信息化时代，数据爆炸，海量的数据创痛的

广义haddoop

狭义上： 就是apache的一个顶级项目：apahce hadoop

　　广义上: 就是指以hadoop为核心的整个大数据处理体系

Hadoop的核心组件：

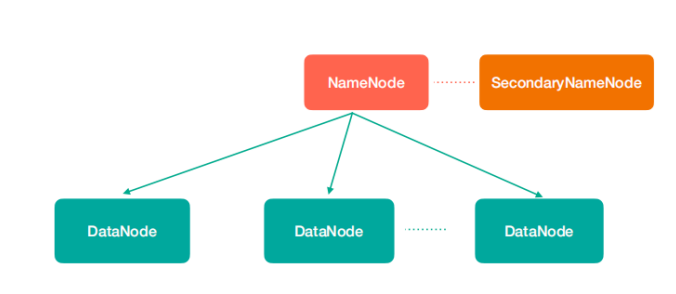
　　1）**Hadoop Common**：支持其他Hadoop模块的常用工具。

　　2)  **Hadoop分布式文件系统（HDFS™）**：一种分布式文件系统，可提供对应用程序数据的高吞吐量访问。

　　3)  **Hadoop YARN**：作业调度和集群资源管理的框架。

**4)  Hadoop MapReduce**：一种用于并行处理大型数据集的基于YARN的系统。





**NameNode（nn）：**存储⽂件的元数据，⽐如⽂件名、⽂件⽬录结构、⽂件属性（⽣成时间、副本

数、⽂件权限），以及每个⽂件的块列表和块所在的DataNode等。

**SecondaryNameNode（2nn）：**辅助NameNode更好的⼯作，⽤来监控HDFS状态的辅助后台程

序，每隔⼀段时间获取HDFS元数据快照。

**DataNode（dn）：**在本地⽂件系统存储⽂件块数据，以及块数据的校验

注意：NN，2NN,DN

Hadoop MapReduce：⼀个分布式的离线并⾏计算框架

拆解任务、分散处理、汇整结果

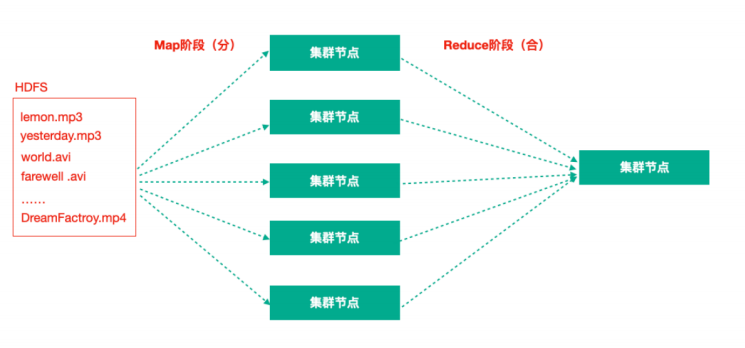
**MapReduce计算 = Map阶段 + Reduce阶段**

Map阶段就是“分”的阶段，并⾏处理输⼊数据；

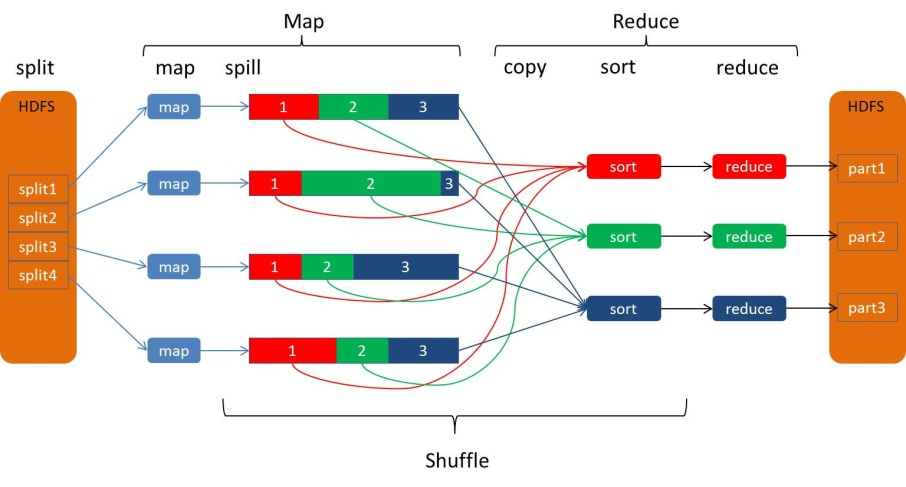
Reduce阶段就是“合”的阶段，对Map阶段结果进⾏汇总；

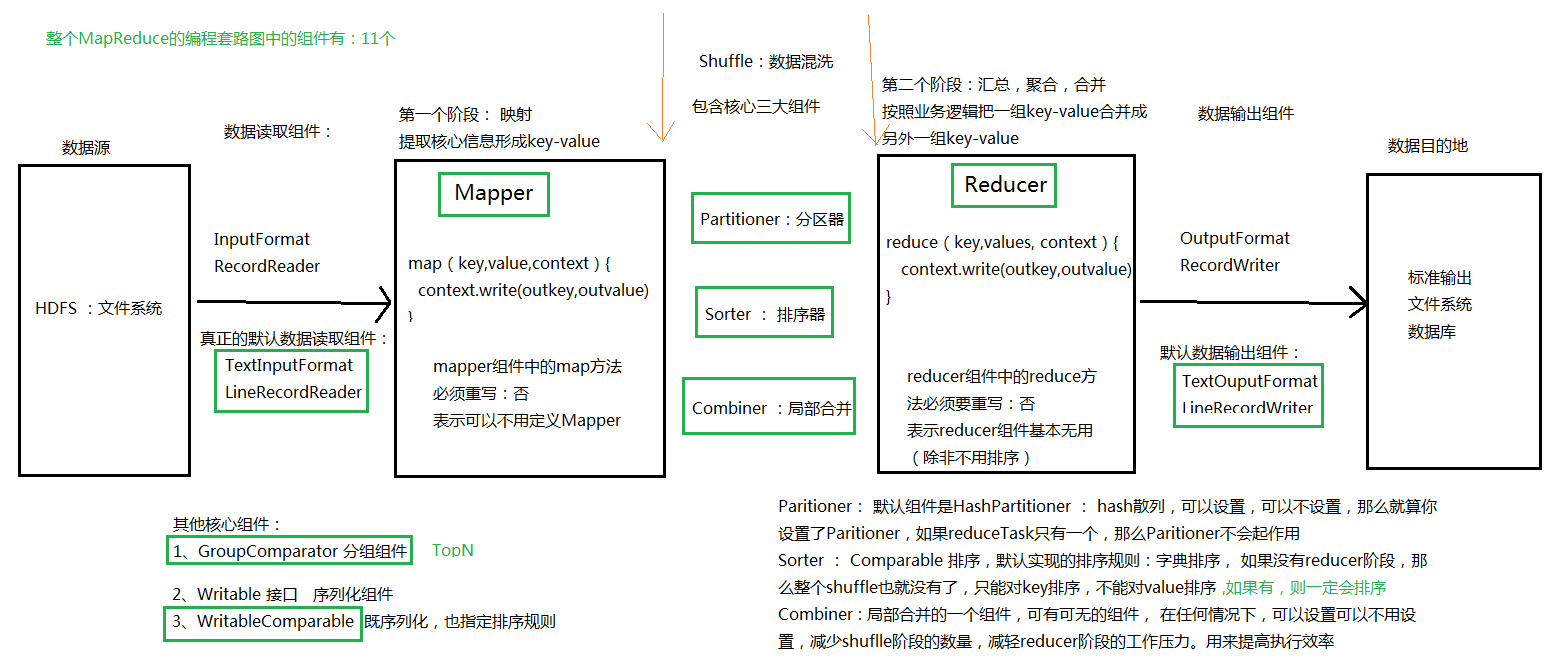
　（1）Mapper负责“分”，即把复杂的任务分解为若干个“简单的任务”来处理。“简单的任务”包含三层含义：一是数据或计算的规模相对原任务要大大缩小；二是就近计算原则，即任务会分配到存放着所需数据的节点上进行计算；三是这些小任务可以并行计算，彼此间几乎没有依赖关系。

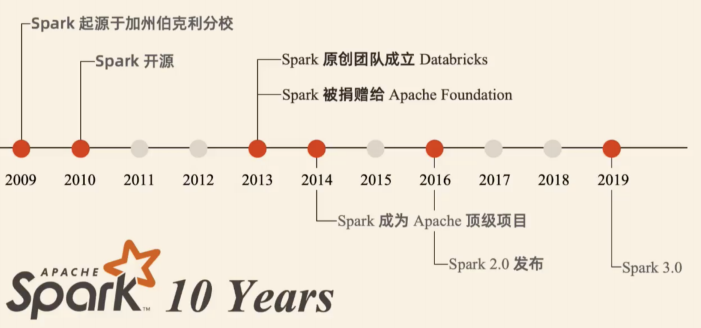
　　（2）Reducer负责对map阶段的结果进行汇总。至于需要多少个Reducer，用户可以根据具体问题，通过在mapred-site.xml配置文件里设置参数mapred.reduce.tasks的值，缺省值为1。



Fsdfs sdsds







Spark是当今大数据领域最活跃、最热门、最高效的大数据通用计算引擎

2009年诞生于美国加州大学伯克利分校AMP 实验室

2010年通过BSD许可协议开源发布

2013年捐赠给Apache软件基金会并切换开源协议到切换许可协议至

1

2

3

4

52014年2月，Spark 成为 Apache 的顶级项目

2014年11月, Spark的母公司Databricks团队使用Spark刷新数据排序世界记录

Spark 成功构建起了一体化、多元化的大数据处理体系。在任何规模的数据计算中，

Spark 在性能和扩展性上都更具优势

速度快。与 MapReduce 相比，Spark基于内存的运算要快100倍以上，基于硬

盘的运算也要快10倍以上。Spark实现了高效的DAG执行引擎，可以通过基于内存来高效处理数据流；

使用简单。Spark支持 Scala、Java、Python、R的API，还支持超过80种高级算

法，使用户可以快速构建不同的应用。而且Spark支持交互式的Python和Scala

的shell，可以非常方便地在这些shell中使用Spark集群来验证解决问题的方法；

通用。Spark提供了统一的解决方案。Spark可以用于批处理、交互式查询

(Spark SQL)、实时流处理(Spark Streaming)、机器学习(Spark MLlib)和图计算

(GraphX)。这些不同类型的处理都可以在同一个应用中无缝使用。Spark统一的

解决方案非常具有吸引力，企业想用统一的平台去处理遇到的问题，减少开发和

维护的人力成本和部署平台的物力成本；

兼容好。Spark可以非常方便地与其他的开源产品进行融合。Spark可以使用

YARN、Mesos作为它的资源管理和调度器；可以处理所有Hadoop支持的数

据，包括HDFS、HBase和Cassandra等。这对于已经部署Hadoop集群的用户特

别重要，因为不需要做任何数据迁移就可以使用Spark的强大处理能力。Spark

也可以不依赖于第三方的资源管理和调度器，它实现了Standalone作为其内置

的资源管理和调度框架，这样进一步降低了Spark的使用门槛，使得所有人都可

以非常容易地部署和使用Spark。此外，Spark还提供了在EC2上部署

Standalone的Spark集群的工具

MapReduce的不足：

表达能力有限

磁盘IO开销大

延迟高

任务之间的衔接有IO开销

在前一个任务执行完成之前，后一个任务无法开始。难以胜任复杂的、多阶

段计算任务

