CS15-319 / 15-619
Cloud Computing

Recitation 13
April 16th, 2013
Announcements

• Open up S3 location of hand ins:
  – Give access to your S3 bucket to:
    • public
    • onlinecloudcomputingcourse@gmail.com
  – You could lose credit or be penalized otherwise
  – See Piazza Post on how to open up your handin directory

• Encounter a general bug:
  – Post on Piazza

• Encounter a grading bug:
  – Post Privately on Piazza

• Post feedback on OLI
Unit 4 – Checkpoint Quiz

• 95% Students completed
• Average score is 94.5% (for students who completed)
New Modules

• Unit 5 – Distributed Programming and Analytics Engines for the Cloud
  – Introduction to Distributed Programming for the Cloud
  – Distributed Analytics Engines for the Cloud: **MapReduce**
    • Introduction
    • The Programming Model
    • The Data Structure and Flow
    • Examples: WordCount, Sort and Sobel
    • The Computation and Architectural Models
    • Job and Task Scheduling
    • Fault-Tolerance
    • YARN: The New Hadoop MapReduce
Project 4, Part b

• Project 4, Part a
  – MapReduce
  – Project 4 Survey

• Project 4, Part b
  – Input Text Predictor: NGram Generation
Common Queries on MapReduce

• Running Hadoop Jobs on a Whirr Cluster
  – You can either run the job from the instance that launched the cluster
    • Need to start an SSH proxy to your cluster
    • Export $HADOOP_CONF_DIR to ~/.whirr/<your-cluster-name>
  – Or you can run the Hadoop job from the Master node
    • SSH to the master node instance
    • List of instances is at ~/.whirr/<your-cluster-name>/instances
    • By default use ssh -i ~/.ssh/id_rsa and **not**
      ~/.ssh/id_rsa.pub
Common Queries on MapReduce

• Job fails because Output Directory already exists:
  – HDFS is an immutable file system
  – Either specify a new output directory
  – Or delete the existing directory `hadoop dfs -rmr`

• Job fails because of lack of memory
  – Hadoop jobs require a lot of ram
  – Launch them on at least m1.small
  – You can modify the heap size in `mapred-site.xml` or pass it to your Hadoop job
    • `mapred.child.java.opts=-Xmx1024m`
MapReduce Tips and Tricks

- Use the Eclipse Java IDE
  - Code completion, options to package JARs etc.
  - MapReduce plugins (Hadoop, and third-party)
- Use byobu to keep your remote sessions alive
- Ensure that whirr has launched all the Hadoop processes properly
  - SSH to the master node to verify Hadoop is installed and run sample jobs to verify cluster is functioning
- Monitor your cluster through the web interfaces
  - `<ec2-public-ip>:50030` and `<ec2-public-ip>:50070`
  - Remember to set the Security Groups to open those two ports for all IPs
Input Text Prediction

• Construct an Input Text Predictor

Google Suggest

WordLogic iKnowU keyboard
How to Construct an Input Text Predictor?

1. **Given a language corpus**
   - Project Gutenberg (2.5GB, already on S3)
   - English Language Wikipedia Articles (30GB, on S3 soon)

2. **Construct an n-gram model of the corpus**
   - An n-gram is a phrase with n words.
   - For example a set of 1,2,3,4,5-grams with counts:
     - this 1000
     - this is 500
     - this is a 125
     - this is a blue 60
     - this is a blue house 20
How to Construct an Input Text Predictor?

3. Build a statistical language model that contains the probability of a word appearing after a phrase

- \( \text{Pr}(\text{is} | \text{this}) = \frac{\text{Count}(\text{this is})}{\text{Count}(\text{this})} = \frac{500}{1000} = 0.5 \)
- \( \text{Pr}(\text{a} | \text{this is}) = \frac{\text{Count}(\text{this is a})}{\text{Count}(\text{this is})} = \frac{125}{500} = 0.25 \)

4. Store and index the words and their probabilities to use in an application
Discussions

• Your questions...
Upcoming Deadlines

• Unit 5:

  UNIT 5: Distributed Programming and Analytics Engines for the Cloud

  Module 18: Introduction to Distributed Programming for the Cloud

  Module 19: Distributed Analytics Engines for the Cloud: MapReduce

• Project 4

  Project 4

  Module 32: Input Text Predictor : Ngram Generation

  NGram Generation  |  Checkpoint  |  Available Now
  Due 4/21/13 11:59 PM