

# Dirty Data

It's a mess. It's **your** problem.

*Friso van Vollenhoven*

*@fzk*

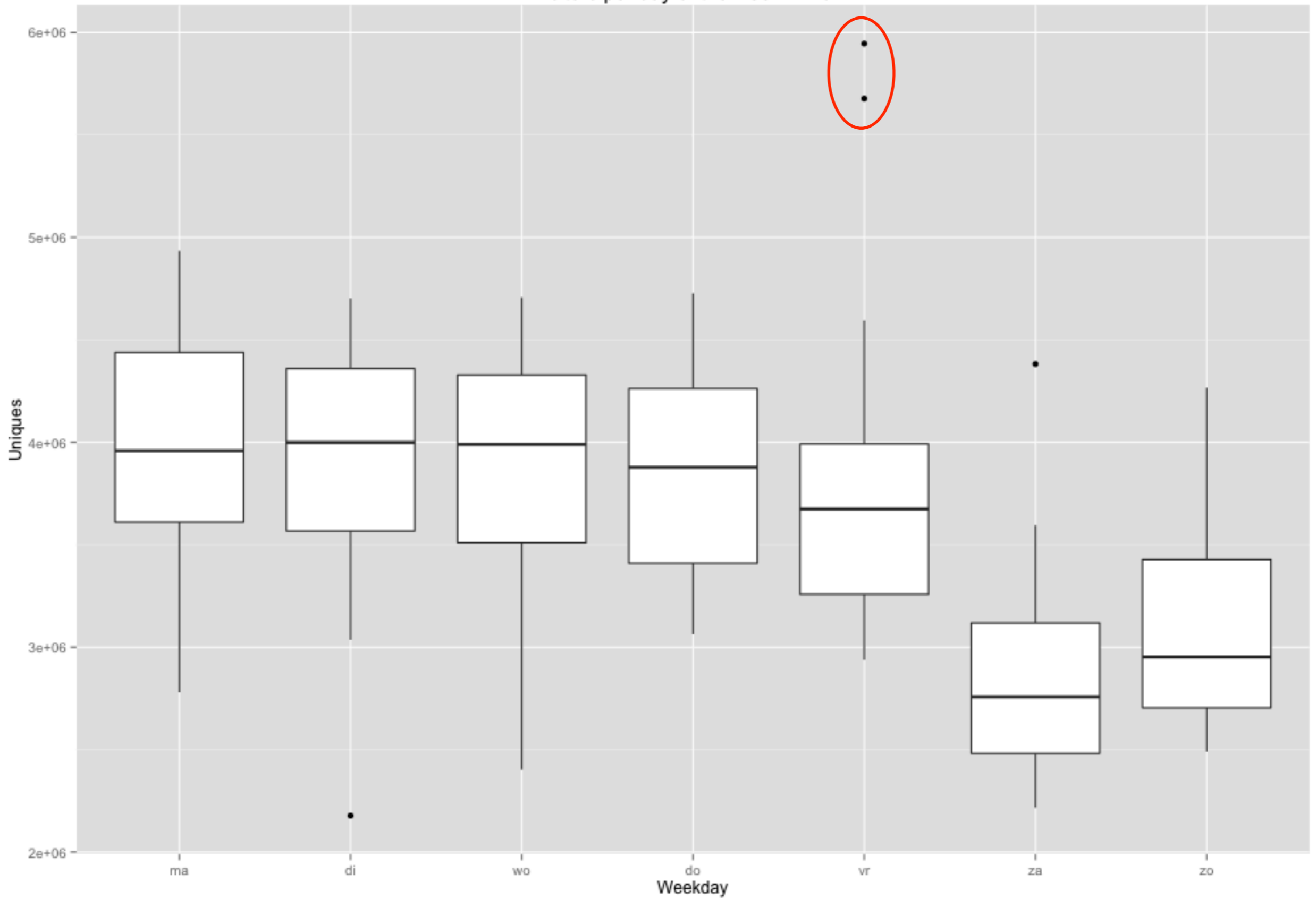
*frisovanvollenhoven@godatadriven.com*



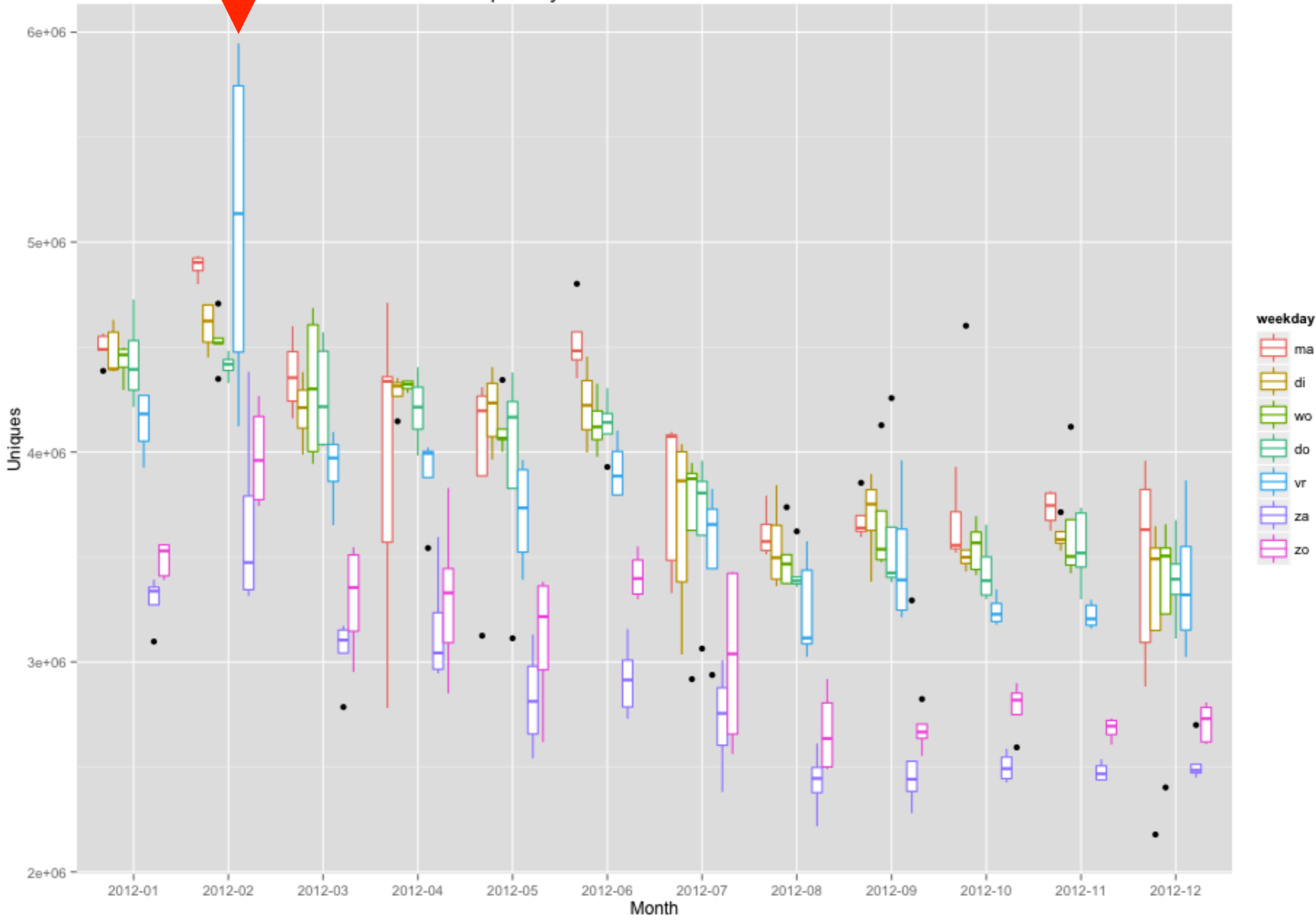
GoDataDriven

PROUDLY PART OF THE XEBIA GROUP

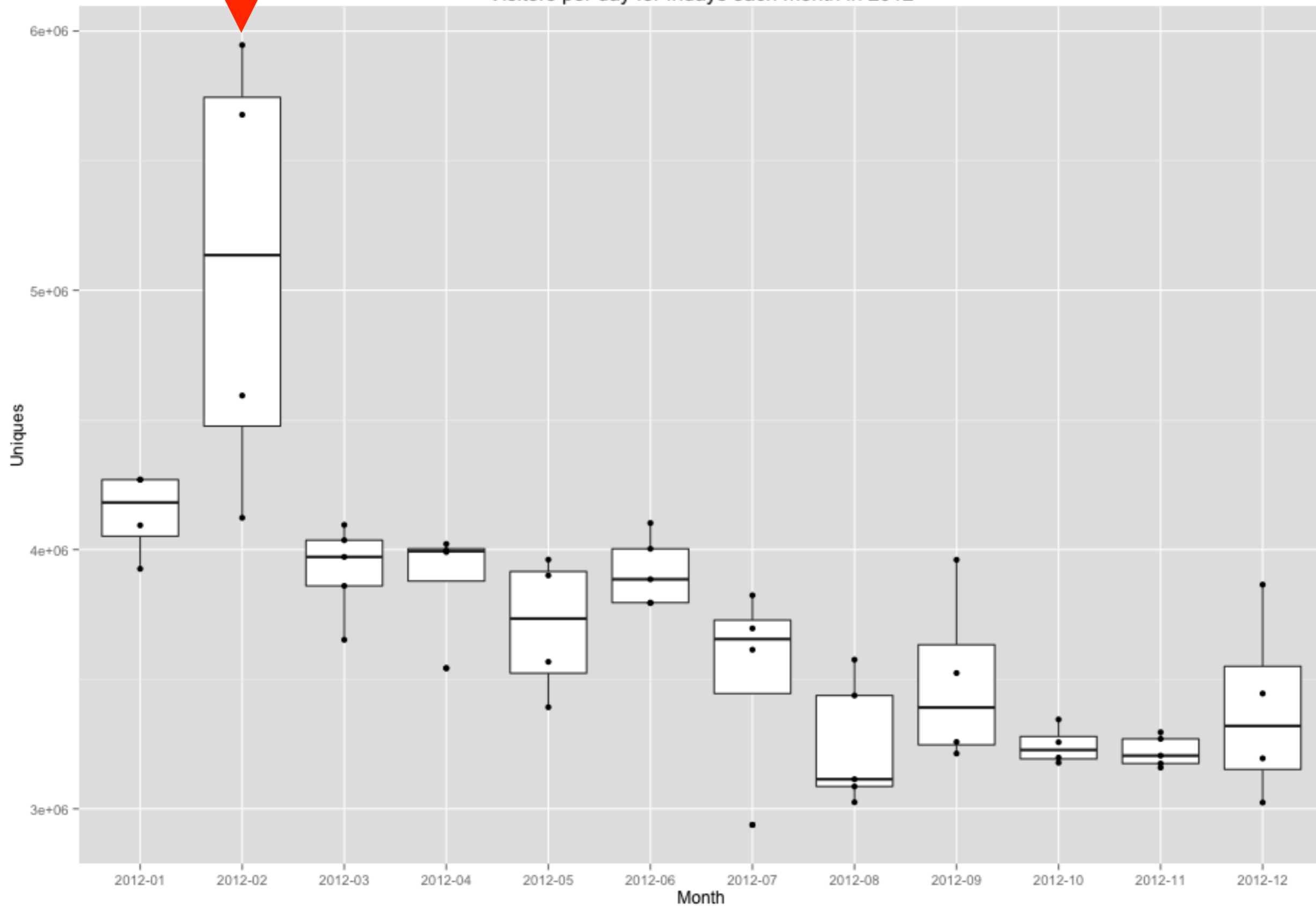
Visitors per day of the week in 2012



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  $\frac{\text{LVII}}{\text{CCCLXV}}$  1330300800

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

10/11011/1101 02/27/20/13  $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 \end{matrix}$

'februari-22 2013'



## Does Amazon S3 fail sometimes? [closed]



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](#)

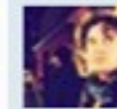
[share](#) | [improve this question](#)

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 $ █
```



**p**

Þ



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

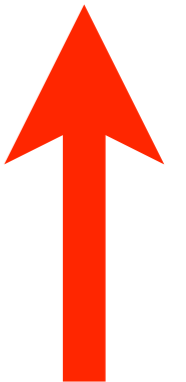
**TSV ==**

**thorn separated values?**

**p == == 0xFFE**

# or -2, in Hive

```
CREATE TABLE browsers (  
    browser_id STRING,  
    browser STRING  
) ROW FORMAT DELIMITED FIELDS TERMINATED BY '-2';
```







WIKIPEDIA The Free Encyclopedia

- Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

- Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

- Toolbox
Print/export

Create account Log in

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



WIKIPEDIA  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

- Interaction
  - Help
  - About Wikipedia
  - Community portal
  - Recent changes
  - Contact Wikipedia


- Toolbox
- Print/export

Create account Log in

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



WIKIPEDIA  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

- Interaction
  - Help
  - About Wikipedia
  - Community portal
  - Recent changes
  - Contact Wikipedia


- Toolbox
- Print/export

Create account Log in

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-



WIKIPEDIA  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

- Interaction
  - Help
  - About Wikipedia
  - Community portal
  - Recent changes
  - Contact Wikipedia

Toolbox


Print/export

Create account Log in

Article **Talk** Read **Edit** View history Search

# 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.<sup>[1][2]</sup> 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** [\[show\]](#)



WIKIPEDIA  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

- Interaction
  - Help
  - About Wikipedia
  - Community portal
  - Recent changes
  - Contact Wikipedia

- Toolbox
- Print/export

Create account Log in

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.

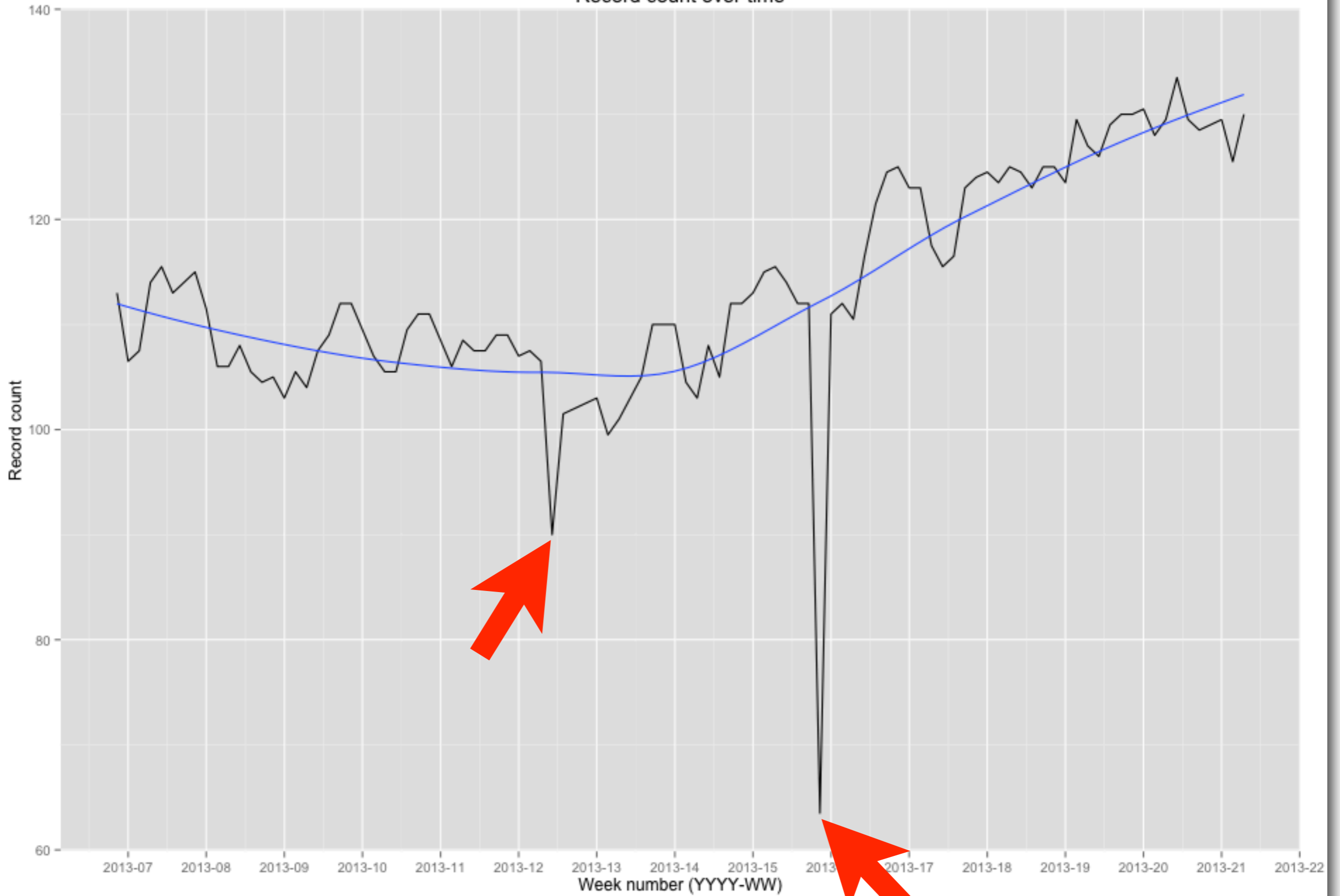
**Contents** [show]

## 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.

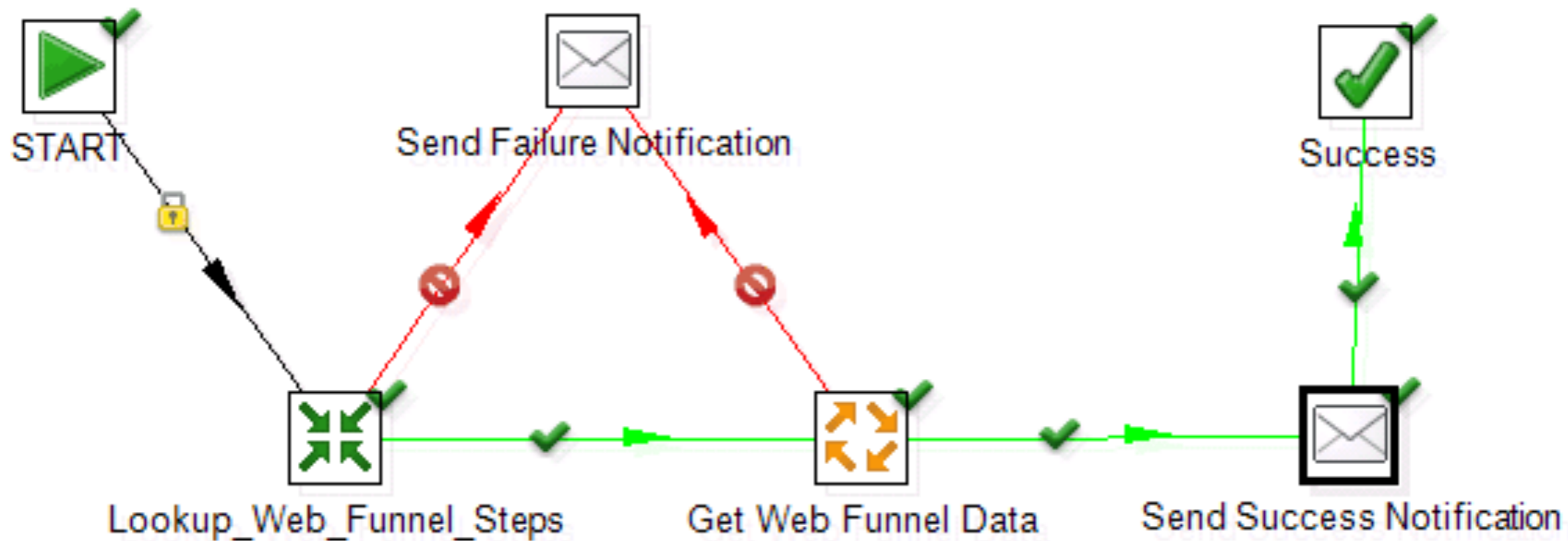
- Simple deployment of ETL code
- Scheduling
- Scalable
- Independent jobs
- Fixable data store
- Incremental where possible
- Metrics

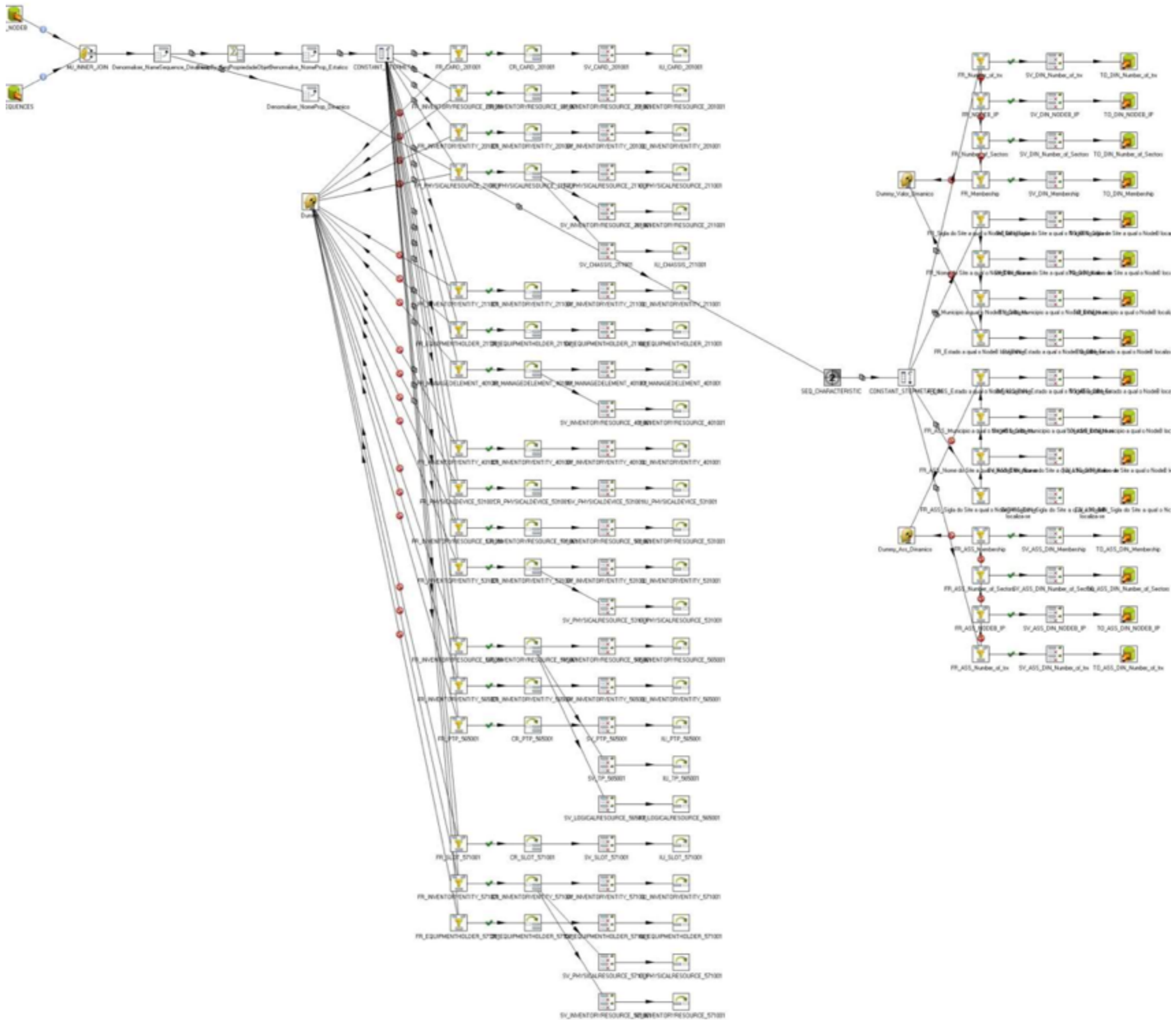


**EXTRACT**

**TRANSFORM**

**LOAD**







Shell

Shell

etl.sh

```
1 #!/bin/bash
2 set -ex
3
4 echo \
5 'open ftp.example.com
6 user username password
7 cd my-files
8 find /' > /tmp/ftp-list-command.txt
9
10 lftp -f /tmp/ftp-list-command.txt | grep "\.gz" > /tmp/full-names-on-ftp.txt
11
12 hadoop fs -ls -R /staging/my-files/ > /tmp/full-names-on-hadoop.txt
13 hadoop fs -ls -R /sources/my-files/ >> /tmp/full-names-on-hadoop.txt
14
15 cat /tmp/full-names-on-hadoop.txt | grep "\.gz" | awk '{print $8}' | awk -F '/' '{print $NF}' > /tmp/basenames-on-hadoop.txt
16
17 cat /tmp/full-names-on-ftp.txt | fgrep -v -f /tmp/basenames-on-hadoop.txt > /tmp/missing-files-on-hadoop.txt
18
19 echo \
20 'open ftp.example.com
21 user username password
22 ' > /tmp/ftp-get-files-command.txt
23 awk '{print "get -c " $0}' /tmp/missing-files-on-hadoop.txt >> /tmp/ftp-get-files-command.txt
24
25
26 cd /local-data/staging/my-files/
27 lftp -vf /tmp/ftp-get-files-command.txt
28 cd -
29
30 while read f
31 do
32     filename=$(echo $f | awk -F '/' '{print $NF}')
33     year=${filename:0:4}
34     month=${filename:4:2}
35     day=${filename:6:2}
36     site=$(echo $f | awk -F '-' '{print $NF}' | sed s/\.gz//g)
37
38     hadoop fs -mkdir "/staging/my-files/$year/$month/$day/$site"
39     hadoop fs -put "/local-data/staging/my-files/$filename" "/staging/my-files/$year/$month/$day/$site/$filename.upload"
40     hadoop fs -mv "/staging/my-files/$year/$month/$day/$site/$filename.upload" "/staging/my-files/$year/$month/$day/$site/$filename"
41 done < /tmp/missing-files-on-hadoop.txt
```

```
1 client = FTPClient(FTP_HOST, FTP_USER, FTP_PASSWORD)
2 remote_files = ftp_file_list(client, FTP_ROOT_DIR)
3
4 fs = hfs.FileSystem.get(hconf.Configuration())
5 hadoop_files = hdfs_file_list(fs, HDFS_ROOT_DIR_STAGING)
6 hadoop_files.extend(hdfs_file_list(fs, HDFS_ROOT_DIR_SOURCES))
7
8 missing_files = filter(lambda f: f not in hadoop_files, remote_files)
9
10 filename_pattern = re.compile(r"(\d{4})(\d{2})(\d{2})\d{4}-\d{12}-(\d{6})\.gz")
11
12 for missing_file in missing_files:
13     temp_file = File(FTP_LOCAL_DOWNLOAD_PATH, missing_file)
14     client.download(missing_file, temp_file)
15
16     year, month, day, site = filename_pattern.match(missing_file[1].name).groups()
17     hdfs_dir = hfs.Path('%s/%s/%s/%s/%s' % (HDFS_UPLOAD_BASE_PATH, year, month, day, site))
18
19     hdfs_dir = hfs.Path('%s/%s/%s/%s/%s' % tuple(filename_pattern.match(missing_file).groups()))
20
21     if (not fs.exists(hdfs_dir)):
22         fs.mkdirs(hdfs_dir)
23
24     fs.copyFromLocalFile(hfs.Path(temp_file.path), hfs.Path(remotedir, remotefilename + '.upload'))
25     fs.rename(hfs.Path(remotedir, remotefilename + '.upload'), hfs.Path(remotedir, remotefilename))
26
27     os.unlink(localfile.path)
28
29
30 fs.close()
31 client.disconnect(False)
32
```



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

CRONTAB(1)

BSD General Commands Manual

CRONTAB(1)

**NAME**

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

**SYNOPSIS**

```
crontab [-u user] file  
crontab [-u user] { -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 /usr/lib/cron/cron.allow file exists, then you must be listed therein in order to be allowed to use this command. If the /usr/lib/cron/cron.allow file does not exist but the

:





# Jenkins

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

http://localhost:8080 (All)

- successful job #380
- successful 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
```

# Independent jobs

source (external)

HDFS upload + move in place

staging (HDFS)

MapReduce + HDFS move

hive-staging (HDFS)

Hive map external table + SELECT INTO

Hive

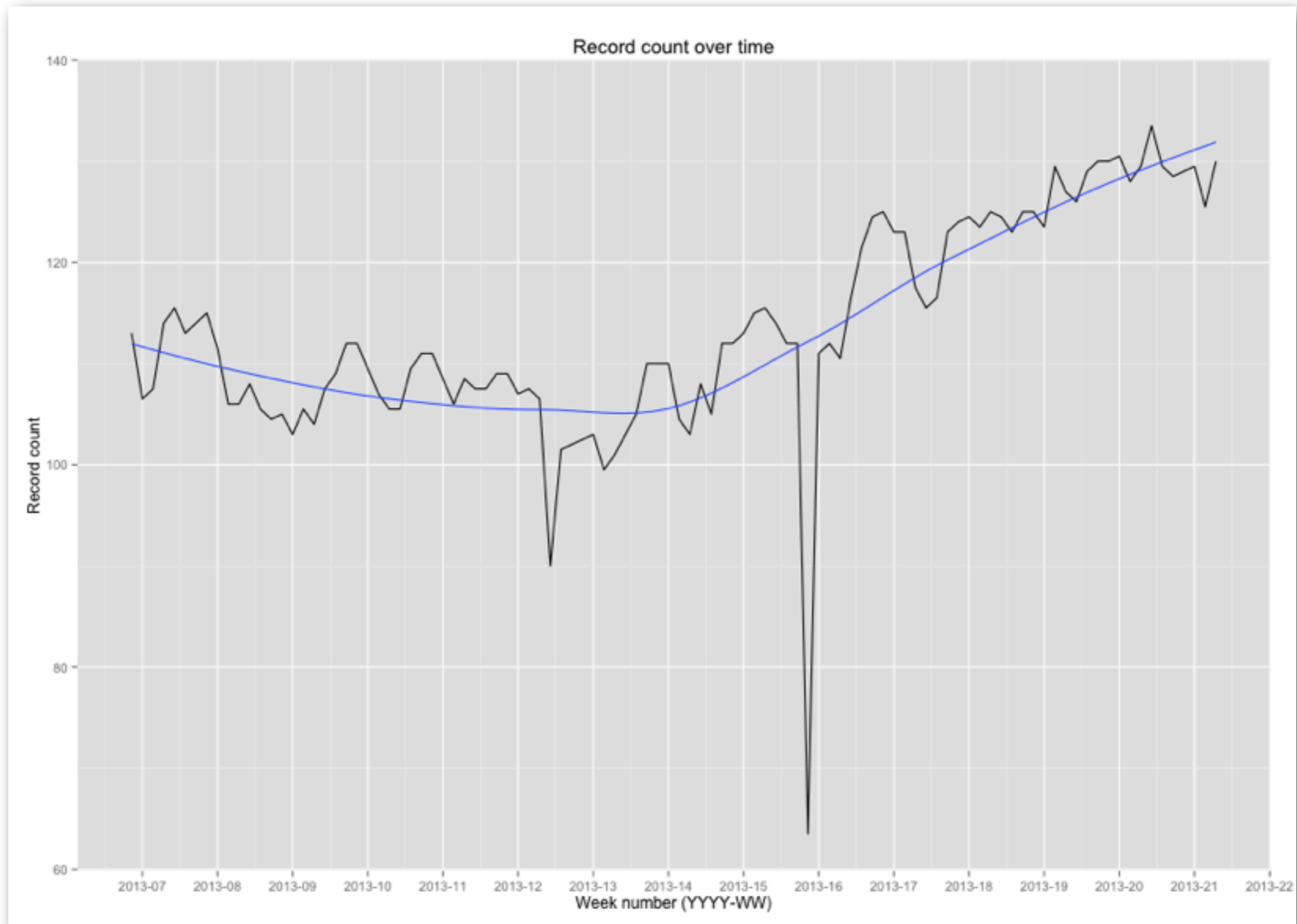
# Out of order jobs

- 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

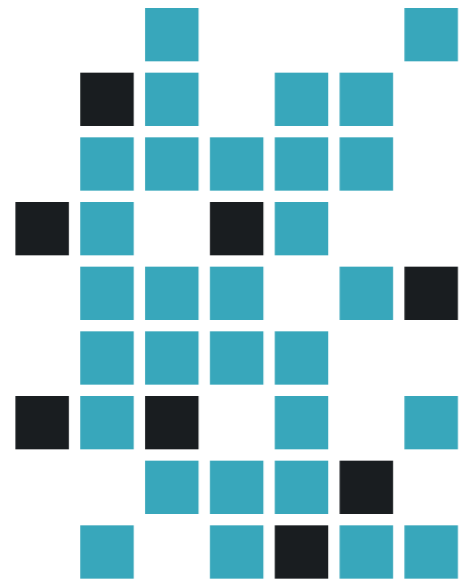
# Metrics



# Metrics service

- Job ran, so many 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!