在大数据时代，传统的数据处理方法还适用吗？

**大数据环境下的数据处理需求**

大数据环境下数据来源非常丰富且数据类型多样，存储和分析挖掘的数据量庞大，对数据展现的要求较高，并且很看重数据处理的高效性和可用性。

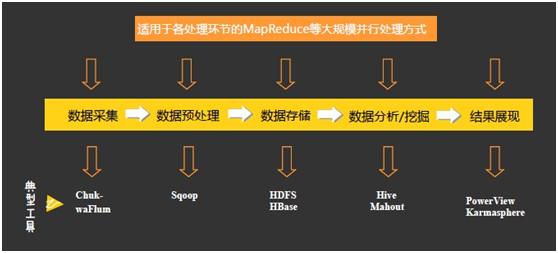
**传统数据处理方法的不足**

传统的数据采集来源单一，且存储、管理和分析数据量也相对较小，大多采用关系型数据库和并行数据仓库即可处理。对依靠并行计算提升数据处理速度方面而言，传统的并行数据库技术追求高度一致性和容错性，根据CAP理论，难以保证其可用性和扩展性。

传统的数据处理方法是以处理器为中心，而大数据环境下，需要采取以数据为中心的模式，减少数据移动带来的开销。因此，传统的数据处理方法，已经不能适应大数据的需求！

**大数据的处理流程包括哪些环节？每个环节有哪些主要工具？**

大数据的基本处理流程与传统数据处理流程并无太大差异，主要区别在于：由于大数据要处理大量、非结构化的数据，所以在各个处理环节中都可以采用MapReduce等方式进行并行处理。



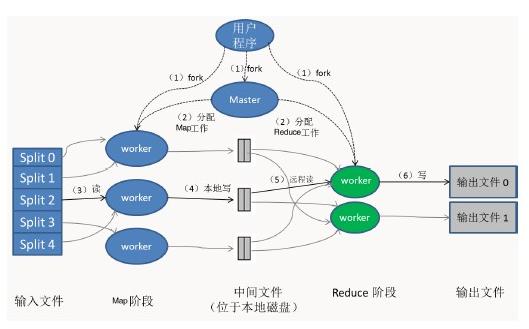
**大数据技术为什么能提高数据的处理速度？**

大数据的并行处理利器——MapReduce

大数据可以通过MapReduce这一并行处理技术来提高数据的处理速度。MapReduce的设计初衷是通过大量廉价服务器实现大数据并行处理，对数据一致性要求不高，其突出优势是具有扩展性和可用性，特别适用于海量的结构化、半结构化及非结构化数据的混合处理。

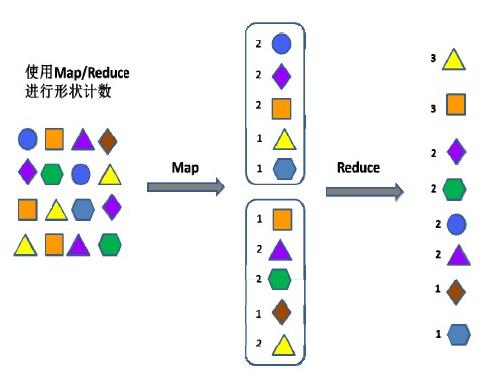
MapReduce将传统的查询、分解及数据分析进行分布式处理，将处理任务分配到不同的处理节点，因此具有更强的并行处理能力。作为一个简化的并行处理的编程模型，MapReduce还降低了开发并行应用的门槛。

MapReduce是一套软件框架，包括Map（映射）和Reduce（化简）两个阶段，可以进行海量数据分割、任务分解与结果汇总，从而完成海量数据的并行处理。



MapReduce的工作原理其实是先分后合的数据处理方式。Map即“分解”，把海量数据分割成了若干部分，分给多台处理器并行处理；Reduce即“合并”，把各台处理器处理后的结果进行汇总操作以得到最终结果。如右图所示，如果采用MapReduce来统计不同几何形状的数量，它会先把任务分配到两个节点，由两个节点分别并行统计，然后再把它们的结果汇总，得到最终的计算结果。

MapReduce适合进行数据分析、日志分析、商业智能分析、客户营销、大规模索引等业务，并具有非常明显的效果。通过结合MapReduce技术进行实时分析，某家电公司的信用计算时间从33小时缩短到8秒，而MKI的基因分析时间从数天缩短到20分钟。



说到这里，再看一看MapReduce与传统的分布式并行计算环境MPI到底有何不同？MapReduce在其设计目的、使用方式以及对文件系统的支持等方面与MPI都有很大的差异，使其能够更加适应大数据环境下的处理需求。



**大数据技术在数据采集方面采用了哪些新的方法**

**系统日志采集方法**

很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如Hadoop的Chukwa，Cloudera的Flume，Facebook的Scribe等，这些工具均采用分布式架构，能满足每秒数百MB的日志数据采集和传输需求。

**网络数据采集方法：对非结构化数据的采集**

网络数据采集是指通过网络爬虫或网站公开API等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

除了网络中包含的内容之外，对于网络流量的采集可以使用DPI或DFI等带宽管理技术进行处理。

**其他数据采集方法**

对于企业生产经营数据或学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。