Big Data Management and Analysis

Spring 2022

The City College of New York
NYU Center for Urban Science + Progress

Course Logistics

- Instructor: Huy T. Vo — Office Hours: Thursday, 2-4pm
- Class Format: mixtures of lectures and hands-on labs
  - Be ready to follow along during class
- Expected to be proficient in Python
  - First half of the course is mostly in notebooks/colab
  - Second half will be run on NYU HPC
- Need to familiarize with shell and command lines

Course Objectives

- Preparations for handling Big Data (in a data science team)
  - Broad understanding of big data and its ecosystem
  - Ability to identify big data challenges and how to tackle them
  - Understanding of big data programming paradigm
  - Knowledge in how to perform computation and analytics using big data platforms

Expected Outcomes

- Big Data Computing Concepts
  - Streaming: Python's generators
  - Functional Programming: map/reduce/filter and beyond
- Big Data Programming Paradigm
  - Hadoop: streaming
  - Spark: RDD, DataFrame, and SQL
- Overview of NoSQL

Preparation

- Make sure to install on your computer:
  - Having a google account for using Google Colab
- We will be using NYU HPC
  - I will request access for everyone

Grading

- 4 Assignments: 60% total (15% each)
- 1 In-class Exam: 15%
- 1 Final Challenge (“mini final project”): 25%
  - Solve a big data problem using the techniques learned in class
- Coding Plagiarism is heavily penalized

Textbooks

- No required textbooks: supplemental materials may be distributed during class
- Suggested readings (most available electronically through NYU Library)
  - *Hadoop : The Definitive Guide-Storage and Analysis at Internet Scale, 4th Edition* (O'Reilly Media, Incorporated, 2015) by T. White
  - *PySpark Recipes A Problem-Solution Approach with PySpark2* (Apress, Berkeley, CA, 2018) by R. K. Mishra
  - *Next generation databases : NoSQL, NewSQL, and Big Data* (Apress, Berkeley, CA, 2015) by G. Harrison
  - *Data Science and Big Data Analytics* (John Wiley & Sons, Indianapolis IN, 2015) by EMC Education Services
  - *Probabilistic Data Structures and Algorithms for Big Data Applications* (Books on Demand, 2019) by A. Gakhov

Data Mining vs. Machine Learning vs. Big Data Analytics

- Modern Data Mining requires both ML and BD
- The differences lie in the learning outcomes
  - We focus on the “structuredness” of big data and the principles of how to handle them efficiently
  - See Syllabus for tentative topics
- Hands-on! Hands-on! Hands-on!
  - Coding is required and crucial
  - We focus on Tools and Technologies

Tentative Schedule

Date	Topic
Week 1	Introduction
Week 2	Streaming
Week 3	Distributed File System & Parallel Computing
Week 4	MapReduce & Apache Hadoop
Week 5	Apache Spark
Week 6	Apache Spark
Week 7	Hadoop/Spark on NYU-HPC
Week 8	Spatial Data
Week 9	Textual Data
Week 10	Network and Graph Data
Week 11	NoSQL
Week 12	Review
Week 13	Exam
Week 14	Cloud Computing
Week 15	Final Challenge & Wrap-Up

Question?

Introduction to Big Data

CUNY The City College of New York
NYU Center for Urban Science + Progress

Big Data Analytics — Warehousing Perspective

Further Readings

- Chapter 1 and 2 of *Data Science and Big Data Analytics*

Big Data Characteristics

[Source: IBM, 2012]
[Source: MapR 2014]

Big Data Characteristics

data format are complex
not everything is relational

[Source: IBM, 2012]
[Source: MapR 2014]

Big Data Characteristics

data format are complex
not everything is relational

data are dynamic

[Source: IBM, 2012]
[Source: MapR 2014]

Big Data Characteristics

data format are complex
not everything is relational

data are dynamic

Veracity

Value

[Source: IBM, 2012]
[Source: MapR 2014]

Big Data Characteristics

data format are complex
not everything is relational

data are dynamic

Veracity

Value

Variability

Visualization

[Source: IBM, 2012]
[Source: MapR 2014]

Big Data Characteristics

data format are complex
not everything is relational

data are dynamic

Veracity

Value

Variability

Visualization

[Source: IBM, 2012]
[Source: MapR 2014]

Big Data — Computing

Big Data — Computing

“Big data is an all-encompassing term for any collection of data sets so **large and complex** that it becomes **difficult** to process using **traditional** data processing **applications**.”

— *Wikipedia:Big_Data (05/2015)*

Big Data “Problems”

- Data that is too large:
  - Excel/ArcGIS couldn’t load the data — *need a more scalable tool?*
  - Too many files to perform analysis manually — *need an automated tool?*
- Data that is too complex:
  - Searching the entire parameter space is too expensive for my simulator — *need a distributed algorithm?*
  - Too expensive to run a sentimental analysis model on millions of tweets — *need to approximation technique or to leverage parallelism?*

Background — Research Interests

- developing **scalable** techniques and systems for **large-scale** data **visualization** and **analysis**
  - Streaming algorithms
  - Big data management
  - Parallel and distributed computing

Background — Research Interests

- developing **scalable** techniques:
  - large-scale data **visualization** and **analysis**

Streaming Algorithms

[IEEE TVCG 2007]

Background — Research Interests

- developing **scalable** techniques:
  - large-scale data **visualization** and **analysis**

Streaming Algorithms

[IEEE TVCG 2007]

Background — Research Interests

- developing **scalable** techniques:
  - large-scale data **visualization** and **analysis**

Big Data Management

Out-of-Core Data Structure

[EG PGV 2007]

Spatio-Temporal Indexing

[IEEE TVCG 2013]

Cache-Coherent Mesh Layout

[JGT 2012]

Background — Research Interests

- developing **scalable** techniques:
  - large-scale data **visualization** and **analysis**

Parallel and Distributed Computing

Multi-core, GPGPUs, and clusters of machines
[EG PGV 2012]

Scalable Displays
[VDA 2012, POWERWALL 2013]

MapReduce
[LDAV 2012]

Background — Research Interests

- developing **scalable** techniques:
  - large-scale data **visualization** and **analysis**

Parallel and Distributed Computing

Multi-core, GPGPUs, and clusters of machines
[EG PGV 2012]

Scalable Displays
[VDA 2012, POWERWALL 2013]

MapReduce
[LDAV 2012]

Background — Research Interests

- developing **scalable** techniques:
  - large-scale data **visualization** and **analysis**

Parallel and Distributed Computing

Multi-core, GPGPUs, and clusters of machines
[EG PGV 2012]

Scalable Displays
[VDA 2012, POWERWALL 2013]

MapReduce
[LDAV 2012]

Computer Graphics

Research using Big Data

Scientific Data

27 Billion Voxels (108GB)

1 Billion Triangles (30GB)

Isosurface Extraction

Volume Rendering

Translucency Rendering

Gigapixel Rendering

Simplification

Urban Data

millions of primitives and visibility tests

hundreds of millions of records

Our Main Objectives

- Improve the performance and quality of Visual Analytic applications, with a focus on *“big” cities, i.e.:*
  - Big (spatio-temporal) data
  - Urban domain (urban science/informatics)
- We make applications:
  - Efficient, and scalable
  - Easy to use for non-experts
  - Can be developed rapidly by domain experts

Examples of Past CUSP Capstone Projects with *Big Data*

Examples of CUSP Co-sponsored Capstone Projects

- 12 data sets
- 40 GB++
- Real-time queries

Piercing the Veil of the Corporate Landlord
(with NYSOAG)

Fighting harmful landlord practices to harass rent-stabilize tenants

Examples of CUSP Co-sponsored Capstone Projects

~150K drivers in NYC, only 69 Relief Stands…

Image capture: Jun 2016 © 2017 Google United States

High violations, drop offs, cheap restaurants

Optimizing Taxi Relief Stands
(with TLC)
300GB, 4 billion data points

Examples of CUSP Co-sponsored Capstone Projects

NYC Bus Profile Dashboard
(with TransitCenter)

A performance measurement tool for riders, agency board members, and staff — data driven

1TB of data

Big Data is not just about size!

- Scalability for batch computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Data were “in good hands”

Big Data is not just about size!

- Scalability for batch computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Data were “in good hands”
- Scalability for people is!
  - Data owners are non-expert

Big Data is not just about size!

- Scalability for batch computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Data were “in good hands”
- Scalability for people is!
  - Data owners are non-expert

Big Data is not just about size!

- Scalability for batch computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Data were “in good hands”
- Scalability for people is!
  - Data owners are non-expert

regardless of whether data are big or small

Big Urban Data — Space + Time

- Several hundred millions to billions of records
- Spatial + Temporal attributes
  - Can have >1 spatial and temporal attributes
- Main objectives in urban informatics: discovering trends and anomalies; and detecting and predicting events
  - Needs to query data intensively
- However, “canned” queries are difficult to frame
  - Needs interactive visualization to explore big urban data

Exploring NYC Taxis

Taxis are ***sensors*** that can provide unprecedented insight into city life: economic activity, human behavior, mobility patterns, …

"What is the average trip time from Midtown to the airports during weekdays?''

"How the taxi fleet activity varies during weekdays?"

"How was the taxi activity in Midtown affected during a presidential visit?''

"How did the movement patterns change during Sandy?"

"Where are the popular night spots?"

[TVCG 2013]

Exploring NYC Taxis

7-8am

8-9am

9-10am

10-11am

NYC Taxi Visualization
Mon 05/02/11 08:13 AM to Mon 05/09/11 06:28 AM
Regular Selection
Recurrent Selection
Start Time
Step Size
End Time
Mon 05/02/11 08:13
1 hour
Mon 05/09/11 06:28
Query
Explore...
Jersey City
Hoboken
Weehawken
Croxton Intermodal Terminal
CSX South Kearny Yard
Liberty Science Center
Ellis Island
Liberty State Park
Whitney Museum of American Art
Newport Centre
Brooklyn Bridge
Flushing Meadows Corona Park
Mt Olivet Cemetery
Map data ©2012 Google Terms of Use
Time Serie
Histogram
ScatterPlot
Attribute:
Num Trips
Num Bins:
96
Number of trips per time
Thu May 5 17:01:08 2011
Num Trips
60000
50000
40000
30000
20000
10000
06:20 May 03
03:33 May 04
01:46 May 05
00:00 May 06
22:13 May 06
20:26 May 07
18:40 May 08

NYC Taxi Visualization
Mon 05/02/11 08:13 AM to Mon 05/09/11 06:28 AM
Regular Selection
Recurrent Selection
Start Time
Step Size
End Time
Mon 05/02/11 08:13
1 hour
Mon 05/09/11 06:28
Query
Explore...
Jersey City
Hoboken
Weehawken
Croxton Intermodal Terminal
CSX South Kearny Yard
Liberty Science Center
Ellis Island
Liberty State Park
Whitney Museum of American Art
Newport Centre
Brooklyn Bridge
Flushing Meadows Corona Park
Mt Olivet Cemetery
Map data ©2012 Google Terms of Use
Time Serie
Histogram
ScatterPlot
Attribute:
Num Trips
Num Bins:
96
Number of trips per time
Thu May 5 17:01:08 2011
Num Trips
60000
50000
40000
30000
20000
10000
06:20 May 03
03:33 May 04
01:46 May 05
00:00 May 06
22:13 May 06
20:26 May 07
18:40 May 08

TaxiVis: Mobility

TaxiVis: Comparing Neighborhoods

Exploring the Effect of Major Events: Sandy

Life of a Taxi in a Day

[IEEE Big Data 2013]

Life of a Taxi in a Day

[IEEE Big Data 2013]

Thank you!