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

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, …

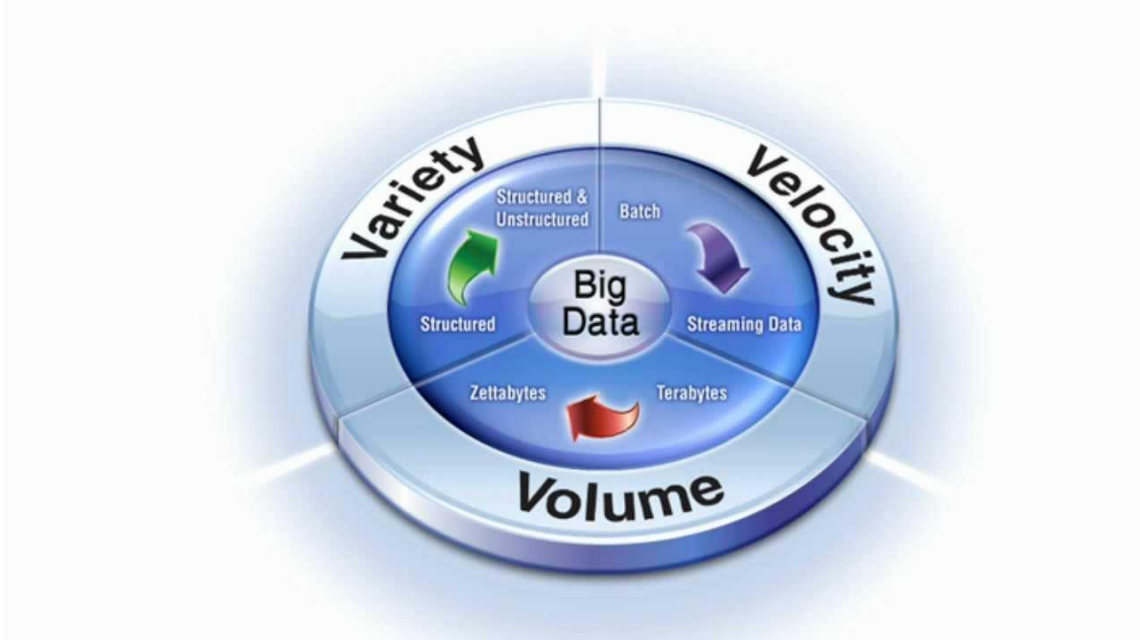
- Hadron Collider - 1 PB/sec
- Boeing jet - 20 TB/hr
- Facebook - 500 TB/day
- YouTube – 1 TB/4 min
- The proposed Square Kilometer Array telescope (the world's proposed biggest telescope) – 1 EB/day

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

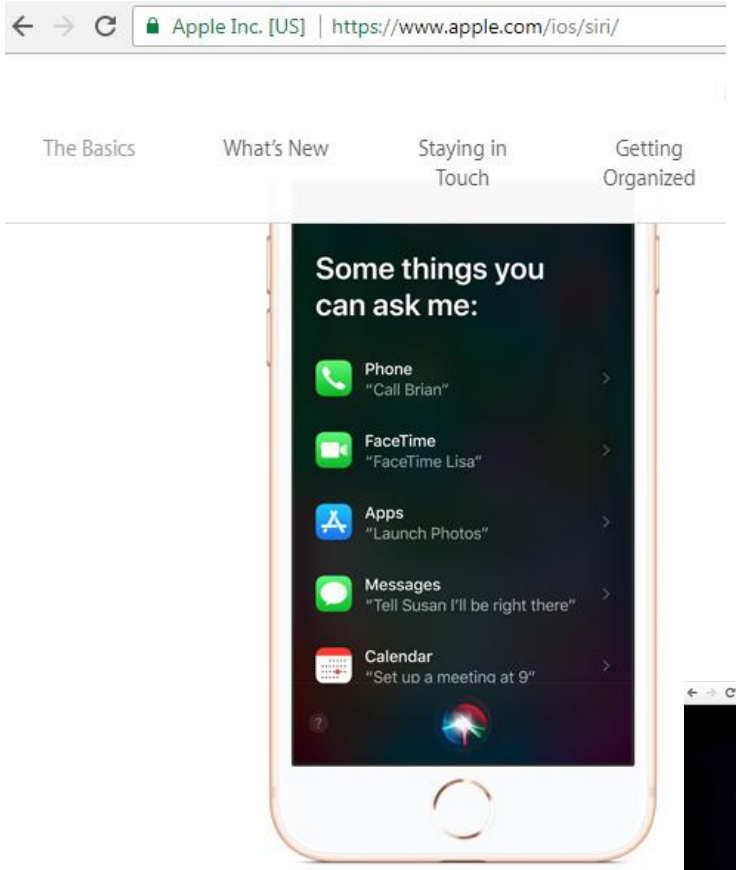
- McKinsey study defines Big Data:

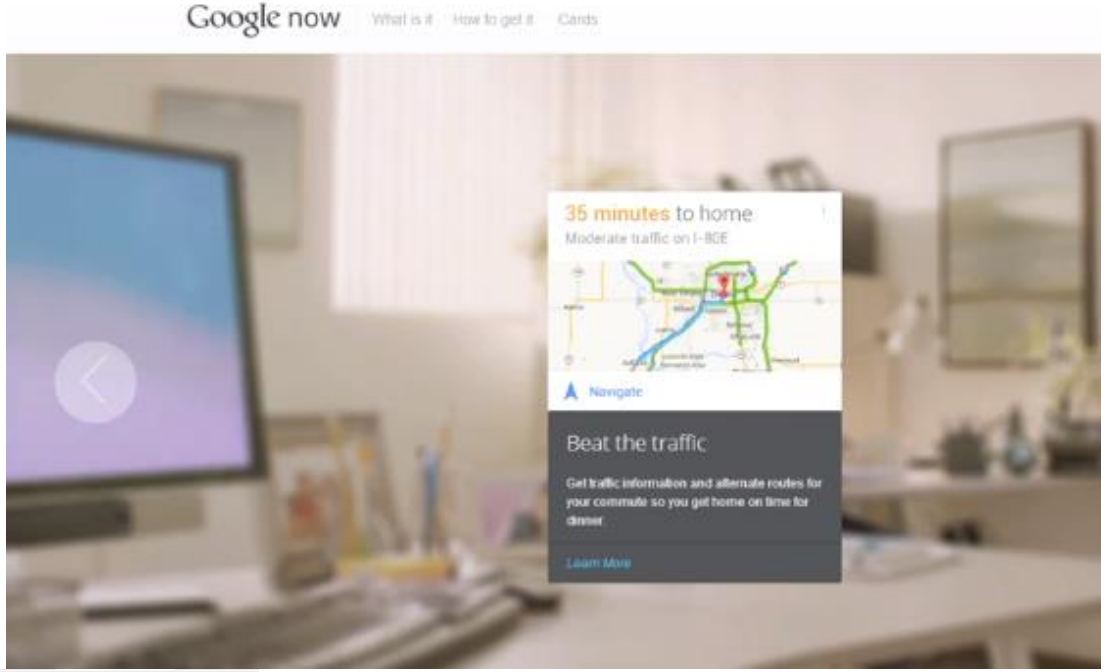
- "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze"

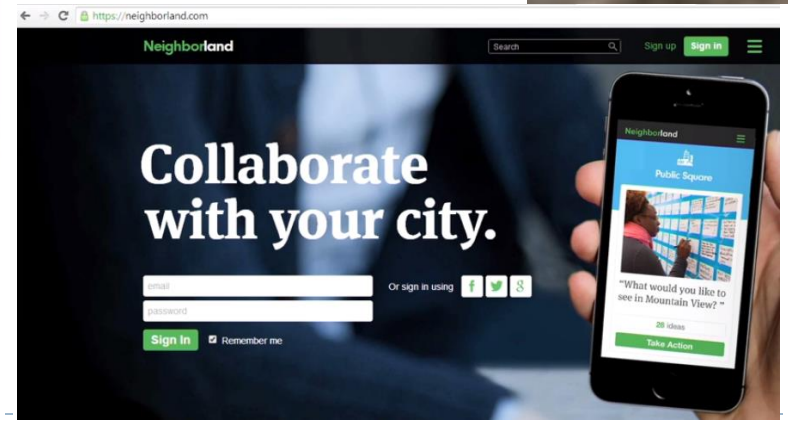
  - The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and petabytes of data now available. That's where Big Data analytics become necessary.
  - Once an organisation is using the technology of Big Data , this can prove to 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

Common applications of Big Data - Consumers



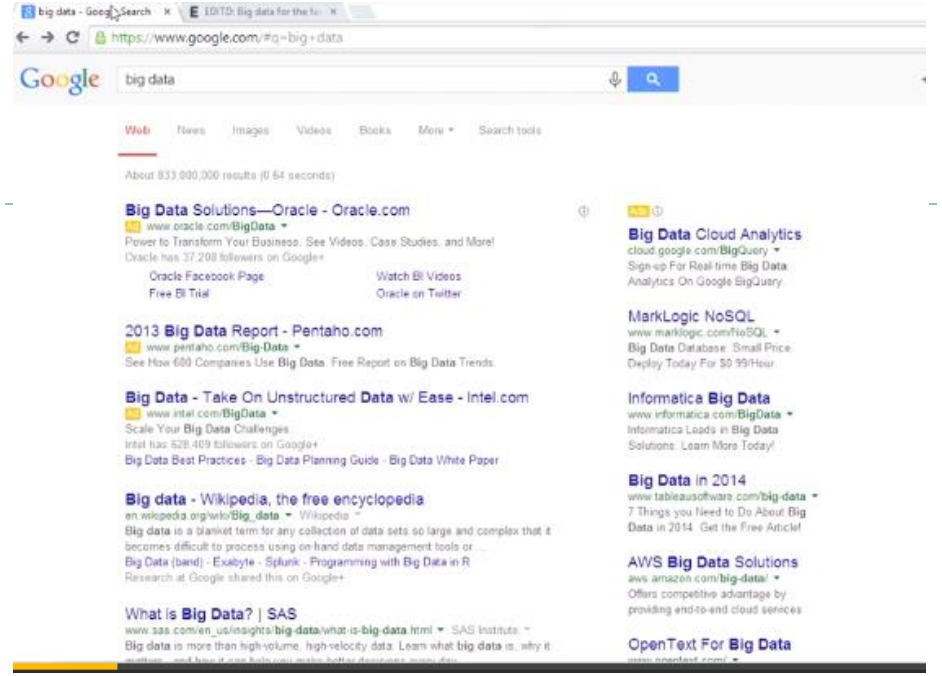




https://www.lynda.com/Data-Science-tutorials

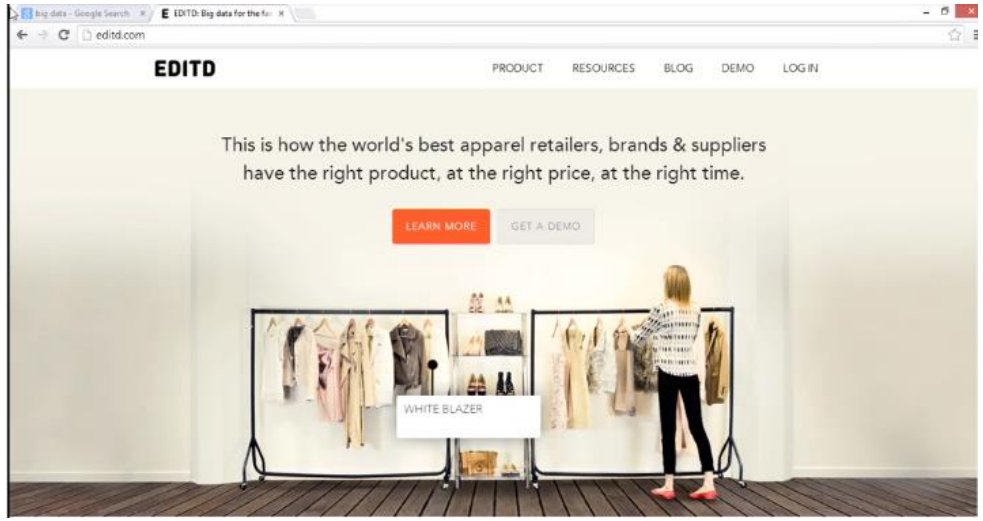
Applications for Business

- Big data in commerce – Google Ad Searches



- Predictive marketing - this is when big data is used to help decide who the audience would be for something before they actually get there.

  - Predict major life events
  - Looks at consumer behaviour
  - Use demographic info
  - Can purchase more data



UPS - success

- 16 million shipments per day
- 40 million tracking requests
- UPS estimated to have 16PBs of data in operations

- Can you guess how much money UPS can save by reducing each driver's route by just 1 mile?

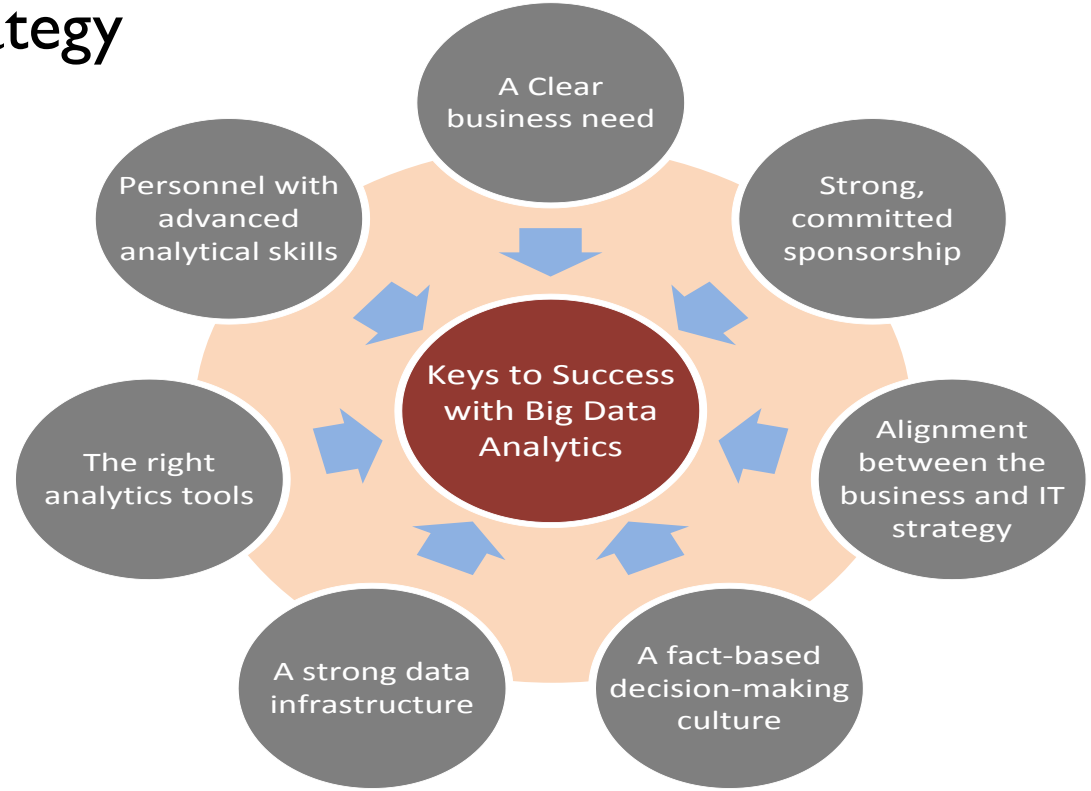
https://www.coursera.org/learn/big-data-introduction/supplement/AmZ98/slides-organizaton-generated-big-data-benefits

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

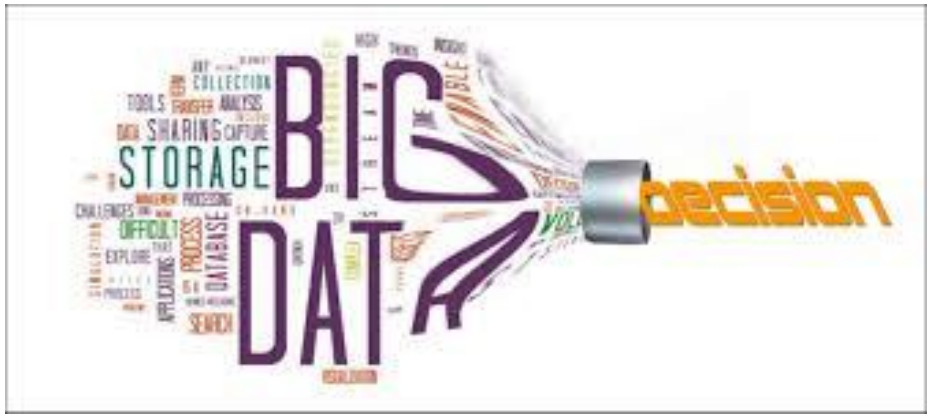


Challenges of Big Data Analytics

- 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)

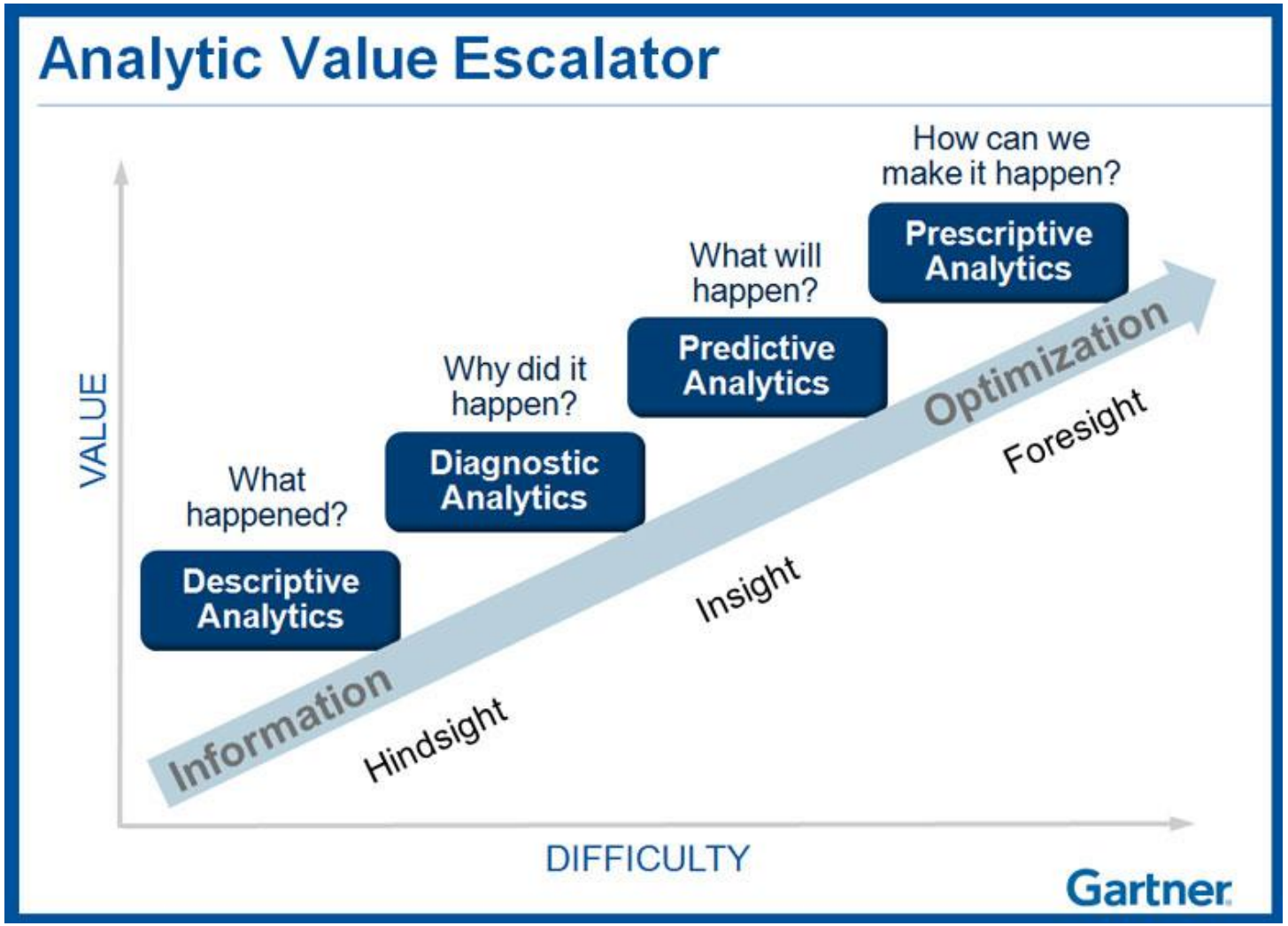
Business Problems Addressed by Big Data Analytics

- 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
- …

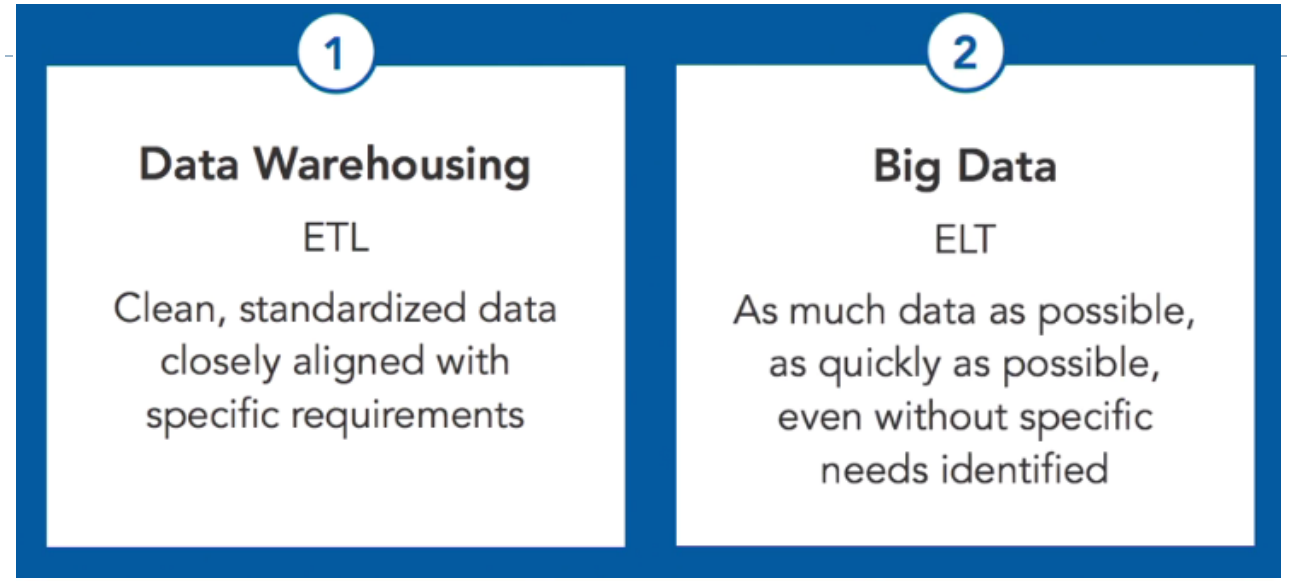


Evolving Analytics

- Descriptive
- Predictive
- Discovery
- Prescriptive



ETL vs ELT

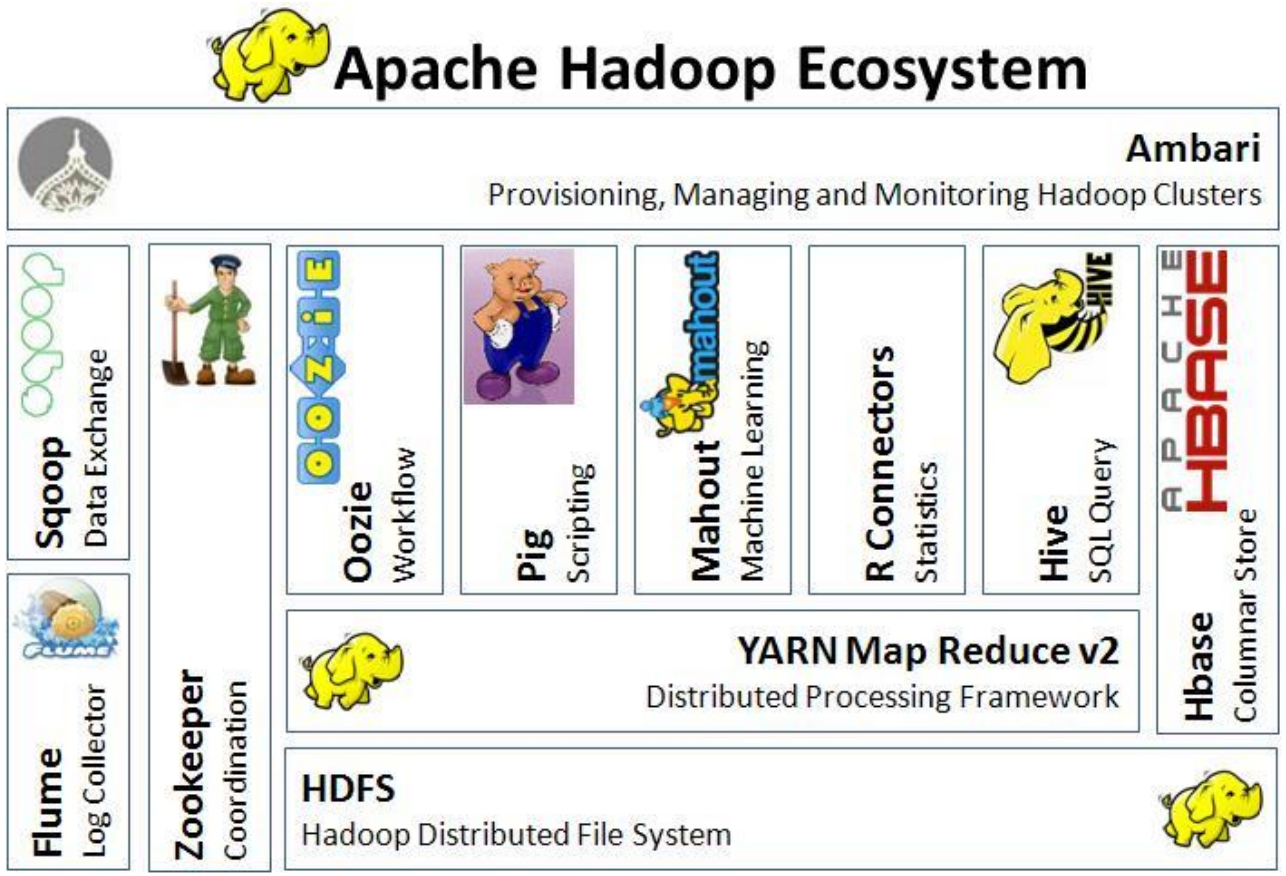


	ETL	ELT
Source data	Some	"All"
Data transfer	Primarily batch	Bulk and streaming
Data cleansing	Before load	Deferred
Master data standardization	Before load	Deferred

BIG Data Technologies

- What is Hadoop?
  - Originally created by Doug Cutting at Yahoo!
  - Not a single product
  - A collection of applications
  - A framework or platform
  - Consists of several modules
  - Hadoop is Open Source
  - Anyone can download or modify
  - Hadoop is now part of Apache Software Foundation

Hadoop Ecosystem and Vendors



Vendor	Enhancement/Extension
Cloudera	Impala
Pivotal	HAWQ
IBM	Big SQL

http://thebigdatablog.weebly.com/blog/the-hadoop-ecosystem-overview

Hadoop is an Entire Ecosystem

1. Data storage environment
2. Languages, tools and APIs
3. Vendor enhancements and extensions
4. Hadoop Distributed File System (core of Hadoop)
   - It stores and distributes files across many different computers
5. Map Reduce
   - Map splits a task into pieces
   - Reduce combines the output
   - Has been replaced by YARN

MapReduce + Hadoop = Big Data core technology

Other Big Data Technologies

- Pig- writes MapReduce programs.
  - Uses the Pig Latin language
- Hive – summarises queries, analyses data, and uses the HiveQL Language.
- Other additional components (major players)
  - Hbase
  - Storm
  - Spark
  - Giraph

Hadoop

Where does Hadoop Go?

- Can be installed on any computer.
- Cloud computing providers are more common.

Who uses Hadoop?

- Yahoo
- LinkedIn
- Facebook
- Quantcast
- Others

- Opensource
  - It's free
  - Anyone can download or modify

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

Relational Database Vs Hadoop Vs NoSQL

- Relational Database
  - - de facto standard for Database Management with a highly structured relational model – this is a burden with dealing with large volumes
- What is Hadoop?
  - It is not a type of database, but a software ecosystem that allows for massively parellel computing. It is an enabler of certain types NoSQL distributed databases (Apache Hbase).
- What is NoSQL?
  - It is a database infrastructure that as been very well-adapted to the heavy demands of big data. The database is unstructured – using the concept of distributed databases (e.g. MongoDB, Cassandra and Hbase)

Resources

- Ted talk – Big Data is Better Data
- http://www.b-eye-network.com/print/17017
- Intro on how Hadoop HDFS works
- Intro on MapReduce
- Hadoop and the Data Warehouse: When to use Whch?
- A 55-minute introduction to NoSQL by Martin Fowler