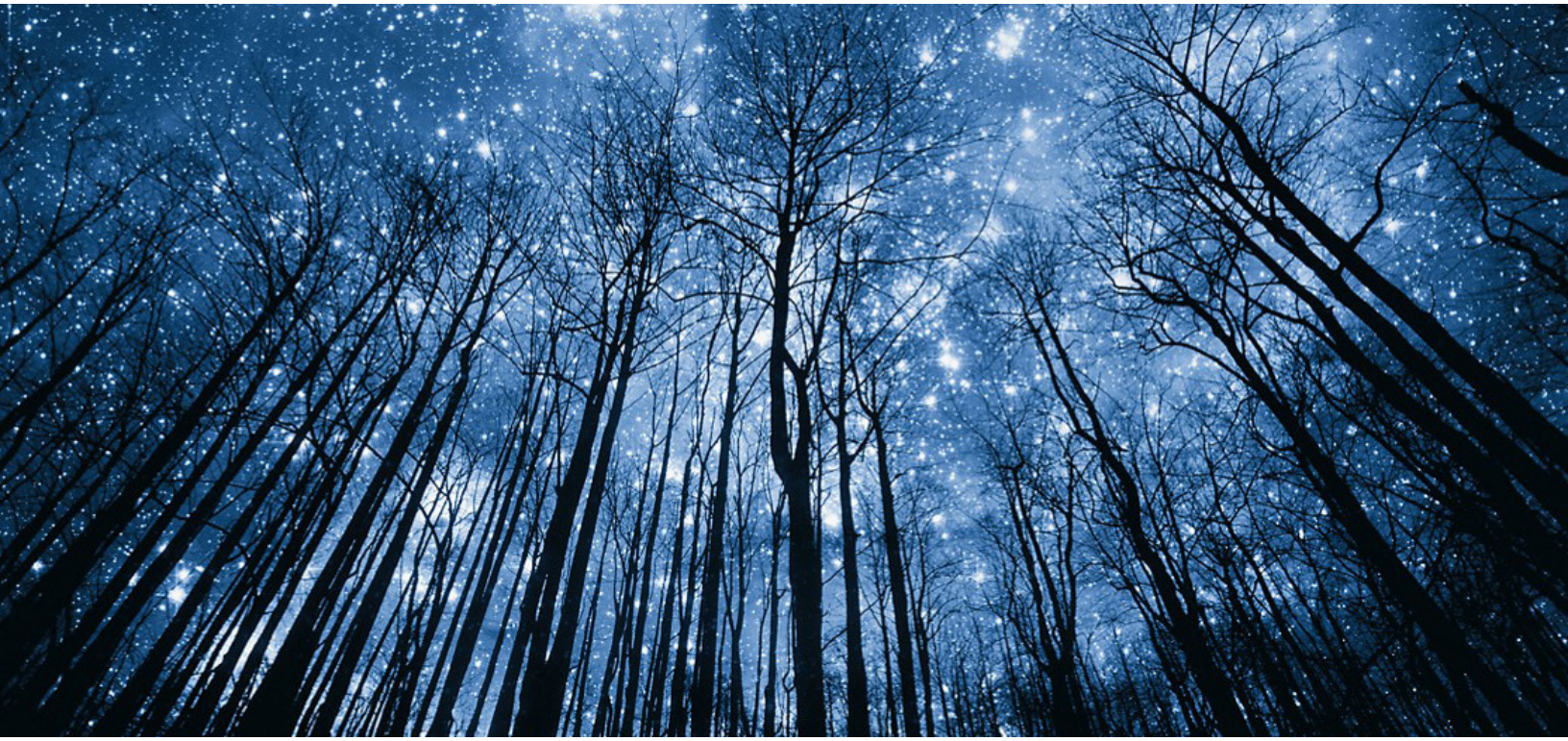
DELL Technologies

CYBER-PHYSICAL SYSTEMS SMART DATA HANDLING



Mohamed Sohail

Advisory Consultant, Data Center & Business Resiliency
Dell Technologies
Mohamed.sohail@dell.com

Knowledge Sharing Article

DELL Technologies Proven Professional

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Cyber-Physical Systems (CPS) security is a serious challenge for all deployed implementations. Gartner predicts by 2025 there will be 58 billion dollars for the IoT IT services with more than 21 billion devices, of which 7.5 billion devices will be for business purposes and 12.8 will be for consumers. With the rise of threats to critical assets, organizations need to expand security programs to include cyber-physical systems. All these devices will heavily interact with many endpoints, most of which with a dynamically changing security profile. Organizations around the world seek better solutions and ways to harden their systems while consistently adding new security controls. By 2025 Gartner predicts that 50% of asset-intensive organizations such as utilities, resources, and manufacturers will unite the security teams of the main three departments (Cyber, Physical, and Supply Chain) under one chief information security officer [CISO] that will directly report to the CEO.

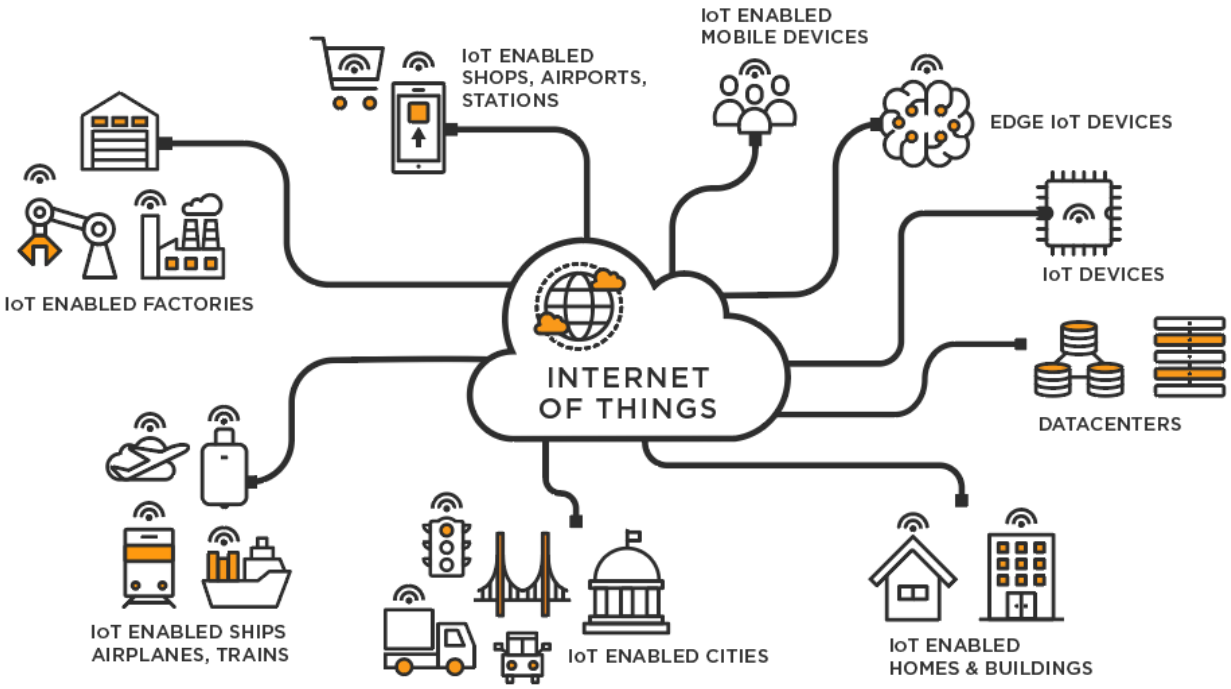


Figure 1: IoT explained[1]

Overview

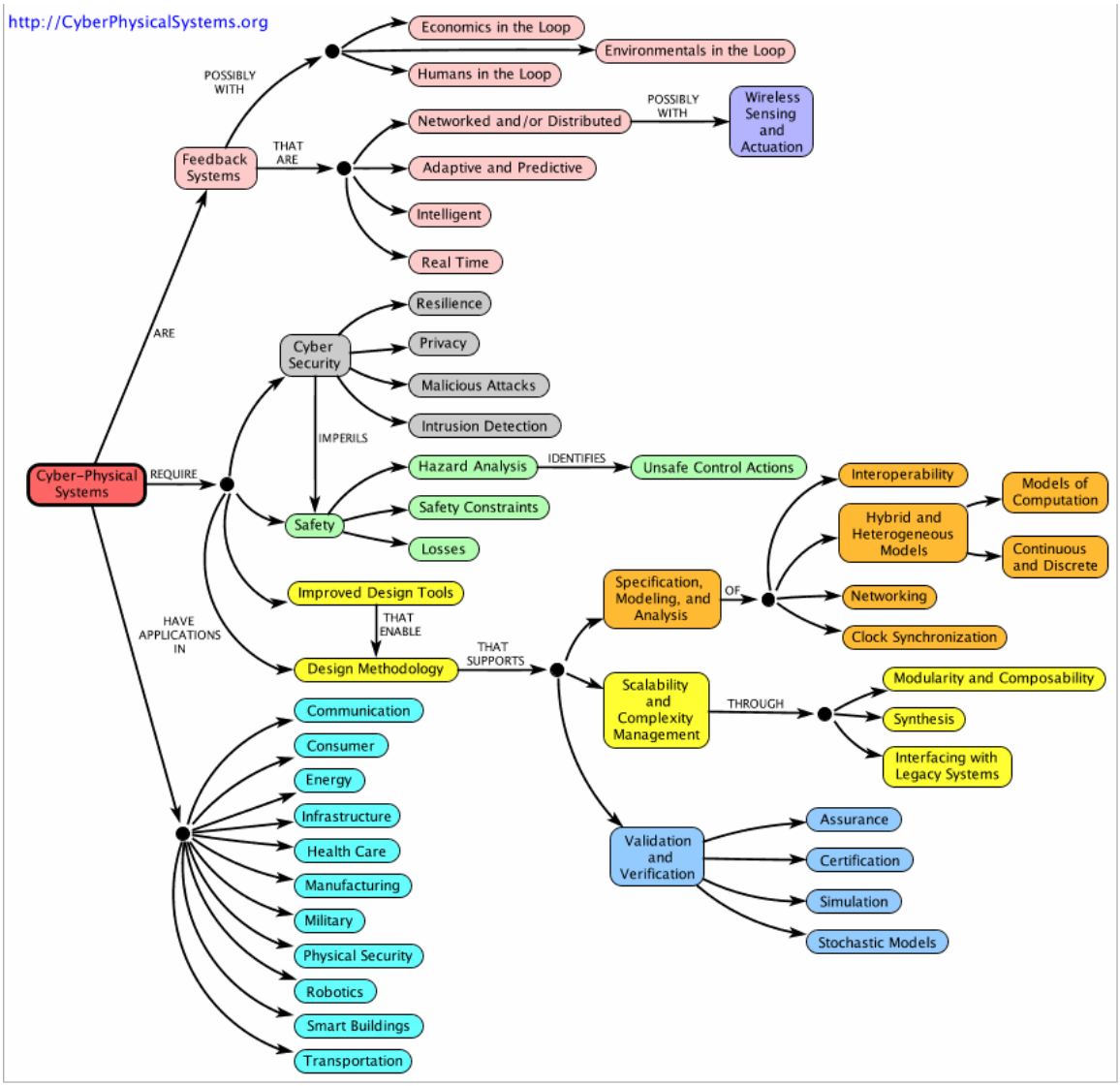


Figure 2: Cyber-Physical Systems Concept map[2]

Internet of Things (IoT) drives numerous innovations and transformations, which come with challenges and risks. The security and safety of data represent one of the top concerns. This disclosure rethinks the veracity and analytics of CPS data handling based on its sensitivity in the context of IoT tiering and classification. Some IoT and edge use-cases, as well as some cloud-based applications, show distinct patterns, such as:

Data-intensive: Massive data could be generated 24x7 from a huge number of sensors/devices, which places a burden on storing and processing it. For example[i], a smart field may have a sensor every 1~10 meter, equating to 1 million sensors for a 10

square KM field. Assuming 1KB data (humidity or temperature) generated per sensor-minute, there will be 1+TB data ingested per day. Another CPS data example is video surveillance which brings even more data. Shouldn't we think about ways to store and analyze this data?

Globally accessed: The huge amount of data, once ingested, is usually accessed globally by different users (Apps) from different locations (across sites, cities, or countries) for different purposes. This will be increasingly important and common as such data, once shared, can be much easier to access or analyze by different organizations/companies which greatly boosts digital transformation and overall efficiency.

Distinct read and write data is generated by edge sensors/devices. Once ingested, those machine-generated data such as to cloud, could be distributed or replicated to multiple data centers or sites. Thereafter, many users or apps global-wide would access or analyze the data usually in reading mode (analyzing the data is the most valuable part of IoT). Such a pattern aligns with many cloud-based Apps including the web, news, photo sharing, etc. In this Knowledge Sharing article we will show a new method for handling sensitive data in the cloud and how total computing and storage costs can be reduced.

As organizations move to the cloud, many wish to implement efficient mechanisms to reduce total operating costs. This creates a need for new security measures and performance optimizations based on the data sensitivity that will be based on data classification from creation to deletion to represent a robust and accurate data life cycle mechanism.

For CPS, we assert that use of traditional encrypt/decrypt became a necessity in any setup where data is generated to ensure that integrity and confidentiality are achieved. In this article, we describe a new way to handle CPS data in a federated way to reduce the cost of computing it and achieve an optimized performance based on the right classification of data.

We propose a new method of classifying data based on its sensitivity – Automated Security Tiering (Figure 3). The proposed functionality includes creation of customized policies based on classifying data flows coming from the underlying CPS (smart homes, manufacturers, aviation, connected cars, etc.) to better leverage various encryption algorithms and classification on the cloud. Figure 3 highlights the dataset handling.

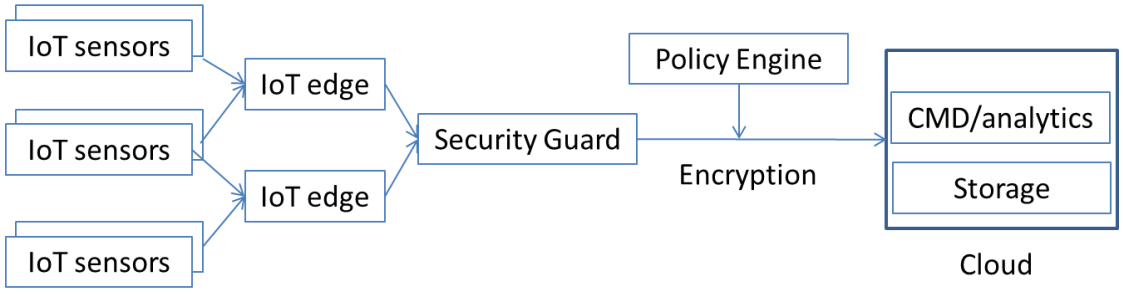


Figure 3: Cyber-Physical Systems dataset handling

What are the problems we currently face?

Industry-leading cloud providers offer storage and/or analytic capability via 3 options:

1) **Persistent store and Analytics**: CPS datasets are secured/stored at rest and have an analyzing interface (depending on user privilege)

2) **Persistent store only**: CPS dataset is secured and stored at rest without analyzing interface.

3) **Analytics only**: CPS dataset can only be analyzed with a specific interface but no store.

The AWS example does not leverage the wide variety of options for secure, encrypted analytic services. Figure 2 highlights this variety.

Figure 4: Generic Devices[3]

CSP typically generate continuous streaming datasets 24x7 from massive sensors and devices. Some datasets have degrees of sensitivity, i.e. personal wearable devices, healthcare devices, video surveillance, car location, etc. Many of these data sets will exist in the cloud for a long time and provide value via real-time, batched, or hybrid analytics for pattern detection and prediction. Others may disappear rather quickly to save on cloud computing costs. Today's systems, however, are often limited to all-or-nothing encryption approaches that do not leverage the variety of options highlighted in Figure 2. This leads to the following problems.

Problem 1 - Homogenous Security for all CPS Data Analytics

The main problem, especially for large data sets, is the "all-or-nothing" policy for encrypting data, is not allowing users to easily perform fine-grained actions such as sharing records or searches.

Figure 5 shows a scenario where all users, and all aspects of the underlying infrastructure, operate under the same encryption paradigm. This can tie the corporation to the same level of leakage for all CPS data, for example.

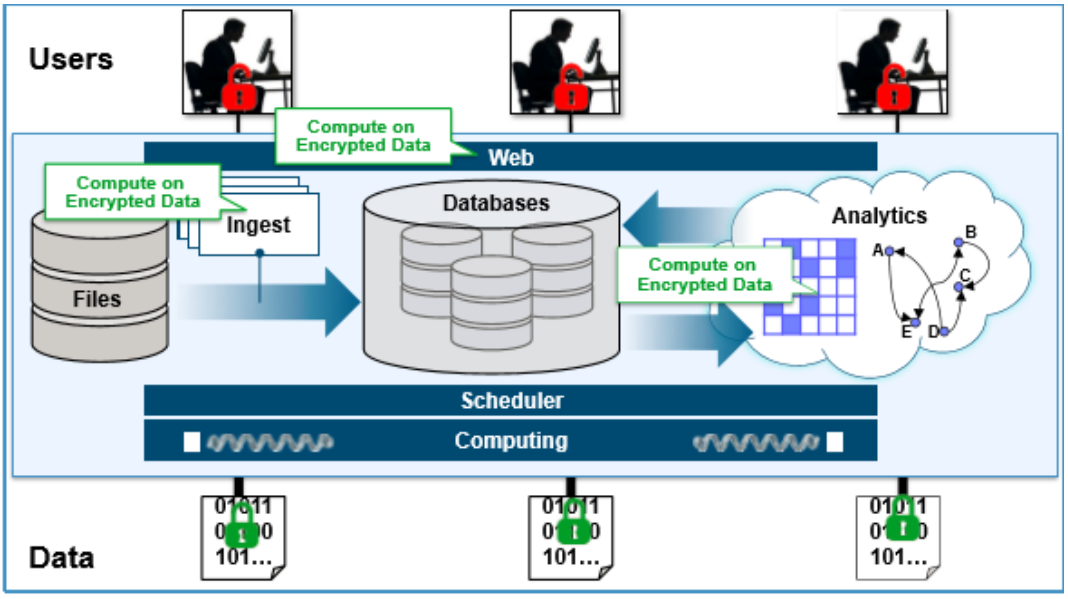


Figure 5: Homogenous Encryption during Analytics

Problem 2 - No Data Classification Security Granularity

Currently, we face challenges when it comes to the data types and how the proper securing tier can be determined when it comes to the encryption level that needs to be applied to specific data types. As shown in Figure 6, once the data type and its level of confidentiality is determined certain levels of security and handling can be enforced.

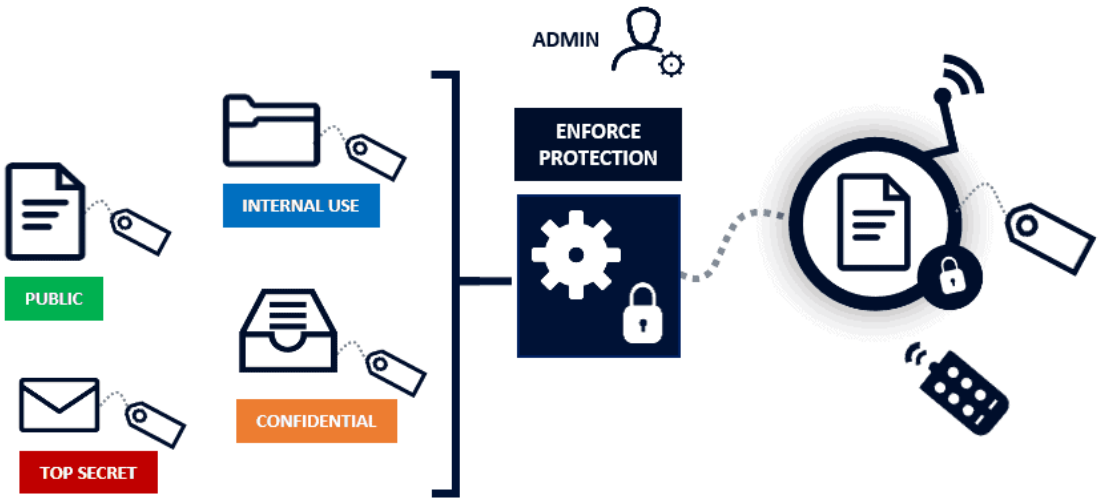


Figure 6: Data Classification[4]

Coarse-grained, homogeneous security mechanisms force data that could otherwise be shared into more restrictive categories. Similarly, data with higher sensitivity is often placed into analytic repositories that have an unacceptable potential level of leakage.

There is no way for providing Heterogeneity for Secure-Based Analytics

There are two problems with leveraging heterogeneous security services for analytics:

1. Advertisement of these options in the cloud is non-standard. In addition, there is no cost model for different forms of heterogeneous security services.
2. The algorithm for a gateway device selecting a security service provider, based on security services for analytics, does not exist.

Data Classification and Mapping at the Gateway

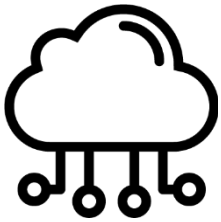
There is currently no way to classify data at the gateway level in a way that maps to security analytic services at the cloud level.

Problem 3 - No Method to Up-level/Down-level security classification

There is currently no way to instruct a cloud provider to provide tiered conversion of data in the cloud to either decrease leakage (at a higher compute cost and fee) or decrease cost by accepting a higher level of leakage.

Nor is there a way to perform this in a time-lapsed fashion (the data only needs to be secure for the first week, when it can then be downgraded to clear text).

How can we solve these problems?



In the section below we will introduce a significant approach for realizing an efficient and economical approach for computing the sensitive encrypted data since we will need to decrypt to read the data from the backend storage. This new method for CPS data handling will ensure the proper security tier of the data residing in the cloud to avoid paying for extra computing power to be read or analyze data when needed. The concept of security tiering is a novel way to reduce the computing cost associated with analytics on masked/encrypted data. This takes into consideration moving the data from different "security" / encryption levels. The goal is to enable the IoT Gateway to determine the security policies and the needed encryption level to apply on a certain data set and when it should be moved from one tier to another.

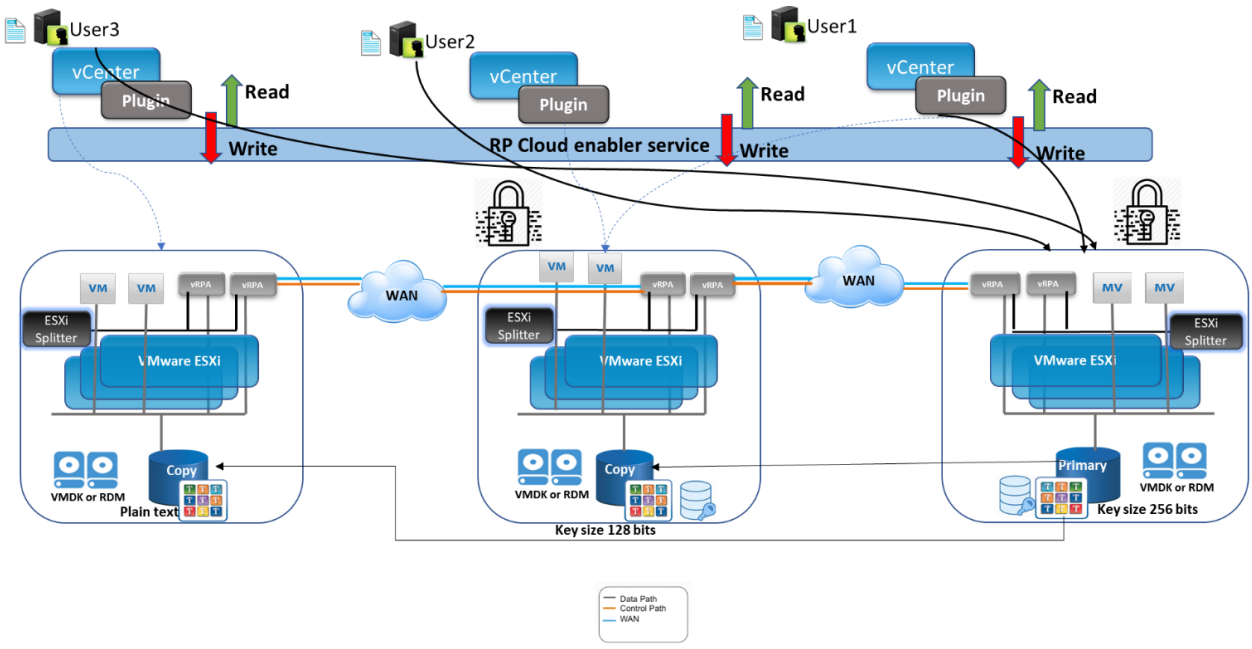


Figure 7: Distribution and data movement for encrypted / read-intensive IoT data

Based on the rules shown in Figure 7, from identifying the most read-intensive CPS data and adjusting it to the proper and performance-optimized storage, the sensitivity based on the nature of the sensor data and how it should be considered when dealing with "read with analytics on masked/encrypted data" can be added.

Below are the novel aspects of the solution.

Cyber-Physical Systems' sensitivity rule creation

In the first step we will need to form a sensitivity-based data classification and tiering routine, we guarantee sufficient data security at flexible and configurable granularity, in order to interoperate with multiple encryption approaches in the cloud and move the data from the different security tiers based on pre-defined security rules. This will include device classification for confidentiality. More information about the initial client configuration and network registration is available in the appendix.

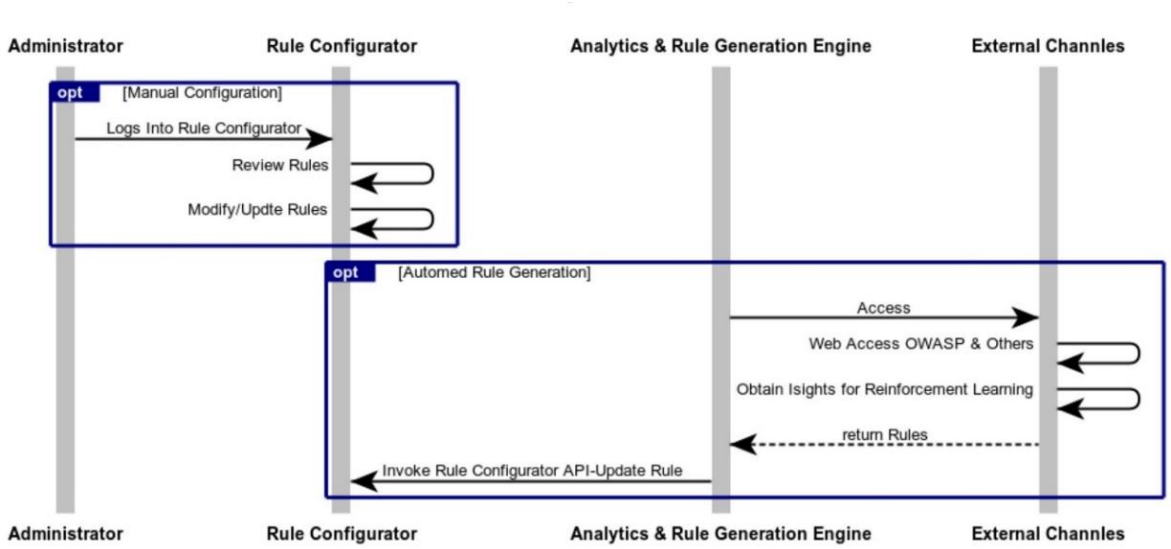


Figure 8: Cyber-Physical Systems IoT sensitivity rule creation

Modifying Classification Based on Usage

The engine will benefit from an audit log that tracks analytic usage. The approach used could leverage techniques from a previous "Governed Replay" article (EMC-15-0079), which is depicted in Figure 9 below.

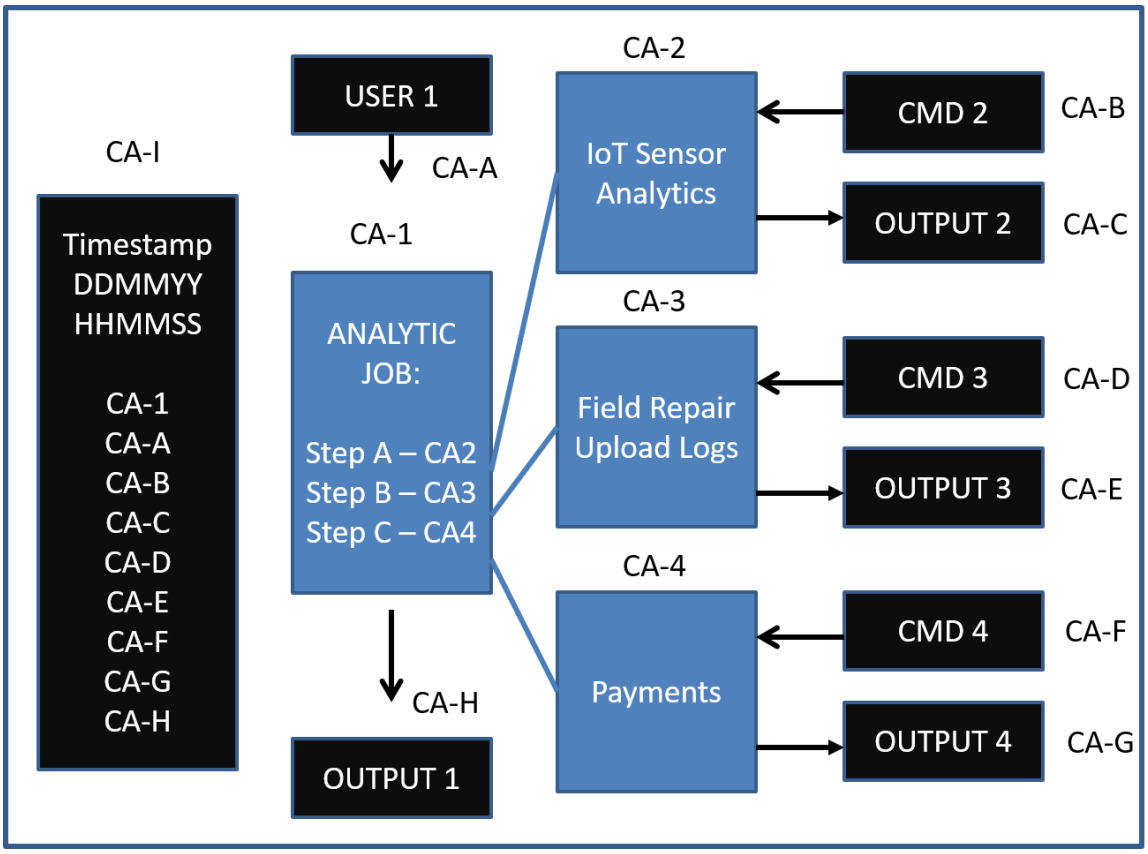


Figure 9 - Tracking User Analytic Jobs in Cloud

Figure 9 depicts an immutable log tracing the execution of User1's analytic job. This job combines 3 separate CMD (COMPUTING ON MASKED DATA)-style secure analytic operations. The user's level (e.g. "top-secret") can be stored, along with the data classification for each job (e.g. "confidential"). In this case, the steps are timestamped and stored with a content address to prove immutability. As hundreds and/or thousands of these operations are recorded, the resulting graphs are analyzed to determine if USER1 doesn't need a top-secret classification, or (for example), USER1 continually fails to try to run analytic jobs that operate on security classifications that are out of bounds.

These types of discoveries can result in changes to user levels and/or classification levels. If the analysis results in a conclusion that a data set can be safely downgraded to a lower-level designation (e.g. top-secret to confidential), it could save the company a significant amount on operational expenses for cloud computing resources.

The corporation may run a cloud-based job to calculate these recommendations and then produce a recommendation (e.g. a script) that can be used to configure gateways to use new classifications on specific data sets.

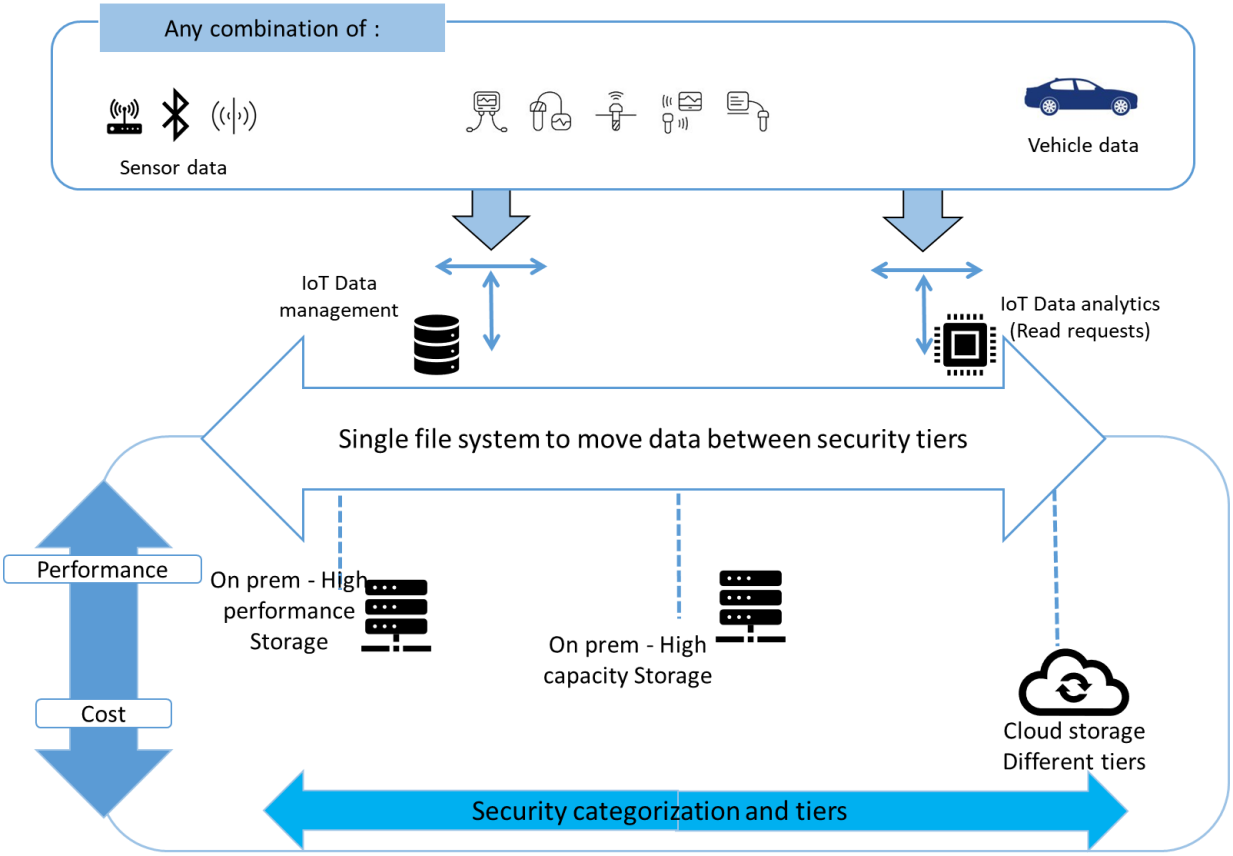


Figure 10: Data movement between security tiers

Automated Security Tiering

A data storage platform with the ability to efficiently store and manage massive structured and unstructured IoT data is required. We are going to propose a novel way to handle the Cyber-Physical Systems IoT Data in a federated way, to store and manage structured/unstructured data, which will facilitate determining the security tier needed. A smart IoT layer needed to be created based on combined algorithms to implement version management of unstructured Cyber-Physical Systems IoT data.

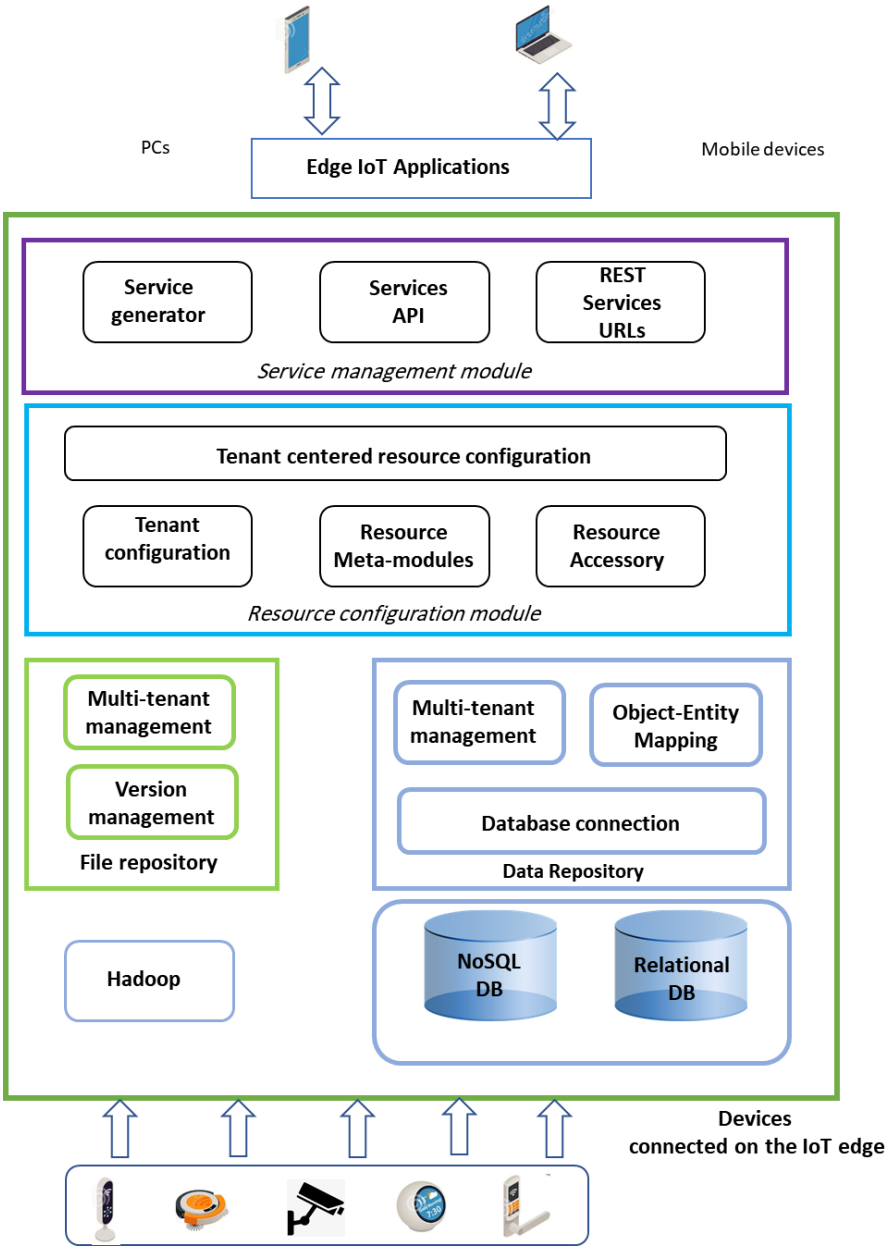


Figure 11: IoT Oriented Data Storage framework

The algorithm will do the following:

- Smart database allocation: dynamically choose the suitable data type to be stored in a certain level of security or encryption hardening. (for ex. Column stores Vs. row stores or in-memory databases for real-time analysis (water system sensors) Vs Hadoop for batch analysis (surveillance videos))
- Pre-data preparation: do in-storage early data preparation for batch processing depending on application type. For example:

a. Camera Data: do pre-structuring for the data making it ready whenever batch processing is required

b. Traffic data: do data cleaning, summarization, and storing it ready for batch processing anytime.

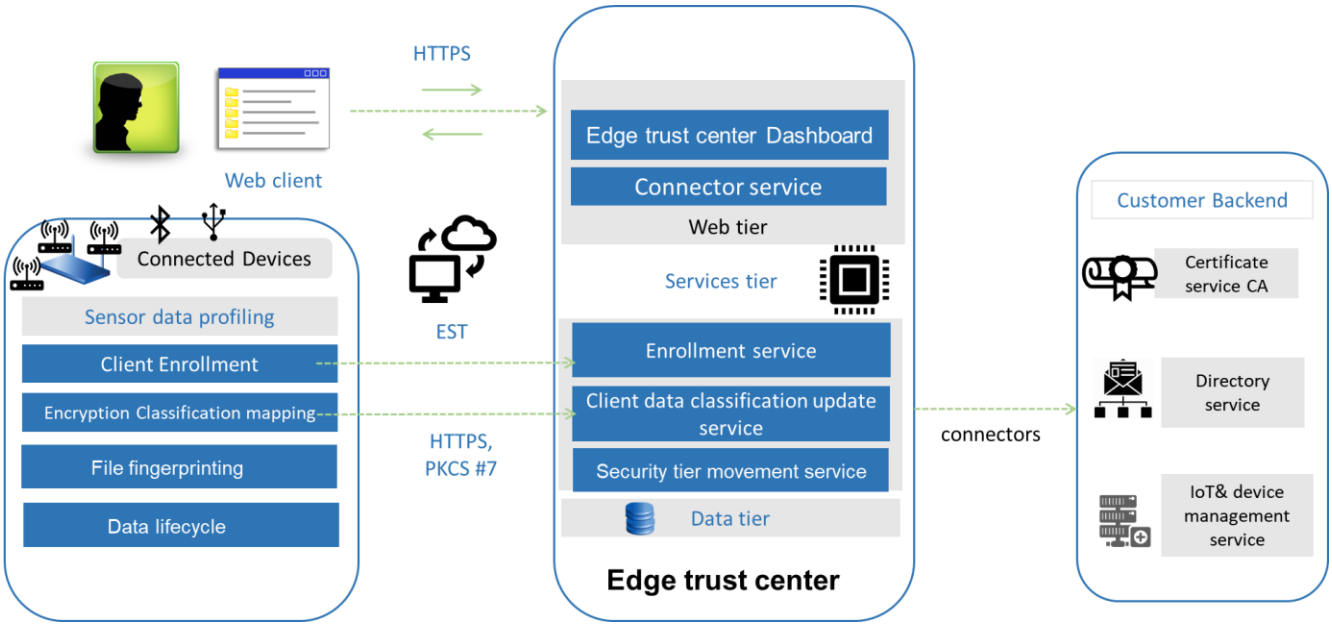


Figure 12: Security Tiering handling

In this diagram, the workflow is as follows

1. Client X creates xxx of files.
2. Edge security software starts to define the Level of sensitivity needed for the data derived from the sensor to assign the workflow to be used “Platinum, Gold, Silver, & Bronze”
3. Edge security software starts to classify the sensors' data based on the Sensors' metadata & the mapping sheet for future written files.
4. From the sensors' metadata, the sensor's profiling and file fingerprinting are performed.
5. Edge security software compares the information from the mapping configurator for the received sensor's data
   a) Data handling on-premises or in the Cloud.
   b) Data lifecycle grouping
6. Edge security software writes the data to the Cyber-Physical Systems IoT data cloud service with the ability to change its Security tier based on the policies updated from the mapping sheet or the administrator.

7. File fingerprinting is performed for every file received from the sensors while tracing the file's activities, versions, near-duplicates "for plain ones", and user's behavior.

To summarize (workflow created, Sensors' data aggregation, file fingerprinting through Sensors' metadata, Mapping sheet updated, Encryption level determined, Sensors' profiling, lifecycle and Security tier level determined, Data written to the cloud, Security tiers handling rules applied to the data on the cloud and data tier residency identified on the cloud).

This will be shown in more detail in the remainder of this article.

Dynamic Cloud Selection based on cost/performance

The policy engine will need awareness of multi-cloud encryption options and associated costs advertised by the provider. When Cyber-Physical Systems data has been classified, the gateway can search for advertised capabilities that map to that level. This may result in multiple provider options that may present options to the gateway. The gateway can choose from similar/identical services being offered by leveraging:

- Cost comparisons
- Distance (latency) comparisons
- History of uptime with a given provider
- Etc.

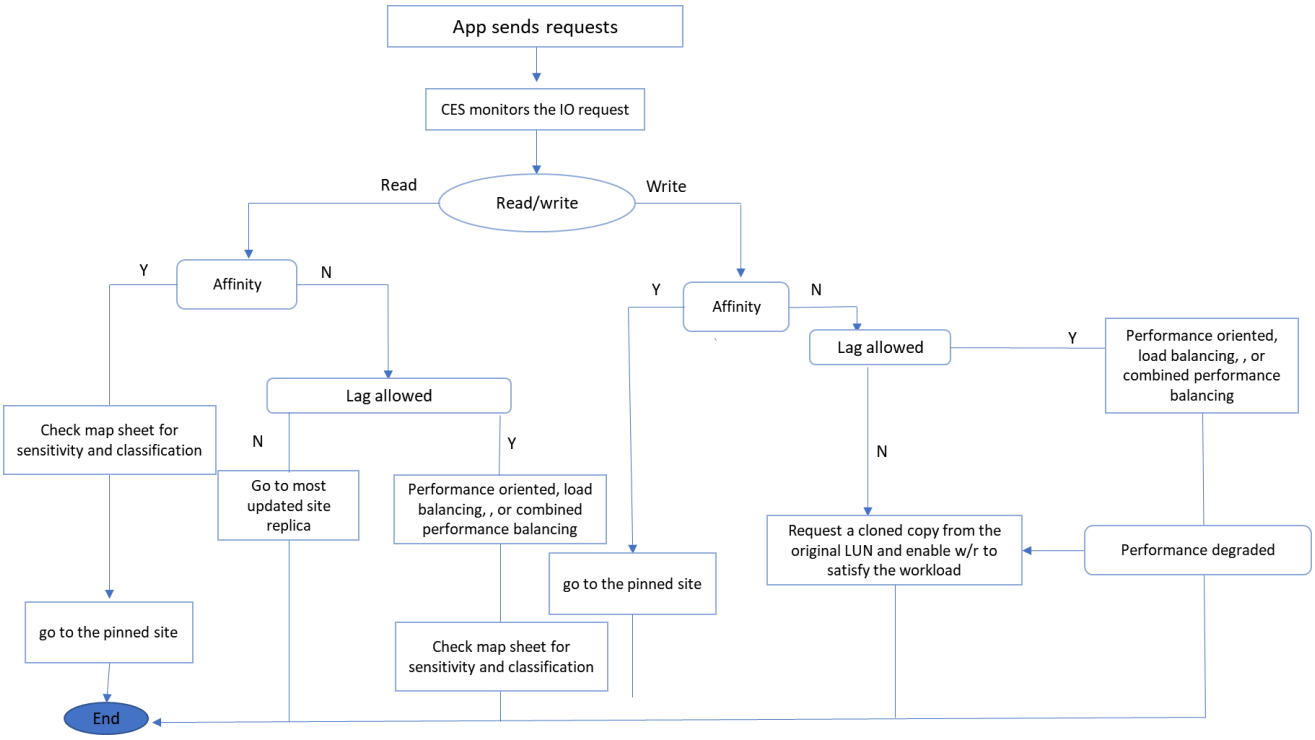


Figure 13: Data Flowchart

Flowchart

The novel approach here uses diagrams for Big Data computation to determine the level of encryption and how the data should be analyzed. Figure 14 highlights the choice to store streaming of the CSP data.

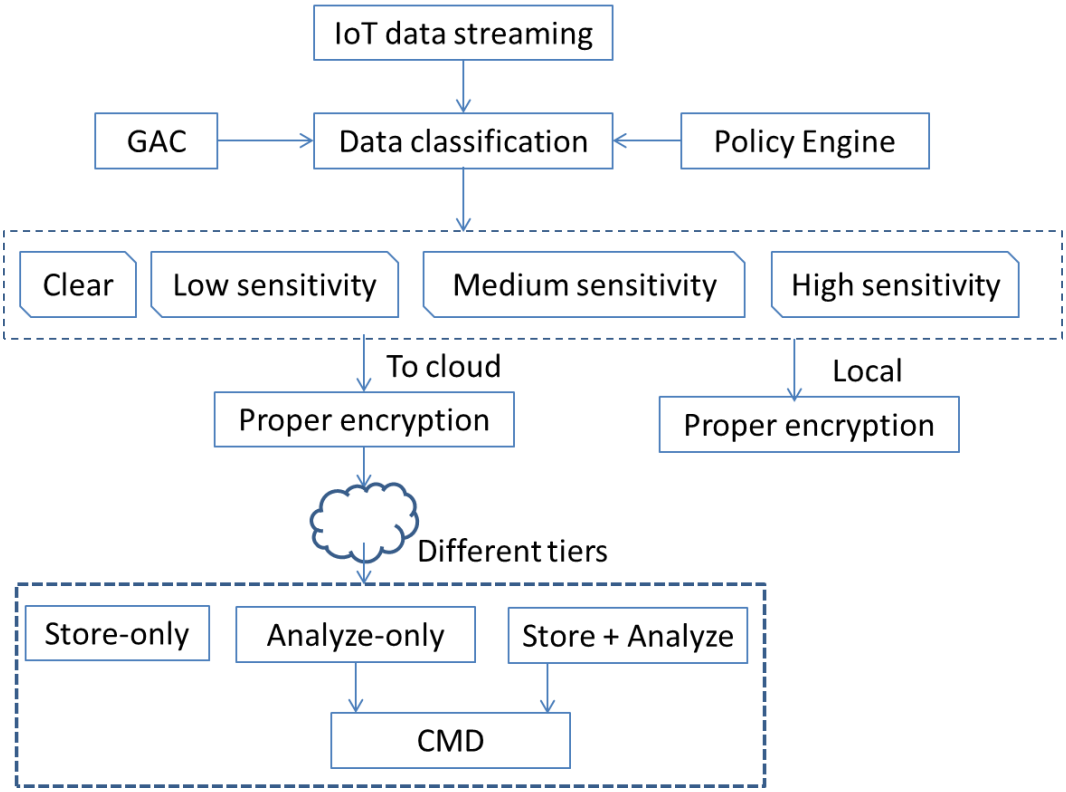


Figure 14: Rule for sensitivity based IoT data moving and analytics

Coordination and advanced control policy

The solution we propose is based on an orchestration layer where a routing layer essentially coordinates with replication system in terms of replication topology to ensure the proper backed storage is correctly selected. The replication load pressure and replication lag status, with such info exported to or collected at close-to-App routing layer with the orchestration layer, can enable advanced control **for purposes like replica-aware, performance-aware, lag-aware, encryption aware, or affinity etc.** Thus, Physical Cyber System App requests can be automatically and intelligently routed to the proper backend site for higher performance, load-balance, etc. See Figure 15 and details of the policy:

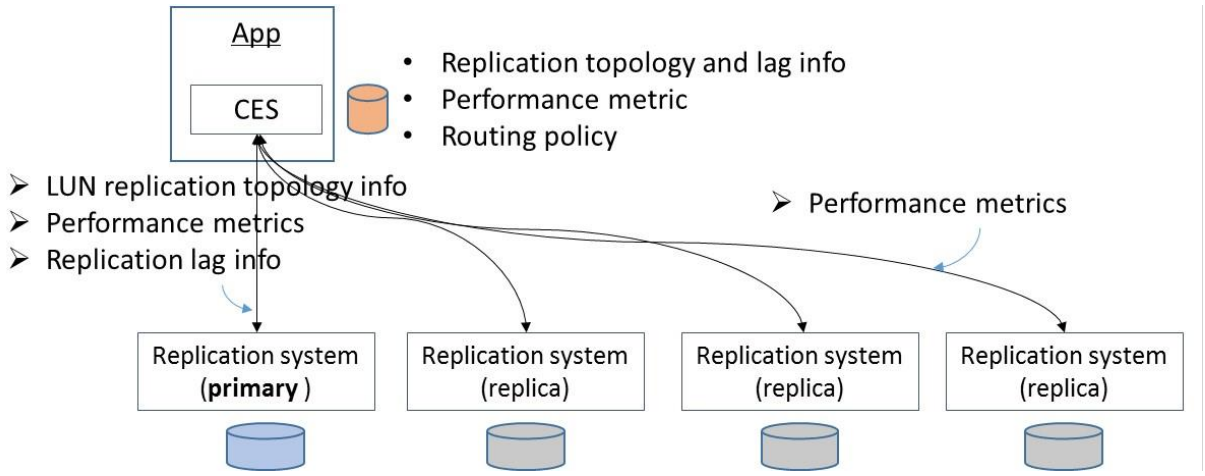


Figure 15: **Flexible and advanced routing policy, and coordination between routing and replication**

1. Replica-aware: Replication systems like Dell EMC RecoverPoint (RP) configures or changes its primary and target and the routing layer may regularly pull or push such information, then update its local database (as metadata). With that, an App from a given location could have more flexible choices to access data either in primary or replica site. If a site is lost or unavailable, RecoverPoint (RP) system (primary site) may also push an urgent message to the routing layer.

2. Performance-aware: The routing layer running on specific host together with App could periodically check the access performance to all available sites, including primary and replica, latency and bandwidth. For example 1) in background, it may ping the primary/replica site to collect networking latency then rank the latency (here latency is more about networking), or 2) in background, it may read some test data from primary or replica site and measure both latency and bandwidth (here performance covers both networking, processing and disk I/O) or 3) for real IO to specific site, it can also measure latency and bandwidth. Those performance data later can be used for read traffic routing, or combined with other policies (as described below) like balancing, affinity and locality.

3. Load-aware: The routing layer may also balance its issued IO to multiple sites. For example, it may take round-robin fashion to all sites or take above performance measurement as weight plus IO amount to a specific site, then use a weighted balance.

4. Lag-aware, Locality and Affinity: The routing layer can also support advanced settings, e.g:

- **Lag-aware**: If replication or CDP system configured multiple replica sites which could be common for enterprise users, even all replica sites running in CDP mode, data replication performance from primary to each replica site may be different, such as due to unpredictable interconnection, which leads to different replication lag. Replication system

(the primary) could monitor such lag (RP already did this), then report it to the routing layer (or routing lay could push the info). Based on that, routing layer may weight the policy, such as if some App needs more real-time data, they could choose primary or the most updated replica site.

- **Locality and consistency**: Write request shall always go to the primary site (i.e. data generation from sensors/devices). Most requests hereafter such as data analytics are read-intensive, but in case there's read-after-write (such as read your own write), then such patterns can be detected (read previous write within a time threshold such as 30sec), and route the read to the primary site as well. This also ensures data consistency (in case write has not replicated to replica site), and likely get higher performance as data probably is cached in networking (CDN) or primary site memory.
- **Affinity**: user may also set pin affinity, so that specific App may always access data from a specific site, such as for cost or compliance purpose.

Steps to initiate and perform initial setup

- ✓ Registration of the sensor or edge into a pre-defined rule based on the designed solution.
- ✓ Network configurations.
- ✓ Creating the SSH key.
- ✓ Creating the FTP public directory to host the SSH public key.
- ✓ Start the Avahi Daemon (Zeroconf) and announce the public FTP path on the network.
- ✓ Creating a client certificate request then send it to the IoT vendor cloud for signing.
- ✓ Install the signed client certificate received by the vendor or any recognized Certificate Authority.
- ✓ The signed client certificate is received. Now the node requests to the security agent are going to be XMLRPC or REST.

1. Once the client joined the network it will search for the security agent's public key through the Zeroconf protocol, and the Edge security software will check the mapping configurator for the previously chosen profile unless the administrator chose a specific tier.
2. Get the key from the Agent's public FTP directory then set it up so the agent can access it without a password to push/pull files and remotely execute commands.
3. Once the agent's SSH key is set up, this action will trigger execution of the registration program that will generate a client SSL certificate request and send the request to the agent's IP address/hostname which was retrieved using the Zeroconf protocol in the step of pulling the agent's public SSH key.
4. The administrator approves/edits the tier category and reviews the workflow.
5. The mapping configuration engine registers the administrator's action and feeds the system for future use.
6. If accepted, the Security Engine will sign the request and send the client certificate to the sensor node using either XMLRPC or REST so in the future they can communicate securely.
7. If denied, the sensor node will be blacklisted. To re-add it, it should be removed from the blacklist.
8. The signed client certificate is received. Now the node requests to the security agent are going to be XMLRPC or REST which will include the node serial number/firmware version with every request it issues to the security agent.
9. Rule configurator takes care of the rule definition.

The approach above suggests a ‘tagging’ approach. Other approaches could be used, e.g. a config file).

Benefits of the approach

- Novel sensitivity-based data moving, tiering, and analytics
- Flexible granularity configuration and access control, a trade-off between computing overhead and sensitivity level
- Can be embedded as an add-on technique to the IoT cloud provider to harden the current data security and reduce cost for end-users.
- Can be provided as a SAAS solution (sensitivity tiering as a service), which can be consumed by multiple tenants, pay per usage.

Conclusion

Cyber Physical systems refer to a wide range of physical devices that create and share data. They are considered one of the major outputs of digital transformation. Given the inclination of Data Center leaders to use a hybrid model to handle data floods, we should consider the different ways to ensure we are optimizing how we are dealing with our data – from the time of data creation to its decommission – to ensure data life cycles have been correctly handled. Most data are created outside the data center. We now have edge computing, and data storage can be hosted on prem, the cloud, or a hybrid model of the two. This compels an intelligent way to deal with this data and its encryption levels to adapt with read-intensive applications in a smart and effective way.

Glossary

Term	Definition
IoT	Internet of Things
M2M	Machine to machine
CMD	Computing on masked data
MPC	multi-party computation
CES	Cloud-enabled service

Table of Figures

References

[1] “IoT Explained diagram.” Accessed: Mar. 20, 2022. [Online]. Available: https://www.tibco.com/reference-center/what-is-the-internet-of-things-iot

[2] “Concept map.” Accessed: Mar. 20, 2022. [Online]. Available: https://cyberphysicalsystems.org/

[3] “Devices.” Accessed: Mar. 20, 2022. [Online]. Available: https://www.opennaukri.com/input-and-output-devices/

[4] “Data Classification.” Accessed: Mar. 20, 2022. [Online]. Available: https://www.sealpath.com/blog/automate-data-classification-protection/

[4] “ATT is Reinventing the Cloud Through Edge Computing,” http://about.att.com/story/reinventing_the_cloud_through_ edge_computing.html.

[5] “Verizon’s cloud-in-a-box pushes the edge with OpenStack,” https://siliconangle.com/blog/2017/07/17/ verizons-cloud-box pushes-edges-openstack-openstacksummit.

[6] Y. Li, Y. Chen, T. Lan, and G. Venkataramani, “Mobiqor: Pushing the envelope of mobile edge computing via quality-of-result optimization,” in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017, pp. 1261–1270.

[7] “Cisco Global Cloud Index: Forecast and Methodology, 2016-2021 White Paper,” https://www.cisco.com/c/en/us/ solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html, 2018.

[8] “IDC Directions 2017: IoT Forecast, 5G & Related Sessions,” http://techblog.comsoc.org/2017/03/04/ idc-directions-2017-iot-forecast-related-sessions/, 2017.

[9] 1 https://www.link-labs.com/blog/iot-agriculture

An IoT-Oriented Data Storage Framework in Cloud Computing Platform IEEE paper

[10] https://support.emc.com/docu96462_CloudBoost-19.2-Release-Notes.pdf?language=en_US

[11]Understanding and Selecting Data Masking Solutions Creating Secure and Useful Data. Securosis, L.L.C

[12]Computing-on-Masked-Data-a-High-Performance_good-in-security. IEEE paper.

[13]CryptDB: protecting confidentiality with encrypted query processing. In Proc. MIT, ACM SOSP, 2011. Open-source, project homepage, Used in Google Encrypted BigQuery; SAP Project SEEED, Used for Healthcare data in the cloud

[14] Big Data Analytics over Encrypted Datasets with Seabed, OSDI16, Microsoft

[15] Top 10 big data security and privacy challenges. Cloud security alliance.

[13] CSI infographic big data.