CSC 369: Distributed Computing

Alex Dekhtyar

Day 1: Welcome

Syllabus

- Teaching and Communication
- Textbook(s)
- Grading
- Exams
- Labs
- Late Policies

Course

- What is “distributed computing”
- Why study it?
- Examples of problems

Syllabus: Teaching and Communication

Lectures are synchronous but recorded

Syllabus: Teaching and Communication

Lectures are synchronous but recorded

Lab periods may be used for guided activities
But often are just for work on lab assignments

Office hour between lecture and lab (M,F)

Syllabus: Teaching and Communication

Mailing list

Slack

Zoom

Static Website

Canvas

Waitlist

Drop/Add deadline: **April 15**

All waitlisted students get full access to class for two weeks

First five (5) days - all adds handled automatically

Everyone else - I will look at the state of affairs next Monday.

Syllabus: Textbooks

NONE

Lecture Notes

Documentation

Original MapReduce and Spark papers

Syllabus: Books

Donald Miner, Adam Shook, ***MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems***, O'Reiley Media, 1st Edition, 2012, ISBN: 978-1449327170.

Mahmoud Parsian, Data Algorithms: ***Recipes for Scaling Up With Hadoop and Spark,*** O'Reiley Media, 2015, ISBN: 978-1491906187.

Christina Chodorow, ***MongoDB: The Definitive Guide***, O'Reiley Media, 2013, ISBN: 978-144924468

Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, ***Learning Spark: Lightining-Fast Big Data Analysis***, Packt, 2015, ISBN: 978- 1449358624

Tomasz Drabas, Denny Lee, ***Learning PySpark***, O'Reiley Media, 2017, ISBN-13: 978-1786463708

Syllabus: Grading

Labs	50-60%
Exams/Written Assessments	35-50%
Homework/Study guides	0-5%

Syllabus: Labs

~ 8 Labs (rougly weekly)

- 1 Intro (Lab 1 starts today)
- 2- 3 MongoDB
- 2-3 Hadoop
- 2-3 Spark

Syllabus: Labs

~ 8 Labs (rougly weekly)

- 1 Intro (Lab 1 starts today)
- 2- 3 MongoDB
- 2-3 Hadoop
- 2-3 Spark

Mostly individual

Some pair programming experiments mid-quarter

Syllabus: Exams

Syllabus: Exams

Syllabus: Exams

Combination of programming and short timed tests.

- **MongoDB programming test + quiz**
- **Hadoop programming test + quiz**
- **Spark programming test + quiz (Final exam time)**

Syllabus: Exams

Combination of programming and short timed tests.

- **MongoDB programming test + quiz**
- **Hadoop programming test + quiz**
- **Spark programming test + quiz (Final exam time)**

Open “most things” on programming tests
Still thinking how to make quizzes work

Syllabus: Late Policies

Step 1. Talk to Me!!!!!

Syllabus: Late Policies

Step 1. Talk to Me!!!!!

- *Deadlines are already lenient*
- *There is a grace period*
- *Deadlines are to prevent you from being bogged down with one problem*
- *Partial credit*

Syllabus

- Teaching and Communication
- Textbook(s)
- Grading
- Exams
- Labs
- Late Policies

Course

- What is “distributed computing”
- Why study it?
- Examples of problems

One small thing: I forgot to ask a couple of questions

https://forms.gle/2vuNJr1nR6FWpioG8

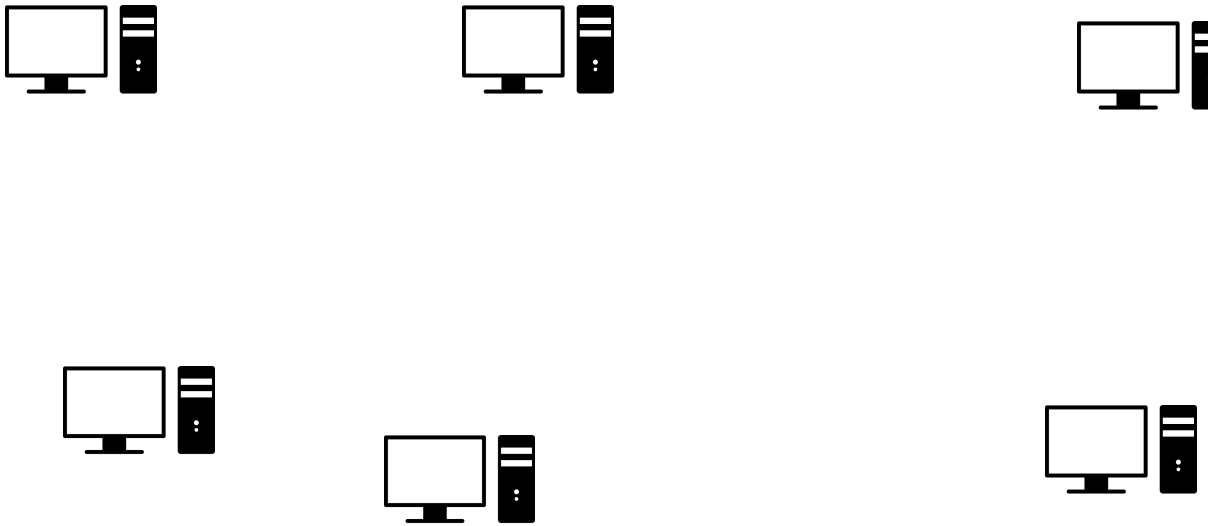
Distributed Computing

Distributed Computing

Multiple independent computers work on the same problem at the same time

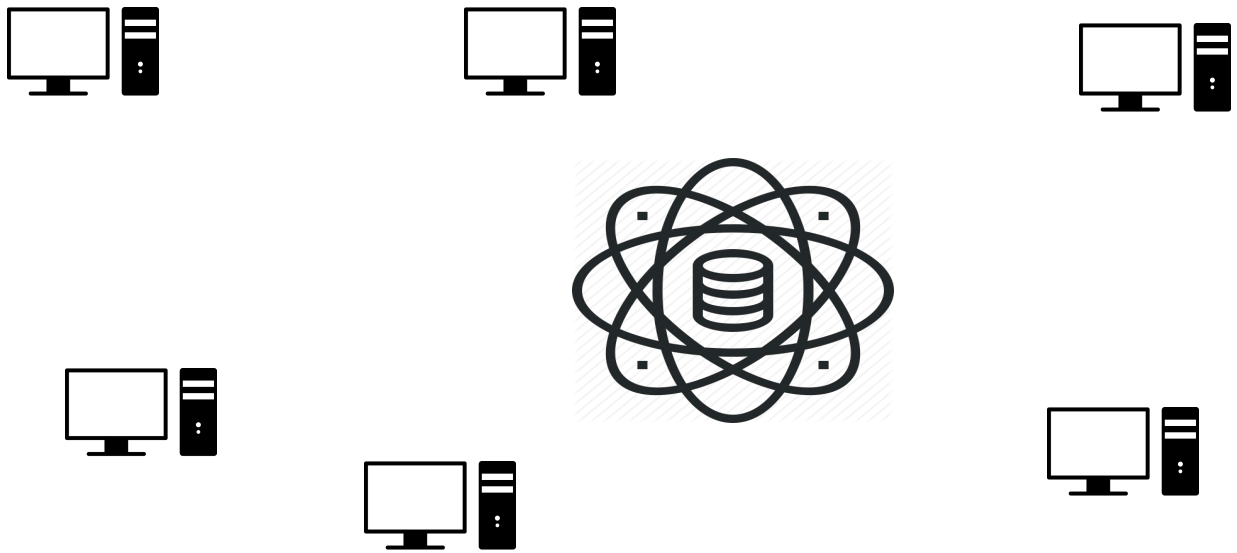
Distributed Computing

Multiple independent computers work on the same problem at the same time



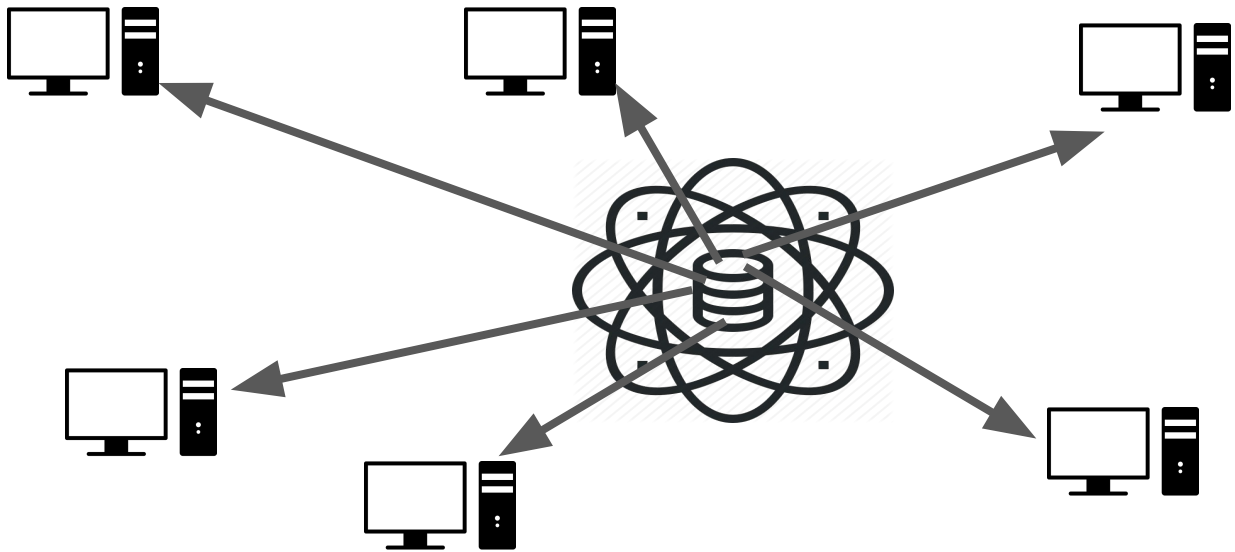
Distributed Computing

Multiple independent computers work on the same problem at the same time



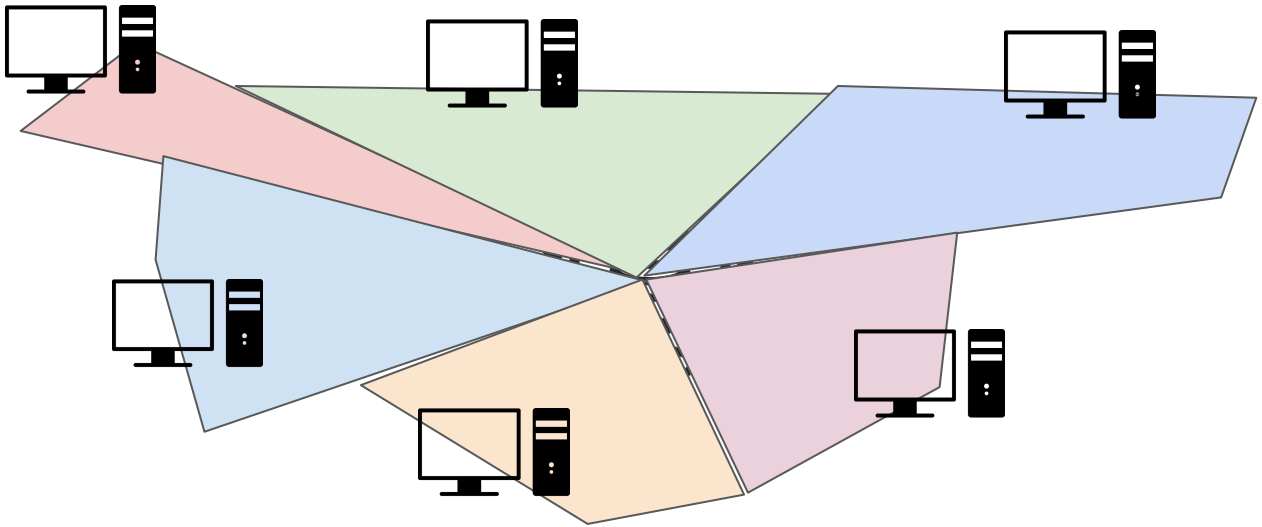
Distributed Computing

Multiple independent computers work on the same problem at the same time



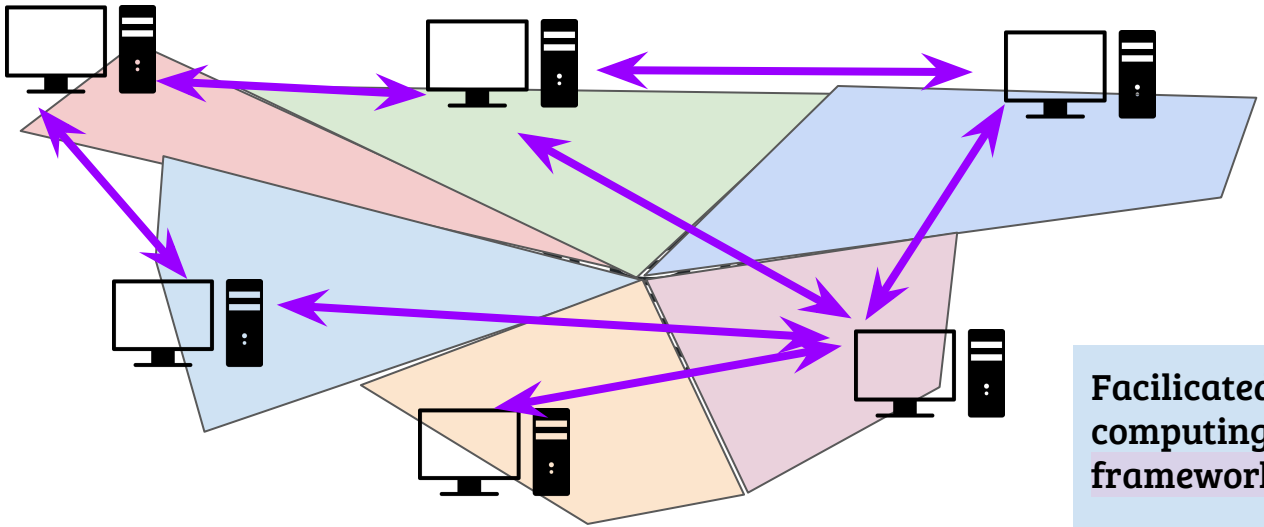
Distributed Computing

Multiple independent computers work on the same problem at the same time



Distributed Computing

Multiple independent computers work on the same problem at the same time



Facilicated by distributed computing systems and frameworks

Distributed Computing

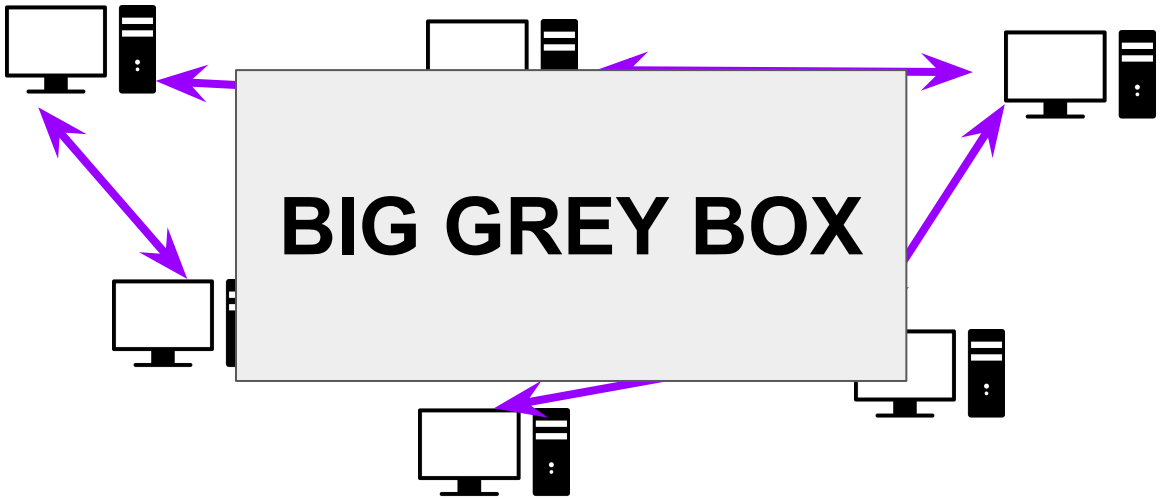
Multiple independent computers work on the same problem at the same time

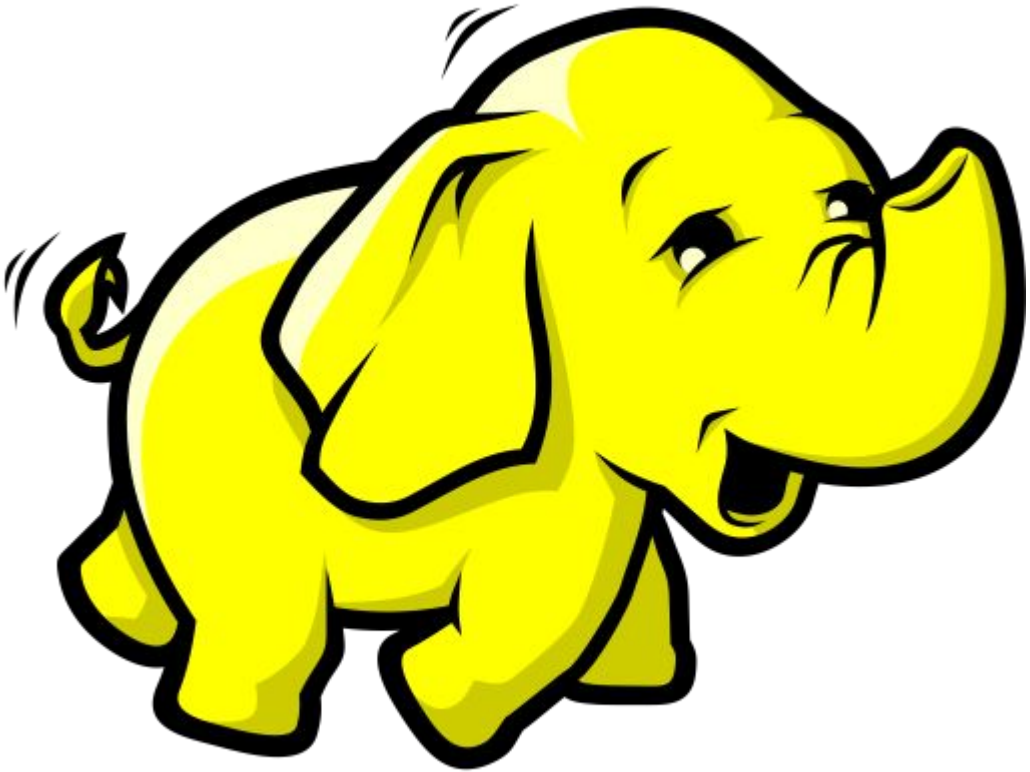
CSC 369: writing **software for solving problems** using existing distributed computing frameworks

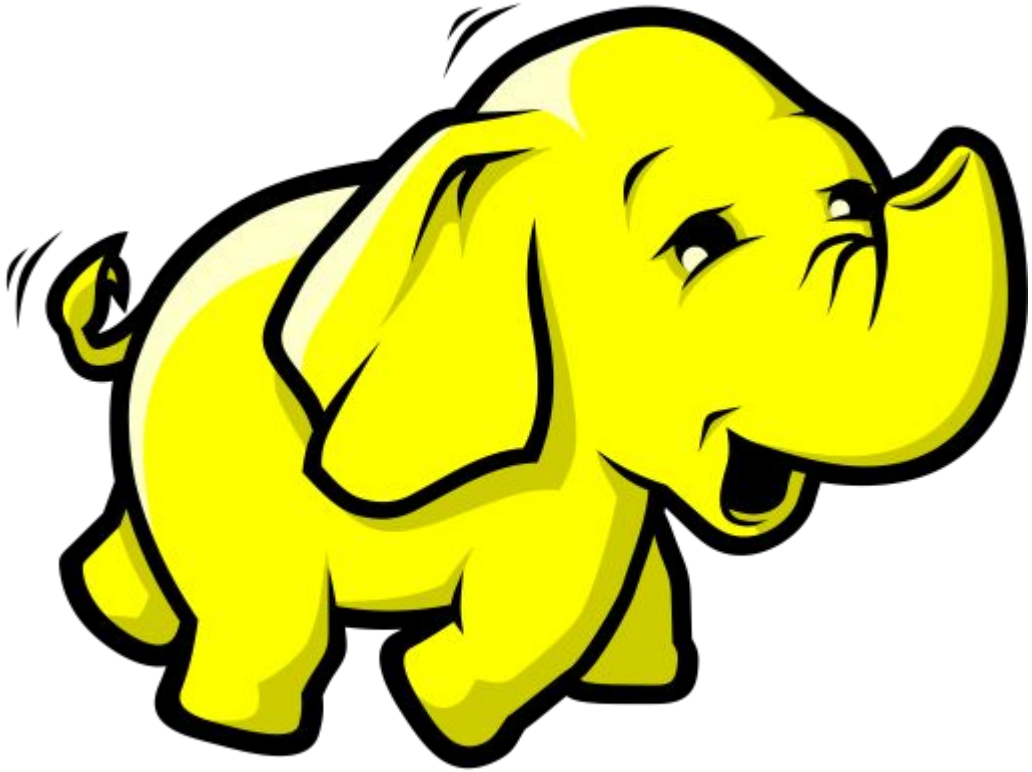
CSC 469: studying how to build distributed computing frameworks

Distributed Computing

CSC 369: writing **software for solving problems** using existing distributed computing frameworks

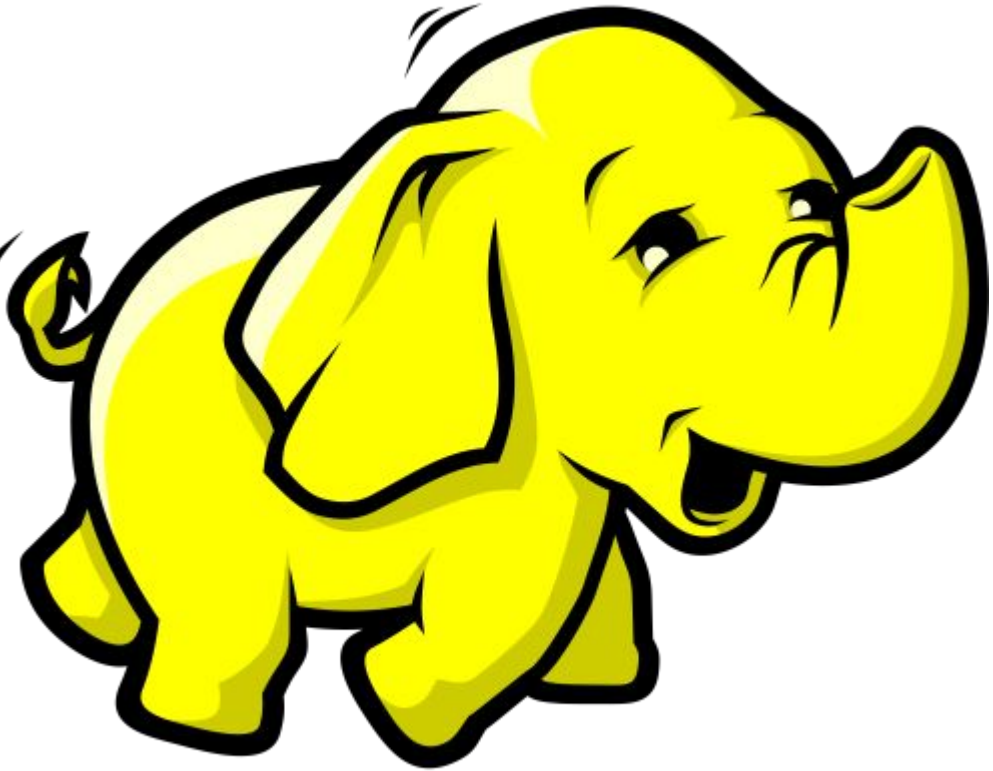






Elephant in the room

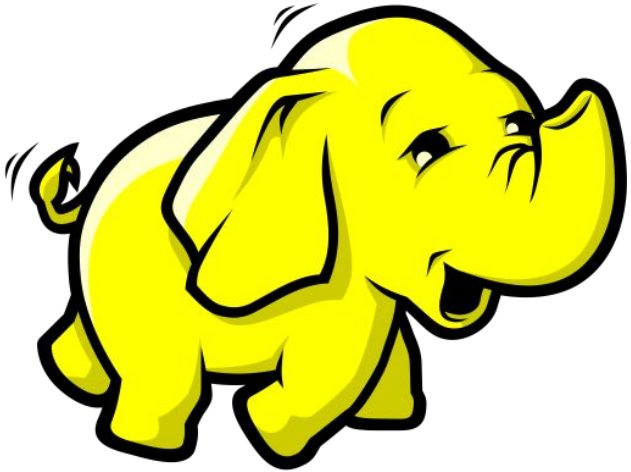
BIG DATA



Elephant in the room

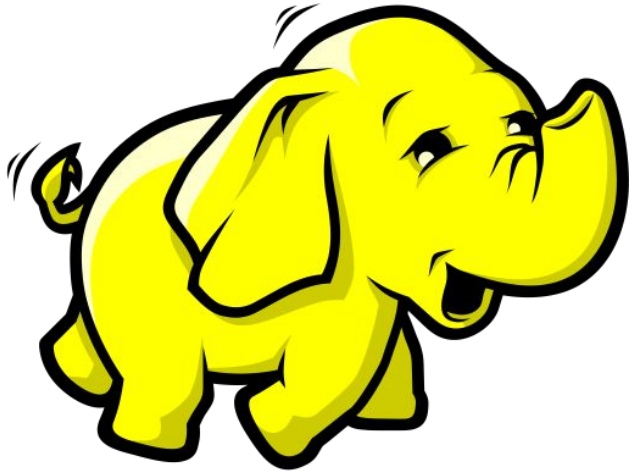
BIG DATA Problems

Big Data = any data collection that is larger than the storage capacity of a single computer system used to process it.



BIG DATA Problems

Big Data = any data collection that is **larger** than the storage capacity of a single computer system used to process it.

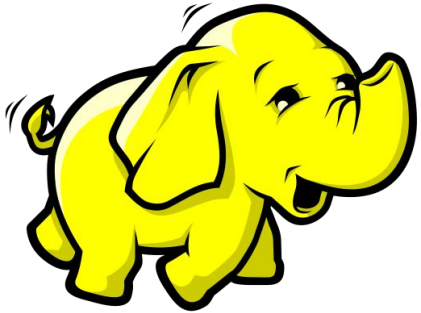


Problems that are easy to solve as small data problems turn out to be difficult as big data problems

This is why we ~~cannot have nice things~~

teach CSC 369

Problems that are easy to solve as small data problems turn out to be difficult as big data problems

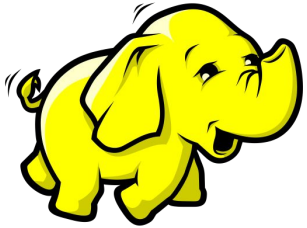


When you have a hammer everything is a nail

I am a “database guy”, so for me “distributed computing problems” = “data management and analysis problems”

Distributed Relational DBMS are not different than regular Relational DBMS and thus are covered in CSC 365

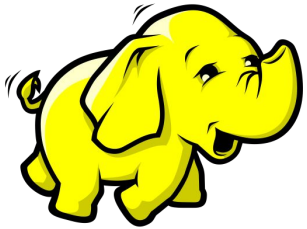
So, we’ll study other distributed frameworks



MongoDB: distributed non-relational document store

Replicates and Shards data

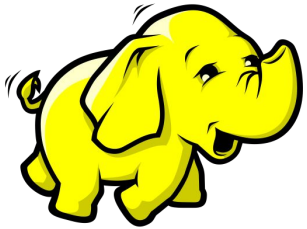
Works with JSON objects



MongoDB: distributed non-relational document store

Replicates and Shards data

Works with JSON objects



Hadoop: open-source implementation of MapReduce framework

MapReduce: distributed computing framework for data processing

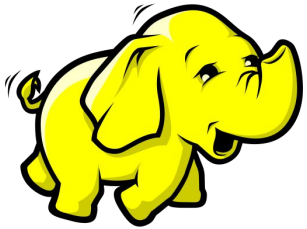
Map: transform data

Reduce: combine information

MongoDB: distributed non-relational document store

Replicates and Shards data

Works with JSON objects



Hadoop: open-source implementation of MapReduce framework

MapReduce: distributed computing framework for data processing

Map: transform data

Reduce: combine information

Spark: lazy evaluation data processing over Hadoop

Resilient Distributed Datasets (RDDs): optimize data processing

Implemented in Scala

PySpark: Python interface to Spark

What Types of Problems?

- Handout #2
- The “Facebook” Example
- The “Google” Example
- The “Twitter” Example
- The “Census” Example
- The “Bioinformatics” Example

I’ll record a 10-15 companion video.

In Lab Today

1. Confirm that everyone has access to ambari-head and MongoDB, change passwords

2. Lab 1: JSON processing