

刘一鸣,周自江,远芳,等. ARQCS 启动策略及其与资源消耗的关系. 应用气象学报,2014,25(4):505-512.

## ARQCS 启动策略及其与资源消耗的关系

刘一鸣\* 周自江 远芳 阮宇智 何文春 孙超 刘媛媛

(国家气象信息中心,北京 100081)

### 摘 要

利用 2012 年 4 月 1 日—9 月 30 日 IBM P570 高性能计算环境 Oracle 11g 数据库平台对全国自动气象站观测资料实时质量控制系统 (ARQCS) 的运行监控数据,探讨了 ARQCS 的启动策略及其与资料解析入库率、ARQCS 的 CPU 耗时、服务时效之间的关系。结果表明:自动气象站资料的解析入库效率呈“几”字型分布,每个观测时次的第 5—20 分钟入库率方差较大,是制约 ARQCS 质量控制时效的主要时间段。设置观测资料入库率不低于 95% 为首次启动条件,不仅比传统的第 15 分钟定时启动提前了 20.6 s,而且首次启动时观测资料入库率不低于 95% 的概率从 66.38% 提升至 95.83%。第 20 分钟后入库率仅增加 1.36%,在此设置首次质量控制的强制启动点,可保证局部异常延时的资料服务时效。动态启动策略使 ARQCS 的启动次数由 5 次降为 2 次,平均每日节约 CPU 时间 391 min。

**关键词:** 自动气象站观测资料; 实时质量控制; 启动策略; 计算资源

### 引 言

全国自动气象站观测资料实时质量控制系统 (简称为 ARQCS) 是一套基于 IBM P570 高性能计算环境 Oracle 11g 数据库平台的实时气象资料业务应用系统,包括对全国 3 万多个自动气象站逐小时观测资料的解析入库、质量控制、存储管理和共享服务等流程节点。随着自动气象站观测资料在实时气象业务服务中应用的广泛和深入<sup>[1-5]</sup>,人们对 ARQCS 效能的要求也越来越高。

ARQCS 始建于 2009 年,当时主要以单一的降水资料为例,面向实时气象预报服务需求,开展自动气象站观测资料质量控制的业务试验<sup>[6-8]</sup>。因质量控制的气象要素单一,算法相对简单,ARQCS 对计算资源的消耗并不明显,同时,考虑到预报服务的资料时效需求,ARQCS 采用每小时后第 15, 25, 35, 45, 55 分钟共计 5 次静态定时启动策略,使资料的时效和完整性尽可能匹配合理。但随着 2010 年 ARQCS 向气温、气压、降水、湿度、风向、风速 6 个

气象要素拓展<sup>[9-10]</sup>、2011 年和 2012 年质量控制算法的两次升级以及全国自动气象站数量的不断增加,ARQCS 中的数学逻辑愈发复杂,运算量急速增多,ARQCS 每次启动会消耗较大的计算资源。与此同时,ARQCS 的 5 次定时启动意味着数据库系统要进行 5 次数据更新。相应地,资料用户有可能会启动 5 次数据库检索访问,数据库系统承载的检索访问压力显著增加,检索时效必然降低<sup>[11-13]</sup>。此外,随着全国地面报文传输改革的推进,ARQCS 需要进行实时质量控制的要素多达 11 类 158 项,质量控制算法也必须进一步升级。如何在现有计算环境与数据库平台不变的前提下,继续使资料质量控制能力和服务时效保持较高水平,ARQCS 的调度策略显得非常重要。

IBM P570 高性能计算环境 Oracle 11g 数据库平台具有完备的数据安全性和完整性控制机制<sup>[14-16]</sup>,系统日志完整记录了自动气象站观测资料的解析入库时效和质量控制时效。初步分析表明,虽然局部样本受观测系统或通信系统延时等因素影响而呈波动状态,但对于全国 3 万多个自动气象站

2013-12-02 收到, 2014-05-05 收到再改稿。

资助项目: 中国气象局气象关键技术集成与应用面上项目 (CMAGJ2013M67), 中国气象局气象关键技术集成与应用重点项目 (CMAGJ2013Z01)

\* email: yimingliu@aliyun.com

的总体样本来说,资料解析入库率随时间变化具有一定规律,很多有关 ARQCS 调度策略的研究<sup>[17-19]</sup>值得借鉴。本文旨在利用 2012 年 4 月 1 日—9 月 30 日数据库平台对 ARQCS 的运行监控数据,探讨实时质量控制系统中资料解析入库效率、质量控制的 CPU 耗时、服务时效之间的内在关系,尝试建立 ARQCS 的动态启动策略,以提升其质量控制效能。

## 1 自动气象站观测资料解析入库时效的统计学特征

### 1.1 资料与方法

目前,ARQCS 对全国自动气象站观测资料的质量控制采用逐观测时次(当前为逐小时)清算机制。当 ARQCS 接收到气象通信系统实时转发的该时次自动气象站观测数据文件时,立即自动解析入库;随后 ARQCS 采用定时启动的机制,基于已成功

完成解析入库的自动气象站观测要素数值完成质量控制方法的判定(图 1);同时,数据库系统时钟会记录下资料入库时间。因此,在每个时次后的 60 min 时间步长内,通过系统时间记录可以计算得到该时次全国自动气象站观测资料在任意时刻的入库时效。为便于分析,本文选取入库率作为资料入库时效的定量表征指标,并以分钟作统计单元,这样第  $t$  分钟的单分钟入库率( $R_t$ )和截至第  $t$  分钟的累计入库率( $S_t$ )分别为

$$R_t = \frac{N_t}{N_0} \times 100\%, t = 1, 2, \dots, 60; \quad (1)$$

$$S_t = \frac{\sum_{i=1}^t N_i}{N_0} \times 100\%, t = 1, 2, \dots, 60. \quad (2)$$

其中, $N_t, N_i$  分别为第  $t, i$  分钟入库的自动气象站数量, $N_0$  为应入库的自动气象站数量。本文  $N_0$  为固定值(即  $N_0 = 31814$  个站)。

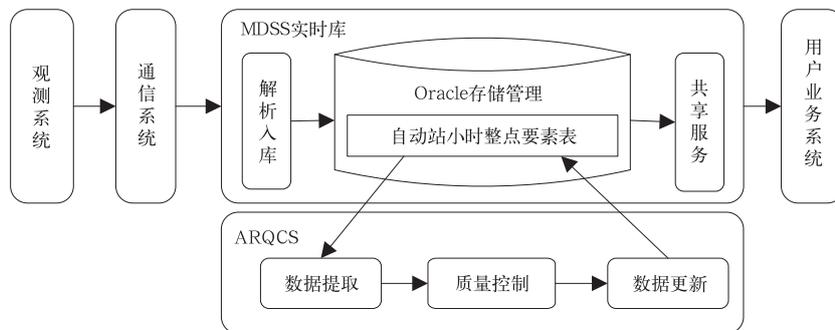


图 1 ARQCS 数据流程图

Fig. 1 ARQCS data flow chart

本文使用 2012 年 4 月 1 日 00:00—9 月 30 日 23:00(世界时,下同)IBM P570 高性能计算环境 Oracle 11g 数据库平台的原始时钟记录。为排除通信系统、数据库系统故障等对 ARQCS 的影响,本文剔除了 2012 年 4 月 16 日、5 月 5 日和 6 月 18 日等 9 d 的样本数据,这样有效小时样本为 4176 个,利用式(1)和式(2)计算得到了 250560 分钟点的单分钟入库率和累计入库率。

此外,为探讨 4176 个有效小时样本入库率的规律性,本文还通过式(3)和式(4)计算出第  $t$  分钟的平均单分钟入库率( $\bar{R}_t$ )和截至第  $t$  分钟的平均累计入库率( $\bar{S}_t$ ),其中  $M$  为固定值 4176。

$$\bar{R}_t = \frac{\sum_{j=1}^M R_{tj}}{M}, t = 1, 2, \dots, 60; \quad (3)$$

$$\bar{S}_t = \frac{\sum_{j=1}^M S_{tj}}{M}, t = 1, 2, \dots, 60. \quad (4)$$

### 1.2 入库时效的统计分布特征

数据解析入库是质量控制的前提(图 1),把握数据入库规律对制定质量控制启动策略具有重要意义。图 2a 为 4176 个有效小时样本的平均单分钟入库率 $\bar{R}_t$ ,在每个时次的前 17 min 平均单分钟入库率变化最为明显,20 min 之后入库率基本保持不变,接近于零。结合业务系统的实际运行情况来看,前

17 min 的单分钟入库率变化可以分成 3 个阶段: 第 1 阶段(第 1—4 分钟), 系统资源相对充足, 通信系统来报数据随时能够入库, 单分钟入库率随着来报数量的增加接近线性增长; 第 2 阶段(第 5—10 分

钟), 来报相对集中, 系统资源遇到瓶颈, 出现数据等待入库的情况, 入库效率基本不变; 第 3 阶段(第 11—17 分钟), 随时间向后推移来报数据越来越少, 单分钟入库率表现为线性下降趋势, 最终接近于零。

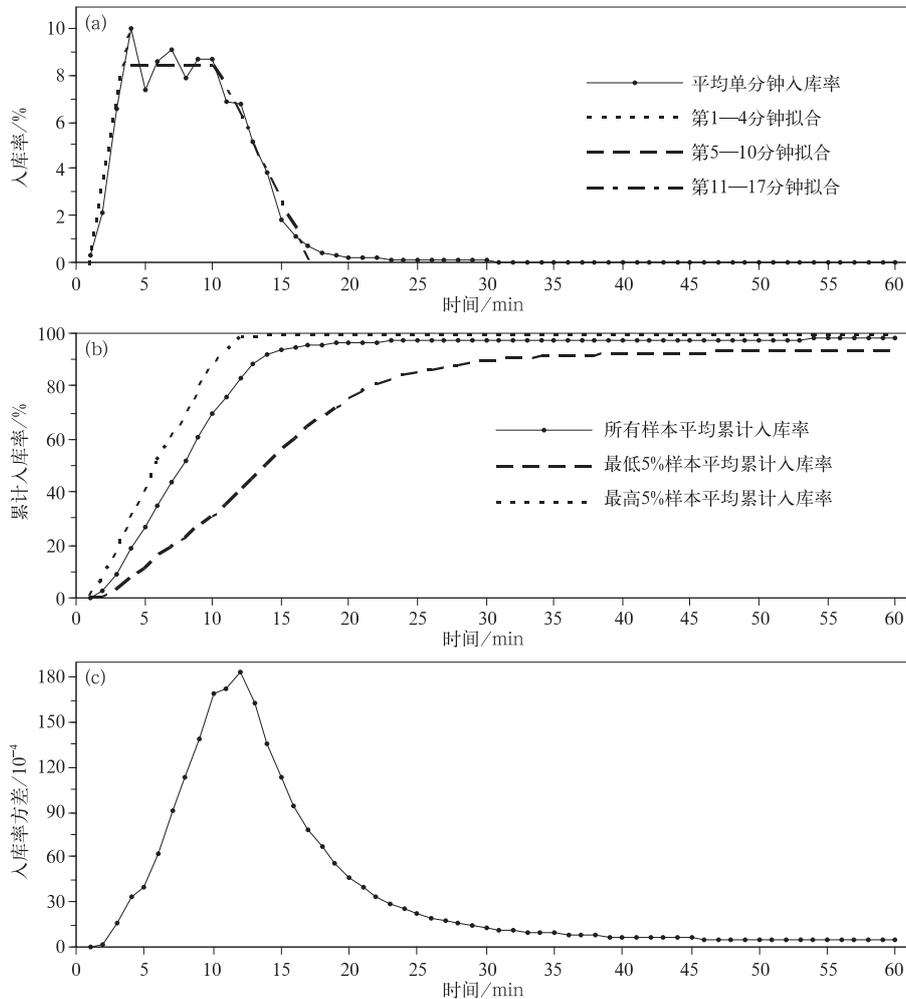


图 2 1 h 内入库率分布

(a) 平均单分钟入库率  $\bar{R}_i$  及第 1—4 分钟、第 5—10 分钟、第 11—17 分钟拟合曲线, (b) 平均累计入库率  $\bar{S}_i$ 、最低和最高 5% 的样本平均累计入库率, (c) 单分钟入库率方差

Fig. 2 Distribution of entry rate

(a) averaged entry rate of all samples for every minute and fitted curves from 1st to 4th minute, from 5th to 10th minute and from 11th to 17th minute, (b) averaged accumulated entry rate of all samples, the lowest 5% and the highest 5% of all samples, (c) variance of entry rate for every minute

从平均累计入库率  $\bar{S}_i$  分布曲线(图 2b)可以看到, 累计入库率在 1 h 内由 0 逐渐增加到 97.80%, 第 14 分钟时超过 90%, 第 17 分钟时超过 95%, 随后基本保持稳定, 从第 20—60 分钟仅微增 1.36%。图 2b 还给出累计入库率最高 5% 和最低 5% 样本的平均值。

对比分析表明, 在第 4—21 分钟二者相差超过 20%, 第 12 分钟时差距最大, 差值达到 58%, 而在第 30 分钟之后二者的差距相对稳定地维持在 7% 左右。与此对应的是较大的入库率方差(图 2c)出现在第 5—20 分钟, 最大值也出现在第 12 分钟, 第 35 分钟之后

方差基本保持不变。图 2a 和图 2c 均表明第 5—20 分钟入库率最不稳定,数据库系统的瓶颈效应明显,受观测系统或通信系统局部异常延时影响,不同时段之间出现较大差异的可能性较大,而在第 35 分钟之后不同时段累计入库率  $S_t$  平均稳定在 97% 的水平,且时段之间的差异较小。

## 2 基于资料完整性和服务时效的 ARQCS 动态启动策略

### 2.1 初次启动时间的确定

原静态启动策略规定 ARQCS 首次启动时间为第 15 分钟。虽然首次质控累计入库率平均可达 93.78%,但在最差的情况下只有 55% 左右(图 2b),并且方差较大(图 2c),表明在该时刻单分钟入库率  $R_t$  变化剧烈,出现较低累计入库率的可能性较高;第 20 分钟后  $R_t$  稳定在较低水平(图 2a),累计入库率  $S_t$  相对稳定(图 2b)。对第 15—20 分钟中的每分钟以及第 20 分钟之后有代表性的几个时间点(选取了原静态启动策略下的后 4 次启动时间点及 1 h 内最后的第 60 分钟)开展  $S_t$  的分段重点分析(表 1)表明:在第 15 分钟时,累计入库率  $S_{15}$  超过 95% 的样本只占总样本量的 66.38%, $S_{15}$  在 90%~95% 之间的样本占 21.19%, $S_{15}$  低于 90% 样本约占 12%。在随后的 5 min 内  $S_t$  超过 95% 的样本显著增加,低于 95% 的样本逐渐下降,第 16 分钟时  $S_t$  在 95% 以上的样本变化最为明显,增加到 74.52%,较前一分钟增加了 8.14%,第 17—20 分钟  $S_t$  在 95% 以上的样本从 79.89% 增加到 84.89%。

表 1 不同时间累计入库率  $S_t$  分段出现频次占总样本量的比例(单位:%)

Table 1 The proportion of accumulated entry rate at different time(unit:%)

时间/ min	$S_t$ 分段比例				
	[0, 50%)	[50%, 80%)	[80%, 90%)	[90%, 95%)	[95%, 100%]
15	1.44	4.14	6.85	21.19	66.38
16	1.15	3.42	4.07	16.83	74.52
17	1.01	2.68	2.63	13.79	79.89
18	0.89	2.30	1.84	12.72	82.26
19	0.81	1.72	1.53	12.17	83.76
20	0.69	1.53	1.15	11.73	84.89
25	0.19	0.96	0.55	10.58	87.72
35	0.10	0.12	0.19	9.56	90.04
45	0.05	0.05	0.10	8.88	90.92
55	0.05	0.02	0.00	8.36	91.57
60	0.02	0.05	0.00	8.07	91.86

上述分析表明,在第 15 分钟时累计入库率  $S_{15}$  在 95% 以上的样本只有六成左右,在随后的几分钟内常出现大量数据入库的情况,原静态启动策略将初次启动时间固定在第 15 分钟,很可能造成第 15 分钟后几分钟内即可完成入库的部分数据未参加首次质量控制。与此同时,95% 的累计入库率是第 15 分钟后几分钟内比较有望达到的一个入库率高值。所以,新启动策略下将累计入库率  $S_t$  达到 95% 的时间点确定为 ARQCS 的首次启动时间。

### 2.2 强制启动时间的时效依据

为探讨每小时入库率超过特定临界值的可能性,本文定义累计入库率在第  $t$  分钟首次超过临界值  $m$  (单位:%) 的概率 ( $P_{m,t}$ ) 及其累计值 ( $Q_{m,t}$ ),即累计入库率在第  $t$  分钟前(含第  $t$  分钟)超过  $m$  的概率:

$$P_{m,t} = \frac{n_{m,t}}{M} \times 100\%, t = 1, 2, \dots, 60; \quad (5)$$

$$Q_{m,t} = \frac{\sum_{i=1}^t n_{m,i}}{M} \times 100\%, t = 1, 2, \dots, 60. \quad (6)$$

其中, $n_{m,t}$  表示累计入库率  $S_t$  在第  $t$  分钟首次超过临界值  $m$  的样本量, $M$  为固定值 4176。最早在第 11 分钟  $P_{95,t}$  大于零值(图 3a),累计入库率  $S_{11}$  达到 95%。并且在图 3b 中,第 20 分钟  $Q_{95,20}$  为 84.89%,这表明有超过八成把握首次质量控制发生在第 20 分钟结束之前,同时保证首次启动时  $S_t$  在 95% 以上。

由图 3b 可知,即使在第 60 分钟, $Q_{95,60}$  也只有 91.86%,这意味着有约 8.14% 的样本在 1 h 之内  $S_t$  达不到新启动策略 95% 的启动标准。第 25 分钟时累计入库率在 90%~95% 之间以及超过 95% 的比例与第 20 分钟时无显著差别(表 1),随后直到第 60 分钟该比例变化相对不明显,且在第 20 分钟之后单分钟入库率  $R_t$  较为稳定(图 2a),累计入库率超过 95% 的概率  $Q_{95,t}$  相对平缓(图 3b),同时考虑到下游实时业务用户对服务时效的最大容忍程度,本文将强制启动 ARQCS 的时间设为第 20 分钟。

### 2.3 ARQCS 的逐小时自动清算机制

从表 1 还可发现,第 25 分钟之后不同时刻的累计入库率  $S_t$  差别较小,结合图 2a 与图 2c 可知,在第 25 分钟之前绝大部分的数据均已完成入库,在第 25—60 分钟的半个多小时里只有个别数据零星入库。这就意味着静态质量控制启动策略在第 25,

35,45,55 分钟的 4 次质量控制过程会对基本相同的已入库数据进行多次重复计算,这会造成较大的计算资源消耗,同时用户检索的多次跟进会造成数据库系统的工作压力随之增加。

同时由图 2b 可以发现,在第 20—60 分钟的 40 min 时间里,还会有平均 1.36% 的数据在 ARQCS 首次启动之后完成入库,而这一比例在最坏情况下高达 17.90%。考虑到目前全国考核自动气象

站已达 3 万余站,首次启动后入库的站数所占比例虽小,但每份站点数据对于提高资料完整性、开展预报与服务均有意义。因此,为了尽可能保证数据的完整性,同时减少重复计算造成的资源浪费,本文保留原有启动策略中的最后一次启动,即在第 55 分钟时完成 ARQCS 的逐小时自动清算,相应的累计入库率平均情况下为 97.77%,最坏情况下为 92.83%,最好情况下为 99.01%。

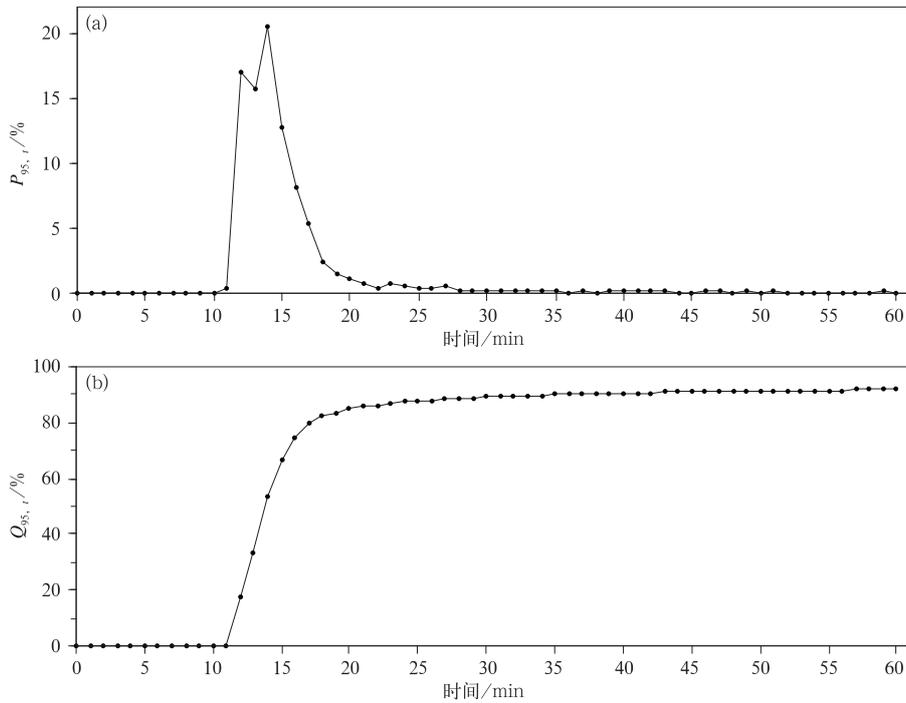


图 3 累计入库率  $S_t$  首次超过 95% 的样本量占总样本量的比例  $P_{95,t}$  (a) 及其累计值  $Q_{95,t}$  (b)

Fig. 3 The proportion of different time when the accumulated entry rate reaches 95% for the first time in an hour(a) and its accumulated value(b)

### 3 ARQCS 动态启动策略的业务试验

#### 3.1 系统实现

为实现 ARQCS 的动态启动策略,在原系统流程(图 1)的数据提取步骤前增加数据提取前统计步骤。根据前文的分析,本文选取了 5 个决定 ARQCS 启动策略的关键参数(表 2)。当满足累计入库率不小于触发启动的入库率临界值时,ARQCS 将首次启动;如果到强制启动时间(第 20 分钟)累计入库率  $S_t$  仍未满足,ARQCS 将强制首次启动。在此之后,系统将在逐小时自动清算启动时间(第 55 分钟)到达时再次启动,以使稍迟入库的数据尽可能完整地参与质量控制。

#### 3.2 业务模拟试验效果

为检验动态启动策略的运行效果,基于 IBM P570 高性能计算环境 Oracle 11g 数据库平台,2012 年 9 月 22—30 日共进行了 216 h 的业务模拟试验。本文从首次启动时间  $T_1$  的变化、新增步骤对单次启动系统开销的影响、系统整体运行效果的提升等方面对试验效果进行测评。

表 2 ARQCS 动态启动策略参数表

Table 2 ARQCS dynamic starting strategy parameters

序号	参数内容	参数设置
1	考核站数	31814 站
2	触发启动的入库率临界值	95%
3	初次启动时间	累计入库率不小于触发启动的入库率临界值
4	强制启动时间	第 20 分钟
5	逐小时自动清算启动时间	第 55 分钟

### 3.2.1 首次启动时间的变化情况

系统首次启动时间  $T_1$  在进行业务模拟试验的 9 d 时间里, ARQCS 在新策略下的实际首次启动时间较原首次启动时间(第 15 分钟)平均提前 20.6 s (图略)。各小时首次启动时间的频次分布如图 4 所示, 最早在第 13 分钟启动, 共出现 5 次, 在 216 h 业务

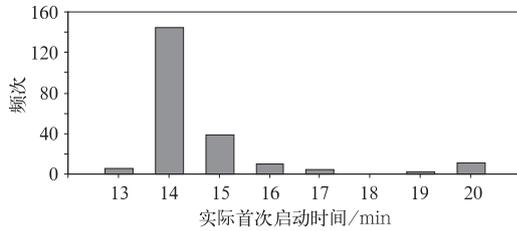


图 4 2012 年 9 月 22—30 日首次启动时间  $T_1$  频次分布图

Fig. 4 Distribution of the 1st starting time from 22 Sep to 30 Sep in 2012

模拟试验中占 2.31%; 最迟在第 20 分钟启动, 共出现 9 次, 占总数的 4.17%, 在绝大多数情况下 ARQCS 首次启动时  $S_i$  超过 95%。

### 3.2.2 新增步骤对单次启动系统开销的影响

流程调整所带来的系统性能变化是评定新启动策略可行性的一项重要指标。试验结果表明, 对原 ARQCS 增加的数据提取前统计步骤单次查询平均耗时为 0.0136 s, 仅占调整前数据提取步骤平均耗时(7.25 s, 2012 年 7 月 23—24 日的 240 次操作的平均值)的 0.19%, 最坏耗时 0.03 s(图略), 基本满足对 ARQCS 的整体运行效果不会造成过多负荷的预期效果。首次启动前平均进行 27 次统计查询(图 5), 最多进行 66 次查询, 即入库量在第 20 分钟前的历次查询中均未满足启动条件, 直到第 20 分钟时系统达到强制启动时间才首次启动(共有 9 次试验为此情况, 仅占 4.17%)。

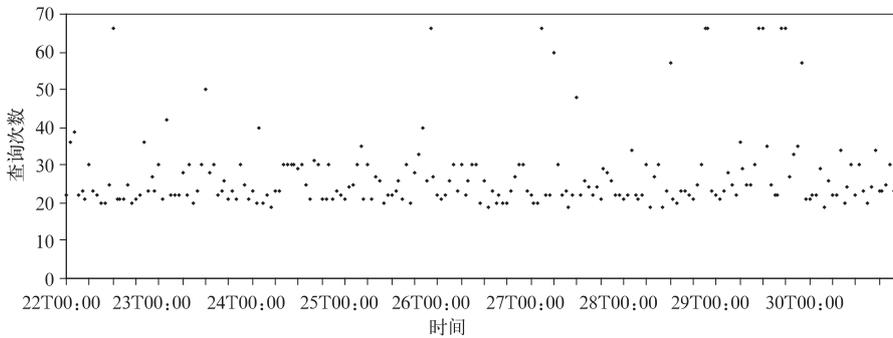


图 5 2012 年 9 月 22—30 日新增数据提取前统计步骤查询次数  
Fig. 5 Search times of newly added statistic step before data acquiring from 22 Sep to 30 Sep in 2012

### 3.2.3 系统整体运行效果的提升

增加数据提取前统计步骤以后, 因 ARQCS 的启动引入了入库时效因子, 后续流程得以优化调整,

原来 5 次系统启动可减少为 2 次。在原静态启动策略下, 平均每小时累计耗时 26.5 min(图 6), 首次启动时间为整点后的第 15 分钟; 采用动态启动策略后,

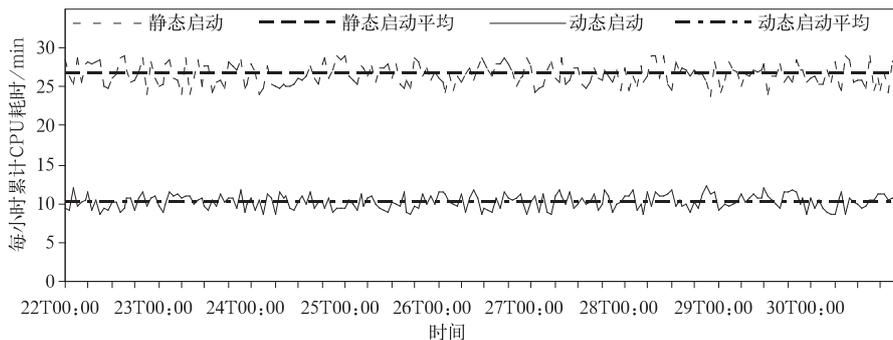


图 6 2012 年 9 月 22—30 日启动策略调整前后每小时累计 CPU 耗时效果对比图  
Fig. 6 Accumulated CPU time costing per hour before and after changing starting strategy from 22 Sep to 30 Sep in 2012

平均每小时累计耗时为 10.2 min。ARQCS 的整体计算性能开销大幅下降,平均每小时节省出的 16.3 min 计算资源可以有效缓解紧张的系统性能消耗。在进行业务模拟试验的 9 d 时间里,采用动态启动策略调整之后首次启动的平均累计入库率  $S_i$  高达 94.91%,较原静态启动策略下首次启动时的平均累计入库率 93.78% 有 1.13% 的提升;在第 55 分钟执行清算时,平均累计入库率为 99.71%,接近 100% 的完整性水平。

#### 4 结论与讨论

本文利用数据库平台对 ARQCS 的运行监控数据,探讨了 ARQCS 的启动策略及其与资料解析入库率、ARQCS 的 CPU 耗时、服务时效之间的关系,得到了以下初步结论:

1) 统计分析发现,自动气象站资料的解析入库效率呈“几”字型分布,具有前段攀升、中段持平、后段回落的特征。每个观测时次的第 5—20 分钟入库率方差较大,系统资源在集中来报时遇到瓶颈,是制约 ARQCS 质量控制效果的主要时间段。

2) 设置观测资料入库率不低于 95% 为首次质量控制启动时间,不仅比传统的第 15 分钟定时启动提前了 20.6 s,而且首次启动时观测资料入库率不低于 95% 的概率从 66.38% 提升至 95.83%。第 20 分钟后平均入库率仅增加 1.36%,在此设置首次质量控制的强制启动点,可保证观测系统或通信系统局部异常延时的资料服务时效。对于平均入库率在第 20 分钟后微增的情况,在第 55 分钟时完成逐小时自动清算,清算启动时的平均累计入库率可达 99.71%。

3) 在动态启动策略的系统实现中,对数据提取步骤进行调整,所增加的数据提取前统计步骤单次查询平均耗时为 0.0136 s,未对 ARQCS 单次启动的系统开销产生过多增加。ARQCS 的启动次数从 5 次优化为 2 次,每个观测时次的总 CPU 耗时由 26.5 min 降为 10.2 min,平均每天节约 CPU 时间 391 min。

在现有计算环境与数据库平台不变的前提下,ARQCS 动态启动策略有效降低了计算资源消耗,提高了系统的整体服务时效。它在虑及自动气象站入库率波动变化的前提下,使 ARQCS 启动时间点

的选择具备一定的自适应能力,业务系统的鲁棒性整体提升。但在目前实时资料存储服务架构下,数据库吞吐率等影响系统整体性能的瓶颈依然存在。如果未来系统整体架构有所改变,能够有效降低或消除数据库吞吐率对系统性能的影响,ARQCS 也将调整以适应新的数据存储服务体系。

#### 参考文献

- [1] 铃伟妙,罗亚丽,张人禾,等. 引发舟曲特大泥石流灾害强降雨过程成因. 应用气象学报,2011,22(4):385-397.
- [2] 易笑园,李泽椿,孙晓磊,等. 渤海西岸暴雨中尺度对流系统的结构及成因. 应用气象学报,2011,22(1):23-34.
- [3] 李瑞萍,程艳芳,赵彩萍,等. 加密自动站资料在强对流天气分析预报中的应用. 气象科技,2012,40(4):596-600.
- [4] 闵晶晶,刘还珠,曹晓钟,等. 天津“6.25”大冰雹过程的中尺度特征及成因. 应用气象学报,2011,22(5):525-536.
- [5] 何立富,周庆亮,陈涛. “05.6”华南暴雨中低纬度系统活动及相互作用. 应用气象学报,2010,21(4):385-394.
- [6] 任芝花,赵平,张强,等. 适用于全国自动站小时降水资料的质量控制方法. 气象,2010,36(7):123-132.
- [7] 任芝花,许松,孙化南,等. 全球地面天气报历史资料质量检查与分析. 应用气象学报,2006,17(4):412-420.
- [8] 刘小宁,鞠晓慧,范邵华. 空间回归检验方法在气象资料质量检验中的应用. 应用气象学报,2006,17(1):37-42.
- [9] 赵煜飞,任芝花,张强. 适用于全国气象自动站正点相对湿度资料的质量控制方法. 气象科学,2011,31(6):687-693.
- [10] 鞠晓慧,任芝花,张强. 自动站小时气压的质量控制方法研究. 安徽农业科学,2010,38(27):15130-15133.
- [11] 林子雨,杨冬青,王腾蛟,等. 基于关系数据库的关键词查询. 软件学报,2010,21(10):2454-2476.
- [12] 崔跃生,张勇,曾春,等. 数据库物理结构优化技术. 软件学报,2013,24(4):761-780.
- [13] 刘波,范士明,刘华. 实时数据库混合索引机制的设计与实现. 计算机应用,2011,31(8):2265-2269.
- [14] 文平. Oracle 大型数据库系统在 AIX/UNIX 上的实战详解. 北京:电子工业出版社,2012:593-605.
- [15] 刘宪军. Oracle RAC 11g 实战指南. 北京:机械工业出版社,2011:6-7.
- [16] 沈宏. Oracle 最新技术战略和产品方向及 11G 新特点. 2011:36. <http://www.oracle.com/technetwork/cn/community/developer-day/5-oracle-db-platform-11g-1432258-zhs.pdf>.
- [17] 杨志邦,徐成,周旭,等. 实时系统中弹性调度策略. 计算机应用,2012(2):573-577.
- [18] 邹勇,淮晓永,李明树. 开放式实时系统中的自适应调度方法. 计算机学报,2004,27(1):58-65.
- [19] 钟樑,刘晓燕,张晶,等. 开放式实时系统资源共享环境下的调度方法分析. 小型微型计算机系统,2012,33(11):2362-2366.

## ARQCS Starting Strategy and Its Relationship with Computing Resource Cost

Liu Yiming Zhou Zijiang Yuan Fang Ruan Yuzhi He Wenchun Sun Chao Liu Yuanyuan

(National Meteorological Information Center, Beijing 100081)

### Abstract

AWS Observation Data Real-time Quality Control System (ARQCS) is an operational real-time meteorological data application system under IBM P570 high performance computing (HPC) Oracle 11g database platform. Functions including data decoding, database inserting, quality control (QC), storage management and share service are provided for more than 30000 AWS all over China. In 2009, when ARQCS is firstly built, QC methods including boundary value check, internal consistency check, time consistency check and spatial consistency check is applied to only 1 element of hourly precipitation. And the starting strategy is a static one, which start ARQCS at the 15th, 25th, 35th, 45th and 55th minute every hour. Later in 2010, QC methods of other important meteorological elements including air temperature, air pressure, humidity, wind direction and speed get to be applied in ARQCS. Meanwhile, the system computing logic is made more complex after 2 times of updating in 2011 and 2012. Now, it is planned to extend ARQCS to 158 elements in 11 classes totally, which need more calculating resources accordingly. To guarantee QC capability and service timeliness of ARQCS in a high level under limited computing resources, a series of schemes are designed and investigated. System log under IBM P570 HPC Oracle database environment from 1st April to 30th Sep in 2012 is used to analyze ARQCS performance. It is found that the database entry rate (ER) of AWS data exhibits a trapezoid shaped distribution, and variance of ER is large from the 5th to the 20th minute in one hour, which means accumulated ER at the 15th minute is unstable and a low accumulated ER may be got if ARQCS starts at this time. It also indicates that an accumulated ER of 95% is very possible (84.89%) to get before the 20th minute, and accumulated ER is increased by only 1.36% after the 20th minute in average. So a new dynamic starting strategy is employed, that ARQCS starts for the first time when accumulated ER gets more than 95% or until the 20th minute, and starts for the second time at the 55th minute. With this approach, the possibility for accumulated ER over 95% at the 1st QC starting is increased by 29% (from 66.38% to 95.83%). And the average 1st QC starting time is 20.6 seconds before the 15th minute in original static starting strategy. Also, less number of starts from 5 to 2 decrease the CPU time cost from 26.5 minutes to 10.2 minutes per hour, which means saving 391 minutes CPU time per day. It is concluded that the dynamic starting strategy is effective for ARQCS starting adaptively and ensures system robustness.

**Key words:** AWS observation data; real-time quality control; starting strategy; computing resource