淘宝数据仓库架构实践

薛奎

2012-04-05

淘宝网

主题

- 概述
- 元数据平台架构
- 存储计算架构
- 开发管理平台架构
- 应用开放平台架构
- 展望

概述

阿里集团未来更像一家数据公司而不是一家电商公司

淘宝数据仓库架构

元数据平台架构

应用平台架构

开发平台架构

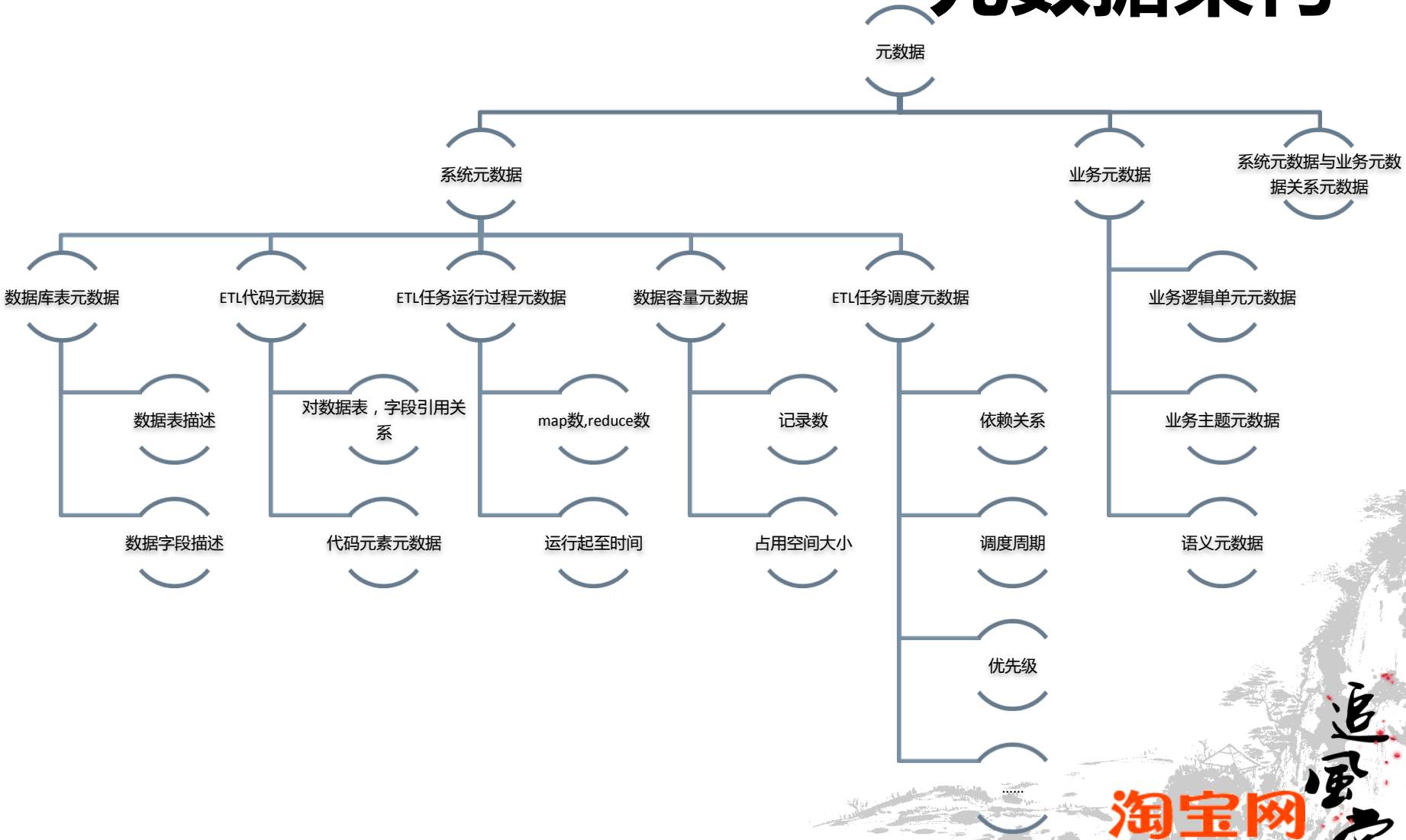
存储、计算平台架构

元数据

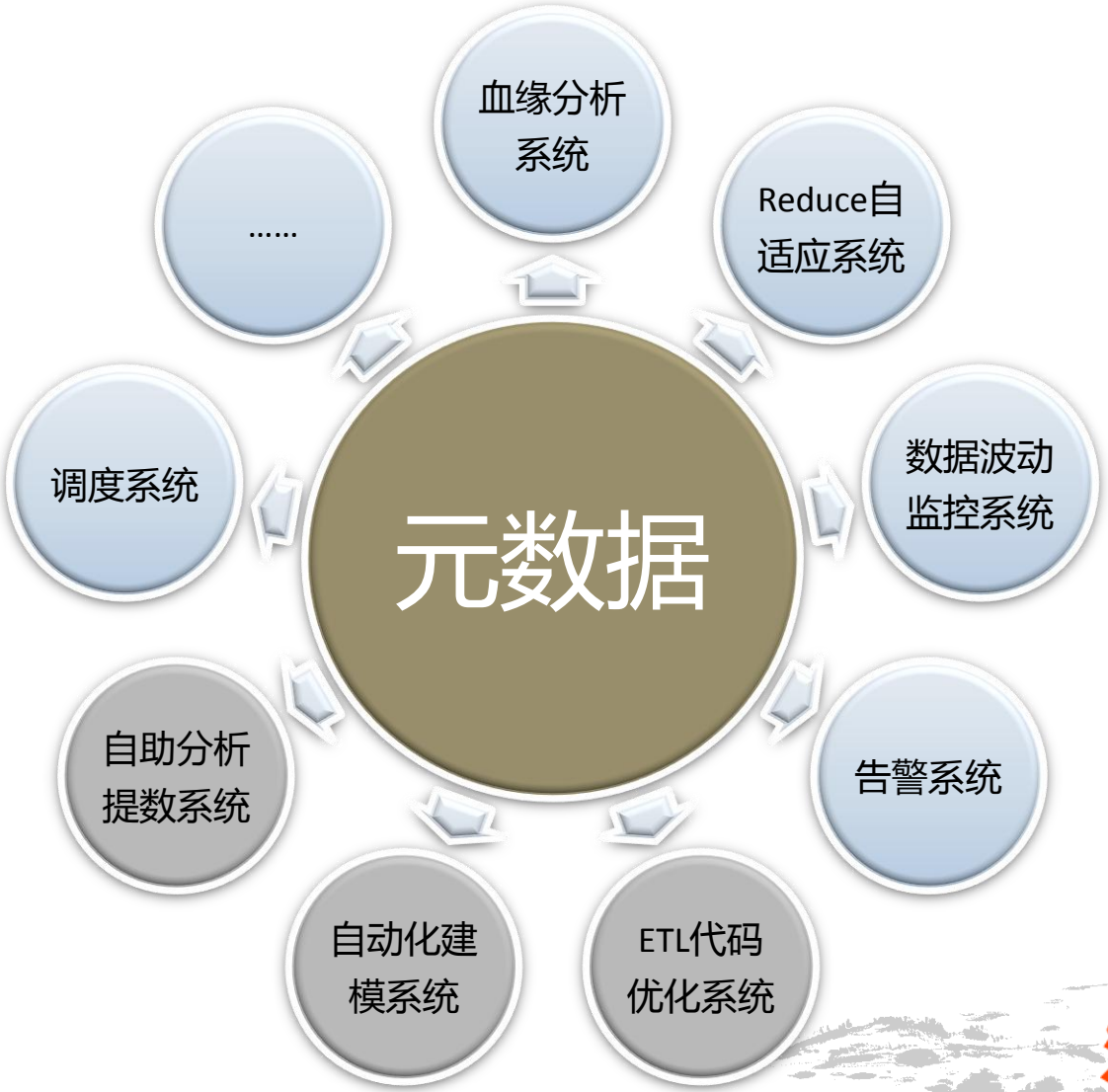
子主题

- **概述**
- **元数据平台架构**
- 元数据在淘宝中的应用

元数据架构



元数据在淘宝中的应用



存储计算架构

子主题

- **存储计算平台选型**
- **传统存储计算平台架构**
- **分布式平台设计理念**
- 淘宝存储计算平台发展

存储计算平台选型

01 规模评估

使用人数、数据量、数据保存周期、数据需求量

平台选型

02 容量评估

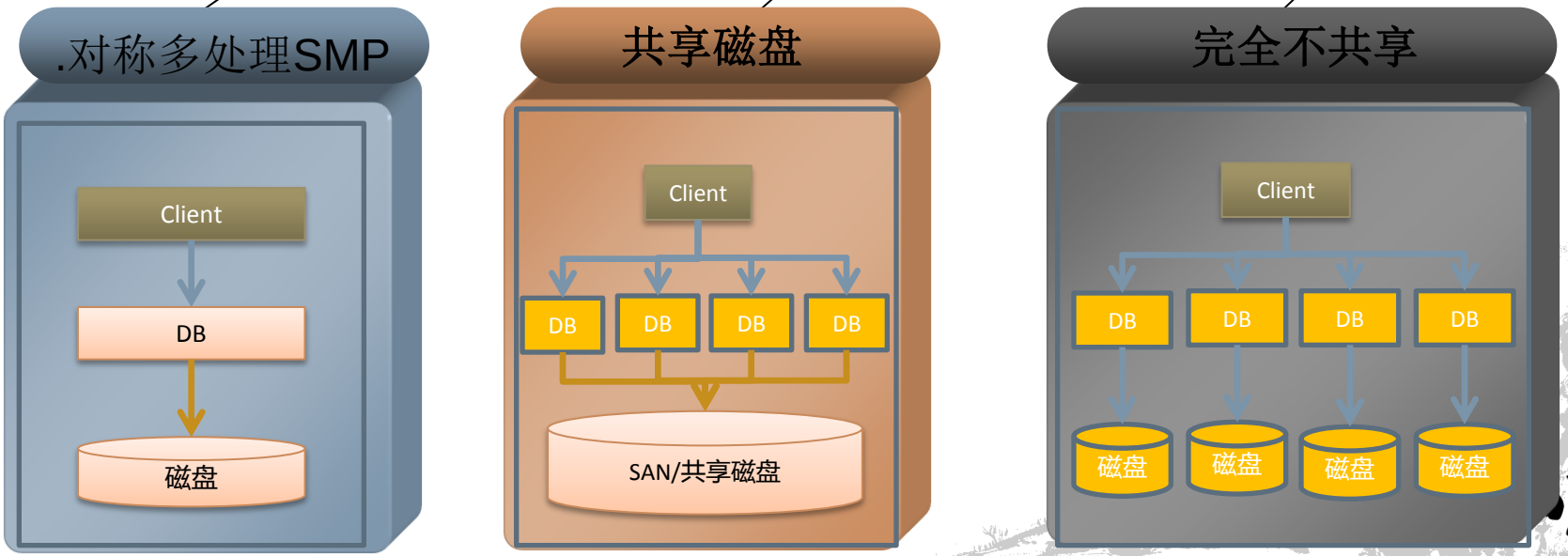
计算(CPU/内存)，存储（磁盘)，网络（网卡，路由器）.

02 需求评估

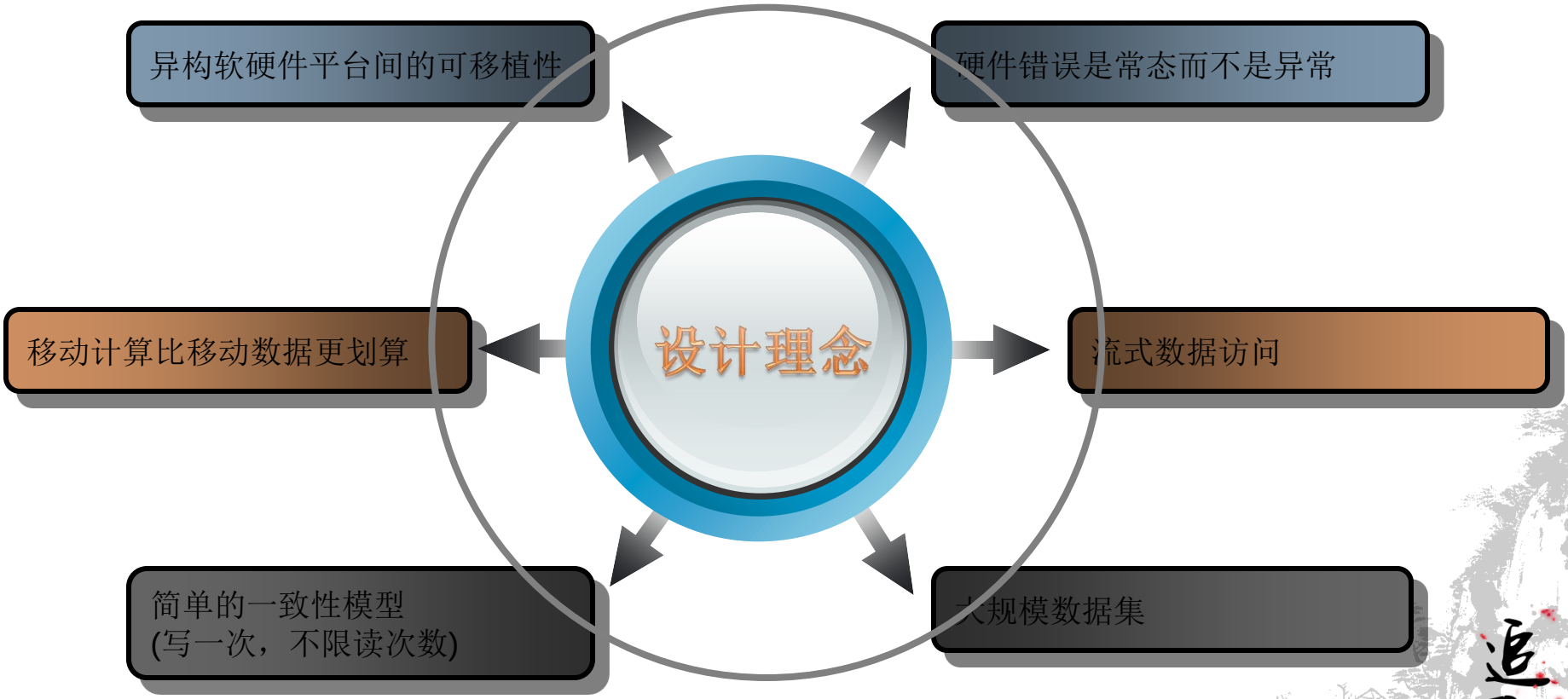
线性扩展、成本、稳定性、性能、运维……

传统数据仓库平台架构

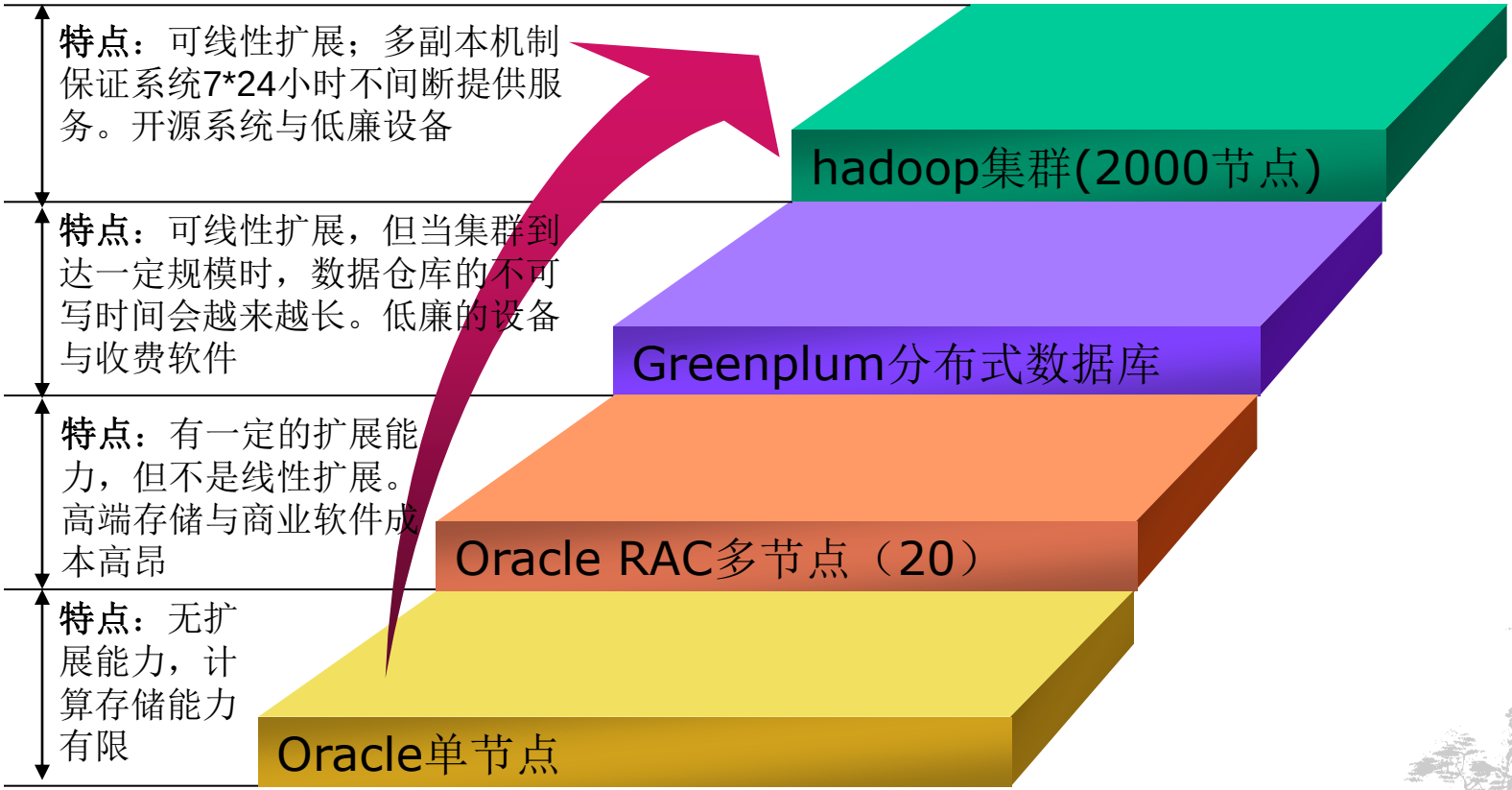
根据对节点(CPU/内存)，磁盘，网络的共享分为完全共享、部分共享与完全不共享几种类型.



分布式平台设计理念



淘宝计算存储平台发展



开发管理平台架构

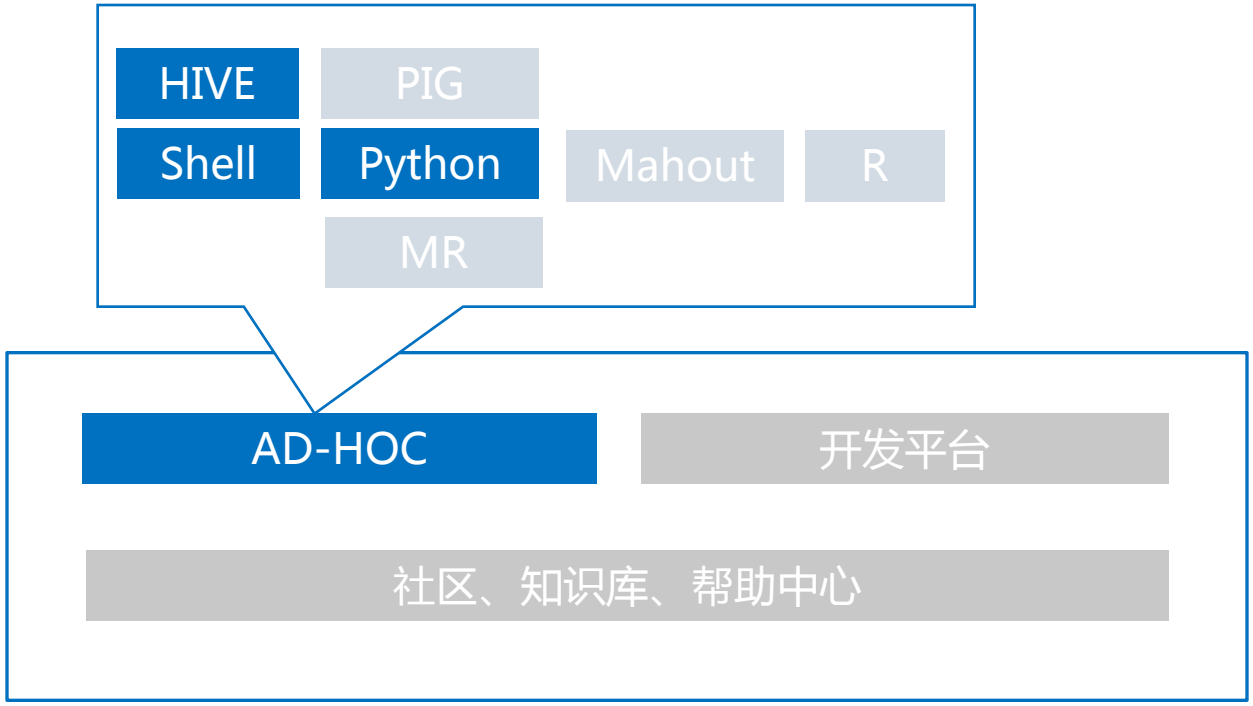
子主题

- 总体规划
- 云分析
- ETL 任务调度

总体规划


开发
云分析
问答
知识中心
测试
自动化测试平台(开发中)
布署上预发
版本/发布功能(开发中)
预发
天网
布署上生产
版本/发布功能(开发中)
冒烟
天网

云分析



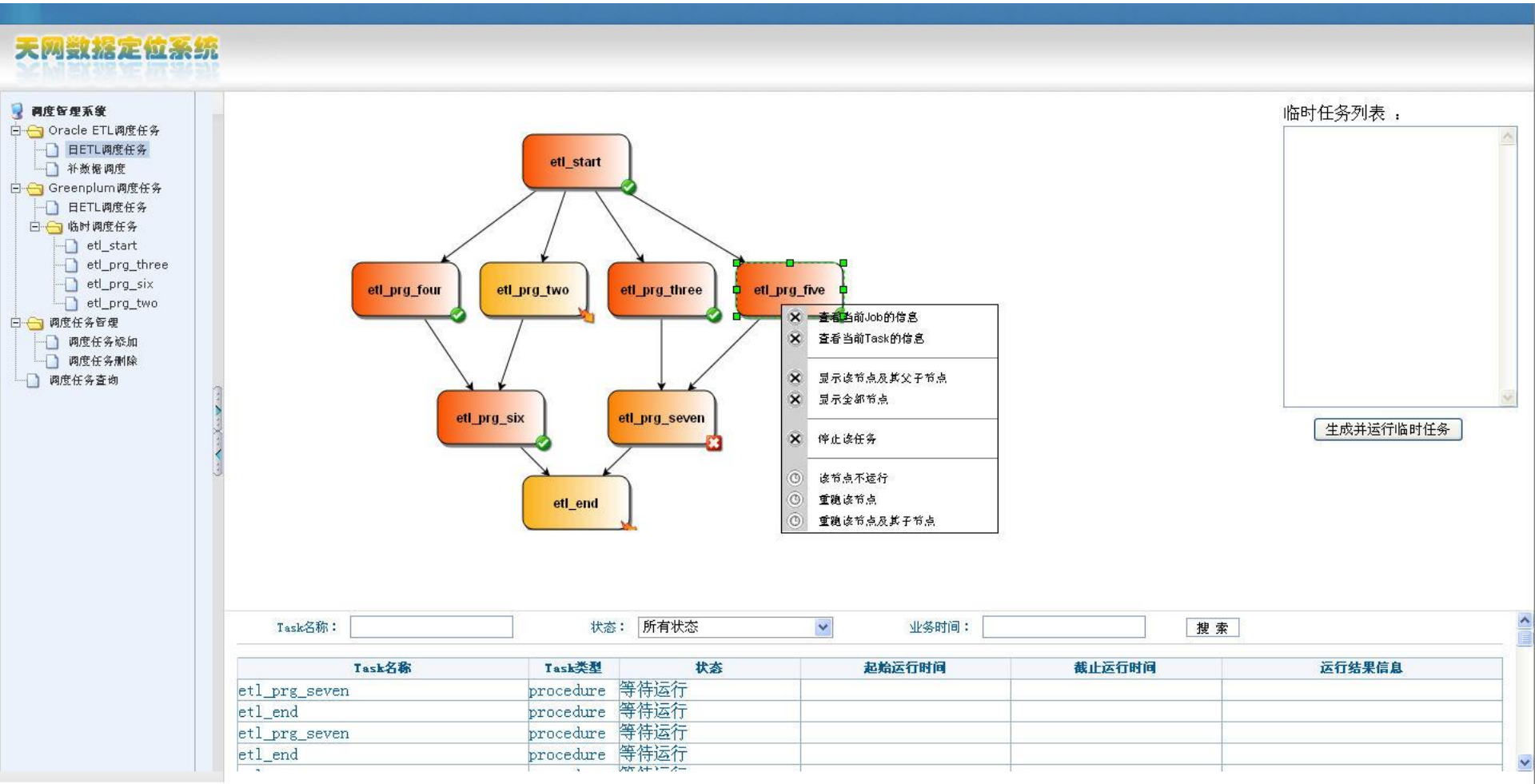
已支持

计划支持

ETL任务调度平台

调度系统之于数据仓库有如大脑于人体一样重要，他是数据仓库所有任务高度协同有序运转的指挥中心.

早期天网原型

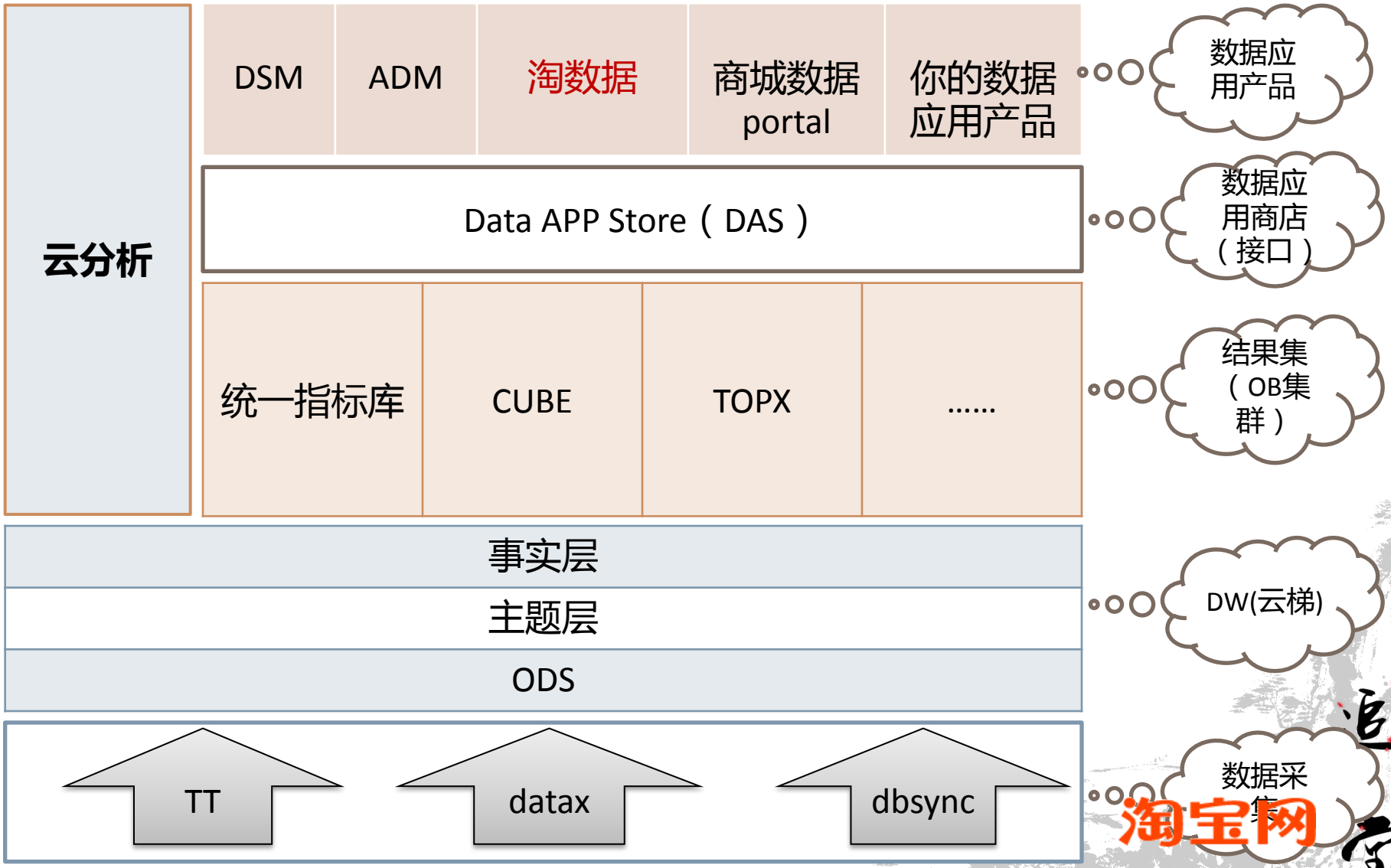

天网数据定位系统
调度管理系统
Oracle ETL调度任务
日ETL调度任务
补数据调度
Greenplum调度任务
日ETL调度任务
临时调度任务
etl_start
etl_prg_three
etl_prg_six
etl_prg_two
调度任务管理
调度任务添加
调度任务删除
调度任务查询
etl_start
etl_prg_four
etl_prg_two
etl_prg_three
etl_prg_five
etl_prg_six
etl_prg_seven
etl_end
查看当前Job的信息
查看当前Task的信息
显示该节点及其父子节点
显示全部节点
停止该任务
该节点不运行
重跑该节点
重跑该节点及其子节点
临时任务列表 ：
生成并运行临时任务
Task名称：
状态：
所有状态
业务时间：
搜 索
Task名称
Task类型
状态
起始运行时间
截止运行时间
运行结果信息
etl_prg_seven procedure 等待运行
etl_end procedure 等待运行
etl_prg_seven procedure 等待运行
etl_end procedure 等待运行

应用开放平台架构

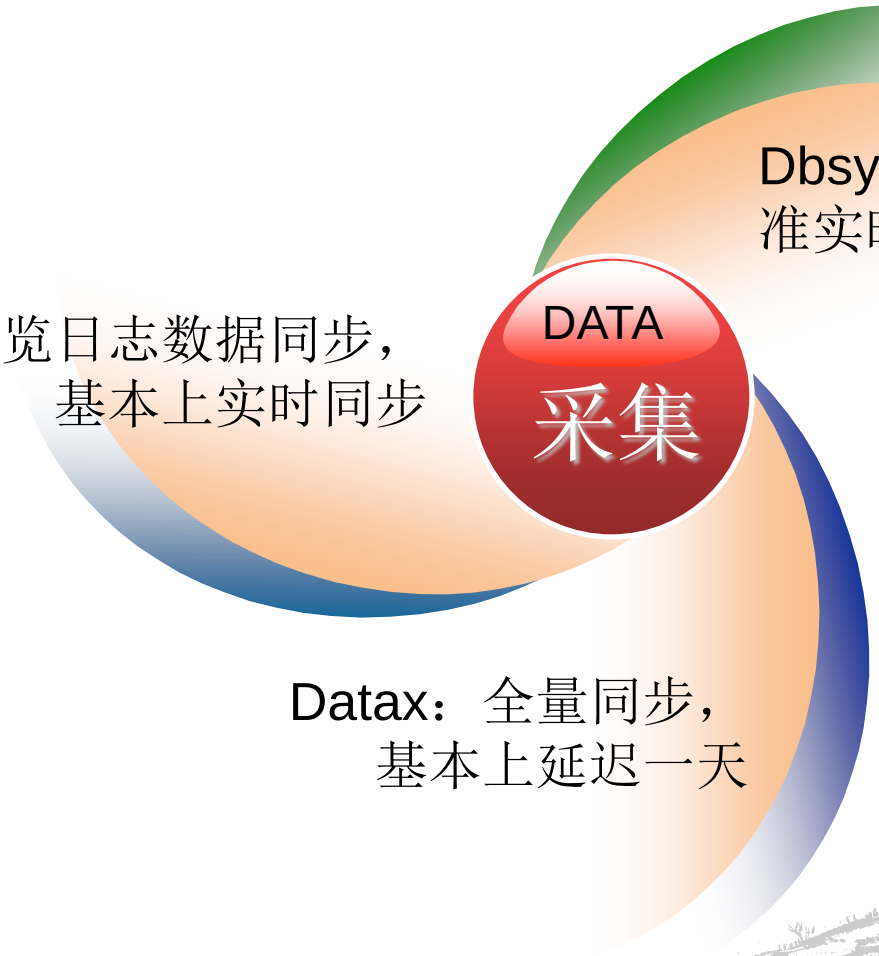
子主题

- 总体规划
- 数据采集
- 统一淘宝数据体系
- 统一指标库、CUBE群、TOP结果集
- 数据应用商店DAS(Data APP Store)
- 官方数据应用：DSM、ADM

总体规划


云分析
DSM
ADM
淘数据
商城数据 portal
你的数据应用产品
数据应用产品
Data APP Store（ DAS ）
数据应用商店（接口）
统一指标库
CUBE
TOPX
……
结果集（OB集群）
事实层
主题层
ODS
DW(云梯)
TT
datax
dbsync
数据采集

数据采集

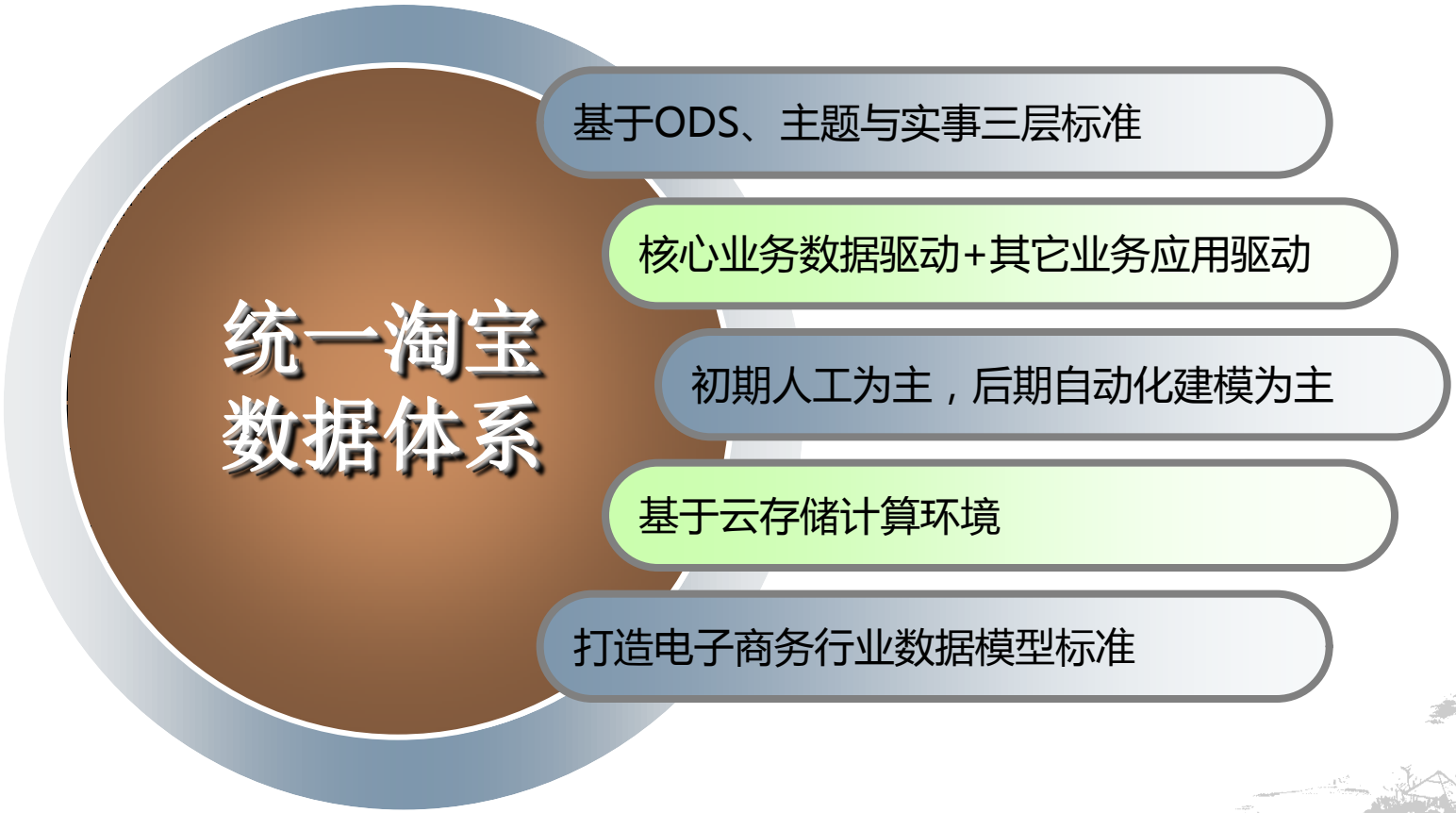


Dbsync：DB log解析，
准实时同步

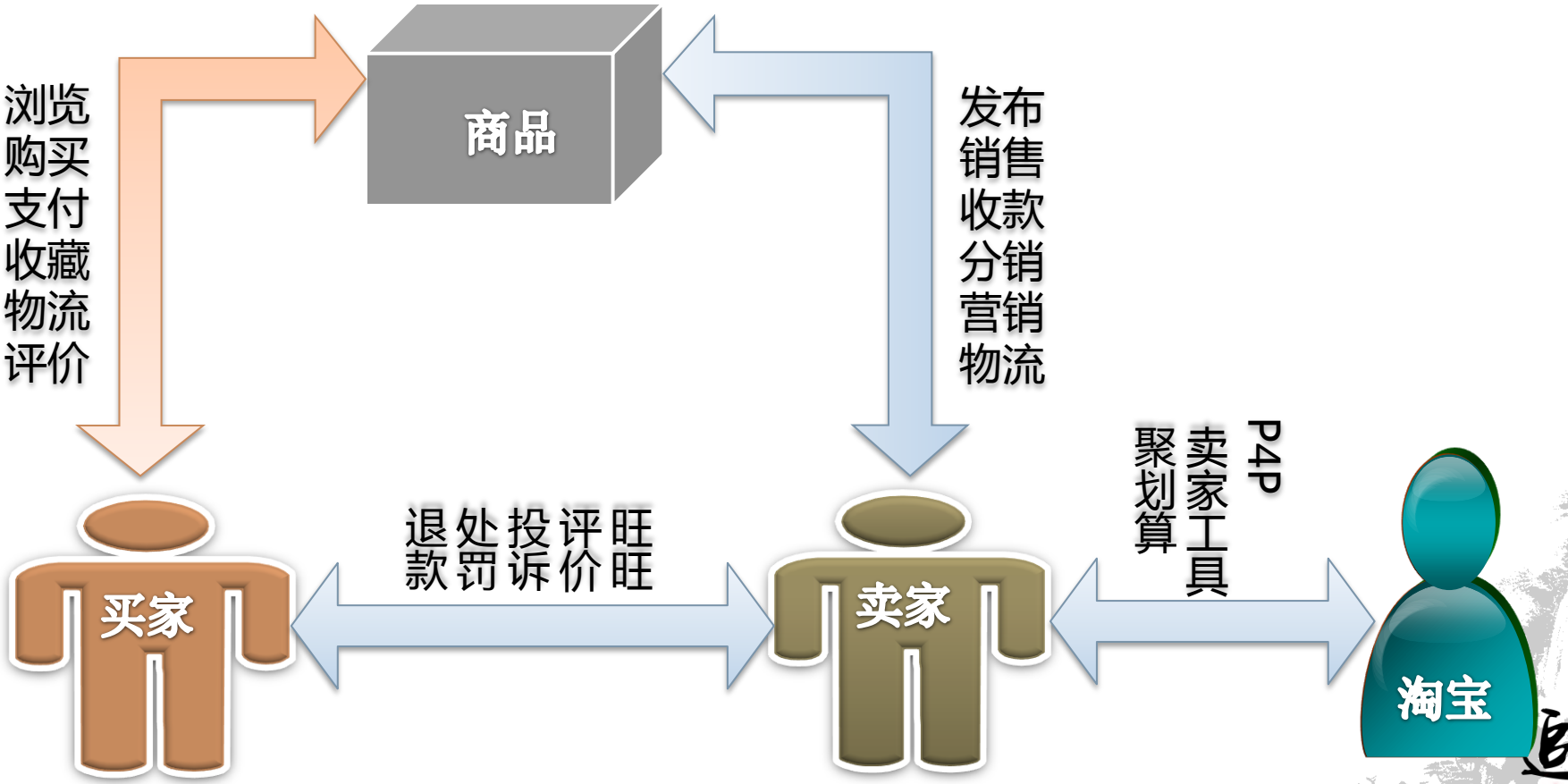
TT：浏览日志数据同步，
基本上实时同步

Datax：全量同步，
基本上延迟一天

统一淘宝数据体系


统一淘宝
数据体系
基于ODS、主题与实事三层标准
核心业务数据驱动+其它业务应用驱动
初期人工为主，后期自动化建模为主
基于云存储计算环境
打造电子商务行业数据模型标准

淘宝业务模型


商品
浏览
购买
支付
收藏
物流
评价
发布
销售
收款
分销
营销
物流
买家
卖家
退款
处罚
投诉
评价
旺旺
P4P
卖家工具
聚划算
淘宝

统一指标库—生成过程

维度								指标		
W1	W2	W3	W4	W5	W6	W7	W8	I1	I2	I3
周期	一级类目	地域	卖家性别	年龄段	卖家星级			GMV	支付宝成交	PV

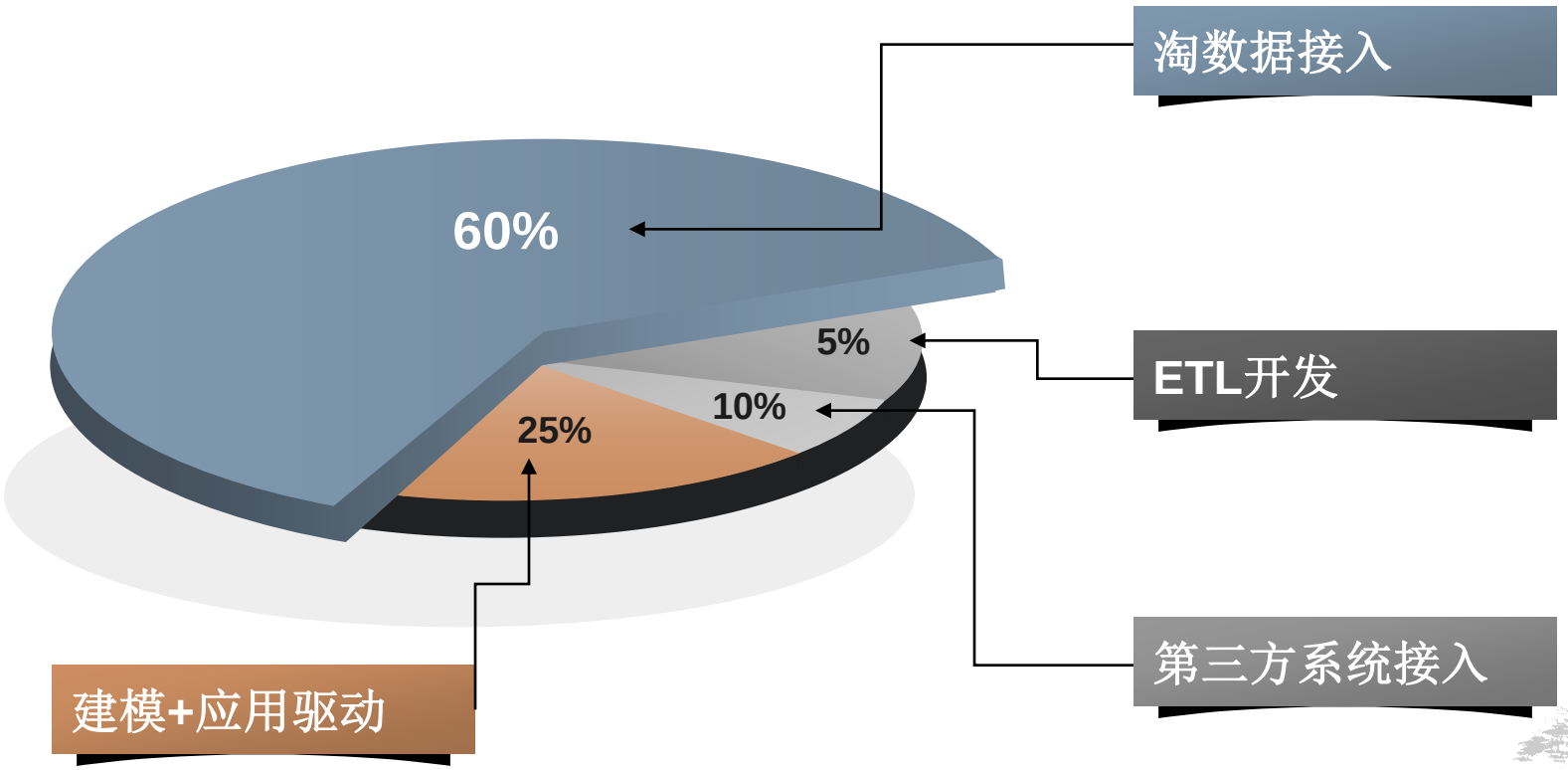
任意维度与指标组合维一确定一个指标

指标库

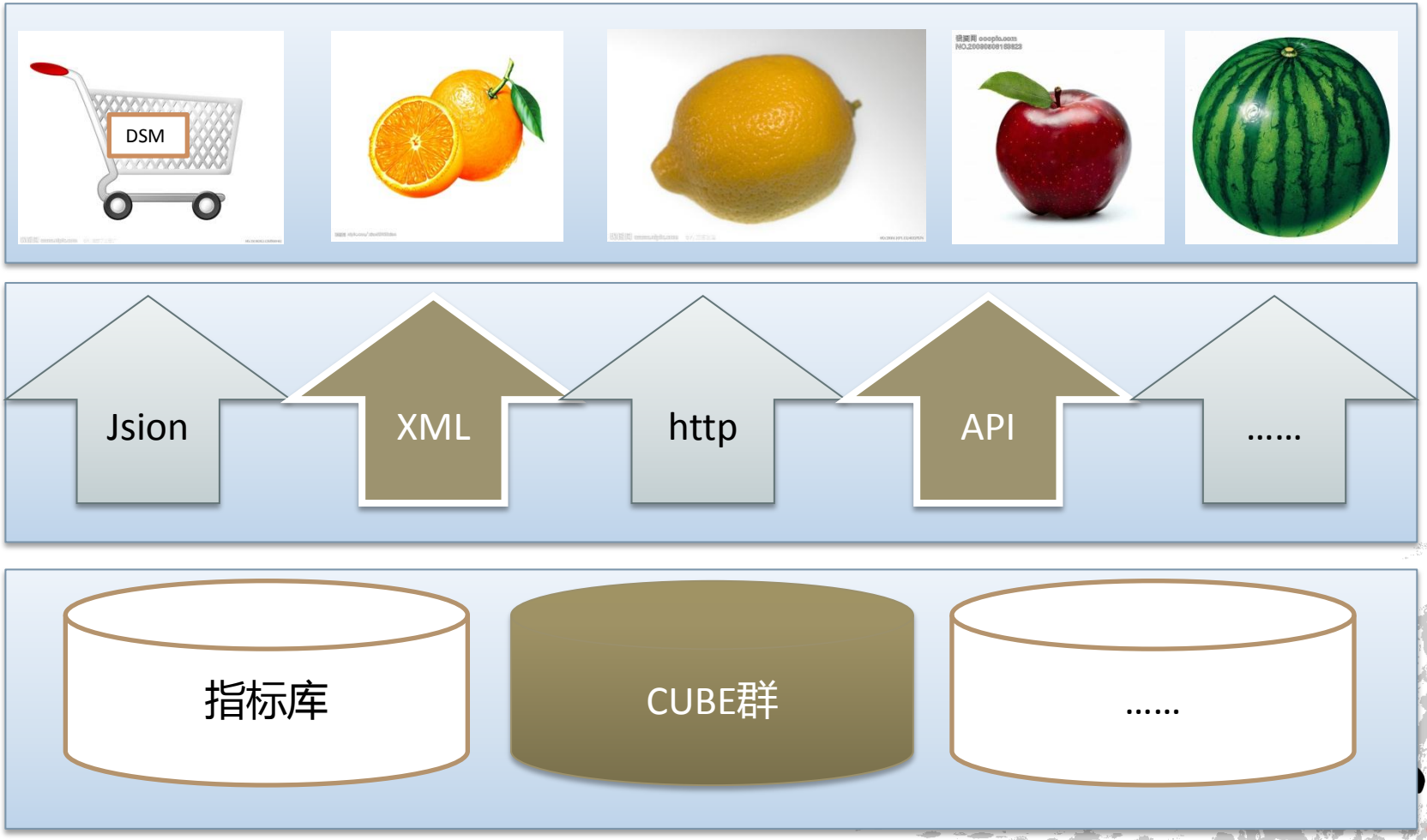
唯一标识	指标名称	度量代码	度量值(元)	标签
20120401001	周期=日 | 一级类目=男装 | 地域=上海 | 日交易大于等于1W的店铺	Sum(GMV)	300000000	GVM|男装|上海|网站运营部|男装运营

统一指标库—目前指标来源

统一指标库：逐步切换为统一建模+应用驱动的来源，目前绝大多数指标库来源为淘数据指标



Data App Store(DAS)



DSM系统

Data Super Market：简称DSM，就像在超市购物一样获取你想要的数据，从此您只需要看一张报表，100%DIY By Yourself.

格式一次定义，永久生成

数据一次定义，定期自动产生

取你所想，用你所用

支持EXCEL的所有编辑功能

支持excel导出

支持定期邮件发送功能

通过业务元数据定位数据

搜索的方法查找数据

展望

数据像水一样，无处不在

联系我们

- **数据平台与产品**
  - ✓ Blog： http://www.tbdata.org/
  - ✓ 百科：
  - ✓ 邮件列表：taobao-dw@list.alibaba-inc.com
- **薛奎**
  - ➢ 微博：淘薛奎
  - ➢ mail：xuekui@taobao.com
  - ➢ 旺旺：薛奎

淘宝网