架构迎接未来变化
IAS 2018

天数智芯
Engine for intelligence

分布式机器学习平台二三事

南京天数智芯科技有限公司-倪岭

机器学习平台

架构迎接未来变化
IAS 2018

1. Overview of the Architecture
2. Distributed Storage
3. Distributed Computing
4. Orchestration and Scheduling
5. Summarize, Q&A

天数智芯

机器学习平台设计目标

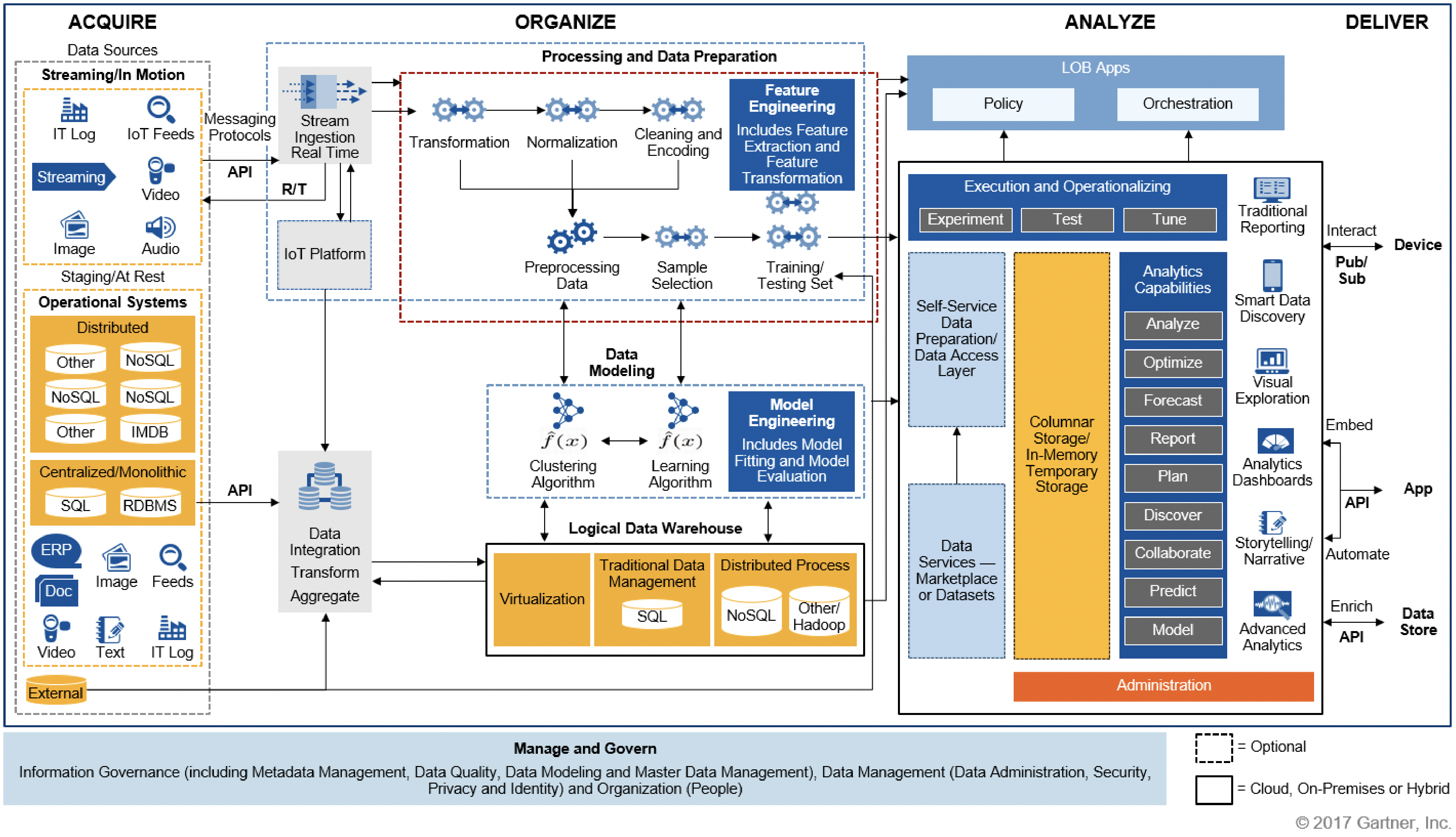
At large companies, machine learning is ***80 percent infrastructure.***

– by a machine learning engineer

- Goal
  - Simplify the workflow, lower the complexity
  - Provide generic, standard, reusable tool and solutions
  - Enable easy and secure sharing and collaboration for data and models
  - Reduce time and effort for data scientists and analysts
  - Optimized for scalability and performance

Make it easy to do right, and hard to go wrong!

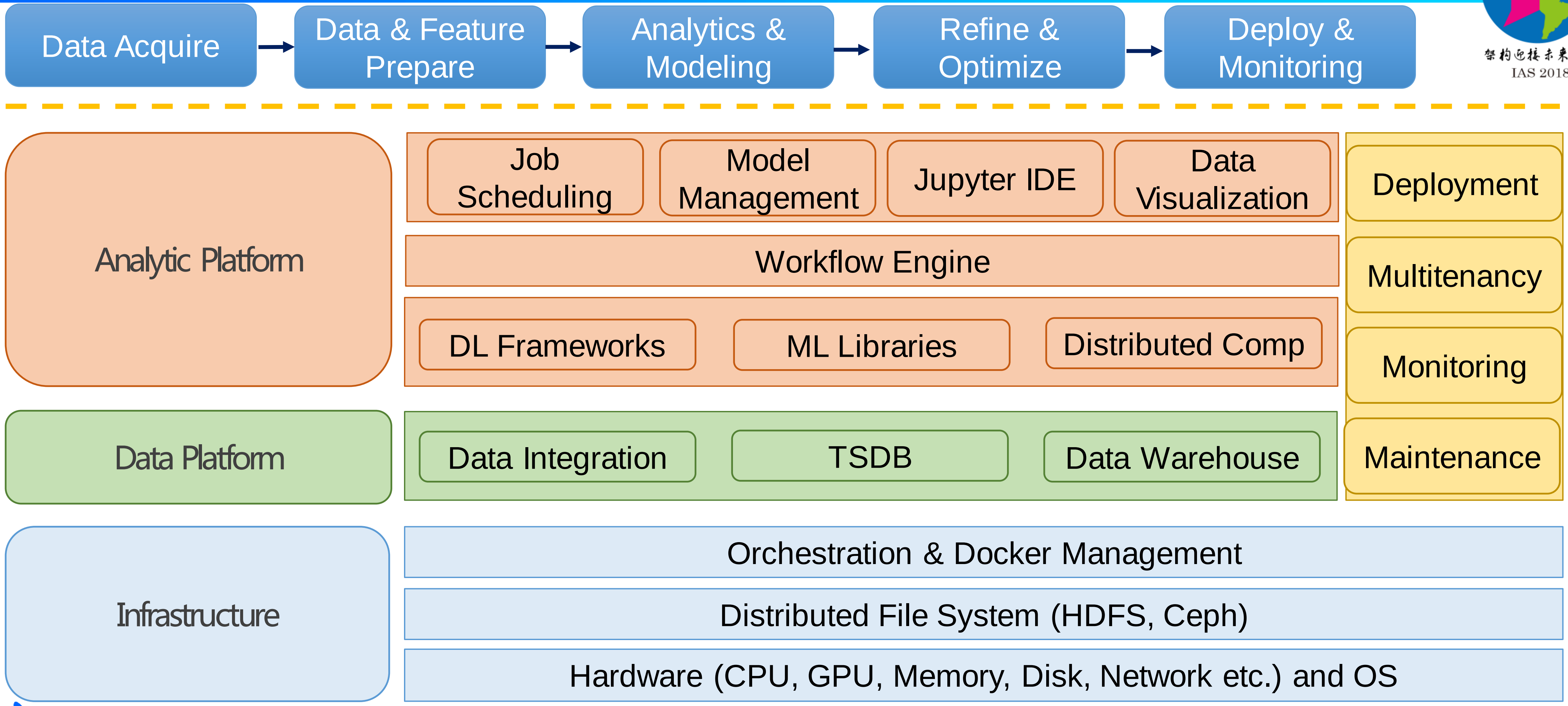
机器学习平台架构 – 四阶段工作流架构



IAS 2018

天数智芯

机器学习平台架构设计 – 分层架构



架构改核未来变化
IAS 2018

天数智芯

机器学习平台 – 关键技术选择

架构匠技术未来变化
IAS 2018

- Configurable and elastic resource management, consistent and reproducible environment
  - Use Kubernetes and Docker
- Easy to scale, customize, function extension and data/model sharing
  - Use modular design/components, reuse mature frameworks as building blocks
- Support common workflow and popular analytics and ML frameworks
  - Integration of popular libraries and frameworks such as Spark, TF, XGBoost, scikit-learn, PyTorch etc.
- High performance
  - Comprehensive optimization and tuning across the whole stack
- Easy to use
  - Intuitive user interface, visualization across the pipeline, Automation, etc.

天数智芯

机器学习平台中的分布式设计原则

架构始终未来变化
IAS 2018

Multiple computers that interact with each other over a network to achieve a common goal.

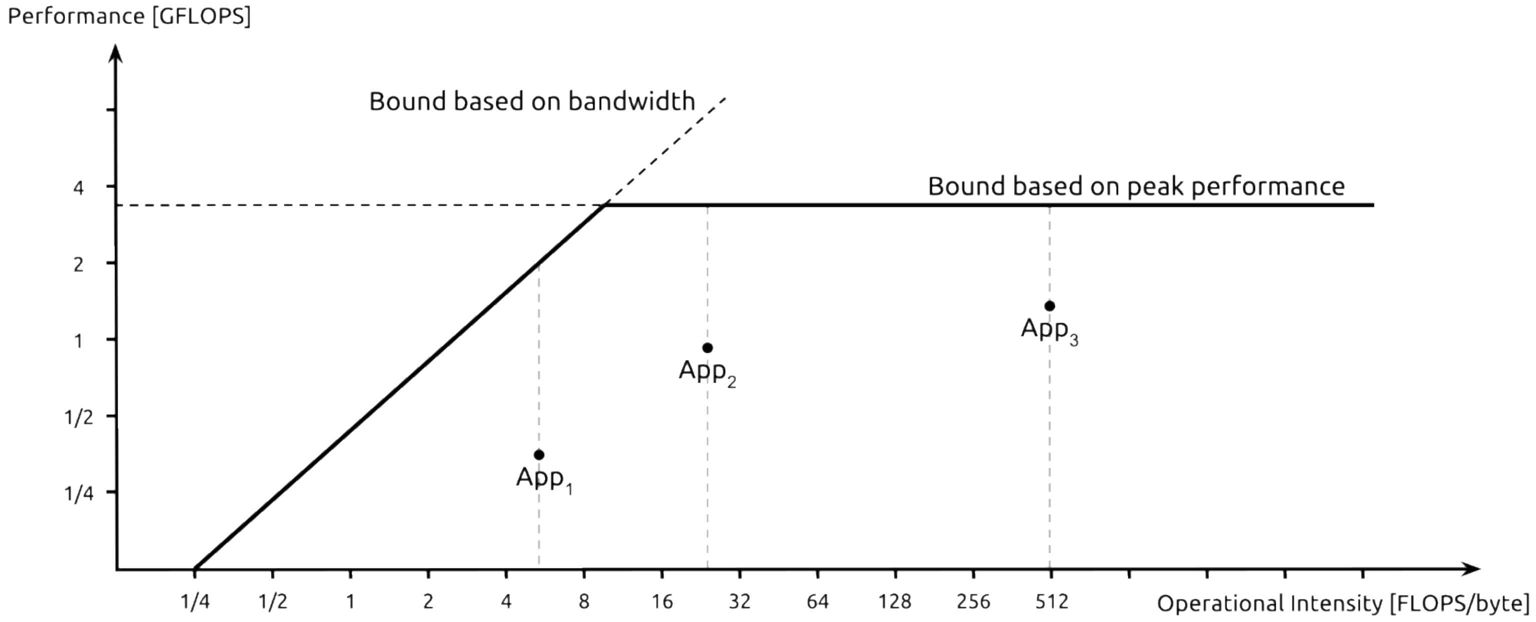
- Incremental scalability – scale for performance and availability
- Symmetry – all nodes are equal
- Decentralization – No central control, avoid single point of failure
- Be Redundant - Use replicas for both reliability and performance
- Asynchronous rather than synchronous
- Strive for statelessness

It depends... And, there is always tradeoff...

天数智芯

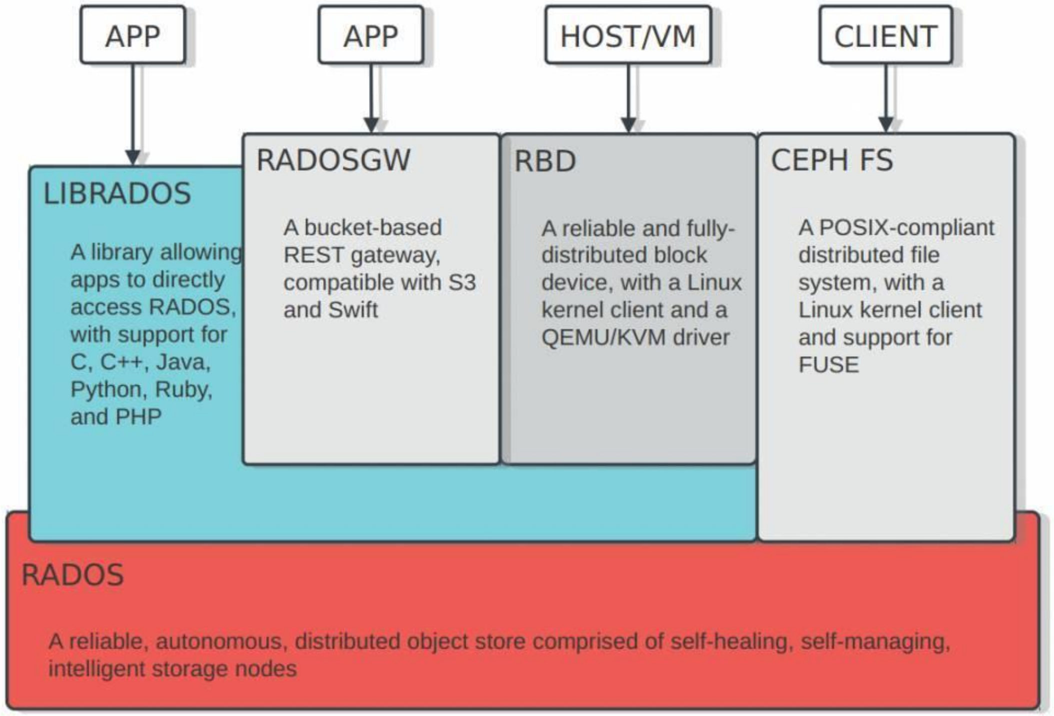
分布式存储 – 存储与计算分离

- Storage + Compute together for Big Data
  - HPC->Hadoop : Move Compute to Data
- Decoupling Storage And Compute (Disaggregated) for ML
  - Elastic, Scalability
  - Manageability
  - Flexibility
  - Hardware requirement
  - Hadoop -> HPC



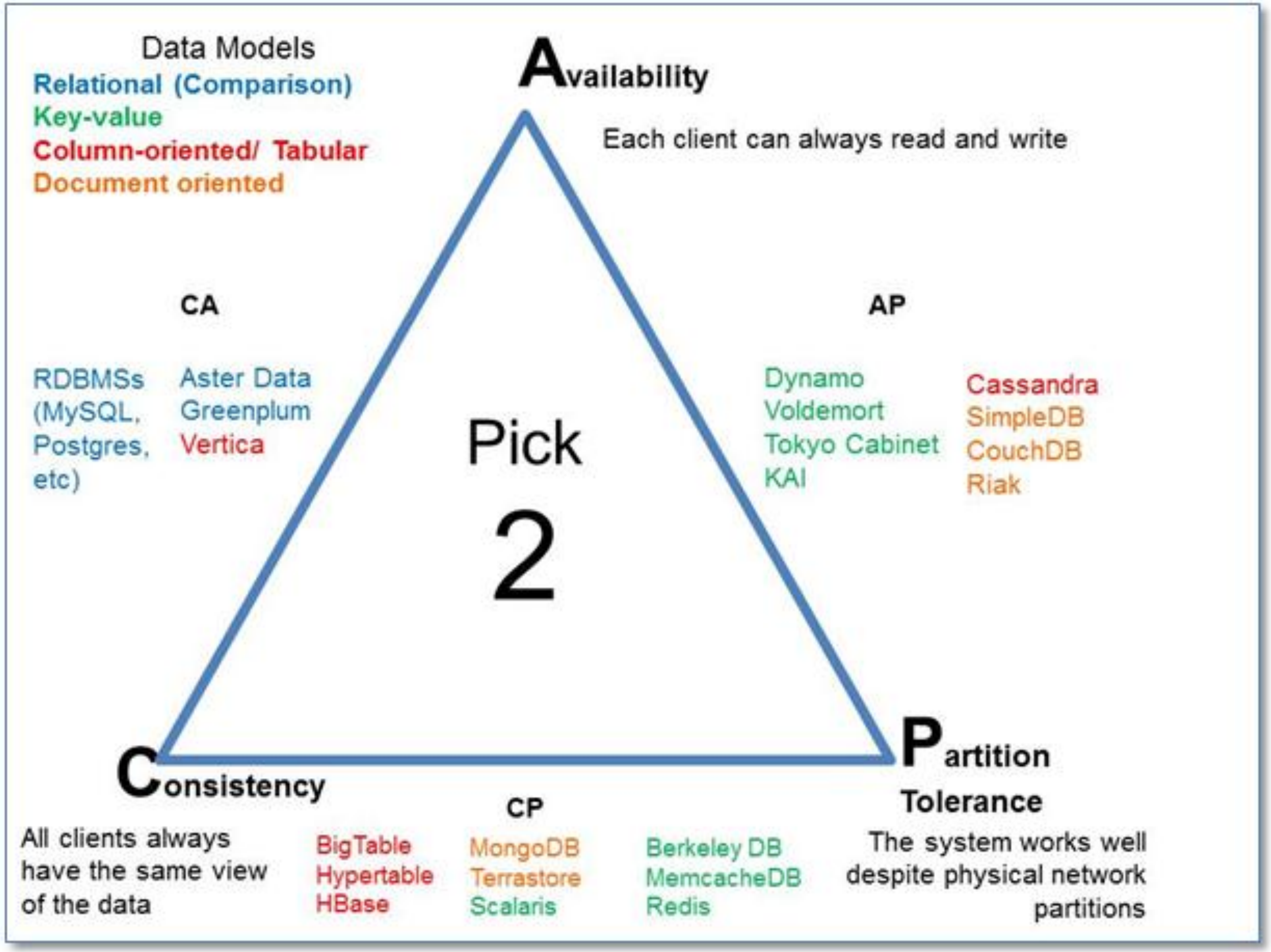
分布式文件系统

- HDFS vs Ceph
- File vs Block vs Object
- Ceph in practice
  - Ceph RBD and Cephfs
  - Erasure coding
  - Ceph with InfiniBand
  - Kubernetes Persistent Volume



分布式数据库

- CAP Theorem
- Choose your data store based on real use cases
  - SQL on Hadoop/NewSQL
  - NoSQL
  - Specialized Database

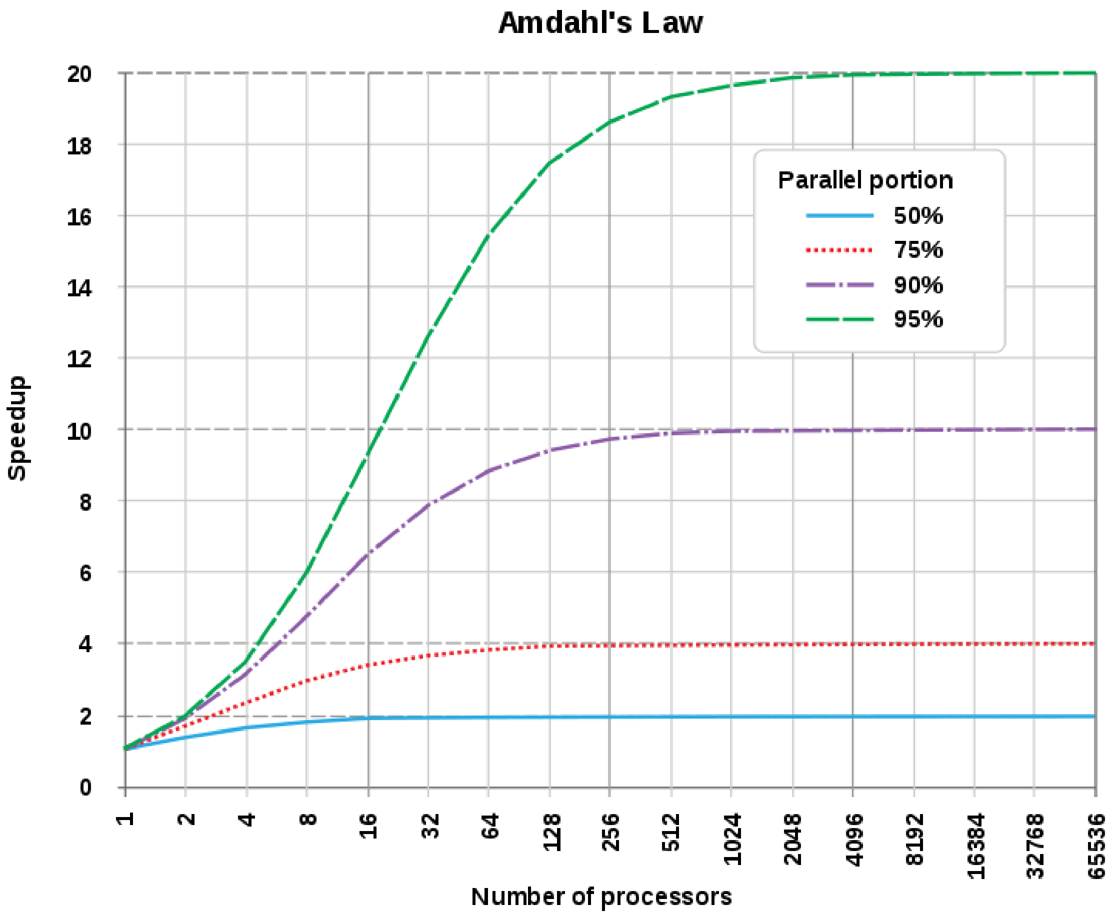


分布式计算 – 加速

- Amdahl’ s Law

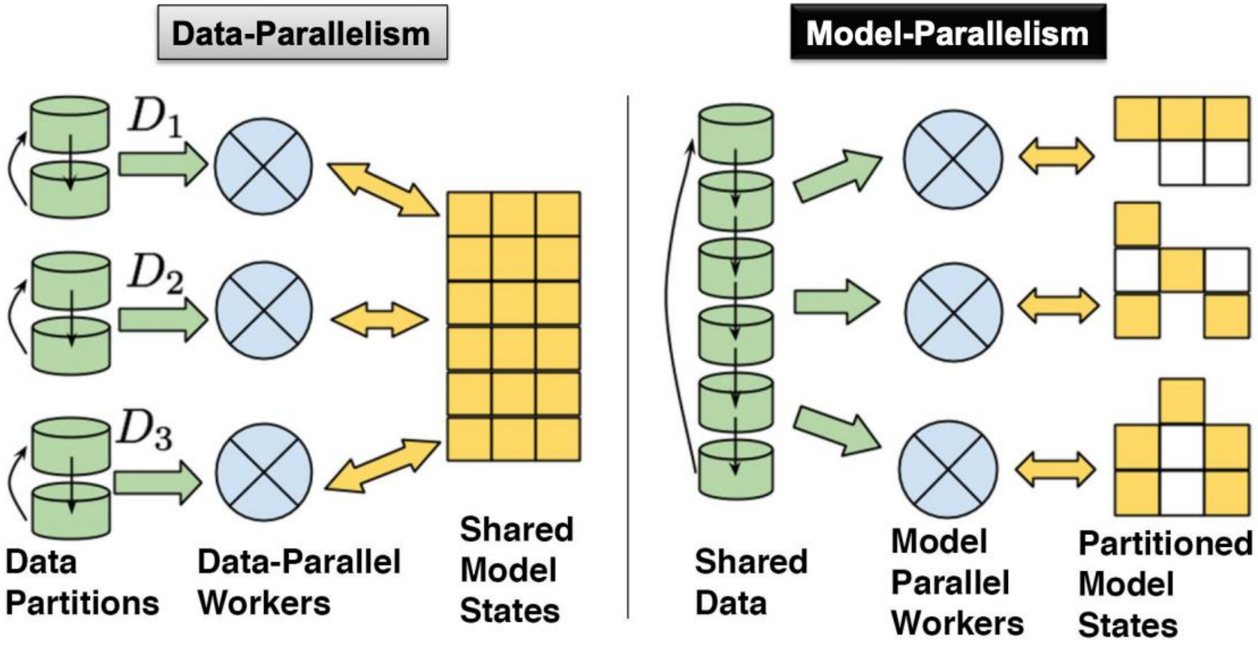
$$S(N) = \frac{1}{(1-P)+\frac{P}{N}}$$

- What else
  - CPU <-> GPU cost
  - Network Communication cost and latency
  - Multi level parallelism



分布式计算 – 深度学习

- What distributed machine learning is meant to resolve
  - Computation Complexity
  - Data volume & Model size
  - Not always the best option
- Challenges
  - Sometimes algorithms specific
  - Data Parallel vs Model Parallel
  - Efficient Communications, Consistency Protocol, User friendly



Credit: *Petuum: A New Platform for Distributed Machine Learning on Big Data* by Eric Xing et al.

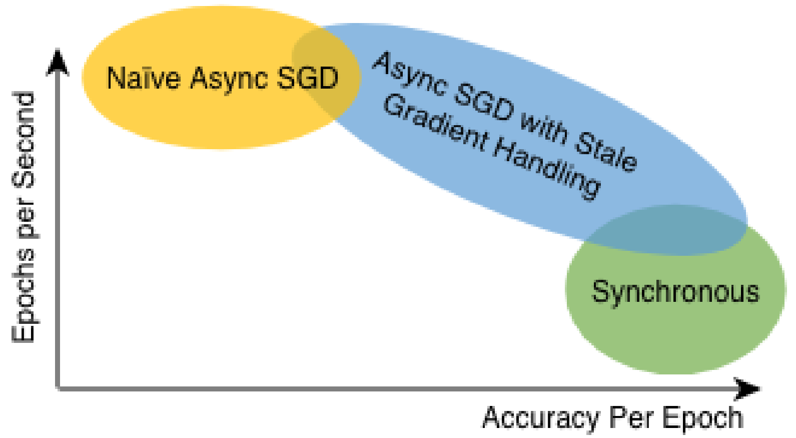
分布式计算 - ImageNet 刷刷刷

Training time and top-1 validation accuracy with ImageNet/ResNet-50

	Batch Size	Processor	DL Library	Time	Accuracy
He et al.	256	Tesla P100 x8	Caffe	29 hours	75.30%
Facebook	8K	Tesla P100 x256	Caffe2	1 hour	76.30%
IBM	8K	Tesla P100 x256	Caffe	50 mins	75.01%
Preferred Networks	32K	Tesla P100 x1024	Chainer	15 mins	74.90%
Tencent	64K	Tesla P40 x2048	TensorFlow	6.6 mins	75.80%
Sony	34K→68K	Tesla V100 x2176	NNL	224 secs	75.03%
Google	32k	TPU V3 x1024	TensorFlow	2.2min	76.30%

分布式计算 - 分布式训练技巧

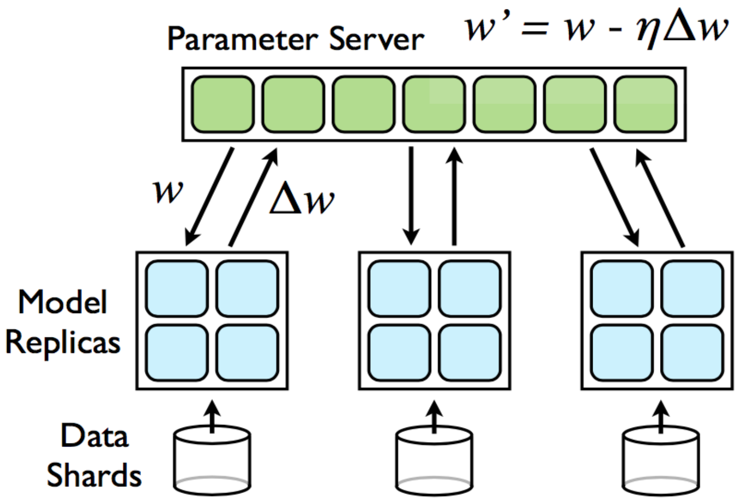
- *Keep It Simple, Stupid.*
- Ensure correctness/convergence, use synchronous SGD
- Large Batch size
  - Learning rate hacks
  - LARS (Layer-wise Adaptive Rate Scaling) Optimizer

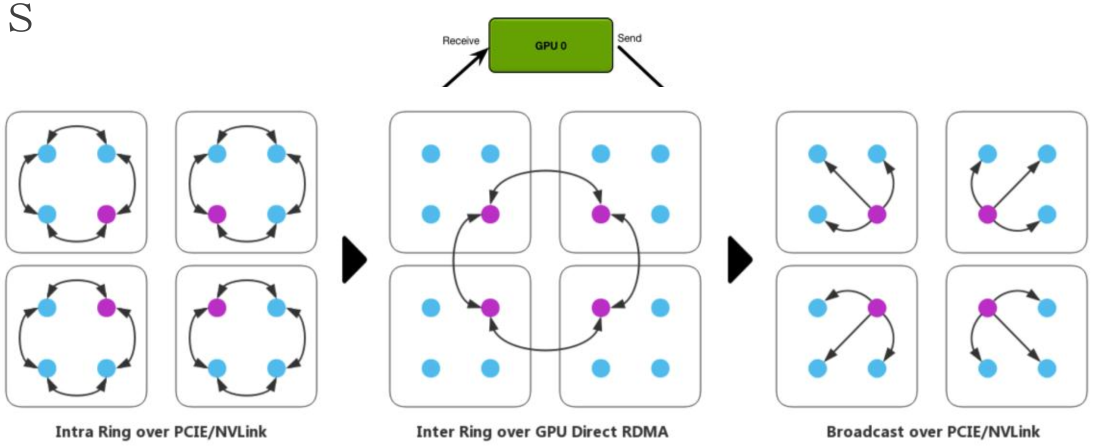


#GPUs	Images per Second	GPU Scaling Efficiency
4	1608	
1088	400778	91.62%
2176	579040	66.19%
2720	729051	66.67%
3264	688504	52.47%

分布式计算 - DNN的网络传输

- Efficient communication protocol
  - MPI Ring All reduce and all kinds of variants
  - Parameter Server
  - P2P with SFB, Mixed/hybrid protocols
- Reduce communication overhead
  - Gradient compression, quantization
  - Mixed precision training
  - Local aggregation, tensor fusion





分布式计算 – 深度学习工程实践

- Horovod library by Uber
  - Implemented Data Parallelism and Ring All Reduce
  - Easy to use with wrapped distributed optimizer
  - Non invasive to deep learning frameworks
- What else
  - How to achieve full user transparency
  - Parameter server still shines
  - Data distribution, pipeline and caching

分布式计算 – 传统数据分析与机器学习

- Spark, H2O distributed engine
- Classic Python
  - Start from single node parallelism
  - Mixed distributed-parallel paradigm
  - Distributed memory vs shared memory
  - Dask and Dask-ML
  - Celery distributed task queue for inference application

分布式调度 – Kubernetes

- Cloud Native challenge
- Resource scheduling
  - Flexible vs templated
  - Overcommit vs exclusive resource usage
- CPU/GPU topology awareness and label management
- Kubernetes networking performance tuning with Calico
- Workflow engine optimization

架构迎接未来变化
IAS 2018

天数智芯

分布式设计误区

- The network is reliable.
- Latency is zero.
- Bandwidth is infinite.
- The network is secure.
- Topology doesn’t change.
- There is one administrator.
- Transport cost is zero.
- The network is homogeneous.

8 fallacies of Distributed Systems

By Peter Deutsch & James Gosling

天数智芯

Engine
for intelligence

Thanks！