CS15-319 / 15-619
Cloud Computing

Recitation 14

April 22nd and 24th, 2014

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

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

Project 4, Module 2 Reflections

Construct an n-gram model of the corpus

- An n-gram is a phrase with n contiguous words
- A example of 1,2,3,4,5-grams with counts:

…… this is a cloud computing ……

#	Example	Count
1	this	1000
2	this is	500
3	this is a	125
4	this is a cloud	60
5	this is a cloud computing	20

Statistical Language Model (SLM)

- Provide a mechanism to solve common natural language processing problems
- Examples: speech recognition, machine translation and intelligent input method
- SLM estimates the probability of a word given the previous phrases and the N-gram count
- N-gram model is one of the most popular mechanisms to generate an SLM today

Project 4 Module 3

- Build a statistical language model (SLM) that reflects the possibility of a word appearing after a word or a phrase

#	Example	Count
1	this	1000
2	this is	500
3	this is a	125
4	this is a cloud	60
5	this is a cloud computing	20

$$\Pr(is|this) = \frac{Count(this\ is)}{Count(this)} = \frac{500}{1000} = 0.5$$

$$\Pr(a|this\ is) = \frac{Count(ths\ is\ a)}{Count(this\ is)} = \frac{125}{500} = 0.25$$

Project 4 Module 3

Example:

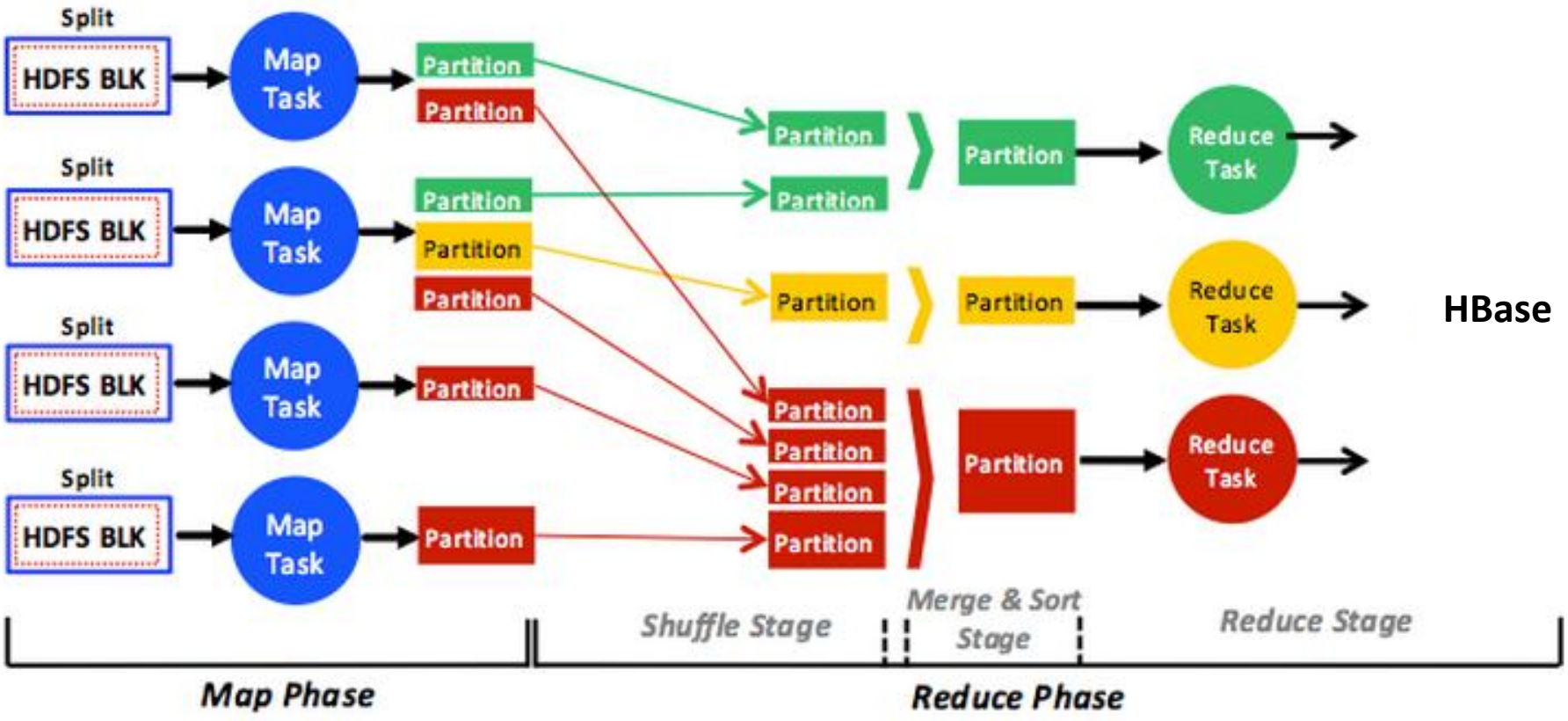
this
this is this day this was

Options	Count	Probability
this was	150	0.15
this is	500	0.50
this day	250	0.25
this kiss	25	0.03
this boy	75	0.08

Options	Count	Probability
this is	500	0.50
this day	250	0.25
this was	150	0.15
this boy	75	0.08
this kiss	25	0.03

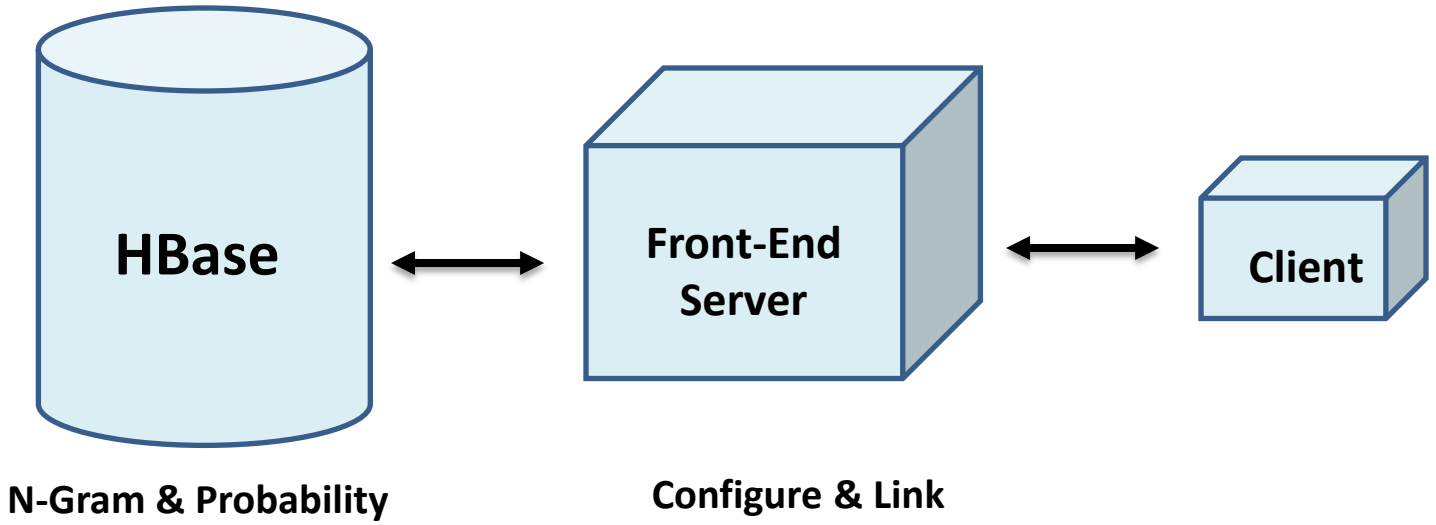
Project 4 Module 3

- Read the input (N-gram) from HDFS and write the output (SLM) to HBase



Project 4 Module 3

- Connect HBase with the Front-End to provide the set of predictions to the web service



Upcoming Deadlines

- Project 4:

Project 4			
Input Text Predictor: Language Model and User Interface			
	Language Model Generation	Checkpoint	4/27/14 11:59PM

- 15-619 Project:
  - Phase 3 (last phase) is due on April 23rd
  - Live-test will be announced

Demo

- Objective:
  - Develop a schema in Hbase to store words and their probabilities
  - Write a MapReduce program to read the n-gram counts and build the statistical language model
  - Render an ordered list of the predicted phrases on the web interface
- Command Line Input Requirement:
  - Ignore phrases that appear below a certain threshold: **t**
  - Store only the top **n** words