丁香园
WWW.DXY.CN

大数据All In Spark实践

自我简介

祝海林，丁香园大数据资深架构师

技术博客：　　http://www.jianshu.com/u/59d5607f1400

开源项目：　　https://github.com/allwefantasy

演讲目录

01

团队架构

团队组成

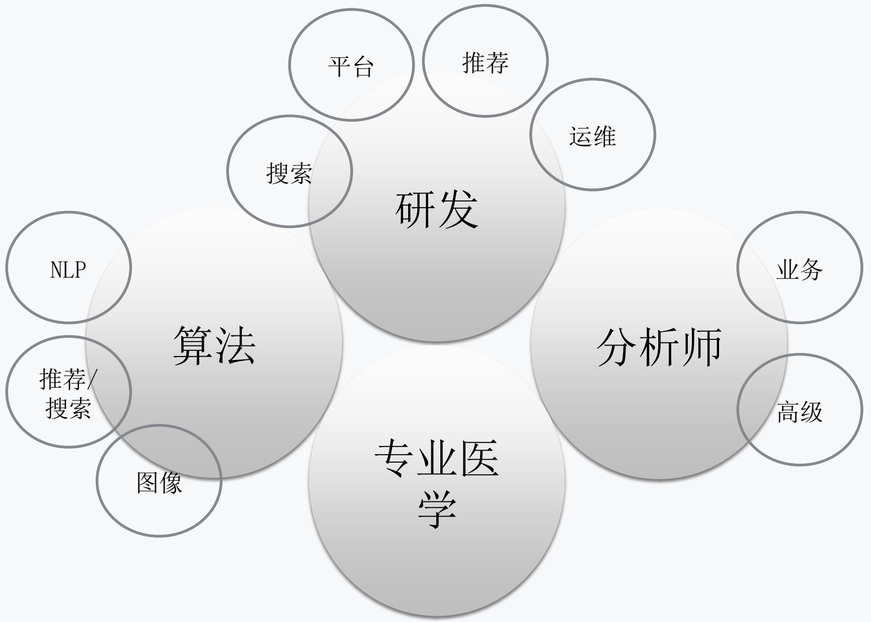
四部分：

研发　大数据基础平台架构，底层画像系统，产品化

算法　产品经理和机器智能

分析师　衔接业务和数据部门，提取，分析数据

专业医学　垂直领域专有团队



初期

四部分：

研发　　=>（3人，1年，5年后端，实习）

算法　　=>（无）

分析师　=>　6-7人

如何形成战力？

现在

四部分：

研发　　=>　9+（资深较多）
算法　　=>　5+（博士硕士）
分析师　=>　7+
知识库 => 2+（大量实习生）

如何形成战力？

如何形成战力

1. 组织架构上，让研发被支持
2. 给研发挑选一款性价比最好的武器

战斗力核心在研发

“

让研发也被支持

”

让分析师支持研发

分析师配置模式：

1. 集中式
2. 分散式
3. 混合式（采用）

混合模式：

1. 高级分析师，集中式
2. 业务分析师，分散式

各模式对比

模式	部门协作	内部协作	全局视野	人数要求
集中式	差	优	中	低
分散式	优	优	差	中
混合式	好	优	中	高

让专业团队支持研发

标签库 -> 高品质标签，便于NLP相关
知识图谱 -> 助力搜索推荐
医学支持 -> 稳步提升研发医疗素养

“

让研发选好武器

”

All In Spark

场景：

1. 批处理
2. 流式
3. 查询
4. 机器学习

为什么：

1. Spark支持Scala/Java/Python
2. Spark支持覆盖大部分场景
3. 招聘&培养成本低
4. 容易形成氛围
5. 内部有很好的工具辅助
6. 生态丰富

ETL 是最繁琐的工作，如何简化

1. SQL化
2. 配置化
3. 简化流式/批处理支持
4. 分析师和研发协作

```
{
  "name": "batch.sql",
  "params": [
    {
      "sql": "\n\nselect a.dxyid, b.action_type, count(*) as count \nfrom piUserBeenActionTable as a where a.kk=\"jack\"\ninner join i
      "outputTableName
    }
  ]
}
],
"configParams": {
}
```

```
select a.dxyid, b.action_type, count(*) as count
from piUserBeenActionTable as a where a.kk="jack"
inner join influence_type as b on a.aa = b.aa
group by a.dxyid, b.action_type
```

参看：http://github.com/allwefantasy/streamingpro 项目

StreamingPro提供了一站式解决方案

1. SQL化（万物皆SQL）
2. 配置化（一个Json文件搞定全部）
3. 服务化（Rest服务，同步异步，结果下载）
4. 跨版本（支持Spark 1.6+/Spark 2.+）
5. 管理化（StreamingPro Manager）

“

副作用：StreamingPro让团队编程开发能力变弱，影响团队发展

”

02

Spark在平台架构使用介绍

“

大数据平台都是套路

”

大数据三件套

平台（存储/计算/展现/BI）

推荐系统

精准投递系统

平台组件

OLAP平台

DA + Hive + Kylin + tableau/stat
分析师平台（此外还有Hue/Impala/Zeppelin等工具）

Skone平台

Spark SQL Server on Parquet/MySQL/ES
类似OLAP平台，为商业部门提供海量查询能力

资源中心

HBase + ES + MySQL
用户画像，内容画像,主数据，你想要的都在这

数据仓库

Hive/Spark/HDFS
数据存储中心

产品组件

推荐系统

基于资源中心，集合Rerank，
Recall等子系统

搜索系统

Based on Solr，架构类似推荐系统

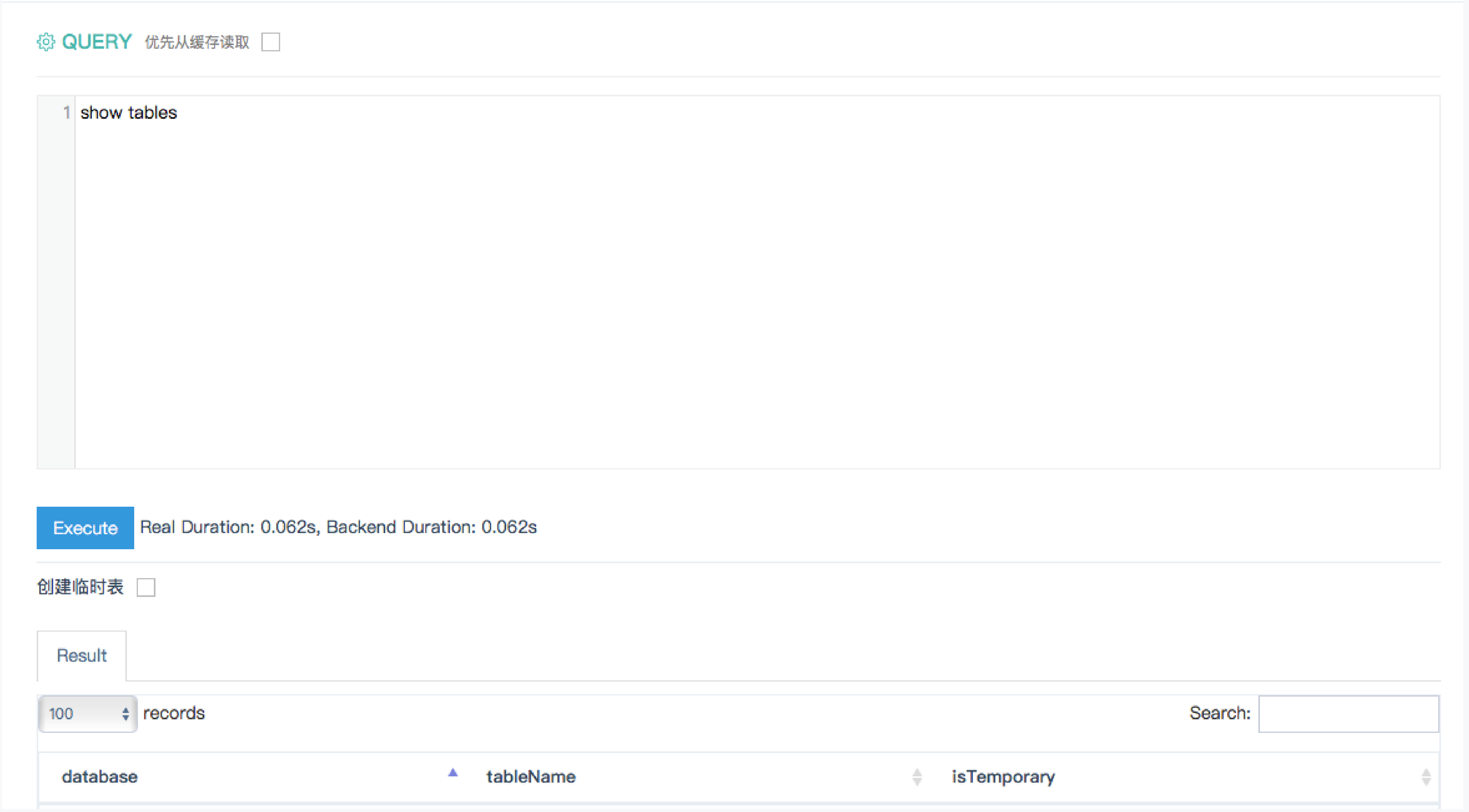
精准投递系统

内容营销，如Push/EDM/短信等

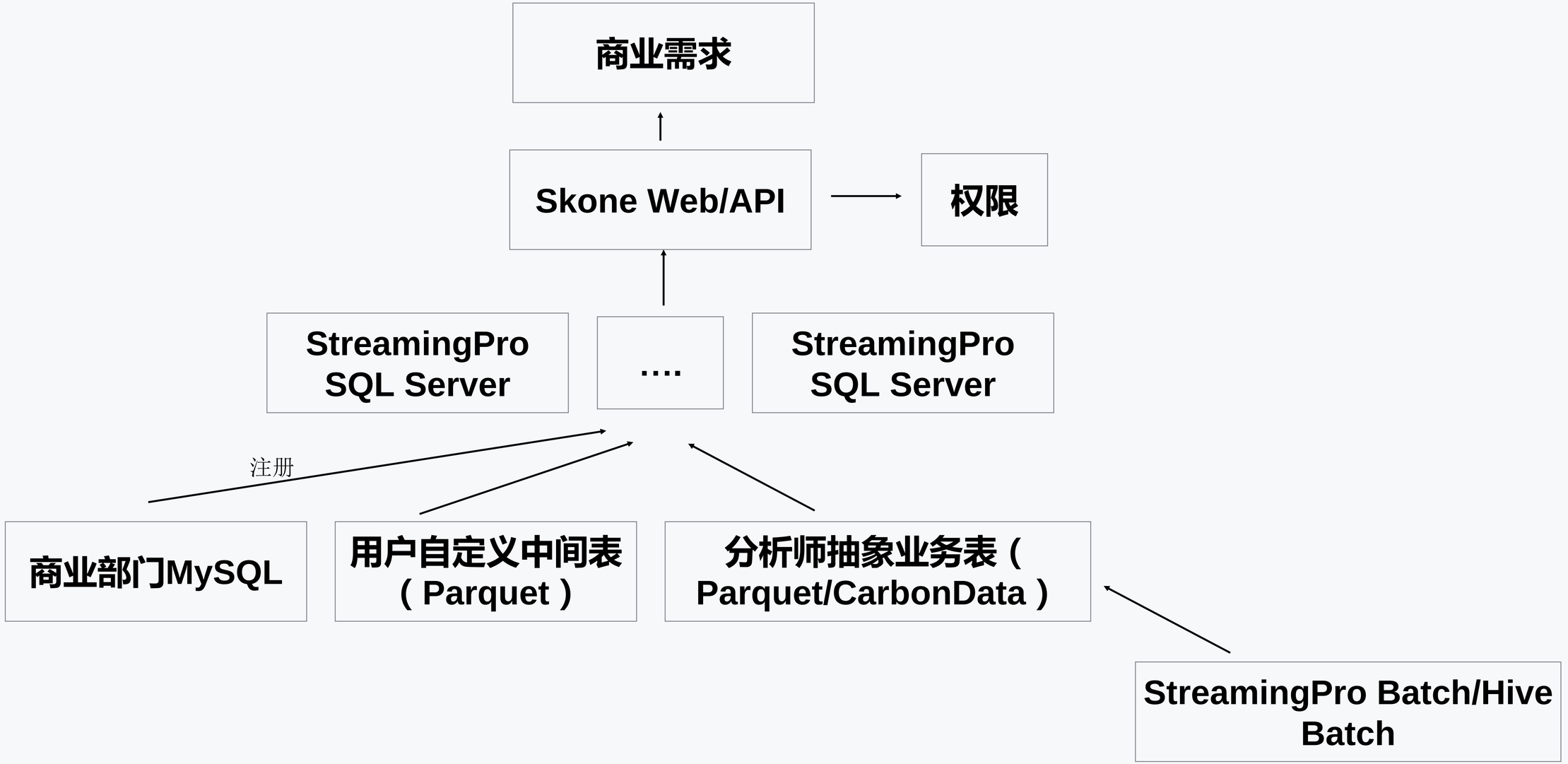
知识库

专业医学标签，知识图谱等

Skone商业支持利器



Skone商业支持利器



Skone商业支持利器（2）

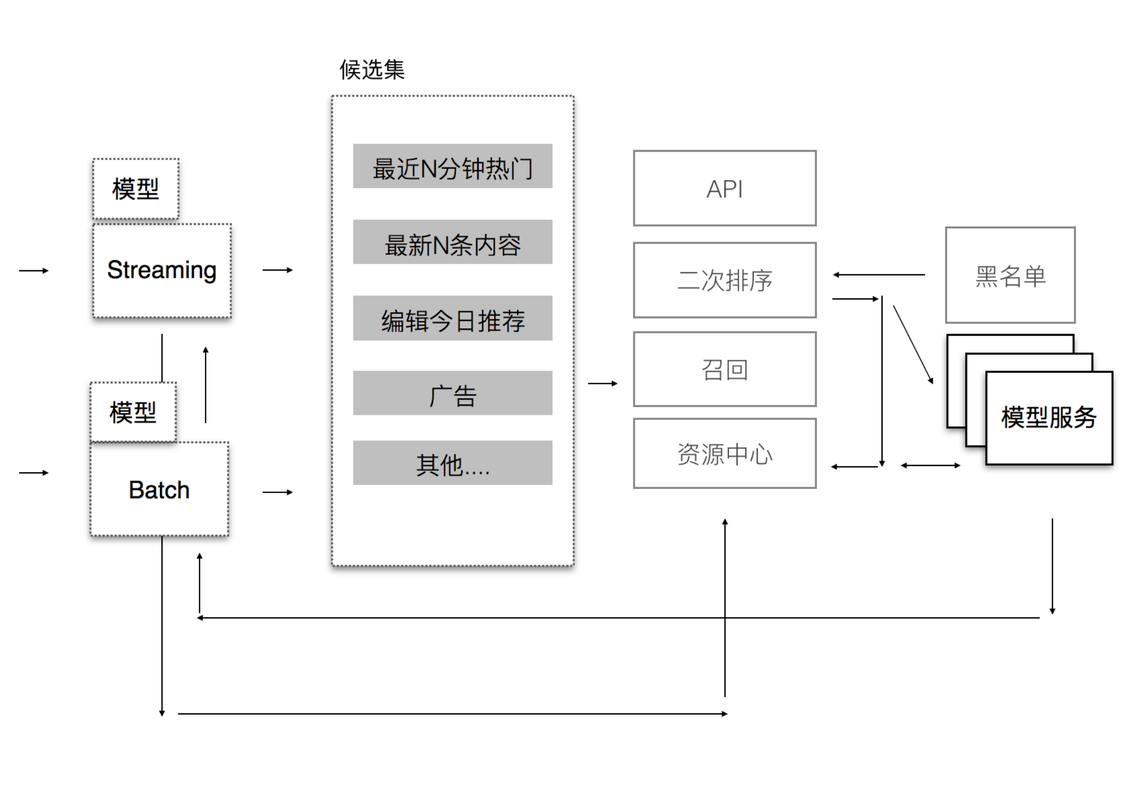
优势：

1. 基础业务和商业业务数据的关联（多数据源关联聚合）
2. 中间表支持，支持灵活数据导出　　（中间数据灵活）
3. 较好的表级权限控制
4. 计算速度较快，能满足业务需求
5. 将商务和分析师平台进行隔离

推荐系统

1. 初始化/批量更新（Batch）
2. 实时更新用户/内容画像（Streaming）
3. Spark 数量 > 10
4. Spark 开发工作量 >1/2

架构图



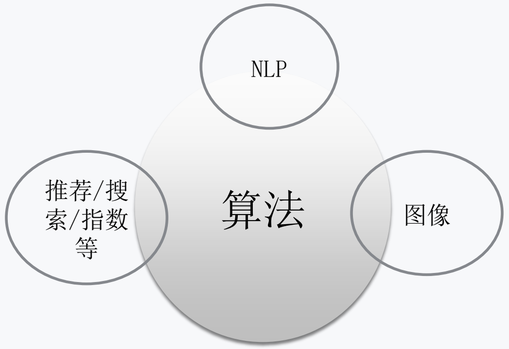
03

Spark在算法领域的使用

团队构成

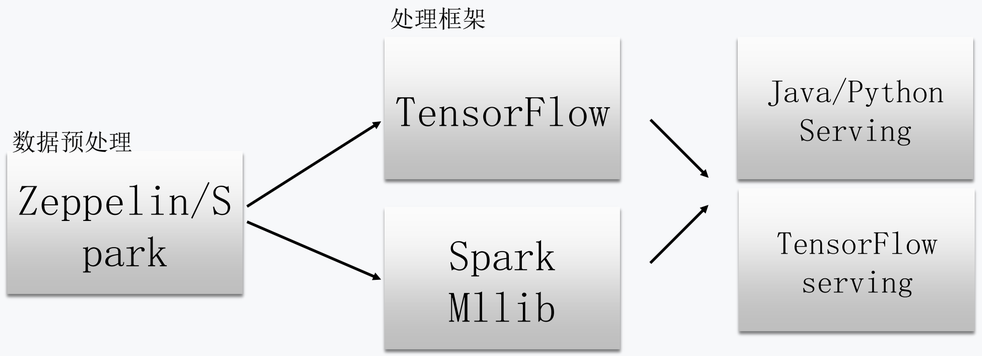
为什么切分成三个组：

1. 大部分业务都是文本相关的。
2. 推荐，搜索，商业，社区是四大客户
3. 图像是医疗领域的爆发点

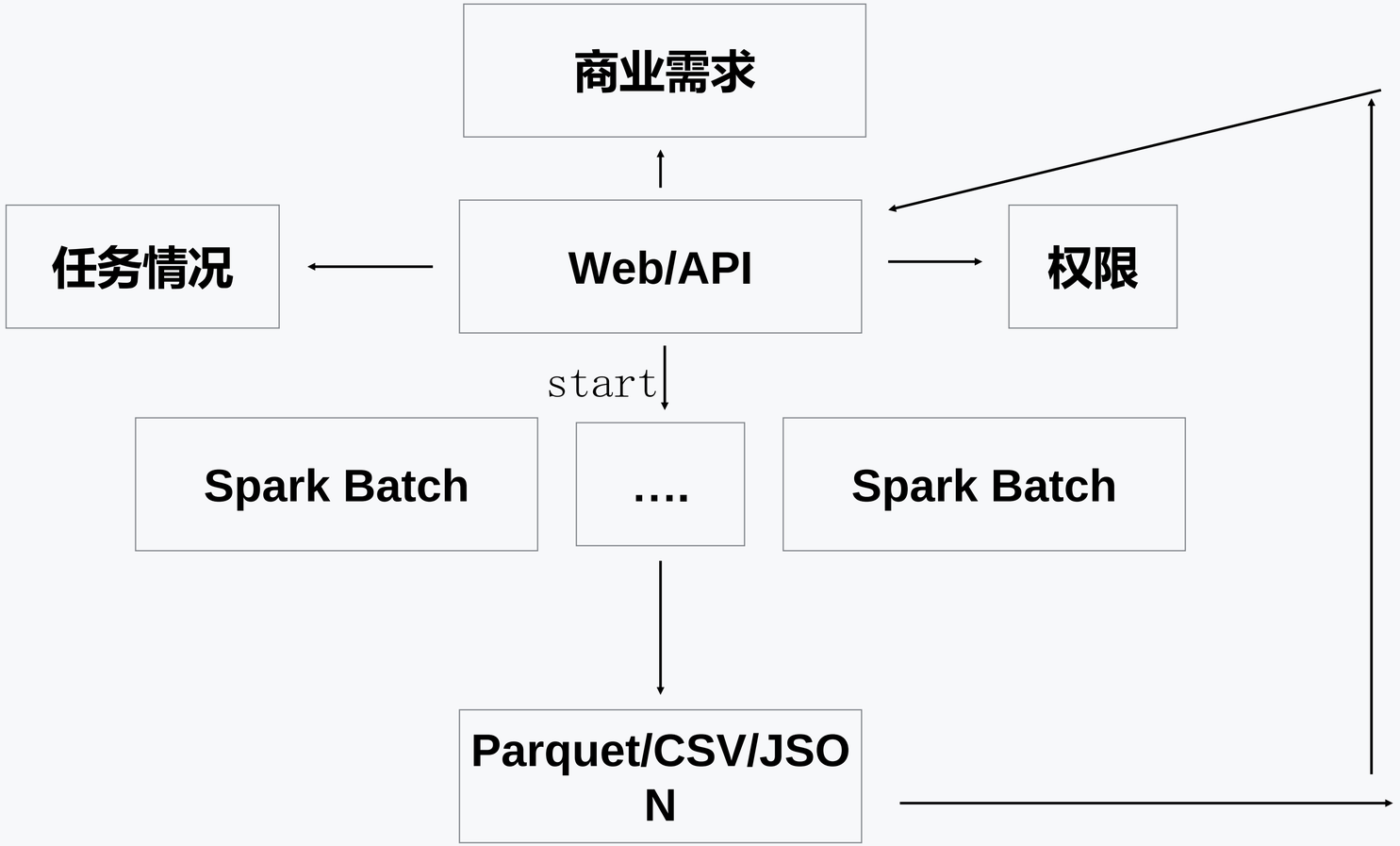


技术构成

1. 使用Zeppelin/Spark 做数据预处理
2. 深度学习框架选型为 Tensorflow
3. 非深度算法：Spark Mllib /Sklearn
4. 语系：Python/Scala



精准营销系统


商业需求
任务情况
Web/API
权限
start
Spark Batch
....
Spark Batch
Parquet/CSV/JSON

例子


运营输入信息
根据行为分析目标用户
根据目标用户走协同，扩大召回
(Spark Mllib)
Web下载(CSV)
保存用户以及分值
注册
Skone

总结

1. 在组织架构上，我们也要让研发被服务
2. 在工具选择上，尽量减少学习成本，foucus在一个综合性能优异的武器上
3. 权衡好效率以及研发的成长
4. 小团队和较大的团队，都适用


Thanks!
欢迎提问
Please make some noise