Realizing the value of combining IoT and big data analytics

Eliano Marques, Global Data Science practice lead, Think Big Analytics
Dr Frank Säuberlich, Director Data Science, Teradata International
Big Data Analytics & IoT

What analytical methodology to use for Data Science for IoT (IoT Analytics) problems?

On one hand, a Data Science for IoT problem is a typical Data Science problem – we can use classical analytics process methodologies like CRISP-DM

On the other hand, there are unique considerations to IoT: high data volumes, use of complex event processing, impact of streaming data etc.

“It’s the Sensor Data, Stupid!”
Five things we never talk about when we talk about sensors…

1. Sensors sometimes lie

Don’t assume that sensor data is accurate, complete and consistent; do apply Information Management best-practices.

2. Even when they are not lying, sensors may not tell the whole truth

Capture raw sensor data wherever possible; otherwise, understand how sensor data has been summarised.

3. Extracting useful signal requires “multi-genre” Analytics – and additional data

Plan to support a wide variety of Analytics – path / pattern / graph / time-series / text - just to prepare an ADS for modelling.

4. Sensors typically don’t measure the quantity of interest directly

Some model scoring may take place on the smart device itself; build models centrally, but be prepared to deploy them widely.

5. By itself, sensor data is frequently not actionable

Integrate your Data Lake and Data Warehouse environments, so that observation and transaction data can be joined.

#StrataHadoop
### Extracting useful signals from sensor data

<table>
<thead>
<tr>
<th>Raw sensor data</th>
<th>Cleansed sensor data</th>
<th>Event detection</th>
<th>Labelled sensor data</th>
<th>Path-to / Event association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw sensor data</td>
<td>Interpolation of missing values, &quot;virtual sensor&quot; correction for drift, re-calibration, etc., etc.</td>
<td>Identification of changes of state; signature matching</td>
<td>Comparison and correlation with human observations, other system / device events</td>
<td>Comparison and correlation with &quot;whole device&quot; and &quot;whole fleet&quot; sensor data</td>
</tr>
<tr>
<td>N/A</td>
<td>Interpolation, neural networks, FFTs, smoothing.</td>
<td>Time-series, Path, Pattern, Similarity</td>
<td>Text, Relational</td>
<td>Path, Graph, Clustering, Co-occurrence</td>
</tr>
<tr>
<td>Data Preparation and Management</td>
<td>Cleansed sensor data</td>
<td>Event detection</td>
<td>Labelled sensor data</td>
<td>Path-to / Event association</td>
</tr>
<tr>
<td>Key Success Factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Key Success Factors
1. Capture full-fidelity data to enable use-case specific event detection.
2. Integration of additional data.
3. Multi-Genre Analytics.
Anything less than high performance = high cost and low customer satisfaction

Post-operation manual analysis was slow and unsatisfactory
- Engine sensor data
- Vehicle master & engineering data

“Problem” Sensor data were a key issue in the current monitoring process

Single system failure detection required time-intensive computation for cross-fleet prediction
Rule-based flagging of “problem” data

- **Bin A – Valid/Good Data**
  - Assumed innocent until proven guilty (fails a rule)
- **Bin B – Errors**
  - Impossible Values
  - Error Codes
  - Nulls
- **Bin C – Outliers**
  - “Are 17 consecutive outliers 17 individual events or 1 single event?”

```
Rule
Bin A
- Valid/Good Data
  Assumed innocent until proven guilty (fails a rule)
Bin B
- Errors
  Impossible Values
  Error Codes
  Nulls
Bin C
- Outliers
  “Are 17 consecutive outliers 17 individual events or 1 single event?”
```
Flagging Error Data – Bin B

Sensor Flatlining
Flagging Outlier Data – Bin C

Sensor Outlier Data
Project Example: Data Integration

Leverage engine sensor data to predict train failure for train fleet in the UK.

Using training set consisting of roughly one million sensor log observations AND several thousand Engineer reports describing failure / fix.

Engineering Reports
- Service logs
- Detailed Failure information
- Categorized by Importance

Sensor Data
- From Trains (>200 variables)
  - Engine Temperature
  - Oil pressure
  - Velocity
  - GPS Position
  - Events (door open y/n)

- Exploration and Visualization using Tableau and Teradata Aster
- Predictive Modeling using Teradata Aster
Analysing Failure data individually and together with Sensor data

Affinity analysis – which components fail in combination (within the same train) – identify candidates for failure prediction

Path analysis – exploring the path to failure (testing different categorizations of sensor readings as events)

Predictive modelling – predict engine failures using random forest algorithm
Project Example: Multi-Genre Analytics

Predict shutdown (paper-jam) events for paper mills

Use machine logs, parsed to create significant machine events. Use pattern matching and affinity analysis to predict shutdown events.

Process Auto Equip
- Shutdown events
- Sensor logs

- Affinity analysis
- Npath Pattern detection
- Naïve Bayes Classifier
Paths and Patterns

Which sequence events most often lead to shutdowns?
Which events often appear together in a no-shutdown or shutdown time window?
## Predict Shutdowns using Naïve Bayes Classifier

<table>
<thead>
<tr>
<th>EVENT 1</th>
<th>EVENT 2</th>
<th>EVENT 3</th>
<th>LOGLIK No Shutdown</th>
<th>LOGLIK MechanicalShutown</th>
<th>PREDICTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>VH2N-N034PC22837_4</td>
<td>VH2N-N034PC22836_1</td>
<td>VH2N-N034PC22842_4</td>
<td>-24.083</td>
<td>-15.5238</td>
<td>Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PC22841_1</td>
<td>VH2N-N034PC22836_1</td>
<td>VH2N-N034PC22837_4</td>
<td>-22.4175</td>
<td>-13.9183</td>
<td>Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PC22836_1</td>
<td>VH2N-N034PC22842_4</td>
<td>VH2N-N034PC22841_1</td>
<td>-25.1901</td>
<td>-16.6981</td>
<td>Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PC22842_4</td>
<td>VH2N-N034PC22837_4</td>
<td>VH2N-N034PC22836_1</td>
<td>-25.8888</td>
<td>-15.5558</td>
<td>Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PC22837_4</td>
<td>VH2N-N034PC22836_1</td>
<td>VH2N-N034PC22841_1</td>
<td>-12.3208</td>
<td>-12.3208</td>
<td>Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PC22837_4</td>
<td>VH2N-N034PC22836_1</td>
<td>VH2N-N034PC22842_4</td>
<td>-18.4708</td>
<td>-18.4708</td>
<td>Shutdown</td>
</tr>
<tr>
<td>H2N-N034PC22842_4</td>
<td>H2N-N034PC22837_4</td>
<td>H2N-N034PC22836_1</td>
<td>-23.1044</td>
<td>-29.2531</td>
<td>No Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PM4058M1-R03</td>
<td>VH2N-N034PM4072M1-R03</td>
<td>VH2N-N034PM4089M1-R03</td>
<td>-24.7597</td>
<td>-29.2759</td>
<td>No Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PM4043M1-R03</td>
<td>VH2N-N034PM4029M1-R03</td>
<td>VH2N-N034PM4074M1-R03</td>
<td>-22.7526</td>
<td>-27.3709</td>
<td>No Shutdown</td>
</tr>
<tr>
<td>VH2N-N034PM4074M1-R03</td>
<td>VH2N-N034PM4043M1-R03</td>
<td>VH2N-N034PM4029M1-R03</td>
<td>-17.4364</td>
<td>-22.0436</td>
<td>No Shutdown</td>
</tr>
</tbody>
</table>

Naïve Bayes makes a prediction as to whether a triplet pattern (sequence or affinity) is more likely to be a Shutdown or No Shutdown triplet
Live Demo
Thank you very much

Frank Säuberlich
Frank.Saeuberlich@teradata.com

Eliano Marques
Eliano.Marques@thinkbiganalytics.com