Instituto Federal de Educação, Ciência e Tecnologia da Bahia - IFBa
Grupo de Sistemas Distribuídos, Otimização, Redes e Tempo-Real - GSORT
Especialização em Computação Distribuída e Ubíqua
INF628 - Engenharia de Software para Sistemas Distribuídos
Prof.: Sandro Santos Andrade

ESPECIFICAÇÃO DE TRABALHO PRÁTICO

Objetivo Geral. O trabalho prático tem como objetivo a aplicação dos *design patterns* para sistemas MapReduce, vistos em sala de aula.

Formato e Produtos a serem Gerados. O trabalho prático será realizado em dupla e consiste na implementação de três *jobs* MapReduce, especificados abaixo.

Objetivos Específicos. Os três *jobs* MapReduce a serem implementados têm como objetivo a realização de consultas no *dataset* do *stackoverflow* fornecidos em sala. Os *jobs* a serem implementados são:

1) Operação de filtragem na base de usuários (stackoverflow.com-Users) para obtenção somente dos usuários brasileiros.
2) Obtenção dos *top ten* usuários brasileiros em relação ao número de postagens realizadas.
3) Uma operação de *join* a sua escolha, apresentando o *id* do usuário, seu *display name* e todas as outras informações da tabela com a qual o *join* foi feito.

No dia da apresentação, a dupla deverá ser capaz de identificar qual(is) *pattern(s)* foi(oram) utilizado(s), quais os componentes da solução (*mappers*, *reducers*, *combiners* e *partitioners*) e de realizar esclarecimentos sobre o código-fonte.

Informações sobre as Máquinas Virtuais

Ao executar as imagens no VirtualBox, configure a rede para operar no modo *Bridged Adapter*. No Linux, isso requer a inicialização do módulo *vboxnetflt*. Para isso execute o seguinte comando como *root*: "modprobe vboxnetflt". Para que a inicialização não seja perdida ao reiniciar a máquina, crie o arquivo /etc/modules-load.d/virtualbox.conf, contendo as seguintes linhas:

vboxdrv
vboxnetflt

Usuários e senhas das imagens VirtualBox:
hadoop/hadoop
root/hadoop

As imagens estão configuradas para automaticamente iniciar os serviços *datanode* e *nodemanager* nas imagens do tipo *datanode* e os serviços *namenode, resourcemanager e history server* na imagem do tipo *namenode*. Após a inicialização das máquinas virtuais você pode acessar os seguintes *frontends* web, a partir do computador *host*:

http://<ip-do-namenode>:50070/dfshealth.jsp → monitor do HDFS (*Hadoop Distributed File System*)
http://<ip-do-namenode>:8088/cluster → monitor do YARN (visualizador dos *jobs* MapReduce em execução)
http://<ip-do-namenode>:19888/jobhistory → servidor de histórico de execuções MapReduce passadas

Os serviços podem ser inicializados, interrompidos ou reinicializados com os seguintes comandos:

sudo systemctl start | stop | restart hadoop-datanode.service
sudo systemctl start | stop | restart hadoop-nodemanager.service
sudo systemctl start | stop | restart hadoop-namenode.service
sudo systemctl start | stop | restart hadoop-resourcemanager.service
sudo systemctl start | stop | restart hadoop-historyserver.service

O arquivo HADOOP_ROOT/etc/hadoop/mapred-site.xml indica qual o *framework* a ser utilizado para a execução de *jobs*. O valor atual contém '*yarn'* que é a execução do *job* no *cluster*. Se você retirar esta configuração ele tentará executar o *job* localmente (para isso você precisará ter o *namenode* e *datanode* rodando numa mesma máquina), porém deve ser inviável devido ao tamanho dos dados do *stackoverflow* a serem processados.

Data de Entrega: 19 de Fevereiro de 2014
Dúvidas devem ser enviadas para sandroandrade@ifba.edu.br