A History of Big Data

Big Data headlines

How Big Data can help save endangered kids

By Naomi Schaefer Riley October 16, 2016

Tech / #BigData

OCT 6, 2016 @ 02:27 AM 18,708 VIEWS

How AI, Drones And Big Data Are Reshaping The Future Of Warfare

Bernard Marr, CONTRIBUTOR

I write about big data, analytics and enterprise perform

Opinions e

Using Big Data to Predict Terrorist Acts Amid Privacy Concerns

By Chris Strohm | October 13, 2016

How complexity is killing big data deployments

Big data applications are 10X more complex than regular apps, and developers often need to know a plethora of technologies just to make big data work.

By Matt Asay | October 13, 2016, 8:12 AM PST

How did big data start?

- Google & Yahoo! need to index web pages faster
- Prepare for web search (i.e., google *search terms*…)
  - Collect all web pages and place into storage
  - Create indices for each page (to be used in search)
    - Count the occurrence of non-stop words (e.g., “the”, “a”) on a page
    - Compute over 200 “signals” about a page to enable search
  - Repeat the process for all new or updated pages
- Processing billions of web pages (2003) was taking too long

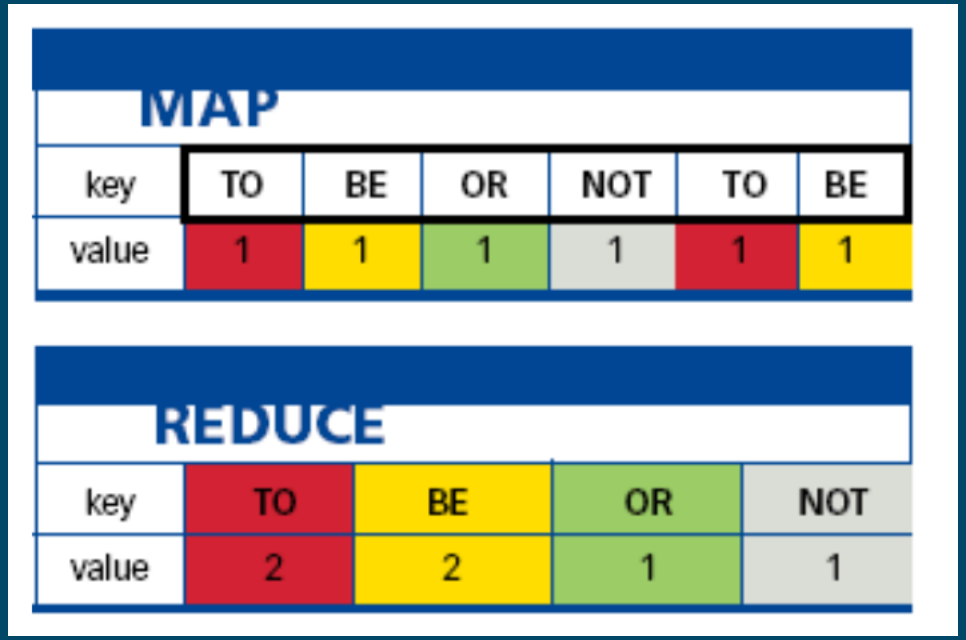
https://searchengineland.com/8-major-google-ranking-signals-2017-278450

“We were trying to handle literally billions of web pages on machines … So there was no option but to latch hundreds to thousands of machines together to build the index. So it was out of desperation that MapReduce was invented.”

https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/

Google MapReduce

- A typical first-week training assignment for a new programmer hired by Google is to write a software routine that uses MapReduce to count all occurrences of words in a set of Web documents.
- 2005, October, Google was running about 3,000 computing jobs per day through MapReduce, representing thousands of machine-days

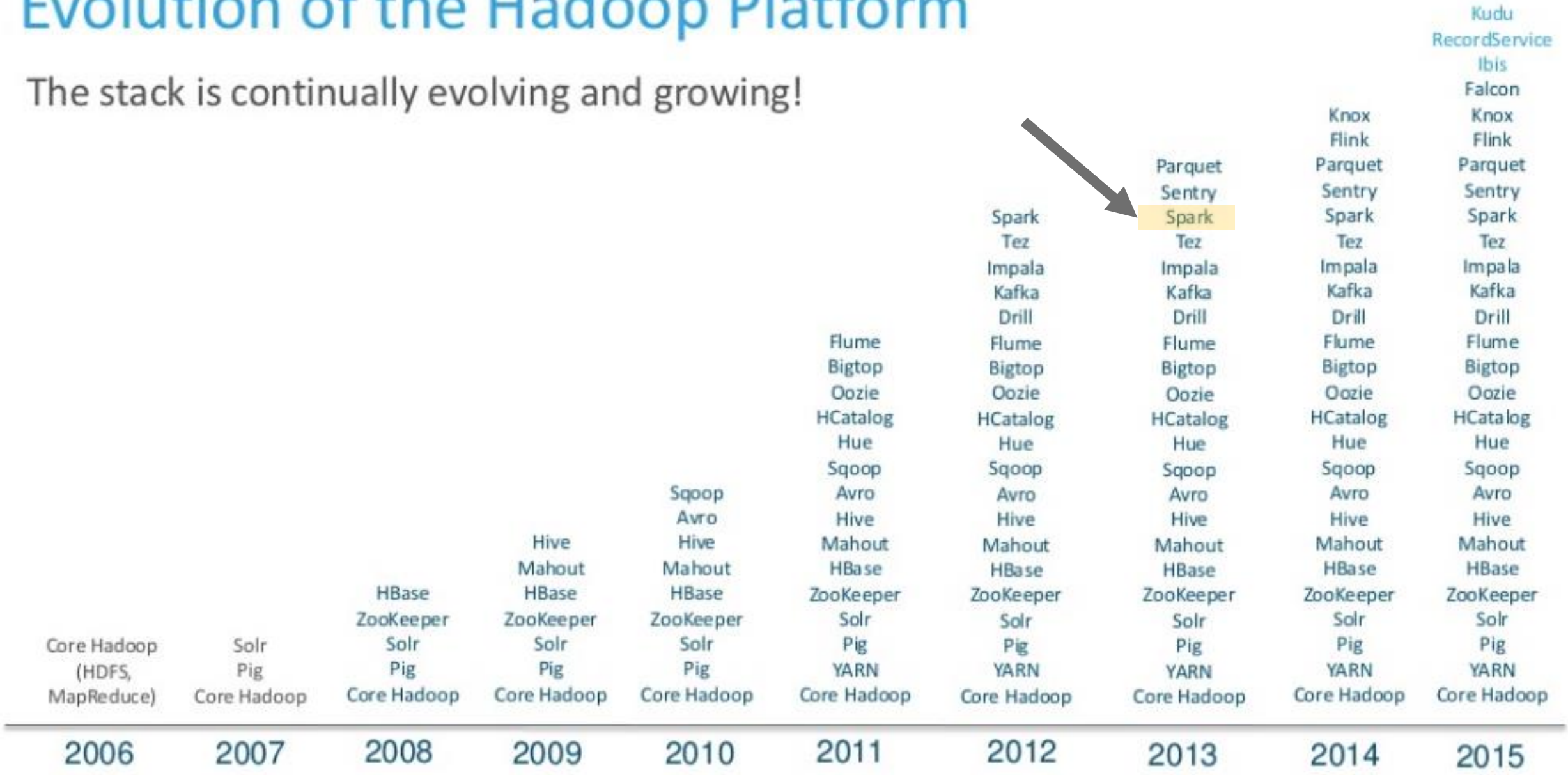


http://www.baselinemag.com/c/a/Infrastructure/How-Google-Works-1/5

Hadoop's Original Architecture

Evolution of the Hadoop Platform

The stack is continually evolving and growing!



Google: From MapReduce to Google Dataflow

- Google big data for search
  - 2004 MapReduce
  - 2010 MapReduce
    - “it got too cumbersome once the size of the data reached a few petabytes.”
  - 2011 (circa) Dataflow
    - Similar to Spark
      - Traditional MapReduce programming model needs hand-crafting a chain of jobs, which are very tricky and hard to read, maintain.
      - Dataflow can take a rather straightforward sequential code, interpret them into MapReduce steps, and optimize them.
    - Computes and processes DAG of data processing

https://www.datacenterknowledge.com/archives/2014/06/25/google-dumps-mapreduce-favor-new-hyper-scale-analytics-system; https://www.quora.com/Why-did-Google-stop-using-MapReduce-and-start-encouraging-Cloud-Dataflow

Big Data Infrastructure

Other Applications

Finance, HR

workday. · ORACLE · NETSUITE · CONCUR

Customer Success

Collaboration

box · Dropbox · zendesk · INTERCOM

CRM

salesforce · ORACLE · Microsoft Dynamics

Visualization, BI

Microsoft · tableau SOFTWARE · Qlik · D3 · DOMO · GoodData · birst

Analytics

MATLAB · SAS · R · Excel · Apache Spark

Application Performance

DATADOG · APPDYNAMICS · New Relic · Information Builders · MicroStrategy

Machine data, Logs

sumologic · splunk>

Geo-spatial

ArcGIS · PARSE.LY

Web, Social Media

Hootsuite · Google Analytics

Data Warehousing, Integration

TERADATA · informatica · Hortonworks · alteryx · Greenplum · talend · cloudera

DevOps

Build

Bamboo · Jenkins

Manage Config

ANSIBLE · puppet labs · CHEF · SALTSTACK

Immutable Infra

VAGRANT · TECTONIC · docker SWARM

Security

RSA

Endpoint

Symantec · intel Security · KASPERSKY

Enterprise MDM

airwatch by vmware · MobileIron

Apps, CASB

netskope · skyhigh · Symantec + BLUE COAT

FireEye

Enterprise VPN

f5 · CITRIX · CISCO AnyConnect Secure Mobility

Network

paloalto NETWORKS · Barracuda

Database as a Service (DBaaS)

Microsoft Azure SQL · amazon web services · RDS, Aurora · Google Cloud Platform · ORACLE Cloud · mongoDB. Atlas

Database Management Systems

SYBASE · *Relational (SQL)* · ORACLE DATABASE 12c · IBM DB2 · SQLite · Microsoft SQL Server · SAP HANA · MySQL · PostgreSQL

NoSQL

mongoDB · cassandra · redis · APACHE HBASE

Graph

neo4j

Virtualization, Containers, Data Center Operating System

CITRIX XenServer · Parallels · vmware · Microsoft Hyper-V · docker · MESOS · kubernetes · ClusterHQ Container Data Management · CoreOS · Joyent · shippable · MESOSPHERE · DC/OS

Host OS, File System, Cluster / Resource Management

Windows Server · redhat · ubuntu · SUSE · HDFS · hadoop YARN

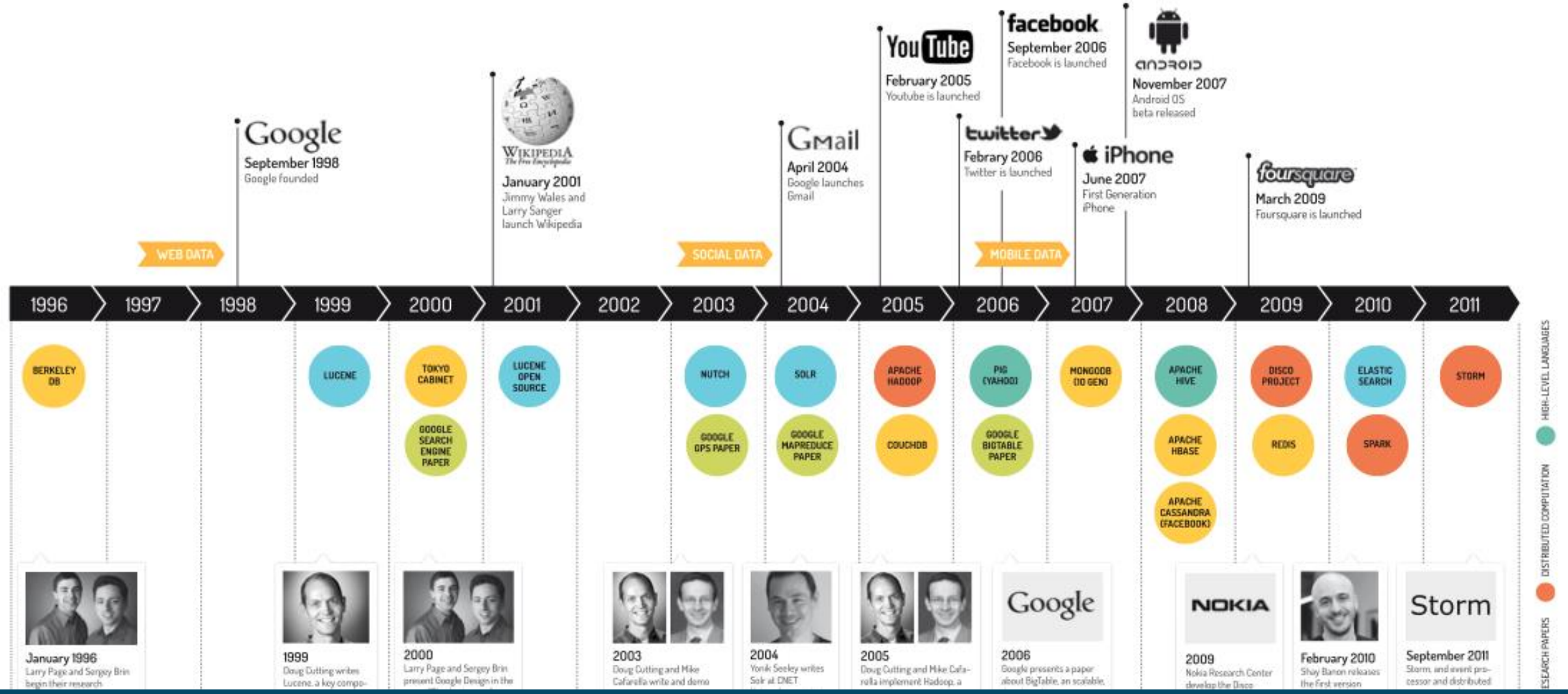
Hardware (Compute, Storage, Networking) / Infrastructure as a Service (Cloud IaaS)

IBM · lenovo · EMC² · DELL · NUTANIX · amazon web services · Microsoft Azure · Google Cloud Platform · CISCO · Hewlett Packard Enterprise · PURESTORAGE · Tintri

Rahul Venkatraj (vrahul@alumni.stanford.edu) Last Updated December, 2016. This is not a comprehensive list of offerings.

BIG DATA

A BRIEF HISTORY



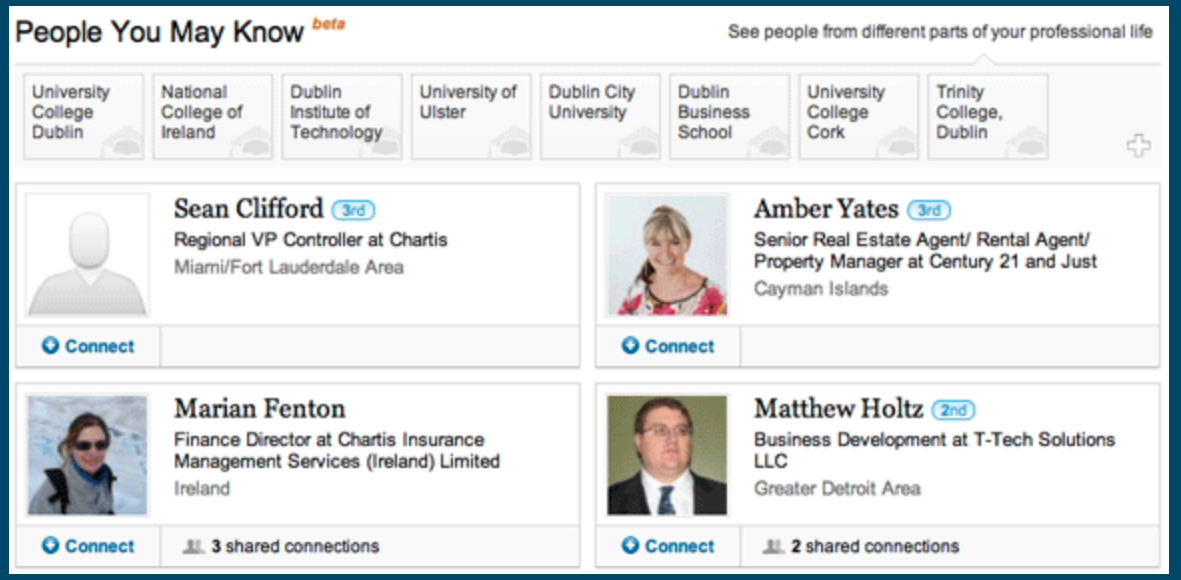

What is
Hadoop?

Big Data applications

- **Facebook** stores and uses Big Data in the form of user profiles, photos, messages, and advertisements.
- **Google** crawls billions of web pages and has a vast array of other Big Data sources.
- **LinkedIn** hosts hundreds of millions of online resumes as well as the knowledge about how people are connected with each other. The company uses all that data to suggest the subset of people with whom we might want to connect.
- **Pandora** uses some 450 song attributes to figure out what songs to recommend..
- **Netflix** is well-known for its movie prediction algorithms, which enable it to suggest to movie viewers what movie to watch next

LinkedIn “People you may know”

- 2008
  - 40 – 50 members
  - Running on Oracle
    - When doesn’t crash, takes 6 weeks to 6 months
- 2009
  - Hadoop 20 node cluster
  - Completes in 3 days!



https://www.slideshare.net/cwsteinbach/the-past-present-and-future-of-hadoop-at-linkedin

Big Data as a computing domain

Big Data infrastructure is…

- software to support processing of very large datasets to exploit (mostly) embarrassingly parallel computation
  - Little or no manipulation is needed to separate the problem into parallel tasks
  - Little or no dependency between those parallel tasks

https://en.wikipedia.org/wiki/Embarrassingly_parallel

Big Data

- the study and applications of data sets that are too complex for traditional data-processing application software to adequately deal with.
- challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source.
- originally associated with three key concepts: volume, variety, and velocity.

Issues addressed by Big Data

- V's
  - Volume – lots of data (lots!)
  - Variety – structure & unstructured
  - Velocity – faster than transactional data (like CC's); think sensor data… so fast that some RDMS has trouble writing it all
  - Value – make sense of the data so that action can be taken (in time)
- Big Data is data that exceeds the processing capacity of conventional database systems.
  - The data is too big, moves too fast, or doesn't fit the strictures of your database architectures.
  - To gain value from this data, you must choose an alternative way to process

Big Data defined as Big Impact (Value)

- What Big Data is really all about is the ability to capture and analyze data and gain actionable insights from that data at a much lower cost than was historically possible.
- No longer do we need complex software that takes months or years to set up and use.
  - Nearly all the analytics power we need is available through software downloads or in the cloud.

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

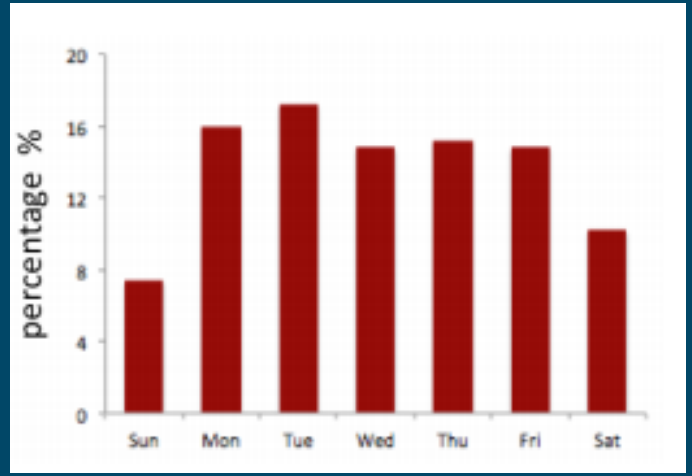
Google Flu Prediction

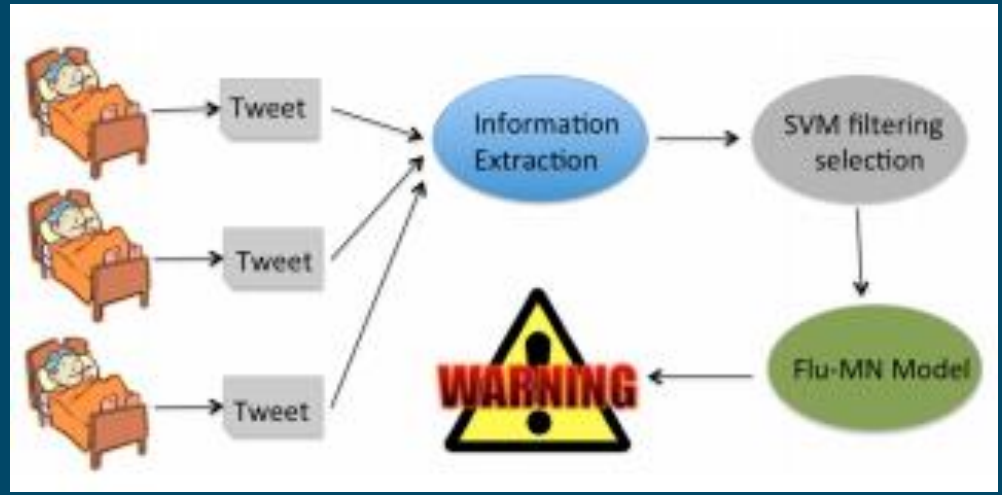
- a strong correlation between a rise in Internet searches for flu information, and a subsequent rise in people coming into a busy urban hospital emergency room complaining of flu-like symptoms
- The initial Google paper stated that the Google Flu Trends predictions were 97% accurate comparing with CDC data

Hopkins Researchers Find "Google Flu Trends" A Powerful Early Warning System for Emergency Departments

Release Date: January 9, 2012

FO



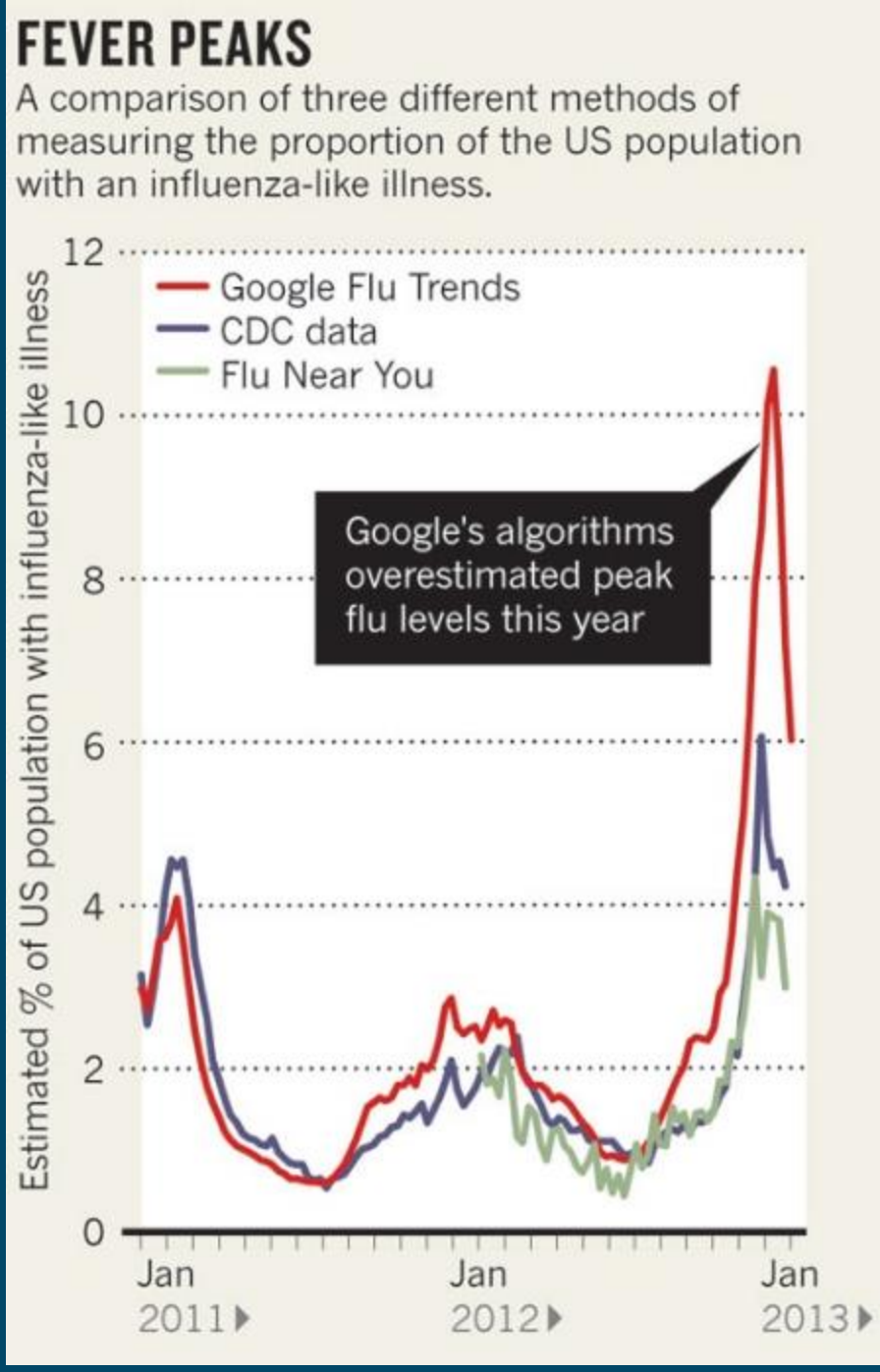


Google Flu Failure

DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

- Some predictions have been very inaccurate—especially over the interval 2011-2013
- Researchers suggest that the problems may be due to widespread media coverage of this year's severe US flu season, including the declaration of a public-health emergency by New York state last month.
  - The press reports may have triggered many flu-related searches by people who were not ill

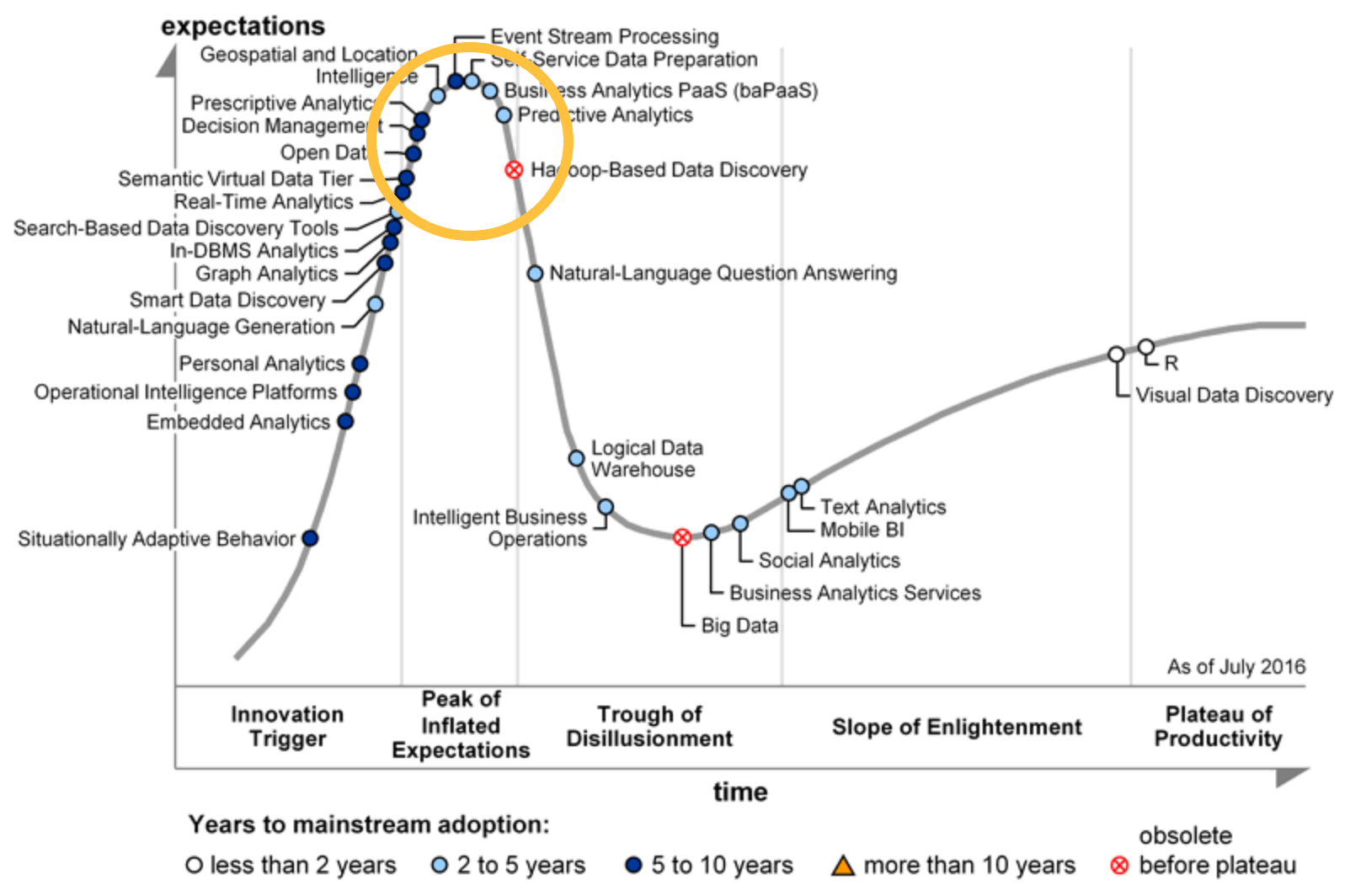


Data warehouse and data management solutions for analytics

- Hadoop technologies has brought new life to an old field (databases)

Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics; **Published:** 25 February 2016; Gartner

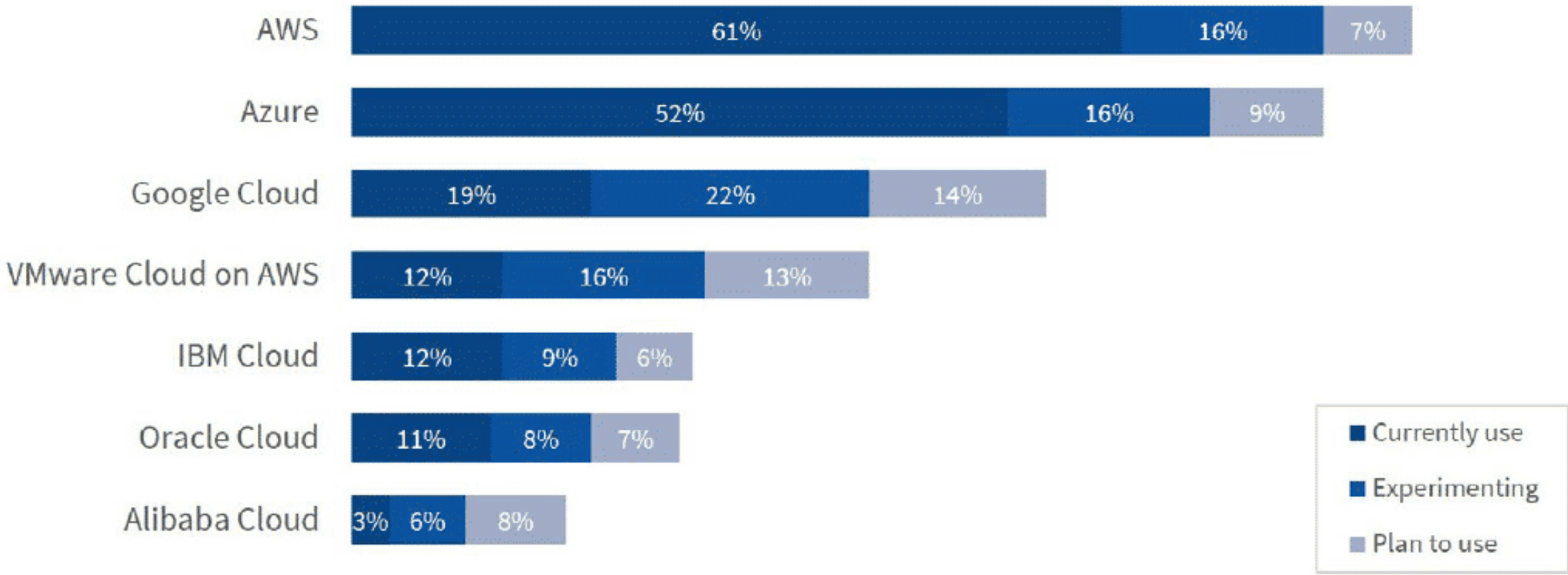
Gartner's Hype cycle for analytics



Hype Cycle for Business Intelligence and Analytics, 2016; **Published:** 25 July 2016; Gartner

Public Cloud Adoption

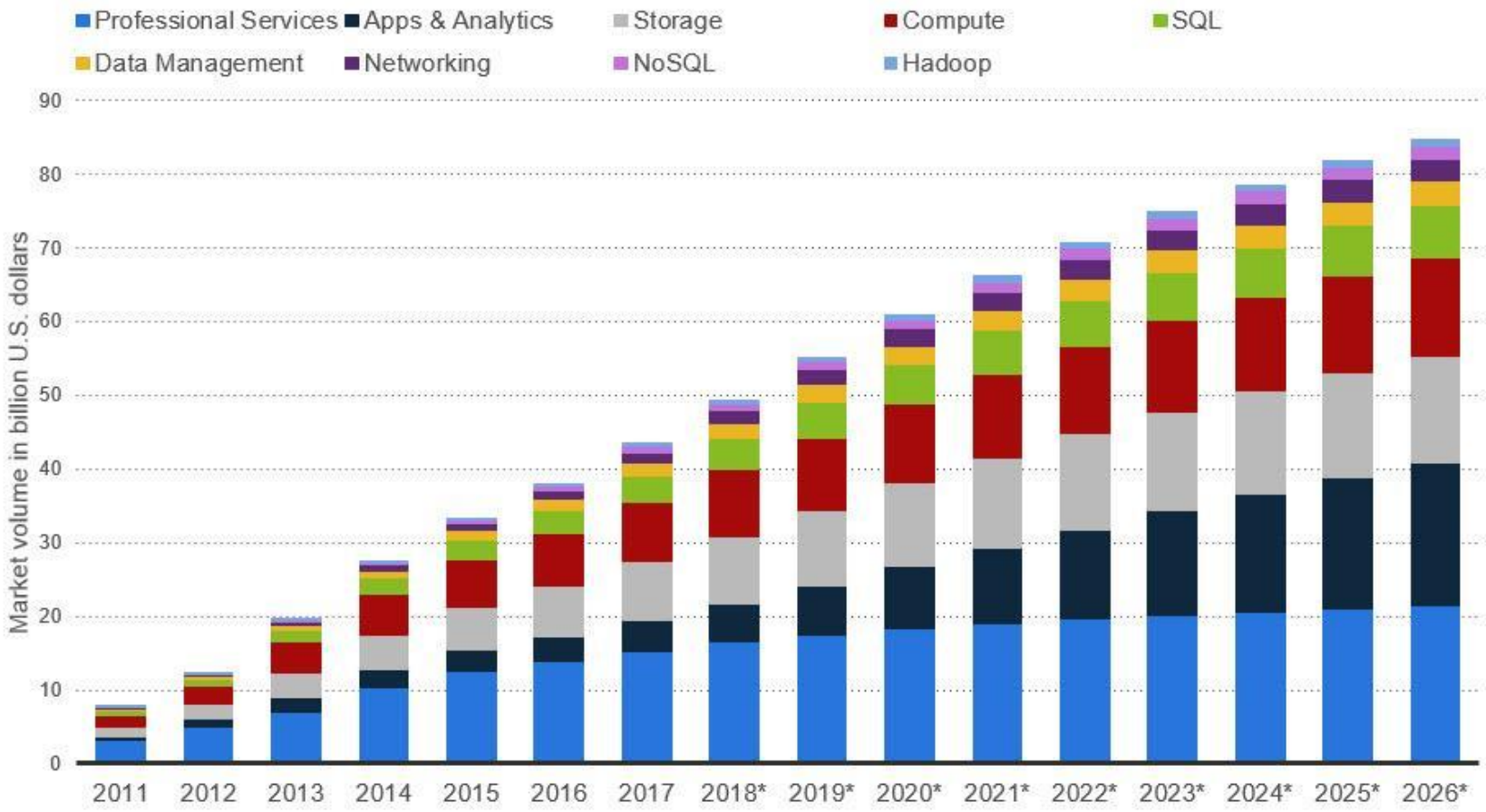
% of All Respondents



Source: RightScale 2019 State of the Cloud Report from Flexera

Big Data Market Worldwide Segment Revenue Forecast 2011-2026

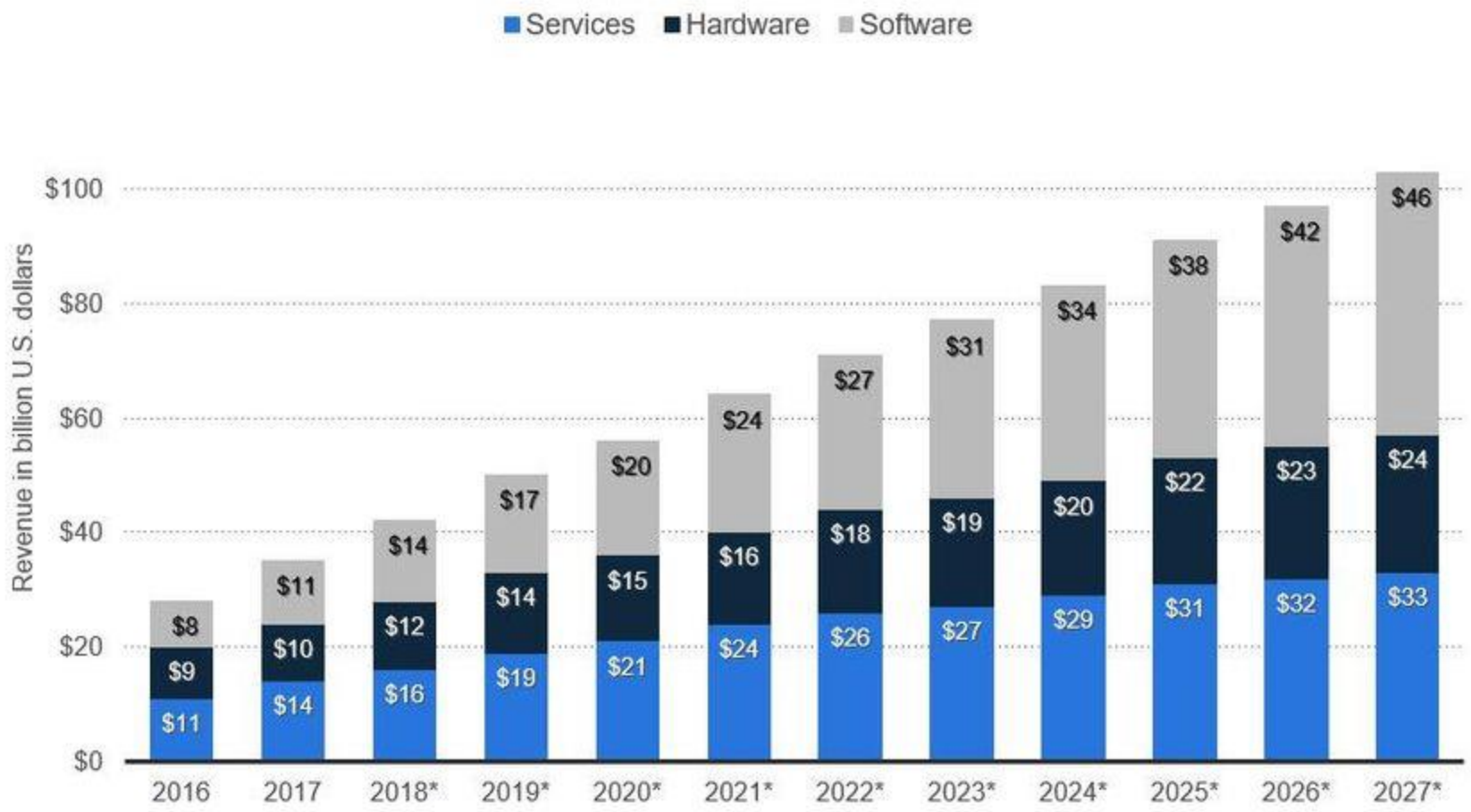
Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



statista

Global Big Data Revenue 2016-2027, by type

Big Data Revenue Worldwide from 2016 to 2027, by major segment (in billion U.S. dollars)



https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#76fc932b2926

Important to remember

- Hadoop was the first Big Data platform (2006)
  - Distributed file system (HDFS)
  - Framework for parallel processing (MapReduce)
- Big data addresses data that are too complex (large) for traditional data-processing software
  - Volume, Variety, Velocity
- Works for embarrassingly parallel computation, which a problem is easily separated into parallel tasks