1. Download the latest Cloudera CDH QuickStart VM from the below link <https://www.cloudera.com/downloads/quickstart_vms/5-10.html>
2. Please select the image based on your platform viz., VMWare/VirtualBox/Docker.
3. The below documentation gives you steps to configure CDH QuickStart VM <https://www.cloudera.com/documentation/enterprise/5-6-x/topics/cloudera_quickstart_vm.html>
4. I would recommend to allocate 4GB to your VM & Manage the apps through command line scripts rather than going for Cloudera Manager (It requires minimum allocation of 8GB)
5. Try to browse Hue using the below URI. You can visualize the files in the HDFS by selecting File Browser option on the right hand side of the dashboard (Default Username/Password : cloudera/cloudera) http://quickstart.cloudera:8888/filebrowser/
6. Make sure you are on latest JAVA & Maven (I had issues when using JAVA 1.7 & older Maven)
7. Download the latest Wikipedia dump from <https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles-multistream.xml.bz2>
8. Push the data from local file system to HDFS **$hdfs dfs -put <local path> <hdfs path>**
9. Download the code from git (Fork of Advanced Analytics using Spark) <https://github.com/YashasRattehalli/aas/tree/master/ch06-lsa>
10. Build the JAR using Maven (Please refer Maven documentation and configure you pom.xml accordingly)
11. Submit the Spark Job & check the results !!! **spark-submit --class com.cloudera.datascience.lsa.RunLSA --master local target/<your jar name/ch06-lsa-2.0.0-jar-with-dependencies.jar> <parameter1> <paramter2> .. <parameterN>**