The Need for Speed: How the Auto Industry Accelerates Machine Learning with Visual Analytics

Zach Izham, VW | Aaron Williams, MapD

March 27, 2018

Introductions

Asghar Ghorbani
Data Scientist

@ghorbani_asghar
asghar.ghorbani@volkswagen.de
/in/aghorbani/
/a-ghorbani

Zach Izham
Legend

@drizham
zach.izham@volkswagen.de
/in/dr-zach-izham-02090b5/
/drizham

Aaron Williams
VP of Global Community

@_arw_
aaron@mapd.com
/in/aaronwilliams/
/williamsaaron

slides: https://speakerdeck.com/mapd/

Volkswagen MAPD

Agenda

A Real World Problem: Churn

- Partial Dependency Analysis - An Accelerated Review
- A Complete Machine Learning Pipeline
- Demo: Data Engineering + Training + Predictive Analytics + Black Box Interrogation

The GPU Data Frame in Action

- GO.ai and MapD

Q&A

"Every business ~~will~~ HAS become a software business, build applications, use advanced analytics and provide SaaS services."

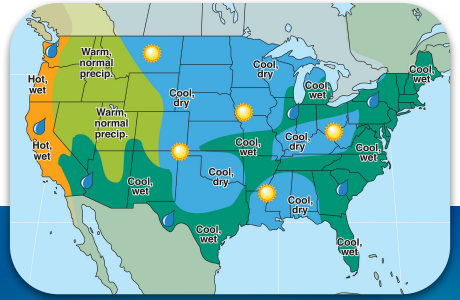
- Smart CEO Guy

Volkswagen MAPD

The Evolution of Competitive Data

Collect It

Make it Predictive



Make It Actionable

Partial Dependency Analysis

An Accelerated Review

Example: Partial Dependency

Assume the following example:

Failure rate of machine component:

Only depends on

hours of work (HoW) of components, *h*

Not (within reason)

Age of components

Assume failure rate, f(x) is only dependent on hours of work and **not** age:

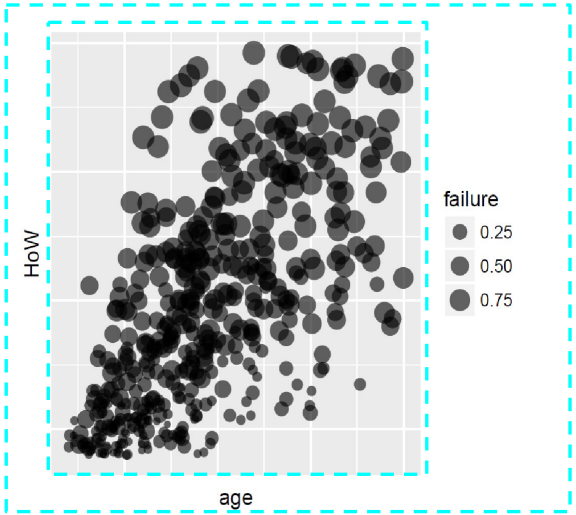
$$f(h) = \frac{1}{1 + \left(\frac{h}{\alpha}\right)^{-\beta}} + noise$$

Where α and β are constants dependent on machine operating conditions

L. Greene et al. Simpson's paradox: A cautionary tale in advanced analytics. *Significance*, 2012.

Example Partial Dependency

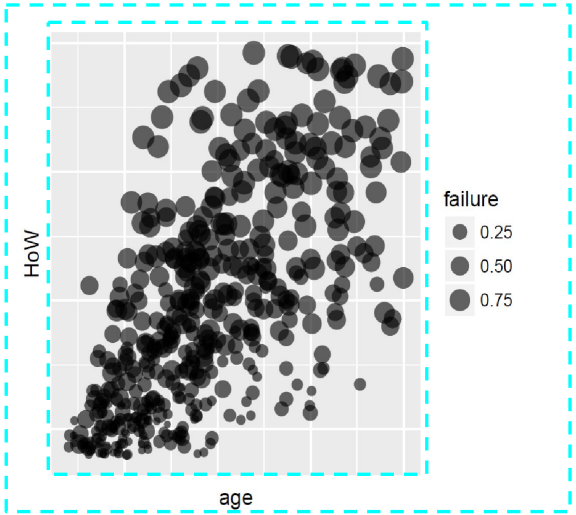
Simpson's Paradox

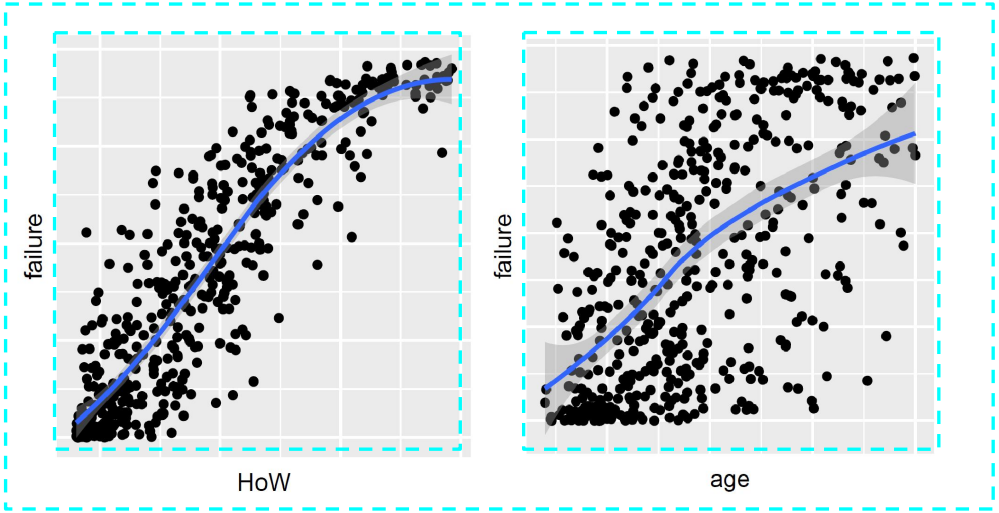


$$f(h) = \frac{1}{1 + \left(\frac{h}{\alpha}\right)^{-\beta}} + noise$$

Example Partial Dependency

Impact of Each Variable on Target Value

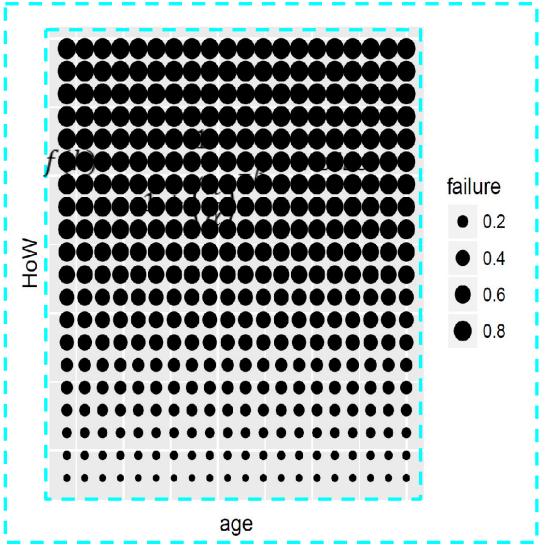




$$f_s(X_s) = E_{X_c}\left(f(X_s, X_c) \mid X_s\right)$$

Generating Data for the Complete State Space

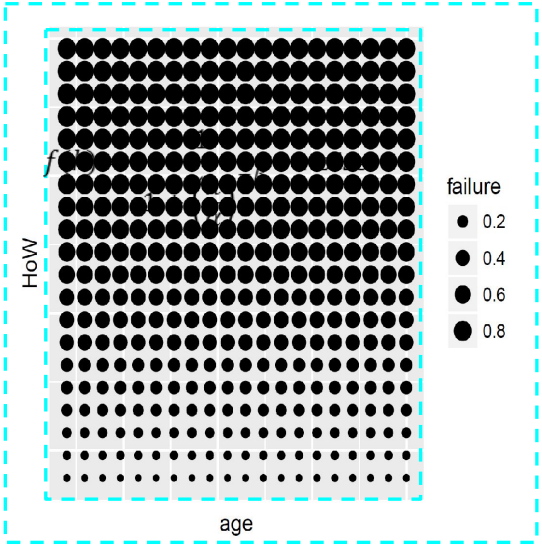
The Failure Rate Generated from the Trained Black Box Model

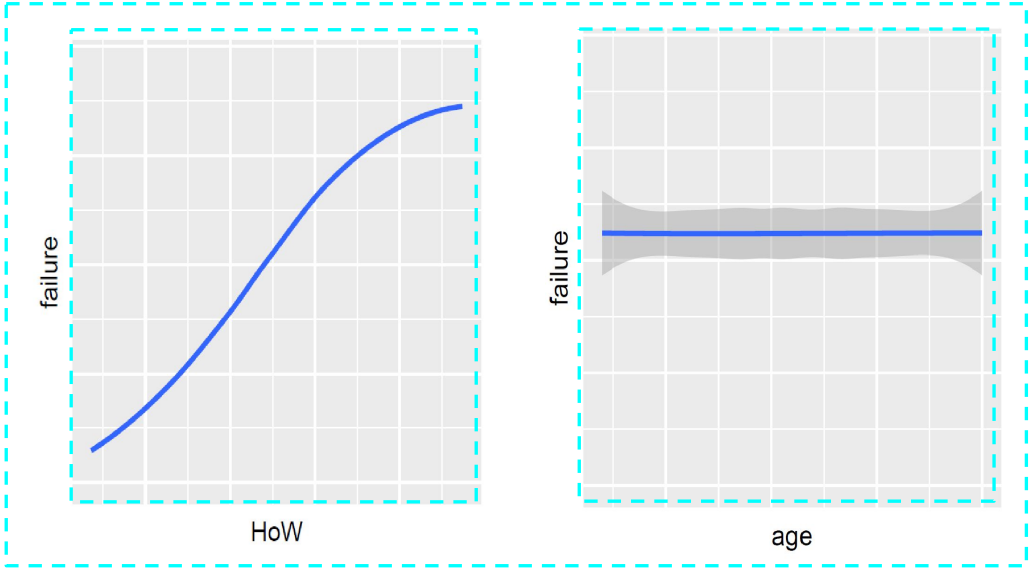


T. Hastie, et al. The Elements of Statistical Learning., 2001

Investigating the System with the Simulated Data

Partial Dependency Analysis



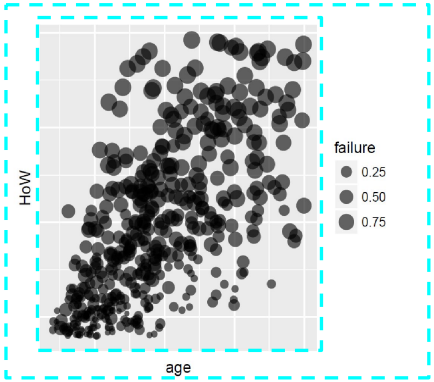


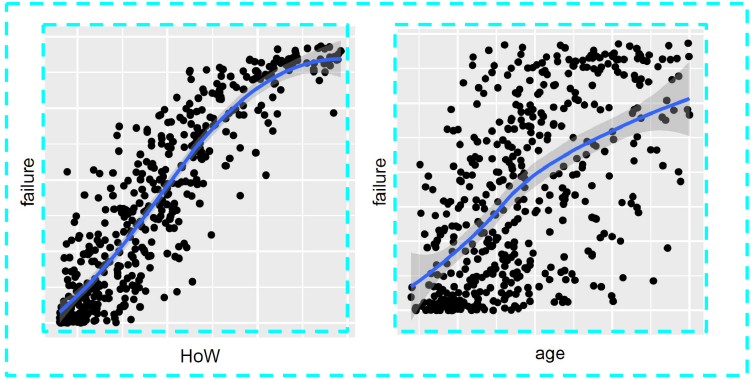
Impact of each variable on target value $f_s(X_s) = E_{X_c} f(X_s, X_c)$

T. Hastie, et al. The Elements of Statistical Learning., 2001

Collect Data to Build a Model

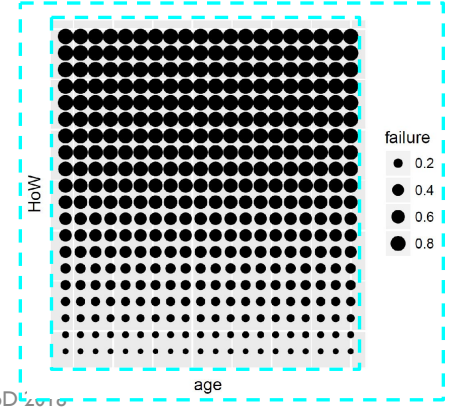
Generate Data for the Whole State Space

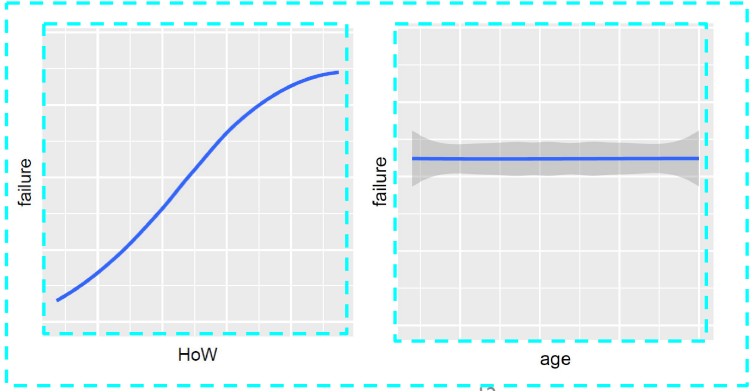




$f_s(X_s) = E_{X_c}\,(f\,(X_s, X_c) | X_s)$

$$f(h) = \frac{1}{1 + \left(\frac{h}{\alpha}\right)^{-\beta}} + noise$$





$f_s(X_s) = E_{X_c}\, f\,(X_s, X_c)$

Volkswagen MAPD

Why do We Need GPUs?

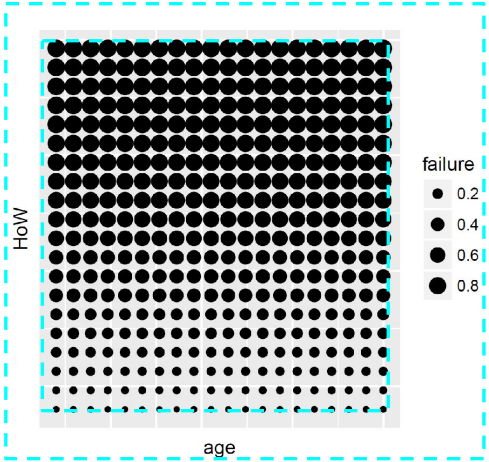
Data Size Explosion

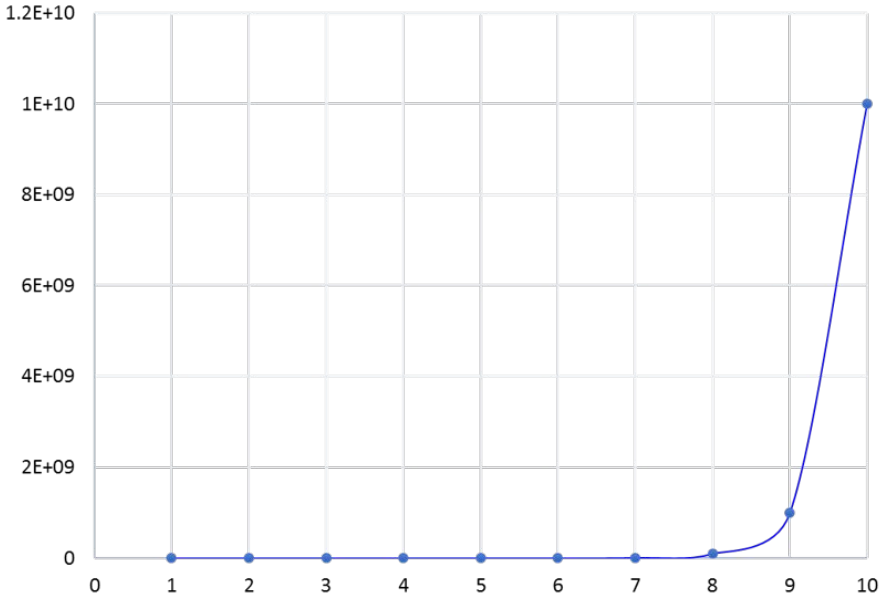
Coarse Grid – Small Data

Dense Grid / Data dimensionality – Large Data

Grid resolution 10:

- 1 variable: 10
- 2 variables: 10 x 10 = 100
- ...
- 10 variables: 10^ 10 = 10,000,000,000





Analysis Logistics

Data Engineering

Some Background - Objective - Creating the Master Data Frame

Relevant data reside on separate tables and databases
Collecting, cleaning and curating the relevant data

Creating a target variable:

Which cars will not be returning to the garage / service center for service

VW Data Pipeline

Getting Data to the Environment

1. Loading data to MapD database:
   a. Table extracted from database and exported as csv.
   b. mapdql used to create table and import in data.

2. Exploratory Data Analysis:
   a. MapD dashboard used to perform exploratory data analysis (EDA), gives: ‘spatial awareness’ of the of data.
   b. Using MapD allows this on the fly investigation on large datasets.

Demo Data Stats

Rough Feeling of Data

Number of rows ~2.2 million
25 relevant columns
2 categorical columns
23 numerical columns

Hardware Stack

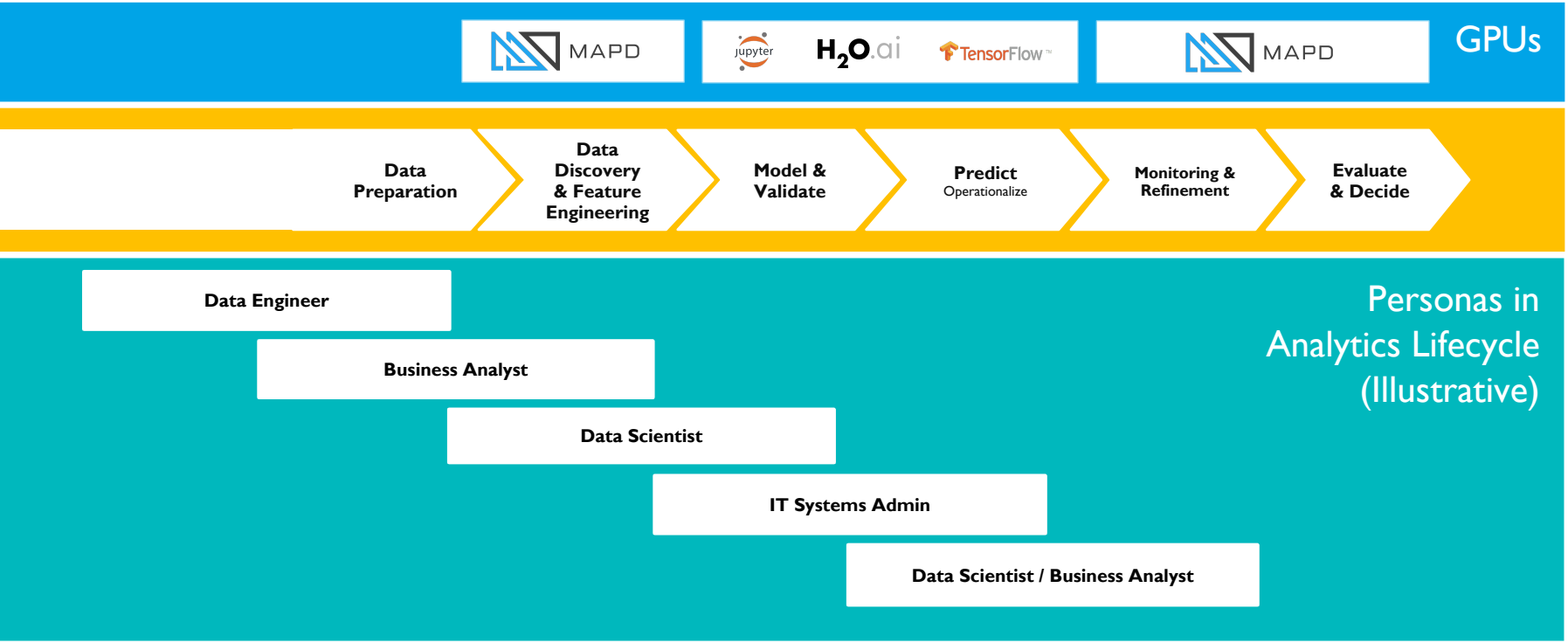
Details of Hardware Setup On Microsoft Azure Instance:

Instance Name: NC24S_V2 Standard

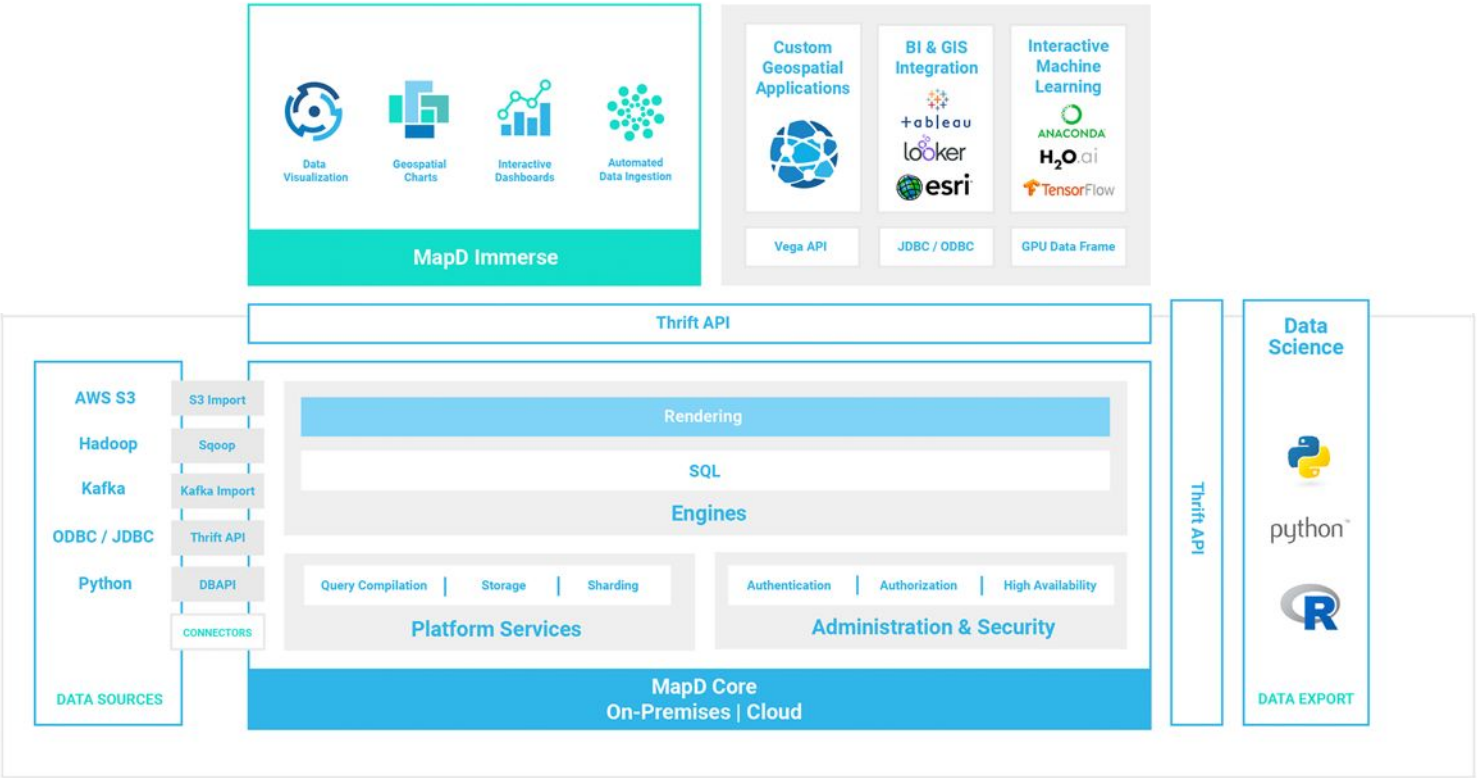
vCPUs (Cores)	24
Storage	448GB
Data Disks	32
GPUs	4 Tesla P100-PCIE-16GB
OS	CentOS

Demo Time!

Machine Learning Pipeline



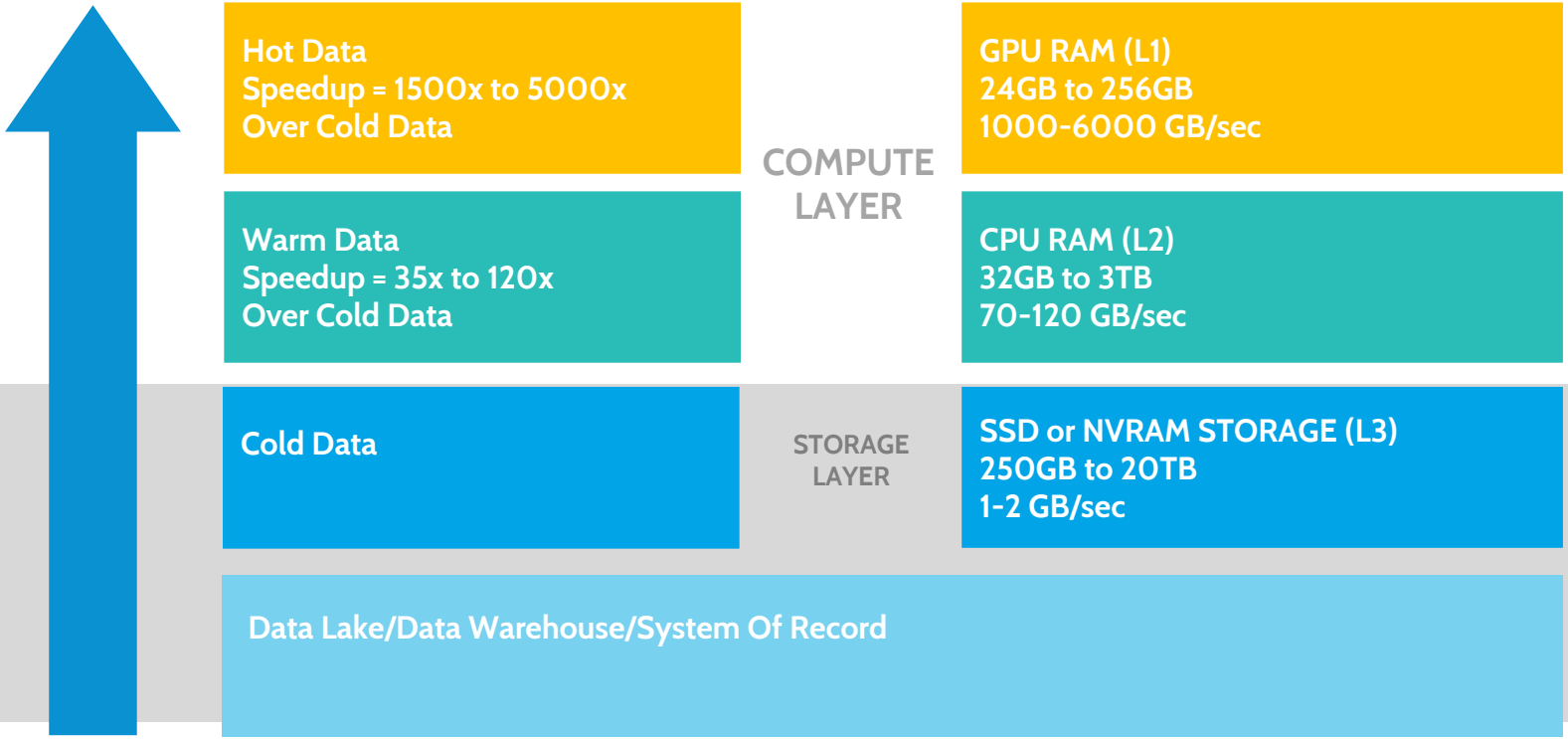
MapD is the analytics platform created for GPUs



Volkswagen MAPD

Advanced memory management

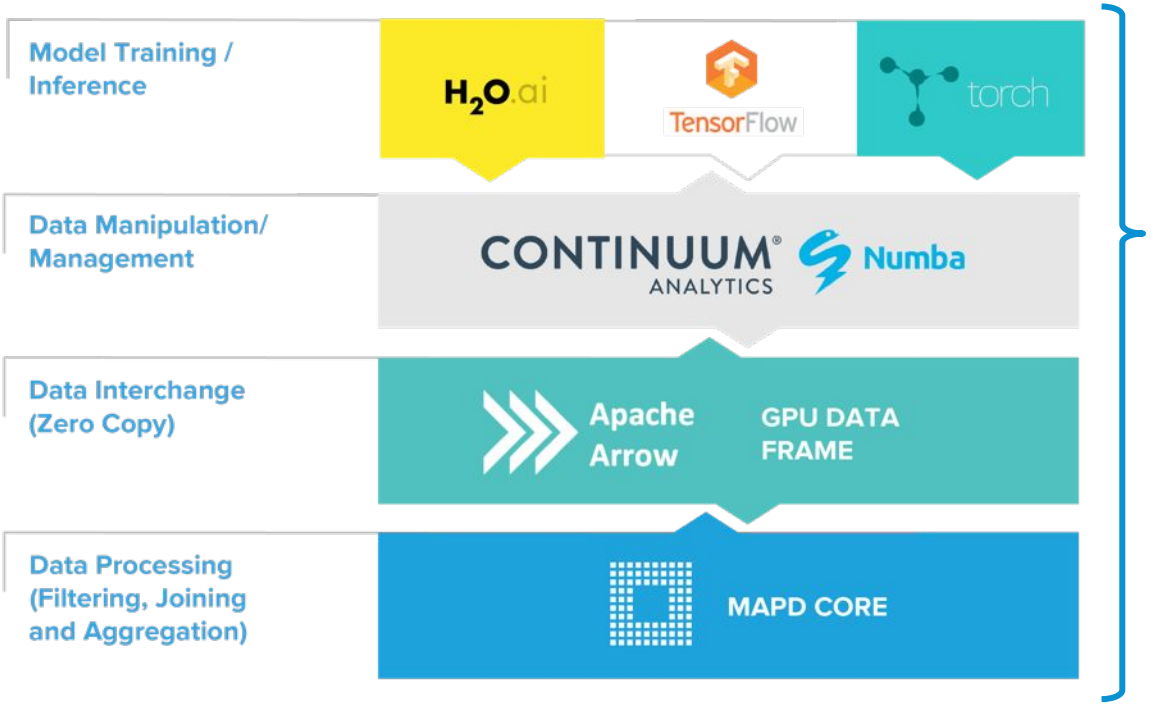
Three-tier caching to GPU RAM for speed and to SSDs for persistent storage



Volkswagen MAPD

The GPU Open Analytics Initiative (GOAI)

Creating common data frameworks to accelerate data science on GPUs



jupyter

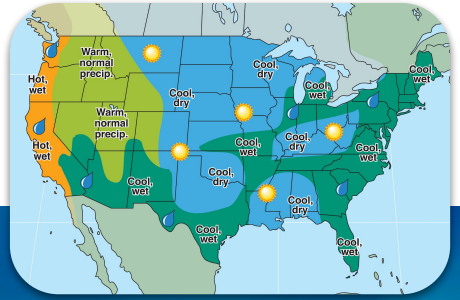
/mapd/pymapd
/gpuopenanalytics/pygdf

Volkswagen MAPD

The Time Is Now

Collect It

Make it Predictive



Make It Actionable

ML Examples

- We’ve published a few notebooks showing how to connect to a MapD database and use an ML algorithm to make predictions

/gpuopenanalytics/demo-docker

- We will also be publishing the notebook from this VW churn example

/mapd

README.md

mapd-ml-demo

Can be run with nvidia-docker-compose. This depends on two containers:

Name	Use	Dockerfile location
`mapd-core`	MapD Database	Defaults to Community Edition on Docker Hub `mapd/mapd-ce-cuda`
`mapd/ml`	Demo notebooks	Dockerfile in top-level of `mapd-ml-demo` repo

Build

mapd-ml-demo Server

To build the `mapd-ml-demo` container, clone the repo and `cd` into it.

To build the container, run:

```
docker build -t mapd/ml .
```

Exporting

If you need to move the containers to a new machine, run:

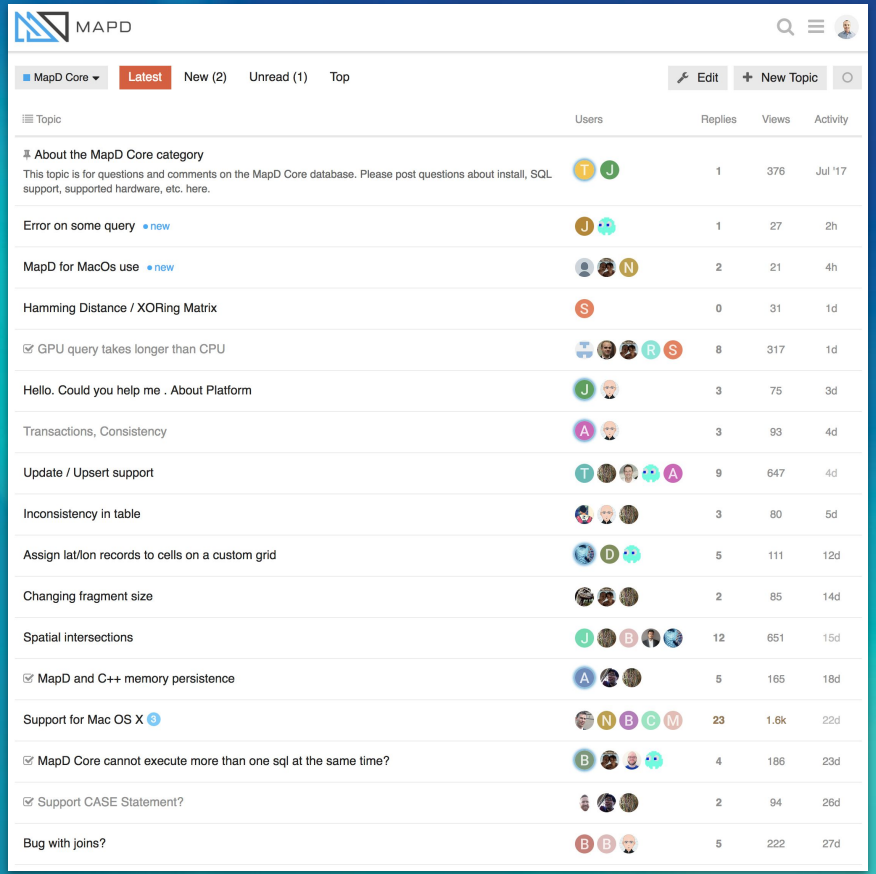
```
docker save -o mapd-ce-cuda.tar mapd/mapd-ce-cuda
docker save -o mapd-ml.tar mapd/ml

gzip mapd-ce-cuda.tar
gzip mapd-iml.tar
```

You will then have files which can be moved to the new machine: `mapd-ce-cuda.tar.gz`, `mapd-ml.tar.gz`. You will also want to grab the `docker-compose.yml` file (but probably not the `nvidia-docker-compose.yml` one).

Run

Next Steps



- **community.mapd.com**
  Ask questions and share your experiences
- **mapd.com/demos**
  Play with our demos
- **mapd.com/platform/download-community/**
  Get our free Community Edition and start playing

Volkswagen MAPD

Thanks to the whole team!

- Asghar Ghorbani
- Wamsi Viswanath
- Abraham Duplaa

Questions?

Asghar Ghorbani
Data Scientist

@ghorbani_asghar
asghar.ghorbani@volkswagen.de
/in/aghorbani/
/a-ghorbani

Zach Izham
Legend

@drizham
zach.izham@volkswagen.de
/in/dr-zach-izham-02090b5/
/drizham

Aaron Williams
VP of Global Community

@_arw_
aaron@mapd.com
/in/aaronwilliams/
/williamsaaron

slides: https://speakerdeck.com/mapd/

Volkswagen MAPD