

MapReduce

CS 475: Concurrent & Distributed Systems (Fall 2021)

Lecture 4

Yue Cheng

Some material taken/derived from:

- Princeton COS-418 materials created by Michael Freedman and Wyatt Lloyd.
- MIT 6.824 by Robert Morris, Frans Kaashoek, and Nickolai Zeldovich.

Licensed for use under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Applications

Web
apps

Data
processing

Data
storage

Emerging
apps?

Resource management

Compute
resources

Memory
resources

Storage
resources

Network
resources



Datacenter infrastructure



Applications

Web
apps

Data
processing

Data
storage

Emerging
apps?

Resource management

Compute

Memory

Storage

Network

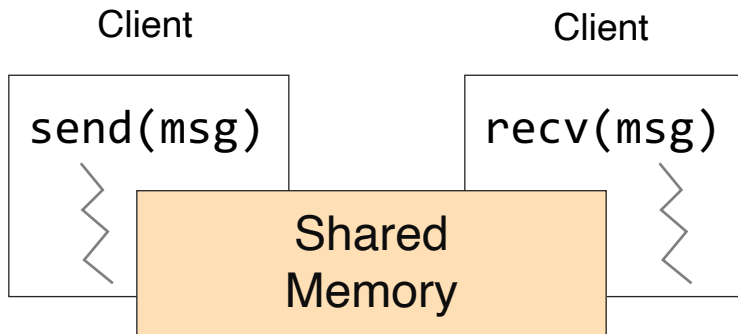
Question: How to program these many computers?



Datacenter infrastructure



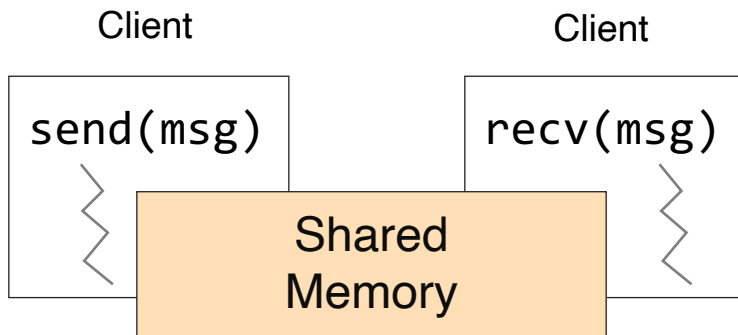
Review: Shared memory



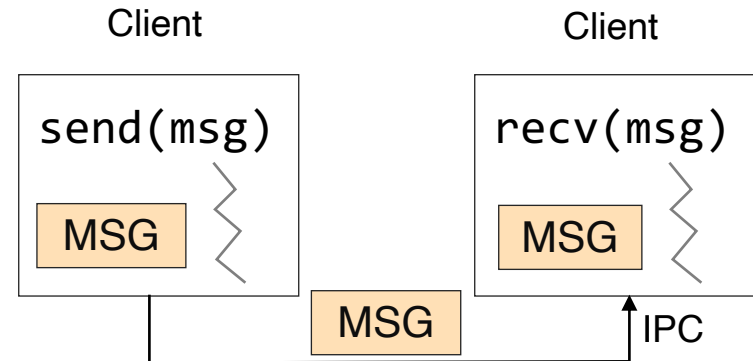
- Shared memory: multiple processes to share data via memory
- Applications must locate and map shared memory regions to exchange data

Review:

Shared memory vs. Message passing



- Shared memory: multiple processes to share data via memory
- Applications must locate and map shared memory regions to exchange data



- Message passing: exchange data explicitly via IPC
- Application developers define protocol and exchanging format, number of participants, and each exchange

Review:

Shared memory vs. Message passing

- Easy to program; just like a single multi-threaded machines
- Hard to write high perf. apps:
 - Cannot control which data is local or remote (remote mem. access much slower)
- Hard to mask failures
- Message passing: can write very high perf. apps
- Hard to write apps:
 - Need to manually decompose the app, and move data
- Need to manually handle failures

Shared memory: Pthread

- A POSIX standard (IEEE 1003.1c) API for thread creation and synchronization
- API specifies behavior of the thread library, implementation is up to development of the library
- Common in UNIX (e.g., Linux) OSes

Shared memory: Pthread

```
void *myThreadFun(void *vargp) {
    sleep(1);
    printf("Hello world\n");
    return NULL;
}

int main() {
    pthread_t thread_id_1, thread_id_2;
    pthread_create(&thread_id_1, NULL, myThreadFun, NULL);
    pthread_create(&thread_id_2, NULL, myThreadFun, NULL);
    pthread_join(thread_id_1, NULL);
    pthread_join(thread_id_2, NULL);
    exit(0);
}
```


Message passing: MPI

- MPI – Message Passing Interface
 - Library standard defined by a committee of vendors, implementers, and parallel programmers
 - Used to create parallel programs based on message passing
- Portable: one standard, many implementations
 - Available on almost all parallel machines in C and Fortran
 - De facto standard platform for the HPC community

Message passing: MPI

```
int main(int argc, char **argv) {
    MPI_Init(NULL, NULL);

    // Get the number of processes
    int world_size;
    MPI_Comm_size(MPI_COMM_WORLD, &world_size);

    // Get the rank of the process
    int world_rank;
    MPI_Comm_rank(MPI_COMM_WORLD, &world_rank);

    // Print off a hello world message
    printf("Hello world from rank %d out of %d processors\n",
           world_rank, world_size);

    // Finalize the MPI environment
    MPI_Finalize();
}
```

Message passing: MPI

```
mpirun -n 4 -f host_file ./mpi_hello_world
```

```
int main(int argc, char **argv) {
    MPI_Init(NULL, NULL);

    // Get the number of processes
    int world_size;
    MPI_Comm_size(MPI_COMM_WORLD, &world_size);

    // Get the rank of the process
    int world_rank;
    MPI_Comm_rank(MPI_COMM_WORLD, *world_rank);

    // Print off a hello world message
    printf("Hello world from rank %d out of %d processors\n",
           world_rank, world_size);

    // Finalize the MPI environment
    MPI_Finalize();
}
```

MapReduce

The big picture (motivation)

- Datasets are **too big** to process using a single computer

The big picture (motivation)

- Datasets are **too big** to process using a single computer
- Good parallel processing engines are **rare (back then in the late 90s)**

The big picture (motivation)

- Datasets are **too big** to process using a single computer
- Good parallel processing engines are **rare (back then in the late 90s)**
- Want a parallel processing framework that:
 - is **general** (works for many problems)
 - is **easy to use** (no locks, no need to explicitly handle communication, no race conditions)
 - can **automatically parallelize** tasks
 - can **automatically handle machine failures**

Context (Google circa 2000)

- Starting to deal with **massive** datasets
- But also addicted to cheap, unreliable hardware
 - Young company, expensive hardware not practical
- Only a few expert programmers can write distributed programs to process them
 - Scale so large jobs can complete before failures



Context (Google circa 2000)

- Starting to deal with **massive** datasets
- But also addicted to cheap, unreliable hardware
 - Young company, expensive hardware not practical
- Only a few expert programmers can write distributed programs to process them
 - Scale so large jobs can complete before failures
- **Key question:** how can every Google engineer be imbued with the ability to write **parallel, scalable, distributed, fault-tolerant** code?
- **Solution:** **abstract out** the redundant parts
- **Restriction:** relies on job semantics, so restricts which problems it works for

Application: Word Count

```
cat data.txt  
  | tr -s '[:punct:][:space:]' '\n'  
  | sort | uniq -c
```

```
SELECT count(word), word FROM data  
      GROUP BY word
```

Deal with multiple files?

1. Compute word counts from individual files

Deal with multiple files?

1. Compute word counts from individual files
2. Then merge intermediate output

Deal with multiple files?

1. Compute word counts from individual files
2. Then merge intermediate output
3. Compute word count on merged outputs

What if the data is too big to fit in one computer?

1. In parallel, send to worker:
 - Compute word counts from individual files
 - Collect results, wait until all finished

What if the data is too big to fit in one computer?

1. In parallel, send to worker:
 - Compute word counts from individual files
 - Collect results, wait until all finished
2. Then merge intermediate output

What if the data is too big to fit in one computer?

1. In parallel, send to worker:
 - Compute word counts from individual files
 - Collect results, wait until all finished
2. Then merge intermediate output
3. Compute word count on merged intermediates

MapReduce: Programming interface

- $\text{map}(k1, v1) \rightarrow \text{list}(k2, v2)$
 - Apply function to $(k1, v1)$ pair and produce set of intermediate pairs $(k2, v2)$

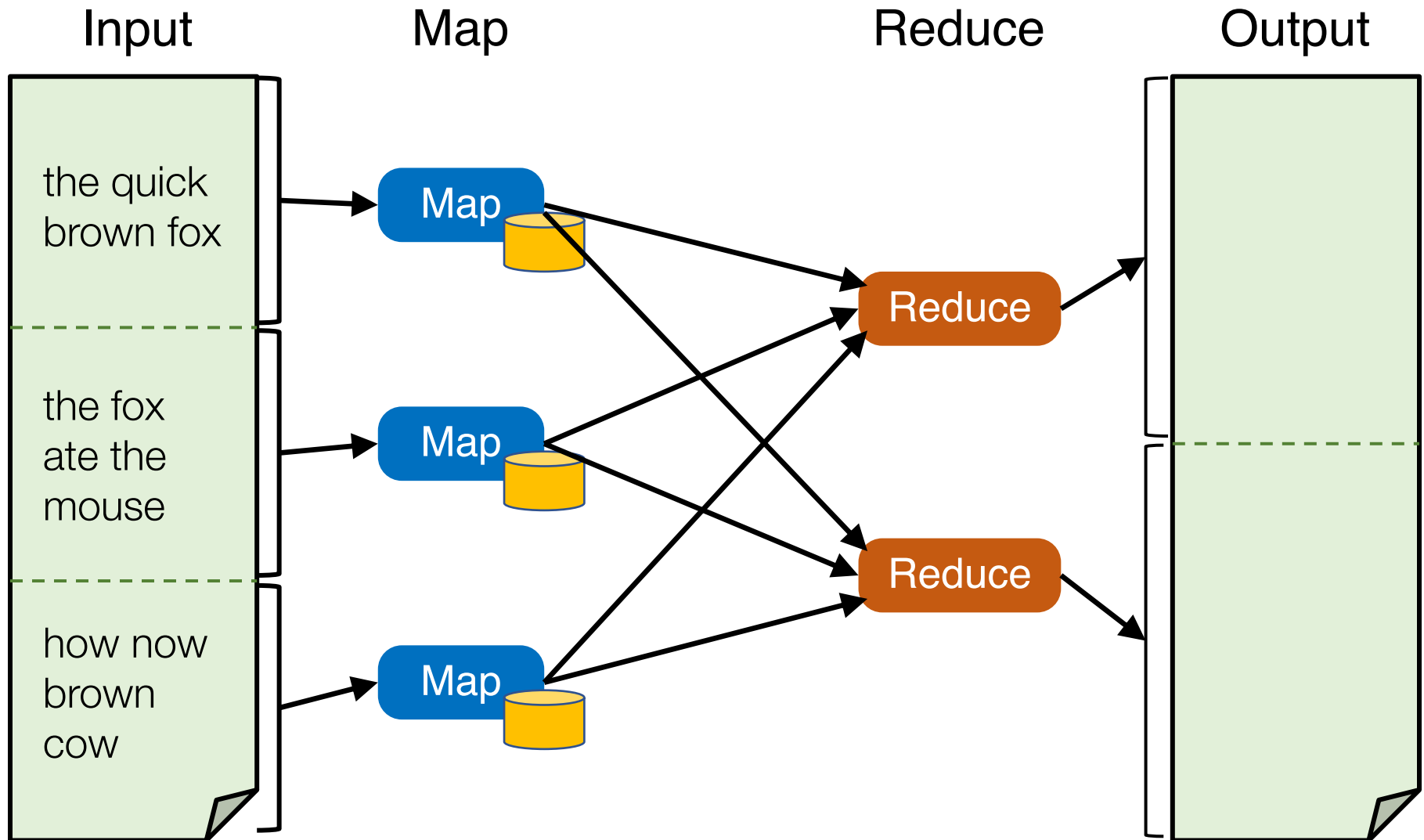
- $\text{reduce}(k2, \text{list}(v2)) \rightarrow \text{list}(k3, v3)$
 - Apply aggregation (reduce) function to values
 - Output results

MapReduce: Word Count

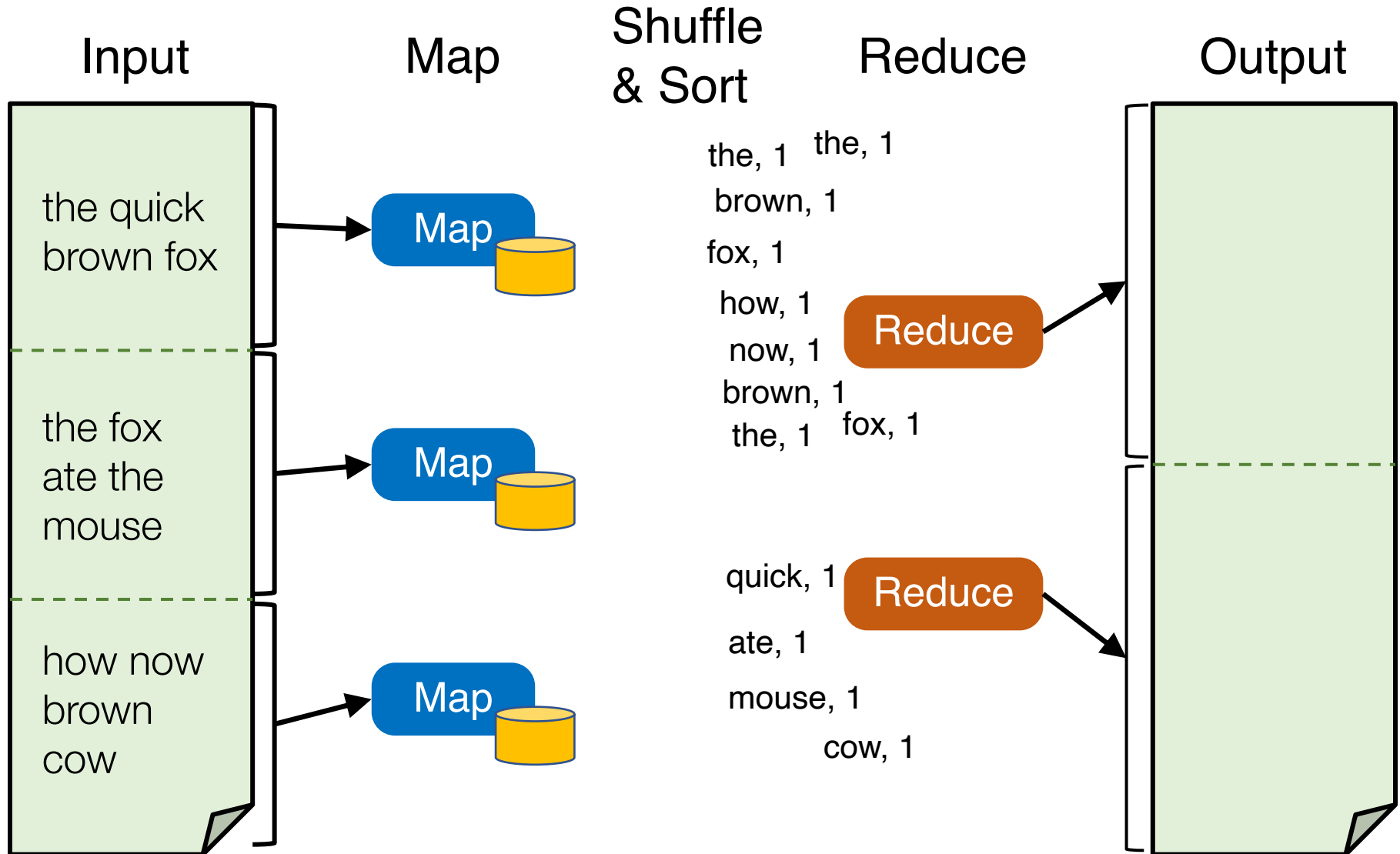
```
map(key, value):  
    for each word w in value:  
        EmitIntermediate(w, "1");
```

```
reduce(key, values):  
    int result = 0;  
    for each v in values:  
        result += ParseInt(v);  
    Emit(AsString(result));
```

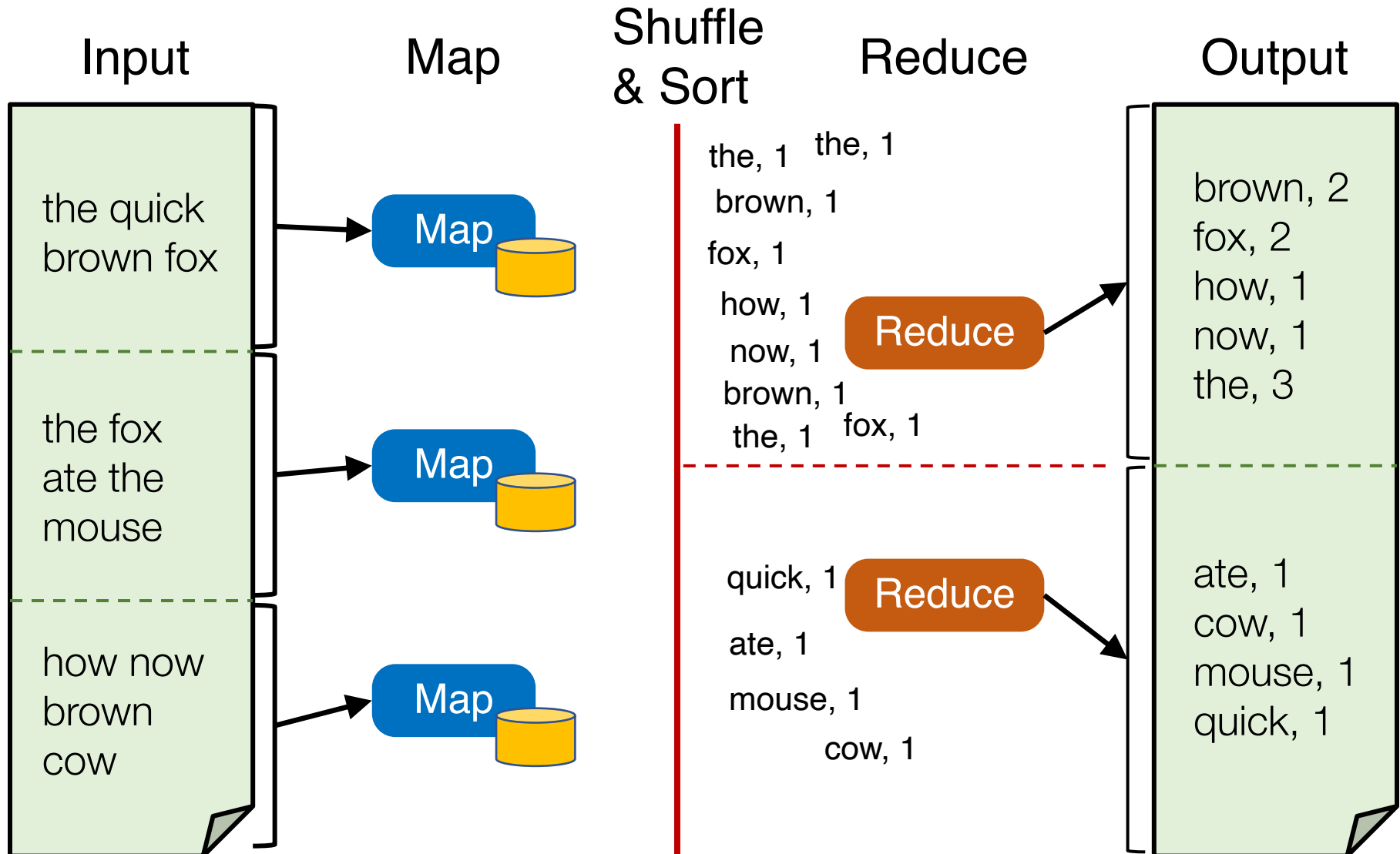
Word Count execution



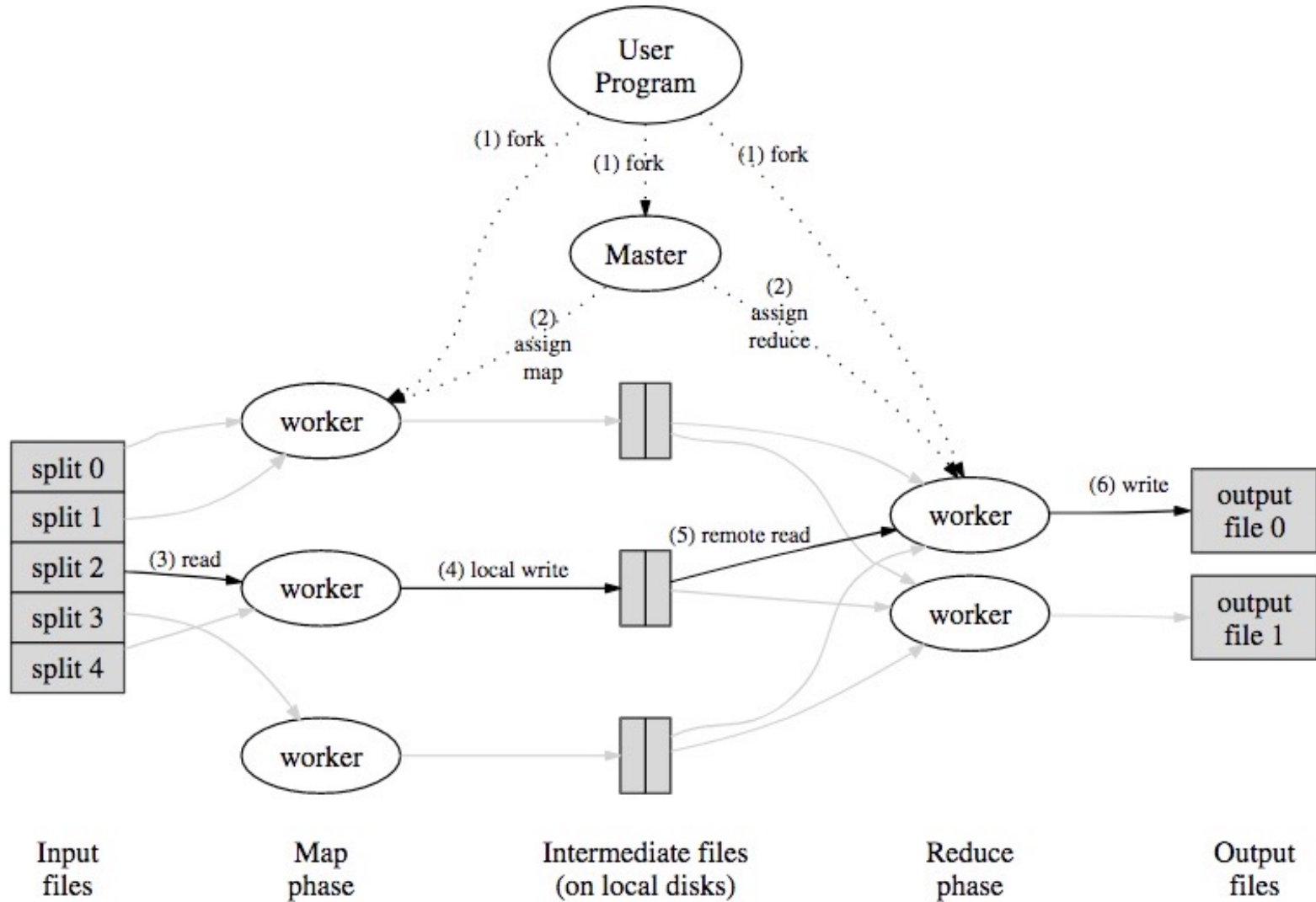
Word Count execution



Word Count execution

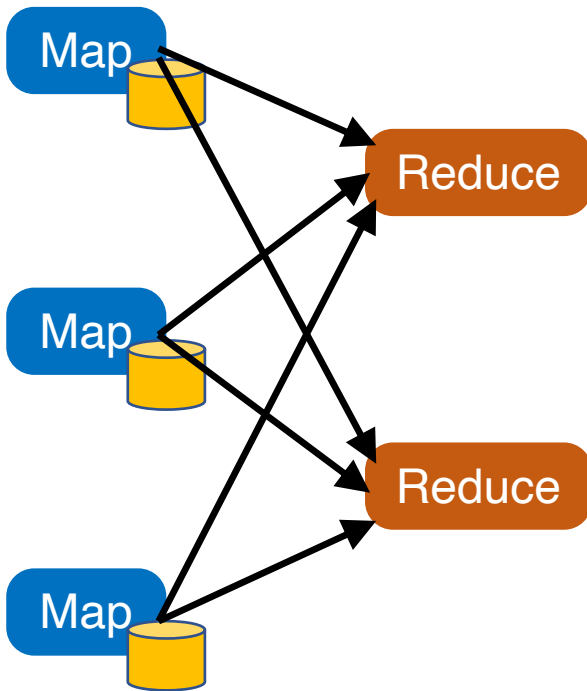


MapReduce data flows



MapReduce processes

Map Shuffle
& Sort Reduce



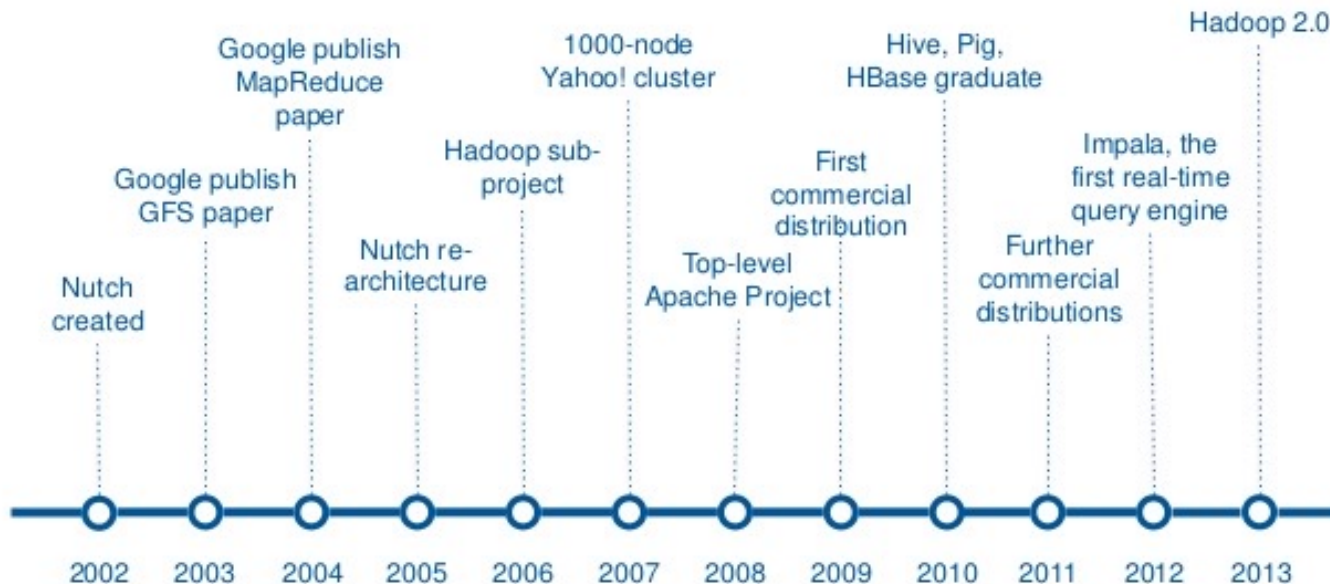
- Map workers write intermediate output to local disk, separated by partitioning. Once completed, tell master node
- Reduce worker told of location of map task outputs, pulls their partition's data from each mapper, execute function across data
- Note:
 - “All-to-all” shuffle b/w mappers and reducers
 - Written to disk (“materialized”) b/w each state

Apache Hadoop



- An open-source implementation of Google's MapReduce framework
 - Hadoop MapReduce atop Hadoop Distributed File System (HDFS)

A Brief History of Hadoop



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
 Oracle
 Databricks CAZENA

STREAMING / IN-MEMORY
 Amazon Kinesis
 SAP Cloud Platform
 Oracle
 Confluent
 Streamio
 hazelcast
 GridGain
 GIGASPACEs
 Wallaroo
 PASTDATA
 KX

NoSQL DATABASES
 Google Cloud AWS
 ORACLE Microsoft Azure
 mongoDB MarkLogic
 Couchbase DATASIX
 redislabs
 ArangoDB SCYLLA

NewSQL DATABASES
 SAP Clustrix
 Pivotal
 Microsoft Azure
 MEMSQL
 Cockroach Labs
 VoltDB
 Paradedata

GRAPH DBs
 Neo4j
 Amazon Neptune
 IBM
 Oracle
 InfoGraphix
 Graphistry

MPP DBs
 TERADATA
 VERTICA
 IBM Data Warehouse Systems
 Kogniton
 Exasol
 dremio
 Yellowbrick

CLOUD EDW
 AWS
 Google Cloud
 Microsoft Azure
 Snowflake
 Infoworks

SERVERLESS
 Amazon
 Google Cloud
 Microsoft Azure
 PULSAR
 nuclio
 Amazon FaaS

DATA TRANSFORMATION
 talend pentaho
 alteryx TRIFACTA
 tmnr Paxata
 StreamSets UNIFI

DATA INTEGRATION
 SAP Data Services Informatica
 Microsoft Azure
 Segment
 Zaloni
 Informatica
 Snowflake
 MANTILLION

DATA GOVERNANCE
 IBM
 Collibra
 Alation
 Microsoft Azure
 OKERA
 MANTA
 dataworld

MGMT / MONITORING
 AWS New Relic
 rubrik
 Dynatrace
 Signalix
 Splunk
 Unavox
 Numerify
 ZENOS

STORAGE
 AWS
 Microsoft Azure
 PURE STORAGE
 ALLUXIO
 QUMON
 COHERITY

CLUSTER SVCS
 Amazon
 IBM
 Microsoft Azure
 ALLUXIO
 WASABI
 QUMON
 COHERITY

DATA GENERATION & LABELLING
 Amazon
 Upwork
 Openden
 Scoble
 HIVE
 Alteryx
 Lionbridge

AI OPS
 ALGORITHMIA
 connect
 Verta.ai
 daimo
 Weights & Biases
 Determined AI
 Fiddler

GPU DBs & CLOUD
 Kinetica
 SOREREM
 BYTILY
 PG-Stream
 LLOYDHUB

HARDWARE
 Google
 Intel
 NVIDIA
 AMD
 IBM
 HANSA
 Movidius
 WAVE
 CRYSTAL BALL
 Intel
 NVIDIA
 AMD
 IBM
 HANSA
 Movidius
 WAVE
 CRYSTAL BALL

CROSS-INFRASTRUCTURE/ANALYTICS

AWS Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS 1010DATA VMware TIBCO TERADATA ORACLE NetApp syncsort MAPR cloudera

ANALYTICS & MACHINE INTELLIGENCE

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

DATA SCIENCE PLATFORMS
 IBM databricks data iku
 DOMINO rapidminer TIBCO
 AMACONDA SAS Altair
 KNIME MathWorks

BI PLATFORMS
 looker
 Domo
 AT SCALE
 GoodData
 MicroStrategy
 Keen IO

VISUALIZATION
 Tableau Power BI
 SAP
 Celonis
 Zepi
 CHARTIO

MACHINE LEARNING
 Azure
 DataRobot
 gamalon
 VISENZE
 ELEMENT

COMPUTER VISION
 Microsoft Azure
 Amazon Rekognition
 Clarifai
 EVER AI
 neuro.ai
 Ubiquity
 YUBI
 synthesis

HORIZONTAL AI
 IBM Watson Cortana
 sentiment
 Affective
 Humenta
 neurologix
 vision

SPEECH & NLP
 Google Cloud
 Amazon Lex Amazon Transcribe
 narrative science
 Movel
 SoundHound Inc
 Bluebird
 cogito slips
 SMARTLY UNL

SEARCH
 ORACLE
 elasticsearch
 algolia
 Lucidworks
 swiftype
 alphaSense
 omni:s

LOG ANALYTICS
 splunk
 sumologic
 solarwinds
 STIMBER
 Hoba
 logz.io

SOCIAL ANALYTICS
 Hootsuite
 NETBASE
 synthesio
 simple reach
 bitly
 SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
 Google Analytics
 mixpanel
 Airtale
 SIGOPT
 granify
 CUSTORA

APPLICATIONS - ENTERPRISE

SALES
 CHORUS
 INSIDESALES.com peopleai
 conversica
 clari
 Avaso
 tactai
 fusejmachines

MARKETING - B2B
 RADIUS
 App Annie
 EVERSTITCH
 HINTIGO
 sense
 tubular
 JENGAO
 KNOTCH

MARKETING - B2C
 Zeta
 bloomreach
 SendGrid
 braze
 ACTIONIQ
 BLUECORE
 CONTENTADIX
 TEALUM
 Amparo
 amperity
 QUANTIFUNO
 Simon
 Jifika
 PERSADO
 remesh

CUSTOMER EXPERIENCE / SERVICE
 qualtrics
 MEDALLIA
 SurveyMonkey
 CLEARBRIDGE
 zendesk
 Customer
 HEAP
 Amplitude
 Pathon Assistant
 DigitalGenia
 ASAP
 ada
 AUTOMAT
 ahni
 CarDesk
 metomi
 Frame AI

ENTERPRISE PRODUCTIVITY
 slack
 ORACLE
 GURU
 lumiatia
 DIFFBOT
 clara
 talla
 Kasisto

HUMAN CAPITAL
 HaeVee
 pynetics
 Everlytic
 Ailyo
 Textio
 Workday
 entelo
 unicom
 beamy

LEGAL
 RAVEL
 Everlaw
 Lexipol
 JUDICATA
 (FEBREVA)
 INTELLECT
 ROSS
 casetext

REGTECH & COMPLIANCE
 text IQ
 Compliance Advantage
 PARTNERSHIPS
 DATA REPUBLIC

FINANCE
 lunaplan
 ZUORO
 SIVAHANA
 TRADESHIFF
 Scale Factor
 Scale Factor
 Scale Factor
 Scale Factor
 Scale Factor

BACK OFFICE AUTOMATION & RPA
 UiPath
 blueprism
 VIDADO
 Workfusion
 workatko
 ANWORKS
 KRYON
 ALKYRI

SECURITY
 TANIUM
 CYCLANCE
 zscaler
 StackPath
 illumio
 CODE42
 CipherCloud
 DARTTRACE
 ANOMALI
 Trend Micro
 VECTRA
 Guardicore
 DATAVISOR
 sift science
 pinpoint
 exabeam
 SCINIFY
 SentinelOne
 SecurityScorecard
 secure
 CodeSecure
 Bibitglass
 BlueTalon
 Recorded Future
 feedzai
 Cybox
 BITSIGHT
 ABBT
 BLUEHEXAGON
 Semble
 VERACLOUD
 XENONIX
 SHIELD51
 Amorlock

APPLICATIONS - INDUSTRY

ADVERTISING
 AppNexus
 criticon
 MOAT
 Madvertise
 distillery
 TAPR
 dataxu
 sumgumtj

EDUCATION
 Lulluhub
 KNEWTON
 Clever
 Cleara
 kidaptive
 PANDORA
 gradscope

REAL ESTATE
 REDFIN
 Opendoor
 VTS
 CREDIFI
 GEOPIY
 Zillow
 Zillow

GOVT
 OPENDOOR
 mark43
 Quid
 PRIMER
 FORGE

INTELLIGENCE
 Palantir
 Dataminr
 Quid
 PRIMER
 FORGE

FINANCE - INVESTING
 KENSCHE
 Quantopian
 ADEPTAS
 ISENTUM
 ALGORIX
 RavenPack
 PAGAYA

FINANCE - LENDING
 ondeck
 Affirm
 KREDITCO
 AVANT
 TALIA
 FRENCH
 Upstart
 CURE
 CLEARBANC
 100Credit
 LendingClub
 MoneyLion
 cire
 cignif

INSURANCE
 Insomnis
 Jannomade
 CYENCE
 Sligo
 Shift Technology
 ROOT
 ZESTIFY
 CAPE

HEALTHCARE
 flatiron
 Clover
 YORUS
 METABOLIA
 Gingeio
 Glow
 Heartly
 3DMed
 zebra
 PEPHA
 ovia
 TEMPUUS
 patientlyme
 AICure
 insitro
 notable
 entic
 mag
 Recursion
 Qventus
 ARTERYS
 IMAGEN
 INNOVACOR
 PAIGE
 DAYAVAN

LIFE SCIENCES
 color
 Genovis
 Verily
 MINDNEXT
 CODE
 Clear Labs
 FINNANOVA
 PHARMACOR
 CITRINE
 DIVERX
 twoAR
 DEEP GENOMICS
 DOWN

TRANSPORTATION
 UBER
 TESLA
 CLEARPATH
 CRUISE
 drive.ai
 CAMBRIDGE
 DRIVE
 nauto
 AMOTIVE
 G7
 PILOT.AI
 NIO
 OPTIMUS
 moovit
 HILTI
 nexar
 Kodiak
 comma.ai
 netradynae
 CIVE
 MAPS
 cognata
 thinl
 INRIX

AGRICULTURE
 FARMERS
 Granular
 JOHN DEERE
 BLUEBRIER
 FarmersEdge
 AgriStar
 FarmLogs
 TARANIS
 GAMAYA
 Electric
 prospero

COMMERCE
 STITCH
 FAIRF
 DIX & CO
 HEALING
 HEURTELIN
 OTHER
 harmony
 stem
 Amper
 ByteDance
 Lockett
 SOIERN
 BOBEVER
 VERDIGIS
 duoeto
 ZINER
 Splice
 Hovers

INDUSTRIAL
 AVEVA
 SIEMENS
 PREDIX
 UPTAKE
 osi
 SCORTEX
 KONIX
 TACHYUS
 Apache Ranger
 KNOX
 SENTRY
 ANACONDA
 ACCUTULO

OPEN SOURCE

FRAMEWORKS
 TensorFlow
 PyTorch
 Keras
 Caffe
 Microsoft Cognitive Toolkit

QUERY / DATA FLOW
 Spark
 SQL
 presto
 SLIMDATA
 GraphQL

DATA ACCESS & DATABASES
 Cassandra
 mongoDB
 redis
 Cockroach Labs
 Druid
 ScioDB
 ScioDB

ORCHESTRATION & MGMT
 talend
 Apache Airflow
 Mesos
 etcd
 Kong

STREAMING & MESSAGING
 Spark
 Flink
 kafka
 Storm
 Apache RocketMQ

STAT TOOLS & LANGUAGES
 Julia
 Scala
 Studio
 SciPy
 Julia

AI OPS & INFRA
 mflow
 kubeflow
 Seldon
 PyFork

AI / MACHINE LEARNING / DEEP LEARNING
 TensorFlow
 Keras
 PyTorch
 Caffe
 Microsoft Cognitive Toolkit

SEARCH
 elasticsearch
 Solr

LOGGING & MONITORING
 elasticsearch
 kibana
 logstash
 fluentbit
 fluentd
 Grafana

VISUALIZATION
 Tableau
 matplotlib
 seaborn
 Folium

COLLABORATION
 BeakerX
 Jupyter
 Anaconda

SECURITY
 Apache Ranger
 KNOX
 SENTRY
 ACCUTULO

DATA SOURCES & APIs

HEALTH
 Apple
 VALIDIC
 practicefusion
 GARMIN
 HUMANA
 KINOSO
 HUMANA

IOT
 GE Digital
 UPTAKE
 thingworx
 helium
 samsara

FINANCIAL & ECONOMIC DATA
 Bloomberg
 THOMSON REUTERS
 DOW JONES
 S&P CAPITAL IQ
 CB Insights
 PLAD
 SECOND MORTGAGE
 INVESTMENT
 STOCKLE
 Stocktwits
 xignite
 Thinknum
 earnest
 predata

AIR / SPACE / SEA
 Orbital Insight
 planet
 AIRBOTICS
 spire
 UNDERSTANDY
 WINDWARD
 MarineTraffic

PEOPLE / ENTITIES
 axclio
 experian
 EPILION
 InsideView
 CRIMSON HEXAGON
 BASIS
 Quantcast
 SAFEGRAPH

LOCATION INTELLIGENCE
 FOURSQUARE
 Mapbox
 sense360
 PlaceIQ
 esri
 factual
 Mapillary
 StreetView
 cuebiq
 Radar
 OpenStreetMap

OTHER
 DATA GOV
 IMAGENET
 CRUX
 Igrafiti.io

DATA RESOURCES

DATA SERVICES
 OPERA
 DATA SCIENCE
 fractal
 DataKind
 INNOPLCXUS

INCUBATORS & SCHOOLS
 PLURALSIGHT
 DataCamp
 DataElite
 INSIGHT
 The Data Incubator
 METIS

RESEARCH
 facebook research
 OpenAI
 MIRI
 VECTOR INSTITUTE
 ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE