Migration of audio files using Hadoop

and Taverna

and xcorrSound waveform-compare

Bolette A. Jurik, baj@statsbiblioteket.dk

The State and University Library, Aarhus, Denmark

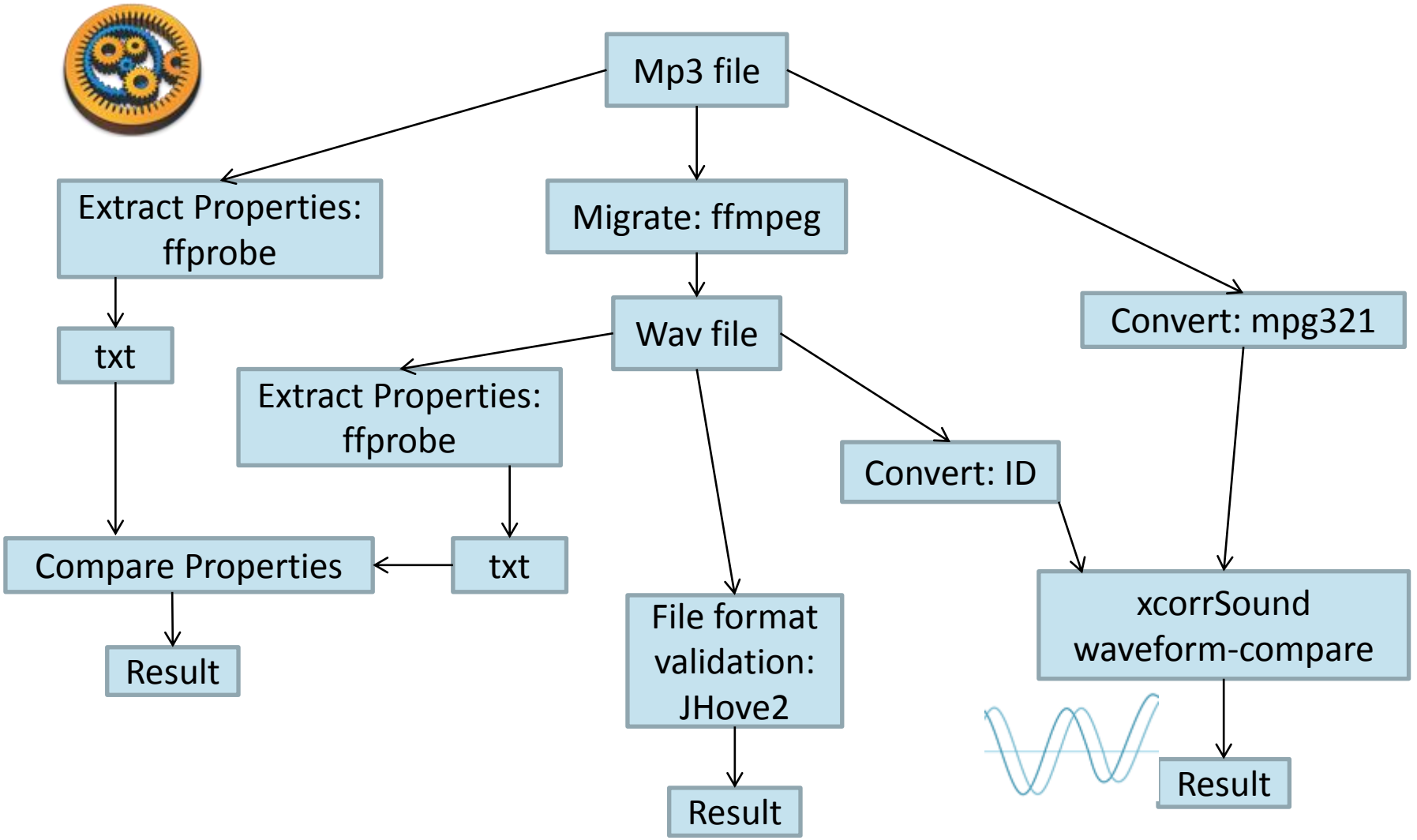
SCAPE Information Day at the Danish State and University Library
Aarhus, Denmark, Wednesday, June 25th, 2014

Background: User Story

- wiki.opf-labs.org/display/SP/Large+Scale+Audio+Migration
- As owner of a large mp3 collection,
  - we need a digital preservation system that can migrate large numbers of mp3s to wav files and
  - ensure that the migration is a good and complete copy of the original.
- Note: at SB we have a 20 TB collection of Danish Radio broadcast mp3s. We used this in a Plato case study in November 2012. Plato recommended the “do nothing” solution…

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

mp3 to wav migration and QA Taverna Workflow



This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Evaluation *mp3 to wav migration and QA* 2012

Metric	Baseline definition	Baseline value	Goal	Evaluation 1 (date)
Number Of Objects Per Hour	Performance efficiency - Capacity / Time behaviour	10 (test 2nd-16th October 2012)	1000	18 (9th-13th November 2012)
QA False Different Percent	Functional suitability - Correctness	5% (test 2nd-16th October 2012)	.1%	0.412 % (5th-9th November 2012)

- a baseline value of 10 objects per hour means that we process 1.18Gb per hour and we produce 28Gb per hour (+ some property and log files).
- The collection that we are targeting is 20 TB or 175.000 files. With baseline value we would be able to process this collection in a little over 2 years. The goal value is set so we would be able to process the collection in a week.
- Evaluation 1 (9th-13th November 2012). Simple parallelisation on one machine. Processed 1756 files (~ 200GB) in a little over 4 days.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

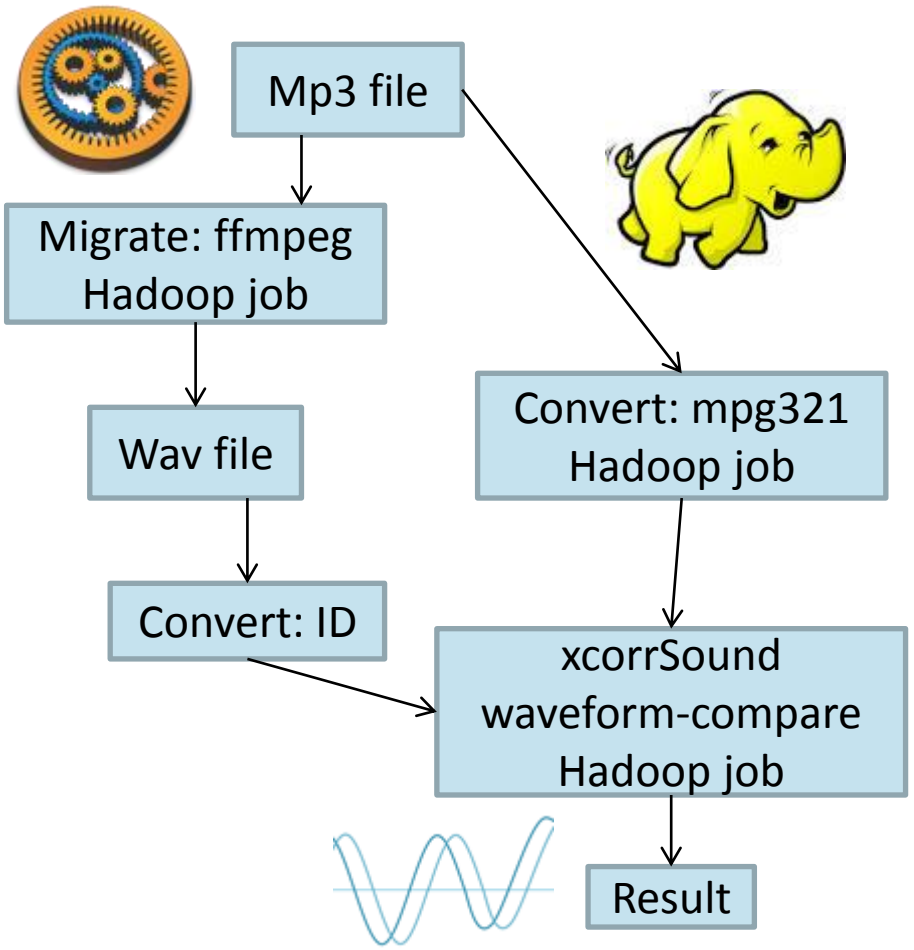
Going large-scale

- github.com/statsbiblioteket/scape-audio-qa

Input
- List of NFS file paths on HDFS (txt file)
- Mp3 files on NFS

Output
- Wav files on NFS
- Log files etc. on HDFS

Tools needed on cluster
- Iapetus: taverna-commandline-2.4.0/executeworkflow.sh
- Nodes on cluster: ffmpeg, mpg321, waveform-compare



This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Demo *mp3 to wav migration and QA* using Hadoop

- Start workflow on iapetus 
- Look at input
- Look at Cloudera Manager: http://cressida:7180/cmf/
- Look at output
- (look at input again)

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Evaluations so far (1)

Small Experiments April 2014
All run on a file list of 58 files (7.2Gb in total).

max split size	duration	launched maps for ffmpeg Hadoop job	Number Of Objects Per Hour
1024	37m, 59s	3	91
512	24m, 2s	6	145
256	18m, 18s	12	190
128	17m, 3s	24	205
64	16m, 55s	47	205
32	17m, 30s	93	199

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Evaluations so far (2)

Large Scale Experiments June 2014
max split size 4414, 12 maps for ffmpeg Hadoop job.

#mp3-files	Duration	Number Of Objects Per Hour	Content comparison Failures	wav files
1000 (~100GB)	4h, 33m	220	63 (6.3%)	~3.1TB
2000 (~200GB)	8h, 56m	224	174 (8.7%)	~6.2TB
2999 (~300GB)	13h, 29m	222	226 (~7.5%)	~9.3TB
3999 (~400GB)	17h, 56m	223	368 (~9.2%)	~12.4TB
4998 (~.5TB)	22h, 24m	223	435 (~8.7%)	~15.5TB

- Is Number of Objects Per Hour acceptable?
- Is Number of Content Comparison Failures acceptable?

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Conclusion

- Writing a challenge for SB Hadoop cluster!
- Performance

2014 JUNE

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

www.free-printable-calendar.net

- Content Comparison Failures and mp3 file quality

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Thanks

- Links
  - User Story: wiki.opf-labs.org/display/SP/Large+Scale+Audio+Migration
  - xcorrSound: openplanets.github.io/scape-xcorrsound/
  - Old Taverna Workflow www.myexperiment.org/workflows/3292.html
  - Experiment Source Code: github.com/statsbiblioteket/scape-audio-qa
  - Danish Radio broadcast mp3 collection http://wiki.opf-labs.org/display/SP/Danish+Radio+broadcasts%2C+mp3
  - 2012 evaluation: wiki.opf-labs.org/display/SP/EVAL-LSDR6-1
  - 2014 evaluation: http://wiki.opf-labs.org/display/SP/Evaluation+-+SB+Experiment+mp3+to+wav+Migration+and+QA+on+Hadoop+Cluster (work in progress)

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).