Hot paths to anomaly detection with TIBCO data science, streaming on AWS

Steven Hillion
Sr Director, Data Science
@StevenHillion

Michael O’Connell
Chief Analytics Officer
@MichOConnell
The Ideal Data Science Platform

- Streaming
- Data Marts
- Transactional Stores
- External Sources

Point-Click | Code
Data Access & Prep
Modeling, ML, DL

- Batch Scoring
- Real-Time Scoring
- Live Applications
- Edge Devices

Offer Cross-Sell Products
Predict Impending Equipment Failure
Optimize Pricing
Prevent Fraud
Manage Inventory

© Copyright 2000-2019 TIBCO Software Inc.
Agenda

- TIBCO Data Science and AWS Marketplace
- The TIBCO Connected Intelligence Cloud
- Anomaly Detection and Analysis
- Demonstration – Spatial Anomaly Analysis
- Links and Assets
## TIBCO Data Science

### Data Access/Prep
- Distributed compute
- Feature engineering
- Reusable templates

### Modeling
- Visual composition
- Multilingual notebook
- Native ML & OS
- Auto-ML, data prep

### Operations
- Model lifecycle management
- Batch automation
- Real-time event processing
- REST, applications, embedding

### Business Apps
- Engineering/IoT
- Customer analytics
- Risk management
- Supply chain

### Function

<table>
<thead>
<tr>
<th>DATA ACCESS/PREP</th>
<th>MODELING</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Distributed compute</td>
<td>+ Visual composition</td>
<td>+ Model lifecycle management</td>
</tr>
<tr>
<td>+ Feature engineering</td>
<td>+ Multilingual notebook</td>
<td>+ Batch automation</td>
</tr>
<tr>
<td>+ Reusable templates</td>
<td>+ Native ML &amp; OS</td>
<td>+ Real-time event processing</td>
</tr>
<tr>
<td>+ Auto-ML, data prep</td>
<td>+ Auto-ML, data prep</td>
<td>+ REST, applications, embedding</td>
</tr>
</tbody>
</table>

### User or Automation

<table>
<thead>
<tr>
<th>USER</th>
<th>AUTOMATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientist</td>
<td>Citizen Data Scientist</td>
</tr>
<tr>
<td>Data Scientist</td>
<td>Citizen Data Scientist</td>
</tr>
<tr>
<td>Analytics Operations</td>
<td>IT / Software Engineer</td>
</tr>
<tr>
<td>Business User</td>
<td>Analytics Operations</td>
</tr>
<tr>
<td>IT / Administration</td>
<td></td>
</tr>
</tbody>
</table>

### Examples
- **Medic;** e.g., researcher on epidemic monitoring
- **Engineer;** e.g., aerodynamics engineer
- **Marketeer;** e.g., customer engagement analyst

© Copyright 2000-2019 TIBCO Software Inc.
AWS Machine Learning Partner Solutions

Provide solutions that help organizations solve their data challenges, enable machine learning and data science workflows or offer SaaS-based capabilities that enhance end applications with machine intelligence.

AWS Machine Learning Competency

Technology Partners

- **MACHINE LEARNING PARTNER SOLUTIONS**
  - Data Services
  - Platform Solutions
  - SaaS and API Solutions
  - Consulting Partners

TIBCO Data Science on AWS

TIBCO DS on AWS Marketplace; Biggest vCPU Grid; Lightest Serverless Footprint

TIBCO Data Science for AWS (10 users)

- Version 6.4.1 (Sold by TIBCO Software Inc.)
- Starting from $16.00/hr to $18.00/hr for software + AWS usage fees
- Scale data science across your organization to solve complex challenges faster and speed innovation with TIBCO Data Science for AWS Marketplace, a collaborative platform for operationalizing data science. Access Amazon EMR and RedShifts and enable data scientists to create innovative solutions using...

- Linux/Unix, CentOS 7 - 64-bit Amazon Machine Image (AMI)

Text Similarity Analyzer

- Version v1.1 (Sold by TIBCO Software Inc.)
- Engineers word/document features on a corpus with NLP methods, and uses these features to compare new text to the corpus.

Autoencoder for Anomaly Detection

- Version v1.1 (Sold by TIBCO Software Inc.)
- Identifies potential anomalies from transaction and or sensor data with a deep learning autoencoder.

Hospital Readmission

- Version v1.1 (Sold by TIBCO Software Inc.)
- Predicts hospital readmission rates from DRG codes, billing and EMR data.

TIBCO Spotfire Analytics for AWS (Hourly)

- Version 10.5.1 (Sold by TIBCO Software Inc.)
- Starting from $1.20/hr or from ($0.40/00/hr (20%) savings) for software + AWS usage fees
- TIBCO Spotfire provides visual analytics for deep insights into data from Amazon Redshift, EMR, Aurora, Databricks, SAP HANA One, Oracle, Microsoft Excel, SQL Server and more. Explore data, and create AI-powered visual dashboards in minutes. Easily scale from a small team to the entire organization...
Data Science in the Cloud: Leidos Healthcare Analytics

Leidos Collaborative Advanced Analytics & Data Sharing Platform (CAADS) uses TIBCO Data Science and AWS to deliver analytics services in Healthcare.

**CDC**
Disease Outbreaks: Determining the cause of an HIV outbreak in the Midwest

**NIH**
Disease Outbreaks: Run simulations of disease propagation to guide public policy, specifically around the Zika virus

**CMS**
Data Governance: Analyzing and consolidating data around emerging Healthcare policies across 56 regions in the United States

**NASA**
Space Exploration: Analyzing human factors that affect the ability to transport astronauts on long fights (e.g., to Mars)

© Copyright 2000-2019 TIBCO Software Inc.
TIBCO Data Science Solutions on AWS

Cloud Apps: Visual Analytics, Data Science, Streaming, Case Management

Anomaly Detection
- Review Status: TIBCO Spotfire
  - Identify issues, sweet spots
- Model: TIBCO Data Science
  - Supervised: Train
  - Unsupervised: Anomalies
- Analyze Event Stream:
  - TIBCO Flogo, Cloud Integration
  - Batch and Real-Time Updates

Risk Management
- Case Manage: TIBCO Live Apps
  - Investigate identified cases
  - Audit trail + recycle

Customer Engagement
- Starter Set
  - Process Mining
  - IoT Analytics
  - Anomaly Detection
  - Risk Management
  - Customer Engagement
  - Blockchain – Dovetail
  - Partner Management
  - Starter Toolkit

© Copyright 2000-2019 TIBCO Software Inc.
Anomaly Analysis Solution Overview

1. Collect data from equipment, normalize, model to predict magnitude of anomaly – TIBCO Data Science & AWS

2. Model detects anomaly – TIBCO Data Science

3. Alert raised and case created – TIBCO Cloud Live Apps

4. Case manager investigates and takes action to the equipment – TIBCO Cloud Integration
Data Challenges in High-Tech Manufacturing

Stack 3D Flash Memory
Cell Layers Vertically

96 Memory Cell Layers
Hot Paths to Anomaly Detection
Longitudinal Anomaly Analysis

pressure, prodPerMinute, baroPressure, wasteGas, lossOfEfficiency, sr, dp1, dp2, dp3, dp4, dp5, dp6 vs. eventDateTime (Day of Month), eventDateTime (Second)
Spatial Anomaly Analysis

One-Valued Wafer Maps

Big Red Spot

Plaid

Geometric Patterns

One Big Blob & One-Valued

One Big Blob

© Copyright 2000-2019 TIBCO Software Inc.
Cross-Sectional Anomaly Detection

Sample
Display results from this sample:
Training Sample - used for building the model

Set Cutoff for Outliers

Reconstruction Error over Time

Components Signature

Reconstruction.MSE, ContributionToError, raw vs. Key (Day of Month), ...
Cross-Sectional Anomaly Detection

Sample
Display results from this sample:
Training Sample: used for building the model

Set Cutoff for Outliers

Incidents: 3
Number of

Reconstruction Error over Time

Components Signature
reconsr_dp5_SE
reconsr_pressure_SE
reconsr_SR_SE

Data table: result
Marking:
- Mark Variability

Reconstruction. MSE, ContributionToError, raw vs. Key (Day of Month),...
Analysis: Cluster Incidents, View Signatures
Analytic workflow – methodologies + demos

Find and cluster anomalous events [Demo #1]
• Transform wafer maps into vectorized coefficients, then cluster on quality
  • Many measured parameters, e.g., quality tests: storage fidelity, logic circuits
• Approach 1: SVD + K-means
  • Focus on failure mode parameters
• Approach 2: Bessel functions + hierarchical clustering
  • Radial basis functions
  • Rotationally invariant | Null-value tolerant | Efficient storage
  • Better than SVD + K-means for multi-parameter analysis

Monitor anomalies [Demo #2]
• Stream wafer data
• Vectorize and cluster

Predict when and why anomalies occur [Demo #3]
• Reduce dimensionality of very wide data
• Train models to determine sensor importance
• Identify responsible process parameters

Process variable corrections/models rebasing
• Identify new patterns as they emerge (e.g., incident analysis)
• Factory monitoring staff can click to characterize the new pattern

Demo data
• 733 dies
• 20 x 40 layout
• 1,500 measured parameters
• Color by die test (white = pass all, colors = fail a test)
Ingest data into Amazon Kinesis for further processing.

Capture and send data to livestream.

Read data from Kinesis into TIBCO Streaming.

Preprocess and transform the data.

Perform SVD on the bin values for wafers using Python.

Send data for live viewing into Spotfire Data Streams.

Refine clusters, identify wafers of interest.

Send data to Spark/Hadoop cluster for more in-depth analysis in TIBCO Data Science.

Retrain models for clustering and root-cause analysis.

Combine data & results.

PMML operator to cluster the data using a model trained in TIBCO Data Science.

(TIBCO Streaming operator for Python)
Anomaly Detection with Spatial Signature Analysis
Anomaly Analysis

- Find and explore anomalous events
- Monitor anomalies
- Predict when and why they occur
Anomaly Detection and Analysis

- Find and explore anomalous events
- Monitor anomalies
- Predict when and why they occur
Anomaly Detection and Analysis

Find and explore anomalous events

Monitor anomalies

Predict when and why they occur
Anomaly Detection and Analysis

Find and explore anomalous events

Monitor anomalies

Predict when and why they occur
Digital Twin for Semiconductor Yield

Digital Twins for Semiconductor Manufacturing Yield: Wide-and-Big Data Analysis Build Models to Relate Product Yield Failure Modes \( (Y_i) \) with Process Parameters \( (P_i) \)

- **Process Optimization**
- **Process Monitoring**
- **Process Control**

\[ Y = f(Y_1, Y_2, Y_3, \ldots, Y_n) \]

- **Run continuously to capture new relationships as they emerge**

- **Product Test:**
  - **Pass / Fail, Failure Modes**
  - **Yield, Yield Components or Clusters**

\[ Y_i = f(P_1, P_2, P_3, \ldots P_n) \]

- **~ 1K Steps, > 1M Process Parameters**
The Extreme Challenge of Big & Wide Data

• **Not just big data** – many rows: lots, wafers, die, units
• **Also wide data** – many columns: > 1M process parameters
  • Sensor traces
    • Time series for every sensor on each machine in each run
  • Physical measurements
    • Film thickness, critical dimensions, layer-to-layer overlay, defect classes & counts
  • Equipment and process attributes
    • Machine and component IDs, process recipe info
  • Supplies
    • Chemical batch IDs, QA sample data

“Today [semiconductor] fabs collect more than 5 billion sensor data points each day. The challenge is to turn massive amounts of data into valuable information.”

—Ann Kellehere, VP of the Technology and Manufacturing Group, Intel
Solution Architecture

Data Prep, Feature Engineering & Selection

- In-database parallelized computing
- Leverages Hadoop, Apache Spark

Computing Cluster

Data Store: Big and Wide Data

Further Feature Selection & Model Building

- In-memory dedicated fast server

In-Memory Fast Analysis Servers

Visualization of Results

- Interactive in-memory visualization environment

Visualization Servers

Visualization of Results
Performance Benchmarks & Conclusions

- Demonstrated performance for time series data from 20,000 sensors, 10,000 wafers in **under 2 minutes**
- Current system scales to time series for 20,000 sensors, 100,000 wafers (**2.5 TB**) with results in **15 minutes**
  - More capacity and better performance can be achieved by adding nodes to the Spark cluster
- Working with top memory manufacturer to deploy production system
- System can provide automated real-time feedback on emerging equipment issues affecting yield

### Big Data Feature Selection Performance Benchmarks – Run Time\(^1\) (minutes)

<table>
<thead>
<tr>
<th>Dataset Size for 1M Variables</th>
<th>20 Sensors (1K variables)</th>
<th>200 Sensors (10K variables)</th>
<th>2K Sensors (100K variables)</th>
<th>20K Sensors (1M variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1K Wafers</td>
<td>0.47</td>
<td>0.48</td>
<td>0.72</td>
<td>1.0</td>
</tr>
<tr>
<td>10K Wafers</td>
<td>0.50</td>
<td>0.53</td>
<td>0.77</td>
<td>1.75</td>
</tr>
<tr>
<td>100K Wafers</td>
<td>0.53</td>
<td>0.67</td>
<td>1.25</td>
<td>15.15</td>
</tr>
</tbody>
</table>

\(^1\) **Test Conditions:**
- Data stored in Hadoop data source
- 25 node Spark cluster – 16 cores, 32 GB for each node
- Each sensor time series compressed to 50 variables with SAX encoder prior to feature selection step
Longitudinal Anomaly Analysis

Subsequence Search

A New Method for Identifying Anomalous Patterns in Time Series (Trace Analytics)
Mueen’s Algorithm for Similarity Search (MASS) is specialized for finding anomalous (versus typical) subsequences of time series.

Extremely fast algorithm for this use case

Suitable for further acceleration using GPU

Material adapted from:
https://www.cs.ucr.edu/~eamonn/matrix_profile_i.pptx
Mueen’s Algorithm for Similarity Search

Quickly create a matrix profile = a partial distance matrix

This uses a sliding window to define a series of subsequences

The Matrix Profile plots the distance of each subsequence to its nearest match, with the time sequence of the start of each subsequence on the x-axis

Material adapted from:
https://www.cs.ucr.edu/~eamonn/matrix_profile_i.pptx
How to “Read” a Matrix Profile

Where you see relatively low values, you know that the subsequence in the original time series must have (at least one) relatively similar subsequence elsewhere in the data (such regions are “motifs” or reoccurring patterns).

Where you see relatively high values, you know that the subsequence in the original time series must be unique in its shape (such areas are “discords” or anomalies).
Manufacturing Batches

Raw Amperage - Each color delimits a batch

Matrix Profile highlights anomalies - set sliding window close to batch size
Community
AI in Operations

Cloud Starters, Accelerators, Analytic Apps

Thoughtleader-Led Solutions

AI on Demand
Data Science in Operations

Driving Customer Engagement

Fraud and Risk Management

Insurance Dynamic Pricing

Digital Twins in Manufacturing

The Industrial Internet - Production Surveillance
Questions & Contact

Steven Hillion
Sr. Director, Data Science
shillion@tibco.com
@StevenHillion

Michael O’Connell
Chief Analytics Officer
moconnell@tibco.com
@MichOConnell

TIBCO Community
community.tibco.com

TIBCO Exchange
community.tibco.com/exchange

TIBCO Tech Blog
community.tibco.com/blog

Acknowledgements
Dr. Tom Hill, Mike Alperin, Nico Rode, David Katz, Jagrata Minardi, Siva Ramalingam
Please complete the session survey in the mobile app.