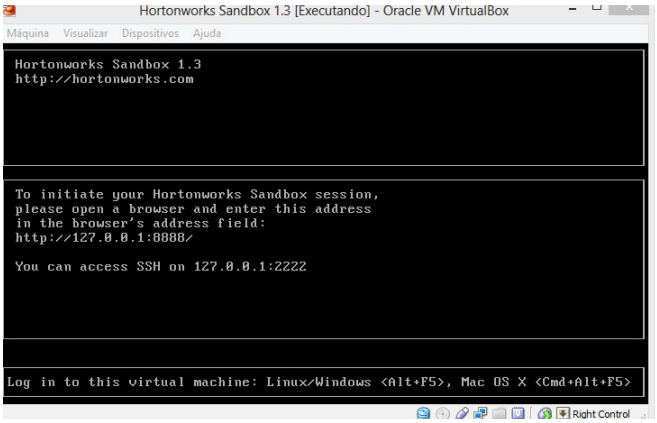
1) Para utilização das práticas com o Hadoop vamos usar uma sandbox (máquina virtual) fornecida pela HortonWorks. Esta máquina virtual possui os seguintes produtos instalados:

Apache Hadoop	1.2.0
Apache Hive includes HCatalog	0.11.0
Apache HBase	0.94.6.1
Apache ZooKeeper	3.4.5
Apache Pig	0.11
Apache Sqoop	1.4.3
Apache Oozie	3.3.2
Apache Ambari	1.2.4
Apache Flume	1.3.1
Apache Mahout	0.7.0

Esta é a versão 2.3.2 do produto e pode ser obtida em: https://hortonassets.s3.amazonaws.com/2.1/virtualbox/Hortonworks_Sandbox_2.1.ova

2) Inicialize a máquina virtual. Espere que a seguinte tela seja exibida:

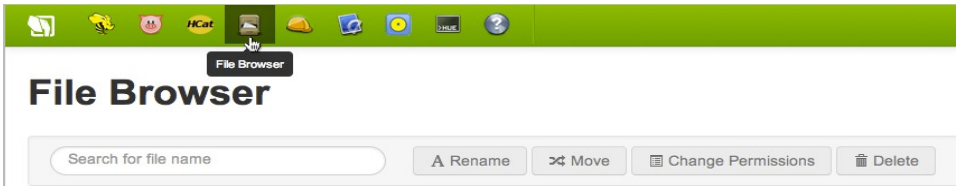


3) Abra um navegador e execute a *url* http://locahost:8000 ;

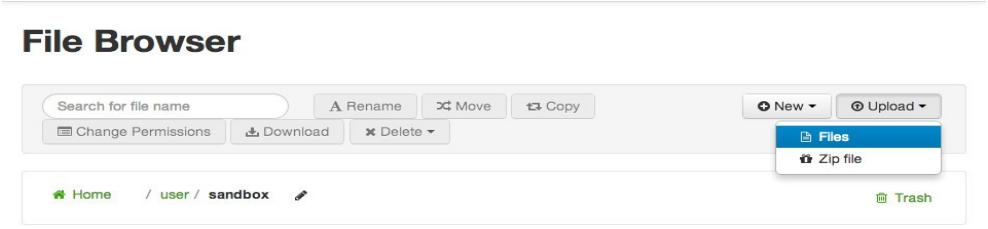
4) Para este tutorial será utilizado uma base de dados com estatísticas de jogos de baseball entre os anos 1871 até 2011. Este arquivo possui 95.000 linhas. A base de dados pode ser acessada em: http://seanlahman.com/files/database/lahman591-csv.zip .

5) Apenas dois arquivos serão utilizados: master.csv e batting.csv

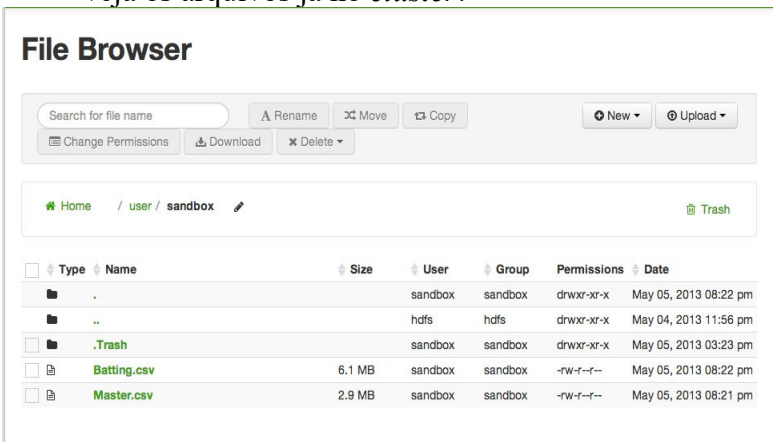
6) Fazendo Upload dos arquivos:
Selecione a opção 'File Browser' no topo da página



- Navegue na estrutura para o root e entre na pasta */data;*
- Clique no botão *upload* e selecione a opção *files*

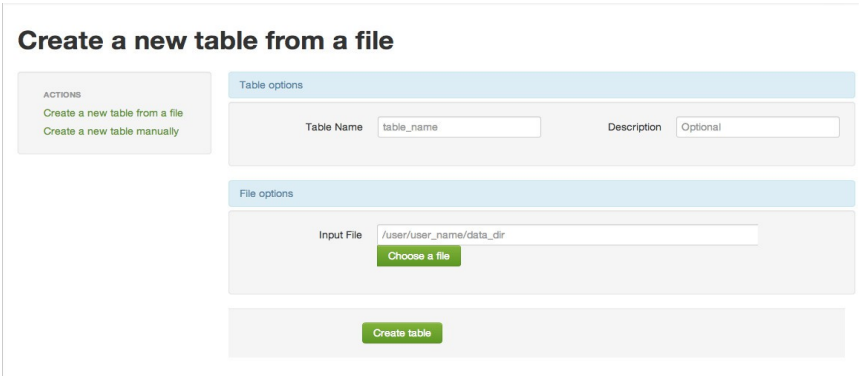


- Selecione os dois arquivos: master.csv e batting.csv;
- Veja os arquivos já no *cluster*:

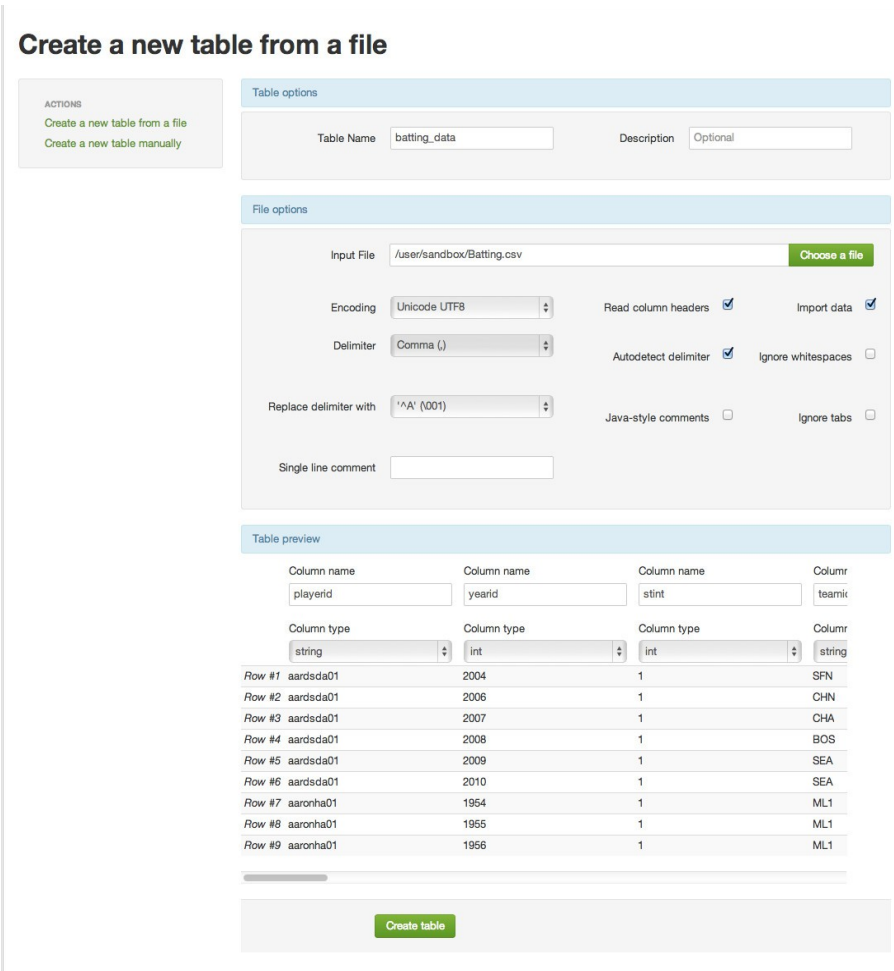


7) Criando as tabelas utilizando HCatalog
   - Com arquivos já carregados no HDFS, agora serão criadas as tabelas utilizando o HCatalog/Hive (metastore). Selecione a opção HCat no menu:

   - No menu '*Actions*' selecione a opção '*Create a new table from a file*';



   - Criar uma tabela para o arquivo batting.csv:
     - nome da tabela: batting_data;
     - O item 'optional' deixe em branco;
     - Selecione o arquivo;
     - Veja que algumas opções já estarãomarcadas, por padrão. E o arquivo será pré exibido;



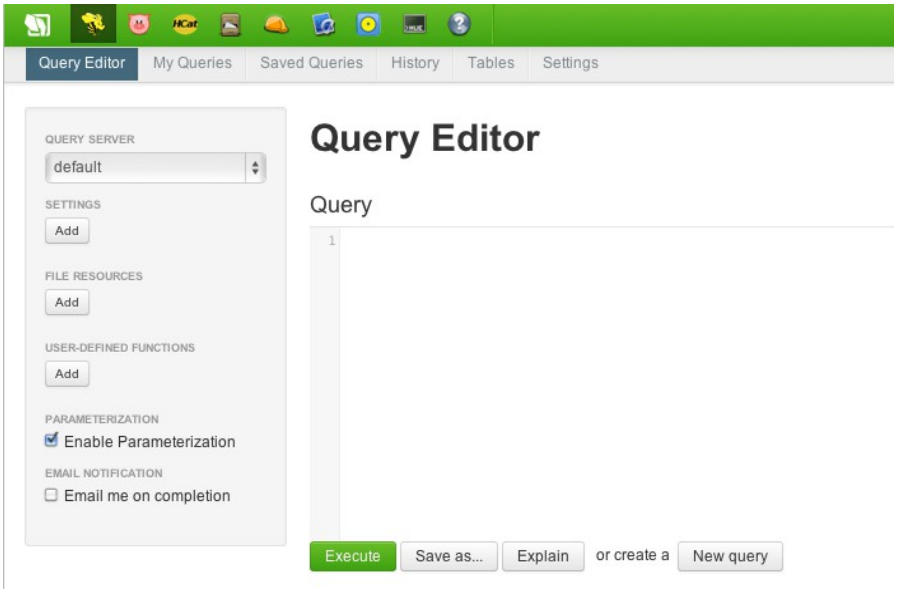
- Altere o nome da coluna 'r' para 'runs' e altere se o tipo de dado para '*int*';
- Clique no botão '*create table*';

- Criar uma tabela para o arquivo master.csv:
  - nome da tabela: master_data;
  - O item 'optional' deixe em branco;
  - Selecione o arquivo;

- Veja as duas tabelas criadas no HCatalog:

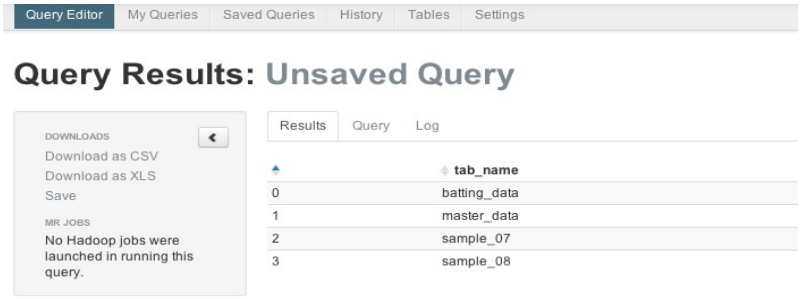
- Clique no botão '*Browse Data*' para visualizar as tabelas criadas;

8) Fazendo algumas análise com o Hive
   - Arquivos já carregados no HDFS;
   - Metadados já criados para o Hive. Desta forma pode-se trabalhar emitindo-se comandos em HiveQL;
   - Para emitir este comando será utilizado uma ferramentas chamada **Beeswax**: interface interativa para o Hive em que poderá ser digitado consultas e as mesmas serão avaliadas pelo Hive e transformadas em uma série de jobs *Map Reduce;*
   - Para executar o Beeswax acesse o menu conforme figura abaixo:



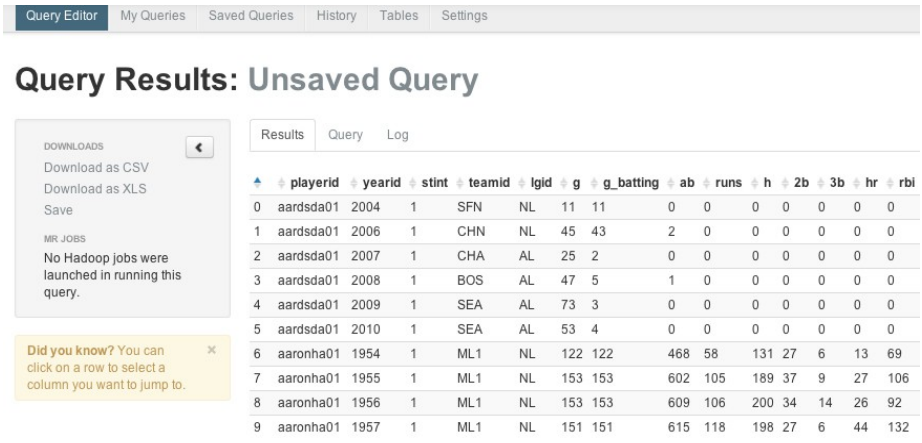
- As consultas podem ser digitadas e para execução acione o botão '*Execute*'. Não é possível digitar mais de uma consulta, separando-as por “;”, por exemplo.
- Como HCatalog e Hive são integrados, nas verdade o mesmo produto, tudo que foi criado no HCatalog é válido para o Hive;
- Digite o comando para visualizar as tabelas:

```
show tables
```



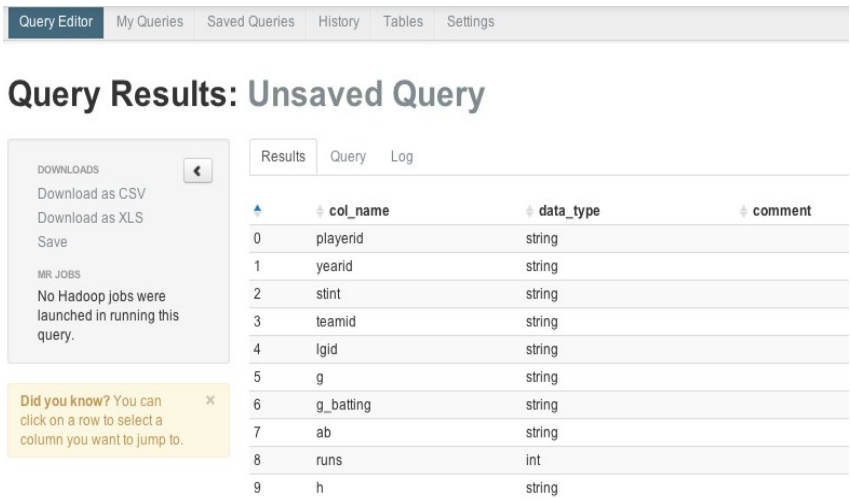
- Hive utiliza o esquema/banco de dados criado no HCatalog. Para estes caso,não é necessário utilizar o nome do esquema. Digite o comando;

```
Select * from batting_data
```



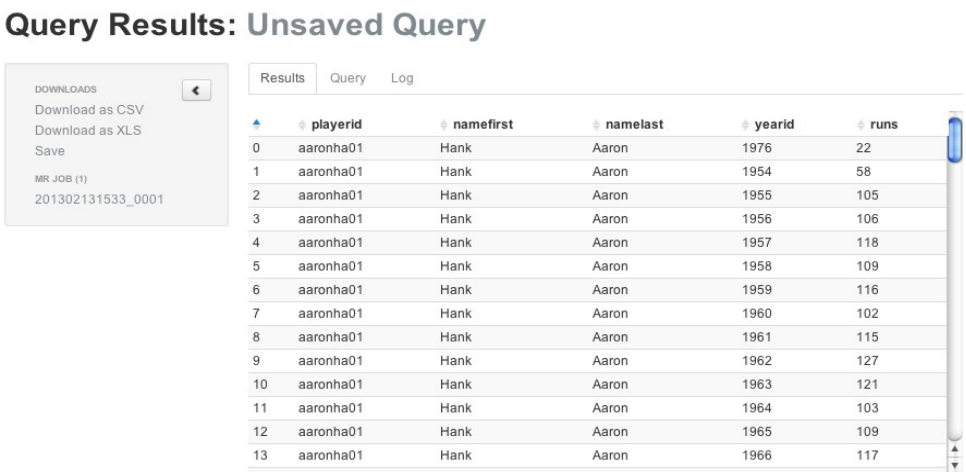
-É também possível visualizar as colunas de uma tabela com o comando;

```
describe from batting_data
```



- Utilizando o Hive também é possível fazer junções dos dados. Digite o comando abaixo para testar uma junção;

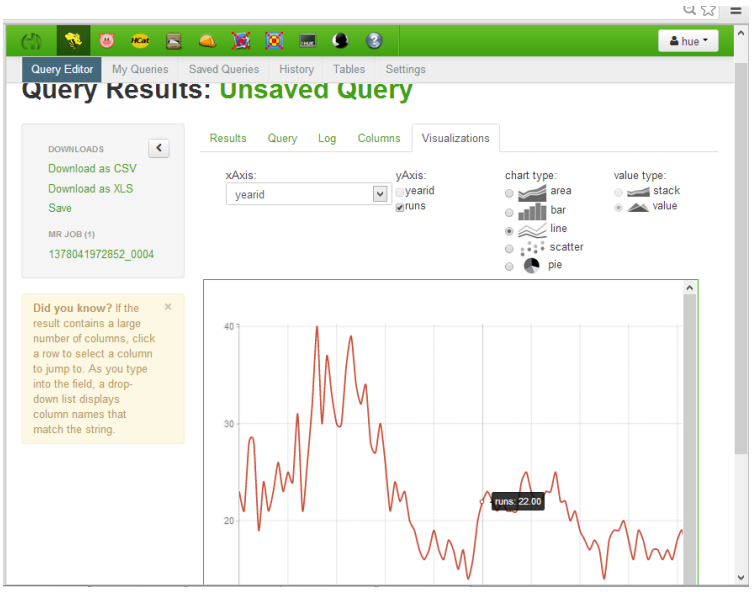
```
select m.playerid, m.namefirst,m.namelast,
       b.yearid,b.runs
from master_data m
join batting_data b on (m.playerid = b.playerid )
```



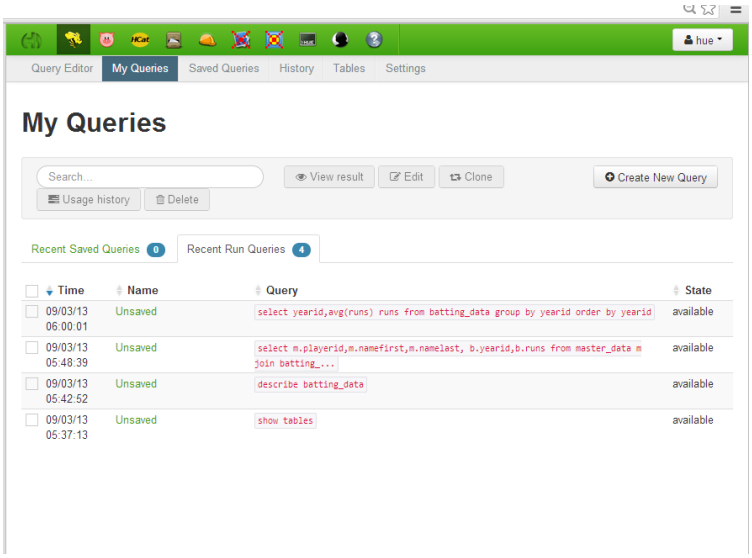
- Observe a geração dos códigos para o jobs Map Reduce. Veja o log. Observe também que é gerado um número para o processo job Map Reduce: '*MR JOB(1)';*
- Outra funcionalidade é a visualização, Clique na aba '*Visualizations*'. Não altere a visualização. Vamos criar um resultado mais adequado para ser apresentado:

```
select yearid,avg(runs) runs
from batting_data
group by yearid
order by yearid
```

- Clique na aba '*Visualizations*' e altere para um gráfico de linhas:



- Outras funcionalidades do Beeswax são armazenar ou visualizar as consultas já emitidas:



- Pode-se ainda consultar um histórico de consultas já emitidas(lembre-se que toda consulta foi convertidas para um job *Map Reduce*);