Alibaba Group
阿里巴巴集团

# 阿里巴巴-开放数据处理服务

## (Open Data Processing Service, ODPS)

数据平台事业部 - 余 波

# 提纲

- **背景与概况**
- 服务架构
- 关键技术
- 服务管理
- 结语

# 背景-业务场景

- 海量数据处理和分享需求
  - 交易数据、日志数据
  - 语音、图像数据
  - 数据的交换和融合

- 典型数据业务
  - 信用贷款
  - 广告CTR

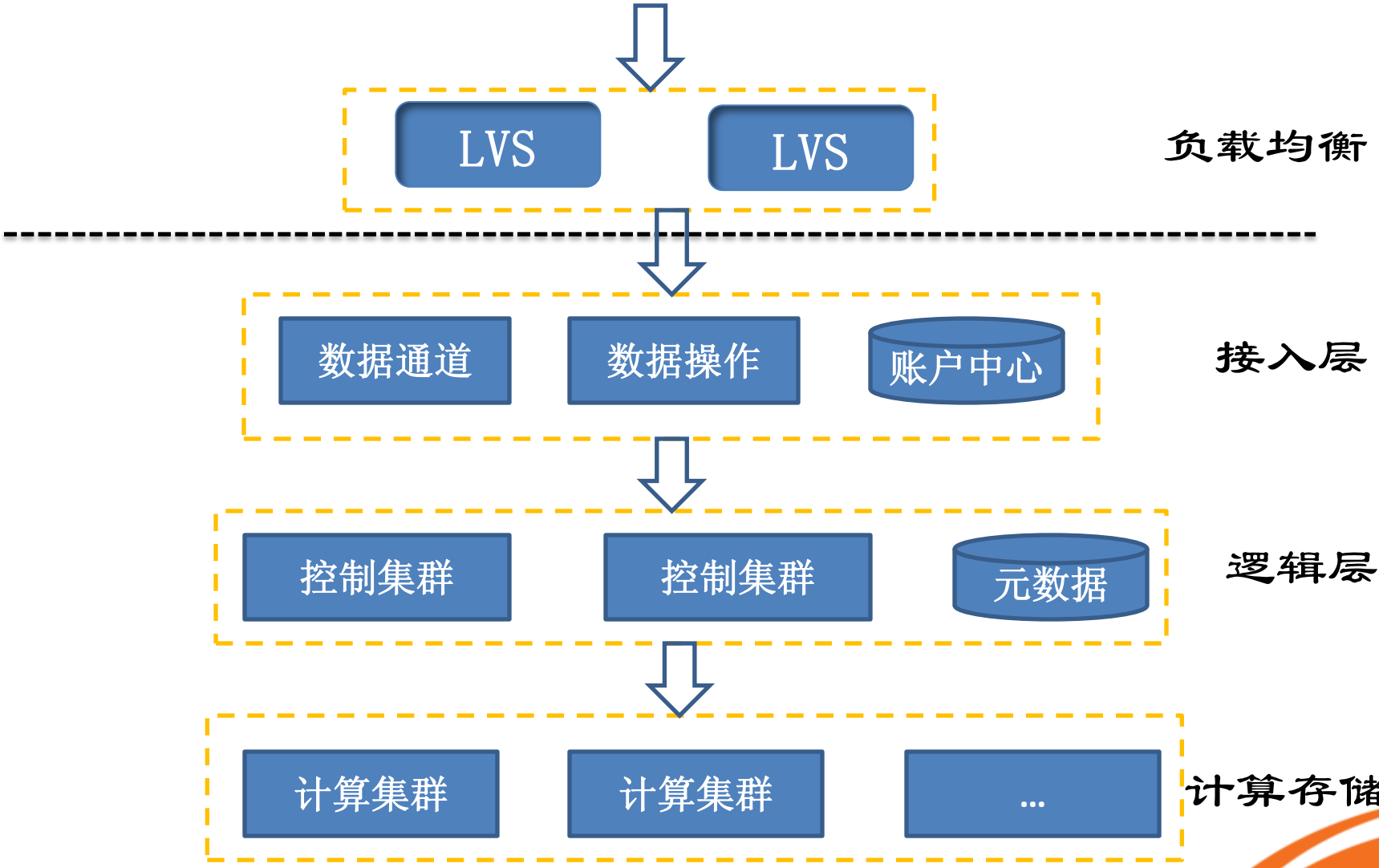
# 背景-技术需求

- 计算能力
- 水平扩展
- 丰富的处理手段
- 服务化
- 安全机制
- 可运维、可管理
- 稳定性

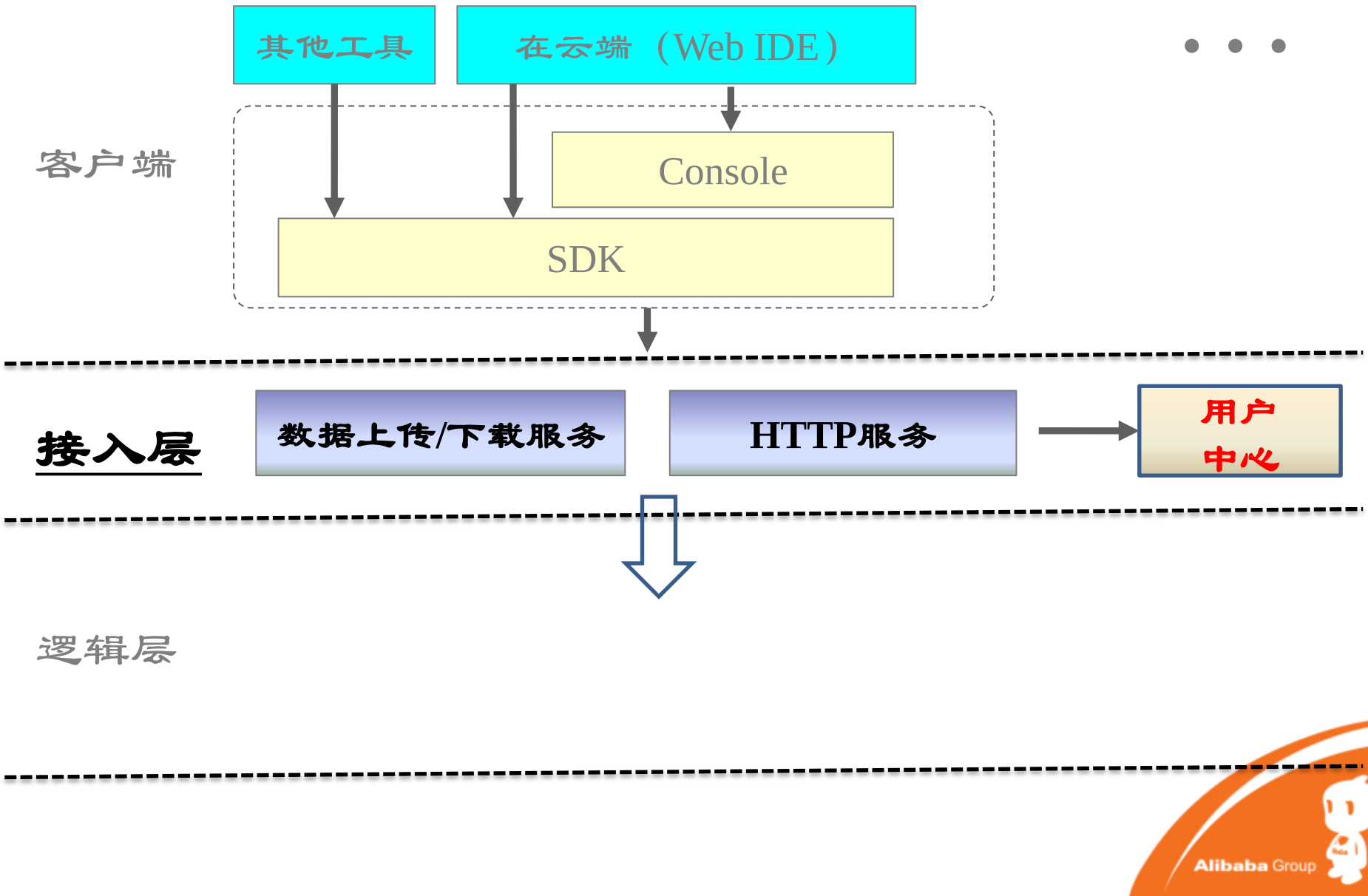
# 提纲

- 背景与概况
- **服务架构**
  - 接入层
  - 逻辑层
  - 存储与计算层
- 关键技术
- 服务管理
- 结语

# 服务架构-整体架构


LVS
LVS
负载均衡
数据通道
数据操作
账户中心
接入层
控制集群
控制集群
元数据
逻辑层
计算集群
计算集群
...
计算存储层

# 服务架构-接入层架构


其他工具
在云端（Web IDE）
客户端
Console
SDK
接入层
数据上传/下载服务
HTTP服务
用户中心
逻辑层

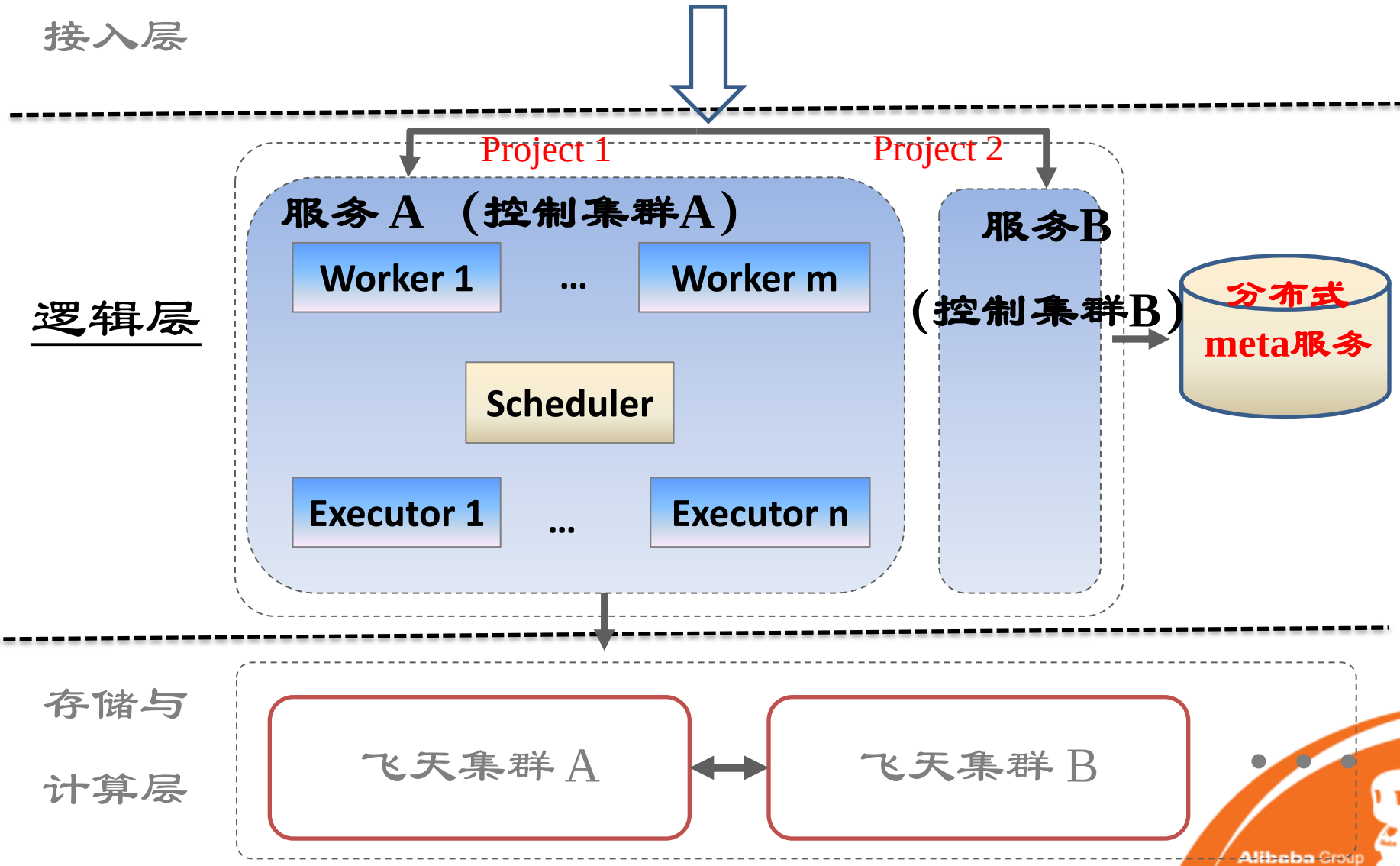
# 服务架构-接入层

- 功能设计
  - 用户认证
  - RESTful API
  - 无状态、水平扩展
- 资源实体
  - Project
    - Table/Partition，数据集合
    - UDF/Resource，文件，jar包，py脚本
    - Job/Instance，抽象可执行实体和运行实例
  - User/Role，用于管理用户对Project内实体的访问控制和授权

# 服务架构-逻辑层功能

- 用户权限管理
- 多个任务执行时序的控制
- 单个任务内部逻辑实现
  - 简单操作的执行
  - 生成飞天作业
- 计算集群的管理

# 服务架构-逻辑层架构


接入层
Project 1
Project 2
逻辑层
服务 A （控制集群A）
Worker 1
...
Worker m
Scheduler
Executor 1
...
Executor n
服务B
（控制集群B）
分布式
meta服务
存储与
计算层
飞天集群 A
飞天集群 B

# 服务架构-逻辑层分析

- Worker/Executor
  - 线性可扩展，负载均衡，无状态
- Scheduler
  - 只维护一组运行实例
- 双ODPS服务
  - 灰度发布，不停服务轮转升级，failover
- 分布式meta服务
  - 使用阿里云OTS分布式存储系统，无需担心空间不够
  - 统一名称空间，双服务和多飞天集群对用户透明

# 服务架构-存储计算层

- 多个飞天集群组成
  - 支持跨集群（机房）数据共享
- 存储
  - 使用盘古分布式文件系统
    - Master-Slave结构
    - 基于Paxos的多Master，故障恢复小于一分钟
    - 文件分块(Chunk)，每块存三份，分布在不同机架
  - 表数据采用统一文件格式：CFile，基于列存储的压缩文件格式
  - 提供数据上传和下载服务，支持PB/天的吞吐量
- 计算
  - 支持多种计算模式：SQL，MR，算法库，图计算（Pregel）
  - 采用伏羲作业，支持DAG，支持基于CPU/MEM的资源调度

# 提纲

- 背景与概况
- 服务架构
- **关键技术**
  - 分布式问题
  - 多集群方案
  - 编程模型
- 服务管理
- 结语

# 关键技术 —分布式问题

- **机器当机**

  各个角色都会当机，包括同时当机。

- **底层系统不稳定**

  依赖的底层系统性能、功能会出现不稳定。

- **时序问题**

  交互过程网络抖动引起的时序混乱。

- **规模问题**

  大规模导致的性能瓶颈。

- **版本升级**

  不同版本的需求和热升级。

# 关键技术 —多集群方案（1）

## ➢ 要解决的问题

- 业务快速增长，单集群扩容受机房容量、飞天规模限制

## ➢ 技术难点

- 数据存储和计算如何划分
- 数据动态变化，需要保证数据读取正确性
- 跨机房带宽如何使用
- 对用户透明

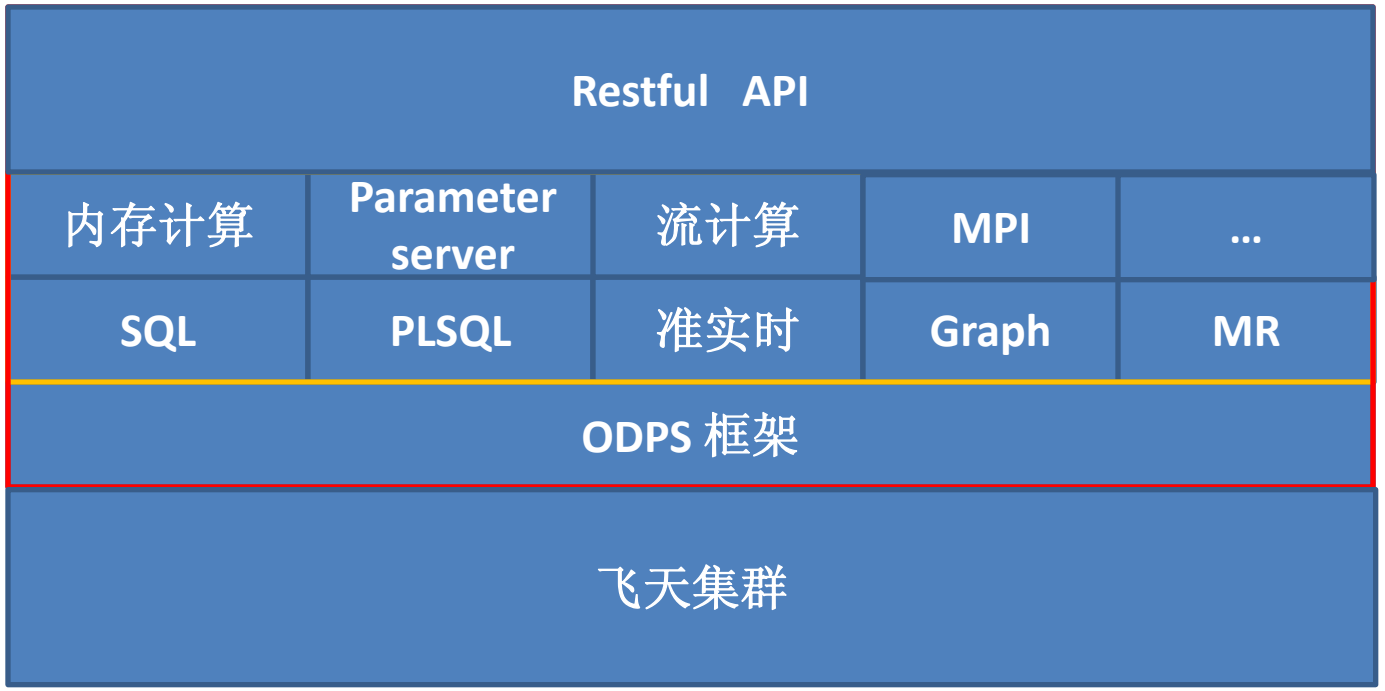
# 关键技术-多集群方案（2）

- 按业务划分集群
  - 关系密切业务的project放在同一个集群
  - 每个project对应一个默认集群，作业总是跑在默认集群上
- 数据版本
  - 同一份数据（表或分区）在多个集群上可能具有不同的版本

```
{"LatestVersion":V1,"Status":{"ClusterA":"V1","ClusterB":"V0"}}
```

  - 当一份数据版本更新后，触发一个跨集群数据复制任务
- 跨集群数据复制
  - 表或分区可以配置是否进行跨集群复制（自动或手工）
  - 流控，优先级
- 直读直写，应对新的跨集群数据依赖，少量任务

# 关键技术 —编程模型


Restful API
内存计算
Parameter server
流计算
MPI
...
SQL
PLSQL
准实时
Graph
MR
ODPS 框架
飞天集群

# 关键技术 —编程模型

- SQL特性
  - 兼容大部分Hive语法
  - 适应大数据量的处理(T到P级别的数据)
  - 延迟较大
  - 不支持并发、无主键
  - 支持Python和Java写UDF，UDAF，UDTF
  - 物理执行方式：DAG，C++实现
  - Code gen
  - 准实时实现 (Service-Mode)

# 关键技术-编程模型

- Service-Mode
  - 常驻服务，预先申请好worker - 减少调度开销
  - Shuffle数据不落地，直接写网络
  - 假设作业规模m*r，要求r个reduce先起，接收map写的数据
  - 内存文件
  - LLVM，减少编译时间
- 根据SQL类型和数据量动态决定是否采用Service-Mode方式
- 未考虑Failover，主要用于开发project和Adhoc数据分析

# 关键技术-编程模型MPI

- **适用场景**

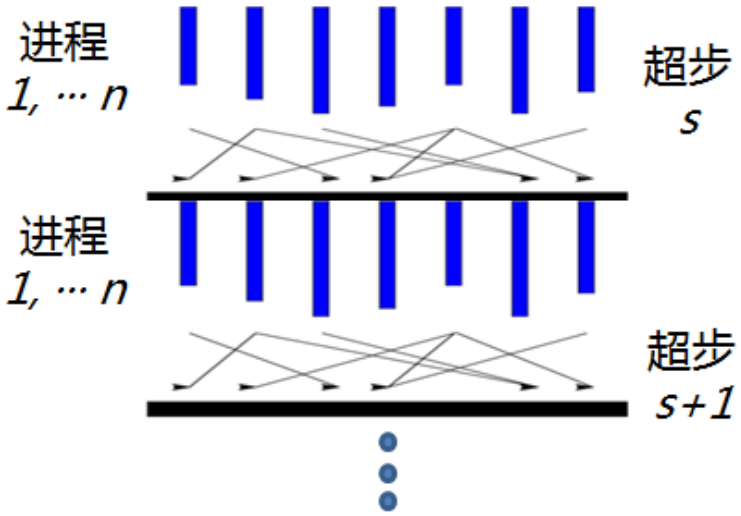
  反复迭代、需要同步类型的大规模机器学习算法。

- **基础算法库**

  逻辑回归、随机森林、贝叶斯、k-means、协同过滤、关联规则、SVD分解等

# 关键技术-编程模型图计算框架

- 海量图结构数据
  - 社交网络（来往，微博），物流信息（菜鸟物流）
  - 电商关系：类目/商品/买家/卖家，交易/浏览
- ODPS图计算框架
  - 支持类似Pregel的Java编程接口，面向迭代类型的作业
  - 磁盘IO→内存网络，换来更快的性能




- 典型应用：
  - PageRank
  - K-均值聚类
  - 非负矩阵分解NMF
  - …
- 算法往往跟业务相关

# 提纲

- 背景与概况
- 服务架构
- 关键技术
- **服务管理**
- **结语**

# 服务管理

- 多租户共享集群
- 基于ACL和Policy的认证授权机制
- 基于project的业务划分
- 基于配额的管理
- 基于历史的优化
- 多种类型计算作业共享集群

# 总结

- 阿里巴巴数据处理服务（ODPS）
  - 支持海量数据的离线存储和计算
  - 以RESTful API的方式提供服务
  - 基于飞天分布式平台
  - 支持跨集群（机房）数据共享
  - 支持SQL、MapReduce、MPI、图计算等编程框架
  - 支持常用的矩阵运算和数据挖掘算法
  - 支持多租户和基于ACL/Policy的权限控制