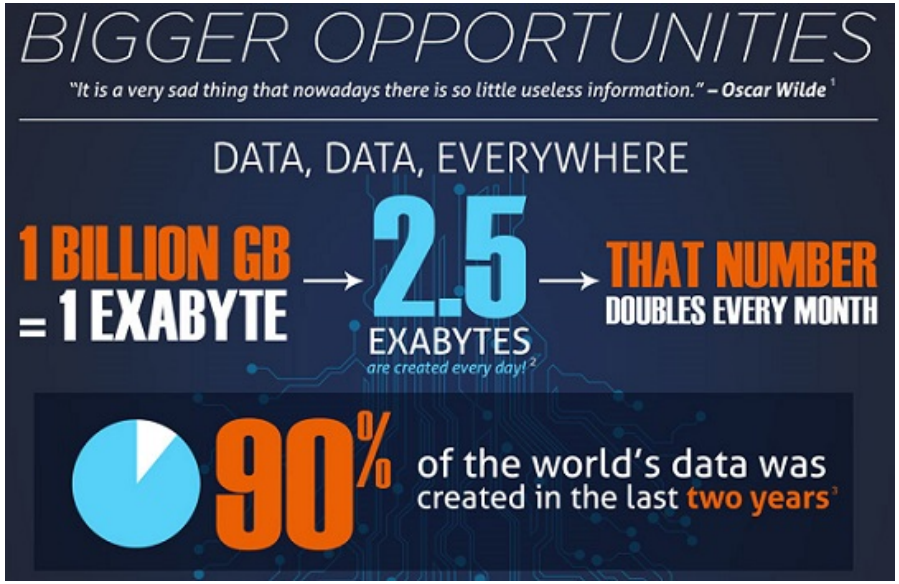
DataInquest

Big Data Technologies and Data Science training for all!!

Big Data 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...



Big Data Buzz!!

$200 million

- The Obama Administration is investing $200 million in big data research projects. *(Source: White House Press)*

1.5m

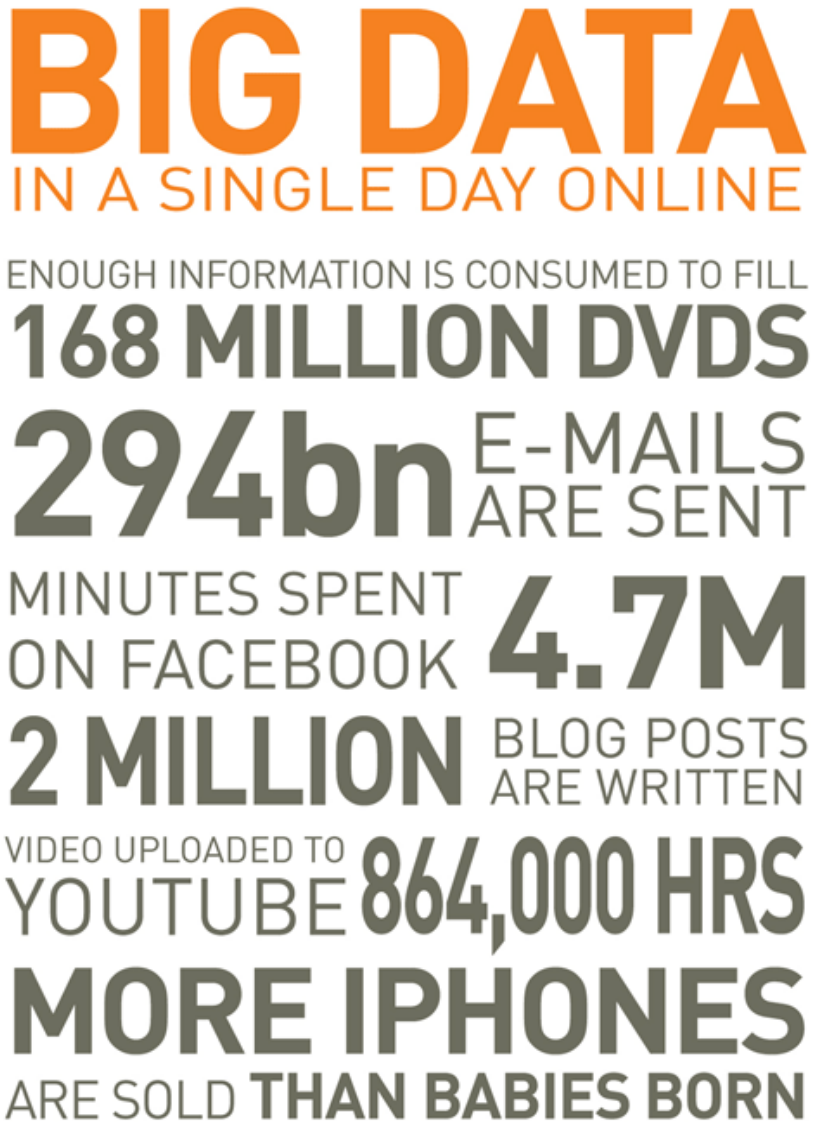
- 140,000 to 190,000 people with deep analytic skills as well as 1.5 million managers and analysts will be needed by 2018 to fill jobs in Big Data *(Source: McKinsey)*

15,000%

- Job postings for data scientists increased 15,000 percent between 2011 and 2012 alone. *(Soure: FICO)*

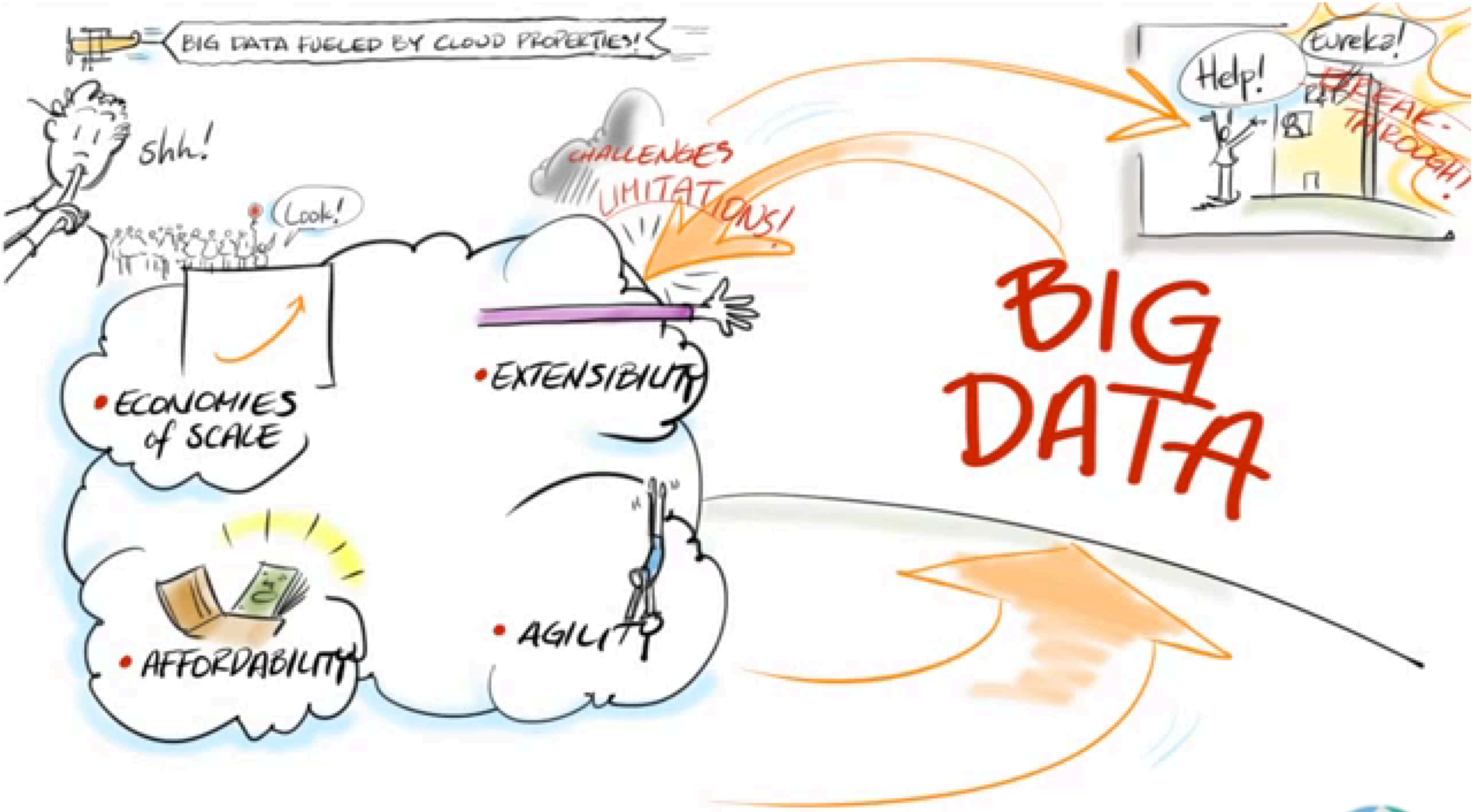
53.4 B

- The big data industry is expected to be a 53.4 billion industry by 2016.*(Source: Domo)*



Big Data Scale!!

How big is Big Data? Cloud Properties

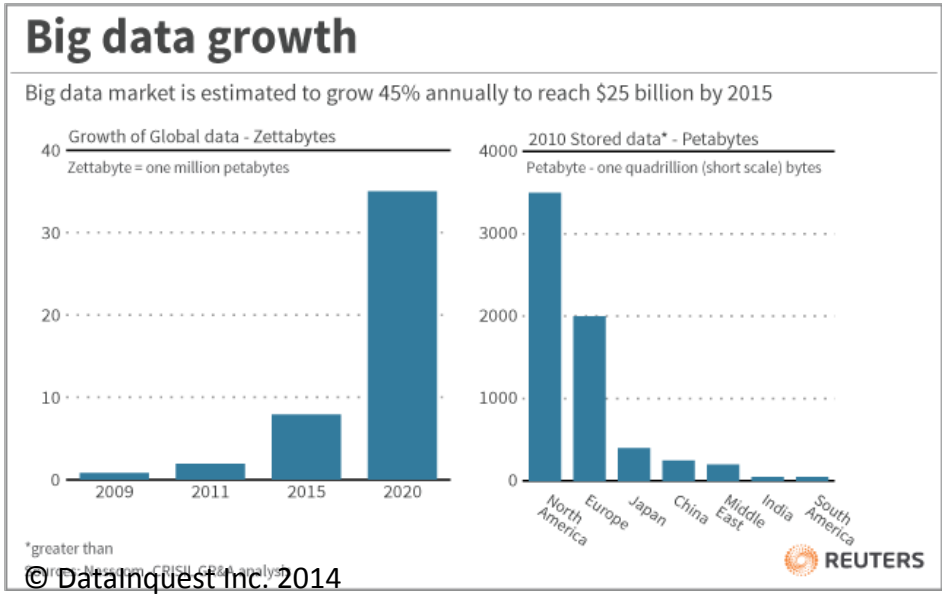


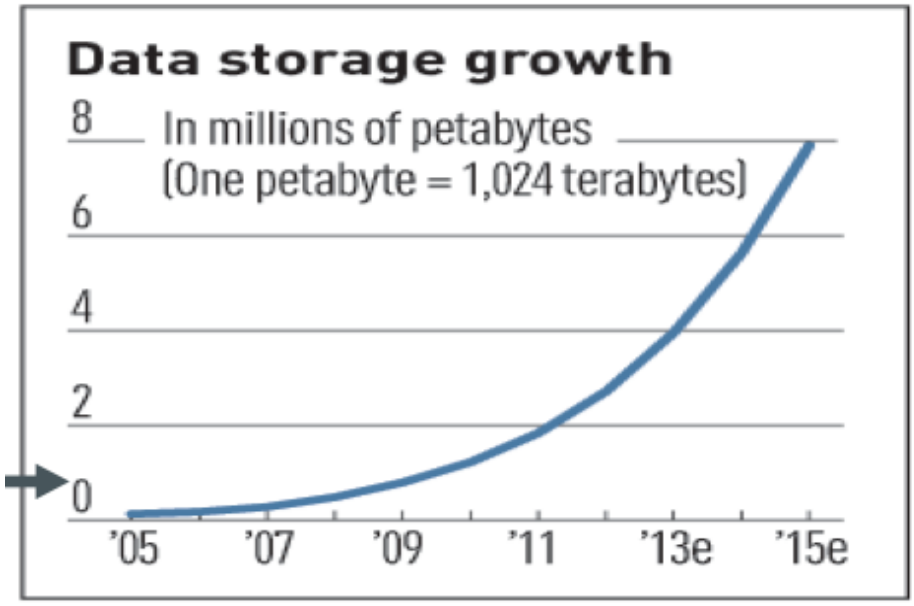
http://www.EMC.com

Characteristics of Big Data – 3Vs

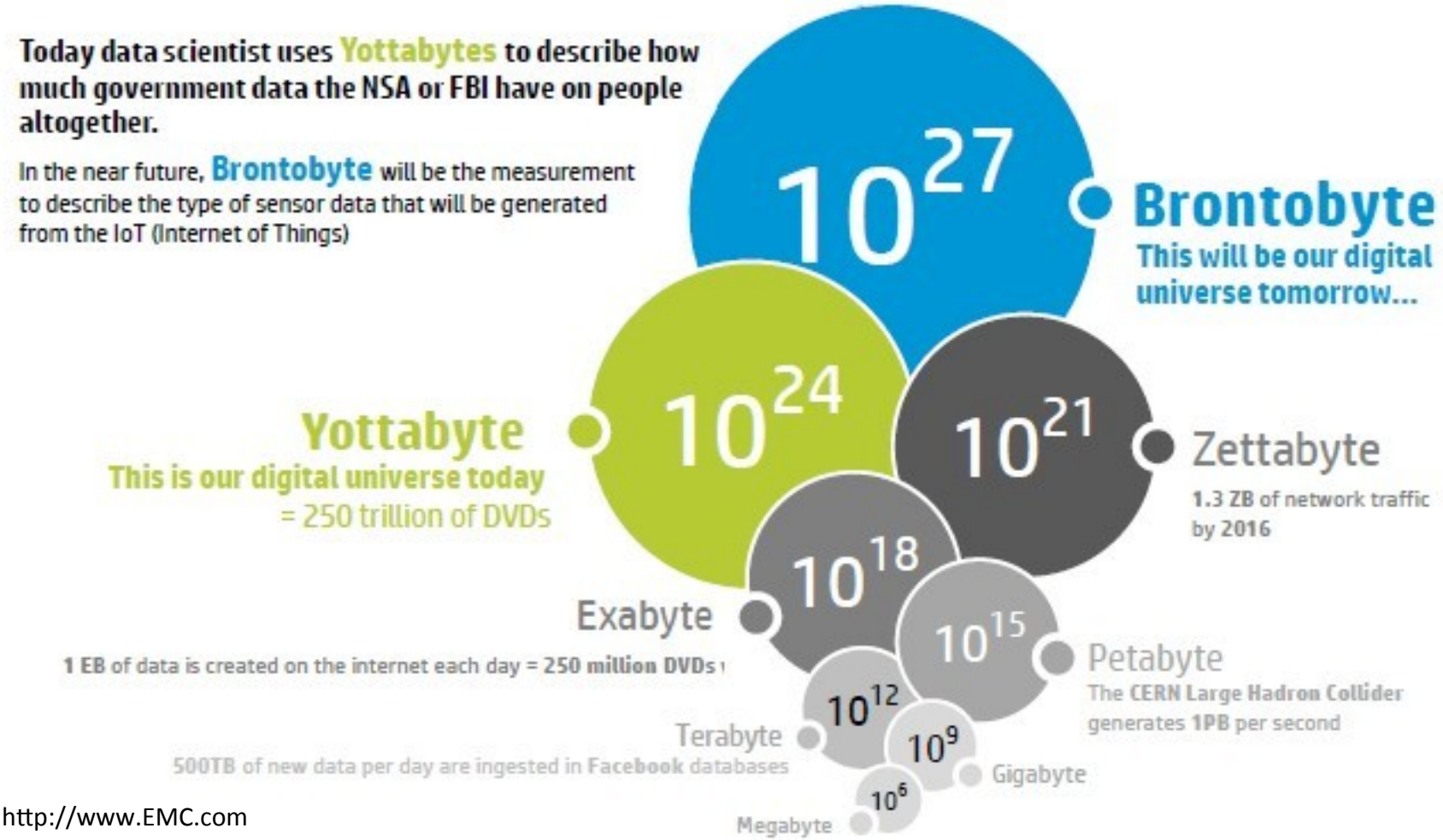
1- Volume

- 90% of the world's data has been created in the last two years - 0.8 zb to 35zb
- 70% of the digital universe—900 exabytes—is generated by users.
- Enterprises store 80% of all data





Big Data Size



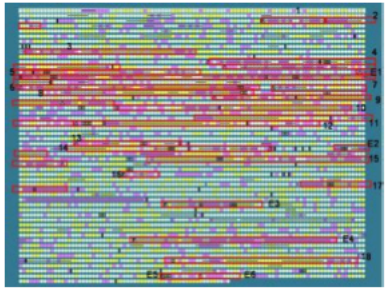
http://www.EMC.com

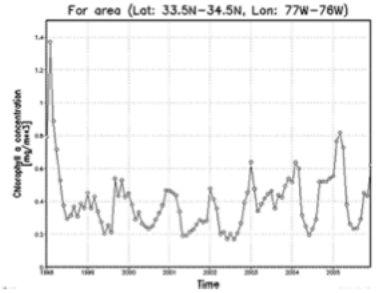
Characteristics of Big Data

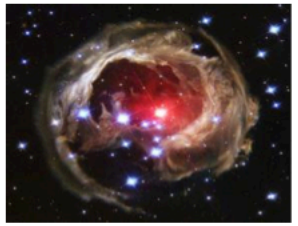
2- Variety

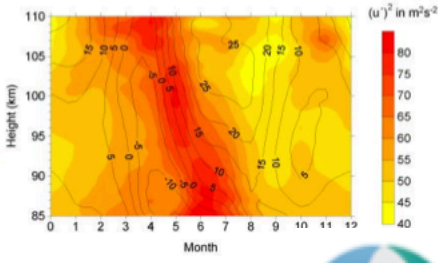
To extract knowledge➔ all these types of data need to linked together

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





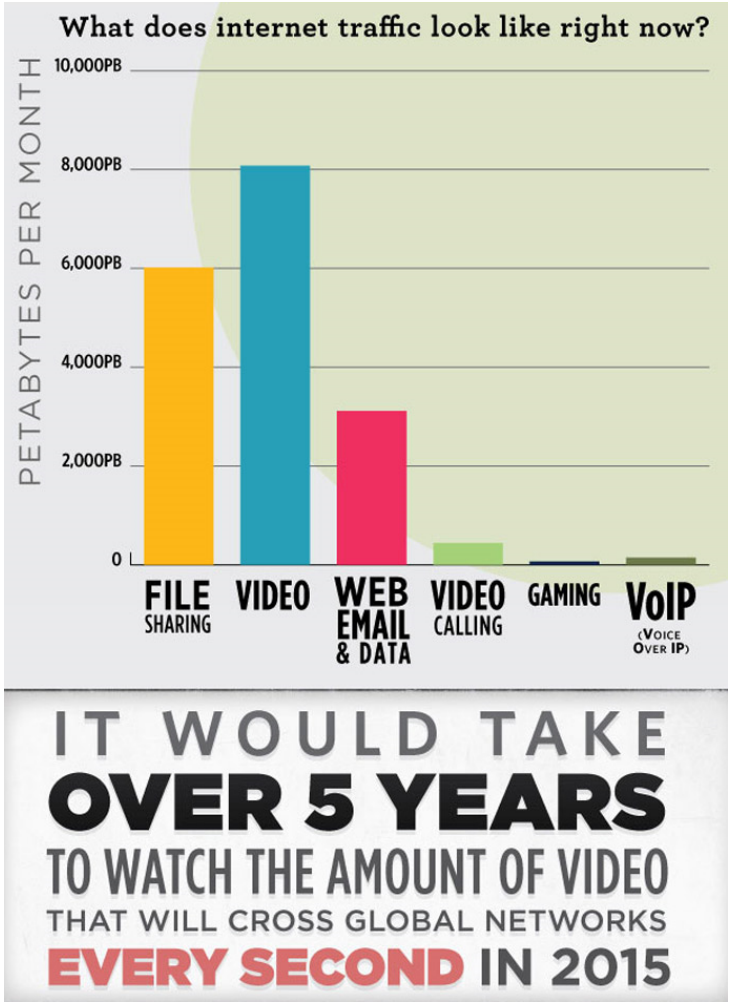




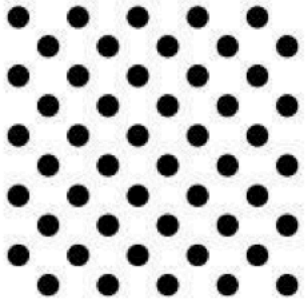
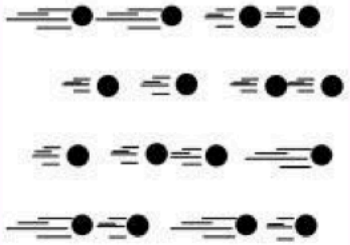
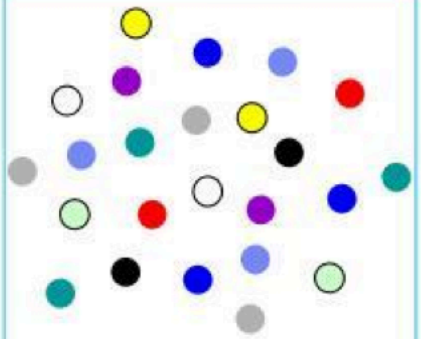
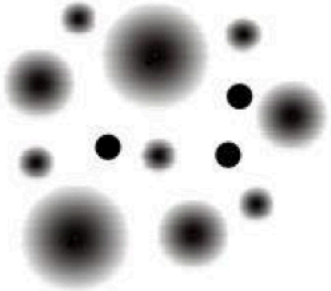
Characteristics of Big Data

3- Velocity

- Data is generated fast and need to be processed fast
- More than 570 new Websites are created every minute of the day
- Internet of things will stream more data in real-time
- Slow Decision → Missed opportunities
  - Realtime promotions based on location, purchase history
  - Device Diagnostics and Real time healthcare monitoring

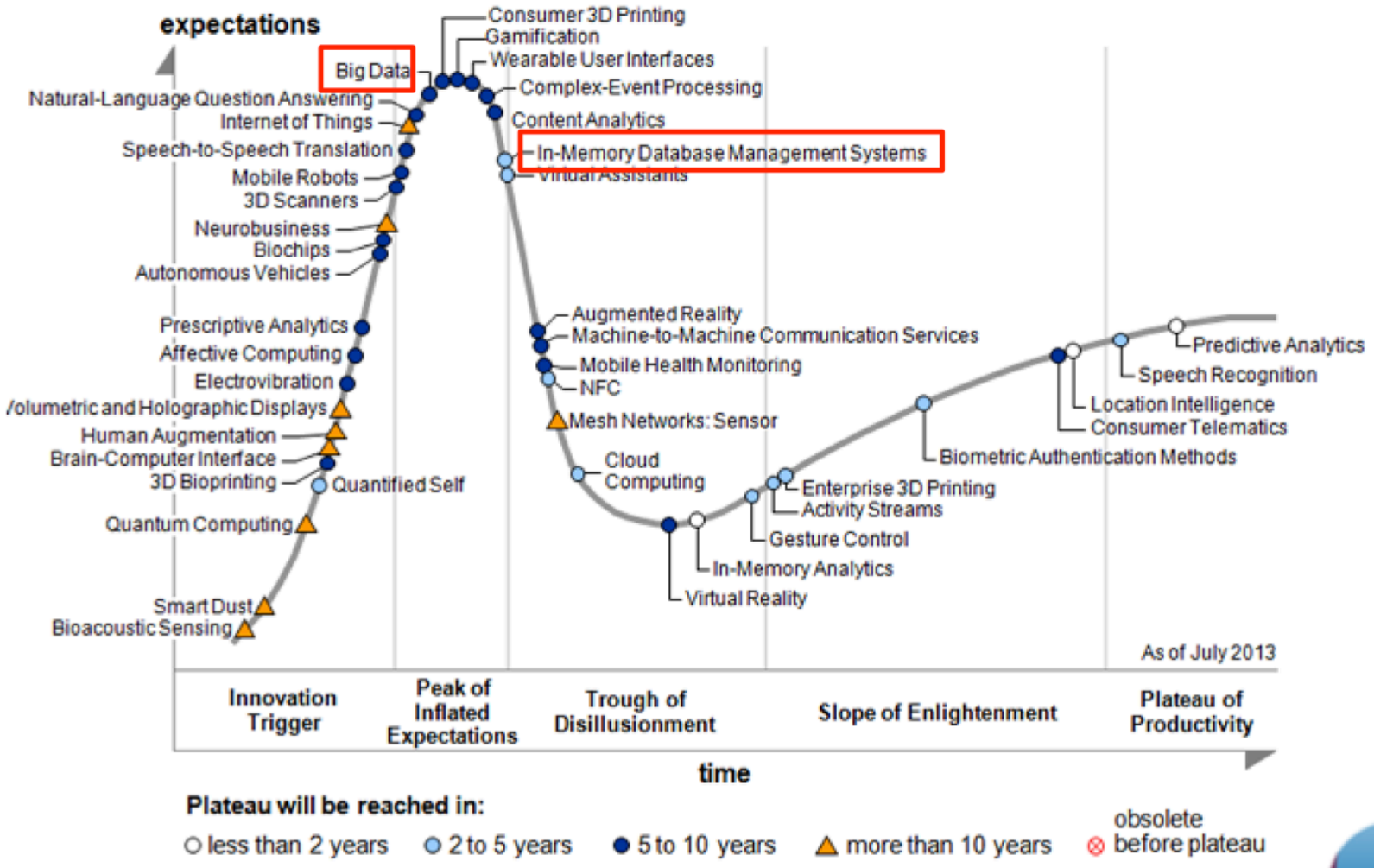


Add 4th 'V' Veracity

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

Big Data – Past Innovation on way to mainstream adoption

NoSQL: Gartner Hype Cycle 2013



Big Data : Value Drivers

REVENUE++

- Pricing and Discounts
- Channel effectiveness
- New Product Development
- E-commerce Analytics
- Advertising & promotions
- Cross-sell Up-sell
- Customer behavior and Loyalty
- Customer Service

COST--

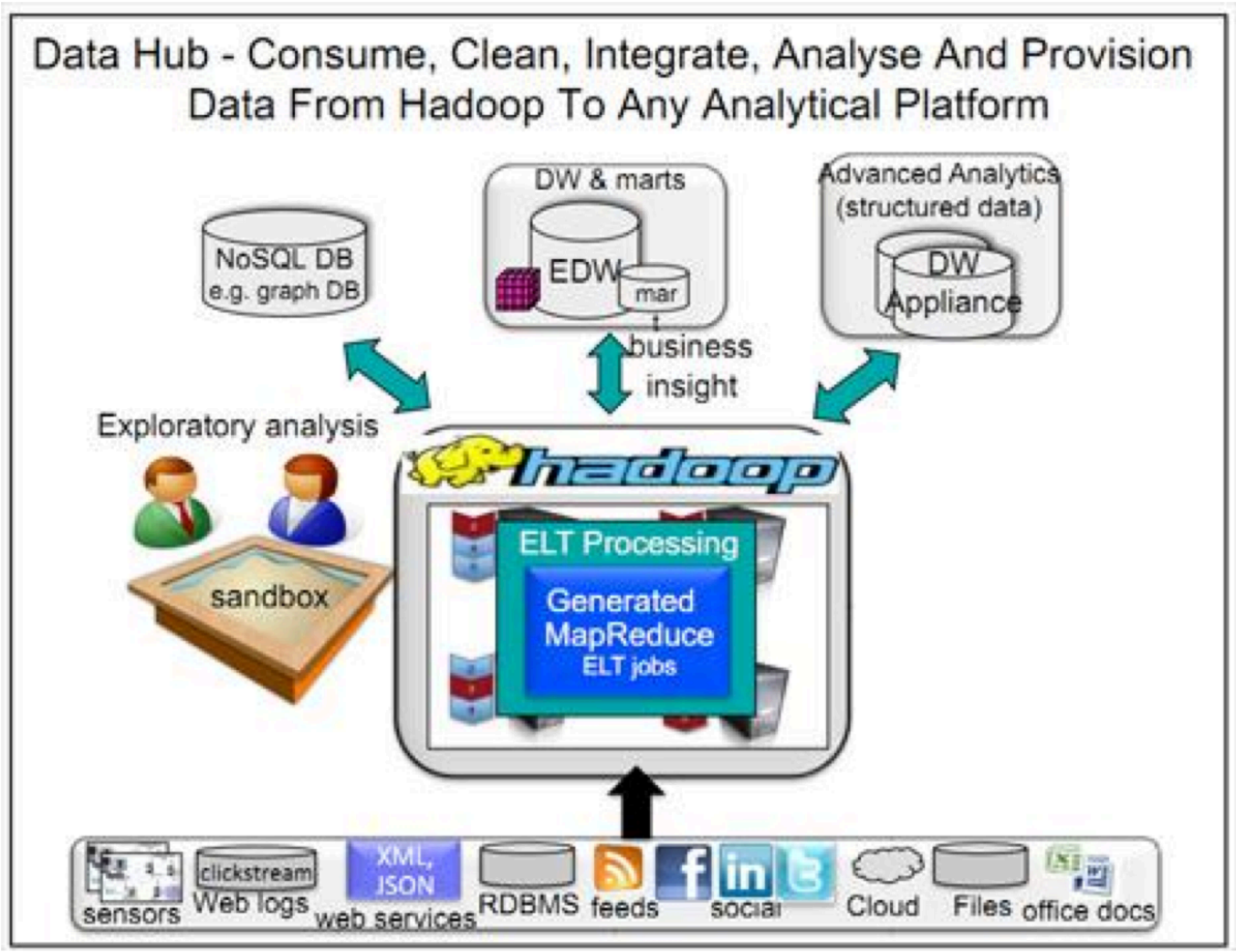
- IT Infra Mgmt
- Operations Analytics
- Supply Chain
- HR & Employee productivity
- Asset Management and utilization
- Claims
- Warranty

RISK-

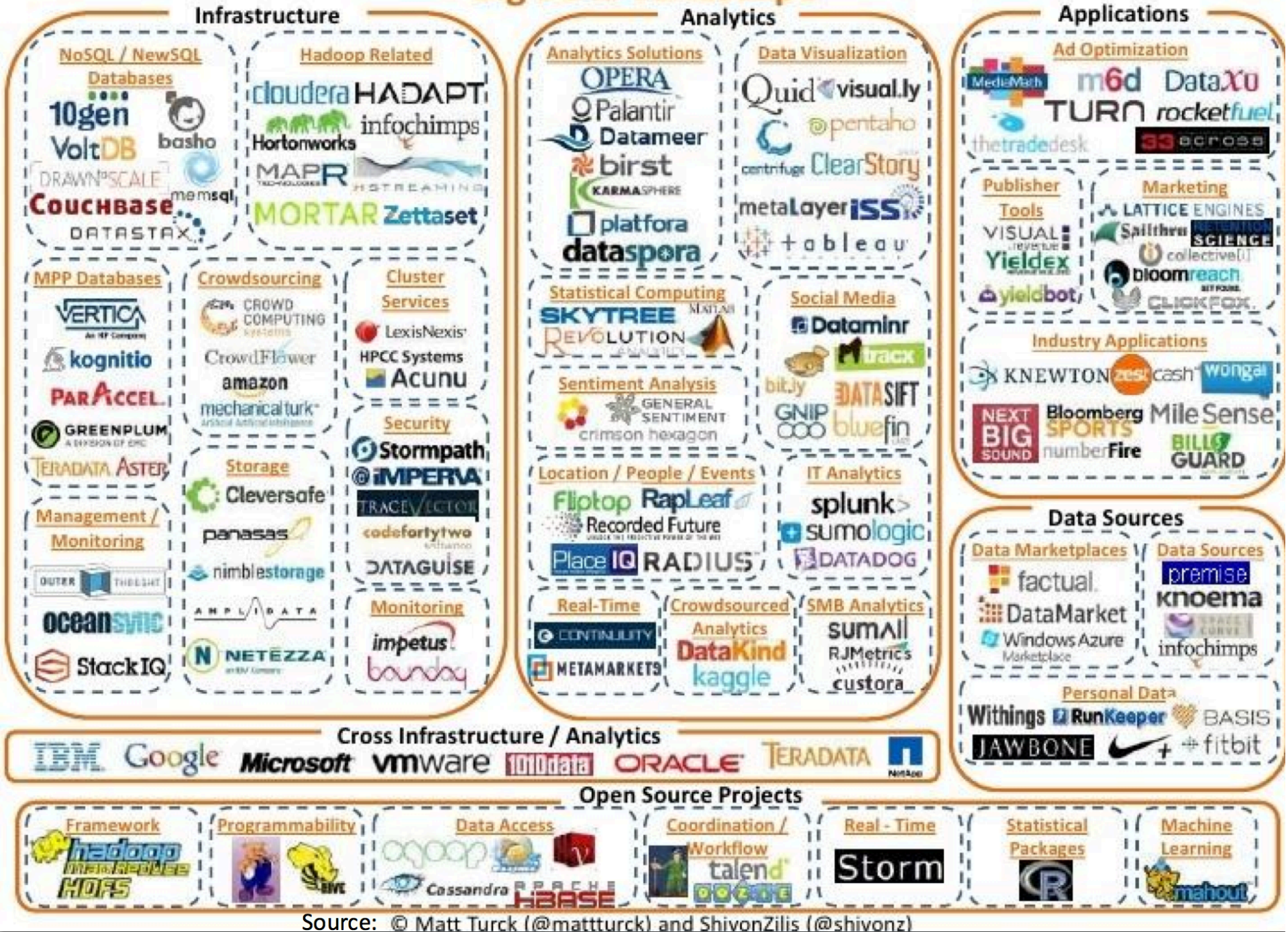
- Customer attrition and Churn
- Fraud
- RISK
- Security
- Revenue leakage
- Asset protection
- Compliance Management

http://www.EMC.com

Modern Data Management Hub



Big Data Landscape



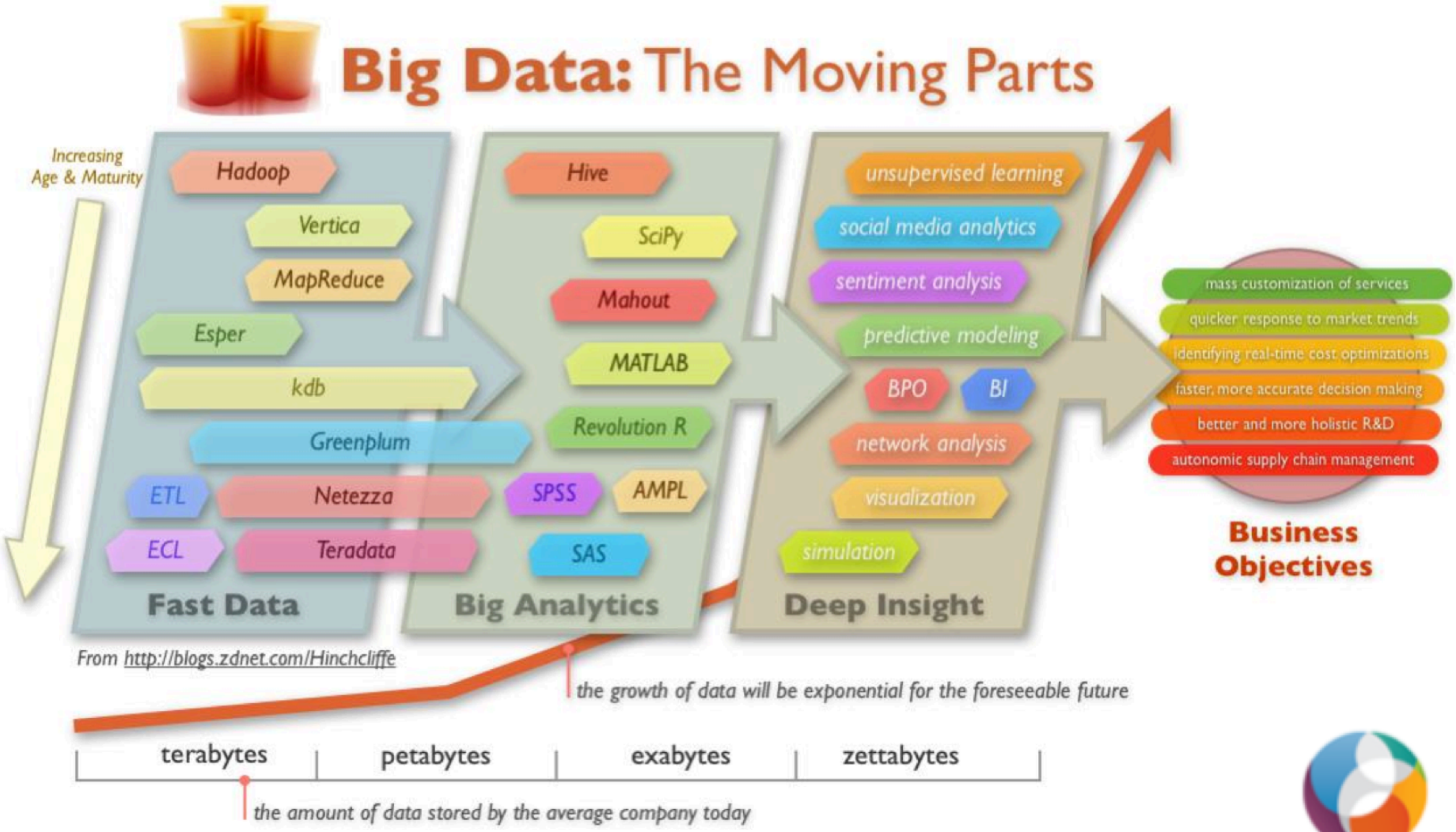
Source: © Matt Turck (@mattturck) and ShivonZilis (@shivonz)

AH.. Simplicity… This looks pretty straight-forward… (but a bit incomplete)

Big Data Market Segments					
Hardware					
Servers (Chips)		**Storage**		**Networking**	
HP		EMC		Cisco	
Dell		NetApp		Arista Networks	
Intel		Fusion-io		Infineta Systems	
Software					
Hadoop	**NoSQL**	**NGDW**	**Analytics & BI**	**Applications**	**Tools**
Hortonworks	DataStax	HP Vertica	Digital Reasoning	Google	Informatica
Cloudera	Sqrrl	EMC Greenplum	Revolution Analytics	Tresata	Talend
MapR	Couchbase	Teradata Aster	Jaspersoft	Opera Solutions	Zettaset
Hadapt	Basho	IBM Netezza	Datameer	SAP	Syncsort
EMC Greenplum	10gen	SAP	Pentaho	DataXu	Vmware
Services					
Cloud Servies		**Technical Services**		**Professional Services**	
Amazon		Hortonworks		Think Big Analytics	
Google		Cloudera		IBM	
MapR		Cloudwick		EMC	
IBM		EMC		Accenture	
Microsoft		IBM		Deloitte	

Source: http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond

Big Data Technology Vendors

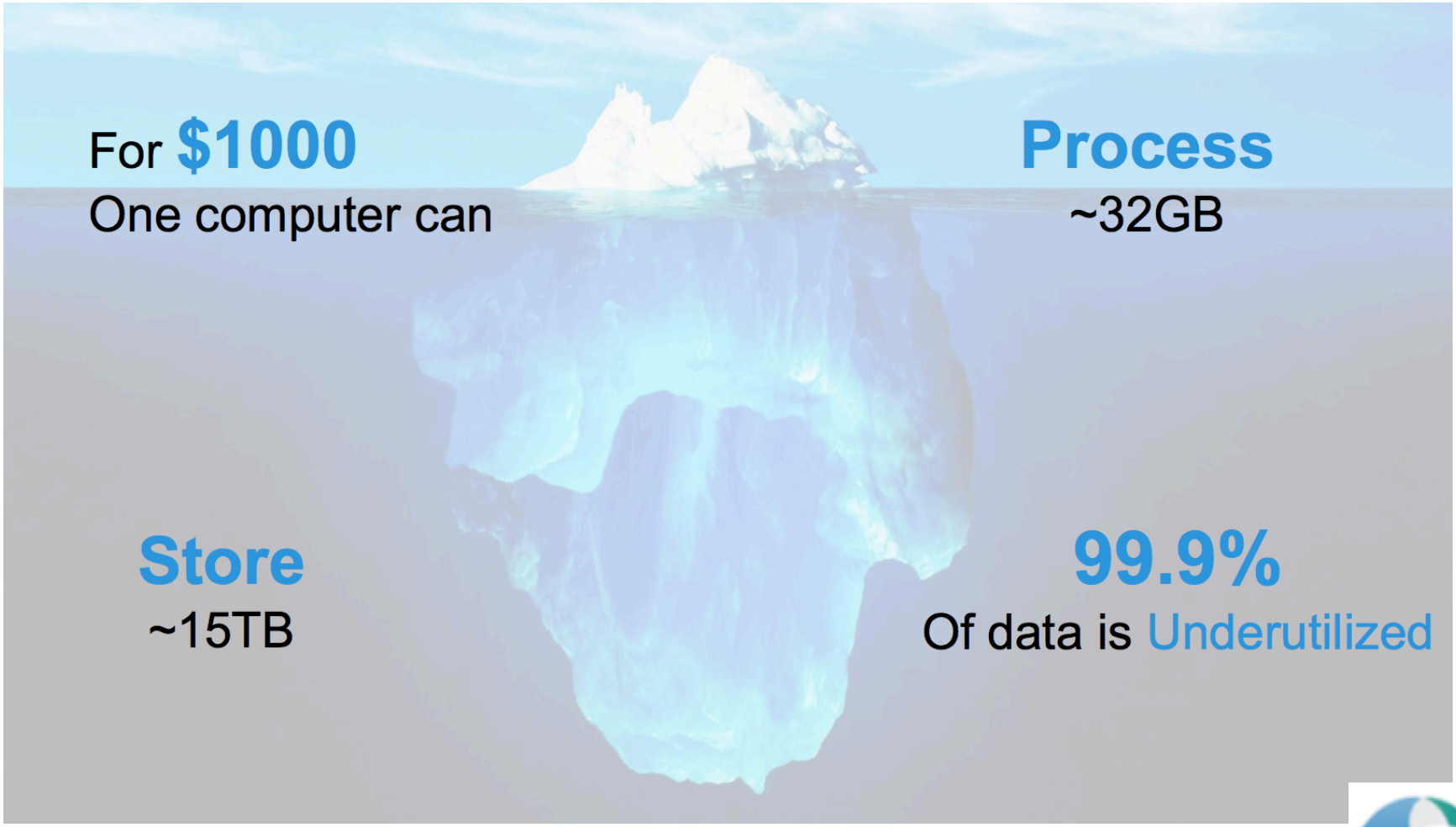


Storage & Memory B/W lagging CPU

	CPU	DRAM	LAN	Disk
Annual bandwidth improvement (all milestones)	1.5	1.27	1.39	1.28
Annual latency improvement (all milestones)	1.17	1.07	1.12	1.11

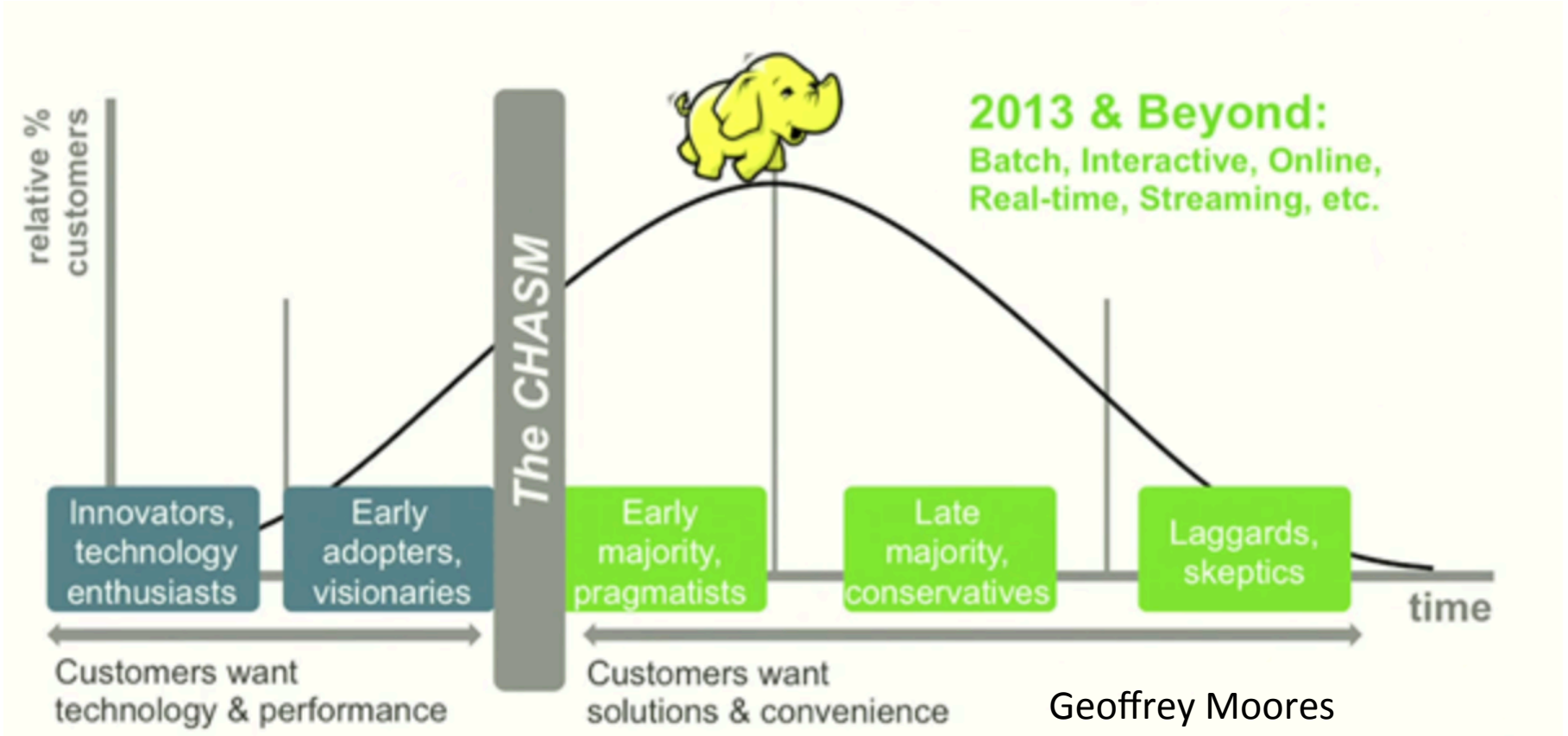
- CPU B/W requirements out-pacing memory and storage
- Disk & memory getting “further” away from CPU
- Large sequential transfers better for both memory & disk

Commodity Hardware Economics



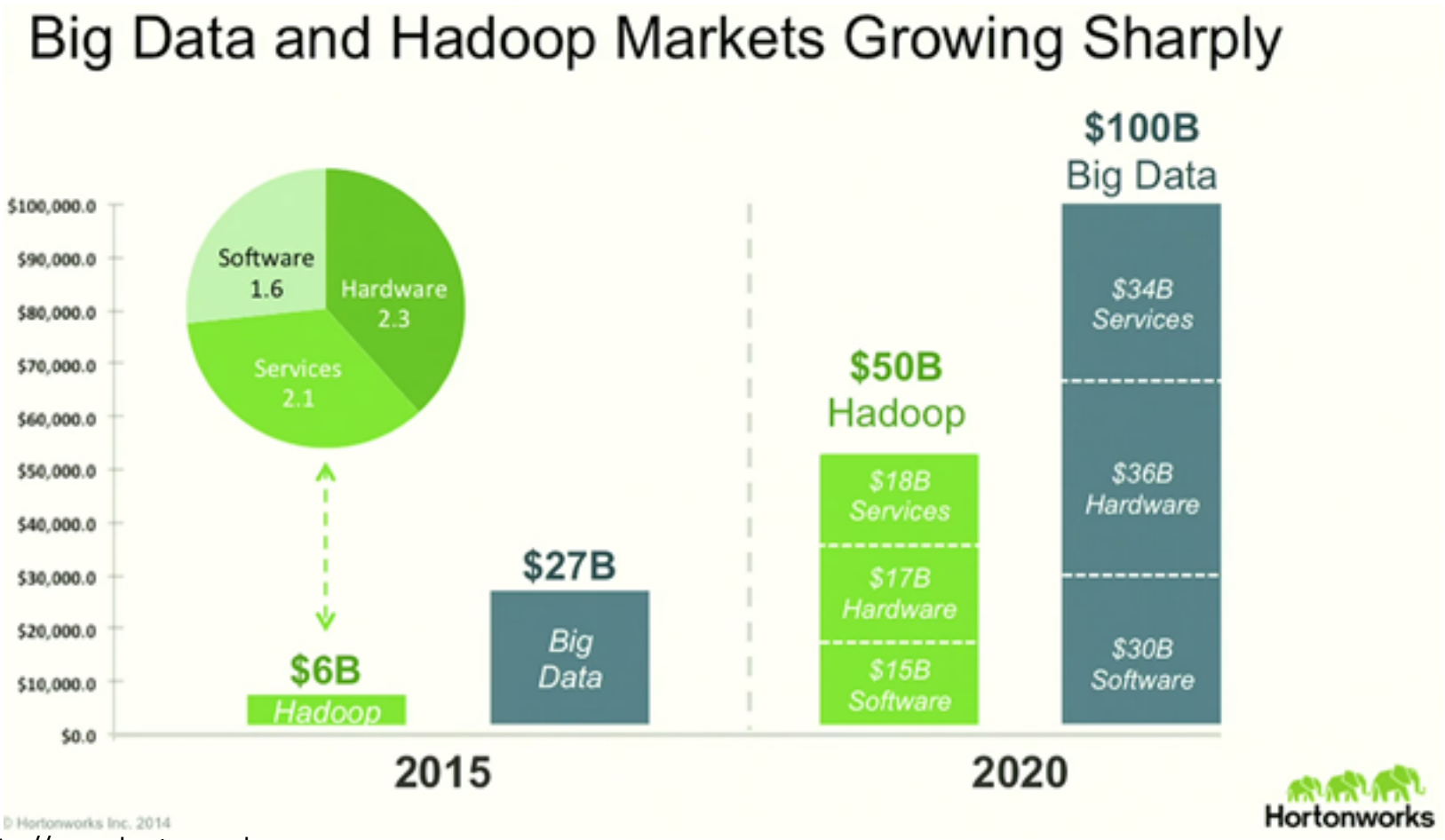
http://www.hortonworks.com

Crossing the Chasm in Big Data



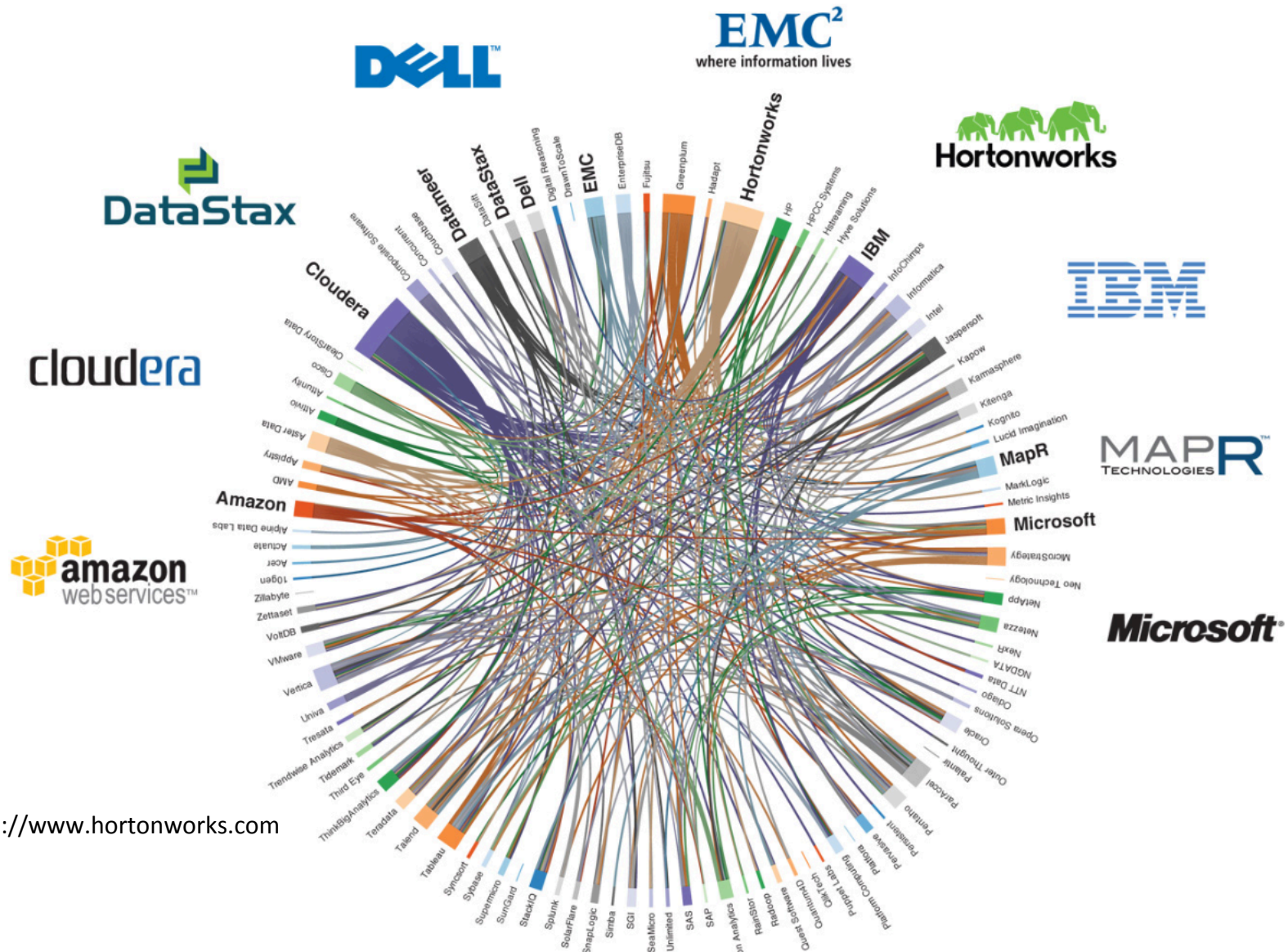
http://www.hortonworks.com

Big Data Market Growth



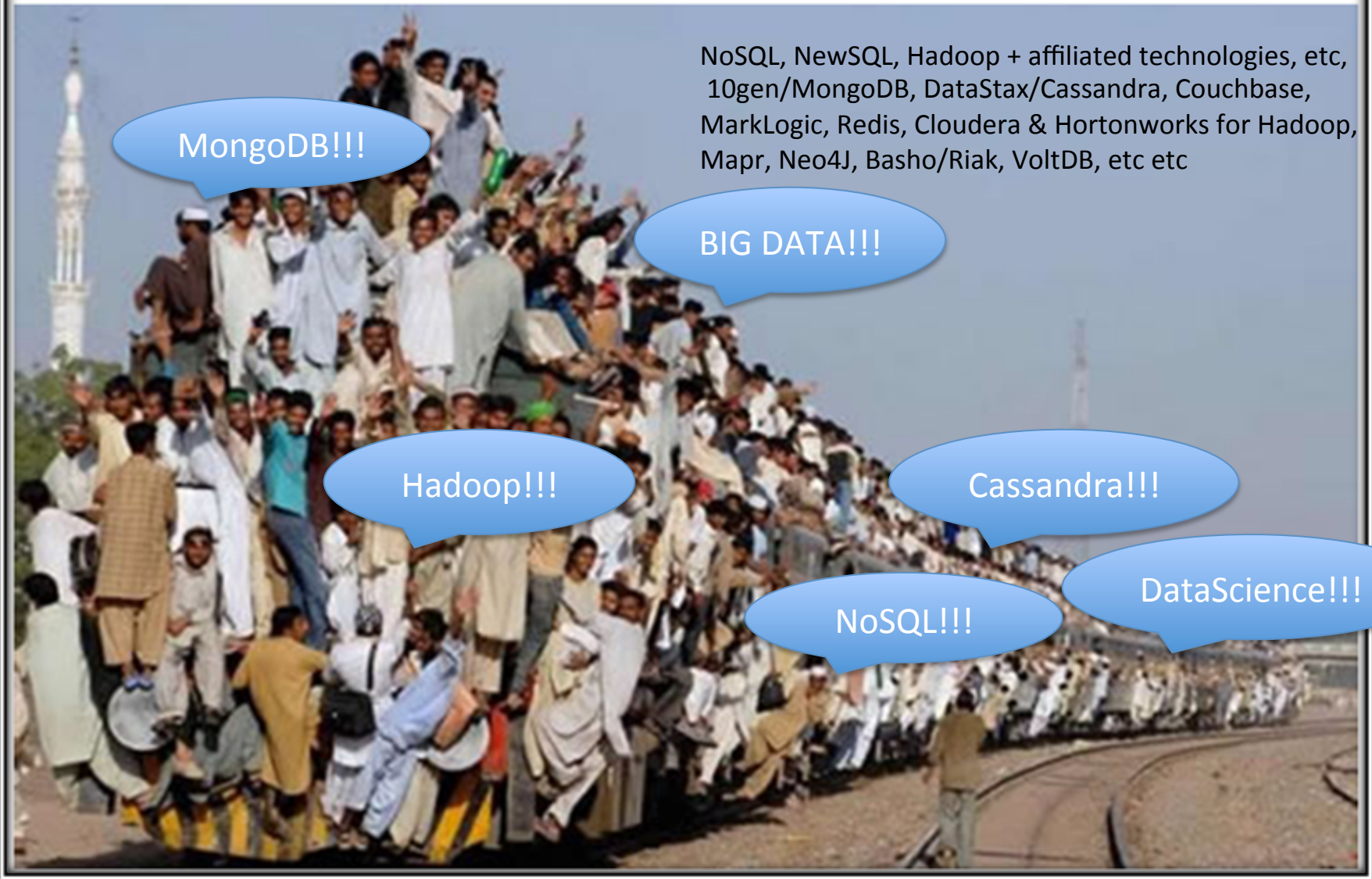
http://www.hortonworks.com

Big Data Vendors



http://www.hortonworks.com

All Aboard The Big Data Bandwagon!



NoSQL, NewSQL, Hadoop + affiliated technologies, etc,
10gen/MongoDB, DataStax/Cassandra, Couchbase,
MarkLogic, Redis, Cloudera & Hortonworks for Hadoop,
Mapr, Neo4J, Basho/Riak, VoltDB, etc etc