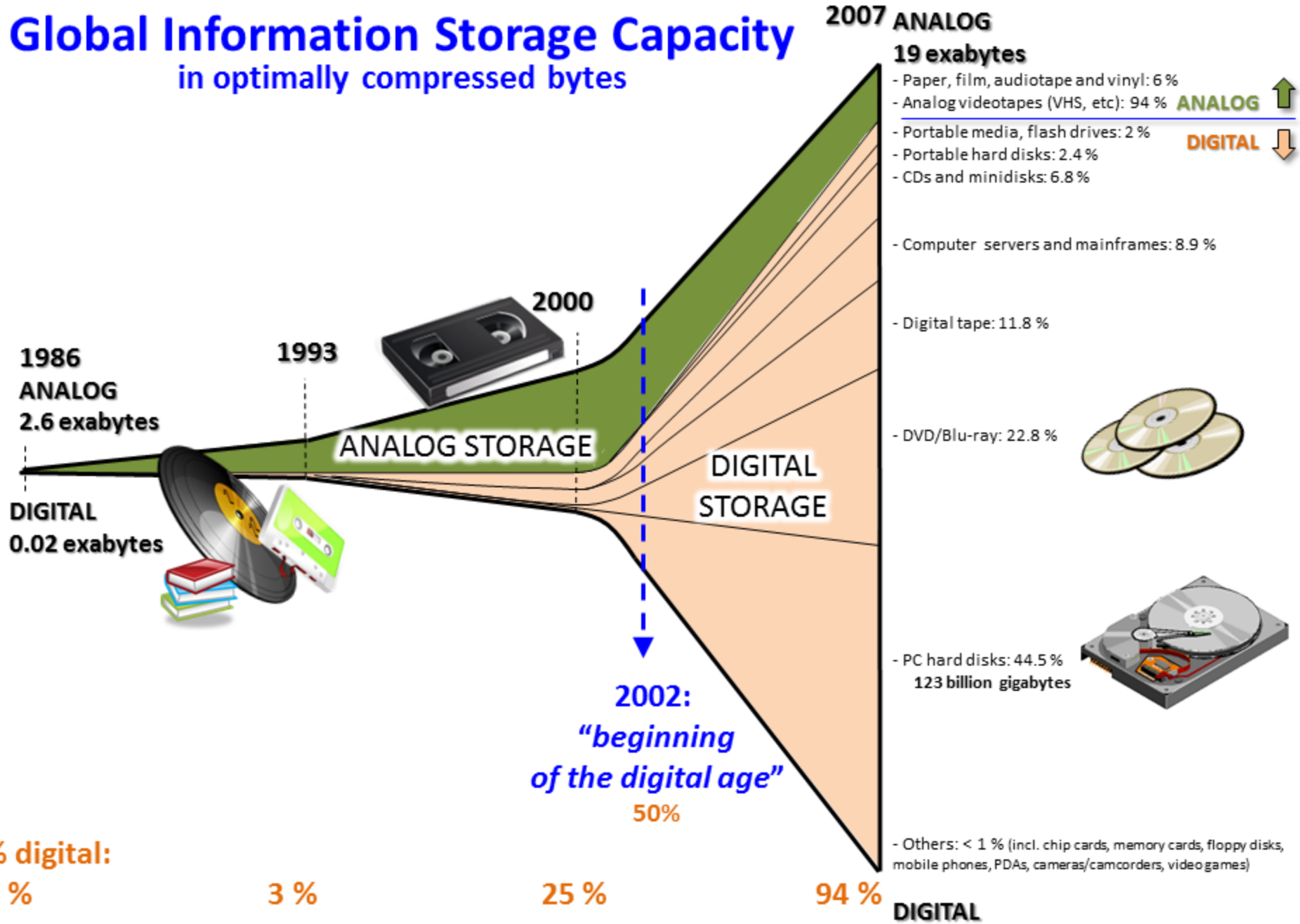
Why Big Data?

Lots of data

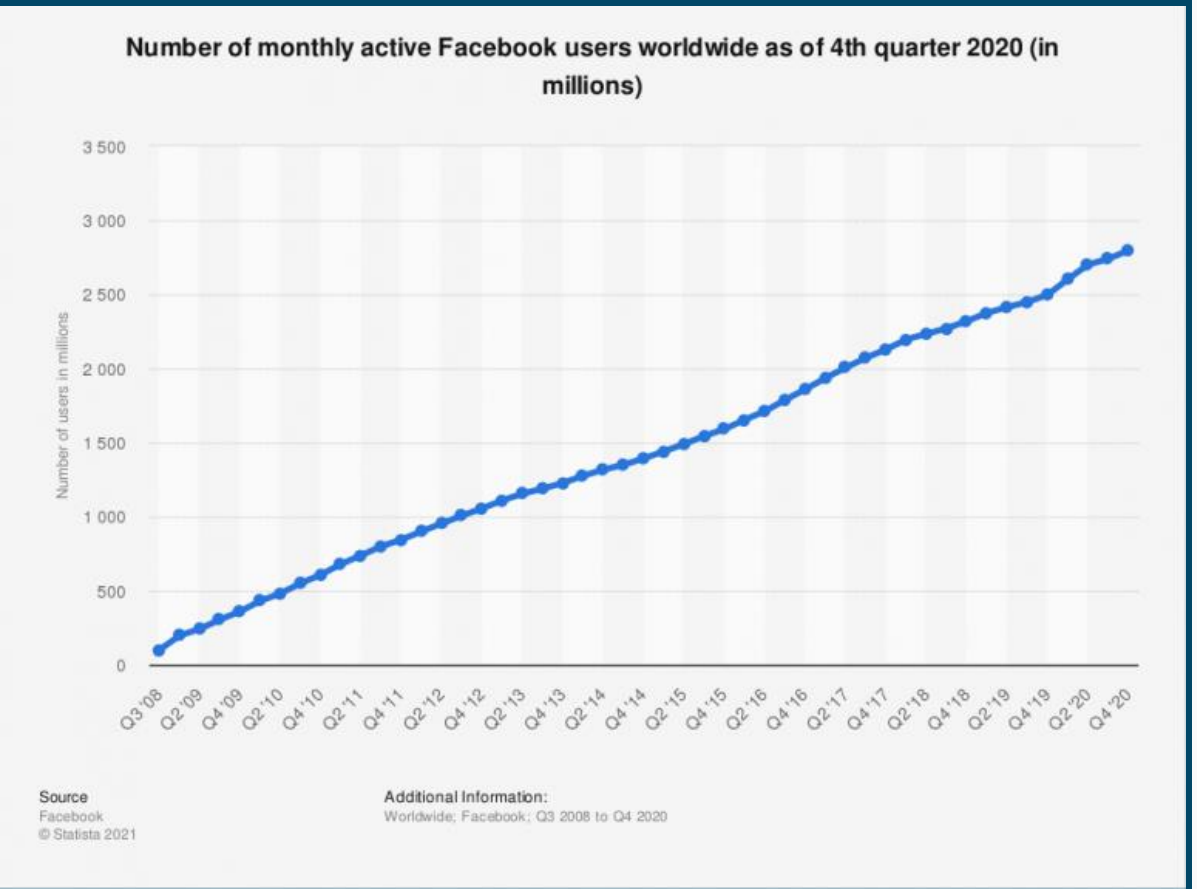
Global Information Storage Capacity

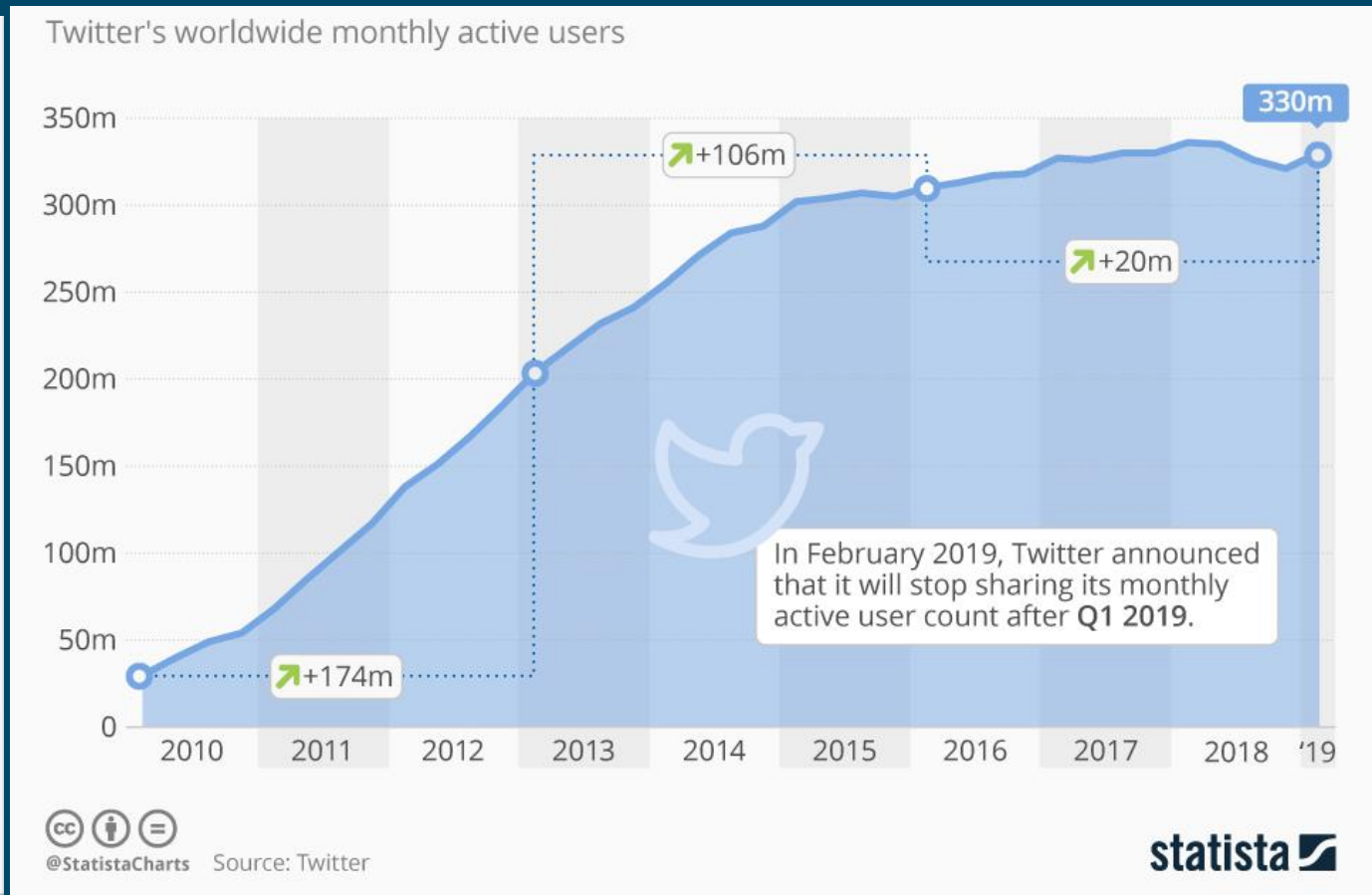
in optimally compressed bytes



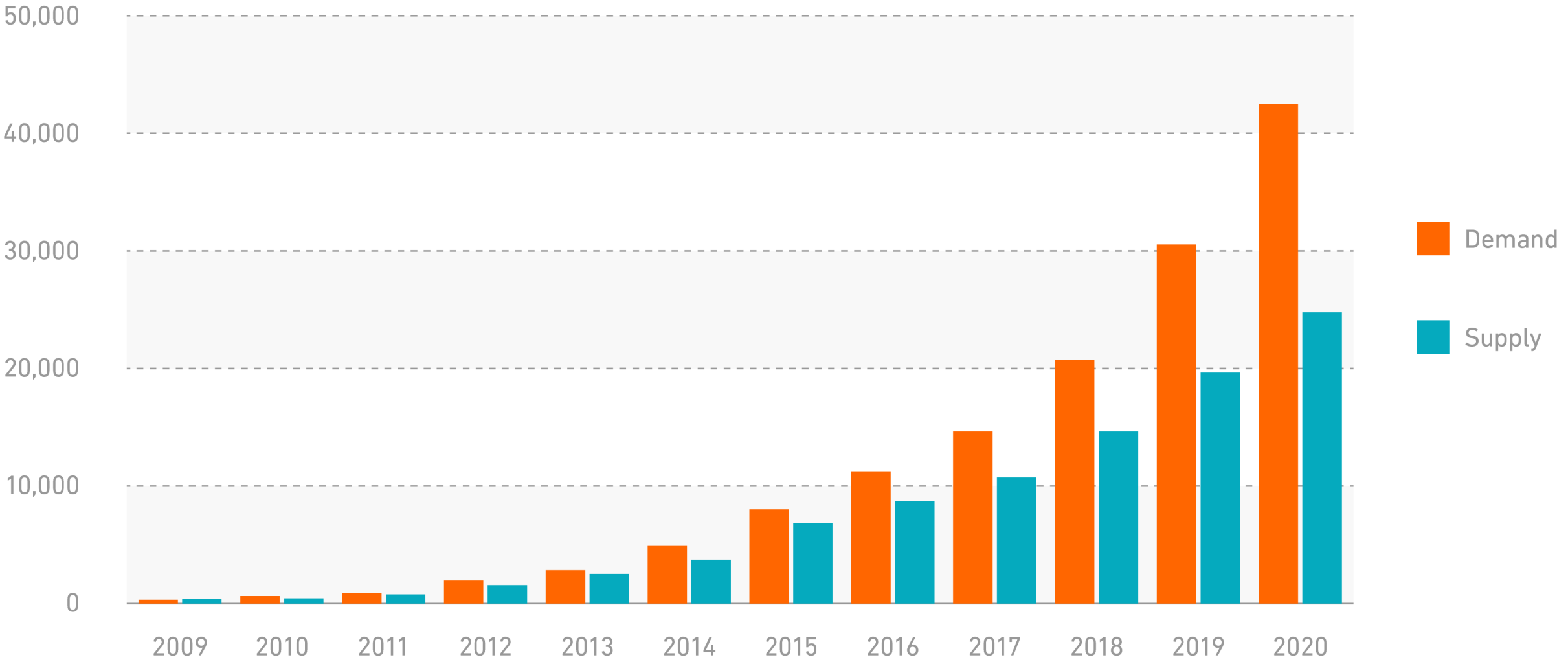
Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

User growth generates data growth



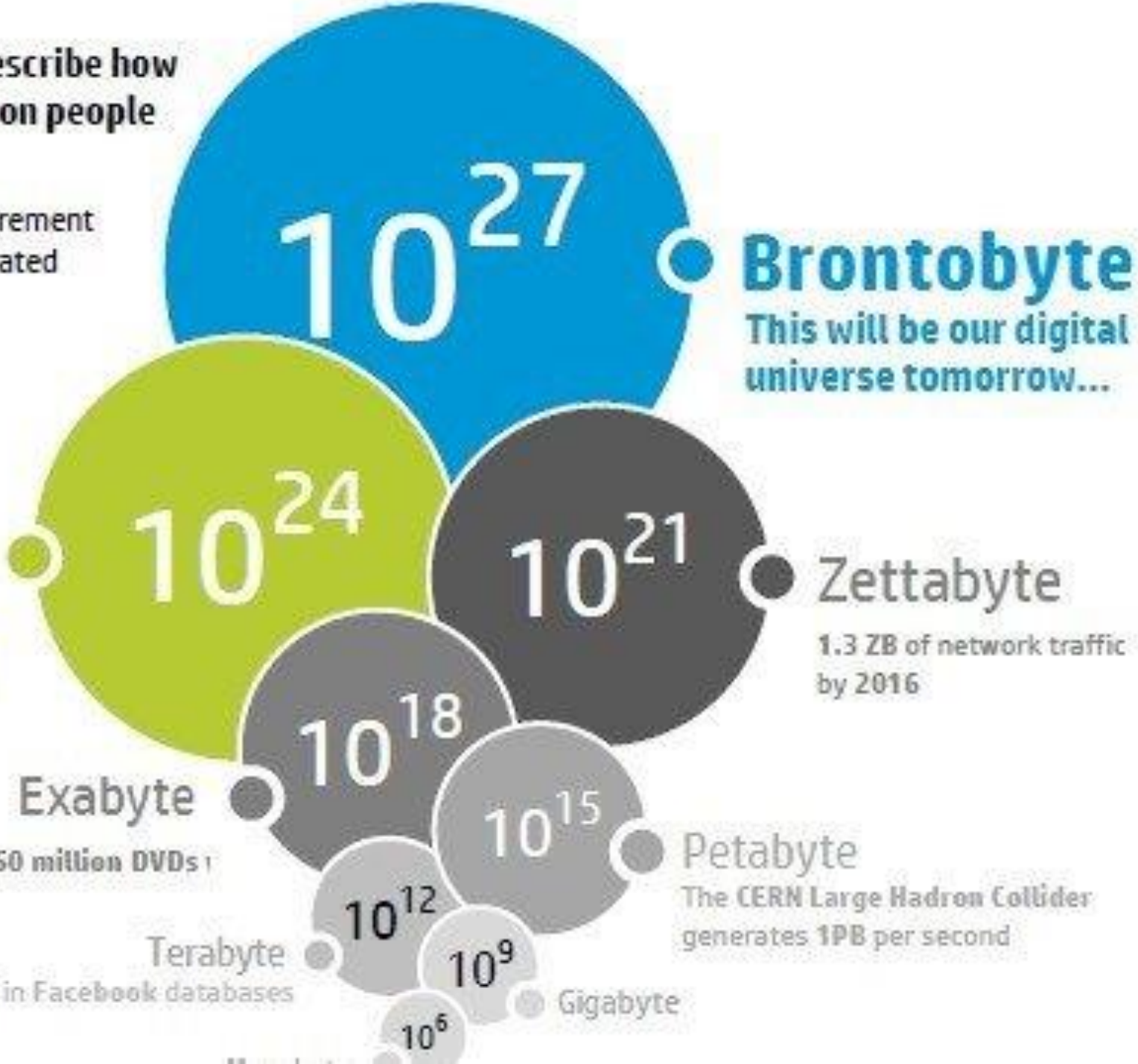


DATA STORAGE SUPPLY AND DEMAND WORLDWIDE, FROM 2009 TO 2020 (IN EXABYTES)



Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)



Demand for analysis

- Consumer companies use Big Data
  - Facebook, Twitter, etc. track user behavior
  - Although they appear to be consumer companies, they monetize user data for advertising
- Now, all businesses expect exploit their company data
  - More efficient transportation (airlines, trucks, etc.)
  - Finding features in data (fraud, drug reactions, cancer prediction, bad drivers, etc.)

Feinleib, D. Big data bootcamp: What managers need to know to profit from the big data revolution. Apress, 2014.

Facebook's use of big data

- Text analysis
  - DeepText analyzes text to see how words are used
- Facial recognition
  - DeepFace learns to recognize faces and their associated tags
- Target advertising
  - Using the above, and user network analysis, AI classifies and clusters users
  - Advertisers pay to put ads in front of certain kinds of users

Massive data requires search

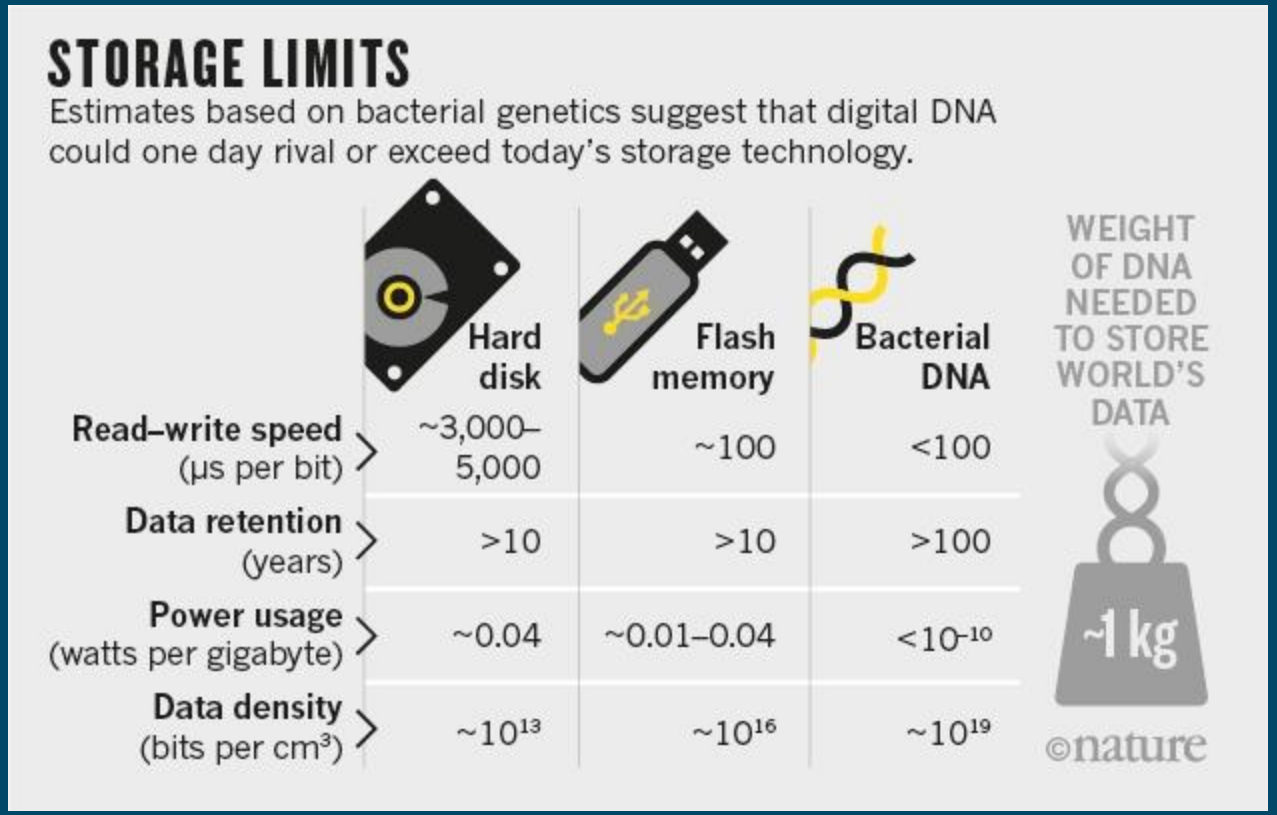
Lots of data requires Search

- Google handles a 100 billion search queries per month
  - Stores web pages and interactions
  - Stores all the searches people make
  - ➔ company has unparalleled insight into the when, what, and how of human search behavior
    - ➔suggests what people are doing at various times of day, locations, etc.
- MapReduce and the Google File System (Hadoop), "reinvented the way Google built its search index“
  --Wired magazine summer of 20I2

Big data can be exploited

More data is the trend

- As technological advances make storing and analyzing data more efficient, companies are doing a lot more analysis, not less.
  - *Jevons paradox*, named for the economist who made this observation about the Industrial Revolution



According to IBM, "every day we create 2.5 quintillion bytes of data—so much that 90% of the data in the world has been created in the last two years alone."

Feinleib, D. *Big data bootcamp: What managers need to know to profit from the big data revolution*. Apress, 2014.

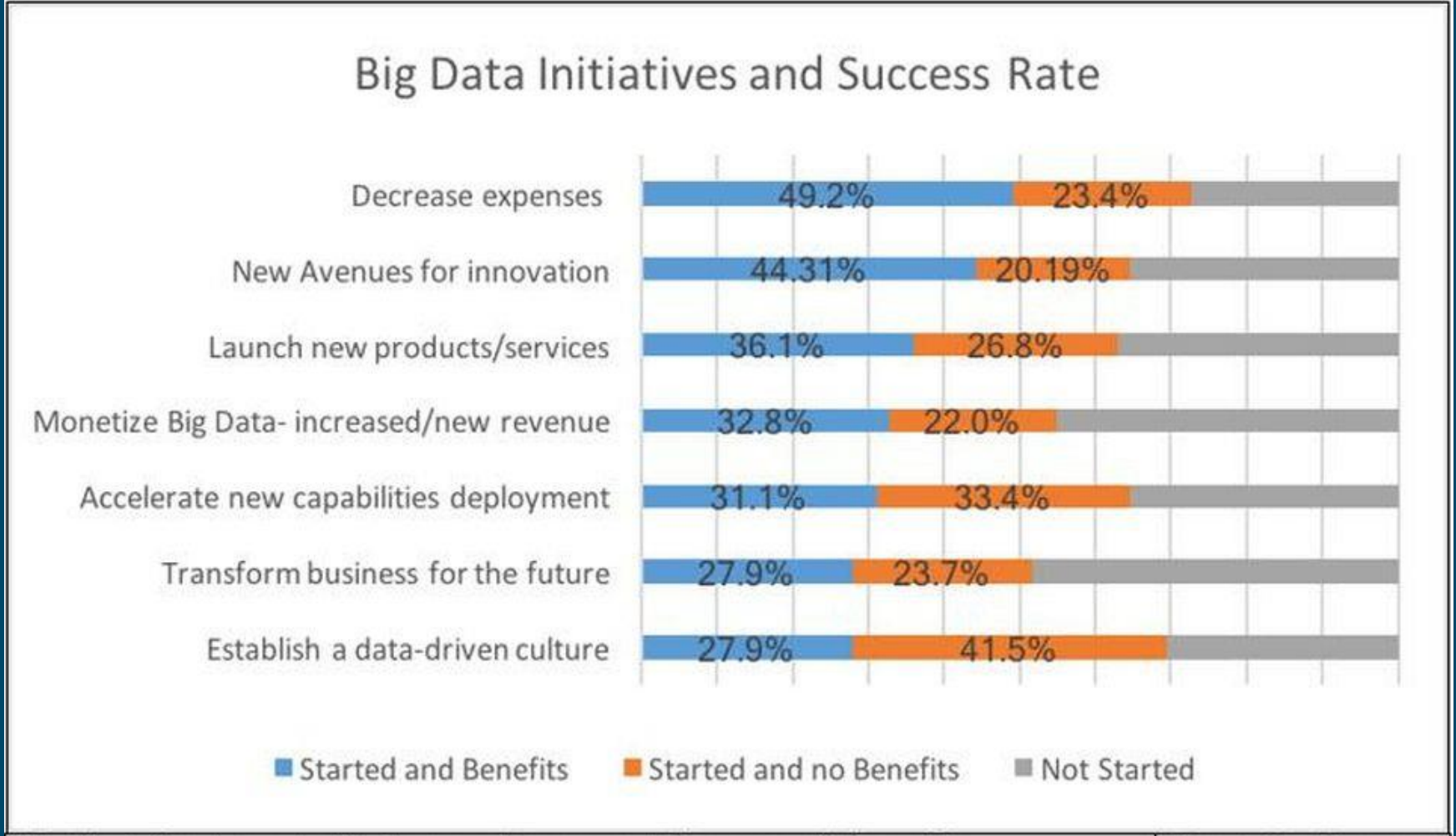
Business demand for Big Data

- **Cost reduction**
  - Big data technologies bring significant cost advantages.
- **Faster, better decision making**
  - With the speed of Hadoop businesses are able to analyze information immediately – and make decisions based on what they've learned.
- **New products and services**
  - With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want.

Big Data = Big Profits

- The **more companies characterized themselves as data-driven, the better** they performed on objective measures of financial and operational results
- Companies in the top third of their industry in the use of data-driven decision making were, on average, **5% more productive and 6% more profitable** than their competitors

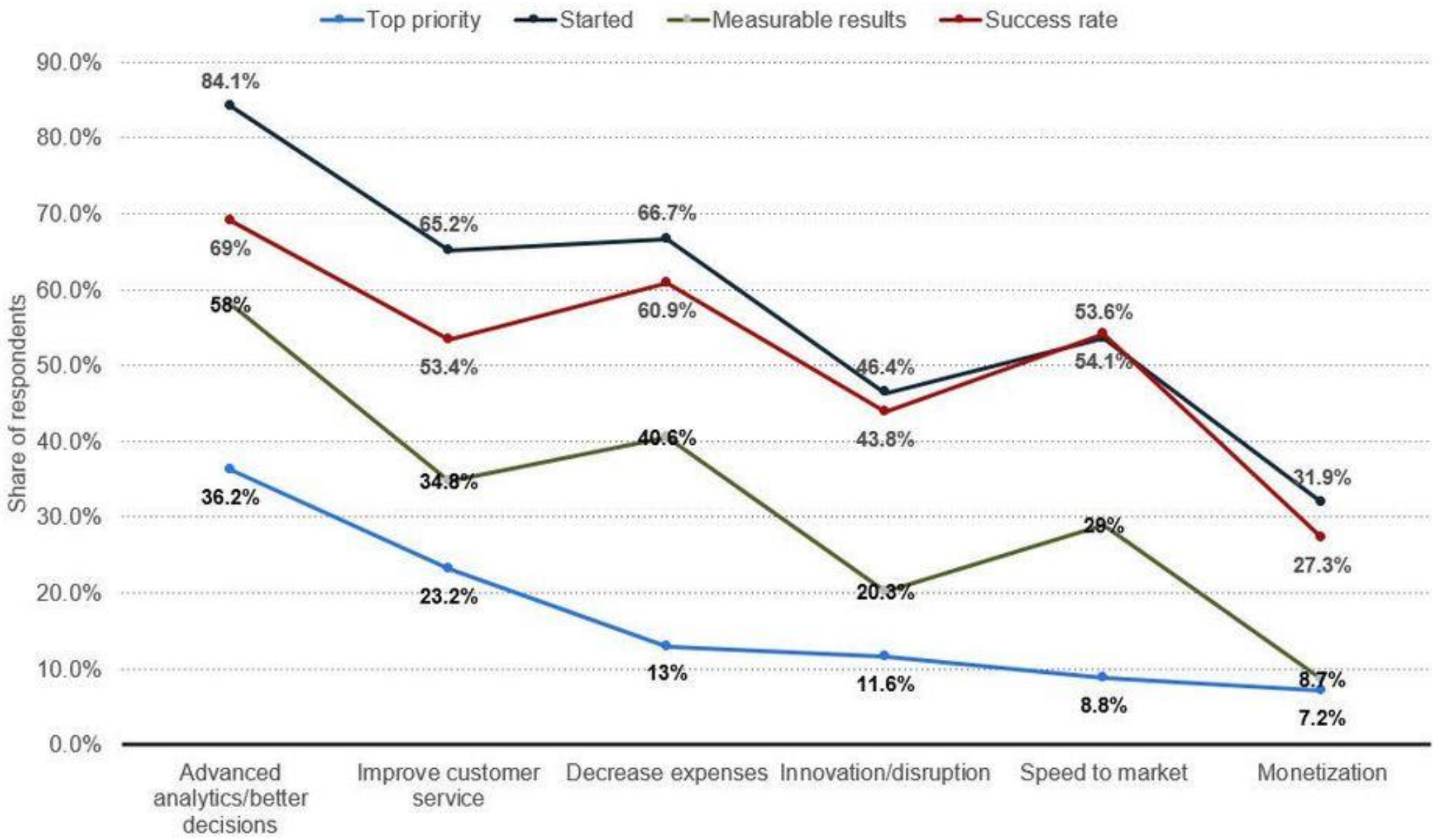
Big Data, Harvard Business Review, October, 2012



Big Data business initiatives underway; with successful results.	**Started**	**Success**
Decrease expenses through operational cost efficiencies	**72.6%**	**49.2%**
Establish a data-driven culture	**69.4%**	**27.9%**
Create new avenues for innovation and disruption	**64.5%**	**44.3%**
Accelerate the speed with which new capabilities and services are deployed	**64.5%**	**31.1%**
Launch new product and service offerings	**62.9%**	**36.1%**
Monetize Big Data through increased revenues and new revenue sources	**54.8%**	**32.8%**
Transform and reposition your business for the future	**51.6%**	**27.9%**

Corporate Big Data Initiative Success Rates U.S. And Worldwide 2018

Big Data Initiatives And Success Rates Among Corporations in the United States and Worldwide, As Of 2018

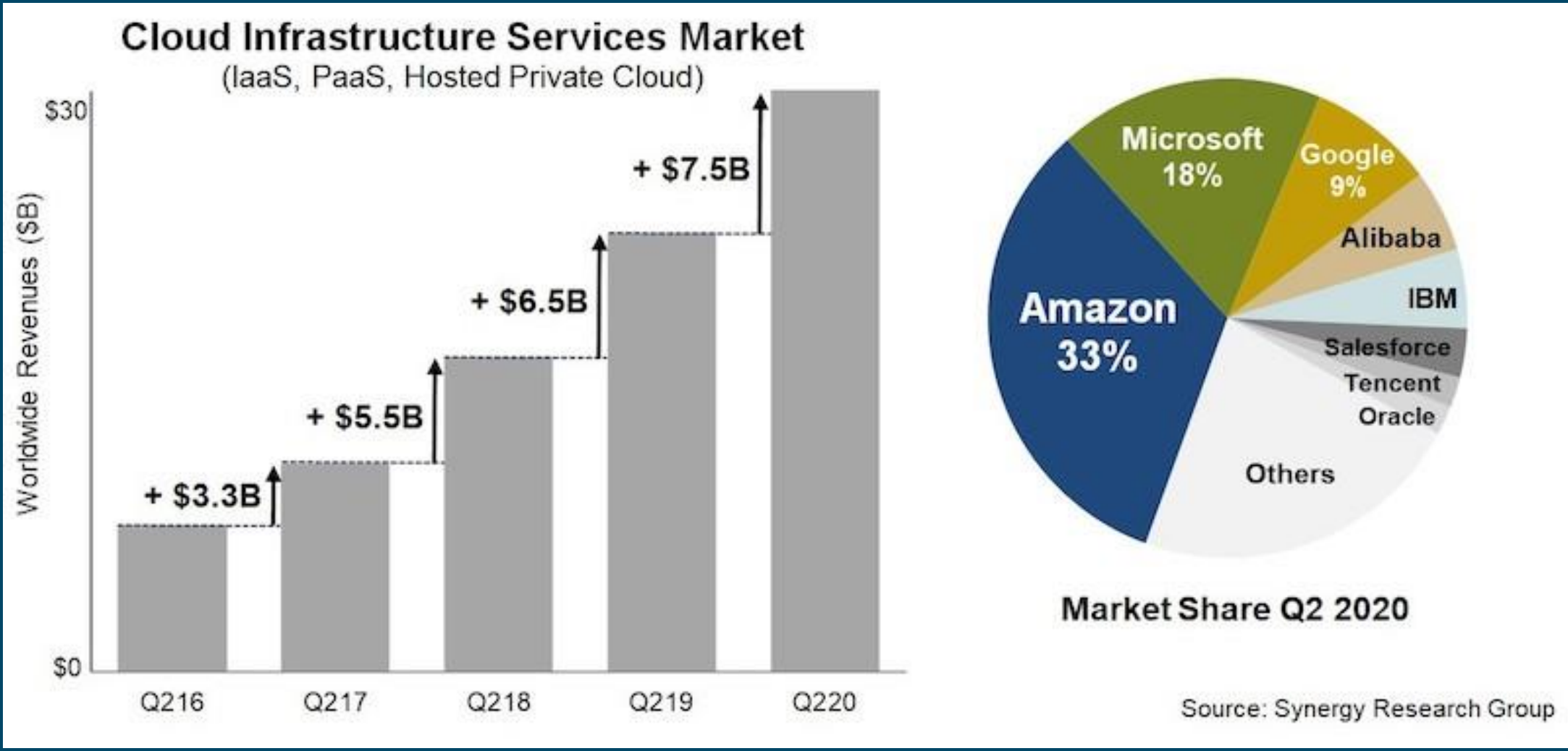


Why big data now?

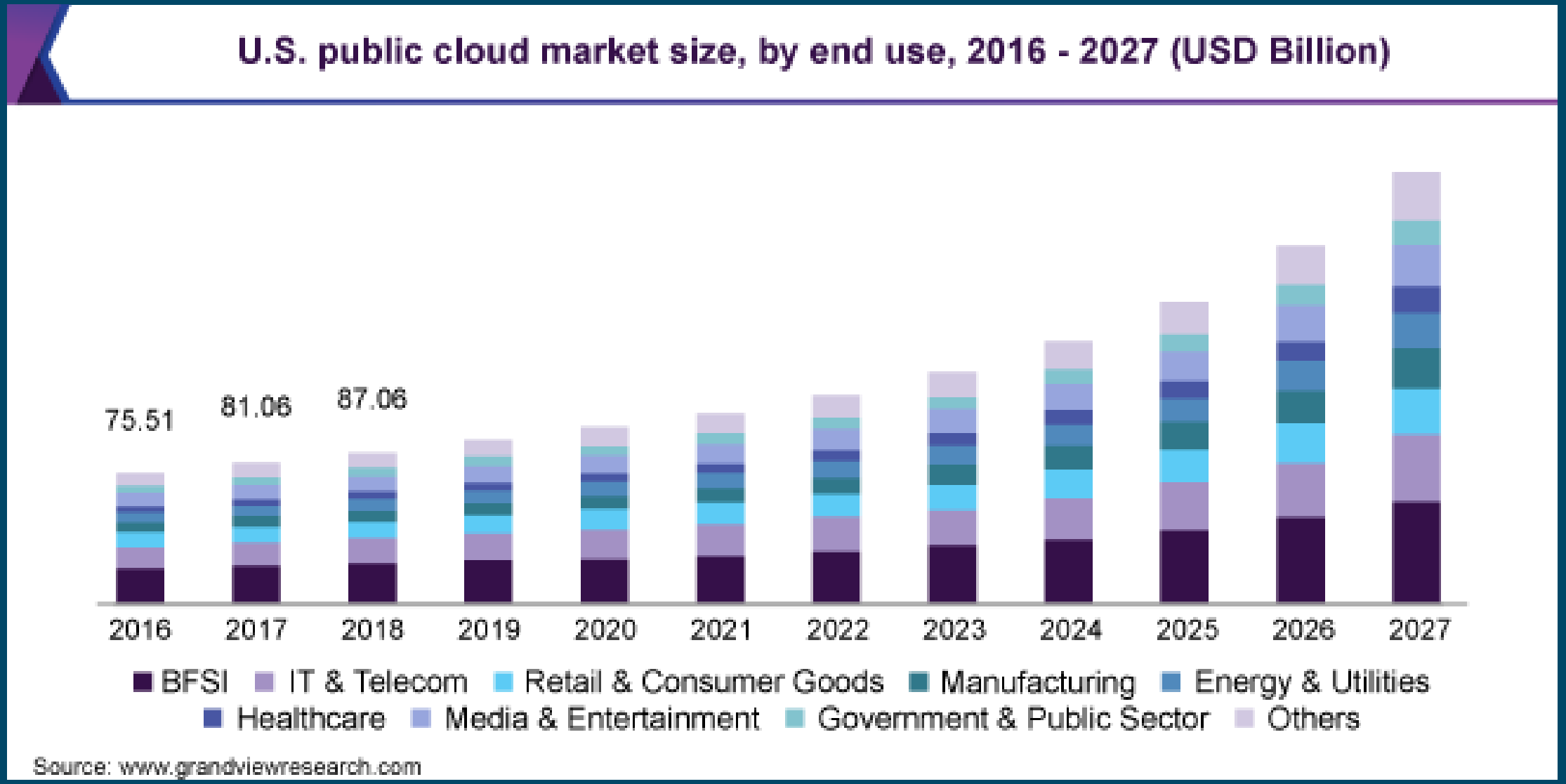
Trends driving Big Data

- Open Source and commodity computing
  - Operating system on low-cost (commodity) hardware to get most of the same functionality that was previously obtained with specialized expensive computers
  - Google, Yahoo!, Facebook extended Linux (e.g, Hadoop) on commodity hardware into distributed databases (historically, an expensive custom solution)
- Commodity Cloud
  - Enterprises are moving computing to the cloud
  - Instead of buying hardware and software and installing it in their own data centers and then maintaining that infrastructure, they're now getting the capabilities they want on demand over the Internet (e.g., AWS)
- In-memory analysis
  - Unlike MapReduce, newer technology (e.g., Spark) makes efficient use of memory to speed up processing
- IoT
  - Internet of things is generating new data and new application opportunities

Let another company manage all the computers



US cloud market



Source: www.grandviewresearch.com

ROBERT MCMILLAN BUSINESS 02.01.13 06:30 AM

MICROSOFT JAMS THOUSANDS OF SERVERS INTO BOXES, BUILDING OUT $1 BILLION DATA CENTER

They're are pre-assembled containers that come jammed with as many as 2,500 servers, that suck cool air in on one side and spit it out on the other.

https://tadviser.com/index.php/Article:Modular_container_DPCs

Examples of using Big Data

Big data processing examples

- Using their customer preferences dataset, Amazon offers access to its self-service ad portal, where companies can buy ad campaigns and target them to ultra-specific demographic
- GE's Flight Efficiency Services can optimize fuel use and more by analyzing the massive volumes of data airplanes generate. How massive? One transatlantic flight generates an average of 1,000 gigabytes.
- Netflix invested $100 million in the first two seasons of *House of Cards*, which premiered in 2013, because consumers who watched the British series *House of Cards* also watched movies directed by David Fincher and starring Kevin Spacey. Executives correctly predicted that a series combining all three would be a hit.

https://builtin.com/big-data/big-data-examples-applications


CROSS-SELLING

More sophisticated models: Cross-selling example

1 Sales History

People who bought product A bought product B.

2 Sales history and event calendar

Modified offer selection based on the fact that the world cup is running.

3 Sales history, events and customer buying history

This customer fits demographic profile A they are buying product X then the model proposes product Y.

4 Sales history, events and customer buying history + social media

Either customer specific or general + what is trending.

Shopping cart analysis

? Where should detergents be placed in the Store to maximize their sales?

? Are window cleaning products purchased when detergents and orange juice are bought together?

? Is soda typically purchased with bananas? Does the brand of soda make a difference?

? How are the demographics of the neighborhood affecting what customers are buying?

Image source: deepclimate.org

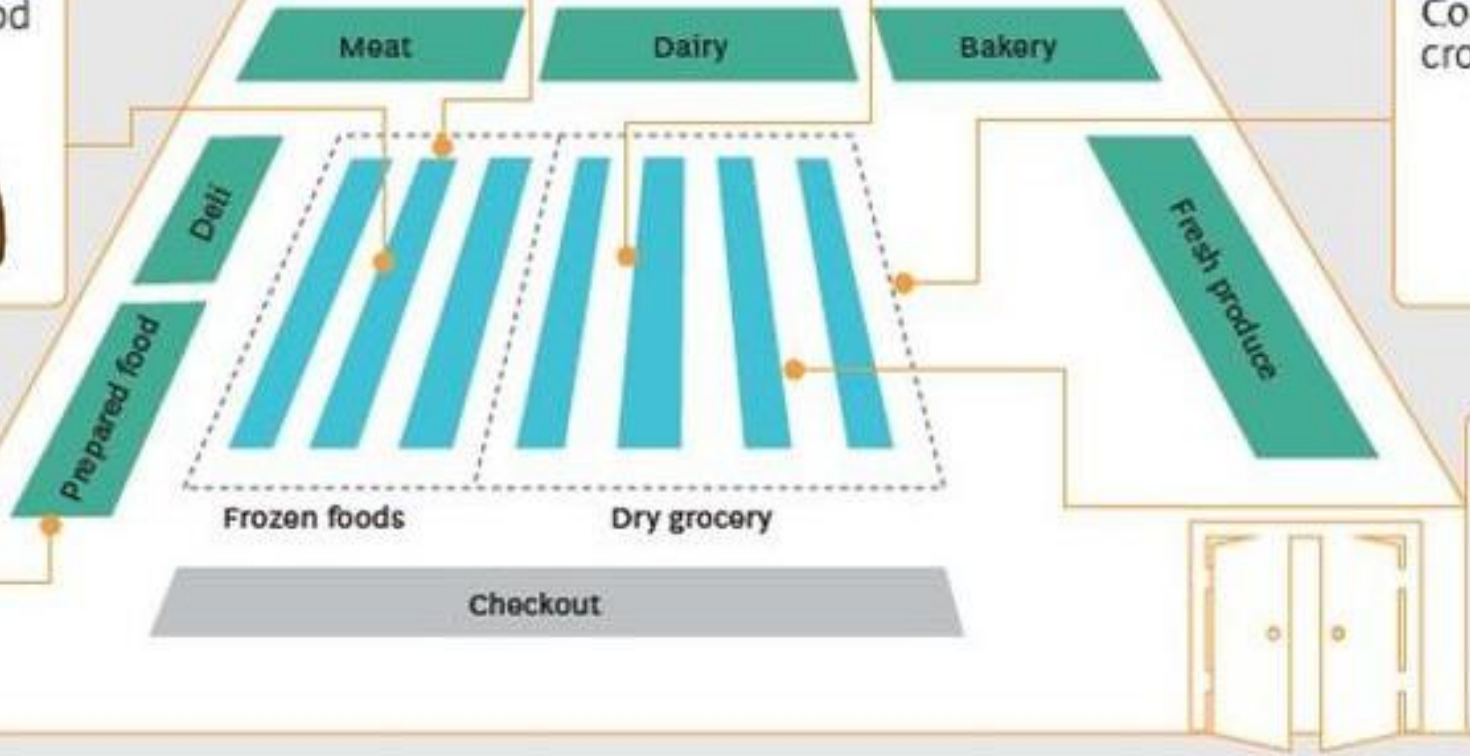
A Grocer's Data-Driven Merchandising Makeover

Reset endcaps to appeal to the local market

Optimize prices on items that increase local foot traffic

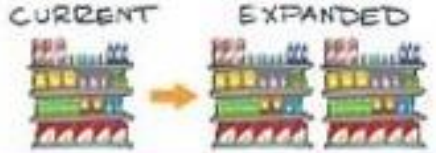
Double-face frozen-food items popular in this location

Co-locate items to drive cross-sales



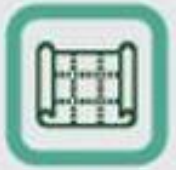
Expand prepared-food and deli sections for heavy lunch traffic

Expand international-foods section to reflect local demand



Source: BCG analysis.

How Big Data Can Enhance Elements of Localization

ELEMENT	FLOOR SPACE	ASSORTMENT	ADJACENCIES	STORE DESIGN	PRICE	PROMOTION
						
VALUE CREATED	• Shift space between categories • Improve physical merchandising • Reduce wasted space • Allocate seasonal space and calendars	• Improve product mix based on purchase history • Add popular local items • Optimize facings • Determine which products should have dual locations • Optimize endcaps, drive aisles, other displays	• Adjust category locations and product positions • Minimize moves to distant aisles • Optimize SKU position, such as height on shelf, order in aisle	• Redesign format or floor space • Change location of departments or services • Add experiential elements for store clustering	• Create zones for stores with similar competitors • Designate willingness-to-pay zones • Define regional known-value items • Define local dynamic pricing	• Select localized promotions • Set level, timing, and frequency of discounts by cluster • Localize seasonal and endcap promotions

Source: BCG analysis.

Amazon and Big Data

- Amazon.com
  - Every time a customer searches for a TV show to watch or a product to buy on the company's web site, Amazon gets a little more insight about that customer.
    - Based on searches and product purchasing behavior, Amazon can figure out what products to recommend next.
    - The data shows what's working and what isn't, and cases for new business investments must be supported by data.
- Amazon Web Services (AWS)
  - Provides infrastructure for rent

AWS Reference Architectures

Amazon EC2
Amazon EC2 Spot
Amazon EMR
Amazon S3
Amazon CloudFront
Amazon RDS

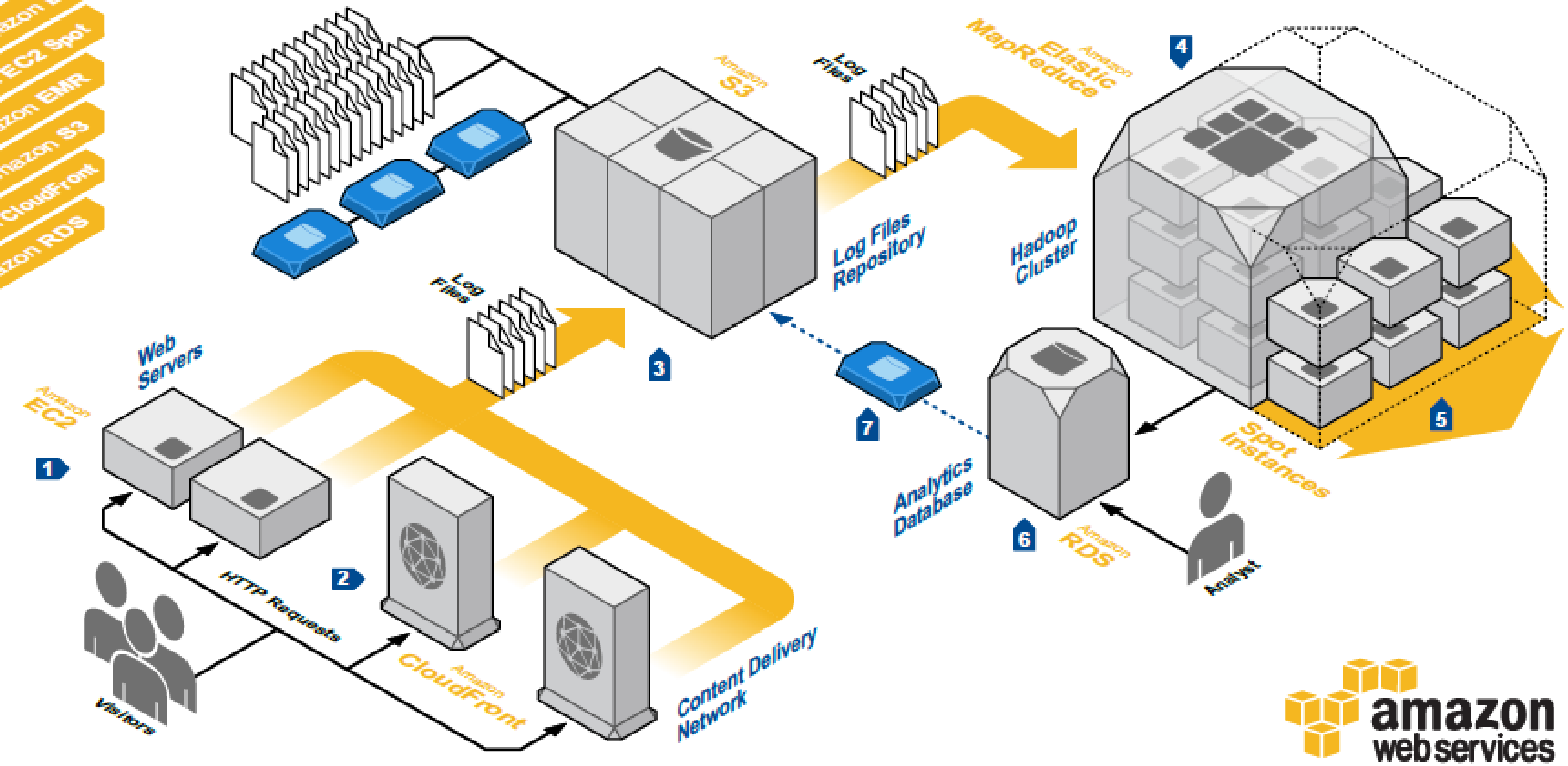
WEB LOG ANALYSIS

Amazon Web Services provides services and infrastructure to build reliable, fault-tolerant, and highly available web applications in the cloud. In production environments, these applications can generate huge amounts of log information.

This data can be an important source of knowledge for any company that is operating web applications. Analyzing logs can reveal information such as traffic patterns, user behavior, marketing profiles, etc.

However, as the web application grows and the number of visitors increases, storing and analyzing web logs becomes increasingly challenging.

This diagram shows how to use Amazon Web Services to build a scalable and reliable large-scale log analytics platform. The core component of this architecture is Amazon Elastic MapReduce, a web service that enables analysts to process large amounts of data easily and cost-effectively using a Hadoop hosted framework.



System Overview

1 The web front-end servers are running on **Amazon Elastic Compute Cloud** (Amazon EC2) instances.

2 **Amazon CloudFront** is a content delivery network that uses low latency and high data transfer speeds to distribute static files to customers. This service also generates valuable log information.

3 Log files are periodically uploaded to **Amazon Simple Storage Service** (Amazon S3), a highly available and reliable data store. Data is sent in parallel from multiple web servers or edge locations.

4 An **Amazon Elastic MapReduce** cluster processes the data set. **Amazon Elastic MapReduce** utilizes a hosted Hadoop framework, which processes the data in a parallel job flow.

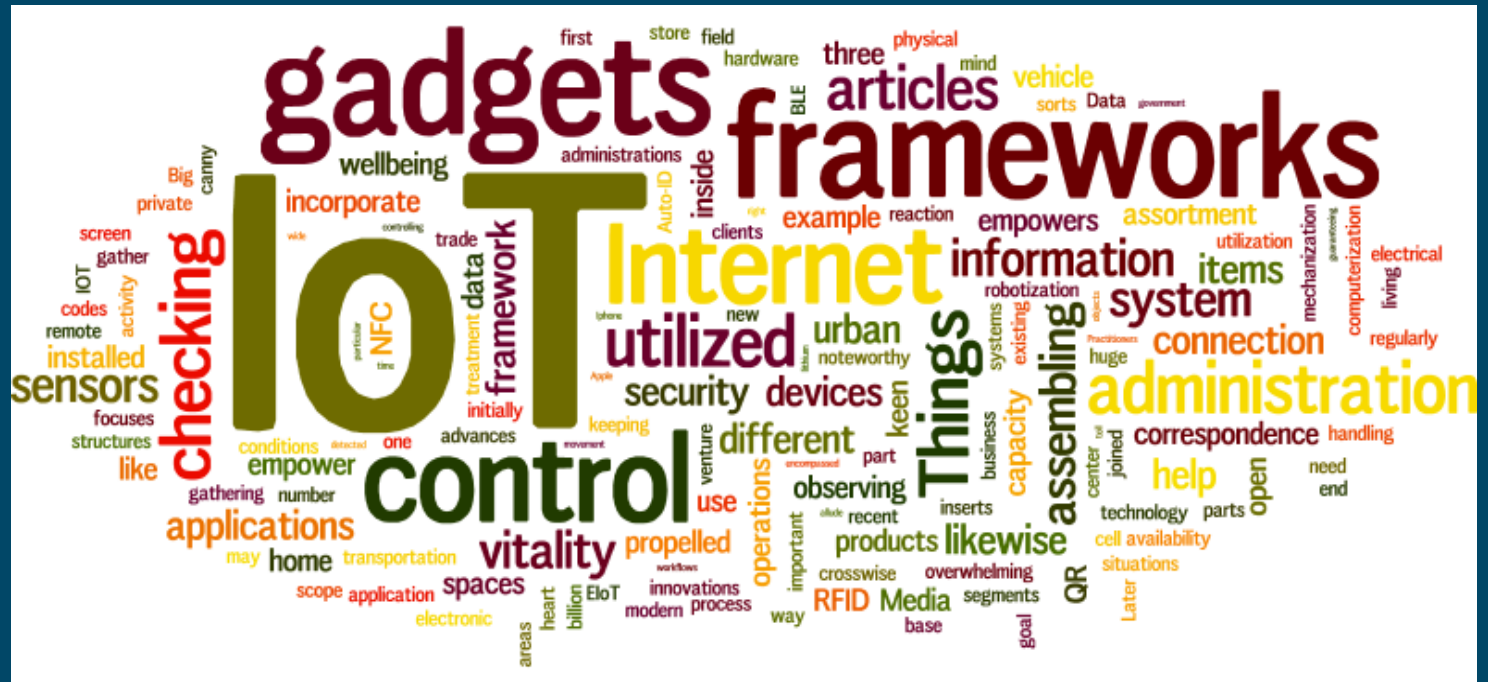
5 When **Amazon EC2** has unused capacity, it offers EC2 instances at a reduced cost, called the *Spot Price*. This price fluctuates based on availability and demand. If your workload is flexible in terms of time of completion or required capacity, you can dynamically extend the capacity of your cluster using **Spot Instances** and significantly reduce the cost of running your job flows.

6 Data processing results are pushed back to a relational database using tools like *Apache Hive*. The database can be an **Amazon Relational Database Service** (Amazon RDS) instance. **Amazon RDS** makes it easy to set up, operate, and scale a relational database in the cloud.

7 Like many services, **Amazon RDS** instances are priced on a pay-as-you-go model. After analysis, the database can be backed-up into **Amazon S3** as a database snapshot, and then terminated. The database can then be recreated from the snapshot whenever needed.

Internet of things (IoT)

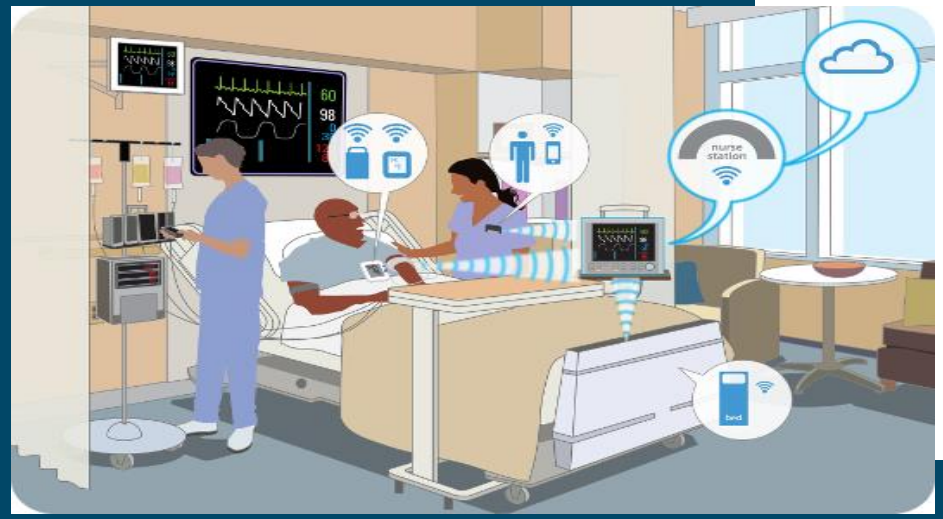
- … is the network of physical objects—devices, vehicles, buildings and other items embedded with electronics, software, sensors, and network connectivity—that enables these objects to collect and exchange data.



IoT = all data, all the time

Wearable Tech

Smart Appliances



Healthcare

What Makes IoT Analytics Different?

High volume, continuous "data in motion" from multiple sensors

Store, blend and manage time-series data

Use of multiple analytics techniques

Distributed analytics (edge)

Integration with operation systems and BPMS

Bidirectional communication and control of endpoints

More data

More complexity

More automation

Source: Gartner

Gartner

SIG LIFE APPROACH TO CONTAINER ORCHESTRATION


Apache Kafka
in the
Automotive Industry
Uber
kafka

Important to remember

- User and devices are generating lots of data
- Trends driving big data analysis
  - Lots of data (IoT), Open-source computing, cloud computing, effective analyses (Spark)
- Big data analysis has good ROI