

Tutorial 04 – Mostrando Valores de "Data Hub"

Cenário

Seu Gerente: não posso acreditar nos truques que você faz com os dados e estou prestes a promovê-lo e a investir em uma nova equipe sob sua liderança ... quando tudo se mostrar possível. Neste meio tempo você recebe uma chamada de emergência - como você agora que detêm a informação - e seu gerente está apavorado sobre a perda de vendas nos últimos três dias ...

Você: foi do céu para a terra com a possibilidade de promoção a chamada de emergência ... bem, sorte pra você, pode haver uma maneira rápida de descobrir o que está acontecendo ...

Explorando Eventos de Log de Modo Interativo

O que você poderia fazer para habilitar a análise guiada e a exploração de dados é fazer com que isto seja pesquisável. Ao indexar seus dados usando qualquer uma das opções de indexação fornecidas pela *Hadoop Search* seus dados podem ser pesquisados para uma variedade de públicos-alvo. Você pode escolher os dados do índice de lote usando a ferramenta *MapReduce Indexing* ou, como no nosso exemplo abaixo, estender a configuração do [Apache Flume](#) que já está obtendo os dados do log da web para também publicar eventos no [Apache Solr](#) para indexação em tempo real.

Os dados de log da web são logs padrão do servidor da *Web*, que podem se parecer como algo do tipo:

```
161.36.171.49 - - [23/Jun/2015:13:33:04 -0800] "GET /product/1166 HTTP/1.1" 200 1394 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
85.157.70.199 - - [23/Jun/2015:13:33:05 -0800] "GET /departments HTTP/1.1" 200 1981 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.77.4 (KHTML, like Gecko) Version/7.0.5 Safari/537.77.4"
113.27.167.86 - - [23/Jun/2015:13:33:06 -0800] "GET /department/outdoors/products HTTP/1.1" 404 1410 "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
113.27.167.86 - - [23/Jun/2015:13:33:07 -0800] "GET /login HTTP/1.1" 200 2196 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3) AppleWebKit/537.76.4 (KHTML, like Gecko) Version/7.0.4 Safari/537.76.4"
221.203.86.126 - - [23/Jun/2015:13:33:08 -0800] "GET /product/255 HTTP/1.1" 200 1236 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"
138.165.114.233 - - [23/Jun/2015:13:33:09 -0800] "GET /categories/international%20soccer/products HTTP/1.1" 200 298 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
84.102.128.216 - - [23/Jun/2015:13:33:10 -0800] "GET /login HTTP/1.1" 200 1702 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"
86.51.171.34 - - [23/Jun/2015:13:33:11 -0800] "GET /department/outdoors/products HTTP/1.1" 200 764 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
73.243.154.157 - - [23/Jun/2015:13:33:12 -0800] "GET /departments HTTP/1.1" 200 452 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.9; rv:30.0) Gecko/20100101 Firefox/30.0"
150.176.228.162 - - [23/Jun/2015:13:33:13 -0800] "GET /department/fan%20shop/products HTTP/1.1" 200 1792 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
69.165.160.102 - - [23/Jun/2015:13:33:14 -0800] "GET /department/team%20sports/categories HTTP/1.1" 200 1988 "-" "Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0"
87.228.159.68 - - [23/Jun/2015:13:33:15 -0800] "GET /product/322 HTTP/1.1" 200 1125 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0) Gecko/20100101 Firefox/30.0"
47.251.217.168 - - [23/Jun/2015:13:33:16 -0800] "GET /departments HTTP/1.1" 200 1953 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.77.4 (KHTML, like Gecko) Version/7.0.5 Safari/537.77.4"
62.139.93.44 - - [23/Jun/2015:13:33:17 -0800] "GET /departments HTTP/1.1" 200 1760 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
33.41.156.71 - - [23/Jun/2015:13:33:18 -0800] "GET /departments HTTP/1.1" 200 1000 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.125 Safari/537.36"
47.251.217.168 - - [23/Jun/2015:13:33:19 -0800] "GET /add_to_cart/354 HTTP/1.1" 200 1372 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
216.90.108.250 - - [23/Jun/2015:13:33:20 -0800] "GET /add_to_cart/134 HTTP/1.1" 200 1222 "-" "Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0"
85.245.45.125 - - [23/Jun/2015:13:33:21 -0800] "GET /department/fan%20shop/products HTTP/1.1" 200 1551 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
25.38.199.191 - - [23/Jun/2015:13:33:22 -0800] "GET /department/apparel/categories HTTP/1.1" 404 1936 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0) Gecko/20100101 Firefox/30.0"
16.62.133.72 - - [23/Jun/2015:13:33:23 -0800] "GET /departments HTTP/1.1" 200 1279 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1985.125 Safari/537.36"
187.210.152.115 - - [23/Jun/2015:13:33:24 -0800] "GET /departments HTTP/1.1" 200 735 "-" "Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0"
84.223.254.95 - - [23/Jun/2015:13:33:25 -0800] "GET /add_to_cart/427 HTTP/1.1" 200 764 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
93.193.116.19 - - [23/Jun/2015:13:33:26 -0800] "GET /departments HTTP/1.1" 200 1437 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
194.31.138.82 - - [23/Jun/2015:13:33:27 -0800] "GET /login HTTP/1.1" 200 938 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"
44.132.157.21 - - [23/Jun/2015:13:33:28 -0800] "GET /departments HTTP/1.1" 200 1010 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
83.7.158.118 - - [23/Jun/2015:13:33:29 -0800] "GET /product/974 HTTP/1.1" 200 1965 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_4) AppleWebKit/537.77.4 (KHTML, like Gecko) Version/7.0.5 Safari/537.77.4"
53.6.243.15 - - [23/Jun/2015:13:33:30 -0800] "GET /department/fitness/products HTTP/1.1" 200 684 "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36"
94.247.156.215 - - [23/Jun/2015:13:33:31 -0800] "GET /departments HTTP/1.1" 200 1264 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"
```

O *Solr* organiza dados de forma semelhante à forma como um banco de dados *SQL* faz. Cada registro é chamado de 'documento' e consiste em campos definidos pelo esquema: apenas como uma linha em uma tabela de banco de dados. Em vez de uma tabela, *Solr* a chama de "coleção" de documentos. A diferença é que os dados em *Solr* tendem a ser mais vagamente estruturados. Os campos podem ser opcionais e em vez de sempre corresponderem aos valores exatos, você também pode inserir consultas de texto que coincidem parcialmente com um campo, assim como você está procurando por páginas da *web*. Você também verá *Hue* se referir a "fragmentos" - e é assim que o *Solr* quebra coleções para espalhá-las em torno do *cluster* para que você possa pesquisar todos os seus dados em paralelo.

Veja como você pode iniciar a indexação em tempo real através da *Cloudera/Hadoop Search* e *Flume* sobre os dados do registro do servidor da *Web* de amostra e usando a *UI* (Interface do Usuário) de Pesquisa no *Hue* para explorá-la:

Criando Seu Índice de Pesquisa

Normalmente, quando se está implantando um novo esquema de pesquisa existem quatro etapas:

1) Criar uma configuração vazia:

Para o bem deste tutorial, você **não precisará realmente executar as etapas 1 ou 2**, já que foi incluída uma configuração e o arquivo de esquema neste cluster. Eles podem ser verificados no seguinte diretório **/opt/examples/flume/solr_configs/conf/**.

Digite o comando abaixo no terminal para gerar as configurações do **solr**:

```
solrctl --zk quickstart:2181/solr instancedir --generate solr_configs
```

// Não precisa fazer isso para este tutorial. A configuração já foi gerada. Esta instrução está aqui caso queira criar seu próprio índice.

O resultado desse comando seria uma configuração de esqueleto que você poderia personalizar ao seu gosto. O primeira a coisa a ser feita é personalizar o **conf/schema.xml** que será discutido na próxima etapa.

2) Editar o esquema:

Como mencionado anteriormente, já geramos os arquivos de configuração para você. Você pode ver o [esquema de amostra modificado aqui](#).

A área mais comum que lhe interessa é a seção **<fields> </ fields>**. A partir desta área, você pode definir os campos que estão presentes e pesquisáveis em seu índice.

3) Fazendo o upload da configuração:

Digite os comandos abaixo no terminal para carregar os *logs* para o **solr** que estão configurados no esquema pré-definido.

```
cd /opt/examples/flume  
solrctl --zk quickstart:2181/solr instancedir --create live_logs ./solr_configs
```

Caso os dados já tenham sido carregados anteriormente e deseja apenas atualizá-los substitua o **--create** por **--update** no comando anterior.

4) Criando sua coleção:

Digite o seguinte comando no terminal para criar sua coleção de dados.

```
solrctl --zk quickstart:2181/solr collection --create live_logs -s 1
```

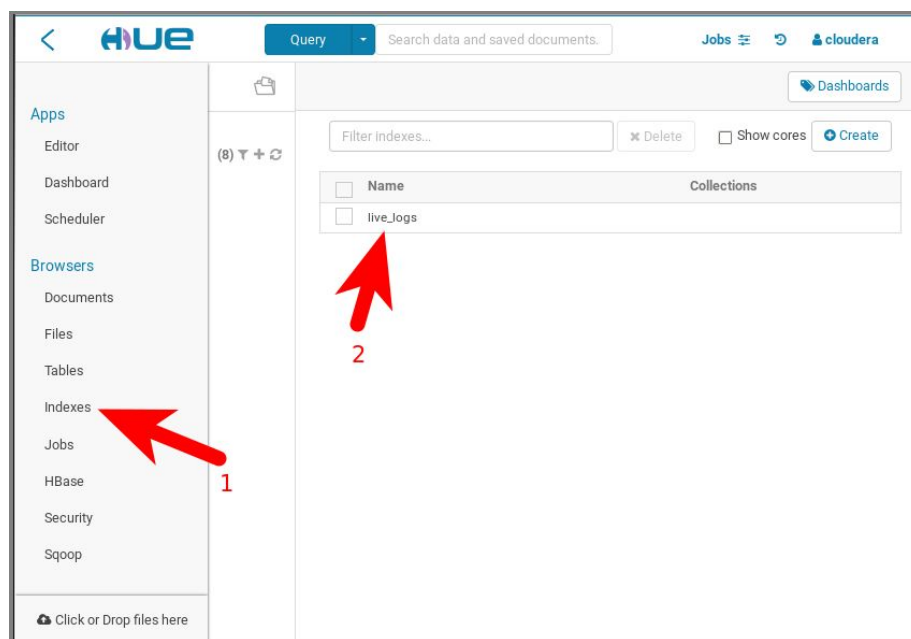
Um pouco sobre o Spark

Se você estiver familiarizado com o *MapReduce*, você notará que este exemplo da *Spark* usa conceitos muito parecidos de operações de "map" e "reduce" (as operações 'join' e 'groupBy' são apenas variações especiais de 'reduce'). A principal vantagem, porém, de usar o *Spark* é que o código é mais conciso e os resultados intermediários podem ser armazenados na memória - permitindo-nos fazer sequências complexas e iterativas muito mais rápidas.

Usar *MapReduce* ainda pode ser uma boa opção para trabalhos em lote que usam muito mais dados do que cabe na memória do *cluster* (por exemplo *petabytes* de dados). Estamos usando *Spark-on-YARN*, o que significa que *MapReduce* e *Spark* (como muitos componentes da CDH) compartilham o mesmo gerenciador de recursos, facilitando o gerenciamento de compartilhamento de recursos entre muitos usuários.

Nota: se deixar parado por algum tempo, o ***prompt scala>*** pode ficar coberto de mensagens de log do cluster. Basta pressionar enter para atualizar o prompt.

Você pode verificar se você criou sua coleção com sucesso em ***Solr***, indo para ***Hue -> Indexes -> live_logs***



Agora que você verificou que sua coleção / índice de pesquisa foi criada com sucesso e podemos começar a colocar dados nela usando ***Flume e Morphlines***. O ***Flume*** é uma ferramenta para obter fluxos de dados em seu *cluster* a partir de fontes como arquivos de *log*, fluxos de rede e muito mais. ***Morphlines*** é uma biblioteca Java para fazer ***ETL on-the-fly***, e é um excelente companheiro para o ***Flume***. Ele permite que você defina uma cadeia de tarefas,

como ler registros, analisar e formatar campos individuais, e decidir onde enviá-los, etc. Nós definimos um *morphline* que lê registros de *Flume*, quebras estes nos campos que queremos procurar, e carrega-os no *Solr* (você pode ler mais sobre *Morphlines* [aqui](#)). Este exemplo, *Morphline* é definido em `/opt/examples/flume/conf/morphline.conf`, e vamos usá-lo para indexar nossos registros em tempo real, pois eles são criados e obtidos pela *Flume*.

Iniciando o Gerador de Log

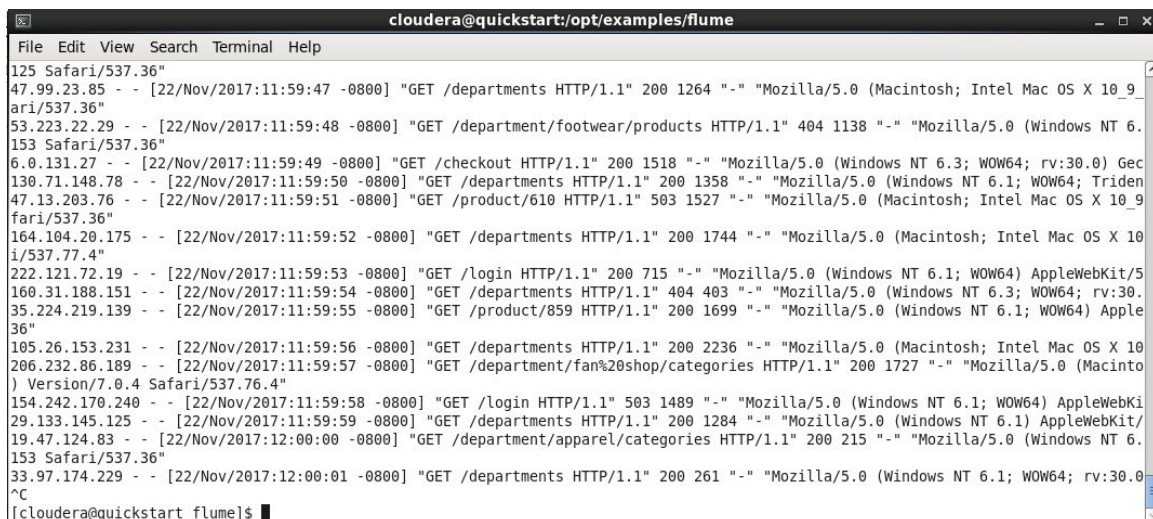
O cluster Cloudera Live possui um gerador de *log* para uso com dados de amostra. Inicie o gerador de *log* executando o seguinte comando no terminal:

```
start_logs
```

Você pode verificar se o gerador de log esta executando

```
tail_logs
```

Pode-se parar de verificar os logs pressionando <Ctrl + C> para retornar ao seu terminal. Você deve ver uma tela semelhante à que está abaixo:



```
cloudera@quickstart:/opt/examples/flume
File Edit View Search Terminal Help
125 Safari/537.36"
47.99.23.85 - - [22/Nov/2017:11:59:47 -0800] "GET /departments HTTP/1.1" 200 1264 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_
ari/537.36"
53.223.22.29 - - [22/Nov/2017:11:59:48 -0800] "GET /department/footwear/products HTTP/1.1" 404 1138 "-" "Mozilla/5.0 (Windows NT 6.
153 Safari/537.36"
6.0.131.27 - - [22/Nov/2017:11:59:49 -0800] "GET /checkout HTTP/1.1" 200 1518 "-" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:30.0) Gec
130.71.148.78 - - [22/Nov/2017:11:59:50 -0800] "GET /departments HTTP/1.1" 200 1358 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Triden
47.13.203.76 - - [22/Nov/2017:11:59:51 -0800] "GET /product/610 HTTP/1.1" 503 1527 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9
fari/537.36"
164.104.20.175 - - [22/Nov/2017:11:59:52 -0800] "GET /departments HTTP/1.1" 200 1744 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10
i/537.77.4"
222.121.72.19 - - [22/Nov/2017:11:59:53 -0800] "GET /login HTTP/1.1" 200 715 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/5
160.31.188.151 - - [22/Nov/2017:11:59:54 -0800] "GET /departments HTTP/1.1" 404 403 "-" "Mozilla/5.0 (Windows NT 6.3; WOW64; rv:30.
35.224.219.139 - - [22/Nov/2017:11:59:55 -0800] "GET /product/859 HTTP/1.1" 200 1699 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) Apple
36"
105.26.153.231 - - [22/Nov/2017:11:59:56 -0800] "GET /departments HTTP/1.1" 200 2236 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10
206.232.86.189 - - [22/Nov/2017:11:59:57 -0800] "GET /department/fan%20shop/categories HTTP/1.1" 200 1727 "-" "Mozilla/5.0 (Macinto
) Version/7.0.4 Safari/537.76.4"
154.242.170.240 - - [22/Nov/2017:11:59:58 -0800] "GET /login HTTP/1.1" 503 1489 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKi
29.133.145.125 - - [22/Nov/2017:11:59:59 -0800] "GET /departments HTTP/1.1" 200 1284 "-" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/
19.47.124.83 - - [22/Nov/2017:12:00:00 -0800] "GET /department/apparel/categories HTTP/1.1" 200 215 "-" "Mozilla/5.0 (Windows NT 6.
153 Safari/537.36"
33.97.174.229 - - [22/Nov/2017:12:00:01 -0800] "GET /departments HTTP/1.1" 200 261 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0
^C
[cloudera@quickstart flume]$
```

Mais tarde, se você quiser parar o gerador de log é só digitar:

```
stop_logs
```

Flume e o morphline

Agora que temos um índice *Solr* vazio e eventos de log sendo gerados que podem ser visualizados em `/opt/gen_logs/logs/access.log` ou com o comando **tail_logs**, podemos usar **Flume** e **morphlines** para carregar o índice com os dados de *log* em tempo real.

A ferramenta chave neste tutorial é o *Flume*. O **Flume** é um sistema para coletar, agregar e mover grandes quantidades de dados de *log* de muitas fontes diferentes para uma fonte de dados centralizada.

Com alguns arquivos de configuração simples, podemos usar *Flume* e uma *morphline* (uma maneira simples de realizar on-line ETL), para carregar nossos dados para nosso índice *Solr*.

Você pode usar o *Flume* para carregar muitos outros tipos de armazenamento de dados; *Solr* é apenas o exemplo que estamos usando para este tutorial.

Nota: posteriormente você pode revisar o arquivo [flume.conf](#) e o [morphline.conf](#) que ele usa.

Inicie o agente *Flume* executando o seguinte comando no terminal:

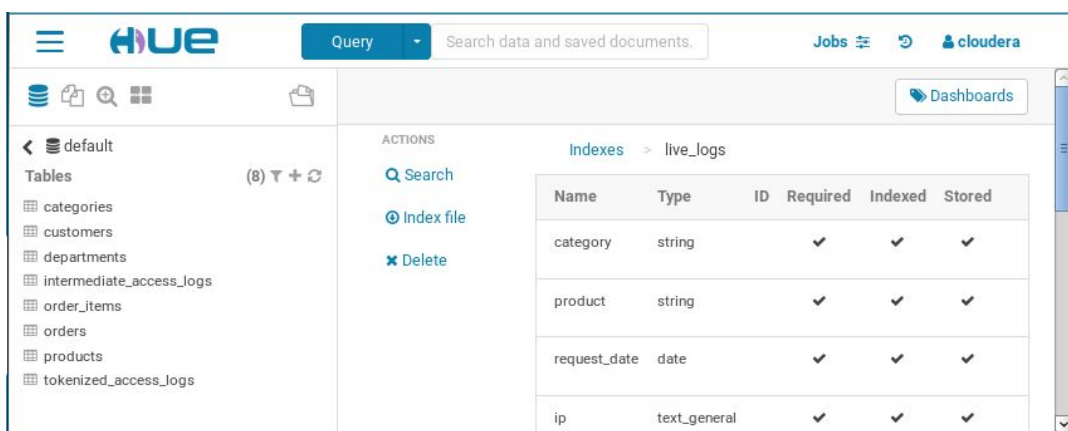
```
flume-ng agent \
--conf /opt/examples/flume/conf \
--conf-file /opt/examples/flume/conf/flume.conf \
--name agent1 \
-Dflume.root.logger=DEBUG,INFO,console
```

Isso começará a executar o agente *Flume* em primeiro plano. Uma vez que começou, e está processando registros, você deve ver algo como:

```
cloudera@quickstart:/opt/examples/flume
File Edit View Search Terminal Help
17/11/22 12:12:04 INFO core.SolrConfig: Adding specified lib dirs to ClassLoader
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../contrib/extraction/lib (resolved as: /tmp/1511381515517-0/../../contrib/extraction/lib).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../dist/ (resolved as: /tmp/1511381515517-0/../../dist).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../contrib/clustering/lib/ (resolved as: /tmp/1511381515517-0/../../contrib/clustering/lib).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../dist/ (resolved as: /tmp/1511381515517-0/../../dist).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../contrib/langid/lib/ (resolved as: /tmp/1511381515517-0/../../contrib/langid/lib).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../dist/ (resolved as: /tmp/1511381515517-0/../../dist).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../contrib/velocity/lib/ (resolved as: /tmp/1511381515517-0/../../contrib/velocity/lib).
17/11/22 12:12:04 WARN core.SolrResourceLoader: Can't find (or read) directory to add to classloader: ../../dist/ (resolved as: /tmp/1511381515517-0/../../dist).
17/11/22 12:12:05 INFO file.Log: Updated checkpoint for file: /home/cloudera/.flume/file-channel/data/log-1 position: 14921 logWriteOrderID: 1511381466146
17/11/22 12:12:05 INFO update.SolrIndexConfig: IndexWriter infoStream solr logging is enabled
17/11/22 12:12:05 INFO core.SolrConfig: Using Lucene MatchVersion: 4.4.0
17/11/22 12:12:06 INFO core.Config: Loaded SolrConfig: solrconfig.xml
17/11/22 12:12:06 INFO schema.IndexSchema: Reading Solr Schema from /tmp/1511381515517-0/conf/schema.xml
17/11/22 12:12:06 INFO schema.IndexSchema: [null] Schema name=example
```

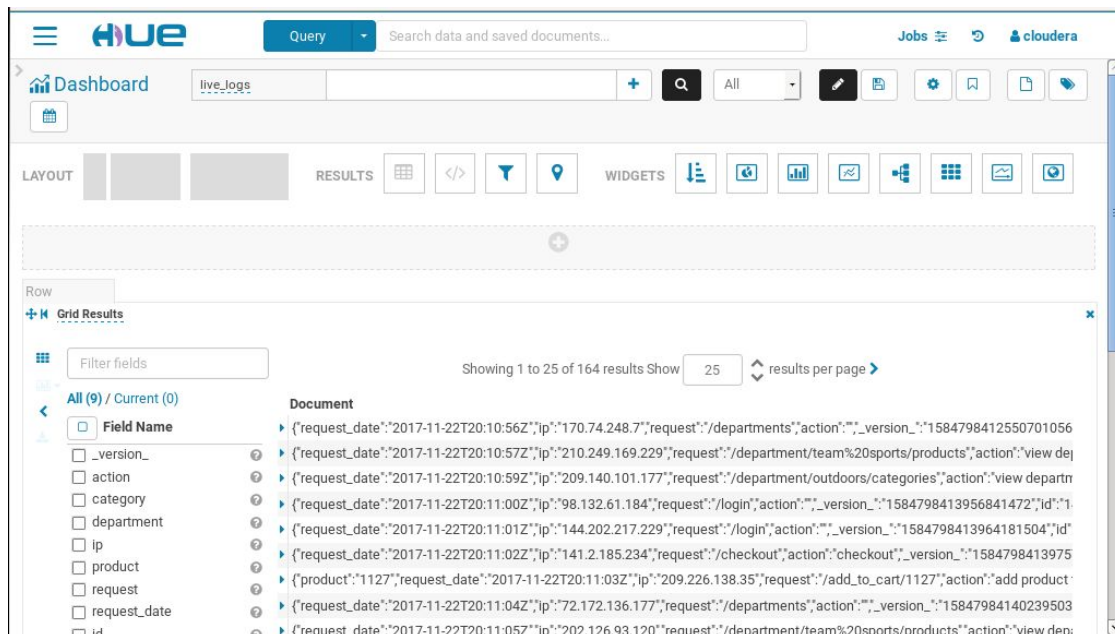
Para parar o processo basta digitar os comandos **Ctrl+C** e em seguida **stop_logs** na linha de comando do terminal para parar a geração de logs.

Agora você pode voltar para a *UI Hue* (consulte a página de orientação do seu cluster para o *link*) e clique em 'Search' na página da coleção (**Hue -> Indexes -> live_logs -> Search**)



Name	Type	ID	Required	Indexed	Stored
category	string		✓	✓	✓
product	string		✓	✓	✓
request_date	date		✓	✓	✓
ip	text_general		✓	✓	✓

Você poderá procurar, detalhar e navegar os eventos que foram indexados e realizar pesquisas com os dados atuais e em tempo real, com isto tendo condições de responder a demanda do cenário apresentado no início deste tutorial.



Pelo amor de nossa história, nós fingimos que você começou a indexar dados ao mesmo tempo em que começou a obter dados (via Flume) para a plataforma, de modo que, quando seu gerente escalasse o problema, você poderia detalhar imediatamente os dados dos últimos três dias e explorar o que aconteceu. Por exemplo, talvez você tenha observado muitos eventos de DDOS (Ataque de Negação de Serviço) e poderia tomar as medidas certas para prevenir o ataque. Problema resolvido! A gerência ficará fantasticamente feliz com suas contribuições recentes, o que naturalmente leva a um grande bônus, algo similar ou mais trabalho.

CONCLUSÃO

Agora você aprendeu a usar **Cloudera/Hadoop Search** para permitir a exploração de dados em tempo real, usando **Flume** e **Solr** e **Morphlines**. Além disso, você agora entende como você pode atender a casos de uso múltiplo nos mesmos dados - bem como a partir de etapas anteriores: sirva vários conjuntos de dados para fornecer maiores informações. A flexibilidade e a capacidade de carga múltipla de um Centro de Dados Empresarial baseado em **Hadoop** são alguns dos principais elementos que tornaram o Hadoop valioso para organizações em todo o mundo.