按照大数据技术流程，由浅入深，逐步引导掌握大数据技术开发
理解 Hadoop 基本概念、分布式存储技术与分布式计算技术
掌握分布式文件系统 HDFS 和分布式并行计算框架 MapReduce

Hadoop 大数据开发实战

HADOOP BIG DATA PRACTICE

杨力 编著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Hadoop
大数据开发实战

HADOOP BIG DATA PRACTICE

杨力◉编著

人民邮电出版社
北京

图书在版编目（CIP）数据

Hadoop大数据开发实战 / 杨力编著. -- 北京 : 人民邮电出版社, 2019.3
ISBN 978-7-115-50217-9

Ⅰ. ①H… Ⅱ. ①杨… Ⅲ. ①数据处理软件－程序设计 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第283129号

内 容 提 要

本书将大数据技术生态圈主流技术框架的应用与发展、搭建 Hadoop 大数据分布式系统集群平台、大数据分布式文件系统 HDFS、大数据分布式并行计算框架 MapReduce、大数据汽车销售数据统计分析项目 5 大模块分为 11 章内容进行阐述。具体分布情况如下：第 1 章是大数据概论，介绍大数据的发展背景及基本概念；第 2 章是搭建 Hadoop 分布式集群；第 3～6 章是 HDFS 分布式文件系统入门、HDFS 接口、HDFS 的运行机制、Hadoop I/O 流操作；第 7～10 章是初识 MapReduce 编程模型、MapReduce 应用编程开发、MapReduce 编程案例、MapReduce 运行机制与 YARN 平台；第 11 章是汽车销售数据统计分析项目实战。本书将理论与实践相结合，介绍了大数据的核心技术，并通过介绍一个企业的开发项目，深入讲解大数据技术在实际工作中的应用。

本书是为所有热爱大数据、打算从事大数据相关工作的读者而编写的，适合有 Java 编程基础的学习者参考使用，也适合作为高等院校、培训机构的大数据技术教材。

◆ 编　　著　杨　力
责任编辑　刘　博
责任印制　陈　犇

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市中晟雅豪印务有限公司印刷

◆ 开本：787×1092　1/16
印张：14.75　　2019 年 3 月第 1 版
字数：381 千字　　2019 年 3 月河北第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

前　言

随着信息技术的发展，以及互联网、移动互联网、可穿戴式互联网时代的来临，数据爆炸式地产生。据统计，近几年人类产生的数据，比人类自有文字记载以来产生的所有数据的总和还要多，而且数据还在以惊人的速度增长着。

过去，各个企业都积累了大量丰富的数据，于是购买服务器来存储这些数据，企业面对不断增长的数据，开始思考：除了需要不断购买服务器，花巨大的硬件成本来存储这些数据，我们能从这些持续不断积累下来的数据中得到什么呢？怎样去挖掘和利用这些数据呢？就在这样一个境遇下，一个全新的技术进入了大众的视野，它提出了海量数据可以分布式存储在成本较低的商用服务器上，并且这些海量数据可以分布式地得到计算处理，这个技术称为大数据技术。本书将要介绍的大数据相关技术，可以帮助企业解决不断增长的海量数据的存储问题和计算处理问题；帮助企业从数据中获取经验，并得到巨大的潜在商业价值。

通过本书的学习，读者将对大数据技术有一个深刻的认识，并且掌握大数据技术中最核心的数据分布式存储系统 HDFS 和数据分布式并行计算框架 MapReduce；再通过对大数据项目案例的开发学习，对大数据技术应用进行训练。

本书共 11 章，第 1 ~ 2 章主要介绍了大数据的背景、大数据的学习基础、大数据的行业案例、大数据技术生态圈以及 Hadoop 的搭建，阅读这部分内容，读者将对大数据及其相关技术有一个全方位的宏观认识；第 3 ~ 6 章主要介绍了大数据存储分布式文件系统 HDFS，通过对这部分内容的学习，读者将学习分布式存储的核心原理，分布式文件系统 HDFS 的操作接口、运行机制及 I/O 操作；第 7 ~ 10 章主要介绍了大数据分布式计算处理框架 MapReduce，通过对这部分内容的学习，读者将理解 MapReduce 编程模型及应用、MapReduce 在 YARN 资源管理平台上的运行机制；第 11 章通过一个企业级的项目，带读者体验大数据技术的应用场景。全书按照大数据的技术流程，由浅入深，逐步引导读者掌握大数据技术的开发。

本书适用于对大数据技术感兴趣的读者。全书的编写力求内容科学准确、系统完整、通俗易懂，让初学者能快速掌握大数据技术，同时对专家级读者也具有一定的参考价值。

感谢曾经和我一起奋战在大数据一线的马延辉、唐刚、游大海、赵明栋、郑思成。最后，特别感谢我的父亲、母亲、岳父、岳母及我的妻子，你们的全力支持才使我能够顺利完成本书。

由于编者水平有限，书中难免出现疏漏和不足，敬请读者批评指正。

编　者

2018 年 8 月

目 录

第 1 章 大数据概论

本章要点

- 大数据的学习基础
- 大数据的背景
- 对大数据的不同认识
- 大数据的行业案例
- 大数据的基本概念
- 大数据技术生态圈

本章将为大家解答以下问题：学习大数据之前应该具备哪些基础知识？大数据出现的时代背景是怎样的？大数据为什么产生？各行业人员对大数据的定义是什么？大数据有哪些实际应用场景？大数据有哪些基本的概念？大数据技术生态圈有哪些常见的应用技术？

1.1 大数据的学习基础

恭喜您，已经迈出学习大数据的第一步，相信通过您的努力，在不久的将来一定会在大数据领域有所成就。

学习大数据之前，读者先要了解一些基础知识，如果这些基础知识掌握得熟练、牢固和深刻，那么将在后续的大数据学习过程中感到得心应手，也会越来越喜欢钻研和探索层出不穷的大数据新技术，为大数据的后续学习奠定坚实可靠的基础。可以说，这些基础知识的掌握程度，直接关乎是否能够坚持学习大数据。

目前，大数据技术领域 80%以上的技术都是运用的 Java 语言。Java 语言自 1995 年诞生之初就备受青睐，后以迅猛之势发展，现已成为编程者的必备技能之一。今天，虽然计算机领域已有几百种编程语言，但 Java 语言依然充满了生命力。

从结构上来看，Java 语言有 3 大模块。

（1）Java 语言第 1 个模块是 Java Standard Edition（Java SE），也就是 Java 标准版，它是 Java 语言最重要、最关键、最能体现 Java 语言编程能力的模块。Java SE 是学习 Java 语言编程开发的第一步，包含 Java 语言的编译运行环境 JDK（Java Developer Kit）、Java 基本数据类型、流程控制、面向对象、I/O 流、网络编程、多线程、反射机制、泛型等非常重要的基础开发知识。

（2）Java 语言第 2 个模块是 Java Enterprise Edition（Java EE），也就是 Java 企业版，也称为 Java Web。它是在 Java SE 的基础上构建起来的基于互联网 Web 应用程序开发的一门语言。Web

应用从 Web 1.0 到 Web 2.0 得到了飞速的发展，Java Web 功不可没，它包含的技术有 HTML、CSS、JavaScript、JQuery、JSP 开发、Servlet 开发、Tomcat 服务器、Struts2、Hibernate、MyBatis、Spring 和 Spring MVC 等，这些都是 Web 开发的主流技术，熟练掌握它们，对大数据技术的学习大有帮助，也有助于大数据可视化、大数据文件系统中的 Web 接口模块等的学习。

（3）Java 语言第 3 个模块是 Java Micro Edition（Java ME），也就是 Java 微缩版，它适合做一些微型平台上的开发。例如，2G 手机中的知名游戏“贪吃蛇”就是用 Java ME 版本开发的。Java ME 也是在 Java SE 的基础上构建的，但后来 Google 发布了一款基于移动平台终端的操作系统——Android 系统，Java ME 因此退出了舞台。

总之，学习大数据技术，一定要先掌握一门操作大数据技术的利器，这个利器就是一门编程语言，比如 Java、Python、R 等。本书就是以 Java 语言为基础编写的。

具备了 Java SE 和 Java EE 的编程技术之后，还需要掌握一门数据库知识，建议学习 MySQL 数据库，包括基本概念、表的设计、视图、索引、函数、存储过程等。

掌握以上技术后，还需掌握一门操作系统技术，那就是在服务器领域占主导地位的 Linux 操作系统，只要能够熟练使用 Linux 常用系统命令、文件操作命令和一些基本的 Linux Shell 编程即可。大数据处理的数据是业务系统服务器产生的海量日志数据信息，这些数据都是存储在服务器端的数据，人们常用的操作系统就是在实际工作中安全性和稳定性都很高的 Linux 或 UNIX 操作系统。大数据 Hadoop 本身提供了 Linux 版本和 Windows 版本。由于数据一般存储在服务器端，因此我们学习大数据也是选择 Linux 版本的 Hadoop，大家学会了 Linux 版本，那么 Windows 版本基本也就掌握了。

1.2 大数据的背景

在讲解“大数据”定义前，首先我们要理解什么是数据。

你用手机发了一条朋友圈，想让大家为你点赞，此时就产生了数据。

你用百度搜索了关键词，找到了想要的结果，此时就产生了数据。

你的智能手环，告诉你一天走了多少步，此时就产生了数据。

……

这样的情况不仅发生在你一个人身上，而且每天发生在几亿甚至十几亿人的身上。可以想象，现在这个时代产生的数据量是多么惊人！也许你对这些数据不太敏感，但是换个角度，假如你是那些提供互联网服务的公司，那么，就需要考虑这些数据的存储问题了。

1.3 对大数据的不同认识

我们所处的时代，数据以惊人的速度产生，数据的存储设备也在以惊人的速度发展，那么到底什么是大数据？这个问题再一次摆在我们眼前，接下来，看看不同领域的人们对大数据的认识。

1.3.1 资深编程者眼中的大数据

图 1-1 所示的都是公司的 Logo，这些是正在使用大数据技术的公司，如 Google、IBM 等世

界著名企业。编程者最关心的是，目前哪些公司在使用大数据技术？这门技术的应用普遍性如何？值不值得学习这门技术？

图 1-1

计算机存储数据的方式是二进制，海量数据存储在一个大型的计算机集群上，在集群上可以搭建各种数据处理平台，比如后面将要讲的 Flume 海量日志采集平台、Hadoop 分布式文件系统、MapReduce 分布式并行处理计算框架、Hive 数据仓库、Storm 流式计算，HBase 分布式实时数据库、Kafka 消息队列、Spark 内存计算等。利用这些平台，可以对数据进行采集、存储、计算和展示，将二进制数处理成人们能够识别的数字，或者人们视觉能够感受的图片或者视频。但是，在这个处理过程中也会出现各种各样的问题，如资源丢失、节点宕机等。

所以，编程者眼中的大数据，其实就是技术。

1.3.2　营销者和学者眼中的大数据

营销者是站在市场前沿的人，他们负责销售大数据产品和宣传大数据的价值；学者是站在科技前沿进行学术研究的人，比如各大研究机构的科研人员、各大高校的教授专家等。他们认为，大数据有 4 个特征，如图 1-2 所示。

第 1 特征是数据体量（Volume）巨大，大到什么程度呢？PB 级别起步！很多人对 PB 可能没什么概念，那么我们就来换算一下：1024MB= 1GB，1024GB =1TB，1024TB=1PB。

第 2 个特征是数据类型多样（Variety），大数据能支持文本、图像、视频、音频等几乎所有的文件类型的存储。关系型数据库只支持结构化的数据存储，而且关系型数据库存储的数据体量的峰值在 GB 级别。

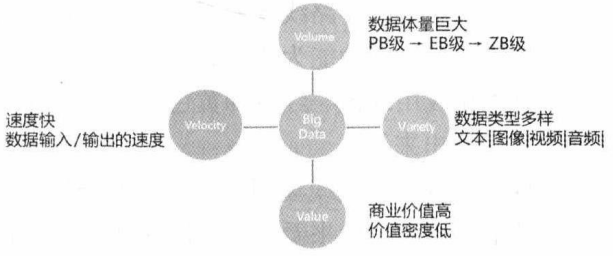

图 1-2

第 3 个特征是商业价值（Value）高，也就是大数据中所蕴含的价值高。

第 4 个特征是速度（Velocity）快，数据输入/输出的速度要快。这也是大数据最核心的一个特征，可以说，如果没有这个特征，就不能称

之为大数据了。从某种意义上讲，前 3 个特征都属于大数据本身的固有特征，只有速度快是大数据技术层面的独有特征。营销者和学者，敏锐地捕捉到了大数据的特征——4 个 V。4 个 V 紧密相连，缺一不可，构成了大数据的初步原型。

1.3.3 商家看大数据

如果买啤酒和尿布这类商品，人们一般会去超市购买。

有一天，美国某沃尔玛分店的数据分析员意外发现，每逢周五，尿布和啤酒的销量便会大大增加，后来他在超市计算机的数据库后台中发现，购买者多为年轻男性。虽然这两种商品似乎“风马牛不相及”，但这名细心的数据分析员在周五进行了现场观察，终于发现了一个秘密。原来这些购买尿布的年轻男性，假日会狂欢玩乐，没时间购买孩子用的东西，所以他们每到周五下班后，会一次买齐孩子周末和下一周使用的尿布，以及聚会时豪饮的啤酒。

原本啤酒在一层摆放，尿布在地下一层摆放。发现这个秘密后，沃尔玛超市及时调整了商店的货品摆放位置，把尿布放在啤酒的旁边卖，这一个小小的位置调整，带来了奇迹，沃尔玛超市的啤酒和尿布的销售业绩增长了十几倍。通过数据分析竟然能发现这么大的潜在商业价值，看来这些数据里藏着很多宝藏，等待着我们去挖掘。自此，超市开始重视积累销售记录数据。

过去，人们不重视数据，因为它们不仅无法为企业创造直接的价值，而且存储数据还要花费很大成本，数据成了企业沉重的包袱。但当我们的思维发生变化后，去挖掘数据，才发现数据的价值极其珍贵。

所以，大数据不仅是技术，是商业价值，它更是一种思维方式。

1.4 大数据的行业案例

前面介绍了学习大数据技术所要具备的基础知识、大数据的背景及不同领域人对大数据的不同认识，本节将通过大数据的行业案例，使读者再一次认识大数据。

1.4.1 电子地图

电子地图，是人们非常熟悉的应用，甚至有的人天天都在使用，如百度地图、高德地图、Google 地图等。基于地图，又涌现出了一个大批优秀的 O2O 应用与服务。利用电子地图，可以导航和获取实时路况信息，可以快速顺利地到达目的地。电子地图已经成为一个公共平台，满足商业和个人的需要。

图 1-3 展示了一个路线规划方案，是从北京的北苑附近驾车到三里屯的行车路线。实际上，电子地图的路线规划功能为我们制定了 3 个行车路线方案，并且将排在第一的路线方案设为推荐路线。推荐路线用绿色、红色和黄色显示，另外两条路线则用灰色表示，并且每条路线的行驶时间都已经估算出来了。

这种习以为常的路线规划和推荐功能是怎么实现的呢？

从功能实现的角度来看，这个路线规划的功能叫实时路况，属于大数据实时计算业务范畴。实时计算业务是对实时性要求很高的数据处理业务。试想一下，如果路况信息做不到实时处理，使用的还是昨天的历史数据，那么它对当前的路况来说还有意义吗？显然没有。

实时路况的底层实现首先需要车辆在行驶过程中产生的 GPS 数据，这些数据可以通过卫星定

位进行采集。注意，在 GPS 数据中有一个很重要的参数值——速度信息。有了速度信息，地图厂商才可以判断某一个路段的拥堵情况。例如，发现某一路段上所有车辆的行驶速度小于 10km/h，在绘制地图的时候就可以用红色表示，告诉使用电子地图的用户，这个路段处于拥堵状态。

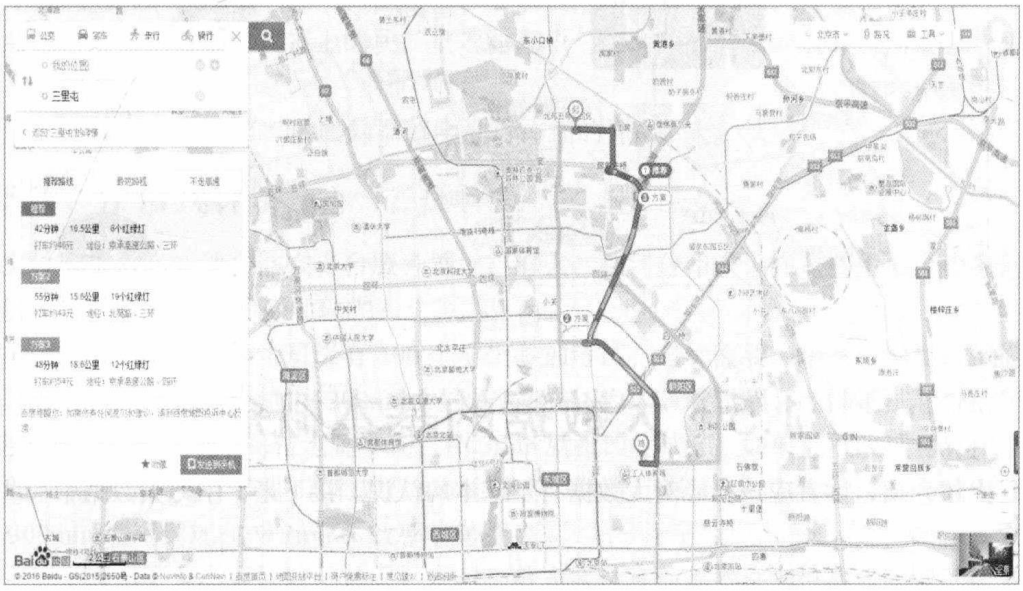

图 1-3

通过各个阶段路况的状态信息及该路段在地图上标注的长度，经过比例尺的计算，就能够计算出实际各个路段长度的总和，再根据各个路段的行驶速度，算出经过各个路程所要花费的时间，从而为使用者寻找一个合适的路线规划。

电子地图实时路况功能的业务实现过程，理解起来并不难，但要真正用大数据技术实现这项功能，就没有那么容易了。如果没有大数据分布式集群的数据处理平台做支撑，单靠传统的数据库技术是做不到上述功能的。在后续的章节中，我们会详细阐述这些技术的实现细节。

1.4.2　电子商务——用户画像

大数据在电子商务平台的应用已经非常成熟，经常在网上购物的你，会发现在电商平台经常会有你喜欢的类似产品的精准推荐。举个例子，最近想购买一双篮球鞋，你在某个电商平台上浏览了很多款式，过一段时间再次打开该电商平台时，你会发现主页上出现了很多你曾经浏览过的篮球鞋或者你喜欢的款式和颜色的篮球鞋，这时你就可以从中挑选一双最喜欢的下单购买了。这里仅简单描述了一下购物场景，但在这背后究竟发生了什么呢？

实际上，该电商平台应用了大数据的用户分析技术，对曾在该平台上浏览或者购买过产品的每个用户信息进行详细分析，如图 1-4 所示。这种将用户的个人信息、家庭信息、喜好信息等进行详细提炼的行为，称为用户画像分析。当然，这些信息都是用户的私密信息，是不对外开放的，但这些信息可以帮助电商平台更好地了解用户，为其提供最好的产品推荐服务，也就是精准营销。正如这个用户购买篮球鞋的案例，如果分析后知道当前用户喜欢的颜色是白色，电商平台就不会为用户推荐黑色或其他颜色的球鞋了。在大数据技术领域，我们可以分析总结出用户的基本信息、购买能力、行为特征、社交网络、心理特征以及兴趣爱好等信息，在绝大多数电商平台中销售额的百分之二十来自大数据电商技术的推荐。

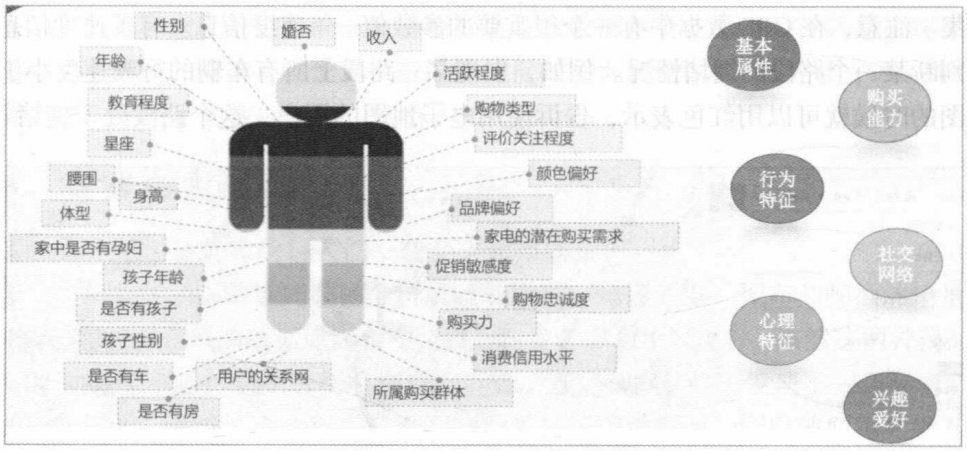

图 1-4

1.5 大数据的基本概念

通过上述介绍，读者应该已经对大数据有了基本的认识。接下来，学习大数据中一些具体的基本概念。

1.5.1 两个核心

大数据的核心技术主要是两大部分内容：一是数据的存储，二是数据的计算（处理）。对于数据的存储和计算处理，传统数据库、数据仓库等产品已经做得非常好了，为什么还要使用大数据技术呢？究其原因，不难发现，传统数据库和数据仓库的底层存储和处理结构采用的是 B+树算法。这种算法有个特性，那就是在数据量不大的时候性能非常好，但当数据量超过某一阈值，此算法的性能就会急剧下降。即使增加服务器扩展集群存储，也不能从根本上解决问题，因为这种解决方案类似于在一个数据库服务器的基础上购买大量磁盘做扩展存储，它不是真正意义上的分布式存储。

这里提到了一个关键词——磁盘。磁盘就是我们经常说的硬盘，也是计算机中常用的一种存储设备。可以向磁盘中存储数据，也可以从磁盘中读取数据。也就是说，处理数据时，是先把数据从磁盘上读到内存中，然后利用 CPU 资源进行计算，从磁盘上读数据的过程称为磁盘寻址。当磁盘中存储了海量数据之后，磁盘寻址的过程将会耗费大量时间。

所以，当数据量非常大时，传统的数据库和数据仓库虽然可以勉强存储，但也很难对这些数据做进一步的统计和分析应用。

大数据要解决的问题就是进行真正意义上的“分布式存储”和“分布式计算”。

1.5.2 分布式存储

“分布式”思想在 20 世纪甚至更早时期就已被提出来了，本节所讲的大数据在架构上并不是一种创新，但要真正实现这个架构并不容易。Google 公司研发出了世界上第一款真正意义上的大数据分布式存储和分布式计算产品，即 Google File System 和 Google MapReduce。

根据分布式思想，当文件数据的体量超过某一台服务器所能够存储的最大容量时，如果要继续存储，则首先根据数据整体规模的大小，以及单台服务器能够存储的最大容量，计算出存储该文件数据需要的服务器总台数，从而实现服务器节点数量的规划；其次将这些规划好的服务器以网络的形式组织起来，变成一个集群，在这个集群上部署一个“分布式文件系统”，统一管理集群

中的各个服务器存储资源；然后，将这个文件数据切分为很多“块”（Block），即计算机操作系统存储文件数据的基本单位，类似于计算机存储数据大小的基本单位字节；最后将这些数据块平均分配到各个服务器节点进行存储，并记录每个块的存放名字及位置信息。

该分布式文件系统提供了统一的操作入口和出口。用户每次访问文件数据的时候，分布式文件系统会临时拼装来自不同机器上的块，呈现给用户一个完整的文件。这样，用户就会感觉自己访问的是一台服务器。

关于分布式存储的细节，后面的章节会进行详细的介绍。

1.5.3　分布式计算

将文件数据分布式地存储在多台服务器上，那么，如何分布式地在这些由多台服务器组成的文件系统上进行数据并行计算处理呢？

举一个简单的例子，一个班级有 100 个学生参加考试，老师需要一个一个地批改他们的试卷并计算其分数，结果花费了将近 300min 才批改完成。为了节省时间，老师把试卷分给年级组的 100 位老师同时批改，结果每位老师平均只用 3min 就批改完成了。如果把批改试卷看作一个作业（Job），该例相当于将这个作业分解成了 100 个任务，并行计算处理，本次批改试卷的完成时间由原来的 300min 缩短到现在的 3min，效率显著提高。

这个例子展示了分布式计算的效果。不过，分布式计算面临着许多挑战：作业的任务如何平均分发到各个节点？计算过程中各个节点上的资源如何统一分配和回收？中间产生的计算结果如何及时地统计汇总？集群服务器计算完成的最终结果如何统一地组织输出？这些令人棘手的问题将在后续的章节中一一得到解答。

1.6　大数据技术生态圈

自然界生态圈和谐统一，为人类提供稳定的自然生态环境。那么，大数据技术生态圈提供了什么呢？首先来看图 1-5，这是一个完整的大数据项目模块设计架构图，要完成图 1-5 所示的各个模块的业务开发，就需要大数据领域中各类技术的支撑，我们把这些为大数据项目提供稳定、安全、可靠的完整技术解决方案的技术总集称为大数据技术生态圈。

图 1-5

此项目模块设计架构自下而上分为 5 个模块，分别介绍如下。

（1）第 1 个模块是数据收集，即考虑数据的种类有哪些，要利用什么样的技术来采集这些数据。数据类型有历史数据/文件、点击流、数据市场、实时日志和数据流等。主流的大数据日志数据采集系统平台有 Flume、kafka、Scribe 和 S-qoop 等。

（2）第 2 个模块是数据存储，其方式有云存储、云数据库、Hadoop 集群、系统管理和自动部署等。从项目的业务角度看，这一块要解决的核心问题是如何存储通过采集平台采集的各种类型的数据。

（3）第 3 个模块是数据分析 BDS、RAS。在大数据领域，对于数据的分析分为两类，一类是离线计算，比如计算电商系统每时每刻产生的历史数据等，这也是目前大数据领域占比最大的一项处理业务；另一类是实时计算，这是相对于离线计算而言的。实时计算的应用，例如实时到账或实时付款这种业务，当业务系统产生数据，大数据平台能够立刻采集、存储并进行计算处理。如今，实时计算的需求越来越多。数据分析领域涌现出了大量优秀的大数据计算框架。离线计算框架有 Hadoop MapReduce 分布式并行计算框架、Hive 分布式数据仓库、Spark-SQL 等；实时计算框架有 HBase 分布式实时数据库、Storm 分布式流式计算框架、Spark-Streaming 等。

（4）第 4 个和第 5 个模块是数据集成 DAG 和数据交易万象。这两个模块侧重于上层的业务处理。经过数据分析处理，会得到不同的结果，将这些结果集根据业务的需求进行组装集成，形成数据网关、开发套件、BI 组件、可视化第三方工具等，为数据交易万象提供服务，形成数据集市层。

然后，用户就可以通过外围的业务系统，根据自己的需要，来这个数据集市上购买需要的数据产品，也就是图 1-4 中的环境数据、运营商数据、征信数据、金融数据、电商数据等。

相信将来会涌现出更多、更优秀的技术框架，大数据的生态圈将会不断更新、不断丰富。

本章总结

本章主要分享了大数据的学习基础、大数据的背景、对大数据的不同认识、大数据的行业案例、大数据的基本概念和大数据技术生态圈，系统全面地剖析了大数据技术的从前、现在和未来，为读者学习大数据技术打下坚实的基础。

本章习题

1. 学习大数据应具备哪些基础知识？
2. 大数据技术生态圈中常见的应用技术有哪些？
3. 简述什么是分布式。
4. 简述什么是分布式存储和计算。

第 2 章 搭建 Hadoop 分布式集群

本章要点

- 云平台
- 安装 CentOS 6
- Linux 系统配置
- Hadoop 的配置部署

在本章，我们将围绕如何搭建 Hadoop 分布式集群环境来讲解大数据技术，这就如同学习开车，得先有一辆车，才能了解车的发动机、变速箱以及方向盘等。

2.1 云平台

2.1.1 了解云平台

读者应该听过阿里云、百度云、京东云等云产品信息，这就是我们常说的云计算，那么大数据和云计算是什么关系呢？实际上，大数据平台软件需要部署在云平台提供的服务器主机上，云计算是大数据的坚实基础。有了云计算，大数据平台才可以稳定、快速地运行。

云产品会为客户提供灵活的服务器主机配置方案。以阿里云为例，客户可以根据自己的需求，在阿里云上选择自己所需的服务器台数，以及每台服务器的配置等。那么，阿里云是如何做到的呢？很简单，阿里云购买了很多台服务器，并在这些服务器上安装了云平台软件，比如 VMware、Docker 等。然后利用这些云平台软件在每一台物理机器上虚拟化出多台虚拟服务器（也叫虚拟机），进而为客户提供灵活的服务器配置，就好像在阿里云可以定制各种各样不同类别的物理主机一样。

也就是说，通过云平台软件，在一台或者多台配置较高的服务器上虚拟化出更多台普通服务器，这种方式和购买多台物理主机的计算、存储性能的效果是完全一样的。

2.1.2 安装 VMware 软件

接下来，请把手里的笔记本电脑看成一台独立的物理主机，我们要在这台独立的物理主机上安装云平台软件 VMware，进而虚拟化出 3 台独立的物理主机，这样就可以搭建 Hadoop 分布式集群环境了。

一台或者两台服务器无法组成集群，集群至少需要 3 台服务器。目前，百度、腾讯等一线互联网公司已经达到了万台集群的规模，携程、去哪儿、苏宁等企业的集群也已有了近千台的规模。而对于我们初学者，搭建一个 3 台服务器的集群就可以了。