数据预处理

1. 数据基本性质

数据保存在basic\_table中

1. 一共有数据820073条数据，解析出的数据的数目是820053条数据， 保存在basic\_table中
2. 时间跨度是：2014年9月17日23时27分19秒 到 2017年11月27日17时11分14秒
3. 无法处理的数据：数据本身不规范
4. 用户与其作业数目

数据保存在user\_jobs中

1. 一共有246个用户
2. 最大的运行任务数目是109756， 最小的运行任务数是1， 均值是3333，

中值是285， 有95%的用户的任务数目小于10000



1. 用户与其jstatus

数据保存在user\_jstatus中

jStatus = 32表示exit, jStatus = 64表示done

1. 一共有246个用户，其中的213个用户既有jstatus=32的状态也有jstatus=64的状态
2. 有20个用户只有jstatus=32的状态
3. 有13个用户只有jstatus=64的状态
4. 大部分的数据jStatus=64的任务数比jStatus=32的任务数多
5. 任务时间分布

程序运行时间 = event time - start time

（用户时间+系统时间）/运行cpu核心数， （用户时间+系统时间）

如何starttime < submittime 那么就将 starttime = submittime;

1. 最大值：15135902秒
2. 最小值：0秒
3. 平均时间：18139.49秒

有819977条数据的运行时间小于4000000，占比为99.99%。有788287条数据的运行时间小于86400秒（也就是一天），占比为96.13%。有766967条数据的运行时间小于43200秒（也就是半天），占比是93.53%。有743014条数据的运行时间小于21600（也就是6个小时），占比是90.61%。有716503条数据的运行时间小于10800（也就是3个小时），占比87.37%。有634096条数据的运行时间小于3600秒（也就是一个小时），占比是77.32%。因此我们可以将任务划分为长任务（大于1天），中等任务（大于1个小时），短任务（小于一个小时）。

1. 不同用户长中短任务分布图

（5）将中等任务和长任务合并之后，称之为长任务。它的分布图如下：

1. 用户早中晚任务分布

（1）6点到14点算是早上，14点到22点算是中午，22点到第二天早上6点算是晚上

从图上可以看出不同用户的早中晚任务分布是不同的。

1. 早中晚长中短任务分布图
2. 长中短任务早中晚分布图
3. 长短任务早中晚分布图

之前直观的感觉是长任务应该在晚上进行，但是从这个统计结果上得不到这个结论。

1. 工作日和周末任务情况
2. 不同用户工作日和周末任务占比

可以看出大部分的用户还是喜欢在工作日工作，少部分的用户喜欢在周末工作。

1. 工作日和周末长中短任务分布

看起来差不多，总体上只能看出周末的任务数会比工作日的任务数少很多，但是长中短任务的占比还是类似的。

1. 不同用户工作日和周末长中短任务分布

不同的用户它在工作日和周日可能进行的任务类型是不同的。

1. 新用户到来时间

从图上可以看出新用户的到来时间基本是呈现线性的