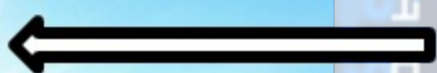
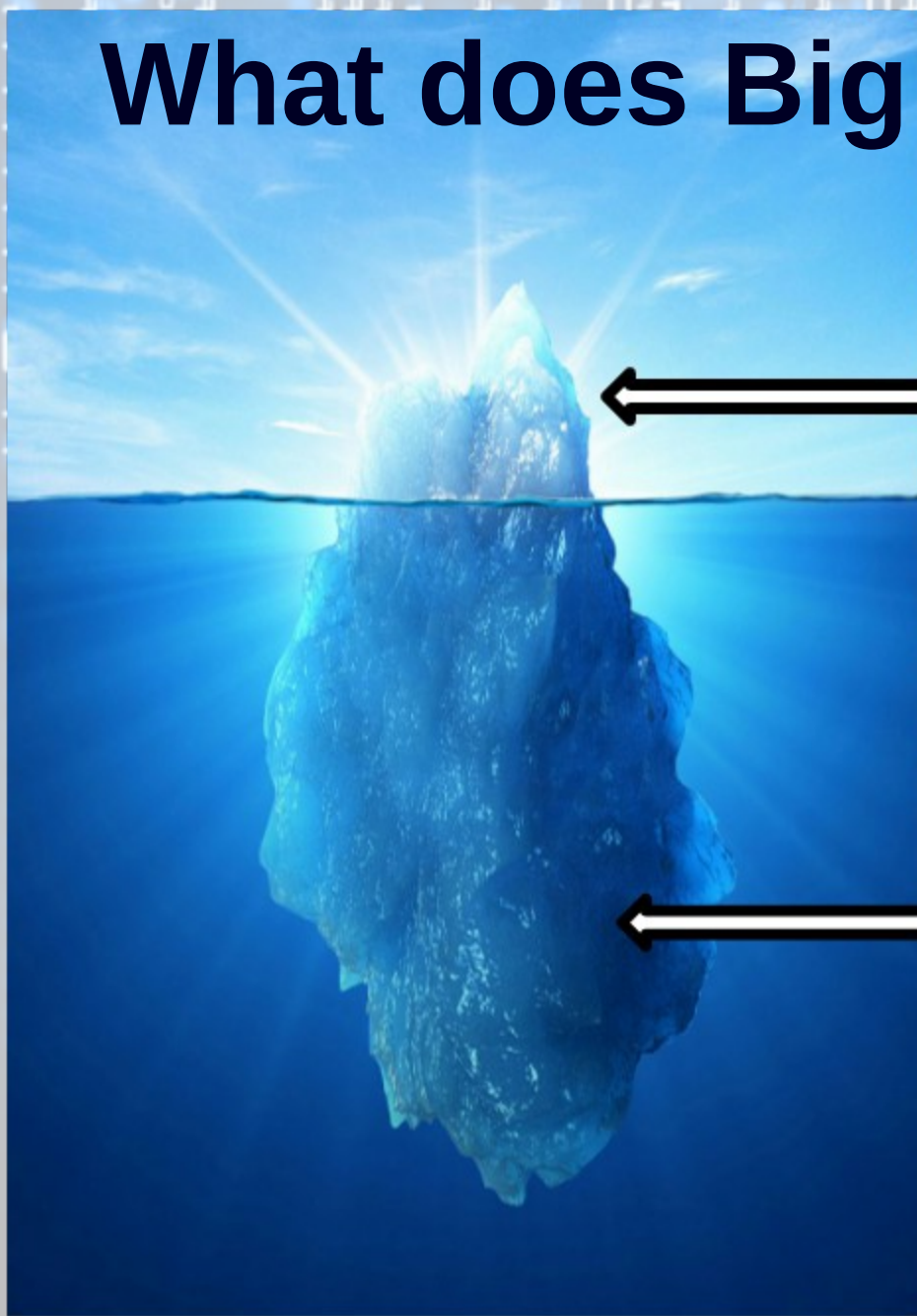


Let's talk about

BIG DATA

~ Neha Mehta

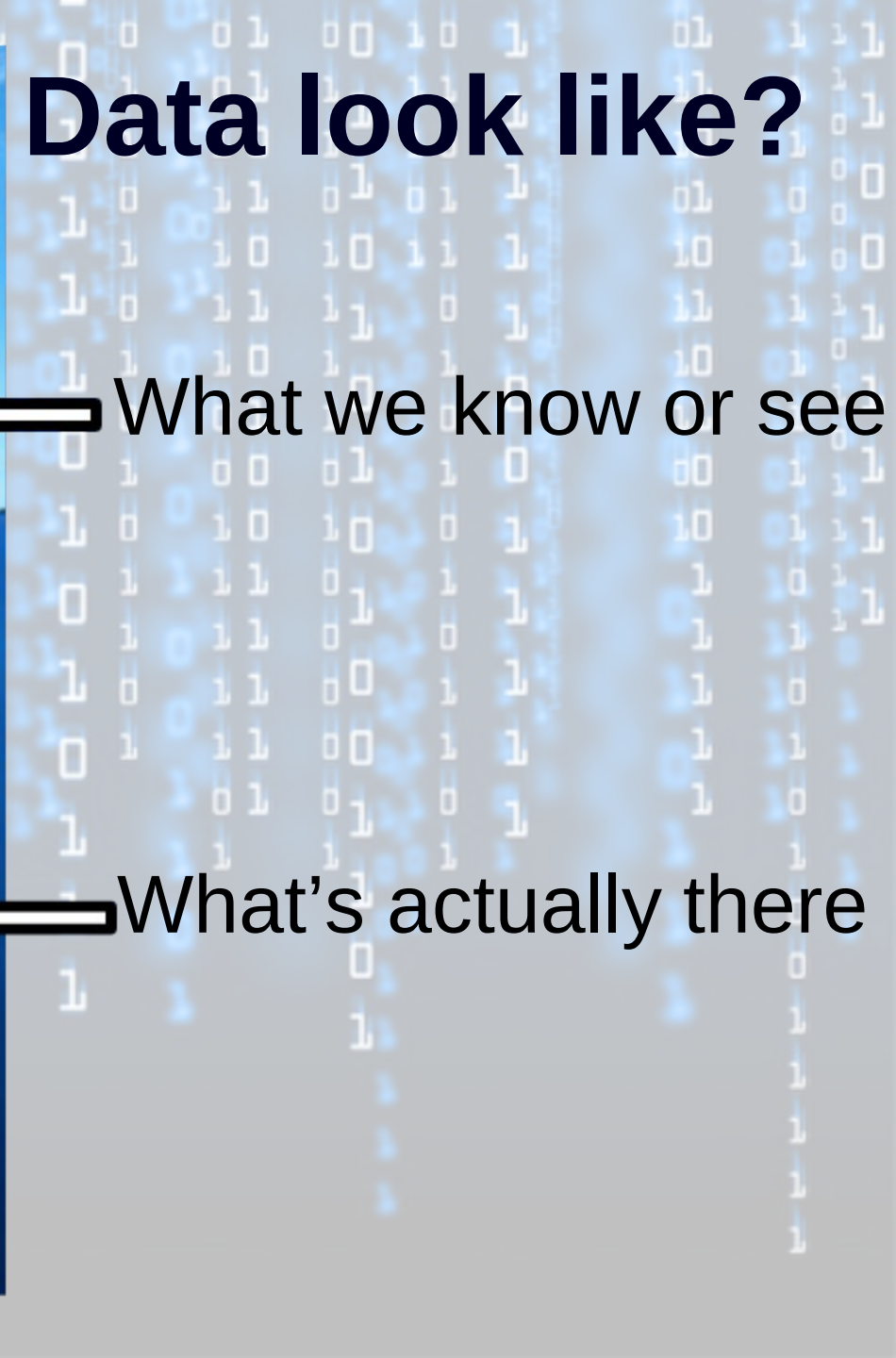
What does Big Data look like?



What we know or see



What's actually there



WHAT IS

ESIA

ENVIRONMENTAL
STATEMENT

- **Wikipedia** : In information technology, big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, curation, storage, search, sharing, analysis, and visualization.
- **IDC** : Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery, and/or analysis.
- **IBM** says that "three characteristics define big data:"
 - Volume (Terabytes -> Zettabytes)
 - Variety (Structured -> Semi-structured -> Unstructured)
 - Velocity (Batch -> Streaming Data)

How BIG is Big Data?



WHAT DOES THE FUTURE LOOK LIKE?

Worldwide IP traffic will **quadruple by 2015.**



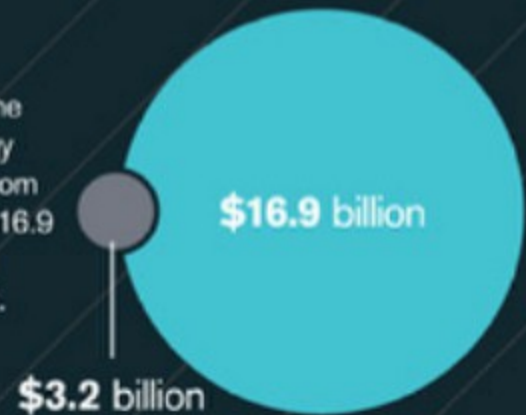
By 2015, nearly **3 billion people**



will be online, pushing the data created and shared to nearly **8 zettabytes.**

HOW IS THE MARKET FOR BIG DATA SOLUTIONS EVOLVING?

A new IDC study says the market for big technology and services will grow from \$3.2 billion in 2010 to \$16.9 billion in 2015. **That's a growth of 40% CAGR.**



1 Petabyte =

1000000000000000
BYTES



1 NEW DEFINITION IS ADDED ON UPDIAT

1,600+ READS ON Scribd

13,000+ HOURS MUSIC STREAMING ON PANDORA

12,000+ NEW ADS POSTED ON craigslist

370,000+ MINUTES VOICE CALLS ON skype

98,000+ TWEETS

20,000+ NEW POSTS ON tumblr

13,000+ iPhone APPLICATIONS DOWNLOADED

320+ NEW twitter ACCOUNTS

100+ NEW Linked in ACCOUNTS

1 associated content NEW ARTICLE IS PUBLISHED

6,600+ NEW PICTURES ARE UPLOADED ON flickr

50+ WORDPRESS DOWNLOADS

695,000+ facebook STATUS UPDATES

1,700+ Firefox DOWNLOADS

125+ PLUGIN DOWNLOADS

79,364 WALL POSTS

510,040 COMMENTS

QUESTIONS ASKED ON THE INTERNET...

100+ Answers.com 40+ YAHOO! ANSWERS

600+ NEW VIDEOS

70+ DOMAINS REGISTERED

60+ NEW BLOGS

168 MILLION EMAILS ARE SENT

694,445 SEARCH QUERIES




25+ HOURS TOTAL DURATION

1,500+ BLOG POSTS



Bliat



SHARE ON:  FACEBOOK
 TWITTER
 YOUTUBE

TAMING BIG DATA

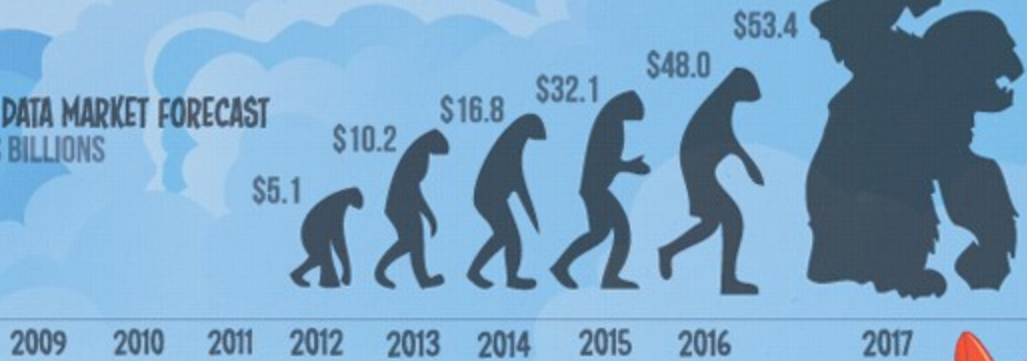
BIG DATA INCLUDES DATA SETS WHOSE SIZE AND TYPE MAKE THEM IMPRACTICAL TO PROCESS AND ANALYZE WITH TRADITIONAL DATABASE TECHNOLOGIES



PRESENTED BY: Wikibon



BIG DATA MARKET FORECAST \$ US BILLIONS



GLOBAL MENTIONS OF "BIG DATA" GOOGLE TRENDS

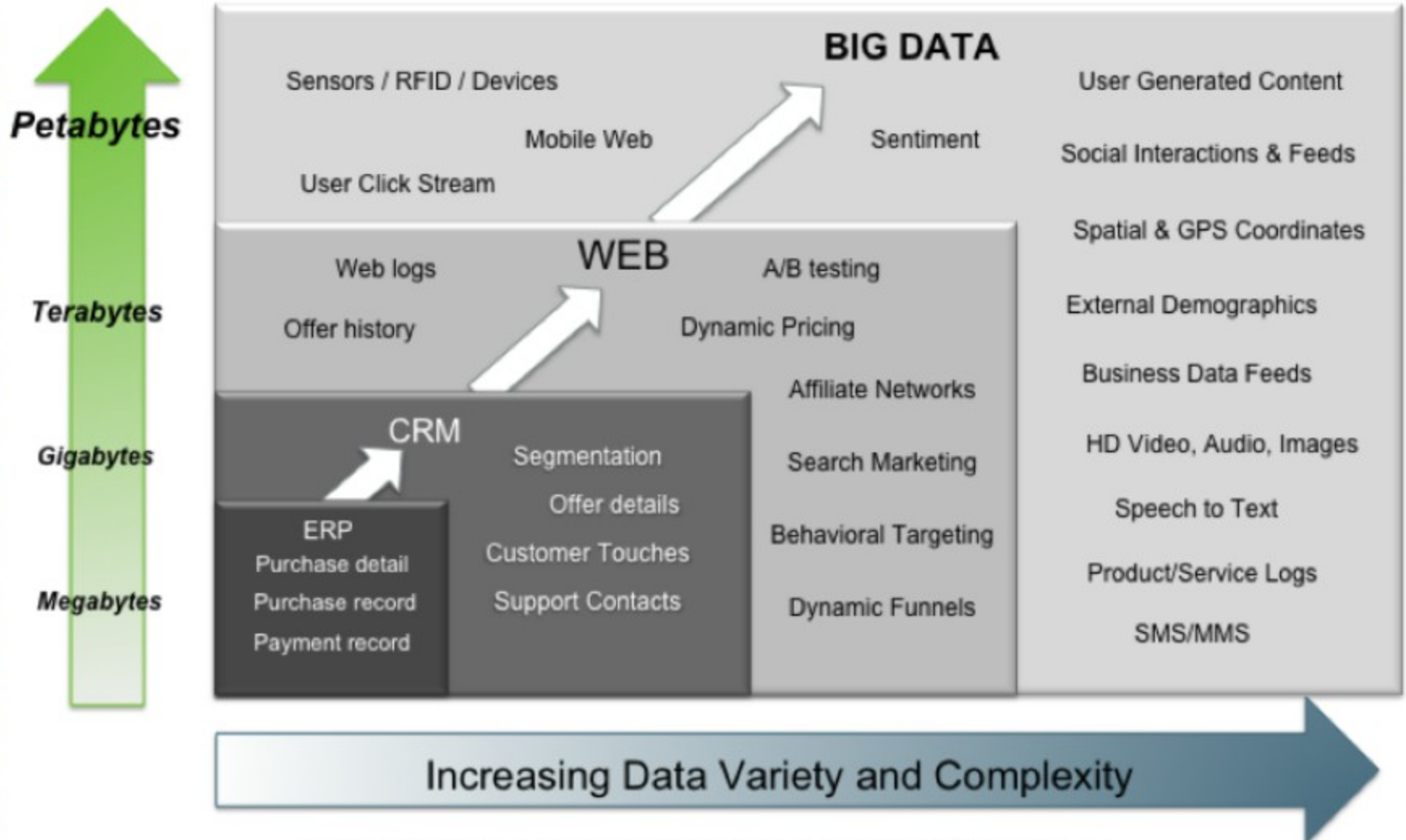


1211.34% INCREASE
OVER BASELINE AVERAGE

"IT'S NO LONGER HARD TO FIND THE ANSWER TO A GIVEN QUESTION; THE HARD PART IS FINDING THE RIGHT QUESTION AND AS QUESTIONS EVOLVE, WE GAIN BETTER INSIGHT INTO OUR ECOSYSTEM AND OUR BUSINESS." - KEVIN WEIL



Big Data = Transactions + Interactions + Observations



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

The "three V's", i.e the Volume, Variety and Velocity of the data coming in is what creates the challenge.

VOLUME



Amount of Big Data stored across the world (in petabytes)

VARIETY



PEOPLE TO PEOPLE

NETIZENS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...



PEOPLE TO MACHINE

ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...



MACHINE TO MACHINE

SENSORS, GPS DEVICES, BAR CODE SCANNERS, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...



2.9 MILLION

EMAILS SENT EVERY SECOND



20 HOURS

OF VIDEO UPLOADED EVERY MIN

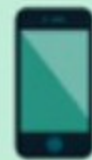


50 MILLION

TWEETS PER DAY

Volume

WHAT MAKES BIG DATA SO BIG?



6 BILLION
mobile subscriptions worldwide



87%
of the world's population



1.01 BILLION
Facebook users worldwide



604 MILLION
users log-in monthly from mobile devices



400 MILLION
Tweets per day



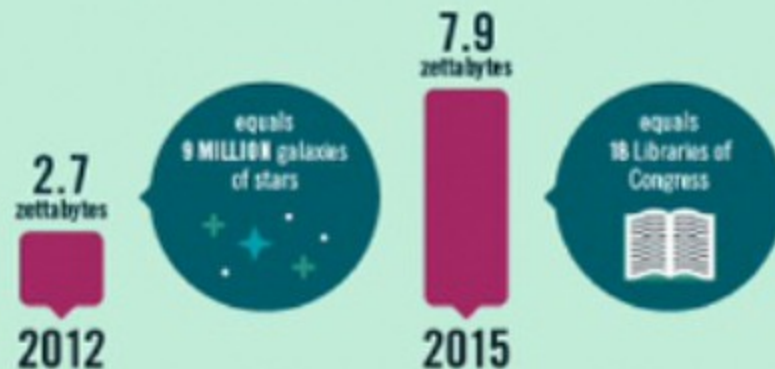
84 MILLION
users access Twitter via mobile

And Big Data will only get bigger as traffic from smartphones and tablets outpaces traditional devices.

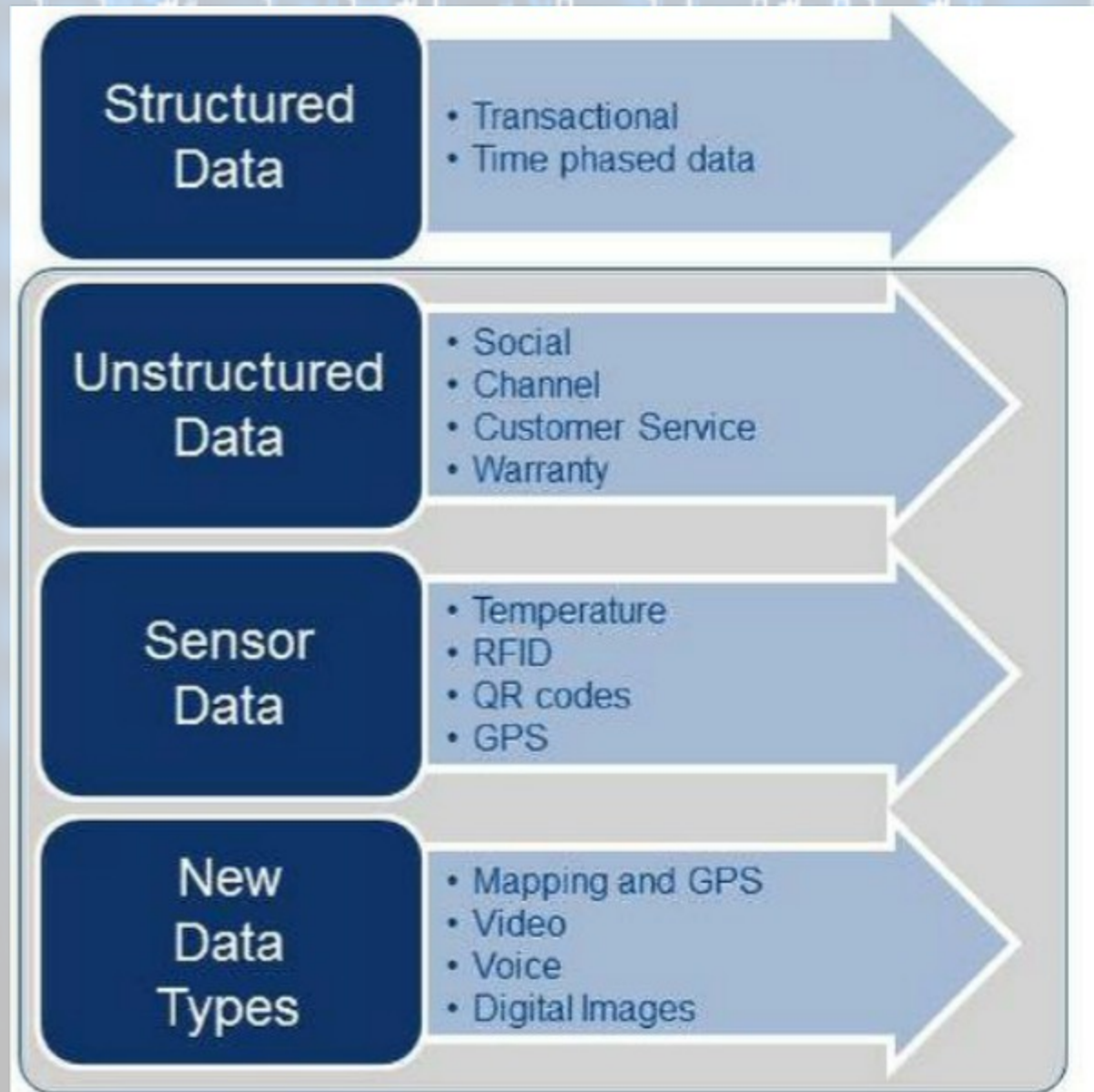
Percentage of Web Traffic by 2016:



Volume of Digital Content:



Variety



Velocity

Time Spend by Average Social networking user per month



Pinterest

405 mins



twitter

89 mins



LinkedIn

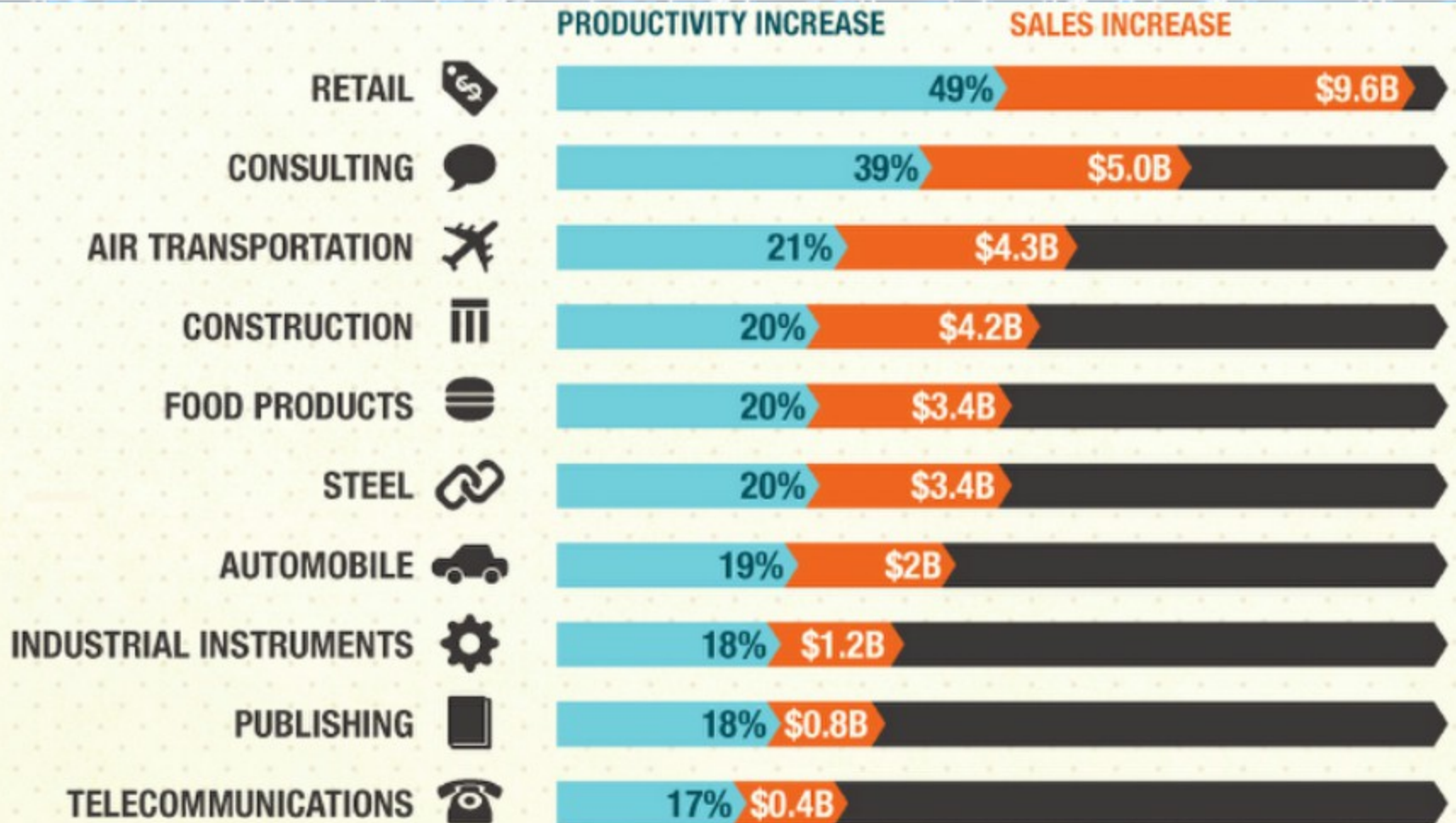
21 mins



Google+

3 mins

Value : *Follow the money!*



Going in: How Supply Chain is affected

Big data levers can deliver value along the manufacturing value chain in terms of cost, revenue, and working capital

	Lever examples	Impact			Subsector applicability
		Cost	Revenue	Working capital	
R&D and design	<ul style="list-style-type: none"> Concurrent engineering/PLM¹ Design-to-value Crowd sourcing 	+20–50% PD ² costs +30% gross margin -25% PD ² costs	-20-50% time to market		High – Low complexity High – Low complexity B2C – B2B
Supply chain management	<ul style="list-style-type: none"> Demand forecasting/shaping and supply planning 	+2–3% profit margin		-3–7% onetime	FMCG ³ – Capital goods
Production	<ul style="list-style-type: none"> Sensor data-driven operations analytics "Digital Factory" for lean manufacturing 	-10–25% operating costs -10–50% assembly costs	Up to +7% revenue +2% revenue		Capital intense – CPG ³ Capital intense – CPG ³
After-sales services	<ul style="list-style-type: none"> Product sensor data analysis for after-sales service 	-10–40% maintenance costs	+10% annual production		Capital intense – CPG ³

1 Product lifecycle management.

2 Product development.

3 Fast-moving consumer goods and consumer packaged goods.

SOURCE: Expert interviews; press and literature search; McKinsey Global Institute analysis

THE FLOOD OF BIG DATA

US\$2.1

BILLION SPENT IN
U.S. ON MOBILE ADS²

IN 2011

MOBILE
AD

4.8

TRILLION ONLINE AD
IMPRESSIONS¹

IN 2011

ONLINE AD
IMPRESSIONS

US\$83.2

BILLION ESTIMATED³

FOR 2012

ONLINE
AD SPEND

100

TERABYTES OF DATA
UPLOADED⁴

DAILY

FACEBOOK

294

BILLION EMAILS
SENT⁵

EVERY DAY

EMAIL

230

MILLION TWEETS⁶

A DAY

TWITTER

BIG DATA = BIG OPPORTUNITY

ISN'T IT GREAT?
WE HAVE TO
PAY NOTHING
FOR THE BARN

YEAH!
AND EVEN
THE FOOD
IS FREE

FACEBOOK AND YOU

If you're not paying for it, you're not the customer. You're the product being sold.

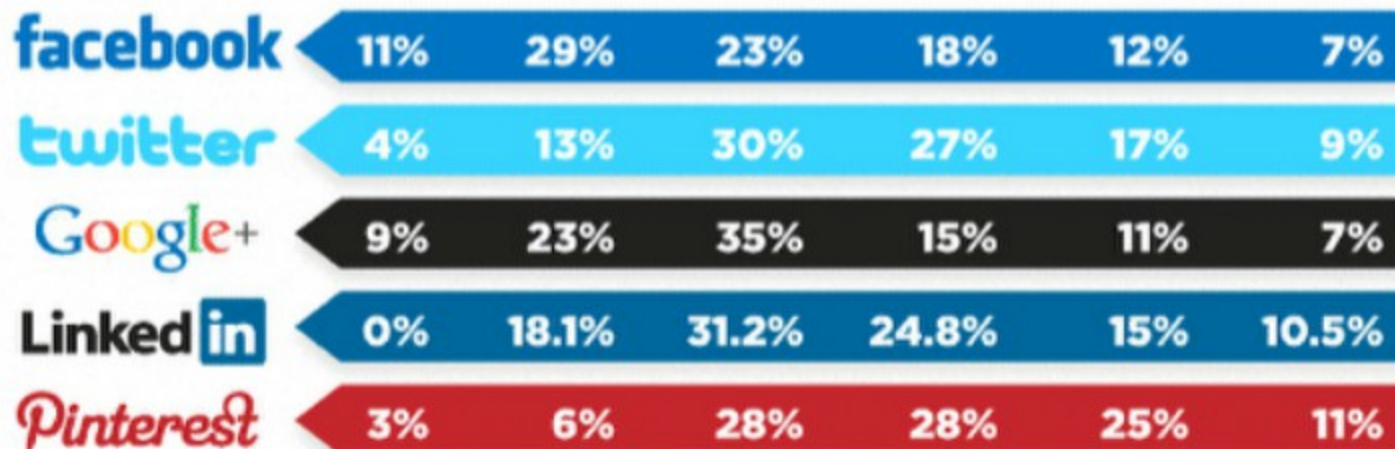
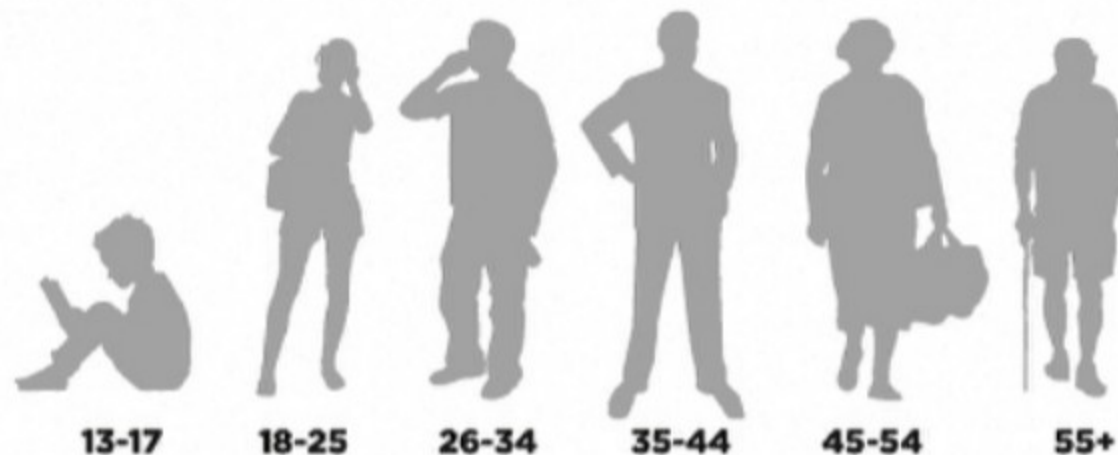
901 Million users



Monthly Visits on Top Social Networking Websites



Age Demographics of Social Networking Users



Estimated User Worth of popular Social networking sites



\$118

facebook



\$71.43

twitter



\$71

Linked in



\$28.09

Pinterest

Challenges

TOP DATA CHALLENGES FOR RETAILERS

● Volume ● Variety ● Velocity



DATA DISCOVERY

How much enterprise data does your company store?



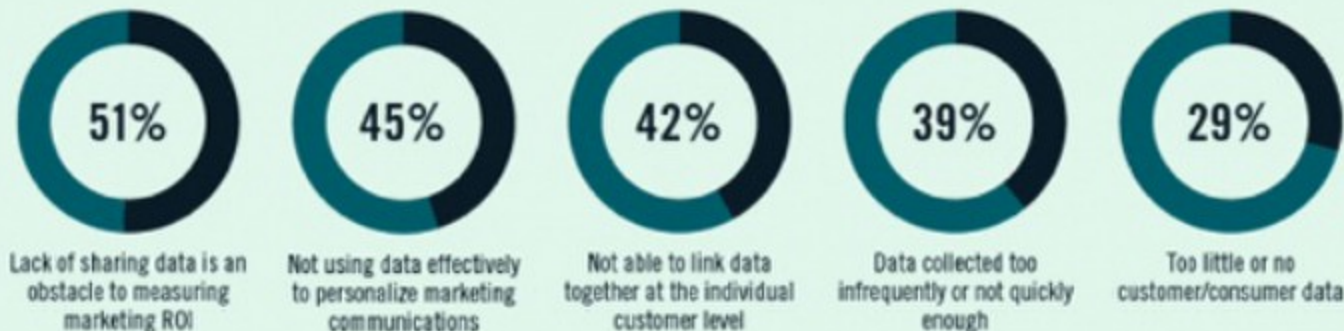
32.8% Don't know
32.8% Less than 50 terabytes
24.6% 51 to 500 terabytes
8.2% More than 1 petabyte
1.6% 500 to 999 terabytes

How much of the data is unstructured?



CHALLENGES TO USING BIG DATA

Given that nearly one-third of retailers are in the dark about their available data, it makes sense that silos are the primary hurdle in using this information.



What will they gain?

GOALS FOR USING BIG DATA

Based on where retailers are investing (or are planning to invest) their resources, they see the value in creating more sophisticated omnichannel marketing efforts.

Retailers plan to focus their Big Data initiatives on improving:



But they expect to deploy their first Big Data projects in:



The GAME PLAN for “Big” gain

HOW TO DEVELOP A BIG DATA GAME PLAN

Keeping up with today’s demanding customers and analytics-savvy competitors (it’s not just Amazon and Expedia anymore) means putting data at the heart of the retail business. Get started with this five-step plan:

1.

Determine the maturity level of your company’s approach to Big Data, then implement proof of concepts to guide your ongoing investments.

2.

Zero in on business functions for which Big Data can drive the greatest improvement, and create detailed use cases for these projects. Three key areas to investigate first are pricing, segmentation, and marketing effectiveness.

3.

Size up your data management and analytics capabilities, identifying gaps and developing the necessary recruitment and training plans.

4.

Make sure your data strategy encompasses customer data/master data management, policy and process rules, and data collection usage and sharing.

5.

Anticipate the hiccups that accompany business process change, helping teams adjust to this new way of incorporating Big Data and analytics into decision-making.

BIG DATA

At IBM, big data is about the 'the art of the possible.'

What Big data should achieve:

Volume: Turn 12 terabytes of Tweets created each day into improved product sentiment analysis

Convert 350 billion annual meter readings to better predict power consumption

Velocity: Sometimes 2 minutes is too late.

Scrutinize 5 million trade events created each day to identify potential fraud

Analyze 500 million daily call detail records in real-time to predict customer churn faster

Variety: Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more.

Monitor 100's of live video feeds from surveillance cameras to target points of interest

Exploit the 80% data growth in images, video and documents to improve customer satisfaction

Veracity: Establishing trust in big data presents a huge challenge as the variety and number of sources grows.

Decisions

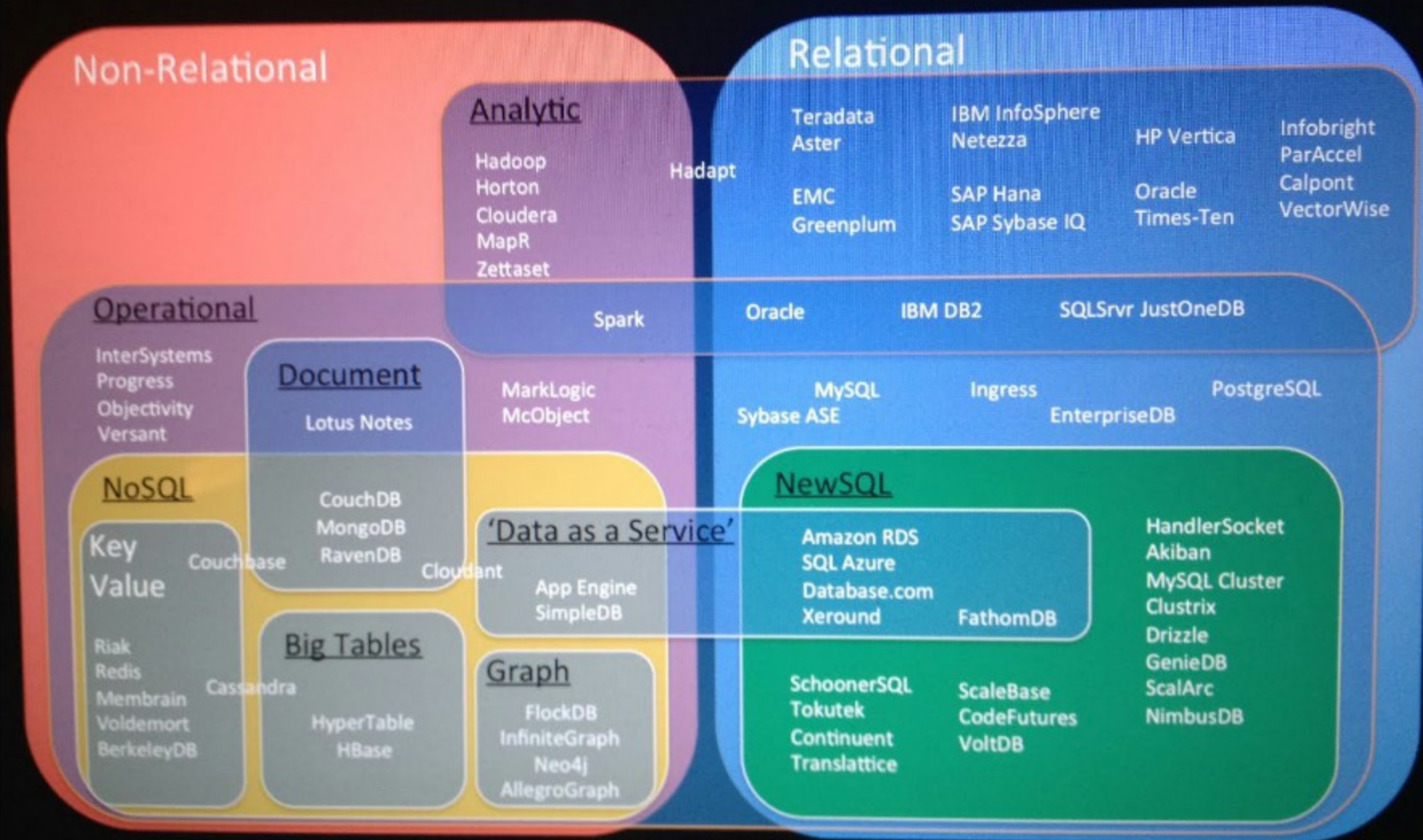
Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach.

Here are three key technologies that can help you get a handle on big data – and even more importantly, extract meaningful business value from it .

- **Information management for big data.** Manage data as a strategic, core asset, with ongoing process control for big data analytics .
- **High-performance analytics for big data.** Gain rapid insights from big data and the ability to solve increasingly complex problems using more data .
- **Flexible deployment options for big data.** Choose between options for on premises or hosted, software-as-a-service (SaaS) approaches for big data and big data analytics

Problem

One Size Does Not Fit All



The big one's that have almost made it:

- **Massively Parallel Processing (MPP)** – This involves a coordinated processing of a program by multiple processors (200 or more in number). Each of the processors makes use of its own operating system and memory and works on different parts of the program. Each part communicates via messaging interface. An MPP system is also known as “loosely coupled” or “shared nothing” system.
- **Distributed file system** or network file system allows client nodes to access files through a computer network. This way a number of users working on multiple machines will be able to share files and storage resources. The client nodes will not be able to access the block storage but can interact through a network protocol. This enables a restricted access to the file system depending on the access lists or capabilities on both servers and clients which is again dependent on the protocol.

The big one's that have almost made it (contd.):

- **Apache Hadoop** is key technology used to handle big data, its analytics and stream computing. Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It can be scaled up from a single server to thousands of machines and with a very high degree of fault tolerance. Instead of relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.
- **Data Intensive Computing** is a class of parallel computing application which uses a data parallel approach to process big data. This works based on the principle of collocation of data and programs or algorithms used to perform computation. Parallel and distributed system of inter-connected stand alone computers that work together as a single integrated computing resource is used to process / analyze big data.

- HDFS:

- **Large files** are split into parts
- Move file parts into a cluster
- Fault-tolerant through **replication** across nodes while being rack-aware
- **Bookkeeping** via NameNode

- MapReduce

- Move **algorithms** close to the data by structuring them for **parallel execution** so that each task works on a part of the data. The power of **Simplicity!**

- noSQL:

- Old technology – before RDBMS
- Sequential files
- Hierarchical DB
- Network DB
- Graph DB (Neo4j, AllegroGraph)
- Document Stores (Lotus Domino, MongoDB, JackRabbit)
- Memory Caches

- **Common Tools:**

- Infrastructure Management:

- Chef & Puppet from DevOps movement

- Grid Monitoring tools:

- Operational insight with Nagios, Ganglia, Hyperic and ZenOSS

- Development support is a need of the hour

- **Operational Tools:**

- Chef & Puppet from DevOps movement + Ganglia

- **Analytics Platforms (DW Solutions, BI solutions and analytics)**

- Netezza, GreenPlum, Vertica

- **Universal Data (All data storage needs at enterprise level)**

- Lily & Spire

Transformation of Retail

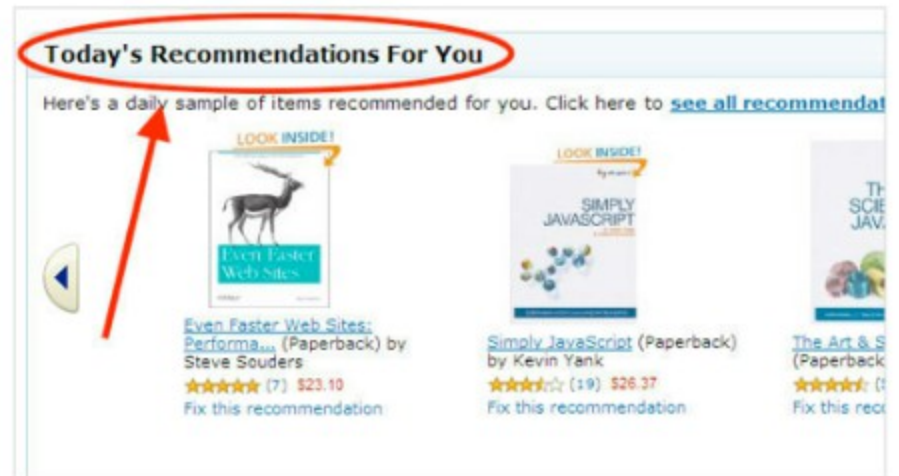
THEN...

Sales



NOW...

Data driven pricing and recommendations



Transformation of Online Marketing

THEN...

Leads

Company	First	Last	Oppty	Created
Acme	Fred	Langan	\$250K	6/08/12
BigCo	Tom	Jones	\$100K	6/17/12
DealCo	Jan	Sedor	\$50K	7/01/12
Stor Works	Liza	Grear	\$750K	7/14/12
RF Group	Carl	Tomer	\$47K	7/18/12

NOW...

Marketing and Sales
Recommendations



Transformation of Customer Service

THEN...

Unhappy
customers



NOW...

Customer insight



Transformation Of Fraud Management

THEN...

Credit databases



NOW...

Social profiles



Transformation of Law Enforcement

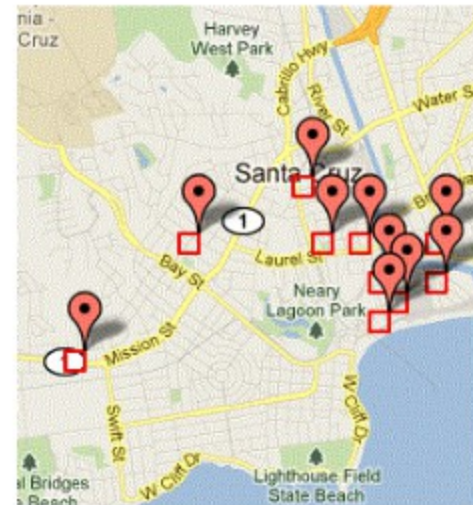
THEN...

Gut instinct



NOW...

Crime hotspot prediction



Big Data & Real Business Issues

Key questions enterprises are asking about Big Data:

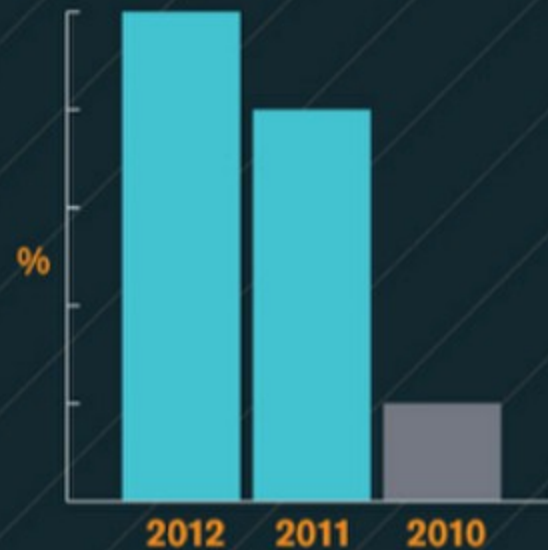
How to store and protect big data?

How to backup and restore big data?

How to organize and catalog the data that you have backed up?

How to keep costs low while ensuring that all the critical data is available when you need it?

Everyday business and consumer life creates **2.5 quintillion** bytes of data per day.



90% of the data in the world today has been created in the last two years alone.



IN THE AGE OF INFORMATION...
IGNORANCE IS A CHOICE

Source links and some more useful information...

- http://www.sas.com/resources/whitepaper/wp_46345.pdf
- <http://wikibon.org/blog/big-data-statistics/>
- <http://wikibon.org/blog/big-data-infographics/>
- <http://www.marketingtechblog.com/ibm-big-data-marketing/>
- <http://blogs.gartner.com/doug-laney/>
- <http://www-01.ibm.com/software/data/bigdata/>
- http://en.wikipedia.org/wiki/Big_data
- http://mike2.openmethodology.org/wiki/Big_Data_Definition
- <http://pinterest.com/rtkrum/cool-infographics-gallery>
- <http://www.go-gulf.com/blog/category/blog/seo-sem/>
- <http://www.forbes.com/sites/davefeinleib/2012/07/24/big-data-trends/>
- http://hortonworks.com/wp-content/uploads/2012/05/bigdata_diagram.png
- <http://visual.ly/big-data>
- http://csrc.nist.gov/groups/SMA/forum/documents/june2012presentations/fcsm_june2012_cooper_mell.pdf
- <http://www.ramcoblog.com/big-data-technologies-a-quick-way-to-get-through-more>



Happy Holidays

