MATLAB EXPO 2018

高可扩展的 MATLAB 云端工程数据分析

陈建平， MathWorks 中国

大纲

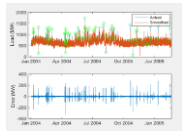
1. 数据访问和探索
2. 数据预处理
3. 开发预测模型
4. 生产服务集成
5. 结果可视化

文件

数据库

传感器

处理大规模数据



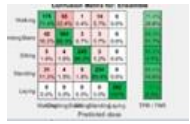
数据归约/变换

特征提取

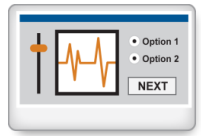
模型构建，机器学习

参数优化

模型验证



桌面 Apps



企业系统应用

kafka

AWS Kinesis

OSIsoft.

嵌入式设备和硬件

第三方面板集成

Spotfire® TIBCO Software

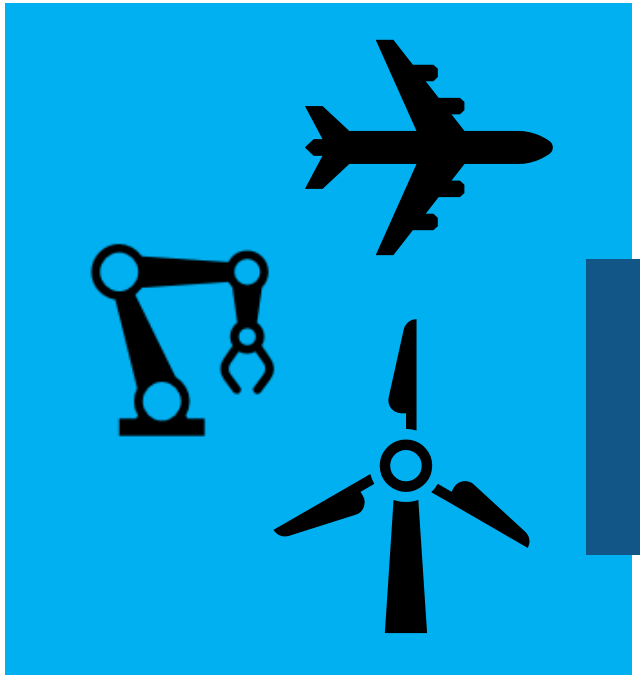
+ a b | e a u

Qlik

Power BI

网络 Apps

大规模流处理的需求



喷气引擎: ~800TB 每天
涡轮引擎: ~ 2TB 每天

预测性维护

增加操作效率

减少计划外的停机

更多的应用需要近乎实时的分析

医疗设备

患者的安全

更加积极的治疗结果

车辆互联

安全

维护

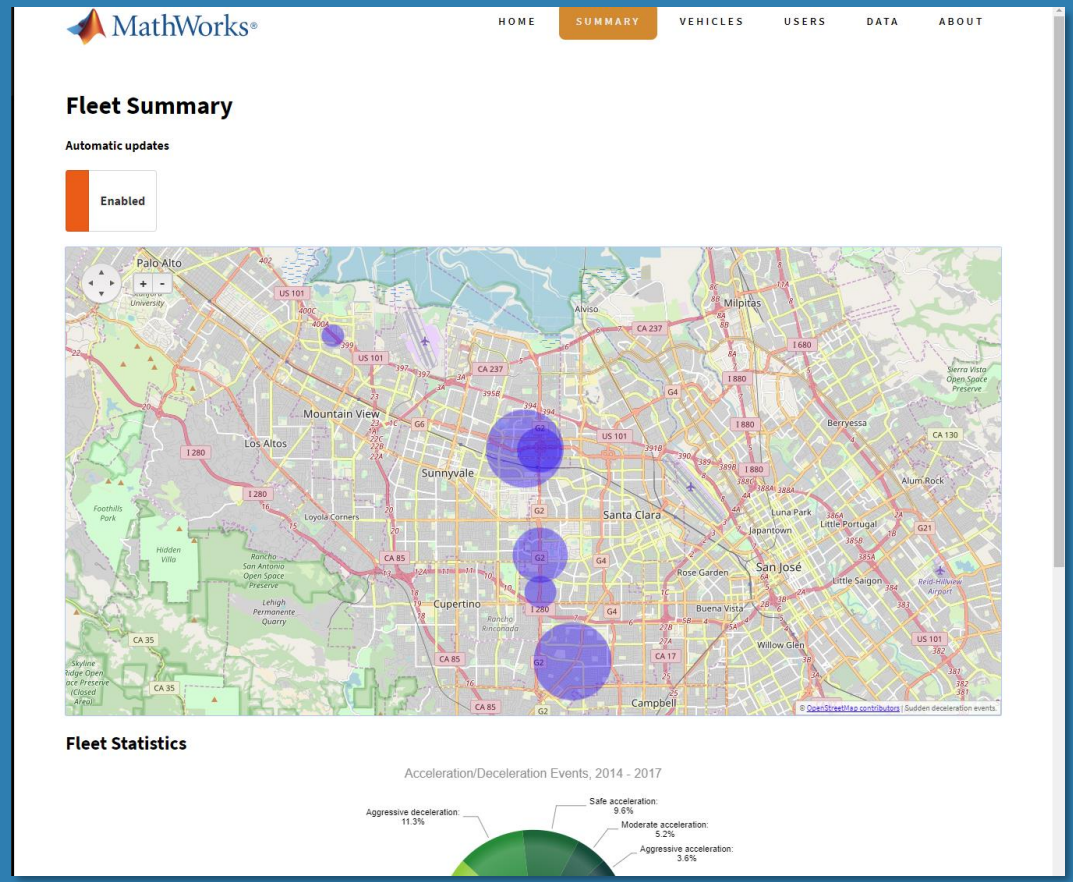
先进的驾驶功能

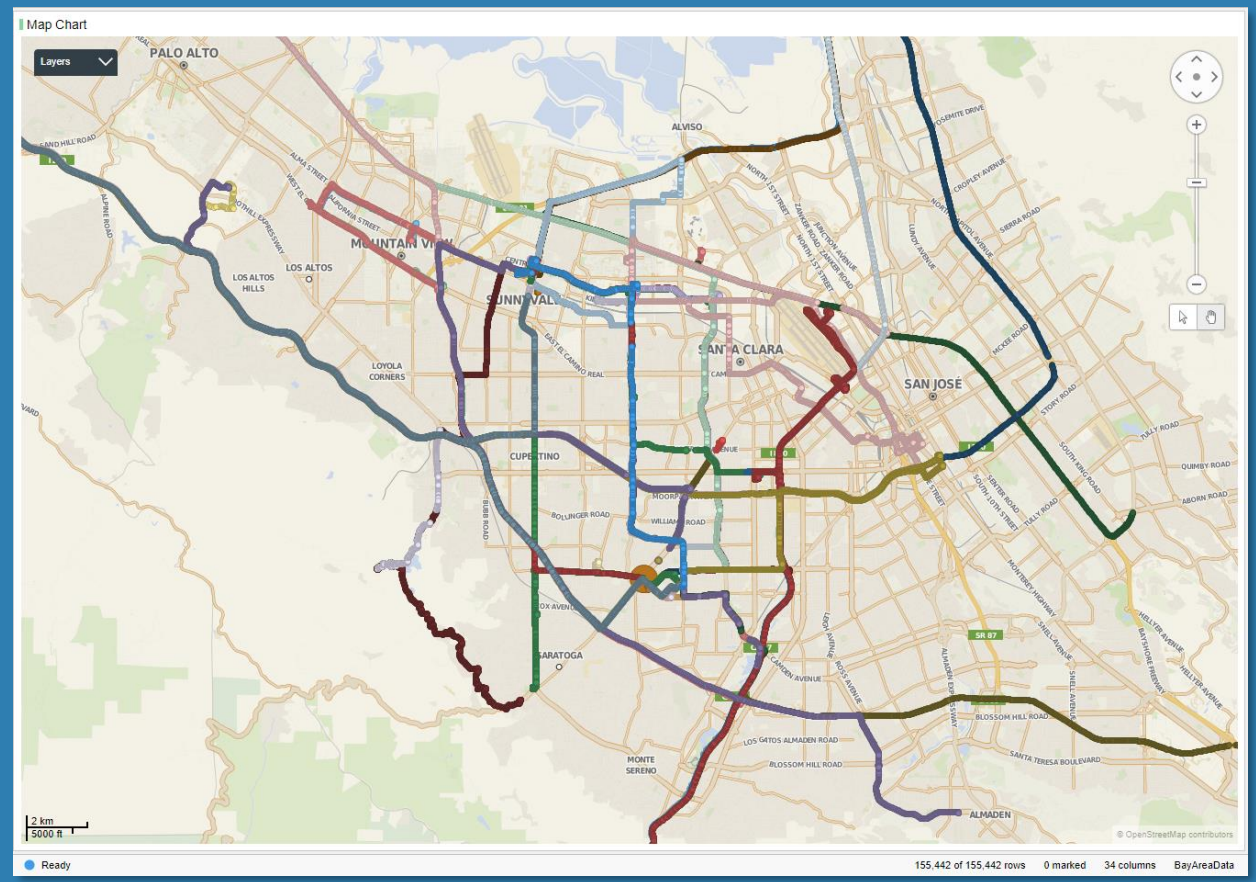
车辆: ~25 GB 每天

实例问题 —— 我的驾驶习惯如何？

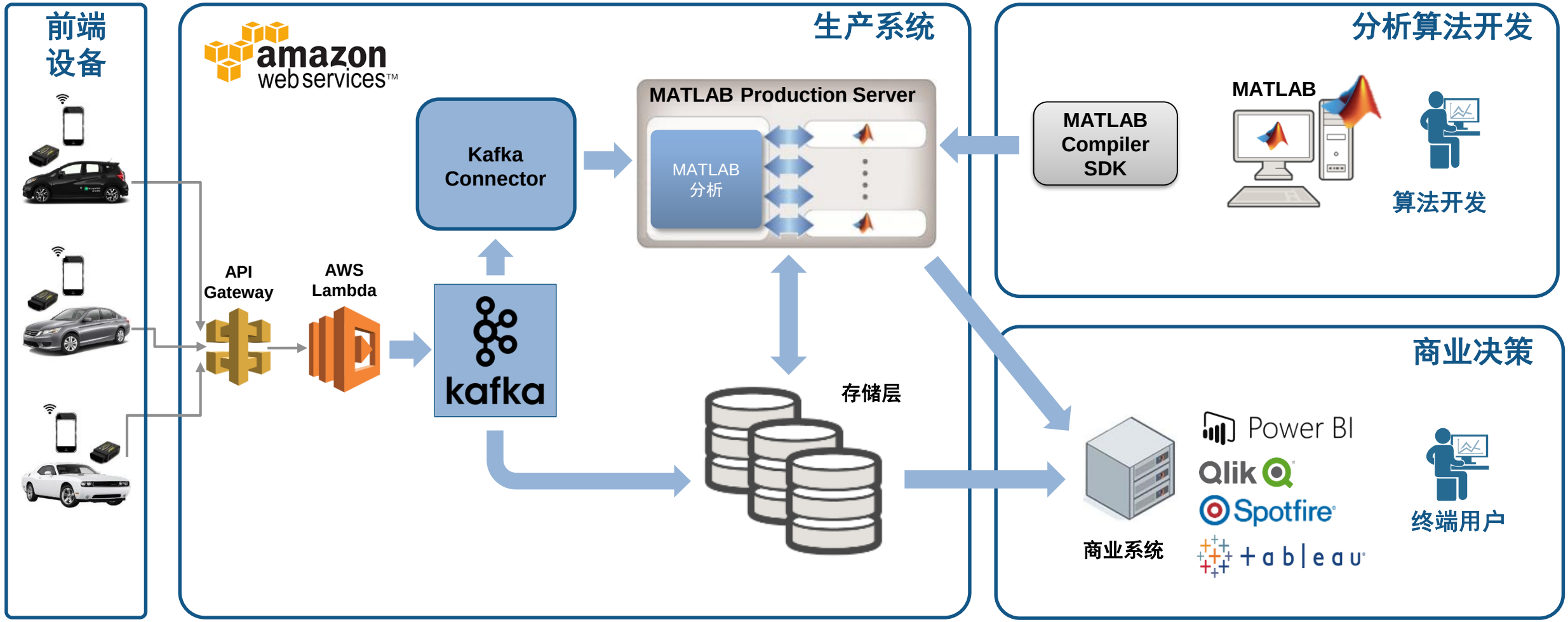
- 一组 MathWorks 员工在车内安装了一个 OBD 狗，用于监控车载系统
- 流式数据传输到云端，汇总并存储
- 我们想用这些数据来评估参与者的驾驶习惯

示例：MATLAB 车队数据分析

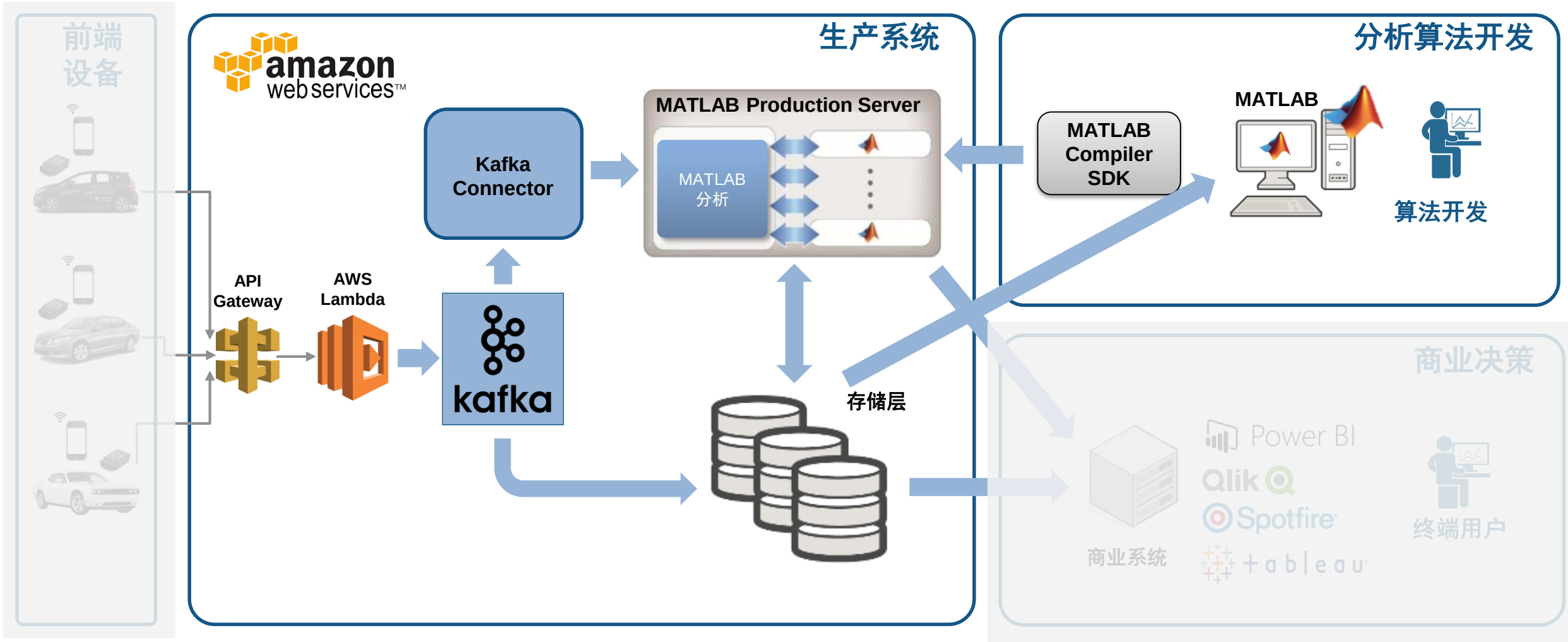




车队数据分析架构


前端
设备
生产系统
amazon
web services™
Kafka
Connector
MATLAB Production Server
MATLAB
分析
API
Gateway
AWS
Lambda
kafka
存储层
分析算法开发
MATLAB
Compiler
SDK
MATLAB
算法开发
商业决策
Power BI
Qlik
Spotfire
tableau
商业系统
终端用户

输入数据清洗


前端
设备
生产系统
amazon
web services™
Kafka
Connector
MATLAB Production Server
MATLAB
分析
API
Gateway
AWS
Lambda
kafka
存储层
分析算法开发
MATLAB
Compiler
SDK
MATLAB
算法开发
商业决策
Power BI
Qlik
Spotfire
tableau
商业系统
终端用户

数据：带时间戳的 JSON 编码消息

```
{
    "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,
      …
}
```

访问数据样点

原始数据

	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'

- ✓ JSON 译码
- ✓ 创建 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

开发一个预处理函数

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					

- ✓ 清洗
- ✓ 增强
- ✓ 重建

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);
```

MATLAB 中随意访问数据

athenaQuery.mlx

Access the data in S3

Bring up the AthenaClient

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

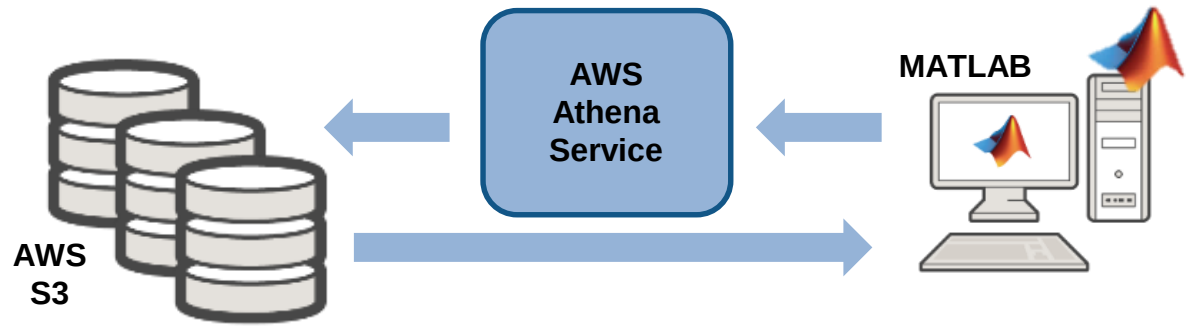
Create a query and submit

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

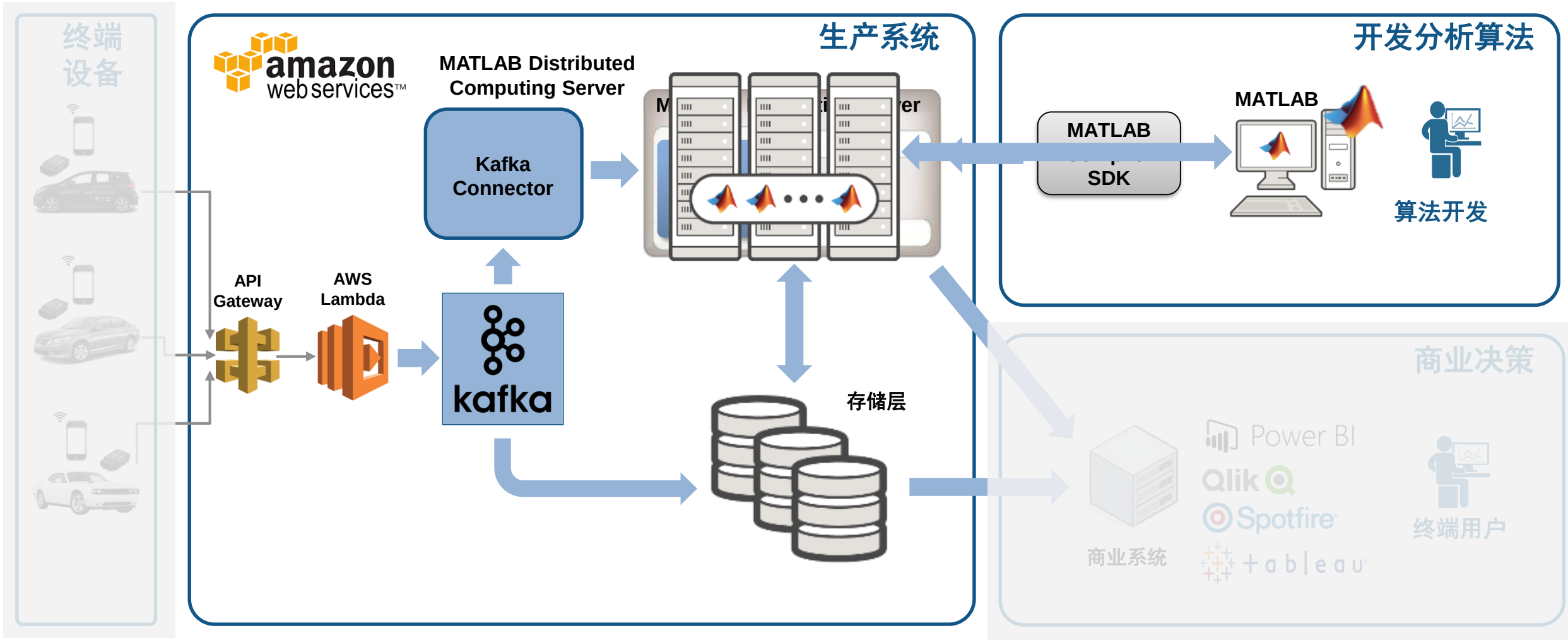
Fetch data as a table for easy analysis

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

Your usual MATLAB workflow goes here



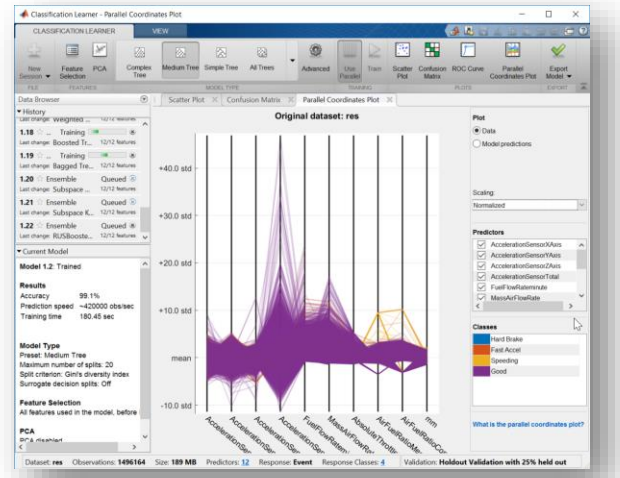
开发预测模型


终端设备
生产系统
amazon web services
MATLAB Distributed Computing Server
Kafka Connector
API Gateway
AWS Lambda
kafka
存储层
开发分析算法
MATLAB
SDK
MATLAB
算法开发
商业决策
商业系统
Power BI
Qlik
Spotfire
tableau
终端用户

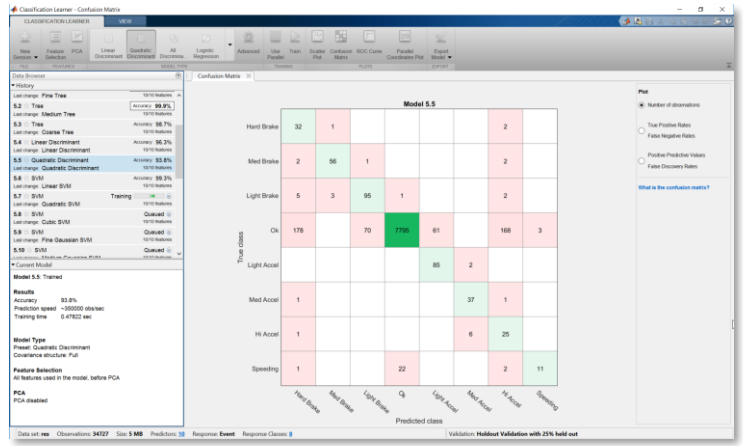
MATLAB 开发预测模型

time	1 Event	2 SpeedGPS	3 AccelerationSensorXAxis	4 AccelerationSensorYAxis	5 AccelerationSensorZAxi
Mon May 11 04:03:15 UTC 2015	Hard Brake	10.8360	-0.6996	0.6014	0.205
Wed May 06 19:09:48 UTC 2015	Hard Brake	27.8280	0.1419	0.9035	-0.526
Sun May 17 17:09:19 UTC 2015	Hard Brake	6.5520	0.9986	-0.0761	-0.004
Fri Jan 16 20:38:37 UTC 2015	Hard Brake	39.6128	0.0999	0.8000	0.367
Sat May 02 14:00:37 UTC 2015	Hard Brake	61.1280	0.4006	-0.4022	0.663
Mon Apr 27 17:54:27 UTC 2015	Fast Accel	37.7640	0.1527	0.4666	0.857
Sun May 03 21:00:42 UTC 2015	Fast Accel	17.2440	1.0235	0.0815	0.304
Mon May 04 11:30:33 UTC 2015	Fast Accel	19.6560	0.1336	0.8932	-0.578

标记事件



信号展示



训练模型

验证模型

```
Evaluating tall expression using the Spark Cluster:
- Pass 1 of 2: Completed in 11 sec
- Pass 2 of 2: Completed in 2.3333 min
Evaluation completed in 2.6167 min
```

扩展

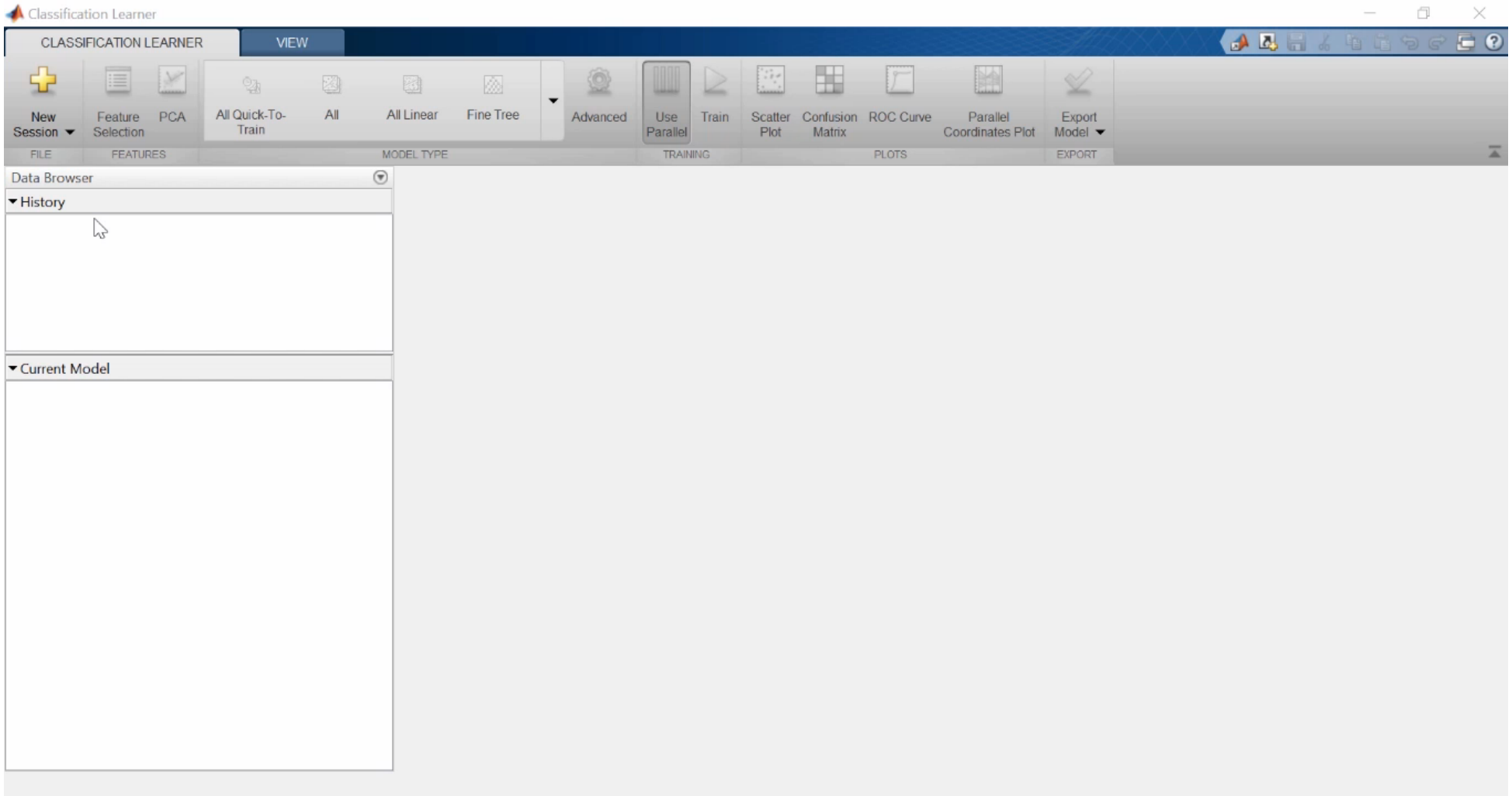
Scale up

```
tt = tall(data);  % test tall array
model = TreeBagger(50,tt,'Event');
```

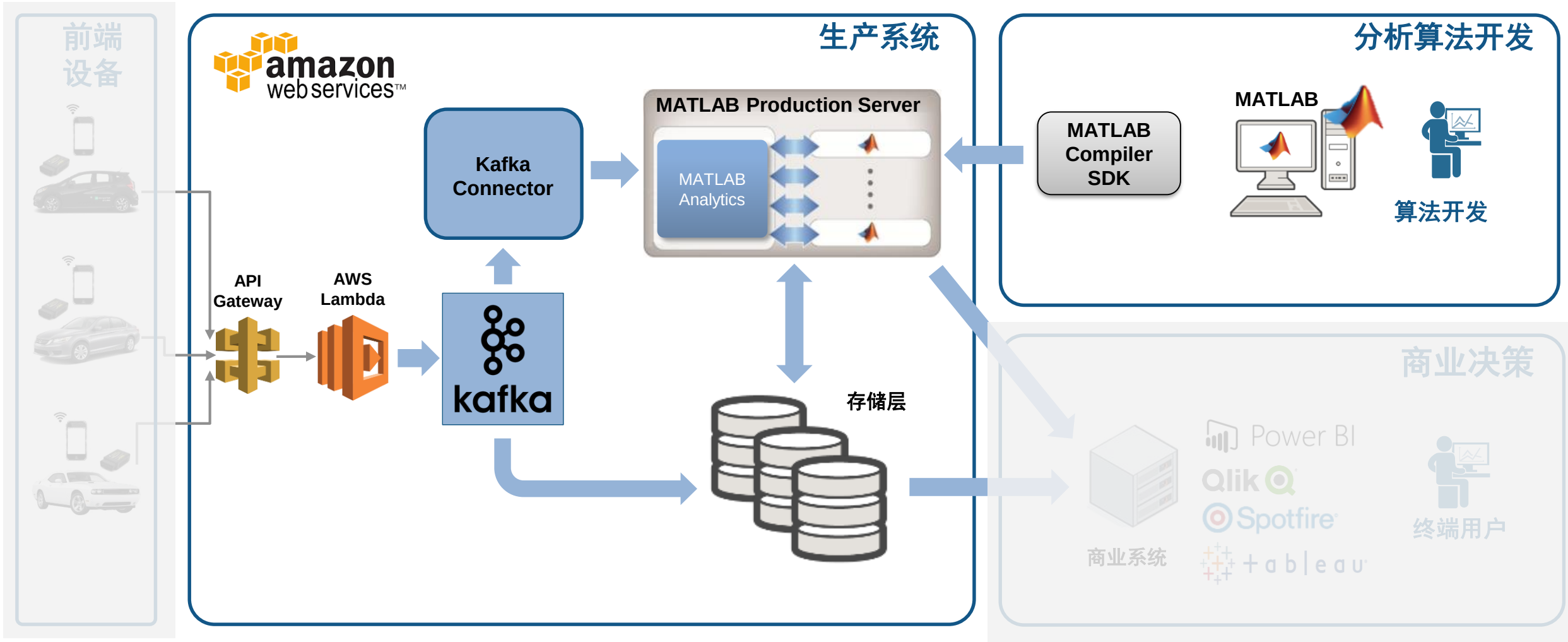
Scale to out of memory data

```
tt = tall(ds);
tt = preprocessData(tt);
model = TreeBagger(50,tt,'Event');
save machineLearningModel model
```

在 MATLAB 中开发预测模型

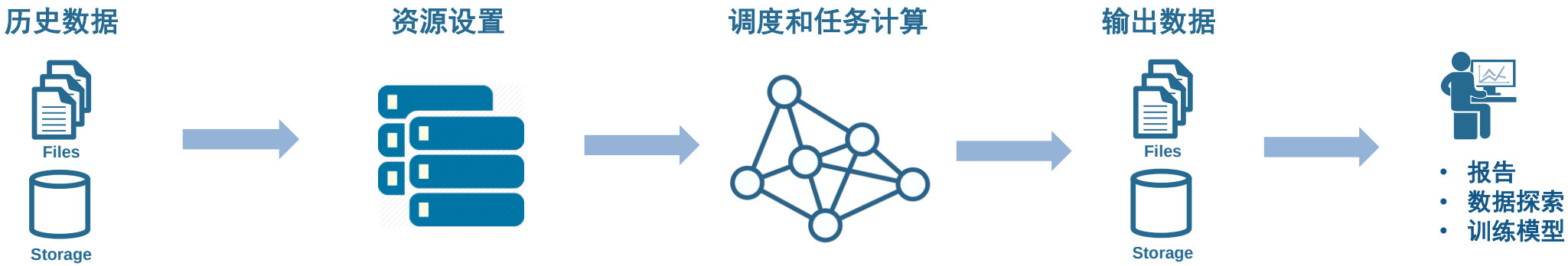


在生产系统中集成分析算法

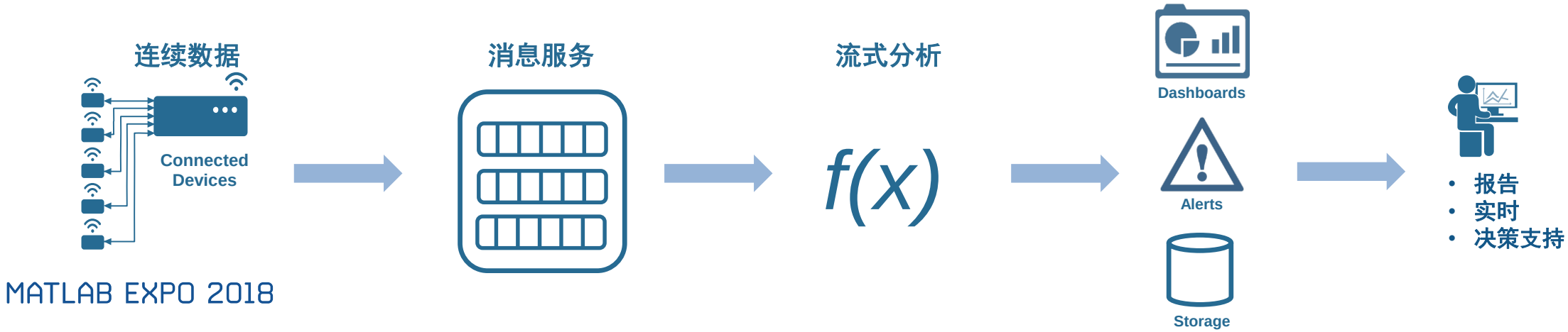


流处理快速入门

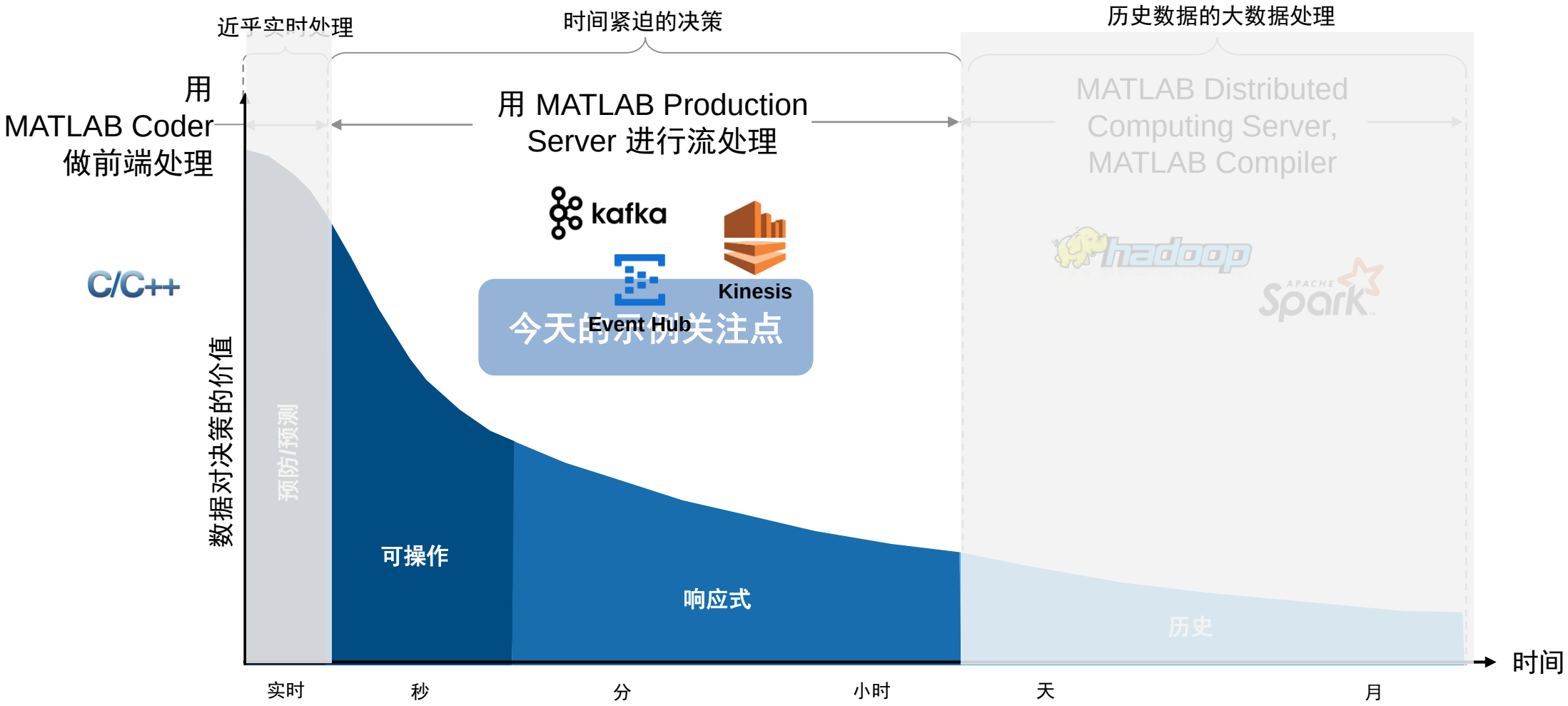
- ***批处理***：应用计算于*过去*采集的*有限大小*的历史数据集



- ***流处理***：应用计算于*连续*产生的*无限*数据集



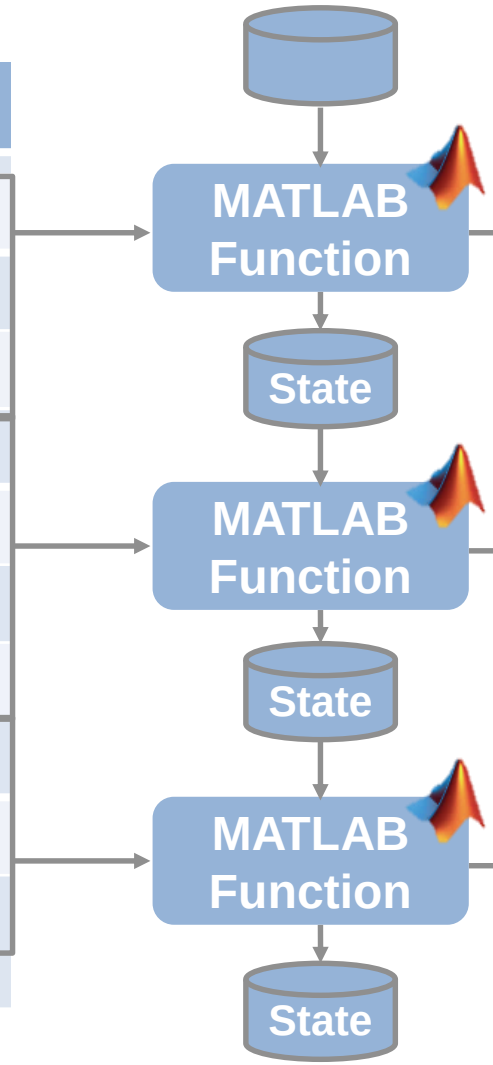
流处理


近乎实时处理
时间紧迫的决策
历史数据的大数据处理
用
MATLAB Coder
做前端处理
用 MATLAB Production
Server 进行流处理
MATLAB Distributed
Computing Server,
MATLAB Compiler
C/C++
kafka
Kinesis
Event Hub
今天的示例关注点
hadoop
Spark
数据对决策的价值
预防/预测
可操作
响应式
历史
时间
实时
秒
分
小时
天
月

流数据被视为无限的时间表

输入表

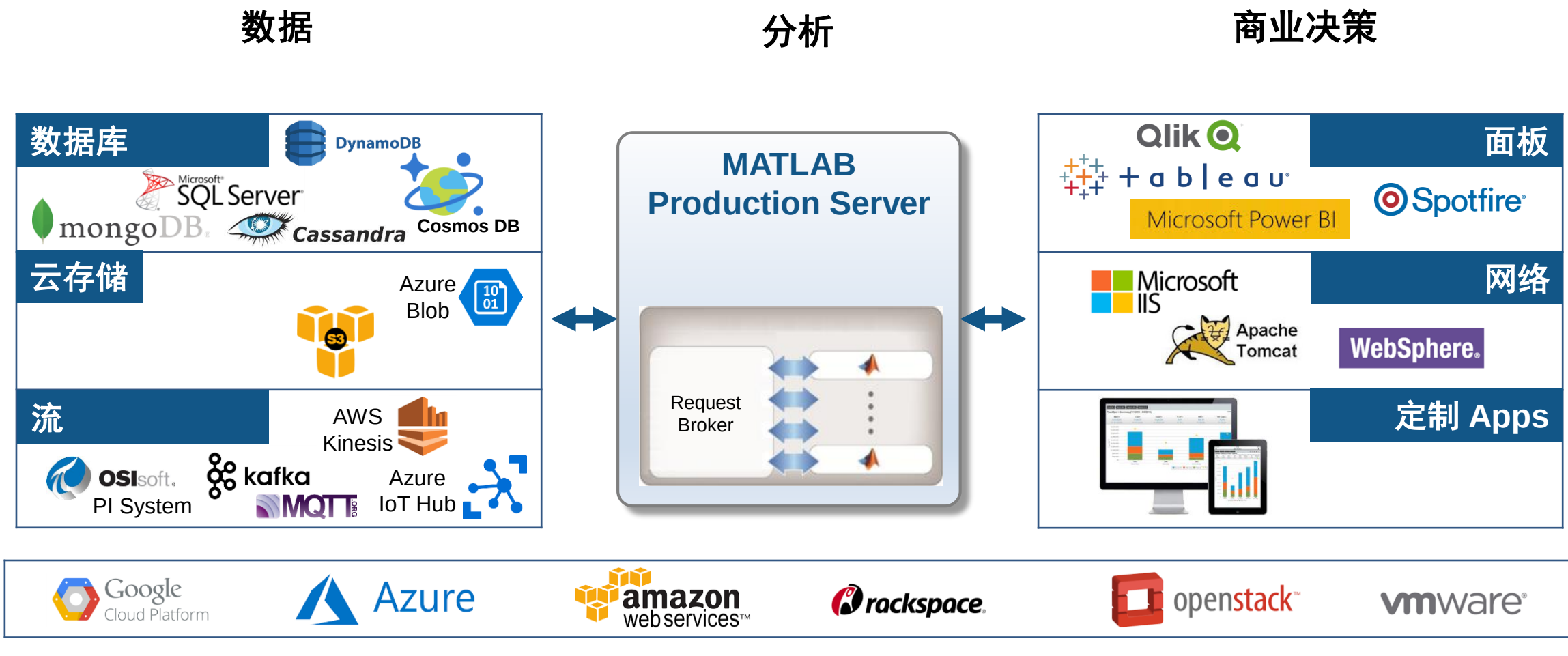
Event Time	Vehicle	RPM	Torque	Fuel Flow
18:01:10	55a3fd	1975	100	110
18:10:30	55a3fe	2000	109	115
18:05:20	55a3fd	1980	105	105
18:10:45	55a3fd	2100	110	100
18:30:10	55a419	2000	100	110
18:35:20	55a419	1960	103	105
18:20:40	55a3fe	1970	112	104
18:39:30	55a419	2100	105	110
18:30:00	55a3fe	1980	110	113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



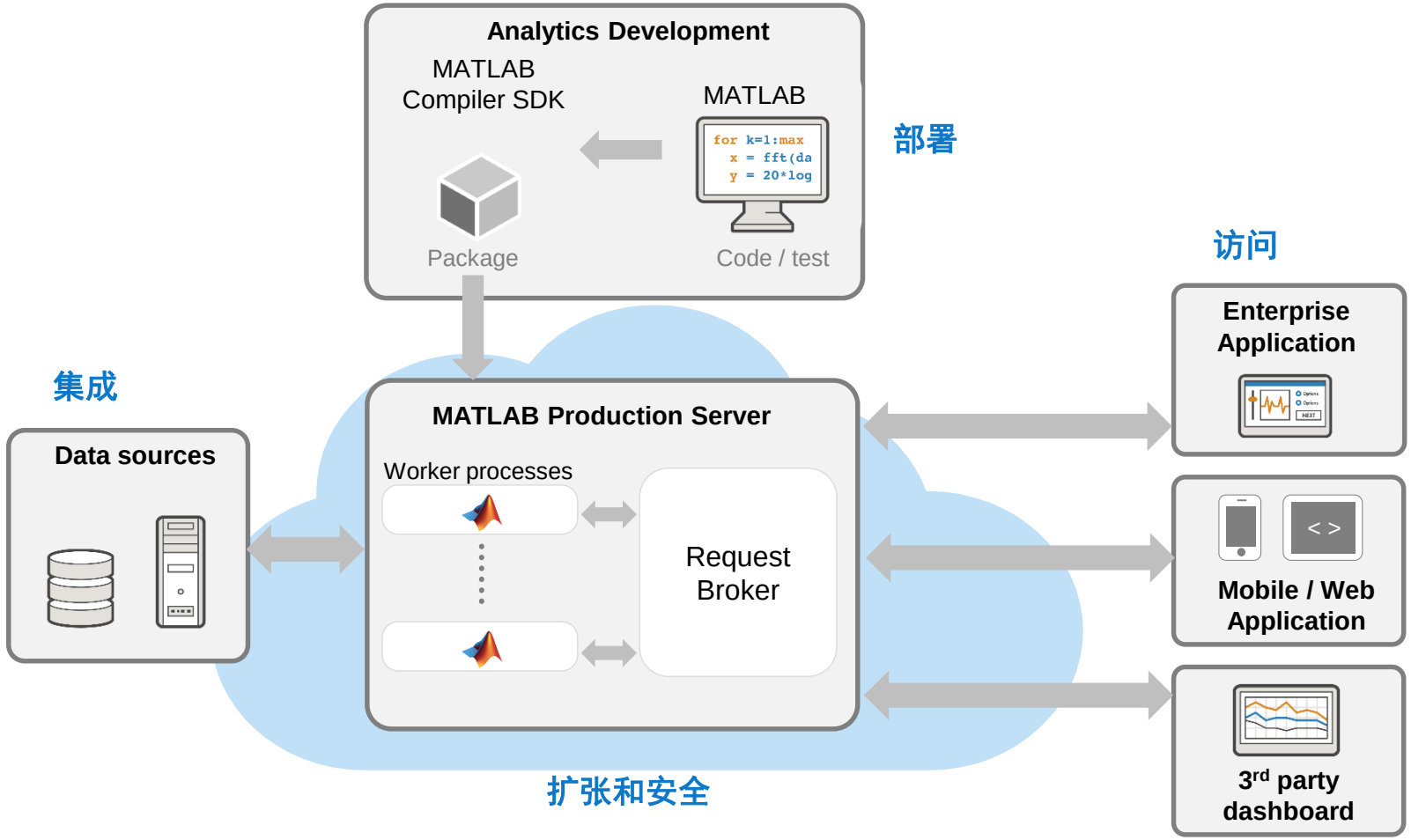
输出表

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

MATLAB Production Server 介绍

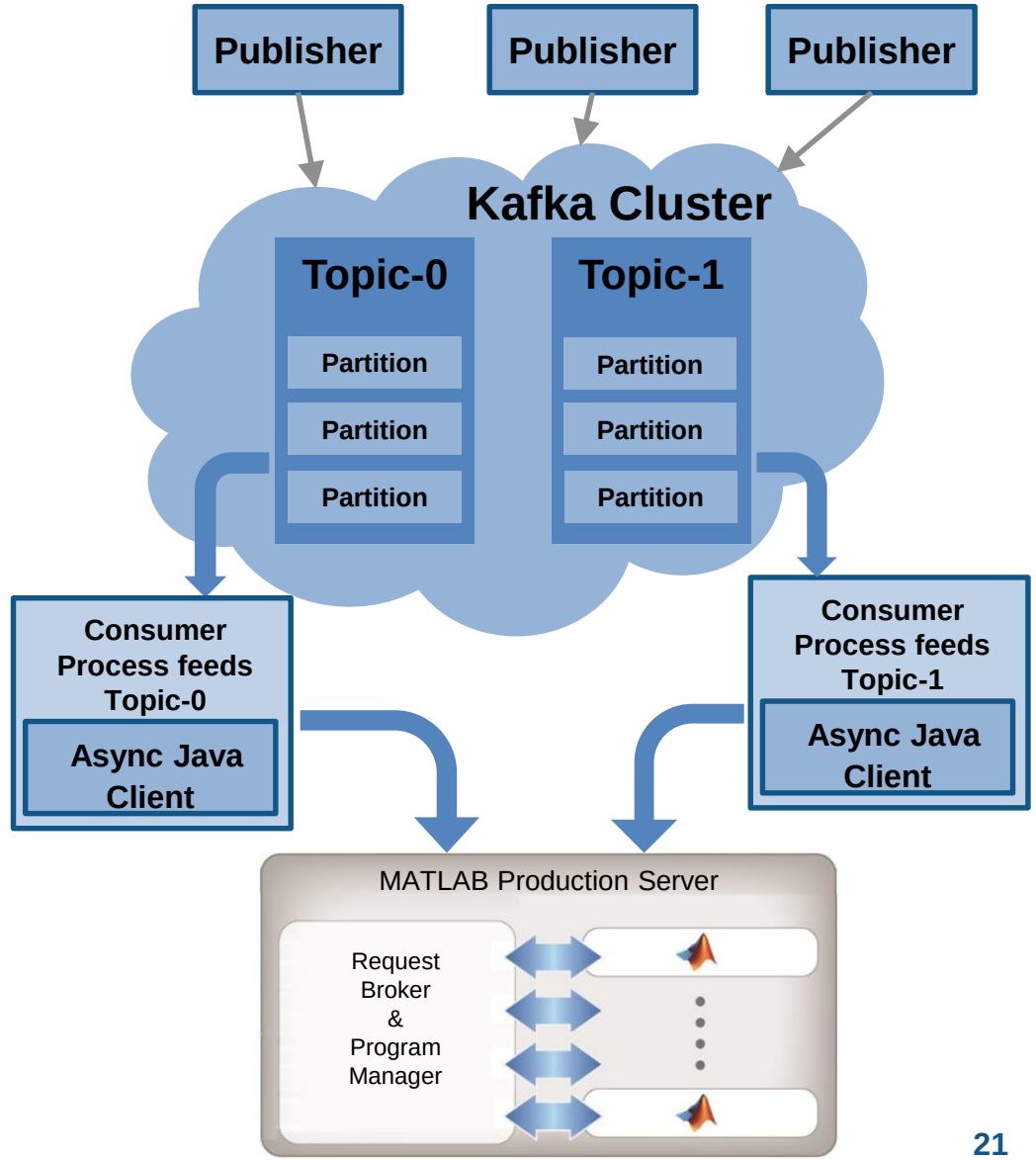


MATLAB Production Server：MATLAB 代码发布成 API 的应用服务器

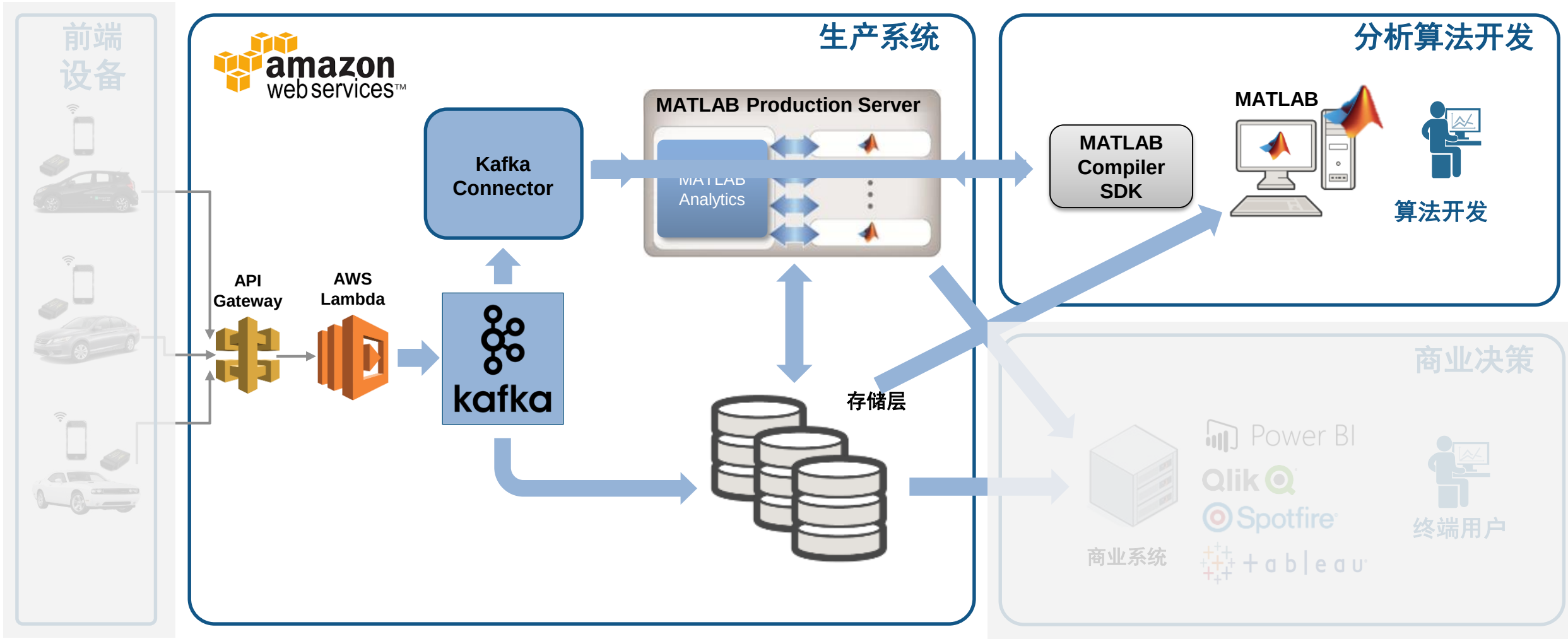


连接 MATLAB Production Server 到 Kafka

- Kafka 客户端把数据发送给部署在服务器中的函数
- 可以把一批消息配置成 MATLAB 时间表传递
- 每个消费者进程把一个主题送给指定的 MATLAB 函数
- 从一个简单的配置程序驱动
  - 除了 MATLAB 无需其他编程



开发和部署流处理函数


前端
设备
生产系统
amazon
web services™
Kafka
Connector
MATLAB Production Server
MATLAB
Analytics
API
Gateway
AWS
Lambda
kafka
存储层
分析算法开发
MATLAB
Compiler
SDK
MATLAB
算法开发
商业决策
Power BI
Qlik
Spotfire
tableau
商业系统
终端用户

在 MATLAB 中开发流处理函数

calculateScores.mlx

Develop a Streaming Function

```
function new_state = calculateScores(car_id, current_data, old_state, resultsStore)
```

Preprocess and perform calculations

```
current_data = preprocessData(current_data);
```

Predict driving events

```
current_data = predictEvents(current_data);
```

Count events for each ten second window

```
countsByTime = countEvents(current_data);
```

Write discrete data to mongodb

```
updateResultsStore(car_id,countsByTime,resultsStore);
```

Update new state

```
new_state = updateState(countsByTime,old_state);
end
```

数据到达时处理每一个数据窗

处理结果

前一个状态

当前待处理数据

在 MATLAB 中开发流处理函数

calculateScores.mlx

Develop a Streaming Function

```
function new_state = calculateScores(car_id, current_data
```

Preprocess and perform calculations

```
current_data = preprocessData(current_data);
```

Predict driving events

```
current_data = predictEvents(current_data);
```

Count events for each ten second window

```
countsByTime = countEvents(current_data);
```

Write discrete data to mongodb

```
updateResultsStore(car_id,countsByTime,resultsStore);
```

Update new state

```
new_state = updateState(countsByTime,old_state);
end
```

```
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)];
```

应用预处理算法

在 MATLAB 中开发流处理函数

calculateScores.mlx

Develop a Streaming Function

```
function new_state = calculateScores(car_id, current_data, old_state, resultsStore)
```

Preprocess and perform calculations

```
current_data = preprocessData(current_data);
```

Predict driving events

```
current_data = predictEvents(current_data);
```

Count events for each ten second window

```
countsByTime = countEvents(current_data);
```

Write discrete data to mongodb

```
updateResultsStore(car_id,countsByTime,resultsStore);
```

Update new state

```
new_state = updateState(countsByTime,old_state);
end
```

使用分类学习 App 习得的模型

```
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'};
predictors = current_data{:,predictorNames};
mdl = load('machineLearningModel.mat');
current_data.Event = predict(mdl.model,predictors);

end
```

在 MATLAB 中开发流处理函数

calculateScores.mlx

Develop a Streaming Function

```
function new_state = calculateScores(car_id, current_data, old_state, resultsStore)
```

Preprocess and perform calculations

```
current_data = preprocessData(current_data);
```

Predict driving events

```
current_data = predictEvents(current_data);
```

Count events for each ten second window

```
countsByTime = countEvents(current_data);
```

Write discrete data to mongodb

```
updateResultsStore(car_id,countsByTime,resultsStore);
```

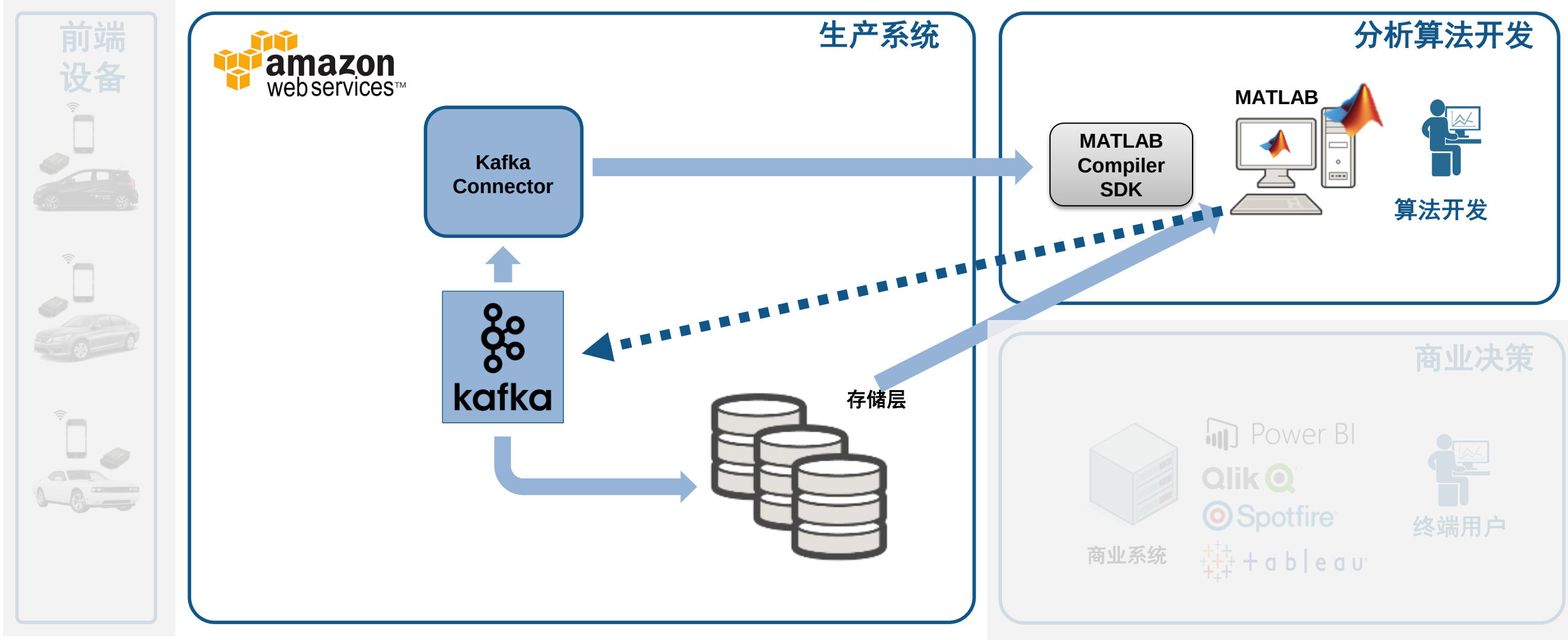
Update new state

```
new_state = updateState(countsByTime,old_state);
end
```

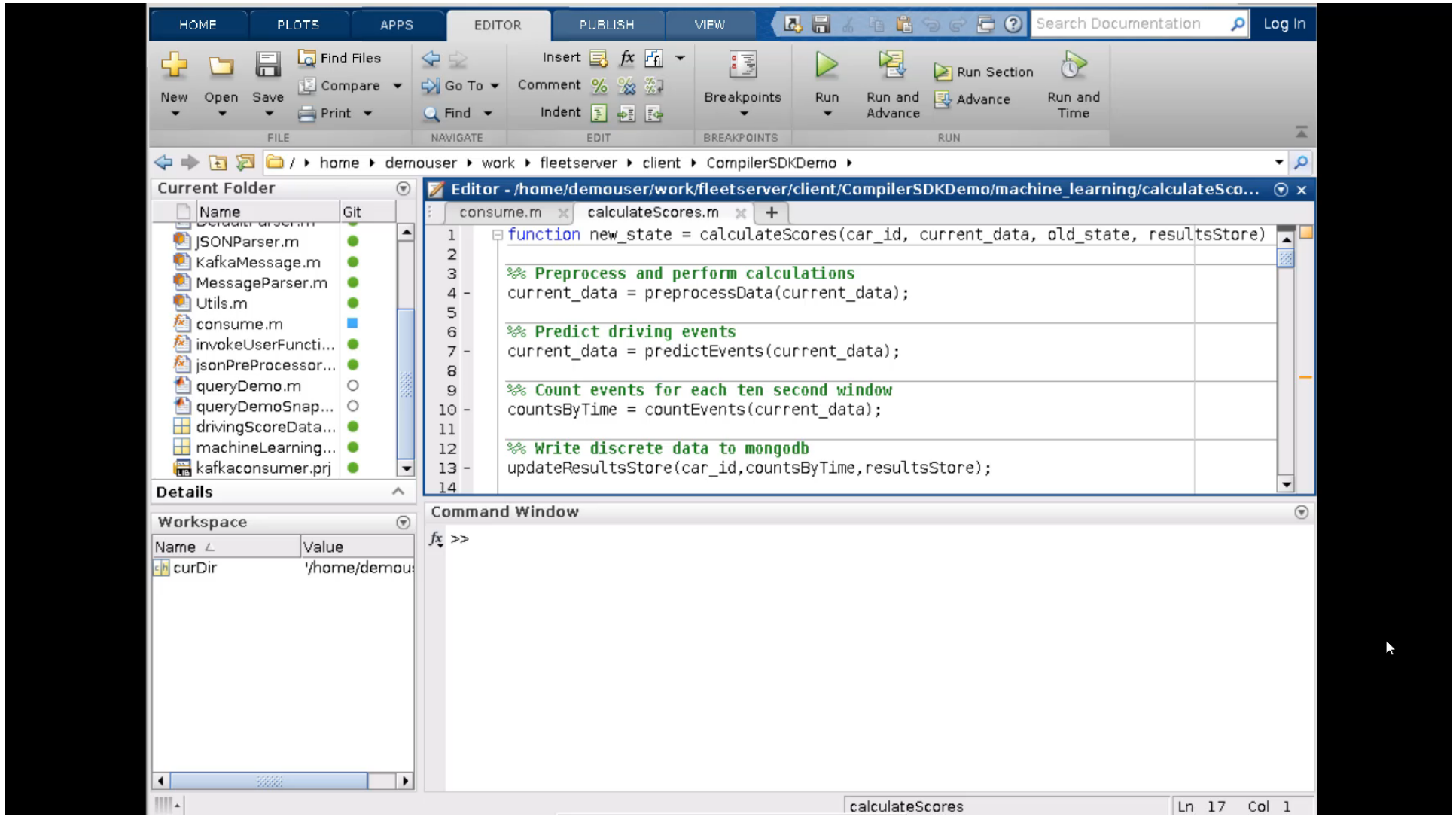
更新 Mongo 数据库

- 针对类型和位置统计事件
- 驾驶习惯估计

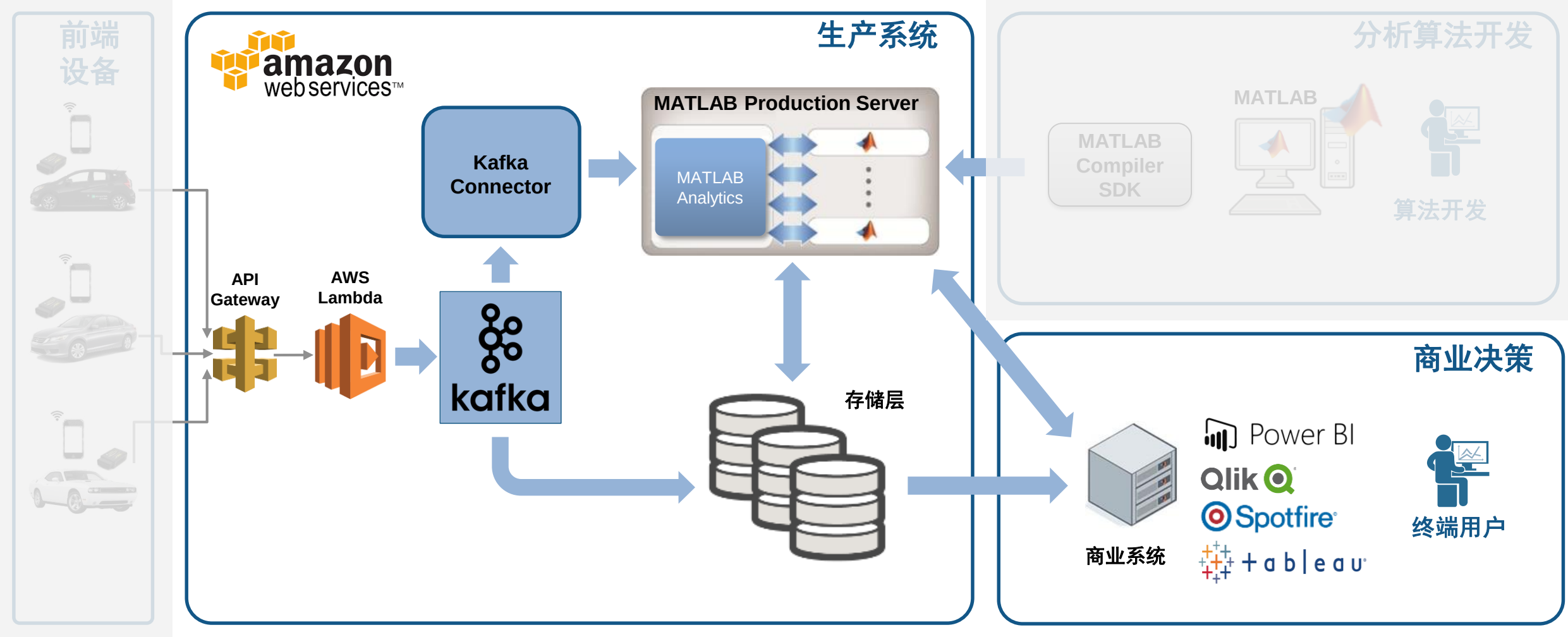
在 MATLAB 中调试流处理函数



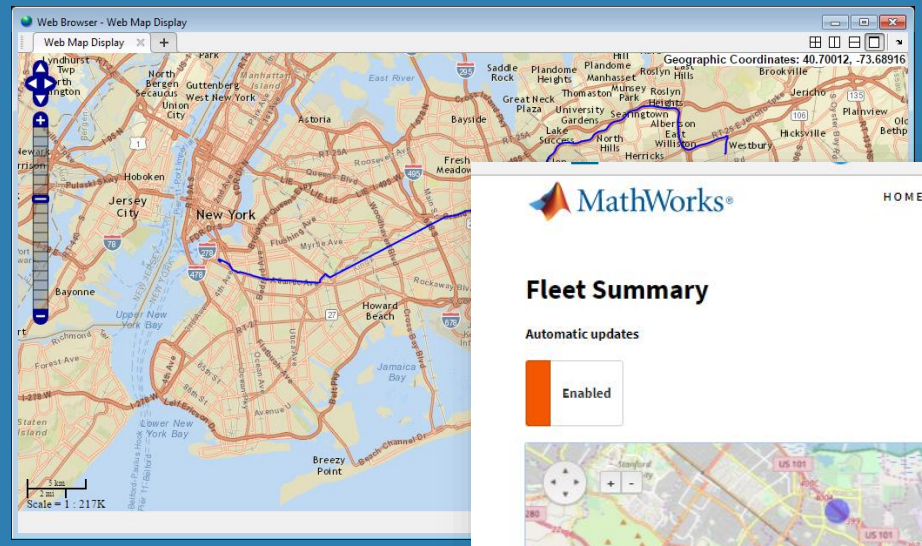
在 MATLAB 中调试流处理函数

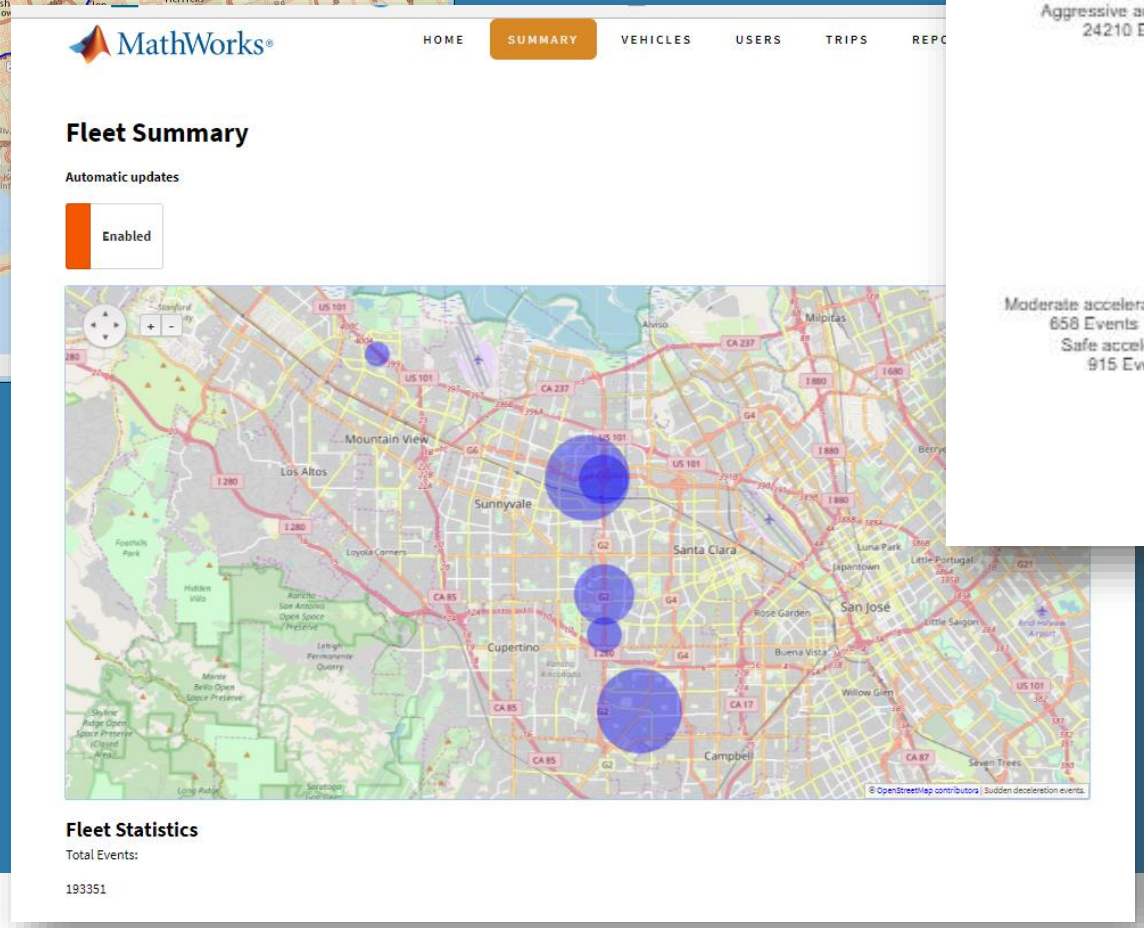


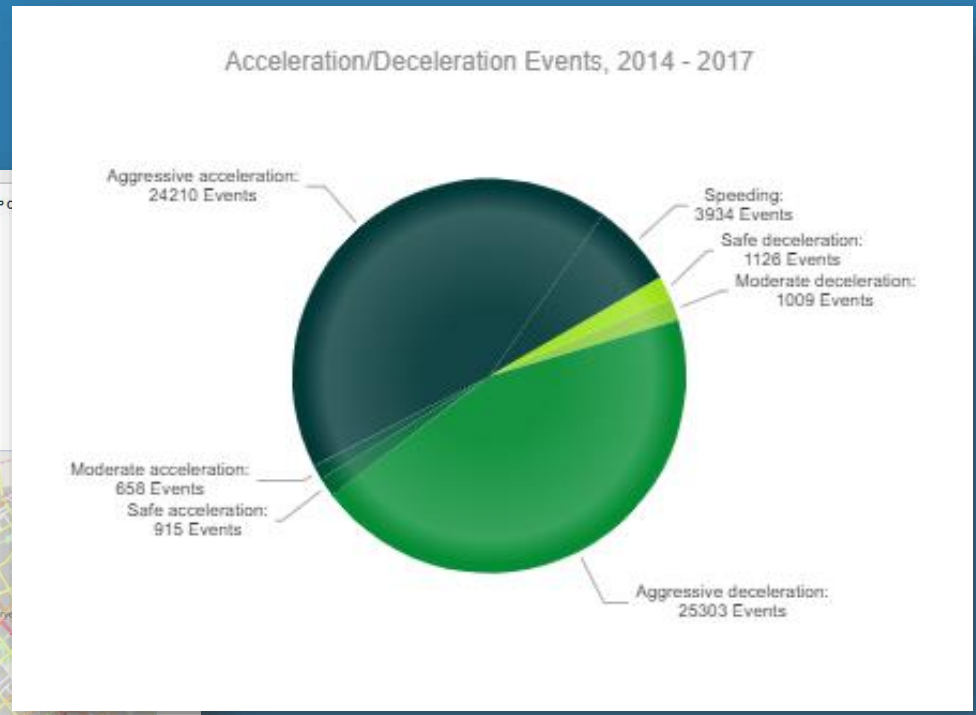
绑定您的仪表盘应用


前端
设备
生产系统
amazon
web services™
Kafka
Connector
MATLAB Production Server
MATLAB
Analytics
API
Gateway
AWS
Lambda
kafka
存储层
分析算法开发
MATLAB
MATLAB
Compiler
SDK
算法开发
商业决策
Power BI
Qlik
Spotfire
tableau
商业系统
终端用户

完成您的应用

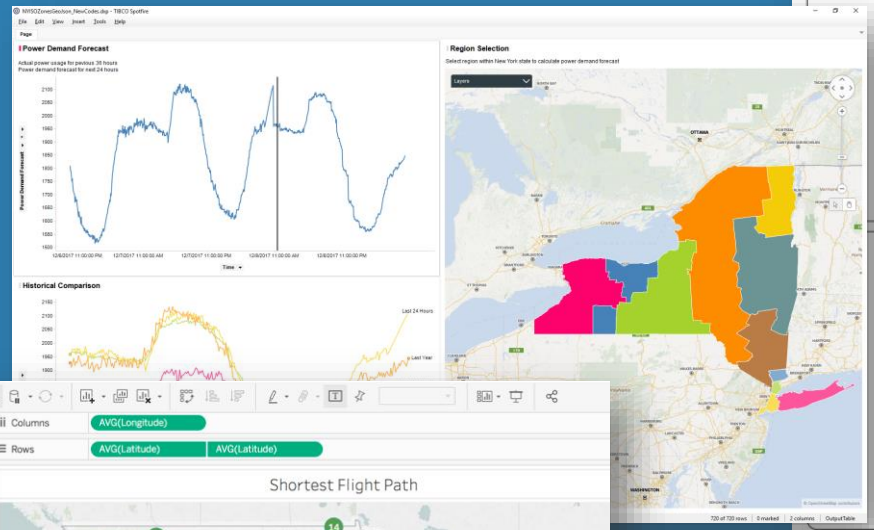

Web Browser - Web Map Display
Web Map Display
Geographic Coordinates: 40.70012, -73.68916
New York
Jersey City
Hoboken
Bayonne
Howard Beach
Breezy Point
Scale = 1 : 217K


MathWorks
HOME
SUMMARY
VEHICLES
USERS
TRIPS
Fleet Summary
Automatic updates
Enabled
Mountain View
Los Altos
Sunnyvale
Santa Clara
San José
Cupertino
Campbell
Fleet Statistics
Total Events:
193351

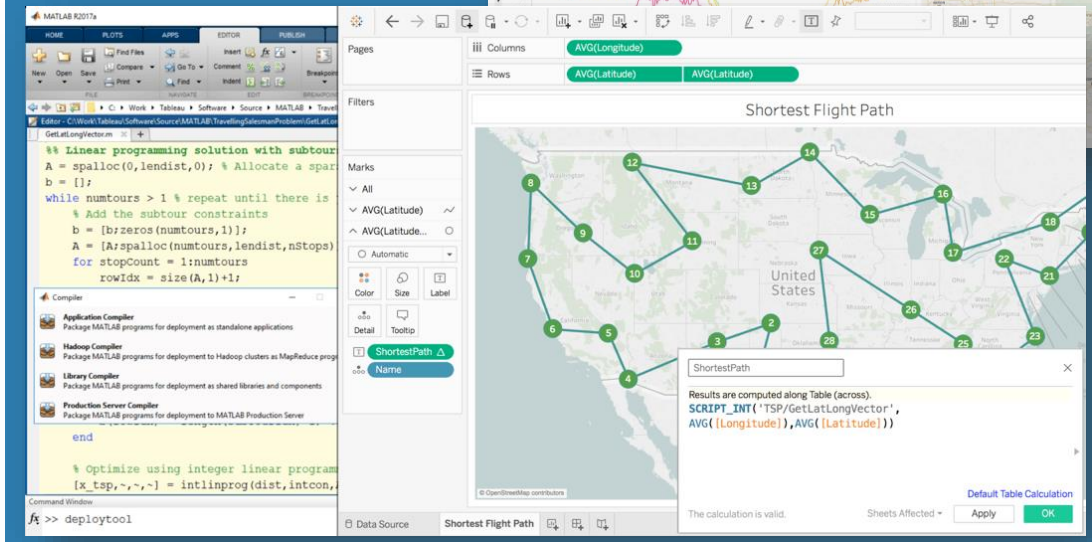

Acceleration/Deceleration Events, 2014 - 2017
Aggressive acceleration:
24210 Events
Speeding:
3934 Events
Safe deceleration:
1126 Events
Moderate deceleration:
1009 Events
Moderate acceleration:
658 Events
Safe acceleration:
915 Events
Aggressive deceleration:
25303 Events

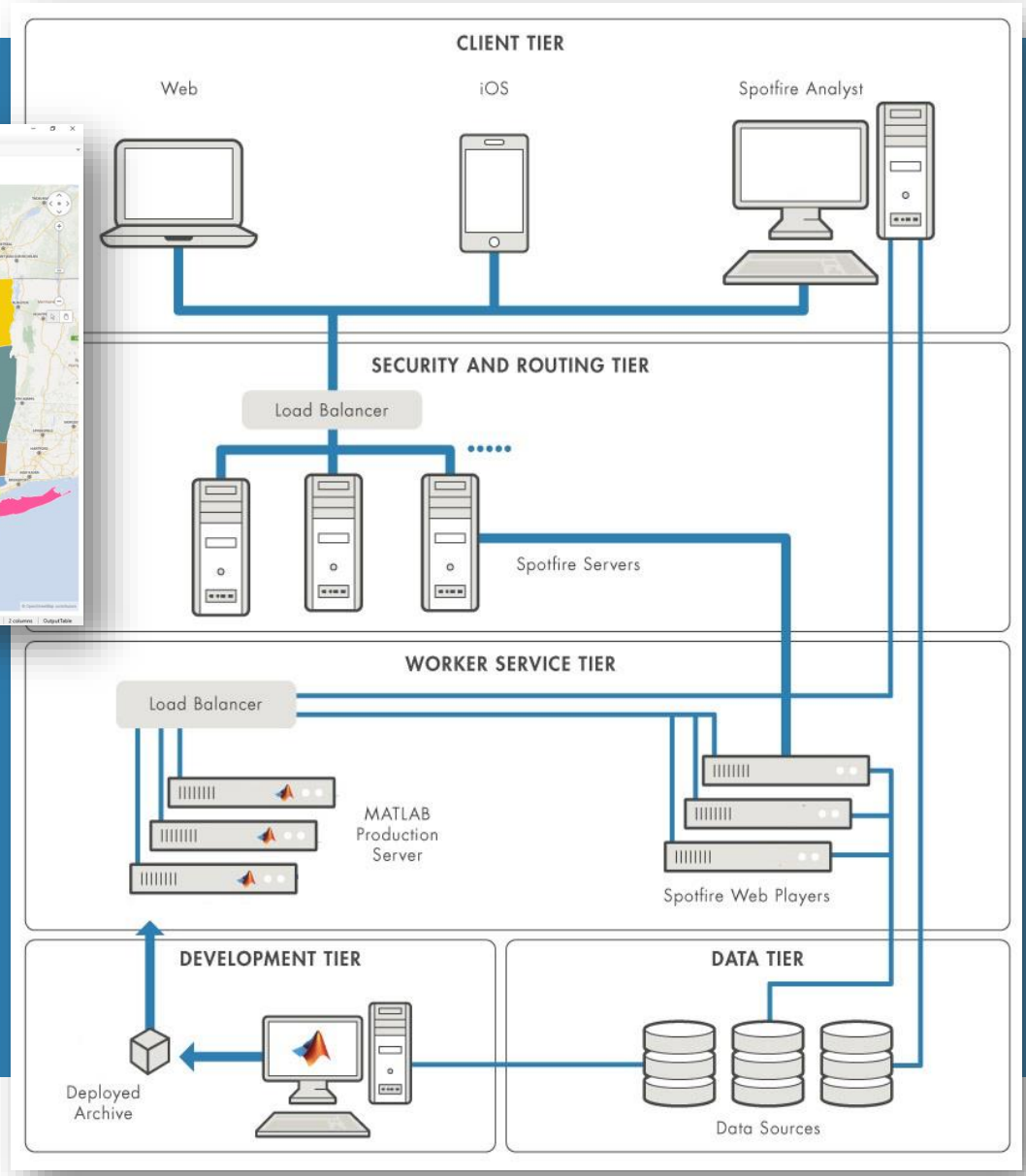
协同企业 BI 工具的扩展分析

TIBCO Spotfire



Tableau





要点回顾

- MATLAB 直接连接到您的数据之中，方便快速设计和验证算法
- MATLAB 语言和 apps 支持快速设计迭代
- MATLAB 产品服务器可以轻松将算法与企业生产系统集成
- 您的时间将聚焦于数据理解和算法设计

学习资源和快速起步

- Data Analytics with MATLAB
- MATLAB Production Server
- MATLAB Compiler SDK
- Statistics and Machine Learning Toolbox
- Database Toolbox
- Mapping Toolbox
- MATLAB with TIBCO Spotfire
- MATLAB with Tableau
- MATLAB with MongoDB