Computing Big Data

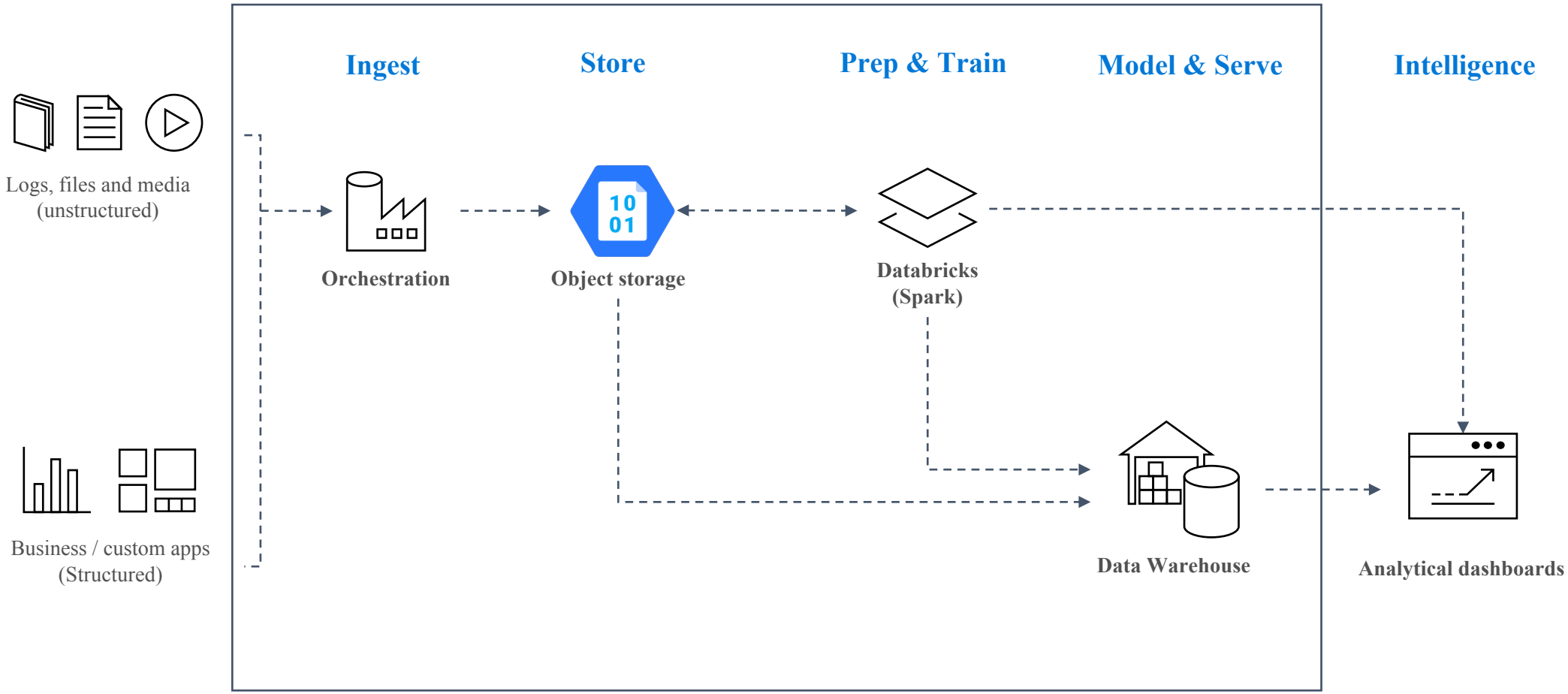
Meirc Training & Consulting

PLUS SPECIALTY TRAINING

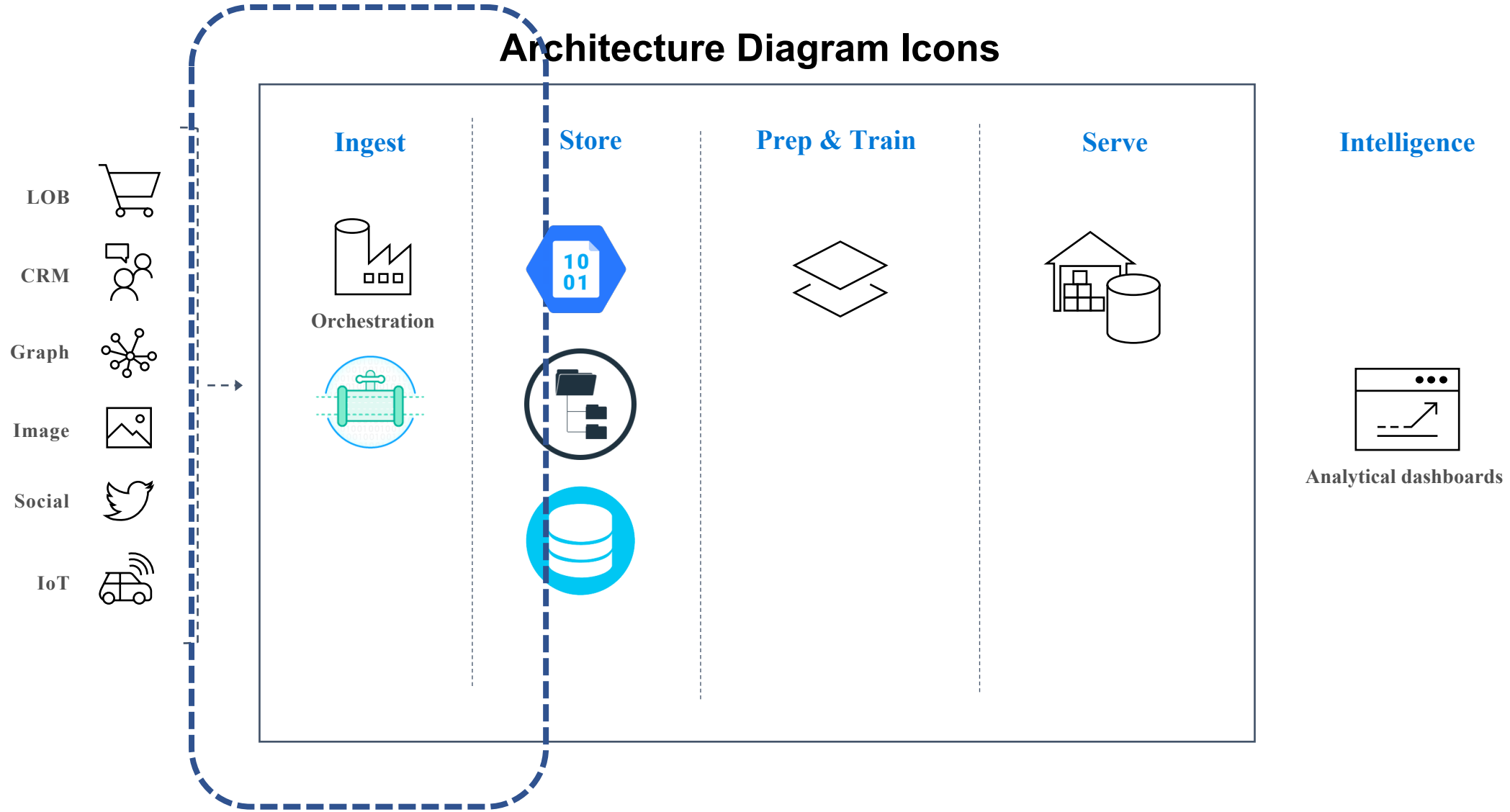
Computing Big Data

- Accessing Big Data
- Role of Cloud Platforms
- Big Data ETL
- Big Data Compute
  - Considerations
  - Technologies
- Hadoop MapReduce
- Hadoop Relevance
- Distributed Compute
  - Apache Spark
  - SAS CAS
- Other Technologies
- High Performance Compute
- Hands-On-Lab

Architecture Diagram Example



Architecture Diagram Icons



Workflow Orchestration

Characteristics

Consistent job execution

Data pipelines

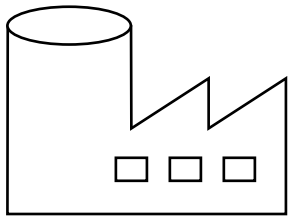
Automating processes

Script management

Notes

Helps monitor and maintain multiple systems for data access and processing.
Graphical interfaces.
Manage interactions and connections of applications.
Scale and pause jobs in production.
DAGs

Icons

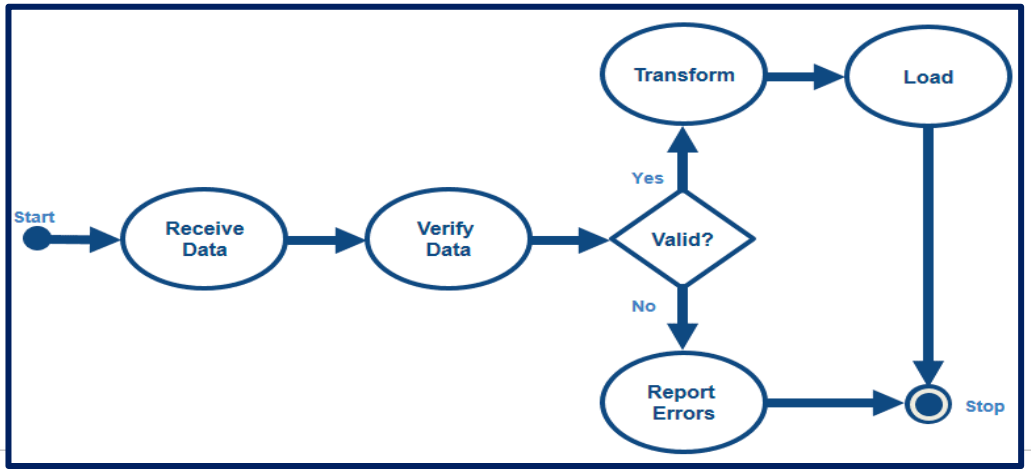


Airflow

AWS Data Pipeline

$

Cost



Streaming Queue

Characteristics

Ordered storage and delivery

Message persistence and queue

Pub/Sub methods

Data buffer

Notes

Consumer and producer queue definitions
Conditional message routing
message rejection and resending.
High read, low write.

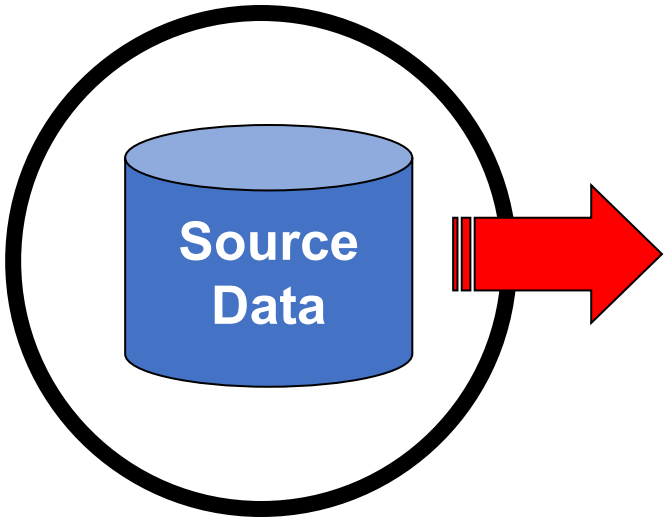
Icons

Cost

$$

Types of Data

Big Data Access Considerations



Considerations

- Asynchronous requests
- Throughput
- Security and Permissions
- Evolving systems
- Expanding pipelines

What Do We Need?

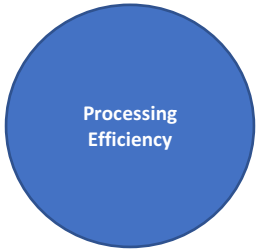
Data Availability & Reliability

Processing Efficiency

Lower Costs

Emergence of Cloud Compute

Mathew Lodge, Anaconda's senior vice president of products and marketing, pointed out in a story on VentureBeat how the center of the big data universe shifted away from Hadoop to the cloud, where storing data in object storage system like Amazon's S3, Microsoft Azure Blob Storage, and Google Cloud Storage is five times cheaper than storing it on HDFS.



Geo-replication of data

PaaS DB offerings

Access to modern hardware

Pay for elasticity

Scale-up and out on-demand

Rent vs Own hardware

Internal management vs services

Cloud Platforms

Market Share:
55%

Pricing:
Per Hour

Offerings:
Compute
Storage
Database
Networking & Content
Development Tools
Security
Analytics

Market Share:
35%

Pricing:
Per Minute

Offerings:
Compute
Storage
Database
Networking & Content
Development Tools
Security
Analytics

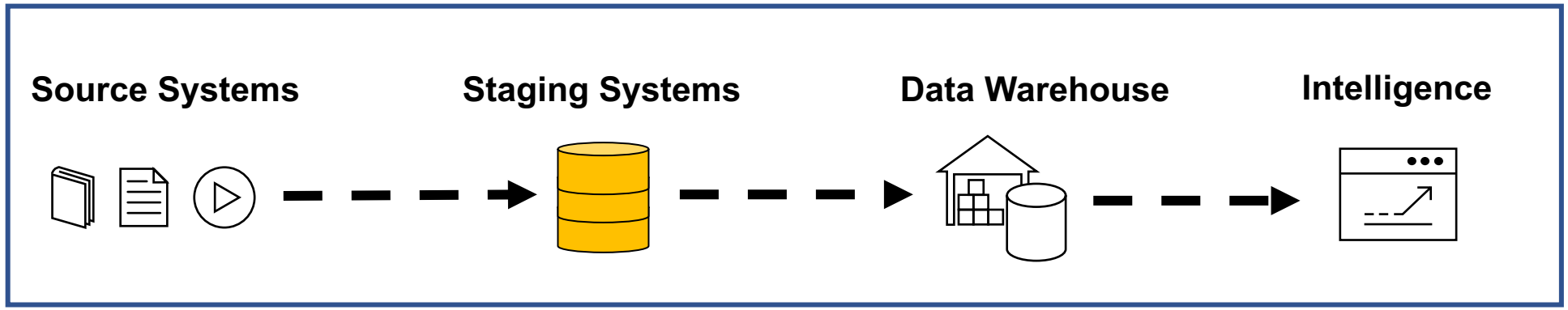
Market Share:
8%

Pricing:
Per Minute

Offerings:
Compute
Storage
Database
Networking & Content
Development Tools
Security
Analytics

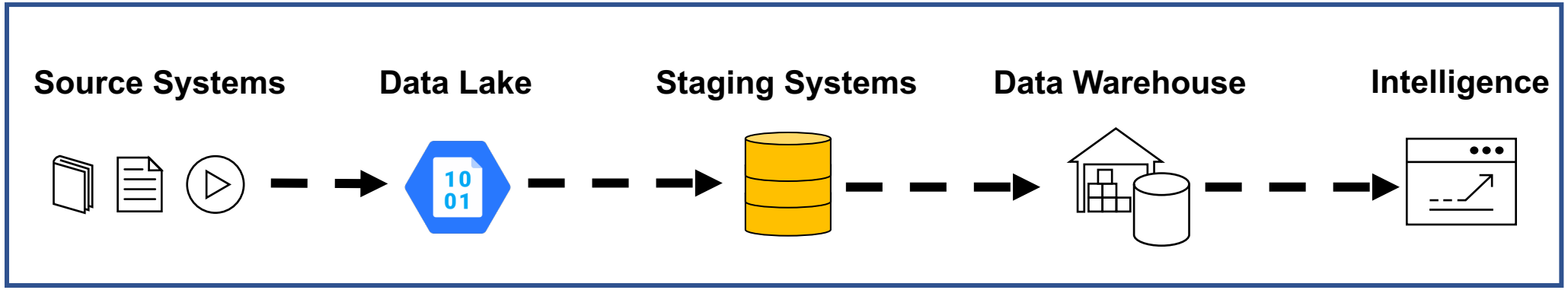
Bigger loads, longer jobs

Traditional ETL



More Jobs, smaller loads

Big Data ETL



Compute Considerations

- Storage Impacts Compute
- Security
- ML & AI Are Iterative
- Compute Is Memory Intensive
- Networking
- Staff Resources

Big Data Compute Technologies

Compute Technologies

Spark

Map Reduce

SAS

Hive

DBMS

Storm

Hadoop Relevance

Cloudera and Hortonworks Merge

datanami

DATA SCIENCE • AI • ADVANCED ANALYTICS

Home About Resources Events Subscribe

HOME | FEATURES | SECTORS | APPLICATIONS | TECHNOLOGIES | VENDORS

Top Stories On

October 18, 2018

Is Hadoop Officially Dead?

Alex Woodie

(Christos-Georghiou/Shutterstock)

The merger of Cloudera and Hortonworks was applauded by many people in the big data community, and even Wall Street liked the news initially. But as the confetti from the party clears some are asking tough questions, like whether the merger signals the death of Hadoop as a viable computer platform moving forward. The answer is probably not. Here's why.

The October 3 announcement that Hortonworks will join forces with its arch-rival Cloudera to create a single company with about $730 million in annual revenue, 2,500 customers, and a $5.2 billion market valuation took a lot of people by surprise.

"My first reaction was I think it's good news for people like us," says Splice Machine CEO Monte Zweben, who had a front-row seat to the first dot-com boom and bust. "We have seen a tremendous opportunity that we didn't see two years ago to operationalize all of the data lakes that those two companies and others have successfully deployed."

CRAY THE SUPERCOMPUTER COMPANY

tellius

Mellanox

(Christos-Georghiou/Shutterstock)

The October 3 announcement that Hortonworks will join forces with its arch-rival Cloudera to create a single company with about $730 million in annual revenue, 2,500 customers, and a $5.2 billion market valuation took a lot of people by surprise.

"My first reaction was I think it's good news for people like us," says Splice Machine CEO Monte Zweben, who had a front-row seat to the first dot-com boom and bust. "We have seen a tremendous opportunity that we didn't see two years ago to operationalize all of the data lakes that those two companies and others have successfully deployed."

"It was a smart move," Jay Kreps, the CEO of Confluent and co-creator of Apache Kafka told ZDNet. "These were two companies competing on the same product, which makes the competition more fierce, ironically."

"I think it's all good," says Kunal Agarwal, the CEO of Unravel Data. "I feel both of these companies putting their technology profiles together, instead of trying to beat each other and create two of everything, they can now focus on providing the right machine learning tool, providing the right IoT platform, providing the right AI tooling."

But not all the comments have been positive. "I am weary about whether the new Cloudera (or even Cloudera and Hortonworks individually) will grow as quickly as its management team and investors expect, which is why I am staying away for now," Virginia Backaitis, a freelance tech reporter, wrote in a Seeking Alpha piece.

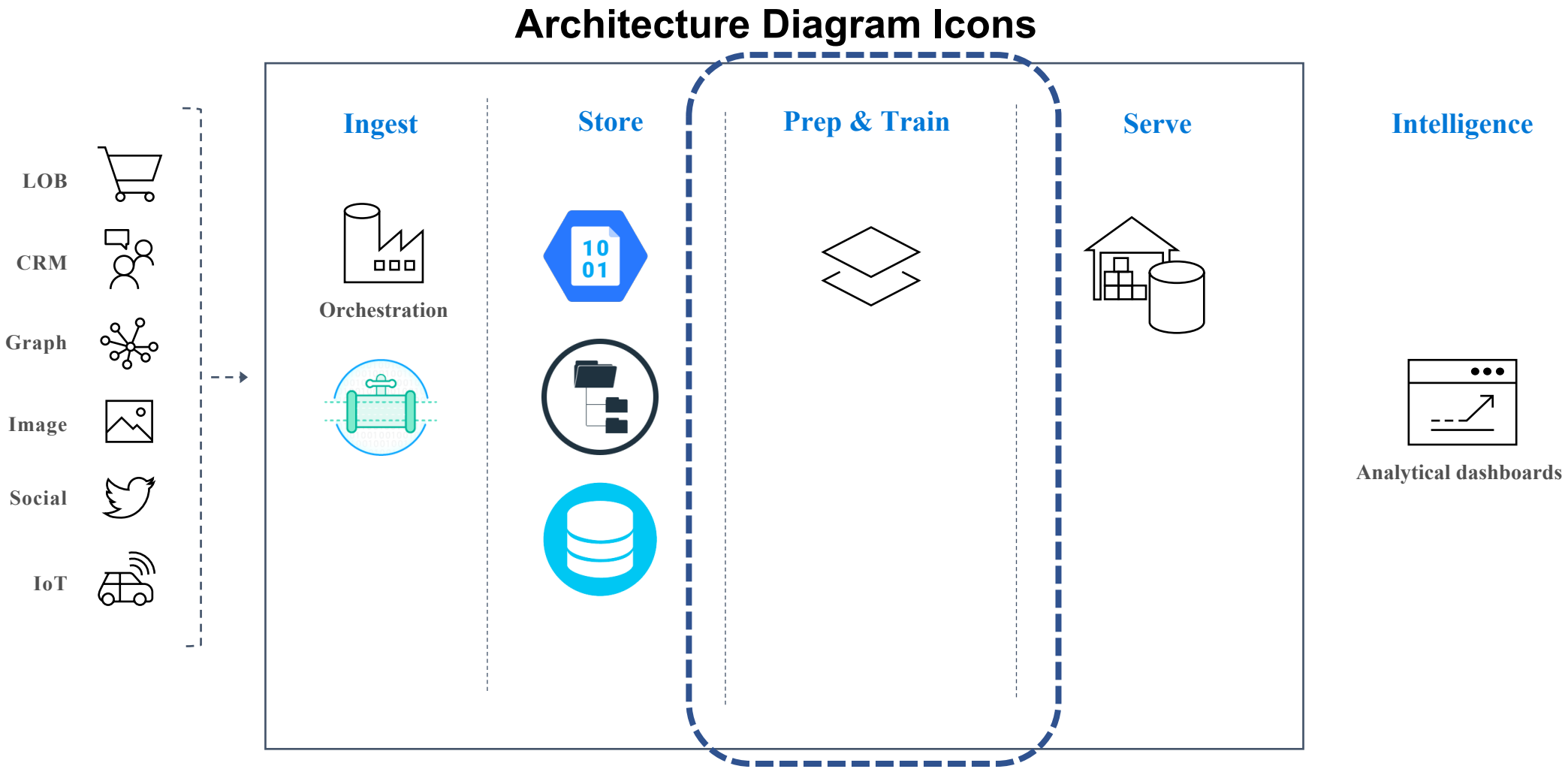
Bloomberg Opinion columnist Shira Ovide was no less derisive, calling the merger of the two perpetually unprofitable tech providers "a seafaring union of two underwater companies."

"It's the Sears-K-Mart merger," Teradata COO Oliver Ratzesberger told *Datanami* this week at the vendor's user conference. "It's the only way for them to potentially survive this. Hadoop in itself has become irrelevant."

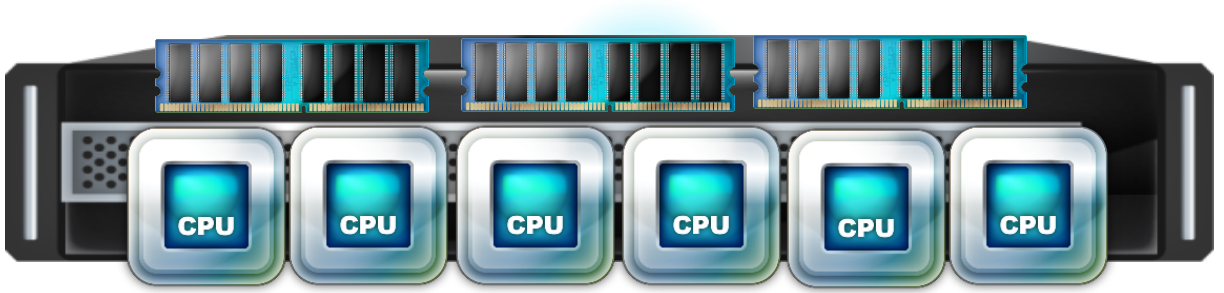
Mathew Lodge, Anaconda's senior vice president of products and marketing, pointed out in a story on VentureBeat how the center of the big data universe shifted away from Hadoop to the cloud, where storing data in object storage system like Amazon's S3, Microsoft Azure Blob Storage, and Google Cloud Storage is five times cheaper than storing it on HDFS.

"The leading cloud companies don't run large Hadoop/Spark clusters from Cloudera and Hortonworks – they run distributed cloud-scale databases and applications on top of container infrastructure," Lodge wrote. "It's time for the Hadoop and Spark world to move with the times."

https://www.datanami.com/2018/10/18/is-hadoop-officially-dead/


Architecture Diagram Icons
LOB
CRM
Graph
Image
Social
IoT
Ingest
Store
Prep & Train
Serve
Intelligence
Orchestration
Analytical dashboards

Workstation



Server

… how does this scale with data?

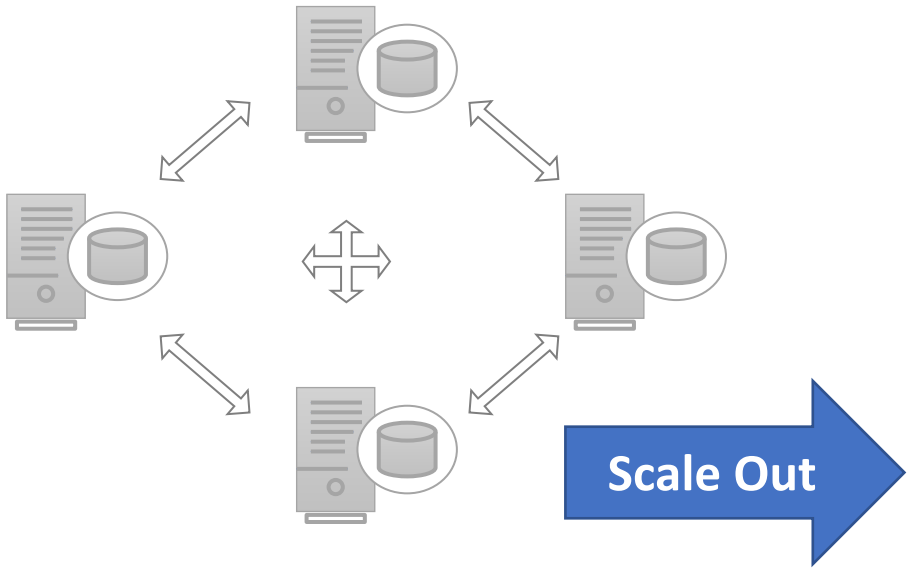
SMP vs MPP

SMP
↓
Symmetric **M**ulti-**P**rocessing

System where processors share resources

Scale Up = larger server

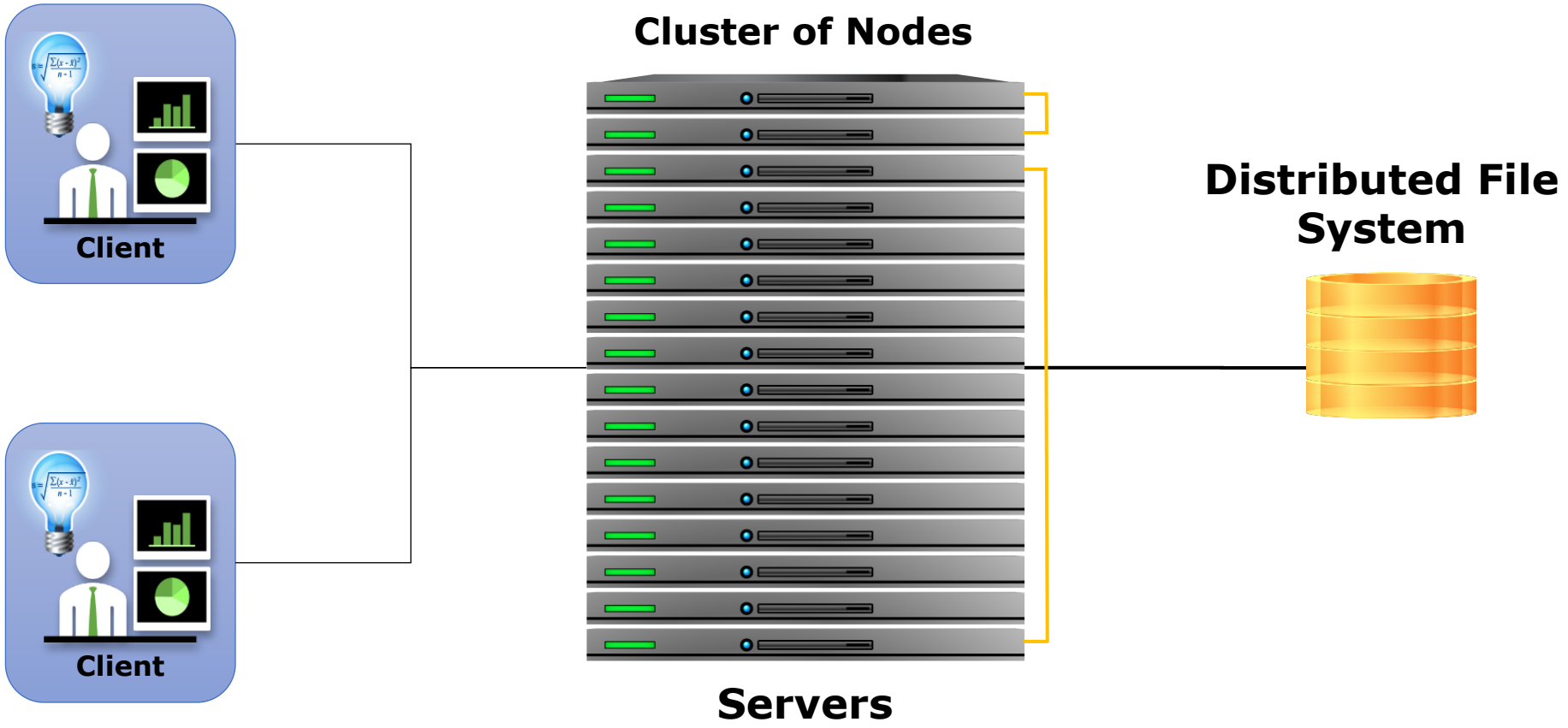
MPP
↓
Massively **P**arallel **P**rocessing



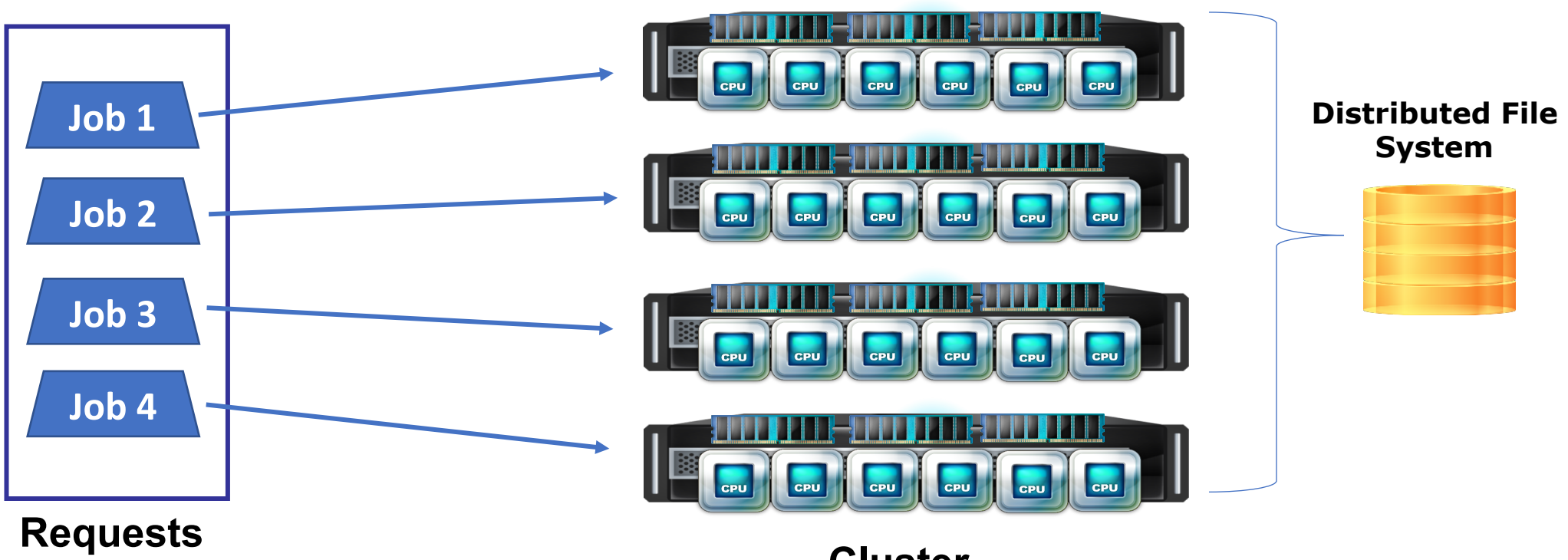
System of processors that **share nothing**

Scale Out = add more servers

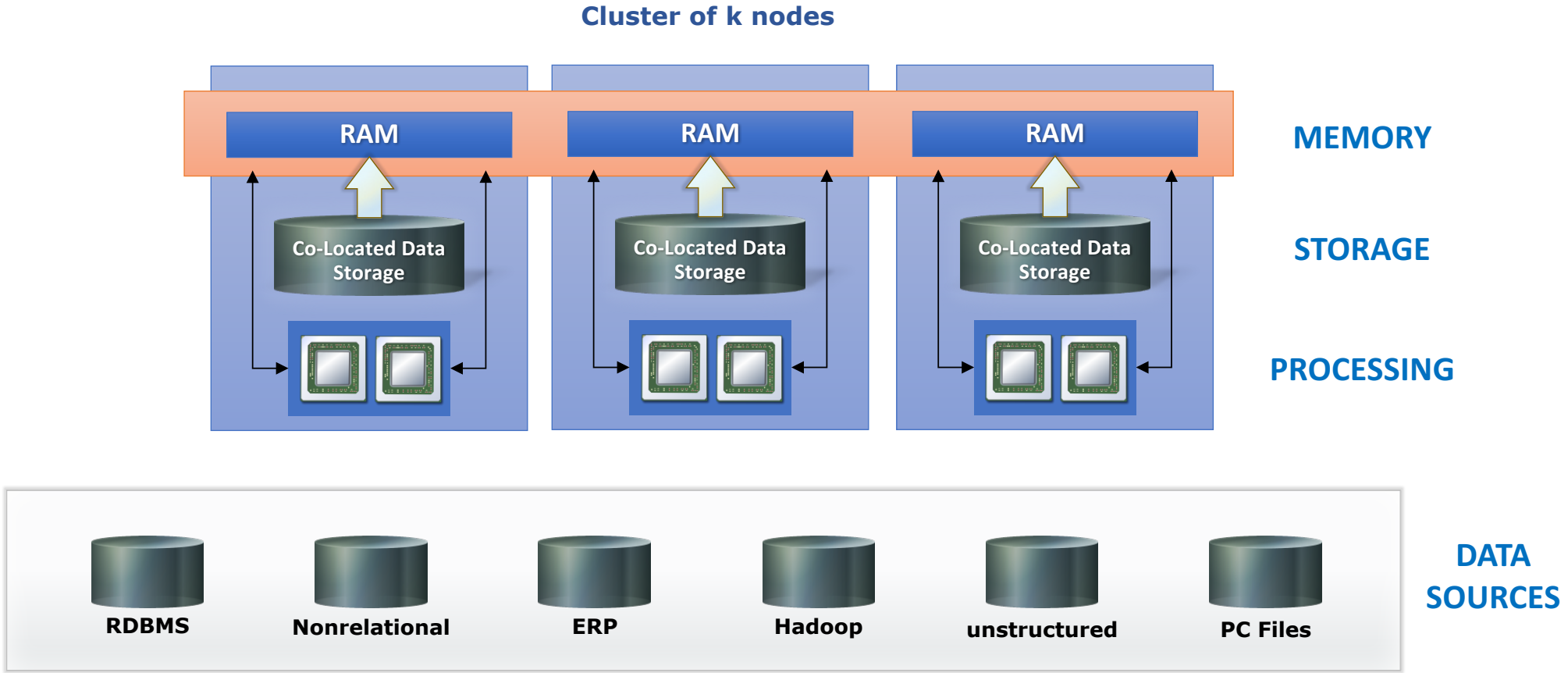
Distributed System



Parallelize Jobs



Distributed In-Memory



Characteristics

Distributed Processing

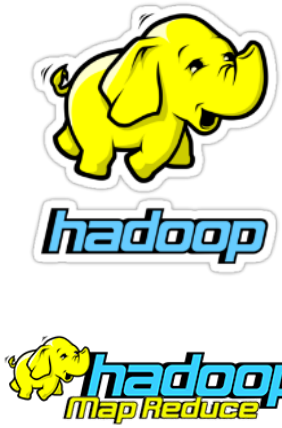
Open sourced technology

Map + Reduce

Notes

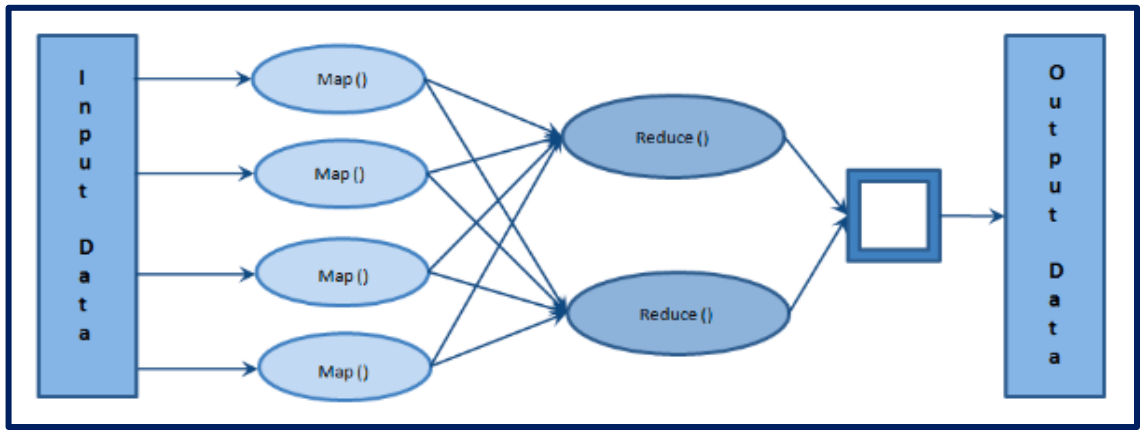
Very inefficient and slow.
Map = find blocks
Reduce = perform computation on blocks.
Original computation paradigm of Hadoop.
Heavy I/O.

Icons



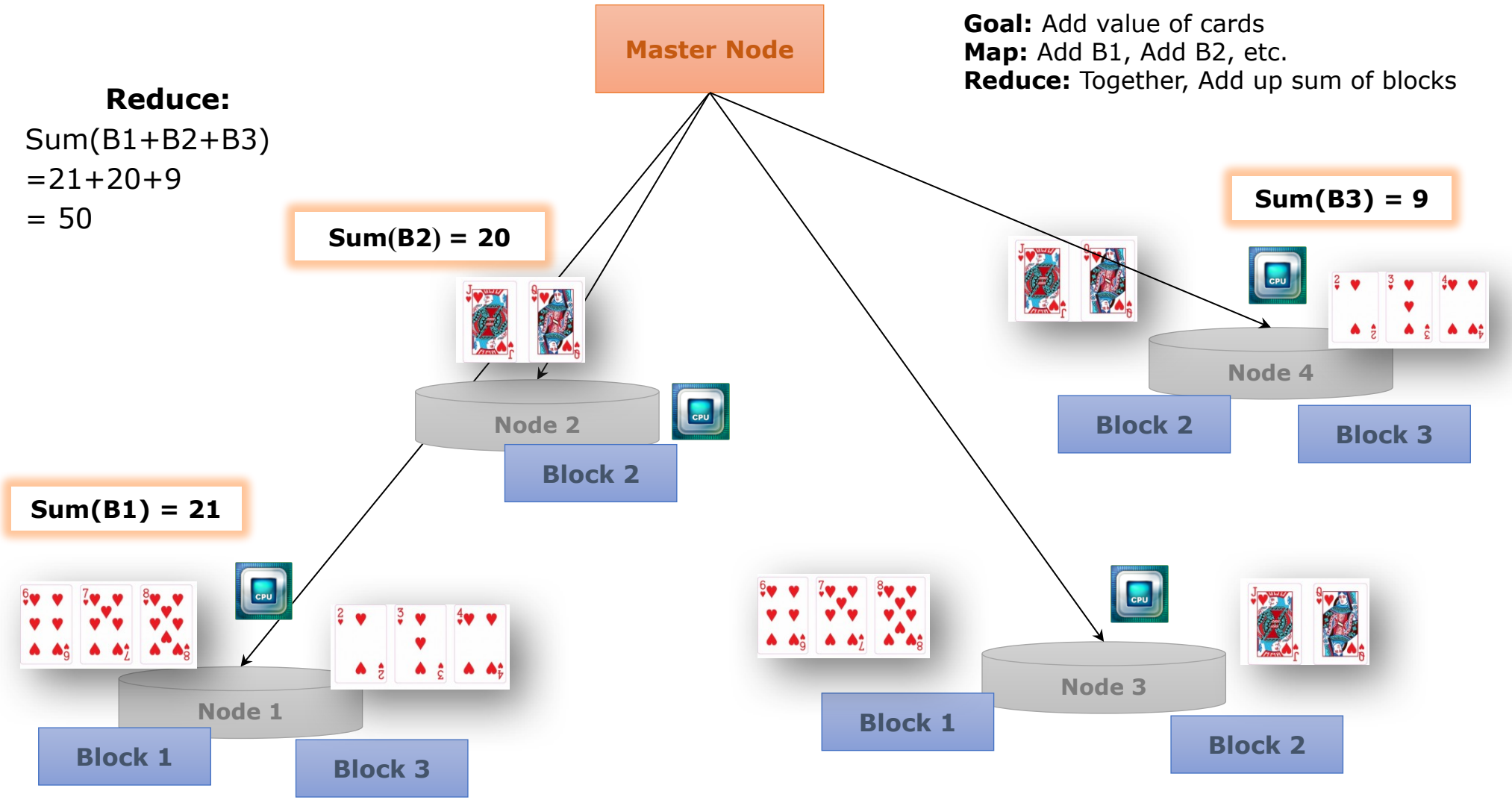
~$

Cost



Goal: Add value of cards
Map: Add B1, Add B2, etc.
Reduce: Together, Add up sum of blocks

Reduce:
Sum(B1+B2+B3)
=21+20+9
= 50



Apache Spark

Apache Spark

Characteristics

Distributed In-memory Compute

Open sourced technology

Map + Reduce

Notes

Very efficient for large datasets.
Lazy execution.
Memory intensive.
Heavy caching.
~100x > MapReduce.
Core facilities: SQL, ML, Streaming and Graph.
Great developer community.

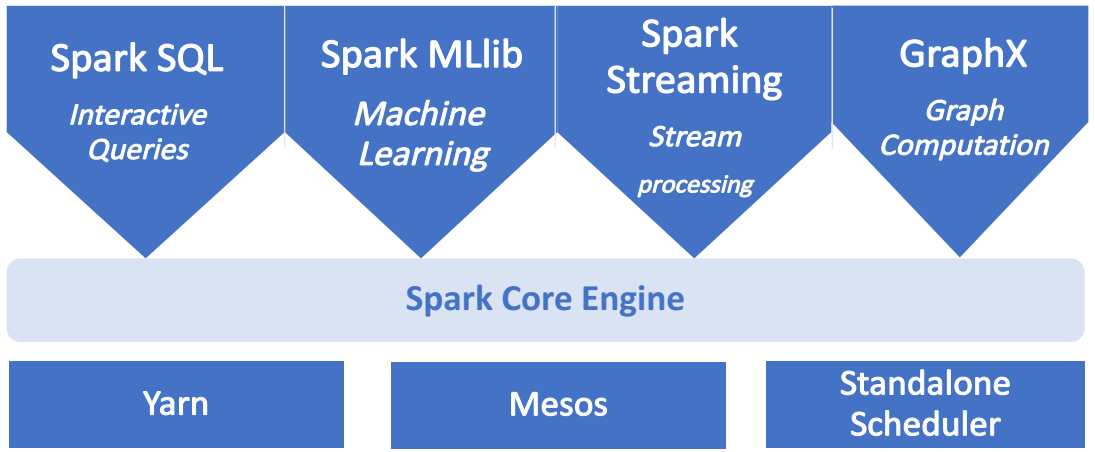
Icons

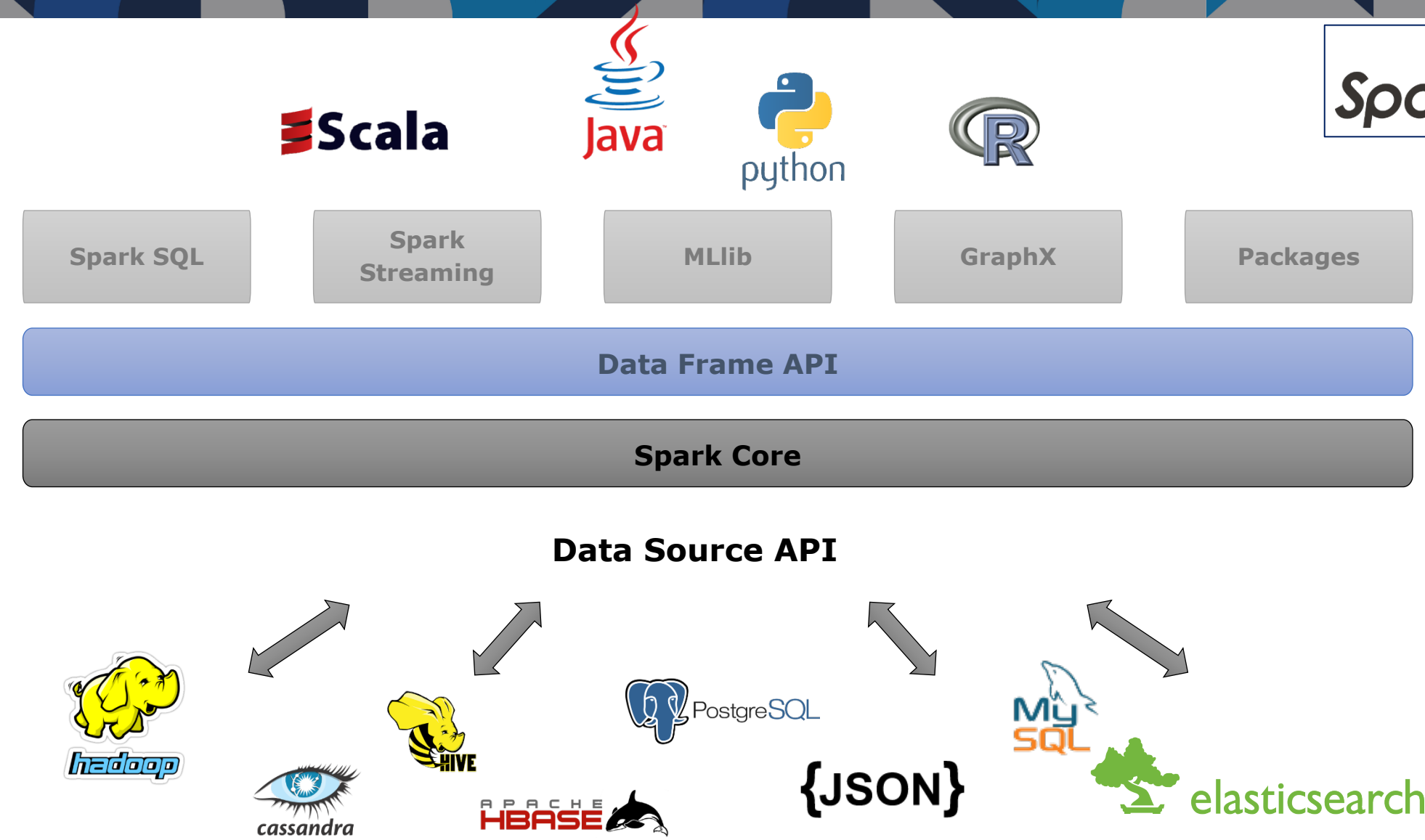
Spark

databricks

$$

Cost




Scala
Java
python
R
Spark
Spark SQL
Spark Streaming
MLlib
GraphX
Packages
Data Frame API
Spark Core
Data Source API
hadoop
cassandra
HIVE
APACHE HBASE
PostgreSQL
{JSON}
MySQL
elasticsearch

IBM Industries & solutions Services Products Support & downloads My IBM

News room > News releases >

IBM Announces Major Commitment to Advance Apache®Spark™, Calling it Potentially the Most Significant Open Source Project of the Next Decade

IBM Joins Spark Community, Plans to Educate More Than 1 Million Data Scientists

Investment of $300 million dollars

News room
News releases
Press kits
Image gallery
Biographies
Background
News room feeds
Global news rooms
News room search
Media contacts

Related links
- IT Analyst support center
- Investor relations

Select a topic or item
- News release
- Related XML feeds
- Contact(s) information
- Related resources

ARMONK, NY - 15 Jun 2015: IBM (NYSE:IBM) today announced a major commitment to Apache®Spark™, potentially the most important new open source project in a decade that is being defined by data. At the core of this commitment, IBM plans to embed Spark into its industry-leading Analytics and Commerce platforms, and to offer Spark as a service on IBM Cloud. IBM will also put more than 3,500 IBM researchers and developers to work on Spark-related projects at more than a dozen labs worldwide; donate its breakthrough IBM SystemML machine learning technology to the Spark open source ecosystem; and educate more than one million data scientists and data engineers on Spark.

IBM News Room Twitter
Join the conversation

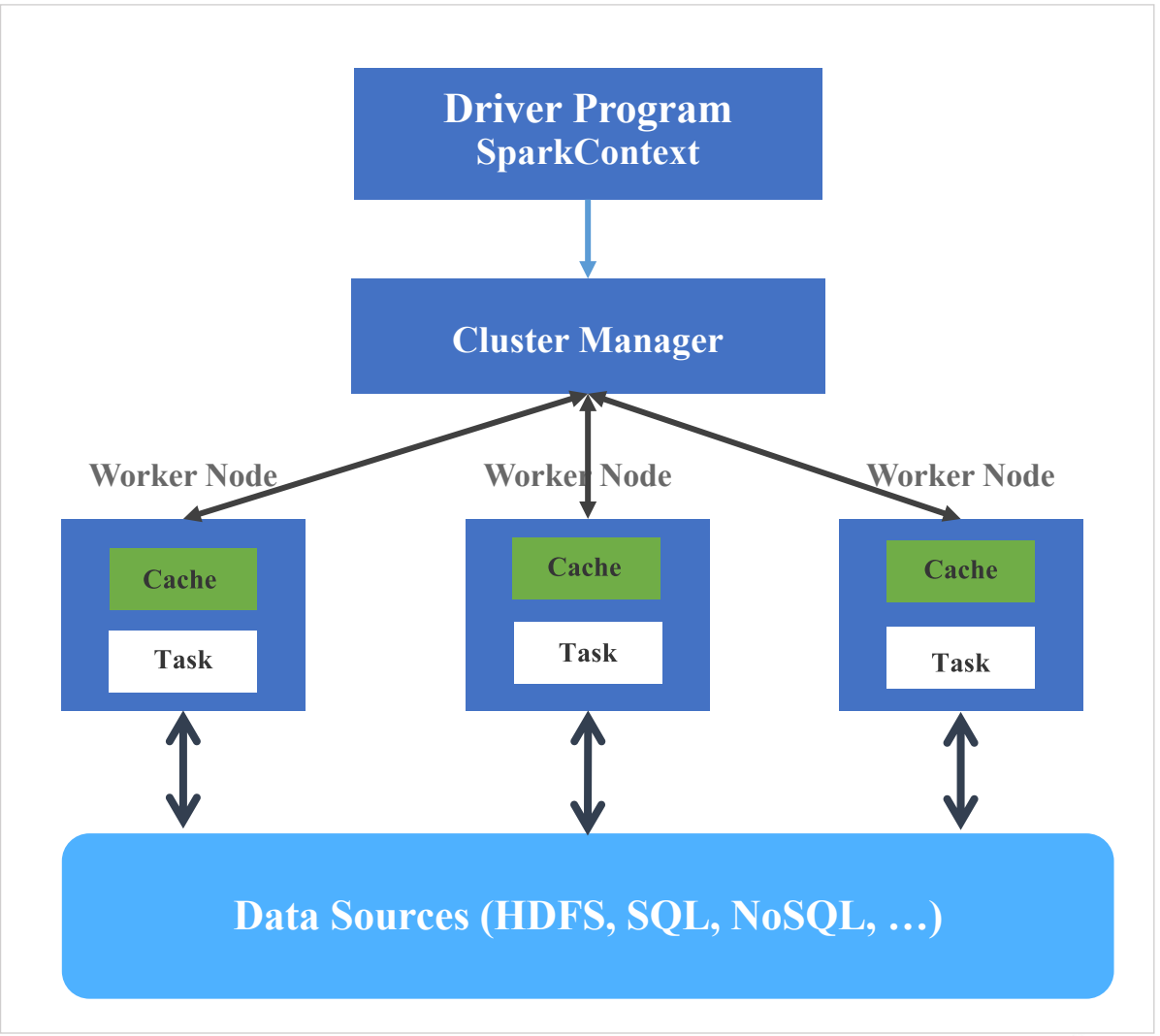
Share
Facebook
E-mail this page
Twitter
LinkedIn

Driver executes the ops on the worker nodes.

Worker nodes read and right data from/to data sources.

Nodes execute as VMs in public clouds.

Results collected by the house.



Spark Use Case Matrix

	Title	Cloud ready	Data Systems	Data	Hadoop / Spark Expertise	Use Case	Time	Pain	Programing Languages:
Great	C-Level, Director, Data Scientist, Data Engineer	Yes - Most data in Azure	Blob Storage, Azure Datalake Event/IoT Hubs, Kafka	1TB +	In Production with Spark	ETL, Ad-hoc Analysis, Machine Learning, Streaming, Data Exploration	ASAP	Defined project, Developing big data strategy, Company-wide deployment	Python, Scala, SQL, R
Good	Lower level Manager, Data Analyst	Yes- Moving to Azure from on prem or other cloud	Cloudera, HDI, CosmosDB, SQL DW, HDFS, MongoDB, Cassandra, Azure DB, SQL Server	100s GB - 1TB	Mostly Hadoop (Hive, Pig, MapR). Moving to Spark	Dashboards, Reports, Text Mining NLP, Deep Learning	1-6 months	Data getting too big for existing tools / ML Team project only, Looking impact to the business	Java
Okay	Other	No - On prem & Evaluating Azure	On-prem sources, Migrating from other clouds (Big Query, GCS, Dataproc, Redshift, S3, EMR, Kinesis)	10GB-100GB	New to Hadoop or Spark, Learning	None	6-12+ months	Personal help, Learning Spark, Science Project	Others
Effort	N/A	No - Only on-prem, not moving	N/A	<10GB (and not expected to grow)	N/A	None	None	None	N/A

SAS Viya (CAS)

Characteristics

Distributed In-memory Compute

SAS R&D technology

Memory persistence

Notes

Very efficient for psuedo-large datasets.
Memory intensive.
~100x > MapReduce.
Core facilities: SAS Procs, ML, Streaming
New SAS CASL language and SAS BASE compatible.

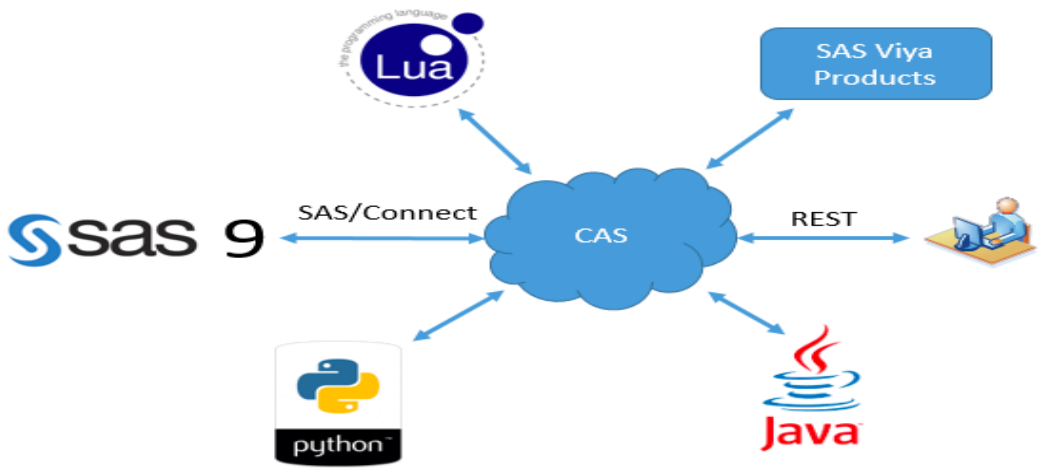
Icons

SAS Viya

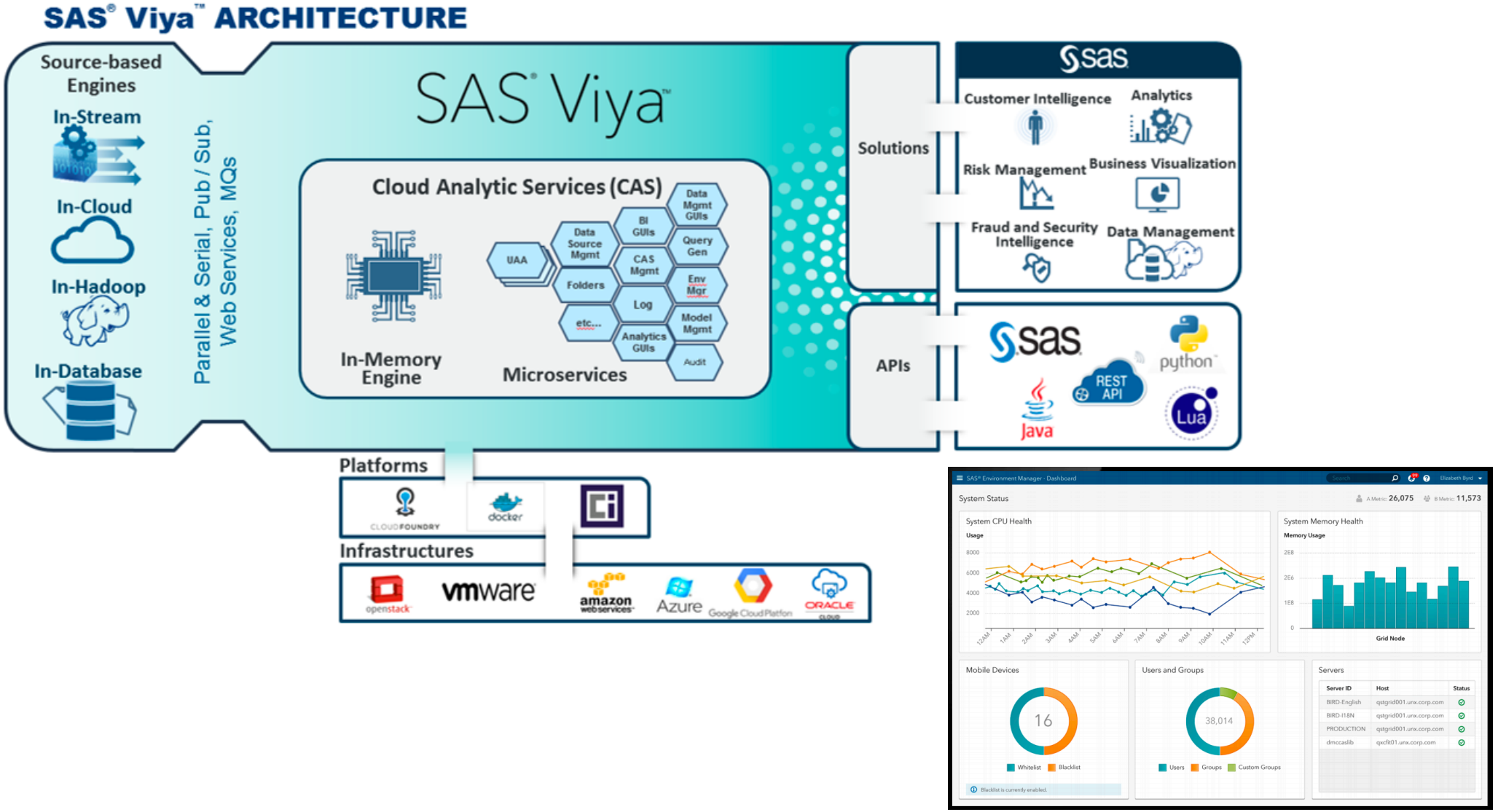
sas

$$$

Cost

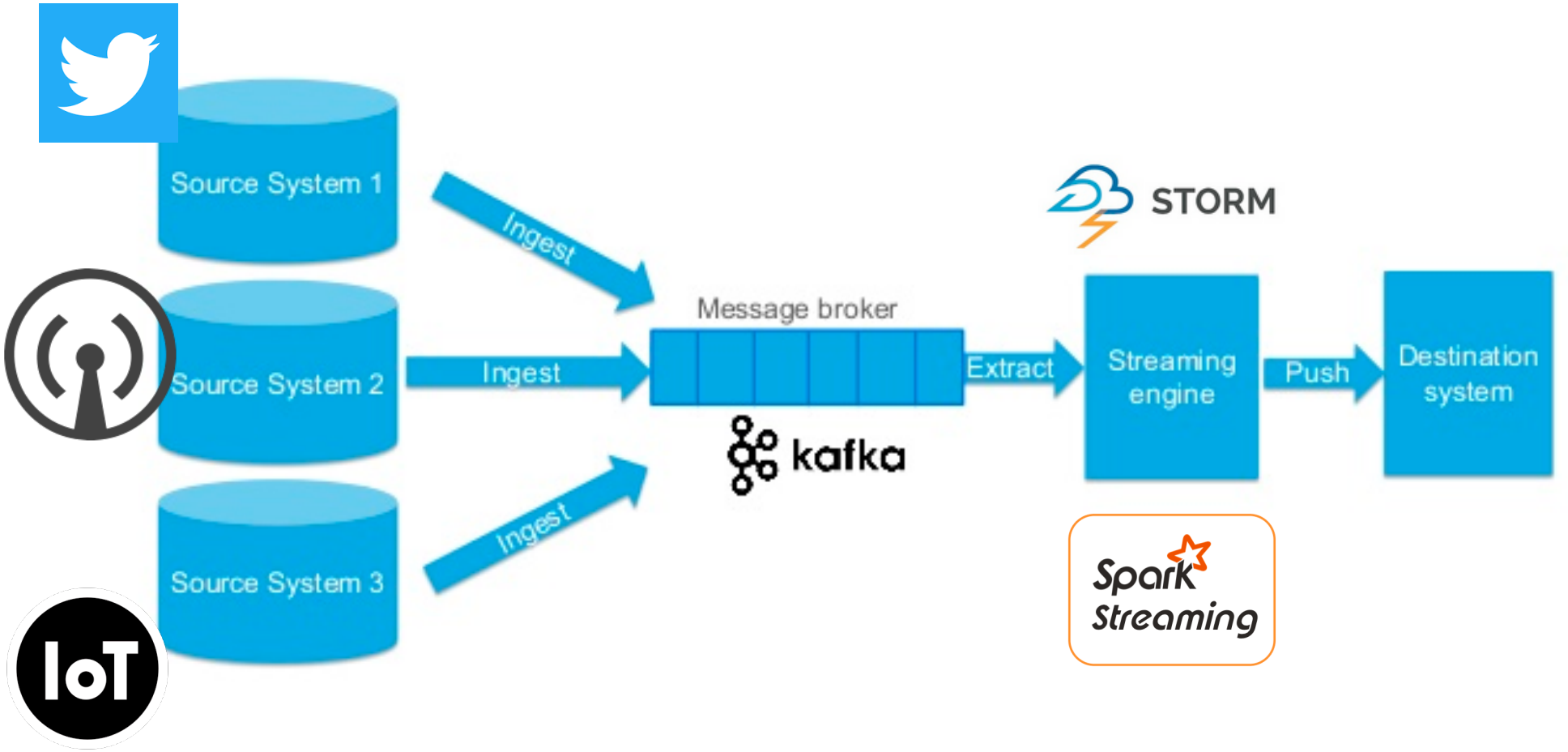


SAS Viya Architecture



Streaming Compute

Streaming Data Flow



Characteristics

Open source, distributed and decentralized

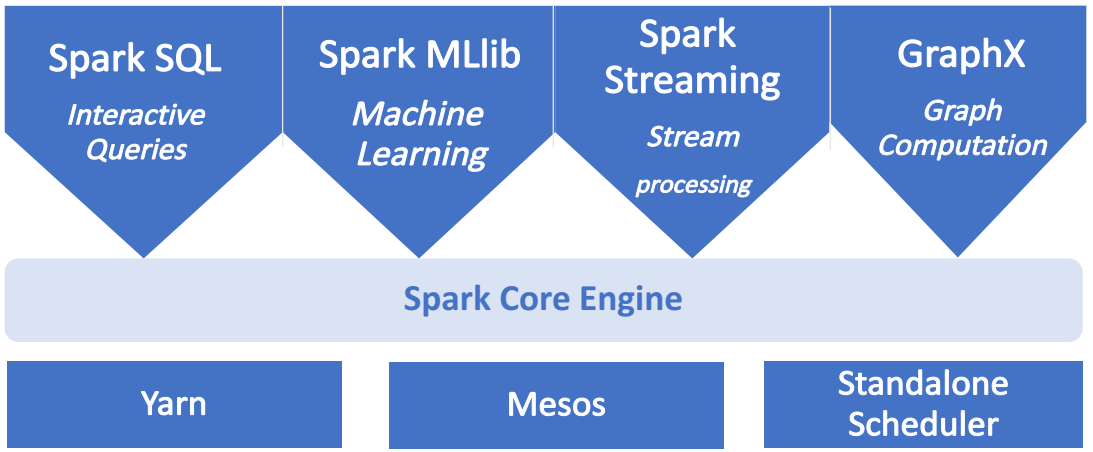
Column-oriented database

Persistent storage

Notes

Same Spark architecture
Low latency.
micro-batch streaming.
Scale on demand.
Compatible with cloud storage platforms.
Handles petabytes of data.
Low development cost.

Icons



Apache Storm

Characteristics

Open source

Distributed stream processing

Clojure programming language

Integrates with queueing systems

Notes

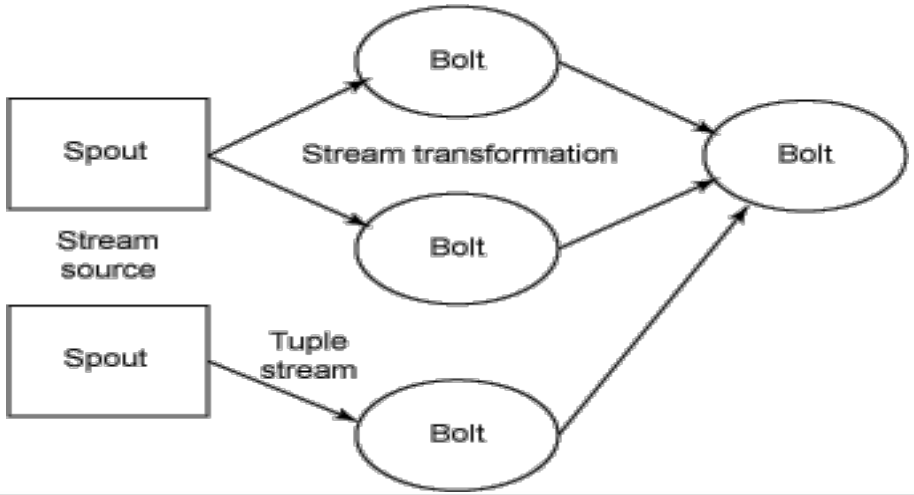
Can be used with any programming language.
Typically sits behind Kafka.
Real-time processing.
Scale on demand.
Compatible with cloud storage platforms.
Master-node topology.

Icons

STORM

$$

Cost



Other Big Data Technologies

Docker Containers

Characteristics

Open source

Virtualized OS-level

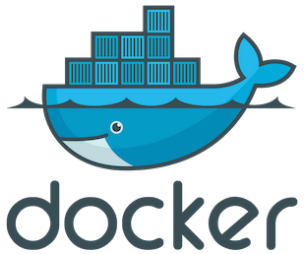
Automates deployment of apps

Portable

Notes

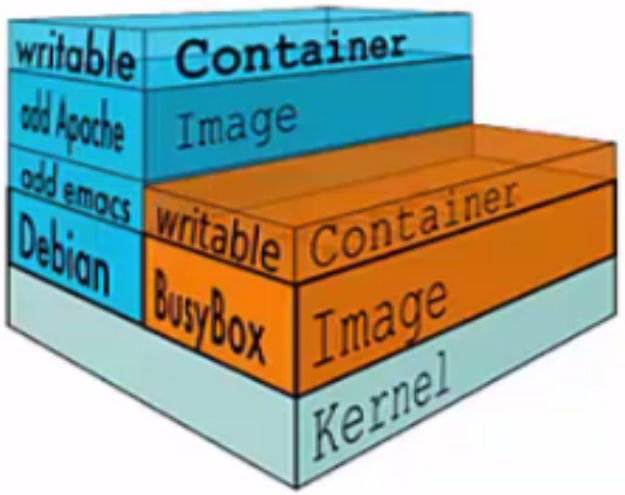
House all needed components within the container.
Consistent deployment.
Docker Hub is repo for public images.
Scale by environment.
IT maintains control of container usage.

Icons



$

Cost



Containers vs Virtual Machines

Containers

Can run multiple containers easier.

Portable between AWS, Google, Azure.

Lightweight

Quick start-up

Effective host resource sharing.

Virtual Machine

Consumes underlying machine

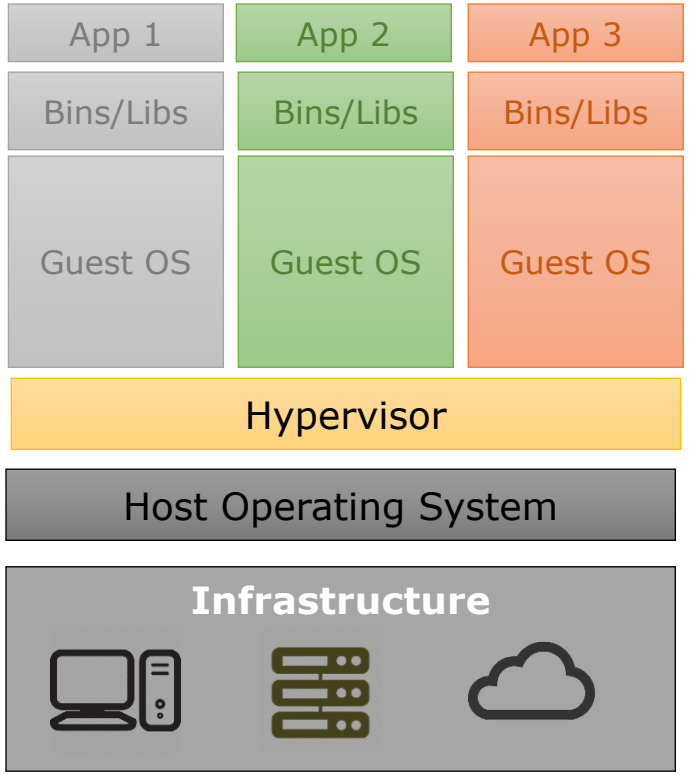
Complicated export process.

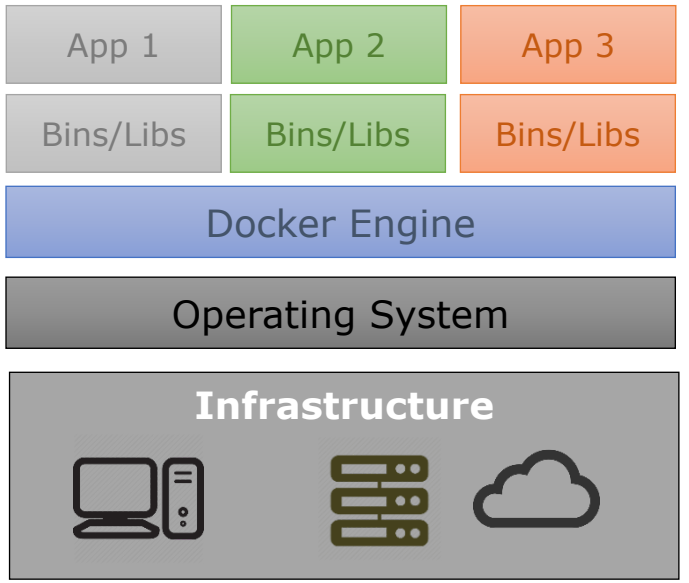
Requires all libraries underlying host

OS Start-up required

Direct consumption of host

Containers vs Virtual Machines





Virtual Machines

Each virtual machine includes the application, the necessary binaries and libraries and an entire guest operating system-all of which may be tens of GBs in size.

Containers

Containers include the application and all of its dependencies, but share the kernel with other containers. They run as an isolated process in user space on the host operating system. They're also not tied to any specific infrastructure- Docker containers run on any computer, on any infrastructure and in any cloud.

Kubernetes

Characteristics

Open source

Manages containerized workloads

Automates deployment of apps

Portable

Notes

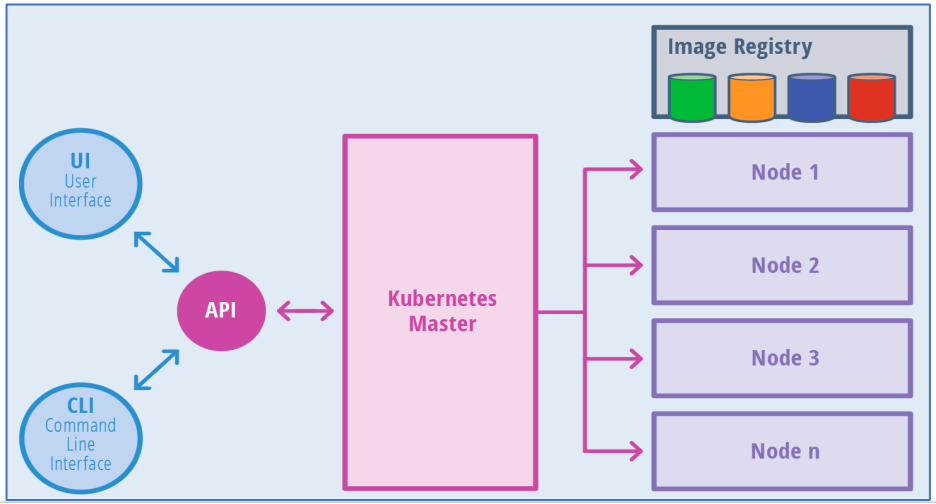
A container platform.
Microservices platform
Portable cloud platform.
Orchestrates computing, storage, networking custom to workload needs.
Scalable at application level.
The future!

Icons

kubernetes

$

Cost



High Performance Compute

Characteristics

Use of super computers

Massively Parallel computing

Most complex and large-scale jobs

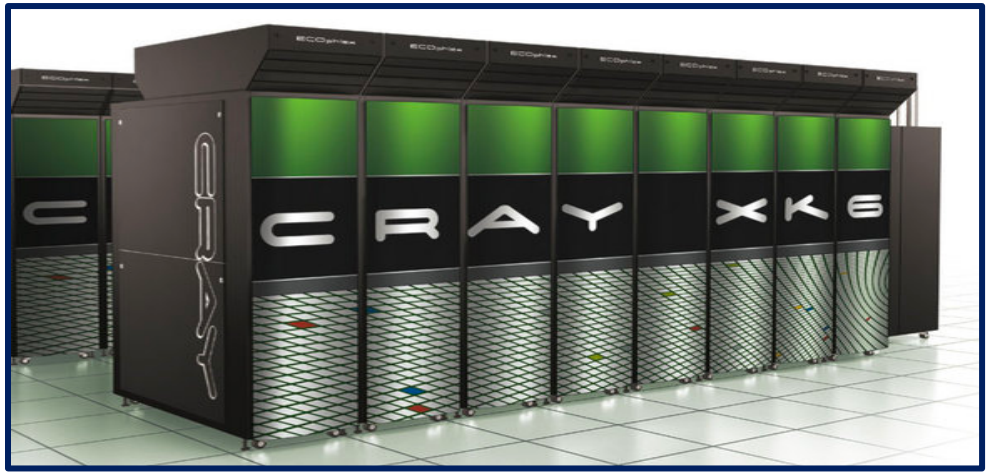
Notes

Hundreds of petabytes per cluster..
Offered as a service from cloud vendors.
Auto monitoring and easy management.
GPU and FPGA enabled.
Complex Deep Learning problems.

Icons

$$$

Cost



Azure Databricks

Hands-on-lab