

## 第十课 大型案例：基于Spark的推荐模型开发

- 1、案例背景
- 2、架构设计
- 3、数据清洗
- 4、模型训练
- 5、模型预测
- 6、脚本封装、部署

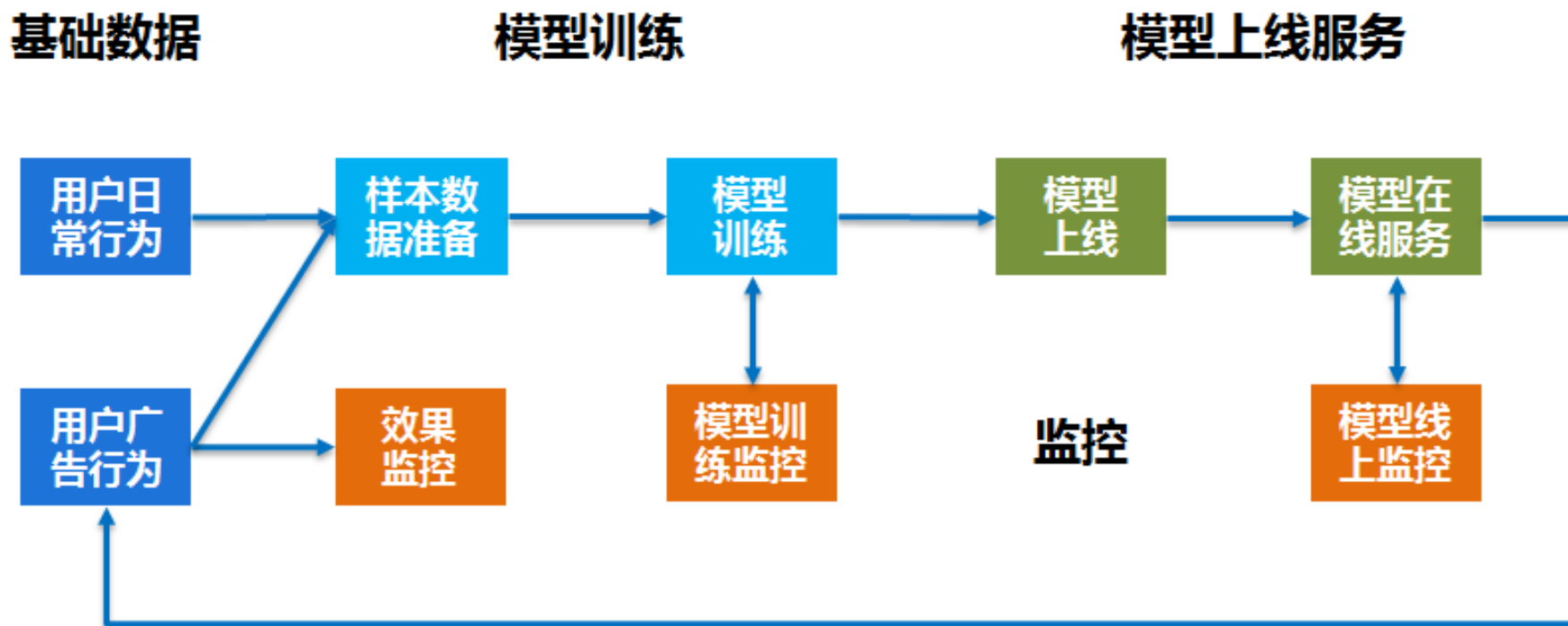
## 1、背景

# 1、背景

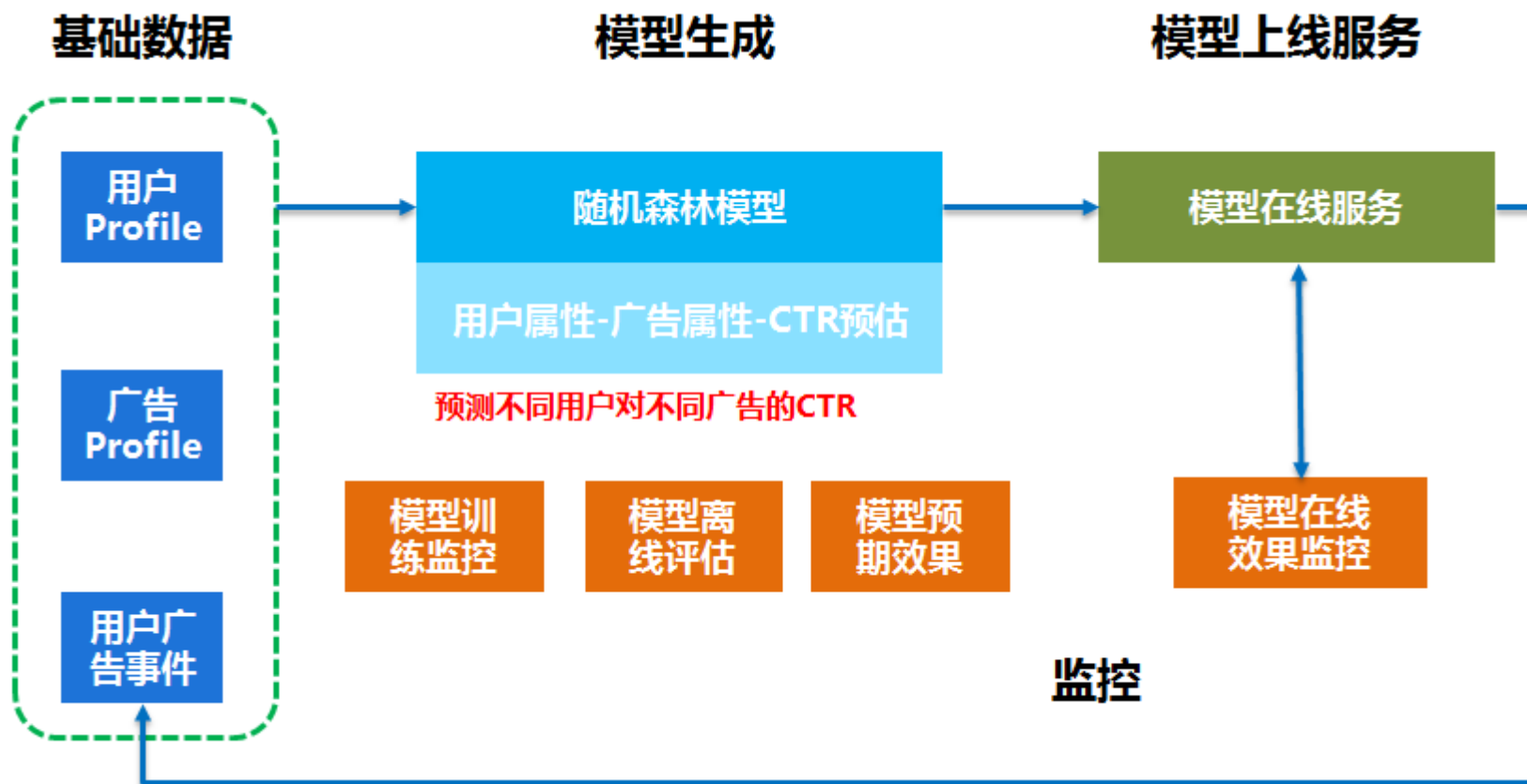


## 2、架构设计

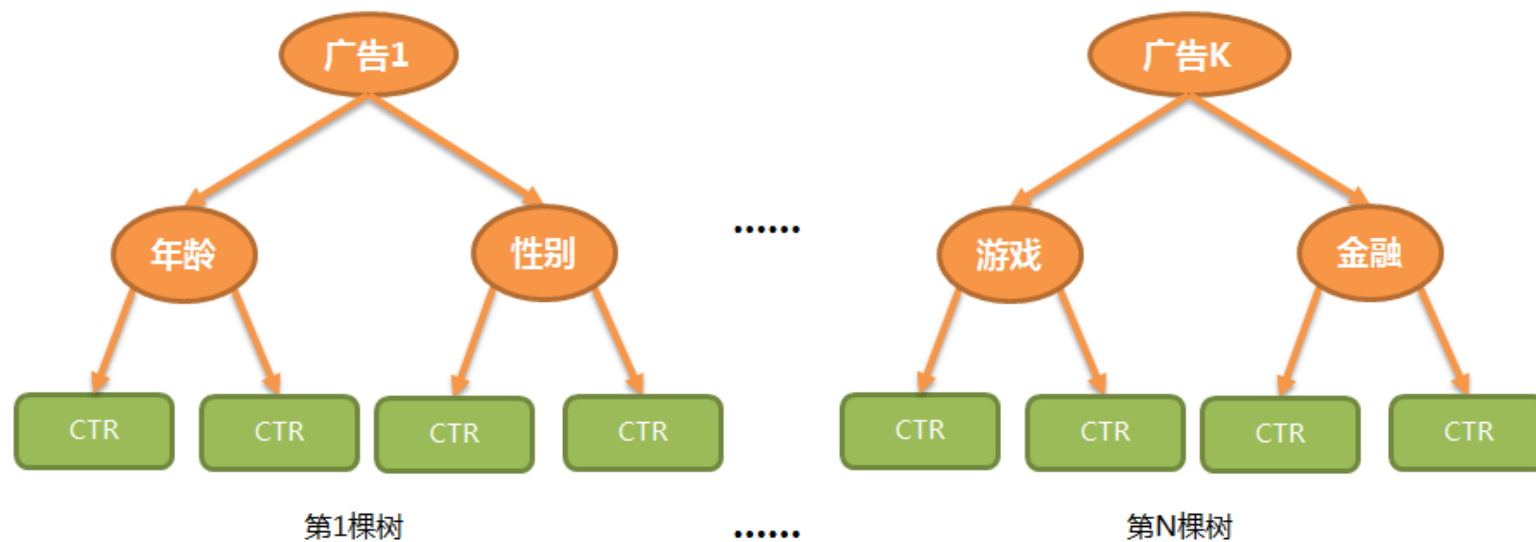
# 1、架构设计



# 1、架构设计



# 1、架构设计



| 场景Profile | 广告Profile | 用户Profile |    |       | 预测   |
|-----------|-----------|-----------|----|-------|------|
| 广告位ID     | 广告ID      | 性别        | 年龄 | ..... | CTR  |
| 热门        | APP1      | 男         | 20 | ..... | 1.5% |
| 热门        | APP2      | 男         | 20 | ..... | 1.8% |
| 热门        | APP3      | 男         | 28 | ..... | 2.0% |
| 热门        | APP4      | 男         | 28 | ..... | 1.5% |
|           |           |           |    |       |      |
|           |           |           |    |       |      |
|           |           |           |    |       |      |
|           |           |           |    |       |      |
|           |           |           |    |       |      |



## 3、数据清洗

# 3、数据清洗

## # 1、合并用户所有标签属性

**insert overwrite table t\_label\_merg partition (ds=DS\_START)**

**select** imei, label\_id, label\_value

**from** (

**select** imei, label\_id, score as label\_value

**from** t\_featruce\_source3

**where** ds = DS\_START

**union all**

**select** imei, label\_id, score as label\_value

**from** t\_featruce\_source2

**where** ds = DS\_START

**union all**

**select** imei, label\_id, score as label\_value

**from** t\_featruce\_source1


**where** ds = DS\_START

) t1

| 标签类别 | 标签ID | 标签名称  | 标签值  |
|------|------|-------|------|
| 101  | 801  | CPU数量 | 0-10 |
| 101  | 802  | CPU频率 | 0-24 |
| 101  | 803  | 分辨率   | 0-11 |
| 101  | 804  | 安卓版本  | 0-30 |
| 101  | 805  | RAM   | 0-11 |
| 101  | 806  | ROM   | 0-18 |
| 101  | 807  | 价格    | 0-8  |
| 101  | 808  | 机型    | 0-51 |

| 标签类别 | 标签ID | 标签名称 | 标签值 |
|------|------|------|-----|
| 102  | 901  | 性别   | 0-2 |
| 102  | 902  | 年龄   | 0-5 |

... ..



| 用户ID     | 标签ID  | 标签值   |
|----------|-------|-------|
| 10000001 | 801   | 2     |
| 10000001 | 802   | 5     |
| 10000001 | 803   | 3     |
| 10000001 | 804   | 1     |
| 10000001 | 901   | 1     |
| 10000001 | 902   | 3     |
| 10000002 | ..... | ..... |
| ... ..   |       |       |

# 3、数据清洗

## # 2、按照libSVM的数据格式合并用户数据，生成训练样本

insert overwrite table t\_train\_sample partition (ds= DS\_START)

select g1.imei, g1.ad\_id, if(g1.click >= 1, 1, 0) click, g2.label

from (

select \*

from t\_ad\_click

where ds = DS\_START

) g1

join (

select imei, concat\_ws(' ', collect\_set(concat(cast(label\_id as string), ':', label\_value))) as label

from t\_label\_merg

where ds = DS\_START

group by imei

) g2

on (g1.imei = g2.imei)

| 用户ID     | 标签ID  | 标签值   |
|----------|-------|-------|
| 10000001 | 801   | 2     |
| 10000001 | 802   | 5     |
| 10000001 | 803   | 3     |
| 10000001 | 804   | 1     |
| 10000001 | 901   | 1     |
| 10000001 | 902   | 3     |
| 10000002 | ..... | ..... |



| 用户ID     | 广告ID  | 点击 | 标签                                  |
|----------|-------|----|-------------------------------------|
| 10000001 | 101   | 0  | 801:2 802:5 803:3 804:1 901:1 902:3 |
| 10000001 | 102   | 1  | 801:2 802:5 803:3 804:1 901:1 902:3 |
| 10000001 | 103   | 0  | 801:2 802:5 803:3 804:1 901:1 902:3 |
| 10000002 | ..... |    |                                     |

... ..

| 用户ID     | 广告ID  | 点击 |
|----------|-------|----|
| 10000001 | 101   | 0  |
| 10000001 | 102   | 1  |
| 10000001 | 103   | 0  |
| 10000002 | ..... |    |

... ..

| sample_lda_libsvm_data.txt |    |     |     |     |     |     |     |     |     |     |      |      |
|----------------------------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| 1                          | 0  | 1:1 | 2:2 | 3:6 | 4:0 | 5:2 | 6:3 | 7:1 | 8:1 | 9:0 | 10:0 | 11:3 |
| 2                          | 1  | 1:1 | 2:3 | 3:0 | 4:1 | 5:3 | 6:0 | 7:0 | 8:2 | 9:0 | 10:0 | 11:1 |
| 3                          | 2  | 1:1 | 2:4 | 3:1 | 4:0 | 5:0 | 6:4 | 7:9 | 8:0 | 9:1 | 10:2 | 11:0 |
| 4                          | 3  | 1:2 | 2:1 | 3:0 | 4:3 | 5:0 | 6:0 | 7:5 | 8:0 | 9:2 | 10:3 | 11:9 |
| 5                          | 4  | 1:3 | 2:1 | 3:1 | 4:9 | 5:3 | 6:0 | 7:2 | 8:0 | 9:0 | 10:1 | 11:3 |
| 6                          | 5  | 1:4 | 2:2 | 3:0 | 4:3 | 5:4 | 6:5 | 7:1 | 8:1 | 9:1 | 10:4 | 11:0 |
| 7                          | 6  | 1:2 | 2:1 | 3:0 | 4:3 | 5:0 | 6:0 | 7:5 | 8:0 | 9:2 | 10:2 | 11:9 |
| 8                          | 7  | 1:1 | 2:1 | 3:1 | 4:9 | 5:2 | 6:1 | 7:2 | 8:0 | 9:0 | 10:1 | 11:3 |
| 9                          | 8  | 1:4 | 2:4 | 3:0 | 4:3 | 5:4 | 6:2 | 7:1 | 8:3 | 9:0 | 10:0 | 11:0 |
| 10                         | 9  | 1:2 | 2:8 | 3:2 | 4:0 | 5:3 | 6:0 | 7:2 | 8:0 | 9:2 | 10:7 | 11:2 |
| 11                         | 10 | 1:1 | 2:1 | 3:1 | 4:9 | 5:0 | 6:2 | 7:2 | 8:0 | 9:0 | 10:3 | 11:3 |
| 12                         | 11 | 1:4 | 2:1 | 3:0 | 4:0 | 5:4 | 6:5 | 7:1 | 8:3 | 9:0 | 10:1 | 11:0 |
| 13                         |    |     |     |     |     |     |     |     |     |     |      |      |

## 4、模型训练

## 4、模型训练

---

**详细见代码**

## 5、模型预测

## 5、模型预测

---

**详细见代码**

## 6、脚本封装



## 6、脚本封装

---

详细见代码

# Thanks

**FAQ时间**