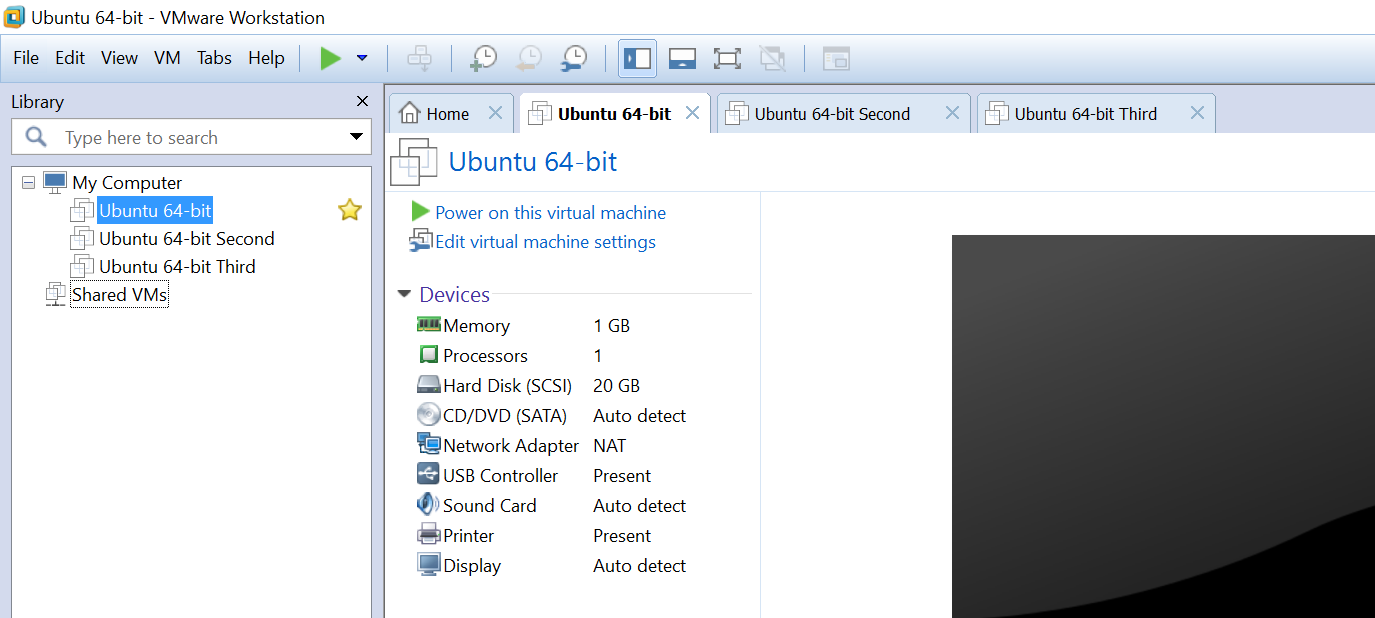
**Wiki species in Fully Distributed mode**

Broadly speaking, following steps were performed in this assignment

1. VM installation

Installed 3 Linux Virtual machines through VMWare on my system.

One as Master and other two as slaves.



Hadoop 0.20.2 was installed on all these 3 machines and multi node cluster was setup.

1. Fix Bugs
   1. Seed modification

One can land on a web page broadly via two methods - out links and in links or other factors like directly typing the URL. Page Rank algorithm has a seed factor to cater the second method. In our program I changed the seed in the Reducer code of Builder and Iterator code of Mapper. The value derived here is by using the n/m formula where m is the total no of pages

* + 1. Builder Reducer

Changed the type of count variable to float. (One can choose double as well)

And increment it by 0.1 inside the loop

* + 1. Iterator Mapper

Change the value of the variable *currscore* to 0.5;

* 1. Other improvements

Changed the code of Builder Mapper to remove the colons and replace them with underscore.

* 1. Other Improvements

Removed the code line that outputs for better results - *new Text(toEmit + ": " + pagetitle)*

After fixing these bugs, I also changed the code to include the URLs that don’t have any inlinks and outlinks to include the pages that can be reached directly and not through links.

Then I ran the builder, iterator and viewer code in Fully distributed mode. Screenshots are attached in separate folder.

The output is much better than before. I have attached the output in separate folder as well.

Future work – Remove some junk records like FR, EN etc.