DS256:Jan17 (3:1)

L1:Introduction

Scalable Systems for Data Science

aka "Big Data Platforms"

Yogesh Simmhan

Plan

1. What is Big Data, and where does it come from?
2. Why is Big Data Challenging, and how can scalable systems help?
3. What is this course about?

Big Data & Why is it Important

What is Big Data?

Image credits: http://www.seekbig.in/1128-tnpsc-economics-questions/

The term ***is*** fuzzy … *Handle with care!*

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Data Generation View

"Big data refers to the approach to data of 'collect now, sort out later'…The low cost of storage and better methods of analysis mean that you generally don't need to have a specific purpose for the data in mind before you collect it."

Rohan Deuskar, CEO and Co-Founder, Stylitics

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: https://community.uservoice.com/wp-content/uploads/benefits-of-effective-questions-800x448-300x168.jpg

Data Systems View

"Big data is when your business wants to use data to solve a problem, answer a question, produce a product, etc., but the standard, simple methods break down on the size of the data set, causing time, effort, creativity, and money to be spent crafting a solution to the problem that leverages the data without simply sampling or tossing out records."

John Foreman, Chief Data Scientist, MailChimp

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: http://heae.tk/question/

Data Analysis View

"While the use of the term is quite nebulous … I've understood "big data" to be about analysis for data that's really messy or where you don't know the right questions or queries to make — analysis that can help you find patterns, anomalies, or new structures amidst otherwise chaotic or complex data points."

Philip Ashlock, Chief Architect, Data.gov

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits:http://www.clipartpanda.com/categories/question-and-answer-clipart

Philosophical View

"Big data, which started as a technological innovation in distributed computing, is now a cultural movement by which we continue to discover how humanity interacts with the world — and each other — at large-scale."

Drew Conway, Head of Data, Project Florida

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: http://en.hdyo.org/fri/questions

So…What is Big Data?

Data whose characteristics exceeds the capabilities of conventional *algorithms, **systems** and techniques* to derive useful value.

https://www.oreilly.com/ideas/what-is-big-data

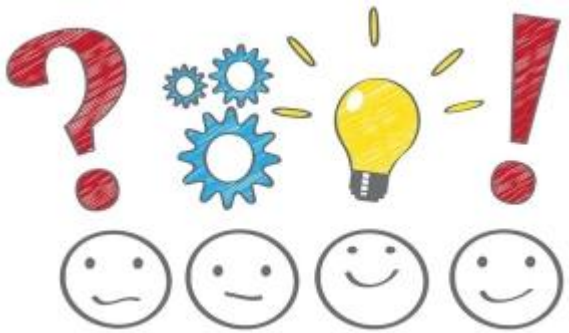
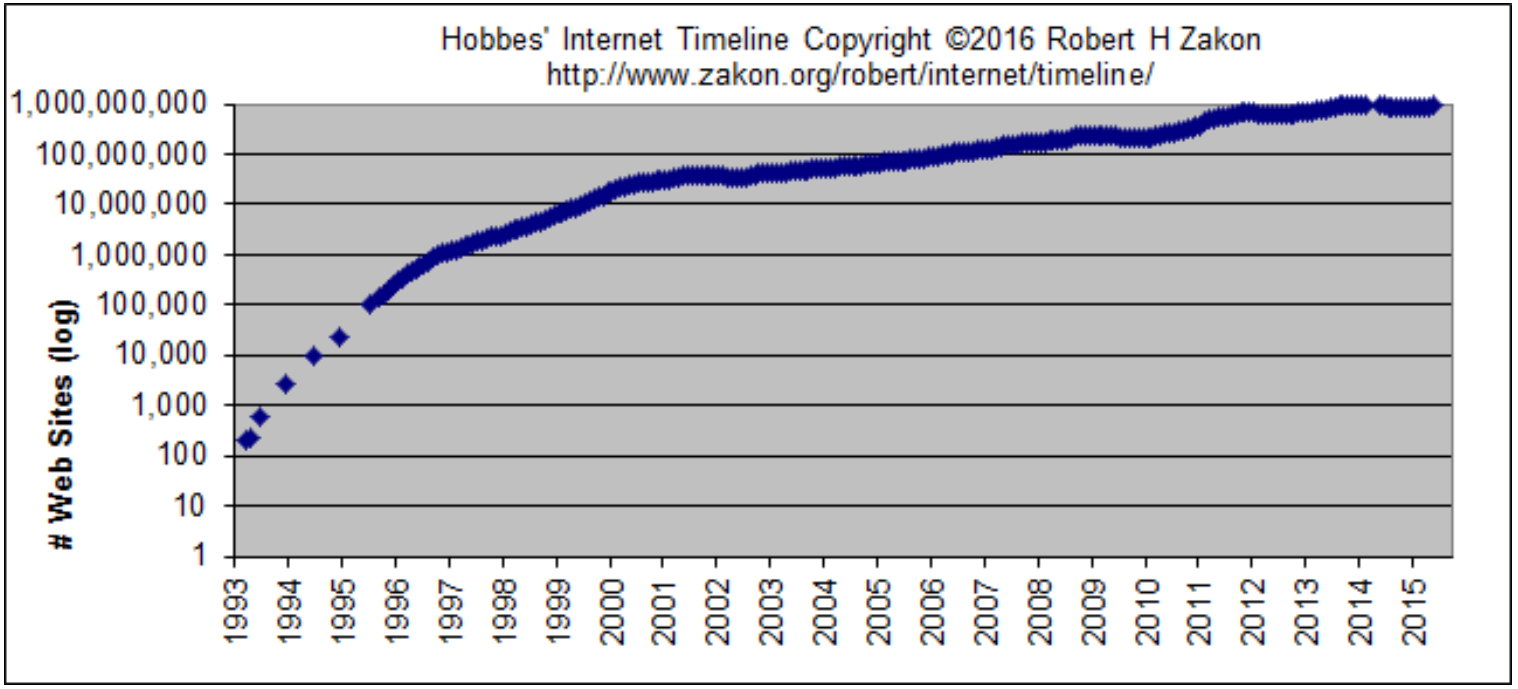


Image Credits: https://community.uservoice.com/wp-content/uploads/benefits-of-effective-questions-800x448-300x168.jpg

And, where does Big Data come from?

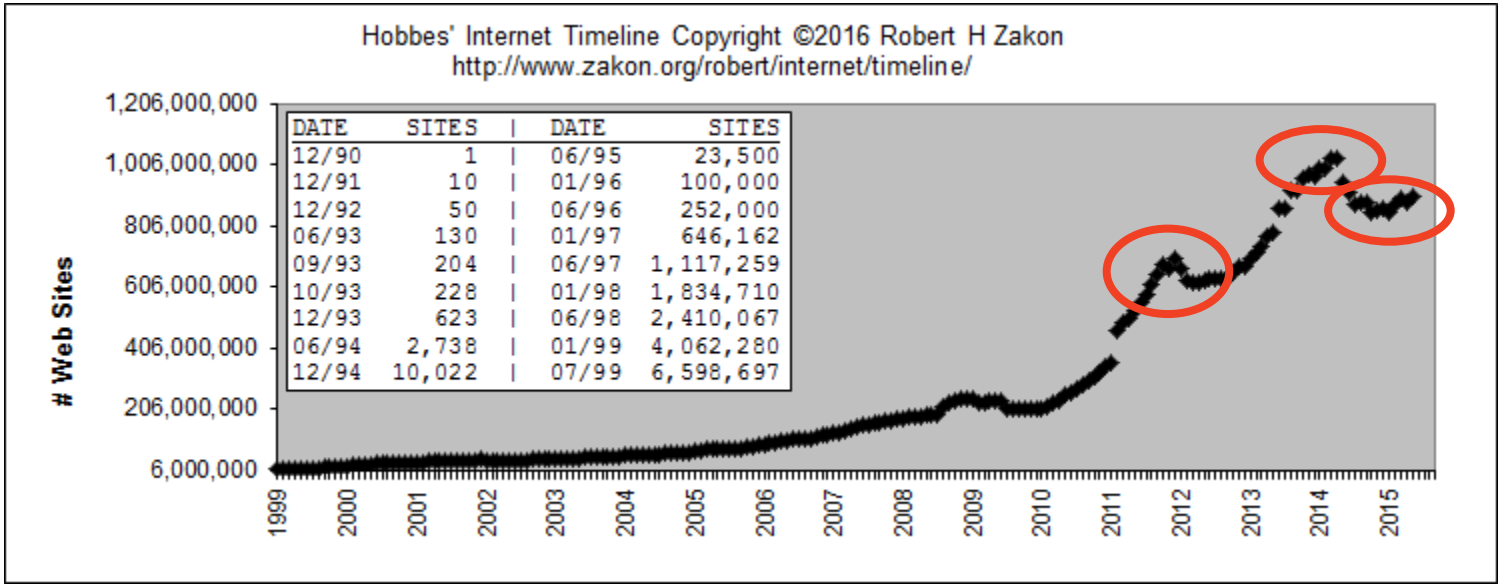
Web & Social Media

- World Wide Web & Search engines (log scale)



https://www.zakon.org/robert/internet/timeline/

Web & Social Media



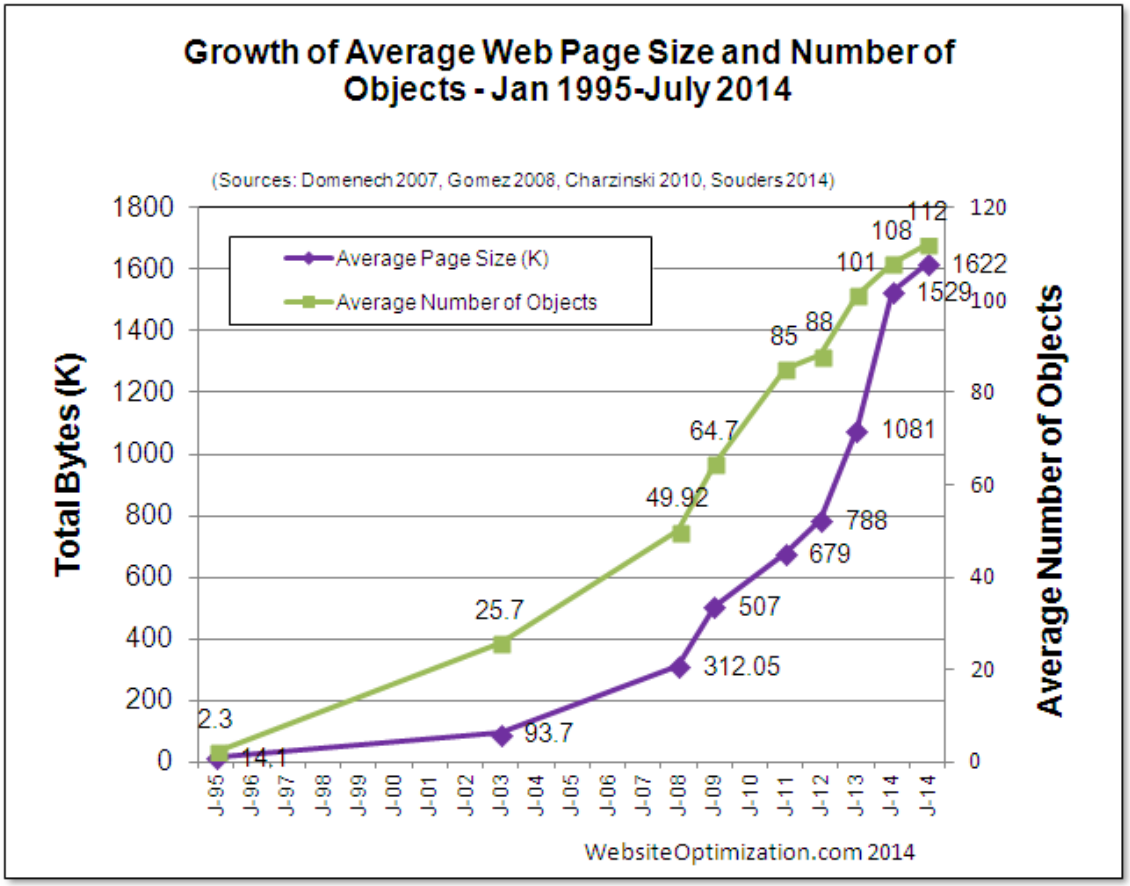
DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

- Constant “arms race” between Search Engine Optimizers and Search Providers
  - Link Farming (clique)
  - Domain Parking, Expiration

https://www.zakon.org/robert/internet/timeline/

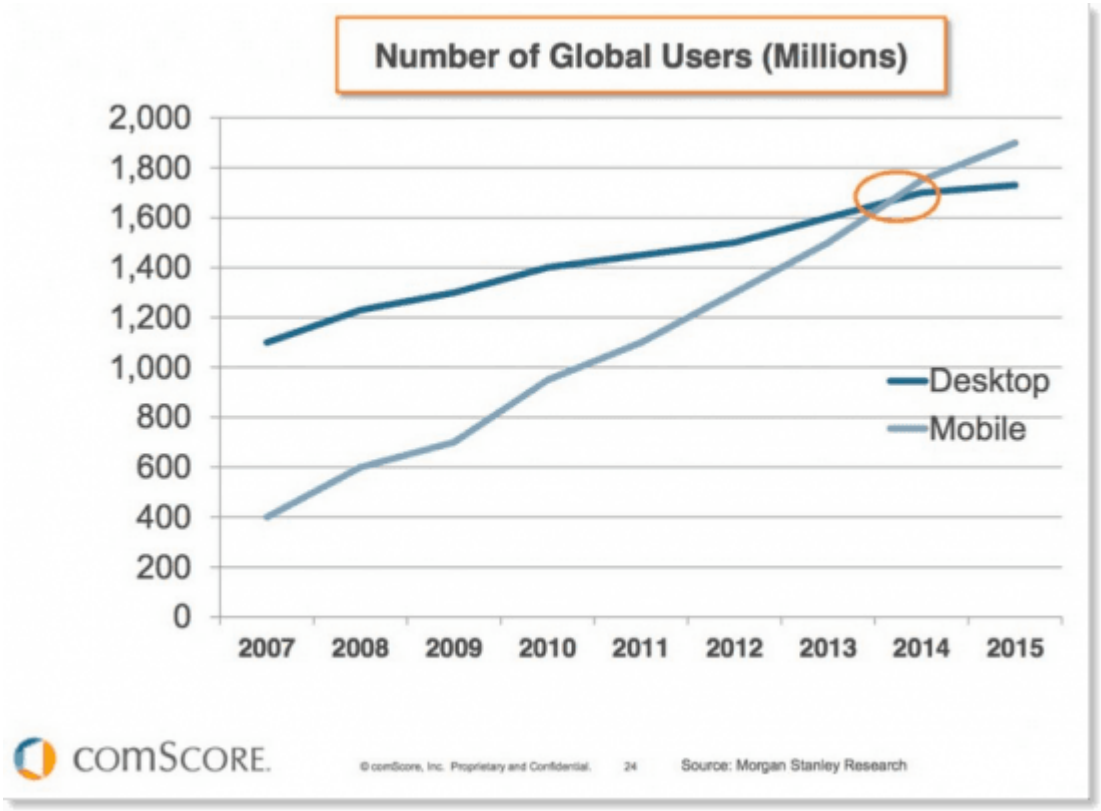
Web & Social Media

- World Wide Web & Search engines



http://www.yottaa.com/company/blog/application-optimization/a-brief-history-of-web-page-size/

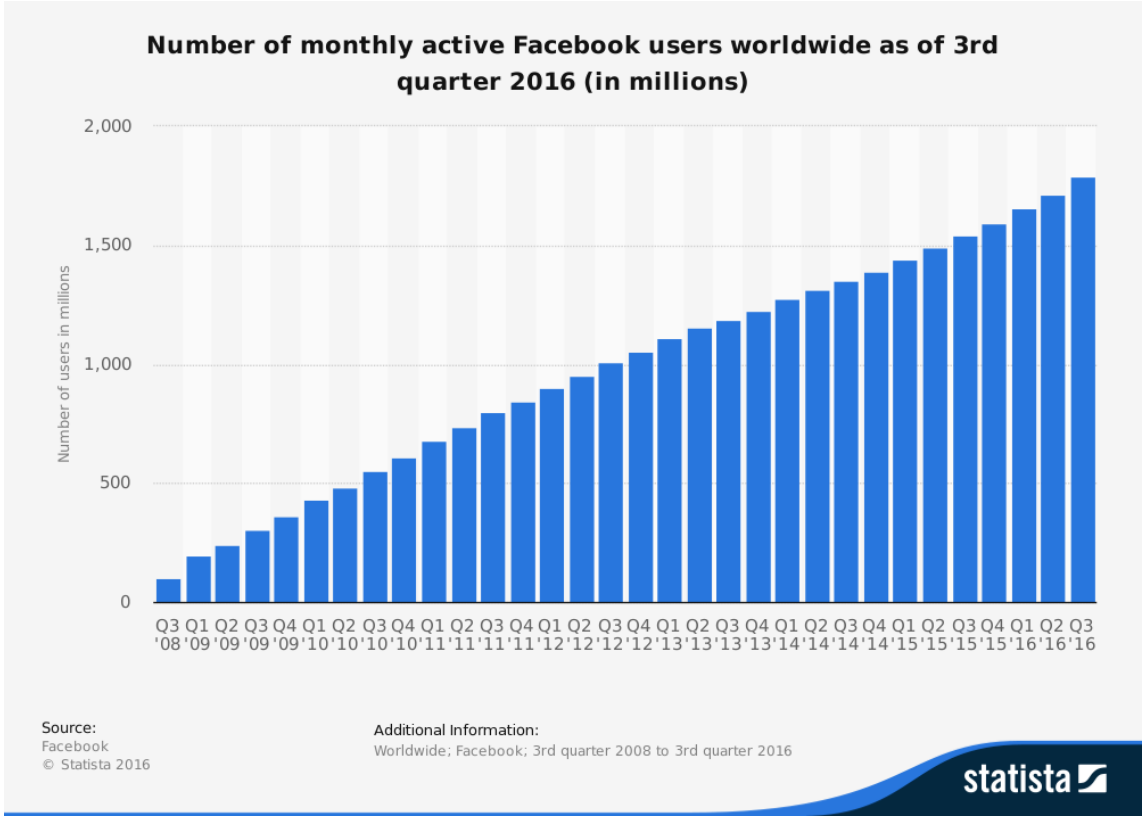
Web & Social Media

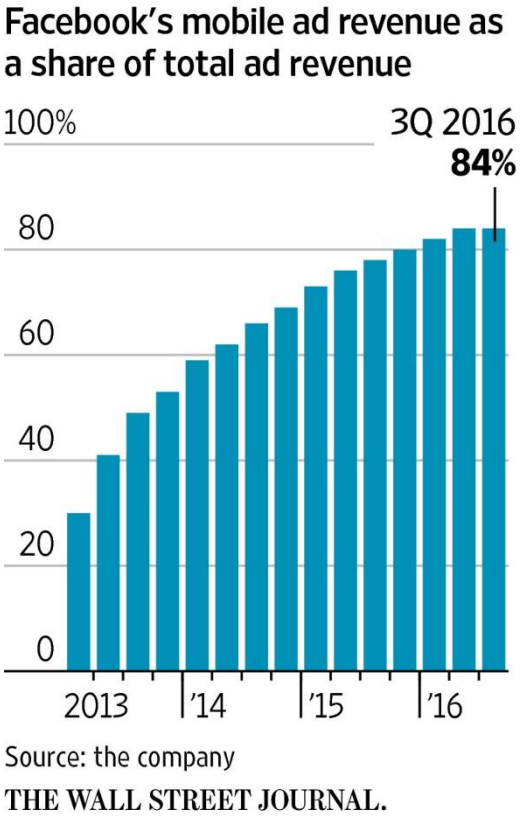


http://www.yottaa.com/a-brief-history-of-web-page-size/

Web & Social Media

- Social Networks & Micro-blogs



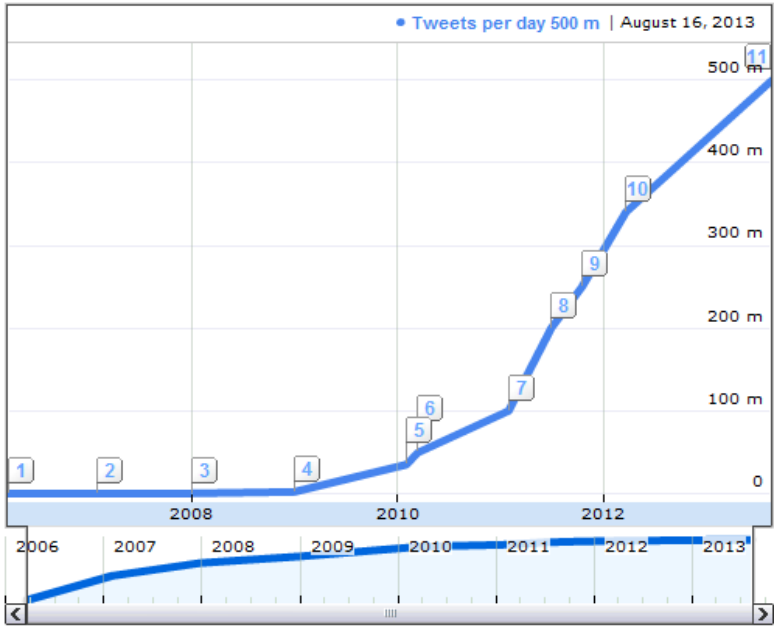


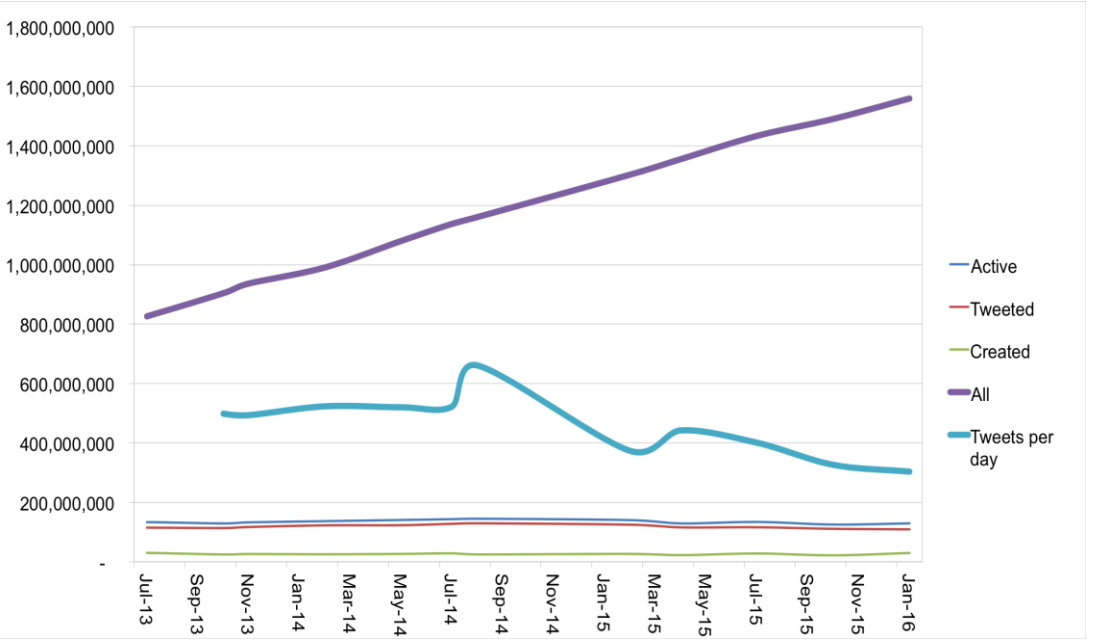
1.79 billion monthly active users as of September 30, 2016

https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
http://www.wsj.com/articles/facebook-profit-jumps-sharply-1478117646
http://newsroom.fb.com/company-info/

Web & Social Media

- Social Networks & Micro-blogs…the rise & plateau

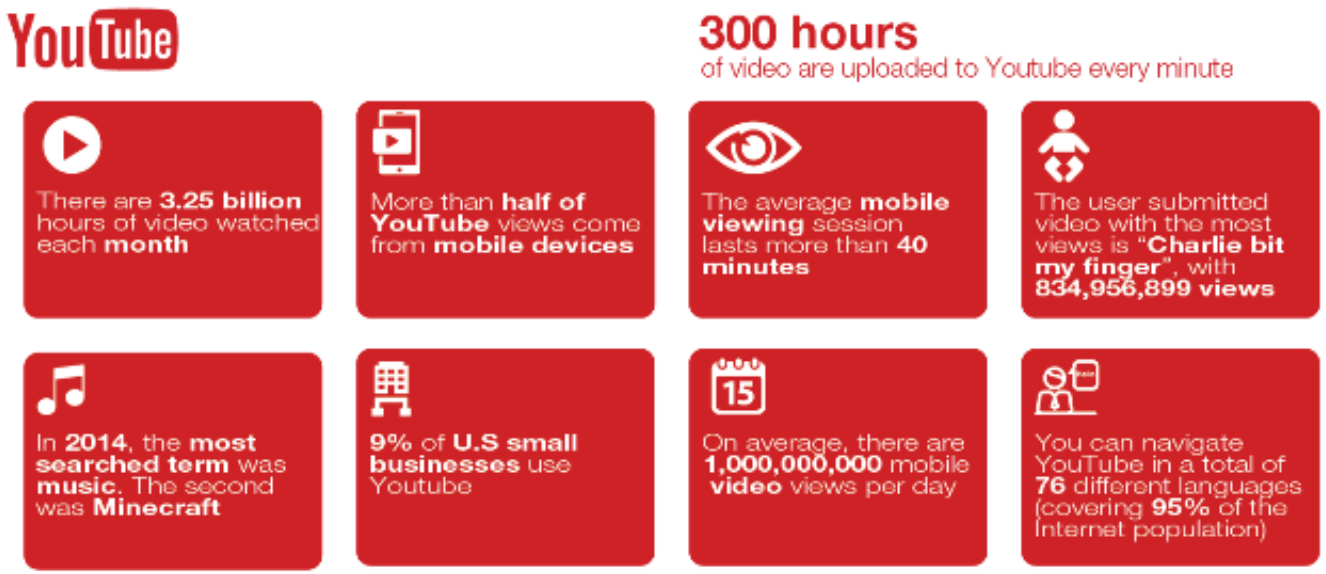




http://static4.businessinsider.com/image/56b089cedd0895437c8b45ef-2390-1265/untitled.png
http://www.internetlivestats.com/twitter-statistics/

Web & Social Media

- Youtube

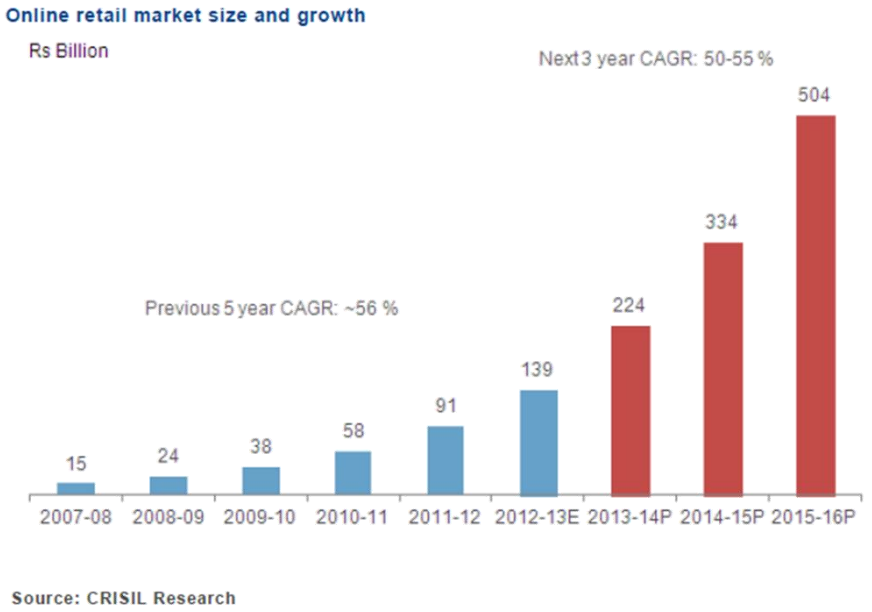


- Siri, Cortana, Now, Alexa, …

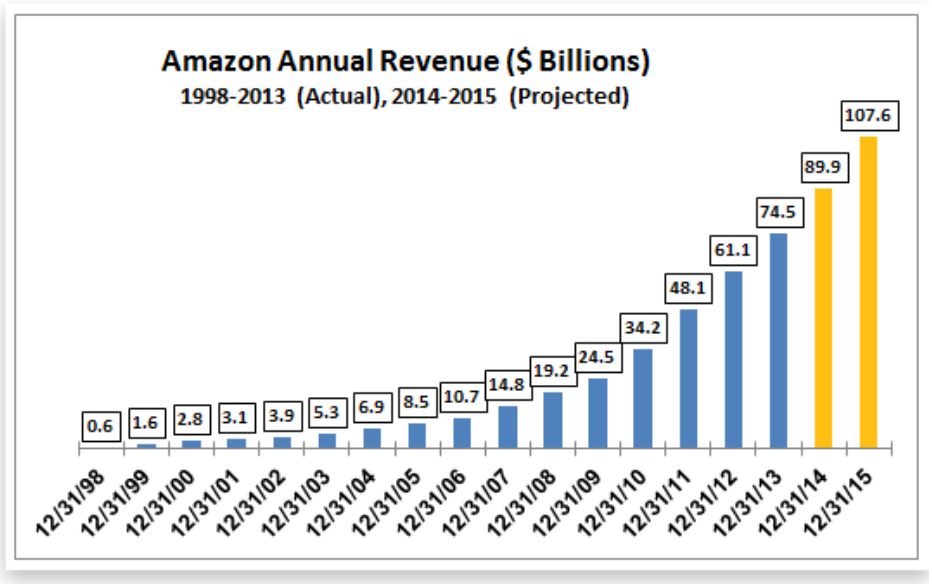
https://viralwoot.com/blog/social-media-stats-2016/

Enterprises & Government

- Online retail & eCommerce



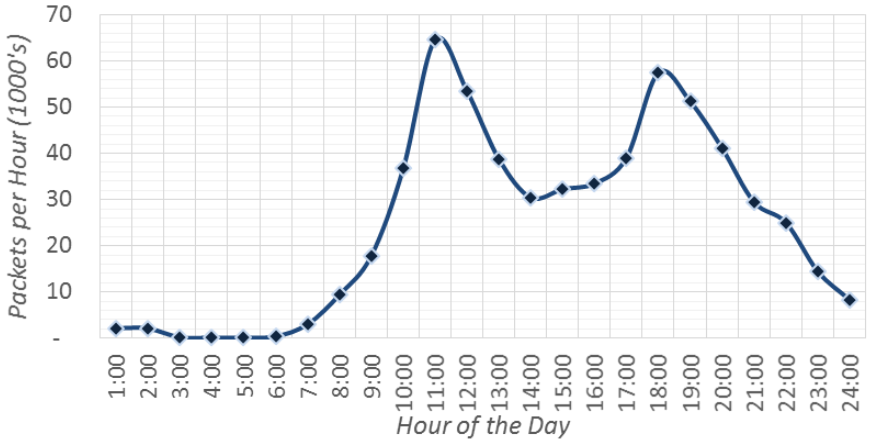
Source: CRISIL Research

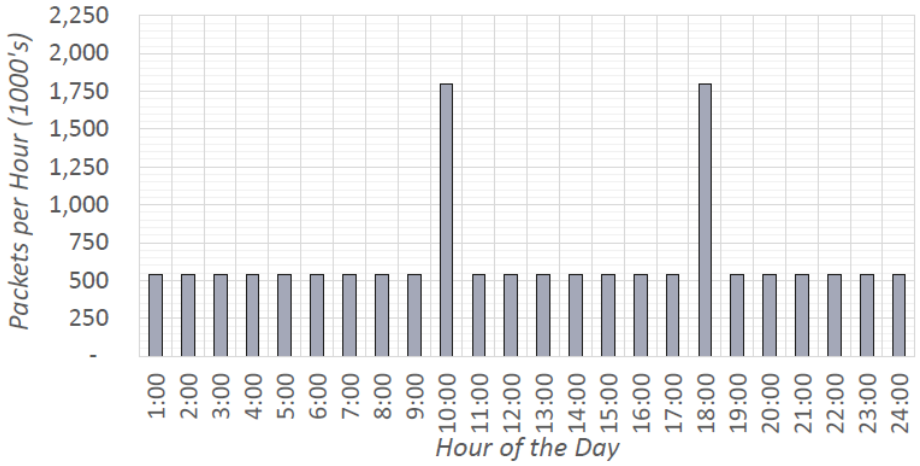


http://blogs.ft.com/beyond-brics/2014/02/28/online-retail-in-india-learning-to-evolve/

http://www.peridotcapital.com/2014/04/amazon-sales-growth-projections-for-next-two-years-appear-overly-optimistic.html

Enterprises & Government: UIDAI

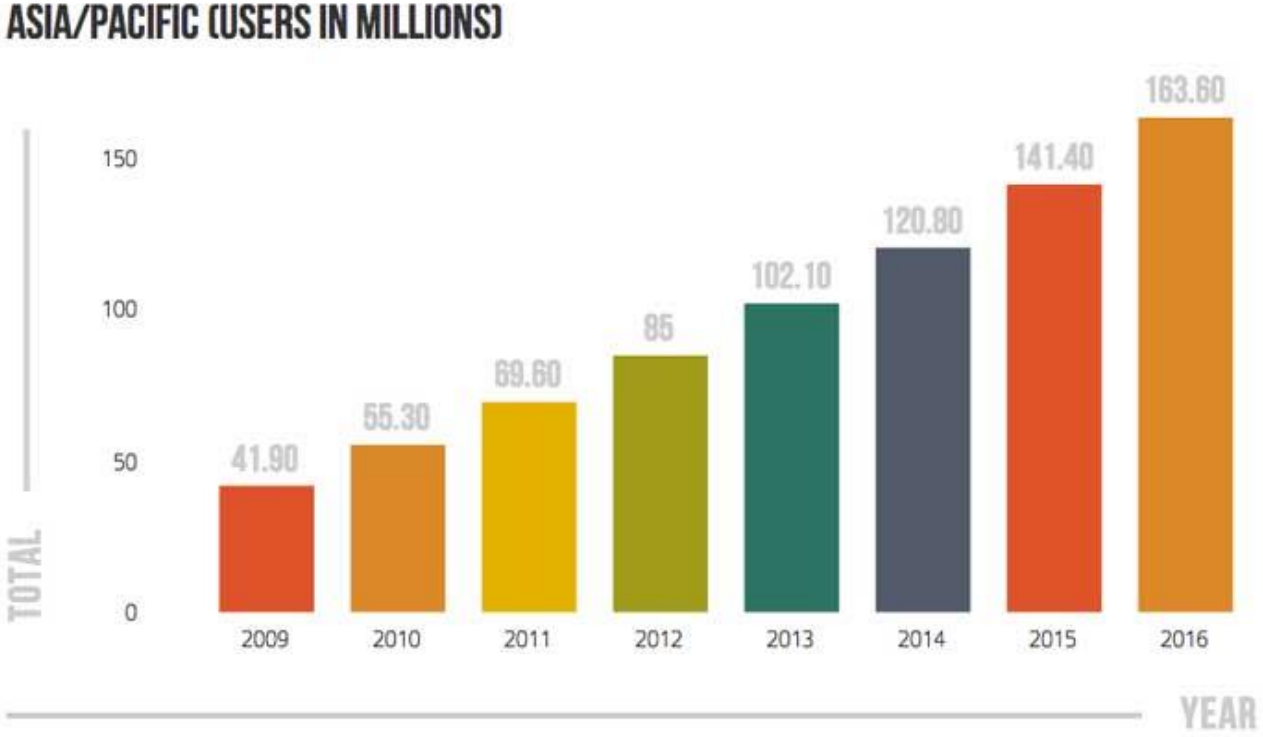




- Aadhaar Enrolment
  - Bi-modal rate distribution
  - Mean packet size is **~3MB**
  - **~600K/day** now. Peak was **1.3M/day** in 2013
- Soon >20 Millions of authentications per day

Enterprises & Government: Finance

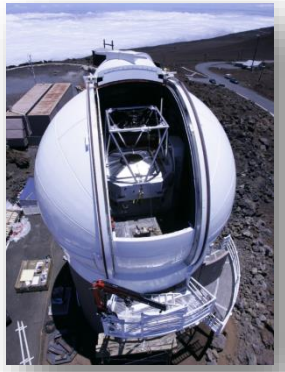
- Mobile Transactions & FinTech



Since November 8, 2016, *Paytm* has surpassed its metrics -tripling *transactions per day* to 7.5 million

http://www.pymnts.com/in-depth/2015/mobile-transactions/
Is Paytm the Xerox of mobile payments?, ETtech.com-03-Jan-2017

Astronomy Sky Survey: PanSTARRS, LSST, SKA, …



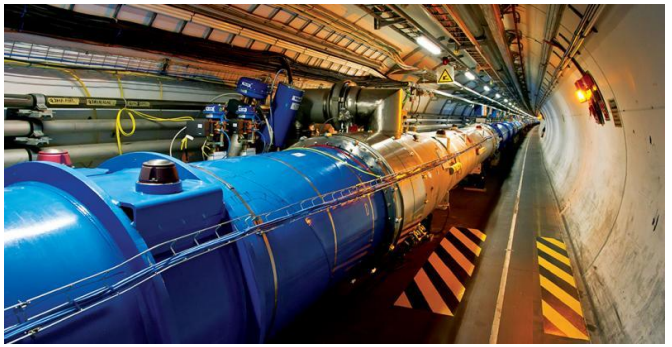
- Sloan Digital Sky Survey…first all digital survey
- PanSTARRS: Scan $^{2}/_{3}^{rds}$ of sky, **3** times/month, **1 PB** images, **30 TB** of data/year
- LSST: 10M events/night. 37B objects, 7T obs, 30T measurements per year*
- Square Km Array: Radio Telescope, 300 dishes
  - 68Tb/s to be sampled, 10PB images/day. 1PB catalog.

Next Gen Sequencers

- TB of data per run

Large Hadron Collider

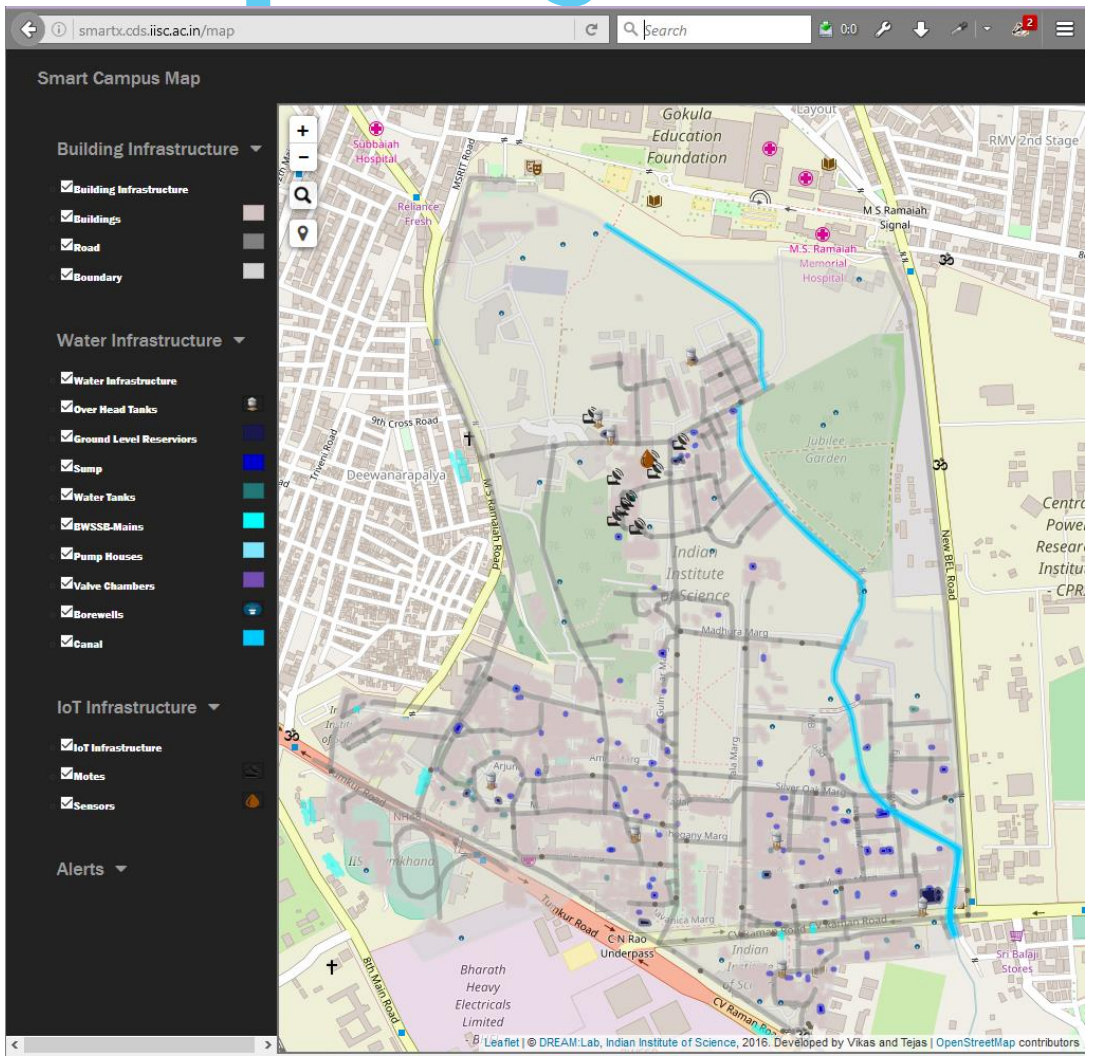
- Petabytes of data per run



*Introduction to LSST Data Management, Jeffrey Kantor
Square Kilometre Array Computational Challenges, Paul Alexander
www.ps1sc.org, home.cern

Internet of Everything

- Personal Devices
  - Smart Phones, Fitbit
- Smart Appliances
- Smart Cities
  - Power, Water, Transportation, Environment
- Smart Retail



Why is Big Data Difficult?

Evolving Nature of Science Data

Large Hadron Collider

PS1 telescope

Illumina NGS @ IISc

Bluetooth Mote @ IISc

Smart Meter @ LADWP

Few Instruments, Large Data **Volume**

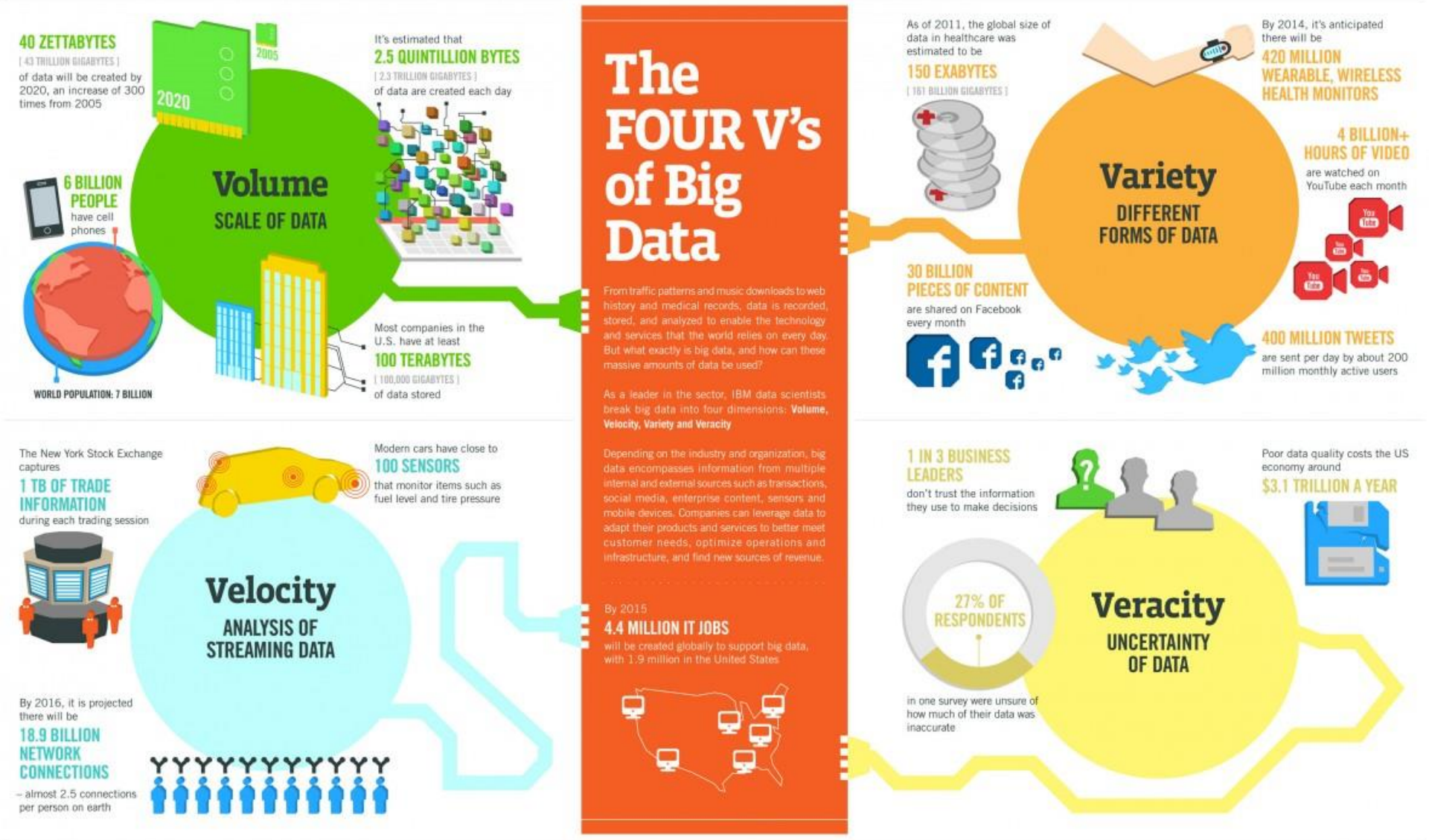
10^2 Sources
TB's Data
Days to Proc.

Many Devices, Volume & Velocity

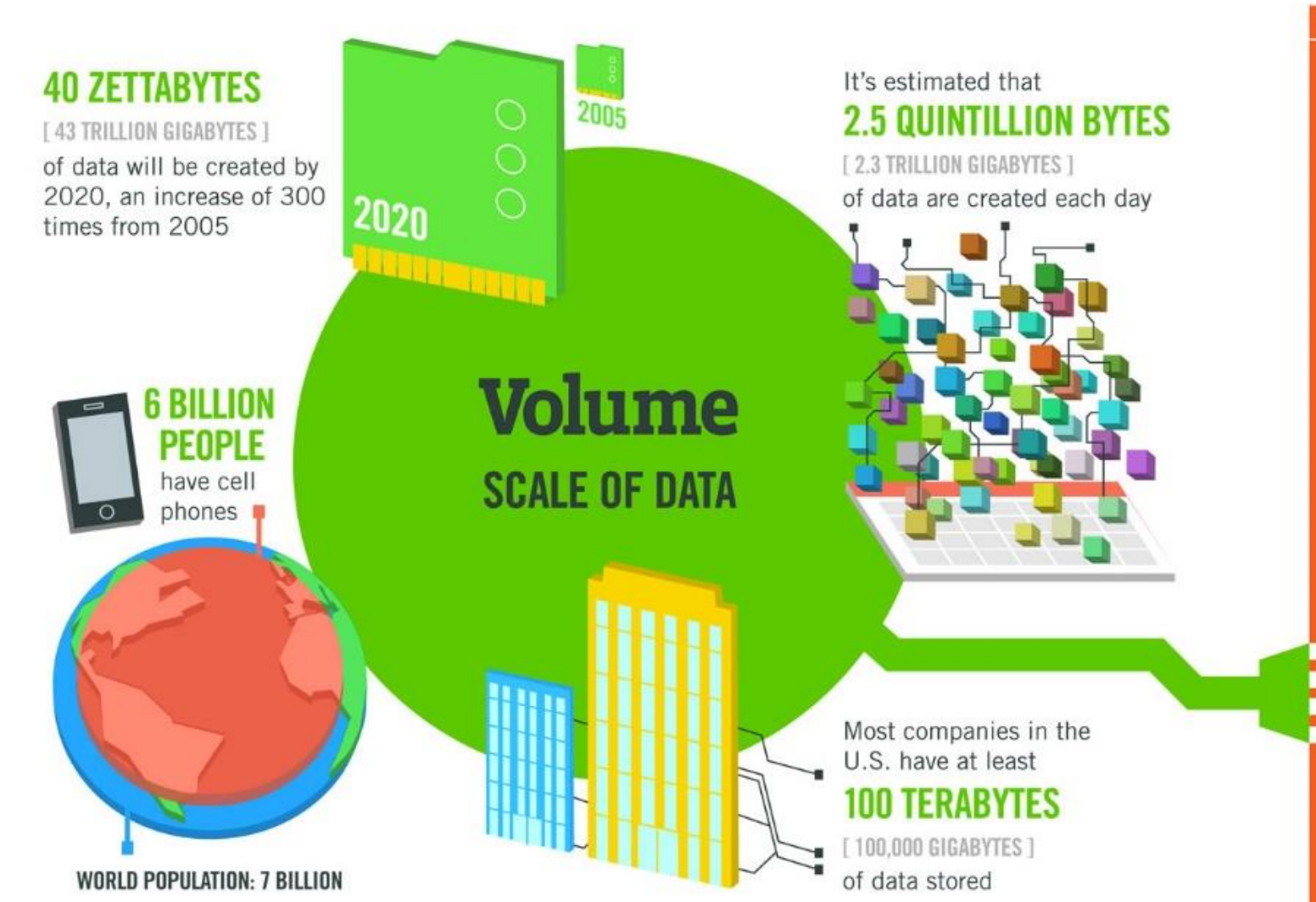
10^5 Sources
GB's Data
Hours to Proc.

Numerous Sensors, High data **Velocity**

10^8 Sources
MB's Data
<Mins to Proc.



http://www.ibmbigdatahub.com/infographic/four-vs-big-data



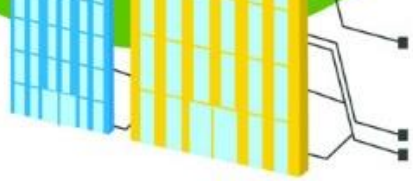
The FOUR of Big Data

From traffic patterns and m
history and medical reco
stored, and analyzed to e
and services that the wor
But what exactly is big da
massive amounts of data

As a leader in the sector
break big data into four
Velocity, Variety and Veraci

Depending on the industry
data encompasses infor
internal and external sourc

WORLD POPULATION: 7 BILLION



U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

and services that the w
But what exactly is big
massive amounts of da

As a leader in the sec
break big data into fo
Velocity, Variety and Ver

Depending on the indus
data encompasses inf
internal and external sou
social media, enterpris
mobile devices. Compa
adapt their products an
customer needs, opt
infrastructure, and find

By 2015

4.4 MILLION IT JO

will be created globally
with 1.9 million in the

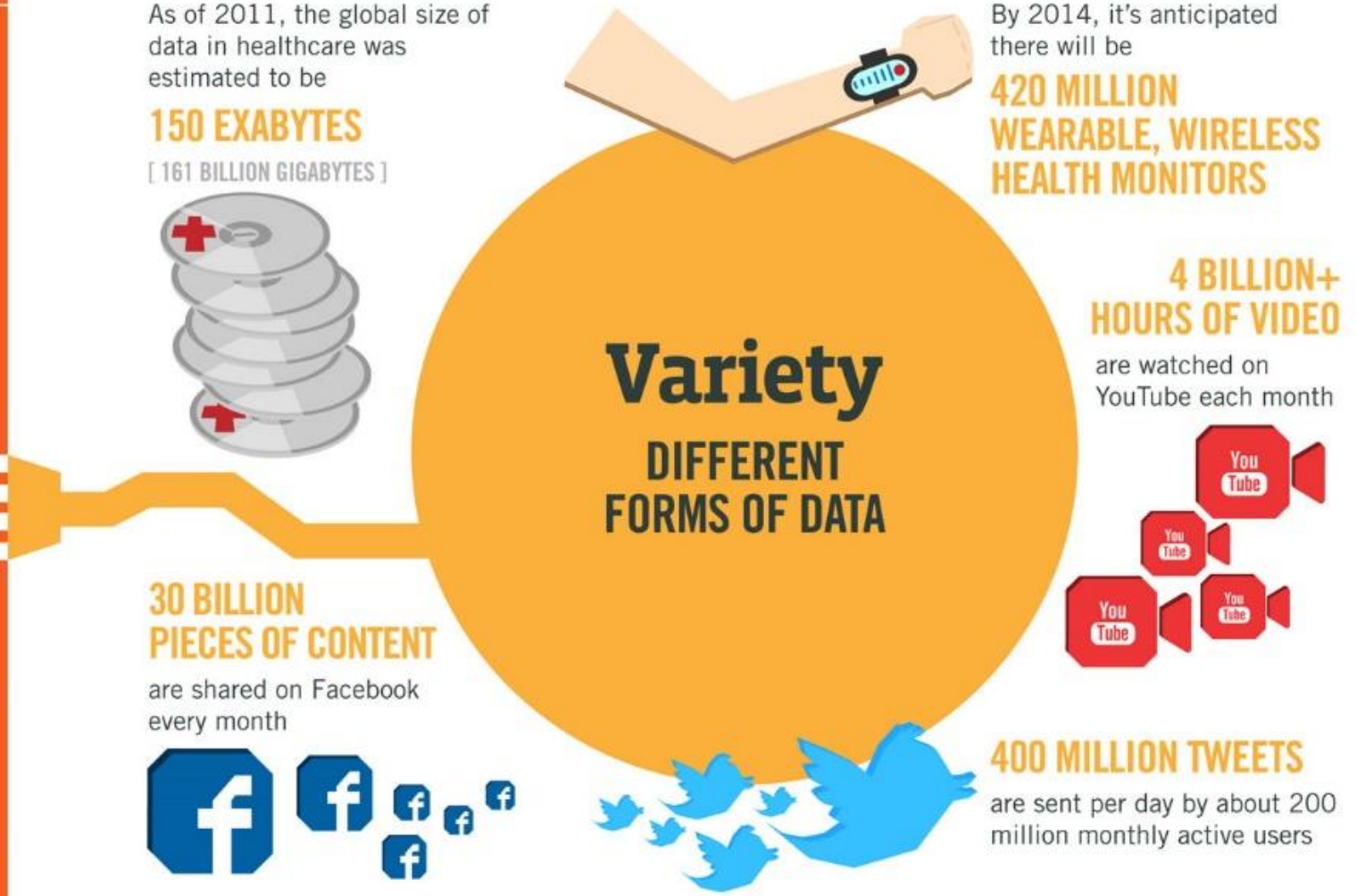
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

R V's ig l

d music downloads to web ecords, data is recorded, to enable the technology world relies on every day. data, and how can these ata be used?

ctor, IBM data scientists our dimensions: **Volume, racity**

stry and organization, big formation from multiple urces such as transactions,



rld relies on every day.
ata, and how can these
be used?

r, IBM data scientists
r dimensions: **Volume,**
ity

y and organization, big
rmation from multiple
ces such as transactions,
content, sensors and
es can leverage data to
services to better meet
mize operations and
ew sources of revenue.

S

o support big data,
Jnited States

every month

400 MILLION TWEETS

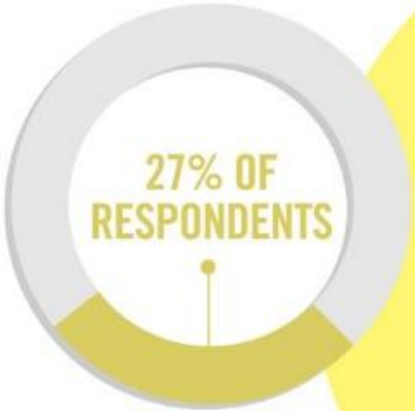
are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR



in one survey were unsure of how much of their data was inaccurate

Veracity

UNCERTAINTY OF DATA

Data Analysis Lifecycle

Acquire

- Acquire Data
  - Sensors, Web logs & crawls, Transactions

Goal

- Define Analytics
  - Trends, Clusters, Outliers, Classification

Process

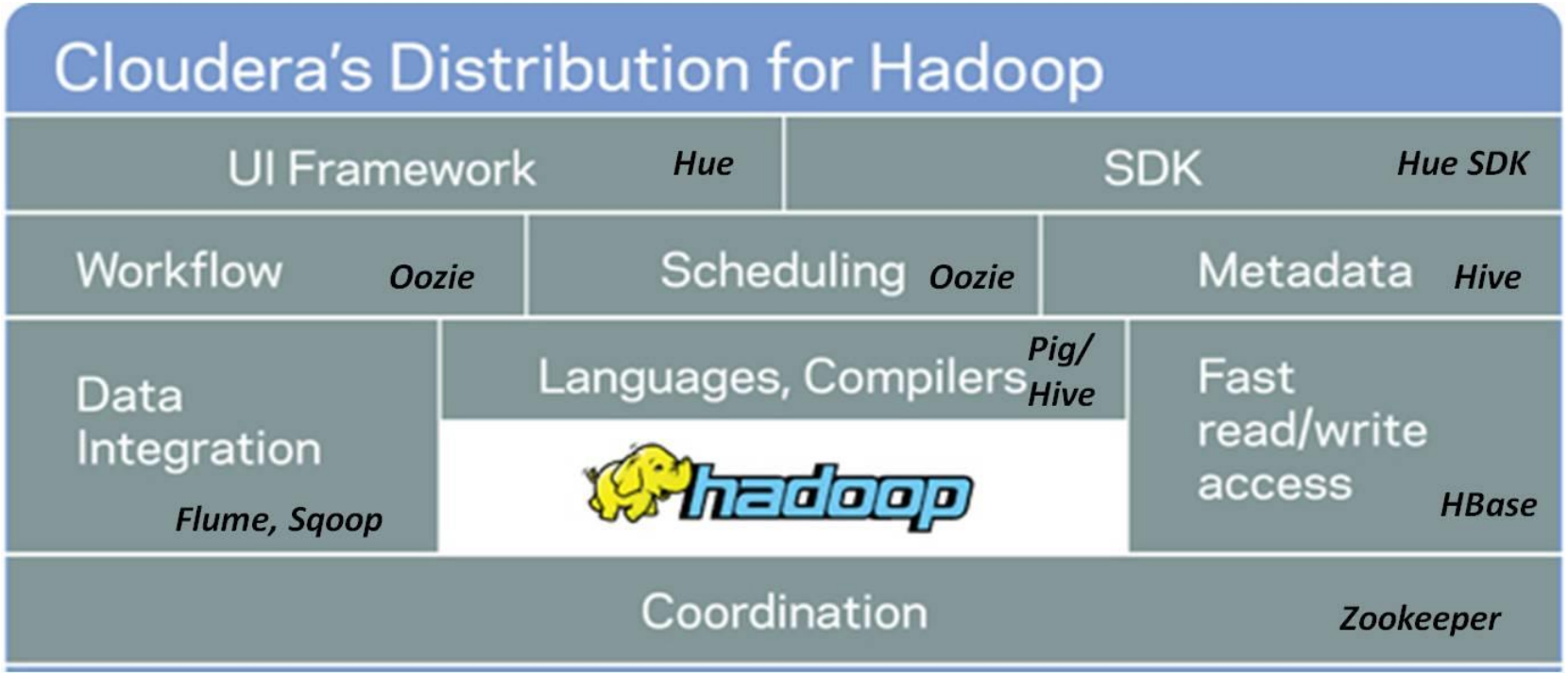
- Translate to Scalable Applications
  - Develop algorithms, Map to abstractions, Implement on Platforms

Data Platforms

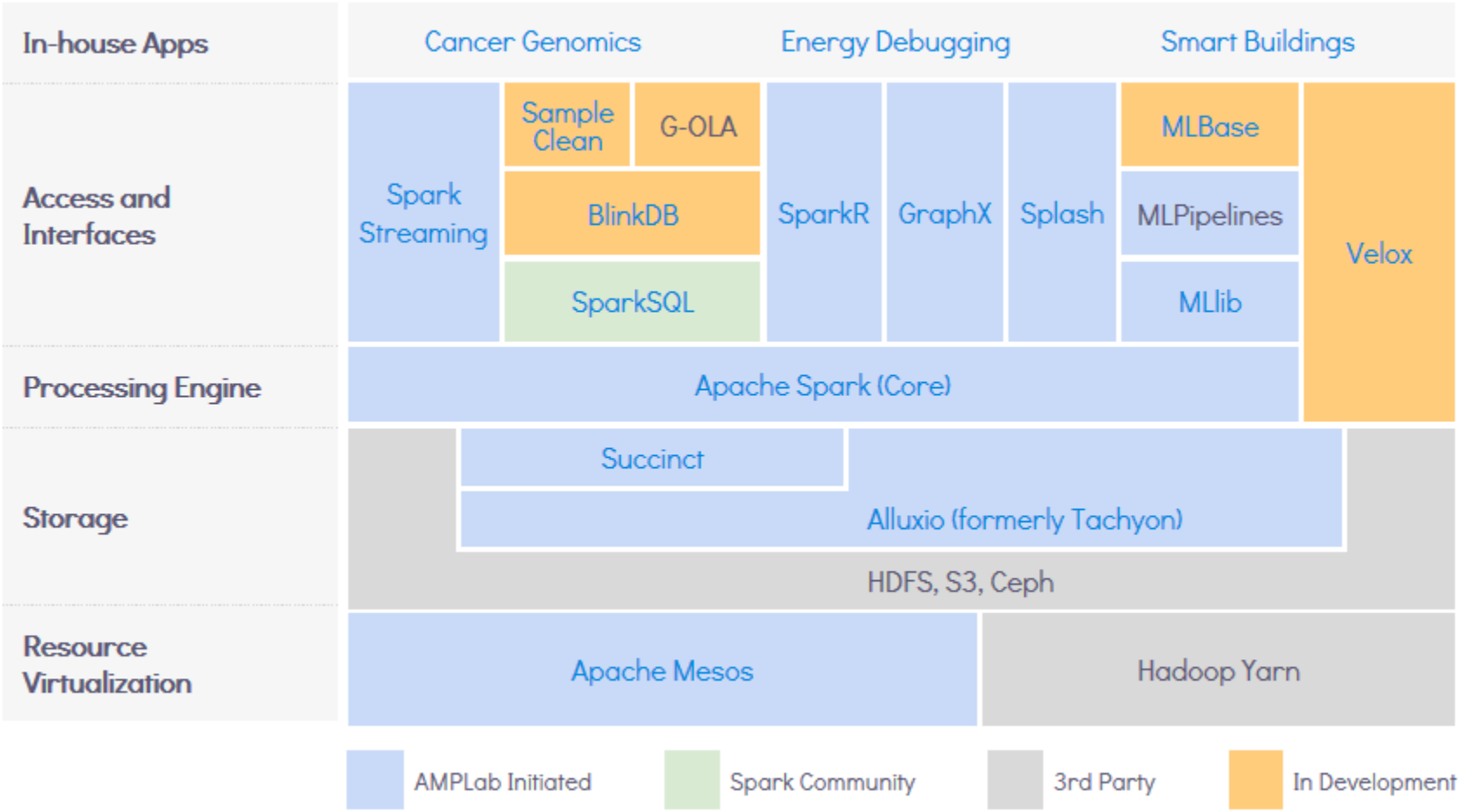
- Acquire, manage, process Big Data
- At large scales
- To meet application needs

- Think in terms of Stacks…

The Platform Stack: Cloudera

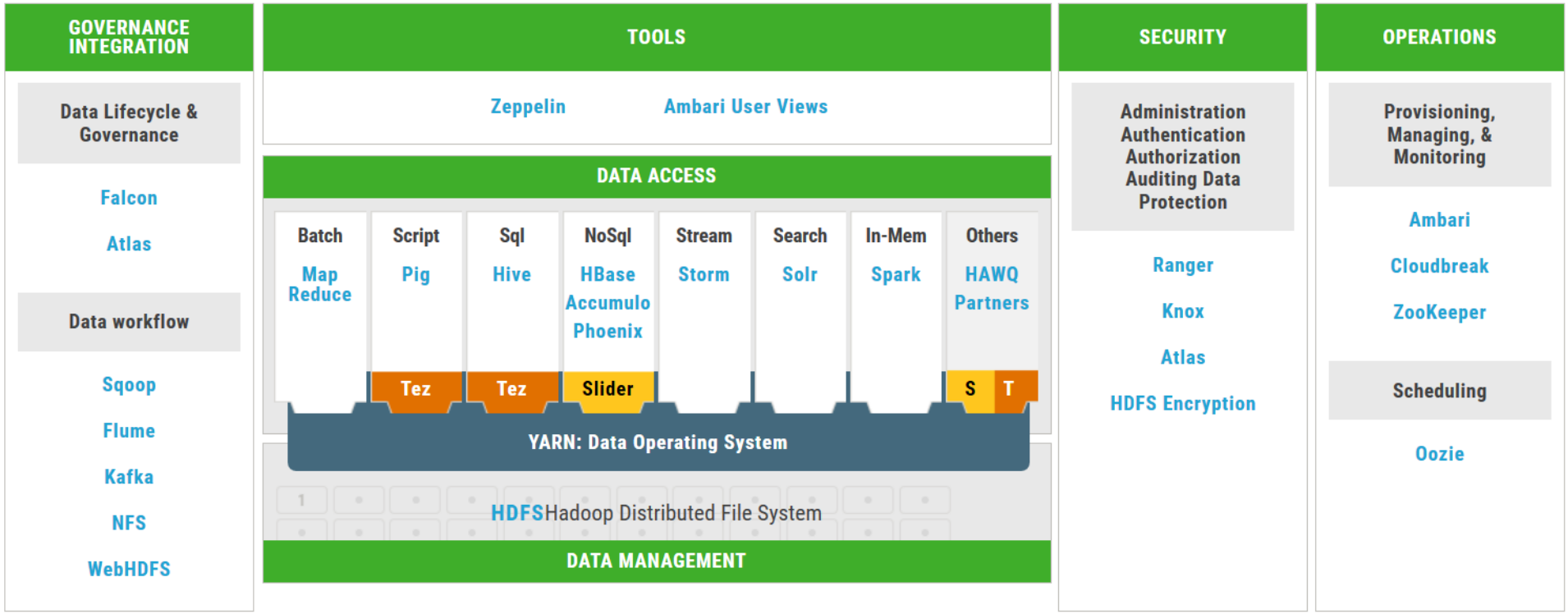


The Platform Stack: BDAS



https://amplab.cs.berkeley.edu/software/

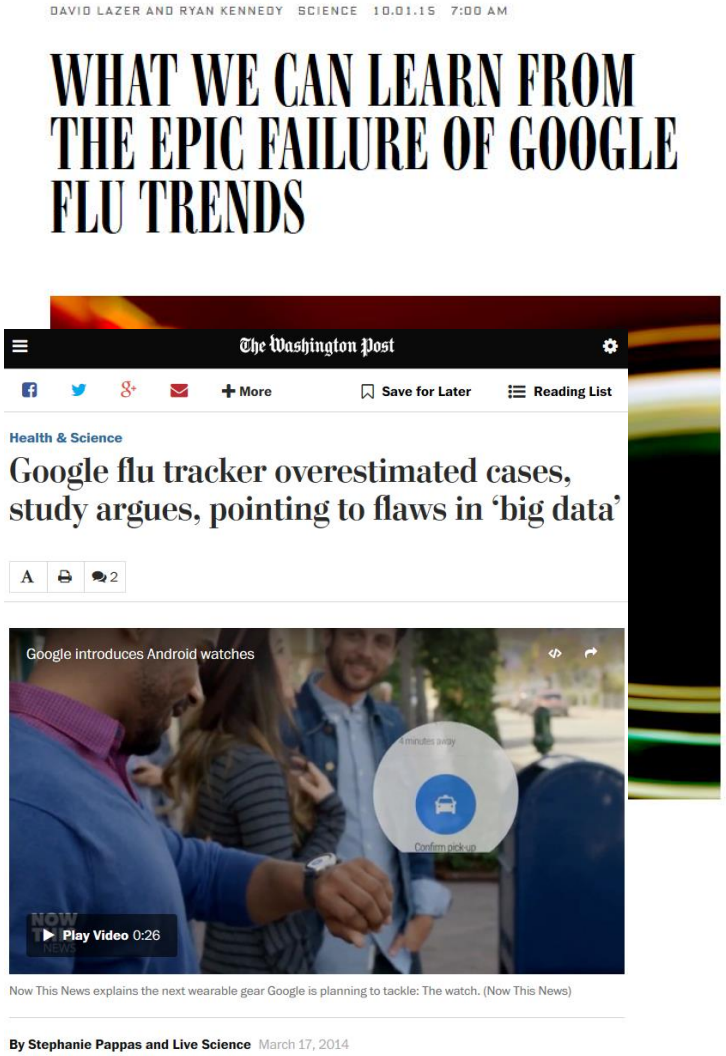
The Platform Stack: HortonWorks



http://hortonworks.com/products/data-center/hdp/

Big Data Analytics is not Perfect

- Google Flu Trends*
  - "When people get sick, they turn to the Web for information"
  - Estimate the start, peak, and duration of each flu season
  - Overestimated flu cases in 2012-2013
  - Missed a nonseasonal flu outbreak of H1N1



DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

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

The Washington Post

+ More Save for Later Reading List

Health & Science

Google flu tracker overestimated cases, study argues, pointing to flaws in 'big data'

Google introduces Android watches

Play Video 0:26

Now This News explains the next wearable gear Google is planning to tackle: The watch. (Now This News)

By Stephanie Pappas and Live Science March 17, 2014

*http://wapo.st/1d7TVLM
*http://blog.google.org/2013/10/flu-trends-updates-model-to-help.html
http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

People are not Perfect

- "Poor grasp of the technology can kill big data investments"*

COMPUTERWORLD UK Follow:

Home 〉 Features 〉 Data 〉 Hotels.com CTO on why big data projects fail

Why do big data projects fail?

Hotels.com CTO explains how it uses Cassandra and Hadoop for excellent customer service

Margi Murphy
June 29, 2015

Big data projects have a lot of promise, but the majority fail. A recent study found that just 11 percent of corporate leaders in the UK haven't generated any cash using data, despite recognising the value it could bring.

*http://www.computerworlduk.com/data/hotelscom-cto-on-why-big-data-project-fail-3617971/

Big Data Systems are not Perfect

Flipkart's Big Billion Days sale fails yet again

Varun Jain | ET Bureau | Oct 16, 2015, 11.43 AM IST

A- A+

According to experts, Flipkart has failed to cope with the huge amount of business the Big Billion Days is gen... Read More

NEW DELHI: Flipkart's 'Big Billion Days' sale has yet again run into glitches this year, and this time the e-commerce giant is tendering apologies to its sellers in Agra whose products have been removed from the site for almost two days now due to delivery issues. Various vendors in the city ET spoke to complained that Flipkart has blocked their products to be sold on its e-commerce site amid a rush of consumers.

About This Course

What is this course about?

- Fundamental systems aspects of big data platforms,
- How these platforms can be used to build large-scale data intensive applications.
- *Why Big Data platforms are necessary?*
- *How they are designed? What are the programming abstractions (e.g. MapReduce) that are used to compose data science applications?*
- *How the programming models are translated to scalable runtime execution on clusters and Clouds (e.g. Hadoop)?*
- *How do you translate algorithms for analyzing large datasets into scalable programs?*

What is this course about?

- Platforms used for developing applications over large columnar/tuple based data,
  - Map Reduce/Apache Hadoop,
  - Streaming data like Apache Storm and Spark Streaming,
  - Graph/linked data like Apache Giraph
- Declarative Tools
  - NoSQL databases, TensorFlow.
- *Why do platforms scale, and how can you use them to scale data science applications?*

Pre-requisites

- String Data Structures, Programming and Algorithm concepts
- Good programming skills required, preferably in Java
  - Large code bases of Big Data platforms
  - Debugging distributed systems

Other Data Science Courses

- Foundations of Data Science
- Data Analytics
- Data Analysis and Visualization

- Introduction to Cloud Computing, Parallel Programming
- Topics in Web-scale Knowledge Harvesting, Data Mining, Deep Learning, Machine Learning, Artificial Intelligence, Probability & Statistics
- Bioinformatics, Chemoinformatics, Video Analytics

Course Metadata

- Webpage: http://cds.iisc.ac.in/courses/ds256/
- Course number: DS256
- Credits: 3:1 *(NOT 2:1)*
- Semester: Jan, 2017
- Lecture: Tue/Thu 330-5PM
- Room: CDS 202
- Instructor: Yogesh Simmhan
- Teaching Assistant: Jayanth Kalyanasundaram
- Mailing List: mailman.serc.iisc.in/mailman/listinfo/ds256.jan17
- Registration Page: coursereg.iisc.ac.in

Lecture	Date	Topic	Task
1	2017-01-05	Introduction to Course. Data, Platforms, Applications. Big Data Stacks.	Assignment 0 Posted
2	2017-01-10	MapReduce: Model	
3	2017-01-12	MapReduce: Basic Algos	
4	2017-01-17	MapReduce: Advanced Algos	Assignment 1 Posted
5	2017-01-19	Tutorial: Hadoop, Turing Cluster	
6	2017-01-24	Invited Talk: MSR Faculty Summit -- Data Science Track	
7	2017-01-28	Hadoop/HDFS	(In lieu of Republic Day)
8	2017-01-31	Hadoop/HDFS	
9	2017-02-02	DSPS	
10	2017-02-07	DSPS	Assignment 1 Due. Assignment 2 Posted
11	2017-02-09	Apache Storm	
*	2017-02-11	*	Assignment 1 Evals
12	2017-02-14	Tutorial: Apache Storm	
13	2017-02-16	Invited Talk: Hortonworks (Storm, Big Picture)	
14	2017-02-21	Invited Talk: Cloud Elasticity	

Lecture	Date	Topic	Task
15	2017-02-23	Pregel: Model	
16	2017-02-28	Pregel: Algorithms	Assignment 2 Due. Assignment 3 Posted.
17	2017-03-02	Pregel: Algorithms, Giraph	Projects Posted.
*	2017-03-04	*	Assignment 2 Evals
18	2017-03-07	Tutorial: Apache Giraph	
19	2017-03-09	Invited Talk: Subgraph Centric	Projects Selected.
20	2017-02-14	Invited Talk: Microsoft (Azure Big Data PaaS)	
21	2017-03-16	Spark, BDAS	
22	2017-03-21	Spark, BDAS	Assignment 3 Due.
23	2017-03-23	Invited Talk: HPE (Spark, etc.)	
*	2017-03-25	*	Assignment 3 Evals
24	2017-03-28	NoSQL HBase/Hive/Cassandra	
25	2017-03-30	NoSQL HBase/Hive/Cassandra	
26	2017-04-04	Invited Talk: Hortonworks (NoSQL)	
27	2017-04-06	Tools: GraphDB, ML, CEP, TensorFlow, Viz	
28	2017-04-11	Invited Talk: Google (ML?)	
29	2017-04-13	Invited Talk: ???	Project Due.
30	2017-04-15	*	Project Demos
*	28-04-17	*	Finals

Assignments & Projects

45% Homework	Three programming assignments *(150 points each)*
30% Project	One final project, to be done individually or in teams *(300 points)*
20% Exams	One Final exam *(200 points)*
5% Participation	Participation (i.e. not just “attendance”) in classroom discussions and online forum for the course *(50 points)*

Text Book

- Select chapters from **Data-Intensive Text Processing with MapReduce**, *Jimmy Lin and Chris Dyer*, 1st Edition, Morgan & Claypool Publishers, 2010
- Select chapters from **Mining of Massive Datasets**, *Jure Leskovec, Anand Rajaraman and Jeff Ullman*, 2nd Edition (v2.1), 2014.
- Current literature and online documentation

Academic Integrity

- Students must uphold IISc's Academic Integrity guidelines. While these are common sense, it is helpful to review them since failure to follow them will lead to sanctions and penalties.
- **This includes a reduced or failing grade in the course. Severe cases of academic violations will be reported to the Institute and may lead to an expulsion.**

http://www.iisc.ernet.in/students-corner/existingstudents-academicintegrity.php

Academic Integrity

- Learning takes place both within and outside the class. Hence, **discussions** between students and reference to **online material** is **encouraged** as part of the course to achieve the intended learning objectives.
- **However**, while you may learn from any valid source, you **must form your own ideas** and **complete problems and assignments by yourself**.
- *All works submitted by the student as part of their academic assessment must be their own!*

Academic Integrity

- **Plagiarism**: *Verbatim* reproduction of material from external sources (web pages, books, papers, etc.) is **not acceptable**.
- If you are *paraphrasing* external content (or even your own prior work) or were otherwise *influenced* by them while completing your assignments, projects or exams, **you must clearly acknowledge them**.
- *When in doubt, add a citation!*

Academic Integrity

- **Cheating**: While *you may discuss* lecture topics and broad outlines of homework problems and projects with others, **you cannot collaborate** in completing the assignments, **copy** someone else's solution or **falsify results**.
- You **cannot use** notes or unauthorized resources during exams, or copy from others.
- The narrow exception to collaboration is between team-mates when competing the project, and even there, *the contribution of each team member* for each project assignment should be **clearly documented**.

Academic Integrity

- **Classroom Behaviour**: Ensure that the course atmosphere, both in the class, outside and on the online forum, is conducive for learning.
- *Participate* in discussions but *do not dominate* or be abusive. **There are no "stupid" questions** ☺ Be considerate of your fellow students and avoid disruptive behaviour.

Thank You