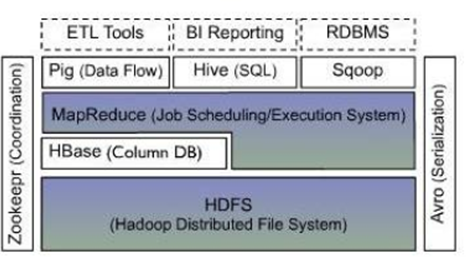
**The Hadoop Ecosystem**



hadoop文件系统（HDFS）是一个运行在普通的硬件之上的分布式文件系统，它和现有的分布式文 件系统有着很多的相似性，然而和其他的分布式文件系统的区别也是很明显的，HDFS是高容错性 的，可以部署在低成本的硬件之上，HDFS提供高吞吐量地对应用程序数据访问，它适合大数据集 的应用程序，HDFS放开一些POSIX的需求去实现流式地访问文件数据，HDFS开始是为开源的 apache项目nutch的基础结构而创建，HDFS是hadoop项目的一部分，而hadoop又是lucene的一 部分。

HBase 是一个分布式的、面向列的开源数据库，该技术来源于 Chang et al 所撰写的 Google 论文 “Bigtable ：一个结 构化数据的分布式存储系统 ” 。就像 Bigtable 利用了 Google 文件系统（ File System ）所提供的分布式数据存储一 样， HBase 在 Hadoop 之上提供了类似于 Bigtable 的能力。 HBase 是 Apache 的 Hadoop 项目的子项目。 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 . 另一个不同的是 HBase 基于列的而不是基于行的模式。

MapReduce是一种编程模型，用于大规模数据集（大于1TB）的并行运算。概念"Map（映射）"和"Reduce（化简）"，和他们的主要思想， 都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。他极大地方便了编程人员在不会分布式并行编程的情况下,将自己的程序运行在分布式系统上。 当前的软件实现是指定一个Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的Reduce（化简）函数，用来保证所有映射的键值对中的每一个共享相同的键组。

Pig是一个基于Hadoop的大规模数据分析平台，它提供的SQL-LIKE语言叫Pig Latin，该语言的编译器会把类SQL的数据分析请求转换为一系列经过优化处理的MapReduce运算。Pig为复杂的海量数据并行计算提供了一个简单的操作和编程接口。用Pig可以对TB级别海量的数据进行查询非常轻松，并且这些海量的数据都是非结构化的数据。

Hive是一个基于Hadoop的数据仓库平台。通过hive，我们可以方便地进行ETL的工作。hive定义了一个类似于SQL的查询语言：HQL，能 够将用户编写的QL转化为相应的Mapreduce程序基于Hadoop执行。

Hive 和Pig都是Hadoop中的项目，并且Hive和pig有很多共同点，但Hive还似乎有点数据库的影子，而Pig基本就是一个对MapReduce实现的工具(脚本)。两者都拥有自己的表达语言，其目的是将MapReduce的实现进行简化，并且读写操作数据最终都是存储在HDFS分布式文件系统上。

Sqoop是一个用来将Hadoop和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（例如 ： MySQL ,Oracle ,Postgres等）中的数据导入到Hadoop的HDFS中，也可以将HDFS的数据导入到关系型数据库中。

ZooKeeper是Hadoop的正式子项目，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、组服务等。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。Zookeeper是Google的Chubby一个开源的实现.是高有效和可靠的协同工作系统。

Avro是一个数据序列化系统，设计用于支持大 批量数据交换的应用。它的主要特点有：支持二进制序列化方式，可以便捷，快速地处理大量数据；动态语言友好，Avro提供的机制使动态语言可以方便地处理 Avro数据。