CSC 369: Distributed Computing
Winter 2019

Instructor: **Alexander Dekhtyar**, *dekhtyar@calpoly.edu*, ***14-210***

Office Hours:

When		Who	Where
Monday	***9:10am - 10:00am***	***Alex***	**14-210**
Tuesday	***1:10pm - 3:00pm***	***Alex***	**14-210**
Friday	***9:30am - 10:00am***	***Alex***	**14-210**

Additional appoinments: send email.

News and Notes

- MongoDB API documentation is now linked from the page
  [January 23, 2019]
- Welcome to CSC 369. The page is now up.

Old News and Notes

Course Materials

Syllabus	Postscript	PDF
Jupyter Server	ambari-head.csc.calpoly.edu	

Labs

Lab 1	**Due: January 14**	**JSON Manipulation**	Postscript	PDF	Lab Data	*[January 9, 2019]*
Lab 2	**Due: January 16**	**MongoDB: first steps**	Postscript	PDF	Lab Data	*[January 14, 2019]*
Lab 3	**Due: January 28**	**MongoDB find() queries/aggregate pipelines**	Postscript	PDF		*[January 23, 2019]*
Lab 4	**Due: February 4**	**MongoDB Aggregate Pipelines**	Postscript	PDF		*[January 31, 2019]*
Lab 5	**Due: February 13**	**Simple Hadoop Programs**	Postscript	PDF	Lab Info	*[February 6, 2019]*
Lab 6	**Due: March 1**	**Intricate Hadoop Programs**	Postscript	PDF	Lab Info (coming up)	*[February 15, 2019]*
Lab 7	**Due: March 7**	**Simple Spark**		PDF		*[March 6, 2019]*
Lab 8	**Due: March 19**	**Real Spark**	Postscript	PDF		*[March 8, 2019]*

MongoDB Resources

Logs

January 14	Jan14-01.log

Python MongoDB API

PyMongo 3.7.2 Documentation	HTML
Tutorial	HTML
Authentication example	HTML
Aggregation pipeline example	HTML

Code and Queries

MongoDB Python API example example.py example.out (example.py output)

Hadoop Resources

Hadoop Resources and code is posted here.

Hadoop Cluster Monitor

Monitor hadoop jobs here	http://ambari-head.csc.calpoly.edu:8088/cluster

Resources

The Original MapReduce paper	PDF	
`org.apache.hadoop` Version 2.7 javadocs	API	
`org.apache.hadoop` Version 2.7 Jar file	hadoop-core-1.2.1.jar	
Bash local variable settings	bashrc-commands.txt	*Paste into the bottom of your `.bashrc` file*
MapReduce (Hadoop v. 2.7) tutorial	HTML	

Code

Code samples discussed in class are posted here

Hadoop program template	template.java
Our first Hadoop program	switchMR.java
Data file for switchMR.java	data
Input Format Tests	
`TextInputFormat` test	FITest.java
`KeyValueTextInputFormat` test	KeyValueTest.java
`FixedRecordInputFormat` test	FixedRecordTest.java
`NLineInputFormat` test	NLTest.java
`NLineInputFormat` test	NLgroup.java

One-mapper/One-reducer version of `NLgroup.java`	SeqGroup.java	
Multiple chained MapReduce jobs	filter.java	words (input file)
Multiple Input Files/Multiple Mappers	multiInMR.java	users.in, messages.in (input files)
Use of JSON		
Using JSON objects	JsonJob.java	json.in,simple.json (input files)
Multiline JSON	MultilineJsonJob.java	test.json (input file)
Multiline JSON Input Format	json-mapreduce-1.0.jar	
Advanced Hadoop Features		
Finding Max	FindMax.java	numbers.txt
Map-Side Join with Distributed Cache	dCacheDemo.java	users.in, messages.in (input files)
Combiner Test: graph scan with no Combiner	TwitterTest.java	
Combiner Test: graph scan with Combiner	CombinerTest.java	

Spark Resources

Spark resources and Spark code discussed in class will go here

Resources

The Original Spark paper (USENIX Cloud Computing'2010)	PDF
Resilient Distributed Datasets (USENIX NSDI'2012)	PDF
PySpark Documentation	HTML
Wienqiang Feng: Learning Apache Spark with Python	PDF
Running PySpark Applications on ambari-head.csc.calpoly.edu	Googledoc

Code

In-class Example (March 1 lecture)	inClass.py
Use of Hadoop Files	htest.py

Homeworks

Lecture Notes

Lecture 1	**What's in this class?**	Postscript	PDF	*[January 4, 2016]*
Lecture 2	**Motivating Examples**	Postscript	PDF	*[January 4, 2016]*
Lecture 3-1	**JSON**	Postscript	PDF	*[January 10, 2017]*
Lecture 3-2	**Maps, Dictionaries, Key-Value Pairs**	Postscript	PDF	*[January 12, 2016]*
Lecture 4	**MongoDB Basics**	Postscript	PDF	*[January 18, 2016]*
Lecture 5	**MongoDB Java Connectivity**	Postscript	PDF	*[January 28, 2016]*
Lecture 6	**MongoDB Aggregation Pipeline**	Postscript	PDF	*[January 27, 2017]*
Lecture 7	**MongoDB Aggregation Pipeline: Part 2**	Postscript	PDF	*[Feb 3, 2017]*
Lecture 8	**Overview of Distributed Systems**	Postscript	PDF	*[February 4, 2017]*
Lecture 9	**MapReduce**	Postscript	PDF	*[January 28, 2016]*

Lecture 10	**Hadoop on our cluster**	Postscript	PDF	*[February 4, 2019]*
Lecture 11	**HDFS commands primer**	Postscript	PDF	*[February 13, 2017]*
Lecture 12	**Hadoop Input Data Formats**	Postscript	PDF	*[February 21, 2017]*
Lecture 14	**Matrix Multiplication in MapReduce**	Postscript	PDF	*[March 5, 2017]*
Lecture 15	**MapReduce for Top K Problem**	Postscript	PDF	*[March 10, 2017]*
Lecture 16	**Resilient Distributed Datasts**	Postscript	PDF	*[February 28, 2019]*

Other Materials

JSON

JSON home page	json.org
JSON specification	ECMA-404: The JSON Data Interchange Format (PDF)
org.json Javadocs	Javadoc

January 7, 2019, dekhtyar ***at*** *csc.calpoly.edu*