

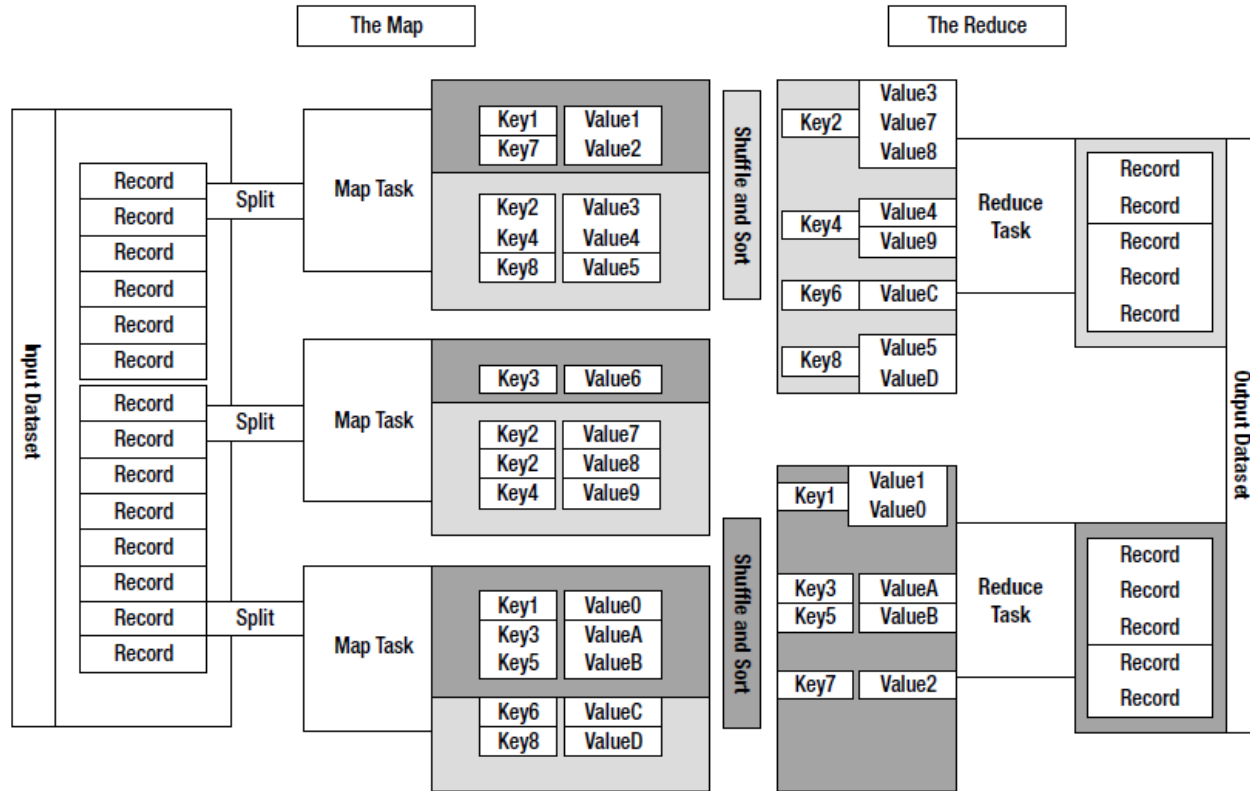
Hadoop in 0 words

Yaroslav Yermilov

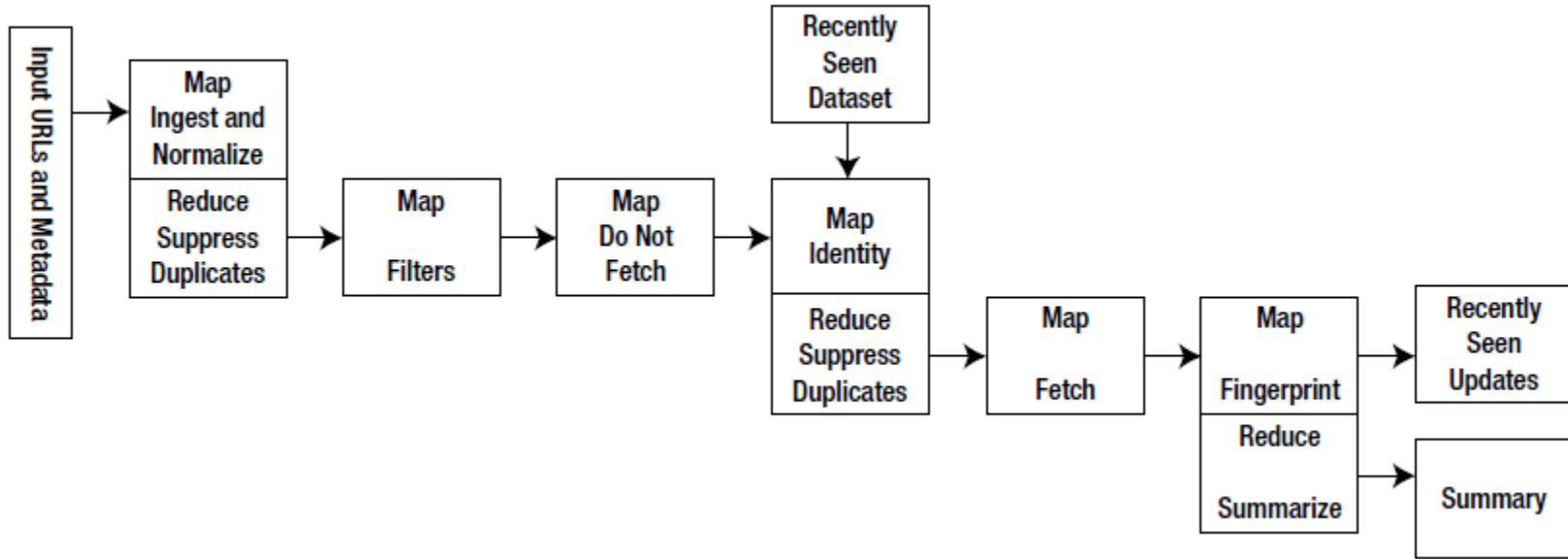
Hadoop MapReduce



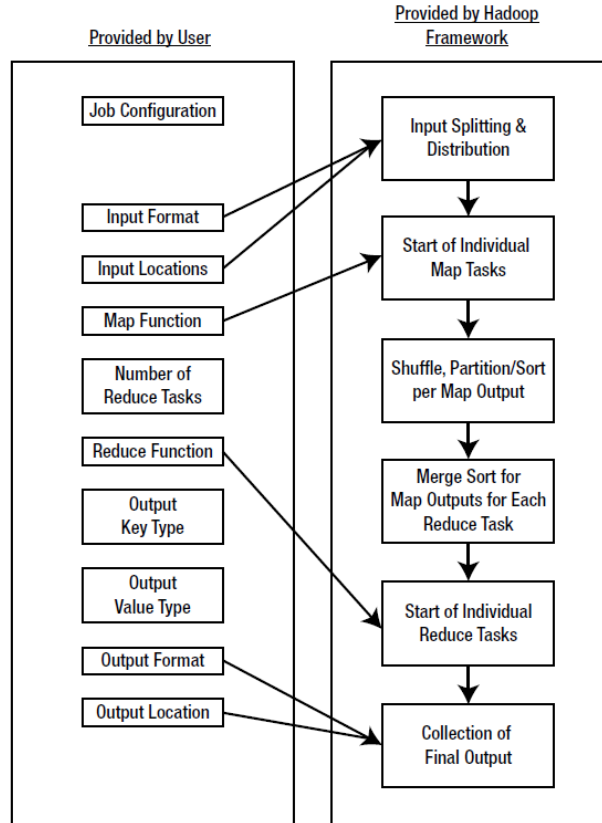
MapReduce model



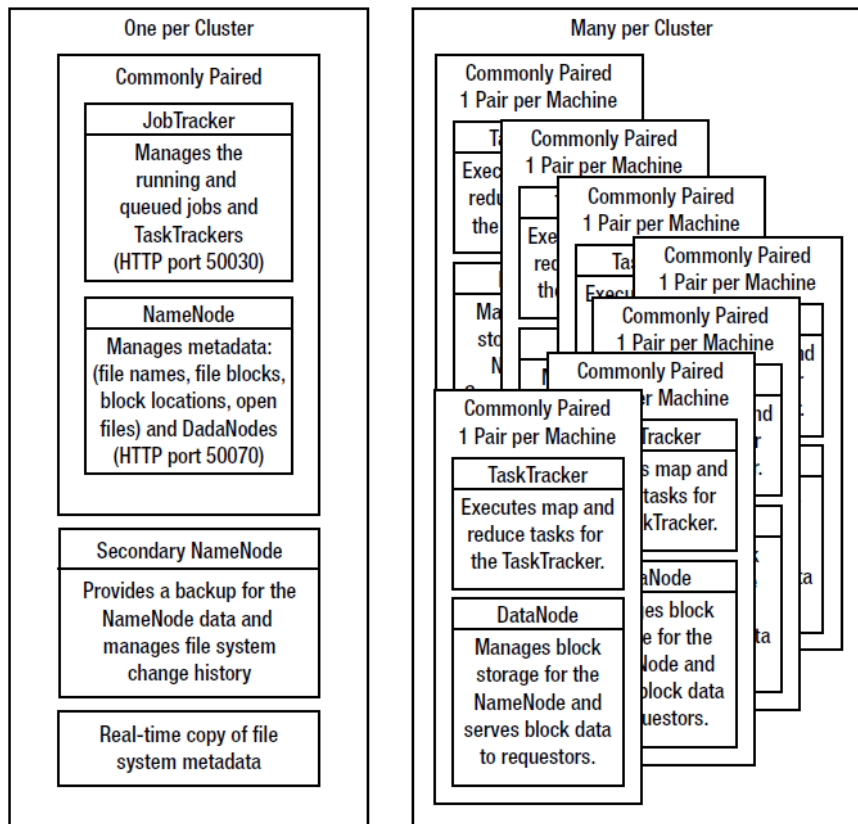
Web crawler example



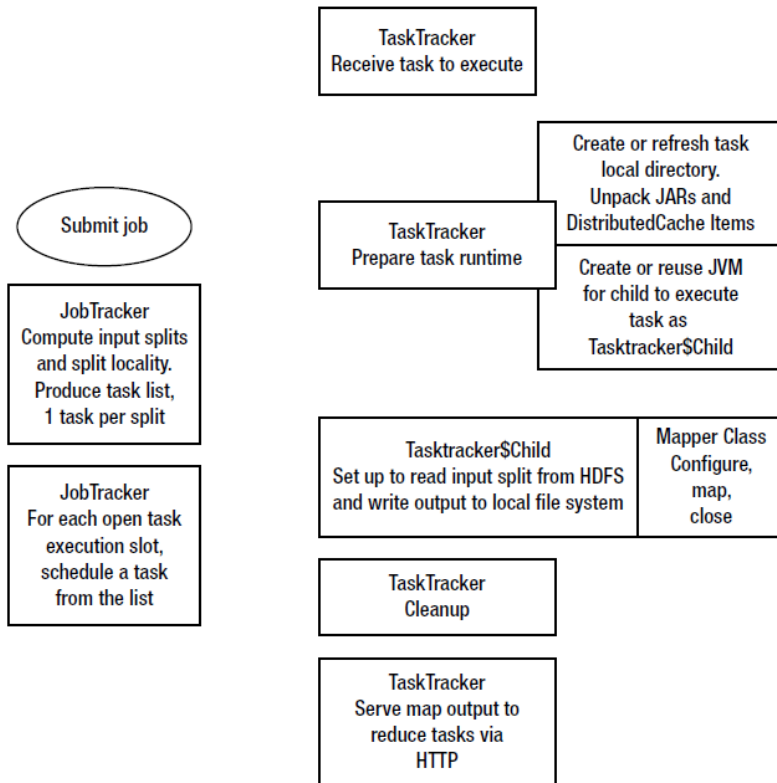
Parts of MapReduce job



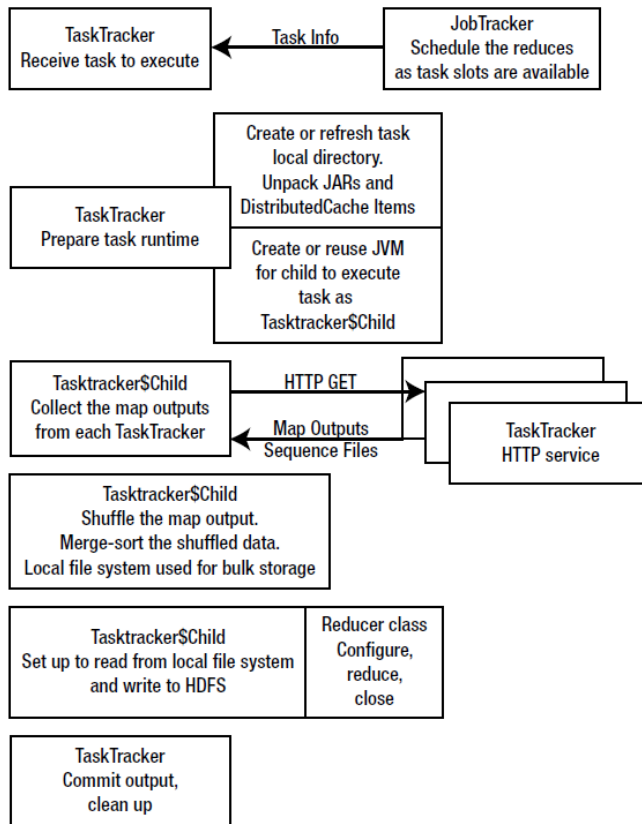
Hadoop cluster



How mapper works



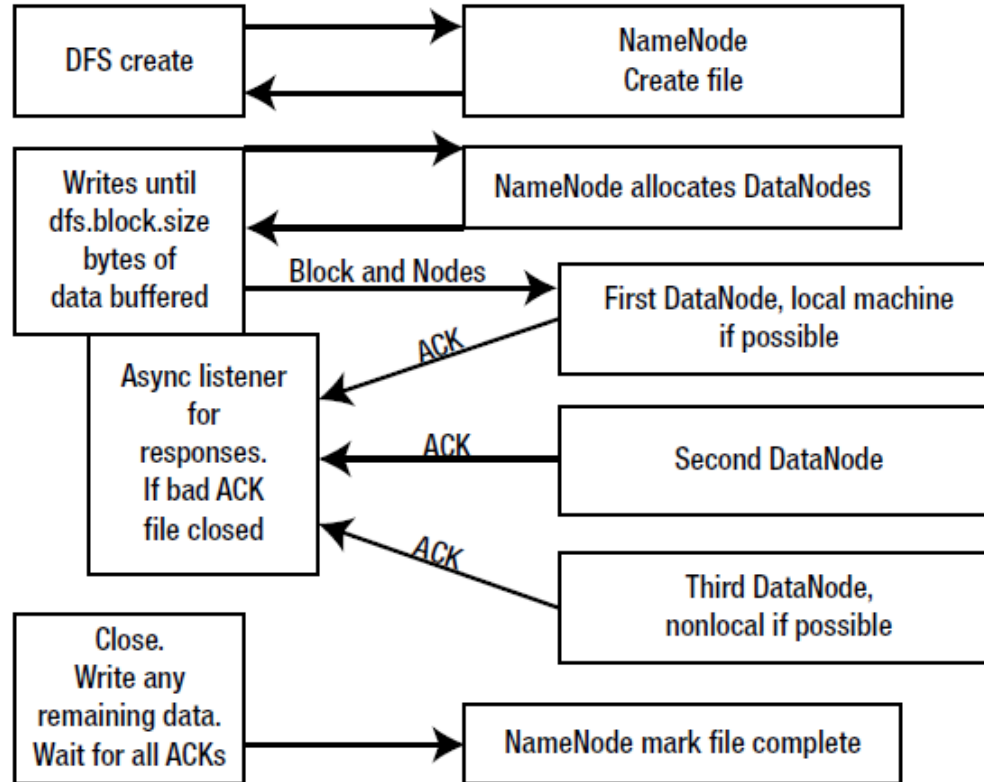
How reducer works



Hadoop HDFS



How HDFS works



Hive



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

Pig



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

HBase



Sqoop



Zookeeper



Oozie



Falcon



Mahout



Avro

