EECS E6893 Big Data Analytics Lecture 1:

Overview of Big Data Analytics

Ching-Yung Lin, Ph.D.
Adjunct Professor, Depts. of Electrical Engineering and Computer Science
IEEE Fellow

September 10th, 2021
Definition and Characteristics of Big Data

“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” -- Gartner

which was derived from:

“While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.” – Doug Laney
What made Big Data needed?

- Innovative techniques
- Cost effective
- Lowered barrier to entry and success
- Increased data volumes
- Increased computation need
- Increased analytics need

“Big Data Analytics”, David Loshin
Key Computing Resources for Big Data

- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network
Scalability — Scale Up & Scale Out

- **Scale out**
  - Use more resources to distribute workload in parallel
  - Higher data access latency is typically incurred
- **Scale up**
  - Efficiently use the resources
  - Architecture-aware algorithm design

Example: Resource utilization for a large production cluster at Twitter data center

- For independent data ==> scale up may not have obvious advantage than scale out
- For linked data ==> utilizing scale up as much as possible before scale out

### Contrasting Approaches in Adopting High-Performance Capabilities

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Typical Scenario</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application development</td>
<td>Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning</td>
<td>A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing</td>
</tr>
<tr>
<td>Platform</td>
<td>Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices</td>
<td>Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology</td>
</tr>
<tr>
<td>Data management</td>
<td>Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts</td>
<td>Alternate models for data management (often referred to as NoSQL or “Not Only SQL”) provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics)</td>
</tr>
<tr>
<td>Resources</td>
<td>Requires large capital investment in purchasing high-end hardware to be installed and managed in-house</td>
<td>The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line</td>
</tr>
</tbody>
</table>

“Big Data Analytics”, David Loshin
Techniques towards Big Data

• Massive Parallelism
• Huge Data Volumes Storage
• Data Distribution
• High-Speed Networks
• High-Performance Computing
• Task and Thread Management
• Data Mining and Analytics
• Data Retrieval
• Machine Learning
• Data Visualization

→ Techniques exist for years to decades. Why is Big Data hot now?
Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud
Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

→ 

- High-Volume
- High-Velocity
- High-Variety

→ Artificial Intelligence
https://www.youtube.com/watch?v=BV8qFeZxZPE
Human brain is a graph/network of 100B nodes and 700T edges.

- **Machine Cognition:**
  - Robot Cognition Tools
  - Feeling

- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools

- **Machine Learning:**
  - Machine Learning Tools
  - Deep Learning Tools

- **Graph Analytics:**
  - Network Analysis
  - Matching and Search
  - Flow Prediction

- **Graph Visualization:**
  - Dynamic Graph
  - Big Graph

- **Graph Database:**
  - Large-Scale Native Store
The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

http://hadoop.apache.org
Four distinctive layers of Hadoop

- API: MapReduce, Pig, Hive, HBase
- Processing Framework: MapReduce v1
- Resource Management
- Distributed Storage: HDFS
Course Main Thrust 2: Apache Spark and ML

Apache Spark™ is a unified analytics engine for large-scale data processing.

**Speed**

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Logistic regression in Hadoop and Spark
Main Spark Stack

Spark SQL
structured data

Spark Streaming
real-time

MLib
machine
learning

GraphX
graph
processing

Spark Core

Standalone Scheduler

YARN

Mesos
Course Main Thrust 3: Streaming Big Data Analytics

[Diagram showing market intelligence and news ranking list with keywords like Corporate, Stock Outlook, Macroeconomics, and Industry.]
Course Main Thrust 4: Big Data Visualization
Human brain is a graph of 100B nodes and 700T edges.
Course Main Thrust 6: Big Data System and AI Solutions

- Big Data Pipeline
- Big Data and AI for Finance
- Big Data and AI for Healthcare
Big Data AI Platform Example: Graphen Ardi

Ardi’s 8 Components

- Graph Database
- Relational Database
- Graph Analytics
- Feature Engineering
- Causality Modeling
- Behavior Prediction
- Deep Language Understanding
- Deep Video Understanding
- Machine Learning
- Deep Learning
- Autonomous Model Optimization
- Production Workflow
- Action Strategy Simulation
Why you want to take this class

• **Key Differentiator of this class:** Focusing on building a full-spectrum understanding of the latest Big Data Analytics technologies and using them to build real industry real-world solutions.

• **Sapphire Big Data Analytics Open Source Applications:** Create a Big Data open source toolsets for various industries (and disciplines)

• **Dataset and Use Cases:** Welcome!!
### Course Grading

- **5 Homewor[s: 50%**
  
  -- **Individual work**: Language Requirement: Python, JavaScript; Get familiar with Linux
  
  -- Report (including description of the work, discussions, experiments, etc) and source code
    - HW #0: Big Data Environment Setup and Testing
    - HW #1: Big Data Analytics and Machine Learning
    - HW #2: Streaming Big Data Analytics
    - HW #3: Big Data Analytics Visualization
    - HW #4: Linked Big Data Analytics
Course Grading

- Final Project: 50%
  -- Teamwork: 2 - 3 students per team (on campus); 1 - 3 students per team for CVN
  - **Proposal** (slides — short presentation in the class)
  - **Progress Presentation** (slides — short presentation in the class)
  - **Progress Report** (report)
  - **Final Report** (paper, up to 10 pages)
  - **Workshop Presentation** (Oral and Demo)
  - **Open Source Codes**
  - **Video Presentation** (on YouTube)
Course Information

▪ Website:

http://www.ee.columbia.edu/~cylin/course/bigdata/

▪ Textbook:

-- None, but reference book(s) and/or articles/papers will be provided each lecture.
### Course Outline

<table>
<thead>
<tr>
<th>Class Date</th>
<th>Class Number</th>
<th>Topics Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>09/10/21</td>
<td>1</td>
<td>Introduction of Big Data Analytics</td>
</tr>
<tr>
<td>09/17/21</td>
<td>2</td>
<td>Big Data Platforms and Data Storage</td>
</tr>
<tr>
<td>09/24/21</td>
<td>3</td>
<td>Big Data Analytics Algorithms I</td>
</tr>
<tr>
<td>10/01/21</td>
<td>4</td>
<td>Big Data Analytics Algorithms II</td>
</tr>
<tr>
<td>10/08/21</td>
<td>5</td>
<td>Real-Time Stream Analysis</td>
</tr>
<tr>
<td>10/15/21</td>
<td>6</td>
<td>Big Data Visualization I</td>
</tr>
<tr>
<td>10/22/21</td>
<td>7</td>
<td>Big Data Visualization II</td>
</tr>
<tr>
<td>10/29/21</td>
<td>8</td>
<td>Linked Big Data Analysis and Graph Computing I</td>
</tr>
<tr>
<td>11/05/21</td>
<td>9</td>
<td>Final Project Proposal Presentation</td>
</tr>
<tr>
<td>11/12/21</td>
<td>10</td>
<td>Linked Big Data Analysis and Graph Computing II</td>
</tr>
<tr>
<td>11/19/21</td>
<td>11</td>
<td>Final Project Progress Proposal</td>
</tr>
<tr>
<td>11/26/21</td>
<td></td>
<td>Thanksgiving Holiday</td>
</tr>
<tr>
<td>12/03/21</td>
<td>12</td>
<td>Big Data Analytics Applications -- AI Finance</td>
</tr>
<tr>
<td>12/10/21</td>
<td>13</td>
<td>Big Data Analytics Applications -- AI Medical</td>
</tr>
<tr>
<td>12/17/21</td>
<td>14</td>
<td>Big Data Analytics Workshop</td>
</tr>
</tbody>
</table>
### Assignments and Submissions

<table>
<thead>
<tr>
<th>Class Date</th>
<th>Assignment</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>09/10/21</td>
<td>HW #0 Big Data Environment Setup and Testing [tutorial]</td>
<td>HW #0</td>
</tr>
<tr>
<td>09/17/21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09/24/21</td>
<td>HW #1 Big Data Analytics and Machine Learning [assignment][tutorial]</td>
<td></td>
</tr>
<tr>
<td>10/01/21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/08/21</td>
<td>HW #2 Streaming Big Data Analytics [assignment] [tutorial]</td>
<td>HW #1</td>
</tr>
<tr>
<td>10/15/21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/22/21</td>
<td>HW #3 Big Data Visualization [assignment][tutorial]</td>
<td>HW #2</td>
</tr>
<tr>
<td>10/29/21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/05/21</td>
<td></td>
<td>HW#3 &amp; Proposal Slides</td>
</tr>
<tr>
<td>11/12/21</td>
<td>HW #4 Linked Big Data Analytics [assignment][tutorial]</td>
<td></td>
</tr>
<tr>
<td>11/19/21</td>
<td></td>
<td>Progress Slides</td>
</tr>
<tr>
<td>11/26/21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/03/21</td>
<td></td>
<td>HW #4 &amp; Progress Report</td>
</tr>
<tr>
<td>12/10/21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/17/21</td>
<td></td>
<td>Final Project Slides and Other Materials</td>
</tr>
</tbody>
</table>
Other Issues

- Professor Lin:
  - Office Hours: After the class or by appointment
  - Contact: c.lin@columbia.edu

- TA (CA/IA/Grader) —
  - Cong Han (ch3212): Tue 4-6pm
  - Yvonne Lee (yl4573): Wed 4-6 pm
  - Guoshiwen Han (gh2567): Mon 9-11am
  - Yiwen Fang (yf2560): Thu 5:30-7:30pm (may change to 3-5pm; please see the course website)
Reading Reference for Lecture 1

Chapter 1: Market and Business Drivers for Big Data Analysis
Chapter 2: Business Problems Suited to Big Data Analytics
Chapter 3: Achieving Organizational Alignment for Big Data Analytics
Chapter 4: Developing a Strategy for Integrating Big Data Analytics into the Enterprise
Chapter 5: Data Governance for Big Data Analytics: Considerations for Data Policies and Processes
Chapter 6: Introduction to High-Performance Appliances for Big Data Management
Chapter 7: Big Data Tools and Techniques
Chapter 8: Developing Big Data Applications
Chapter 9: NoSQL Data Management for Big Data
Chapter 10: Using Graph Analytics for Big Data
Chapter 11: Developing the Big Data Roadmap
5 Example Big Data Use Case Categories

**Big Data Exploration**
Find, visualize, understand all big data to improve decision making

**Enhanced 360° View of the Customer**
Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

**Security/Intelligence Extension**
Lower risk, detect fraud and monitor cyber security in real-time

**Operations Analysis**
Analyze a variety of machine data for improved business results

**Data Warehouse Augmentation**
Integrate big data and data warehouse capabilities to increase operational efficiency
1. Expertise Location
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Healthcare Analysis
8. Data Exploration and Visualization
9. Personalized Search
10. Anomaly Detection
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring (Smarter another Planet)
14. Cellular Network Monitoring
15. Cloud Monitoring
17. Traffic Navigation
18. Image and Video Semantic Understanding
19. Genomic Medicine
20. Brain Network Analysis
21. Data Curation
22. Near Earth Object Analysis
Category 1: 360° View

Recommendation

item

user

Enhancing:

Graph Visualizations

Communities

Centralities

Ego Net Features

Graph Search

Graph Query

Graph Matching

Graph Sampling

Network Info Flow

Shortest Paths

Latent Net Inference

Bayesian Networks

Markov Networks

Middleware and Database
Use Case 1: Social Network Analysis in Enterprise for Productivity

Production Live System used by IBM GBS since 2009 – verified ~$100M contribution

- 15,000 contributors in 76 countries; 92,000 annual unique IBM users
- 25,000,000+ emails & SameTime messages (incl. Content features)
- 1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, ..., access data
- 1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data
- 200,000 people’s consulting project & earning data

- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: $7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings and
- APQC (WW leader in Knowledge Practice) April 2013: “The Industry Leader and Best Practice in Expertise Location”
Use Case 3: Customer Behavior Sequence Analytics

- Behavior Pattern Detection
- Help Needed Detection

- Markov Network
- Latent Network
- Bayesian Network
Use Case 4: Graph Analytics for Financial Analysis

**Goal:** Injecting Network Graph Effects for Financial Analysis. Estimating company performance considering correlated companies, network properties and evolutions, causal parameter analysis, etc.

- IBM 2003
- IBM 2009

**Data Source:**
- Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

**Network feature:**
- \( s \) (current year network feature),
- \( t \) (temporal network feature),
- \( d \) (delta value of network feature)

**Financial feature:**
- \( p \) (historical profits and revenues)

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 33%.

Targets: 20 Fortune companies’ normalized profits

Goal: Learn from previous 5 years, and predict next year

Model: Support Vector Regression (RBF kernel)
Use Case 5: Social Media Monitoring

Live Tweets, Sentiment, Keywords, Graphs, Zooming/Panning, Real-Time Translation, Locations, Top Retweets.
Use Case 6: Customer Social Analysis for Telco

**Goal:** Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

- **Applications based on the extracted social profiles**
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign

- **Approach**
  - Construct social graphs from CDRs based on \{caller, callee, call time, call duration\}
  - Extract customer social features (e.g. influence, communities, etc.) from the constructed social graph as customer social profiles
  - Build analytics applications (e.g. personalized advertisement) based on the extracted customer social profiles

**Applications**
- Personalized Advertisement
- High Value Customer Identification & targeting
- Viral marketing campaign

**System G Analysis**
- Degree Centrality
- Weakly Connected Component
- Maximal Cliques
- Pagerank
- Community Detection
- K-core

**Customer Profiles** (influence, community, etc.)

PoCs with Chinese and Indian Telecomm companies
Category 2: Data Exploration

Huge Network Visualization
Network Propagation
I2 3D Network Visualization
Geo Network Visualization
Graphical Model

Communities
Graph Search
Network Info Flow
Bayesian Networks

Centralities
Graph Query
Shortest Paths
Latent Net Inference

Ego Net Features
Graph Matching
Graph Sampling
Markov Networks

Middleware and Database
Use Case 7: Graph Analytics and Visualization

Graph Matching

(a) Query

(b) Matches

(c) Graph Communities

(entity)

headache

high fever

chill

temperature

migraine

stomachache

cough

Matches

(d) Graph Matching

- cancer
- diabetes
- kidney disorder
- heart disease
- HIV
- cold
User Case 8: Visualization for Navigation and Exploration

Whisper: Tracing the information diffusion in Social Media

SocialHelix: Visualization of Sentiment Divergence in Social Media

- Belts: communities
- Color: sentiments
- Bars: keywords & hashtags of an event
Use Case 9: Graph Search

existing search engine

query -> index -> ranking -> re-ranking

Graph analysis

query context

Improved search results:

Interest / social network based content recommendations

Info-Socio networks

Use Case 9: Graph Search

Practitioner Portal

Search criteria

Search results

18,577 results found

IBM Social Business Adoption QuickStart (U.S. English). Proposal Insert | Proposal and Presentation Accelerator (PPAQ)

Drive the successful launch and adoption of social business software throughout your organization with a structured engagement comprised of assessments, planning and design consultation, online workshops, and team- and skills-building activities.
Category 3: Security

Network Info Flow

Ponzi scheme Detection

Ego Net Features

Attacker: Near-Star

Normal:
1. Clique-like
2. Two-way links

Detecting DoS attack

Graph Visualizations

Communities

Centralities

Ego Net Features

Graph Search

Graph Query

Graph Matching

Graph Sampling

Network Info Flow

Shortest Paths

Latent Net Inference

Bayesian Networks

Markov Networks

Middleware and Database
Use Case 10: Anomaly Detection at Multiple Scales

Based on President Executive Order 13587

**Goal:** System for Detecting and Predicting Abnormal Behaviors in Organization, through large-scale social network & cognitive analytics and data mining, to decrease insider threats such as espionage, sabotage, colleague-shooting, suicide, etc.

“Enterprise Information Leakage Impacted economy and jobs” Feb 2013

“What's emerged is a multibillion dollar detective industry” npr Jan 10, 2013

**Infrastructure + ~ 490 Analytics**
Use Case 11: Fraud Detection for Bank

Ponzi scheme Detection

Attacker: Near-Star
Normal:
1. Clique-like
2. Two-way links

(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge
Use Case 12: Detecting Cyber Attacks

Detecting DoS attack

(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.
Category 4: Operations Analysis

Cloud Service Placement

KPI time series (e.g., server performance/load, network performance/load)

Causality analyzer

- KPI (a time series)
- (potential) pairwise relationship (e.g., causality)

Varying over time

Graph Visualizations

- Communities
- Centralities
- Ego Net Features
- Graph Search
- Graph Query
- Graph Matching
- Graph Sampling
- Network Info Flow
- Shortest Paths
- Bayesian Networks
- Latent Net Inference
- Markov Networks
Use Case 13: Smarter another Planet

**Goal:** Atmospheric Radiation Measurement (ARM) climate research facility provides 24x7 continuous field observations of cloud, aerosol and radiative processes. **Graphical models** can automate the validation with improvement efficiency and performance.

**Approach:** BN is built to represent the dependence among sensors and replicated across timesteps. BN parameters are learned from over 15 years of ARM climate data to support distributed climate sensor validation. Inference validates sensors in the connected instruments.

**Bayesian Network**
- 3 timesteps
- 63 variables
- 3.9 avg states
- 4.0 avg indegree
- 16,858 CPT entries

**Junction Tree**
- 67 cliques
- 873,064 PT entries in cliques
Use Case 14: Cellular Network Analytics in Telco Operation

**Goal:** Efficiently and uniquely identify *internal* state of Cellular/Telco networks (e.g., performance and load of network elements/links) using probes between monitors placed at selected network elements & endhosts

- Applied Graph Analytics to telco network analytics based on CDRs (call detail records): estimate traffic load on CSP network with low monitoring overhead
  1. CDRs, already collected for billing purposes, contain information about voice/data calls
  2. Traditional NMS* and EMS** typically lack of end-to-end visibility and topology across vendors
  3. Employ graph algorithms to analyze network elements which are not reported by the usage data from CDR information

- **Approach**
  - Cellular network comprises a hierarchy of network elements
  - Map CDR onto network topology and infer load on each network element using graph analysis
  - Estimate network load and localize potential problems
**Use Case 15: Monitoring Large Cloud**

**Goal:** Monitoring technology that can track the time-varying state (e.g., causality relationships between KPIs) of a large Cloud when the processing power of monitoring system cannot keep up with the scale of the system & the rate of change

- **Causality relationships (e.g., Granger causality) are crucial performance monitoring & root cause analysis**
- **Challenge:** easy to test pairwise relationship, but hard to test multi-variate relationship (e.g., a large number of KPIs)

![Diagram of KPIs and causality relationships]

**KPI time series** (e.g., server performance/load, network performance/load) → **Causality analyzer** → **Basic analytics engine** (e.g., pairwise granger causality) → **Link sampling & estimation** → **Overall graph**

**Select KPI pairs (sampling) → Test link existence → Estimate unsampled links based on history → Overall graph**
Category 5: Data Warehouse Augmentation
Use Case 16: Code Life Cycle Improvement

- Advantages of working directly with graph DB for graph applications
  1. Smaller and simpler code
  2. Flexible schema → easy schema evolution
  3. Code is easier and faster to write, debug and manage
  4. Code and Data is easier to transfer and maintain
Use Case 17: Smart Navigation Utilizing Real-time Road Information

**Goal:** Enable unprecedented level of accuracy in *traffic scheduling* (for a fleet of transportation vehicles) and navigation of individual cars utilizing the *dynamic real-time information* of changing road condition and predictive analysis on the data.

- Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shortest path computation) on time-varying graphs (order of magnitudes improvement over existing solutions).

- High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall.

Our approach:
Querying over dynamic graph + predictive analytics on graph properties.
Use Case 18: Graph Analysis for Image and Video Analysis
Use Case 19: Graph Matching for Genomic Medicine

Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.
Use Case 20: Data Curation for Enterprise Data Management

Prior Collaborative Use

torkroth@us.ibm.com

eser@us.ibm.com

works on

Project

uses

BE/BSA-I-3

Semantic Knowledge

Topic Cancer

has topic

Topic Diseases

has topic

Cancer Drugs

has topic

Supervised Curation

Curation User Interface

Drug Table

Manufactured In

has Table

Hospital Table

Location

joins with

has Table

has Column

Extracted Metadata

has Table

Cancer Drugs

has Table

manufactured in
Use Case 21: Understanding Brain Network
Use Case 22: Planet Security

• Big Data on Large-Scale Sky Monitoring

Dangers from space
Learn about the threat to Earth from asteroids & comets and how the Pan-STARRS project is designed to help detect these NEOs. Learn more...

1,400,000,000 pixels
Pan-STARRS has the world’s largest digital cameras. Read about them here...

The PS1 Prototype
PS1 goes operational and begins science mission
PS1 Science Consortium formed...
PS1SC Blog
PS1 image gallery
Homework #0: Big Data Environment Setup and Test (due September 24, 5pm)

1. Warm-Up Exercises:
   • Setup Google Cloud account and environment
   • Install Google Cloud SDK
   • Create a Spark cluster
   • Word Count using Google Cloud Storage and Spark
   • Hive and BigQuery

2. Data Analysis — NYC Bike Expert:
   • Load data to a Cloud Storage
   • Simple Analyses through BigQuery

3. Data Analysis — Understanding Shakespeare:
   • Load data to a Cloud Storage
   • Simple Analyses through Word Counts
   • Analyses after running Natural Language Toolkit
Homework Late Submission Policy

5pm: submission deadline
Next Day 5pm: 10% penalty
Two Days late 5pm: 20% penalty
Three Days late 5pm: 30% penalty
Any late submission more than 3 days (5pm) will not be accepted.