Composing Mahout clustering jobs

Frank Scholten

frank@jteam.nl

Bio

- Frank Scholten
- Developer at 

Amsterdam, NL
- Mahout user / contributor
- http://blog.jteam.nl/author/frank

Agenda

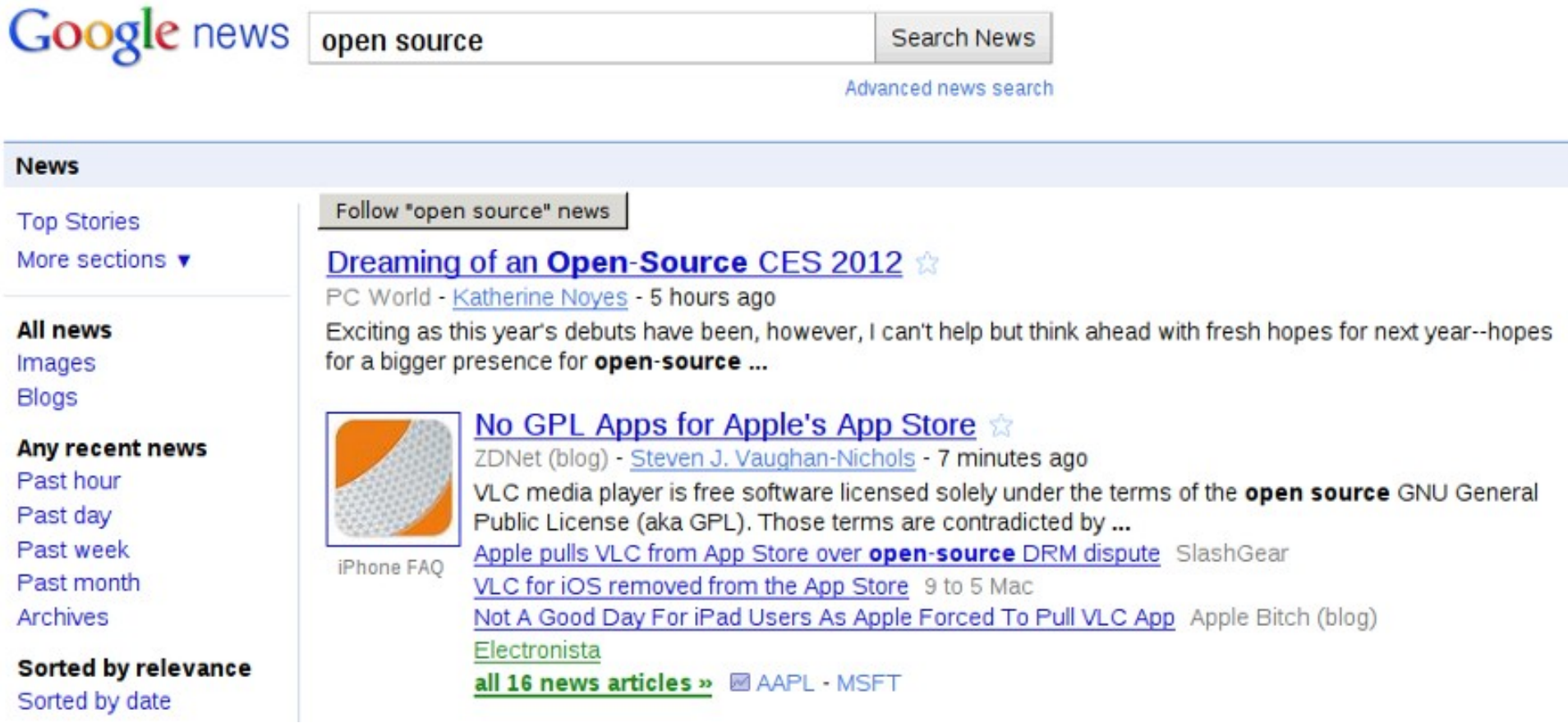
What is clustering?

Introducing mahout

Clustering stackoverflow

What is clustering?

Clustering - Google News



Why clustering?

- Summarizing data

- Applications

  Market analysis – identify customer groups

  Biology – identify species

  Image compression

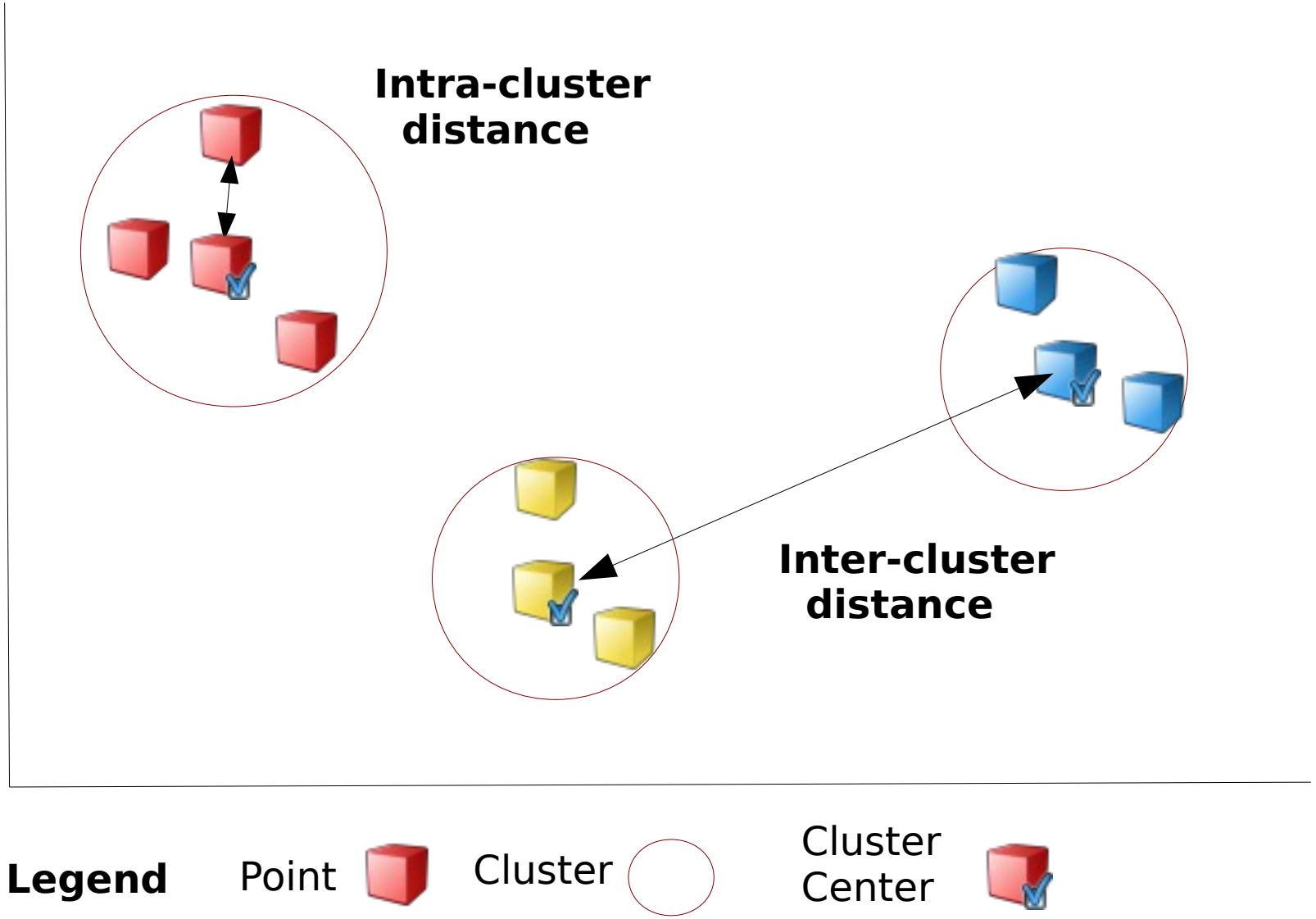
  many more applications!

Definition

“Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense.”

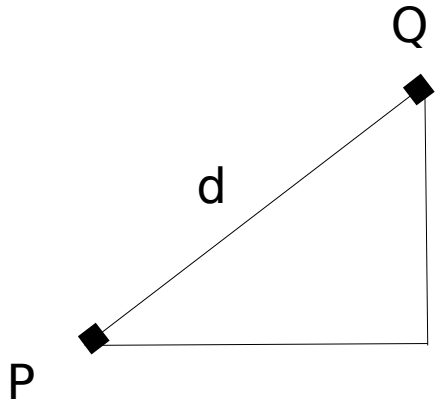
Source: *Wikipedia*

2-D Clustering Example



Distance Measures

Euclidian distance measure



$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Vectorization

Vectorize data to measure distances

'The fox chased the dog'

- [the => 2, fox => 1, chased => 1, dog => 1]

#0000CD → [wavelength => 475]

“Amsterdam” → [lat => 52, long => 4]

K-Means Algorithm

Select K random vectors

Specify distance measure + threshold

Every iteration

- Add vector closest to cluster
- Recompute center
- Converged if no vectors within threshold

Introducing

Scalable machine learning

On top of Hadoop, for the most part

Started in 2008

Version 0.5 released last week!

Collaborative Filtering

Is this SPAM?

Classification

Clustering

And much more!

Composing several jobs

```
$ mahout seqdirectory  <options>
$ mahout seq2sparse        <options>
$ mahout kmeans                <options>
$ mahout clusterdump  <options>
```

bin/mahout

$ mahout seqdirectory --help

Running on hadoop, using HADOOP_HOME=/usr/local/hadoop

HADOOP_CONF_DIR=/usr/local/hadoop/conf

Usage:

[--keyPrefix <keyPrefix> --chunkSize <chunkSize> --charset <charset> --output

<output> --fileFilterClass <fileFilterClass> --help --input <input>]

bin/mahout

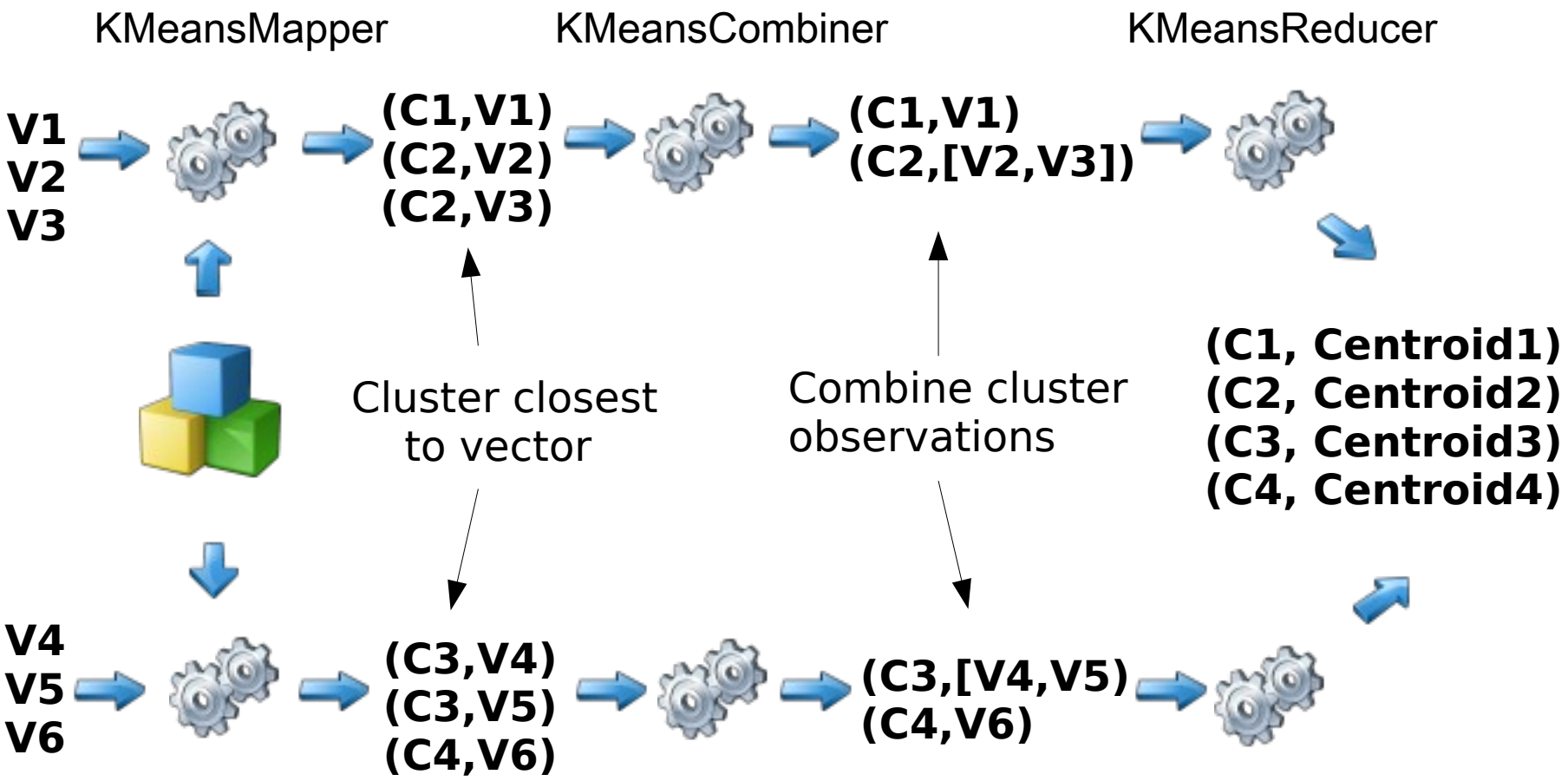
- bin/mahout calls MahoutDriver
- MahoutDriver
  - Parses options
  - Configures other Drivers

KMeansDriver

```
String[] args = new String[] {
  "--input", input,
  "--output", output,
  "--clusters",  clusters,
  "--clustering",
  "--numClusters", “10”
};

ToolRunner.run(conf,new KmeansDriver(),args);
```

KMeansDriver

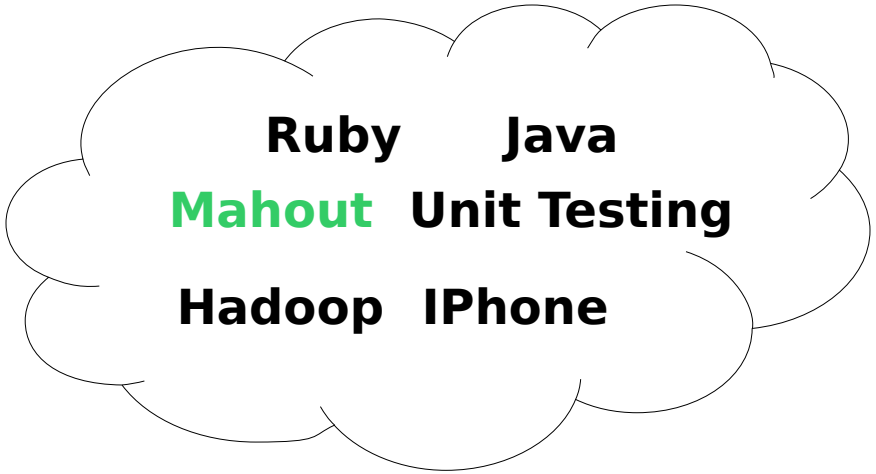


Clustering

Clustering

- Publicly available monthly dumps
- Posts ~ **5.5 GB** ~ **1.4 M** questions
- Inspired by **Mahout in Action** book

Goal - Tag cloud



(250 tags)

Questions

Deploying Mahout on a Hadoop cluster

Datasets for Apache Mahout

Classifying data using Apache Mahout

...

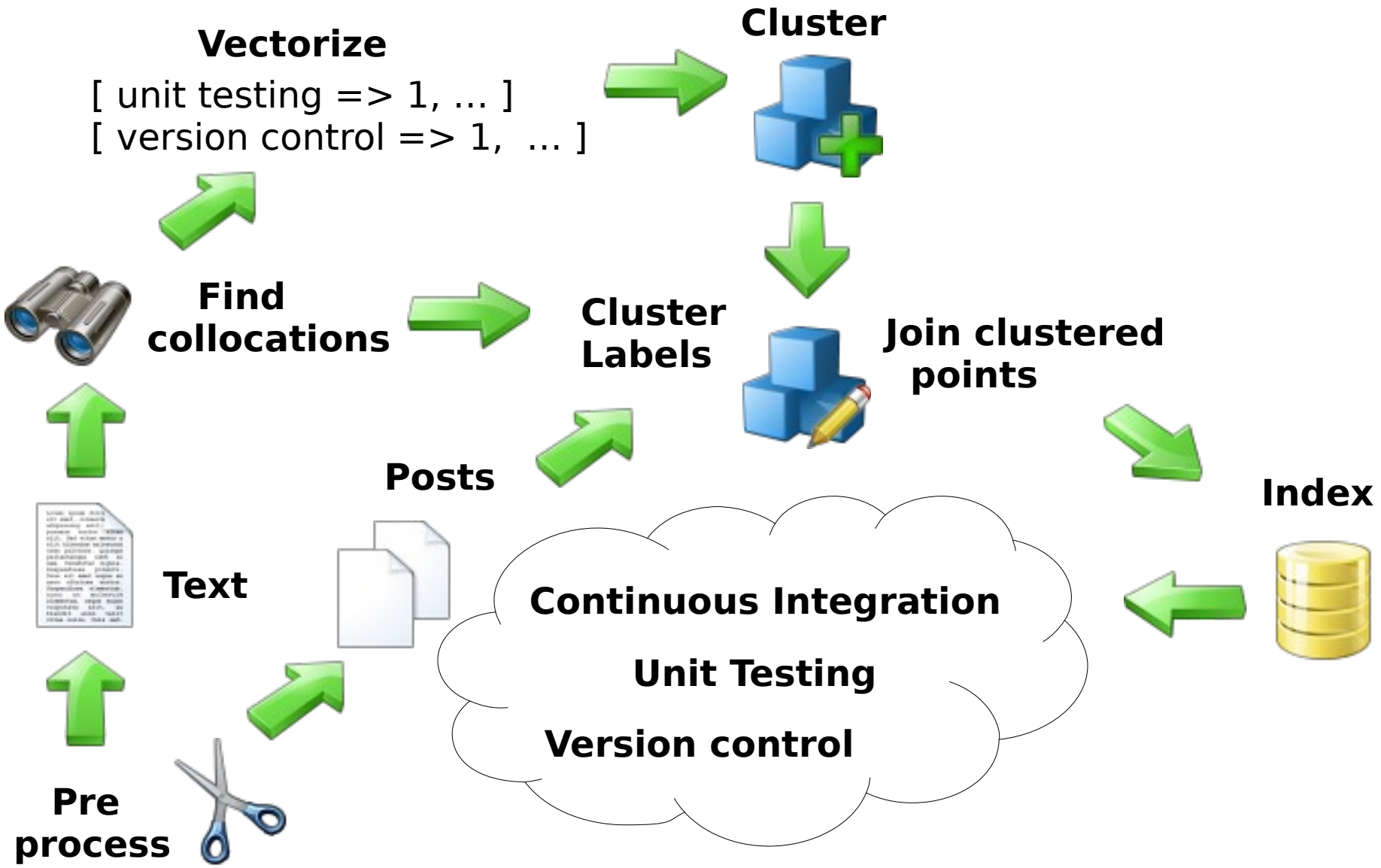
(unknown #)

How to cluster?

- Stopwords influence clustering big time!
- Option?
  - Cluster on **collocations**, e.g *“Unit Testing”*
- **MAHOUT-415**

  Lucene filter for collocations

Clustering on collocations



Pre process

StackOverflow posts

Use Mahout's XMLInputFormat

<row Id="4234" PostTypeId="1" content=”...”>
<row Id="136" PostTypeId="2" content=”...”>
<row Id="985" PostTypeId="1" content=”...”>

“
”<p
How to use Mahout</p>

Extract post body

Interpret & Strip HTML

How to use Mahout

Plain text

Find collocations

Unigrams

- **“Java”**, **“Ruby”**

Bigrams

- **“Continuous Integration”**, **“Unit Testing”**

Find collocations 

- Use Mahout's CollocDriver
- Compute LogLikelihood Ratio (LLR) (Dunning)
- Select bigrams with high LLR

Find collocations 

Save collocations in **Bloom Filter**

Add collocation
"Unit Testing"

Bloom Filter

0, 2, 5, 4

[1, 0, 1, 0, 1, 1, 0, 0, ...]

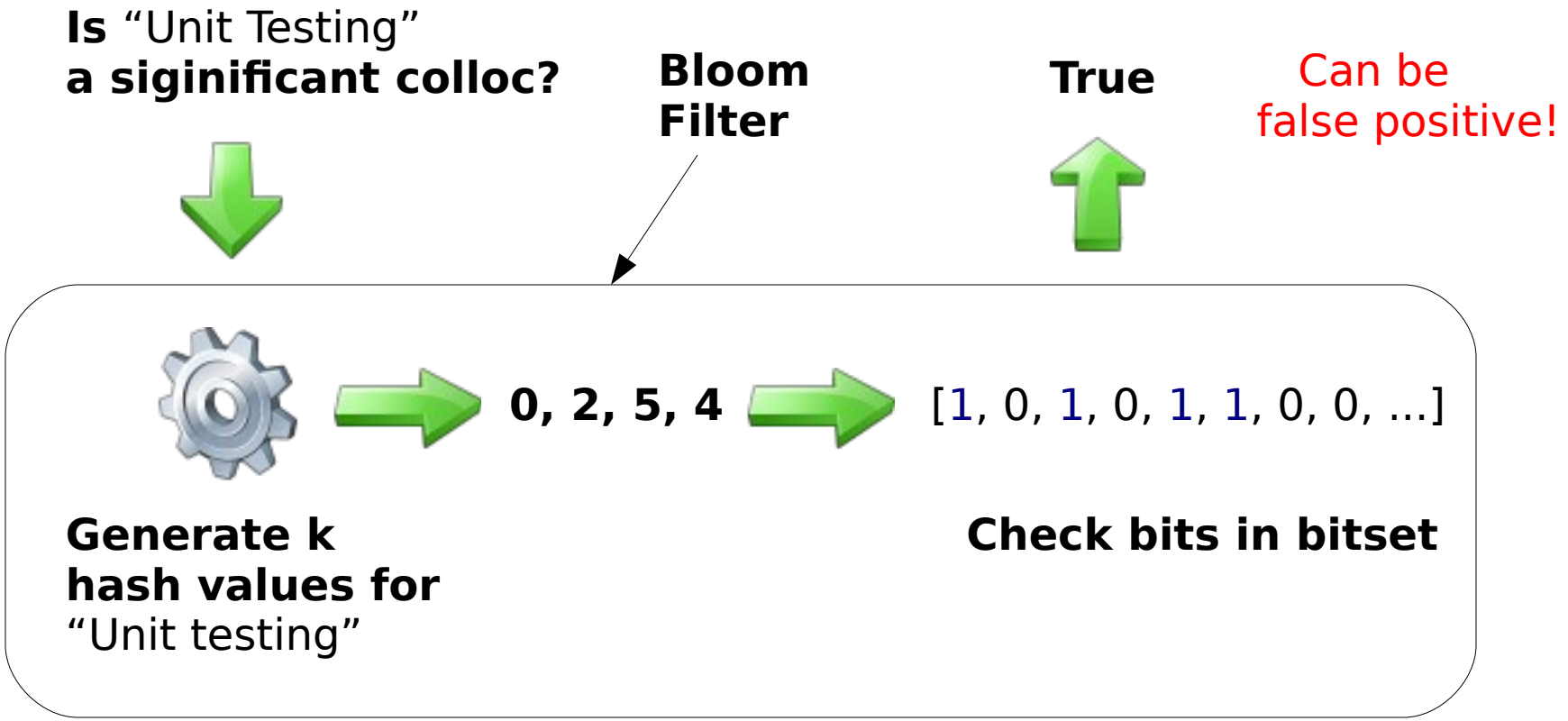
Generate k hash values for
"Unit testing"

Set bits in bitset

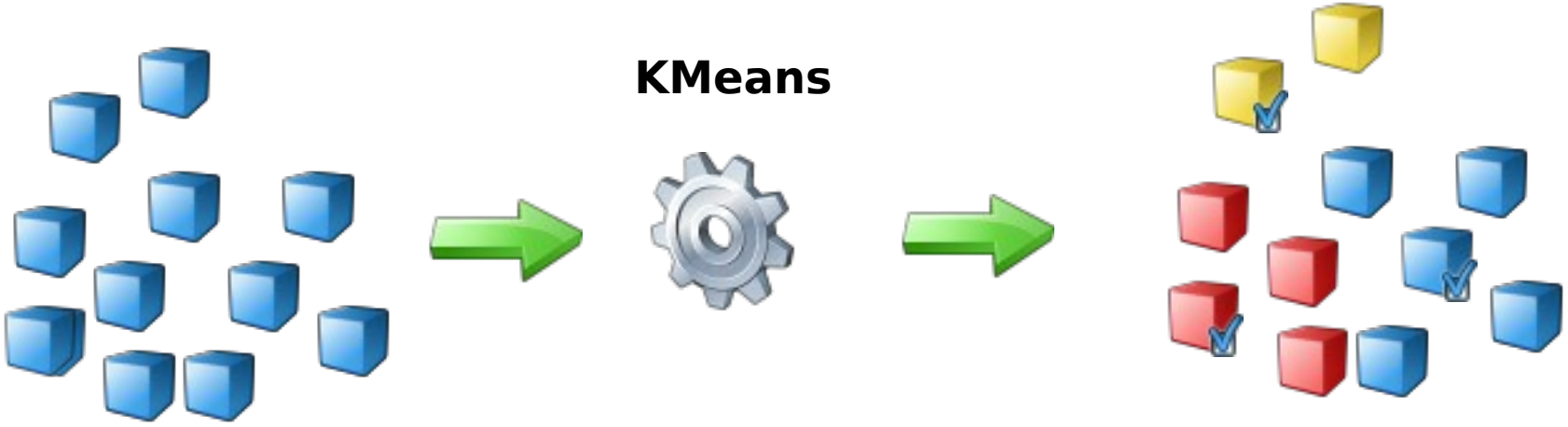
Vectorize

[Unit testing => 1, ...]
[Version control => 1, ...]

Lucene analyzer emits collocations



Cluster



Join clustered points

Posts
Sequence file

[(id), (title, content)]

Clustered points

(id = 23 ,)

(id = 78 ,)

(id = 51 ,)

(id = 23 ,)

(id = 34 ,)

Map-side join

[(id), (title, content, clusterId)]

Cluster labels

Cluster centroid

[0.56, 0.32, 0.98, ...]

Term with highest weight

0.98

Dictionary file

0 => Unit testing
1 => Version control
2 => Continuous integration
...

Label “Continuous integration”

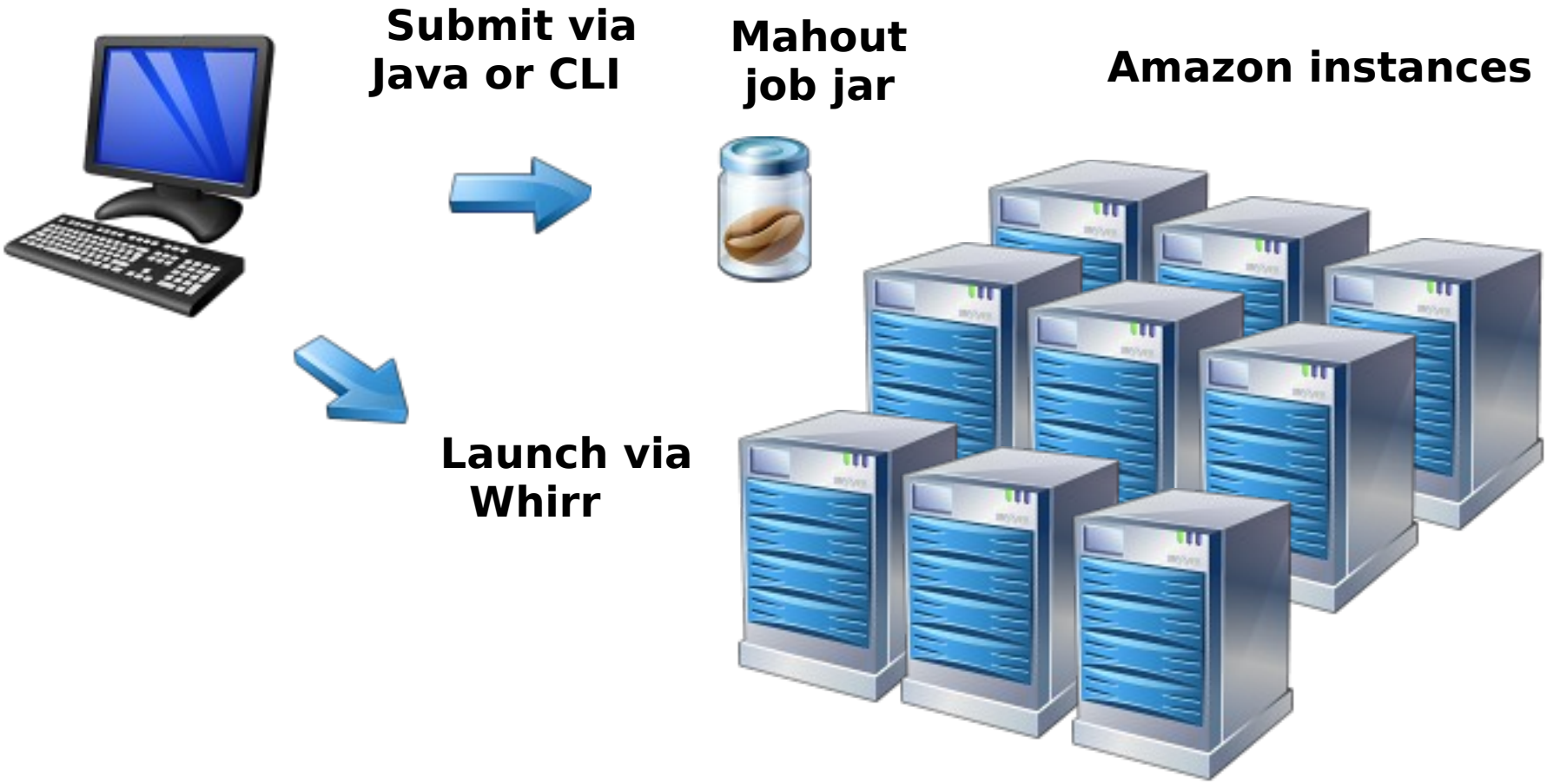
Index 

Index

[id,title,content,clusterId,clusterLabel]

View with web app & Solr

Running the job on



Apache Whirr

- Tool for launching clusters

- Whirr property file

whirr.provider=aws-ec2

whirr.instance-templates=
1 hadoop-jobtracker+hadoop-namenode,
10 hadoop-datanode+hadoop-tasktracker

whirr.identity=topsecret

Apache Whirr

Launch!

```
$ whirr launch-cluster \
        --config so-cluster.properties
$ export HADOOP_CONF_DIR=.whirr/so-cluster
```

Run!

```
$ hadoop fs -put posts.xml input
```

```
$ mahout seq2sparse ...
```

Whirr 

Launch!

Configuration prop = new PropertiesConfiguration(whirrConfigFile);

ClusterSpec spec = new ClusterSpec(prop);

Service service = new Service();

Cluster cluster = service.launchCluster(clusterSpec);

Whirr

Submit!

```
Configuration configuration = new Configuration();

configuration.addResource(
new Path(
  “/home/frank/.whirr/so/hadoop-site.xml"
));

Job job = new Job(conf);

job.submit();
```

Demo Time!

Conclusions

References

Mahout mailinglist

http://blog.jteam.nl/author/frank

Q&A