Building a Big Data DWH

Data Warehousing on Hadoop

Friso van Vollenhoven   *@fzk*

CTO   *frisovanvollenhoven@godatadriven.com*

GoDataDriven

PROUDLY PART OF THE XEBIA GROUP

"In computing, a data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis."

-- Wikipedia

Article   Talk   Read   Edit   View history   Search

Business reporting

From Wikipedia, the free encyclopedia

Business Reporting or **Enterprise Reporting** is a fundamental part of the larger movement towards improved business intelligence and knowledge management. Often implementation involves extract, transform, and load (ETL) procedures in coordination with a data warehouse and then using one or more reporting tools. While reports can be distributed in print form or via email, they are typically accessed via a corporate intranet.

With the dramatic expansion of information technology, and the desire for increased competitiveness in corporations, there has been an increase in the use of computing power to produce unified reports which join different views of the enterprise in one place.[1] This reporting process involves querying data sources with different logical models to produce a human readable report—for example, a computer user has to query the Human Resources databases and the Capital Improvements databases to show how efficiently space is being used across an entire corporation.

References   [edit]

1. ^ Moeller, Robert (2007). COSO Enterprise Risk Management: Understanding the New Integrated

Article   Talk   Read   Edit   View history   Search

Data analysis

From Wikipedia, the free encyclopedia

> This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(December 2008)*

Analysis of data (datalytics) is a process of inspecting, cleaning, transforming, and modeling **data** with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facts and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive
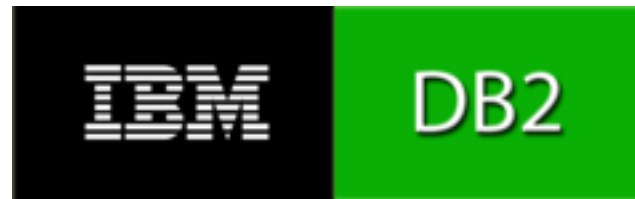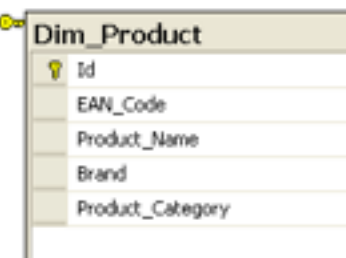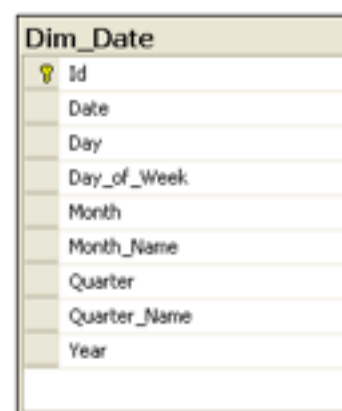
Dim_Date

🔑	Id
	Date
	Day
	Day_of_Week
	Month
	Month_Name
	Quarter
	Quarter_Name
	Year

Fact_Sales

	Date_Id
	Store_Id
	Product_Id
	Units_Sold

Dim_Store

🔑	Id
	Store_Number
	State_Province
	Country

Dim_Product

🔑	Id
	EAN_Code
	Product_Name
	Brand
	Product_Category

How to:

- Add a column to the facts table?
- Change the granularity of dates from day to hour?
- Add a dimension based on some aggregation of facts?

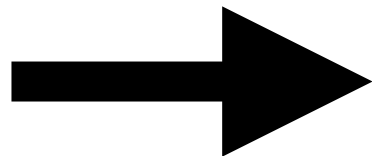Schema's are designed
with questions in mind.

Changing it requires to
redo the ETL.
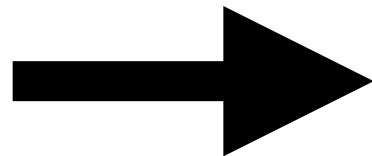
Schema's are designed with questions in mind. ➡ Push things to the facts level.

Changing it requires to redo the ETL. ➡ Keep all source data available all times.

MORE
DATA,
PLEASE

And now?

- MPP databases?
- Faster / better / more SAN?
- (RAC?)

metadata + query engine
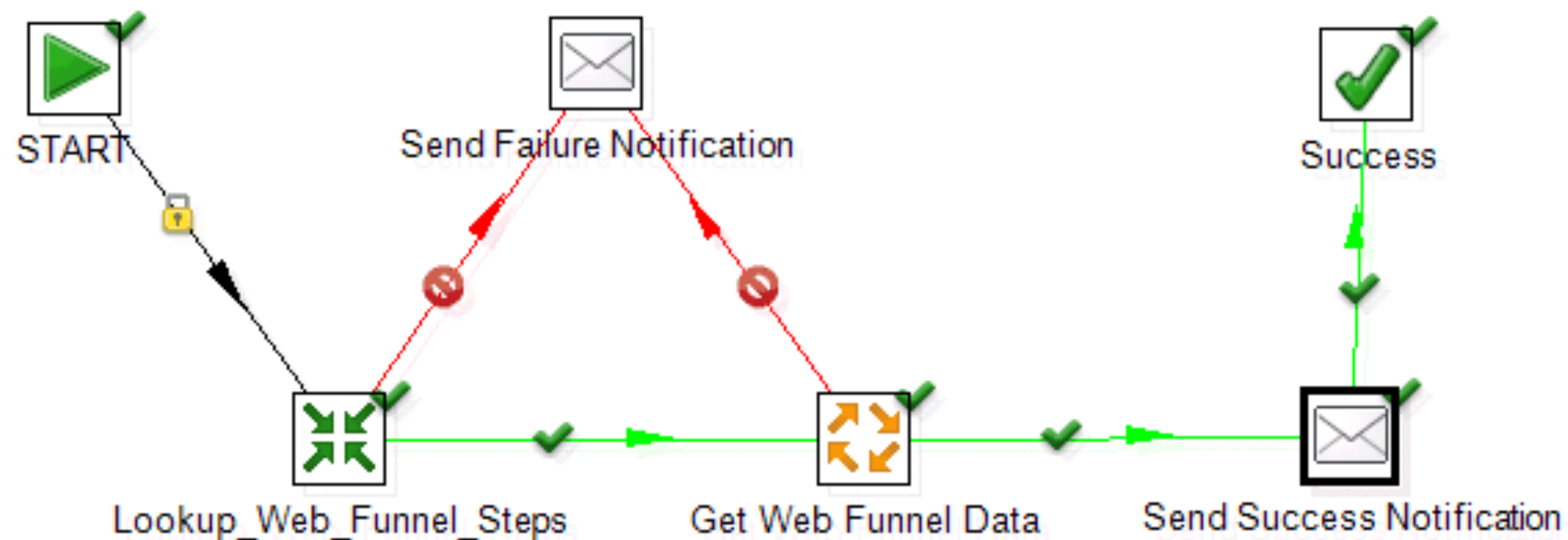
distributed processing

distributed storage

create-table.sql

```sql
1   CREATE EXTERNAL TABLE source_data (
2     host STRING,
3     dt STRING,
4     request STRING,
5     status_code STRING,
6     response_length STRING,
7     referrer STRING,
8     user_agent STRING)
9   ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
10  WITH SERDEPROPERTIES (
11    'input.regex' = '(\\d{1,3}.\\d{1,3}.\\d{1,3}.\\d{1,3}) - - \\[(\\d{1,2}/[a
12    'output.format.string' = '%1$s %2$s %3$s %4$s %5$s %6$s %7$s')
13  STORED AS TEXTFILE
14  LOCATION '/source_data/myfiles/';
15
16
17  SELECT host, count(*) AS request_count FROM source_data
18  WHERE dt > '2013-01-01'
19  GROUP BY host;
20
```

EXTRACT
TRANSFORM
LOAD

START

Send Failure Notification

Success

Lookup_Web_Funnel_Steps

Get Web Funnel Data

Send Success Notification

Shell

```bash
#!/bin/bash
set -ex

echo \
'open ftp.example.com
user username password
cd my-files
find /' > /tmp/ftp-list-command.txt

lftp -f /tmp/ftp-list-command.txt | grep "\.gz" > /tmp/full-names-on-ftp.txt

hadoop fs -ls -R /staging/my-files/ > /tmp/full-names-on-hadoop.txt
hadoop fs -ls -R /sources/my-files/ >> /tmp/full-names-on-hadoop.txt

cat /tmp/full-names-on-hadoop.txt | grep "\.gz" | awk '{print $8}' | awk -F '/' '{print $NF}' > /tmp/basenames-on-hadoop.txt

cat /tmp/full-names-on-ftp.txt | fgrep -v -f /tmp/basenames-on-hadoop.txt > /tmp/missing-files-on-hadoop.txt

echo \
'open ftp.example.com
user username password
' > /tmp/ftp-get-files-command.txt
awk '{print "get -c " $0}' /tmp/missing-files-on-hadoop.txt >> /tmp/ftp-get-files-command.txt


cd /local-data/staging/my-files/
lftp -vf /tmp/ftp-get-files-command.txt
cd -

while read f
do
  filename=$(echo $f | awk -F '/' '{print $NF}')
  year=${filename:0:4}
  month=${filename:4:2}
  day=${filename:6:2}
  site=$(echo $f | awk -F '-' '{print $NF}' | sed s/\.gz//g)

  hadoop fs -mkdir "/staging/my-files/$year/$month/$day/$site"
  hadoop fs -put "/local-data/staging/my-files/$filename" "/staging/my-files/$year/$month/$day/$site/$filename.upload"
  hadoop fs -mv "/staging/my-files/$year/$month/$day/$site/$filename.upload" "/staging/my-files/$year/$month/$day/$site/$filename"
done < /tmp/missing-files-on-hadoop.txt
```

```python
client = FTPClient(FTP_HOST, FTP_USER, FTP_PASSWORD)
remote_files = ftp_file_list(client, FTP_ROOT_DIR)

fs = hfs.FileSystem.get(hconf.Configuration())
hadoop_files = hdfs_file_list(fs, HDFS_ROOT_DIR_STAGING)
hadoop_files.extend(hdfs_file_list(fs, HDFS_ROOT_DIR_SOURCES))

missing_files = filter(lambda f: f not in hadoop_files, remote_files)

filename_pattern = re.compile(r"(\d{4})(\d{2})(\d{2})\d{4}-\d{12}-(\d{6})\.gz")

for missing_file in missingfiles:
    temp_file = File(FTP_LOCAL_DOWNLOAD_PATH, missing_file)
    client.download(missing_file, temp_file)

    year, month, day, site = filename_pattern.match(missing_file[1].name).groups()
    hdfs_dir = hfs.Path('%s/%s/%s/%s/%s' % (HDFS_UPLOAD_BASE_PATH, year, month, day, site))

    hdfs_dir = hfs.Path('%s/%s/%s/%s/%s' % tuple(filename_pattern.match(missing_file).groups()))

    if (not fs.exists(hdfs_dir)):
        fs.mkdirs(hdfs_dir)

    fs.copyFromLocalFile(hfs.Path(temp_file.path), hfs.Path(remotedir, remotefilename + '.upload'))
    fs.rename(hfs.Path(remotedir, remotefilename + '.upload'), hfs.Path(remotedir, remotefilename)

    os.unlink(localfile.path)


fs.close()
client.disconnect(False)
```

- No JVM startup overhead for Hadoop API usage
- Relatively concise syntax (Python)
- Mix Python standard library with any Java libs

NAME

    crontab -- maintain crontab files for individual users (V3)

SYNOPSIS

    crontab [**-u** <u>user</u>] <u>file</u>
    crontab [**-u** <u>user</u>] { **-l** | **-r** | **-e** }

DESCRIPTION

    The **crontab** utility is the program used to install, deinstall or list the
    tables used to drive the cron(8) daemon in Vixie Cron. Each user can
    have their own crontab, and they are not intended to be edited directly.

    (Darwin note: Although cron(8) and crontab(5) are officially supported
    under Darwin, their functionality has been absorbed into launchd(8),
    which provides a more flexible way of automatically executing commands.
    See launchctl(1) for more information.)

    If the <u>/usr/lib/cron/cron.allow</u> file exists, then you must be listed
    therein in order to be allowed to use this command. If the
    <u>/usr/lib/cron/cron.allow</u> file does not exist but the

:

- Flexible scheduling with dependencies
- Saves output
- E-mails on errors
- Scales to multiple nodes
- REST API
- Status monitor
- Integrates with version control

successfull job #380

successfull job running #291

unstable job #552

unstable job running #237

aborted or disabled job #307

aborted Job running #256

failed job #593

failed job running #308

queued job number 1 #464

queued job number 2 #462

queued job number 3 #394

job with failed api call

Deployment

`git push jenkins master`

- Scheduling
- Simple deployment of ETL code
- Scalable
- Developer friendly

Visitors per day of the week in 2012
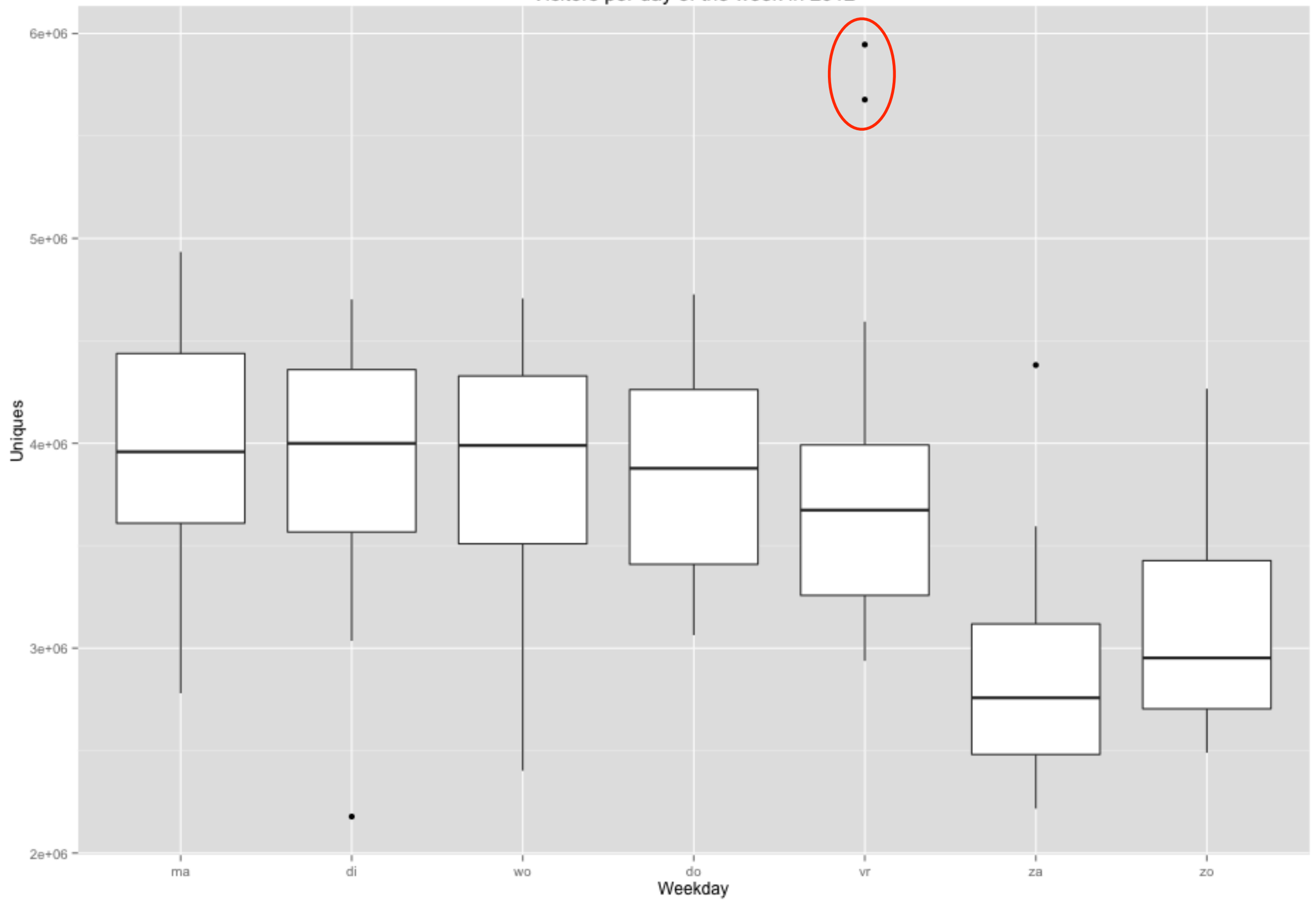
Visitors per day of the week for each month in 2012
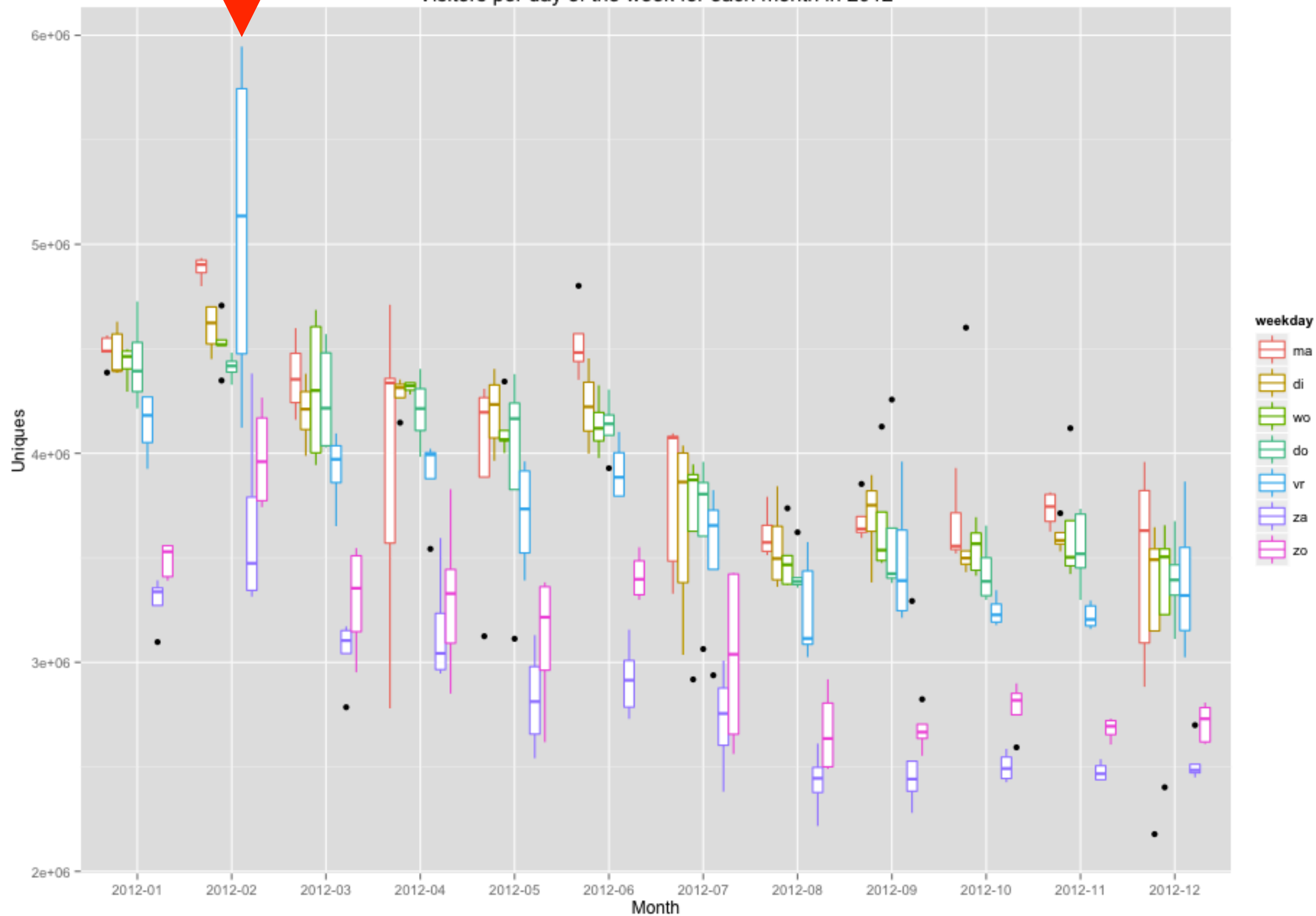
Visitors per day for fridays each month in 2012
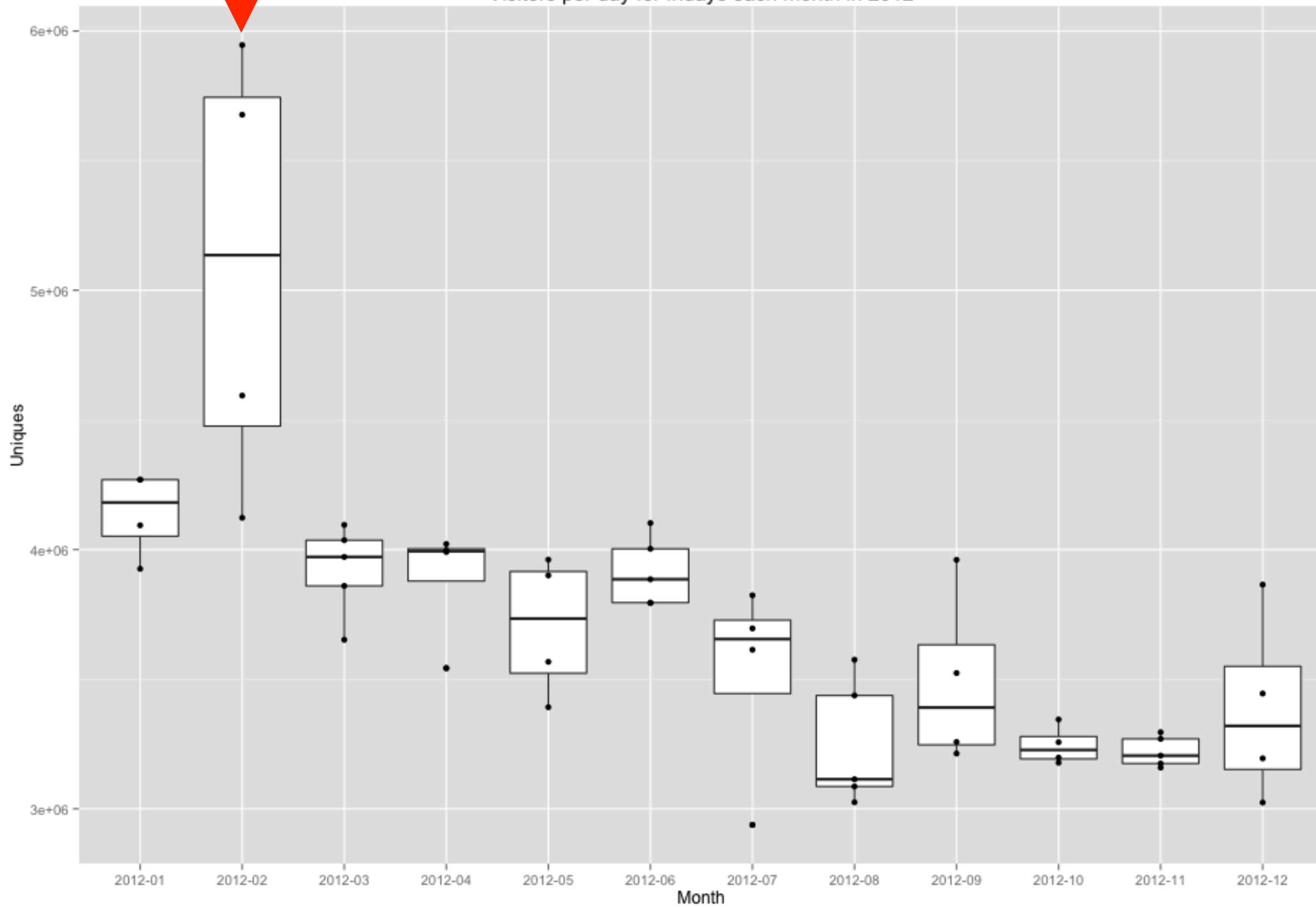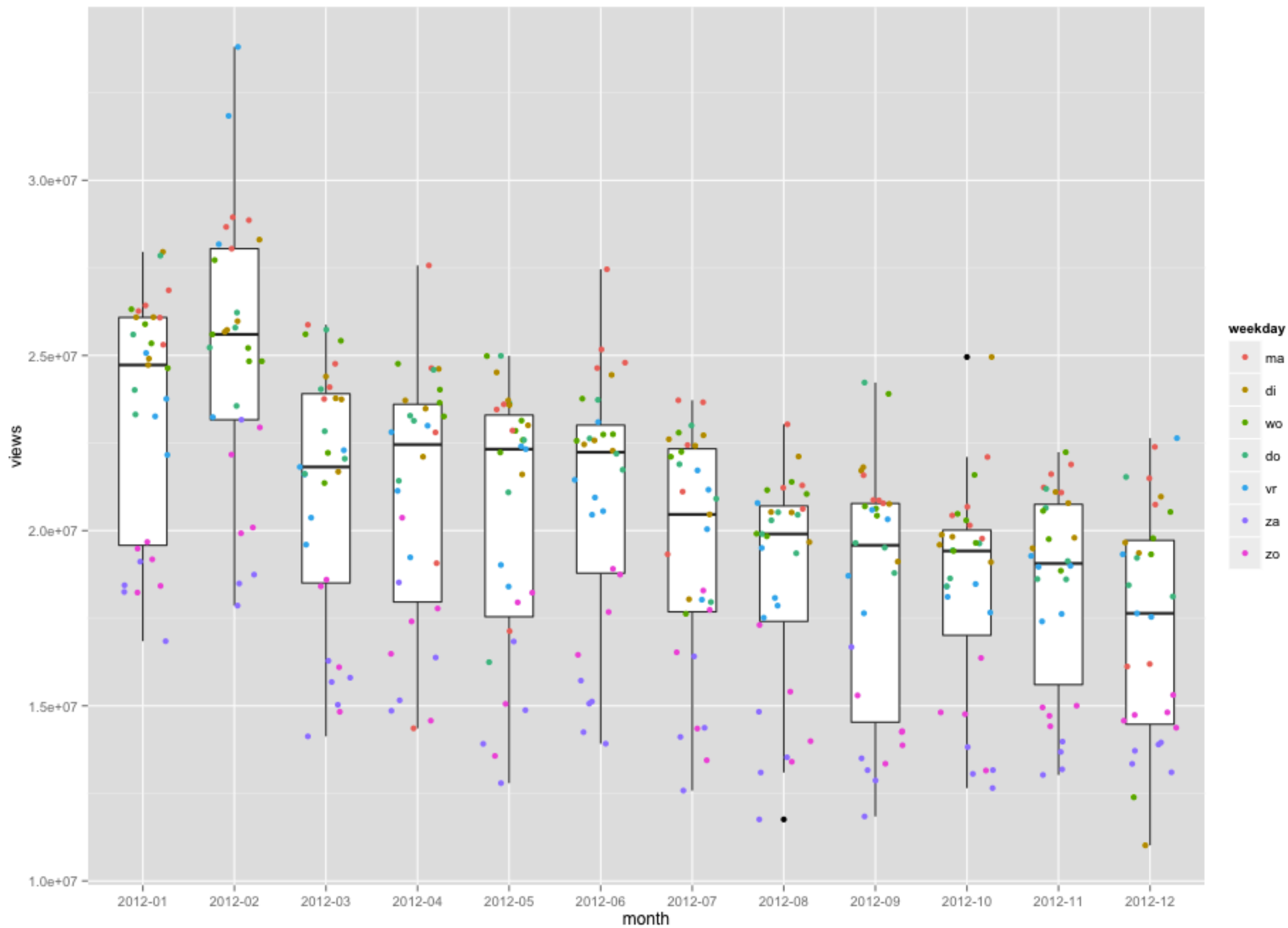
PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013   02/27/13   27/02/2013   27/02/13

20130227   2013.02.27   27.02.13   27-02-13

27.2.13   2013.II.27. $27\frac{1}{2}$-13   2013.158904109

MMXIII-II-XXVII   MMXIII $\frac{LVII}{CCCLXV}$ 1330300800

$((3+3)\times(111+1)-1)\times3/3-1/3^3$ 2013

10/11011/1101   02/27/20/13 $\overset{2\ 3\ 1\ 4}{\underset{5\ 67\ 8}{0123 7}}$

Hissss

`'februari-22 2013'`

Does Amazon S3 fail sometimes? [closed]

2

▲ ▼ ☆ 1

We just added an autoupdater in our software and got some bug report saying that the autoupdate wouldn't complete properly because the downloaded file's sha1 checksum wasn't matching. We're hosted on Amazon S3...

That's either something wrong with my code or something wrong with S3.

I reread my code for suspicious stuff and wrote a simple script downloading and checking the checksum of the downloaded file, and indeed got a few errors once in while (1 out of 40 yesterday). Today it seems okay.

Did you experience that kind of problem? Is there some kind of workaround ?

extra info: test were ran in Japan.

download   amazon-s3

share | improve this question

edited May 3 '12 at 0:18
Chilledrat
1,598 ●2 ●9 ●17

asked Aug 28 '08 at 1:25
poulejapon
4,174 ●1 ●18 ●39
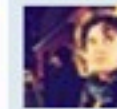
closed as not constructive by the Tin Man, vstm, EJP, 0x7fffffff, xdazz
Oct 6 '12 at 6:40

As it currently stands, this question is not a good fit for our Q&A format. We expect answers to be supported by facts, references, or specific expertise, but this question will likely solicit debate, arguments, polling, or extended discussion. If you feel that this question can be improved and possibly reopened, see the FAQ for guidance.

A: Yes, sometimes as often as 1 in every 10K calls. Or about once a week at 3K files / day.

```
friso@fvv:/tmp/data $ ls
total 8
-rw-r--r--  1 friso  wheel     0B 29 mei 10:57 i-am-void.gz
-rw-r--r--  1 friso  wheel    20B 29 mei 10:58 i-am.gz
friso@fvv:/tmp/data $
```

```
friso@fvv:/tmp/data $ ls
total 8
-rw-r--r--   1 friso   wheel      0B 29 mei 10:57 i-am-void.gz
-rw-r--r--   1 friso   wheel     20B 29 mei 10:58 i-am.gz
friso@fvv:/tmp/data $ gunzip -c i-am.gz
friso@fvv:/tmp/data $
friso@fvv:/tmp/data $
friso@fvv:/tmp/data $ gunzip -c i-am-void.gz

gzip: i-am-void.gz: unexpected end of file
friso@fvv:/tmp/data $
```

þ

Þ

character	Þ	þ
Unicode name	LATIN CAPITAL LETTER THORN	LATIN SMALL LETTER THORN
Unicode	00DE	00FE
Character entity reference	Þ	þ
Windows-1252, ISO-8859-1, ISO-8859-15	DE	FE
LaTeX	\TH	\th

TSV ==
thorn separated values?

þ == 0xFE

or -2, in Hive

```
CREATE TABLE browsers (
    browser_id STRING,
    browser STRING
) ROW FORMAT DELIMITED FIELDS TERMINATED BY '-2';
```
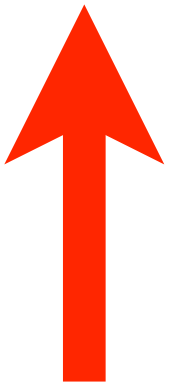
Article   Talk   Read   Edit   View history   Search

Data governance

From Wikipedia, the free encyclopedia

Data governance is an emerging discipline with an evolving definition. The discipline embodies a convergence of data quality, data management, data policies, business process management, and risk management surrounding the handling of data in an organization. Through data governance, organizations are looking to exercise positive control over the processes and methods used by their data stewards and data custodians to handle data.

Data governance is a set of processes that ensures that important data assets are formally managed throughout the enterprise. Data governance ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality. It is about putting people in charge of fixing and preventing issues with data so that the enterprise can become more

Part of a series on

Governance

| **Models** |
| Collaborative · Good · Multistakeholder · Open-source · Private |

| **By level** |
| Local · Global |

| **By field** |
| Climate · Clinical · Corporate · **Data** · Earth system · Ecclesiastical · Environmental · Higher education · Information · Network · Ocean · Political party · Project · Self · Service-oriented architecture · Soil · Technology · Transnational · Website |

| **Measures** |

Article   Talk   Read   Edit   View history   Search

Master data management

From Wikipedia, the free encyclopedia

> This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(April 2012)*

In computing, **Master Data Management (MDM)** comprises a set of processes, governance, policies, standards and tools that consistently defines and manages the master data (i.e. non-transactional data entities) of an organization (which may include reference data).

An MDM tool can be used to support Master Data Management by removing duplicates, standardizing data (Mass Maintaining), incorporating rules to eliminate incorrect data from entering the system in order to create an authoritative source of master data. Master data are the products, accounts and parties for which the business transactions are completed. The root cause problem stems from business unit and product line segmentation, in which the same customer will be serviced by different product lines, with redundant data being entered about the customer (aka party in the role of customer) and account in order to process the transaction. The redundancy of party and account data is compounded in the front to back office life cycle, where the authoritative single source for the party account and product data is needed but is often

Article   Talk   Read   Edit   View history   Search

Meta-data management

From Wikipedia, the free encyclopedia

> This article **may be too technical for most readers to understand**. Please help improve this article to make it understandable to non-experts, without removing the technical details. The talk page may contain suggestions.
>
> *(January 2013)*

Meta-data management (also known as metadata management, without the hyphen) involves storing information about other information. With different types of media being used, references to the location of the data can allow management of diverse repositories.

Metadata management can be defined as the end-to-end process and governance framework for creating, controlling, enhancing, attributing, defining and managing a metadata schema, model or other structured aggregation system, either independently or within a repository and the associated supporting processes (often to enable the management of content).

URLs, images, video etc. may be referenced from a triples table of object, attribute and value.

With specific knowledge domains, the boundaries of the metadata for each must be managed, since a general ontology is not useful to experts in one field whose language is knowledge-

Data custodian

From Wikipedia, the free encyclopedia

> This article includes a list of references, but **its sources remain unclear because it has insufficient inline citations**. Please help to improve this article by introducing more precise citations. *(August 2010)*

In Data Governance groups, responsibilities for data management are increasingly divided between the business process owners and information technology (IT) departments. Two functional titles commonly used for these roles are Data Steward and Data Custodian.

Data Stewards are commonly responsible for data content, context, and associated business rules. Data Custodians are responsible for the safe custody, transport, storage of the data and implementation of business rules.[1][2] Simply put, Data Stewards are responsible for what is stored in a data field, while Data Custodians are responsible for the technical environment and database structure. Common job titles for data custodians are Database Administrator (DBA), Data Modeler, and ETL Developer.

Contents [show]

Article | Talk   Read | Edit | View history   Search   🔍

Data steward

From Wikipedia, the free encyclopedia

In metadata, a **data steward** is a person that is responsible for maintaining a data element in a metadata registry. A data steward may share some responsibilities with a data custodian.

Data stewardship roles are common when organizations are attempting to exchange data precisely and consistently between computer systems and reuse data-related resources. Master data management often makes references to the need for data stewardship for its implementation to succeed.
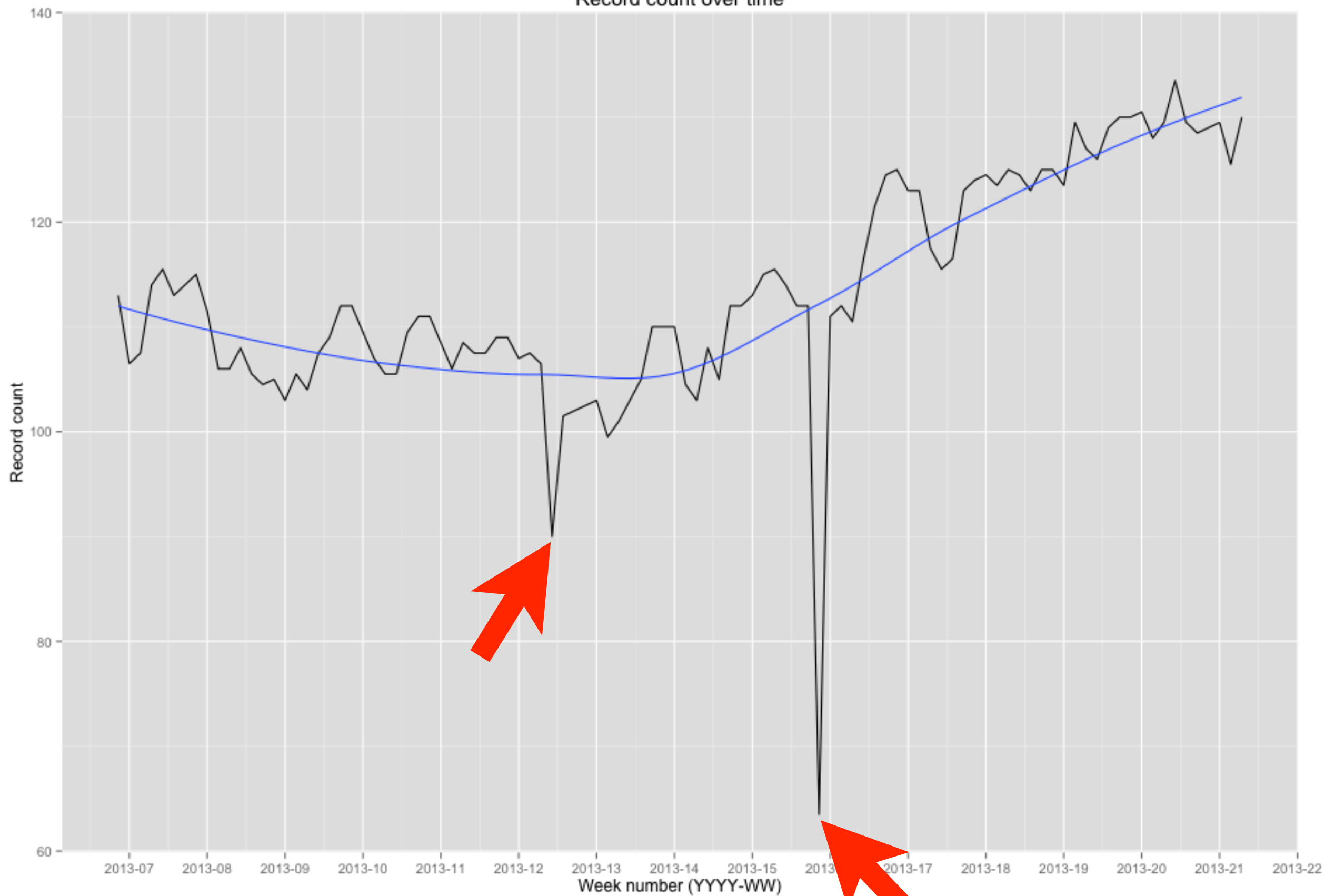
Contents [show]

Data Steward Responsibilities   [edit]

A data steward ensures that each assigned data element:

1. Has clear and unambiguous data element definition.
2. Does not conflict with other data elements in the metadata registry (removes duplicates, overlap etc.)
3. Has clear enumerated value definitions if it is of type Code.

Record count over time

- The format will change
- Faulty deliveries will occur
- Your parser will break
- Records will be mistakingly produced (over-logging)
- Other people test in production too (and you get the data from it)
- Etc., etc.

- Simple deployment of ETL code
- Scheduling
- Scalable
- Independent jobs
- Fixable data store
- Incremental where possible
- Metrics

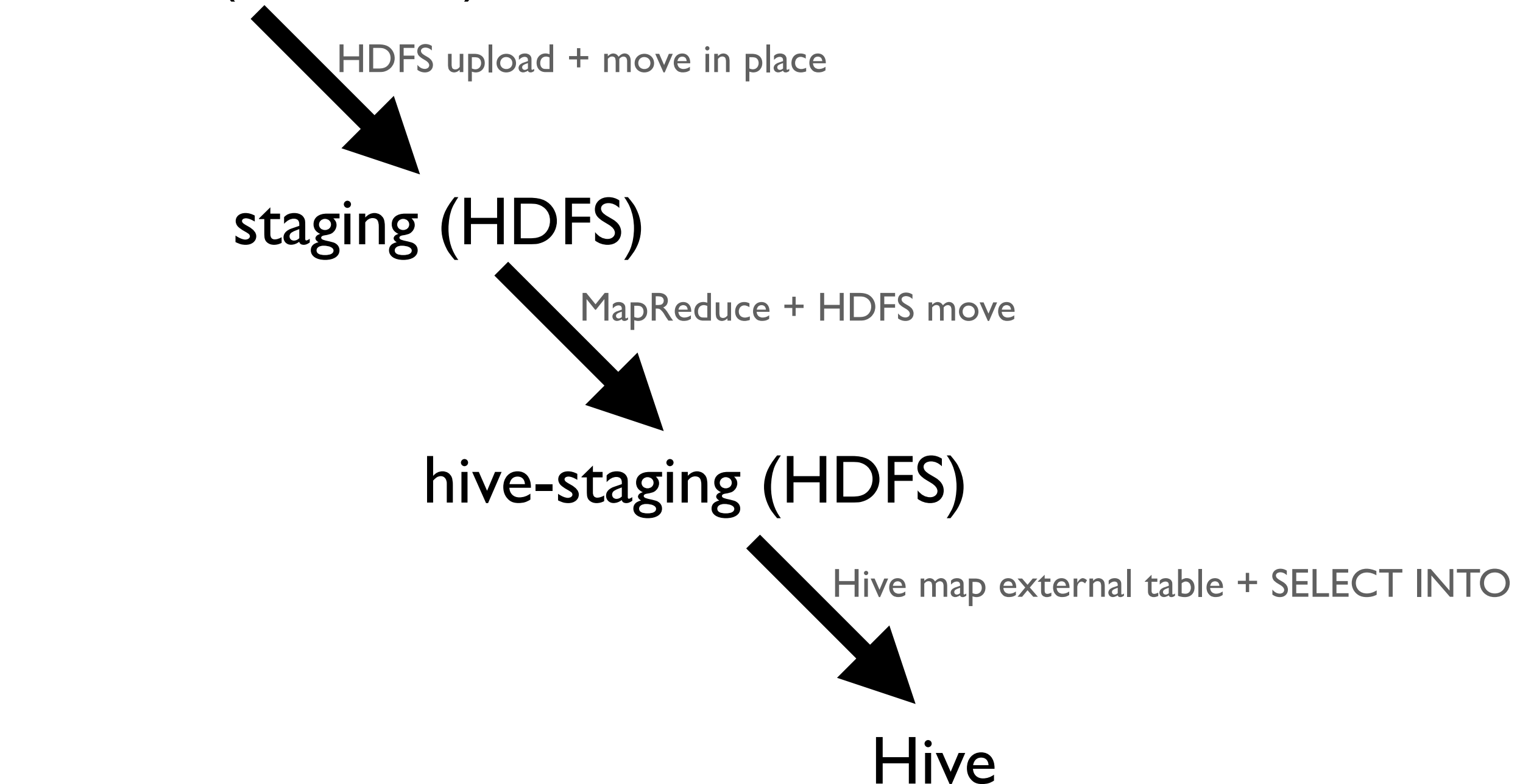# Independent jobs

source (external)

→ HDFS upload + move in place

staging (HDFS)

→ MapReduce + HDFS move

hive-staging (HDFS)

→ Hive map external table + SELECT INTO

Hive

Out of order jobs

- At any point, you don't really know what 'made it' to Hive
- Will happen anyway, because some days the data delivery is going to be three hours late
- Or you get half in the morning and the other half later in the day
- It really depends on what you do with the data
- This is where metrics + fixable data store help...

Fixable data store

- Using Hive partitions
- Jobs that move data from staging create partitions
  - When new data / insight about the data arrives, drop the partition and re-insert
  - Be careful to reset any metrics in this case
- Basically: instead of trying to make everything transactional, repair afterwards
- Use metrics to determine whether data is fit for purpose

Metrics

Record count over time
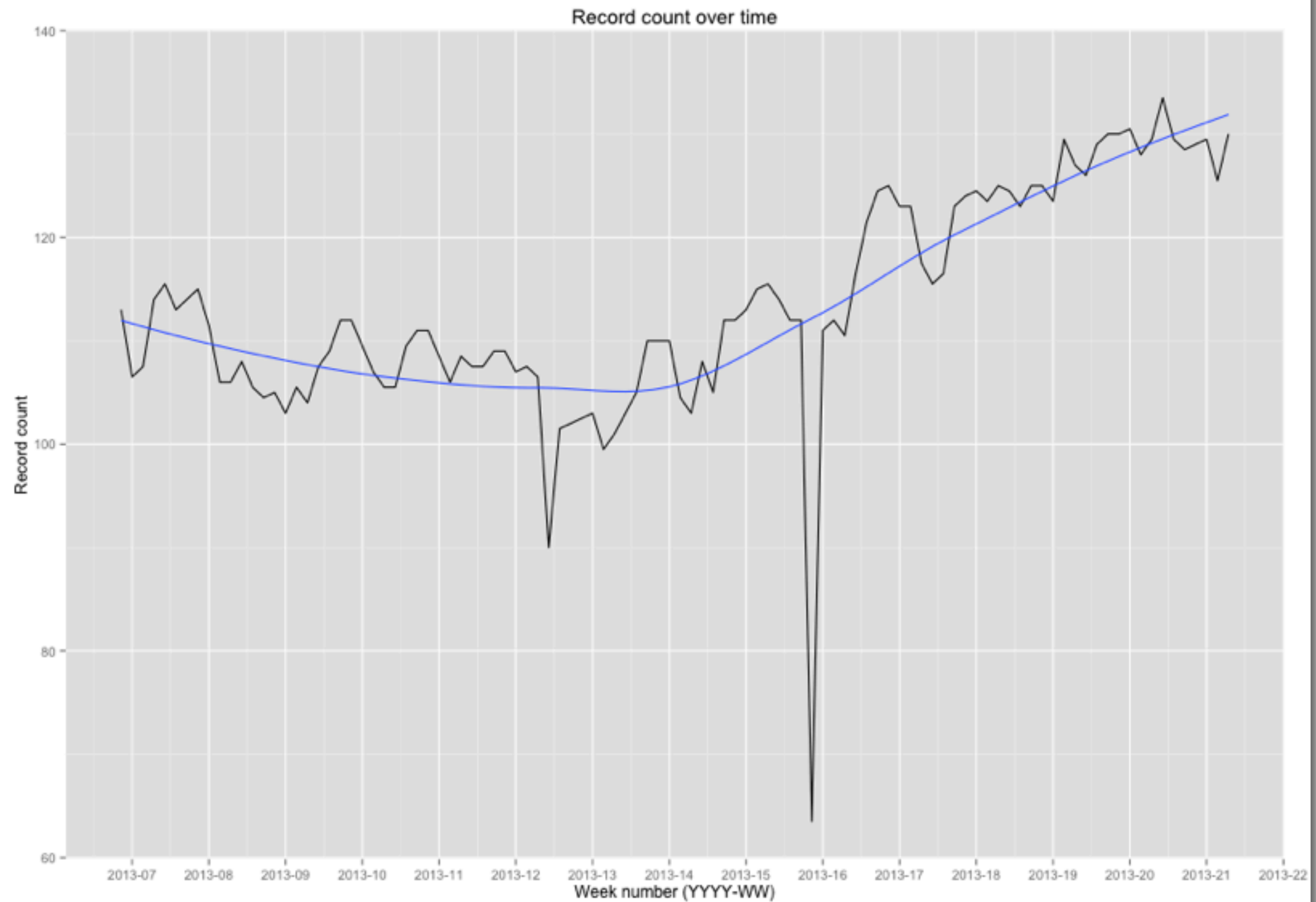
Metrics service

- Job ran, so may units processed, took so much time
  - e.g. 10GB imported, took 1 hr
  - e.g. 60M records transformed, took 10 minutes
- Dropped partition
- Inserted X records into partition
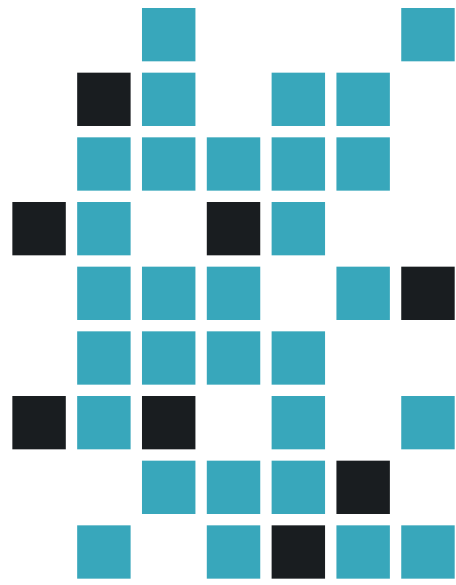
GoDataDriven

We're hiring / Questions? / Thank you!

Friso van Vollenhoven   *@fzk*
CTO   *frisovanvollenhoven@godatadriven.com*