# Data Analytics: The Next Big Thing in Information

June Crowe and Joseph R. Candlish *(United States)*

## Abstract

Information is now available in an overabundance, so much so, that distinguishing the noise from the signal has become very problematic. In the past, the collection and storage of information was the primary issue. Currently, there are massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In response to this need, data analytical tools and services have emerged as a means to solve this problem.

Grey literature repositories, libraries, and information centers are well positioned to take advantage of these new tools and services. The current trend is to make grey literature more easily discoverable, accessible, and with the new data analytical tools and services, more easily analyzed.

The intent of our survey of the Grey Literature community was to provide a snapshot of the Community's use, planned use, and knowledge of data analytical tools/services for big data as it affects grey literature. The survey summary that follows indicates where the Community currently stands in regards to the use of data analytical tools and services. The poster slides presented at the Grey Literature conference in Rome, Italy are in Appendix A.

## Represented Industries

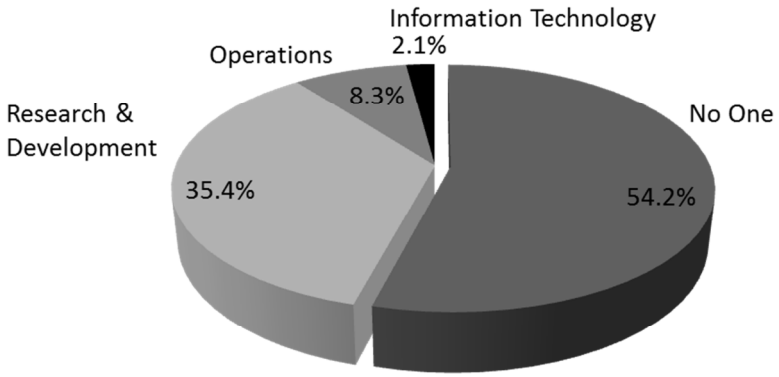
From September 13 through October 31, 2012, an online survey was conducted and made available through two internet vehicles from: (1) the GreyNet Group on LinkedIn®, and (2) the GreyNet listserv. Forty eight responses scattered across South America (1), New Zealand (1), Africa (2), Asia (4), Australia (8), Europe (16) and North America (16), yielded insight into the Grey Literature Communitys' knowledge of the Big Data construct. Overall, academia represented 50% of responses, with

government and private industry composing the remaining half. Within these industries, nearly 42% of the respondents were at the staff level, indicating that there's great understanding of the Big Data landscape, especially in academia.

## The Current Landscape

The current landscape of Big Data products and services revealed several key significant points. First and foremost, a large majority (73%) of respondents indicated that their organization does not currently use Big Data products and services. This situation is partly due to the lack of drivers/champions to adopt them (>54%). However, in the area of Research and Development, there proved to be a significant contributing driver (35%) among the survey population that did denote the current existence and utilization of Big Data discovery and analytical tools (Figure 1).

**Figure 1: Survey responses of current drivers for the adoption of Big Data services/products. (N=48)**




Since such a large percentage of the Community has not adopted Big Data capabilities, it was not surprising to see that the majority (74%) indicated a novice expertise level. Moreover, 80% of the respondents had not seen any data analytical products/services demonstrated but were planning to use such products for web analytics, predictive analytics, and real-time analytics. For those who were familiar with existing analytical tools, SAP, SAS, and Google BigQuery were among the most popular. The

survey question  concerning the near future impact of big data analytical platforms, databases, services, and data analytical tools on grey literature (some impact-27%, moderate impact-19%, high impact-33%) clearly shows that the Grey Literature Community is expecting these services/products to provide some solution to the problem of Big Data.

## Importance of Big Data

Big data is important primarily because it is growing at an exponential rate. Over five exabytes is created every two days. The problem with Big Data is not just data analysis, but with discovering, harvesting, curating, storing and its management. Steve Pederson, CEO of BrightPlanet Corporation, stated that 90% of Big Data content lies in the expanding universe of unstructured content; the vast majority of that information is hidden and unknown in the Deep Web segment of the Internet (Pederson, 2012). Not surprisingly, much of grey literature is found in the Deep Web.

The Grey Literature community strongly felt that Big Data will be a huge positive for society just as it will be for science (56%). George Strawn of the Networking and Information Technology Research and Development (NITRD) Program identified four trends in Big Data in science and business:

(1) bigger data;

(2) increase in unstructured data;

(3) increase in distributed data; and

(4) increase in distributed computing.

These trends will spawn new tools and services for data sharing and collaboration, for data analytics, and for the management of data (Strawn, 2012).

Mobile devices are quickly becoming a primary means of accessing data. However, the survey results indicated that less than half of the respondents responded that it was only somewhat important (36%),

on a 1-5 scale, as a method of accessing Big Data results. Figure 2 outlines levels of importance on a Likert scale of 1-5.

**Figure 2: Likert scale based on levels of importance.**

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Not Important at All | Indifferent | Somewhat Important | Moderately Important | Very Important |
| 19.4% | 19.4% | 36.1% | 19.4% | 5.7% |

This result indicated that many in the Community do not see the mobile phone as a key tool for accessing data. This was another surprise because Juniper Research is predicting a greater demand in 2013 for data analytics solutions across mobile devices (Juniper Research, 2013).

## Barriers and Potential Concerns to Adopting Big Data Products/Services

The two primary barriers to adopting big data services/products are the lack of skilled personnel and/or the lack of sufficient resources. In the Grey Literature community, the lack of sufficient resources proved to be the greatest barrier at 45.7%, with the lack of skilled personnel being a secondary barrier at 33.3%. Currently there is a shortage of personnel skilled in data curation, data integration and re-use, as well as data analysts (Schindel, 2012).

The barriers could conceivably be overcome if there were drivers/champions to led the cause of adopting these technologies/services, and to clearly identify the return on investments to management. Dennis Gannon, Director of Cloud Research Strategy pointed out that every area of science is now engaged in data-intensive research. Thus, the need for these technologies, products and services are not going away but will continue to increase (Gannon, 2012). As the volume of digital data continues to grow exponentially, there will be a continued need for skilled personnel and a need for adequate

financial support. Both will be challenges the Grey Literature Community face in order to reap the benefits of big data analytical products and services.

Biased reporting has proved to be a leading issue to consider as well. Over 90% of respondents were concerned that biased reporting will be a cause of concern across multiple sectors (economic, political, health, scientific, social, etc.). From an historical perspective, an abundant amount of evidence supports this concern. For example, outcome biased reporting within the medical community has been a legitimate pubic concern when newly developed pharmaceuticals are trial tested then submitted to the Food and Drug Administration (FDA) for approval via New Drug Application (NDA) (Rising et al., 2008).

## Big Data Goals

Results from the survey ranked the overall goals of utilizing Big Data products and services. The Grey Literature community favored data discovery (47.2%) with data mining analytics (44.4%) and data visualization (38.9%). The majority of responses emphasized the importance of these three categories by denoting a five on a 1-5 scale (Table 1).

**Table 1: Survey responses on a 1-5 scale (1 being least important, 5 being most important) of which Big Data products/services would be most relevant to your organization's data goals. (N=36)**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Data discovery** | 5.6% | 8.3% | 8.3% | 30.6% | **47.2%** |
| **Data mining analytics** | 2.8% | 2.8% | 27.8% | 22.2% | **44.4%** |
| **Data visualization** | 2.8% | 8.3% | 19.4% | 30.6% | **38.9%** |

## Summary of Survey and Future Considerations

Of the 48 survey takers, a trend in the responses revealed that respondents became increasingly impatient/distracted as the survey progressed. In the first third of the survey, responses were nearly 100% participation. Responses decreased by an average of 10 throughout the second third and plummeted by nearly 20 according to the last third of the survey. This indicated that the survey should

have been shorter and the options for answers more limited. The survey administrators will take these findings into consideration for future surveys.

Overall, the Grey Literature respondents are keenly aware of the benefits of using Big Data services and products but have yet to identify people within their organizations as drivers/champions to make it reality. The lack of identified backers who can clearly and consistently make the case to show the immediate benefits and ultimate return on investments in these technologies and services to management has impacted their adoption or hindered their implementation. As indicated from the survey, the Grey Literature community is not using these products in substantial numbers nor have they seen these products/services demonstrated. Yet, the Community sees great value in these products/services for their local economy (>68% of survey takers), and they are planning to use these tools for web analytics, predictive analytics, and real-time analytics. Additionally, if the Community could select big data products/services for common data goals, they would select them first of all for data discovery and then for data mining analytics. Lastly, the lack of adequate financial resources is the greatest barrier to adopting these products/services.

In terms of future considerations, re-distributing the survey in three to five years may yield interesting responses as Big Data initiatives are readily explored. Additionally, as Big Data products and services mature, a better understanding of the developing landscape may reveal insight into trends that cannot yet be foreseen.

## References

Gannon, D. (2012), Science as a service: Data Analytics and Data Mining - The Approaching Tidal Wave. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Gannon_Data_Analytics.pdf)

Juniper Research. 2012. Juniper Research's Top 10 Mobile Trends for 2013. (http://www.juniperresearch.com/shop/download_whitepaper.php?whitepaper=198)

Pederson, S (2012), Exploiting Big Data from the Deep Web: The new frontier for creating intelligence. BrightPlanet, Sioux Falls, South Dakota. White paper available (http://www.brightplanet.com/2012/07/creating-intelligence-from-big-data-whitepaper/)

Schindel, D.E. (2012), Data Curation: Skill-sets and Workforce Needs. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Schindel_Workforce_Needs.pdf)

Strawn, G.O. (2012), Big Data. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Strawn_Big_Data_Overview.pdf)

Rising K., P. Bacchetti, and L. Bero (2008), Reporting Bias in Drug Trials Submitted to the Food and Drug Administration: Review of Publication and Presentation. PLoS Med 5(11): e217. – doi:10.1371/journal.pmed.0050217

Appendix A


June Crowe and Joseph R. Candlish
Information International Associates, Inc., Oak Ridge, TN.



## Abstract

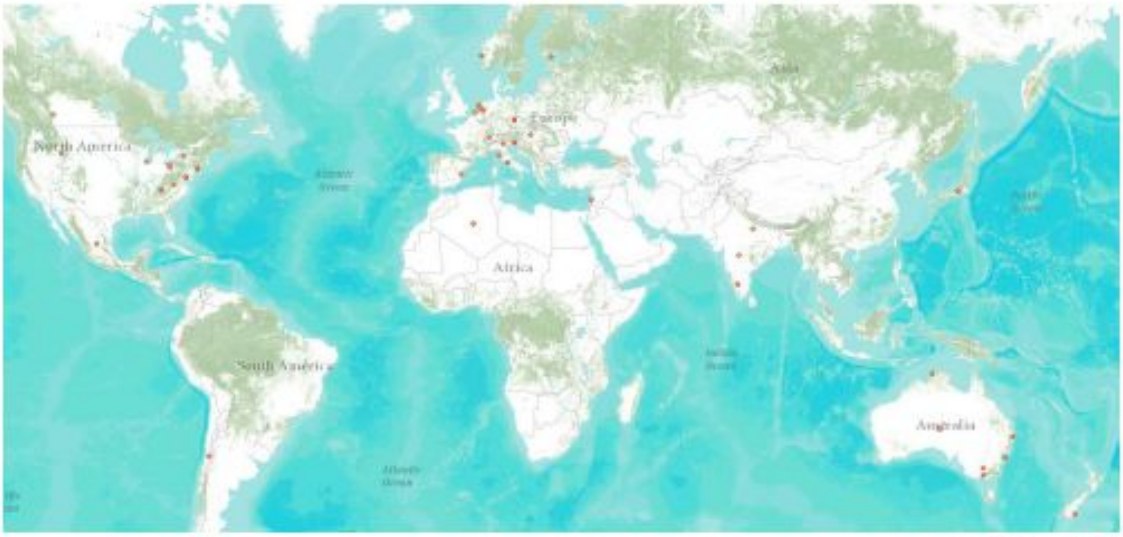
**Data Analytics: The Next Big Thing in Information**

Information is now available in an overabundance, so much so, that distinguishing the noise from the signal has become very problematic. In the past, the collection and storage of information was the primary issue. Currently, we have massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In the meantime, data analytical tools have emerged to solve this problem.

Grey literature repositories, libraries, and information centers are well positioned to take advantage of these new tools. The current trend is to make grey literature more easily discoverable, accessible, and with the new data analytical tools, more easily analyzed.

We created and administered a survey to the Grey Literature Community to get a snapshot of the Community's use, planned use, and knowledge of Big Data as it affects grey literature. The results are reported in these slides.

## Survey Participants from Around the World




Total No.: 48



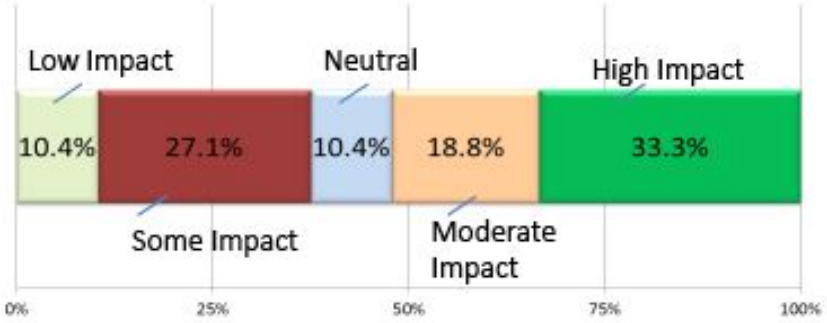
## Represented Industries

| | |
|---|---|
| **Academia** | **50%** |
| **Government agency** | 25% |
| **Private industry** | 25% |

- Within these industries, nearly 42% are staff level

## Potential Impact of Big Data

In the next 2-3 years, do you see Big Data products/services (analytical platforms, databases, services, appliances) impacting grey digital collections?






## Current Landscape

Does your organization currently use Big Data products and services?

- No, 73%
- Yes, 27%

Why such a low percentage?

- No one is driving the adoption of products/services

Here are the drivers

## Current Landscape

Indicate your level of expertise in Big Data: **Novice, 74.4%**

Within the breadth of product/services currently available, the three most recognized were:

1. SAP
2. SAS
3. Google BigQuery

Interesting Finding: 81.4% of survey takers have not seen any Big Data products/services demonstrated.

However: current/planned analytic Big Data projects favored Web analytics (30.4%), Predictive analytics (25.6%), and Real-time analytics (15.4%) respectively.



## Importance of Big Data

According to a survey The Pew Research Center undertook about Big Data, most respondents think that the rise of Big Data is a huge positive for society. Do you think that Big Data will be a huge positive for your organization?

| | |
|---|---|
| **Definitely (value of 5)** | **30.8%** |
| **Moderate (4)** | **25.6%** |

***>56% of responses!***
Only 10% neutral

Regarding Mobile Devices: A strong percentage (36.1%) of survey takers feel it is only "*Somewhat Important*" to access Big Data Results via mobile devices.

## Barriers to Adopting Big Data Products/Services

Please rank the following potential barriers within your organization that may impede the adoption of Big Data products/services. Please rank each on a scale of 1-5 with 1 being the least and 5 the greatest.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Lack of Personnel** | 11.1% | 11.1% | 19.4% | 25.0% | **33.3%** |
| **Lack of Resources** | 8.6% | 11.4% | 14.3% | 20.0% | **45.7%** |



## Goals to Utilize Big Data Products/Services

If it were possible to select any of the Big Data products/services, which of the following common data goals would be most relevant in your situation? Please rank each on a scale of 1-5 with 1 being the least and 5 the greatest.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Data visualization** | 2.8% | 8.3% | 19.4% | 30.6% | **38.9%** |
| **Data discovery** | 5.6% | 8.3% | 8.3% | 30.6% | **47.2%** |
| **Data mining analytics** | 2.8% | 2.8% | 27.8% | 22.2% | **44.4%** |

## Potential Concerns and Local Economy

As Big Data capabilities evolve, do you think that biased reporting will be a cause of concern across multiple facets of applicability (economic, political, social, scientific, health, etc.)?

| Least Concern | Somewhat Concerned | **Moderate Concern** | Moderately High Concern | High Concern |
|---|---|---|---|---|
| 2.9% | 2.9% | **52.9%** | 26.5% | 14.7% |

In your opinion, will Big Data initiatives benefit your local economy?

| Least | Somewhat | **Moderate** | Moderately High | Most |
|---|---|---|---|---|
| 5.7% | 25.7% | **42.9%** | 20.0% | 5.7% |



## Metrics on Survey Participants

Total Number of Survey Takers: **48**

Number of Completed Surveys: **32 (66.7%)**

Trend in Responses

- Of the 24 questions, the first third of the questionnaire had the most participation.
- The second third had an average of 36-39 participants
- The last third had a range of responses from 19-36.

## Summary

Of the 48 survey takers, a trend in the responses revealed that respondents became increasingly impatient/distracted as the survey progressed. The first third of the responses were nearly 100% participation. Responses decreased by an average of 10 throughout the second third and plummeted by nearly 20 according to the last third of the survey. This indicated that the survey should have been shorter and the options for answers more limited. These are points that the survey administrators will heed with the next survey.

Overall, the Grey Literature respondents are keenly aware of the benefits of using big data services and products but yet to identify people within their organizations as drivers to make it reality. For many in the Grey Literature community as well as in other communities, the lack of clarity of the immediate benefits, along with the unknown timeframe for seeing a return on the investment in these products/services has probably delayed or otherwise hindered their implementation. As was indicated from the survey, the Grey Literature community is not using these products in substantial numbers nor have they seen these products/services demonstrated.


## Summary Cont.

Yet, the Community sees great value in these products/services for their local economy (>68% of survey takers), and they are planning to use these tools for web analytics, predictive analytics, and real-time analytics. Additionally, if the Community could select big data products/services for common data goals, they would select them first of all for data discovery and then for data mining analytics. Lastly, the lack of adequate financial resources is the greatest barrier to adopting these products/services.

In terms of future considerations, re-distributing the survey in three to five years may yield interesting responses as Big Data initiatives are readily explored. As Big Data products and services mature, a better understanding of the developing landscape may reveal insight into trends that cannot yet be foreseen.