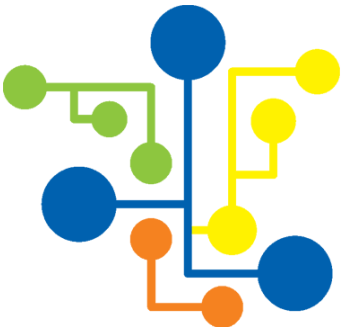


Data Science Academy

www.datascienceacademy.com.br

Formação Cientista de Dados

Projeto com Feedback 6

Processamento de Logs em Tempo Real Com Flume, Spark Streaming e Hbase

Na computação, um arquivo de log é um arquivo que registra eventos que ocorrem em um sistema operacional ou outro software, ou mensagens entre diferentes usuários de um software de comunicação. Logging é o ato de manter um log. No caso mais simples, as mensagens são gravadas em um único arquivo de log. Exemplos de arquivos de log, incluem:

- Proxy log
- Transaction log
- Event log
- Message log
- Application log
- Web server log

A análise de logs tem várias aplicações práticas e os principais casos de uso de processamento de log de servidor web por exemplo, incluem:

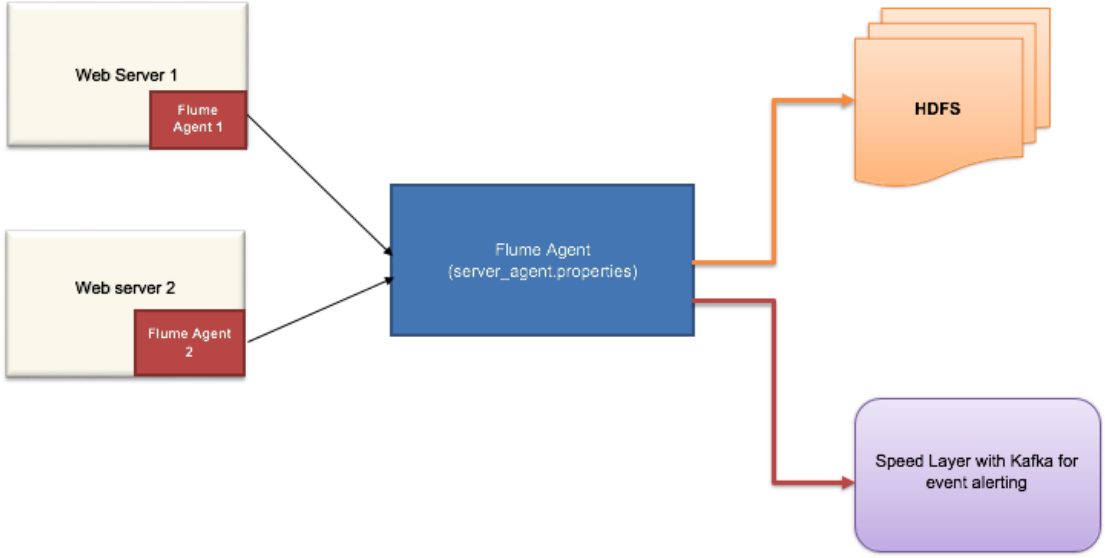
- Monitoramento de integridade de aplicativos
- Fraude - Segurança
- Padrão do usuário (sessão de um fluxo de cliques)
- Experiência de usuário
- Triagem de Suporte
- Coleta de dados métricos

Neste projeto de Big Data, você deverá construir um pipeline de monitoramento em tempo real de logs de aplicação usando Apache Flume para aquisição de dados, Spark Streaming para processamento e Apache Hbase para armazenamento.

Dado o conjunto de dados de log de dois servidores da Web disponível para download em http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html, precisamos fazer análises, processamento de eventos e recuperação de dados de log, além claro de armazenamento. Exemplo extraído do dataset:

```
in24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0" 200 1839
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0" 304 0
```

Arquitetura sugerida:



Você pode usar uma Sandbox Cloudera ou Hortonworks ou pode construir toda sua infraestrutura em nuvem, assim como usar máquinas virtuais. Sugerimos a utilização de pelo menos 3 nodes no cluster.

Você deve enviar os scripts usados para construir o ambiente e evidências de que o pipeline funciona!

Quando concluir o projeto, envie os scripts e datasets para projeto@dsacademy.com.br. Caso os datasets usados sejam muito grandes, armazene em um diretório virtual (existem vários na internet, como Google Drive ou Dropbox) e envie o link para que nossa equipe possa baixar os datasets. Se os arquivos foram pequenos (uma amostra do dataset original), envie no anexo junto com o script. Documente seu script tanto quanto possível.

Caso prefira, disponibilize seu projeto no Github e envie o link do seu repositório para nossa equipe no e-mail projeto@dsacademy.com.br. Nesse caso, o Readme do repositório deve constar que este trata-se de um projeto da Formação Cientista de Dados da Data Science Academy.

Em até 24 horas, daremos o feedback!

Bom trabalho!