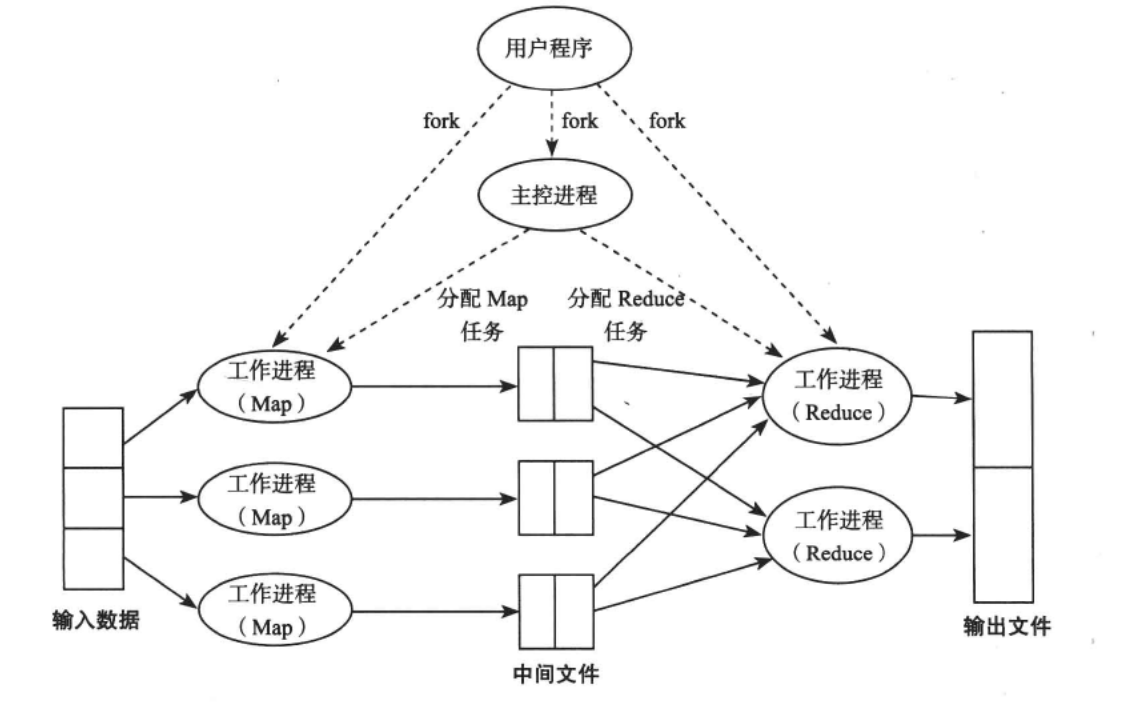
# 概述

提到数据库，大多数人首先想到的就是MapReduce。MapReduce使得普通程序员可以在不了解分布式底层细节的前提下开发分布式程序。使用者只需要编写两个称为Map和Reduce的函数即可，MapReduce框架会自动处理数据划分、多机并行执行、任务之间的协调，并且能够处理某个任务执行失败或者机器出现故障的情况。

# 原理

MapReduce的执行流程如下所示：



MapReduce框架包含三种角色：主控进程（Master）用于执行任务划分、调度、任务之间的协调等；Map工作进程（Map Worker，简称Map进程）以及Reduce工作进程（Reduce Worker，简称Reduce进程）分别用于执行Map任何和Reduce任务。

# 扩展

MapReduce框架有效地解决了海量数据的离线批处理问题，在各大互联网公司得到广泛的应用。事实已经证明了MapReduce的巨大影响力，以至于引发了一系列的扩展和改进。这些扩展包括：

1. Google Tenzing：基于MapReduce模型构建SQL执行引擎，使得数据分析人员可以直接通过SQL语言处理大数据。
2. Microsoft Dryad：将MapReduce模型从一个简单的两步工作流扩展为任何函数集的组合，并通过一个有向无环图来表示函数之间的工作流。
3. Google Pregel：用于图模型迭代计算，这种场景下Pregel的性能远远好于MapReduce。

## Google Tenzing

## Microsoft Dryad

## Google Pregel