Ontožúr kickoff

Tomáš Kramár, Dušan Zeleník

Čo ideme riešiť?

- spolupráca s Azetom
- umiestňovanie reklám
- zvyšovanie klikanosti (CTR, click-through rate)

Impresie

datetime	Časová pečiatka zobrazenia reklamy
oaid	Identifikátor používateľa
age	Vek používateľa
sex	Pohlavie používateľa
region	Geografický región z ktorého používateľ pochádza
town	Mesto, z ktorého používateľ pochádza
zone_id	Číslo plochy, de facto portál (cas.sk, azet.sk, lesk.sk...)

Impresie

campaign_id	Reklamná kampaň (jedna kampaň = viacero banerov)
advertiser_id	Zadávateľ reklamy
banner_position	Pozícia v rámci stránky (1..8)
views_campaign	Koľko krát používateľ videl reklamu z tejto kampane
views_banner	Koľko krát používateľ videl tento banner
time_of_week	hodina:minúta od polnoci pondelka

Impresie

prev_click_banner	Klikol už niekedy používateľ na tento banner?
prev_click_campaign	Klikol už niekedy používateľ na banner tohto zadávateľa?
prev_click_advertiser	Klikol už niekedy používateľ na banner tohto klienta?
page_url	URL na ktorej sa zobrazuje reklama (prekvapivo, môže byť NULL)
img_url	URL na obrázok reklamy
advert_text	Text reklamy
advert_url	Cieľová URL reklamy

Kliky

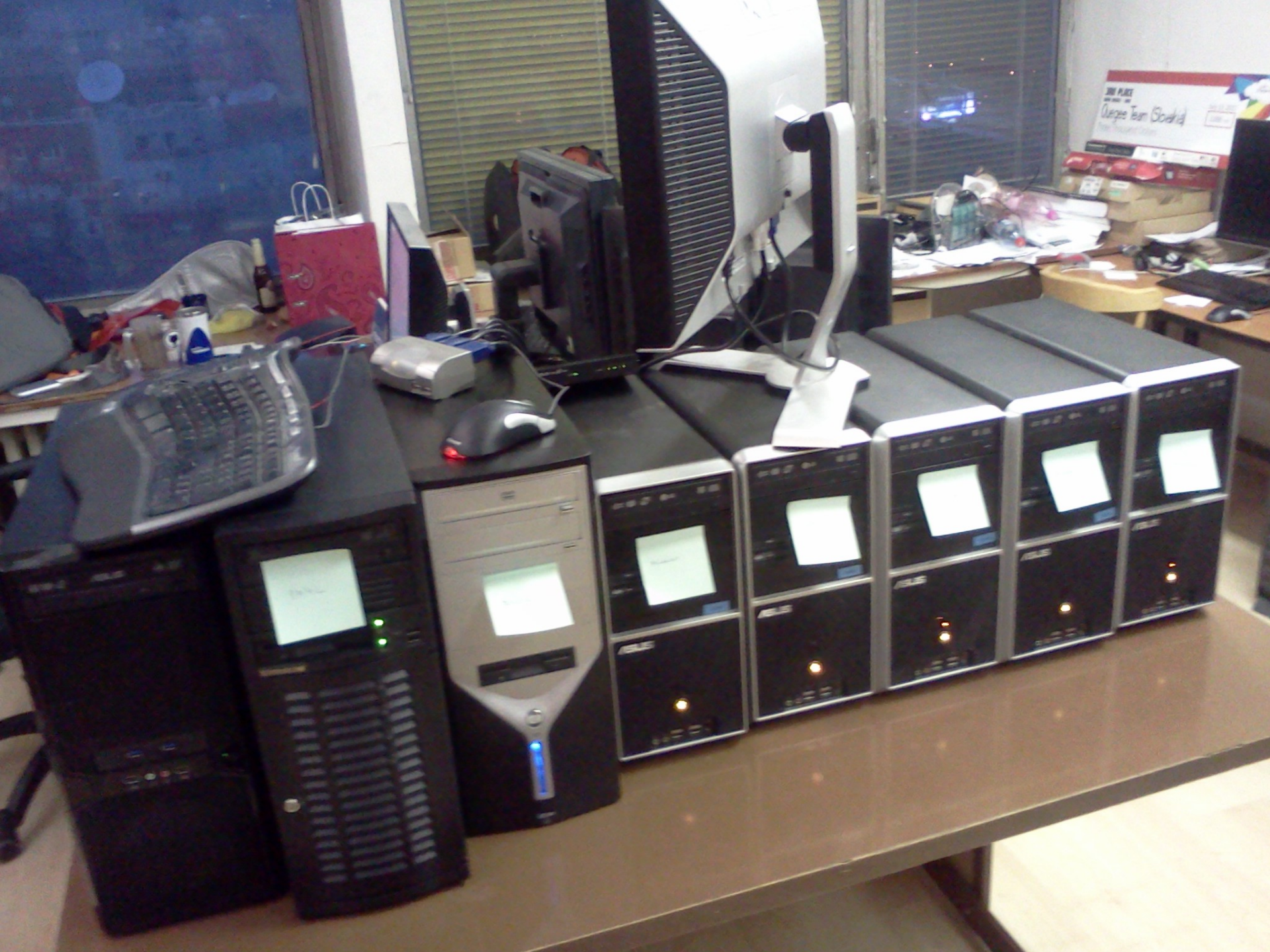
- Identifikátory potrebné pre párovanie s impresiami

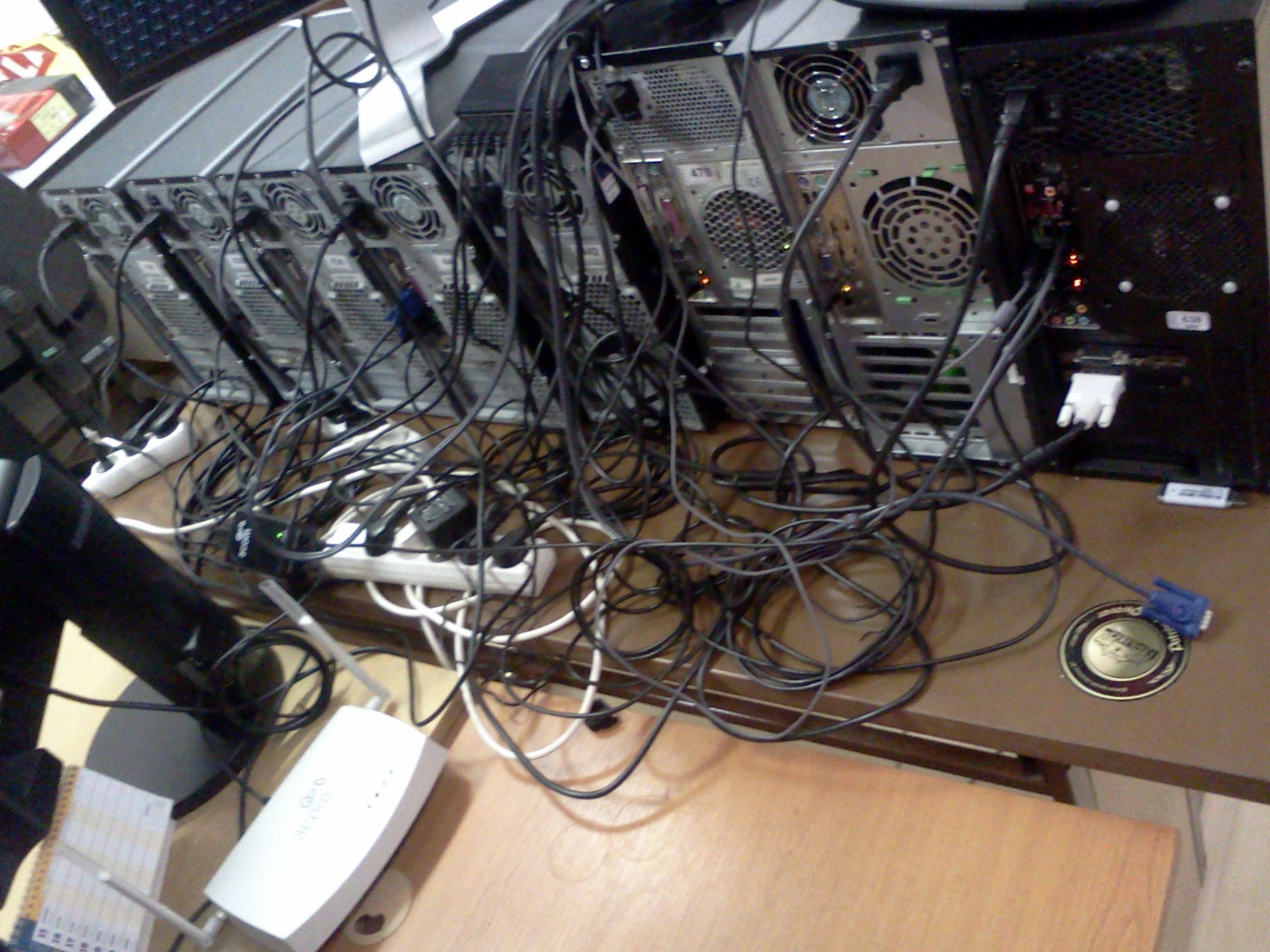
CTR = 0.03%

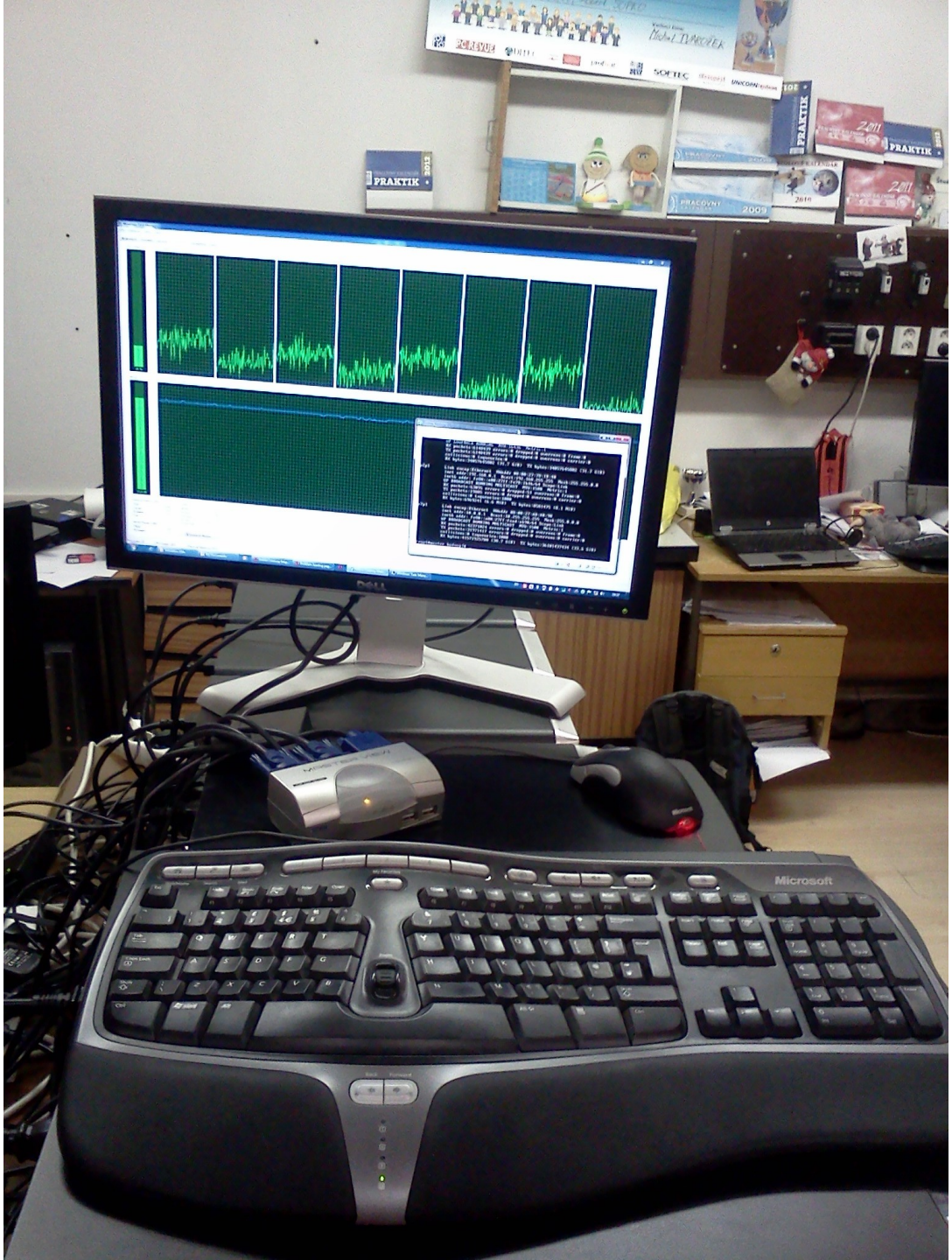
Dáta

- pol hodinové dumpy
- impresie a kliky
- deň =~ 900M bzip2, 11G rozpakované
- zatiaľ 14 dní, 32G zbalených dát a stále pribúdajú

Silné stroje






PRAKTIK
PRAKTIK
PRAKTIK
DELL
Microsoft

Lokálny hadoop

Máte?

Ako na svoj single node cluster?

- stačí naštartovať

  start-all.sh
- kontrolovať joby

  http://localhost:50070
- kontrolovať data

  http://localhost:50030

  alebo aj cez konzolu

  hadoop dfsadmin -report
- zabiť svoj ultimately useless job

  hadoop job -kill <job-id>

HDFS - súborový systém

- prezeranie hdfs

  hadoop fs -ls <args>

  hadoop fs -lsr <args>

- vkladanie a vyberanie dát

  hadoop fs -put <localsrc> <dst>

  hadoop fs -get <src> <localdst>

- mazanie

  hadoop fs -rm <file>

  hadoop fs -rmr <dir>

- ďalšie užitočné...

  hadoop fs -cat <file>

  hadoop fs -tail -f <file>

  hadoop fs -cp <file1> <file2>

Hive

- spustenie hive

  hive
- na čo si dať pozor
  - dáta pre jednu tabuľku musia byť v jednom hdfs adresári
  - ideálne vo viacerých súboroch, ale nie vo vnorených adresároch
  - jeden adresár = jedna tabuľka
- skoro ako SQL, pozri example

https://cwiki.apache.org/confluence/display/Hive/GettingStarted#GettingStarted-ExampleQueries

Ako si vytvoriť hive tabuľku

```
create external table impressions
(id int, happened_at string, oaid string, age int, sex int, region int, town int,
zone_id int, campaign_id int, banner_id int, advertiser_id int, banner_position
int, views_campaign int, views_banner int, time_of_week string,
prev_click_banner int, prev_click_campaign int, prev_click_advertiser int,
page_url string, img_url string, advert_text string, advert_url string)
row format delimited
fields terminated by '\t'
stored as textfile
location '/data/azet/impressions';
```

Mahout

- rada implementácií pre KDD
  - clustering, classification, recommendation, fpgrowth
- priprav dáta
  napr. mahout seqdirectory & mahout seq2sparse
- a pusti napr. kmeans

  mahout kmeans
  -i testdata
  -o output
  -c clusters
  -dm org.apache.mahout.common.distance.CosineDistanceMeasure
  -x 5 -ow -cd 1 -k 25
- naštuduj, potom pomachruješ na ontožúre
  https://cwiki.apache.org/MAHOUT/quickstart.html

Vlastný map reduce job

hadoop jar <my.jar> <args>

- môžete skúsiť napr. grep z príkladov
  hadoop jar *examples*.jar grep inputdir outputdir 'hello'
- ale môžete si spraviť aj vlastné jarko

http://hadoop.apache.org/common/docs/current/mapred_tutorial.html

```
public static class Map
        extends MapReduceBase
        implements Mapper{
  public void map(...){
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                output.collect(word, one);
        }
  }
}
```

```
public static class Reduce
        extends MapReduceBase
        implements Reducer{
  public void reduce(...) {
        int sum = 0;
        while (values.hasNext()) {
                sum += values.next().get();
        }
        output.collect(key, new IntWritable(sum));
  }
}
```

Kde sa dá zapojiť v rámci trénovania modelu

- prehľad dát, štatistika
- analýza korelácií
- vzory
- model

Nenechajte sa obmedzovať Hadoopom, kľudne prineste R, Weku, RapidMiner...

Kde sa dá zapojiť aj mimo trénovania modelu

- Extrakcia odvodených atribútov
  - analýza textov reklám
  - cieľových stránok
  - obrázkov

Choďte a rozmýšlajte

Ale dáta si nechajte pre seba

Ako nastaviť hadoop

- rozbaliť archív
- Upraviť conf/hadoop-env.sh a zmeniť cestu k JAVA_HOME
- Upraviť conf/core-site.xml a pridať

```
<property>
  <name>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
</property>
```

- Je dôležité, aby adresár, ktorý ste uviedli vyššie existoval, a používateľ, pod ktorým spúštate hadoop mal do neho plný prístup
- Spustiť príkaz:
  - bin/hadoop namenode -format

Ako nastaviť hadoop

- Nastaviť passwordless ssh prístup na localhost
  - potrebujete sshd (ssh server), nainštalujte balík a spustite démonka
  - ssh-keygen (ak ešte žiadny kľúč nemáte)
  - ssh-copy-id localhost (zadajte heslo)
  - overte, že ssh localhost nepýta heslo
- Spustiť klaster
  - bin/start-all.sh
- Namenode a tasktracker
  - http://localhost:50070
  - http://localhost:50030
  - tu by ste mali vidieť jeden Live node, chvíľu trvá kým to nabehne, a možno tam budete chvíľu vidieť 0

Ako nastaviť hadoop

- Pridať hadoop na PATH
  - Upraviť súbor ~/.bashrc
  - PATH=$PATH:/home/tomas/hadoop-1.0.1/bin
  - cestu upravte podľa seba, bacha na medzery okolo =, nemozu tam byt
  - Zmena sa prejaví po reštarte terminálu, alebo spustení príkazu
    - source ~/.bashrc
  - Odteraz môžete spustiť hadoop z akéhokoľvek adresára
- Stopnutie klastra
  - bin/stop-all.sh

Ako nastaviť hive

- Stiahnuť hive
  - http://sk.freebsd.org/pub/apache/dist/hive/hive-0.8.1/hive-0.8.1.tar.gz
- Rozbaliť
- Pokiaľ máte hadoop v PATH (predošlý slide), tak netreba robiť nič, a pôjde to
- Pridajte si na PATH aj hive
  - PATH=$PATH:/home/tomas/hadoop-1.0.1/bin:/home/tomas/hive-0.8.1
  - Odteraz môžete spustiť hive z akéhokoľvek adresára

Ako nastaviť mahout

- Stiahnite mahout
  - http://sk.freebsd.org/pub/apache/dist/mahout/0.6/mahout-distribution-0.6.tar.gz
- Rozbaľte
- Upravte bin/mahout a niekde na vrch doplňte JAVA_HOME=, rovnako ako pre hadoop
  - alternatíva je pridať toto do .bashrc
    - export JAVA_HOME=/usr/lib/jvm/java-6-openjdk
- Pridajte si na PATH aj mahout
  - PATH=$PATH:/home/tomas/hadoop-1.0.1/bin:/home/tomas/hive-0.8.1:/home/tomas/mahout-distribution
  - Odteraz môžete spustiť mahout z akéhokoľvek