



《**数据科学与大数据技术专业概述**》

|  |  |
| --- | --- |
| **专 业** | **数据科学与大数据** |
| **姓 名** | **杨心怡** |
| **日 期** | **2022/9/7** |

**摘 要**

随着近年来科技的不断发展与互联网行业的持续进步，越来越多的社会资源被网络化和数字化，大数据正在成为互联网技术发展的重要动力，其价值正在得到越来越多的重视。针对近年哈尔滨工业大学响应国家政策最新开设在计算机方向的数据科学与大数据技术专业，本文从专业研究内容，专业主要课程内容及专业主要技术栈，专业学习建议与指导，目前专业的前沿研究方向和专业本科生毕业去向这五个方面对该专业进行了较为全面的介绍，目的在于为外专业学生提供专业概述，同时为本专业新生的专业学习提供一定的指导与建议。

**关键词:** 数据科学与大数据技术 专业概述 互联网

**目 录**

[**第1章 专业研究内容介绍** - 1 -](#_Toc113483410)

[**第2章 专业主要课程内容及技术栈介绍** - 2 -](#_Toc113483411)

[2.1 专业主要课程介绍 - 2 -](#_Toc113483412)

[2.1.1 大数据计算基础 - 2 -](#_Toc113483413)

[2.1.2 大数据分析 - 2 -](#_Toc113483414)

[2.1.3 数据挖掘 - 3 -](#_Toc113483415)

[2.2 大数据常用技术栈介绍 - 3 -](#_Toc113483416)

[2.2.1 数据采集和传输层 - 4 -](#_Toc113483417)

[2.2.2 数据存储层 - 4 -](#_Toc113483418)

[2.2.3 数据分析层 - 4 -](#_Toc113483419)

[2.2.4 资源管理层 - 5 -](#_Toc113483420)

[2.2.5 工作流调度器 - 5 -](#_Toc113483421)

[**第3章 专业学习指导和建议** - 6 -](#_Toc113483422)

[3.1 计算机基础知识 - 6 -](#_Toc113483423)

[3.2 大数据框架技术 - 6 -](#_Toc113483424)

[3.3 项目实践 - 6 -](#_Toc113483425)

[**第4章 专业前沿研究方向** - 7 -](#_Toc113483426)

[4.1 可视化大数据分析 - 7 -](#_Toc113483427)

[4.2 AI - 7 -](#_Toc113483428)

[4.3 跨学科领域交叉的数据融合分析与应用 - 7 -](#_Toc113483429)

[4.4 大数据安全与隐私 - 7 -](#_Toc113483430)

[4.5 大数据治理 - 7 -](#_Toc113483431)

[**第5章 专业本科生毕业去向** - 9 -](#_Toc113483432)

[**结 论** - 10 -](#_Toc113483433)

[**参考文献** - 11 -](#_Toc113483434)

**第1章 专业研究内容介绍**

哈尔滨工业大学数据科学与大数据技术专业隶属于国家重点学科“计算机科学与技术”，于2017年获批，2018年首次招生，在2020年获批省级一流本科专业建设点。

该专业是以大数据为中心，以大数据管理和分析处理的基本理论和方法（即“数据科学”），以及海量、异构、多源数据全生命周期管理驱动的系统设计、开发、实施、评测、优化和运维（即“数据工程”）为基础的新兴学科。

它从科学和技术角度概括了“大数据”这一热点问题。主要研究方向包括大数据管理与分析、大数据治理、数据科学、生物大数据等[1]，期望培养能够综合运用大数据计算和大数据管理基础理论，深度分析与数据挖掘方面的知识，独立解决与大数据相关的复杂工程技术问题，具有大数据获取，建模，管理，分析挖掘与应用等方面的理论知识与工程能力的人才。

**第2章 专业主要课程内容及技术栈介绍**

## 2.1 专业主要课程介绍

哈尔滨工业大学的数据科学与大数据专业开设的本专业学生必选专业核心课有大数据计算基础、大数据分析与数据挖掘。下面将分别对这三门课程进行介绍。

### 2.1.1 大数据计算基础

大数据时代的一大特征即为海量数据的产生，面对极为庞大的数据量，如何存储这些数量高达PB（2^50字节）级甚至更多的数据以及如何运用已有的硬件和软件资源来处理这些数据就成为了一个亟待解决的问题。

大数据计算基础课程的开设为数据科学与大数据专业的学生提供了学习如何解决类该问题的机会。这门课程的内容主要包括了两方面，大数据算法和具体大数据技术。

大数据算法部分主要对外存模型算法、亚线性空间算法、亚线性时间算法、并行模型算法等算法将进行讲解，大数据技术则讲解一些在大数据领域常用的技术，如开源的分布式计算和存储框架Hadoop，大数据分析引擎Spark等。

### 2.1.2 大数据分析

目前，大数据领域每年都会涌现出大量新的技术，成为大数据获取、存储、处理分析或可视化的有效手段。大数据分析技术能够将大规模数据中隐藏的信息和知识挖掘出来，为人类社会经济活动提供依据，提高各个领域的运行效率，甚至整个社会经济的集约化程度。

大数据分析有六个基本方面，分别为：

1. **可视化分析**：数据可视化是数据分析工具最基本的要求。可视化可以直观的展示数据，让数据自己说话，让观众听到结果。
2. **数据挖掘算法**：集群、分割、孤立点等数据挖掘算法可以深入数据内部，挖掘数据的价值。
3. **预测性分析能力**：预测性分析可以人们根据可视化分析和数据挖掘的结果做出一些预测性的判断。
4. **语义引擎**：[非结构化数据](https://baike.baidu.com/item/%E9%9D%9E%E7%BB%93%E6%9E%84%E5%8C%96%E6%95%B0%E6%8D%AE/309808?fromModule=lemma_inlink)的多样性带来了数据分析的新的挑战，需要一系列的工具去解析，提取，分析数据。语义引擎即是被设计成能够从“文档”中智能提取信息的工具。
5. **数据质量和数据管理**：数据质量和数据管理是一些管理方面的最佳实践。通过标准化的流程和工具对数据进行处理可以保证一个预先定义好的高质量的分析结果。
6. **数据存储和数据仓库**：数据仓库是为了便于多维分析和多角度展示数据按特定模式进行存储所建立起来的[关系型数据库](https://baike.baidu.com/item/%E5%85%B3%E7%B3%BB%E5%9E%8B%E6%95%B0%E6%8D%AE%E5%BA%93/8999831?fromModule=lemma_inlink)。

### 2.1.3 数据挖掘

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

数据挖掘涉及的技术方法很多，有多种分类法。根据挖掘任务可分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等等；根据挖掘对象可分为关系[数据库](https://www.zhihu.com/search?q=%E6%95%B0%E6%8D%AE%E5%BA%93&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D" \t "_blank)、[面向对象数据库](https://www.zhihu.com/search?q=%E9%9D%A2%E5%90%91%E5%AF%B9%E8%B1%A1%E6%95%B0%E6%8D%AE%E5%BA%93&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D)、空间数据库、时态数据库、文本数据源、多媒体数据库、异质数据库、遗产数据库以及环球网Web；根据挖掘方法分，可粗分为：机器学习方法、统计方法、神经网络方法和数据库方法。

机器学习中，可细分为归纳学习方法([决策树](https://www.zhihu.com/search?q=%E5%86%B3%E7%AD%96%E6%A0%91&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D)、规则归纳等)、基于范例学习、遗传算法等。统计方法中，可细分为：回归分析(多元回归、自回归等)、[判别分析](https://www.zhihu.com/search?q=%E5%88%A4%E5%88%AB%E5%88%86%E6%9E%90&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D" \t "_blank)(贝叶斯判别、费歇尔判别、[非参数判别](https://www.zhihu.com/search?q=%E9%9D%9E%E5%8F%82%E6%95%B0%E5%88%A4%E5%88%AB&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D)等)、聚类分析(系统聚类、[动态聚类](https://www.zhihu.com/search?q=%E5%8A%A8%E6%80%81%E8%81%9A%E7%B1%BB&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D" \t "_blank)等)、探索性分析(主元分析法、[相关分析法](https://www.zhihu.com/search?q=%E7%9B%B8%E5%85%B3%E5%88%86%E6%9E%90%E6%B3%95&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D)等)等。

神经网络方法中，可细分为：[前向神经网络](https://www.zhihu.com/search?q=%E5%89%8D%E5%90%91%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A847210034%7D)(BP算法等)、自组织神经网络(自组织特征映射、竞争学习等)等。数据库方法主要是多维数据分析或OLAP方法，另外还有面向属性的归纳方法。

## 2.2 大数据常用技术栈介绍

提起大数据，不得不提由IBM提出的关于大数据的5V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性），大数据领域的从业人员的日常工作也与这5V密切相关。

大数据技术在过去的几十年中取得非常迅速的发展，尤以Hadoop和Spark最为突出，已构建起庞大的技术生态体系圈[2]。

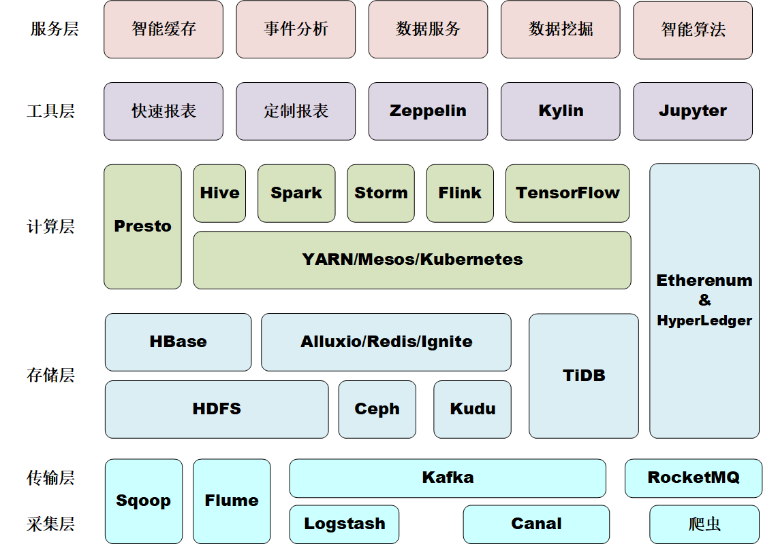


图1.大数据技术栈

下面分不同层次介绍每个层次及一些常用的技术栈（层次并非严格划分，一些技术可以运用于多个层次）。

### 2.2.1 数据采集和传输层

数据采集通常是指将我们需要的数据从外界收集到我们大数据平台上并保存下来。数据传输则是数据流通的管道，如果管道出了问题，前端看到数据的及时性和正确性就会受到影响，因此必须有人对管道的通畅进行管理和维护[3]。

在数据采集与传输方面，常用的技术包括：

* + Flume：可靠的、高可用的用于数据采集、聚合和传输的分布式系统
  + Logstash: 常用于数据采集，开源的服务器端数据处理管道
  + Sqoop：主要用于Hadoop（如HDFS、Hive、HBase）和RDBMS（如mysql、oracle）之间的数据导入导出
  + Kafka: 分布式消息系统，主要应用在数据缓冲、异步通信、汇集数据、系统接偶等方面
  + Pulsar：pub-sub模式的分布式消息平台，拥有灵活的消息模型和直观的客户端API。

### 2.2.2 数据存储层

大数据存储通常是指将数量巨大、难于收集、处理、分析的数据集持久化到计算机中。

在大数据存储方面，常用的技术包括：

* + HBase：基于Google Bigtable的开源实现，是一个具有高可靠性、高性能、面向列、可伸缩性、典型的key/value分布式存储的nosql数据库系统，主要用于海量结构化和半结构化数据存储。
  + HDFS：分布式文件存储系统，具有高容错（high fault-tolerant）、高吞吐（high throughput）、高可用（high available）的特性。
  + Kudu: 介于HDFS和HBase之间的基于列式存储的分布式数据库。兼具了HBase的实时性、HDFS的高吞吐，以及传统数据库的sql支持。

### 2.2.3 数据分析层

大数据分析即是指对规模巨大的数据进行分析，从中挖掘出隐藏的信息和知识。

在大数据分析方面，常用的技术包括：

* + Spark：快速、通用、可扩展、可容错的、内存迭代式计算的大数据分析引擎。
  + Flink：分布式的大数据处理引擎，可以对有限数据流和无线数据流进行有状态的计算。
  + MapReduce：分布式运算程序的编程框架，适用于离线数据处理场景。
  + Hive：基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供HQL语句（类SQL语言）查询功能，存储依赖于HDFS。

### 2.2.4 资源管理层

数据资源管理是应用数据库管理、数据仓库等信息系统技术和其他数据管理工具，完成组织数据资源管理任务，满足企业信息需求的管理活动。

在资源管理方面，常用的技术包括：

* + Yarn：资源调度平台，负责为运算程序分配资源和调度，不参与用户程序内部工作。
  + Kubernetes：又称K8s，为容器化的应用提供资源调度、部署运行、均衡容灾、服务注册、扩容缩容等功能的自动化容器操作的开源平台。

### 2.2.5 工作流调度器

工作流，是对工作流程及其各操作步骤之间业务规则的抽象、概括描述。工作流调度，即将工作流程中的工作如何前后组织在一起的逻辑和规则，在计算机中以恰当的模型表达并对其实施计算。

常用的工作流调度器有：

* + Oozie：基于工作流引擎的任务调度框架，能够提供能够提供对MapReduce和Pig 任务的调度与协调。
  + Azkaban：由LinkedIn开源，相对Oozie更轻量级。用于在一个工作流内以一个特定顺序运行一组任务。

**第3章 专业学习指导和建议**

对于大数据专业的同学来说，在认真学习课内内容的基础上，还需要学习较多课外知识，在此提出一条推荐的偏向于就业的专业学习路线（路线内容中部分课内课程中即有涉及，部分需要在课外额外花费时间进行学习）：

1. 巩固计算机基础知识。2.学习大数据框架的技术原理。3.进行项目实践。

下面将分条详细介绍这三步包含的具体内容。

## 3.1 计算机基础知识

* 精通一门语言：Java，C，C++，Python，Go，Scala，等等。（大数据建议选择 Java、Scala或者Python）。
* 数据结构与算法：链表，队列，堆，二叉树，排序，查找，贪心，回溯等。
* 计算机网络与基础以及操作系统相关知识。
* 数学：高等数学，线性代数，概率论与数理统计。

## 3.2 大数据框架技术

大数据方向的技术栈在上文中的大数据常用技术栈中已经进行了介绍，其中，Linux操作系统相关知识，Hadoop，ZooKeeper，Hive，HBase，Redis，Kafka，Spark，Flink建议重点学习，在学习时，可以参考网络上的视频资料以及一些推荐度较高的书籍。

## 3.3 项目实践

毕业生在进行企业应聘时，项目实践经历是简历上非常重要的部分，而在学校内，本科生比较难有机会接触到大型的商业实践项目，因此，想要有较好的项目实践经验，一般推荐提前准备，进入企业进行实习或者个人主动寻找一些项目实践机会，比如参加相关比赛，上网参考一些较好的开源项目或者网络上的项目实践视频来亲自实现一些项目，以达到积累经验，将学习的知识融会贯通，综合运用的目的。

**第4章 专业前沿研究方向**

目前大数据的前沿研究方向主要聚焦于下面五个方面：

## 4.1 可视化大数据分析

在进行数据分析之前，需要对数据进行探索式地考查，在此过程中，可视化会发挥很大的作用。对大数据进行分析以后，为了方便用户理解结果，也需要把结果展示出来。尤其是可视化移动数据分析工具，能追踪用户行为，让应用开发者得以从用户角度评估自己的产品，通过观察用户与一款应用的互动方式，开发者将能理解用户为何执行某些特定行为，从而为自己完善和改进应用提供依据。

## 4.2 AI

包括[大数据与神经计算](https://www.zhihu.com/search?q=%E5%A4%A7%E6%95%B0%E6%8D%AE%E4%B8%8E%E7%A5%9E%E7%BB%8F%E8%AE%A1%E7%AE%97&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)、深度学习、语义计算以及人工智能其他相关技术结合。得益于以[云计算](https://www.zhihu.com/search?q=%E4%BA%91%E8%AE%A1%E7%AE%97&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)、大数据为代表的计算技术的快速发展，使得信息处理速度和质量大为提高，能快速、并行处理海量数据。

## 4.3 跨学科领域交叉的数据融合分析与应用

由于现有的大数据平台易用性差，而垂直应用行业的数据分析又涉及领域专家知识和领域建模，目前在大数据[行业分析](https://www.zhihu.com/search?q=%E8%A1%8C%E4%B8%9A%E5%88%86%E6%9E%90&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)应用与通用的[大数据技术](https://www.zhihu.com/search?q=%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%8A%80%E6%9C%AF&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)之间存在很大的鸿沟，缺少相互的交叉融合。

因此，迫切需要进行跨学科和跨领域的大数据技术和应用研究，促进和推动大数据在典型和重大行业中的应用和落地，尤其是与物联网、移动互联、云计算、[社会计算](https://www.zhihu.com/search?q=%E7%A4%BE%E4%BC%9A%E8%AE%A1%E7%AE%97&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)等热点技术领域相互交叉融合。

## 4.4 大数据安全与隐私

[大数据时代](https://www.zhihu.com/search?q=%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%97%B6%E4%BB%A3&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)，各网站均不同程度地开放其用户所产生的实时数据，一些监测数据的[市场分析](https://www.zhihu.com/search?q=%E5%B8%82%E5%9C%BA%E5%88%86%E6%9E%90&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)机构可通过人们在社交网站中写入的信息、智能手机显示的位置信息等多种数据组合进行分析挖掘。然而，大数据时代的数据分析不能保证个人信息不被其他组织非法使用，用户隐私安全问题的解决迫在眉睫。

[安全智能](https://www.zhihu.com/search?q=%E5%AE%89%E5%85%A8%E6%99%BA%E8%83%BD&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)更加强调将过去分散的安全信息进行集成与关联，独立的分析方法和工具进行整合形成交互，最终实现智能化的安全分析与决策。

## 4.5 大数据治理

大数据将打开各行各业的数据“[潘多拉魔盒](https://www.zhihu.com/search?q=%E6%BD%98%E5%A4%9A%E6%8B%89%E9%AD%94%E7%9B%92&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)”。社交网站、电商巨头、电信运营商乃至金融、医疗、教育等行业，都将加入大数据的“淘金”热潮，政府部门同样会从大数据中获益匪浅。如何将海量数据应用于决策、营销和产品创新？如何利用大数据平台优化产品、流程和服务？如何利用大数据更科学地制定[公共政策](https://www.zhihu.com/search?q=%E5%85%AC%E5%85%B1%E6%94%BF%E7%AD%96&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)、实现社会治理？所有这一切，都离不开[大数据治理](https://www.zhihu.com/search?q=%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%B2%BB%E7%90%86&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A1505783265%7D)。可以说，在大数据战略从顶层设计到底层实现的“落地”过程中，治理是基础，技术是承载，分析是手段，应用是目的。

**第5章 专业本科生毕业去向**

哈尔滨工业大学数据科学与大数据专业每年读研深造和出国留学的人数占应届毕业生的70%以上，就业率100%。就业去向主要以国内外知名企业为主，也可到航空航天、通信、金融和保险等行业从事与数据科学与大数据技术相关的工作。毕业生的就业城市主要分布于北京、上海、深圳、杭州等。

该专业与美国、英国、加拿大、新加坡、澳大利亚等多个国家的20余所著名高校建立了合作关系。学生在读期间和毕业后，均可通过这些合作关系到国外访学或进行合作研究[1]。

**结 论**

数据科学与大数据作为近年来开设的专业，紧追当下互联网发展的热点方向，致力于培养学生的大数据思维方式和分析能力，正吸引着越来越多的哈工大学子。本文从专业研究内容，专业主要课程内容及专业主要技术栈，专业学习建议与指导，目前专业的前沿研究方向和专业本科生毕业去这五个方面简单介绍了该专业，希望能够为将来有意向报名该专业并从事相关工作的学生提供一个较为全面的专业概述，加深对该专业的了解；也希望文中的专业学习路径和热点研究方向能为本专业学子提供学习上的参考。

**参考文献**

1. 数据科学与大数据[EB/OL].(2022-05-24)[2022-09-05].[数据科学与大数据技术 - 哈尔滨工业大学招生办 (hit.edu.cn)](https://zsb.hit.edu.cn/article/read/5314044d4eeaa1db2e8743bf361bb6ce)
2. 大数据常用技术栈[EB/OL].(2020-11-01)[2022-09-06]. [大数据常用技术栈 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/95926504)
3. 数据采集和数据传输[EB/OL].(2022-02-24)[2022-09-06]. [数据采集和数据传输 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/471778662)