CS15-319 / 15-619
Cloud Computing

Recitation 13

November 19th and Nov 22nd, 2013

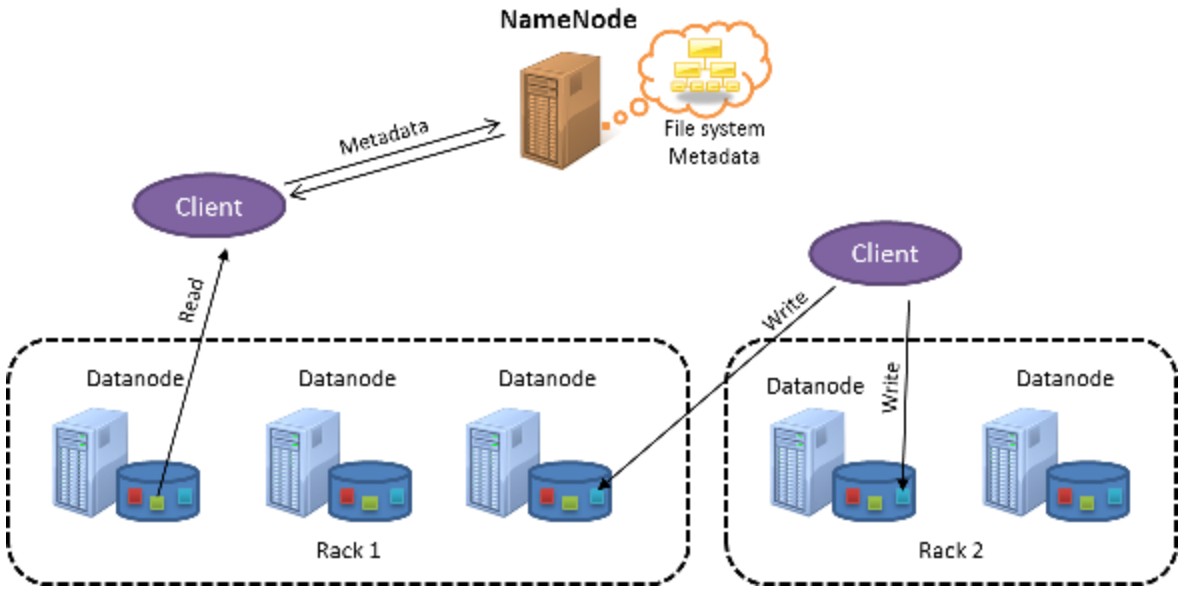
Announcements

- Encounter a general bug:
  - Post on Piazza
- Encounter a grading bug:
  - Post Privately on Piazza
- Don’t ask if my answer is correct
- Don’t post code on Piazza
- Search before posting
- Post feedback on OLI

Piazza Questions

- Program taking a long time to run
  - Please check your code for complex data structures
- Moving data to HDFS
  - HDFS is not directly mountable
    - User Space vs. Linked to O/S Kernel

Hadoop Distributed File System



Writing never completes until replication is finished

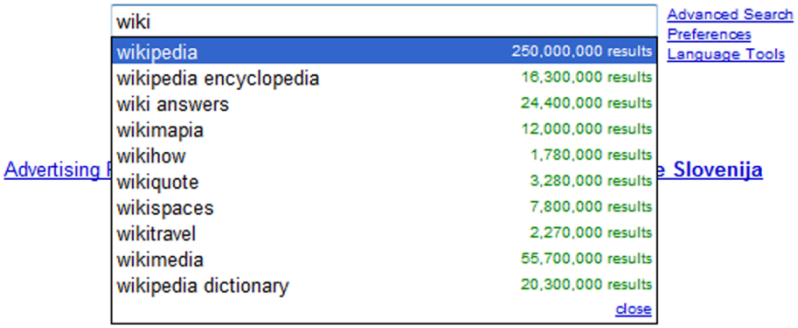
– Replication: 3, data block will reside at 3 different data nodes

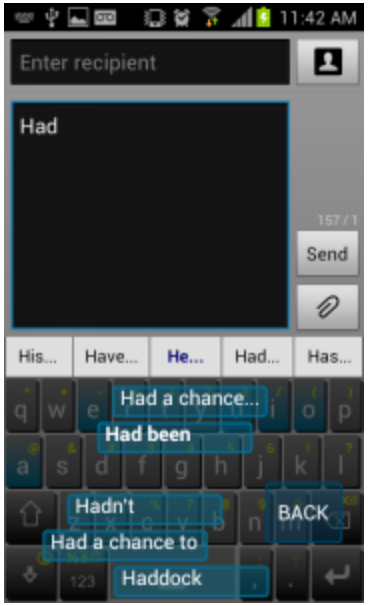
Module to Read

- UNIT 5: Distributed Programming and Analytics Engines for the Cloud
  - Module 16: Introduction to Distributed Programming for the Cloud
  - Module 17: Distributed Analytics Engines for the Cloud: MapReduce
  - Module 18: Distributed Analytics Engines for the Cloud: Pregel
  - Module 19: Distributed Analytics Engines for the Cloud: GraphLab

Input Text Predictor

- Suggest words based on letters already typed





n-gram

- An *n*-gram is a phrase with *n* contiguous words

Example Phrase: This is interesting because this is a cloud computing course

#	1-gram	Count
1	this	2
2	is	2
3	interesting	1
4	because	1
5	a	1
6	cloud	1
7	computing	1
8	course	1

#	2-gram	Count
1	this is	2
2	is interesting	1
3	interesting because	1
4	because this	1
5	is a	1
6	a cloud	1
7	cloud computing	1
8	computing course	1

#	3-gram	Count
1	this is interesting	1
2	is interesting because	1
3	interesting because this	1
4	because this is	1
5	this is a	1
6	is a cloud	1
7	a cloud computing	1
8	cloud computing course	1

#	4-gram	Count
1	this is interesting because	1
2	is interesting because this	1
3	interesting because this is	1
4	because this is a	1
5	this is a cloud	1
6	is a cloud computing	1
7	a cloud computing course	1
8		

#	5-gram	Count
1	this is interesting because this	1
2	is interesting because this is	1
3	interesting because this is a	1
4	because this is a cloud	1
5	this is a cloud computing	1
6	is a cloud computing course	1
7		
8		

#	6-gram	Count
1	this is interesting because this is	1
2	is interesting because this is a	1
3	interesting because this is a cloud	1
4	because this is a cloud computing	1
5	this is a cloud computing course	1
6		
7		
8		

How to Construct an Input Text Predictor?

1. Given a language corpus
   - Project Gutenberg (2.5 GB)
   - English Language Wikipedia Articles (30 GB)
2. Construct an n-gram model of the corpus
   - An n-gram is a phrase with n contiguous 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 cloud 60
     - this is a cloud computing 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
   - $\Pr(is|this) = \frac{Count(this\ is)}{Count(this)} = \frac{500}{1000} = 0.5$
   - $\Pr(a|this\ is) = \frac{Count(this\ is\ a)}{Count(this\ is)} = \frac{125}{500} = 0.25$
4. Store and index the words and their probabilities to use in an application

This Week's Goal

Construct an n-gram model of the corpus

- An n-gram is a phrase with n contiguous 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 cloud 60
  - this is a cloud computing 20

Upcoming Deadlines

- Project 4:

Project 4		
Input Text Predictor: NGram Generation		
NGram Generation	Checkpoint	11:59PM 11/24/2013

- Unit 5:

UNIT 5: Distributed Programming and Analytics Engines for the Cloud
Module 16: Introduction to Distributed Programming for the Cloud
Module 17: Distributed Analytics Engines for the Cloud: MapReduce
Module 18: Distributed Analytics Engines for the Cloud: Pregel
Module 19: Distributed Analytics Engines for the Cloud: GraphLab

Demo Outline

- 1. Hadoop Commands
  - hadoop fs -put
  - hadoop fs -get
  - hadoop distcp
  - http://hadoop.apache.org/docs/r1.0.4/commands_manual.html
- 2. N-Gram Generation
  - Google Instant
  - Input Text Predictor
  - N-Gram Generation