神马搜索大数据基础架构

黄锐华/阿里巴巴(神马搜索)

预留位置，不放PPT内容

内容提纲

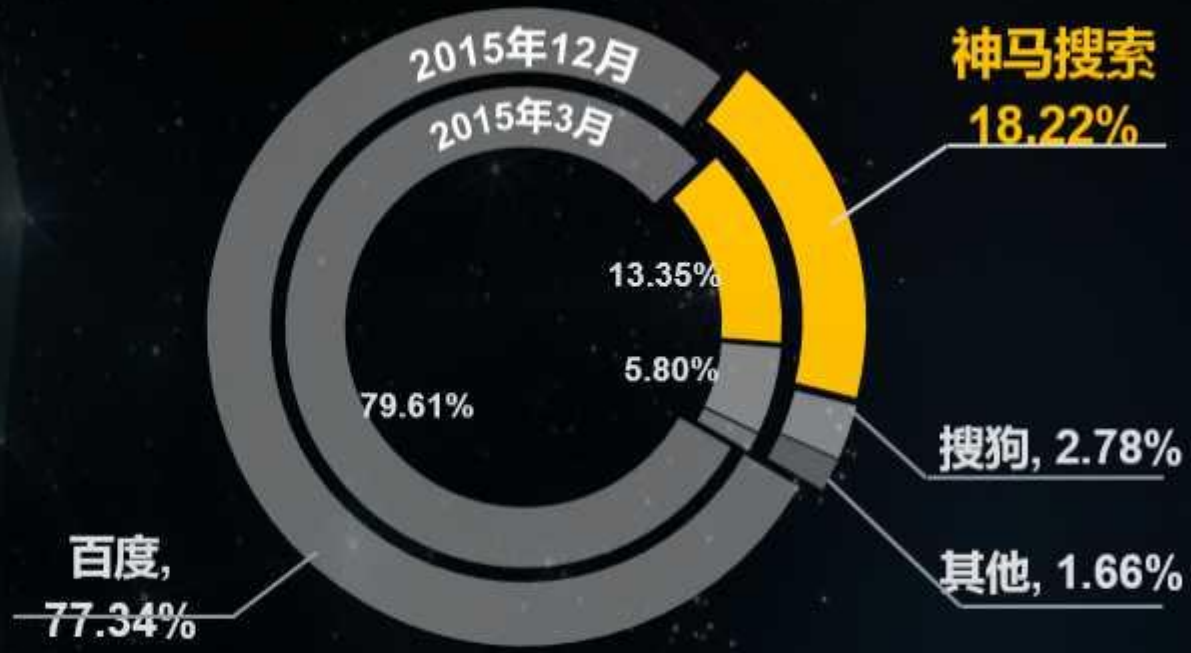
预留位置，不放PPT内容

神马大数据平台整体架构

预留位置，不放PPT内容

关于神马搜索

神马搜索
移动搜索第二



数据来源：CNZZ & 全球流量监测机构StatCounter 2015

预留位置，不放PPT内容

平台发展3个阶段

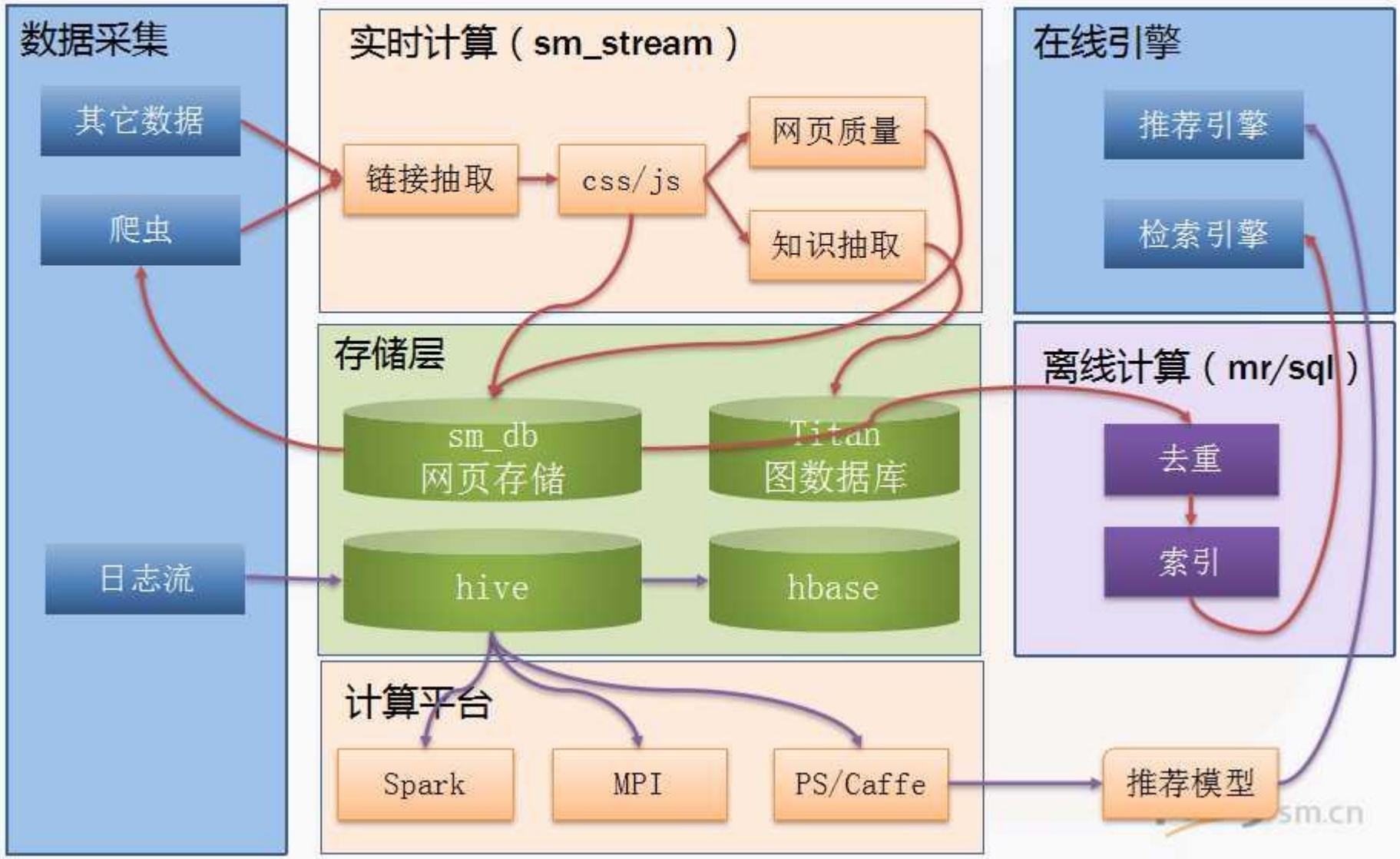
阶段1，满足大规模数据存储和（时效性）计算。
业务代表：抓取、索引、排序

阶段2，满足复杂的数据挖掘和处理需求。
业务代表：日志挖掘、数据融合、推荐模型

阶段3，功能和平台的标准化
代表：通用算法、通用调度、流程语言

预留位置，不放PPT内容

大数据平台整体架构



预留位置，不放PPT内容

网页存储系统（sm_db）

预留位置，不放PPT内容

存储需求

时效性需求

- 每天百亿级实时读写
- 热点新闻、股票、实时评论的秒级展现

超大规模存储

- 千亿网页，万亿元素
- 网页、新闻、小说、app、人物、旅游、财经 等类型

扫描性能

- 各索引流程的超大规模扫描
- 网页计算的扫描需求

数据管理

- 复杂的数据合并清理逻辑
- 重复抓取带来的大量冗余

预留位置，不放PPT内容

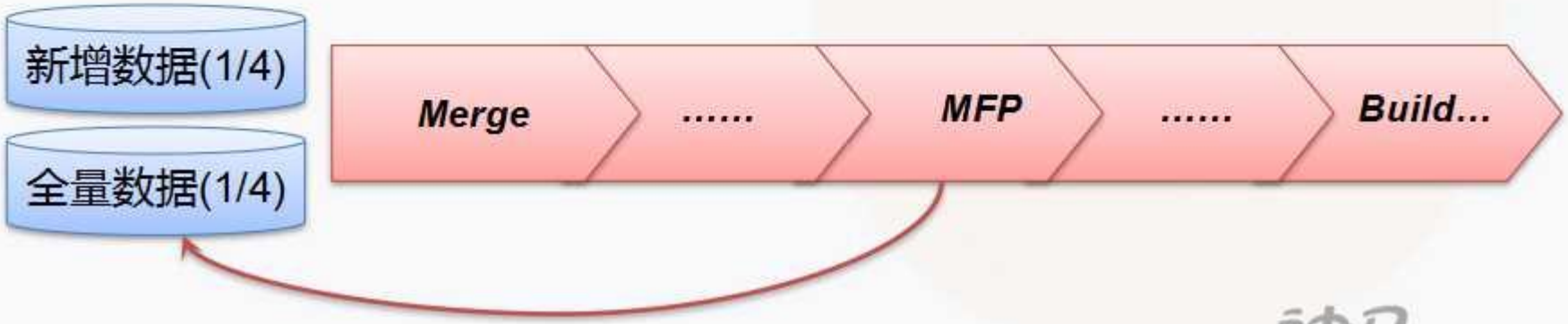
hbase使用体验

Hbase是最优选择，但依然无法直接使用。

预留位置，不放PPT内容

基于DFS的流程搭建方式

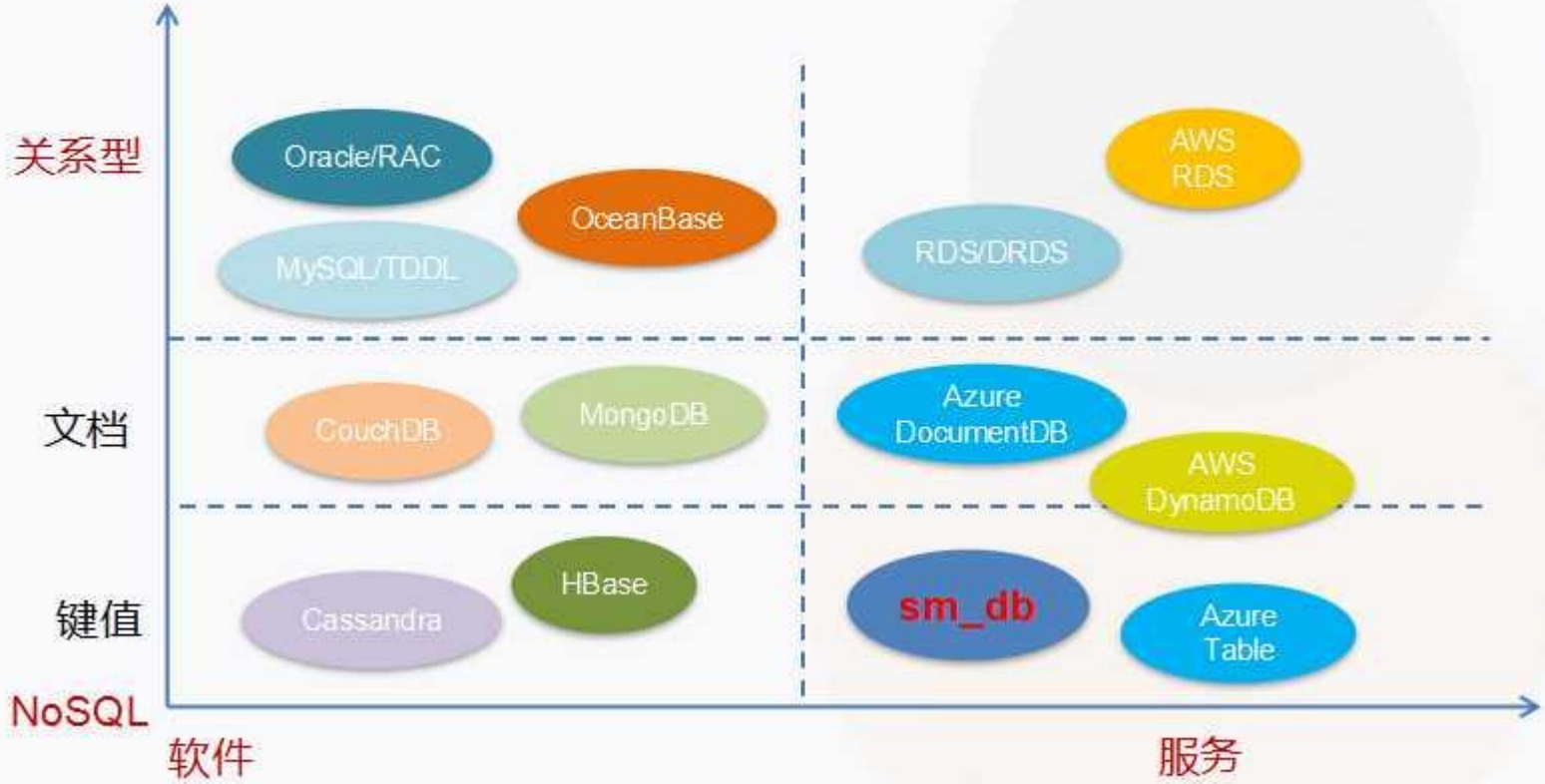
- 直接基于DFS的流程（不使用存储服务）
  - 只能分片处理
  - 无法满足时效性需求
  - 数据和流程无法共享
  - 严重限制业务灵活性



预留位置，不放PPT内容

神马存储解决方案（sm_db）

自建**分布式结构化存储系统**



预留位置，不放PPT内容

sm_db设计

vs hbase

功能简化

- 固定分区
- 固定key-range
- Buffer替换Cache

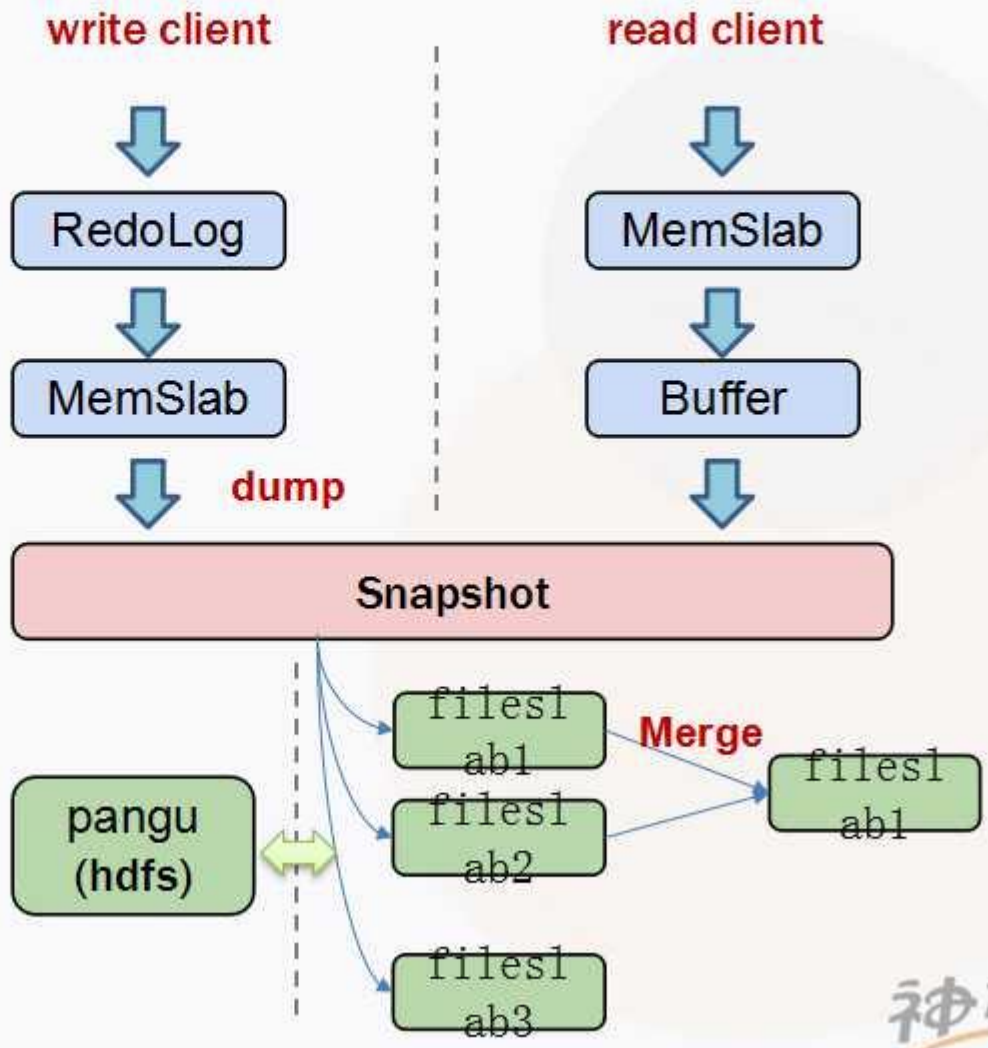
功能增强

- 开放Merge、Clean策略
- snapshot机制
- 多partition

预留位置，不放PPT内容

sm_db结构

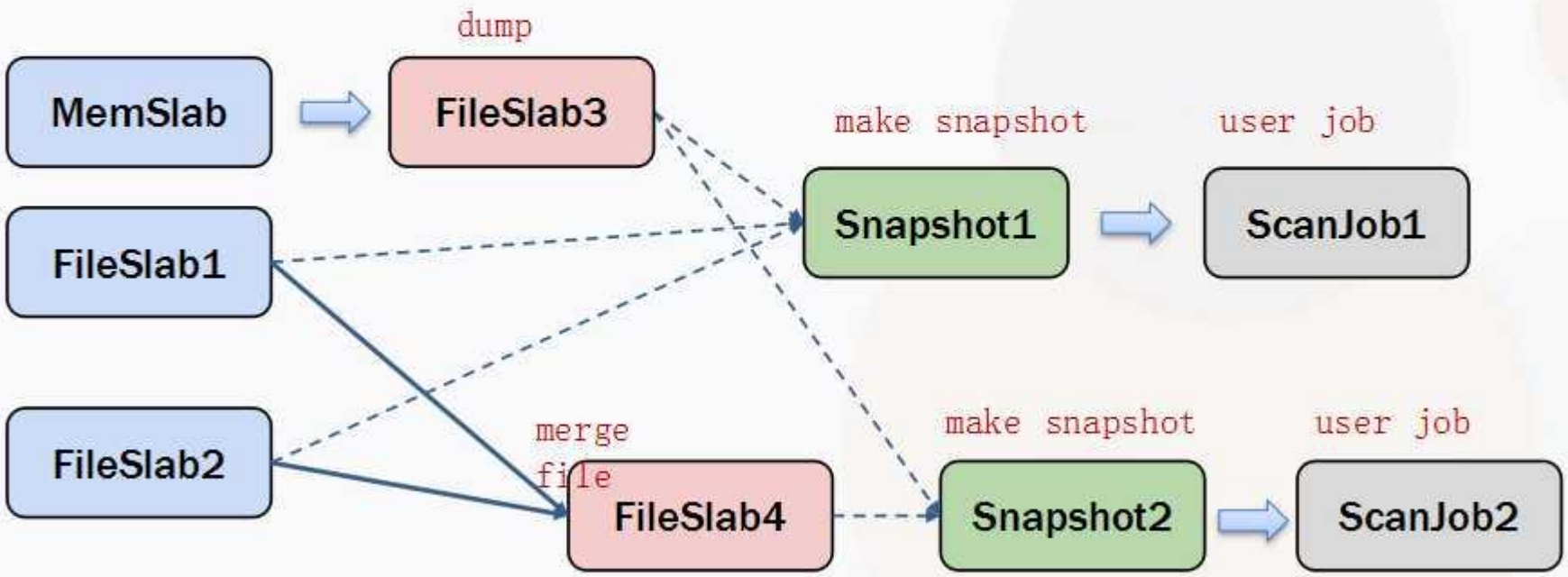
Partition内部结构



预留位置，不放PPT内容

sm_db扫表

通过Snapshot机制，直接扫描文件本身，并保证不同扫描任务并行。



预留位置，不放PPT内容

sm_db应用规模

数据大小：>> 10PB
数据数量：千亿 +
每天写入：百亿级别
每秒读取：百万级别
表规模　：100+
机器规模：1000+

预留位置，不放PPT内容

实时计算平台（sm_stream）

预留位置，不放PPT内容

实时计算需求

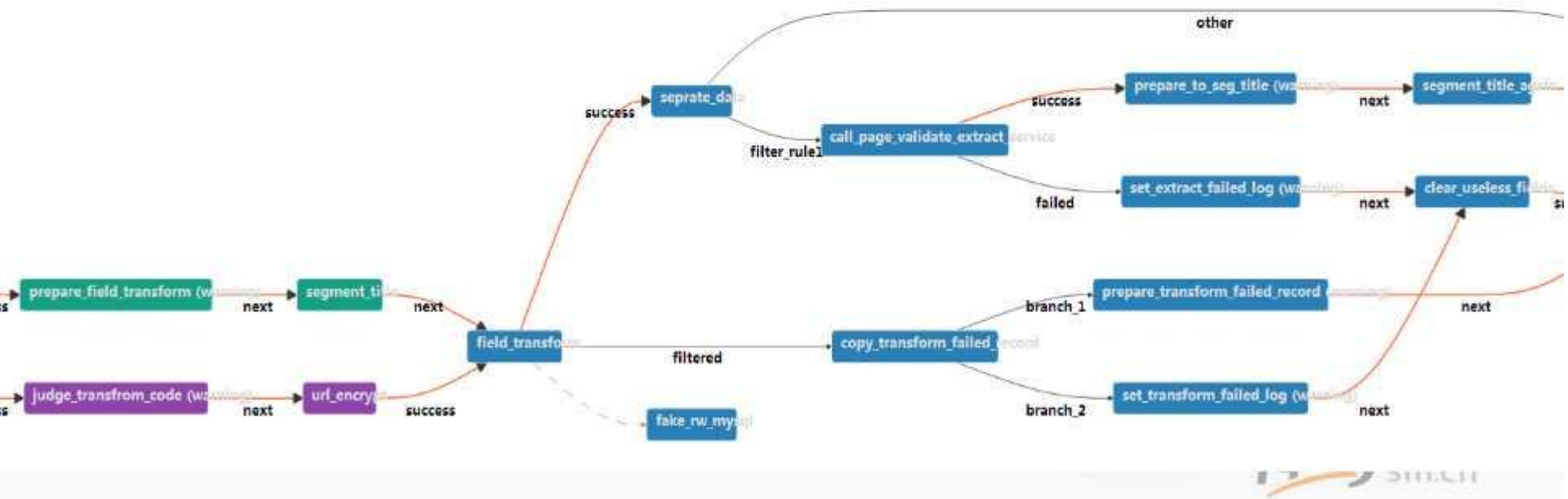
- 业务时效性需求
  - 股票、评论、新闻
- 管理复杂流程
- 数据/流程共享
- 提升开发效率
- 降低维护成本

预留位置，不放PPT内容

sm_stream 可视化

- 流程可视化
- 性能可视化
- 实时sample数据

预留位置，不放PPT内容



sm_stream 性能分析

节点性能指标

请选择instance: 0

节点名	线程数	CPU/线程	节点CPU	队列百分比	实际队列长度	配置队列长度	输入OPS
page_changed	10	0.06%	0.64%	100.00%	50	50	32.26
set_db_read_timeout	1	0.14%	0.14%	100.00%	10	10	32.26
qps_control	1	1.02%	1.02%	100.00%	50	50	22.79
url_backup	1	0.41%	0.41%	100.00%	20	20	32.26
crawler_field_process	1	0.34%	0.34%	100.00%	20	20	32.26
generate_timestamp	1	0.51%	0.51%	0.00%	0	10	21.85
page_intention_detect	20	2.04%	40.80%	0.00%	0	50	21.85

预留位置，不放PPT内容

sm_stream 标准化

- 插件实现标准化
- 基于配置使用，免去代码依赖
- 流程即服务
- 丰富功能库

	[illegible]	eature的计算后，更新到在线的Tair
	add_simhash	生成新闻的simhash值，用于新闻去重
mysql读写	mysql_reader	mysql读取插件，一般使用call_processor调用，提供select功能。
	mysql_writer	mysql写入插件，一般使用call_processor调用，提供update、insert功能
	add_group	field添加group

预留位置，不放PPT内容

sm_stream 配套体系

- 监控运维体系
- 多机群调度
- 版本管理
- 多平台支持

预留位置，不放PPT内容

sm_stream应用

- 部署多个机群，几千台机器
- 服务神马、uc、高德等业务
- 服务几百个线上业务流程
- 提供几百个标准功能

预留位置，不放PPT内容

图数据库（Titan）

预留位置，不放PPT内容

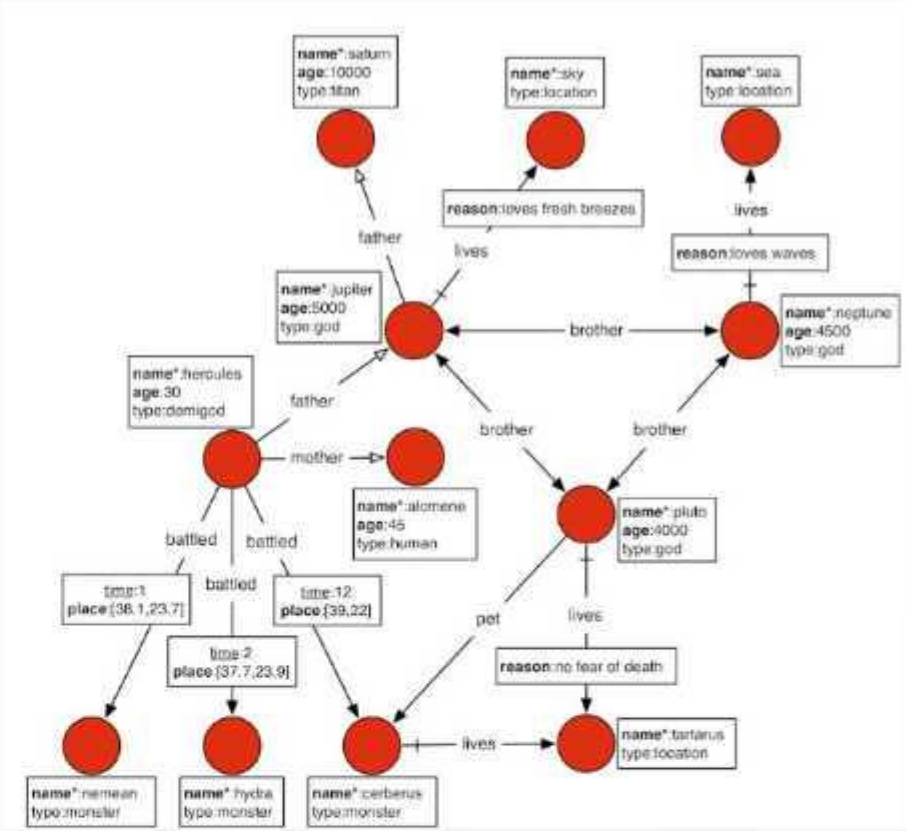
业务场景

- 满足知识化搜索需求

预留位置，不放PPT内容

图数据存储特性

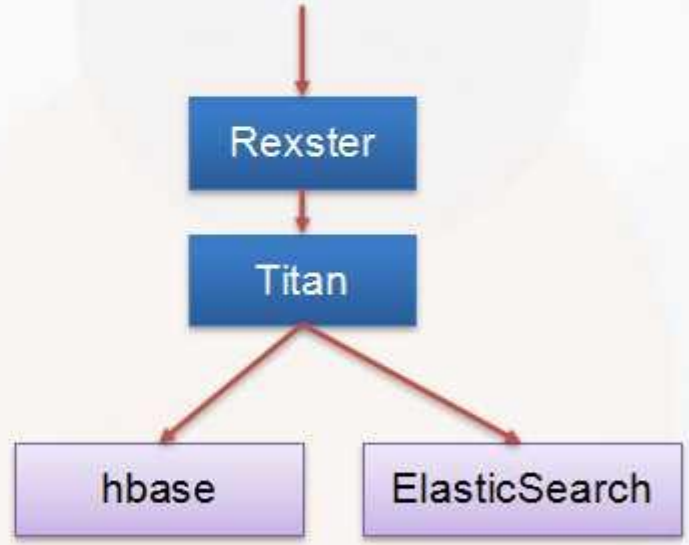
- 存储特性
  - 大量实体、关系、属性
  - 不同行业差异大
  - 超级节点
- 存储方案
  - **Titan-with-HBase-and ElasticSearch**
    - **Titan：分布式的图数据库**
    - **Hbase：结构化存储服务**
    - **ElasticSearch：分布式搜索引擎**
  - 超级节点分partition存储



预留位置，不放PPT内容

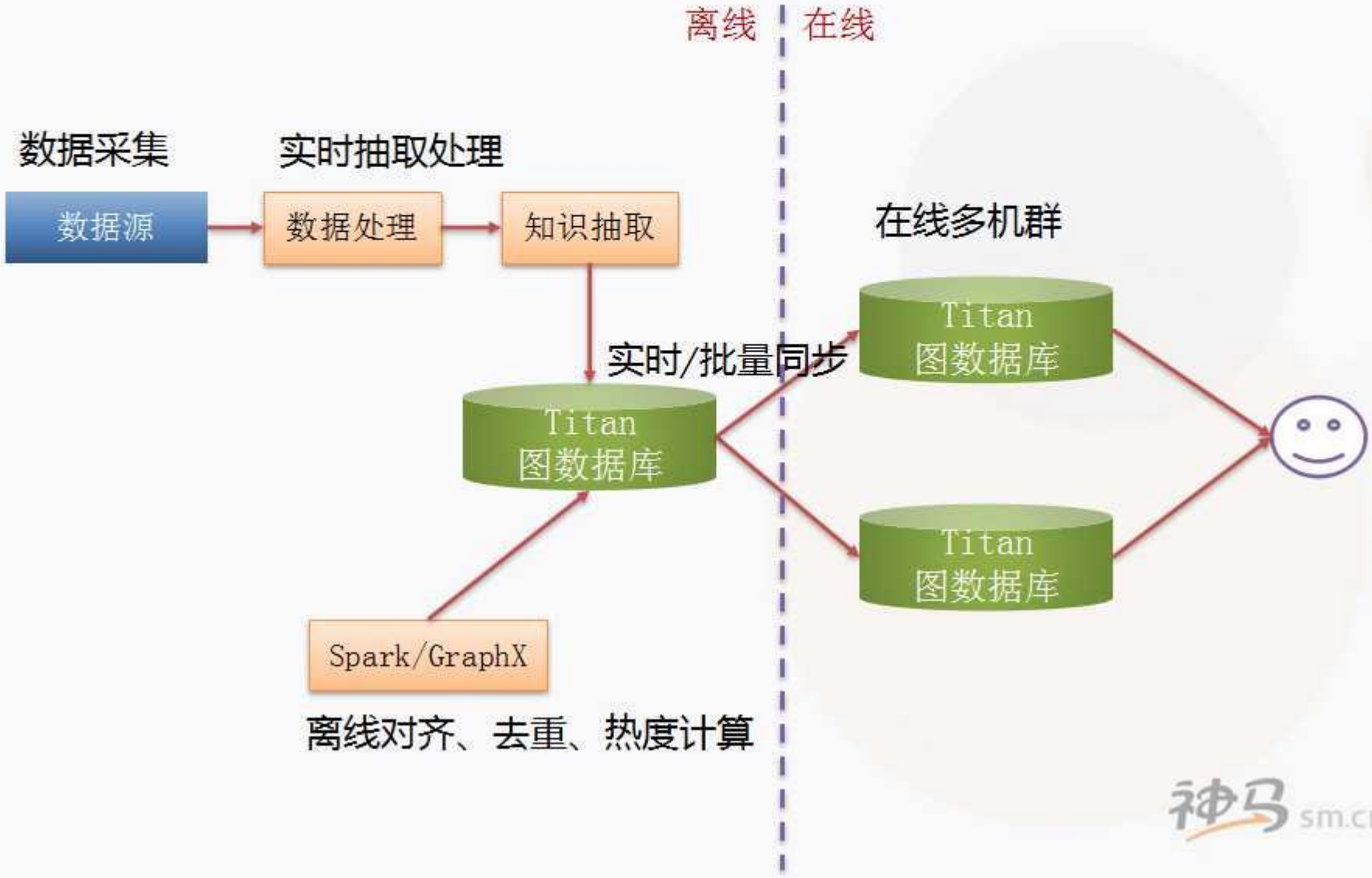
图数据计算特性

- 基于图的推导
  - 姚明的女儿的母亲的老公
- 基于图的查询
  - 周杰伦的个人资料
- 基于图的检索
  - 80年代的港台男歌星
- 数据融合
  - Spark GraphX



预留位置，不放PPT内容

业务流程



预留位置，不放PPT内容

Titan改进

- 邻接点属性排序
- 超时保护机制
- 解决内存泄漏
- 过滤条件读取

预留位置，不放PPT内容

应用效果

- 结果直达，无需二次点击
- 搜索结果更精确
- 更丰富的内容展现
- 服务几十个行业

预留位置，不放PPT内容

计算平台

预留位置，不放PPT内容

多平台共存

- 不同业务对平台需求不同

日志挖掘
- MR job
- Hive

推荐算法
- Spark
- mpi

数据融合
- Titan
- Graph

深度学习
- Caffe
- TensorFlow

预留位置，不放PPT内容

计算平台整体架构

预留位置，不放PPT内容

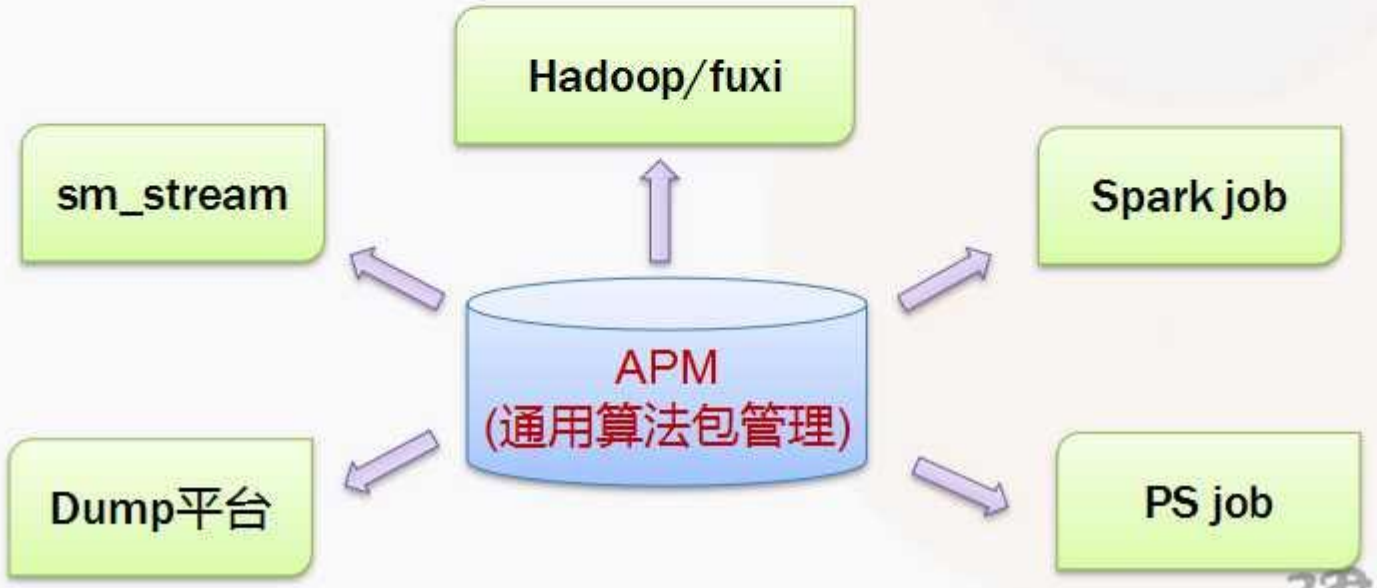
多平台挑战

多平台带来执行复杂度。

预留位置，不放PPT内容

功能标准化

- 目标
  - 功能标准化（一次开发，无限使用）
  - 跨网络环境



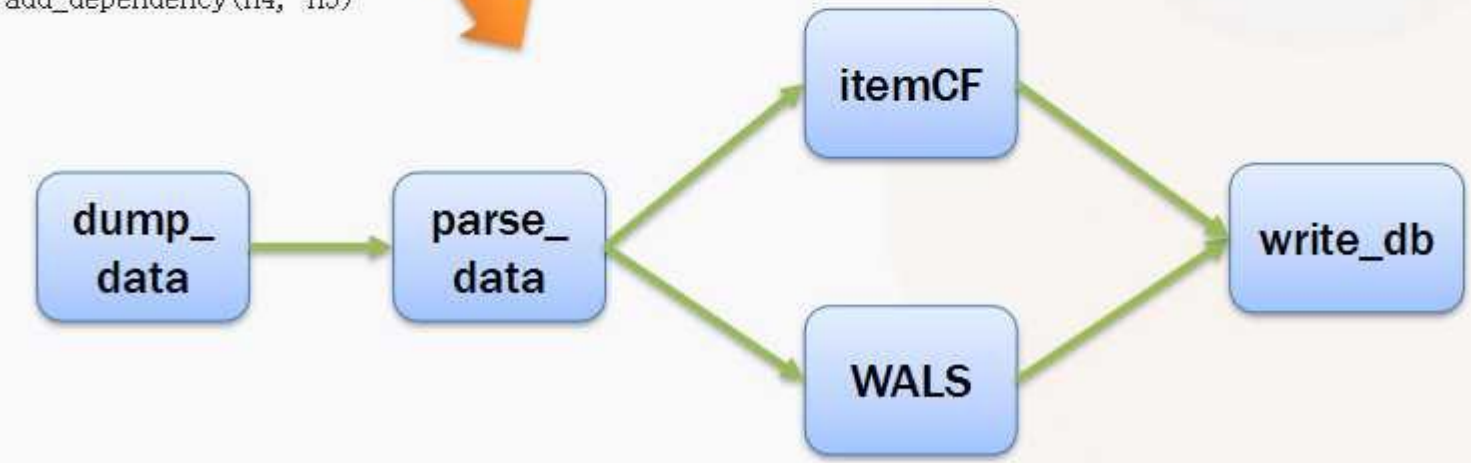
预留位置，不放PPT内容

FlowLanguage

```
flow = Flow(rec_flow, 推荐流程)
n1 = flow.add_node(dump_data)
n2 = flow.add_node(parse_data)
n3 = flow.add_node(itemCF)
n4 = flow.add_node(wals)
n5 = flow.add_node(write_db)

flow.add_dependency(n1, n2)
flow.add_dependency(n2, n3,
MyBoolCondition())
flow.add_dependency(n2, n4,
MyBoolCondition())
flow.add_dependency(n3, n5)
flow.add_dependency(n4, n5)
```

平台无关使用，通用流程语法

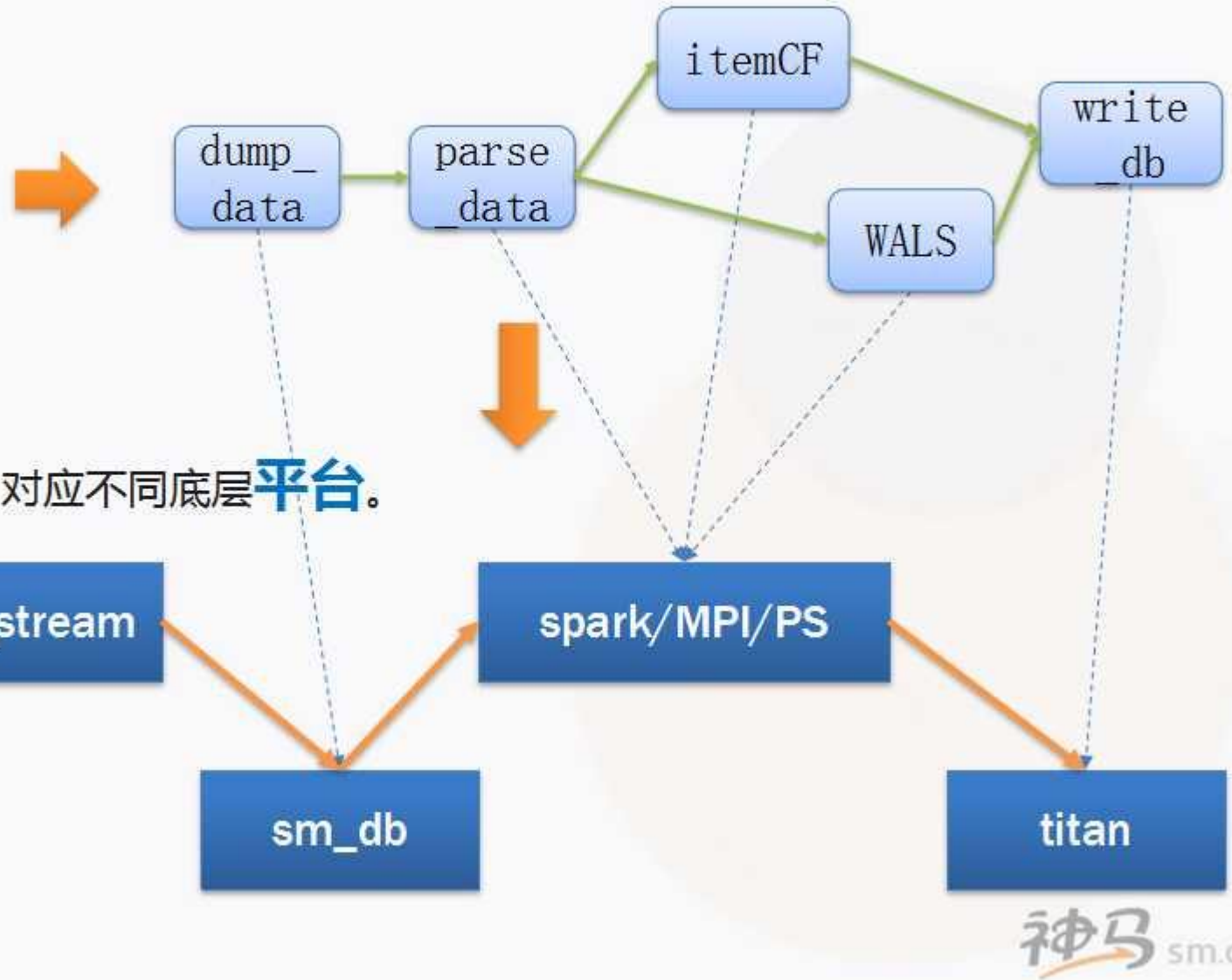


预留位置，不放PPT内容

FlowLanguage

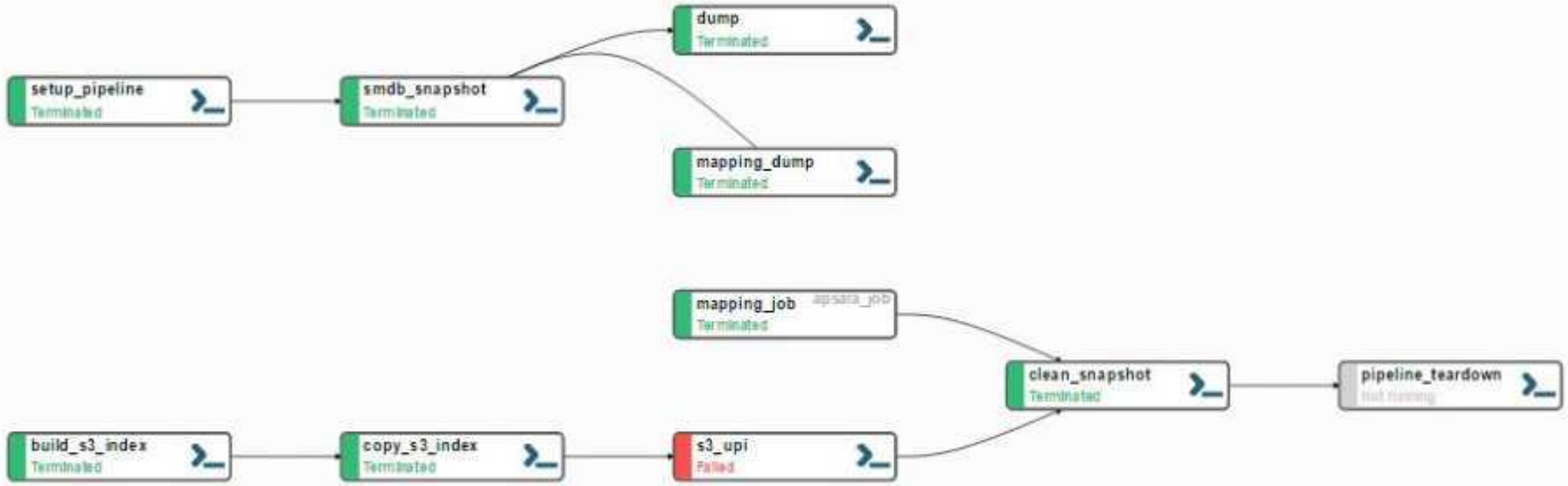
```
flow = Flow(src_flow, 推荐问题)
n1 = flow.add_node(dump_data)
n2 = flow.add_node(parse_data)
n3 = flow.add_node(itemCF)
n4 = flow.add_node(wals)
n5 = flow.add_node(write_db)

flow.add_dependency(n1, n2)
flow.add_dependency(n2, n3,
MyBoolCondition())
flow.add_dependency(n2, n4,
MyBoolCondition())
flow.add_dependency(n3, n5)
flow.add_dependency(n4, n5)
```



预留位置，不放PPT内容

工作流程



预留位置，不放PPT内容

Spark发展

- 平台开发
  - 基础功能：文件系统扩展、跨平台调度、资源和权限
  - 算法开发
- 发展

预留位置，不放PPT内容

其他

预留位置，不放PPT内容

未来思路

- 以深度学习为中心
  - 语音、图像、长尾、推荐、NLP
- 平台多样性vs统一
  - 引入新的框架，快速验证
  - 上层统一，降低使用门槛
- 时效性要求
  - 快速反馈

预留位置，不放PPT内容

一些心得

- 关注业务，更要关注人
  - 大家解决问题的方式往往能给人以启发
- 不要轻信直觉
  - 逆向思考比经验更宝贵
- 不要迷信开源
- 多平台是不得已的选择
- 服务一线员工，而非取悦boss
- 建平台容易，推广难
- 积累你的口碑

预留位置，不放PPT内容

预留位置，不放PPT内容