**电工导实验报告10**

**一、实验目的**

1. **利用hadoop进行mapreduce**

**二、实验内容**

1. **什么是hadoop**
2. **mapreduce编程模型**

**三、实验环境**

1. **ubuntu + hadoop**

**四、实验原理**

**Apache Hadoop**是一款支持数据密集型[分布式应用](https://zh.wikipedia.org/w/index.php?title=%E5%88%86%E4%BD%88%E5%BC%8F%E6%87%89%E7%94%A8&action=edit&redlink=1)并以Apache 2.0许可协议发布的[开源](https://zh.wikipedia.org/wiki/%E5%BC%80%E6%BA%90)[软件框架](https://zh.wikipedia.org/wiki/%E8%BB%9F%E9%AB%94%E6%A1%86%E6%9E%B6)。它支持在商品硬件构建的大型集群上运行的应用程序。Hadoop是根据[Google公司](https://zh.wikipedia.org/wiki/Google%E5%85%AC%E5%8F%B8)发表的[MapReduce](https://zh.wikipedia.org/wiki/MapReduce)和[Google文件系统](https://zh.wikipedia.org/wiki/Google%E6%AA%94%E6%A1%88%E7%B3%BB%E7%B5%B1)的论文自行实现而成。

Hadoop框架透明地为应用提供可靠性和数据移动。它实现了名为MapReduce的[编程范式](https://zh.wikipedia.org/wiki/%E7%BC%96%E7%A8%8B%E8%8C%83%E5%BC%8F)：应用程序被分区成许多小部分，而每个部分都能在集群中的任意节点上运行或重新运行。此外，Hadoop还提供了分布式文件系统，用以存储所有计算节点的数据，这为整个集群带来了非常高的带宽。MapReduce和分布式文件系统的设计，使得整个框架能够自动处理节点故障。它使应用程序与成千上万的独立计算的电脑和PB级的数据。现在普遍认为整个Apache Hadoop“平台”包括Hadoop内核、MapReduce、Hadoop分布式文件系统（HDFS）以及一些相关项目，有Apache Hive和Apache HBase等等。

Hadoop是一个能够让用户轻松架构和使用的[分布式计算](http://baike.baidu.com/subview/30655/11213586.htm)平台。用户可以轻松地在Hadoop上开发和运行处理海量数据的[应用程序](http://baike.baidu.com/view/330120.htm)。它主要有以下几个优点：

1. 高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖。

2. 高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

3. 高效性。Hadoop能够在节点之间动态地移动数据，并保证各个节点的[动态平衡](http://baike.baidu.com/view/1603996.htm)，因此处理速度非常快。

4. 高容错性。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

低成本。与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低。

Hadoop带有用[Java](http://baike.baidu.com/view/29.htm)语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。Hadoop 上的[应用程序](http://baike.baidu.com/view/330120.htm)也可以使用其他语言编写，比如 [C++](http://baike.baidu.com/view/824.htm)。

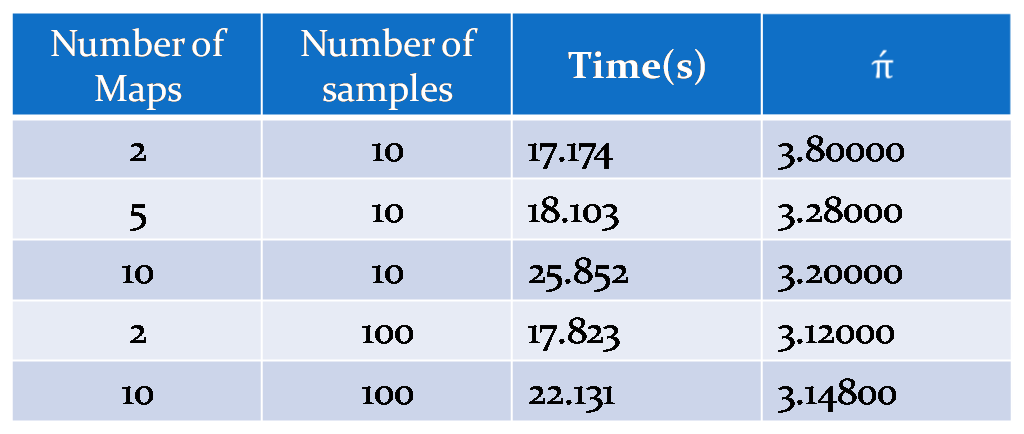
**hadoop大数据处理的意义**

Hadoop得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载(ETL)方面上的天然优势。Hadoop的分布式架构，将大数据处理引擎尽可能的靠近存储，对例如像ETL这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接走向存储。Hadoop的MapReduce功能实现了将单个任务打碎，并将碎片任务(Map)发送到多个节点上，之后再以单个数据集的形式加载(Reduce)到数据仓库里。

**五、Mini Exercise**

**Exercise 1:**

practise using basic hadoop command and fill in the following table



观察表格前三行，在sample number 一定的情况下，map number越多，运行时间越长，π的精度越高。

通常情况下，运行时间和精度是成正比的。

**Exercise 2:**

Work out a solution to make the computed π approximate the 5th digit after the decimal dot correctly.

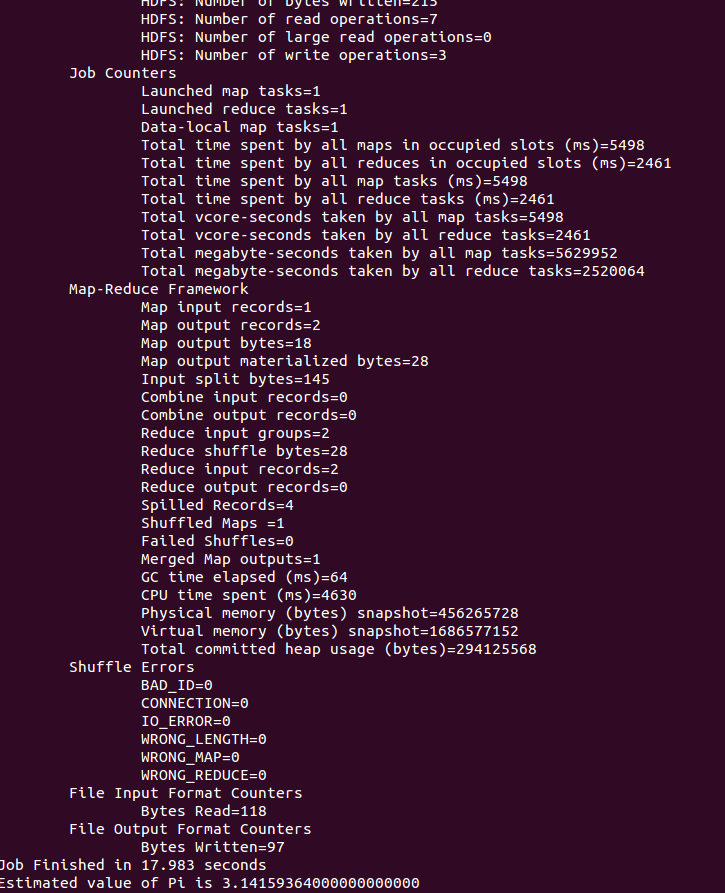
通过以下代码对π的值进行估算：

*$ hadoop jar /usr/local/Hadoop/share/Hadoop/mapreduce/Hadoop-mapreduce-examples-2.2.0.jar pi <nMaps> <nSamples>*

\*where <nMaps> is the number of mapper jobs and <nSamples> is the number of samples

经过修改Number of Maps 和 Number of samples, 我们可以得到不同的精度。

运行结果如下：



小数点后五位精度

实验结果是3.1415936400000000000， 实验结果与π的小数点后的五位是吻合的。

**七、实验总结**

Hadoop的作业并不是我想象的那么一帆风顺，从一开始安装虚拟机，到最后利用Hadoop解决一些问题，中间历经了各种困难。但是从实验中我的确看到了Hadoop将程序分解为多个小部分分别进行处理的强大功能。在当今这个数据爆炸的时代，Hadoop等构架的运用能大幅提高效率，而且能为许多大型的工程提供基础。

**八、结语**

一周的电工导C迎来尾声。这一学期我们学了很多，而且有用的东西，比如开始的爬取网页，通过web.py进行网页的开发，再到后来对图片进行处理，canny检测，sift匹配等，让我们接触到了更多的领域，也对我们之后研究方向的寻找提供了帮助。

最后也感谢张娅老师和何大治老师带来的精彩课程，也感谢助教们的耐心解答。

F1403023 5140309534 韩坤言