大数据概念

维基百科对大数据的定义：

大数据，或称海量数据，指的是所涉及的数据量规模巨大到无法通过人工，在合理的时间内达到截取、管理、处理、并整理成为人类所能解读的信息。

大数据是互联网发展到现阶段的一种表象或特征，在以云计算为代表的技术创新大幕的衬托下，这些原本很难收集和使用的数据开始容易被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

大数据的特征，4V，（Volume大量、Variety多样、Velocity高速、Value价值），第一，数据体量巨大，大数据的起始计量单位至少是PB（1024TB）；第二，数据类型繁多，比如网络日志、视频、图片、地理位置信息等等；第三，价值密度低，商业价值高。第四，处理速度快。最后这一点也和传统的数据挖掘技术有着本质的不同。

大数据相关的技术

虚拟化技术、分布式处理技术、海量数据的存储和管理技术、NoSQL、实时流数据处理等。

分布式处理技术

分布式处理系统可以将不同地点的或具有不同功能的或拥有不同数据的多台计算机用通信网络连接起来，在控制系统的统一管理下，协调地完成信息处理任务，这就是分布式处理系统的定义。

Hadoop的构成，Hadoop=HDFS（文件系统，数据存储技术相关）+HBase（数据库）+MapReduce（数据处理）+……Others。

Hadoop用到的一些技术有：

　　HDFS: Hadoop分布式文件系统(Distributed File System) － HDFS (HadoopDistributed File System)

　　MapReduce：并行计算框架

　　HBase: 类似Google BigTable的分布式NoSQL列数据库。

　　Hive：数据仓库工具，由Facebook贡献。

　　Zookeeper：分布式锁设施，提供类似Google Chubby的功能，由Facebook贡献。

　　Avro：新的数据序列化格式与传输工具，将逐步取代Hadoop原有的IPC机制。

　　Pig:大数据分析平台，为用户提供多种接口。

　　Ambari：Hadoop管理工具，可以快捷的监控、部署、管理集群。

　　Sqoop：用于在Hadoop与传统的数据库间进行数据的传递。