大数据技术学习路线指南

大数据技术作为决策神器，日益在社会治理和企业管理中起到不容忽视的作用，美国，欧盟都已经将大数据研究和使用列入国家发展的战略，类似谷歌，微软，百度，亚马逊等巨型企业也同样把大数据技术视为生命线以及未来发展的关键筹码。这个系列的教程将从技术和应用的角度解读大数据与云计算里的具体内容，和你一起拔高人生的视野。

第一节：大数据是什么

首先，大数据技术是什么？

简而言之，从大数据中提取大价值的挖掘技术。专业的说，就是根据特定目标，从数据收集与存储，数据筛选，算法分析与预测，数据分析结果展示，以辅助作出最正确的抉择，其数据级别通常在 PB 以上，复杂程度前所未有。

关键作用是什么？

挖掘出各个行业的关键路径，帮助决策，提升社会（或企业）运作效率。

最初是在怎样的场景下提出？

在基础学科经历信息快速发展之后，就诞生了“大数据”的说法。但其实是随着数据指数级的增长，尤其是互联网商业化和传感器移动化之后，从大数据中挖掘出某个事件现在和未来的趋势才真正意义上被大众所接触。

大数据技术包含的内容概述？

非结构化数据收集架构，数据分布式存储集群，数据清洗筛选架构，数据并行分析模拟架构，高级统计预测算法，数据可视化工具。

大数据技术的具体内容？

分布式存储计算架构（强烈推荐：Hadoop）
分布式程序设计（包含：Apache Pig 或者 Hive）
分布式文件系统（比如：Google GFS）
多种存储模型，主要包含文档，图，键值，时间序列这几种存储模型（比如：BigTable，Apollo，DynamoDB 等）
数据收集架构（比如：Kinesis，Kafla）
集成开发环境（比如：R-Studio）
程序开发辅助工具（比如：大量的第三方开发辅助工具）
调度协调架构工具（比如：Apache Aurora）
机器学习（常用的有 Apache Mahout 或 H2O）
托管管理（比如：Apache Hadoop Benchmarking）
安全管理（常用的有 Gateway）
大数据系统部署（可以看下 Apache Ambari）
搜索引擎架构（ 学习或者企业都建议使用 Lucene 搜索引擎）
多种数据库的演变（MySQL/Memcached）
商业智能（大力推荐：Jaspersoft ）
数据可视化（这个工具就很多了，可以根据实际需要来选择）
大数据处理算法（10 大经典算法）

大数据中常用的分析技术？

A/B 测试、关联规则挖掘、数据聚类、
数据融合和集成、遗传算法、自然语言处理、
神经网络、神经分析、优化、模式识别、
预测模型、回归、情绪分析、信号处理、
空间分析、统计、模拟、时间序列分析

大数据未来的应用趋势预测？

每个人健康和生活都需要的个性化建议；
企业管理中的选择和开拓新市场的可靠信息来源；
社会治理中大众利益的发现与政策满足。

第二节：实践原型

引言：大数据的目的在于挖掘价值，而它的本质与 OODA 循环决策模型非常相似。用 OODA 这个原型来理解大数据是最合适的了！在战场上，OODA 循环决策的周期越短，胜算越大；在市场中，大数据收集和反馈信息最快，效果越好！

OODA 模型

概而论之，OODA 指的是在充分观察了解你和对手的环境的前提下，模拟对手在特定环境下的行为，进而做出一系列的对策，并且快速响应执行！之后又迅速收集反馈信息，进入下一个 OODA 循环决策。

OODA 与大数据

OODA 的整个处理流程，其实就是一个运动控制系统。大数据也是类似，从手机信息、处理分析到决策执行，这些都与 OODA 有异曲同工之妙！大数据的运算速度与 OODA 的循环速度一样，都提前决定着结果。

第三节：大数据的内幕

引言：接着前两篇对大数据的介绍之后，本篇从实际操作的角度分享大数据内部关键的运作机制，这是在真正开始学习大数据之前对大数据的一个概览。为的是让我们成为大数据的主人。

大数据运行机制

这是对大数据运行机制的概览，如果你阅读过上一篇（OODA），就会感觉非常熟悉。不错，他们在概念上是如出一撤的！不过实际操作却又有巨大的不同。

收集数据：

大数据的第一站就是收集和存储海量数据（公开/隐私）。现在每个人都是一个巨大的数据源，通过智能手机和个人笔记本释放出大量的个人行为信息。获取数据似乎已经变得越来越容易，数据收集这一模块最大的挑战在于获取海量数据的高速要求以及数据的全面性考虑。

清洗数据：

传统商业智能在数据清洗处理的做法（ETL）是，把准确的数据放入定义好的格式中，通过基础的抽取统计生成高维度的数据，方便直接使用。然而大数据有个最突出的特征——数据非结构化或者半结构化。因为数据有可能是图片，二进制等等。数据清洗的最大挑战来了——如何转化处理大量非结构数据，便于分布式地计算分析。

。

。

。

。

。

。

分享与反馈：

随着大数据分析结果的产生，决策者需要的旺旺不是一堆僵硬的数据，而是一张直观动态的决策建议视图。并且在决策之后，需要一个执行反馈系统来评估大数据分析结果的准确性。不断地去优化大数据分析的架构和算法！使得大数据架构更加智能!!

最后请你再次阅读这个系列的上一篇文章，对比大数据与 OODA 之间的异同点，并且在图纸上画出你对大数据的理解！

第四节：Hadoop 是什么

引言：Hadoop 作为大数据工业中的主引擎，了解 Hadoop 就像是在打开大数据这扇门。首

先它本身是一个分布式计算架构，更重要的是它是一个可扩展的生态系统，像 IBM，EMC，Amazon，微软，甲骨文等大型 IT 公司都已经有了基于 Hadoop 的商业化大数据产品。虽然现在还有比 Hadoop 更为先进的分布式架构（Dremel，DataFlow 等），但也都是基于 Hadoop 的改进升级，因此也说 Hadoop 是大数据的基础，基础的稳固决定了未来能走多远!!

Hadoop 是一个大家族，是一个开源的生态系统，是一个分布式运行系统，是基于 Java 编程语言的架构。不过它最高明的技术还是 HDFS 和 MapReduce，使得它可以分布式处理海量数据。

它与现存的文件系统不同的特性有很多，比如高度容错（即使中途出错，也能继续运行），支持多媒体数据和流媒体数据访问，高效率访问大型数据集合，数据保持严谨一致，部署成本降低，部署效率提交等，如图是 HDFS 的基础架构

MapReduce（并行计算架构）,它可以将计算任务拆分成大量可以独立运行的子任务，接着并行运算，另外会有一个系统调度的架构负责收集和汇总每个子任务的分析结果。其中 包含映射算法与规约算法。如图是 MapReduce 的内部计算步骤

。

。

。

。

。

。

两本最重要的书籍（这两本基本已经可以满足大部分你对 Hadoop 的需要）：

Hadoop 权威指南/Hadoop 最佳实践

第五节：大数据服务比较

第六节：大数据平台实例

第七节：为什么是 Hadoop

第八节：MapReduce 是什么

第九节：HDFS 是什么

。

。

。

。

。

。

本系列后续章节请前往官网阅读！