高等学校应用型新工科创新人才培养计划指定教材
高等学校云计算与大数据专业“十三五”课改规划教材

“新工科”
配套教材
五大专业集群
教学资源完备

大数据开发与应用

青岛英谷教育科技股份有限公司
山东工商学院 编著

立体化教辅

- 教学PPT
- 教学大纲
- 考试大纲
- 开源代码
- 配套设备
- 在线题库
- 视频讲解

西安电子科技大学出版社
http://www.xduph.com

高等学校应用型新工科创新人才培养计划指定教材

高等学校云计算与大数据专业“十三五”课改规划教材

大数据开发与应用

青岛英谷教育科技股份有限公司

编著

山东工商学院

西安电子科技大学出版社

内 容 简 介

本书系统讲解了目前大数据开发领域的主流技术与实用技能，尤其侧重于对 Hadoop 生态系统的讲解，包括 Hadoop 框架的运作流程、执行原理及数据工具等内容。

全书共分 12 章，分别对大数据概论、Hadoop 集群环境搭建以及 HDFS、MapReduce、ZooKeeper、HBase、Hive、Storm、Sqoop、Kafka、Spark 和 ElasticSearch 的核心知识进行了介绍，同时辅以对各种 API 及实例的深入解析与实践指导，旨在使读者迅速理解并掌握大数据的相关知识框架体系，提高动手能力，熟练使用 Hadoop 集成环境等大数据开发工具，完成大数据相关应用的开发、调试和运行工作。

本书适用面广，可作为高等学校大数据专业、计算机类专业的教材，也可作为大数据从业者、软件开发人员以及程序设计爱好者的参考用书。

图书在版编目(CIP)数据

大数据开发与应用 / 青岛英谷教育科技股份有限公司，山东工商学院编著. —西安：西安电子科技大学出版社，2018.8(2018.9 重印)

ISBN 978-7-5606-5015-9

Ⅰ. ① 大… Ⅱ. ① 青… ② 山… Ⅲ. ① 数据处理—研究 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字(2018)第 163984 号

策　　划　毛红兵
责任编辑　刘炳桢　毛红兵
出版发行　西安电子科技大学出版社(西安市太白南路 2 号)
电　　话　(029)88242885　88201467　　邮　　编　710071
网　　址　www.xduph.com　　电子邮箱　xdupfxb001@163.com
经　　销　新华书店
印刷单位　北京虎彩文化传播有限公司
版　　次　2018 年 8 月第 1 版　2018 年 9 月第 2 次印刷
开　　本　787 毫米×1092 毫米　1/16　印　张 18
字　　数　421 千字
定　　价　50.00 元

ISBN 978-7-5606-5015-9/TP

XDUP 5317001-2

高等学校云计算与大数据专业
“十三五”课改规划教材编委会

❖❖❖ 前　言 ❖❖❖

当今社会是一个高速发展的时代，也是一个数据爆炸的时代，企业内部的经营信息、互联网中的商品物流信息、人与人的交互信息和位置信息等，每时每刻都产生着大量的数据，而这些数据的集合就被称为大数据。如今，大数据已在全球得到广泛应用，包括金融、汽车、餐饮、电信、能源、体能和娱乐等在内的各行各业都已经融入了大数据生态圈当中。

大数据相关产业在世界范围内发展迅猛。美国在 2012 年就开始着手大数据的开发与利用工作，奥巴马政府投入 2 亿美元支持大数据相关产业的发展，并强调大数据会是“未来的石油”，是在国与国的竞争当中具有重要战略意义的资源。2014 年 3 月，“大数据”一词首次写入我国《政府工作报告》，李克强总理在多个场合反复强调：要开发应用好大数据这一基础性战略资源。2015 年 8 月 31 日，国务院正式发布《促进大数据发展行动纲要》，从政府大数据、大数据产业、大数据安全保障体系三个方面提出了未来 5～10 年我国大数据发展的具体目标和任务，是我国大数据行业发展的权威性、纲领性、战略性文件，为我国大数据应用、产业和技术的发展提供了行动指南，标志着我国大数据战略部署的基本确立。

大数据的迅猛发展是数字设备计算能力增长的必然结果，但巨大的数据量也给数据的存储、管理以及分析带来了极大的挑战。一天之中，互联网产生的全部内容可以刻满 1.68 亿张 DVD，发出的邮件有 2940 亿封之多（相当于美国两年的纸质信件数量），发出的社区帖子达 200 万个（相当于《时代》杂志 770 年的文字量）。显然，依赖单个设备处理能力的传统数据处理技术早已无法满足如此大规模数据的存储和处理需求，数据管理方式的变革已呼之欲出。

鉴于此，以 Google 等为代表的一些数据处理公司研发了横向的分布式文件存储、分布式数据处理和分布式数据分析技术，很好地解决了数据爆炸所产生的各种问题，并由此开发了以 Hadoop 为核心的开源大数据处理系统和以 HBase 为核心的开源数据库系统等。这些系统的共同特点是：通过采用计算机节点集群横向扩展数据处理能力，使程序能在集群上并行执行，从而实现了对海量数据的存储、处理和检索。目前，这些系统已成为大数据开发领域的主流技术。

本书共分为 12 章，分别对大数据概论、Hadoop 集群环境搭建以及 HDFS、MapReduce、ZooKeeper、HBase、Hive、Storm、Sqoop、Kafka、Spark 和 ElasticSearch 的核心知识点进行了介绍，并侧重讲解 Hadoop 生态系统的相关理论与实践知识，包括 Hadoop 框架的执行原理、运作流程以及各组成部分的功能等。本书兼顾系统性与实用性，在全面介绍当前大数据开发的主流技术与实用技巧的基础上，通过对各种 API 和实例的讲解、剖析及上机练习，注重实践能力的提升，旨在使读者通过对本书的学习，深入理解和掌握大数据开发的相关知识，并能熟练使用 Hadoop 集成环境完成大数据应用的开

发、调试和运行工作。

本书由青岛英谷教育科技股份有限公司和山东工商学院共同编写，参与本书编写工作的有张伟洋、焦裕朋、侯方超、孟洁、刘鹰子、韩小雨、刘峰吉、金成学、王燕等。本书在编写期间得到了各合作院校专家及一线教师的大力支持。在此，要特别感谢给予我们开发团队大力支持和帮助的领导和同事，感谢合作院校的师生给予我们的支持和鼓励，更要感谢开发团队每一位成员所付出的艰辛劳动。

教材问题反馈

由于水平有限，书中难免有不足之处，欢迎大家批评指正。读者在阅读过程中如发现问题，可通过邮箱(yinggu@121ugrow.com)联系我们，或扫描右侧二维码进行反馈，以期不断完善。

本书编委会

2018 年 3 月

❖❖❖ 目　录 ❖❖❖

第1章 概论

本章目标

- 了解大数据技术的来源和应用领域
- 了解大数据基础设施的基本架构
- 了解大数据开发的概念
- 了解大数据开发的作用
- 了解大数据开发所涉及的主要技术和分类
- 了解大数据开发和大数据分析之间的关系

大数据指规模超出了传统技术工具收集、存储、管理、分析能力的数据集。随着信息技术的发展，越来越多的信息以数据的形式被记录下来，人类正在步入“大数据时代”。为了应对大数据所带来的挑战，把握其中的机遇，大数据处理技术应运而生，其中，大数据开发技术占有非常重要的地位。

本章将带领读者初步了解大数据开发技术，希望在学习完本章内容之后，读者能够对大数据开发的技术版图有一个清晰的认识，从而明确自己的学习目标和学习路径，在后续的学习过程中能够做到心中有数、事半功倍。

1.1 大数据技术简介

学习大数据开发首先要了解大数据技术。本节将从大数据技术的起源、应用领域和基础设施三个方面对大数据技术进行简要介绍。

1.1.1 大数据技术的起源

传统的数据处理技术究竟是如何过渡为大数据处理技术的？搞清楚这个问题，有助于我们更加清晰地理解大数据技术的概念。

大约在 2000 年前后，互联网行业高速发展，全球范围内各种网站、网页数量急剧增加，全球最大的全文搜索引擎公司 Google 不得不考虑如何应对和处理急剧增长的海量数据(爬虫获取的网页、Web 请求日志等)和各种类型的衍生数据(索引、元数据、网页的各种图结构等)。因为，虽然相关计算工作的大部分复杂度都不是太高，但由于数据的量很大，仍然需要非常多的计算资源来支持，但在那个年代，计算资源是非常昂贵的。于是，为了节省开支，Google 开始寻求其他的数据处理方法。

此时，有人提出了一种思路——把计算任务拆分，然后分布到不同的机器上进行，最后再将各个机器的计算结果汇总起来，由此便实现了分布式的并行数据处理。为了实现这个想法，Google 的科学家们开始对如何进行并行计算、如何分配数据以及如何处理失败等技术问题进行研究和探索。最终，Google 公司在对这些研究成果进行总结的基础上，发表了一篇经典的论文——《MapReduce：Simplied Data Processing on Large Clusters》，奠定了现代大数据处理技术的基础。

相应地，Google 还需要设计一套抽象计算模型来实践这种数据处理方式，为了降低使用门槛，这套模型必须能够隐藏并发、容错、数据和均衡负载等方面的细节。MapReduce 计算模型由此诞生。

MapReduce 模型使得大规模的并行计算变得简单，但事实上，MapReduce 对大数据计算的最大贡献并非是它名字中直观显示的 Map 和 Reduce 思想(类似的计算思想在 Lips 等函数式编程语言中早已存在)，而是这个计算框架可以运行在一群廉价的 PC 机上(注意这里的“廉价”是相对于配备有超高频率中央处理器的大型计算机而言的)。

MapReduce 的伟大之处在于向大众普及了工业界对于大数据计算的认识——良好的横向扩展性和容错处理机制。从前，想对更多的数据进行计算，只能通过制造更快的计算

机，而现在只需要不断添加计算节点就能处理等量的数据。从此，大数据计算的主流模式由集中式开始过渡至分布式。

MapReduce、GFS 和 BigTable 是当时 Google 大数据处理技术的三个核心工具，但是，虽然这些工具很强大，其他的公司或个人却无法使用，原因很简单——这些工具都不开源。但在 2006 年，一个叫 Doug Cutting 的人将该技术的开源框架——Hadoop——贡献给了开源社区。初代 Hadoop 中的 MapReduce 和 HDFS 即为 Google 的 MapReduce 和 GFS 的开源实现，而 BigTable 的开源实现则是同样知名的 HBase。自此，大数据处理工具开始逐渐普及，“大数据时代”的历史大幕正式拉开。

1.1.2　大数据应用领域

大数据技术发展迅速，各行各业都在积极探索大数据的应用领域，其中的许多应用已经日趋成熟。本小节简要介绍当前主要的大数据应用领域。

1. 搜索引擎

大数据技术来源于 Google，而 Google 公司的最核心业务就是它的搜索引擎，因此，可以说搜索引擎是大数据最早的成熟应用领域。

一个成熟的商用搜索引擎，需要具备海量的多元数据、高效的索引、复杂的数据处理模型和应对大量并发的能力，因此需要非常强大的计算能力作为支持。如果想要获取足够的计算能力，一种方式是购买大型服务器，但这样做成本非常高。相比之下，大数据技术为搜索引擎提供了廉价的计算资源，大大节省了搜索引擎的开发和维护成本，因此，目前搜索引擎与大数据技术的结合非常紧密。

2. 推荐引擎

推荐引擎需要对海量用户数据进行实时采集、实时处理，并基于复杂的推荐算法生成相应的模型，然后再利用该模型向特定用户群体推荐他们可能感兴趣的内容。

在这个过程中，推荐引擎需要采集、存储和处理大量的复杂数据，计算的复杂度也比较高，更重要的是它必须具备较高的实时性。针对这些需求，大数据技术都能提供较为成熟的解决方案。因此，当前的各大主流推荐引擎也都普遍基于大数据技术开发，如亚马逊的商品推荐、百度的推荐产品组合、豆瓣电台、优酷的猜你喜欢、网易云音乐的私人 FM 等。

3. 分布式爬虫

爬虫技术是一项非常重要的数据采集和处理技术。例如，在搜索引擎技术体系中，海量的网页信息通常需要通过爬虫来获取。随着互联网的高速发展，网页的数量爆炸式增长，爬虫需要面对越来越多的数据，而大数据技术可以高效支持高并发的网页数据爬取作业，还可以根据作业要求对大数据集群进行弹性伸缩，方便管理计算和存储资源，分布式文件系统和非关系型数据库也都非常适合爬虫业务场景的要求。

大名鼎鼎的 Hadoop 项目就是由一个叫做 Nutch 的分布式爬虫项目衍化而来的，可见大数据技术与分布式爬虫技术之间有着非常紧密的联系。

4. 电信和金融

电信运营商拥有多年的数据积累，其中既有诸如财务收入、业务发展量等结构化数据，也有图片、文本、音频、视频等非结构化数据。这些数据来源于移动语音、固定电话、固网接入和无线上网等各种电信业务，以及运营商收集的实体渠道、电子渠道、直销渠道等各种类型渠道的接触信息，涉及公众客户、政企客户和家庭客户。

整体来看，电信行业的大数据应用仍处在探索阶段。目前，国内运营商对大数据的应用主要集中在五个方面：

(1) 网络管理和优化，包括基础设施建设优化和网络运营管理与优化。

(2) 市场与精准营销，包括客户画像、关系链研究、精准营销、实时营销和个性化推荐。

(3) 客户关系管理，包括客服中心优化和客户生命周期管理。

(4) 企业运营管理，包括业务运营监控和经营分析。

(5) 数据商业化，指将数据作为商品来交易，单独营利。

在金融行业，大数据的应用范围较广。典型的案例如花旗银行利用 IBM 沃森电脑为财富管理客户推荐产品；美国银行利用客户点击数据集为客户提供特色服务(如有竞争的信用额度)；招商银行利用客户的刷卡、存取款、电子银行转账、微信评论等行为数据分析顾客可能感兴趣的产品和优惠信息，每周给客户发送针对性广告，等等。

5. 机器学习

人类大脑本身可看做是一台模式分类器，接收各种传感器(眼、耳、鼻、舌、皮肤)输入的信息，加以融合处理后进行正负反馈，在信息塑造神经元结构以及神经元信息处理结构的反复迭代过程中，最终诞生了人类的智能。

作为人工智能的核心技术之一，机器学习模仿了人类大脑的工作方式：计算机程序可以不断从经验中获取知识、学习策略，当再次遇到类似的问题时，能运用经验知识解决问题并积累新的经验。显然，经验越多，越有利于机器学习模型解决问题能力的提升，人工智能也就越聪明。而经验本质上就是数据，数据的量很大时，就需要用大数据技术来处理，因此机器学习离不开大数据技术。

机器学习的重要分支——深度学习技术——为强人工智能的实现提供了一种极大的可能。深度学习是指通过神经网络的方式，基于大量标注后的数据集进行监督学习，通过训练数据的前向拟合和反向传播，迭代训练、获得模型，然后利用模型在生产环节中进行数据推理。深度学习对数据量和计算能力的依赖性极强，一个成熟的深度学习模型需要海量的经验数据，并且要经历长时间的迭代计算去训练，数据越多，训练的时间越长，塑造培养出的神经网络性能就越强。因此，基于大数据技术的深度学习技术成为当前非常流行的机器学习解决方案。

1.1.3 大数据基础设施

大数据基础设施包括进行大数据处理时需要用到的各种资源、载体、硬件、软件和功能，由物理层和平台层两部分组成。

当前，大数据基础设施都是以大数据集群的方式呈现的。所谓集群，是指将一组相互独立的计算机通过网络互相连接，并使用软件使其中的所有计算机协同工作、共同对外提供服务的资源和功能集合。

本书中所提到的“大数据集群”和“大数据平台”是等价的，可以互相替换。

1. 物理层

大数据集群的物理层指为大数据处理提供硬件保障的部分，是数据存储和计算的底层载体。通常来说，大数据物理层由若干台计算机以及支持它们进行通信的网络设备组成。

每台计算机就是基础设施中的一个计算或存储单元，构成这些计算机的各种硬件(如中央处理器、内存、磁盘等)是具体承载计算和存储等各项功能的组件。这些组件可以由真实的硬件设备提供，也可以通过虚拟化技术实现的虚拟设备(当然虚拟化的背后仍然需要底层的硬件提供资源)提供。如果这些设备是虚拟的，则该计算机被称为虚拟机，否则被称为实体机或物理机。

网络通信设备通常包括路由器、交换机等硬件设备，这些设备同样可以通过虚拟化技术实现。当计算机通过网络通信设备连接到一起可以互相通信时，无论它们是否在空间上被组织在一起，实际上就已经形成了一个计算机网络，大数据基础设施的物理层也就形成了。

2. 平台层

平台层即大数据平台，是指部署在物理层之上的框架和软件的集合，用于实现管理集群资源、分配集群任务、承载集群功能、协调集群角色、监控集群状态等功能。在整个大数据技术体系中，大数据平台主要为基础设施提供功能服务，使数据在其内部流动、算法在其中运行、应用在其上实现。

典型的大数据平台架构如图 1-1 所示，其中标明了该平台涉及的所有组件及其相互关系。

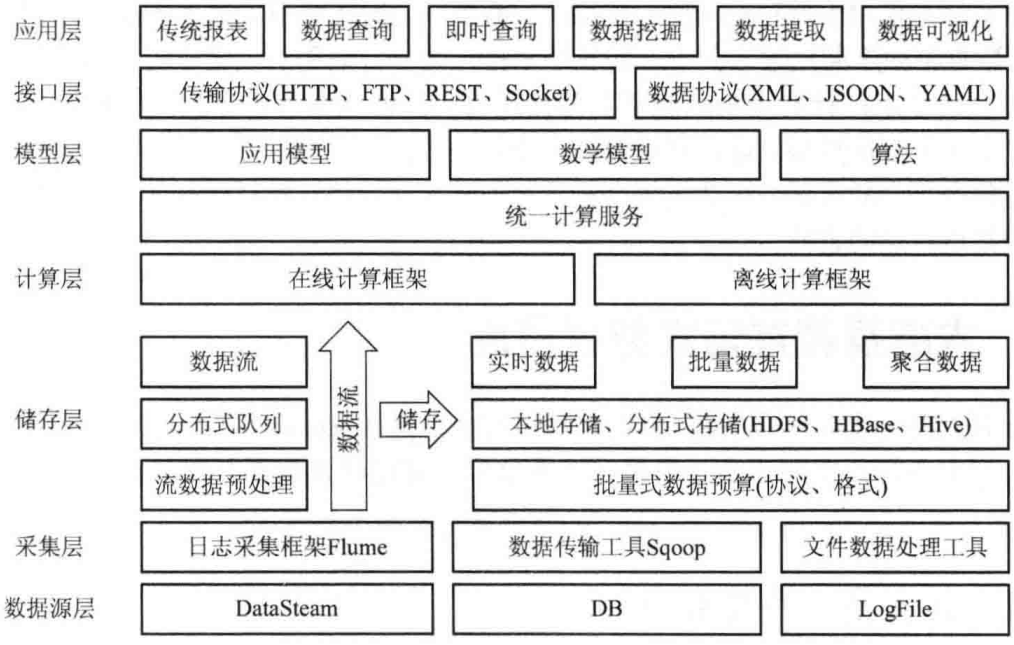

图 1-1　大数据平台架构图

下面按照数据的流动方向，依次介绍大数据平台中各组件的功能：

(1) 数据源层。该层是原始数据存在的地方。这些原始数据往往是在业务过程中产生的(例如程序或设备的日志数据等)，里面既包含了有价值的信息，也包含了大量噪音以及错误的脏数据。这些数据或是被储存在第三方的数据库或存储器中，或是被储存在我们自己的设备中，或是在采集到的时候就被直接发送到大数据平台中。

(2) 采集层。用于将数据源中的数据引入平台中。根据不同的数据源类型和业务需求，应选择不同的工具进行采集。例如爬虫用于采集网页数据，日志采集工具用于采集日志，数据传输工具用于从数据库中获取数据等。

(3) 储存层。用于对平台的存储资源进行管理，并将数据储存在大数据平台中。在这一层中，通常要至少包含一个缓存器(消息队列)和一个存储器(分布式文件系统、分布式数据库或数据仓库)，同时进行数据的初步清洗工作，主要作用是去除明显错误或没有价值的数据。

(4) 计算层。完成数值计算任务。当前大数据计算的模式主要分为流式计算(也称在线计算)和批量计算(也称为离线计算)两种。在流式计算中，数据是流动着的，因此无法确定某一组数据会在何时到来，也无法在计算过程中将数据储存到硬盘中或者进行二次读取；而批量计算要求数据已经被静态地储存在硬盘中，然后才能将数据读取出来进行计算。

(5) 模型层。该层定义了相关应用模型和算法模型的接口，主要用于给计算层提供逻辑支持，完成特定的数值计算。例如数据挖掘算法、机器学习算法等都位于该层。

(6) 接口层。对外提供接入协议，将平台内的数据处理结果向外输出。

(7) 应用层。图 1-1 中的最顶层，主要承载基于大数据平台的各种应用。

事实上，一个大数据平台可以看做是一个容器，它将大量的服务器节点有机地组织起来，形成一个集群，这个集群基于自身的计算与储存的性能优势，承载了两大核心内容：数据和功能。相对应的，大数据平台最核心的部分就是储存层(承载数据)和计算层(承载功能)。

储存层不仅用于保存最重要的资源——数据，同时还承担了数据 I/O 的功能，而数据 I/O 将数据源源不断地输送到平台的各个角落，是平台效率的关键决定因素之一，可谓是整个平台的心脏；计算层则承载了数据处理模块，即从数据中寻找规律和价值的功能模块，决定了平台的核心功能，其运算能力也是大数据平台整体效率的重要决定因素，因此可谓是整个平台的大脑。大数据平台的其他部分都是围绕这两层展开的，或是为这两层服务，或是这两层的延伸。

1.2 大数据技术与大数据开发

在对大数据技术有了一定了解之后，应对大数据开发的内容有一个整体的认识。本节详细介绍大数据开发的概念、作用、技术分类，并比较大数据开发与大数据分析之间的异同。

1.2.1 什么是大数据开发

截至目前为止，我们已经对大数据基础设施、大数据集群、大数据平台等相关概念进

行了介绍，对大数据的应用领域也有了一定了解。在此基础上，我们给出本书中大数据开发的明确定义，即：使用程序语言和大数据技术框架，将与大数据相关的需求实现为一个系统、软件或模块的开发过程。

为了进一步明确这个概念，请注意以下几种情况：

(1) 不使用程序开发语言的，不属于大数据开发的范畴。例如用 Excel 分析数据的过程。

(2) 功能需求与大数据无关的，不属于大数据开发的范畴。例如用一台服务器就可以承载所有功能的需求。

(3) 最终产品并非是一个系统、软件或模块的，不属于大数据开发的范畴。例如最终产品是一份数据分析报告，或使用 Spark Shell 命令行完成的数据处理过程。

(4) 需求被明确前或需求被满足后的工作，不属于大数据开发的范畴。例如大数据平台已经按照需求开发完成，数据分析师利用平台中储存的数据进行算法研发。

值得注意的是，大数据开发是一个完整的系统性工程，应该用整体观念来看待，不能把其中的某项工作单独割裂出来进行界定。例如，操作 Linux Shell 或使用图形界面来部署调试集群、查看日志等工作，虽然不符合上述定义，但却是整个系统性开发工作中不可分割的一部分，因此仍然在大数据开发工作的范畴之中；另一方面，虽然我们试图尽可能清晰地界定大数据开发与其他工作之间的边界，但这个边界仍然是模糊的，需要在实际开发工作中灵活变通，如向 Hadoop 集群中提交一个实现某种数据挖掘功能的 MapReduce 任务，即便该任务与整个平台的耦合性并不强，可以被割裂出来界定为数据挖掘工作，但若被界定为数据开发工作，也并没有明显的不妥。

1.2.2　大数据开发的作用

从 1.1.3 小节中我们得知，大数据基础设施的核心——大数据平台——要包含诸多模块、承载若干功能，本质上可以看做是一个容器。而大数据开发就是对这个容器的实现，也就是将这些模块和功能进行搭建和实现，并保证其正常运行。

准确地说，大数据开发涵盖了图 1-1 中从下到上各层的实现，其中主要的部分是采集层、储存层、计算层、模型层和接口层，核心部分是储存层和计算层。各层中功能模块的技术实现会根据实际业务场景不同而有所变化，但仍然是围绕着储存数据和数值计算这两大核心功能来进行的。因此，大数据开发的作用主要集中在以下几个方面。

1. 资源配置

大数据处理系统面向的是大体量、多来源、多类型的数据。因此，大数据开发需要综合考虑系统资源的合理设计和分配，综合考虑节点数量和角色的分配、硬盘容量和可能的扩展、后台任务和内存空间的分配以及程序设计时内存和并发量等问题。如果这些资源问题没有处理好，会导致整个大数据集群性能和稳定性下降，极端情况下可能会导致集群部分服务异常关闭，甚至整个集群宕机。

2. 数据移动

数据移动问题包括数据从外部流入到平台、数据从平台流出到外部、数据在平台内的