MATLAB EXPO 2018

Scaling up MATLAB Analytics with Kafka and Cloud Services

Olof Larsson
Agenda

1. Access and Explore Data
   - Files
   - Databases
   - Sensors

2. Preprocess Data
   - Working with Messy Data
   - Data Reduction/Transformation
   - Feature Extraction

3. Develop Predictive Models
   - Model Creation e.g. Machine Learning
   - Parameter Optimization
   - Model Validation

4. Integrate with Production Systems
   - Desktop Apps
   - Enterprise Scale Systems
     - kafka
     - AWS Kinesis
   - Embedded Devices and Hardware

5. Visualize Results
   - 3rd party dashboards
   - Web apps
The Need for Large-Scale Streaming

Predictive Maintenance
*Increase Operational Efficiency*
*Reduce Unplanned Downtime*

More applications require near real-time analytics

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

Medical Devices
*Patient Safety*
*Better Treatment Outcomes*

Connected Cars
*Safety, Maintenance*
*Advanced Driving Features*

Car: ~25 GB per hour

MATLAB EXPO 2018
Example Problem – How’s my driving?

- 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

- We would like to use this data to score the driving habits of participants
Example: Fleet Analytics with MATLAB
Fleet Analytics Architecture

Edge Devices

API Gateway
AWS Lambda

Production System

Kafka Connector

MATLAB Production Server

MATLAB Analytics

Storage Layer

Analytics Development

MATLAB Compiler SDK

Algorithm Developers

Business Decisions

Power BI
Qlik
Spotfire
Tableau

Business Systems

End Users

MATLAB EXPO 2018
The first step is to clean up the incoming data.

Access and Explore Data

1. Edge Devices
   - API Gateway
   - AWS Lambda

2. Kafka
   - Kafka Connector

Production System
   - MATLAB Production Server
     - MATLAB Analytics
     - Storage Layer

3. Analytics Development
   - MATLAB Compiler SDK
     - MATLAB
     - Algorithm Developers

Business Decisions
   - Power BI
   - Qlik
   - Tableau

End Users
   - Business Systems
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

```
{
    "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

1. Access and Explore Data

Raw Data

- Decode JSON data
- Create Timetable

Timetable

MATLAB EXPO 2018
Develop a Preprocessing Function

Preprocess Data

Timetable

```matlab
% 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);
```
Develop a Predictive Model

Production System

MATLAB Distributed Computing Server

Analytics Development

Algorithm Developers

Business Decisions

Power BI
Qlik
Tableau
End Users

API Gateway
AWS Lambda
kafka
Storage Layer
Everything you need to develop a predictive model is found in MATLAB

3 Develop Predictive Models

Label Events

Represent Signals

Train Model

Validate Model

MATLAB EXPO 2018

Evaluate 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

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
```
Develop a Predictive Model in MATLAB
Integrate Analytics with Production Systems

Production System

Analytics Development

Integrate with Production Systems

Edge Devices

AWS Lambda

API Gateway

kafka

Kafka Connector

MATLAB Production Server

MATLAB Compiler SDK

Algorithm Developers

Business Decisions

End Users

Business Systems

MATLAB

Power BI

Qlik

Spotfire

tableau

MATLAB EXPO 2018
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?

Stream Processing with MATLAB Production Server

Edge Processing with MATLAB Coder

Near Real time decisions

Time critical decisions

Big Data processing on historical data

Value of data to decision making

Preventive / Predictive

Actionable

Reactive

Real-Time

Seconds

Minutes

Hours

Days

Months

Time

Today's example focuses here

MATLAB Distributed Computing Server, MATLAB Compiler

Integrate with Production Systems

Kinesis Event Hub

Hadoop

Spark

MATLAB EXPO 2018
### Input Table

<table>
<thead>
<tr>
<th>Event Time</th>
<th>Vehicle</th>
<th>RPM</th>
<th>Torque</th>
<th>Fuel Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>18:01:10</td>
<td>55a3fd</td>
<td>1975</td>
<td>100</td>
<td>110</td>
</tr>
<tr>
<td>18:10:30</td>
<td>55a3fe</td>
<td>2000</td>
<td>109</td>
<td>115</td>
</tr>
<tr>
<td>18:05:20</td>
<td>55a3fd</td>
<td>1980</td>
<td>105</td>
<td>105</td>
</tr>
<tr>
<td>18:10:45</td>
<td>55a3fd</td>
<td>2100</td>
<td>110</td>
<td>100</td>
</tr>
<tr>
<td>18:30:10</td>
<td>55a419</td>
<td>2000</td>
<td>100</td>
<td>110</td>
</tr>
<tr>
<td>18:35:20</td>
<td>55a419</td>
<td>1960</td>
<td>103</td>
<td>105</td>
</tr>
<tr>
<td>18:20:40</td>
<td>55a3fe</td>
<td>1970</td>
<td>112</td>
<td>104</td>
</tr>
<tr>
<td>18:39:30</td>
<td>55a419</td>
<td>2100</td>
<td>105</td>
<td>110</td>
</tr>
<tr>
<td>18:30:00</td>
<td>55a3fe</td>
<td>1980</td>
<td>110</td>
<td>113</td>
</tr>
<tr>
<td>18:30:50</td>
<td>55a3fe</td>
<td>2000</td>
<td>100</td>
<td>110</td>
</tr>
</tbody>
</table>

---

### Output Table

<table>
<thead>
<tr>
<th>Time window</th>
<th>Vehicle</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>18:00:00</td>
<td>55a3fd</td>
<td>5</td>
</tr>
<tr>
<td>18:10:00</td>
<td>55a3fe</td>
<td>...</td>
</tr>
<tr>
<td>18:20:00</td>
<td>55a419</td>
<td>...</td>
</tr>
<tr>
<td>18:10:00</td>
<td>55a3fd</td>
<td>7</td>
</tr>
<tr>
<td>18:20:00</td>
<td>55a3fe</td>
<td>3</td>
</tr>
<tr>
<td>18:20:00</td>
<td>55a419</td>
<td>...</td>
</tr>
<tr>
<td>18:30:00</td>
<td>55a3fe</td>
<td>4</td>
</tr>
<tr>
<td>18:30:00</td>
<td>55a419</td>
<td>...</td>
</tr>
</tbody>
</table>

---

Streaming data is treated as an unbounded Timetable.
Introducing MATLAB Production Server

Data

- Databases
  - DynamoDB
  - SQL Server
  - Cassandra
  - Cosmos DB

Cloud Storage

- Azure Blob

Streaming

- AWS Kinesis
- Azure IoT Hub

Analytics

MATLAB Production Server

Business System

- Dashboards
  - Qlik Sense
  - Tableau
  - Microsoft Power BI
  - Spotfire

Web

- Microsoft IIS
- Apache Tomcat
- WebSphere

Custom Apps

- Google Cloud Platform
- Azure
- Amazon Web Services
- Rackspace
- OpenStack
- VMware

Platform

MATLAB EXPO 2018
MATLAB Production Server is an application server that publishes MATLAB code as APIs
Connecting MATLAB Production Server to Kafka

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

Integrate with Production Systems

Edge Devices

Production System

Analytics Development

Business Decisions

Power BI

Tableau

End Users

API Gateway

AWS Lambda

Kafka Connector

MATLAB Production Server

MATLAB Compiler SDK

MATLAB

Algorithm Developers

Business Systems

kafka

Storage Layer

Production System

4
Develop a Stream Processing Function in MATLAB

Develop a Streaming Function

```matlab
function new_state = calculateScore(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
```

Current window of data to be processed

Current score

Previous state

Process each window of data as it arrives
Develop a Streaming Function

```
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)];
end
```

Apply your pre-processing algorithm
Develop a Stream Processing Function in MATLAB

Use the model you created with Classification Learner App

MATLAB EXPO 2018
Develop a Stream Processing Function in MATLAB

```matlab
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
```

Update Mongo database
- Count of events by type and location
- Results of driver scoring

MATLAB EXPO 2018
Debug a Stream Processing Function in MATLAB

Integrate with Production Systems

Edge Devices

Production System

Analytics Development

Kafka Connector

Storage Layer

MATLAB Compiler SDK

Algorithm Developers

Business Decisions

MATLAB EXPO 2018
Integrate with Production Systems

Debug a Stream Processing Function in MATLAB
Tie in your Dashboard Application
Visualize Results

Complete Your Application
Scalable Analytics with Enterprise BI Tools

TIBCO Spotfire

Tableau

MATLAB EXPO 2018
Key Takeaways

➢ MATLAB connects directly to your data so you can quickly design and validate algorithms

➢ The MATLAB language and apps enable fast design iterations

➢ MATLAB Production Server enables easy integration of your MATLAB algorithms with enterprise production systems

➢ Allows you to spend your time understanding the data and designing algorithms
Resources to learn and get started

- 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