Big Data

Hamza Ahmed

Definition

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. Why? More data may lead to more accurate analyses.

Volume Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data are streaming in from social media. Also increasing amounts of sensor and machine-to-machine data are being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

Velocity Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

Variety, Data today comes in all types of formats Structured, numeric data in traditional databases. Information created from line-of-business applications unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data are something many organizations still grapple with.

At SAS, we consider two additional dimensions when thinking about big data:

Variability, In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

- **Complexity.** Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

Introduction

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance. In 2010, enterprises and users stored more than 13 exabytes of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [PCAST2010] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b].

Why big data should matter to you

The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smarter business decision making. For instance, by combining big data and high-powered analytics, it is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use clickstream analysis and data mining to detect fraudulent behavior.

Big Data: Who's in Charge

A hot topic in the big data world is how to make sense of this information. One major hurdle when facing big data challenges is determining who is in charge of data management strategy. In the past, IT groups have done the bulk of data management strategy, but recent trends in data governance and data stewardship have given business analysts and business managers a seat at the table

Big data in action

Perspective: UPS

UPS is no stranger to big data, having begun to capture and track a variety of package movements and transactions as early as the 1980s.The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores more than 16 petabytes of data. Much of its recently acquired big data, however, comes from telematics sensors in more than 46,000 vehicles. The data on UPS trucks, for example, includes their speed, direction, braking and drive train performance. The data in not only used to monitor daily performance, but to drive a major redesign of UPS drivers' route structures. This initiative, called ORION (On-Road Integration Optimization and Navigation), is arguably the world's largest operations research project. It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time.

http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

February 2001.

Data Integration, Aggregation, and Representation

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure. Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution. Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other purposes. Witness, for instance, the tremendous variety in the structure of bioinformatics databases with information regarding substantially similar entities, such as genes. Database design is today an art, and is carefully executed in the enterprise context by highly-paid professionals. We must enable other professionals, such as domain scientists, to create effective database designs, either through devising tools to assist them in the design process or through forgoing the design process completely and developing

techniques so that databases can be used effectively in the absence of intelligent database design.

BIG DATA TRANSFORMS BUSINESS

Businesses that exploit Big Data to improve strategy and execution are distancing themselves from competitors. The Big Data solution from EMC provides market-leading, scale-out storage, a unified analytics platform, and business process and application development tools. Together, these allow organizations to draw deeper insights and become a more predictive organization.

Looking back at the 2013 Big Data trends

Last year I noted that on on-the-go Big Data, meaning being able to view Big Data visualizations on mobile devices, will become important and in 2013 we saw the rise of a bunch of new mobile devices including smart watches and Google Glass. There are some Big Data startups, such as Roambi, who have very clear understanding of on-the-go Big Data and are capable of bringing real-time interactive visualizations to mobile devices. On-the-go Big Data really took off in 2013 and is here to stay.

The second trend was that Big Data does not require big bucks because of the plethora of Big Data open source tools that are available in the market as well as the decreasing costs of storage. The price of storage does indeed continue to decrease in costs, but the amount of data also grows exponentially. Will we be able to keep up with this or will the amount of data outgrow the available storage? The amount of open source tools is growing rapidly, but there is also a rise in licensed Big Data solutions, because open source tools do require experience Big Data personnel and many organisations do not yet have these staff available. So, to start with Big Data it does not have to cost the world, but to develop and implement a complete Big Data solution can be expensive, although the results can also be significant.

The third trend was big real-time data and 2013 did indeed show an important growth in real-time analytics. More and more tools become available that create a layer on top of Hadoop to be able to deal with real-time data and Hadoop 2.0's YARN framework enables real-time data analysis. In the coming years this will become more important as many industries see the advantages of real-time analytics.

Big consumer data, or the quantified-self movement, was the fourth trend and this really took off in 2013. Wearable technologies that can measure everyday life have started to appear massively and more and more consumers are measuring at least something of their behavior, be it their sleeping patterns or the running results. Recent research from Pew Research Centre revealed that 69% of U.S. adults keep track of at least one health indicator such as weight, diet, exercise routine or symptom.

The final trend was Big Data related to privacy. Last year I wrote that "it feels like we have fallen in love with Big Data and that we are blind for the pitfalls in Big Data. We do not want to see the backside of Big Data and the effect it will have on our privacy." Of course, in 2013 we saw the PRISM leak by Edward Snowden, which showed that privacy in the Big Data world is indeed an endangered species and there is probably a lot more to be revealed in the coming months. Privacy is indeed affected by Big Data and as consumers, and companies, we will have to get used to this new reality.

http://smartdatacollective.com/bigdatastartups/174741/seven-big-data-trends-2014

Conclusion

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

Refernces

Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert

Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.

[Gar2011] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press

release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916

[Gon2008] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. Nature 453, 779-782 (5 June 2008)

[LP+2009] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer,Nicholas Christakis, Noshir Contractor,James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Science 6 February 2009: 323 (5915), 721-723.

[McK2011] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.