数据预处理-说明版

1. 数据基本性质
2. **数据来源**：中科大超算中心运行的历史任务。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

**数据格式**：原始的数据是txt格式，为了方便分析，利用python转存到了Mysql数据库（usermodel），并保存在表basic\_table中。

**数据规模**：原始的数据有820073条，但是由于有些数据不规范，解析之后保存在数据库中的数据缺失20条，现有820053条。

1. 用户与其作业数目

数据保存在表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。

（3）有159个用户执行的任务数超过100。



表1 用户运行任务数目统计表

图1 不同用户运行任务数目散点图

从图1可以看出不同用户运行任务数目大致呈现长尾分布

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

图2 不同用户不同运行状态分布图

从图2可知大部分的用户jStatus=64的任务数比jStatus=32的任务数多

1. 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

图3 所有运行任务的时间分布散点图

从图5可知，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

图4 不同用户长中短任务数量占比

从图4可知，不同用户运行长中短任务的占比有很大的不同，有些用户只运行长任务，有些用户只运行短任务。

1. 用户早中晚任务分布
2. 用户24小时任务量分布

图5 不同用户24小时任务分布图

从图5 可以看出一天24小时大概可以分成三个区间段：0-8,8-16,16-24

（2）8点到16点算是早上，16点到24点算是中午，24点到第二天早上8点算是晚上

图6 用户早中晚任务占比

从图6可以看出不同用户的早中晚任务分布是不同的。

1. 工作日和周末任务情况
2. 不同用户周内的任务占比

图7 不同用户周内任务提交量

一峰突起的数据是用户ID为1323的用户，它之所以将任务的提交量都集中在周二，是因为它在2017年10月10日这一天批量提交了59669条数据。

图8 删除1323用户的批处理任务后的周内任务分布

从图8可以看出不同的用户在周内的任务提交量是不同的

1. 不同用户工作日和周末任务占比

图9 不同用户工作日和周末任务量占比

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. 新用户到来时间

新用户的定义：首次提交运行任务的用户

图10 新用户到来的时间折线图

从图10可以看出新用户的到来时间基本是呈现线性的

1. 批处理任务识别
2. 如果将单位小时内提交的任务数超过一百任务看做是一个批处理任务的话，那么这样的批处理任务有867个；非批处理任务有320705个。总数为321572个。
3. 从以上数据可以看出原始数据中有很多是批处理任务，并且有些用户特别喜欢提交批处理任务。
4. 以上的批处理任务的定义不太准确，需要进一步的完善。