Introduction on MapReduce programming model with Hadoop

OUTLINE

Note: the outline is subject to change based on background/interest level of attendees.

Day 1 Introduction
Background on common parallel and distributed computing systems
  Common types of system
  Programming models with parallel and distributed computing system.
Overview of Map/Reduce programming model.
  Key ideas in MapReduce programming model.
  Comparing MapReduce model vs. other models.
Introduction on Hadoop Project
  What’s Hadoop is and what is good for.
Tutorial on using shared cyberinfrastructure at TACC

Day 2: Programming with Hadoop
More on Hadoop project
  Structure of a hadoop cluster,
  HDFS
  Data flow in hadoop cluster
Key programming components
  e.g. map, combine, partition, reduce etc.
A walkthrough on “word count” example
Running exemplar program with Hadoop distribution.

Day 3: Data Analysis with Hadoop
Introduction on Mahout library for data mining
  Overview of the library.
  Walkthrough on selected algorithms available in Mahout
A few research project examples of working with Hadoop cluster.
  Example project on using Hadoop for data distribution.
  Example project on implementing new algorithm within Hadoop,
  Example project on organizing analytic processing using Hadoop

Day 4 Introduction on selected projects related with Hadoop
Related projects from Apache
  Pig, Hbase, Hive, etc.
Other related projects for benchmarking and tuning.
Hibench, starfish.
Hadoop based domain specific application
  Crossbow (Bioinformatics), Mavuno(text mining) etc.