Introduction to
Big Data and Big Data Analytics

N C Chauhan

Contents

- Introduction
- Big Data and its importance
- Four Vs
- Drivers for Big data
- Big data analytics
- Big data applications

Types of Digital Data

- Structured data:
  - Data that follow a precise data model
  - Data which is in an organized form (rows and columns) and can be easily used by computer program
  - E.g. Data stored in databases
- Unstructured data:
  - Data that does not conform to a data model or is not in a form which can be used easily by a computer program.
  - Data of any organization like memo, chat rooms, ppts, images, videos, letters, body of emails, etc.
  - E.g. data on social media, Log files, etc.
- Semi-structured data:
  - Data that does not conform to a data model but has some structure.
  - It is not in a form which can be used easily by a computer program.
  - Emails, markup languages like XML and HTML, etc.
  - Meta data for this data is available but not sufficient.

Definitions of Big Data

- Big data is an evolving term used to describe any voluminous amount of structured, semi-structured and unstructured data that has potential to be mined for information.
- Big data is high-volume, high velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making. (Gartner – IT Glossary)

Big data is a collection of data sets that are large and complex in nature.

They constitute both structured and unstructured data that grow large so fast that they are not manageable by traditional relational database systems or conventional statistical tools.

How much data? - Facts

- We create 2.5 quintillion ($1x10^{18}$) bytes every day
- 90% of world's data was created in the last 2 years
- 80% of world's data is unstructured
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data
- Facebook processes 500TB per day
- 72 hours of video are uploaded to youtube every minute
- Over 5 billion people use cell phones to call, send SMS, email, browse Internet, and interact via social networking sites.

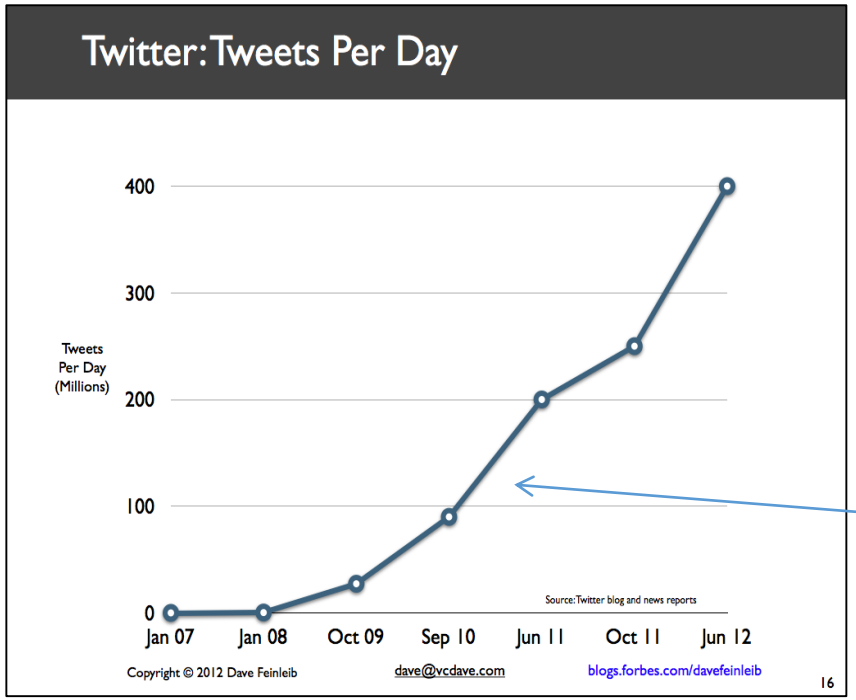
Analytics Challenges with Big Data

- Traditional RDBMS fail to handle Big Data
- Big Data (terabytes) can not fit in the memory of a single computer
- Processing of Big Data in a single computer will take a lot of time
- Scaling with the traditional RDBMS is expensive

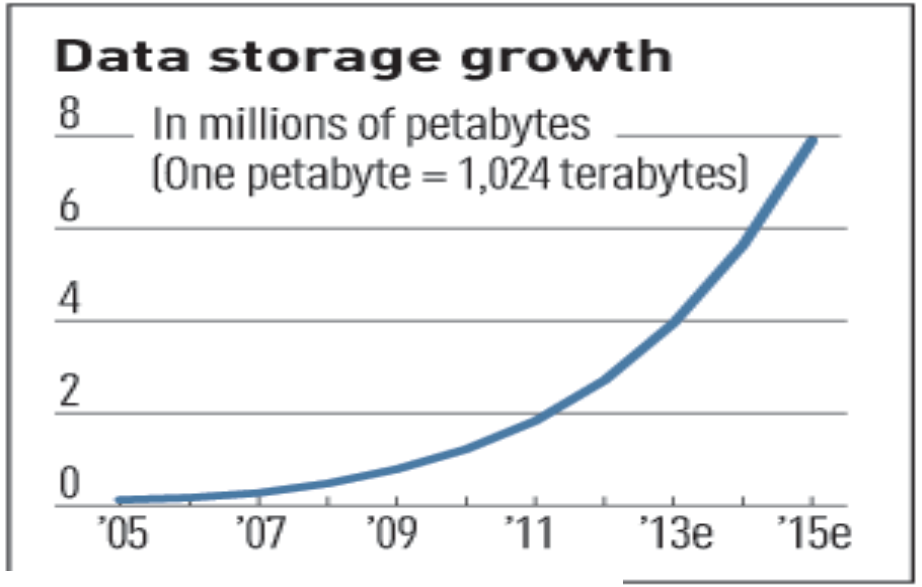
Characteristics of Big Data (3 V’s)

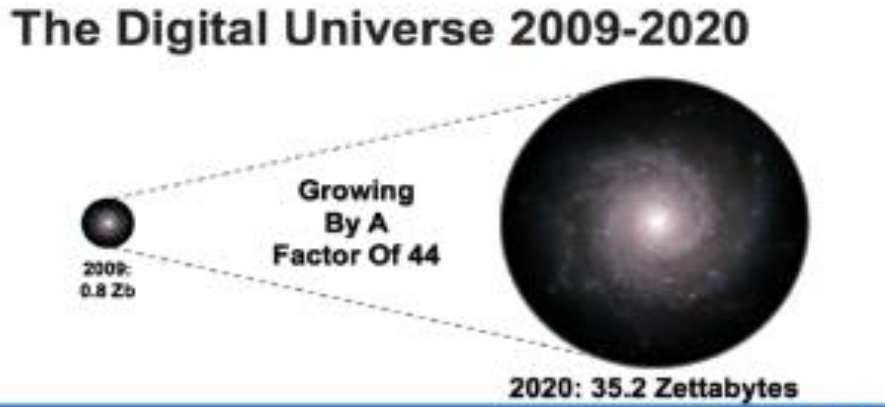
- Volume
- Velocity
- Variety

Volume (Scale)



Exponential increase in collected/generated data





- Dat
  -
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

terabytes | petabytes | exabytes | zettabytes

the amount of data stored by the average company today

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

76 million smart meters in 2009… 200M by 2014

2+ billion people on the Web by end 2011

Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

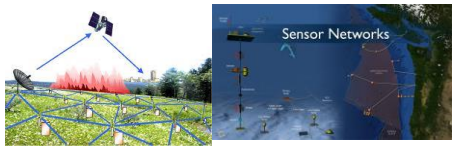
- **Examples**
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction
  - **Theft Detection:**

Real-time/Fast Data

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)

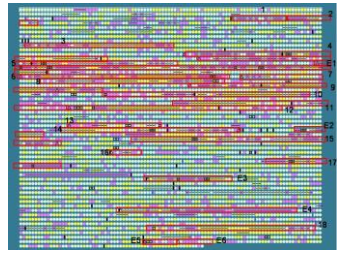


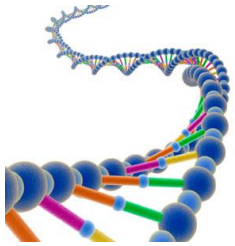
Sensor technology and networks
(measuring all kinds of data)

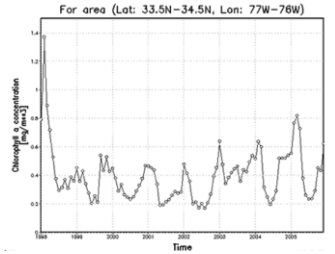
- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

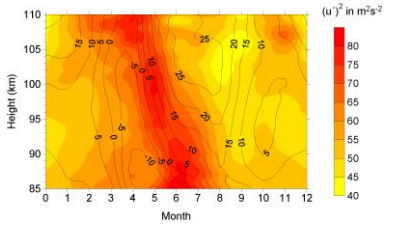
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



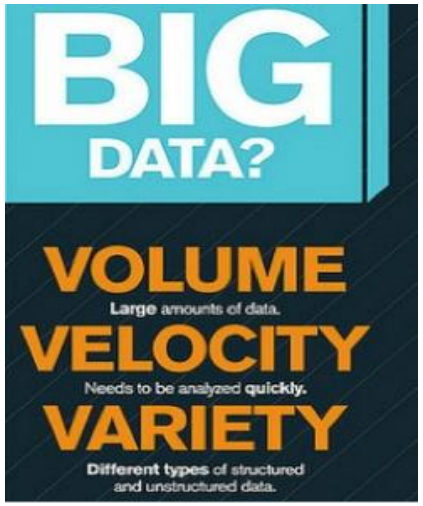


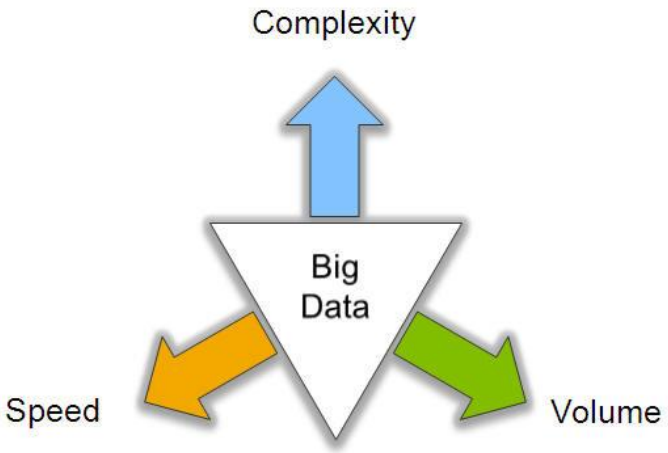




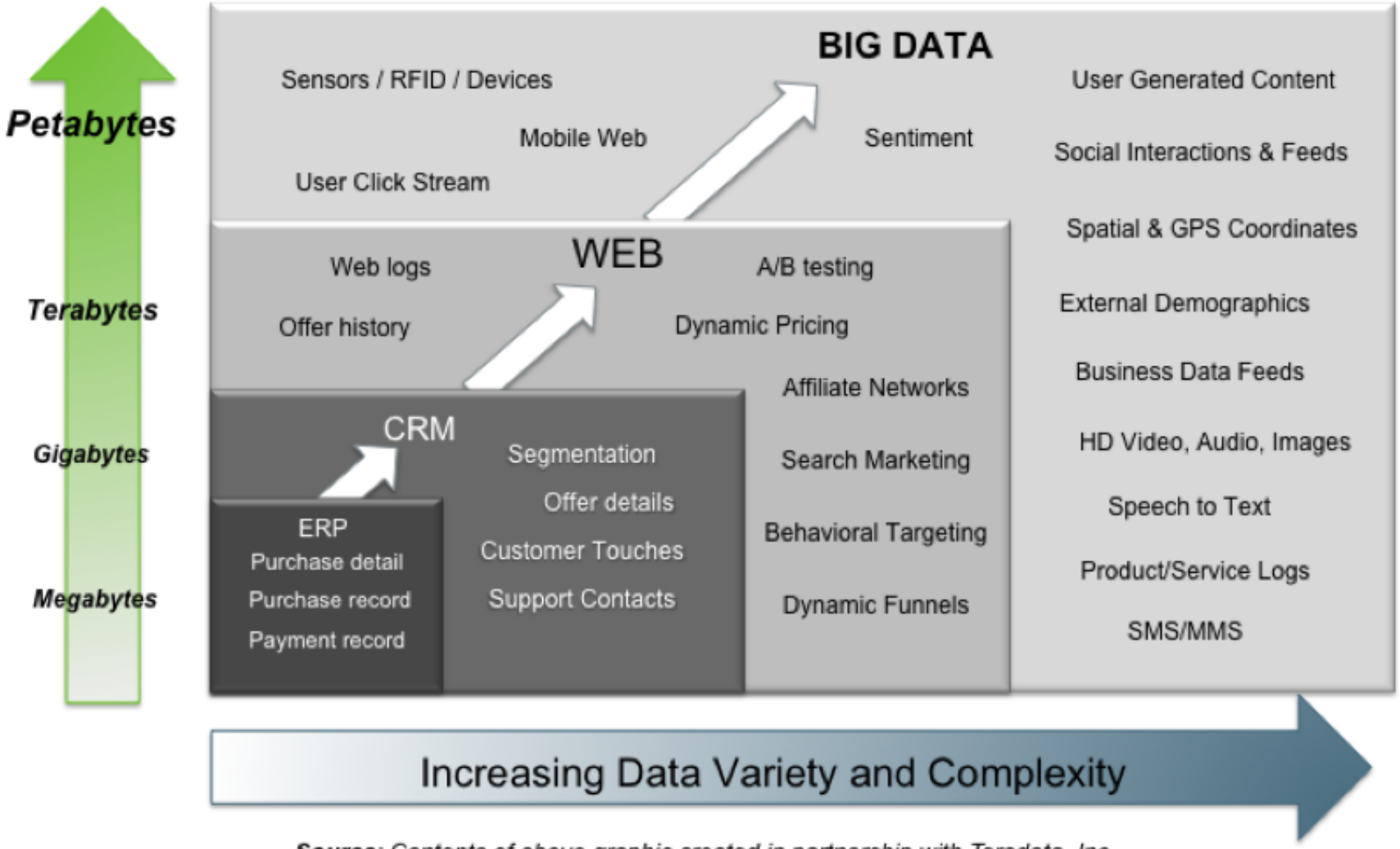
To extract knowledge➔ all these types of data need to linked together
Challenge: Data Integration

Big Data: 3V's



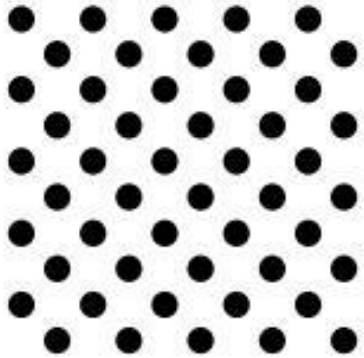
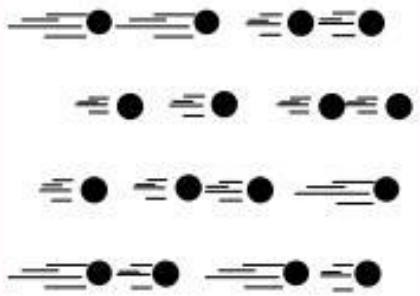
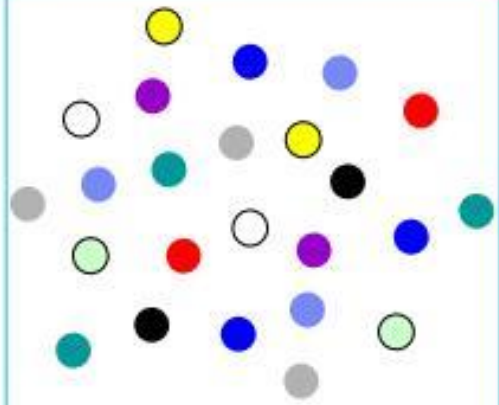
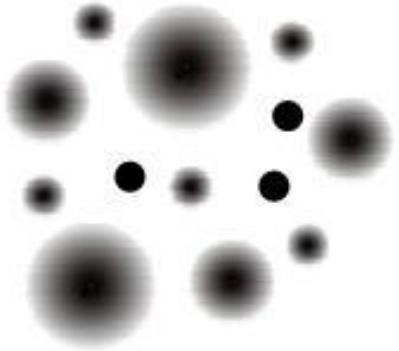


Big Data = Transactions + Interactions + Observations



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

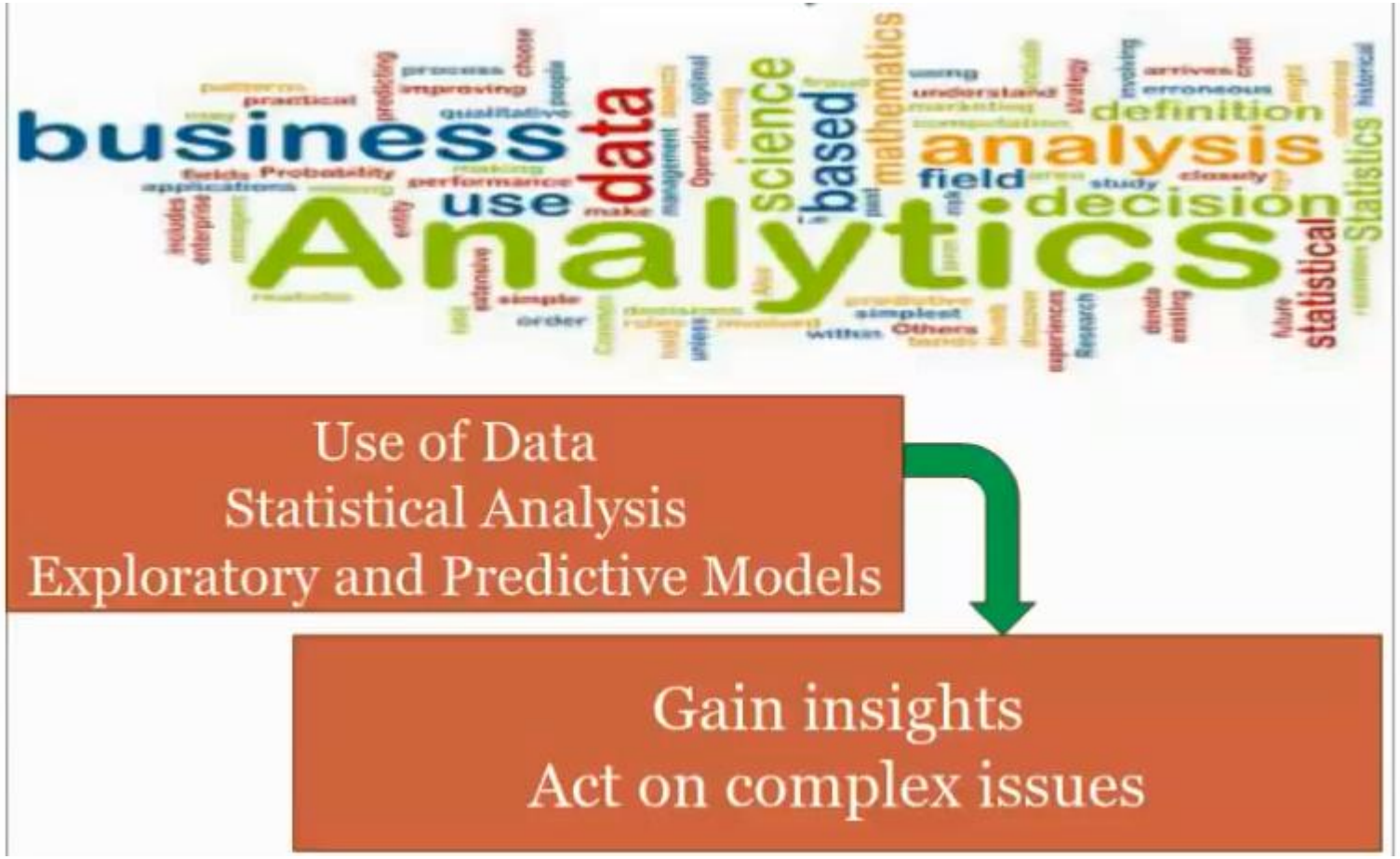
Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Some Big Data Issues Affecting Analytics

- Volume:
  - How much data is really relevant to the problem solution? Cost of processing?
  - *So, can you really afford to store and process all that data?*
- Velocity:
  - Much data coming in at high speed
  - Need for streaming versus block approach to data analysis
  - *So, how to analyze data in-flight and combine with data at-rest*
- Variety:
  - A small fraction is structured formats, Relational, XML, etc.
  - A fair amount is semi-structured, as web logs, etc.
  - The rest of the data is unstructured text, photographs, etc.
  - *So, no single data model can currently handle the diversity*
- Veracity: cover term for …
  - Accuracy, Precision, Reliability, Integrity
  - *So, what is it that you don't know you don't know about the data?*
- Value:
  - How much value is created for each unit of data (whatever it is)?
  - *So, what is the contribution of subsets of the data to the problem solution?*

What is Analytics?



4 Types of Analytics

Types of Analytics

- ***Descriptive***: A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance.
- ***Diagnostic***: A set of techniques for determine what has happened and why
- ***Predictive***: A set of techniques that analyze current and historical data to determine what is most likely to (not) happen
- ***Prescriptive***: A set of techniques for computationally developing and analyzing alternatives that can become courses of action – either tactical or strategic – that may discover the unexpected
- ***Decisive***: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making when presented with a set of alternatives.

What is Big Data Analytics?

- Big data analytics is a process of:
- Collecting
- Organizing and
- Analyzing

of large sets of data ("big data") to

- Discover patterns and
- Other useful information.

What is Big Data Analytics?

- Big data analytics is:
  - Technology-enabled analytics (through automated tools such as SAS, SPSS, R, Statistica, etc)
  - Quicker and better decision making in real time
  - Richer, deeper insights into customers, partners, and the business
  - IT's collaboration with business users and data scientists
  - Working with datasets whose volume and variety is beyond storage and processing capability of a typical database software
  - Moving code to data for greater speed and efficiency

What Big Data Analytics is not?

- Only about volume
- Just about technology
- Meant to replace RDBMS
- Meant to replace data warehouse
- Only used by huge online companies like Google or Amazon
- “One-size fit all” traditional RDBMS built on shared disk and memory

Why sudden hype around BDA?

- Data is growing at a 40% compound annual rate.
- Cost per gigabyte of storage has hugely dropped.
- Availability of number of user-friendly analytical tools in the market
- Opportunities for real time applications in practice

Motivation: Big Data Analytics in Practice

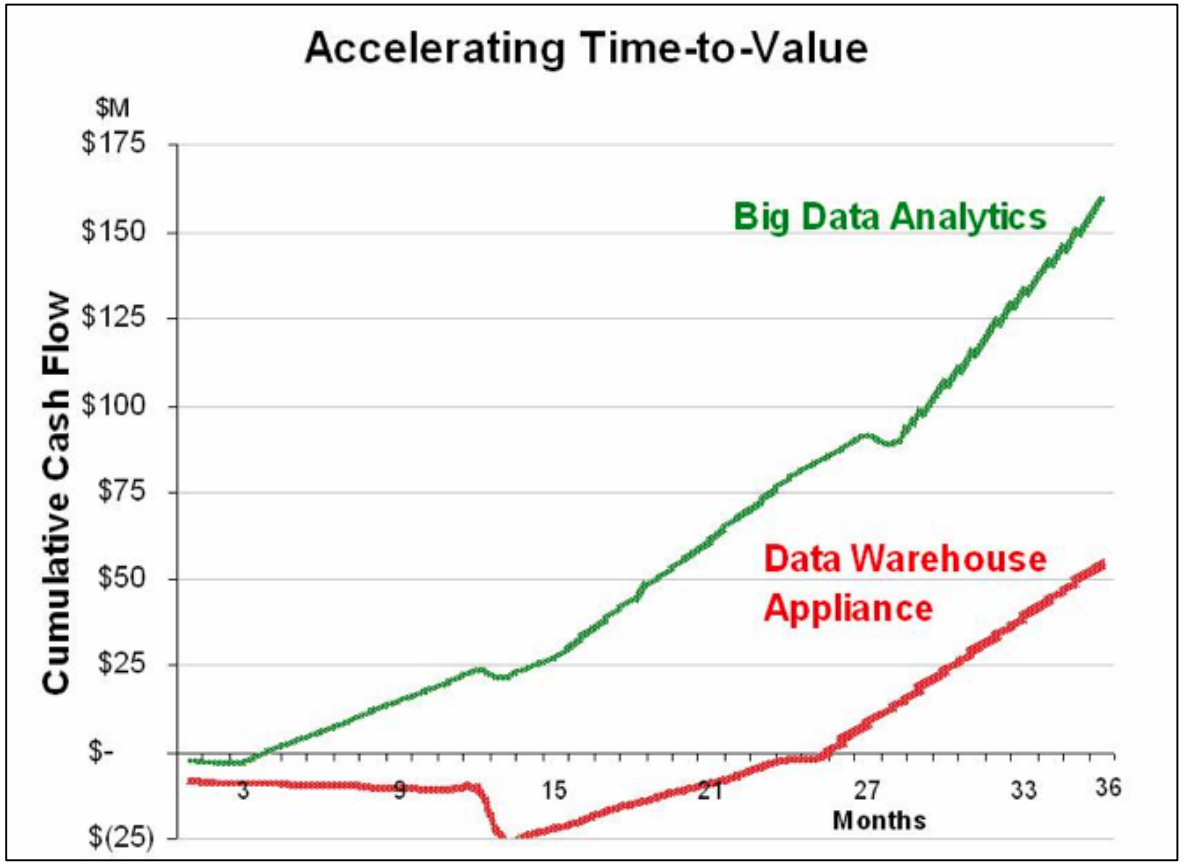
Etihad Airways uses technology to harvest and analyze gigabytes of data generated by hundreds of sensors working inside its planes. This allows to monitor planes in real time, reduce fuel costs, manage plane maintenance, and even spot problems before they happen.

Many people use Facebook to update their status, share photos, and "like" content.
The Obama presidential campaign used all that data on the social network to not just find voters but to assemble an army of volunteers.

One of India's highest-rated TV shows aggregates and analyzes the millions of messages it receives from viewers on controversial issues like female feticide, caste discrimination and child abuse — and uses that data to push for political change.

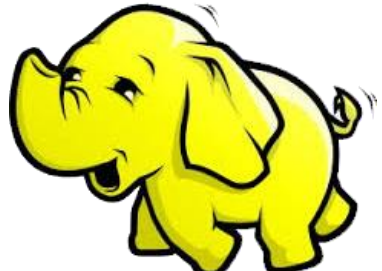
Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



What is Apache Hadoop?

- Open source software framework designed for storage and processing of large scale data on clusters of commodity hardware
- Created by Doug Cutting and Mike Carafella in 2005.
- Cutting named the program after his son's toy elephant.



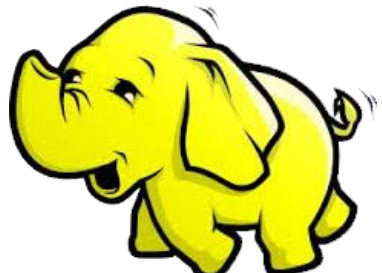
Hadoop's Developers

Doug Cutting

2005: Doug Cutting and Michael J. Cafarella developed Hadoop to support distribution for the Nutch search engine project.

The project was funded by Yahoo.

2006: Yahoo gave the project to Apache Software Foundation.



Who Uses Hadoop?

Key Advantages of Hadoop

- Stores data in its native format
  - No loss of information as there is no translation/transformation to any specific schema
- Scalability:
  - can store and distribute very large datasets across hundreds of inexpensive servers that operate in parallel
  - Proven to scale by companies like Facebook and Yahoo
- Cost effective:
  - Reduced cost/terabyte of storage and processing
- Resilient to failure:
  - higher availability – Faulty tolerance through replication of data
- Flexibility:
  - ability to work with all kind of data
  - Can derive meaningful business insights
- Fast:
  - processing is fast compared to other conventional systems (“move code to data paradigm”)

Analytics Tools

- Most used statistical programming tools
  - IBM SPSS
  - SAS
  - Stata
  - R
  - MATLAB

R and MATLAB have the most comprehensive support of statistical functions.