

CS460: Intro to Database Systems

Database Systems and Beyond

Instructor: Manos Athanassoulis

<https://midas.bu.edu/classes/CS460/>

Database Systems

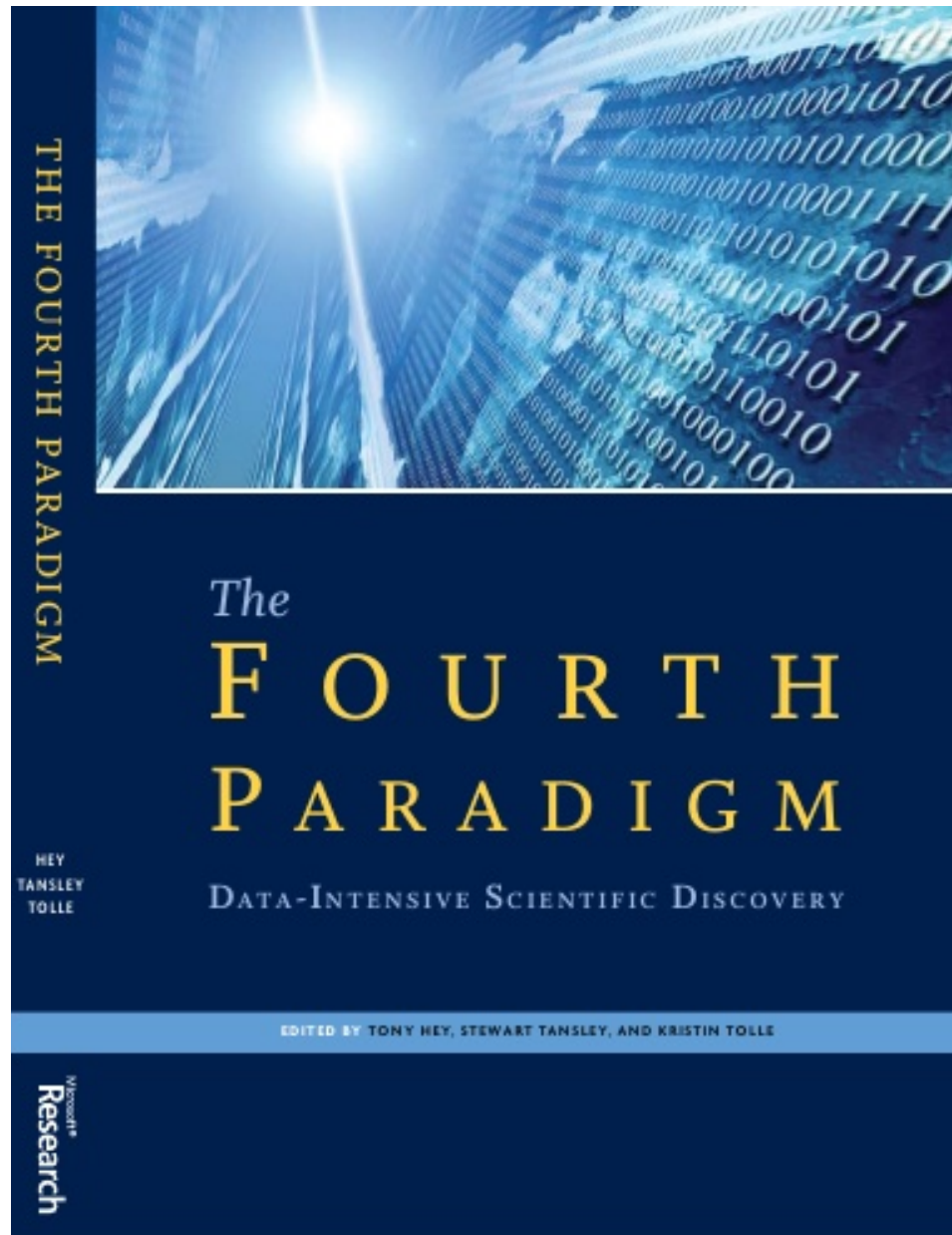
we spent a whole semester on Database Systems
what is next?

what can we do with data?

data-driven science

data-driven discovery

data-driven governance



*“Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, ‘**data-intensive**’ science paradigm is emerging.*

*The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. **Lots of new tools are needed to make this happen.**”*

Faster Innovation through Data-Intensive Approaches

Need for Innovation in Data Management!

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
Duolet CAZENA

STREAMING / IN-MEMORY
Amazon Kinesis
SAP Cloud Platform
ORACLE
confluent
strim
hazelcast
GridGain
GIGASPACE
Wallaroo
FASTDATA
lx

NoSQL DATABASES
Google Cloud AWS
ORACLE
MongoDB MarkLogic
Couchbase ORSTAX
redislabs
ArangoDB SCYLLA

NewSQL DATABASES
SAP Clustrix
Pivotal
MEMSQL
Cockroach Labs
VOLTDB
paradigm

GRAPH DBs
Amazon Neptune
ORACLE
InfinitaGraph
IBM
OrientDB
Objectivity

MPP DBs
TERADATA
IBM Data Warehouse Systems
Action
Kognitio
Exasol
dremio
Yellowbrick

CLOUD EDW
aws
Google Cloud
Microsoft Azure
Pivotal
snowflake
Infoworks

SERVERLESS
AWS Lambda
Google Cloud Functions
Microsoft Azure Functions
Pulsar
nuclio
Polar Function Service

DATA TRANSFORMATION
talend pentaho
alteryx TRIFACTA
tamr Paxata
StreamSets UNIFI

DATA INTEGRATION
SAP Data Services
MuleSoft
snoplogic
segment
ZALONI
Import.io
Infoworks
SNOWFLOW
MATILLION

DATA GOVERNANCE
Informatica
IBM
collibra
Alation
OKERA
MANTA
dataworld

MGMT / MONITORING
aws New Relic
rubrik
dynatrace
SignalFx
splunk
univention
zenoss
OpsRamp
MAGNITUDE

STORAGE
aws Google Cloud
Microsoft Azure
PURESTORAGE
ALLUXIO
Qumulo
COHERITY

CLUSTER SVCS
Amazon EKS
IBM
Google Cloud
Microsoft Azure
Google Cloud
Microsoft Azure
Google Cloud
Microsoft Azure

DATA GENERATION & LABELLING
amazon mechanical turk
upwork
oppen
hive
Mighty AI
LIONBRIDGE

AI OPS
ALGORITHMIA
VERTA.AI
datmo
datastion
Determined AI
tidder

GPU DBs & CLOUD
kinetica
SQream
bryllyt
PG-Stream
LOYDHUB

HARDWARE
Google TPU
ARM
Intel AI
NVIDIA
GRAPHICORE
MYTHIC
Movidius
Habana
WAVE
CERAMIC
PANGLOSS
DEFINIX

ANALYTICS & MACHINE INTELLIGENCE

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

DATA SCIENCE PLATFORMS
IBM databricks dataiku
DOMINO rapidminer TIBCO
ANACONDA SAS
KNIME MathWorks

BI PLATFORMS
looker
amazon analytics
DOME
ARCADIA DATA
ThoughtSpot
ATSCALE
Qlik
GoodData
Information Builders
birst
MicroStrategy
Keen IO

VISUALIZATION
tableau
Power BI
SAP
Google Cloud
celonis
Periscope Data
zepl
CHARTIO

MACHINE LEARNING
Amazon SageMaker
Google Cloud
DataRobot
gamalon
VISENZE
ELEMENT

COMPUTER VISION
Microsoft Azure
Amazon Rekognition
clarifai
EVER AI
neura
UBIQUITY
traq
synthesia

HORIZONTAL AI
IBM Watson
Cortana
sentient
Voyager
Affectiva
Numenta
petuum
naralogs
BLUE VISION

SPEECH & NLP
Google Cloud
Amazon Translate
narrative
science
Mobval
SoundHound Inc
Mindfield
SMARTLING
Unbabel

SEARCH
elasticsearch
algolia
Lucidworks
swiftype
alphasense
omni:us

LOG ANALYTICS
splunk
sumologic
solarwinds
TIMBER
lobano
logz.io

SOCIAL ANALYTICS
Hootsuite
sprinklr
NETBASE
synthesio
simplereach
bitly
SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics
mixpanel
Airtable
SIGOPT
granify
custora

APPLICATIONS - ENTERPRISE

SALES
CHORUS
INSIDESALES.COM
people.ai
conversica
clari
fuse machines

MARKETING - B2B
RADIUS
EVERSTRING
Lattice
MINTIGO
sense
tubular
NGAGIO
KNOTCH

MARKETING - B2C
Zeta
bloomreach
SendGrid
braze
ACTIONIQ
BLUECORE
CONTENT SQUARE
TEALUM
mparticle
Ampero
amperity
QUANTIFIND
Simon
Lytica
PERSADO
remesh

CUSTOMER EXPERIENCE / SERVICE
qualtrics
MEDALLIA
SurveyMonkey
Zendesk
Intercom
Drift
LIFEPEPPER
Gainsight
pendo
HEAP
Amplitude
Watson Assistant
Vialogflow
DigitalGenius
A.S.A.P.P
ada
AUTOMAT
ahni
CaDesk
reforms
frame.ai

ENTERPRISE PRODUCTIVITY
slack
ORACLE
GURU
lumiata
DIFFBOT
clara
talla
Kasisto

HUMAN CAPITAL
HireVue
pymetrics
hiQ
GOSTER
mya
Aillyo
textio
Wade & Wendy
Stella
entelo
uncommon
beomey

LEGAL
RAVEL
Everlaw
EDISCO
KIRO
JUDICATA
BREVIA
IRONCLAD
PREMPTION
ROSS
Casetext

REGTECH & COMPLIANCE
BigID
Tessian
text IQ
Comply Advantage

FINANCE
Anaplan
ZUORA
SAHANA
TRADESHIFT
mineral tree
SCALE FACTOR
SCALE FACTOR
butkeeper
pilot

BACK OFFICE AUTOMATION & RPA
UiPath
Blue Prism
VADO
AppZen
WorkFusion
workato
Remedy
Catalytic
ANTWORKS
KRYON
ALUKY

SECURITY
TANUIM
CYLANCE
zscaler
StackPath
illumio
CODE42
CIPHERCloud
DARKTRACE
ANOMALI
ThreatMetrix
VECTRA
pindrop
exabeam
SIGNIFY
SentinelOne
SecurityScorecard
SECURE
CodeSecure
bitglass
BlueTalon
Recorded Future
feedzai
Cyber
BITSIGHT
spartan
sparkcognition
CyberArk
FORTEK
riskrecon
JASK
ARMORBLADE

APPLICATIONS - INDUSTRY

ADVERTISING
AppNexus
criteo
ORACLE
MOAT
theTradeDesk
dstillery
LiveIntent
TAPAD
datax
gumgum
Cupier

EDUCATION
Lullishoo
Knewton
Clever
edX
KIDAPPE
PANORAMA
knowre
gradescope

REAL ESTATE
REDFIN
Opndoor
VTS
CREDFI
GEOPHY
reonomy
COMPSTAK
SPACEMAKER
SKYLINE

GOV'T
OPENGOV
mark43
FiscalNote
LiveStories
Passport
SmartProcure
STREETLIGHTDATA
OpenDataSoft

INTELLIGENCE
Palantir
Dataminr
Quid
PRIMER
FORGE

FINANCE - INVESTING
KENSHC
Quantopian
ADDEPAR
NUMERA
SENTIUM
ALGORIT
HavenPack
PAGAYA

FINANCE - LENDING
ondeck
JIANPU.AI
Kreditech
AVANT
TALA
CLEARBANC
Upgrade
100Credit
WeLab
WeCASH
TrueAccord
MoneyLion
Active AI
aire
cignifi

INSURANCE
Metromile
Lemonade
CYENCE
Hippo
Shift Technology
ROOT
Zesty AI
TRAVELER
CAPE

HEALTHCARE
flatiron
Clover
KYRUS
HealthTap
METABIOTA
Gingerio
Glow
babylon
3D Med
zebra
PathAI
ovia
TEMPUS
patientslikeme
AiCure
insitro
LIMAGO
citizen
notable
humandot
RECURSION
prognos
enlitic
Imag
BAYLABS
Qventus
ARTERS
IMAGEN
PAIGE
DATAYANI
innovaccer
Leonardo

LIFE SCIENCES
Xenoma
color
BenevolentAI
verily
WuXiNextCODE
SIBIRIN
Clear Labs
fraxnoma
MANOPHORE
DNVexus
Phosphor
CITRINE
twoAR
deep genomics
QWIK
QWIK

TRANSPORTATION
UBER
TESLA
CLEARPATH
CRUISE
nuvo
Aurora
drive.ai
CAMBRIDGE
nauto
AMOTIVE
G7
PILOT.AI
NIO
OPTIMUS
moovit
Ike
nexar
Kodiak
comma.ai
netradyn
Civil Maps
cognata
thind
INRIX

AGRICULTURE
FARMERS
Granular
JOHN DEERE
BLUE RIVER
FarmersEdge
AgroStar
FarmLogs
TARANIS
GAMAYA
terrotron
prospera

COMMERCE
instacart
FAIRFIRE
STITCH FIX
Doo & Co
heuristic
Other

INDUSTRIAL
AVEVA
SIEMENS
PREDIX
UPTAKE
SCORTEX
TACHYUS
ByteDance
Stem
Amper
SOJERN
BBOXEVER
Verdigris
duetto
Jukebox
Electric
ZINER
Spoke

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS IOI DATA vmware TIBCO TERADATA ORACLE NetApp syncsort MAPR cloudera

OPEN SOURCE

FRAMEWORKS
Spark
Flink
YARN
TEZ
MESOS
docker
CDAP
HELD

QUERY / DATA FLOW
Spark SQL
presto
SLAM DATA
GraphQL
Flink

DATA ACCESS & DATABASES
cassandra
mongoDB
redis
Cockroach Labs
druid
CouchDB
SciDB
riak
HBASE
Cloud Spanner
ACCUTRA

ORCHESTRATION & MGMT
talend
Apache Airflow
Apache Ambari
Apache Zookeeper
etcd
Kong

STREAMING & MESSAGING
Spark
nifi
Flink
beam
kafka
STORM
Apache RocketMQ

STAT TOOLS & LANGUAGES
python
Scala
R
Studio
SciPy
julia

AI OPS & INFRA
miflow
Kubeflow
leap
SELOON
Polyaxon

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow
Keras
PyTorch
OpenAI
DM TK
theano
mxnet
VELES
Chainer
PyTorch
neon
DSSTNE
mlib
DL4J
MAHOUT
Aerosolve
FastAI
mir

SEARCH
elasticsearch
Solr

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: Apple, VALIDIC, practicefusion
IOT: GE Digital, Uptake
FINANCIAL & ECONOMIC DATA: Bloomberg, Refinitiv, S&P Global
AIR / SPACE / SEA: Inmarsat, Iridium, SpaceX
PEOPLE / ENTITIES: LinkedIn, Facebook, Twitter
LOCATION INTELLIGENCE: Google Maps, HERE, TomTom
OTHER: Various niche data providers

DATA RESOURCES

DATA SERVICES: DataCamp, Coursera, edX
INCUBATORS & SCHOOLS: Y Combinator, Techstars, 500 Startups
RESEARCH: OpenAI, Facebook Research, MIRI, Galvanize

see full version at: http://mattturck.com/wp-content/uploads/2019/07/2019_Matt_Turck_Big_Data_Landscape_Final_Fullsize.png

DATA & AI LANDSCAPE 2019



increase throughput by parallelization

“scale-up”

use more powerful machines (>#CPUs, >RAM)

“scale-out”

use more machines

Scale Up Execution

how to use more cores (threads)?

inter-query parallelism

each query runs on one processor

inter-operator parallelism

each query runs on multiple processors
an operator runs on one processor

intra-operator parallelism

An operator runs on multiple processors

Scale Up Storage

needs more disks!

how to distribute data?

block partition

hash partition

range partition

how to distribute data accesses?

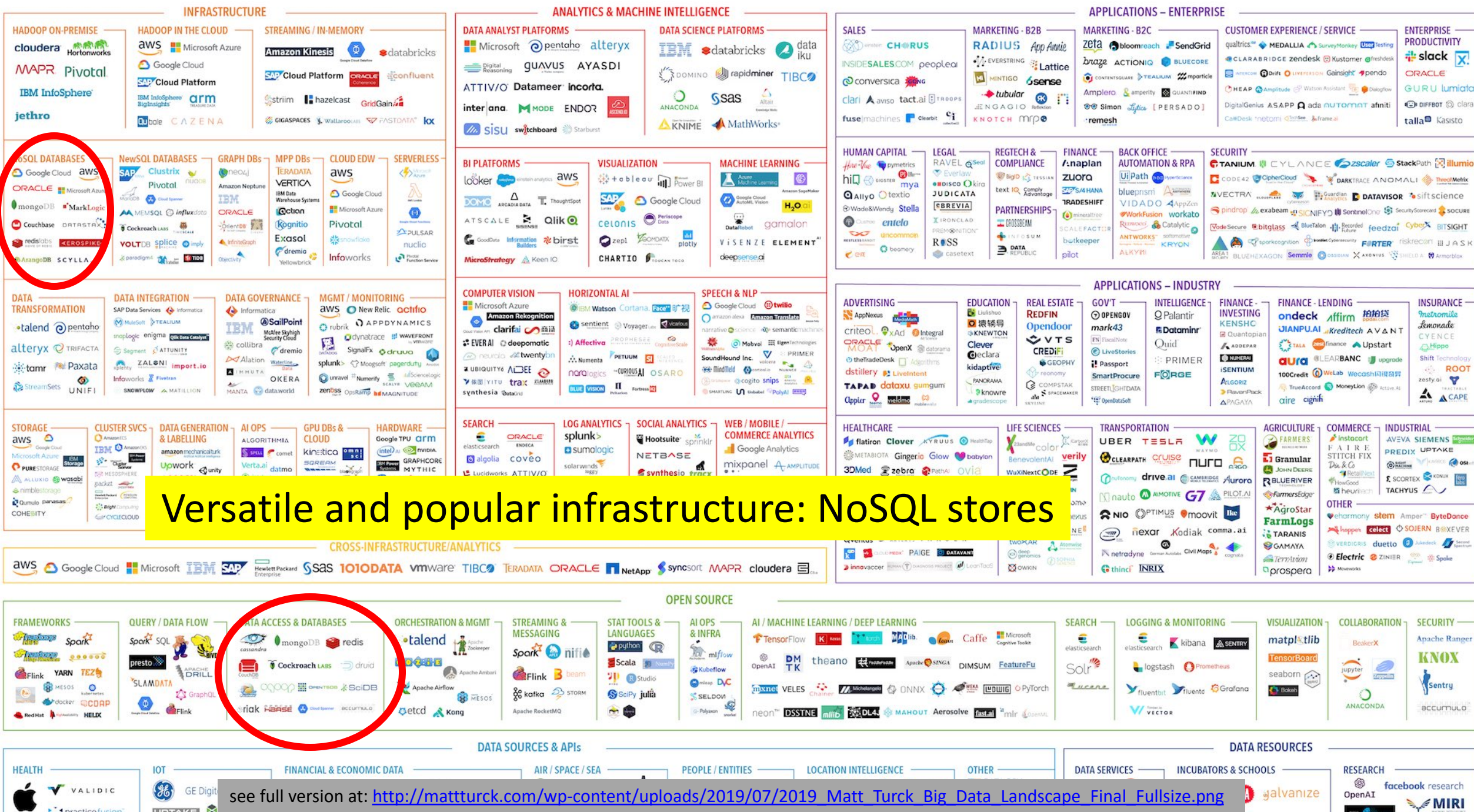
Scale Out

similar questions across machines

new bottlenecks?

move data across machines: network!

DATA & AI LANDSCAPE 2019



see full version at: http://mattturck.com/wp-content/uploads/2019/07/2019_Matt_Turck_Big_Data_Landscape_Final_Fullsize.png

diving into the internals of modern data systems

cutting-edge designs / *research* projects / *engineering* projects

CS 591: Data Systems Architectures

Spring 2020

Today: more discussion on
simplified and efficient storage models
“NoSQL stores”

What is NoSQL?

from "Geek and Poke"

HOW TO WRITE A CV



Leverage the NoSQL boom

What is NoSQL?

An emerging “movement” around non-relational software for Big Data

Roots are in the Google and Amazon homegrown software stacks

Wikipedia: “A NoSQL database provides a mechanism for storage and retrieval of data that use looser consistency models than traditional [relational databases](#) in order to achieve [horizontal scaling](#) and higher availability. Some authors refer to them as "Not only SQL" to emphasize that some NoSQL systems do allow [SQL](#)-like query language to be used.”

NoSQL Stores

offer an easy to program storage model

simplification of relational

two attributes (a key and a value)

value has variable size

NoSQL features

Scalability is crucial!

- load increased rapidly for many applications

Large servers are expensive

Solution: use clusters of small commodity machines

- need to partition the data and use replication (sharding)
- cheap (usually open source!)
- cloud-based storage

NoSQL features

Sometimes not a well defined schema

Allow for semi-structured data

- still need to provide ways to query efficiently (use of index methods)
- need to express specific types of queries easily

Scalability

Often cited as the main reason for moving from DB technology to NoSQL

DB Position: there is no reason a parallel DBMS cannot scale to 1000's of nodes

NoSQL Position: a) Prove it; b) it will cost too much anyway

Flavors of NoSQL

Four main types:

- key-value stores
- document databases
- column-family (aka big-table) stores
- graph databases

Here we will talk more about “Document” databases (MongoDB)

Key-Value Stores

There are many systems like that:

Redis, MemcachedDB, Amazon's DynamoDB, Voldemort

Simple data model: key/value pairs

the DBMS does not attempt to interpret the value

Queries are limited to query by key

- get/put/update/delete a key/value pair
- iterate over key/value pairs

Document Databases

Examples include:

MongoDB, CouchDB, Terrastore

Special type of key/value that value is a document.

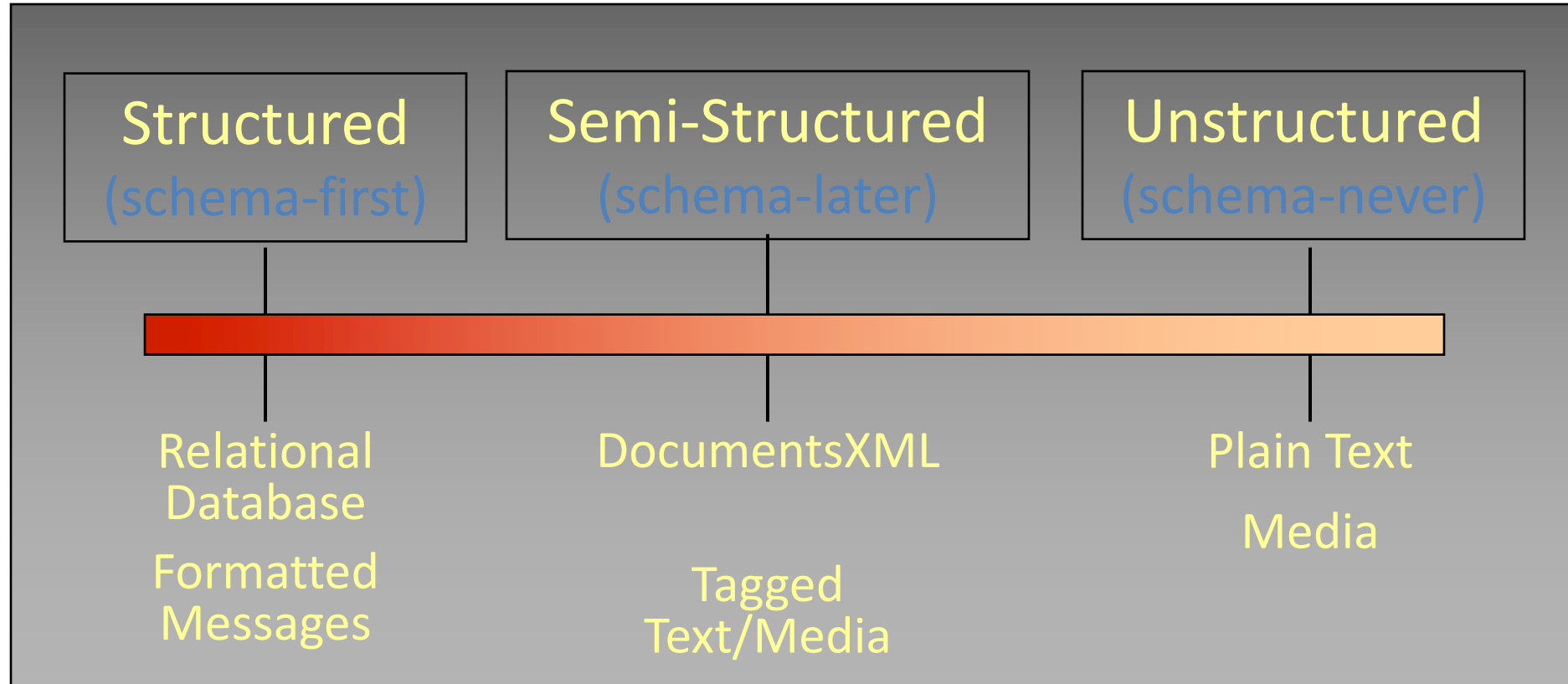
- use some sort of semi-structured data model: XML/JSON
- the value can be examined and used by the system (unlike in key/data stores)

Queries based on key (as in key/value stores), but also on the document (value).

Here again, there is support for sharding and replication.

- the sharding can be based on values within the document

The Structure Spectrum



MongoDB (An example of a Document Database)

Data are organized in collections. A collection stores a set of documents.

Collection (like table) and document (like record)

- but: each document can have a different set of attributes even in the same collection
- Semi-structured schema!

Only requirement: every document should have an “_id” field

- hum**ong**ous => Mongo

Example mongodb

```
{  "_id": ObjectId("4efa8d2b7d284dad101e4bc9"),  
    "Last Name": " Cousteau",  
    "First Name": " Jacques-Yves",  
    "Date of Birth": "06-1-1910" },  
  
{  "_id": ObjectId("4efa8d2b7d284dad101e4bc7"),  
    "Last Name": "PELLERIN",  
    "First Name": "Franck",  
    "Date of Birth": "09-19-1983",  
    "Address": "1 chemin des Loges",  
    "City": "VERSAILLES" }
```

Example Document Database: MongoDB

Key features include:

- JSON-style documents

 - actually uses BSON (JSON's binary format)

- replication for high availability

- auto-sharding for scalability

- document-based queries

- can create an index on any attribute for faster reads

under the hood, a simple key-value store called WiredTiger!
design based on LSM-trees

MongoDB Terminology

relational term \Leftrightarrow MongoDB equivalent

database \Leftrightarrow database

table \Leftrightarrow collection

row \Leftrightarrow document

attributes \Leftrightarrow fields (field-name:value pairs)

primary key \Leftrightarrow the `_id` field, which is the key associated with the document

JSON

JSON is an alternative data model for semi-structured data

- JavaScript Object Notation

Built on two key structures:

- an object, which is a sequence of name/value pairs
`{ "_id": "1000", "name": "Sanders Theatre", "capacity": 1000 }`
- an array of values `["123", "222", "333"]`

A value can be:

- an atomic value: string, number, true, false, null
- an object
- an array

The `_id` Field

Every MongoDB document must have an `_id` field.

- its value must be unique within the collection

- acts as the primary key of the collection

- it is the key in the key/value pair

If you create a document without an `_id` field:

- MongoDB adds the field for you

- assigns it a unique BSON (binary JSON) ObjectId

- example from the MongoDB shell:

```
> db.test.save({ rating: "PG-13" })
```

```
> db.test.find() { "_id" : ObjectId("528bf38ce6d3df97b49a0569"), "rating" : "PG-13" }
```

Note: quoting field names is optional (see rating above)

Capturing Relationships in MongoDB

Two options:

1. store references to other documents using their `_id` values
2. embed documents within other documents

Example relationships

```
{
  "_id": ObjectId("52ffc33cd85242f436000001"),
  "name": "Tom Hanks",
  "contact": "987654321",
  "dob": "01-01-1991"
}
```

Here is an example of embedded relationship:

```
{
  "_id": ObjectId("52ffc33cd85242f436000001"),
  "contact": "987654321",
  "dob": "01-01-1991",
  "name": "Tom Benzamin",
  "address": [
    {
      "building": "22 A, Indiana Apt",
      "pincode": 123456,
      "city": "Los Angeles",
      "state": "California"
    },
    {
      "building": "170 A, Acropolis Apt",
      "pincode": 456789,
      "city": "Chicago",
      "state": "Illinois"
    }
  ]
}
```

```
{
  "_id": ObjectId("52ffc4a5d85242602e000000"),
  "building": "22 A, Indiana Apt",
  "pincode": 123456,
  "city": "Los Angeles",
  "state": "California"
}
```

And here an example of reference based

```
{
  "_id": ObjectId("52ffc33cd85242f436000001"),
  "contact": "987654321",
  "dob": "01-01-1991",
  "name": "Tom Benzamin",
  "address_ids": [
    ObjectId("52ffc4a5d85242602e000000"),
    ObjectId("52ffc4a5d85242602e000001")
  ]
}
```

Queries in MongoDB

Each query can only access a single collection of documents.

Use a method called

```
> db.collection.find(<selection>, <projection>)
```

Example: find the names of all R-rated movies:

```
> db.movies.find({ rating: 'R' }, { name: 1 })
```


Projection

Specify the name of the fields that you want in the output with 1 (0 hides the value)

Example:

```
> db.movies.find({}, {"title":1, _id:0})
```

(will report the title but not the id)

Selection

You can specify the condition on the corresponding attributes using the find:

```
> db.movies.find({ rating: "R", year: 2000 }, { name: 1, runtime: 1 })
```

Operators for other types of comparisons:

MongoDB	SQL equivalent
\$gt, \$gte	>, >=
\$lt, \$lte	<, <=
\$ne	!=

Example: find the names of movies with an earnings <= 200000

```
> db.movies.find({ earnings: { $lte: 200000 } })
```

For logical operators \$and, \$or, \$nor

use an array of conditions and apply the logical operator among the array conditions:

```
> db.movies.find({ $or: [ { rating: "R" }, { rating: "PG-13" } ] })
```

Aggregation

Recall the aggregate operators in SQL: AVG(), SUM(), etc.

More generally, aggregation involves computing a result from a collection of data.

MongoDB supports several approaches to aggregation:

- single-purpose aggregation methods
- an aggregation pipeline
- map-reduce

Aggregation pipelines are more flexible and useful (see next):

<https://docs.mongodb.com/manual/core/aggregation-pipeline/>

Simple Aggregations

db.collection.count(<selection>)

returns the number of documents in the collection
that satisfy the specified selection document

Example: how many R-rated movies are shorter than 90 minutes?

```
> db.movies.count({ rating: "R", runtime: { $lt: 90 } })
```

db.collection.distinct(<field>, <selection>)

returns an array with the distinct values of the specified field
in documents that satisfy the specified selection document
if omit the query, get all distinct values of that field

Example: which actors have been in one or more of the top 10 grossing movies?

```
> db.movies.distinct("actors.name", { earnings_rank: { $lte: 10 } })
```

Aggregation Pipeline

A very powerful approach to write queries in MongoDB is to use pipelines.

We execute the query in stages.

Every stage gets as input some documents, applies filters/aggregations/projections and outputs some new documents.

These documents are the input to the next stage (next operator) and so on

Similar to a traditional query plan. But always with one child (no joins!)

Aggregation Pipeline example

Example for the zipcodes database:

```
> db.zipcodes.aggregate( [
  { $group: { _id: "$state", totalPop: { $sum: "$pop" } } },
  { $match: { totalPop: { $gte: 10*1000*1000 } } }
] )
```

```
{
  "_id": "10280",
  "city": "NEW YORK",
  "state": "NY",
  "pop": 5574,
  "loc": [
    -74.016323,
    40.710537
  ]
}
```

Here we use `group_by` to group documents per state, compute sum of population and output documents with `_id`, `totalPop` (`_id` has the name of the state). The next stage finds a match for all states the have more than 10M population and outputs the state and total population.

More here: <https://docs.mongodb.com/manual/tutorial/aggregation-zip-code-data-set/>

continued:

In SQL:

Output example:

```
{  
  "_id" : "AK",  
  "totalPop" : 550043  
}
```

```
SELECT state, SUM(pop) AS totalPop  
FROM zipcodes  
GROUP BY state  
HAVING totalPop >= (10*1000*1000)
```


```
db.zipcodes.aggregate( [  
  { $group: { _id: "$state", totalPop: { $sum: "$pop" } } },  
  { $match: { totalPop: { $gte: 10*1000*1000 } } }  
] )
```

more examples:

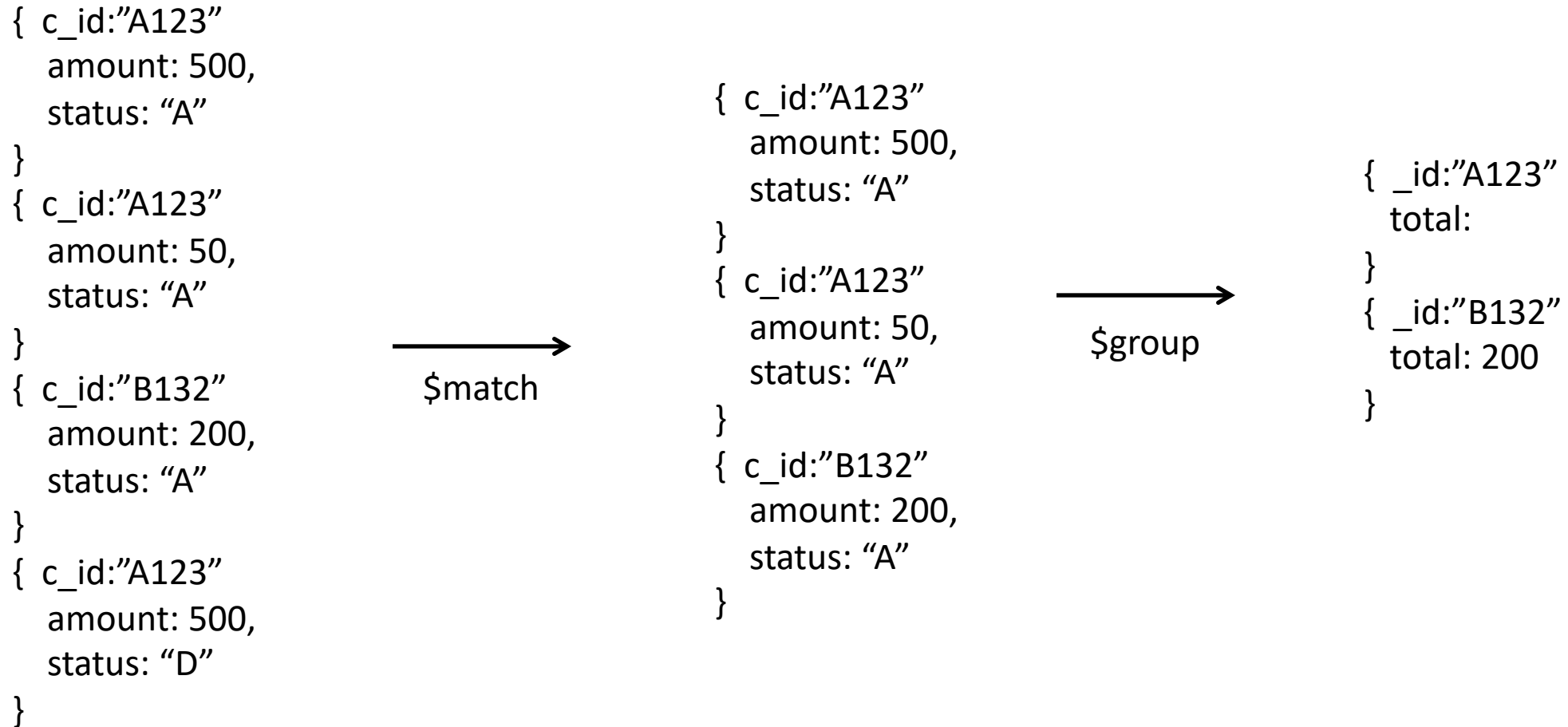
```
db.zipcodes.aggregate( [
  { $group: { _id: { state: "$state", city: "$city" }, pop: { $sum: "$pop" } } },
  { $group: { _id: "$_id.state", avgCityPop: { $avg: "$pop" } } }
] )
```

What we compute here?

First we get groups by city and state and for each group we compute the population.
Then we get groups by state and compute the average city population

<pre>{ "_id" : { "state" : "CO", "city" : "EDGEWATER" }, "pop" : 13154 }</pre>		<pre>{ "_id" : "MN", "avgCityPop" : 5335 }</pre>
--	---	--

Aggregation Pipeline example



```
db.orders.aggregate([ { $match: {status: "A"}}
                      { $group: {_id:"c_id", total: {$sum: $amount}}
                      ])
```

Other Structure Issues

NoSQL

- a) Tables are unnatural
- b) “joins” are evil
- c) need to be able to “grep” my data

DB

- a) Tables are a natural/neutral structure
- b) data independence lets you precompute joins under the covers
- c) this is a price of all the DBMS goodness you get

This is an Old Debate – Object-oriented databases, XML DBs, Hierarchical, ...

Fault Tolerance

DBs: coarse-grained FT – if trouble, restart transaction

- Fewer, Better nodes, so failures are rare
- Transactions allow you to kill a job and easily restart it

NoSQL: Massive amounts of cheap HW, failures are the norm and massive data means long running jobs

- So must be able to do mini-recoveries
- This causes some overhead (file writes)

DATA & AI LANDSCAPE 2019

INFRASTRUCTURE

The diagram is divided into three main sections by vertical lines:

- HADOOP ON-PREMISE:** Includes logos for Cloudera, Hortonworks, MapR, Pivotal, IBM InfoSphere, and jethro.
- HADOOP IN THE CLOUD:** Includes logos for AWS, Microsoft Azure, Google Cloud, SAP Cloud Platform, IBM InfoSphere BigInsights, ARM Telemetry Stack, Databricks, and CAZENA.
- STREAMING / IN-MEMORY:** Includes logos for Amazon Kinesis, Google Cloud Platform, Databricks, SAP Cloud Platform, Oracle Coherence, Confluent, Striim, Hazelcast, GridGain, GIGASPACE, Wallaroo, FASTDATA, and KX.

ANALYTICS & MACHINE INTELLIGENCE

The diagram is divided into two main sections by a horizontal line. The left section is titled 'DATA ANALYST PLATFORMS' and the right section is titled 'DATA SCIENCE PLATFORMS'. Both sections contain logos of various companies.

DATA ANALYST PLATFORMS:

- Microsoft
- pentaho
- alteryx
- Digital Reasoning
- QUAVUS
- AYASDI
- ATTIVIO
- Datameer
- incorta.
- interana
- MODE
- ENDOR
- ASPENDO

DATA SCIENCE PLATFORMS:

- IBM
- databricks
- dataiku
- DOMINO
- rapidminer
- TIBCO
- ANACONDA
- SAS
- Altaira
- KNIME
- MathWorks

APPLICATIONS – ENTERPRISE

Category	Count
SALES	10
MARKETING - B2B	10
MARKETING - B2C	10
CUSTOMER EXPERIENCE / SERVICE	10
ENTERPRISE PRODUCTIVITY	10
Other	10

The collage is organized into three distinct sections, each with a header in red capital letters:

- BI PLATFORMS:** This section includes logos for Looker, Amazon, Microsoft, and SAP Analytics Cloud (SAP), followed by AWS, Domo, Arcadia Data, ThoughtSpot, AT&T, Qlik, SAS, GoodData, Information Builders, and Birst.
- VISUALIZATION:** This section features logos for Tableau, Microsoft Power BI, SAP Lenses, Google Cloud, Celonis, Periscope Data, Zepi, Xoriant, and ChARTIO.
- MACHINE LEARNING:** This section displays logos for Azure Machine Learning, Amazon SageMaker, Google Cloud AI Platform, H2O, DataRobot, Gamalon, ViSenze, Element, and DeepSense.ai.

At the bottom of the collage, there are logos for MicroStrategy, Keen IO, and PowerToGo.

HUMAN CAPITAL    
LEGAL    
REGTECH & COMPLIANCE    
FINANCE    
BACK OFFICE AUTOMATION & RPA   
SECURITY

JUDICATICA    
PARTNERSHIPS    
TRADESHIFF       

The collage displays logos for various companies in the data and cloud space, organized into four categories:

- DATA TRANSFORMATION:** Includes logos for Talend, Pentaho, Alteryx, Trifacta, TMR, Paxata, StreamSets, and UNIFI.
- DATA INTEGRATION:** Includes logos for SAP Data Services, Informatica, MuleSoft, Tealium, Snaplogic, Enigma, Data Connect, Segment, Attunity, Alation, Zeleni, Import.io, InfoWorks, Fivetran, Snowplow, and Matillion.
- DATA GOVERNANCE:** Includes logos for Informatica, SailPoint, IBM, McAfee Skyhigh Security Cloud, Collibra, Dremio, Alation, Waterline, IBM, OKERA, MANTA, and data.world.
- MGMT / MONITORING:** Includes logos for AWS, New Relic, Actifio, Rubrik, AppDynamics, Dynatrace, Wavefront, SignalFx, Druva, Splunk, Moogsoft, PagerDuty, Unwired, Numerify, Scalyst, Veeam, Zenoss, OpsRamp, and MAGNITUDE.

COMPUTER VISION

- Microsoft Azure
- Amazon Rekognition
- Cloud Vision API
- Clarifai
- Ever AI
- Deepomatic
- Neura
- Twinty
- Ubiquity
- Adee
- Synthesia
- Yitu
- Trax
- OpenGrid
- Cameras

HORIZONTAL AI

- IBM Watson
- Cortana
- Face++
- Sentient
- Voyager
- Prophesize
- Affectiva
- Numenta
- Naturallogics
- Blue Vision
- Petuum
- Forbes
- Prophesize
- Cognitive Scale
- Curious AI
- OSARO

SPEECH & NLP

- Google Cloud
- Amazon Alexa
- Amazon Translate
- Twilio
- Semantic Machine
- Moovio
- Eigen Technologies
- SoundHound Inc.
- PRIMO
- Midfield
- Volterra
- Cognito
- Smiling
- UbiLog

APPLICATIONS - INDUSTRY

Industry	Startups
ADVERTISING	AppNexus, xAd, Integral Ad System, OpenX, DoubleClick, Adform, Adgains, Intent, GumGum
EDUCATION	Ludishuo, 猿辅导, KNEWTON, Clever, eCleria, kidaptive, PANORAMA, Knowre
REAL ESTATE	REDFIN, Opendoor, VTS, CREDIFI, GEOPHY, reonomy, COMPUSTAT, SPACEMAKER
GOV'T	OPENGOV, mark43, FSI, FocusNote, LiveStories, Passport, SmartProcure, STREETLIGHTDATA, iRIS, iRIS2
INTELLIGENCE	Palantir, Dataminr, Quid, PRIMER, FORGE
FINANCE - INVESTING	KENSHC, Quantopian, ADDEPAR, NUMERAM, ISENTIUM, ALGORIT, RavenPack
FINANCE - LENDING	ondeck, Affirm, 拍拍贷, JIANPU.AI, Kreditech, AVANT, TALAI, ONE finance, Upstart, aURA, LEARBANC, upgrade, 100Credit, WeLab, Wechat微众银行, TrueAccord, MoneyLion, ActiveAI
INSURANCE	metromile, Lemonade, CYENCE, Hippo, Shift Technology, ROOT, zesty.ai, A CAPA

The collage is organized into six main categories, each with a header and a collection of logos:

- STORAGE**: Includes logos for AWS, Google Cloud, Microsoft Azure, IBM Storage, Pure Storage, Alluxio, Wasabi, Nimbustorage, Datto, Paraverse, and Cohesity.
- CLUSTER SVCS**: Includes logos for Amazon EC2, Amazon EKS, Databricks, Google Kubernetes Engine, Mesosphere, Docker, Kubernetes, and AWS CloudCloud.
- DATA GENERATION & LABELLING**: Includes logos for Amazon Mechanical Turk, Upwork, Unity, Scale, Hive, Labelbox, Alereverse, and Lionbridge.
- AI OPS**: Includes logos for Algorithmia, SAP, Comet, Verta.ai, Datto, Datastrato, Weight & Dimensions, and Fiddler.
- GPU DBs & Cloud**: Includes logos for Kinetica, SQRIRM, Blazeblock, Brylryt, Blazegob, PG-Stream, FLOYDLUB, and Pondero.
- HARDWARE**: Includes logos for Google TPU, ARM, Intel, NVIDIA, GRAPHCORE, MYTHIC, IBM Power, Xilinx, Cerebras, Habana, Movidius, Vayve, CERNIAI, and DEFENIX.

The collage displays logos for the following companies:

- SEARCH:** Elasticsearch, Algolia, Lucidworks, ATTIVO, Swifttype, EXALAND, Alphahense, MAANA, omni:us, SINEOUA.
- LOG ANALYTICS:** Oracle EMECA, Splunk, Sumologic, Solarwinds, TIMBER, Hband, logz.io.
- SOCIAL ANALYTICS:** Hootsuite, Sprinklr, Netbase, Synthesio, Tracx, SimpleReach, Bitly, SimilarWeb.
- WEB / MOBILE / COMMERCE ANALYTICS:** Google Analytics, Mixpanel, AMP, Airtable, RES, SIGOPT, gra, Custora.

HEALTHCARE

- Clover
- Kyrus
- HealthTap
- Ginger
- Glow
- babylon
- zbra
- Patni
- ovia
- patientslikeme
- AiCure
- insitro
- runo
- citizen
- prognos
- reursion
- manipone
- DNAexus

LIFE SCIENCES

- standi
- color
- Karogen
- BenevolentAI
- verily
- WuXiNextCODE
- iscrip
- Clear Labs
- freemove
- MANIPONE
- DNAexus

TRANSPORTATION

- UBER
- TESLA
- CLEARPATH
- CRUISE
- nuro
- ASGO
- drive.ai
- CAMBRIDGE
- Aurora
- navto
- AMOTIVE
- G7
- PILOT.AI
- NIO
- OPTIMUS
- moovit
- Ike

AGRICULTURE

- FARMERS
- Granular
- JOHN DEERE
- BLUE RIVER
- HowGood
- heureka
- AgroStar
- FarmLogs

COMMERCE

- instacart
- FAIRE
- STITCH FIX
- Di & Co
- RetailNext
- SCORTEX
- TACHYUS
- other
- eharmony
- stem
- Amper
- ByteDance

INDUSTRIAL

- AVEVA
- SIEMENS
- PREDIX
- UPTAKE
- OSI
- ByteDance

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS 1010DATA vmware TIBCO TERADATA ORACLE NetApp syncsoft MAPR cloudera

OPEN SOURCE

The diagram illustrates the landscape of open-source technologies across several key areas:

- FRAMEWORKS**: Includes Hadoop, Spark, Flink, YARN, TEZ, Mesos, Kubernetes, Docker, COAP, Redhat, and HELIX.
- QUERY / DATA FLOW**: Features Spark SQL, Presto, Apache Drill, SLAMDATA, GraphQL, and Flink.
- DATA ACCESS & DATABASES**: Lists Cassandra, MongoDB, Redis, Cockroach Labs, Druid, CouchDB, OrientDB, SciDB, Riak, HBase, Cloud Spanner, and BigTable.
- ORCHESTRATION & MGMT**: Shows Talend, Apache Zookeeper, Apache Ambari, Apache Airflow, Mesos, etcd, and Kong.
- STREAMING & MESSAGING**: Includes Spark Streaming, NiFi, Flink, Beam, Kafka, Storm, and Apache RocketMQ.
- STAT TOOLS & LANGUAGES**: Covers Python, R, Scala, Jupyter Studio, SciPy, Julia, and R.
- AI OPS & INFRA**: Mentions mlflow, Kubeflow, MLsec, SELDON, PyTorch, and Polyaxon.
- AI / MACHINE LEARNING**: Highlights TensorFlow, OpenAI, DMTK, Weaviate, mxnet, VESLS, Chainer, Microsoft SENSE, DIMSUM, FeatureFu, ONNX, WEKA, Ludwig, and PyTorch.
- SEARCH**: Includes Elasticsearch and Solr.
- LOGGING & MONITORING**: Features Elasticsearch, Kibana, Sentry, Logstash, Prometheus, Fluentbit, Fluentd, Grafana, and Vector.
- VISUALIZATION**: Lists matplotlib, TensorBoard, seaborn, and Bokeh.
- COLLABORATION**: Includes BeakerX, Jupyter, and Anaconda.
- SECURITY**: Shows Apache Ranger, Knox, Sentry, and Accumulo.

DATA SOURCES & APIs

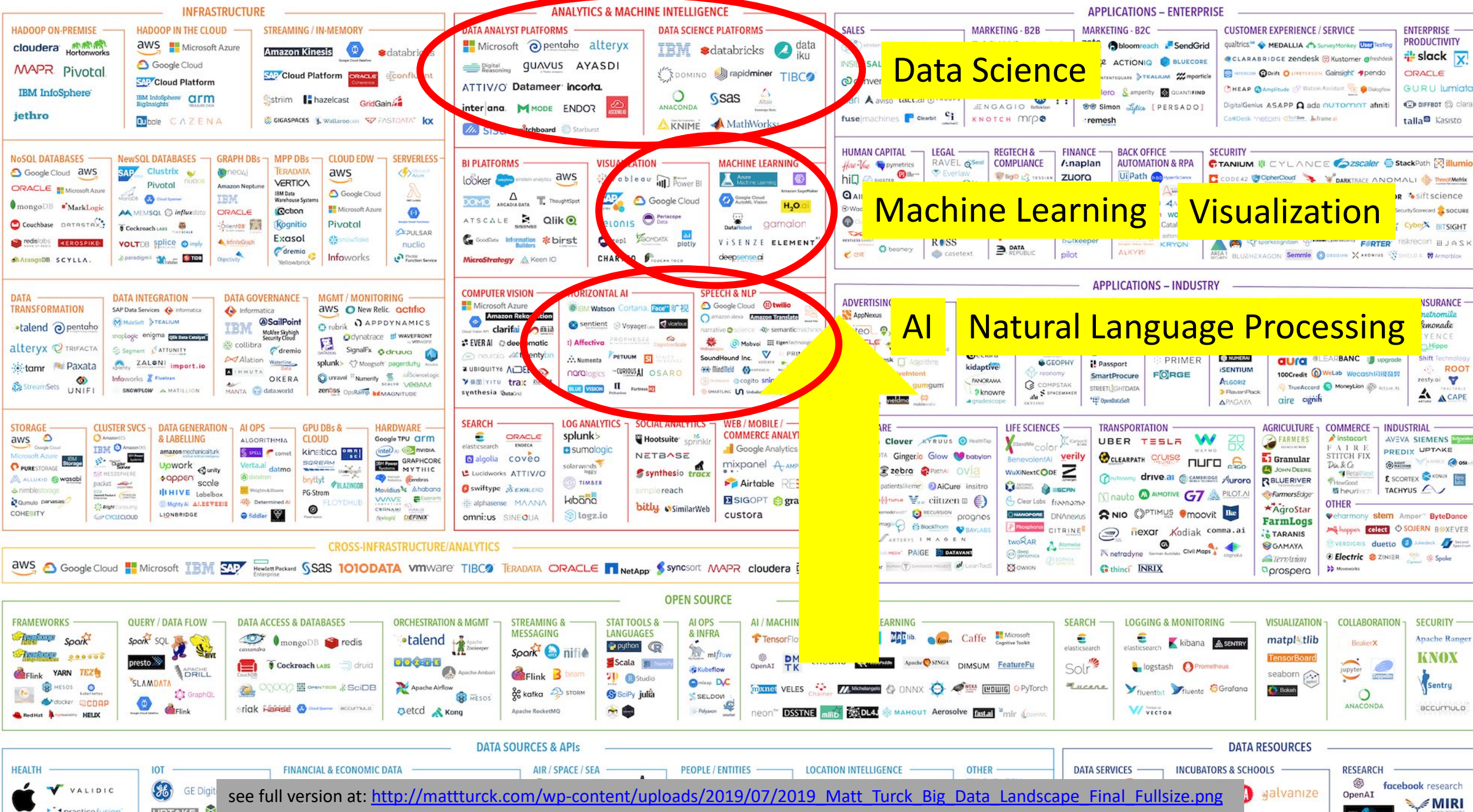
HEALTH — IOT — FINANCIAL & ECONOMIC DATA — AIR / SPACE / SEA — PEOPLE / ENTITIES — LOCATION INTELLIGENCE — OTHER — DATA SERVICES — INCUBATORS & SCHOOLS — RESEARCH —

see full version at: http://mattturck.com/wp-content/uploads/2019/07/2019_Matt_Turck_Big_Data_Landscape_Final_Fullsize.png

DATA RESOURCES

TEACHERS & SCHOOLS **RESEARCH**

DATA & AI LANDSCAPE 2019



A path in data science

(1) strong data systems skills

- (i) coding skills
- (ii) system architecture insights
performance tradeoffs

(2) application domain knowledge

(3) statistics, machine learning, math tools

Academic Research

Industry