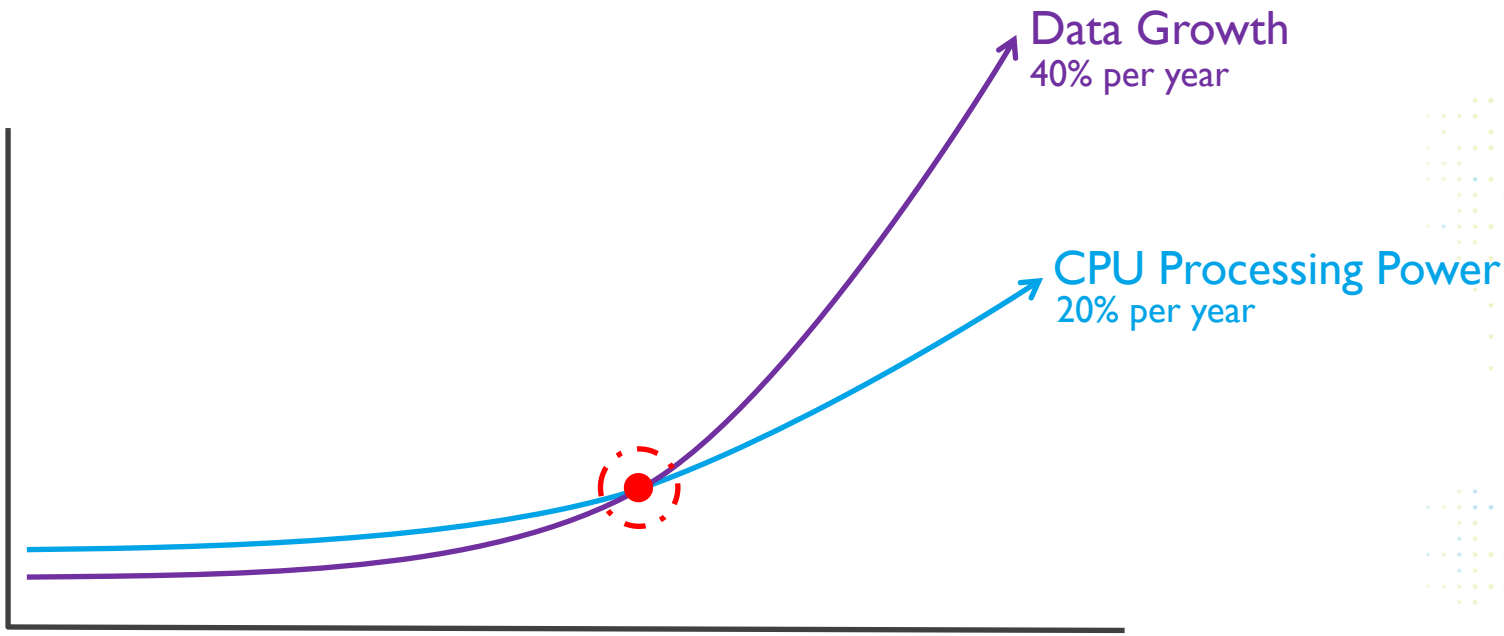
LEVERAGING GPU-ACCELERATED ANALYTICS ON TOP OF APACHE SPARK

June 6, 2017

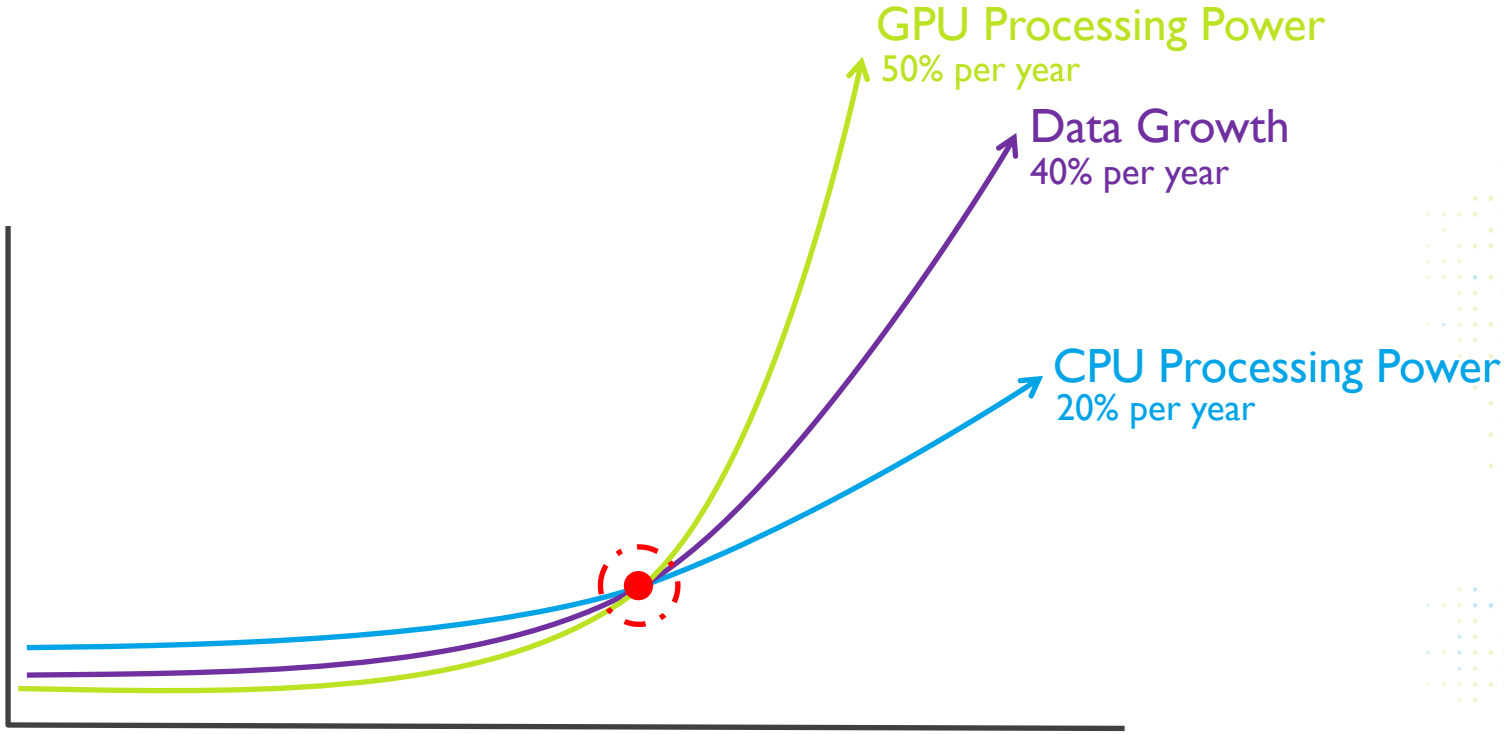
Todd Mostak | Co-founder + CEO, MapD

@toddmostak | @mapd

A compute inflection point



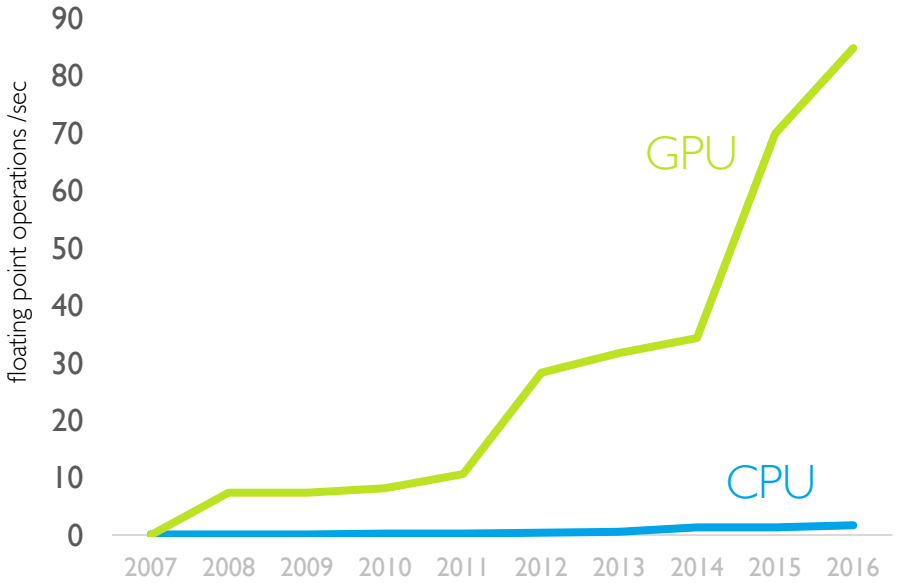
GPUs offer a way forward



GPUs outperform CPUs in data critical areas

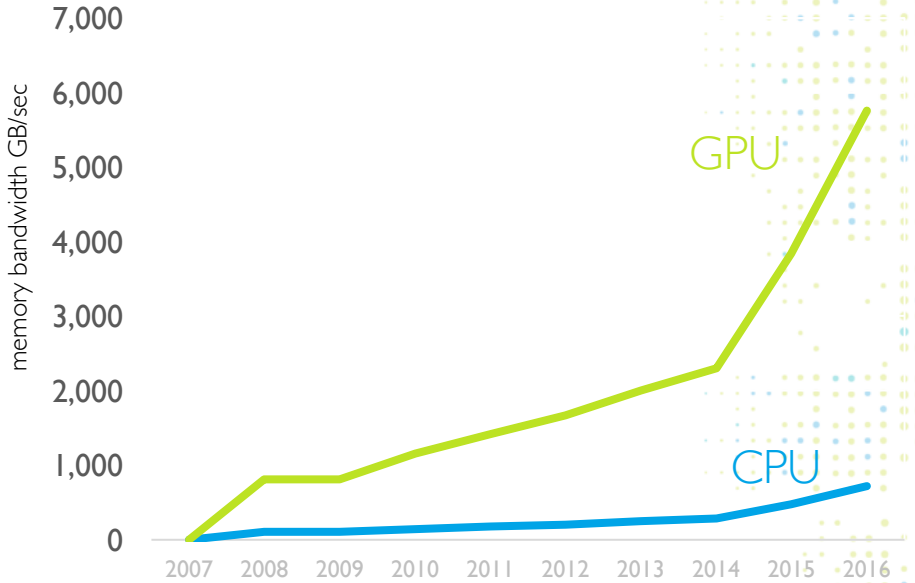
Compute Power

Teraflops



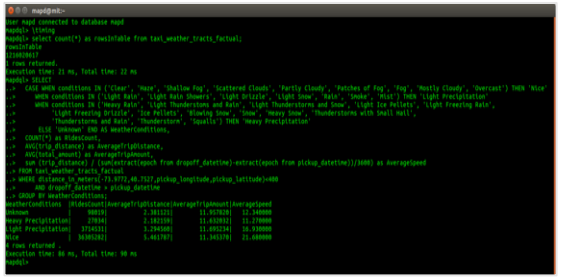
Ability to Read Data

Memory Bandwidth

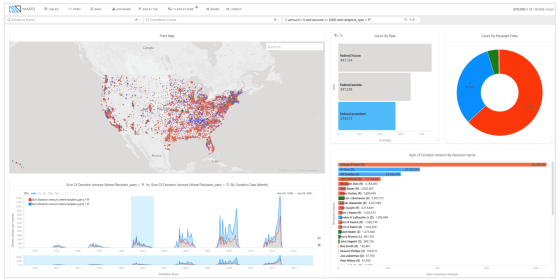


MapD: software optimized for the fastest hardware

100x Faster Queries



Speed of Thought Visualization



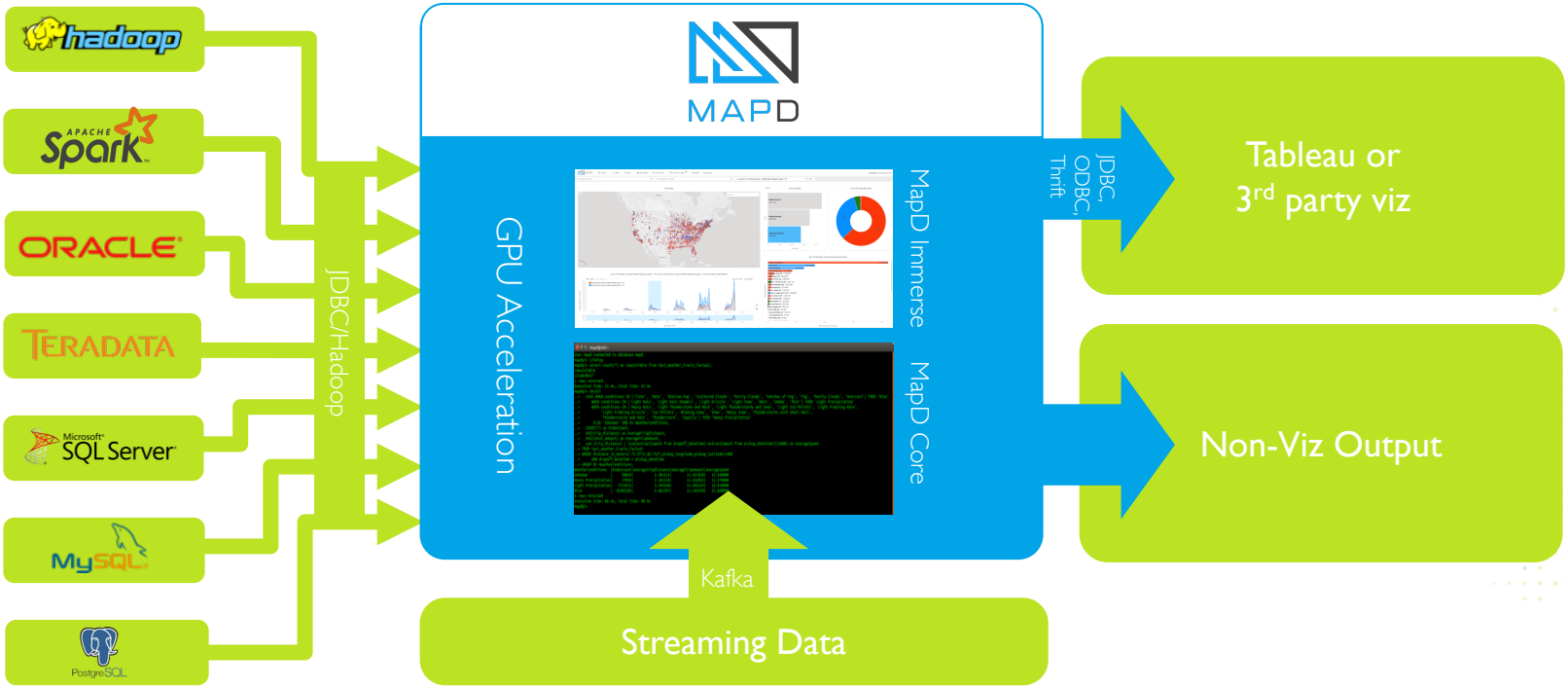
MapD Core

An in-memory, relational, column store database powered by GPUs

MapD Immerse

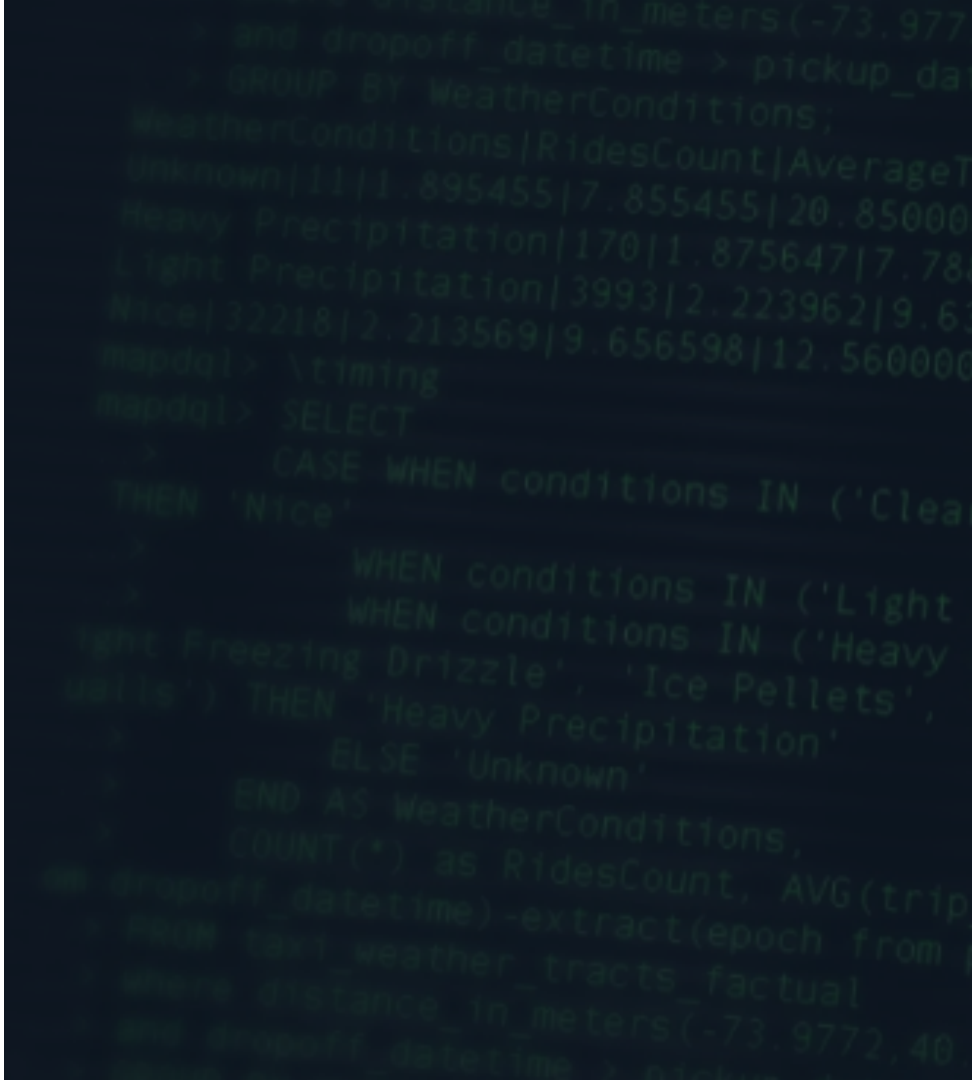
A visual analytics engine that leverages the speed + rendering capabilities of MapD Core

Where MapD sits

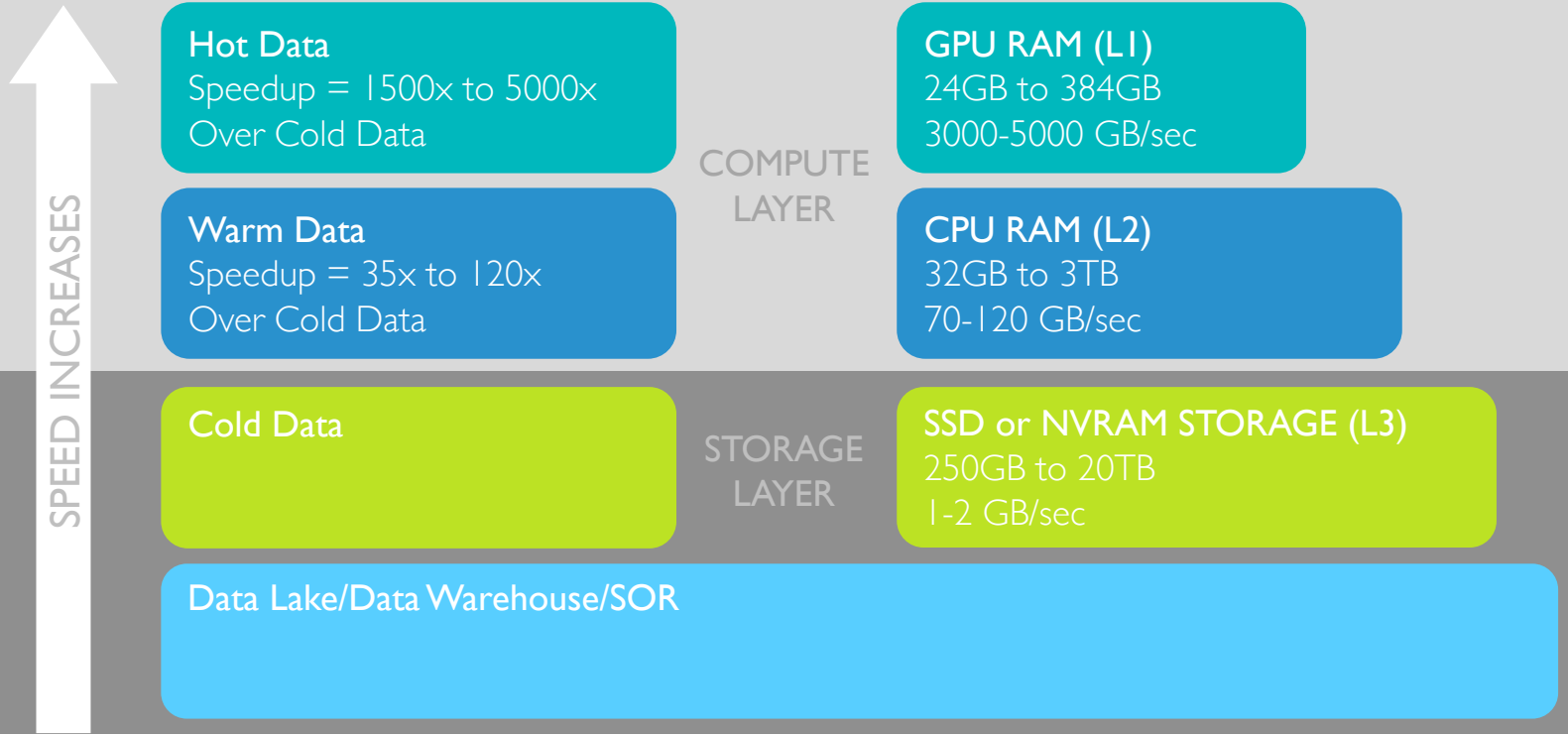


MapD Core

The world's fastest in-memory GPU database powers the world's most immersive data exploration experience



Performance starts with memory management



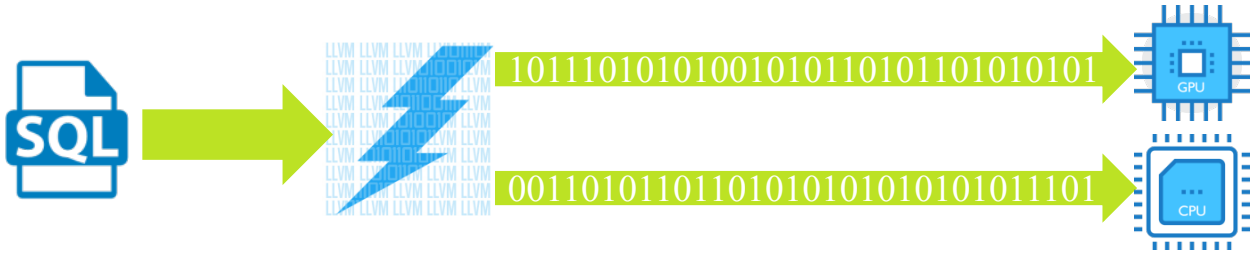
Query Compilation with LLVM

Traditional DBs can be highly inefficient

- each operator in SQL treated as a separate function
- incurs tremendous overhead and prevents vectorization

MapD compiles queries w/LLVM to create one custom function

- Queries run at speeds approaching hand-written functions
- LLVM enables generic targeting of different architectures (GPUs, X86, ARM, etc).
- Code can be generated to run query on CPU and GPU simultaneously



These innovations drive exceptional speed + scale

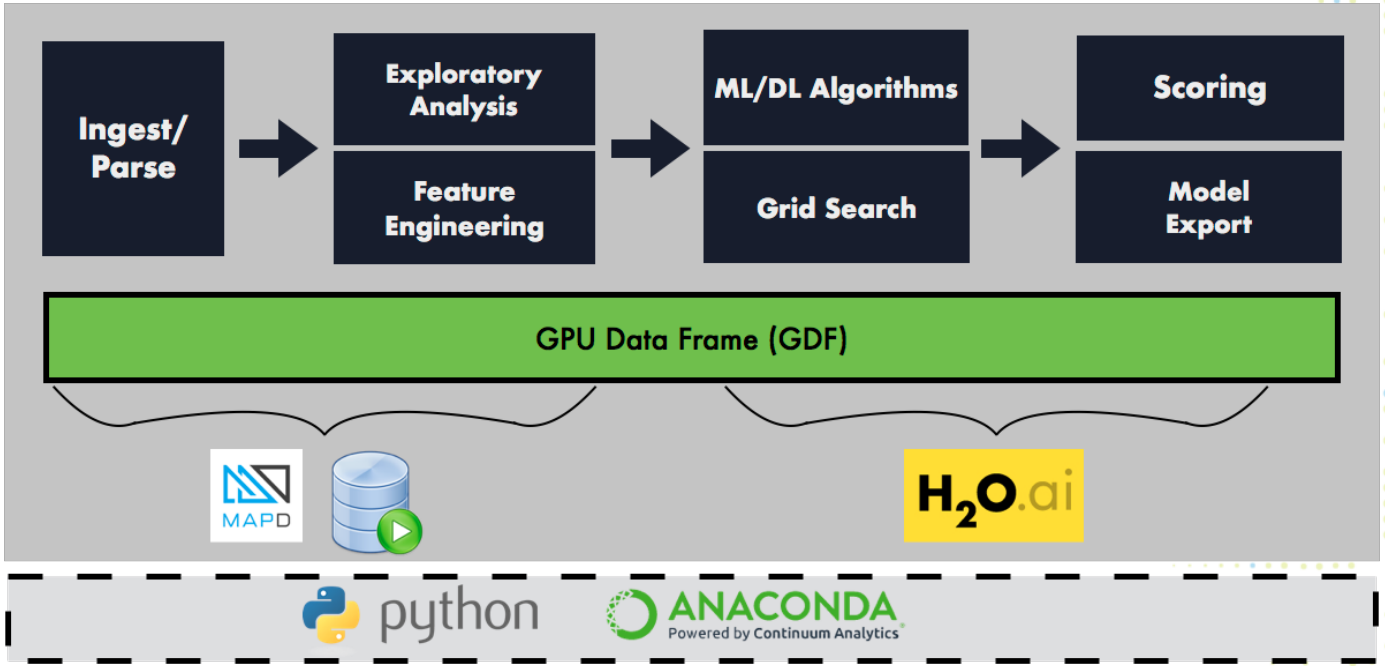
Noted DB blogger, Mark Litwintschik has benchmarked MapD vs. major CPU systems and found it to be between 74x to 3,500x faster than CPU DBs.

The table is sorted by the fastest time query 1 finished in (measured in seconds).

Query 1	Query 2	Query 3	Query 4	Setup
0.021	0.053	0.165	0.51	MapD & 8 Nvidia Pascal Titan Xs
0.027	0.083	0.163	0.891	MapD & 8 Nvidia Tesla K80s
0.028	0.2	0.237	0.578	MapD & 4-node g2.8xlarge cluster
0.036	0.131	0.439	0.964	MapD & 4 Nvidia Titan Xs
0.051	0.146	0.047	0.794	kdb+/q & 4 Intel Xeon Phi 7210 CPUs
1.034	3.058	5.354	12.748	ClickHouse, Intel Core i5 4670K
1.56	1.25	2.25	2.97	Redshift, 6-node ds2.8xlarge cluster
2	2	1	3	BigQuery
4	4	10	21	Presto, 50-node n1-standard-4 cluster
6.41	6.19	6.09	6.63	Amazon Athena
8.1	18.18	n/a	n/a	Elasticsearch (heavily tuned)
10.19	8.134	19.624	85.942	Spark 2.1, 11 x m3.xlarge cluster w/ HDFS
11	10	21	31	Presto, 10-node n1-standard-4 cluster
14.389	32.148	33.448	67.312	Vertica, Intel Core i5 4670K
34.48	63.3	n/a	b/a	Elasticsearch (lightly tuned)
35	39	64	81	Presto, 5-node m3.xlarge cluster w/ HDFS
43	45	27	44	Presto, 50-node m3.xlarge cluster w/ S3
152	175	235	368	PostgreSQL 9.5 & cstore_fdw
264	313	620	961	Spark 1.6, 5-node m3.xlarge cluster w/ S3

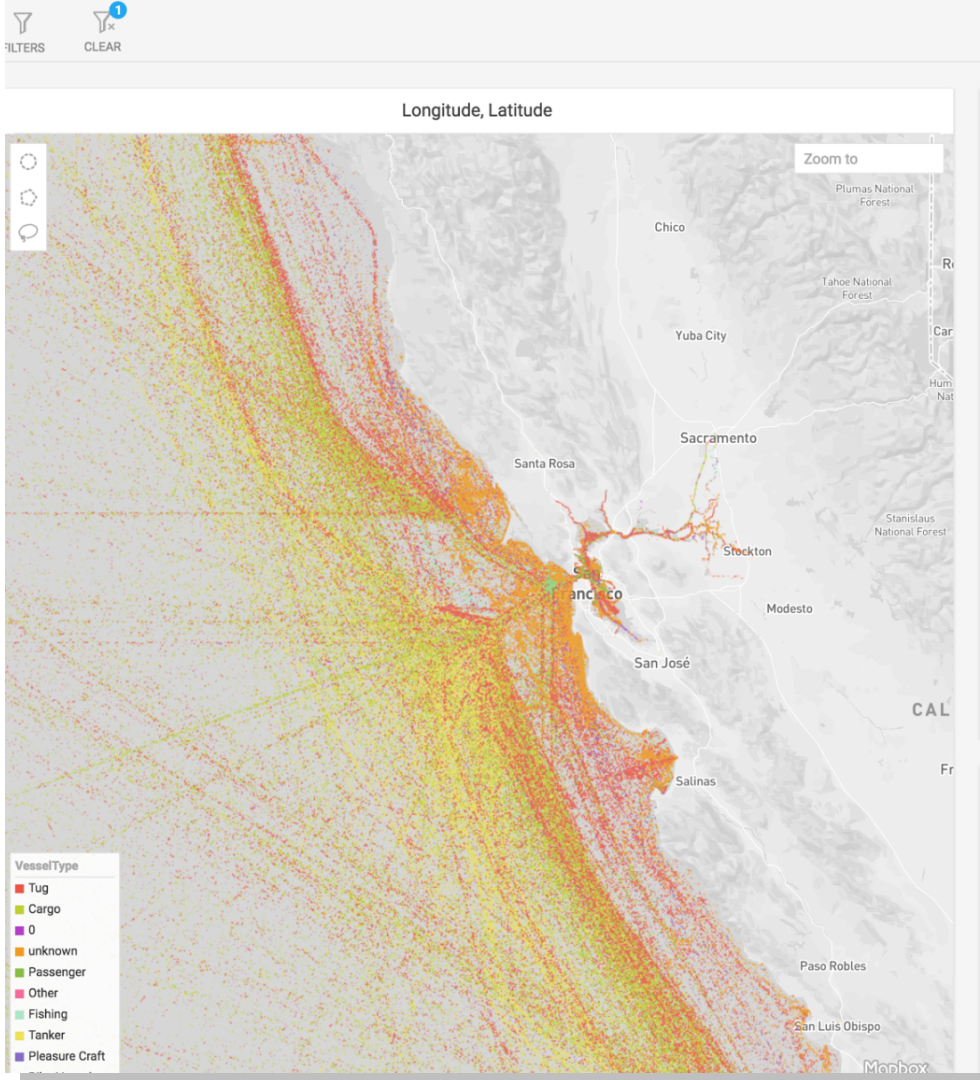
The GPU Open Analytics Initiative (GOAI) and the GPU Data Frame (GDF)

goai
GPU OPEN ANALYTICS INITIATIVE

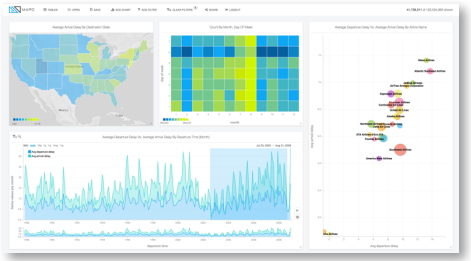


MapD Immerse

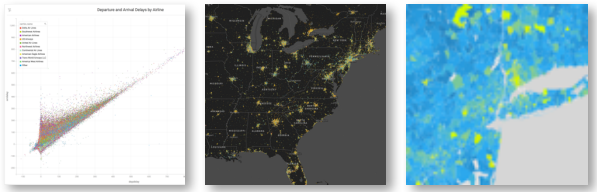
Lightning fast visual analytics for the MapD Core database

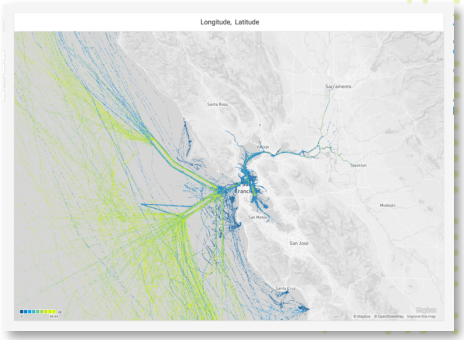


MapD Immerse: our hybrid approach



MapBox React



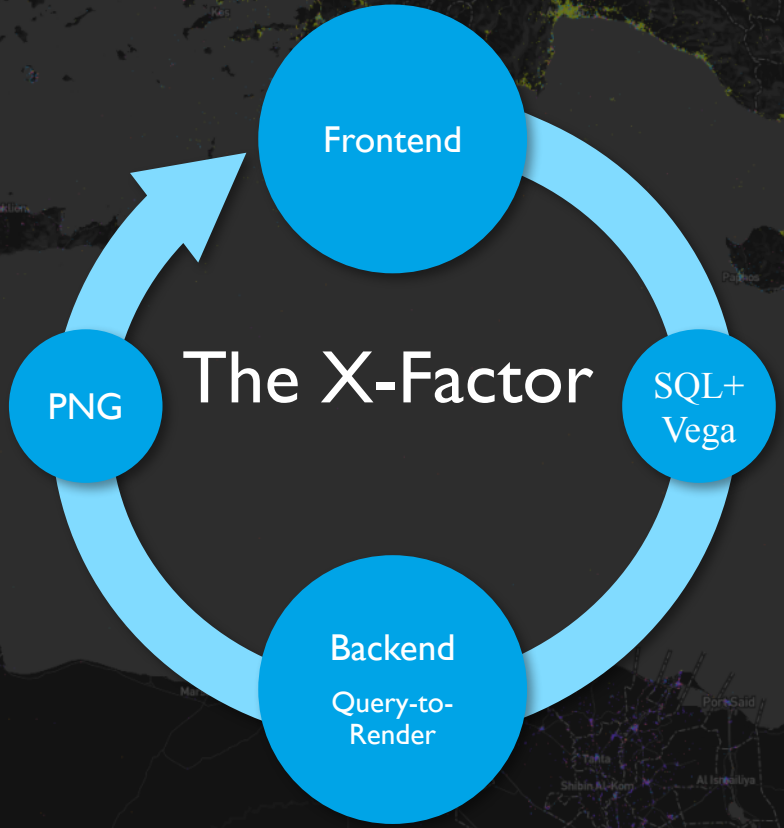


Basic charts are frontend rendered using D3 and other related toolkits

Scatterplots, pointmaps + polygons are backend rendered using the Iris Rendering Engine on GPUs

Geo-Viz is composited over a frontend rendered basemap

Server side rendering



Data goes from compute (CUDA) to graphics (OpenGL)
pipeline without copy and comes back as compressed PNG (~100 KB) rather than raw data (> 1GB)

Vega Spec (a visualization grammar)

- A declarative JSON format for creating visualization designs
- Used to describe backend visualizations
- Defines attributes of render primitives which can be driven by data columns and mapped by scales

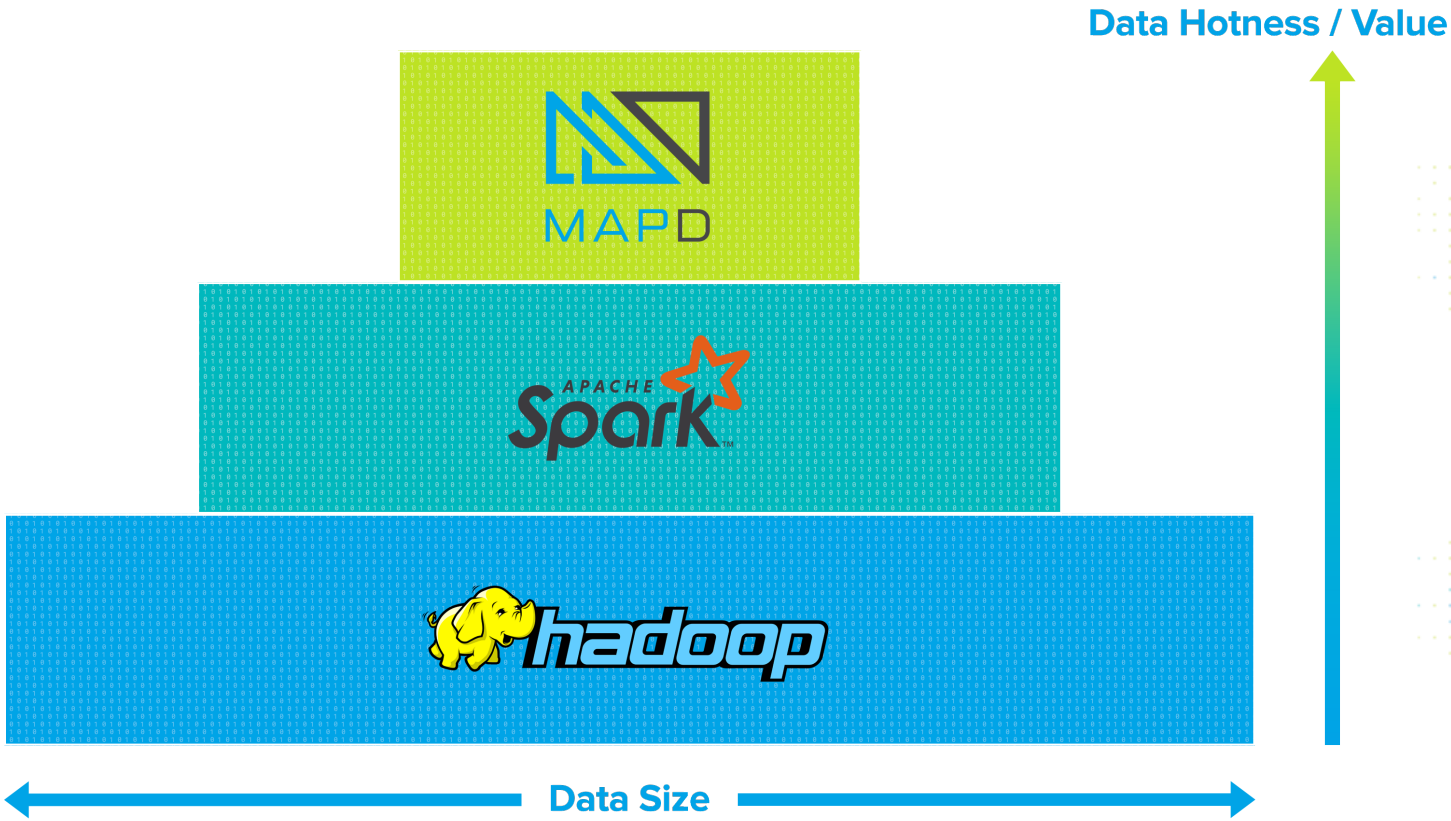
Shader Compilation Framework

- Templatized: supports multiple types (ints, floats, colors, etc), and multiple continuities (discrete, continuous)

Supercharging Spark

Using MapD as a small-footprint, high-power cache on top of Apache Spark

A Modern Analytics Stack



Demo

MapD, Now Open Source

Features	MapD Core Open Source	MapD Community	MapD Enterprise
Core Database	✓	✓	✓
Thrift API	✓	✓	✓
JDBC	✓	✓	✓
Charting API	✓	✓	✓
Immerse Visual Analytics Client	Non-Commercial Binary	Non-Commercial	✓
Rendering Engine	Non-Commercial Binary	Non-Commercial	✓
High Availability	✗	✗	✓
Distributed	✗	✗	✓
LDAP	✗	✗	✓
ODBC	✗	✗	✓
Warranty	✗	✗	✓
Support	Community Support	Community Support	Phone & Email

Delivering significant customer value across industries

verizon

HUAWEI

eog resources

Polling smartphones on demand to assess network health.

Took hours on Oracle previously.

Running complex queries to analyze wireless QOS data.

Previous queries were non-interactive

EOG uses MapD to query and visualize well data

Other BI and GIS systems could not scale to their data sizes

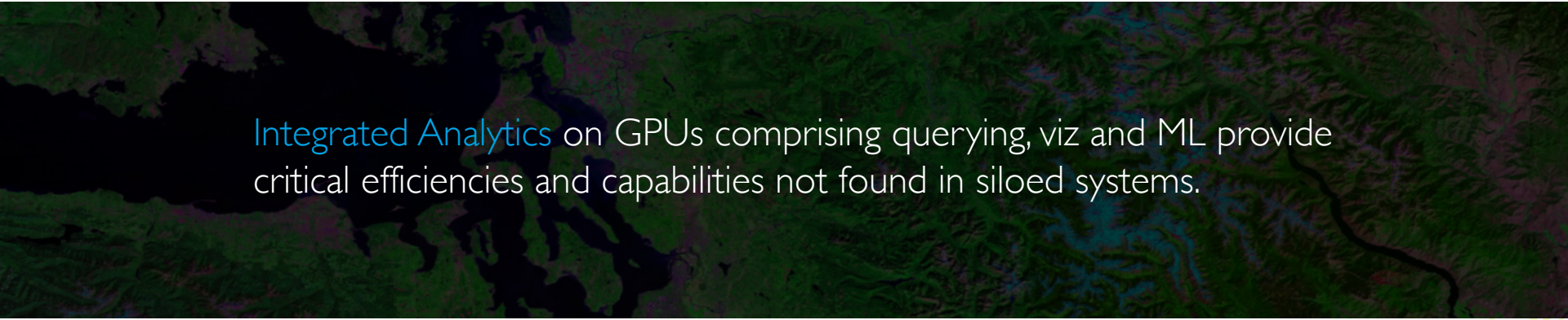
Closing thoughts

We are at an inflection point in compute and GPUs are set to become mainstream in analytics.

Closing thoughts

GPUs allow users to scale up before needing to scale out, making GPU-accelerated analytics an excellent high-performance cache on a larger data lake or Spark deployment.

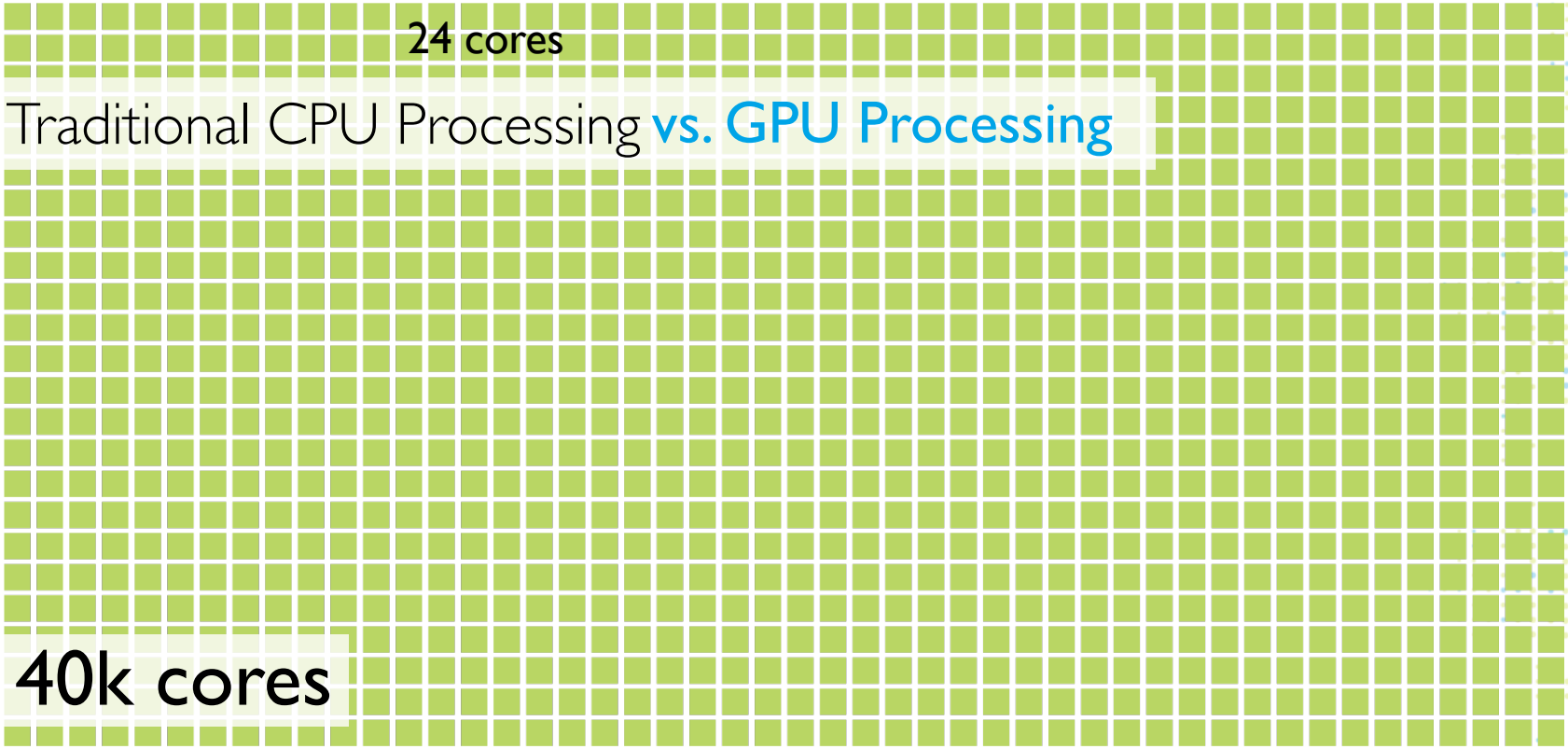
Closing thoughts



Integrated Analytics on GPUs comprising querying, viz and ML provide critical efficiencies and capabilities not found in siloed systems.

MapD Technologies, Inc.
One Front Street
Suite 2650
San Francisco, CA 94111
mapd.com

Cores and memory define the difference



(actually it would take a bit more than 24 of these slides to show 40k cores)