

数据分析平台

平台演进及数据分析方法应用

演讲人：杨军 @蚂蚁金服-数据平台部

* 仅限内部交流使用
如果需要公开，请联系文档作者

目录.CONTENTENTS

PART / 01

- 我是谁：个人简介
- 我们是谁：数据平台部简介

PART / 02

- 做什么：数据分析领域简介

PART / 03

- 怎么来：数据分析平台演进历史
- 怎么做：数据分析平台3.0详解

PART / 04

- 能干什么：数据分析驱动数据
分析平台性能优化

* 仅限内部交流使用，如果需要公开，请联系文档作者



01 / 简介

个人介绍及数据平台部介绍

* 仅限内部交流使用，如果需要公开，请联系文档作者

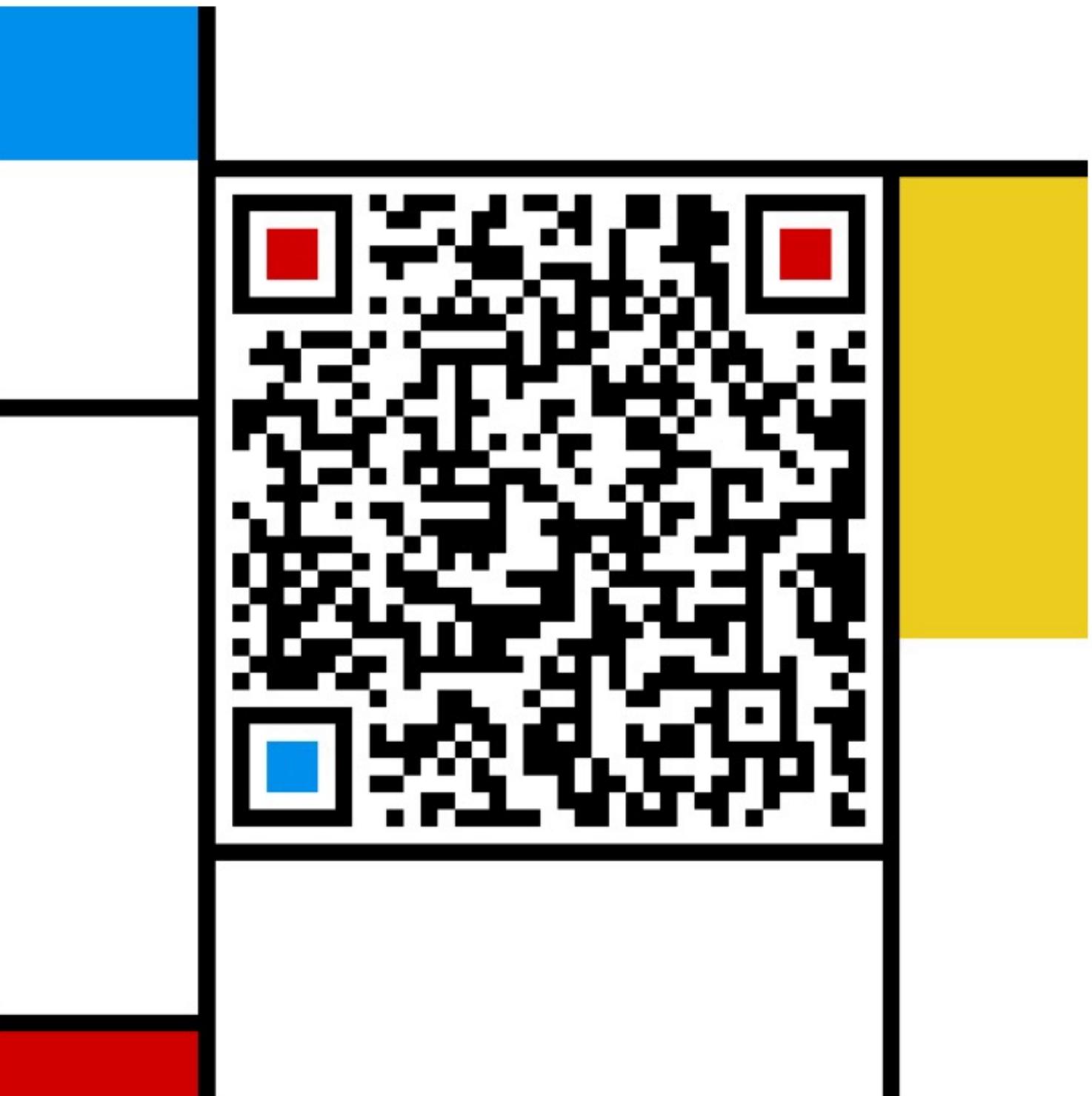


01 | 个人简介

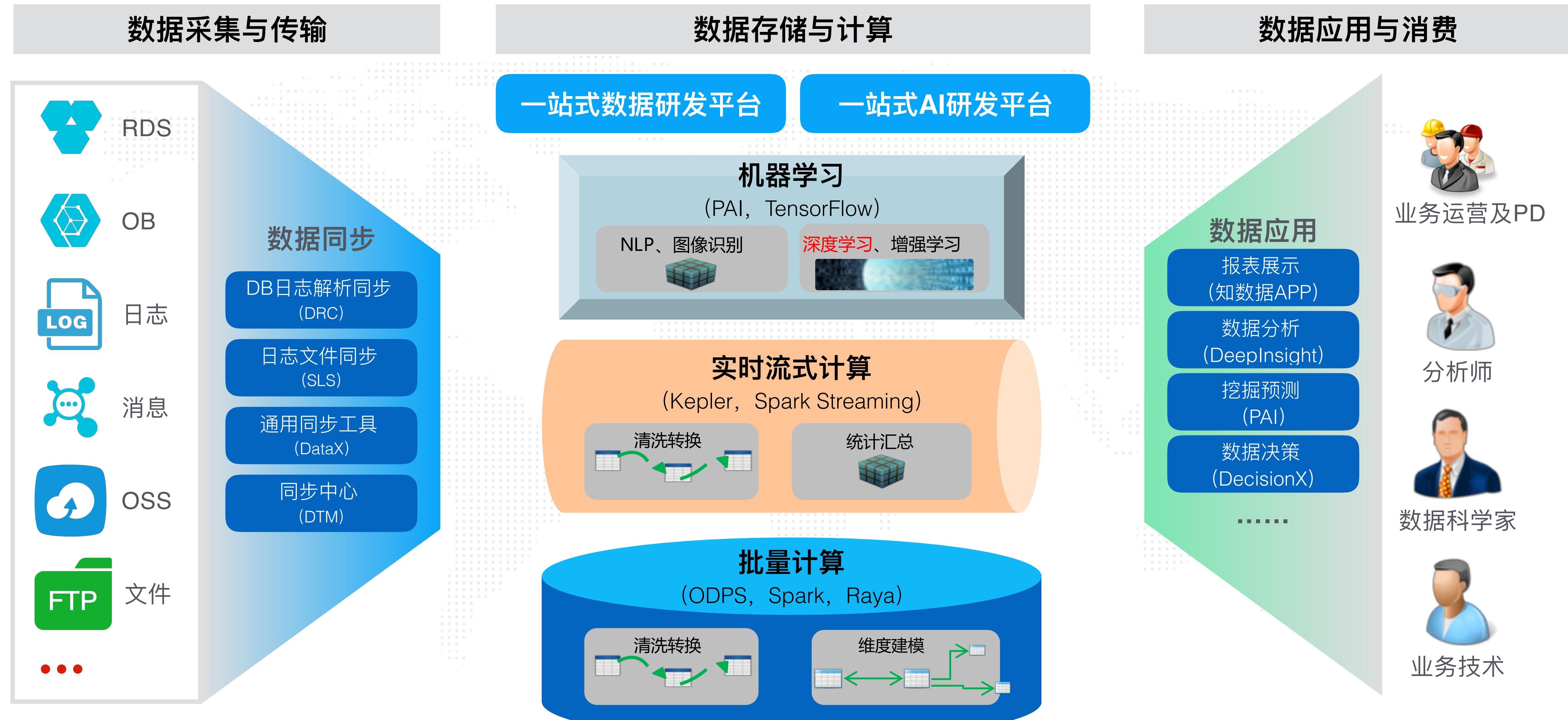
杨军

花名：悟迷（心悟成佛，心迷成魔，^_^）

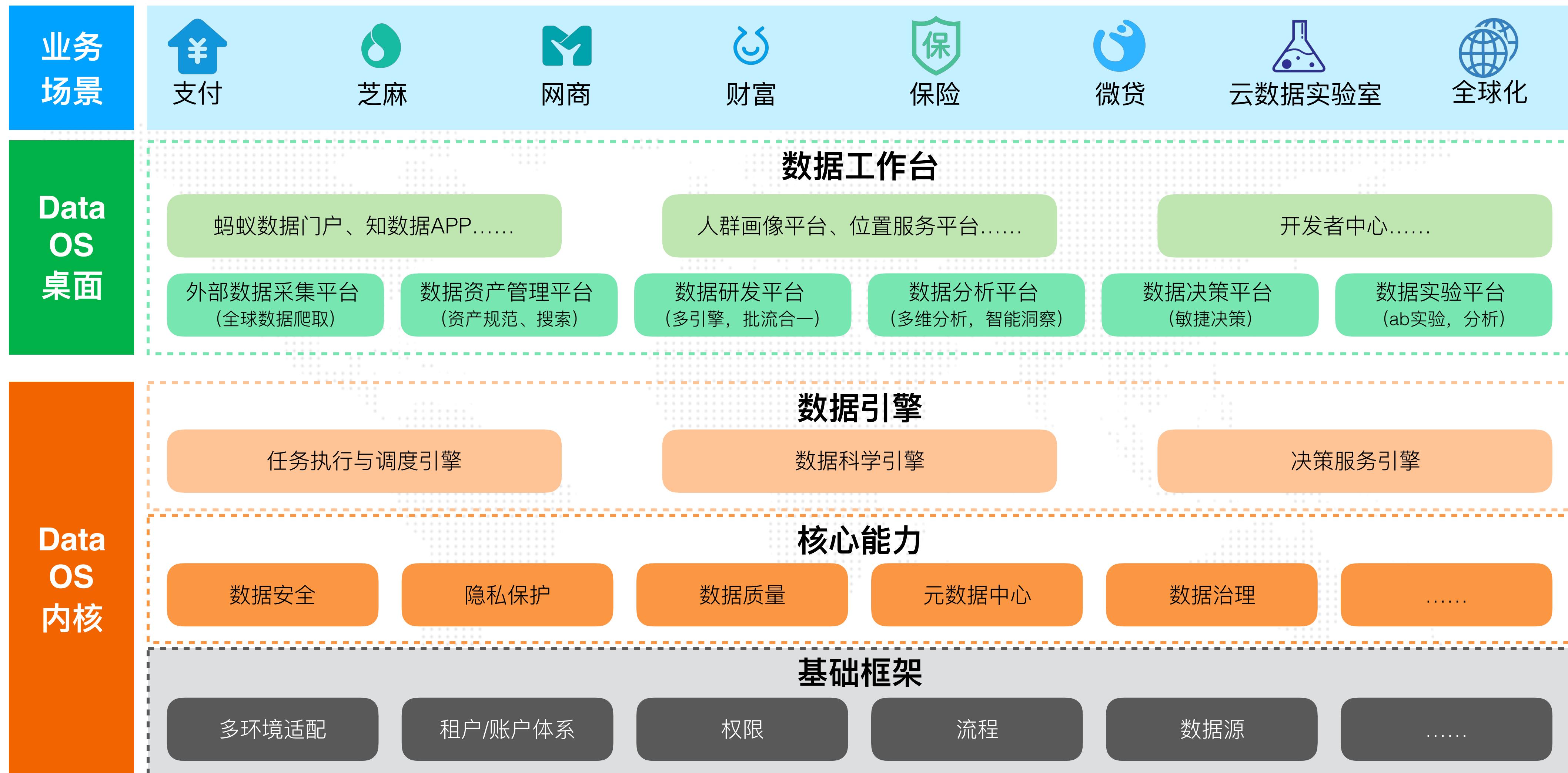
- ◆ 11年实习入职，12年正式入职，主要经历了ETL、实时计算、批流统一的网站日志处理框架等方面的工作。
- ◆ 14年进入蚂蚁财富，主要参与招财宝，保险，众筹，基金等核心业务建设，在基金主导建设并切换蚂蚁自己的基金销售交易清算平台阿基米德。
- ◆ 16年回归数据进入蚂蚁数据平台，带领团队建设数据分析平台引擎层，落地数据分析方法论。
- ◆ 目前负责数据安全与合规，在大数据的道路上继续潜行。



02 | 数据平台部简介(1/3)



02 | 数据平台部简介(2/3)



02 | 数据平台部简介(3/3)

每一个微小的念头 都值得用数据浇灌



02 / 数据分析

数据分析领域体系化结构

* 仅限内部交流使用，如果需要公开，请联系文档作者

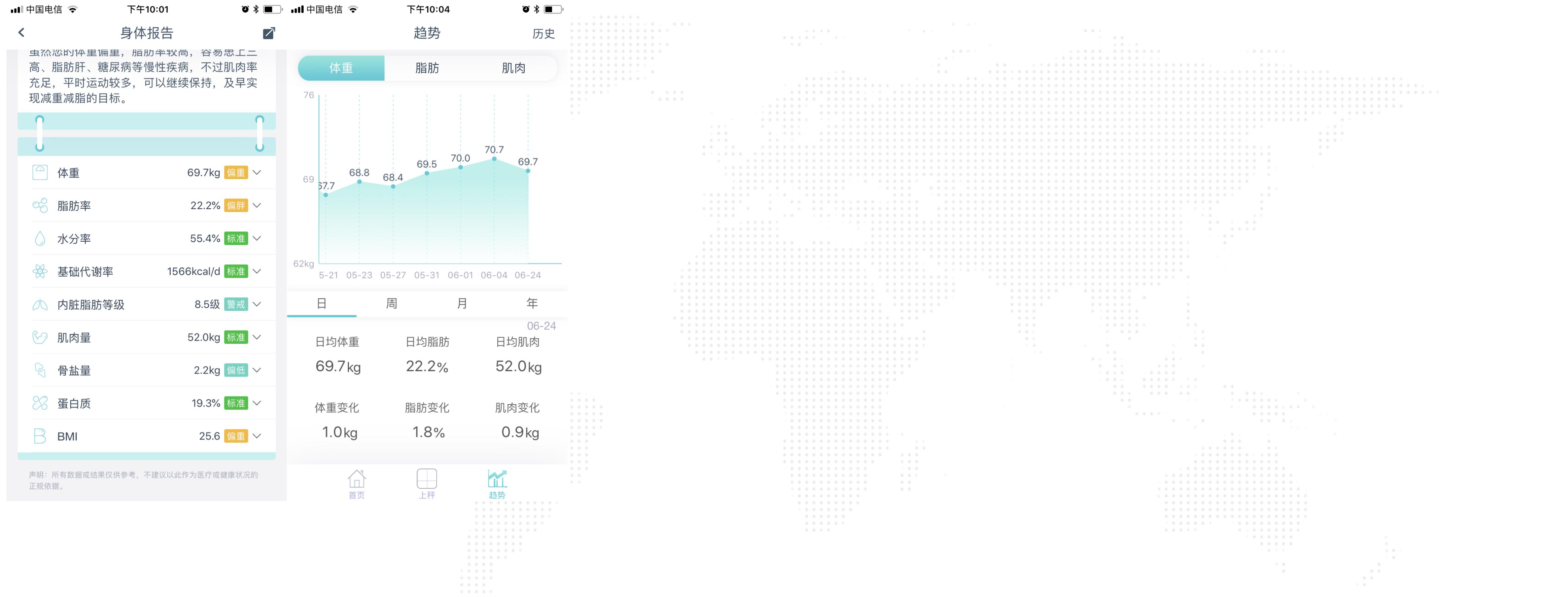




01 | 身边的数据分析



01 | 身边的数据分析



01 | 身边的数据分析



01 | 身边的数据分析



01 | 身边的数据分析



01 | 身边的数据分析

下午10:01 下午10:04 下午10:04

身体报告

忠实地1公里，脂肪率较高，容易患上三高、脂肪肝、糖尿病等慢性疾病，不过肌肉率充足，平时运动较多，可以继续保持，及早实现减重减脂的目标。

体重	69.7kg 偏重
脂肪率	22.2% 偏胖
水分率	55.4% 标准
基础代谢率	1566kcal/d 标准
内脏脂肪等级	8.5级 警戒
肌肉量	52.0kg 标准
骨盐量	2.2kg 偏低
蛋白质	19.3% 标准
BMI	25.6 偏重

声明：所有数据或结果仅供参考，不建议以此作为医疗或正规依据。

趋势

体重、脂肪、肌肉趋势图（单位：kg）

日期	体重	脂肪	肌肉
5-21	67.7	22.2%	55.4%
05-23	68.8	22.2%	55.4%
05-27	68.4	22.2%	55.4%
05-31	69.5	22.2%	55.4%
06-01	70.0	22.2%	55.4%
06-04	70.7	22.2%	55.4%
06-24	69.7	22.2%	55.4%

日 周 月 年

日均体重 日均脂肪 日均肌肉
69.7kg 22.2% 52.0kg

体重变化 脂肪变化 肌肉变化
1.0kg 1.8% 0.9kg

絕密

下午10:13

尴尬

2018年6月26日 星期二

订单

历史订单

- 一人宴(聚龙大厦店) > 2018-06-20 11:06 订单已送达 古法炖牛坑腩-古法炖牛坑腩 ¥30.1
- 二师兄的脊梁 > 2018-05-17 12:18 订单已送达 归去来兮·脊骨饭 (脊骨、随机咸菜...) ¥20.3
- 二师兄的脊梁 > 2018-04-23 11:37 订单已送达 春游西湖·脊骨套餐 (脊骨、米饭青...) ¥20.3
- 二师兄的脊梁 > 2018-03-26 19:00 订单已送达 归去来兮·脊骨套餐 ¥20.3
- 二师兄的脊梁 > 2018-03-23 11:30 订单已送达 二师兄骨肉饭 等2件商品 ¥38.1
- 二师兄的脊梁 > 2018-03-07 12:15 订单已送达

健身记录

- 步行 + 跑步距离 1.9 公里 今天下午6:22
- 步数 2,851 步 今天下午6:22
- 已爬楼层 4 层 今天下午3:35

睡眠状况

- 睡眠分析 6 小时 31 分 今天上午6:33

健康数据

体重 69.7kg 偏重

脂肪率 22.2% 偏胖

医疗急救卡

外卖 发现 订单 我的

下午10:02

身体报告

杨 偏重型 76岁 身体得分 身体年龄

对比变化 与32天前 (2018.05.23) 晚上的测量相比，体重上升0.9kg，脂肪率下降0.5%。

健康报告

最近这段时间，体重上升了一些呢，继续保持良好运动习惯，坚持规律的有氧运动，每周游泳2次，每次30分钟。或者每周跳健美操2次，每次30分钟。

健康分析

虽然您的体重偏重，脂肪率较高，容易患上三高、脂肪肝、糖尿病等慢性疾病，不过肌肉率充足，平时运动较多，可以继续保持，及早实现减重减脂的目标。

体重 69.7kg 偏重

脂肪率 22.2% 偏胖



01 | 身边的数据分析

下午10:01

身体报告

虽然您的体重偏重，脂肪率较高，容易患上三高、脂肪肝、糖尿病等慢性疾病，不过肌肉率充足，平时运动较多，可以继续保持，及早实现减重减脂的目标。

体重	69.7kg 偏重
脂肪率	22.2% 偏胖
水分率	55.4% 标准
基础代谢率	1566kcal/d 标准
内脏脂肪等级	8.5级 警戒
肌肉量	52.0kg 标准
骨盐量	2.2kg 偏低
蛋白质	19.3% 标准
BMI	25.6 偏重

声明：所有数据或结果仅供参考，不建议以此作为医疗或正规依据。

绝密



下午10:13



身体报告

杨 偏重型 76岁 身体得分 身体年龄

对比变化
与32天前（2018.05.23）晚上的测量相比，体重上升0.9kg，脂肪率下降0.5%。

健康报告
最近这段时间，体重上升了一些呢，继续保持良好运动习惯，坚持规律的有氧运动，每周游泳2次，每次30分钟。或者每周跳健美操2次，每次30分钟。

健康分析
虽然您的体重偏重，脂肪率较高，容易患上三高、脂肪肝、糖尿病等慢性疾病，不过肌肉率充足，平时运动较多，可以继续保持，及早实现减重减脂的目标。

历史订单

- 一人宴(聚龙大厦店) > 2018-06-20 11:06 订单已送达 古法炖牛坑腩-古法炖牛坑腩 ¥30.1
- 二师兄的脊梁 > 2018-05-17 12:18 订单已送达 归去来兮·脊骨饭 (脊骨、随机咸菜...) ¥20.3
- 二师兄的脊梁 > 2018-04-23 11:37 订单已送达 春游西湖·脊骨套餐 (脊骨、米饭青...) ¥20.3
- 二师兄的脊梁 > 2018-03-26 19:00 订单已送达 归去来兮·脊骨套餐 ¥20.3
- 二师兄的脊梁 > 2018-03-23 11:30 订单已送达 二师兄骨肉饭 等2件商品 ¥38.1
- 二师兄的脊梁 > 2018-03-07 12:15 订单已送达

订单

帮买帮送

健身记录

- 步行 + 跑步距离 1.9 公里 今天下午6:22
- 步数 2,851 步 今天下午6:22
- 已爬楼层 4 层 今天下午3:35

睡眠状况

- 睡眠分析 6 小时 31 分 今天上午6:33

外卖

发现

订单

我的

下午10:02

身体报告

体重	69.7kg 偏重
脂肪率	22.2% 偏胖

01 | 身边的数据分析

下午10:01 下午10:04 下午10:04

身体报告

虫森恋的1公里里1公里，脂肪率较高，容易患上三高、脂肪肝、糖尿病等慢性疾病，不过肌肉率充足，平时运动较多，可以继续保持，及早实现减重减脂的目标。

体重	69.7kg 偏重
脂肪率	22.2% 偏胖
水分率	55.4% 标准
基础代谢率	1566kcal/d 标准
内脏脂肪等级	8.5级 警戒
肌肉量	52.0kg 标准
骨盐量	2.2kg 偏低
蛋白质	19.3% 标准
BMI	25.6 偏重

声明：所有数据或结果仅供参考，不建议以此作为医疗或正规依据。

绝密

体重 趋势 历史

体重趋势图（单位：kg）

日期	体重
5-21	67.7
05-23	68.8
05-27	68.4
05-31	69.5
06-01	70.0
06-04	70.7
06-24	69.7

日 周 月 年

日均体重 日均脂肪 日均肌肉
69.7kg 22.2% 52.0kg

体重变化 脂肪变化 肌肉变化
1.0kg 1.8% 0.9kg

上秤 趋势

下午10:13

尴尬

2018年6月26日 星期二

订单 **帮买帮送**

历史订单

- 一人宴(聚龙大厦店) > 2018-06-20 11:06 订单已送达 古法炖牛坑腩-古法炖牛坑腩 ¥30.1
- 二师兄的脊梁 > 2018-05-17 12:18 订单已送达 归去来兮·脊骨饭 (脊骨、随机咸菜...) ¥20.3
- 二师兄的脊梁 > 2018-04-23 11:37 订单已送达 春游西湖·脊骨套餐 (脊骨、米饭青...) ¥20.3
- 二师兄的脊梁 > 2018-03-26 19:00 订单已送达 归去来兮·脊骨套餐 ¥20.3
- 二师兄的脊梁 > 2018-03-23 11:30 订单已送达 二师兄骨肉饭 等2件商品 ¥38.1
- 二师兄的脊梁 > 2018-03-07 12:15 订单已送达

健身记录

- 步行 + 跑步距离 1.9 公里 今天下午6:22
- 步数 2,851 步 今天下午6:22
- 已爬楼层 4 层 今天下午3:35

睡眠状况

- 睡眠分析 6 小时 31 分 今天上午6:33

今天 健康数据 数据来源 医疗急救卡 外卖 发现 订单 我的

下午10:02

身体报告

杨 偏重型 76岁 身体得分 身体年龄

对比变化 与32天前 (2018.05.23) 晚上的测量相比，体重上升0.9kg，脂肪率下降0.5%。

健康报告 最近这段时间，体重上升了一些呢，继续保持良好运动习惯，坚持规律的有氧运动，每周游泳2次，每次30分钟。或者每周跳健美操2次，每次30分钟。

健康分析 虽然您的体重偏重，脂肪率较高，容易患上三高、脂肪肝、糖尿病等慢性疾病，不过肌肉率充足，平时运动较多，可以继续保持，及早实现减重减脂的目标。

体重 脂肪率

01 | 身边的数据分析





下午10:04

重 趋势 历史

脂肪 肌肉

日期	重量 (kg)
05-23	68.8
05-27	68.4
05-31	69.5
06-01	70.0
06-04	70.7
06-24	69.7

周 月 年

06-24

指标	值
日均脂肪	22.2%
日均肌肉	52.0kg
脂肪变化	1.8%
肌肉变化	0.9kg

重 上秤 趋势



尴尬

搜索

下午10:12

④ 7 8 9 10 11 12

六月

日 一 二 三 四 五 六

24 25 26 27 28 29 30

2018年6月26日 星期二

健身记录

步行 + 跑步距离 1.9 公里
今天 下午6:22

步数 2,851 步
今天 下午6:22

已爬楼层 4 层
今天 下午3:35

睡眠状况

睡眠分析 6 小时 31 分
今天 上午6:33

今天 健康数据 数据来源 医疗急救卡

搜索

下午10:13

④ 7 8 9 10 11 12

订单

帮买帮送

历史订单

一人宴(聚龙大厦店) > 2018-06-20 11:06 订单已送达
古法炖牛坑腩-古法炖牛坑腩 ¥30.1

二师兄的脊梁 > 2018-05-17 12:18 订单已送达
归去来兮·脊骨饭 (脊骨、随机咸菜...) ¥20.3

二师兄的脊梁 > 2018-04-23 11:37 订单已送达
春游西湖·脊骨套餐 (脊骨、米饭青...) ¥20.3

二师兄的脊梁 > 2018-03-26 19:00 订单已送达
归去来兮·脊骨套餐 ¥20.3

二师兄的脊梁 > 2018-03-23 11:30 订单已送达
二师兄骨肉饭 等2件商品 ¥38.1

二师兄的脊梁 > 2018-03-07 12:15 订单已送达

外卖

发现

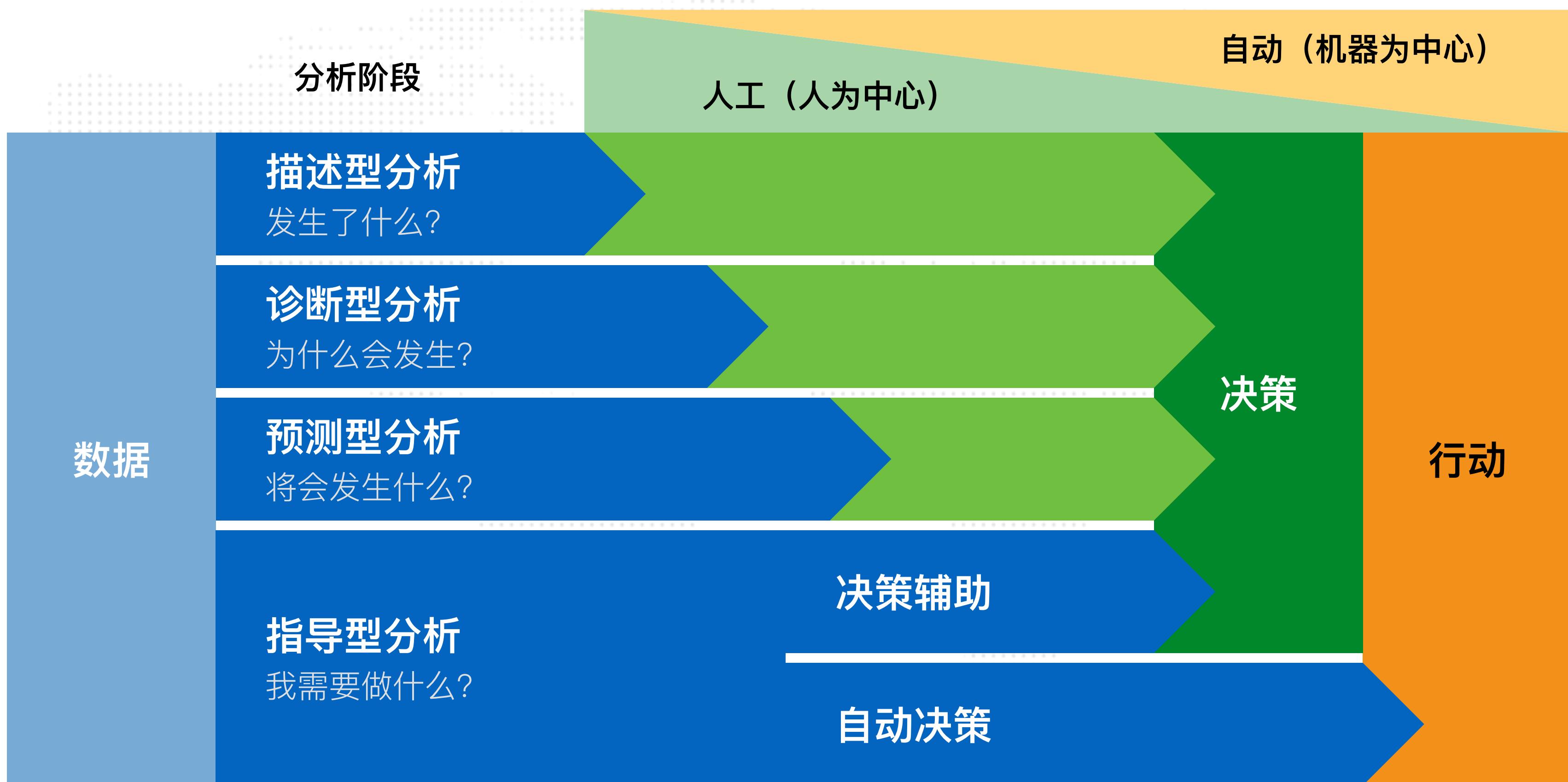
订单

我的

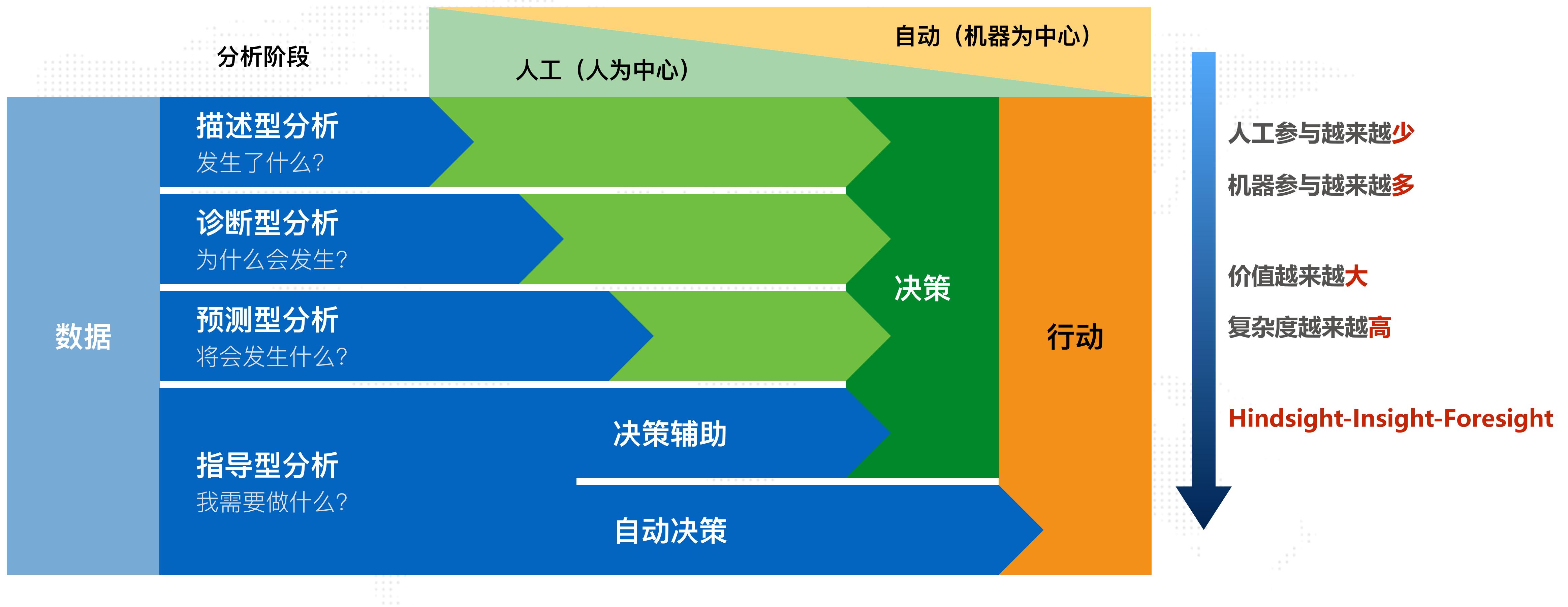
The image is a screenshot of a mobile application interface. At the top, there's a header with the time '下午10:02' (PM 10:02). Below the header, the title '身体报告' (Health Report) is centered. On the left side, there's a circular profile picture of a young man with short dark hair, wearing a blue shirt. To the right of the profile picture, the name '杨' (Yang) is displayed, followed by the text '偏重型' (Slightly Heavy Type) with an information icon (a white 'i' inside a blue circle). Further to the right are the numbers '76' and '30岁' (30 years old), with the text '身体得分' (Health Score) and '身体年龄' (Body Age) positioned below them. Below this main header, the section '对比变化' (Comparison Change) is shown, stating that compared to 32 days ago (May 23, 2018), the user's weight increased by 0.9kg and their fat percentage decreased by 0.5%. The '健康报告' (Health Report) section contains a paragraph encouraging continued exercise and healthy habits. A red rectangular box highlights a note about potential health risks due to high weight and fat levels. On the left, there's a sidebar with icons for '体重' (Weight) and '脂肪率' (Fat Percentage). On the right, there's a large, partially visible image of a baby's face.

* 仅限内部交流使用，如果需要公开，请联系文档作者

02 | 数据分析领域



02 | 数据分析领域



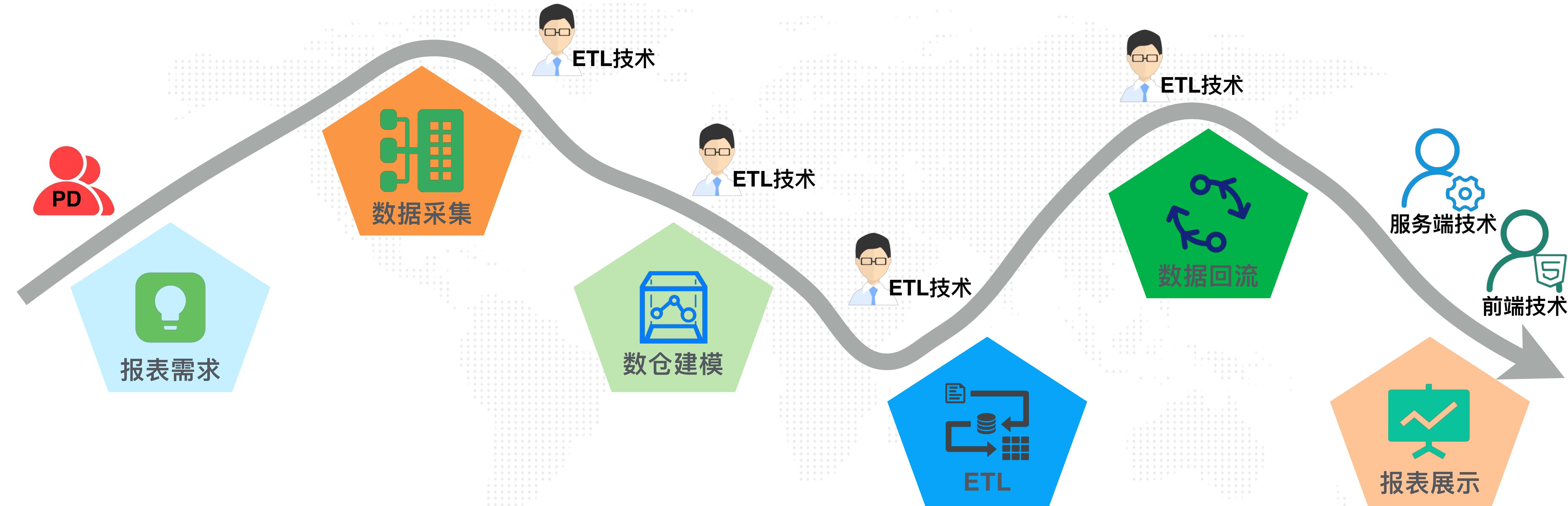
03 / 数据分析平台

蚂蚁数据分析平台演进及技术详解

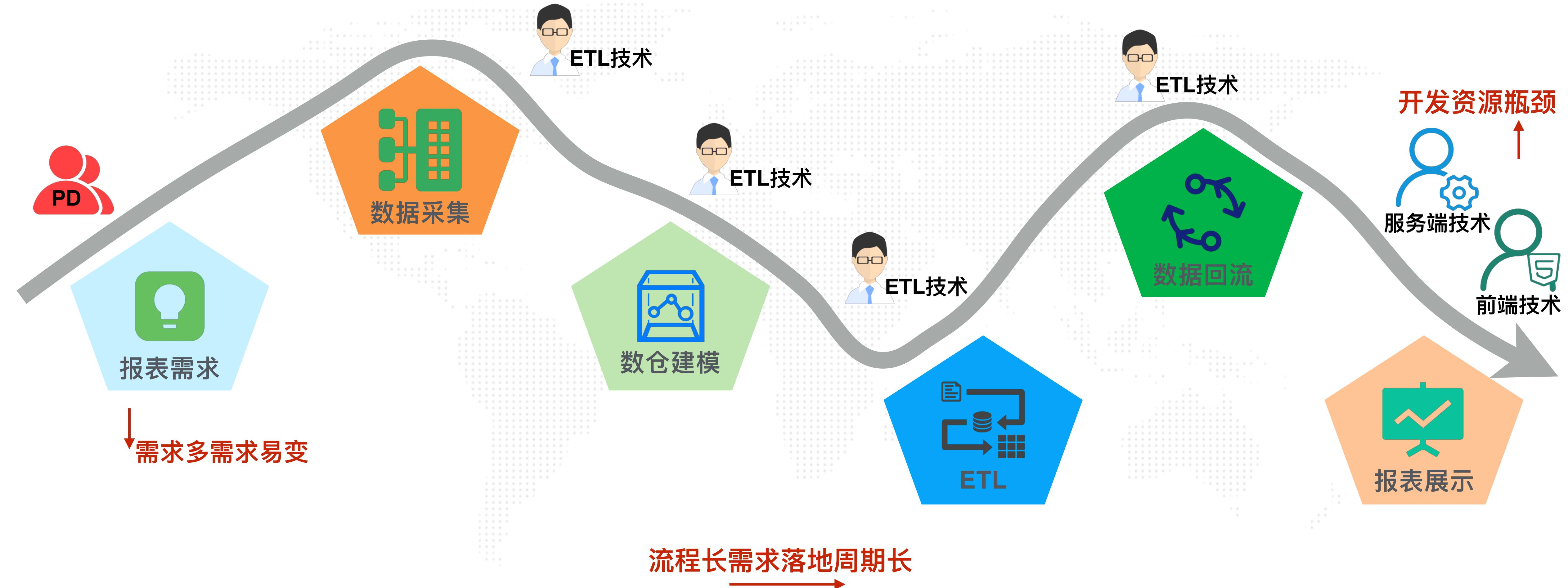
* 仅限内部交流使用，如果需要公开，请联系文档作者



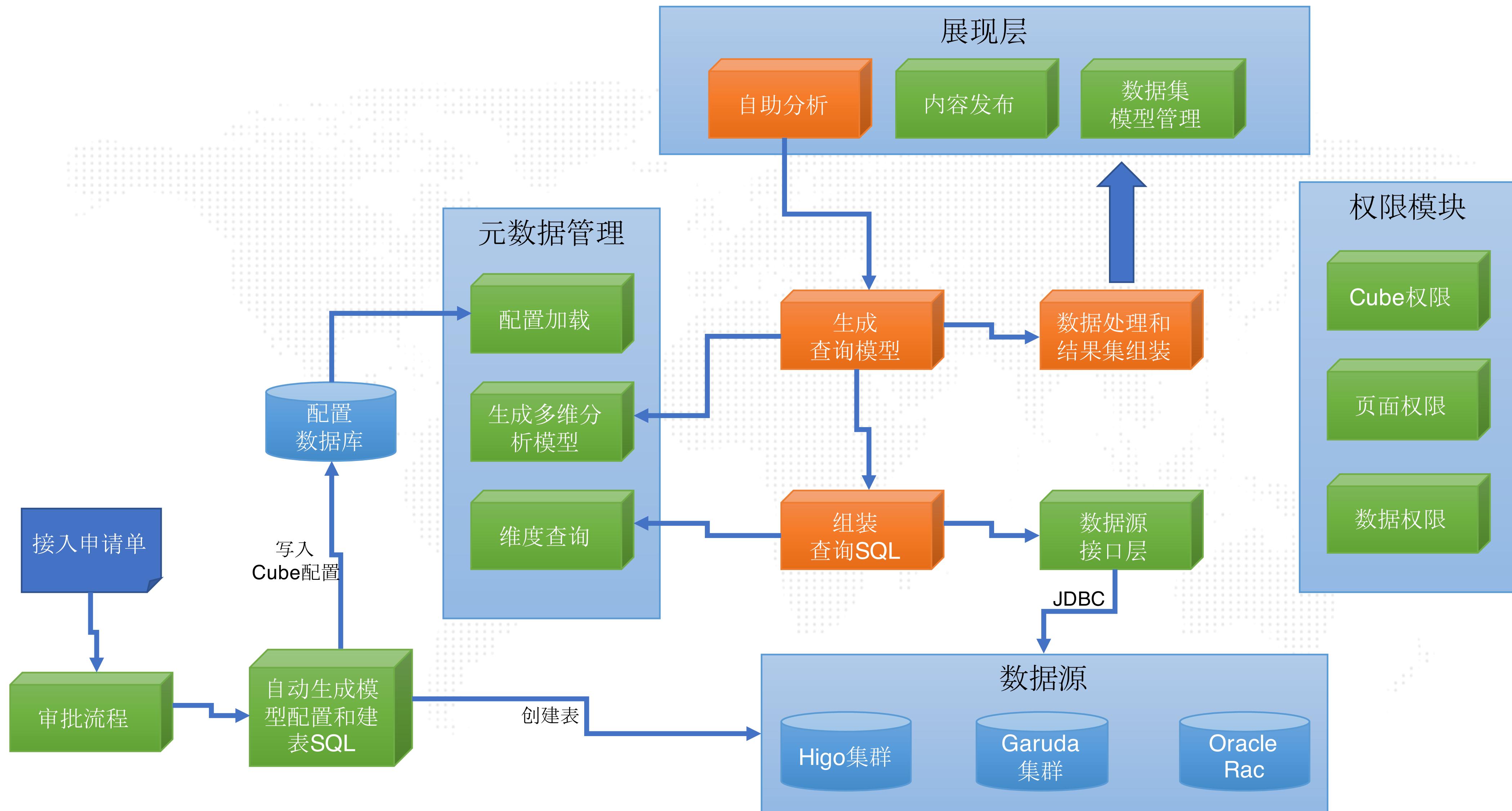
01 | 传统数据分析流程与矛盾



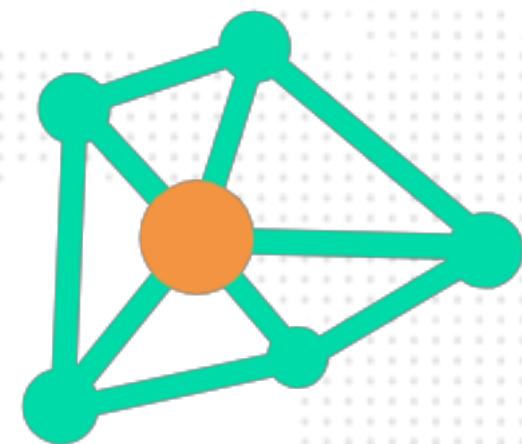
01 | 传统数据分析流程与矛盾



02|数据分析平台2013 V1.0



03 | 数据分析平台1.0新的矛盾



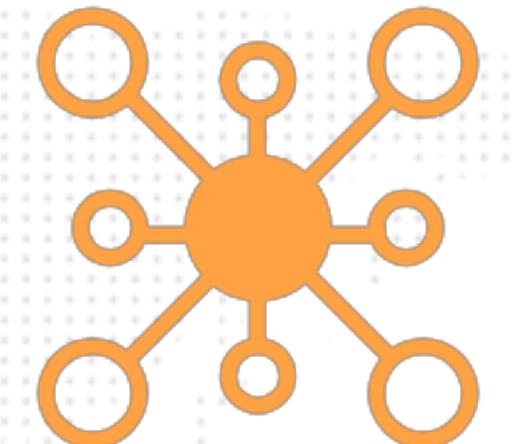
分析功能不足 靠ETL加工

星型模型、雪花模型不支持
明细处理函数不支持
需要ETL加工
ETL资源瓶颈



分析性能不足 靠半自动回流

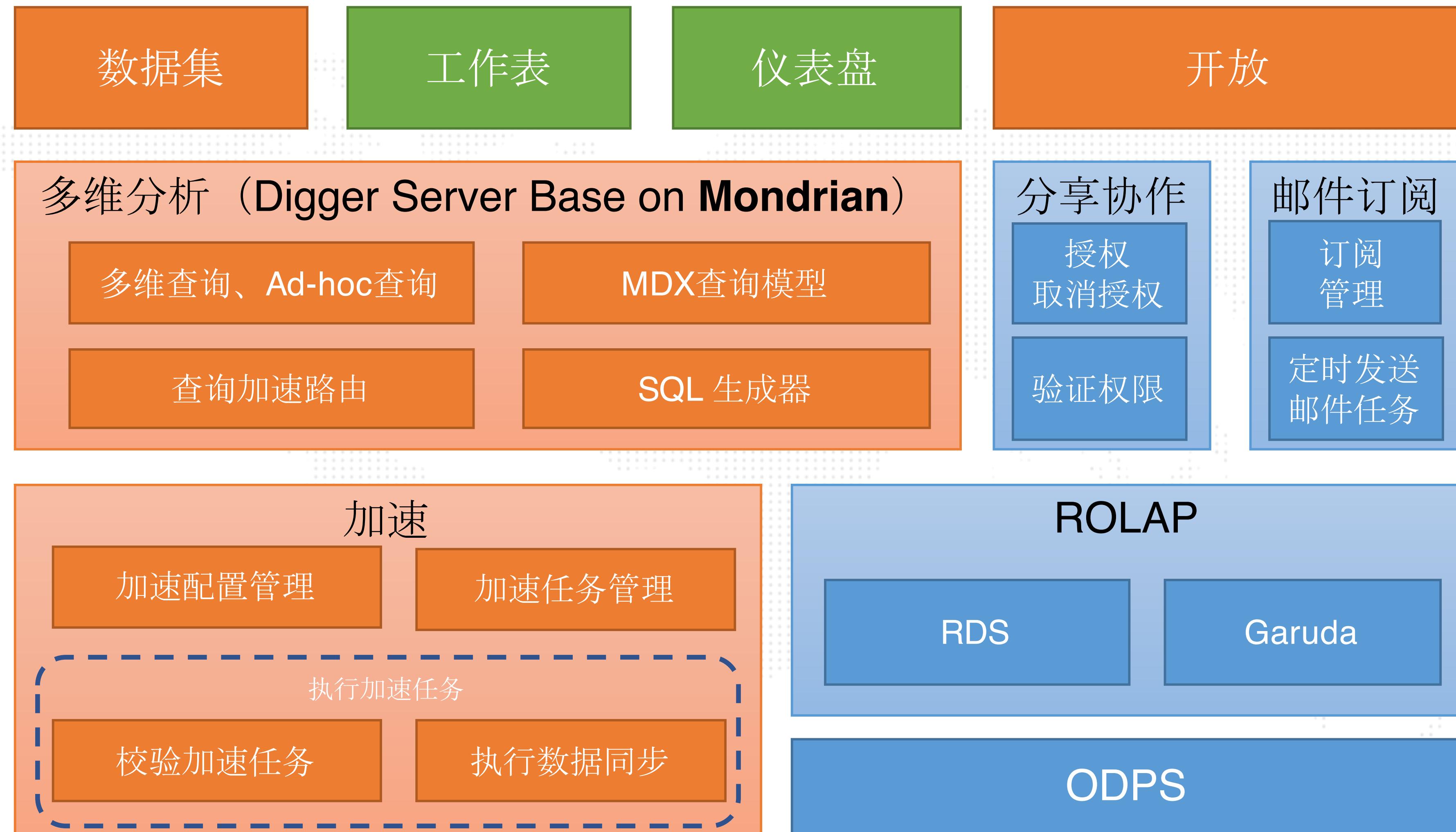
ODPS即时分析性能不足
用户人工回流数据到其他数据源
用户门槛较高
依赖ETL资源



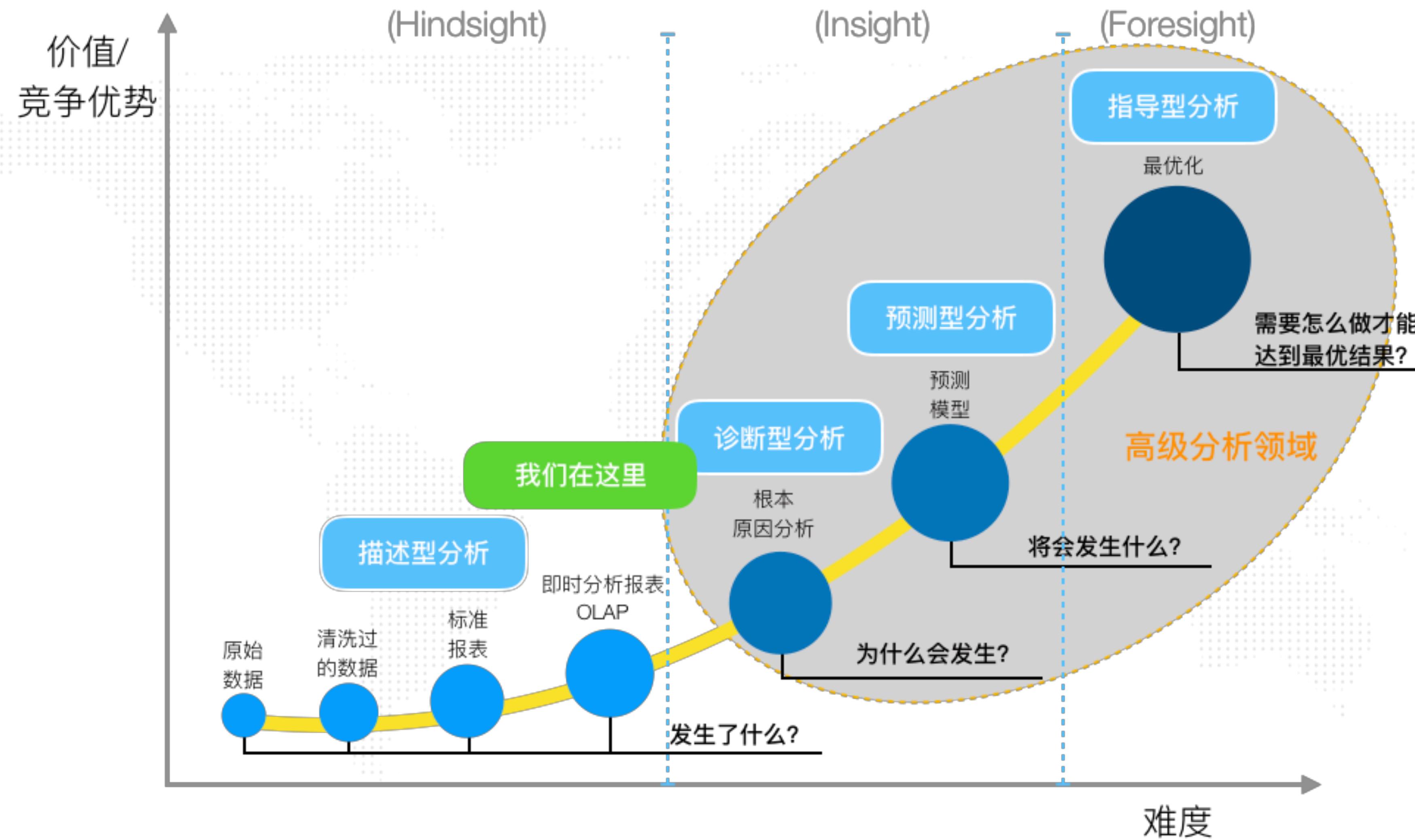
数据能力与 业务工作台分裂

数据分析平台独立系统
小二有自己的业务工作台
数据能力与业务工作台分裂
用户切来切去

04 | 数据分析平台2014-2016 V2.0



05 | 重新定义新分析洞察(1/3)



05 | 重新定义新分析洞察(2/3)

我们曾经生活在一个非此即彼的世界。

您要么懂得如何编程，要么与高级分析技术无缘。

要么学习R、Python 和/或 SAS，要么请人帮您攻克难关。

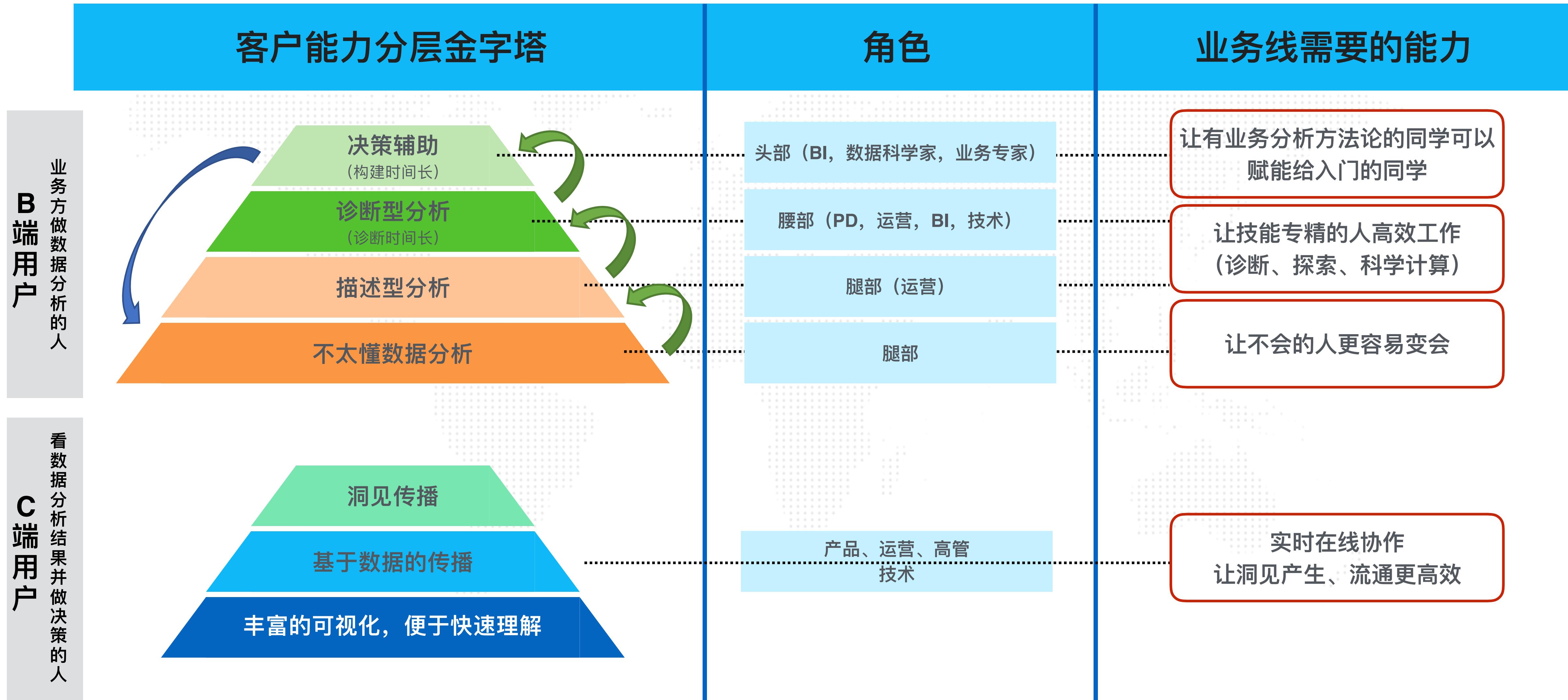
Tableau相信，为了真正地扩充人类智能，我们需要为技术能力千差万别的用户提供丰富的功能。

我们信奉的原则是，让每个技能级别的人都能够从数据中获取见解和证据。

—— Tableau



05 | 重新定义新分析洞察(3/3) / 客户分层及不同层次需求



06 | 数据分析平台V3.0

场景
应用层

会员增长

智慧人群

金融核心平台

蚂蚁保险

蚂蚁客权

蚂蚁达尔文

蚂蚁风控

通用
模版层

分析可视化组件模板

树图

流程图

父子图

等等

分析算法模板

诊断算法

Z检验

相关度
算法

分布检测

分析洞察解决方案模板

全链路监控诊断

人群洞察

中台技术
能力层

拖拽分析

可视化

协作

开放平台IDE
(含分析流程编排引擎)

智能同步

查询路由

数据预警

数据权限

数据集加工

OLAP引擎
Connector

智能预算

科学计算引擎

底层计算&存储能力



06 | 数据分析平台V3.0核心能力细化

Intelligence, **S**elf-Service, **E**nd-to-End Solution, **E**mbedded



分析洞察平台

会员增长平台

智慧人群服务

AB实验平台

国际增长地图

开放服务门面 : SDK / API / DSL

数据科学平台

语言

- 解释编译优化调试
- 多语言 (R , Python 等)
- 数据集集成
- 开发者工具

能力

- 轻加工能力 : 明细
- 多维分析能力 : 钻取
- 科学分析能力 : 检验、模型
- 复合分析模型 : 人群、留存

运行

- 智能路由
- 智能优化
- 多源适配 (SPI)
- 缓存&队列管控

统计支撑 / 健康检查

iSync (智能同步中心)

- ❖ 同步任务管理
- ❖ 同步任务调度执行
- ❖ 多引擎支持 (SPI)
- ❖ 智能任务优先级
- ❖ 智能选源选格式
- ❖ 在线查询路由

iPrepare (智能预算算, 原 iCube)

- ❖ 多策略计算逻辑定义
- ❖ 智能生命周期管理
- ❖ 任务自动运维和管控
- ❖ 智能合并路径最优化
- ❖ 多引擎支持 (SPI)
- ❖ 查询最优路由及改写

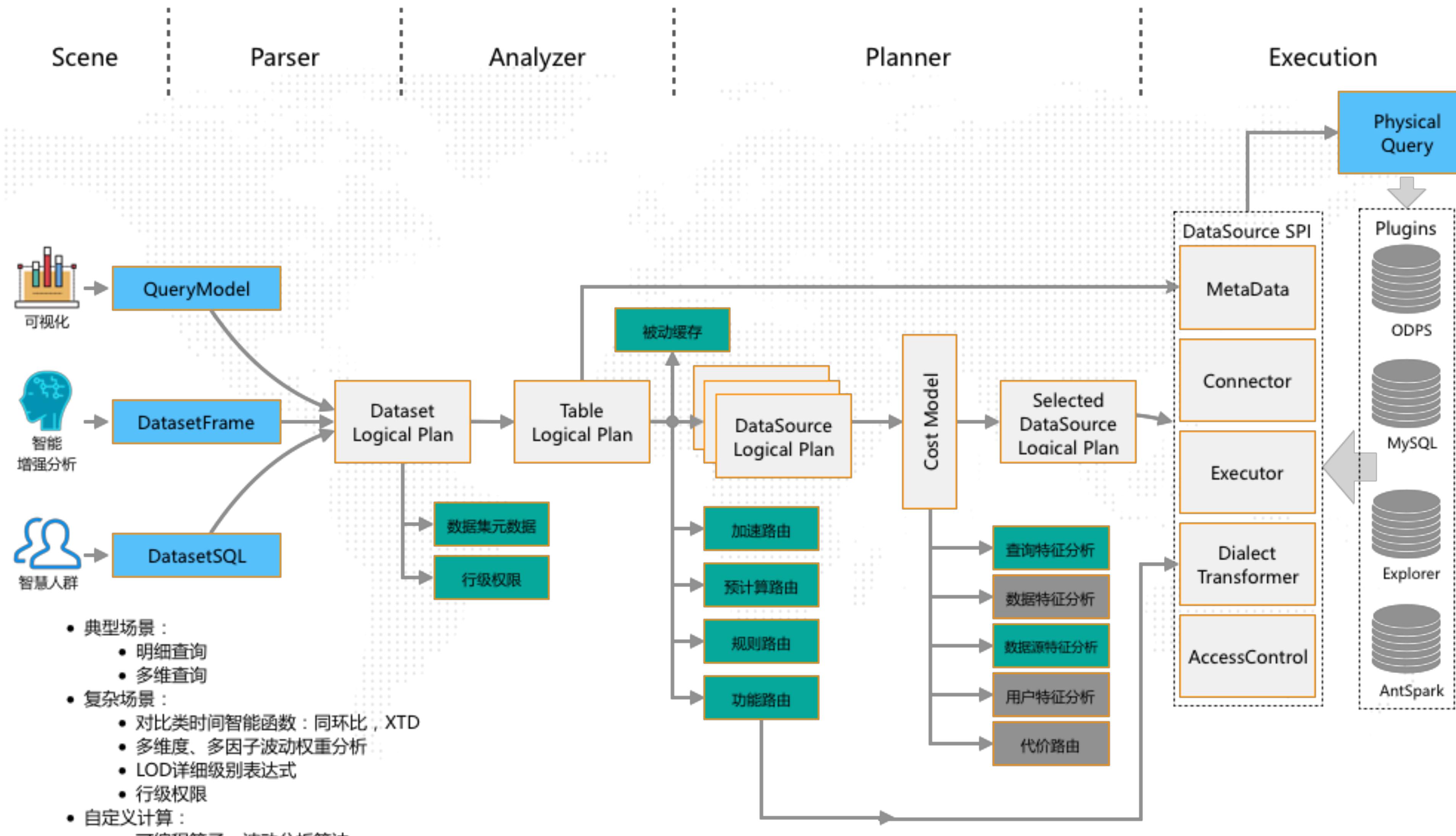
iEngine (计算引擎)

- ❖ 科学计算运行容器
- ❖ 算子可插拔热升级
- ❖ 多版本多上下文隔离
- ❖ 多租户隔离
- ❖ 资源自适应
- ❖ 大数据量高性能

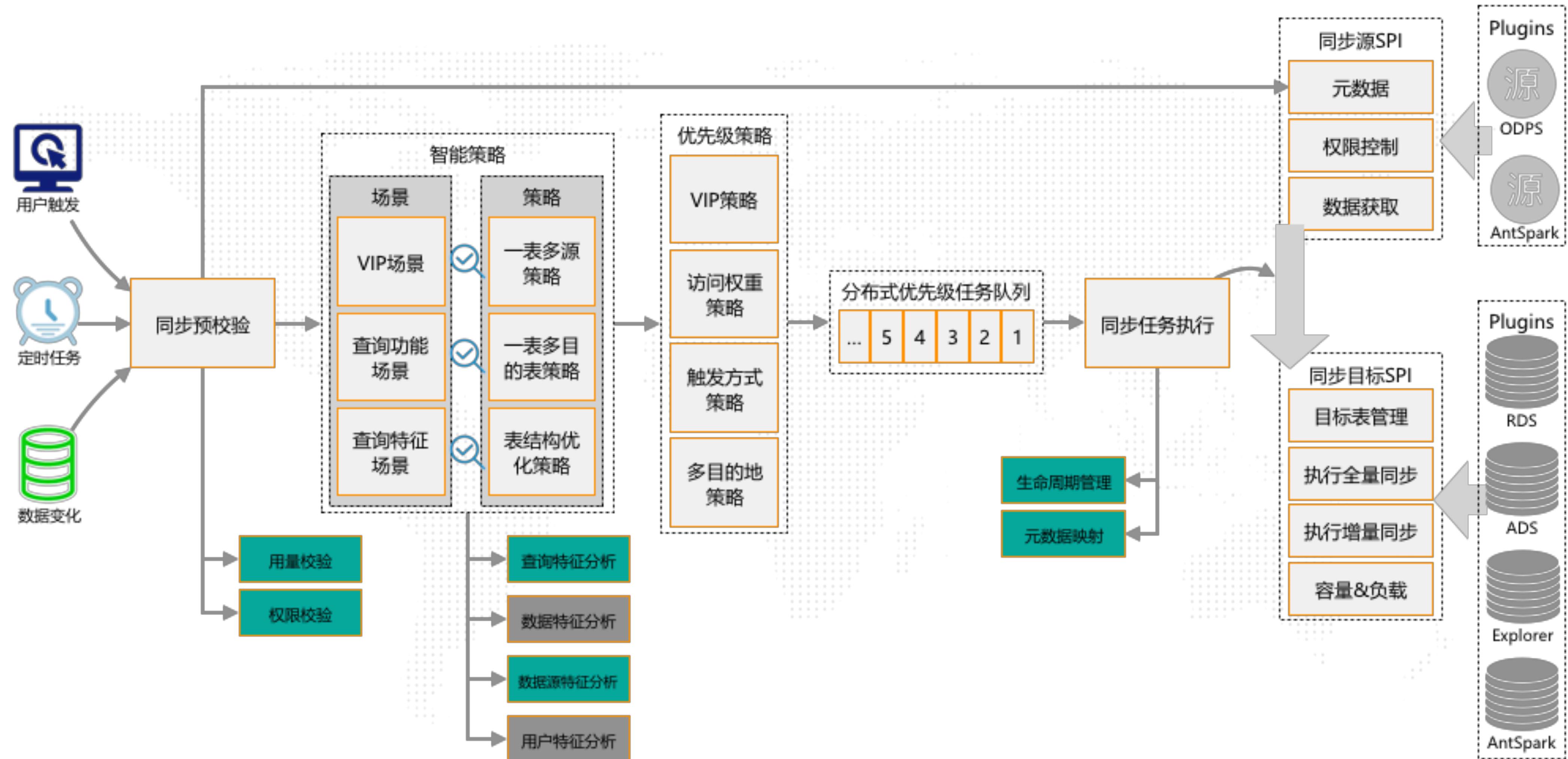
统一元数据中心 / 全链路血缘

基础设施 : ODPS / RDS / ADS / Explorer / AntSpark / SparkOnODPS / Kudu / Pangu /

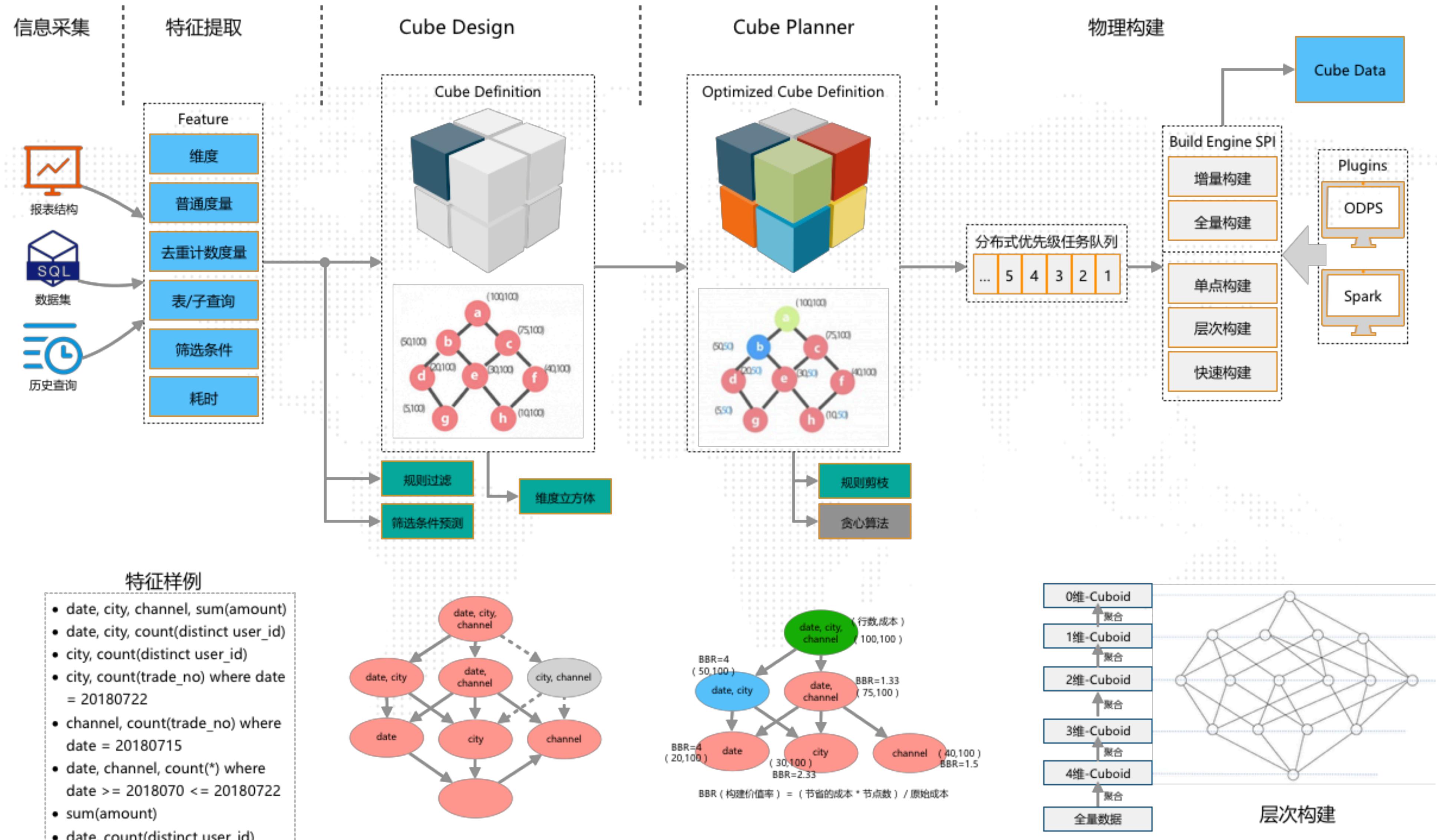
07 | 数据分析平台核心技术(1/3) / 查询全貌



07 | 数据分析平台核心技术(2/3) / 智能同步

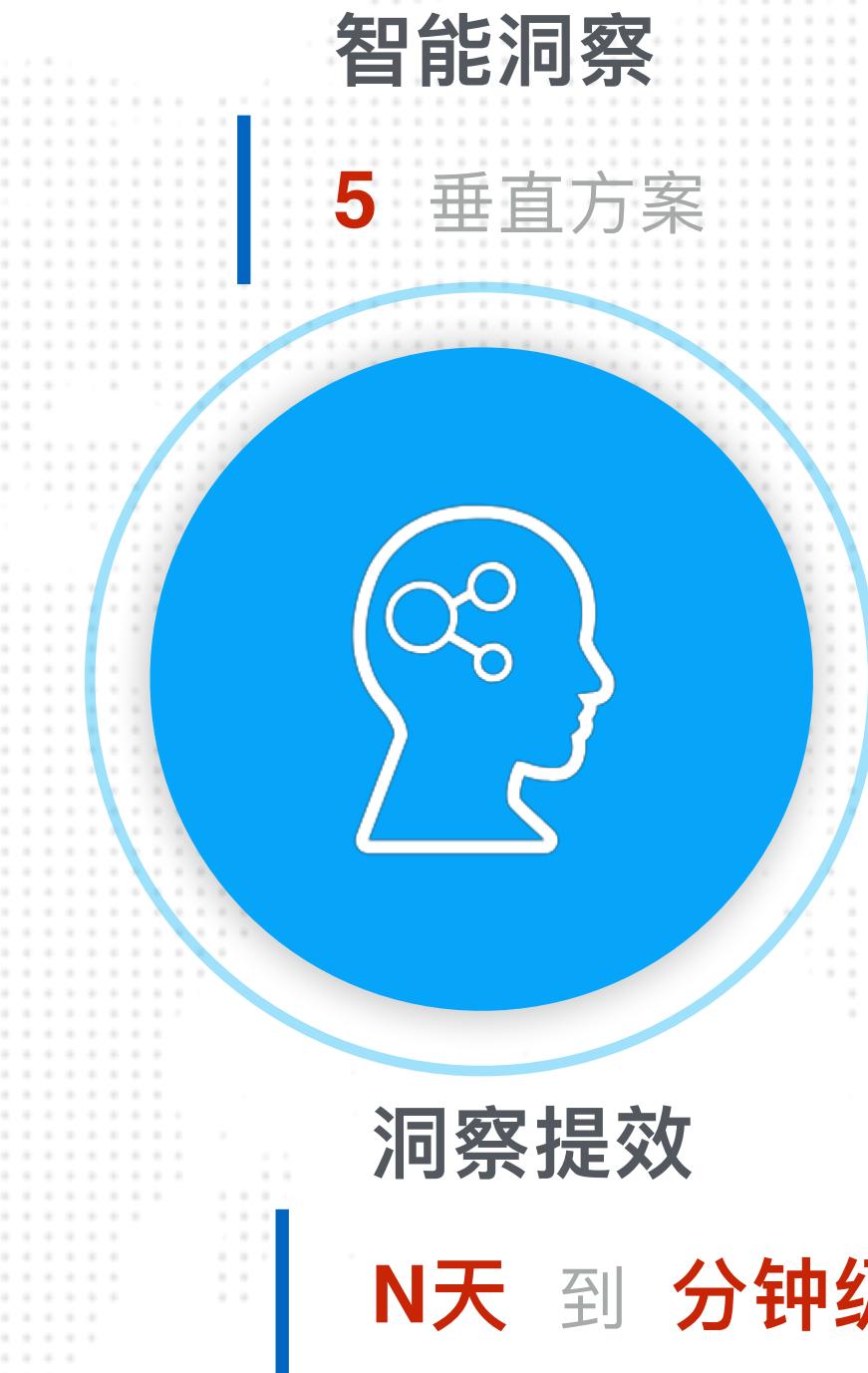


07 | 数据分析平台核心技术(3/3) / 智能预算计算



* 仅限内部交流使用，如果需要公开，请联系文档作者

08 | 数据分析平台成果



04 / 数据分析应用

数据分析驱动数据分析平台性能优化

* 仅限内部交流使用，如果需要公开，请联系文档作者



01 | 问题定义

报表查询15秒

太慢了



打开一些页面要8秒

期望提升到秒级



个别报表查询要90秒

简直不能忍



显然这里的问题是RT的问题：

用户的期望是能够达到秒级响应，但是我们知道，就像稳定性一样，实际困难是不可能100%达到秒级的。

02 | 定义指标



什么是一个好的指标：

一个好的指标应该简单易懂，一个好的指标应该是个比率，一个好的指标可以指导行为改变，例如汽车里程和速度。

03 | 依赖业务流程和物理架构来分解指标



特点与挑战

- ◆ 查询链路非常复杂
- ◆ 数据源多种多样，能力参差不齐
- ◆ 查询是用户自助通过界面拖拽生成，会造成形形色色的查询，规律非常不明显



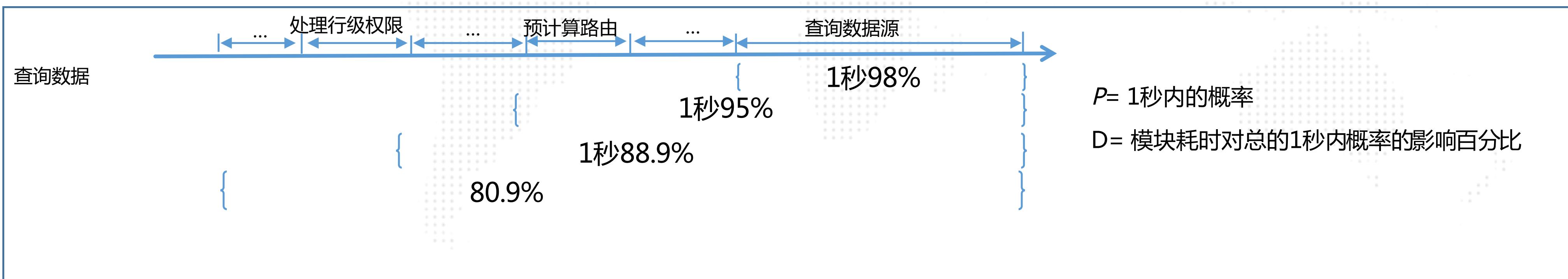
04 | 分解后用数学的方式对指标进行抽象

目标

$$1\text{秒内查询RT占比} = \frac{\sum_{i=1}^n x_i}{\text{total query}}$$

$x_1 = \text{缓存访问次数} * \text{路由到缓存查询1秒内的概率}$
 $x_2 = (\sum_{i=1}^n \text{每张表recall} * \text{该表慢查询pattern命中次数} * \text{构建成功率} * \text{及时加速成功率}) * \text{报表预计算缓存查询1秒内的概率}$
 $x_3 = \underline{\text{mysql及时加速成功率}} * (\sum_{i=1}^n \text{需要加速到Mysql的每张表访问的次数}) * \text{路由到Mysql链路查询1秒内的概率}$
 $x_4 = \underline{\text{greenplum及时加速成功率}} * (\sum_{i=1}^n \text{需要加速到greenplum的每张表访问的次数}) * \text{路由到greenplum链路查询1秒内的概率}$
 $x_5 = \text{MR访问次数} * \text{路由到MR链路查询1秒内的概率}$
 $x_6 = (\sum_{i=1}^n \text{直连MySQL每张表访问的次数}) * \text{路由到ODPS链路查询1秒内的概率}$
 $x_7 = (\sum_{i=1}^n \text{直连EXPLORER每张表访问的次数}) * \text{路由到ODPS链路查询1秒内的概率}$
 $x_8 = (1 - \text{路由成功率}) * \text{查询次数} * \text{路由到ODPS链路查询1秒内的概率}$

报表预计算缓存查询1秒内的比例



05 | 利用数据分析找到问题并制定行动策略

$$RT\text{1秒内报表占比} = \sum_{n=1}^n (X_i \cdot P_i)$$

X_1 = ADS访问次数占比 ,
 X_2 = RDS访问次数占比 ,
 X_3 = EXPLORER访问次数占比 ,
 X_4 = ICUBE访问次数占比 ,
 X_5 = ODPS访问次数占比 ,
 X_6 = CACHE访问次数占比 ,
 X_7 = MYSQL访问次数占比 ,

P_1 = ADS一秒内返回占比
 P_2 = RDS一秒内返回占比
 P_3 = EXPLORER一秒内返回占比
 P_4 = ICUBE一秒内返回占比
 P_5 = ODPS一秒内返回占比
 P_6 = CACHE一秒内返回占比
 P_7 = MYSQL一秒内返回占比



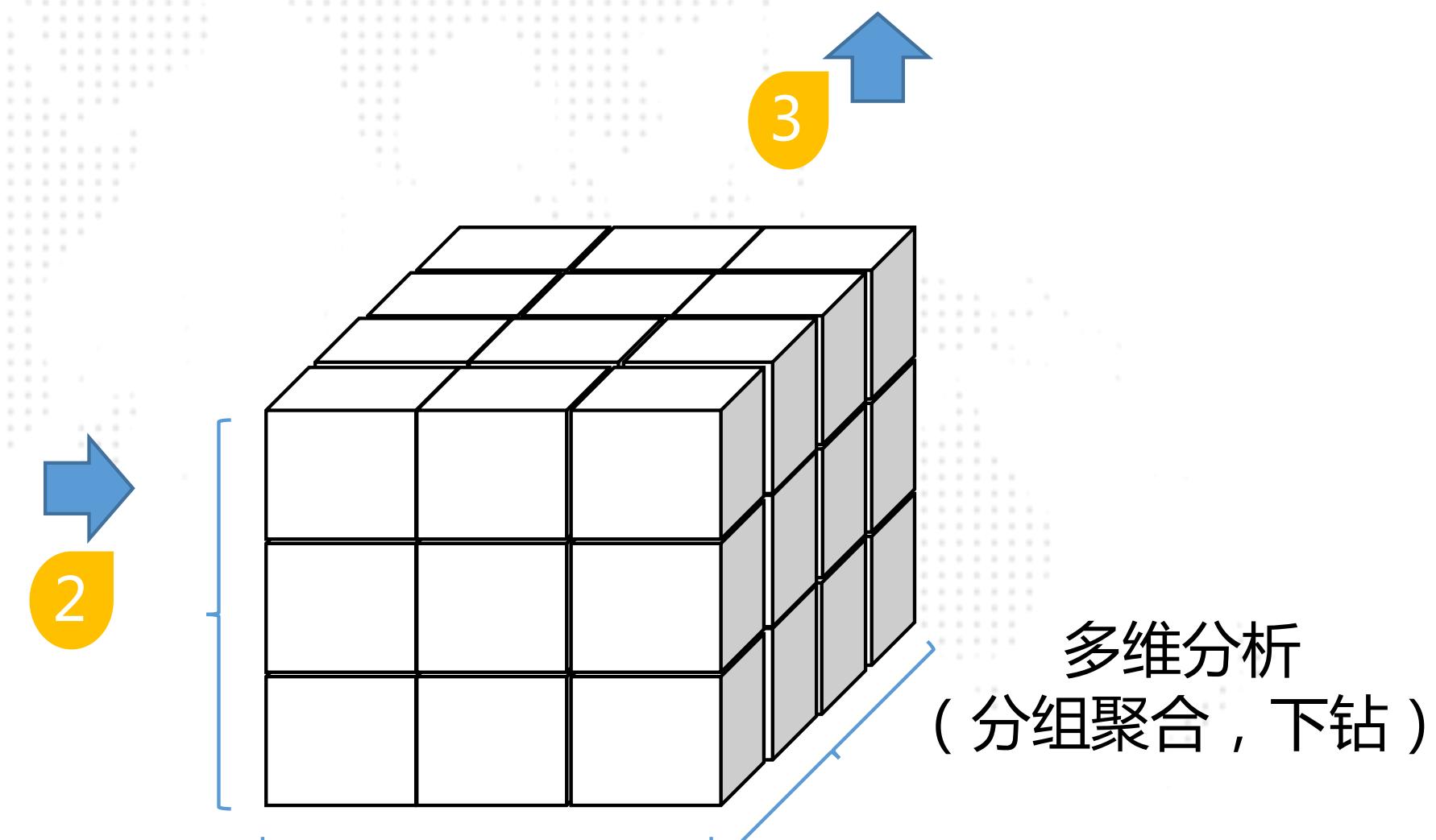
统计分析 (直方图)

Server RT <=1秒比例: ? %
Server RT <=10秒比例: ? %

4

问题	收益预估	owner	deadline

问题列表及行动策略

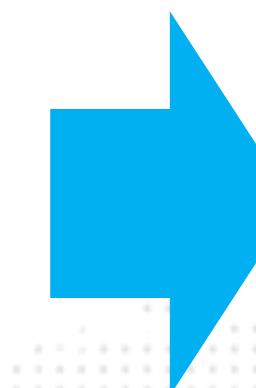


多维分析
(分组聚合 , 下钻)

06 | 效果案例



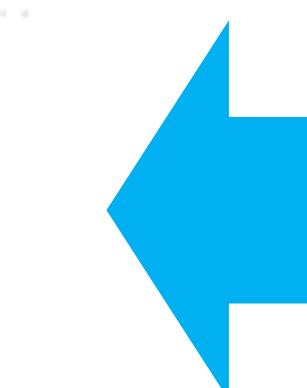
↑ 导出 ① 说明 ↗ 最大



	query_source_type	cache_source_type	count	ratio
1	garuda	tair	972	92.48
2	odps	\N	72	6.85
3	garuda	\N	7	0.67



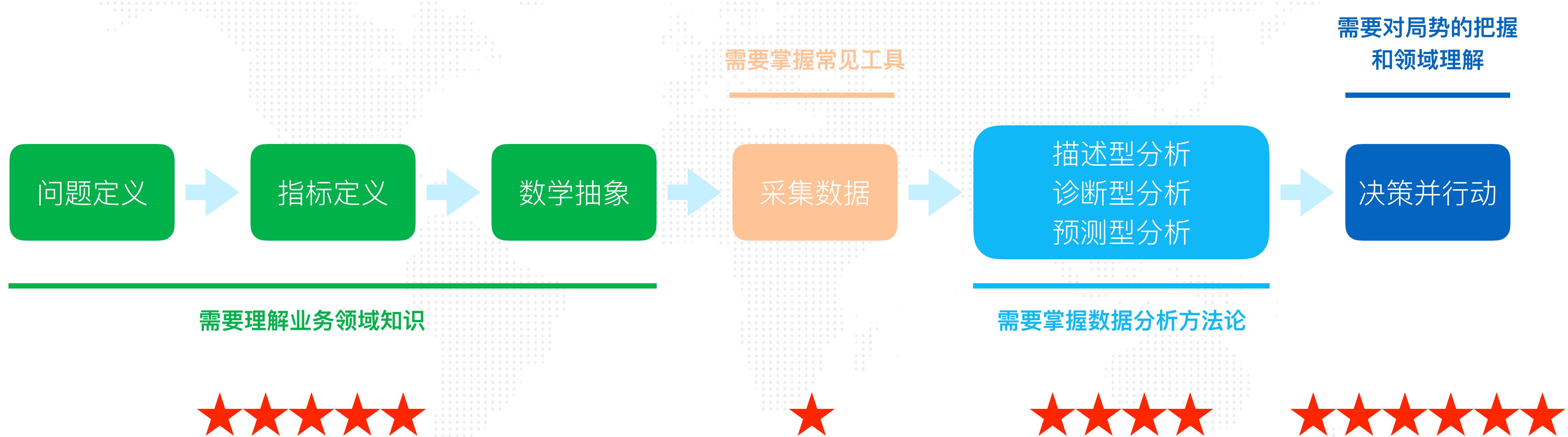
	query_source_t	cache_sou	query_mode	count ↓	ratio
1	garuda	tair	count distinct	968	92.1
2	odps	\N	other	58	5.52
3	odps	\N	count distinct	14	1.33
4	garuda	\N	other	5	0.48
5	garuda	tair	other	4	0.38
6	garuda	\N	count distinct	2	0.19



- ADS源?
- COUNT DISTINCT度量?
- 对整个链路的影响程度?



07 | 数据分析应用模式总结



05/回顾

总结回顾

* 仅限内部交流使用，如果需要公开，请联系文档作者



01 | 回顾



数据平台部

每一个微小的念头
都值得用数据浇灌



数据分析平台

报表工具
自助多维分析
智能增强分析



数据分析方法

问题定义，指标定义，数学抽象
数据采集，数据分析，决策行动
ROI





We are hiring!

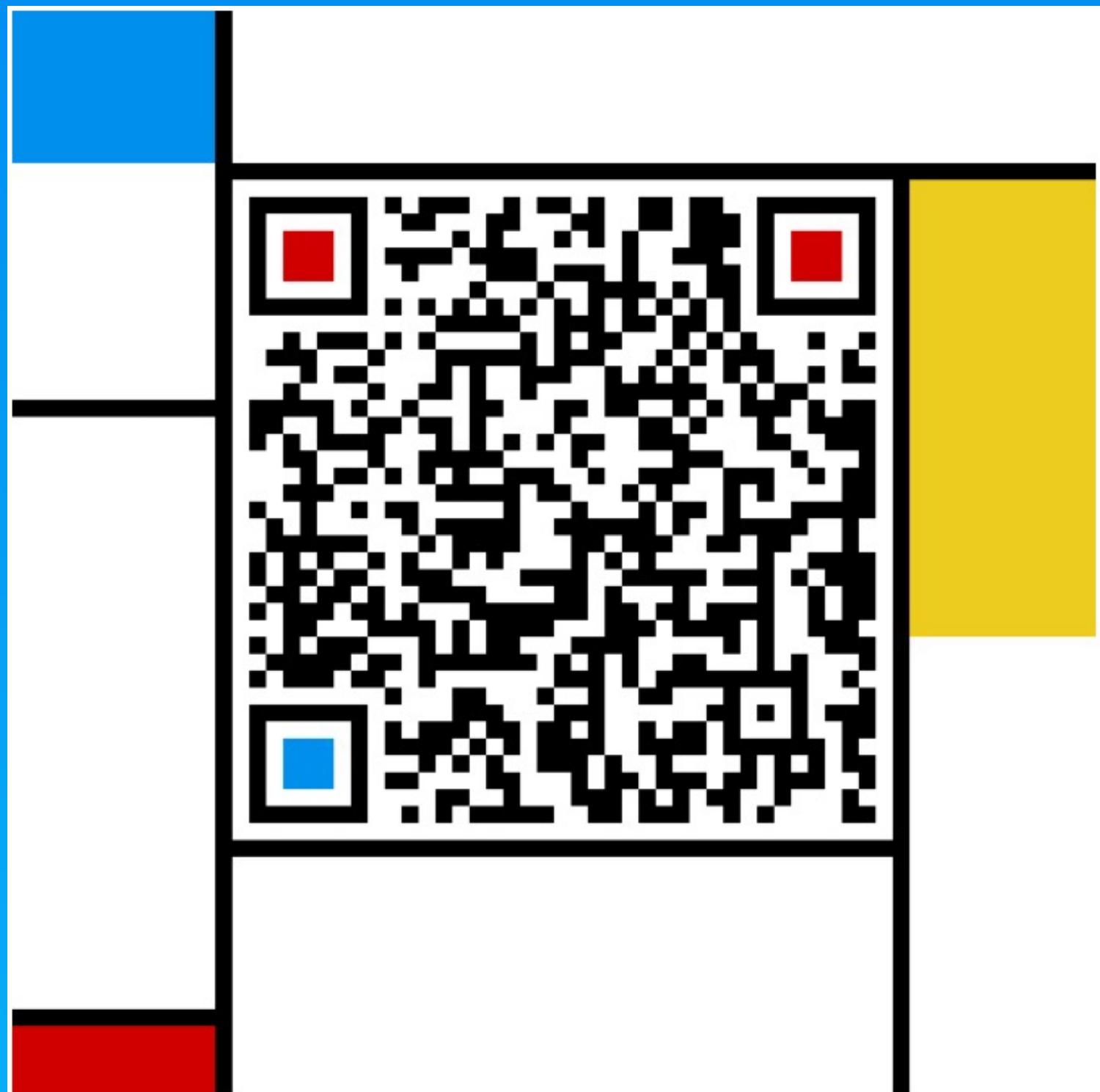
THE END

THANK YOU!

演讲人：杨军 @蚂蚁金服-数据平台部

Tel : 18157139520

E-mail : debugcool.yangj@antfin.com



* 仅限内部交流使用
如果需要公开，请联系文档作者

附 | Java技术方向

