

Pandora 大数据平台

XSpark

崔文正



XSpark 数据分析平台

- ❖ 常规大数据组件和XSpark的背景
- ❖ XSpark 的系统架构和功能特点
- ❖ 基于XSpark的应用和实践

常规大数据组件

数据可视化



批量/实时计算



集群调度



Yarn



存储



HDFS

监控



Prometheus



influxdata



Grafana



Zookeeper

高可用

为什么做XSpark

- ❖ 大数据组件必须配套成体系才能发挥作用
- ❖ 搭建一整套可靠的大数据分析平台很繁琐、困难
- ❖ 运维大数据组件稳定性需要丰富的经验和很大的人力投入
- ❖ 存储和计算成本很高

为什么做XSpark

XSpark 定位是什么？

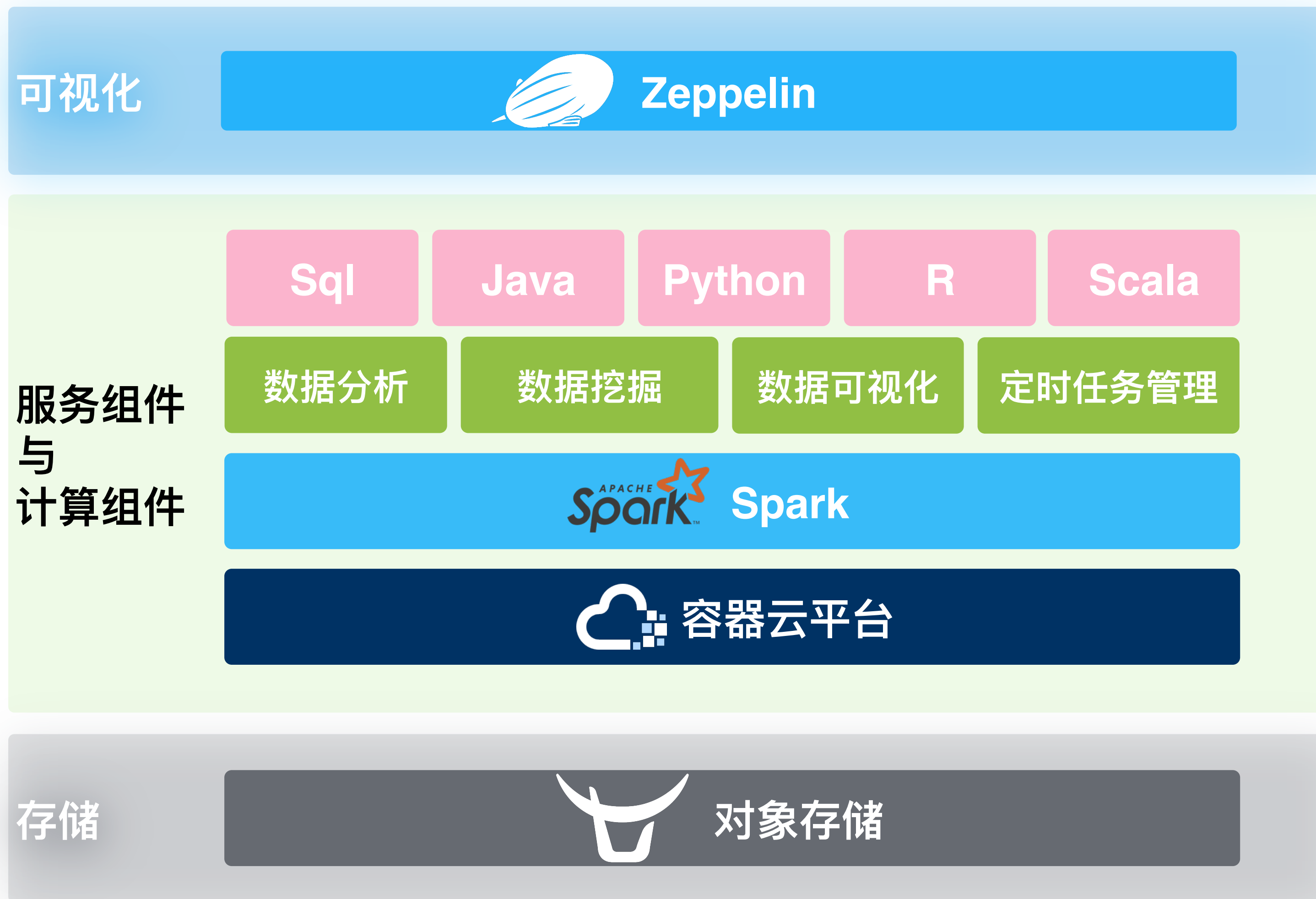
- ❖ 从存储到数据可视化，全栈的大数据分析产品
- ❖ 基于容器云平台，用户可以自助操作一键创建使用
- ❖ 配套集群监控、管理可视化工具，能有效降低运维成本
- ❖ 集成优秀社区组件，优化并能做的更好

XSpark 数据分析平台

- ❖ 常规大数据组件和XSpark的背景
- ❖ **XSpark 的系统架构和功能特点**
- ❖ 基于XSpark的应用和实践

XSpark架构

XSpark 的系统架构



XSpark 云存储

1. 扩展性

- ❖ 基于云存储，根据用量动态扩充

2. 稳定性

- ❖ 数据可靠性不低于 99.99999999%

3. 高性能

- ❖ 比HDFS平均高30%-50%的性能

4. 低成本

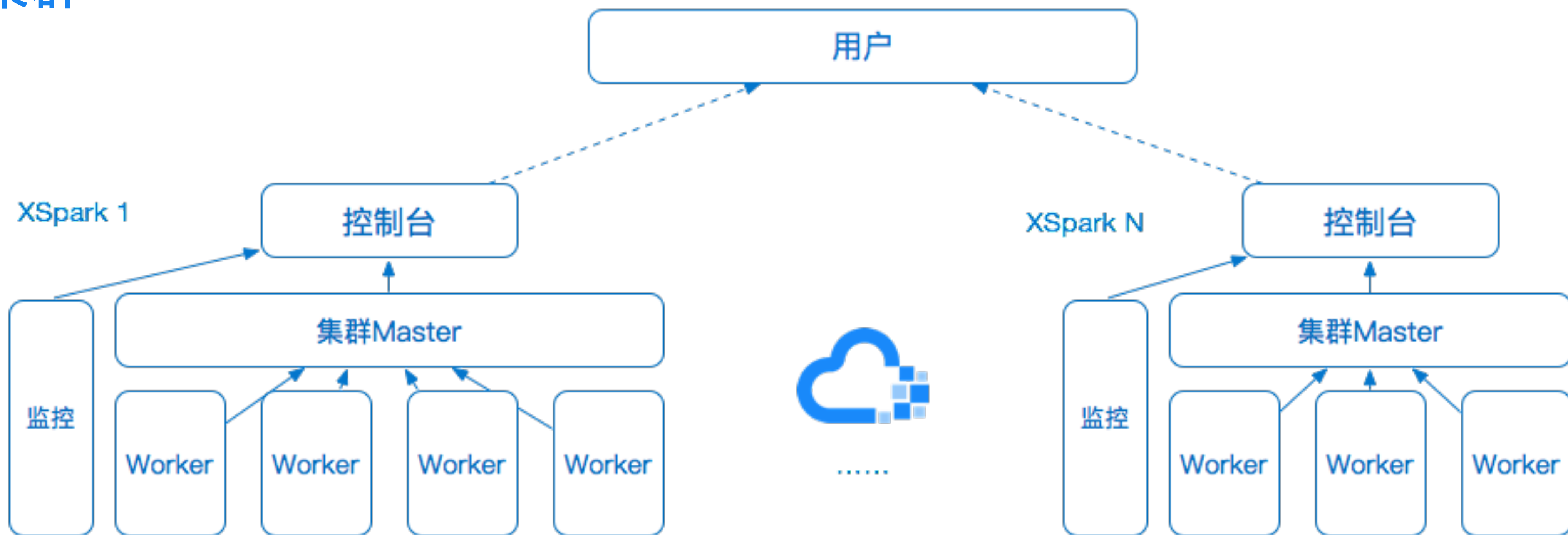
- ❖ 和HDFS 1.14:3 的副本数量

HDFS对比

存储对比

| | 对象存储 | HDFS |
|----------|--------------|--------------|
| 存储容量 | 动态扩容 | 集群规模上限 |
| 可靠性 | 10个9 | <u>约6个9</u> |
| inode 数量 | 接近无限 | NameNode内存上限 |
| 备份数量(成本) | 1.14 | 3备份 |
| 安全性 | 多租户隔离 | 共享集群 |
| 文件读取特点 | 完全兼容HDFS文件格式 | |
| 读写性能 | 单节点1Gbps | 单节点670Mbps |
| 局限性 | 文件夹操作较重 | 小文件灾难 |

XSpark 容器云集群



XSpark 控制台



XSpark Service Stack

| Service名称 | Container总量 | 状态 |
|-----------------------|-------------|----|
| spark-master-service | 1 | ▶ |
| spark-worker-service | 1 | ▶ |
| spark-monitor-service | 1 | ▶ |

开始使用

集群监控

扩容缩容

 更改配置

| | | |
|------|-------------------------------------|----------|
| 配置规格 | spark-master: 四核(CPU), 8GB(内存) * 1 | SSD1_20G |
| | spark-worker: 二核(CPU), 4GB(内存) * 1 | SSD1_10G |
| | spark-monitor: 一核(CPU), 1GB(内存) * 1 | SSD1_10G |

集群指令:

重启XSpark

磁盘清理

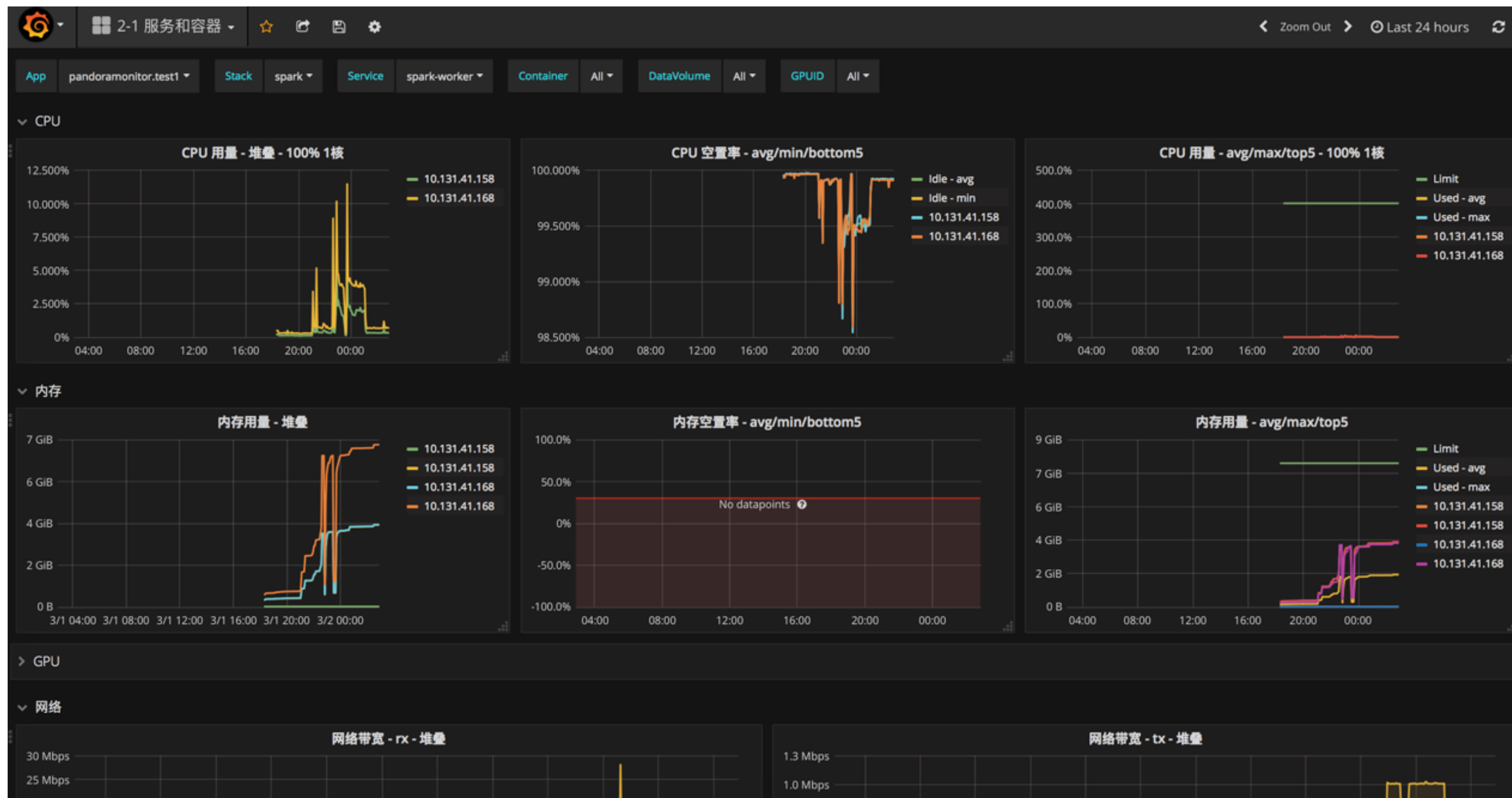


Apache Spark 集群

集群状况

访问SparkUI

XSpark 监控



容器云平台

1. 高效性

- ❖ 分钟级创建集群，秒级扩容缩容
- ❖ 镜像智能预热，减少等待时间

2. 成本低

- ❖ 多调度策略，可以充分使用物理资源
- ❖ 按需使用，需时扩容，闲时缩容

3. 维护性

- ❖ 一站式监控平台

XSpark 开源组件

1. Spark

- ❖ 性能优化，结合云存储特点，对元数据操作进行加速
- ❖ 优化集群调度方案

2. Zeppelin

- ❖ 解决死锁、资源无法释放等众多功能问题
- ❖ 易用性：增加邮件提醒、一键重启等丰富功能

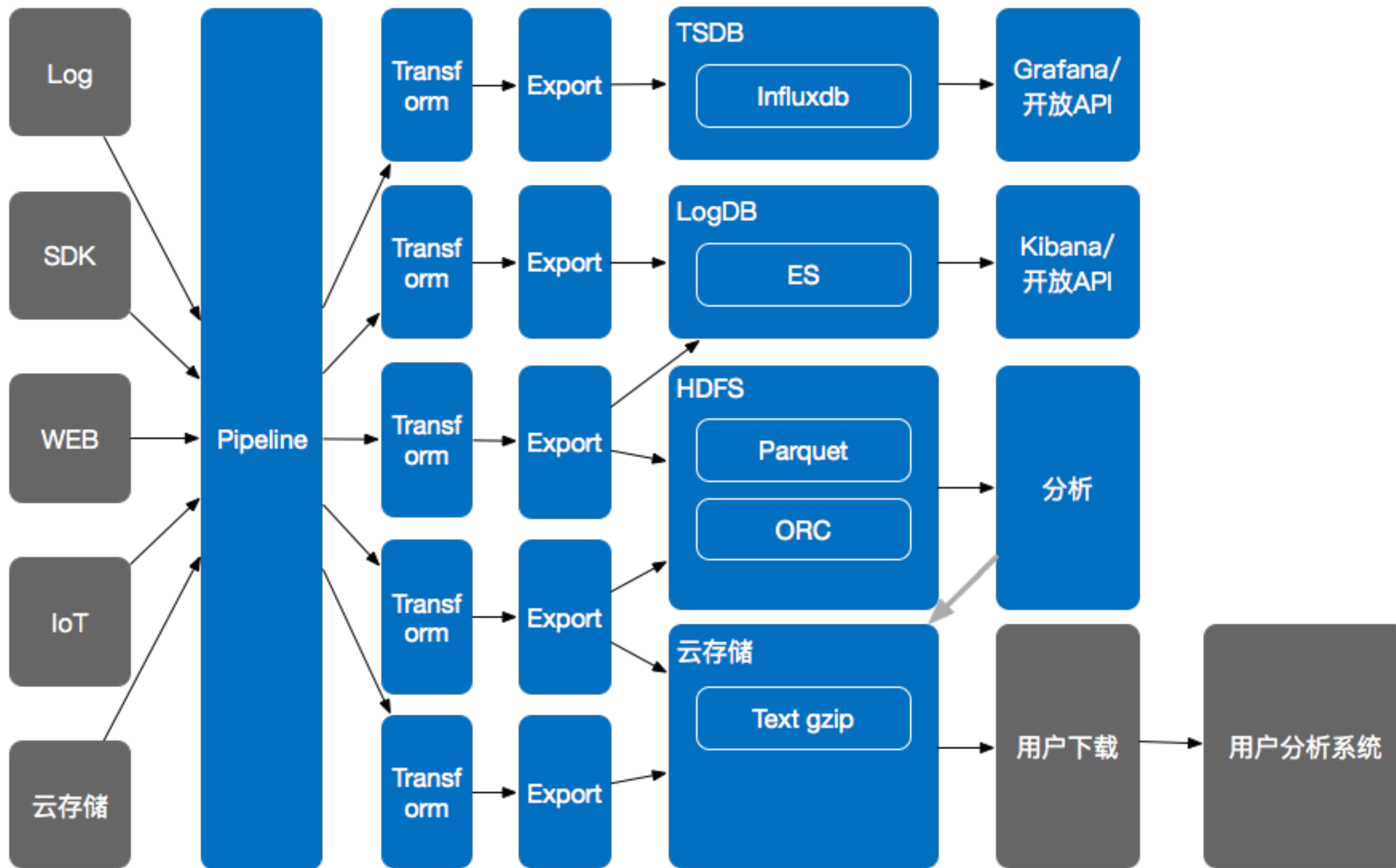
XSpark 数据分析平台

- ❖ 常规大数据组件和XSpark的背景
- ❖ XSpark 的系统架构和功能特点
- ❖ **基于XSpark的应用和实践**

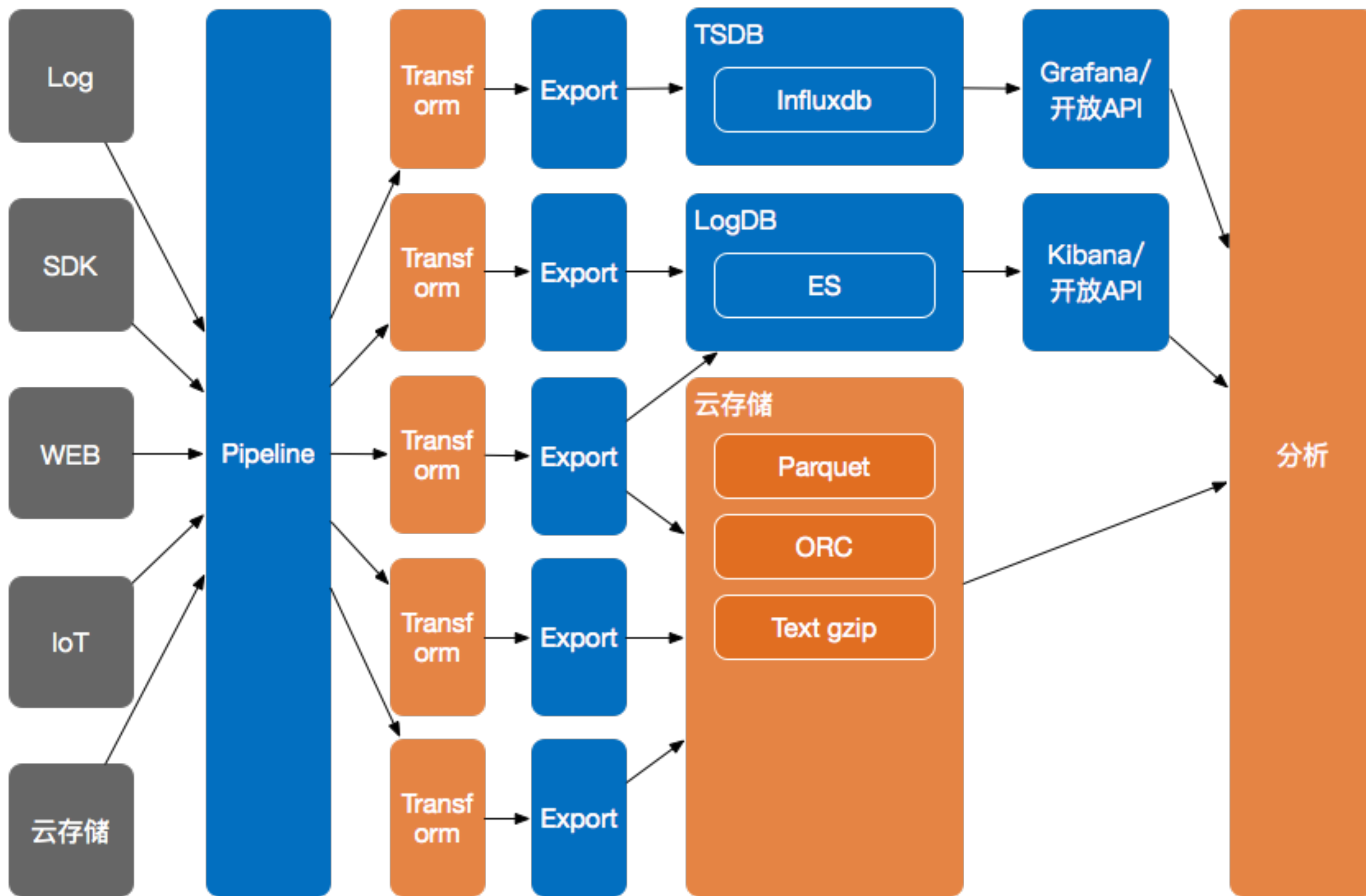
基于XSpark的应用和实践

- ❖ Pandora 架构升级, 产品、技术优化
- ❖ CDN 日志自助分析
- ❖ 机器学习的一个小实例

应用和实践 Pandora 架构升级



应用和实践 Pandora 架构升级



应用和实践 Pandora 架构升级

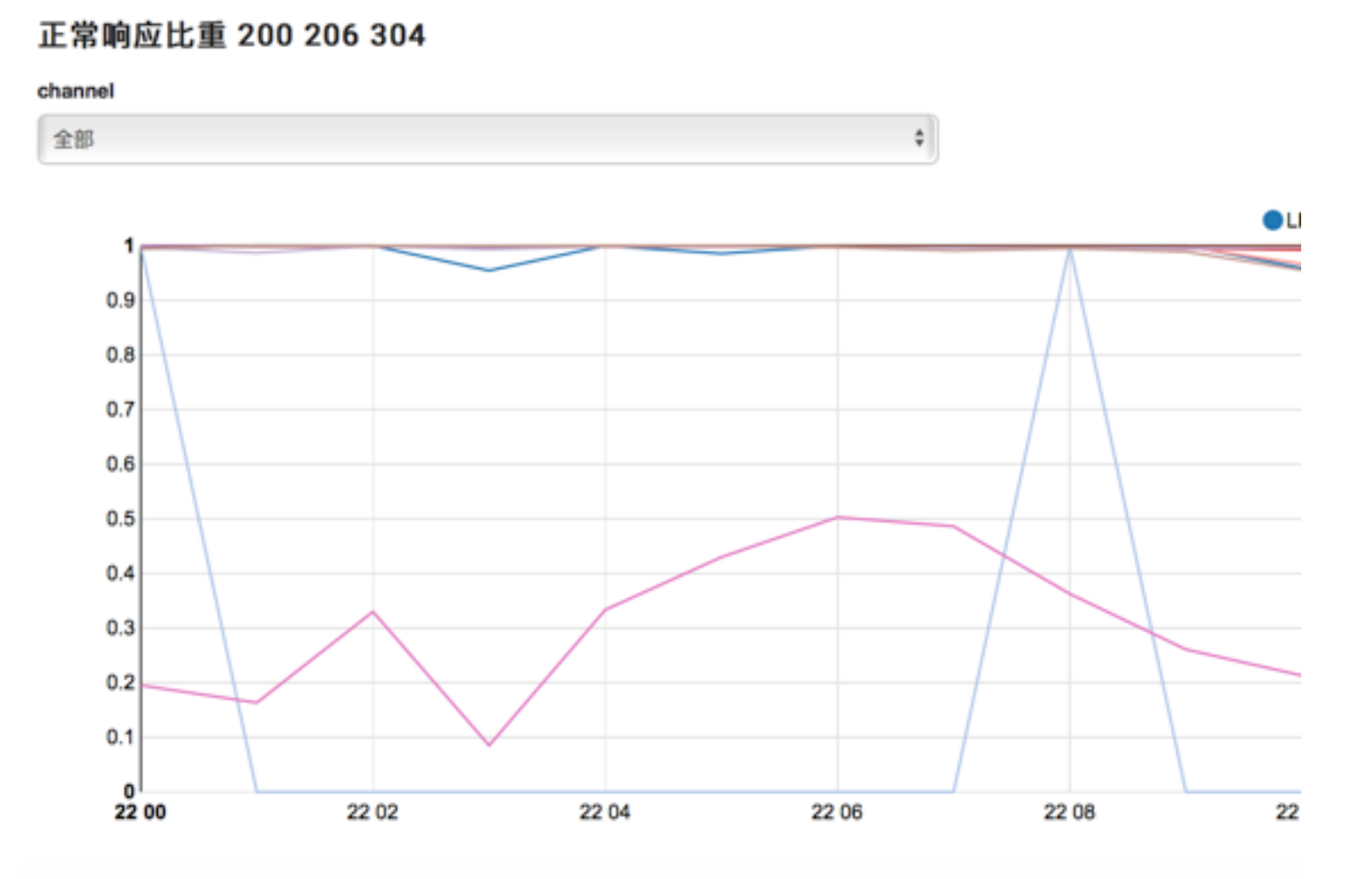
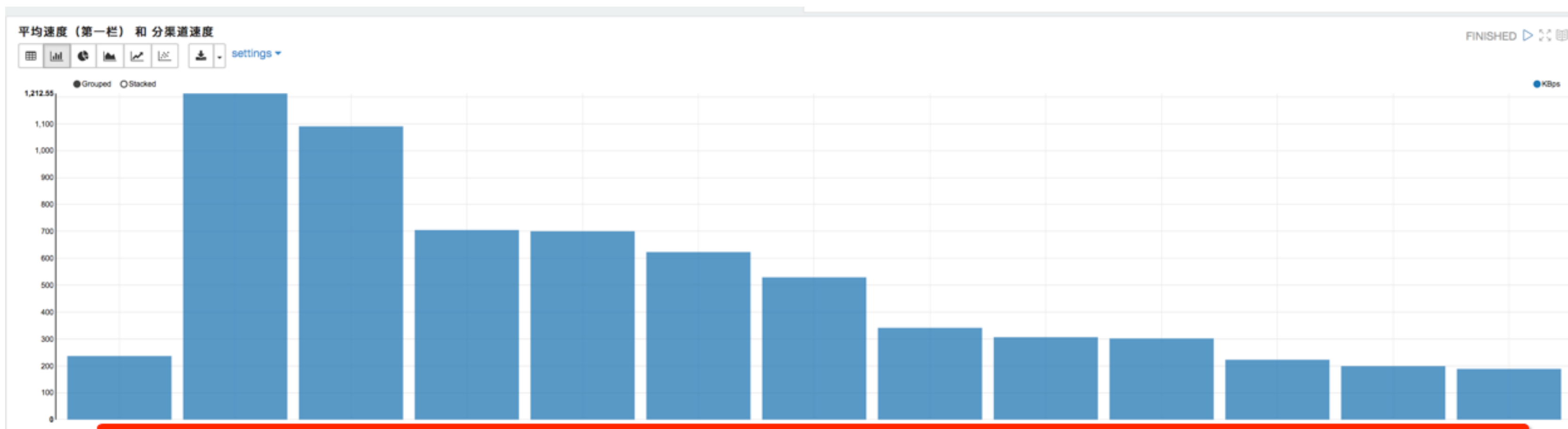
XSpark 带来的收益

- ❖ 简化了用户的分析数据场景，无需下载，直接云上大数据分析
- ❖ 云存储全面替代HDFS，性能、存储量和运维性全面提升
- ❖ 全部容器化，轻量级集群，充分利用物理资源
- ❖ 支持Pandora所有大数据产品开放接口，所有数据分析一站式

基于XSpark的应用和实践

- ❖ Pandora 架构升级，产品、技术优化
- ❖ **CDN 日志自助分析**
- ❖ 机器学习的一个小实例

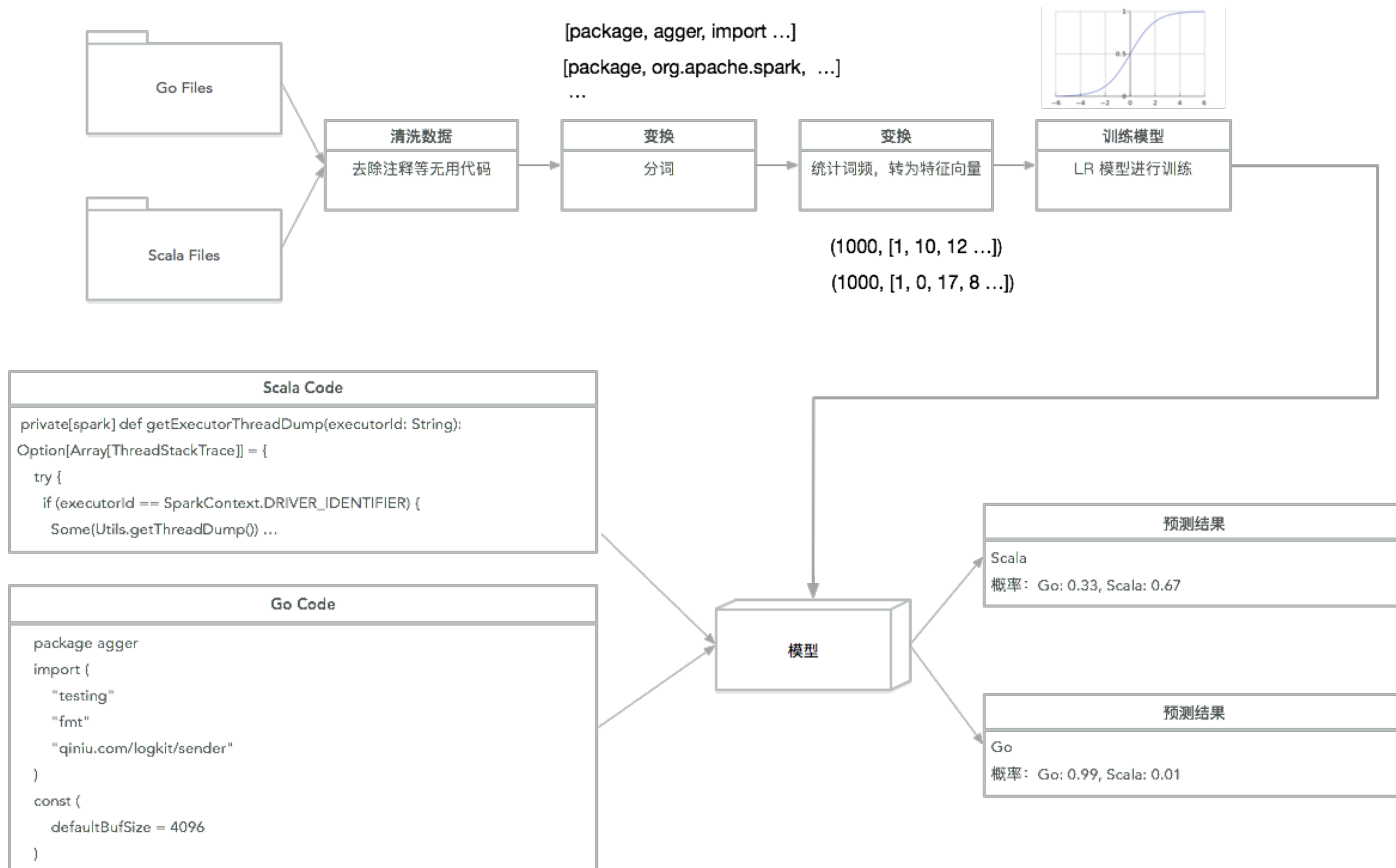
CDN日志自助分析



机器学习小实践

- ❖ Pandora 架构升级，产品、技术优化
- ❖ CDN 日志自助分析
- ❖ 机器学习实践，通过代码内容预测使用语言

机器学习小实践



谢谢!

