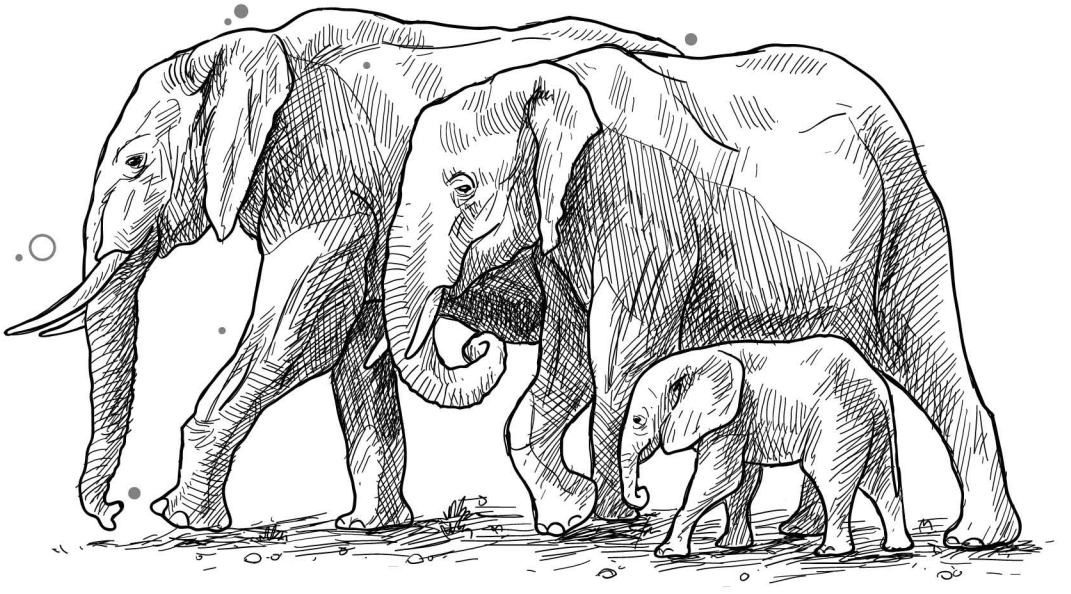
企业级大数据项目实战

用户搜索行为分析系统从0到1

张伟洋 / 著



清華大學出版社
北京

内 容 简 介

本书基于真实业务场景，以项目导向为主线，从0到1全面介绍“企业级大数据用户搜索行为分析系统”的搭建过程。全书共6章，第1章讲解项目需求与架构设计，详细阐述项目数据流与系统架构；第2章介绍大数据项目开发环境配置，手把手带领读者配置操作系统、Hadoop集群与相关工具，为后续项目实施打下基础；第3～5章逐步实现项目需求，第3章讲解“用户行为数据采集模块”的开发，第4章讲解“用户行为数据离线分析模块”的开发，第5章讲解“用户行为数据实时分析模块”的开发，这3章采用项目导向的方式，让读者参与实际开发过程；第6章讲解“数据可视化模块”的开发，并整合各模块，测试数据流转，完成项目的开发与部署。

本书项目源自真实业务场景，目的是使读者通过实际项目来理解理论知识并提高实践能力。本书适合缺乏大数据项目经验的从业者阅读，也适合作为高等院校大数据专业的教学用书。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。举报：010-62782989，beiqinquan@tup.tsinghua.edu.cn。

图书在版编目（CIP）数据

企业级大数据项目实战：用户搜索行为分析系统从 0 到 1/张伟洋著. —北京：清华大学出版社，2023.6

ISBN 978-7-302-63090-6

Ⅰ. ①企… Ⅱ. ①张… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国国家版本馆 CIP 数据核字（2023）第 047338 号

责任编辑：王金柱
封面设计：王　翔
责任校对：闫秀华
责任印制：刘海龙

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：010-83470000　　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
印 装 者：三河市人民印务有限公司
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　　张：**17　　**字　　数：**459 千字
版　　次：2023 年 6 月第 1 版　　**印　　次：**2023 年 6 月第 1 次印刷
定　　价：89.00 元

产品编号：101069-01

前　言

当今互联网已进入大数据时代，大数据技术已广泛应用于各行各业。不同领域每天都会产生海量数据，数据计量单位已从 TB 发展到 ZB，未来数据量还将爆发式增长。谷歌、阿里巴巴、百度、京东等互联网公司都急需掌握大数据技术的人才，目前这类人才出现了供不应求的状况。

市面上大数据相关的图书不少，但以真实项目贯穿全书来介绍技术和提升应用能力的并不多。事实上大数据技术涉及的内容和工具非常多，学习曲线也并不顺滑，而且仅仅了解理论知识但没有经过项目开发的检验，也无法很好地掌握这些知识。如何事半功倍地学习和掌握大数据技术，这正是本书编写的初衷。本书以实现一个“企业级大数据用户搜索行为分析系统”项目为导向，内容涵盖当前主流的大数据开发框架 Hadoop、ZooKeeper、Kafka、Hive、HBase、Spark 等。

本书借助真实业务场景，详解项目实现过程，插入案例分析与动手练习，帮助读者提高动手能力。

本书内容

全书共 6 章，内容如下：

第 1 章讲解项目需求与架构设计，概览项目整体规划。

第 2 章讲解大数据项目开发之前对操作系统集群环境的配置，包括虚拟机的创建、CentOS 7 操作系统的安装、Hadoop 的安装等，手把手带领读者搭建项目所使用的大数据开发基础环境，为后续的项目实操打下坚实的基础。

第 3~5 章逐步实现项目需求。以项目为导向，讲解“用户行为数据采集模块”“用户行为数据离线分析模块”和“用户行为数据实时分析模块”的开发。

第 6 章讲解“数据可视化模块”的开发，并将所有模块整合到一起，测试数据的流转，最终完成了整个项目的开发与系统的搭建。

本书各章均采用手把手的教学方式，读者可以对照书中的步骤从 0 到 1 成功搭建属于自己的大数据分析系统。

本书特点

- 由一线资深大数据专家根据真实业务场景编撰，使用当前流行的大数据开发技术、工具和框架。
- 使用简明的语言进行描述，易于理解与实践。
- 以完整项目为主线，帮助读了解企业级项目的构建流程。
- 讲解项目开发的同时介绍相关理论知识，有助于读者查漏补缺，深入理解。
- 手把手教学，可以边学习边实践，有效提高实践能力。

学习本书的建议

推荐按章节顺序阅读并上机实操，因为后续章节是建立在前面的基础上的，这种循序渐进的方式能让读者更加容易地掌握大数据开发技能，并完成项目开发。

首先阅读第 1、2 章，了解项目的主要功能、技术架构与集群规划，以对项目有整体认识，学习搭建开发环境。

然后依次学习第 3~6 章。学习每章前，先了解操作目的与该章内容在项目中的位置，然后实操搭建环境，编写应用程序，如此效果更佳。

按书中步骤实操，便可成功搭建本书的大数据项目。搭建成功后，读者还可以举一反三，加入一些创意功能，以令本书项目更完善、功能更强大。读者也可在网站与他人讨论，互帮互助。

本书适合的读者

- 了解大数据框架，想通过项目实践提高技术应用水平的人员。
- 缺乏项目经验的大数据从业者。
- 培训机构与高校大数据专业的学生。

本书配套资源

本书还提供了源码和教学课件，读者可以扫描下面的二维码，按照页面提示把下载链接转发到自己的邮箱进行下载。如果在阅读本书的过程中发现问题，请用电子邮件联系 booksaga@126.com，邮件主题写“企业级大数据项目实战：用户搜索行为分析系统从 0 到 1”。





读者如对书中知识有疑问，可通过微信公众号“奋斗在 IT”联系作者，该公众号提供 Java 大数据学习教程与视频资源。

因时间与水平限制，书中难免存在疏漏，欢迎读者批评指正。

张伟洋

2023.3

目　录

第 1 章　项目需求描述 …… 1

1.1　项目需求 …… 1
1.2　项目数据流设计 …… 2
1.3　项目架构设计 …… 3
1.4　集群角色规划 …… 6
1.5　项目开发环境介绍 …… 7

第 2 章　项目开发环境准备 …… 9

2.1　VMware 中安装 CentOS 7 操作系统 …… 9
2.1.1　下载 CentOS 7 镜像文件 …… 10
2.1.2　VMware 新建虚拟机 …… 12
2.1.3　安装 CentOS 7 …… 16
2.1.4　启动 CentOS 7 …… 21
2.1.5　打开 Shell 终端 …… 22
2.2　Linux 系统环境配置 …… 22
2.2.1　新建用户 …… 23
2.2.2　修改用户权限 …… 23
2.2.3　关闭防火墙 …… 24
2.2.4　设置固定 IP …… 24
2.2.5　修改主机名 …… 28
2.2.6　新建资源目录 …… 28
2.3　安装 JDK …… 29
2.4　克隆虚拟机 …… 30
2.5　配置主机 IP 映射 …… 33
2.6　配置集群各节点 SSH 无密钥登录 …… 34
2.6.1　SSH 无密钥登录原理 …… 34
2.6.2　SSH 无密钥登录操作步骤 …… 35
2.7　搭建 Hadoop 分布式集群 …… 36
2.7.1　搭建思路 …… 37
2.7.2　搭建 Hadoop 集群 …… 37
2.8　动手练习 …… 45

第 3 章　用户行为数据采集模块开发 …… 46

3.1　用户行为数据来源 …… 47
3.1.1　构建测试数据 …… 47
3.1.2　数据预处理 …… 48
3.2　使用 Flume 采集用户行为数据 …… 48
3.2.1　Flume 采集架构 …… 49
3.2.2　Flume 组件 …… 51
3.2.3　Flume 的安装与测试 …… 52
3.2.4　配置 Flume 多节点数据采集 …… 55
3.3　使用 Kafka 中转用户行为数据 …… 57
3.3.1　ZooKeeper 集群的搭建 …… 58
3.3.2　ZooKeeper 集群的启动与连接 …… 60
3.3.3　Kafka 集群的搭建 …… 61
3.3.4　Kafka 集群的启动与查看 …… 63
3.3.5　Kafka 主题操作 …… 64
3.4　Flume 数据实时写入 Kafka …… 67
3.4.1　数据流架构 …… 67
3.4.2　配置 centos03 节点的 Flume …… 67
3.4.3　启动 Flume …… 68
3.4.4　测试数据流转 …… 69
3.5　使用 HBase 存储用户行为数据 …… 69
3.5.1　HBase 集群的架构 …… 70
3.5.2　HBase 集群的搭建 …… 73

3.5.3 HBase 集群的启动、查看与停止……75
3.5.4 测试 HBase 数据表操作……77
3.5.5 创建 HBase 用户行为表结构……80
3.6 Flume 数据实时写入 HBase……81
3.6.1 数据流架构……81
3.6.2 配置 centos03 节点的 Flume……81
3.6.3 Flume 写入 HBase 原理分析……83
3.6.4 用户行为日志匹配测试……84
3.6.5 启动 Flume……85
3.6.6 测试数据流转……86
3.7 动手练习……87

第 4 章 用户行为数据离线分析模块开发……88

4.1 Hive 安装……88
4.1.1 Hive 内嵌模式安装……89
4.1.2 Hive 本地模式安装……92
4.1.3 Hive 远程模式安装……94
4.2 Hive 数据库操作……97
4.2.1 创建数据库……97
4.2.2 修改数据库……97
4.2.3 选择数据库……99
4.2.4 删除数据库……99
4.2.5 显示数据库……99
4.3 Hive 表操作……100
4.3.1 内部表操作……101
4.3.2 外部表操作……105
4.4 Hive 离线分析用户行为数据……107
4.4.1 创建用户行为表并导入数据……107
4.4.2 统计前 10 个访问量最高的用户 ID 及访问数量……108
4.4.3 分析链接排名与用户点击的相关性……109
4.4.4 分析一天中上网用户最多的时间段……109
4.4.5 查询用户访问最多的前 10 个网站域名……110
4.5 Hive 集成 HBase 分析用户行为数据……110
4.5.1 Hive 集成 HBase 的原理……111
4.5.2 Hive 集成 HBase 的配置……111
4.5.3 Hive 分析 HBase 用户行为表数据……112
4.6 Spark 集群的搭建……114
4.6.1 应用提交方式……114
4.6.2 搭建集群……116
4.7 Spark 应用程序的提交……118
4.7.1 spark-submit 工具的使用……118
4.7.2 执行 Spark 圆周率程序……119
4.7.3 Spark Shell 的启动……120
4.8 Spark RDD 算子运算……121
4.8.1 Spark RDD 特性……121
4.8.2 创建 RDD……123
4.8.3 转换算子运算……124
4.8.4 行动算子运算……130
4.9 使用 IntelliJ IDEA 创建 Scala 项目……131
4.9.1 在 IDEA 中安装 Scala 插件……132
4.9.2 创建 Scala 项目……133
4.10 Spark WordCount 项目的创建与运行……134
4.10.1 创建 Maven 管理的 Spark 项目……135
4.10.2 编写 WordCount 程序……137
4.10.3 提交 WordCount 程序到集群……138
4.10.4 查看 Spark WebUI……139

4.10.5　查看程序执行结果 …… 141
4.11　Spark RDD 读写 HBase …… 141
4.11.1　读取 HBase 表数据 …… 142
4.11.2　写入 HBase 表数据 …… 144
4.12　使用 Spark SQL 实现单词计数 …… 151
4.12.1　Spark SQL 编程特性 …… 151
4.12.2　Spark SQL 的基本使用 …… 153
4.12.3　Spark SQL 实现单词计数 …… 155
4.13　Spark SQL 数据源操作 …… 159
4.13.1　基本操作 …… 159
4.13.2　Parquet 文件 …… 164
4.13.3　JSON 数据集 …… 166
4.13.4　Hive 表 …… 167
4.13.5　JDBC …… 169
4.14　Spark SQL 与 Hive 整合分析 …… 170
4.14.1　整合 Hive …… 171
4.14.2　操作 Hive …… 173
4.15　Spark SQL 整合 MySQL 存储分析结果 …… 175
4.15.1　MySQL 数据准备 …… 175
4.15.2　读取 MySQL 表数据 …… 176
4.15.3　写入结果数据到 MySQL 表 …… 177
4.16　Spark SQL 热点搜索词统计 …… 179
4.16.1　开窗函数的使用 …… 179
4.16.2　热点搜索词统计实现 …… 181
4.17　Spark SQL 搜索引擎每日 UV 统计 …… 184
4.17.1　内置函数的使用 …… 184
4.17.2　搜索引擎每日 UV 统计实现 …… 186
4.18　动手练习 …… 187

第 5 章　用户行为数据实时分析模块开发 …… 189

5.1　Spark Streaming 程序编写 …… 189
5.1.1　Spark Streaming 工作原理 ……189
5.1.2　输入 DStream 和 Receiver……191
5.1.3　第一个 Spark Streaming 程序 ……191
5.2　Spark Streaming 数据源……193
5.2.1　基本数据源 ……193
5.2.2　高级数据源 ……195
5.2.3　自定义数据源 ……196
5.3　DStream 操作 ……199
5.3.1　无状态操作 ……199
5.3.2　状态操作 ……200
5.3.3　窗口操作 ……202
5.3.4　输出操作 ……203
5.3.5　缓存及持久化 ……205
5.3.6　检查点 ……205
5.4　Spark Streaming 按批次累加单词数量……207
5.4.1　编写应用程序 ……207
5.4.2　运行应用程序 ……209
5.4.3　查看 Spark WebUI ……210
5.5　Spark Streaming 整合 Kafka 计算实时单词数量……211
5.5.1　整合原理 ……212
5.5.2　编写应用程序 ……213
5.5.3　运行应用程序 ……216
5.6　Structured Streaming 快速实时单词计数……217
5.7　Structured Streaming 编程模型 ……220
5.8　Structured Streaming 查询输出 ……221
5.8.1　输出模式 ……222
5.8.2　外部存储系统与检查点 ……223
5.9　Structured Streaming 窗口操作 ……224
5.9.1　事件时间 ……225
5.9.2　窗口聚合单词计数 ……226
5.9.3　延迟数据和水印 ……229

5.10 Structured Streaming 消费 Kafka 数据实现单词计数 …… 232
5.11 Structured Streaming 输出计算结果到 MySQL …… 235
5.11.1 MySQL 建库、建表 …… 235
5.11.2 Structured Streaming 应用程序的编写 …… 236
5.11.3 打包与提交 Structured Streaming 应用程序 …… 239
5.12 动手练习 …… 242

第 6 章 数据可视化模块开发 …… 244

6.1 IDEA 搭建基于 SpringBoot 的 Web 项目 …… 244
6.1.1 创建 Maven 项目 …… 245
6.1.2 项目集成 SpringBoot …… 246
6.2 WebSocket 数据实时推送 …… 249
6.2.1 WebSocket 推送原理 …… 249
6.2.2 项目集成 WebSocket …… 249
6.2.3 创建 JDBC 查询工具类 …… 250
6.2.4 创建 WebSocket 服务处理类 …… 251
6.3 使用 ECharts 进行前端视图展示 …… 253
6.4 多框架整合实时分析用户行为日志数据流 …… 258
6.4.1 项目实时处理工作流程 …… 258
6.4.2 模拟实时产生用户行为数据 …… 259
6.4.3 集群数据流转 …… 261
6.5 动手练习 …… 264

第 1 章

项目需求描述

本章主要对本书所要实现的项目需求进行描述，让读者了解项目包括的主要功能、项目所使用的技术架构、项目的集群规划等。读者在项目开发前，应仔细阅读本章内容，对项目的整体规划有一个全面的了解，便于后续的项目开发。

本章目标：

- 了解项目的需求设计
- 了解项目的数据流设计
- 了解项目的架构设计
- 了解项目的集群角色规划
- 掌握大数据技术的生态系统架构

1.1 项 目 需 求

随着互联网的迅速发展，Web系统在满足大量用户访问的同时，几乎每天都在产生大量的用户行为数据（用户在使用系统时通过点击、浏览等行为产生的日志数据）及业务交互数据。通过对这些行为数据进行分析可以获取用户的浏览行为，挖掘数据中的潜在价值，从而更好地、有针对性地进行系统的运营。然而随着日志数据每天上百吉字节地增长，传统的单机处理架构已经不能满足需求，此时就需要使用大数据技术并行计算来解决。

本书通过“用户搜索行为分析系统”项目从0到1、手把手讲解如何使用大数据技术对搜索引擎中的海量用户搜索日志数据进行用户行为分析，最终实现以下需求：

- 实时统计前 10 名流量最高的搜索词。
- 使用可视化图表实时展示统计结果。
- 统计一天中上网用户最多的时间段。
- 统计用户访问最多的前 10 个网站域名。
- 分析链接排名与用户点击的相关性。

- 统计每天搜索数量前 3 名的搜索词（热点搜索词统计）。
- 搜索引擎每日 UV（Unique Visitor，独立访客）统计。

对于实时统计，最终将使用柱形图以可视化的形式在浏览器中实时动态展示并排名，展示效果如图1-1所示。

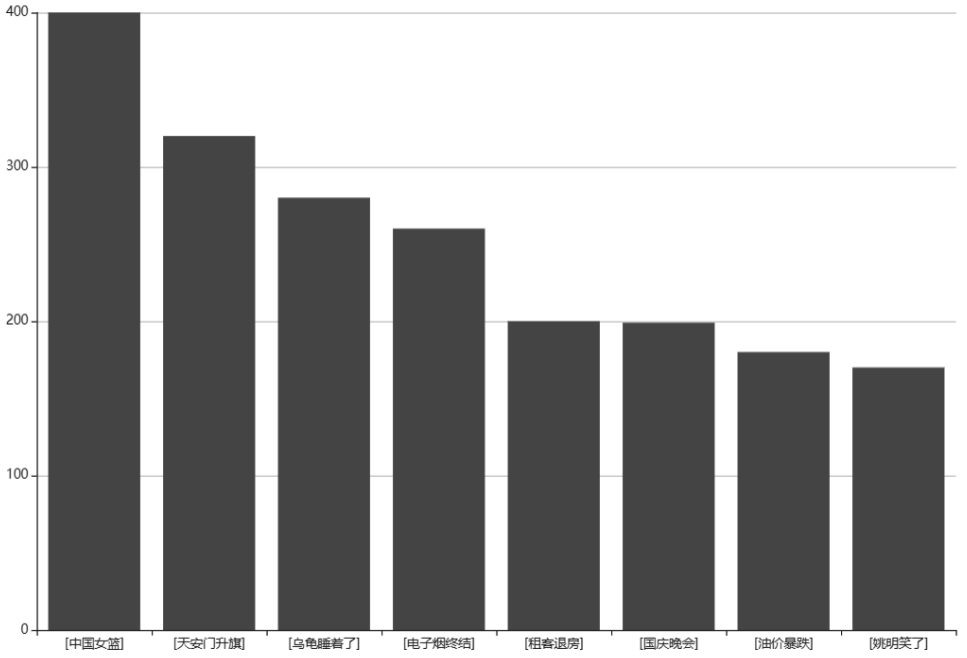


图 1-1　可视化柱形图实时展示搜索词访问量

图1-1柱形图的横坐标表示用户搜索的关键词，纵坐标表示关键词对应的搜索访问数量。

1.2　项目数据流设计

为了实现项目需求，我们需要搭建大数据分析系统，对海量数据进行分析与计算。项目需求涉及离线计算和实时计算，由于Spark是一个基于内存计算的快速通用计算系统，既拥有离线计算组件又拥有实时计算组件，因此以Spark为核心进行数据分析会更加容易，且易于维护。整个系统数据流架构的设计如图1-2所示。

从图1-2可以看出，本系统整体上共分为三个模块：数据采集模块、数据分析模块、数据可视化模块。数据分析模块又分为离线分析模块和实时分析模块，这两个模块在本书中都会进行讲解，并且重点讲解实时分析模块。因此，本书的讲解思路及顺序如下：

（1）数据采集模块
（2）数据离线分析模块
（3）数据实时分析模块
（4）数据可视化模块

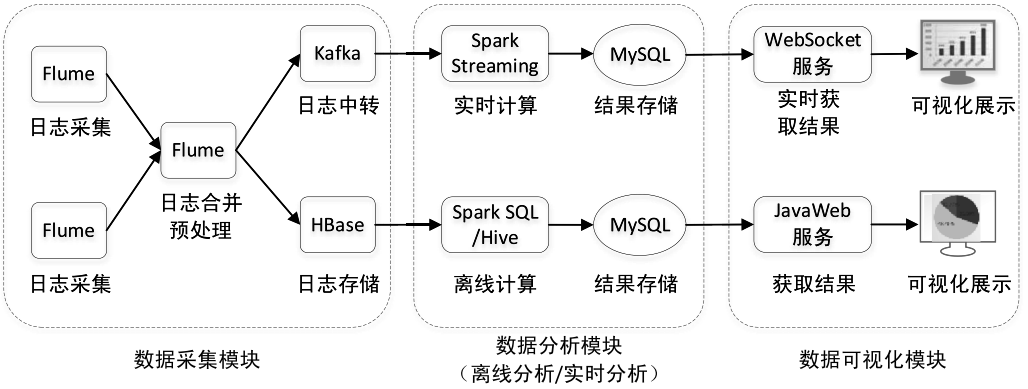


图 1-2　日志分析系统数据流架构设计

读者可按照上述顺序并参考本书的章节逐步进行项目的开发，直到完成并成功运行整个项目。图1-2所示的日志数据流转步骤如下：

01 在产生日志的每台服务器上安装Flume进行日志采集，然后把各自采集到的日志数据发送给同一个Flume服务器进行日志的合并。

02 将合并后的日志数据以副本的方式分成两路（两路数据相同）：一路进行实时计算，另一路进行离线计算。将需要实时计算的数据发送到实时消息系统Kafka进行中转，将需要离线计算的数据存储到HBase分布式数据库中。

03 使用Spark Streaming作为Kafka的消费者，按批次从Kafka中获取数据进行实时计算，并将计算结果存储于MySQL关系数据库中。

04 使用Spark SQL（或Hive）查询HBase中的日志数据进行离线计算，并将计算结果存储于MySQL关系数据库中。通常的做法是使用两个关系数据库分别存储实时和离线的计算结果。

05 使用WebSocket实时获取MySQL中的数据，然后通过可视化组件（ECharts等）进行实时展示。

06 当用户在前端页面单击以获取离线计算结果时，使用Java Web获取MySQL中的结果数据，然后通过可视化组件（ECharts等）进行展示。

1.3　项目架构设计

本节对“用户搜索行为分析系统”项目中所使用的大数据框架Hadoop、ZooKeeper、Kafka、Flume、HBase、Spark等进行简要介绍，为后续的系统搭建打下理论基础。框架的详细介绍读者可查阅笔者的《Hadoop 3.x大数据开发实战（视频教学版）》一书。

Apache Hadoop是大数据开发所使用的一个核心框架，是一个允许使用简单编程模型跨计算机集群分布式处理大型数据集的系统。使用Hadoop可以方便地管理分布式集群，将海量数据分布式地存储在集群中，并使用分布式并行程序来处理这些数据。它被设计成从单个服务器扩展到数千台计算机，每台计算机都提供本地计算和存储。Hadoop本身的设计目的是不依靠硬件来提供高可用性，而是在应用层检测和处理故障。

随着Hadoop生态系统的成长，出现了越来越多的新项目，这些项目有的需要依赖Hadoop，有的可以独立运行，有的对Hadoop提供了很好的补充。

本书讲解的“用户搜索行为分析系统”项目的架构设计如图1-3所示。

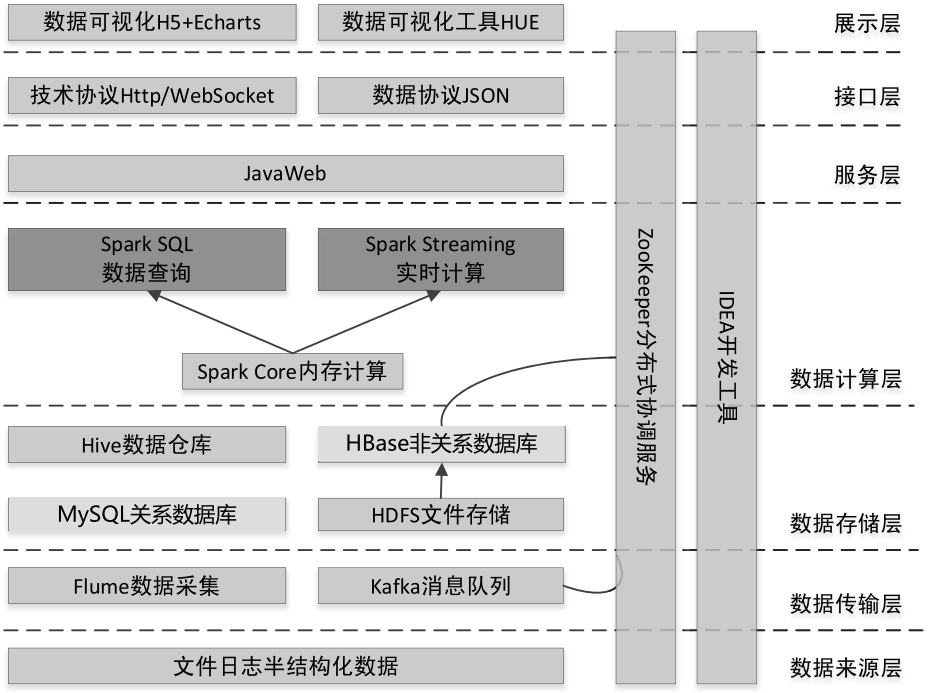


图 1-3　“用户搜索行为分析系统”技术架构

1. 数据来源层

用户在Web网站和手机App中浏览相关信息，服务器端会生成大量的日志文件记录用户的浏览行为。日志文件属于半结构化数据，和普通纯文本相比，半结构化数据具有一定的结构性，有基本固定的结构模式，例如日志文件、XML文档、JSON文档、Email等都属于半结构化数据。

2. 数据传输层

使用Flume+Kafka构建数据传输层。

Apache Flume是一个分布式的、可靠和易用的日志收集系统，用于将大量日志数据从许多不同的源进行收集、聚合，最终移动到一个集中的数据中心进行存储。Flume的使用不仅限于日志数据聚合，由于数据源是可定制的，因此Flume可以用于传输大量数据，包括但不限于网络流量数据、社交媒体生成的数据、电子邮件消息和几乎所有可能的数据源。

Kafka是一个基于ZooKeeper的高吞吐量低延迟的分布式的发布与订阅消息系统，它可以实时处理大量消息数据以满足各种需求。即便使用非常普通的硬件，Kafka每秒也可以处理数百万条消息，其延迟最低只有几毫秒。

为了使Flume收集的数据和下游系统之间解耦合，保证数据传输的低延迟，采用Kafka作为消息中间件进行日志的中转。

3. 数据存储层

大数据项目使用HDFS、HBase、Hive和MySQL构成数据存储层。

HDFS（Hadoop Distributed File System）是Hadoop项目的核心子项目，在大数据开发中通过分布式计算对海量数据进行存储与管理。它基于流数据模式访问和处理超大文件的需求而开发，可以运行在廉价的商用服务器上，为海量数据提供了不惧故障的存储方法，进而为超大数据集的应用处理带来了很多便利。

HBase是一个分布式的、非关系型的列式数据库，数据存储于分布式文件系统HDFS中，并且使用ZooKeeper作为协调服务。HDFS为HBase提供了高可靠性的底层存储支持，ZooKeeper则为HBase提供了稳定的服务和失效恢复机制。

HBase的设计目的是处理非常庞大的表，甚至可以使用普通计算机处理超过10亿行的、由数百万列组成的表的数据。

Hive是一个基于Hadoop的数据仓库架构，使用SQL语句读、写和管理大型分布式数据集。Hive可以将SQL语句转换为MapReduce（或Apache Spark和Apache Tez）任务执行，大大降低了Hadoop的使用门槛，减少了开发MapReduce程序的时间成本。

我们可以将Hive理解为一个客户端工具，它提供了一种类SQL查询语言，称为HiveQL。这使得Hive十分适合数据仓库的统计分析，能够轻松使用HiveQL开启数据仓库任务，如提取/转换/加载（ETL）、分析报告和数据分析。Hive不仅可以分析HDFS文件系统中的数据，也可以分析其他存储系统（例如HBase）中的数据。

MySQL是最好的关系数据库管理系统应用软件。本书使用MySQL存储用户行为数据的计算结果，便于后续通过Web应用等轻量级框架读取结果进行用户端展示。

4. 数据计算层

数据计算层的核心是Spark计算引擎。

Spark是一个快速通用的集群计算系统。它提供了Java、Scala、Python和R的高级API，以及一个支持通用的执行图计算的优化引擎。它还支持一组丰富的高级工具，包括使用SQL进行结构化数据处理的Spark SQL、用于机器学习的MLlib、用于图处理的GraphX，以及用于实时流处理的Spark Streaming。

Spark 的核心（Spark Core）是一个对由很多计算任务组成的、运行在多个工作机器或者一个计算机集群上的应用进行调度、分发以及监控的计算引擎。在Spark Core的基础上，Spark提供了一系列面向不同应用需求的组件，例如Spark SQL和Spark Streaming。

Spark SQL和Spark Streaming都属于Spark系统的组件，它们都依赖于底层的Spark Core。Spark SQL可结合HBase进行数据的查询与分析，Spark Streaming可以进行实时流数据的处理。这两种不同的处理方式可以在同一应用中无缝使用，大大降低了开发和维护的人力成本。

5. 服务层

使用JavaWeb构建系统顶层服务，方便用户通过浏览器访问系统、查看分析结果等。将数据计算层的分析结果存储于关系数据库MySQL中，JavaWeb程序只需读取MySQL中的结果数据进行展示即可。

6. 接口层

服务层使用JavaWeb获取结果数据后，需要使用前端技术展示在浏览器的网页中，便于用户查看。若需要实时展示最新结果数据，则需要使用WebSocket技术。WebSocket是基于TCP的一种新的网络协议，它可以使客户端和服务器之间的数据交换变得更加简单，允许服务器端主动向客户端推送数据。在WebSocket API中，浏览器和服务器只需要完成一次握手，两者之间就可以直接创建持久性的连接，并进行双向数据传输，从而实现数据的前端实时展示。

7. 展示层

为了让结果数据在浏览器网页中的展示更加直观、易于理解，往往需要使用图表进行展示，例如柱形图、饼形图等。HTML5对图表的支持非常友好，因此可以借助目前比较流行的HTML5图表组件ECharts实现数据图表化。

ECharts是一个使用 JavaScript 实现的开源可视化库，可以流畅地运行在个人计算机（PC）和移动设备上，兼容当前绝大部分浏览器（IE9/10/11、Chrome、Firefox、Safari等），提供直观、交互丰富、可高度个性化定制的数据可视化图表。

1.4　集群角色规划

本书项目使用3个节点（主机名分别为centos01、centos02、centos03）搭建大数据集群，各节点的角色（守护进程或框架名称）规划如表1-1所示。

表 1-1　集群角色规划

集　　群	centos01 节点	centos02 节点	centos03 节点
HDFS	NameNode SecondaryNameNode DataNode	DataNode	
YARN	ResourceManager NodeManager	NodeManager	NodeManager
ZooKeeper	QuorumPeerMain	QuorumPeerMain	QuorumPeerMain
HBase	HMaster		
HRegionServer	HRegionServer	HRegionServer	
Hive	Hive		
Kafka	Kafka	Kafka	Kafka
Flume	Flume	Flume	Flume
Spark	Master	Worker	Worker

centos01节点为集群主节点，各个框架的主进程安装在该节点上，其他进程则安装在centos02和centos03节点上。

在后续章节会对各个框架的安装及相关守护进程进行详细讲解。

1.5　项目开发环境介绍

本书项目所使用的开发环境及版本介绍如表1-2所示。

表 1-2　项目开发环境

软　　件	版　　本	作　　用
VMware Workstation	12.5.2	桌面虚拟计算机软件，提供用户可在单一的桌面上同时运行不同的操作系统和进行开发、测试、部署新的应用程序的最佳解决方案。 本书使用 VMware 创建虚拟机，安装 Linux 操作系统，搭建大数据集群
CentOS 操作系统	7.3	CentOS 7 是一个企业级的 Linux 发行版本。本书用 CentOS 7 操作系统搭建大数据集群，部署大数据应用程序
JDK	1.8	Hadoop 等很多大数据框架使用 Java 开发，依赖于 Java 环境，因此需要安装 Java 环境 JDK
Hadoop	3.3.1	Hadoop 是大数据开发所使用的一个核心框架，是一个跨计算机集群分布式处理大型数据集的系统。使用 Hadoop 可以方便地管理分布式集群，将海量数据分布式地存储在集群中，并使用分布式并行程序来处理这些数据。 本书的离线数据分析模块所使用的框架 HBase 和 Hive 都依赖于 Hadoop
Flume	1.9.0	Flume 是一个分布式的、可靠和易用的日志收集系统，用于将大量日志数据从许多不同的源进行收集、聚合，最终移动到一个集中的数据中心进行存储。 本书使用 Flume 进行用户行为日志数据的实时采集
Kafka	3.1.0	Kafka 是一个高吞吐量低延迟的分布式的发布与订阅消息系统，它可以实时处理大量消息数据以满足各种需求。Kafka 每秒也可以处理数百万条消息，其延迟最低只有几毫秒。简单来说，Kafka 是消息中间件的一种。 本书使用 Kafka 进行用户行为数据的实时中转，充当数据中转站的角色
ZooKeeper	3.6.3	ZooKeeper 是一个分布式应用程序协调服务，主要用于解决分布式集群中应用系统的一致性问题。 本书使用 ZooKeeper 协调集群服务，进行集群状态一致性管理。Kafka、HBase 都依赖于 ZooKeeper
HBase	2.4.9	HBase 是一个分布式的非关系型的列式数据库。HBase 数据存储于分布式文件系统 HDFS，并且使用 ZooKeeper 作为协调服务。 HBase 的设计目的是处理非常庞大的表，甚至可以使用普通计算机处理超过 10 亿行的、由数百万列组成的表的数据。 本书使用 HBase 存储数以亿计的用户行为日志数据，用于进行离线统计分析

（续表）

软　件	版　本	作　用
Hive	2.3.3	Hive 是一个基于 Hadoop 的数据仓库工具，提供了一种类 SQL 查询语言，称为 HiveQL。这使得 Hive 十分适合数据仓库的统计分析。Hive 不仅可以分析 HDFS 文件系统中的数据，也可以分析其他存储系统（例如 HBase）中的数据。 本书使用 Hive 进行用户行为数据的离线分析
Spark	3.2.1	Spark 是一个快速通用的集群计算系统。它支持一组丰富的高级工具，例如使用 SQL 进行结构化数据处理的 Spark SQL、用于实时流处理的 Spark Streaming。 本书使用 Spark SQL 进行用户行为数据的离线分析。当然也可以整合将 Spark SQL 与 Hive 进行数据的离线分析。多种方式本书都会讲解
MySQL	8.0	MySQL 是一种关系数据库管理系统，通过将数据保存在不同的表中提高数据的存储速度与灵活度。在 Web 应用方面，MySQL 是最好的关系数据库管理系统应用软件，本书使用 MySQL 存储用户行为数据的计算结果，便于后续通过 Web 应用等轻量级框架读取结果进行用户端展示
IntelliJ IDEA	2021.3.3	IDEA 在业界被公认为最好的 Java 开发工具之一。其功能强大，符合人体工程学的设计，让开发不仅高效，更成为一种享受。 本书使用 IDEA 开发 Spark 应用程序、Web 应用程序，并对应用程序进行打包等