Karlsruhe Institute of Technology

Karlsruhe Institute of Technology

Hadoop tutorial

1 - Introduction to Hadoop

A. Hammad, A. García | September 7, 2011

STEINBUCH CENTRE FOR COMPUTING (SCC)

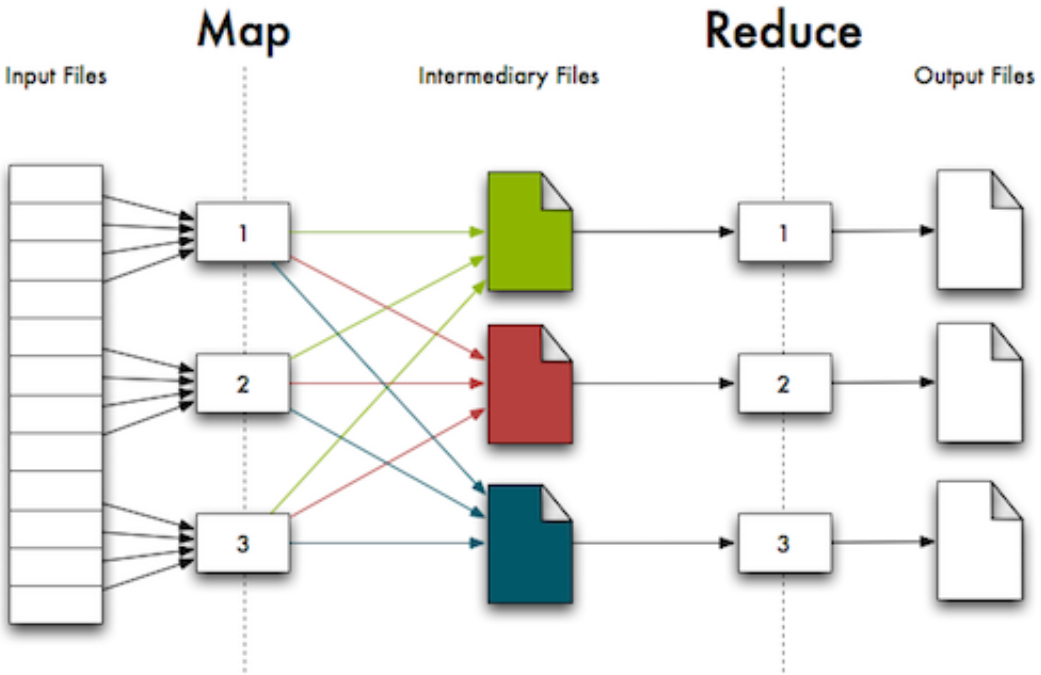
The Hadoop Framework

What is it for?

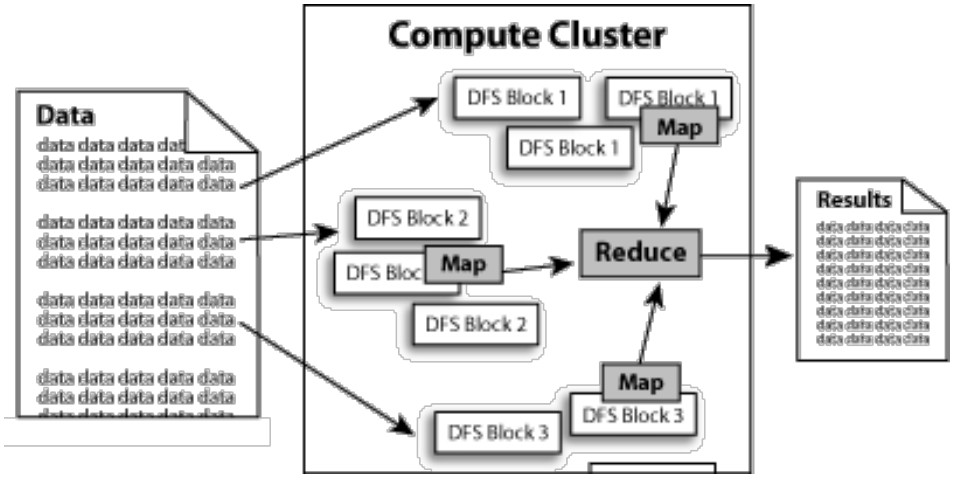
- Data intensive computing on commodity hardware
  - Yahoo's (re)implementation of Google's **Map-Reduce**
    $\Rightarrow$ simple-process huge amounts of data in efficient way
  - highly scalable filesystem, computing coupled to storage

The Map-Reduce paradigm

- (Split step)
- Map step
- (Shuffle step)
- Reduce step



Scalability



- Scalability achieved through data locality
- computing goes to data, not otherwise

Scalability

Daily production usage in Yahoo!, Facebook

- clusters with thousands of nodes
- **30 PB** of data and growing
  - whole dataset processed daily!
  - sorting benchmarks winners, e.g. 1 TB data sorted in 1 minute by 3800 nodes (2009)

Applications

- Programming language Java
- Hadoop Pipes API for C++
- Streaming for any executables (e.g. shell utilities) as mapper or reducer

Example

```
hadoop jar $HADOOP_LIB/hadoop-streaming.jar
    -input /dfsInputDir/myInputData -mapper "shellMapper.sh"
    -reducer "shellReducer.sh" -output /dfsOutputDir/myResults
```

Use cases

- Data intensive computing (high IO)
- High parallelization

Complementary to

- message passing (MPI, ...)
- RDBMS, traditional databases

	Traditional RDBMS	MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

The Hadoop ecosystem

- Apache project, **open source**
- many subprojects
  - common, HDFS, MapReduce
  - Pig: data flow language
  - Hive: a distributed data warehouse, SQL-based language inspired by Google's Bigtables; billions rows, million columns
  - HBase: a distributed, column-oriented database
  - Zookeeper: a distributed, highly available coordination service
  - Oozie: a MapReduce workflow service
  - ...
- **backed by big web players** (Yahoo!, Facebook, Amazon, Twitter, ...)
  $\Rightarrow$ available as a Service: Amazon's Elastic MapReduce

More info

http://hadoop.apache.org/common/docs/current/mapred_tutorial.html

Thanks for listening!

Karlsruhe Institute of Technology

Questions?