International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Challenges with Big Data Analytics

Saumya Salian

Abstract: ***Big Data Analytics is high-focus of data science. Huge volumes of business data collected from diverse sources can provide a tremendous scope to understand business trends and customer behavior. Big Data has become significant as many organizations both public and private have been collecting voluminous amount of domain-specific information, which can contain useful information about problems such as national intelligence, marketing, cyber security, fraud detection and medical informatics. The potential of data-driven decision-making is now being acknowledged broadly, and there is expanding exuberance for the notion of ``Big Data.'' Diversity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases that can create value from data. The aim of this research is to investigate specific challenges introduced with Big Data Analytics.***

Keywords: Analytics, Data Modeling, Visualization, Integration

1. Introduction

Big Data has the potential to revolutionize many business driven opportunities. By successfully analyzing information on customer behavior, social media trends, product development, sales and marketing effectiveness, or any of a thousand other data types, new insights can be uncovered to help steer the business. Organizations that implemented a Big Data initiative reported an improvement in their operating profit.

The information collected will not be in a format ready for analysis. Most of data today is not in structured format; for example, tweets are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data increases when it can be linked with other data, thus data integration is a major creator of value. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, visualization of the results and its interpretation is crucial to extracting actionable knowledge.

Big data technologies are evolving to a point in which more organizations are prepared to model and adopt big data as a core component of the information management and analytics infrastructure. Big data is emerging as the next great step in enabling integrated analytics in many common business scenarios. As big data wends its way into the enterprise, information technology (IT) practitioners and business sponsors will have to manage number of challenges that must be addressed before any big data program can be successful.

Following are the common challenges in Big data Analytics
1) Data extraction and cleaning
2) Data integration and Aggregation
3) Meeting the need for Processing Big Data at High Speed with In-memory
4) Query Processing, Data Modeling and Analysis
5) Making Sense of Big Data with Visualization

2. Big Data Challenges

2.1 Data Extraction and Cleaning

Mostly, the information collected will not be in a format ready for analysis. For example, consider information collected from health care system, consisting of handwritten dictations from several physicians, structured data from sensors and image data such as x-rays. It is difficult to analyze such types of unstructured data. Requirement of extraction process that retrieves required information from diverse sources and illustrates it in a structured form suitable for analysis is necessary.

Data cleaning is an integral part of data analysis. In fact, it takes more time to clean the data than to perform statistical analysis on it. Data cleaning is a major challenge in Big Data analytics and the value of data for decision-making purposes will be jeopardized if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the variety and voluminous information involved in big data projects, it becomes even more pronounced. Data visualization will only provide factual and correct analysis if the data quality is assured. To deal this issue, organization need to have data governance or information management process in place to ensure the data is clean.

2.2 Data Integration and Data Aggregation

Given the variety, volume, velocity of data, big data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. One of the most predictable business pressures driving Big Data investment is having too many data sources.

In most of the cases, important business data is spread out over too many locations, from databases to file stores to collaborative web portals to multiple versions of enterprise applications. For effective large-scale analysis, data integration has to happen in a completely automated manner. The value of big data comes from its variety, but so, too, does its complexity.

The proliferation of data sources, types and stores is increasing the challenge of combining data into meaningful, valuable information. While companies are investing in

Licensed Under Creative Commons Attribution CC BY

initiatives to increase the amount of data at their disposal, most are spending more time finding the data they need than putting it to work. The challenges involved with data integration are growing in step.

Data aggregation is any process in which information is expressed in a summary form for purposes such as reporting or analysis. Ineffective data aggregation is currently a major component that limits query performance. And, with up to 90 percent of all reports containing aggregate information, it becomes clear why proactively implementing an aggregation solution can generate significant performance benefits, opening up the opportunity for companies to enhance their organizations' analysis and reporting capabilities.

2.3 Meeting the need for Processing Big Data at High Speed with In-memory

The most common business pressure is an inability to deliver information as quickly as business users need it. In today's competitive business environment, companies have to find and analyze the relevant data quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is to manage volumes of data and acquiring the level of detail needed, all at a high speed. The challenge increases as the degree of granularity increases. One possible method is hardware. Some vendors are using increased memory and powerful parallel processing to crunch sheer volumes of data rapidly.

Alternative method is putting data in-memory and using a grid computing approach, where many machines are used to solve a problem. In this method the target data is loaded directly into the random access memory (RAM) of a server or desktop close to the processer. This eliminates the need to connect to a storage array or disk, locate the desired information, and channel it over a network to the server doing the processing. The potential of the processors can be directly utilized to retrieve and manipulate the desired information. Effective technologies for processing data at high speed is in-memory computing. Both approaches allow organizations to explore huge data volumes and gain business insights in near-real time

2.4 Query Processing, Data Modeling and Analysis

Technique for querying and mining Big Data differs from traditional statistical analysis on small samples .Data is often noisy, complex, dynamic, heterogeneous, inter-related. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis is powerful and reveal more reliable hidden patterns and knowledge. Data Mining requires integrated, cleaned, ethical, and efficiently accessible data, computing query and mining programs, scalable mining algorithms, and big-data computing environments[7]. Using efficient data modeling techniques helps to improve the quality and integrity of the data, understand its semantics, and impart intelligent querying functions.

As noted previously, real-life medical records have errors, are heterogeneous, and frequently are distributed across multiple systems. Challenge here is to scale complex query processing techniques to terabytes while enabling interactive response times. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

2.5 Making Sense of Big Data with Visualization

The decisive step in any analytics project is to take the processed information and make it easily understandable and available for its intended audience. Visualization provides business insight in the form of reports, dashboards, or charts which provides self-service and interactive solutions to the business user.

In Big Data environments this visualization is even more pronounced, for the simple fact that the heterogeneity of the data infrastructure places a premium on IT resources making it ineffective for them to manage BI reporting. Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information. Systems with a rich palette of visualizations become important in conveying to the users the results of the queries in a way that is best understood in the particular domain.

3. Conclusion

All the challenges reflect different facets of a more fundamental issue: the absence of a strategy for integrating big data into the enterprise environment. Efficiently managing today's big data challenges requires a robust data integration strategy backed by leading-edge data technologies and services. This can be achieved through the development of informed processes that take advantage of best-of-breed data integration technologies that address your evolving challenges and eliminate the correlated risks.

References

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs,C. Roxburgh, and A. H. Byers. "Big data: The next frontier for innovation, competition, and productivity." McKinsey Global Institute, May 2011.
[2] Y. Noguchi. "Following Digital Breadcrumbs to Big Data Gold. National Public Radio," Nov 2011.
[3] Y. Noguchi. "The Search for Analysts to Make Sense of Big Data," Nov 2011.
[4] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou , J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom. "Challenges and Opportunities with Big Data," Mar 2012.
[5] S. Lohr. "The age of big data," Feb 2012.
[6] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., And Murthy, R

Licensed Under Creative Commons Attribution CC BY

Hive—a warehousing solution over a Map-Reduce framework. In VLDB, 2009.

[7] H.W. Ian, E.F., "Data mining: Practical machine learning tools and techniques," 2005: Morgan Kaufmann.

Licensed Under Creative Commons Attribution CC BY