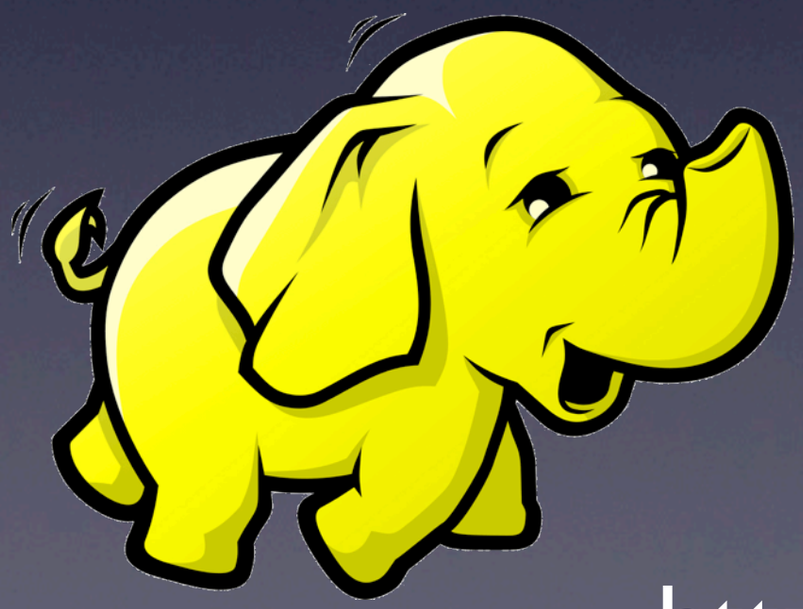
CADEC 2013

Hadoop och Pig

Jacob Tardell, Callista Enterprise



http://www.callistaenterprise.se/cadec2013

Hadoop Summit #1

Announcing the Hadoop Summit at Yahoo, March 25th, 2008

by Jeremy Zawodny (@jzawodn)

Wed February 20, 2008

3 Comments Bookmark Share

Tweet Gilla 0

With all the growing interest in Hadoop (especially after <a href="That'd be this...

http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html

APACHE HADOOP

hadoop

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

Vad är problemet?

Mer data

Mer avancerade beräkningar

Datorerna blir fler

Traditionell parallellprogrammering är svår

Inspirationen

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling ma-

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp

Googla "google mapreduce paper"

Hadoopplattformen

Tre delar:

1. MapReduce för beräkningar
2. Koordinering
3. Lagring

MapReduce

MapReduce är en problemlösningsstrategi.

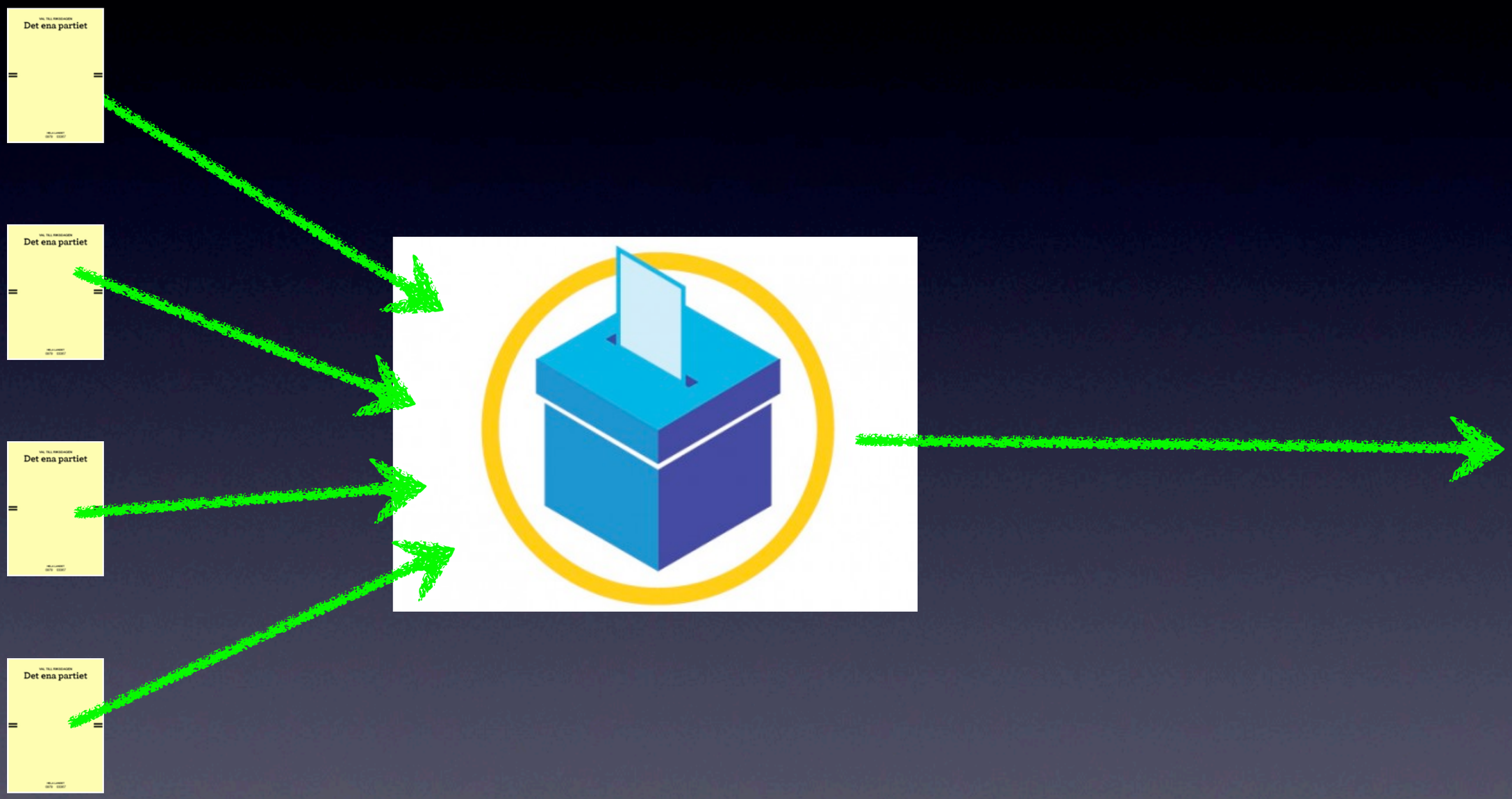
Många, men inte alla, algoritmer kan skrivas som MapReduce-algoritmer.

Många små problem

En variant på söndra och härska:

1. Del upp problemt
2. Lös delarna var för sig
3. Samla ihop resultaten

Ett försök till förklaring



Räkna röster i valet

A 1
B 3
C 2

A 4
B 1
D 2

B 3
C 2
D 2

A 5
B 7
C 4
D 4

En första ansats

```
public static class Map
        extends Mapper<LongWritable, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, Context context)
             throws IOException, InterruptedException {

                String line = value.toString();
                StringTokenizer tokenizer = new StringTokenizer(line);
                while (tokenizer.hasMoreTokens()) {
                        word.set(tokenizer.nextToken());
                        context.write(word, one);
                }
        }
}
```

```
public static class Reduce
                extends Reducer<Text, IntWritable, Text, IntWritable> {

        public void reduce(Text key, Iterable<IntWritable> values,
                Context context) throws IOException, InterruptedException {

                int sum = 0;
                for (IntWritable val : values) {
                        sum += val.get();
                }
                context.write(key, new IntWritable(sum));
        }
}
```

Pig

Dataflow-språk (Pig Latin)

Kompilerar till MapReduce

Ovanpå Hadoop

Kommer från Yahoo!

Räkna ord i Strindberg

word-count.pig

```
/**
* Räkna ord
*
*/

input_lines = LOAD '$INPUT' AS (line);

words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;
filtered_words = FILTER words BY word MATCHES '\\w+';
word_groups = GROUP filtered_words BY word;

word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;
ordered_word_count = ORDER word_count BY count DESC;

STORE ordered_word_count INTO '$OUTPUT';
```

Line: 15 Column: 41 | Pig | Tab Size: 4 | COUNT

Köra skriptet

```
pig-0.10.1/bin/pig -x local
-param INPUT=strindberg/*
-param OUTPUT=wc_res
word-count.pig
```

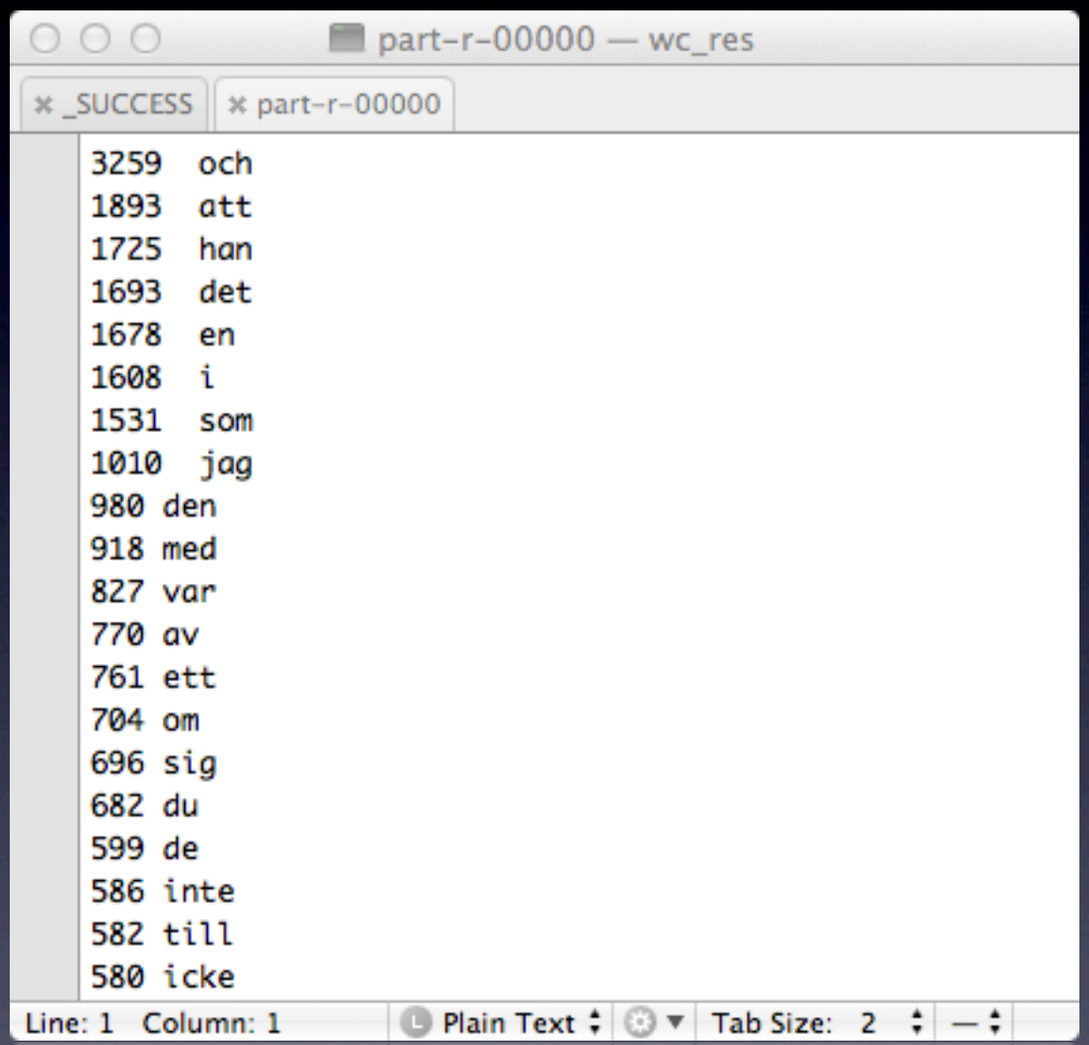
Körning

pig-wc.log

```
2013-01-09 13:55:24,892 [main] INFO  org.apache.pig.Main - Apache Pig version 0.10.1 (r1426677) compiled Dec 28 2012, 16:46:13
2013-01-09 13:55:24,893 [main] INFO  org.apache.pig.Main - Logging error messages to: /Users/jacob/Developer/hadoop/pig-101/pig_135773
2013-01-09 13:55:25.167 java[28569:1703] Unable to load realm info from SCDynamicStore
2013-01-09 13:55:40,795 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system
2013-01-09 13:55:41,674 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2013-01-09 13:55:41,675 [main] WARN  org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
2013-01-09 13:55:41,683 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,ORDER_BY,FI
2013-01-09 13:55:41,911 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thres
2013-01-09 13:55:41,929 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move
2013-01-09 13:55:41,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size be
2013-01-09 13:55:41,956 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size af
2013-01-09 13:55:42,020 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2013-01-09 13:55:42,047 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduc
2013-01-09 13:55:42,081 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up sing
2013-01-09 13:55:42,109 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - BytesPerReducer=
2013-01-09 13:55:42,109 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Neither PARALLEL
2013-01-09 13:55:42,162 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job
```

Line: 172 Column: 127 Plain Text Tab Size: 2

Resultat


part-r-00000 — wc_res
_SUCCESS
part-r-00000
3259 och
1893 att
1725 han
1693 det
1678 en
1608 i
1531 som
1010 jag
980 den
918 med
827 var
770 av
761 ett
704 om
696 sig
682 du
599 de
586 inte
582 till
580 icke
Line: 1 Column: 1
Plain Text
Tab Size: 2

Köra i molnet

Amazon Web Services (AWS)

Amazon Elastic MapReduce (EMR)

Några exempel från verkligheten

Webben (Twitter, Yahoo! etc)

Säkerhetslösning för bank i Utah

Elmätare i Frankrike

Prisberäkningsmodell för Sear's

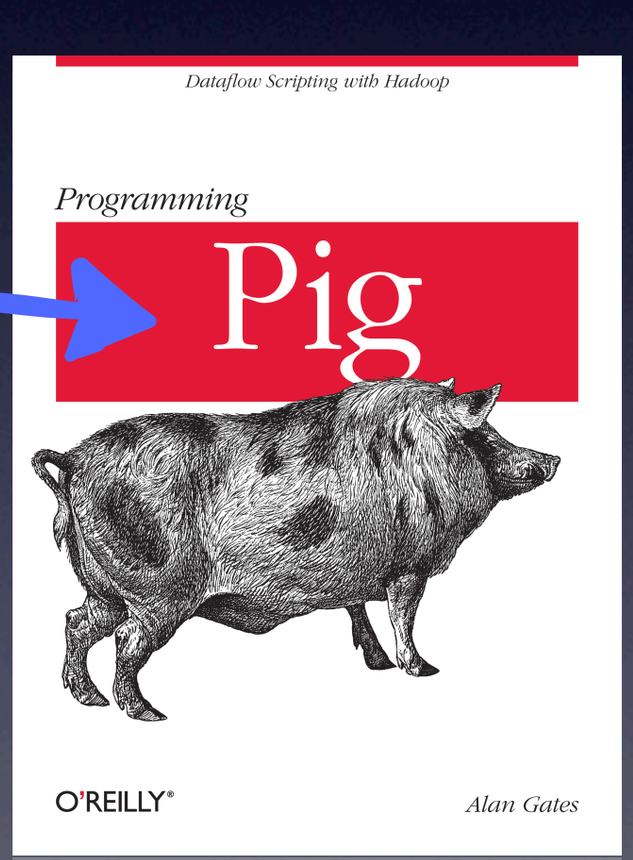
Ekosystemet

Hadoop Commons, MapReduce, HDFS, Hbase, Pig, Hive, Mahout, Zookeeper, Cassandra Sqoop, Avro, Kafka, Lily, Hcatalog, Giraph...

Yahoo!, Hortonworks, Cloudera, Amazon AWS, Facebook, IBM, Microsoft, Twitter

Läs och lek!

http://pig.apache.org



Hadoop Summit 2013
EU: Amsterdam 20-21 mars
US: San Jose, Kalif. 26-27 juni

Frågor?

jacob.tardell@callistaenterprise.se

@jata