ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

The Future Revolution on Big Data

Abdul Raheem Syed[1], Kumar Gillela[2], Dr. C. Venugopal[3]

Assistant Professor, ECM, SNIST, Hyderabad, India [1]

Assistant Professor, CSE, MRITS, Hyderabad, India [2]

Professor, ECM, SNIST, Hyderabad, India [3]

Abstract: Big data is a popular, but poorly defined marketing buzzword, that describe the exponential growth, availability and use of information, both structured and unstructured. The big data trend and how it can serve as the basis for innovation, differentiation and growth. Looking at big data is that it represents the large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. Examples of this data include high-volume sensor data and social networking information from web sites such as Google, Face Book, LinkedIn, Yahoo, Amazon and Twitter. The exponential growth in the amount of biological data needed for data management, analysis and accessibility.

Keywords: Social Media, structured data, unstructured data, Hadoop, NoSQL, Data Warehousing

I. Introduction

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it. The value of big data to an organization falls into two categories: analytical use, and enabling new products. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analysing shoppers' transactions, social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports.

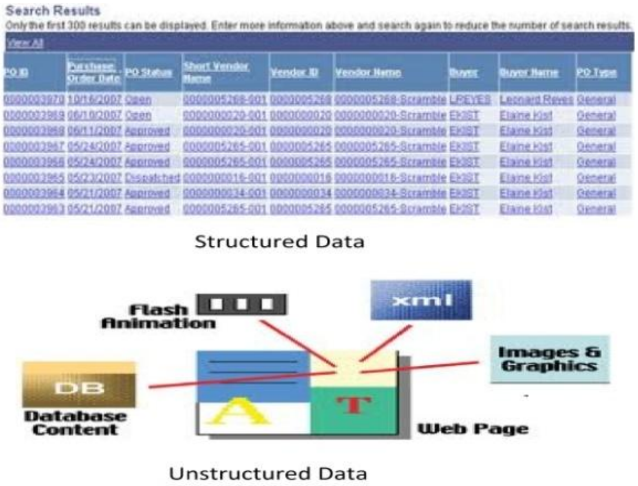
1.1Types of Data

Big Data is made of structured and unstructured information. The term structured data and unstructured data refers to that is identifiable based on it is organized in a structure or not (See figure1).

1.1.1 Structured data

The most common form of structured data or structured data records is a database where specific information is stored based on a methodology of columns and rows. Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers. Relational databases and spreadsheets are examples of structured data.

In contrast, unstructured data has no identifiable structure. Structured information is the data in databases and is about 10% of the story.



Search Results

Only the first 300 results can be displayed. Enter more information above and search again to reduce the number of search results.

View All

PO ID	Purchase Order Date	PO Status	Short Vendor Name	Vendor ID	Vendor Name	Buyer	Buyer Name	PO Type
0000003970	10/16/2007	Open	0000005268-001	0000005268	0000005268-Scramble	LREYES	Leonard Reyes	General
0000003969	06/10/2007	Open	0000000020-001	0000000020	0000000020-Scramble	EKIST	Elaine Kist	General
0000003968	06/11/2007	Approved	0000000020-001	0000000020	0000000020-Scramble	EKIST	Elaine Kist	General
0000003967	05/24/2007	Approved	0000005265-001	0000005265	0000005265-Scramble	EKIST	Elaine Kist	General
0000003966	05/24/2007	Approved	0000005265-001	0000005265	0000005265-Scramble	EKIST	Elaine Kist	General
0000003965	05/23/2007	Dispatched	0000000016-001	0000000016	0000000016-Scramble	EKIST	Elaine Kist	General
0000003964	05/21/2007	Approved	0000000034-001	0000000034	0000000034-Scramble	EKIST	Elaine Kist	General
0000003963	05/21/2007	Approved	0000005265-001	0000005265	0000005265-Scramble	EKIST	Elaine Kist	General

Figure1. Structured and Unstructured Data

1.1.2 Unstructured data

The term Unstructured Data refers to information that either does not have a pre-defined data model and/or does not fit well into relational tables. The term unstructured data refers to any data that has no identifiable structure. Unstructured information is 90% of Big Data and is 'human information' like emails, videos, tweets, Face book posts, call centre conversations, closed circuit TV footage, mobile phone calls, website clicks. Big Data is only getting bigger 90% of the data in the world today was created within the last two years. While each individual document may contain its own specific structure or formatting that based on the software program used to create the data, unstructured data may also be considered “loosely structured data” because the data sources do have a structure but all data within a dataset will not contain the same structure.

II. WHAT DOES BIG DATA LOOK LIKE

In 2012, Gartner formalized their Big Data definition as a "3V" framework - high Volume, high Velocity and high Variety information asset, requiring new forms of processing to enable enhanced decision making, insight discovery and process optimization. The IBM adds a fourth "V" of Veracity to add trust and noise filtering to the challenge of Big Data analysis. Input data to big data systems could be chatter from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, MP3s of rock music, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data, the list goes on (See figure2).

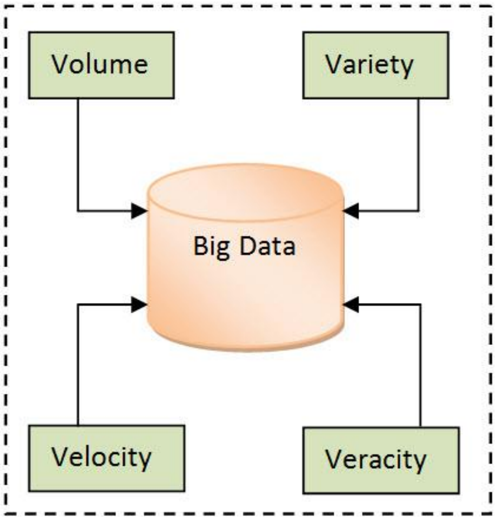


Figure 2 Dimensions of Big data

2.1 Volume

There are many factors contribute to the increase in data volume – transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, etc. In the past, excessive data volume created a storage issue. But with today's decreasing storage costs, other issues emerge, including how to determine relevance among the large volumes of data and how to create value from data that is relevant. This volume presents the most immediate challenge to conventional IT structures. It calls for scalable storage, and a distributed approach to querying. Many companies already have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it.

2.2 Variety

Big data is any type of data; today data comes in all types of formats – from traditional databases to hierarchical data stores created by end users and OLAP systems, to text documents, email, meter-collected data, video, audio, stock ticker data and financial transactions. By some estimates, 80 percent of an organization's data is not numeric. But it still must be included in analyses and decision making. Different browsers send different data, users withhold information, and they may be using differing software versions or vendors to communicate with you. A common use of big data processing is to take unstructured data and extract ordered meaning, for consumption either by humans or as a structured input to an application.

2.3 Velocity

Velocity "means both how fast data is being produced and how fast the data must be processed to meet demand". Problems previously restricted to segments of industry are now presenting themselves in a much broader setting. Specialized companies such as financial traders have long turned systems that strive with fast moving data to their advantage. The Internet and mobile era generating a data flow back to the provider. The Retailers who are able to quickly utilize that information, by recommending additional purchases, for instance, gain competitive advantage. The Smartphone era increases again the rate of data inflow, as consumers carry with them a streaming source of reallocated images and audio data.

2.4 Veracity

Veracity deals with uncertain or imprecise data. In traditional data warehouses there was always the assumption that the data is certain, clean, and precise. That is why so much time was spent on Master Data Management, Metadata management, Identity Insight/Assertion, etc. However, when we start talking about social media data like Tweets, Facebook posts, etc. how much faith should we put in the data. This data can be used as a count toward your sentiment, but you would not count it toward your total sales and report on that. Due to the velocity of data like stock trades, machine/sensor generated events, you cannot spend the time to “cleanse” it and get rid of the uncertainty, so you must process it as is understanding the uncertainty in the data. And as you bring multi-structured data together, determining the origin of the data, and fields that correlate becomes nearly impossible.

III. THE BIG DATA TREND AND GROWTH

Companies worldwide are already implementing their plans to store and analyze big data. The majority of IT decision-makers surveyed work for companies in the manufacturing, healthcare, financial services industries, and other companies. More than half of is 64 percent are in the manufacturing, 29 percent in production, 19 percent healthcare and 16 percent financial services industries, but only 13 percent of those have a big data solution that is fully implemented (See figure3).

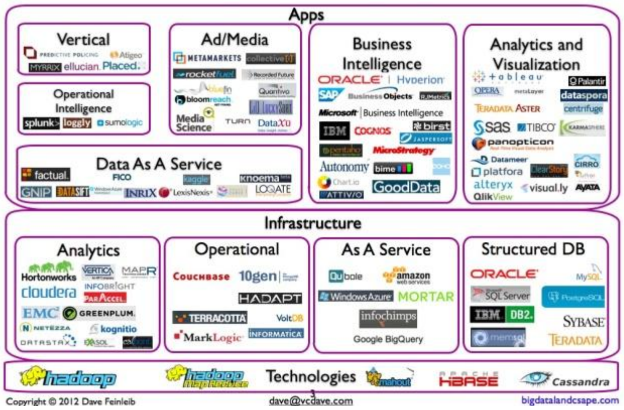


Figure 3 Trends of Big Data

Nearly half rated managing data growth as extremely challenging when managing their companies' business intelligence (BI), 41 percent rated integrating Business Intelligence tools as extremely challenging, and 40 percent rated the need for tools to analyze data and glean insights as extremely challenging. In practice/Practical we have explored the nature of big data, and surveyed the landscape of big data from a high level. As usual, when it comes to deployment there are dimensions to consider over and above tool selection.

3.1 Cloud

The majority of big data solutions are now provided in three forms: software-only, as an appliance and cloud-based. Decisions between which routes to take will depend, among other things, on issues of data locality, privacy and regulation, human resources and project requirements.

3.2 Big data is big

It is a fundamental fact that data that is too big to process conventionally is also too big to transport anywhere. Information Technology is undergoing an inversion of priorities: it is the program that needs to move, not the data. Financial trading systems crowd into data centers to get the fastest connection to source data, because that millisecond difference in processing time equates to competitive advantage.

3.3 Big data is messy

Big data practitioners consistently report that 80% of the effort involved in dealing with data is cleaning it up in the first place, as Pete Warden observes in his Big Data Glossary: "I probably spend more time turning messy source data into something usable than I do on the rest of the data analysis process combined." Because of the high cost of data acquisition and cleaning, it's worth considering what you actually need to source yourself. Quality can of course be variable, but will increasingly be a benchmark on which data marketplaces compete.

3.4 Culture

The phenomenon of big data is closely tied to the emergence of data science, a discipline that combines math, programming and scientific instinct. The benefits from big data means investing in teams with this skill set, and surrounding them with an organizational willingness to understand and use data for advantage.

IV. DATA FROM TRILLIONS OF SENSORS AND INTERNET

It is projected that within a few years we will have to talk in Bronto bytes (10^{27}) when we discuss data coming from sensors. The Internet of Things, or Machine to Machine (M2M) communication, connects billions of devices with each other and thereby generates an unfathomable amount of data. In 2020, 40% of all data in the world will be M2M data. This data of course has to be processed, stored, analyzed and visualized to have any meaning and to drive your business cases. Sensor data, or M2M data, is data from readings made by machine sensors that measure pre-set conditions at regular intervals. Examples of the data that is collected include log data, geo location data, CPU utilization, temperature, rules, etc. These data can be linked to Key Performance Indicators that send an alarm when a certain threshold is passed and an action is required.

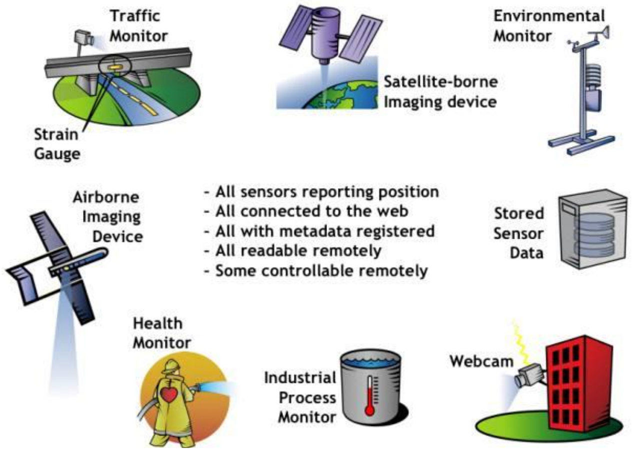

Traffic Monitor
Strain Gauge
Satellite-borne Imaging device
Environmental Monitor
Airborne Imaging Device
- All sensors reporting position
- All connected to the web
- All with metadata registered
- All readable remotely
- Some controllable remotely
Stored Sensor Data
Health Monitor
Industrial Process Monitor
Webcam

Figure 4 Big data on Sensors

V. BIG DATA AND BIGGER OPPORTUNITY

Even though "Big Data" has now been around for a few years, the opportunities for startups seem to keep growing, just as the amount of data keeps growing. According to IBM, companies have captured more data in the last two years than in the previous 2000 years. This data comes from sensors, social media posts, digital pictures and videos, purchase transactions, everywhere. Every day, we create 2.5 quintillion bytes of data, much of it unstructured and far beyond the capability of conventional databases. Hence one segment of the opportunity is the need for new database technologies, like Hadoop, a distributed file system originally designed for indexing the Web. Data capacity is measured in petabytes (1000 terabytes), or soon even yottobytes (10^{24}).

By any definition, the opportunities from Big Data have the potential to create a next wave of successful technology companies that could change the way we all live and work. I will summarize here some of the key business domains with large opportunities, based on a McKinsey Global Institute study and other sources:

1. Targeted marketing. Big data can mean big profits. By transforming a single shopper's path into data points, companies can see how you move through a store, and how that tracks with sales.
2. Protecting the environment. Analyzing the massive sets of data available on toxic emissions and weather patterns can help us understand environmental threats on a systemic level. We just need big data tools to do the analysis.
3. Health care in the U.S. Health care is a large and important segment with huge data challenges, mostly not structured or linked. It has multiple and varied stakeholders, including the pharmaceutical and medical products industries, providers, payers, and patients. Each of these has different interests and incentives, with real money to spend.
4. Social media and web data. Social media postings and e-Commerce transactions are just a couple of the sources of external data that are of great interest to many companies. Facebook now exceeds a billion users posting, the Internet has 650 million websites, and there are 200 e-Commerce items ordered every second. That's a lot of data to analyze.
5. Automated device generated data. Another big data opportunity is the vast amount of sensor data, machine generated data that exists and is growing at an exponential pace as more machines become internet-enabled. Examples include data generated from traffic cameras, parking meters, toll collection, and black boxes in airplanes.
6. Scientific research in overdrive. Data has long been the cornerstone of scientific discovery, with big data, and the big computing power necessary to process it, research can move at an exponentially faster clip.
7. Global personal location tracking. Processing personal location data is a domain that includes child safety, law enforcement, tracking terrorists, and travel planning.
8. Global manufacturing. Manufacturing is a global industry with complex and widely distributed value chains and a large amount of data available. This domain therefore offers opportunities at multiple points in the value chain, from bringing products to market and research and development (R&D), RFID tracking, to after-sales services.
9. Data is the new weapon of defense. The traditional battlefield has dissolved into thin air. In the Big Data era, information is the deadliest weapon and leveraging massive amounts of it is this era's arms race. But current military tech is buckling under the sheer weight of data collected from satellites, unmanned aircraft, and intercepted messages.
10. Public sector administration. The public sector is another large part of the global economy facing tremendous pressure to improve its productivity. Governments have access to large pools of digital data but have hardly begun to take advantage of the powerful ways in which they could use this information.

These large opportunities are why the Big Data market is expected to be worth $50 billion by 2017, up from $5 billion today in products, with services likely to represent another 40 percent of that figure. Compared to the yet another dating site proposal that I see most often, Big Data is an exciting opportunity for entrepreneurs and investors alike.

VI. SOCIAL NETWORK'S AND BIG DATA

The use of big data and social networks for public health purposes has gained attention in recent years with the spread of social networks and with the availability of on-line data systems. Analysis and data exchange in health systems can help prevention and monitoring of diseases. Social media contains useful information about customers, prospects and competitors, but surfacing that intelligence requires a focused approach. Big data by the sources, including

Figure 5 Big data on Social Media

6.1 Social Networking and Media: There are currently over 700 million Facebook users, 250 million Twitter users and 156 million public blogs. Each Facebook update, Tweet, blog post and comment creates multiple new data points - structured, semi-structured and unstructured - sometimes called Data Exhaust.

6.2 Mobile Devices: There are over 5 billion mobile phones in use worldwide. Mobile devices, particularly smart phones and tablets, also make it easier to use social media and other data-generating applications.

6.3 Internet Transactions: Billions of online purchases, stock trades and other transactions happen every day, including countless automated transactions. Each creates a number of data points collected by retailers, banks, credit cards, credit agencies and others.

6.4 Networked Devices and Sensors: Electronic devices including servers and other IT hardware, smart energy meters and temperature sensors. All create semi-structured log data that record every action.

6.5 Big Data is the Future of Healthcare: With big data poised to change the healthcare, organizations need to understanding this phenomenon and realizing the envisioned benefits.

VII. BIG DATA PROBLEMS ORBIG DATA OPPOTUNITY

The advent of the Web, mobile devices and other technologies has caused a fundamental change to the nature of data. Big Data has qualities that differentiate it from traditional corporate data. No centralized, highly structured and easily manageable. Now data is highly distributed, loosely structured, and increasingly large in volume. Specifically:

Volume – The amount of data created both inside corporations and outside the firewall via the web, mobile devices, IT infrastructure, and other sources is increasing exponentially each year.

Type – The variety of data types is increasing, namely unstructured text-based data and semi-structured data like social media data, location-based data, and log-file data.

Speed – The speed at which new data is being created. The need for real-time analytics to derive business value from it -- is increasing thanks to digitization of transactions, mobile computing and the sheer number of internet and mobile device users.

In digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society. For example, search engine companies such as Google, Yahoo!, and Microsoft have created an entirely new business by capturing the information freely available on the World Wide Web and providing it to people in useful ways. These companies collect trillions of bytes of data every day and continually add new services such as satellite images, driving directions, and image retrieval. Just as search engines have transformed how we access information, other forms of big data computing can and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations.

7.1 Big Data: Hadoop, Business Analytics

The old way traditionally, data processing for analytic purposes followed a fairly static blueprint. Business enterprises create modest amounts of structured data with stable data models via enterprise applications like CRM, ERP and financial systems. Data integration tools are used to extract, transform and load the data from enterprise applications and transactional databases to a staging area where data quality and data normalization occur and the data is modelled into neat rows and tables. The modelled, cleansed data is then loaded into an enterprise data warehouse. This routine usually occurs on a scheduled basis, usually daily or weekly, sometimes more frequently.

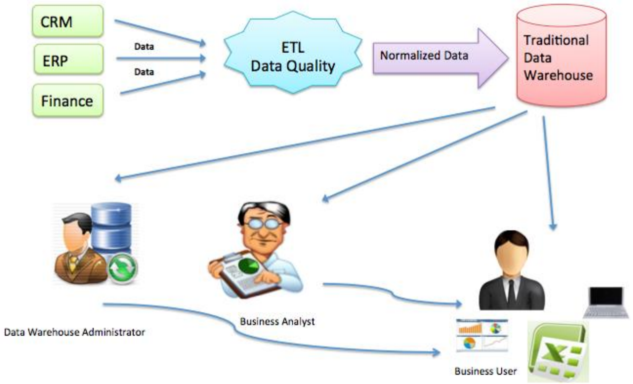


Figure 6 Data Processing and Analytics

7.2 New Approaches Processing

There are number of approaches to processing and analyzing Big Data, but most have some common characteristics.

7.2.1 Hadoop

Hadoop is an open source framework for processing, storing and analyzing massive amounts of distributed unstructured data. Originally created by Doug Cutting at Yahoo!, Hadoop was inspired by MapReduce, a user-defined function developed by Google in early 2000s for indexing the Web. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Hadoop clusters run on inexpensive commodity hardware so projects can scale-out without breaking the bank. Hadoop is now a project of the Apache Software Foundation, where hundreds of contributors continuously improve the core technology. Hadoop breaks up Big Data into multiple parts so each part can be processed and analyzed at the same time.

7.2.2 NoSQL

A related new style of database called NoSQL (Not Only SQL) has emerged to, like Hadoop, process large volumes of multi-structured data. However, whereas Hadoop is adept at supporting large-scale, batch-style historical analysis and NoSQL databases are aimed, for the most part at serving up discrete data stored among large volumes of multi-structured data to end-user and automated Big Data applications. This capability is sorely lacking from relational database technology, which simply can't maintain needed application performance levels at Big Data scale.

7.2.3 Hybrid Big Data Integration

So while some companies may specifically be looking for Hadoop or NoSQL data integration, others are looking for a comprehensive solution that will connect on the back end to multiple big data sources as well as their enterprise data stores.

An all-round big data integration offering will enable you to input, output, manipulate and report on data using Hadoop and NoSQL stores, including: Apache Cassandra, Hadoop HDFS, Apache Hive, Apache HBase and MongoDB. It should provide easy job orchestration across Hadoop, Amazon EMR, MapReduce, Pig scripts, NoSQL databases and traditional data stores

7.3 Big data impact on Data Warehousing

The "Big Data" movement has taken the information technology world by storm. Fueled by open source projects emanating from the Apache Foundation, the big data movement offers a cost-effective way for organizations to process and

Store large volumes of any type of data: Structured, semi-structured and unstructured

- Despite problems, Big Data makes it Huge
- Two markets for Big Data: Comparing Value Propositions
- Categorizing Big Data Processing Systems
- The New Analytical Ecosystem: Making Way for Big Data

VIII. CONCLUSIONS

New systems using big data will extend, and possibly replace, our traditional DBMS's. There is no question that there is enough data available that traditional database management systems will be defeat completely. Moving forward with big data systems, and that the best way is to start small and prove the benefits. While this is not much different from any other new technology, it might be an especially good strategy to apply to big data applications. Cloud computing may also prove valuable for big data. Currently available systems for health care domains and Social Media and Retail limits the functionalities due to the ever increases in demands. As a result the integration of new technologies is necessary to cope up with on-demand. In the future, further studies will be conducted for improving the big data platforms through theory and experiments.

XI. REFERENCES

1. Pantelopoulos, Nikolaos G, A survey on wearable sensor-Based Systems for health Monitoring and Prognosis, IEEE Tran Sys Man and Cybernetics, vol40(1), pp: 1-12, 2010.
2. Zhoung Liu,Dong-sheng Yang, Ding Wen and Wei-ming Zhang, Cyber-Physical-Social Systems for Command and Control, IEEE Intelligent Systems, Vol 26,no.4,pp 92-96,2011.
3. Lee Insup and Oleg Sokolsky, Medical Cyber Physical Systems, in Proc.of DAC, USA 2012.
4. Dean, Je_rey and Ghemawat, Sanjay, MapReduce: simpli_ed data processing on large clusters, OSDI,2004.
5. http://www.Google.com
6. Lee, Edward A, Cyber Physical Systems: Design Challenges, ISORC, pp 363-369, 2008.
7. Onella, Jukka- Pekka. "Social Networks and Collective Human Behavior." UN Global Pulse. 10 Nov. 2011.
8. Lohr, Steve. "The Age of Big Data." New York Times. 11 Feb, 2012.
9. Global Internet Usage by 2015 Alltop.
10. Toyama, Kentaro. "Can Technology End Poverty?" Boston Review. Dec 2010.
11. Global Monitoring Report 2009: A Development Emergency. Rep. Washington DC: International Bank for Reconstruction and Development/ the World Bank, 2009.
12. Laney, D. (2001), \3-D Data Management: Controlling Data Volume, Velocity and Variety, META Group Research Note, February 6.
13. Sala-i-Martin, X. (1997), \I Just Ran Two Million Regressions," American Economic Review, 87 (May), 187{183.
14. Tilly, C. (1984), \The Old New Social History and the New Old Social History," Review.