



**大数据的介绍、应用和前景**

|  |  |
| --- | --- |
| **专 业** | **数据科学与大数据技术** |
| **姓 名** | **黄俊方** |
| **日 期** | **2022/8/31** |

**摘 要**

本文介绍了数据科学与大数据技术专业研究内容、专业主要课程内容和专业主要技术栈介绍。并根据自身经验给出了专业学习建议与指导，探索了目前专业的前沿研究方向。最后对整个计算机大类专业本科生毕业去向展开了调研。

**关键词：数据科学，大数据技术，技术栈**

目录

[第1章 大数据应用介绍 4](#_Toc112870373)

[第2章 专业主要课程内容 4](#_Toc112870374)

[第3章 专业主要技术栈介绍 6](#_Toc112870375)

[3.1 Hadoop 6](#_Toc112870376)

[3.2 Hive 6](#_Toc112870377)

[3.3 Spark 7](#_Toc112870378)

[3.4 Flink 7](#_Toc112870379)

[3.5 Hbase 7](#_Toc112870380)

[3.6 Zookeeper 7](#_Toc112870381)

[3.7kafka 8](#_Toc112870382)

[3.8 Azkaban 8](#_Toc112870383)

[3.9 Flume 8](#_Toc112870384)

[第4章 专业学习建议与指导 8](#_Toc112870385)

[4.1 打基础，注重实践 8](#_Toc112870386)

[4.2 学好数学课 9](#_Toc112870387)

[4.3 学好英语 9](#_Toc112870388)

[4.4 广泛参与涉猎，不设限 9](#_Toc112870389)

[第5章 目前专业的前景 9](#_Toc112870390)

[第6章 专业本科生毕业去向 10](#_Toc112870391)

[6.1就业特点 10](#_Toc112870392)

[6.2专业就业优势 10](#_Toc112870393)

# 第1章 大数据应用介绍

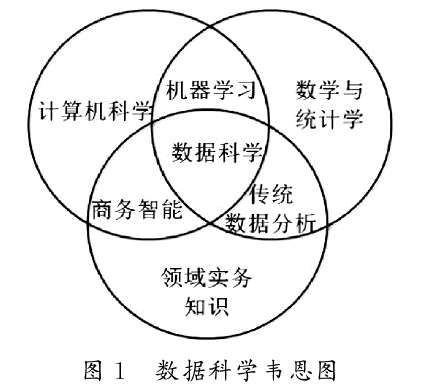
目前，大数据的主要来源和应用都是来自于企业内部。企业内部大数据的应用，可以在多个方面提升企业的生产效率和竞争力。具体而言：市场方面，利用大数据关联分析，更准确地了解消费者的使用行为，挖掘新的商业模式；销售规划方面，通过大量数据的比较，优化商品价格；运营方面，提高运营效率和运营满意度，优化劳动力投入，准确预测人员配置要求，避免产能过剩，降低人员成本；供应链方面，利用大数据进行库存优化、物流优化、供应商协同等工作，可以缓和供需之间的矛盾、控制预算开支，提升服务。

最典型的应用还是在电子商务领域，每天有数以万计的交易在淘宝上进行，与此同时相应的交易时间、商品价格、购买数量会被记录，更重要的是，这些信息可以与买方和卖方的年龄、性别、地址、甚至兴趣爱好等个人特征信息相匹配。淘宝数据魔方是淘宝平台上的大数据应用方案，通过这一服务，商家可以了解淘宝平台上的行业宏观情况、自己品牌的市场状况、消费者行为情况等，并可以据此进行生产、库存决策，而与此同时，更多的消费者也能以更优惠的价格买到更心仪的宝贝。而阿里信用贷款则是阿里巴巴通过掌握的企业交易数据，借助大数据技术自动分析判定是否给予企业贷款，全程不会出现人工干预。据透露，截至目前阿里巴巴已经放贷３００多亿元，坏账率约０。３％左右，大大低于商业银行。

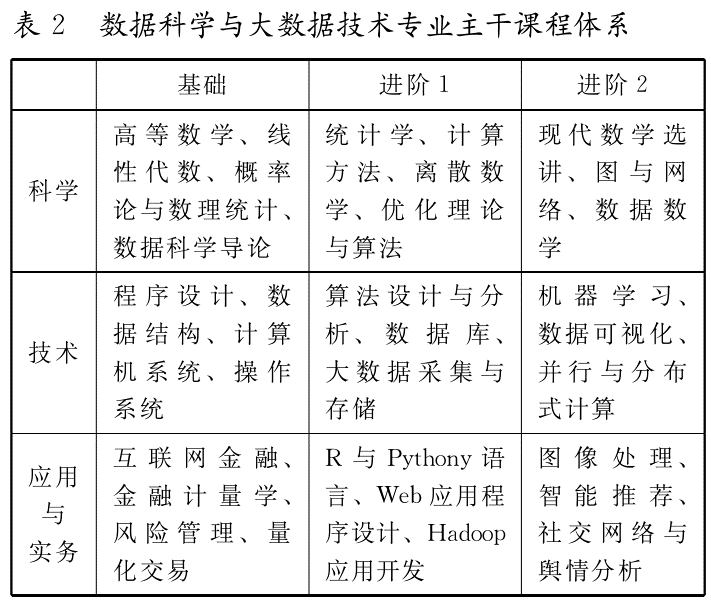
在线社交网络，是一种在信息网络上由社会个体集合及个体之间的连接关系构成的社会性结构。在线社交网络大数据主要来自即时消息、在线社交、微博和共享空间４类应用。由于在线社交网络大数据代表了人的各类活动，因此对于此类数据的分析得到了更多关注。在线社交网络大数据分析是从网络结构、群体互动和信息传播３个维度，通过基于数学、信息学、社会学、管理学等多个学科的融合理论和方法，为理解人类社会中存在的各种关系提供的一种可计算的分析方法。目前，在线社交网络大数据的应用包括网络舆情分析、网络情报搜集与分析、社会化营销、政府决策支持、在线教育等。

# 第2章 专业主要课程内容

数据科学的方法体系与理论基础与现有统计学、计算机科学等学科联系紧密，亦有明显区别。统计学一直是研究数据的科学，但在数据科学框架下，统计学自身发展也受到挑战。一方面，数据科学中的“科学”部分会跳出统计学，诸如拓扑学、随机场、几何甚至表示论等更一般的数学概念与理论也成为数据科学理论基础的重要组成部分。另一方面，大数据背景下，数据具有多类型、大规模、高速度与低价值密度等新特点，相关算法的设计及其实现是数据科学的重要研究内容，现有学科并不能全面覆盖。而计算机科学的研究对象是计算机软硬件系统本身，并不直接以数据为研究对象。在其基础上增加数据科学相关内容，只会使之臃肿并偏移自身的研究核心。数据科学是一门独立学科，与统计学、计算机科学及其他相关学科是补充而非简单的替代关系。



数据科学与大数据技术专业旨在培养全面掌握数据科学与大数据基础理论、基本技术与常用工具，了解应用领域大数据，能胜任大数据分析与挖掘、大数据处理系统开发与构建等工作的专门性科学技术人才，其课程体系应完整覆盖所涉基础学科以及大数据对象引发的专门性问题。为求不失一般性，我们从“科学”（数学、统计学、“数据数学”）、“技术”（计算机科学与大数据技术）及“应用与实务”三个基本方面给出数据科学与大数据技术本科专业主干课程体系的构想



# 第3章 专业主要技术栈介绍

### 3.1 Hadoop

Hadoop是一个由Apache基金会所开发的分布式系统基础架构。Hadoop实现了一个分布式文件系统，其中一个组件是HDFS。HDFS有高容错性的特点，并且设计用来部署在低廉的硬件上；而且它提供高吞吐量来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。

### 3.2 Hive

Hive是基于Hadoop的一个数据仓库工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。hive数据仓库工具能将结构化的数据文件映射为一张数据库表，并提供SQL查询功能，能将SQL语句转变成MapReduce任务来执行。

### 3.3 Spark

Spark是专为大规模数据处理而设计的快速通用的计算引擎。

Spark拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是其Job中间输出结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的MapReduce的算法。

### 3.4 Flink

Flink是由Apache软件基金会开发的开源流处理框架，其核心是用Java和Scala编写的分布式流数据流引擎。Flink以数据并行和流水线方式执行任意流数据程序，Flink的流水线运行时系统可以执行批处理和流处理程序。此外，Flink的运行时本身也支持迭代算法的执行。

### 3.5 Hbase

HBase是一个分布式的、面向列的开源数据库，该技术来源于 Fay Chang 所撰写的Google论文“Bigtable：一个结构化数据的分布式存储系统”。 HBase不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是HBase基于列的而不是基于行的模式。

### 3.6 Zookeeper

ZooKeeper是一个分布式的，开放源码的分布式应用程序协调服务，是Google的Chubby一个开源的实现，是Hadoop和Hbase的重要组件。它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等。

### 3.7kafka

Kafka是由Apache软件基金会开发的一个开源流处理平台，由Scala和Java编写。Kafka是一种高吞吐量的分布式发布订阅消息系统，它可以处理消费者在网站中的所有动作流数据。

### 3.8 Azkaban

Azkaban是由Linkedin公司推出的一个批量工作流任务调度器，用于在一个工作流内以一个特定的顺序运行一组工作和流程。Azkaban使用job配置文件建立任务之间的依赖关系，并提供一个易于使用的web用户界面维护和跟踪你的工作流。

### 3.9 Flume

Flume是Cloudera提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

# 第4章 专业学习建议与指导

### 4.1 打基础，注重实践

与其他行业相比，计算机专业无疑是看重专业实践的。只说不练假把戏，有理论却什么都不会用不会写，相当于什么都不会。

### 4.2 学好数学课

大数据与数学密切相关，学习与数学相关的学科可以让你轻松学习。同时，数学不仅是为了打牢专业基础，学习数学也能很好地锻炼我们的思维能力，而可编程文本并不是对我们逻辑思维能力的考验。特别对于大数据专业，概率统计和线性代数相当重要。

### 4.3 学好英语

学好英语，这不仅是对计算机相关专业的学生，对所有的大学生都是重要的。英语的意义应该超越一切想象，找份工作看看你的英语水平，还有英语面试，英语考试是必修课。而且，去工作或读研究后再深入研究计算机，基本上不会有中文材料供你选择，所有材料都是英文的。

### 4.4 广泛参与涉猎，不设限

计算机专业是一门关系广泛的专业，刚开始的时候不一定要限制自己，学习尽可能多的基础知识，比如数据结构，计算机组成原理、操作系统等，如果不知道将来要做什么，最好还是广涉猎，因为不管是学习还是工作时，计算机技术都会交叉。

# 第5章 目前专业的前景

大数据是现有产业升级与新产业诞生的重要推动力量。数据为王的大数据时代的到来，产业界需求与关注点发生了重大转变：企业关注的重点转向数据，计算机行业正在转变为真正的信息行业，从追求计算速度转变为关注大数据处理能力，软件也将从编程为主转变为以数据为中心。数据已成为与矿物和化学元素一样的原始材料，未来可能形成数据服务、数据探矿、数据化学、数据材料、数据制药等一系列战略性的新兴产业。

无论是国外政府的大数据研究计划，还是国内外大公司的大数据研发，当前最重视的都是大数据分析算法和大数据系统的效率。因此，当工业界把主要精力放在应对大数据的工程技术挑战的时候，科技界应开始着手关注大数据的基础理论研究。大数据科学作为一个新兴的交叉学科方向，其共性理论基础将来自多个不同的学科领域，包括计算机科学、统计学、人工智能、社会科学等。因此，大数据的基础研究离不开对相关学科的领域知识与研究方法论的借鉴。在大数据的基础研究方面，建议研究大数据的内在机理，包括大数据的生命周期、演化与传播规律，数据科学与社会学、经济学等之间的互动机制，以及大数据的结构与效能的规律性（如社会效应、经济效应等）。在大数据计算方面，研究大数据表示、数据复杂性以及大数据计算模型。在大数据应用基础理论方面，研究大数据与知识发现（学习方法、语义解释），大数据环境下的实验与验证方法，以及大数据的安全与隐私等。

# 第6章 专业本科生毕业去向

### 6.1就业特点

计算机专业毕业生的就业率发展相对平稳。根据每一届计算机专业毕业生的就业情况统计来看，就业情况还是很可观的。但是衡量就业情况，不能仅仅看就业率，还要考虑到薪资水平。随着计算机专业的不断发展，各大高校都在扩张计算机专业学生，因此现在的计算机专业学生的数量在持续增加，就业的整体质量不如从前。企业对毕业生的选择性加大，就加剧了就业的难度，社会对计算机专业人才的要求越来越高，要求专业知识和职业素养同时具备，并且水平要高。综上所述，计算机专业的就业岗位多，但是竞争也是十分激烈，计算机专业的学生不可掉以轻心。

### 6.2专业就业优势

（1）平均起步收入较高。许多专业的毕业生在刚刚踏上工作岗位的时候薪资水平都很低，但是计算机专业的学生在刚毕业的时候就能拿到较高的工资。这是一个其他专业所不具备的优势所在。

（2）就业面广。计算机领域涉及到的东西很多，各行各业都会有涉及到计算机专业的就业岗位。毕业生可以结合自己的兴趣和实际情况，来选择一个适合自己的职位。选择性很大，不会让学生有就业的拘束感。

（3）对计算机有兴趣的毕业生可以很快在工作中找到乐趣。这是一门专业性很高的工作，如果真心的喜欢，就会在工作中发现极大的乐趣，爱上工作，让自己的工作生活更有激情。

**参考文献**