北京交通大学

基于大数据的智能推荐系统设计与实现

日志数据处理

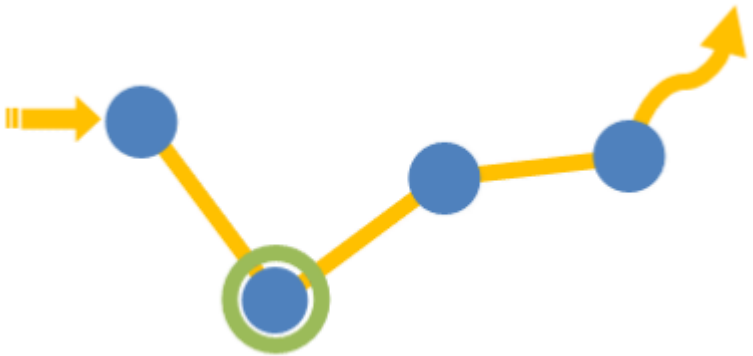
张迪
dizhang@bjtu.edu.cn

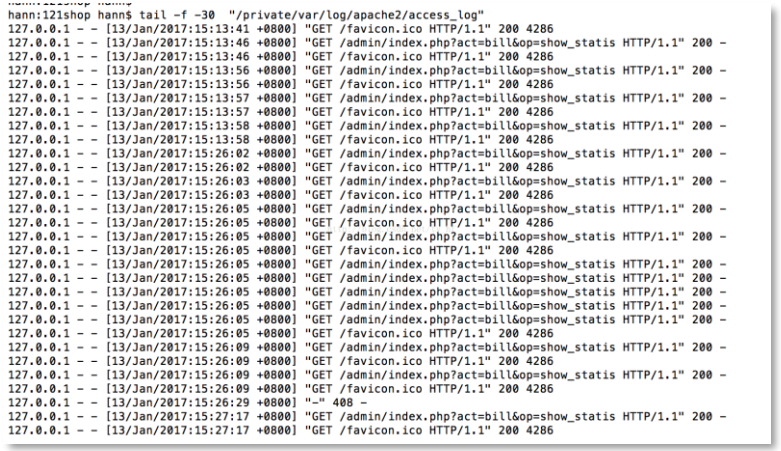
回顾：日志数据

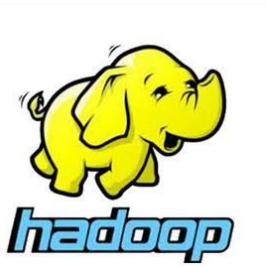
- 数据来源是由本项目的Web端程序运行所产生的访问日志
- 修改Tomcat日志格式，使Tomcat服务器产生的日志格式符合上面表格中的格式要求：
  - 如果使用的是纯Apache服务器，那么访问日志在Apache根目录的conf文件夹内的httpd.conf文件中配置
  - 如果使用的是Apache tomcat服务器，那么访问日志在Apache根目录的conf文件夹内的server.xml文件中配置
  - Apache Tomcat访问日志的最终格式设置为： pattern="%h %l %u %t "%r" %s %b "%{Referer}i" "%{User-Agent}i" "%{uuid}c;%{userId}c;%{st}c;" %I %O %{X-Forwarded-For}i %v %D"

为什么要处理日志数据？

- 日志数据包含很多有价值的用户信息
  - 最典型的信息：访问网站的用户数
  - 用户的访问来源
    - 搜索引擎
    - 地址导航页
    - 自主访问
  - 访问方式
    - 普通浏览器
    - 手机APP
  - 用户的点击流信息







如何处理日志数据？

- 日志数据的特征
  - 以时间为标签的数据
  - 相互之间的耦合度不高
- 日志数据的处理方式
  - 当数据量很小时
    - 简单的单机程序即可以很容易的完成处理
  - 当数据量很大时，该如何处理？
    - 基于Mapreduce的日志处理

```
hann:121shop hann$
hann:121shop hann$ tail -f -30  "/private/var/log/apache2/access_log"
127.0.0.1 - - [13/Jan/2017:15:13:41 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:13:46 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:13:46 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:13:56 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:13:56 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:13:57 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:13:57 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:13:58 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:13:58 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:02 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:02 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:03 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:03 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:05 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:09 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:09 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:09 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:26:09 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
127.0.0.1 - - [13/Jan/2017:15:26:29 +0800] "-" 408 -
127.0.0.1 - - [13/Jan/2017:15:27:17 +0800] "GET /admin/index.php?act=bill&op=show_statis HTTP/1.1" 200 -
127.0.0.1 - - [13/Jan/2017:15:27:17 +0800] "GET /favicon.ico HTTP/1.1" 200 4286
```

示例程序

- 工具类：HdfsUtil.java，对hdfs的一些操作命令进行封装
- 工具类：WebLogUtil.java，主要包括两个方法，用于对日志进行过滤和为日志添加SessionID
- Mapper类： WebLogMapper.java
- Reducer类： WebLogReducer.java
- 驱动类： WebLogDriver.java

相关任务

- 完成基于Hadoop的日志处理
  - 以示例程序为基础
  - 基于自己搭建的Hadoop平台
- 封装自己的数据分析和处理工具
  - 主要目的：通过程序直接执行系统的命令
  - HdfsUtil工具类，用于执行HDFS的相关操作
  - SqoopUtil工具类，用于实现在Python中调用Sqoop完成数据导入的操作
  - HiveUtil工具类，用于执行Hive查询

Thank you!

Q & A