

Hadoop for Users Training

The Hadoop for Users training will be held in November 2008. This training will be conducted by Yahoo! instructors in Singapore and co-organised by the Infocomm Development Authority of Singapore (IDA) and Yahoo! as part of the Cloud Computing Testbed Initiative.

About Cloud Computing Testbed

In July 2008, IDA Singapore became a Centre of Excellence for Cloud Computing, in partnership with Hewlett-Packard, Yahoo! and Intel, where it will create opportunities for research and development in cloud computing, enhance local capabilities and enable users gain easy access to this next generation service.

About Hadoop

Hadoop is a framework for running applications on large clusters built of commodity hardware. The framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework.

Applications of Hadoop can be found in Internet scale data intensive applications, such as distributed grep, distributed sort, web link-graph reversal, term-vector per host, web access log stats analysis, inverted index construction, document clustering, machine learning, machine translation, and natural language processing. Users of Hadoop include Yahoo!, eBay, Amazon, Facebook, NYTimes, and ChinaMobile.

About Pig and ZooKeeper

Pig was developed at Yahoo! to enable users to do more powerful workflows than simple map/reduce. Tasks that require joining data sources or setting up a pipeline of tasks to do can be difficult to setup using just Hadoop. Pig provides a high-level language for expressing processing steps and compiles Pig programs into map/reduce jobs that can be run on a Hadoop cluster. Not only does this simplify the work of the programmer, but can also generate more efficient map/reduce jobs.

ZooKeeper was also developed at Yahoo! to coordinate processes of distributed applications. Multiple processes mean that one of the processes may fail and comeback, system views may get out of sync, configuration needs to be agreed upon, etc. Even something as simple as knowing which process is alive and who is not can become problematic. ZooKeeper provides a simple interface for handling these kinds of coordination tasks and backs it with a robust and replicated service backend.

Pre-requisites

This 1-day course will cover the basic concepts of parallel and distributed programming and computing. It will be conducted by instructors from the Institute for High Performance Computing (IHPC).

Dates

There will be 2 runs of the training in Nov 2008:

- 12 Nov – Hadoop for Users (for audience with pre-requisites)
- 12 – 13 Nov – Pre-requisites & Hadoop for Users

Training Venue

Executive Seminar Rooms 4.1 and 4.2 (on Level 4), Singapore Management University, Administration Building, 81 Victoria Street, Singapore 188065.

Registration Fee

The above training is free-of-charge to all Singapore-based organisations and companies. Interested institutes of higher learning (IHLs) in Singapore should consolidate staff and student registration.

Please register via email to ida_grid@ida.gov.sg by 1200 hours on 10 Nov. Indicate clearly whether you are registering for the 12 Nov course or the 12 – 13 Nov course. The event URL is at www.ngp.org.sg.

Agenda (please refer to event URL for latest updates)

Hadoop for Users (on 12 Nov and 13 Nov)

[Attendees are required to bring their own laptop for the afternoon hands-on session]

0900-0930 Overview of Hadoop: History, Hadoop at Yahoo! (30 min)
0900-0950 Overview of MapReduce (20 min)
0950-1000 Q&A / Break
1000-1050 Hadoop - DFS, MapReduce API, program debugging, performance tuning (50 min)
1050-1100 Q&A / Break
1100-1125 Pig (25 min)
1125-1150 ZooKeeper (25 min)
1150-1200 Q&A
1200-1300 Lunch
[1300-1700 Lab session]
1300-1330 Accessing & running jobs on Hadoop cluster (to be conducted by IDA/SCS)
1330-1430 “Hello world” Hadoop/map-reduce program (1-2 programming exercises)
1430-1530 “Hello world” Pig program (1-2 programming exercises)
1530-1600 Break
1600-1700 “Hello world” ZooKeeper program (1-2 programming exercises)

Pre-requisite Course (on 12 Nov)

0900-1000 Motivating Parallelism
1000-1100 Parallel & Distributed Computing Platforms
1100-1200 Performance Measures for Parallel Systems
1200-1300 Lunch
1300-1400 Principles of Parallel Algorithms Design & Embarrassingly Parallel Computations
1400-1500 Distributed File Systems & Scheduling
1500-1530 Break
1530–1630 Programming Using the Message Passing Paradigm
1630-1730 Programming Shared Address Space Platforms

Organised by



Facility Sponsor



In Cooperation With



