ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 1, pp: (96-105), Month: January - March 2015, Available at: www.researchpublish.com

Novel Approaches to Big Data Management

JAVED MOHAMMED

Department of Computer Science, New York Institute of Technology, Old Westbury, NY, United States.

Abstract: **Fueled by the pervasiveness of the Internet, unprecedented computing power, ubiquitous sensors and meters, addictive consumer gadgets, inexpensive storage and (to-date) highly elastic network capacity, digital information is streaming at a mind-boggling rate. Big Data is not a precise term; rather, it is a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes data sets that are growing exponentially and that are too large, too raw, or too unstructured for analysis using relational database techniques. Whether terabytes or peta bytes, the precise amount is less the issue than where the data ends up and how it is used. Today's applications are expected to manage a variety of structured and unstructured data, accessed by massive networks of users, devices, and business locations, or even sensors, vehicles and Internet-enabled goods. As the need for data access expands to the network edge, most databases are still grounded in a central data center. DBAs continue to toil, as they have for decades, moving databases to more powerful hardware, or bigger clusters, and constantly re-designing them in order to accommodate business growth [8]. Big data has the potential to improve the quality of services; enable infrastructure that businesses depend on to adapt continually and efficiently; improve the performance of employees; help organizations better understand customers; and reduce liability risks. Big Data Analytics and marketing models presents Next generation of networks that are in a prime position to monetize rich supplies of customer information, while being mindful of legal and privacy issues. As data assets are transformed into new revenue streams will become integral to high performance. The speed of business these days and the amount of data that we are now swimming in mean that we need to have new ways and new techniques of getting at the data, finding out what's in there, and figuring out how we manage it.**

Keywords**: Big Data, Next Generation Networks, Network Transformation, High Performance, Data Center.**

1. INTRODUCTION

Recent rapid advances in the proliferation of smart devices and popularity of social networking is generating unprecedented amounts of data, both structured and unstructured, whether it be text, audio or video in information technologies both hardware and software that are creating a new revolution in discovery and learning. Enormous numbers of tiny but powerful sensors are being deployed to gather data deployed on the sea floor, in the forest canopy, on the sides of volcanoes, in buildings and bridges, in living organisms. Modern scientific instruments, from gene sequencers to telescopes to particle accelerators, generate unprecedented amounts of data. Other contributors to the data tsunami include point-of-sale terminals, social networks, the World Wide Web, mobile phones (equipped with cameras, accelerometers, and GPS technology), and electronic health records. These sensors, instruments, and other information sources and, indeed, simulations too produce enormous volumes of data that must be captured, transported, stored, organized, accessed, mined, visualized, and interpreted in order to extract knowledge and determine action. Key advances include jumps in the availability of rich streams of data, precipitous drops in the cost of storing and retrieving massive amounts of data, exponential increases in computing power and memory, and jumps in the prowess of methods for performing machine learning and reasoning. These advances have come together to create an inflection point in our ability to harness large amounts of data for generating insights and guiding decision making. More data means more opportunities for analysts to gain insight about customers, and hence new ways to serve customers, and to offer well-tailored products and services.

Logs of interactions are regularly captured by web services such as search engines and online shopping sites and with specialized point-of-sale terminals. Outside the web, organizations regularly capture transactions and operational data on enterprise databases. Sensor networks are sprouting in multiple venues in support of a wide spectrum of applications, from sets of traffic flow sensors in highways, to meshes that track moisture in vineyards. Beyond fixed sensors, dynamic sensor networks are being created by constellations of connected mobile devices. Analysts hold a treasure trove of customer-behavior data, and there is gold in Big Data. Next-generation analytics can help mobile operators mine and refine the value of this new economic asset. The race is on to collect as many details as possible and mobile network operators are in a prime position to know the most, much of it available who wishes to mine it for insights.

New evidential paradigms and sensing technologies in managing big data are transforming all of the major sciences into e Sciences – reshaping the way science is carried out in the 21st century. Scientific efforts have long relied on extracting insights from measurements, models, and simulations. However, data-centric methodologies have grown in importance in the sciences with the growing flood of data, and concomitant rise in power of computing resources. Data has grown by orders of magnitude over the last 20 years with the use of such tools as automated sky surveys in astronomy, gene sequencing in biology, web logs of behavioral data in sociology, and the deployment of sensor networks in agriculture and the environment. New information-centric sub disciplines, such as bioinformatics, astro informatics, and matin formatics have arisen. Computational tools critically enable discovery and inference with methods for simulating phenomena, visualizing processes, extracting meaning from massive data sets – and directing the collection of new data [6]. Computational models for learning and inference also show great promise in their ability to provide guidance to scientists on the overall hypothetical deductive cycle of discovery, where an initial predictive model learned from data guides the iterative collection of new data, followed by model revision and new inferences that once again point the way to new data to collect. Advances in machine learning have multiple touch points with the process of science. The methods even cut to the core goals of scientific inference with algorithms that can elucidate or rule out the existence of causality in processes. The melding of new data resources with methods of simulation, visualization, and machine learning will enable scientists to pursue insights and comprehension at a faster pace than ever before in many key areas of scientific discovery.

Opportunities abound for tapping our new capabilities more broadly to provide insights to decision makers and to enhance the quality of their actions and policies. Great progress has been made in assessing the value of collecting additional information before taking action. The likelihoods output by the predictive models lay at the heart of decision analyses that weigh the costs and benefits of different actions. We describe below several key areas of national priority in which applying predictive models and decision analysis is critical to ensuring our nation’s prosperity and wellbeing.

1.1 KEY AREAS OF NATIONAL PRIORITY:

➢ ***Enabling Evidence-Based Healthcare:***

Critically valuable data continues to fall on the floor in healthcare, where paper records and manual procedures for information capture still dominate the landscape. There is tremendous upside to investing in data collection and machine learning in healthcare. Predictive models can be harnessed to simultaneously enhance the quality of healthcare and lower its costs. Opportunities for machine learning in healthcare include the construction of predictive diagnostic models that are pressed into service when performing automated triage, inter-specialty referral, and diagnosis of patients presenting with sets of complaints, signs and symptoms. Broader applications include the use of predictive models to optimize the management of chronic diseases, a chief source of healthcare costs and the root of poor quality of life for our aging population. In one approach, predictive models can inform how scarce resources should be applied for maximal benefit. As an example, a recent study of 300,000 electronic records of visits to emergency departments in the Washington, DC [10], area highlights the value of machine learning to lower costs by minimizing re-hospitalizations within 30 days through selective investment in post-discharge support where it will be most useful. Machine learning will also provide a foundation for clinical discovery that promises to identify associations between gene composition and risks of illness, disease progression, and the efficacy of pharmacological agents, and the broader hope of identifying interventions and cures. Within the next 12-36 months, people will be able to have their entire genome sequenced for less than $1000 a near future that one expert referred to as a “a 10-mile-wide asteroid heading toward us.” There are great opportunities to learn from the massive quantity of data that will likely become available, and multiple issues, from storage to privacy to methods for learning from large amounts of genomic data, must be confronted in the near term.

➢ ***Enabling the New Biology:***

Over the last several decades, new experimental techniques coupled with advanced data analytics strategies have utterly transformed the way we study biology. Pursuit of this field has become an increasingly data-centric endeavor, now comprising iterative stages of computation and experiment. Today, genomes of plants and animals are typically captured with gigabytes of data, and machine learning is used to identify networks and modules at the core of the operation of cells. As our knowledge has grown, so, too, has our appreciation for the complexity of biological systems. For example, at a minimum, there is simply no way to grasp epigenetic processes without computation. Indeed, biological function arises from the complex interplay of individual components, and the emerging science of systems biology which strives to integrate biological structures and create predictive models representing holistic functions and behaviors offers the potential to radically alter how we understand disease, develop new drugs and interventions, and engineer organisms to synthesize important byproducts such as biofuels. Biological data capture and analytical requirements quickly scale to multiple terabytes of information capturing information at each time slice of periods of study and computation therefore sits at the leading edge of discovery in biology; it will continue to be front and center in building insights and moving forward with models and understandings amidst complexity.

- ***Enabling a Revolution in Transportation:***

Opportunities abound for enhancing the efficiency and reliability of multiple modalities of our transportation system with data-centric approaches. Recent efforts with machine learning on large-scale data sets about flows of traffic demonstrate how we can use predictive models to predict future road flows – and flows on roads that are not sensed in real time. Such predictions can be harnessed for generating context-sensitive directions, real-time routing, and load balancing. Predictive models for traffic flows in greater city regions have leveraged years of heterogeneous data, including in-road sensors, GPS devices in vehicles (mass transit and crowd-sourced from volunteers), information about road topology, accidents, weather, and major events such as sporting events. This work has led to fielded services that provide flow prediction and directions in major cities of the United States. Other opportunity areas in transportation include integrating predictive models with fluid dynamics and queue-theoretic models to build tools that help engineers to understand the value of new roadways and road upgrades. We can also apply simulations to understanding the real-world implications of different high-occupancy vehicle (HOV) and high-occupancy toll (HOT) policies and provide insights about the demand and flows on roads expected with different designs for dynamic tolling systems. In another area, meshing predictive models with automated planning systems can help to shift people from private vehicles to shared transportation solutions. For example, automated planning systems can assemble multimodal transportation plans that mix public and private transportation, offering people more flexible, end-to-end transportation alternatives. Predictive models combined with optimization can also be employed to mediate and optimize the operation and ease of participating in ridesharing systems. Data collection and learning models for driving, combined with advances in real-time sensing, will also play an important role in building safer cars that employ collision warning and avoidance systems. Such systems promise to reduce the unacceptably large number of deaths on U.S. roadways each year (more than 40,000 deaths and even higher numbers of disabling injuries per year). Methods for learning automated driving competencies from data will additionally be crucial in the development of autonomous vehicles that drive without human intervention. Beyond enhancing the safety of transport in private vehicles and enabling greater densities of vehicles on freeways, the development of autonomous vehicles will enable the creation of large-scale public micro transit systems that flexibly transport people from point to point within city regions, continuing to execute prediction and optimization for maximizing their availability and efficacy. The latter potential shift in the way we do most of our travel has deep implications for energy and the environment.

- ***Enabling Advanced Intelligence and Decision-Making for America's Security:***

Data-centric learning and modeling have an important and promising role to play at multiple levels in America’s security. Intelligence, surveillance, and reconnaissance (ISR) poses challenges that are well suited for data-centric computational analyses. Statistical methods for fusing information have undoubtedly been used for evidential analysis by government agencies engaged on security. However, late-breaking computing research on such topics as active learning, learning from network features, and the use of reliability indicators, can enhance our ability to effectively piece together data from heterogeneous sources and sensors. Machine learning and reasoning can be applied in both exploratory investigations and in focused predictive modeling. Methods for computing the value of information for real-time prediction or for offline planning of additional investments in data collection an direct dollars and assets to where they will have the highest expected value. Models can include machinery for effectively computing confidences, allowing for explicit consideration of false positive rates, so as to mediate the invasiveness of actions, and to gauge the impact of policies and investigations on the public. There is overall great opportunity for better coordinating national technical assets to fuse traditionally disparate data sets into richer models for pursuing answers to questions, as well as monitoring and exploration. There are opportunities also for enhancing sensory fusion and visualization of sensor data and inferences for commanders and soldiers using our defense systems and platforms, for enhanced situation awareness in peaceful times and during engagements. And while much of this work involves applying computational tools to classified information, it is possible to develop the underlying approaches and strategies on unclassified data sets and challenge problems.

- ***Enabling Personalized Education:***

Data-centric modeling in education is a challenging area that promises to enhance the way we educate our children. Several communities of computer scientists have been working to understand how computational models can assist with education. In particular, efforts in the cognitive science, user modeling, and intelligent tutoring systems communities include building predictive models and decision policies that are used in computer-based tutoring systems and also for developing more ideal teaching strategies. This research can be viewed as seeking bridges between machine learning and educational psychology. Other studies to date highlight opportunities for computational methods and analyses to serve as accessories to teachers. For example, predictive models may one day diagnose students' learning styles and then personalize presentation and content in a way that is engaging and efficient. Such predictive models have already been constructed from data collected from thousands of K-12 students using computer-based instructional tools in their schools, yielding predictions about which educational material will be most effective for each individual student based on his or her particular errors and understanding of the material. In addition, predictive models in educational settings can be used to forecast levels of engagement that students have with educational material, and actions that can be taken to enhance engagement or to re-engage students. An important, related area of work involves generating and operating simulations that react intelligently to the actions of individuals or groups of individuals trying to solve a problem in a simulated space, i.e., reacting to gestures, behaviors, or questions, etc., in a way that responds to the individual students, the subject matter, and the pedagogy thereby shedding light on the various factors and approaches being simulated. Beyond usage in these systems, data-centric analyses can also advise designs for the best ways to distribute and combine scarce expert pedagogy with always-available online content and tutoring. Importantly, any research into education data analytics requires deeply rooted involvement of teachers and tutors from the start – as they can not only help steer these analyses from their past experiences but they can also take the new knowledge that is generated and incorporate it into their teaching moving forward.

- ***Enabling the Smart Grid:***

Our nation's electric power grid is an antiquity of networking and distribution that can and should be significantly upgraded based on design principles drawn from modern information networking and communications. Numerous opportunities for enhancing the reliability and efficiency of the power system will require new overlays and architectures that make possible finer-grained control of distribution and metering. Fine-grained control of distribution will require new kinds of innovations that provide greater transparency and informational awareness and that allow for modulation of flows and payments at multiple levels of the power grid. Prediction of load over time plays a central role in today's power generation industry, where ongoing agreements are made among power suppliers on a daily or hourly manner for transferring large quantities of power among regions of the country. Decisions are based on forecasts about future demand and availabilities of power from different sources. These forecasts of future demands are still largely done manually and in a coarse manner. There are significant opportunities for collecting data on power loads over time, learning predictive models from that data, and employing decision analyses to generate intelligent power distribution strategies. For example, researchers have shown that instrumenting a home with just three sensors for electricity, power, and water makes it possible to determine the resource usage of individual appliances. There is an opportunity to transform homes and larger residential areas into rich sensor networks, and to integrate information about personal usage patterns with information about energy availability and energy consumption, in support of evidence-based power management. Directions include developing new systems that provide people with tools to encode preferences about the timing of device usage and interfaces/communication protocols that enable devices to "talk" independently with the grid. Many rich scenarios for power load balancing and cost optimization are enabled by such tools. For example, on a hot summer day, HVAC compressors at multiple homes and facilities in a region can be coordinated and sequenced rather than being allowed to operate in a haphazard, independent manner so as to minimize peak loads while delivering effective cooling. Such systems can gracefully back off during high load situations, according to cooling preferences and willingness to pay, rather than impose large-scale brownouts on a region. As another example, multiple tasks can be deferred to times when power is less expensive, and the decisions about the best timing of usage in advance of a preferred deadline can be supported by predictive models. We believe that urgent challenges are on the horizon in this realm and these will drive efforts in the near term. Such challenges include developing approaches to scheduling the charging of large numbers of electric vehicles in a manner that balances the load within neighborhoods, across towns, and across different regions of the larger power grid.

- ***Technical Approaches to Enhancing Data Privacy:***

The variety and volume of data collected, and the potential to use this to improve our daily lives, will continue to grow for the foreseeable future. While the potential benefits are great, for numerous application areas, data privacy and data security must be addressed to achieve the greatest benefits while protecting civil liberties. These privacy and civil liberties issues are central, and will require a modernization of existing policies for collecting and using data about individuals. Importantly, there is a key role for technology as well as political process in managing the tradeoff between privacy and the benefits of collecting and using data. For example, consider the potential for training a predictive model to discover which treatment works best for a new flu strain that is spreading across the country, based on real-time medical records from thousands of emergency rooms and hospitals. In the past, the only methods to train such a predictive model would have required first combining the data from these thousands of organizations into a central data base, potentially raising privacy concerns. Today privacy enhanced machine learning methods exist that enable training such a model without the need to create a central database. Instead of centralizing the data, these approaches distribute the machine learning computation, sharing encrypted intermediate results so that the data need never leave the hospital that collected it. Another approach to privacy is differential privacy, where special forms of noise are adding to data to obscure peoples' identities without incurring significant losses in the accuracy of inferences. Other approaches aim to make sharing transparent and controllable, allowing people to both view and make decisions about the data they share. Providing such views and controls can allow people to make tradeoffs based on their preferences. For example, people may be comfortable with providing specific types of data in return for enhanced personalization that is enabled by sharing information. At times, people may wish to share data in an altruistic manner to assist, for example, with healthcare studies or with the crowdsourcing of traffic flows in a city. Computational tools can allow people to specify how and when their data can be used, including the type and quantity of data that may be collected. At the same time, service providers who seek access to personal data can work to understand the varying preferences about privacy and sharing and work to optimize their services based on data that people are willing to share. To manage the important privacy issues, it will be important to invest in further research into technical approaches to privacy and sharing, to add flexibility as society determines how we can best achieve the benefits of this new data while protecting civil liberties.

2. RECENT INNOVATIONS TRENDS AND CONCERNS

Data is increasingly being recognized as a rich resource flowing through organizations from a continually growing range of sources. But to realize its full potential, this data must be accessed by an array of users to support both real-time decision making and historical analysis, integrated with other information, and still kept safe from hackers and others with malicious intent. Fortunately, leading vendors are developing products and services to help. Here, DBTA presents the list of Trend-Setting Products in Data and Information Management for 2015.

2.1 Predictions for Data and Analytics:

As analytics continues to play a larger role in the enterprise, the need to leverage and protect the data looms larger. According to the IDC, the big data and analytics market will reach $125 billion worldwide in 2015 [9]. Here are 10 predictions from industry experts about the data and analytics in 2015 [3].

1. Hadoop – Hadoop will become a worldwide phenomenon, believes Concurrent CEO Gary Nakamura, who notes that Hadoop has shown tremendous growth throughout Europe and Asia, and that expansion will only continue. A key to Hadoop being able to become an enterprise backbone, is the ROI businesses can expect from using it, and products and tools continue to evolve to keep pace with the technology's trajectory. According to Actian, SQL will be a "must-have" to get the analytic value out of Hadoop data. We'll see some vendor shake-out as bolt-on, legacy or immature SQL on Hadoop offerings cave to those that offer the performance, maturity and stability organizations need.

2. Enterprise Security – With the seemingly never-ending stream of news reports of hacks and data leaks, one of the major data issues of 2014 that we can expect to continue in 2015 is big data breaches. "There is nothing you can do to stop a zero-day vulnerability, but the question is what do we do about it," stated Walker White, president of data as a service providerBDNA. At this point it isn't about keeping the hackers out, but how companies react to protect

their data once the hackers have penetrated their systems. “Security ultimately is an arms race, there are very few mechanisms that simply can’t be broken, it tends to just be how far ahead can you stay of the people that are trying to break in,” agreed Seth Proctor, CTO of NuoDB.

3. Business Intelligence – The growth of BI tools that are more user friendly for the average business employee will help to take some of the burden of IT teams. To do this, more BI providers will incorporate search into their interfaces to make the tools more accessible to average business users, according to Thought spot CEO Ajeet Singh.

4. Cloud – The cloud will increasingly become the deployment model for BI and predictive analytics, particularly with the private cloud powered by the cost advantages, according to Actian.

5. Hybrid Architecture – Hybrid architectures will become the norm for many organizations, according to Steven Riley of Riverbed Technology. Even though cloud computing and third-party hosting will continue their rapid expansion, on-premise IT will remain a reality for 2015 and beyond. “In the coming year, analytics will have the power to become the next killer app to legitimize the need for hybrid cloud solutions,” adds Revolution Analytics CEO Dave Rich. Analytics has the ability to mine vast amounts of data from diverse sources, deliver value and build predictions without huge data landfills [5]. In addition, the ability to apply predictions to the myriad decisions made daily – and do so within applications and systems running on-premises – is unprecedented.”

6. Medical Data – When the average person thinks of their most important personal data security, most think about credit card information. Bit glass, which provides security for cloud apps and mobile devices, believes that medical records are 50 times more valuable on the black market than credit cards. Bit glass predicts that medical records will become a bigger target for data attacks than traditional methods such as credit cards. This will result in scrutiny pertaining to HIPAA regulations. Regulations stipulate that health organizations must report data breaches that affect more than 500 people.

7. Data Science – As organizations gain a greater appreciation of the role that that data is playing data scientists are in greater demand, yet there are not enough qualified data scientists, according to EXASOL, an in-memory database company. Joe Caserta of Caserta Concepts believes that chief analytics officers (CAO) will now play role in the enterprise. As data-rich organizations continue to adopt a more strategic approach to big data, it makes sense that the responsibility for all that information needs to sit with someone who can apply the analytics big picture to all parts of the organization - the CAO. The coming year will be time for data-driven organizations to dedicate resources and executive commitment to the function.

8. Internet of Things - OpenText predicts consumers will begin to become more aware of the IoT all around them - from smart watches to cars with built-in sensors, and Vormetric, a provider of security solutions, believes that the IoT will trigger a greater enterprise emphasis on securing big data using encryption. More personalized private data will be stored and analyzed by data analysis tools in the future.

9. Location Data – Technologies will emerge in 2015 – full stack virtualization, pervasive visibility, and hybrid deployments – that will create a form of infrastructure mobility that allows organizations to optimize for location of data, applications, and people, says Riley of Riverbed. He predicts that organizations that begin to disperse their data to multiple locations will begin to gain significant competitive advantages.

10. NewSQL – NewSQL will start taking the place of some RDBMSs, according to Morris of NuoDB, who believes that NewSQL will begin to support enterprise-scale applications that traditionally were only held by RDBMSs.

Data and analytics will only become more important and valuable to the enterprise. As the technologies for putting data to greater use continue to multiply, it is clear that those opportunities also carry risk, and there is the need to better protect the data that is being amassed.

2.2 Changing Economics, New Opportunities:

The economics of the data center are changing. Today, we have cheaper hardware, cheaper disk storage, and cloud providers that are lowering prices every day. The biggest cost of information technology is no longer with storage, compute power, or data transfer speeds, it is in the time and development effort of building new applications and the opportunity cost of not moving at the speed of the business [7]. The data-centered data center makes organizations more agile. It makes it easier to find new applications for available data and helps deliver applications more quickly. It lowers the overall IT cost and provides a competitive advantage.

2.3 The Changing Nature of Data:

It used to be the case that most data we wanted to analyze came from sources of the data center: transactional systems, enterprise resource planning (ERP) applications, customer relationship management (CRM) applications, and the like. The structure, volume, and rate of the data were all fairly predictable and well known. Today a significant and growing share of data application logs, web applications, mobile devices, and social media comes from outside the data center, even outside our control. That data constantly evolves and as a result frequently uses newer, more flexible data formats such as JSON and Avro. That increases demands on both the systems themselves and on the people who manage and use them

3. TECHNOLOGY AND PRACTICAL CHALLENGES

It is hard to avoid mention of Big Data anywhere we turn today. There is broad recognition of the value of data, and products obtained through analyzing it [1]. Industry is abuzz with the promise of Big Data [2]. Government agencies have recently announced significant programs towards addressing challenges of Big Data. Yet, many have a very narrow interpretation of what that means, and we lose track of the fact that there are multiple steps to the data analysis pipeline, whether the data are big or small. At each step, there is work to be done, and there are challenges with Big Data.

- The first step is data acquisition. Some data sources, such as sensor networks, can produce staggering amounts of raw data. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information.

- The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. This metadata is likely to be crucial to downstream analysis. Frequently, the information collected will not be in a format ready for analysis. The second step is an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis.. Furthermore, we are used to thinking of Big Data as always telling us the truth, but this is actually far from reality. We have to deal with erroneous data: some news reports are inaccurate.

- 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. 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.

- Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and Big Data computing environments. 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 nonSQL 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.

- The 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.

- 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 [11]. Much more powerful and general techniques must be developed to fully realize the power of big-data computing across multiple domains

- 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 other catastrophes. But, this requires much greater levels of interoperability and data mobility.
- As a scientific discipline, for machine learning 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.
- 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.

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, users will try to understand, and verify, the results produced by the computer. There is a multi-step pipeline required to extract value from data. Heterogeneity, incompleteness, scale, timeliness, privacy and process complexity give rise to challenges at all phases of the pipeline. Furthermore, this pipeline is not a simple linear flow – rather there are frequent loops back as downstream steps suggest changes to upstream steps.

Today, a single database can no longer meet enterprise requirements. The requirements of interactive applications, enterprise, mobile, and we are met by operational databases. The requirements of business intelligence applications are met by analytical databases and/or Apache Hadoop distributions. Finally, off the shelf and legacy application requirements are met by relational databases. While a NoSQL database like Couch base Server is a general-purpose database, it works well with Apache Hadoop distributions for analysis and Apache Lucene distributions for full text search.

4. RECOMMENDED SOLUTIONS ADOPTED

Big-data computing is perhaps the biggest innovation in computing in the last decade. We have only begun to see its potential to collect, organize, and process data in all walks of life. A modest investment by the federal government could greatly accelerate its development and deployment. Investments in big-data computing will have extraordinary near-term and long-term benefits. The technology has already been proven in some industry sectors; the challenge is to extend the technology and to apply it more widely. Below we list specific actions that would greatly accelerate progress.

- Invest in higher capacity networking infrastructure, both in the backbones as well as access by major research universities and government labs. Assist cloud service providers in connecting their systems to this high capacity network backbone (e.g., with tax incentives).
- Establish a networked collection of cluster systems with an initiative, but with greater coordination, oversight, and network bandwidth between them. These systems should be geographically dispersed, e.g., at multiple universities and research laboratories. Some of these machines would be made available to researchers pursuing applications of big-data computing, while the others would be made available to systems researchers, exploring both the operations of individual clusters, as well as connecting multiple clusters into a cloud
- 0Research thrusts within computing must cover a wide range of topics, including: hardware and system software design; data-parallel programming and algorithms; automatic tuning, diagnosis and repair in the presence of faults; scalable machine learning algorithms; security and privacy; and applications such as language translation and computer vision. Interdisciplinary programs should marry technologists with applications experts with access to extremely large datasets in other fields of science, medicine, and engineering

- Construct special-purpose data centers for the major e Science programs. Possible economies of scale could be realized by consolidating these into a small number of "super data centers" provisioned as cloud computing facilities. This approach would provide opportunities for technologists to interact with and support domain scientists more effectively. These efforts should be coupled with large-scale networking research projects.
- Encourage the deployment and application of big-data computing in all facets of the government, ranging from the IRS, intelligence agencies (multimedia information fusion), the CDC (temporal and geographic tracking of disease outbreaks), and the Census Bureau (population trends).

It is imperative to design and create a Modern Data Architecture that is flexible and powerful to support future requirements. This architecture need to leverage polyglot persistence, support agile development, is provided as a service, enable a flow data, is adaptive, and is elastic. By doing so, it enables enterprises to meet future requirements with less overhead, less cost, and less time. Couch base Server is a scalable, enterprise-grade NoSQL database. It supports a flexible data model via documents and JSON. It integrates with Apache Hadoop distributions, and is certified by Cloud era for Cloud era Enterprise [4]. It integrates with Lucid Works and Elastic search products. It can be integrated with Apache Storm. It is a key component of real-time big data architectures at eBay, PayPal, Live Person, LinkedIn, and more. It's the foundation of modern data architecture.

5. THE FUTURE

Data-driven science is reshaping the everyday life of humans in the 21st century. The combination of rich data sources, new computing technology, and advanced data mining and machine learning algorithms allows scientists to gain much deeper insights into many scientific phenomena than ever before possible. We are just beginning to revolutionize how we study the workings of the Universe; how we diagnose and treat medical ailments; how we generate, price, and deliver energy to homes and businesses; how we gather intelligence and "connect the dots" between multiple sources of information; etc. Fundamental research in computer science has played a critical role in creating the computing technology and the analysis techniques underlying the transformation in discovery and learning embodied by eScience. Continued forward progress is essential to our nation's leadership and essential to a broad spectrum of federal agencies fulfilling their missions. This progress requires increased investment in the tools and techniques of eScience, and close partnership between disciplinary scientists and computer scientists. America's competitiveness depends upon a vigorous eScience effort.

6. CONCLUSION

Today, data available via the Internet, sensor networks, and new and higher resolution sensors across the sciences allow us to capture more data about people and the world than ever before and the quantities of data available are accelerating. Coupled with recent advances in machine learning and reasoning, as well as rapid rises in computing power and storage, we are transforming our ability to make sense of these increasingly large, heterogeneous, noisy or incomplete datasets collected from a variety of sources; to visualize and infer important new knowledge from the data; and to guide action and policies in mission-critical situations, enabling us to make the best decisions. The predictive models and decision analyses will transform many facets of our daily lives, from healthcare delivery to transportation to energy and the environment. These methods will be critical for protecting America from threats, and they have the potential to alter how we educate the next generation, how we interact with one another, and how we protect our personal privacy and security in an era of constant connectivity and unfiltered access. In this article we illustrate how data analytics is critical to address our nation's priorities and to ensure our nation's prosperity well into the 21st century.

REFERENCES

[1] Big Data. Nature http://www.nature.com/news/specials/bigdata/index.html, Sep 2008.

[2] 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.

[3] PARKHILL, D. The Challenge of the Computer Utility. Addison- Wesley Educational Publishers Inc., US, 1966.

[4] Cloudera, Hadoop training and support [online]. Available from: http://www.cloudera.com/.

[5] BECHTOLSHEIM, A. Cloud Computing and Cloud Networking. talk at UC Berkeley, December 2008.

[6] CHANG, F., DEAN, J., GHEMAWAT, S., HSIEH, W., WALLACH, D., BURROWS, M., CHANDRA, T., FIKES, A., ANDGRUBER, R. Bigtable: A distributed storage system for structured data. In Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06) (2006).

[7] CHENG, D. PaaS - onomics: A CIO's Guide to using Platform-as-a-Service to Lower Costs of Application Initiatives While Improving the Business Value of IT. Tech. rep., Long Jump, 2008.

[8] GHEMAWAT, S., GOBIOFF, H., AND LEUNG, S.-T. The Google file system. In SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles (New York, NY, USA, 2003), ACM, pp. 29–43. Available from: http://portal.acm.org/ft_gateway.cfm?id=945450& type=pdf & coll=Portal &dl=GUIDE&CFID =19219697& CFTO KEN=50259492

[9] RANGAN, K. The Cloud Wars: $100+ billion at stake. Tech. rep., Merrill Lynch, May 2008.

[10] SIEGELE, L. Let It Rise: A Special Report on Corporate IT. The Economist (October 2008).

[11] DEAN, J., AND GHEMAWAT, S. Mapreduce: simplified data processing on large clusters. In OSDI'04: Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation (Berkeley, CA, USA, 2004), USENIX Association, pp. 10–10.