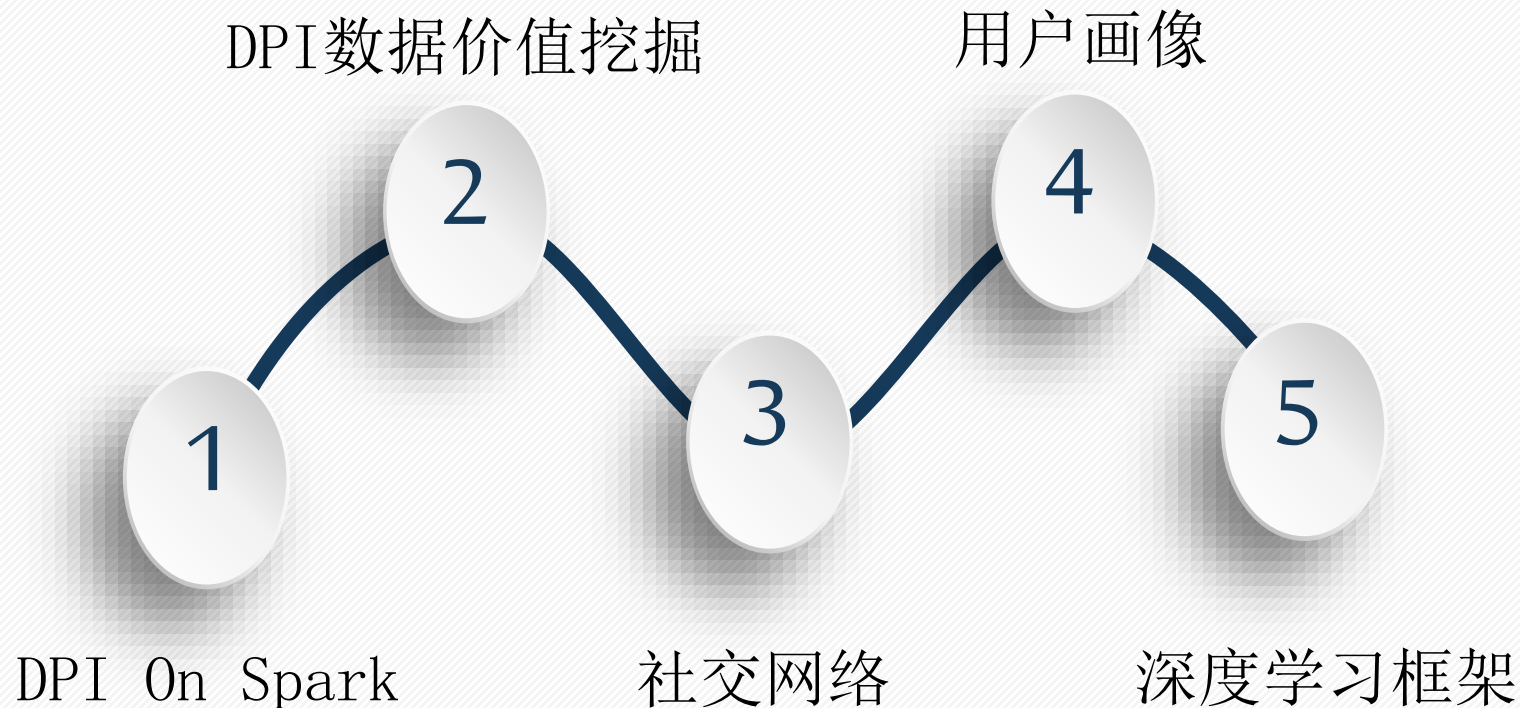




# 中国电信在大数据领域上的探索与创新

# Timeline





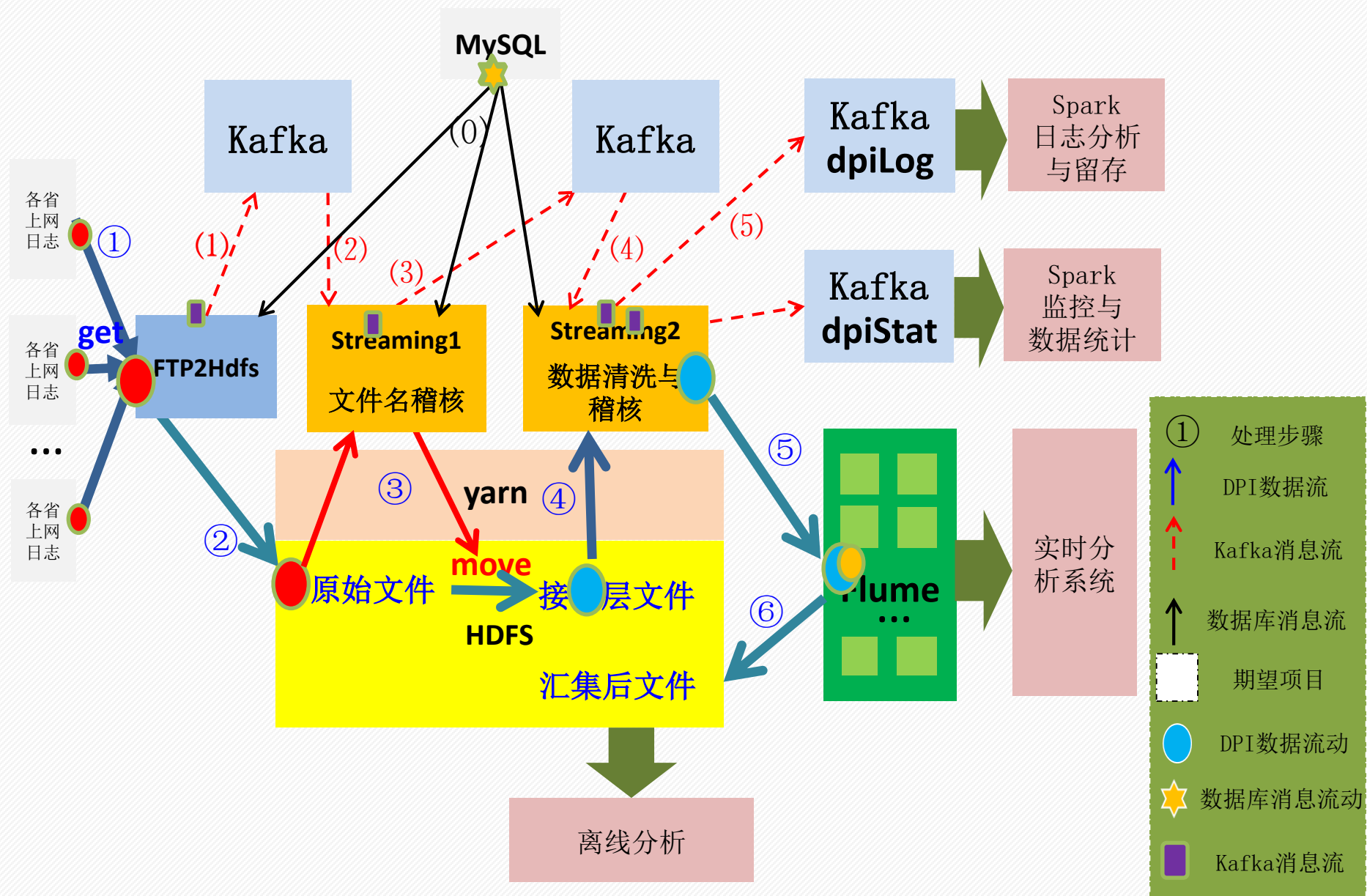
# DPI On Spark

# Apache Streaming



Streaming Model	Native		Micro-batching		Native	Native	
API	Compositional		Declarative		Compositional	Declarative	
Guarantees	At-least-once		Exactly-once		At-least-once	Exactly-once	
Fault Tolerance	Record ACKs		RDD based Checkpointing		Log-based	Checkpointing	
State Management	Not build-in		Dedicated Operators		Stateful Operators	Stateful Operators	
Latency	Very Low		Medium		Low	Low	
Throughput	Low		Medium		High	High	
Maturity	High		High		Medium	Low	

# 数据流处理





# DPI数据价值挖掘

# 处理流程

## 01 URL提取解码

中国邮速递物流  
股份有限公司上

## 02 分词及去噪

中国, 速递, 物  
流, 股份, 公司

## 03 词嵌入

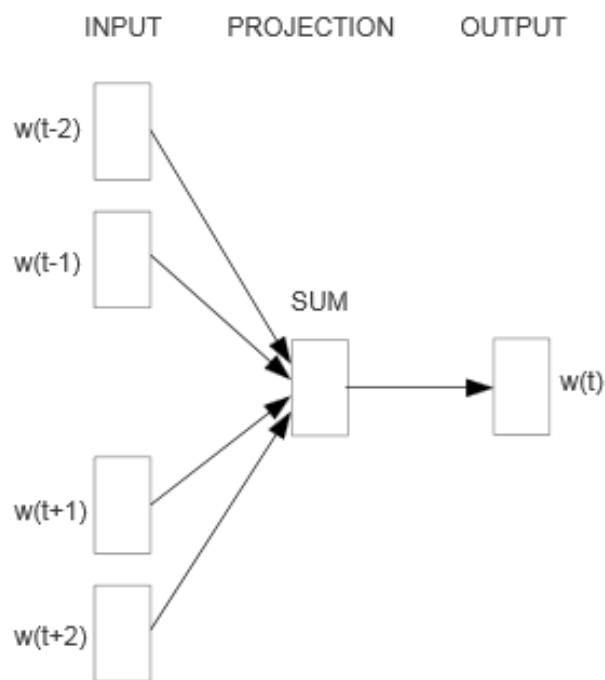
中国 -> Array(-  
0.092944495,  
0.009013829,  
0.59998745,  
0.30139154,  
0.22618507,  
0.4328063, ...

## 04 分类

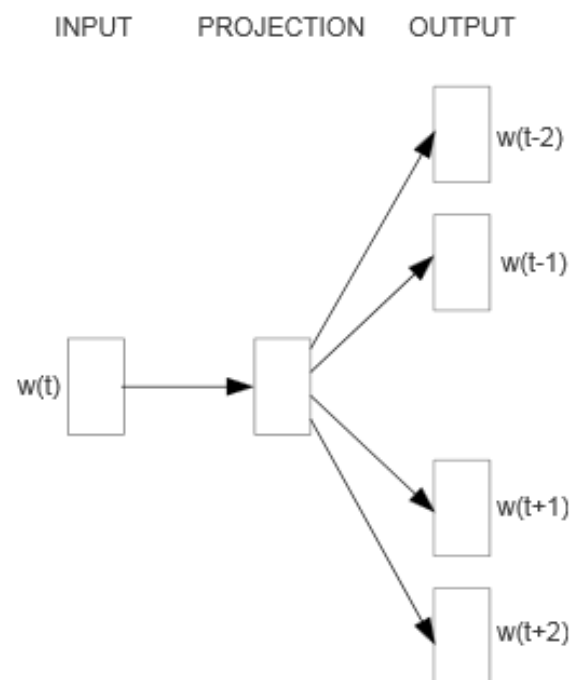
计算与核心词  
距离, 比较选  
择, 确定类别

# 核心算法(word2vector)

模型	CBOW		Skip-Gram	
方法	Hierarchical Softmax	Negative Sampling	Hierarchical Softmax	Negative Sampling



**CBOW**



**Skip-gram**



# 项目结果

任务一：

实时识别用户的搜索需求

任务二：

筛选公司的潜在价值用户

价值用户	搜索内容	商业价值
180XXXX8739	移动的投诉电话	用户发展
181XXXX8878	手机提示未发送的彩信过多	服务提升
181XXXX9025	电信全球通可以用移动联通的手机吗？	客户挽留
133XXXX6792	电信流量是全国的吗	套餐推荐



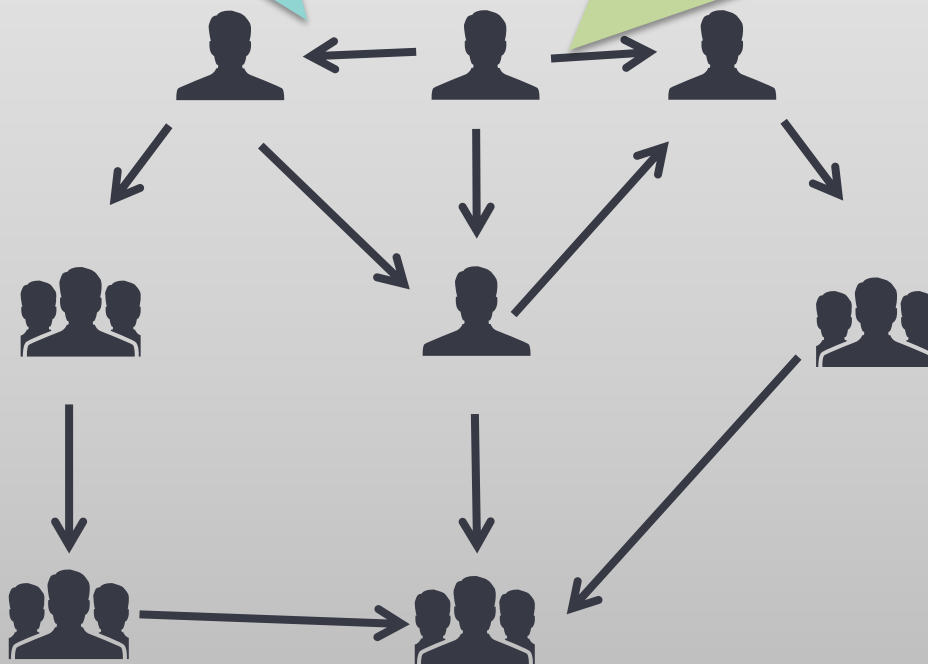
# 社交网络

# 网络结构

开始时间、通话时长、区号、基站、  
业务类型、通话费用 ...

IMSI、IMEI、号码、付费模式、  
归属地 ...

社交网络  
(通话数据)



# 功能服务



- SVDPlusPlus

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T (p_u + |N(u)|^{-\frac{1}{2}} \sum_{j \in N(u)} y_j)$$

- 共同好友

```
aggregateMessages[List[Double]]()  
triplets.map{}
```

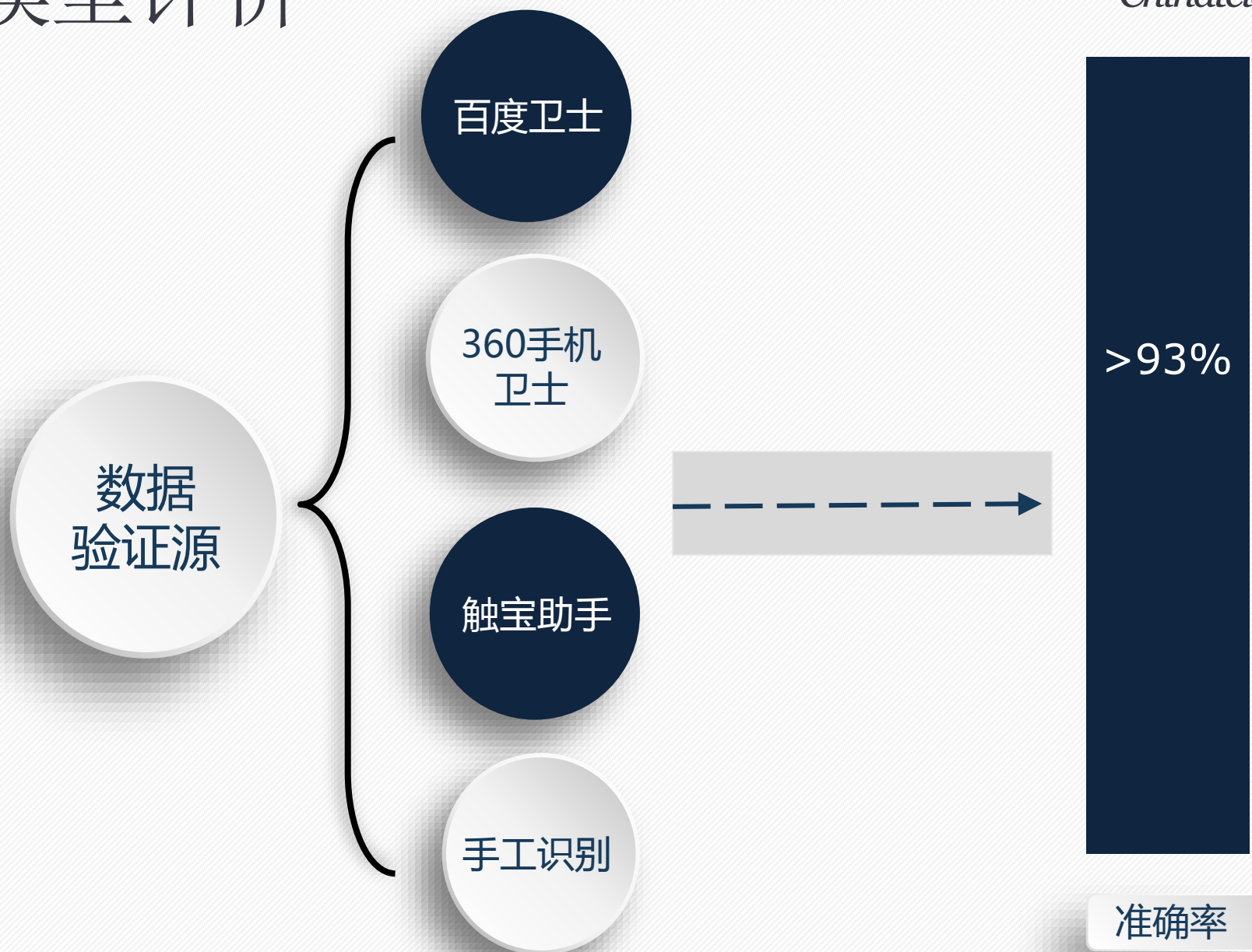
- LabelPropagation

```
Pregel(lpaGraph, initialMessage, maxIterations =  
    maxSteps)(  
    vprog = vertexProgram,  
    sendMsg = sendMessage,  
    mergeMsg = mergeMessage)
```

# 骚扰电话



# 模型评价

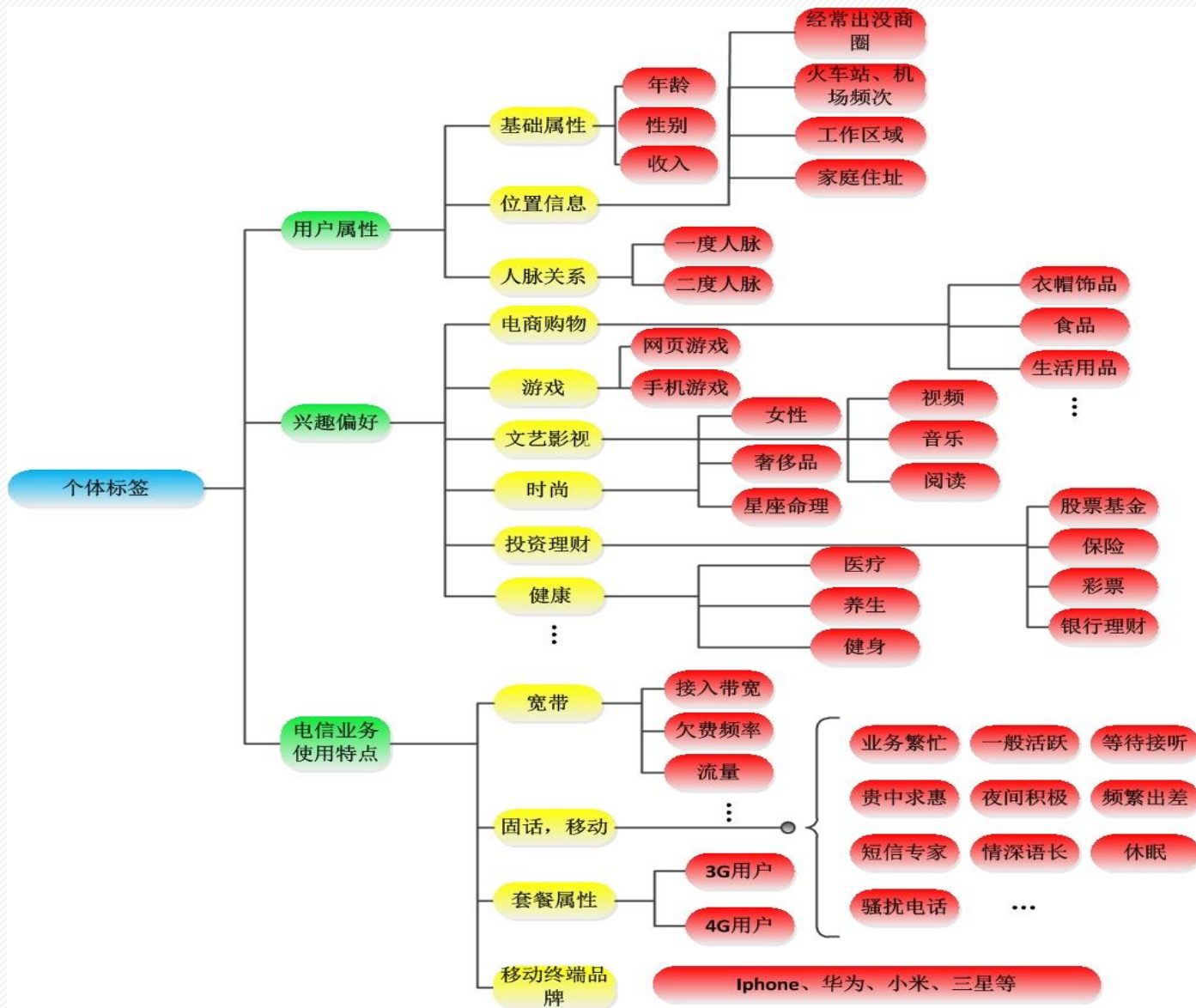




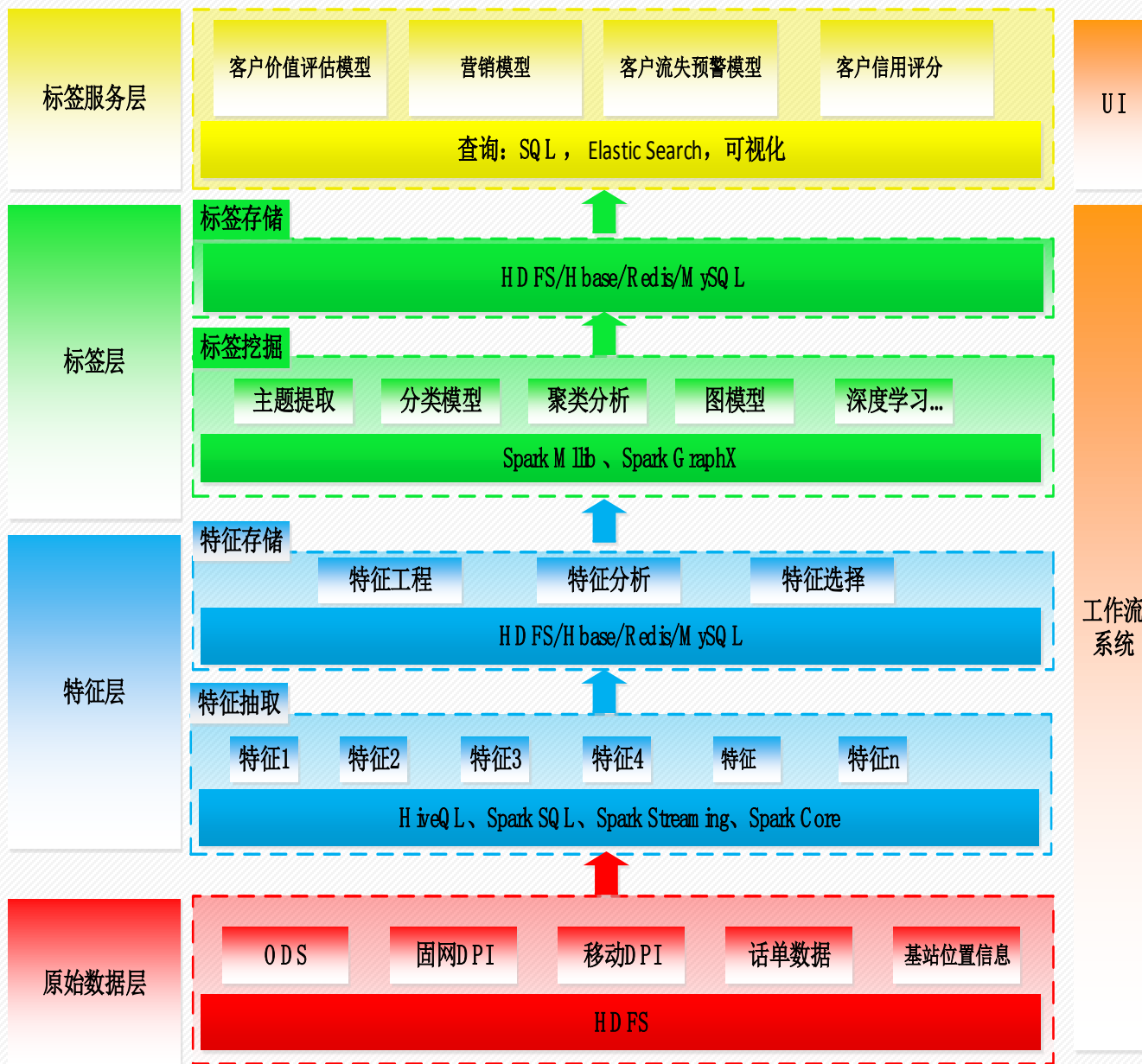
# 用户画像



# 标签体系

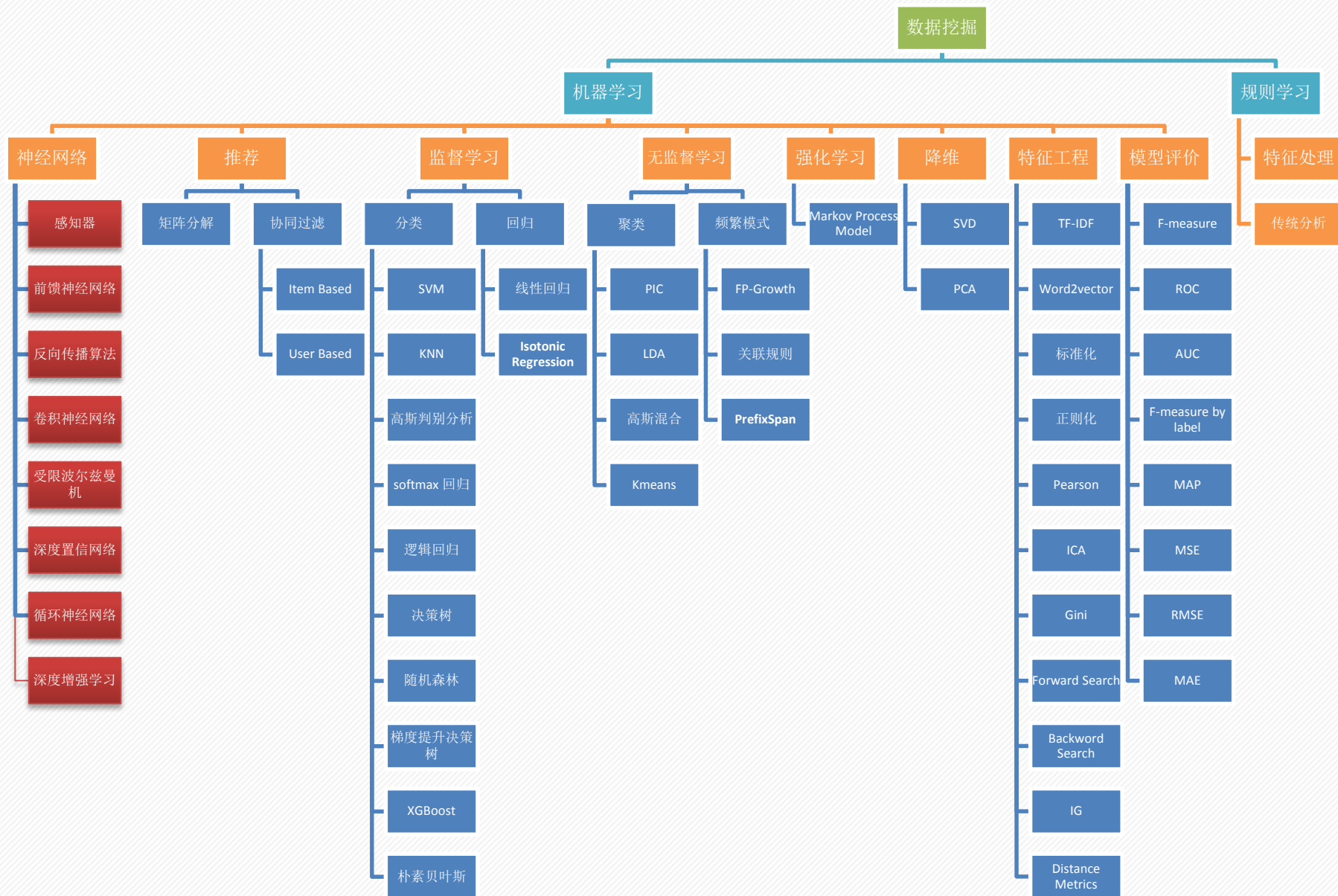


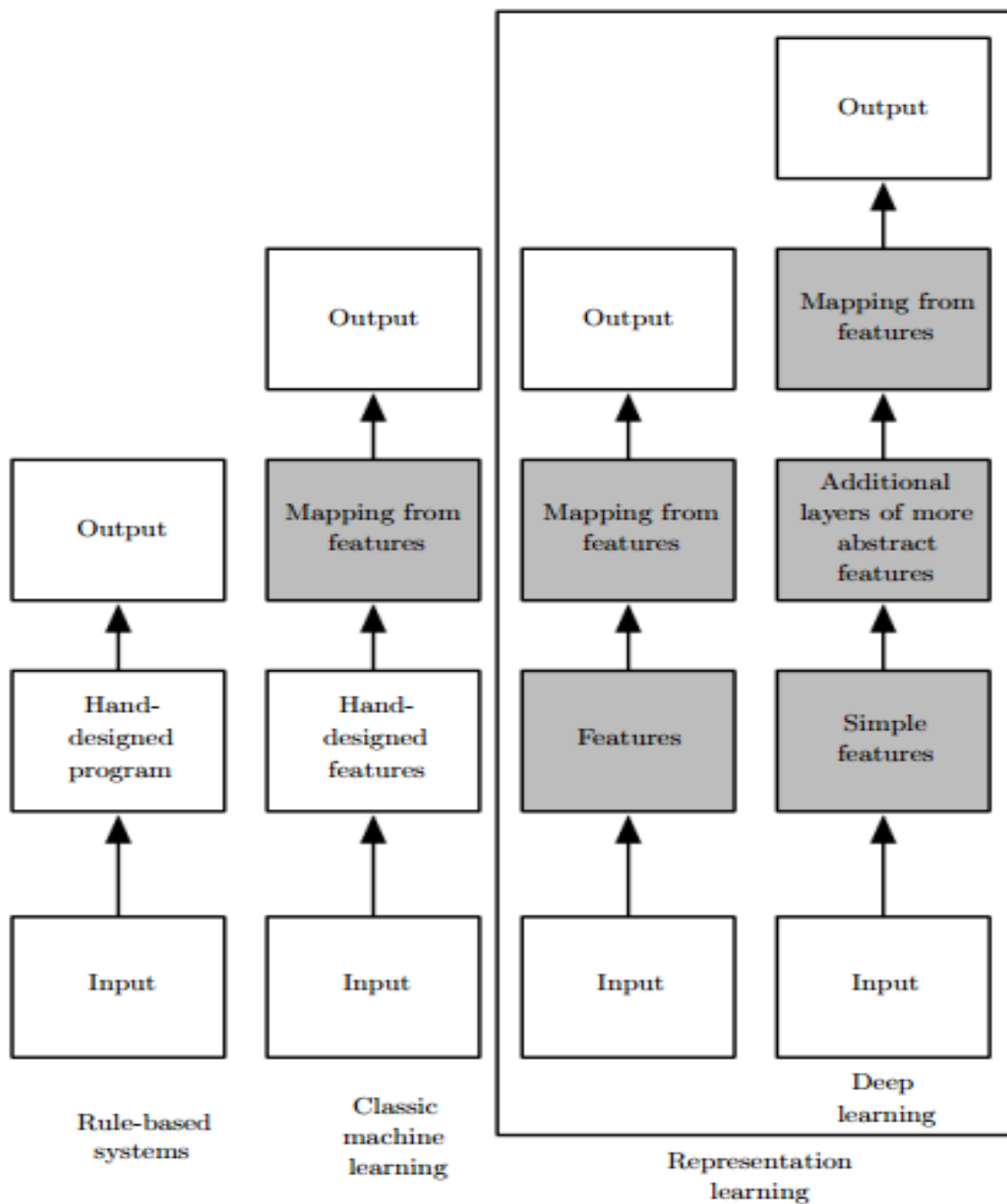
# 用户画像系统架构





# 基于Spark的深度学习框架





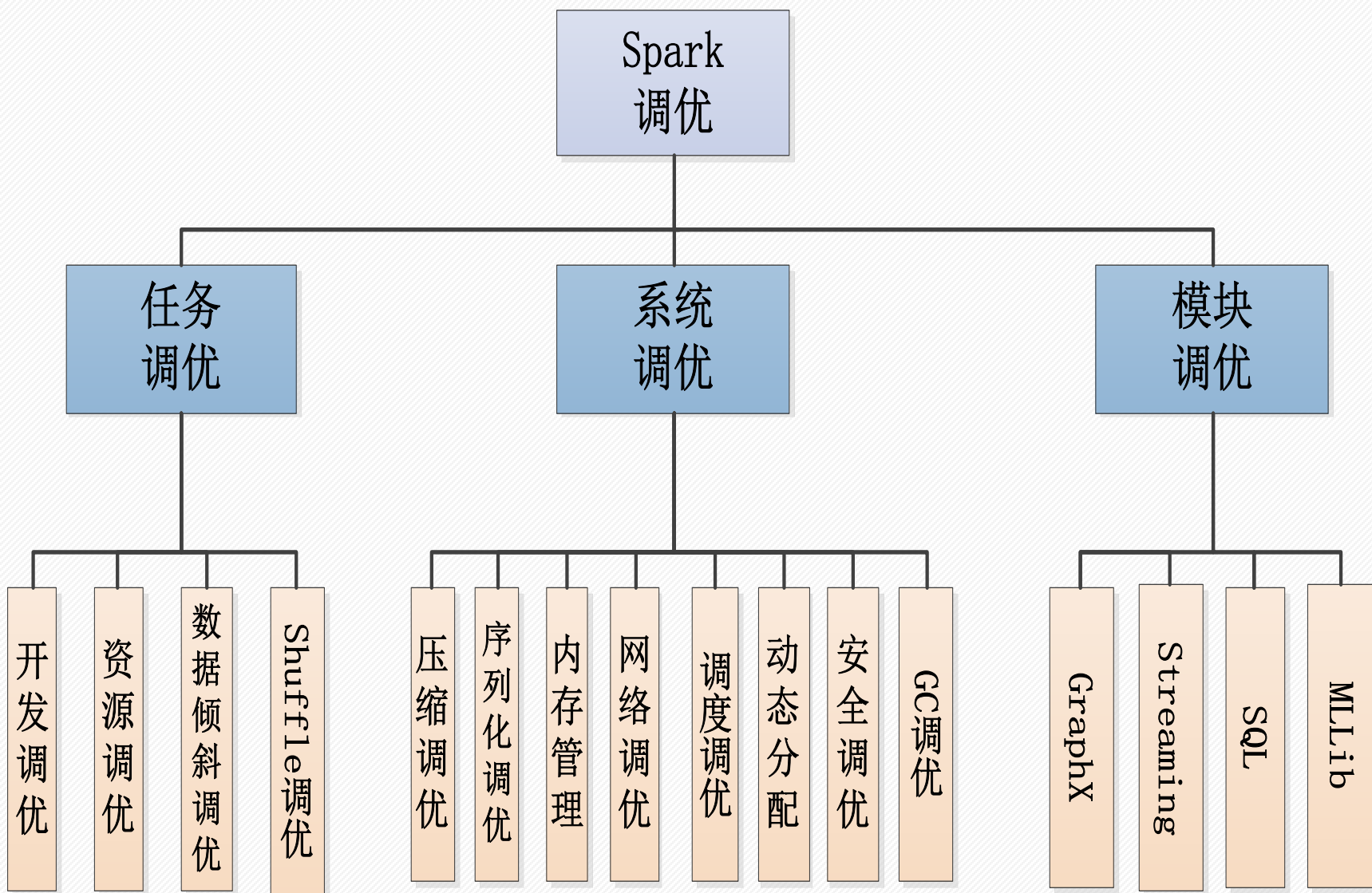
# Deep Learning On Spark

## Spark + DL: 大数据下的深度学习框架

DL框架	开源项目	开源团队
TensorFlow	TensorSpark	Arimo/Databricks
Caffe	CaffeOnSpark	Yahoo !
MXNet	MXNet	DMLC
	Spark MLlib	Baidu/Yahoo!/Berkeley/Databricks
Paddle	Spark On Paddle	Baidu



# Spark 调优





# Q&A

**We are hiring**

**EMAIL**

dingth@chinatelecom.cn



# *China Telecom*