# CS 6350 BIG DATA MANAGEMENT & ANALYTICS

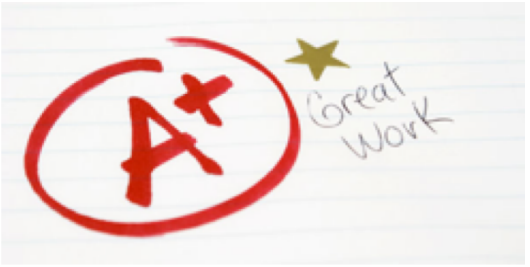
Anurag Nagar

# ADMINISTRATIVE ISSUES

- Syllabus is uploaded
- Extensive use of eLearning
- Use of Piazza for discussion
- Please respect class times, office hours, policies.
- Use Piazza for sending message to instructor, or if you send email mention class and section number in subject:

e.g. CS 6350.00x – Assignment 1 doubts

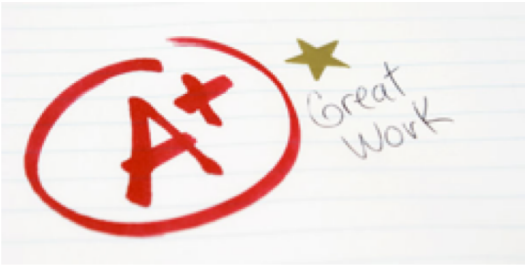
# GRADING




- Weightage:

25% Homework
10% Project
25% Midterm
25% Final
15% Quizzes & Class Participation

# GRADING POLICIES




- Collaboration:
  - Assignments:
  Pair programming => you can team up with another student from same class and section
  * many benefits
  - Project:
  You can work in a teams of 1 – 4 students
  * allows you to build a better product that you can be proud of and display to employers.

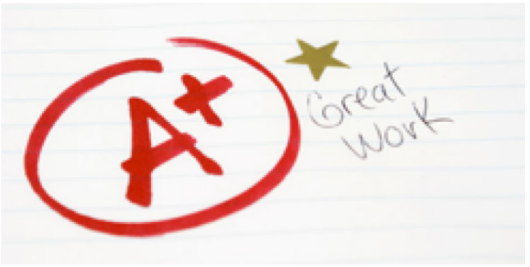
# GRADING


## How is your grading??

Based on following ideas:

- Grade on a curve – relative
- Fair to everyone - no exceptions - nobody gets preferential treatment
- Reward for regular class attendance & participation
- Not a way to punish or give free rides to students
- Students are given freedom to choose projects, experiment with various technologies, libraries and packages

# HOW TO SUCCEED




- Keep up with the class. Don't get behind.
- Come prepared to class. Read the previous topics from reference material and textbook.
- Bring your laptop to class. There will be surprise quizzes and in-class assignments.
- Start your projects and assignments early. Do not wait till the last minute. Remember, there will be no extension of deadlines under any circumstances.
- Choose your projects according to your interest. Once chosen, topic cannot be changed.

# TEXTBOOKS

- B1: Jimmy Lin and Chris Dyer, Data-Intensive Text Processing with MapReduce, Morgan & Claypool Publishers, 2010.
  https://lintool.github.io/MapReduceAlgorithms/ Free Download
- B2: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Addison-Wesley April 2005. Optional
  https://www-users.cs.umn.edu/~kumar001/dmbook/index.php
- B3: Anand Rajaraman and Jeff Ullman, Mining of Massive Datasets, Cambridge Press, http://www.mmds.org/ Free Download
- B4: Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Cambridge University Press, Optional
  http://hanj.cs.illinois.edu/book
- B5: Tom White, Hadoop: The Definitive Guide, O'Reilly Media 4th edition, April 2015 Available as eBook through UTD library
  http://shop.oreilly.com/product/0636920033448.do
- • B6: Matei Zaharia, Spark: The Definitive Guide, O'Reilly Media 1st edition, Feb 2018 Available as eBook through UTD library
  http://shop.oreilly.com/product/0636920034957.do

# CLASS POLICIES

- Late Submission Policy:

  - Every student has 4 free days that can be used over the entire semester. You can use at most 2 days for any one assignment/project.
  After you have used up the 4 free days, there will be a penalty of 10% per late day.

  - All submissions close 2 days after the deadline.

  - Deadlines cannot be extended in any situation. Please start your work early.

# CLASS POLICIES

- Classroom Quiz and Lab Policy:

  - Quizzes will be administered through eLearning. You are required to bring your laptop or other electronic device. Also, it is up to you to manage your computer. You should ensure that you have sufficient charge on your battery, your computer doesn't start updating itself, etc.

  - You have to be on UTD Wifi (CometNet) only. Your IP address will be checked and if it's not from within UTD, you will not get any credit.

  - Sometimes quizzes will be collaborative in which we solve questions together.

# CLASS POLICIES

- Exam Policy:

- The date and time of midterm and final exam are mentioned in the syllabus. You can only take exams during that period. The only exceptions are:
1. You have a medical emergency certified by a doctor
2. You have 3 exams on the same day (sorry 2 exams don't count)
3. There is a time conflict with other class or exam

This is according to the official UTD policy

- Exams will be comprehensive. The instructor will provide a list of important topics before the exam.

# CLASS POLICIES

- Communication Policy:

- All administrative announcements will be posted on eLearning. You should automatically get an email when it is posted.

- We will use Piazza discussion board. You should use it for getting help on assignments or projects. Technical questions can be asked there. Your peers, TA, or the instructor will try to answer them.

- Individual communication will be sent to your UTD email only.

- All notes, assignments, and projects will be posted on eLearning.

# CLASS POLICIES

- How to get help:

For getting help, you can use the following resources in order of preference:

1. Research online, including forums such as stackoverflow, etc

2. Through discussion with your peers via Piazza. Remember, you are free to discuss concepts and ideas, but the final submission must be your own work.

3. In case option 1 and 2 don't work, contact the TA for help.

4. In case options any of the above options don't work, contact the instructor.

# CLASS POLICIES

- How to contact the instructor:

  - Walk-in during office hours.

  - For technical questions related to assignments, use Piazza. I will try to answer most questions there.

  - Via email. I normally reply to emails promptly during normal business hours.

  - If none of the above work, you can set up an appointment during business hours.

# CLASS POLICIES

- Academic Honesty:

  - Academic honesty is taken very seriously at UTD.

  - The official policy is listed here.

  - Do not engage in academic dishonesty. It can land you in serious trouble. Consequences at UTD are mentioned here.

  - Any act of dishonesty will be reported to the department.

# CLASS POLICIES

- Classroom Citizenship:

- Please be responsible citizens and professionals.

- Within and outside the classroom, be respectful to each other and the instructor.

- This class is very interactive and you are encouraged to ask questions. If you still don't get a concept or need additional help, please see the instructor after class.

- Do not argue unnecessarily about your score or grade. Harassing the TA or instructor can get you in trouble.

- Remember, there cannot be individual exceptions or preferential treatments towards anyone.

# NON-NEGOTIABLE


- Date of midterm & final

** Exception in case of verified sickness**

- Final grade after posting

** We get requests, for reasons such as internship, visa, jobs, TA requirement, etc, they will not be entertained. **

# COMPUTING RESOURCES FOR BIG DATA

# COMPUTING RESOURCES FOR HDFS

- This class is very practical and requires you to work hands-on with various Big Data computing technologies.
- You will be provided with a list of software and installation instructions.
- **You** are responsible for software installation, troubleshooting, and debugging.
- If you run into problem or have issues, please contact the **Teaching Assistant (TA).**
- The instructor can only provide instructions and hints for resolving issues, but cannot perform installation for you.
- For cloud computing resources, you are responsible for managing your account including billing. **The instructor or the department is not liable for any financial charges.**

# COMPUTING RESOURCES FOR HDFS

We will start with Hadoop Distributed File System (HDFS)

How to get access?

- UTD Cluster – available through node cs6360.utdallas.edu (Accessible from within campus)
- AWS EMR - https://aws.amazon.com/emr/
- Google Cloud - https://cloud.google.com/hadoop/
- Microsoft Azure

# COMPUTING RESOURCES FOR APACHE SPARK

Next will come Apache Spark

How to get access?

- UTD Cluster – available through node cs6360.utdallas.edu (Accessible from within campus)
- AWS EMR - https://aws.amazon.com/emr/
- Google Cloud - https://cloud.google.com/hadoop/
- Databricks community edition – https://databricks.com/try-databricks

# COMPUTING RESOURCES FOR NOSQL

- We will discuss in class
- For Cassandra, https://medium.com/@sreekar.anugu/spinning-up-a-cassandra-cluster-on-google-cloud-for-free-with-just-a-browser-8dd5f64b426d

# NO EXCUSES

Saying that the UTD cluster is down or slow is NOT a valid excuse

# REWARDS

# BIG DATA (BD)…


- What will I learn?
- Do I really need BD?
- Salary?

**AVERAGE SALARY FOR High Paying Skills and Experience**

| SKILL | 2013 | YR/YR CHANGE |
|---|---|---|
| R | $ 115,531 | n/a |
| NoSQL | $ 114,796 | 1.6% |
| MapReduce | $ 114,396 | n/a |
| PMBok | $ 112,382 | 1.3% |
| Cassandra | $ 112,382 | n/a |
| Omnigraffle | $ 111,039 | 0.3% |
| Pig | $ 109,561 | n/a |
| SOA (Service Oriented Architecture) | $ 108,997 | -0.5% |
| Hadoop | $ 108,669 | -5.6% |
| Mongo DB | $ 107,825 | -0.4% |

# Also, Scala, ML using Spark

## *Is Apache Spark the Next Big Thing in Big Data?*

R. Emmett O'Ryan  03/12/2014  Big Data  4

In any article or blog post, any mention of Big Data usually includes something about Hadoop. When it comes to Big Data, Apache Hadoop has been the big elephant in the room, and the release of Hadoop 2.0 in 2013 made the environment easier and more stable. But even with the inclusion of Impala for querying stored information real-time, Hadoop is still a batch-based system that processes data in, well, batch mode.

Big Data processing is said to have three main characteristics, a.k.a. the 3Vs: volume, velocity and variety. Many data scientists and Big Data engineers will argue that Apache Hadoop processes data fast enough to meet the criteria of velocity. Perhaps so. However, coming from a real-time and a high-performance computing background, this argument about Hadoop and

ECONOMIC SNAPSHOT MAR 19

# Want to start out making $100K? Take a look at these tech jobs

## DALLAS

**Starting salary**

1. Data warehouse developer: $108,953
2. Hadoop: $108,953
3. ETL developer: $106,296
4. SAP apps: $106,296
5. Program manager: $106,296

**Senior salary**

1. Agile coach: $156,224
2. Data scientist: $146,587
3. Vendor risk auditor: $143,553
4. Data warehouse developer: $143,500
5. Hadoop: $143,500

Covered in this class

Source: https://www.dallasnews.com/business/economic-snapshot/2018/03/19/want-tostart-making-100k-take-look-tech-j

# WHAT BACKGROUND DO I NEED?

- Refresh your UNIX skills. Need to know at least:
  - How to SSH into a server
  - How to copy/move files
  - wget command

- Refresh your basic and advanced Java skills
  - Understand concept of streams
  - Class methods and signatures

- Functional Programming. We will be using Scala later on in the course

- SQL knowledge for SQL on Hadoop databases e.g. Hive, Impala, etc