Big Data Manipulation- A new concern to the ICT world
(A massive Survey/statistics along with the necessity)

Syed Jamaluddin Ahmad, Roksana Khandoker Jolly

Abstract— **Big Data is a new concept in the global arena. Data creates values in the economy and very parts of life. When we do some things, it creates some data. In the early history of computing data is a valuable thing to develop some new technique or new idea generation.**

In the beginning, there was data – first in file systems, and later in databases as the need for enterprise data management emerged. In 1970 with the rules of the relational model, began to gain commercial traction in the early 1980's. As for "Big Data", at least beyond the accumulation of scientific data, the need to manage "large" data volumes came later, first impacting the database world and then more recently impacting the systems community in a big way. 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.

Turning Big Data—call logs, mobile-banking transactions, online user-generated content such as blog posts and Tweets, online searches, satellite images, etc.—into actionable information requires using computational techniques to unveil trends and patterns within and between these extremely large socioeconomic datasets. New insights gleaned from such data mining should complement official statistics, survey data, and information generated by Early Warning Systems, adding depth and nuances on human behaviours and experiences—and doing so in real time, thereby narrowing both information and time gaps. Data have become a torrent flowing into every area of the global economy. Companies churn out a burgeoning volume of transactional data, capturing trillions of bytes of information about their customers, suppliers, and operations. millions of networked sensors are being embedded in the physical world in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data in the age of the Internet of Things.

Big Data has important, distinct qualities that differentiate it from conventional source data. The data from these innovative sources are highly distributed, loosely structured, large in volume, and often available in real-time. Big Data is also an important part of the data revolution as referenced in the recommendations made to the Secretary General by the High Level Panel of Eminent Persons on the post 2015 development agenda in their report "A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development".

Better data and statistics will help governments track progress and make sure decisions are evidence based; they can also strengthen accountability. Mobile devices, sensors, tracking devices and other technologies have caused a fundamental change to the availability of source data.

Digital data is now everywhere—in every sector, in every economy, in every organization and user of digital technology. While this topic might once have concerned only a few data geeks, big data is now relevant for leaders across every sector, and consumers of products and services stand to benefit from its application. The ability to store, aggregate, and combine data and then use the results to perform deep analyses has become ever more accessible as trends such as Moore's Law in computing, its equivalent in digital storage, and cloud computing continue to lower costs and other technology barriers.

A true data revolution would draw on existing and new sources of data to fully integrate statistics into decision making, promote open access to, and use of, data and ensure increased support for Analytic systems". "Big Data" being a hot topics in the field of data mining, varies seminars, symposiums and workshops are being arranged in Bangladesh, so that the section dealing with data analysis can have a better idea about the sources of data generation the huge volumes, the results that may be derived and the threats lying in Big Data analysis. Bangladesh Bureau of Statistics(BBS) being a sloe organization mandated for official Statistics, huge volume of data is generated every year, so a large section of officials is engaged in data analysis, as a result a workshop on "Big Data" was held to give some idea of 'what', 'how', and 'where' about " Big Data" . Likewise a workshop on Advance Data Management(ADM) including 'Big Data" was held

in BUET on June 28 and 29' 2013 in which several research papers from home and abroad were presented in this context.

Our Thesis paper does not offer a grand theory of technology-driven social change in the Big Data era, rather it aims to highlight the main development and uses raised by "Big Data" management. This thesis paper covers three main issues: Big Data sources, Main challenges, and Areas of use.

Index Terms— **Hadoop, API, SAS, HDFS, SME,SID.**

Syed Jamaluddin Ahmad, Assistant Professor & Chairman, Department of Computer Science & Engineering, University of South Asia, City: Dhaka, Country: Bangladesh, Mobile No.: +8801633628612

Roksana Khandoker Jolly, Senior Lecturer, Department of Computer Science & Engineering, University of South Asia, City: Dhaka, Country: Bangladesh, Mobile No.: +8801737157856

Analytics :	Information resulting from the systematic analysis of data or statistics1.
Big data:	Data that is complex in terms of volume, variety, velocity and/or its relation to other data, which makes it hard to handle using tradition database management or tools.
Big data analytics:	Refers to analysis techniques operated on data sets classified as "big data".
Cloud:	The presence of IT services such as computing power and storage as a service accessible via a network such as the Internet.
Data governance:	A technique to manage data within an organization efficiently and effectively ("Jemoet met geweld de board in!," 2012)
Hadoop:	An open-source analytics toolset that supports running data-intensive applications on different nodes.
MapReduce :	A model, mostly known as a part of Hadoop, used to distribute the processing of a large dataset across different nodes by using map and reduce jobs.
Multi-structured:	Since data is often somehow structured, the term unstructured is misleading in these cases. Therefore multi-structured is a better term, referring to content not having a fixed structure. Terms like semi-structured and "grey data" are also referring to this.
Node:	A node refers to a (virtual) terminal (or computer machine) in a computer network.
SAS	SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market.

ABBREVIATIONS

ACID	Atomicity, consistency, isolation and durability: a set of properties guaranteeing basic functionalities of most databases.
API	Application Programming Interface: a programmed specification to enable easy communication with other software components.
A2I	Access to Information (A Program under Prime Minister's Office of Bangladesh)
BASE	Basic availability, soft-state and eventually consistency: a successor and looser version of ACID making horizontal scaling more feasible.
BBS	Bangladesh Bureau of Statistics
BI	Business Intelligence: analyzing and combining data in order to create knowledge which helps the organization to create and exploit opportunities.
CRM	Customer Relationship Management: managing organization's interactions with customers, clients and sales prospects.
EDW	Enterprise Data Warehouse: a centralized database used for reporting and analysis.
ERP	Enterprise Resources Planning
HDFS	Hadoop Distributed File System: part of the Hadoop toolset making distributed storage of data possible.
IoT	Internet of Things: the phenomenon of connecting devices to a global network such as the Internet resulting in a interconnected digitally world.
IMDB	In-memory database: a database management system that primarily relies on main memory (e.g. RAM) to execute processing tasks at very high speed.
JSON	Java script Object Notation: mostly used to exchange data between web applications.
MIS	Management Information System: provides information needed to manage an organization efficiently and effectively. Examples are enterprise resources planning (ERP) and customer relationship management (CRM) systems.
MPP	Massively Parallel Processing: processing tasks by using different nodes (distributed computing).
NoSQL	Not only SQL: a new generation of horizontal scalable databases often compliant to the BASE rule set and often capable to handle unstructured and multi-structured data.
SME	Small and medium-sized enterprises.
SID	Statistics & Informatics Division

I. INTRODUCTION

The year is 2012, and everyone everywhere is buzzing about "Big Data". Virtually everyone, ranging from big Web companies to traditional enterprises to physical science researchers to social scientists, is either already experiencing or anticipating unprecedented growth for data available in their world, as well as new opportunities and great untapped value that successfully taming the "Big Data" beast will hold [9]. It is almost impossible to pick up an issue of anything from the trade press [8, 34], or even the popular press [53, 18, 28], without hearing something about "Big Data". Clearly it's a new era! Or is it...? The database community has been all about "Big Data" since its inception, although the meaning of "Big" has obviously changed a great deal since the early 1980's when the work on parallel databases as we know them today was getting underway. Work in the database community continued until "shared nothing" parallel database systems were deployed commercially and fairly widely accepted in the mid-1990's.

Most researchers in the database community then moved on to other problems. "Big Data" was reborn in the 2000's, with massive, Web-driven challenges of scale driving system developers at companies such as Google, Yahoo!, Amazon, Face- book, and others to develop new architectures for storing, accessing, and analyzing "Big Data". This rebirth caught most of the database community napping with respect to parallelism, but now the database community has new energy and is starting to bring its expertise in storage, indexing, declarative languages, and set oriented processing to bear on the problems of "Big Data" analysis and management. In this paper we review the history of systems for managing "Big Data" as well as today's activities and architectures from the (perhaps biased) perspective of three "database guys" who have been watching this space for a number of years and are currently working together on "Big Data" problems. The remainder of this paper is organized as follows. In Section 2, we briefly review the history of systems for managing "Big Data" in two worlds, the older world of databases and the newer world of systems built for handling Web scale data. Section 3 examines systems from both worlds from an architectural perspective, looking at the components and layers that have been developed in each world and the

roles they play in "Big Data" management. Section 4 then argues for rethinking the layers by providing an overview of the approach being taken in the ASTERIX project at UC Irvine as well as touching on some related work elsewhere. Section 5 presents our views on what a few of the key open questions are today as well as on how the emerging data intensive computing community might best go about tackling them effectively.

- **Big data creates value in several ways**

We have identified five broadly applicable ways to leverage big data that offer transformational potential to create value and have implications for how organizations will have to be designed, organized, and managed. For example, in a world in which large-scale experimentation is possible, how will corporate marketing functions and activities have to evolve? How will business processes change, and how will companies value and leverage their assets (particularly data assets)? Could a company's access to, and ability to analyze, data potentially confer more value than a brand? What existing business models are likely to be disrupted?

For example, what happens to industries predicated on information asymmetry—e.g., various types of brokers—in a world of radical data transparency? How will incumbents tied to legacy business models and infrastructures compete with agile new attackers that are able to quickly process and take advantage of detailed consumer data that is rapidly becoming available, e.g., What they say in social media or what sensors report they are doing in the world? In addition, what happens when surplus starts shifting from suppliers to customers, as they become empowered by their own access to data, e.g., comparisons of prices and quality across competitors?

- **Creating transparency**

Simply making big data more easily accessible to relevant stakeholders in a timely manner can create tremendous value. In the public sector, for example, making relevant data more readily accessible across otherwise separated departments can sharply reduce search and processing time. In manufacturing, integrating data from R&D, engineering, and manufacturing units to enable concurrent engineering can significantly cut time to market and improve quality.

- **Replacing/supporting human decision making with automated algorithms**

Sophisticated analytics can substantially improve decision-making, minimize risks, and unearth valuable insights that would otherwise remain hidden. Such analytics have applications for organizations from tax agencies that can use automated risk engines to flag candidates for further examination to retailers that can use algorithms to optimize decision processes such as the automatic fine-tuning of inventories and pricing in response to real-time in-store and online sales. In some cases, decisions will not necessarily be automated but augmented by analyzing huge, entire datasets using big data techniques and technologies rather than just smaller samples that individuals with spreadsheets can handle and understand. Decision-making may never be the same; some organizations are already making better decisions by analyzing entire datasets from customers, employees, or even sensors embedded in products.

- **Innovating new business models, products, and services**

Big data enables companies to create new products and services, enhance existing ones, and invent entirely new business models. Manufacturers are using data obtained from the use of actual products to improve the development of the next generation of products and to create innovative after-sales service offerings. The emergence of real-time location data has created an entirely new set
of location-6 based services from navigation to pricing property and casualty insurance based on where, and how, people drive their cars.

- **Use of big-data will become a key basis of competition and growth and growth for individual firms**

The use of big data is becoming a key way for leading companies to outperform their peers. For example, we estimate that a retailer embracing big data has the potential to increase its operating margin by more than 60 percent Big data will also help to create new growth opportunities and entirely new categories of companies, such as those that aggregate and analyze industry data. Many of these will be companies that sit in the middle of large information flows where data about products and services, buyers and suppliers, and consumer preferences and intent can be captured and analyzed

- **How do we measure the value of big data?**

When we set out to size the potential of big data to create value, we considered only those actions that essentially depend on the use of big data—i.e., actions where the use of big data is necessary (but usually not sufficient) to execute a particular lever. We did not include the value of levers that consist only of automation but do not involve big data (e.g., productivity increases from replacing bank tellers with ATMs). Note also that we include the gross value of levers that require the use of big data. We did not attempt to estimate big data's relative contribution to the value generated by a particular lever but rather estimated the total value created.

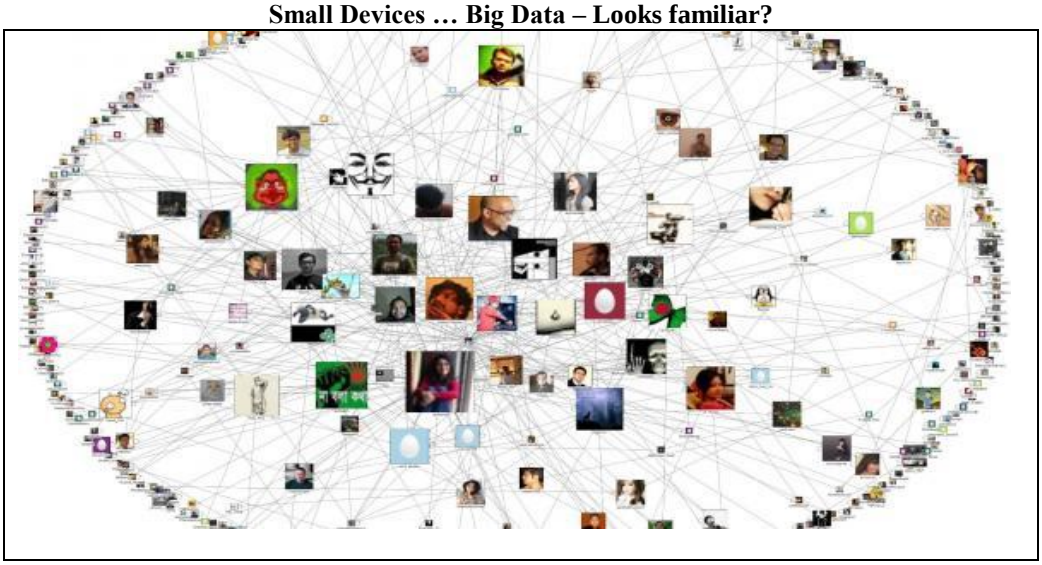
Exhibit 1Big data can generate significant financial value across sectors

		
US health care▪$300 billion value per year▪~0.7 percent annual productivity growth	**Europe public sector administration▪€250 billion value per year▪~0.5 percent annual productivity growth**	**Global personal location data▪**$100 billion+ revenue for service providers▪Up to $700 billion value to end users

l**US retail**▪60+% increase in net margin possible▪0.5–1.0 percent annual productivity growth

Manufacturing Up to 50 percent decrease in product development, assembly costs Up to 7 percent reduction working capital

Small Devices … Big Data – Looks familiar?



II. BIG DATA IN THE DATABASE WORLD

In the database world, a.k.a. the enterprise data management world, "Big Data" problems arose when enterprises identified a need to create data warehouses to house their historical business data and to run large relational queries over that data for business analysis and reporting purposes. Early work on support for storage and efficient analysis of such data led to research in the late 1970's on "database machines" that could be dedicated to such purposes.

Early database machine proposals involved a mix of novel hardware architectures and designs for prehistoric parallel query processing techniques [37]. Within a few years it became clear that neither brute force scan-based parallelism nor proprietary hardware would become sensible substitutes for good software data structures and algorithms. This realization, in the early 1980's, led to the first generation of software-based parallel databases based on the Architecture now commonly referred to as "shared-nothing" [26]. The architecture of a shared-nothing parallel database system, as the name implies, is based on the use of a networked cluster of individual machines each with their own private processors, main memories, and disks.

All inter-machine coordination and data communication is accomplished via message passing. Notable first-generation parallel database systems included the Gamma system from the University of Wisconsin [27], the GRACE system from the University of Tokyo [29], and the Teradata system [44], the first successful commercial parallel database system (and still arguably the industry leader nearly thirty years later).

These systems exploited the declarative, set-oriented nature of relational query languages And pioneered the use of divide-and-conquer parallelism based on hashing in order to partition data for storage as well as relational operator execution for query processing. A number of other relational database vendors, including IBM [13], successfully created products based on this architecture, and the last few years have seen a new generation of such systems (e.g., Netezza, Aster Data, Datallegro, Greenplum, Vertica, and ParAccel). Major hardware/software vendors have recently acquired many of these new systems for impressively large sums of money, presumably driven in part by "Big Data" fever. So what makes "Big Data" big, i.e., just how big is "Big"?

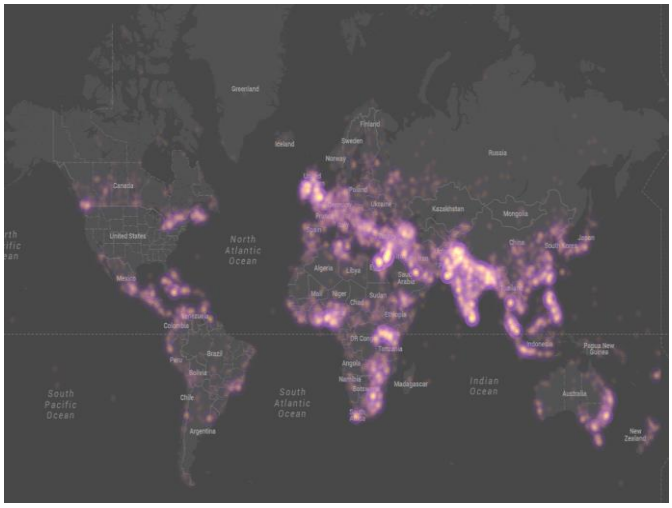
III. BIG DATA IN THE SYSTEMS WORLD

In the distributed systems world, "Big Data" started to become a major issue in the late 1990's due to the impact of the worldwide Web and a resulting need to index and query its rapidly mushrooming content. Database technology (including parallel databases) was considered for the task, but was found to be neither well suited nor cost-effective [17] for those purposes. The turn of the millennium then brought further challenges as companies began to use information such as the topology of the Web and users' search histories in order to provide increasingly useful search results, as well as more effectively-targeted advertising to display alongside and fund those results. Google's technical response to the challenges of Web-scale data management and analysis was simple, by database standards, but kicked off what has

become the modern "Big Data" revolution in the systems world (which has spilled back over into the database world). To handle the challenge of Web-scale storage, the Google File System (GFS) was created [31]. GFS provides clients with the familiar OS-level byte-stream abstraction, but it does so for extremely Large files whose content can span hundreds of machines in shared-nothing clusters created using inexpensive commodity hardware. To handle the challenge of processing the data in such large files, Google pioneered its Map Reduce programming model and platform [23]. This model, characterized by some as "parallel programming for dummies", enabled Google's developers to process large collections of data by writing two user-defined functions, map and reduce, that the Map Reduce framework applies to the instances (map) and sorted groups of instances that share a common key (reduce) – similar to the sort of partitioned parallelism utilized in shared-nothing parallel query processing. Driven by very similar requirements, software developers at Yahoo!, Facebook, and other large Web companies followed suit. Taking Google's GFS and Map Reduce papers as rough technical specifications,

IV. BIG DATA TODAY

If a company has batch-style semi structured data analysis challenges, they can instead opt to enter the Hadoop world by utilizing one or several of the open-source technologies from that world. A lively early "parallel databases vs. Hadoop" debate captured the field's attention in the 2008-2009 timeframe and was nicely summarized in 2010 in a pair of papers written by the key players from the opposing sides of the debate [46, 24].



V. SOME STATISTICS ON BIG DATA

According to Neilson Online currently there are more than 1,733,993,741 internet users. Few numbers to understand how much data is generated every year-

- Email
  - ✓ 90 trillion – The number of emails sent on the Internet in 2009.
  - ✓ 247 billion – Average number of email messages per day.
  - ✓ 1.4 billion – The number of email users worldwide.
  - ✓ 100 million – New email users since the year before.
- Websites
  - ✓ 234 million – The number of websites as of December 2009.
  - ✓ 47 million – Added websites in 2009.
- Web servers
  - ✓ 13.9% – The growth of Apache websites in 2009.
  - ✓ 22.1% – The growth of IIS websites in 2009.
  - ✓ 35.0% – The growth of Google GFE websites in 2009.
  - ✓ 384.4% – The growth of Nginx websites in 2009.
  - ✓ 72.4% – The growth of Lighttpd websites in 2009.
- Domain names
  - ✓ 81.8 million – .COM domain names at the end of 2009.
  - ✓ 12.3 million – .NET domain names at the end of 2009.
  - ✓ 7.8 million – .ORG domain names at the end of 2009.
  - ✓ 76.3 million – The number of country code top-level domains (e.g. .CN, .UK, .DE, etc.).
  - ✓ 187 million – The number of domain names across all top-level domains (October 2009).
  - ✓ 8% – The increase in domain names since the year before.
- Internet users
  - ✓ 1.73 billion – Internet users worldwide (September 2009).
  - ✓ 18% – Increase in Internet users since the previous year.
  - ✓ 738,257,230 – Internet users in Asia.
  - ✓ 418,029,796 – Internet users in Europe.
  - ✓ 252,908,000 – Internet users in North America.
  - ✓ 179,031,479 – Internet users in Latin America / Caribbean.
  - ✓ 67,371,700 – Internet users in Africa.
  - ✓ 57,425,046 – Internet users in the Middle East.
  - ✓ 20,970,490 – Internet users in Oceania / Australia.
- Social media
  - ✓ 126 million – The number of blogs on the Internet (as tracked by BlogPulse).
  - ✓ 84% – Percent of social network sites with more women than men.
  - ✓ 27.3 million – Number of tweets on Twitter per day (November, 2009)
  - ✓ 57% – Percentage of Twitter's user base located in the United States.
  - ✓ 4.25 million – People following @aplusk (Ashton Kutcher, Twitter's most followed user).
  - ✓ 350 million – People on Facebook.
  - ✓ 50% – Percentage of Facebook users that log in every day.
  - ✓ 500,000 – The number of active Facebook applications.

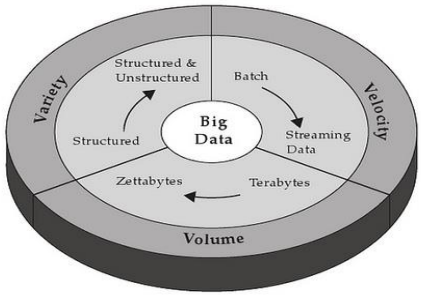
❑ Images
- ✓ 4 billion – Photos hosted by Flickr (October 2009).
- ✓ 2.5 billion – Photos uploaded each month to Facebook.
- ✓ 30 billion – At the current rate, the number of photos uploaded to Facebook per year.

❑ Videos
- ✓ 1 billion – The total number of videos YouTube serves in one day.
- ✓ 12.2 billion – Videos viewed per month on YouTube in the US (November 2009).
- ✓ 924 million – Videos viewed per month on Hulu in the US (November 2009).
- ✓ 182 – The number of online videos the average Internet user watches in a month (USA).
- ✓ 82% – Percentage of Internet users that view videos online (USA).
- ✓ 39.4% – YouTube online video market share (USA).
- ✓ 81.9% – Percentage of embedded videos on blogs that are YouTube videos.

❑ Web browsers
- ✓ 62.7% – Internet Explorer
- ✓ 24.6% – Firefox
- ✓ 4.6% – Chrome
- ✓ 4.5% – Safari
- ✓ 2.4% – Opera
- ✓ 1.2% – Other

VI. BIG DATA DEFINITION:

Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes, exabytes and zettabytes of data in a single data set.
No single standard definition.



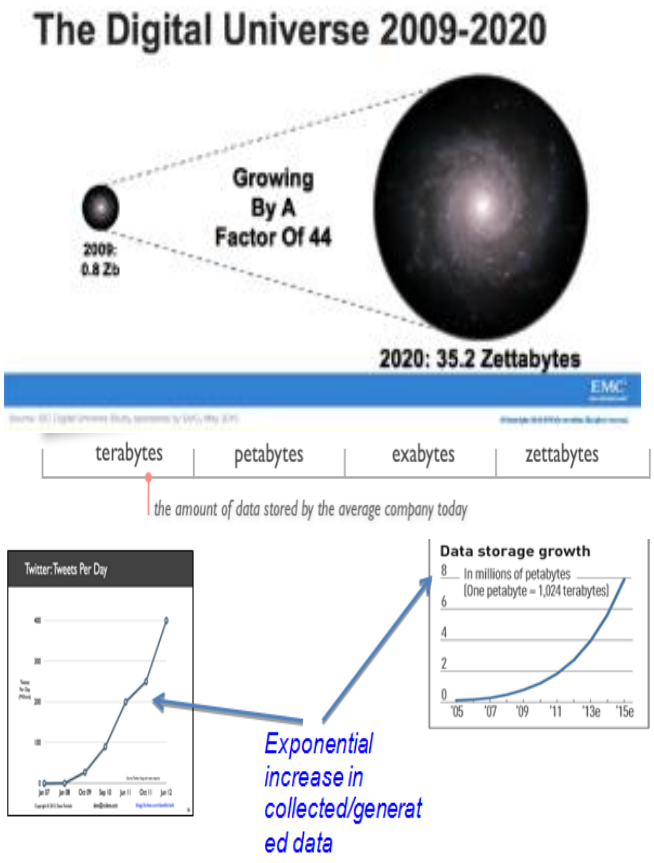
IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.

- "*Big Data*" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

Characteristics of Big Data:
1-Scale (Volume)

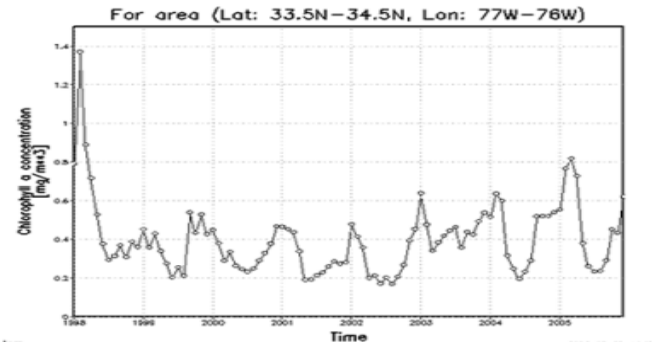
- Data Volume
  - ◦ 44x increase from 2009 to 2020
  - ◦ From 0.8 zettabytes to 35zb

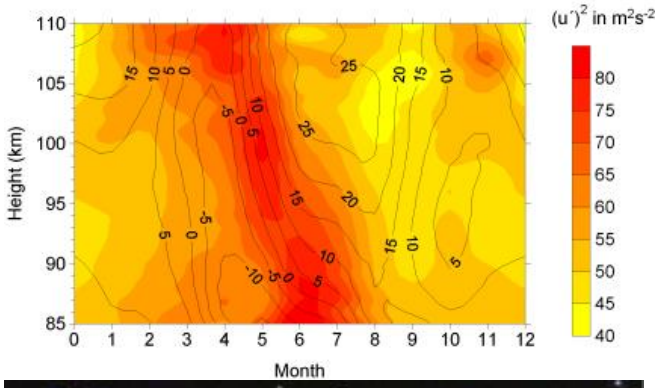
Data volume is increasing exponentially.

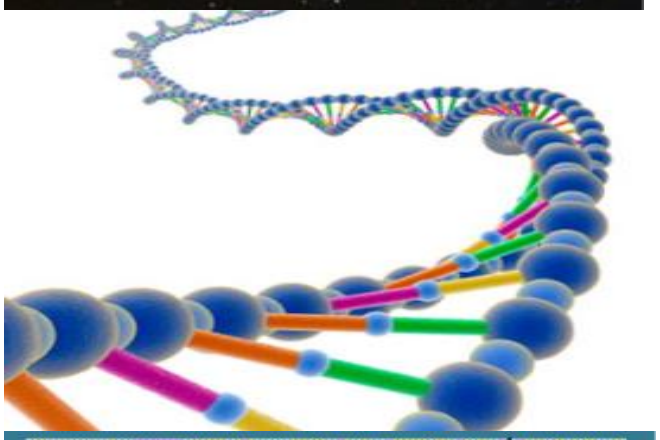


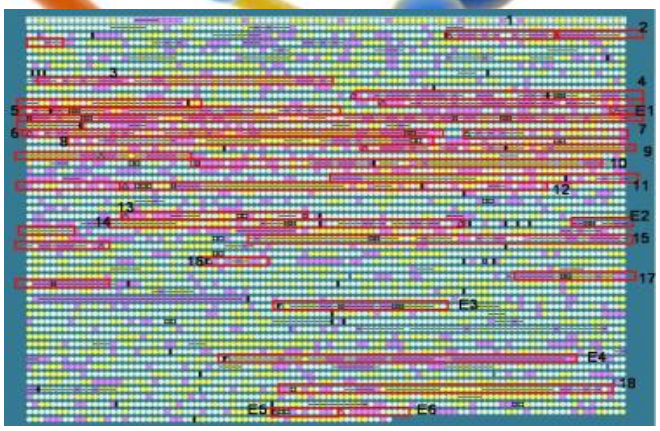
VII. CHARACTERISTICS OF BIG DATA

2-Complexity (Varity)
- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









To extract knowledge➔ all these types of data need to linked together

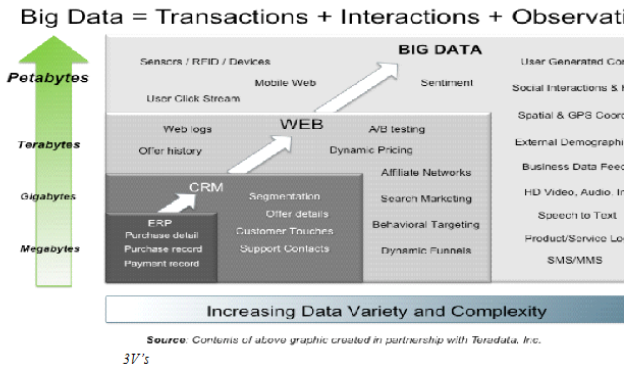
3-Speed (Velocity)

- Data are generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

a) THE DATA REVOLUTION

The world is experiencing a data revolution, or *"data deluge*" (Figure 1). 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 exabytes in 2005 to 1200 exabytes 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.



a)

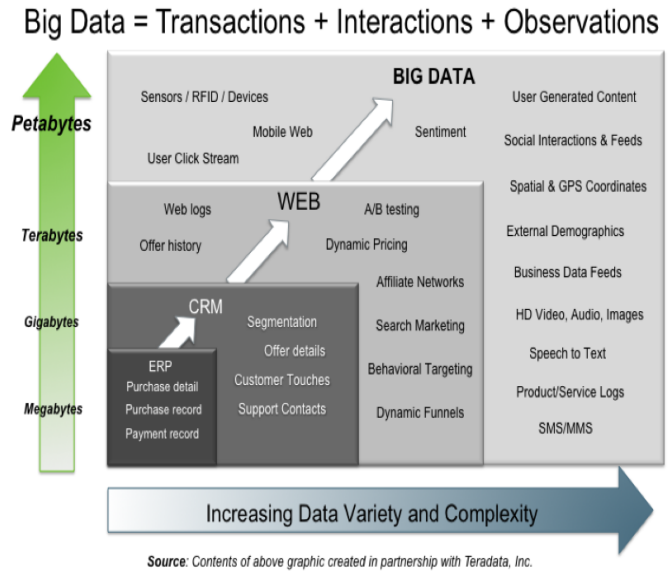
3V's

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you

- **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

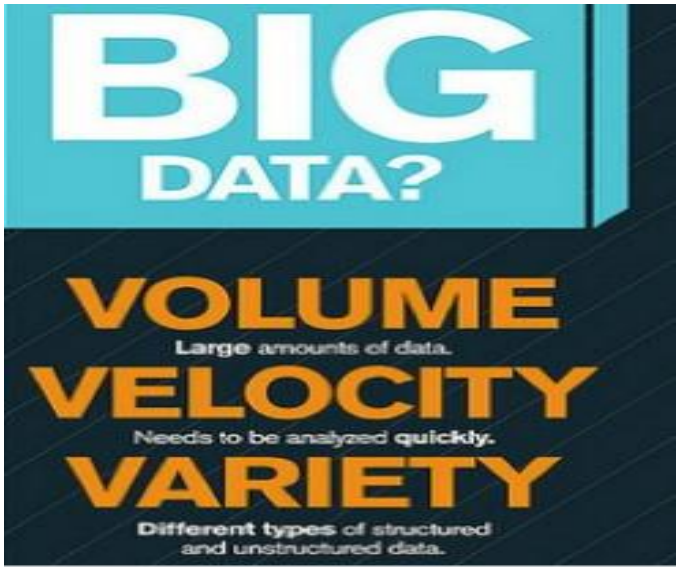
THE DATA REVOLUTION

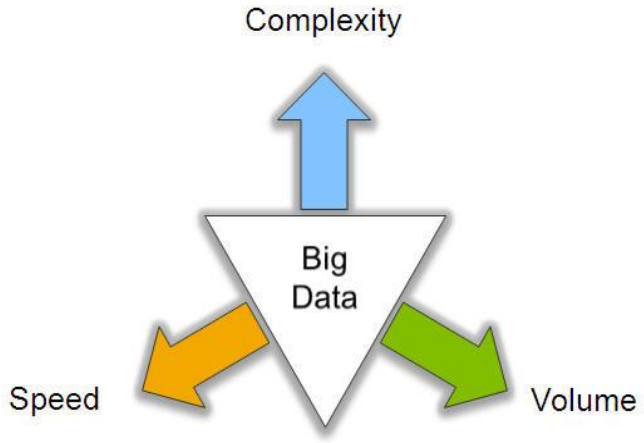
a) **The world is experiencing a data revolution, or *"data deluge***" (Figure 1). Whereas in previous generations, a relatively small volume of analog data was produced and made available through a limited number of channels, today a massive amount of data is regularly being generated and flowing from 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 exabytes in 2005 to 1200 exabytes 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.



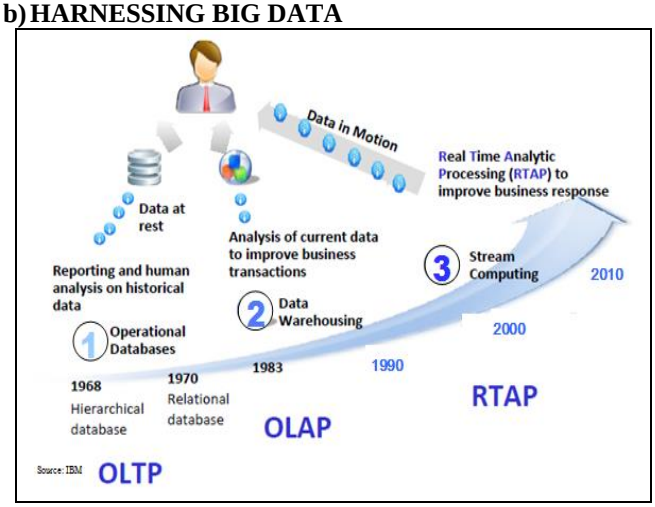
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

3V's



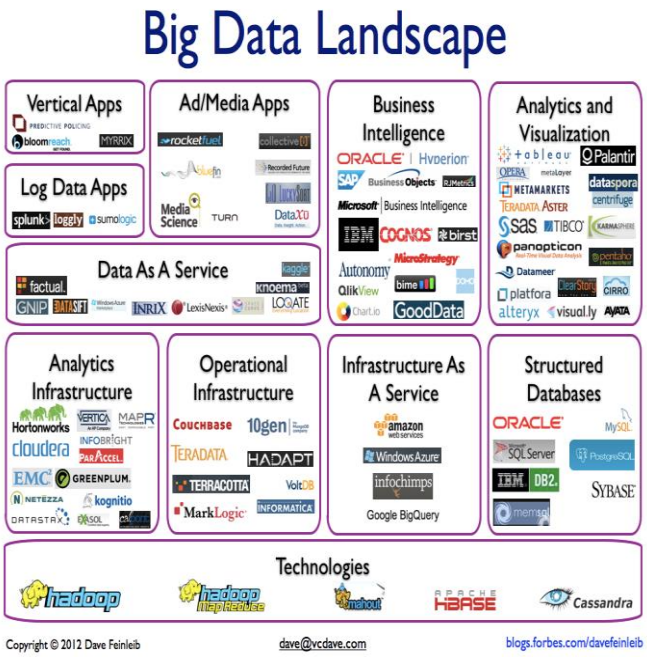


b) HARNESSING BIG DATA



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

c) WHAT TECHNOLOGY DO WE HAVE FOR BIG DATA??



BIG DATA TECHNOLOGY

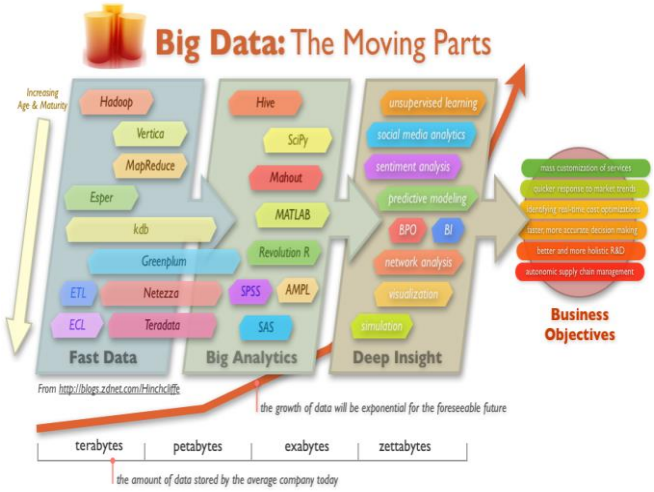
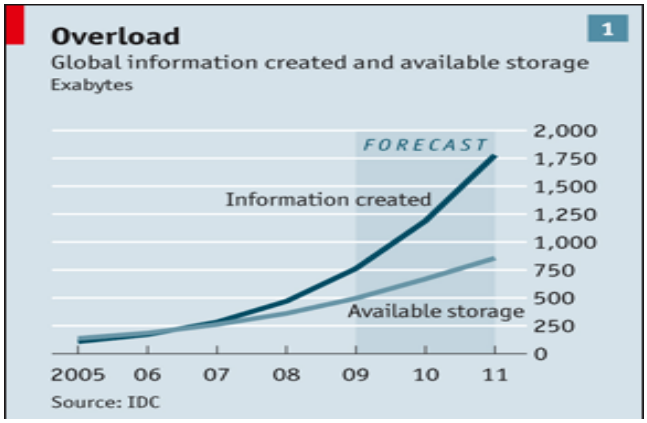


Figure 1: The Early Years of the Data Revolution



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

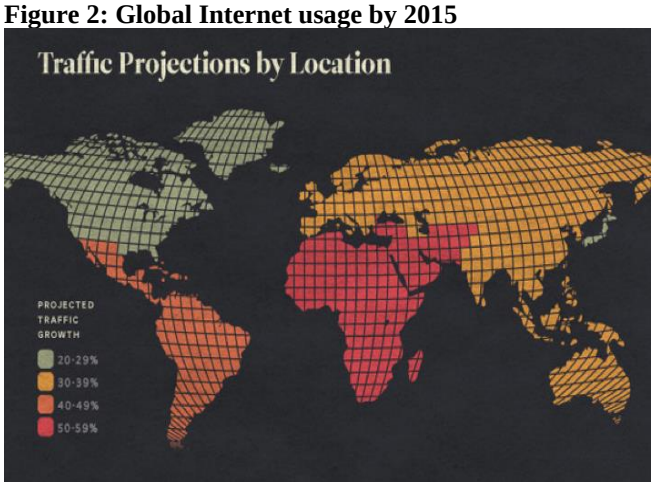
b) THE REVOLUTION HAS VARIOUS FEATURES AND IMPLICATIONS.

The stock of available data gets younger and younger, i.e. the share of data that is "less than a minute old" (or a day, or a week, or any other time benchmark) rises by the minute.iii Further, a large and increasing percentage of this data is both produced and made available real-time (which is a related but different phenomenon).iv The nature of the information is also changing, notably with the rise of social media and the spread of services offered via mobile phones. The bulk of this information can be called "data exhaust," in other words, *"the digitally trackable or storable actions, choices, and preferences that people generate as they go about their daily lives."10* At any point in time and space, such data may be available for thousands of individuals, providing an opportunity to figuratively take the pulse of communities. The significance of these features is worth re-emphasising: this revolution is extremely recent (less than one decade old), extremely rapid (the growth is exponential), and immensely consequential for society, perhaps especially for developing countries.

c) OTHER REAL-TIME INFORMATION STREAMS ARE ALSO GROWING IN DEVELOPING REGIONS.

The use of social media such as Facebook and Twitter is also growing rapidly; in Senegal, for example, Facebook receives about 100,000 new users per month.13 tracking trends in online news or social media can provide information on emerging concerns and patterns at the local level, which can be highly relevant to global development. Furthermore, programme participation metrics collected by UN agencies and other development organisations providing services to vulnerable populations is another Promising source of real-time data, particularly in cases where there is an Information and Communications Technology (ICT) component of service delivery and digital records are generated.

Figure 2: Global Internet usage by 2015



Source: The Atlantic, "Global Internet Traffic Expected to Quadruple by 2015." http://bit.

Big Data for Development: Getting Started "**Big Data**" is a popular phrase used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process with traditional database and software techniques. The characteristics which broadly distinguish Big Data are sometimes called the "3 V's": more volume, more variety and higher rates of velocity. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and from cell phone GPS signals to name a few. This data is known as "Big Data" because, as the term suggests, it is huge in both scope and power.

To illustrate how Big Data might be applicable to a development context, imagine a hypothetical household living in the outskirts of a medium-size city a few hours From the capital in a developing country.

The head of household is a mechanic who owns a small garage. His wife cultivates vegetables and raises a few sheep on their plot of land as well as sews and sells drapes in town. They have four children aged 6 to 18. Over the past couple of months, they have faced soaring commodity prices, particularly food and fuel. Let us consider their options.

The couple could certainly reduce their expenses on food by switching to cheaper Alternatives, buying in bulk, or simply skipping meals. They could also get part of their Food at a nearby World Food Programme distribution center. To reduce other expenses, The father could start working earlier in the morning in order to finish his day before Nightfall to lower his electricity bill. The mother could work longer hours

and go to town Everyday to sell her drapes, rather than twice a week. They could also choose to top-off.

Their mobile phone SIM cards in smaller increments instead of purchasing credit in larger sums and less-frequent intervals. The mother could withdraw from the savings Accumulated through a mobile phone-based banking service which she uses. If things get worse they might be forced to sell pieces of the garage equipment or a few Sheep, or default on their microfinance loan repayment. They might opt to call relatives in Europe for financial support. They might opt to temporarily take their youngest child out of school to save on tuitions fees, school supplies and bus tickets. Over time, if the Situation does not improve, their younger children may show signs of anaemia, prompting them to call a health hotline to seek advice, while their elder son might do online searches, or vent about his frustration on social media at the local cybercafé. Local aid workers and journalists may also report on increased hardships online.

Such a systemic—as opposed to idiosyncratic—shock will prompt dozens, hundreds or thousands of households and individuals to react in roughly similar ways.

Over time, these collective changes in behaviour may show up in different digital data sources. Take this series of hypothetical scenarios, for instance:

(1) The incumbent local mobile operator may see many subscribers shift from adding an average denomination of $10 on their SIM-cards on the first day of the month to a pattern of only topping off $1 every few days; The data may also show a concomitant significant drop in calls and an increase in the use of text messages;

(2) Mobile banking service providers may notice that subscribers are depleting their mobile money savings accounts; A few weeks into this trend, there may be an increase in defaults on mobile repayments of microloans in larger numbers than ever before;

(3) The following month, the carrier-supported mobile trading network might record. Three times as many attempts to sell livestock as are typical for the season;

(4) Health hotlines might see increased volumes of calls reporting symptoms Consistent with the health impacts of malnutrition and unsafe water sources;

(5) Other sources may also pick up changes consistent with the scenario laid out Above. For example, the number of Tweets mentioning the difficulty to "*afford Food"* might begin to rise. Newspapers may be publishing stories about rising Infant mortality;

(6) Satellite imaging may show a decrease in the movement of cars and trucks travelling in and out of the city's largest market;

(7) WFP might record that it serves twice as many meals a day than it did during the same period one year before. UNICEF also holds daily data that may indicate that school attendance has dropped.

The list goes on. This example touches on some of the opportunities available for harnessing the power of real-time, digital data for development. But, let us delve a little deeper into what the relevant characteristics, sources, and categories of Big Data, which could be useful for global development in practice, might be**. Big Data *for* the purposes of *development*** relates to, but differs from, both 'traditional development data' (e.g. survey data, official statistics), and what the private sector and Mainstream media call 'Big Data' in a number of ways. For example, microfinance data (e.g. number and characteristics of clients, loan amounts and types, repayment defaults) falls somewhere between 'traditional development data' and 'Big Data.' It is similar to 'traditional development data' because the nature of the information is important for development experts. Given the expansion of mobile and Online platforms for giving and receiving microloans means that today a large amount of Microfinance data is available digitally and can be analysed in real time, thus qualifying it to be considered Big Data for Development. At the other end of the spectrum, we might include Twitter data, mobile phone data, online queries, etc. These types of data can firmly be called 'Big Data', as popularly defined (massive amounts of digital data passively generated at high frequency). And, while these streams of information may not have traditionally been used in the field of development, but they could prove to be very useful indicators of human well-being. Therefore, we would consider them to be *relevant* Big Data sources for development. Big Data for Development sources generally share some or all of these features:

(1) **Digitally generated** – i.e. the data are created digitally (as opposed to being Digitised manually), and can be stored using a series of ones and zeros, and thus Can be manipulated by computers;

(2) **Passively produced** – a by product of our daily lives or interaction with digital Services;

(3) **Automatically collected** – i.e. there is a system in place that extracts and stores. The relevant data as it is generated;

(4) **Geographically or temporally trackable** – e.g. mobile phone location data or Call duration time;

(5) **Continuously analysed** – i.e. information is relevant to human well-being and Development and can be analysed in real-time;

It is important to distinguish that for the purposes of global development, "real-time" Does not always mean occurring immediately. Rather, "real-time" can be understood as Information which is produced and made available in a relatively short and *relevant Period of time* and information which is made available within a timeframe that allows Action to be taken in response i.e. creating a feedback loop. Xiii Importantly, it is the Intrinsic time dimensionality of the data, *and* that of the feedback loop that jointly define. It's characteristic as real-time. (One could also add that the real-time nature of the data is ultimately contingent on the analysis being conducted in real-time, and by extension, where action is required, used in real-time.)

With respect to spatial granularity, finer is not necessarily better. Village or community Level data may in some cases be preferable to household or individual level data because it can provide richer insights and better protect privacy. As per the time dimensionality, any immediacy benchmark is difficult to set precisely, and will become out-dated, as higher frequency data are made available in greater volumes and with a higher degree of immediacy in the next few years. It must also be noted that real-time is an attribute that doesn't last long: sooner or later, it becomes contextual, i.e. non-actionable data. These include data made available on the spot about average rainfalls or prices, or phone calls made over a relatively long period of time in the past (even a few months), as well as the vast majority of official statistics—such as

GDP, or employment data. Without being too caught up in semantics at length, it is important to recognise that Big Data for Development is an evolving and expanding universe best conceptualised in terms of continuum and irrelativeness. For purposes of discussion, Global Pulse has developed a loose taxonomy of types of new, digital data sources that are relevant to global development:

(1) **Data Exhaust** – passively collected transactional data from people's use of digital Services like mobile phones, purchases, web searches, etc., and/or operational Metrics and other real-time data collected by UN agencies, NGOs and other aid Organisations to monitor their projects and programmes (e.g. stock levels, school Attendance); these digital services create networked sensors of human behaviour;

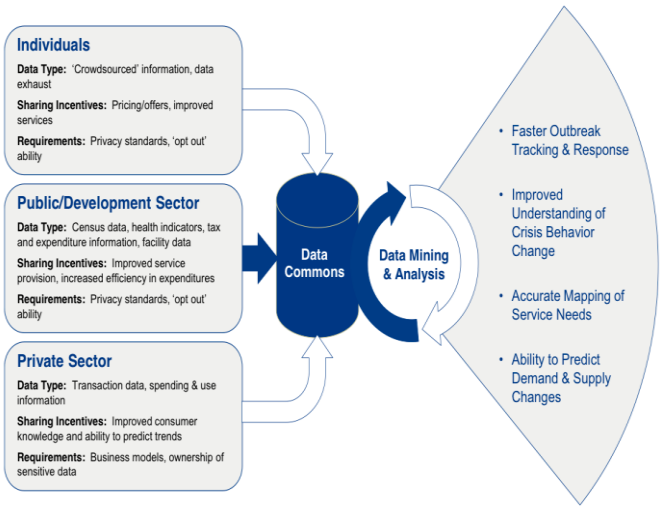
(2) **Online Information** – web content such as news media and social media Interactions (e.g. blogs, Twitter), news articles obituaries, e-commerce, job Postings; this approach considers web usage and content as a sensor of human Intent, sentiments, perceptions, and want;

(3) **Physical Sensors** – satellite or infrared imagery of changing landscapes, traffic Patterns, light emissions, urban development and topographic changes, etc; this Approach focuses on remote sensing of changes in human activity;

(4) **Citizen Reporting or Crowd-sourced Data** – Information actively produced or Submitted by citizens through mobile phone-based surveys, hotlines, user generated Maps, etc; While not passively produced, this is a key information source for verification and feedback.

Yet another perspective breaks down the types of data that might be relevant to International development by how it is produced or made available: by individuals, by the public/development sector, or by the private sector (Figure 3).

Figure 3: "Understanding the Dynamics of the Data Ecosystem"



VIII. CAPACITY OF BIG DATA ANALYTICS

a) The expansion of technical *capacity* to make sense of Big Data in various sectors and academia abounds.

Initially developed in such fields as computational biology, biomedical engineering, Medicine, and electronics, Big Data analytics refers to tools and methodologies that aim to transform massive quantities of raw data into "data about the data"—for analytical purposes. They typically rely on powerful algorithms that are able to detect patterns, trends, and correlations over various time horizons in the data, but also on advanced visualization techniques as "*sense-making tools.*"27 Once trained (which involves having training data), algorithms can help make predictions that can be used to detect anomalies in the form of large deviations from the expected trends or relations in the data. Discovering patterns and trends in the data from the observation and juxtaposition of different kinds of information requires defining a common framework for information processing. At minimum, there needs to be a simple lexicon that will help tag each datum. This lexicon would specify the following:

(1) ***What*:** i.e. the type of information contained in the data;
(2) ***Who*:** the observer or reporter;
(3) ***How*:** the channel through which the data was acquired;
(4) ***How much*:** whether the data is quantitative or qualitative;
(5) ***Where and when*:** the spatio-temporal *granularity* of the data—i.e. the level of Geographic disaggregation (province, village, or household) and the interval at which data is collected.

Then, the data that will eventually lend itself to analysis needs to be adequately prepared. This step may include:

(1) ***Filtering***—i.e. keeping instances and observations of relevance and getting rid of Irrelevant pieces of information;
(2) ***Summarising***—i.e. extracting keyword or set of keywords from a text;
(3) *Categorizing, and/or turning the raw data into an appropriate set of indicators—*

I.e. assigning a qualitative attribute to each observation when relevant—such as 'Negative' vs. 'positive' comments, for instance. Yet another option is simply to Calculate indicators from quantitative data such as growth rates of price indices (I.e. inflation). This is necessary Because these advanced models—non-linear models with many heterogeneous interacting elements—require more data to calibrate them with a data-driven approach. This intensive mining of socioeconomic data, known as "reality mining,"29 can shed light. On processes and interactions in the data that would not have appeared otherwise. Reality mining can be done in three main ways:

(1) *"Continuous data analysis over streaming data,"* using tools to scrape the Web to Monitor and analyse high-frequency online data streams, including uncertain, Inexact data. Examples include systematically gathering online product prices in Real-time for analysis;
(2) *"Online digestion of semi-structured data and unstructured ones"* such as news Items, product reviews etc., to shed light on hot topics, perceptions, needs and wants;
(3) "*Real-time correlation of streaming data (fast stream) with slowly accessible historical data repositories.*" This terminology refers to "*mechanisms for correlating and integrating real-time (fast streams) with historical records…in order to deliver a*

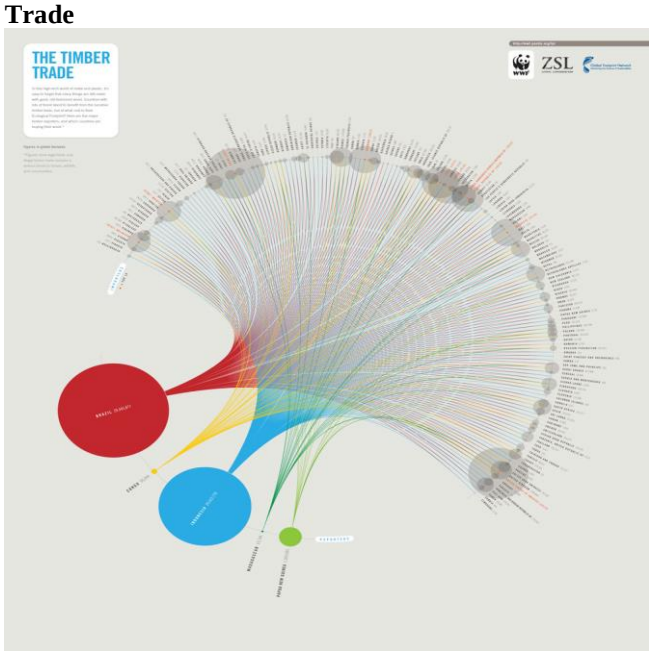
contextualised and personalised information space [that adds] considerable value to the data, by providing (historical) context to new data."

Big Data for Development could use all three techniques to various degrees depending on the availability of data and the specific needs. Further, an important feature of Big Data analytics is the role of visualisation, which can provide new perspectives on findings that would otherwise be difficult to grasp. For example, "word clouds" (Figure 4), which are a set of words that have appeared in a certain body of text – such as blogs, news articles or speeches, for example – are a simple.

Figure 4: A word cloud of this paper

Source: The full text of this paper; word cloud created using www.Wordle.net

Figure 5: Data Visualization of the Global Legal Timber Trade



b) Analysis Working with new data sources brings about a number of analytical challenges

The relevance and severity of those challenges will vary depending on the type of analysis being conducted, and on the type of decisions that the data might eventually inform. The question "*what is the data really telling us*?" is at the core of any social science research and evidence-based policymaking, but there is a general perception that "new" digital data sources poses specific, more acute challenges. *Getting the Picture Right* One is reminded of Plato's allegory of the cave: the data, as the shadows of objects passing in front of the fire, is all the analyst sees.55 But how accurate a reflection is the data? Sometimes the data might simply be false, fabricated.

c) Interpreting Data

In contrast to user-generated text, as described in the section above, some digital data sources—transactional records, such as the number of microfinance loan defaults, number of text messages sent, or number of mobile-phone based food vouchers activated—are as close as it gets to indisputable, hard data. But whether or not the data under consideration is thought to be accurate, interpreting it is never straightforward.

IX. BIG DATA- TOOLS

Open source solutions for processing big data:

- **Hadoop:** Hadoop project develops open-source software for reliable, scalable, distributed computing. Hadoop includes few sub-projects. Hadoop ecosystem consists as follows-
- **HDFS:** Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations.
- **Map Reduce:** MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers.
- **Pig:** Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
- **Hive:** Hive is a data warehouse infrastructure built on top of Hadoop that provides tools to enable easy data summarization, adhoc querying and analysis of large datasets data stored in Hadoop files.

a) HADOOP

1. **Introduction:**

Apache Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware. Hadoop was derived from Google's Map. Reduce and Google File System (GFS) papers. The Hadoop framework transparently provides both reliability and data motion to applications. Hadoop implements a computational paradigm named Map. Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both map/reduce and the distributed file systems are designed so that the framework

automatically handles node failures. It enables applications to work with thousands of computation-independent computers and petabytes of data. The entire Apache Hadoop "platform" is now commonly considered to consist of the Hadoop kernel, MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects – including Apache Hive, Apache HBase, and others.Hadoop is written in the Java programming language and is an Apache top-level project being built and used by a global community of contributors.Hadoop and its related projects (Hive, HBase, Zookeeper, and so on) have many contributors from across the ecosystem. Though Java code is most common, any programming language can be used with "streaming" to implement the "map" and "reduce" parts of the system.

2. Architecture

Hadoop consists of the Hadoop Common package which provides filesystem and OS level abstractions, a MapReduce engine (either MapReduce or YARN) and the Hadoop Distributed File System (HDFS). The Hadoop Common package contains the necessary Java ARchive (JAR) files and scripts needed to start Hadoop. The package also provides source code, documentation and a contribution section that includes projects from the Hadoop Community. For effective scheduling of work, every Hadoop-compatible file system should provide location awareness: the name of the rack (more precisely, of the network switch) where a worker node is. Hadoop applications can use this information to run work on the node where the data is, and, failing that, on the same rack/switch, reducing backbone traffic. HDFS uses this method when replicating data to try to keep different copies of the data on different racks. The goal is to reduce the impact of a rack power outage or switch failure, so that even if these events occur, the data may still be readable. A small Hadoop cluster will include a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode and DataNode. A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications. Hadoop requires Java Runtime Environment (JRE) 1.6 or higher. The standard start-up and shutdown scripts require Secure Shell (ssh) to be set up between nodes in the cluster. In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the file system index, and a secondary NameNode that can generate snapshots of the namenode's memory structures, thus preventing file-system corruption and reducing loss of data. Similarly, a standalone JobTracker server can manage job scheduling. In clusters where the HadoopMapReduce engine is deployed against an alternate file system, the NameNode, secondary NameNode and DataNode architecture of HDFS is replaced by the file-system-specific equivalent.

3. File systems

HDFS is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single namenode; a cluster of datanodes form the HDFS cluster. The situation is typical because each node does not require a datanode to be present. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other. HDFS stores large files (typically gigabytes to terabytes), across multiple machines. It achieves reliability by replicating the data across multiple hosts, and hence does not require RAID storage on hosts. With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack. Data nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high. HDFS is not fully POSIX compliant, because the requirements for a POSIX file system differ from the target goals for a Hadoop application. The tradeoff of not having a fully POSIX-compliant file system is increased performance for data throughput and support for non-POSIX operations such as Append.HDFS added high-availability capabilities, as announced for release 2.0 in May 2012 allowing the main metadata server (the NameNode) to be failed over manually to a backup in the event of failure. Automatic fail-over is being developed as well. Additionally, the file system includes what is called a secondary namenode, which misleads some people into thinking that when the primary namenode goes offline, the secondary namenode takes over. In fact, the secondary namenode regularly connects with the primary namenode and builds snapshots of the primary namenode's directory information, which is then saved to local or remote directories. These checkpointed images can be used to restart a failed primary namenode without having to replay the entire journal of file-system actions, then to edit the log to create an up-to-date directory structure. Because the namenode is the single point for storage and management of metadata, it can be a bottleneck for supporting a huge number of files, especially a large number of small files. HDFS Federation is a new addition that aims to tackle this problem to a certain extent by allowing multiple name spaces served by separate namenodes.An advantage of using HDFS is data awareness between the job tracker and task tracker. The job tracker schedules map or reduce jobs to task trackers with an awareness of the data location. An example of this would be if node A contained data (x,y,z) and node B contained data (a,b,c). Then the job tracker will schedule node B to perform map or reduce tasks on (a,b,c) and node A would be scheduled to perform map or reduce tasks on (x,y,z). This reduces the amount of traffic that goes over the network and prevents unnecessary data transfer. When Hadoop is used with other file systems this advantage is not always available. This can have a significant impact on job-completion times, which has been demonstrated when running data-intensive jobs.HDFS was designed for mostly immutable files and may not be suitable for systems requiring concurrent write operations.Another limitation of HDFS is that it cannot be mounted directly by an existing operating system. Getting data into and out of the HDFS file system, an action that often needs to be performed before and after executing a job, can be inconvenient. A Filesystem in Userspace (FUSE) virtual file system has been developed to address this problem, at least for Linux and some other Unix systems.File access can be achieved through the native Java API, the Thrift API to generate a client in the language of the users' choosing (C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml), the command-line interface, or browsed through the HDFS-UI webapp over HTTP.

4. Other file systems

By May 2011, the list of supported file systems included:

- **HDFS:** Hadoop's own rack-aware file system. This is designed to scale to tens of petabytes of storage and runs on top of the file systems of the underlying operating systems.
- **Amazon S3 file system.** This is targeted at clusters hosted on the Amazon Elastic Compute Cloud server-on-demand infrastructure. There is no rack-awareness in this file system, as it is all remote.
- **MapR'smaprfs file system.** This system provides inherent high availability, transactionally correct snapshots and mirrors while offering higher scaling than HDFS while giving higher performance. Maprfs is available as part of the MapR distribution and as a native option on Elastic Map Reduce from Amazon's web services.
- **CloudStore** (previously Kosmos Distributed File System), which is rack-aware.
- **FTP File system:** this stores all its data on remotely accessible FTP servers.
- **Read-only HTTP and HTTPS file systems.** Hadoop can work directly with any distributed file system that can be mounted by the underlying operating system simply by using a file:// URL; however, this comes at a price: the loss of locality. To reduce network traffic, Hadoop needs to know which servers are closest to the data; this is information that Hadoop-specific file system bridges can provide.Out-of-the-box, this includes Amazon S3, and the CloudStore filestore, through s3:// and kfs:// URLs directly.

A number of third-party file system bridges have also been written, none of which are currently in Hadoop distributions.

- In 2009 IBM discussed running Hadoop over the IBM General Parallel File System. The source code was published in October 2009.
- In April 2010, Parascale published the source code to run Hadoop against the Parascale file system.
- In April 2010, Appistry released a Hadoop file system driver for use with its own CloudIQ Storage product.
- In June 2010, HP discussed a location-aware IBRIX Fusion file system driver.
- In May 2011, MapR Technologies, Inc. announced the availability of an alternative file system for Hadoop, which replaced the HDFS file system with a full random-access read/write file system,with advanced features like snaphots and mirrors, and get rid of the single point of failure issue of the default HDFS NameNode.

5. Job Tracker

Above the file systems comes the Map Reduce engine, which consists of one JobTracker, to which client applications submit Map Reduce jobs. The JobTracker pushes work out to available Task Tracker nodes in the cluster, striving to keep the work as close to the data as possible. With a rack-aware file system, the Job Tracker knows which node contains the data, and which other machines are nearby. If the work cannot be hosted on the actual node where the data resides, priority is given to nodes in the same rack. This reduces network traffic on the main backbone network. If a Task Tracker fails or times out, that part of the job is rescheduled. The TaskTracker on each node spawns off a separate Java Virtual Machine process to prevent the TaskTracker itself from failing if the running job crashes the JVM. A heartbeat is sent from the TaskTracker to the JobTracker every few minutes to check its status. The Job Tracker and TaskTracker status and information is exposed by Jetty and can be viewed from a web browser.If the JobTracker failed on Hadoop 0.20 or earlier, all ongoing work was lost. Hadoop version 0.21 added some checkpointing to this process; the JobTracker records what it is up to in the file system. When a JobTracker starts up, it looks for any such data, so that it can restart work from where it left off.

Known limitations of this approach are:The allocation of work to TaskTrackers is very simple. Every TaskTracker has a number of available slots (such as "4 slots"). Every active map or reduce task takes up one slot. The Job Tracker allocates work to the tracker nearest to the data with an available slot. There is no consideration of the current system load of the allocated machine, and hence its actual availability.If one TaskTracker is very slow, it can delay the entire MapReduce job - especially towards the end of a job, where everything can end up waiting for the slowest task. With speculative execution enabled, however, a single task can be executed on multiple slave nodes.

6. Scheduling

By default Hadoop uses FIFO, and optional 5 scheduling priorities to schedule jobs from a work queue. In version 0.19 the job scheduler was refactored out of the JobTracker, and added the ability to use an alternate scheduler (such as the Fair scheduler or the Capacity scheduler).

Fair scheduler: The fair scheduler was developed by Facebook. The goal of the fair scheduler is to provide fast response times for small jobs and QoS for production jobs. The fair scheduler has three basic concepts.

- Jobs are grouped into Pools.
- Each pool is assigned a guaranteed minimum share.
- Excess capacity is split between jobs.

By default, jobs that are uncategorized go into a default pool. Pools have to specify the minimum number of map slots, reduce slots, and a limit on the number of running jobs.

Capacity scheduler: The capacity scheduler was developed by Yahoo. The capacity scheduler supports several features which are similar to the fair scheduler.

- Jobs are submitted into queues.
- Queues are allocated a fraction of the total resource capacity.
- Free resources are allocated to queues beyond their total capacity.
- Within a queue a job with a high level of priority will have access to the queue's resources. There is no preemption once a job is running.

7. Other Applications

The HDFS file system is not restricted to MapReduce jobs. It can be used for other applications, many of which are under development at Apache. The list includes the HBase database, the Apache Mahout machine learning system, and the Apache Hive Data Warehouse system. Hadoop can in theory be used for any sort of work that is batch-oriented rather than real-time, that is very data-intensive, and able to work on

pieces of the data in parallel. As of October 2009, commercial applications of Hadoop included:

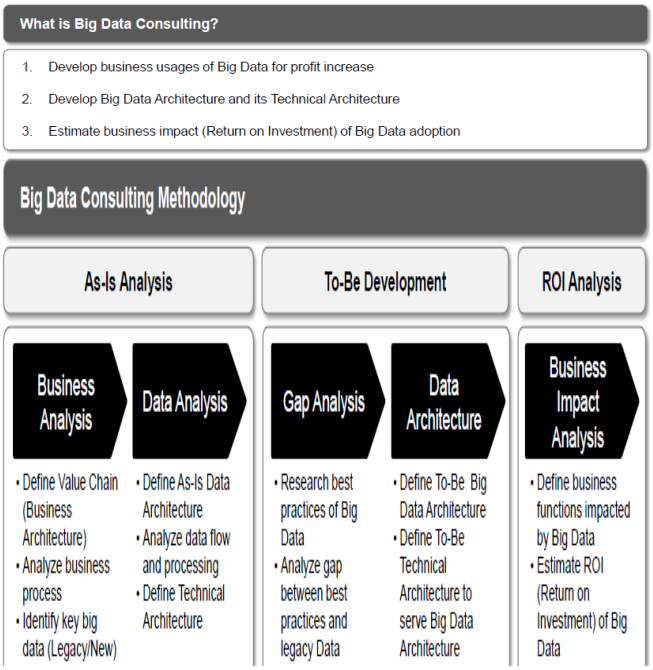
- Log and/or click stream analysis of various kinds
- Marketing analytics
- Machine learning and/or sophisticated data mining
- Image processing
- Processing of XML messages
- Web crawling and/or text processing
- General archiving, including of relational/tabular data, e.g. for compliance

X. Remarks on the Future of Big Data for Development:

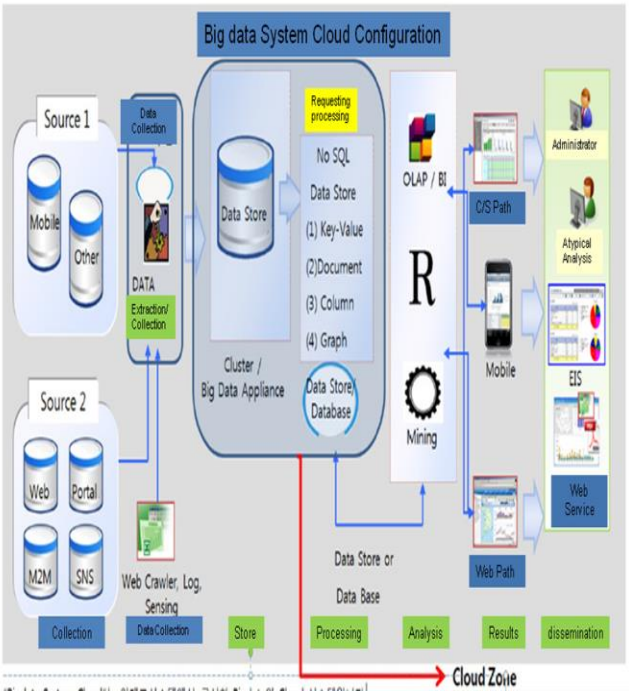
Big Data is a sea change that, like nanotechnology and quantum computing, will shape the twenty-first century. According to some experts, "*[by] employing massive data mining, science can be pushed towards a new methodological paradigm which will transcend the boundaries between theory and experiment."* Another perspective frames this new ability to unveil stylized facts from large datasets as "*the fourth paradigm of science*". This paper does not claim that Big Data will simply replace the approaches, tools and systems that underpin development work. What it does say, however, is that 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. The question is neither **"if," nor "when," but "how."**

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. First, because the new types of data that people will produce in ten years is unknown. Second, because the same uncertainty holds for computing capacities, given that Moore's Law xxxii with certainly not hold in an era of quantum computing. Third, because a great deal will depend on the future strategic decisions taken by a myriad of actors—chief of which are policymakers. 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, based on a full recognition of the opportunities and challenges. Specifically, its success hinges on two main factors. One is the level of institutional and financial support from public sector actors, and the willingness of private corporations and academic teams to collaborate with them, including by sharing data and technology and analytical tools. Two is the development and implementation of new norms and ontology's for the responsible use and sharing of Big Data for Development, backed by a new institutional architecture and new types of partnerships. Our hope is that this paper contributes to generating exchanges, debates and interest among a wide range of readers to advance Big Data for Development in the twenty-first century.

Annex 1. Big Data



Annex 2. Big Data Configuration



XI. ACTION PLAN FOR BIG DATA MANAGEMENT IN BANGLADESH:

SL	Activities	Main Actors	MoI	Timeline		Remarks
				Start	End	
1	2	3	4	5	6	7
01	Listing of Big Data producer and user in the Government sector.	BBS		June 15 2013		Proposed
02	Study of Statistics Low in terms of Big data Management	SID, BBS		June 15 2013		Proposed
03	Setup a core committee regarding Big Data Rights/ Law-Rules implementation.	SID, BBS, A2I		July 1 2013		Proposed
04	Study of Foreign Govt Strategy to develop Big Data Governance strategy.	BBS		July 1 2013		Proposed
05	Develop Governance strategy for big data management with collaboration of SID, BBS & A2I.	SID		Sept 1 2013		Proposed
06	Dissemination of Govt. Strategy, Law, policy through a workshop	BBS		Nov 1 2013		Proposed
07	Policy & Advisory assistance to Big Data producers and users.	BBS		Jan 1 2014		Proposed
08	Study/use BANBES,BCC intranet-infrastructure for e service delivery in Uzila level.	BBS		Feb 1 2014		Proposed
09	Setup a tire 3 DR-site in an earth-quake free zone.	A2I		Oct 1 2014		
10	Setup an ICT-Cell in every stakeholder including Digital archiving, processing, storage and security. Facilitates the building of Ministry's/ Agency's own service database	A2I		Aug 1 2014		Proposed
11	Setup Big Data Management tools and Business Intelligence using BBS ICT infrastructure	BBS, A2I		July 1 2014		Proposed
12	Develop the common e-service delivery framework (Guided by National Information Management Committee)	SID, BBS		Jan 1 2015		Proposed

XII. BBS CREATE HUGE VOLUME OF DATA:

Historical & Time series Perspective of BBS Data are divided into following major groups:

a) Census:

Serial No.	Periodicity	Census Bangladesh	Status of Publications	Data State
1	2	3	4	5
a) Population & Housing Census				
1)	Decennial	1974 (Census was due in 1971, could not be held due to our liberation war)	Published	Hard-copy Publication.
2)	Decennial	1981	Published	Total 4.0 GB., Text format + Meta data
3)	Decennial	1991	Published	Total 6.0 GB. Text format + Meta data
4)	Decennial	2001	Published	Total 8.0 GB. Text format + Meta data
5)	Decennial	2011	Partially Published	
b) Agriculture Census				
1)	Irregular Interval	1977	Published	Total 6.00 GB. Text, FoxPro format + Meta data
2)	Irregular Interval	1983-84	Published	Total 7.50 GB. Text, FoxPro format + Meta data
3)	Irregular Interval	1996	Published	Total 11.00 GB. Text, FoxPro format + Meta data

Serial No.	Periodicity	Census Bangladesh	Status of Publications	Data State
1	2	3	4	5
c) Economic Census				
1)	1st Time in the country	1986	Published	Total 6.00 GB. Text, FoxPro format + Meta data
2)	Irregular Interval (Phased)	2001 (Urban) 2003 (Rural)	Published	Total 8.00 GB. Text, FoxPro format + Meta data
3)	Decennial	2013	Preliminary report Published	Oracle Database + Meta data
d) Slum Census				
1)	Irregular Interval	1986	Published	Total 0.50 GB. Text, FoxPro format + Meta data
2)	Irregular Interval	1997	Published	Total 1.00 GB. Text, FoxPro format + Meta data
3)	Irregular Interval	2013	Preliminary Report is expected to be published in December 2014 and Final Report in June 2015	Expected vol. 2.00GB Text format + Meta data

b) Surveys:

Serial No.	Name of Survey	Purpose	Timeline	Data State
1	2	3	4	5
1)	Labour Force Survey (LFS)	To determine total labour force, unemployment rate, employment by sex, industry and occupation	1999 to 2010	Total 1.00 GB. Foxpro,STATA format + Meta data
2)	Multiple Indicator Cluster Survey (MICS)	To know the health, education and nutrition status of women and children	1998 to 2009	Total 1.50 GB. Foxpro,STATA format + Meta data
3)	Survey of Manufacturing Industries (SMI)	To estimate the gross value addition and gross fixed capital formation in industry sector	2012-13	Total 2.00 GB. STATA format + Meta data
4)	Agriculture Crop Production Survey	To estimate the crop-wise production and yield rate of 6 major crops and 118 minor crops	Done annually (2012)	Total 3.50 GB. STATA format + Meta data
5)	Sample Vital Registration Survey (SVRS)	To estimate intercensal population growth and other demographic indicators	Done annually (2002 to 2011)	Total 1.50 GB. Foxpro, STATA format + Meta data
6)	Health and Mortality Survey (HMS)	To estimate fertility , mortality, morbidity, causes of death and health expenditure	2012	Total .50 GB. Foxpro, STATA format + Meta data
7)	Literacy Assessment Survey (LAS)	To assess reading, writing, numeracy and comprehension level of the population	2011	Total .30 GB. Foxpro, STATA format + Meta data
8)	Child and Mother Nutrition Survey (CMNS)	To estimate nutrition status of women and children	1992 to 2012	Total .50 GB. FoxPro, STATA format + Meta data
9)	Price and Wage Rate Survey	To compute CPI and determine food inflation, nonfood inflation and Wage Rate Index (WRI)	Monthly	Total .50GB/Y Foxpro,STATA format + Meta data
10)	Household Income and Expenditure Survey (HIES)	To estimate poverty and inequality level as well as income, expenditure and consumption of household	1981 to 2010	Total 5.00 GB Foxpro,STATA format + Meta data

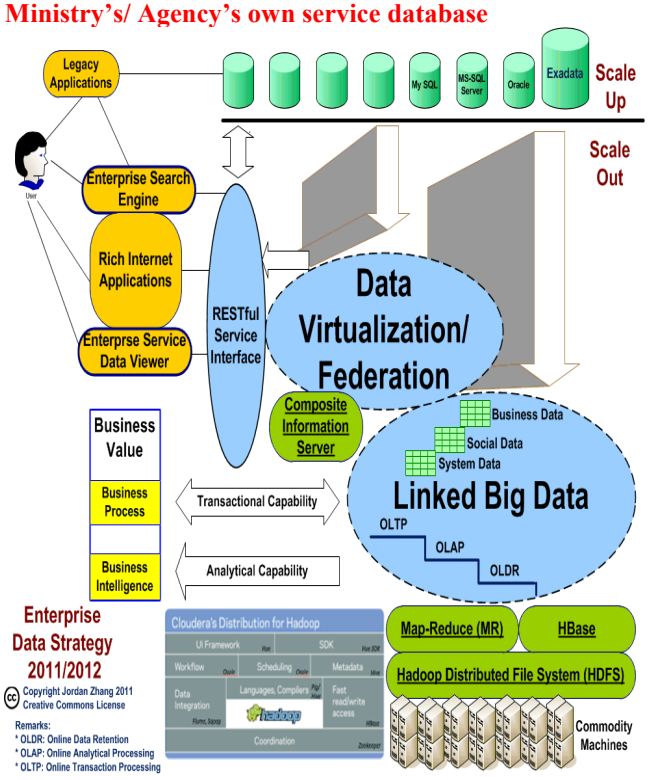
c) Other Census & Survey Programs:

Name of Programs	Current Status
Slum Census and Census of Floating Population	**Pre-test on questionnaire has been done and discussed in a workshop It is being finalized now**
Survey on Rural Credit	**Draft questionnaire has been prepared and discussed in the technical committee It is being finalized now.**
Survey on Remittance	**Preparation of questionnaire, sample design, Training of supervisors completed, Data collection is being done now.**
Survey on Production cost of selected crops	**Preparation of questionnaire, sample design, Training of supervisors and enumerators completed, Data collection on selected crops is being done now.**
Test phase for National Population Register (Pilot)	**Data collection has been completed. It is being processed now. The result will be available by June 30, 2013**

I. Develop the necessary information/databases in various sectors of the economy. For planning and decision making at the district level.
II. Promote informatics culture at the district level.
III. Improve the analysis capacity and the presentation of the statistics utilized for national, regional and district planning.
IV. Develop modeling and forecasting techniques that are required for decision making for socio-economic development.

DISNIC facilitates easy collection, compilation, dissemination and on-line accessibility of information on several sectors of the economy at country level with the availability of qualitative information at all possible levels. DISNIC also facilitates the building up of databases of national importance through active co-operation of the Governments.

Ministry's/ Agency's own service database



Questionnaire on use of Big Data by United Nations

XIII. <QUESTIONNAIRE ON USE OF BIG DATA<

Use of Big Data sources in the National Statistical System
Please indicate which of the following Big Data sources will likely be used in the *next 12 months* by your office or other agencies that are part of the National Statistical System of your
Country. Please indicate which of the following Big Data sources will likely be used in the *next 12 months* by your office or other agencies that are part of the National Statistical System of your country. Administrative sources arising from the administration of a program, be it governmental or not; for instance, electronic medical records, hospital visits, insurance records, bank records, or food banks Commercial or transactional arising from the transaction between two entities; for instance, credit card transactions, or other online transactions (including from mobile devices) Sensor networks; for instance, satellite imaging, road sensors, or climate sensors Tracking devices; for instance, tracking data from mobile telephones, or the Global Positioning System (GPS).

Legislation in some countries (e.g. Canada) may provide the right to access data from both government and nongovernment organizations while others (e.g. Ireland) may provide the right to access data from public authorities only. This can result in limitations for accessing certain types of Big Data. It is recognized that the right of NSOs to access administrative data, established in principle by the law, often is not adequately supported by specific obligations for the data holders. Even if legislation has provision to access all types of data, the statistical purpose for accessing the data might need to be demonstrated to an extent that may be different from country to country.

Definitions may vary from country to country but privacy is generally defined as the right of individuals to control or influence what information related to them may be disclosed. The parallel can be made with companies that wish to protect their competitiveness and consumers. Privacy is a pillar of democracy. The problem with Big Data is that the users of services and devices generating the data are most likely unaware that they are doing so, and/or what it can be used for. The data would become even bigger if they are pooled, as would the privacy concerns. Are privacy issues, such as managing public trust and acceptance of data reuse and its link to other sources, a major challenge for use of Big Data by the National Statistical System in your country?

There is likely to be a cost to the NSOs to acquire Big Data, especially Big Data held by the private sector and particularly if legislation is silent on the financial modalities surrounding acquisition of external data. As a result, the right choices have to be made by NSOs, balancing quality (which encompasses relevance, timeliness, accuracy, coherence, accessibility and interpretability) against costs and reduction in response burden. Costs may even be significant for NSOs but the potential benefits far outweigh the costs, with Big Data potentially providing information that could increase the efficiency of government programs (e.g. health system). Rules around procurement in the government may come into play as well. One of the findings in the report prepared by TechAmerica Foundation's Federal Big Data Commission in the United States was that the success of transformation to Big Data lies in: Understanding a specific agency's critical business imperatives and requirements, developing the right questions to ask and understanding the art of the possible, and taking initial steps focused on serving a set of clearly defined use cases. This approach can certainly be transposed in an NSO environment. Are financial issues, such as potential costs of sourcing data versus benefits, a major challenge for use of Big Data by the National Statistical System in your country?

No, this issues do not constitute a major challenge
No opinion: this issues have not been considered yet
Yes, this issues are a major challenge (please explain)

Big Data for official statistics means more information coming to NSOs that is subject to policies and directives to which NSOs must adhere. Another management challenge refers to human resources, as the data science associated with Big Data that is emerging in the private sector does not seem

to have connected yet with the official statistics community. The NSOs may have to invest in inhouse training for data exploration or acquire data scientists. Are management issues, such as adhering to new policies and regulations, and developing human resources with the necessary set of skills and expertise, a major challenge for use of Big Data by the National Statistical System in your country?
No, this issues do not constitute a major challenge
No opinion: this issues have not been considered yet
Yes, this issues are a major challenge (please explain)

Representativeness is the fundamental issue with Big Data. The difficulty in defining the target population, survey population and survey frame jeopardizes the traditional way in which official statisticians think and do statistical inference about the target (and finite) population. With a traditional survey, statisticians identify a target/survey population, build a survey frame to reach this population, draw a sample, collect the data etc. They will build a box and fill it with data in a very structured way. With Big Data, data come first and the reflex of official statisticians would be to build a box! This raises the question is this the only way to produce a coherent and integrated national system of official statistics? Is it time to think outside of the box? Another issue is both related to information technology (IT) and methodology in nature. When more and more data is being analyzed, traditional statistical methods, developed for the very thorough analysis of small samples, run into trouble; in the most simple case they are just not fast enough. There comes the need for new methods and tools:

a. Methods to quickly uncover information from massive amounts of available data, such as visualization methods and data, text and stream mining techniques, which are able to 'make Big Data small'. Increasing computer power is a way to assist with this step at first;
b. Methods capable of integrating the information uncovered in the statistical process, such as linking at massive scale, macro/mesointegration, and statistical methods specifically suited for large datasets. Methods need to be developed that rapidly produce reliable results when applied to very large datasets.

The use of Big Data for official statistics definitely triggers a need for new techniques. Methodological issues that these techniques need to address are:

(a) Measures of quality of outputs produced from hard to manage external data supply. The dependence on external sources limits the range of measures that can be reported when compared with outputs from targeted information gathering techniques;
(b) Limited application and value of externally sourced data;
(c) Difficulty of integrating information from different sources to produce high value products;
(d) Difficulty of identifying a value proposition in the absence of the closed loop feedback seen in commercial organizations.

Are methodological issues, such as data quality and suitability of statistical methods, a major challenge for use of Big Data by the National Statistical System in your country?
No, this issues do not constitute a major challenge
No opinion: this issues have not been considered yet
Yes (please explain)
Improving data velocity of accessing administrative data means to also use intensively standard Application Programme Interfaces (APIs) or (sometimes) streaming APIs to access data. In this way it is possible to connect applications for data capturing and data processing directly with administrative data. Collecting data in real time or near real time maximizes in fact the potential of data, opening new opportunities for combining administrative data with highvelocity data coming from other different sources, such as:

a) Commercial data (credit card transactions, on line transactions, sales, etc.);
b) Tracking devices (cellular phones, GPS, surveillance cameras, apps) and physical sensors (traffic, meteorological, pollution, energy, etc.);
c) Social media (twitter, Facebook, etc.) and search engines (online searches, online page view);
d) Community data (Citizen Reporting or Crowd sourced data). In an era of Big Data this change of paradigm for data collection presents the possibility to collect and integrate many types of data from many different sources. Combining data sources to produce new information is an additional interesting challenge in the near future. Combining "traditional" data sources, such as surveys and administrative data, with new data sources as well as new data sources with each other provide opportunities to describe behaviors of "smart" communities. It is yet an unexplored field that can open new opportunities.

Are information technology issues a major challenge for use of Big Data by the National Statistical System in your country?

No, this issues do not constitute a major challenge
No opinion: this issues have not been considered yet
Yes, this issues are a major challenge (please explain)
Are there other major challenges for use of Big Data by the National Statistical System in your country?
No, this issues do not constitute a major challenge
No opinion: this issues have not been considered yet
Yes, this issues are a major challenge (please explain)
Please specific whether the following statistical domains are areas of either use, or research into use, of Big Data in official statistics in next 12 months:
Demographic and social statistics (including subjective wellbeing)
Vital and civil registration statistics
Economic and financial statistics
Price statistics
Areas of use (or research into use) of Big Data in official statistics in n...
No, this is not a specific area of use (or research into use) of Big Data in the next 12 months
Yes, this is a specific area of use (or research into use) of Big Data in the next 12 months (please explain)
No, this is not a specific area of use (or research into use) of Big Data in the next 12 months
Yes, this is a specific area of use (or research into use) of Big Data in the next 12 months (please explain)
No, this is not a specific area of use (or research into use) of Big Data in the next 12 months
Yes, this is a specific area of use (or research into use) of Big Data in the next 12 months (please explain)

No, this is not a specific area of use (or research into use) of Big Data in the next 12 months
Yes, this is a specific area of use (or research into use) of Big Data in the next 12 months (please explain)
Transportation statistics
Environmental statistics
Other domains of official statistics

No, this is not a specific area of use (or research into use) of Big Data in the next 12 months
Yes, this is a specific area of use (or research into use) of Big Data in the next 12 months (please explain)
No, this is not a specific area of use (or research into use) of Big Data in the next 12 months
Yes, this is a specific area of use (or research into use) of Big Data in the next 12 months (please explain)
No, there are no other domains of official statistics that constitute a specific area of use (or research into use) of Big Data in the next 12 months
Yes, the following domain(s) of official statistics constitute a specific area of use (or research into use) of Big Data in the next 12 months (please explain)

Are you aware of any documents describing experiences in your country regarding the use of Big Data in official statistics?
No
Yes (please specify, and, if available, please provide the Web links where these documents can be obtained.

XIV. BEST PRACTICES IN DATA MANAGEMENT FOR BIG DATA

a) Big Data Happens when Storage and Compute Demand Increase

In traditional data storage environments, servers and computation resources are in place to process the data. However, even using today's traditional data storage mechanisms, there are data challenges that can stretch the capacity of storage systems. Tasks such as simulations and risk calculations, which work on relatively small amounts of data, can still generate computations that can take days to complete, placing them outside the expected decision-making window needed. Other business processes may require long-running ETL-style processes or significant data manipulation. When traditional data storage and computation technologies struggle to provide either the storage or the computation power required to work with their data, an organization is said to have a big data issue.

b) Accurate and Timely Decision Making

Ultimately, the goal of most data processing tasks is to come to a business decision. An organization is deemed to have big data when any of the above factors, individually or in combination, make it difficult for an organization to make the business decisions needed to be competitive. While a large organization may have different big data issues compared to the big data concerns of a small firm, ultimately the problem comes down to the same set of challenges. We will now discuss the technologies being used to address big data challenges and how you can bring the power of SAS to help solve those challenges.

c) SAS and Big Data Technologies

SAS is pursuing a number of complementary strategies for big data, enabling you to decide which approach is right for your enterprise. These strategies are:

- Using emerging big data platforms (Apache Hadoop).
- Creating new technology for problems not well-addressed by current big data platforms (SAS LASR and SAS High-Performance Analytics).
- Moving more computation to traditional databases (SAS In-Database).
- Implementing data virtualization (SAS Federation Server).

Let's look at each of these in turn, and discuss how SAS Data Management make dealing with big data easier.

d) Apache Hadoop

The most significant new technology that has emerged for working with big data is Apache Hadoop. Hadoop is an open-source set of technologies that provide a simple, distributed storage system paired with a parallel processing approach well-suited to commodity hardware. Based on original Google and Yahoo innovations, it has been verified to scale up to handle big data. Many large organizations have already incorporated Hadoop into their enterprise to process and analyze large volumes of data with commodity hardware using Hadoop. In addition, because it is an open and extensible framework, a large array of supporting tools are available that integrate with the Hadoop framework.

e) Hadoop Technology Overview

Table 1 describes some of the available Hadoop technologies and their purpose in the Hadoop infrastructure.

Hadoop Technology	Purpose
HDFS	Hadoop Distributed File System (HDFS) is a distributed, scalable and portable file system written in Java for the Hadoop framework. Users load files to the file system using simple commands, and HDFS takes care of making multiple copies of data blocks and distributing those blocks over multiple nodes in the Hadoop system to enable parallel operation, redundancy and failover.
MapReduce	The key programming and processing algorithm in Hadoop. The algorithm divides work into two key phases: Map and Reduce. Not all computation and analysis can be written effectively in the MapReduce approach, but for analysis that can be converted, highly parallel computation is possible. MapReduce programs are written in Java. All the other languages available in Hadoop ultimately compile down to MapReduce programs.
Pig	Pig Latin is a procedural programming language available for Hadoop. It provides a way to do ETL and basic analysis without having to write MapReduce programs. It is ideal for processes in which successive steps operate on data. Here is a Pig Latin program example: A = load 'passwd' using PigStorage(':'); B = For each A generate $0 as id; dump B; store B into 'id.out';
Hive	Hive is another alternative language for Hadoop. Hive is a declarative language very similar to SQL. Hive incorporates HiveQL (Hive Query Language) for declaring source tables, target tables, joins and other functions similar to SQL that are applied to a file or set of files available in HDFS. Most importantly, Hive allows structured files, such as comma-delimited files, to be defined as tables that

Hadoop Technology	Purpose
	HiveQL can query. Hive programming is similar to database programming. Here is a Hive program example: INSERT OVERWRITE TABLE pages SELECT redirect_table.page_id, redirect_table. redirect_title, redirect_table.true_title, redirect_table.page_latest, raw_daily_ stats_table.total_pageviews, raw_daily_stats_table.monthly_trend FROM redirect_table JOIN raw_daily_stats_table ON (redirect_table.redirect_title = raw_daily_stats_table.redirect_title);

SAS® and Hadoop Integration

Figure 1 indicates the integration of various components of SAS and Hadoop. The
Hadoop technologies are indicated in grey, traditional SAS technologies in light blue and newer SAS technologies in dark blue.

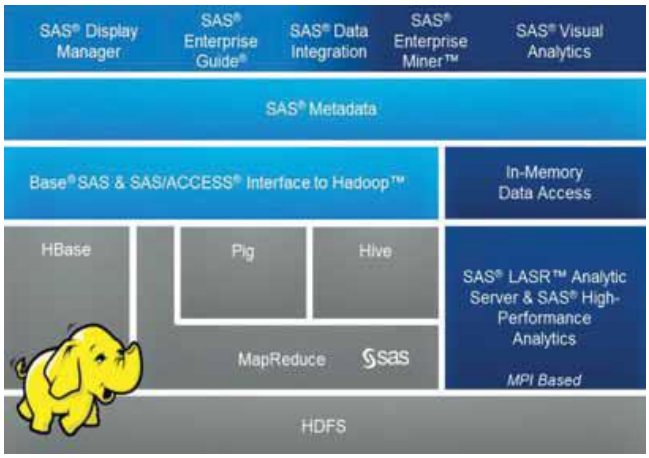


Figure 1: SAS and Hadoop Integration

SAS Data Integration Studio enables organizations to use Hadoop in the following ways:

1) As a data environment, by using a SAS/ACCESS® engine.
2) As a file-based storage environment, by using SAS file I/O capabilities.
3) As a computation environment, by using Hadoop transformations in SAS Data Integration Studio for Pig, HIVE and MapReduce programming.
4) As a data preparation environment for SAS LASR Analytic Server with conversion capabilities to SAS LASR Hadoop storage.

f) Accessing Data in HDFS Using SAS® File I/O : SAS can access data stored in the HDFS in several ways. The first uses file input and output. The SAS file input/output capabilities have been enhanced to read and write directly to the HDFS. Using the SAS infile statement, the File Reader and File Writer transformations in SAS Data Integration Studio can directly read and write HDFS files. Using SAS in combination with Hadoop also adds several unique capabilities that are not part of the Hadoop language itself. This helps you bring the power of SAS to your Hadoop programs. These capabilities are:

The HDFS is a distributed file system, so components of any particular file may be separated into many pieces in the HDFS. Using SAS, you do not need to know details about the HDFS files or how they are distributed. SAS is able to interact with the distributed components of any file on the HDFS as if it were one consolidated file. You can work with HDFS distributed files like you would work with any other file coming from any other system. HDFS does not provide metadata about the structure of the data stored in the HDFS. Using SAS, you can apply SAS formats and automatically discover the structure of any data contained in the HDFS.

Access Data in HDFS Using SAS/ACCESS® Interface to Hadoop. The second technique available to interact with files stored in HDFS uses the new SAS/ACCESS for Hive engine. The new engine provides libname access to any data stored in Hadoop. It uses the SAS Metadata Server to provide additional control and manageability of resources in Hadoop. The engine is designed to use the Hadoop Hive language. Using Hive, SAS can treat comma-separated or other structured files as tables, which can be queried using SAS Data Integration Studio, and ETL can be built using these tables as any other data source. Figure 2 shows a SAS Data Integration Studio job performing a join using Hadoop Hive. Notice the indicators on the top right of the tables that show the tables to be Hadoop data.

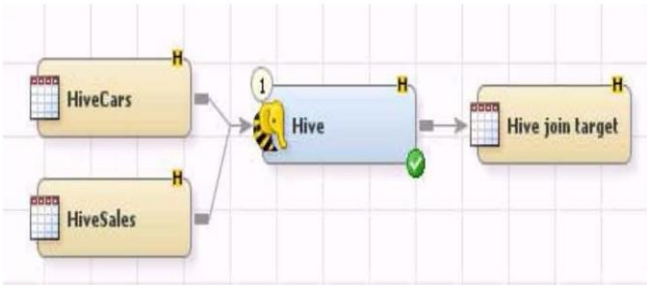


Figure 2: SAS Data Integration.

g) Computation in Hadoop

SAS Data Integration Studio provides a series of transformations shown in Figure 3 that are useful for working with data in Hadoop.

Figure 3: Hadoop transforms that are available in SAS Data Integration Studio

More details of these transforms are shown in Table 2.

Transform	Function
Hadoop Container	Convenience transformation allowing multiple Hadoop programs to be bundled into one transformation.
Hadoop File Writer	Move a structured file in the local system to a file in HDFS.
Hadoop File Reader	Move a file in HDFS to a structured file in the local system.
Pig	Choose from a set of available program templates in the Pig language that help you write ETL programs in Pig, and/or write your own Pig Latin program to manipulate and process data in Hadoop using the Pig language.
Hive	Choose from a set of available program templates in the Hive language that help write ETL programs in Hive, and/ or write your own Hive code to query, subset, filter or otherwise process Hadoop data using the Hive language
MapReduce	Choose a Java jar file containing MapReduce programs to be submitted to the Hadoop system.
Transfer from Hadoop	Transfer one or more files in the HDFS to the local system.
Transfer to Hadoop	Transfer one or more files on the local system to the Hadoop HDFS.

Table 2: Transform details

h) SAS® Data Management and SAS® LASRTM

The SAS LASR Analytic Server provides an in-memory, distributed computational system similar to Hadoop. SAS LASR is ideal for analytic algorithms for which the MapReduce paradigm is not well-suited. As an in-memory server, you still need to feed data to SAS LASR, and SAS Data Integration Studio simplifies this process. You register tables using the new SAS/ACCESS to the SAS LASR engine and then SAS Data Integration Studio can be used to perform a diverse set of tasks on SAS LASR data, just like it would for other data sources. Figure 4 shows a SAS LASR table being loaded with SAS Data Integration.

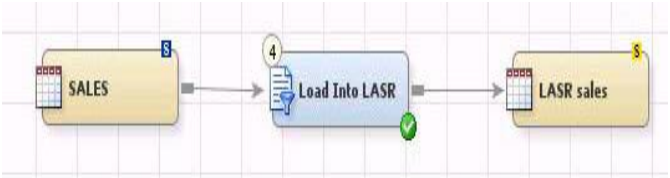


Figure 4: SAS LASR Table.

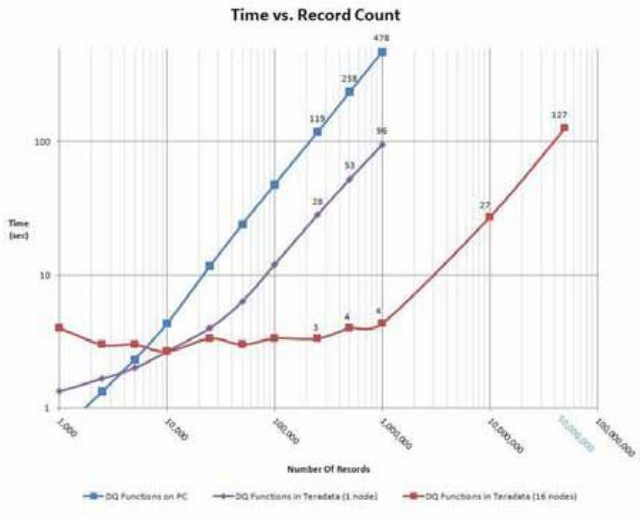
SAS LASR does not support joins or pushdown optimization as other databases do. Therefore, if you have data that needs to be joined or modified, you need to perform the data manipulation prior to loading the data into SAS LASR. You can do this in SAS using standard PROC SQL; or, if your data is already in Hadoop, you might want to directly perform the joins in Hadoop. You can create joins in Hadoop using one of the SAS Data Integration Studio Hadoop transforms. There are examples and templates available to help you build your code. Once you have completed the data preparation stage in Hadoop, you can convert the Hadoop files or tables to SAS LASR format using the Convert to SAS LASR template available in the Pig transform.

i) SAS® In-Database Data Quality

SAS, by means of the SAS/ACCESS technologies and accelerator products, has been optimized to push down computation to the data. By reducing data movement, processing times decrease and users are able to more efficiently use, compute resources and database systems. SAS/ACCESS engines already do implicit pass through to push joins, where clauses, and even Base SAS procedures such as SORT, TABULATE and other operations down to databases. SAS Scoring Accelerator and SAS Analytics Accelerator provide additional capabilities by providing a SAS Embedded Process that actually runs SAS code in a target database system, enabling orders of magnitude performance improvements in predictive model scoring and in the execution of some algorithms. SAS has added the ability to push data quality capabilities into the database. The SAS Data Quality Accelerator for Teradata enables the following data quality operations to be generated without moving data by using simple function calls:

- Parsing.
- Extraction.
- Pattern Analysis.
- Identification Analysis.
- Gender Analysis.
- Standardization.
- Casing.
- Matching.

Example performance improvements are indicated in the graph in Figure 6. Both scales are logarithmic. For example, we can see that data quality functions were performed on 50 million records in just more than two minutes on a 16-node Teradata cluster, while a PC was only able to process approximately 200,000 records in the same amount of time. The graph shows that performance improvements scale linearly, which means that as you add more nodes and processing power, the performance of your in-database data quality programs continues to improve.



j) Data Federation and Big Data

Data federation is a data integration capability that allows a collection of data tables to be manipulated as if they were a single table, while retaining their existing autonomy and integrity. It differs from traditional ETL/ELT methods because it pulls only the data needed out of the source system. Figure 7 is an illustration of the differences between traditional ETL/ELT and data federation.

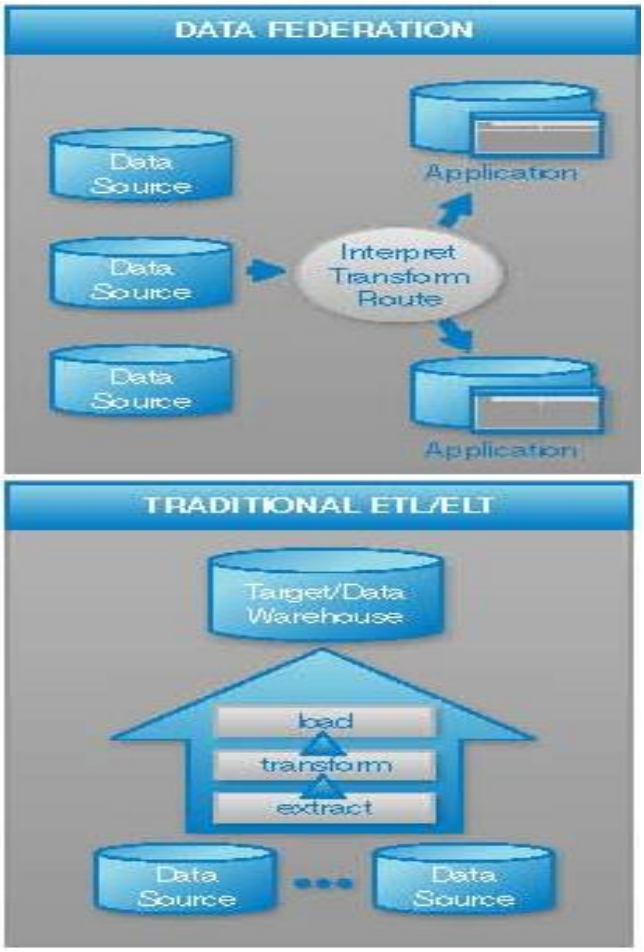


Figure 6: The differences between ETL and data federation

Data federation is ideally suited when working with big data because the data federation technique allows you to work with data stored directly in the source systems. Using data federation, you only pull the subset of data that you need when you need it. The SAS Federation Server is the heart of the SAS data federation capabilities. Using the SAS Federation Server you can combine data from multiple sources, manage sensitive data through its many security features and improve data manipulation performance through in-database optimizations and data caching. The server has fully threaded I/O, push-down optimization support, in-database caching, many security features (including row-level security), an integrated scheduler for managing cache refresh, a number of native data access engines for database access, full support for SAS data sets, auditing and monitoring capabilities, and a number of other key features. Using the SAS Federation Server, you can gain centralized control of all your underlying data from multiple sources. Figure 7 is a high-level overview of the SAS Federation Server.

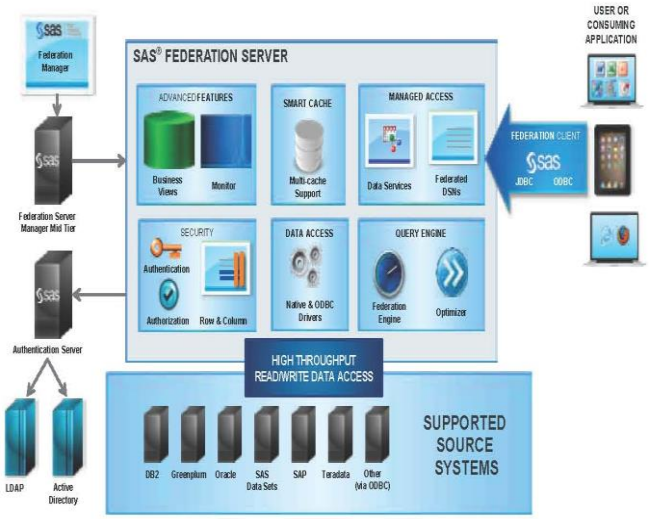


Figure 7: Overview of the SAS Data Federation Server architecture

There are a number of big data scenarios for which the SAS Federation Server is ideally suited. The following use cases illustrate some of these scenarios.

Data Federation Server Use Case 1: - Data Is Too Sensitive

In this scenario, illustrated in Figure 9, data is owned by organizations that do not want to grant direct access to their tables. The data may be owned by organizations that charge for each access, or is deemed mission critical. Users are not allowed to go directly against the actual tables. SAS Federation Server provides an ideal answer to this problem because it funnels the data access through the federation server itself, so multiple users do not have or need access to the base tables. A data cache can be optionally inserted into the result stream, so that even if the underlying tables are not accessible (for example, if the source system is down), data can still be supplied to users. Also, the federation server provides a single point for managing all security so users do not have to be granted direct access to the underlying tables.

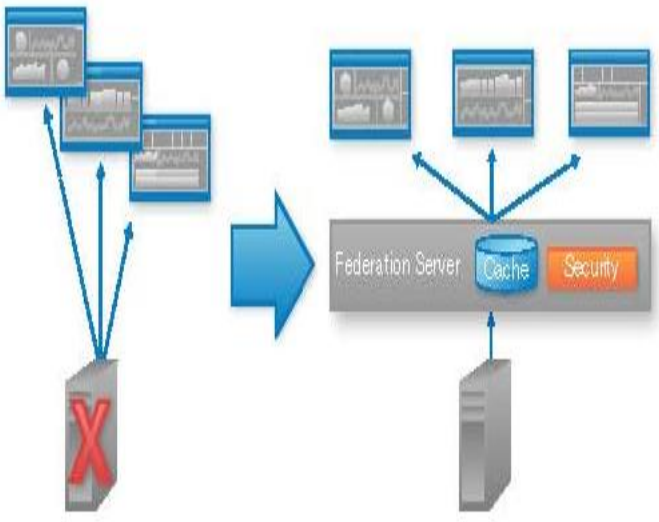


Figure 9: Example of federation scenario when data is too sensitive

Data Federation Use Case 2: - Data Is Too Diverse

In this use case, illustrated in Figure 10, the data is stored in multiple source systems that all have different security models, duplicate users and different permissions. This makes it hard to control permissions in a consistent way and requires every application to be customized to handle every data source. If there are changes needed (for example, if a user has to be removed or added to a new system), each application must be updated to handle the change. In a large system with a lot of data, this can become increasingly difficult to manage, especially over time. SAS Federation Server solves this problem. All the data access security can be managed singly in the SAS Federation Server, so multiple users do not go through to the base tables and security and permissions are managed in a single place for all target applications. In addition, by using the optional data cache, you can provide access to data for multiple users without having to recalculate the result sets every time for each individual user. This adds to system efficiency.

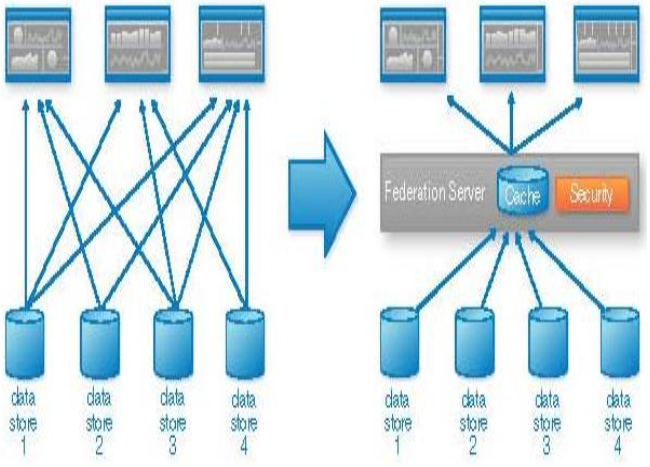


Figure 10: Example of federation scenario when data is too diverse.

Data Federation Use Case 3: - Data Is Too Ad Hoc

When data is changing frequently, constant updates are needed to maintain integration logic. It becomes difficult to make a repeatable integration process, especially if there are many data integration applications that need access to the same data. The application logic for accessing the data must be distributed into each application, and any changes in the data require corresponding changes in every application. As data grows in volume and complexity, and the number of applications that need to have access to the data grows, it becomes increasingly difficult to manage all the distributed data access logic in all the various applications, especially if the data is frequently changing.

SAS Federation Server solves this use case well. Using SAS Federation Server, it is easy to insert new views or modify existing views to accommodate different applications and users without requiring changes to the underlying data sources or applications. SAS Federation Server provides a single point of control for all integration logic. Plus, since the data access is managed through the SAS Federation Server, data integration as well as security and permissions are managed in a single place for all target applications. Figure 11 is an illustration of this use case.

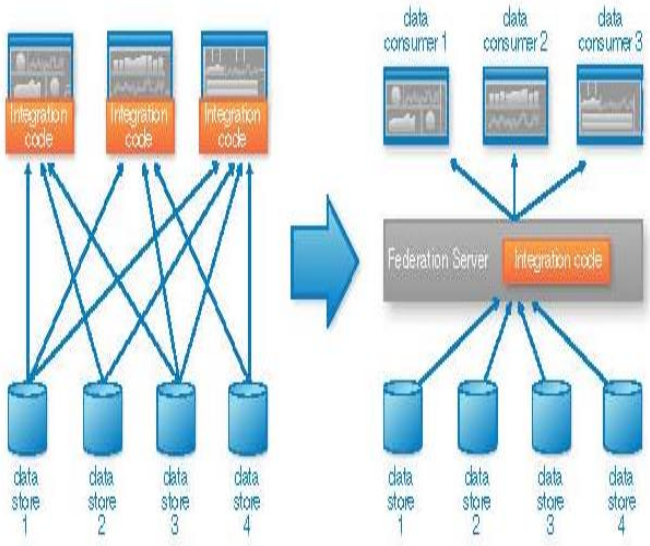


Figure 11: Example of federation scenario when data is too ad hoc.

SAS Federation Server supports access by using ODBC, JDBC or through the SAS/ACCESS to SAS Federation Server libname engine. A server management Web client is available for administering and monitoring the server. From this interface you can monitor multiple servers from the Web client interface. Figure 12 is an example of the manager Web client.

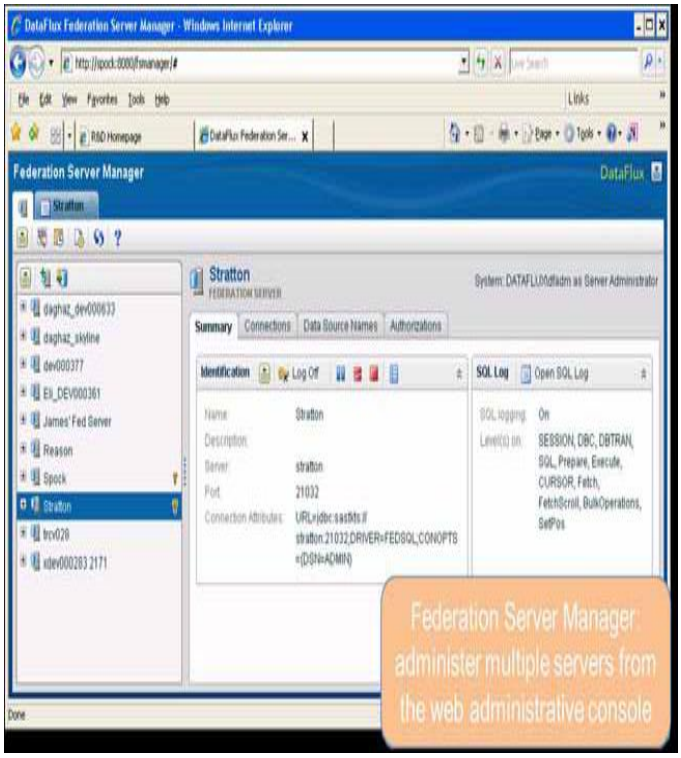


Figure 12: Federation Server Manager overview.

XV. CONCLUDING REMARKS

Big-Data has unveiled a new horizon of researcher regarding data analysis that is going to commence a radical change in the twenty-first century. Literally 'Big-Data" means that has no defined boundary in case of velocity, volume and variety

i.e. complexity. More specifically, it will transcend the limits of human experience. This compact thesis paper does not claim that "Big-data" will be in the driving seat of all development works in the years to come, rather an endeavor our has been made to build a caution so that human race can save and protect themselves by understanding the information they increasingly produce in digital forms.

The analysis of "Big-Data" has evolved a new branch of science in ICT field. It may be called a nascent science. In our perspective, it is very much new. The sources of generation are so rapidly changing it has become every much difficult to cope with. Human being has an inborn tendency to play with challenges always, here is the case for 'Big-Data also. It is a well-known maxim that everything whether that is good a bad must have merits and demerits also.

The misuse of 'Big-data' may be in micro and macro sectors. The threat to a company or group of people may be macro threats, whether for a lone person it is called micro threat i.e. breach of personal security. To drive maximum benefit there must be provided public Collaboration. In fine, it may be concluded that in this marrow thesis paper an effort has been made to identify the sources, to find out some analytical process involved, to determine the outcome of the analysis and threat to breach of personal security in analyzing "Big-data". In conclusion it may be motioned that further researches may be conducted in the new arena like 'Big-data' for which this thesis paper may be used as reference in the work, if desired.

REFERENCES:

[1] **Big Data for Development: Challenges & Opportunities**
U N Global Pulse
370 Lexington Ave, Suite 1707
New York, New York 10017
E-mail: info@unglobalpulse.org
http://unglobalpulse.org/

[2] Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute

[3] **EFFECTS OF BIG DATA ANALYTICS ON ORGANIZATIONS' VALUE CREATION**
Niels Mouthaan, Business Information Systems
University of Amsterdam

[4] **Bigtable: A Distributed Storage System for Structured Data**
Fay Chang, Jeffrey Dean, Sanjay Ghemawat
Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber
ffay,jeff,sanjay,wilsonh,kerr,m3b,tushar,_kes,gruberg@google.com

[5] **Inside "Big Data Management": Ogres, Onions, or Parfaits**

[6] **MAD Skills: New Analysis Practices for Big Data**
Jeffrey Cohen, Greenplum, Brian Dolan
Fox Audience Network

[7] Big Data Meets Big Data Analytics
Three Key Technologies for Extracting Real-Time Business Value from the Big Data,That Threatens to Overwhelm Traditional Computing Architectures

[8] Questionnaire on use of Big Data
http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html

[9] Why Big Data for Bangladesh?
A small-talk on Big Data
Fokhruz Zaman, SID Strategy Workshop, 7th Sep, 2013, Dhaka

[10] **Historical Perspective of BBS Data**
Workshop of WADM 2013, CSE, BUET, June 28-29, 2013
Keynote Speech
Md. Nazrul Islam, Senior System Analyst
Md. Karamat Ali, Senior program
Mr. Chandra Shekhur Roy, Senior Maintenance Engineer

[11] Big Data Management and Analytics
Workshop on Advanced Data Management (WAMD)
CSE, BUET, June 28-29, 2013
Keynote Speech
Dr. Latifur Khan
Univ. of Texas at Dallas
Dr. Mohammad Mehedy Masud
United Arab Emirates University

[12] Alexander, Malcolm and Nancy Rausch. 2013. *What's New in SAS® Data Management*. Proceedings of the SAS Global Forum 2013 Conference. Cary, NC: SAS Institute Inc. Available at support.sas.com/resources/papers/proceedings13/070-2013.pdf.

[13] Rausch, Nancy, et al. 2012. *What's New in SAS® Data Management.* Proceedings of the SAS Global Forum 2012 Conference. Cary, NC: SAS Institute Inc. Available at support.sas.com/resources/papers/proceedings12/110-2012.pdf.

[14] Rausch, Nancy and Tim Stearn. 2011. *Best Practices in Data Integration: Advanced Data Management.* Proceedings of the SAS Global Forum 2011 Conference. Cary, NC: SAS Institute Inc. Available at support.sas.com/resources/papers/proceedings11/137-2011.pdf.

[15] Best Practices in SAS® Data Management for Big Data

[16] Hazejager, Wilbram and Pat Herbert. 2011. *Innovations in Data Management* – Introduction *to Data Management Platform.* Proceedings of the SAS Global Forum 2011 Conference. Cary, NC: SAS Institute Inc. Available at support.sas.com/resources/papers/proceedings11/137-2011.pdf.

[17] Hazejager, Wilbram and Pat Herbert. 2011. *Master Data Management, the Third Leg of the* Data *Management Stool: a.k.a. the DataFlux® qMDM Solution.* Proceedings of the SAS Global Forum 2011 Conference. Cary, NC: SAS Institute Inc. Available at support.sas.com/resources/papers/proceedings11/146-2011.pdf.

Syed Jamaluddin Ahmad, achieved Bachelor of Science in Computer Science and Engineering (BCSE) from Dhaka International University, Masters of Science in Computing Science Associates with research: Telecommunication Engineering from Athabasca University, Alberta, Canada and IT-Pro of Diploma from Global Business College, Munich, Germany. Presently Working as an Assistant Professor, Computer Science and Engineering, Shanto-Mariam University of Creative Technology, Dhaka, Bangladesh. Formerly, was head of the Department of Computer Science & Engineering, University of South Asia from 2012-2014, also Lecturer and Assistant Professor at Dhaka International University from 2005-2007 and 2011-2012 respectively and was a lecturer at Loyalist College, Canada, was Assistant Professor at American International University, Fareast International University, Royal University, Southeast University and Many more. He has **alread**y 15th international publications, 12th seminar papers, and conference articles. He is also a founder member of a famous IT institute named Arcadia IT (www.arcadia-it.com). Achieved Chancellor's Gold Crest in 2010 for M.Sc. in Canada and Outstanding result in the year of 2005. and obtained "President Gold Medal" for B.Sc.(Hon's). Best conductor award in Germany for IT relevant works. Membership of "The NewYork International Thesis Justification Institute, USA, British Council Language Club, National Debate Club, Dhaka, English Language Club and DIU . Developed projects: Mail Server, Web Server, Proxy Server, DNS(Primary, Secondary, Sub, Virtual DNS), FTP Server, Samba Server, Virtual Web Server, Web mail Server, DHCP Server, Dial in Server, Simulation on GAMBLING GAME Using C/C++, Inventory System Project, Single Server Queuing System Project, Multi Server Queuing System Project, Random walk Simulation Project, Pure Pursuit Project (Air Scheduling), Cricket Management Project, Daily Life Management Project, Many Little Projects Using Graphics on C/C++, Corporate Network With Firewall Configure OS:LINUX (REDHAT) Library Management Project Using Visual Basic, Cyber View

Network System:Tools:Php OS: Windows Xp Back-end: My SQL Server, Online Shopping: Tools: Php, HTML, XML. OS:Windows Xp, Back-end: My SQL and Cyber Security" Activities-'Nirapad Cyber Jogat, Atai hok ajker shapoth'-To increase the awareness about the laws, 2006 (2013 amendment) of Information and Communication and attended Workshop on LINUX Authentication"-Lead by- Prof. Andrew Hall, Dean, Sorbon University, France, Organized By- Athabasca University, CANADA, April, 2009. His areas of interest include Data Mining, Big Data Management, Telecommunications, Network Security, WiFi, Wimax, 3g, 4g network, UNIX, LINUX Network Security,Programming Language(C/C++ or JAVA), Database (Oracle), Algorithm Design, Graphics Design & Image Processing and Algorithm Design.

Roksana Khandoker Jolly, achieved Bachelor of Science in Computer Science and Engineering (BCSE) from United International University, Masters of Science in Computer Science and Engineering from University of South Asia. Presently Working as a Senior Lecturer, Computer Science and Engineering, University of South Asia, Dhaka, Bangladesh. Formerly, was also a lecturer at different poly-technique institutes. She has 4th international journals and attended different international and national conferences. She is the Chairman of the famous IT institute named Arcadia IT and Chairman of Brighton International Alliance. Her areas of interest include Data Mining, Big Data Management, Telecommunications, Network Security, WiFi, Wimax, 3g and 4g network