链滴

不懂这 25 个名词，好意思说你懂大数据？

作者：zsr251
原文链接：https://ld246.com/article/1519304953417
来源网站：链滴
许可协议：署名-相同方式共享 4.0 国际 (CC BY-SA 4.0)

如果你刚接触大数据，你可能会觉得这个领域很难以理解，无从下手。近日，Ramesh Dontha在DatConomy上连发两篇文章，扼要而全面地介绍了关于大数据的75个核心术语，这不仅是大数据初学者好的入门资料，对于高阶从业人员也可以起到查漏补缺的作用。

本文先介绍 了25 个基本大数据术语，帮助你温故知新，那么开始吧~

01 算法(Algorithm)

算法可以理解成一种数学公式或用于进行数据分析的统计学过程。那么，「算法」又是何以与大数据上关系的呢?要知道，尽管算法这个词是一个统称，但是在这个流行大数据分析的时代，算法也经常提及且变得越发流行。

02 分析(Analytics analyze)

让我们试想一个很可能发生的情况，你的信用卡公司给你发了封记录着你全年卡内资金转账情况的邮，如果这个时候你拿着这张单子，开始认真研究你在食品、衣物、娱乐等方面消费情况的百分比会怎?你正在进行分析工作，你在从你原始的数据(这些数据可以帮助你为来年自己的消费情况作出决定)中掘有用的信息。

那么，如果你以类似的方法在推特和脸书上对整个城市人们发的帖子进行处理会如何呢?在这种情况，我们就可以称之为大数据分析。所谓大数据分析，就是对大量数据进行推理并从中道出有用的信息以下有三种不同类型的分析方法，现在我们来对它们分别进行梳理。

03 描述性分析法(Descriptive Analytics)

如果你只说出自己去年信用卡消费情况为：食品方面 25%、衣物方面 35%、娱乐方面 20%、剩下 2% 为杂项开支，那么这种分析方法被称为描述性分析法。当然，你也可以找出更多细节。

04 预测性分析法(Predictive Analytics)

如果你对过去 5 年信用卡消费的历史进行了分析，发现每年的消费情况基本上呈现一个连续变化的势，那么在这种情况下你就可以高概率预测出：来年的消费状态应该和以往是类似的。这不是说我们预测未来，而是应该理解为，我们在「用概率预测」可能发生什么事情。在大数据的预测分析中，数科学家可能会使用先进的技术，如机器学习，和先进的统计学处理方法(这部分后面我们会谈到)来预天气情况、经济变化等等。

05 规范性分析(Prescriptive Analytics)

这里我们还是用信用卡转账的例子来理解。假如你想找出自己的哪类消费(如食品、娱乐、衣物等等)以对整体消费产生巨大影响，那么基于预测性分析(Predictive Analytics)的规范性分析法通过引入动态指标(action)」(如减少食品或衣物或娱乐)以及对由此产生的结果进行分析来规定一个可以降低你体开销的最佳消费项。你可以将它延伸到大数据领域，并想象一个负责人是如何通过观察他面前多种态指标的影响，进而作出所谓由「数据驱动」的决策的。

06 批处理(Batch processing)

尽管批量数据处理从大型机(mainframe)时代就已经存在了，但是在处理大量数据的大数据时代面前批处理获得了更重要的意义。批量数据处理是一种处理大量数据(如在一段时间内收集到的一堆交易数)的有效方法。分布式计算(Hadoop)，后面会讨论，就是一种专门处理批量数据的方法。

07 Cassandra

是一个很流行的开源数据管理系统，由Apache Software Foundation 开发并运营。Apache掌握了多大数据处理技术，Cassandra就是他们专门设计用于在分布式服务器之间处理大量数据的系统。

08 云计算(Cloud computing)

虽然云计算这个词现在已经家喻户晓，这里大可不必赘述，但是为了全篇内容完整性的考虑，笔者还在这里加入了云计算词条。本质上讲，软件或数据在远程服务器上进行处理，并且这些资源可以在网上任何地方被访问，那么它就可被称为云计算。

09 集群计算(Cluster computing)

这是一个来描述使用多个服务器丰富资源的一个集群(cluster)的计算的形象化术语。更技术层面的理是，在集群处理的语境下，我们可能会讨论节点(node)、集群管理层(cluster management layer)、载平衡(load balancing)和并行处理(parallel processing)等等。

10 暗数据(Dark data)

这是一个生造词，在笔者看来，它是用来吓唬人，让高级管理听上去晦涩难懂的。基本而言，所谓暗据指的是，那些公司积累和处理的实际上完全用不到的所有数据，从这个意义上来说我们称它们为「」的数据，它们有可能根本不会被分析。这些数据可以是社交网络中的信息，电话中心的记录，会议录等等。很多估计认为所有公司的数据中有60%到90%不等可能是暗数据，但实际上没人知道。

11 数据湖(Data lake)

当笔者第一次听到这个词时，真的以为这是个愚人节笑话。但是它真的是一个术语。所以一个数据湖(ata lake)即一个以大量原始格式保存了公司级别的数据知识库。这里我们介绍一下数据仓库(Data w rehouse)。数据仓库是一个与这里提到的数据湖类似的概念，但不同的是，它保存的是经过清理和并其它资源整合后的结构化数据。

数据仓库经常被用于通用数据(但不一定如此)。一般认为，一个数据湖可以让人更方便地接触到那些真正需要的数据，此外，你也可以更方便地处理、有效地使用它们。

12 数据挖掘(Data mining)

数据挖掘关乎如下过程，从一大群数据中以复杂的模式识别技巧找出有意义的模式，并且得到相关洞。它与前文所述的「分析」息息相关，在数据挖掘中，你将会先对数据进行挖掘，然后对这些得到的果进行分析。为了得到有意义的模式(pattern)，数据挖掘人员会使用到统计学(一种经典的旧方法)、器学习算法和人工智能。

13 数据科学家

数据科学家是时下非常性感的一门行业。它指那些可以通过提取原始数据(这就是我们前面所谓的数据)进而理解、处理并得出洞见的这样一批人。部分数据科学家必备的技能可以说只有超人才有：分析能、统计学、计算机科学、创造力、讲故事能力以及理解商业背景的能力。难怪这帮人工资很高。

14 分布式文件系统(Distributed File System)

大数据数量太大，不能存储在一个单独的系统中，分布式文件系统是一个能够把大量数据存储在多个储设备上的文件系统，它能够减少存储大量数据的成本和复杂度。

15 ETL

ETL代表提取、转换和加载。它指的是这一个过程：「提取」原始数据，通过清洗/丰富的手段，把数「转换」为「适合使用」的形式，并且将其「加载」到合适的库中供系统使用。即使ETL源自数据仓，但是这个过程在获取数据的时候也在被使用，例如，在大数据系统中从外部源获得数据。

16 Hadoop

当人们思考大数据的时候，他们会立即想到Hadoop。Hadoop是一个开源软件架构(logo是一头可爱大象)，它由 Hadoop分布式文件系统(HDFS)构成，它允许使用分布式硬件对大数据进行存储、抽象分析。如果你真的想让某人对这个东西印象深刻，你可以跟他说 YARN(Yet Another Resource Scheuler)，顾名思义，就是另一个资源调度器。我确实被提出这些名字的人深深震撼了。提出 Hadoop 的 Apache 基金会，还负责 Pig、Hive 以及 Spark(这都是一些软件的名字)。你没有被这些名字惊艳到?

17 内存计算(In-memory computing)

通常认为，任何不涉及到 I/O访问的计算都会更快一些。内存计算就是这样的技术，它把所有的工作据集都移动到集群的集体内存中，避免了在计算过程中向磁盘写入中间结果。Apache Spark 就是一内存计算的系统，它相对 Mapreduce 这类 I/O 绑定的系统具有很大的优势。

18 物联网(IoT)

最新的流行语就是物联网(IoT)。IoT 是嵌入式对象中(如传感器、可穿戴设备、车、冰箱等等)的计算备通过英特网的互联，它们能够收发数据。物联网生成了海量的数据，带来了很多大数据分析的机遇。

19 机器学习(Machine Learning)

机器学习是基于喂入的数据去设计能够学习、调整和提升的系统的一种方法。使用设定的预测和统计法，它们持续地逼近「正确的」行为和想法，随着更多的数据被输入到系统，它们能够进一步提升。

20 MapReduce

MapReduce可能有点难以理解，我试着解释一下吧。MapReduce是一个编程模型，最好的理解就要注意到Map和Reduce是两个不同的过程。在 MapReduce中，程序模型首先将大数据集分割成一小块(这些小块拿技术术语来讲叫做「元组」，但是我描述的时候会尽量避免晦涩的技术术语)，然后些小块会被分发给不同位置上的不同计算机(也就是说之前描述过的集群)，这在Map过程是必须的。后模型会收集每个计算结果，并且将它们「reduce」成一个部分。MapReduce的数据处理模型和Haoop分布式文件系统是分不开的。

21 非关系型数据库(NoSQL)

这个词听起来几乎就是「SQL，结构化查询语言」的反义词，SQL 是传统的关系型数据管理系统(RDMS)必需的，但是 NOSQL 实际上指的是「不止SQL」。

NoSQL实际上指的是那些被设计来处理没有结构(或者没有「schema」，纲要)的大量数据的数据库理系统。NoSQL 适合大数据系统，因为大规模的非结构化数据库需要 NoSQL的这种灵活性和分布式先的特点。

22 R语言

这还有人能给一个编程语言起一个更加糟糕的名字吗?R 语言就是这样的语言。不过，R 语言是一个统计工作中工作得很好的语言。如果你不知道 R 语言，别说你是数据科学家。因为 R 语言是数据科中最流行的编程语言之一。

23 Spark(Apache Spark)

Apache Spark 是一个快速的内存数据处理引擎，它能够有效地执行那些需要迭代访问数据库的流处、机器学习以及SQL负载。Spark通常会比我们前面讨论过的MapReduce 快好多。

24 流处理(Stream processing)

流处理被设计来用于持续地进行流数据的处理。与流分析技术(指的是能够持续地计算数值和统计分析

能力)结合起来，流处理方法特别能够针对大规模数据的实时处理。

25 结构化vs非结构化数据(Structured v Unstructured Data)

这是大数据中的对比之一。结构化数据基本上是那些能够被放在关系型数据库中的任何数据，以这种式组织的数据可以与其他数据通过表格来关联。非结构化数据是指任何不能够被放在关系型数据库中数据，例如邮件信息、社交媒体上的状态，以及人类语音等等。