[1. 需求 1](#_Toc30349)

[1.1. 需求定义 1](#_Toc7883)

[1.2. 用例图 1](#_Toc4741)

[2. 开发视图 2](#_Toc16232)

[2.1. 开发简图 2](#_Toc30778)

[2.1.1. wifiprob数据接收流程描述 3](#_Toc18501)

[3. 部署视图 4](#_Toc32202)

[4. 引用 4](#_Toc27837)

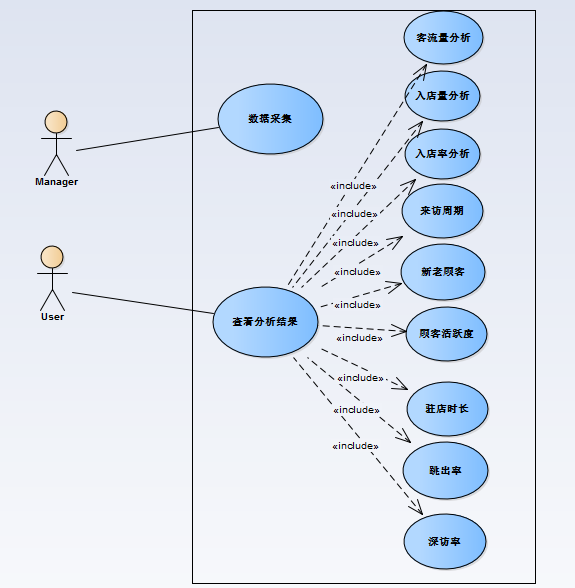
### 1. 需求

* 1. 需求定义

具体见需求文档。

|  |  |  |
| --- | --- | --- |
| **需求** | **需求描述** | **需求优先级** |
| WIFI探针 |  | 高 |
| 数据采集 |  | 高 |
| 客流量分析 |  | 中 |
| 入店量分析 |  | 中 |
| 入店率分析 |  | 中 |
| 来访周期分析 |  | 中 |
| 新老顾客分析 |  | 中 |
| 顾客活跃度 |  | 中 |
| 驻店时长分析 |  | 中 |
| 跳出率分析 |  | 中 |
| 深访率分析 |  | 中 |
| 安全性 |  | 高 |
| 可维护性 |  | 中 |
| 易用性 |  | 中 |
| 可靠性 |  | 中 |
|  |  |  |

* 1. 用例图



### 2. 开发视图

2.1. 开发简图

系统分为四部分：

\* **web服务器**

\* **wifi 探针接收服务器（acceptor）**

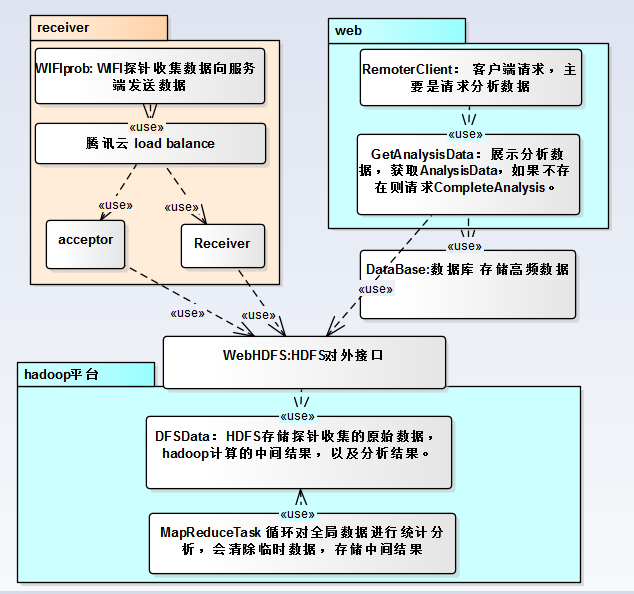
**\*关系型数据库**

\***计算与存储平台**

- hadoop计算集群 对数据进行全局的统计与分析

- hdfs文件系统 按块存储所有的数据

- 关系型数据库 存储高频度使用的数据，主要是对探针数据的统计结果



2.1.1. wifiprob数据接收流程描述

- 探针发送数据

- 探针服务器接收数据

- 探针接收服务器会对每条数据进行初步处理

- 这里会做一些统计分析，尤其是比较好进行增量分析的、不需要所有数据即可分析的统计。（比如总客户统计是很好进行增量分析的，但是回访量分析却没有很好的算法进行增量分析）

- 处理后会将数据存入队列

- 队列中数据达到一定量时会异步的写入HDFS文件系统（这里是为了尽可能的减少与HDFS的交互次数，也是为了较少HDFS的小文件数量）

2.1.2. 用户请求分析数据流程描述

- 用户发起请求

- web服务器收到请求

- web 服务器查询本地缓存是否有用户所需数据，有则返回

- web 服务器查询数据库是否有用户所需数据，有则返回

- web 服务器读取HDFS文件系统中的数据，并根据需要存入数据库与本地缓存中，都到后即返回。这里HDFS通过WEBHDFS以及Kerberos授权的方式提供远程调用。

2.1.3 hadoop平台描述

- hadoop集群会循环执行以下任务

- 合并hdfs中的小文件

- 执行mapreduce任务

- 清除临时文件

- 清除过期统计数据

2.1.4模块接口描述

2.1.4.1 hadoop提供接口

|  |  |  |
| --- | --- | --- |
| 1. | 接口 | readStatistic(Path) |
|  | 接口描述 | 通过WEBHDFS,根据路径读取HDFS文件中的统计数据 |
|  | 前置条件 | Path对应文件已存在，hadoop平台已启动 |
| 2. | 接口 | saveData(Data,Path) |
|  | 接口描述 | 通过WEBHDFS将数据存储至HDFS平台 |
|  | 前置条件 | Hadoop平台已启动，path指定的父目录存在 |

2.2. 详细设计

2.2.1 Web

2.2.2 Receiver

探针接收服务器以分布式部署，HTTP post的方式接收数据。

1. 负载均衡：使用腾讯云负载均衡，每个探针想服务器发送的数据都会均匀的负载到各台服务器上。
2. Session一致性：搭建分布式redis集群存储session信息，由于receiver服务器目前只接受wifi探针发送的数据，所以对session的一致性要求比较低。
3. 数据存储：由于HDFS不适合频繁写、小文件存储，所以每个节点都会对本地数据缓存，达到一定规模后才会写入HDFS文件。

2.2.3 MapReduce

2.2.3.1 算法描述

系统使用map-reduce算法统计当前为止的客户流量，新老客户，回访周期，入店时长。

算法使用两趟map-reduce来得出统计结果。

运行map-reduce程序前会清空HDFS临时文件夹。

**第一次map-reduce**，我们会把HDFS文件系统的所有探针数据按照用户MAC地址分类，并按照时间排序，存入HDFS临时文件夹中。

所有的用户行踪均根据探测到的指针来判断。

**第二次map-reduce**，我们会对每个用户(mac)已排序好的探针数据，统计探针数据在时间轴上的出现密度，由此得出用用户在店内的时间段，继而得出根据单个用户数据统计出来的客户流量，新老客户，回访周期，入店时长这几个统计数据。

统计单个用户的数据后，然后需要进行归并。

1. 对于客户流量与新老客户，是分时间段归并的。
2. 对于回访周期与入店时长，是分时间长短归并的。

从语言形式上与实际处理上来说，这两者的归并方式都差不多。

举列如下：

**(对应下表第二个map的结果)**对单个用户数据统计结果如下：

*用户A: 9：30点入店1次，作为新客户入店1次，作为老客户入店0次*

*11：20点入店1次，**作为新客户入店0次，作为老客户入店1次*

*11：30点入店1次，作为新客户入店0次，作为老客户入店1次*

*用户B: 9:40点入店1次，作为新客户入店1次，作为老客户入店0次*

*11:40点入店1次，作为新客户入店0次，作为老客户入店1次*

那么

**(对应第二个combiner的结果)**对多个用户数据在单台（mapper）机器上统计结果如下：

*9-10点入店2次，作为新客户入店2次，作为老客户入店1次*

*10-11点入店3次，作为新客户入店0次，作为老客户入店3次*

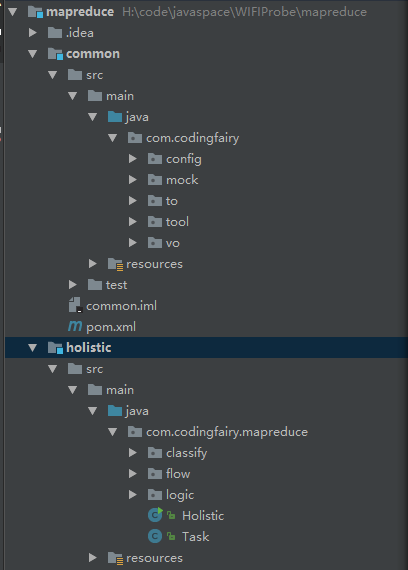
**(对应第二个reduce)**当用户A,B的数据被不同的mapper读取时，以上combiner的工作应当由reducer做。

如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | key | value | | 备注 |
|  | | | | | |
| 第一次读入数据 | | 行号 | 探针数据包 | |  |
| 第一个map输出 | | 用户的mac地址 | 探针数据包 | | 探针数据包中含有mac地址,我们把这个mac地址拿出来作为key，数据包作为value |
| combiner输出 | | 用户的mac地址 | 探针数据包 | |  |
| 第二个reducer输出 | | 用户的mac地址 | 探针数据包 | | 这里把每个用户的发送的数据包都归一后按照数据包的时间戳排序后存储，存储格式为：*用户mac地址【空格】用户所有数据包json列表* |
|  | | | | | |
| 第二次读入数据 | | 行号 | *用户mac地址【空格】用户所有数据包* | 这里的所有数据包都是时间上排过序的 | |
| 第二个map输出 | 客户流量 | “Flow”+Hour | (Hour,Hour+1)期间入店数，出店数，深访率，跳出率 | | 都是统计每一个时间区间内的响应统计项的数量。如：*2016.5.30日4到5点的入店数为12，出点数为45，深访率0.34，跳出率0.12，新客户7，老客户5。* |
| 新老客户 | “newOld”+Hour | (Hour,Hour+1)期间新客户数，老客户数 | |
| 回访周期 | “cycle”+时间区间（t,t+1） | 回访周期时间长度居于（t,t+1）区间的客户数 | | 都是统计落在某个区间的的客户数。  比如：*回访周期为（6-12小时）的客户有122个，入店时长为（3-5分钟）的客户有1241个* |
| 入店时长 | “inStore”+时间区间（t,t+1） | 入店时长时间长度居于（t,t+1）区间的客户数 | |
| combiner输出 | 客户流量 | 同map输出 | 同map输出 | | 对来自同一个map同一个KEY的数据累加,比如在1:05跟1:45的分别有一人入店，那么输入为{(1,2)=>1},{(1,2)=>1}（表示两个人都是在1-2点进入了店铺，但是没有合并），那么combiner会合并为{(1,2)=>2} |
| 新老客户 |
| 回访周期 |
| 入店时长 |
| 第二个reducer输出 | 客户流量 | 同map输出 | 内容同map输出  但会转换成json格式的Text | 同上，对来自不同的map输出同一个KEY的数据累加 | |
| 新老客户 |
| 回访周期 |
| 入店时长 |
|  | |  |  | |  |

Map reduce的统计结果会被存入HDFS文件系统中。

2.2.3.2 模块设计



**-- common 公共依赖的代码**

-- config HDFS各个文件夹路径配置，字符串值，服务器节点配置

-- to 实现了Writable与cloneable的value包装类，实现了ComparableWritable的key 包装类。方便的自定义包装各种类。

-- tool 常用工具类：如文件读写，日志打印，日期格式转换等。

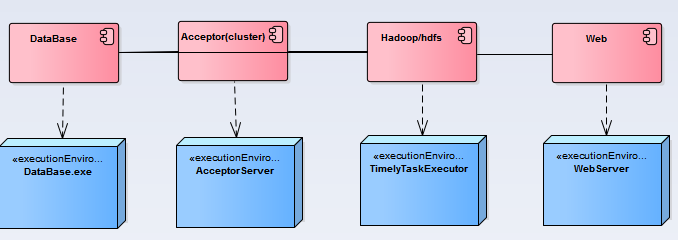
**-- holistic**

-- classify 用于根据mac地址分类的map-reduce程序，其中包括了Mapper、 Reducer,Combiner三个类

-- flow 用于统计主要分析项。包括了Mapper、Reducer,Combiner三个部分以及辅助类 MapperWriter

-- logic 主要是实现一线复杂的逻辑，是flow的辅助包

### 3. 部署视图



### 4. 引用