CS226
Big-Data Management

Instructor:

Ahmed Eldawy

CELEBRATING 30 YEARS
UCR Marlan and Rosemary Bourns
College of Engineering

Welcome (back) to UCR!

Class information

- Classes: MWF 10:00 – 10:50 AM
- Zoom Link: on iLearn. Do not distribute.
- Instructor: Ahmed Eldawy
- TA: Akil Sevim
- Office hours: MW 11:00 – 11:50 AM
- Website: http://www.cs.ucr.edu/~eldawy/20FCS226/
  - iLearn (Any UCRX students?)
- Piazza: https://piazza.com/ucr/fall2020/cs226
- Email: eldawy@ucr.edu
  - Subject: “[CS226]”

Course work

- Active participation in class (5%)
- Reading and review tasks (10%)
- Assignments (20%)
- Mid-terms (15%)
- Project (50%)

Project

- Groups of 4-5 students
- Milestones
  - Group Selection
  - Project proposal (5%)
  - Project proposal presentation (10%)
  - Literature survey (10%)
  - Report outline (5%)
  - Final report (10%)
  - Final presentation (10%)

Project: Coordination between CS 226 and CS 225

- You can share project groups between CS 226 (Big Data Management) and CS 225 (Spatial Computing), given:
  - You work on one project with double in size, this gives opportunity to focus on one big project for the two courses
  - All group members must be taking the two courses, except one member at maximum.
  - The project must have a spatial component and a big data component
  - The team must submit two separate reports, one for each course. Each report must focus on the component relevant to the course.

- Example projects:
  - **City ranking**: incorporate spatial factors in ranking cities for quality of live
  - **Satellite imagery analysis**: use large satellite images datasets in any societal application
  - **Contact tracing for epidemics control**: use spatial data to trace contacts to patients of COVID-like epidemics

R'GEOSPATIAL

WHAT'S GIS?

GIS stands for Geographic Information system. The software is used for both gathering and visualizing data. The most common example that is applicable to our lives is Google maps. Whenever you look up directions or try to figure out where you are, that's the byproduct of GIS.

GOAL OF THE CLUB:

R'geospatial is a relatively new club on campus that aims to show the utility and transferability of GIS skills to students' careers. The club will cover how to use the software and also how it applies to your major and future professional endeavors.

If interested, please add us on Highlanderlink and feel free to reach out to us if you have any questions!

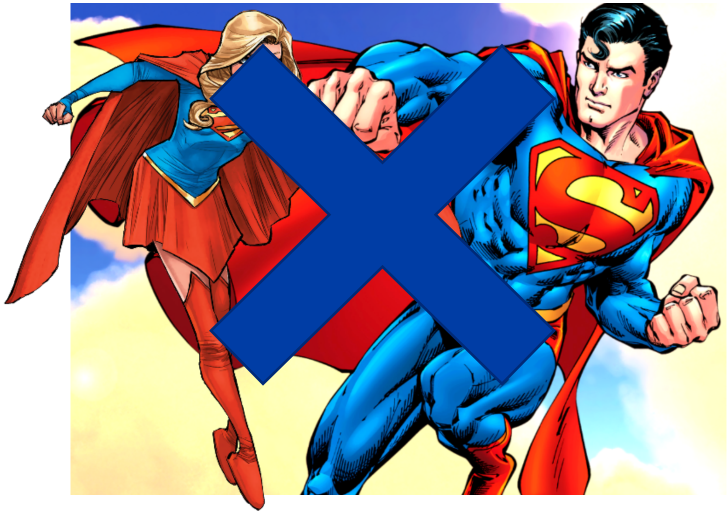
Survey & Breakout

Course goals

- What are your goals?

- Understand what big data means
- Identify the internal components of big data platforms
- Recognize the differences between different big data platforms
- Explain how a distributed query runs on big data

Super Hero



MARVEL
1
MARK WAID • JAVIER GARRÓN • ISRAEL SILVA
ANT-MAN
AND THE
WASP

Ant-Man/Wasp

Get smaller to understand how ants work and what they are capable of.

Use this knowledge to control thousands of ants and do amazing things!

Optional task: Watch Ant-Man and the Wasp this weekend ☺

Big-data Expert

- Understand how the big-data platforms really work
- Control those thousands of processors efficiently to carry out your task

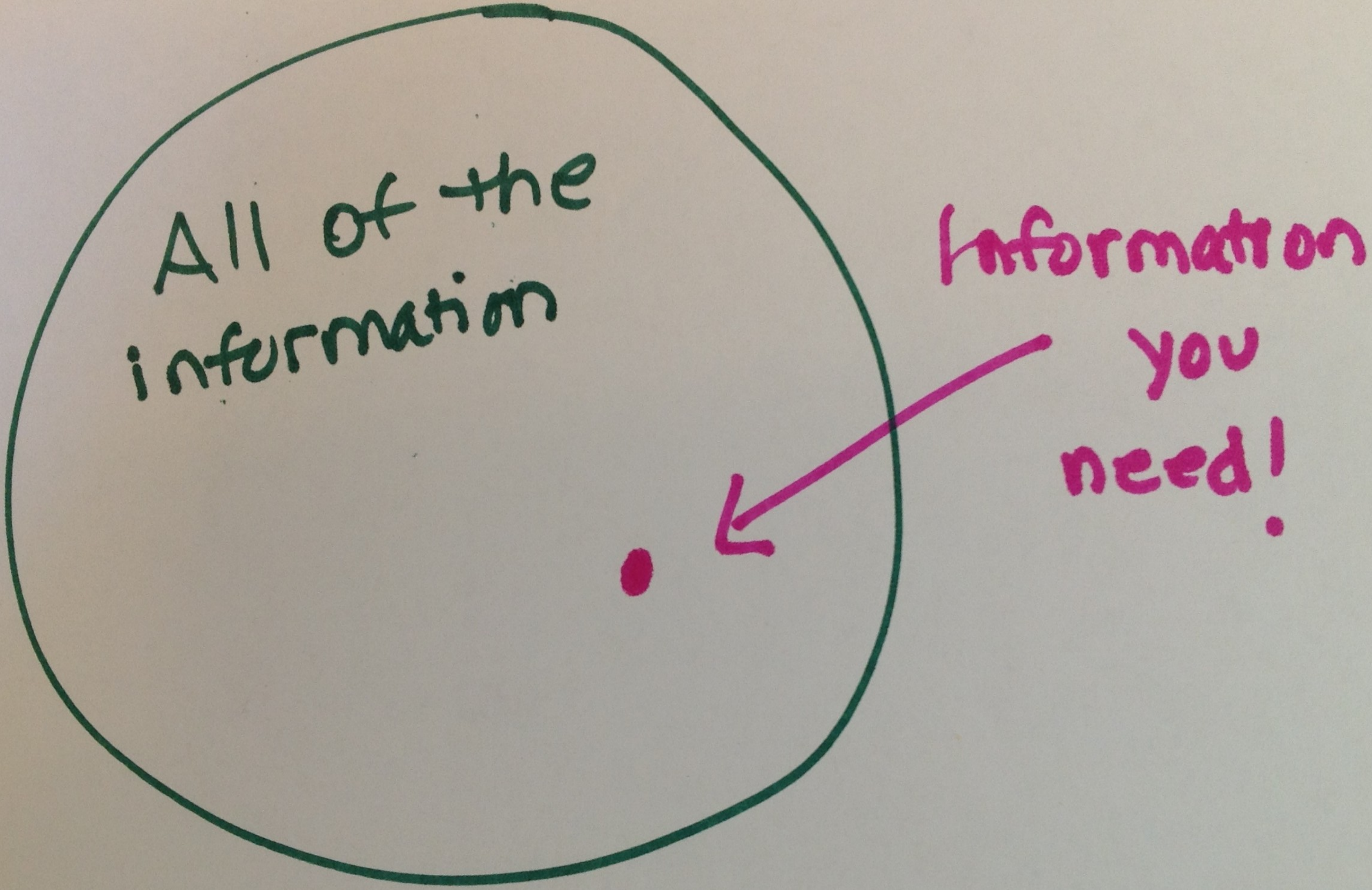
CELEBRATING 30 YEARS
Marlan and Rosemary Bourns College of Engineering

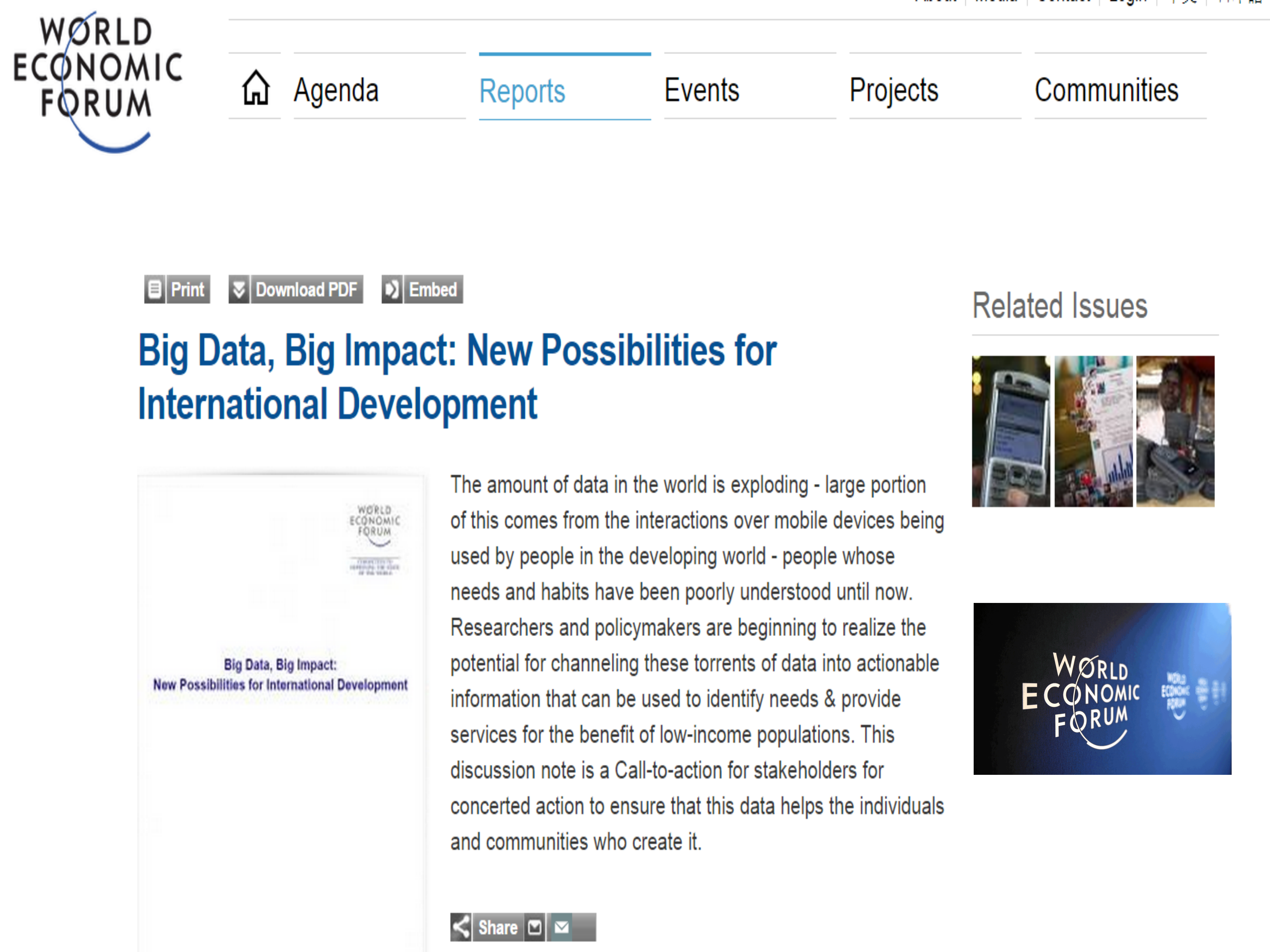
Syllabus

- Overview of big data
- Big-data storage
- Big-data processing
- Big-data indexing
- Big-SQL processing
- Programming packages

Introduction

Big Data
Straight Ahead


All of the information
Information you need!



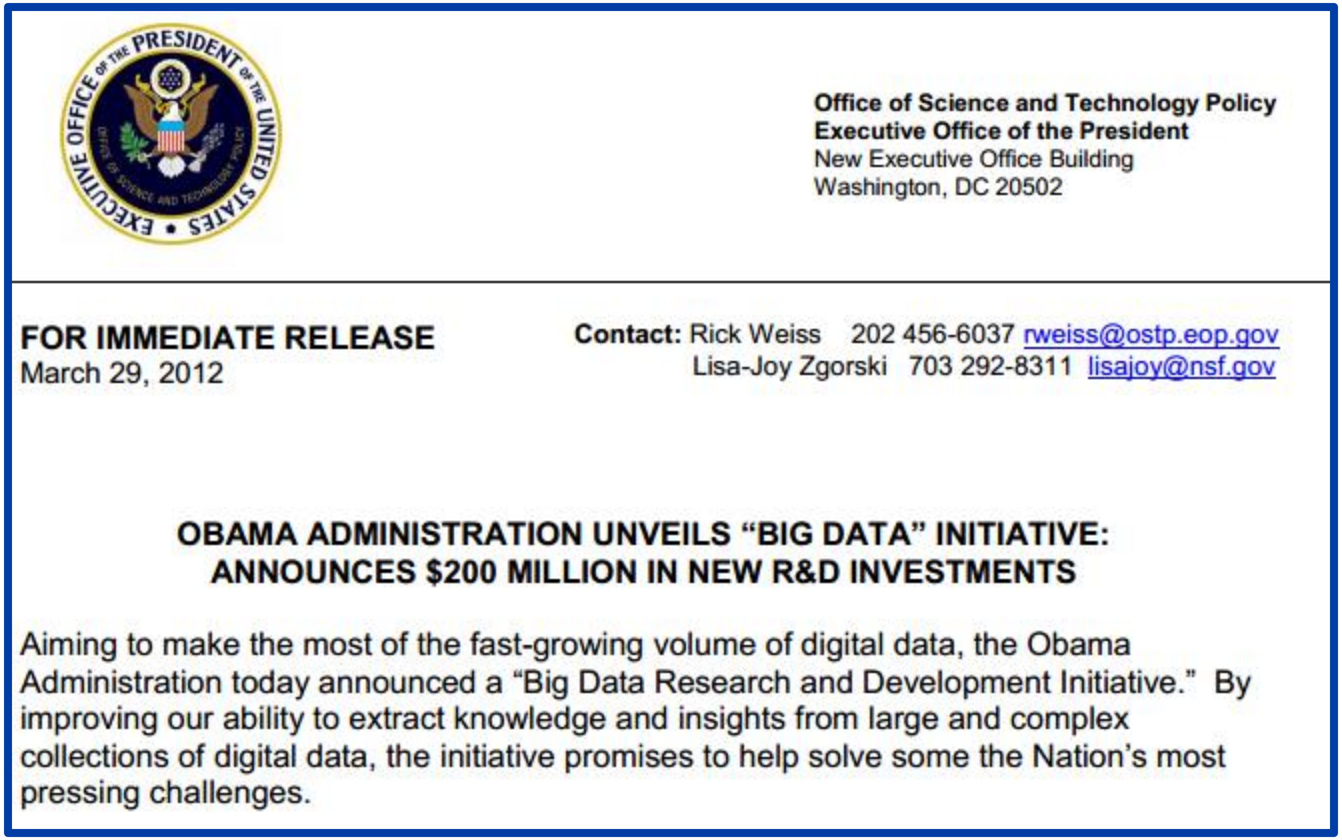
Big Data, Big Impact: New Possibilities for International Development

Big Data, Big Impact:
New Possibilities for International Development

The amount of data in the world is exploding - large portion of this comes from the interactions over mobile devices being used by people in the developing world - people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realize the potential for channeling these torrents of data into actionable information that can be used to identify needs & provide services for the benefit of low-income populations. This discussion note is a Call-to-action for stakeholders for concerted action to ensure that this data helps the individuals and communities who create it.

Interest in Big Data in the US

- **March 2012:** Obama administration unveils **BIG DATA** initiative: $200 Million in R&D investment



Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

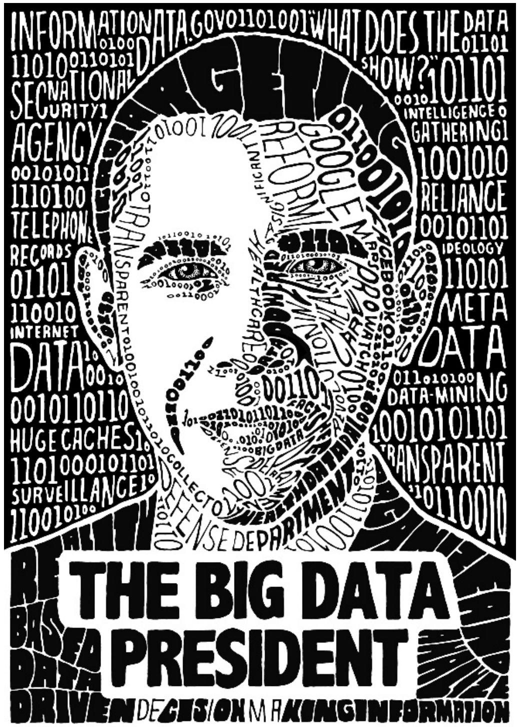
FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

**OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS**

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges.

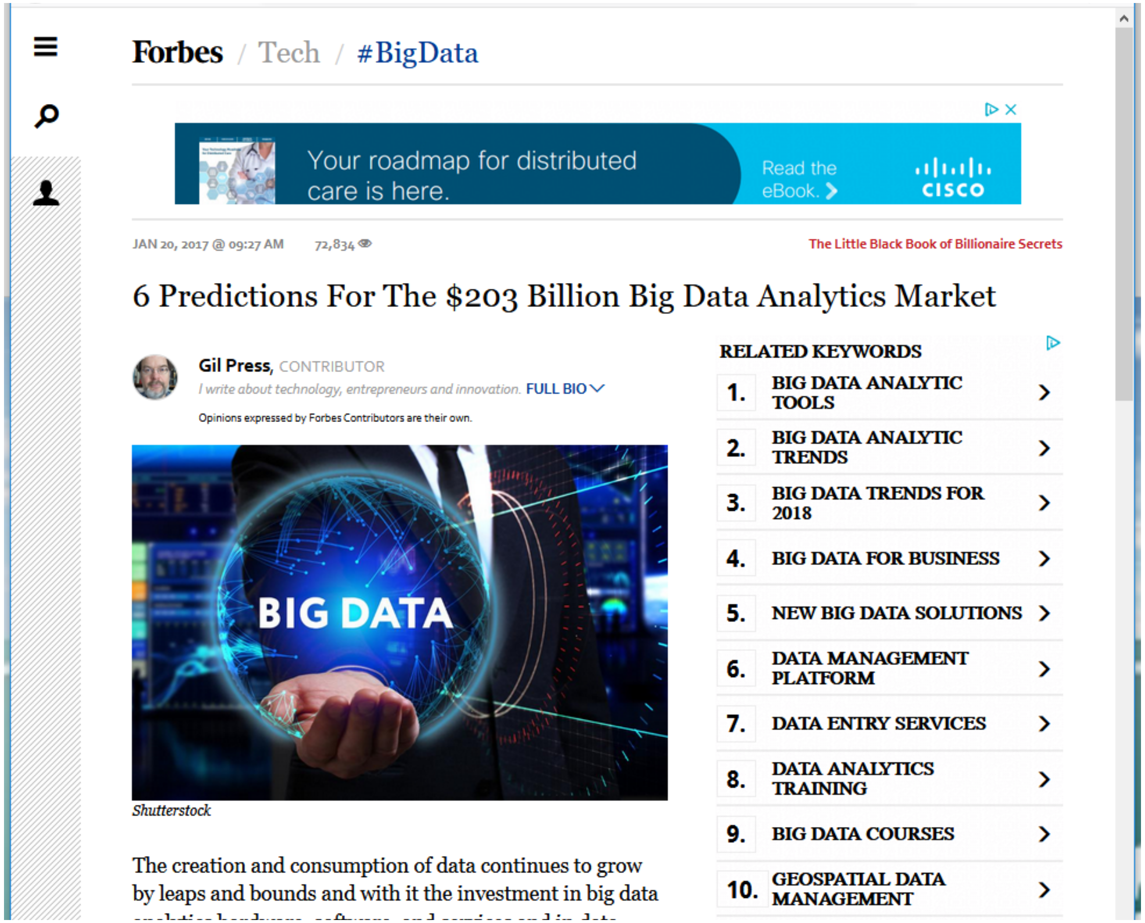
- **June 2013:** Washington Post is calling Obama "**The Big Data President**"



Interest in Big Data in Europe

- March 2014: David Cameron and Angela Merkel talking about Big Data in a Computer Expo in Hannover, Germany

The Market of Big Data



Forbes / Tech / #BigData

JAN 20, 2017 @ 09:27 AM 72,834

6 Predictions For The $203 Billion Big Data Analytics Market

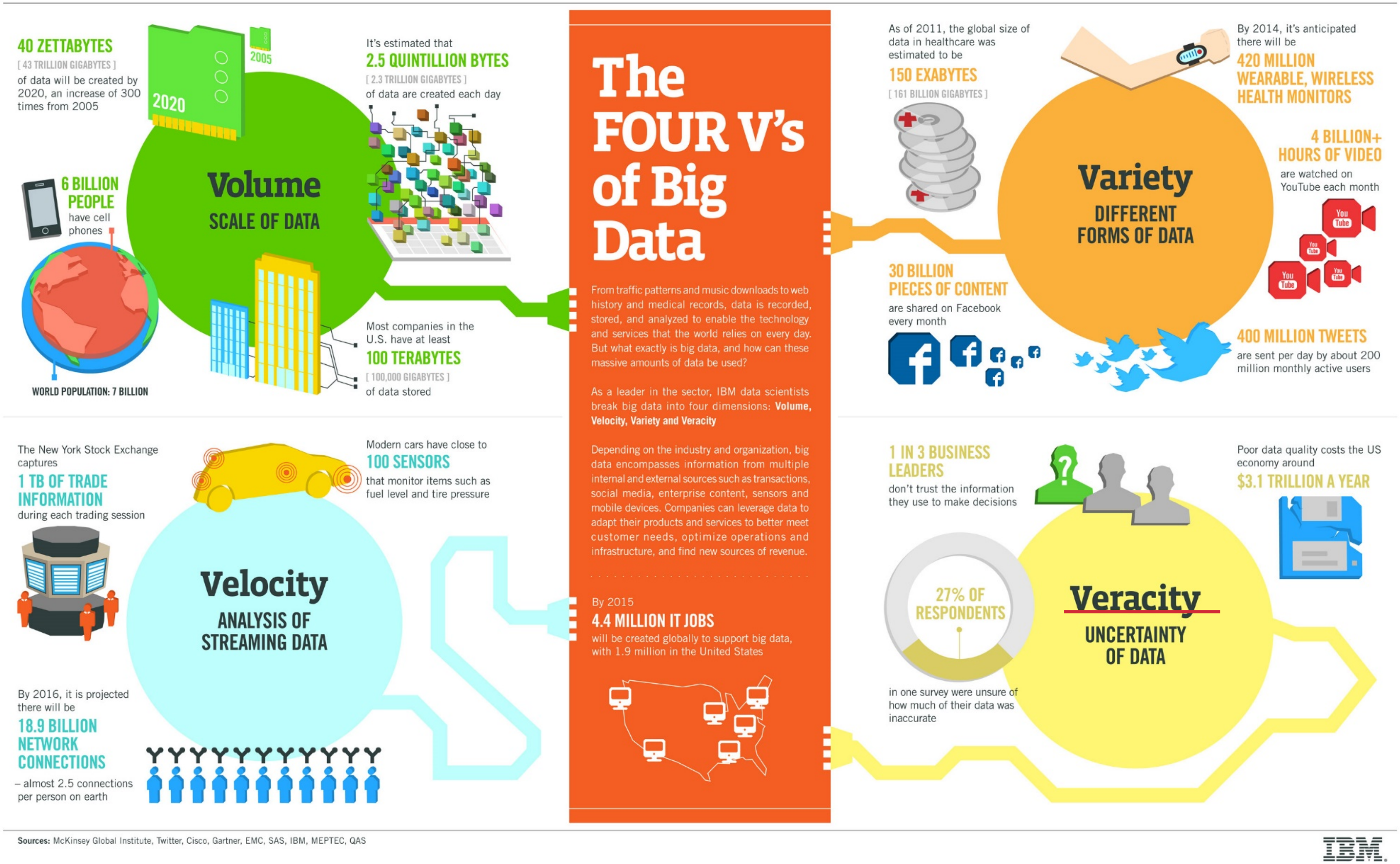
Gil Press, CONTRIBUTOR
I write about technology, entrepreneurs and innovation. FULL BIO

Opinions expressed by Forbes Contributors are their own.

Shutterstock

The creation and consumption of data continues to grow by leaps and bounds and with it the investment in big data

~~Four~~ Three V’s of Big Data



The FOUR V’s of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It’s estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity
ANALYSIS OF STREAMING DATA

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

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it’s anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Veracity
UNCERTAINTY OF DATA

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

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

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

IBM.

Big Data Vs Big Computation

- Full scans (e.g., log processing)
- Range scans
- Point lookups
- Iterations
- Joins (self, binary, or multiway)
- Proximity queries
- Closures and graph traversals

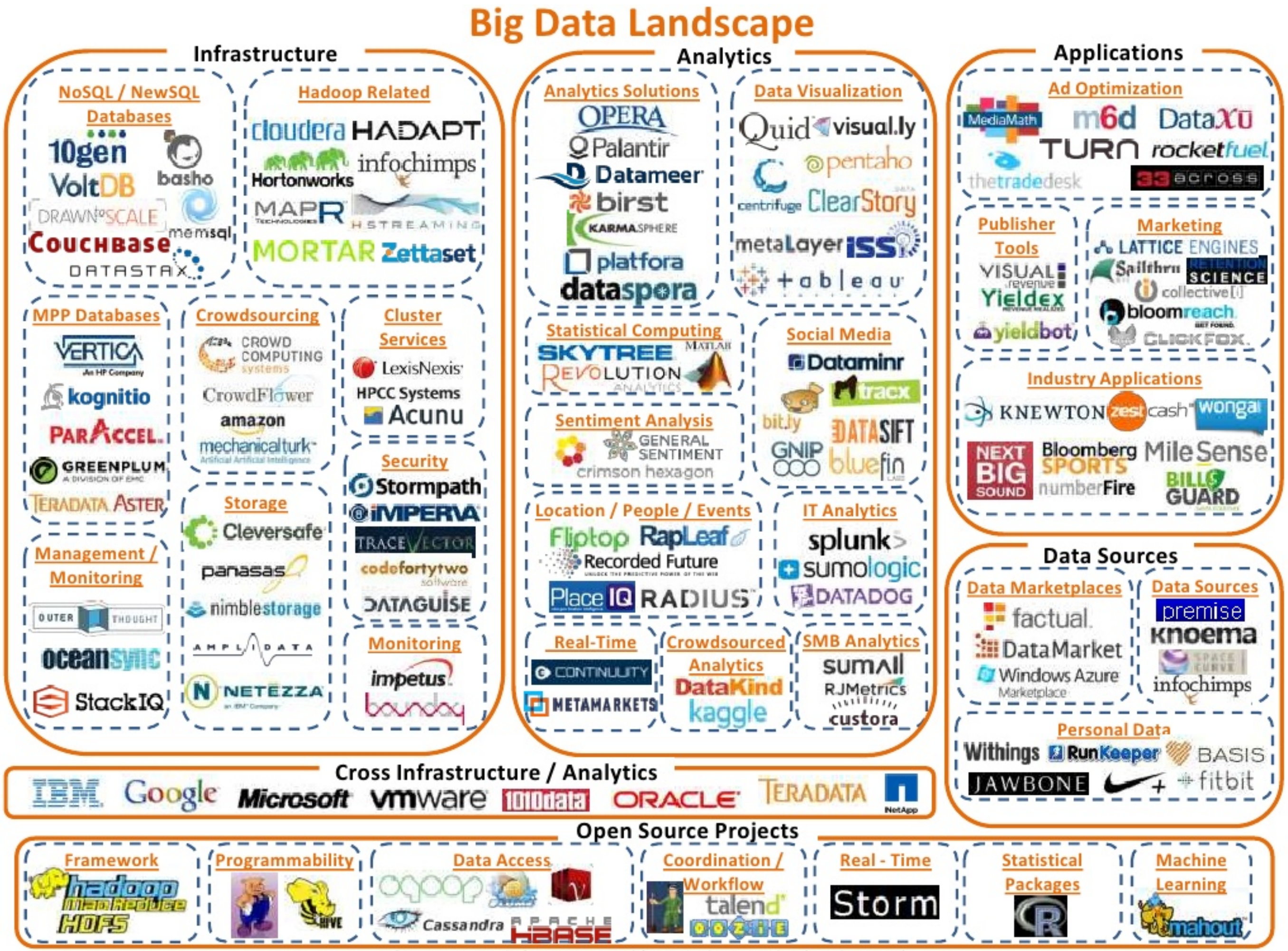
Big Data Applications

- Web search
- Marketing and advertising
- Data cleaning
- Knowledge base
- Information retrieval
- Internet of Things (IoT)
- Visualization
- Behavioral studies

Publicly Available Datasets

- Data.gov
- UCR Star [https://star.cs.ucr.edu]
  - facebook.com/ucrstar & 👍
- Twitter Streaming API
- Yahoo! Webscope [http://webscope.sandbox.yahoo.com/]
- GDELT [http://www.gdeltproject.org/]
- Instagram API

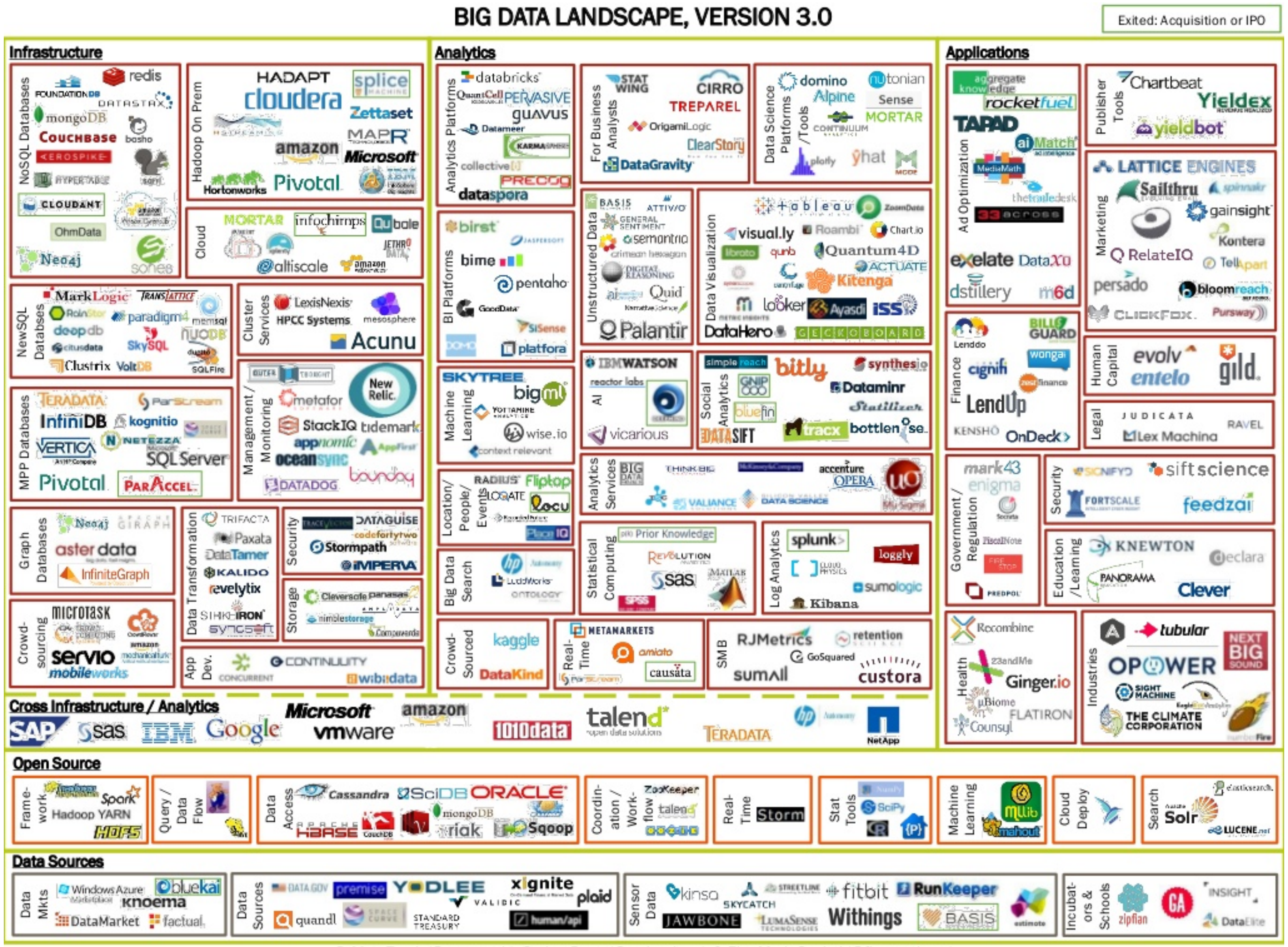
Big Data Landscape 2012



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

http://mattturck.com/2012/06/29/a-chart-of-the-big-data-ecosystem/

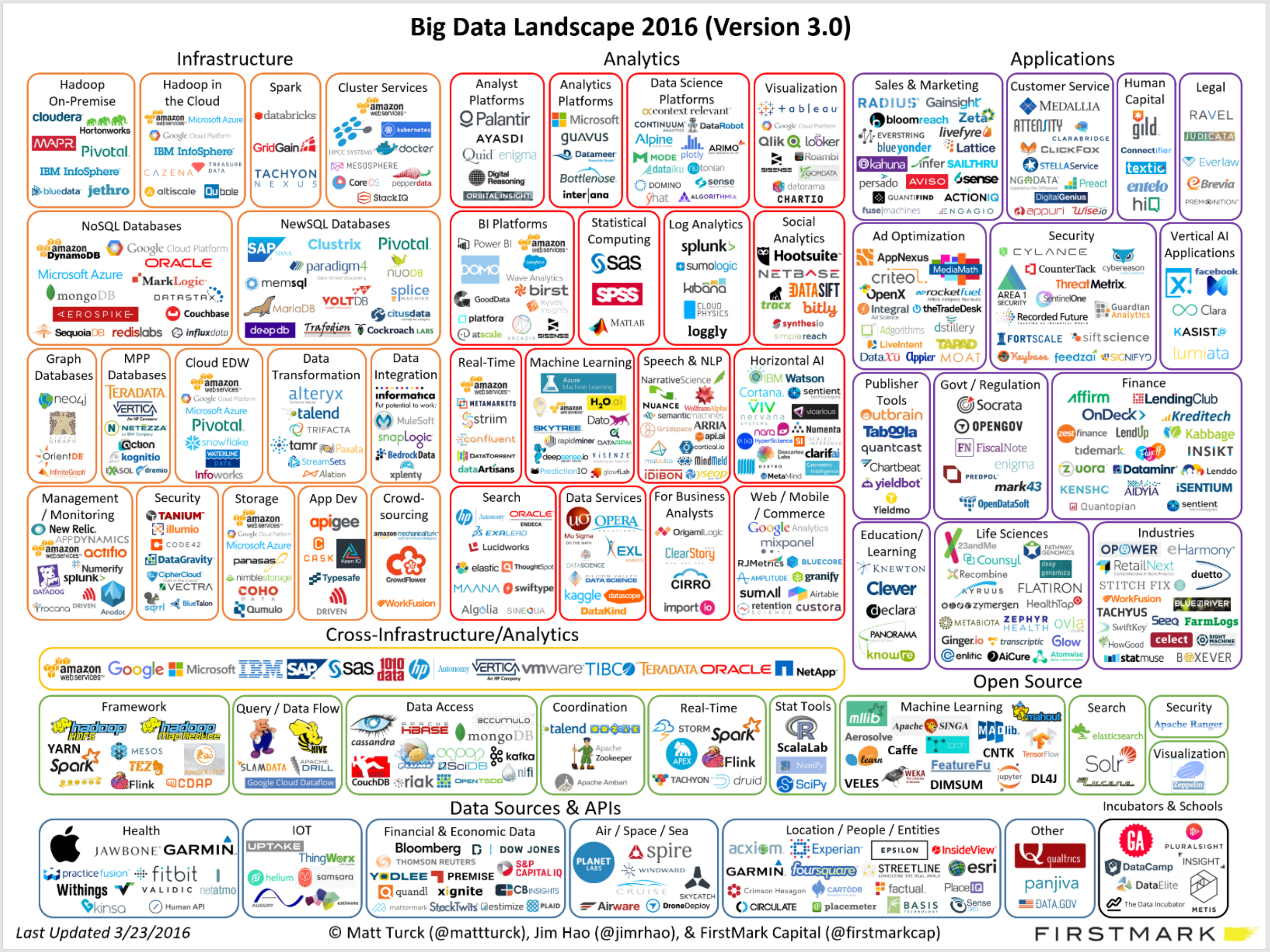
Big Data Landscape 2014



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

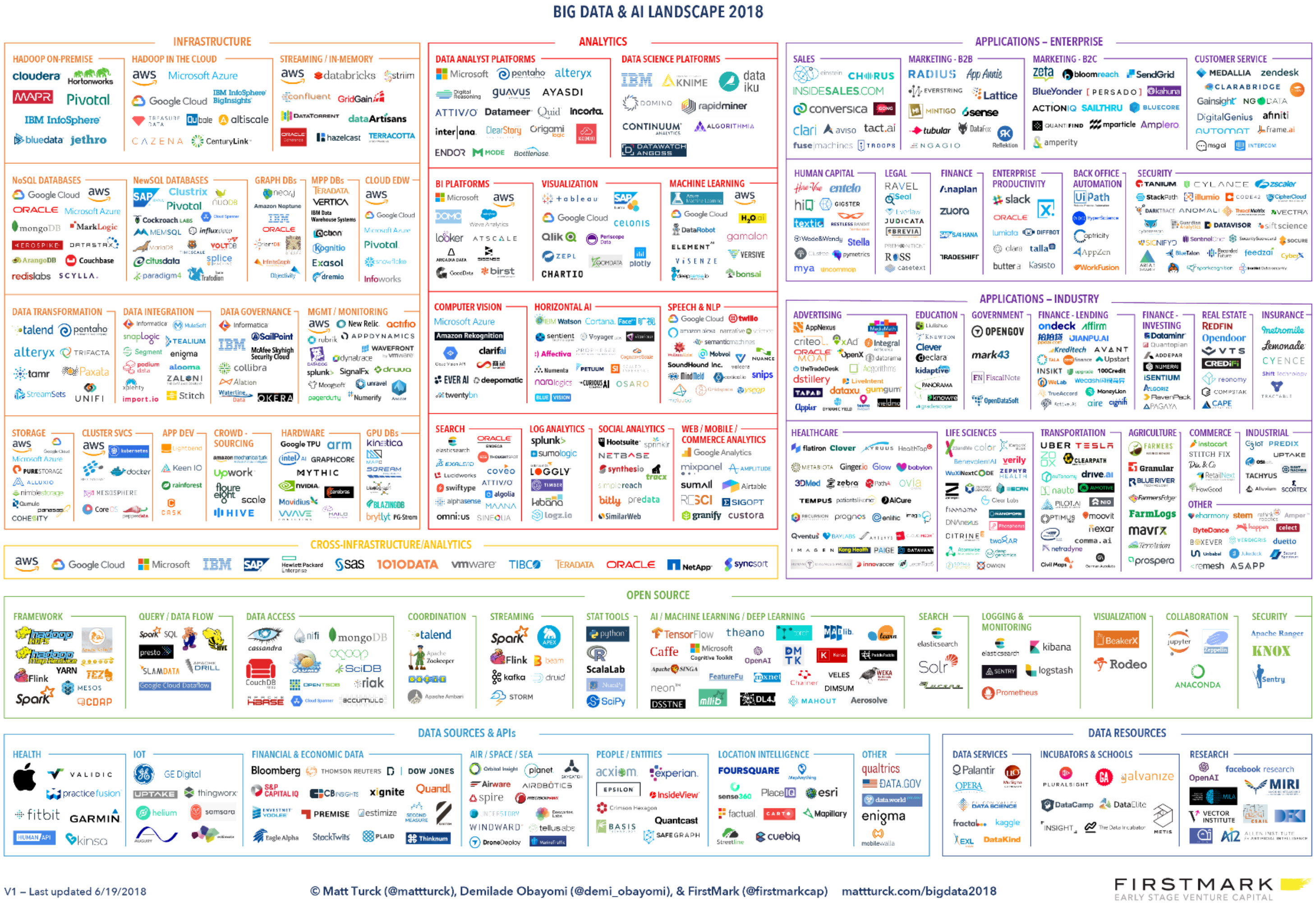
http://mattturck.com/2014/05/11/the-state-of-big-data-in-2014-a-chart/

Big Data Landscape 2016



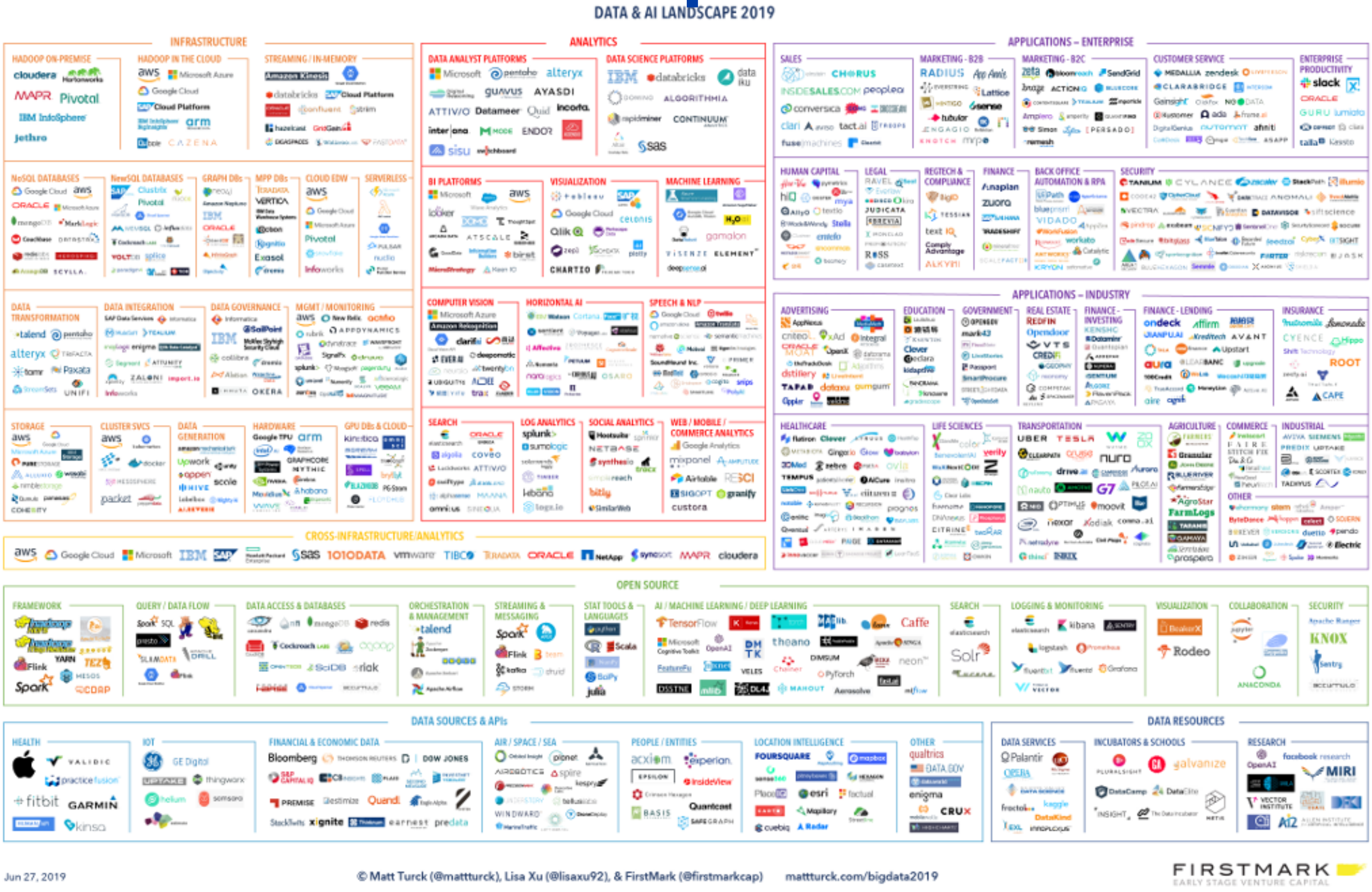
http://mattturck.com/2016/02/01/big-data-landscape/

Big Data Landscape 2018



CELEBRATING 30 YEARS
Marlan and Rosemary Bourns College of Engineering

Data & AI Landscape 2019

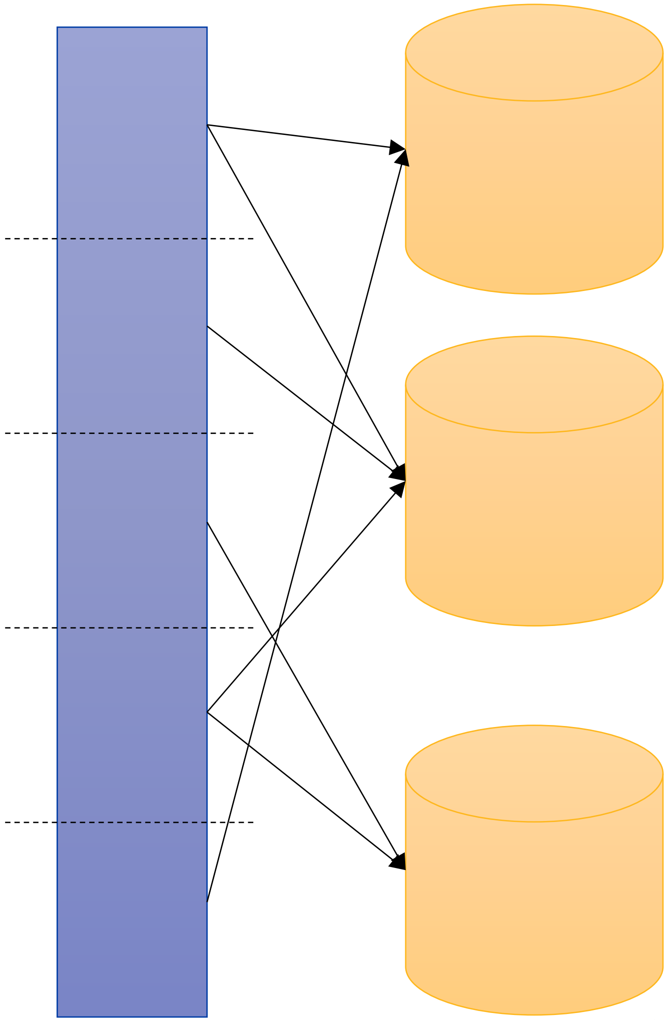


ırns

Components of Big Data

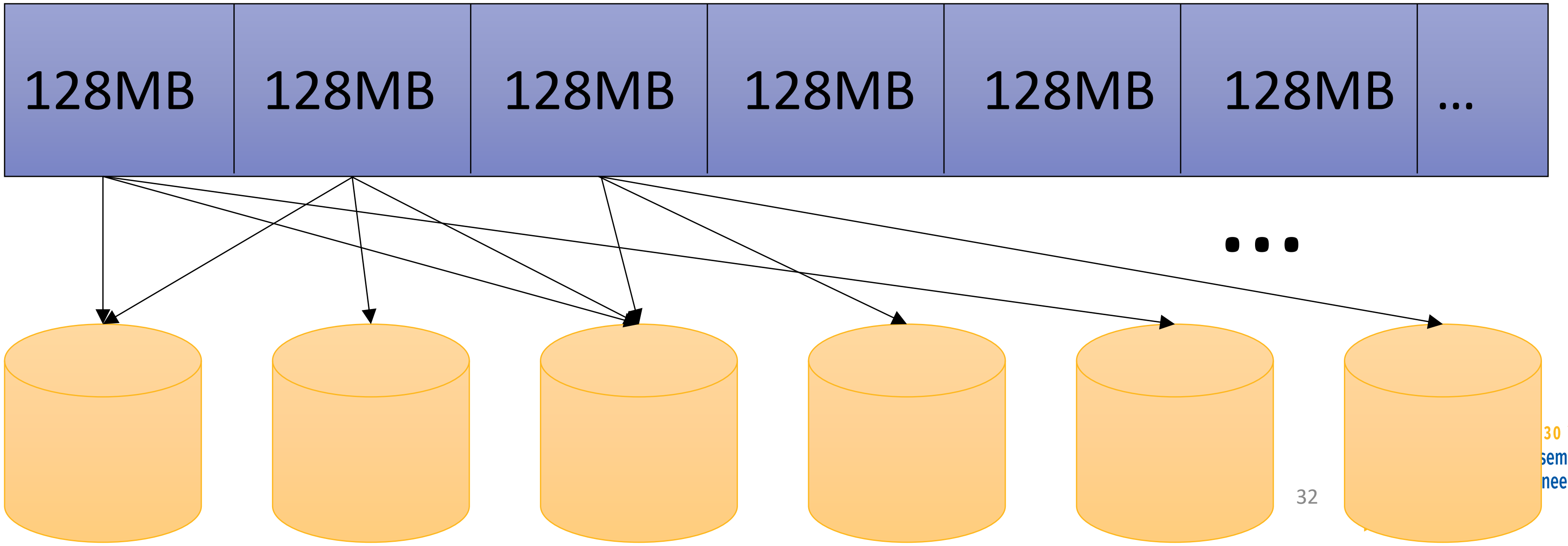
Storage of Big Data

- Data is growing faster than Moore's Law
- Too much data to fit on a single machine
- Partitioning
- Replication
- Fault-tolerance



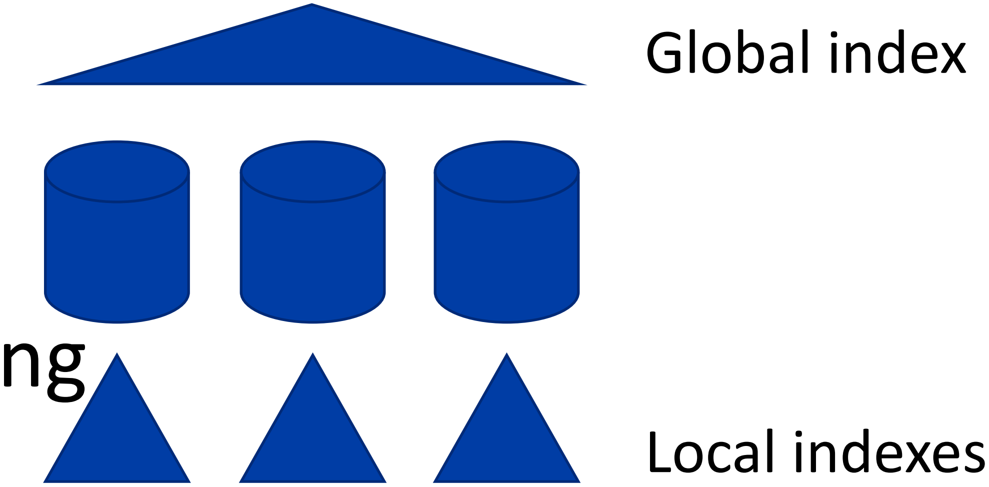
Hadoop Distributed File System (HDFS)

- The most widely used distributed file system
- Fixed-sized partitioning
- 3-way replication
- Write-once read-many
- See also: GFS, Amazon S3, Azure Blob Store



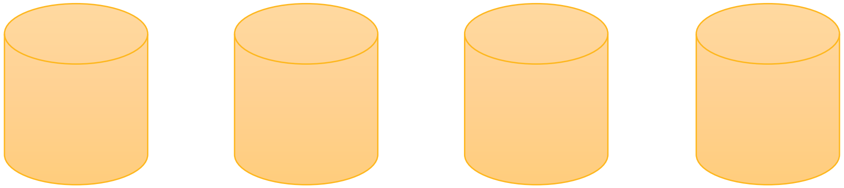
Indexing

- Data-aware organization
- Global Index **partitions** the records into blocks
- Local Indexes organize the records in a partition
- Challenges:
  - Big volume
  - HDFS limitation
  - New programming paradigms
  - Ad-hoc indexes

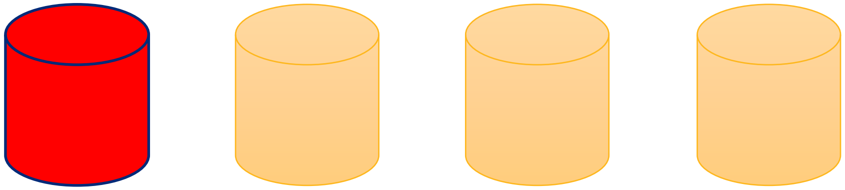


Fault Tolerance

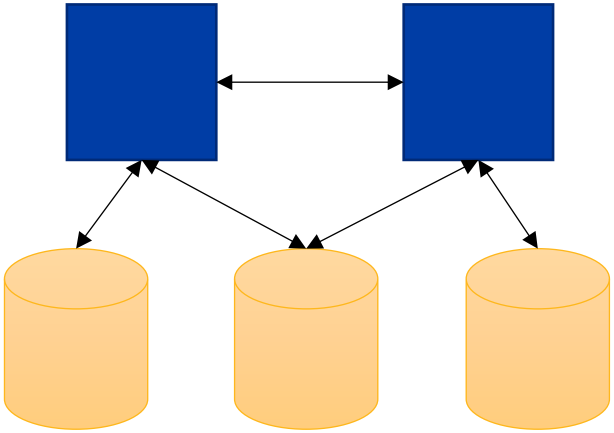
- Replication



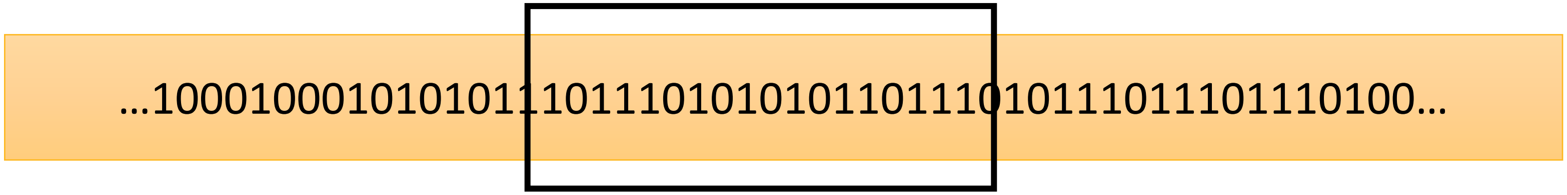
- Redundancy



- Multiple masters



Streaming

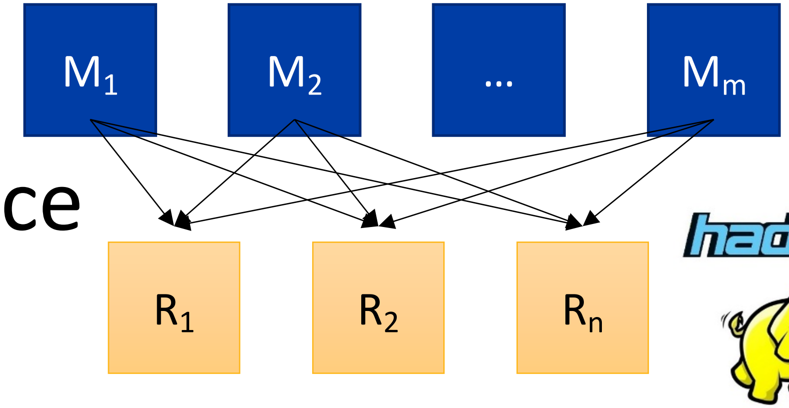


Processing window

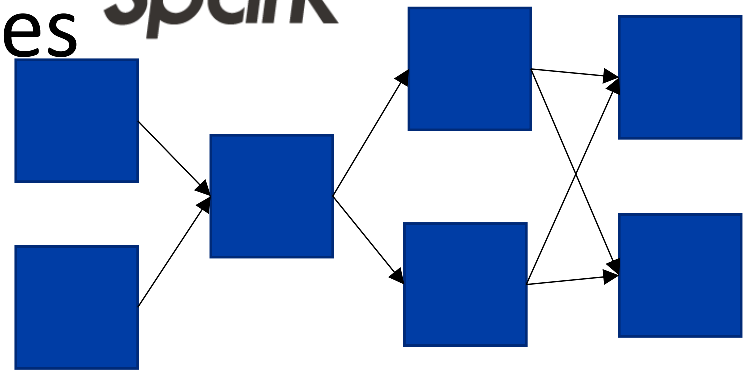
- Sub-second latency for queries
- One scan over the data
- (Partial) preprocessing
- Continuous queries
- Eviction strategies
- In-memory indexes

Task Execution

- MapReduce
  - Map-Shuffle- Reduce
  - Resiliency through materialization



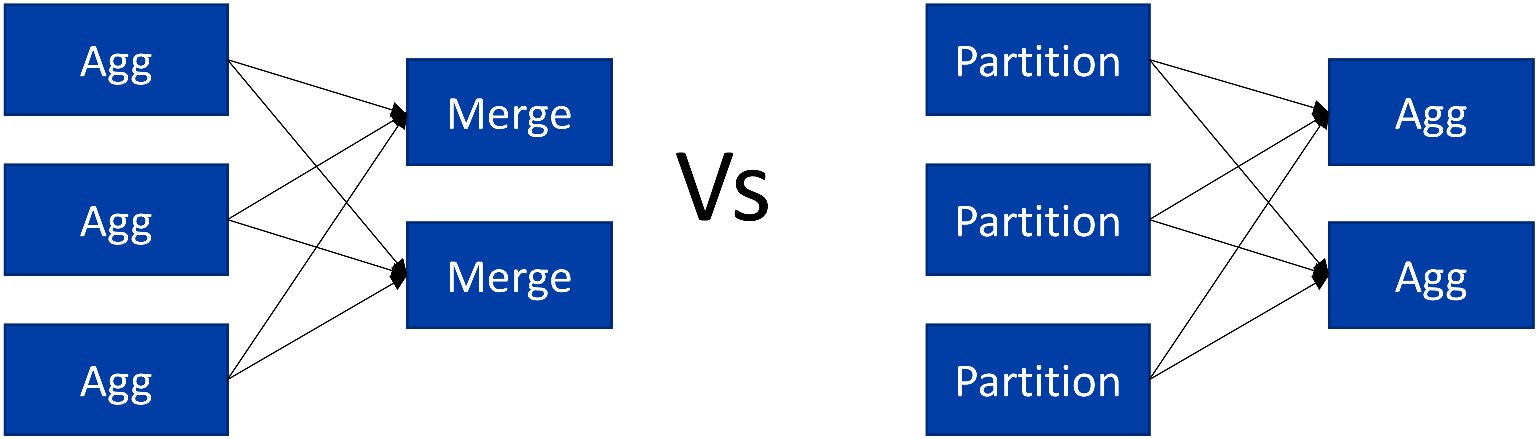
- Resilient Distributed Datasets (RDD)
  - Directed-Acyclic-Graph (DAG)
  - In-memory processing
  - Resiliency through lineages



- Hyracks
- Stragglers
- Load balance

Query Optimization

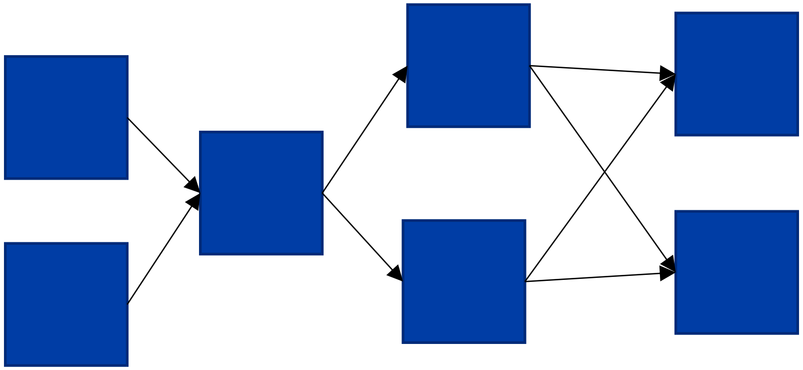
- Finding the most efficient query plan
- e.g., grouped aggregation



- Cost model (CPU – Disk – Network)

Provenance

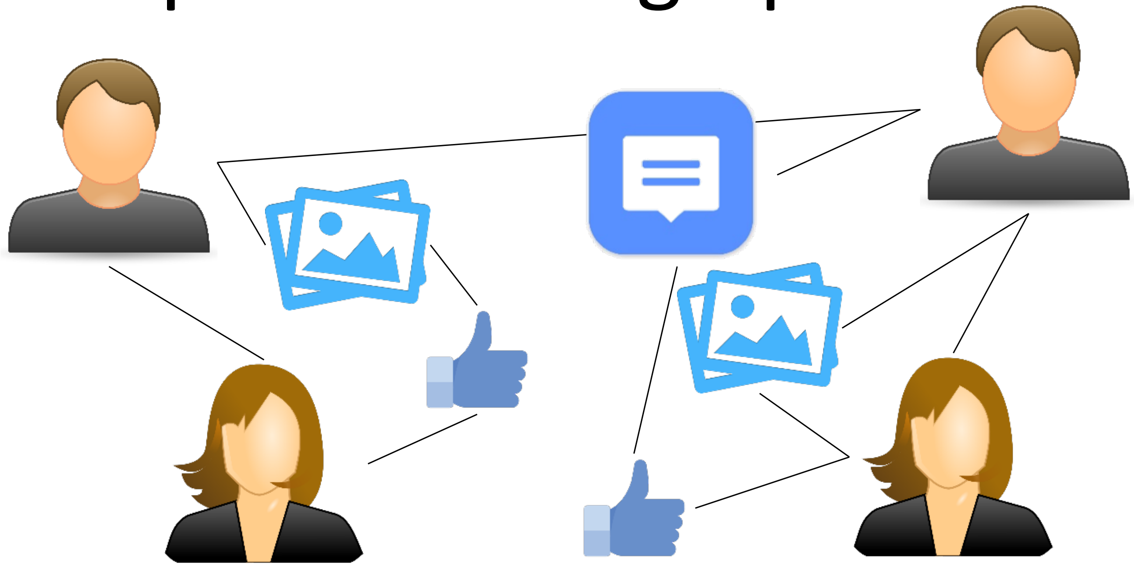
- Debugging in distributed systems is painful



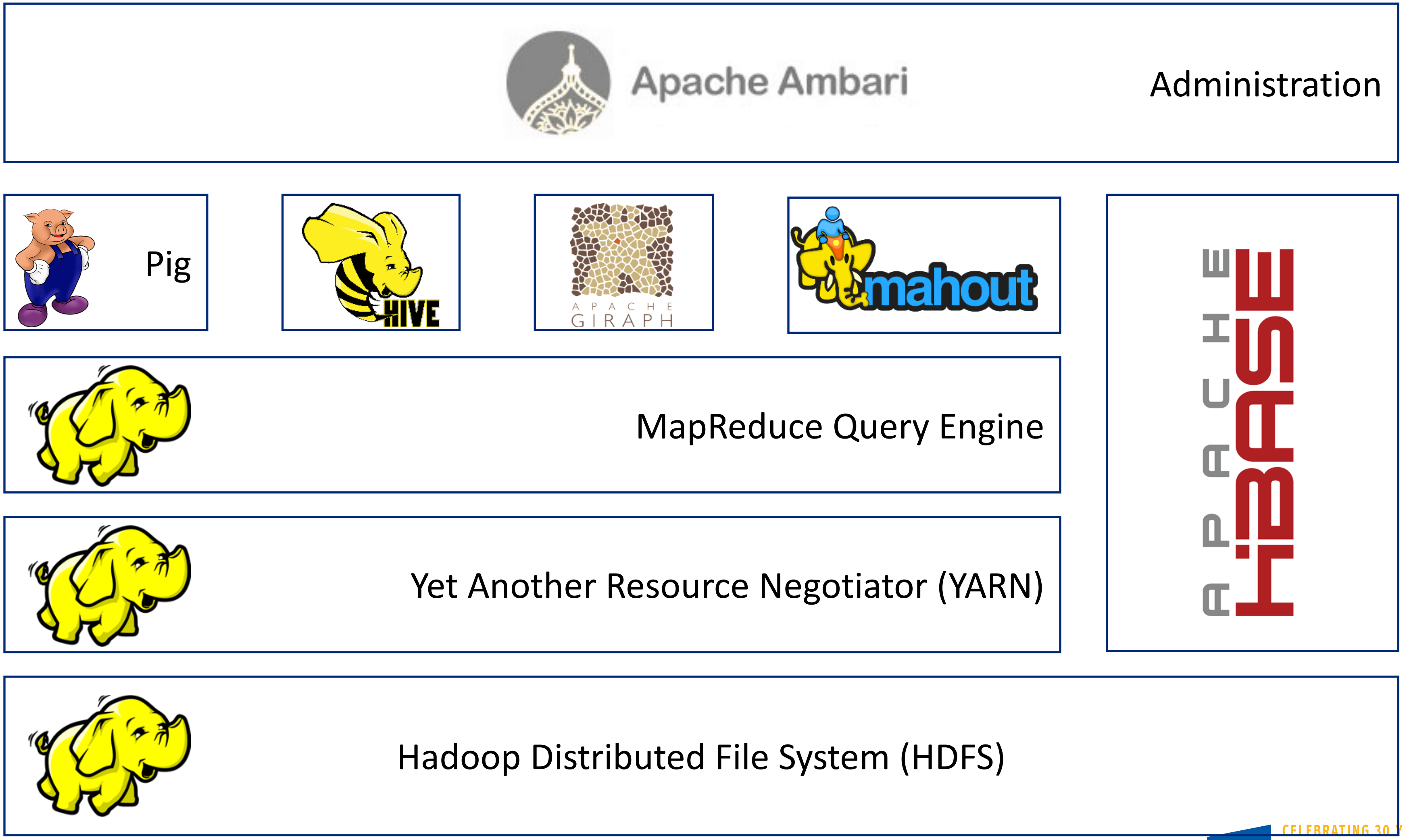
- We need to keep track of transformations on each record

Big Graphs

- Motivated by social networks
- Billions of nodes and trillions of edges
- Tens of thousands of insertions per second
- Complex queries with graph traversals



Hadoop Ecosystem



Administration

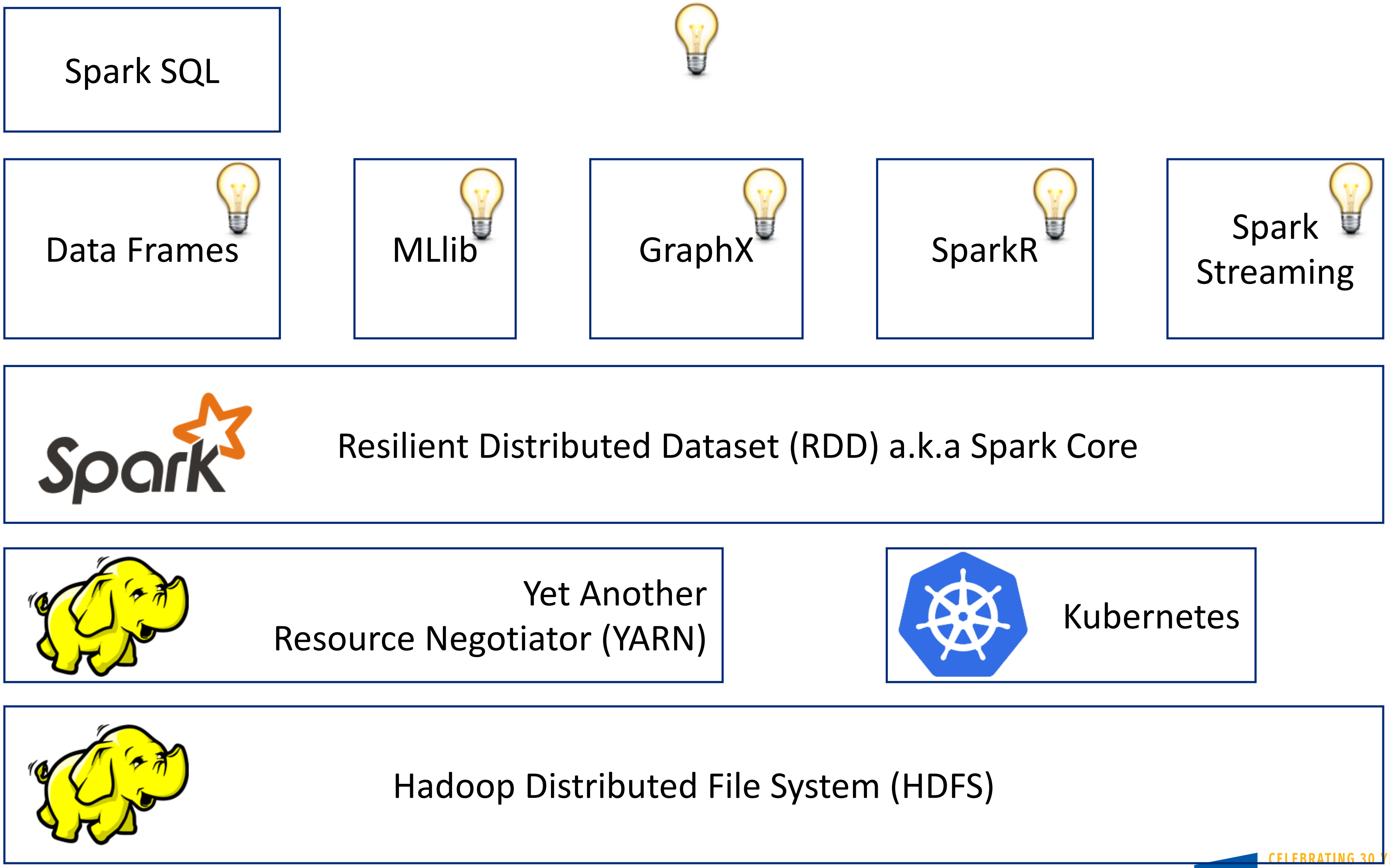
Pig

MapReduce Query Engine

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

Spark Ecosystem



Spark SQL

Data Frames

MLlib

GraphX

SparkR

Spark Streaming

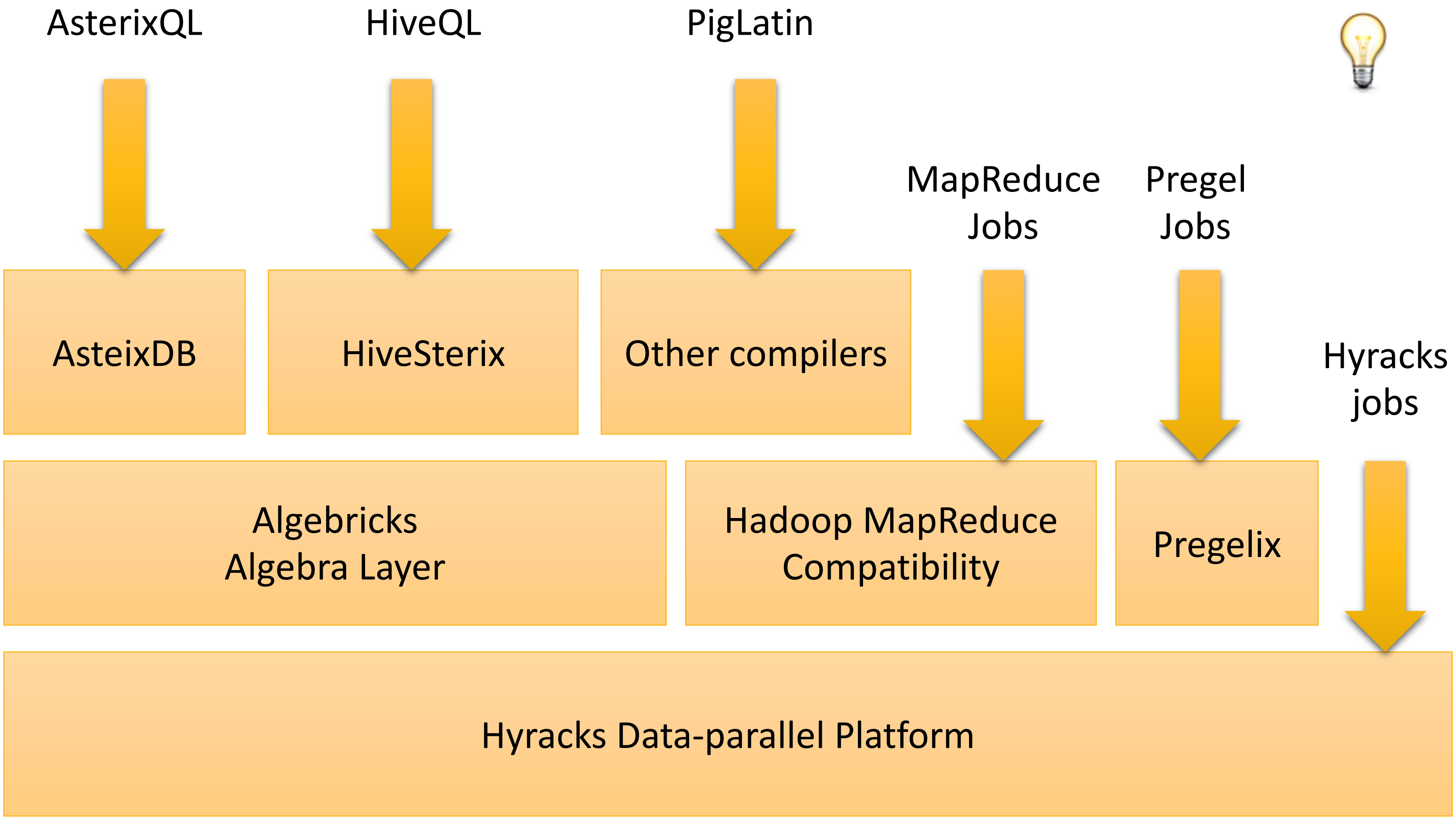
Resilient Distributed Dataset (RDD) a.k.a Spark Core

Yet Another Resource Negotiator (YARN)

Kubernetes

Hadoop Distributed File System (HDFS)

AsterixDB



AsterixQL

HiveQL

PigLatin

MapReduce Jobs

Pregel Jobs

AsteixDB

HiveSterix

Other compilers

Hyracks jobs

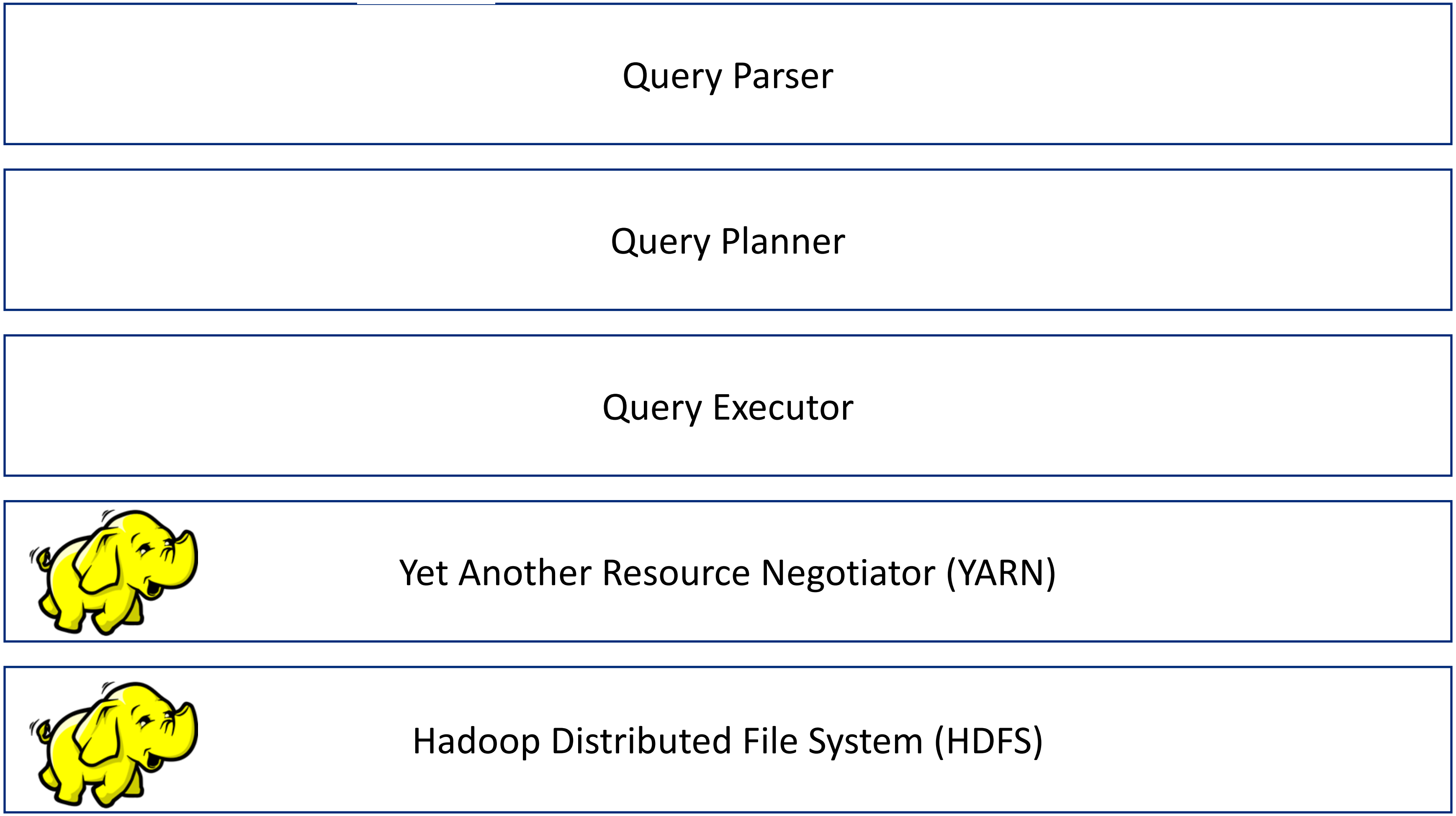
Algebricks Algebra Layer

Hadoop MapReduce Compatibility

Pregelix

Hyracks Data-parallel Platform

Impala



Query Parser

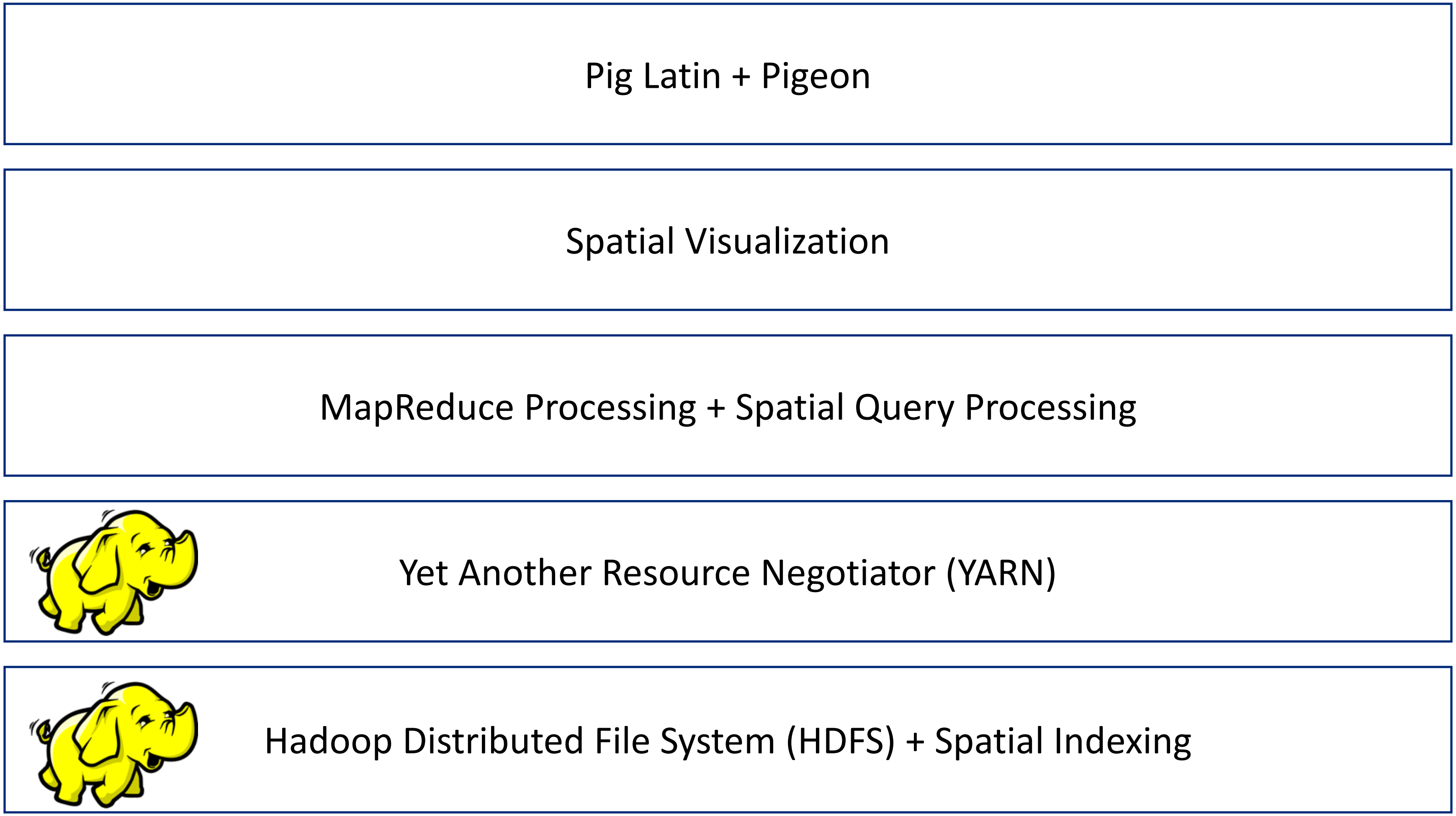
Query Planner

Query Executor

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

SpatialHadoop



Pig Latin + Pigeon

Spatial Visualization

MapReduce Processing + Spatial Query Processing

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS) + Spatial Indexing

Reading Material

- "The Age of Analytics in a Data-driven World" [Executive Summary] by McKinsey & Company