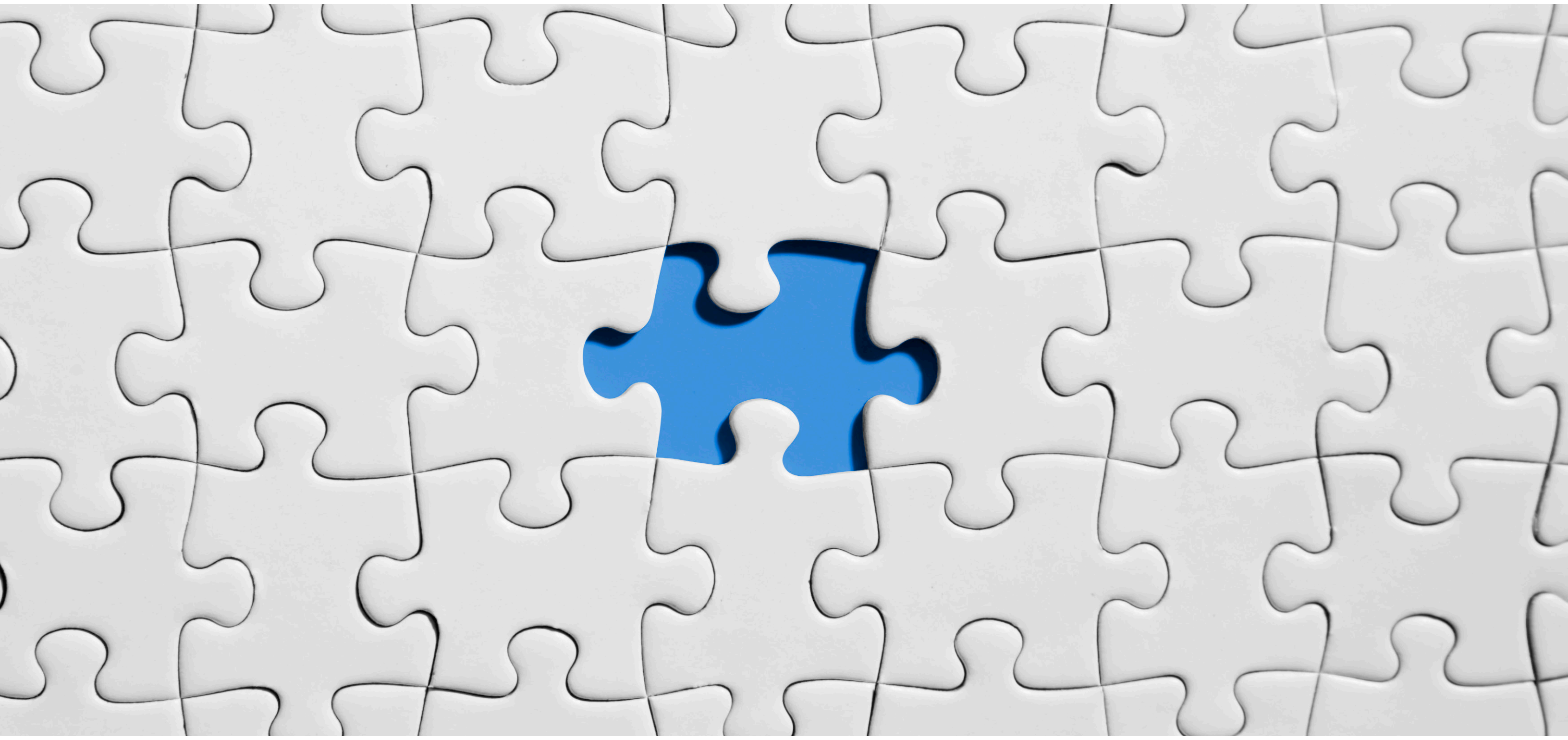
PROJECT ALVARIUM:
THE FUTURE OF EDGE DATA



Steve Todd

Dell Technologies Fellow

Dell Technologies

Steve.todd@dell.com

Knowledge Sharing Article

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

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

1 Introduction

This Knowledge Sharing paper introduces Project Alvarium, an open community of companies and technologists committed to collaborating on an emerging technology known as a Data Confidence Fabric (DCF). Readers will learn about the technology and gain an understanding of the community's goals.

Dell Technologies' historical emphasis on trusted data delivery serves as the backdrop for this paper. A recent corporate press release conveys the company's commitment to data.

On November 12, 2019, Dell Technologies announced a set of "moonshot" goals for 2030. With a keen focus on "Social Impact," the company "will use its global scale, broad technology portfolio, and expertise" to (1) advance sustainability, (2) cultivate inclusion, and (3) transform lives. [1]

At the heart of each goal lies an indisputable common denominator: data. The press release highlights the industry's lack of tools to properly manage the world's data:

> *"Ethics and privacy are foundational to Dell Technologies' corporate and social impact strategies and are essential to executing against the 2030 goals. The company is setting the pace in privacy and transparency by fully automating data control processes making it easier for customers to access, delete or share their personal data. To amplify team members' and partners' passion for ethics and integrity, the company will use digital tools to make it easier to get insights from, measure and monitor compliance issues using digital data."*[2]

Protecting the world's mission-critical data is nothing new for Dell Technologies, referred to as the "world's largest data protection vendor."[3]

Although not exclusively, the company has largely supported this claim by protecting enterprise data (data born and managed within an enterprise network). Highly-skilled security teams deploy, configure, and oversee products from Dell Technologies that deliver trusted enterprise data to applications.

What does it mean for an application to trust the delivery of data? To answer this question, we turn to the following definition of "trust" from the National Institute of Science and Technology (NIST).[4]

> *Trust: A characteristic of an entity that indicates its ability to perform certain functions or services correctly, fairly, and impartially, along with assurance that the entity and its identifier are genuine.*

The assurance of "genuine" data delivery better positions enterprise applications to perform functions and services correctly. The word "genuine" can mean different things to different applications, e.g.:

- Always available
- Correct
- Valid

- Timely
- Private/confidential
- Compliant
- Secure
- Etc.

Enterprise IT architects tune and manage enterprise data delivery so that their applications perform correctly when processing that data.

The delivery of genuine data will become harder as 2030 approaches. IT departments will be increasingly challenged to deliver trusted edge data (data that is born, analyzed, and managed outside of a traditional enterprise network) to applications. Edge data sources can include IoT devices, sensors, gateways, cell phones, augmented reality/virtual reality devices, and employee laptops/tablets. Data from these devices originates outside of an enterprise perimeter. The data exists beyond the traditional reach of enterprise security teams.

Indeed, IDC predicts[5] that in the next five years, the need for enterprise-class IT at the edge will grow dramatically.

> *By 2023, over 50% of new enterprise IT infrastructure deployed will be at the edge rather than corporate datacenters, up from less than 10% today; by 2024, the number of apps at the edge will increase 800%.*

Dell Technologies' success in delivering trusted (i.e., genuine, per NIST), mission-critical data to enterprise applications does not always translate to every form of edge data. There are many reasons for this (discussed below). The ecosystem over which edge data travels is of primary concern. This ecosystem spans multiple heterogeneous systems, networks, vendors, and geographies.

It is for this reason that Dell Technologies includes data ethics and privacy as one of its moonshot goals. Enterprise data is centralized; edge data is decentralized. Proprietary data protection architectures cannot always operate on data that spans the edge and the enterprise. An augmented approach is required, one that can handle edge use cases in ways compatible with the enterprise.

While enterprise architectures cannot solve all edge data use cases, they can certainly inform them.

In the spirit of delivering trusted data, Dell Technologies is donating code and expertise to the LINUX Foundation as part of Project Alvarium[6].

> *Project Alvarium will focus on building the concept of a Data Confidence Fabric (DCF) to facilitate measurable trust and confidence in data and applications spanning heterogeneous systems. The project will be seeded by code from Dell Technologies, with support from industry leaders including Arm, IBM, IOTA Foundation, MobiledgeX, OSIsoft, Unisys, and more.*

The term "alvarium" is a Latin word referring to a beehive. The project is so-named to emphasize the importance of workers (bees) collaborating to execute a common task:

creating a community specification (a hive) that builds trust into edge data delivery. Cooperation and trust occur within the beehive.

Project Alvarium is defined as *an open community that is building Data Confidence Fabric technology.*

In response to the growing importance of delivering trusted data from the edge, Dell Technologies built the first example of a Data Confidence Fabric.[7] A DCF is defined as follows:[8]

A Data Confidence Fabric delivers trusted data to applications with measurable confidence.

It is important to note that a DCF measures confidence in the data delivery mechanisms (as opposed to just the data itself).

How does a DCF measure the level of trust applied to the delivery of data? It does so by (a) specifying how trust should be "inserted" during delivery and (b) providing an equation that "scores" the delivery confidence. Figure 1 depicts an example DCF configuration and how it might apply to the delivery of data in an edge ecosystem.

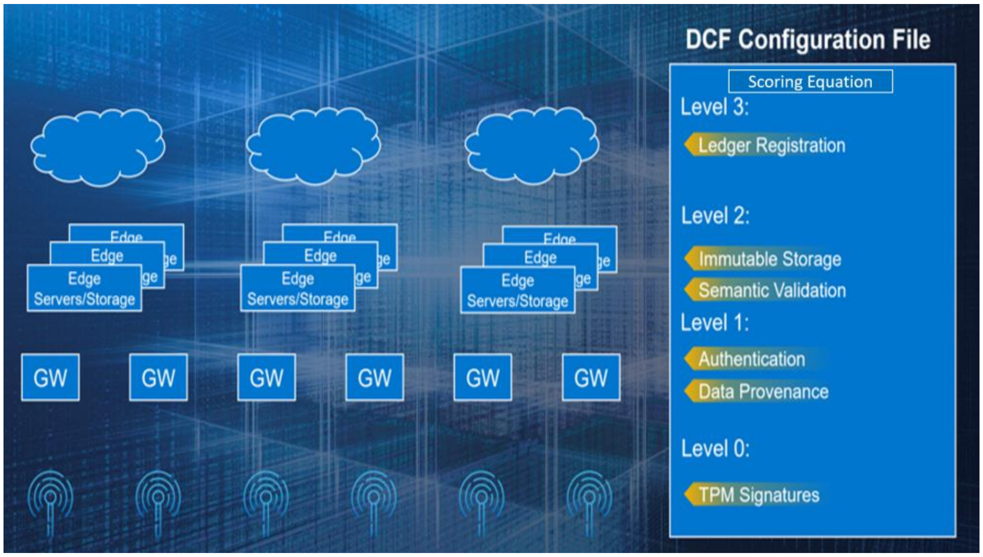


Figure 1 - Data Confidence Fabric Configuration and Equation

In the enterprise, it is the security team that assures trusted data delivery. For edge data, a confidence score will play that role. The DCF configuration file pictured above enables a definition of what constitutes the delivery of trusted (or "genuine," per NIST) data across an edge ecosystem. Hardware and software technologies participate in the fabric's attempt to deliver data under this definition (and are scored on their success or failure). Multiple DCFs can be overlaid across an edge ecosystem and provide differing levels of confidence.

During the next decade, with billions of devices sending edge data across heterogeneous networks, the ability to trust that data's delivery with a measurable degree of confidence will be paramount. No one vendor can own or generate this trust; a community of collaborators must create a standard.

This paper explores the future of edge data and how the Project Alvarium community can advance the adoption of Data Confidence Fabrics.

Section two of this paper will highlight the differences between edge data and enterprise data. The complexity of geographically-distributed systems and their large attack surface will be emphasized. The section concludes with a Forrester Consulting survey highlighting that the industry struggle is real, and the need for a holistic solution is substantial.

The third section will review the enterprise approach of inserting trust into mission-critical data and discuss how those techniques might apply to edge data.

Section four provides more detail on building a DCF. An overview of the DCF example donated by Dell Technologies to the Project Alvarium community is provided. This overview will describe DCF configurations and scoring techniques in more detail.

Section five describes a long-term benefit of DCFs: the ability to increase the value of data through the avoidance of fines and monetization in emerging data marketplaces.

The conclusion of the paper summarizes the work and includes an invitation to join the Project Alvarium community.

2 Checklist for Edge Data

Figure 2 paints a picture of bringing edge data and business applications together. This picture helps identify a checklist of items that, when in place, improve the trustworthiness of data delivery.

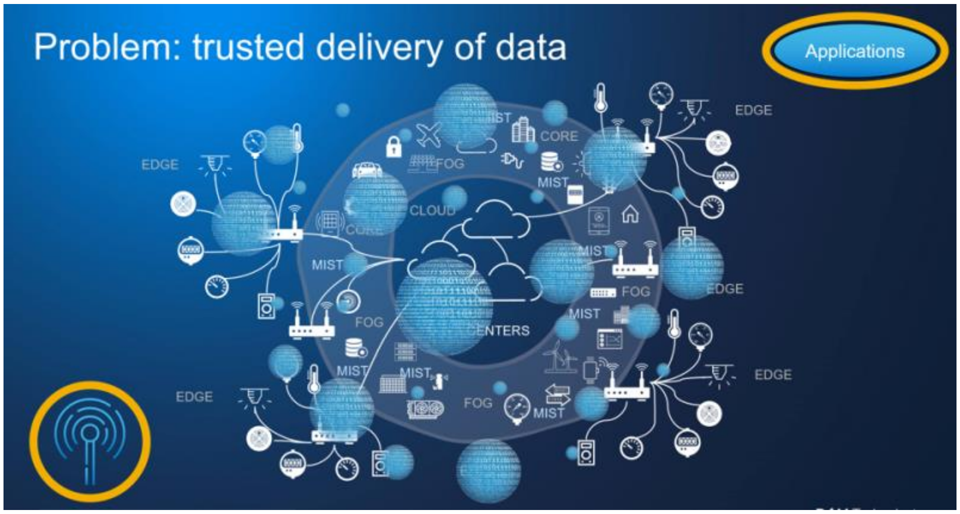


Figure 2 - The Scope of Delivering Trusted Edge Data

Figure 2 depicts applications (top right) that wish to analyze data coming from a wireless device (bottom left, e.g., cell phone, laptop, IoT sensor, etc.). There are two problems, if left unsolved, that can result in adverse economic impact[9] on the corporation deploying the applications.

1. Plumbing problem: how are the data and applications brought together across heterogeneous networks and infrastructure?
2. Trust problem: how do the applications know that the delivered data is trustworthy?

Solutions to the first problem are well on their way. As IDC points out, the number of applications running on the edge is estimated to grow by 800% in less than five years. Established infrastructure vendors are already figuring out ways to transition their existing technologies to the edge, and startups will cover gaps. A substantial amount of plumbing is already in place.

Trusting the delivery of edge data (e.g., data coming from the device in Figure 1), however, is not as straightforward.

Why is this?

Enterprise data has historically been generated and managed within the walls of a corporation. Consider an application that retrieves data from an enterprise storage system. Security best-practices in enterprise application deployment and network management ensure that applications analyze data within a boundary that doesn’t compromise trust.

When *edge* data arrives at an application, however, a significant number of unknowns exist. These unknowns are described below.

2.1 Hardware root of trust

Applications that ingest edge data do not currently have robust guarantees about the trustworthiness of the data’s source. Data can be sourced from a multitude of devices. An application may find itself operating somewhere in the middle of a collection of nodes that are moving data from a device to a cloud. These applications must evolve to receive assurances about the hardware environment in which the data initially originated.

Edge devices that implement a “root of trust” can be helpful. NIST provides a definition:[10]

> *Roots of trust are highly reliable hardware, firmware, and software components that perform specific, critical security functions. Because roots of trust are inherently trusted, they must be secure by design. As such, many roots of trust are implemented in hardware so that malware cannot tamper with the functions they provide. Roots of trust provide a firm foundation from which to build security and trust.*

What are some examples of critical security functions that a hardware root of trust solution can provide? Consider a gateway device (e.g., a Dell Gateway 3000 Series) that contains an embedded Trusted Platform Module (TPM) chip with root of trust capabilities.

A DCF configuration file can specify that all gateways in an edge ecosystem implement the following three root of trust features:

1. Signatures on device data. The gateway uses the TPM's unique private key, assigned at manufacturing, to sign any data that passes through.
2. Secure boot. The TPM chip on the gateway verifies that the software (e.g., drivers, operating systems, etc.) has not been tampered with or changed.
3. Secure onboarding. Management software that oversees the health of gateway devices can perform a handshaking protocol with the TPM as a way of on-boarding the gateway after initial installation.

As part of a Data Confidence Fabric, the gateway described above confirms the presence/activity of these three roots of trust features. It attaches DCF metadata as a form of assurance to upstream applications.

This DCF metadata also enables the creation of a confidence score (e.g., if the device was not securely onboarded, the confidence will be less than one hundred percent).

There are other forms of hardware roots of trust that can be used (in addition to TPMs). One example is Trusted Execution Environments (TEEs) that securely run applications in the context of a Trusted Application Manager (TAM).

Depending on the use case, a DCF can require the existence of specific hardware root of trust capabilities, and then measure whether those capabilities were in operation during data delivery.

The Industrial Internet Consortium (IIC) is paying close attention to the security of the endpoint devices described above. IIC has generated a best practices document that strongly recommends the use of root of trust and defines levels of trust according to a standard known as IEC 62443.[11] DCF scoring methodologies can incorporate trust levels coming from standards such as IEC 62443.

2.2 Security profile at ingest

Enterprise data storage systems have robust security monitoring (e.g., a Security Operation Center may monitor activity on networks, endpoints, platforms, etc.). In these environments, only authorized access to data is allowed.

Devices that generate and process edge data may not support enterprise-class authentication approaches. A lack of authentication puts edge data at risk from the unique threats (described below) that can occur outside of the enterprise perimeter.[12]

Researchers at the Carnegie Mellon University (CMU) Software Engineering Institute are working on solutions that are specific to the constraints and threats of an IoT environment.[13] They are developing methods for authentication and authorization for IoT devices that consider:

- High-priority threats of tactical environments such as node impersonation and capture

- Operations in disconnected, intermittent, limited (DIL) environments
- Resource constraints of IoT devices

Leaving authentication unimplemented may result in *any* client (authorized or not) accessing data at will.

The researchers at CMU have proposed architectures that extend enterprise authentication techniques (e.g., OAuth + JWT) to edge devices. Figure 3 highlights their architectural proposal[14].

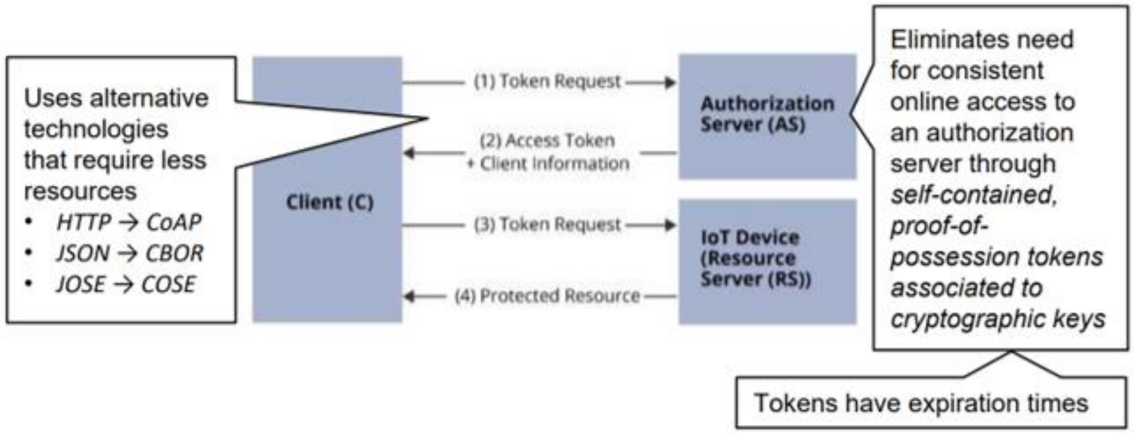

Uses alternative technologies that require less resources
• HTTP → CoAP
• JSON → CBOR
• JOSE → COSE
Client (C)
(1) Token Request
(2) Access Token + Client Information
(3) Token Request
(4) Protected Resource
Authorization Server (AS)
IoT Device (Resource Server (RS))
Eliminates need for consistent online access to an authorization server through self-contained, proof-of-possession tokens associated to cryptographic keys
Tokens have expiration times

Figure 3 - Protecting IoT Data Resources via Authorization Servers

When applications access edge data, solutions to the following concerns may be critical:

- Have robust authentication/authorization techniques verified all applications attempting to access the data?
- Are failed/successful logins audited?
- Who or what has been accessing data from this device?
- Has access been correlated to specific users?

Currently, there is no way for an edge application to know whether the data under analysis flowed through an authenticated channel.

If a DCF requires authenticated channels, assurance of authentication from edge devices can be requested, scored, and communicated to an application. Similarly, a DCF may specify that audit logs describing access attempts be kept and made available.

2.3 Detection of data tampering

Prevention of tampering in an enterprise context is often the responsibility of an IT department; corporate security practitioners restrict physical and programmatic access to critical data. IT departments have a wide variety of tools from which to choose.

Enterprise-class storage systems can guard against data tampering by using a variety of different cryptographic techniques. For example, content-addressable storage systems (e.g., Elastic Cloud Storage (ECS) by Dell EMC) calculate a hash during data creation and validate the hash when the content is requested. The use of a hash allows an application to determine whether the data has unexpectedly changed (e.g., via tampering).

In edge environments, however, there may not be any comprehensive IT department or security oversight. Enterprise-class, immutable storage systems (e.g., ECS) cannot be placed (and managed) everywhere on the edge.

One of the emerging technologies offering solutions in this space is the InterPlanetary File System (IPFS). This solution borrows from edge-friendly BitTorrent technology in which lightweight nodes communicate using peer-to-peer, decentralized messaging. IPFS (like enterprise-class content-addressable storage systems) uses hashes to store and retrieve data. These hashes can also assist in the detection of tampering. What follows is a definition of IPFS[15]:

> *The InterPlanetary File System (IPFS) is a protocol and peer-to-peer network for storing and sharing data in a distributed file system. IPFS uses content-addressing to uniquely identify each file in a global namespace connecting all computing devices.*

The deployment of a decentralized, open-source object store holds great promise to detect tampering. If device data immediately enters a “nearby” content-addressable object store and receives a hash value, the ability to tamper with the data minimizes.

It is currently not possible, however, for an edge application to know whether this form of protection occurred.

When using a DCF, however, an application can inspect DCF metadata to determine if hashes protected the data at some point on its journey. Note that this approach can be augmented with other techniques (such as per-message security) to help protect the data end-to-end.

If the DCF metadata is granular enough, examining the type of hash algorithm used can also influence the confidence score.

2.4 Data ownership and Governance

In an enterprise application scenario, the corporation itself often owns the data. There is no need for an application to worry about ownership. In some cases, a corporation may purchase or license data from another source. Either way, applications are typically unaware of data ownership in an enterprise context.

For edge data, ownership is more heavily nuanced and complex. When machines generate data, ownership can be murky. Who owns the data? The device vendor? The purchaser of the device? The user? Who owns the insights originating from the data?

Corporations, device manufacturers, or consumers may own the data. If applications lack knowledge of this ownership, disastrous business results can occur. For example, if the owner of edge-generated data lives in California or Europe (areas with strict data privacy laws), an application may introduce considerable risk by using the data in violation of regulations relevant to those geographies.

As the industry progresses towards 2030, the importance of consumer ownership of data will only increase. For Dell Technologies to achieve privacy goals, applications must develop business logic that knows how to handle ownership nuances and regulations. At the same time, these same applications must still recognize corporate ownership of data.

Several different technologies are looking to address these issues.

For corporate data, vendors manufacture devices (e.g., gateways) with ownership metadata "baked into" the device. Intel, for example, uses the approach shown in Figure 4 during secure device onboarding (SDO). SDO can establish a foundation of ownership during installation[16].

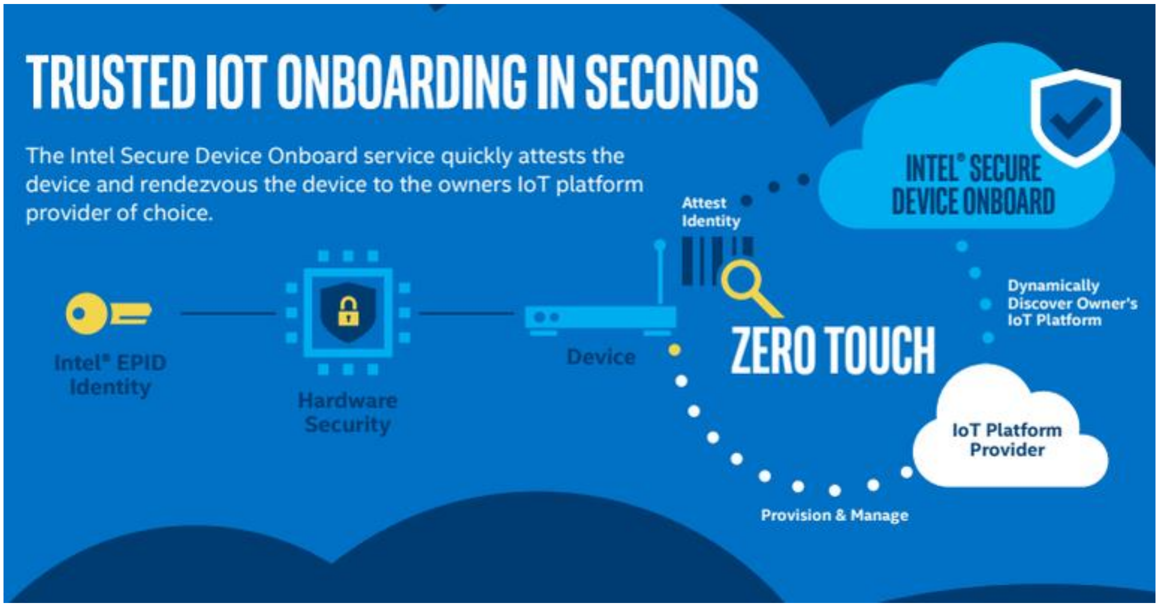

TRUSTED IOT ONBOARDING IN SECONDS
The Intel Secure Device Onboard service quickly attests the device and rendezvous the device to the owners IoT platform provider of choice.
Intel® EPID Identity
Hardware Security
Device
Attest Identity
INTEL® SECURE DEVICE ONBOARD
ZERO TOUCH
Dynamically Discover Owner's IoT Platform
IoT Platform Provider
Provision & Manage

Figure 4 - Embedding Ownership into Edge Devices

For consumers, there are emerging forms of identity that help establish data ownership. One such effort is known as Decentralized Identities (DIDs). DIDs are also called "self-sovereign identity." They are viewed as a promising solution to the General Data Protection Regulation (GDPR) issues relating to consumer data in Europe[17] (and the California Consumer Privacy Act as well).

Microsoft has been active in the development of decentralized identities and has proposed an architecture to move the industry forward.[18]

> *A new form of identity is needed, one that weaves together technologies and standards to deliver key identity attributes—such as self-ownership and censorship resistance—that are difficult to achieve with existing systems. To deliver on these promises, we need a technical foundation made up of seven key innovations—most notably, identifiers that are owned by the user, a user agent to manage keys associated with such identifiers, and encrypted, user-controlled datastores.*

This statement by Microsoft, when applied in the context of data ownership, means that consumers will interact with different entities (e.g., applications) via their decentralized identity (as opposed to a corporately-assigned identity). Figure 5 depicts Microsoft's vision.[19]

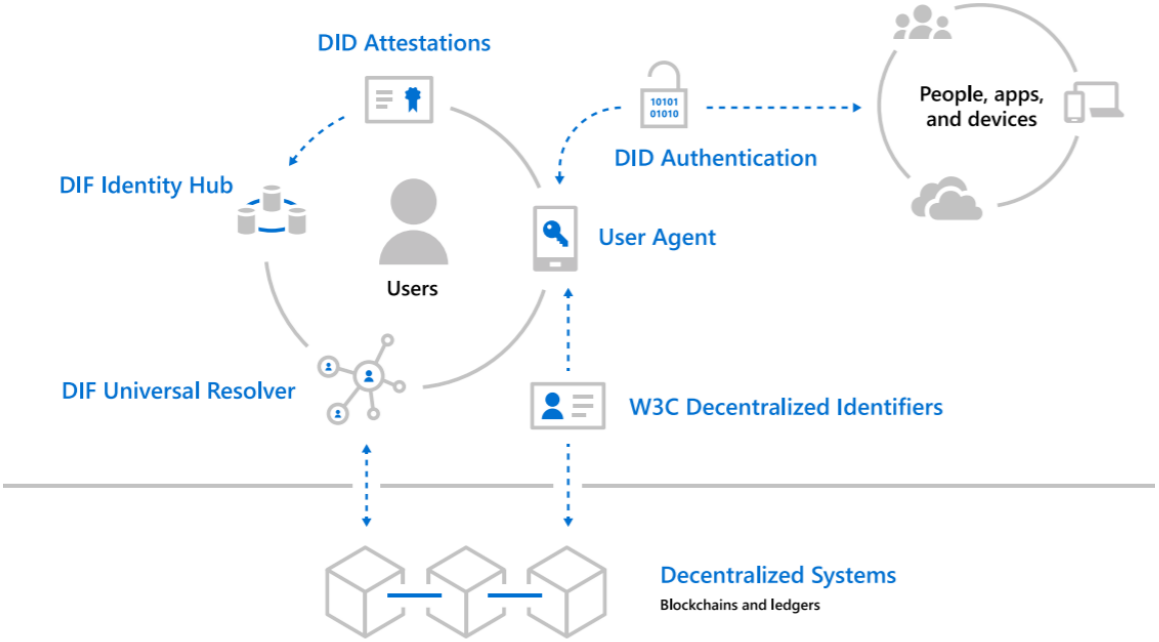


Figure 5 - Microsoft Proposed Architecture for Decentralized Identities (DIDs)

The bottom layer of Figure 5 shows identity information stored in a decentralized ledger. These ledgers exist much closer to the edge than traditional enterprise identity services (e.g., centralized AD or LDAP servers).

The Decentralized Information Group at MIT CSAIL is solving (among other things) data ownership problems[20]:

We're exploring how to radically change the way Web applications work today, resulting in decentralized architectures that enable true data ownership; working on frameworks that ensure information can be shared, used, and manipulated in a way that is compliant with regulation, business rules, social norms, and user preferences; as well as investigating methodologies to make algorithms trustworthy and accountable.

One of the data ownership solutions at MIT is known as Solid. Professor Tim Berners-Lee (inventor of the World Wide Web) leads the Solid project.[21]

Users should have the freedom to choose where their data resides and who is allowed to access it. By decoupling content from the application itself, users are now able to do so.

While MIT Solid and DIDs are both based on W3C standards, it is unclear whether the two implementations will converge. If a DCF specifies that identity/ownership should always be associated with edge data, it must be able to support either of these implementations.

Data ownership highlights a significant difference between edge and enterprise data: governance and compliance. Corporations have invested significantly in the use of enterprise-class tools to prove their compliance with corporate, national, and international regulations.

These tools often assume that the corporation owns the data. This won't work for edge data.

A DCF can be helpful in this regard by requiring the registration of ownership during the data's journey. The owner of the information has the right to expect that the handling of the data will comply with all regulations, and a DCF can specify which policies must be adhered to when the data is born.

Also, when the owner of data is well-known, the value of that data can be more accurately expressed (e.g., clearly-established ownership can remove friction when selling data into a data marketplace).

Standard data ownership semantics, however, continue to be fragmented across the industry. Lack of ownership semantics can result in significant fines (see Figure 19 for a depiction of a DCF solution to avoid these fines).

2.5 Point of Origin Capture

The history of data's creation is often not documented in an enterprise setting. In an edge context (e.g., a temperature sensor on a manufacturing floor), knowing the point of origin and additional information can be critical (e.g., which one of my devices is currently overheating).

The data that originates from these edge devices may not contain information about its surrounding geography or compute infrastructure. It is, therefore, incumbent upon the ingest environment to attach trustworthy provenance information. Provenance is defined as follows.[22]

Data provenance provides a historical record of the data and its origins.

Provenance about the origin of edge data can provide a rich context to an application (and therefore improve business insights). Unfortunately, there are no existing standards for generating and attaching provenance.

Researchers at the Universities of Central Queensland and Western Sydney have been studying the attachment of provenance in the Internet of Things and use the diagram below to propose a framework for provenance generation and attachment. The IoT application (top center) desires to analyze provenance about the Managed Device (right-hand side). The logic in-between these two is responsible for adding that provenance.[23]

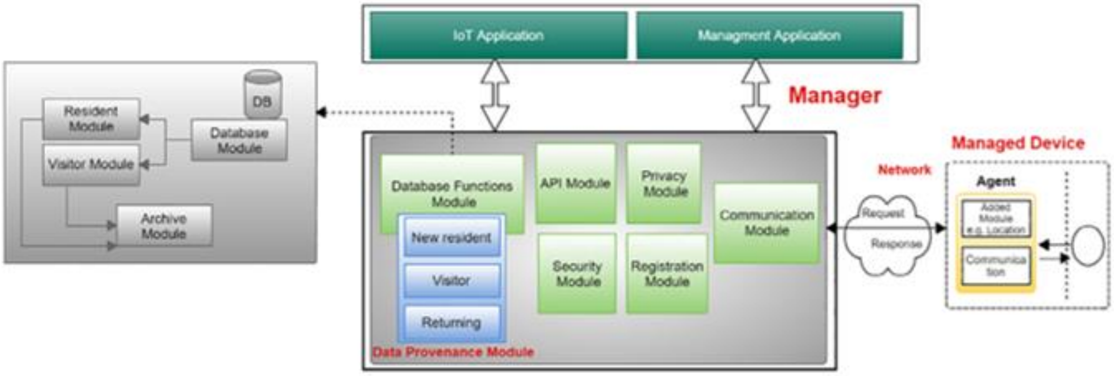


Figure 6 - Research Architecture for Managing the Attachment of IoT Provenance

Figure 6 highlights the requirement for a Management Application (top right) to configure the process of data provenance insertion (a DCF aspires to play that role).

Once provenance information is securely attached to edge data, applications will have much higher confidence in their analytic insights. Provenance attachment increases the value of edge data.

A DCF can record the attachment of provenance information to the data. Applications processing that data will know about this attachment and achieve higher confidence in their business output.

2.6 Data Lineage

Enterprise governance tools track data from its point of origin to its consumption by an application. Business users rely on the ability to inspect this path when incorrect data produces poor business insights.

This problem becomes more complicated in the era of edge data. Infogix describes the challenge:[24]

> *Thanks to big data and new data sources like IoT devices, the amount of information organizations consume continues to grow exponentially. Not only is it critical to know where that data came from, but it's also important to understand where it has been within the data supply chain, and how it has changed along the way. Data's path can be winding and complex, but understanding its usage and flow is a critical part of any enterprise data governance program.*

Tracking the data supply chain of IoT sensor data, for example, cannot be accomplished using a centralized database.

Instead, the industry is beginning to explore the use of distributed ledger technology (DLT). DLTs can record the lineage of data: points of origin, where it's been, who has looked at it, etc. IOTA is a DLT company that is creating a ledger (known as a Tangle) for just this purpose. Figure 7 highlights the use of a DLT that allows a car insurance company to view and track the history of an international automobile incident.[25]

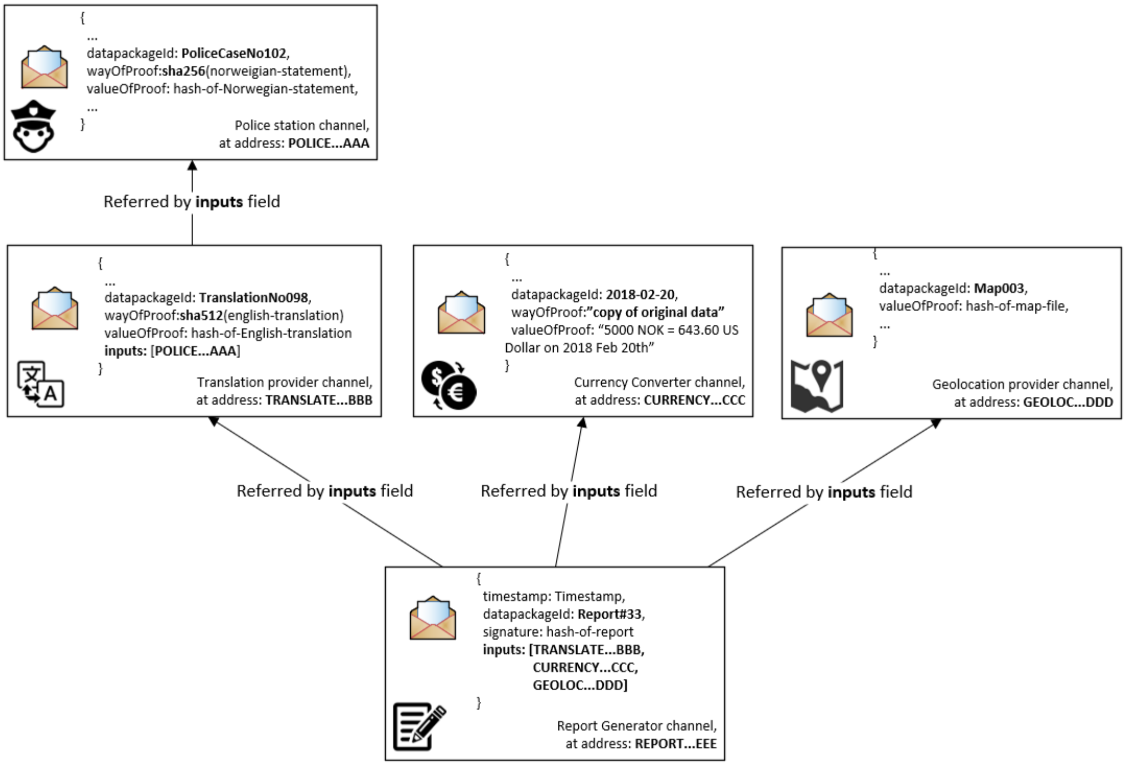

{
...
datapackageId: PoliceCaseNo102,
wayOfProof:sha256(norweigian-statement),
valueOfProof: hash-of-Norwegian-statement,
...
}
Police station channel,
at address: POLICE...AAA
Referred by inputs field
{
...
datapackageId: TranslationNo098,
wayOfProof:sha512(english-translation)
valueOfProof: hash-of-English-translation
inputs: [POLICE...AAA]
}
Translation provider channel,
at address: TRANSLATE...BBB
{
...
datapackageId: 2018-02-20,
wayOfProof:"copy of original data"
valueOfProof: "5000 NOK = 643.60 US Dollar on 2018 Feb 20th"
}
Currency Converter channel,
at address: CURRENCY...CCC
{
...
datapackageId: Map003,
valueOfProof: hash-of-map-file,
...
}
Geolocation provider channel,
at address: GEOLOC...DDD
Referred by inputs field
Referred by inputs field
Referred by inputs field
{
timestamp: Timestamp,
datapackageId: Report#33,
signature: hash-of-report
inputs: [TRANSLATE...BBB,
CURRENCY...CCC,
GEOLOC...DDD]
}
Report Generator channel,
at address: REPORT...EEE

Figure 7 - Using a Ledger to Track Data Lineage

The path depicted in Figure 7 is a graph of ledger entries. The entries show the origin of data (a police report of a traffic accident in Norway) and subsequent forwarding to an insurance company in a different country. A variety of parties (translators, currency converters, and mapping software) handle the data along its way.

While ledgers may hold promise to track data handling from its point of origin, there are currently no application programmer's interfaces (APIs) that allow inspection of the data supply chain in a standard way.

DCFs, by their very nature, record and measure data delivery, and therefore can be used to keep track of this type of lineage.

2.7 Deterministic delivery

Enterprise IT systems can be tweaked and tuned to minimize data latencies. This tuning often enables the data to be in the right place at the right time for mission-critical applications.

At the edge, networking technology often gets in the way of deterministic data delivery. In 2012, an IEEE 801.2 working group began creating a set of networking standards known as Time-Sensitive Networking (TSN). The group recognized that enterprise IT networking technologies could not deliver data precisely at the time that it was needed.[26]

Standard IT network equipment has no concept of "time" and cannot provide synchronization and precision timing. Delivering data reliably is more important than delivering within a specific time, so there are no constraints on delay or synchronization precision.

One of the problems with TSN, however, is that it must bridge into the enterprise space. This bridging would allow all applications (including enterprise apps) to benefit from TSN. Standards are emerging that attempt to cross this bridge. The CC-Link partner association recently announced a new specification that attempts to unite these worlds. Figure 8 shows TSN characteristics extending between Information Technology (IT) systems and Operational Technology (OT) systems.[27]

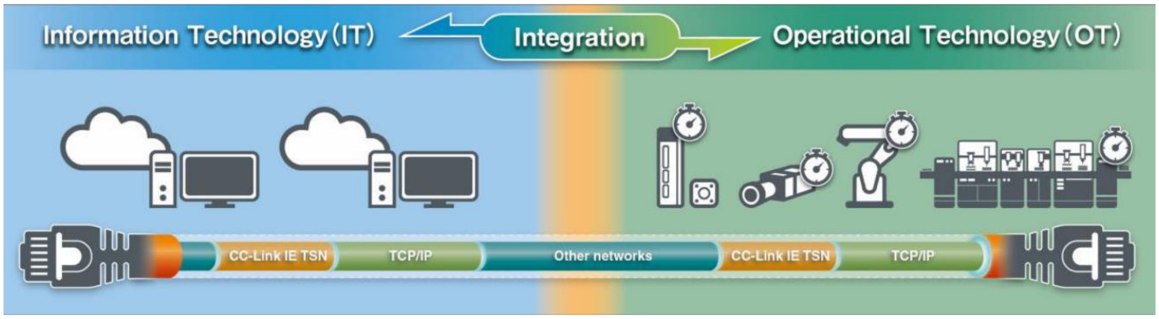


Figure 8 - IT / OT Bridging of Time-Sensitive Networks

There are authorities (Time-Stamping Authorities or TSAs) and other protocols (IETF Time-stamp Protocols – RFC 3161) that are also relevant in this environment. [28]

While bridging standards are forming, however, enterprise-level applications will continue to face significant challenges in making time-critical business decisions using edge data.

Time-sensitive networking is just one example of a networking technology that can provide an increased level of trusted delivery to an application. As a DCF is made aware of the use of various networking primitives, it can record their usage and thus improve the overall score.

2.8 Attack surface

The enterprise goes to great lengths to thwart malicious actors that attempt to penetrate corporate IT systems.

With edge data, the attack surface exponentially grows, and traditional methods of exposing threats are limited in their capabilities. RSA has provided enterprise-class security for decades and has recently turned their attention to securing the edge and, in particular, IoT devices.[29]

Traditionally, security and identity systems have operated separately from IoT systems. Cybersecurity teams secure and monitor IT systems; IoT systems are often managed by lines-of-business (LoBs) with separate engineering teams. Additionally, IoT devices may be deployed in the field and in potentially hostile locations with no physical security guarantees (e.g. an unmanned wind turbine or traffic sensors in a smart city use case).

In such scenarios, the IoT devices require additional protection measures against physical attacks such as manipulating, replacing, or spoofing devices.

RSA has created a lightweight container that can run on constrained edge devices. Project Iris aspires to bring enterprise-class threat monitoring and detection to the edge. Figure 9 depicts Iris detecting threats on an Edge Gateway (the left -side of the figure) in partnership with the enterprise (Project Iris Cloud).[30]

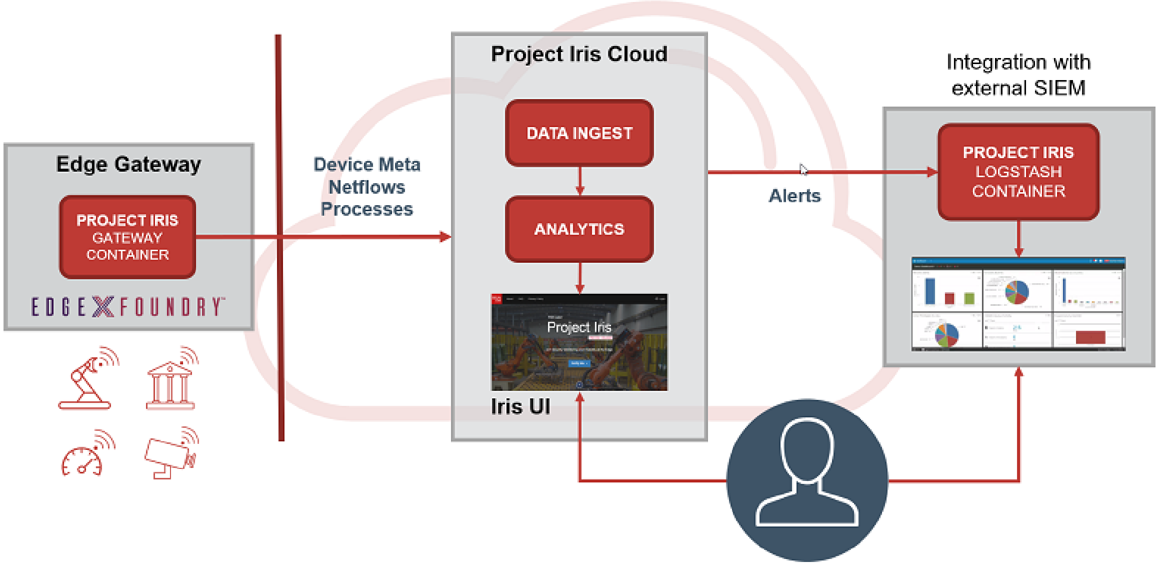


Figure 9 - Project Iris and Protecting Edge Devices

Using IoT operational and security analytics, Project Iris profiles devices, baselines normal behavior, and detects and alerts on anomalous activities and compromised devices. Leveraging machine learning and with no requirement for changing the edge devices, Iris can secure large deployments of IoT sensors and actuators.[31]

In addition to RSA Iris, Manufacturer Usage Descriptions (MUDs) “allow end devices to signal to the network what sort of access and network functionality they require to properly function.”[32] This standard can also help build behavioral profiles.

A DCF can help an application understand what kinds of threats are present, or what type of behavior edge devices see, as edge data travels across a distributed ecosystem. A DCF can annotate any increase in threat activity and, if desired, affect application confidence in the delivery of the data.

2.9 Validation and Measurement

Edge data can travel a long way, from a sensor on a manufacturing floor to an application in a cloud. It is critical that the data is semantically valid and interpreted correctly.

How can a DCF be used to determine if faulty data is coming from a failing sensor?

One method is to perform range checking against the manufacturer's specifications (see IEEE 1451). Alternatively, the DCF can request a comparison of an event against historical readings. Either approach can occur at the point of ingestion or higher up in the stack.

These techniques allow a DCF to confirm, with some level of confidence, that a device is emitting sensible readings.

How can a DCF be used to require that a proper unit of measurement is associated with the data (helping applications to interpret the data correctly)?

VMware's CTO of Edge and IoT, Dr. Greg Bollella, stresses the importance of associating sensor data with the System of International Units (SI Units). The use of SI Units increases the confidence that applications are correctly interpreting readings. He warns of the perils of improperly interpreting data coming from IoT sensors.[33]

> *As streams of measured values from IoT sensors start to pervade everything humans do, the chance for errors and subsequent serious and bad consequences will grow dramatically.*

If a DCF requires the association of SI Units with all incoming sensor data, that data will have a much higher chance of being interpreted correctly by waiting applications.

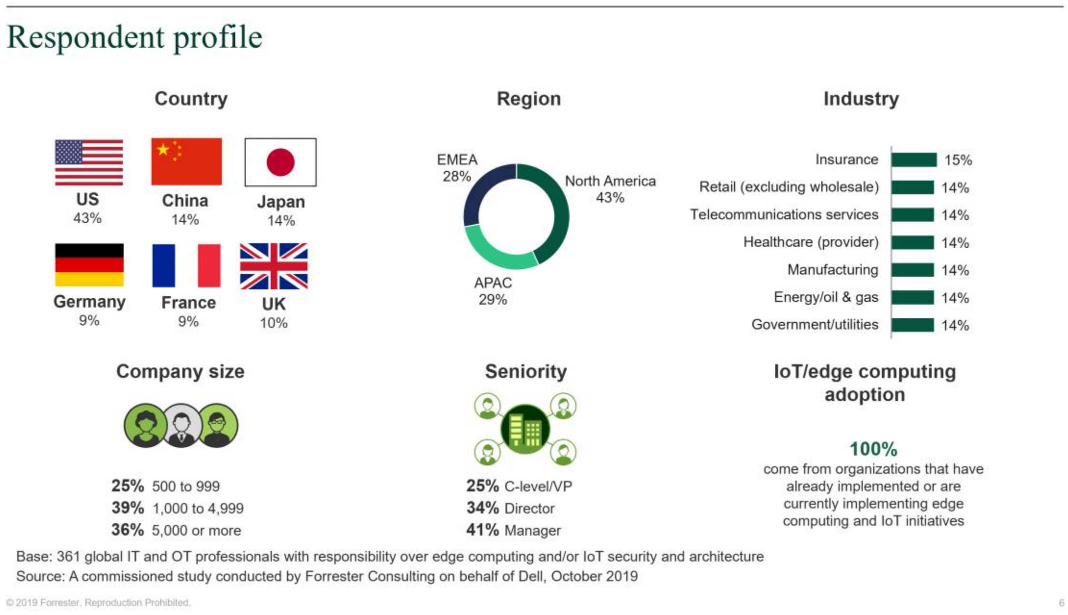
2.10 Industry findings

The nine areas described above represent a broad range of edge data items to address.

A recent survey, commissioned by Dell Technologies and conducted by Forrester Consulting, reinforces that the concerns above are real.[34]

The research hypothesis for this survey states that enterprise companies are ill-equipped to handle the security risks inherent in edge computing and IoT deployments. To test this hypothesis, Forrester surveyed managers, directors, and C-level executives from large (500 employees or more) companies that had already implemented (or are implementing) edge computing and IoT initiatives.

Figure 10 shows the respondent profile.



Respondent profile

Company size

25% 500 to 999
39% 1,000 to 4,999
36% 5,000 or more

Seniority

25% C-level/VP
34% Director
41% Manager

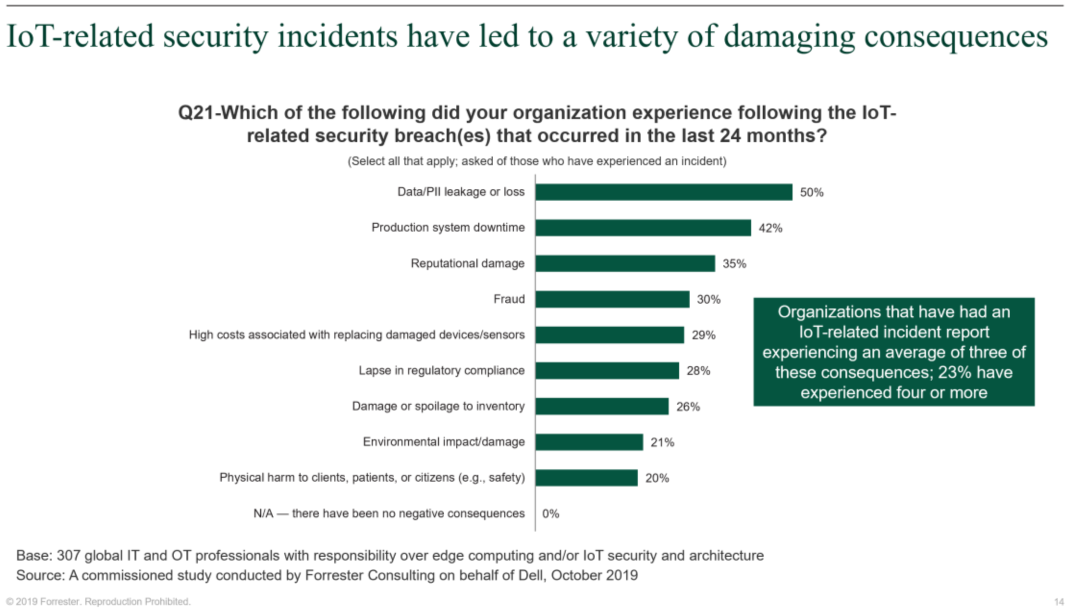
IoT/edge computing adoption

100% come from organizations that have already implemented or are currently implementing edge computing and IoT initiatives

Base: 361 global IT and OT professionals with responsibility over edge computing and/or IoT security and architecture
Source: A commissioned study conducted by Forrester Consulting on behalf of Dell, October 2019

Figure 10 - Respondent Profile from Forrester Edge/IoT Survey

What makes the survey impactful is that Forrester interviewed people from the front line of edge computing. Indeed, 86% of them had already experienced a breach in their edge ecosystems. And during those breaches, data leakage or loss topped their list of damages.



IoT-related security incidents have led to a variety of damaging consequences

Q21-Which of the following did your organization experience following the IoT-related security breach(es) that occurred in the last 24 months?

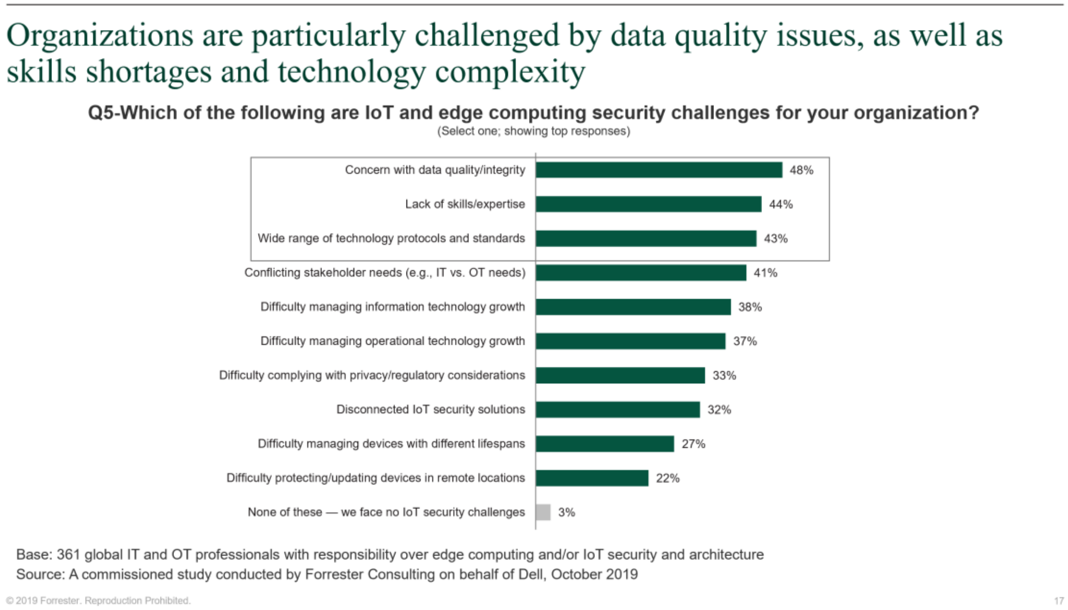
(Select all that apply; asked of those who have experienced an incident)

Organizations that have had an IoT-related incident report experiencing an average of three of these consequences; 23% have experienced four or more

Base: 307 global IT and OT professionals with responsibility over edge computing and/or IoT security and architecture
Source: A commissioned study conducted by Forrester Consulting on behalf of Dell, October 2019

Figure 11 - Data leakage or loss experienced in 50% of edge security incidents

It is no surprise, then, that survey respondents put data quality and integrity at the top of their list of concerns (just above data privacy/regulatory issues).



Organizations are particularly challenged by data quality issues, as well as skills shortages and technology complexity

Q5-Which of the following are IoT and edge computing security challenges for your organization?

(Select one; showing top responses)

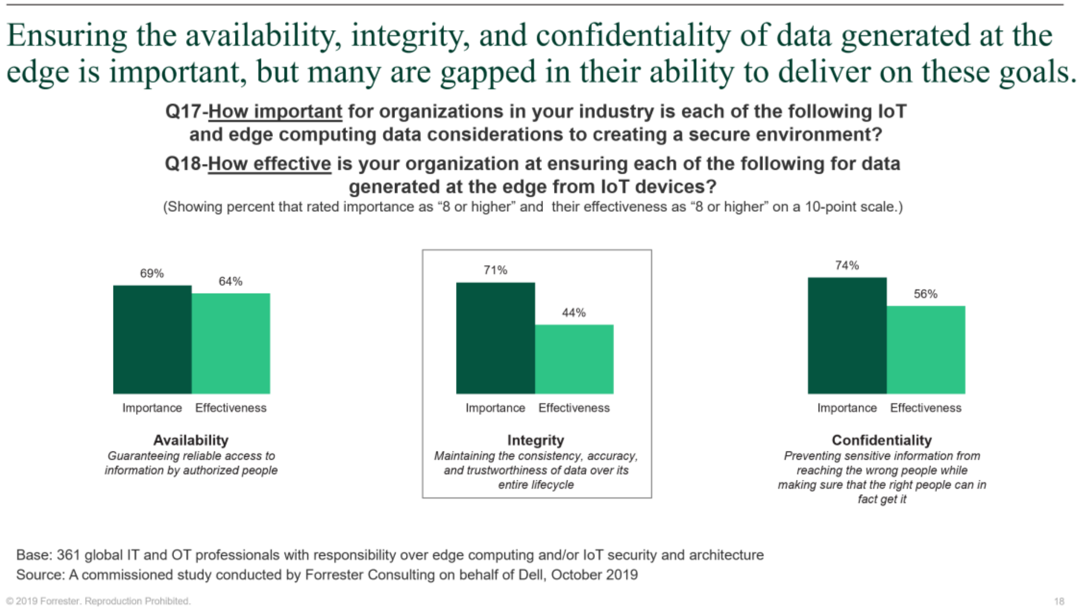
Base: 361 global IT and OT professionals with responsibility over edge computing and/or IoT security and architecture

Source: A commissioned study conducted by Forrester Consulting on behalf of Dell, October 2019

17

Figure 12 - Integrity and Quality of Edge Data is a Top Concern

When explicitly asked about handling the integrity of edge data, from birth to delivery to applications and beyond, respondents indicated that the most significant gap is the inability to manage edge data with integrity.



Ensuring the availability, integrity, and confidentiality of data generated at the edge is important, but many are gapped in their ability to deliver on these goals.

Q17-How important for organizations in your industry is each of the following IoT and edge computing data considerations to creating a secure environment?

Q18-How effective is your organization at ensuring each of the following for data generated at the edge from IoT devices?

(Showing percent that rated importance as "8 or higher" and their effectiveness as "8 or higher" on a 10-point scale.)

Availability
Guaranteeing reliable access to information by authorized people

Integrity
Maintaining the consistency, accuracy, and trustworthiness of data over its entire lifecycle

Confidentiality
Preventing sensitive information from reaching the wrong people while making sure that the right people can in fact get it

Base: 361 global IT and OT professionals with responsibility over edge computing and/or IoT security and architecture

Source: A commissioned study conducted by Forrester Consulting on behalf of Dell, October 2019

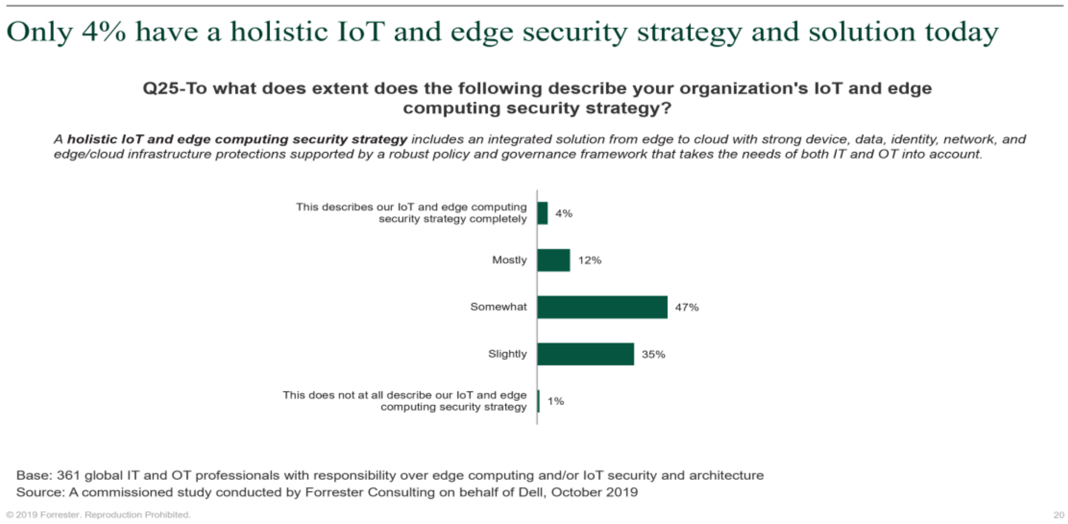
18

Figure 13 - Lack of Industry Confidence in the Handling of Edge Data

The final survey question strikes directly at the heart of the hypothesis that the industry is currently ill-equipped to handle edge security concerns. Forrester defines a "holistic IoT and edge computing strategy" as follows:

*A **holistic IoT and edge computing security strategy** includes an integrated solution from edge to cloud with strong device, data, identity, network, and edge/cloud infrastructure protections supported by a robust policy and governance framework that takes the needs of both IT and OT into account.*

This definition hits at many of the checklist items that were described in sections 2.1-2.9: identity, network determinism, governance, etc. When IoT and Edge practitioners were asked whether their organizational strategy was <u>entirely</u> consistent with a holistic strategy, the answer was a resounding "no."



Only 4% have a holistic IoT and edge security strategy and solution today

Q25-To what does extent does the following describe your organization's IoT and edge computing security strategy?

*A **holistic IoT and edge computing security strategy** includes an integrated solution from edge to cloud with strong device, data, identity, network, and edge/cloud infrastructure protections supported by a robust policy and governance framework that takes the needs of both IT and OT into account.*

This describes our IoT and edge computing security strategy completely
4%
Mostly
12%
Somewhat
47%
Slightly
35%
This does not at all describe our IoT and edge computing security strategy
1%

Base: 361 global IT and OT professionals with responsibility over edge computing and/or IoT security and architecture
Source: A commissioned study conducted by Forrester Consulting on behalf of Dell, October 2019

Figure 14 - The Need for a Holistic Edge Security Strategy

The issues described in Sections 2.1-2.9 have potential solutions in-flight. But none of them work together holistically.

A Data Confidence Fabric (DCF) aspires to address this gap. If as successful as the code anticipates, this technology will bring the same level of trustworthiness to edge data as is experienced in the enterprise.

The good news here is that the DCF design derives from enterprise-class data delivery primitives.

How can this be? Enterprise infrastructure is primarily based on centralized, homogenous technology. Highly-skilled enterprise architects configure the data delivery path, surrounding the data with secure, trusted hardware and software components. In this environment, enterprise applications don't trust the data itself; they trust that the underlying data delivery primitives are sound.

By contrast, edge infrastructure is intrinsically heterogeneous and too decentralized for one security team. The data delivery primitives are unknown (and therefore untrusted).

Can a record of the deployed data delivery primitives travel with the data itself?

To answer this question, let's look at how enterprise data is delivered.

3 Lessons from Enterprise Data Delivery

The era of enterprise storage began in the late 1980s as companies became more reliant on applications, and applications became ever-hungrier for disk capacity and performance.

Over the next three decades, customers spent tens of billions of dollars on enterprise-class storage systems. For example, in 1994, EMC's Symmetrix storage system reached one billion dollars in sales. By the time Dell had announced the acquisition of EMC in 2016, IDC had estimated that the size of the external enterprise storage market had reached twenty-four billion dollars.[35]

Why was the industry willing to invest so heavily in enterprise storage? Because delivering trusted data to enterprise applications was increasingly impacting the balance sheets of major corporations. The data protection market grew alongside enterprise storage, for example, because applications could not afford any downtime. Business interruptions to trusted data delivery could result in millions of dollars of lost revenue.

The hypothesis, therefore, is that enterprise applications that analyze edge data will still have the same need for trusted data delivery.

How does enterprise-class storage deliver trusted data? Figure 15 depicts the answer. Enterprise-class storage delivers trusted data via layered trust insertion.[36]

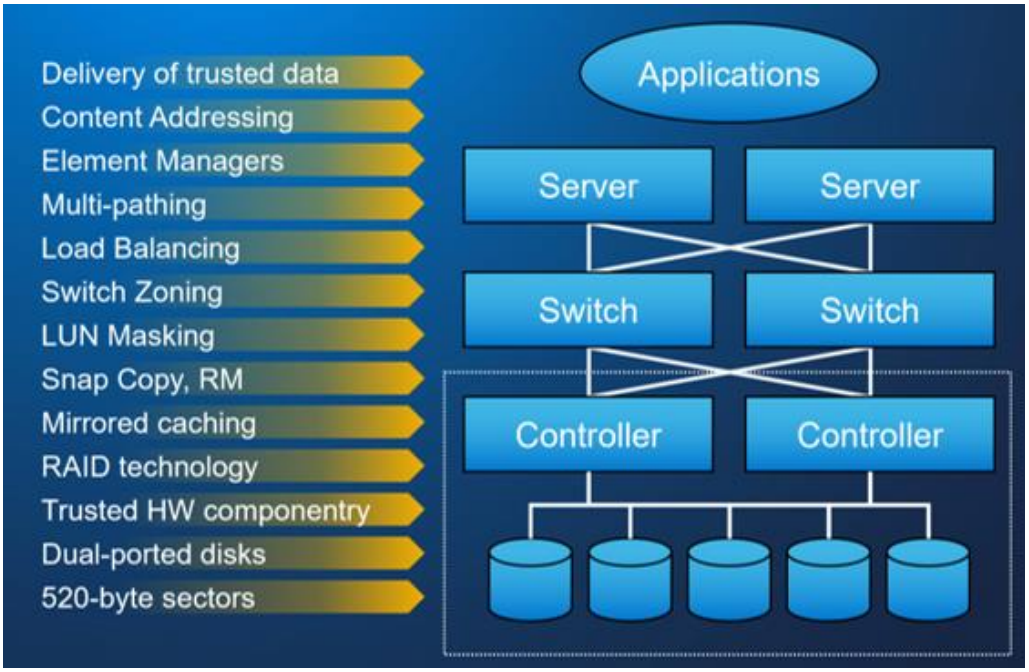


Figure 15 - Enterprise-class Trust Insertion

Figure 15 highlights that as enterprise data flows from individual disk devices (bottom layer) to applications (top layer), various forms of hardware and software trust insertion occur. As data arrives at the top, confidence in the data is assumed to be high. This assurance results from the care the data received on its journey.

When data confidence is high, the data's value is high. And when the data's value is high, business processes are optimized, the risk reduces, and revenue increases. Application processing of reliable data brings a myriad of benefits to corporate balance sheets.

To further illustrate this benefit, the list below highlights the trust insertion layers depicted in Figure 15. Keep in mind that a team of experts configured these layers. The descriptions start at the disk level and progress up to the application. Each technology "inserts trust," and applications implicitly assume the data is genuine.

- Disk vendors allowed the re-formatting of the disk surface to include eight extra bytes (increasing sector sizes from 512 to 520 bytes). These bytes served as "data integrity bits" that provided additional protection for the data (e.g., the storing of checksum values).
- Disk vendors also added an extra access port to disk drives. This hardware trust insertion technique increased the availability of the data in case of failures.
- Disk controllers (depicted just above the disk drives in Figure 15) contained trusted hardware componentry (e.g., non-volatile RAMs for surviving power failures) to increase the trustworthiness of the data path.
- RAID software (e.g., redundant array of inexpensive disk level 5) ran on these disk controllers. These algorithms could continue delivering trusted data in the event of a disk failure.
- Mirrored caching software delivered trusted data faster while simultaneously protecting against disk controller failures.
- Snap copy (additional local copies) and remote mirroring (extra remote copies) were data protection techniques that enabled business continuance in the face of failures of the original (or additional availability via offline analysis).
- LUN Masking software "hid" data from some applications while allowing access to trusted applications.
- Network switches could be placed side-by-side (fault tolerance), and zones created to provide "trusted paths" to specific data sets.
- Load-balancing could detect how much data was flowing over specific paths and balance the traffic to deliver data more quickly.
- Multi-pathing software could detect failures in pathways to the data and route around them, enabling continuity in the delivery of the data to applications.
- Element managers configured trusted data delivery paths via user interfaces.
- Content addressing (as described in section 2) provided the ability to confirm that the data had not been tampered with or changed.

The stack described in the bulleted list above can result in significant delays in data delivery. However, for every trust insertion delay (e.g., the write penalty of RAID-5 algorithms), an optimization emerged to overcome that delay (e.g., a write cache that sits atop RAID-5 configurations).

The combination of the trust insertion components described above (both hardware and software) results in the delivery of trusted data in the enterprise.

Consider an edge data delivery environment in which the problems described in Sections 2.1-2.9 were solved:

1. Data originates in a hardware root of trust environment with verifiable digital signatures and confirmation of secure boot and device onboarding.
2. Robust access control and authentication prevent access to this data by denying unauthorized parties.
3. Hash values on ingested data are generated early in the ingest process, allowing upstream applications to detect tampering, modification, or corruption.
4. The identity of the data's owner is well-known and securely associated with the data. Proper governance policies can be enforced based on ownership.
5. Provenance about the point of origin is attached to the data and is cryptographically verifiable.
6. Lineage metadata records the path that the data takes on its way to the application.
7. Deterministic data delivery techniques move data across networks in a reliable and timely fashion.
8. Threats to the data are monitored, detected, and increasingly prevented.
9. Semantic validation and proper labeling of the data occur, increasing the odds that the application will function correctly when analyzing the data.

This list displays all the hallmarks of trusted enterprise data delivery. Not described above are additional enterprise storage techniques (e.g., data-at-rest encryption, data sanitization). Many of them would also provide value at the edge.

Unfortunately, edge applications that process this data cannot prove that these steps occurred. The level of trust assumed to exist in enterprise data delivery contexts does not translate to the edge. But this does not mean that trusted data delivery is unachievable. It just means that the application needs a new way to determine that it happened.

Therefore, we must record the existence of trusted data delivery on the edge by formally annotating the delivery process.

For example, evidence of the execution of all nine steps can be documented, digitally signed, and delivered to an application.

Data Confidence Fabrics aim to provide this solution by annotating the trust insertion process. As edge data flows through a DCF framework, a record of trust insertion will appear alongside it, and the record is permanently associated with the data.

The following section provides an overview of how a Data Confidence Fabric can annotate the edge data delivery process.

4 Data Confidence Fabrics: An Overview

The introduction to this paper defined a Data Confidence Fabric (DCF):

A Data Confidence Fabric delivers trusted data to applications with measurable confidence.

What does it mean when edge data “enters” into a DCF?

As data travels across a distributed (e.g., edge) environment, on its way to one or more applications, a DCF facilitates a well-defined (measurable) annotation process. Figure 16 depicts this flow.

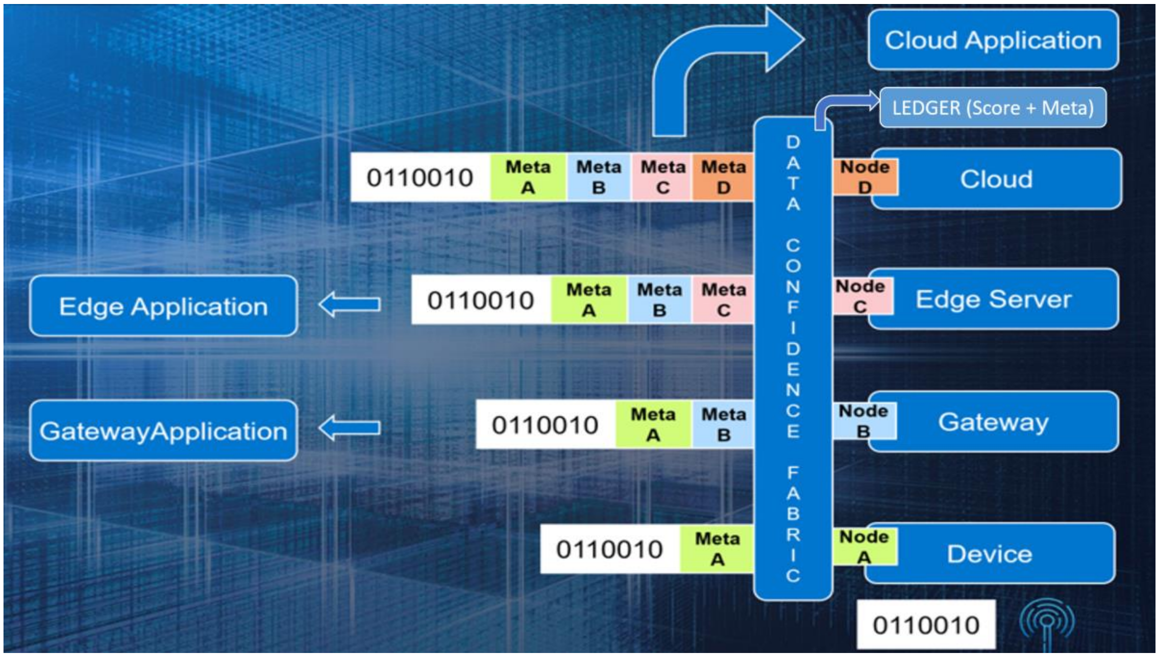


Figure 16 - Annotation in a Data Confidence Fabric

Figure 16 presents an example in which a device (e.g., the Wi-Fi-enabled sensor shown at the bottom right) generates an event (e.g., a bitstream). The event travels across a DCF via a distributed edge ecosystem from the device (Node A) to a gateway (Node B) to an edge server (Node C) to a cloud (Node D).

Each node actively participates in the Data Confidence Fabric by (a) performing trust insertion and (b) annotating the results of the trust insertion along the way.

For example, Node A may possess a Trusted Platform Module (TPM) chip that digitally signs data (e.g., for proving data ownership). After performing the signature, Node A documents the results and attaches the annotation (Meta A) to the device data.

Node B (the gateway) receives both the event (the bitstream) and the annotation (Meta A) and performs additional trust insertion. For example, Node B can validate the TPM signature (if it knows the TPM’s public key), and it can attach provenance about the ingest environment (hardware and software revisions, geographic location of the device, etc.). As Node B executes trust insertion on the event data, it records the results, generates the annotation (Meta B), and sends everything up the chain.

The event destination is a cloud application. Along the way, trusted applications can also process the event (and the associated annotations) as it passes through the fabric.

All DCF metadata, and the resulting confidence score, is inserted into a secure ledger. The ledger entry ultimately provides assurance of trusted data delivery to applications.

Figure 16 highlights that trust insertion increases the overall payload size. The method of handling/forwarding of the trust insertion metadata (e.g., over the same path or not) will be a primary area of discussion in the Alvarium community.

A trust insertion configuration file defines the operation of a DCF. This file can be applied across a distributed set of nodes, as depicted in Figure 17.

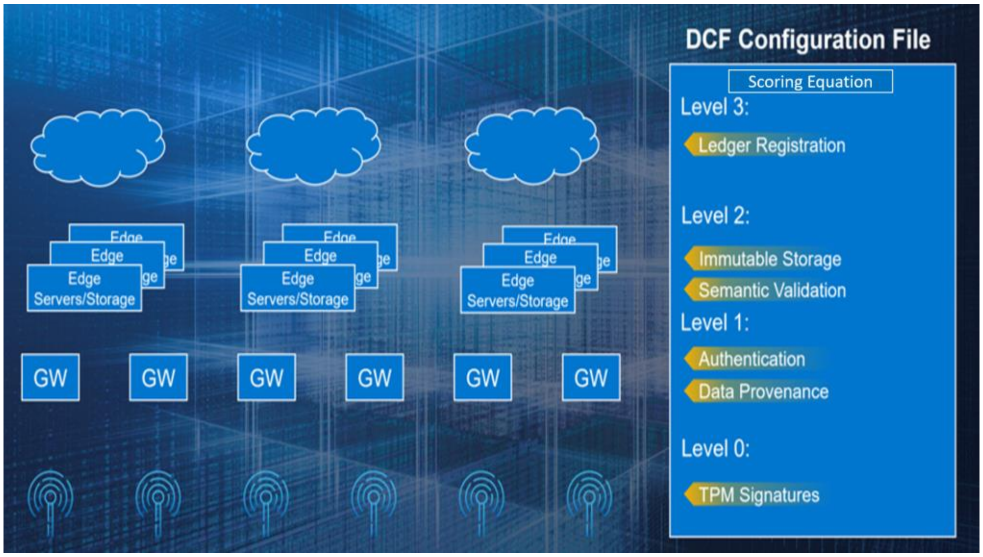


Figure 17 - DCF Configuration File

The configuration file depicted in Figure 17 specifies various forms of trust insertion at each level. The edge ecosystem attempts to insert trust based on this file:

- Level zero devices perform digital signature operations on device data using their local TPMs.
- Level one gateways (a) attach data provenance information to any device data, and (b) authenticate any attempt to inspect device data (or associated trust metadata).
- Level two edge servers (a) perform semantic validation on the device data, and (b) store a copy of the device data into an immutable object-store. Note that edge servers are likely to have more compute and storage capability to perform these types of trust insertion operations.
- Level three cloud applications (a) validate the TPM signature, and (b) register the trust metadata, and a reference to the data, in a ledger.

Multiple DCFs can be configured and co-exist simultaneously. Peer DCFs can also co-exist across different companies or organizations, facilitating data exchange opportunities. It is also possible for third parties to install and operate DCFs on behalf of others (e.g., Trust-as-a-Service).

One final aspect of a DCF is the ability to define equations that lead to a DCF-specific "confidence score." A confidence score is valuable for several reasons.

1. Confidence scores reflect the investment made into the delivery of trusted data.
2. Confidence scores establish a baseline of trust that can improve over time.
3. A decrease in confidence scores can signify that trust insertion operations failed or did not execute.
4. Data set delivery methods can be compared against each other via confidence scores.
5. It is possible to generate high-level portfolio views of trusted data assets.
6. Statements of data's value can derive from confidence scores.

The Scoring Equation displayed at the top of the DCF Configuration File in Figure 17 produces a numeric value. The equation is designed to be flexible and specific to any given fabric.

If each trust insertion technology in Figure 17 produces an output of "1" (success) or "0" (failure), an equation can use these values as inputs producing a confidence score.

An equation that uses addition, for example, could have a target confidence score of "6", which means that all trust insertion technologies executed properly in the delivery of the data (6 out of 6 means 100% confidence in the data's delivery).

$$\begin{aligned}\mathit{ConfidenceScore} = {} & \mathit{Result\ (TPM\ Signature)} + \\ & \mathit{Result\ (Data\ Provenance)} + \\ & \mathit{Result\ (Authentication)} + \\ & \mathit{Result\ (Semantic\ Validation)} + \\ & \mathit{Result\ (Immutable\ Storage)} + \\ & \mathit{Result\ (Ledger\ Registration)}\end{aligned}$$

An equation that uses multiplication could have a confidence score of "1" if all methods executed properly and "0" if any of them did not.

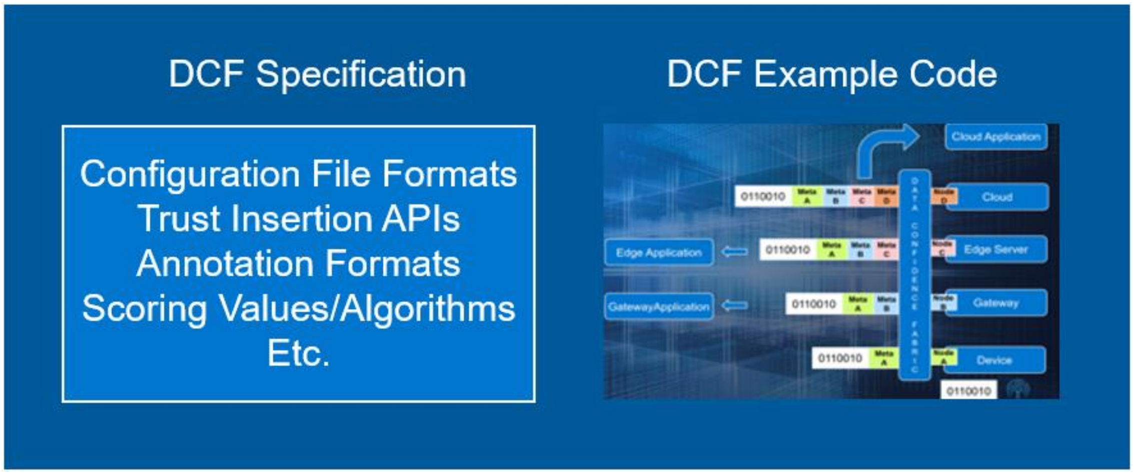
$$\begin{aligned}\mathit{ConfidenceScore} = {} & \mathit{Result\ (TPM\ Signature)} * \\ & \mathit{Result\ (Data\ Provenance)} * \\ & \mathit{Result\ (Authentication)} * \\ & \mathit{Result\ (Semantic\ Validation)} * \\ & \mathit{Result\ (Immutable\ Storage)} * \\ & \mathit{Result\ (Ledger\ Registration)}\end{aligned}$$

If any given method wishes to provide some variability in the reporting of its results, it may return a result that falls within a specific range (e.g., a number between 0.0 and 1.0). For example, if the "Data Provenance" method is only able to gather and attach four out of five desired metadata values, it can contribute a score of 0.8 (which impacts the overall ConfidenceScore result).

The Alvarium community can also explore the inclusion of DCF metadata in the equations. For example, if the DCF Authentication metadata contains a pointer to a list of failed attempts at accessing the data, the overall confidence score can be adjusted accordingly. Similarly, if a stronger hashing algorithm was used while storing data, the score may increase.

The inclusion of a DCF equation, and its resulting confidence score, touches on a primary benefit of data confidence fabrics: measurable impact to corporate balance sheets. The next section discusses confidence scores and their relation to the data’s value.

Project Alvarium is an open community of companies and technologists committed to collaborating on Data Confidence Fabrics. In 2020, member companies can analyze and improve the two assets depicted in Figure 18.



DCF Specification

Configuration File Formats
Trust Insertion APIs
Annotation Formats
Scoring Values/Algorithms
Etc.

DCF Example Code

Figure 18 – Initial Contributions to Project Alvarium

The Project Alvarium community will explore a specification for edge data delivery that, when implemented, delivers annotated edge data. When data moves across geographies, vendors, or clouds, the annotations join it, and both are available for inspection (e.g., through consultation of a ledger). This approach advocates for a standard metadata format that spans the enterprise and the edge.

With the general understanding of how a Data Confidence Fabric works, it is now appropriate to review how they can positively impact corporate balance sheets.

Perhaps the most immediate and impactful way that a DCF can affect the bottom line is through the avoidance of fines. As more companies undergo audits of their data assets, they are challenged to prove that they did not violate compliance laws.

If they fail at proving their compliance, significant fines can result (as discussed later in this paper). To avoid these fines, organizations must build auditable levels of trust into their edge data delivery mechanisms.

Figure 19 shows an additional trust insertion component known as a “policy assignment.”

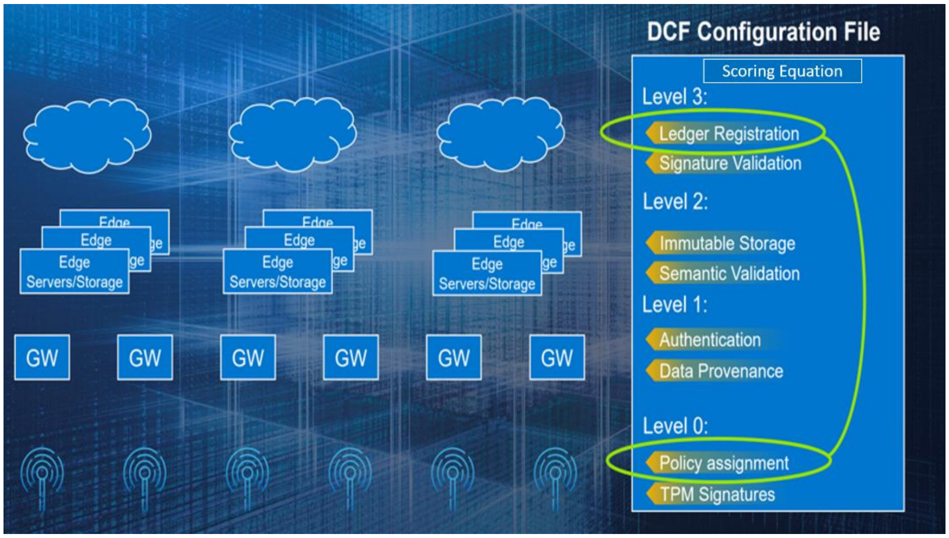


Figure 19 - Assignment of Data Policies in a DCF

Figure 19 shows the assignment of a data policy (green oval on the bottom) to new edge data originating from a sensor. The policy describes the constraints applicable to the incoming data. As the data moves through the Data Confidence Fabric, the policy is visible to every node. When the data registers in the ledger (green oval on top), the policy permanently associates itself with the data in a ledger entry.

With a DCF, any application wishing to use the data is intrinsically aware of the policy and can create new business logic that is compliant with the regulations.

The top-level ledger entry is strong evidence that the company has proper safeguards in place in the face of an audit. Business applications that leverage this ledger entry to access the data can similarly log their adherence to the policy (providing further proof of compliance).

In the same way that banking ledgers prove financial compliance, DCF ledgers can prove data compliance.

How much corporate money might be saved by applications that attach compliance policies to newly-created edge data (and then enforce those policies)? A quick survey of the 2019 regulatory fines landscape reveals that the risk profile for data violations runs into the billions of dollars.[37]

Another way that a DCF can impact the bottom line is by feeding more reliable data to AI algorithms. These algorithms often strive to make automated business decisions that increase operational efficiency and reduce corporate spending. When untrustworthy data feeds into these algorithms, automated business decisions can do more damage than good. The bottom left-hand circle in the Venn diagram (Figure 20) highlights this problem.[38]

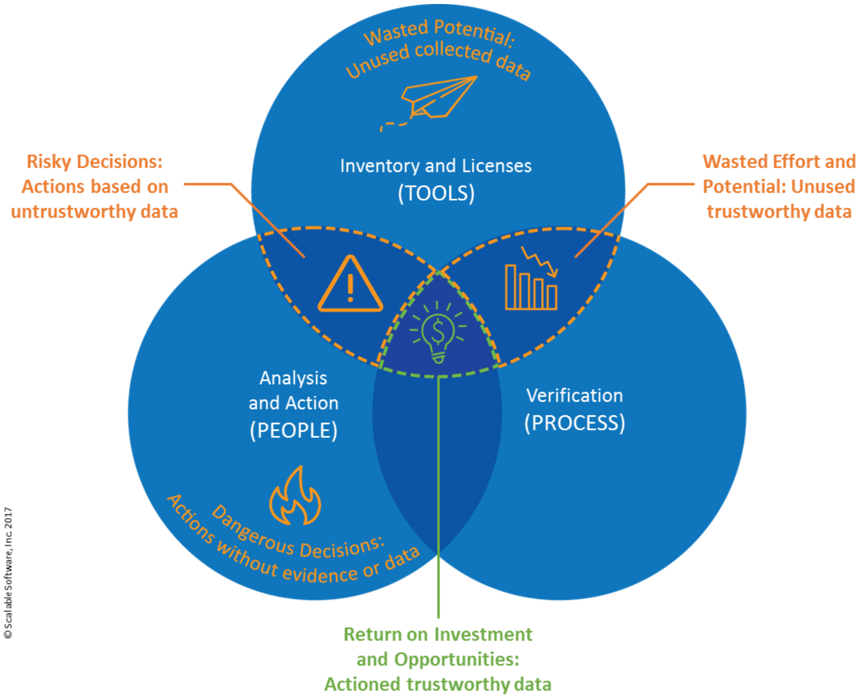

Wasted Potential: Unused collected data
Risky Decisions: Actions based on untrustworthy data
Inventory and Licenses (TOOLS)
Wasted Effort and Potential: Unused trustworthy data
Analysis and Action (PEOPLE)
Verification (PROCESS)
Dangerous Decisions: Actions without evidence or data
© Scalable Software, Inc 2017
Return on Investment and Opportunities: Actioned trustworthy data

Figure 20 - Dangerous, Risky, or Profitable Decisions Based on Data

The diagram highlights that analysis (either by people or algorithms) leads to business actions. In the IT world, using data to make decisions has three possible outcomes:

1. A lack of data results in dangerous decisions.
2. The use of untrustworthy data leads to risky choices.
3. Trustworthy data yields a return on investment.

With DCF, an application has access to the data, the confidence scores, and the policy leading to the score. One hypothesis to explore with a DCF is whether increased data confidence scores yield a higher corporate return on investment.

It makes sense that highly annotated data would yield more accurate (and profitable) insights. A DCF, therefore, can assist in the reduction of operational expenditures.

In summary, a DCF is a holistic trust framework used to annotate the trustworthiness of the data delivery process. It uses distributed ledger technology to record the annotations permanently. For example, a DCF ledger can adequately record the components used to address the nine checklist items described in Sections 2.1-2.9:

1. The use of hardware root of trust results in a DCF ledger entry that records the successful use of a TPM to sign the data. The entry also records successful secure boot and onboarding.
2. An authentication and access control framework can record the existence and enforcement of authentication policies.
3. When data is persisted into an IPFS data store, the hash value can then be entered into a DCF ledger as a way for future applications to directly reference the ledger (and validate that nothing was altered post-storage).

4. The identity of the data's owner can also be stored in a DCF, permanently associating the owner with the data. Privacy policies based on identity are attached to the data and adhered to throughout the delivery process.
5. Provenance about the point of origin can be added to the DCF.
6. Lineage metadata can be accumulated along the DCF path and associated with the data.
7. The use of any deterministic data delivery techniques (e.g., TSN) can be recorded in the DCF, as well as the actual time that the data was delivered.
8. The threat profile at the time of ingestion can be detected by tools such as RSA Iris and associated with the data.
9. Validation of data ranges and unit labeling can occur as part of a DCF configuration.

One final DCF benefit to explore is the ability to use confidence scores for an emerging revenue stream: the sale of data. Confidence scores can be used as part of a calculation to determine the data's value. As will be shown below, knowing the value of data can open the door for selling data as a new form of revenue.

5 Data Confidence Fabrics and Data Value

Data Confidence Fabrics have research roots in the study of data's value.

In 2017, Dell Technologies contributed to a research paper exploring the topic of data valuation. The paper concluded that most corporations were unprepared to place a value on data.[39]

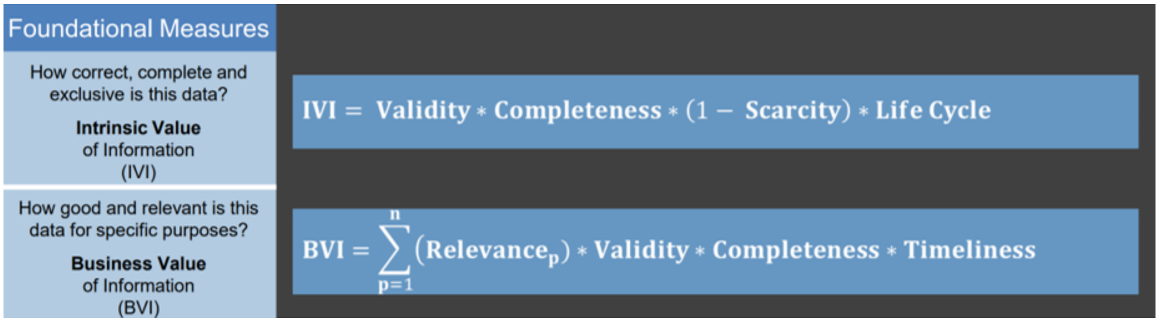
> *All the companies we studied were awash in data, and the volume of their stored data was growing on average by 40% per year. We expected this explosion of data would place pressure on management to know which data was most valuable. However, the majority of companies reported they had no formal data valuation policies in place.*

This lack of formal data valuation policies leaves companies in a position to miss out on creating new revenue streams. Accenture predicts that by 2030 twelve exabytes of edge (IoT) data will be monetized every day. The company believes that the monetization will generate trillions of dollars of value through data exchange.[40]

Analyst company Gartner has been a thought leader in formalizing the approach to data valuation. Analyst Doug Laney published a Gartner report that advised corporations to implement valuation models:[41]

> *CDOs and CAOs, with the guidance of CFOs, should establish a standard methodology for measuring the actual and potential economic value of key information assets to their organizations. Adopt one or more of Gartner's suggested information valuation models and perform these measurements periodically.*

Figure 21 depicts two of the models that Gartner suggests.[42]



Foundational Measures

How correct, complete and exclusive is this data?

Intrinsic Value of Information (IVI)

$$\text{IVI} = \text{Validity} * \text{Completeness} * (1 - \text{Scarcity}) * \text{Life Cycle}$$

How good and relevant is this data for specific purposes?

Business Value of Information (BVI)

$$\text{BVI} = \sum_{p=1}^{n} (\text{Relevance}_p) * \text{Validity} * \text{Completeness} * \text{Timeliness}$$

Figure 21 - Example Gartner Data Valuation Equations

Data characteristics like validity, completeness, and timeliness, as defined by Gartner, are related to trust. However, implementing Gartner's recommendations can be challenging; data is typically not annotated with any of these characteristics. Or, if there is annotation, it is often done after the fact (which can affect timeliness).

DCF confidence scores can feed into many of the variables used in data valuation calculations. A DCF also supports a programmable framework that allows the execution of dynamic equations on-the-fly. Figure 22 provides a graphical view of the addition process described in the previous section, along with a view of the trust insertion technologies contributing to the score.

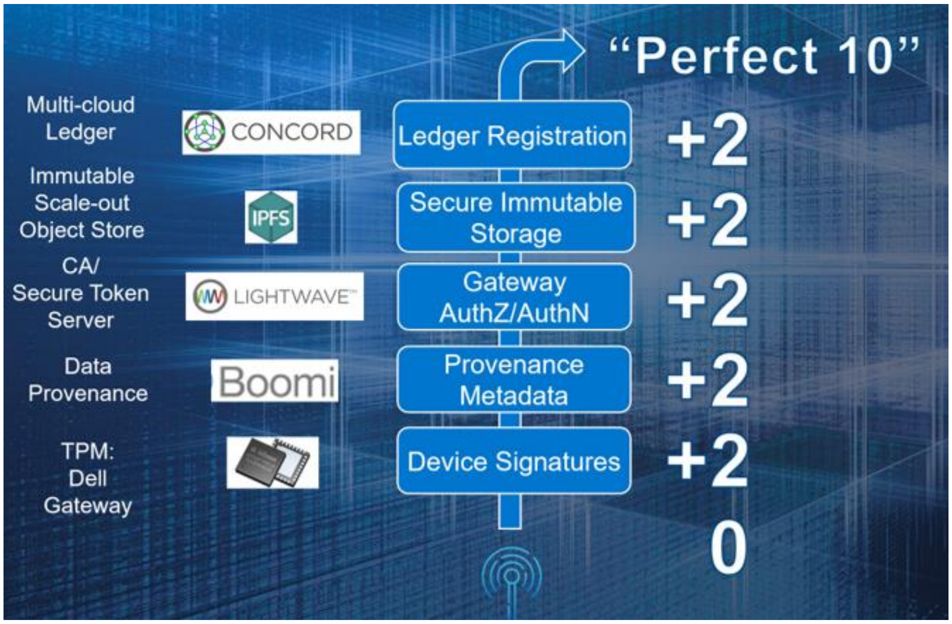


Figure 22 – DCF Trust Insertion Technologies and Scoring

The example in Figure 22 came from a Dell Technologies lab in which a mix of open and proprietary trust insertion components contributed scores to the first-ever DCF. These components included:

- Signing data via a TPM chip on a Dell Technologies Gateway 3000.
- Capturing the provenance of the gateway's ingest environment using Dell Technologies' Boomi product.

- Using a certification authority and secure token server (VMware's Open-Source Lightwave technology) to authenticate requests to inspect the data.
- Immutably storing the data in an open-source object store (IPFS).
- Storing the score and IPFS reference in an immutable ledger (using VMware's open-source Project Concord consensus algorithm).

Note that a DCF allows the mixing and matching of different trust insertion technologies. The same result was also achieved when the IOTA ledger was substituted for VMware.

Figure 22 highlights the promise of a Data Confidence Fabric. A ledger entry records newly-created data. This entry contains both a pointer to the data (or batches of data) and a statement of value (the confidence score and associated trust delivery metadata). Ledgers have immutability characteristics as well as implied ownership (e.g., the owner of the private key signs the ledger entry).

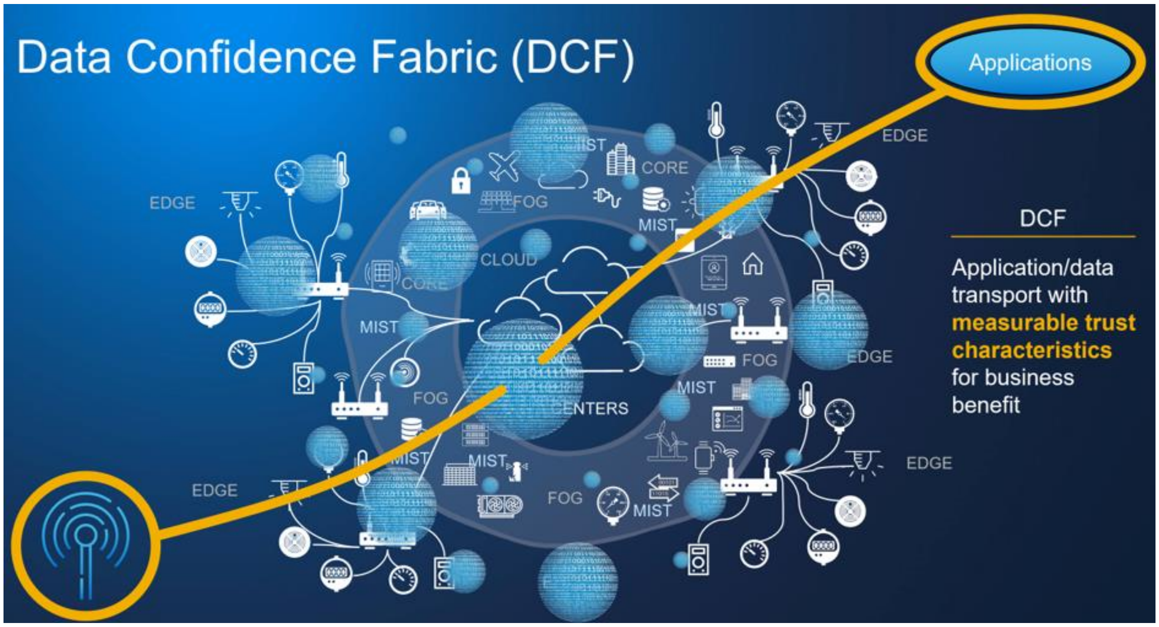
An open Data Confidence Fabric specification is a win-win-win:

- Consumers win. As their data enters a DCF, data ownership and consumer privacy policies can be applied immediately. If the consumer calls into question the treatment of their data, DCF ledgers reconstruct the lineage over which their data flowed. The ownership established in those ledgers can also position consumers to monetize the assets that they own (and high confidence scores make the data asset more attractive to potential buyers).
- Enterprise companies win. Providing a DCF deployment for consumers to use can lead to higher consumer confidence. Decentralized edge-based infrastructures will begin to exhibit the characteristics of the enterprise-class systems with which they interact. DCF confidence scores establish baseline trust instrumentation that can be measured and improved over time. DCF operators, like consumers, can position themselves to derive business value out of a ledger of owned data assets. These companies can also explore the reduction in cost when building and running a DCF with open trust components (e.g., IPFS).
- Vendors win when their trust insertion technologies appear in Data Confidence Fabrics. As ledger technologies mature, vendor payment for trust services can be built on top of (or into) the ledgers themselves.

Dell Technologies' DCF contributions to the Project Alvarium community is not just a step towards the company's data vision for 2030. It also represents a future business boost for customers wishing to ride the anticipated wave of data marketplaces (trillions of dollars in revenue) while avoiding the wave of regulatory fees (billions of dollars in fines).

6 Conclusion

A Data Confidence Fabric spans the complexity of edge ecosystems, annotating edge data with confidence scores that bring measurable business benefits. Figure 23 depicts how a DCF cuts through the complexity. Data flows through a secure, open fabric as it undergoes trust annotation.



Data Confidence Fabric (DCF)

DCF

Application/data transport with **measurable trust characteristics** for business benefit

Figure 23 – DCF Improves Business Results for Edge Data

As we saw in section two of this paper, the difficulties inherent in the management of edge data are numerous. The Forrester survey reinforced this view and called for a holistic solution.[43]

Organizations, regardless of their level of maturity, should consolidate the number of vendors they work with and prioritize those that can curate a holistic solution – ones with proven integrations and broad technology ecosystems that can address multiple needs.

This consolidation is happening within the Project Alvarium community, whose purpose it is to curate a holistic solution: The Data Confidence Fabric.

DCFs have evolved from enterprise storage systems. These systems have reliably supplied data to enterprise applications for decades. By distributing trust insertion to the edge, and annotating the process, corporations can experience significant business benefits.

Dell Technologies has brought its data protection expertise to the Project Alvarium community in the form of the industry’s first Data Confidence Fabric. This fabric has demonstrated the feasibility of performing trust insertion and highlighted the possibility of generating confidence scores.

Confidence scores bring new opportunities to bolster corporate balance sheets. They enable the reduction of fines through regulatory compliance and increase the odds of prudent business decisions via optimized insights on more reliable data.

A DCF can be built on top of existing IoT and edge ecosystems by adding trust insertion components over time (in adherence to the Alvarium specification). Consumers are then drawn to these fabrics due to their disciplined, measured treatment of edge data. While a

DCF is initially targeted for the annotation of edge data, it can also provide the same benefits inside an enterprise (providing even more benefits to consumers)

Finally, Data Confidence Fabrics guide the industry on the long-term path of data trust and data monetization: selling trustworthy data for profit in data marketplaces.

This long-term vision brings us back to Dell Technologies' audacious 2030 goal for consumer data. One of the reasons for creating a Data Confidence Fabric in the first place is to solve the many problems associated with handling precious consumer data from the edge.

As Dell Technologies enters 2020, the company is investing in Data Confidence Fabrics for the following reasons:

- DCFs support consumer data coming from Hardware Root of Trust devices and can appropriately label the data that originates from those devices.
- DCFs can validate that access and authentication technology is in operation on the edge systems that ingest data from consumers. There is also a potential tie to attestation (whether the type of device is appropriate for the data or workload).
- DCFs allow applications to check whether consumer data has undergone change or tampering.
- A DCF records the environment in which consumer data originates (e.g., such as a hospital). This provenance will include specific information about the device capturing the data (e.g., similar to keeping track of the equipment and people that were present in a hospital delivery room), enabling additional insights.
- DCFs can record the original owner of the data (e.g., the consumer).
- DCFs can track what happens to the data (lineage) as it makes its way towards applications and beyond. It can also track data transformations and protection (e.g., replication).
- A DCF can keep track of whether consumer data was delivered on time (e.g., deterministic delivery via the TSN described in Section 2.7) and in compliance with the appropriate policy.
- DCFs can acknowledge the presence of threats to consumer data and the handling of them.
- DCFs can require that data is properly labeled and validated, increasing the confidence that applications output correct business insights.
- DCFs have the potential to address many more use cases, including chain of custody, data supply chains, and tracking derivative works.

That said, the company recognizes that it cannot accomplish the 2030 goals on its own. Organizations that are serious about capitalizing on the rising value of edge data would be wise to join and contribute to the Project Alvarium community at https://alvarium.org/.

Welcome to Project Alvarium: Help define the Future of Edge Data.

Endnotes

[1] DellTechnologies.com. November 12, 2019 press release. https://corporate.delltechnologies.com/en-us/newsroom/announcements/detailpage.press-releases~usa~2019~11~20191112-new-2030-goals-for-societal-change-top-dell-technologies-strategic-agenda.htm

[2] Ibid.

[3] Dell EMC. Five Reasons to Choose Dell EMC for Data Protection. June 2017. https://www.dellemc.com/content/dam/digitalassets/active/en/unauth/briefs-handouts/products/data-protection/h15318-top-5-reasons-why-dell-emc-data-protection.pdf.

[4] NIST Computer Security Resource Center. Glossary. Csrc.nist.gov. January 20, 2020. https://csrc.nist.gov/glossary/term/Trust.

[5] Press, Gil. Top Ten Tech Predictions for 2020 From IDC. Forbes.com. October 29, 2019. https://www.forbes.com/sites/gilpress/2019/10/29/top-10-tech-predictions-for-2020-from-idc/.

[6] The LINUX Foundation. New Linux Foundation Effort to Focus on Data Confidence Fabrics to Scale Digital Transformation Initiatives. LinuxFoundation.org. October 28, 2019. https://www.linuxfoundation.org/press-release/2019/10/new-linux-foundation-effort-to-focus-on-data-confidence-fabrics-to-scale-digital-transformation-initiatives/.

[7] Todd, Steve. Building the First Data Confidence Fabric. Blog.DellEMC.com. October 28, 2019. https://blog.dellemc.com/en-us/building-the-first-data-confidence-fabric/.

[8] Todd, Steve. IoT Data Confidence Fabrics. Stevetodd.typepad.com. May 30, 2019. https://stevetodd.typepad.com/my_weblog/2019/05/iot-data-confidence-fabrics.html.

[9] Todd, Steve. Enterprise Trust Insertion and IoT. Stevetodd.typepad.com. August 5, 2019. https://stevetodd.typepad.com/my_weblog/2019/08/enterprise-trust-insertion-and-iot.html.

[10] NIST Computer Security Resource Center. Roots of Trust. Csrc.nist.gov. January 20, 2020. https://csrc.nist.gov/Projects/Hardware-Roots-of-Trust.

[11] Hanna, Steve, Kumar, Srinivas, and Weber, Dean. IIC Endpoint Security Best Practices. Industrial Internet Consortium. March 12, 2018. https://www.iiconsortium.org/pdf/Endpoint_Security_Best_Practices_Final_Mar_2018.pdf.

[12] Lewis, Grace A. Authentication and Authorization of IoT Devices in Edge Environments. Carnegie Mellon University Software Engineering Institute. 2017. https://resources.sei.cmu.edu/asset_files/Presentation/2017_017_001_506478.pdf.

[13] Lewis, Grace A. Authentication and Authorization of IoT Devices in Edge Environments. Carnegie Mellon University Software Engineering Institute. 2017. https://resources.sei.cmu.edu/asset_files/Presentation/2017_017_001_506478.pdf.

[14] Ibid.

[15] Wikipedia.org. Interplanetary File System. December 10, 2019. https://en.wikipedia.org/wiki/InterPlanetary_File_System.

[16] Condon, Stephanie. Intel aims to scale IoT deployments with Secure Device Onboarding. ZDNET.October 3, 2017. https://www.zdnet.com/article/intel-aims-to-scale-iot-deployments-with-secure-device-onboarding/.

[17] Broudy, Alex. Self-sovereign identity and GDPR. Hackernoon.com. March 7, 2019. https://hackernoon.com/gdpr-and-self-sovereign-identity-what-lies-ahead-56de20055d5c.

[18] Microsoft. Decentralized Identity. Own and control your identity. Microsoft.com. 2018. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE2DjfY.

[19] Ibid.

[20] MIT CSail. Decentralized Information Group Home Page. December 10, 2019. https://www.csail.mit.edu/research/decentralized-information-group.

[21] Ibid.

[22] Wikipedia.org. Data Lineage. December 11, 2019. https://en.wikipedia.org/wiki/Data_lineage#Data_provenance.

[23]Elkhodr, Mahmoud, and Alsinglawi, Belal. Data provenance and trust establishment in the Internet of Things. October 18, 2019. Wiley Research Article. https://onlinelibrary.wiley.com/doi/pdf/10.1002/spy2.99.

[24] Ortmann, Mike. Business Data Lineage vs. Technical Data Lineage: What's the Difference? Infogix.com. September 18, 2019. https://www.infogix.com/business-data-lineage-versus-technical-data-lineage-whats-the-difference/.

[25] Feng, Lu. Data integrity and data lineage by using IOTA. Feng.blog(). April 16, 2018. http://feng.lu/2018/04/16/Data-integrity-and-data-lineage-by-using-IOTA/.

[26] Time Sensitive Networking. Wikipedia.org. January 23, 2020. https://en.wikipedia.org/wiki/Time-Sensitive_Networking.

[27] Eitel, Lisa. The new CC-Link IE TSN — for time-sensitive networking for open industrial Ethernet. MotionControlTips.com. December 5, 2018. https://www.motioncontroltips.com/the-new-cc-link-ie-tsn-for-time-sensitive-networking-for-open-industrial-ethernet/.

[28] Furlong, Judy. Dell EMC Distinguished Engineer. January 23, 2020.

[29] Bowers, Kevin. RSA Labs Project Iris: Edge Monitoring and Analytics for IoT. RSA.com. August 7, 2019. https://www.rsa.com/en-us/blog/2019-08/rsa-labs-project-iris-edge-monitoring-and-analytics-for-iot.

[30] Ibid.

[31] Zolfonoon, Riaz. RSA Distinguished Engineer. January 20, 2020.

[32] IETF. Manufacturer Usage Description Specification RFC 8520. IETF.org. January 20, 2020. https://datatracker.ietf.org/doc/rfc8520/.

[33] Bollella, Greg. So You Got '42', Now What? VMware.com. December 11, 2018. https://blogs.vmware.com/edge/2018/12/11/so-you-got-42-now-what/.

[34] Blackborow, Josh, and Christakis, Sophia. A Holistic Security Strategy Is Essential To Edge and IoT Success. Forrester. A commissioned study conducted by Forrester Consulting on behalf of Dell, January 2020.

[35] Press, Gil. A Very Short History of EMC Corporation. Forbes.com. September 6, 2016. https://www.forbes.com/sites/gilpress/2016/09/06/a-very-short-history-of-emc-corporation/.

[36] Todd, Steve. Enterprise Trust Insertion and IoT. Stevetodd.typepad.com. August 5, 2019. https://stevetodd.typepad.com/my_weblog/2019/08/enterprise-trust-insertion-and-iot.html.

[37] Swinhoe, Dan. The biggest data breach fines, penalties and settlements so far. CSOOnline.com. October 31, 2019. https://www.csoonline.com/article/3410278/the-biggest-data-breach-fines-penalties-and-settlements-so-far.html.

[38] Best, Adam. The Importance of Trustworthy Data! Validate and Verify. Scalable.com. March 21, 2017. https://www.scalable.com/importance-trustworthy-data-validate-verify/.

[39] Short, Jim, and Todd, Steve. What's Your Data Worth? MIT Sloan Management Review. March 3, 2017. http://ilp.mit.edu/media/news_articles/smr/2017/58331.pdf.

[40] Accenture Research Report. Value of data: the dawn of the data marketplace. Accenture.com. September 7, 2018. https://www.accenture.com/us-en/insights/high-tech/dawn-of-data-marketplace.

[41] Laney, Doug. Why and How to Measure the Value of your Information Assets. Gartner Research Report. August 4, 2015.

[42] Laney, Doug. Infonomics: The New Economics of Information. Caserta Presentation. December 16, 2019. https://static1.squarespace.com/static/5699038c3b0be3b876587b6f/t/5ce33b53362fa200016b8c4e/1558395735709/Infonomics.pdf.

[43] Blackborow, Josh, and Christakis, Sophia. A Holistic Security Strategy Is Essential To Edge and IoT Success. Forrester. A commissioned study conducted by Forrester Consulting on behalf of Dell, January 2020.

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.