. Ltd, **(2006)**, pp. 27-31.
[4] W. Stallings and L. Brown, "Computer Security", PEARSON, **(2012)**, pp. 377-382.
[5] H. Lim, S. Hye Baek, Security 3.0, IDam, **(2011)**, pp. 40-42.
[6] W. Soo Cho, "Information System Security", HongReng Science Publishing Company, **(2003)**, pp. 264-283.
[7] IBM project management report, 2004. 02.
[8] http://www.netcraft.com.
[9] J. Kim, Korea Information Science Journal, vol. 3, no. 5, **(2010)**, pp. 163-275.
[10] http://www.libelium.com/resources/top_50_iot_sensor_applications_ranking/.
[11] http://www.computerworlduk.com/galleries/cloud-computing/internet-of-things-best-business-enterprise-offerings/.
[12] T. K. Landauer, P. W. Foltz and D. Laham, "Introduction to Latent Semantic Analysis", Discourse Processes, vol. 25, **(1998)**, pp. 259-284.
[13] V. Ranadive, "The Power of Now", McGraw-Hill, USA, **(1999)**.
[14] C. Kuei-Chen, H. Yeu-Shiang and L. Tzai-Zang, "A study of software reliability growth from the perspective of learning effects", Reliability Engineering and System Safety, vol. 93, **(2008)**, pp. 1410-1421.
[15] S. R. K. Venkata and B. Raveendra Babu, "A log based approach for software reliability modeling", Advanced Computer Science Software Eng, vol. 4, no. 2, **(2014)**, pp. 49-51.
[16] K. G. Manton, E. Stallard and J. W. Vaupel, "Alternative Models for the Heterogeneity of Mortality Risks Among the Aged", Journal of the American Statistical Association, vol. 81, no. 395, **(1986)**, pp. 635-644.
[17] J. P. Kotter, "Leading Change", Havard Business School Press, Boston, **(1996)**, pp. 71-76.
[18] P. Drucker, "Managing in a time of Great Change", Truman Talley/Dutton, New York, **(1995)**.
[19] J. E. Hanke and D. W. Wichern, "Business Forecasting", Upper Saddle River, NJ, **(2009)**.
[20] K. Krishnan, "Data Warehousing in the age of BIG DATA", Morgan Kaufmann Publishers, **(2013)**, pp. 219-221.

