

# Big Data Technologies

## General Introduction

Lionel Fillatre

Polytech Nice Sophia

[lionel.fillatre@univ-cotedazur.fr](mailto:lionel.fillatre@univ-cotedazur.fr)

# Class website

- Class website

<https://lms.univ-cotedazur.fr/2024/course/view.php?id=1849>

- Password for the course registration: dhP57M&up
- **You must register yourself!**

# Outlines

- Course Logistics
- Basic Definitions
- Big Data Landscape
- Big Data Framework
- Big Data with a Cluster
- Introduction to Docker
- Install Docker
- Conclusion

# Course Logistics



# Timeline of this course

1. **16/09/2024**: General Introduction + Introduction to Docker and Scala
2. **23/09/2024**: Hadoop Distributed File System
3. **30/09/2024**: MapReduce
4. **07/10/2024**: Spark + **1<sup>st</sup> scored lab**
5. **14/10/2024**: NoSQL Databases
6. **21/10/2024**: SQL in Big Data
7. **04/11/2024**: Advanced Analytics + **2<sup>nd</sup> scored lab**
8. **18/11/2024**: **1 hour written exam (8h-9h)**

# Grading

- Week 4: 1<sup>st</sup> scored lab (25% of the final grade)
- Week 7: 2<sup>nd</sup> scored lab (25% of the final grade)
- Week 8: final written exam (50% of the final grade)
- **Scored labs**
  - Work in pairs (or solo if the number of students is not even)
  - You need a personal laptop with Docker and the dedicated container installed
  - Results are important (of course!) but writing clear code comments is the key to get a good score!
  - Lab 3 is useful for scored lab 4
  - Lab 6 is useful for scored lab 7
  - Participation in labs might be considered for scoring
- **Final exam (no document)**
  - The final exam concerns all the course topics
  - Some technical questions on the course and some general questions (architecture, NoSQL, etc.)
  - Writing code is possible but very limited

# Classrooms for labs

- E106 (in French)
  - Teacher: Antonin Bavoil
  - Students: MAM5
- B218 (in English)
  - Teacher: Antonio SOGGIA
  - Students: M2 IM + EIT Digital M1 + EIT Digital M2
- Classrooms may change: <https://sco.polytech.unice.fr/>

# Basic Definitions

# Big Data Definition

- No single standard definition!
- “**Big Data**” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

# The 10 Vs of Big Data



# Who's Generating Big Data



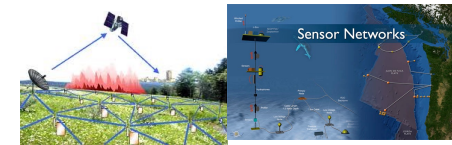
**Social media and networks**  
(all of us are generating data)



**Scientific instruments**  
(collecting all sorts of data)



**Mobile devices**  
(tracking all objects all the time)



**Sensor technology and networks**  
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

# The Model Has Changed...

- **The Model of Generating/Consuming Data has Changed**

**Old Model:** Few companies are generating data, all others are consuming data



**New Model:** all of us are generating data, and all of us are consuming data





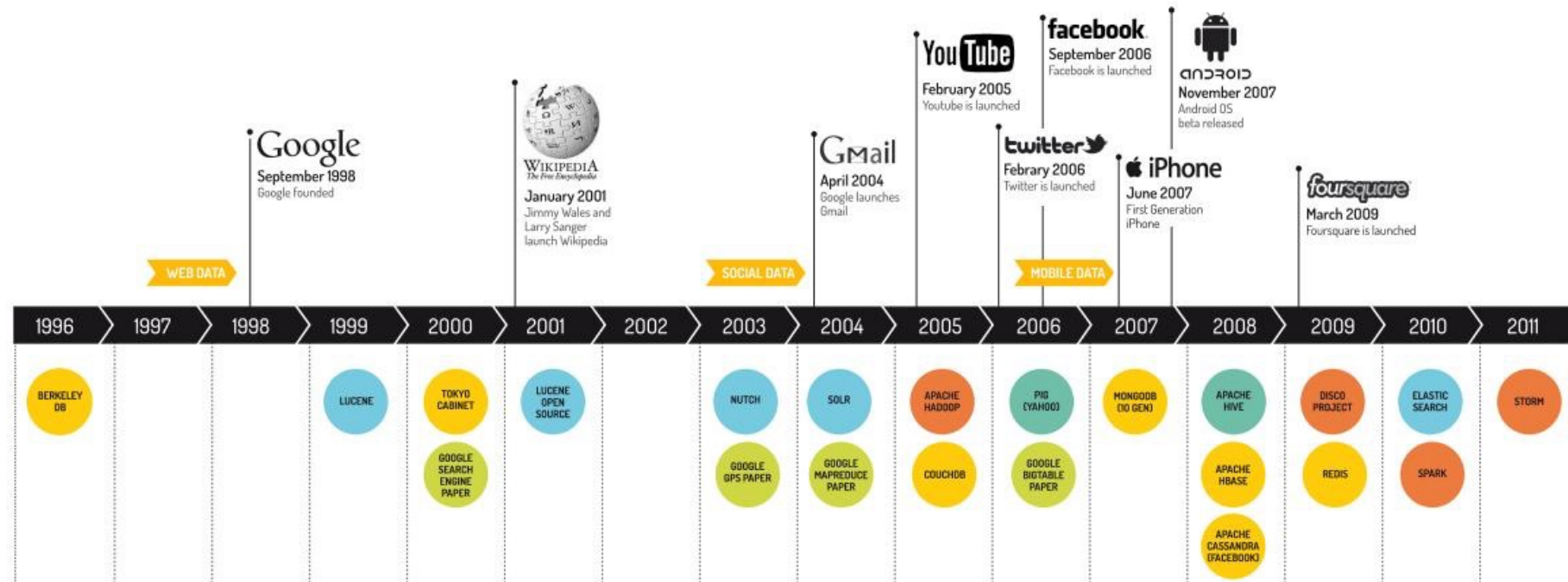
# Challenges in Handling Big Data

- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data
- **Recently, we observe a convergence to Artificial Intelligence**
  - Big data needs artificial intelligence to analyze the complex volume of data
  - This gives birth to advanced analytics

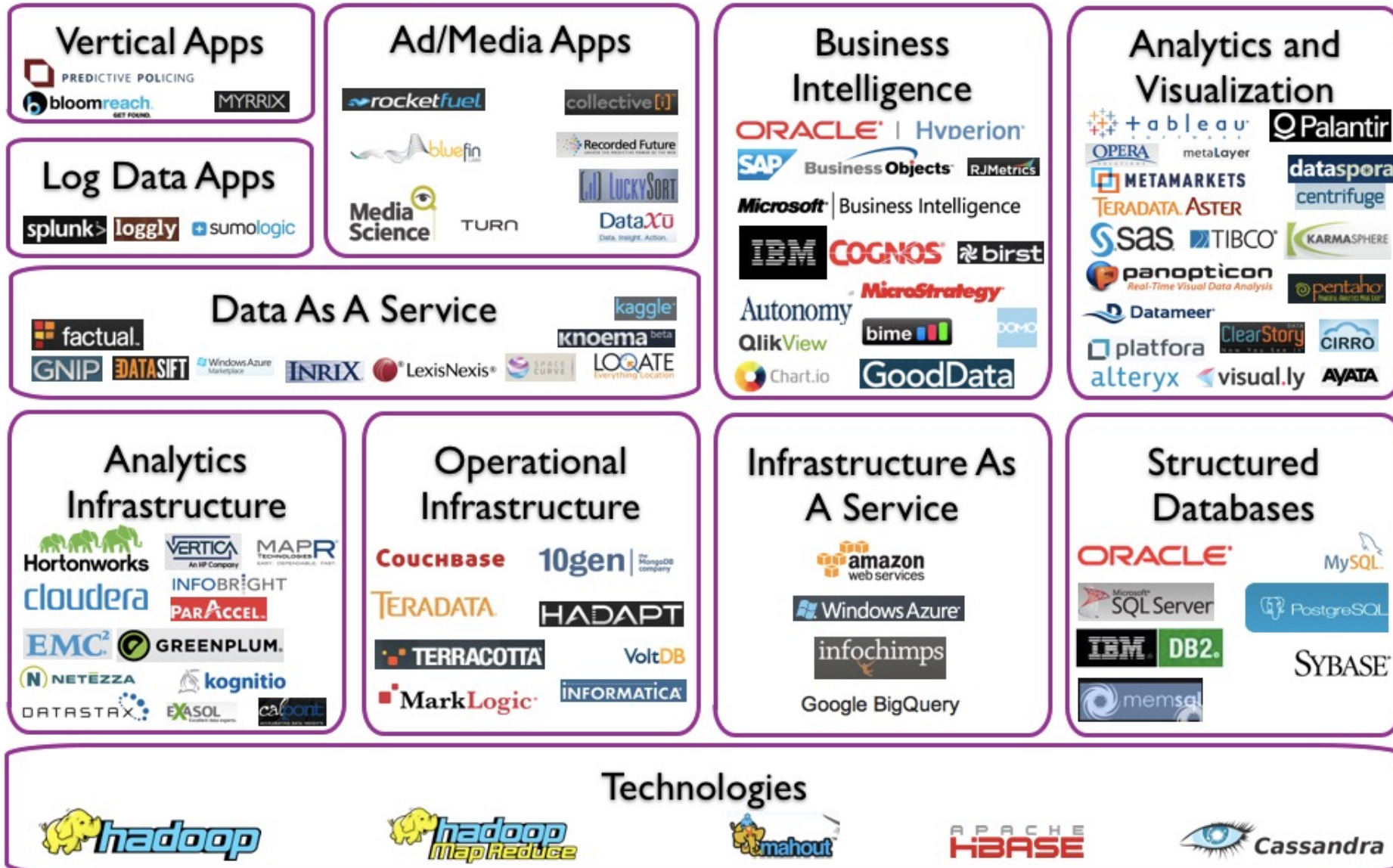
# Big Data Landscape

# What technology for Big Data?

## BIG DATA A BRIEF HISTORY



# Big Data Landscape





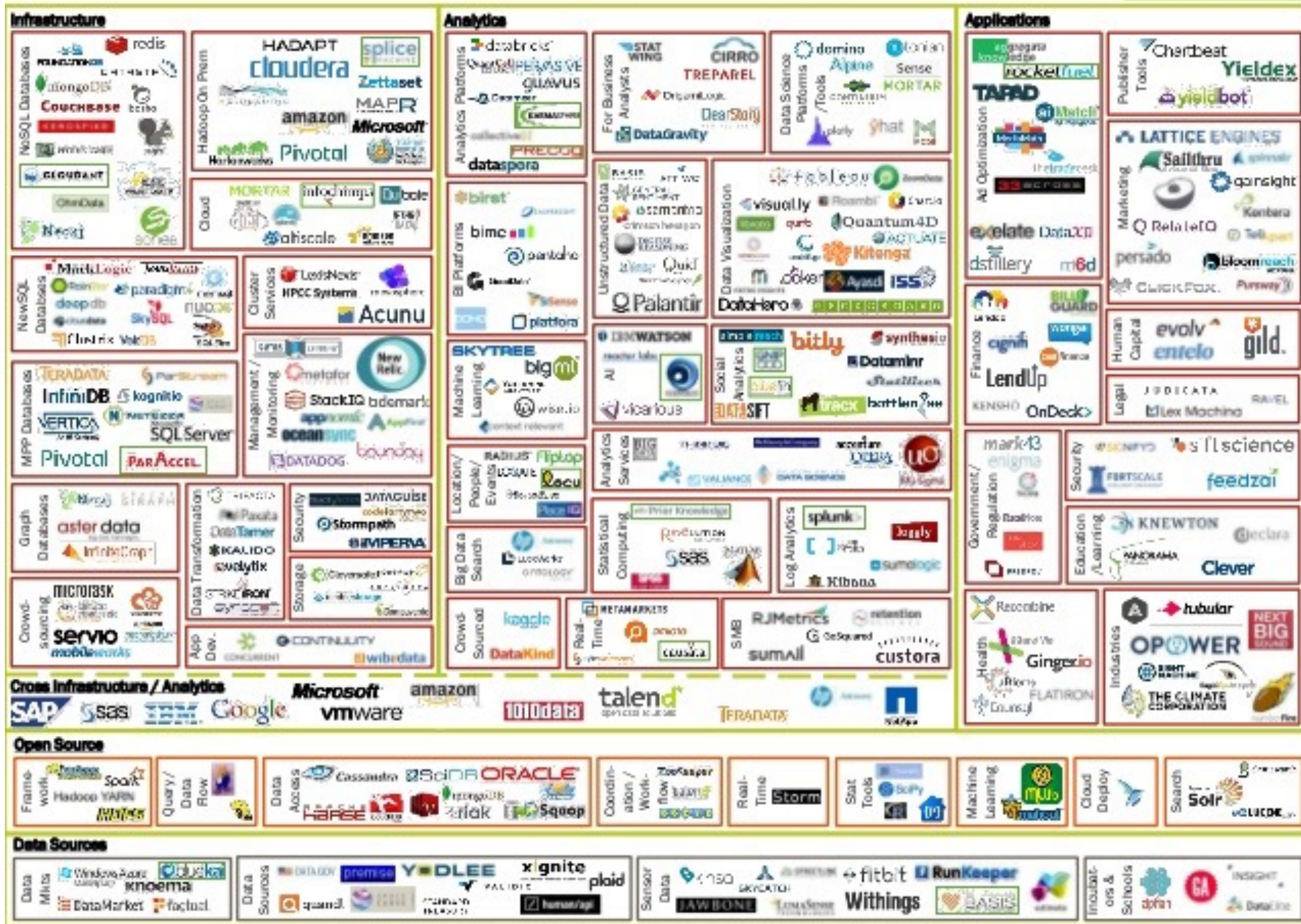
# Big Data Landscape (Version 2.0)





# BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO





## BIG DATA LANDSCAPE 2017

## INFRASTRUCTURE

|                                                                                                                                                                                                             |                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                        |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <h3>HADOOP ON-PREMISE</h3> <p>cloudera Hortonworks</p> <p>MAPR Pivotal</p> <p>IBM InfoSphere</p> <p>bluedata jethro</p>                                                                                     | <h3>HADOOP IN THE CLOUD</h3> <p>amazon web services Microsoft Azure</p> <p>Google Cloud Platform</p> <p>IBM InfoSphere BigInsights</p> <p>treasure data</p> <p>Qubole atlascale</p> <p>CAZENA CenturyLink</p>           | <h3>STREAMING / IN-MEMORY</h3> <p>amazon web services databricks</p> <p>confluent stream</p> <p>GridGain</p> <p>DATA TORRENT</p> <p>dataArtisans</p> <p>ORACLE</p> <p>hazelcast</p> <p>TEPRACOTTA</p>                                                                                                                                                                                                  |
| <h3>NOSQL DATABASES</h3> <p>Google Cloud Platform</p> <p>ORACLE</p> <p>amazon dynamodb</p> <p>Microsoft Azure MarkLogic</p> <p>mongodb DATASTAX</p> <p>HEROSPHERE Couchbase</p> <p>redislabs influxdata</p> | <h3>NEWSQL DATABASES</h3> <p>SAP</p> <p>Clustrix Pivotal</p> <p>nuodb</p> <p>Cockroach LABS</p> <p>memsql</p> <p>splice</p> <p>scrimba</p> <p>YOLDB</p> <p>ciudata</p> <p>Trifolion</p> <p>dataops</p> <p>paradigm4</p> | <h3>GRAPH DBS</h3> <p>neob4j</p> <p>IBM</p> <p>ORACLE</p> <p>orientdb</p> <p>orientdb</p> <p>InfraGraphX</p> <p>Objectivity</p> <h3>MPP DBS</h3> <p>TERADATA</p> <p>VERTICA</p> <p>NETEZZA</p> <p>Cloudera</p> <p>kognitio</p> <p>SOL</p> <p>dremio</p> <h3>CLOUD EDW</h3> <p>amazon Redshift</p> <p>Google Cloud Platform</p> <p>Microsoft Azure</p> <p>Pivotal</p> <p>snowflake</p> <p>Infoworks</p> |
| <h3>DATA TRANSFORMATION</h3> <p>talend</p> <p>pentaho</p> <p>alteryx</p> <p>TRIFACTA</p> <p>tmnr</p> <p>Paxata</p> <p>StreamSets</p> <p>UNIFI</p>                                                           | <h3>DATA INTEGRATION</h3> <p>informatica</p> <p>spagoLOG</p> <p>Segment</p> <p>enigma</p> <p>alooma</p> <p>ZALONI</p> <p>splintery</p> <p>Import</p> <p>Stitch</p> <p>MuleSoft</p> <p>TEALUM</p>                        | <h3>DATA GOVERNANCE</h3> <p>informatica</p> <p>ibm</p> <p>skyhigh</p> <p>collibra</p> <p>Alation</p> <p>Waterline Data</p> <h3>MGMT / MONITORING</h3> <p>amazon CloudWatch</p> <p>New Relic</p> <p>APDYNAMICS</p> <p>actifio</p> <p>WAVEFRONT</p> <p>unavrel</p> <p>splunk</p> <p>ooC</p> <p>Yrrocano</p> <p>Numeryfi</p> <p>pagerduty</p>                                                             |
| <h3>STORAGE</h3> <p>amazon S3</p> <p>Google Cloud Platform</p> <p>Microsoft Azure</p> <p>ALLUXIO</p> <p>nimbustorage</p> <p>Qumulo</p> <p>COHO</p> <p>panasas</p>                                           | <h3>CLUSTER SERVICES</h3> <p>amazon EMR</p> <p>EMR</p> <p>ibm</p> <p>docter</p> <p>MESOSPHERE</p> <p>CoreOS</p> <p>perisdata</p>                                                                                        | <h3>APP DEV</h3> <p>Lightbend</p> <p>APPTO</p> <p>rainforest</p> <p>CASK</p> <h3>CROWDSOURCING</h3> <p>amazon mechanical turk</p> <p>workwork</p> <p>WorkFusion</p> <p>CloudPionier</p> <h3>HARDWARE</h3> <p>Google GPU</p> <p>nervana</p> <p>ARM</p> <p>Graphcore</p> <p>MYTHIC</p> <p>Graphcore</p> <p>nvidia</p> <p>Movidius</p> <p>SCORTEX</p>                                                     |

## ANALYTICS

**DATA ANALYST PLATFORMS**

Microsoft pentaho alteryx  
Digital  
guavus AYASDI  
WATTIVO Datameer Quid  
ClearStory OrigamiLogic interana  
Bottlenose ARIMO ENDOR MODE

**DATA SCIENCE PLATFORMS**

IBM KNIME dataiku  
DOMINO yhat  
CONTINUUM ANALYTICS Alpine  
ALGORITHMIA Angoss

**BI PLATFORMS**

Microsoft amazon redshift  
looker Wave Analytics  
ARCADA DATA atscale  
GoodData SIGMA

**VISUALIZATION**

tableau SAP  
Google Cloud Platform  
Qlik CELONIS  
Periscope ZEPL  
CHARTIO plotly  
VIZOP

**VERTICAL ANALYTICS**

PREDIX Cjio CAPE  
UPTAKE  
TACHYUS Alliumium  
deborama

**STATISTICAL COMPUTING**

sas SPSS  
MATLAB

**DATA SERVICES**

Palantir MapInfo  
SCIENCE OPENA  
DATA SCIENCE kaggle  
Fractal EXL  
DataKind FF

**MACHINE LEARNING**

IBM Watson Cortana.  
Face++ 旷视 sentiment  
Voyager  
Affective CognitiveScale SenseMe  
aroonomic LogicalScale Numenta  
petuum  
OSARO  
CURIOUS AI  
Holo VISION

**SPEECH & NLP**

Google Cloud Platform  
amazon Alexa NarrativeScience  
Wabai Labs semantic machines  
Mindfield ARRIA  
Talkio IDIBON  
maubai snippets  
Grindspace Soundhound Inc.

**SEARCH**

elastic  
ORACLE  
Allegia  
swiftype MAANA  
alphenase  
Searchlink SINEQUA

**LOG ANALYTICS**

splunk  
sumologic  
loggly kibana  
logzio

**SOCIAL ANALYTICS**

Hootsuite  
NETBASE  
DATASIFT track  
synthesis  
bitly predata

**WEB / MOBILE / COMMERCE ANALYTICS**

Google Analytics  
mixpanel AMPITUDE  
sumall Airtable  
retention SIGOPT  
granify custora

## APPLICATIONS – ENTERPRISE

**SALES**

- Chorus
- INSIDESALES.COM
- conversica
- clari
- AVISO
- TACT
- fusemachines
- TROOPS

**MARKETING - B2B**

- RADIUS
- App Annie
- EVERSTITCH
- Lattice
- infer
- HINTIGO
- sense
- tubular
- Datafox
- ENGAGIO

**MARKETING - B2C**

- Zeta
- bloomreach
- blue yonder
- PERSADO
- actionIQ
- kahuna
- BLUECORE
- SAILTHRU
- QUANTIFIND
- mparticle
- Amperio

**CUSTOMER SERVICE**

- MEDALLIA
- zendesk
- Gainsight
- NGDATA
- CLICKFOX
- DigitalGenius
- appuri
- automatt
- Frame AI
- msqi
- INTERCOM

**HUMAN CAPITAL**

- HireVue
- enelo
- hiQ
- GIGSTER
- textio
- RESTLESS HANDOUT
- Wade & Wendy
- Qustreet
- Stella
- pymetrics

**LEGAL**

- RAVEL
- Seal
- Everlaw
- JUDICATA
- Brevia
- R&SS
- casetext

**FINANCE**

- anaplan
- zuora
- tidemark
- zebra
- HANA
- TRADESHIFT

**ENTERPRISE PRODUCTIVITY**

- slack
- facebook
- ORACLE
- lumata
- Clara
- talia
- butter AI
- KASISTO

**BACK OFFICE AUTOMATION**

- iHub
- HyperScience
- captricity
- AppZen

**SECURITY**

- TANIUM
- StackPath
- DARKTRACE
- Illumio
- CODE42
- DATAgravity
- VECTRA
- CipherCloud
- Guardian
- ANOMALI
- sift science
- SYNPIV
- MimeZero
- SecurityScorecard
- Talon
- Recorded Future
- feedzai
- SOCRATE
- AESATA
- FORTISCALE
- YesHub
- HyperCoreLogic

## – APPLICATIONS – INDUSTRY

|                                                                                                                                                                                                                 |                                                                                                                                                                                         |                                                                                                                                                              |                                                                                                                                                              |                                                                                                                                                                       |                                                                                                                     |                                                                                                                  |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| <h3>ADVERTISING</h3> <p>  AppNexus<br/>  Criteo<br/>  RocketHub<br/>  OpenX<br/>  The Trade Desk<br/>  Tango<br/>  Appier<br/>  DYNAMILIS<br/>  Yieldmo         </p>                                            | <h3>EDUCATION</h3> <p>  Knewton<br/>  Clever<br/>  edX<br/>  Khan Academy<br/>  edmentum<br/>  FutureLearn<br/>  FutureLearn<br/>  FutureLearn         </p>                             | <h3>GOVERNMENT</h3> <p>  Socrata<br/>  OPENGOV<br/>  mark43<br/>  FNI<br/>  OpenDataSoft         </p>                                                        | <h3>FINANCE - LENDING</h3> <p>  OnDeck<br/>  Kredtech<br/>  Affirm<br/>  AVANT<br/>  INSIKT<br/>  TALA<br/>  TrueAccord<br/>  ignify<br/>  aire         </p> | <h3>FINANCE - INVESTING</h3> <p>  Dataminr<br/>  KENSCH<br/>  Quantopian<br/>  NUMERA<br/>  iSTOCKphoto<br/>  claritymoney<br/>  ALGORIX<br/>  RavenPack         </p> | <h3>REAL ESTATE</h3> <p>  Opendoor<br/>  VTS<br/>  CREDIFI<br/>  Reonomy<br/>  COMPASS         </p>                 | <h3>INSURANCE</h3> <p>  Metromile<br/>  Lemonade<br/>  CYNCE<br/>  Shift Technology<br/>  Tractable         </p> |
| <h3>HEALTHCARE</h3> <p>  FLATIRON<br/>  HealthTap<br/>  Ginger<br/>  COTA<br/>  Aiora<br/>  iMagine<br/>  CYTOSURE<br/>  METABIO<br/>  Glow<br/>  Aiora<br/>  iMagine<br/>  CYTOSURE<br/>  METABIO         </p> | <h3>LIFE SCIENCES</h3> <p>  Color Genomics<br/>  ZEPHYR<br/>  Citrine<br/>  twoAR<br/>  Atomwise<br/>  Benevolent AI<br/>  ZEPHYR<br/>  Citrine<br/>  twoAR<br/>  Atomwise         </p> | <h3>TRANSPORTATION</h3> <p>  UBER<br/>  TESLA<br/>  CLEARPATH<br/>  drive.ai<br/>  pilot.ai<br/>  OPTIMUS<br/>  nexar<br/>  comma.ai<br/>  Nuro         </p> | <h3>AGRICULTURE</h3> <p>  FARMERS EDGE<br/>  FarmLogs<br/>  BLUE RIVER<br/>  mavix<br/>  Terrotron<br/>  prospero         </p>                               | <h3>COMMERCE</h3> <p>  Instacart<br/>  STITCH FIX<br/>  RetailNext<br/>  GoodGood         </p>                                                                        | <h3>OTHER</h3> <p>  eHarmony<br/>  stem<br/>  BOXEVER<br/>  deepto<br/>  Unbabel<br/>  Second Spectrum         </p> |                                                                                                                  |

## CROSS-INFRASTRUCTURE/ANALYTICS

OPEN SOURCE

The diagram illustrates a comprehensive ecosystem of data science and engineering tools, organized into 12 functional categories:

- FRAMEWORK**: Includes Hadoop, MapReduce, YARN, Flink, Spark, Tez, and CDAP.
- QUERY / DATA FLOW**: Includes Spark SQL, Presto, SLAMDATA, ARACHNE DRILL, MESOS, and Google Cloud Dataflow.
- DATA ACCESS**: Includes nifi, mongoDB, cassandra, HBase, CouchDB, OPENSTACK, riak, Spanner, and accumulo.
- COORDINATION**: Includes talend, Apache Zookeeper, and Apache Ambari.
- STREAMING**: Includes Spark, Flink, B beam, kafka, druid, and STORM.
- STAT TOOLS**: Includes python, R, ScalaLab, Numpy, and SciPy.
- AI / MACHINE LEARNING / DEEP LEARNING**: Includes tensorflow, theano, Caffe, CNTK, DM TK, Keras, PyTorch, Apache SINGA, OpenAI, FeatureFu, neon, milib, DL4J, Chainer, DUMSUM, DSSTNE, WEKA, and Aerosolve.
- SEARCH**: Includes elasticsearch, Solr, and Lucene.
- LOG ANALYSIS**: Includes elasticsearch, kibana, logstash, and BEAKER.
- VISUALIZATION**: Includes Rodeo.
- COLLABORATION**: Includes jupyter, Zeppelin, and ANACONDA.
- SECURITY**: Includes Apache Ranger, KNOX, and Sentry.

## DATA SOURCES & APIS

HEALTH IOT FINANCIAL & ECONOMIC DATA AIR / SPACE / SEA PEOPLE / ENTITIES LOCATION INTELLIGENCE OTHER

Apple Jawbone Validic Garmin Fitbit Human API Kinso

GE Digital Uptake ThingWorx Helium Samsara

Bloomberg Thomson Reuters Dow Jones S&P Capital IQ CB Insights Xignite Quandl Yodlee Premise Estimize Second Measure Forium Eagle Alpha StockTwits Plaid Maternmark Thinknum

Planet Labs Airware Spire SkyCatcher Airbotics JPL-Caltech Tellicus Labs WinWarden DroneDeploy ManuTraffic

Axiom Experian Epsilon InsideView Crimson Hexagon Quantcast Basis SafeGraph

Foursquare Sense Mapbox PlaceIQ Esri Factual Carto Mapillary StreetLine

Qualtrics Data.gov Data.world Panjiva Enigma

## DATA RESOURCES

**INCUBATORS & SCHOOLS**

- PLURALSIGHT
- GA
- galvanize
- DataCamp
- DataElite
- The Data Incubator
- METIS

**RESEARCH**

- facebook research
- OpenAI
- MIRI  
BOLD NEW APPROACHES TO AI
- CSAIL
- ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE



## BIG DATA &amp; AI LANDSCAPE 2018





## DATA &amp; AI LANDSCAPE 2019

## INFRASTRUCTURE

The diagram illustrates the ecosystem of data processing technologies, categorized into three main groups:

- HADOOP ON-PREMISE:** Includes Cloudera, Hortonworks, MAAP, Pivotal, IBM InfoSphere, and jethro.
- HADOOP IN THE CLOUD:** Includes AWS, Microsoft Azure, Google Cloud, SAP Cloud Platform, IBM InfoSphere BigInsights, and CRM.
- STREAMING / IN-MEMORY:** Includes Amazon Kinesis, Databricks, Azure Cloud Platform, Oracle, and Confluent.

Below these categories, a row of additional tools is shown, including Strim, Hazelcast, GigaSpaces, Wallaroo, GridGain, and others.

The diagram illustrates the database landscape, categorized into six main groups:

- NOSQL DATABASES:** Includes Google Cloud, AWS, Oracle, Microsoft Azure, MongoDB, MarkLogic, Couchbase, DataStax, Redis, Aerospike, and ArangoDB.
- NEWSQL DATABASES:** Includes SAP, Clustrix, Pivotal, Microsoft, IBM, MEMSQL, InfluxDB, Cockroach Labs, VoltDB, and Splice.
- GRAPH DBs:** Includes Neo4j, Amazon Neptune, Oracle, and InfoGraph.
- MPP DBs:** Includes Teradata, Vertica, SAP HANA, and Exasol.
- CLOUD EDW:** Includes Google Cloud, Microsoft Azure, Pivotal, Snowflake, and Informatica.
- SERVERLESS:** Includes Amazon, Microsoft Azure, Pulsar, and Nuclia.

**DATA TRANSFORMATION**

- talend
- pentaho
- alteryx
- trifacta
- tm9
- Paxata
- StreamSets
- UNIFI

**DATA INTEGRATION**

- SAP Data Services
- Informatica
- Microsoft
- TEALUM
- inlogic
- enigma
- Segment
- ATTENTIFY
- Informatica
- ZALONI
- import.io
- InfoWorks
- Fluxion
- SNOWFLAKE
- MATILLION

**DATA GOVERNANCE**

- Informatica
- IBM
- Scale
- Spiky
- Cloud
- collibra
- Alation
- Warcorn
- collibra
- dremio
- Alation
- Warcorn
- OKERA
- data.world

**MGMT / MONITORING**

- AWS
- New Relic
- actifio
- rubrik
- APPDYNAMICS
- dynabrace
- WAFERLESS
- Signifia
- druid
- splunk
- Microsoft
- pagerduty
- Sumo Logic
- Namur
- substack
- zenoss
- OpsRamp
- MAGNETUDE

[illegible]

## ANALYTICS & MACHINE INTELLIGENCE

The diagram is divided into two main sections by a horizontal line. The top section is titled 'DATA ANALYST PLATFORMS' and contains logos for Microsoft, Pentaho, Alteryx, Digital Reasoning, QlikView, SAS, Tableau, and IBM. The bottom section is titled 'DATA SCIENCE PLATFORMS' and contains logos for IBM, Databricks, Dataiku, Domino, Rapidminer, Tibco, Anaconda, SAS, Alteryx, Cerner, H2O, Endor, and Kognition. Below the logos, there are two rows of text: 'ATTN/O Datameer incorta.' and 'inter/ana. H2O MODE ENDOR'.

The collage is organized into three vertical columns, each with a header in red capital letters. The first column, 'BI PLATFORMS', features logos for Looker, Amazon Analytics, AWS, Domo, Alteryx, Tableau, ThoughtSpot, ATSCale, IBM, Qlik, Google Data, Informatica, and Birst. The second column, 'VISUALIZATION', includes logos for Tableau, Power BI, SAP, Google Cloud, Celonis, Zephi, Xoriant, and Chartio. The third column, 'MACHINE LEARNING', displays logos for Azure, Amazon SageMaker, Google Cloud, H2O, DataRobot, gamalon, VIZEN, and Element.

**COMPUTER VISION**

- Microsoft Azure
- Amazon Rekognition
- Clarifai
- EVERAI
- deepomatic
- neurala
- twentyton
- UBIQUITY
- AEE
- YITU
- TRAX
- FLANNET
- Synthesia
- Quintana

**HORIZONTAL AI**

- IBM Watson
- Cortana
- Baidu 百度
- Sentient
- Voyager Labs
- Altera
- Affective
- PROPHESIE
- Numenta
- PETUUM
- norlogics
- ORORO.AI
- OSARO
- BLUE GENOMICS
- Fukuda
- Red Hat

**SPEECH & NLP**

- Google Cloud
- twilio
- amazon alexa
- Amazon Translate
- narative speech
- semanticsearch
- Mobvoi
- EigenTechnologies
- PRIMER
- SoundHound Inc.
- VOICETECH
- Mindfield
- VOICEBOX
- Source
- cogito
- sprinklr
- SHALUTALK
- UP
- Paychex

**SEARCH**

- elasticsearch
- ORACLE  
EMERGA
- algolia
- COVEO
- Lucidworks
- ATTIVO
- swiftype
- EXALTO
- alphasense
- MAANA
- omni-us
- SINEOQA

**LOG ANALYTICS**

- splunk
- sumologic
- solarwinds
- LOGICMONITOR
- TIMBER
- wobano
- logz.io

**SOCIAL ANALYTICS**

- Hootsuite
- sprinklr
- NETBASE
- synthesio
- TRUCK
- simplereach
- bittly
- SimilarWeb

**WEB / MOBILE /  
COMMERCE ANALYTICS**

- Google Analytics
- mixpanel
- AMPLITUDE
- Airtop
- RESCLOUD
- ixigpoint
- gratify
- custora

## APPLICATIONS – ENTERPRISE

[illegible]

## — APPLICATIONS – INDUSTRY

[illegible]

**HEALTHCARE**

flatiron Clover Xyus DayStar  
 METABOTA Ginkgo Goldy babybon  
 3Med zebra PINDA OVIA  
 TEMPUUS patens/kine AIcure insiro  
 HANCOCK citizen  
 notable Kinetix/CTF reuron  
 e-nthic magenta Biochthon BAYLARS  
 Qventus ARTERY IMAGEN  
 BLUEbird PAIGE DAYMANT  
 innovacore 3D Systems LocusTus

**LIFE SCIENCES**

Amendable Cargene  
 bioinventAI verity  
 WuXiNextCODE  
 ZINC  
 Clear Labs Finnmark  
 DYNASYS  
 Phosphor  
 CITRINE  
 twoSTAR  
 Altemus  
 COWIN  
 Sphera

**TRANSPORTATION**

UBER TESLA  
 CLEARPATH Cruise  
 nuro  
 drive.ai  
 nauto AMMOTIVE  
 G7  
 NIOM OPTIMUS  
 moovit ilac  
 nexar Kodiak comma.ai  
 nethadine  
 think INRIX  
 CIVIL Maps

**AGRICULTURE**

FARMERS  
 Granular  
 John Deere  
 BLUEBERRYER  
 Farmersedge  
 AgroStar  
 FarmLogs  
 TARANIS  
 GAMMA  
 CROPSHOP  
 CROSPERO

**COMMERCE**

Instacart  
 K STITCH FIX  
 De & Co  
 Bluebird  
 Newfrontier  
 Neudorff  
 Other

**INDUSTRIAL**

AVEVA SIEMENS  
 PREDIX UPTAKE  
 SCORTEX KENEXA  
 TACHYUS  
 eharmony stem Amper ByteDance  
 veridien select SOIERN BUXEVEY  
 HODGKINS duoeto  
 GAYMA  
 Electric ZINIR  
 Moxiva Spole

## CROSS-INFRASTRUCTURE/ANALYTICS



OPEN SOURCE

The diagram is a horizontal bar chart representing various data science and machine learning frameworks, tools, and languages. The bar is divided into 12 colored segments, each representing a different category. Each segment contains logos for tools and frameworks in that category.

- FRAMEWORKS**: Spark, Flink, Hadoop, Mesos, YARN, Tez, MapReduce, HBase, HDFS, Hive, Pig, Mahout, Hama, HBase, HDFS, Hive, Pig, Mahout, Hama.
- QUERY / DATA FLOW**: Spark, SQL, Presto, Trino, Flink, Kudu, HBase, HDFS, Hive, Pig, Mahout, Hama.
- DATA ACCESS & DATABASES**: Cassandra, MongoDB, Redis, Cockroach Labs, Druid, HBase, HDFS, Hive, Pig, Mahout, Hama.
- ORCHESTRATION & MGMT**: Talend, Apache Airflow, Apache Beam, Apache Spark, Apache Flink, Apache Hama, Apache Mahout, Apache Pig, Apache Hive, Apache HDFS.
- STREAMING & MESSAGING**: Spark, Flink, Kafka, Storm, Apache RocketMQ, Apache Pulsar, Apache Kudu, Apache HBase, Apache HDFS, Apache Hive, Apache Pig, Apache Mahout, Apache Hama.
- STAT TOOLS & LANGUAGES**: Jupyter, R, Scala, Python, Julia, SAS, SPSS, Stata, RStudio, JupyterLab, RStudio, JupyterLab, RStudio, JupyterLab.
- AI OPS & INFRA**: MLOps, MLflow, DVC, Seldon, Pachyderm, Kubeflow, Argo, Tekton, Jenkins, GitLab CI, GitHub Actions, CircleCI, Travis CI, Jenkins, GitLab CI, GitHub Actions, CircleCI, Travis CI.
- AI / MACHINE LEARNING / DEEP LEARNING**: TensorFlow, Keras, PyTorch, MXNet, Caffe, Theano, PySpark, Databricks, AWS SageMaker, Google Cloud AI, Microsoft Azure ML, IBM Watson Studio, SAP HANA Cloud, Oracle Data Science, SAP HANA Cloud, Oracle Data Science.
- SEARCH**: Elasticsearch, Solr, OpenSearch, Meilisearch, Algolia, Elasticsearch, Solr, OpenSearch, Meilisearch, Algolia.
- LOGGING & MONITORING**: ELK Stack (Elasticsearch, Logstash, Kibana), Prometheus, Grafana, InfluxDB, DistroX, ELK Stack (Elasticsearch, Logstash, Kibana), Prometheus, Grafana, InfluxDB, DistroX.
- VISUALIZATION**: Matplotlib, Tableau, Power BI, QlikView, Qlik Sense, Qlik Nxt, Qlik Sense, Qlik Nxt, Qlik Sense, Qlik Nxt.
- COLLABORATION**: BeakerX, Jupyter, JupyterLab, Jupyter Notebook, Jupyter Notebook, Jupyter Notebook, Jupyter Notebook.
- SECURITY**: Apache Ranger, Apache Knox, Apache Sentry, Apache Hadoop, Apache Spark, Apache Flink, Apache Beam, Apache Airflow, Apache Airflow, Apache Airflow.

## DATA SOURCES & APIs

**HEALTH**

- Apple
- VALIDIC
- practicefusion
- fitbit
- GARMIN
- HUMAN API
- kinsa
- MIMIC

**IOT**

- GE Digital
- UPTAKE
- thingworx
- helium
- samsara
- edwards

**FINANCIAL & ECONOMIC DATA**

- Bloomberg
- THOMSON REUTERS
- DOW JONES
- SEI CAPITAL IQ
- CB INSIGHTS
- PLAID
- SECOND MEASURE
- YODLEE
- YODLEE
- estimize
- PREMISE
- Quandl
- Engel Alpha
- StockTwits
- xignite
- Thinkbox
- earnest
- predata

**AIR / SPACE / SEA**

- Orbital Insight
- planet
- SEYCAT
- AIRBOTICS
- spire
- PROTEOMICS
- keospy
- UNDERSTORY
- telluslabs
- WINDWARD
- brunoDeploy
- MarineTraffic
- GOAT GLOBAL
- JS FRONTS

**PEOPLE / ENTITIES**

- axiom
- experian
- EPILON
- InsideView
- Crimson Hexagon
- BASIS
- Quantcast
- SAFE GRAPH

**LOCATION INTELLIGENCE**

- FOURSQUARE
- Mapbox
- mapbox
- sense360
- pluribus
- HEXAON
- PlaceIQ
- esri
- factual
- CART
- Mapillary
- Streetside
- cuebiq
- ARadar
- OpenStreetMap

**OTHER**

- DATA.GOV
- IMAGENET
- LORE
- IMMUNOCORE
- CRUX
- graffiti.io

## DATA RESOURCES

**DATA SERVICES**

- OPERA
- UCD
- DATA SCIENCE
- fractalx
- kaggle
- DataKind
- EXEL
- innoplexus

**INCUBATORS & SCHOOLS**

- PLURALSIGHT
- DataCamp
- DataKind
- INSIGHT
- The Data Incubator
- METIS

**RESEARCH**

- OpenAI
- facebook research
- MIRI
- VECTOR INSTITUTE
- CSAIL
- AI2
- ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE

# DATA & AI LANDSCAPE 2020

## INFRASTRUCTURE

### STORAGE

Amazon S3, Google Cloud Storage, Microsoft Azure Storage, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

### HADOOP CLOUDERA

Google Cloud Dataflow, Amazon EMR, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA LAKES

Amazon Lake Formation, AWS Lake, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA WAREHOUSES

Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### STREAMING / IN-MEMORY

Amazon Kinesis, Google Cloud Dataflow, Microsoft Azure Stream Analytics, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## ANALYTICS & MACHINE INTELLIGENCE

### BI PLATFORMS

Looker, Tableau, Power BI, Qlik, Alteryx, MicroStrategy, Kognia

### VISUALIZATION

Tableau, Power BI, Qlik, Alteryx, MicroStrategy, Kognia

### DATA ANALYST PLATFORMS

Microsoft, Pentaho, Alteryx, Qlik, Alteryx, MicroStrategy, Kognia

## APPLICATIONS - ENTERPRISE

### SALES

Salesforce, HubSpot, Pardot, Marketo, Eloqua, Braze, Braze

### MARKETING - B2B

HubSpot, Pardot, Marketo, Eloqua, Braze, Braze

### MARKETING - B2C

HubSpot, Pardot, Marketo, Eloqua, Braze, Braze

### CUSTOMER EXPERIENCE / SERVICE

HubSpot, Pardot, Marketo, Eloqua, Braze, Braze

### HUMAN CAPITAL

HubSpot, Pardot, Marketo, Eloqua, Braze, Braze

## DATA SCIENCE & MACHINE INTELLIGENCE

### NoSQL DATABASES

Amazon DynamoDB, Google Cloud Firestore, Microsoft Azure Cosmos DB, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

### NewSQL DATABASES

Amazon DynamoDB, Google Cloud Firestore, Microsoft Azure Cosmos DB, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

### GRAPH DBs

Amazon DynamoDB, Google Cloud Firestore, Microsoft Azure Cosmos DB, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

### MPP DBs

Amazon DynamoDB, Google Cloud Firestore, Microsoft Azure Cosmos DB, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

### SERVER-LESS

Amazon DynamoDB, Google Cloud Firestore, Microsoft Azure Cosmos DB, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

### CLUSTER SVCS

Amazon DynamoDB, Google Cloud Firestore, Microsoft Azure Cosmos DB, IBM Cloud Object Storage, Oracle Cloud Infrastructure, VAST, CUMMIS

## DATA SCIENCE & MACHINE INTELLIGENCE

### DATA SCIENCE NOTEBOOKS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA SCIENCE PLATFORMS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### MACHINE LEARNING

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## APPLICATIONS - ENTERPRISE

### LEGAL

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### REGTECH & COMPLIANCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### FINANCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AUTOMATION & RPA

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### SECURITY

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## ETL / DATA TRANSFORMATION

### ETL / DATA TRANSFORMATION

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA INTEGRATION

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA GOVERNANCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA QUALITY

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## COMPUTER VISION

### COMPUTER VISION

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### HORIZONTAL AI

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### SPEECH & NLP

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## APPLICATIONS - INDUSTRY

### ADVERTISING

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### EDUCATION

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### REAL ESTATE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### GOVT & INTELLIGENCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### COMMERCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### FINANCE - LENDING

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### INSURANCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## MGMT / MONITORING

### MGMT / MONITORING

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA GENERATION & LABELLING

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AI OPS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### GPU DBs & CLOUD

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AI HARDWARE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## SEARCH

### SEARCH

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### LOG ANALYTICS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### SOCIAL ANALYTICS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### WEB / MOBILE / COMMERCE ANALYTICS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## HEALTHCARE

### HEALTHCARE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### LIFE SCIENCES

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### TRANSPORTATION

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AGRICULTURE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### INDUSTRIAL

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### OTHER

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## OPEN SOURCE

### FRAMEWORKS

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### QUERY / DATA FLOW

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### DATA ACCESS & DATABASES

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### ORCHESTRATION & PIPELINES

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### STREAMING & MESSAGING

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### STAT TOOLS & LANGUAGES

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AI OPS & INFRA

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AI / MACHINE LEARNING / DEEP LEARNING

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### SEARCH

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### LOGGING & MONITORING

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### VISUALIZATION

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### COLLABORATION

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### SECURITY

TensorFlow, PyTorch, Keras, Caffe, MXNet, PySpark, Databricks, Hadoop, Spark, Flink, Storm, Kafka, RabbitMQ, Redis, MongoDB, PostgreSQL, MySQL, Oracle, IBM DB2, SAP HANA, Microsoft SQL Server, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse, IBM Power Systems, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## DATA SOURCES & APIs

### DATA MARKETPLACES & DISCOVERY

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### FINANCIAL & ECONOMIC DATA

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### AIR / SPACE / SEA

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### PEOPLE / ENTITIES

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### LOCATION INTELLIGENCE

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### OTHER

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

## DATA RESOURCES

### DATA SERVICES

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

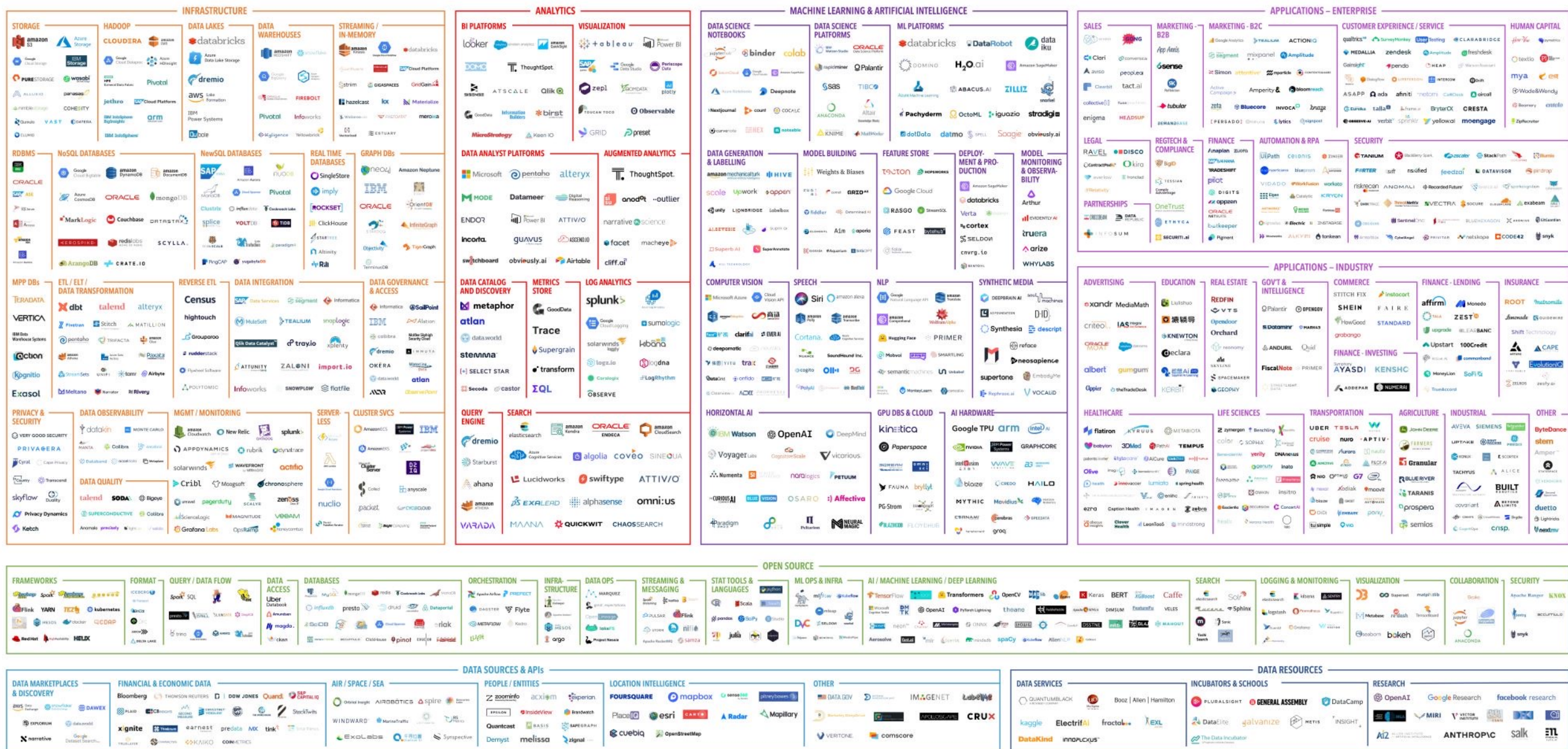
### INCUBATORS & SCHOOLS

Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM

### RESEARCH

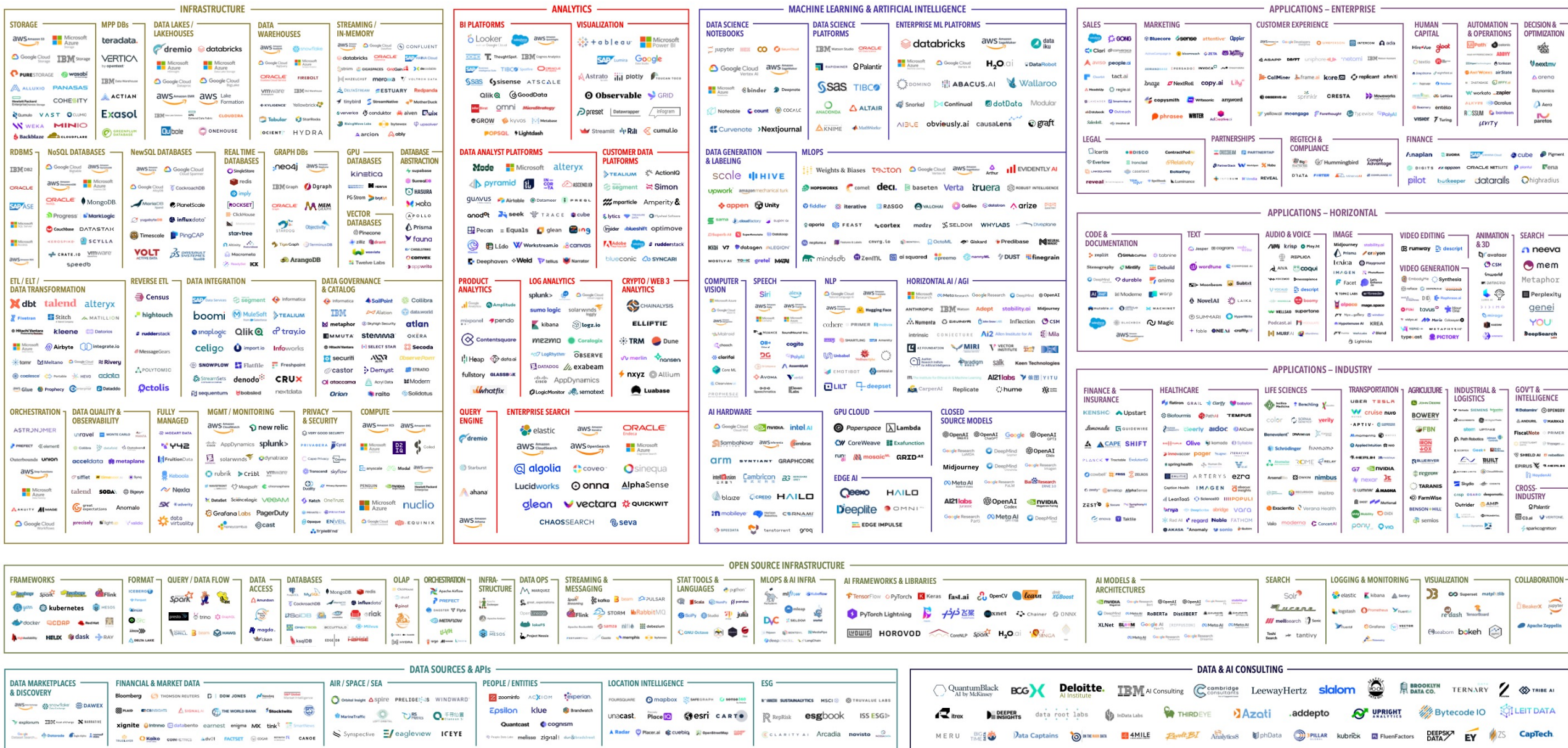
Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, IBM Watson Studio, SAP Cloud Platform, Oracle Cloud Infrastructure, IBM InfoSphere, QRM





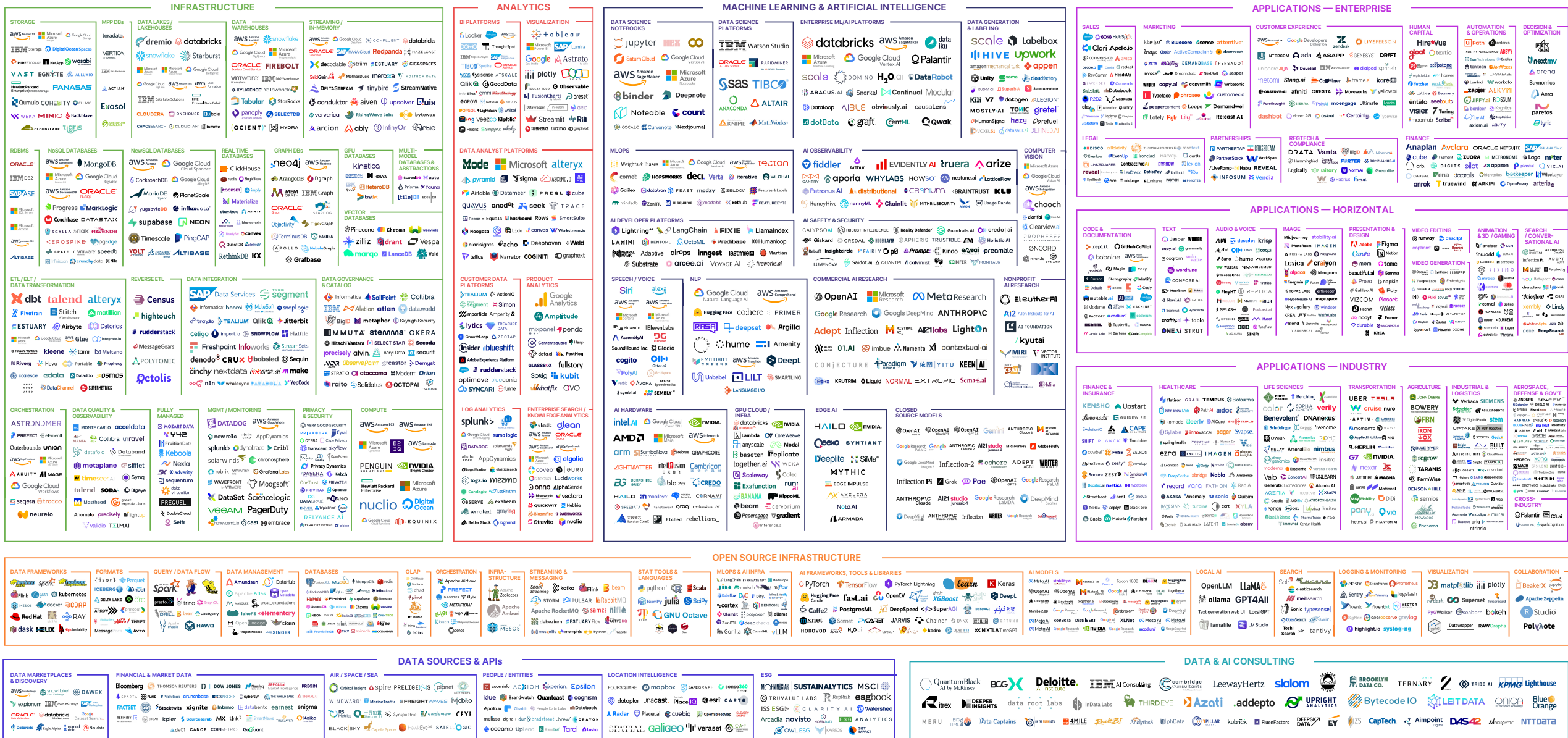


# THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE





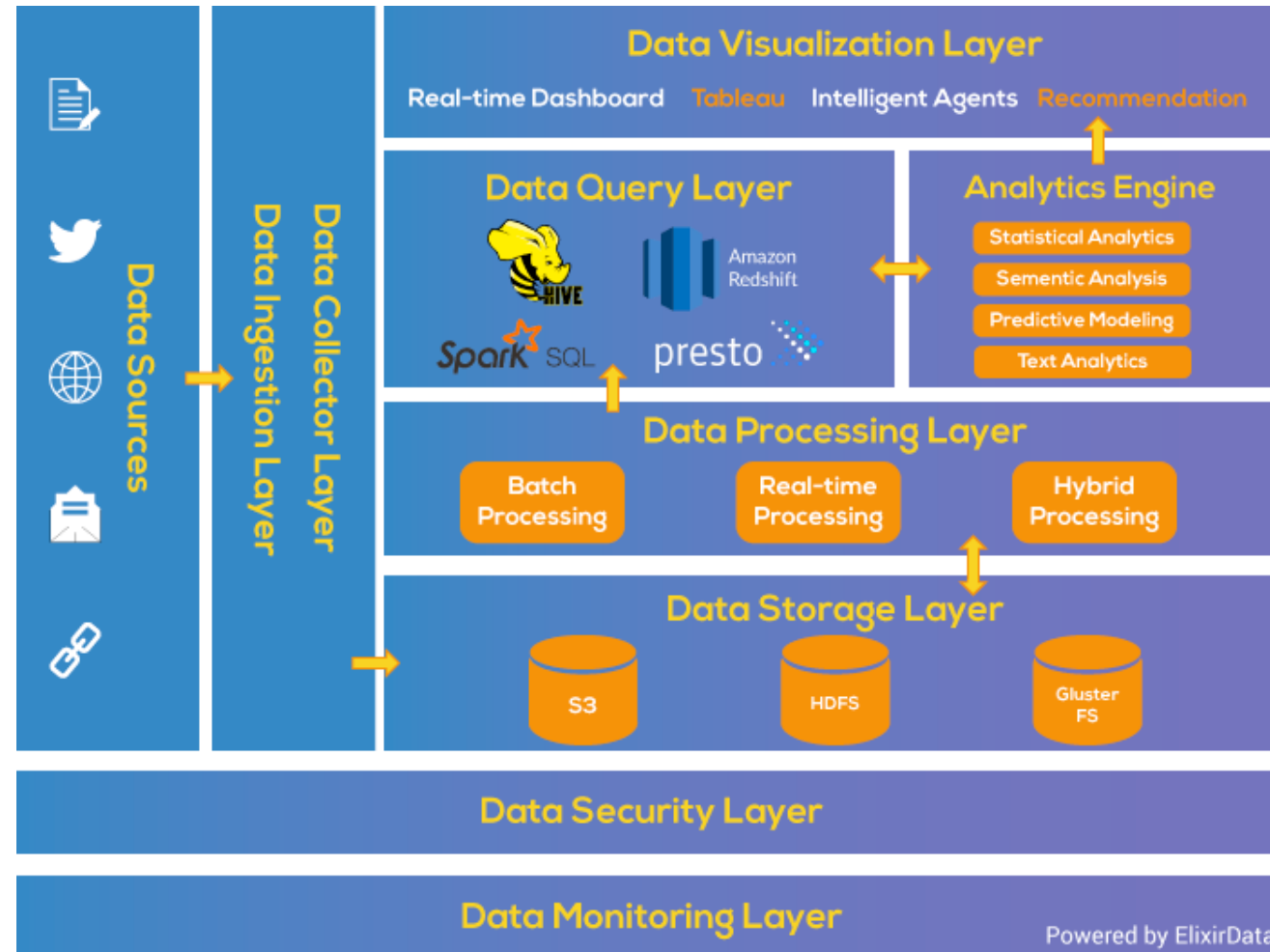
## THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE &amp; DATA) LANDSCAPE





# Big Data Framework

# Big Data Framework



# Data Ingestion Layer

- Ingest data flows from hundreds or thousands of various sources into data center (variable speed, different formats, etc.), extracting the data, and detecting the changed data.
- It's rightly said that "If starting goes well, then, half of the work is already done"
- Data can be streamed in real time or ingested in batches
- Prioritizing data sources, validating individual files and routing data items to the correct destination.
- Data serialization (Apache Thrift, Apache Avro, etc.) is the concept of converting data into a (compressed) format that allows it to be shared or stored such that its original structure can be recovered.
- Few data ingestion tools: Apache Flume, Apache Nifi, Elastic Logstash



# Data Collector Layer

- The focus is on transportation data from ingestion layer to rest of data pipeline.
- Often, we use a messaging system that will act as a mediator between all the programs that can send and receive messages.
- Data Collector Tools: Apache Kafka

# Data Processing Layer

- The focus is to specialize data pipeline processing system
- Processing can be done in 3 ways:
  - Batch Processing System : A pure batch processing system for off-line analytic. For example, Apache Sqoop efficiently transfers bulk data between Apache Hadoop and structured datastores such as relational databases.
  - Near Real Time Processing System: A pure online processing system for on-line analytic. For example, the Apache Storm cluster makes decisions about the criticality of the event and sends the alerts to the alert system (dashboard, e-mail, other monitoring systems).
  - Hybrid Processing system: This consist of batch and real-time processing system capabilities. For this type of processing, tool used can be Apache spark and apache Flink.

# Data Storage Layer

- Keep data in the right place based on usage
- A new concept is useful: **polyglot persistence** is the idea of using multiple databases to power a single application.
- Polyglot persistence is the way to share or divide your data into multiple databases and leverage their power together.
- Tools used for Data Storage: HDFS, Gluster file systems (GFS), Amazon S3

# Data Query Layer

- This is the layer where strong analytic processing takes place.
- This is a field where interactive queries are necessities and it's a zone traditionally dominated by SQL expert developers.
- Companies from all industries use analytics queries to:
  - Increase revenue
  - Decrease costs
  - Increase productivity
- Tools Used for Analytics Query: Apache Hive, Spark SQL, Amazon Redshift



# Analytics Engine

- Analytics engine is the use of advanced analytic techniques against very large, diverse data sets.
- Data analytics became an essential step while computing such a large amount of data.
- Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable.
- Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing.
- Analytics tools: Apache Spark, Machine learning libraries, R

# Data Visualization Layer

- This layer focus on Big Data Visualization. We need something that will grab people's attention, pull them in, make your findings well-understood.
- That's why it provides full business infographics: your own findings from your own data need the annotation and the bold canvas.
- The data visualization layer often is the thermometer that measures the success of the project. This is the where the data value is perceived by the user.
- Few tools used for dashboards: Tableau, Kibana, AngularJS

# Data Security Layer

- Security is the primary task for any kind of work.
- Security should be implemented at all layers of the lake starting from ingestion, through storage, analytics, discovery, all the way to consumption.

# Data Monitoring Layer

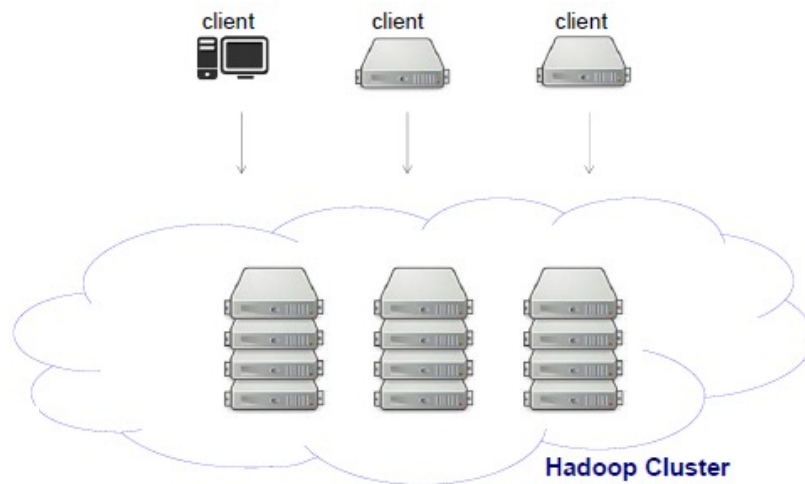
- Data in enterprise systems is like food: it has to be kept fresh, also it needs nourishment, otherwise it goes bad and doesn't help you in making strategic and operational decisions. Just as consuming spoiled food could make you sick, using “spoiled” data may be bad for your organization's health.
- Continuous monitoring of data is an important part of the governance mechanisms.



# Big Data with a Cluster

# Main Example: Hadoop Cluster

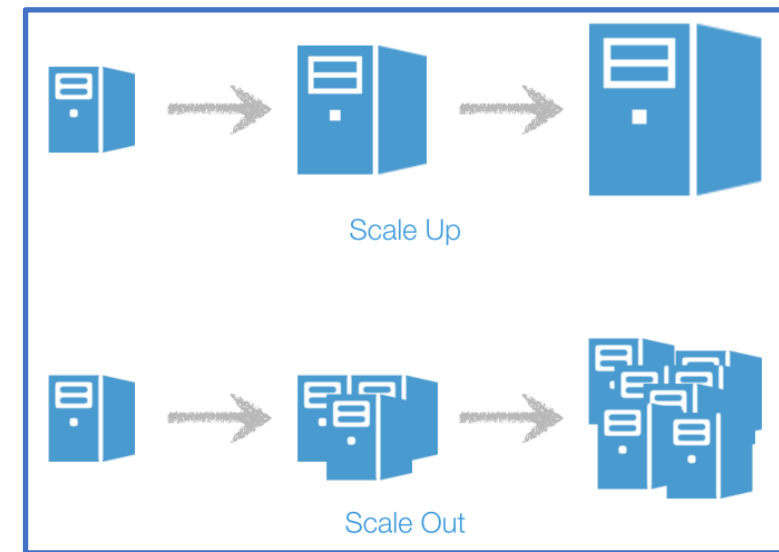
- **Many details on** <http://hadoop.apache.org>
- **A set of "cheap" commodity hardware**
  - No need for super-computers, use commodity unreliable hardware
  - Not desktops
- **Networked together**
- **May reside in the same location**
  - Set of servers in a set of racks in a data center



# Big Data System Principles with a Cluster

- Scale-Out rather than Scale-Up
- Bring code to data rather than data to code
- Deal with failures – they are common
- Abstract complexity of distributed and concurrent applications

# Scale-Out Instead of Scale-Up



- **It is harder and more expensive to scale-up**
  - Add additional resources to an existing node (CPU, RAM)
  - New units must be purchased if required resources can not be added
- **Scale-Out**
  - Add more nodes/machines to an existing distributed application
  - Software Layer is designed for node additions or removal
  - A set of nodes are bonded together as a single distributed system
  - Very easy to scale down as well

# Code to Data

- **Traditional data processing architecture**

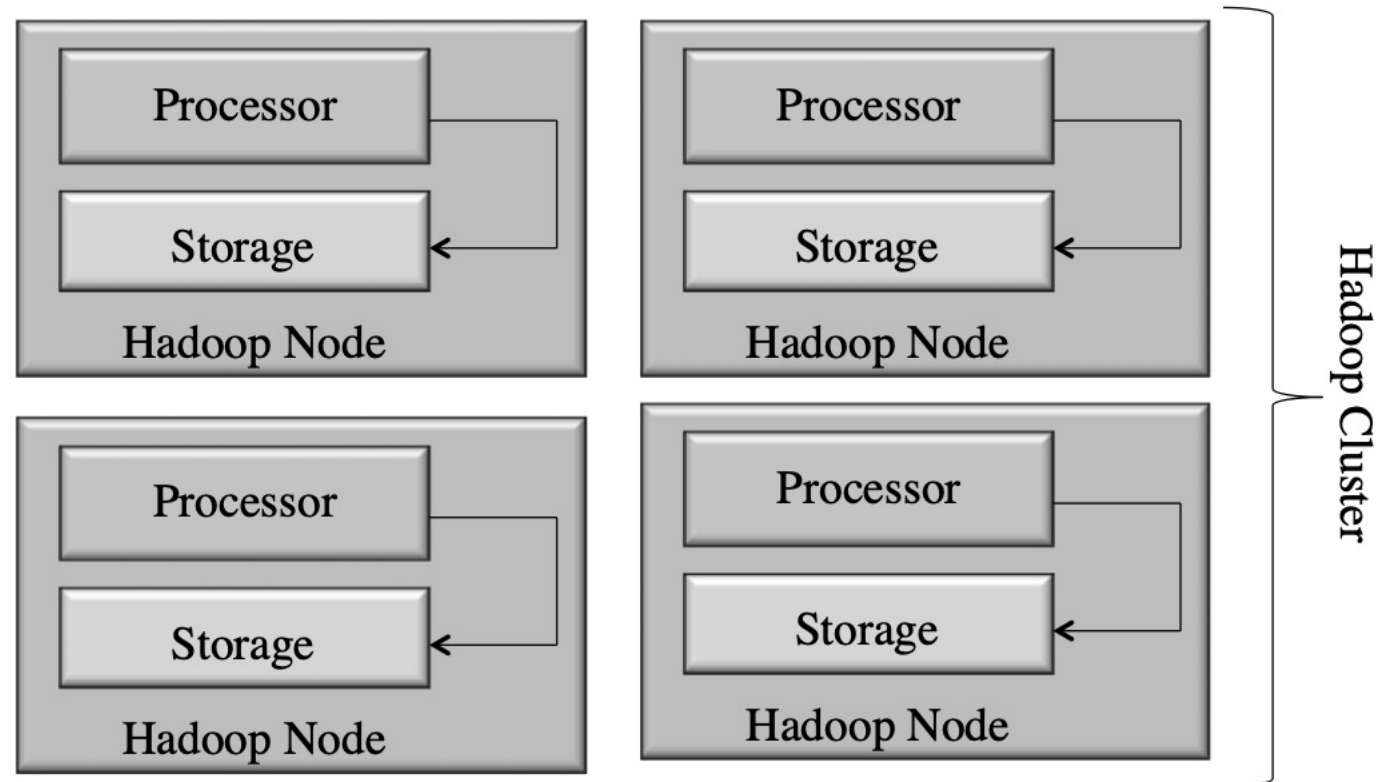
- Nodes are broken up into separate processing and storage nodes connected by high-capacity link
- Many data-intensive applications are not CPU demanding causing bottlenecks in network





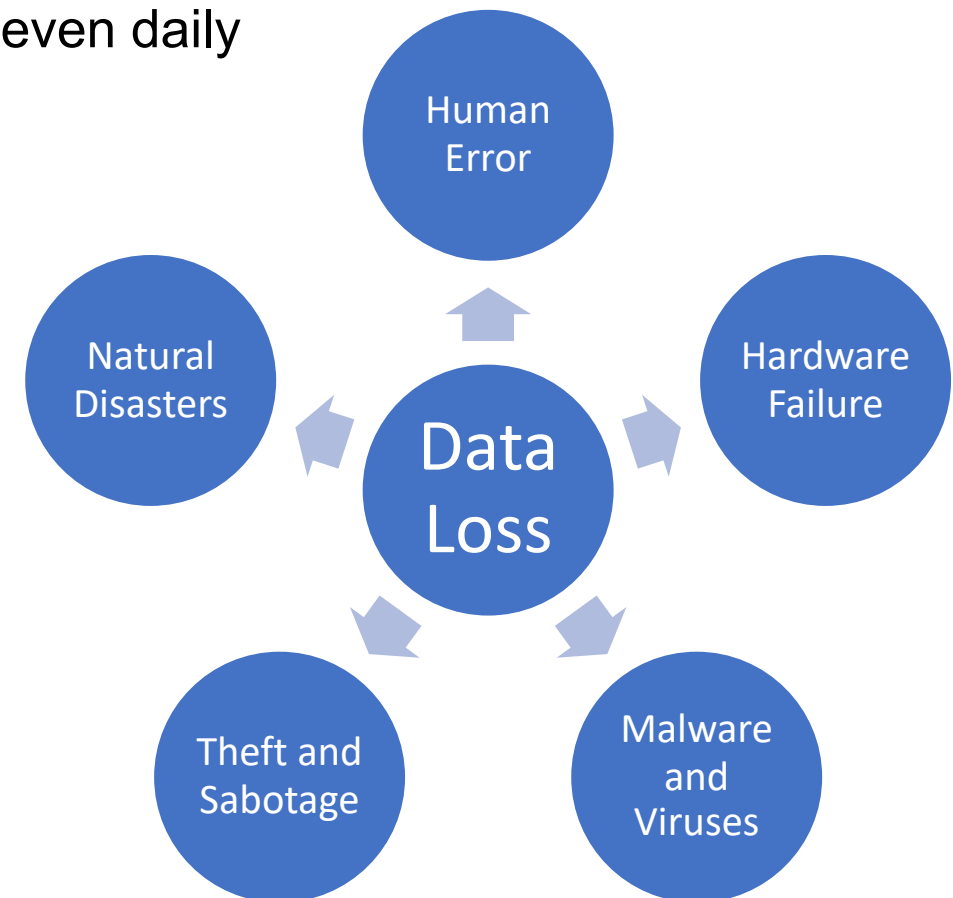
# Code to Data

- **Co-locate processors and storage**
  - Code is moved to data (size is tiny, usually in KBs)
  - Processors execute code and access underlying local storage



# Failures are Common

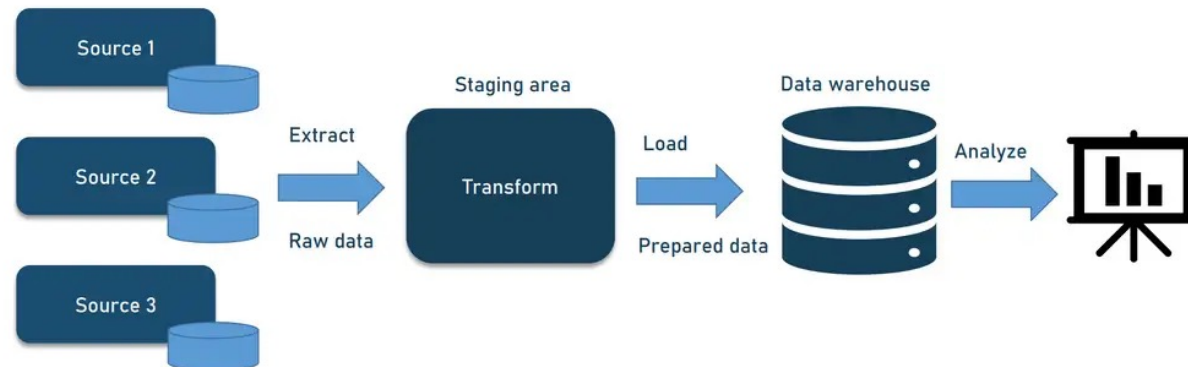
- **Given a large number machines, failures are common**
  - Large warehouses may see machine failures weekly or even daily
- **Hadoop is designed to cope with node failures**
  - Data is replicated
  - Tasks are retried



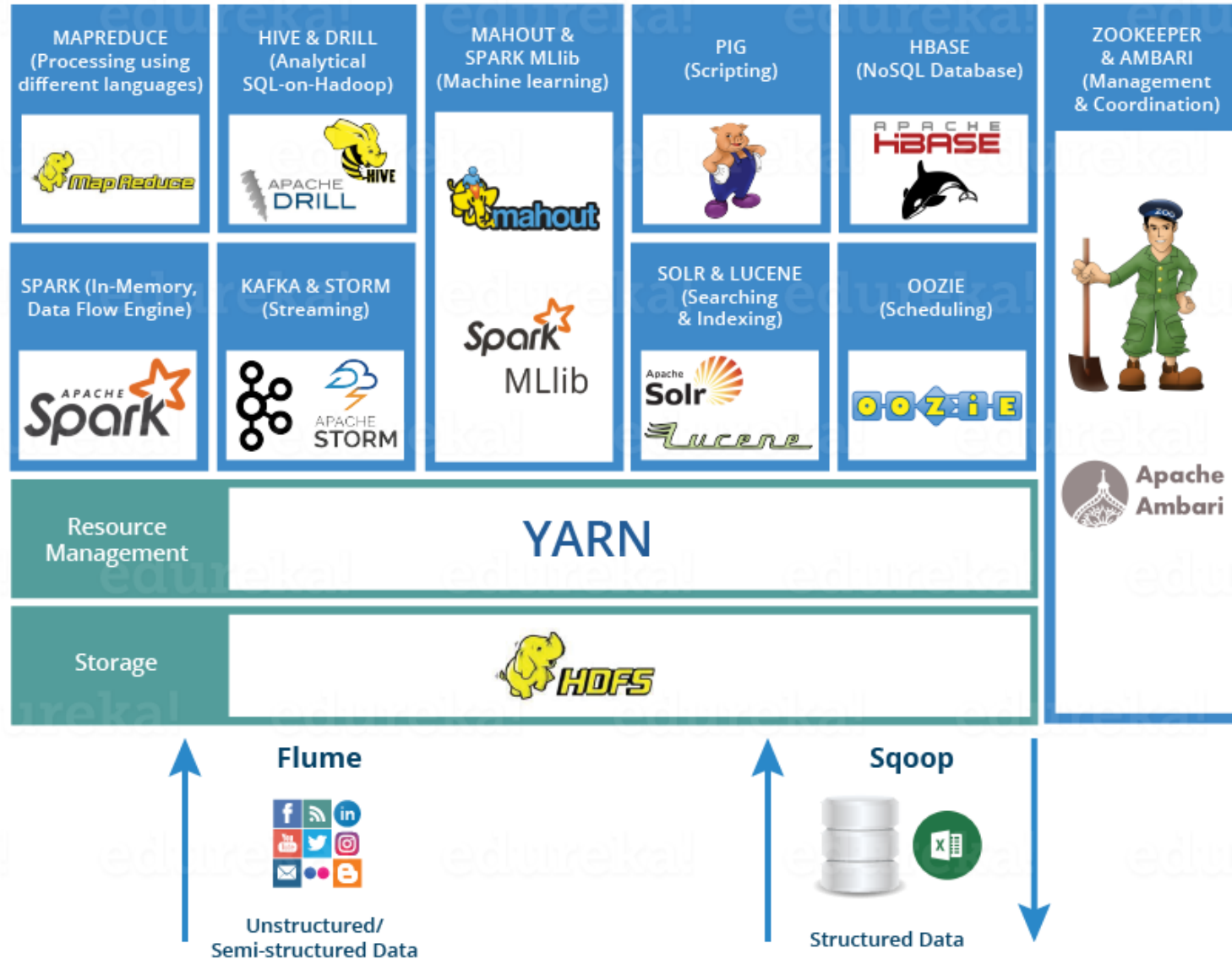
# Abstract Complexity

- **Abstract many complexities in distributed and concurrent applications**
  - Defines small number of components
  - Provides simple and well-defined interfaces of interactions between these components
- **Frees developer from worrying about system level challenges**
  - Some issues: Race conditions, data starvation
  - Some solutions: Processing pipelines, data partitioning, code distribution
- **Allows developers to focus on application development and business logic**

ETL PIPELINE



# Big Data Ecosystem (few elements)



# Many Distribution Vendors

- Cloudera Distribution for Hadoop
- MapR Distribution
- IBM Open Platform
- VMware Tanzu
- Google Cloud Platform
- Microsoft Azure



VMware Tanzu™



Google Cloud Platform





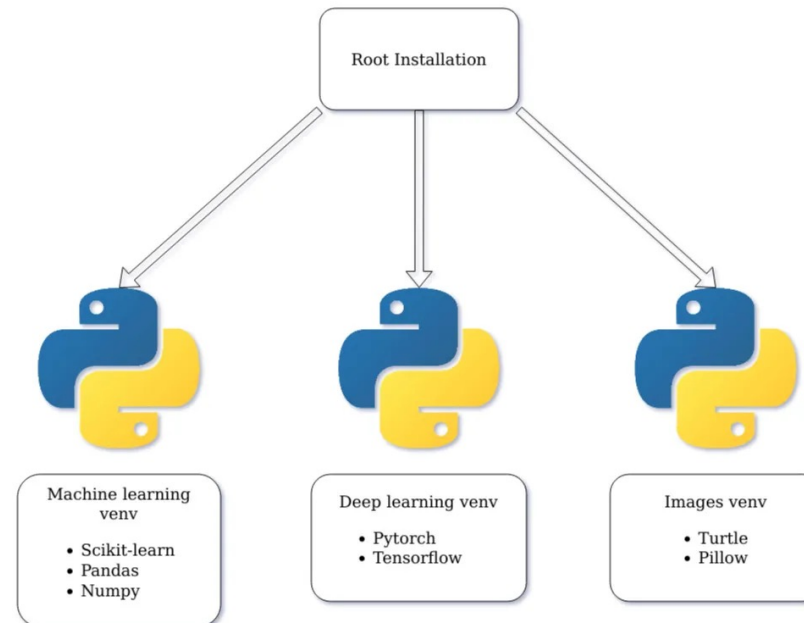
# Introduction to Docker

# Virtual Environment

# Why should we use virtual environment?

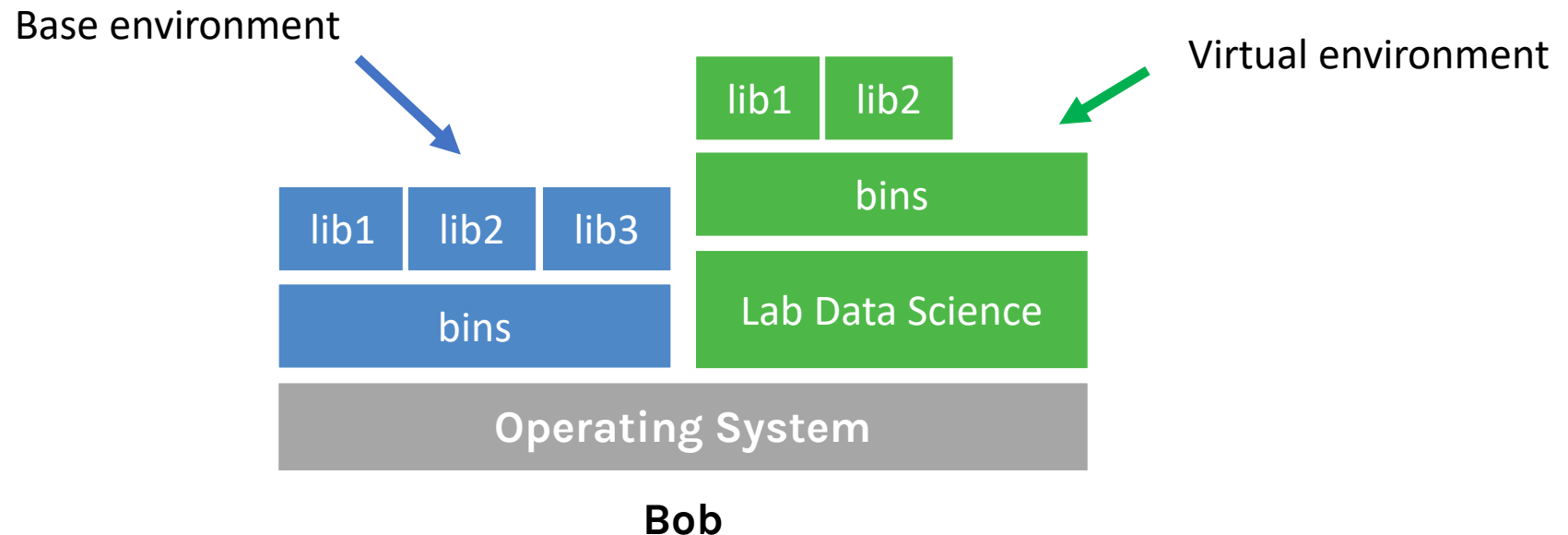
- Virtual environments help to make development and use of code more streamlined.
- Virtual environments keep dependencies in separate “sandboxes” so you can switch between both applications easily and get them running.
- Given an operating system and hardware, we can get the exact code environment set up using different technologies.

Example in Python



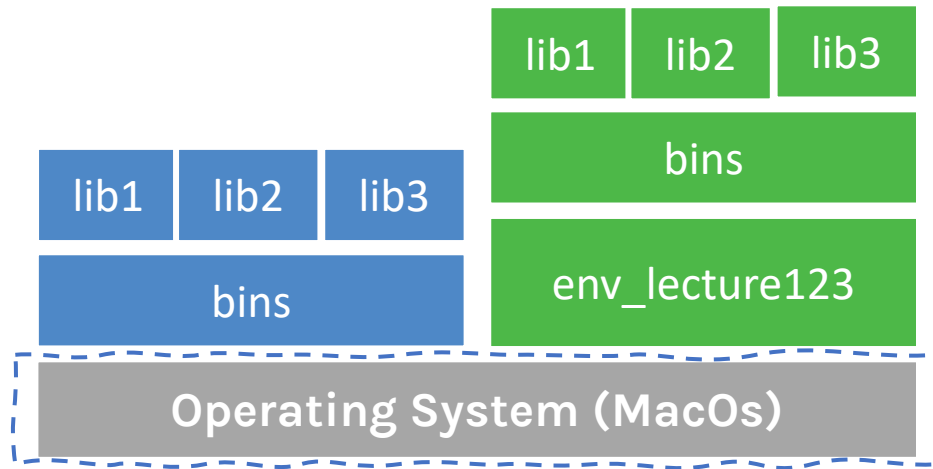
# Example with Python

- For its lab in data science, Bob thinks that it would be good to isolate the new environment from the base environment avoiding any conflict with the installed packages.
- He adds a layer of abstraction called **virtual environment** that helps him keep the modules organized and avoid misbehaviors while developing new projects.

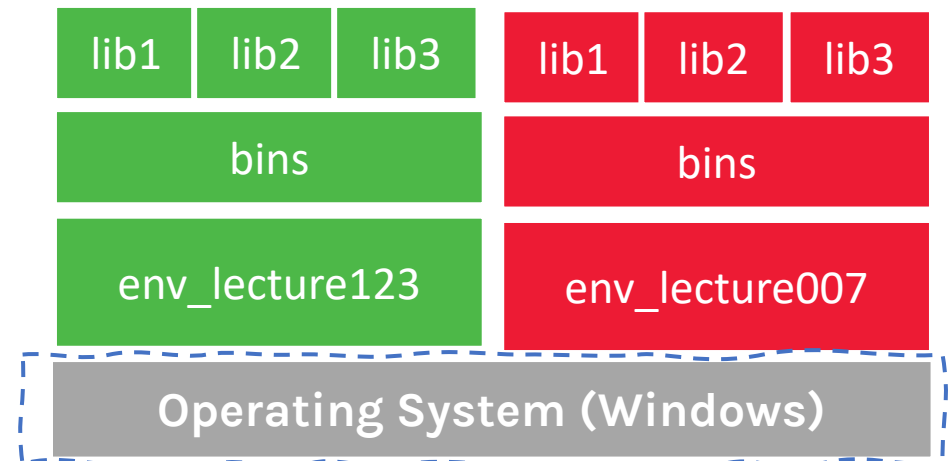


# Main drawback

- What could go wrong? Unfortunately, Bob and Alice reproduce different results and they think the issue relates to their operating systems. Indeed while Bob has a MacOS, Alice uses a Windows OS.



**Bob**



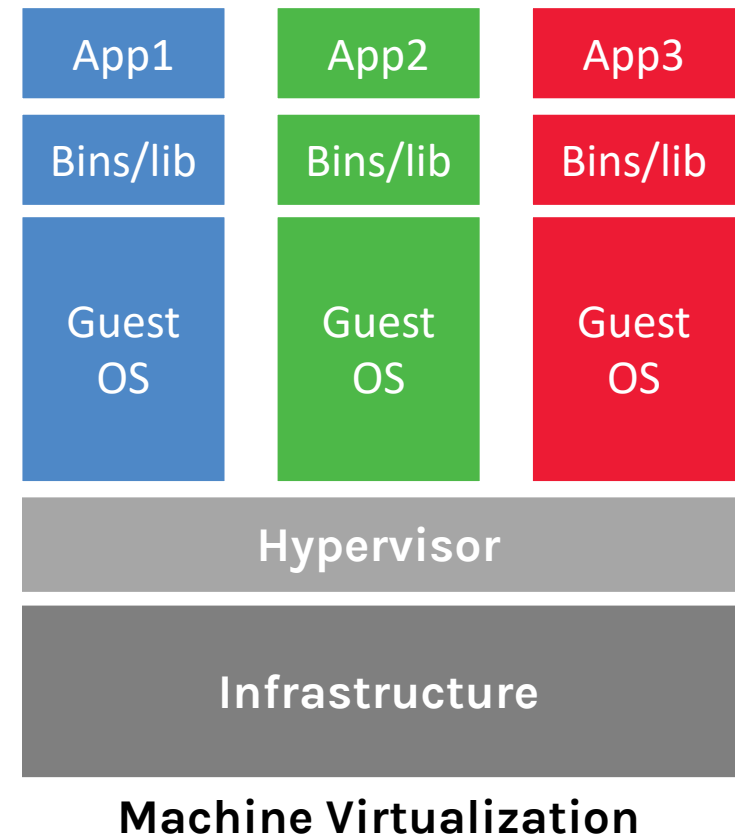
**Alice**



# Virtual Machine

# What are Virtual Machines (VM)?

- Virtual machines have their own virtual hardware: CPUs, memory, hard drives, etc.
- VMs need a hypervisor that manages different virtual machines on server
- **Hypervisor** can run as many virtual machines as you wish
- Operating system is called the « **host** » while those running in a virtual machine are called « **guest** »
- You can install a completely different operating system on this virtual machine

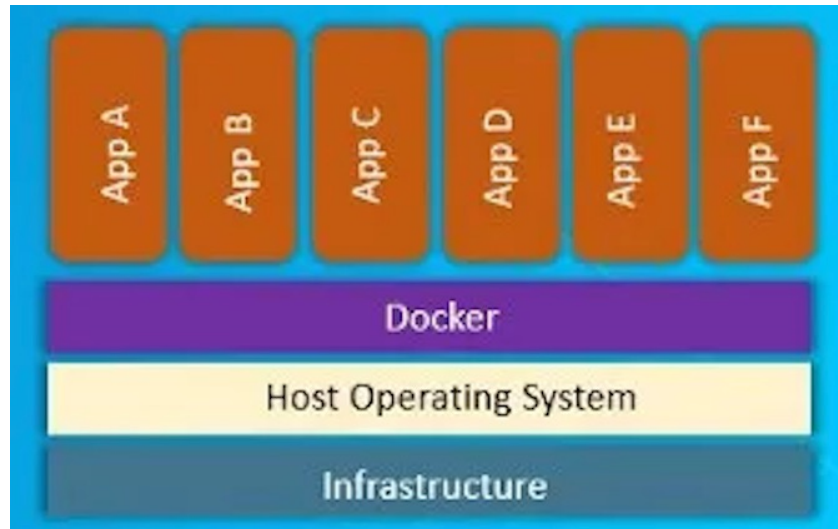


# Why should we use virtual machines?

- Even though by using virtual environments we isolate our computations, we might need to use the same operating system that must run regardless of the physical machine
- **Isolation!**
  - Each virtual machine is kept isolated from its host physical system and other virtualized machines.
  - If one virtual model crashes, the other virtual machines and the host system don't affect each other.
  - In addition, data isn't shared between one virtual model and another.
- **Full autonomy:** it works like a separate computer system, it is like run a computer within a computer.
- **Very secure:** the software inside the virtual machines can't affect the actual computer.
- **Lower costs:** buy one machine and run multiple operating systems.

# Container

# What is a container?



- **Standardized** packaging for software dependencies
- **Isolate** apps from each other
- **Works** for all major Linux distributions, MacOS, Windows

# Docker Containers are not Virtual Machines

Virtual Machines



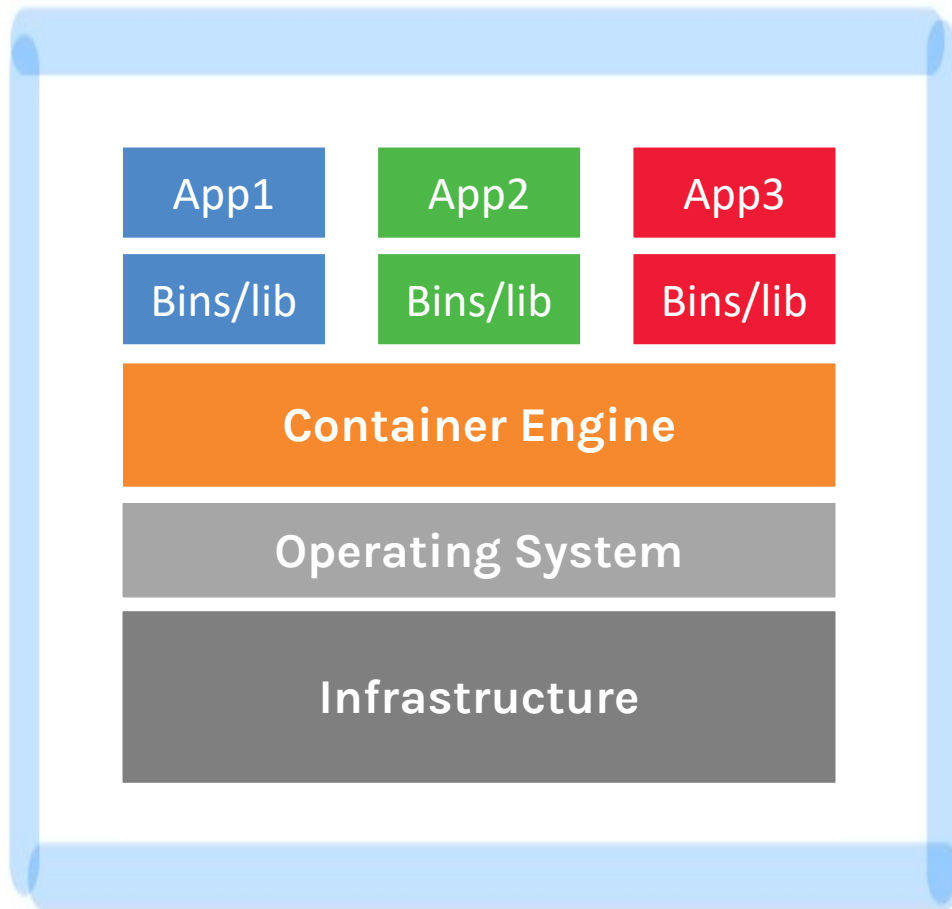
Containers



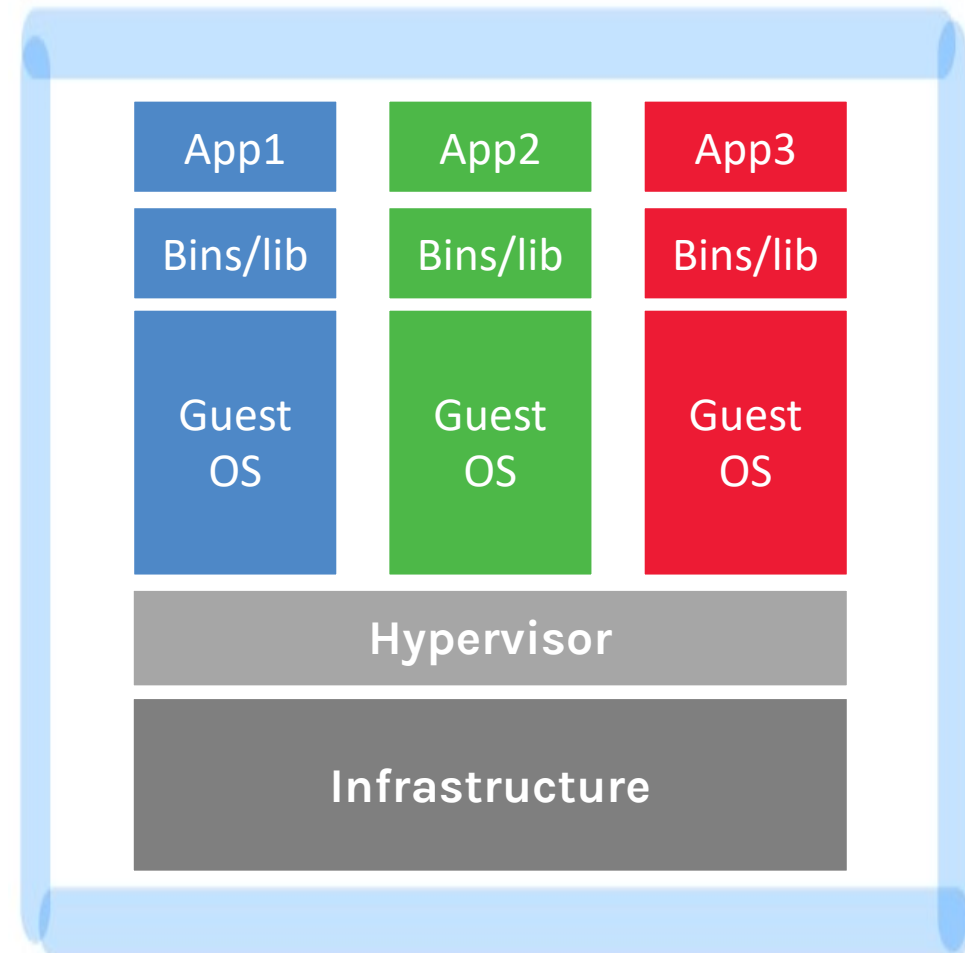


# Docker Container vs Virtual Machines

Container



Virtual Machine



# Docker Container vs Virtual Machines (VM)

## VMs

- Each VM runs its own OS
- Boot up time is in minutes
- Not version controlled
- Cannot run more than couple of VMs on an average laptop
- Only one VM can be started from one set of VMX and VMDK files

## Docker

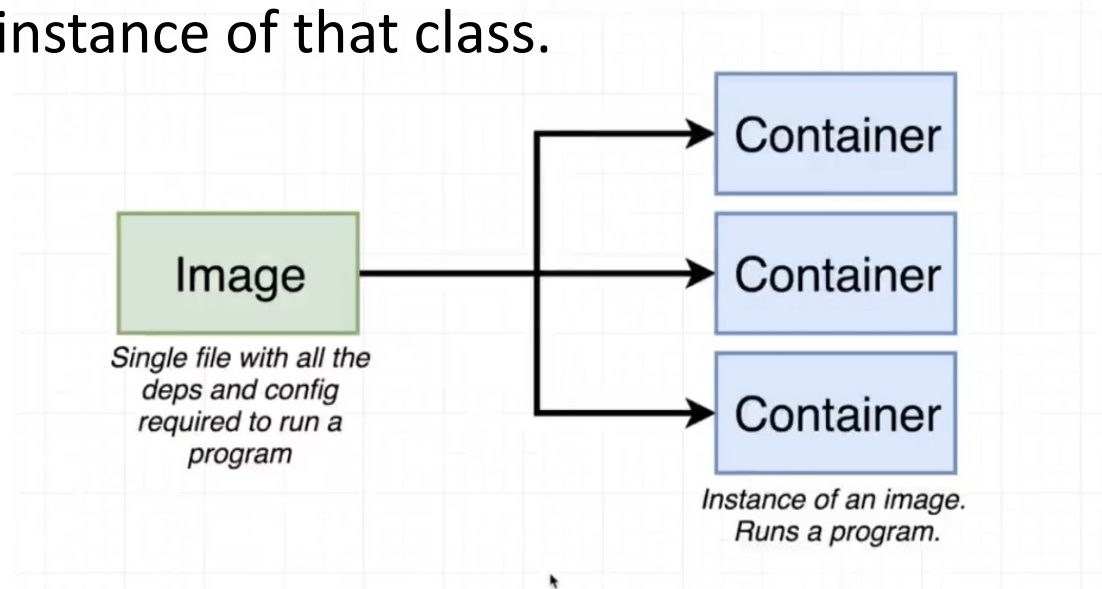
- Container is just a user space of OS
- Containers instantiate in seconds
- Images are built incrementally on top of another like layers.
- Images can be diffed and can be version controlled.
- Can run many Dockers in a laptop
- Multiple docker containers can be started from one Docker image

# History of Containers (just a few dates)

- 1979: Unix V7 (beginning of process isolation)
- 2000: FreeBSD Jails
- 2001: Linux VServer
- 2004: Solaris Containers
- 2008: LXC (Linux Containers)
- 2013: **Docker**
- 2018: The Gold Standard (containerization becomes the foundation for modern software infrastructure)
- 2020: Kubernetes Grows Up and **Docker Desktop** for Windows Home is released
- 2022: Record Adoption of Container Technologies

# Images and Containers

- Docker **Image** is a template aka blueprint to create a running Docker **container**.
- Docker uses the information available in the Image to create (run) a container.
- Possible interpretations
  - Image is like a recipe, container is like a dish
  - You can think of an image as a class (object-oriented programming) and a container is an instance of that class.



# How to build an image

- We use the Dockerfile, a simple text file, to build the Docker Image, which are iso files and other files.
- We run the Docker Image to get Docker Container.

```
FROM ubuntu:22.04

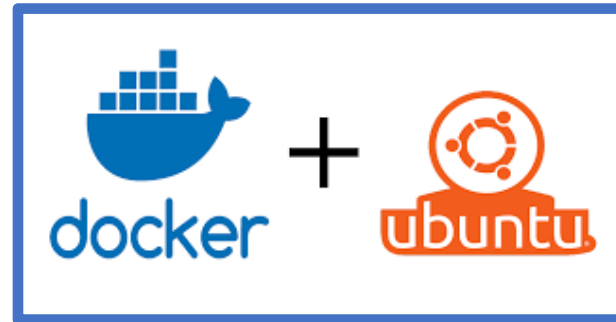
# install packages
RUN apt-get update

# install python and Jupyterlab
RUN apt-get install -y python3-pip

# expose a specific port
EXPOSE 8888

# Command to run when the container starts
CMD ["/bin/bash"]
```

Dockerfile



Docker image  
(Ubuntu for example)



Docker container

# Inside the Dockerfile

```
FROM alpine:latest
RUN apk update
RUN apk add nmap
ENTRYPOINT [``nmap``]
CMD [``localhost``]
```

Dockerfile

**FROM:** This instruction in the Dockerfile tells the daemon which base image to use while creating our new Docker image.

- In the example here, we are using a very minimal OS image called alpine (just 5 MB of size).

**RUN:** This command instructs the Docker daemon to run the given commands as it is while creating the image.

- A Dockerfile can have multiple RUN commands

**ENTRYPOINT:** The ENTRYPOINT instruction is used when you would like your container to run the same executable every time. Usually, ENTRYPOINT is used to specify the binary and CMD to provide parameters.

**CMD:** The CMD sets default command and/or parameters when a docker container runs.



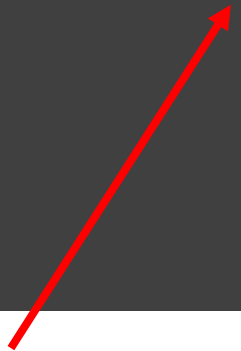
# Multiple containers from same image

- You could think of an image as instating a class.
- You could instate it with different parameters using the CMD and therefore different containers will be different.

```
FROM ubuntu:latest  
RUN apt-get update  
ENTRYPOINT ["/bin/echo", "Hello"]  
CMD ["world"]
```

Dockerfile

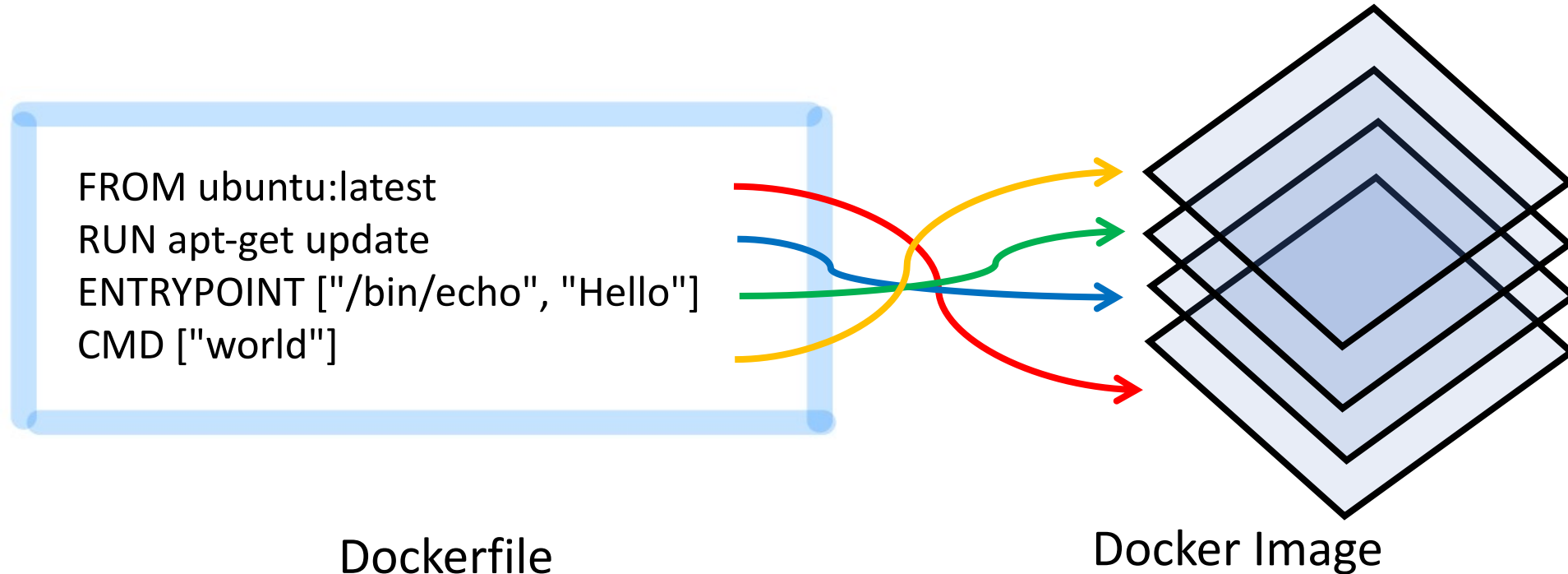
```
> docker build -t hello:first -f Dockerfile .  
  
> docker run -it hello:first  
> Hello world  
> docker run -it hello:first Bob  
> Hello Bob
```



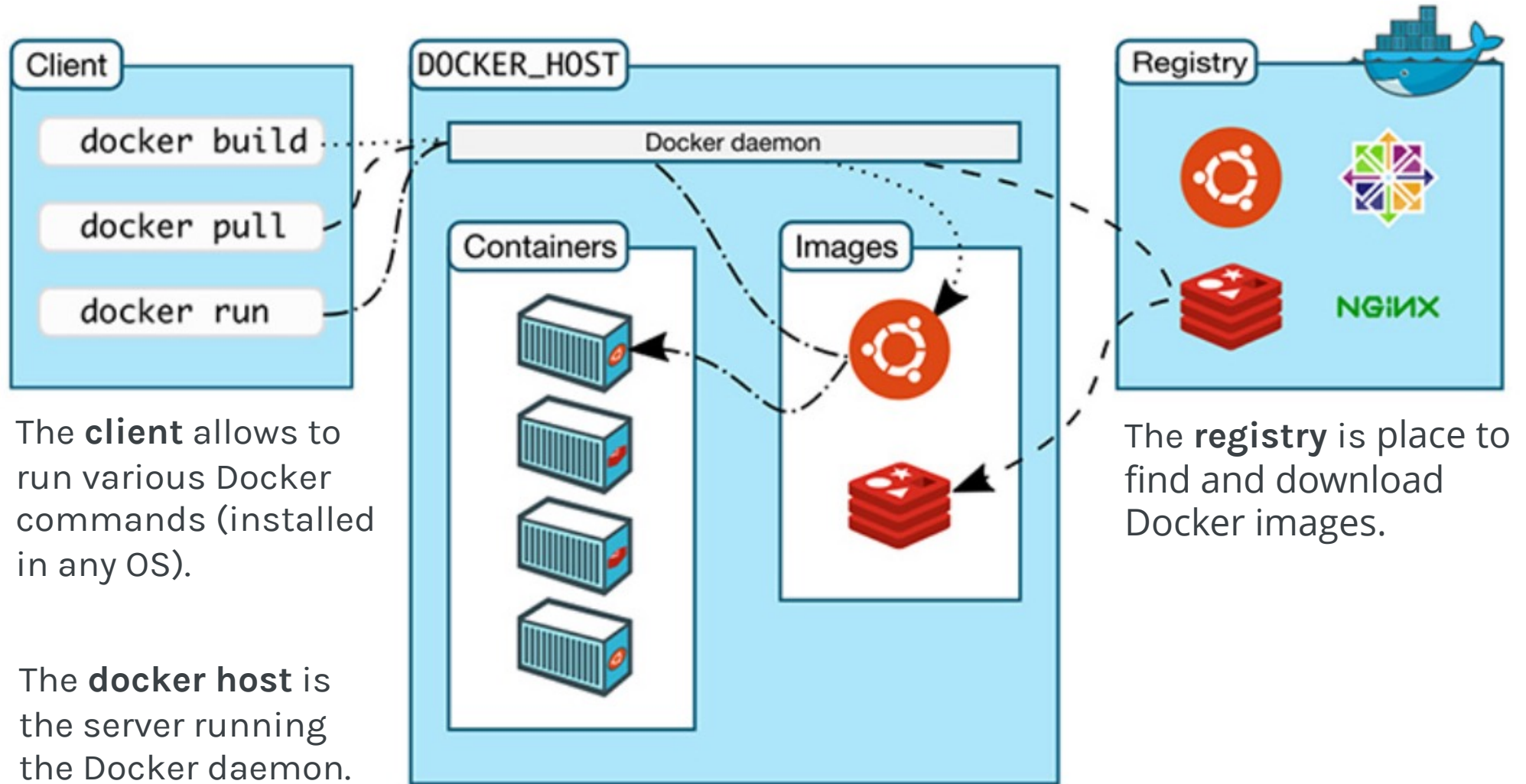
The period means the current working directory.

# Docker Image as Layers

- When we execute the build command, the daemon reads the Dockerfile and creates a layer for every command.
- The builder attempts to reuse layers from earlier builds.

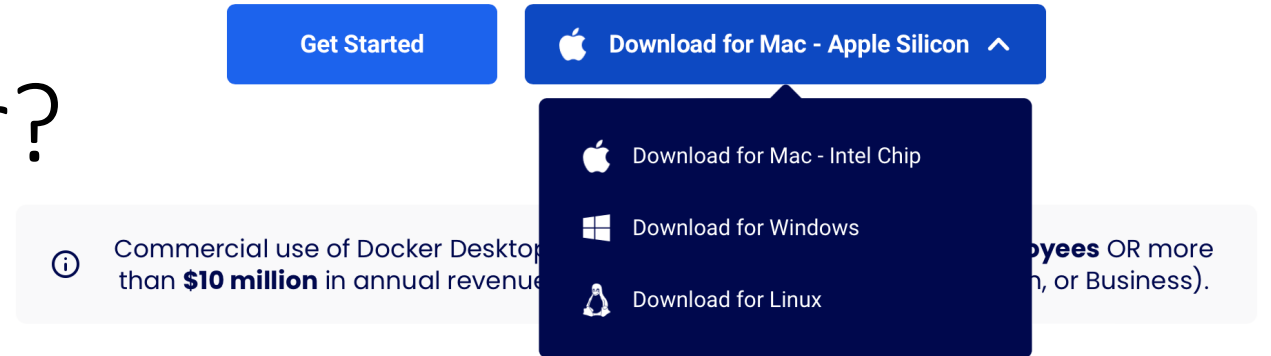


# The Docker Engine Architecture



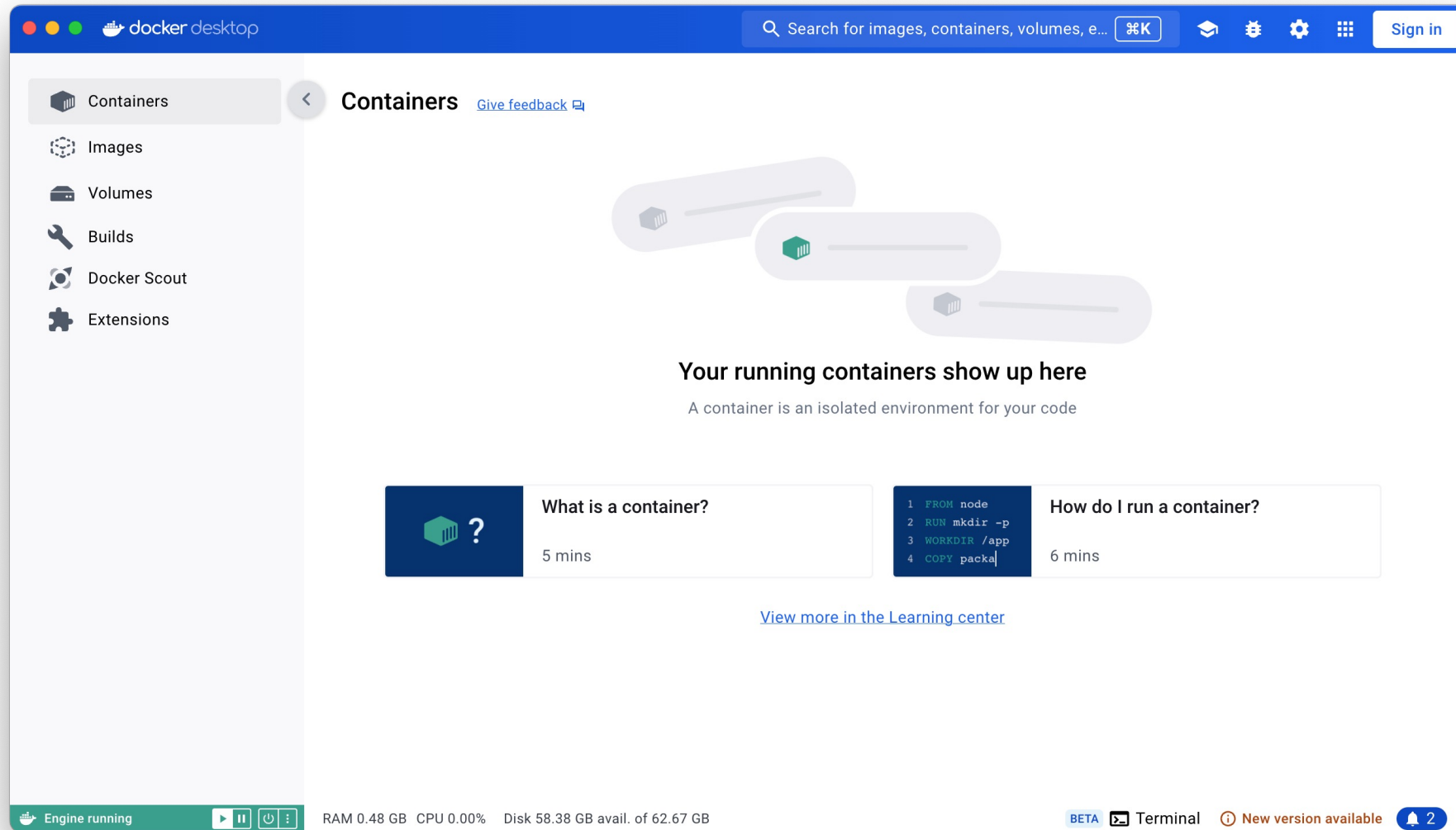
# Install Docker

# How to install Docker?



- Go to the webpage  
<https://www.docker.com/products/docker-desktop/>
- Download Docker Desktop ; choose the version appropriate for your OS (Mac, Windows, Linux)
- Install Docker Desktop
- Open Docker Desktop
  - Accept “Docker Subscription Service Agreement”
  - Use recommended settings (requires password)
- It is not required to sign in the application
- You can skip the survey
- You may have to restart your computer
- Start Docker Desktop

# Docker Desktop



# Create your first container

- Open a terminal
- Create a folder « BigData » where you wish

```
mkdir BigData
```
- Create a subfolder « Lab0 » and go inside it

```
cd BigData
mkdir Lab0
cd Lab0
```
- Edit and save the textfile « Dockerfile » in « Lab0 » with the content

```
FROM ubuntu:22.04
ENTRYPOINT ["/bin/sh", "-c", "echo 'It works!'"]
```
- Build your first Docker image by running the command

```
docker build -t linux-docker .
```
- Run your first container based on this image

```
docker run linux-docker
```
- In the terminal, you should read « It works! »



# Create your second container (for Lab 1)

- Open a terminal
- Create a subfolder « Lab1 » inside the folder « BigData » and go inside it

```
mkdir Lab1
```

```
cd Lab1
```

- Copy the textfile « Dockerfile » from the footnote url to « Lab1 »
- Build your first image

```
docker build -t lab1:v1 .
```

- Run your first contained based on this image

```
docker run -it -p 8888:8888 lab1:v1 /usr/bin/bash
```

- Open a browser and go to

```
http://127.0.0.1:8888/lab?token=datatoken
```

# Some useful Docker commands

- List all your images

```
docker images
```

- To view the list of running containers

```
docker ps
```

- Remove all unused containers, networks, images (both dangling and unused), and optionally, volumes

```
docker system prune
```

# Recommandation

- **First installation**

- We can help you (if possible) during the lab to finish/test your installation
- **BUT you must have downloaded Docker desktop** (because the download time might be very long in the lab room)

- **Before each lab**

- Before each lab, **build the image the day before** at home with high speed internet connexion and time!
- Some builds need more than 10 minutes, even 20 minutes!

# Conclusion

# Conclusion

- Big Data frameworks are composed of many tools.
- Labs on Big Data Technology will studied some of these tools.
- Tools will run inside containers.
- Containers can be deployed from a specific image on many platforms, such as workstations, virtual machines, public cloud, etc.
- Containers are extremely popular, and their popularity is growing.
- Multiple containers can be managed by a service called Kubernetes (not studied in this course).