CS167
Introduction to Big-data

Instructor: Ahmed Eldawy

Welcome to UCR! (Virtually)

Class information

- Classes: Tuesday, Thursday 2:00 – 3:20 PM via Zoom
- Instructor: Ahmed Eldawy
- Office hours: Monday, Thursday 11:00-11:50 (Conflicts?)
- TA: Payas Rajan and Xin Zhang
- Website: http://www.cs.ucr.edu/~eldawy/21SCS167/
- Email: eldawy@ucr.edu Subject: "[CS167] …"
- Slack workspace https://join.slack.com/t/cs167s21/shared_invite/zt-odmd6bxu-x4rLDpFmDlXoRRuRv3iulA

Class Logistics

- All classes will be recorded and published after class
- Ask questions in the chat window
- Raise your hand (in Zoom) if you have a question that you would like to ask verbally

Lab Logistics

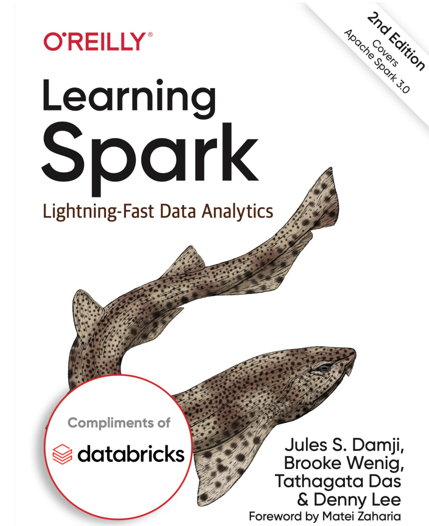
- All labs will be on Zoom
- Attend the session that you are enrolled in
- The TA will share their screen
- Students will follow the instructions on their machines
- Ask questions in the chat
- If you have a question, you can share your screen (privately) with the TA to get help!!

Course work

- Active participation (10%)
- Assignments (15%) (3% x 5)
- Labs (30%) (3% x 10)
- Mid-terms (15% x 3)

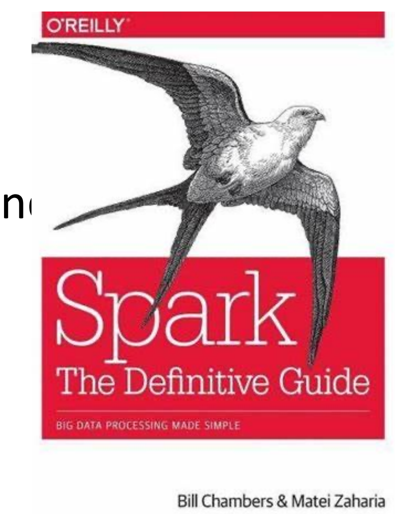
- All exams will be open slides, notes, and book.

Textbook

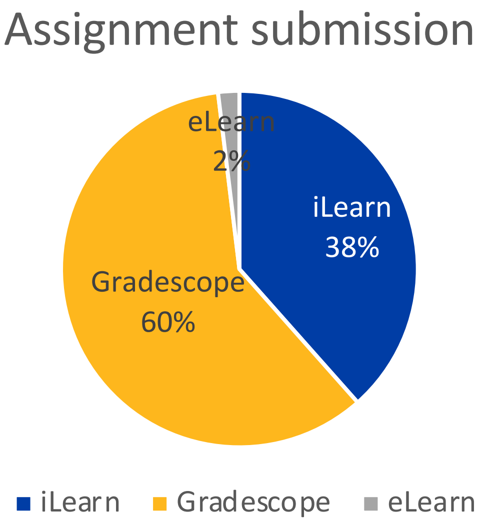


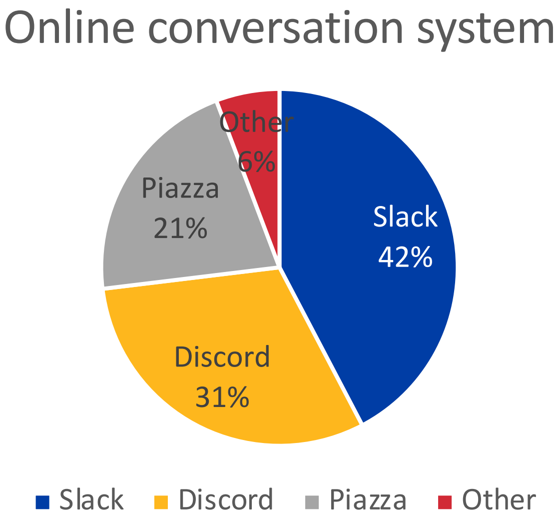
- No required textbook
- Recommended textbooks

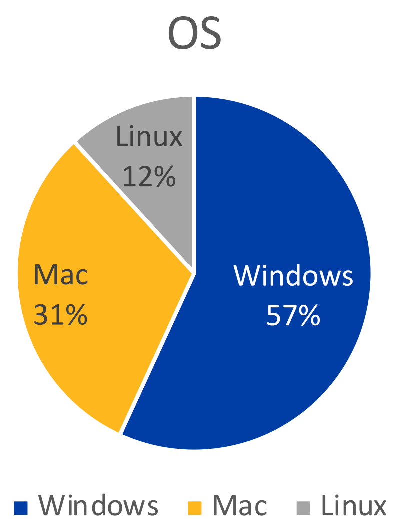
1. "Learning Spark Lightning-Fast Data Analytics" 2nd Edition by Jules S. Damji, Brooke Wenig, Tathagata Das & Denny Lee.
2. "Spark: The Definitive Guide: Big Data Processing Made Simple": 1st Edition, by Bill Chambers an Matei Zaharia
   ISBN-13: 978-1491912218
   ISBN-10: 1491912219



Presurvey Results





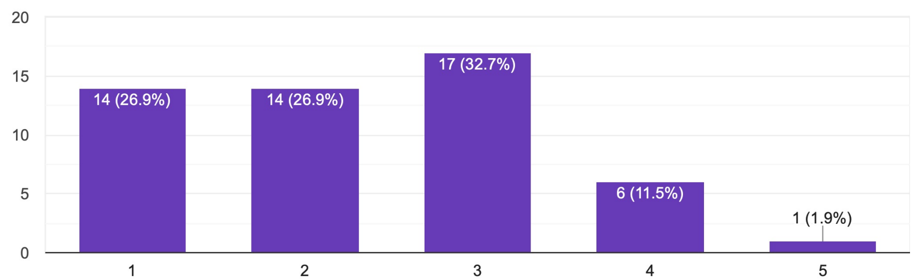


Background

Java

52 responses

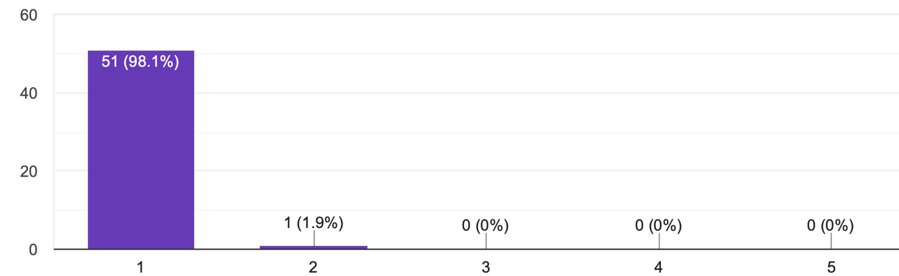
Java



Scala

52 responses

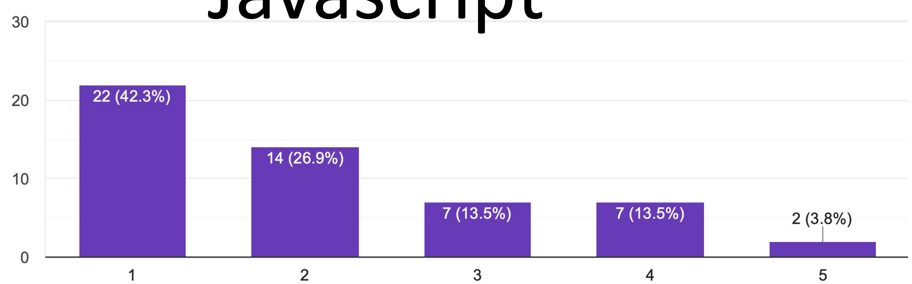
Scala



Javascript

52 responses

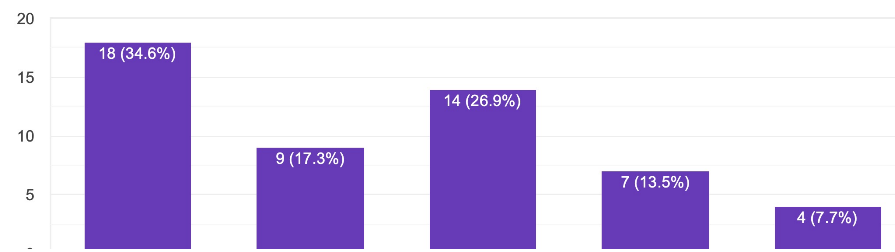
Javascript



SQL

SQL

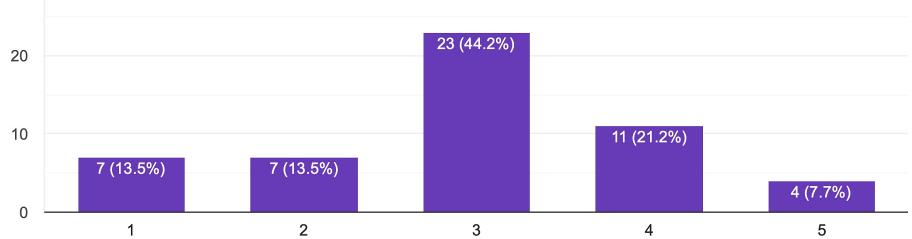
52 responses



Linux command-line tools

52 responses

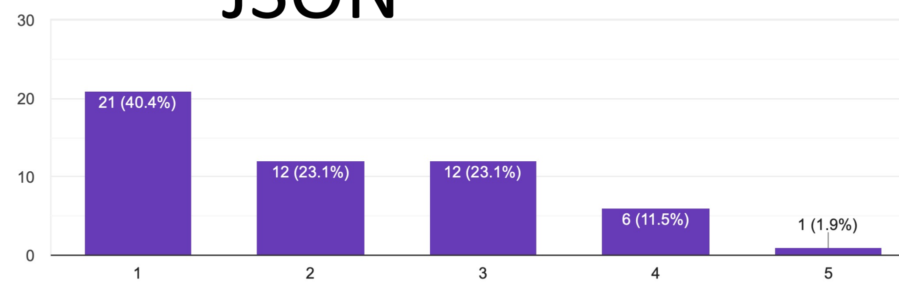
Linux command-line



JSON data format

52 responses

JSON



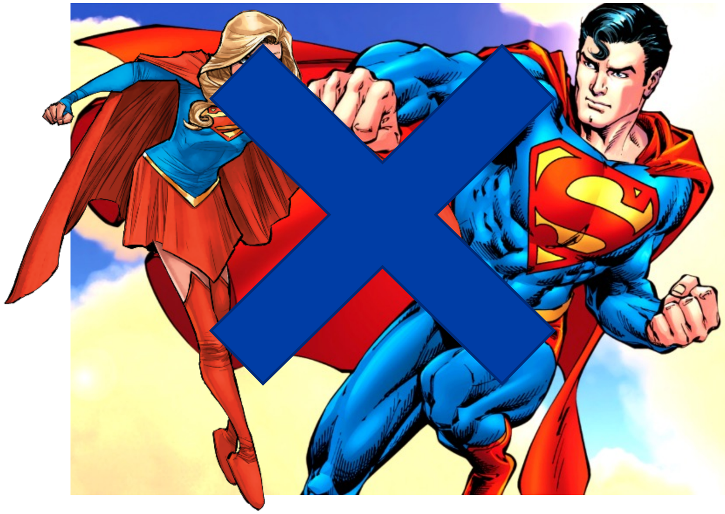
Excitements/Concerns

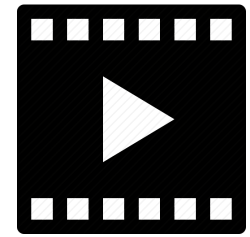
- Excited
  - Play with large amounts of data
  - Not sure/Exploring
  - Distributed frameworks
  - Learn about machine learning
  - Search/Sort big-data
  - Internals of big-data systems
  - Move from DBMS to big-data
  - Learn new technology
- Concerns
  - Steep learning curve
  - Extreme workload
  - Not having the necessary background
  - Not being the right course for me
  - Online teaching

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

Superhero






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!

Big-data Expert

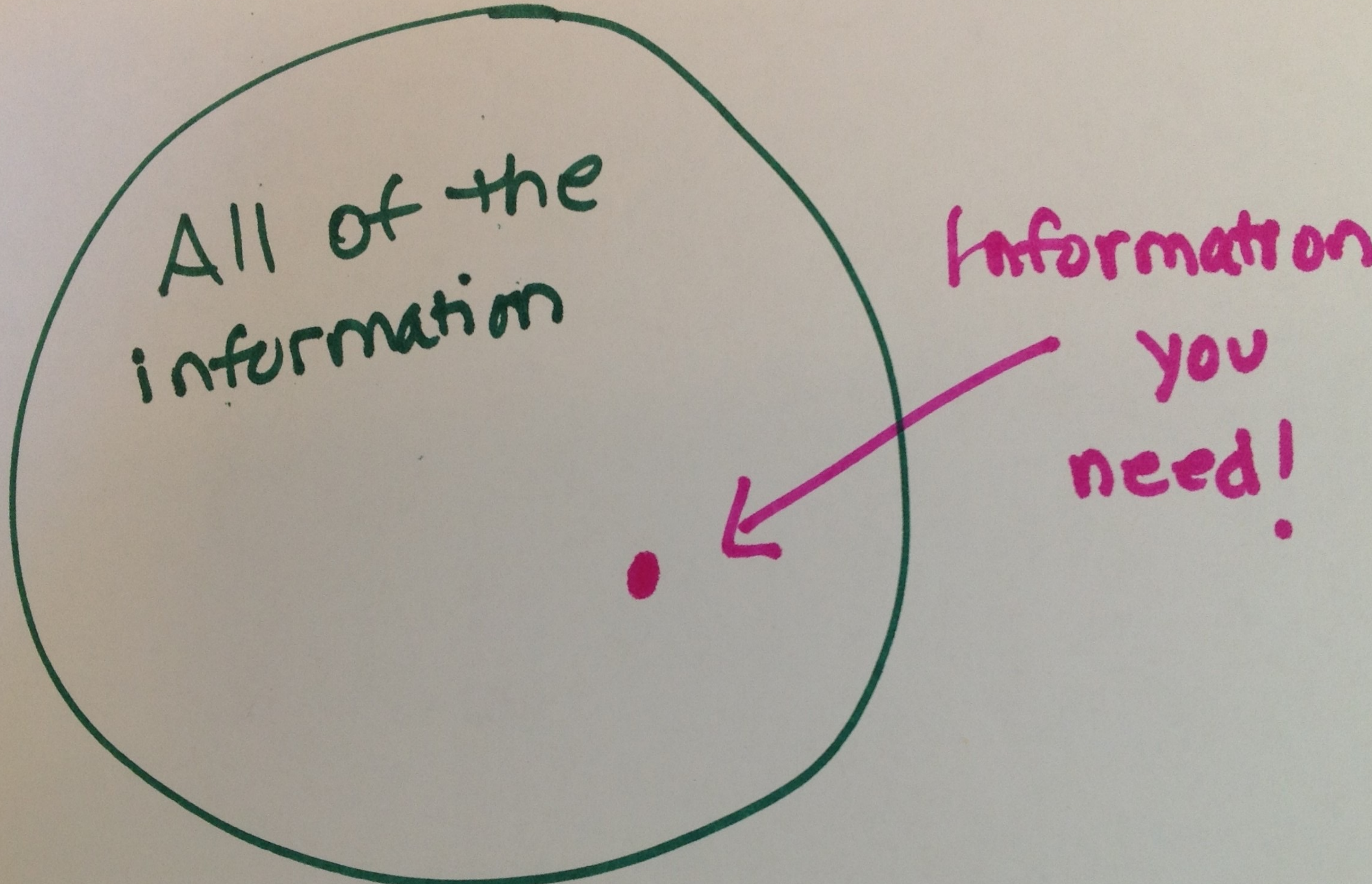
- Understand how the big-data platforms really work
- Control those thousands of processors efficiently to carry out your ta

Syllabus

- Overview of big data
- Big-data storage
- Big-data processing
- Structured data processing
- Column-based storage and retrieval
- Big-spatial data
- Document databases
- Machine learning on big data
- Big-data visualization

Introduction


Big Data
Straight Ahead


All of the information
Information you need!

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

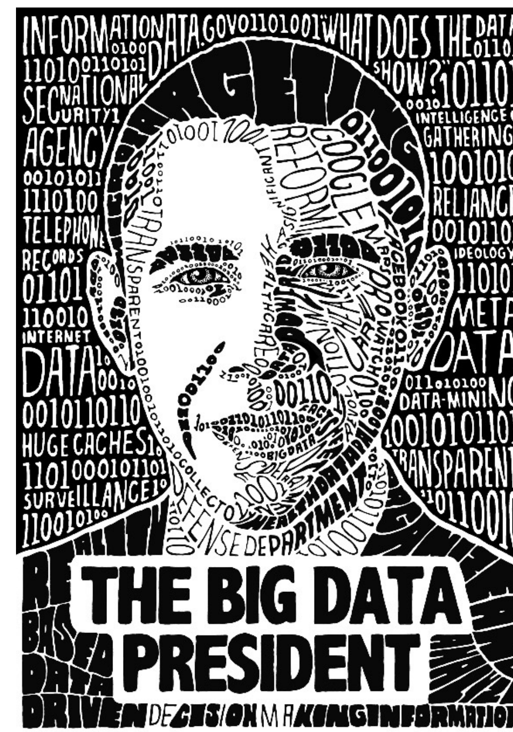
FOR IMMEDIATE RELEASE
March 29, 2012

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.

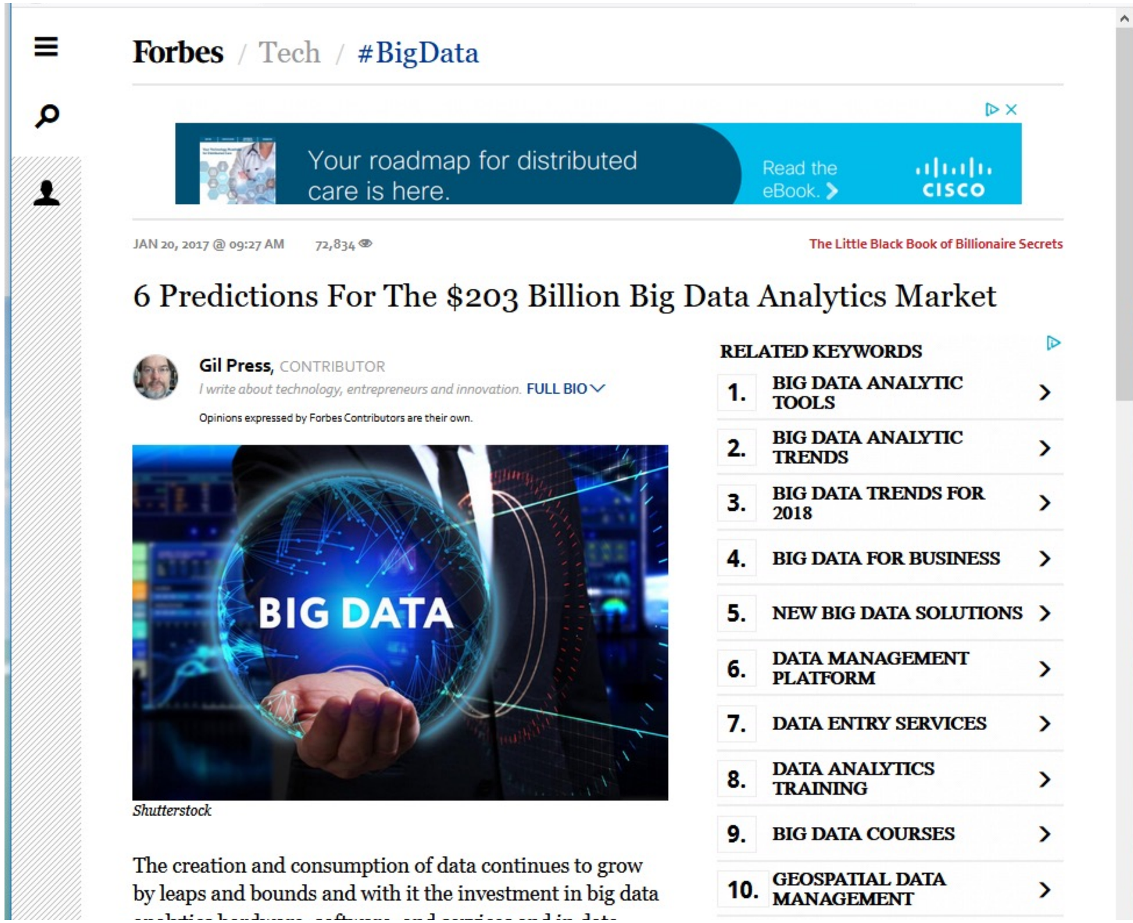
- **June 2013:** Washington Post is calling Obama "**The Big Data President**"



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

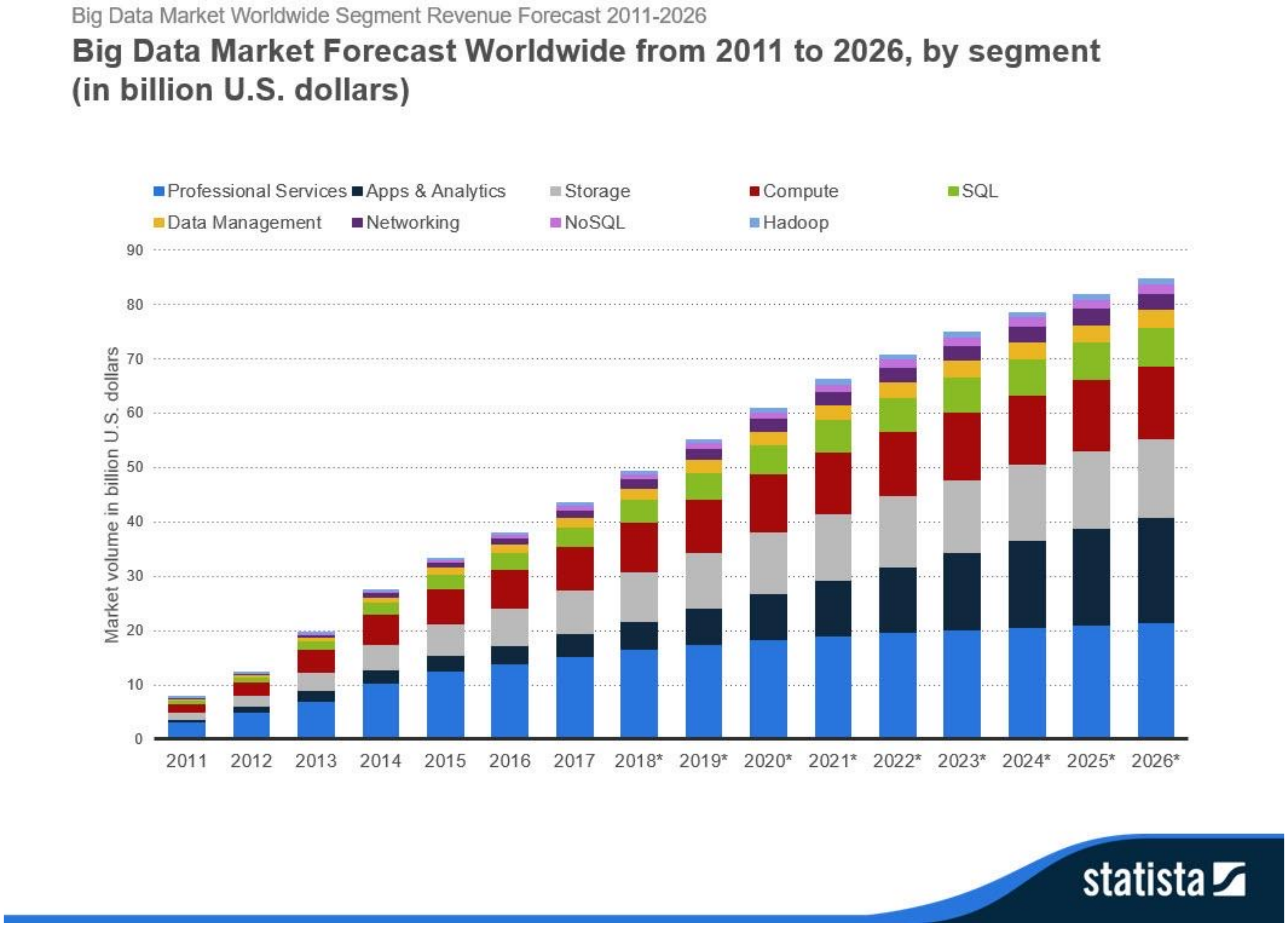
I write about technology, entrepreneurs and innovation. FULL BIO

Opinions expressed by Forbes Contributors are their own.

Shutterstock

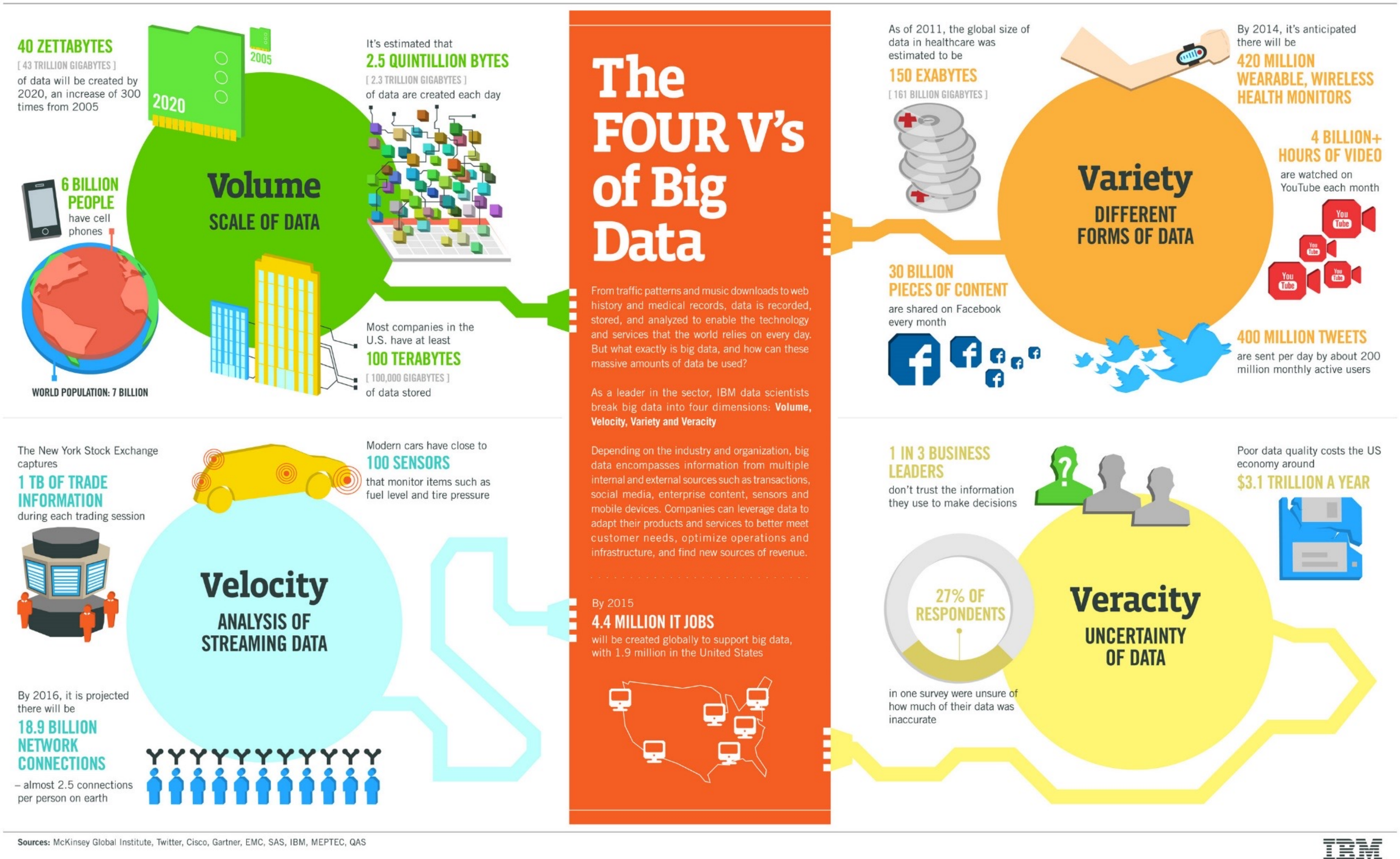
The creation and consumption of data continues to grow by leaps and bounds and with it the investment in big data

Job Market



https://www.techicy.com/5-best-programming-languages-to-watch-out-in-2019-for-data-science.html

~~Three~~ Four V's of Big Data



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

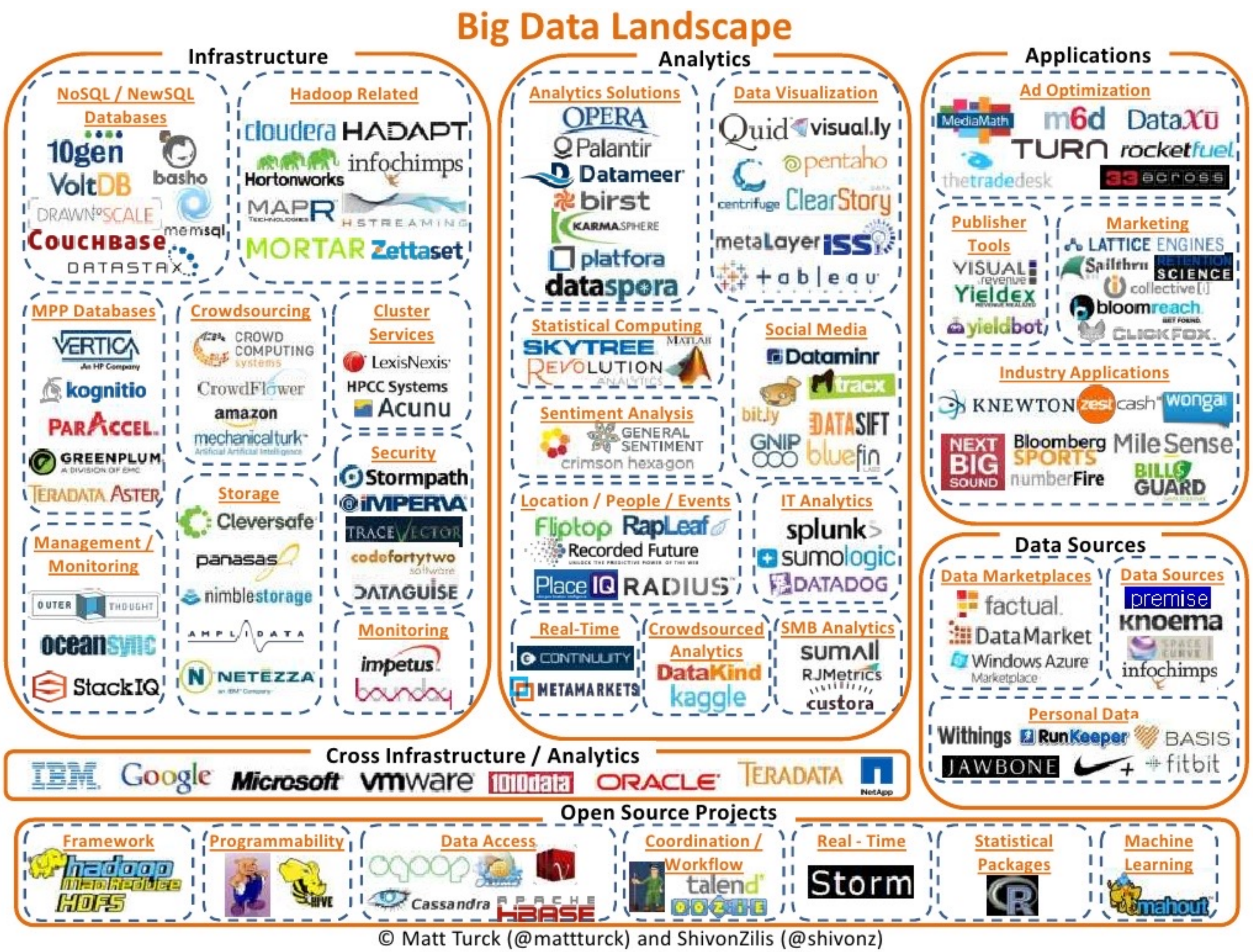
Big Data Applications

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

Publicly Available Datasets

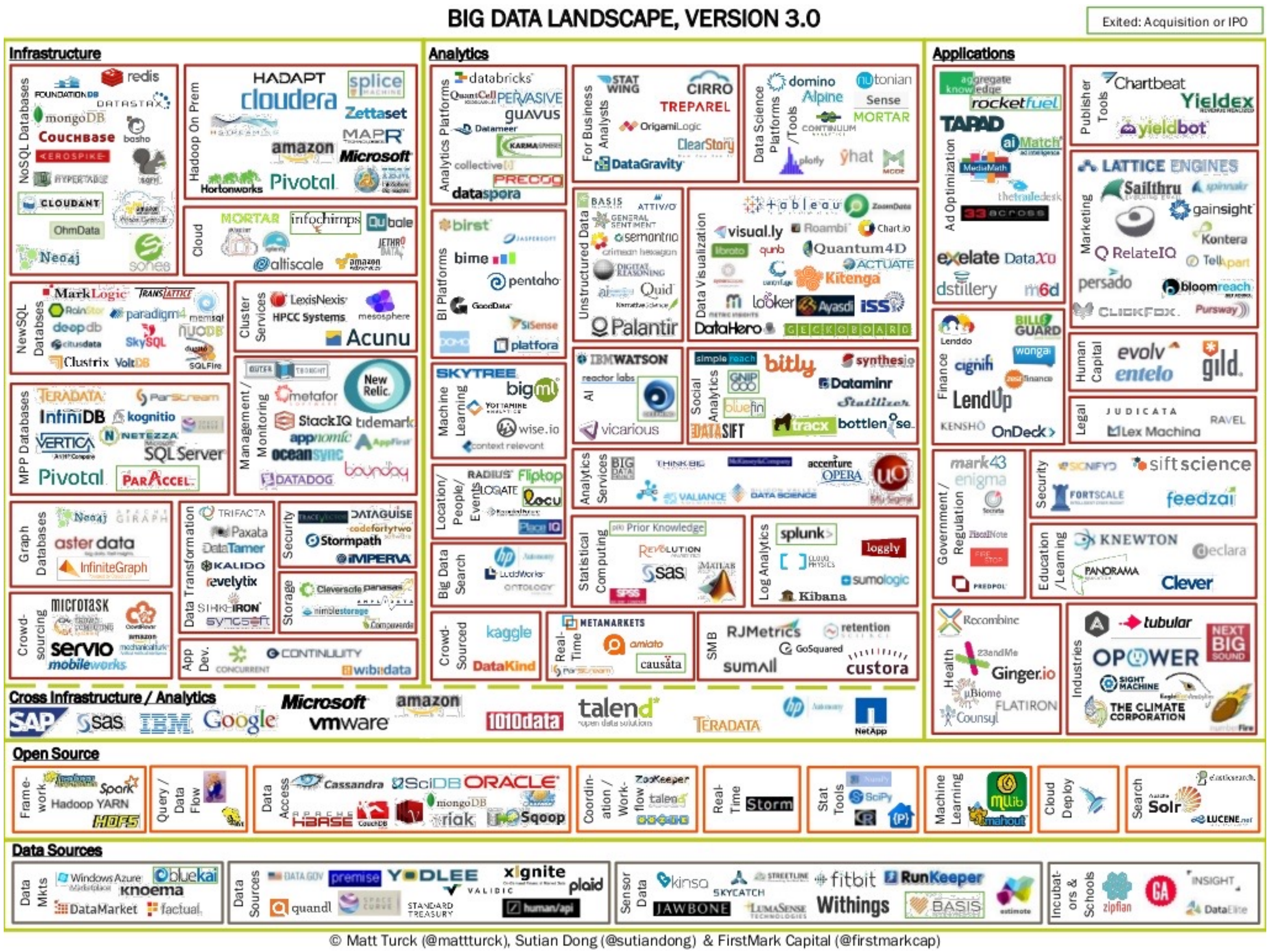
- Data.gov
- Data.gov.uk
- UCR STAR [https://star.cs.ucr.edu]
- Twitter Streaming API
- GDELT [http://www.gdeltproject.org/]
- Kaggle.com

Big Data Landscape 2012



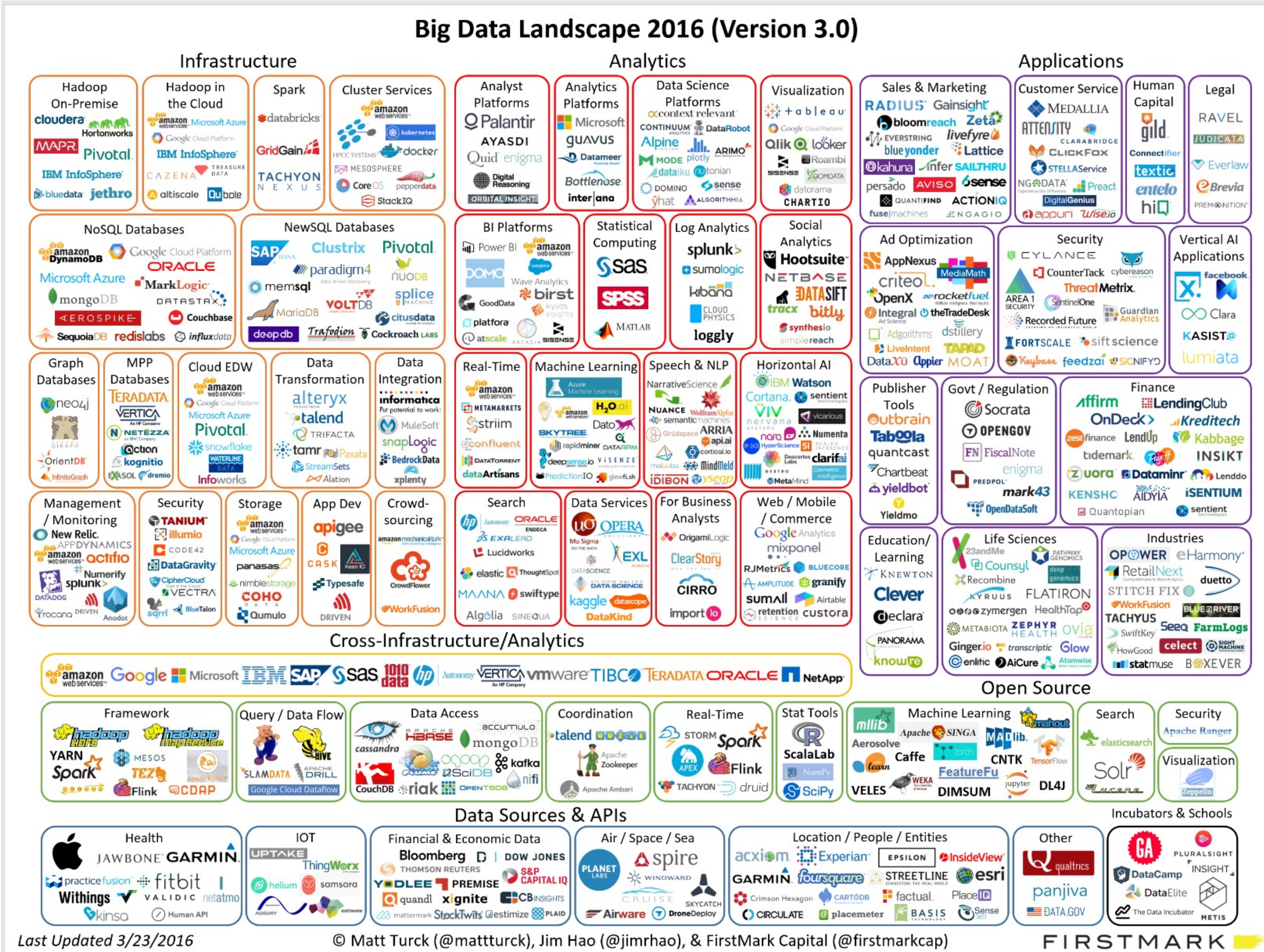
http://mattturck.com/2012/06/29/a-chart-of-the-big-data-ecosystem/

Big Data Landscape 2014



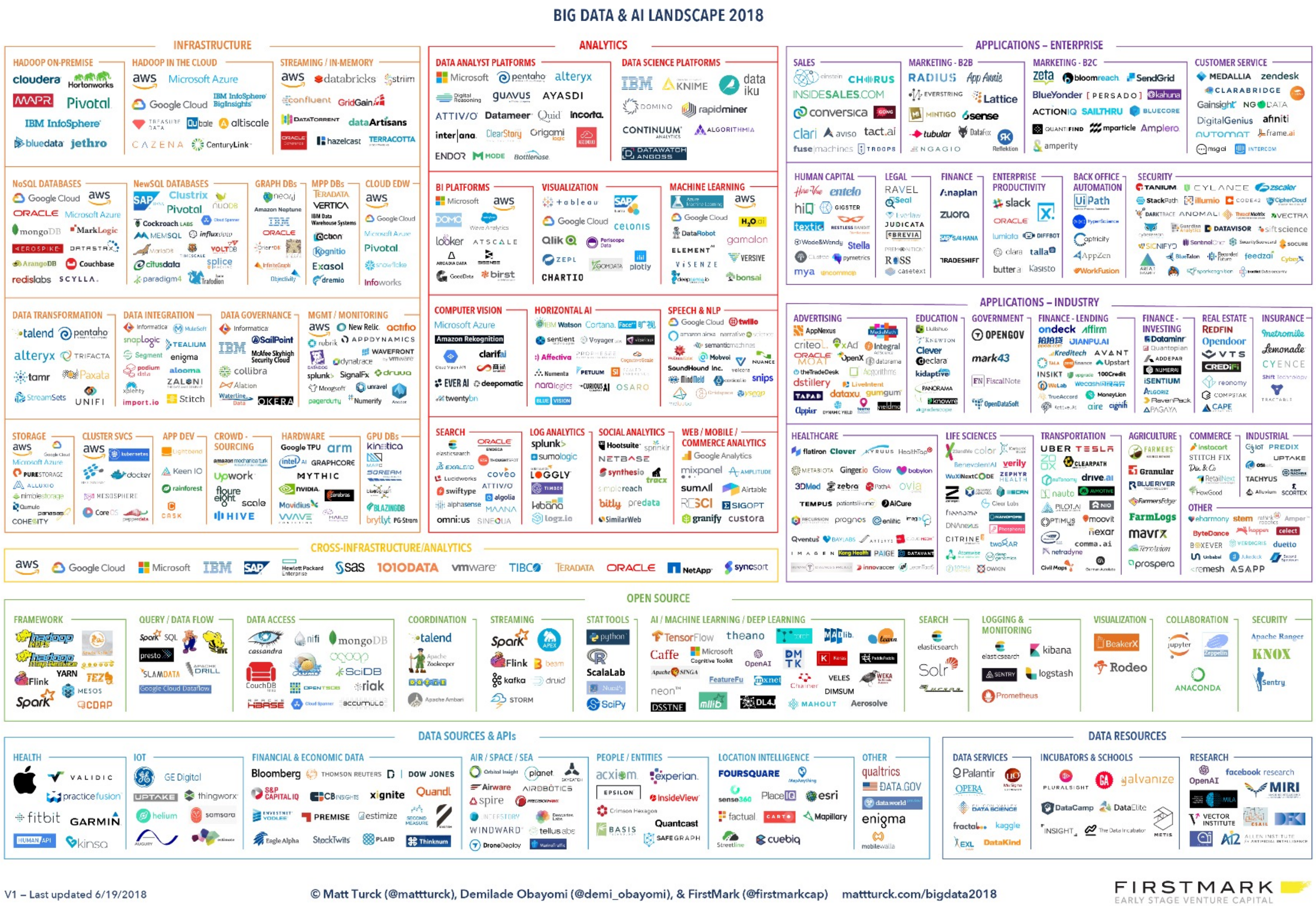
http://mattturck.com/2014/05/11/the-state-of-big-data-in-2014-a-chart/

Big Data Landscape 2016

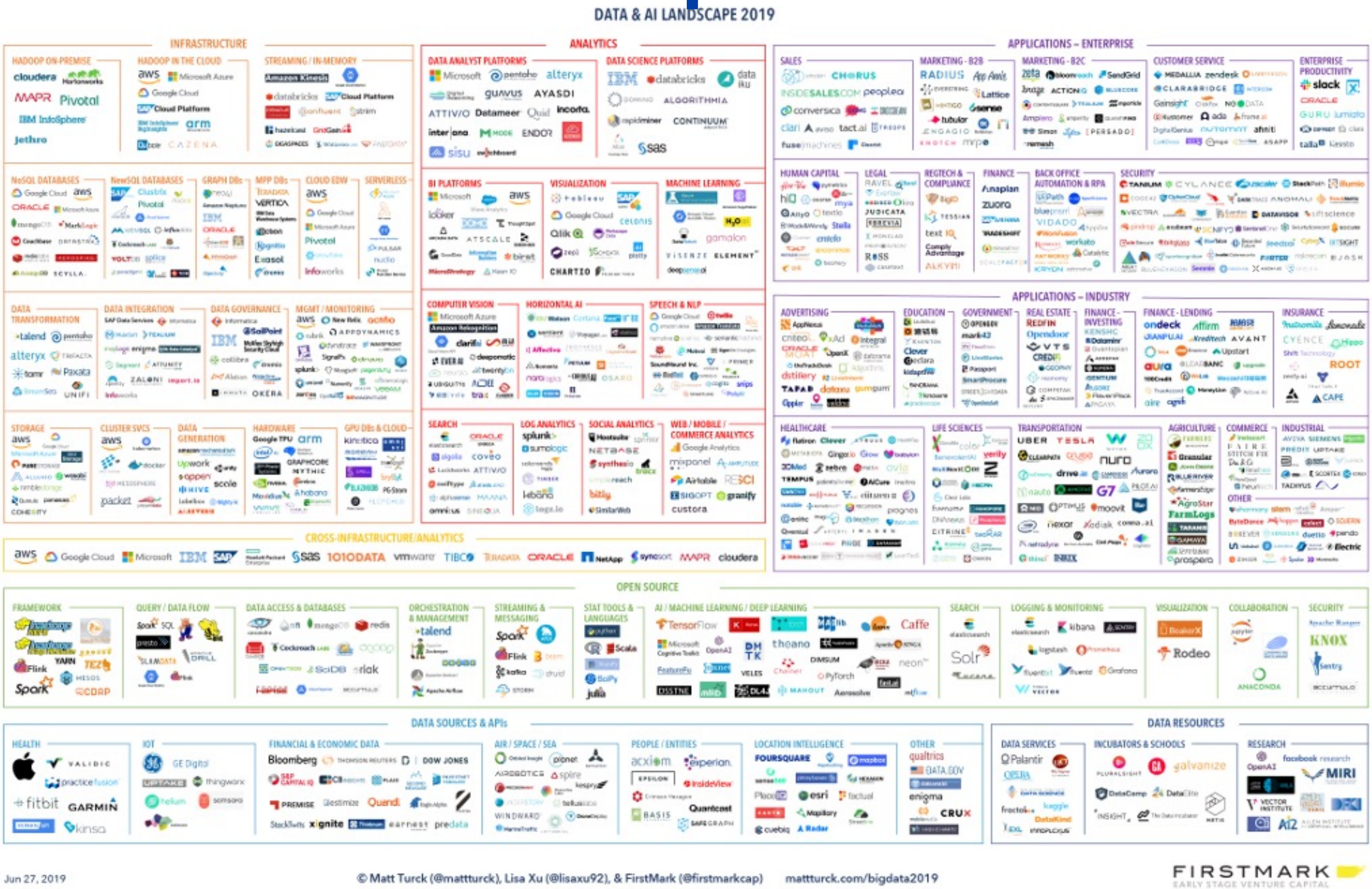


http://mattturck.com/2016/02/01/big-data-landscape/

Big Data Landscape 2018



Data & AI Landscape 2019



Components of Big Data

Components of Big Data

Big-data Libraries

MLlib (Machine Learning), GraphX, Visualization

High-level Languages

SparkSQL, Pig, SQL++, HiveQL

Distributed Computing

MapReduce (Hadoop and Google), Resilient Distributed Dataset (Spark), Hyracks (AsterixDB)

Big Data Distributed Storage

Hadoop Distributed File System, Cloud storage systems (Amazon S3 and Google File System), Key-value stores

Cloud Services

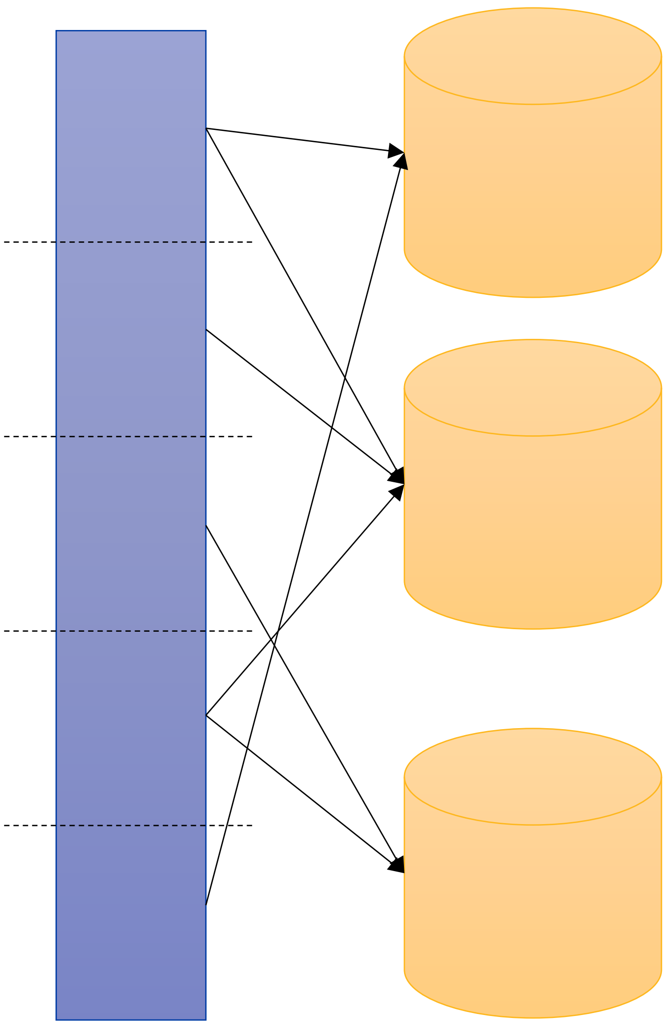
Amazon Web Services, Microsoft Azure, and Google Cloud Platform

Coordination/Cluster Management

Oozie, Yarn, Kubernetes

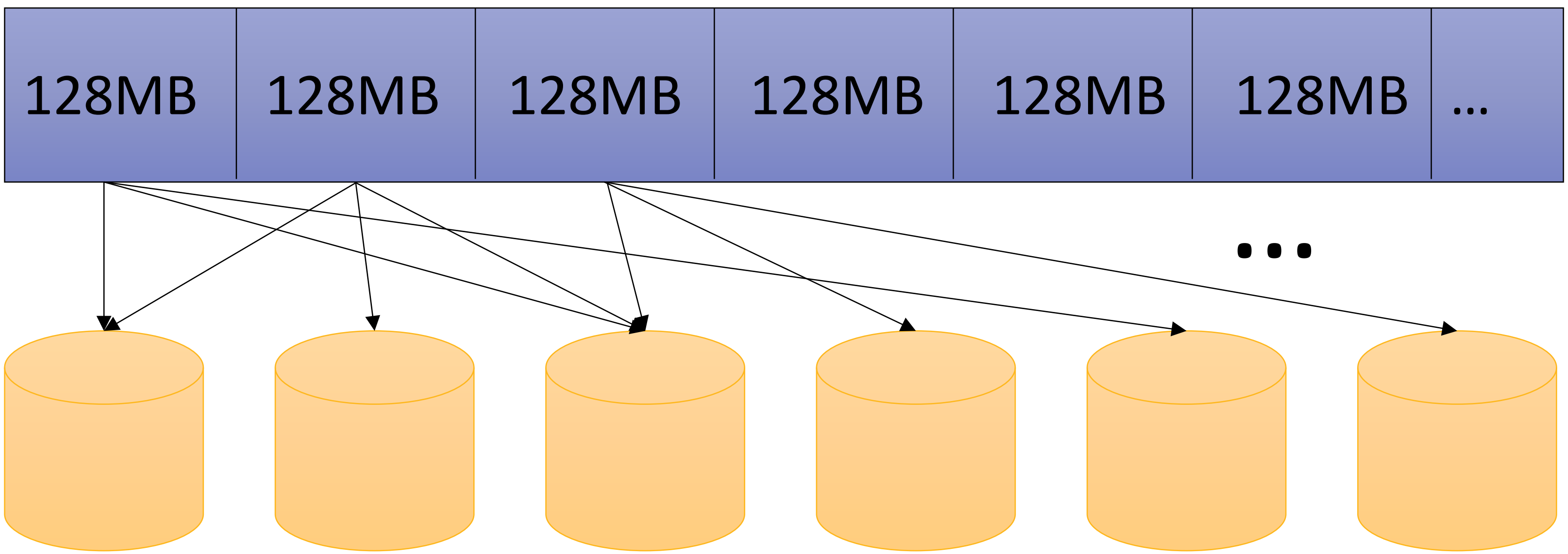
Storage of Big Data

- 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

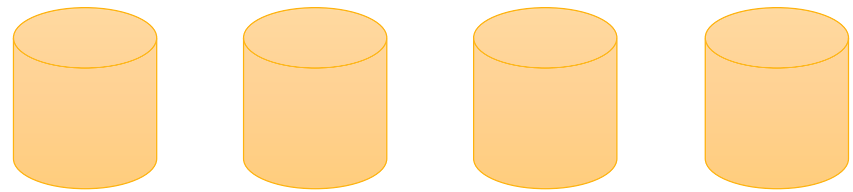


Indexing

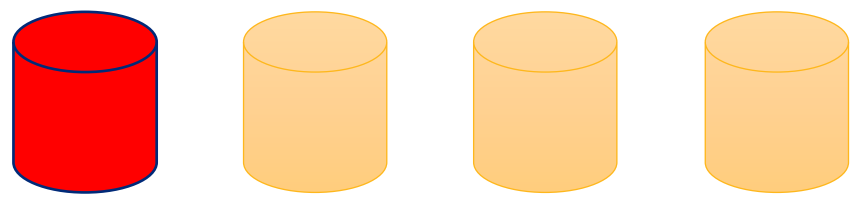
- Data-aware organization
- Global Index **partitions** the records into blocks
- Local Indexes organize the records in a partition
- Challenges:
  - Big volume
  - HDFS limitation
  - New programming paradigms
  - Ad-hoc indexes

Fault Tolerance

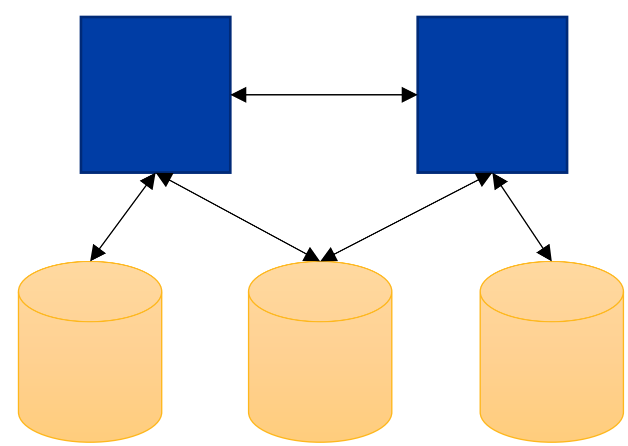
- Replication



- Redundancy



- Multiple masters

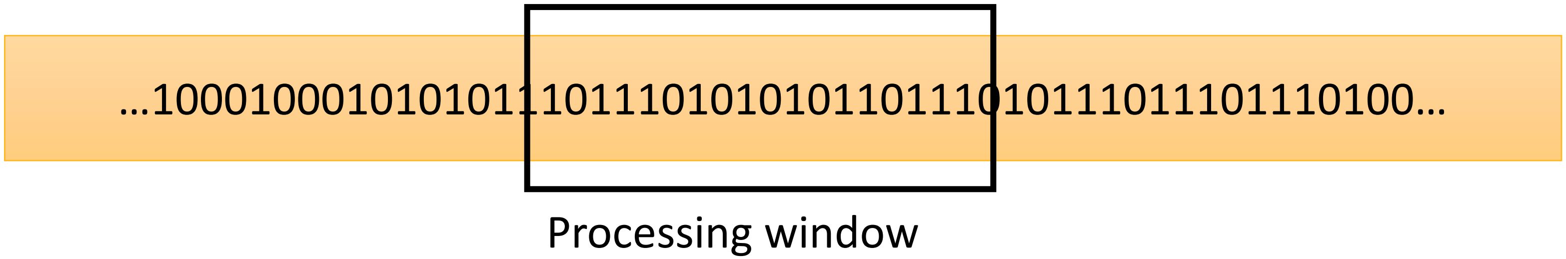


Key-value Stores

ID	Name	Email	...
1	Jack	jack@example.com	
2	Jill	jill@example.net	
3	Alex	alex@example.org	

1	→	Jack	jack@example.com	...
2	→	Jill	jill@example.net	...
3	→	Alex	alex@example.org	...

Streaming



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

Structured/Semi-structured

ID	Name	Email	…
1	Jack	jack@example.com	
2	Jill	jill@example.net	
3	Alex	alex@example.org	

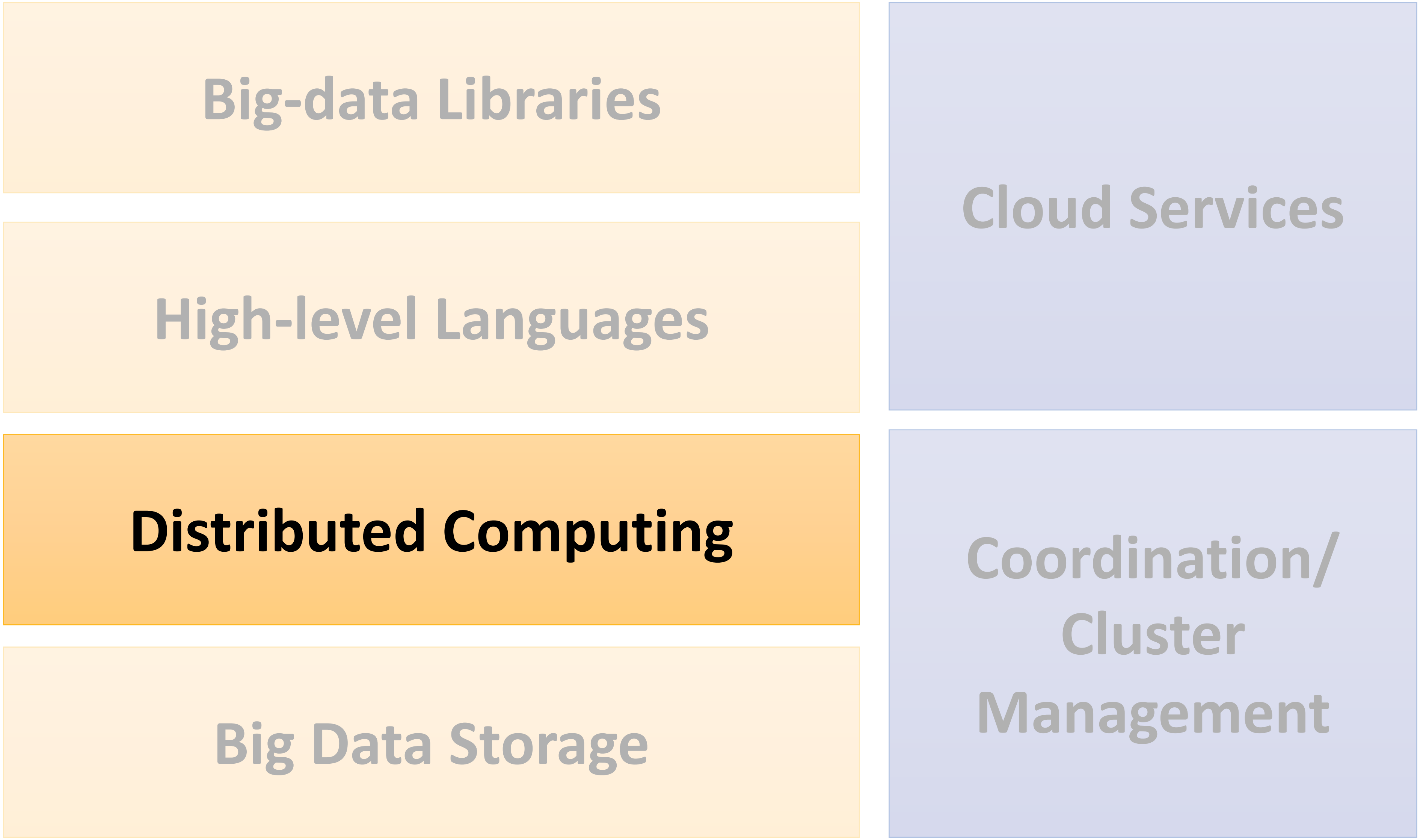
Document 1

{ “id”: 1, “name”:”Jack”, “email”: “jack@example.com”,
“address”: {“street”: “900 university ave”, “city”: “Riv
state: “CA”}, “friend_ids”: [3, 55, 123]}

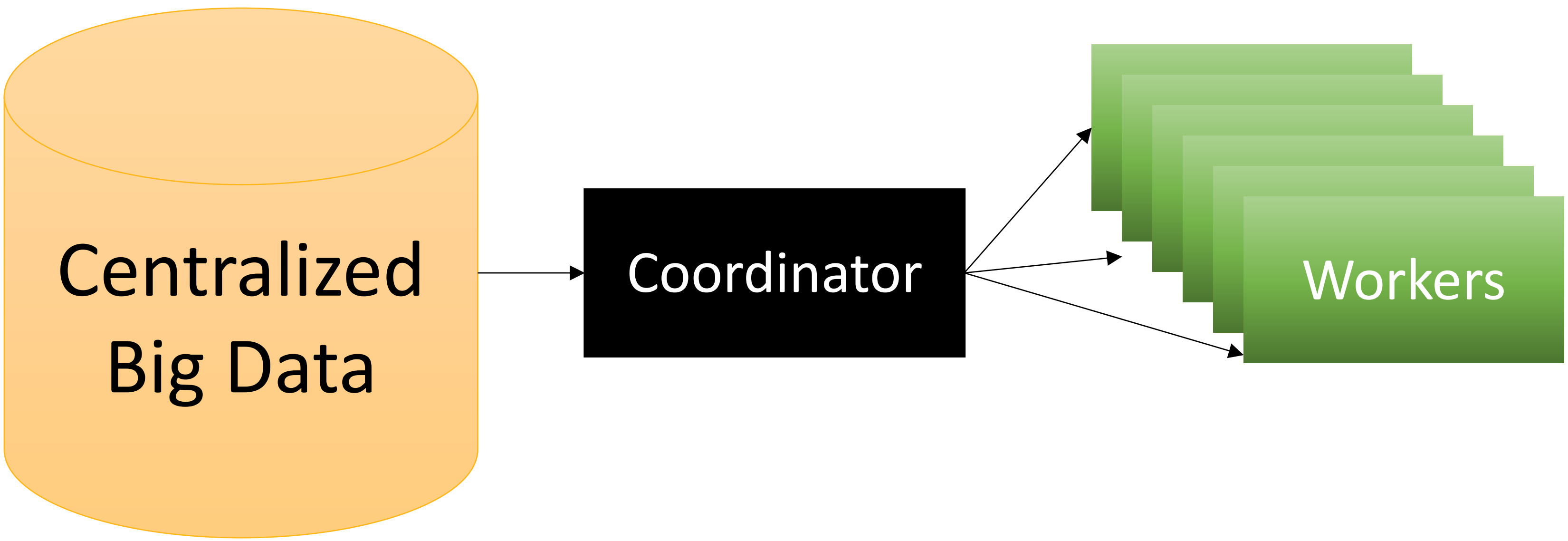
Document 2

{ “id”: 2, “name”: “Jill”, “email”: “jill@example.net”, “hobbies”: [“hiking”, “cooking”]}

Distributed Computing



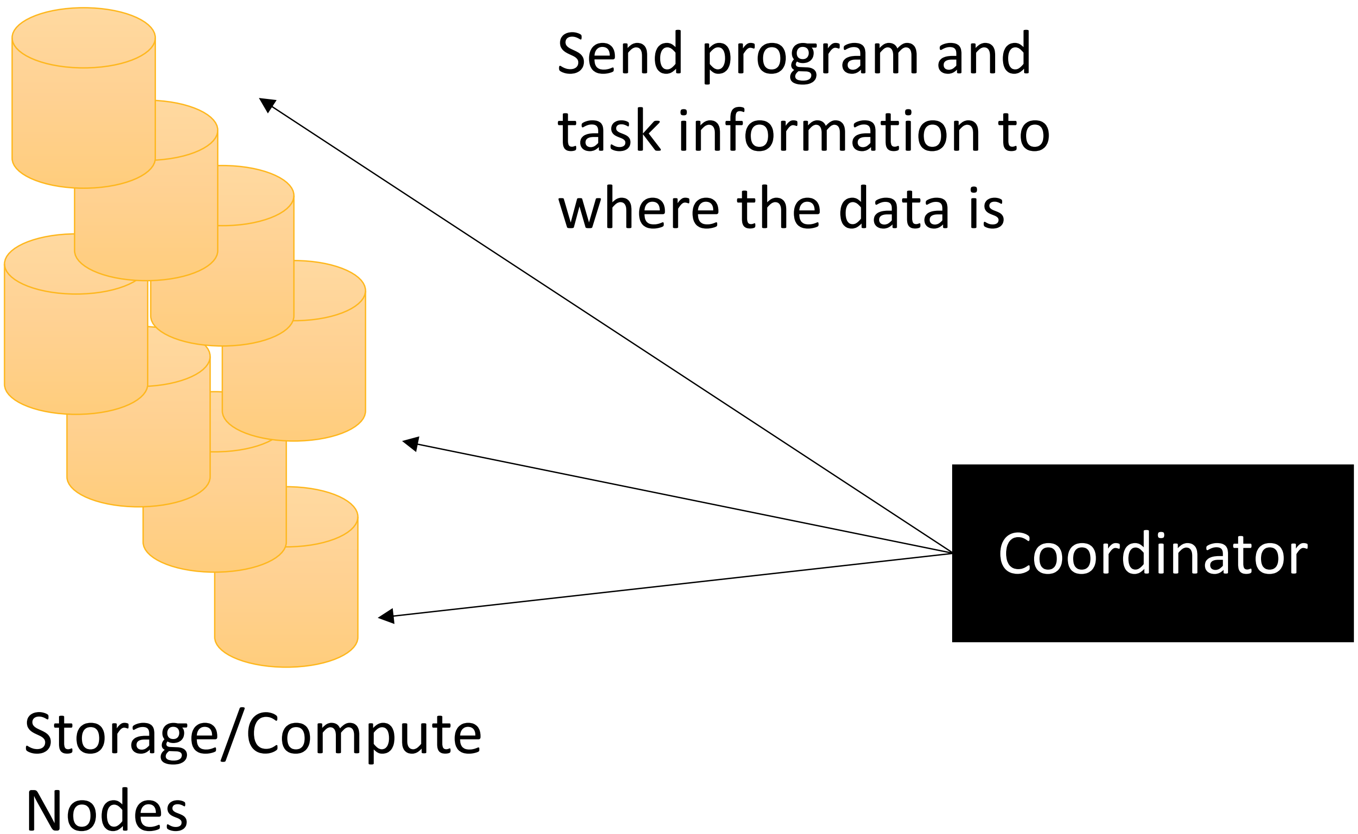
Traditional Distributed Computing



Ship data to computation paradigm

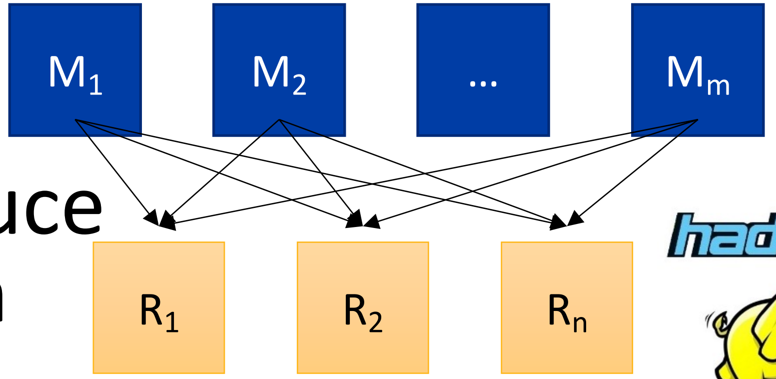
e.g., High performance computing (HPC)

Big-data Computing

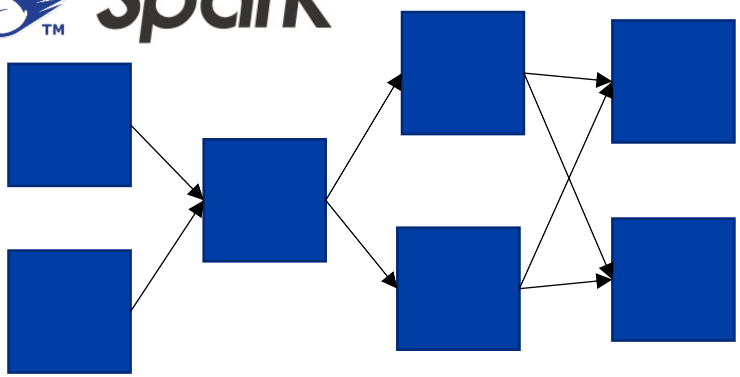


Ship compute to data paradigm

Task Execution

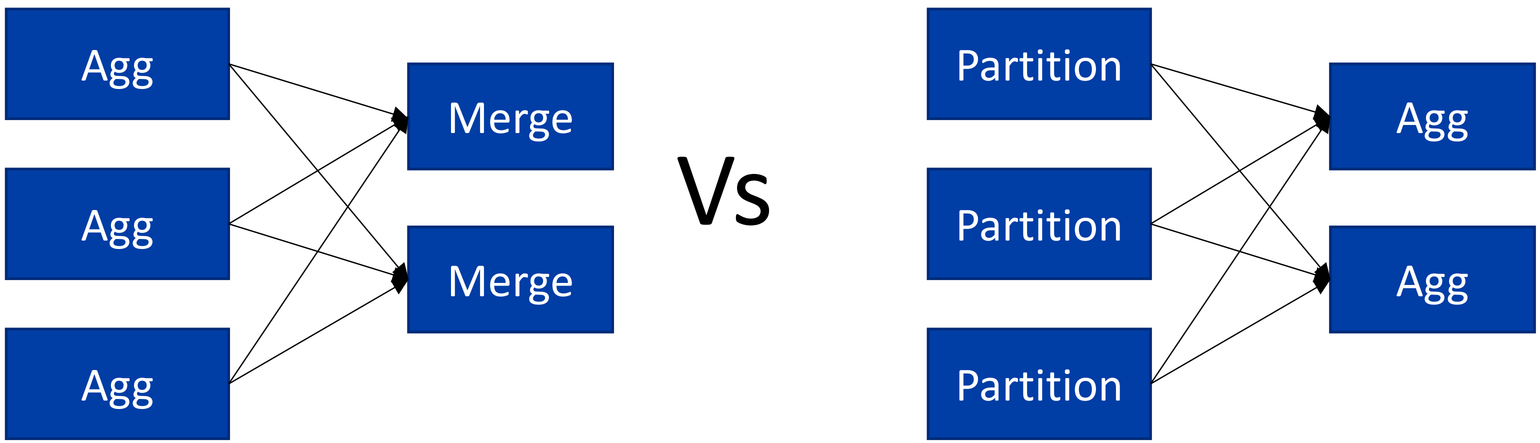


- MapReduce
  - Map-Shuffle- Reduce
  - Resiliency through materialization
- Resilient Distributed Datasets (RDD)
  - Directed-Acyclic-Graph (DAG)
  - In-memory processing
  - Resiliency through lineages
- Hyracks
- Stragglers
- Load balance



Query Optimization

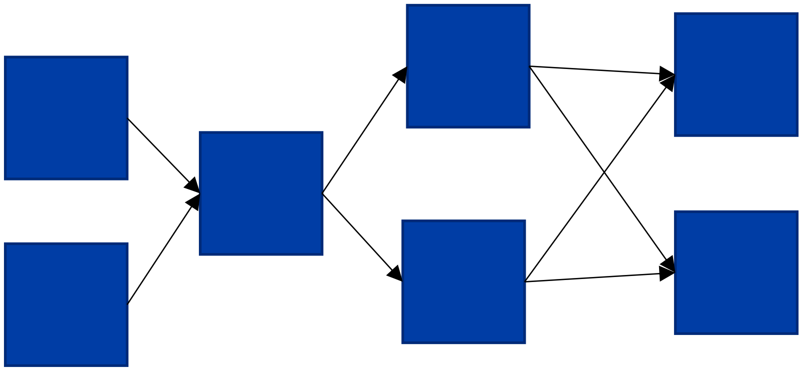
- Finding the most efficient query plan
- e.g., grouped aggregation



- Cost model (CPU – Disk – Network)

Provenance

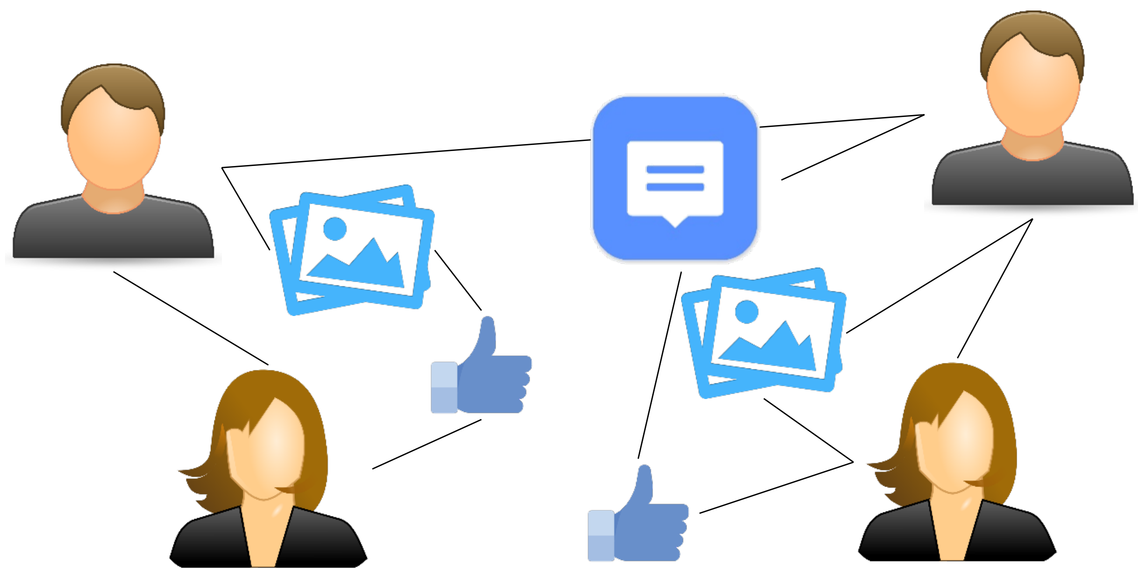
- Debugging in distributed systems is painful



- 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



Structured Data Processing

- There is a need for processing structured and semi-structured data
- Let the big-data system know about the structure of the data and processing
- Allow the system to optimize query processing
- Examples: Algebricks, SparkSQL, and Pig

Hadoop Ecosystem

Apache Ambari

Administration

Pig

HIVE

mahout

MapReduce Query Engine

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

Spark Ecosystem

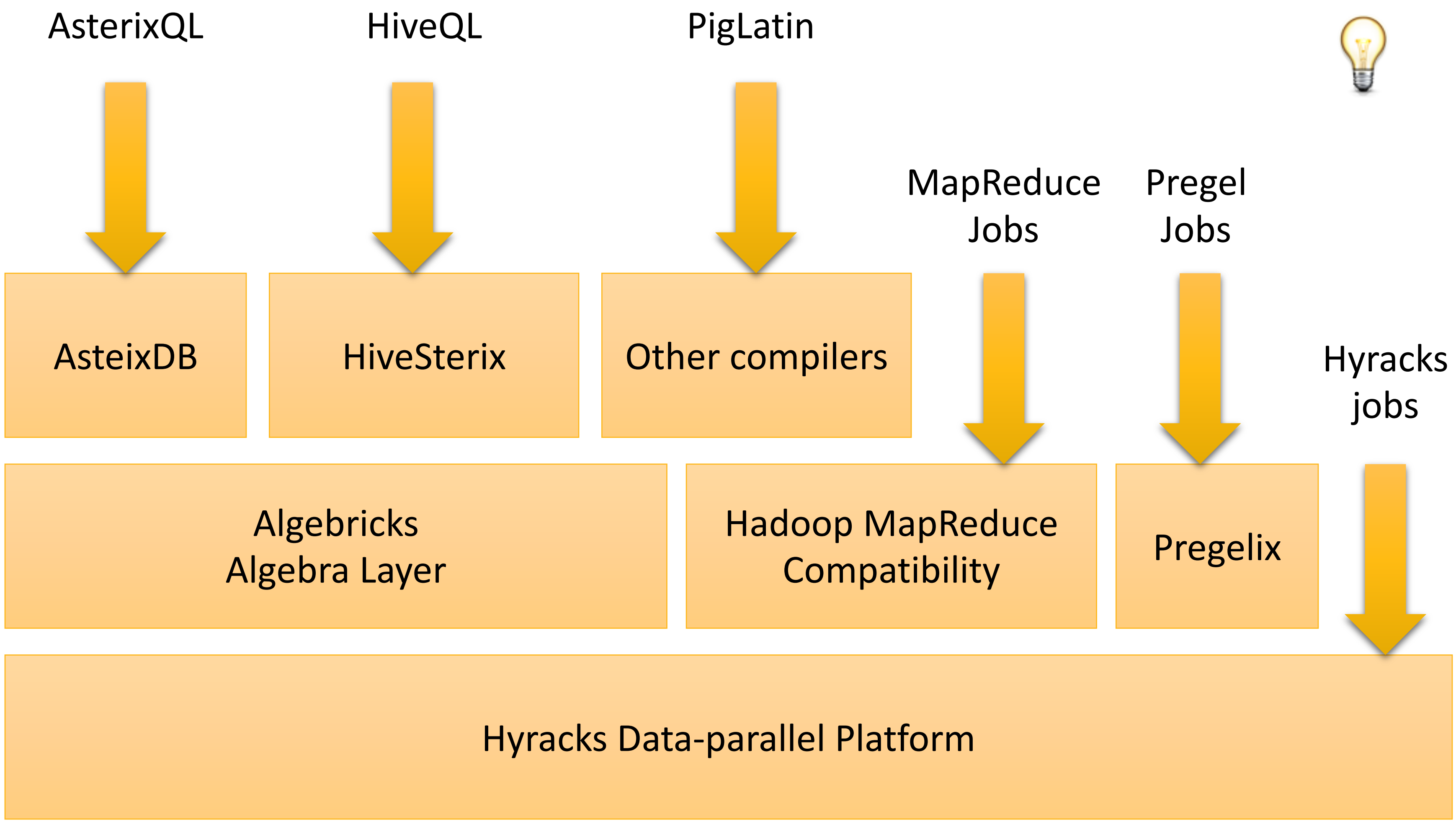
Spark SQL | MLlib | GraphX | SparkR | Spark Streaming

Data Frames

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

Yet Another Resource Negotiator (YARN) | Kubernetes

Hadoop Distributed File System (HDFS)


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