# 头条推荐业务架构介绍

## 业务

在头条APP海量⽤户与海量⽂章之上，使⽤lambda⼤数据实时和离线计算整体架构，利⽤头条⽤户

在APP上的点击⾏为、浏览⾏为、收藏⾏为等建⽴⽤户与⽂章之间的画像关系，通过机器学习推荐算法

## 架构与业务流

1、用户的⾏为收集，业务数据收集

2、批量计算(离线计算)：⽤用户⽂文章画像

3、⽤用户的召回结果、排序精选过程

4、grpc的实时推荐业务流的搭建缓存

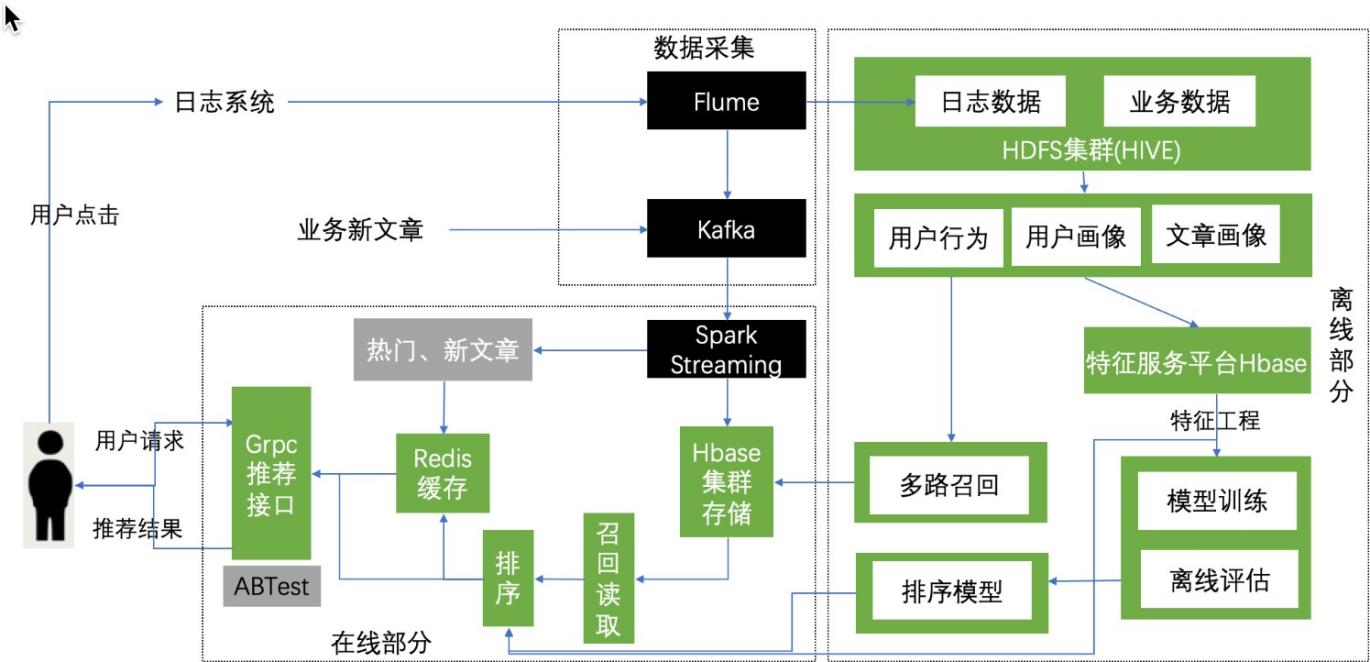
# 1.3 开发环境介绍

## 1.3.1 虚拟机设备

**1.3.3 python环境**

分布式环境：Hbase会遇到三台时间同步的问题

python环境：三台也都必须安装



# 2.2 数据库迁移

## 2.2.1 数据库迁移需求

业务数据：133，134，135， 136

web

推荐系统：137，138，139

导⼊过来，增量更新

hadoop的hive数据仓库

同步⼀份数据在集群中⽅便进⾏数据分析操作

⽤户资料信息呢两张表：user\_profifile，user\_basic

⽂章内容基本信息、频道三张表：news\_article\_basic，news\_article\_content，news\_channel

启动数据库步骤

systemctl start docker

docker start mysql

docker exec -it mysql bash

### Sqoop 迁移业务数据导⼊入问题

新增的⽤用户、⽂文章

修改的⽤用户信息、⽂文章信息

两种导⼊形式，我们选择增量，定期导⼊新数据

sqoop全量导⼊

不需要创建HIVE表

sqoop增量导⼊入

append

#### incremental

直接sqoop导⼊入到hive(–incremental lastmodiﬁed模式不不⽀支持导⼊入Hive )

#### sqoop导⼊入到hdfs，然后建⽴立hive表关联

* + - 1. **Sqoop 迁移案例例**

sqoop 导出的 hdfs 分⽚片数据，都是使⽤用逗号 , 分割于 hive 默认的分隔符是 /u0001 （Ctrl+A）

Mysql导⼊入对应hive类型:

MySQL(bigint) --> Hive(bigint) MySQL(tinyint) --> Hive(boolean) MySQL(int) --> Hive(int) MySQL(double) --> Hive(double) MySQL(bit) --> Hive(boolean) MySQL(varchar) --> Hive(string) MySQL(decimal) --> Hive(double) MySQL(date/timestamp) --> Hive(string)

注意：1、连接JDBC的IP 地址 或者主机名是否错误 2、确认mysql数据库打开并且能够sqoop测试成功

3、⽂文章表导⼊入news\_article\_basic，news\_article\_content、news\_channel

sqoop迁移⼀一些有特殊字符的数据导⼊入hadoop

加⼊入—query参数，从数据库中选中对应字段，过滤相应内容，使⽤用REPLACE、CHAR(或者

CHR)进行替换字符

HIVE读取存在问题

HIVE默认分割符，**"\t", "\n"**这些特殊字符出不不了了

--query

select article\_id, user\_id, channel\_id, REPLACE(REPLACE(REPLACE(title, CHAR(13), ''), CHAR(10), ''), ',', ' ') title, status, update\_time from news\_article\_basic limit 1;

+------------+---------+------------+-----------------------------+---

-----+---------------------+

| article\_id | user\_id | channel\_id | title | status | update\_time |

+------------+---------+------------+-----------------------------+---

-----+---------------------+

| 1 | 1 | 17 | Vue props⽤用法⼩小结原荐 | 2 | 2019-02-18 11:08:16 |

+------------+---------+------------+-----------------------------+---

-----+---------------------+

并且mysql表中存在tinyibt,必须在connet中加⼊入: ?tinyInt1isBit=false

防⽌止默认到HIVE中,字段默认会被转化为boolean数据类型

4、news\_channel与⽤用户两张表⼀一起导⼊入

5、news\_article\_content

全量量导⼊入(表只是看结构，不不需要在HIVE中创建，因为是直接导⼊入HIVE，会⾃自动创建

news\_article\_content)

## crontab-shell脚本定时运⾏行行

创建⼀一个定时运⾏行行的脚本

crontab -e

写⼊入定时执⾏行行的命令：

\*/30 \* \* \* \* /root/toutiao\_project/scripts/import\_incremental.sh

启动服务

service crond start/stop

## 2.2.3 总结

sqoop导⼊入业务数据到hadoop操作append, lastmodiﬁeld

增量量导⼊入形式

# ⽤户⾏为收集到HIVE

## 为什么要收集⽤用户点击⾏为⽇志

便于了解分析⽤用户的行为、喜好变化为⽤户建立画像提供依据

## ⽤户⽇志如何收集

### 埋点开发测试流程

埋点参数:

#### 就是在应⽤用中特定的流程收集⼀一些信息，⽤用来跟踪应⽤用使⽤用的状况，后续⽤用来进⼀一步优化产 品或是提供运营的数据⽀支撑

1、PM（项⽬目经理理）、算法推荐⼯工程师⼀一起指定埋点需求⽂文档

2、后端、客户端 APP集成

3、推荐⼈人员基于⽂文档埋点测试与梳理理

### 头条⽂章推荐埋点需求整理理

埋点事件号:

停留留时间

read

点击事件

click

曝光事件(相当于刷新⼀一次请求推荐新⽂文章) exposure

收藏事件

collect

分享事件

share

埋点参数⽂文件结构

曝光的参数：下拉刷新，推荐新的若⼲干篇⽂文章 我们将埋点参数设计成⼀一个固定格式的json字符串串

## 离线部分-⽤用户⽇日志收集

### 通过ﬂume将业务数据服务器器A的⽇日志收集到hadoop服务器器hdfs的hive中

根据课件设置 sources，channels，sinks

开启收集命令

/root/bigdata/flume/bin/flume-ng agent -c /root/bigdata/flume/conf -f /root/bigdata/flume/conf/collect\_click.conf -Dflume.root.logger=INFO,console -name a1

注意收集后，自动生成的时间是当前的最新时间

然后执行

alter table user\_action add partition (dt='2018-12-11') location "/user/hive/warehouse/profile.db/user\_action/2018-12-11/"

来进行关联

## Supervisor进程管理理

很⽅方便便的监听、启动、停⽌止、重启⼀一个或**多个**进程 使⽤用

1、配置 supervisor开启配置⽂文件在哪⾥里里

/etc/supervisor/

2、配置.conf ，reco.conf

3、写⼊入配置格式

4、开启supervisor, 启动 supervisord -c /etc/supervisord.conf 5、supervisorctl管理理进程

## supervisor 启动监听ﬂume收集⽇日志程序

1、添加配置

[program:collect-click]

command=/bin/bash /root/toutiao\_project/scripts/collect\_click.sh user=root

autorestart=true redirect\_stderr=true

stdout\_logfile=/root/logs/collect.log loglevel=info

stopsignal=KILL stopasgroup=true killasgroup=true

1. update更更新我们要监控管理理的程序

在执行hadoop dfs -put ./hadoopbak/profile.db/user\_action/ /user/hive/warehouse/profile.db/

注意需要先create database profile;创建数据库

数据上传后，可以通过http://192.168.19.137:50070/explorer.html#/来查看hadoop的数据情况

## 2.3.6 总结

用户⾏为日志收集的相关工作流程

Fluume收集到hive配置

supervisor进程管理工具使⽤

# 2.1 离线画像业务介绍

#### ⽂章内容标签化：内容标签化，根据内容定性的制定⼀一系列列标签，这些标签可以是描述性标签。针对于

**⽂章就是⽂章相关的内容词语。**

文章：频道ID内容，关键词、主题词

用户画像：研究用户对内容的喜好程度

# 离线⽂文章画像计算

## 离线⽂文章画像组成需求

文章画像，就是给每篇文章定义一些词。

关键词：文章中⼀一些词的权重（TFIDF与texrank）高的。

主题词：是进⾏规范化处理的，⽂文章中出现的同义词，计算结果出现次数高的词。共性的词

#### 1、原始文章表数据合并得到文章所有的词语句信息

文章标题+文章频道名称+文章内容组成文章完整内容

## 原始文章数据的合并

初始化spark信息配置，定义一个基类，这样

### 创建Spark初始化相关配置

合并三张表内容，到一张表当中，写入到HIVE中

article数据库：存放文章计算结果article\_data

建议初始化spark ， SparkSessionBase jupyter notebook先把代码写好，测试好

### 进行合并计算

3、新建merge\_data.ipynb⽂文件初始化spark信息

读取文章进行处理合并

DF 进⾏注册一个表，temp合并⽂文章频道名称

## Tﬁdf计算

article\_data

结巴分词

<https://github.com/fxsjy/jieba>



### ⽬目的

计算出每篇⽂文章的词语的TFIDF结果⽤用于抽取画像

结果：计算得出N篇⽂文章的TFIDF值，IDF索引结果合并得到词最终表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| hive> | select | \* from tfidf\_keywords\_values | where | article\_id=3; |
| OK |  |  |  |  |
| 3 | 17 | var 22.4813 |  |  |
| 3 | 17 | &# 41.1171 |  |  |
| 3 | 17 | 对象 14.0423 |  |  |
| 3 | 17 | 节点 56.4314 |  |  |
| 3 | 17 | 中⽂文 7.0541 |  |  |
| 3 | 17 | stringify 22.5552 |  |  |

#### 2、所有历史⽂文章Tﬁdf计算

* + - 1. **TFIDF模型的训练步骤**

读取N篇⽂文章数据

文章数据进⾏分词处理得到分词结果 分词的词如果是分布式，分词用的词库txt三台机器都需要上传

#### 先计算分词之后的每篇⽂文章的词频，得到CV模型

**然后根据词频计算IDF以及词，得到IDF模型**

训练idf模型，保存

cv \* log(C1出现文章次数/总文章数量)

不同词列表，所有的词的IDF的值利⽤用模型计算N篇文章数据的TFIDF值

tfidf\_keywords\_values：结果结果

#### ⽤用到idf\_keywords\_values这个表: 词以索引的对应关系

对于每篇⽂文章的每个词的权重做排序筛选

最终文章，词，tfidf权重都存在tfidf\_keywords\_values表中

idf\_keywords\_values 存的是词，和词的index

#### 3、所有历史文章TextRank计算

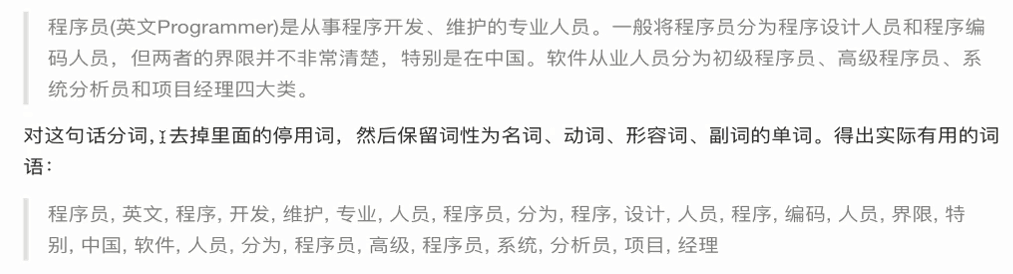
定义:通过词之间的相邻关系构建网络，然后用PageRank迭代计算每个节点的rank值，排序rank值即可得到关键词

词之间的相邻关系构建⽹网络

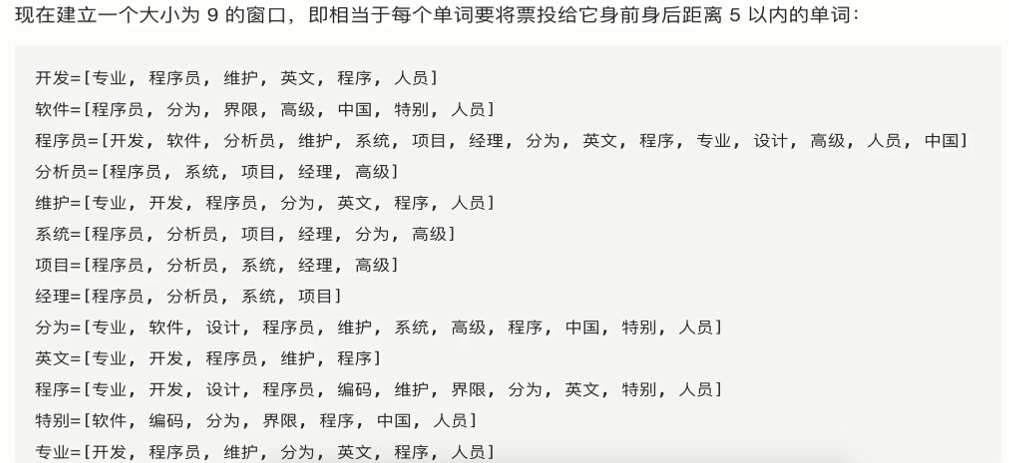
PageRank迭代计算每个节点的rank值(每个词，⽹⻚利⽤用图模型来提取文章中的关键词

举例10篇⽂文章

1. 分词过滤



1. 图建立，K窗⼝



3、计算权重排序

程序员=1.9249977,

人员=1.6290349, 分为=1.4027836, 程序=1.4025855,

⾼级=0.9747374,

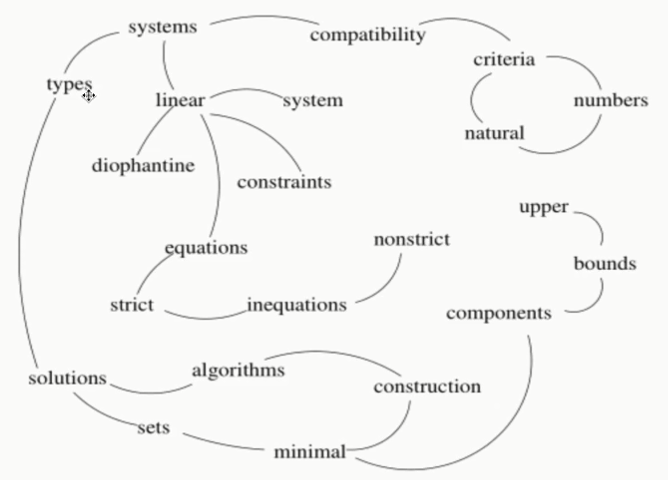
软件=0.93525416,

中国=0.93414587, 特别=0.93352026, 维 护 =0.9321688, 专 业 =0.9321688, 系 统 =0.885048, 编码=0.82671607, 界限=0.82206935, 开发=0.82074183,

分析员=0.77101076, 项⽬目=0.77101076, 英⽂文=0.7098714,

设计=0.6992446, 经理理=0.64640945

关键词评分：权重，那个词指向它的词比较多，权重就大



Github地址

https://github.com/baidu/lac

Pangbank是谷歌创始人发明的，可以了解一下

### 2.4.3.1 ⽂文章的TextRank计算

jieba.analyse.TextRank （我们textrank依靠结巴的）textrank\_model

textrank\_model = TextRank(window=5, word\_min\_len=2)

textrank\_keywords\_values tfidf\_keywords\_values