Course Notes - Data Analytics and Security

Nolan MacDonald

Fall 2024

Contents

Chapter 1

Module 1 - Introduction

1.0.1 Overview

Big data has become a buzzword. Its hype has risen, peaked, and now fallen a bit. But in the wake of the hype cycle, there are a number of enduring changes to the technological and social landscape. Where do these changes come from? What are the underlying technical features that made this happen?

- Name underlying technical and social aspects underlying big data
- Define big data

1.0.2 Resources

[1.] The Parable of Google Flu: Traps in Big Data Analysis

1.1 What is Big Data?

Module 1.1 – What is Big Data? (9:36)

1.1.1 Three Organizing Questions

- Who uses big data?
- What does big data do for them (as opposed to regular data)?
- What are some other associated topics or buzzwords?

1.1.2 Big Data - "3 V's"

Big Data is known by "3 V's" and is also a catch all buzz word

- Volume, Velocity, Variety
- Aggregation of data all in one place
- Strong implications for international affairs e.g., Election interference
- Wiki: 4th concept, veracity – quality or insightfulness of the data, errors in data

1.1.3 Data

Data is information encoded in a series of bits

- Bits can take a value of 0 or 1 integers
- Bits are encoded as a series of 0s and 1s (typically 32)
- Text is also a series of 0s and 1s, but the schemes are more complicated
- ASCII is a 7 bit scheme for Latin alphabet (upper and lower case), punctuation and digits
- Unicode is more complex, variable length scheme for characters in many languages, plus emoji-like characters
- Example: "Jeffrey" has 7 characters, at 7 bits each in ASCII, total 49 bits

1.1.4 Volume

- How much data is there? 12 zetabytes (1.2E22)
- Data is growing at an increasing rate: 90% of data was created in the last 2 years

1.1.5 Units of Data

- 1 bit is one piece of binary information (zero or one)
- 1 byte is a set of 8 bits – Can be one of 256 combinations (28)
- 104 (10 kilobytes) – Couple pages of text, long email
- 106 (1 megabyte) – Roughly 1 min. of compressed music
- 109 (1 gigabyte) – Roughly a compressed but decent hour video
- 3.2E9 bytes (3.2 gigabytes) – Amount of data in your DNA
- All text on Wikipedia is about 9.5GB
- 1012 bytes (1 terabyte) is about the size of an external hard drive (in 2014)
- 3E12 bytes (3 TB) is the approximate amount of storage in your brain

1.1.6 Server Rack

- 4E14 bytes (400 TB) is about as much data that can be housed in a server rack
- Can be up to a petabyte
- 400 TB is the amount of data in all books ever written

1.1.7 Data Center

- 1018 bytes (1 exabyte) is how much could be stored in a data center
- 5E18 bytes (5 exabytes) would be the size of all words ever spoken, if transcribed

1.1.8 Variety

Aspects of life recorded in 1980 - Plane tickets, taxes, interactions with the largest companies - Major companies had mainframe computers to keep records of specific interactions

1.1.9 Aspects of life recorded in 2020

- Location (from cell phones)
- Attitudes on social media (from posts, likes, etc.)
- For some people, "big data" is nearly synonymous with social media/internet technologies
- Workflow and interactions (from emails, internet search logs)
- Speech – From smart devices like Alexa and Siri

By combining different types of information, lots of different possibilities to impact society

1.1.10 Velocity

- Primary example: Twitter/X
- Twitter saves a tweet and sends it to mobile devices, browsers, etc.
- Twitter has to manage multiple tweets before it's done with saving and sending

1.1.11 Data in Real Time

Some web related technologies need to keep up with data in real time, even as it comes faster

- Is a web request part of a denial of service attack?
- Is the email spam or phishing?
- High peak loads – What is Healthcare.gov needs to serve 10 million people in 1 day?
- Number of people per day visiting a site can vary widely creating issues with velocity
- Outside of the web, there is threat monitoring software
- Catalog all the threats or sources in real time

1.1.12 Big Data - A Broad Term

Think of big data as something broader than just data size and data type

- "Big Data" will include things that big data might enable or portend as it enables society at large
- Power to collect (and lose) personal info
- Power to predict
- Artificial intelligence (AI) and how it relates to the power of prediction
- Driverless cars

1.1.13 Concepts that underlie big data

- Computing – Changes in computing, key technologies
- Statistics – Make predictions and inferences of something that might happen
- Applications – What are the new things that big data enables that allow us to change the world and what does it mean for the rest of the world

1.2 Google Flu Trends

Module 1.2 – Google Flu Trends (7:03)

1.2.1 Google Flu Trends (GFT)

Google Flu Trends (GFT) – Predicted flu based on search queries

- Original paper (2008) – Launched in 2008, lecture says 2009 put in production
- Starts with CDC's system of cataloguing Influenza-Like-Illness (ILI)
- Weekly, state-level number of incident account
- Reported how many people got the flu at the doctor's office
- Aggregates data at state-level and uses statistics to estimate relative prevalence between the different doctors
- Lag real-time by a few weeks
- Calculate weekly, state-level search term prevalence
- Used top 50 million terms like cough, aching, terms related to the flu
- Each state in each week and matched to CDC ILI data
- Tried a lot of combinations of these terms to pick the group which best matched inputs

1.2.2 Number of top queries and correlation

- Use the top 45 queries able to predict the ILI data variation between state and time
- Mean correlation above 0.90 (very high)
- Calculations in real-time, faster frequency, more up to date
- Produce data at a smaller level – Not just Georgia but Atlanta or subsections of Atlanta

1.2.3 Issues

- Had trouble in 2008-2009 with the onset of H1N1 (new flu strain), model was refit
- Did well for a couple years
- Trouble in 2012-2013 season due to high press coverage of flu

1.2.4 Assumptions and Issues

- Flu trends depend on CDC ILI reporting
- There is no Google Flu Trends without actual measured flu trends
- The relationship between search behavior and flu is the same over time
- Google grew quickly from 2003-2008 - Biggest issue with the 94% correlation model
- 2000 – 22B (22E9) searches
- 2007 – 438B (438E9) searches
- 2008 – 637.2B (637.2E9) searches
- 2009 – 953.7B (953.7E9) searches
- Search behavior and searches changed over time – was not stable
- Search behavior and searches is a common problem with big data
- GFT could not have been made without CDC data
- Made it quicker and more granular but did not make anything new

Shut down in 2015 and provided to researchers at CDC

Continuing experiments to incorporate GFT but CDC does not really use GFT

1.2.5 Summary

GFT is big data – Depends on search queries, primary example of big data (1 trillion per year) Benefit – Finer granularity of place and time Set of key assumptions – The way that people search for IFI and searches should not change over time End result – Not a useful product, not able to latch onto any trends because searches changed over time (search behavior, queries, number of searches) GFT represents an interesting and useful new product of big data but shows how insights and products on big data might be fragile

1.3 Required Reading Notes

[1]. The Parable of Google Flu: Traps in Big Data Analysis

1.3.1 Introduction

- February 2013 - Google Flu Trends (GFT) predicted more than double doctor visits for the flu compared to the Centers for Disease - Control and Prevention (CDC)
- Estimates based on surveillance reports across the US
- GFT was built to predict CDC reports
- Research on whether search or social media can predict x has become commonplace and is often put in sharp contrast with traditional methods and hypotheses
- Search or social media is far from supplanting more traditional methods or theories
- GFT mistakes – Big data hubris (exaggerated self-confidence) and algorithm dynamics

1.3.2 Big data hubris

- Big data hubris is the Implicit assumption that big data is a substitute not a supplement to traditional data collection and analysis
- Enormous scientific possibilities in big data
- Quantity of data doesn't mean you can ignore issues of measurement, construct validity, reliability and dependencies among data
- GFT problematic using big and small data – Find best matches among 50 million search terms to fit 1152 data points
- Flu data structurally unrelated – weed out seasonal search terms such as high school basketball
- Big data was overfitting the small number of cases – a standard concern in data analysis
- GFT became a part flu detector, part winter detector
- GFT had to update algorithm in 2009 and noted a few changes in October 2013
- GFT study in 2010 demonstrated GFT was not much better than a simple projection forward using CDC data (on a 2-week lag)
- Lagged models over time significantly outperforming GFT
- Still useful if combined with other near-real-time health data

1.3.3 Algorithm Dynamics

- Algorithm dynamics – Changes made by engineers to improve the commercial service and by consumers in using that service

- GFT was an unstable reflection of the prevalence of the flu because of algorithm dynamics affecting Google's search algorithm

- Missed by large amount in 2011-2012 Flu season

- GFT issues may be from changes to Google's search algorithm, which is constantly testing and improving search
- Replicating original algorithm is difficult – GFT never documented the 45 search terms used
- GFT errors are closely aligned with searches for flu treatments and information on differentiating the cold from the flu
- "Blue Team" Dyanmics – Algorithm producing the data (and user utilization) has been modified by the service provider in accordance with their business model
- GFT search model changes to support Google's business model – Providing information quickly and to promote more advertising revenue (e.g., Recommended searches)
- "Red Team" Dynamics – Occur when research subjects (e.g., web searchers) attempt to manipulate the date-generating process to meet their own goals
  - E.g., economic or political gain, twitter polling, using tactics to make sure their candidate or product is trending, twitter or FB spreading rumprs about stock prices and markets

1.3.4 Transparency, Granularity and All-Data

- Hard to replicate GFT without core search terms, collection of searches
- Using Google Correlate with concepts on how GFT was built will not replicate their findings
- The few search terms in the GFT papers do not seem to be strongly related with either GFT or CDC data
- Use Big Data to Understand the Unknown
- Valuable to understand the prevalence of flu at very local levels, not practical for CDC to produce
- Constantly changing algorithms need to be better understood how they change over time
- Need to replicate findings across time and use other data sources to ensure they are observing robust patterns and not evanescent trends (quickly fading trends)
- Big data offers possibilities for understanding human interactions at a societal scale
- We should focus on big and small data through innovative analytics using data from traditional and new sources

1.3.5 Bytes

- Byte, Kilobyte (kb), Megabyte (mb), Gigabyte (gb), Terrabyte (tb), Petabyte (pb)
- 1 yottabyte – Storage capacity of NASA datacenter
- Structured data can contain unstructured components
- Data growth comes from unstructured data

1.3.6 5 "V's"

- Volume, Variety, Velocity, Veracity, Value
- Velocity – Increasing speed of which data is created and the speed at which it can be processed, stored, and analyzed
- Veracity – quality or insightfulness of the data, errors in data

1.3.7 Data analytics types

- Descriptive – What happened? Requires most amount of human input
- Diagnostic – Why did it happen?
- Predictive – What will happen?
- Prescriptive – How can we make it happen? No human input, most value, optimization, decision support and decision automation, most difficult

Chapter 2

Module 2 - Hardware Trends

2.0.1 Overview

The improvements in computer hardware technology in the last 50 years have been incredible. We identify which aspects of computing have evolved quickly and which have not, and trace what this means for big data technologies.

2.0.2 Resources

[1]. Understanding Hadoop Clusters and the Network

[2]. Apache Hadoop

[3]. Apache Spark

[4]. Developing for the Intelligent Cloud and Intelligent Edge Rohan Kumar (Microsoft)

2.1 Trends in Computer Hardware

Module 2.1 - Trends in Computer Hardware (9:58)

How trends in computer hardware are shaping what Big Data is

2.1.1 Parts of a Computer

- CPU – executes instructions
- Memory – store information “for a little while” during computation
- This is RAM, temporary storage
- Disk – Store information for longer periods of time
- Motherboard architecture driven by heat dissipation

2.1.2 CPU

- Read an instruction from memory and decode it
- Find any associated data that is needed to process the instruction
- Process the insruction

- Write the results out
- CPU implementation process is called microarchitectures e.g., Xeon, Whiskey Lake
- Key Challenges: Waiting for data to complete an instruction, conditional execution/branches
- Key Solution: Multiple queues, guess at what branches will be followed

2.1.3 Storing Information

- To work quickly, computers need to access information quicklu
- Physical constraunts: you can't store all your data right next to your CPU

2.1.4 Level of closeness to look up a byte

- L1 cache – 0.5 ns
- L2 cache – 7 ns
- Memory – 100 ns
- Disk – 10 million ns
- In memory elsewhere in data center – 500,000 ns
- Solid state hard drives – 500,000 ns
- Note: A 1 ghz CPU executes an instruction ever nanosecond, including potentially looking up data

Moore's Law: The number of transistors per unit area can double every 18 months **Kryder's Law:** Storage density on magnetic disks doubles every 18 months

Processing and storage is getting cheaper and faster Accessing hard disks is NOT getting faster (bottleneck of Big Data) Increasing processor performance is exponentially more costly

2.1.5 CPU Limits

- Moore's Law
- Power Dissipation – If you shrunk transistors, power dissipation would decrease
- Typically processors do more, so power per area increases
- Single cores can't go faster without taking too much power
- Instead of more speed – Multiple cores, more cache

2.1.6 Memory Physical Limits

- Also depends on Moore's law-like progress
- Doesn't have acute overheating problems like CPUs – Capacity just keeps growing

2.1.7 Hard Disk Physical Limits

- Kryder's law suggests total storage will increase
- Platters can only spin so fast - accessing this large amount of data is a bottleneck
- One Exception
- Sequential reads will get faster since the head can read more data in a single rotation
- This can benefit performance of massive data operations which often need lots of sequential data

2.1.8 Summary

Computer has 3 components – CPU, Memory, Disk

Major technological trends - CPU is growing cheaper but not faster - Memory/Disk space is growing - Disk access is NOT growing

2.2 Computer Architecture for Big Data

2.2.1 Big Data Performance

Module 2.2 - Big Data Performance (10:10)

- Many computations are fixed size
- Some things may grow faster than computational resources o Aspects of social network data o Log data
- As disk space in particular gets cheaper, new things that were previously not storable will be stored on disk

2.2.2 Discussion Questions

- How would you build a system which can handle more data than fits on a single hard drive?
- What if you need a system which will serve a website to more customers at once?
- How would you sort a list which doesn't fit on one computer?

2.2.3 Strategies

- Better Hardware - Supercomputers
- Exponentially more expensive
- Petabytes of memory, 10,000s of CPU/GPU cores
- Cost: $100M+
- Distributed Architectures – Using multiple computers together

2.2.4 Load Balancing

- Using computer architectures in parallel
- The best way to distribute work depends on what the work is

2.2.5 Databases for Big Data: Master/Worker Architecture

- Master has list of which data is on which computers
- When requesting the data the workers provide their parts (other computers)
- Issues
- What is the master gets overloaded?
- What if the master fails?

2.2.6 Databases for Big Data: Eventual Consistency

- What happens if the leader node gets overwhelmed?
- We could have multiple leaders, each of which can task worker nodes
- Eventual consistency
- Leaders know what other works are doing by sending messages
- Messages are not instant, so first leader can be requested data information that the second knows changed but the first hasn't been notified
- This is OK in many contexts but not in things like e.g., banking

2.2.7 Serving Multiple Customers

- Each computer receives requests, but can only process so many per unit time
- Multiple machines can handle more requests
- But what if you're a website which needs to look up information about a customer?
- If many servers deal with the same database it can be overloaded
- If just reading data you can have many copies of the database
- If you sometimes change data, you need to make sure there aren't multiple changes at the same time

2.2.8 Sorting a Big List

A loose algorithm for sorting a list on N computers

- On computer 1, sort the data and take each 1/N partition and assign it to a computer
- Tell each computer to send its values less than 1/N to the first computer, those between 1/N and 2/N to the second, etc.
- Now we know each computer has data which is either all greater or all less than the data on each other computer
- So we just sort the data on each computer, and we're done!
- We have ways of sorting large amounts of data
- What probably slows this down?
- Sending data from one computer to another
- Reading data from disk – This always was going to happen

2.2.9 Summary

- Supercomputers v. Software Architecture
- Introduction of Big Data architectural elements
- Load balancing
- Master/Worker
- Eventual consistency
- Custom algorithms

2.3 Software Architecture for Big Data

Module 2.3 - Software Architecture for Big Data (11:39)

Chapter 3

Module 3 - Software Trends

3.0.1 Overview

In order to handle large amounts of data, computers require software. How do computers actually deal with all that data? We start by discussing data storage and retrieval technologies and show how software or algorithmic "tricks" can substantially speed up computations. Then we talk about the fundamental drivers of cost and performance in software systems.

- Discuss how software fits into analyzing large amounts of data
- Use python to count the number of words and the most common word in a book
- Discuss tradeoffs in distributed algorithms

3.0.2 Resources

[1]. Apache Lucene - Free, open-source search engine library that provides full-text search capabilities

[2]. Elastic Search - Distributed search and analytics engine that stores and indexes data for fast search, analytics, and relevancy

[3]. Elastic Search Guide - Search and analytics engine that powers the Elastic Stack

3.1 The Importance of Software

Module 3.1 - The Importance of Software: An Example of Search Engines (8:47)

3.1.1 How Would You Make a Search Engine?

3.1.2 Indexing

3.1.3 Relevance Rankings

3.1.4 Relevance + Indexing

Relevance + Indexing = Modern Search

3.1.5 Purpose

3.1.6 Summary

3.2 Big Data or Just Data

Module 3.2 - Big Data or Just Data (12:54)

3.3 Programming Languages and Tradeoffs

Module 3.3 - Programming Languages and Tradeoffs (8:19)