map-D A GPU Database for Real-Time Big Data Analytics and Interactive Visualization
map-D A GPU Database for Real-Time Big Data Analytics and Interactive Visualization

SC13 Denver
#mapDsc13

Tom Graham
Todd Mostak
map-D? super-fast database built into GPU memory

Do? world’s fastest real-time big data analytics interactive visualization

Demo? twitter analytics platform 1 billion+ tweets milliseconds
Location services Switch on

- GPS
- Lat/Lon
- Metadata
- Twitter’s API
- Map-D
- Tweetmap
Core Innovation

Map-D’s database architecture is integrated into the memory on GPUs

Takes advantage of the memory bandwidth and massive parallelism on multiple GPUs and clusters

Runs 70-1000x faster than other in-memory databases and analytics platforms

Any kind of data
#HAIYAN

1 billion+ tweets on 8 NVIDIA Tesla K40s

2,880 x 8 = 23,040 cores
12 x 8 = 96 GB memory

2.3 TB/sec memory bandwidth
>30 teraflops compute power

Nothing is pre-computed!
Streaming live tweets
Interactive and real-time analytics
map-D overview

- SQL-enabled database (not a GPU accelerator)
- Real-time search of any size dataset in milliseconds
- Interactive visualizations generated on the fly
- Compatible with any type of data
- Scales to any size of dataset
- Live data streams onto the system
- Powered by inexpensive, off-the-shelf hardware
- 1000+ analytic/visualization queries per second
- Optimized for GPUs but also runs on CPUs, Phi, AMD and mobile chips
1 billion+ Tweetmap

500 million tweets a day = 7-10 million ‘geocoded’

Tweet = more than just 140 characters:
- geo coordinates
- timestamp
- user and follower information
- reply information
- #hashtags
- host platform

Tweet volume and velocity is a massive challenge

Need new tools to interactively visualize data
1 billion+ Tweetmap

- Search tweet text
- Search by user
- Live streaming tweets
- Stats + census data
- Animate over time
- Identify trends
- Heatmap
- Chloropleth
- Point map
- Base maps
- Share maps

Chart: Cloud
Variables: Word
Scale: Count
1 billion+ Tweetmap

Correlate with external and internal data sets
• Brand preference vs census district income
• Tweet density by region (chloropleth)

Deep analysis of content
• What product, show, or person is discussed over time
• What opinion is being expressed ‘sentiment analysis’
“Shared Nothing” Processing
Multiple GPUs, with data partitioned between them

Filter

text ILIKE ‘rain’

Node 1

Node 2

Node 3
Tweet Indexing on GPU

Encode tweets using a “dictionary”

Filter
text ILIKE ‘rain’

Filter
SELECT tweetid FROM words
WHERE id = 57663

<table>
<thead>
<tr>
<th>Word</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Rain</td>
<td>57663</td>
</tr>
<tr>
<td>Rainbow</td>
<td>57664</td>
</tr>
<tr>
<td>Rainman</td>
<td>57665</td>
</tr>
<tr>
<td>Rainy</td>
<td>57666</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Filtering in Parallel

- Column-oriented execution avoids wasting memory bandwidth

- Plan:
  - Produce bitmap of tweets to read
  - Read tweets, increment output bins in bitmap

```
SELECT tweet_id FROM words WHERE id = 57663
```

<table>
<thead>
<tr>
<th>TweetId</th>
<th>WordId</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>57663</td>
</tr>
<tr>
<td>2</td>
<td>57664</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>8841</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TweetId</th>
<th>Lat</th>
<th>Lon</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-41.5</td>
<td>23.1</td>
</tr>
<tr>
<td>2</td>
<td>-41.7</td>
<td>77.4</td>
</tr>
<tr>
<td>3</td>
<td>-37.4</td>
<td>48.2</td>
</tr>
<tr>
<td>4</td>
<td>28.4</td>
<td>-44.0</td>
</tr>
</tbody>
</table>

Data Tables Reside in GPU Memory
Filtering in Parallel

• 1000+ GPU threads
• Running in “warps”
• Threads in same warp run the exact same instructions
  • Need same amount of data to be efficient
Filtering in Parallel

- 1000+ GPU threads
- Running in “warps”
- Threads in same warp run the exact same instructions
  - Need same amount of data to be efficient

```
TweetId  WordId
...
1       57663
2       57664
2       27
3       8841
...
```

```
Bitmap
1
0
0
1
0
...
```

Tweet 1
Tweet n
Filtering in Parallel

- 1000+ GPU threads
- Running in “warps”
- Threads in same warp run the exact same instructions
  - Need same amount of data to be efficient

<table>
<thead>
<tr>
<th>TweetId</th>
<th>WordId</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57663</td>
</tr>
<tr>
<td>2</td>
<td>57664</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>8841</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bitmap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Tweet 1
Tweet n
Filtering in Parallel

- 1000+ GPU threads
- Running in “warps”
- Threads in same warp run the exact same instructions
  - Need same amount of data to be efficient

<table>
<thead>
<tr>
<th>TweetId</th>
<th>WordId</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>57663</td>
</tr>
<tr>
<td>2</td>
<td>57664</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>8841</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

```
Tweet 1
1
0
1
1
1
0
...
```

```
Tweet n
```
Filtering in Parallel

- 1000+ GPU threads
- Running in “warps”
- Threads in same warp run the exact same instructions
  - Need same amount of data to be efficient
Filtering in Parallel

- 1000+ GPU threads
- Running in “warps”
- Threads in same warp run the exact same instructions
  - Need same amount of data to be efficient
Filtering in Parallel

- 1000+ GPU threads
- Running in “warps”
- Threads in same warp run the exact same instructions
  - Need same amount of data to be efficient
Effective big data tools

Democratization of big data analytics

Interaction with live data streams

Socialization of data driven insight

Map-D is open source
Map-D is a startup

Supported enterprise-grade database
• Appliance or in the cloud

Platform integration
• Cloudera  |  NVIDIA  |  Software AG

Tailored database and analytics solutions
• Twitter  |  Major League Baseball
           |  Sunlight Foundation  |  Leidos

Free, public big data tools powered by Map-D
• Harvard’s Worldmap  |  National Geographic
              |  Smithsonian Center for Astrophysics  |  MIT CSAIL
Play with our live demo

mapd.csail.mit.edu
Who has been tweeting at SC13?

#mapDsc13
Special thanks

Prof Sam Madden, MIT CSAIL
map-D

1 billion+ Demo in NVIDIA booth

@datarefined

info@map-d.com

map-d.com