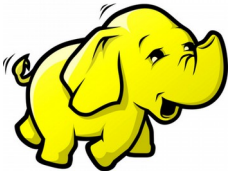


Cloudera Administrator Apache Hadoop

Parte 04-3 Gerenciamento de Jobs



Marco Reis
<http://marcoreis.net>

Agenda



- Criação de workflows
- Gerenciamento e agendamento de jobs no Oozie

Oozie



- O Apache Oozie é um workflow e coordenador para gerenciamento de jobs Hadoop
- Um workflow Oozie é um Directed Acyclical Graphs (DAGs), ou seja, é uma sequência de ações, como programas MapReduce, streaming, Hive, Sqoop, etc.
- O coordenador do Oozie Coordinator pode disparar os jobs com base em periodicidade e por disponibilidade de dados

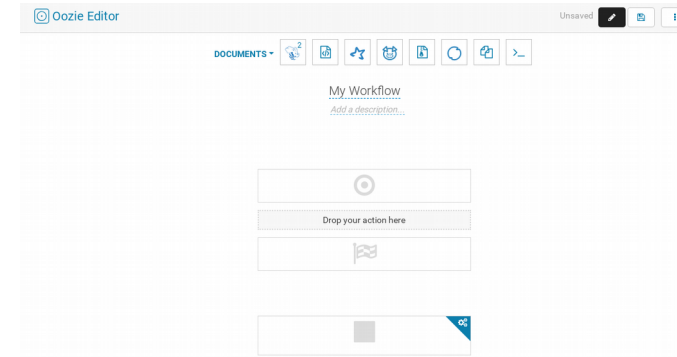
ETL



- O processo de ETL é um dos casos de uso para um cluster Hadoop
 - Extract – extrair os dados da fonte (Shell, Sqoop, ssh etc.)
 - Transform – selecionar e enriquecer os dados para atender aos requisitos (MapReduce, Java, Spark etc.)
 - Load – carregar os dados no armazenamento (Sqoop, Hive etc.)
- Nosso exemplo de ETL segue o roteiro:
 - Criar a tabela externa userstackoverflow_temp com os dados no HDFS
 - Criar a tabela Parquet userstackoverflow
 - Carregar a tabela Parquet com os registros da tabela externa
 - Remover a tabela externa e o arquivo XML

Editor do Oozie

- O editor do Oozie faz parte do Hue e é totalmente gráfico, no estilo arrastar e soltar
- Para as atividades, vamos usar o usuário “hive”
- São 2 conjuntos de operações:
 - Documentos: scripts já gravados no HDFS (imagem ao lado)
 - Actions: scripts usados apenas neste workflow (imagem abaixo)
- Todas as operações do fluxo devem estar gravadas em arquivos de script no HDFS
- Nosso exemplo vai usar 3 ações HiveServer2 e 1 ação Shell
 - Serão 3 operações no Hive e 1 operação no HDFS





Preparação dos arquivos

- Os scripts do workflow devem estar no diretório `/user/hive/scripts/`
- Os arquivos de dados devem estar no diretório `/user/hive/dados/staging/`
- Crie os scripts localmente e faça o upload com o próprio Hue → Files com o conteúdo a seguir
- Copie um dos arquivos de dados para o staging
 - Este será usado para a primeira carga do workflow

Scripts



01-create-table-user-temp.sh

```
drop table if exists datalake.userstackoverflow_temp;
create external table datalake.userstackoverflow_temp (
  id bigint,
  reputation bigint,
  creationdate string,
  displayname string,
  lastaccessdate string,
  websiteurl string,
  location string,
  views bigint,
  upvotes bigint,
  downvotes bigint,
  accountid bigint )
ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'
WITH SERDEPROPERTIES (
  "column.xpath.id"="/row/@id",
  "column.xpath.reputation"="/row/@Reputation",
  "column.xpath.creationdate"="/row/@CreationDate",
  "column.xpath.displayname"="/row/@DisplayName",
  "column.xpath.lastaccessdate"="/row/@LastAccessDate",
  "column.xpath.websiteurl"="/row/@WebsiteUrl",
  "column.xpath.location"="/row/@Location",
  "column.xpath.views"="/row/@Views",
  "column.xpath.upvotes"="/row/@UpVotes",
  "column.xpath.downvotes"="/row/@DownVotes",
  "column.xpath.accountid"="/row/@Accountid"
)
STORED AS
INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
OUTPUTFORMAT 'org.apache.hadoop.hive.q1.io.IgnoreKeyTextOutputFormat'
LOCATION '/user/hive/dados/staging'
TBLPROPERTIES (
  "xmlinput.start" = "<row",
  "xmlinput.end" = ">"
);
```


02-create-table-user-parquet.sh

```
SET parquet.compression=snappy;  
create table if not exists datalake.userstackoverflow  
stored as parquet  
as select * from datalake.userstackoverflow_temp limit 0;  
  
insert into datalake.userstackoverflow  
select * from datalake.userstackoverflow_temp;
```

03-drop-table-user-temp.sh

```
drop table datalake.userstackoverflow_temp;
```

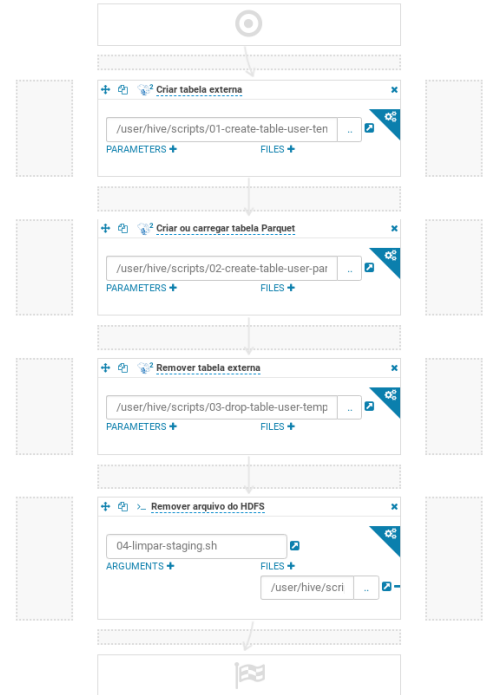
04-limpar-staging

Limpar diretório de staging

```
hdfs dfs -rm /user/hive/dados/staging/*
```

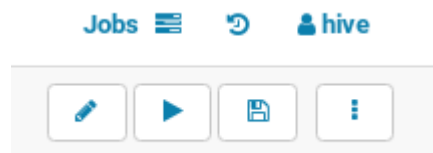
Workflow

- No final da atividade, o objetivo é ter um workflow como o mostrado ao lado
- O nome do workflow é etl-user-stackoverflow
- Arraste uma action HiveServer2 e selecione o arquivo correspondente no HDFS
- Observação: na action Shell temos seleccione o caminho do arquivo no link Files+ e na caixa de texto principal indique apenas o nome do script
- Não se esqueça de gravar o script!!!




Submissão do Job

- Para criar workflows complexos, o ideal é testar cada uma das etapas
- Como nosso workflow já foi testado, podemos simplesmente executá-lo com o botão submit
- Não precisa selecionar a opção dryrun, que é usada para testar sem confirmar as operações



Execução

- Durante a execução do workflow podemos acompanhar a evolução do processamento
- Os workflows executados estão disponíveis no menu Job

Job Browser

Jobs

Queries

Workflows

Schedules

Bundles

SLAs

Hive jobs are running as the 'hive' user

user:hive

Succeeded

Running

Failed in the last

7

days

Resume

Suspend

Kill

<input type="checkbox"/>	<div>Id</div>	<div>Name</div>	<div>User</div>	<div>Type</div>	<div>Status</div>	<div>Progress</div>	<div>Group</div>	<div>Started</div>	<div>Duration</div>
<input type="checkbox"/>	0000003-180917070315533-oozie-oozi-W	etl-user-stackoverflow	hive	workflow	SUCCEEDED	100%		September 17, 2018 7:46 AM	6m, 58s
<input type="checkbox"/>	0000002-180917070315533-oozie-oozi-W	etl-user-stackoverflow	hive	workflow	KILLED	100%		September 17, 2018 7:43 AM	2m, 15s
<input type="checkbox"/>	0000001-180917070315533-oozie-oozi-W	etl-user-stackoverflow	hive	workflow	KILLED	100%		September 17, 2018 7:39 AM	2m, 40s
<input type="checkbox"/>	0000000-180917070315533-oozie-oozi-W	etl-user-stackoverflow	hive	workflow	KILLED	100%		September 17, 2018 7:36 AM	2m, 45s

etl-user-stackoverflow

ID:
0000003-180917070315533-oozie-oozi-W

DOCUMENT
etl-user-stackoverflow

STATUS
SUCCEEDED

USER
hive

PROGRESS
100%

DURATION
0s

SUBMITTED
September 17, 2018 7:4...

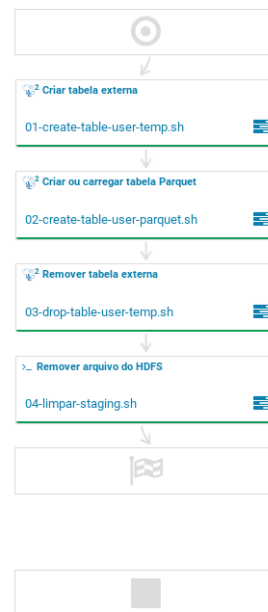
Graph

Properties


Logs

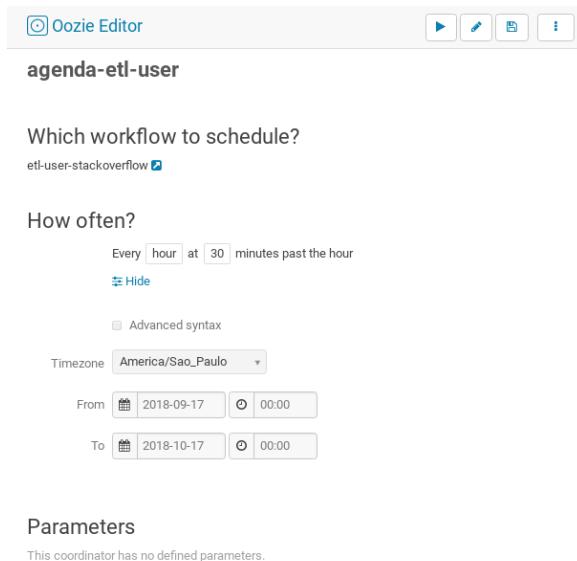
Tasks

XML



Editor de Agendamento

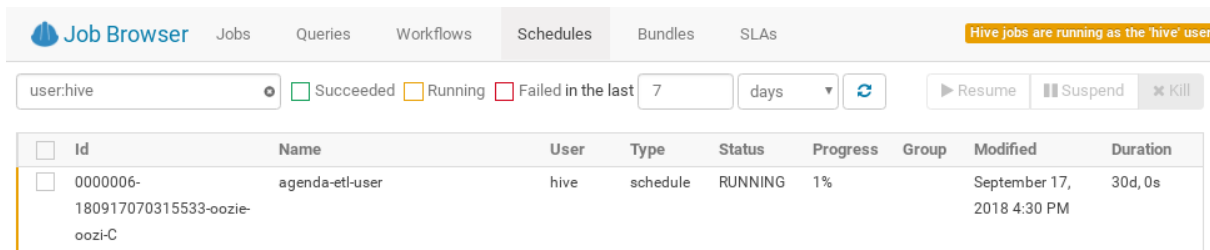
- O Oozie permite o agendamento dos workflows em períodos de tempo pré-definidos
- Vamos criar uma rotina para o workflow etl-user-stackoverflow que será executado a cada 30 minutos, durante 1 mês
- Para criar o agendamento, clique no botão Edit, depois no botão 
- Selecione a opção Schedule e preencha os dados como na tela ao lado
- Não se esqueça de salvar o agendamento!!!
- Para ativar a nova rotina, clique em Submit



The screenshot shows the Oozie Editor interface for scheduling a workflow. The title bar reads "Oozie Editor" with navigation icons. The main heading is "agenda-etl-user". Below it, the question "Which workflow to schedule?" is followed by the selected workflow "etl-user-stackoverflow". The "How often?" section shows a frequency of "Every hour at 30 minutes past the hour" with a "Hide" link. There is an "Advanced syntax" checkbox. The "Timezone" is set to "America/Sao_Paulo". The "From" date is "2018-09-17" at "00:00", and the "To" date is "2018-10-17" at "00:00". The "Parameters" section at the bottom states "This coordinator has no defined parameters."

Agendamentos

- Os agendamentos em execução, mesmo aqueles com estado de WAITING, estão mostrados no menu Jobs → Schedules



The screenshot shows the 'Job Browser' interface with the 'Schedules' tab selected. A notification at the top right states 'Hive jobs are running as the 'hive' user'. Below the navigation bar, there is a search filter 'user:hive', a legend for job states (Succeeded, Running, Failed in the last 7 days), and action buttons for Resume, Suspend, and Kill. The main table lists a single schedule job with the following details:

<input type="checkbox"/>	Id	Name	User	Type	Status	Progress	Group	Modified	Duration
<input type="checkbox"/>	0000006-180917070315533-oozie-oozi-C	agenda-etl-user	hive	schedule	RUNNING	1%		September 17, 2018 4:30 PM	30d, 0s

Dúvidas?

Marco Reis
<http://marcoreis.net>