Insights into Customer Behavior from Clickstream Data

Ronald J. Nowling
Red Hat, Inc.
rnowling@redhat.com
http://rnowling.github.io/

Who Am I?

- Software Engineer at Red Hat
- Data Science Team, Emerging Technologies
  - Evaluate solutions in open-source Big Data space
  - Ensure software works for Red Hat customers
  - Promote data science internally through consulting projects
- Apache Bigtop PMC

Clickstream Data

Clickstream Data

61 million page views

Clickstream Data

61 million page views

125,000 registered users

Clickstream Data

61 million page views

125,000 registered users

500,000 pages

Clickstream Data

61 million page views

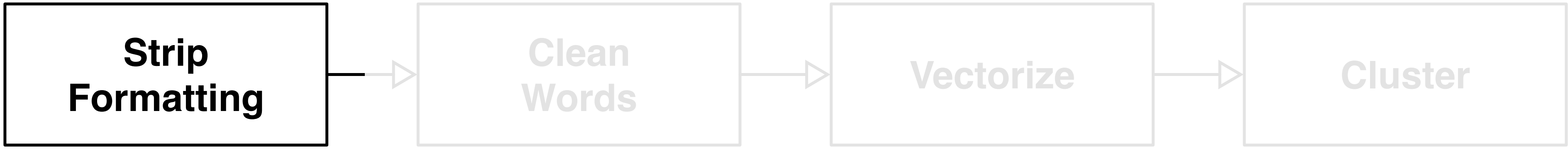
125,000 registered users

500,000 pages

125,000 knowledgebase articles

Potential Applications

- Build customer profiles to aid sales teams
- Recommendation system for knowledgebase
- Improve customer portal search
- Guide selection of new knowledgebase topics by content writers



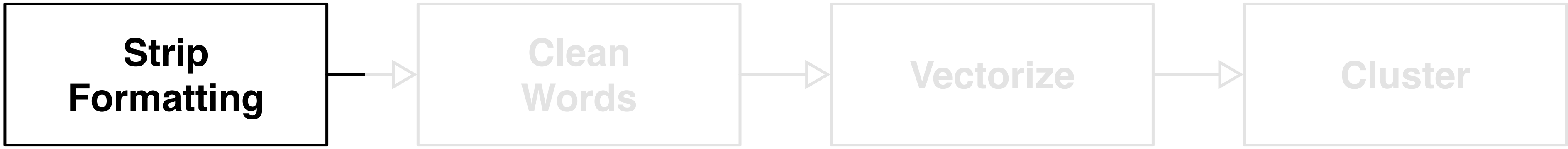
Strip Formatting

Clean Words

Vectorize

Cluster

```
What are the different types of kernel packages in Red Hat
Enterprise Linux?
=============================================================
Issue
------
What are the different types of kernel packages in Red Hat
Enterprise Linux?

Environment
---------------
Red Hat Enterprise Linux

Resolution
------------
Red Hat Enterprise Linux contains the following kernel
packages:
```



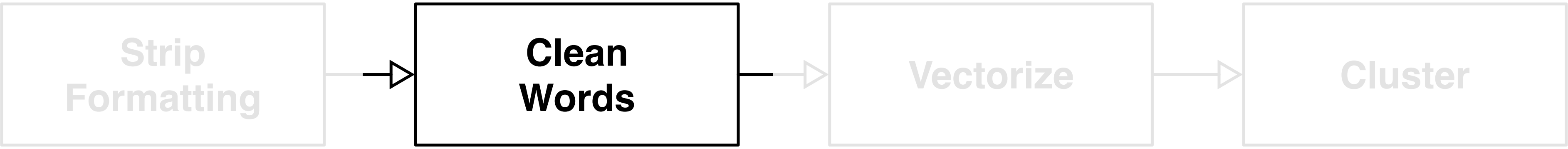
Strip Formatting

Clean Words

Vectorize

Cluster

```
What are the different types of kernel packages in Red Hat
Enterprise Linux

Issue
What are the different types of kernel packages in Red Hat
Enterprise Linux

Environment
Red Hat Enterprise Linux

Resolution
Red Hat Enterprise Linux contains the following kernel
packages some may not apply to your architecture and not all
are available in all major releases kernel contains the
kernel and following key features
```


Strip Formatting

Clean Words

Vectorize

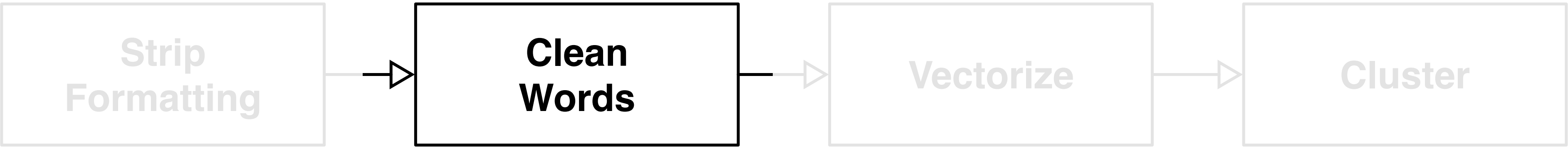
Cluster

What are the different types of kernel packages in Red Hat Enterprise Linux

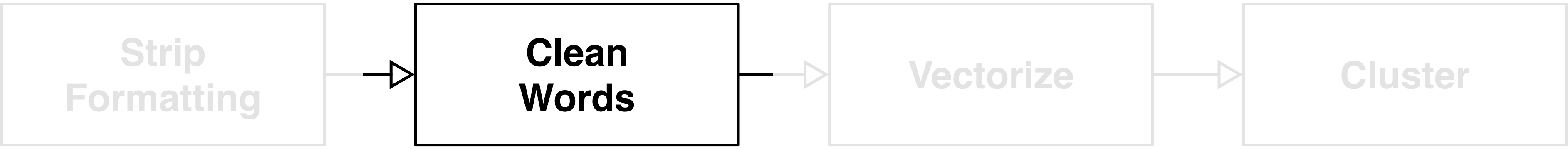
Issue
What are the different types of kernel packages in Red Hat Enterprise Linux

Environment
Red Hat Enterprise Linux

Resolution
Red Hat Enterprise Linux contains the following kernel packages some may not apply to your architecture and not all are available in all major releases kernel contains the kernel and following key features



```
What are the different type  of kernel package  in Red Hat
Enterprise Linux

Issue
What are the different type  of kernel package  in Red Hat
Enterprise Linux

Environment
Red Hat Enterprise Linux

Resolution
Red Hat Enterprise Linux contain  the follow    kernel
package  some may not apply to your architecture and not all
are available in all major release  kernel contain  the
kernel and follow    key feature
```



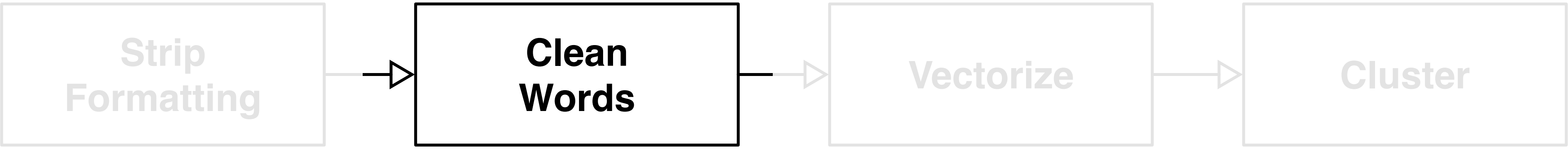
Strip Formatting

Clean Words

Vectorize

Cluster

```
What are the different type  of kernel package  in Red Hat
Enterprise Linux

Issue
What are the different type  of kernel package  in Red Hat
Enterprise Linux

Environment
Red Hat Enterprise Linux

Resolution
Red Hat Enterprise Linux contain  the follow    kernel
package  some may not apply to your architecture and not all
are available in all major release  kernel contain  the
kernel and follow    key feature
```

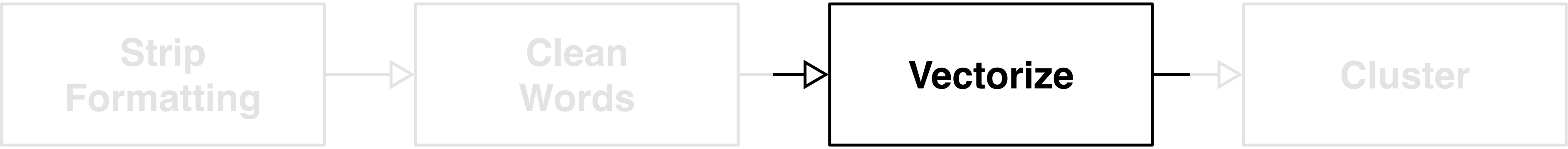

Strip Formatting

Clean Words

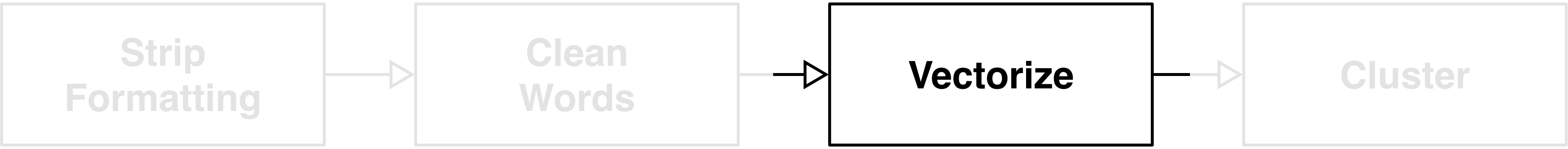
Vectorize

Cluster

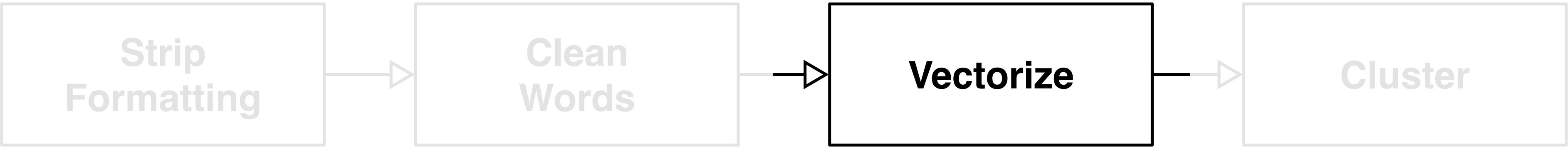
```
             different type     kernel package     Red Hat
Enterprise Linux

Issue
             different type     kernel package     Red Hat
Enterprise Linux

Environment
Red Hat Enterprise Linux

Resolution
Red Hat Enterprise Linux contain                kernel
package               apply         architecture
    available        major release  kernel contain
kernel     follow    key feature
```



```
kernel: 5
red: 4
hat: 4
enterprise: 4
linux: 4
package: 3
contain: 3
different: 2
type: 2
intel: 2
environment: 1
resolution: 1
follow: 1
system: 1
```



```
kernel: 5                  different: 2
red: 4                     type: 2
hat: 4                     intel: 2
enterprise: 4              environment: 1
linux: 4                   resolution: 1
package: 3                 follow: 1
contain: 3                 system: 1
```



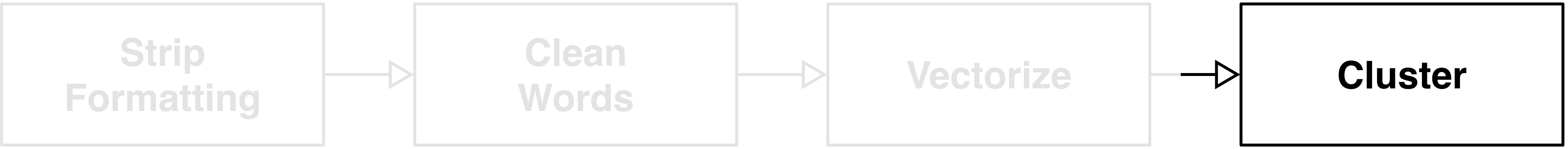
Strip Formatting

Clean Words

Vectorize

Cluster

```
kernel: 5
red: 4
hat: 4
enterprise: 4
linux: 4
package: 3
contain: 3
```

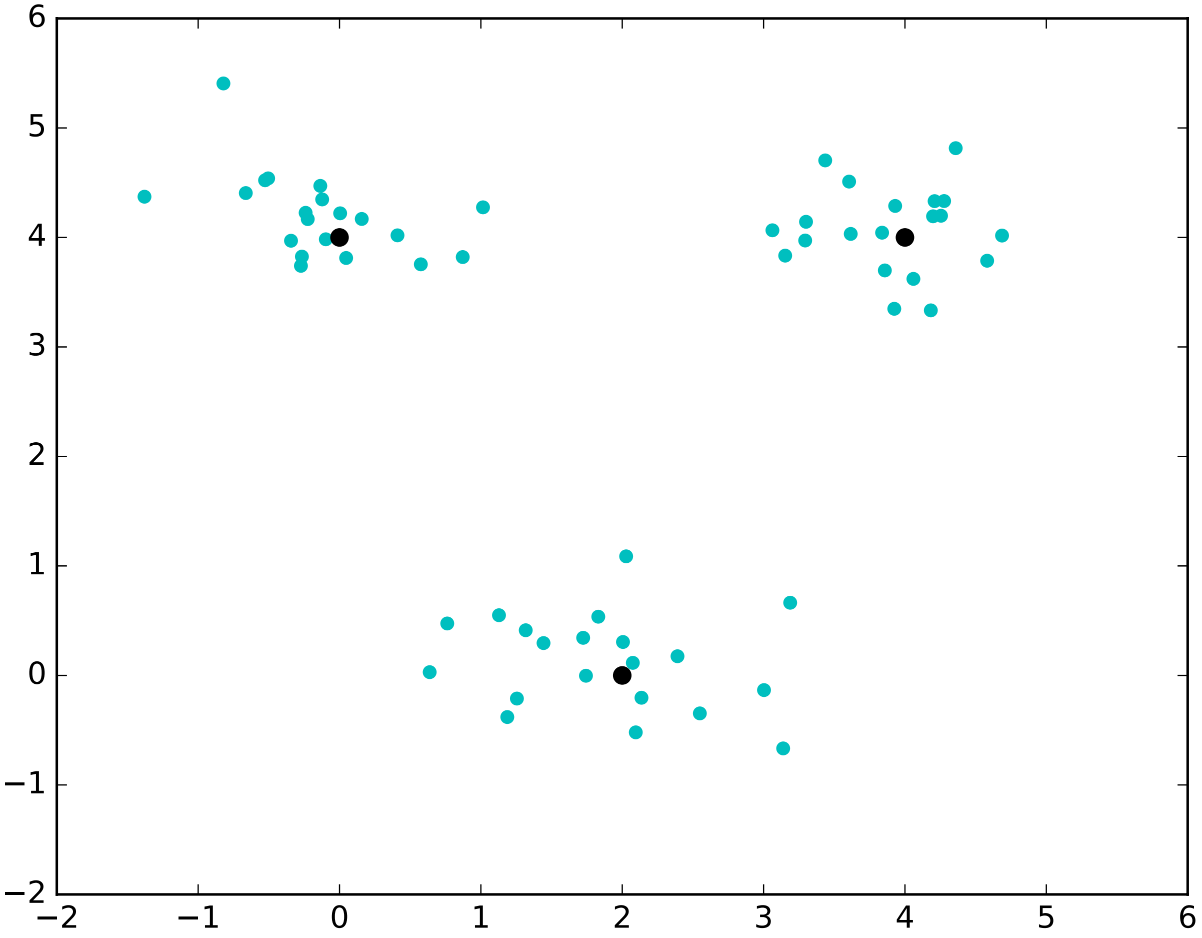


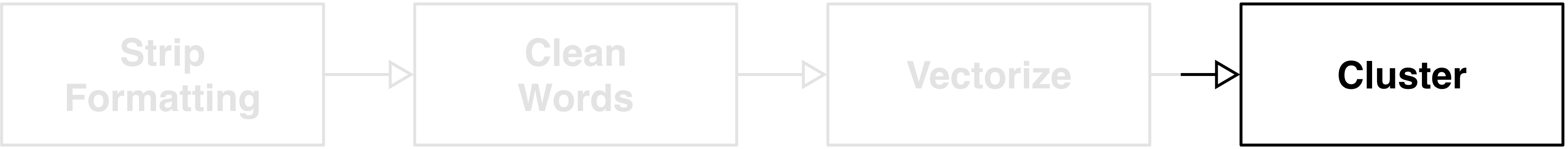
Strip Formatting

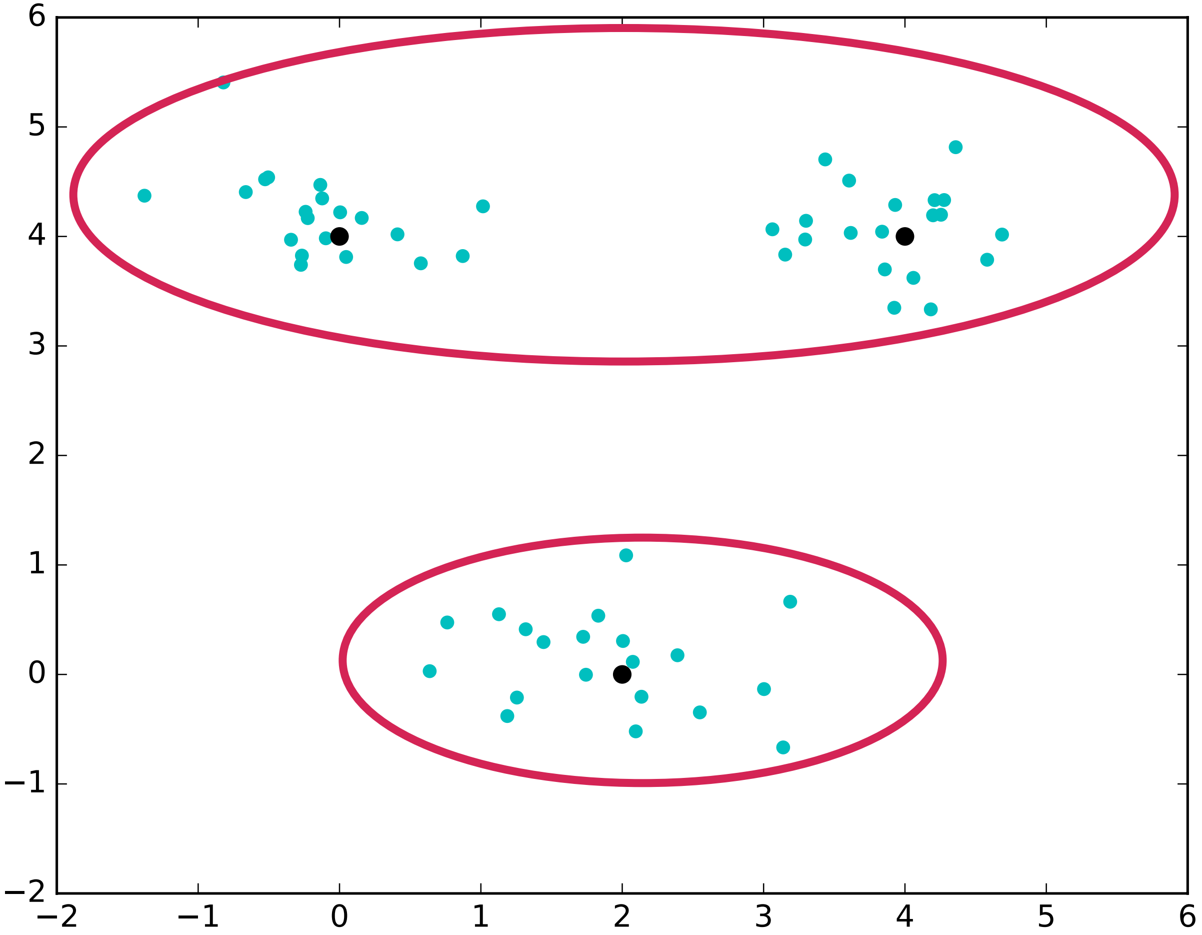
Clean Words

Vectorize

Cluster




Strip Formatting
Clean Words
Vectorize
Cluster


6
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
5
6

Topics

openshift gear cartridge online
node broker

vm rhev virtualization disk

glusterfs storage volume brick rhs
glusterd node client mount geo

rhel support driver hp hardware
version firmware card intel

Topics

openshift gear cartridge online node broker

vm rhev virtualization disk

glusterfs storage volume brick rhs glusterd node client mount geo

rhel support driver hp hardware version firmware card intel

Topics

openshift gear cartridge online
node broker

vm rhev virtualization disk

glusterfs storage volume brick rhs
glusterd node client mount geo

rhel support driver hp hardware
version firmware card intel

Topics

openshift gear cartridge online node broker

vm rhev virtualization disk

glusterfs storage volume brick rhs glusterd node client mount geo

rhel support driver hp hardware version firmware card intel

Topics

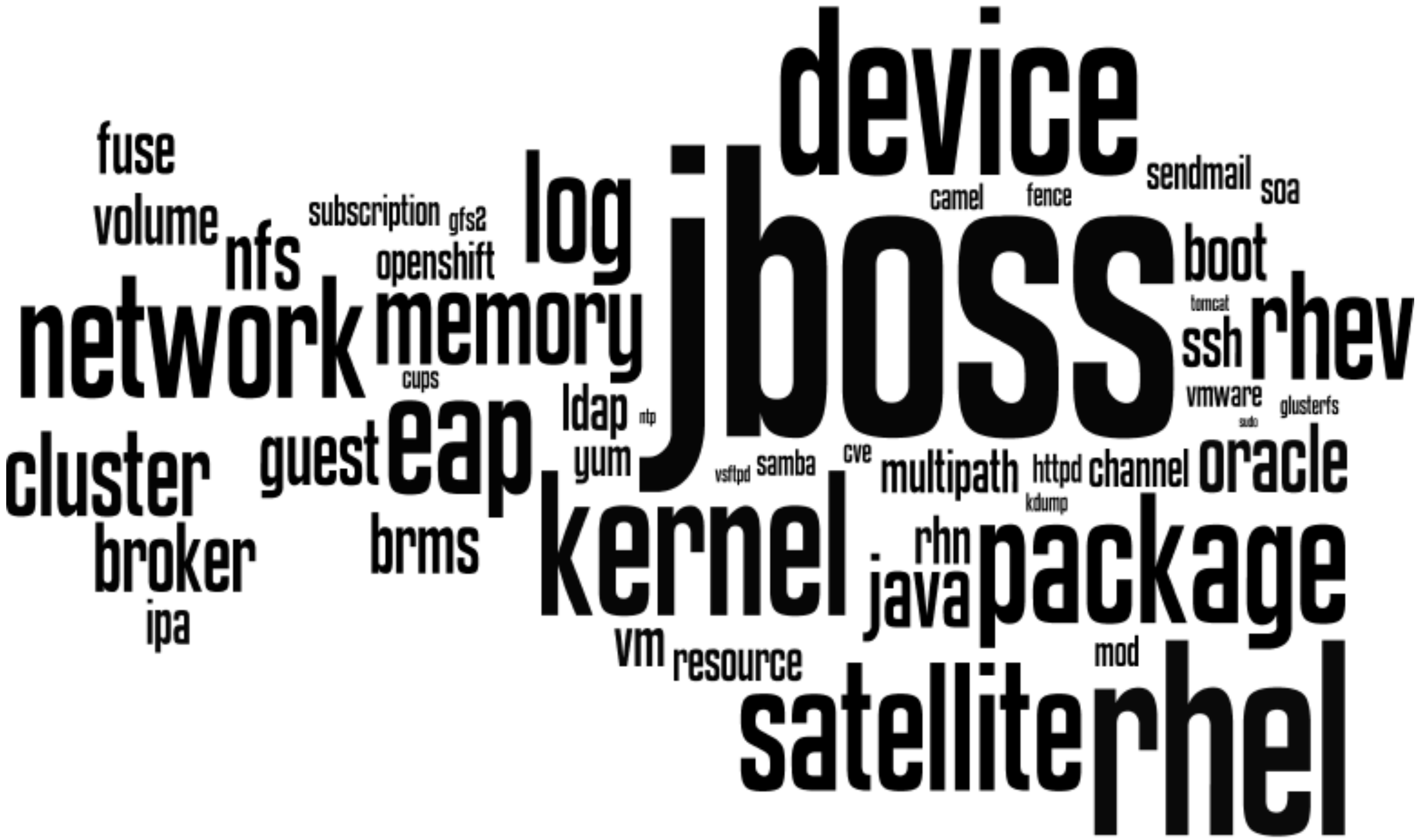
openshift gear cartridge online node broker

vm rhev virtualization disk

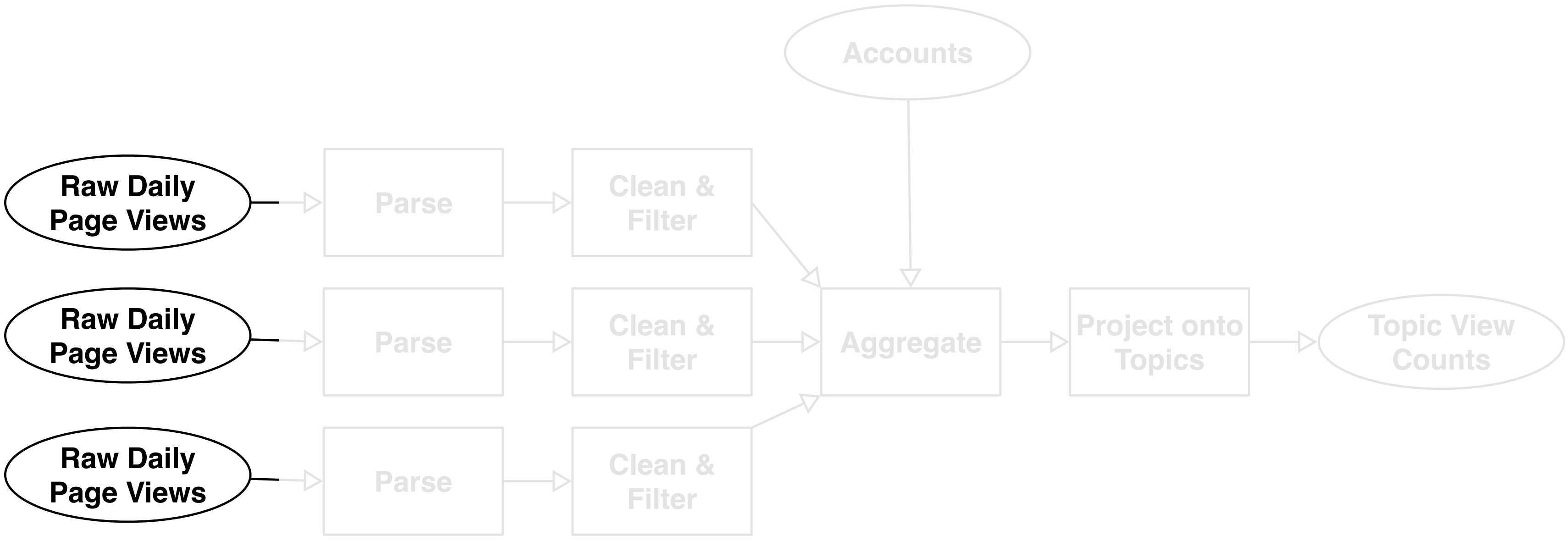
glusterfs storage volume brick rhs glusterd node client mount geo

rhel support driver hp hardware version firmware card intel

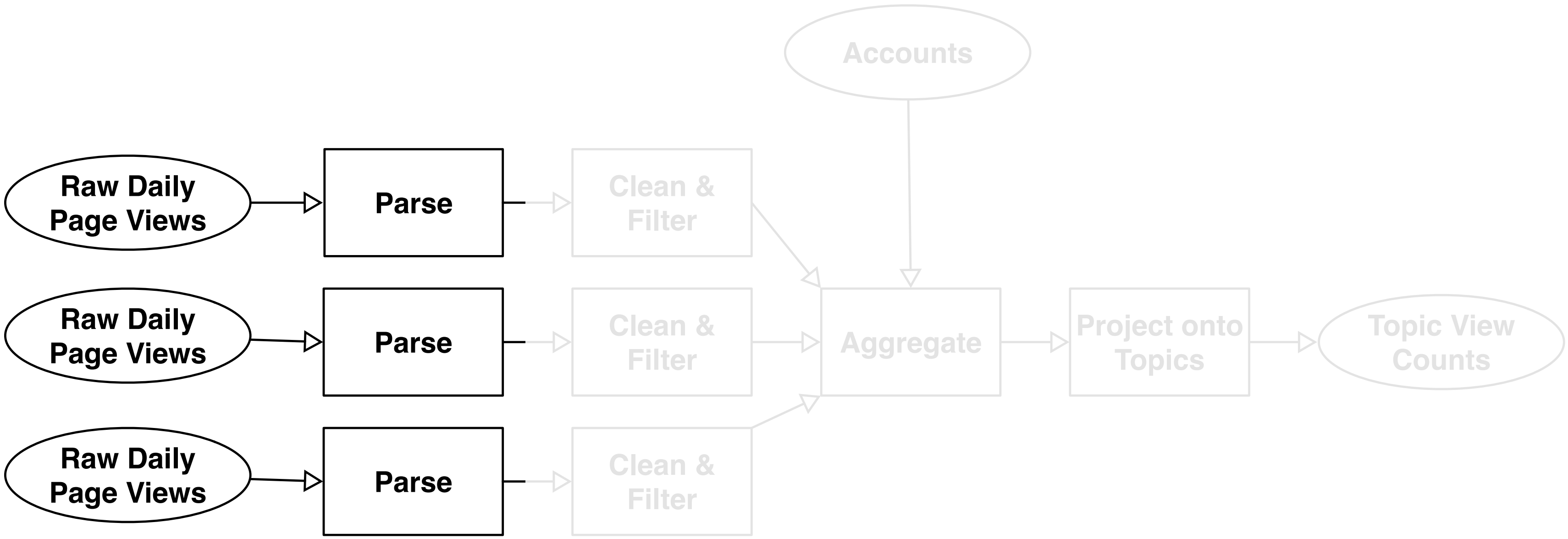
Topic Article Counts



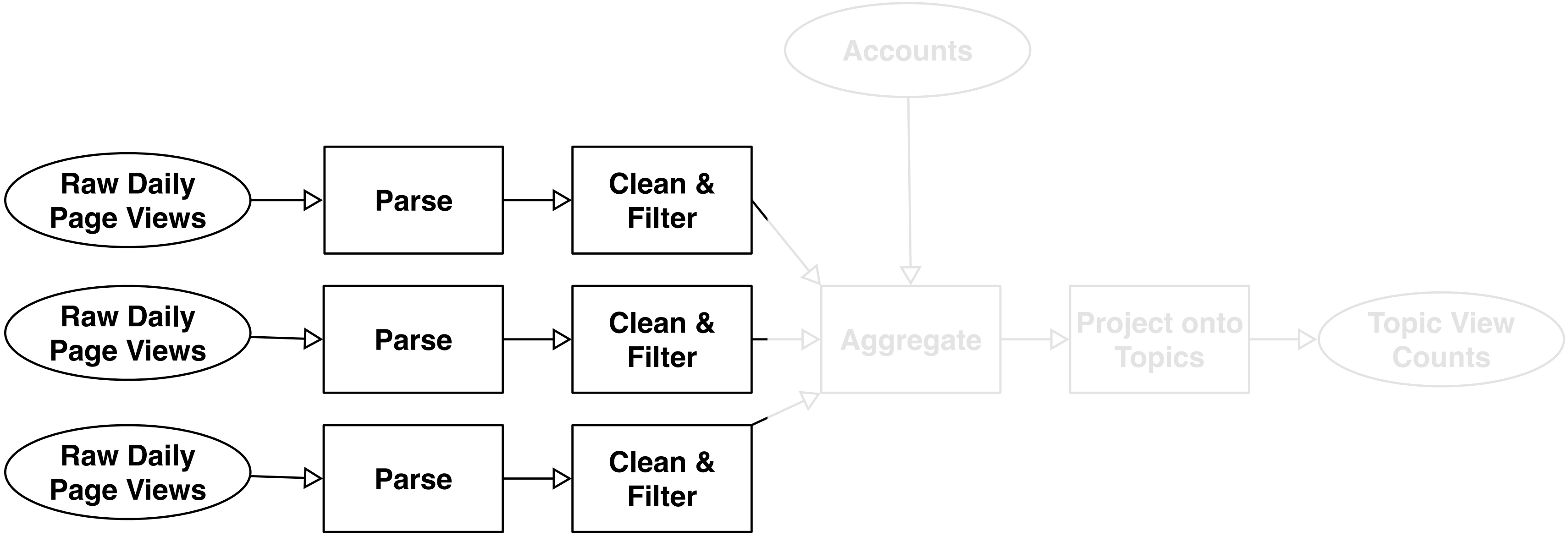
Clickstream Processing



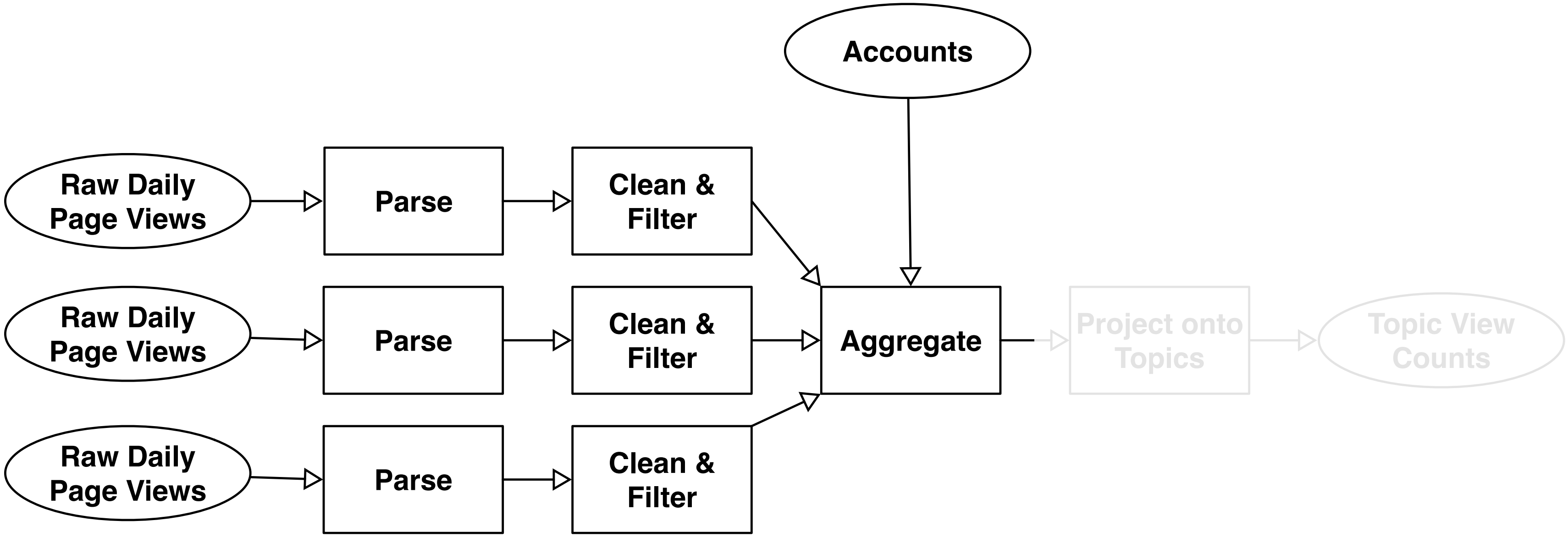
Clickstream Processing



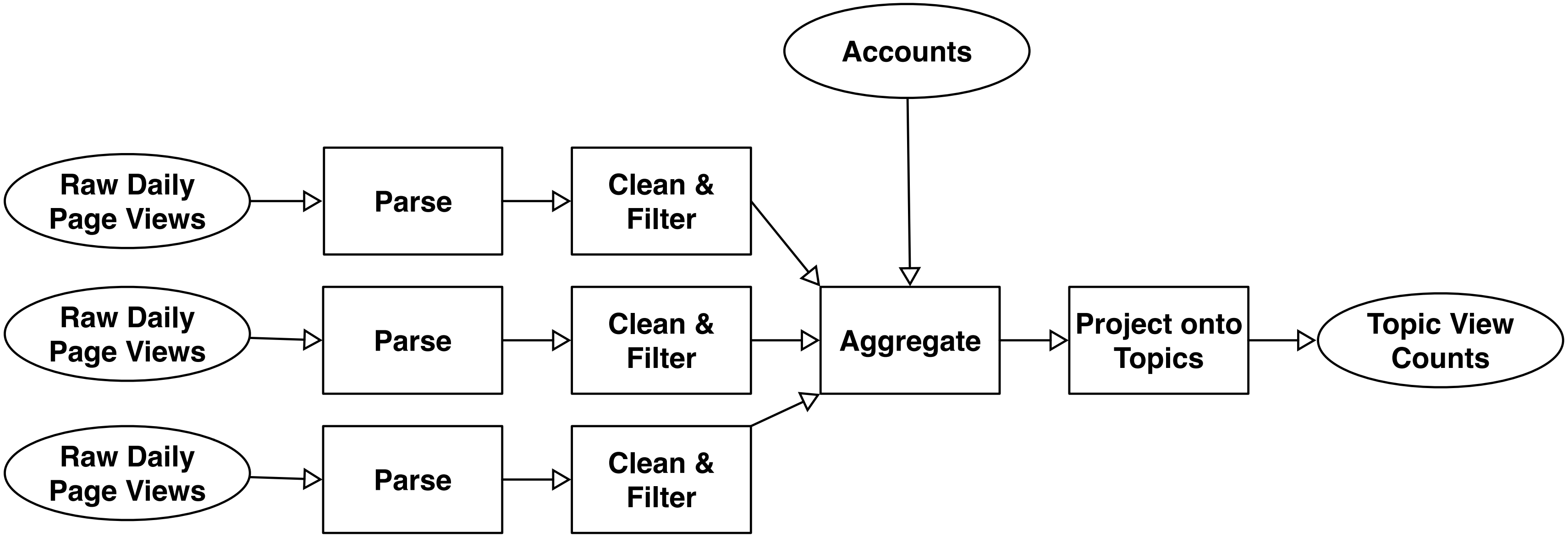
Clickstream Processing



Clickstream Processing



Clickstream Processing



Customer Profiles

- Dominant topics
  - JBoss
  - Red Hat Enterprise Virtualization
  - Hardware support
  - Gluster
  - Booting into rescue mode
  - Packages

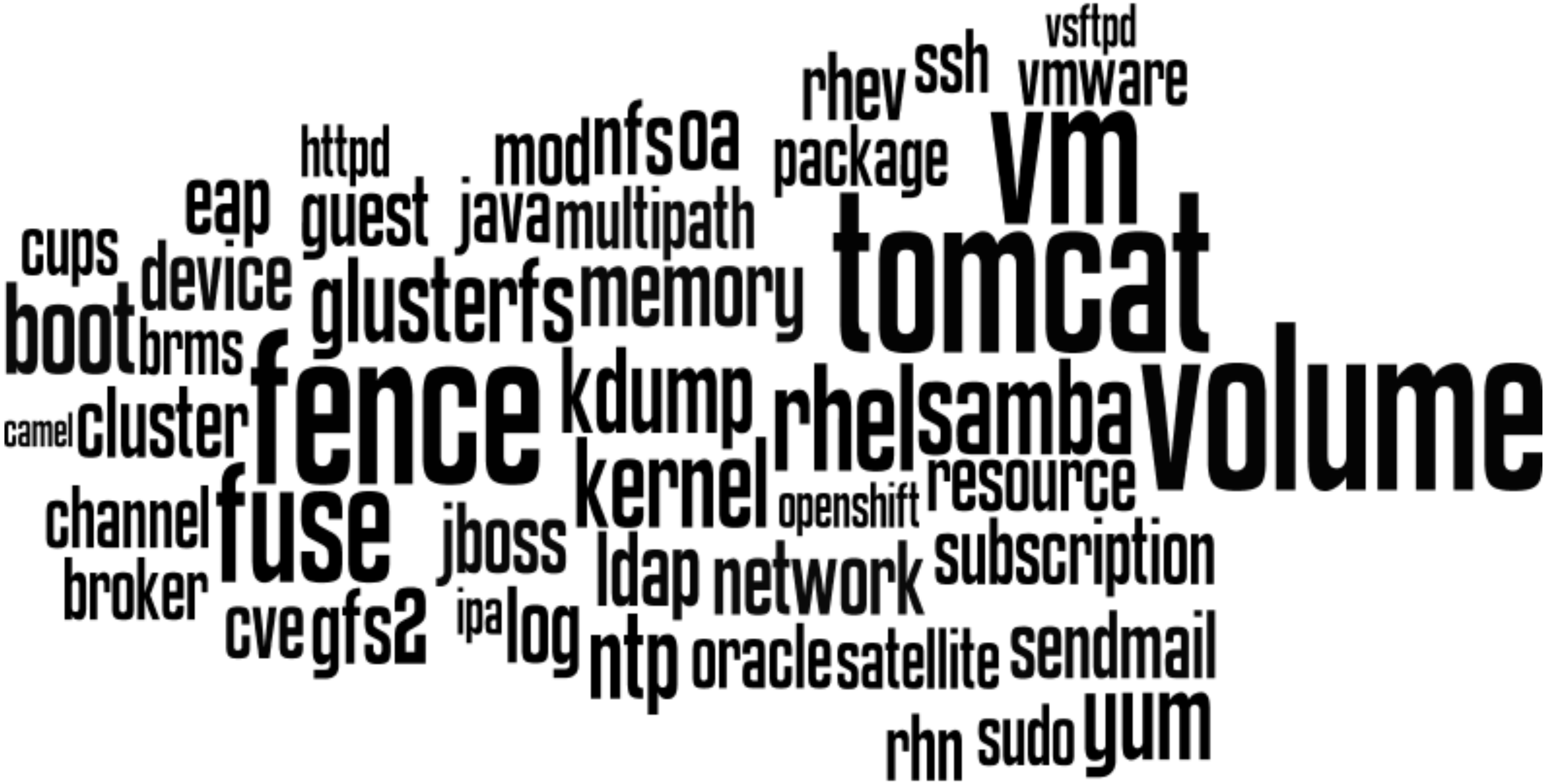
Customer Profiles

- Supporting topics
  - Logging
  - LDAP
  - Samba
  - High resource usage
  - File systems / LVM / block devices
  - Networking

Customer Profiles

- JBoss and RHEV appear in combination with a number of other products
- Some products only appear by themselves with supporting topics (logging, networking, filesystems)
  - OpenShift
  - Gluster

Topic Enrichments



Malformed TSV Files

- Gzip files need to be read sequentially
- Tab-separated, no quoting (in theory!)
- Escaped tabs and newlines within records
  - E.g., `\\n` or `\\t`
- Improperly escaped tabs and newlines
  - E.g., `\\\t` vs `\\\\t`
- Extraneous unmatched quote marks
  - E.g., ‘`some_user`

Lessons Learned

- Consider custom Hadoop input formats for tricky file formats
- Verify everything – what works in general may not work for you
  - Stemming
  - Filtering most frequent words
  - K-Means vs LDA

Lessons Learned

- K-Means
  - Improve accuracy: Multiple runs, more iterations
- Watch out for memory leaks
  - Un-persist cached RDDs
  - Un-persist broadcasted variables
- Parquet for performance

Potential Applications

- Build customer profiles to aid sales teams
- Recommendation system for knowledgebase
- Improve customer portal search
- Guide selection of new knowledgebase topics for content writers

Resources

http://rnowling.github.io/

QUESTIONS