Scholars Journal of Engineering and Technology (SJET)
Sch. J. Eng. Tech., 2014; 2(3B):412-414
©Scholars Academic and Scientific Publisher
(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

ISSN 2321-435X (Online)
ISSN 2347-9523 (Print)

Research Article

"BIG DATA" Galaxies (A New Definition for Assembling "BIG DATA")

David Fogarty
University of Phoenix, Phoenix, Arizona

***Corresponding author**
David Fogarty
Email: dfogarty2@gmail.com

Abstract: "BIG DATA" is a key topic in many areas such as government, industry, healthcare, education and data economies are increasingly being hailed as a way to create more jobs and completely new industries. Recently, the world Economic Forum in Davos, Switzerland featured "BIG DATA" as a key topic of interest. As organizations become increasingly dependent on the use of "BIG DATA" to drive their activities this paper explores the process of identifying the compilation of "BIG DATA" which are the unique data assets an organization possesses and no other organizations can duplicate. In this paper we will explore the compilation or bringing together data from various sources within an organization and through its unique network and introduce the term "BIG DATA" Galaxy for the first time.

Keywords: "BIG DATA", Data, Galaxies, Data Marts, Data Bases, Relational, Multi-Structured, Compilation, Cloud, Computing.

INTRODUCTION

"BIG DATA" is defined by Thomas Davenport and the International Institute of Analytics as "data that is either too voluminous or too unstructured to be analyzed through traditional means" [1]. Arthur describes "BIG DATA" as "composed of digital information, including unstructured and multi-structured data, often derived from interactions between people and machines, such as web applications, social networks, genomics and sensors. [2]" The global analyst firm Gartner defines "BIG DATA" as in adjectives describing the data itself. This description includes Volume, Velocity, Veracity and Variety. The major question becomes how to think about what to name a conglomeration or a set of all of this huge data. Normally one would use the word data set, data mart or data warehouse to describe a compilation of data but it was the opinion of this researcher that the term "BIG DATA" requires a different term which not only can represent the volume, velocity and variety of the data itself but can also be useful in describing how the data is to be organized.

In this research we will explore a new definition for a set of "BIG DATA" which truly reflects the volume, velocity and variety characteristic of "BIG DATA" and also how this data needs to be organized.

REVIEW OF THE LITERATURE

The definition of "BIG DATA" has many variations but similar themes across the expert community [3-5]. "BIG DATA is often defined as more data made possible by the internet and every time somebody uses a wireless device. "BIG DATA" was also the topic of interest at recent World Economic Forum in Davos, Switzerland which points to how important this subject is to our world leaders which included representatives from both business and governments. "BIG DATA" was also featured as a strategic direction for firms in recent editions of the Harvard Business Review [6] and the McKinsey Quarterly[7], The Economist [8-9] describes how new discoveries and insights can be obtained from an "overwhelming amount of web-based, mobile, and sensor-generated data arriving at a terabyte and even exabyte scale. Finally, most recently Gobble [5] discussed how "BIG DATA" represents the next evolution in innovation.

ANALYSIS

Given the above introduction to "BIG DATA" in the last section the author proposes a need for a new term to describe the compilation of huge amounts of multi-structured data. Databases have been around since the dawn of digital computing. Where data needed to be stored for subsequent retrieval a database served its purpose. Flat files were the original popular storage units for a database. A flat file means that the data was stored into one long file of text delimited by a tab or other characters. One example of flat files was data stored in magnetic cylinders or tape and accessed through IBM 3090 databases. This was characteristic of data stored in the 1970's and 1980s. In order to search a flat file to obtain information on a field for example length of residence the computer was forced to sequentially search throughout the entire file in order to retrieve the information.

Next on the scene in data storage was the relational database. Relational databases are developed by creating tables to store information in rows (representing records) and columns (representing fields). The tables were linked by one or more common columns which allowed users to easily find information by allowing the various tables to key on when making commands. Commands in a relational database come in the form of a standardized structured query language or (SQL). SQL is the foundation of all popular commercial relational databases. One of the more versatile advantages of the relational database was that it allows the user to perform a sort on a given field and generate a report that contains specific fields from each record. This process results in a savings in processing time and therefore faster queries for the user.

.

DISCUSSION

In all of the various databases in the previous section the various database types were created to either solve a problem of data storage or the ability to rapidly retrieve large amounts of data. The term database is used in all of these cases as a way to indicate that it is a centralized repository of data. Hence the term base which comes from the Latin word basis which stands for base or pedestal. In the context of "BIG DATA" we don't want a base but more of a universe of data . Moreover, given the promise and potential of "BIG DATA" to solve some of the world's most vexing problems and the fact that we desire to retain this information forever even if has no current utility the new term "BIG DATA" Galaxy was coined in response to having to name a set of data within the "BIG DATA" environment.

The term galaxy is proposed which is defined in by the Random House Dictionary as 1; "a large system of stars held together by mutual gravitation and isolated from similar systems by vast regions of space" and 2; "any large and brilliant or impressive assemblage of persons or things." While the second definition could include "BIG DATA" if we define data as an assemblage of things we have a problem since a "thing" is defined as a "material object. As data is not a material object and instead is described by Merriam Webster as "information in numerical form that can be digitally transmitted or processed" we therefore cannot use the second definition. In addition, it would be useful to also incorporate components of the first definition of the galaxy related to astrophysics which includes the gravitational pull that would bring a group of stars together. One can make the case for combining "BIG DATA" with this definition since "BIG DATA tends to be unique to an organization's data collection processes, associations and partnerships. On the commercial side "BIG DATA" is relevant to a firm's commercial footprint which is similar to a gravitational pull that would tend to bring a group of stars together. A commercial footprint could be the firm in question and a single customer view along with all the data from its clients and even external data which can be purchased like for example from Dunn and Bradstreet, Acxiom or Experian. On the government side a public footprint could be represented by a "BIG DATA" Galaxy consistent of data collected on the local, state and federal levels of government.

In summary, we ask the question why do we need to use the term "BIG DATA" Galaxies instead of flat files, relational databases or other types of data storage schema? Firstly, galaxies are beyond large in terms of their relative size which is in agreement with the concept of what we think of when we hear "BIG DATA". Moreover, galaxies are held together by a single force which is gravity. "BIG DATA" should also be constructed with a single force in mind which is to bring data together regardless of its structure or compatibility with the objective of being able to generate useful insights. In "BIG DATA" we are no longer worried or constrained about linking similar records or a single customer view as with relational or flat file databases. Moreover we are no longer concerned with just storing specific information. "BIG DATA" is all about storing as much information as possible with the prospect of someday being able to make sense out of this information.

Storing a "BIG DATA" Galaxy should be on either a private or public cloud depending on the sensitivity of the particular data being assembles. Firms like Google and Amazon offer their massive data infrastructure which are primarily used to run their online operations for a fee so that organizations interested in creating "BIG DATA" galaxies can participate without making huge fixed investment costs. Many healthcare companies have already set up private clouds for their electronic medical records efforts and these can be also utilized for "BIG DATA" Galaxies. Recent abuses of power by the NSA and the IRS which has been uncovered and extensively discussed by the press means that the government should be soon having excess capacity in computing power and storage. The term cloud is often associated with water vapor clouds as the popular media and the advertising tries to depict in imagery. However, given the new term "BIG DATA" galaxy it is proposed that the term data cloud can now be associated with a nebula cloud which represents multiple galaxies. This would be a fitting term for a system which stored multiple "BIG DATA" galaxies.

CONCLUSIONS

Given the potential of "BIG DATA" to solve some of the world's most vexing problems and the lack of the theoretical framework in the literature for defining "BIG DATA" this new framework for assembling and storing data in "BIG DATA" galaxies will be potentially useful for researchers and practitioners alike to think about and organize their "BIG DATA".

Pursuing the creation of "BIG DATA" galaxies will also stimulate research in data fusion and not allow researchers and practitioners to assume that just because there are is no matching information similar to in a relational database the data cannot be linked together. In the cosmos gravity is the common linking factor and I propose in "BIG DATA" the ability to generate current and future insights from the data be the common linking factor. This should even hold true if one is not aware of the current value of the data.

Applying this definition to "BIG DATA" now paves the path for further research on how to create "BIG DATA" Galaxies. Currently this is accomplished through converting unstructured data to structured data using domain knowledge and then bringing these sets together into a common structured dataset. It is proposed by this researcher that this is not the optimal path since specific domain knowledge cannot always be applied at once and there is a need to have a general "galaxy" of data from which multiple domains can draw upon for research in the present time and the future. This problem represents a ripe opportunity for further research and future commercialization.

REFERENCES

1. Davenport TH. (ed.)' Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data, New Jersey: Pearson Education. 2012.
2. Arthur L; Big Data Marketing, New York, Wiley. 2013.
3. Dumbill E; What is Big Data? An Introduction to the Big Data Landscape. 2012. O'Reilly Strata . Available online at http://strata.oreilly.com/2012/01/whatis- big-data.html
4. Franks B; Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics. New York: Wiley. 2012.
5. Gobble MA; Big Data: The Next Big Thing in Innovation. Research and Technology Management, 2012; 56(1): 64-66.
6. Harvard Business Review; Spotlight on Big Data. Harvard Business Review, 2012; 90(10).
7. Brown B; Chui M, Manyika J; Are you Ready for the Era of "Big Data". McKinsey Quarterly, 2012.
8. The Economist; The Data Deluge. Special Report on Managing Information, Technology Section, February 25 , 2010a. Available online at http://www.economist.com/node/15579717).
9. The Economist; All Too Much. Special Report on Managing Information, Technology Section, February 25, 2010b. Available online at http://www.economist.com/node/15557421.