大数据Hadoop高薪直通车课程

Hadoop 2.x 实战应用

讲师：轩宇（北风网版权所有)

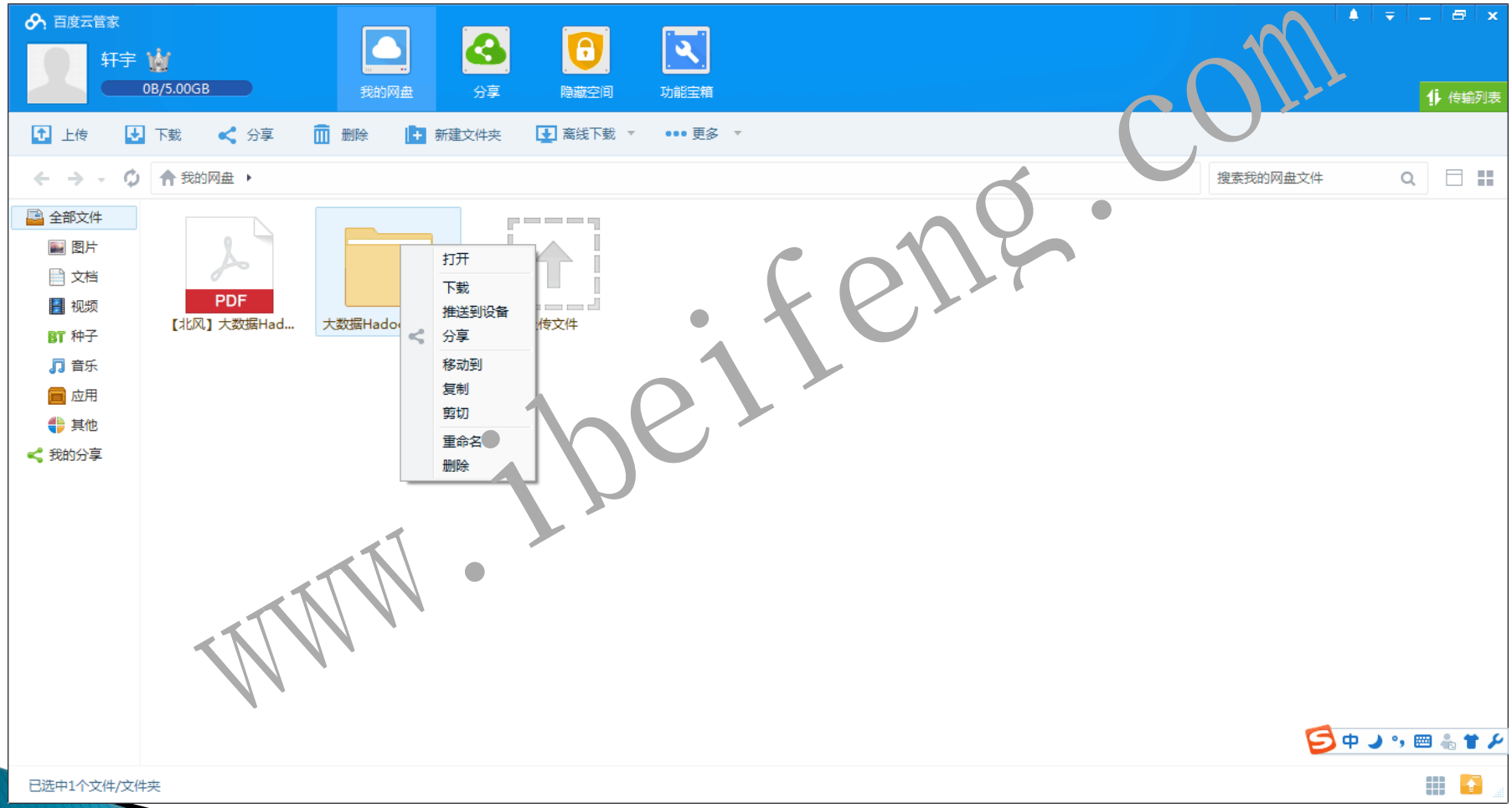
课程大纲

1. 基于HDFS云盘存储系统
2. Hadoop 三大发行版本
3. 项目实战之日志数据收集
4. 项目实战之日志数据预处理
5. 项目实战之日志数据分析

课程大纲

1. 基于HDFS云盘存储系统
2. Hadoop 三大发行版本
3. 项目实战之日志数据收集
4. 项目实战之日志数据预处理
5. 项目实战之日志数据分析

基于HDFS云盘存储系统



基于HDFS云盘存储系统

课程大纲

1. 基于HDFS云盘存储系统
2. Hadoop 三大发行版本
3. 项目实战之日志数据收集
4. 项目实战之日志数据预处理
5. 项目实战之日志数据分析

Cloudera Hadoop

- 2008 年成立的 **Cloudera** 是最早将 Hadoop 商用的公司，为合作伙伴提供 Hadoop 的商用解决方案，主要是包括**支持，咨询服务，培训**。
- 2009年Hadoop的创始人 Doug Cutting也加盟 Cloudera公司。Cloudera 产品主要为CDH，Cloudera Manager，Cloudera Support
- CDH是Cloudera的Hadoop发行版，**完全开源**，比Apache Hadoop在兼容性，安全性，稳定性上有所增强。
- Cloudera Manager是集群的软件分发及管理监控平台，可以在几个小时内部署好一个Hadoop集群，并对集群的节点及服务进行实时监控。Cloudera Support即是对Hadoop的技术支持。
- Cloudera 的标价为**每年每个节点4000美元**。Cloudera开发并贡献了可**实时处理大数据的Impala项目**。

Hortonworks Hadoop

- 2011年成立的**Hortonworks**是雅虎与硅谷风投公司Benchmark Capital合资组建
- 公司成立之初就吸纳了大约25名至30名专门研究Hadoop的雅虎工程师，上述工程师均在2005年开始协助雅虎开发Hadoop，**贡献了Hadoop 80%的代码**。
- 雅虎工程副总裁、雅虎Hadoop开发团队负责人Eric Baldeschwieler出任Hortonworks的首席执行官。
- Hortonworks 的主打产品是Hortonworks Data Platform (HDP)，也同样是**100%开源的产品**，HDP除常见的项目外还包含了**Ambari，一款开源的安装和管理系统**
- HCatalog，一个元数据管理系统，HCatalog现已集成到Facebook 开源的Hive中。Hortonworks的Stinger开创性地极大地优化了Hive项目。Hortonworks为入门提供了一个非常好的，易于使用的沙盒。
- Hortonworks开发了很多增强特性并提交至核心主干，这使得Apache Hadoop能够在包括Windows Server和Windows Azure在内的Microsoft Windows平台上本地运行。定价以集群为基础，**每10个节点每年为12500美元**。

版本选择

市面上有很多种 HBase 发行套件（或是软件包），并且每种有多个版本。目前最有名的发行套件是原生的 Apache 发行套件和 Cloudera 公司的 CDH。

- Apache——Apache HBase 项目是所有 HBase 开发的父项目。所有代码都集中到那里，各个公司的开发者给它贡献代码。和其他开源项目一样，版本发行周期取决于参与者（也就是雇佣开发人员从事项目开发的公司）和他们想把什么特性放进一个特定的版本。一般来说，HBase 社区和它们的版本是保持一致的。其中值得注意的版本包括 0.20.x、0.90.x、0.92.x 和 0.94.x。本书专注于 0.92.x。
- Cloudera 公司的 CDH——Cloudera 是一家在生态系统中有自己发行版本的公司，包括 Hadoop 和其他模块（包含 HBase）。这个套件被称为 CDH（Cloudera's distribution including Apache Hadoop）。CDH 建立在 Apache 的代码基础上，采用了特殊发行版本，并在里面添加了没有包含在任何 Apache 官方发行版本中的许多补丁。Cloudera 也根据客户需求添加额外的特性。在 Apache 代码基础里的补丁不一定出现在 CDH 所基于的同一代码基础分支里。

我们推荐使用 Cloudera 的 CDH 套件。通常它比原生的 Apache 发布版包含更多补丁，用来增加稳定性、改善性能、有时候增加功能特性。CDH 也比 Apache 版本被更好地测试过，并且比原生 Apache 运行在更多的生产集群上。在你为集群选择发行版本前，我们建议考虑这些要点。

对于提供的安装步骤，我们假设你已经安装了 Java、Hadoop 和 ZooKeeper。关于安装 Hadoop 和 ZooKeeper 的操作指南，请参考你选择的发行版本的文档。

课程大纲

1. 基于HDFS云盘存储系统
2. Hadoop 三大发行版本
3. 项目实战之日志数据收集
4. 项目实战之日志数据预处理
5. 项目实战之日志数据分析

Nginx服务器

在Nginx中日志文件是由log_format这个指令来定义的，它的语法如下：

```
log_format   name   format
##   name：指的是日志格式的名称（后面调用）
##   format：设置日志具体格式的
```

在Nginx中有自己默认的日志格式，如下内容

```
#log_format  main  '$remote_addr  $remote_user [$time_local] "$request" '
#                  '$status $body_bytes_sent "$http_referer" '
#                  '"$http_user_agent" "$http_x_forwarded_for"';
```

Nginx服务器

$remote_addr	客户端的ip地址（如果中间有代理服务器那么这里显示的ip就为代理服务器的ip地址）
$remote_user	用于记录远程客户端的用户名称（一般为“-”）
$time_local	用于记录访问时间和时区
$request	用于记录请求的url以及请求方法
$status	响应状态码
$body_bytes_sent	给客户端发送的文件主体内容大小
$http_user_agent	用户所使用的代理（一般为浏览器）
$http_x_forwarded_for	可以记录客户端IP，通过代理服务器来记录客户端的ip地址
$http_referer	可以记录用户是从哪个链接访问过来的

【北风网】日志格式

```
log_format  main  '"$remote_addr" "$remote_user" "$time_local" "$request" '
          '"$status" "$body_bytes_sent" $request_body "$http_referer" '
          '"$http_user_agent" "$http_x_forwarded_for" "$host"'
```

$remote_addr
    客户端的ip地址（如果中间有代理服务器那么这里
    显示的ip就为代理服务器的ip地址）
$remote_user
    用于记录远程客户端的用户名称（一般为“-”）
$time_local
    用于记录访问时间和时区
$request
    用于记录请求的url以及请求方法
$status
    响应状态码
$body_bytes_sent
    给客户端发送的文件主体内容大小
$request_body
    为post的数据
$http_referer
    可以记录用户是从哪个链接访问过来的
$http_user_agent
    用户所使用的代理（一般为浏览器）
$http_x_forwarded_for
    可以记录客户端IP，通过代理服务器来记录客户端的ip地址
$host
    服务器主机名称

业务需求之IP地址

◆ 【$remote_addr】

客户端的ip地址（如果中间有代理服务器那么这里显示的ip就为代理服务器的ip地址）。

◆ 业务

➢ 依据ip地址确定区域，定向营销，【IP地址 -> 地域】

中国 IP 地址段

```
******************************************************************
[211.64.0.0 - 211.71.255.255] 中国
[211.123.0.0 - 211.255.255.255] 中国
[210.25.0.0 - 210.47.255.255] 中国
```

➢ 用户统计，访问某一网站数

● 准确性（同一外网，不同内网）

业务需求之访问时间

- 【$time_local】

  用于记录访问时间和时区。

- 业务
  - 分析用户访问网站的时间段
  - 针对销售来说，合理安排值班，销售课程

业务需求之请求地址

- 【$request】

  用于记录请求的url以及请求方法。

- 业务
  - 用户最关注的网站 -> 课程
  - 定向投放此套课程，做好的相关课程

业务需求之转入链接

- 【$http_referer】

  可以记录用户是从哪个链接访问过来的。

- 业务

  - 关注用户如何，访问我们的课程，定向某个区域，进行广告投放

【北风网】日志存储

- 日志文件
  - 每天的文件按照【日期】存放在对应的文件夹中
  - 一天之内只产生一个文件，以每天零点为准

【北风网】日志收集

- 【日志文件】类型数据收集
  - 程序
    - 定时程序，如QUARTZ
  - 脚本
    - 简单的shell脚本

【北风网】日志预处理

- 数据预处理
  - 大数据处理的核心关键点

“数据质量”

  - 数据清洗、过滤，剔除【脏数据】
    - MapReduce程序
    - Hive
      - 原表 -> 业务表：使用HQL语句、python脚本、udf

【北风网】日志分析

◆ 特殊处理的字段

- $time_local，转换为Long类型数字，或者符合某个格式
- $request，提取出请求的URL
- $http_referer，上一个连接，计算网页中的【二级跳等】

```
1  "111.161.98.42"
2  "-"
3  "10/Aug/2015:00:00:01 +0800"
4  "GET /tech-15714393.html HTTP/1.1"
5  "302"
6  "0"
7  -
8  "http://www.ibeifeng.com/tech-15714393.html"
9  "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)"
10 "-"
11 "www.ibeifeng.com"
```

本课程版权归北风网所有

欢迎访问我们的官方网站

www.ibeifeng.com