CS 744: SUMMARY

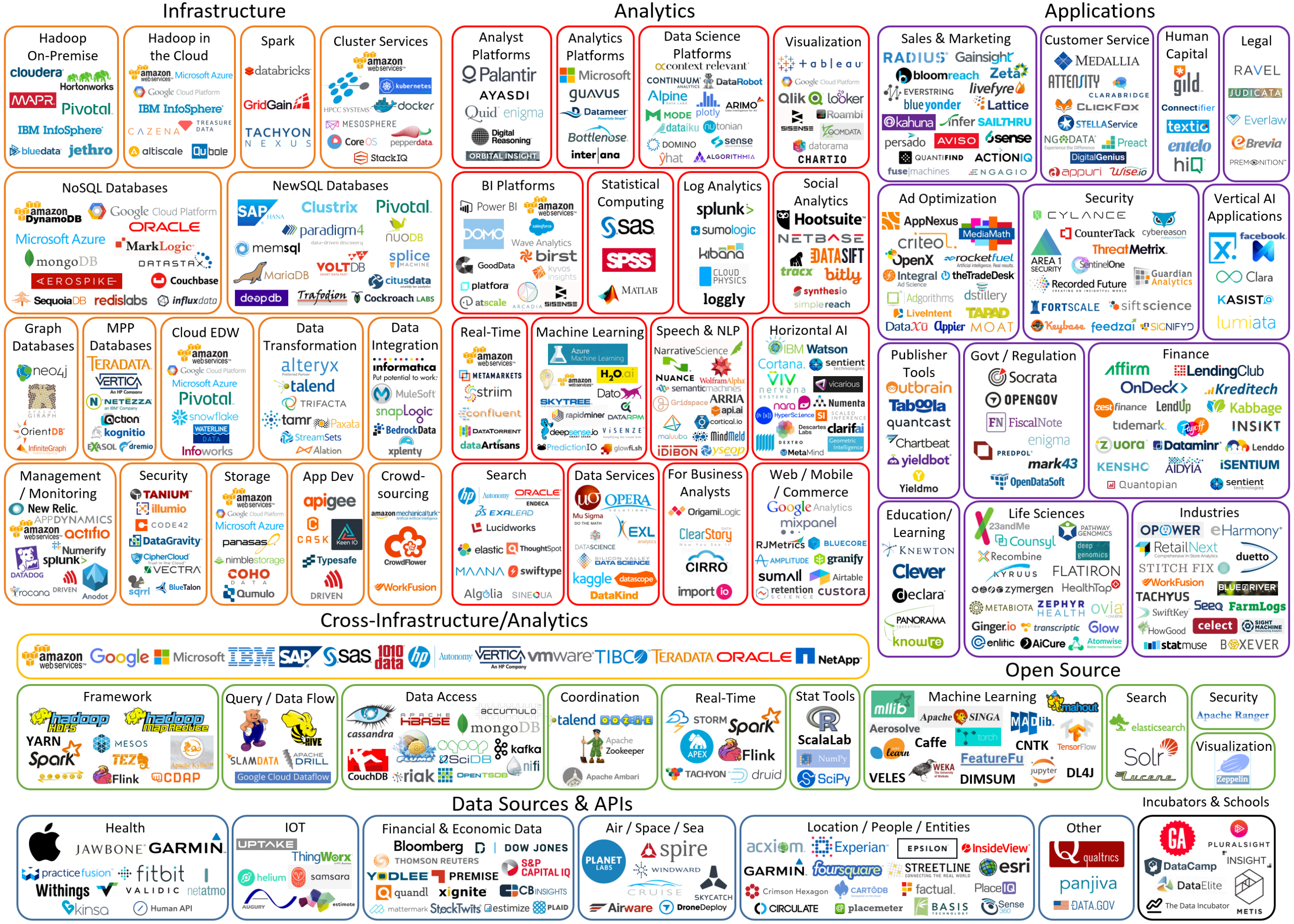
Shivaram Venkataraman

Fall 2019

ADMINISTRIVIA

- Midterm 2 on Tuesday
- Poster session Dec 13th, 3-5pm details on Piazza
- Final report Dec 17th
- AEFIS Course feedback form!

Big Data Landscape 2016 (Version 3.0)



Infrastructure

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

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

Spark
databricks, GridGain, TACHYON NEXUS

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

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

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, NUODB, memsql, splice MACHINE, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, GIRAPH, OrientDB, InfiniteGraph

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

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

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

Data Integration
informatica, MuleSoft, snapLogic, Bedrock Data, xplenty

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

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

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

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

Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

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

Analytics Platforms
Microsoft, GUAVUS, Datameer, Bottlenose, inter|ana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, MODE, plotly, ARIMO, dataiku, Nutonian, DOMINO, sense, yhat, ALGORITHMIA

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

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

Statistical Computing
sas, SPSS, MATLAB

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

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

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

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

Speech & NLP
NarrativeScience, NUANCE, WolframAlpha, semanticmachines, Gridspace, ARRIA, api.ai, maluuba, cortical.io, MindMeld, iDiBON, yseop

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

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

Data Services
Mu Sigma, OPERA SOLUTIONS, EXL analytics, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

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

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

Applications

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

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

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

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

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

Vertical AI Applications
X.ai, facebook M, Clara, KASISTO, lumiata

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

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

Finance
affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, Payoff, INSIKT, Zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient technologies

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

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

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

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

Open Source

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

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

Data Access
cassandra, ACCUMULO, HBASE, mongoDB, kafka, SciDB, CouchDB, riak, OPENTSDB, nifi

Coordination
talend, Apache Zookeeper, Apache Ambari

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

Stat Tools
R, ScalaLab, NumPy, SciPy

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

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

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

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

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

Location / People / Entities
acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense360

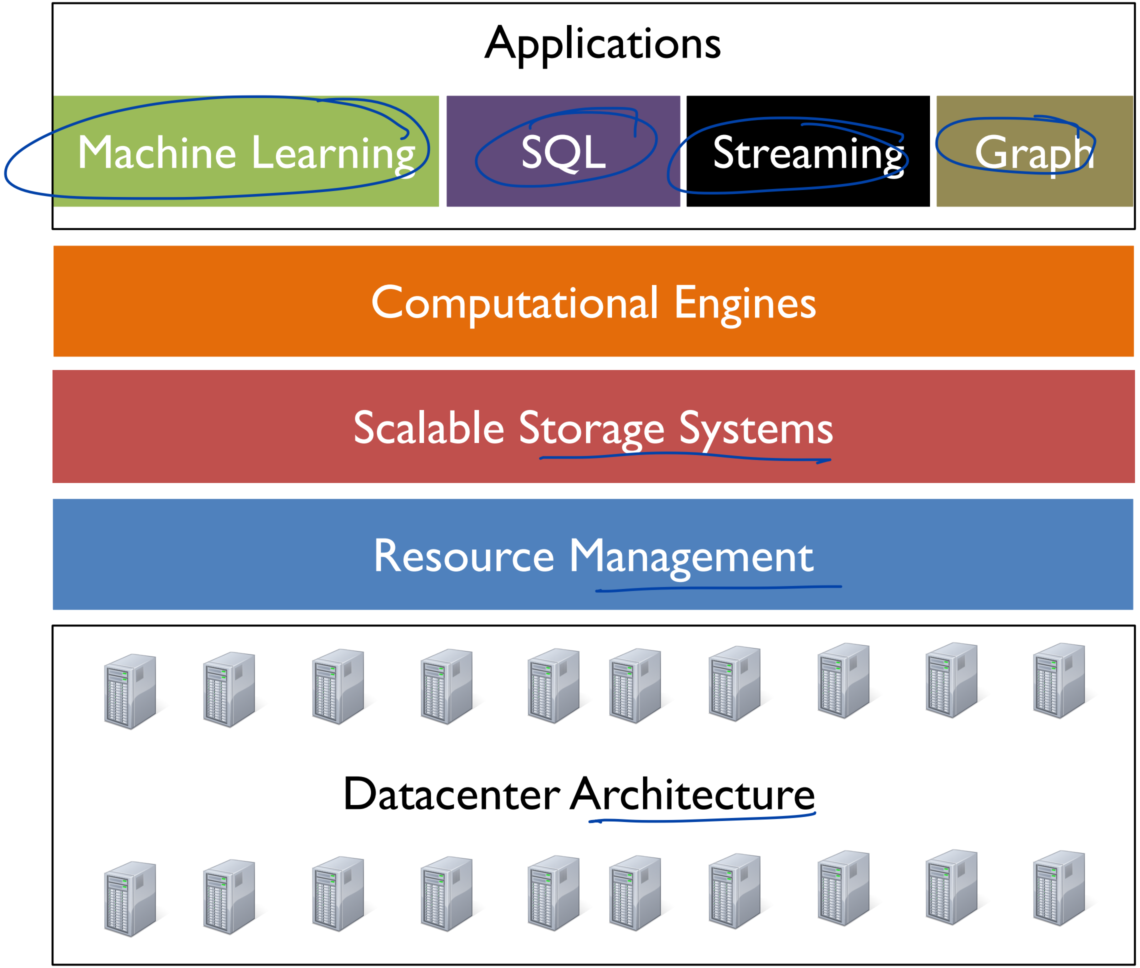
Other
qualtrics, panjiva, DATA.GOV

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

Last Updated 3/23/2016

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

FIRSTMARK



Applications

Machine Learning

SQL

Streaming

Graph

Computational Engines

Scalable Storage Systems

Resource Management

Datacenter Architecture

TensorFlow

Spark

hadoop

Apache MESOS

Open Compute Project

OUTLINE

Unification vs Specialization

Survey results, Discussion

Big data systems: Looking forward

SPECIALIZATION VS UNIFICATION

GENERALITY: "ONE SIZE FITS ALL" DBMS

1970s

IBM

Research prototypes: SystemR and INGRES

Main function: OLTP

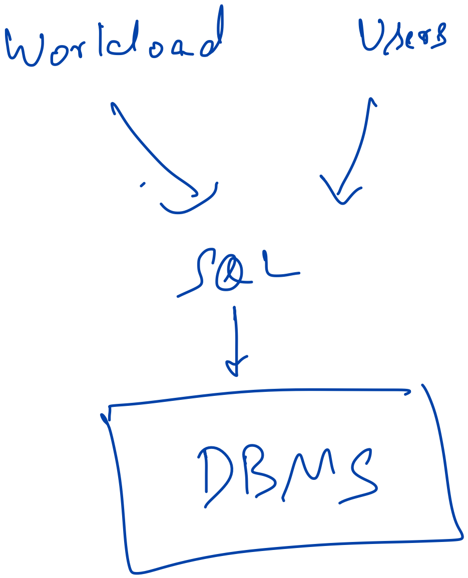
Process data to derive decisions

From 1990s

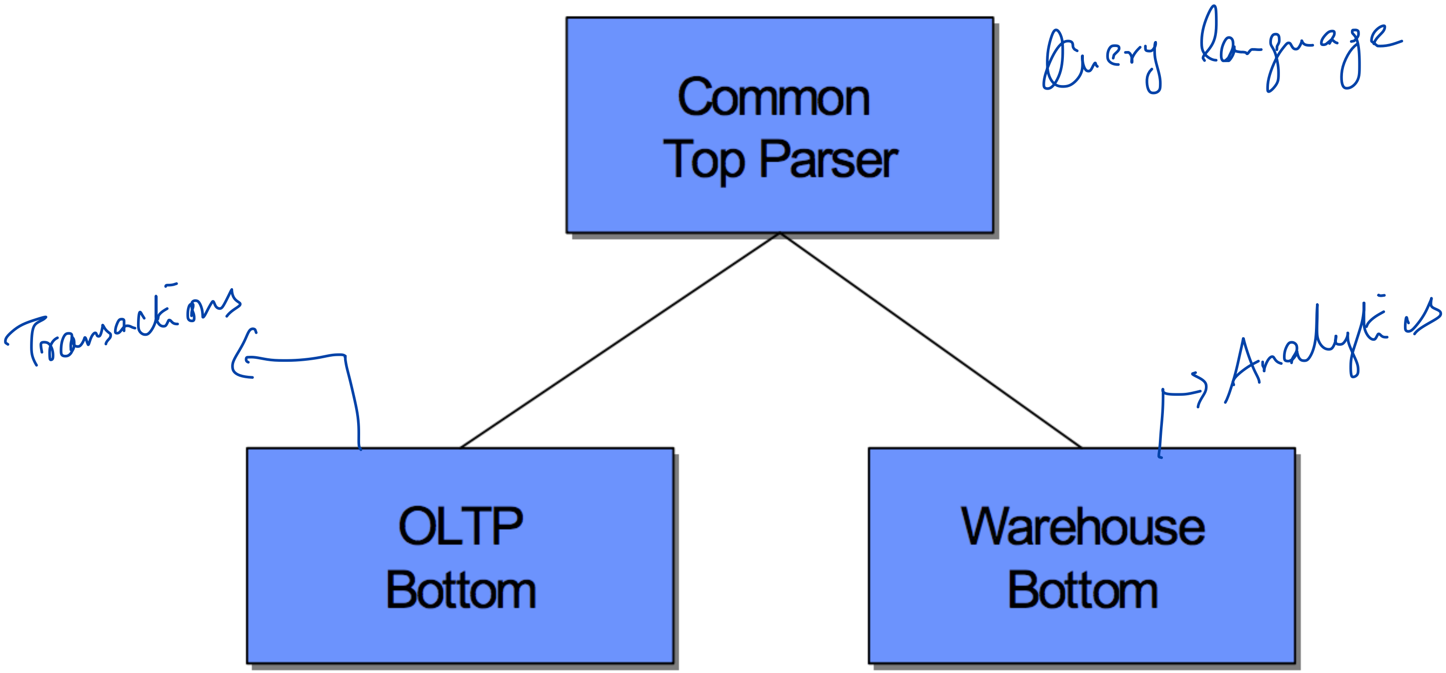
Rise of business intelligence workloads

OLAP workloads need to be isolated from OLTP

Solution: Scrape data into data warehouses.



DBMS IMPLEMENTATION



Common Top Parser

OLTP Bottom

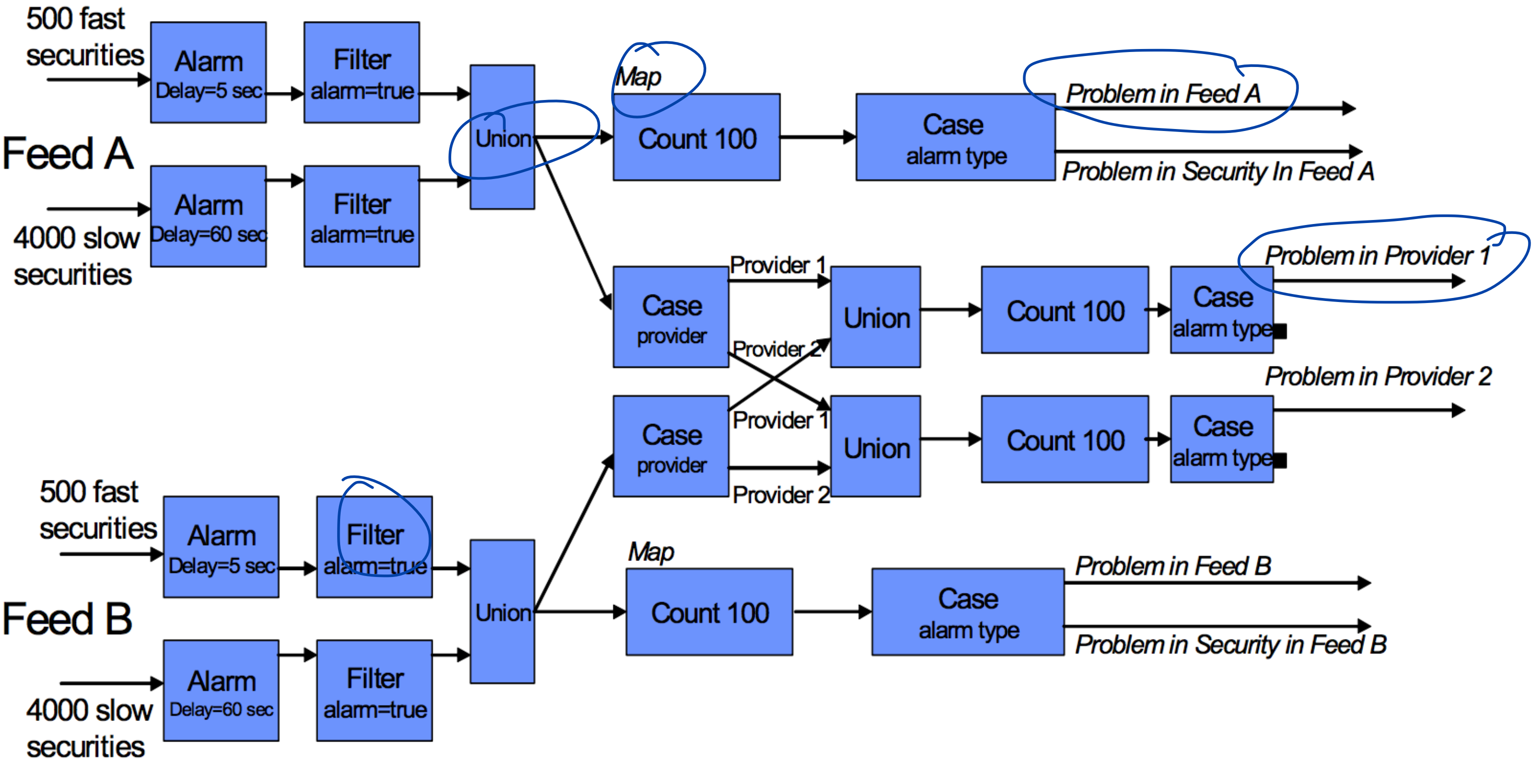
Warehouse Bottom

STREAM PROCESSING ?

early 2000s

every 1s

Example: Financial feed processing (Bloomberg, Reuters)



EXAMPLE WORKLOAD

Goals: Maximize message processing throughput on single machine

Scenario: Stock tick is late is if it occurs more than X secs from previous tick

Performance comparison:

2.8 GHz, 512 MB memory, single SCSI disk

160,000 messages per second with StreamBase → Custom stream processing

900 messages per second with DBMS → with support for streaming

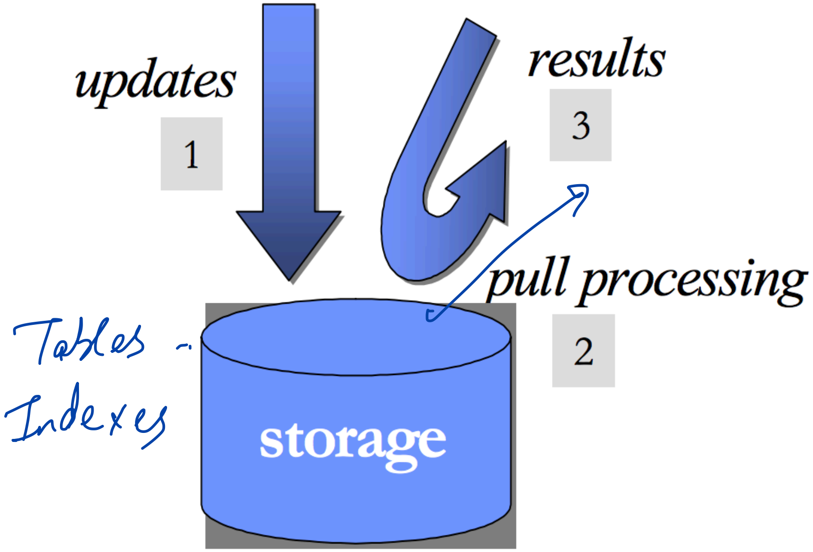
WHY IS IT SLOW ?

DBMS: "Outbound" processing model

1. Insert data
2. Index data, commit transaction
3. Process query, return results

Process after store

First store data
and
then Process the data



WHY IS IT SLOW ?

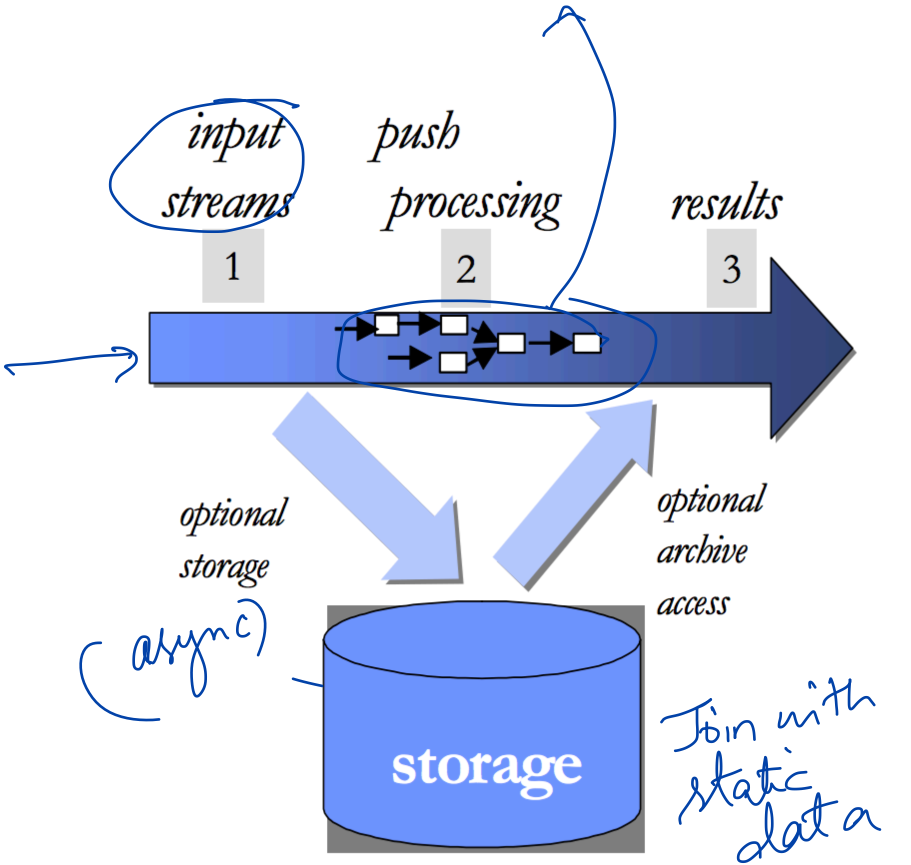
First store the query

"Inbound" data processing

1. Push inputs into system
2. Process query
3. Return results
4. Optionally store (async)

Only way to do this in DBMS: Triggers

Not performant



(async)

Join with static data

OUTBOUND

on top of this model

INBOUND

"Pull" records given query

Store data, run any query

trigger

"Push" records into query

Store queries, pass data through

IS IT JUST STREAMING ?

Sensor Networks: TinyDB

→ IoT or edge-computing

Text Search: GFS / MapReduce

→ Google

Scientific databases: SciDB

↳ multi dimensional array

Data warehouses

Column stores, read-oriented vs. write oriented

TF ← unification

BIG DATA SYSTEMS

Unified systems

Naiad - Timely → Batch → stream → Graph

PyWren | MapReduce → Arch → Map/Reduce API

SQL → Spark
Streaming → Spark
GraphX → Spark

Stuff

Clipper ← TF
Clipper ← Pytorch
Scikit Learn

Specialized systems

TPU → ML inference workloads

PS → very large models

Weld

Ray ? - RL applications

API [functions
Actors
get]

Powergraph → Power-law graphs

BENEFITS

Unified systems

Developer ease of use

→ No need to stitch thing together

Additional workloads

Hard to build → Abstractions
→ Perform

Complexity

Specialized systems

↳ Performance ! → Exploit workload

Simple code

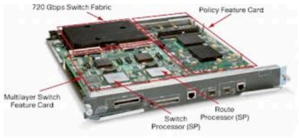
Industry specific
↓
Vendor choice ?

IS IT JUST A CYCLE ?

80s

SPARC

MIPS



PowerPC

90s

Intel® Core™ i7 processor

→ high end

→ Mobile phones

GPUs

TPUs

?

ASICs

WHERE ARE WE IN THE CYCLE ?

Oracle

PostgreSQL

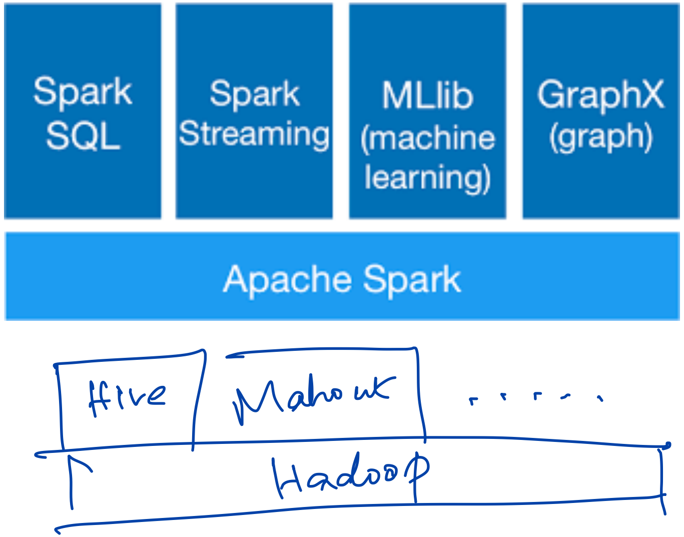
Dryad

CIEL

hadoop

Spark

2004 - 2011



Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

Hive

Mahout

. . . . .

Hadoop

2011 - 2015

TensorFlow

PyTorch

Streaming

Apache Flink

2015 - now

BOOTSTRAPPING UNIFIED SYSTEMS ?

1. Implement a system/app/functionality that is superior to what is out there
2. Rapidly build an ecosystem providing additional functionalities

Example:

Tensorflow initially target SGD/deep learning

Unifies number of other features

- tf.data supporting map, flat_map etc.
- tf.linalg implementing linear algebra
- tf.sparse for sparse data / shallow models

Apache Arrow

Protobuf

SURVEY RESULTS

LEARNING OBJECTIVES

At the end of the course you will be able to

- Explain the design and architecture of big data systems
- Compare, contrast and evaluate research papers
- Develop and deploy applications on existing frameworks
- Design, articulate and report new research ideas

Paper Review

Discussion

Assignment

Project

DISCUSSION

https://forms.gle/sQFiAKwiQfHEKkPd8

What were some of your goals when you started the course? (Think about the first survey.) Reflect on what part of your goals have been achieved and how.

- Arch design patterns
- Historical lineage of why we use what we use.
- What are the metrics which matter?
- Critically evaluate, Compare → good
  → short comings
- How to build such a system

In the class, we discussed one trend across systems of unification vs. specialization. What are some other trends you have noticed across the papers in the class?

→ Eager vs. lazy execution

↳ Optimizations

Stateful | Stateless ↳ elastic

↓ memory

more computation

→ Latency vs. Tput vs. Correctness → Trade-off Space

→ Fault tolerance → Checkpointing

Lineage

Ignore it

Single point of failure - Centralized or not

→ Hardware → Commodity - stragglers, fault

↓

Specialized - TPUs (7ms)

Sync vs. Async

ML/Graph

Open Source vs. Closed source

↓ design evolves influenced

↓ Fixed design

Data proc.

Spark
TF
Pytorch
Flink

Storage

S3 → 99.99%
GFS
⋮

Mutability ↳ State

vs.

Immutability ↳ lineage

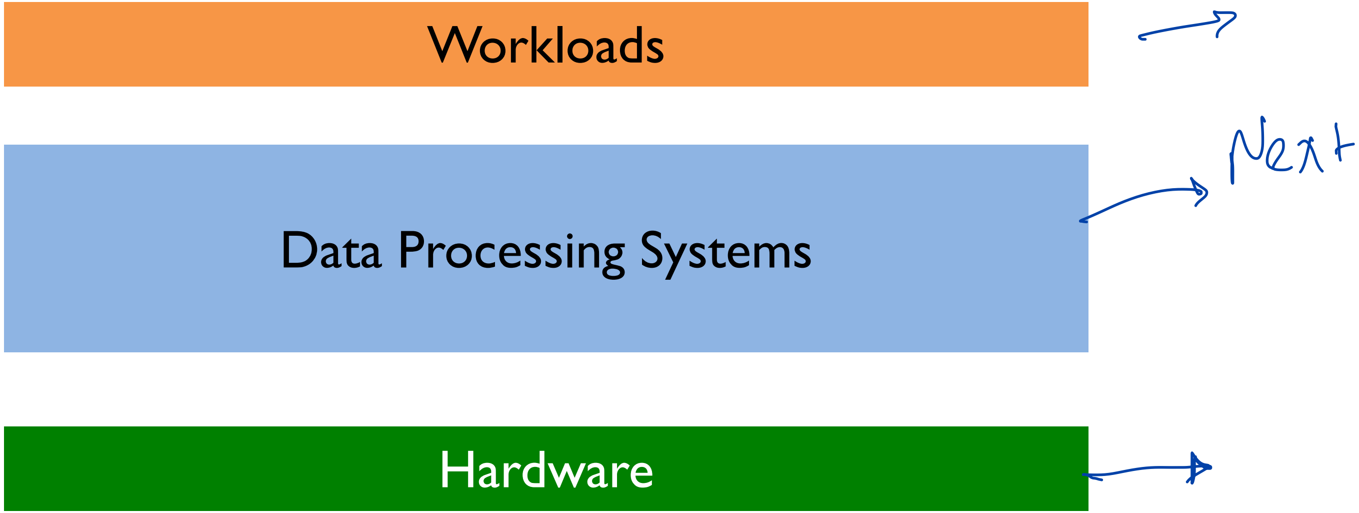
API design

Naiad ↳ low-level
vs.
TF/Keras high-level

Driver → managing computation or not

LOOKING FORWARD

NEXT-GENERATION BIG DATA SYSTEMS ?



Workloads

Data Processing Systems

Hardware

Next

TRENDS IN WORKLOADS

Workloads

Infrastructure

New functionalities

- Data science / AI
- Robotics → RL

New data sources

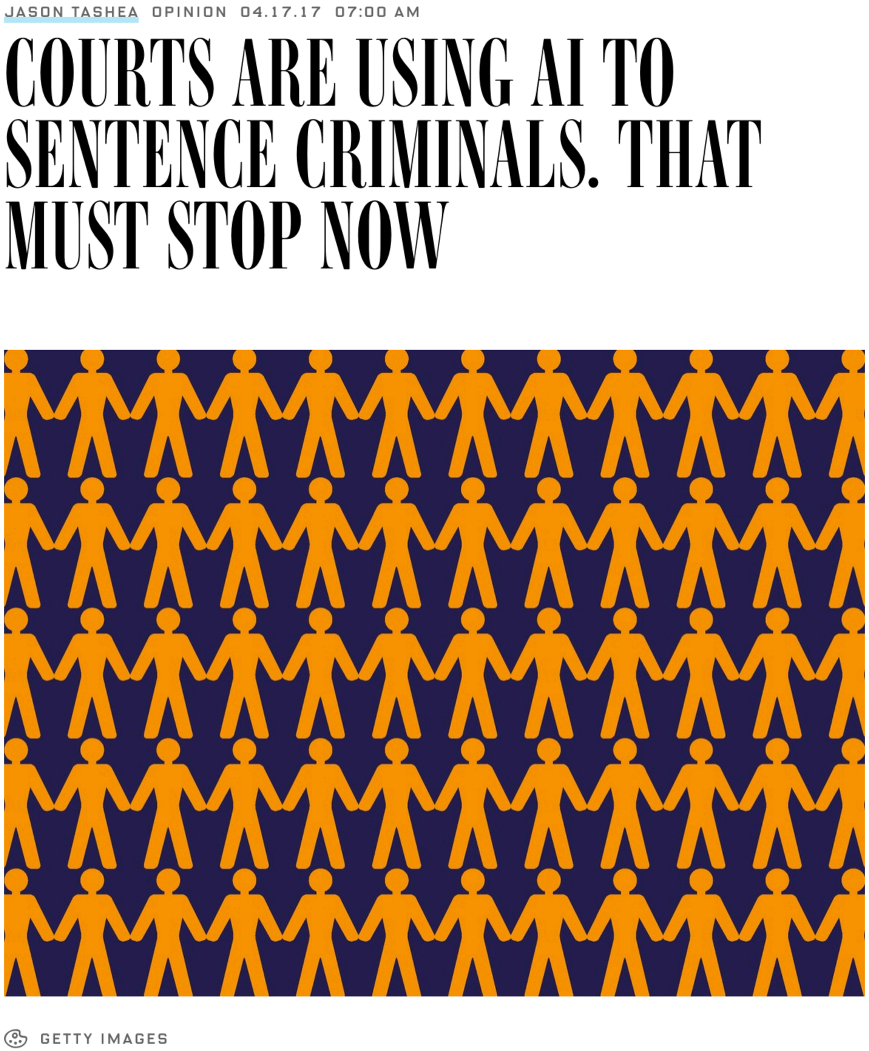
- Bio-medical data → Sequence genomes → MRI
- Video streams → richer
- IoT / edge devices

tuples → HD videos

larger & richer

DIVERSITY ?

Fairness in ML?



JASON TASHEA OPINION 04.17.17 07:00 AM

COURTS ARE USING AI TO SENTENCE CRIMINALS. THAT MUST STOP NOW

GETTY IMAGES

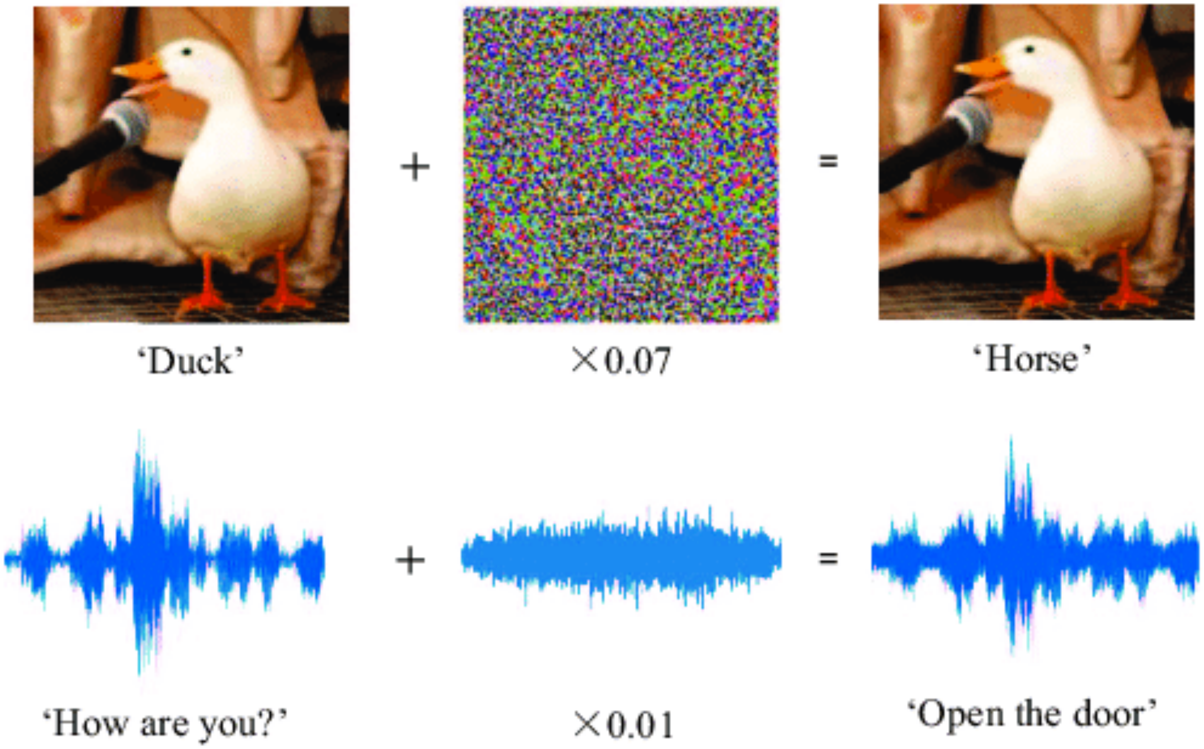
HOW ROBUST IS YOUR SYSTEM ?

Failure

infrastructure

data analysis

"Adversarial examples"



WHAT CAN SYSTEMS RESEARCH DO ?

More than performance?

- Latency, throughput, efficiency
- Ease of use

Some other goals to consider ?

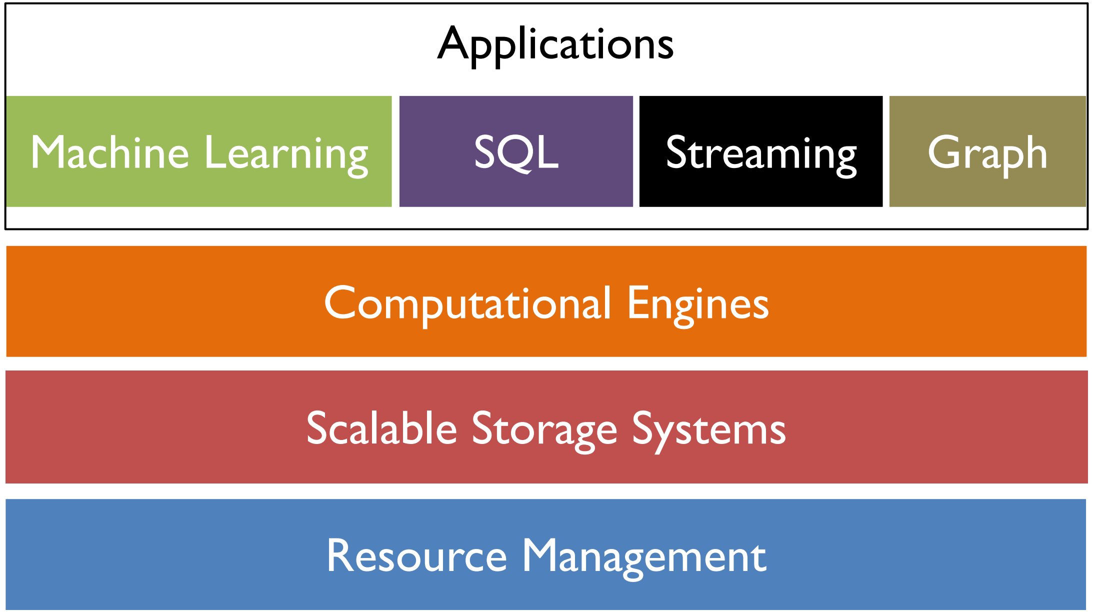
- Security, Privacy
- Robustness
- Data bias / ethics

COURSE SUMMARY

Large scale data analysis has changed the world

→

COURSE SUMMARY



Applications

Machine Learning | SQL | Streaming | Graph

Computational Engines

Scalable Storage Systems

Resource Management

Your System Here ?

Spark

TensorFlow

hadoop

kubernetes

Apache MESOS