CHALLENGES AND OPPORTUNITIES WITH BIG DATA

[1]Dr.SunilTekale, [2]Mr.P.Amarnath, [3]A S GousiaBanu, [4]P.Pavani
[1]Professor, [2,3,4]Assistant Professor
Department of Computer Science and Engineering,
Malla Reddy College of Engineering, Hyderabad

ABSTRACT

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.

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs 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.

Review:

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[3]. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge. 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. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process and bringing the data back[5]. This is an obstacle to carrying over the interactive elegance of the first generation of SQL driven OLAP systems into the data mining type of analysis that is in increasing demand. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

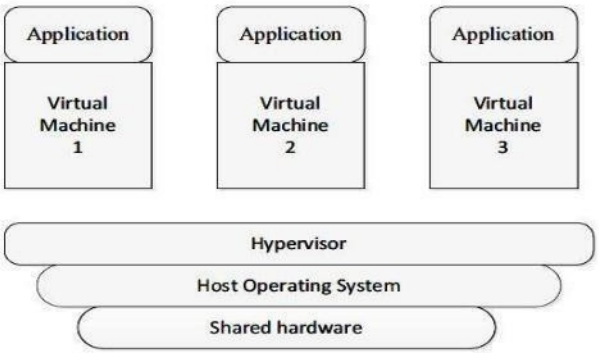
Keywords: Semantic, data base, data mining, data acquisition, Data extraction, cleaning

Introduction

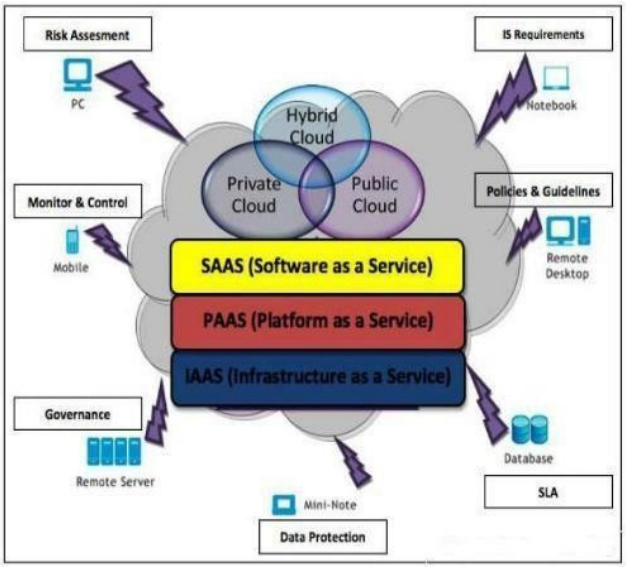
Big Data has the potential to revolutionize not just research, but also education [1]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [2]. 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

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-11, 2017
DOI: 10.21276/ijcesr.2017.4.11.26

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. It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [3], by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates a savings of 300 billion dollars every year in the US alone. In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data) [4], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative , computational social sciences 2 (a new methodology fast growing in popularity because of the dramatically lowered cost of obtaining data), financial systemic risk analysis (through integrated analysis of a web of contracts to find dependencies between financial entities), homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged information and other events, known as Security Information and Event Management (SIEM)), and soon.



Challenges in Big Data Analysis

There are various challenges that needs to be addressed in handling Big Data. As the name suggest big data, its really very big to mange because of

1. Heterogeneous Formats
2. Scale
3. Timeliness
4. Incompleteness
5. Privacy
6. Human Collaboration

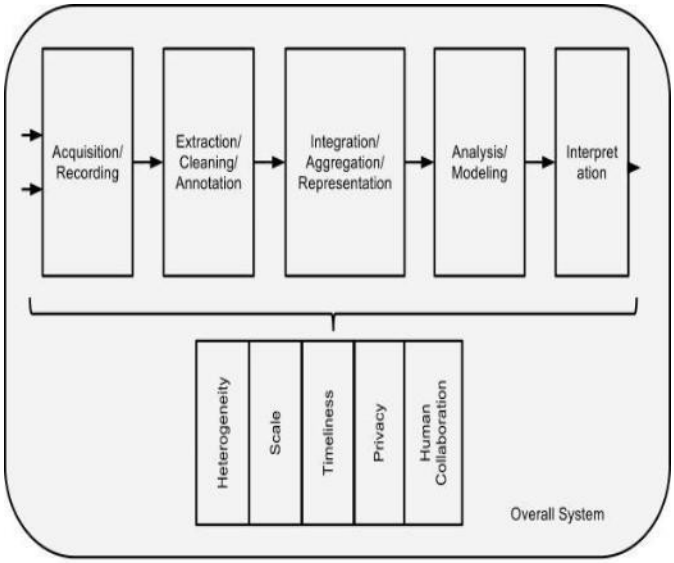


Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Data Acquisition and Recording

Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate

of an elderly citizen, and presence of toxins in the air we

Information Extraction and Cleaning

Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements (possibly with some associated uncertainty), and image data such as x-•rays. We cannot leave the data in this form and still effectively analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge. Note that this data also includes images and will in the future include video; such extraction is often highly application dependent (e.g., what you want to pull out of an MRI is very different from what you would pull out of a picture of the stars, or a surveillance photo). In addition, due to the ubiquity of surveillance cameras and popularity of GPS enabled mobile phones, cameras, and other portable devices, rich and high fidelity location and trajectory (i.e., movement in space) data can also be extracted.

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.

Query Processing, Data Modeling, and Analysis

Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-•-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trust worthy breathe, to the planned square kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce peta bytes of data today. relationships, to disclose inherent clusters, and to uncover hidden relationships and models.

Interpretation

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so.

Opportunities in Big Data

1. Soaring Demand for Analytics Professionals

2. Huge Job Opportunities & Meeting the Skill Gap

3. Big Data Analytics: A Top Priority in a lot of Organizations

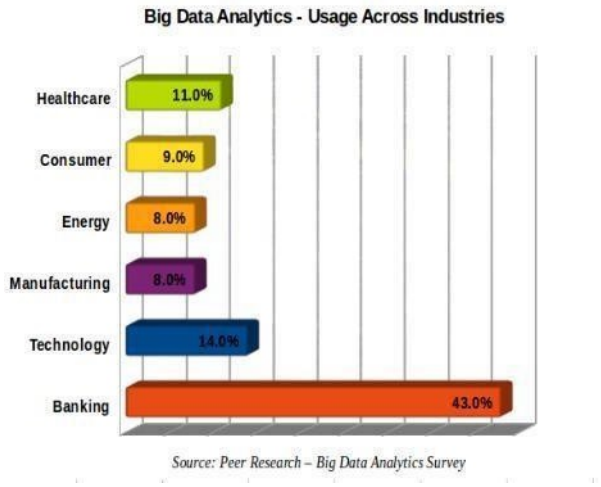
4. Adoption of Big Data Analytics is Growing:

5. Analytics: A Key Factor in Decision Making

6. The Rise of Unstructured and Semi structured Data Analytics:

7. Surpassing Market Forecast / Predictions for Big Data Analytics:

8. Numerous Choices in Job Titles and Type of Analytics:



Proposed System:
Roughly there are two types of approaches for big data analytics

1. Parallelize existing (single-machine) algorithms.

2. Design new algorithms particularly for distributed settings

The problem now is we take many things for granted on one computer On one computer, have you ever worried about calculating the average of some numbers? Probably not. You can use Excel, statistical software (e.g., R and SAS), and many things else We seldom care internally how these tools work Can we go back to see the early development on one computer and learn some lessons/experiences.
Consider the example of matrix-matrix products C = A × B, A ∈ R n×d , B ∈ R d×m where Cij = X d k=1 AikBkj This is a simple operation. You can easily write your own code
A segment of C code (assume n = m here) for (i=0;i<n;i++)
for (j=0;j <N;J++)
{ C[I][J]=0;
For(k=0;k<n;k++) C[i][j]+=a[i][k]*b[k][j];
}
But on Matlab (single-thread mode) $ matlab –single Comp Thread>> tic; c = a*b; toc Elapsed time is 4.095059 seconds.

CPU ↓ Registers ↓ Cache ↓ Main Memory ↓ Secondary storage (Disk) ↑ : increasing in speed ↓ : increasing in capacity Optimized BLAS: try to make data

For big-data analytics, we are in a similar situation We want to run mathematical algorithms (classification and clustering) in a complicated architecture (distributed system) But we are like at the time point before optimized BLAS was developed.

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.

REFERENCES

[1] Big Data. Nature (http://www.nature.com/news/specials/bigdata/index.html), Sep 2008.
[2] Data, data everywhere. The Economist (http://www.economist.com/node/15557443), Feb2010.
[3] Drowning in numbers – Digital data will flood the planet—and help us understand it better. The Economist(http://www.economist.com/blogs/dailychart/2011/11/bigdata-0), Nov2011.

[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 – A community white paper developed by leading researchers across the UnitedStates.http://cra.org/ccc/docs/init/bigdata whitepaper.pdf, Mar 2012.
[5] S. Lohr. The age of big data. New York Times(http://www.nytimes.com/2012/02/12/sun day- review/bigdatas-impact-in-the-world.html), Feb2012.
[6]J.Manyika,M.Chui,B.Brown,J.Bughin,R.Dob bs,C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.
[7] Y. Noguchi. Following Digital Breadcrumbs to Big Data Gold. National Public Radio(http://www.npr.org/2011/11/29/1425219 10/thedigital breadcrumbs -that-lead-to-big-data), Nov2011.
[8] Y. Noguchi. The Search for Analysts to Make Sense of Big Data. National Public Radio (http://www.npr.org/ 2011/11/30/142893065