Chaitanya Durgesh Nynavarapu
A20561894

ASSIGNMENT-1
BIG DATA (CSP 554)

3.

The Parable of Google Flu: A Cautionary Tale in Big Data Analysis

Ever wondered what happens when a tech giant tries to outsmart traditional public health methods? Well, buckle up, because the story of Google Flu Trends (GFT) is a wild ride that'll make you think twice about the power of big data.

The Big Idea

Back in the day, Google had this brilliant idea: use people's search queries to predict flu outbreaks. Sounds clever, right? They thought so too. The goal was to revolutionize how we track and respond to flu epidemics. Spoiler alert: things didn't quite go as planned.

What Went Wrong?

- Big Data Hubris: Google's team got a bit too cocky with their mountain of data. They figured, "Who needs those old-school health reports when we've got all these searches?" Classic case of putting all your eggs in one (very big) basket.
- Correlation ≠ Causation: Just because people were searching for flu-related terms doesn't mean they actually had the flu. Sometimes, a sneeze is just a sneeze, you know?
- The Algorithm Tango: Google's search algorithms are always changing, and so are people's search habits. This constant dance threw GFT's predictions are way off-balance.

The Lessons Learned

1. **Don't Ditch the Classics:** Traditional Data sources like CDC reports still have their place. It's not about replacing them but complementing them.
2. **Stay Flexible:** The digital world changes fast. Any model based on it needs to be ready to change just as quickly.
3. **Context is King:** Big data without context is like a map without a compass – you might have a lot of information, but you are still lost.

The Takeaway

The GFT story isn't just about a failed project; it's a humbling reminder that even with all our technological advances, predicting human behavior and health trends is incredibly complex. It's a call for a more balanced approach, combining the power of big data with good old-fashioned scientific rigor. So, next time you hear someone singing the praises of big data as the solution to all our problems, remember the parable of Google Flu. It might just save you from catching a case of big data fever!

4.

A Brief Survey on Big Data: Technologies, Terminologies, and Data-Intensive Applications

Introduction

Big data has emerged as a prominent research topic due to its widespread applications across various domains. This survey, conducted by Hemn Barzan Abdalla, explores the significance of big data, its taxonomy, and the technologies used in big data applications.

Research Question and Significance

The study aims to provide a comprehensive overview of big data technologies, terminologies, and applications. This research is important because big data has become crucial in decision-making, data sciences, business applications, and government sectors.

Key Concepts and Definitions

Big data is characterized by the "5 V's":

1. Volume: Massive amounts of data, often in terabytes or petabytes
2. Variety: Diverse types of data (structured, unstructured, and semi-structured)
3. Velocity: Speed of data generation and processing
4. Veracity: Accuracy and reliability of data
5. **Value:** Insights and knowledge extracted from the data

Methodology

The author conducted a literature review, analyzing various big data technologies, tools, and applications. The study focused on processing techniques, security measures, and storage solutions for big data.

Results:

The survey identified several key technologies and tools used in big data processing:

- NoSQL: For handling unstructured data models
- Cassandra: Distributed database system for large datasets
- Hadoop: Open-source framework for distributed storage and processing
- Apache Spark: Fast cluster computing system
- Apache Storm: Real-time computation system for data streams
- Apache Hive: Data warehouse software for querying large datasets

Big data applications were found in various domains, including:

- Organizational management
- Media and entertainment
- Environmental monitoring
- Education
- Healthcare and life sciences
- Social media and networking
- Smart city transportation
- Data transfer and communication

Discussion

The study highlighted several challenges associated with big data applications:

1. Data quality assurance and validation
2. Security and privacy concerns
3. Metadata generation

4. Real-time processing of large volumes of data
5. Scalability issues

The author emphasized the importance of visual analytics in big data, consisting of three main layers:

1. Visualization: Representing data visually
2. Analytics: Drawing conclusions from data
3. Data Management: Managing the lifecycle of data

Implications and Future Research

The survey identified several areas for future research and development:

- Improving data processing techniques and algorithms
- Enhancing security and privacy measures
- Developing more efficient storage solutions
- Advancing real-time analytics capabilities
- Exploring new applications in emerging fields like IoT and AI

Conclusion

This comprehensive survey provides valuable insights into the current state of big data technologies, terminologies, and applications. It highlights the growing importance of big data across various sectors and the need for continued research to address challenges and harness its full potential.

5.

Answers to the Questions

1. **What was the problem with the Google flu detection algorithm?**

   ➢ The problem with the Google flu detection algorithm was that it significantly overestimated flu cases for extended periods, sometimes by more than double the actual numbers reported by the CDC. It failed to maintain accuracy over time due to changes in search patterns and Google's own algorithms.

2. **What is big data hubris?**

   ➢ Big data hubris refers to the overconfidence in the ability of big data to solve problems without needing traditional methods of data collection and analysis. It's the assumption that large datasets are inherently superior to smaller, more carefully curated data

3. **What approach could have been used to improve the Google flu detection algorithm?**

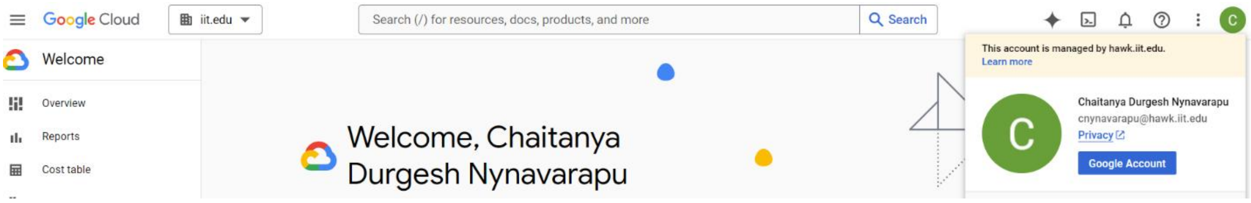
   ➢ To improve the Google flu detection algorithm, researchers suggested combining GFT data with traditional CDC reports and applying more sophisticated statistical techniques. This "mash-up" approach of integrating big data with conventional surveillance methods could have yielded more accurate results.

4. **What is “algorithm dynamics”?**

   ➢ Algorithm dynamics refers to how changes in Google's search algorithms and shifts in user behavior affected the consistency and accuracy of GFT's predictions. These ongoing modifications to the search ecosystem made it difficult for the original GFT model to maintain its predictive power over time.

6.

- ***Main page of the GCP management console***



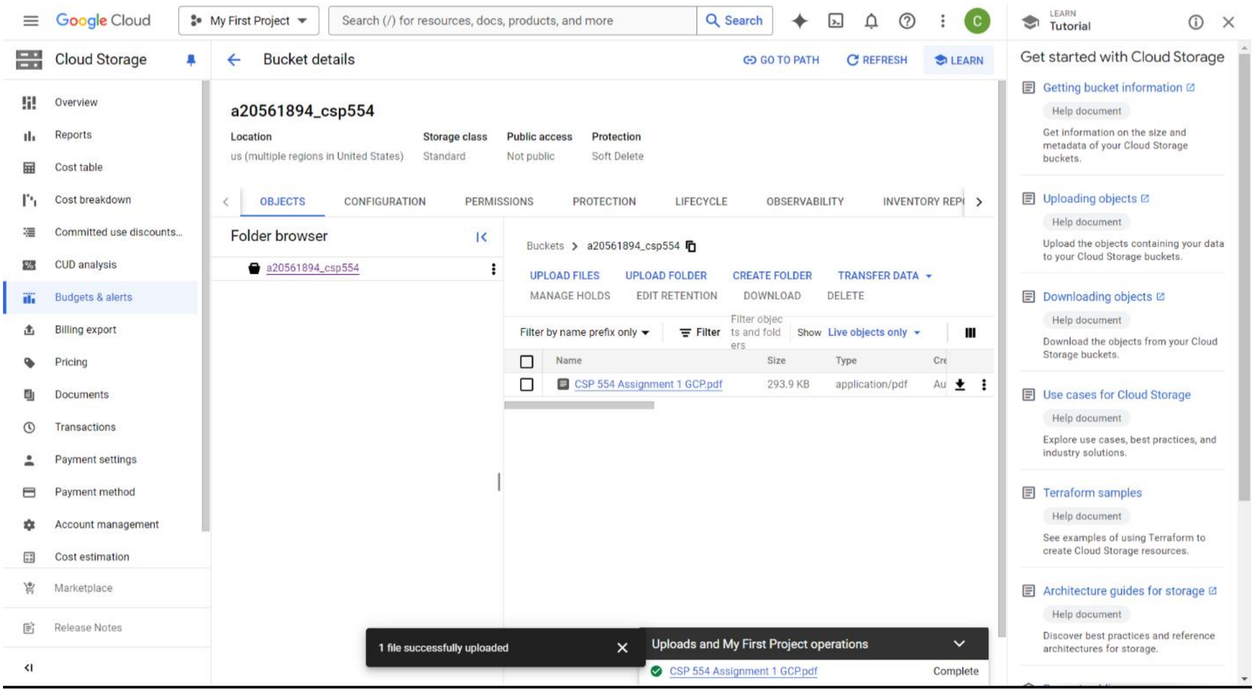
Welcome, Chaitanya Durgesh Nynavarapu

This account is managed by hawk.iit.edu.
Learn more

Chaitanya Durgesh Nynavarapu
cnynavarapu@hawk.iit.edu
Privacy

7.

- ***Object in the bucket***



a20561894_csp554

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

☐	Name	Size	Type	Cre
☐	CSP 554 Assignment 1 GCP.pdf	293.9 KB	application/pdf	Au

1 file successfully uploaded

Uploads and My First Project operations

✓ CSP 554 Assignment 1 GCP.pdf Complete

Get started with Cloud Storage

Getting bucket information
Help document
Get information on the size and metadata of your Cloud Storage buckets.

Uploading objects
Help document
Upload the objects containing your data to your Cloud Storage buckets.

Downloading objects
Help document
Download the objects from your Cloud Storage buckets.

Use cases for Cloud Storage
Help document
Explore use cases, best practices, and industry solutions.

Terraform samples
Help document
See examples of using Terraform to create Cloud Storage resources.

Architecture guides for storage
Help document
Discover best practices and reference architectures for storage.