Alluxio - 开源AI和大数据存储编排平台

顾 荣

Alluxio PMC & Maintainer

南京大学 计算机系副研究员、博士

提纲

1. Alluxio项目&系统简介
2. Alluxio 2.0新特性概览
3. Alluxio未来发展趋势快览
4. 总结

数据处理的四大趋势驱动了新型基础架构的需求

Data Ecosystem - *Beta*

COMPUTE

hadoop Map Reduce

hadoop HDFS

STORAGE

Data Ecosystem 1.0

COMPUTE

presto Spark Flink hadoop Map Reduce

HIVE Caffe APACHE HBASE TensorFlow

amazon web services S3 Microsoft Azure Google Cloud Storage ceph hadoop HDFS

IBM NetApp Hewlett Packard Enterprise DELL EMC ECS HITACHI Hitachi Data Systems MINIO

STORAGE

ALLUXIO

大数据之路与企业创新的选择

同置 (Co-located)

Co-located
compute & HDFS
on the same cluster

分散 (Disaggregated)

Disaggregated
compute & HDFS
on the same cluster

混合云化部署HDFS

Burst HDFS data in the cloud,
public or private

支持更多计算框架

Support Presto, Spark and other computes without app changes

向对象存储过渡

Enable & accelerate big data on
object stores

技术转变中的挑战

混合云部署HDFS

- Accessing data over WAN too slow
- Copying data to compute cloud time consuming and complex
- Using another storage system like S3 means expensive application changes
- Using S3 via HDFS connector leads to extremely low performance

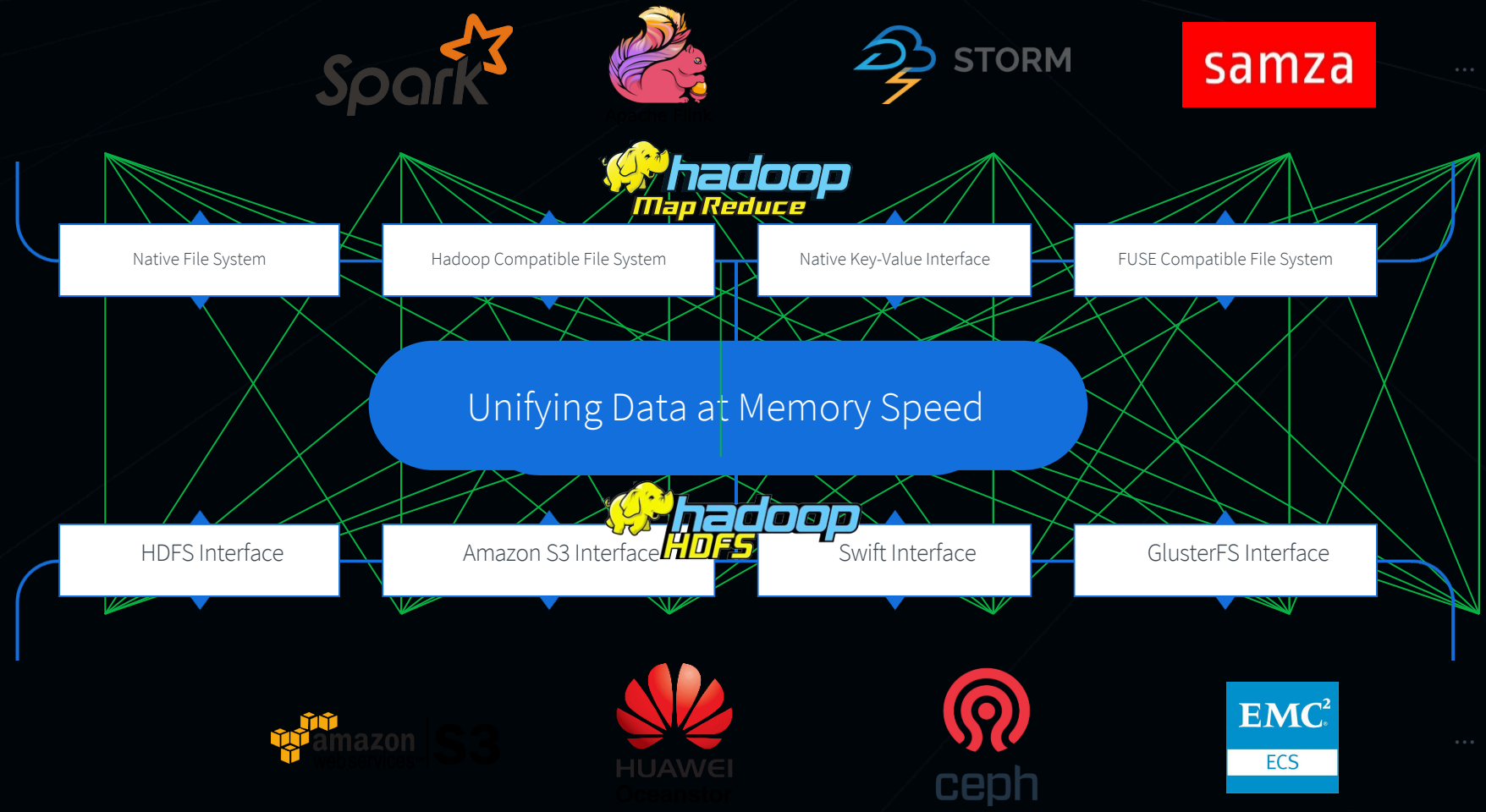
支持更多计算框架

- Copying data to multiple compute clouds time consuming and error prone
- Migrating applications for new storage systems is complex & time consuming
- Storing and managing multiple copies of the data becomes expensive

向对象存储过渡

- Object stores performance for big data workloads can be very poor
- No native support for popular frameworks
- Expensive metadata operations reduce performance even more
- No support for hybrid environments directly

计算与存储实现独立可扩展性



计算与存储实现独立可扩展性

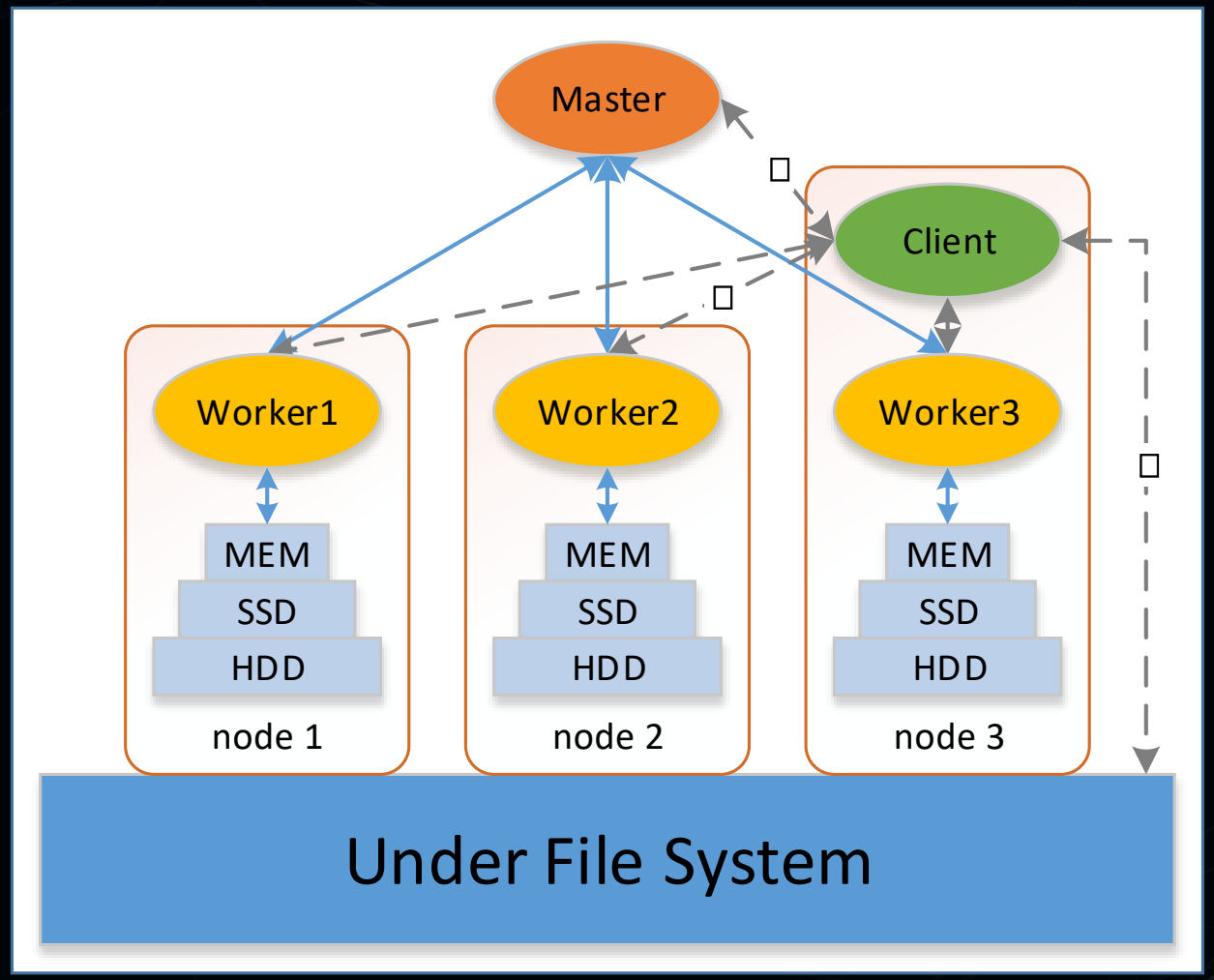
Java File API | HDFS Interface | FUSE Interface | S3 Interface | REST API

Alluxio: a Virtual Distributed File System (VDFS)

HDFS Driver | S3 Driver | Swift Driver | NFS Driver

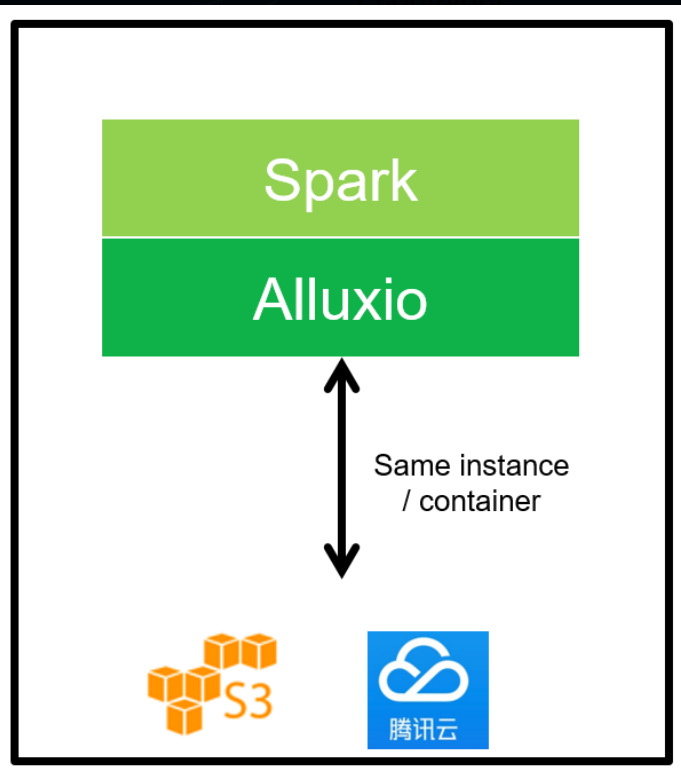
Alluxio系统内部整体架构

- Master-Worker
  - Master
    - 管理全部元数据
    - 监控各个Worker状态
  - Worker
    - 管理本地MEM、SSD和HDD
- Client
  - 向用户和应用提供访问接口
  - 向Master和Worker发送请求
- Under File System
  - 一般用于备份

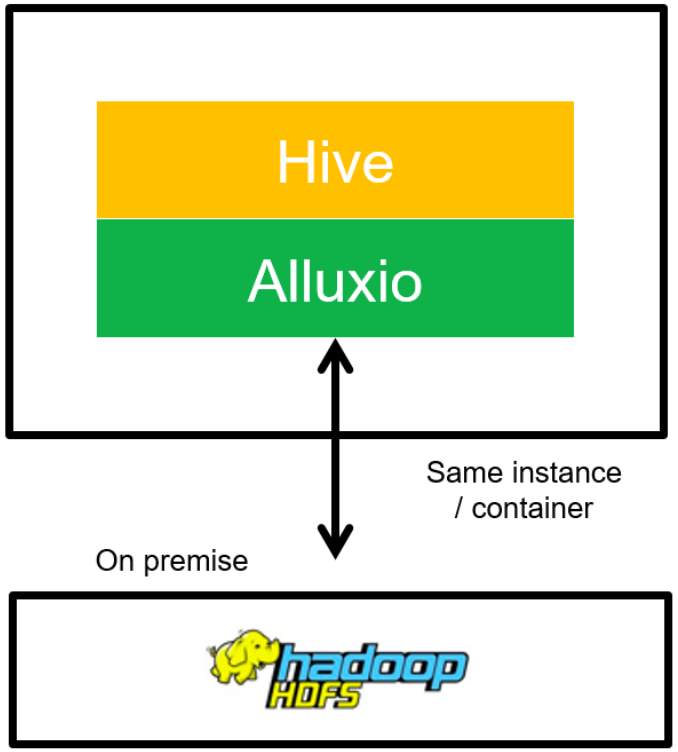


Alluxio数据编排赋能的几类场景

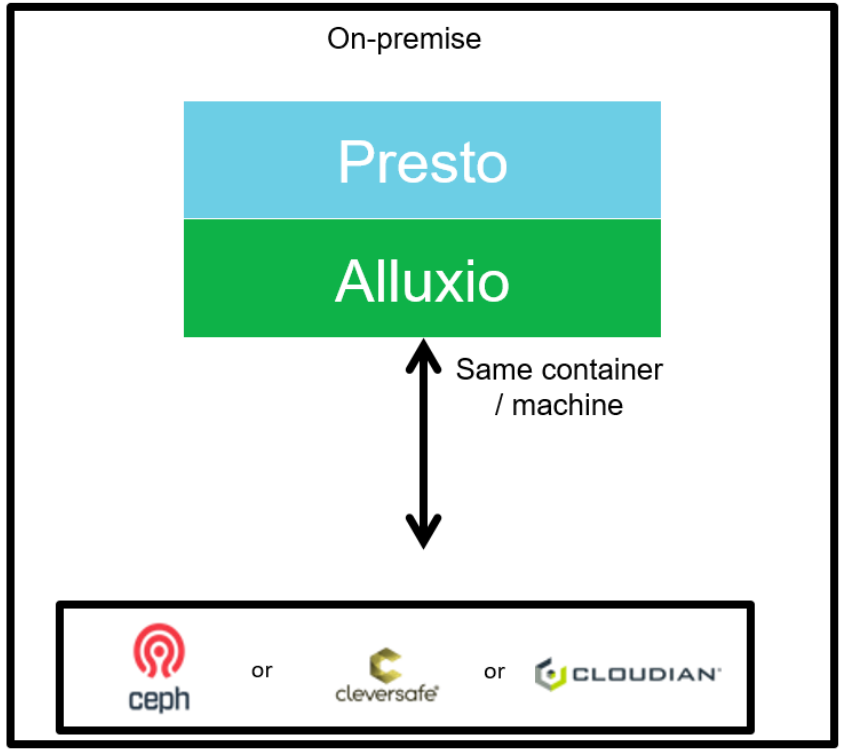
Accelerate big data frameworks on the public cloud



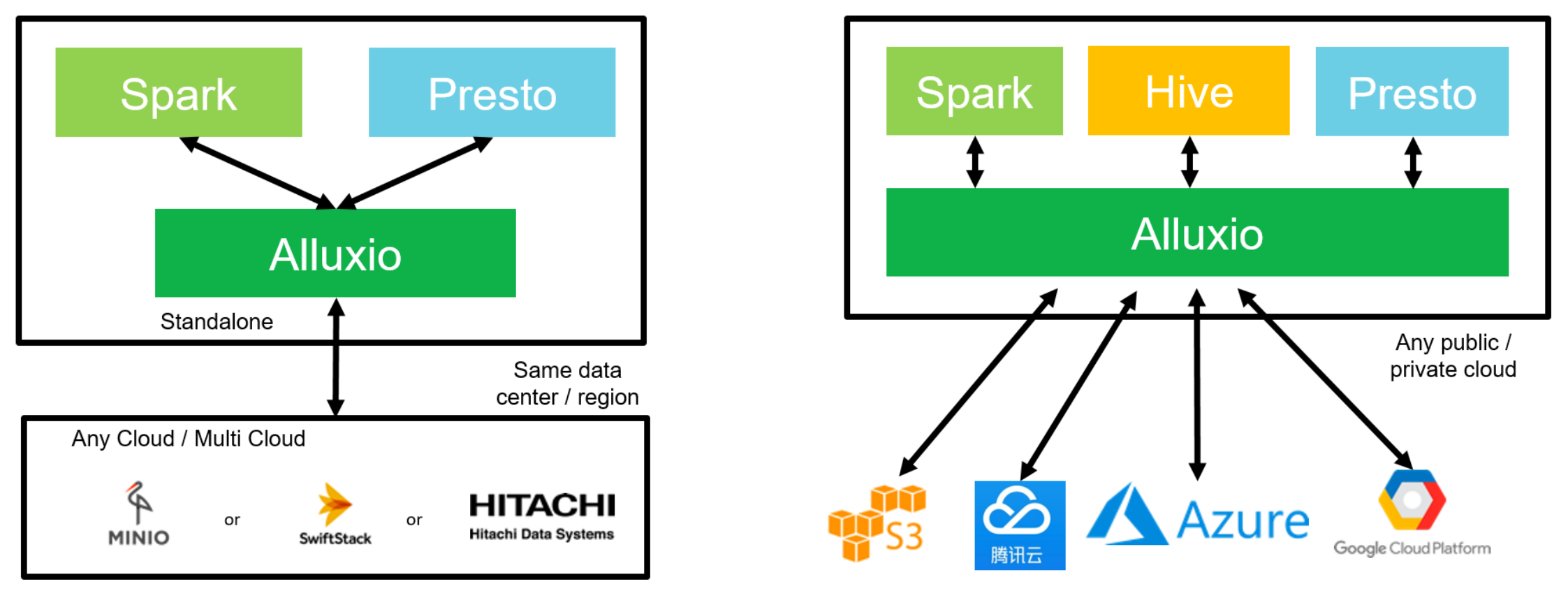
Burst big data workloads in hybrid cloud environments



Dramatically speed-up big data on object stores on premise



高级使用场景



Enable big data on object stores across single or multiple clouds

Orchestrate data frameworks on the public cloud

Alluxio的核心创新

数据本地性
Data Locality
with Intelligent
Multi-tiering

Accelerate big data workloads with transparent tiered local data

数据可访问性
Data Accessibility
for popular APIs &
API translation

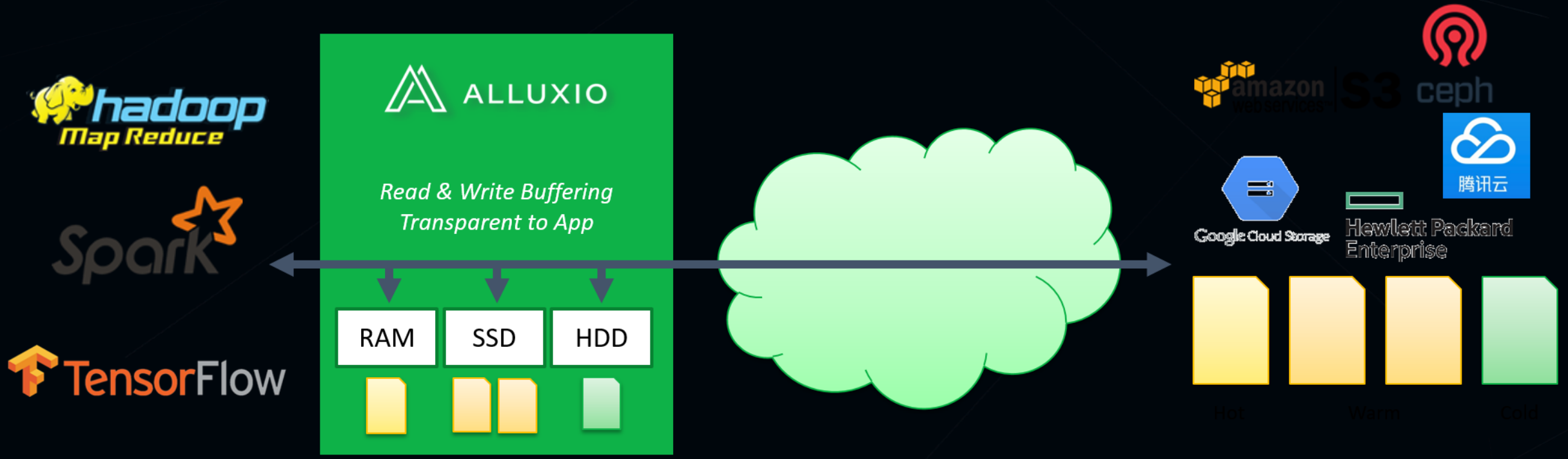
Run Spark, Hive, Presto, ML workloads on your data located anywhere

数据伸缩性
Data Elasticity
with a unified
namespace

Abstract data silos & storage systems to independently scale data on-demand with compute

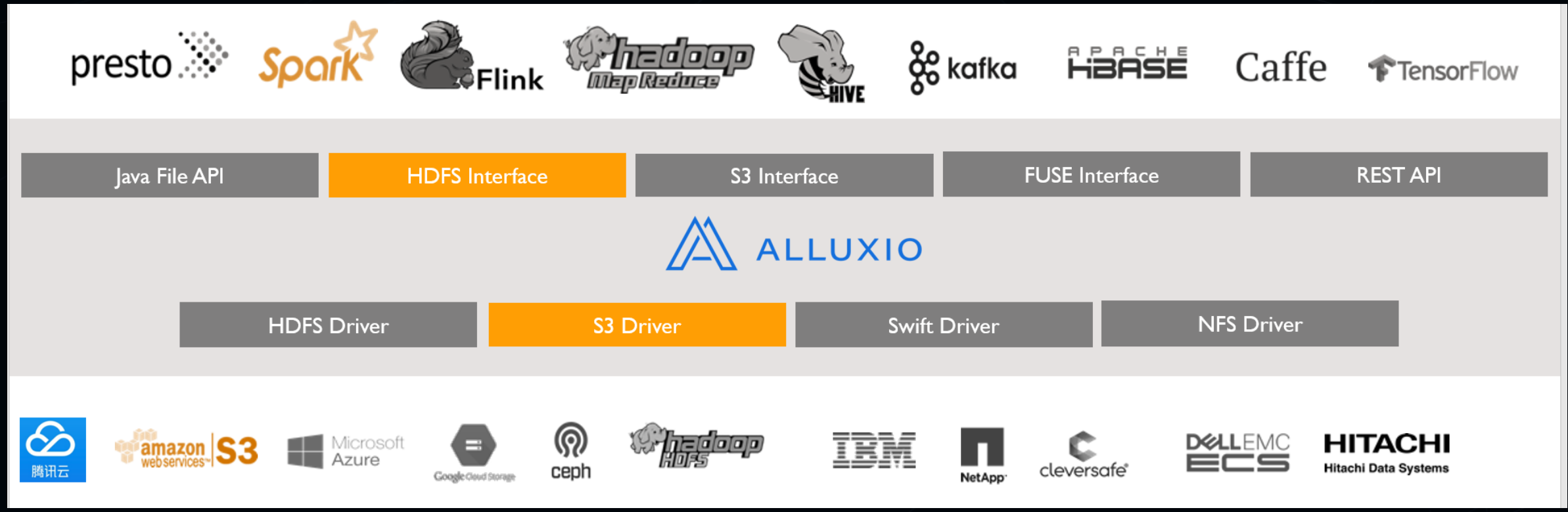
基于智能多层缓存实现数据本地性

Local performance from remote data using multi-tier storage



通过提供流行APIs和API转换实现数据可访问性

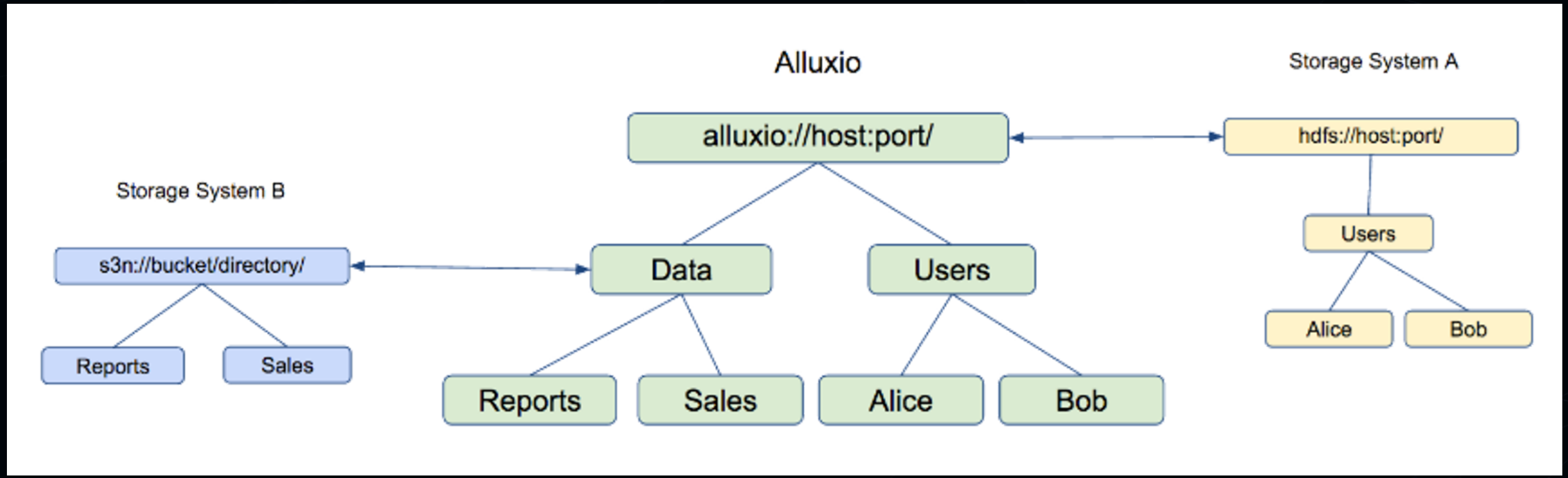
Convert from Client-side Interface to native Storage Interface



通过统一命名空间实现数据可伸缩性

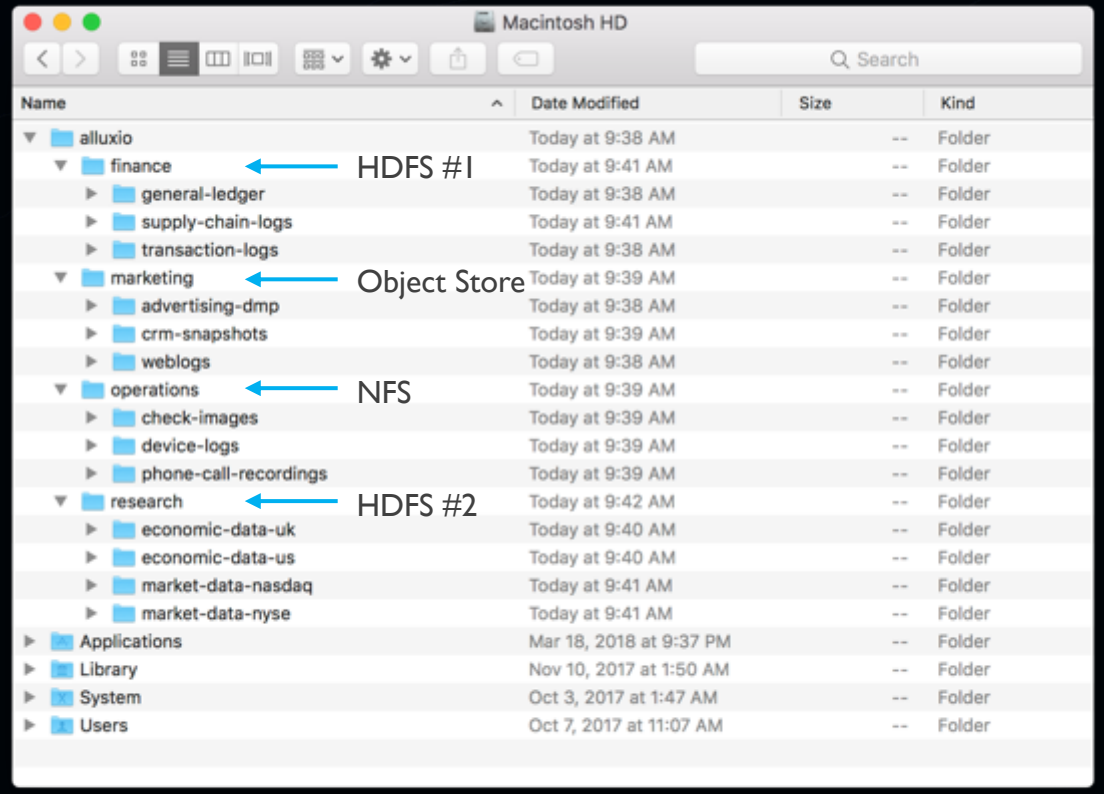
Enables effective data management across different Under Store

Uses Mounting with Transparent Naming



统一命名空间（Unified Namespace）

Transparent access to understorage makes all enterprise data available locally



SUPPORTS

- HDFS
- NFS
- OpenStack
- Ceph
- Amazon S3
- Azure
- Google Cloud

IT OPS FRIENDLY

- Storage mounted into Alluxio by central IT
- Security in Alluxio mirrors source data
- Authentication through LDAP/AD
- Wireline encryption

100+ Known Production Deployments

Financial Services

DBS
ING
BARCLAYS
WELLS FARGO
PayPal
HUATAI SECURITIES

Retail & Entertainment

Walmart
Shopee
ROBLOX
CAESARS ENTERTAINMENT
NetEase Games
Myntra
JD.COM 京东
苏宁易购 suning.com
唯品会 VIPSHOP.COM

Data & Analytics Services

Aunalytics
LigaData
SAMSUNG SDS
TalkingData
RYTE

Technology

Tencent 腾讯
PERCEPTIN
esri
Tencent Cloud
@WalmartLabs
ESENTIRE
Baidu 百度
ORACLE
Alibaba Cloud
Baidu Cloud
Sogou 搜狗
CUELOGIC
Lucidworks
QINIU 七牛云存储
walkme
MOMO

Consumer

SAMSUNG
nielsen
YAHOO!
链家
Lenovo
bazaarvoice
GUARDANT HEALTH
网易 严选

Telco & Media

COMCAST
HUAWEI
虎牙直播 HUYA.com
swisscom
中国移动 China Mobile
China unicom 中国联通

Travel & Transportation

waze
Ctrip
DiDi
去哪儿 Qunar.Com

TalkingData

Incredible Open Source Momentum with growing community

1000+ contributors & growing

4278+ Git Stars

Github: https://github.com/Alluxio/alluxio
Join the conversation on Slack
alluxio.org/slack

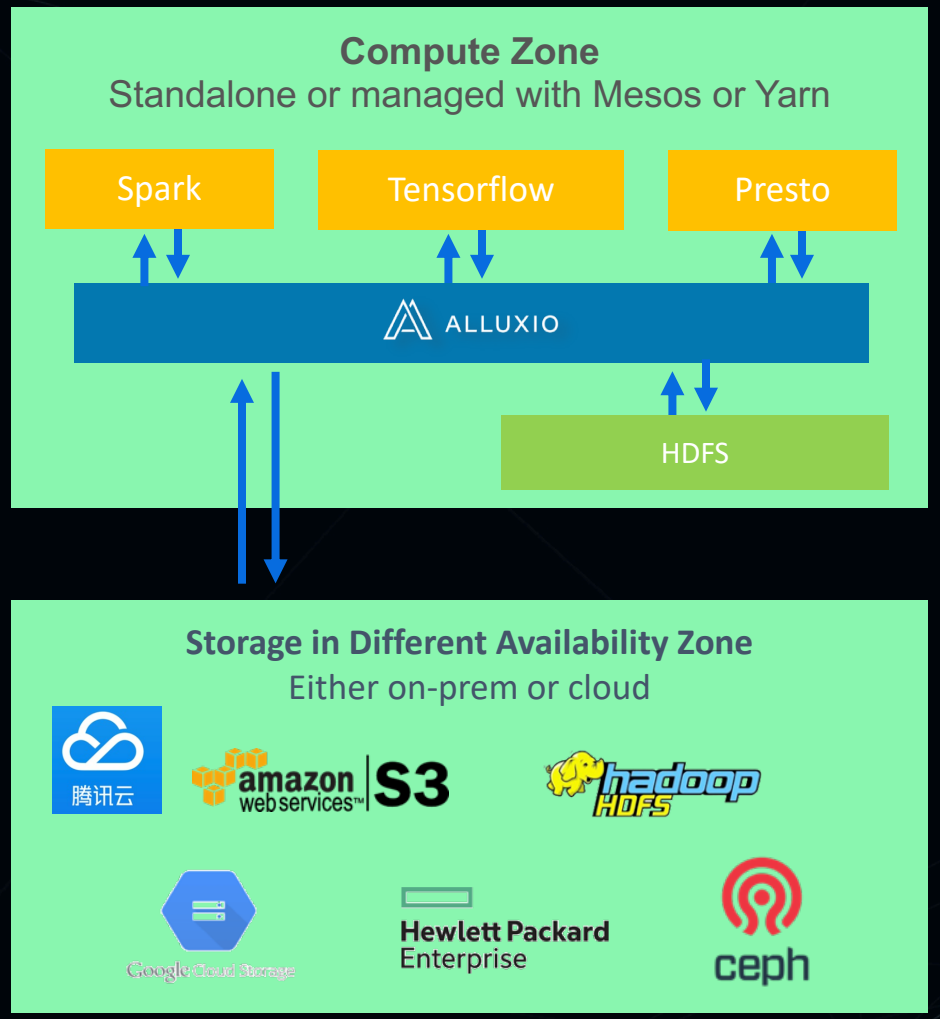
Apache 2.0 Licensed

Hundreds of thousands of downloads

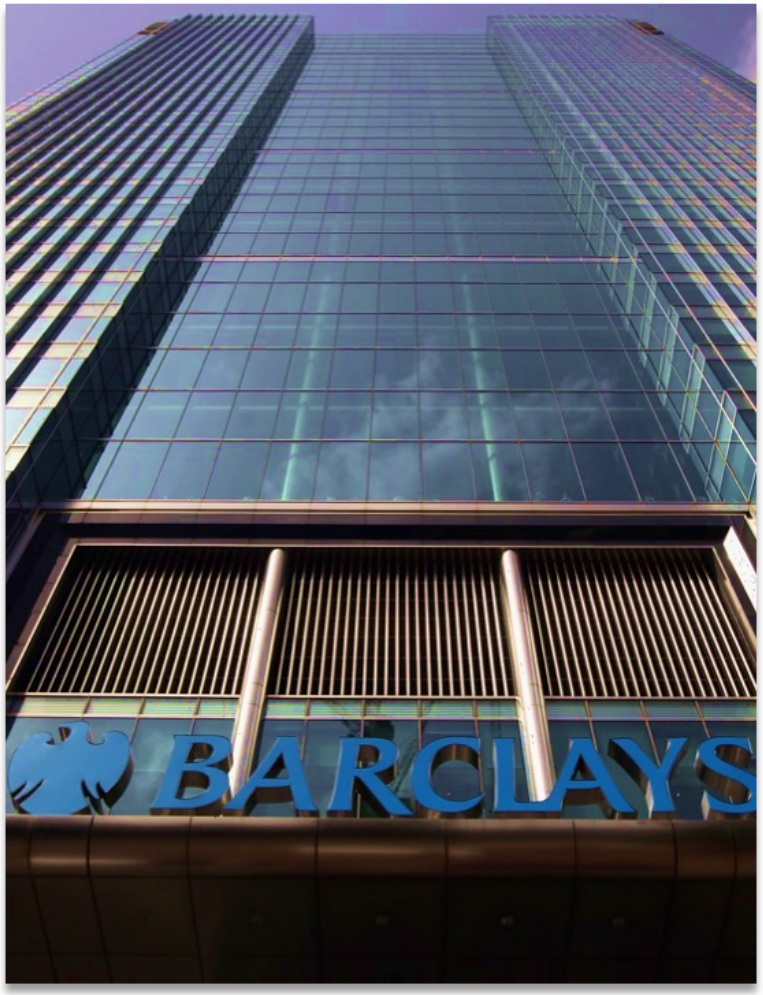
Alluxio适用场景分析

- Finding high-fit use-cases
- Example First Projects
  - Enterprise Storage & Big Data Teams
    - Virtual Data Lakes
    - Gradual transition to low cost storage
    - Unify hybrid-cloud storage
  - Machine Learning & Data Science Teams
    - Accelerate training
    - Improve productivity

Alluxio is installed with or near compute to unify data stores, stage remote data, and improve system performance.



Machine Learning Case Study – BARCLAYS



SPARK

TERADATA

Challenge –

Gain end to end view of business with large volume of data

Queries were slow / not interactive, resulting in operational inefficiency

SPARK

ALLUXIO

TERADATA

Solution –

ETL Data from Teradata to Alluxio

Impact –

Faster Time to Market – “Now we don’t have to work Sundays”

Use Case: http://bit.ly/2oMx95W

Analytics Use Case – Top Retailer

SPARK

HDFS

Challenge –

Bottleneck in Trend Analysis of mission critical daily sales and inventory management

Queries were slow / not interactive, resulting in operational inefficiency

SPARK

ALLUXIO

HDFS

Solution –

With Alluxio, data queries are 10X faster

Impact –

Higher operational efficiency

Use case: http://bit.ly/2ook8Nh

Alluxio 2.x新特性介绍

支持超大规模数据工作负载

- **支持超过10亿+个文件**
  - ✓ 2.0引入了分层元数据管理(tiered metadata management)这一新选项，以支持包含超过10亿个文件的单群集部署。
  - ✓ 我们现在默认使用RocksDB进行堆外存储。
  - ✓ 热数据的元数据继续存储在堆内的进程内存中，而其余元数据由Alluxio在进程内存外进行管理。
  - ✓ alluxio.master.metastore可以配置为仅使用堆内存储。

- **高度分布式数据服务**
  - ✓ 2.0引入了Alluxio作业服务(Job Service)，这是一种分布式集群服务，可以实现复制、持久化、跨存储移动和分布式加载等数据操作，从而实现高性能和大规模扩展。

Alluxio 2.x新特性介绍

支持超大规模数据工作负载

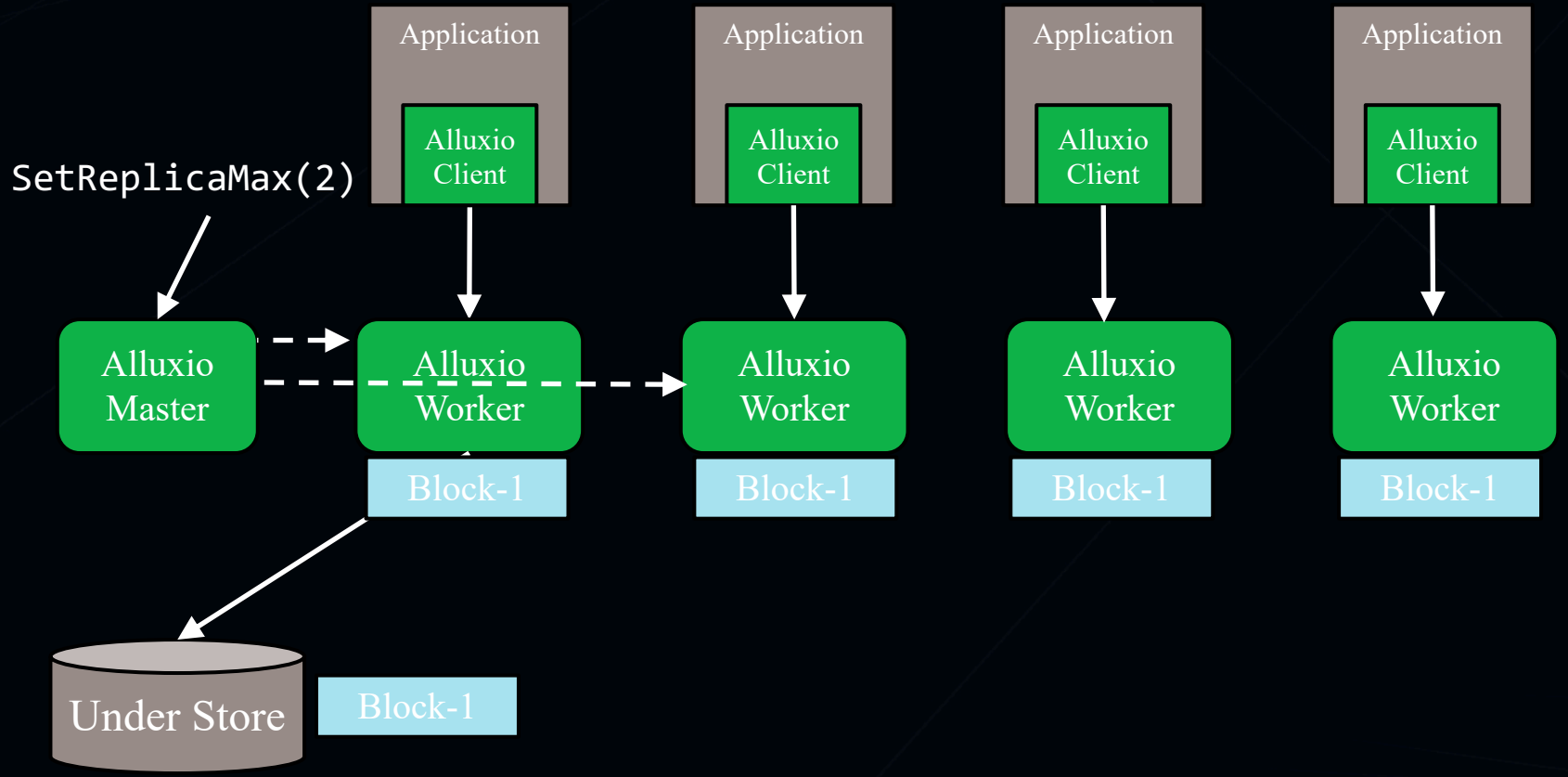
- **自适应副本以增强数据本地性**
  - ✓ 该功能为Alluxio配置一定数量范围的自动管理的存储数据副本数。
  - ✓ alluxio.user.file.replication.max和alluxio.user.file.replication.min可用于指定该范围。

- **内嵌式日志以达到高可用性**
  - ✓ 2.0设计了一种称为内嵌式日志(embedded journal)的面向文件/对象元数据的新容错和高可用模式。
  - ✓ 内嵌式日志使用RAFT共识算法，并且实现方面独立于任何其他外部存储系统。这对于抽象对象存储特别有用。

Alluxio 2.x新特性介绍

支持超大规模数据工作负载

- **自适应副本以增强数据本地性**



Alluxio 2.x新特性介绍

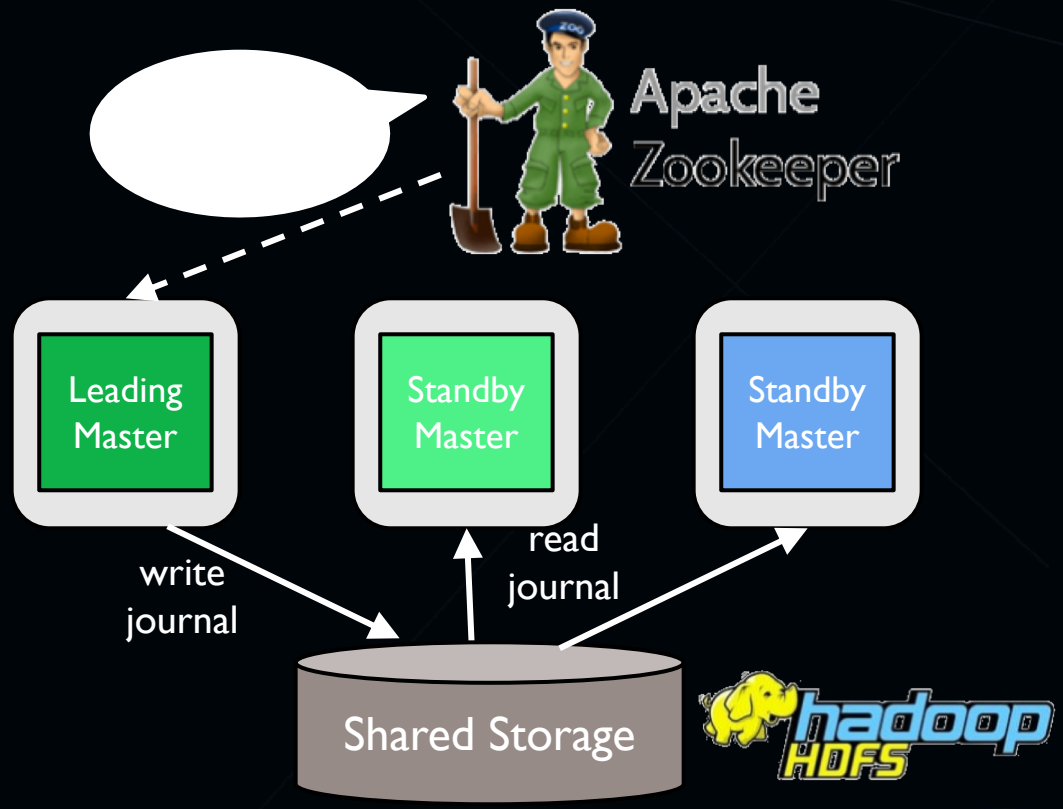
支持超大规模数据工作负载

- **内嵌式日志以达到高可用性**

Alluxio 1.x HA依赖ZK/HDFS组件

- **Alluxio HA运行模式**
  - ✓ Zookeeper: 负责选择leader master
  - ✓ HDFS: 负责存储日志文件，并在多个masters直接共享

- **存在的问题**
  - ✓ 日志存储的选择受限
  - ✓ 依赖于第三方组件，服务的调试恢复都比较困难。
  - ✓ HDFS集群本身的不稳定，会使得Alluxio集群维护成本变大



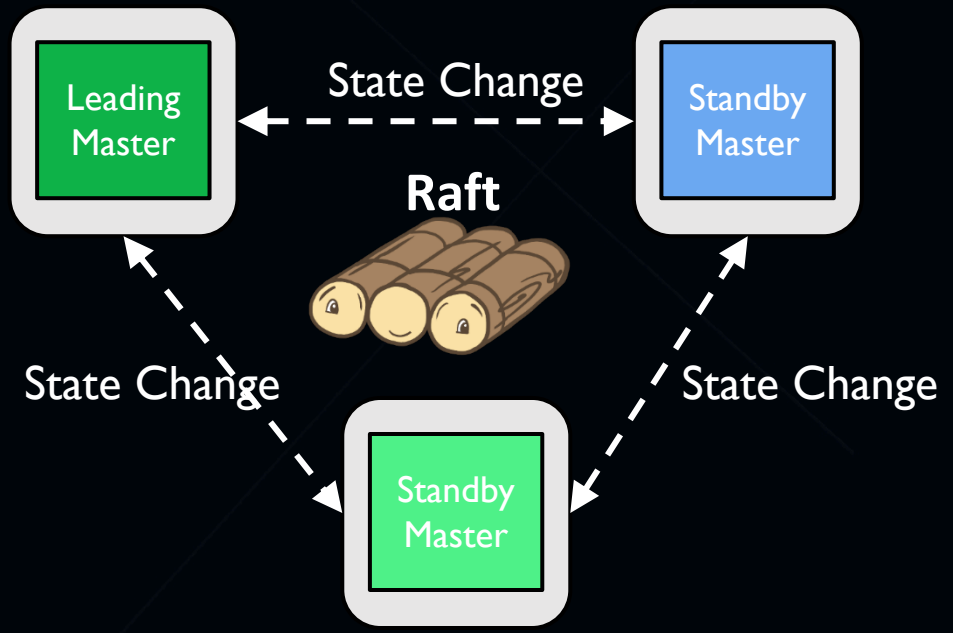
Alluxio 2.x新特性介绍

支持超大规模数据工作负载

- **内嵌式日志以达到高可用性**

Alluxio 2.x去除了ZK/HDFS依赖

- **在Alluxio三个Master内部利用RAFT算法达成共识（ Consensus ）状态**
  - 只有Leading master提交状态变化，Standby masters保持同步
- **优势**
  - 可以采用本地磁盘存储日志（Master节点间作副本）
- **挑战**
  - 性能调优



Alluxio 2.x新特性介绍

更好的存储抽象，实现完全独立和弹性的计算

- **支持跨不同版本的HDFS集群**
  - ✓ 数据的爆炸式增长导致企业通常会拥有许多数据仓库，包括采用跨不同版本的多个Hadoop集群。目前，跨这些集群的统一访问非常困难。使用Alluxio 2.0，用户可以使用Alluxio连接到多个多种版本的HDFS集群，并实现统一的数据访问。

- **与Hadoop主动同步**
  - ✓ 该新功能是与HDFS iNotify进行对接集成，可对存储在Hadoop中的文件所发生的任何数据和元数据更改进行更新，允许通过Alluxio访问数据的应用程序能够主动接收最新更新。

Alluxio 2.x新特性介绍

对机器学习、数据查询等系统更强的支撑

- **支持在任意存储上运行机器学习和深度学习工作负载**
  - ✓ 机器学习和深度学习框架往往需要从Hadoop或对象存储中提取大规模数据，这通常是手动且非常耗时的过程。
  - ✓ Alluxio的FUSE功能支持POSIX兼容的API，因此通过Alluxio，TensorFlow、Caffe等框架以及其他基于Python的模型可以使用传统文件系统的访问方式直接访问任何存储系统中的数据。

- **与结构化数据管理与查询系统进行深度整合**
  - ✓ 在Alluxio层面提供Catalog Service，提供了对结构化数据的抽象，添加Hive MetaStore到Alluxio中就像挂载一个文件系统。
  - ✓ Alluxio感知文件和对象的数据存储结构和模式(schema)，从而更好地提供服务，提供了Alluxio Data Transformation服务，例如：
    - ✓ 自动将CSV格式的文件转成Parquet格式
    - ✓ 将很多小的表文件整合成大文件，减少查询耗时等

Direction: Structured Data Service

- **Alluxio Catalog Service (Target 2.1)**
  - Serve Metadata of Tables (like Hive Meta Store)
    - Highly Efficient by using Apache Iceberg (e.g., no slow dir listing)
  - Speed up query planning, independent of speeding up by caching files in Alluxio File System

- **Alluxio Connector for Presto (Target 2.1)**
  - Presto connects to Alluxio directly without Hive Connector
  - Enable push downs to Alluxio layer

Direction: Alluxio on Kubernetes

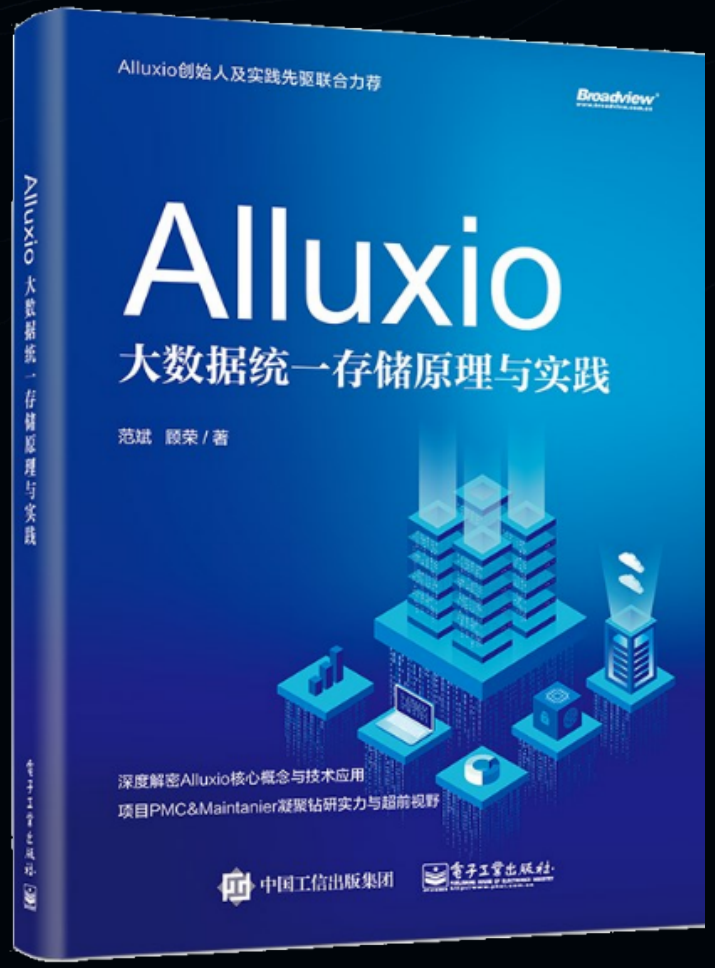
- Call for Community Contribution!
- **Productionize Helm Chart**
  - https://github.com/Alluxio/alluxio/issues/9616
- **K8S csi-driver/provisioner**
  - https://github.com/Alluxio/alluxio/issues/9599
- **Alluxio K8S Operator**

Direction: File System and Cloud Integration

- **Automatic & Transparent Caching (Target 2.1)**
  - Use Alluxio as a caching layer for Presto, Spark or Hive without modifying HMS
  - https://github.com/Alluxio/alluxio/issues/9231

- **AWS/GCP Integration**
  - Improve EMR bootstrap script
  - Images on AWS / GCP marketplace

新出版的Alluxio中文书籍



《Alluxio：大数据统一存储原理与实践》

范斌 **顾荣**/著
出版社：电子工业出版社. 出版时间：2019年8月
ISBN: 978-7-121-36782-3. 字数：242千字

国内首本大数据存储系统Alluxio书籍

TalkingData

欢迎加入Alluxio开源社区！

www.alluxio.org

扫描关注丰富的Alluxio中文技术材料与案例

TalkingData

基于Presto+Alluxio的adhoc查询方案在网易游戏的实践

基于Presto+Alluxio的adhoc查询方案在网易游戏的实践

Alluxio助力华尔街大型量化基金公司提升模型处理效率

量化对冲基金公司使用Alluxio提升机器学习模型处理速度近4倍！

Alluxio提升对象存储平台Hitachi Content Platform的访问性能

日立数据系统公司（日立旗下子公司）的Hitachi Content Platform (HCP)基于Alluxio构建私有云基础架...

Alluxio加速腾讯云EMR服务

Alluxio作为大数据领域存储层解决方案，专注于使用分布式的内存存储技术，为大数据分析和机器学习等围绕...

美国领先营销公司BV基于Alluxio分层存储10倍提速AWS S3上的S...

美国领先营销公司BV基于Alluxio分层存储10倍提速AWS S3上的Spark和Hive作业

这家神秘的东南亚最大银行如何深度拥抱大数据创新？

新加坡发展银行（DBS）举办技术工程创新活动专访Alluxio创始人李浩源博士！

案例分析：联想利用Alluxio分析多位置来源的PB级智能手机数...

本文介绍了联想利用Alluxio分析多位置来源的PB级智能手机数据并消除ETL的案例。

案例分享：FineBI基于alluxio实现亿级热数据高性能分析

FineBI基于alluxio实现亿级热数据高性能分析

腾讯案例研究：使用Alluxio为每月超过1亿用户提供个性化新闻

本文介绍使用Alluxio为每月超过1亿用户提供个性化新闻的腾讯案例研究。

陌陌:使用Spark SQL和Alluxio加速Ad Hoc查询

本文介绍了陌陌使用Spark SQL和Alluxio加速Ad Hoc查询的实践案例。

Alluxio助力中国联通构建高效Spark计算服务平台

Alluxio助力中国联通构建高效Spark计算服务平台！

Arimo利用Alluxio的内存能力提升深度学习模型的结果效率(Ti...

本文介绍了Arimo如何利用Alluxio的内存能力来提升深度学习模型的效率。

Microsoft Azure云平台上的TensorFlow：通过Alluxio启用Bl...

Microsoft Azure云平台上的TensorFlow：通过Alluxio启用Blob存储服务

干货 | ALLUXIO在携程大数据平台中的应用与实践

进入大数据时代，实时作业有着越来越重要的地位...

ArcGIS与Alluxio - 使用Alluxio提高ArcGIS的数据访问能力，加速对...

本文讲述了使用Alluxio提高ArcGIS的数据访问能力，加速对数据的理解与分析！

助力存储成本优化，京东、陌陌、TalkingData共同探讨Alluxio...

Alluxio在京东、陌陌、TalkingData的使用分享！

去哪儿网：使用Alluxio（前Tachyon）实现300倍提升

概述互联网公司同质应用服务竞争日益激烈，业务部门亟需利用线上实时反馈数据辅助决策支持以提高服务水...

百度案例：使用Alluxio提速数据查询30倍

百度案例：使用Alluxio提速数据查询30倍

变不可能为可能：巴克莱银行使用Alluxio提速小时级Spark作业...

巴克莱银行使用Alluxio提速小时级Spark作业至秒级，变不可能为可能！

Myntra案例分析：为定制化的移动电子商务加速云上分析

本文介绍了印度领先的电子商务服装零售商Myntra使用Alluxio的案例！

搜狗实战案例：基于Alluxio优化Spark Shuffle性能

搜狗基于Alluxio提升Spark大规模Shuffle的性能和稳定性

Alluxio应用实践

本文主要介绍了Alluxio分布式内存文件系统，及其在QueryEngine Spark S...

HashData实战案例：使用Alluxio构建云原生分析型MPP数据库

Alluxio数据编排层为MPP数据库高效运行在云对象存储上保驾护航！

TalkingData

顾荣
gurong@nju.edu.cn

2019

TalkingData