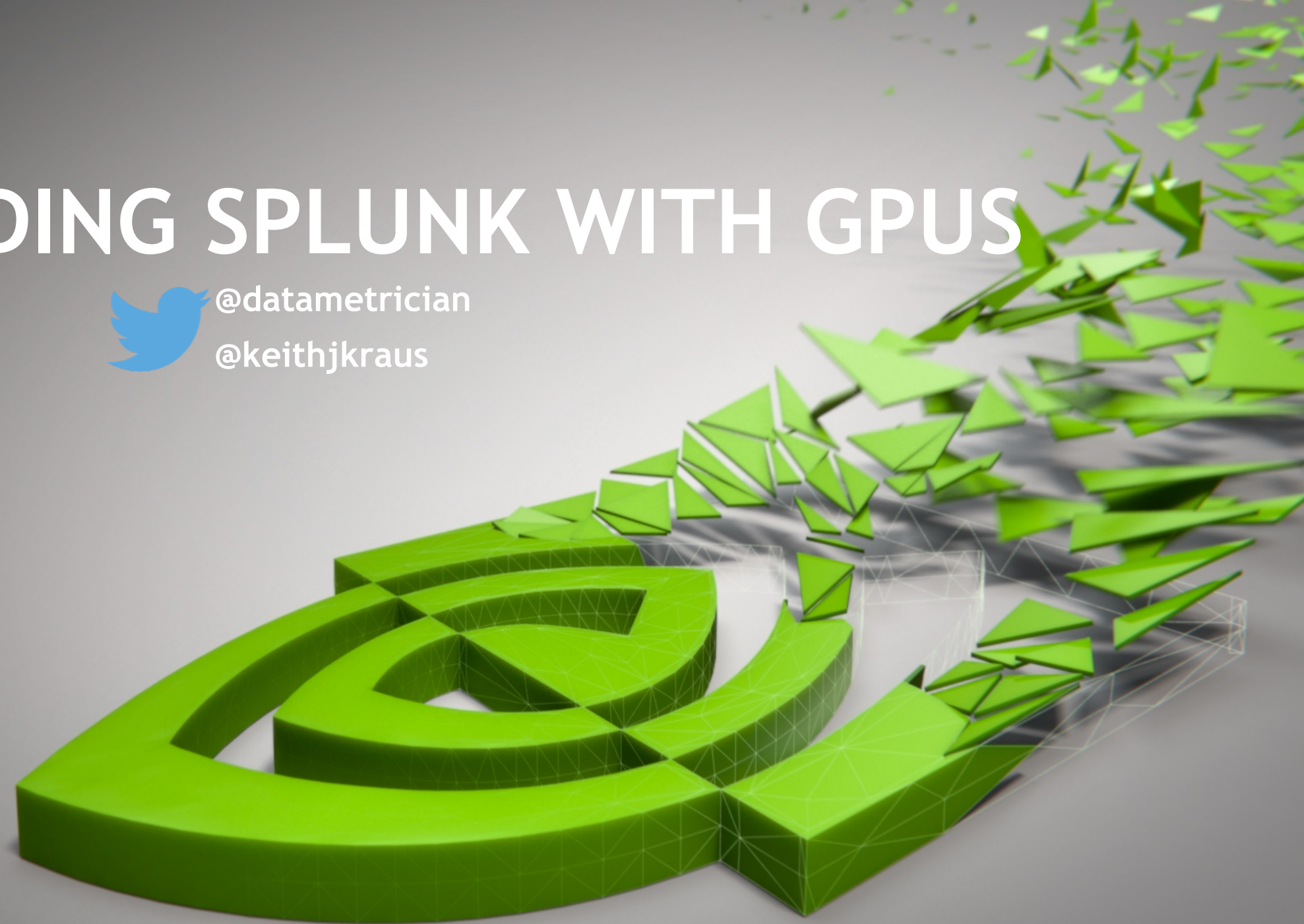
EXTENDING SPLUNK WITH GPUS

Joshua Patterson @datametrician
Keith Kraus @keithjkraus
NVIDIA

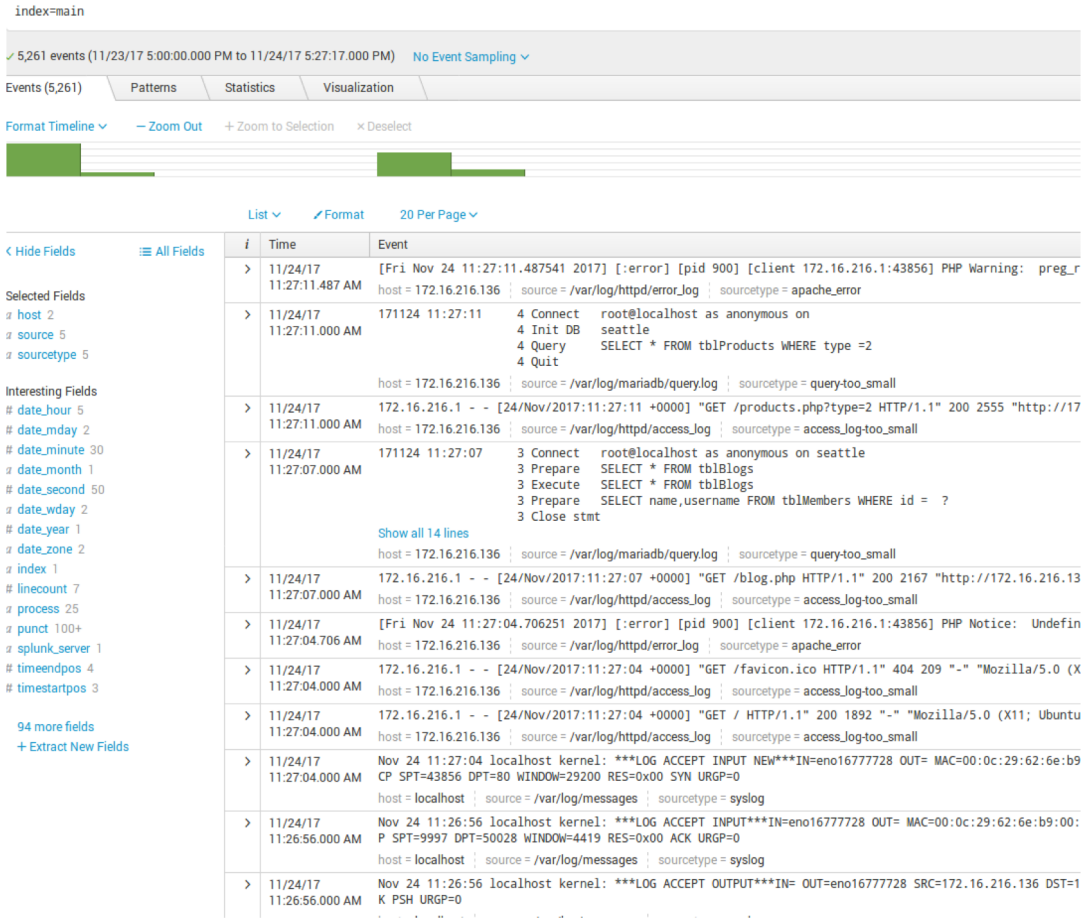


SPLUNK

Industry leading machine data platform

Splunk is a software platform to search, analyze, and visualize the machine-generated data gathered from the websites, applications, sensors, devices, and so on, that comprise your IT infrastructure or business.

splunk>



http://docs.splunk.com/Documentation/Splunk/7.0.2/Overview/AboutSplunkEnterprise

SPLUNK

Industry leading machine data platform turned industry leading SIEM

Security Information and Event Management (SIEM), provides security monitoring, advanced threat detection, forensics and incident management and more.

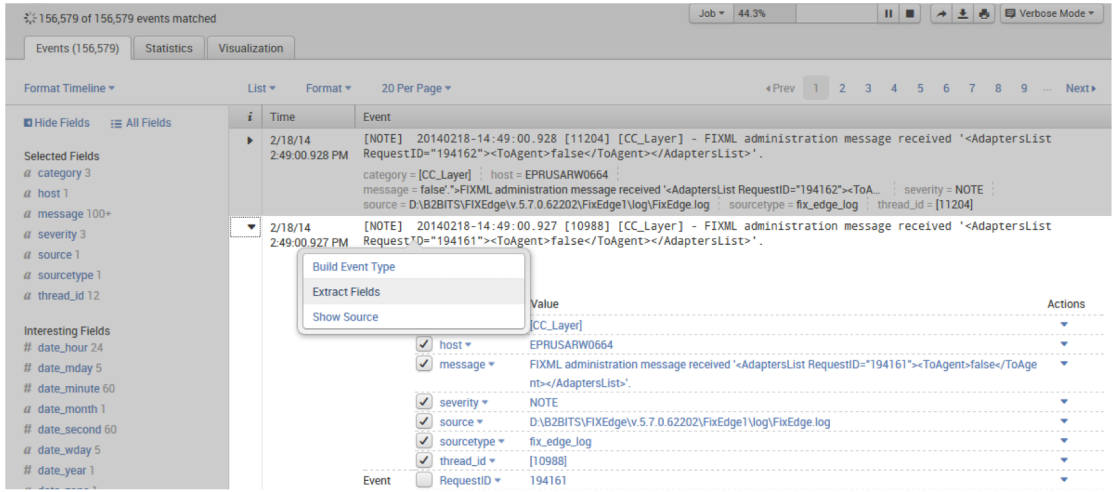
splunk>

https://www.splunk.com/en_us/cyber-security.html

SPLUNK

What makes it an appealing platform?

- Fast transactional searching and querying in a user friendly language
- Security Analysts, Incident Responders, Auditors, etc. are familiar and comfortable using it
- **Turns an unstructured data problem into a structured data problem**



SPLUNK

What could be improved?

- Prohibitively expensive to scale hardware
  - Most enterprise only keep 60-90 days “hot”
- Analytical querying is slow
- Machine learning capabilities based on Scikit-Learn and Apache Spark

FIRST PRINCIPLES OF CYBER SECURITY

Where the industry must go

1. **Indication of compromise needs to improve as attacks are becoming more sophisticated, subtle, and hidden in the massive volume and velocity of data. Combining machine learning, graph analysis, and applied statistics, and integrating these methods with deep learning is essential to reduce false positives, detect threats faster, and empower analyst to be more efficient.**

2. **Event management is an accelerated analytics problem, the volume and velocity of data from devices requires a new approach that combines all data sources to allow for more intelligent/advanced threat hunting and exploration at scale across machine data.**

3. **Visualization will be a key part of daily operations, which will allows analyst to label and train Deep Learning models faster, and validate machine learning prediciton.**

FIRST PRINCIPLES OF CYBER SECURITY

Where the industry must go

1. **Indication of compromise needs to improve as attacks are becoming more sophisticated, subtle, and hidden in the massive volume and velocity of data. Combining machine learning, graph analysis, and applied statistics, and integrating these methods with deep learning is essential to reduce false positives, detect threats faster, and empower analyst to be more efficient.**

2. Event management is an accelerated analytics problem, the volume and velocity of data from devices requires a new approach that combines all data sources to allow for more intelligent/advanced threat hunting and exploration at scale across machine data.

3. Visualization will be a key part of daily operations, which will allows analyst to label and train Deep Learning models faster, and validate machine learning prediciton.

RULES DON'T SCALE

Current methods are too slow

Right now, financial services reports it takes an average of **98 days to detect an Advance Threat** but retailers say it **can be about seven months.**

Once the **security community moves beyond** the mantras **"encrypt everything"** and **"secure the perimeter,"** it can begin developing ***intelligent prioritization and response plans to*** various kinds of **breaches** – with a strong focus on integrity.

The challenge lies in **efficiently scaling** these technologies for ***practical deployment***, and making them **reliable for large networks.** This is where the security community should focus its efforts.

http://www.wired.com/2015/12/the-cia-secret-to-cybersecurity-that-no-one-seems-to-get/

ATTACKS ARE MORE SOPHISTICATED

How Hackers Hijacked a Bank's Entire Online Operation

https://www.wired.com/2017/04/hackers-hijacked-banks-entire-online-operation/

DISCOVERING UNKNOWN THREATS

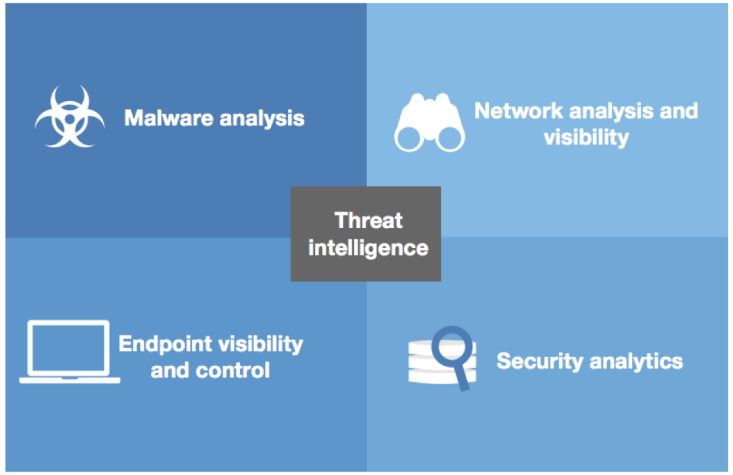
Current methods aren't fast enough

The SIEM & Advanced Analytics layer is where Cyber Security Analytics primarily focuses (all CPU based):

- Apache Spot
- Apache Metron
- ELK

The final stage is Deep Learning:

- Fortune 500 companies have outgrown traditional SIEM and need to move to AI quickly to identify threats
- New technologies are emerging in anomaly detection and network analysis, but they still rely on CPU-based architectures. End to end GPU acceleration will allow them to migrate to an accelerate platform.
- A need to bring it all together, but hyper scale is expensive.



DISCOVERING UNKNOWN THREATS

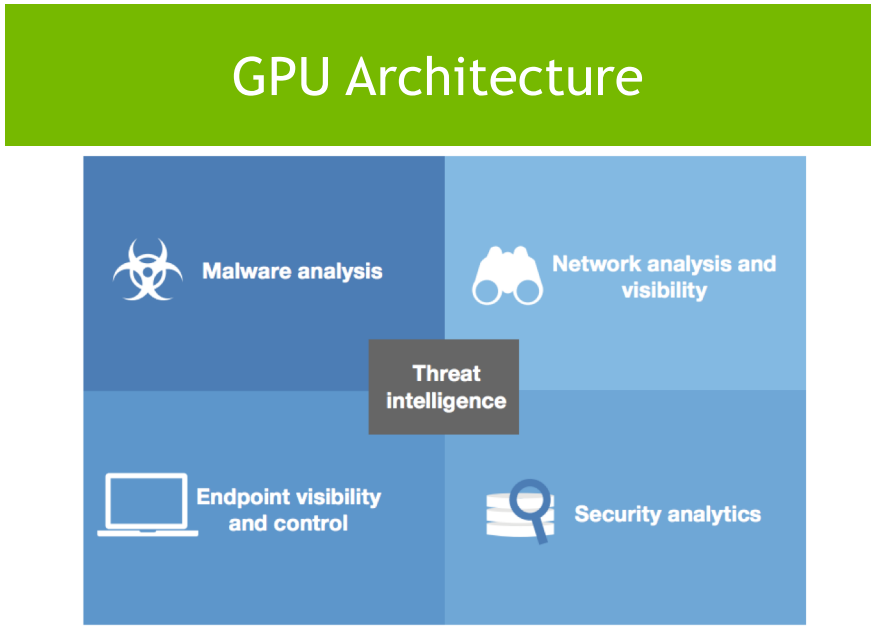
Bringing the data pipeline together with GPU

The SIEM & Advanced Analytics layer is where Cyber Security Analytics primarily focuses (all CPU based):

- Apache Spot
- Apache Metron
- ELK

The final stage is Deep Learning:

- Fortune 500 companies have outgrown traditional SIEM and need to move to AI quickly to identify threats
- New technologies are emerging in anomaly detection and network analysis, but they still rely on CPU-based architectures. End to end GPU acceleration will allow them to migrate to an accelerate platform.
- A need to bring it all together, but hyper scale is expensive.



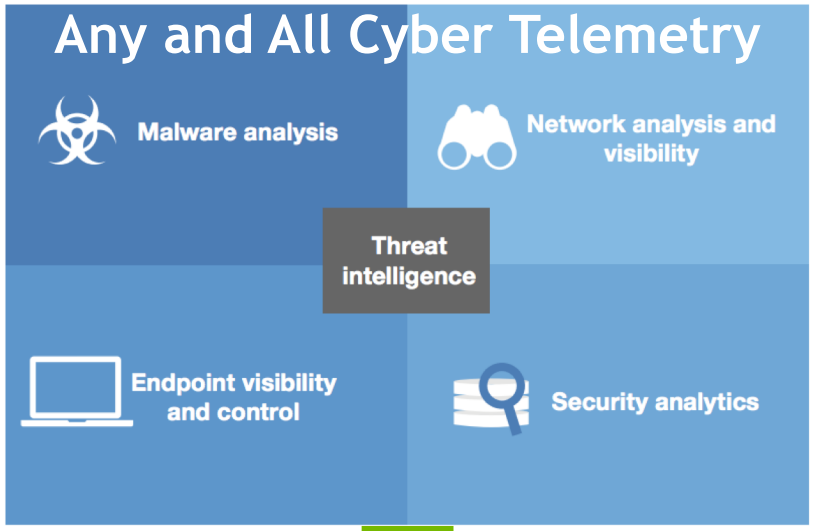
We're building a platform for GPU-Accelerated Machine Learning and Data Analytics.

Not just for cybersecurity, but for other machine data, log, and event problems in general. This architecture will allow speed, scale, and efficiency required for cybersecurity, IOT, and more.

The ultimate goal is GPU acceleration at every level, from streaming to deep learning, in an integrated hardware and software solution.

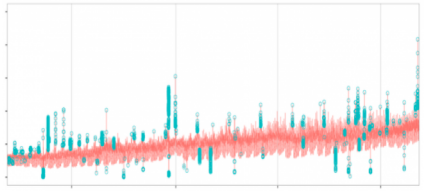
BUILDING INTELLIGENT DEFENSE

AI platform for Machine Data



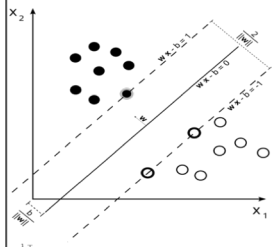
All steps will be GPU accelerated

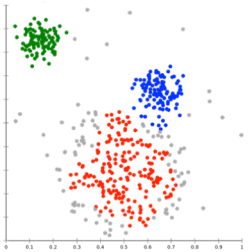
Analysis



TIME SERIES

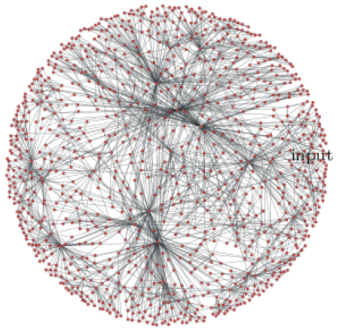
SUPERVISED ML



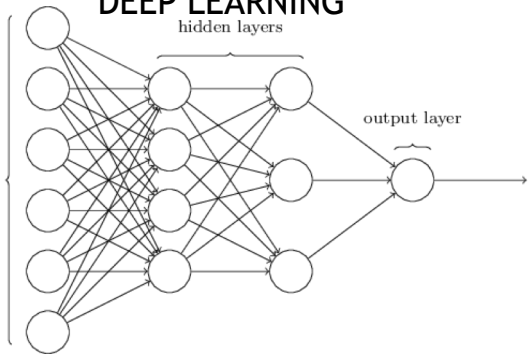


UNSUPERVISED ML

GRAPH ANALYTICS



DEEP LEARNING



Analytics Progression

H2O.ai graphistry splunk> GUNROCK cuSTINGER MAPD ANACONDA NVIDIA nvGRAPH BlazingDB

NVIDIA AND BOOZ ALLEN HAMILTON

Partnership to build enterprise ready cyber security solutions

Booz | Allen | Hamilton

FIRST PRINCIPLES OF CYBER SECURITY

Where the industry must go

1. Indication of compromise needs to improve as attacks are becoming more sophisticated, subtle, and hidden in the massive volume and velocity of data. Combining machine learning, graph analysis, and applied statistics, and integrating these methods with deep learning is essential to reduce false positives, detect threats faster, and empower analyst to be more efficient.

2. Event management is an accelerated analytics problem, the volume and velocity of data from devices requires a new approach that combines all data sources to allow for more intelligent/advanced threat hunting and exploration at scale across machine data.

3. **Visualization will be a key part of daily operations, which will allows analyst to label and train Deep Learning models faster, and validate machine learning predictions.**

VISUALIZATION WITH GPU

Less hardware, more performance, more scale

splunk>

MAPD

1/10th the hardware
1-2 orders of magnitude more performance

MALTEGO

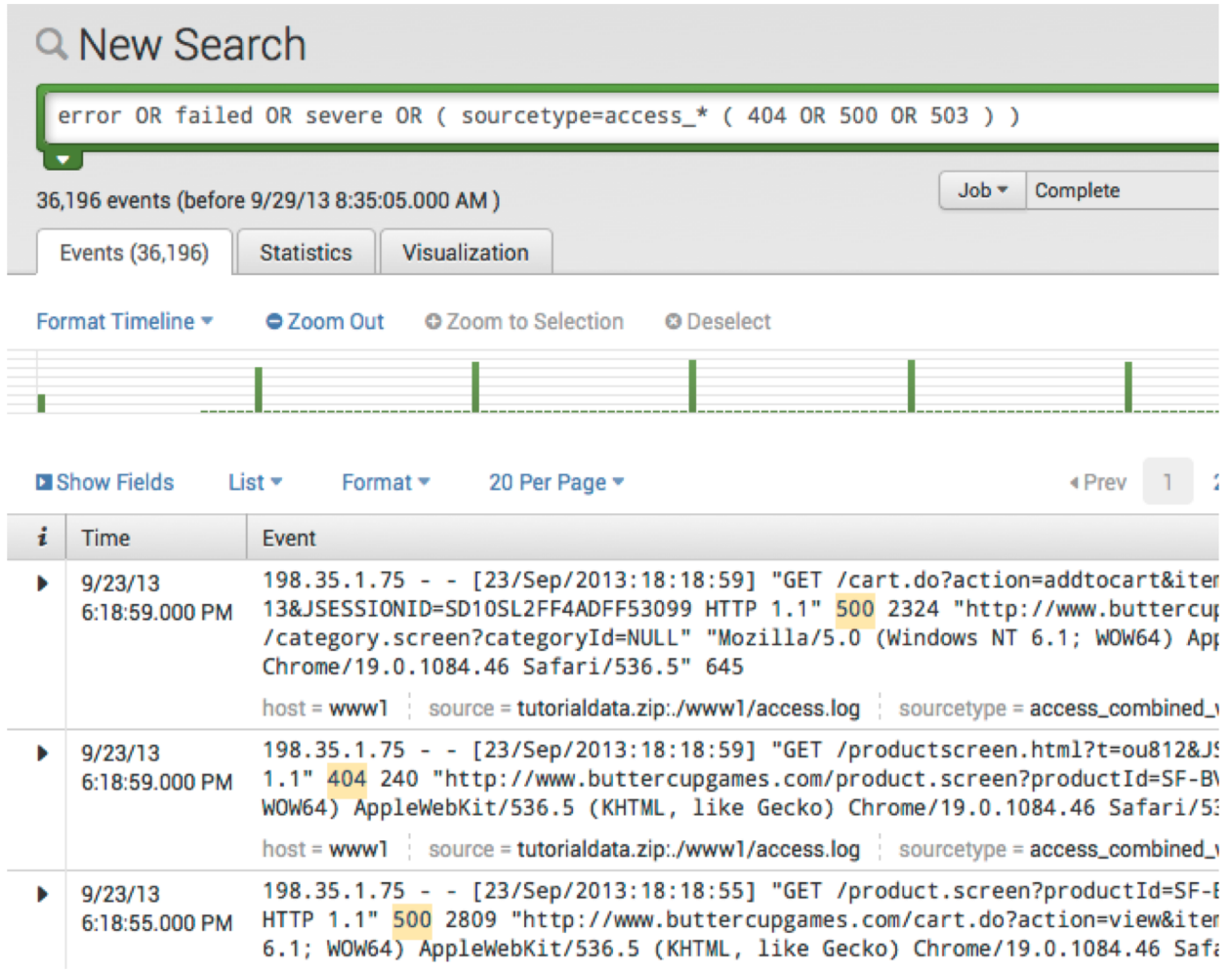
Gephi

graphistry

Real time visualization of 100K+ nodes 1M+ Edges
50-100x faster clustering than other solutions

LISTS DO NOT VISUALLY SCALE

Text search is a great starting point!

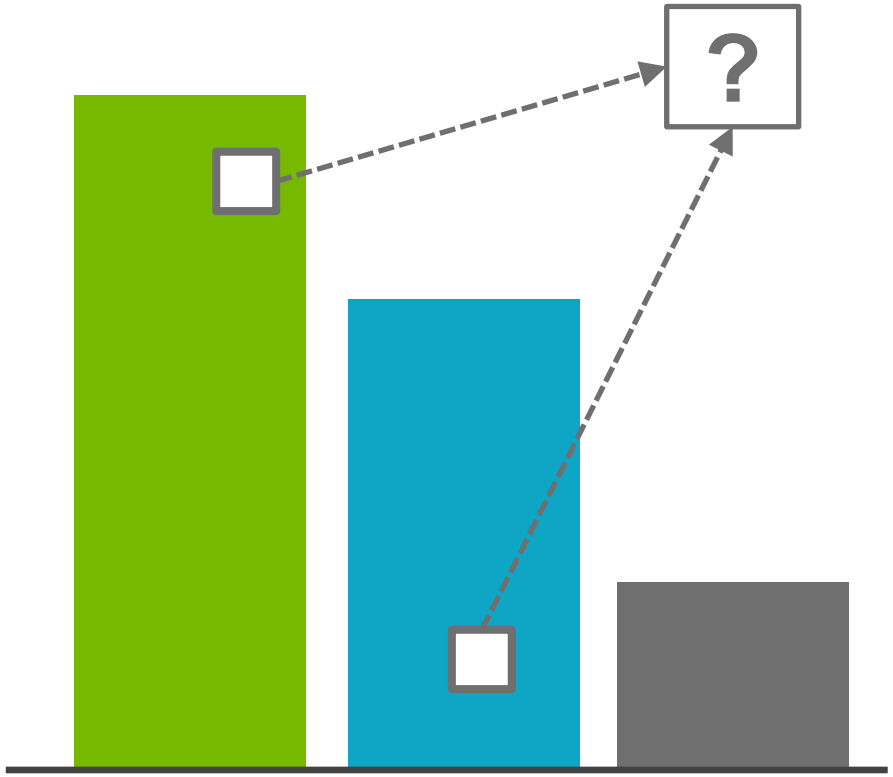


Does not scale
Do not see the 30K+ events nor the IPs, users, nor how they relate…

TRADITIONAL VISUALIZATIONS

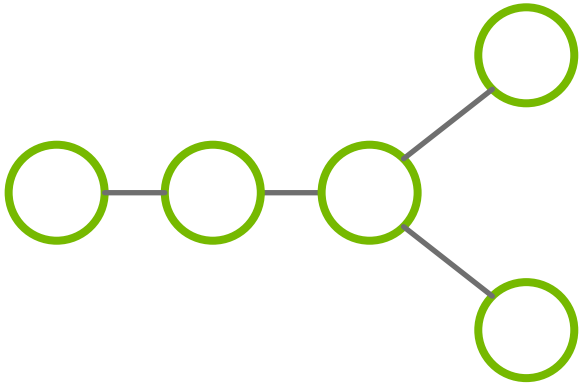
Great for summaries

- Gives overview and ideas for next steps
- Next steps often need granularity that isn't given
- Lose important information about behaviors, relationships, patterns, outliers, etc.

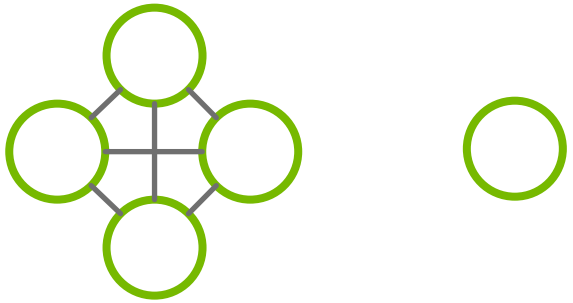


GRAPHS ANSWER IMPORTANT QUESTIONS

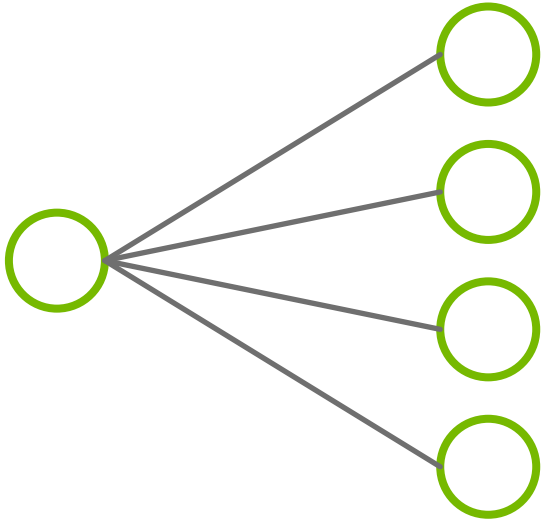
Whereas tables struggle to answer many of these questions effectively



Progression & Behavior



Patterns, Correlations, & Outliers



Entities & Scope

FIRST PRINCIPLES OF CYBER SECURITY

Where the industry must go

1. Indication of compromise needs to improve as attacks are becoming more sophisticated, subtle, and hidden in the massive volume and velocity of data. Combining machine learning, graph analysis, and applied statistics, and integrating these methods with deep learning is essential to reduce false positives, detect threats faster, and empower analyst to be more efficient.

2. **Event management is an accelerated analytics problem, the volume and velocity of data from devices requires a new approach that combines all data sources to allow for more intelligent/advanced threat hunting and exploration at scale across machine data.**

3. Visualization will be a key part of daily operations, which will allows analyst to label and train Deep Learning models faster, and validate machine learning prediciton.

CPUS ARENT FAST ENOUGH

CPUs are the new bottleneck

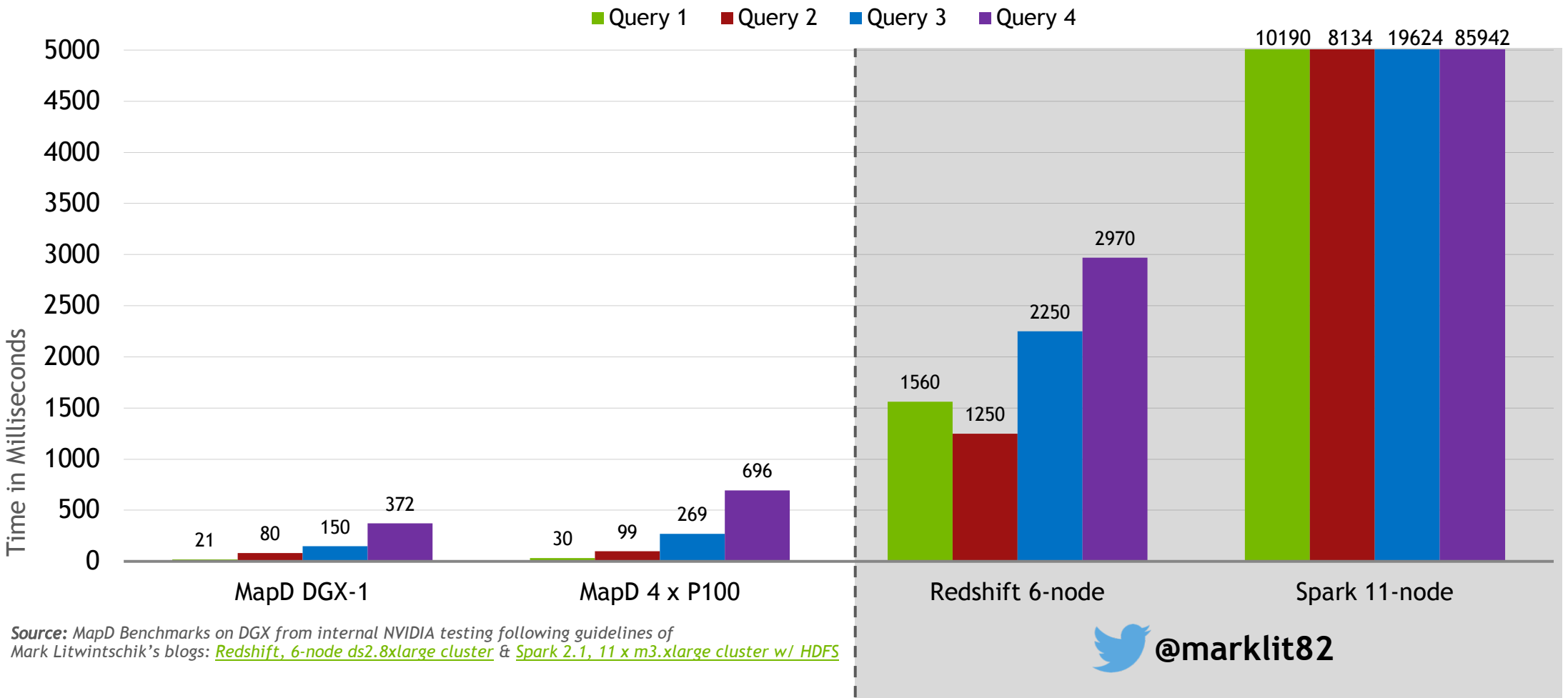
- In a simple benchmark consisting of aggregating data, the CPU is the bottleneck
- The CPU bottleneck is even worse in more complex workloads!

```
top - 08:54:14 up  1:50,  4 users,  load average: 0.20, 1.64, 6.43
Tasks: 360 total,   2 running, 358 sleeping,   0 stopped,   0 zombie
%Cpu0  : 94.7 us,  1.7 sy,  0.0 ni,  3.3 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu1  : 95.0 us,  1.7 sy,  0.0 ni,  3.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu2  : 98.3 us,  0.3 sy,  0.0 ni,  1.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu3  : 87.3 us,  4.3 sy,  0.0 ni,  8.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu4  : 95.0 us,  1.3 sy,  0.0 ni,  3.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu5  : 98.3 us,  0.0 sy,  0.0 ni,  1.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu6  : 96.7 us,  1.3 sy,  0.0 ni,  2.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu7  : 92.7 us,  1.0 sy,  0.0 ni,  5.6 id,  0.3 wa,  0.0 hi,  0.3 si,  0.0
%Cpu8  : 93.7 us,  1.3 sy,  0.0 ni,  5.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu9  : 92.3 us,  0.7 sy,  0.0 ni,  7.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu10 : 97.3 us,  0.7 sy,  0.0 ni,  2.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu11 : 97.3 us,  0.7 sy,  0.0 ni,  2.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu12 : 92.0 us,  3.0 sy,  0.0 ni,  5.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu13 : 94.9 us,  1.0 sy,  0.0 ni,  4.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu14 : 88.3 us,  3.0 sy,  0.0 ni,  8.7 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu15 : 92.6 us,  2.3 sy,  0.0 ni,  4.7 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu16 : 94.7 us,  2.3 sy,  0.0 ni,  2.6 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu17 : 93.0 us,  0.7 sy,  0.0 ni,  6.0 id,  0.0 wa,  0.0 hi,  0.3 si,  0.0
%Cpu18 : 93.0 us,  3.7 sy,  0.0 ni,  3.3 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
%Cpu19 : 91.2 us,  0.7 sy,  0.0 ni,  8.1 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0
```

Source: *Mark Litwintschik's blog: 1.1 Billion Taxi Rides: EC2 versus EMR*

GPUS ARE FAST

1.1 Billion Taxi Ride Benchmark



Source: *MapD Benchmarks on DGX from internal NVIDIA testing following guidelines of Mark Litwintschik's blogs: Redshift, 6-node ds2.8xlarge cluster & Spark 2.1, 11 x m3.xlarge cluster w/ HDFS*

@marklit82

GPUS ARE FAST

TPC-H Join Query Benchmark

TPCH Query 21 – End to End Results Using 32-bit Keys*

TIME (MS)	SF1	SF10	SF100
CPU (single-threaded)	1329	31731	465064
V100 (PCIe3)	22	164	1521
V100 (3xNVLINK2)	12	45	466

300x

3.2x

TPCH Query 4 – End to End Results Using 32-bit Keys*

TIME (MS)	SF1	SF10	SF100
CPU (single-threaded)	150	2041	24960
V100 (PCIe3)	13	105	946
V100 (3xNVLINK2)	7	23	308

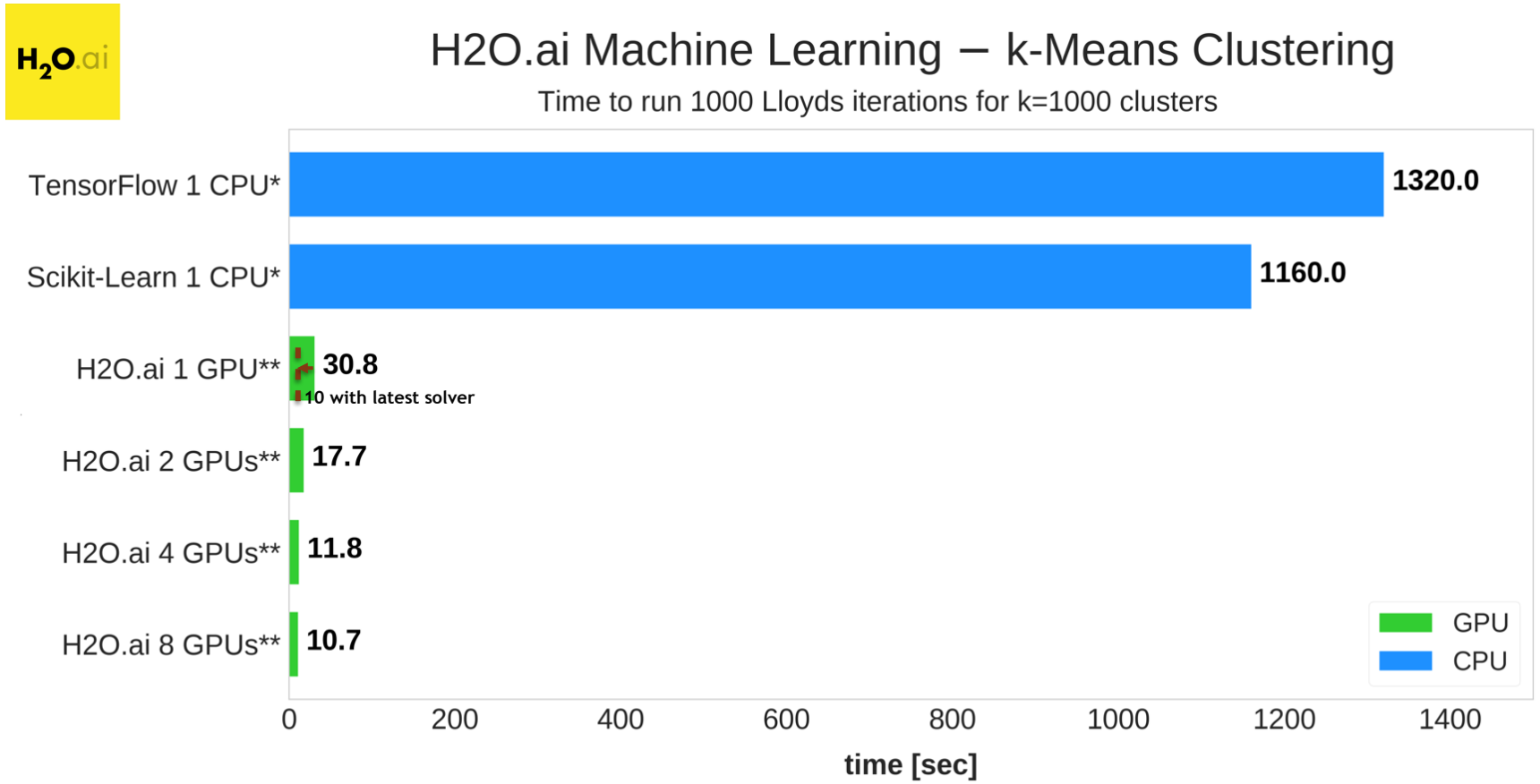
26x

3.1x

Assuming the input tables are loaded and pinned in system memory

GPUS ARE FAST

K-Means Benchmark

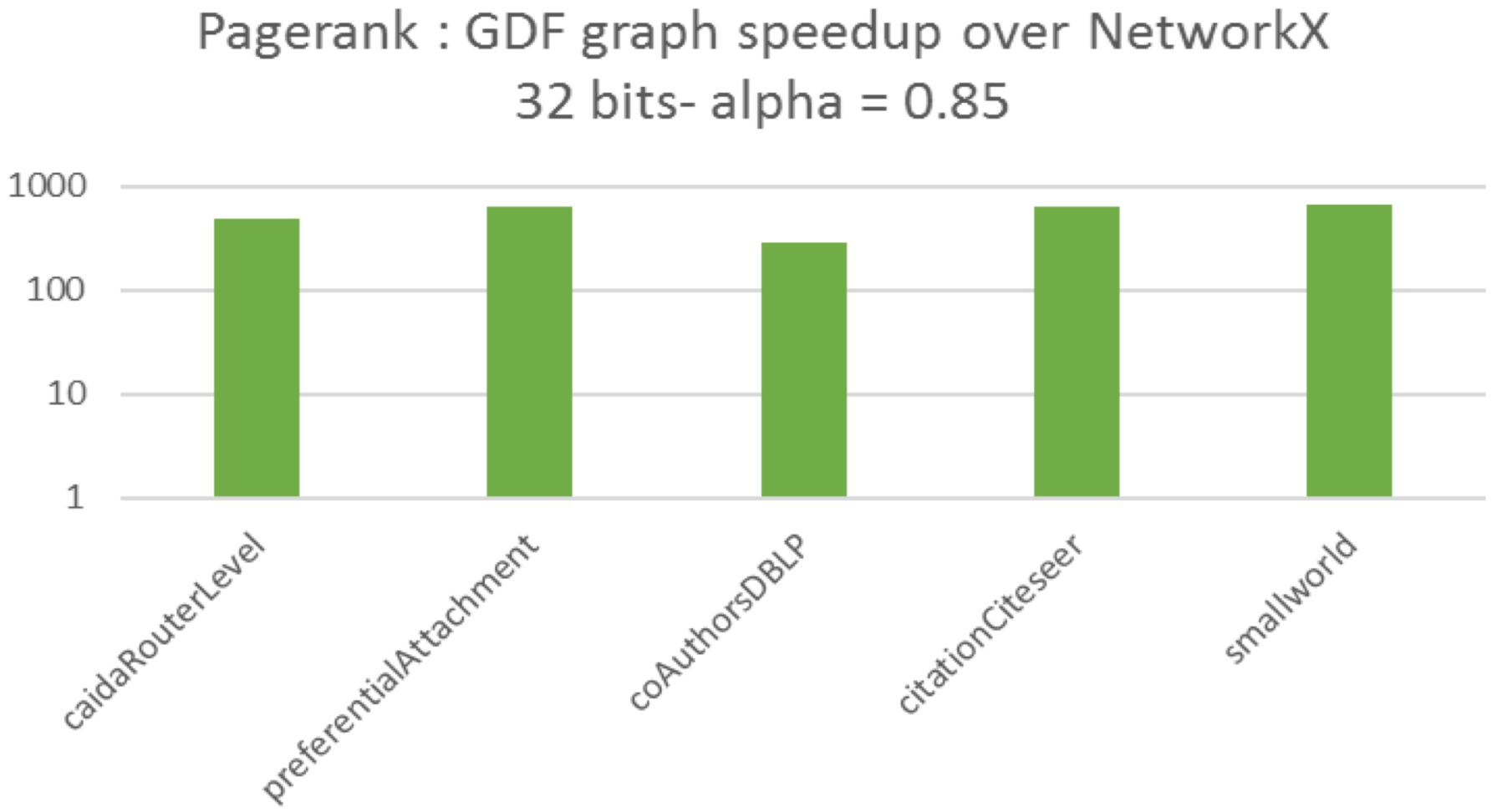


Kaggle Homesite Home Insurance Claims Predictions Dataset (261k rows, 298 cols)
k-Means Clustering (Lloyds), random initialization, 1000 centroids, 1000 iterations
Hardware: *Intel i7 5820k (6-core), **NVIDIA Tesla P100 (DGX-1)

http://github.com/h2oai/perf/

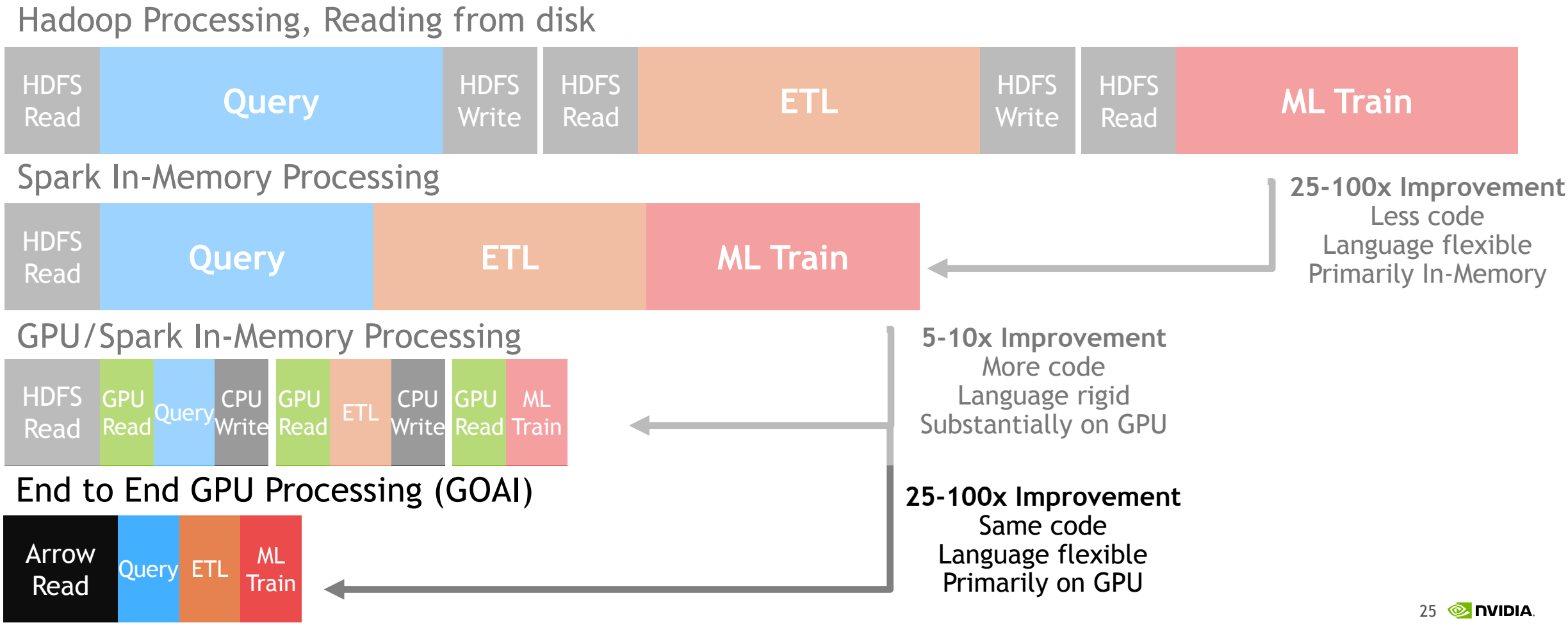
GPUS ARE FAST

nvGRAPH Benchmark



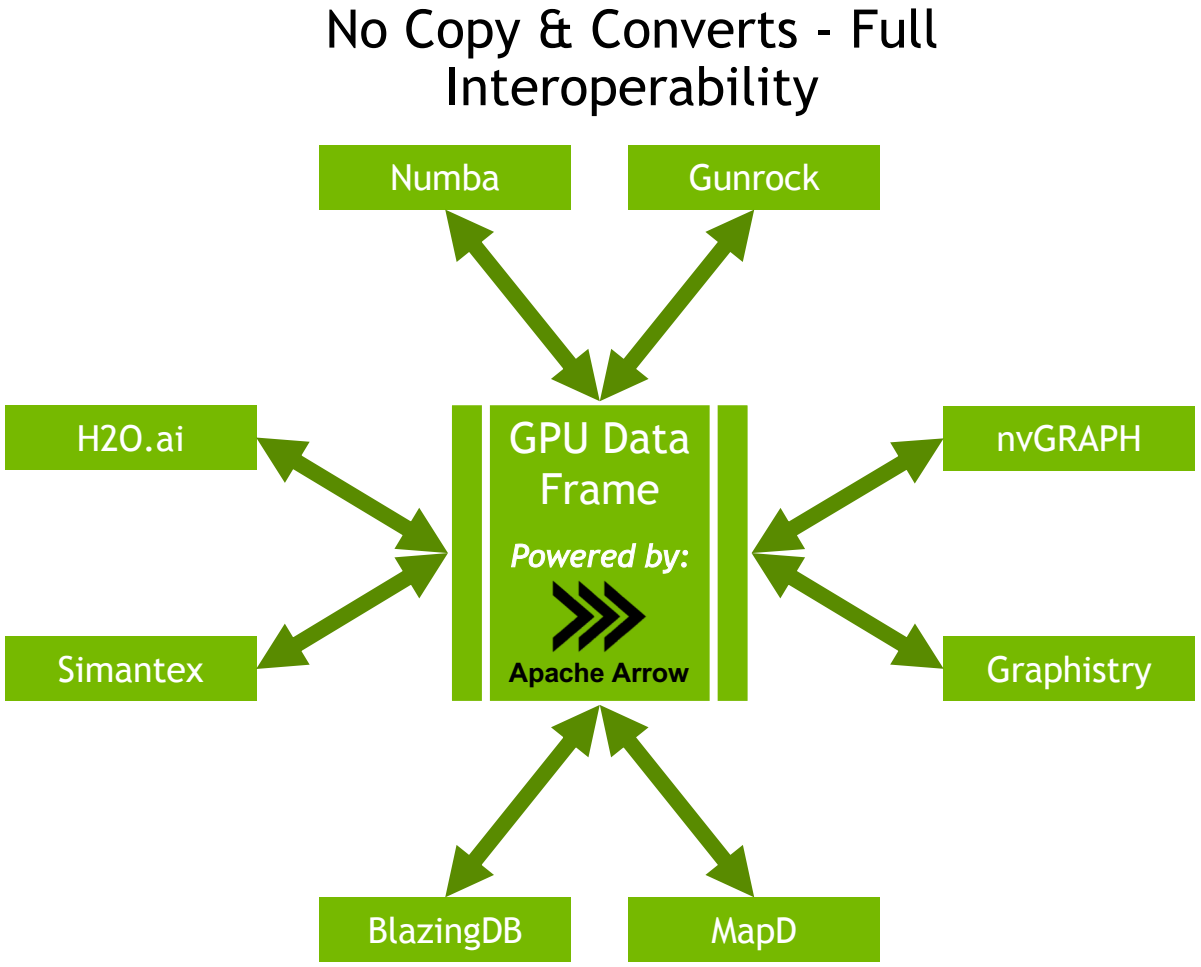
GPU DATA FRAME

Faster Data Access Less Data Movement



GPU OPEN ANALYTICS INITIATIVE

First Project, the GPU Data Frame



github.com/gpuopenanalytics

- GPU Data Frame is the first project of GOAI
- Apache Arrow for GPU
- libgdf: A C library of helper functions, including:
  - Copying the GDF metadata block to the host and parsing it to a host-side struct.
  - Importing/exporting a GDF using the CUDA IPC mechanism.
  - CUDA kernels to perform element-wise math operations on GDF columns.
  - CUDA sort, join, and reduction operations on GDFs.
- pygdf: A Python library for manipulating GDFs
  - Python interface to libgdf library with additional functionality
  - Creating GDFs from Numpy arrays and Pandas DataFrames
  - JIT compilation of group by and filter kernels using Numba
- dask_gdf: Extension for Dask to work with distributed GDFs.
  - Same operations as pygdf, but working on GDFs chunked onto different GPUs and different servers.
  - Will bring the same Kubernetes support that Dask already has.

GOAI ECOSYSTEM

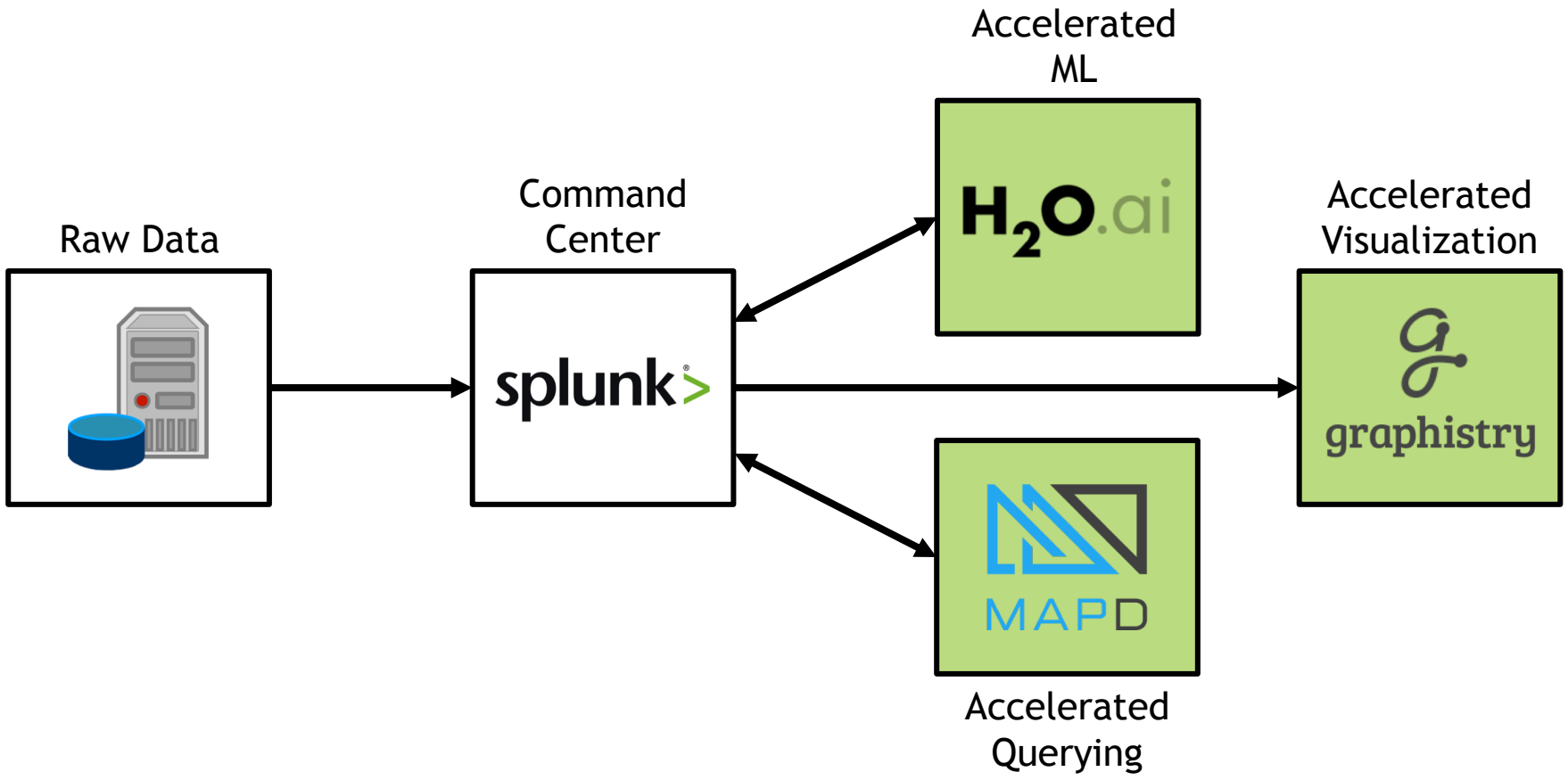
GPU DATABASES

PROCESSING ANALYTICS

GRAPH

INTEGRATING SPLUNK AND GOAI V1

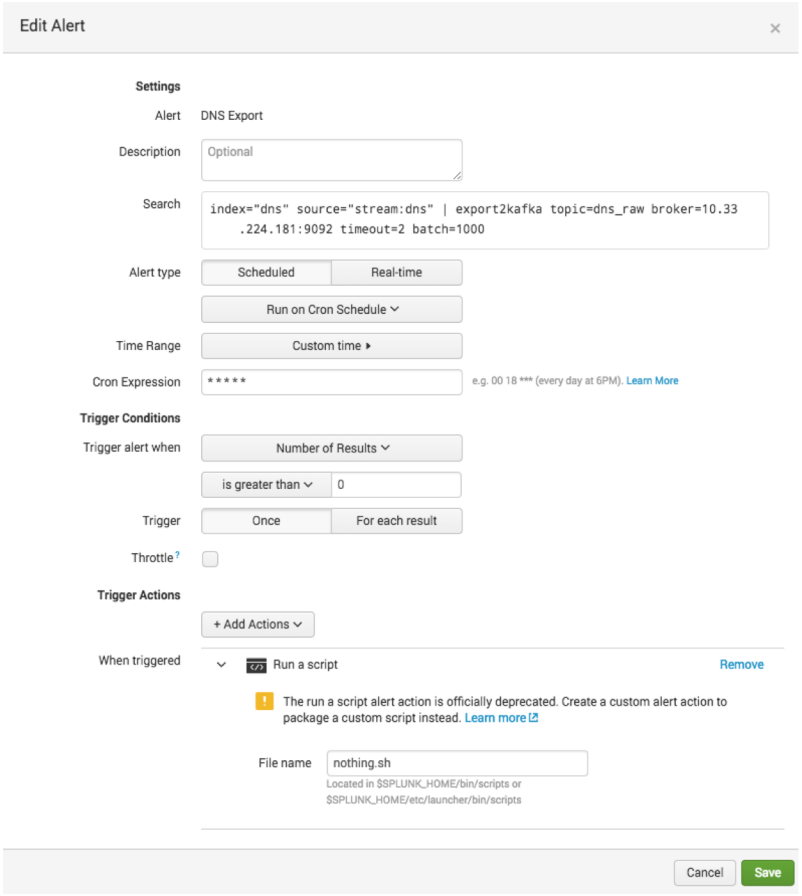
Utilizing GPU-enabled technologies from Splunk



INTEGRATING SPLUNK AND GOAI

Using Splunk alerts to consistently export structured data

- Create an alert with a custom search that you've created
- Create a custom search that exports data to an arbitrary system or location, i.e. Kafka
- Run the alert on a schedule and trigger **Once** when the number of results is greater than 0
  - The trigger action should be an empty script



INTEGRATING SPLUNK AND GOAI

Creating a custom search command for exporting data

- The Splunk Python SDK allows us to do whatever we want with Splunk **structured** search results
- I.E. the code to the right pushes the messages to a Kafka topic
- GPU-accelerated data manipulation and an in-GPU-memory data pipeline powered by GOAI is a possibility in the future

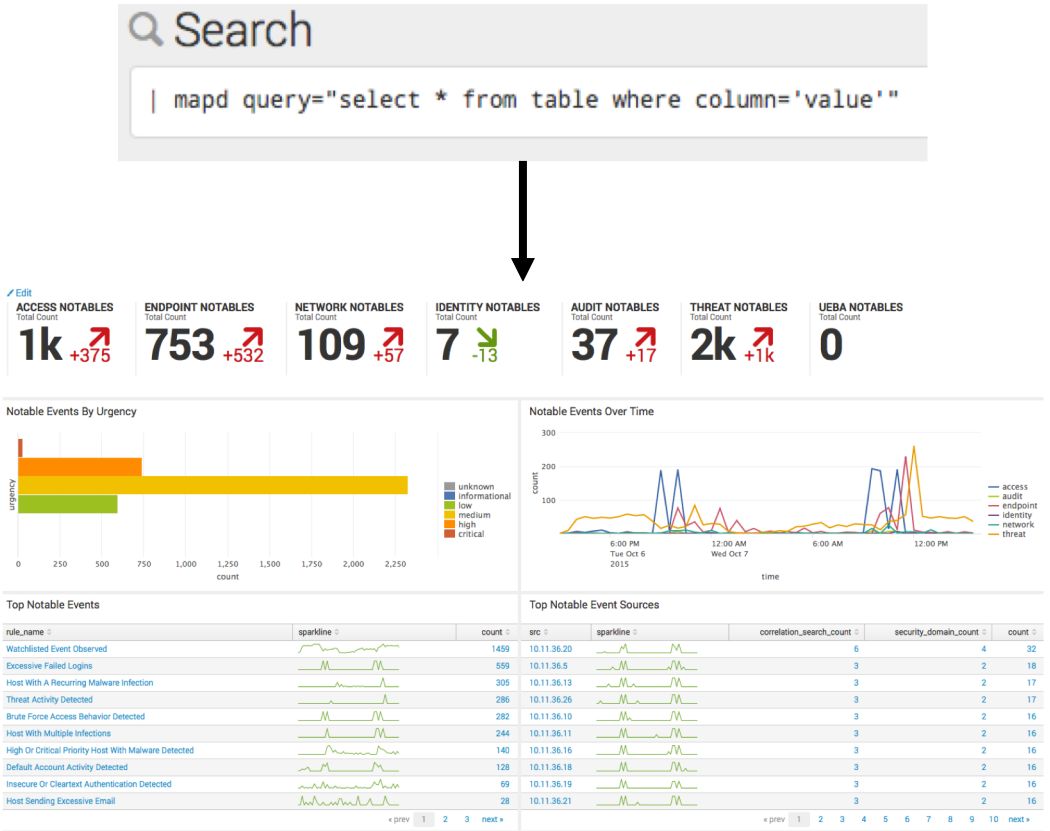
```
class FileSinkCommand(StreamingCommand):
  broker = Option(require=True)
  topic = Option(require=True)
  batch = Option(require=False, default=2000)
  timeout = Option(require=False, default=60)
  pool = Option(require=False, default=2)
  start_time = int(time.time())

  def create_producers(self, pool, broker):
    producers = []
    for i in range(pool):
      producers.append(Producer({'bootstrap.servers': broker,
'session.timeout.ms': 10000}))
    return producers
  def stream(self, records):
    topic = str(self.topic)
    broker = str(self.broker)
    batch = int(self.batch)
    timeout = int(self.timeout)
    pool = int(self.pool)
    producers = self.create_producers(pool, broker)
    cnt = 0

    for record in records:
      trimmed = {k: v for k, v in record.iteritems()}
      producers[cnt % pool].produce(topic, json.dumps(trimmed))
      cnt += 1
      if cnt % batch == 0:
        # batch level reached poll to get producer to move messages out
        for p in producers:
          p.poll(0)

      if cnt % 10 == 0 and int(time.time()) > (60 * timeout) + self.start_time:
        # quit after timeout has been reached, only check every 10 records
        break

      # return record for display in Splunk
      yield record

    for p in producers:
      p.flush()
```

SPLUNK AND MAPD

Accelerating analytical queries

- The Splunk Python SDK allows us to create a custom search command that queries a database such as MapD and return the results as Splunk events

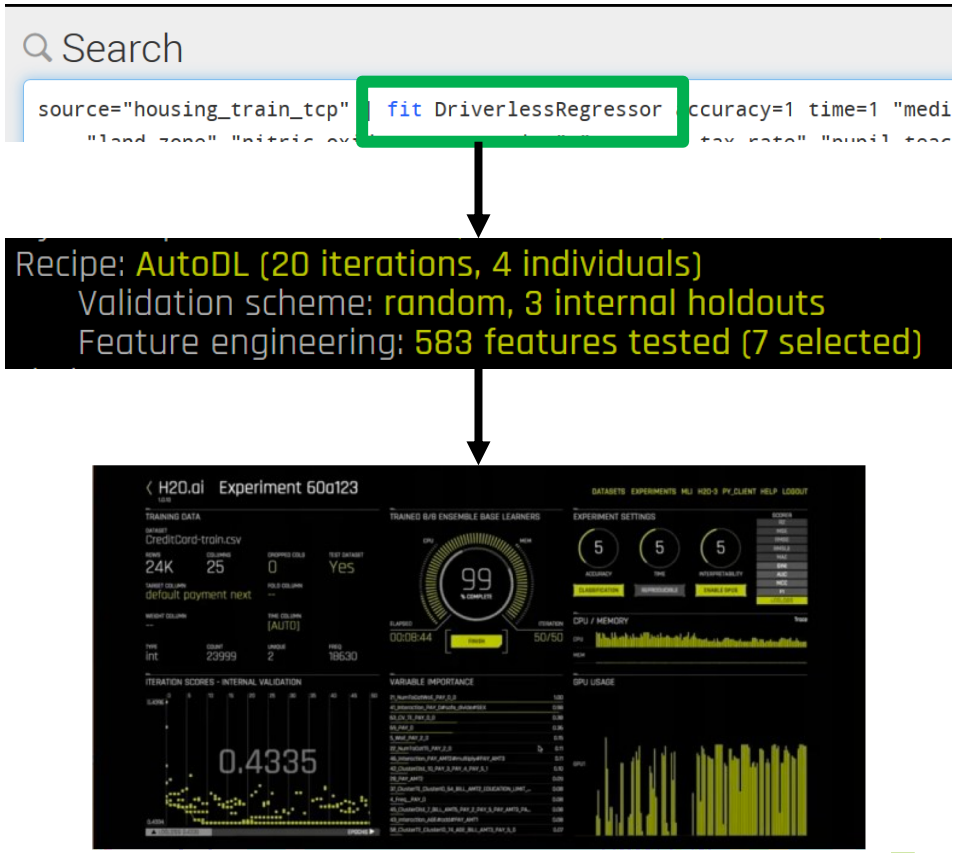
- **The results from the database query are not imported or indexed by Splunk,**
  - **Don’t count against license usage**
  - **Aren’t limited by Search performance**



SPLUNK AND DRIVERLESS AI

Automated machine learning for smarter rules and detections

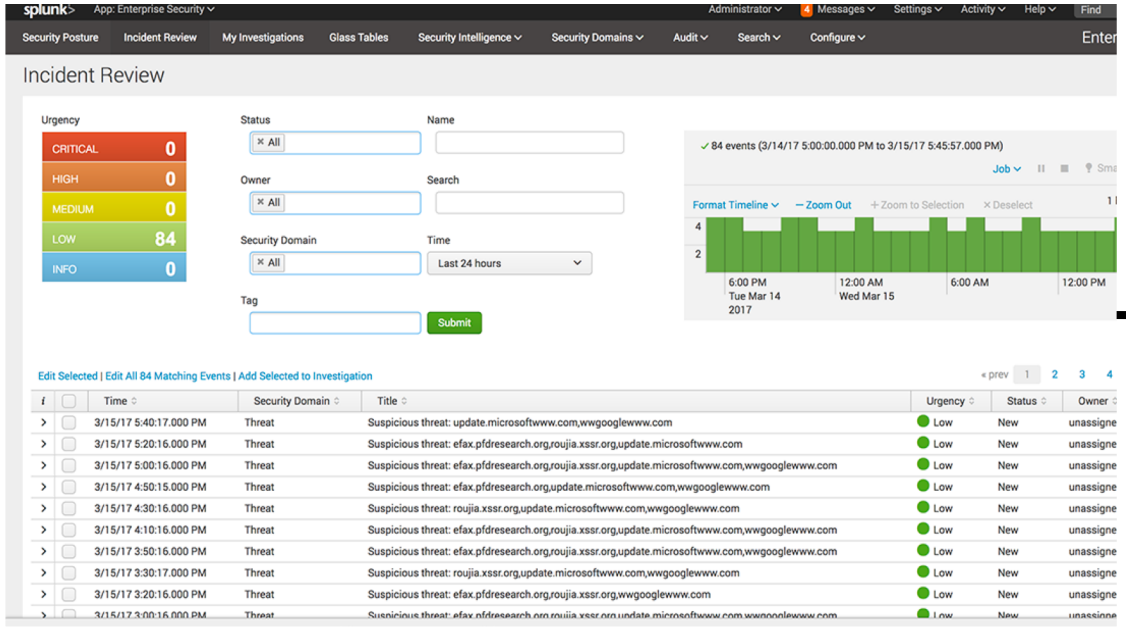
- Gives analysts the power of automated feature engineering and machine learning
- Seamless integration with Splunk in an easy to use custom search command
- Integration allows it to be used anywhere searches are used: dashboards, alerts, reports, etc.



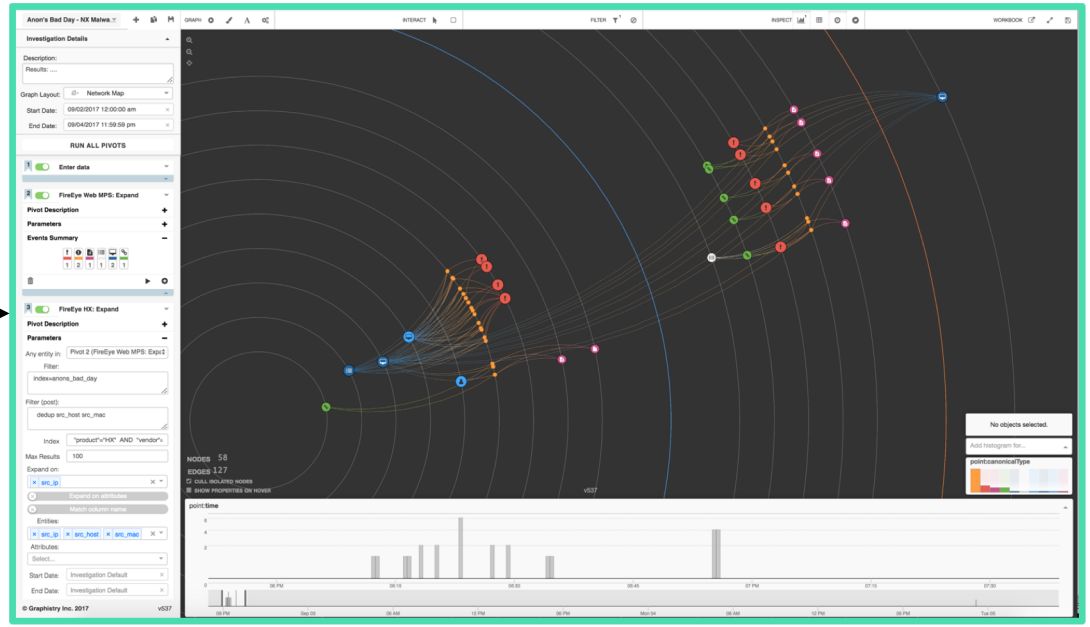
SPLUNK AND GRAPHISTRY

Empowering analysts with an accelerated visual investigation platform

- One click to jump into a visual investigation that allows for interactive drilldown into an incident



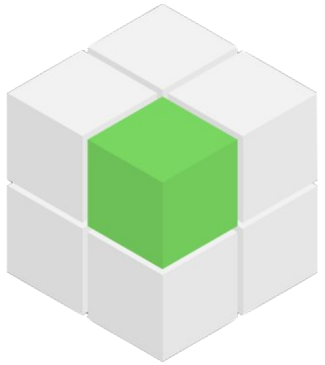
- Intuitive graph layouts give analysts valuable insights as to where to focus their drilldown efforts

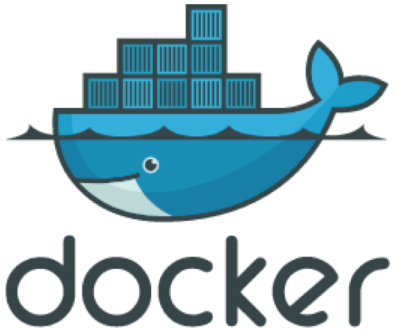


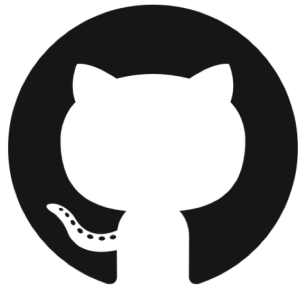
BRINGING IT ALL TOGETHER

Releasing the GPU-ML Container and Splunk code examples

- **GPU-ML Container:** Docker container that has everything to get started with GPU-accelerated data analysis today
- **Splunk code examples:** Kafka export example, and MapD custom search command in the near future
- Check the GPU Open Analytics Initiative Twitter / Github in the coming weeks







FUTURE

MORE GPU ACCELERATION

Continue integrating GPU-accelerated technologies with Splunk

GPU-Accelerated Data Manipulation

Accelerate data massaging in Splunk export process and build towards an in-GPU-memory data pipeline

GPU-Accelerated Scale Out Data Warehousing

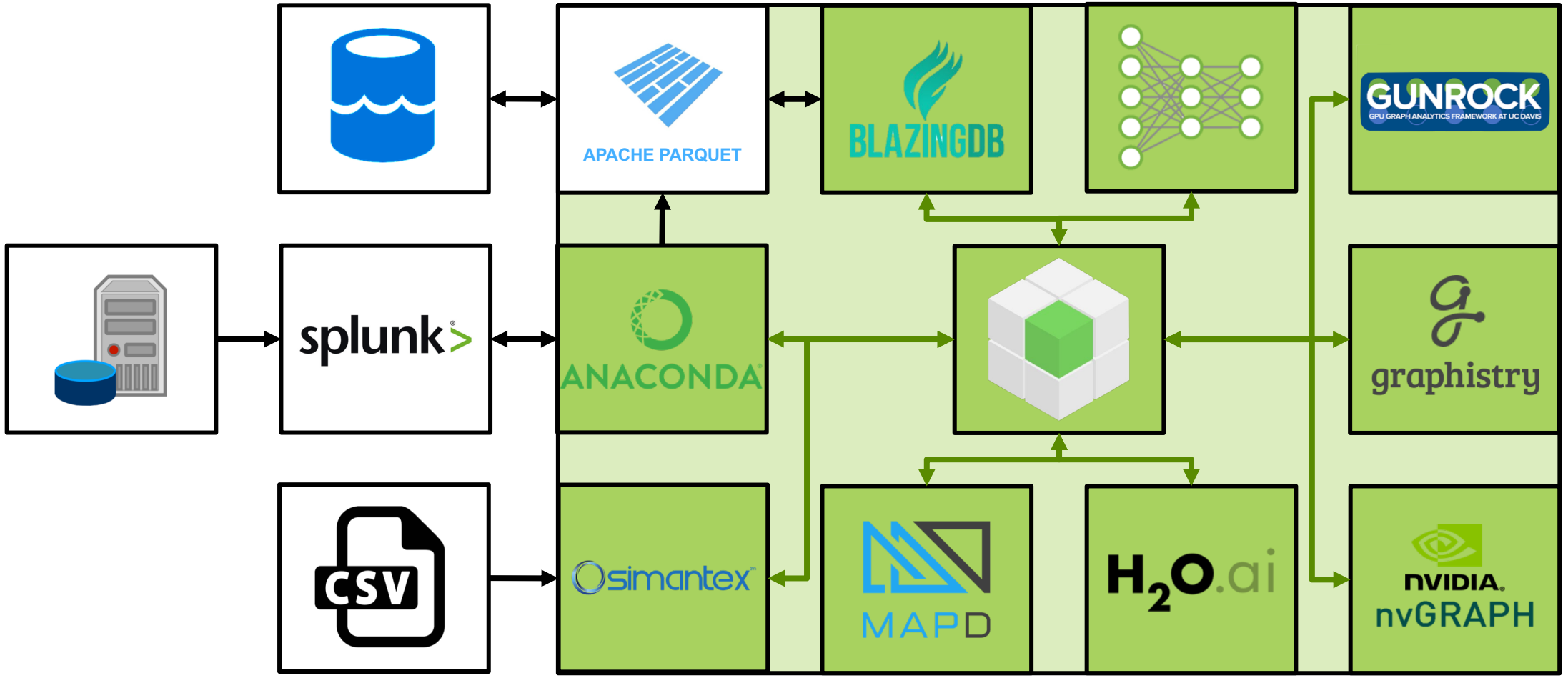
Accelerate queries on large amounts of historical data from both Splunk and traditional data lake

GPU-Accelerated Graph Analytics

Accelerate graph analysis and feature creation for better rules and detection

INTEGRATING SPLUNK AND GOAI V2

Use Splunk as a command center



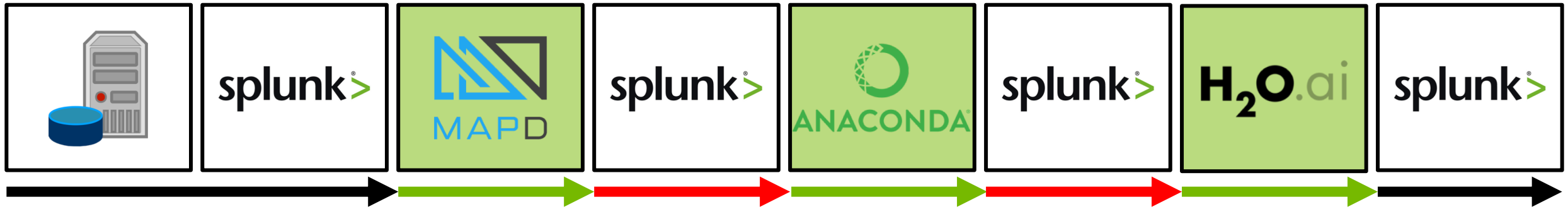
GOAI Powered GPU-Accelerated Core

INTEGRATING SPLUNK AND GOAI V2

Use Splunk as a command center… but not as a data middleman

Search

```
index="mydata" | mapd query="select * from table where column='value'" | custom_feature_engineering | fit DriverlessRegressor
```



INTEGRATING SPLUNK AND GOAI V2

Use Splunk as a command center… but not as a data middleman

Search

```
index="mydata" | mapd query="select * from table where column='value'" | custom_feature_engineering | fit DriverlessRegressor
```

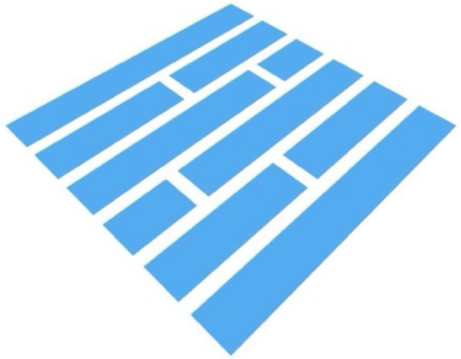
JOIN THE REVOLUTION

Everyone Can Help!

APACHE ARROW

https://arrow.apache.org/

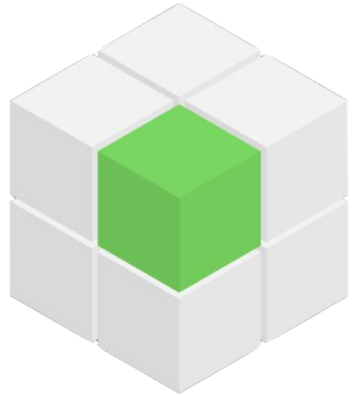
@ApacheArrow



APACHE PARQUET

https://parquet.apache.org/

@ApacheParquet



GPU Open Analytics Initiative

http://gpuopenanalytics.com/

@Gpuoai

Integrations, feedback, documentation support, pull requests, new issues, or donations welcomed!

QUESTIONS?

Joshua Patterson @datametrician
Keith Kraus @keithjkraus
NVIDIA