Big Data Analytics

Data Science and Analytics

Large-Scale Data Management

- How to manage very large amounts of data and extract value and knowledge from them

Introduction to Big Data

What is Big Data?
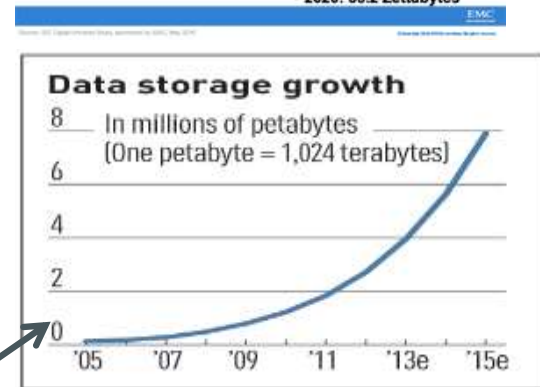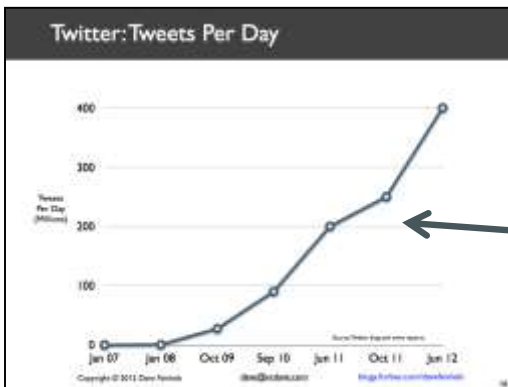
What makes data, "Big" Data?

Big Data Definition
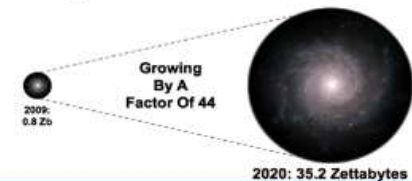
- No single standard definition…

"**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…

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

The Digital Universe 2009-2020

Growing By A Factor Of 44

2009: 0.8 Zb

2020: 35.2 Zettabytes

Data storage growth
In millions of petabytes (One petabyte = 1,024 terabytes)

terabytes	petabytes	exabytes	zettabytes

the amount of data stored by the average company today

Twitter: Tweets Per Day
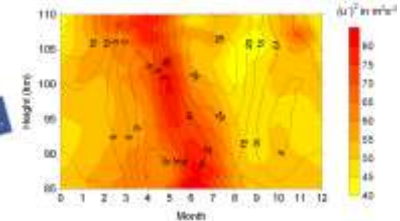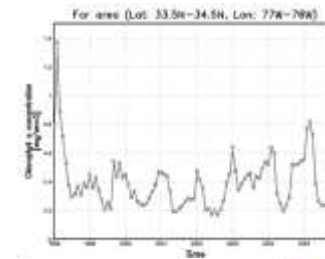
Exponential increase in collected/generated data

Characteristics of Big Data:
2-Complexity (Varity)

- 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
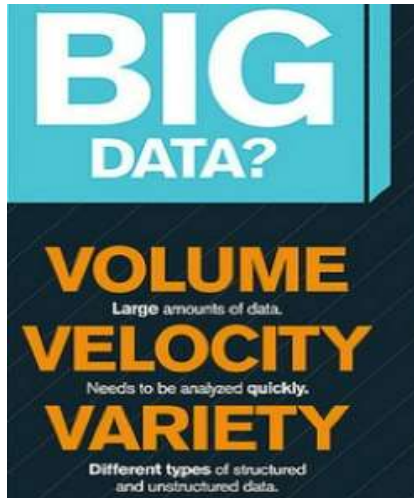
- Online Data Analytics

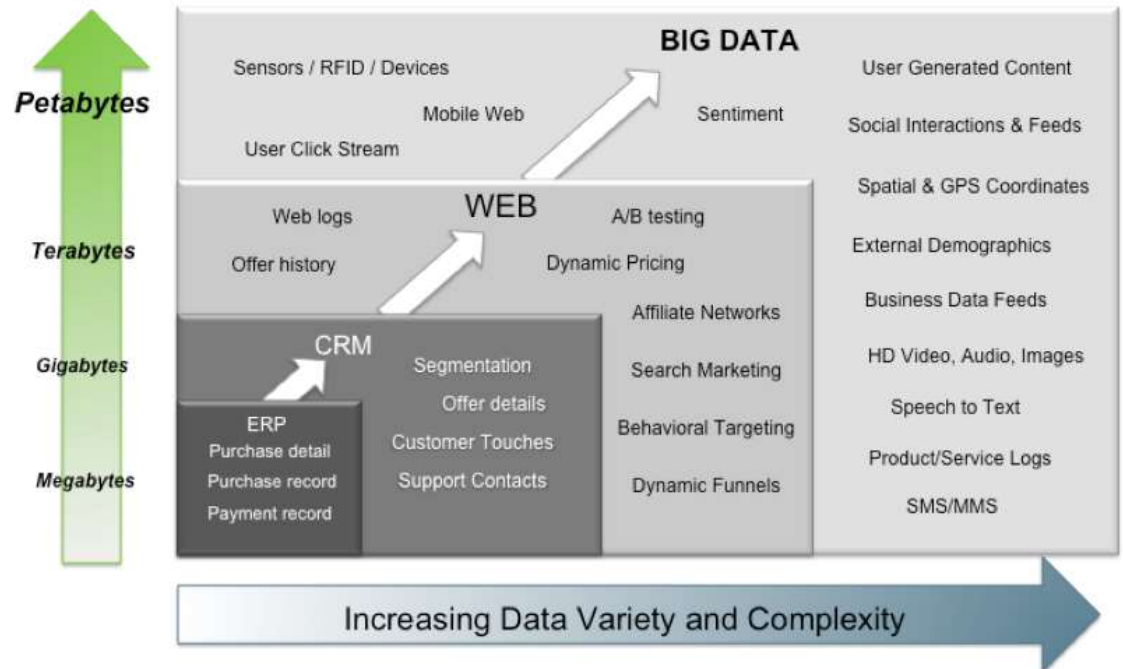- Late decisions ➜ missing opportunities

- **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

Some Make it 4V's
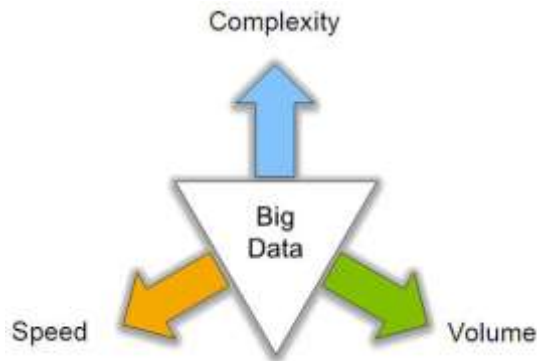
Volume	Velocity	Variety	Veracity*
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
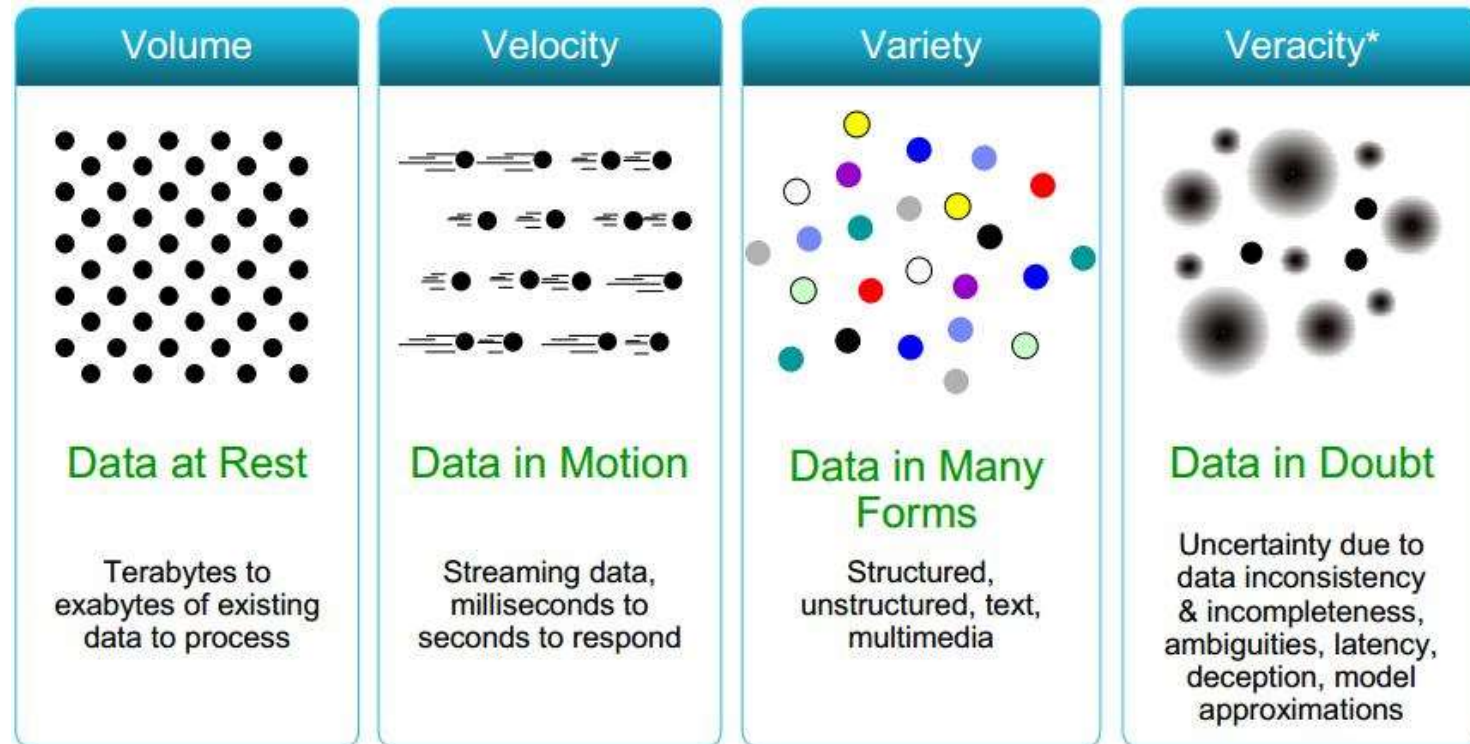
Harnessing Big Data
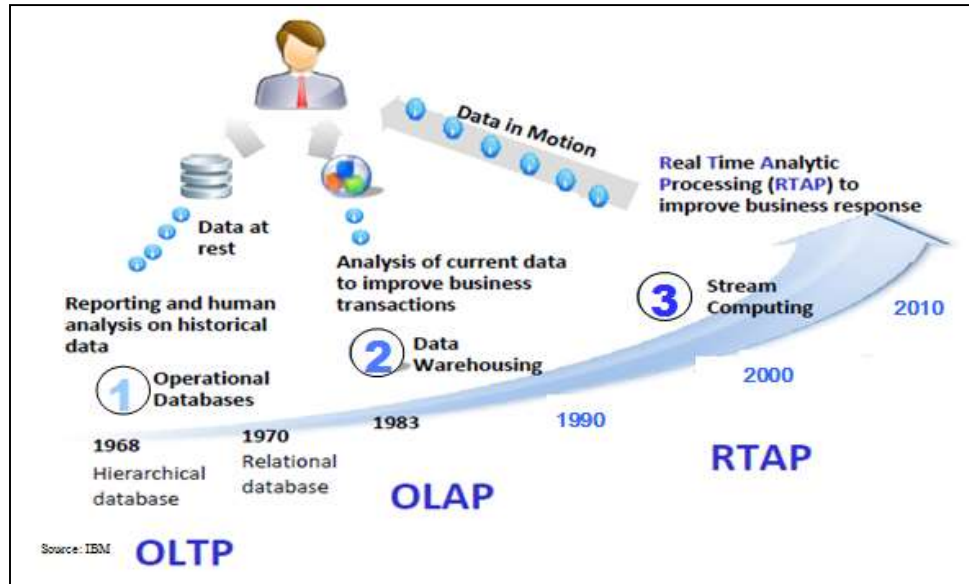
- **OLTP:** Online Transaction Processing   (DBMSs)

- **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

COMPLEXITY

HIGH

Predictive Analytics and Data Mining

Business Intelligence

LOW

BUSINESS VALUE

HIGH

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, scale out architectures are well-suited for big data apps

Accelerating Time-to-Value

Cumulative Cash Flow ($M): $175, $150, $125, $100, $75, $50, $25, $-, $(25)

Months: 3, 9, 15, 21, 27, 33, 36

Big Data Analytics

Data Warehouse Appliance
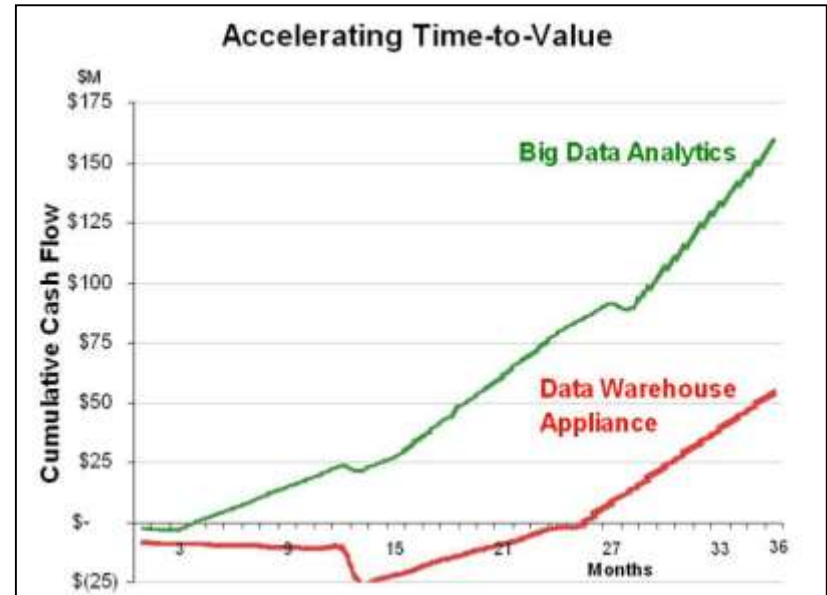
Challenges in Handling Big Data
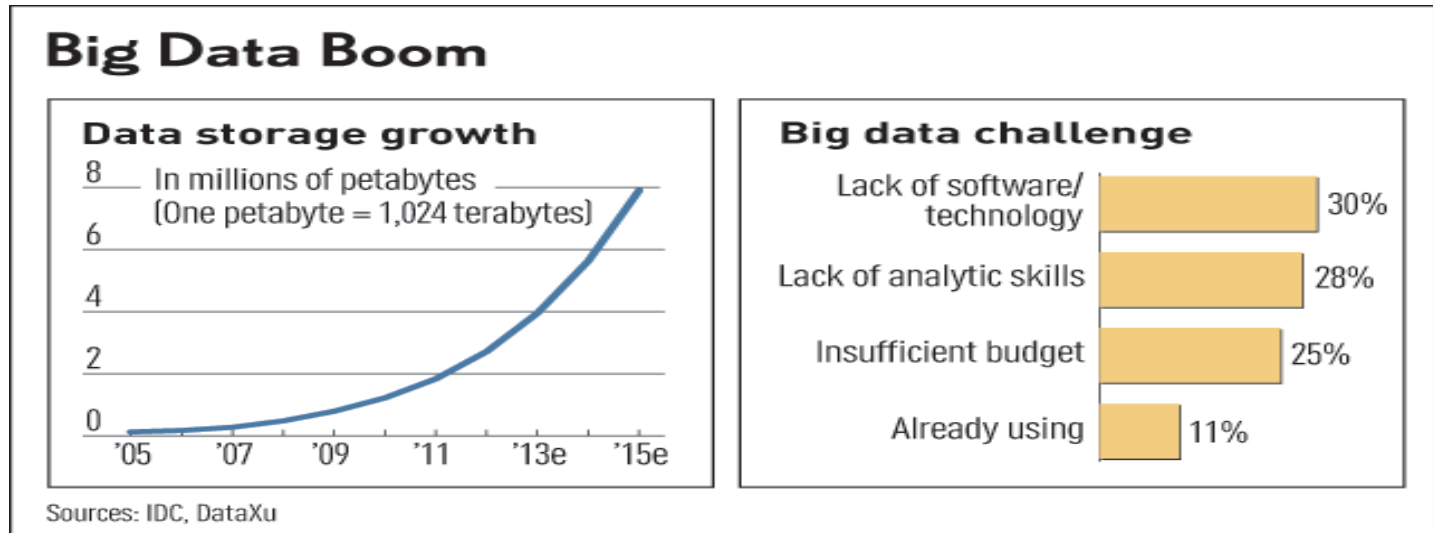
- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed

- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

What Technology Do We Have

For Big Data ??

Big Data Landscape

Vertical Apps
PREDICTIVE POLICING
bloomreach GET FOUND.
MYRRIX

Log Data Apps
splunk> loggly sumologic

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

Data As A Service
factual.
kaggle
knoema beta
GNIP DATASIFT Windows Azure Marketplace INRIX LexisNexis LOQATE

Business Intelligence
ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
QlikView bime DOMO
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

Analytics Infrastructure
Hortonworks VERTICA An HP Company MAPR
cloudera INFOBRIGHT
EMC² GREENPLUM ParAccel
NETEZZA kognitio
DATASTAX EXASOL cal

Operational Infrastructure
COUCHBASE 10gen MongoDB company
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 Map Reduce
mahout
APACHE HBASE
Cassandra

16

dave@vcdave.com
blogs.forbes.com/davefeinleib

Big Data Technology

Big Data: The Moving Parts

From http://blogs.zdnet.com/Hinchcliffe
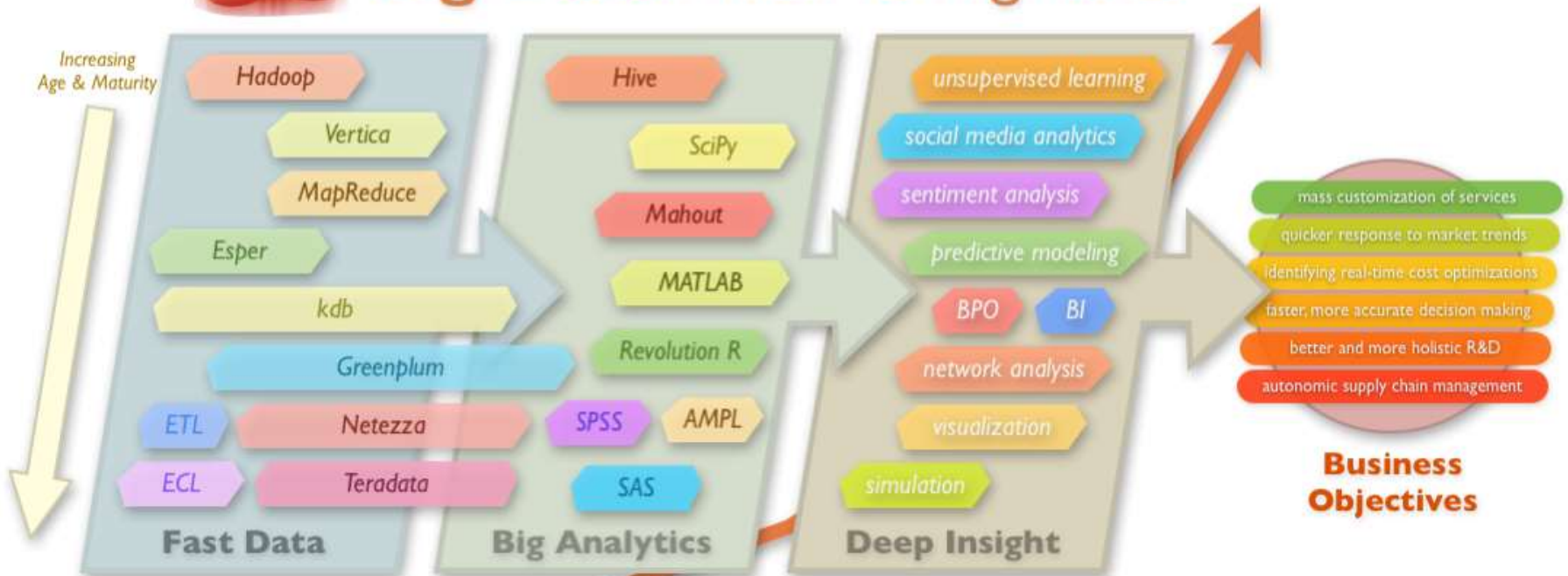
the growth of data will be exponential for the foreseeable future

| terabytes | petabytes | exabytes | zettabytes |

the amount of data stored by the average company today

17