Handling Big Data

Humera Noor Minhas

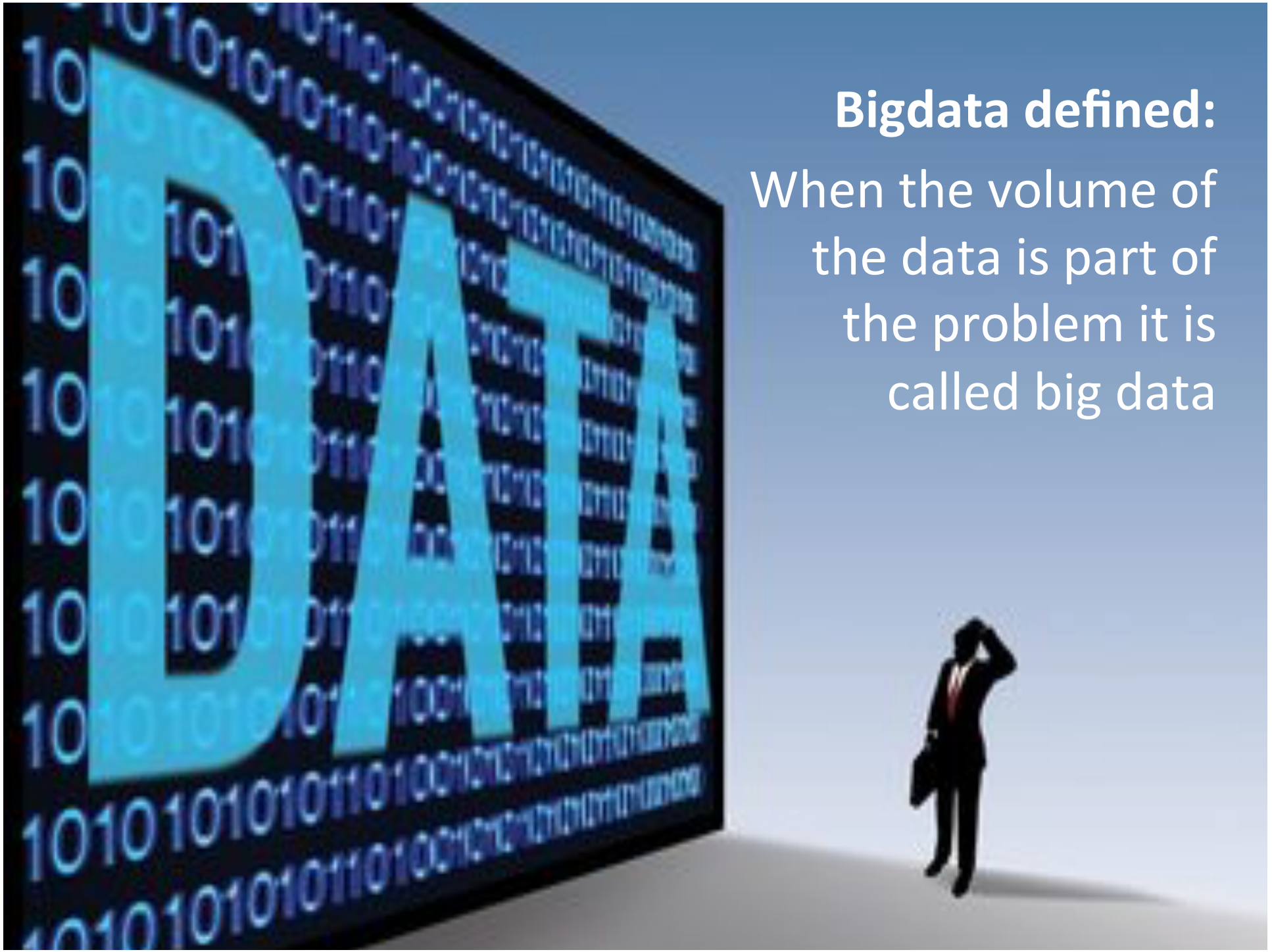
Cliqz

رَبِّ اشْرَحْ لِیْ صَدْرِیْ o وَ یَسِّرْلِیْٓ اَمْرِیْ o وَاحْلُلْ عُقْدَةً مِّنْ لِّسَانِیْ o یَفْقَهُوْا قَوْلِیْ o (طٰہٰ:28-25)

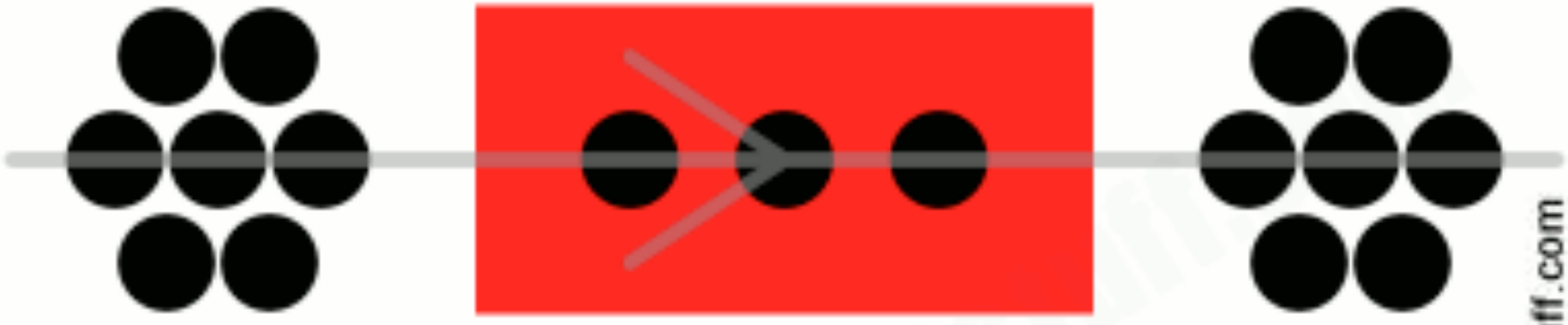
اے میرے رب! میرا سینہ کھول دے اور میرے لیے میرا کام آسان کر دے اور میری زبان کی گرہ کھول دے تاکہ لوگ میری بات سمجھ سکیں۔

Agenda

- Serial processing
- Parallel processing
- MapReduce
- Counting example (offline mode)
- Explore big data (online mode)


DATA
Bigdata defined:
When the volume of the data is part of the problem it is called big data

Serial processing



Parallel processing

A counting example


teasers
MilkyWay
Twix
Galaxy
Caramel
Galaxy
mars


Mars


Quality
Street


Mars


Quality
Street

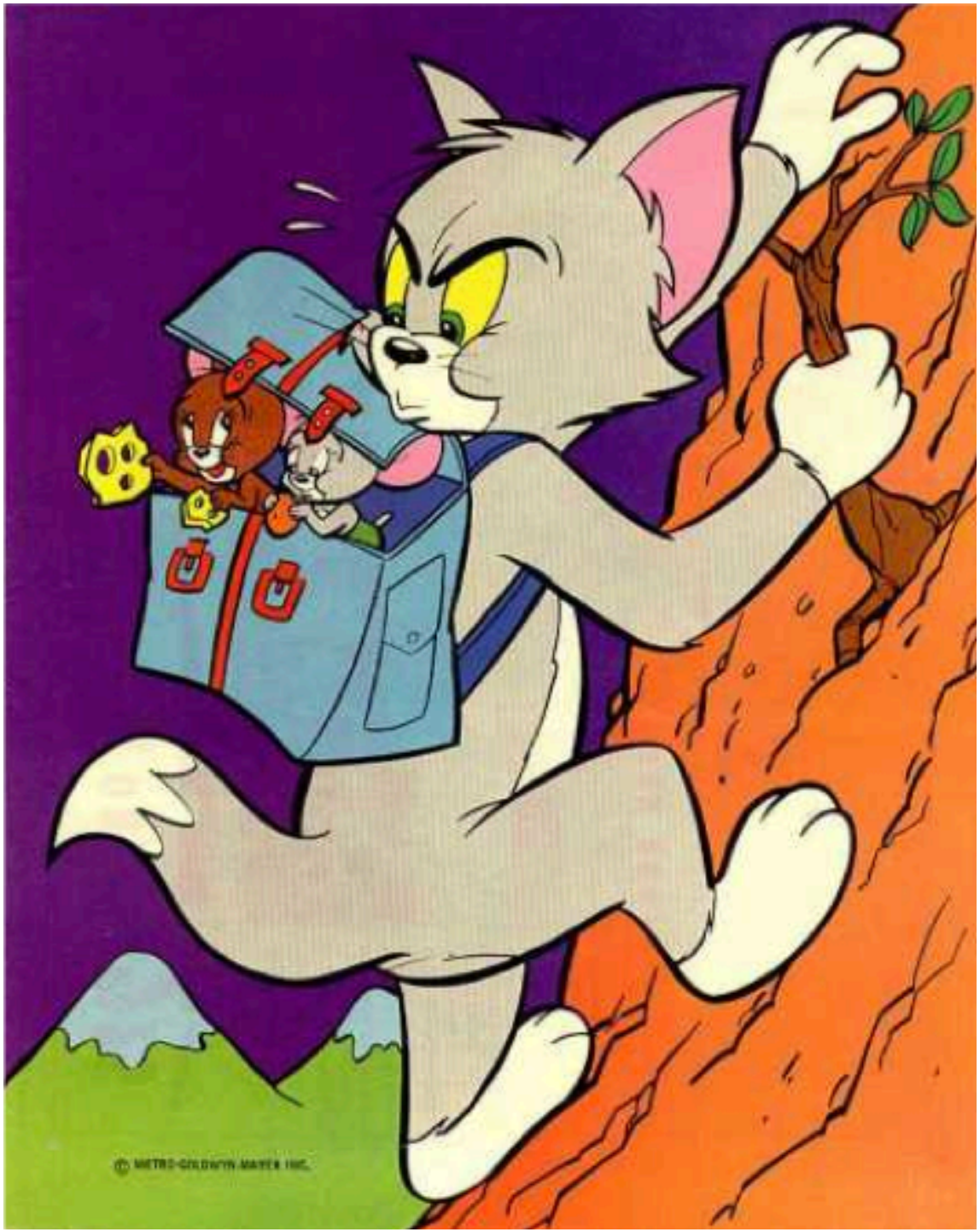

Mars

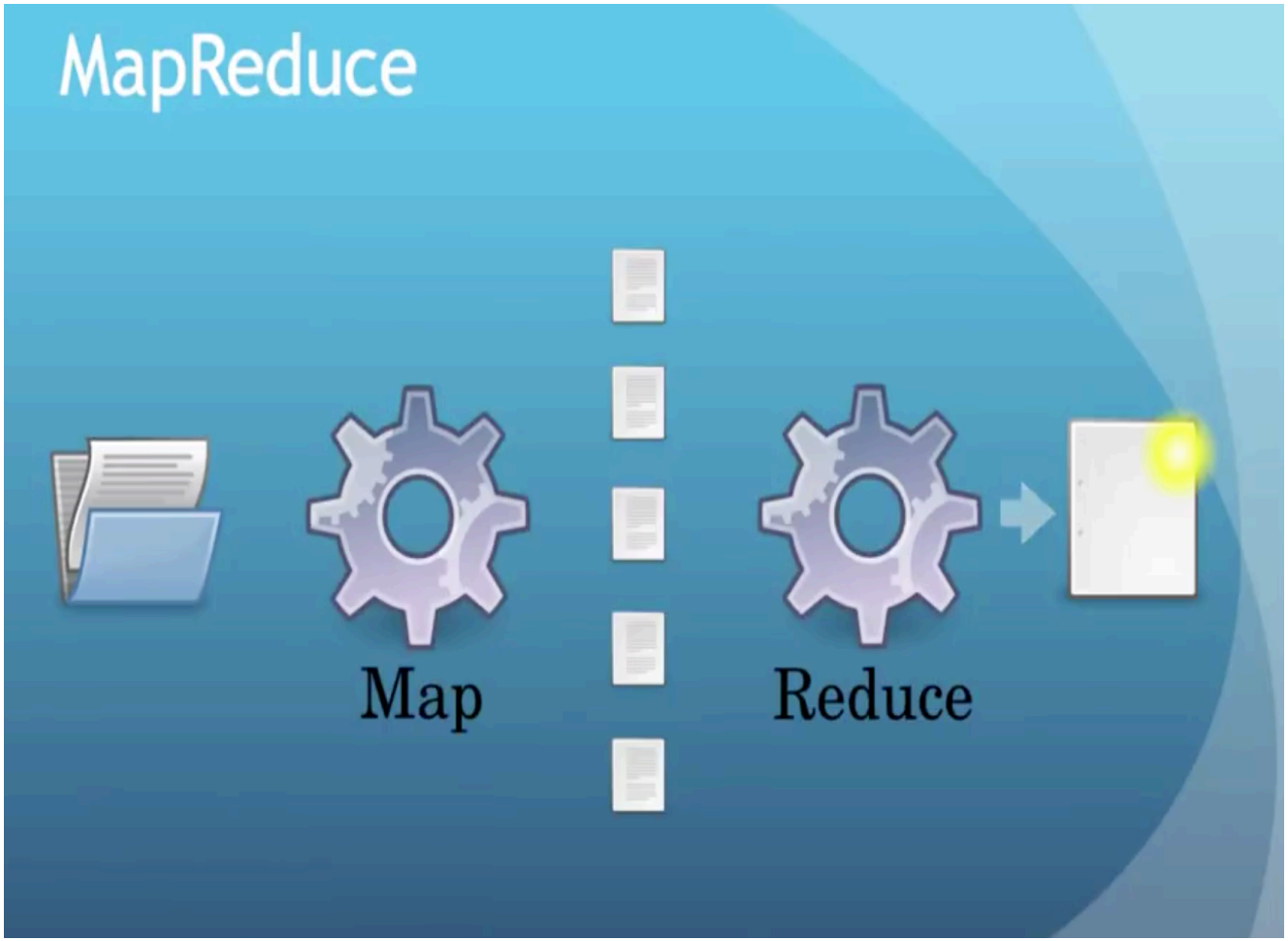

Quality Street
Quality Street


SNICKERS


Maltesers
teasers

MapReduce





https://**www.youtube.com**/watch?v=bcjSe0xCHbE

Let’s code a real-world problem

Data

```
643 240 1498075864.0    p4-relatedvideos    https://www.youtube.com/watch?v=Q2Dd8EinF-k https://www.youtube.com/watch?v=EmPRxKZYQSI
586 240 1498082886.0    p3-youtube  z12+qmobile https://www.youtube.com/watch?v=QyBNoal7vWA
410 808 1498035863.0    p0-googlequery  %EC%A0%9C%EC%A3%BC+%EB%B2%A4%EC%B2%98%EB%A7%88%EB%A3%A8 https://www.google.co.kr/?gws_rd=ssl#q=%EC%A0%9C%EC%A3%BC+%EB%B2%A4%EC%B2%98%EB%A7%88%EB%A3%A8
586 240 1498070324.0    p0-googlequery  Trolls  https://www.google.com.pk/search?q=latest+comedies&rlz=1C1LENN_enPK510PK512&oq=latest+comedies&aqs=chrome.0.0l6.7259j0j7&sourceid=chrome&ie=UTF-8#q=Trolls&stick=H4sIAAAAAAAAAGWRPWvbQBjHcyYv9mOHOhcXgijEZCgeCopOlmKytqEUEgJuhm6mOp1k2Xq_w5Y8ZevcUOhH6NKpkI_QofkAGUPolm-QMZJrX2iy3e95_vfj-F91bQ_UQN2Pe5ljTEnL_ywYF20aBcz2GG-Tfc28QvMId4eTcb4Ak8zCrnWF6iVopJdmeZHbVF1V06zcjgejlMutSQI-kVszTnuDrrMQEUP4Y1dG9XHmaIvVbJqEWWGplWAN0iRZxjSDkkyXxh510oHNJR8QU2jWSEonSTCiC2lGrANhyGjmuLM4sZdM9SCxcm7Kq6411LNHsS_Gw-ny6f86-4u-VurNu_ubbeVL5eLnn2t0XoHGu6LHftFk6lGBO7BaMm5DVVktG-1U9lqAYaN6-6KJdhCeHxo7u_g3gvpHJs6ik8j2nBxfIvwLwXqfURYK_OPJ9jvC3xA0j6OIMz_vs_Lz7LMIv4b1o1B4IsevYFOZ12fEhqF3VpQ6PCL-9L_uA34PtRMWWCzlpw4-BHgb-T6jwotC_AZeKlsqlQPV8fyAF8ZteD5uPADmWsPDVwIAAA
586 23017    1498023367.0    p1-googlesearch biometric+attendance+machine+pakistan    http://www.srt.com.pk/biometric-attendance-system.html  https://www.google.com.pk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=11&cad=rja&uact=8&ved=0ahUKEwiO0v3Sk87UAhWMsY8KHZiHBZU4ChAWCD4wAA&url=http%3A%2F%2Fwww.srt.com.pk%2Fbiometric-attendance-system.html&usg=AFQjCNG5O4Z89-92Nl3WFQNbc9etk3oogw    simple1
158 808 1498036586.0    p4-relatedvideos    https://www.youtube.com/watch?v=O9opzluPJU0 https://www.youtube.com/watch?v=xK0Q0UaNT2E
643 23017    1498062201.0    p0-googlequery  %D1%86%D1%81%D0%BF+%D1%82%D0%BE%D0%BC%D1%81%D0%BA+%D0%BB%D0%BE%D0%B3%D0%BE  https://www.google.ru/search?q=%D1%86%D1%81%D0%BF+%D1%82%D0%BE%D0%BC%D1%81%D0%BA+%D0%BB%D0%BE%D0%B3%D0%BE&newwindow=1&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiD2pXJrM_UAhVhMZoKHSJNBMsQ_AUIBigB&biw=1073&bih=586
862 23017    1498085868.0    p4-relatedvideos    https://www.youtube.com/watch?v=_ff7AZtLb08 https://www.youtube.com/watch?v=_ff7AZtLb08
643 240 1498065586.0    p1-googlesearch %D0%B1%D0%B3%D0%BC%D1%83+%D0%BC%D0%B5%D0%B4%D0%B8%D0%BA%D0%BE-%D0%BF%D1%80%D0%BE%D1%84%D0%B8%D0%BB%D0%B0%D0%BA%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9+%D1%84%D0%B0%D0%BA%D1%83%D0%BB%D1%8C%D1%82%D0%B5%D1%82   https://www.bsmu.by/page/6/91/  https://www.google.ru/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiY6tqXuc_UAhXpF5oKHYRpAFoQFggnMAA&url=https://www.bsmu.by/page/6/91/&usg=AFQjCNGr6X0Mqpf8JKCf_MV4YCb5dgWvKQ&sig2=5p-X1aKgJTXhyFUZNdLgEw&cad=rjt    simple1
586 23017    1498043305.0    p4-relatedvideos    https://www.youtube.com/watch?v=b6-YdTjQOT4 https://www.youtube.com/watch?v=WkOXzsG_iQA
586 23017    1498059429.0    p4-relatedvideos    https://www.youtube.com/watch?v=l5VmV8FmCgI https://www.youtube.com/watch?v=l5VmV8FmCgI
586 240 1498068933.0    p1-googlesearch metafont    https://tex.stackexchange.com/questions/107489/what-is-metapost-metafont-and-how-can-i-get-started-using-it     noref
```

Dataset Explained

Country code	484	608
Application	23017	815
Timestamp	0.0	1405374769.41
Pattern	p1-googlesearch	p2-googleimage
Query	mortal+kombat	gundam+for+trash
URL	https://es.wikipedia.org/wiki/Mortal_Kombat_(serie)	http://www.ufunk.net/wp-content/uploads/2012/10/makaon-can-sculptures-7.jpg
Additional URL		http://www.ufunk.net/en/artistes/makaon-can-sculptures/

Goal

- Count queries
- Final result should look like:

  facebook 598342

  youtube 23454

  ebay 13984

  …

Helpful commands (Linux/Mac)

- Linux:
  - Run MapReduce:

```
cat dataset |python mapper.py |sort -k1,1 |python reducer.py > mapreduceoutput
```

  - Sort File:

```
sort -rn -t $'\t' -k2 unsorted >sorted
```

Helpful commands (Windows)

- Option1:
  - Use anaconda console (not normal bash or jupyter)
  - Run MapReduce:
    - type dataset | python mapper.py | sort | python reducer.py > mapreduceoutput
  - Sort:
    - Use Excel ☺
- Option2:
  - Use CYGWIN tool to run Mac commands on Windows. https://cygwin.com/

https://tinyurl.com/HUDW-HandleBigData

humera.noor@gmail.com
https://de.linkedin.com/in/humeranoor