Tackling The Challenges of Big Data
Visualizing Twitter

Samuel Madden

Professor and Director of Big Data at CSAIL

Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Visualizing Twitter
Introduction to Twitter Data

Samuel Madden

Professor and Director of Big Data at CSAIL

Massachusetts Institute of Technology

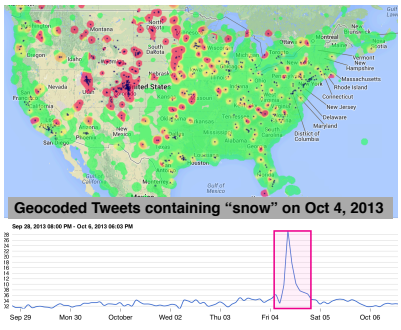
This Module

1. Understanding Twitter data
2. Demonstration of MapD, a Twitter visualization system
3. Technology behind MapD
4. Other Approaches to Interactive Analytics

Part 1: Twitter

- **500 million tweets a day**
  - > 8 million/day "geocoded"
- **More than just 140 characters:**
  - Geo Coordinates
  - Timestamp
  - User and Follower information
  - Reply Information
  - Hashtags
  - Device/Platform Used to Post

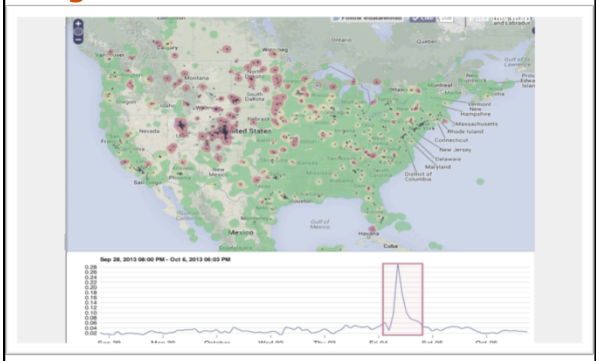
Geocoded Data



"Big Data" and Twitter

- **Volumes and rates are massive; need new tools to interactively visualize data**

"Big Data" and Twitter

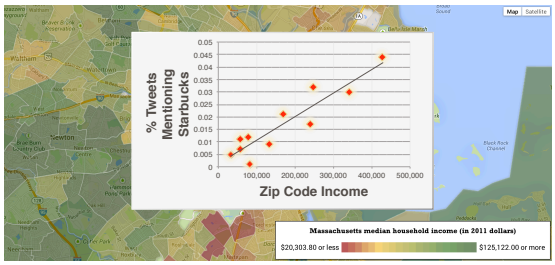


"Big Data" and Twitter

- **Volumes and rates are massive; need new tools to interactively visualize data**
- **Want to correlate with external and internal data sets**
  - E.g., brand preference vs census district income
- **Want to do deep analysis of content**
  - What product, show, or person is being discussed
  - What opinion is being expressed ("sentiment analysis")

Spatial Correlations

Requirement: Interactive learning and statistics



Source: http://www.richblockspoorblocks.com

"Big Data" and Twitter

- **Volumes and rates are massive; need new tools to interactively visualize data**
- **Want to correlate with external and internal data sets**
  - E.g., brand preference vs census district income
- **Want to do deep analysis of content**
  - What product, show, or person is being discussed
  - What opinion is being expressed ("sentiment analysis")

Tackling The Challenges of Big Data

Use Case: Visualizing Twitter

Introduction to Twitter

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling The Challenges of Big Data

Visualizing Twitter

MapD Demo

Samuel Madden

Professor and Director of Big Data at CSAIL

Massachusetts Institute of Technology

Part 2: MapD Demo

- **What is MapD?**
  - GPU Accelerated Database
  - With a Twitter Visualization in Front of it
- **What Will You See**
  - ~ 50 M tweets displayed
  - Nothing is pre-computed
- **Designed to tackle the "volume" and "velocity" challenges**

Tackling The Challenges of Big Data

Visualizing Twitter

MapD Demo

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling The Challenges of Big Data

Visualizing Twitter

How MapD Works

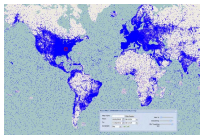
Samuel Madden

Professor and Director of Big Data at CSAIL

Massachusetts Institute of Technology

How Does MapD Work?

- ***Key insight:*** **GPUs have enough memory that a cluster of them can store substantial amounts of data**
- **Not an accelerator, but a full blown SQL Database!**



147,201,658 tweets from Oct 1, 2012 to Nov 6, 2012

Relative intensity of "tornado" on Twitter (with point overlay) from Febuary 29, 2012 to March 1, 2012

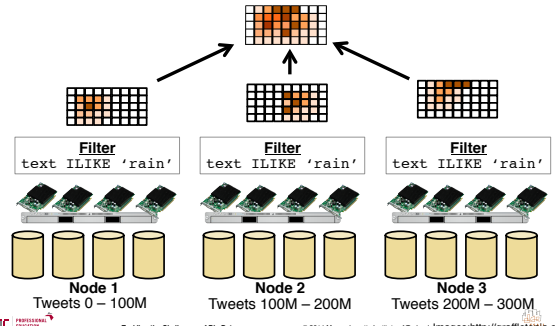
GPUs

- **Massive parallelism enables interactive browsing interfaces**
  - High End GPUs can provide
    - **250 GB/sec of bandwidth**
      - 5X conventional microprocessors
    - **4 Teraflops compute**
      - 10X conventional multi-core microprocessor
  - 7 – 70x speedup in database ops
- **Challenges**
  - Limited Memory on GPUs – But growing!
  - Limited bandwidth between CPU and GPU – Will Change!

"Shared Nothing" Processing

Multiple GPUs, with data partitioned between them

```
Query: Heatmap tweets containing "rain"
```



Images:http://graffletopia.com

Tweet Indexing on GPU

- **Encode tweets using a "dictionary"**

Filter
`text ILIKE 'rain'`

→

Filter
`SELECT tweetid FROM words WHERE id = 57663`

Word	Encoding
…	…
Rain	57663
Rainbow	57664
Rainman	57665
Rainy	57666
…	…

Example: Filtering in Parallel

- **oriented execution**
  - Avoids wasting memory bandwidth
- **Plan:**
  - Reduce bitmap of tweets to read
  - Read tweets, increment output bins

Filter
`SELECT tweetId FROM words WHERE id = 57663`

TweetId	WordId
…	…
1	57663
2	57664
2	27
3	8841
…	…

TweetId	Lat	Lon
…	…	
1	-41.5	23.1
2	-41.7	77.4
3	-37.4	48.2
4	28.4	-44.0
…		

Data Tables Reside in GPU Memory

Example: Filtering in Parallel

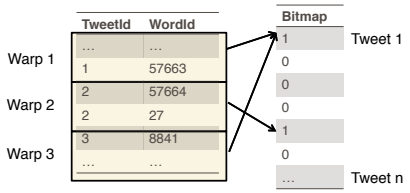
- **1000+ GPU Threads**
- **Running in "Warps"**
- **Threads in same warp run exactly the same instructions**
  - Balanced data per thread improves performance

	TweetId	WordId
	…	…
Warp 1	1	57663
Warp 2	2	57664
	2	27
Warp 3	3	8841
	…	…

Bitmap	
0	Tweet 1
0	
0	
0	
0	
0	
…	Tweet n

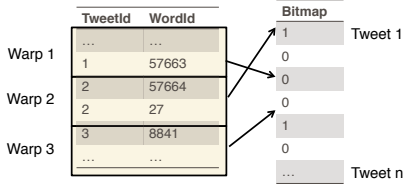
Example: Filtering in Parallel

- 1000+ GPU Threads
- Running in "Warps"
- Threads in same warp run exactly the same instructions
  - Need same amount of data to be efficient



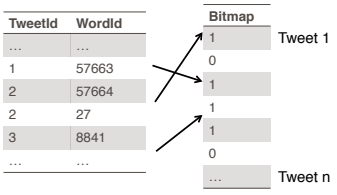
Example: Filtering in Parallel

- 1000+ GPU Threads
- Running in "Warps"
- Threads in same warp run exactly the same instructions
  - Need same amount of data to be efficient



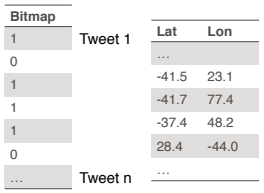
Example: Filtering in Parallel

- 1000+ GPU Threads
- Running in "Warps"
- Threads in same warp run exactly the same instructions
  - Need same amount of data to be efficient



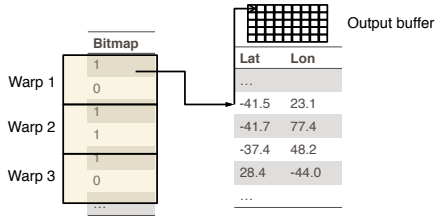
Example: Filtering in Parallel

- **1000+ GPU Threads**
- **Running in "Warps"**
- **Threads in same warp run exactly the same instructions**
  - Need same amount of data to be efficient



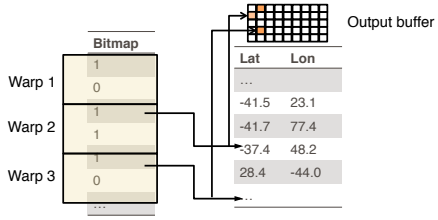
Example: Filtering in Parallel

- **1000+ GPU Threads**
- **Running in "Warps"**
- **Threads in same warp run exactly the same instructions**
  - Need same amount of data to be efficient



Example: Filtering in Parallel

- **1000+ GPU Threads**
- **Running in "Warps"**
- **Threads in same warp run exactly the same instructions**
  - Need same amount of data to be efficient



Parallel Plumbing

Once parallel versions of all operators are built, programmers can just think in terms of SQL

E.g.,

```
SELECT heatmap(lat,lon)
WHERE tweettext ILIKE "rain"
```

Scaling to 100's of millions of points in milliseconds

Summary

- ***Interactivity*** **can create a qualitative difference**
- **GPUs can dramatically accelerate some tasks**
  - **Compute intensive** tasks via parallelism
  - **Data intensive** tasks via increased memory bandwidth
- **Clever use of hardware enables dramatic speedups**
- **Arrays of multiple GPUs help solve large problems**

Try it yourself:
http://mapd.csail.mit.edu

Tackling The Challenges of Big Data

Visualizing Twitter

How MapD Works

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling The Challenges of Big Data

Visualizing Twitter

Other Approaches to Interactive Analytics

Samuel Madden

Professor and Director of Big Data at CSAIL

Massachusetts Institute of Technology

Other Approaches to Interactive Analytics

- **Interactive Analytics Important in Many Situations**
- **Examples**
  - Fault Diagnostics
  - Intrusion Detection
  - Financial Analysis
  - Online Advertising
- **GPUs & Massive Parallelism are Just One Way Approach**

Other Techniques We'll Cover

- **Brute Force / Massive Parallelism**
  - MapD (Hadoop, and many others)
- **Partitioning**
  - Split the data, operate on the part you need
- **Sampling**
  - Operate on a Subset of the Data
- **Summarization**
  - Operate on a Summary of the Data

Partitioning

Sampling

Summarization

Other Techniques We'll Cover

- **Brute Force / Massive Parallelism**
  - MapD (Hadoop, and many others)
- **Partitioning**
  - Split the data, operate on the part you need
- **Sampling**
  - Operate on a Subset of the Data
- **Summarization**
  - Operate on a Summary of the Data

Tackling The Challenges of Big Data
Visualizing Twitter

THANK YOU

PROFESSIONAL EDUCATION

CSAIL