Karlsruhe Institute of Technology

Karlsruhe Institute of Technology

Hadoop tutorial

5 - The Pig data processing language

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

STEINBUCH CENTRE FOR COMPUTING (SCC)

The background

Hadoop MapReduce

- + data intensive computing on commodity hardware
- + great scalability
- - too low level for some analysis tasks
- - gets messy if need to combine several MR steps

$\Rightarrow$ **need** to wrap complexity in **higher level framework**

The Pig data processing language

Pig

- higher level language
- **data flow** language
- adapted for processing big amounts of moderately structured data
- more structured data than MapReduce, less structured than RDBMS
- powerful transformation functions and data model
  “Pigs eat anything” :-)

- Pig Latin language
- execution environment/interpreter
  $\Rightarrow$ converts Pig instructions into (series of) MapReduce programs
  transparently

The Pig data processing language

- Pig is a scripting language for exploring large data sets
  $\Rightarrow$ 40% of Yahoo! Hadoop usage
- has powerful debugging concepts: `ILLUSTRATE` to operate on representative sample subset
- extensible: **User Defined Functions** for loading, evaluating, storing
- language structure dictated by MapReduce paradigm: ideal for batch processing large dataset, touching most of the data (cf, databases optimized for tiny operations)

Hands on!

Karlsruhe Institute of Technology

Questions?

Let's try it!