(Large Scale Computing)

# Data Science Architecture

## Introduction

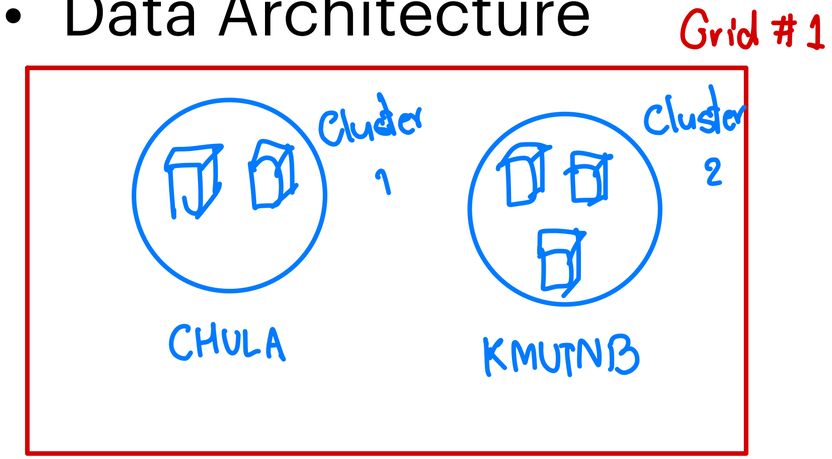
เน้นสร้างเครื่อง สร้างระบบ

ทำไม Big Data Analytic ได้ผลดีแบบนี้

Krerk Piromsopa, Ph.D.

# Agenda

- Introduction
- BigData Revisit
- Data Analytics
  - Scaling Data Analytics
- Data Architecture




Parallel คือ เรียกในเครื่องเดียว

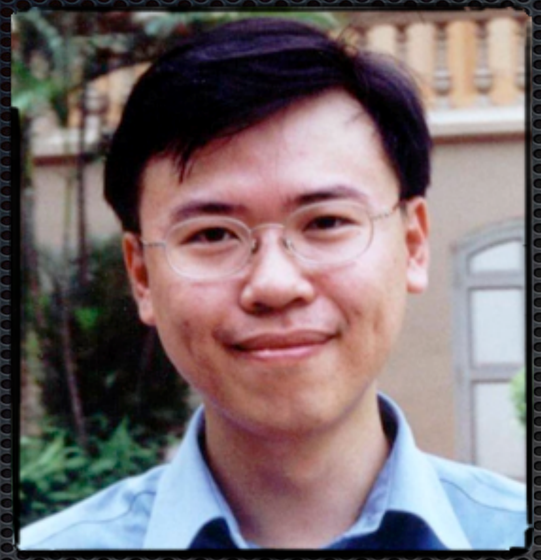
หลุดจากเครื่องนี้ไปเครื่องนึงเรียก Distributed

โดยส่งผ่าน network

1. Grid หลายเครื่องช่วยกันทำงาน แต่อยู่คนละที่
2. Cluster หลายเครื่องอยู่ใน DC วงเดียวกัน ส่งผ่าน switch เดียวกัน

# Who is in charge?

**Mr. Grid**
Data Processing



**Mr. Architecture**
Architecture

**Mr. Cluster**
Data Storage (NoSQL)

# Big Data

## 3/4/5 V(s)

How is Big Data?

Big Data ≠ Analytic with Excel

- Volume → เกินความสามารถที่ประมวลผลใน 1 เครื่อง
- Velocity → data ไม่ยาว แต่เกิดความเร็วสูง ต้องคิดนอกกรอบ
- Variety → data มีความหลากหลาย เช่น เสียงร้องไห้ เสียงดู social media ของผู้ซื้อ เอา text + image
- Veracity?
- Value?

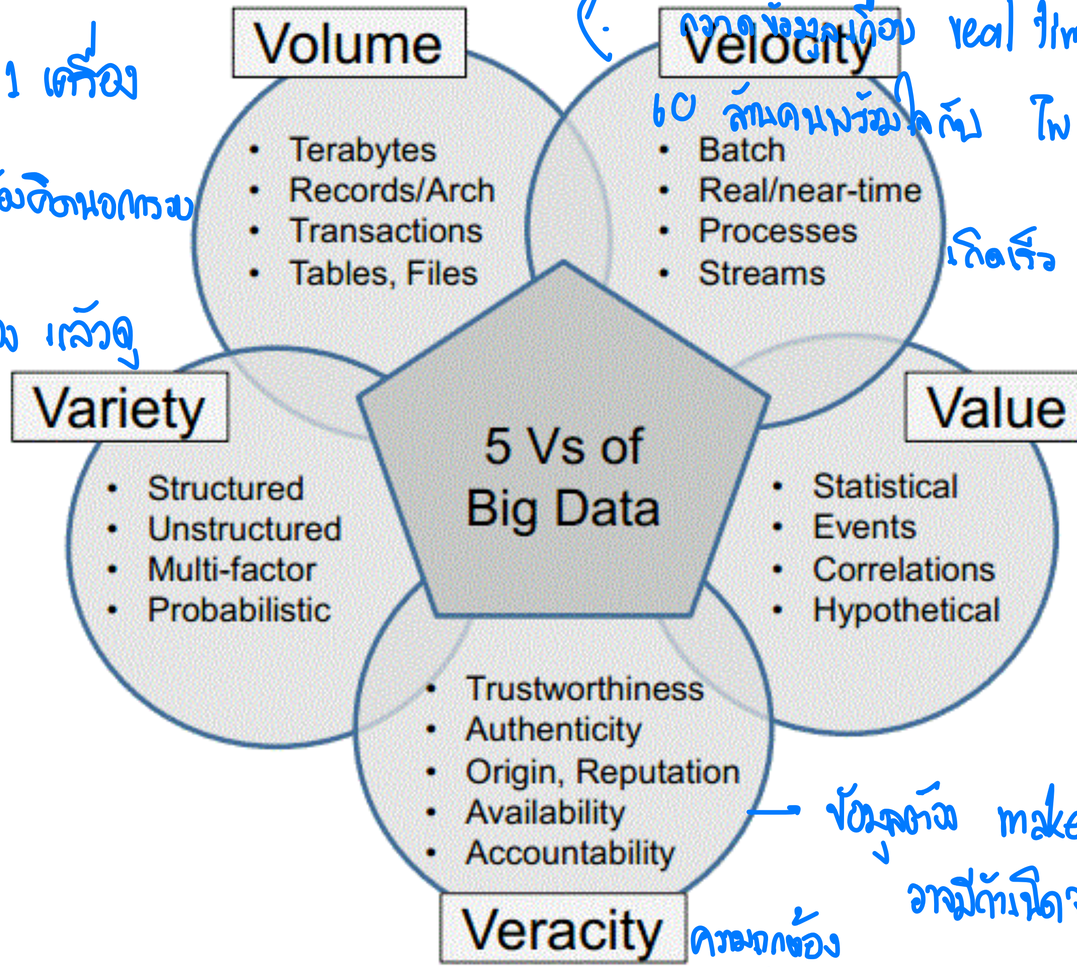
ใช้ประโยชน์ต่อไป

3 V อาจไปได้ข้อมูลดีเสมอไป

เช่น ข้อมูลแข่ง Super Bowl ใน Twitter ?

ดูข้อมูลเกือบ real time

60 ล้านคนพร้อมใจกัน Tweet, กด trade หุ้น เกิดเร็ว




ต้องทำให้งานได้, เกิด business

ข้อมูลต้อง make sence

อาจมีต้นกำเนิดจาก Boston MIT school

ความถูกต้อง

# What will we learn?

เอาเลข 3 คนรวมกัน แล้วเฉลี่ย

ทำ 3 ชุด 3 อาจารย์

- Big Data Architecture สร้างเครื่องแบบไหนถึงดีที่
- Big Data Processing programming model อย่างไร
- Big Data Storage เก็บข้อมูลใหญ่ Petabytes อย่างไร

Let’s assume that we have big data?
How to process them?

# Demand in Computing

- nuclear weapons stewardship,
- cryptology
- climate and environmental modeling
- 3D protein molecule reconstruction
- severe storm forecasting
- design of advanced aircraft
- molecular nanotechnology
- intelligent planetary spacecraft
- real-time medical imaging

Taken from Quest for pete-scale computing

# Solutions

- Cluster ?
- Grid ?

- Volunteer Computing

ระหว่างที่เครื่องเราว่าง จะส่ง project มารันที่เครื่องเรา

# Volunteer Computing

- Desktop Grid เช่น BOINC framework ใช้ implement volunteer
- A type of distributed computing
- Owners donate their computing resources (power and storage) to projects
- Popular projects
  - SETI @ home (Search for Extra-Terrestrial Intelligence) หามนุษย์ต่างดาว, ดาวดวงใหม่
    ถ้าเจอดาวดวงนั้นในเครื่องเรา ให้ตั้งชื่อดาว
    ใช้ processing power
  - Folding @ home (protein folding, molecular dynamics)
  - More....

ในอดีต PS3 GPU แรงกว่าคอมจนนำไปส่งไปยังดาวนอกกลาง เพราะไปทำวิจัยรวด

# Framework for Volunteer Computing

- XGrid (MAC only) - no longer support
- etc...
- Berkeley Open Infrastructure for Network Computing (BOINC)

# *Distributed Computing*

- Volunteer Computing/ Desktop Grid
  - computing power gathered from volunteers' computer.

# BOINC

- Easy to join ?
  - Simple Client/Server model
  - Use GCC, so work with most programming languages
- Native performance
  - Support Mac, Windows, Linux, PS3, etc..

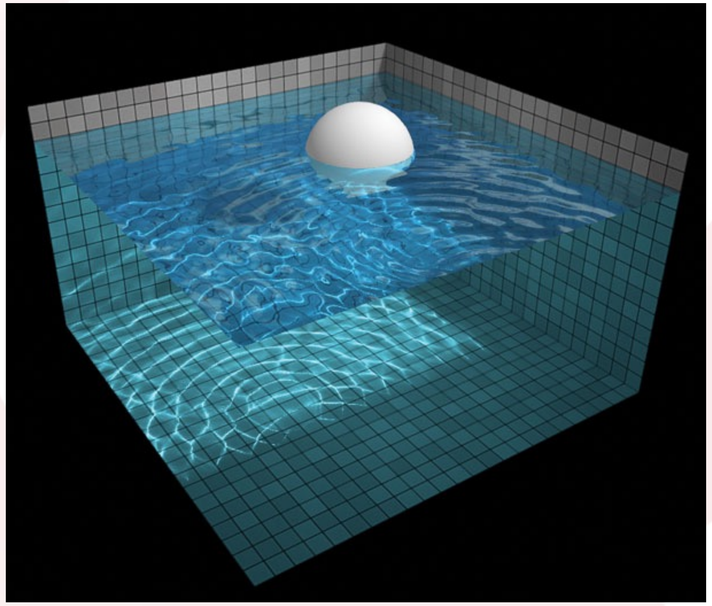
# *Web as Distributed Computing*

- Low barrier of entrance
  - No client software install
  - Easy to participate
  - Large user base
- Platform and OS independent
- Security (Sandbox environment)

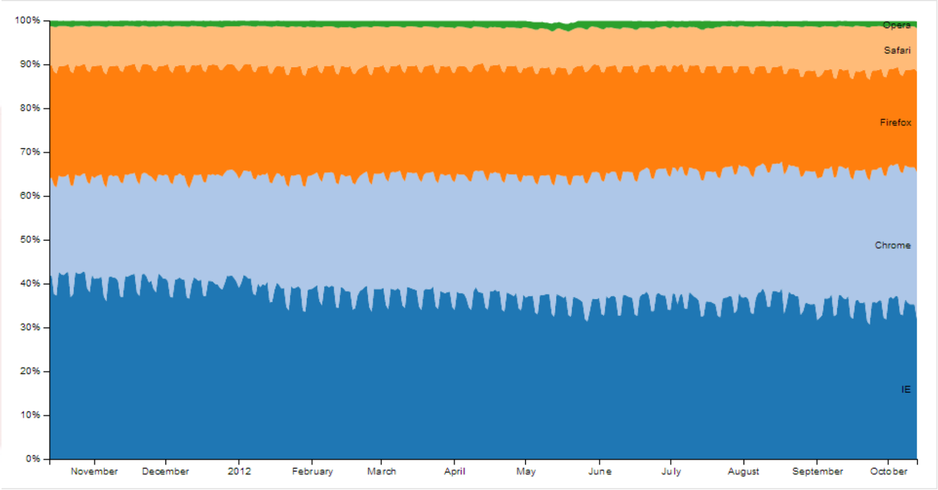
# *Emerging Web Technology*

- Web 2.0 -> HTML5, JavaScript 2.0, WebGL
- Gap between web application and legacy desktop application have been reduced

# *Emerging Web Technology*



physic simulation*



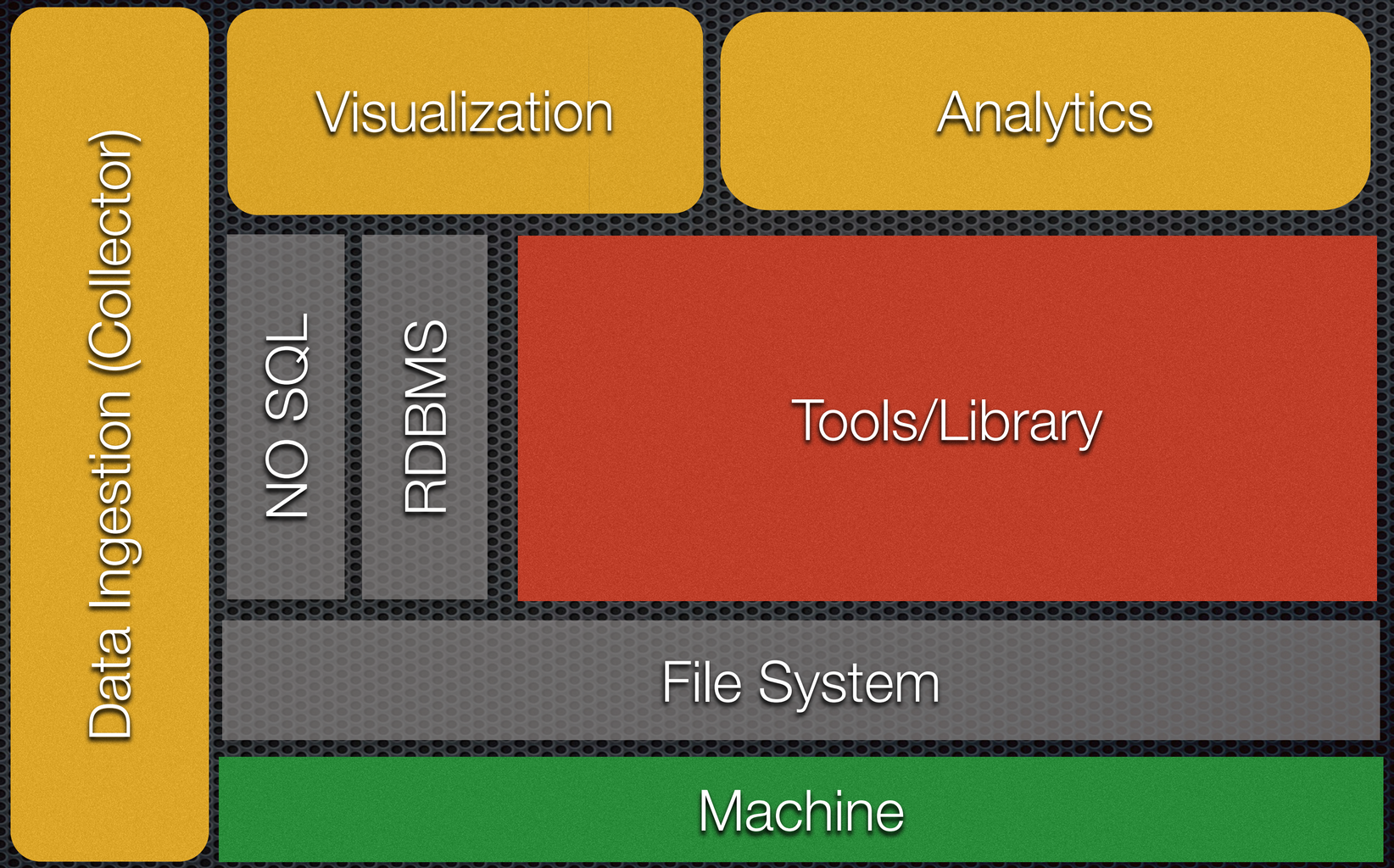

data virtualization**

* http://www.webdesignermag.co.uk/blog/going-3d-with-webgl/
** http://bl.ocks.org/mbostock/3885211

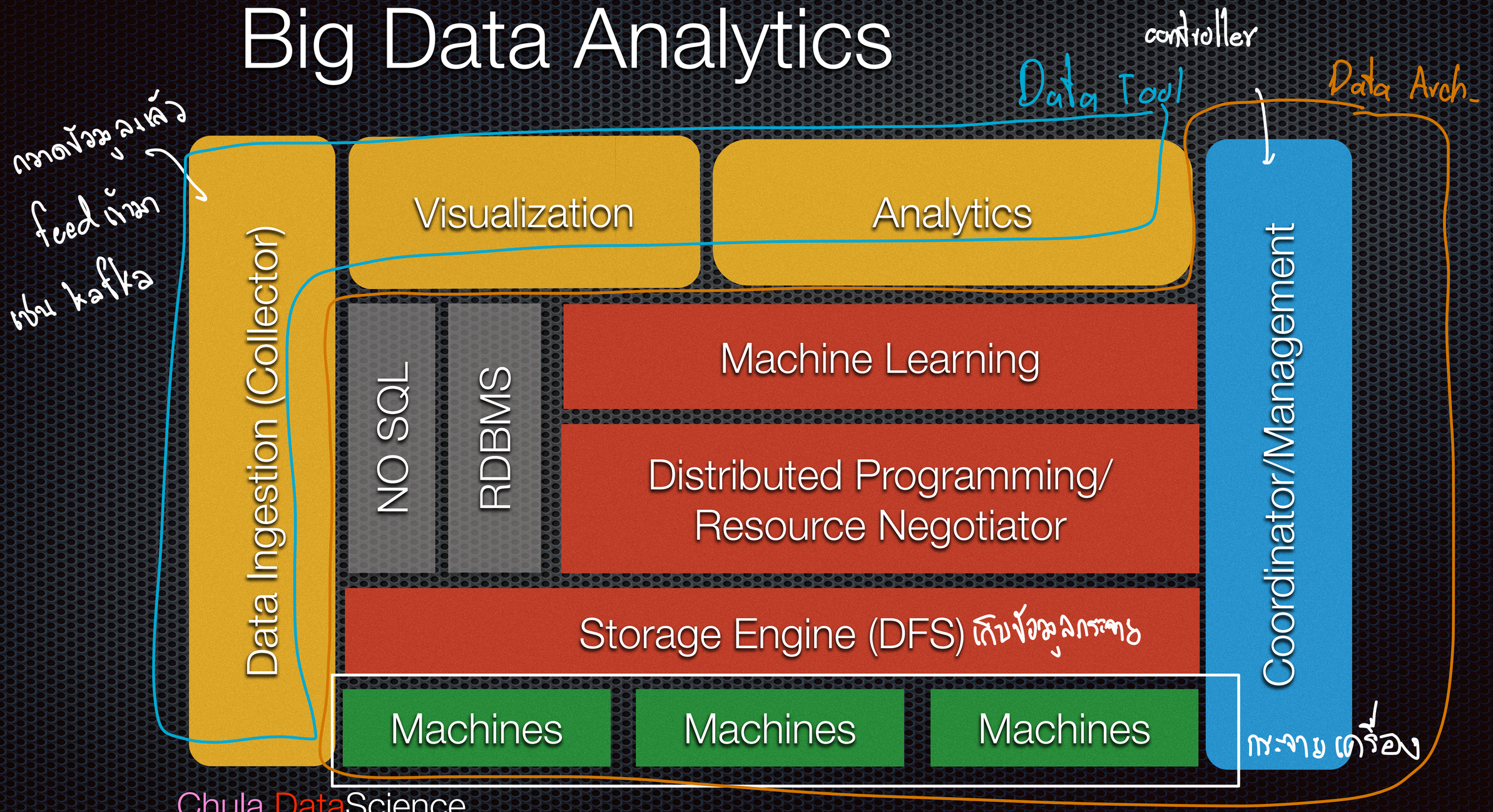
- How about performance?
  - What tools for web application for using as distributed computing platform?
- What is the difference from traditional desktop grid?
  - Tradeoff

# How to scale your data analytics ?

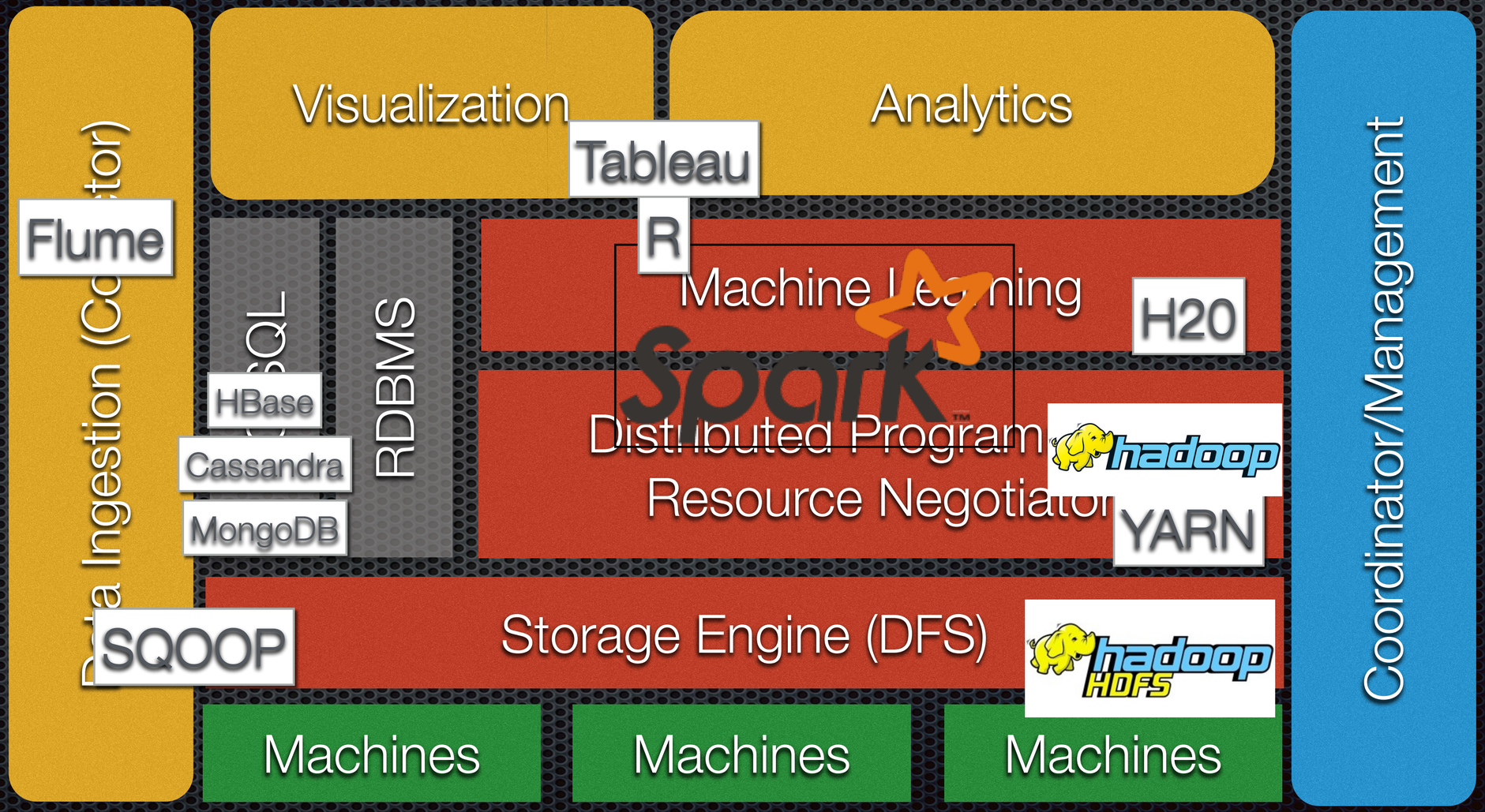
# Data Analytics (classic)



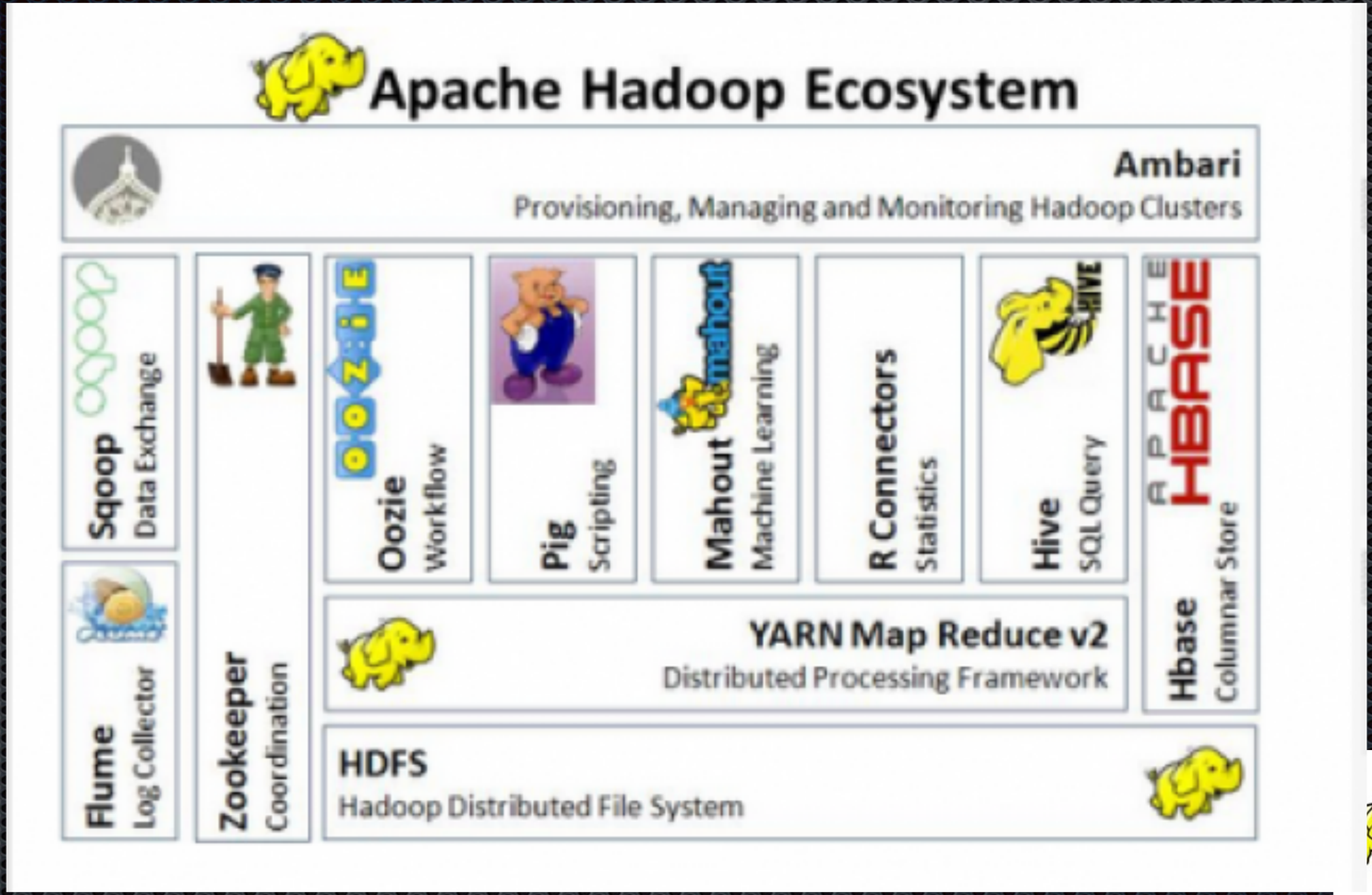
# Big Data Analytics



# Big Data Analytics (ctd.)

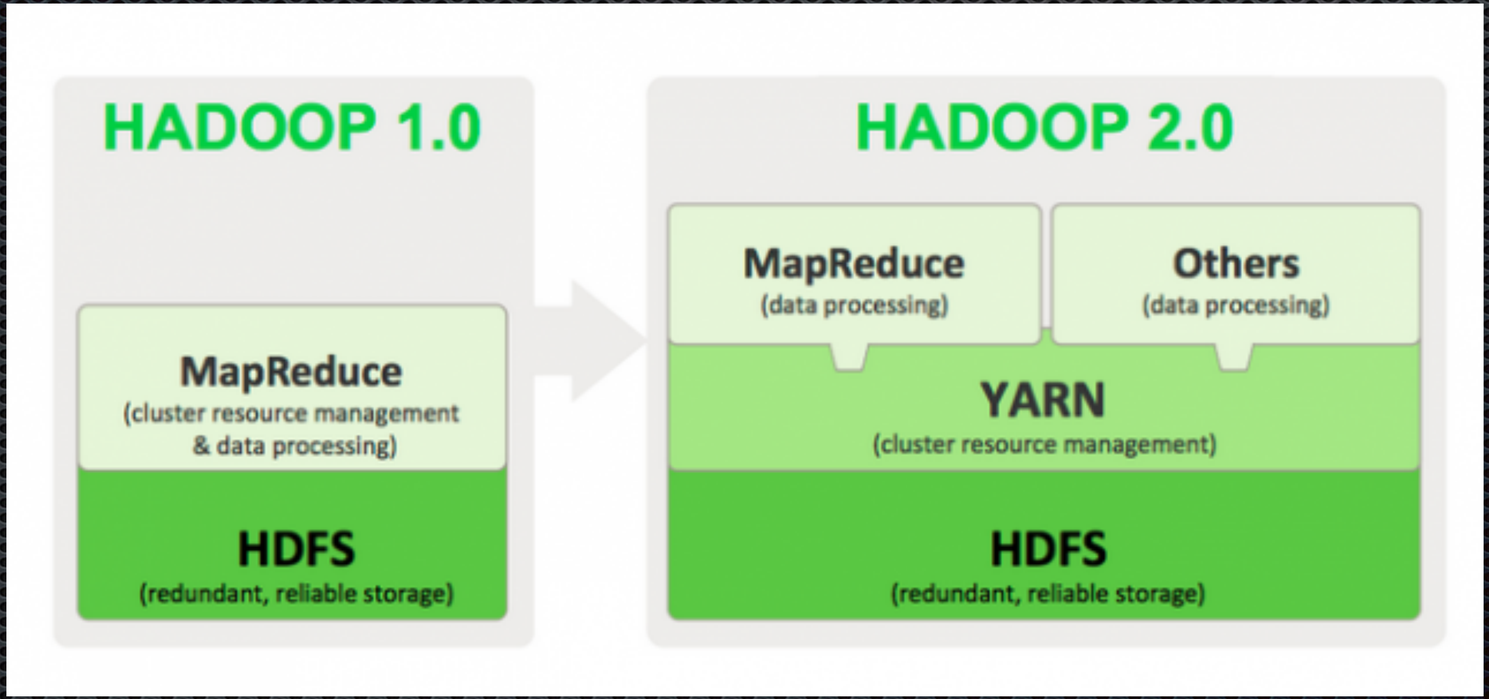


# Hadoop Ecosystem



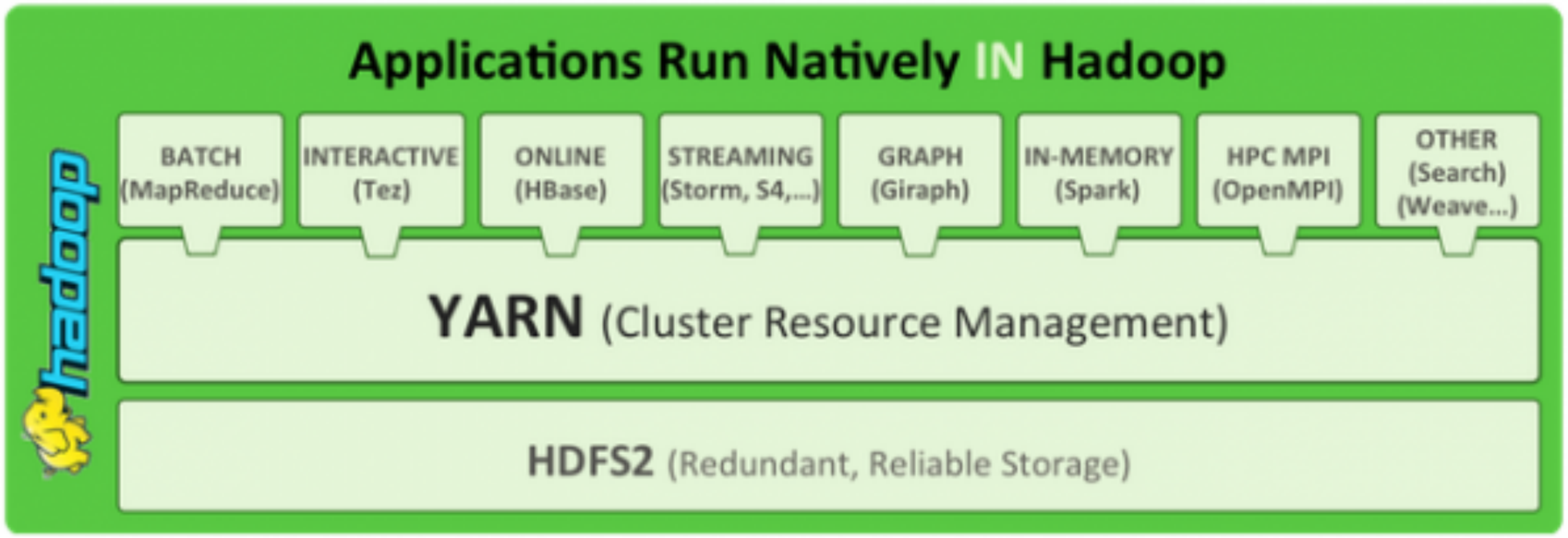

https://opensource.com/life/14/8/intro-apache-hadoop-big-data

# Hadoop 1.0 vs. 2.0




https://opensource.com/life/14/8/intro-apache-hadoop-big-data

# Hadoop YARN

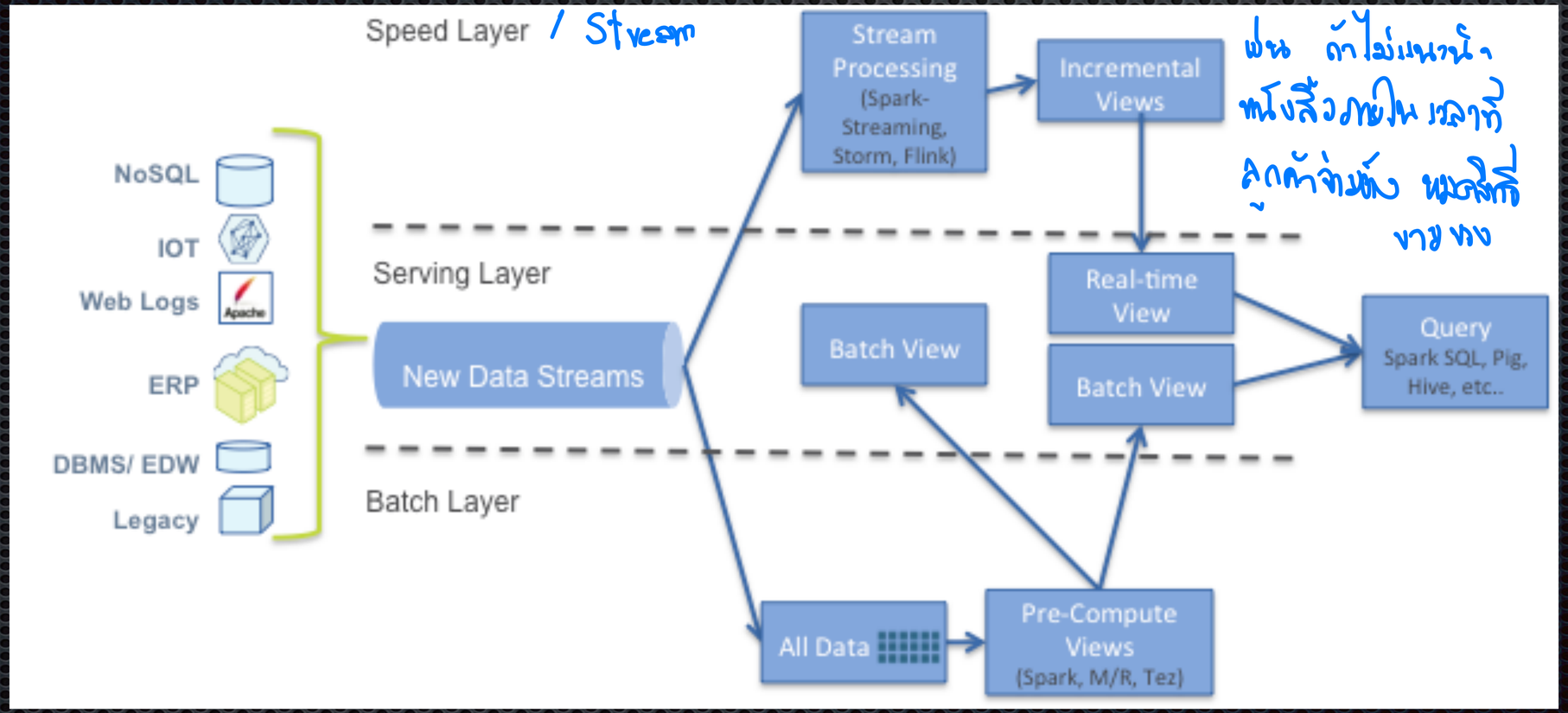



https://opensource.com/life/14/8/intro-apache-hadoop-big-data

# Architecture

# Lambda Architecture

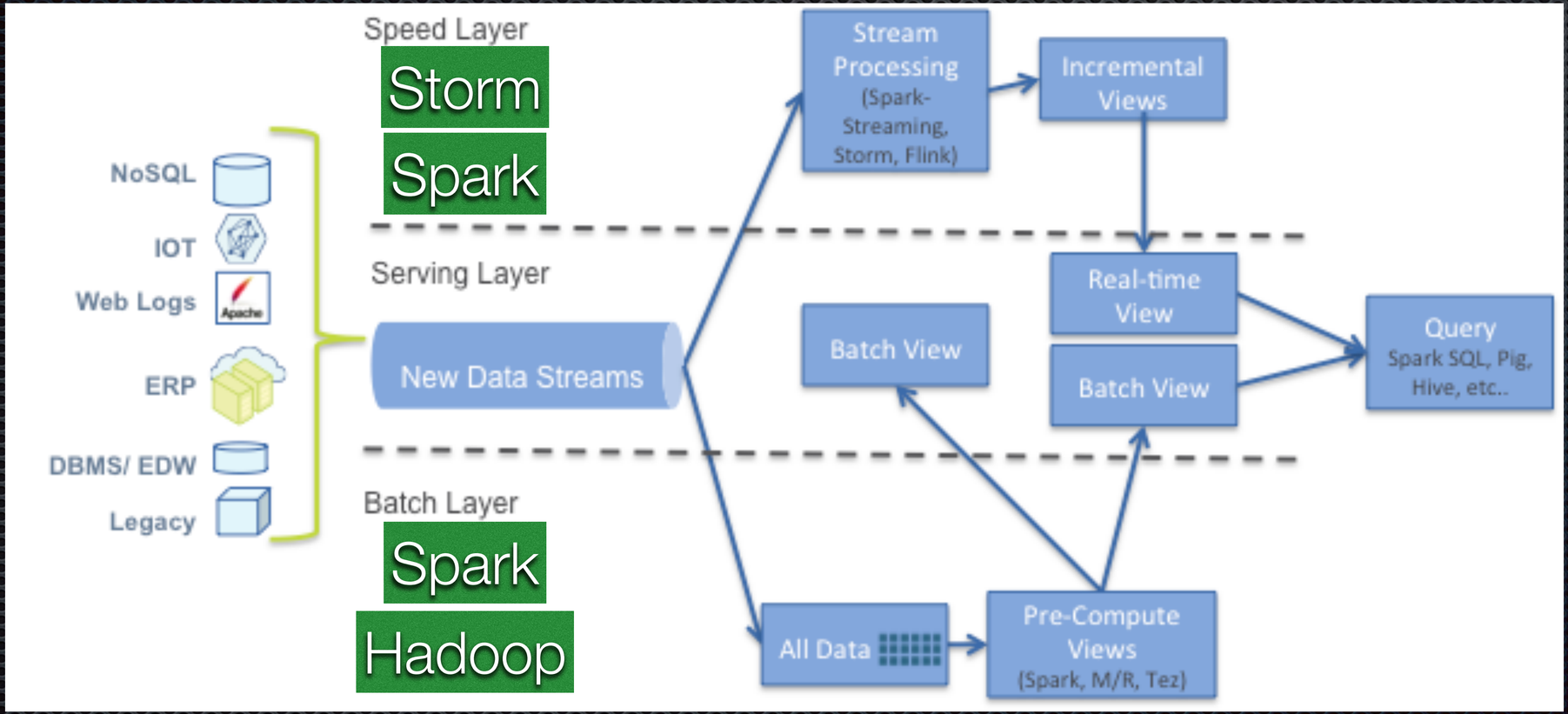
Nathan Marz, 2012




https://www.talend.com/blog/2015/07/15/hadoop-summit-2015-takeaway-the-lambda-architecture

# Lambda Architecture

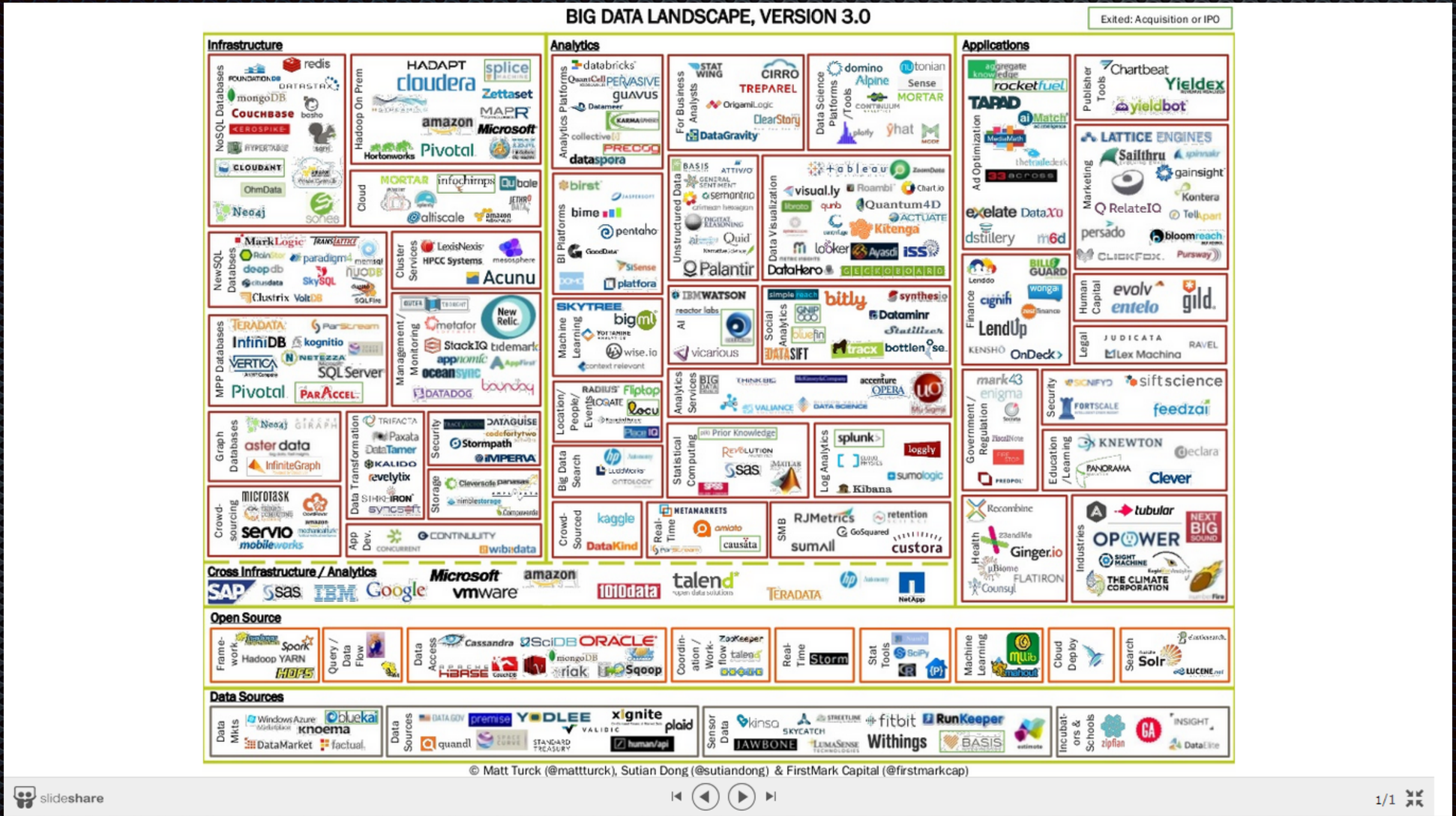
Nathan Marz, 2012




https://www.talend.com/blog/2015/07/15/hadoop-summit-2015-takeaway-the-lambda-architecture

# Big Data Landscape

# Company




Chula DataScience

http://dataconomy.com/understanding-big-data-ecosystem/

# Big Data Landscape 2016 (Version 2.0)

## Infrastructure

### Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

### Hadoop in the Cloud
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, TREASURE DATA, altiscale, qubole

### Spark
databricks, GridGain, TACHYON NEXUS

### Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

### NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

### NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, nuoDB, memsql, splice machine, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

### Graph Databases
neo4j, GIRAPH, OrientDB, InfiniteGraph

### MPP Databases
TERADATA, VERTICA, NETEZZA, Actian, kognitio, EXASOL, dremio

### Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

### Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

### Data Integration
informatica, Put potential to work:, MuleSoft, snapLogic, BedrockData, xplenty

### Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk>, DATADOG, rocana, DRIVEN, Anodot

### Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

### Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

### App Dev
apigee, CASK, Keen IO, Typesafe, DRIVEN

### Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

## Analytics

### Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

### Analytics Platforms
Microsoft, GUAVUS, Dataameer, Bottlenose, inter|ana

### Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine DATA LABS, plotly, ADATAO, MODE, dataiku, nutonian, DOMINO, sense, yhat, ALGORITHMIA

### Visualization
tableau, Google Cloud Platform, Qlik, looker, Roambi, SISENSE, GOOMDATA, datorama, CHARTIO

### BI Platforms
Power BI, amazon web services, DOMO, salesforce, Wave Analytics, birst, GoodData, kyvos insights, platfora, atscale, ARCADIA DATA, SISENSE

### Statistical Computing
sas, SPSS, MATLAB

### Log Analytics
splunk>, sumologic, kibana, CLOUD PHYSICS, loggly

### Social Analytics
Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthesio, simplereach

### Real-Time
amazon web services, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

### Machine Learning
Azure Machine Learning, amazon web services, H2O.ai, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfi.sh

### Speech & NLP
NarrativeScience, ARRIA, NUANCE, semantic machines, Gridspace, api.ai, cortical.io, maluuba, MindMeld, iDIBON, vsenso

### Horizontal AI
IBM Watson, Cortana, sentient technologies, VIV, vicarious, nervana SYSTEMS, nara, Numenta, HyperScience, SI SCALED INFERENCE, Descartes Labs, clarifai, DEXTRO, Geometric Intelligence, MetaMind

### Search
hp, Autonomy, ORACLE, ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

### Data Services
uo, Mu Sigma, OPERA, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

### For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

### Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumall, Airtable, retention SCIENCE, custora

## Applications

### Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, 6sense, QUANTIFIND, ACTIONIQ, fuse machines, ENGAGIO

### Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

### Human Capital
gild, Connectifier, textio, entelo, hiQ

### Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

### Ad Optimization
AppNexus, criteo, MediaMath, OpenX, rocketfuel, Integral Ad Science, theTradeDesk, TAdgorithms, dstillery, LiveIntent, DataXu, Appier, TAPAD

### Security
CYLANCE, CounterTack, cybereason, AREA 1 SECURITY, ThreatMetrix, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

### Vertical AI Applications
facebook, Clara, KASISTO, lumiata

### Publisher Tools
outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

### Govt / Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

### Finance
affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tidemark, Payoff, INSIKT, uora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient technologies

### Education/ Learning
KNEWTON, Clever, declara, PANORAMA EDUCATION, knowre

### Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

### Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

## Cross-Infrastructure/Analytics
amazon web services, Google, Microsoft, IBM, SAP, sas, 1010data, hp, Autonomy, VERTICA An HP Company, vmware, TIBCO, TERADATA, ORACLE, NetApp

## Open Source

### Framework
hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Apache Kylin, Flink, CDAP

### Query / Data Flow
SLAMDATA, HIVE, APACHE DRILL, Google Cloud Dataflow

### Data Access
cassandra, APACHE HBASE, accumulo, mongoDB, CouchDB, riak, OPENTSDB, sqoop, SciDB, kafka, nifi

### Coordination
talend, ZOOKEEPER, Apache Zookeeper, Apache Ambari

### Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

### Stat Tools
R, Scala, NumPy, SciPy

### Machine Learning
mllib, mahout, Aerosolve, Apache SINGA, MADlib, Caffe, learn, torch, CNTK, TensorFlow, FeatureFu, WEKA, VELES, DIMSUM, jupyter, DL4J

### Search
elasticsearch, Solr, lucene

### Security
Apache Ranger

### Visualization
Zeppelin

## Data Sources & APIs

### Health
JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

### IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

### Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMISE, quandl, xignite, CB INSIGHTS, mattermark, estimize, PLAID

### Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

### Location / People / Entities
acxiom, Experian, EPSILON, GARMIN, foursquare, InsideView, esri, STREETLINE, CARTODB, factual, PlaceIQ, Crimson Hexagon, placemeter, BASIS TECHNOLOGY, Sense 360

### Other
qualtrics, panjiva, DATA.GOV

## Incubators & Schools
GA, DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

*Last Updated 2/12/2016*

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

# DATA & AI LANDSCAPE 2019

## INFRASTRUCTURE

### HADOOP ON-PREMISE
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, jethro

### HADOOP IN THE CLOUD
aws, Microsoft Azure, Google Cloud, SAP Cloud Platform, IBM InfoSphere BigInsights, arm Treasure Data, Qubole, CAZENA

### STREAMING / IN-MEMORY
Amazon Kinesis, Google Cloud Dataflow, databricks, SAP Cloud Platform, ORACLE Cloud Service, confluent, striim, hazelcast, GridGain, GIGASPACES, Wallaroo Labs, FASTDATA, kx

### NoSQL DATABASES
Google Cloud, aws, ORACLE, Microsoft Azure, mongoDB, MarkLogic, Couchbase, DATASTAX, redislabs, KEYSPIKE, ArangoDB, SCYLLA

### NewSQL DATABASES
SAP, Clustrix, NuoDB, MariaDB, Cloud Spanner, MEMSQL, influxdata, Cockroach Labs, TIMESCALE, VOLTDB, splice machine, imply, paradigm4, Datastax, TiDB

### GRAPH DBs
neo4j, Amazon Neptune, IBM, ORACLE, OrientDB, InfoGrach, Objectivity

### MPP DBs
TERADATA, VERTICA, IBM Data Warehouse Systems, actian, Kognitio, Exasol, dremio, Yellowbrick

### CLOUD EDW
aws, Google Cloud, Microsoft Azure, Pivotal, snowflake, Infoworks

### SERVERLESS
Google Cloud Functions, AWS Lambda, IBM Cloud Functions, PULSAR, nuclio, Pivotal Function Service

### DATA TRANSFORMATION
talend, pentaho, alteryx, TRIFACTA, tamr, Paxata, StreamSets, UNIFI

### DATA INTEGRATION
SAP Data Services, Informatica, MuleSoft, TEALIUM, snaplogic, enigma, Qlik Data Catalyst, Segment, ATTUNITY, Talend, ZALONI, import.io, Infoworks, Fivetran, SNOWPLOW, MATILLION

### DATA GOVERNANCE
Informatica, IBM, SailPoint, McAfee Skyhigh Security Cloud, collibra, dremio, Alation, IMHUTA, OKERA, MANTA, data.world

### MGMT / MONITORING
aws, New Relic, actifio, rubrik, APPDYNAMICS, dynatrace, WAVEFRONT by VMWARE, DATADOG, SignalFx, druva, splunk>, Moogsoft, pagerduty, Anodot, unravel, Numerify, SCALYR, ScienceLogic, zenoss, OpsRamp, VEEAM, MAGNITUDE

### STORAGE
aws, Google Cloud, Microsoft Azure, IBM Storage, PURESTORAGE, ALLUXIO, wasabi, nimblestorage, Qumulo, panasas, COHESITY

### CLUSTER SVCS
Amazon EKS, IBM, Amazon EMR, Qubole, HESOSPHERE, packet, Cloudera, Bright Computing, CYCLECLOUD

### DATA GENERATION & LABELLING
amazon mechanical turk, Upwork, unity, appen, scale, HIVE, Labelbox, Mighty AI, ALEGION, LIONBRIDGE

### AI OPS
ALGORITHMIA, SPELL, comet, Vertaa.ai, datmo, datafold, Weights & Biases, Determined AI, fiddler

### GPU DBs & CLOUD
kinetica, OMNISCI, SQREAM, Blazingdb, brytlyt, PG-Strom, FLOYDHUB

### HARDWARE
Google TPU, arm, intel AI, NVIDIA, IBM Power Systems, GRAPHCORE, MYTHIC, Graphic Remedies, cerebras, Movidius, habana, WAVE COMPUTING, Esperanto, CORNAMI, flexlogix, DEFINIX

## CROSS-INFRASTRUCTURE/ANALYTICS
aws, Google Cloud, Microsoft, IBM, SAP, Hewlett Packard Enterprise, sas, 1010DATA, vmware, TIBCO, TERADATA, ORACLE, NetApp, syncsort, MAPR, cloudera

## ANALYTICS & MACHINE INTELLIGENCE

### DATA ANALYST PLATFORMS
Microsoft, pentaho, alteryx, Digital Reasoning, guavus, AYASDI, ATTIVIO, Datameer, incorta, interana, MODE, ENDOR, sisu, switchboard, Starburst

### DATA SCIENCE PLATFORMS
IBM, databricks, data iku, DOMINO, rapidminer, TIBCO, ANACONDA, sas, Altair, KNIME, MathWorks

### BI PLATFORMS
looker, sisense, tableau, aws, DOMO, ARCADIA DATA, ThoughtSpot, ATSCALE, SISENSE, Qlik, GoodData, Information Builders, birst, MicroStrategy, Keen IO

### VISUALIZATION
tableau, Power BI, SAP Lumira, Google Cloud, celonis, Periscope Data, zepl, GODATA, plotly, CHARTIO, TOUCAN TOCO

### MACHINE LEARNING
Azure Machine Learning, Google Cloud AutoML Vision, Amazon SageMaker, H2O.ai, DataRobot, gamalon, ViSENZE, ELEMENT AI, deepsense.ai

### COMPUTER VISION
Microsoft Azure, Amazon Rekognition, Cloud Vision API, clarifai, 商汤, EVER AI, deepomatic, neurala, twenty20n, UBIQUITY6, ADCE, 依图 YITU, trax, ALMERIA, synthesia, DataGrid

### HORIZONTAL AI
IBM Watson, Cortana, Facebook AI Research, sentient, Voyager Labs, vicarious, :) Affective, PROPHESEE, Cognitive Scale, Numenta, PETUUM, SENSEI, narologics, CURIOUS AI, OSARO, BLUE VISION, Fortress IQ

### SPEECH & NLP
Google Cloud, twilio, amazon alexa, Amazon Translate, narrative science, semantic machines, Mobvoi, EigenTechnologies, SoundHound Inc., VOCIVA, PRIMER, MindMeld, cogito, snips, SMARTLING, Unbabel, PolyAI

### SEARCH
elasticsearch, ORACLE ENDECA, algolia, coveo, Lucidworks, ATTIVIO, swiftype, EXALEAD, alphasense, MAANA, omni:us, SINEQUA

### LOG ANALYTICS
splunk>, sumologic, solarwinds, TIMBER, kibana, logz.io

### SOCIAL ANALYTICS
Hootsuite, sprinklr, NETBASE, synthesio, tracx, simplereach, bitly, SimilarWeb

### WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics, mixpanel, AMPLITUDE, Airtable, RESCI, SIGOPT, granify, custora

## APPLICATIONS – ENTERPRISE

### SALES
crunch, CHORUS, INSIDESALES.COM, peopleai, conversica, clari, aviso, tact.ai, TROOPS, fuse machines, Clearbit, ei

### MARKETING - B2B
RADIUS, App Annie, EVERSTRING, Lattice, MINTIGO, sense, tubular, ENGAGIO, KNOTCH, mrp

### MARKETING - B2C
zeta, bloomreach, SendGrid, braze, ACTIONIQ, BLUECORE, CONTENTSQUARE, TEALIUM, marticle, Amplero, amperity, QUANTIFIND, Simon, Lytics, [PERSADO], remesh

### CUSTOMER EXPERIENCE / SERVICE
qualtrics, MEDALLIA, SurveyMonkey, User Testing, CLARABRIDGE, zendesk, Kustomer, freshdesk, intercom, Drift, LIVEPERSON, Gainsight, pendo, HEAP, Amplitude, Watson Assistant, Dashflow, DigitalGenius, ASAPP, ada, AUTOMAT, affiniti, CaseDesk, netomi, frame.ai

### ENTERPRISE PRODUCTIVITY
slack, X, ORACLE, GURU, lumiata, DIFFBOT, clara, talla, Kasisto

### HUMAN CAPITAL
HireVue, pymetrics, hiQ, GIGSTER, mya, Allyo, textio, Wade&Wendy, Stella, Gusto, entelo, RESTFUL EXPERT, incommon, beamery, eightfold

### LEGAL
RAVEL, Seal, Everlaw, DISCO, kira, JUDICATA, BREVIA, IRONCLAD, PREMONITION, ROSS, casetext

### REGTECH & COMPLIANCE
BigID, TESSIAN, text IQ, Comply Advantage

### PARTNERSHIPS
CROSSBEAM, INFOSUM, DATA REPUBLIC

### FINANCE
snaplan, zuora, S/4 HANA, TRADESHIFT, mineraltree, SCALEFACTOR, bookkeeper, pilot

### BACK OFFICE AUTOMATION & RPA
UiPath, HyperScience, blue prism, AIPHERE, VIDADO, AppZen, WorkFusion, workato, Remoday, Catalytic, ANTWORKS, automation anywhere, ALKYMI, KRYON

### SECURITY
TANIUM, CYLANCE, zscaler, StackPath, illumio, CODE42, CipherCloud, DARKTRACE, ANOMALI, ThreatMetrix, VECTRA, CYBERARK, Guardian Analytics, DATAVISOR, siftscience, pindrop, exabeam, SIGNIFYD, SentinelOne, SecurityScorecard, SOCURE, Vade Secure, bitglass, BlueTalon, Recorded Future, feedzai, CyberX, BITSIGHT, ARIA, sparkcognition, broadbit Cybersecurity, FORTER, riskrecon, JASK, BLUEHEXAGON, Semmle, OBSIDIAN, AKONIS, SHIELD, Armorblox

## APPLICATIONS – INDUSTRY

### ADVERTISING
AppNexus, MediaMath, criteo, xAd, Integral, ORACLE MOAT, OpenX, datorama, theTradeDesk, Adjoiner, dstillery, LiveIntent, TAPAD, dataxu, gumgum, Appier, Sizmek, yieldmo, mediaalpha

### EDUCATION
Luminosity, 猿辅导, KNEWTON, Clever, declara, kidaptive, PANORAMA, knowre, gradescope

### REAL ESTATE
REDFIN, Opendoor, VTS, CREDIFI, GEOPHY, reonomy, COMPSTAK, SPACEMAKER, SKYLINE

### GOV'T
OPENGOV, mark43, PSI, FiscalNote, LiveStories, Passport, SmartProcure, STREETLIGHT DATA, OpenDataSoft

### INTELLIGENCE
Palantir, Dataminr, Quid, PRIMER, FORGE

### FINANCE - INVESTING
KENSHO, Quantopian, ADDEPAR, NUMERAI, iSENTIUM, ALGORIZ, FlexePack, PAGAYA

### FINANCE - LENDING
ondeck, affirm, 拍拍贷 PPDAI, JIANPU.AI, Kreditech, AVANT, TALA, zest finance, Upstart, aura, LEARBANC, upgrade, 100Credit, WeLab, Wecashlight, MoneyLion, ACTIVE.AI, aire, cognifi

### INSURANCE
Instrumentle, Lemonade, CYENCE, Hippo, Shift Technology, ROOT, zesty.ai, TRUEBILLS, CAPE

### HEALTHCARE
flatiron, Clover, KYRUS, HealthTap, METABIOTA, Ginger.io, Glow, babylon, 3DMed, zebra, PatnAI, OVIA, TEMPUS, patientslikeme, AiCure, insitro, LinkDoc, enlitic, citizen, notable, RECURSION, prognos, Qventus, ARTERYS, IMAGEN, BAYLABS, PAIGE, DATAVANT, innovaccer, LeanTaaS

### LIFE SCIENCES
23andMe, color, Karkinos Healthcare, BenevolentAI, verily, WuXiNextCODE, SCRIPT, Clear Labs, freenome, NANOPORE, Phosphorus, DNAnexus, CITRINE, twoXAR, Atomwise, deep genomics, OWKIN

### TRANSPORTATION
UBER, TESLA, WAYMO, ZOOX, CLEARPATH, cruise, nuro, AGILITY, nutonomy, drive.ai, CAMBRIDGE, Aurora, nauto, AIMOTIVE, G7, PILOT.AI, NIO, OPTIMUS RIDE, moovit, Ike, nexar, Kodiak, comma.ai, netradyne, Civil Maps, cognia, thinci, INRIX

### AGRICULTURE
FARMERS, Granular, JOHN DEERE, BLUE RIVER, FarmersEdge, AgroStar, FarmLogs, TARANIS, GAMAYA, Terrafuture, prospera

### COMMERCE
Instacart, FAIRE, STITCH FIX, Dia & Co, RetailNext, HowGood, heuritech, SCORTEX, KONUX, TACHYUS

### INDUSTRIAL
AVEVA, SIEMENS, Bright Machines, PREDIX, UPTAKE, PARAVIS, OSIsoft, SCORTEX, KONUX, TACHYUS

### OTHER
eharmony, stem, Amper, ByteDance, happen, celect, SOJERN, BOXEVER, VERDIGRIS, duetto, Jukebox, Second Spectrum, Electric, ZINIER, Spoke, Moveworks

## OPEN SOURCE

### FRAMEWORKS
Hadoop, Spark, Tensorflow, Keras, Theano, Torch, MAGMA, fastai, Caffe, Microsoft Cognitive Toolkit, Flink, YARN, TEZ, MESOS, kubernetes, docker, CDAP, Red Hat, HELIX

### QUERY / DATA FLOW
Spark SQL, Hive, presto, SLAMDATA, APACHE DRILL, GraphQL, Flink

### DATA ACCESS & DATABASES
cassandra, mongoDB, redis, CouchDB, Cockroach Labs, druid, OpenTSDB, SciDB, riak, Hansle, Cloud Spanner, ACCUMULO

### ORCHESTRATION & MGMT
talend, Zookeeper, OOZIE, Ambari, Airflow, MESOS, etcd, Kong

### STREAMING & MESSAGING
Spark, nifi, Flink, beam, kafka, STORM, Apache RocketMQ

### STAT TOOLS & LANGUAGES
python, R, Scala, Julia, Studio, SciPy, julia

### AI OPS & INFRA
mlflow, Kubeflow, mlengine, DVC, SELDON, Polyaxon

### AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, Keras, Torch, MAGMA, fastai, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, mxnet, VELES, Chainer, Michelangelo, ONNX, WEKA, LUDWIG, PyTorch, neon, DSSTNE, mlib, DL4J, MAHOUT, Aerosolve, Dist-ML, mlr, OpenML

### SEARCH
elasticsearch, Solr, Lucene

### LOGGING & MONITORING
elasticsearch, kibana, SENTRY, logstash, Prometheus, fluentbit, fluentd, Grafana, VECTOR

### VISUALIZATION
matplotlib, TensorBoard, seaborn, Bokeh

### COLLABORATION
BeakerX, jupyter, ANACONDA

### SECURITY
Apache Ranger, KNOX, Sentry, accumulo

## DATA SOURCES & APIs

### HEALTH
VALIDIC, practicefusion, fitbit, GARMIN, HUMAN API, kinsa, MIMIC

### IOT
GE Digital, UPTAKE, thingworx, helium, samsara, estimote

### FINANCIAL & ECONOMIC DATA
Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB INSIGHTS, PLAID, SECOND MEASURE, INVESTNET YODLEE, THE WORLD BANK, estimize, PREMISE, Quandl, Eagle Alpha, StockTwits, xignite, Thinknum, earnest research, predata

### AIR / SPACE / SEA
Orbital Insight, planet, SKYSAT, AIROBOTICS, spire, PRECISIONHAWK, Descartes Labs, kespry, UNDERSTORY, tellusLabs, WINDWARD, DroneDeploy, MarineTraffic, LOFT ORBITAL, RS Metrics

### PEOPLE / ENTITIES
acxiom, experian, EPSILON, InsideView, Crimson Hexagon, BASIS TECHNOLOGY, Quantcast, SAFEGRAPH

### LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, mapbox, sense360, pitney bowes, HEXAGON, Placer.io, esri, factual, CARTO, Mapillary, Streetline, cubiq, Radar, OpenStreetMap

### OTHER
DATA.GOV, IMAGENET, wiki-links, LabelBox, Microsoft Research Open Data, APOLLOSCAPE, CRUX, MEMORIALDATA, grafiti.io

## DATA RESOURCES

### DATA SERVICES
OPERA, UCI Machine Learning Repository, DATA SCIENCE, fractal, kaggle, DataKind, EXL, INNOPLEXUS

### INCUBATORS & SCHOOLS
PLURALSIGHT, GA, galvanize, DataCamp, DataElite, INSIGHT, The Data Incubator, METIS

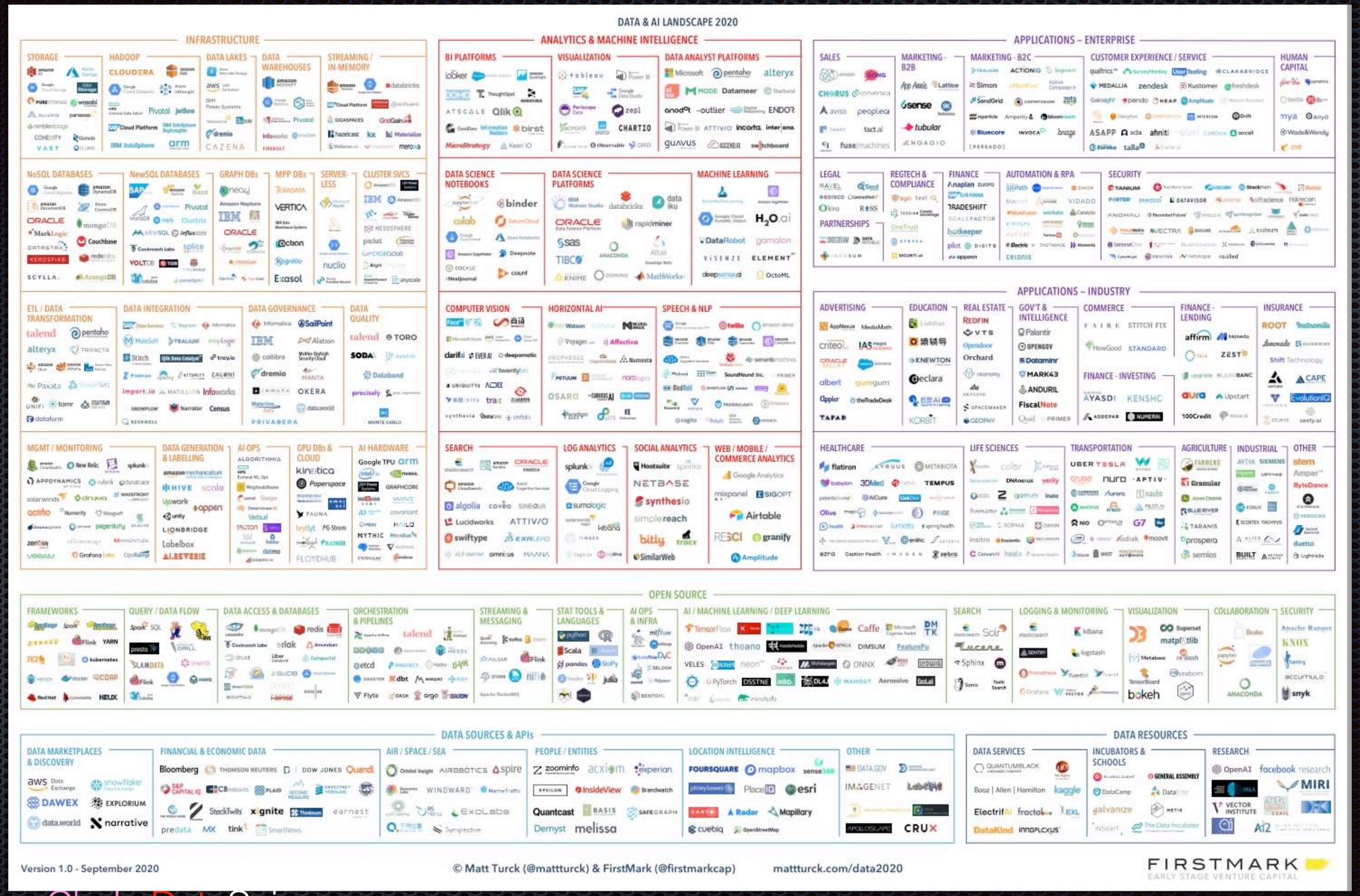
### RESEARCH
OpenAI, facebook research, MILA, MIRI, VECTOR INSTITUTE, CSAIL, DeepMind, Allen Institute for Artificial Intelligence

July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)

mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

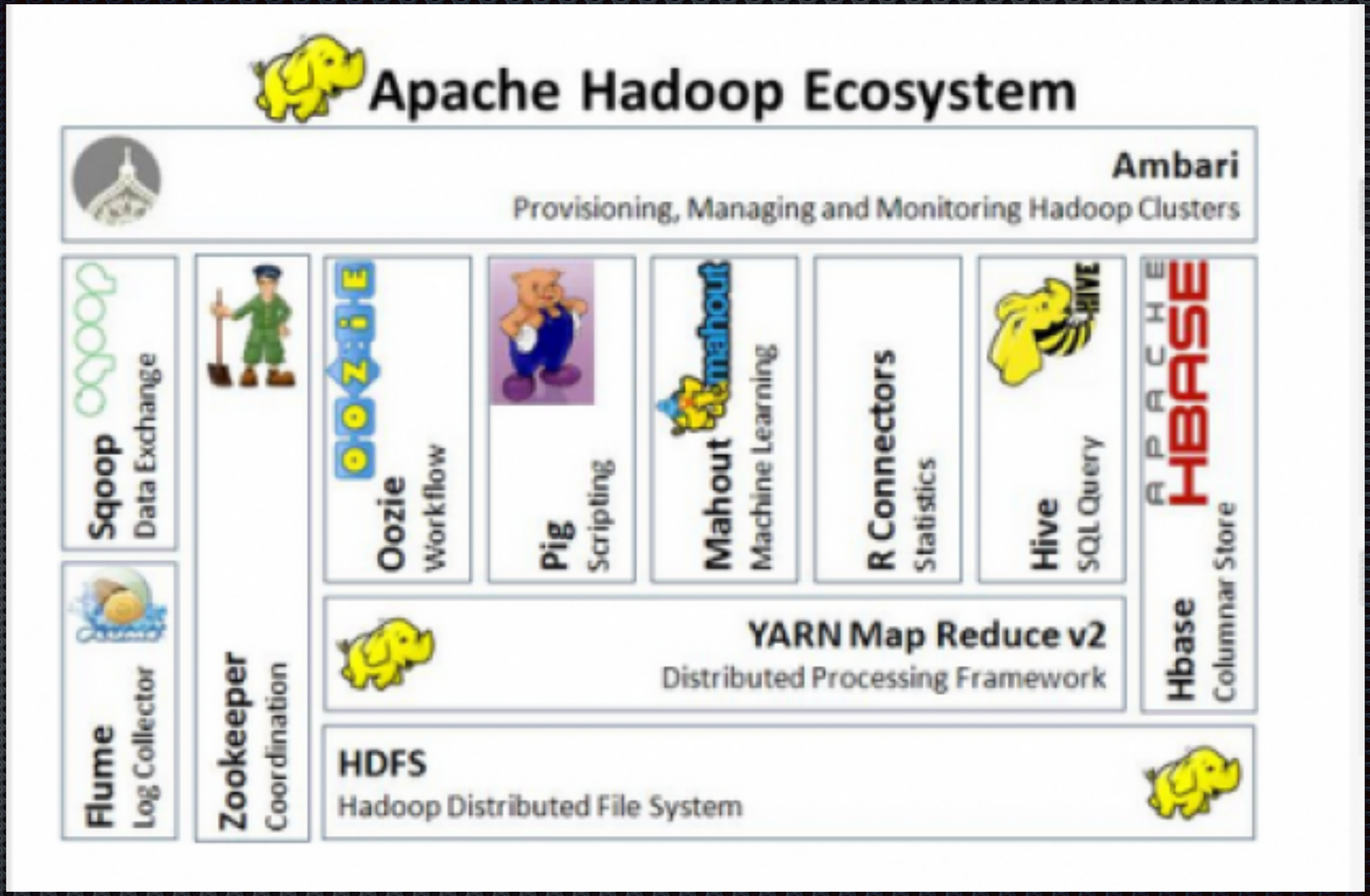



Chula DataScience

# Put them all together (Batch)

# Steps (Batch)

- Transfer data to (H)DFS
- Perform data analytics
  - BI Tools: Tableau, Pentaho, etc.
  - Frameworks: Spark, MapReduce, Hive (Shark), etc.
  - Languages: R, Python, Java, Scala, HiveSQL, etc.
- Use Visualization tool to view results


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
APACHE
HBASE
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

# Ecosystem

- Getting Data into HDFS
- Querying data in HDFS
- SQL on Hadoop / HBase
- Stream Processing
- NoSQL stores
- Data Analytics Tools
- Business Intelligence (BI) Tools

# Data Formats

- Text ?
- Serialization
  - binary protocol?
  - JSON, XML?
  - Java Serialization?
- None is good for storing meta data in HDFS

# MapReduce revisit (Embalancing Parlell)

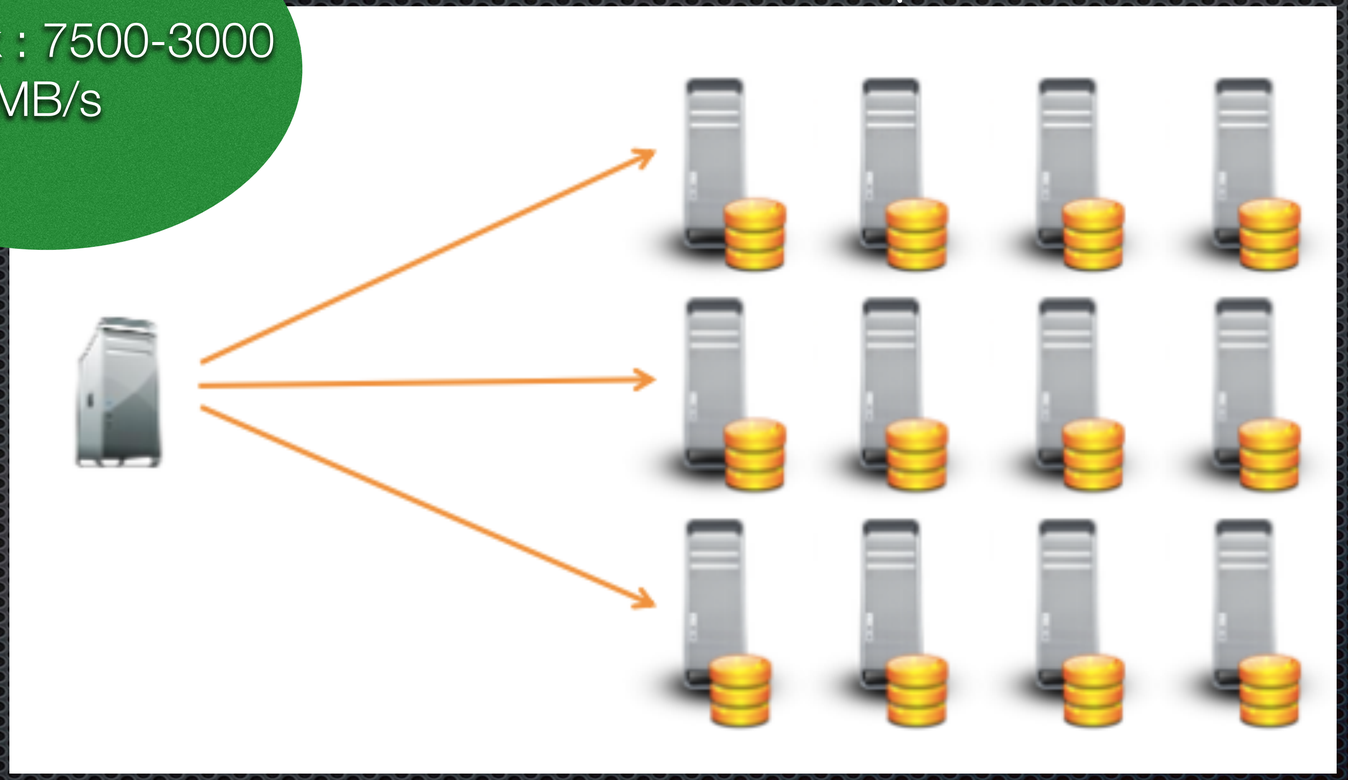
ทำให้ big data จู่จี้มากขึ้น

# MapReduce?

อดีตคือ cluster แย่งกันอ่านไฟล์

ปัจจุบันคือ การกระจาย storage

100 Disk : 7500-3000 MB/s



Let’s distribute the data/storage.

# Example (Word frequency count)

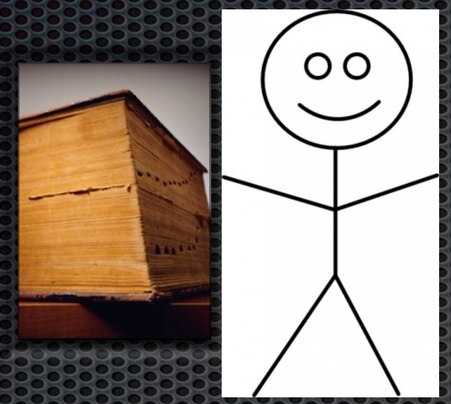
Find frequency of each word in this big book?

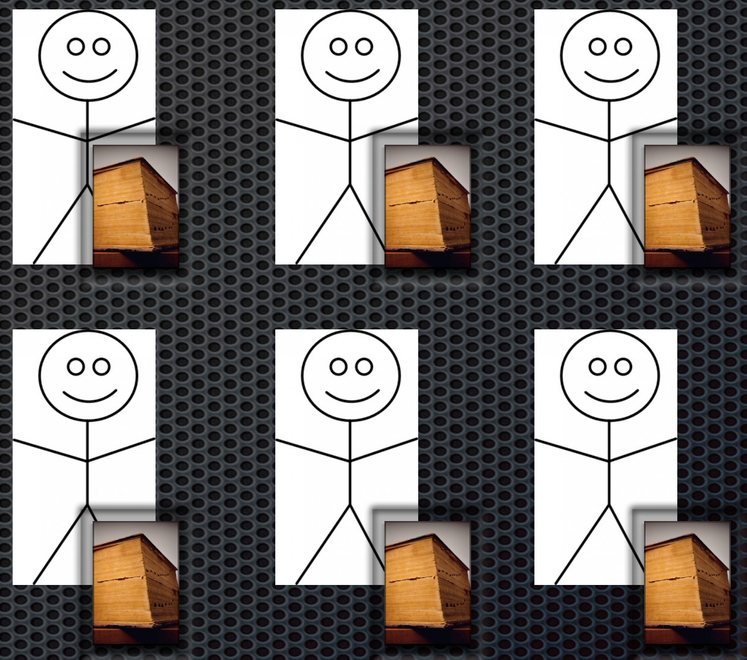
Sol 1

ทำต้องมีน file server xelox หนังสือทุกคน

# Word freq.: Grid/Cluster

Here is the data.
Please count.





รอบ 2

ดึกมปงหน้า

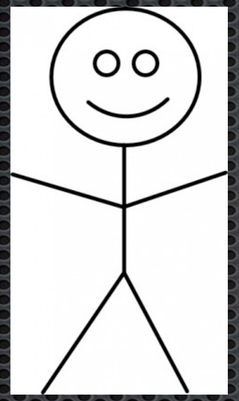
1000 คน เร็ว 1000 เท่า
10000 คน เร็ว 10000 เท่า

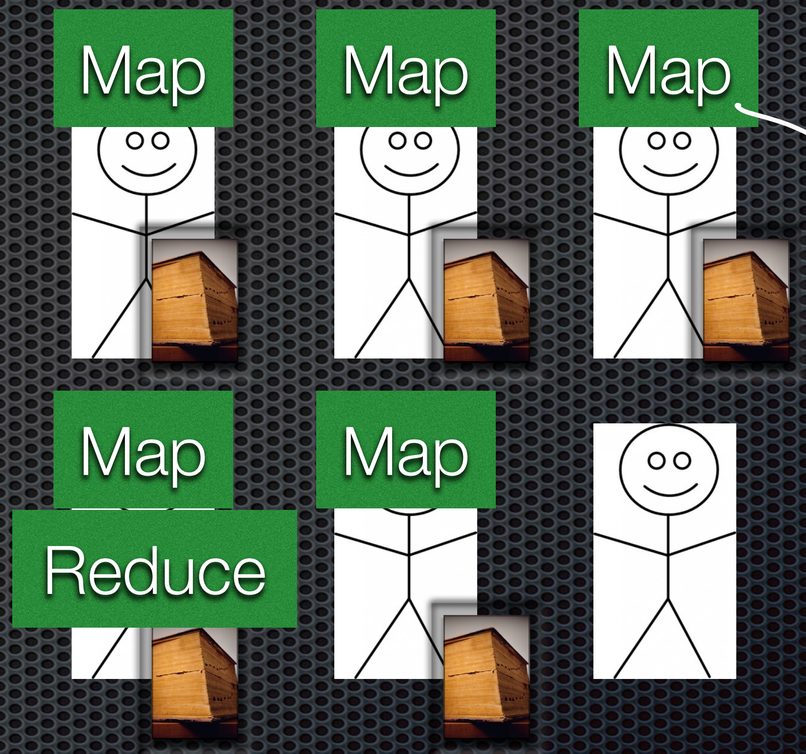
# Word freq.: MapReduce

1000 หน้าแรก 1001 - 2000 pages

Store a part of data.

With your data, please count.



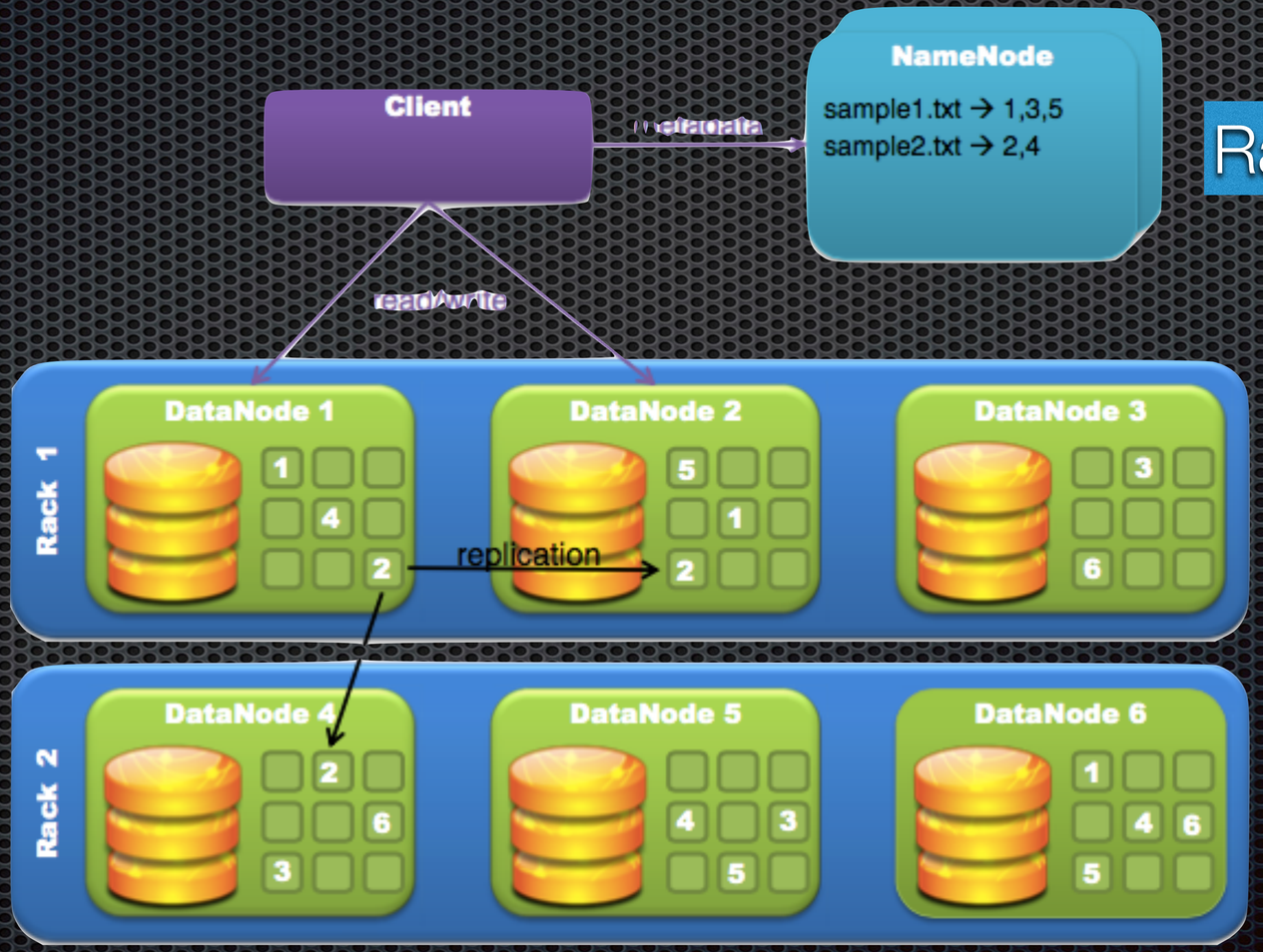



โปรแกรมนับคำ

2001 - 4000 pages.

Google เก็บเศษสิ่งของเว็บใน cluster ทำให้ Google เกิดขึ้นทุกวันนี้

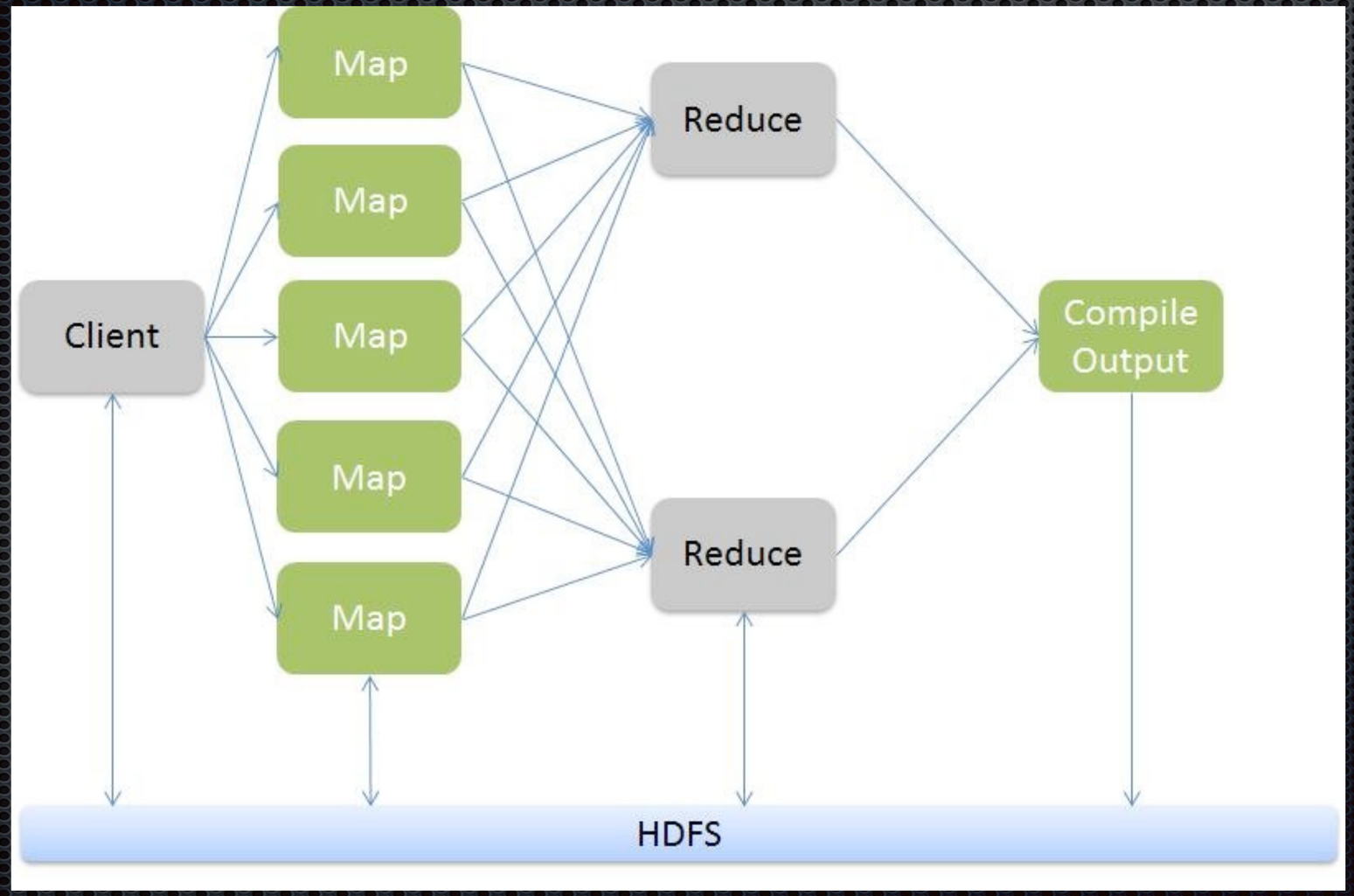
# Hadoop HDFS




Rackaware

3 copy

# How MapReduce work?



# Hadoop Architecture

# What if it is JSON/XML data?

```
1  {
2      "id" : 100,
3      "message" : "Sales target reached.",
4      "recipients" : [
5          {
8              "title" : "VP Sales and Marketing"
9          },
10         ...,
15         {
16             "name" : "Margo Haim",
17             "email" : "margoh@bigcompany.com",
18             "title" : null
19         }
20     ]
```

Expression expected.

# Data Formats (ctd.)

- Writeable
  (original hadoop)
- SequenceFiles
- AVRO
  (designed by/for hadoop)
- Apache Thrift
  (designed by Facebook)

- Used by Hive
- Protocol Buffers
  (designed by Google)
- (similar to Thrift)
- Apache Parquet
- (Columnar data)

# Alternative solutions

- Apache HBase
- Apache Cassandra
- Apache Accumulo
- Apache Kudu
- Apache Hive (Shark)
- Apache Drill
- Cloudera Impala

# Transfer data to HDFS

- SQOOP for RDBMS (and JDBC/ODBC)
- Flume for log (multiple sources)
- Chukwa (data collection system)
- Kafka for messaging (high-throughput distributed messaging system)

# Querying data in HDFS

- Java MapReduce
- Hadoop Streaming
- Spark
- Hive, Shark
- Cloudera Search (text search)
- Impala
  (Real-time query over big data)
- Presto
  (fast SQL query over Hadoop)

# SQL on Hadoop / HBase

- Hive/Shark
- Apache Drill
  (Interactive Analysis)
- Impala
  (Real-time query over big data)
- Presto
  (fast SQL query over Hadoop)

# Stream Processing

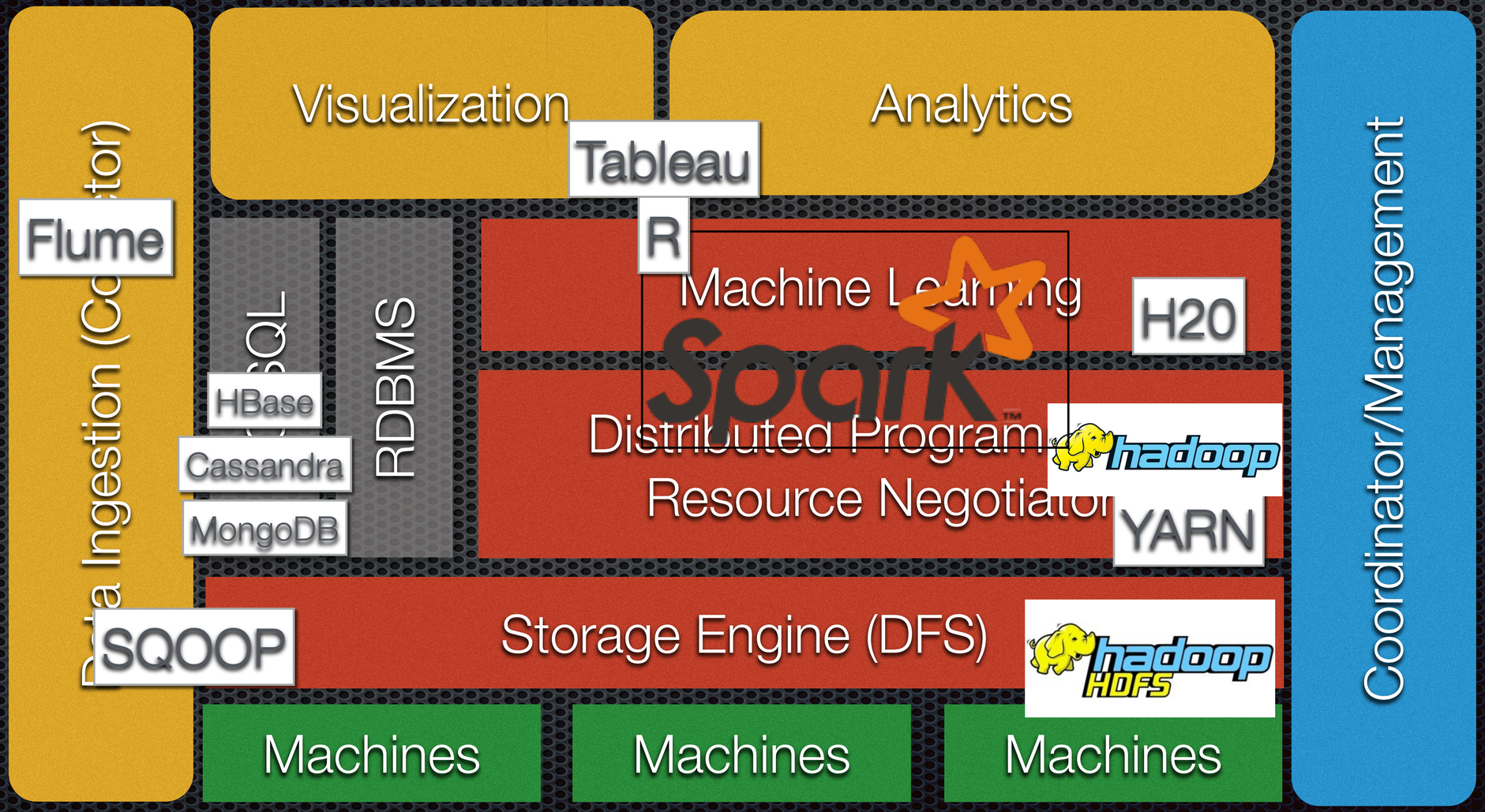
- Storm
- Spark

# NoSQL

- Base
- Cassandra
- etc.

# Put the jigsaw puzzle together

# Big Data Analytics

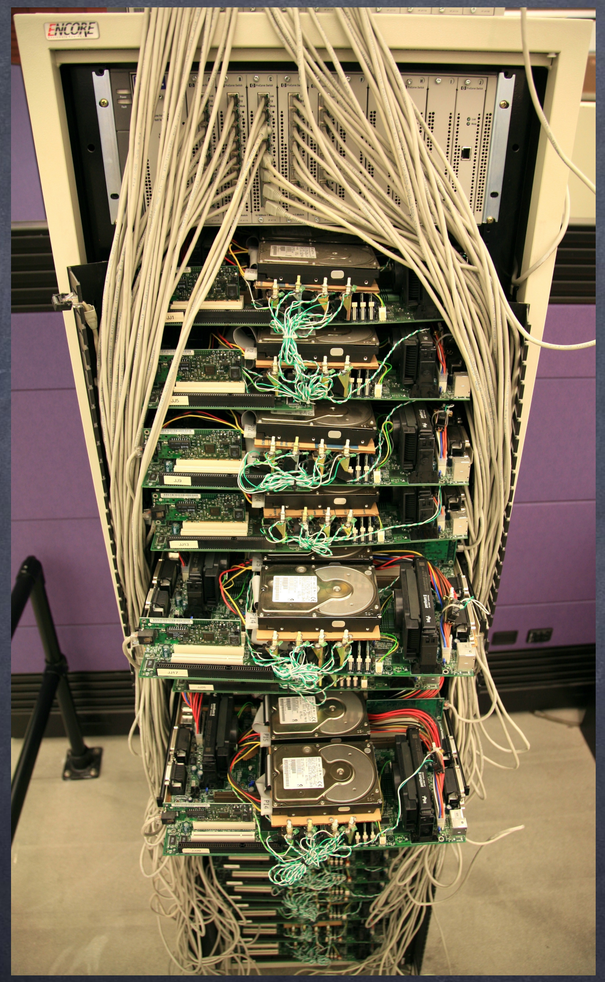


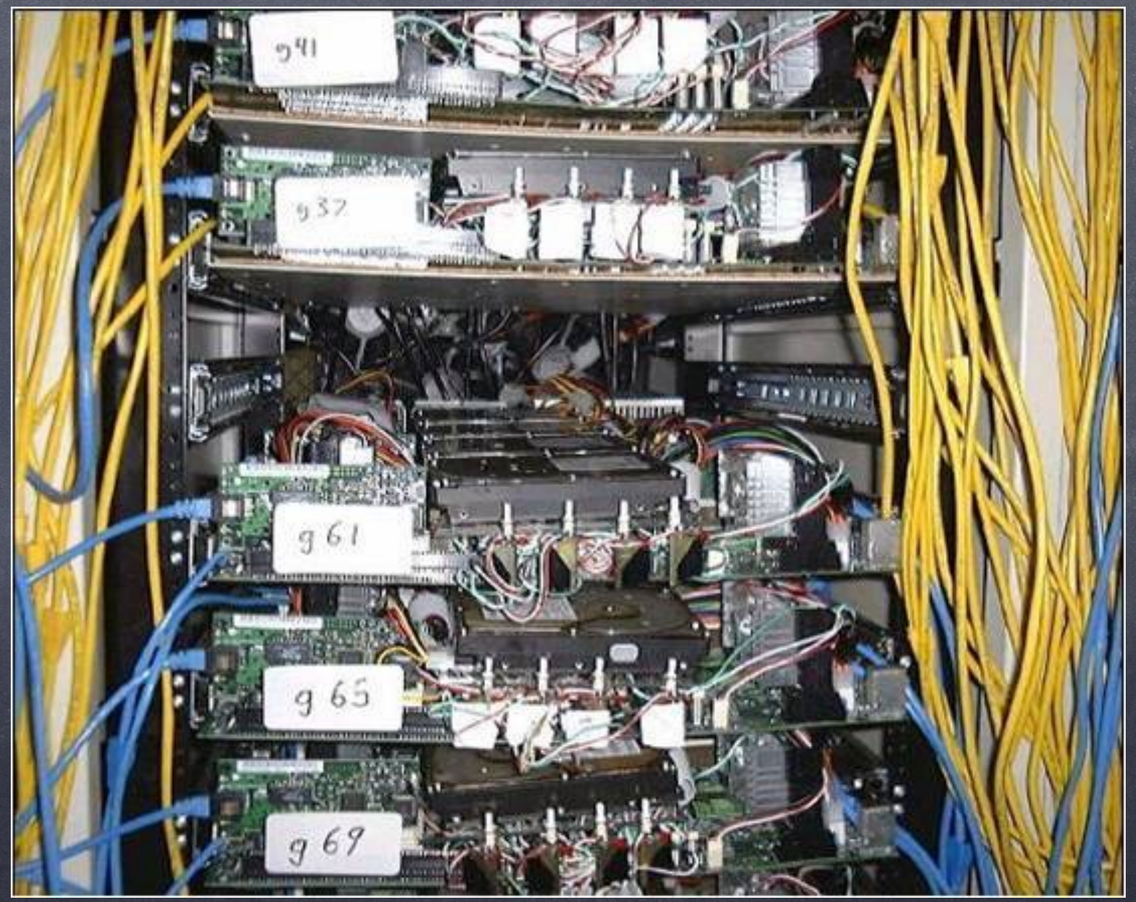
# References

- http://hadoopecosystemtable.github.io
- http://hadoopilluminated.com
- https://opensource.com/life/14/8/intro-apache-hadoop-big-data
- http://www.ibm.com/developerworks/views/analytics/libraryview.jsp?search_by=Data+warehouse+augmentation

# Pictures from Google Cluster

Taken from various web sites..


ENCORE


941
937
961
965
969

Rackable SYSTEMS
2 8:26

Rackable SYSTEMS
2 8:26