CS226
Big-Data Management
Instructor: Ahmed Eldawy
Class information

- Classes: Tuesday and Thursday 9:40 AM – 11:00 AM at WCH 142
- Instructor: Ahmed Eldawy
- Office hours: Tuesday and Thursday 11:00 AM – 12:00PM @357 WCH. Conflicts?
- Website: http://www.cs.ucr.edu/~eldawy/18WCS226/
- Email: Eldawy@ucr.edu
- Subject: “[CS226] …”
Course work

- Active participation in the class (10%)
- Class presentation (10%)
- Reading and review tasks (5%)
- Assignments (15%)
- Project (35%)
- Final exam (25%)
  - Date: Thursday, March 22, 2018
  - Time: 8:00 a.m. - 11:00 a.m.
  - Location: TBD
Project

- Groups of 2-3 students
- Project proposal. Due on Thursday, 1/25
- Literature survey. Due on Thursday, 2/8
- Paper outline. Due on Thursday, 2/22
- Final paper. Due on Thursday, 3/8
- Project presentation. On 3/13 and 3/15

01/09/2018
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
Super Hero
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 - a 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

**Office of Science and Technology Policy**
Executive Office of the President
New Executive Office Building
Washington, DC 20502

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 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 market, hardware, software and services and in data analytics tools.
Four Three V’s of Big Data

**Volume**
- 40 Zettabytes (45 quadrillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005
- 6 billion people have cell phones
  - World population: 7 billion

**Velocity**
- The New York Stock Exchange captures 1 TB 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

**Variety**
- Most companies in the U.S. have at least 100 terabytes (100 billion gigabytes) of data stored
- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (161 quadrillion gigabytes)
- By 2014, it’s anticipated there will be 420 million wearable, wireless health monitors
- 4 billion+ hours of video are watched on YouTube each month
- 400 million tweets are sent per day by about 200 million monthly active users

**Veracity**
- As of 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States
- 27% of respondents in one survey were unsure of how much of their data was inaccurate
- Poor data quality costs the US economy around $3.1 trillion a year

*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.*

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, ISOPECC, QAG
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
- Data.gov.uk
- Twitter Streaming API
- Yahoo! Webscope [http://webscope.sandbox.yahoo.com/]
- GDELT [http://www.gdeltproject.org/]
- Instagram API
Big Data Landscape 2014

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

01/09/2018
Hadoop Distributed File System (HDFS)

- The most widely used distributed file system
- Fixed-sized partitioning
- 3-way replication
- Write-once read-many
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-Acyclic-Graph (DAG)
  - In-memory processing
  - Resiliency through lineages
- Hyracks
- Stragglers
- Load balance

01/09/2018
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
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)
AsterixQL -> AsteixDB
HiveQL -> HiveSterix
PigLatin -> Other compilers

MapReduce Jobs -> Hadoop MapReduce Compatibility
Pregel Jobs -> Pregelix

Hyracks jobs

Hyracks Data-parallel Platform
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