**Overview**

[**Apache Pig**](http://hortonworks.com/hadoop/pig/) is a platform for analyzing large data sets. It comprises of a high-level language named 'Pig Latin' for expressing data analysis programs, coupled with the infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

In this tutorial, you will learn the following topics:

1. Load a data file into HDFS.
2. Learn about 'FILTER, FOREACH' with examples.
3. Storing values into HDFS.
4. Learn about Grunt shell's File Commands.

**Prerequisite**

A working Hadoop cluster – the easiest way to get a pre-configured and fully functional Hadoop cluster is to download the [Hortonworks Sandbox here](http://hortonworks.com/sandbox).

After downloading the Sandbox and running the VM, we will SSH in to follow the rest of the steps.

ssh root@127.0.0.1 -p 2222;

the password is hadoop

**Step 1:**

**Creating a data file.**

**a.** Create a text file named "movies.txt" in your local file system on the sandbox and add the following content:

1,The Nightmare Before Christmas,1993,3.9,4568

2,The Mummy,1932,3.5,4388

3,Orphans of the Storm,1921,3.2,9062

4,The Object of Beauty,1991,2.8,6150

5,Night Tide,1963,2.8,5126

6,One Magic Christmas,1985,3.8,5333

7,Muriel's Wedding,1994,3.5,6323

8,Mother's Boys,1994,3.4,5733

9,Nosferatu: Original Version,1929,3.5,5651

10,Nick of Time,1995,3.4,5333

**b.** Load the file "movies.txt" into a directory on HDFS named 'user/hadoop' using the shell command as shown.

hadoop fs -put movies.txt /user/hadoop

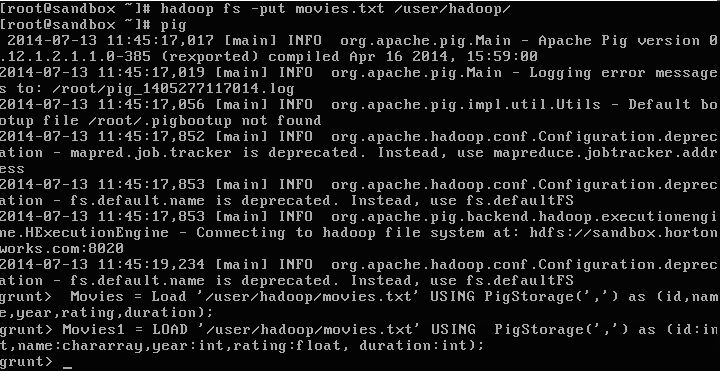
**c.** Using Pig's Grunt Shell Interface.  
Type "pig" at the local shell prompt to get into Pig's Grunt shell.  
Load the content of "movies.txt" into a variable named "Movies".

grunt> Movies = LOAD '/user/hadoop/movies.txt' USING PigStorage(',') as (id,name,year,rating,duration);

Or, if you want to assign types:

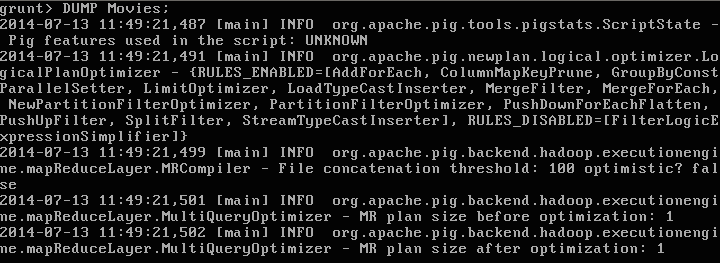
grunt> Movies = LOAD '/user/hadoop/movies.txt' USING PigStorage(',') as (id:int,name:chararray,year:int,rating:float, duration:int);

The commands are executed as below:



To see the content of the variable "Movies", use the following command:

DUMP Movies;

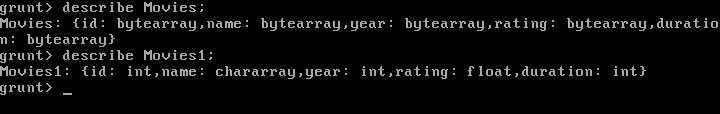


The command 'DUMP' would execute as follows:



To check the format of the variable "Movies", use the following command:

Describe Movies;



**Step 2:**

**Filtering Data**

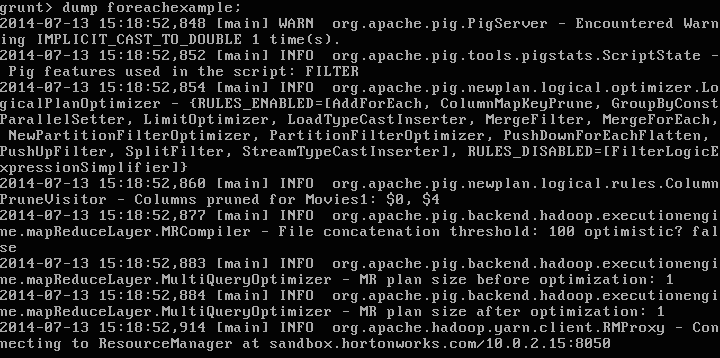
Now that the data is loaded into the variable "Movies", let's filter the data for movies with a rating of greater than 3.5 using the following command:

grunt>movies\_greater\_than\_three\_point\_five = FILTER Movies BY rating>3.5;

From the variable 'movies\_greater\_than\_three\_point\_five', let's' extract the values for 'year','rating', and 'moviename' and save them in another variable named 'foreachexample'.

grunt> foreachexample= foreach movies\_greater\_than\_three\_point\_five generate year,rating,name;

grunt> dump foreachexample;

foreachexample  


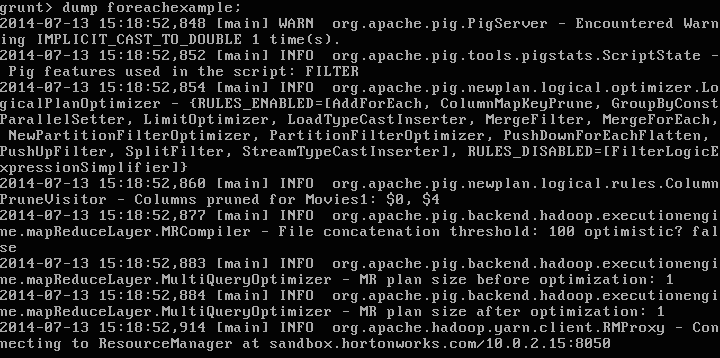
Filter result

**Step 3:**

**Storing variable values into HDFS.**

Let's store the values of variable 'movies\_greater\_than\_three\_point\_five' into HDFS:

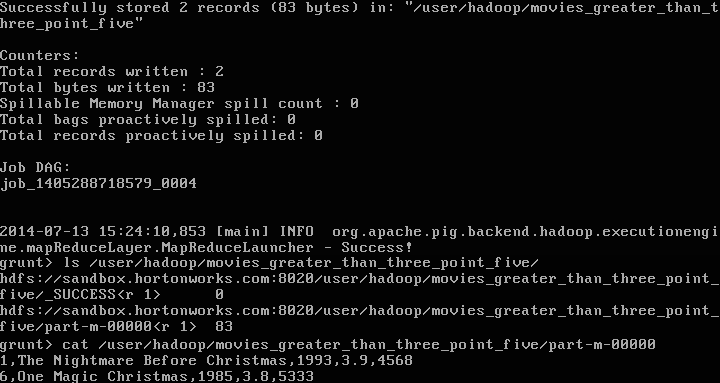
grunt> STORE movies\_greater\_than\_three\_point\_five INTO '/user/hadoop/movies\_greater\_than\_three\_point\_five' USING PigStorage (',');



On any error conditions, Hadoop immediately throws an exception.  
In the above case, there was a 'file not found' error.

Now that we have the data in HDFS, use the 'cat' command to open the processed file:

grunt> cat /user/hadoop/movies\_greater\_than\_three\_point\_five/part-m-00000



**Step 4:**

**File Commands**

Pig's Grunt shell has commands that can run on HDFS as well as on the local file system.

grunt> cat /user/hadoop/movies.txt

grunt> ls /user/hadoop/

grunt> cd /user/

grunt> ls

grunt> cd /user/hadoop

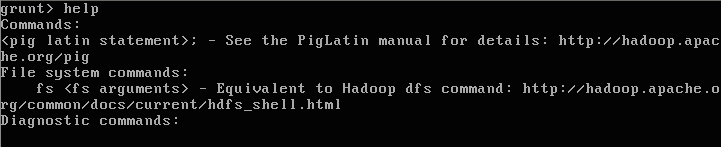
grunt> ls

grunt> copyToLocal /user/hadoop/movies.txt /home/

grunt> pwd

**Step 5:**

**To get help in Pig, simply type "help" in the Grunt shell.**



This tutorial gives you a glimpse of how to explore data with Apache Pig.