从0到1，成为大数据行业领袖

2018-03-25 JAVA高级架构

目前最火的大数据，很多人想往大数据方向发展，想问该学哪些技术，学习路线是什么样的，觉得大数据很火，就业很好，薪资很高。如果自己很迷茫，为了这些原因想往大数据方向发展，也可以，那么我就想问一下，你的专业是什么，对于计算机/软件，你的兴趣是什么?是计算机专业，对操作系统、硬件、网络、服务器感兴趣?是软件专业，对软件开发、编程、写代码感兴趣?还是数学、统计学专业，对数据和数字特别感兴趣。

其实这就是想告诉你的大数据的三个发展方向，平台搭建/优化/运维/监控、大数据开发/ 设计/ 架构、数据分析/挖掘。请不要问我哪个容易，哪个前景好，哪个钱多。

先扯一下大数据的4V特征：

数据量大，TB->PB

数据类型繁多，结构化、非结构化文本、日志、视频、图片、地理位置等;

商业价值高，但是这种价值需要在海量数据之上，通过数据分析与机器学习更快速的挖掘出来;

处理时效性高，海量数据的处理需求不再局限在离线计算当中。

现如今，正式为了应对大数据的这几个特点，开源的大数据框架越来越多，越来越强，先列举一些常见的：

文件存储：Hadoop HDFS、Tachyon、KFS

离线计算：Hadoop MapReduce、Spark

流式、实时计算：Storm、Spark Streaming、S4、Heron

K-V、NOSQL数据库：HBase、Redis、MongoDB

资源管理：YARN、Mesos

日志收集：Flume、Scribe、Logstash、Kibana

消息系统：Kafka、StormMQ、ZeroMQ、RabbitMQ

查询分析：Hive、Impala、Pig、Presto、Phoenix、SparkSQL、Drill、Flink、Kylin、Druid

分布式协调服务：Zookeeper

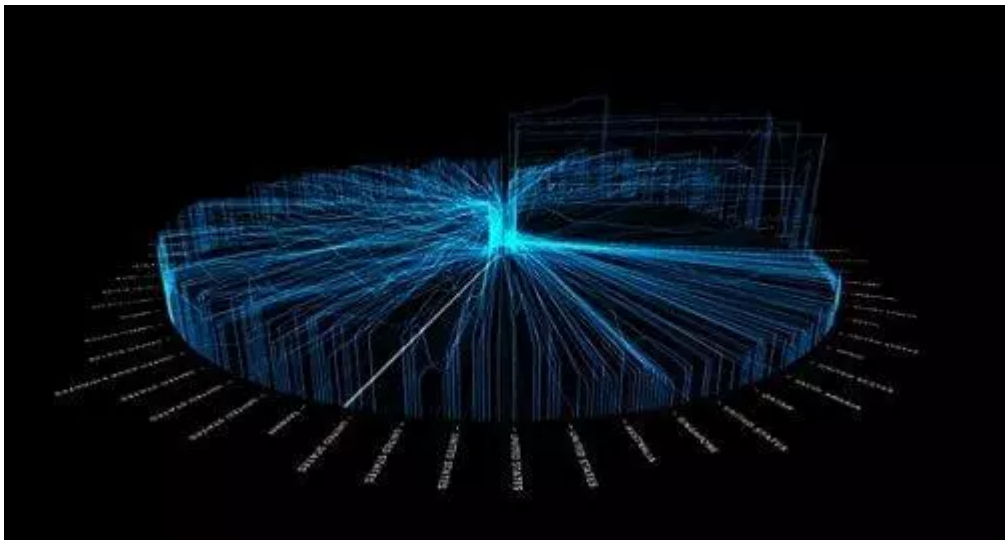
集群管理与监控：Ambari、Ganglia、Nagios、Cloudera Manager

数据挖掘、机器学习：Mahout、Spark MLLib

数据同步：Sqoop

任务调度：Oozie

眼花了吧，上面的有30多种吧，别说精通了，全部都会使用的，估计也没几个。就我个人而言，主要经验是在第二个方向(开发/设计/架构)，且听听我的建议吧，有安装教程。



初识Hadoop

1.1 学会百度与Google

不论遇到什么问题，先试试搜索并自己解决。Google首选，翻不过去的，就用百度吧。

1.2 参考资料首选官方文档

特别是对于入门来说，官方文档永远是首选文档。相信搞这块的大多是文化人，英文凑合就行，实在看不下去的，请参考第一步。

1.3 先让Hadoop跑起来

Hadoop可以算是大数据存储和计算的开山鼻祖，现在大多开源的大数据框架都依赖Hadoop或者与它能很好的兼容。

关于Hadoop,你至少需要搞清楚以下是什么：

Hadoop 1.0、Hadoop 2.0

MapReduce、HDFS

NameNode、DataNode

JobTracker、TaskTracker

Yarn、ResourceManager、NodeManager

自己搭建Hadoop，请使用第一步和第二步，能让它跑起来就行。建议先使用安装包命令行安装，不要使用管理工具安装。另外：Hadoop1.0知道它就行了，现在都用Hadoop 2.0.

1.4 试试使用Hadoop

HDFS目录操作命令;上传、下载文件命令;提交运行MapReduce示例程序;打开Hadoop WEB界面，查看Job运行状态，查看Job运行日志。知道Hadoop的系统日志在哪里。

1.5 你该了解它们的原理了

MapReduce：如何分而治之;HDFS：数据到底在哪里，什么是副本;

Yarn到底是什么，它能干什么;NameNode到底在干些什么;Resource Manager到底在干些什么;

1.6 自己写一个MapReduce程序

请仿照WordCount例子，自己写一个(照抄也行)WordCount程序，

打包并提交到Hadoop运行。你不会Java?Shell、Python都可以，有个东西叫Hadoop Streaming。
如果你认真完成了以上几步，恭喜你，你的一只脚已经进来了。

大数据方向的工作目前分为三个主要方向:
01.大数据工程师
02.数据分析师
03.大数据科学家
04.其他（数据挖掘本质算是机器学习，不过和数据相关，也可以理解为大数据的一个方向吧）

总结如下:

必须技能10条:
01.Java高级(虚拟机、并发)
02.Linux 基本操作
03.Hadoop（此处为侠义概念单指HDFS+MapReduce+Yarn ）
04.HBase（JavaAPI操作+Phoenix ）
05.Hive(Hql基本操作和原理理解）
06.Kafka
07.Storm
08.Scala需要
09.Python
10.Spark (Core+sparksql+Spark streaming ）
高阶技能6条:
11.机器学习算法以及mahout库加MLlib
12.R语言
13.Lambda 架构
14.Kappa架构
15.Kylin
16.Aluxio
二、学习路径

由于本人是从Java开发通过大概3个月的自学转到大数据开发的。所以我主要分享一下自己的学习路劲。

第一阶段:

01.Linux学习（跟鸟哥学就ok了）

02.Java 高级学习（《深入理解Java虚拟机》、《Java高并发实战》）

第二阶段:

03.Hadoop （董西成的书）

04.HBase（《HBase权威指南》）

05.Hive（《Hive开发指南》）

06.Scala（《快学Scala》）

07.Spark （《Spark 快速大数据分析》）

08.Python （跟着廖雪峰的博客学习就ok了）

第三阶段:

对应技能需求，到网上多搜集一些资料就ok了，

我把最重要的事情(要学什么告诉你了),

剩下的就是你去搜集对应的资料学习就ok了

当然如果你觉得自己看书效率太慢，你可以网上搜集一些课程，跟着课程走也OK 。这个完全根据自己情况决定。如果看书效率不高就很网课，相反的话就自己看书。

三，学习资源推荐:

01.Apache 官网

02.Stackoverflow

04.github

03.Cloudra官网

04.Databrick官网

05.过往的记忆（技术博客）

06.CSDN，51CTO

07.至于书籍当当一搜会有很多，其实内容都差不多。

最后但却很重要一点:要多关注技术动向，持续学习。

扫描下方二维码，会给你一个大数据学科体系

添加美女微信号获取学习体系资料