

Introduction to Big Data Systems

CS 448 - Spring 2019

March 18th

Thamir Qadah

Overview

- **Discussion on:**

- *Motivation for Big Data*
- *The MapReduce Model*
- *Hadoop distributed file system*
- *Spark data processing framework*

- **Think-Pair-Share Sessions, given a few discussion question:**

- *2 minutes of thinking*
- *2-4 minutes discuss with partner*
- *2-4 minutes class-wide discussion*

Discussion on Big Data

What are the characteristics of Big Data?

How are they different from traditional database applications?

Why do we need different data management systems for them?

What are the characteristics of Big Data?

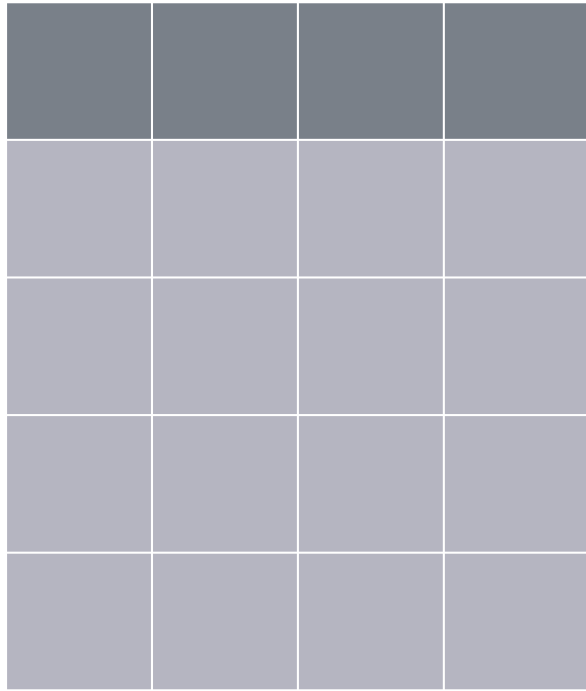
Volume: Size of data

Velocity: Rate of data

Variety: Types of data

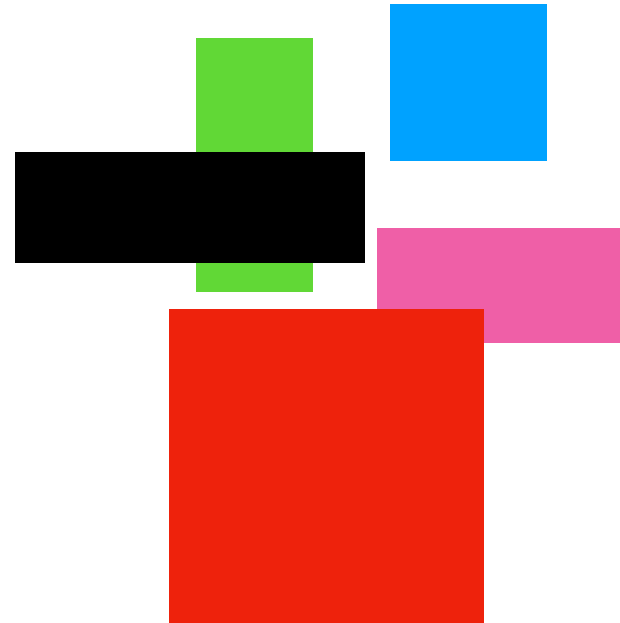
Veracity: Quality of data

How are they different from traditional database applications?



Structured

e.g. Database tables



Semi- or Un-structured

e.g. JSON, XML,
Images, Videos ...

Why do we need different data management systems for Big Data?

Traditional DBMSs require some form of ETL

Not ideal for certain use-cases (e.g.,
Build an inverted index of webpages,
Page-rank of web-pages)

One size **does not** fit all

BIG DATA & AI LANDSCAPE 2018

INFRASTRUCTURE

The diagram illustrates the landscape of data infrastructure, categorized into four main areas:

- HADOOP ON-PREMISE:** Includes logos for cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, and bluescale.
- HADOOP IN THE CLOUD:** Includes logos for AWS, Microsoft Azure, Google Cloud, IBM InfoSphere BigInsights, Treasure Data, Databricks, Altimetric, CAZENA, and CenturyLink.
- STREAMING / IN-MEMORY:** Includes logos for AWS, Databricks, Streamlit, Confluent, GridGain, Oracle, dataArtisans, Hazelcast, TERRACOTTA, KX, FASTDATA, and WallarooLabs.
- NOSQL DATABASES:** Includes logos for Google Cloud, AWS, ORACLE, Microsoft Azure, mongoDB, MarkLogic, CEROPIKE, DATASTAX, ArangoDB, Couchbase, redislabs, and SCYLLA.

DATA TRANSFORMATION

- talend
- pentaho
- alteryx
- TRIFORCE
- tm1
- Paxata
- StreamSets
- UNIFI

DATA INTEGRATION

- SAP Data Services
- Informatica
- TEALium
- spnLogic
- enigma
- podium data
- Segment
- alooma
- xplenty
- ZALONI
- Stitch
- import.io
- infoWorks
- ATTUNITY

DATA GOVERNANCE

- Informatica
- SailPoint
- IBM
- McAfee Stiglych Secure Cloud
- collibra
- Alation
- Waterline Data
- IMHUTA
- OKERA

MGMT / MONITORING

- AWS
- New Relic
- actio
- APDYNAMICS
- rubrik
- WAVEFORM
- dynatrace
- VMWARE
- DATAVOX
- splunk
- SignalFx
- druid
- Moogsoft
- unravel
- pagerduty
- Numeryt
- Acronis

A collage of logos for various cloud storage, cluster services, app development, crowd-sourcing, hardware, and GPU DBs. The logos are arranged in a grid-like fashion, with each category having a header and several logos below it. The categories are: STORAGE (aws, Google cloud, IBM Storage, Microsoft Azure, PURE STORAGE, ALLUXIO, hubblestorage, Qumulo, panteras, COHESITY), CLUSTER SVCS (aws, kubernetes, docker, mesosphere, Core OS, prestodb), APP DEV (lightbend, Keen IO, rainforest, CRK), CROWD-SOURCING (amazonmechanicurk, upwork, figure eight, scale, HIVE), HARDWARE (Google TPUs, arm, intel AI GRAPH, IBM Power Systems, MYTHIC, nvidia, Movidius, WAVE, HAILLO), and GPU DBs (Kinetica, INFINIO, SPORE, blazingsdb, blazingsdb, brytlyt PG-Storm).

ANALYTICS

DATA ANALYST PLATFORMS

- Microsoft
- pentaho
- alteryx
- Digital Reasoning
- guavus
- AYASDI
- ATTIVO
- Datameer
- Quid
- incorta.
- interana
- ClearStory
- Origami logic
- ADSS
- ENDOR
- MODE
- Bottlenose
- switchboard

DATA SCIENCE PLATFORMS

- IBMX
- KNIME
- dataiku
- DOMINO
- rapidminer
- CONTINUUM ANALYTICS
- ALGORITHMIA
- DATAWATCH
- ANGIOS
- SAS

BI PLATFORMS

- Microsoft
- aws
- DOMO
- Wave Analytics
- looker
- THoughtSpot
- ARCADA DATA
- ATSCALE
- Information Builders
- GoodData
- VERISSE
- MicroStrategy
- birst

VISUALIZATION

- tableau
- SAP LUMINA
- Google Cloud
- celonis
- Qlik
- Periscope Data
- ZEPL
- XcodaData
- plotly
- CHARTIO
- TOUGCAN TOCO

MACHINE LEARNING

- Azure Machine Learning
- aws
- Google Cloud
- H2O
- DataRobot
- gamalot
- ELEMENT AI
- VISENZE
- VERSIVE
- deepgenio
- bonsai

The diagram illustrates the AI startup landscape, categorized into three main areas:

- COMPUTER VISION:**
 - Microsoft Azure
 - Amazon Rekognition
 - Cloud Vision API
 - clarifai
 - EVER AI
 - deeppomatic
 - twentybn
 - neurana
- HORIZONTAL AI:**
 - IBM Watson
 - Cortana
 - Face++ 旷视
 - sentient
 - Voyager Labs
 - Microsoft
 - Affective
 - PROPOSEE
 - CognitiveScale
 - Numenta
 - PETUUM
 - SCALED FINANCE
 - naraqi
 - curious AI
 - OSARO
 - BLUE VISION
- SPEECH & NLP:**
 - Google Cloud
 - twilio
 - Amazon Alexa
 - narrative science
 - semantic search
 - WellnessAI
 - Mobvoi
 - Speechmatics
 - PRIMER
 - SoundHound Inc.
 - volica
 - nuance
 - deep mind lab
 - sentral ai
 - snips
 - yoqap
 - Grindge

SEARCH

- elasticsearch
- ORACLE EDGEC
- EXPLORER
- Lucidworks ATTIVO
- swiftype
- alpinahsense
- algolia
- MAANA
- omni:us
- SINEQUA

LOG ANALYTICS

- splunk
- sumologic
- LOGGLY
- TIMBER
- hibana
- logz.io

SOCIAL ANALYTICS

- Hootsuite
- NETBASE
- synthesio
- simplereach
- bitly
- predata
- SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS

- Google Analytics
- mixpanel
- AMPLITUDE
- sumall
- Airtable
- RESOL
- SIGOPT
- granify
- custora

APPLICATIONS – ENTERPRISE

SALES

- einstein
- CHORUS
- INSIDESALES.COM
- conversica
- clari
- aviso
- tact.ai
- fuse|machines
- TROOPS

MARKETING - B2B

- RADIUS
- EVERSTRING
- MINTIGO
- sense
- tubular
- ENGAGIO
- App Annie
- Lattice
- Datafox
- mpo

HUMAN CAPITAL

- HireVue
- entelo
- hiQ
- GIGSTER
- textIQ
- RESTLESS HANDBUILD
- Wade&Wendy
- Stella
- mya
- uncommapp

LEGAL

- RAVEL
- Seal
- Everlaw
- JUDICATA
- IRONCLAD
- PROMOTION
- Ross
- casetext

FINANCE

- naplan
- zuora
- SAS HANA
- TRADESHIFF

ENTERPRISE PRODUCTIVITY

- slack
- ORACLE
- lumiata
- clara
- talla
- butter.ai
- kasisto

BACK OFFICE - AUTOMATION

- UiPath
- hyperScience
- blueprintsri
- electricly
- AppZen
- WorkFusion

APPLICATIONS - ENTERPRISE

- zeta
- bloomreach
- SendGrid
- BlueYonder
- ACTIONIQ
- SAILTHRU
- BLUECORE
- QUANTIFIED
- Impartice
- Amplero
- amperity
- TEALIUM
- Simon
- Lyfycie

CUSTOMER SERVICE

- MEDALLIA
- zendesk
- CLARABRIGHT
- Gainsight
- NGDATA
- DigitalGenius
- afiniti
- AUTOMAT
- frame.ai
- mgma
- INTERCOM
- CalDes

APPLICATIONS – INDUSTRY

ADVERTISING

- AppNexus
- critico
- ORACLE ad services
- TheTradeDesk
- TAPAB
- xAd
- OpenX
- Integral Ad Solutions
- datarena
- Adapthor
- Livestories
- dataxum
- gumgum

EDUCATION

- Livelihood
- KNEWTON
- Clever
- edexra
- kidaptive
- PANORAMA
- KNOWME

GOVERNMENT

- OPENGV
- mark43
- FiscalNote
- GRIDSMAAT
- Livestories
- Passport
- SmartProcure
- STREETLIGHTDATA

REAL ESTATE

- REDFIN
- Opendoor
- VTS
- CREDIF
- economy
- COMPSTAT
- CONF.

FINANCE - INVESTING

- KENSCH
- Dataminr
- Quantopian
- Astara
- SENIUMT
- AlgoSecur
- RiverPack

FINANCE - LENDING

- ONDECK
- Jianpu
- 拍拍贷
- 信用
- Credittech
- AVANT
- UPSTART
- INSIKT
- Skyline
- 100Credit
- Weebao
- Webao
- Weebao
- Monetylon
- credit

INSURANCE

- Metromile
- Lemonade
- CYBERACE
- Shift Technology
- TRACTABLE

HEALTHCARE

- flatiron** **Clover** **KYRUUS** **HealthTap**
- METABIOTA** **Gingerio** **Glow** **oVia**
- 3DME** **zebra** **PathAI** **obvian**
- TEMPUS** **patientsLikeMe** **AICure**
- RECURSION** **prognos** **@enitic** **magiQ**
- Qventus** **BAYLABS** **ARTERIS** **3DOLIMEX**
- IMAGEN** **LongHealth** **PAIGE** **DATAVANT**
- BIOMAT** **DANADIS PREP** **innovacore** **LeanTox**

LIFE SCIENCES

- Exosome** **Color** **Carigene** **bioRxiv**
- Benelabor** **verily**
- WuxiNextCODE** **ZEPHYR HEALTH**
- Novartis** **AMGEN** **Clear Labs** **BIOSPIN**
- freedom** **1HANOPORES**
- DINAVEX** **Phosphorix**
- CITRINE** **two4AR**
- Atomwise** **genomics**
- OWKIN**

TRANSPORTATION

- UBER** **TESLA**
- GO** **CLEARPATH**
- noTonomy** **drive.ai**
- PILOT** **AMANG**
- NAVIO** **NIO**
- PTIMUS** **moovit**
- riemaa** **comax a.i**
- netadryde**
- Civil Maps** **German Autobahn**

AGRICULTURE

- FARMERS** **RESEARCH**
- Granular** **JOHN DEERE**
- BLUE RIVER**
- FarmersEdge**
- FarmLogs**
- TARANIS**
- GAMAYA**
- terramet**
- prospera**

COMMERCE

- Postscript FIX** **Da & C**
- RetailNext**
- HowGood**
- heureka**
- Other**
- eharmony** **stem** **netflix** **Amper**
- ByteDance** **happn** **select**
- BOXYEVOL** **VERODIGOS** **duetto**
- Unbabel** **Juliekick**
- remesh** **ASAPP**

INDUSTRIAL

- SIEMENS** **PREDICI**
- Cyjet** **UPTAK**
- TACHYUS** **SCORTEK**

CROSS-INFRASTRUCTURE/ANALYTICS

CROSS-INDUSTRY/ANALYTICS

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

OPEN SOURCE

The diagram illustrates the popularity of various data science and engineering tools, categorized into 12 segments. Each segment contains logos for tools within that category:

- FRAMEWORK**: Hadoop, Tez, YARN, Flink, Mesos, Spark, CDAP
- QUERY / DATA FLOW**: Spark SQL, Presto, SLAM DATA, Google Cloud Dataflow, Flink
- DATA ACCESS**: Cassandra, Nifi, MongoDB, ScioDB, CouchDB, Riak, HBase, Cloud Spanner, Accumulo
- COORDINATION**: Talend, Apache Toolkeeper, Apache Ambari, Apache Airflow
- STREAMING**: Spark, APEX, Flink, Beam, Kafka, Druid, Storm
- STAT TOOLS**: Python, ScalaLab, SciPy, Julia
- AI / MACHINE LEARNING / DEEP LEARNING**: TensorFlow, Theano, PyTorch, MxNet, Caffe, Microsoft Cognitive Toolkit, OpenAI, DM TK, Keras, PySpark, Apache SINGA, FeatureFu, Mxnet, Chainer, VELES, DIMSUM, neon TM, DSSTNE, milib, DL4J, MAHOUT, Aerosolve
- SEARCH**: Elasticsearch, Solr, Scylla
- LOGGING & MONITORING**: Elasticsearch, Kibana, SENTRY, Logstash, Prometheus
- VISUALIZATION**: BeakerX, Rodeo
- COLLABORATION**: Jupyter, Karpenter, ANA CONDA
- SECURITY**: Apache Ranger, KNOX, Sentry

DATA SOURCES & APIs

DATA SOURCES & APIs

- HEALTH**: Apple, Validic, Practice Fusion, Fitbit, Garmin, Human API, Kinsa
- IOT**: GE Digital, Uptake, Thingworx, Helium, Samsara, Magistry, Estimate
- FINANCIAL & ECONOMIC DATA**: Bloomberg, Thomson Reuters, Dow Jones, S&P Capital IQ, CB Insights, Xignite, Quandl, Investing.com, Premise, Estimate, Second Measure, Eagle Alpha, Stocktwits, Plaid, Thinkium, Earnest
- AIR / SPACE / SEA**: Orbital Insight, Planet, Skytech, Airware, AeroBotics, Spire, Kespri, Microcosm, Understory, Descartes Labs, Windward, Tellus Labs, DroneDeploy, MarineTraffic
- PEOPLE / ENTITIES**: Axiom, Experian, Epsilon, InsideView, Crimson Hexagon, Basis, Quantcast, SafeGraph
- LOCATION INTELLIGENCE**: Foursquare, MapAnything, Mapbox, Sense360, Pluribus, Hexagon, PlaceIQ, Esri, Factual, Carty, Mapillary, Streetline, Cuebiq, Radar
- OTHER**: Qualtrics, Data.gov, Data.world, Enigma, MobileVella, CRUX, HighCharts

DATA RESOURCES

Discussion on MapReduce

What are the main pieces of logic a programmer needs to specify?

What are the benefits of the MapReduce and Hadoop?

What are the main pieces of logic a programmer needs to specify?

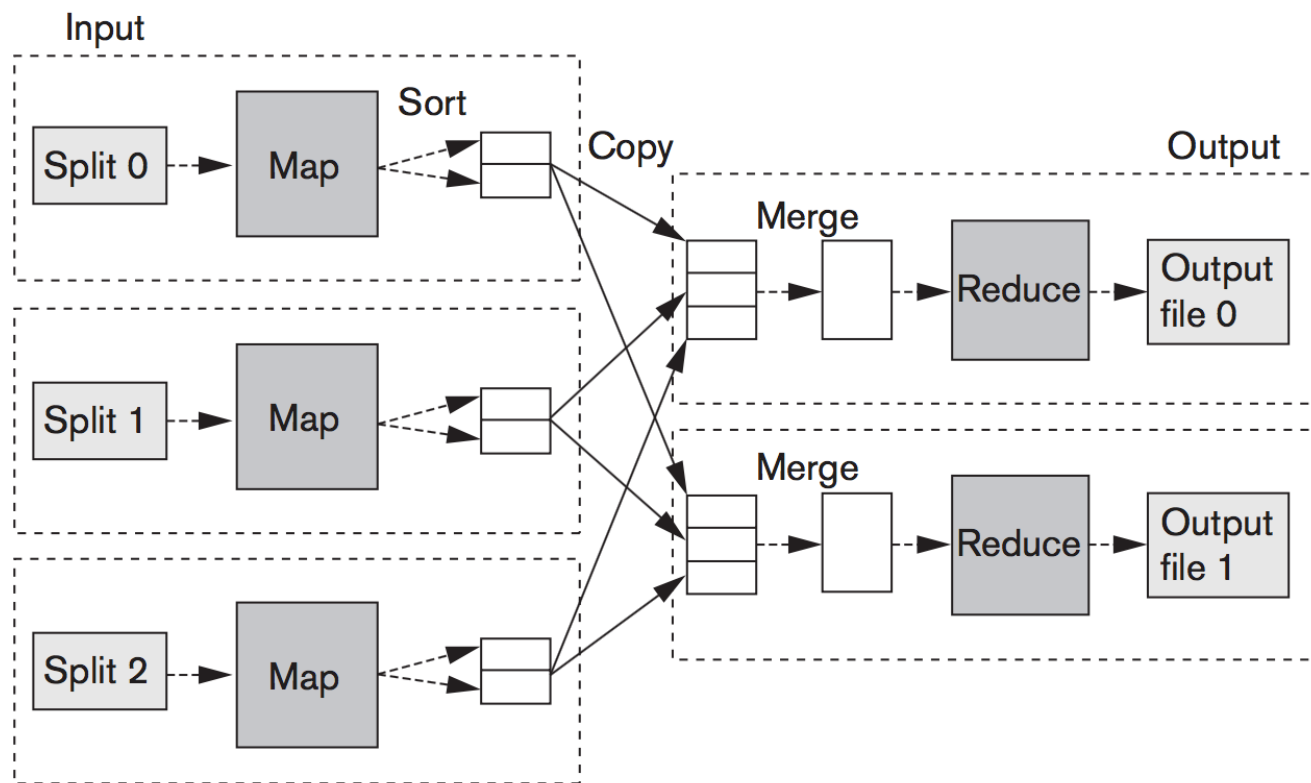


Figure 25.1
Overview of MapReduce execution. (Adapted from T. White, 2012)

MapReduce Model

`map(K1,V1) : List[K2,V2]`

`reduce(K2, List[V2]) : List[K3,V3]`

MapReduce Example

```
map[LongWritable,Text](key, value) : List[Text, LongWritable] = {  
  String[] words = split(value)  
  for(word : words) {  
    context.out(Text(word), LongWritable(1))  
  }  
}  
reduce[Text, Iterable[LongWritable]](key, values) : List[Text, LongWritable] = {  
  LongWritable c = 0  
  for( v : values) {  
    c += v  
  }  
  context.out(key,c)  
}
```

What does this code compute?

What are the benefits of the MapReduce and Hadoop?

Simple distributed programming

Allows for highly parallel and distributed and reliable data processing

Free and open source

Discussion on HDFS

What are the design goals for HDFS?

What are the main architectural components of HDFS?

What are the design goals for HDFS?

Fault-tolerance

Throughput-optimized

Support for large files

Append-only data write model

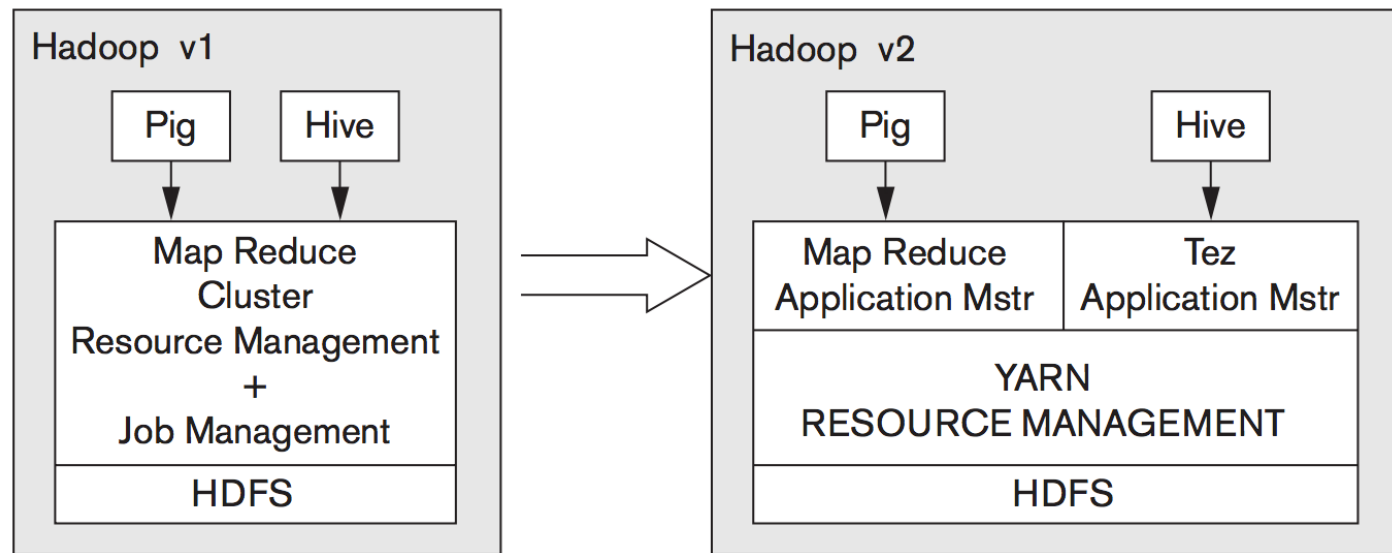
What are the main architectural components of HDFS?

Name Node (+ secondary)

Data Nodes

Discussion on YARN

Figure 25.3
The Hadoop v1 vs.
Hadoop v2
schematic.



What is the key concept behind YARN?

What are the benefits?

Discussion on YARN

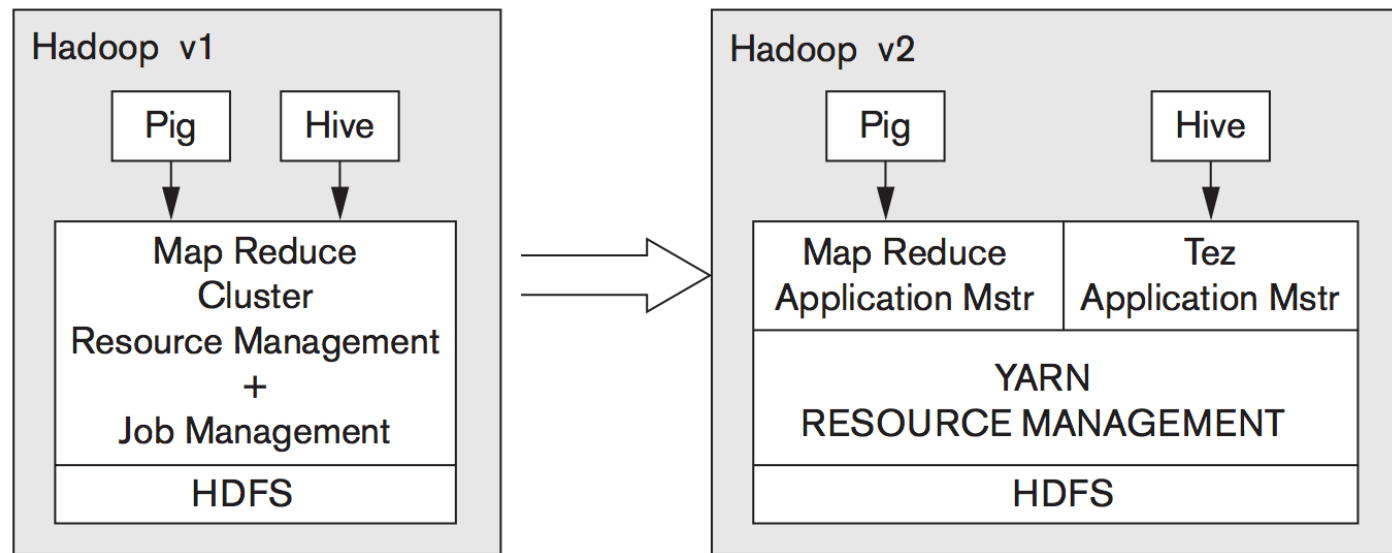


Figure 25.3

The Hadoop v1 vs. Hadoop v2 schematic.

Separation of Concerns

Improved resource utilization

Allow other applications to run on cluster



Zaharia et al. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. NSDI 2012

Shi et al. Clash of the Titans: MapReduce vs. Spark for Large Scale Data Analytics, VLDB 2015



What are the elements of the vision behind Spark?

What is the key feature introduced in Spark 2.0?



What are the elements of the vision behind Spark?

Functional High-level API to support data scientists workflows

Unified data processing

What is the key feature introduced in Spark 2.0?

Structured APIs

What technology is better?

	Parallel Databases	MapReduce
Structured Data		
Unstructured Data		
Fault-tolerance		
Query Expressiveness		
Simple Usage		
Support for Novel Applications		

Project 4

Use a real cluster environment (RCAC Scholar)

Practice with HDFS

Practice with Spark and Spark-SQL
(possibly Spark-Streaming too!)