Large Scale Processing with Hadoop

William Palmer

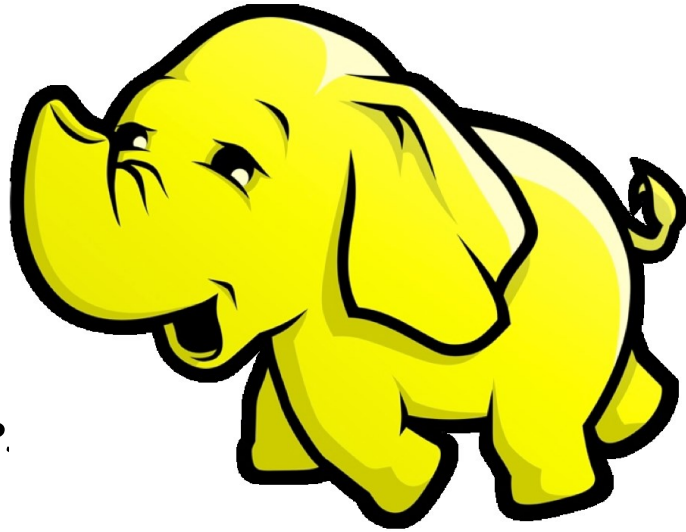
Some slides courtesy of Per Møldrup-Dalum (State and University Library, Denmark) and Sven Schlarb (Austrian National Library)

SCAPE Information Day

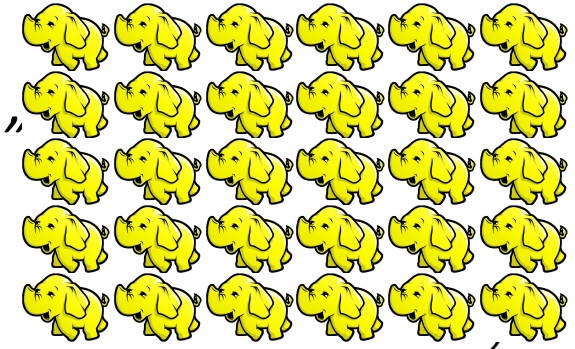
British Library, UK, 14th July 2014

Large Scale Processing Methodologies

- Traditional
  - One central large processor capability
  - One+ central storage instance
  - Data stored away from processor
  - Paradigm: *"Move the data to the proce*



- Hadoop
  - Many smaller commodity computers/CPUs
  - Storage capacity in all computers, federated together
  - Easily expandable
  - Paradigm: *"Move the processor to the data"*



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).

Example

- The New York Times + Hadoop on Amazon Web Services
- 11 million articles (1851-1980) that need to be converted to PDF
- 4TB TIFF data
- 24 hours wall time to complete the migration
- Cost: $240 (not including bandwidth)

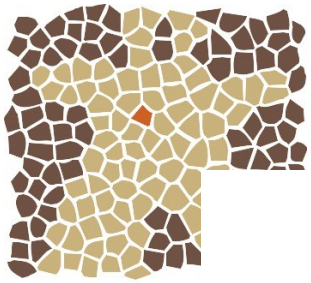
- http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/
- http://cse.unl.edu/~byrav/INFOCOM2011/workshops/papers/p1099-xiao.pdf

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).

Hadoop Ecosystem: The Zoo

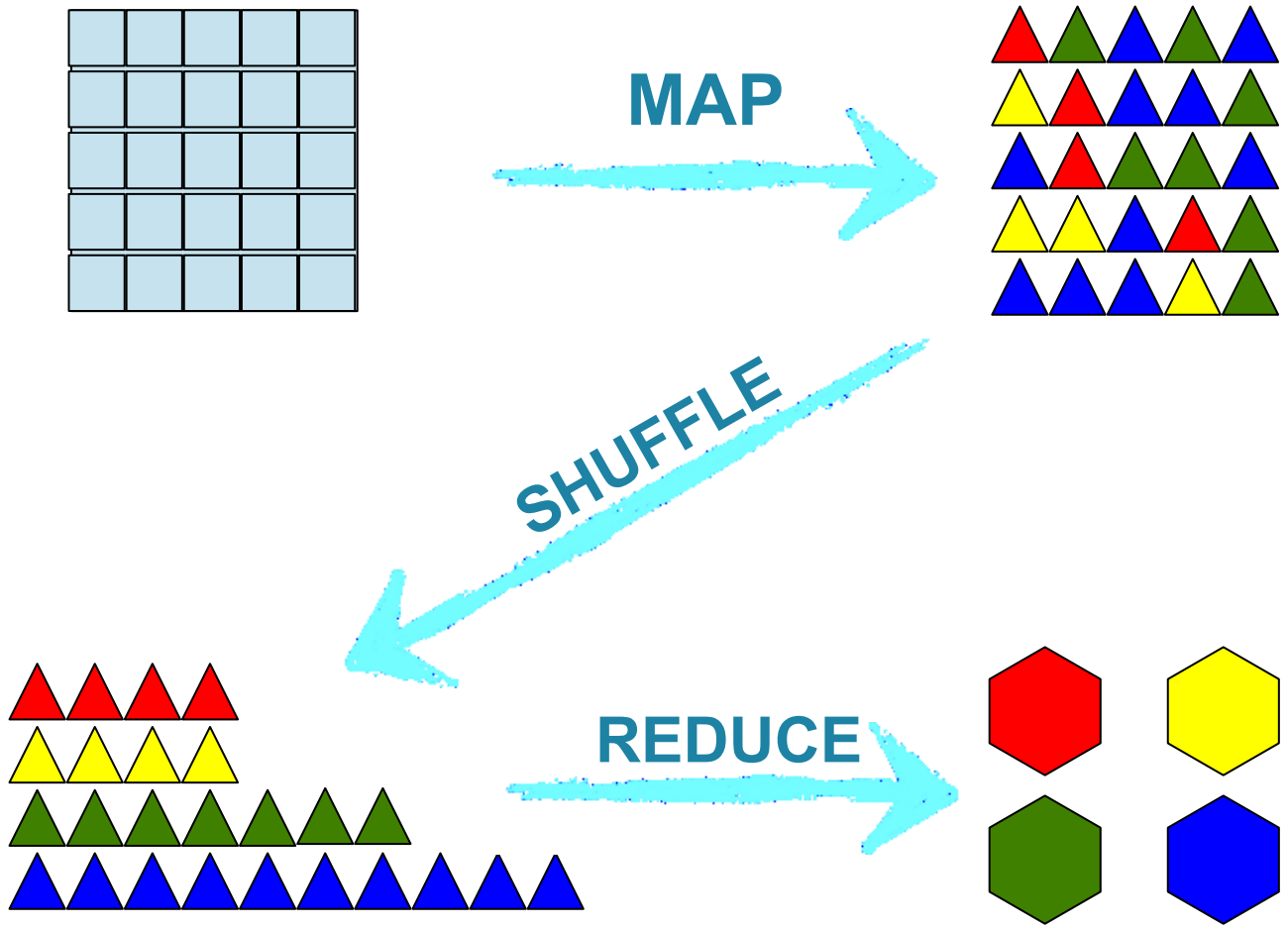
HDFS – data locality

MapReduce



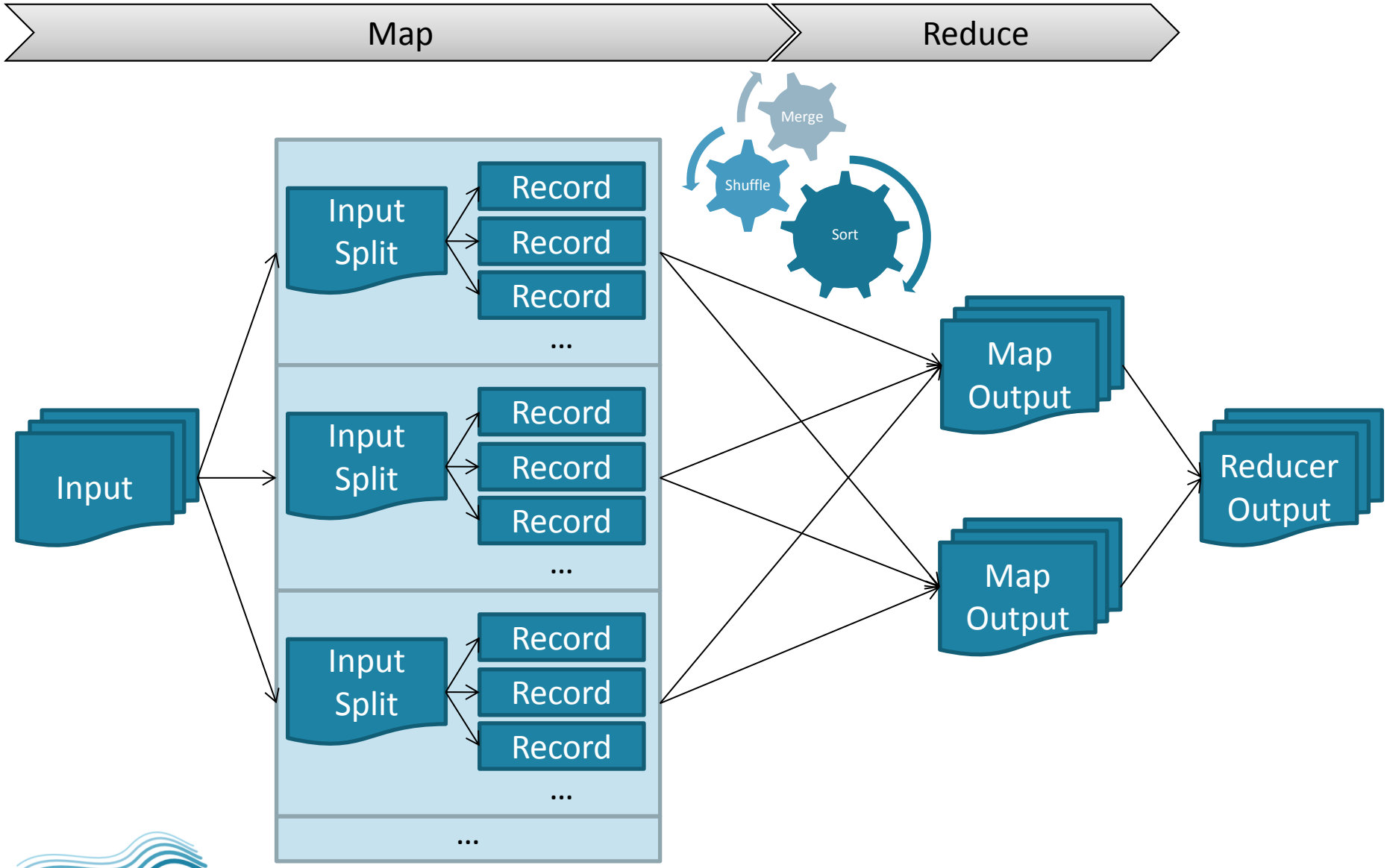
er 270137).

MapReduce



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).

MapReduce in detail



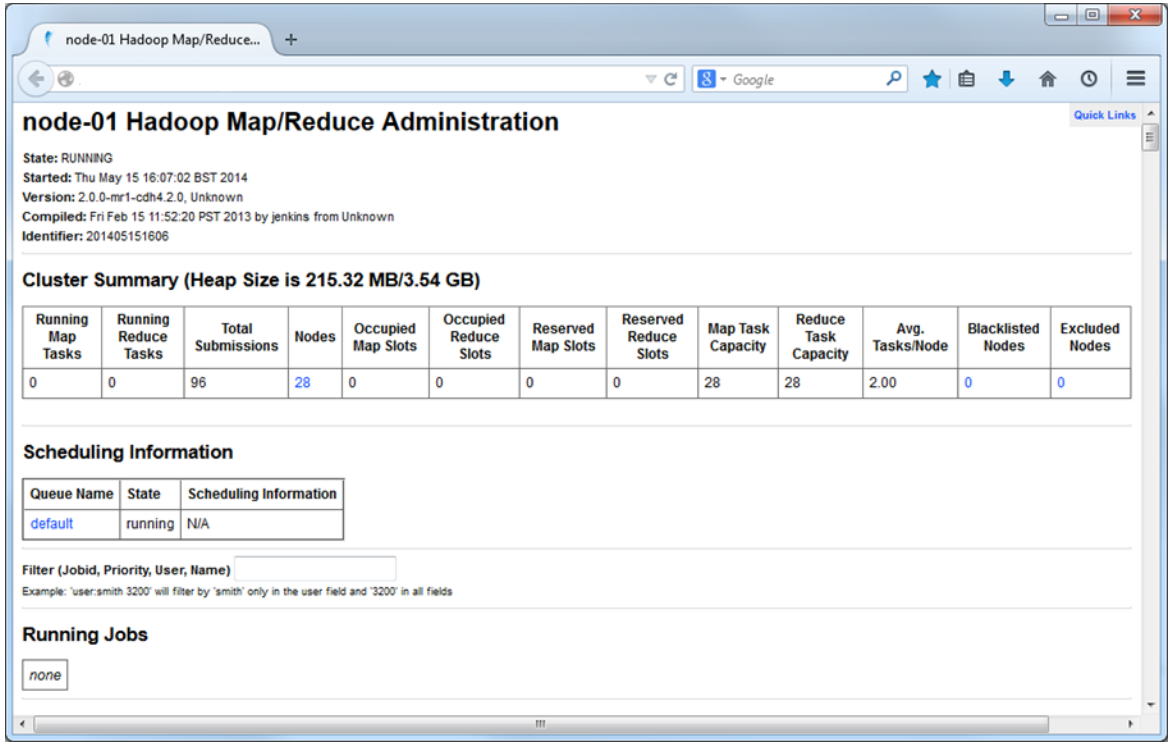
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Hadoop In Action

- Designed for processing text
- Capacity can be reduced/expanded
- Comes with HDFS filesystem, with federation and redundancy (three copies of data by default)
  - Using commodity hardware node failures are expected
  - A node being down should not affect the cluster
- Data locality is considered when distributing computation, processing data where it is stored, reducing the need to transfer it
- Very large community and ecosystem

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).

(Obligatory) Hadoop Screenshots




```
14/02/13 11:22:33 INFO gzchecker.GZChecker: Loading paths...
14/02/13 11:22:36 INFO gzchecker.GZChecker: Setting paths...
14/02/13 11:22:37 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. …
14/02/13 11:22:39 INFO mapred.FileInputFormat: Total input paths to process : 1
14/02/13 11:22:40 INFO mapred.JobClient: Running job: job_201401131502_0058
14/02/13 11:22:41 INFO mapred.JobClient:  map 0% reduce 0%
…
```

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).

Hadoop In Action

- We are using Hadoop/MapReduce for parallelisation
  - Non standard use case
  - As a parallelisation method costs are associated …
  - … but get a lot of well supported features for free
    - HDFS
    - Administration
    - Support
  - Once a MapReduce program is developed scalability just happens
    - Can theoretically prototype on a Raspberry Pi and run on a 3000 node super cluster

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).

Hadoop In Action

- Do I have to copy data to HDFS for processing?
  - 1TB of data took 8 hours to copy from NAS to HDFS
  - Image format migration (TIFF-JP2) took ~57hours
  - … still got to get the data back to the NAS
- What if I don't?
  - Same image format migration code accessing/posting data directly from/to Repository took ~58hours
  - No copying data before/after
  - More efficient as processing time is greater per file
  - Won't necessarily hold for different preservation actions (see: "small files problem")

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).

Hadoop at The British Library

- Two Hadoop clusters:
  - Digital Preservation Team Cluster
    - Virtualised hardware
    - 1 management node, 1 master node
    - 28 worker nodes (1 core/1 CPU, 6GB RAM each)
    - 14TB raw storage, 5TB useable @ replication of 3
    - Cloudera Hadoop (CDH4)
    - For testing/R&D
  - Web Archiving Team Cluster
    - Physical hardware
    - 80 nodes (8 cores/2CPUs, 16GB RAM)
    - 700TB raw storage, 233TB useable @ replication of 3
    - Cloudera Hadoop (CDH3)
    - In production use

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).

SCAPE Workflow Results

- TIFF->JP2 migration with QA
  - Single node @ 26 files/hour (with OpenJPEG)
  - 28 nodes @ 735 files/hour (with OpenJPEG)
    - 2409 files/hour with Kakadu
- Detecting DRM in PDF files
  - 28 nodes @ 51869 files/hour
- Identifying web content
  - 5.3million files/hour

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).

Other Large Scale Execution Platforms

- SCAPE tools are treated as individual components and should be reusable on other large scale execution platforms (all tools described today are, at least)
- British Library Digital Library System (DLS) has a bespoke workflow execution system where SCAPE tools have been integrated
- Other platforms: GNU Parallel …
- Tools can be integrated with your own systems

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).