Data gathering

1. Plan to go for the extra credit so we collect the data early.

As first step we collect 100 k data and do the coding based on those data.

1. How we collect the data

Collect in the Hadoop cluster and get a zip file contains the data.

1. Format of data

1::google.com::http://google.com::206710::Google

Delimiter - ::

non English characters

1. What type of data is there
   1. UNKNOWN – meaning
   2. Loading... – meaning

How to run Hadoop

Put those file in the Hadoop cluster and unzip.

Commands used for transfer files form one server to another.

//Transfer directory to another linux server

scp -r /home/cmahapat/unzip/ hadoop1:. :

// Copy directory contaning files to hadoop cluster

hadoop fs -copyFromLocal unzip /cmahapat/

Then ran the commands.

// Run the jar file

hadoop jar Homework2.jar Question1 /cmahapat/unzip /cmahapat/output1/1

// Copy the output file from hadoop cluster to local machien

hadoop fs -copyToLocal /cmahapat/output1/1/part-r-00000

// Open the output file with use of nano editor

nano part-r-00000

Output

Question 1 – output

UNKNOWN 3.88812182E8

Google 4.0678177E7

Yahoo 2.6078382E7

Cartoon Network | Free Games, Online Videos, Full Episodes, and Kids TV Shows 1.917354E7

Home 1.8064942E7

Universitas Negeri Jakarta 1.415988E7

アルバイト・バイト＆派遣･正社員の求人情報【求人アスコム】 1.1749869E7

Estate Sales, Downsizing and Liquidations via Fully Managed Online Auctions | MaxSold 1-888-672-3677 - There&#039;s an auction near you! 9879165.0

BSNL Employees Union... 7986210.0

Classtize - Post Free Classifieds Without Registration, Free Online Advertisment 7457105.0

Question 2 – output

UNKNOWN 595

Google 13

Loading... 4

Search 4

Welcome to nginx! 4

Home 3

Charter Spectrum® Official - Internet, Cable TV, and Phone Service 3

�ʺ��g - �K�O�ʵe���e���ɪ��s ! 2

ImageShack - Best place for all of your image hosting and image sharing needs 2

Live webcams sex chat - live sex shows! ... Xcams ... 2