CS226
Big-Data Management
Instructor:
Ahmed Eldawy
Welcome (back) to UCR!
Class information

- Classes: MWF 10:00 – 10:50 AM
- Zoom Link: on iLearn. Do not distribute.
- Instructor: Ahmed Eldawy
- TA: Akil Sevim
- Office hours: MW 11:00 – 11:50 AM
  - iLearn (Any UCRX students?)
- Piazza: [https://piazza.com/ucr/fall2020/cs226](https://piazza.com/ucr/fall2020/cs226)
- Email: [eldawy@ucr.edu](mailto:eldawy@ucr.edu)
  - Subject: “[CS226]”
Course work

- Active participation in class (5%)
- Reading and review tasks (10%)
- Assignments (20%)
- Mid-terms (15%)
- Project (50%)
Project

• Groups of 4-5 students
• Milestones
  ▪ Group Selection
  ▪ Project proposal (5%)
  ▪ Project proposal presentation (10%)
  ▪ Literature survey (10%)
  ▪ Report outline (5%)
  ▪ Final report (10%)
  ▪ Final presentation (10%)
Project: Coordination between CS 226 and CS 225

- You can share project groups between CS 226 (Big Data Management) and CS 225 (Spatial Computing), given:
  - You work on one project with double in size, this gives opportunity to focus on one big project for the two courses
  - All group members must be taking the two courses, except one member at maximum.
  - The project must have a spatial component and a big data component
  - The team must submit two separate reports, one for each course. Each report must focus on the component relevant to the course.

- Example projects:
  - **City ranking**: incorporate spatial factors in ranking cities for quality of life
  - **Satellite imagery analysis**: use large satellite images datasets in any societal application
  - **Contact tracing for epidemics control**: use spatial data to trace contacts to patients of COVID-like epidemics
WHAT’S GIS?

GIS stands for Geographic Information system. The software is used for both gathering and visualizing data. The most common example that is applicable to our lives is Google maps. Whenever you look up directions or try to figure out where you are, that’s the byproduct of GIS.

GOAL OF THE CLUB:

R’geospatial is a relatively new club on campus that aims to show the utility and transferability of GIS skills to students’ careers. The club will cover how to use the software and also how it applies to your major and future professional endeavors.

If interested, please add us on Highlanderlink and feel free to reach out to us if you have any questions!
Survey & Breakout
Course goals

• What are your goals?

• Understand what big data means
• Identify the internal components of big data platforms
• Recognize the differences between different big data platforms
• Explain how a distributed query runs on big data
Ant-Man/Wasp

Get smaller to understand how ants work and what they are capable of.

Use this knowledge to control thousands of ants and do amazing things!

Optional task: Watch Ant-Man and the Wasp this weekend 😊
Big-data Expert

• Understand how the big-data platforms really work
• Control those thousands of processors efficiently to carry out your task
Syllabus

• Overview of big data
• Big-data storage
• Big-data processing
• Big-data indexing
• Big-SQL processing
• Programming packages
Introduction
Big Data

Straight Ahead
All of the information

Information you need!
Big Data, Big Impact: New Possibilities for International Development

The amount of data in the world is exploding - large portion of this comes from the interactions over mobile devices being used by people in the developing world - people whose needs and habits have been poorly understood until now. Researchers and policymakers are beginning to realize the potential for channeling these torrents of data into actionable information that can be used to identify needs & provide services for the benefit of low-income populations. This discussion note is a Call-to-action for stakeholders for concerted action to ensure that this data helps the individuals and communities who create it.
Interest in Big Data in the US

- **March 2012**: Obama administration unveils **BIG DATA** initiative: $200 Million in R&D investment

- **June 2013**: Washington Post is calling Obama “The Big Data President”

---

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a “Big Data Research and Development Initiative.” By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some of the Nation’s most pressing challenges.
Interest in Big Data in Europe

• **March 2014:** David Cameron and Angela Merkel talking about Big Data in a Computer Expo in Hannover, Germany
The creation and consumption of data continues to grow by leaps and bounds and with it the investment in big data infrastructure, software, and services.
Four Three V’s of Big Data

**Volume**

- 40 Zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
- 6 billion people have cell phones.

**Velocity**

- The New York Stock Exchange captures 1 terabyte of trade information during each trading session.
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.
- By 2016, it is projected there will be 18.9 billion network connections – almost 2.5 connections per person on earth.
- By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.5 million in the United States.

**Variety**

- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 billion gigabytes).
- 30 billion pieces of content are shared on Facebook every month.
- 400 million tweets are sent per day by about 200 million monthly active users.

**Veracity**

- 27% of respondents in one survey were unsure of how much of their data was inaccurate.
- By 2014, it's anticipated there will be 420 million wearable, wireless health monitors.

**The Four V’s of Big Data**

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety, and Veracity.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, ICOTEQ, GAO
Big Data Vs Big Computation

• Full scans (e.g., log processing)
• Range scans
• Point lookups
• Iterations
• Joins (self, binary, or multiway)
• Proximity queries
• Closures and graph traversals
Big Data Applications

• Web search
• Marketing and advertising
• Data cleaning
• Knowledge base
• Information retrieval
• Internet of Things (IoT)
• Visualization
• Behavioral studies
Publicly Available Datasets

• Data.gov
• UCR Star [https://star.cs.ucr.edu]
  ▪ facebook.com/ucrstar & 👍
• Twitter Streaming API
• Yahoo! Webscope [http://webscope.sandbox.yahoo.com/]
• GDELT [http://www.gdeltproject.org/]
• Instagram API
Big Data Landscape 2014
Big Data Landscape 2018
Components of Big Data
Storage of Big Data

• Data is growing faster than Moore’s Law
• Too much data to fit on a single machine
• Partitioning
• Replication
• Fault-tolerance
Hadoop Distributed File System (HDFS)

- The most widely used distributed file system
- Fixed-sized partitioning
- 3-way replication
- Write-once read-many
- See also: GFS, Amazon S3, Azure Blob Store
Indexing

• Data-aware organization
• Global Index **partitions** the records into blocks
• Local Indexes organize the records in a partition
• Challenges:
  ▪ Big volume
  ▪ HDFS limitation
  ▪ New programming paradigms
  ▪ Ad-hoc indexes
Fault Tolerance

• Replication

• Redundancy

• Multiple masters
Streaming

- Sub-second latency for queries
- One scan over the data
- (Partial) preprocessing
- Continuous queries
- Eviction strategies
- In-memory indexes
Task Execution

- MapReduce
  - Map-Shuffle-Reduce
  - Resiliency through materialization

- Resilient Distributed Datasets (RDD)
  - Directed-Ayclic-Graph (DAG)
  - In-memory processing
  - Resiliency through lineages

- Hyracks
- Stragglers
- Load balance
Query Optimization

• Finding the most efficient query plan
• e.g., grouped aggregation

• Cost model (CPU – Disk – Network)
Provenance

• Debugging in distributed systems is painful

• We need to keep track of transformations on each record
Big Graphs

- Motivated by social networks
- Billions of nodes and trillions of edges
- Tens of thousands of insertions per second
- Complex queries with graph traversals
Hadoop Ecosystem

- Pig
- Hive
- Giraph
- Mahout
- Apache Ambari
- Administration
- MapReduce Query Engine
- Yet Another Resource Negotiator (YARN)
- Hadoop Distributed File System (HDFS)
Spark Ecosystem

- Spark SQL
- Data Frames
- MLlib
- GraphX
- SparkR
- Spark Streaming

Resilient Distributed Dataset (RDD) a.k.a Spark Core

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)

Kubernetes
Impala

Query Parser

Query Planner

Query Executor

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)
SpatialHadoop

- Pig Latin + Pigeon
- Spatial Visualization
- MapReduce Processing + Spatial Query Processing
- Yet Another Resource Negotiator (YARN)
- Hadoop Distributed File System (HDFS) + Spatial Indexing
Reading Material

• “The Age of Analytics in a Data-driven World” [Executive Summary] by McKinsey & Company