This white paper was created through a distributed conversation among many prominent researchers listed below. This conversation lasted a period of approximately three months from Nov. 2011 to Feb. 2012. Collaborative writing was supported by a distributed document editor. See last page for a complete list of contributors.

Challenges and Opportunities with Big Data

A community white paper developed by leading researchers across the United States

Executive Summary

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data.'' While the promise of Big Data is real -- for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 -- there is currently a wide gap between its potential and its realization.

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. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

During the last 35 years, data management principles such as physical and logical independence, declarative querying and cost-based optimization have led, during the last 35 years, to a multi-billion dollar industry. More importantly, these technical advances have enabled the first round of business intelligence applications and laid the foundation for managing and analyzing Big Data today. The many novel challenges and opportunities associated with Big Data necessitate rethinking many aspects of these data management platforms, while retaining other desirable aspects. We believe that appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and systems.

We believe that these research problems are not only timely, but also have the potential to create huge economic value in the US economy for years to come. However, they are also hard, requiring us to rethink data analysis systems in fundamental ways. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business.

Challenges and Opportunities with Big Data

1. Introduction

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

Scientific research has been revolutionized by Big Data [CCC2011a]. The Sloan Digital Sky Survey [SDSS2008] has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is an entire discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.

Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance.

It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [CCC2011c], by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates [McK2011] 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) [CCC2011d], energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative [MGI2011]), computational social sciences

(a new methodology fast growing in popularity because of the dramatically lowered cost of obtaining data) [LP+2009], financial systemic risk analysis (through integrated analysis of a web of contracts to find dependencies between financial entities) [FJ+2011], 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 so on.

In 2010, enterprises and users stored more than 13 exabytes of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [PCAST2010] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b].

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modeling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. For example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all be laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it [NYT2012]. We can drastically increase this number by supporting

many levels of engagement with the data, not all requiring deep database expertise. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis.

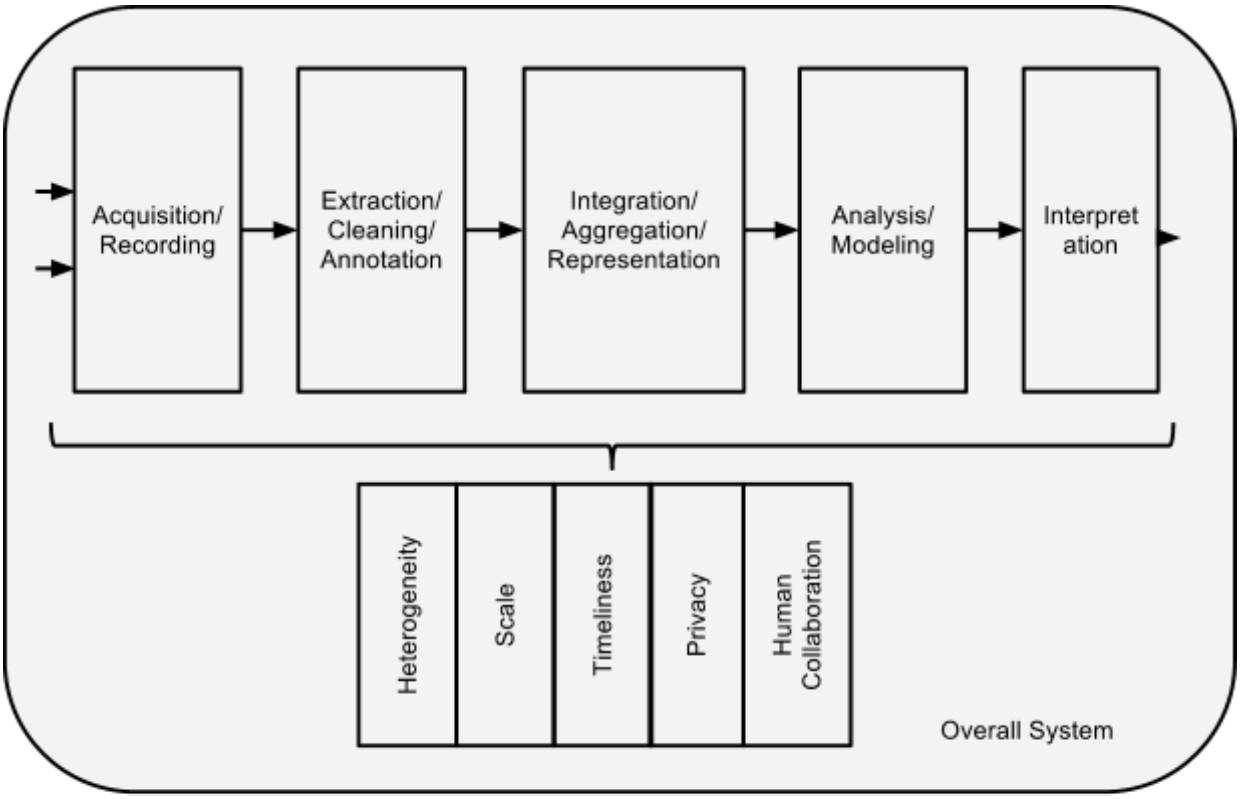

Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Fortunately, existing computational techniques can be applied, either as is or with some extensions, to at least some aspects of the Big Data problem. For example, relational databases rely on the notion of logical data independence: users can think about what they want to compute, while the system (with skilled engineers designing those systems) determines how to compute it efficiently. Similarly, the SQL standard and the relational data model provide a uniform, powerful language to express many query needs and, in principle, allows customers to choose between vendors, increasing competition. The challenge ahead of us is to combine these healthy features of prior systems as we devise novel solutions to the many new challenges of Big Data.

In this paper, we consider each of the boxes in the figure above, and discuss both what has already been done and what challenges remain as we seek to exploit Big Data. We begin by considering

the five stages in the pipeline, then move on to the five cross-cutting challenges, and end with a discussion of the architecture of the overall system that combines all these functions.

2. Phases in the Processing Pipeline

2.1 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 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 petabytes of data today.

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. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artifact that deserves attention? In addition, the data collected by these sensors most often are spatially and temporally correlated (e.g., traffic sensors on the same road segment). We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack. Furthermore, we require "on-line" analysis techniques that can process such streaming data on the fly, since we cannot afford to store first and reduce afterward.

The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. For example, in scientific experiments, considerable detail regarding specific experimental conditions and procedures may be required to be able to interpret the results correctly, and it is important that such metadata be recorded with observational data. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data provenance. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. For example, a processing error at one step can render subsequent analysis useless; with suitable provenance, we can easily identify all subsequent processing that dependent on this step. Thus we need research both into generating suitable metadata and into data systems that carry the provenance of data and its metadata through data analysis pipelines.

2.2 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.

We are used to thinking of Big Data as always telling us the truth, but this is actually far from reality. For example, patients may choose to hide risky behavior and caregivers may sometimes mis-diagnose a condition; patients may also inaccurately recall the name of a drug or even that they ever took it, leading to missing information in (the history portion of) their medical record. Existing work on data cleaning assumes well-recognized constraints on valid data or well-understood error models; for many emerging Big Data domains these do not exist.

2.3 Data Integration, Aggregation, and Representation

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other purposes. Witness, for instance, the tremendous variety in the structure of bioinformatics databases with information regarding substantially similar entities, such as genes. Database design is today an art, and is carefully executed in the enterprise context by highly-paid professionals. We must enable other professionals, such as domain scientists, to create effective database designs, either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so that databases can be used effectively in the absence of intelligent database design.

2.4 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 trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models.

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. As noted previously, real-life medical records have errors, are heterogeneous, and frequently are distributed across multiple systems. The value of Big Data analysis in health care, to take just one example application domain, can only be realized if it can be applied robustly under these difficult conditions. On the flip side, knowledge developed from data can help in correcting errors and removing ambiguity. For example, a physician may write “DVT” as the diagnosis for a patient. This abbreviation is commonly used for both “deep vein thrombosis” and “diverticulitis,” two very different medical conditions. A knowledge-base constructed from related data can use associated symptoms or medications to determine which of two the physician meant.

Big Data is also enabling the next generation of interactive data analysis with real-time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, to populate hot-lists or recommendations, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today.

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. 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.

2.5 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. This

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. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision-maker diligence -- rather than accept the stated solvency of a financial institution at face value, a decision-maker has to examine critically the many assumptions at multiple stages of analysis.

In short, it is rarely enough to provide just the results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data. By studying how best to capture, store, and query provenance, in conjunction with techniques to capture adequate metadata, we can create an infrastructure to provide users with the ability both to interpret analytical results obtained and to repeat the analysis with different assumptions, parameters, or data sets.

Systems with a rich palette of visualizations become important in conveying to the users the results of the queries in a way that is best understood in the particular domain. Whereas early business intelligence systems' users were content with tabular presentations, today's analysts need to pack and present results in powerful visualizations that assist interpretation, and support user collaboration as discussed in Sec. 3.5.

Furthermore, with a few clicks the user should be able to drill down into each piece of data that she sees and understand its provenance, which is a key feature to understanding the data. That is, users need to be able to see not just the results, but also understand why they are seeing those results. However, raw provenance, particularly regarding the phases in the analytics pipeline, is likely to be too technical for many users to grasp completely. One alternative is to enable the users to "play" with the steps in the analysis – make small changes to the pipeline, for example, or modify values for some parameters. The users can then view the results of these incremental changes. By these means, users can develop an intuitive feeling for the analysis and also verify that it performs as expected in corner cases. Accomplishing this requires the system to provide convenient facilities for the user to specify analyses. Declarative specification, discussed in Sec. 4, is one component of such a system.

3. Challenges in Big Data Analysis

Having described the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases. These are shown as five boxes in the second row of Fig. 1.

3.1 Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

Consider an electronic health record database design that has fields for birth date, occupation, and blood type for each patient. What do we do if one or more of these pieces of information is not provided by a patient? Obviously, the health record is still placed in the database, but with the corresponding attribute values being set to NULL. A data analysis that looks to classify patients by, say, occupation, must take into account patients for which this information is not known. Worse, these patients with unknown occupations can be ignored in the analysis only if we have reason to believe that they are otherwise statistically similar to the patients with known occupation for the analysis performed. For example, if unemployed patients are more likely to hide their employment status, analysis results may be skewed in that it considers a more employed population mix than exists, and hence potentially one that has differences in occupation-related health-profiles.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress.

3.2 Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But,

there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

First, over the last five years the processor technology has made a dramatic shift - rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different; for example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node. Furthermore, the move towards packing multiple sockets (each with 10s of cores) adds another level of complexity for intra-node parallelism. Finally, with predictions of "dark silicon", namely that power consideration will likely in the future prohibit us from using all of the hardware in the system continuously, data processing systems will likely have to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how we design, build and operate data processing components.

The second dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters (that are required to deal with the rapid growth in data volumes). This places a premium on declarative approaches to expressing programs, even those doing complex machine learning tasks, since global optimization across multiple users' programs is necessary for good overall performance. Reliance on user-driven program optimizations is likely to lead to poor cluster utilization, since users are unaware of other users' programs. System-driven holistic optimization requires programs to be sufficiently transparent, e.g., as in relational database systems, where declarative query languages are designed with this in mind.

A third dramatic shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, hard disk drives (HDDs) were used to store persistent data. HDDs had far slower random IO performance than sequential IO performance, and data processing engines formatted their data and designed their query processing methods to "work around" this limitation. But, HDDs are increasingly being replaced by solid state drives today, and other technologies such as Phase Change Memory are around the corner. These newer storage technologies do not have the same large spread in performance between the sequential and random I/O performance, which requires a rethinking of how we design storage subsystems for data processing systems. Implications of this changing storage subsystem potentially touch every aspect of data processing, including query processing algorithms, query scheduling, database design, concurrency control methods and recovery methods.

3.3 Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge as described in Sec. 2.1, and a timeliness challenge described next.

There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

3.4 Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

Consider, for example, data gleaned from location-based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address these privacy concerns. An attacker or a (potentially malicious) location-based server can infer the identity of the query source from its (subsequent) location information. For example, a user's location information can be tracked through several stationary connection points (e.g., cell towers). After a while, the user

leaves “a trail of packet crumbs” which could be associated to a certain residence or office location and thereby used to determine the user’s identity. Several other types of surprisingly private information such as health issues (e.g., presence in a cancer treatment center) or religious preferences (e.g., presence in a church) can also be revealed by just observing anonymous users’ movement and usage pattern over time. In general, Barabási et al. showed that there is a close correlation between people’s identities and their movement patterns [Gon2008]. Note that hiding a user location is much more challenging than hiding his/her identity. This is because with location-based services, the location of the user is needed for a successful data access or data collection, while the identity of the user is not necessary.

There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but gets larger and changes over time; none of the prevailing techniques results in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

3.5 Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

In today’s complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. Wikipedia, the online encyclopedia, is perhaps the best known example of crowd-sourced data. We are relying upon information provided by unvetted strangers. Most often, what they say is correct. However, we should expect there to be individuals who have other motives and abilities – some may have a reason to provide false information in an intentional attempt to mislead. While most

such errors will be detected and corrected by others in the crowd, we need technologies to facilitate this. We also need a framework to use in analysis of such crowd-sourced data with conflicting statements. As humans, we can look at reviews of a restaurant, some of which are positive and others critical, and come up with a summary assessment based on which we can decide whether to try eating there. We need computers to be able to do the equivalent. The issues of uncertainty and error become even more pronounced in a specific type of crowd-sourcing, termed participatory-sensing. In this case, every person with a mobile phone can act as a multi-modal sensor collecting various types of data instantaneously (e.g., picture, video, audio, location, time, speed, direction, acceleration). The extra challenge here is the inherent uncertainty of the data collection devices. The fact that collected data are probably spatially and temporally correlated can be exploited to better assess their correctness. When crowd-sourced data is obtained for hire, such as with "Mechanical Turks," much of the data created may be with a primary objective of getting it done quickly rather than correctly. This is yet another error model, which must be planned for explicitly when it applies.

4. System Architecture

Companies today already use, and appreciate the value of, business intelligence. Business data is analyzed for many purposes: a company may perform system log analytics and social media analytics for risk assessment, customer retention, brand management, and so on. Typically, such varied tasks have been handled by separate systems, even if each system includes common steps of information extraction, data cleaning, relational-like processing (joins, group-by, aggregation), statistical and predictive modeling, and appropriate exploration and visualization tools as shown in Fig. 1.

With Big Data, the use of separate systems in this fashion becomes prohibitively expensive given the large size of the data sets. The expense is due not only to the cost of the systems themselves, but also the time to load the data into multiple systems. In consequence, Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads. The challenge here is not to build a system that is ideally suited for all processing tasks. Instead, the need is for the underlying system architecture to be flexible enough that the components built on top of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads. The effects of scale on the physical architecture were considered in Sec 3.2. In this section, we focus on the programmability requirements.

If users are to compose and build complex analytical pipelines over Big Data, it is essential that they have appropriate high-level primitives to specify their needs in such flexible systems. The Map-Reduce framework has been tremendously valuable, but is only a first step. Even declarative languages that exploit it, such as Pig Latin, are at a rather low level when it comes to complex analysis tasks. Similar declarative specifications are required at higher levels to meet the programmability and composition needs of these analysis pipelines. Besides the basic technical need, there is a strong business imperative as well. Businesses typically will outsource Big Data processing, or many aspects of it. Declarative specifications are required to enable technically meaningful service level agreements,

since the point of the out-sourcing is to specify precisely what task will be performed without going into details of how to do it.

Declarative specification is needed not just for the pipeline composition, but also for the individual operations themselves. Each operation (cleaning, extraction, modeling etc.) potentially runs on a very large data set. Furthermore, each operation itself is sufficiently complex that there are many choices and optimizations possible in how it is implemented. In databases, there is considerable work on optimizing individual operations, such as joins. It is well-known that there can be multiple orders of magnitude difference in the cost of two different ways to execute the same query. Fortunately, the user does not have to make this choice – the database system makes it for her. In the case of Big Data, these optimizations may be more complex because not all operations will be I/O intensive as in databases. Some operations may be, but others may be CPU intensive, or a mix. So standard database optimization techniques cannot directly be used. However, it should be possible to develop new techniques for Big Data operations inspired by database techniques.

The very fact that Big Data analysis typically involves multiple phases highlights a challenge that arises routinely in practice: production systems must run complex analytic pipelines, or workflows, at routine intervals, e.g., hourly or daily. New data must be incrementally accounted for, taking into account the results of prior analysis and pre-existing data. And of course, provenance must be preserved, and must include the phases in the analytic pipeline. Current systems offer little to no support for such Big Data pipelines, and this is in itself a challenging objective.

5. 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.

Bibliography

[CCC2011a] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.

[CCC2011b] Advancing Personalized Education. Computing Community Consortium. Spring 2011.

[CCC2011c] Smart Health and Wellbeing. Computing Community Consortium. Spring 2011.

[CCC2011d] A Sustainable Future. Computing Community Consortium. Summer 2011.

[DF2011] Getting Beneath the Veil of Effective Schools: Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.

[Eco2011] Drowning in numbers -- Digital data will flood the planet—and help us understand it better. *The Economist*, Nov 18, 2011. http://www.economist.com/blogs/dailychart/2011/11/big-data-0

[FJ+2011] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.

[Gar2011] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916

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

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

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

[MGI2011] Materials Genome Initiative for Global Competitiveness. National Science and Technology Council. June 2011.

[NPR2011a] Folowing the Breadcrumbs to Big Data Gold. Yuki Noguchi. *National Public Radio*, Nov. 29, 2011. http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data

[NPR2011b] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. *National Public Radio*, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data

[NYT2012] The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[PCAST2010] Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. PCAST Report, Dec. 2010. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf

[SDSS2008] SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. Jan. 2008. Available at http://www.sdss3.org/collaboration/description.pdf

About this Document

This white paper was created through a distributed conversation among many prominent researchers listed below. This conversation lasted a period of approximately three months from Nov. 2011 to Feb. 2012. Collaborative writing was supported by a distributed document editor.

Divyakant Agrawal, UC Santa Barbara
Philip Bernstein, Microsoft
Elisa Bertino, Purdue Univ.
Susan Davidson, Univ. of Pennsylvania
Umeshwar Dayal, HP
Michael Franklin, UC Berkeley
Johannes Gehrke, Cornell Univ.
Laura Haas, IBM
Alon Halevy, Google
Jiawei Han, UIUC
H. V. Jagadish, Univ. of Michigan (Coordinator)
Alexandros Labrinidis, Pittsburgh Univ.
Sam Madden, MIT
Yannis Papakonstantinou, UC San Diego
Jignesh M. Patel, Univ. of Wisconsin
Raghu Ramakrishnan, Yahoo!
Kenneth Ross, Columbia Univ.
Cyrus Shahabi, Univ. of Southern California
Dan Suciu, Univ. of Washington
Shiv Vaithyanathan, IBM
Jennifer Widom, Stanford Univ.