产学合作协同育人项目 高等院校数据科学与大数据技术系列规划教材

教育部-普开数据教学内容与课程体系改革项目数据科学与大数据技术专业示范课程

全国大学生创新创业实践联盟大数据专业委员会推荐教材

大数据离线分析

傅德谦 主编

赵向兵 张林涛 刘鸣涛 副主编

Hadoop + Hive + Sqoop + Pig + Oozie

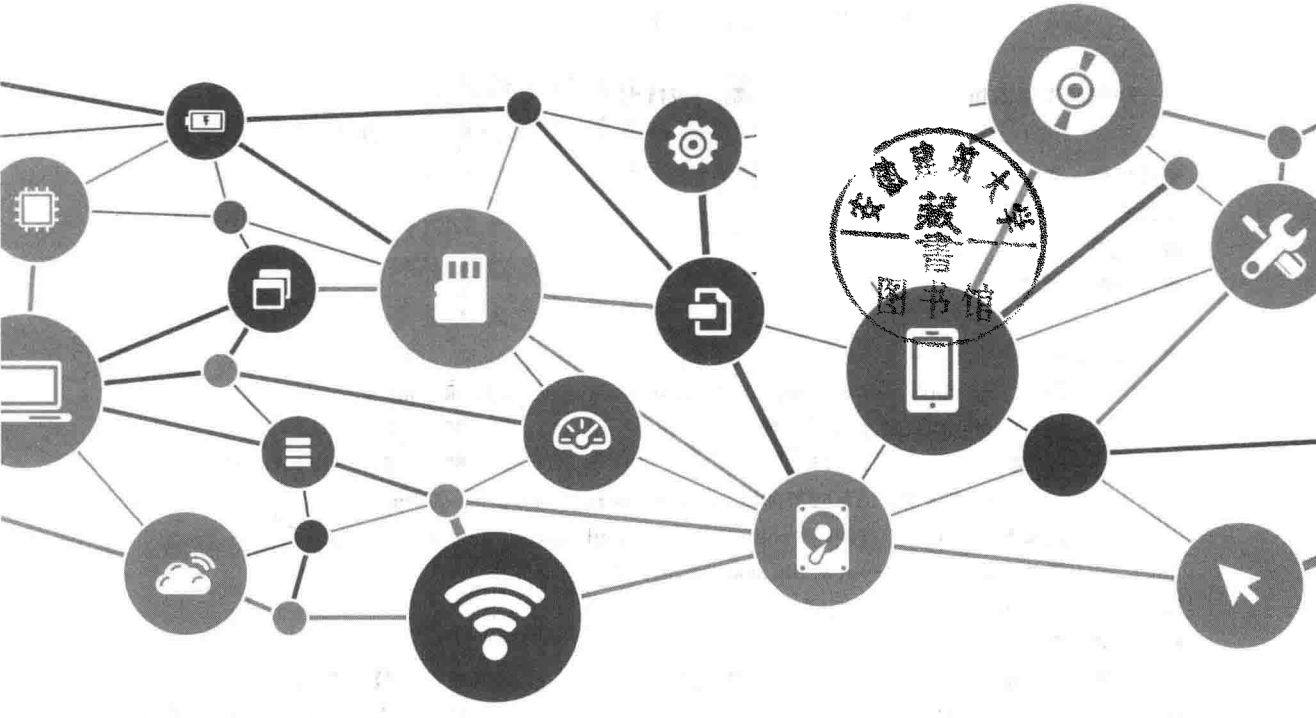
微博历史数据分析 + 电商销售数据分析

清华大学出版社

高等院校数据科学与大数据技术系列规划教材

大数据离线分析

傅德谦　主编
赵向兵　张林涛　刘鸣涛　副主编



清华大学出版社
北京

内 容 简 介

本书基于开源 Hadoop 大数据生态圈的主流离线分析工具 Hive 和 Pig，通过技术讲解和案例实战相结合的方式，介绍了海量数据离线分析的技术方法。本书内容主要包括 Hive 数据库表、基于 HiveQL 的常规操作、视图、索引和 Pig 等数据处理分析和基础工具知识，Hive 函数、Pig Latin 编程、ETL 工具 Sqoop 和工作流引擎 Oozie 等相关高级技术，以及实际项目案例。

本书既可供学习大数据离线分析技术的本科和高职高专学生作为教材，也可供从事数据分析相关工作的技术人员作为参考资料。

图书在版编目(CIP)数据

大数据离线分析/傅德谦主编. —北京：清华大学出版社，2017
(高等院校数据科学与大数据技术系列规划教材)
ISBN 978-7-302-48329-8

Ⅰ. ①大… Ⅱ. ①傅… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 215771 号

责任编辑：刘翰鹏
封面设计：傅瑞学
责任校对：袁　芳
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62770175-4278
印 装 者：三河市吉祥印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：11.5　　字　　数：273 千字
版　　次：2017 年 8 月第 1 版　　印　　次：2017 年 8 月第1 次印刷
印　　数：1～2000
定　　价：35.00 元

产品编号：076512-01

丛书编委会

（排名不分先后，按姓名汉语拼音排序）

前言

为什么要写这本书

数据时代(Data Time)的到来使大数据技术得到了学术界和产业界的重视，并获得了快速发展。随着全球数字化、移动互联网和物联网在各行各业的应用发展，使累积的数据量越来越大。诸多先行的企业、行业和国家已经证明，利用大数据技术可以更好地服务客户、发现新商业机会、扩大新市场、转换新动能。

当前正处于大数据产业发展的前期，市场需求日趋旺盛，但是人才缺口巨大，技术支撑严重不足，大数据专业知识的广泛传播非常紧迫。

本书基于教育部“2016年产学合作协同育人项目”——普开数据教学内容和课程体系改革项目，作为项目成果公开出版。北京普开数据技术有限公司在多届全国高校教师培训工作中起到了“种子”教师培养的作用，本书编者都是在培训过程中结识并展开合作的；同时在本书编写过程中，公司给予了强力支持，在此表示感谢。

读者对象

(1) 学习大数据离线分析的本科和高职高专学生。

(2) 从事数据分析相关工作的技术人员。

如何阅读本书

本书主要介绍了基于Hadoop生态圈的大数据离线处理技术。主流的大数据离线分析技术一般包括：使用HDFS存储数据，使用MapReduce做批量计算；需要数据仓库的存入Hive，从Hive进行分析和展现；涉及复杂业务场景时，使用Sqoop、Pig、Oozie等工具会更加灵活方便。

本书略过了HDFS存储数据、MapReduce批量计算的相关内容。HDFS是Hadoop提供的分布式存储框架，它可以用来存储海量数据，MapReduce是Hadoop提供的分布式计算框架，它可以用来统计和分析HDFS上的海量数据。该部分内容为Hadoop基础知识，读者如果需要深入学习，可以参考其他书籍或材料(如清华大学出版社2016年6月出版的《大数据技术基础》)。

本书内容是重点围绕Hive数据仓库展开的，Hive在Hadoop上提供了SQL接口，开发人员只需要编写简单易上手的SQL语句就可以实现创建表、删除表、加载数据、下载数据、分析数据等功能，读者可以从目录的章节名称中快速检索并学习各方面的知识。

同时，本书针对离线分析过程中的工程任务场景还提供了一些辅助工具介绍。Sqoop解决在Hadoop和关系数据库之间传递数据的问题，如果读者有这方面的基础或对其他ETL工具更熟悉，可以略过。Pig为大型数据集的处理提供了更高层次的抽象，以更灵活方便的方法实现加载数据、表达转换数据和存储最终结果，有这方面基础或暂无需求的读者可以略过书中第6、7章。Oozie实现对系统中多任务的管理，当平台中任务数量很大、需要维

护和运行时，Oozie 可以方便地完成调度监控这些任务的功能，对于仅处理简单任务场景的读者可以略过该部分内容。

偏重实践操作是本书的特色，书中所讲内容基本都配有实践操作演示。通过每部分知识的学习和相应操作环节，可以很快地掌握技术，并有很强的工程应用场景感。本书最后提供了一个综合应用案例，读者可以应用所学知识实现一个工程项目，从而有效训练工程应用开发能力。

勘误和支持

由于本书编者水平有限，书中难免会出现一些错误或者不准确的地方，恳请读者批评、指正。如果在教材使用中遇到问题，或者要学习更多相关内容，请关注微信号 lemonedu 或联系普开数据在线实验平台(lab. zkpk. org)。

编　者

2017 年 6 月

绪　论

1. 大数据离线分析的知识背景

大数据技术有两个主要的方向：大数据平台相关的构建、优化、运维和监控；大数据的ETL(Extract-Transform-Load，抽取-转换-加载)、存储、计算和分析挖掘。在大数据分析的技术中，根据数据特点和应用需求又分为离线分析和实时计算两类技术，本书是针对前者。大数据离线处理主要的特点有：数据量巨大(静态的冷数据或温数据)且保存时间长；以复杂的批量运算为主要运算类型；以应用需求为导向产生中间数据或最终结果。

大数据离线分析的场景是数据总量很大、类型复杂，但真正有价值的数据可能占比例很小，需要通过从大量不相关的、各种类型的数据中去分析、挖掘，发现新的规律和新的价值，给业务提供发展趋势、模型预测等决策支持。

大数据离线分析要解决的另一个痛点是进行快速处理和分析。由于数据量很大、逻辑复杂，一个流程需要跑几十个或上百个包，使用传统技术耗时太长，有些情况下计算结果没出来就已失效。当然，传统分析技术知识对于读者在学习本书时仍然会有很大的帮助，技术架构虽然不同，但分析方法还是相近的。

2. 大数据离线分析的应用场景

目前大数据做得比较深的行业主要还是集中在互联网、电商、金融、银行和生产制造等领域。离开业务应用场景谈大数据分析就是一个空中楼阁，每一个行业、同一个行业不同的企业，对大数据的应用需求和流程都不相同，需要懂其自身的行业、懂其自身的业务。

懂行业、深挖业务是大数据离线分析项目所必需的基础。在大数据项目的规划和落地的过程中，首先要分析业务场景有哪些？确定了场景，就基本有了调研需求细节的方向。然后，考虑这些业务场景需要哪些数据资源可以支撑？哪些数据资源是可以内部解决的，哪些数据资源是需要通过外部合作的，有哪些必要的数据资源是现在没有但是可以通过增加数据获取渠道来获取的？

大数据的应用主要分为 OLTP 和 OLAP 两种。大数据离线分析支撑 OLAP。OLTP 用于支持线上业务，需要快速响应用户请求、支持事务，对于容错性和稳定性要求非常高；OLAP 主要是离线计算，用来做数据分析，实现推荐、统计、预测、决策等业务，这也是本书讨论的主题。

对于大多数从业者来讲，一般没有太深的行业背景；当然，有行业背景的人也一般不具备大数据技术，这是实际情况。需要注意的是，大数据技术工作者要有与行业专家密切配合的思维，在组建团队和开发过程中需要时刻保持这种协作意识。

3. 大数据离线分析的工具介绍

大数据离线处理目前技术上已经成熟。Hadoop 框架是主流技术，使用 HDFS 存储数

据，使用 MapReduce 做批量计算；需要数据仓库的存入 Hive，从 Hive 进行分析和展现；涉及复杂业务场景时，使用 Sqoop、Pig、Oozie 等工具会更灵活方便。其中，MapReduce 批量计算是基于 Hadoop 大数据离线分析的基础，HDFS 是 Hadoop 提供的分布式存储框架，它可以用来存储海量数据；MapReduce 是 Hadoop 提供的分布式计算框架，它可以用来统计和分析 HDFS 上的海量数据；Hive 分析的最终执行也会转换成 MapReduce 的分布式批量计算过程（该部分为 Hadoop 基础知识，本书不做介绍）。本书重点介绍 Hive 相关知识和常用的大数据离线分析辅助工具。

Hive 作为一种基于 Hadoop 的数据仓库工具，通过 Hive（SQL on Hadoop）在 Hadoop 上提供 SQL 接口，开发人员只需要编写简单易上手的 SQL 语句就可以实现创建表、删除表、加载数据、下载数据、分析数据等功能。不同于传统数据仓库技术的，Hive 会负责把 SQL 翻译成 MapReduce，在 Hadoop 分布式平台上以并行化的方式运行，效率会大大提升。

Sqoop 是在 Hadoop 和关系数据库之间传递数据的工具。通过 Sqoop 可以方便地将数据从关系数据库导入 HDFS、Hive 表，或者将数据从 HDFS 导出到关系数据库。就像 Hive 把 SQL 翻译成 MapReduce 一样，Sqoop 把指定的参数翻译成 MapReduce，提交到 Hadoop 运行，完成 Hadoop 与其他数据库之间的数据交换。

Pig 为大型数据集的处理提供了更高层次的抽象，简化了 Hadoop 常见的工作任务，以更灵活方便的方法实现加载数据、表达转换数据和存储最终结果。

Oozie 是实现对系统中多任务管理的工具。大数据分析、数据采集、数据交换等都是一个个的任务，这些任务中有的是定时触发，有的需要依赖其他任务来触发，当平台中有几百上千个任务需要维护和运行时，Oozie 可以方便地完成调度和监控这些任务的功能。

第1章

走进 Hive

本章摘要

在系统地学习了 Hadoop 之后，还需要了解 Hadoop 生态圈里面的很多组件，Hive 就是其中之一，扮演数据仓库的角色。Hive 建立在 Hadoop 集群的上层，侧重于离线分析，对存储在 Hadoop 集群上的数据提供类 SQL 的接口进行操作，实现简单的 MapReduce 统计。

本章将从 Hive 的发展史讲起，然后讲解 Hive 的安装部署，接着会对 Hive 的一些命令以及它的数据类型和文件格式进行介绍，让读者在学习本章的内容后能对 Hive 组件有一个初步的认识。

1.1 Hive 简介

Hive 是建立在 Hadoop 上的开源数据仓库基础构架，用于存储和处理海量结构化数据。作为一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制，它提供了一系列的工具用来进行数据提取、转化、加载(ETL)，定义了简单的类 SQL 查询语言(称为 HQL)，允许熟悉 SQL 的用户方便地使用 Hive 查询数据；同时也允许熟悉 MapReduce 的开发者开发自定义的 Mapper 和 Reducer 来处理内建的 Mapper 和 Reducer 无法完成的复杂的分析工作。可以把 Hive 中海量结构化数据看成一张张的表，而实际上这些数据是分布式存储在 HDFS 中的。Hive 经过对语句进行解析和转换，最终生成一系列基于 Hadoop 的 Map/Reduce 任务，通过执行这些任务完成数据处理。

1.1.1 Hive 发展史

1. Hive 的诞生

Hive 诞生于 Facebook 中的日志分析需求，其设计目的是让精通 SQL 技能的分析师能够在 Facebook 存放在 HDFS 的大规模数据集上进行查询。面对海量的结构化数据，Hive 以较低的成本完成了以往需要大规模数据库才能完成的任务，并且学习门槛相对较低，应用开发灵活而高效。

2. Hive 的历史

Hive 自 2009 年 4 月 29 日发布第一个官方稳定版 0.3.0 到今天，在短短的几年时间里，一直在逐步的完善之中。从 2010 年下半年开始，Hive 成为 Apache 顶级项目。今天，Hive 已经是一个成功的 Apache 项目，很多组织把它用作一个通用的、可伸缩的数据处理

平台。

1.1.2 体系结构

Hive从外部接口中获取用户提交的HQL命令，然后对用户指令进行解析(需要元数据信息)，实例化成一个MapReduce可执行计划，按照该计划生成MapReduce任务后交给Hadoop集群基于用户指定的数据进行处理，最终反馈结果给用户。

Hive的架构如图1-1所示，主要由以下4个部分组成。

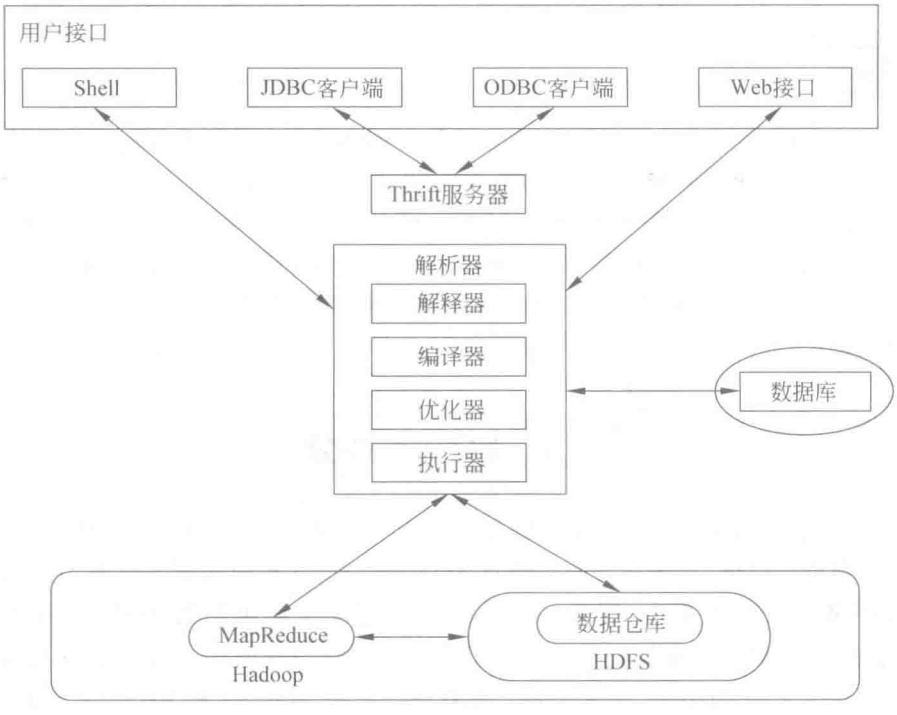


图1-1 Hive架构图

(1) 用户接口。用户接口主要有CLI、Client和Web UI。其中，最常用的是CLI，CLI启动时会同时启动一个Hive副本。Client是Hive的客户端，帮助用户连接至Hive Server。在启动Client模式时，需要指出Hive Server所在节点，并且在该节点启动Hive Server。Web UI提供通过浏览器访问Hive的方式。

(2) 数据库。Hive将元数据存储在数据库中，如MySQL、Derby。Hive中的元数据包括表的名字、表的列、分区及其属性、表的属性(是否为外部表等)和表的数据所在目录等。

(3) 解析器。解析器包括解释器、编译器、优化器、执行器。前三个完成HQL查询语句从词法分析、语法分析、编译、优化以及查询计划的生成。生成的查询计划存储在HDFS中，并在随后由MapReduce调用执行。

(4) Hadoop。Hive的数据存储在HDFS中，大部分的查询、计算由MapReduce完成。

1.2 Hive 的安装部署

1.2.1 安装配置 Hive

Hive 的安装需要在 Hadoop 已经成功安装的基础上，并且要求 Hadoop 已经正常启动（选用 Hadoop 2.7.2）。因为将 Hive 安装在 HadoopMaster 节点上，所以下面的所有操作都在 HadoopMaster 节点上进行。

Hadoop 默认的系统用户是 zkpk，密码也是 zkpk，下面所有的操作都使用 zkpk 用户，切换 zkpk 用户的命令是：

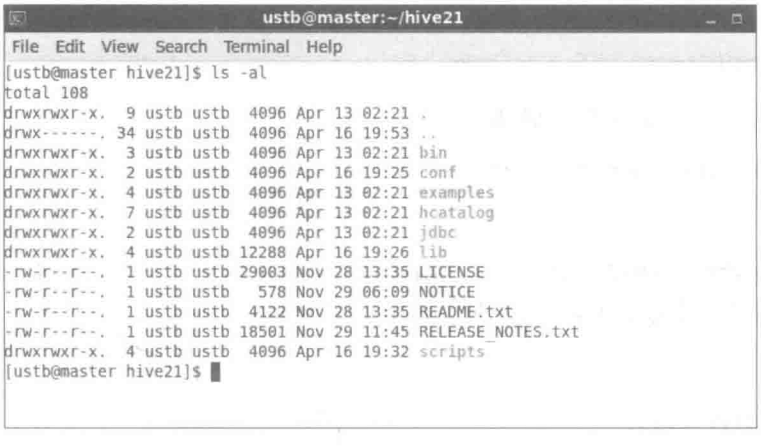
```
[zkpk@master ~]$su - zkpk
```

1. 解压并安装 Hive

使用下面的命令解压 Hive 2.1.1 安装包。

```
[zkpk@master ~]$cd /home/zkpk/resources/software/hadoop/apache
[zkpk@master apache]$mv apache-hive-2.1.1-bin.tar.gz ~/
[zkpk@master apache]$cd
[zkpk@master ~]$tar -zxvf ~/ apache-hive-2.1.1-bin.tar.gz
[zkpk@master ~]$mv apache-hive-2.1.1-bin hive21
[zkpk@master ~]$cd hive21
```

执行 ls -al 命令检查解压内容，会看到下面所示内容都是 Hive 包含的文件。



2. 安装配置 MySQL

注意：安装和启动 MySQL 服务需要 root 权限，因此首先要切换成 root 用户，然后用 yum 命令（需要联网）安装。

```
[zkpk@master ~]$su root
```

输入密码：

```
zkpk
```

(1) 安装 MySQL

方法一：通过 yum 安装 MySQL。

```
[root@master zkpk]$yum install mysql-server
[root@master zkpk]$yum install mysql-client
[root@master zkpk]$yum install mysql-devel
```

注意：如果命令执行过程中提示软件包已经安装了，则不需要重新安装。

方法二：通过 rpm 安装包安装 MySQL。

首先检查 MySQL 是否已经安装(注意命令中间有“|”)。

```
[root@master zkpk]$rpm -qa | grep mysql
```

删除 Linux 上已经安装的 MySQL 相关库信息。

```
[root@master zkpk]$rpm -e 上步查询出的结果-nodeps
```

删除后再次通过 rpm -qa 命令检查是否卸载干净。

然后通过如下命令安装 MySQL(其中 Mysql-server-***需要写实际安装的压缩包名)。

```
[root@master zkpk]$rpm -ivh Mysql-server-***
```

(2) 启动 MySQL 服务

安装之后，用如下方式启动服务。

```
[root@master zkpk]$/etc/init.d/mysqld restart
```

或

```
[root@master zkpk]$service mysqld restart
```

如果看到如下的打印输出，表示启动成功。

```
[root@master zkpk]# /etc/init.d/mysqld restart
Stopping mysqld:                                           [  OK  ]
Starting mysqld:                                           [  OK  ]
```

以 root 用户登录 MySQL(注意这里的 root 是数据库的 root 用户，不是系统的 root 用户)。默认情况下，root 用户没有密码，可以通过下面的方式登录。

```
[root@master zkpk]$mysql -uroot
```

然后创建 Hadoop 用户并授权。

```
mysql>grant all on *.* to hadoop@'%' identified by 'hadoop';
mysql>grant all on *.* to hadoop@'localhost' identified by 'hadoop';
mysql>grant all on *.* to hadoop@'master' identified by 'hadoop';
mysql>flush privileges;
```

创建数据库。

```
mysql>create database hive21;
```

输入命令退出 MySQL。