文章编号:1005-1228(2013)06-0044-02

S4 开发及性能调优设计思想

赵思焰

(爱立信(中国)通信有限公司,北京,100102)

摘 要:文章对于实时流数据分析框架 S4 在开发和性能调优提出了新的设计思想,以便更适合在各种应用中进行整合和高效实现的探讨。

关键词:Apache S4;实时流框架;性能调优;开发设计

中图分类号: TP311.52 **文献标识码**:A

S4 Development Design and Performance Tunning Design Thinking

ZHAO Si-yan

(Ericsson (China) Communications Co. Ltd,Beijing 100102,China)

Abstract: This article describe the thinking of development design and performance tunning design for S4 which is the real time data analysis framework, in order to apply sorts of applications in integrating other batch system and highly efficient implementations.

Key words: Apache S4; real time stream framework; performance tunning; development design

S4 [1] 是支持无限实时流的开发应用系统,2010 年发起于 yahoo,2011 年正式成为 apache 孵化器项目。S4 的目的就是简化开发并行处理系统所固有的复杂性,倡导模块化和热插拔性,使应用动态和更方便的装载到处理流中。现在最新版本是 0.6。弹性的系统设计可以应用不同数据场景中,也可以作为 batch 系统的增强实现。

1 S4 的开发

由于 s4 是基于纯 java 实现个人 gradlew [2] 工程构建,按照各子项目划分成不同功能模块,大大简化了系统的应用整合能力和接口开放灵活性。掌握开发 S4 模块功能和开发设计流程,便于对应用系统对需求的掌控和数据调优都会起到事半功倍的作用。

2.1 相关开发源码项目及功能

S4 的源码包包含有多个子项目和 Twitter 应用实例,其工程构建具有从底层,工具到应用的清晰层次,基于 S4 开发应用架构,必须掌握以下相关源码工程和实例,才能从设计到性能有准确的把握。

s4-benchmarks: 用于测试部署服务器及节点的性能,以便评估自定义项目的 performance。

s4-core:核心子工程项目,用于在 eclipse 中调试和启动应用程序用。

s4-tools: 用于构建 application 项目, 并 build 到 eclipse 项目中。

test-apps: 学习编写 PE 输入输出流很好的范例,结合 walkthrough 学习其应用设计。

1.2 开发流程设计

开发流程设计图, 如图 1 所示, 图中主要描述在开发流程中必须的步骤。整个开发流程基于 S4 的 Gradlew 工程开发,建议用 Eclipse 做为加速开发,调试以及模拟运行的有力工具。

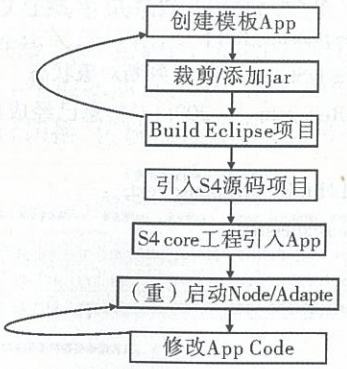

图 1 开发流程设计图

收稿日期:2013-07-16

作者简介:赵思焰(1973-),男,北京人,研究员,主要研究方向:大数据及实时分析框架。

流程简述如下:

(1)创建模板应用(App)。利用 S4 自带 s4-tool 工程构建模板应用,帮助建立应用的最简运行包。

(2)裁剪 / 添加 jar。S4 自带模板项目会带来多余的 jar 文件或者你需要添加自定义 jar 包,则需要修改应用(App)的 build.gradle 文件进行重编译,例如设计单独的适配器(Adapter)工程。

(3)Build Eclipse 项目。利用 gradlew 命令将自定义应用及 S4 主要工程项目导出到 Eclipse 中,如果自定义第三方包,在需要重新修改模板应用直到加载完成。

(4)引入 S4 源码。S4 源码工程主要为 Core,Base,Comm 子项目的引入,便于调试节点或适配器(Adapter)工程内部机制,以上所有工程都需进行 Eclipse 工程构建。

(5)S4 core 工程引入应用(App)。将应用作为 s4-core 子项目的类路径便于开发调试。

(6)启动 Node/Adapter。利用 s4-core 工程可以启动应用节点(Node),多节点和多适配器留均可以在控制台参数里设置与调试。

(7)修改 App 代码。在应用中直接修改逻辑,然后执行 6 操作,直到打包成 s4r 部署文件,最后调试后将 s4r 文件部署到相应的分布式系统中即可。

2　S4 的性能影响因素

S4 性能等级分类,如图 2 所示:

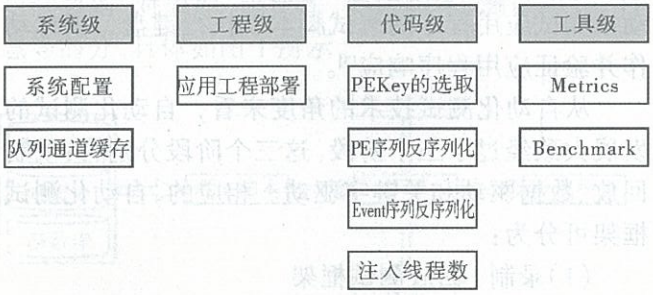

图 2　S4 性能等级分类

由于系统中对于 s4 或其自身的架构特点对应用的各种影响。我将 s4 性能[3]影响分为四个等级:

(1)系统级。

在多个应用系统中,系统配置文件(default.s4.comm.properties)起到了全局作用,即改变任意属性参数会影响每个应用的性能,在发送端并发数和工作队列可略作调整,提高系统级的整体提升。

由于 s4 采用 kyro 进行高速序列化,其缓冲池(buffer)也可按服务器配置性能进行设置修改。

(2)工程级。

S4 的模板(template)工程缺省将适配器(Adapter)和应用(App)应用捆绑在一起,虽然失去了一定的适配的灵活性,但对于单个输入流在性能上要优于独立适配器工程发布

(3)代码级。

由于 PE Key 决定了 PE 的 instances 的个数,决定了 PE 的序列化操作次数,所以这里 key 并不代表唯一,可以表示为某一类的 key。

PE 的序列化和反序列化中可以适当增加并行处理机制,以提高逻辑操作能力。

对于适配器 (Adapter) 发送端,最好以缺省 s4 Event 类作为发送对象,扩展 Event 对象会大大增加序列化操作时间。

利用 s4 注入机制,灵活调整应用节点或发送端工作线程数,找到最优平衡数字。

(4)工具级。

利用 Metrics Library,可将应用层工作状态直接输出到 csv Report 中分析例如每秒种处理 Event 个数,也可以自定义 Gauge 输出某 PE 工作时间,但注意输出的频度也会造成性能改变。

同时我们可以利用 s4 benchmark 项目来测试运行在工作环境下机器的吞吐量,改变相应的节点数和修改发送 Event 最大速率,将会评估到数据流的最大变化。

3　结束语

总之,s4 的调优工作涉及到各个环节的方方面面,例如网络负载和机器配置,针对自己系统中的特点需要制定有效的设计计划,才能抓住问题的关键,找到性能提升适合的办法,更好的实施于离线(Batch)系统中或者新的实时专用系统。

参考文献:

[1] Apache S4. The Apache Software Foundation[EB/OL]. http://incubator.apache.org/s4/.

[2] Leo Neumeyer.Standford Infolab[EB/OL].http://www.slideshare.net/leoneu/20111104-s4-overview.

[3] LeonardoNeumeyer,BruceRobbins,AnishNair,AnandKesari.S4:Distributed Stream Computing Platform[EB/OL]. http://www.4lunas.org/pub/2010-s4.pdf.