Produced by:
Dr. Brenda Mullally bmullally@wit.ie

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

HDip Busines Systems Analaysis
Data Analytics

Big Data and Emerging Trends:

- Learn what Big Data is and how it is changing the world of analytics
- Understand the motivation for and business drivers of Big Data analytics
- Become familiar with the wide range of enabling technologies for Big Data analytics

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, …

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

Examples

- Boeing jet - 20 TB/hr
- Facebook - 500 TB/day.
- YouTube – 1 TB/4 min.

Fundamentals of Big Data Analytics

- Big Data by itself, regardless of the size, type, or speed, is worthless
- Big Data + “big” analytics = value
- With the value proposition, Big Data also brought about big challenges
  - Effectively and efficiently capturing, storing, and analyzing Big Data
  - New breed of technologies needed (developed (or purchased or hired or outsourced …)

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 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

Enablers of 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

Challenges of Big Data Analytics

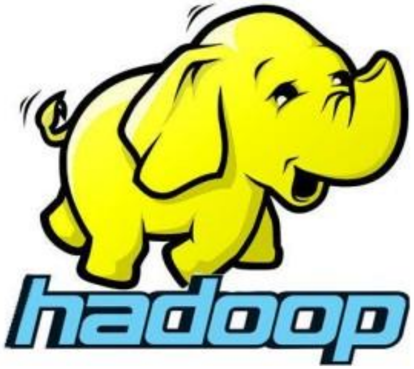
- Data volume
  - The ability to capture, store, and process the huge volume of data in a timely manner
- Data integration
  - The ability to combine data quickly and at reasonable cost
- Processing capabilities
  - The ability to process the data quickly, as it is captured (i.e., stream analytics)
- 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,



Big Data Technologies
MapReduce

- MapReduce distributes the processing of very large multi-structured data files across a large cluster of ordinary machines/processors
- Goal - achieving high performance with “simple” computers
- Developed and popularized by Google
- Good at processing and analyzing large volumes of multi-structured data in a timely manner
- Example tasks: indexing the Web for seearch, graph analysis, text analysis, machine learning, …

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 Technologies
Hadoop

- How Does Hadoop Work?
  - Access unstructured and semi-structured data (e.g., log files, social media feeds, other data sources)
  - Break the data up into "parts," which are then loaded into a file system made up of multiple nodes running on commodity hardware using HDFS
  - Each "part" is replicated multiple times and loaded into the file system for replication and failsafe processing
  - A node acts as the Facilitator and another as Job Tracker
  - Jobs are distributed to the clients, and once completed the results are collected and aggregated using MapReduce

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

Coexistence of Hadoop and DW

1. Use Hadoop for storing and archiving multi-structured data
2. Use Hadoop for filtering, transforming, and/or consolidating multi-structured data
3. Use Hadoop to analyze large volumes of multi-structured data and publish the analytical results
4. Use a relational DBMS that provides MapReduce capabilities as an investigative computing platform
5. Use a front-end query tool to access and analyze data

Big Data Vendors

- Big Data vendor landscape is developing very rapidly
- A representative list would include
  - Cloudera - cloudera.com
  - MapR – mapr.com
  - Hortonworks - hortonworks.com
  - Also, IBM (Netezza, InfoSphere), Oracle (Exadata, Exalogic), Microsoft, Amazon, Google, …

Software, Hardware, Service, …