Analyzing 6Vs of Big Data using System Dynamics
استخدام ديناميكية النظام في تحليل 6Vs لتضخم البيانات

Alla Talal Yassin – ICCI- Ministry of High Education
Eaesha Hamoudy- Baghdad Univercity
www.allatalal.com ،allatalal@yahoo.com

Abstract

Big Data is not just about lots of data, it is actually a concept providing an opportunity to find new insight into the existing data as well guidelines to capture and analysis future data. This paper deals with the Big data ideas, which is refers to data sets whose size is beyond the ability of commonly used softeware tools to capture,manage , and process the data within a tolerable elapsed time. In this paper we design a System Dynemic (SD) model for the 6Vs analyzing which is the key characteristics of Big data .finally we illustrate some cost examples for the object.

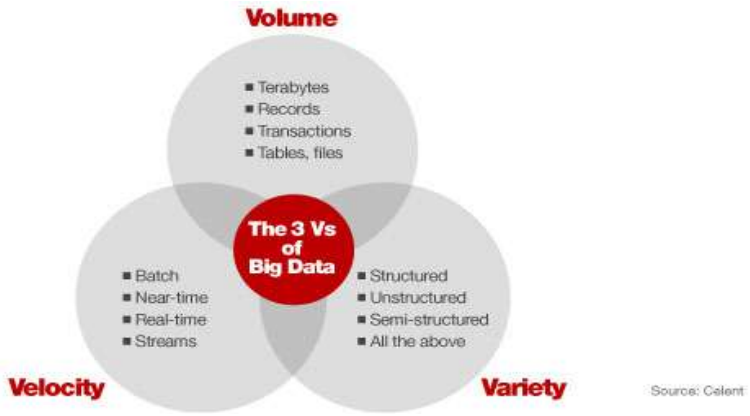
Keywords—*Big Data ,6Vs, System Dynemic, cost*

مستخلص

تضخم البيانات لا يعني الكم الهائل من البيانات ، هو مفهوم لفرص ايجاد مرشد للبيانات الموجودة لتحليلها في المستقبل . والبحث يعنى بمسالة القدرة على استخدام ادوات البرامجيات لادارة عمليات البيانات في وقت الاجراء. ويقدم البحث فكرة تضخم البيانات ويشير الى الحجم والقدرة والعمليات الاجرائية . واستعرض في البحث تصميم انموذج باستخدام ديناميكية النظم لـ 6Vs والتي تعد مفتاح خواص تضخم البيانات ، و اخيرا تم عرض بعض الامثلة حول كلفة موضوع البحث .

1.Introduction

While there are several definitions of big data, This leaves many companies and industries with an open and lingering question about what they should be doing with Big Data. The most common reference focuses on data that reflects figure (1) illustrate the added Volume (terabytes,beyond,records;transactions), additional Variety(internal, external, behavioral, social, structured & unstructured), and increased Velocity (near- or real-time assimilation). However, to rethink data analysis systems in fundamental ways. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business.



Fig(1) :The most common Big Data diffention for 3Vs [1]

3Vs is a term used to define the different attributes of big data - volume, variety and velocity. In 2001, the 3Vs term was coined to define the constructs or attributes that make up an organization's stored and owned data repositories. 3Vs is now used to define the trends and dimensions of big data[1].

2. 4Vs consept

4Vs is a data management trend that was conceived to help organizations realize and cope with the emergence of big data. The 3Vs compare the storage, utilization and consumption of data with regards to the three base dimensions, and it encompasses all data forms, regardless of storage location or format, that are eventually compiled as a big data repository. More recently, additional Vs have been proposed for addition to the model, including variability -- the increase in the range of values typical of a large data set and value, which addresses the need for valuation of enterprise data. Figure (2) illustrate the 4Vs consept for Big Data .

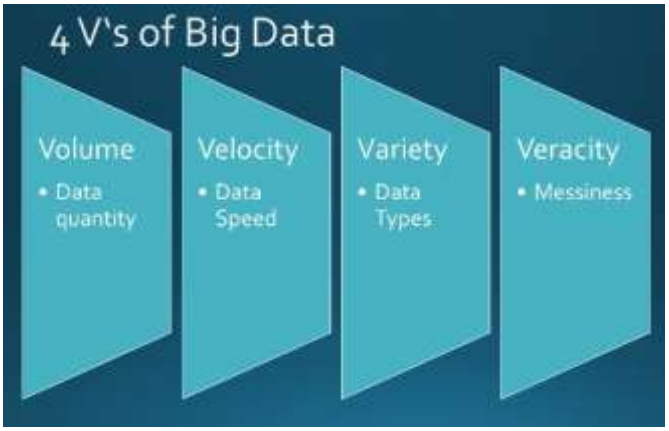


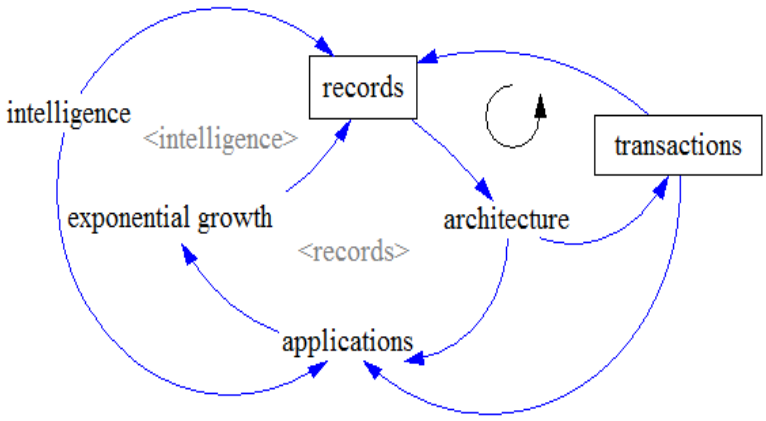
Figure (2) The 4Vs consept for Big Data [2].

Many literature review illustrate that there are additional Vs that IT, business and data scientists need to be concerned with, most notably big data *Veracity*. Other big data V's getting attention at the summit are: validity and volatility. Here is an overview the 6V's of big data. However , we try to analze these Vs in SD models as shows in the next paragraph[2].

3. SD models for 6Vs

3.1 Volume

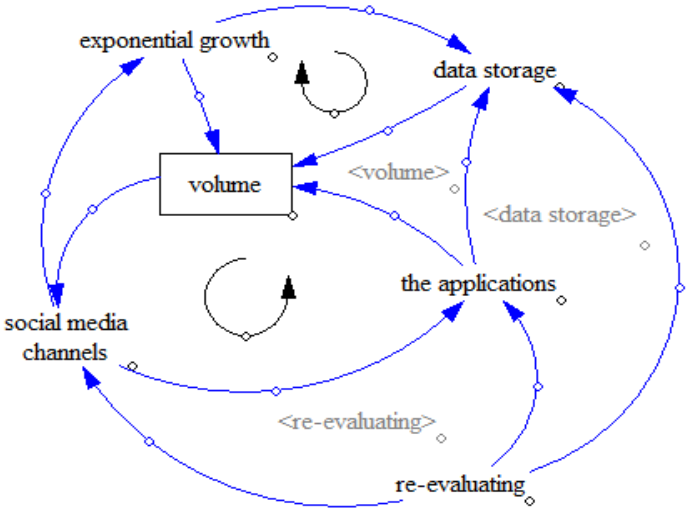
Currently the exponential growth in the data storage as the data is now more than text data. We can find data in the many formats on social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents ***Big Data***. Big data implies enormous volumes of data. .It used to be employees created data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive. Yet, Inderpal states that the volume of data is not as much the problem as other V's like veracity[3]. Figure (3) shows the SD model for Volume Factors for Big Data charecters.



Fig(3) : Volume Factors for Big Data charecters

3.2 Variety

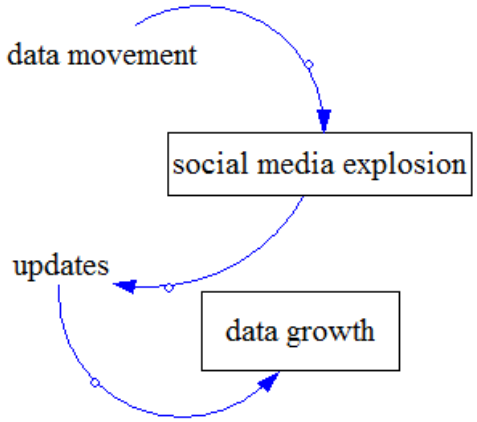
Data can be stored in multiple format. Data could be either in traditional or non- traditional format and it depend on the institution requirmrnt . Figure (4) illustrate the most common factors in variety one of the Big Data charecters .Many many different formats make the challenge to overcome with the *Big Data*. It makes any business more agile and robust so it can adapt and overcome business challenges. Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and ROI if you are able to handle the velocity. Inderpal suggest that sampling data can help deal with issues like volume and velocity[4].



Fig(4) : variety Factors for Big Data charecters

3.3 Velocity

Variety refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data. Figure(5) illustrate velocity Factors for Big Data charecters. The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. However, news channels and radios have changed how fast we receive the news[5].

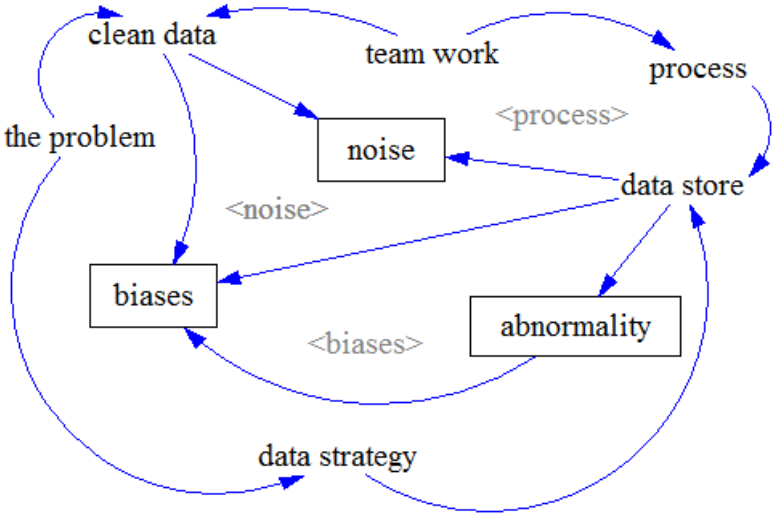


Fig(5) : Velocity Factors for Big Data charecters

Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent ***Big Data***.

3.4 veracity

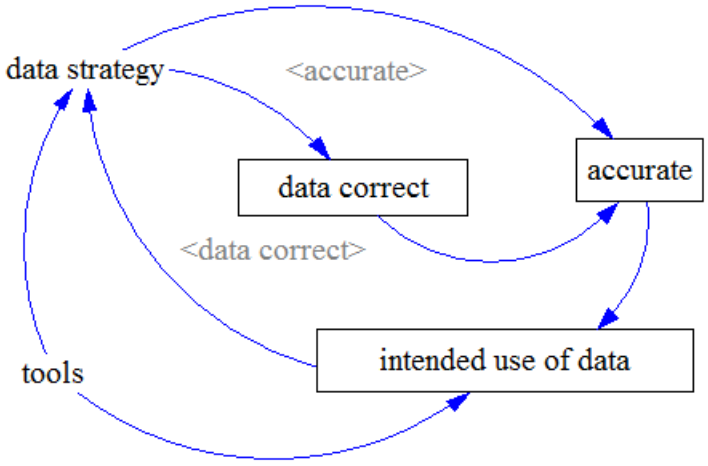
Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Inderpal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity. Figure(6) illustrate veracity factors for Big Data charecters. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems [5].



Fig(6) : veracity Factors for Big Data charecters

3.5 Validity

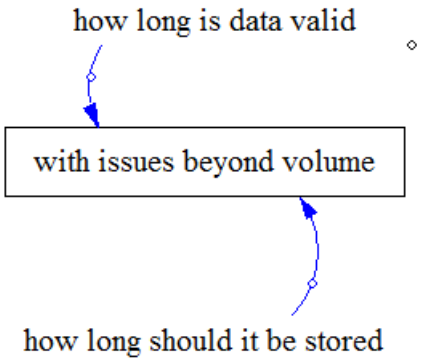
Like big data veracity is the issue of validity meaning is the data correct and accurate for the intended use. Figure (7) illustrate validity factors for Big Data charecters.Clearly valid data is key to making the right decisions[6].



Fig(7) : Validity Factors for Big Data charecters

3.5 Volatility

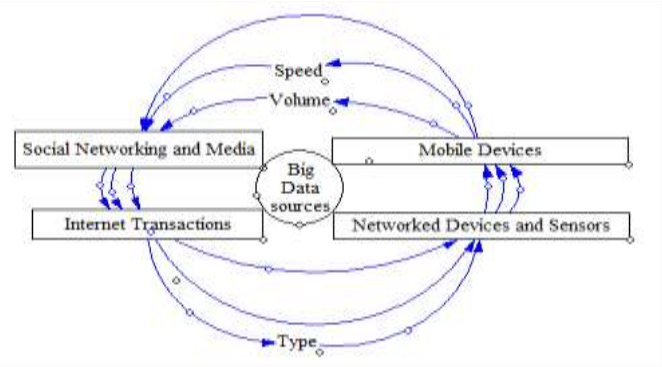
Big data volatility refers to how long is data valid and how long should it be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis. Figure (8) illustrate volatility factors for Big Data charecters.Big data clearly deals with issues beyond volume, variety and velocity to other concerns like veracity, validity and volatility[7].



Fig(8) : Volatility Factors for Big Data charecters

4. Big Data source generation

The size of available data has been growing at an increasing rate. More sources of data are added on continuous basis. Volume is the amount of data created both inside corporations and outside the firewall via the web, mobile devices, IT infrastructure, and other sources is increasing exponentially each year. While the type mean the variety of data types is increasing, namely unstructured text-based data and semi-structured data like social media data, location-based data, and log-file data. Speed – The speed at which new data is being created – and the need for real-time analytics to derive business value from it is increasing thanks to digitization of transactions, mobile computing and the sheer number of internet and mobile device users[8]. Figure (9) illustrate a number of Big Data sources generation.



Fig(9) : A number of Big Data sources generation

5. Cost Problem (examples)

Many researches found that the appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and systems. It has been found that the Cost of processing 1 Petabyte of data with 1000 node could be cost (6,120,000$) as illustrate in example (1).

5.1 Example :

As we know that 1 PB = 1015 B = 1 million gigabytes = 1 thousand terabytes so for 9 hours for each node to process 500GB at rate of 15MB/S and for 15*60*60*9 = 486000MB ~ 500 GB, if we sapuse that 1000 * 9 * 0.34$ = 3060$ for single run, it will be 1 PB = 1000000 / 500 = 2000 * 9 = 18000 h /24 = 750 Day , finally The cost for 1000 cloud node each processing 1PB will be 2000 * 3060$ = 6,120,000$[9].

5.2 Example :

The US 2012 Election could be an example for the cost of big data . in the Facebook page (33 million "likes") , YouTube channel (240,000 subscribers and 246 million page views). Every single night, the team ran 66,000 computer simulations[10].

5.3 Example :

In this Example we analyze a compartion of two cost items for wintercorp report , it has been shows that the cost for data warehouse application was about ($30 million) while by using Hodoop for Big Data management it was just ($9.3 million). Figure(10) illustrate the cost items of the two Big Data methodology.

	Data Warehouse Appliance	Hadoop
Volume of Data	500 TB	500 TB
System Cost[1]	$22.7	$1.4
Initial Acquisition Cost	$5.5[2]	$0.2[3]
Upgrades at 26% CAGR	$8.4	$0.3
Maintenance/Support[4]	$8.2	$0.2
Power/Space/Cooling	$0.6	$0.7
Admin	$0.8	$0.8
Application Development	$6.6	$7.2
Total Cost of Data	**$30 million**	**$9.3 million**

Fig(10): The cost items of the two Big Data methodology[11].

5.4 Example

In india, Annual potential talent pool available for Big data for IT professionals is about (2,800,000) person , PhDs (14,000) person and about (5000) statestics. Figure (11) illustrate the india labour for Big Data Management.

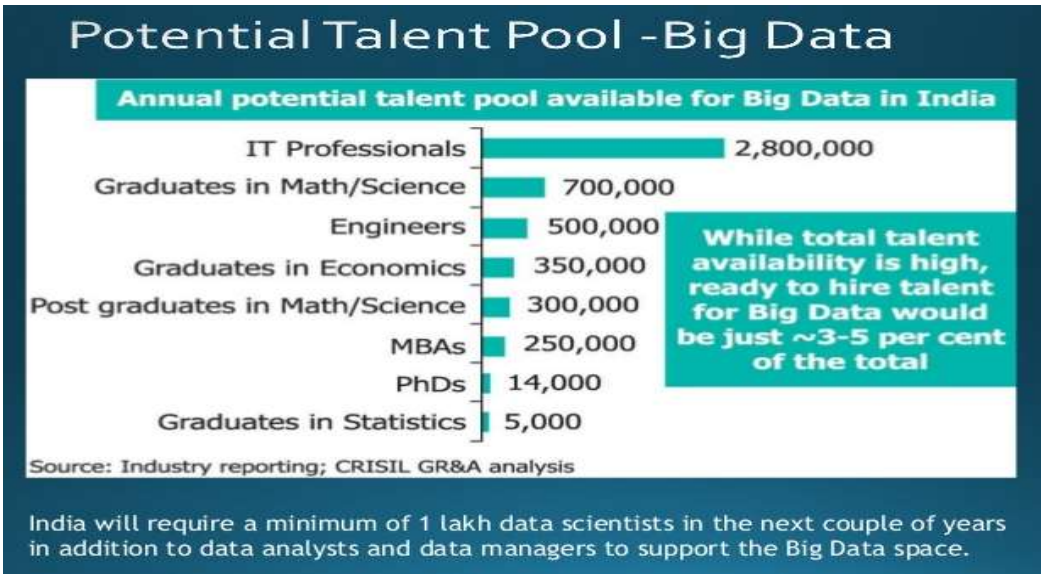


Fig (11) : The india labour for Big Data Management[5].

Conclusion(s)

- SD could be a good tools to analyse the Vs key charectes for Big Data
- We believe that these research problems are not only timely in the literature review, but also have the potential to create huge economic value in the economy for years to come.
- More recently, additional Vs have been proposed for addition to the model , today they are 6Vs from many Big Data source generation.

Refrence

1. http://thefinancialbrand.com/26363/big-data-analytics-retail-banking-jm/
2. Dave Anderson , The Value of Big Data - Defining It Once and For All,2013, http://www.builtinchicago.org/blog/value-big-data-defining-it-once-and-all.
3. http://hbr.org/2012/10/big-data-the-management-revolution
4. www. Colder.com
5. Timos Sellis, CSIT, Big Data and Complex ,Networks Analytics, http://mams.rmit.edu.au/dgmuzdc443n4.pdf
6. http://globalsp.ts.fujitsu.com/dmsp/Publications/public/br-bigdata- revolution.pdf
7. Jinchuan CHEN, Big data challenge: a data management perspective, DOI 10.1007/s11704-013-3903-7, Front. Comput. Sci., 2013, 7(2): 157–164, ttp://iir.ruc.edu.cn/~jchchen/FCSBigData.pdf
8. http://www.freeformdynamics.com/fullarticle.asp?aid=1451
9. http://www.slideshare.net/remyavivek/big-data-ppt-23276173
10. http://www.ibmbigdatahub.com/infographic/four-vs-big-data
11. Winter corporonreport, Cambridge MA, Big Data :what does it relly cost., v1.08, 2013.