BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

阿里实时计算

和仲

- 简介
- 模型
- 架构
- 未来

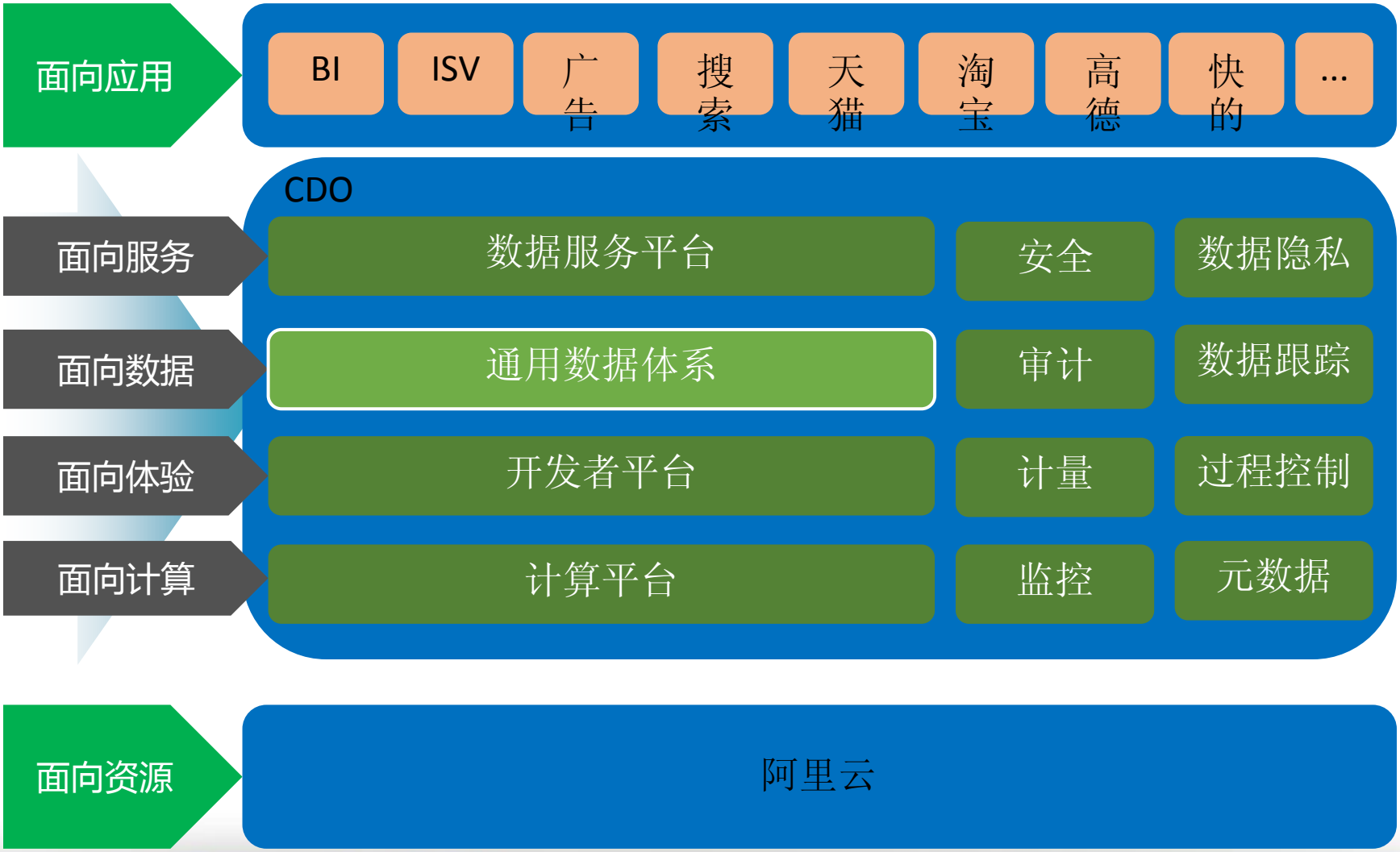
简介

- 花名：和仲
- 姓名：强琦
- 个人介绍：读书的研究方向是机器学习基础理论，毕业后一直从事搜索技术的研发，08年进入阿里后也一直在搜索和广告技术领域，12年加入集团数据平台事业部，致力于打造开放的大数据供应链基础设施平台。对机器学习，分布式计算，搜索广告技术都有浓厚的兴趣。
- 微博：和仲Q

简介


支付宝
小贷
保险
基金
小微金服
淘宝
天猫
聚划算
Alibaba.com
1688.com
Aliexpress
电子商务
菜鸟物流
物流
数字娱乐
高德
友盟
其它领域
数据平台
阿里云
aliyun.com

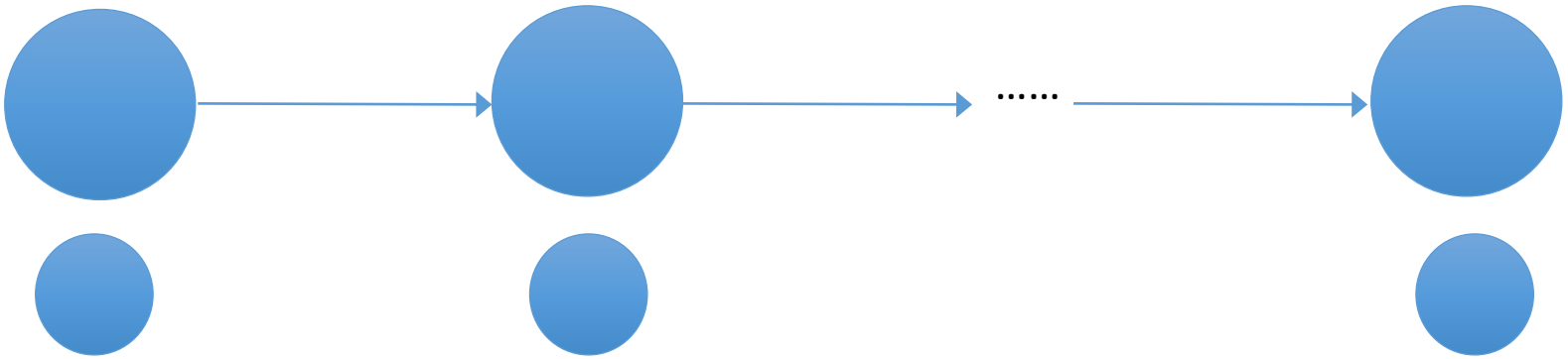
简介


面向应用
BI
ISV
广告
搜索
天猫
淘宝
高德
快的
...
CDO
面向服务
数据服务平台
安全
数据隐私
面向数据
通用数据体系
审计
数据跟踪
面向体验
开发者平台
计量
过程控制
面向计算
计算平台
监控
元数据
面向资源
阿里云

简介

- 实时
  - ✓ 数据的时效性
  - ✓ 计算的时效性
- 计算
  - ✓ 可枚举
  - ✓ 不可枚举
  - ✓ 交互式（增量）
  - ✓ 无状态
  - ✓ 有状态
- 成本模型
  - ✓ 数据复用程度
- Pattern
  - ✓ 预知pattern(数据，计算)
  - ✓ 不可知
- 数据规模
  - ✓ 大数据
  - ✓ “小”数据
- 实时数据的实时计算

模型



假设有N条数据，M个资源，共有n个module。第i个module的吞吐为*OI*,调度的资源数为*Pi*

串行模型的延时为：$\sum_{i=0}^{n-1}\frac{N}{M.O_I}$

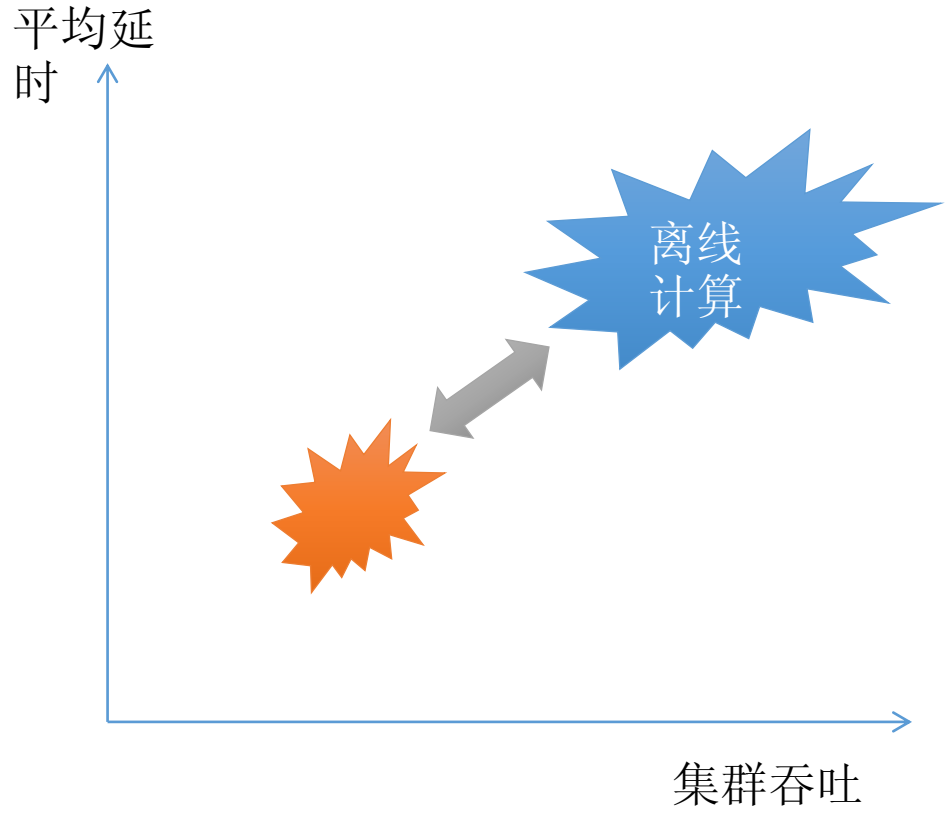
并行模型的延时为：$\frac{N}{P_0O_0}$，满足：$P_0O_0=P_1O_1=P_IO_I$；$\sum_{i=0}^{n-1}P_i$=M。

$$\frac{N}{P_0O_0}=\frac{N\sum_{i=0}^{n-1}P_i}{MP_0O_0}=\sum_{i=0}^{n-1}\frac{N}{M.O_I}$$

模型

串行模型的延时为：$\sum_{i=0}^{n-1} \frac{N}{M.O_I}$

- 优：模型简单；吞吐；
- 劣：数据时效性；倾斜；
- 面向吞吐；兼顾延时



- 优：数据时效；倾斜友好
- 劣：建模复杂；调度复杂
- 面向延时；兼顾吞吐

并行模型的延时为：$\frac{N}{P_0 O_0}$，满足：$P_0O_0=P_1O_1=P_IO_I$；$\sum_{i=0}^{n-1} P_i=\mathrm{M}$

模型

- 增量模型
  - 确定性
  - 可加性
  - 可逆性
- 交互式计算
- 并行DAG

✧ Case

```
t1 = select a, sum(b) as b' from t0 group by a;
t2 = select count(a) from t1 group by b' /10;
```

	粒度	计算	生命周期	容错监控	面向	DAG
全量	Partition/文件/pull	局部	数据处理完，进程”退出”	进程	吞吐	串行
增量(流)	Batch/内存/push	有状态	Keep alive	数据-中间结果不落地	延时	并行

模型

Map

```
void map(GalaxyRecord record);
```

Reduce

```
void reduce(X key, List<Y> values);
```

Merge

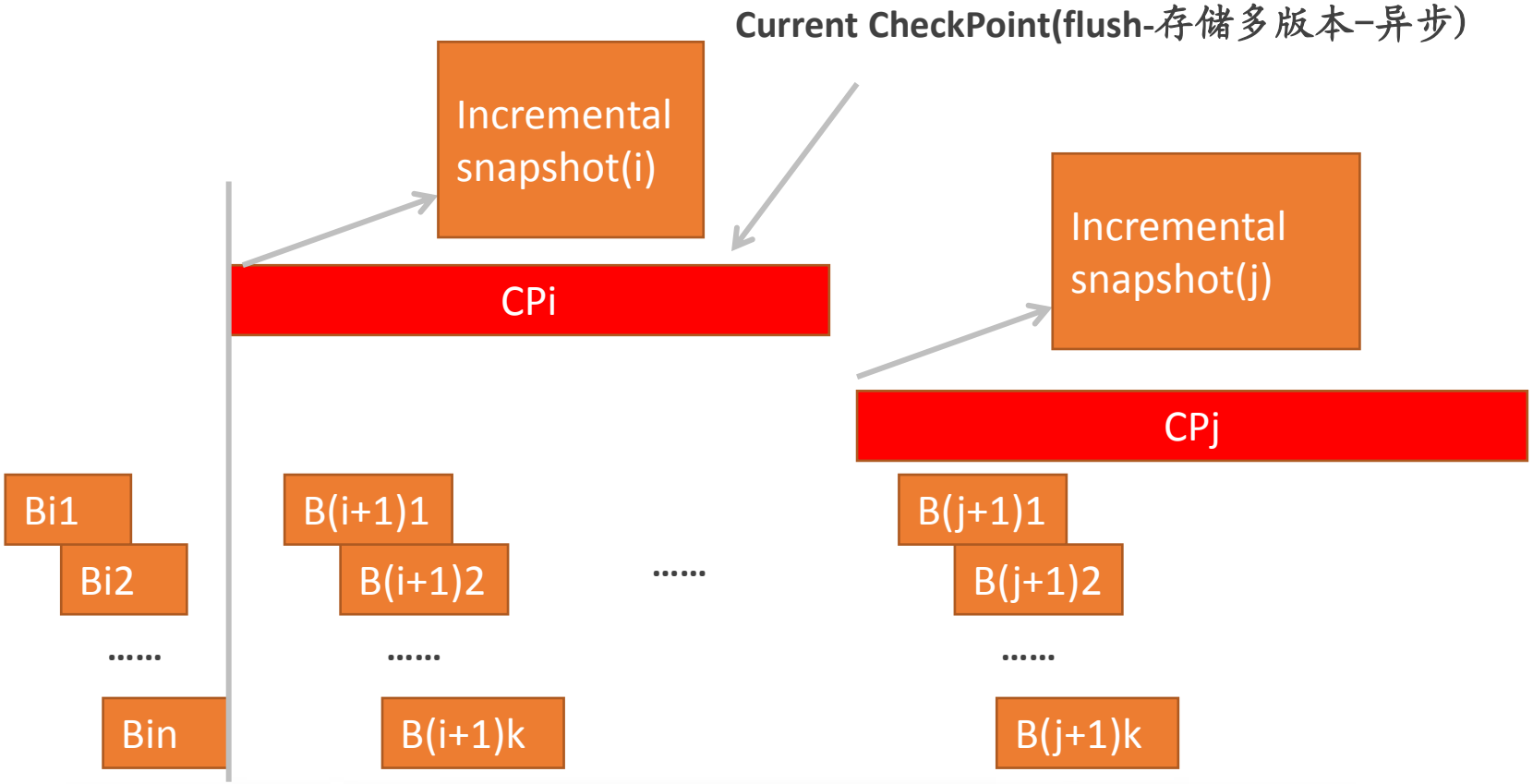
```
T merge(T oldValue, X key, Z value, StateGroup stateGroup);
T rollback(T oldValue, X key, Z value, StateGroup stateGroup);
```

mapOnly, mapreduce, mrm

- 源表
- 目标表
- 纬度表(本地化)
- 临时表（本地化）
- 中间表（本地化）

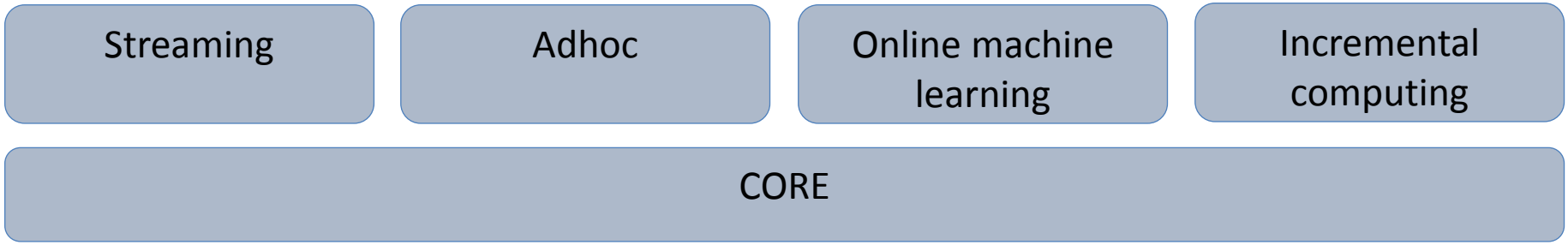
模型

有状态—>全局—> 性能->局部解

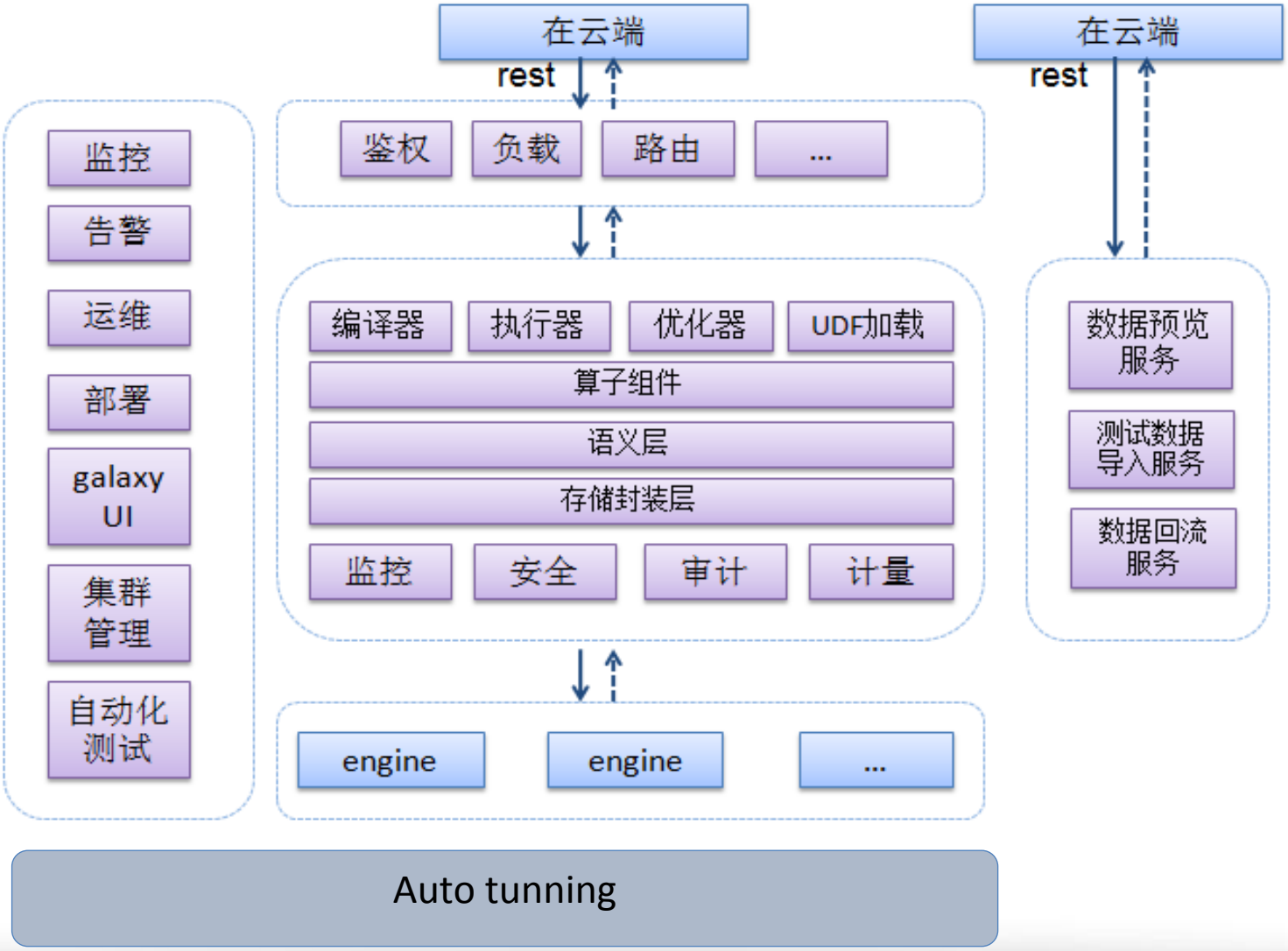


架构

架构


Streaming
Adhoc
Online machine learning
Incremental computing
CORE

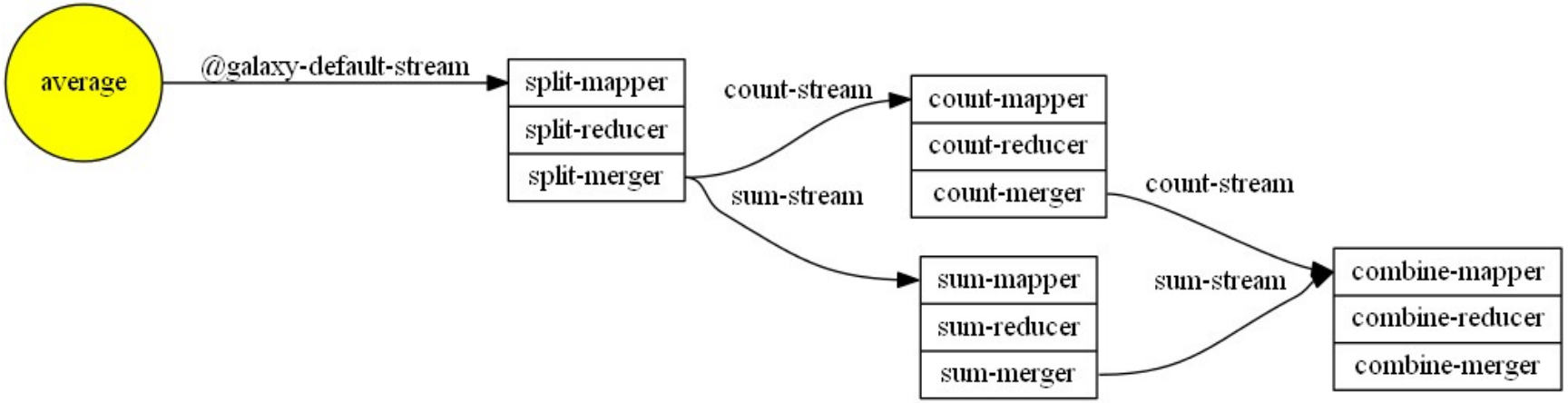
架构

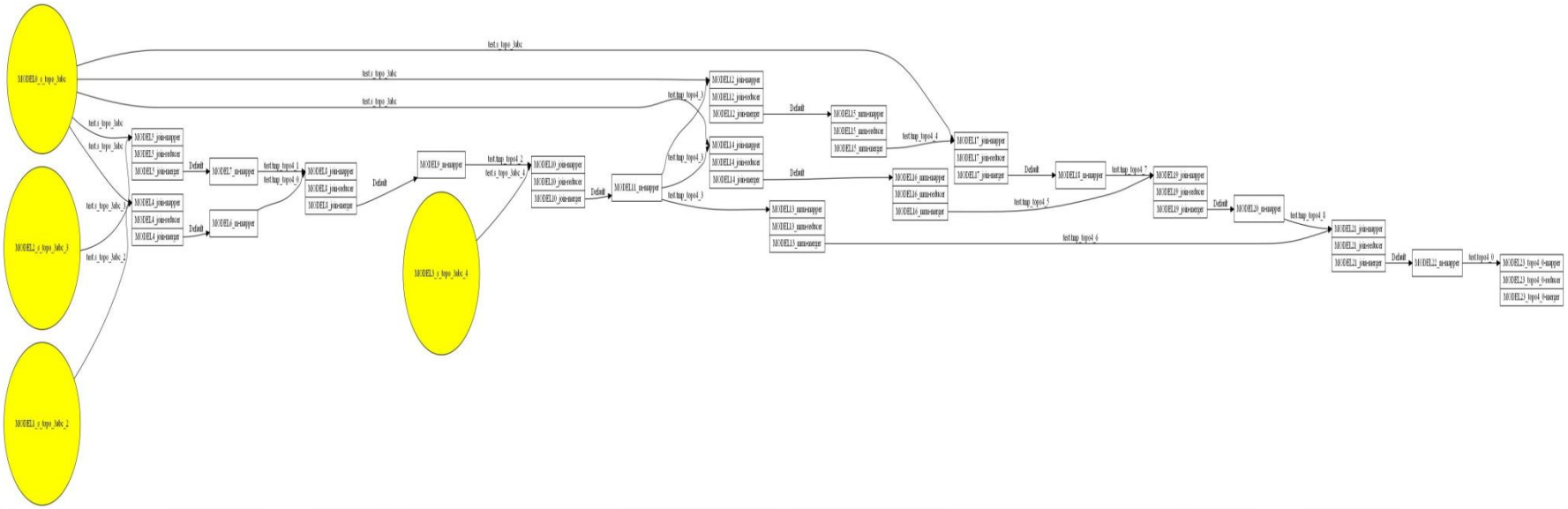

在云端
rest
在云端
rest
监控
告警
运维
部署
galaxy UI
集群管理
自动化测试
鉴权
负载
路由
...
编译器
执行器
优化器
UDF加载
算子组件
语义层
存储封装层
监控
安全
审计
计量
数据预览服务
测试数据导入服务
数据回流服务
engine
engine
...
Auto tunning

架构

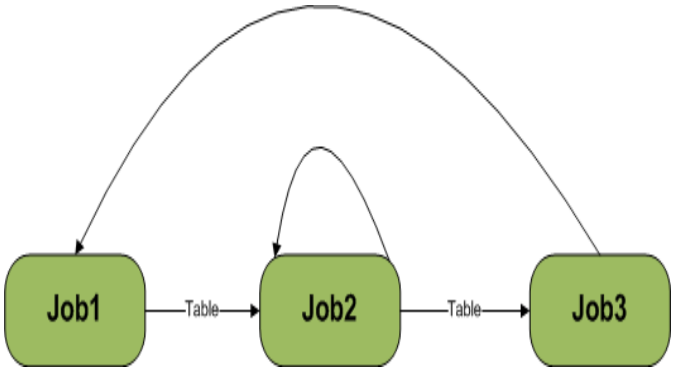
- 原语
  - Map
  - Reduce
  - Shuflle
  - Union
  - Merge
- 高级算子
  - Topk
  - Join
  - _windows

架构


average
@galaxy-default-stream
split-mapper
split-reducer
split-merger
count-stream
sum-stream
count-mapper
count-reducer
count-merger
sum-mapper
sum-reducer
sum-merger
count-stream
sum-stream
combine-mapper
combine-reducer
combine-merger



架构




```
val input_table1=loadTable()
for (i => 1 to 10) {
  val input_table2 =
  job1(input_table1)
  for (j => 1 to 5) {
        input_table2 =
  job2(input_table2)
  }

  table3 = job3(input_table2)
  input_table1 = table3
  table3.checkpoint();
}
table3.output()
```

架构—调度

- 实时调度系统
  - 在线服务调度/隔离
    - Min
    - Max
  - 上云适配
  - 通用运维

架构—分析引擎

- ADS(分析数据库服务）

架构—分析引擎

极速的计算能力

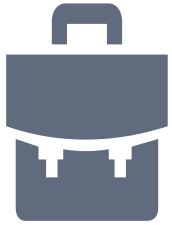
SQL/UDF/Join

自由的查询能力

like/in/contains

智能的优化

列顺序/表顺序/……



分层的安全

列授权/公私钥

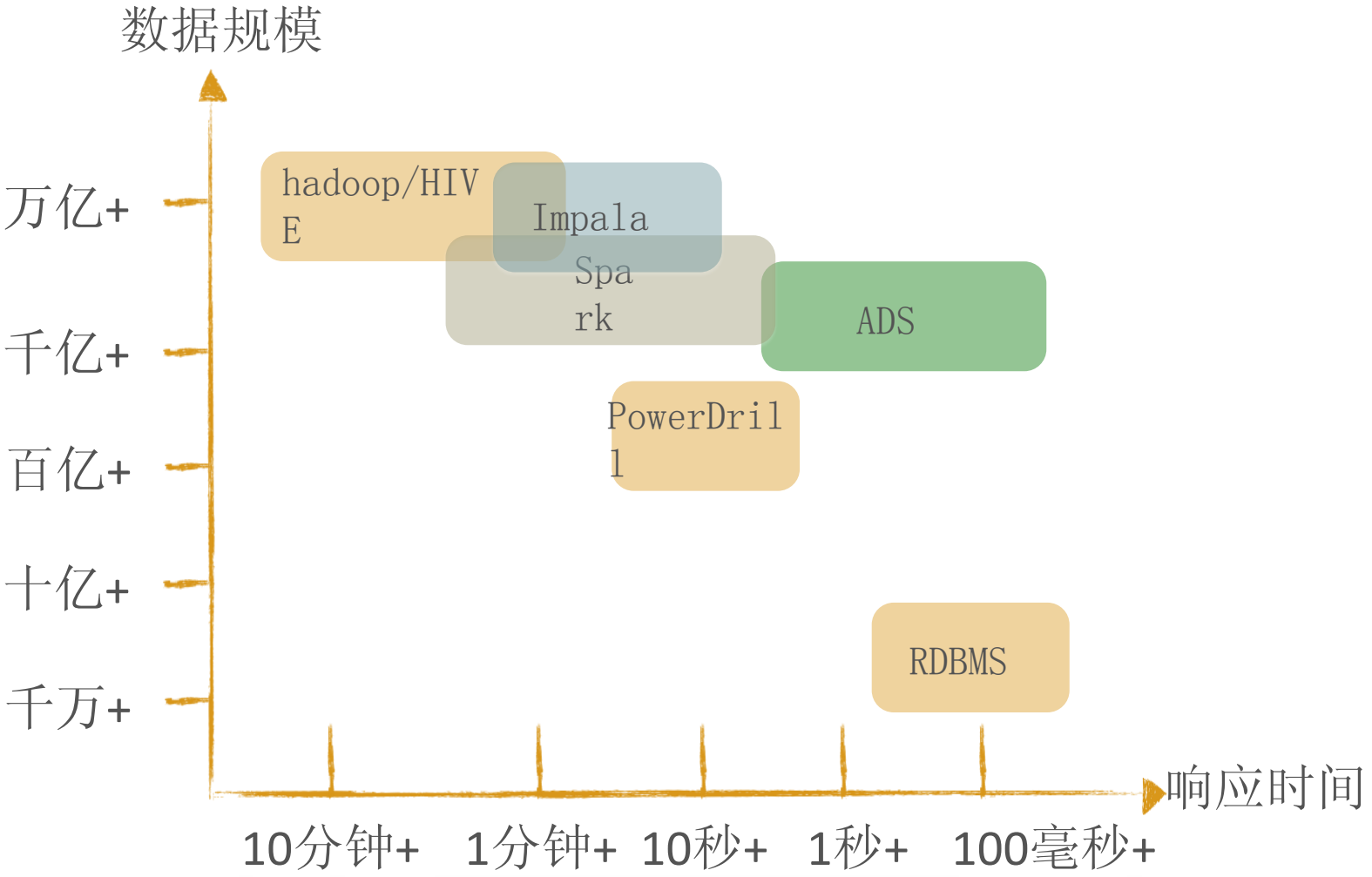
方便的接口

REST/MySQL/MDX
ODPS/RDS/OSS

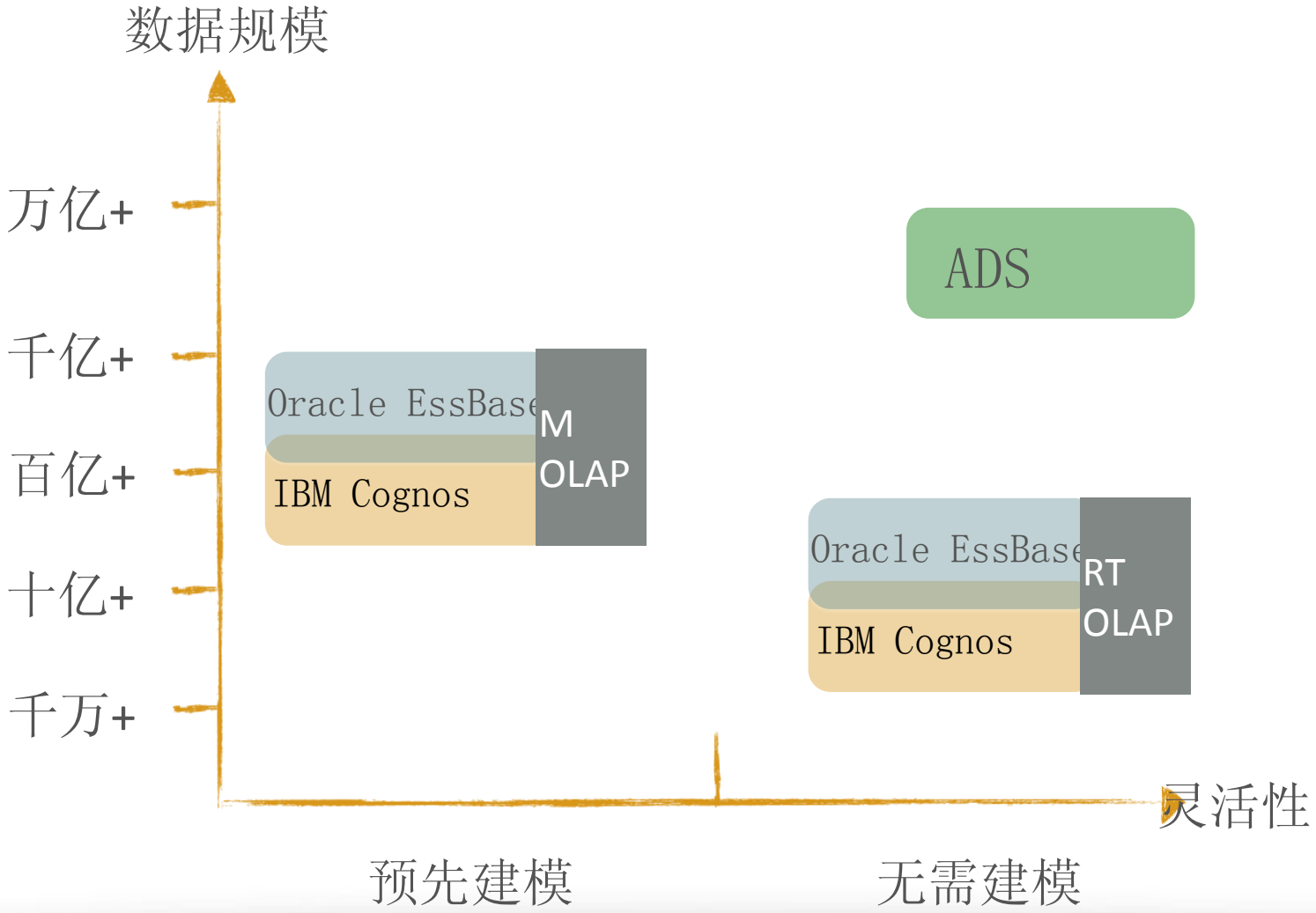
弹性的多租户

资源隔离/元数据/……

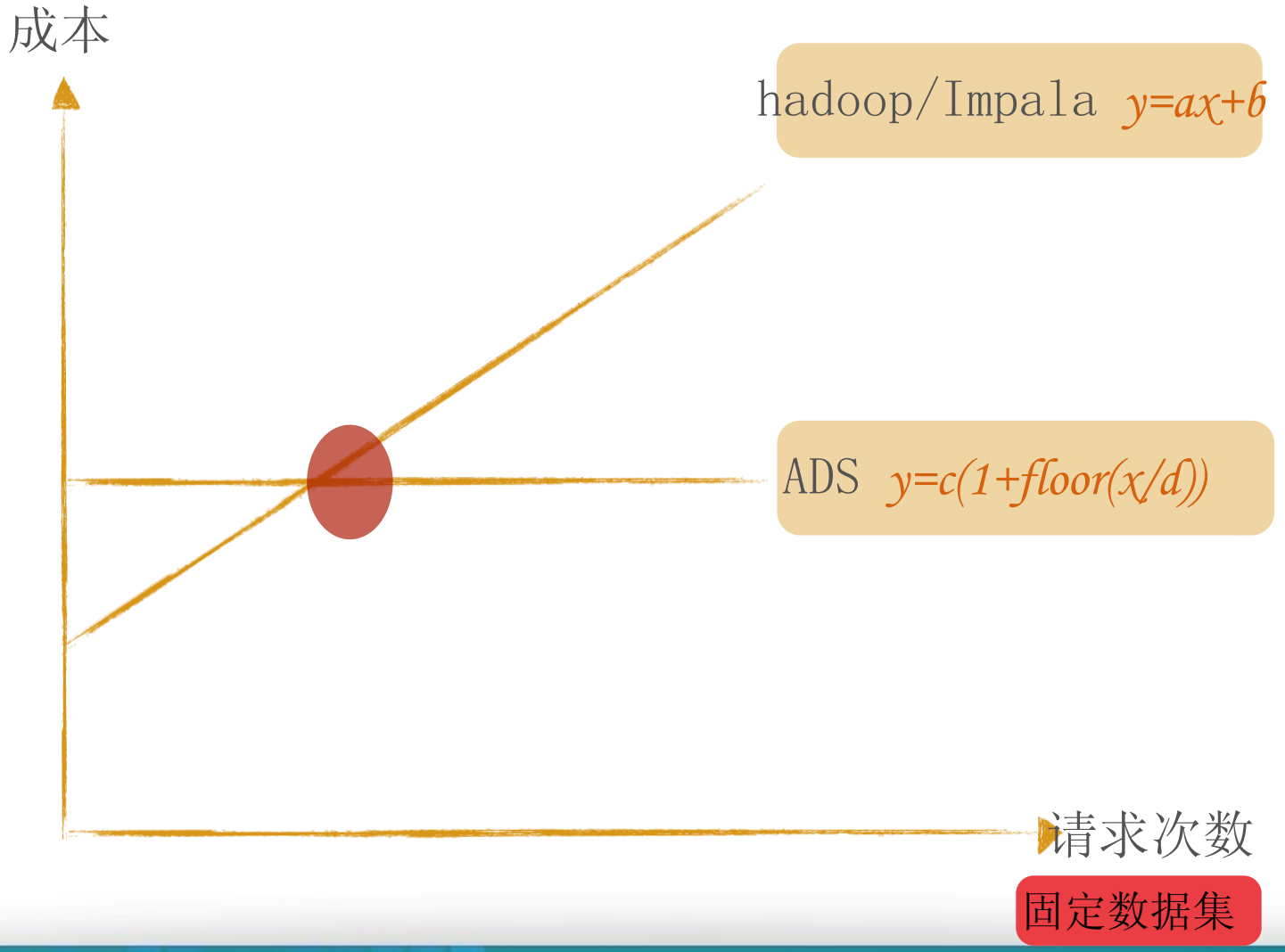
架构—分析引擎


数据规模
万亿+
千亿+
百亿+
十亿+
千万+
hadoop/HIVE
Impala
Spark
ADS
PowerDrill
RDBMS
响应时间
10分钟+
1分钟+
10秒+
1秒+
100毫秒+

架构—分析引擎


数据规模
万亿+
千亿+
百亿+
十亿+
千万+
ADS
Oracle EssBase
IBM Cognos
M
OLAP
Oracle EssBase
IBM Cognos
RT
OLAP
灵活性
预先建模
无需建模

架构—分析引擎



架构—分析引擎


Request
70M+


RT
60ms


QPS
10000+


Developer
User
400+


Product
60+


Storage
100TB+


Records
500B+


Availability
99.99%
双机房热备

架构—分析引擎

文档 23 +

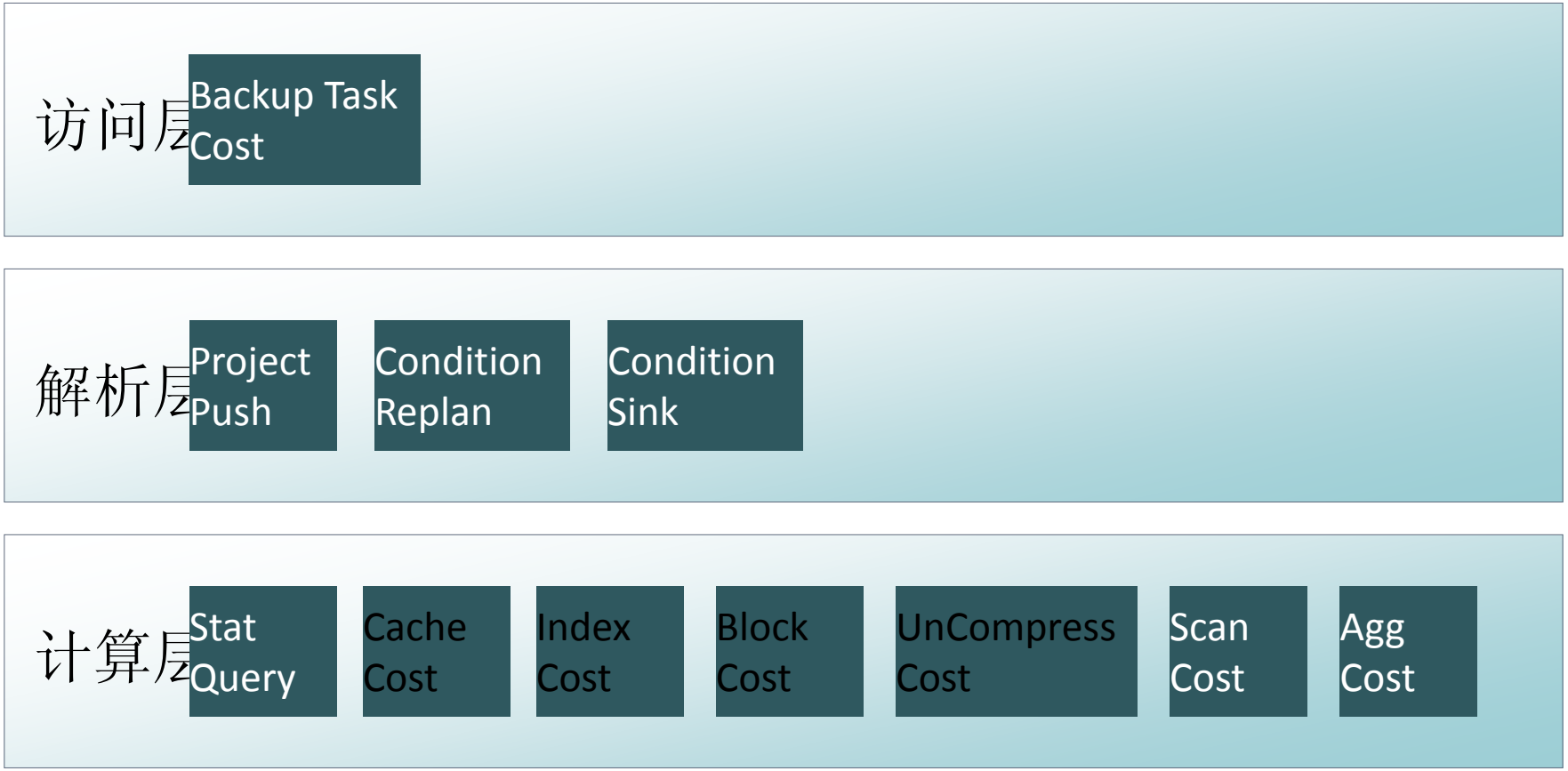
应用 45 +

表 390 +

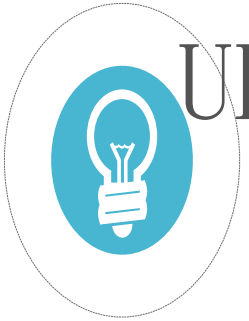
架构—分析引擎


访问层
DB 隔离
分区结果 Cache
路由 Cache
异步 长连接
局部聚合
计算层
DB 隔离
CBO
流式 计算
异步IO
内存 对齐
Cache
并发写 DFS
数据层
链式压缩
预排序
GCIH
原生类型
长尾

架构—分析引擎


访问层
Backup Task Cost
解析层
Project Push
Condition Replan
Condition Sink
计算层
Stat Query
Cache Cost
Index Cost
Block Cost
UnCompress Cost
Scan Cost
Agg Cost

架构—分析引擎



UDTF

DMP lookalike

空间检索

观象台 DMP

多值列

10x Perf intersection

未来

- 场景
  - ✓ 延时
  - ✓ 交互式
- 开放
  - ✓ benchmark
  - ✓ 基础设施
  - ✓ 数据服务
- 技术
  - ✓ 统一计算框架

BDTC
2014 中国大数据技术大会
BIG DATA TECHNOLOGY CONFERENCE
暨第二届CCF大数据学术会议

谢谢