Setting up hadoop cluster

We have configured 2 node pseudo distributed cluster in hadoop which implies that both the nodes run on the same system in multiple JVMs.The datanodes have been configured to run on port number 50022 and on default port.Other important configurations have been specified in the conf(default properties) and conf2 folder supplied with the documents.

Starting hadoop daemons:

Once the cluster has been configured , all the required daemons are started.The default configuration daemons are started by running the predefined script start-all.sh.The daemons on the extra node are started by running hadoop-daemon.sh with extra arguments for starting datanode and its tasktracker.

Transferring data from local machine to HDFS:-

The hadoop framework's map reduce classes i.e. Mapper and Reducer can process data that is present in the HDFS. So before submitting a map reduce job , we have to transfer our aadhar data from our local machine to HDFS. During the transfer process the HDFS breaks the file into 5 blocks and stores all of them on both nodes.This is consistent with the fact that replication factor is 2. We have specified this replication factor in hdfs-site.xml in configurations (both) folders.The default block size has been 64 MB.

Submitting Map reduce Job:

To process the data we supplied a jar file containing classes from aadhar.java and Tokens.java. Aadhar.java contains job description as well as mapper and reducer classes. Tokens.java contains personalised class which works somewhat similar to string tokenizer class. Once the job has been successfully submitted ,it will start running on the datanodes available and percentage of maps and reduce completed are shown simultaneously.

Retrieving Results:

Once the job has been successfully completed the results from the job are stored on HDFS in a direcctory specified in job run parameters. The data is trasferred from HDFS to local filesystem where it is accompanied by history of logs and a success document.

Visualising the exctracted data:

We have used R language to perform statistical calculations on the results. Once the results have taken the desired form(i.e. In the form of matrix or dataframe) the plots are generated by writing suitable R scripts.

Publishing the results:

There are numerous scenarios in which the results from our project can be used. They can be used one time by a single person or committee ,in that case it is best suitable to provide the report consisting of processed results and their respective graphs. But in a scenario in which the extracted results are used over a period of time by a group of users, it is best suitable to publish the results on a web application(asp.net application in our case). It enables best use of the results to design better policies related to aadhar id cards.

We have plotted the graphs on

1. Total aadhar cards gernerated to not generated
2. State wise Male to female aadhar id cards generated
3. Agewise aadhar id cards generated
4. Overall male to female participation in all states.

One of the graphs that we had plotted is shown below

