# Программа учебного курса «Параллельное программирование в Hadoop»

Созыкин А.В., Гольдштейн М.Л.

Институт математики и механики УрО РАН, г. Екатеринбург

# Hadoop

- Hadoop – один из сервисов OpenCirrus и «Университетский кластер»
- Особенности:
  - Ориентация на обработку больших объемов данных (Терабайты)
  - Автоматическое распараллеливание на тысячи узлов (текущая архитектура до 4 000[1])
  - Работа на кластере машин стандартной архитектуры
  - Перемещение вычислений к данным
  - Автоматическая обработка отказов оборудования

[1] http://developer.yahoo.com/blogs/hadoop/posts/2011/02/mapreduce-nextgen/

# Ограничения Hadoop

- Программная модель Google Map/Reduce:
  - Фиксированная структура выполнения задачи
  - Одна фиксированная фаза коммуникаций
  - Программист пишет только функции Map и Reduce (аналоги из функциональных языков (Lisp, Haskell и т.п.):
    - Map: Список $\rightarrow$ Список
    - Reduce: Список $\rightarrow$ Значение
- Сложность в освоении

# Использование Hadoop


- Hadoop чрезвычайно популярен в США и Западной Европе
- Коммерческие компании:
  - Yahoo, EBay, Facebook, Google, IBM, The New York Times, Twitter
- Научные задачи: обработка естественных языков, машинное обучение, генетические алгоритмы, биоинформатика
- В России Hadoop используется мало

# University Initiative to Address Internet-Scale Computing Challenges

- Совместная программа IBM, Google и университетов США, старт октябрь 2007
- Цель – развитие знаний студентов о современных подходах к параллельным вычислениям
- Предоставляемые ресурсы:
  - Кластер с Apache Hadoop (~1600 ядер)
  - Средства разработки для Hadoop
  - Учебные курсы совместной разработки Google+Университеты

# University Initiative to Address Internet-Scale Computing Challenges

- Университеты:
  - University of Washington
  - Carnegie Mellon University
  - Massachusetts Institute of Technology
  - Stanford University
  - University of California at Berkeley и др.
- Учебные курсы:
  - Доступны бесплатно через Web по лицензии Creative Commons
  - http://code.google.com/intl/ru-RU/edu/parallel/
- Результат программы: рост популярности Hadoop/MapReduce

# Учебный курс Hadoop

- «Университетский кластер»
  - Предоставляется кластер Hadoop
  - Нужно разрабатывать учебные курсы!
- Курс «Параллельное программирование в hadoop»
- Кафедра «Высокопроизводительные компьютерные технологии»
  - Институт математики и механики УрО РАН
  - Уральский федеральный университет

# Программа курса

- Структура курса:
  - 36 академических часов
  - 10 лекций
  - 7 лабораторных работ
- Предварительные требования:
  - Java
  - Linux (желательно)
  - Основы параллельных вычислений (желательно)

# Использованные источники

- Учебные курсы:
  - University of Washington: Problem Solving on Large Scale Clusters
  - California Polytechnic State University: CPE 458-Parallel Programming
  - University of California, San Diego: MapReduce
  - Google: Cluster Computing and MapReduce
  - Google: MapReduce in a Week
- Учебные материалы вендоров:
  - Yahoo! Hadoop Tutorial
  - Apache MapReduce Tutorial

# Темы лекций. Основы

- Введение в Hadoop
- Распределенная файловая система HDFS
- Алгоритм MapReduce
- Разработка и запуск программ в Hadoop
- MapReduce API

# Темы лекций. Прикладные задачи

- Применение Hadoop для автоматической обработки текстов
- Применение Hadoop в научных исследованиях
- Основы администрирования Hadoop
- Apache Pig - язык, управляемый потоком данных
- Распределенная база данных Apache HBase

# Лабораторные работы

- Распределенная файловая система HDFS
- Запуск программ в Hadoop
- Использование Eclipse для разработки ПО для Hadoop
- Hadoop Streaming - программирование для MapReduce не на Java
- Построение простого инвертированного индекса
- Основы работы с Pig
- Основы работы с HBase

# Оборудование

- Ресурсная база: суперкомпьютерный центр ИММ УрО РАН
- Оборудование для работы Hadoop:
  - Кластер из 4-х узлов по 2 двуядерных процессора AMD Opteron 2ГГц, 8 ГБ памяти, 250 ГБ жесткий диск, Ubuntu Linux
  - Сервер 2 двенадцатиядерных прцессора AMD Opteron 2 ГГц, 24 ГБ памяти, 250 ГБ жесткий диск, Ubuntu Linux (Pseudo-distributed mode)
- Для студентов: персональный компьютер с Java под Windows или Linux

# Среда разработки

- Основной вариант: Karmasphere + Eclipse:
  - Включает дистрибутив Hadoop
  - Полностью готова к использованию, не надо ничего настраивать и устанавливать
  - Платформы Windows (нужен cygwin) и Linux
  - Karmasphere Community Edition бесплатна
- Альтернативные варианты:
  - Eclipse + стандартный plugin из Hadoop
  - Karmasphere + Netbeans

# Материалы курса

- Страница курса:
  - http://www.asozykin.ru/courses/hadoop
- Программа доступна сейчас
- Материалы (лекции + лабораторные) будут выставлены до 1 сентября 2011 г.
- Возможна передача ранней версии курса для ознакомления:
  - Запрос на адрес: avs@imm.uran.ru
  - Просьба написать рецензию на курс

# Итоги

- Программа курса «Параллельное программирование в Hadoop»
- Основа: курсы по Hadoop университетов США
- Использование в «Университетском кластере»:
  - Передача материалов курса
  - Обучение через Web, лабораторные на кластере ИММ УрО РАН
  - Личное обучение

Вопросы?

# Примеры научных работ

- **MrsRF: an efficient MapReduce algorithm for analyzing large collections of evolutionary trees.** Suzanne J Matthews and Tiffani L Williams. BMC Bioinformatics 2010, 11(Suppl 1):S1.
- **MapReduce-Based Pattern Finding Algorithm Applied in Motif Detection for Prescription Compatibility Network.** Yang Liu, Xiaohong Jiang, Huajun Chen, Jun Ma, and Xiangyu Zhang. APPT 2009, LNCS 5737, pp. 341 – 355, 2009.
- **MapReduce for Data Intensive Scientific Analysis.** Jaliya Ekanayake and Shrideep Pallickara. 2008. Fourth IEEE International Conference on e-Science. pp. 277--284.
- **Scaling Genetic Algorithms Using MapReduce**. Verma, A., Llora, X., Goldberg, D. E., and Campbell, R. H. 2009. Proceedings of the 2009 Ninth international Conference on intelligent Systems Design and Applications.
- **Grammar based statistical MT on Hadoop: An end-to-end toolkit for large scale PSCFG based MT**. Ashish Venugopal, Andreas Zollmann. The Prague Bulletin of Mathematical Linguistics No.91, 2009, 67–78.