Hadoop

Prof. Tahar Kechadi
School of Computer Science & Informatics

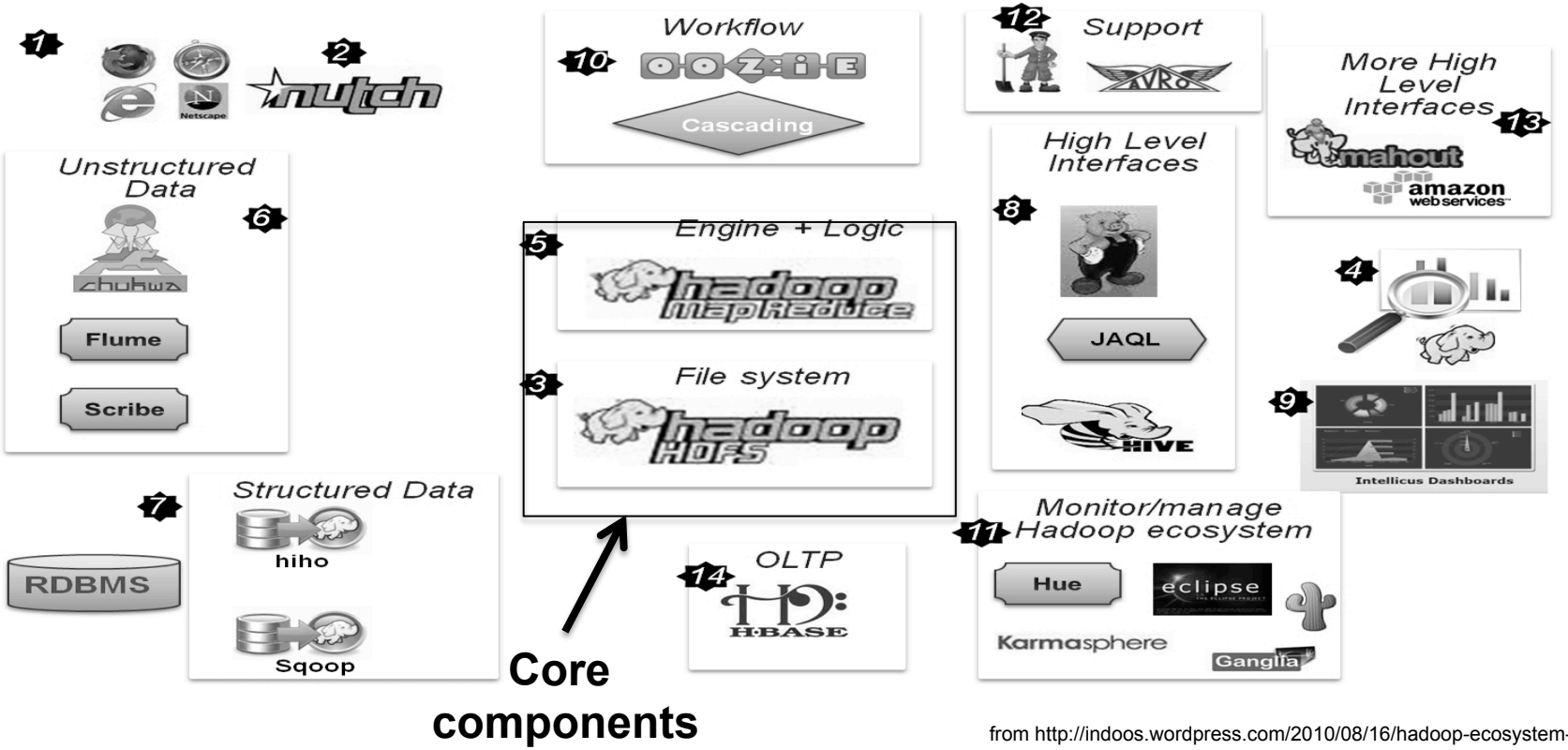
Outline

- **Hadoop**, an open-source implementation

- **Pig**, a high-level abstraction layer

Hadoop & the ecosystem

* Hadoop is an open-source implementation of MapReduce
  * September 2007 – release 0.14.1
  * Latest stable release is 2.2.0 (2.3.0 02/2014)



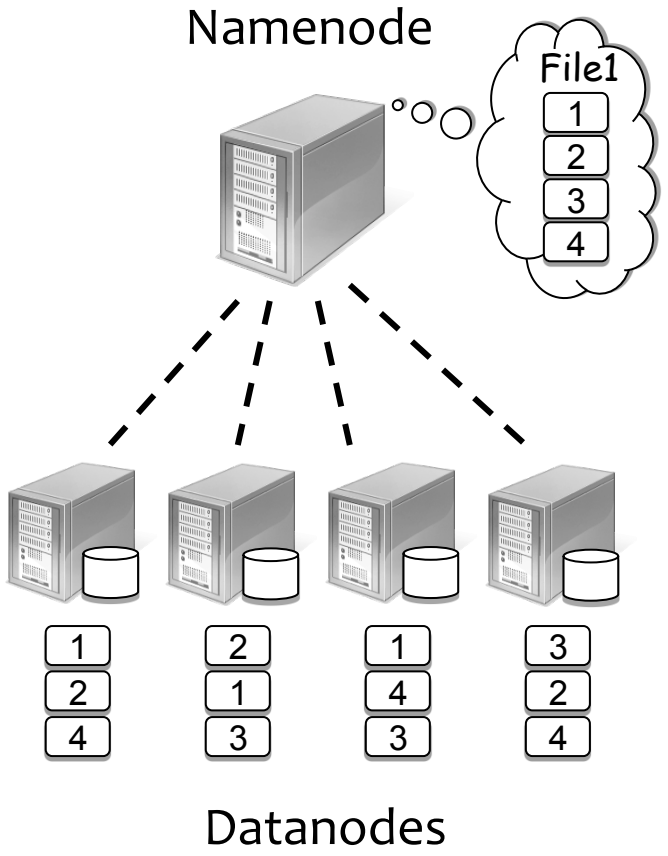
from http://indoos.wordpress.com/2010/08/16/hadoop-ecosystem-world-map/

Hadoop & the ecosystem

* The Hadoop ecosystem is quite large and growing fast
* There is already an explosion of business-focused applications and platforms using Hadoop, including:

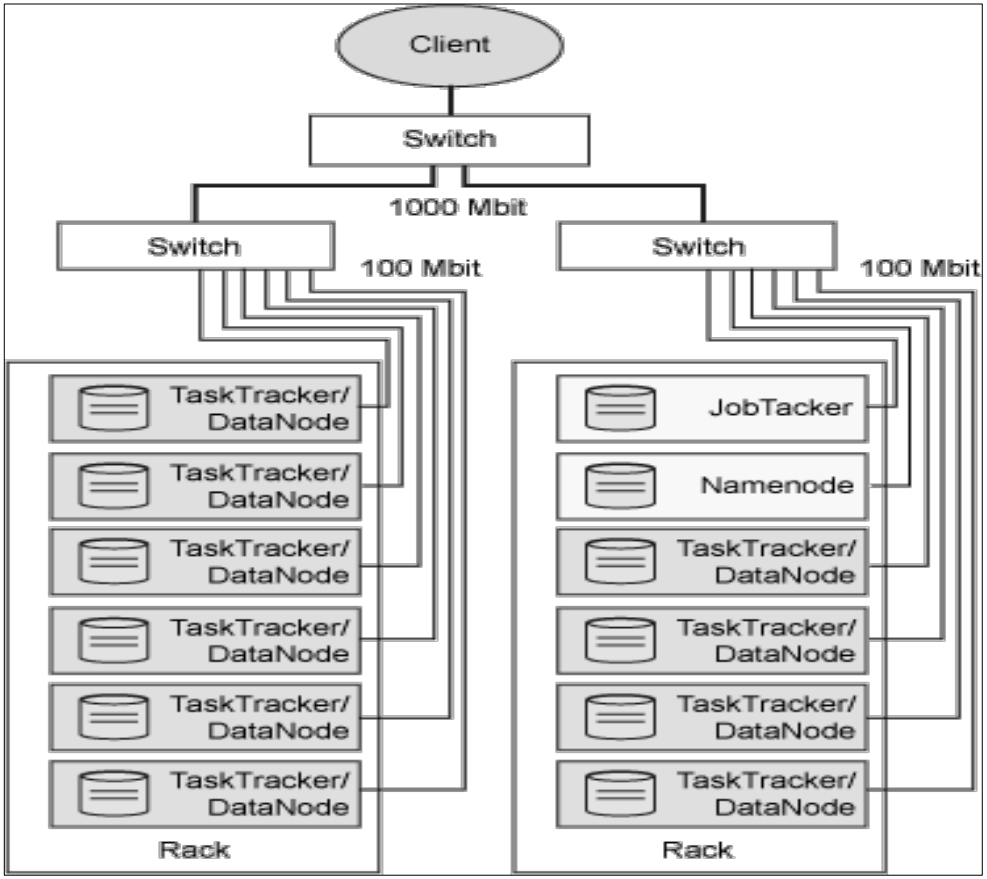
HDFS

- Highly scalable and fault-tolerant
- Blocks replicated across several *datanodes* (usually 3+)
- Single *namenode* stores metadata (file names, block locations, etc.).
- Optimised for large files, sequential reads
- Files written once (no append)
- Files split into 64MB chunks (typical size)



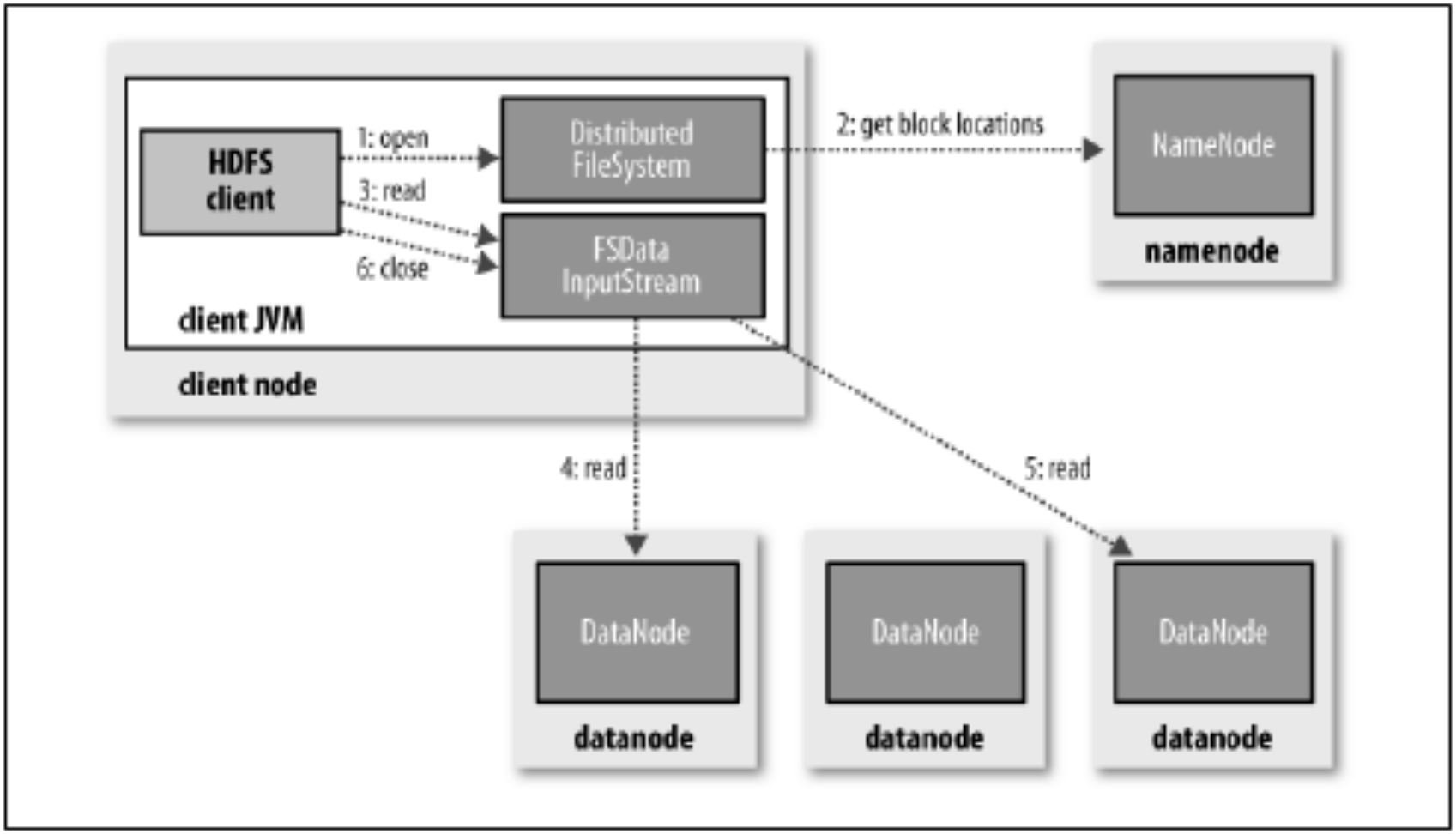
General architecture

- MapReduce layer
  - JobTracker
  - TaskTrackers
- HDFS layer
  - Namenode
  - Datanode



Example of a physical distribution within a hadoop cluster

Reading data



(Tom White)

HDFS

- Pros:
  - Very large files
  - Streaming data access
  - Commodity hardware
  - Fault-tolerance
- Cons:
  - Low-latency data access
  - Lots of small files
  - Multiple writers

Getting Started…

- Download the raw Apache version, or one of the numerous existing distros
  - hadoop.apache.org
  - www.cloudera.com - A set of VMs is also provided
  - http://www.karmasphere.com/
- Three ways to write jobs:
  - Java API
  - Hadoop Streaming (for Python, Perl, etc.)
  - Pipes API (C++)

Hadoop Setup

- Prerequisite: Java
- Create Hadoop Group and User
- Setup ssh certificate
- Setup hadoop
- Setup Hadoop Environment Variables
- Configure Hadoop
- Format namenode
- Start Hadoop service