大数据开发需要用到的技术

大数据是对海量数据进行存储、计算、统计、分析处理的一系列处理手段，处理的数据量通常是TB级，甚至是PB或EB级的数据，这是传统数据处理手段所无法完成的，其涉及的技术有分布式计算、高并发处理、高可用处理、集群、实时性计算等，汇集了当前IT领域热门流行的IT技术。

**想要学好大数据需掌握以下技术：**

**1. Java编程技术**

Java编程技术是大数据学习的基础，Java是一种强类型语言，拥有极高的跨平台能力，可以编写桌面应用程序、Web应用程序、分布式系统和嵌入式系统应用程序等，是大数据工程师最喜欢的编程工具，因此，想学好大数据，掌握Java基础是必不可少的！

**2.Linux命令**

对于大数据开发通常是在Linux环境下进行的，相比Linux操作系统，Windows操作系统是封闭的操作系统，开源的大数据软件很受限制，因此，想从事大数据开发相关工作，还需掌握Linux基础操作命令。

**3. Hadoop**

Hadoop是大数据开发的重要框架，其核心是HDFS和MapReduce，HDFS为海量的数据提供了存储，MapReduce为海量的数据提供了计算，因此，需要重点掌握，除此之外，还需要掌握Hadoop集群、Hadoop集群管理、YARN以及Hadoop高级管理等相关技术与操作！

**4. 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，十分适合数据仓库的统计分析。对于Hive需掌握其安装、应用及高级操作等。

**5. Avro与Protobuf**

Avro与Protobuf均是数据序列化系统，可以提供丰富的数据结构类型，十分适合做数据存储，还可进行不同语言之间相互通信的数据交换格式，学习大数据，需掌握其具体用法。

**6.ZooKeeper**

ZooKeeper是Hadoop和Hbase的重要组件，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、组件服务等，在大数据开发中要掌握ZooKeeper的常用命令及功能的实现方法。

**7. HBase**

HBase是一个分布式的、面向列的开源数据库，它不同于一般的关系数据库，更适合于非结构化数据存储的数据库，是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，大数据开发需掌握HBase基础知识、应用、架构以及高级用法等。

**8.phoenix**

phoenix是用Java编写的基于JDBC API操作HBase的开源SQL引擎，其具有动态列、散列加载、查询服务器、追踪、事务、用户自定义函数、二级索引、命名空间映射、数据收集、行时间戳列、分页查询、跳跃查询、视图以及多租户的特性，大数据开发需掌握其原理和使用方法。

**9. Redis**

Redis是一个key-value存储系统，其出现很大程度补偿了memcached这类key/value存储的不足，在部分场合可以对关系数据库起到很好的补充作用，它提供了Java，C/C++，C#，PHP，JavaScript，Perl，Object-C，Python，Ruby，Erlang等客户端，使用很方便，大数据开发需掌握Redis的安装、配置及相关使用方法。

**10. Flume**

Flume是一款高可用、高可靠、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。大数据开发需掌握其安装、配置以及相关使用方法。

**11. SSM**

SSM框架是由Spring、SpringMVC、MyBatis三个开源框架整合而成，常作为数据源较简单的web项目的框架。大数据开发需分别掌握Spring、SpringMVC、MyBatis三种框架的同时，再使用SSM进行整合操作。

**12.Kafka**

Kafka是一种高吞吐量的分布式发布订阅消息系统，其在大数据开发应用上的目的是通过Hadoop的并行加载机制来统一线上和离线的消息处理，也是为了通过集群来提供实时的消息。大数据开发需掌握Kafka架构原理及各组件的作用和使用方法及相关功能的实现！

**13.Scala**

Scala是一门多范式的编程语言，大数据开发重要框架Spark是采用Scala语言设计的，想要学好Spark框架，拥有Scala基础是必不可少的，因此，大数据开发需掌握Scala编程基础知识！

**14.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，其提供了一个全面、统一的框架用于管理各种不同性质的数据集和数据源的大数据处理的需求，大数据开发需掌握Spark基础、SparkJob、Spark RDD、spark job部署与资源分配、Spark shuffle、Spark内存管理、Spark广播变量、Spark SQL、Spark Streaming以及Spark ML等相关知识。

**15.Azkaban**

Azkaban是一个批量工作流任务调度器，可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，可以利用Azkaban来完成大数据的任务调度，大数据开发需掌握Azkaban的相关配置及语法规则。

**16.Python与数据分析**

Python是面向对象的编程语言，拥有丰富的库，使用简单，应用广泛，在大数据领域也有所应用，主要可用于数据采集、数据分析以及数据可视化等，因此，大数据开发需学习一定的Python知识。

只有完整的学完以上技术，才能算得上大数据开发人才，真正从事大数据开发相关工作，工作才更有底气，升职加薪不成问题！