A SURVEY ON BIG DATA

Amegha.K, Sowmya.B, Apoorva M.P

Abstract: Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. While the term may seem to reference the volume of data, that isn't always the case. The term big data especially when used by vendors --may refer to the technology that an organization requires to handle the large amounts of data and storage facilities. This paper discusses the characteristics, utilities and opportunities and some of the challenges of big data.

Index: MapReduce, NLP, NoSQL, Cloud Computing, Streaming data

1. Introduction

The amount of data in the present world has been exploding tremendously, and these ever-growing data are extremely diverse. A flood of data is generated every day by the interactions of billions of people using mobile devices and Internet. The increasing volume of enterprise information, multimedia, social media, information-sensing mobile devices, genomics and medical records will fuel exponential growth in data in the future. Analyzing large data sets—so-called big data—will become a key basis of competition in business and technology.

Big data refers to the large data sets which are very difficult to store, analyze and manage due to their size as well as complexity. Their size ranges from thousands of terrabytes to peta-, and exa-bytes. More often, however, Big Data is defined situationally rather than by size. As technology advances over time, the size of the data sets that qualify as big data will also increase. Traditional data processing and database management systems have failed to process this big data.

When data scale to “big data”, many challenges arise in order to deal with it. Scaling with traditional database and queue, scaling by sharding –horizontal partitioning-, maintenance, fault-tolerance, replication etc., are some of the challenges in managing extremely large data sets. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world. Access to data is critical – information from multiple data sources, often from third parties need to be integrated. Maintenance of big data is a challenge not only because of its volume but also because of its variety and velocity. Variety refers to the heterogeneity of data types, representation, and semantic interpretation and both the rate at which data arrive and the time in which it must be acted upon together constitutes velocity.

Big data or the huge collection of information across the world can create significant value for the world economy and play an important role in enhancing the productivity and competitiveness of companies and public sector and creating substantial economic surplus for consumers. For decades, IT was all about systems, networks and software and data was treated simply as bi-product of information processing activities. “Big Data” forces a shift in focus from IT assets deployment and administration to the management of high-value data assets.

The past decade’s successful web startups are prime examples of big data used as an enabler of new products and services. For example, by combining a large number of signals from a user’s actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It’s no coincidence that the lion’s share of ideas and tools underpinning big data has emerged from Google, Yahoo, Amazon and Facebook.

Big Data has been of concern to organizations working in select fields for some time, such as the physical sciences (meteorology, physics), life sciences (genomics, biomedical research), government (defense, treasury), finance and banking (transaction processing, trade analytics), communications (call records, network traffic data), and, of course, the Internet (search engine indexation, social networks). Now, however, due to our digital fecundity, Big Data is becoming an issue for organizations of all sizes and types.

In this paper, we present a survey of big data, its characteristics, opportunities, technology and application challenges.

2. Big Data Opportunities

Some of the specific Big Data opportunities they are capitalizing on include:

- Faceted search at scale
- Multimedia search
- Sentiment analysis
- Automatic database enrichment
- New types of exploratory analytics
- Improved operational reporting

We’ll now look more closely at these opportunities, with each accompanied by a brief example of an opportunity realized using a technology whose role is often overlooked or misunderstood in the context of Big Data: the search engine. We’ll then review the full range of tools available to organizations seeking to exploit Big Data, followed by further examples from the search world.

A. Faceted Search at Scale

Faceted search is the process of iteratively refining a search request by selecting (or excluding) clusters or categories of results. In contrast to the conventional method of paging through simple lists of results, faceted search (also referred to as parametric search and faceted navigation) offers a remarkably effective means of searching and navigating large volumes of information—especially when combined with user aids like type-ahead query suggestions, auto-spelling correction and fuzzy matching.

Until recently, faceted search could only be provided against relatively small data sets because the data classification and descriptive meta-tagging upon which faceted search depends were largely manual processes. Now, however, industrial-grade Natural language processing (NLP) technologies are making it possible to automatically classify and categorize even Big Data-size collections of unstructured content, and hence to achieve faceted search at scale. We can see industrial faceting at work in the dual Web/enterprise search engine EXALEAD CloudView, in other public Web search engines like Google, Yahoo! and Bing, and, to varying degrees of automation and scale, in search utilities from organizations like HP, Oracle, Microsoft and Apache. Look for this trend to accelerate and to bring new accessibility to unstructured Big Data.

B. Multimedia Search

Multimedia content is the fastest growing type of user-generated content, with millions of photos, audio files and videos uploaded to the Web and enterprise servers daily. Exploiting this type of content at Big Data scale is impossible if we must rely solely on human tagging or basic associated metadata like file names to access and understand content.

However, recent technologies like automatic speech-to-text transcription and object-recognition processing (called Content-Based Image Retrieval or CBIR) are enabling us to structure this content from the inside out, and paving the way toward new accessibility for large-volume multimedia collections.

C. Sentiment Analysis

Sentiment analysis uses semantic technologies to automatically discover, extract and summarize the emotions and attitudes expressed in unstructured content. Semantic analysis is sometimes applied to behind-the-firewall content like email messages, call recordings and customer/constituent surveys. More commonly, however, it is applied to the Web, the world’s first and foremost Big Data collection and the most comprehensive repository of public sentiment concerning everything from ideas and issues to people, products and companies.

Sentiment analysis on the Web typically entails collecting data from select Web sources (industry sites, the media, blogs, forums, social networks, etc.), cross-referencing this content with target entities represented in internal systems (services, products, people, programs, etc.), and extracting and summarizing the sentiments expressed in this cross-referenced content.

This type of Big Data analysis can be a tremendous aid in domains as diverse as product development and public policy, bringing unprecedented scope, accuracy and timeliness to efforts such as:

- Monitoring and managing public perception of an issue, brand, organization, etc. (called “reputation monitoring”)
- Analyzing reception of a new or revamped service or product
- Anticipating and responding to potential quality, pricing or compliance issues

- Identifying nascent market growth opportunities and trends in customer demand.

D. Database Enrichment

Once you can collect, analyze and organize unstructured Big Data, you can use it to enhance and contextualize existing structured data resources like databases and data warehouses. For instance, you can use information extracted from high-volume sources like email, chat, website logs and social networks to enrich customer profiles in a Customer Relationship Management (CRM) system. Or, you can extend a digital product catalog with Web content (like, product descriptions, photos, specifications, and supplier information). You can even use such content to improve the quality of your organization's master data management, using the Web to verify details or fill in missing attributes.

E. Exploratory Analytics

Exploratory analytics has aptly been defined as "the process of analyzing data to learn about what you don't know to ask." It is a type of analytics that requires an open mind and a healthy sense of curiosity. In practice, the analyst and the data engage in a two-way conversation, with researchers making discoveries and uncovering possibilities as they follow their curiosity from one intriguing fact to another (hence the reason exploratory analytics are also called "iterative analytics"). In short, it is the opposite of conventional analytics, referred to as Online Analytical Processing (OLAP). In classic OLAP, one seeks to retrieve answers to precise, pre-formulated questions from an orderly, well-known universe of data. Classic OLAP is also sometimes referred to as Confirmatory Data Analysis.

3. Characteristics of Big Data

As shown in Figure 1, volume, velocity and variety make up three key characteristics of big data:

Volume: Rather than just capturing business transactions and moving samples and aggregates to another database for analysis, applications now capture all possible data for analysis.

Velocity: Traditional transaction-processing applications might have captured transactions in real time from end users, but newer applications are increasingly capturing data streaming in from other systems or even sensors. Traditional applications also move their data to an enterprise data warehouse through a deliberate and careful process that generally focuses on historical analysis.

Variety: The variety of data is much richer now, because data no longer comes solely from business transactions. It often comes from machines, sensors and unrefined sources, making it much more complex to manage.

A. Volume

The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics. Having more data beats out having better models: simple bits of math can be unreasonably effective given large amounts of data. If you could run that forecast taking into account 300 factors rather than 6, could you predict demand better?

This volume presents the most immediate challenge to conventional IT structures. It calls for scalable storage, and a distributed approach to querying. Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it.

Assuming that the volumes of data are larger than those conventional relational database infrastructures can cope with, processing options break down broadly into a choice between massively parallel processing architectures - data warehouses or databases such as Greenplum-and Apache Hadoop-based solutions. This choice is often informed by the degree to which the one of the other "Vs"-variety comes into play. Typically, data warehousing approaches involve predetermined schemas, suiting a regular and slowly evolving dataset. Apache Hadoop, on the other hand, places no conditions on the structure of the data it can process.

At its core, Hadoop is a platform for distributing computing problems across a number of servers. First developed and released as open source by Yahoo, it implements the MapReduce approach pioneered by Google in compiling its search indexes. Hadoop's MapReduce involves distributing a dataset among multiple servers and operating on the data: the "map" stage. The partial results are then recombined: the "reduce" stage.

To store data, Hadoop utilizes its own distributed file system, HDFS, which makes data available to multiple computing nodes. A typical Hadoop usage pattern involves three stages:

- Loading data into HDFS,
- MapReduce operations, and
- Retrieving results from HDFS.

This process is by nature a batch operation, suited for analytical or non-interactive computing tasks. Because of this, Hadoop is not itself a database or data warehouse solution, but can act as an analytical adjunct to one. One of the most well-known Hadoop users is Facebook, whose model follows this pattern. A MySQL database stores the core data. This is then reflected into Hadoop, where computations occur, such as creating recommendations for you based on your friends' interests. Facebook then transfers the results back into MySQL, for use in pages served to users.

B. Velocity

The importance of data's velocity — the increasing rate at which data flows into an organization — has followed a similar pattern to that of volume. Problems previously restricted to segments of industry are now presenting themselves in a much broader setting. Specialized companies such as financial traders have long turned systems that cope with fast moving data to their advantage. Now it's our turn. Why is that so? The Internet and mobile era means that the way we deliver and consume products and services is increasingly instrumented, generating a data flow back to the provider. Online retailers are able to compile large histories of customers' every click and interaction: not just the final sales. Those who are able to quickly utilize that information, by recommending additional purchases, for instance, gain competitive advantage. The smartphone era increases again the rate of data inflow, as consumers carry with them a streaming source of geolocated imagery and audio data.

It’s not just the velocity of the incoming data that’s the issue: it’s possible to stream fast-moving data into bulk storage for later batch processing, for example. The importance lies in the speed of the feedback loop, taking data from input through to decision. A commercial from IBM makes the point that you wouldn’t cross the road if all you had was a five-minute old snapshot of traffic location. There are times when you simply won’t be able to wait for a report to run or a Hadoop job to complete. Industry terminology for such fast-moving data tends to be either “streaming data,” or “complex event processing.” This latter term was more established in product categories before streaming processing data gained more widespread relevance, and seems likely to diminish in favor of streaming.

There are two main reasons to consider streaming processing. The first is when the input data are too fast to store in their entirety: in order to keep storage requirements practical some level of analysis must occur as the data streams in. At the extreme end of the scale, the Large Hadron Collider at CERN generates so much data that scientists must discard the overwhelming majority of it — hoping hard they’ve not thrown away anything useful. The second reason to consider streaming is where the application mandates immediate response to the data. Thanks to the rise of mobile applications and online gaming this is an increasingly common situation.

Product categories for handling streaming data divide into established proprietary products such as IBM’s InfoSphere Streams, and the less-polished and still emergent open source frameworks originating in the web industry: Twitter’s Storm, and Yahoo S4. As mentioned above, it’s not just about input data. The velocity of a system’s outputs can matter too. The tighter the feedback loop, the greater the competitive advantage.

The results might go directly into a product, such as Facebook’s recommendations, or into dashboards used to drive decision-making. It’s this need for speed, particularly on the web, that has driven the development of key-value stores and columnar databases, optimized for the fast retrieval of precomputed information. These databases form part of an umbrella category known as NoSQL, used when relational models aren’t the right fit.

Microsoft SQL Server is a comprehensive information platform offering enterprise-ready technologies and tools that help businesses derive maximum value from information at the lowest TCO. SQL Server 2012 launches next year, offering a cloud-ready information platform delivering mission-critical confidence, breakthrough insight, and cloud on your terms.

C. Variety

Rarely does data present itself in a form perfectly ordered and ready for processing. A common theme in big data systems is that the source data is diverse, and doesn’t fall into neat relational structures. It could be text from social networks, image data, a raw feed directly from a sensor source. None of these things come ready for integration into an application.

Even on the web, where computer-to-computer communication ought to bring some guarantees, the reality of data is messy. Different browsers send different data, users withhold information, they may be using differing software versions or vendors to communicate with you. And you can bet that if part of the process involves a human, there will be error and inconsistency. A common use of big data processing is to take unstructured data and extract ordered meaning, for consumption either by humans or as a structured input to an application. One such example is entity resolution, the process of determining exactly what a name refers to. Is this city London, England, or London, Texas? By the time your business logic gets to it, you don't want to be guessing.

The process of moving from source data to processed application data involves the loss of information. When you tidy up, you end up throwing stuff away. This underlines a principle of big data: when you can, keep everything. There may well be useful signals in the bits you throw away. If you lose the source data, there's no going back.

Despite the popularity and well understood nature of relational databases, it is not the case that they should always be the destination for data, even when tidied up. Certain data types suit certain classes of database better. For instance, documents encoded as XML are most versatile when stored in a dedicated XML store such as MarkLogic. Social network relations are graphs by nature, and graph databases such as Neo4J make operations on them simpler and more efficient.

Even where there's not a radical data type mismatch, a disadvantage of the relational database is the static nature of its schemas. In an agile, exploratory environment, the results of computations will evolve with the detection and extraction of more signals. Semi-structured NoSQL databases meet this need for flexibility: they provide enough structure to organize data, but do not require the exact schema of the data before storing it.

4. Big data Analysis

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges.

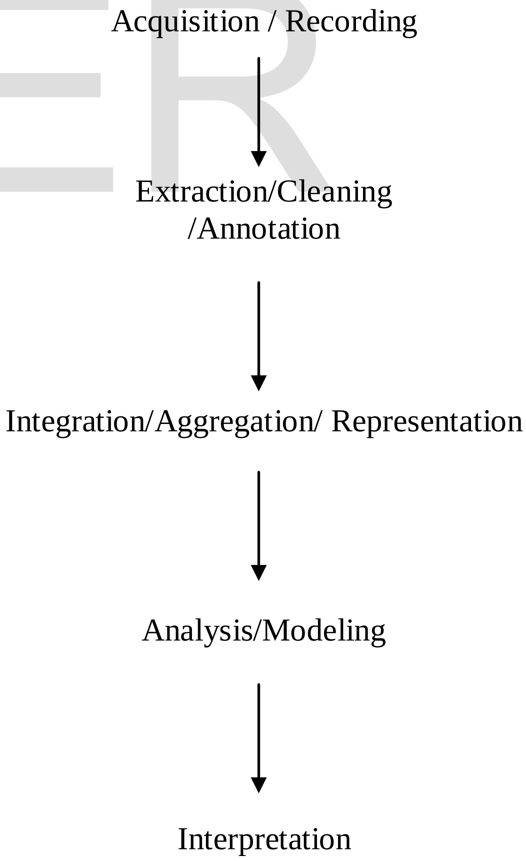
Acquisition / Recording
Extraction/Cleaning /Annotation
Integration/Aggregation/ Representation
Analysis/Modeling
Interpretation

Figure 1: Major steps in Big data analysis

The big data needs- heterogeneity, scale, timeliness, privacy and human collaboration- make these tasks challenging. The following are the five phases in big data analysis pipeline.

Data Acquisition and Recording

Not all the data in the world is valuable. From the available huge amount of data in which much of them are of no interest, the useful information must be filtered out and compressed by orders of magnitude. This data reduction should be performed so carefully and intelligently by processing this raw data to a size that users can handle while preserving all the necessary information. The next challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. Metadata acquisition systems can minimize the human burden in recording metadata.

Information Extraction and Cleaning

The collected information will not be in a format ready for analysis. So an information extraction process is needed to pull out the required information from the underlying sources and expresses it in a structured form suitable for analysis.

Data Integration, Aggregation and Representation

Often, it is not enough merely to record the data and then move it into a repository. It is due to the heterogeneity of the data flood. To an extent, metadata help to overcome this difficulty but still it poses some challenges 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. Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Effective database designs need to be created, 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.

Data Modeling and analysis

Big data requires specific methods for querying and mining which are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today.

Interpretation

Interpretation of the big data analysis results usually involves all the assumptions made and retracing the analysis. The provenance of the result also should be provided along with the results which are the supplementary information that explains how each result was derived and, based upon precisely what inputs.

5. Technology and Application Challenges

Much of the technology required for big-data computing is developing at a satisfactory rate due to market forces and technological evolution. For example, disk drive capacity is increasing and prices are dropping due to the ongoing progress of magnetic storage technology and the large economies of scale provided by both personal computers and large data centers. Other aspects require more focused attention, including:

High-speed networking:

Although one terabyte can be stored on disk for just $100, transferring that much data requires an hour or more within a cluster and roughly a day over a typical “high-speed” Internet connection. (Curiously, the most practical method for transferring bulk data from one site to another is to ship a disk drive via Federal Express.) These bandwidth limitations increase the challenge of making efficient use of the computing and storage resources in a cluster. They also limit the ability to link geographically dispersed clusters and to transfer data between a cluster and an end user. This disparity between the amount of data that is practical to store, vs. the amount that is practical to communicate will continue to increase. We need a “Moore’s Law” technology for networking, where declining costs for networking infrastructure combine with increasing bandwidth.

Cluster computer programming:

Programming large-scale, distributed computer systems is a longstanding challenge that becomes essential to process very large data sets in reasonable amounts of time. The software must distribute the data and computation across the nodes in a cluster, and detect and remediate the inevitable hardware and software errors that occur in systems of this scale. Major innovations have been made in methods to organize and program such systems, including the MapReduce programming framework introduced by Google. Much more powerful and general techniques must be developed to fully realize the power of big-data computing across multiple domains.

Extending the reach of cloud computing:

Although Amazon is making good money with AWS, technological limitations, especially communication bandwidth, make AWS unsuitable for tasks that require extensive computation over large amounts of data. In addition, the bandwidth limitations of getting data in and out of a cloud facility incur considerable time and expense. In an ideal world, the cloud systems should be geographically dispersed to reduce their vulnerability due to earthquakes and other catastrophes. But, this requires much greater levels of interoperability and data mobility. The OpenCirrus project is pointed in this direction, setting up an international testbed to allow experiments on interlinked cluster systems. On the administrative side, organizations must adjust to a new costing model. For example, government contracts to universities do not charge overhead for capital costs (e.g., buying a large machine) but they do for operating costs (e.g., renting from AWS). Over time, we can envision an entire ecology of cloud facilities, some providing generic computing capabilities and others targeted toward specific services or holding specialized data sets.

Machine learning and other data analysis techniques:

As a scientific discipline, machine learning is still in its early stages of development. Many algorithms do not scale beyond data sets of a few million elements or cannot tolerate the statistical noise and gaps found in real-world data. Further research is required to develop algorithms that apply in real-world situations and on data sets of trillions of elements. The automated or semi-automated analysis of enormous volumes of data lies at the heart of big-data computing for all application domains.

Widespread deployment:

Until recently, the main innovators in this domain have been companies with Internet-enabled businesses, such as search engines, online retailers, and social networking sites. Only now are technologists in other organizations (including universities) becoming familiar with the capabilities and tools. Although many organizations are collecting large amounts of data, only a handful are making full use of the insights that this data can provide. We expect "big-data science" – often referred to as eScience – to be pervasive, with far broader reach and impact even than previous-generation computational science.

Security and privacy:

Data sets consisting of so much, possibly sensitive data, and the tools to extract and make use of this information give rise to many possibilities for unauthorized access and use. Much of our preservation of privacy in society relies on current inefficiencies. For example, people are monitored by video cameras in many locations – ATMs, convenience stores, airport security lines, and urban intersections. Once these sources are networked together, and sophisticated computing technology makes it possible to correlate and analyze these data streams, the prospect for abuse becomes significant. In addition, cloud facilities become a cost-effective platform for malicious agents, e.g., to launch a botnet or to apply massive parallelism to break a cryptosystem. Along with developing this technology to enable useful capabilities, we must create safeguards to prevent abuse.

6. Conclusion

Big data is going to be a revolution in the IT world. Proper analysis and processing of the available information make faster advances in scientific applications and help the enterprises and public sector to increase their profit and success. It has got applications in vast majority of areas including business as well as science and technology. However many technical challenges mentioned in this paper must be addressed before it can be utilized fully. Advanced research and experiments are to be conducted and encouraged to extract the benefits of big data fully, by overcoming the existing technological and application challenges. This survey paper enlightens the opportunities of big data and also addresses some of the challenges of it.

References

[1] "Big data-extracting the value from your digital landfills" AIIM.

[2] "A guide to big data workload management challenges" By George Gilbert, underwritten by Datastrax.

[3] "A Practical Guide to Big Data Opportunities, Challenges & Tools" Laura Wilber

[4] "Big Data for Development: Challenges & Opportunities" by Emmanuel Letouzé, May 2012.

[5] "Big data: The next frontier for innovation, competition and productivity" by McKinsey global institute May 2011.

[6] "Big Data and Cloud Computing: Current State and Future Opportunities" Divyakant Agrawal , Sudipto Das, Amr El Abbadi University of California, Santa Barbara.

[7] " Big data: definition and charecteristics" A community white paper developed by leading researchers across the United States.