ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Big Data – A Pilot Study on Scope and Challenges

Dona Sarkar[1]
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata - India

Dr. Asoke Nath[2]
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata - India

Abstract: ***Innovations in technology and greater affordability of digital devices have presided over today's Age of Big Data, an umbrella term for the explosion in the quantity and diversity of high frequency digital data. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, and manage, sharing, storage, transfer, visualization, and privacy violations and process data within a tolerable elapsed time. Big data "size" is a constantly moving target. Generally size of the data is Petabytes and Exabyte. Traditional database systems are not able to capture, store and analyze this large amount of data. As the internet is growing, amount of big data continue to grow. Big data analytics provide new ways for businesses and government to analyze unstructured data. Numerous technological innovations are driving the dramatic increase in data and data gathering. This is why big Data is creeping in the World as a new moving target.***

Keywords: Digitization; data deluge: Hadoop; Expansion of Hadoop; heterogeneity;

I. INTRODUCTION

As we move towards an era where the digitization of information is more and more on demand, the overall amount of data takes an exponential growth. The research area Big Data has emerged to precisely tackle the vast amounts of data generated, as well as to investigate the strong societal impacts incurred by the explosion of data in the society. A concept refers to data that is so large in volume, moving at an unforeseen velocity, with such a high variation in structure and very often veracious in nature that - to be fully exploited, explored and to derive its value - the development of new techniques and systems is required. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data. Big Data is a concept that refers to oceanic volume of data, moving at an unforeseen velocity, with such a high variation in structure and very often veracious in nature that - to be fully exploited, explored and to derive its value. These '**4V**'s are the essential dimensions of Big data. They're a helpful lens through which to view and understand the nature of the data and the software platforms available to exploit them. The main characteristics of these 4Vs are illustrated as follows:

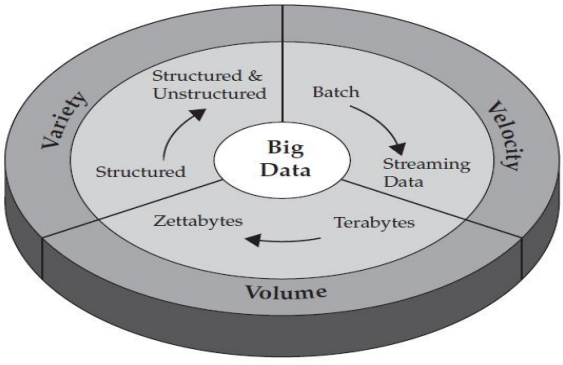


Figure 1-1 *IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.*

» ***Velocity***

Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. The importance of data's velocity — the increasing rate at which data flows into an organization — has followed a similar pattern to that of volume.

» ***Volume***

Big data implies enormous volumes of data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive. This volume presents the most immediate challenge to conventional IT structures. It calls for scalable storage, and a distributed approach to querying.

» ***Variety***

Variety refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDF, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data.

» ***Veracity***

Big Data veracity refers to the biases, noise and abnormality in data. Veracity refers to the messiness or trustworthiness of the data. Veracity in big data analysis is the biggest challenge when compares to things like volume and velocity. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems. The volumes often make up for the lack of quality or accuracy.

Another important characteristic of the Bigdata is ***Complexity*** - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

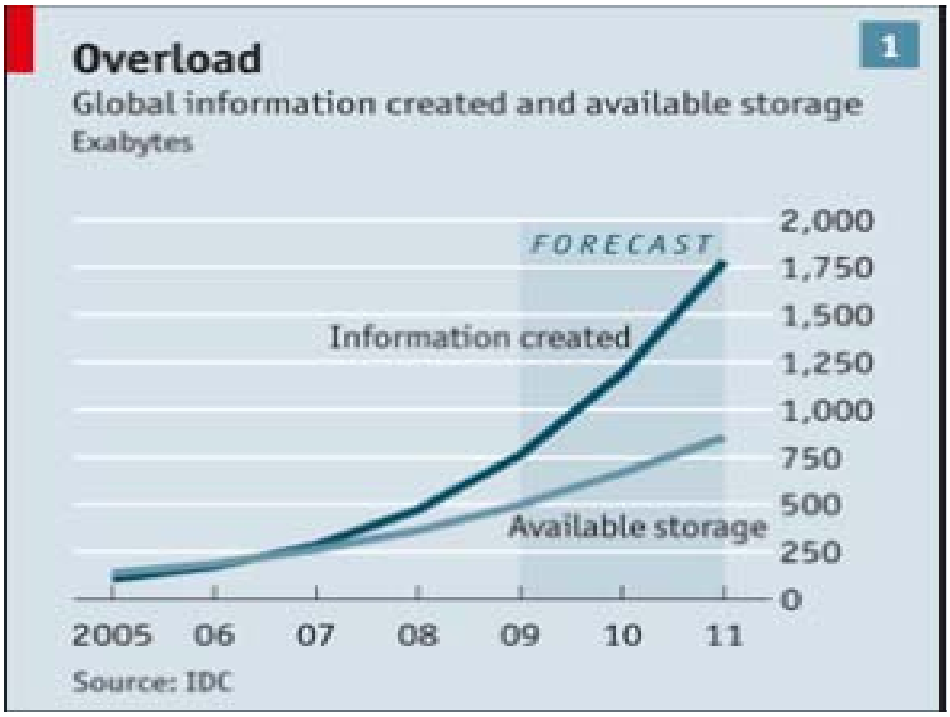
II. WHY BIG DATA ?

Comparison between Traditional analytics & big data analytics

	Traditional Data Analytics	**Big data Analytics**
Data features	» Environment suitable for structured data only. » Usual unit of volume is megabyte/gigabyte.	» Environment suitable for any data structured, semi-/multi/unstructured data from multiple sources. » Usual unit of volume is terabyte or petabyte.
Population of analysis and questions to ask	» Sample data analysis of Known populations. » Answering questions we know that we don't know.	» Non-sample data analysis of unknown populations. » Answering questions we don't know we don't know.

Technologies	» SQL approach to data. » Relational database (data to function model) » No open source. » Batch processing (offline) of "historical," static data.	» Massively parallel processing and NoSQL approach to data, but almost SQL compliant. » Hadoop framework (function to data model). » Open source. » Stream processing (online) of (near) real time, live data.
Research & Development	» Individual manufacturer/ developer can work independently	» Nobody works alone; all related parties must work together

III. The Data Revolution



Source: "The Leaky Corporation." The Economist. http://www.economist.com/node/18226961.

The world is experiencing a data revolution, or *"data deluge*" . Whereas in previous generations, a relatively small volume of analog data was produced and made available through a limited number of channels, today a massive amount of data is regularly being generated and flowing from various sources, through different channels, every minute in today's Digital Age.. It is the speed and frequency with which data is emitted and transmitted on the one hand, and the rise in the number and variety of sources from which it emanates on the other hand, that jointly constitute the data deluge. The amount of available digital data at the global level grew from 150 Exabyte in 2005to 1200 Exabyte in 2010. It is projected to increase by 40% annually in the next few years, which is about 40 times the much-debated growth of the world's population. This rate of growth means that the stock of digital data is expected to increase 44 times between 2007 and 2020, doubling every 20 months.

Challenges in Big Data:

Big data analytics faces different challenges. These are described as follows:

» **Heterogeneity and Incompleteness**

Machine analysis algorithms expect homogeneous data, and cannot understand nuance. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain.

» **Timeliness**

There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements init that meet a specified criterion The larger the data set to be processed, the longer it will take to analyze. It is difficult to design a structure when data is growing in very high speed.

» **Human Collaboration**

A Big Data analysis system must support input from multiple human experts, and shared exploration of results.

» **Privacy and security**

This is another big challenge preserving individual privacy. For example in the healthcare industry, record of individual is very personal. But it can be available from multiple sources. So, it is difficult to maintain privacy and security.

» **Data Quality**

A large volume of data is processed. Analyzing which data is important and to capture it is a big challenge.

» **Analysis**

Big data is coming from various data sources. So analytics is a challenge.

» **Skill**

Big data require people with new skill sets. Managing big data effectively requires the right people.

Scope of Big Data:

» **Access to and impact of Big Data**

Big Data is increasingly used as an umbrella term that includes the connected societal, ethical and legal issues that are arising as ever increasing amounts of data are collected and analyzed by society. This theme also touches upon the industrial and economic drivers and impacts.

» **Making Big Data really work**

Research into the framework of how to use heterogeneous and complex data to real benefit in a wide range of problems spanning industry to the social sciences.

» **Big Data in Imaging**

Many application areas produce datasets that need to be visualized to be useful. Mathematical imaging offers rigorous and efficient methodologies for retrieving meaningful information from (Big) Data – covering noise removal, segmentation, recognition and imprinting. It also enables useful approaches such as Visual Analytics to become possible.

» **Big Data Science and Infrastructure**

Vast amount of data is mined, processed, data warehoused and analyzed. All this is done to achieve a goal of making the best use of data. When the data sets were small and the various permutations lower, it was easier for it to be made available in a spread sheet or specialized report. These reports are usually fixed and very inflexible. However, these also had limitations of the depth of parameters to be mined or relationship to be established

» **Scalable Data Infrastructure:**

Another unique characteristic of big data is that, unlike large data sets that have historically been stored and analyzed, often through data warehousing, big data is made up of discretely small, incremental data elements with real-time additions or modifications. It does not work well in traditional, online transaction processing data stores or with traditional SQL analysis tools. Big data requires a flat, horizontally scalable database, often with unique query tools that work in real time with actual data.

» **Big Data Search and Mining**

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

Need of Big Data:

"Big Data" is changing our world. This revolution is being driven by many factors:

» A proliferation of sensors

» More generally, the creation of almost all information in digital form

» Dramatic cost reductions in storage and scalability improvements in computation

» Dramatic algorithmic breakthroughs in machine learning and other areas

» Abundance of computing & storage of generated data (estimated 8ZB)

» More data provides greater value whereas; the traditional data doesn't scale well.

» Increase of storage capacities, increase of processing power, rapid availability of data

IV. BIG DATA RESEARCHES

In recent years, "Big Data" has become a new ubiquitous term. Big Data is transforming science, engineering, medicine, healthcare, finance, business, and ultimately society itself. The IEEE Big Data has established itself as the top tier research conference in Big Data. The first conference IEEE Big Data 2013was held in Santa Clara, CA from Oct 6-7, 2013. 259 paper submissions for the main conference and 32 paper submissions for the industry and government program. Of those, 44 regular papers and 53 short papers were accepted. Also, there were 14 workshops associated with IEEE Big Data 2013 covering various important topics related to various aspects of Big Data research, development and applications, and more than 400 participants from 40 countries attend the 4-day event. The recent availability of huge amounts of data, along with advanced tools of exploratory data analysis, data mining/machine learning and data visualization, offers a whole new way of understanding the world. In order to exploit these huge volumes of data, new techniques and technologies are needed**.**

Big Data in Defense and Security:

Intelligence and Operations

For many intelligence experts, automated analysis technology is the top intelligence, surveillance and reconnaissance (ISR) priority. The UK Maritime Intelligence Fusion Centre recently highlighted the imbalance between investment in collectors and in the tools to support its analysis, rendering analysts incapable of taking into account all available sources when performing

their assessment**.** Similarly in the field of cyber-security, where network managers can be dealing with millions of attacks every day, Big Data analytics are being applied to spot advanced persistent threats – such as socially engineered attacks designed to steal corporate intellectual property or government information – above the ever-growing background noise of everyday nuisance or opportunistic attacks. Most hackers have a modus operandi, which once identified can be used to predict the form of future attacks and put appropriate defensive measures in place.

V. BIG DATA AND HADOOP

Apache **Hadoop** is a fast-growing big-data processing platform defined as "an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. Developed by Doug Cutting, Cloudera's Chief Architect and the Chairman of the Apache Software Foundation, Apache Hadoop was born out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it. Hadoop can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email - regardless of its native format. Even when different types of data have been stored in unrelated systems, it is possible to store it all into Hadoop cluster with no prior need for a schema. The Hadoop ecosystem is evolving and becoming the \standard" technology for big data analysis

Why Hadoop?

Social media/web is unstructured. Amount of data is immense. New data source arise quickly.

Apache Hadoop has two main subprojects:

» MapReduce - The framework that understands and assigns work to the nodes in a cluster. Has been defined by Google in 2004 and is able to distribute data workloads across thousands of nodes. It is based on "break problem up into smaller sub-problems" strategy and can be exposed via SQL and in SQL-based BI tools

» Hadoop Distributed File System (HDFS) - An Apache open source distributed file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big. it is known for highly scalable storage and automatic data replication across three nodes for fault tolerance.

Hadoop changes the economics and the dynamics of large scale computing, having a remarkable influence based on four salient characteristics. Hadoop enables a computing solution that is:

» **Scalable:** New nodes can be added as needed and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.

» **Cost effective:** Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.

» **Flexible:** Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.

» **Fault tolerant:** When you lose a node, the system redirects work to another location of the data and continues processing without missing a beat.

VI. TECHNOLOGY OUTLOOK

The following section discusses some of the Big Data technologies along with the projected time frame for mainstream adoption.

Hadoop MapReduce and Hadoop Distributed File System (HDFS)

Hadoop is a framework that provides open source libraries for distributed computing using MapReduce software and its own distributed file system, simply known as the Hadoop Distributed File System (HDFS). It is designed to scale out from a few computing nodes to thousands of machines, each offering local computation and storage. One of Hadoop's main value propositions is that it is designed to run on commodity hardware such as commodity servers or personal computers, and has high tolerance for hardware failure. In Hadoop, hardware failure is treated as a rule rather than an exception.

Technologies used for Big data Analytics:

HDFS

The HDFS is a fault-tolerant storage system that can store huge amounts of information, scale up incrementally and survive storage failure without losing data. Hadoop clusters are built with inexpensive computers. If one computer (or node) fails, the cluster can continue to operate without losing data or interrupting work by simply re-distributing the work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking files into small blocks and storing duplicated copies of them across the pool of nodes. The figure below illustrates how a data set is typically stored across a cluster of five nodes. In this example, the entire data set will still be available even if two of the servers have failed. Compared to other redundancy techniques, including the strategies employed by Redundant Array of Independent Disks (RAID) machines, HDFS offers two key advantages. Firstly, HDFS requires no special hardware as it can be built from common hardware. Secondly, it enables an efficient technique of data processing in the form of Map Reduce.

MapReduce35

Map Reduce is a programming model and software framework first developed by Google. It works like a UNIX pipeline. Most enterprise data management tools (database management systems) are designed to make simple queries run quickly. Typically, the data is indexed so that only small portions of the data need to be examined in order to answer a query. This solution, however, does not work for data that cannot be indexed. To answer a query in this case, all the data has to be examined. Hadoop uses the Map Reduce technique to carry out this exhaustive analysis quickly. Map Reduce is a data processing algorithm that uses a parallel programming implementation. In simple terms, Map Reduce is a programming paradigm that involves distributing a task across multiple nodes running a "map" function. The map function takes the problem, splits it into sub-parts and sends them to different machines so that all the sub-parts can run concurrently. The results from the parallel map functions are collected and distributed to a set of servers running "reduce" functions, which then takes the results from the sub-parts and re-combines them to get the single answer. A Map Reduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result. Map Reduce framework consists of a single master Job Tracker and one slave Task Tracker per cluster node. The master is responsible for scheduling the job's component tasks on the slave, re-executing the failed task. The slave executes the task as directed by the master.

NoSQL

NoSQL databases are highly scalable, non-relational databases (such as columnar, document, key-value, object and graph databases) designed to handle large volumes of data, particularly in applications requiring near real time processing. A **NoSQL** (often interpreted as **N**ot **O**nly **SQL**[1][2]) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling and finer control over availability. The data structure (e.g. key-value, graph, or document) differs from the RDBMS, and therefore some operations are faster in NoSQL and some in RDBMS. There are differences though, and the particular suitability of a given NoSQL DB depends on the problem it must solve (e.g., does the solution use graph algorithms?).

NoSQL databases are increasingly used in big data and real-time web applications.[3] NoSQL systems are also called "Not only SQL" to emphasize that they may also support SQL-like query languages. Many NoSQL stores compromise consistency in favor of availability and partition tolerance. Many NoSQL databases have excellent integrated caching capabilities. So,the frequently used data is kept in system memory. NoSQL database types are:

1. ***Document database***: pair each key with complex data structure known as document. Document may contain nested document. This type of database store documents which are usually hierarchal in nature.
2. ***Graph stores***: Graph database is based on graph theory. It is used to store information about network.
3. ***Key value stores***: Every single item is stored as an attribute name together with its value.
4. ***Wide column stores***: They are optimized for queries over large datasets and store column of data together instead of rows.

Cloud Computing:

The rise of cloud computing and cloud data stores has been a precursor and facilitator to the emergence of big data. Cloud computing is the commoditization of computing time and data storage by means of standardized technologies.

It has significant advantages over traditional physical deployments. However, cloud platforms come in several forms and sometimes have to be integrated with traditional architectures.

Cloud computing employs visualization of computing resources to run numerous standardized virtual servers on the same physical machine. Cloud providers achieve with this economies of scale, which permit low prices and billing based on small time intervals, e.g. hourly.

This standardization makes it an elastic and highly available option for computing needs. The availability is not obtained by spending resources to guarantee reliability of a single instance but by their interchangeability and a limitless pool of replacements. This impacts design decisions and requires dealing with instance failure gracefully.

Private Cloud

Private clouds are dedicated to one organization and do not share physical resources. The resource can be provided in-house or externally. A typical underlying requirement of private cloud deployments are security requirements and regulations that need a strict separation of an organization’s data storage and processing from accidental or malicious access through shared resources. Private cloud setups are challenging since the economical advantages of scale are usually not achievable within most projects and organizations despite the utilization of industry standards

Another reason for private cloud deployments are legacy systems with special hardware needs or exceptional resource demand, e.g. extreme memory or computing instances which are not available in public clouds. These are valid concerns however if these demands are extraordinary the question if a cloud architecture is the correct solution has to be raised. One reason can be to establish a private cloud for a transition period to run legacy and demanding systems in parallel while their services are ported to a cloud environment culminating in a switch to a cheaper public or hybrid cloud.

Public Cloud

Public clouds share physical resources for data transfers, storage, and processing. However, customers have private visualized computing environments and isolated storage. Security concerns, which entice a few to adopt private clouds or custom deployments, are for the vast majority of customers and projects irrelevant. Visualization makes access to other customers’ data extremely difficult.

The copying of data out to local systems or other providers is often more expensive. This is not an insurmountable problem and in practice encourages utilizing more services from a cloud provider instead of moving data in and out for different services or processes.

The available resources for each customer on a physical machine are usually throttled to ensure that each customer receives a guaranteed level of performance. Larger resources generally deliver very predictable performance since they are much closer aligned with the physical instance's performance. Horizontally scaling projects with small instance should not rely on an exact performance of each instance but be adaptive and focus on the average performance required and scale according to need.

Hybrid Cloud

The hybrid cloud architecture merges private and public cloud deployments. This is often an attempt to achieve security and elasticity, or provide cheaper base load and burst capabilities. Some organizations experience short periods of extremely high loads, e.g. as a result of seasonality like black Friday for retail, or marketing events like sponsoring a popular TV event. These events can have huge economic impact to organizations if they are serviced poorly.

The hybrid cloud provides the opportunity to serve the base load with in-house services and rent for a short period a multiple of the resources to service the extreme demand. This requires a great deal of operational ability in the organization to seamlessly scale between the private and public cloud. Tools for hybrid or private cloud deployments exist like Eucalyptus for Amazon Web Services. On the long-term the additional expense of the hybrid approach often is not justifiable since cloud providers offer major discounts for multi-year commitments. This makes moving base load services to the public cloud attractive since it is accompanied by a simpler deployment strategy.

Expansion of Hadoop:

Hadoop creates *clusters* of machines and coordinates work among them. If any of the clusters fails, then Hadoop continues to operate the cluster without losing data.

Some of the Hadoop related projects are described as:

***Pig*:** It is a Scripting language and run time environment. It allows users to execute MapReduce on a Hadoop cluster. Pig's language layer currently consists of a textual language called Pig Latin.

Hive: It provides SQL access for data in HDFS. Hive's query language, HiveQL, compiles to MapReduce. It also allows user-defined functions.

***HBase**:* A scalable, distributed database that supports structured data storage for large tables. It is column based rather than row based.

***Mahout*:** Library of machine learning and data mining algorithm. It has four types of algorithm.

Oozie: Oozie is a Java Web-Application that runs in a Java servlet-container – Tomcat. It is job coordinator and workflow manager.

BigTop: It is used for for packaging and testing the Hadoop ecosystem.

Opportunities in Big Data:

» The use of big data will become a key basis of competition and growth for individual firms. All the companies will use big data. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value from deep and up-to-real-time information.

» Big data has opportunities in the field of education. More detailed information for school can be generated. This is beneficial for teacher and parents

» Almost all sectors like computer and electronic products, insurance, and government will increase their productivity from the use of big data

» Concept of big data has practical application in the area of healthcare research. In health care, data is coming from medical records, radiology images, human genetics etc. More information is analyzed regarding patient care and disease. Hence studies can be completed faster. Big data will help better future diagnoses and treatment of the patient.

» Use of smart phone and tablet leads to high amount of mobile data traffic. Big Data is important for mobile networks. It is useful to improve network quality, traffic planning, prediction of hardware maintenance etc

» Various branches of science generates large amount of experimental data. Fulfilling the demands of science requires a new way of handling data

Future Scope of Criteria:

Big Data is a sea change that, like nanotechnology and quantum computing, will shape the twenty-first century. According to some experts, *"[by] employing massive data mining, science can be pushed towards a new methodological paradigm which will transcend the boundaries between theory and experiment."*

Another perspective frames this new ability to unveil stylized facts from large datasets as "*the fourth paradigm of science*".

Big Data constitutes an historic opportunity to advance our common ability to support and protect human communities by understanding the information they increasingly produce in digital forms.

If we ask how much development work will be transformed in 5 to 10 years as Big Data expands into the field, the answer is not straightforward. Big Data will affect development work somewhere between significantly and radically, but the exact nature and magnitude of the change to come is difficult to project.

» Because the new types of data that people will produce in ten years is unknown.

» Because the same uncertainty holds for computing capacities, given that Moore's Law with certainly not hold in an era of quantum computing.

» Because a great deal will depend on the future strategic decisions taken by a myriad of actors—chief of which are policymakers.

VII. Limitations of Big Data

Many open questions remain—including the potential misuse of Big Data, because information is power. If, however, we ask how Big Data for Development can fulfill its immense potential to enhance the greater good, then the answer is clearer. What is needed is both *intent* and *capacity* to be sustained and strengthened, on the basis of a full recognition of the opportunities and challenges.

Specifically, its success hinges on two main factors.

1. The level of institutional and financial support from public sector actors, and the willingness of private corporations and academic teams to collaborate with them, are including by sharing data and technology and analytical tools.

2. The development and implementation of new norms and ontology for the responsible use and sharing of Big Data for Development, backed by a new institutional architecture and new types of partnerships

VIII. Conclusion And Future Scope

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many

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

References

1. www.cloudera.com/content/cloudera/en/.../hadoop-and-big-data.html
2. http://en.wikipedia.org/wiki/Big_data
3. www.planet-data.eu/sites/default/files/.../Big_Data_Tutorial_part4.pdf
4. www.j2eebrain.com/java-J2ee-hadoop-advantages-and-disadvantages.html
5. www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf
6. www.purdue.edu/discoverypark/cyber/.../pdfs/BigDataWhitePaper.pdf
7. https://www.rgpv.ac.in/iccbdt/papers/34.pdf
8. http://web.stanford.edu/~jdlevin/Papers/BigData.pdf
9. www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf
10. http://en.wikipedia.org/wiki/Apache_Hadoop.
11. www.ibm.com/software/data/infosphere/hadoop/
12. http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
13. http://radar.oreilly.com/2013/10/dealing-with-data-in-the-hadoop-ecosystem.html

AUTHOR(S) PROFILE

Miss Dona Sarkar is a student pursuing her M.Sc course in the Department of Computer Science, St. Xavier's College (Autonomous), Kolkata. She is currently working on the project in the improvisation of a renowned algorithm in the field of Cryptography under the supervision of Dr. Asoke Nath.

Dr. Asoke Nath is Associate Professor in the Department of Computer Science, St. Xavier's College (Autonomous), Kolkata. His major research areas are Cryptography and Network Security, Steganography, Green Computing, MOOCs, e-learning Methodologies, Artificial Neural networks. He has published more than 104 papers in various International Journals and conference proceedings.

