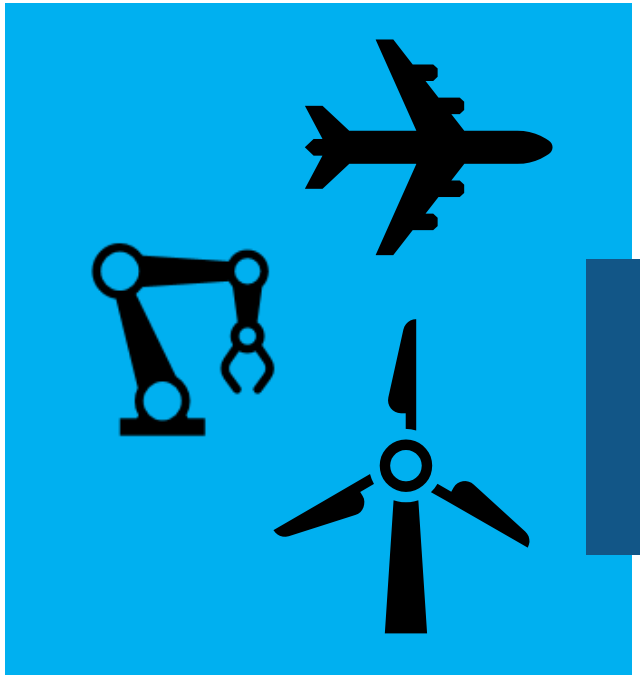
MATLAB EXPO 2018

Scaling up MATLAB Analytics with Kafka and Cloud Services

Pallavi Kar

The Need for Large-Scale Streaming



Jet engine: ~800TB per day
Turbine: ~ 2 TB per day

Predictive Maintenance

Increase Operational Efficiency

Reduce Unplanned Downtime

Many applications require near real-time analytics

Medical Devices

Patient Safety

Better Treatment Outcomes

Connected Cars

Safety, Maintenance

Advanced Driving Features

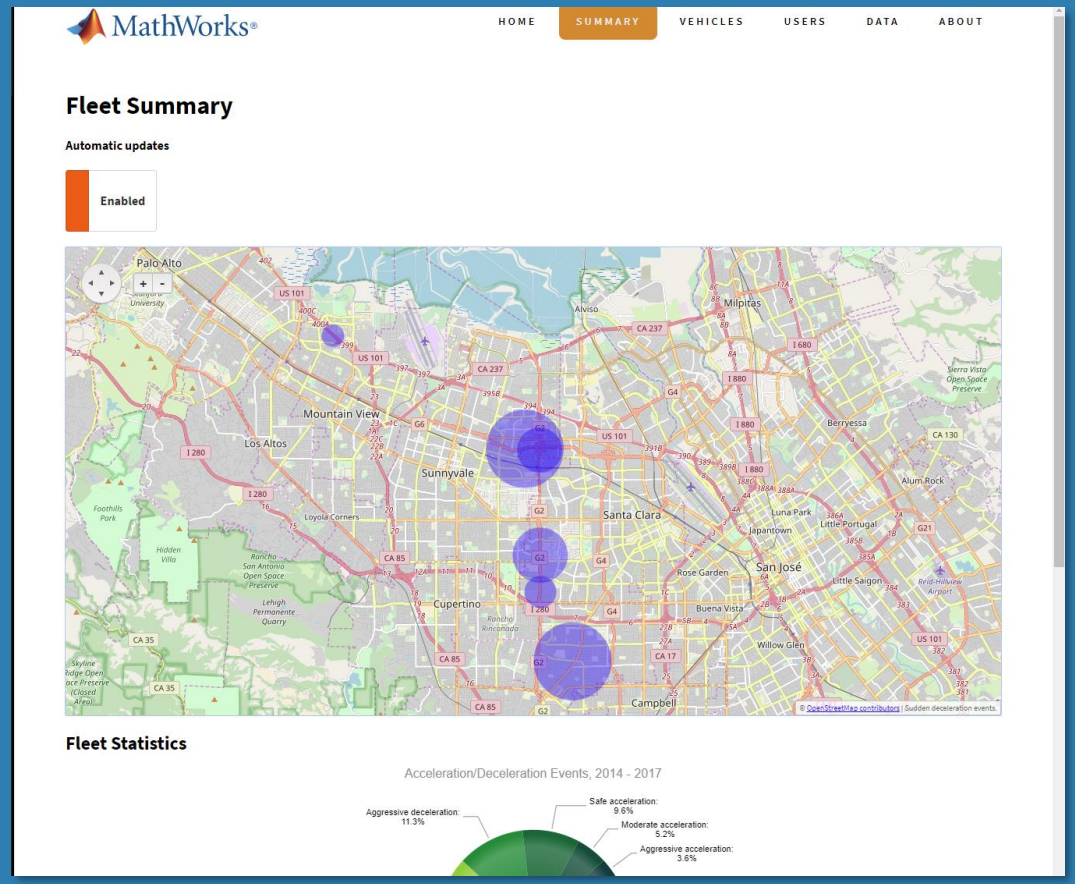
Car: ~25 GB per hour

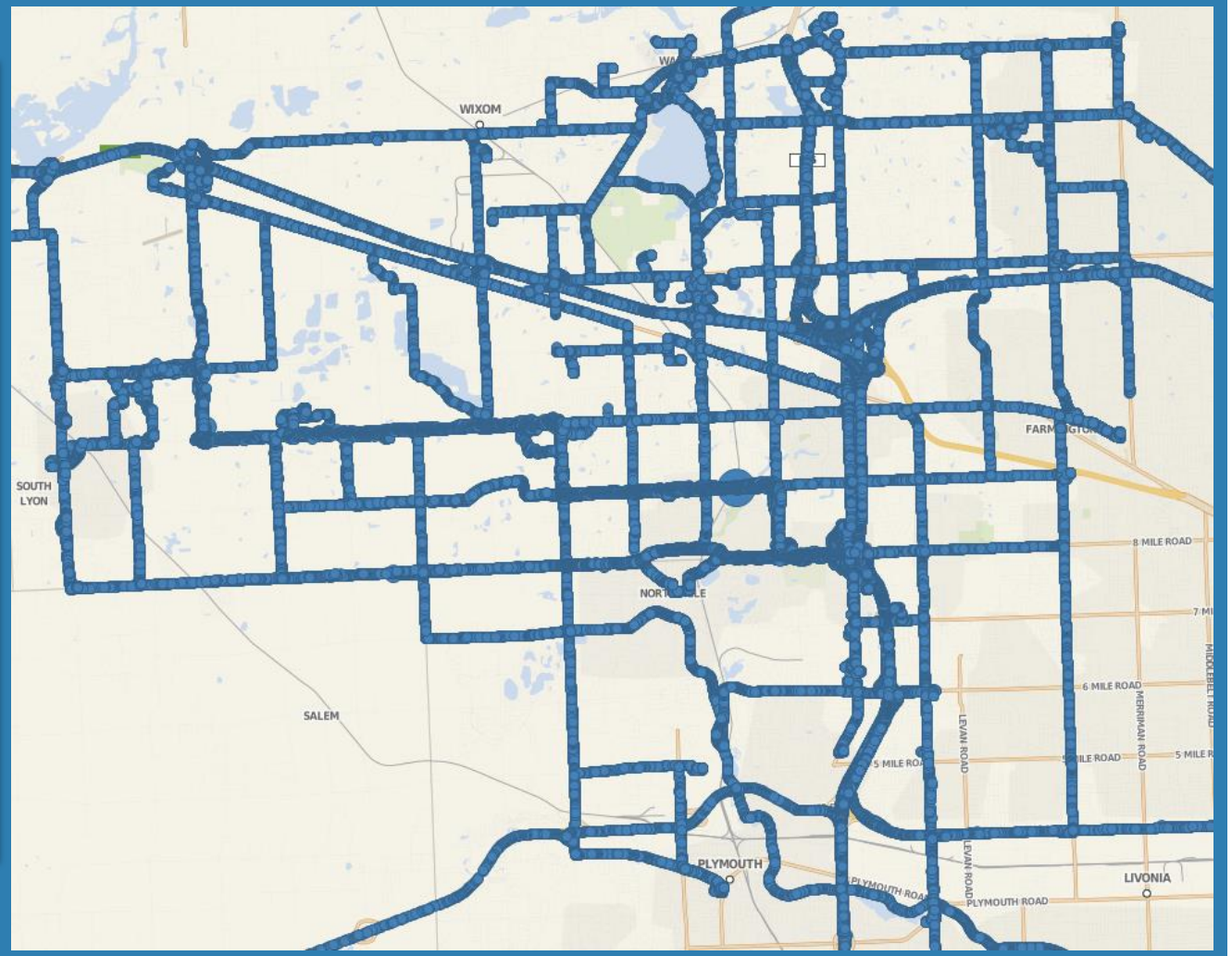
Example Problem – How's my driving?

Analyzing vehicle data to score driving habits

- A group of MathWorks employees installed an OBD dongle in their car that monitors the on-board systems
- Data is streamed to the cloud where it is aggregated and stored

Case Study: Stream based Analytics on drive data with MATLAB





Challenges in building such a system…

Edge Devices

- Aggregate data from multiple devices
- Build accurate algorithms
- Integrate algorithms into production systems
- Run analytics on multiple streams of data

Analytics Development

MATLAB

Algorithm Developers

Business Decisions

Business Systems

Power BI

Qlik

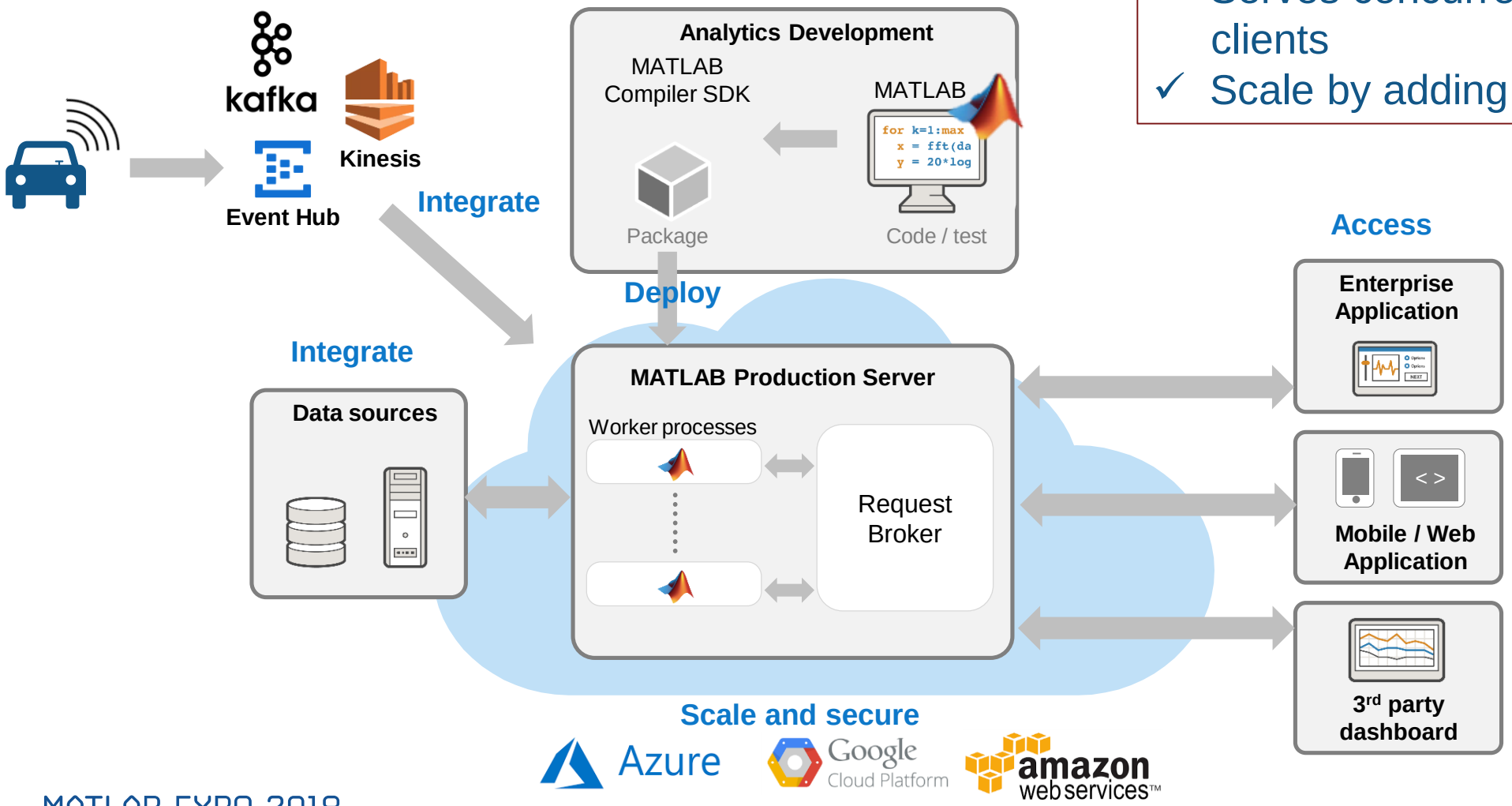
Spotfire

tableau

End Users

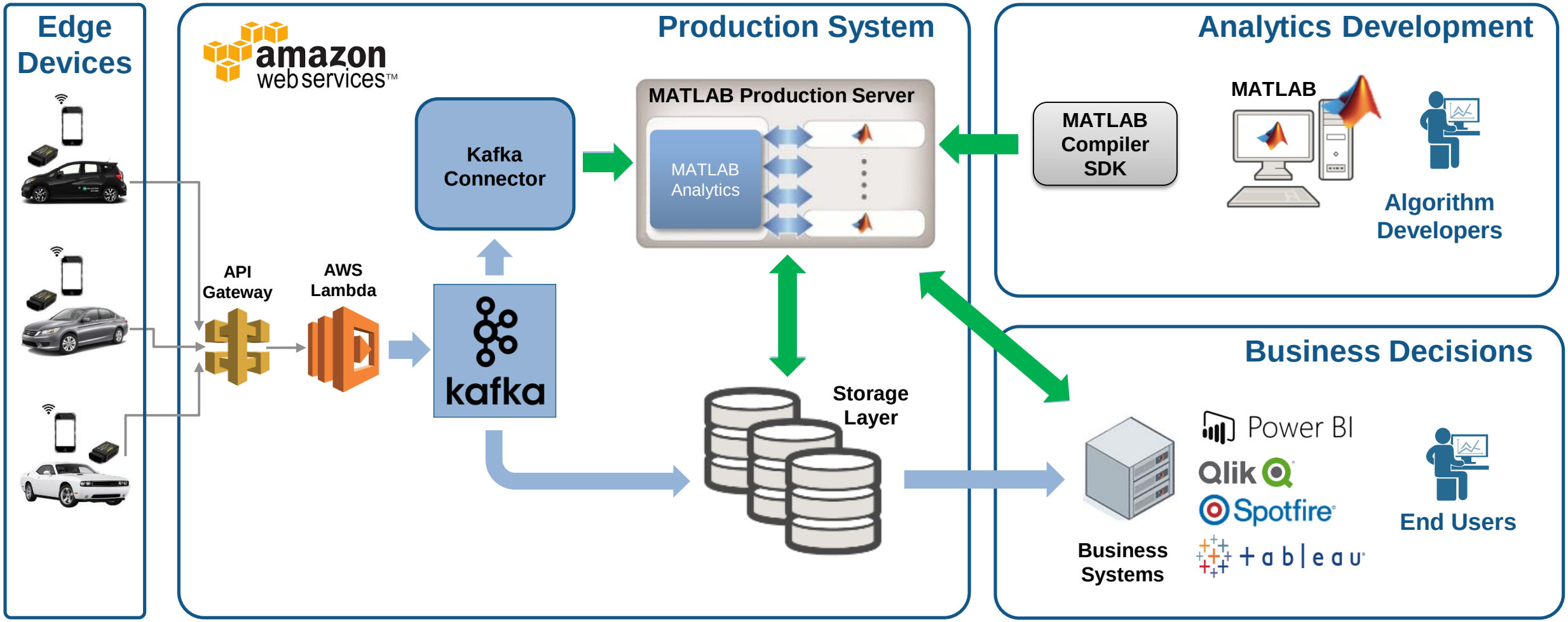
Solution : MATLAB Production Server and Streaming engine

- ✓ MATLAB Analytics to REST APIs
- ✓ Serves concurrent requests from web clients
- ✓ Scale by adding workers



Fleet Analytics Architecture

Connectors provided by MathWorks



Development to Deployment Workflow

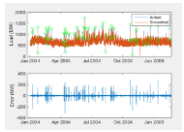
1. Access and Explore Data
2. Preprocess Data
3. Develop Predictive Models
4. Integrate with Production Systems
5. Visualize Results

Files

Databases

Sensors

Working with Messy Data



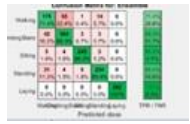
Data Reduction/ Transformation

Feature Extraction

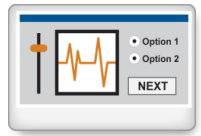
Model Creation e.g. Machine Learning

Parameter Optimization

Model Validation



Desktop Apps



Enterprise Scale Systems

MATLAB Excel .NET C/C++ .exe Java .dll

Embedded Devices and Hardware

3rd party dashboards

Spotfire® TIBCO Software

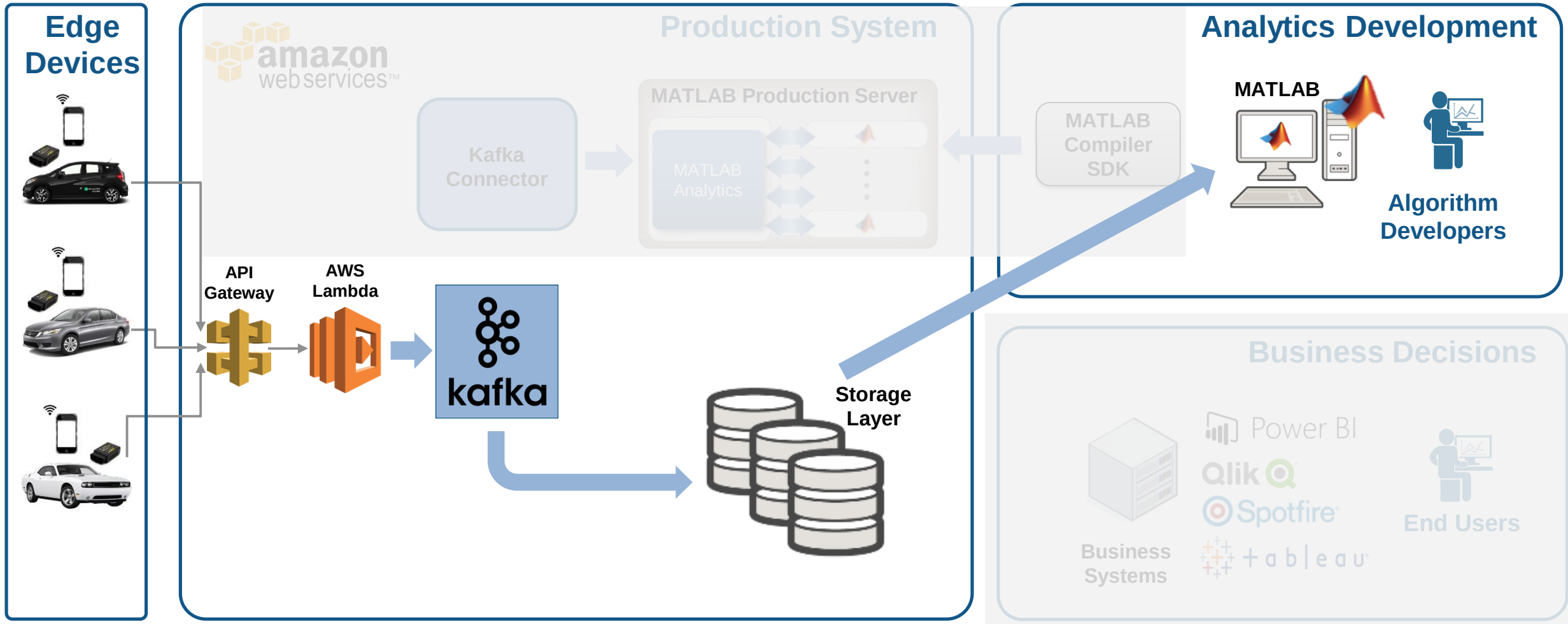
+ableau

Qlik

Power BI

Web apps

Accessing data in MATLAB



The Data: Timestamped messages with JSON encoding

```
{
    "vehicles_id": {"$oid":"55a3fd0069702d5b41000000"},
    "time" : {"$date":"2015-07-13T18:01:35.000Z"},
    "kc" : 1975.0, "kff1225" : 100.65293, "kff125a" : 110.36619, …
}
```

Key

Timestamp

Values

```
{
    "vehicles_id": {"$oid":"55a3fe3569702d5c5c000020"}
    "time":{"$date":"2015-07-13T18:01:53.000Z"},
    "kc" : 2000.0, "kff1225" : 109.65293, "kff125a" : 115.36619,
      …
}
```

```
{
    "vehicles_id": {"$oid":"55a4193569702d115b000001"}
    "time":{"$date":"2015-07-12T19:04:04.000Z"}
    "kc":2200.0, "kff1225" : 112.65293, "kff125a" : 112.36619,
      …
}
```

Access a Sample of Data

Raw Data

	timestamp	1 value	2 key
1	15-Jan-2015 22:12:23	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee0" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'
2	15-Jan-2015 22:12:24	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee1" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'
3	15-Jan-2015 22:12:25	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee2" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'
4	15-Jan-2015 22:12:26	'{ "_id" : { "$oid" : "55a41cb069702d115b059ee3" }, "trip_id" : { "$oid"...	'55a41cb069702d115b059ede'

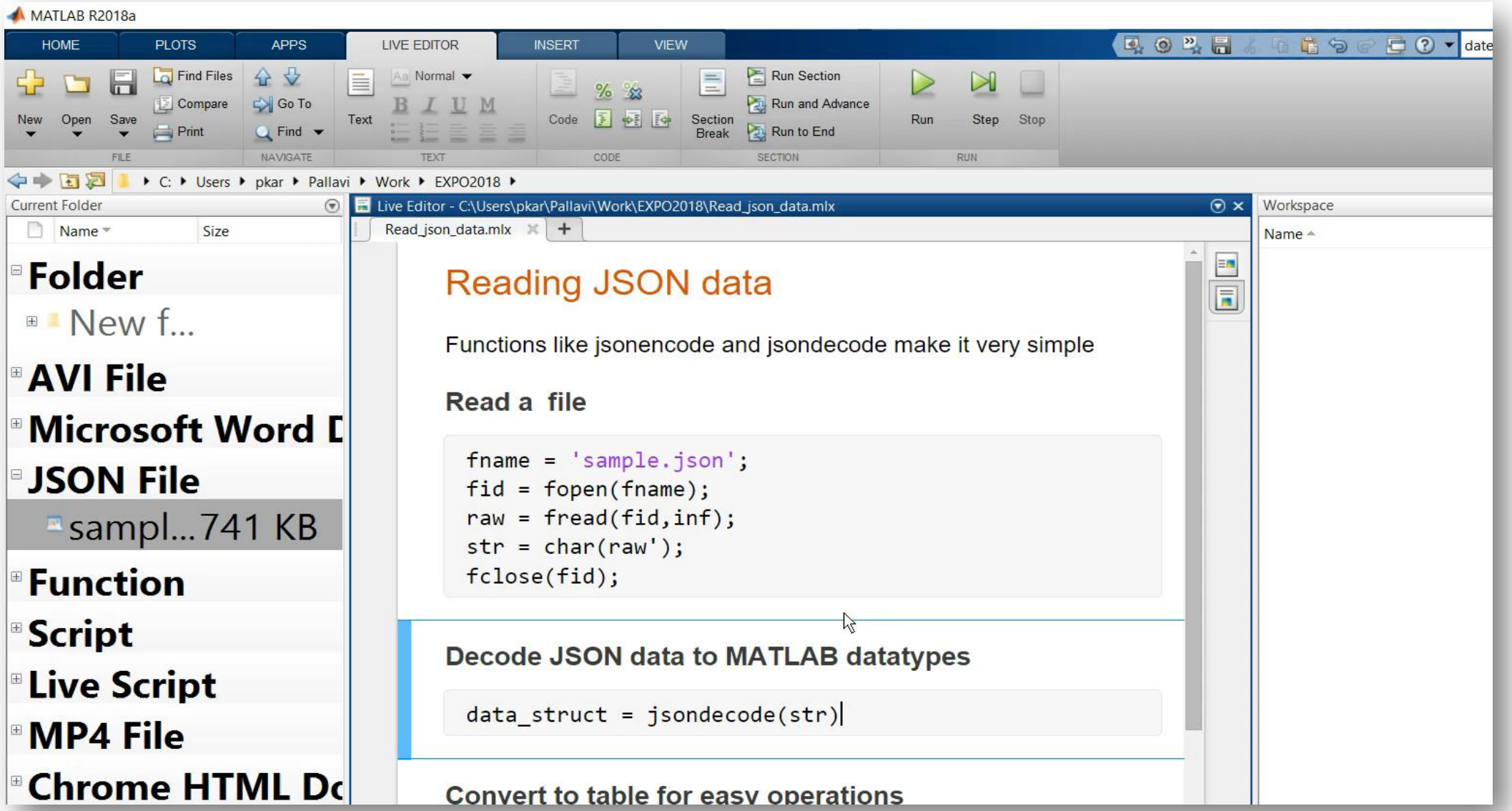
- ✓ Decode JSON data
- ✓ Create Timetable

Timetable

t = 4647×40 timetable

	trip_id	VIN	kff1001	kff1005	kff1006	kff1220	kff1221	kff1222	kff1223	kff125a
1 Sun Jul 12 16:18:41 UTC 2015	55a3fe356...	55a3fe356...	17.1000	-84.9323	45.4704	NaN	NaN	NaN	NaN	59.0434
2 Sun Jul 12 16:18:42 UTC 2015	55a3fe356...	55a3fe356...	17.1000	-84.9322	45.4704	NaN	NaN	NaN	NaN	57.8609
3 Sun Jul 12 16:18:43 UTC 2015	55a3fe356...	55a3fe356...	18.9000	-84.9322	45.4705	NaN	NaN	NaN	NaN	52.7147
4 Sun Jul 12 16:18:44 UTC 2015	55a3fe356...	55a3fe356...	18.9000	-84.9322	45.4705	NaN	NaN	NaN	NaN	51.1983
5 Sun Jul 12 16:18:45 UTC 2015	55a3fe356...	55a3fe356...	18.0000	-84.9321	45.4706	NaN	NaN	NaN	NaN	49.1095
6 Sun Jul 12 16:19:13 UTC 2015	55a3fe356...	55a3fe356...	58.5000	-84.9305	45.4686	NaN	NaN	NaN	NaN	73.2005
7 Sun Jul 12 16:19:14 UTC 2015	55a3fe356...	55a3fe356...	56.7000	-84.9304	45.4685	NaN	NaN	NaN	NaN	75.3612
8 Sun Jul 12 16:19:15 UTC 2015	55a3fe356...	55a3fe356...	57.6000	-84.9304	45.4683	NaN	NaN	NaN	NaN	70.7542
9 Sun Jul 12 16:19:16 UTC 2015	55a3fe356...	55a3fe356...	56.7000	-84.9303	45.4682	NaN	NaN	NaN	NaN	62.8340

Working with JSON in MATLAB



Ad Hoc Access to Data from MATLAB

athenaQuery.mlx

Access the data in S3

Bring up the AthenaClient

```
athenaClient = aws.athena.Client();
athenaClient.Database = 'trainingdata';
athenaClient.initialize();
```

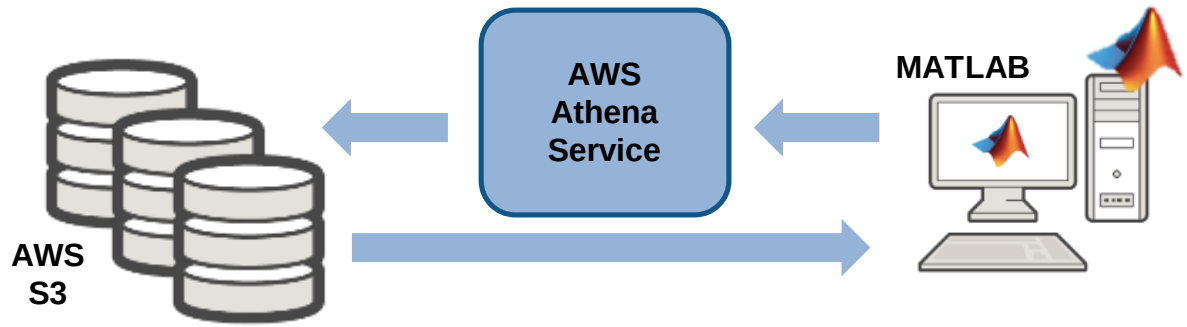
Create a query and submit

```
athenaClient.submitQuery('SELECT * FROM "trainingdata"."sampledata" limit 100','s3://fleettrainingdata')
```

Fetch data as a table for easy analysis

```
ds = datastore('s3://fleettrainingdata/*.csv');
ds.NumHeaderLines = 2;
data = table(ds);
```

Your usual MATLAB workflow goes here



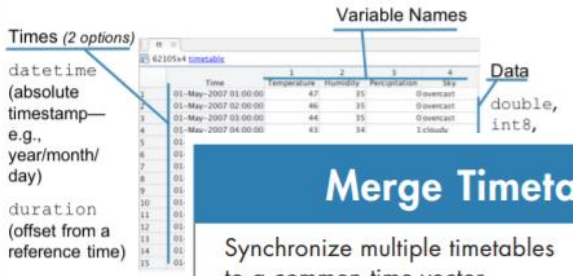
2 Preprocess Data

Pre-processing and Feature Engineering

Timetable

MATLAB datatype designed to organize and work with time series data.

Components of a Timetable



Create Timetabl

```
tt = timeta
     ... ,va
```

(All variables m of rows.)

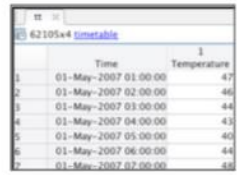
```
tt = table2
```

(The first datetir becomes the ro

Timetable Manipulation

Access Data These return the same array:

```
tt.Temperature
tt{:,'Temperature'}
tt{:,1}
```



Add a New Variable

```
r = zeros(height(tt),1);
```

riable Nar

rties.Va
mes;

st be valic

tlab.la
names fr

Data Using

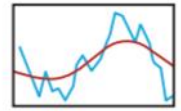
ime(tt,r

used to fil
ptions as
').

Data Cleaning

Smooth Data

```
B = smoothdata(A,method);
```



Smooth noisy data with methods:

```
'movmean','movmedian','gaussian',
'lowess','loess','rlowess',
'rloess','sgolay'
```

lier(A,meth

s with method

ean','quart
esd'

Points

nge(A,metho

anges with me

iance','line

Merge Timetables

Synchronize multiple timetables to a common time vector.

```
tt = synchronize(tt1,tt2,...,ttN);
```

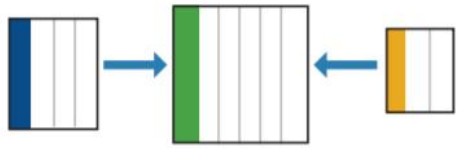
Synchronizing often results in missing data points (times at which a variable was not measured). `synchronize` supports several methods for adjusting data to fill in gaps:

Fill: `'fillwithmissing','fillwithconstant'`

Interpolation: `'linear','spline','pchip'`

Nearest Neighbor: `'previous', 'next','nearest'`

Aggregation: `'mean','min','max',@func,...`



Missing Data

Find Missing Values

```
TF = ismissing(tt);
```

Fill Missing Values

```
tt = fillmissing(tt,method);
```

Replace missing values with values calculated from nearby points with methods:

```
'previous','next','nearest',
'linear','spline','pchip'
```

Remove Rows Containing Missing Values

```
tt = rmmissing(tt);
```

Time	1 Temperature	2 Humidity
01-May-2007 01:00:00	47	35
01-May-2007 02:00:00	46	NaN
01-May-2007 03:00:00	NaN	35
01-May-2007 04:00:00	NaN	34
01-May-2007 05:00:00	40	34
01-May-2007 06:00:00	44	35
01-May-2007 07:00:00	48	35

Big Data

Tall arrays extend MATLAB functions to work on data too big to load into memory.

Create a "tall" timetable:

```
% Create a datastore that points to
% the data
ds = datastore('*.csv');

% Create a tall table from the
% datastore
t = tall(ds);

% Convert to a timetable
tt = table2timetable(t);
```

M×5 tall timetable

Time	LATP	LONP	ALT	PTCH	ROLL
10-May-2001 16:24:12	39.055	-84.661	866	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	NaN	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	866	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	NaN	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	866	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	NaN	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	866	-0.37352	0.076902
10-May-2001 16:24:12	NaN	NaN	NaN	-0.37352	0.076902
:	:	:	:	:	:
:	:	:	:	:	:

Develop a Preprocessing Function

Timetable

t = 4647×40 timetable

	trip_id	VIN	kff1001	kff1005	kff1006	kff1220	kff1221	kff1222	kff1223	kff125a
1 Sun Jul 12 16:18:41 UTC 2015	55a3fe356...	55a3fe356...	17.1000	-84.9323	45.4704	NaN	NaN	NaN	NaN	59.0434
2 Sun Jul 12 16:18:42 UTC 2015	55a3fe356...	55a3fe356...	17.1000	-84.9322	45.4704	NaN	NaN	NaN	NaN	57.8609
3 Sun Jul 12 16:18:43 UTC 2015	55a3fe356...	55a3fe356...	18.9000	-84.9322	45.4705	NaN	NaN	NaN	NaN	52.7147
4 Sun Jul 12 16:18:44 UTC 2015	55a3fe356...	55a3fe356...	18.9000	-84.9322	45.4705	NaN	NaN	NaN	NaN	51.1983
5 Sun Jul 12 16:18:45 UTC 2015	55a3fe356...	55a3fe356...	18.0000	-84.9321	45.4706	NaN	NaN	NaN	NaN	49.1095
6 Sun Jul 12 16:19:13 UTC 2015	55a3fe356...	55a3fe356...	58.5000	-84.9305	45.468	NaN	NaN	NaN	NaN	73.2005
7 Sun Jul 12 16:19:14 UTC 2015	55a3fe356...	55a3fe356...	56.7000	-84.9304	45.468					
8 Sun Jul 12 16:19:15 UTC 2015	55a3fe356...	55a3fe356...	57.6000	-84.9304	45.468					
9 Sun Jul 12 16:19:16 UTC 2015	55a3fe356...	55a3fe356...	56.7000	-84.9303	45.468					

- ✓ Clean up
- ✓ Enrich
- ✓ Restructure

Preprocess data

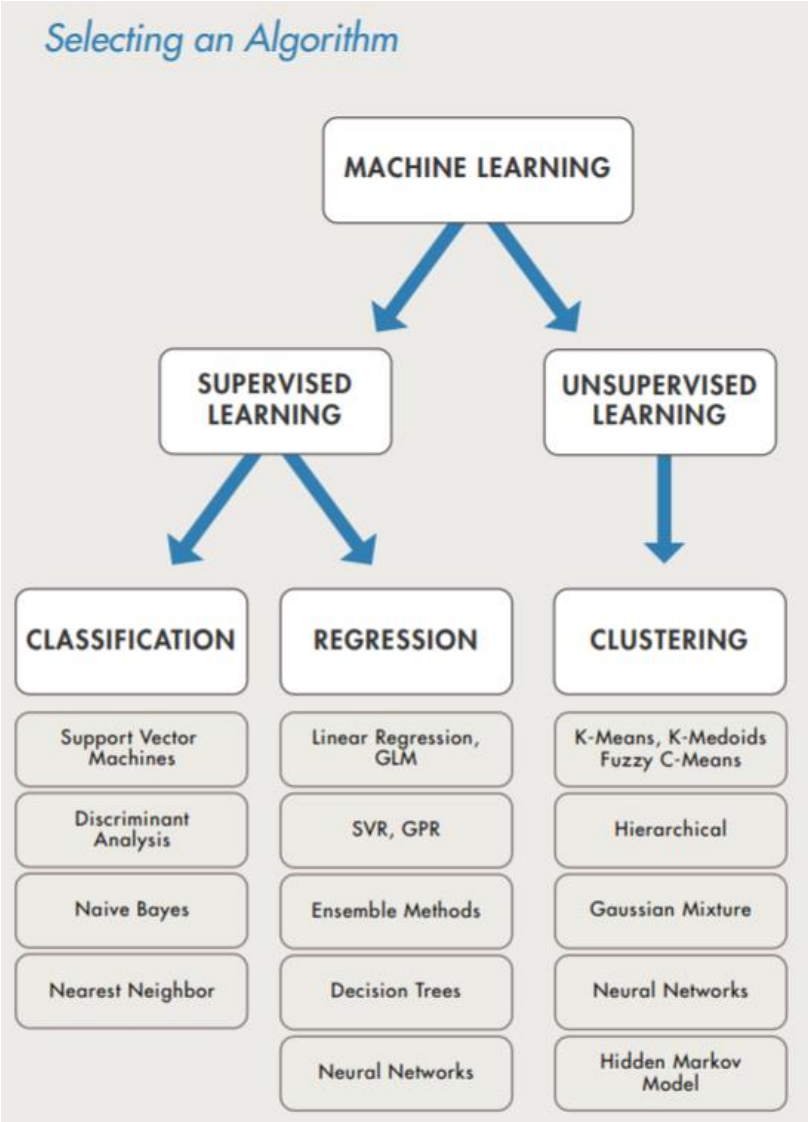
```
t = sortrows(t);
t = rmmissing(t,'MinNumMissing',width(t)-2);
```

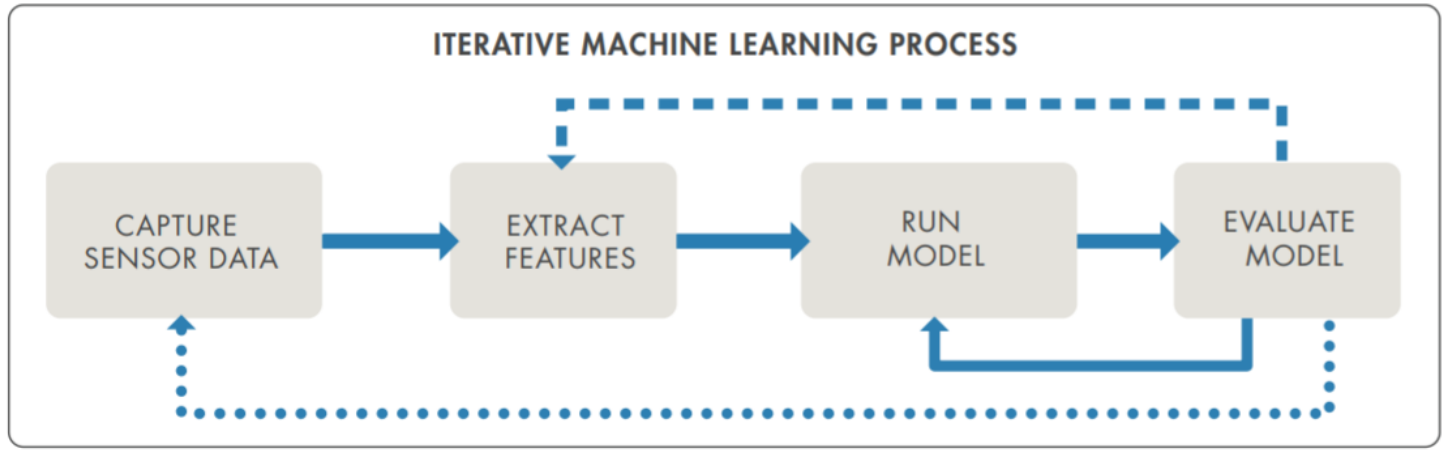
Perform windowed calculations

```
t.Speed = movmedian(t.SpeedGPS,3);
t.D1 = [0;diff(t.SpeedGPS)];
```

```
[tmin,tmax] = bounds(t.time);
tnew = tmin:seconds(10):tmax;
countsByTime = retime(t(:,'Event'),tnew,@histcounts);
```

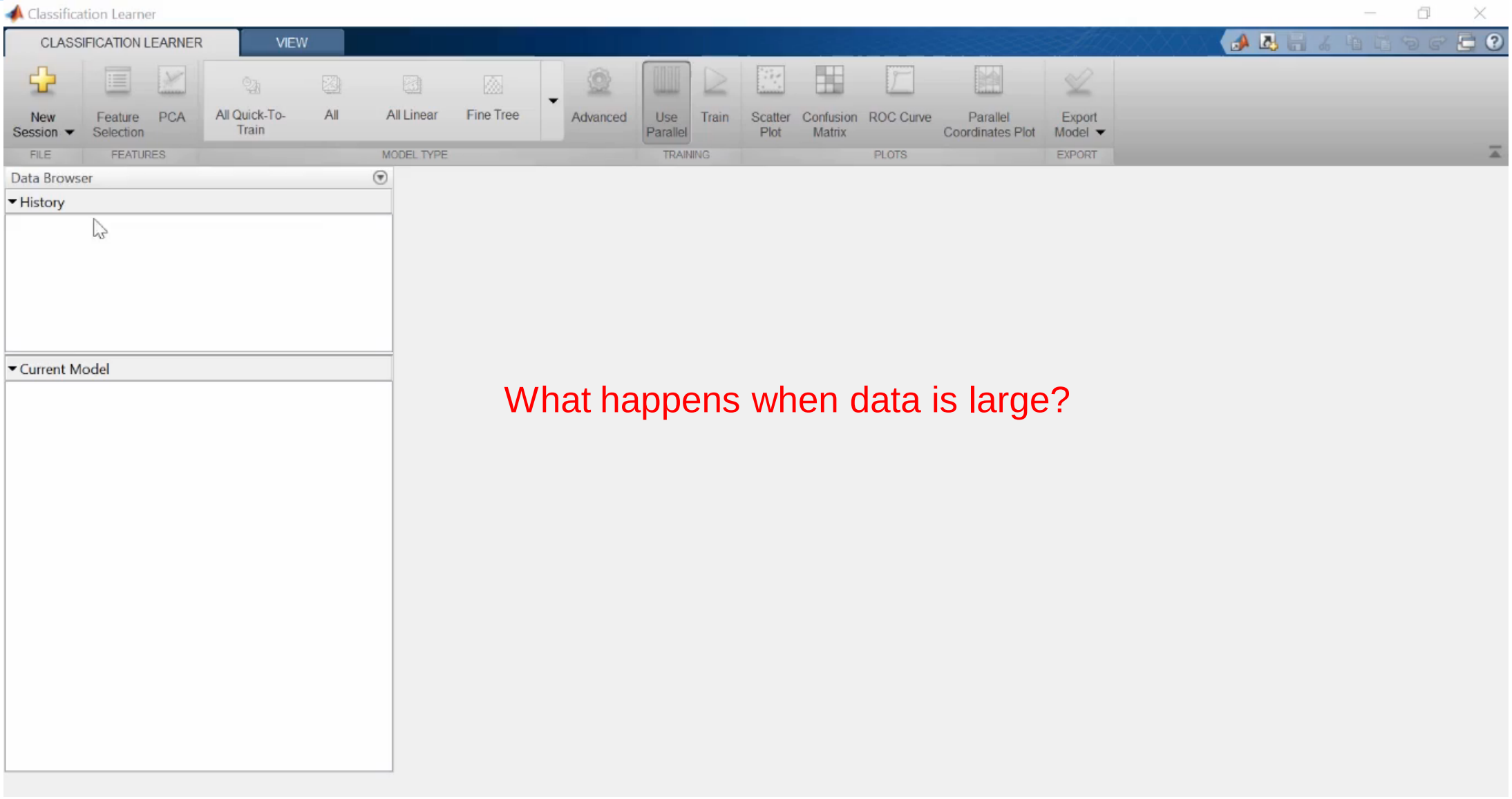
Building predictive models





Algorithm	Prediction Speed	Training Speed	Memory Usage	Required Tuning	General Assessment
Logistic Regression (and Linear SVM)	Fast	Fast	Small	Minimal	Good for small problems with linear decision boundaries
Decision Trees	Fast	Fast	Small	Some	Good generalist, but prone to overfitting
(Nonlinear) SVM (and Logistic Regression)	Slow	Slow	Medium	Some	Good for many binary problems, and handles high-dimensional data well
Nearest Neighbor	Moderate	Minimal	Medium	Minimal	Lower accuracy, but easy to use and interpret
Naïve Bayes	Fast	Fast	Medium	Some	Widely used for text, including spam filtering
Ensembles	Moderate	Slow	Varies	Some	High accuracy and good performance for small- to medium-sized datasets
Neural Network	Moderate	Slow	Medium to Large	Lots	Popular for classification, compression, recognition, and forecasting

Develop a Predictive Model in MATLAB

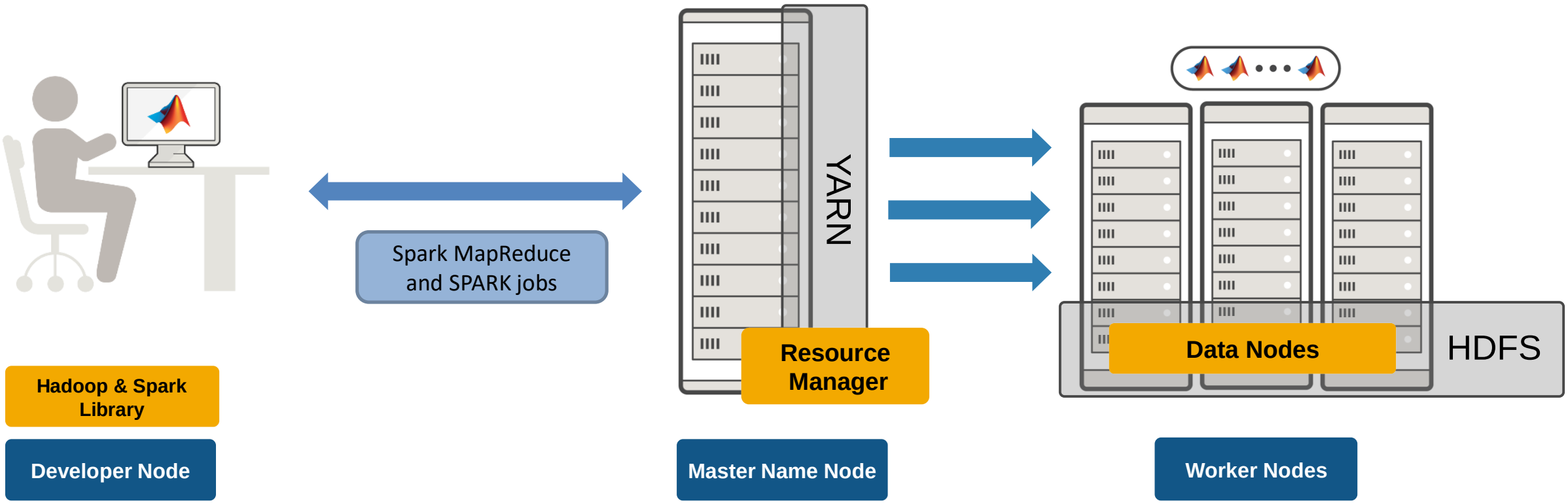


What happens when data is large?

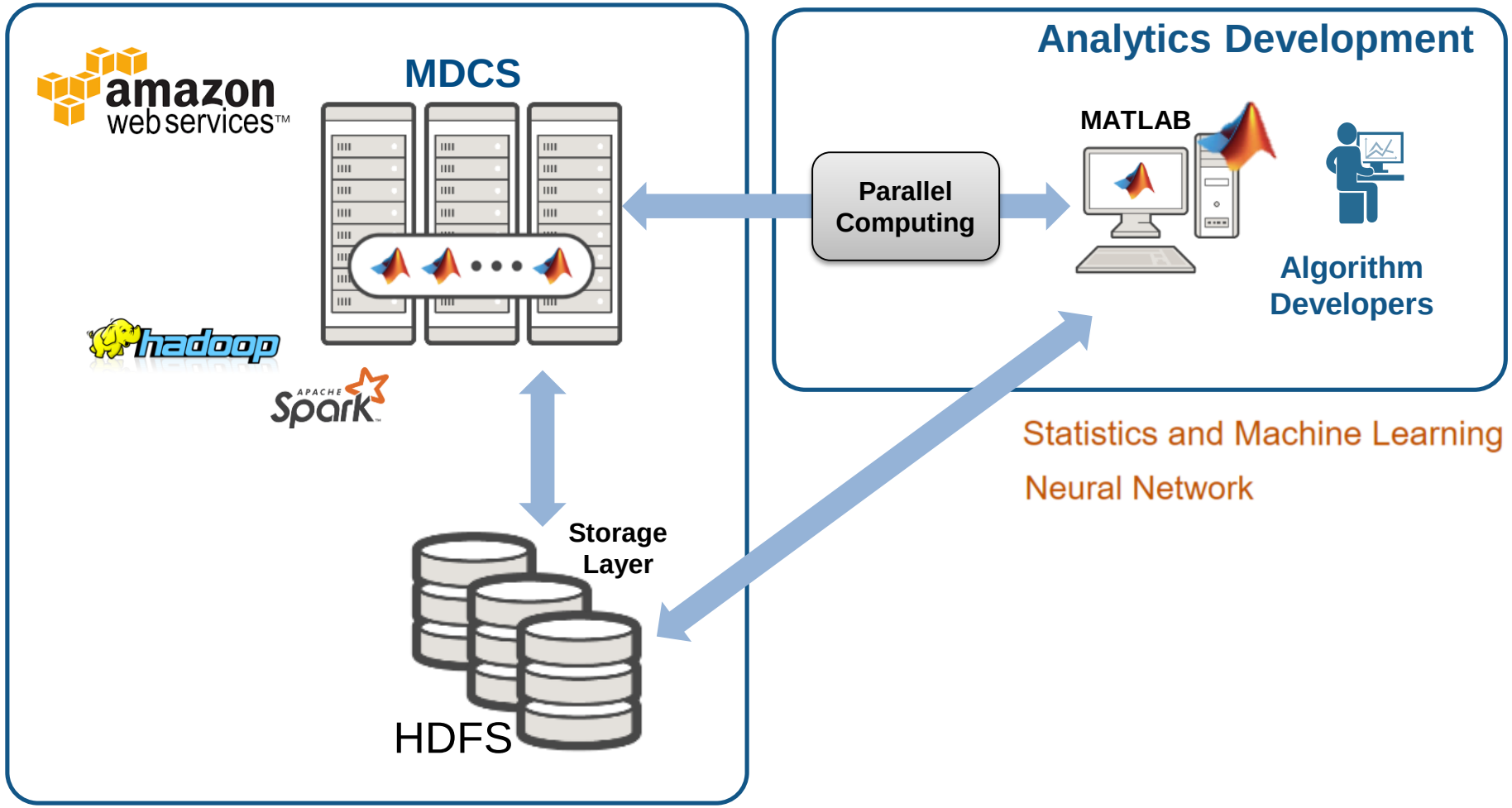
Submit Big Data jobs from MATLAB on HADOOP & SPARK

MATLAB workers on worker nodes in the cluster

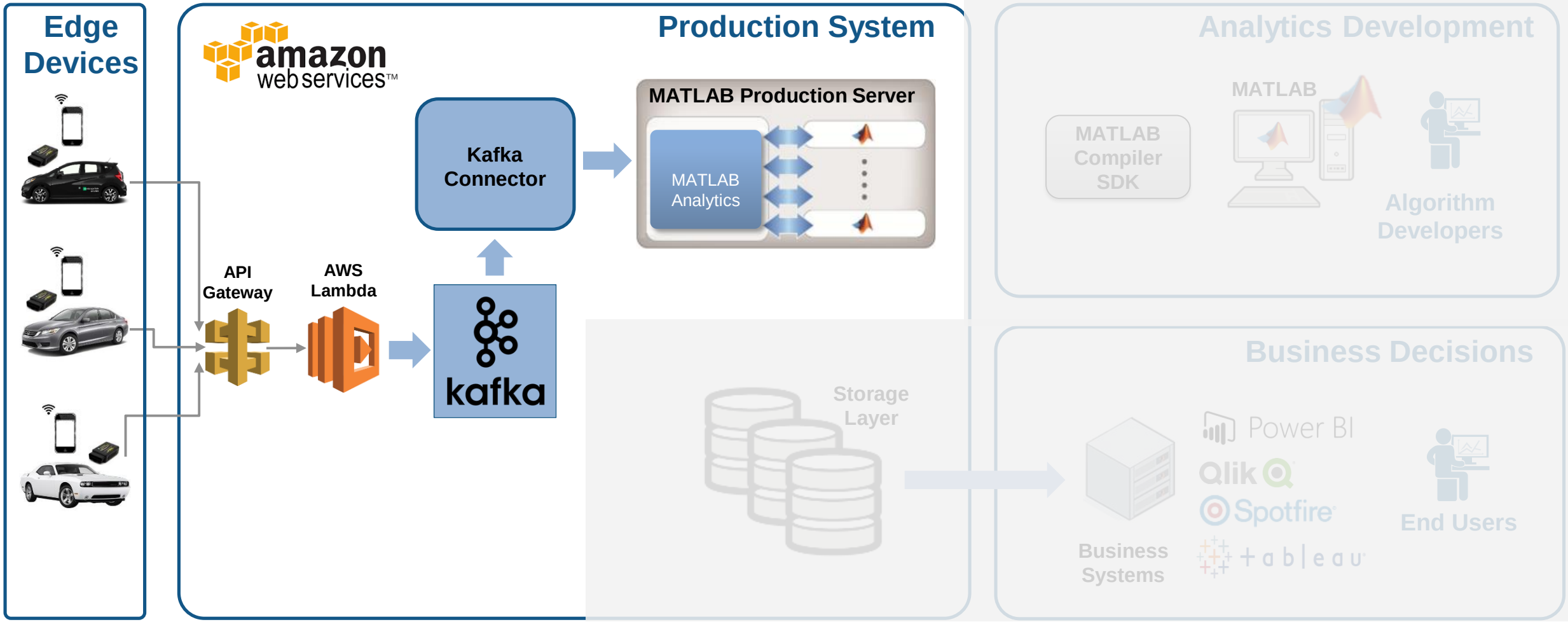
- MDCS workers (working from MATLAB)



Develop a Predictive Model

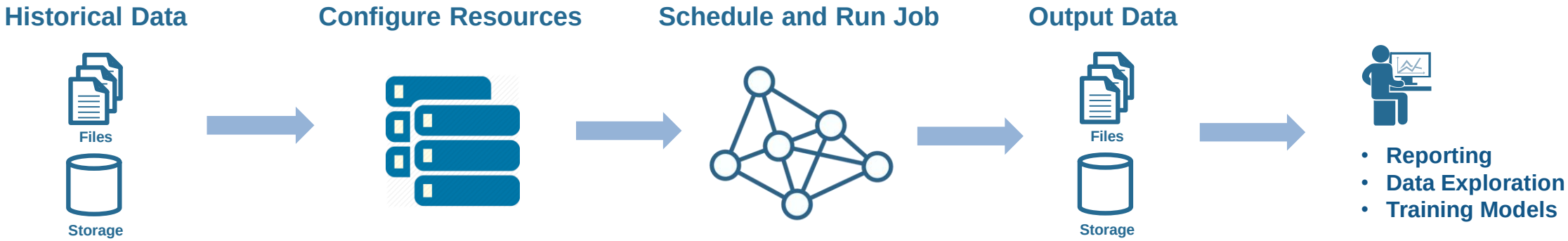


Develop and Deploy a Stream Processing Function

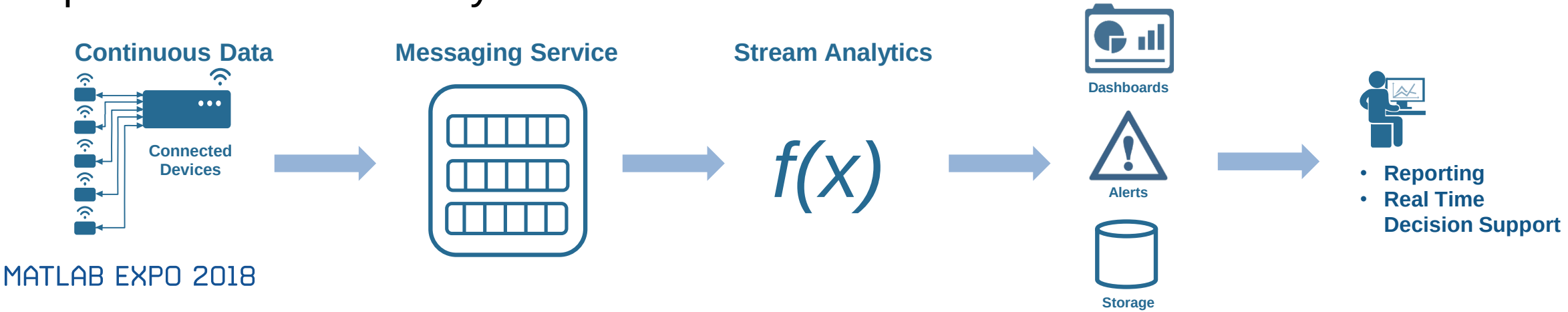


A quick Intro to Stream Processing

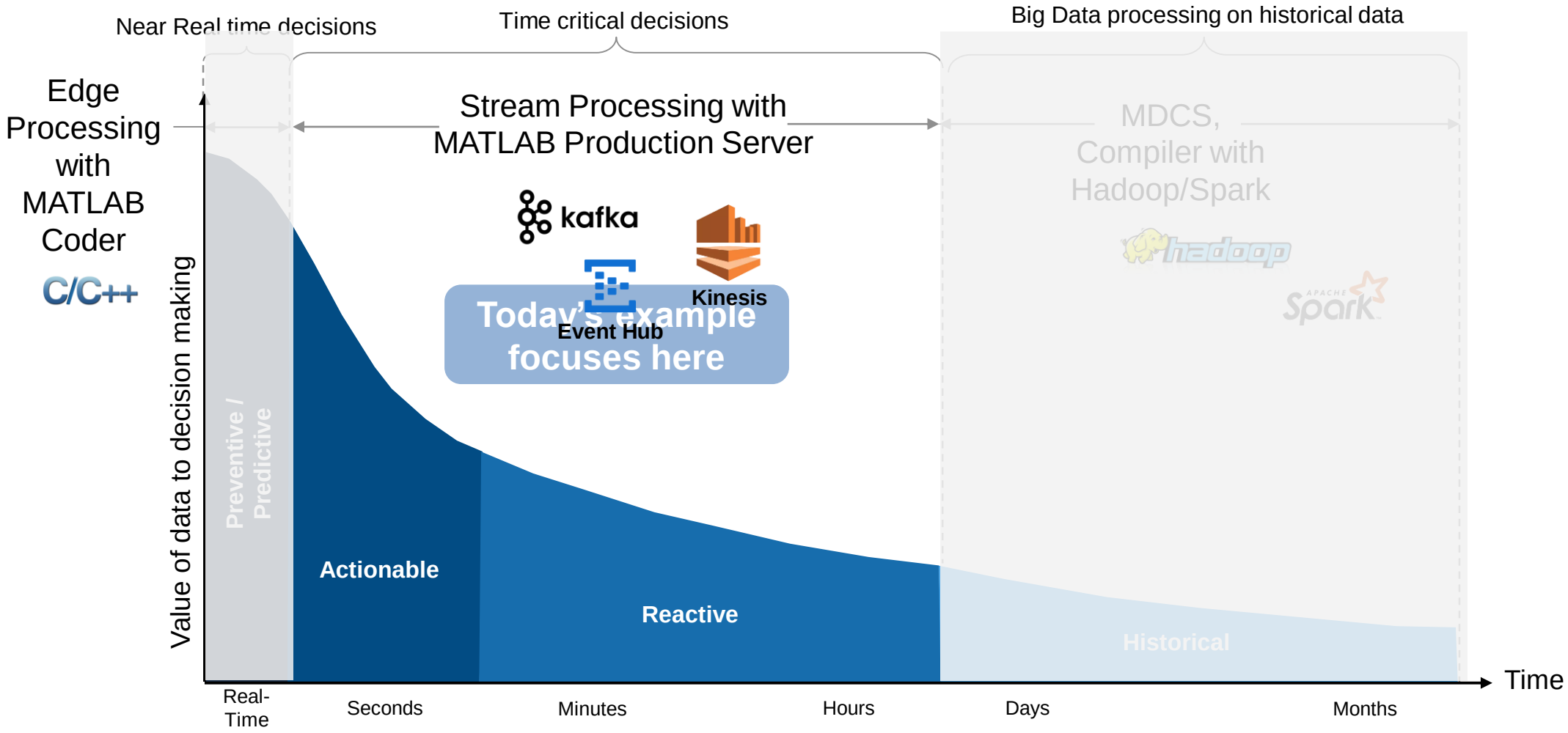
- ***Batch Processing*** applies computation to a finite sized historical data set that was acquired in the past



- ***Stream Processing*** applies computation to an unbounded data set that is produced continuously

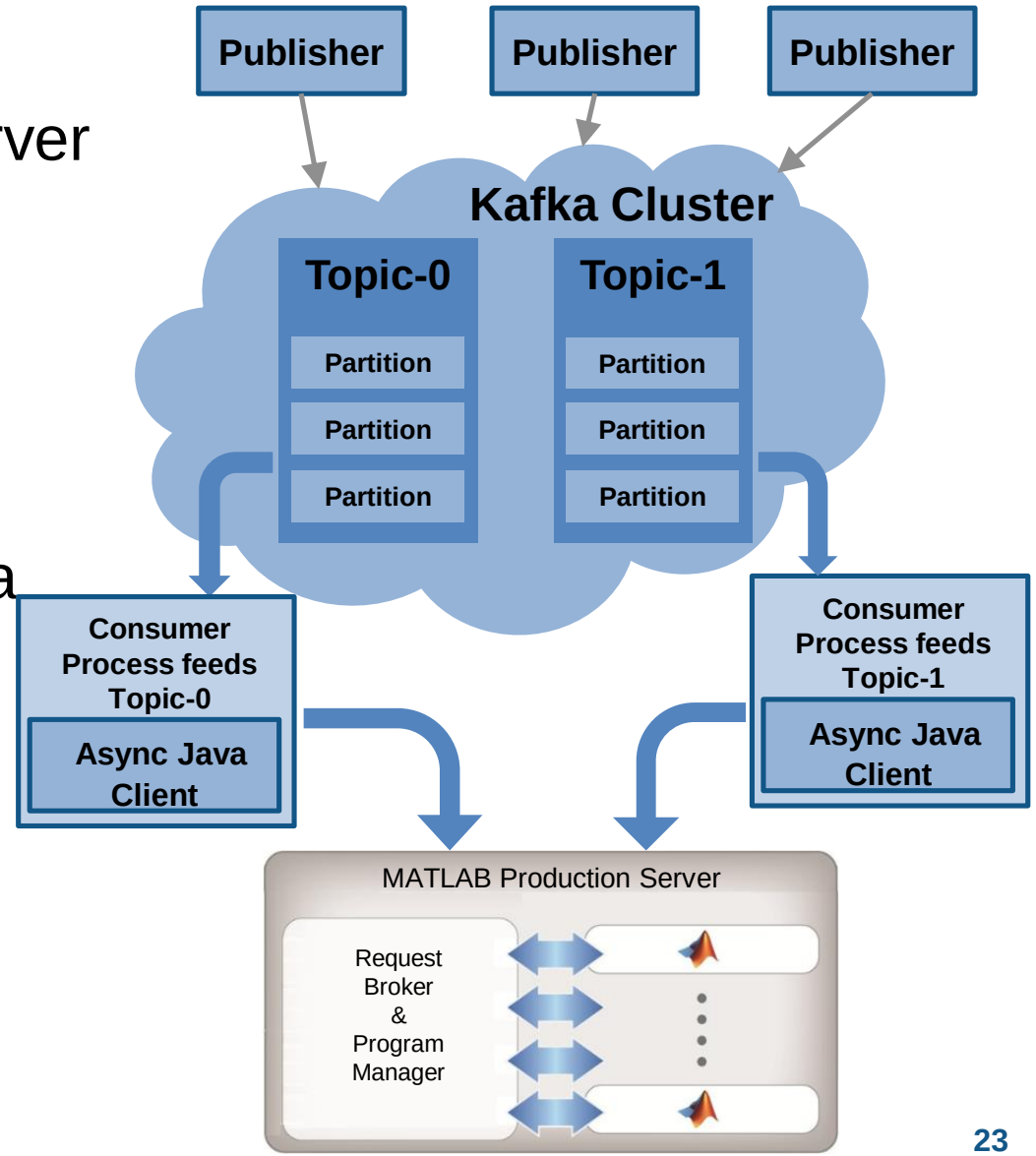


Why stream processing?



Connecting MATLAB Production Server to Kafka

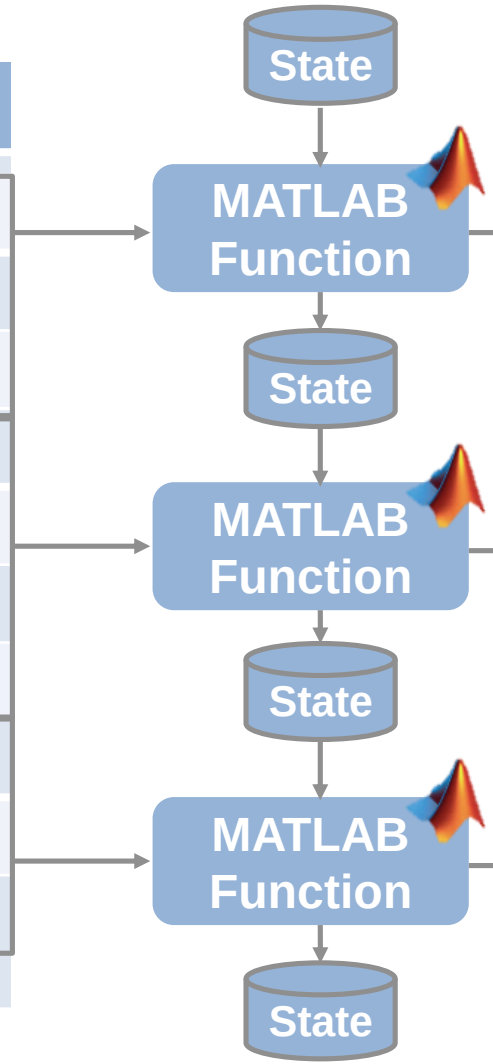
- Kafka client for MATLAB Production Server feeds topics to functions deployed on the server
- Each consumer process feeds one topic to a specified function
- Configurable batch of messages passed as a MATLAB Timetable
- Drive everything from a simple config file
  - No programming outside of MATLAB!



Streaming data is treated as an unbounded Timetable

Input Table

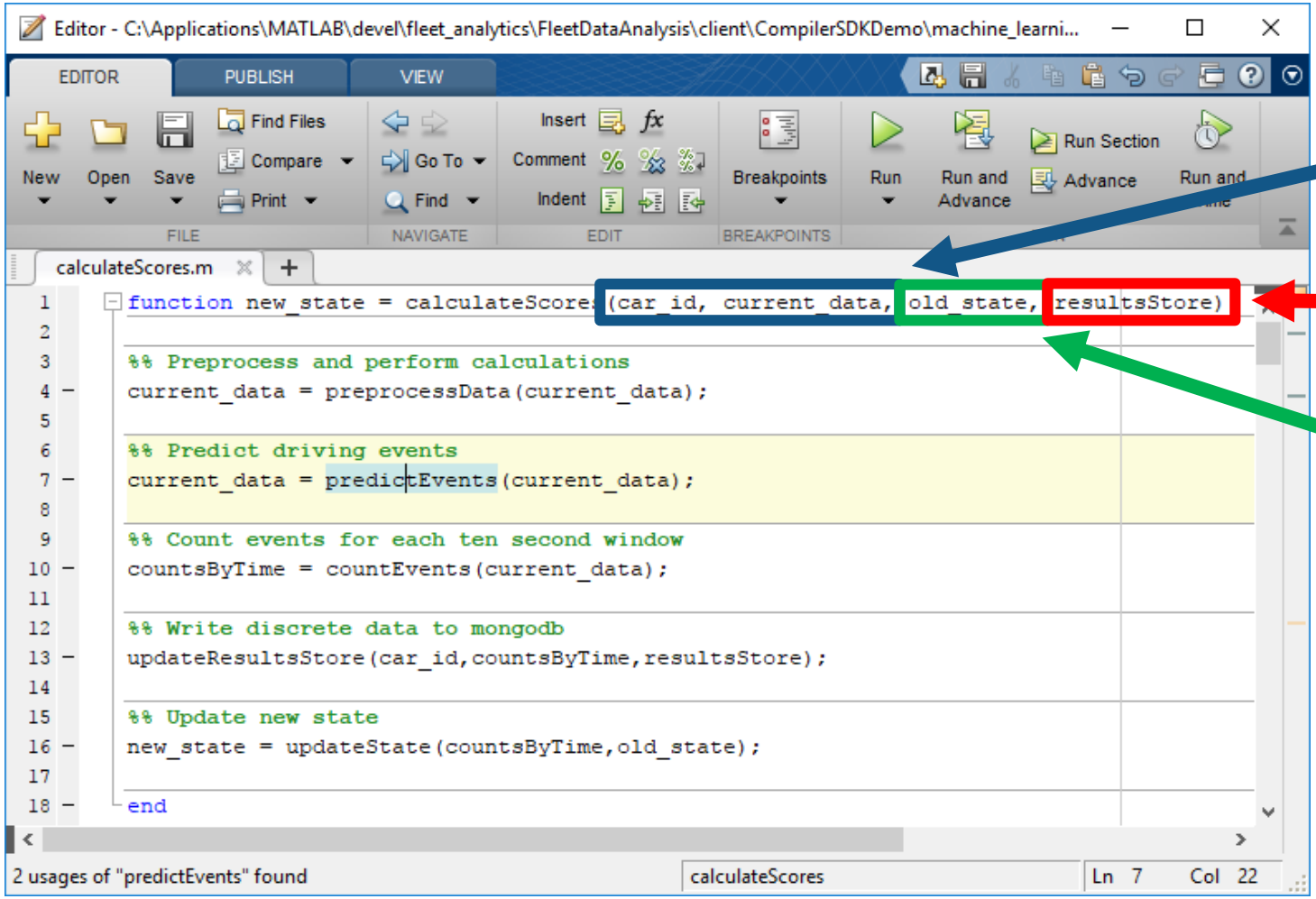
Event Time	Vehicle	RPM	Torque	Fuel Flow
18:01:10	55a3fd	1975	100	110
18:10:30	55a3fe	2000	109	115
18:05:20	55a3fd	1980	105	105
18:10:45	55a3fd	2100	110	100
18:30:10	55a419	2000	100	110
18:35:20	55a419	1960	103	105
18:20:40	55a3fe	1970	112	104
18:39:30	55a419	2100	105	110
18:30:00	55a3fe	1980	110	113
18:30:50	55a3fe	2000	100	110
…	…	…	…	…



Output Table

Time window		Vehicle	Score
…		…	…
18:00:00	18:10:00	55a3fd	5
		55a3fe	…
		55a419	…
18:10:00	18:20:00	55a3fd	7
		55a3fe	3
		55a419	…
18:20:00	18:30:00	55a3fd	…
		55a3fe	4
		55a419	…
18:30:00	18:40:00	55a3fd	…
		55a3fe	5
		55a419	8

Develop a Stream Processing Function in MATLAB



Process each window of data (input table) as it arrives

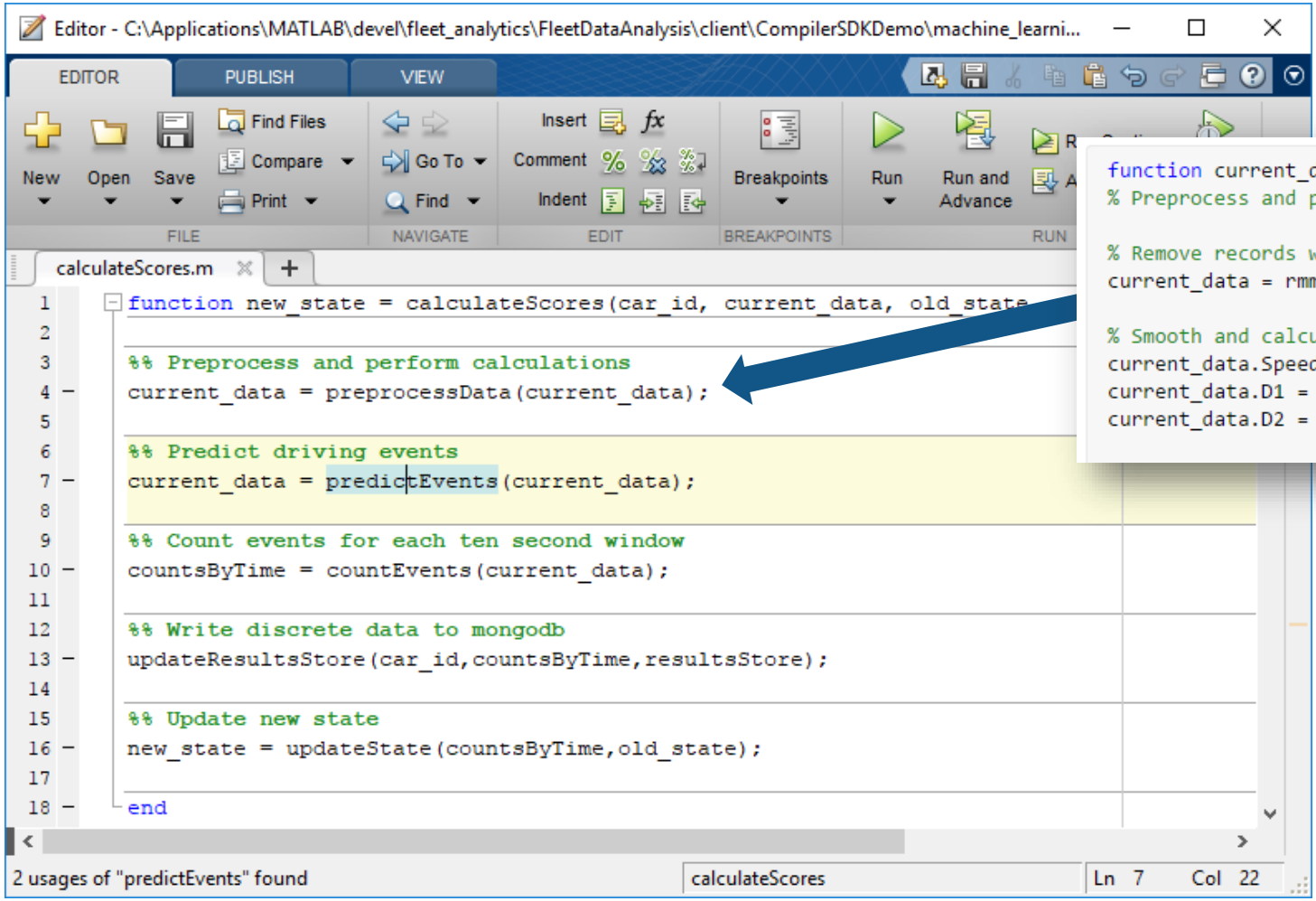
Current window of data to be processed

Current score (pointer)

Previous state (pointer)

Develop a Stream Processing Function in MATLAB

Process each window of data (input table) as it arrives



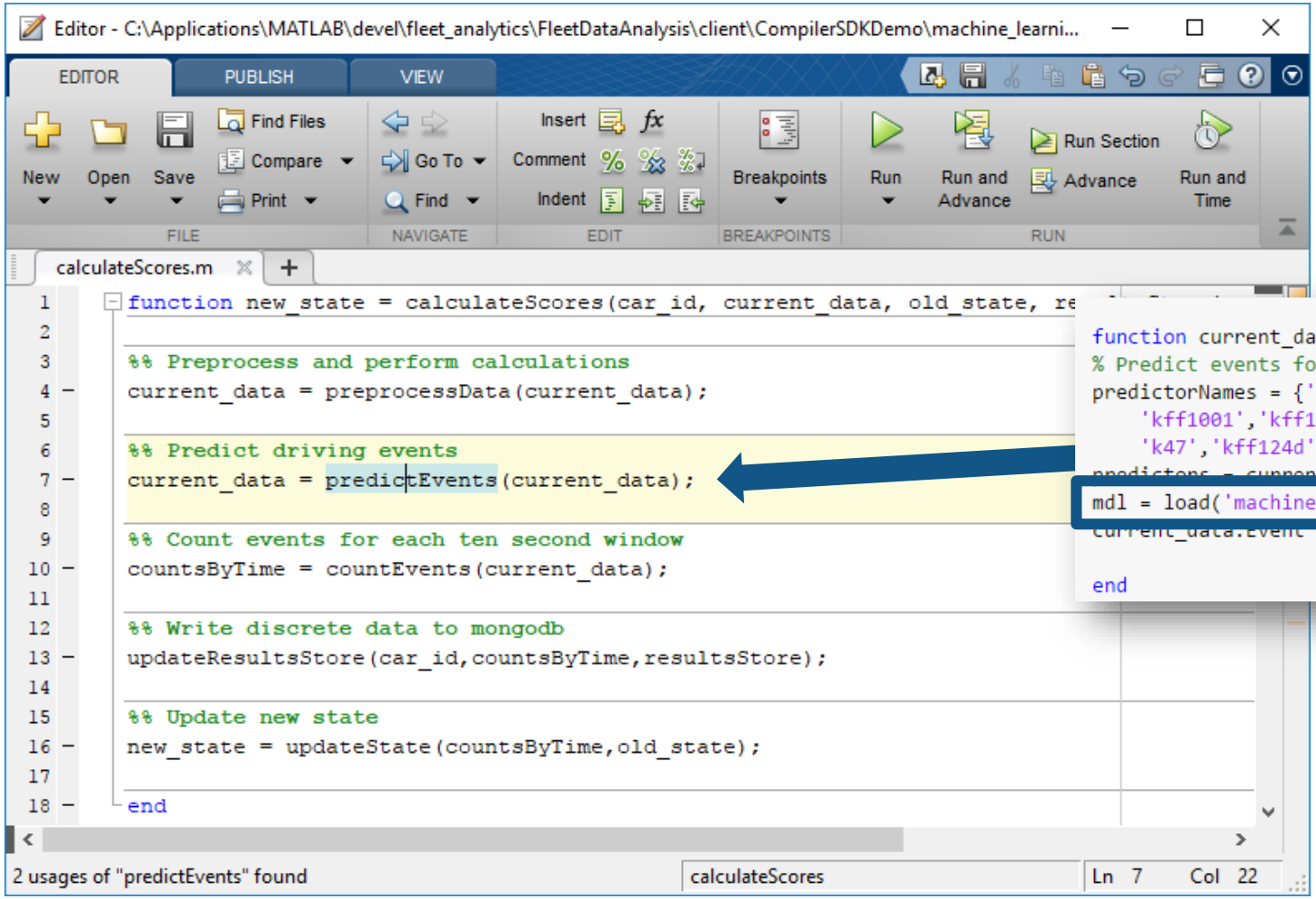

```
function current_data = preprocessData(current_data)
% Preprocess and perform calculations

% Remove records with all missing data
current_data = rmmissing(current_data,'MinNumMissing',width(current_data)-1);

% Smooth and calculate approximate gradients
current_data.Speed = movmedian(current_data.kff1001,5);
current_data.D1 = [0;diff(current_data.kff1001)];
current_data.D2 = [0;0;diff(current_data.kff1001,2)];
```

Develop a Stream Processing Function in MATLAB

Process each window of data (input table) as it arrives



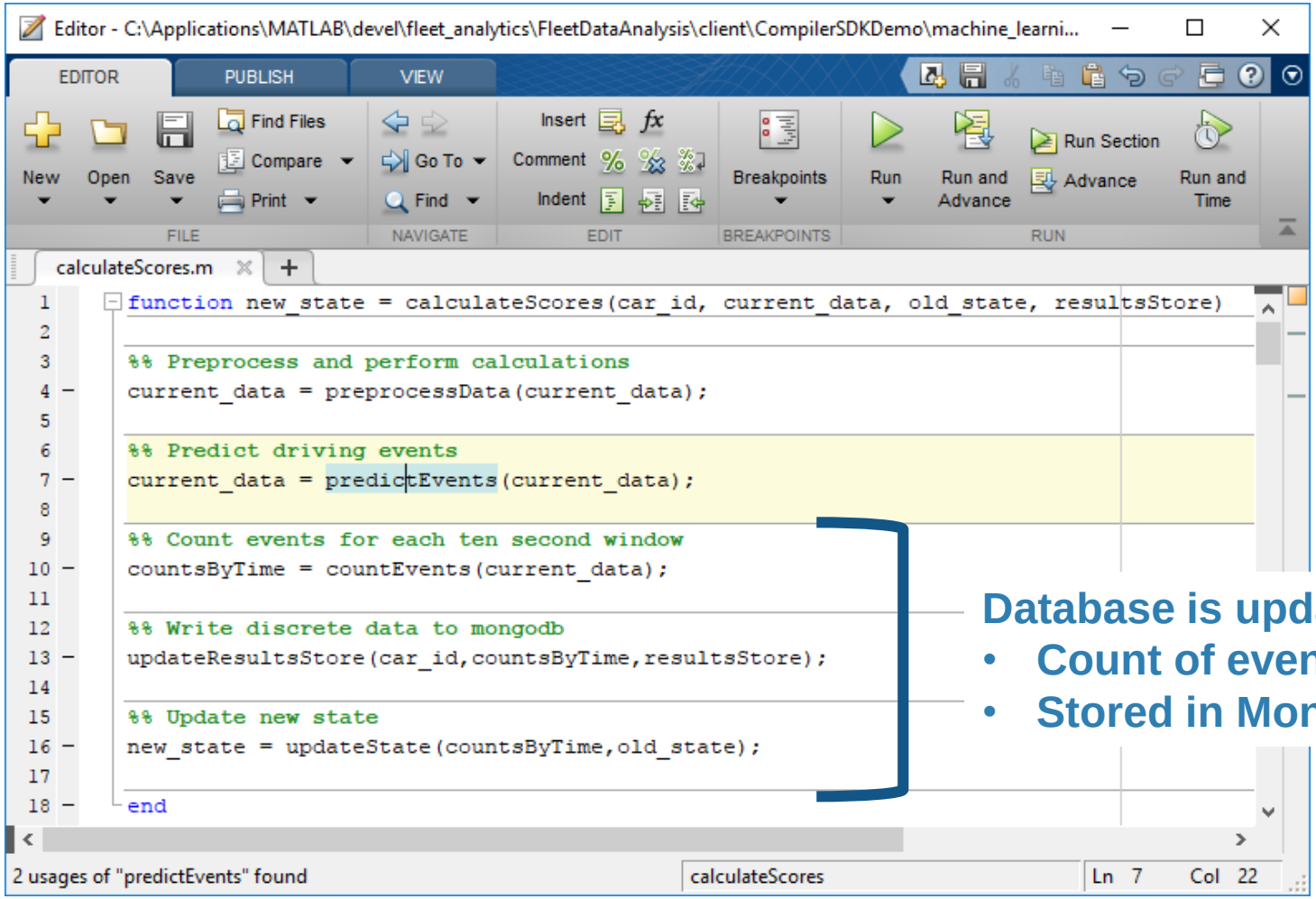

```
function current_data = predictEvents(current_data)
% Predict events for current data based on machine learning model
predictorNames = {'kff1005','kff1006','kff125a','k10','kff1249','Speed','D1','D2',...
    'kff1001','kff1220','kff1221','kff1222','kff1223',...
    'k47','kff124d'};
mdl = load('machineLearningModel.mat');

end
```

Use the model you created with Classification Learner App

Develop a Stream Processing Function in MATLAB

Process each window of data (input table) as it arrives

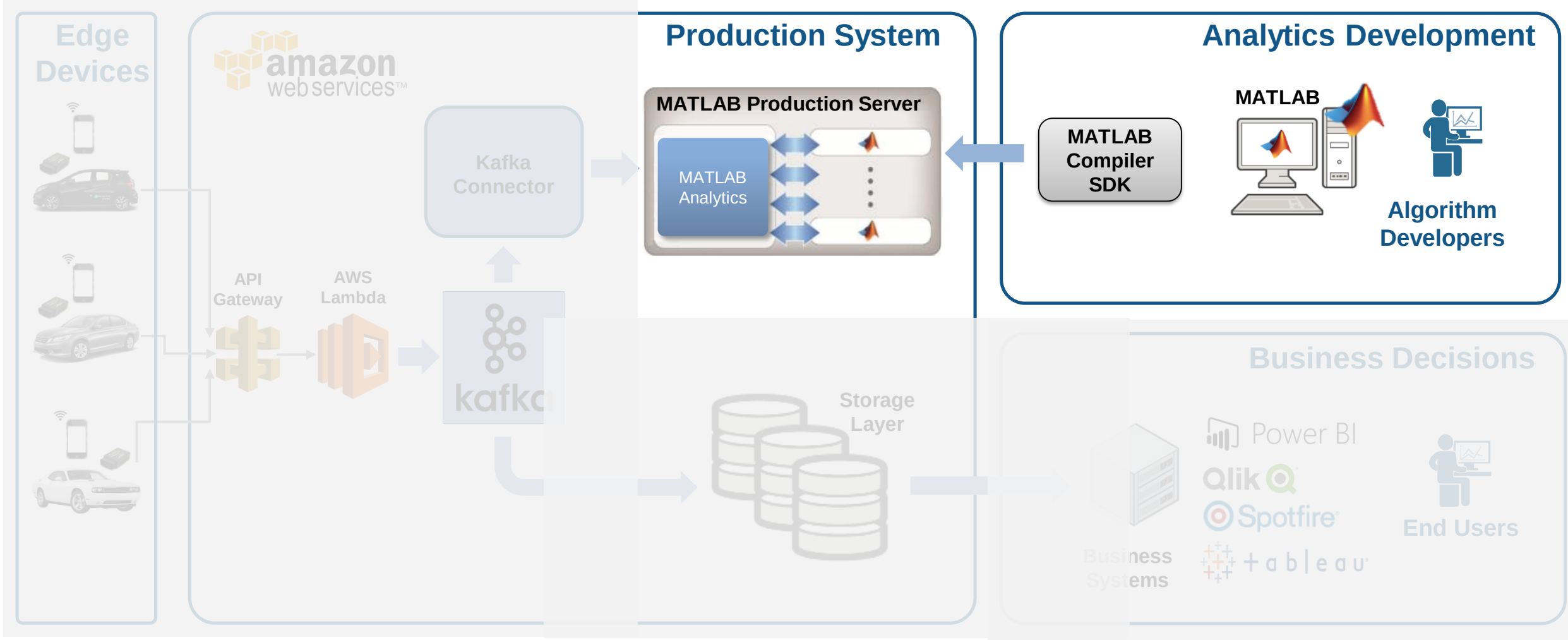


Database is updated with results of driver scoring

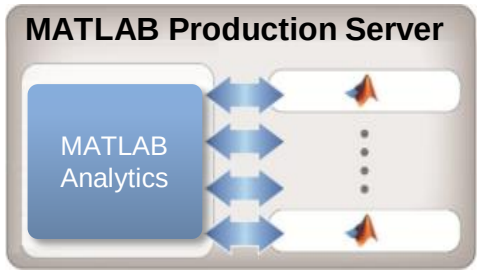
- **Count of events by type and location**
- **Stored in MongoDB instance**

4 Integrate with Production Systems

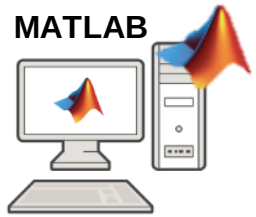
Operationalize Analytics into Production Systems

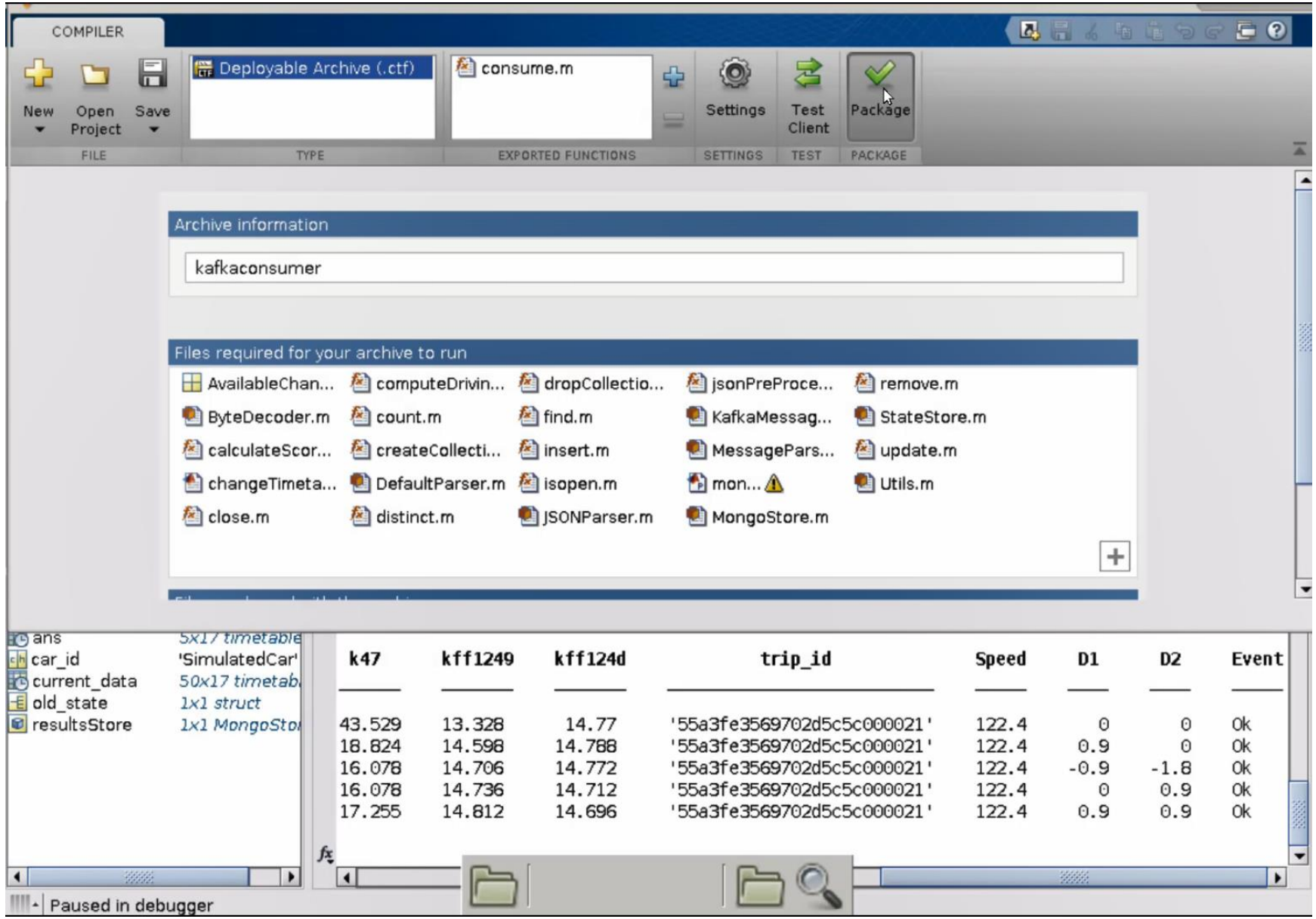


MATLAB Compiler SDK Workflow



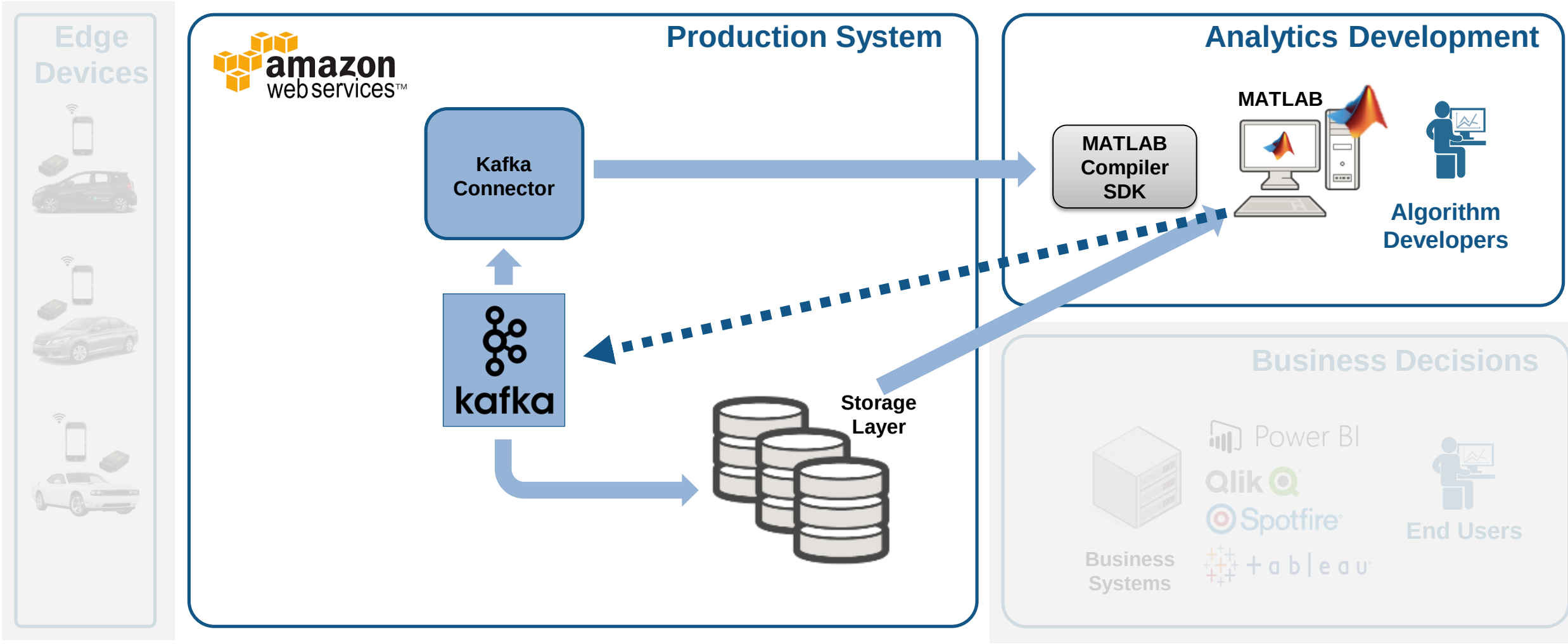
MATLAB Compiler SDK



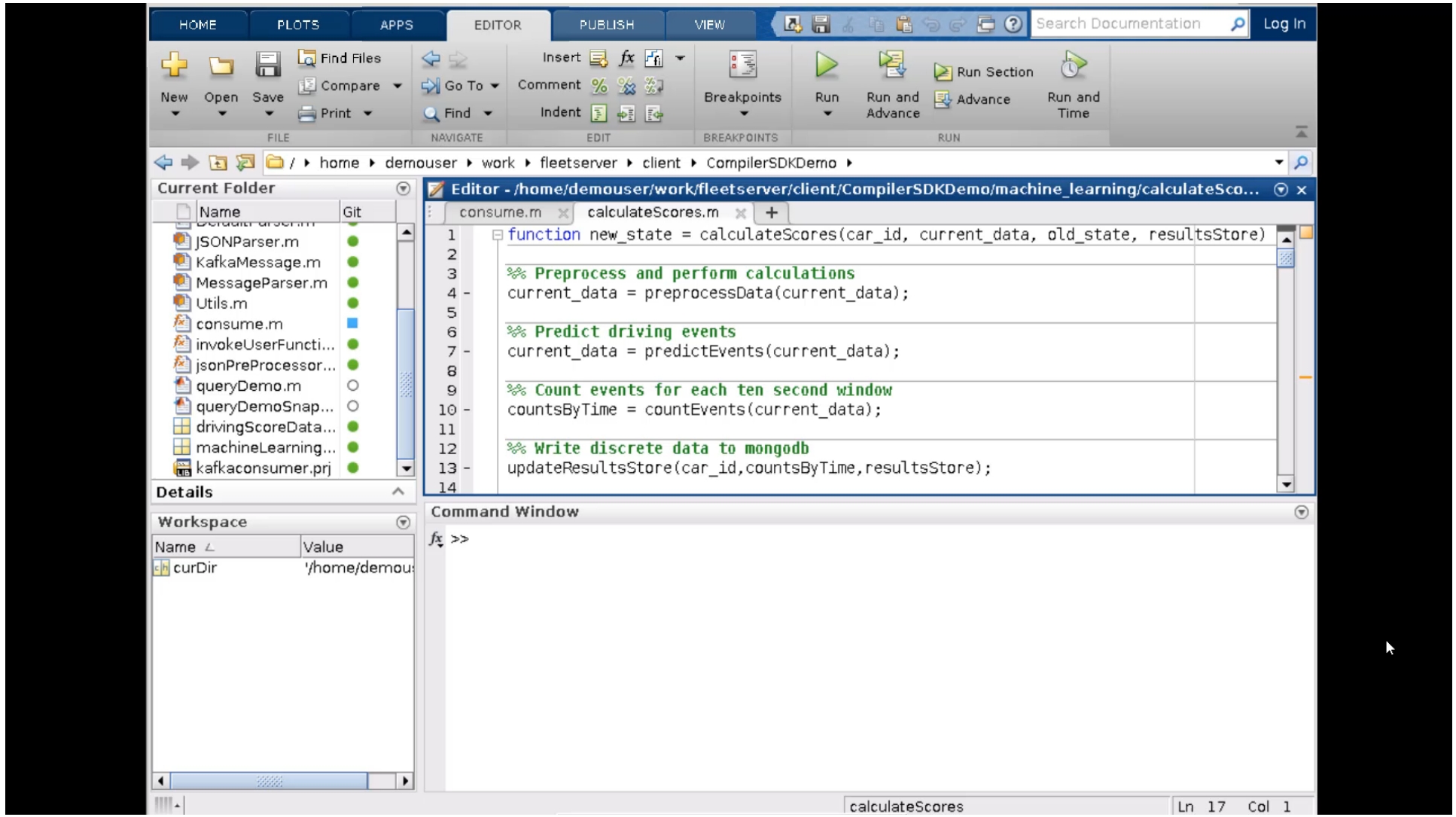


4 Integrate with Production Systems

Debug a Stream Processing Function in MATLAB

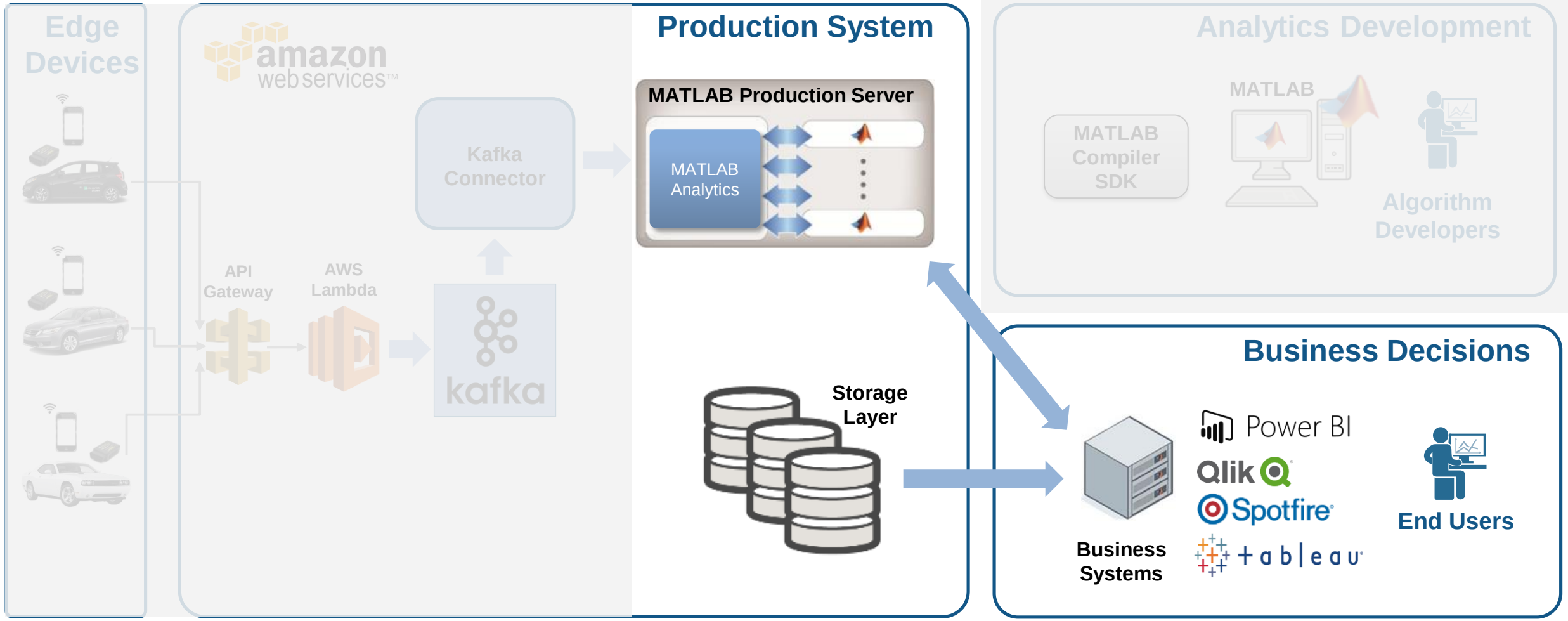


Debug a Stream Processing Function in MATLAB



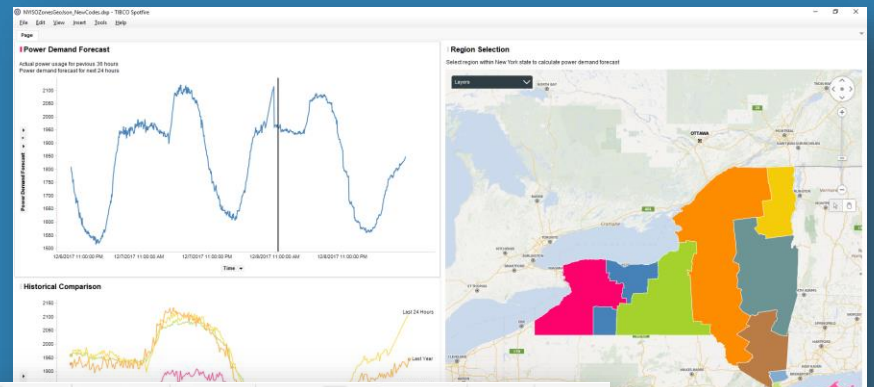
4 Integrate with Production Systems

Tie in your Dashboard Application

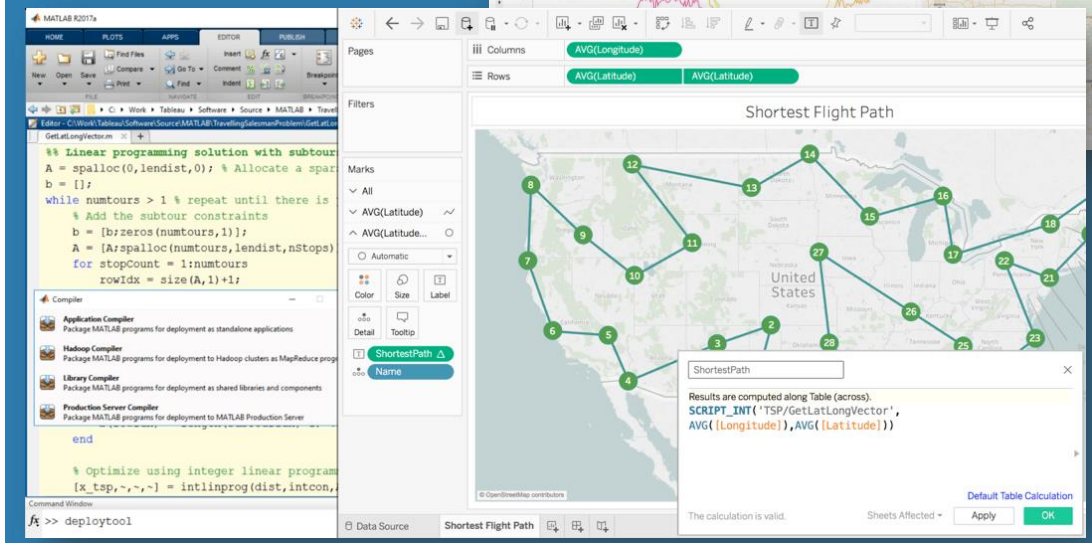


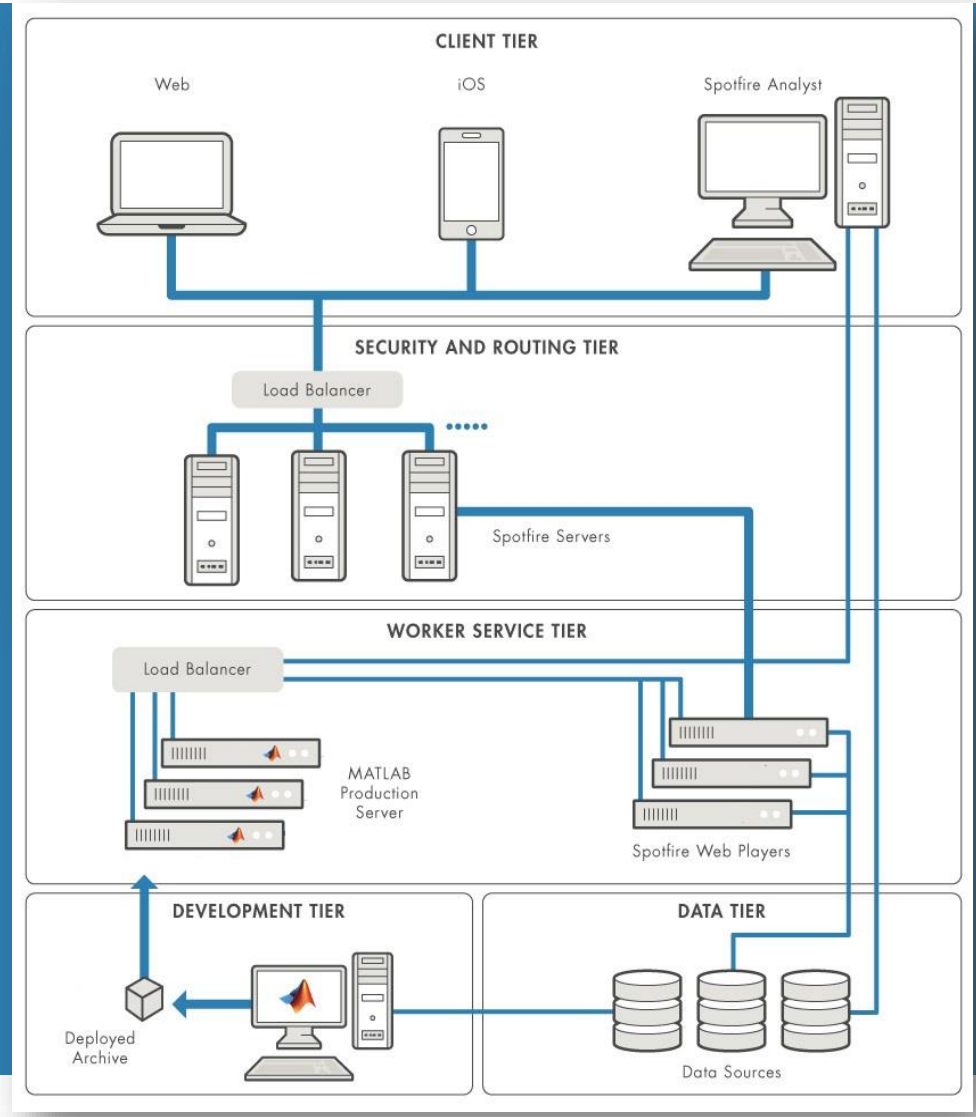
Scalable Analytics with Enterprise BI Tools

TIBCO Spotfire

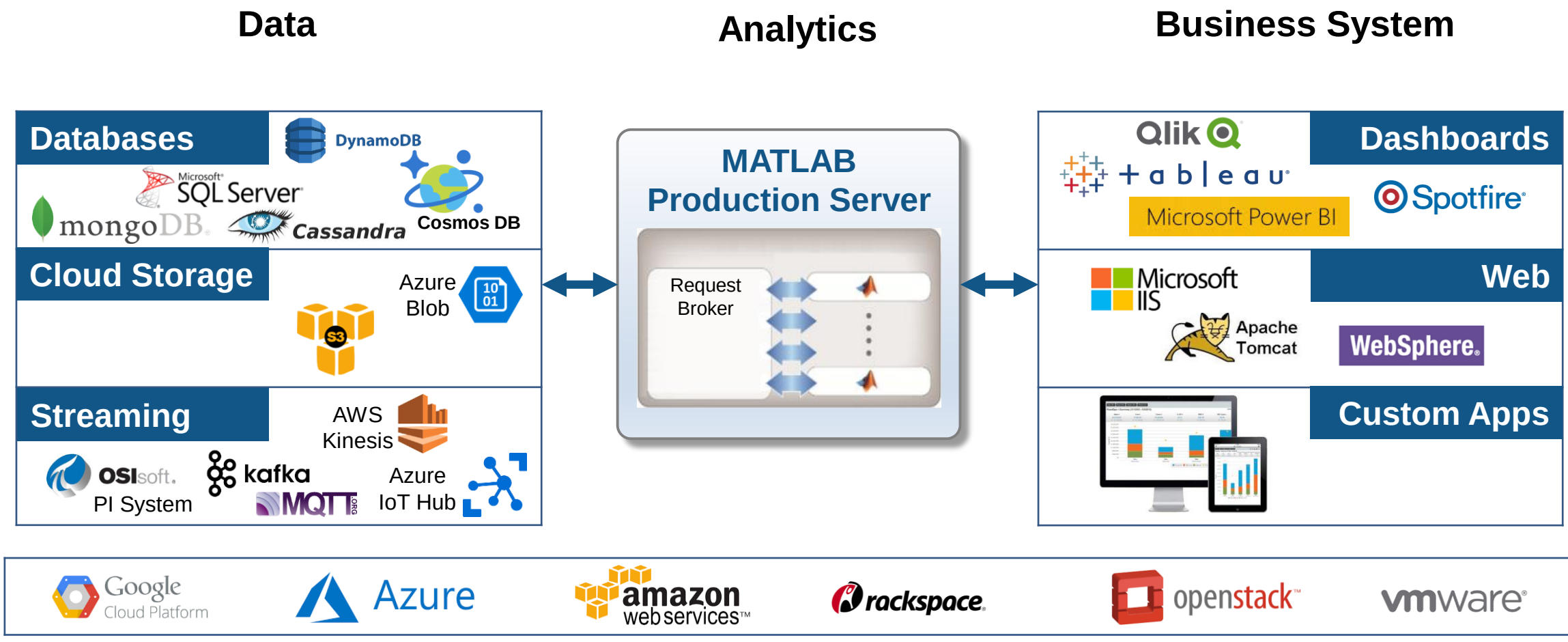


Tableau





MATLAB based applications in Production Level Ecosystem



Volkswagen Data Lab develops driver recognition algorithms with MATLAB

Develop technology building block for tailoring car features and services to individual

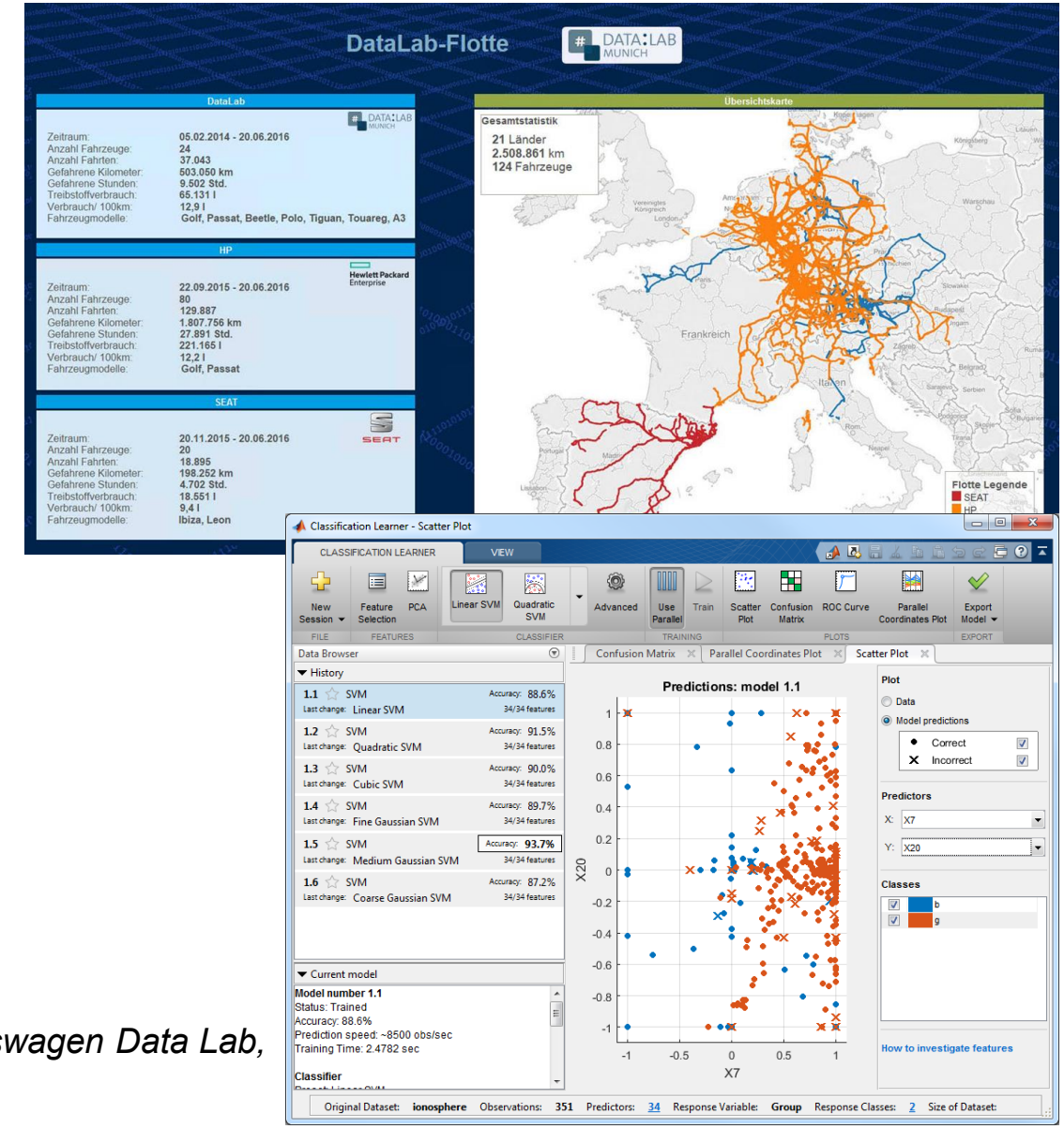
- Need to identify individual drivers based on their driving behavior using collected data

Challenges

- Accuracy despite low training data
- Robustness despite environmental conditions
- Computing time

Data sources

- Logged CAN bus data and travel record



Source: „Connected Car – Fahrererkennung mit MATLAB“ Julia Fumbarev, Volkswagen Data Lab, MATLAB EXPO Germany, June 27, 2017, Munich Germany

Key Takeaways

MATLAB users

- Spend your time understanding the data and designing algorithms
- You can run MATLAB on any development engine, desktop, server or cloud

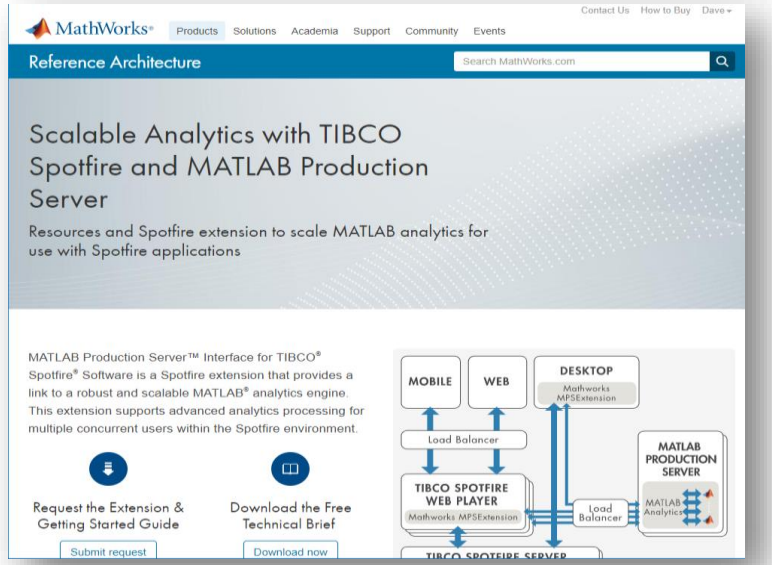
Solution architects

- MATLAB can connect directly to your data repositories
- MATLAB can deploy within your ecosystem and on platform of your choice using MATLAB Production Server

Resources to learn and get started

- Data Analytics with MATLAB
- Statistics and Machine Learning Toolbox
- Database Toolbox
- Mapping Toolbox

- MATLAB Production Server
- MATLAB Compiler SDK
- MATLAB with TIBCO Spotfire
- MATLAB with Tableau
- MATLAB with MongoDB



Speaker Details

Email: Pallavi.Kar@mathworks.in

LinkedIn: https://www.linkedin.com/in/pallavi-kar-2a591518

Twitter: @PallaviKar2512

Contact MathWorks India

Products/Training Enquiry Booth

Call: 080-6632-6000

Email: info@mathworks.in

Your feedback is valued.

Please complete the feedback form provided to you.

Data Analytics Applications	Controls and Embedded Systems	Signal Processing Systems	Robotics and Autonomous Systems
Predictive Maintenance Using MATLAB and Simulink *Amit Doshi, MathWorks*	Designing Efficient Power Electronics Systems Using Simulation *Vivek Raju, MathWorks* *Naga Chakrapani Pemmaraju, MathWorks*	Designing and Testing Voice Interfaces through Microphone Array Modeling, Audio Prototyping, and Text Analytics *Vidya Viswanathan, MathWorks*	Automated Driving Development with MATLAB and Simulink *Manohar Reddy, MathWorks*
Exhibition Break			
Using Fleet Analytics and MATLAB to Build Strategies for BS-VI Development *Shubham Garg, Honda*	Lithium-Ion Battery Parameter Estimation for HIL, SIL, and MIL Validation *Thayalan Shanmugam, RNTBCI*	Verifying the Hardware Implementation of Automotive Radar Signal Processing with MATLAB *Sainath K and Shashank Venugopal, NXP*	Autonomous Drive *Gopinath Chidambaram,L&T Technology Services*
Lunch and Exhibition			
Scaling up MATLAB Analytics with Kafka and Cloud Services *Pallavi Kar, MathWorks*	Full Vehicle Simulation for Electrification and Automated Driving Applications *Prasanna Deshpande, MathWorks* *R V*	5G: What's Behind the Next Generation of Mobile Communications? *Tabrez Khan, MathWorks*	Demystifying Deep Learning *Dr. Amod Anandkumar, MathWorks*
Exhibition Break			
Developing Optimization Algorithms for Real-World Applications *Dr. Lakshminarayan Ravichandran, MathWorks* *Gautam Ponnappa PC, MathWorks*	Verification and Validation of High-Integrity Systems *Chethan CU, MathWorks* *Vaishnavi H R, MathWorks*	Designing and Integrating Antenna Arrays with Multi-Function Radar Systems *Shashank Kulkarni, Ph.D., MathWorks* *Swathi Balki, MathWorks*	Deploying Deep Neural Networks to Embedded GPUs and CPUs *Rishu Gupta, Ph.D, MathWorks*
Exhibition Break			
Tackling Big Data Using MATLAB *Alka Nair, MathWorks*	Generating Industry Standards Production C Code Using Embedded Coder *Rajat Arora, MathWorks* *Durvesh Kulkarni, MathWorks*	Designing and Verifying Digital and Mixed-Signal Systems *Aniruddha Dayalu, MathWorks*	Developing Algorithms for Robotics and Autonomous Systems *Dhirendra Singh, MathWorks* *Abhisek Roy, MathWorks*

Upcoming Sessions

THANK YOU