大数据的前景以及在企业中的运用

编者话：这是一篇从大数据从业者角度看大数据的文章，主要面向对象是大数据学习，准备转行大数据的工程师，准备面试大数据工程师或者大数据初级入门等人员。所以没有过多地从技术层面分析大数据架构，而是从大数据工程师的实用的角度给出一些建议。

主要内容：

1. 大数据的前世今生概况。
2. 大数据的主流架构。
3. 学大数据要学什么，转行难不难？
4. 大数据工程师日常工作。
5. 大数据面试 tips。

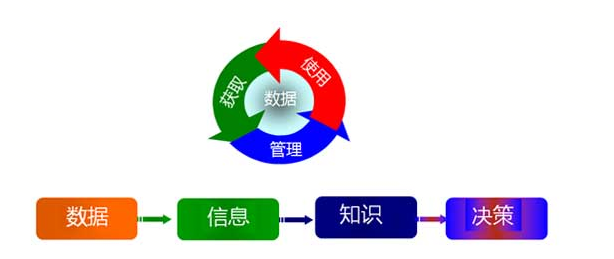
1. 大数据的前世

说起大数据，不得不说大数据的前辈：BI（Business Intelligence）商业智能。在大数据出现之前几乎所有的大型公司，都会有BI框架。商业智能的概念最早在 1996 年提出。

当时将商业智能定义为一类由数据仓库（或数据集市）、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成的、以帮助企业决策为目的技术及其应用。而这些数据可能来自企业的 ERP、CRM、SCM 等业务系统。

**BI 主要解决的是从数据到决策的这么一个过程**

说到底也是一个从数据到分析结果的过程。所以大数据可以看做 BI 的一个加强版。



那么既然已经有了 BI 了，为什么还需要大数据呢？关于大数据麦肯锡全球研究所给出的定义是：**一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合**，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

是的，大数据的出现是因为随着社会的发展，企业搜集到的需要处理的数据让传统的 BI 框架不能够满足日常分析。所以技术有了一个新的迭代：大数据生态系统

2.大数据的今生

2.1 大数据是干什么的？

大数据简单一句话：海量数据存储和海量数据计算。以我们公司来说，每天搜集的数据大概 2 PB（1024 \* 1024 G 的大小）。从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘。但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

2.2 大数据的处理框架：数据存储和计算

大数据既然是海量数据的存储和海量数据的计算。那么就要面对两个问题，大数据的存储问题和大数据的计算问题。

目前主流的分布式存储系统有 apache 的 HDFS，和亚马逊的 S3。那么为什么需要分布式存储呢？什么是分布式存储？

**大数据的分布式存储**

在这个数据爆炸的时代，产生的数据量不断地在攀升，从 GB，TB，PB，ZB。

挖掘其中数据的价值也是企业在不断地追求的终极目标。但是要想对海量的数据进行挖掘，首先要考虑的就是海量数据的存储问题，比如 Tb 量级的数据。

谈到数据的存储，则不得不说的是磁盘的数据读写速度问题。早在上个世纪 90 年代初期，普通硬盘的可以存储的容量大概是 1 G 左右，硬盘的读取速度大概为 4.4 MB/s.读取一张硬盘大概需要 5 分钟时间，但是如今硬盘的容量都在 1 TB 左右了,相比扩展了近千倍。

但是硬盘的读取速度大概是 100 MB/s。读完一个硬盘所需要的时间大概是 2.5 个小时。所以如果是基于 TB 级别的数据进行分析的话，光硬盘读取完数据都要好几天了，更谈不上计算分析了。那么该如何处理大数据的存储，计算分析呢？

一个很简单的减少数据读写时间的方法就是同时从多个硬盘上读写数据，比如，如果我们有 100 个硬盘，每个硬盘存储 1 % 的数据 ，并行读取，那么不到两分钟就可以完成之前需要 2.5 小时的数据读写任务了。这就是大数据中的分布式存储的模型。

当然实现分布式存储还需要解决很多问题，比如硬件故障的问题，使用多台主机进行分布式存储时，若主机故障，会出现数据丢失的问题，所以有了副本机制：系统中保存数据的副本。

一旦有系统发生故障，就可以使用另外的副本进行替换（著名的 RAID 冗余磁盘阵列就是按这个原理实现的）。其次比如一个很大的文件如何进行拆分存储，读取拆分以后的文件如何进行校验都是要考虑的问题。比如我们使用 Hadoop 中的 HDFS 也面临这个问题，只是框架给我们实现了这些问题的解决办法，开发中开发者不用考虑这些问题，底层框架已经实现了封装。

同样假如有一个 10 TB 的文件，我们要统计其中某个关键字的出现次数，传统的做法是遍历整个文件，然后统计出关键字的出现次数，这样效率会特别特别低。

基于分布式存储以后，数据被分布式存储在不同的服务器上，那么我们就可以使用分布式计算框架（比如 MapReduce，Spark 等）来进行并行计算（或者说是分布式计算），即：每个服务器上分别统计自己存储的数据中关键字出现的次数，最后进行一次汇总，那么假如数据分布在 100 台服务器上，即同时 100 台服务器同时进行关键字统计工作，效率一下子可以提高几十倍。这也是分布式存储和分布式计算让人着迷的地方。

**大数据的分布式计算**

同样因为大数据的数据之大，决定了分布式存储，所以同样因为计算量之大，需要分布式计算（但因为是分布式计算，即批处理要求大数据的同一批次处理的数据结构要一样）大数据的分布式计算框架从时效角度来看，分离线计算框架和实时计算框架。目前主流的离线计算框架是 MapReduce，实施计算框架是 Storm 和 Spark，Flink。具体原理参考我的博客

2.3 大数据的生态系统

所谓的大数据生态系统简单可以理解成集成了大数据的存储和计算以及分析等常见技能于一身的技术框架。

比如目前主流的 Hadoop 生态系统和 Spark 生态系统。Hadoop 自身提供了 HDFS 用来数据存储，提供了 MapReduce 用来数据处理（离线计算）。而 Spark 提供了各种实时计算，机器学习的计算框架。但是 Spark 没有提供了数据存储的框架，所以还是依赖于第三方的存储框架，比如 Hadoop 的 HDFS，比如亚马逊的 S3 存储系统。

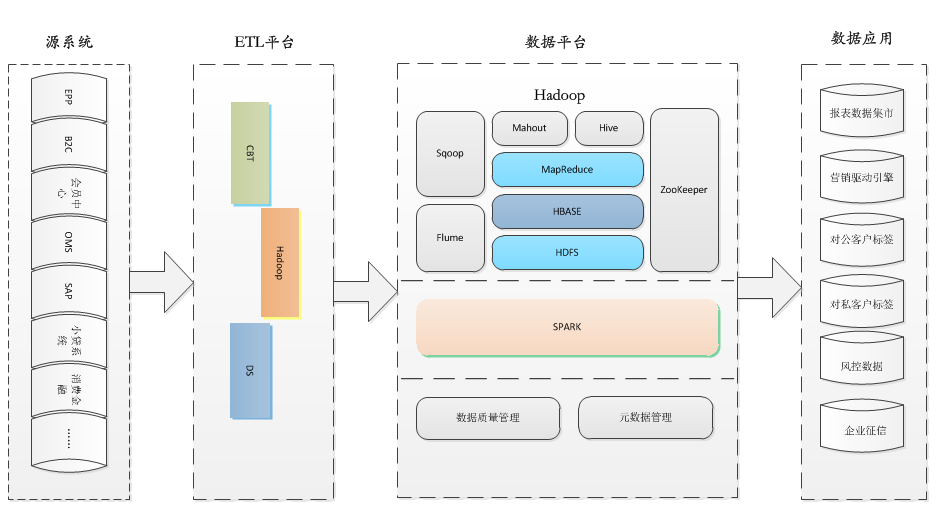
因为各种框架基本都是开源的，所以各大厂商，开源基金都会开发维护，然后发布。造成比如 Hadoop 就有很多版本可供使用。主流的是 apache 的版本。具体后面介绍

3.大数据的技术框架

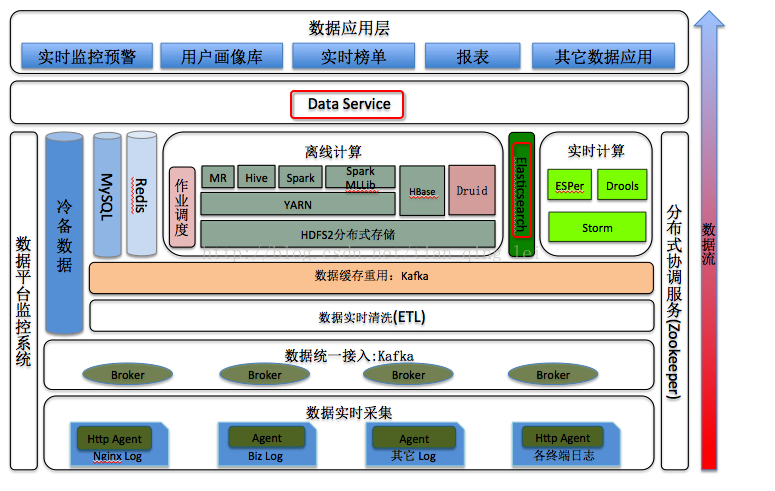
3.1 公司大数据技术架构：存储 + 计算

所谓大数据架构：就是公司大数据计算平台的抽象化。因为分布式存储有很多技术可以实现，如 HDFS，S3 等，分布式计算框架也有很多技术可以实现，如 Flink，Spark，MR，同样任务调度平台也有很多技术实现，如 oozie，azakaban，比如我们公司的任务调度平台就是自己封装的。所以不同公司可能选择的技术框架不一样。

**但是所谓技术架构核心都是存储+计算+任务调度。**



实际各个公司根据自己的需求不同，会对架构的框架做一些增删。但是核心的存储和计算肯定都会有，只是选择技术框架可能不同。比如有的公司只需要离线计算，难么可能只需要配置 HDFS，Hive，Zookeeper 等即可。如果有的公司需要实时计算，可能需要 Storm，Kafka 等。



3.2 介绍主流框架 Hadoop

1. Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构,是由 Lucene--DougCutting 开创的开源软件，用 java 书写代码，实现与 Google 类似的全文搜索功能，它提供了全文检索引擎的架构，包括完整的查询引擎和索引引擎。
2. 主要解决，海量数据的存储和海量数据的分析计算问题。
3. 广义上来说，HADOOP 通常是指一个更广泛的概念 —— HADOOP 生态圈。

3.3 Hadoop 的版本选择

Hadoop 三大发行版本：Apache、Cloudera、Hortonworks。

1.Apache 版本最原始（最基础）的版本，对于入门学习最好。  
2.Cloudera 在大型互联网企业中用的较多。Cloudera 开发的 hadoop 版本被称为 CDH 版本（CDH 版本号和 hadoop 版本号不同）。  
3.Hortonworks 文档较好。

3.4 大数据架构里技术框架介绍

1. **Sqoop**：sqoop 是一款开源的工具，主要用于在 Hadoop(Hive)与传统的数据库 (mysql) 间进行数据的传递，可以将一个关系型数据库（例如：MySQL，Oracle 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导进到关系型数据库中。
2. **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；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。
3. **Kafka**：Kafka 是一种高吞吐量的分布式发布订阅消息系统，有如下特性：

（1）通过 O(1) 的磁盘数据结构提供消息的持久化，这种结构对于即使数以 TB 的消息存储也能够保持长时间的稳定性能。  
（2）高吞吐量：即使是非常普通的硬件Kafka也可以支持每秒数百万的消息。  
（3）支持通过 Kafka 服务器和消费机集群来分区消息。  
（4）支持 Hadoop 并行数据加载。

1. **Storm**：Storm 为分布式实时计算提供了一组通用原语，可被用于“流处理”之中，实时处理消息并更新数据库。这是管理队列及工作者集群的另一种方式。Storm 也可被用于“连续计算”（continuous computation），对数据流做连续查询，在计算时就将结果以流的形式输出给用户。
2. **Spark**：Spark 是当前最流行的开源大数据内存计算框架。可以基于 Hadoop 上存储的大数据进行计算。
3. **Oozie**：Oozie 是一个管理Hdoop作业（job）的工作流程调度管理系统。Oozie 协调作业就是通过时间（频率）和有效数据触发当前的 Oozie 工作流程。
4. **Hbase**：HBase 是一个分布式的、面向列的开源数据库。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
5. **Hive**：hive 是基于 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，并提供简单的 sql 查询功能，可以将 sql 语句转换为 MapReduce任务进行运行。其优点是学习成本低，可以通过类 SQL 语句快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计分析。
6. **R 语言**：R 是用于统计分析、绘图的语言和操作环境。R 是属于 GNU 系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。
7. **Mahout**: Apache Mahout 是个可扩展的机器学习和数据挖掘库，当前 Mahout 支持主要的 4 个用例：
   * 推荐挖掘：搜集用户动作并以此给用户推荐可能喜欢的事物。
   * 聚集：收集文件并进行相关文件分组。
   * 分类：从现有的分类文档中学习，寻找文档中的相似特征，并为无标签的文档进行正确的归类。
   * 频繁项集挖掘：将一组项分组，并识别哪些个别项会经常一起出现。
8. **ZooKeeper**：Zookeeper 是 Google 的 Chubby 一个开源的实现。它是一个针对大型分布式系统的可靠协调系统，提供的功能包括：配置维护、名字服务、分布式同步服务等。ZooKeeper 的目标就是封装好复杂易出错的关键服务，将简单易用的接口和性能高效、功能稳定的系统提供给用户。

3.5 对大数据的认识

这里介绍的大数据技术框架只需要有一个感性的认识，知道大数据是干嘛的，核心的东西是什么即可。所谓的大数据架构是个怎么回事。对自己学习大数据心里有底即可。做到先见森林即可，至于见树木那是后面学习的事，也可以关注我的博客，后续会更新全部的教程，有问题也可以一起探讨。

4. 大数据工程师日常工作内容是什么？

现在几乎稍微大一点 IT 公司都会有大数据研发部，就连不是 IT 的公司也在拥抱大数据，比如飞利浦，恒瑞医药等非 IT 公司（因为这两家公司曾邀请我去面试大数据，组建大数据研发部）。目前大数据按时效来分：实时计算和离线计算。基本来说，市面上除了一些电商公司，部分实时交易公司（如滴滴，携程），大部分公司的公司的数据计算都是离线计算。

**毫不夸张地说，绝大部分大数据工程师出去找工作最终的归宿也是离线计算一块，因为实时计算公司用的不多。**

**离线计算：**一般来说除非你去的公司是刚组建大数据部门，需要你搭建集群，负责参与架构，技术实施一块。当然这种公司不是太多，一般以小公司多。更多的是你去的公司都已经搭建好了大数据开发平台，并且已经有专门的人参与运维。

**你的工作主要是面对复杂的业务逻辑，在大数据平台行进行项目开发实施，任务调度，将结果反馈给业务方**。

而对于离线来说，目前大数据公司，比如京东，离线计算的主流开发语言是 hql（Hive SQL），即更多的面对的是 Hive 一块的开发，需要你有很深的 SQL 功底（换句话来说：甚至可以不需要 java 功底）。如果你工作的内容是平台研发部，那么工作的内容更多的是集群运维，调优，节点的部署，技术的更新，平台架构的运维和开发。

因为很多大公司的技术框架都不是直接使用的开源版本，而是在此基础上自己公司二次开发的。

**实时计算：** 尽管是实时计算，需要比如什么 Flume，Kafka，Storm 等技术框架，除了小型公司，一般中大型公司都是直接搭建好的集群，做实时开发的时候更多的是申请资源，按要求配置使用。比如 kafka 的使用直接申请 topic 即可。同样如果是平台开发部，则需要开发框架之类的，需要较深的 java 等功底。

但是相对业务开发人员来说，平台开发人员要远少于业务一块的开发人员。但是如果是小公司，可能需要你即面对业务的开发，也负责的平台的开发。

**机器学习：**这也算是一个单独的方向吧，主要是人工智能一类的公司或者大公司的一些研发部门会做的业务，主要用 Spark，R 语言，Python，Mahout 等技术实现，相对来说，门槛稍微高一些。当然也算大数据的分支，应用的也是大数据的技术。

5. 学大数据要学什么？转行难不难？

虽然大数据已经火了很多年，但是目前国内所有高校还没有正规大数据专业的毕业生。直到 2017 年全国部分高校才开始开设大数据专业。

说明了两点：

1. 大数据专业未来的前景可期，未来的技术趋势，薪水也会很可观。
2. 国内高校技术与商业发展的脱钩，这个确实做得不如人家美国的高校。

几乎可以说目前各大公司的大数据从业者都是从其他方向转过来的的，所以大家也都是半路出家的。市面上资料也很多，培训机构也很多。我的大数据技术一部分是自己学习的，一部分是公司研发中学习的，包括公司组织的技术培训。也看过市面上的大数据培训机构的视频，也招聘过类似机构人学员。

**负责任的说很多大数据培训机构的老师绝对没有做过实际大数据开发，我猜很多也是自己看资料学习然后去培训别人的，因为教的很多技术不是开发中使用的，或者说实际开发中压根就不是这样使用的，同时很多开发中使用的技术根本没有教授**。

尽管如此，基础视频的学习比看资料效果好很多。

**大数据系统学习要学哪些知识点？**

* Java SE 基础 （可以包扣部分框架 SSM）
* Mysql，SQL 编程基础
* linux 基础入门，Shell 编程；
* Hadoop 集群的搭建，HDFS，MR，YARN 的工作原理，写出基础的 MR 程序（实际开发中没有公司再写 MR 了，更多的是用 Hive，但是面试会问）。
* Hive的使用与编程（离线开发的核心技术），部分数据仓库的建模的知识；
* Flume，kafka 的使用；
* 任务调度平台 Oozie，azakaban 学习；
* Storm（准实时计算，部分公司不用）；
* Scala，Spark 编程（实时，机器学习用的多）；
* 熟练掌握两个离线项目，一个实时项目。

**Warm Tips**：如果有 Java 基础的话，转行大数据相对来说难度不大。当然这里的指的是入门大数据开发工程师，可以找到工作。大数据另一个应用方向：人工智能，机器学习那一块则需要很深的数学基础，算法学习，需要靠长期的不断积累的，不断地学习才能企及。

如果小白的话，转行大数据的话，找有工作经验的人带你入门，效果会好很多。或者网上找些视频先学习一下，打个基础。关于报培训机构的问题，如果有基础的人想转行大数据，个人觉得没必要。培训机构的老师很多也不是很 6，甚至都没有做过大数据开发吧，教授给你的东西有限。主要是学费那么黑。

6. 大数据面试一点建议

**1. 简历问题：**

简历的书写要用心哇，要用心哇，要用心哇**（重要的事说三遍）**面试中收到过很多大数据面试简历，说实在，乱七八糟，看都不想看。最起码排班要整洁好看，可以网上花个几块钱到专门简历模板的网站买个模板，不用花哨，但是要清爽。简历是你呈递给面试官的第一印象。个人觉得简历 3 张 A4 纸最好，少了显得内容不够丰富，多了冗余。彩色打印，稍微配个 1 块钱的塑料壳，主要是让别人觉得你用心了。

其次，简历的内容，一般来说自己掌握的核心知识点做一个简洁的概况。项目写三到四个，项目的排版很重要（具体参考下图）。项目是干什么的不重要，重要的你在项目的中角色，用了什么技术。注意面试中：**大公司会问的很细，包扣知识点等。但小公司都是直接问项目，要对简历上的项目特别熟悉。**



一般关于项目的面试都是：介绍一下项目-->项目的架构--->用到了什么技术，然后根据你在项目中角色问技术点（所以不熟的知识点不要写在自己的角色里）--->项目开发中遇到的问题等等，以及一些延伸。

**2. 面试常见知识点**

1. 根据工作性质的不同，有些公司会问 Java 基础等，但是很多公司也不一定会问，具体看招聘要求，好好研究一下。但是 Java 基础，框架还是有机会的话，要多学习，因为有利于提高自己的水平。
2. linux 常用命令基本必问，基础的 Shell 编程，看的懂，写的出来
3. HDFS 存储原理，Namenode 原理，MR 的 Shuffle 过程描述，Yarn 资源调度等原理，有没有看过源码之类的，调优等。Hadoop 这一块的面试几乎所有公司都会问，具体面试题，我的博客里有一部分，网上也有很多。可能工作中，你不一定用的到，但是面试必问，所以要掌握。否则连面试都过不了的话，还谈什么未来。
4. Flume，Kafka 等工作原理，如果面试的是非实时，可以说明用的不多。比如 Flume 如果解决丢数据问题？Flume 拥堵怎么办？
5. Spark 几乎大数据公司都要求了解，很多工作用的不是太多。国内来说，最大的 Spark 集群在腾讯。要了解 Spark 编程，Scala 等。其实很多大厂的实时框架用的还是 Storm，比如阿里，苏宁易购。因为 Spark 不是准实时。
6. 面试内容太多，博客后期会逐步出一个面试的集锦，感兴趣的可以关注。面试之前最好找个同事朋友先模拟面试一下最好。