Machine learning in the cloud with Mahout and Whirr

Frank Scholten @ Berlin Buzzwords 2012

Committer Apache Whirr
Software Developer Orange11
@Frank_Scholten

frankscholten@apache.org
frank.scholten@orange11.nl

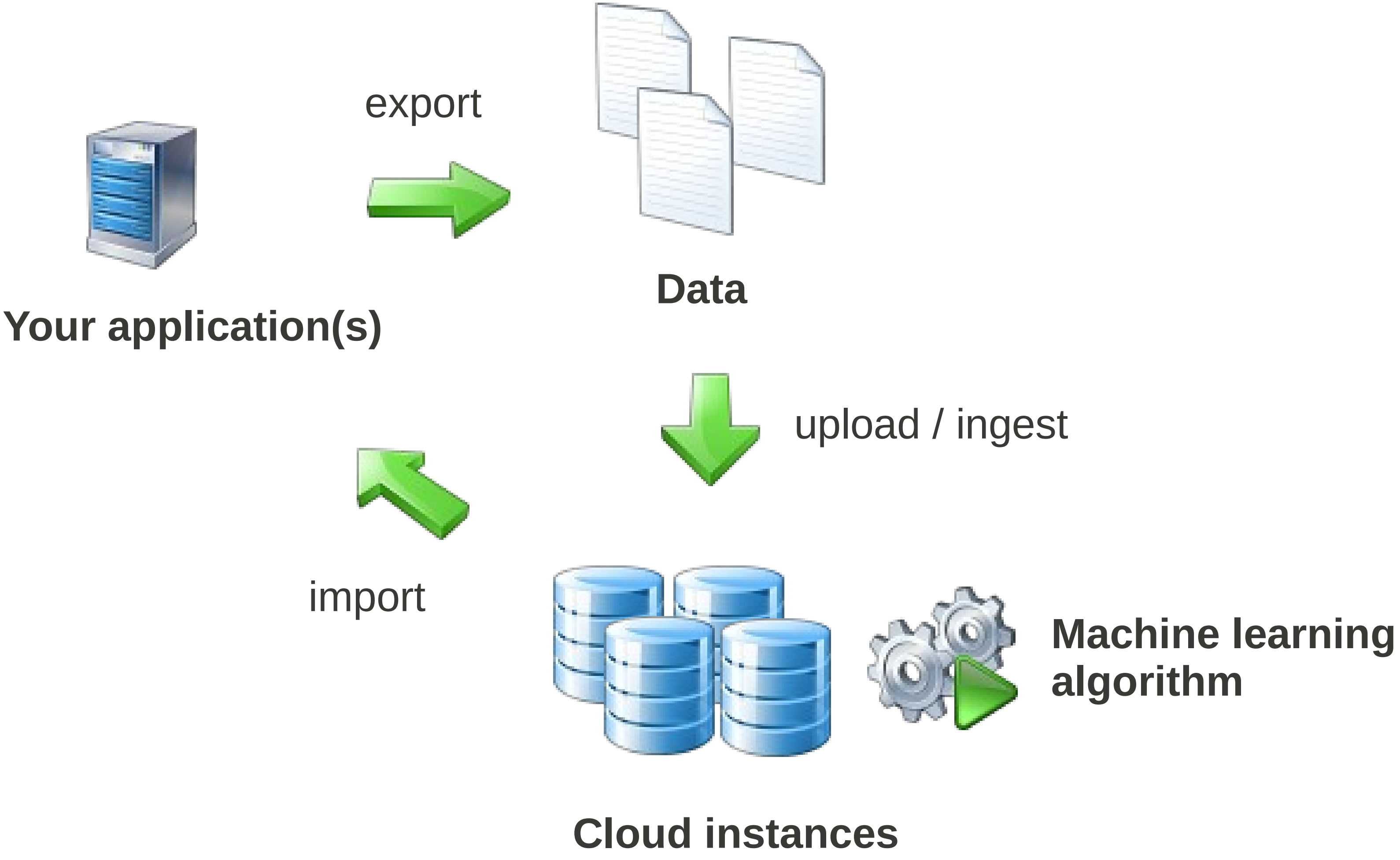
Agenda

- Machine learning and the cloud
- Apache Mahout
- Apache Whirr
- Zoom-in on Whirr's Mahout service

Machine learning – Data + Algorithm + Infrastructure

- Data
  - Events – User clicks, log entries
  - Text – Blogs, Tweets, articles, e-mail
- Algorithm
  - Use a clustering algorithm to group related documents
  - Recommend products to users with collaborative filtering algorithm
- Infrastructure
  - Private datacenter with your own hardware
  - Cloud instances
-

Machine learning in the cloud – Workflow



export

Data

Your application(s)

upload / ingest

import

Machine learning algorithm

Cloud instances

What is Apache Mahout?

- Scalable machine learning on Hadoop **(0.20.204.0)**
  - User & Item-based recommendation
  - SGD classification
  - Clustering algorithms

- Version 0.7 to be released any time now
  - Cleanup and refactoring release

Running Mahout

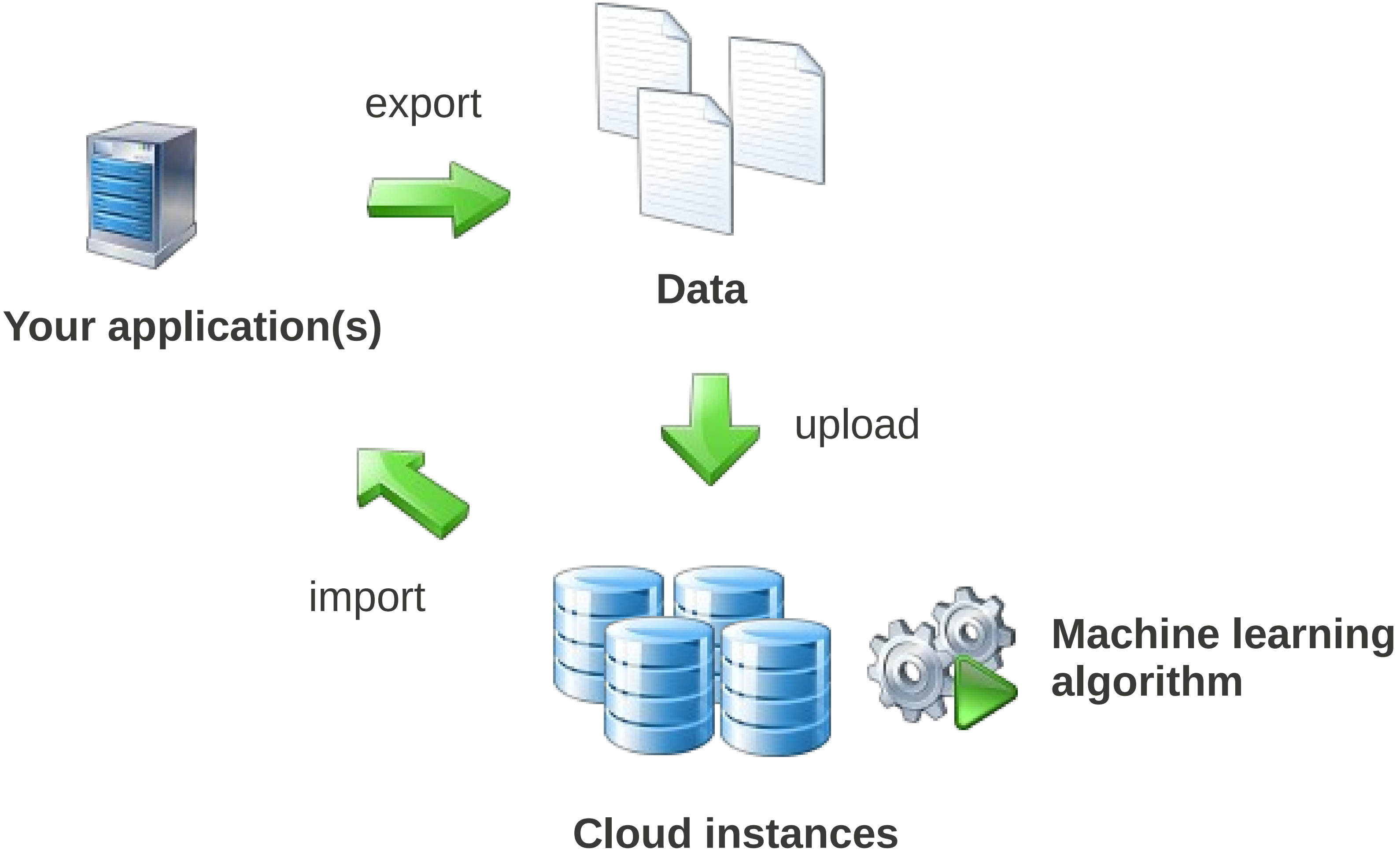
- bin/mahout script

```
$ mahout kmeans \
--input input   \
--output output \
--numClusters 10 ...
```

- Java driver class

```
String[] args = {“--input”, “input”, ... };
ToolRunner.run(conf, new KmeansDriver(), args);
```

Machine learning in the cloud – Workflow



Your application(s)

export

Data

upload

import

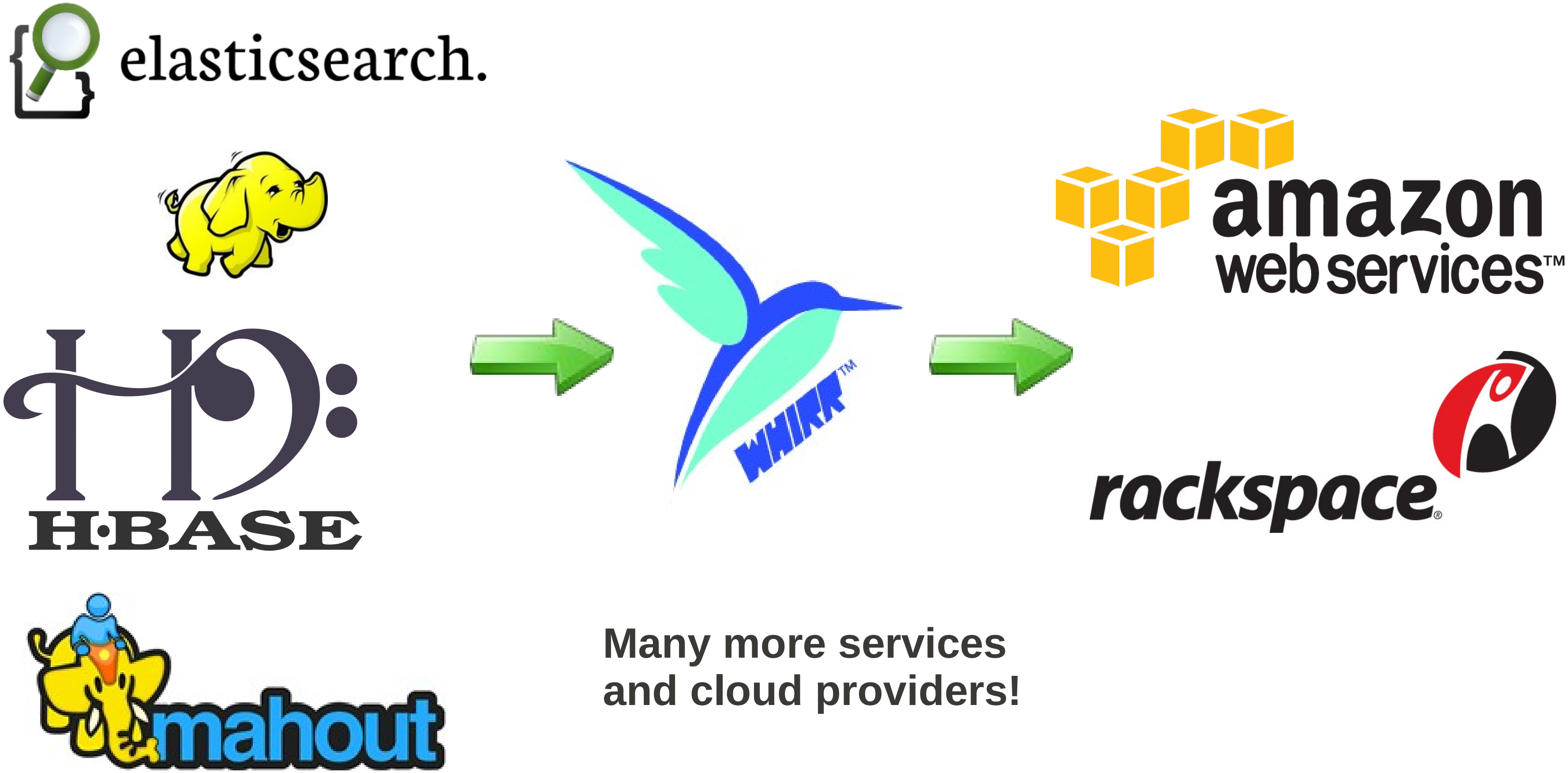
Cloud instances

Machine learning algorithm

What is Apache Whirr?

Services

Cloud providers

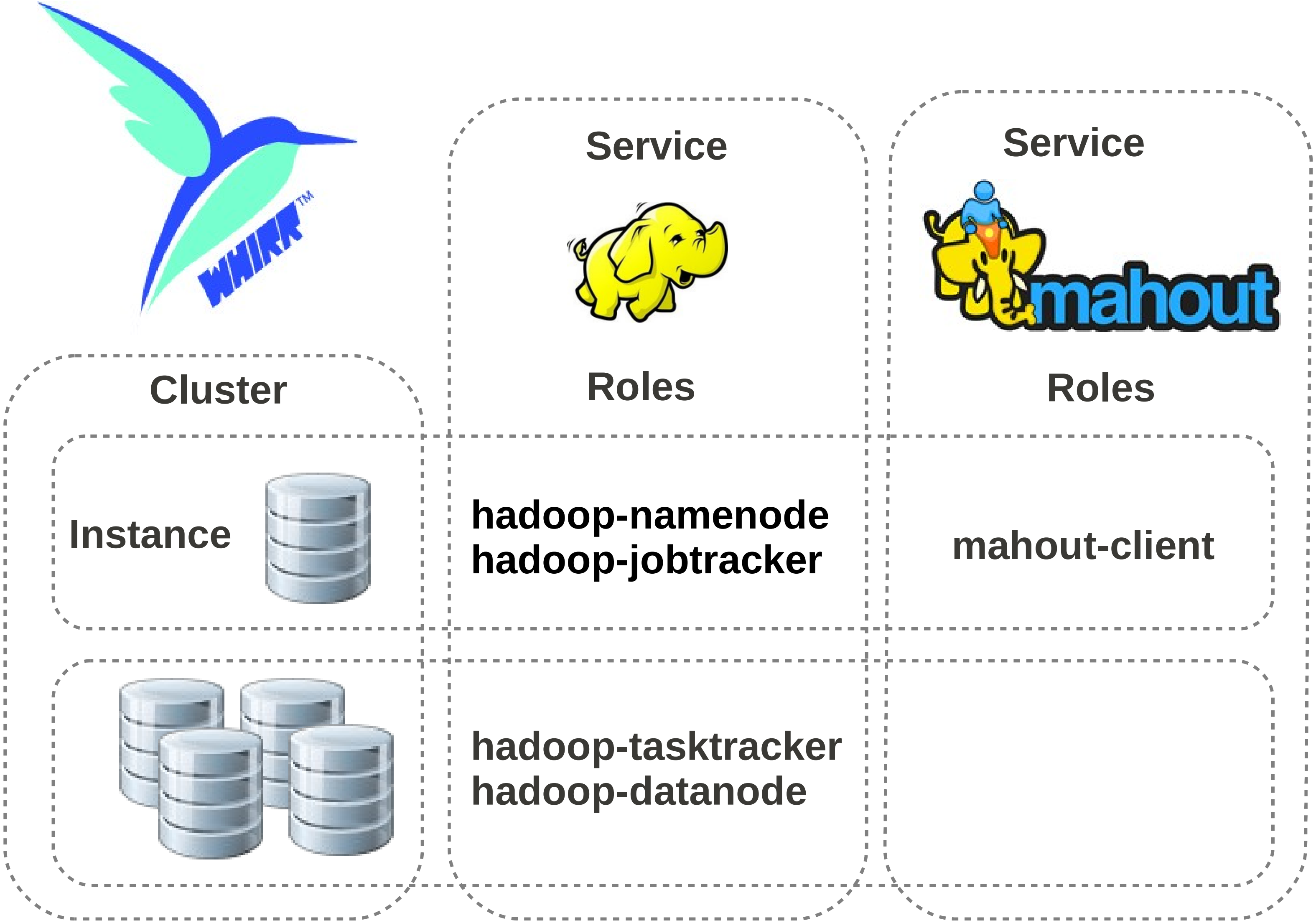


Many more services and cloud providers!

Whirr details

- Started in 2010 from scripts to setup Hadoop on EC2
- Apache Top Level Project since September 2011
- Uses **jClouds** under the hood **(1.4.0)**
- Version 0.7.2 coming soon – Apache Solr, Apache Pig

Whirr concepts



Cluster

Instance

Service

Roles

hadoop-namenode
hadoop-jobtracker

hadoop-tasktracker
hadoop-datanode

Service

Roles

mahout-client

Whirr's Mahout service

- Installs Mahout binary distribution tarball
- Single role **`mahout-client`**
- 2 properties
  - **`whirr.mahout.tarball.url`**
  - **`whirr.mahout.version`**
- Use **`file://`** urls to deploy local mahout distros! **(WHIRR-220)**

Starting a Mahout cluster with `bin/whirr`

- Create a cluster specification property file

```
whirr.instance-templates=

1  hadoop-jobtracker+hadoop-namenode+mahout-client,

10 hadoop-datanode+hadoop-tasktracker
```

- Override Hadoop configuration

```
hadoop-mapreduce.mapred.child.java.opts=-Xmx1024m
```

- `$ bin/whirr launch-cluster --config mahout-cluster.properties`

Using the Mahout cluster

- Whirr writes configuration locally under **`~/.whirr/<cluster-name>`** directory
  - **`instances`** – instance IP addresses and roles
  - **`hadoop-core.xml`** – Same config as on the cluster
- Login via authorized public key
  - **`$ ssh <public-ip>`**
- Copy data to the cluster
  - **`$ hadoop distcp <src> <dst>`**
- Run your Mahout job!
  - **`$ mahout seqdirectory ... ; mahout seq2sparse ... ; mahout kmeans ... ;`**

Building a Whirr service

- Java + Bash

- Extend a **`ClusterActionHandlerSupport`** per role
  - Configure & boostrap phase
  - Before and after hooks
  - **`addStatement(event,call(script,”-u”,tarball));`**

- Provided by Whirr **`install_tarball install_openjdk ...`**

DIY Mahout example

- Mahout example - ASF public mail archive **(WHIRR-594)**

- Scripts + config
  - Launch a Mahout/Hadoop cluster
  - Attaches and mounts EBS with 200 GB ASF e-mails **(@gsingers)**

- You
  - **`$ hadoop distcp`** the data into the cluster
  - Run mahout!

Whirr tips and tricks

- Environment variables
  - `WHIRR_CREDENTIAL WHIRR_IDENTITY WHIRR_PROVIDER`
- Free-style provisioning
  - `whirr.instance-templates=1 noop`
  - `$ bin/whirr run-script foo.sh --role noop`
- Starting clusters via Java API
  - `ClusterSpec clusterSpec = ClusterSpec.withTemporaryKeys(config);`
  - `ClusterController controller = new ClusterController();`
  - `controller.launchCluster(clusterSpec);`

Conclusions

- Whirr enables automatic deployment of Mahout

- Related topics
  - Cluster performance tuning via Whirr configuration
  - Integrating with your application
  - Visualizing results

- Upcoming Whirr features
  - Adding and removing nodes from running cluster **(WHIRR-214)**
  - Deploy to local Virtualbox cloud **(WHIRR-379)**
  - Optionally control ordering of services **(WHIRR-221)**

References

- Websites and mailinglists
  - **{user|dev}@mahout.apache.org**
  - **{user|dev}@whirr.apache.org**
- ' Mahout in Action' book
- Orange11 & Searchworkings blogs
  - **http://blog.orange11.nl**
  - **http://www.searchworkings.org/blog**

Thank you!

Questions?

orange11™