Cross match of 10 billions photometric records using Hadoop

Yuji Shirasaki (JVO NAOJ)

Digital Universe @ JVO

- A big table :
  - 20 billions of photometric data from various survey
    - SDSS, TWOMASS, USNO-b1.0,GSC2.3,Rosat, UKIDSS, SDS(Subaru Deep Survey), VVDS (VLT), GDDS (Gemini), RXTE, GOODS, DEEP2 …
  - ONLY coordinate + brightness + band ID (+ link to original resource)
- Currently provides small region search
- Plan to extend to all sky search with color condition
  - Require Cross ID among all the photometric records
  - Distributed data processing → Hadoop

Hadoop

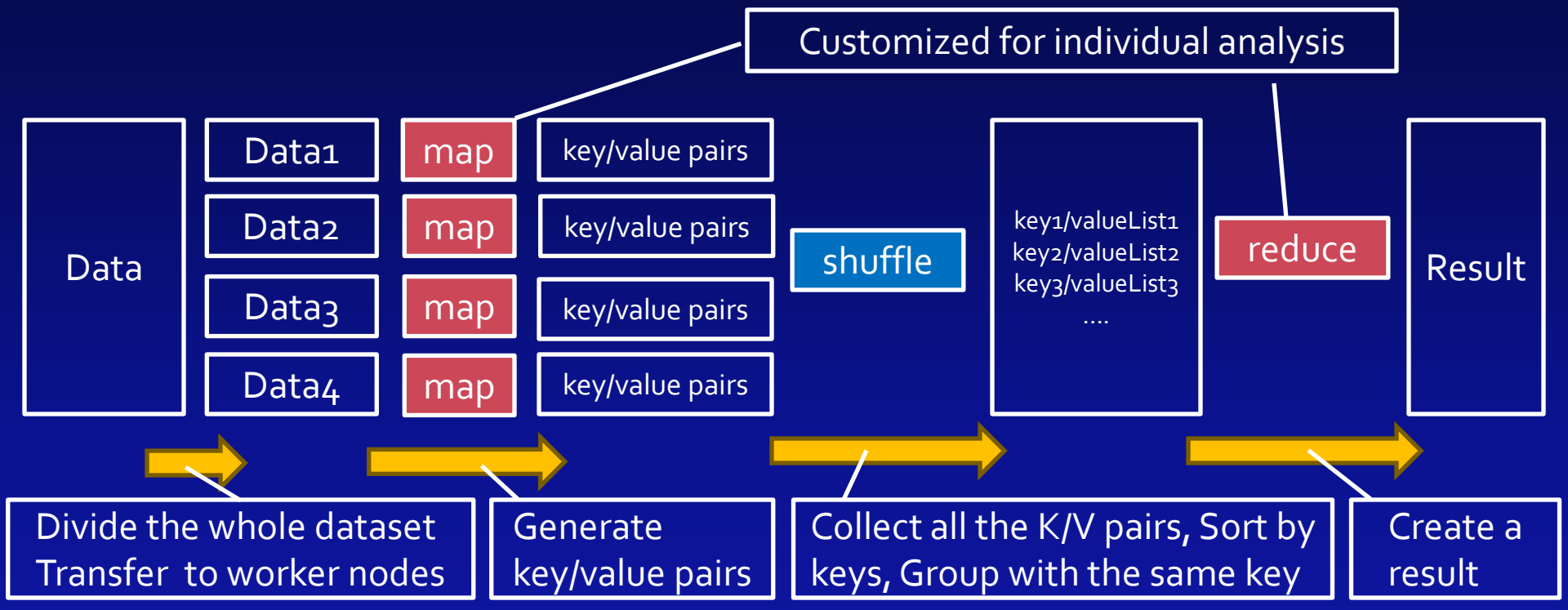
- What is ?
  - Java software framework for distributed data processing
  - data is processed where the data resides
  - One of the apache top level projects.
  - http://hadoop.apache.org/
- Applications
  - Facebook : log analysis, machine learning
  - The New York Times: generate PDF of 11 million articles from 1851-1980
  - Yahoo: genrate ranking
  - Many other companies are using Hadoop in their system.

Components of Hadoop

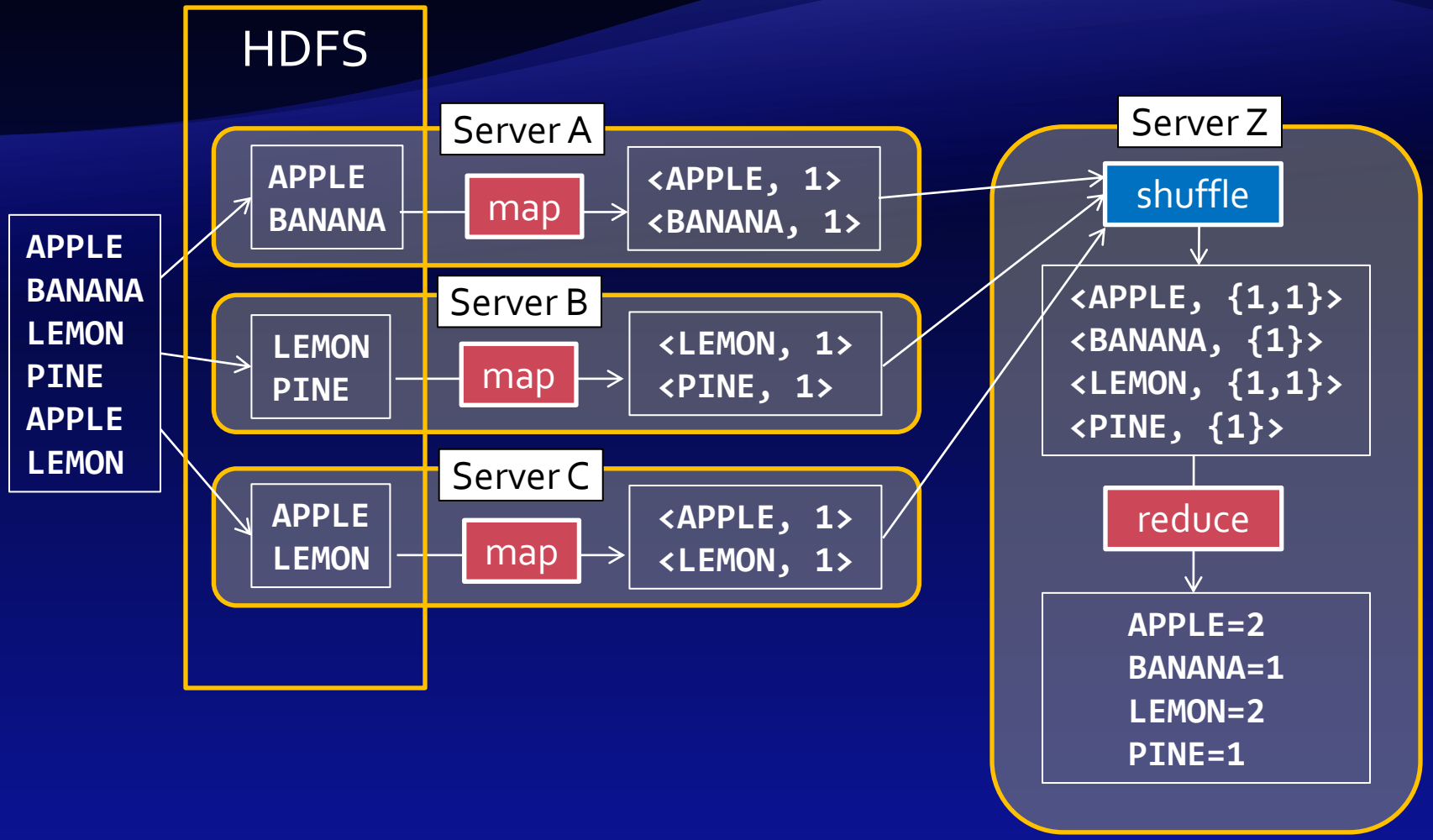
- Hadoop Distributed File System (HDFS)
  - Combine storages attached to separate servers
  - A file is divided into blocks with the same size, stored over multiple servers, replicated on three (default) servers.
  - One namenode server & datanode server(s).
- Job Tracker & Task Tracker
  - Execute a job following the MapReduce procedure.
  - User submits a Job to the Job Tracker server
  - Job Tracker breaks down the single Job into multiple Tasks, and submit the tasks to Task Tracker server.
  - Task is scheduled so that it is executed on the datanode which stores the data to be processed.

MapReduce

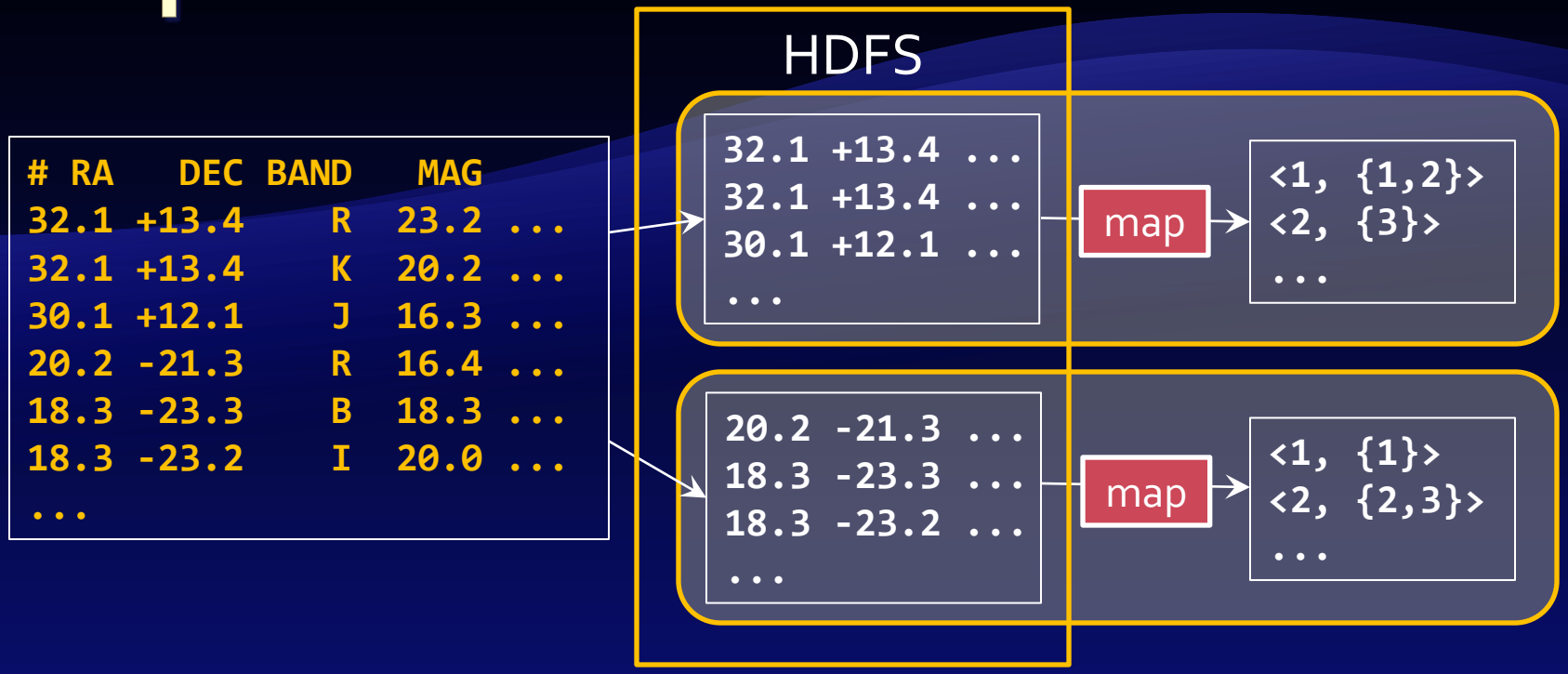
- A programming model for processing large data sets
- MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean and Sanjay Ghemawat (Google Inc.)
- Most of the computations at Google can be represented as a sequence of map, shuffle and reduce operations



An example: word count



MapReduce for Cross Match



- ✓ Divide the whole dataset into subsets based on a region of sky.
- ✓ The Map function processes whole of the input file to produce cross match result (list of matched record ids)
- ✓ The Reduce function is not executed, since each subset is independent each other.

Hadoop Installation

- Download hadoop-common package and unpack at arbitrary directory
- Java is required
- Configure
  - core-site.xml : default filesystem
  - hdfs-site.xml : data dir for namenode & datanode, block size
  - mapred-site.xml : job tracker node, data (system) dir for mapreduce, max # of map task
  - copy hadoopdir to all nodes of the cluster, create data directories
- HDFS format
- start-all.sh

Implementation on Hadoop

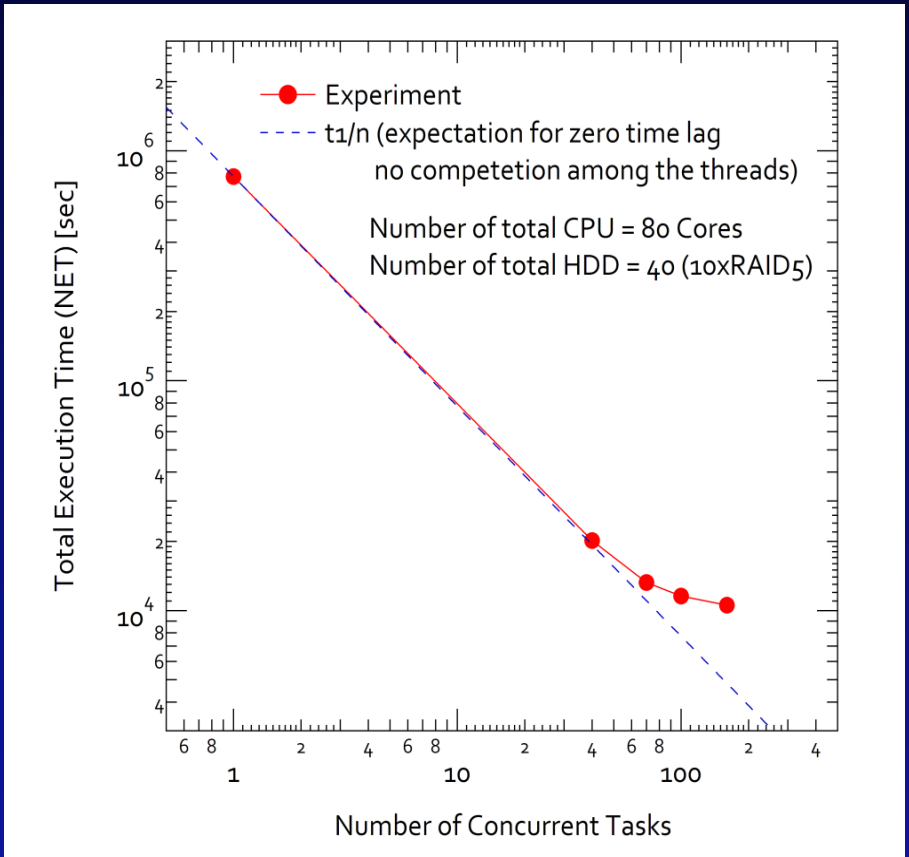
- Three Java classes
  - ✓ MapperForXMach.java
    - override map method of org.apache.mapreduce.Mapper
    - Execute cross match for whole input file and write the result
  - ✓ WholeFileInputFormat.java
    - override createRecordReader of org.apache.mapreduce.lib.input.FileInputFormat
    - Read whole file as one record
  - ✓ Xmatcher.java
    - Implements run method of org.apache.hadoop.util.Tool
    - Submit the Cross match job to the Hadoop cluster

Experiment

- 1 billions records (1/20 of whole data)
- Divided into 6112 files. ~3MB/file
- Each file contains records of which pos error circle overlaps with the same region specified with an HTM index (level 6).
- Each file are gziped and copied to HDFS.
- Max number of task executed in parallel
  - 1, 40, 70, 100, 160
- Hardware
  - 10 servers: each has 2x4 core and 4 SATA HDD (RAID5)

Result

- ✓ If executed by a single task
  - 9 dyas for 1G records → 180 days for whole dataset (20G rec.)
- ✓ Parallel execution of 70 tasks (~# of cores, twice # of HDD)
  - 3.7 hours for 1G rec. → 3 days for whole
- ✓ Scaling relation breaks around ~40 tasks
  - Overhead of writing to the local FS.
  - Writing time occupies ~60% of the total.



Conclusion

- Hadoop can be a solution for processing large amount astronomical dataset
- A key feature of Hadoop "do the work in the same server as the data" is adequate for data insensitive processing
- It may be applicable also to data reduction system of large format mosaic camera (Suprime-Cam, Hyper SC …)
- Hadoop is designed to scale to thousands of nodes & petabytes of storage, also to provide a failover mechanism
- Scalable and reliable system in low cost & short time