SATHYABAMA UNIVERSITY

COURSE MATERIAL - BIG DATA (SIT1606)

FACULTY OF COMPUTING

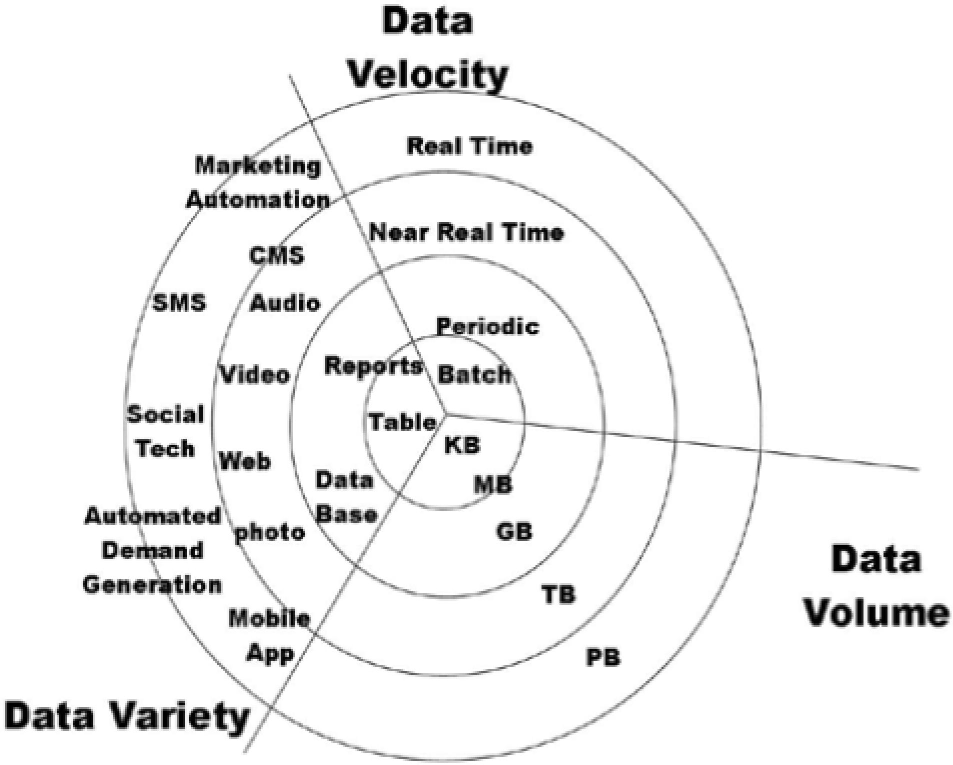
COURSE OBJECTIVES

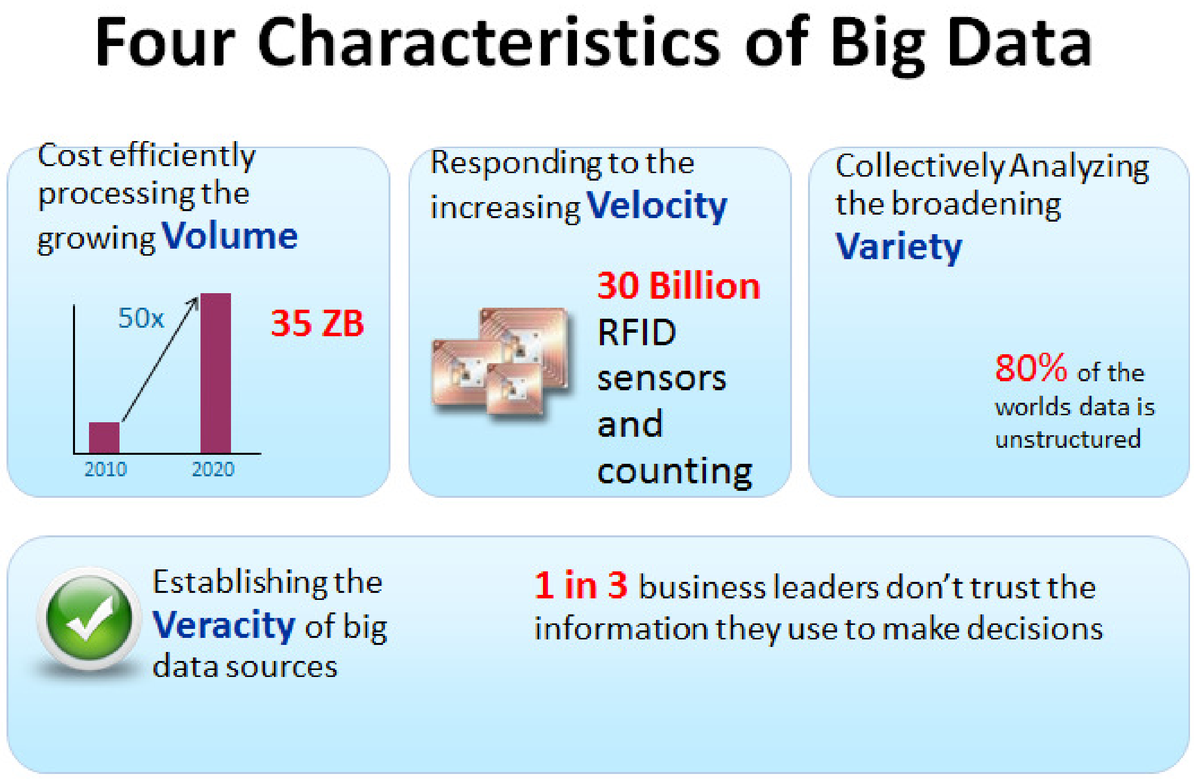
1) To understand the dominant software systems and algorithms for coping with Big Data.
2) Apply appropriate analytics techniques and tools to analyze big data, create statistical models, and identify insights
3) To explore the ethical implications of big data research, and particularly as they related to the web

Unit I – Introduction

Introduction to Big Data:

Big Data has to deal with large and complex datasets that can be structured, Semi-structured, or unstructured and will typically not fit into memory to be Processed.





Four Characteristics of Big Data

Cost efficiently processing the growing **Volume**

35 ZB

Responding to the increasing **Velocity**

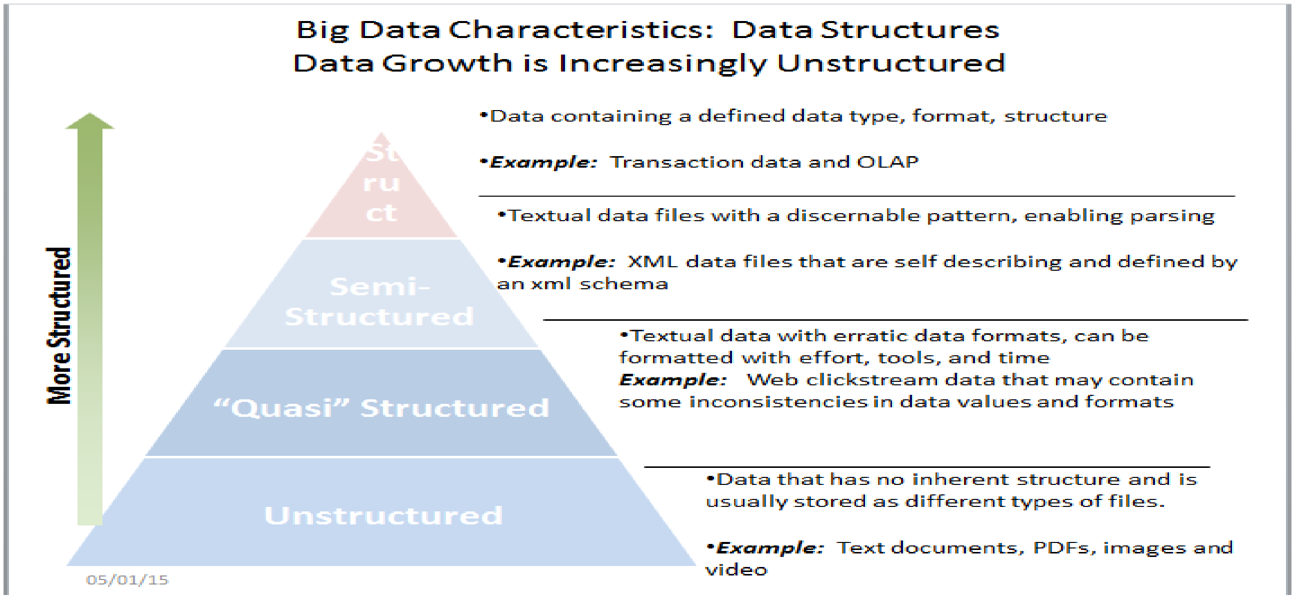
30 Billion RFID sensors and counting

Collectively Analyzing the broadening **Variety**

80% of the worlds data is unstructured

Establishing the **Veracity** of big data sources

1 in 3 business leaders don't trust the information they use to make decisions



Big Data Characteristics: Data Structures
Data Growth is Increasingly Unstructured

- Data containing a defined data type, format, structure
- ***Example:*** Transaction data and OLAP

- Textual data files with a discernable pattern, enabling parsing
- ***Example:*** XML data files that are self describing and defined by an xml schema

- Textual data with erratic data formats, can be formatted with effort, tools, and time
Example: Web clickstream data that may contain some inconsistencies in data values and formats

- Data that has no inherent structure and is usually stored as different types of files.
- ***Example:*** Text documents, PDFs, images and video

05/01/15

Challaenches of Conventional System:

Fundamental challenges

- How to store
- How to work with voluminous data sizes,
    - and more important, how to understand data and turn it into a competitive advantage.

How about Conventional system technology?

• CPU Speeds:

– 1990 - 44 MIPS at 40 MHz

– 2000 - 3,561 MIPS at 1.2 GHz

– 2010 - 147,600 MIPS at 3.3 GHz

• RAM Memory

– 1990 – 640K conventional memory (256K extended memory recommended)

– 2000 – 64MB memory

– 2010 - 8-32GB (and more)

• Disk Capacity

– 1990 – 20MB

– 2000 - 1GB

– 2010 – 1TB

• Disk Latency (speed of reads and writes) – not much improvement in last 7-10 years, currently around 70 – 80MB / sec

How long it will take to read 1TB of data?

- 1TB (at 80Mb / sec):
- – 1 disk - 3.4 hours
- – 10 disks - 20 min
- – 100 disks - 2 min
- – 1000 disks - 12 sec

What do we care about when we process data?

• Handle partial hardware failures without going down:

– If machine fails, we should be switch over to stand by machine

– If disk fails – use RAID or mirror disk

• Able to recover on major failures:

– Regular backups

– Logging

– Mirror database at different site

• Capability:

– Increase capacity without restarting the whole system

– More computing power should equal to faster processing

• Result consistency:

– Answer should be consistent (independent of something failing) and returned in reasonable amount of time

Nature of Data:

Big data is a term thrown around in a lot of articles, and for those who understand what big data means that is fine, but for those struggling to understand exactly what big data is, it can get frustrating. There are several definitions of big data as it is frequently used as an all-encompassing term for everything from actual data sets to big data technology and big data analytics. However, this article will focus on the actual types of data that are contributing to the ever growing collection of data referred to as big data. Specifically we focus on the data created outside of an organization, which can be grouped into two broad categories: structured and unstructured.

Structured Data

1. Created

Created data is just that; data businesses purposely create, generally for market research. This may consist of customer surveys or focus groups. It also includes more modern methods of research, such as creating a loyalty program that collects consumer information or asking users to create an account and login while they are shopping online.

2. Provoked

A Forbes Article defined provoked data as, “Giving people the opportunity to express their views.” Every time a customer rates a restaurant, an employee, a purchasing experience or a product they are creating provoked data. Rating sites, such as Yelp, also generate this type of data.

3. Transacted

Transactional data is also fairly self-explanatory. Businesses collect data on every transaction completed, whether the purchase is completed through an online shopping cart or in-store at the cash register. Businesses also collect data on the steps that lead to a purchase online. For example, a customer may click on a banner ad that leads them to the product pages which then spurs a purchase.

As explained by the Forbes article, “Transacted data is a powerful way to understand exactly what was bought, where it was bought, and when. Matching this type of data with other information, such as weather, can yield even more insights. (We know that people buy more Pop-Tarts at Walmart when a storm is predicted.)"

4. Compiled

Compiled data is giant databases of data collected on every U.S. household. Companies like Acxiom collect information on things like credit scores, location, demographics, purchases and registered cars that marketing companies can then access for supplemental consumer data.

5. Experimental

Experimental data is created when businesses experiment with different marketing pieces and messages to see which are most effective with consumers. You can also look at experimental data as a combination of created and transactional data.

Unstructured Data

People in the business world are generally very familiar with the types of structured data mentioned above. However, unstructured is a little less familiar not because there's less of it, but before technologies like NoSQL and Hadoop came along, harnessing unstructured data wasn't possible. In fact, most data being created today is unstructured. Unstructured data, as the name suggests, lacks structure. It can't be gathered based on clicks, purchases or a barcode, so what is it exactly?

6. Captured

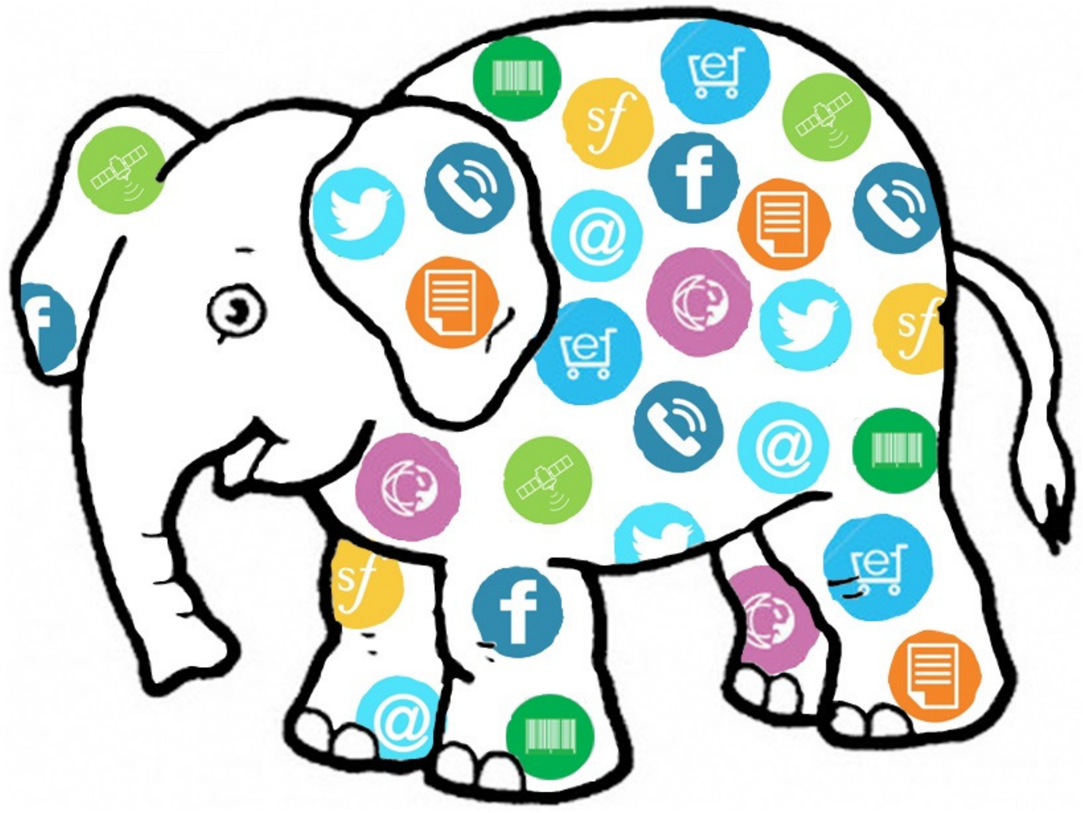
Captured data is created passively due to a person's behavior. Every time someone enters a search term on Google that is data that can be captured for future benefit. The GPS info on our smartphones is another example of passive data that can be captured with big data technologies.

7. User-generated

User-generated data consists of all of the data individuals are putting on the Internet every day. From tweets, to Facebook posts, to comments on news stories, to videos put up on YouTube, individuals are creating a huge amount of data that businesses can use to better target consumers and get feedback on products.

Big data is made up of many different types of data. The seven listed above comprise types of external data included in the big data spectrum. There are, of course, many types of internal data that contribute to big data as well, but hopefully breaking down the types of data helps you to better see why combining all of this data into big data is so powerful for business.

Sources Of Big Data :



Classification of Types of Big Data

The following classification was developed by the Task Team on Big Data, in June 2013. Comments and feedback are welcome.

1. Social Networks (human-sourced information): this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.

1100. Social Networks: Facebook, Twitter, Tumblr etc.

1200. Blogs and comments

1300. Personal documents

1400. Pictures: Instagram, Flickr, Picasa etc.

1500. Videos: Youtube etc.

1600. Internet searches

1700. Mobile data content: text messages

1800. User-generated maps

1900. E-Mail

2. Traditional Business systems (process-mediated data): these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions,reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data").

21. Data produced by Public Agencies

    2110. Medical records

22. Data produced by businesses

    2210. Commercial transactions

    2220. Banking/stock records

    2230. E-commerce

    2240. Credit cards

3. Internet of Things (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

31. Data from sensors

 311. Fixed sensors

  3111. Home automation

  3112. Weather/pollution sensors

  3113. Traffic sensors/webcam

  3114. Scientific sensors

  3115. Security/surveillance videos/images

 312. Mobile sensors (tracking)

  3121. Mobile phone location

  3122. Cars

  3123. Satellite images

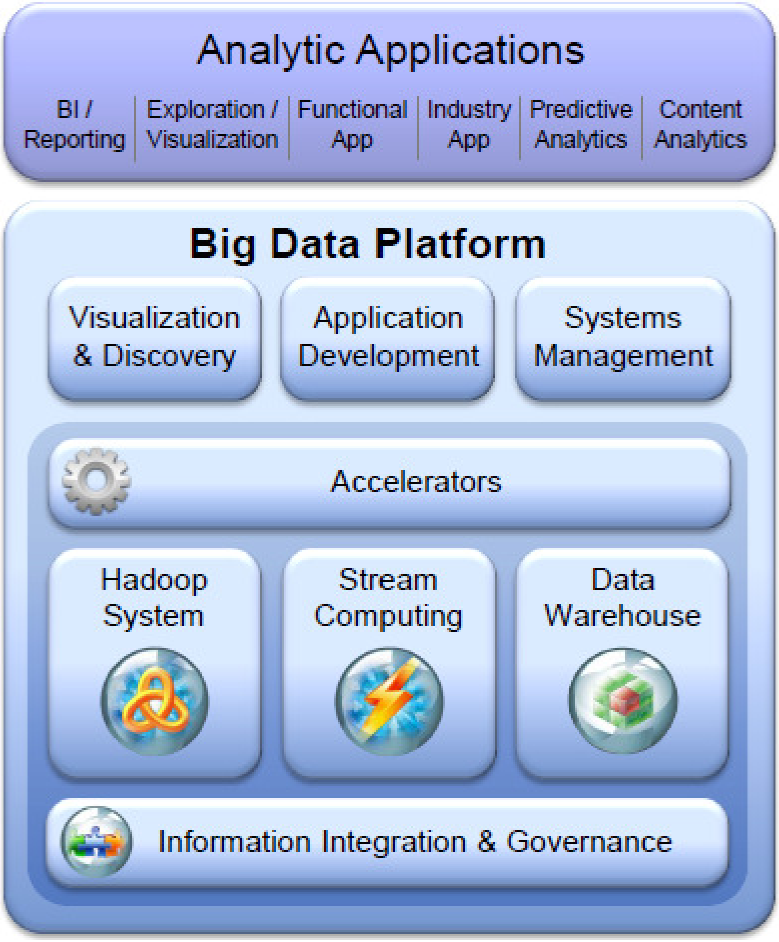
32. Data from computer systems

 3210. Logs

 3220. Web logs

Building a Big Data Platform:

Big Data Platform - Hadoop System:



New analytic applications drive the requirements for a big data platform

•Integrate and manage the full variety, velocity and volume of data

•Apply advanced analytics to information in its native form

•Visualize all available data for ad-hoc analysis

•Development environment for building new analytic applications

•Workload optimization and scheduling

•Security and Governance

Augments open source Hadoop with enterprise capabilities :

–Enterprise-class storage

–Security

–Performance Optimization

–Enterprise integration

–Development tooling

–Analytic Accelerators

–Application and industry accelerators

–Visualization

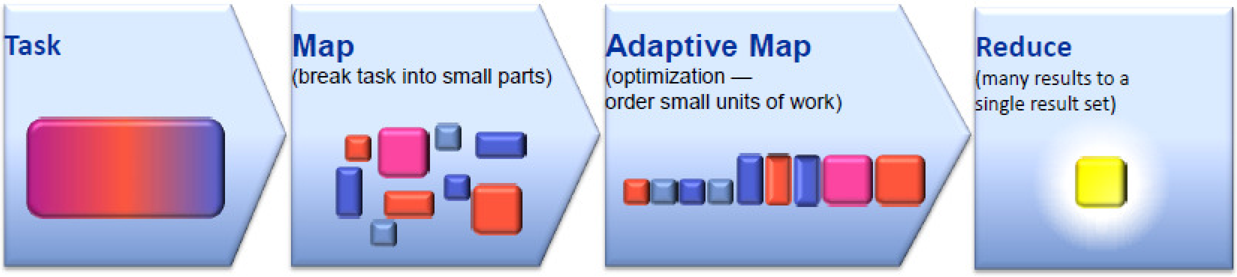
Workload Optimization:

Adaptive MapReduce
•Algorithm to optimize execution time of multiple small and large jobs

•Performance gains of 30% reduce overhead of task startup

Hadoop System Scheduler
•Identifies small and large jobs from prior experience

•Sequences work to reduce overhead



Big Data Platform - Stream Computing :

□**Built to analyze data in motion**

•Multiple concurrent input streams

•Massive scalability

□**Process and analyze a variety of data**

•Structured, unstructured content, video, audio

•Advanced analytic operators

Big Data Platform - Data Warehousing :

□Workload optimized systems

–Deep analytics appliance

–Configurable operational analytics appliance

–Data warehousing software

□Capabilities

•Massive parallel processing engine

•High performance OLAP

•Mixed operational and analytic workloads

Big Data Platform - Information Integration and Governance

□Integrate any type of data to the big data platform
–Structured

–Unstructured

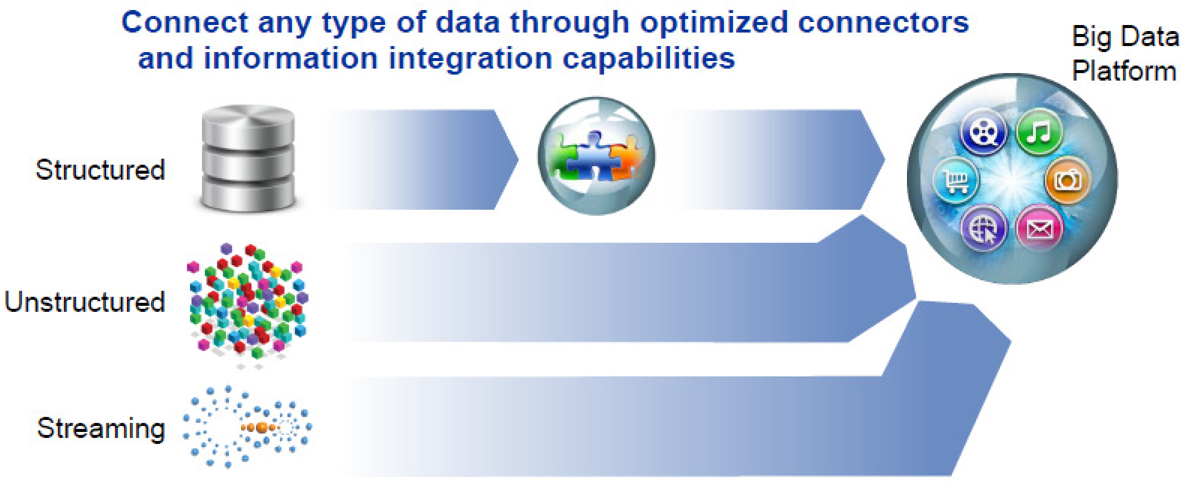
–Streaming
□Governance and trust for big data
–Secure sensitive data

–Lineage and metadata of new big data sources

–Lifecycle management to control data growth

–Master data to establish single version of the truth

Leverage purpose-built connectors for multiple data sources :


Connect any type of data through optimized connectors
and information integration capabilities
Big Data
Platform
Structured
Unstructured
Streaming

□Massive volume of structured data movement

•2.38 TB / Hour load to data warehouse
•High-volume load to Hadoop file system

□Ingest unstructured data into Hadoop file system

□Integrate streaming data sources

Big Data Platform - User Interfaces

•Business Users

•Visualization of a large volume and wide variety of data

•**Developers**

•Similarity in tooling and languages

•Mature open source tools with enterprise capabilities

•Integration among environments

•**Administrators**

•Consoles to aid in systems management

Big Data Platform –Accelerators :

Analytic accelerators

–Analytics, operators, rule sets

Industry and Horizontal Application Accelerators

–Analytics

–Models

–Visualization / user interfaces

–Adapters

Big Data Platform - Analytic Applications :

Big Data Platform is designed for analytic application development and integration

BI/Reporting – Cognos BI, Attivio

Predictive Analytics – SPSS, G2, SAS

Exploration/Visualization – BigSheets, Datameer

Instrumentation Analytics – Brocade, IBM GBS

Content Analytics – IBM Content Analytics

Functional Applications – Algorithmics, Cognos Consumer Insights, Clickfox, i2, IBM GBS

Industry Applications – TerraEchos, Cisco, IBM GBS

Big Data Enterprise Architecture:

The 5 V's of Big Data:

Too often in the hype and excitement around Big Data, the conversation gets complicated very quickly. Data scientists and technical experts bandy around terms like Hadoop, Pig, Mahout, and Sqoop, making us wonder if we're talking about information architecture or a Dr. Seuss book. Business executives who want to leverage the value of Big Data analytics in their organisation can get lost amidst this highly-technical and rapidly-emerging ecosystem. In an effort to simplify Big Data, many experts have referenced the "3 V's": Volume, Velocity, and Variety. In other words, is information being generated at a high volume (e.g. terabytes per day), with a rapid rate of change, encompassing a broad range of sources including both structured and unstructured data? If the answer is yes then it falls into the Big Data category along with sensor data from the "internet of things", log files, and social media streams. The ability to understand and manage these sources, and then integrate them into the larger Business Intelligence ecosystem can provide previously unknown insights from data and this understanding leads to the "4th V" of Big Data – *Value*.

There is a vast opportunity offered by Big Data technologies to discover new insights that drive significant business value. Industries are seeing data as a market differentiator and have started reinventing themselves as "data companies", as they realise that information has become their biggest asset. This trend is prevalent in industries such as telecommunications, internet search firms, marketing firms, etc. who see their data as a key driver for monetisation and growth. Insights such as footfall traffic patterns from mobile devices have been used to assist city planners in designing more efficient traffic flows. Customer sentiment analysis through social media and call logs have given new insights into customer satisfaction. Network performance patterns have been analysed to discover new ways to drive efficiencies. Customer usage patterns based on web click-stream data have driven innovation for new products and services to increase revenue. The list goes on.

Key to success in any Big Data analytics initiative is to first identify the business needs and opportunities, and then select the proper fit-for-purpose platform. With the array of new Big Data technologies emerging at a rapid pace, many technologists are eager to be the first to test the latest Dr. Seuss-termed platform. But each technology has a unique specialisation, and might not be aligned to the business priorities. In fact, some identified use cases from the business might be best suited by existing technologies such as a data warehouse while others require a combination of existing technologies and new Big Data systems.

With this integration of disparate data systems comes the 5th V – *Veracity*, i.e. the correctness and accuracy of information. Behind any information management practice lies the core doctrines of Data Quality, Data Governance, and Metadata Management, along with considerations for Privacy and Legal concerns. Big Data needs to be integrated into the entire information landscape, not seen as a stand-alone effort or a stealth project done by a handful of Big Data experts.

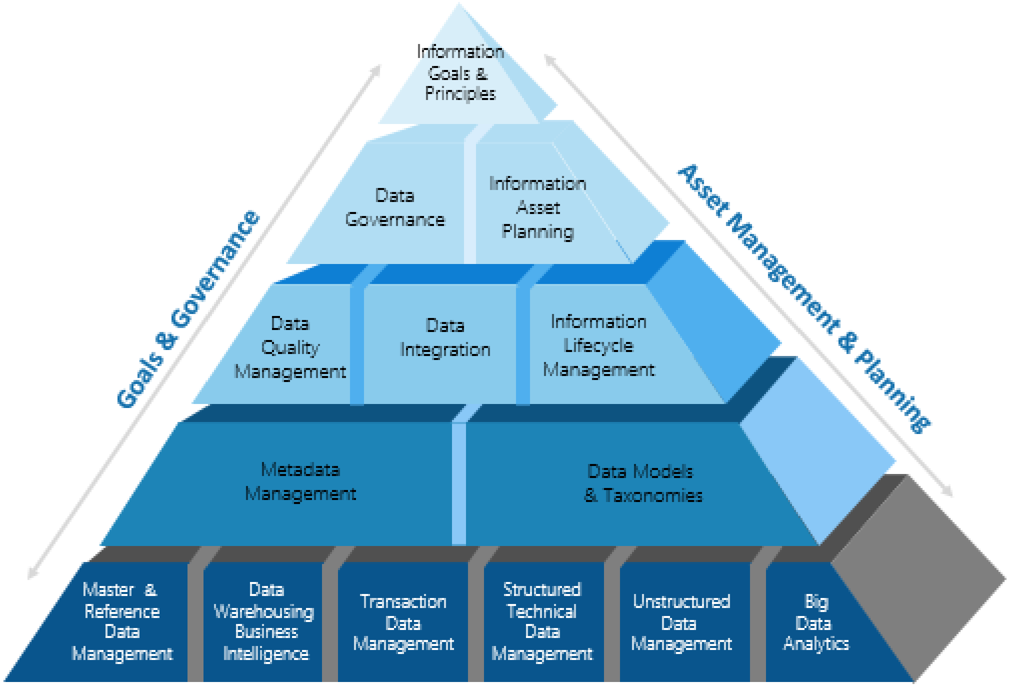

Information Goals & Principles
Data Governance
Information Asset Planning
Data Quality Management
Data Integration
Information Lifecycle Management
Metadata Management
Data Models & Taxonomies
Master & Reference Data Management
Data Warehousing Business Intelligence
Transaction Data Management
Structured Technical Data Management
Unstructured Data Management
Big Data Analytics
Goals & Governance
Asset Management & Planning

Figure 1. Enterprise Architects Information Management Framework

In the excitement and hype around Big Data analytics, it's easy to see this emerging technology as a "silver bullet" that can magically generate new insights solely through powerful technology and smart data scientists. As in any age of change, however, core principles still apply, and in order to gain insights from Big Data, you need to make sure your "little data" is correct. Many of the "golden nuggets" of discovery are obtained through an intersection of Big Data analytics with traditional sources such as a data warehouses or master data management hubs.

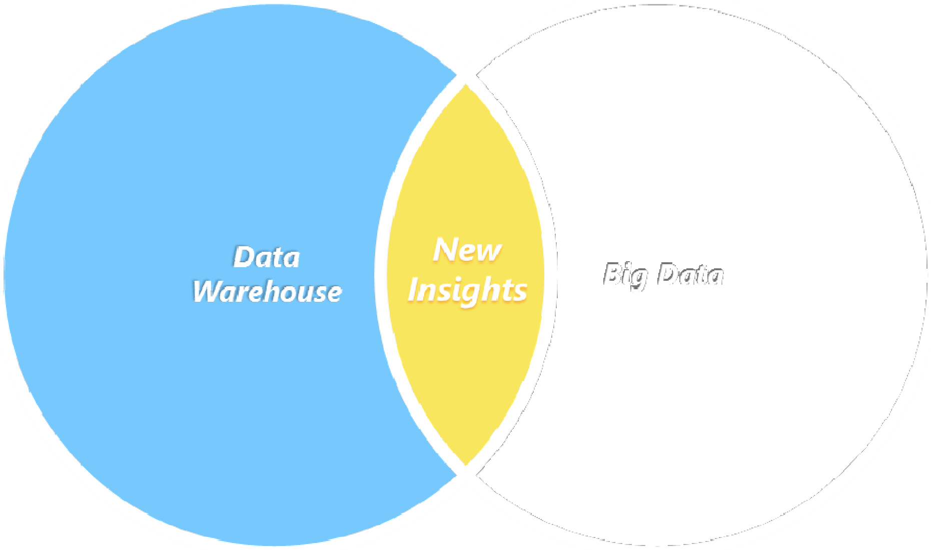


Figure 2. The intersection of Big Data analytics with traditional sources such as a data warehouse

Customer sentiment analysis is a common use-case for Big Data analytics—i.e. what are our customers saying about our products in social media and/or call log records? And how can we leverage this information to improve our business? Unless you have a robust 'single source of record' for customer information, new discoveries from Big Data analytics will be of little use. Was it Jane R. Doe or Jane P. Doe complaining about the new luxury sedan model? With data properly managed within an information management framework, the full value of Big Data becomes apparent and "golden nuggets" of information can appear. For example, not only did Jane R. Doe complain about the new luxury sedan, but she had five service calls about her transmission. She has purchased five high-priced sedans from us in the past ten years and has an income of over $750,000. Jane R. Doe recently followed our competitor on Twitter and has asked several questions about new features. It might be worth having a representative call her personally.

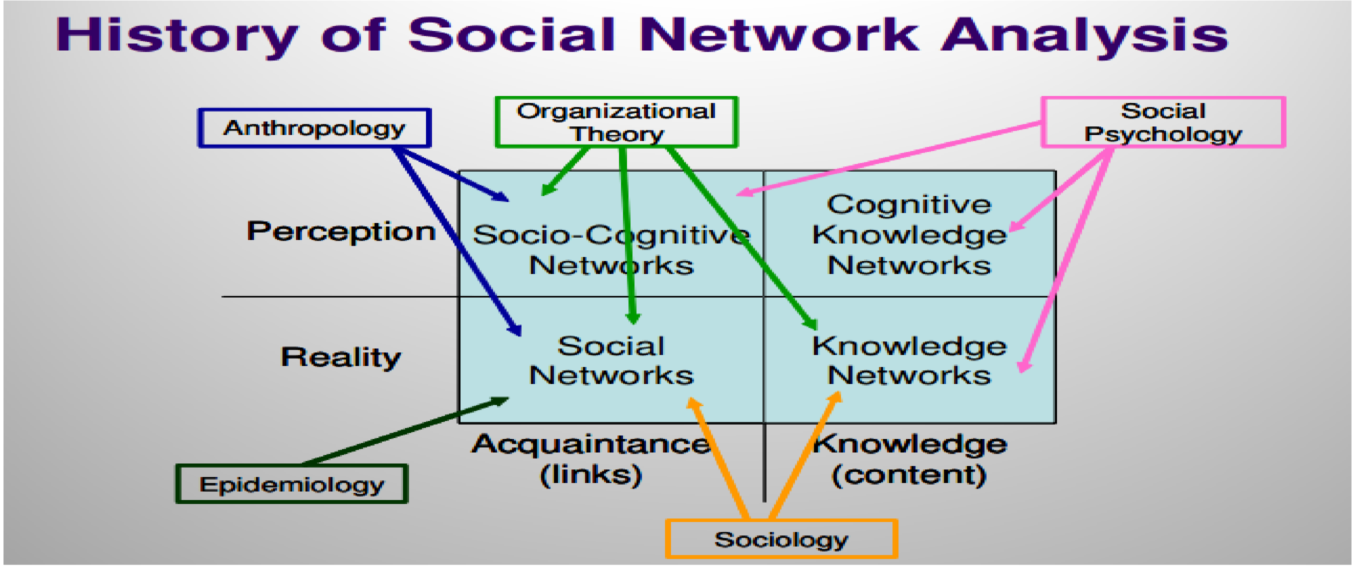
Big Data analytics is an exciting development in the field of information management and, if used properly, can generate a wealth of opportunity. In order to discover the "golden nuggets" in your organisation, remember these guiding principles:

- Start with your business goals and drivers and align them to fit-for-purpose technologies (not the other way around)
- Integrate your Big Data initiatives with core information management practices
- Build your information management practice on a core framework that includes data governance, data quality management, data quality, and the other principles that create a trusted source of information

Lastly, have fun—this is an exciting time to be in information management. New technologies are emerging almost daily that can add significant value to your organisation, particularly in the Big Data space.

Big Data Analytics for Social and Behavioral Sciences

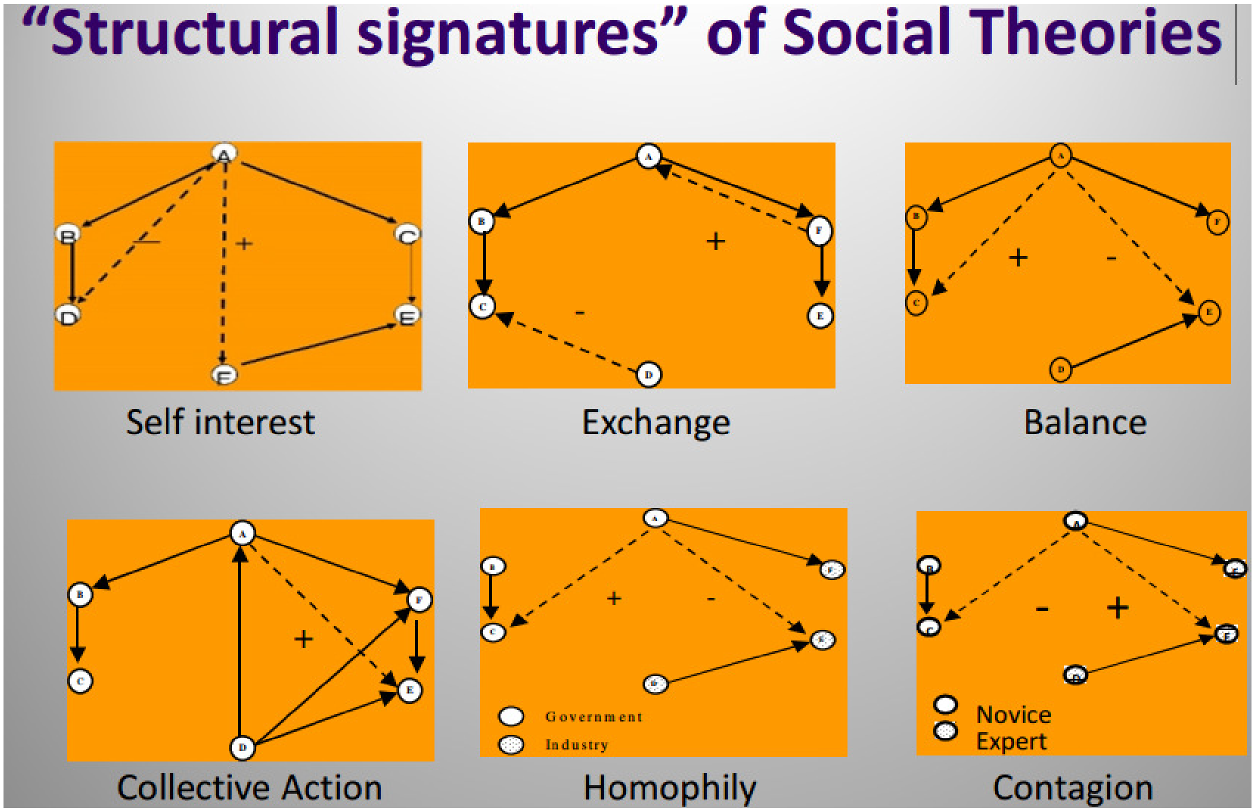
What Social & Behavioral Sciences Tell Us?



- Social science networks have widespread application in various fields
- Most of the analyses techniques have come from Sociology, Statistics and Mathematics
- For a comprehensive introduction to social network analysis

Why do we create and sustain networks?

- Theories of self-interest
- Theories of social and resource exchange
- Theories of mutual interest and collective action
- Theories of contagion
- Theories of balance
- Theories of homophily
- Theories of proximity
- Theories of co-evolution So



"Structural signatures" of Social Theories

Self interest

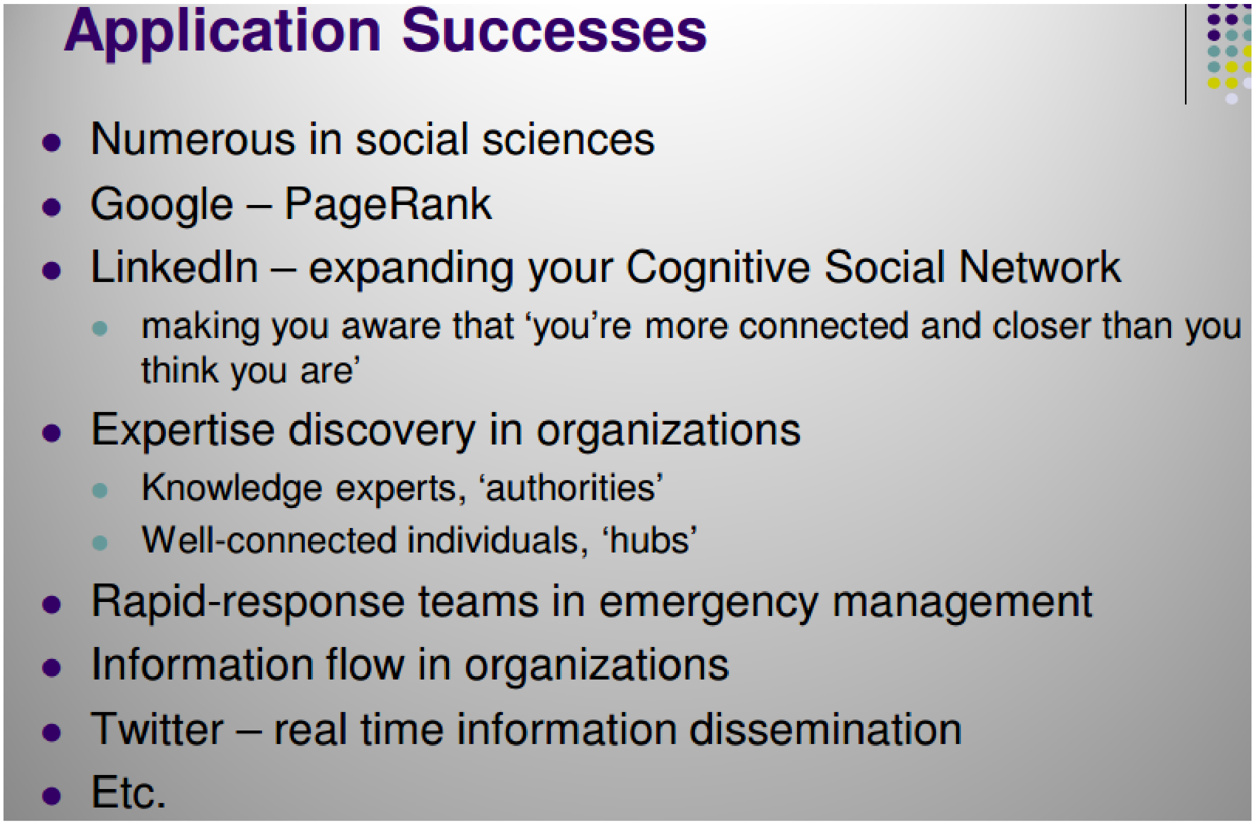
Exchange

Balance

Collective Action

Homophily

Contagion



Application Successes

- Numerous in social sciences
- Google – PageRank
- LinkedIn – expanding your Cognitive Social Network
  - making you aware that 'you're more connected and closer than you think you are'
- Expertise discovery in organizations
  - Knowledge experts, 'authorities'
  - Well-connected individuals, 'hubs'
- Rapid-response teams in emergency management
- Information flow in organizations
- Twitter – real time information dissemination
- Etc.