Big Data and its Implications

D. Dhana Sekagar,
Assistant Professor,
Department of Computer Applications,
SVCET, Chittoor.

S. Ismail Basha,
Assistant Professor**,**
Department of Computer Applications,
KMM Institute of Post Graduate Studies, Tirupati.

***Abstract-* “Big data” appears to have become a buzzword overnight. The term describes innovative techniques and technologies to capture, store, distribute, manage and analyze peta byte-or larger-sized datasets with high-velocity and diverse structures that conventional data management methods are incapable of handling. Big data has demonstrated the capacity to improve predictions, save money, boost efficiency and enhance decision - making in fields as disparate as traffic control, weather forecasting, disaster prevention, finance, fraud control, business transaction, national security, education, and health care**

Keywords: Big data platform, Dimension, Generating, Induction, Generating, Some Concepts Why Big Data.

INDUCTION

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, curation, storage, search, sharing, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

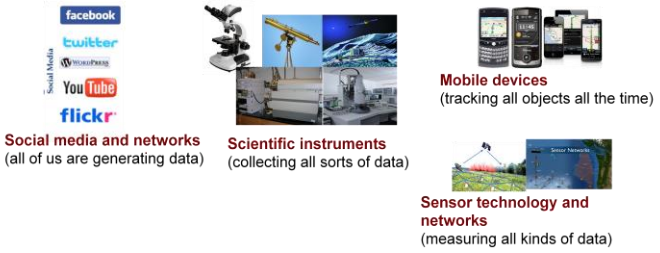
(or)

Put another way, big data is the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies.

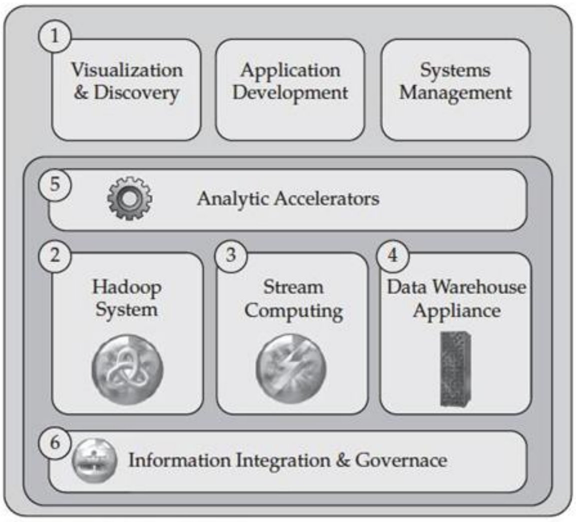
The four dimensions of Big Data

- Volume: Large volumes of data
- Velocity: Quickly moving data
- Variety: structured, unstructured, images, etc.
- Veracity: Trust and integrity is a challenge and a must and is important for big data just as for traditional relational DBs .

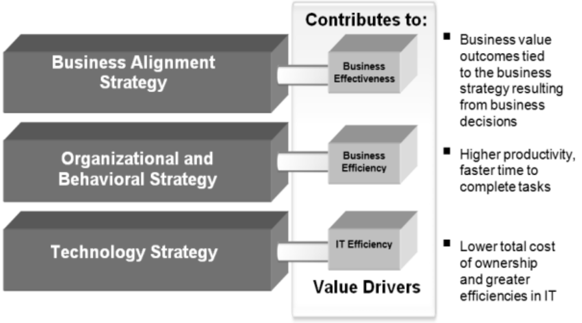
Who’s Generating Big Data


facebook
twitter
WordPress
Social Media
YouTube
flickr
Social media and networks
(all of us are generating data)
Scientific instruments
(collecting all sorts of data)
Mobile devices
(tracking all objects all the time)
Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion



Why Big Data and BI


Contributes to:
Business Alignment Strategy
Business Effectiveness
Organizational and Behavioral Strategy
Business Efficiency
Technology Strategy
IT Efficiency
Value Drivers
Business value outcomes tied to the business strategy resulting from business decisions
Higher productivity, faster time to complete tasks
Lower total cost of ownership and greater efficiencies in IT

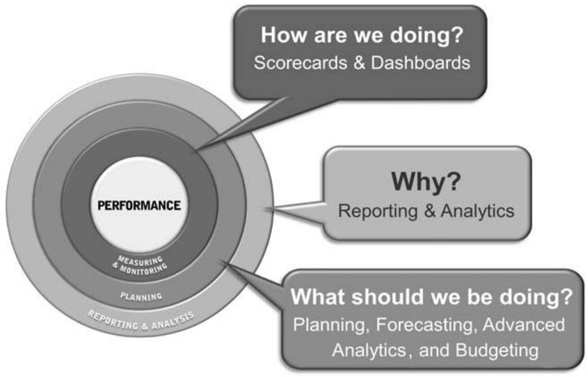
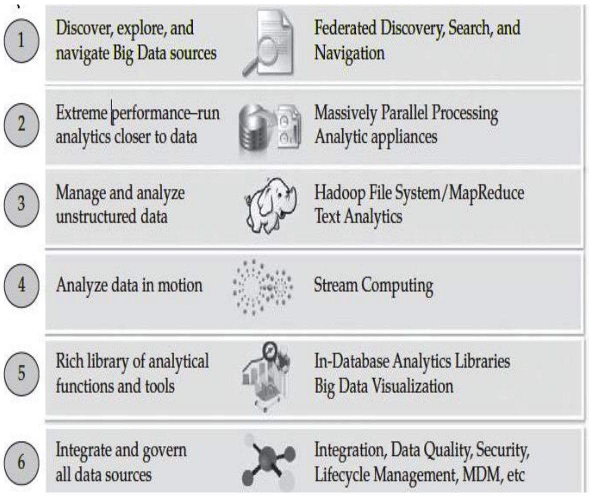


Figure : The Big Data platform Manifesto imperatives and underlying technologies



1	Discover, explore, and navigate Big Data sources	Federated Discovery, Search, and Navigation
2	Extreme performance–run analytics closer to data	Massively Parallel Processing Analytic appliances
3	Manage and analyze unstructured data	Hadoop File System/MapReduce Text Analytics
4	Analyze data in motion	Stream Computing
5	Rich library of analytical functions and tools	In-Database Analytics Libraries Big Data Visualization
6	Integrate and govern all data sources	Integration, Data Quality, Security, Lifecycle Management, MDM, etc

Figure : IBM's Big Data Platform

Figure --*The IBM Big Data platform*

Some concepts

- NoSQL (Not Only SQL): Databases that "move beyond" relational data models (i.e., no tables, limited or no use of SQL)
  - Focus on retrieval of data and appending new data (not necessarily tables)
  - Focus on key-value data stores that can be used to locate data objects
  - Focus on supporting storage of large quantities of unstructured data
  - SQL is not used for storage or retrieval of data
  - No ACID (atomicity, consistency, isolation, durability)

NoSQL

- NoSQL focuses on a schema-less architecture (i.e., the data structure is not predefined)
- In contrast, traditional relation DBs require the schema to be defined before the database is built and populated.
  - Data are structured
  - Limited in scope
  - Designed around ACID principles.

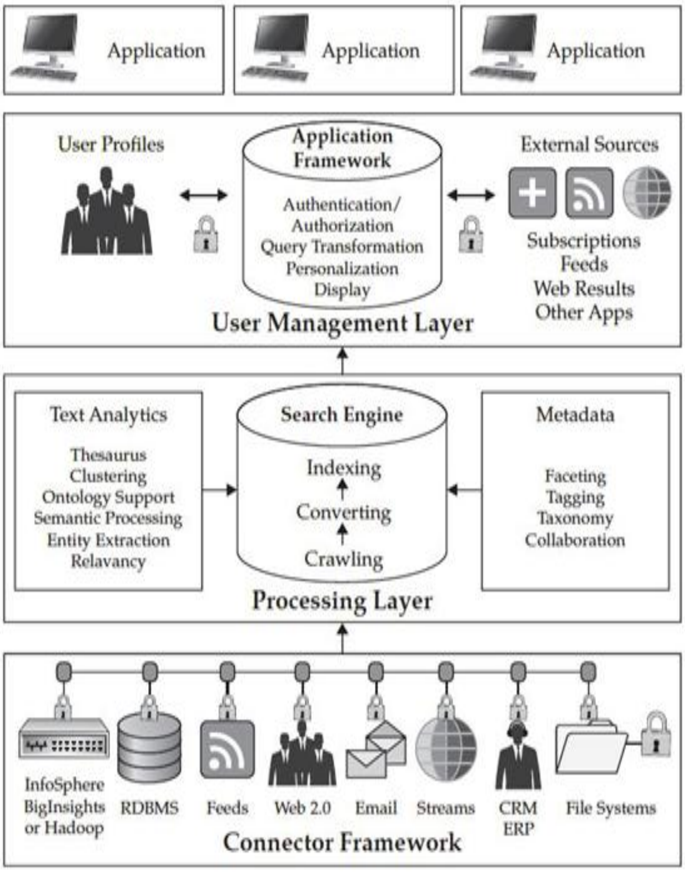
Hadoop

- Hadoop is a distributed file system and data processing engine that is designed to handle extremely high volumes of data in any structure.
- Hadoop has two components:
  - The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between
  - The MapReduce programing paradigm for managing applications on multiple distributed servers
- The focus is on supporting redundancy, distributed architectures, and parallel processing

Some Hadoop Related Names to Know

- Apache Avro: designed for communication between Hadoop nodes through data serialization
- Cassandra and Hbase: a non-relational database designed for use with Hadoop
- Hive: a query language similar to SQL (HiveQL) but compatible with Hadoop
- Mahout: an AI tool designed for machine learning; that is, to assist with filtering data for analysis and exploration
- Pig Latin: A data-flow language and execution framework for parallel computation
- ZooKeeper: Keeps all the parts coordinated and working together

What to do with the data

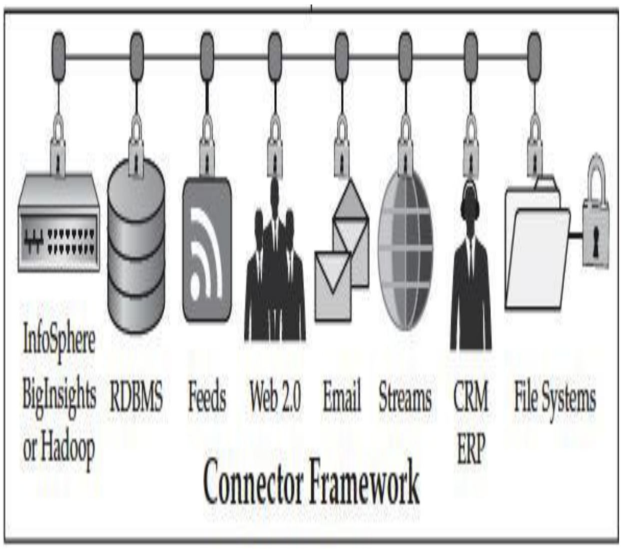


Parallels with Data Warehousing
Data Warehouses

- Extraction
- Transformation
- Load
- Connector
- Processing
- User Management
- Connector
- Processing
- User Management
- Connector
- Processing
- User Management

Connector Framework

- Supports access to data by creating indexes that can be used for access to the data in its native repository (i.e., it does not manage the data, it keeps track of where it is located)



Processing Layer

- Two primary functions:
  - Indexes content: data are crawled, parsed, and analyzed with the result that contents are indexed and located
- Processes queries
  - Manages access to various servers hosting the indexed and searchable content

Challenges and Opportunities with Big Data

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

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

CONCLUSION

Big data transforms the data management land scape by changing fundamental notions of data governance and IT delivery. Though big data is still at its early stage, the advantages of big data will feed the development of new capabilities in sensing, understanding, and playing an active role in the world for the next 20 years and will change all walks of life. However, the underlying analytics and interpretations of results will still require human cognition to connect the dots and see the big picture. An organization needs a strategic plan to adopt the big data technologies. The ability to collect and analyze massive amounts of data will be a key competitive advantage across all industries, including government. Such analytics projects can be complicated, I idiosyncratic, and disruptive — thus they require a strategic plan to be successful. It takes time to change the culture of depending only on traditional data analytics.

There will be occasions of unethical, abuse or misuse of big data applications as big -data analytics and technologies are implemented. Therefore, it is better to be cautious and start small and simple. It takes the whole society to implement big data technologies. Since big data will affect all of us in our life and collaboration and partnership are essential to make big data successful, we are all responsible to work to gather in dealing with the issues raised along with application of the technologies on this journey for the next decade or two.

REFERENCES

1. Big data: The next frontier for innovation, competition, and productivity,
McKinsey Global Institute, May 2011.
2. Big Data' Is Only the Beginning of Extreme Information Management by Gartner
(Mark A. Beyer, Anne Lapkin, Nicholas Gall, Donald Feinberg, Valentin T. Sribar)
3. Privacy-Preserving Data Mining: Models and Algorithms. Edited by Charu C. Agarwal and Phillip S .U, Kluwer Academic Publishers, 2007.
4. Privacy Preserving Data Mining, Rakesh Agarwal and Ramakrishan Srikanth, IBM Almaden Research Center, 2000.
5. SQL Server 2008 R2 Glossary StreamInsight
http://msdn.microsoft.com/en-us/
library/ee378962.aspx
6. Demand Response Measurement & Verification -
http://www.smartgrid.gov/sites/default/files/pdfs/demand_response.pdf
7. T. Hey, S. Tansley, and K. Tolle, Eds., The Fourth Paradigm: Data-Intensive Scientific Discovery.,2010.
8. A. Wagner, S. Speiser, and A. Harth, "Semantic web technologies for a smart energy grid: Requirements and challenges," in ICWS, 2010.
http://iswc2010.semanticweb.org/pdf/506.pdf
9. Hadoop - The definitive guide by Tom White
10. Massive Data Analytics and the Cloud A Revolution in Intelligence Analysis by Booz Allen Hamilton (Michael Farber, Mike Cameron, Christopher Ellis, Josh Sullivan, Ph.D.).