Big Data - Introduction

Anurag Nagar

Big Data

- Data is all around you.
- In recent years there has been a shift in the type of data:

Structured -> Unstructured

Fixed, pre-determined units -> Variable units

Think of Facebook posts, tags, likes

Twitter posts, re-tweets, etc

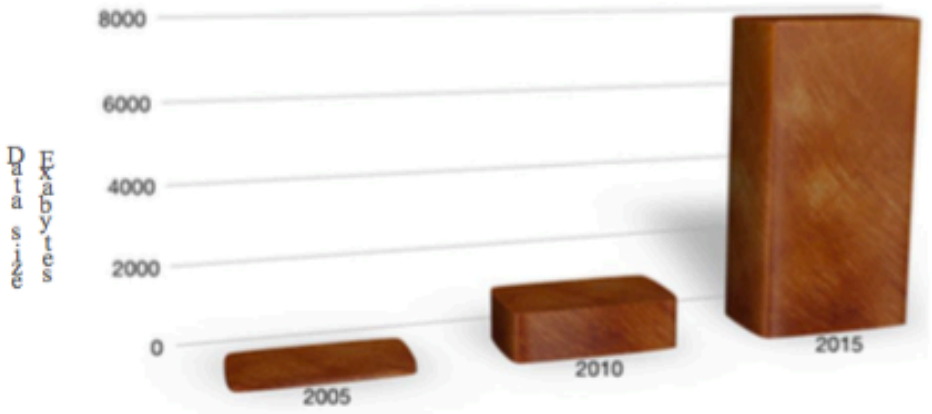
Can you determine beforehand the numbers?

Smaller size -> Very large sizes

Lot of time for analysis -> Instant analysis

How Big is Big Data?

- Big is a fast moving target: kilobytes, megabytes, gigabytes, terabytes (10^{12}), petabytes (10^{15}), exabytes (10^{18}), zettabytes (10^{21}),……
- Over 1.8 zb created in 2011; ~8 zb by 2015



Source: IDC's Digital Universe study, sponsored by EMC, June 2011
http://idcdocserv.com/1142
http://www.emc.com/leadership/programs/digital-universe.htm

As of June 2012

Nature of Big Data: Volume, Velocity and Variety

Big Data on the Web

As of the third quarter of 2017, Facebook had 2.07 billion monthly active users
Facebook users send on average 31.25 million messages and view 2.77 million videos every minute.

Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.

Source:

http://www.internetlivestats.com/twitter-statistics/
https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Big Data on the Web

Over 50 billion pages indexed and more than 2 million queries/min

Articles from over 10,000 sources in real time

In 2015, a staggering 1 trillion photos were taken and billions of them will be shared online. By 2017, nearly 80% of photos will be taken on smart phones.

Every minute up to 300 hours of video are uploaded to YouTube alone.

Source:

https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5305a62c17b1

Defining Big Data

- Various definitions exist:
  Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

  - Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science 7: 1–5.

Defining Big Data

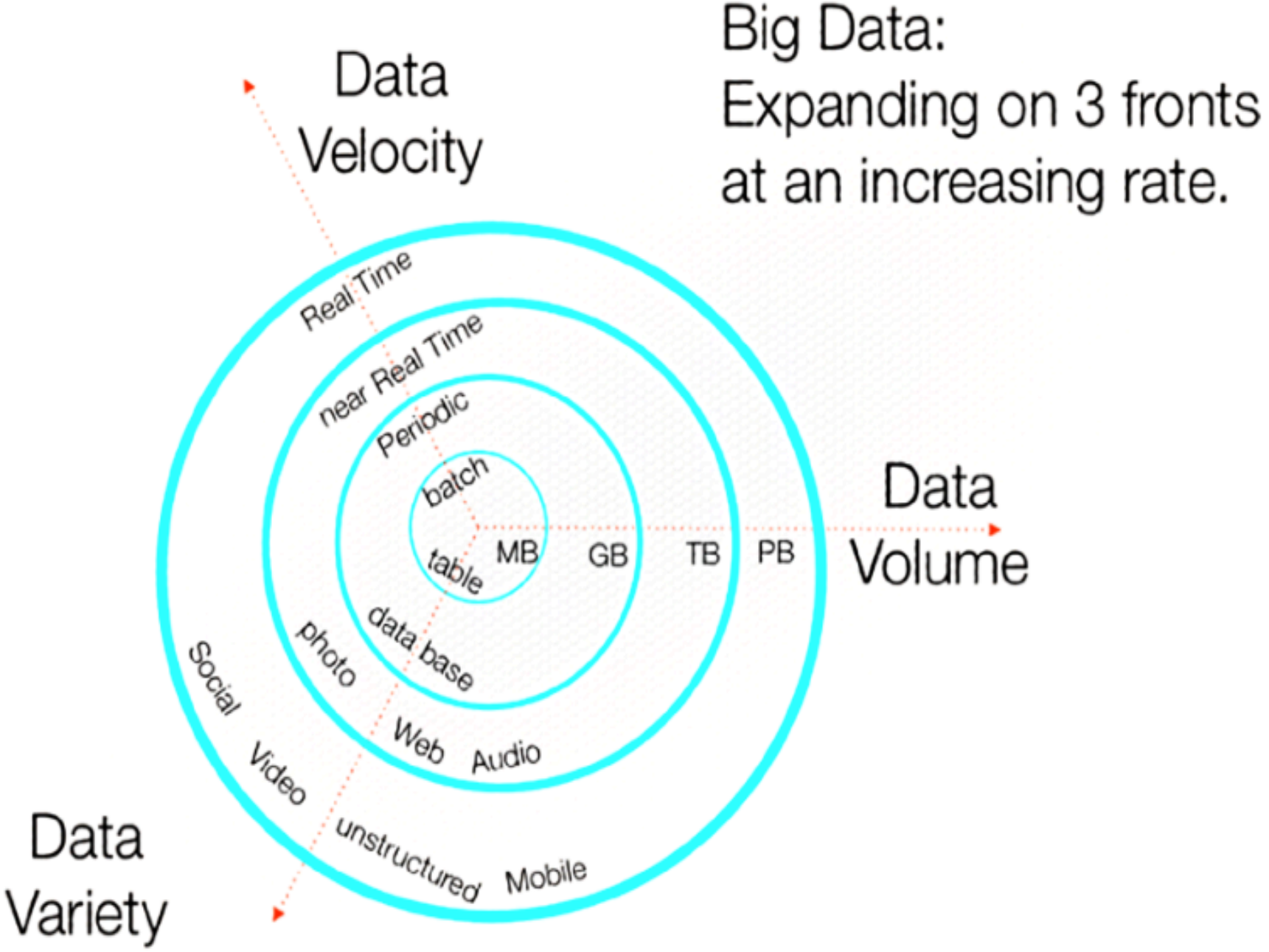
“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

Defining Big Data

- Gartner defined it as:

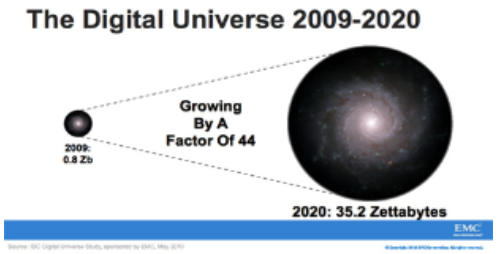
Data growth challenges and opportunities are three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).

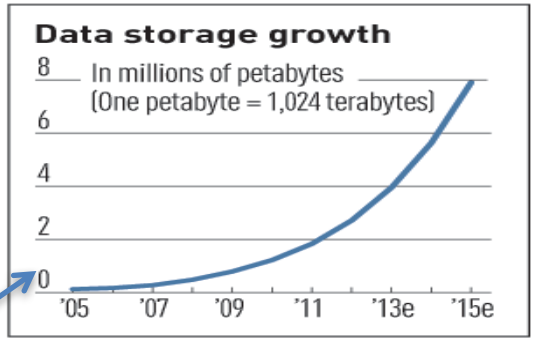
- Doug Laney of Gartner group

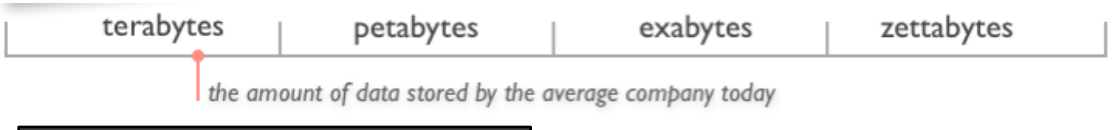

Big Data:
Expanding on 3 fronts
at an increasing rate.
Data
Velocity
Data
Volume
Data
Variety
Real Time
near Real Time
Periodic
batch
table
MB
GB
TB
PB
data base
photo
Social
Web
Audio
Video
unstructured
Mobile

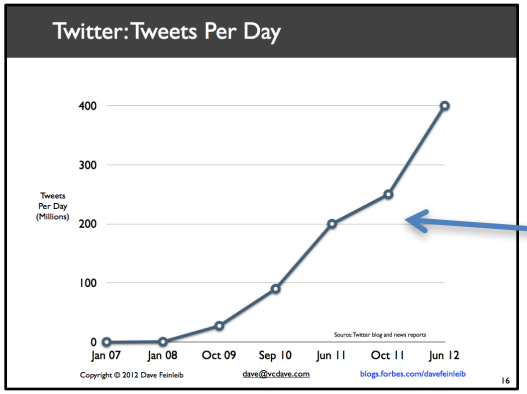
Characteristics of Big Data:
1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





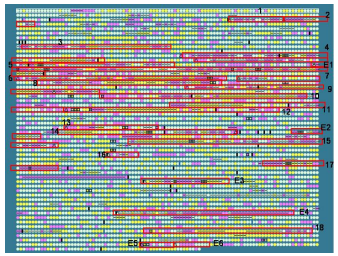


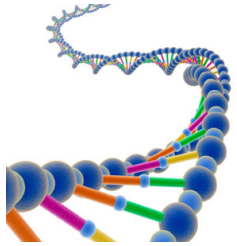


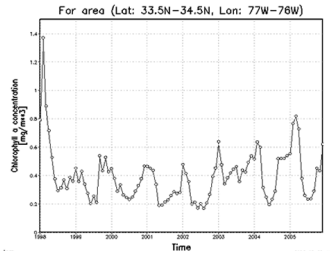
Exponential increase in collected/generated data

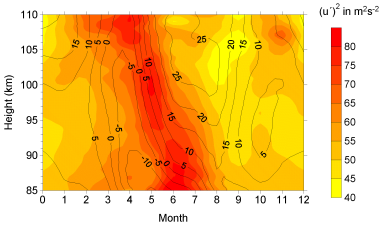
Characteristics of Big Data:
2-Complexity (Variety)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









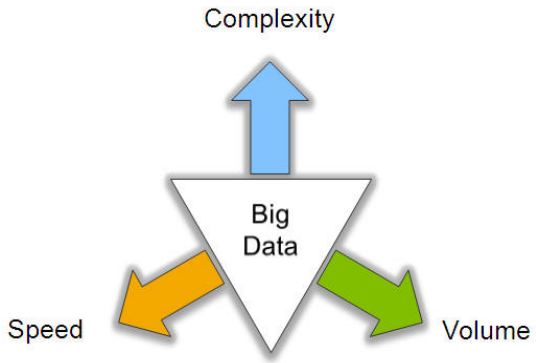
To extract knowledge➔ all these types of data need to linked together

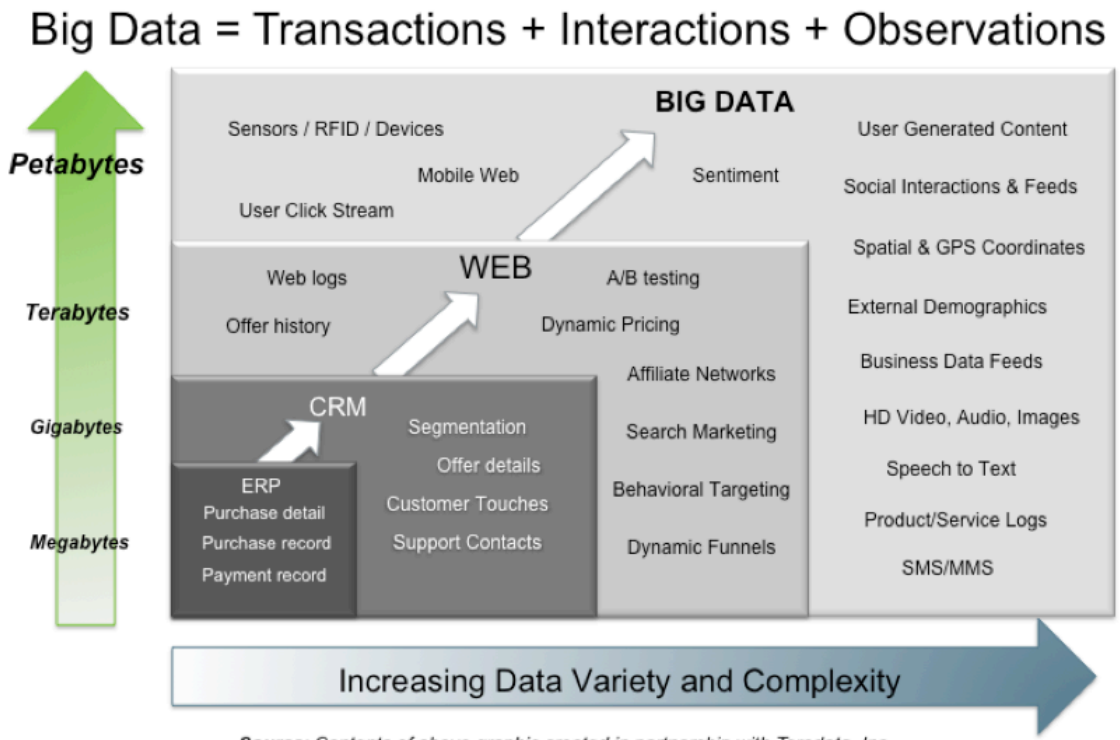
Characteristics of Big Data: 3-Speed (Velocity)

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

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

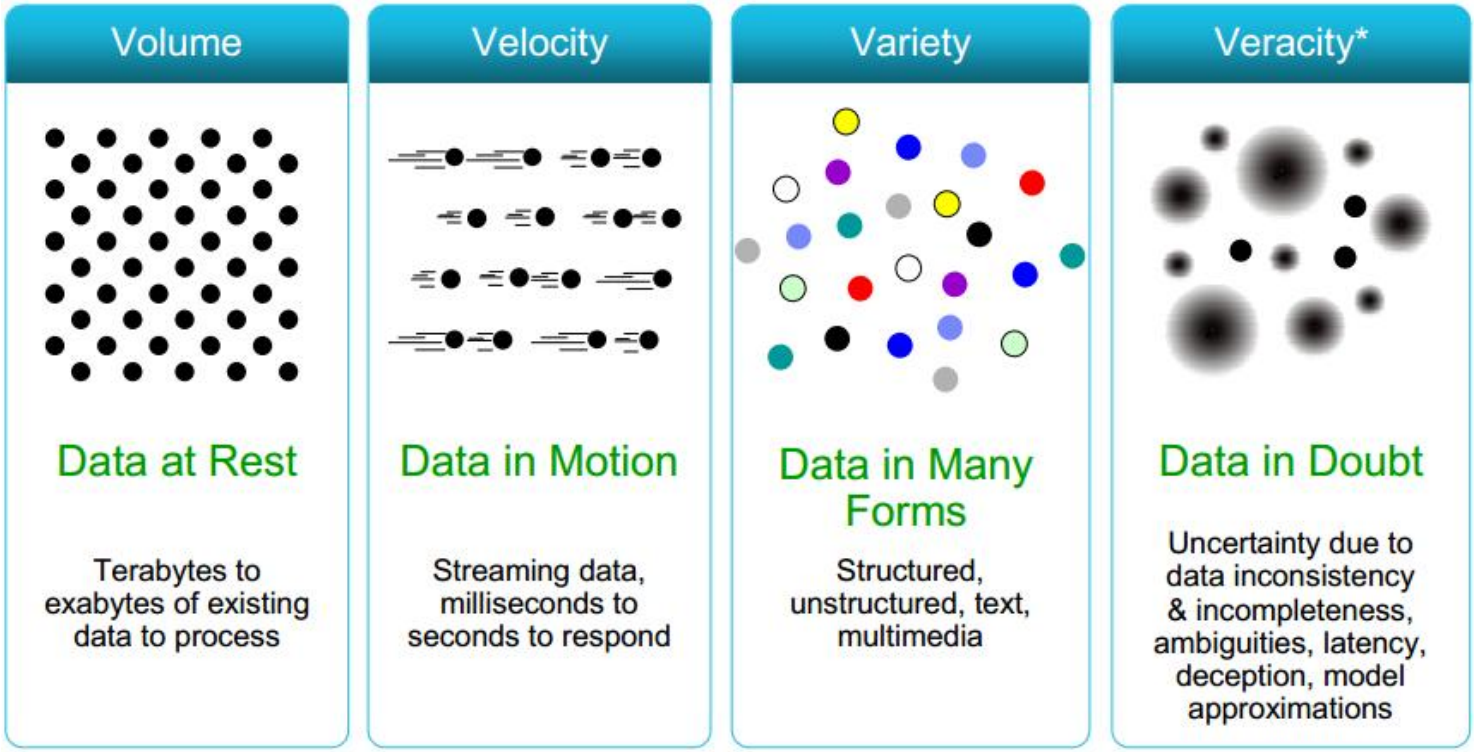
Big Data: 3V's



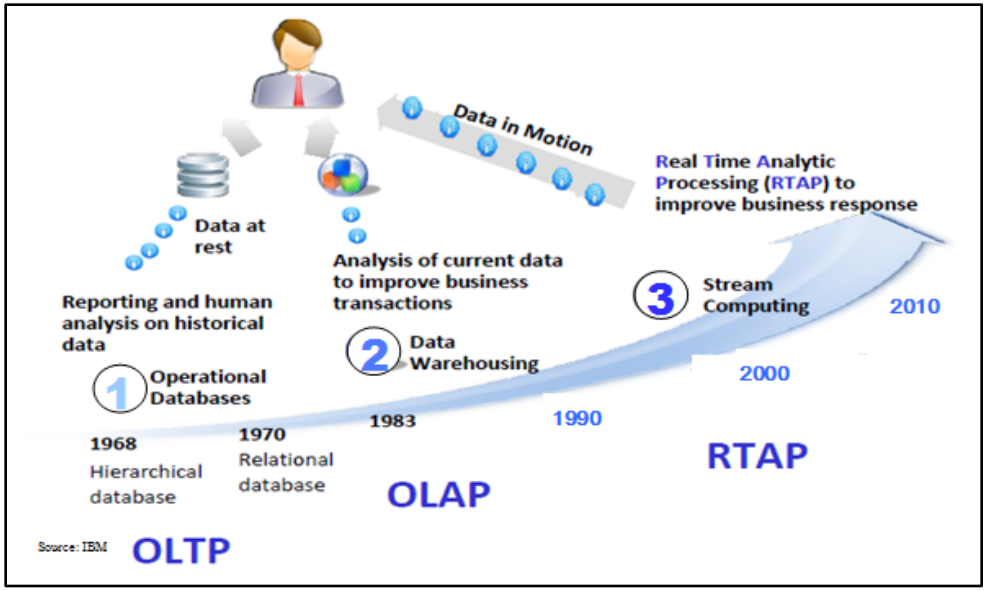


***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Some Make it 4V's



Harnessing Big Data



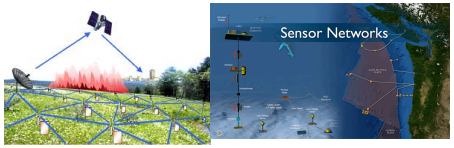
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Who's Generating Big 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)



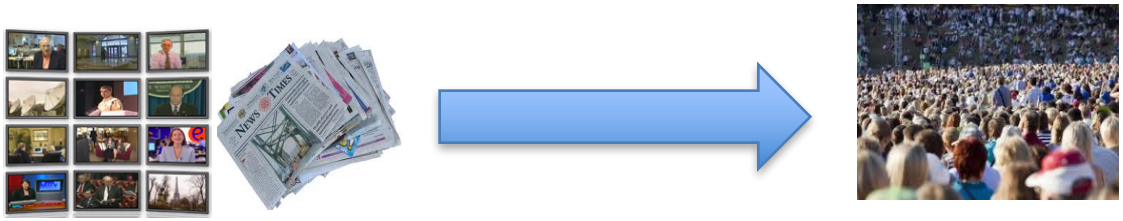
Sensor technology and networks
(measuring all kinds of data)

- 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

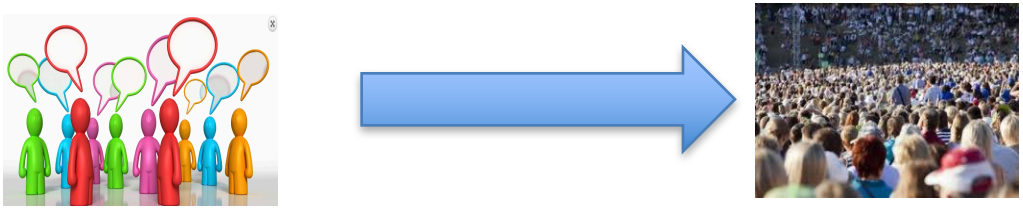
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

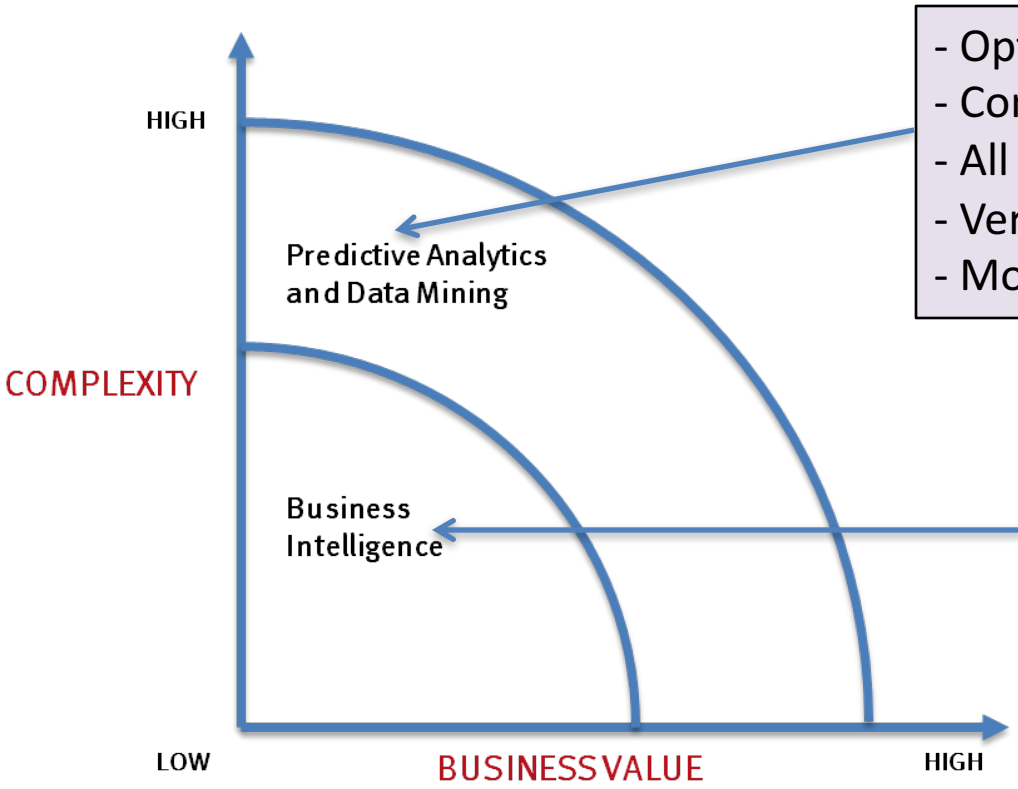
Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data



What's driving Big Data

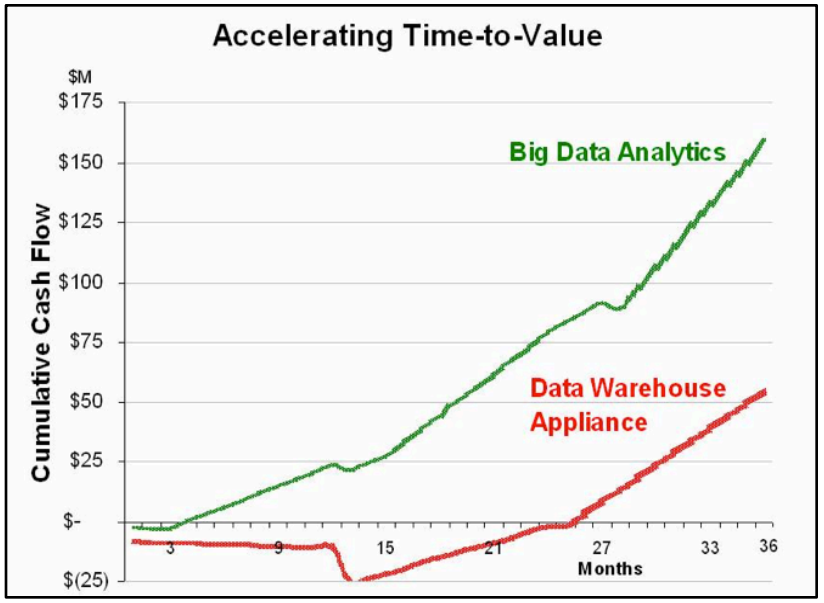


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

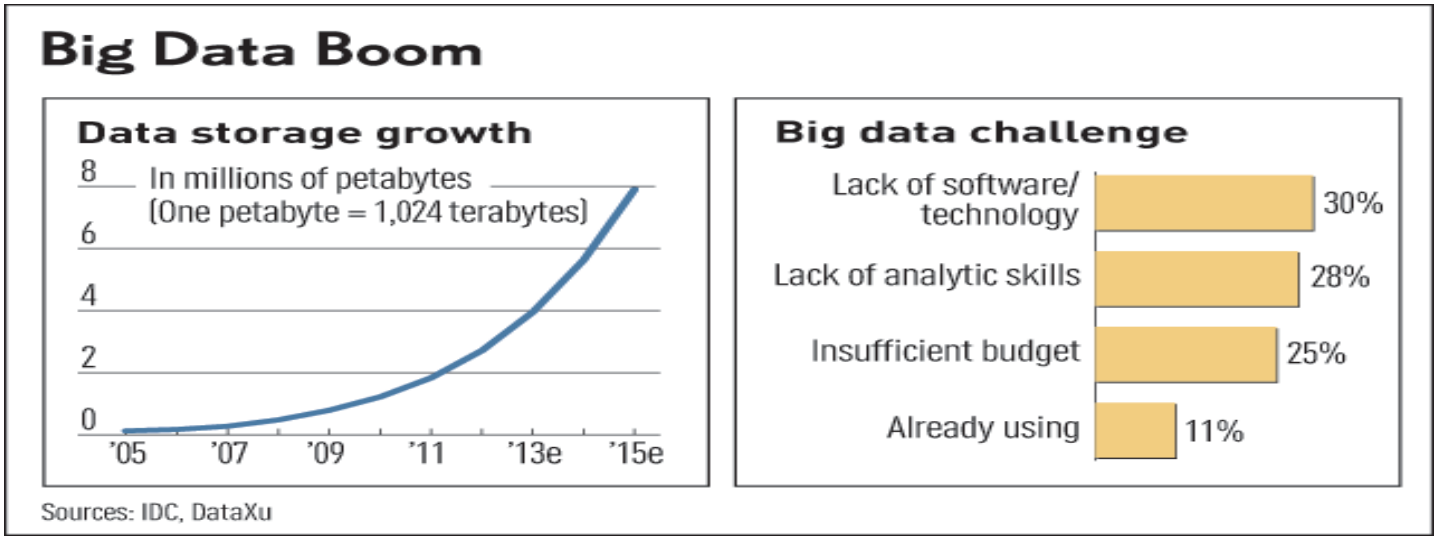
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Value of 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, scaled out architectures are well-suited for big data apps



Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

Type of Data

- 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

What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

What Technology Do We Have For Big Data ??

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft | Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS TIBCO KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx visual.ly AYATA

Data As A Service

factual
kaggle
knoema
GNIP DATASIFT Windows Azure Marketplace INRIX LexisNexis LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
ParAccel
EMC² GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

Couchbase
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies

hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

What is this class about?

- Learning foundations of Big Data
- Learning Hadoop Distributed File System
  - forms the basis of what we will cover
  - makes processing of Big data easy, affordable, and fault-tolerant.
- Learning MapReduce programming framework
  - splits task across multiple nodes
  - advantages over traditional programming
- Higher level frameworks for Big Data
  - Pig, Hive,
- Non-relational datastores
  Cassandra, ..

What is this class about?

- Machine Learning and Analysis using Big Data
  - Assumption: Students know basis of ML
- Mix of theoretical and practical work
- Projects

What this class is NOT

- A theoretical study of distributed systems
- An in-depth study of just one of the topic areas. e.g. Spark Machine Learning, Design of recommender systems, Mahout classification techniques

$\Rightarrow$ You can choose one of these areas for your project.

- A course in Machine Learning, Data Mining, AI

$\Rightarrow$ You should already have some knowledge

- Study of the *entire* Hadoop stack in depth