Big Data -- What Data and Why?

Badie N. Farah
Eastern Michigan University

Data is collected daily and automatically from multitude of sources using all possible formats, stored, and backed up for future use and applications. However, it is less obvious that such an activity will generate a positive economic value when the data is used in mining for strategic advantage. Therefore, the current upward trajectory for collecting and storing data with the expectation of great economic returns might not pan out. The rational way is to analyze all data collection, storage, and analysis proposals using cost and benefit metrics to guarantee that economic benefits for the organization will be the outcome.

INTRODUCTION

Data is collected daily from multitude of sources using all possible formats and stored (and backed up) for future use and applications (Laurilla, et al, 2012; Ferris, et al, 2014). The intent of storing data is multifaceted. One reason for collecting and storing data is historical preservation; where the data might be interrogated for reflection on what had taken place sometime ago. Processing of data for such usage might be limited to organizing, cataloguing, and reproducing. Another reason for collecting and storing data is to produce summary reports to comply with the requirements of the organization. Yet still another reason for collecting and storing of data is for the purpose of mining such data for possible discoveries that might be of value to the organization (Oracle, 2013; Taylor, 2012). These reasons have been around for a very long time. What fueled the magnitude of collecting and storing of data today is the tremendous processing power of computers, the availability of such powerful computers to almost all organizations, the availability of tremendous data storage devices with great capacities at a very reasonable price, and above all the availability of advanced statistical methods that make processing the data and extracting information possible (Bollier, et al, 2010). Businesses, certainly large ones, are to a large extent, convinced that collecting and storing data for future use can solve problems and generate economic payback (Lohr, 2012; Boyd, et al, 2011). In certain situations, such as generating summary reports or keeping data because of legal requirements, the value of such data is obvious. However, it is less obvious that such an activity will generate an economic value when the data is used for mining (Wu, et al, 2014). Such a return may, or may not, be of a positive economic value. Therefore, the current upward trajectory for collecting and storing data with the expectation of great economic returns might not pan out. It is the herd mentality what driving this activity. The rational way is to analyze all data collection, storage, and analysis proposals using cost and benefit metrics to guarantee that economic benefits will be the outcome of such an undertaking (Podesta, et al, 2014).

The love of generating, storing, and analyzing data at organizations is very obvious. Most of large organizations, if not all, have certain entities in charge of maintaining data operations. That requires employees with specialties of data storage, security, and manipulation (Chen, et al, 2012). Somehow the herd mentality dictates the desirability of having such an activity simply because every other organization (or the competition) is also on the same path. As such, organizations produce data about all activities. In addition, many organizations subscribe to external data sources related to their activities, businesses, and their environments. But no one necessarily points to the evidence of the value of such an activity; nor, that the collected data lead to better decisions (Trnka, 2014).

In some instances, it is the desire of the executive of the company to run the business in his own way. To do that sometimes the executive might require certain data to be collected and summarized in a certain way. Such an activity could require the work of hundreds of people over extended periods of times to produce such summaries and comparative analysis. This also could be repeated over and over again on monthly, quarterly, or yearly basis. This requirement may not be beneficial to the organization wellbeing, it is simply how the executive performs his job. Another executive might do away with all these summaries thusly eliminating the wasted resources for data collection and analysis. That is, in the perspective of the new executive, the business will be fine without all these reports. In some sense we can argue that data collection, storage, and analysis is based (not entirely) on the executive in charge attitude, which in turn create a culture within the organization that is reflective of his attitude. There is no doubt that data is essential for the success of organizations. None the less, the important question remains the same; what data and why?

Organizations in general and managers in particular should ask themselves the following questions before they commit to any particular program of data collection, storage, and analysis to improve their return on investment.

1. Does such an activity lead to better decisions?
2. At what cost such an activity should be undertaken?
3. Does this activity create confusion?
4. Is certain data unnecessary for the effective operation of the organization?

It might be beneficial for organizations to ask such questions before they embark on an extensive gunshot approach to data collection simply because they can. We advocate the position that data collection, storage, and analysis should be based on feasibility analysis rather than mere possibilities. The feasibility study should consist of:

a) Technical feasibility of data collection, storage, and analysis. This addresses the availability of devices to collect the data (the source), to store the data (enough storage, for how long to store such a data, accessibility of data, data format), and to process the data (software to process the data and possibly mine the data).
b) Economic feasibility of data collection of data collection, storage, and analysis. This addresses the economic value of the data to the organization. A detailed analysis is advocated in this paper and is elaborated upon in future sections.
c) Personnel feasibility of data collection, storage, and analysis. This addresses the viability and expertise of the current personnel of the organization with respect to the proposed data. If such expertise does not currently exist, can the organization attract such talent?

In the following sections of this paper we will discuss the above questions in detail and develop some possible answers.

Does Such an Activity Lead to Better Decisions?

Organizations seem to collect most, if not all, the data that is available in addition to acquiring external data. This data comes in a variety of format, quantity, and complexity (Agrawal, et al, 2011). Answering this question will be the first step in improving the data acquisition process. In other words, the organization first should decide on the questions that they hope the data will answer, and then tailor the data collection toward answering these questions. A question might take the form of hypothesis and therefore the answer might be the outcome of hypothesis testing.

These questions may be related to the various functional areas of the organization. In other words, some of the questions may be generated by the marketing activities of the organization. Other questions may be related to the accounting function of the organization. Also some questions may be generated by the financial activities of the organizations. Other questions may rise as a result of the production activities of the organization. Yet others could be the result of current activities of the information technology of the organization.

A Strategic Planning Process for Data Acquisition

One possibility to ascertain that data collection, storage, and analysis could lead to better decision is to process the questions, generated by the various functions of the organization, by a task force (or standing committee) with members who have different specialties and interest from across the organization. This process may be structured like a strategic planning for data collection, storage, and analysis. It filters up and consolidates from the department level of the organization to the executive level of the organization. Such a strategic process for data acquisition takes place yearly during the planning period which could extend for several months.

This process may coincide with the strategic planning process for the organization. Each department (or interest group) generates statements of needs for certain data to be collected, stored, and analyzed. These statements should contain rationale for such requests and should be augmented with value analysis for such endeavor. Value analysis is discussed in the next section of this paper. The rationale for the acquisition of data should explain why such an activity will lead to better decisions for that particular department or interest group. Such explanation should examine the current decision processes, their deficiencies, and how the proposed data acquisition will mitigate these deficiencies and lead to better decisions. In addition, the proposed data acquisition should articulate what data to be collected, the source of such data, the storage form and length of storage of data, how often such data will be collected, how to dispose of data, who has access to such data, and what kind of access is granted.

During the data acquisition strategic planning period the department accumulate all requests in a fashion that is defined by the organization data acquisition task force following its required content and format. This document is then forwarded to the next level prescribed by the data acquisition task force process. During the same planning period the possible elimination of currently existing data acquisition projects are discussed, and when necessary requests to terminate such projects are included in the departmental data acquisition plan for the current planning year. This allows the wider organization to comment, concur, or object to terminating existing data acquisition projects. Since it might be the case that one department is not interested in an existing data program no longer, while another department still see value in such a program for making better decisions.

As the departmental plans move up the organization structure they are compared to other plans and differences and similarities are noted. After all plans are received, the task force will then have the final authority to approve, consolidate, or disapprove any of these plans. This process, when executed properly, should eliminate redundancy or duplication in data acquisition plans. By following this process, the organization will continue to have the widest support for its data acquisition program.

At What Cost Such an Activity Should be Undertaken?

Given that it has been determined that such a data acquisition project is desirable because it leads to better decisions; it is important to figure out at what cost. This cost includes entities such as data collection, data storage and maintenance; and data analysis. There are two major components to each one of these costs. The technical component – cost of hardware and software; and the human costs which pays for some data generation and data analysis. This last component could be the most expensive and that is why it pays to be very cautious in deciding whether to undertake such an activity.

A Total Cost of Ownership (TCO) calculation may be appropriate for computing the cost the organization will incur to adopt a particular data acquisition project. A TCO formulation may be advanced, or required, by the data task force of the organization. This approach will provide consistency in calculating data acquisition cost across all similar projects, and thus lead to meaningful comparative analysis, as to what data acquisition project to undertake, when necessary. A sample formula for the TCO is:

$$\textbf{TCO = TDC + TDS + TDA} \tag{1}$$

Where: TDC is the total cost of Data Collection, TDS is the total cost of Data Storage, and TDA is the total cost of Data Analysis. TDC, TDS, and TDA may be further decomposed into their cost components.

$$\textbf{TDC} = \sum \textbf{tdc (i)} \qquad \textbf{for all i,} \tag{2}$$

Where: tdc (i) represents the total cost of collecting say data set i.

$$\textbf{TDS} = \sum \textbf{tds (i)} \qquad \textbf{for all i,} \tag{3}$$

Where: tds (i) represents the total cost of storing say data set i.

$$\textbf{TDA} = \sum \textbf{tda (i)} \qquad \textbf{for all i,} \tag{4}$$

Where: tda (i) represents the total cost of data analysis on say data set i.

The total costs in formulas (2), (3), and (4) may further be given in their cost components. These components include the cost of human operator, the cost of the necessary software, the cost of the necessary hardware, and other cost that are related to infrastructure and the like.

The above cost must be balanced with comparable benefits for the organization. The benefits minus the cost will then determine the value of the data acquisition for the organization. The benefits may assume one of two forms. Tangible and intangible benefits. Tangible benefits may be calculated in terms of a dollar value; while intangible benefits may be analyzed separately or assigned a dollar value to complete the cost/benefit analysis. Sometimes the intangible benefits that accrue to the organization is of such a great magnitude that it is the major deciding factor for undertaking a data acquisition project. The following is a formula for calculating the benefits of a data acquisition project.

$$\textbf{BDA = TBDA + IBDA} \tag{5}$$

Where: BDA is the total benefits; TBDA is the tangible benefits; and IBDA is the intangible benefits of data acquisition project respectively. In the case where the organization does not assign a dollar value for IBDA, the term will be dropped out of the formula and the analysis of IBDA is done separately.

A further decomposition of the TBDA and IBDA may be formulated as follows:

$$\textbf{BDA} = \sum \textbf{tbda (i,j)} + \sum \textbf{ibda (i,j)} \qquad \textbf{for all i and all j,} \tag{6}$$

Where: tbda (i,j), ibda (i,j) represent the tangible and intangible benefits of data set i to project j respectively. This formulation allow for the case where a particular data set benefits more than one project within the organization.

Does This Activity Create Confusion?

Data is usually fragmented and might come in bits and pieces from different sources. Sources such as mobile devices, Internet services like Google and Yahoo, stock markets, weather satellites, and internal accounting and production activities. In addition, the data has different formats from text, to voice, to picture, and video.

It is important that the organization create a story from the collected data so it is appropriate to answering the posed questions rather than creating confusion. To a certain extent, this is the job of the organization's managers since automated data systems (such as enterprise data systems) cannot (on their own) create a coherent representation of the story of the organization. Managers should strive to ascertain that the new data would contribute to generating a clearer picture, of the organization, than the existing one. Absent this assertion, the new data might be at best useless, or at worst introduce confusion among the decision makers and their understanding of the current status of the organization.

An entity within the organization such as a department, a team in charge of a particular project, or a managerial task force that has interest in data acquisition needs to construct a story around the data. In other words, what does such an entity see the data reflecting on the organization? For example, sales data might tell a story about the velocity of an item, the category of customers, or a relationship between the categories of customers and the configuration of the sold item (say a tablet, smartphone, or even a service contract).

In the following section we will discuss a measure of confusion that new data might introduce in the decision making process. Uncertainty will be used as a measure of confusion and a mathematical model to calculate uncertainty will be introduced.

Uncertainty as a Measure of Confusion

Decision making in organization is fraught with uncertainty particularly when the decisions are of strategic nature and span a long period of time. Decision makers use data to alleviate or decrease uncertainty. Therefore, any data introduced in the decision making process must contribute to decreasing uncertainty for the decision maker rather than contribute to increasing such uncertainty (increasing confusion), or at best be neutral.

Let **u (i)** be the uncertainty associated with decision i under the current state of decision making and the data associated with the decision. Let us also assume that introducing new data (with appropriate analysis) will change this uncertainty by a certain percentage **x (i)**. Therefore, the uncertainty of a decision after the new data is incorporated in the decision making process is given by the following equation:

$$\mathbf{U(i) = u(i) + x(i) * u(i) = u(i) * [1 + x(i)] \quad for\ all\ i,} \tag{7}$$

x (i) may assume negative, positive, or zero values. If **x (i) < 0,** then the new data contribute to better decisions by decreasing the uncertainty associated with the decision. In this case the organization benefits from the added data because it enhances the certainty of the decision. However, if **x (i) > 0**, then the new data adds to the uncertainty of the decision. In other words, the new data may add confusion rather than further illuminate the decision. In this case, the organization is better off not spending any resources or efforts collecting and analyzing the new data. In the case that **x (i) = 0**, then augmenting the new data with the existing data does not contribute to better decision, and simply it should be ignored. In summary for **x (i) ≥ 0**, the decision maker should forgo the proposed new data and save the resources of the organization.

Since it is possible (may be even probable) that certain data could be used in multiple decisions, it is therefore appropriate to sum equation **(7)** over all i. The result, given by equation **(8)**, represents the total uncertainty of all decisions after new data is introduced in the decision making process of the organization.

$$\mathbf{U = \sum U\ (i) \qquad for\ all\ i,} \tag{8}$$

The total uncertainty **U** may be examined (compared to a certain threshold) to determine if such a data collection, storage, and analysis is beneficial to the organization. In the case where **U** is determined to be very high (a high level of confusion), then the data is more confusing than helping the decision maker and a different data should be sought.

Measuring the uncertainty (confusion) of data with respect to decision making is an added tool to determine if collecting, storing, and analyzing data sheds better light on the decision making process; or to the contrary it adds to the confusion of such a process. Quantifying the confusion that data collection, storage, and analysis might add to the decision process gives the decision maker the opportunity to decide if it is at all valuable or necessary before undertaking such an activity. It is not always the case that more data is better.

Is Certain Data Unnecessary for the Effective Operation of the Organization?

Collected data is historic by its nature. It is important for the organization to determine if the data mainly looks back, or it is helpful for predicting the future. Some historic data is essential for forecasting the future such as Time Series Analysis. However, not all historic data can provide basis for predicting a direction for the organization. The organization needs to determine what data is essential for the effective operation of the organization, and what data is simply historic and has no predictive value. If the process that generate the data is none recurring, then probably the data will have no value for the future operation of the organization and should be tagged as such so it does cause any confusion within the organization. Data (both qualitative and quantitative) that is predictive of the future is helpful for the effective operations of the organization. How much qualitative vs. quantitative data to be collected is highly dependent on the type of questions the organization need to be answered. For the purpose of the effective operation of the organization, data may be considered as "Backward Looking Data", or "Forward Looking Data".

Backward Looking Data

Most of the data collected by an organization is of historic nature. Production and operation data, sales data, accounting data, financial data, and inventory data are some examples. This historic data provide the organization with snapshots of the past which will help in adjusting the operations within the organization to enhance its performance, correct mistakes, and even enlarge certain departments while discontinuing others. This data has one or more of the following objectives among others.

- Data that are required by law to be maintained for a specific period of time. Such as, tax withholding data, data that is covered by current litigation, and email messages.
- Data that is necessary for the daily operation of the organization. Such as, quality control data, back order data, payroll data, and sales data.

None the less, some of this data may also be used to discover trends with respect to products, sales, and services. In this sense it may be considered forward looking data and used, as described in the following section, in telling a story about the organization and formulating questions about the direction the organization might assume in the future.

Forward Looking Data

Forward looking data provides the organization with a possible view of the future. This data might highlight a particular trend, a new market, a new product, a new pricing policy, or even an opportunity for changing the total direction of the organization to steer a new course. All of these decisions are strategic in nature and tremendously important for the wellbeing, even the survival, of the organization. As such, collecting, storing, and analyzing data is justified, as long as, it is the appropriate data. If it happens that the data is the wrong data, the result will be confusion at worst, or missing an opportunity at best. So in this case, the return on investment to the organization is none existent to disastrous. That is why an organization must collect, store, and analyze the correct data at all times. From all available data, the organization must select what is plausible for assisting in making the correct decisions. Making the correct decisions depends on asking the correct questions, collecting the necessary data, and using the appropriate tools for analyzing the data to help answer these questions. Asking the correct questions should be based on the ability of the organization's management to tell a story about the current state of the organization and where they see the future for the organization. .Most of the times, the organization might find that the future is merely an extension of the present, rather than a total reinvention of the organization. For this reason, an organization needs to tell a story using its current internal data and augment it with external data and pose questions which are then translated into hypothesis. These hypothesis will determine if the current data is sufficient for testing the hypothesis, or more data need to be collected. Results from testing the hypothesis will shed light on the questions asked, or may provide answers.

SUMMARY

Organizations constantly collect data from multiple sources using a variety of available devices, venues, and formats. The data comes from mobile devices, Internet services such as Google and Yahoo, stock markets, weather satellites, and from the internal operations of the organization. The data is then stored (and backed up) for future use, analysis, and applications. What fueled the magnitude of collecting and storing of data today is the tremendous processing power of computers, the availability of such powerful computers to almost all organizations, the availability of tremendous data storage devices with great capacities at a very reasonable price, and the sophistication and power of statistical analysis algorithms. The value of such activity is well demonstrated in certain instances where the data is necessary for the operation of the organization, such as, production data, accounting and finance data, and compliance data. However, it is less obvious that such an activity will generate an economic value when the data is used for mining. Such a return may, or may not, be of a positive economic value. Therefore, the current upward trajectory for collecting and storing data with the expectation of great economic returns might not pan out. The rational way is to analyze all data collection, storage, and analysis proposals using cost and benefit metrics to guarantee that economic benefits will be the outcome of such an undertaking. In this paper we provided formulas for measuring such economic value. When mathematical formulas were not appropriate we articulated how an organization can demonstrate the value of such an activity.

REFERENCES

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M. & Widom, J. (2011). Challenges and Opportunities with Big Data 2011-1.

Bollier, D., & Firestone, C. M. (2010). The promise and peril of big data (p. 56). Washington, DC, USA: Aspen Institute, Communications and Society Program. Chicago.

Boyd, D., Crawford, K. (2011). Six provocations for big data. Paper to be presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 36(4), 1165-1188.

Ferris, A., Moore, D., Pohle, N., & Srivastava, P. (December 2013/January 2014). Big Data What Is It, How Is It Collected and How Might Life Insurers Use It? The Actuary Magazine, 28-32, 10(6).

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T. M. T., Dousse, O., & Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing* (No. EPFL-CONF-192489).

Lohr, S. (2012). The age of big data. New York Times, February 11, 2012.

Oracle. Ideas Economy: Finding Value in Big Data; A Summary of an Economist discussion on Big Data sponsored by Oracle. (2013, June 4).

Podesta, J., Pritzker, P., Moniz, E., Holdren, J., & Zients, J. (2014). Big data: seizing opportunities, preserving values. Executive Office of the President, The White House Washington, Study.

Taylor, C. (2012, May 29). Big Data for the Rest of Us. Harvard Business Review.

Trnka, A. (2014). Big Data Analysis. European Journal of Science and Theology, 10(1), 143-148.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97-107.