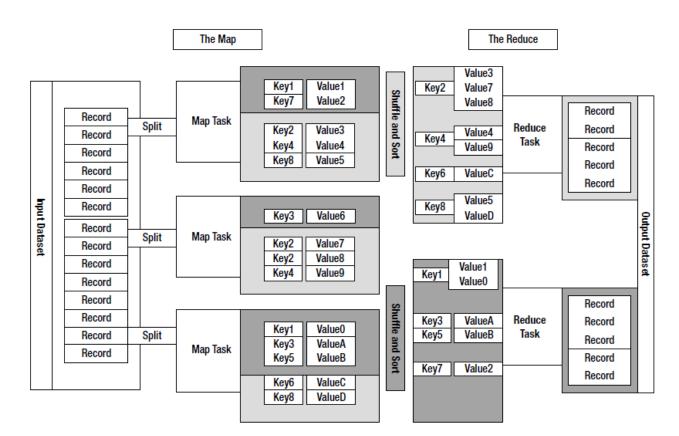
Hadoop in 0 words

Yaroslav Yermilov

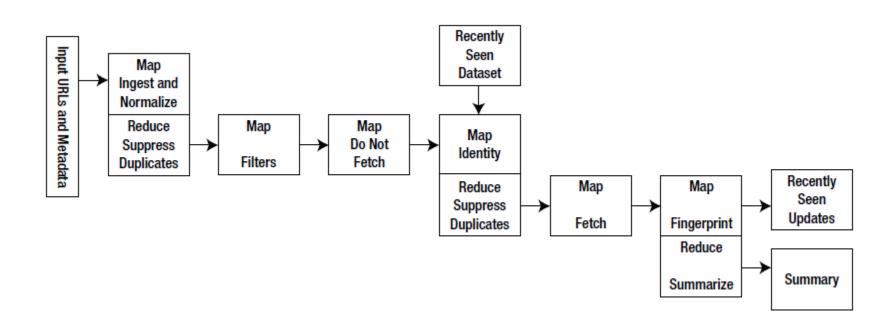
Hadoop MapReduce



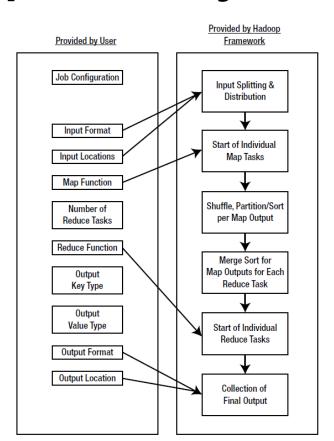
MapReduce model



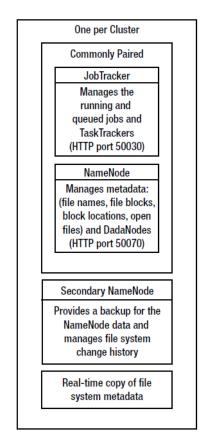
Web crawler example

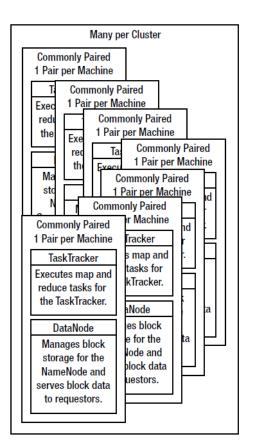


Parts of MapReduce job



Hadoop cluster





How mapper works

Submit job

JobTracker Compute input splits and split locality. Produce task list, 1 task per split

JobTracker
For each open task
execution slot,
schedule a task
from the list

TaskTracker Receive task to execute

TaskTracker Prepare task runtime Create or refresh task local directory. Unpack JARs and DistributedCache Items

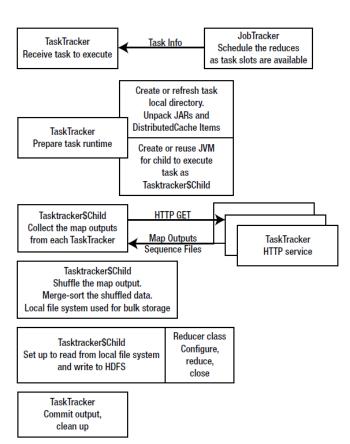
Create or reuse JVM for child to execute task as Tasktracker\$Child

Tasktracker\$Child Set up to read input split from HDFS and write output to local file system Mapper Class Configure, map, close

TaskTracker Cleanup

TaskTracker Serve map output to reduce tasks via HTTP

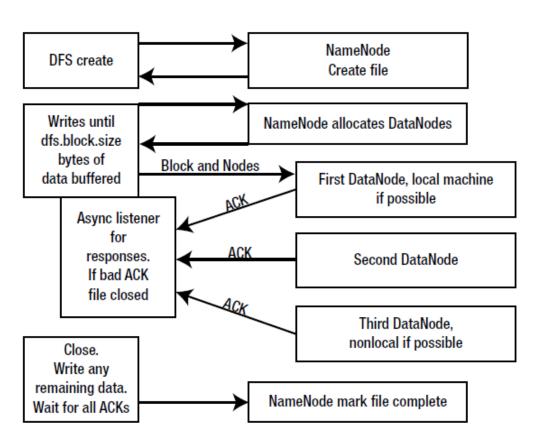
How reducer works



Hadoop HDFS



How HDFS works



Hive



```
SELECT a.year, a.player_id, a.runs FROM batting a
JOIN (SELECT year, max(runs) runs FROM batting GROUP BY year ) b
ON (a.year = b.year AND a.runs = b.runs);
```

Pig



```
batting = load 'Batting.csv' using PigStorage(',');
runs = FOREACH batting GENERATE $0 as playerID, $1 as year, $8 as runs;
grp_data = GROUP runs BY (year);
max_runs = FOREACH grp_data GENERATE group as grp, MAX(runs.runs) as max_runs;
join_max_run = JOIN max_runs BY ($0, max_runs), runs by (year,runs);
join_data = FOREACH join_max_run GENERATE $0 as year, $2 as playerID, $1 as runs;
dump_join_data;
```

HBase



Sqoop



Zookeeper



Oozie



Falcon



Mahout



Avro

