Produced by:
Dr. Brenda Mullally bmullally@wit.ie
Ruth Barry rbarry@wit.ie

Department Computing Maths and Physics
Waterford Institute of Technology
www.wit.ie
moodle.wit.ie

MSc Enterprise Software Systems Business Intelligence

Big Data and Integration technologies

Big Data - Definition and Concepts

- Big Data means different things to people with different backgrounds and interests
- Traditionally, "Big Data" = massive volumes of data
  - E.g., volume of data at NASA, Google, …
- Where does the Big Data come from?
  - Everywhere! Web logs, RFID, GPS systems, sensor networks, social networks, Internet-based text documents, Internet search indexes, detail call records, astronomy, atmospheric science, biology, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, multimedia archives, …

The Data Size Is Getting Big, Bigger, …

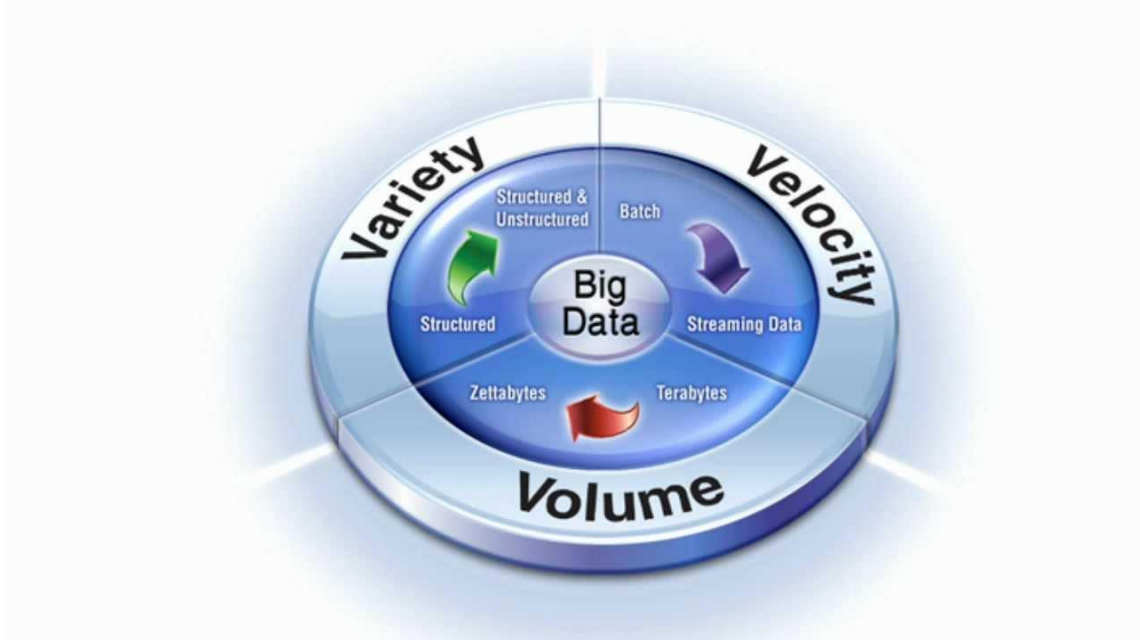
- Hadron Collider - 1 PB/sec
- Boeing jet - 20 TB/hr
- Facebook - 500 TB/day
- YouTube – 1 TB/4 min
- The proposed Square Kilometer Array telescope (the world's proposed biggest telescope) – 1 EB/day

Name	Symbol	Value
Kilobyte	kB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zettabyte	ZB	10^{21}
Yottabyte	YB	10^{24}
Brontobyte*	BB	10^{27}
Gegobyte*	GeB	10^{30}

*Not an official SI (International System of Units) name/symbol, yet.

Big Data - Definition and Concepts

- Big Data is a misnomer!
- Big Data is more than just "big"
- The Vs that define Big Data
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Variability
  - Value
  - *…It's About Variety, not Volume: companies are focused on the variety of data, not its volume. The most important goal and potential reward of Big Data initiatives is the ability to analyze diverse data sources and new data types, not managing very large data sets.*



Big Data definition

- McKinsey study defines Big Data:

- "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze"

  - The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.
  - Once an organisation is using the technology of Big Data , this can prove to be the easy part—the hard part is figuring out what you are going to do with the output generated by your Big Data analytics. As the ancient Greek philosophers said, "Action is character." It's what you do that counts.

  - www.mckinseyquarterly.com/home.aspx

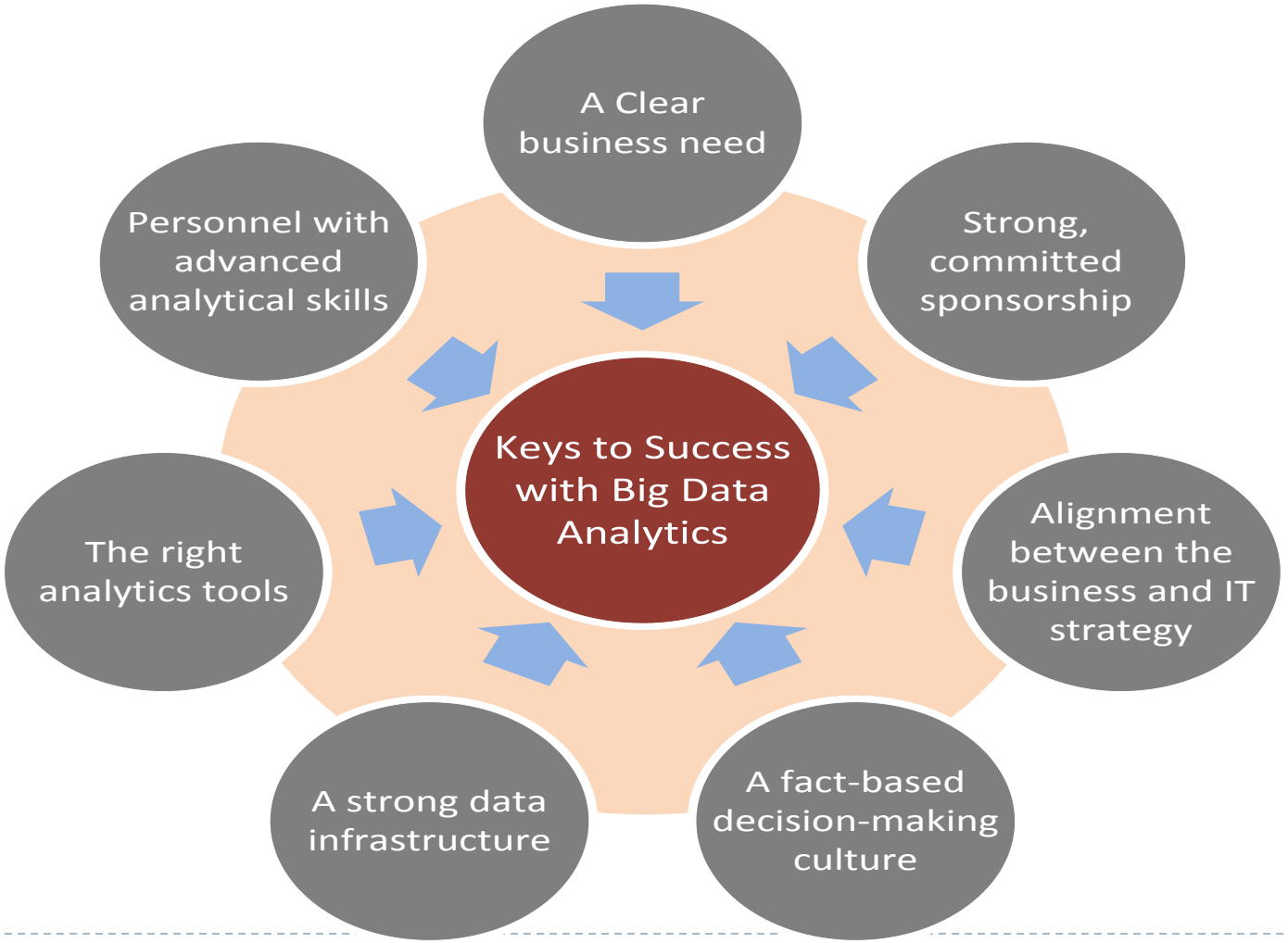
Big Data Considerations

- You can't process the amount of data that you want to because of the limitations of your current platform.
- You can't include new/contemporary data sources (e.g., social media, RFID, Sensory, Web, GPS, textual data) because it does not comply with the data storage schema
- You need to (or want to) integrate data as quickly as possible to be current on your analysis.
- You want to work with a schema-on-demand data storage paradigm because of the variety of data types involved.
- The data is arriving so fast at your organization's doorstep that your traditional analytics platform cannot handle it.
- …

Critical Success Factors for Big Data Analytics

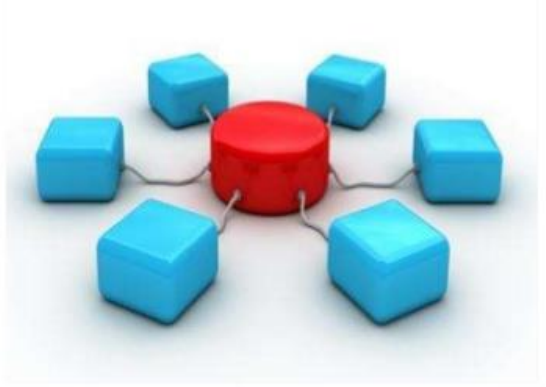
- A clear business need (alignment with the vision and the strategy)
- Strong, committed sponsorship (executive champion)
- Alignment between the business and IT strategy
- A fact-based decision-making culture
- A strong data infrastructure
- The right analytics tools
- Right people with right skills

Critical Success Factors for Big Data Analytics



Data Integration Techniques and technologies

- Challenges – integrating disparate data (internal) + big data
  - Data quality and security
  - Lack of a business case and funding
  - Poor data integration infrastructur



White, Colin, "Data Integration: Using ETL, EAI, and EII Tools TDWI"

Data Integration Techniques

- Data consolidation
  - ETL
  - ECM (enterprise content management) - integration of unstructured data
- Data Federation
  - Enterprise information integration (EII)
- Data propagation
  - Enterprise application integration (EAI)
  - Enterprise data replication (EDR)

White, Colin, "Data Integration: Using ETL, EAI, and EII Tools TDWI"

Data Integration Application Variables

- Source data type
  - Structured
  - Semi-structured (e.g., XML)
  - Unstructured
  - Packaged application
  - EAI
  - Web service
  - Metadata
- Source data organization
  - Homogeneous or heterogeneous
  - Centralized or distributed (integrated data and metadata)
  - Federated (integrated metadata) or dispersed (no integrated metadata)
- Source data transformation requirements
  - Data restructuring
  - Data cleansing
  - Data reconciliation
  - Data aggregation
- Target data currency (latency) and access
  - Real time
  - Near real time
  - Point in time
  - Read-only or read-write
- Data integration technique and mode
  - consolidation, federation, propagation, changed data capture
  - event push or on-demand pull
  - synchronous or asynchronous
- Data integration technology
  - ETL, EII, EAI, EDR, ECM
- Data scale
  - Number of data sources
  - Data store size
  - Data store volatility

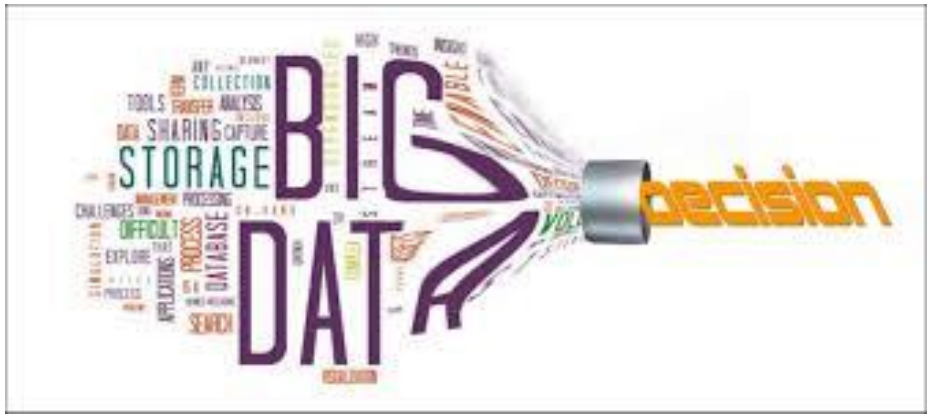
White, Colin, "Data Integration: Using ETL, EAI, and EII Tools TDWI"

Challenges of Big Data Analytics

- Data volume
  - The ability to capture, store, and process the huge volume of data in a timely manner
- Data integration
  - The ability to combine data quickly and at reasonable cost
- Processing capabilities
  - The ability to process the data quickly, as it is captured (i.e., stream analytics)
- Data governance (… security, privacy, access)
- Skill availability (… data scientist)
- Solution cost (ROI)

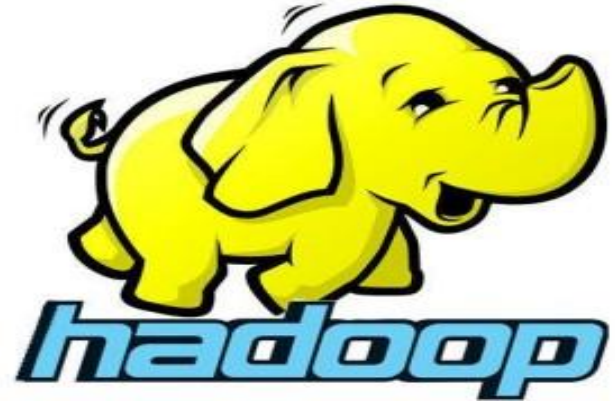
Business Problems Addressed by Big Data Analytics

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling/up-selling
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities
- …



Big Data Technologies

- MapReduce …
- Hadoop …
- Hive
- Pig
- Hbase
- Flume
- Oozie
- Ambari
- Avro
- Mahout, Sqoop, Hcatalog, ….



http://www.cignex.com/platform/big-data-integration-analytics

Big Data Technologies
Hadoop

- Hadoop is an open source framework for storing and analyzing massive amounts of distributed, unstructured data
- Originally created by Doug Cutting at Yahoo!
- Hadoop clusters run on inexpensive commodity hardware so projects can scale-out inexpensively
- Hadoop is now part of Apache Software Foundation
- Open source - hundreds of contributors continuously improve the core technology
- MapReduce + Hadoop = Big Data core technology

Big Data And Data Warehousing

- What is the impact of Big Data on DW?
  - Big Data and RDBMS do not go nicely together
  - Will Hadoop replace data warehousing/RDBMS?
- Use Cases for Hadoop
  - Hadoop as the repository and refinery
  - Hadoop as the active archive
- Use Cases for Data Warehousing
  - Data warehouse performance
  - Integrating data that provides business value
  - Interactive BI tools

Data warehouse is the architecture, Big Data solution is a technology *(Inmon, 2013)*

- *http://www.b-eye-network.com/print/17017*

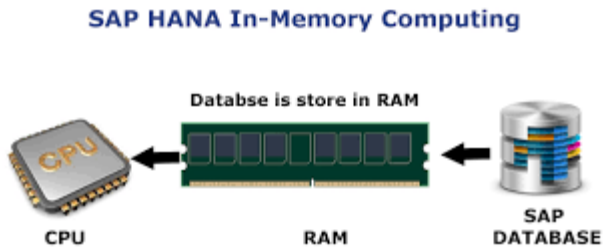
Big Data Vendors

- Big Data vendor landscape is developing very rapidly
- A representative list would include
  - Claudera - claudera.com
  - MapR – mapr.com
  - Hortonworks - hortonworks.com
  - Also, IBM (Netezza, InfoSphere), Oracle (Exadata, Exalogic), Microsoft, Amazon, Google, …
  - http://www.ibm.com/analytics/us/en/technology/data-integration/
  - http://www.datameer.com/product/data-integration.html

Software, Hardware, Service, …

New and innovative techniques to enable Big Data Analytics

- In-memory analytics
  - Storing and processing the complete data set in RAM
- In-database analytics
  - Placing analytic procedures close to where data is stored
- Grid computing & MPP
  - Use of many machines and processors in parallel (MPP - massively parallel processing)
- Appliances
  - Combining hardware, software, and storage in a single unit for performance and scalability



NoSQL Databases

- A 55-minute introduction to NoSQL by Martin Fowler is available to view here:
  - http://www.youtube.com/watch?v=qI_g07C_Q5I
- Try out a demo version of a document database (MongoDB) for yourself here:
  - http://try.mongodb.org/
- Suitable for BigData, scalability, flexible data models, less costly