Using Machine Learning for Intelligent Storage Performance Anomaly Detection

Ramakrishna Vadla, IBM

Archana Chinnaiah, IBM

Acknowledgement : Sumant Padbidri, Anbazhagan Mani

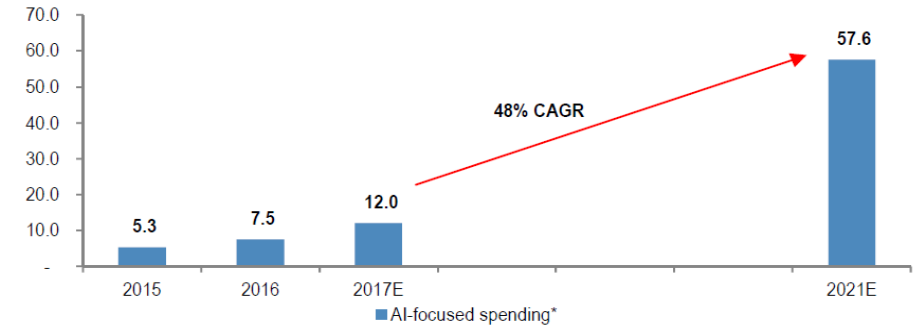
Agenda

- Market Estimates & Forecasts
- Applications in Storage
- Cloud Architecture
- Anomaly Detection
- Performance Anomaly Detection

AI & ML - Market Estimates & Forecasts

- ✓Worldwide revenues for cognitive and AI systems will increase from $12.5B in 2017 to more than $46B in 2020
- ✓IDC forecasts spending on AI and ML will grow from $12B in 2017 to $57.6B by 2021.

Figure 2: Global AI-focused spending* ($, bn)



Source: AI-spending estimates from IDC. *Includes AI-focused spending on hardware, software (applications + software platforms), and services (IT consulting & system implementation).

Number 3
Machine Learning

5 Year Growth Rate: 34%

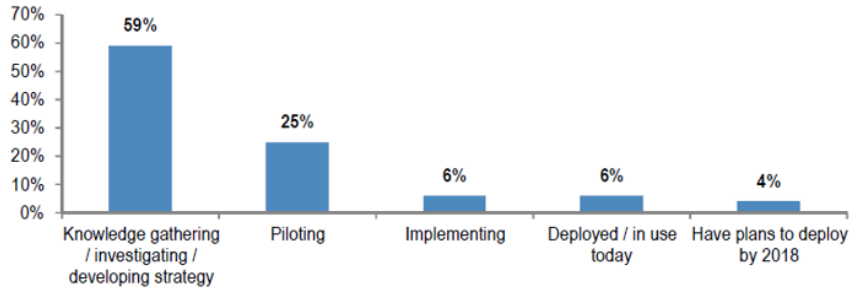
- Published patent applications for Patent Classification G06N "Computer Systems Based on Specific Computational Models" grew at a compound annual rate of 34% from 2013 to 2017.
- This includes machine learning and artificial neural networks.

Company	2017 published applications
IBM	654
Microsoft	139
Google	127
LinkedIn	70
Facebook	66
Intel	52
Fujitsu	49

Source: IFI Claims Patent Services (Patent Analytics). 8 Fastest Growing Technologies SlideShare Presentation.

- ✓ Machine learning patents grew at a 34% between 2013 and 2017, 3rd-fastest growing category of all patents granted.

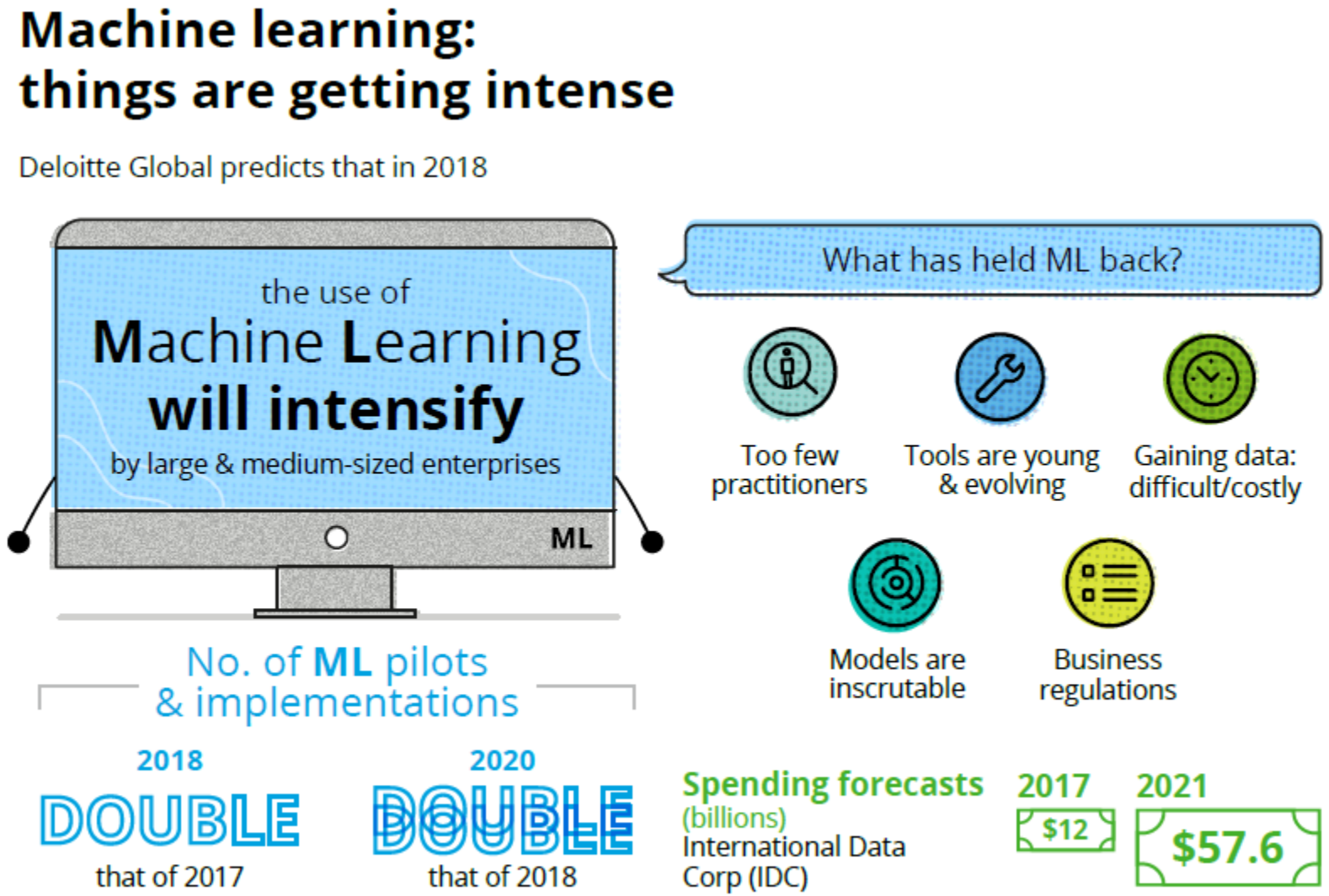
Figure 18: Current stage of enterprise AI adoption



Source: Gartner survey. Responses: What is the current stage of artificial intelligence solutions adoption within your organization?

Source:http://www.forbes.com

AI & ML - Market Estimates & Forecasts



Why Now?

- ✓ Enormously increased data - 90% data created in last couple of years
- ✓ Substantially more-powerful computer hardware – CPU, GPU
- ✓ Cloud makes big data more widely accessible
- ✓ Significantly improved algorithms

Source: Deloitte Global Predictions 2018 Infographics

Machine Learning Applications in Storage

Applications

- Predictive Analytics
  - Capacity Forecasting – (Regression)
  - Power consumption in data centers – (Regression)
  - Tracking of known issues - Learn from other customer issues - (Classification)
  - Predicting blocks to be accessed in near future (Recommendations)
- Performance anomaly detection
  - Performance metrics analysis (Time-series data analysis)
  - Automated Triaging and Root Cause Analysis (Classification)
  - Log analysis - (Clustering)
- Configuration best practices recommendations
  - Manual upgrades/Automated upgrades
  - Configuration validation to avoid interruptions in service
- Intelligent Performance Tuning

Value Proposition

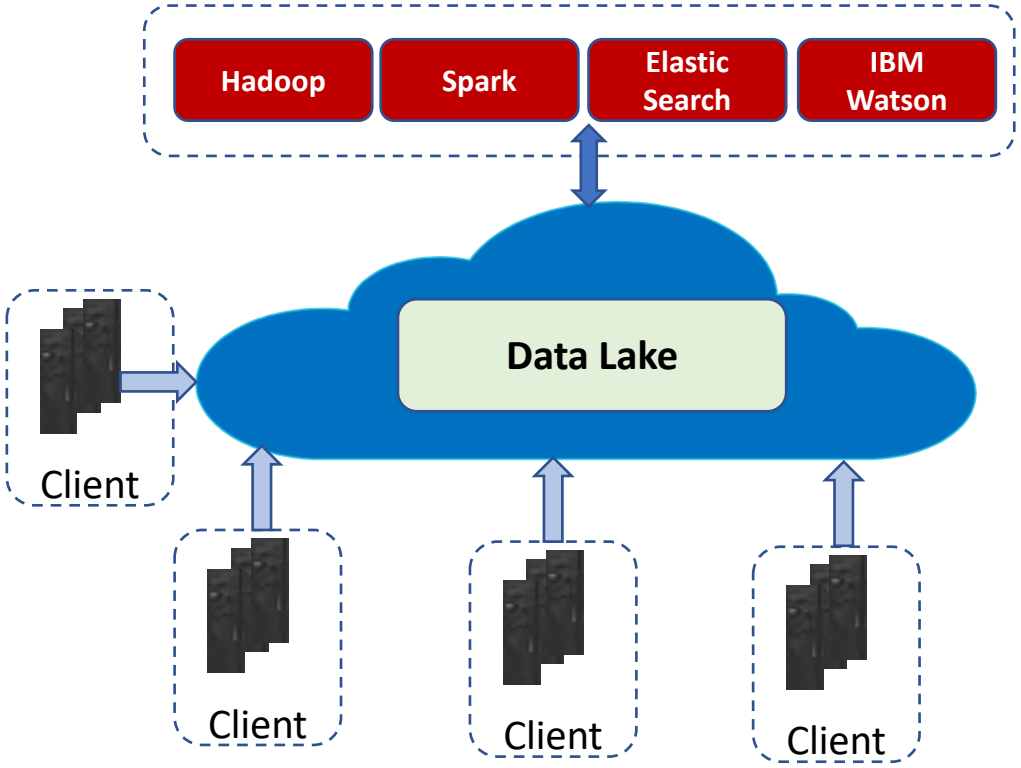
- ✓ **Prevent Issues proactively before they occur.**
- ✓ **Avoid downtime & Achieve uptime 99.999%**
- ✓ **Cost efficiency - Reduce storage & operational costs**
- ✓ **Data Storage Optimization**
- ✓ **Simplifying the support**
- ✓ **Proactive notification of risks and health checks**

Cloud Architecture - Storage Analytics

The world's most valuable resource is no longer oil, but data

www.economist.com

- ✓ Cloud based scale-out architecture.
- ✓ Storage systems support data collection with high frequencies, seconds, minutes.
- ✓ More data available for analysis.
- ✓ Data lake based on NoSQL such as Cassandra deployed on the cloud.
- ✓ All clients send storage metric data to cloud – performance, config and health data.
- ✓ Multi-tenancy support.
- ✓Support for integration of ML tools.



Machine Learning – Anomaly Detection

Supervised Learning	Predict based on training data containing desired outputs. • Training data contains normal and anomaly labelling • Regression, Classification, Decision trees, Random forests, K-Nearest Neighbor, SVM
Unsupervised Learning	Doesn’t include desired outputs, goal to discover patterns • No labels provided – assumption anomalies are very rare compare to normal • Clustering - K-Means, Hierarchical, DBSCAN, Time-series analysis, ARIMA
Semi-supervised Learning	Training data includes a few desired outputs Training data contains only normal labelling
Reinforcement Learning	Rewards from sequence of actions Agent -> Action - > Environment -> Reward & State -> Agent (Markov Decision Process)

Storage Performance Challenges

Bottlenecks

- Disk failure/Inaccessible disks
- Read/Write I/O errors
- Volume issues
- Port masking
- Configuration issues – Host, Storage subsystem, port, Interoperability
- Network congestion
- Workload configurations
- UPS battery failure
- Port protocol errors,
- Port congestion

Metrics

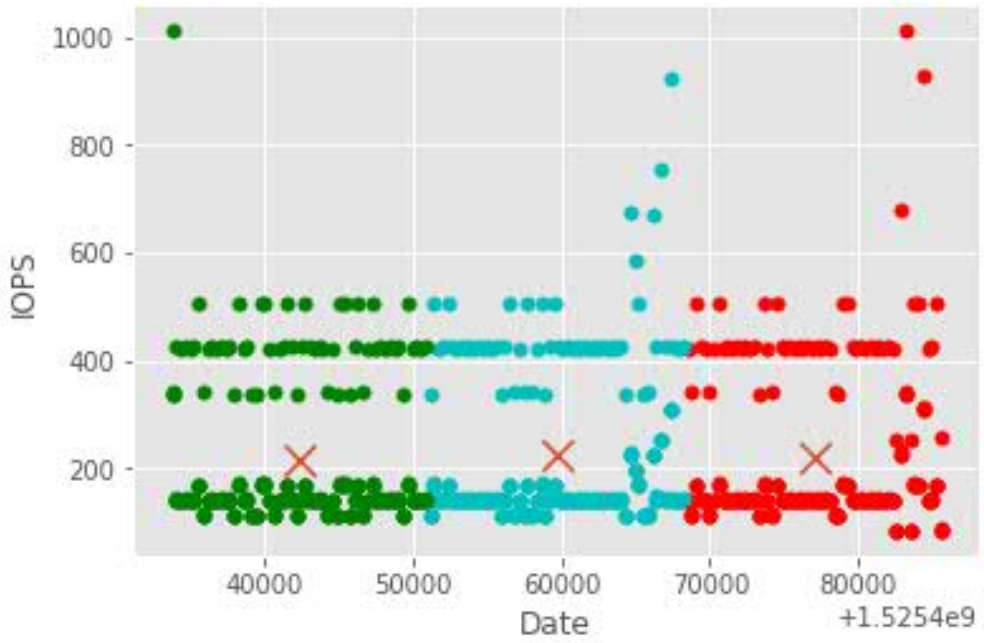
- I/O Rate R/W,
- Data Rate R/W,
- Response time R/W,
- Cache hit R/W,
- Data block size R/W,
- Porta data rate R/W,
- Port-local node queue time

Correlations

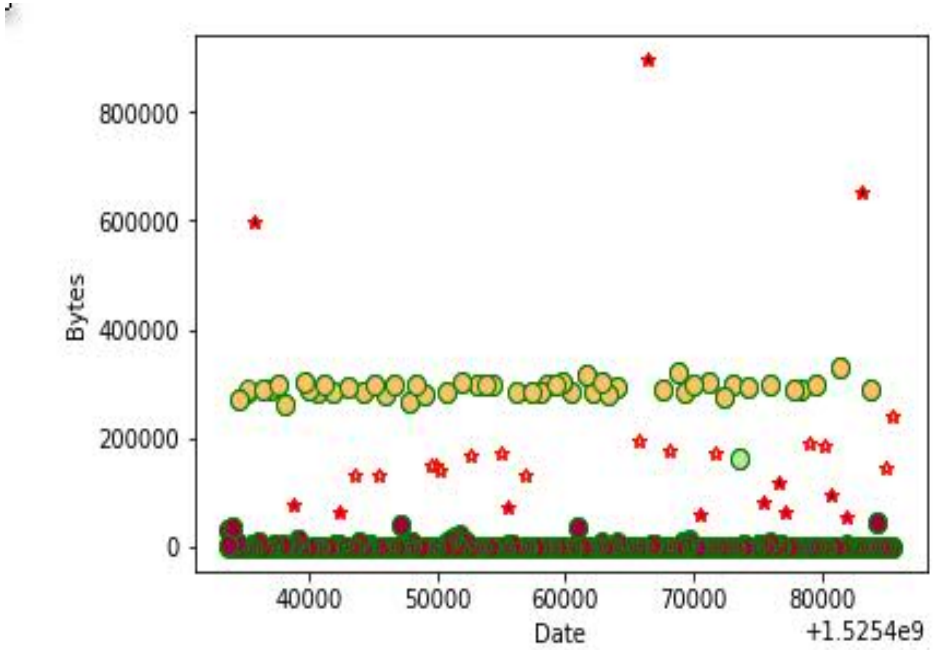
- CPU & Network Traffic
- CPU & Memory
- Port & Host counters
- IOPs, read rate, & CPU, memory

Performance Anomaly Detection

Clustering – Outlier detection



K-Means

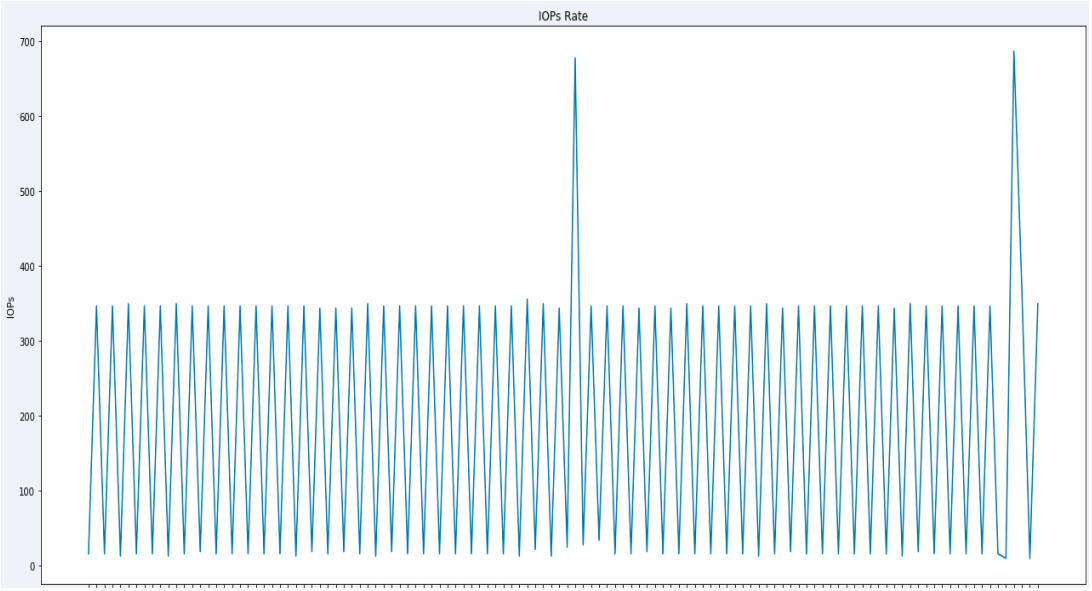


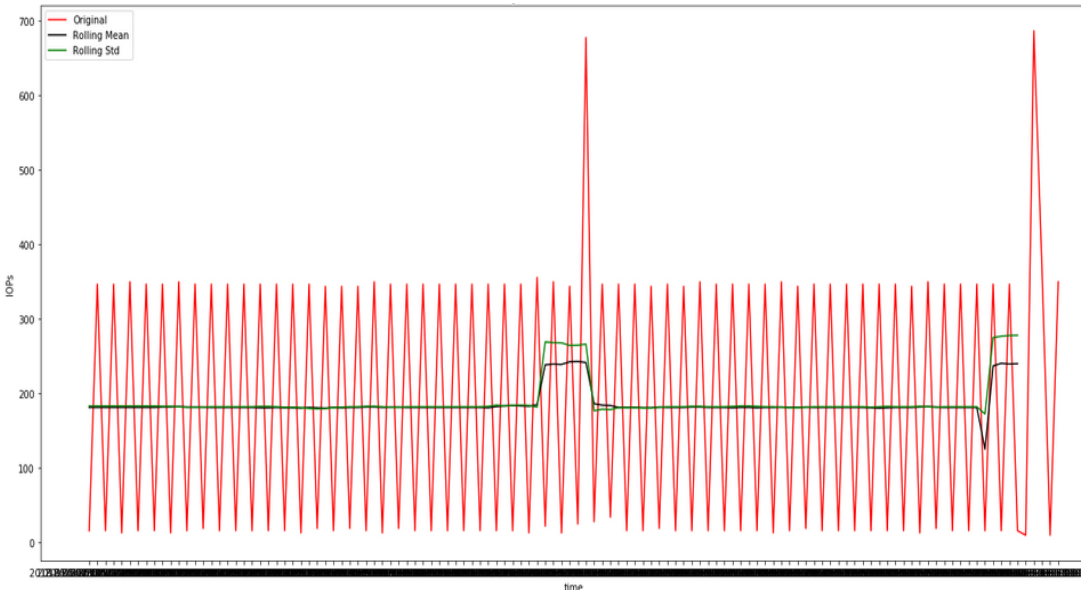
DBSCAN

Performance Anomaly Detection

Time Series Anomaly Detection

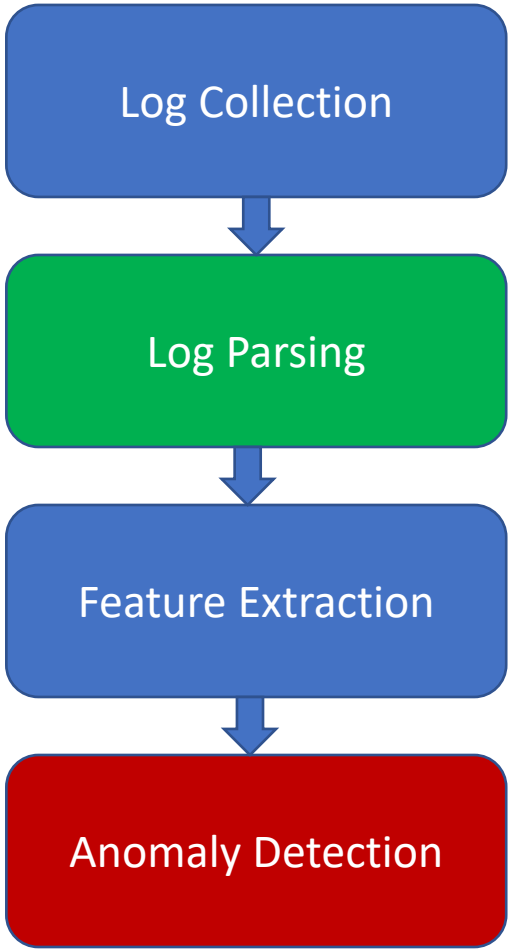
- ARIMA - AutoRegressive Integrated Moving Average





IOPs Rate Anomaly

Log Analysis – Anomaly Detection

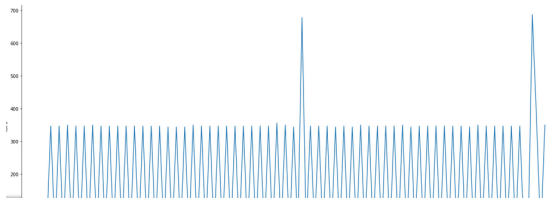


2018-05-05 09:11:20.672 [<Device>] [<Thread>] [INFO] Processing complete.

2018-05-05 09:11:20.672 [<Device>] [<Thread>] [INFO] **Processing complete**.

[timestamp, device, process state].

Time-series Analysis



Q & A
Thank You