LEARNING MADE EASY

Hewlett Packard Enterprise
Special Edition

Data Centric Architecture

for dummies®

Why it will help your business

What components and principles it is made of

How to master your transformation

Brought to you by:

Hewlett Packard
Enterprise

Stefan Brock
Bernd Bachmann
Johann Kortsch
Lukas Ritter

Data Centric Architecture For Dummies

Stefan Brock, Bernd Bachmann,
Johann Kortsch and Lukas Ritter

Data Centric Architecture

for dummies®

Sonderausgabe

Hewlett Packard
Enterprise

WILEY

WILEY-VCH GmbH

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über `http://dnb.d-nb.de/` abrufbar.

© 2022 Wiley-VCH GmbH, Boschstraße 12, 69469 Weinheim, Germany

All rights reserved including the right of reproduction in whole or in part in any form.

Wiley, the Wiley logo, Für Dummies, the Dummies Man logo, and related trademarks and trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries. Used by permission.

Cover photo and all figures in the book: © Hewlett Packard Enterprise
Layout: Straive, Chennai, India

At a Glance

Table of Contents

Introduction

The world is full of buzzwords: Internet of Things, Big Data, Machine Learning, Artificial Intelligence, Digital Transformation, and so many more. All of them are intended to bring added value and, ultimately, a better future. Many companies and organizations are starting their digital journey, like a trip to outer space. But are they prepared for it? Do they have the right preconditions? Do they have the fuel of their journey - their data - accessible and under control?

Many initiatives realize halfway through that the idea of generating value from data is not that simple. The desired value creation does not come from the data itself. Access to data is a necessary prerequisite for the success of the journey but correct management of data is equally important. Data is like a raw resource which needs to be refined in order to reach its full potential. We can mine the ore but there are many additional steps required before we have a car. Not only processing but also logistics. You need to introduce those data refinement (processing) and management (logistics) steps in your enterprise architecture to monetize data in a targeted and efficient manner. That's why some smart people have come up with the idea of Data Centric Architecture. Some people say, just a new buzzword. Others might say, let's forget about data, let's focus on customer value. HPE answers those questions and describes how data should be managed, how data governance should be handled, and how data can generate value without violating the privacy requirements of individuals.

You might think, however, I am not in the open and green field. There is reality – the brownfield of legacy applications and traditional architecture. How can I overcome that? True **Data Centric Architecture** is not a greenfield-only approach; it has to cope with the existing legacy environment. Only if the investments, know-how, and data are leveraged can your company realize the full potential of digitization. Therefore, it is essential to know that data-centricity enables easy and fast integration of legacy data into the Data Centric Architecture. User-centric apps and analytics services can be powered by data from IoT, digital applications and the legacy world.

Overall, the critical aspects in Data Centric Architecture are:

- ✔ The know-how about data and metadata
- ✔ Simplification of connection
- ✔ Support for near-real-time access wherever possible
- ✔ Strong governance to protect and provide data and enable role-based access
- ✔ Controlling legal compliance and retention periods
- ✔ Potential data monetization

About this Book

This book helps you, if you are a practitioner or manager or just interested to gain insights into the Data Centric Architecture and its value to the business and digital transformation.

- ✔ You can take an in-depth look at the key principles and components of the Data Centric Architecture.
- ✔ You will gain insight into the transformation journey to data-centricity.
- ✔ You will look at some myths about the approach.
- ✔ You will get comprehensive information on how HPE can help organizations like yours prepare for this change.

Foolish Assumptions about the Reader

To understand this book, you should have a basic understanding of digital services and information technology. Some more detailed technical stuff will be marked as such. The focus will be on the business value generation and the fundamental principles everyone who works with Data Centric Architecture should know about. This book is designed in a modular way so that you can jump between the chapters or skip what might be too technical for your interest.

Icons Used in this Book

As in all Dummies books you'll find a couple of icons in the text:

The magnifying glass leads you to definitions of technical terms.

Wherever this icon turns up you'll find useful advice or are remined to note something.

This is a warning sign. Pay attention whenever you see it in this book.

Here we go into details, provide background information or explain rather technical matters

IN THIS CHAPTER

The main idea and technical overview of Data Centric Architecture

Key feature highlights of Data Centric Architecture

Pain relievers of Data Centric Architecture

Business value of Data Centric Architecture

Chapter 1
Introducing Data Centric Architecture: How Does It Help?

You might already have some idea of what Data Centric Architecture could look like. In the following we will introduce you to our approach as well as potential benefits for you and your company.

Data Centric Architecture »in a Nutshell«

Putting data at the core of your business and IT enables the organization to compete in the digital era. This idea of data-centricity is at the heart of Data Centric Architecture.

So, what does Data Centric Architecture actually look like? The HPE approach to data-centricity can be described in a picture with ten building blocks, which are briefly introduced in the following and described in detail in Chapter 3. Doing this right leverages the existing data and information gathered. Moreover, it brings the advantage of including IoT and other digital services in the portfolio. The intrinsic knowledge and value of the company can be leveraged.

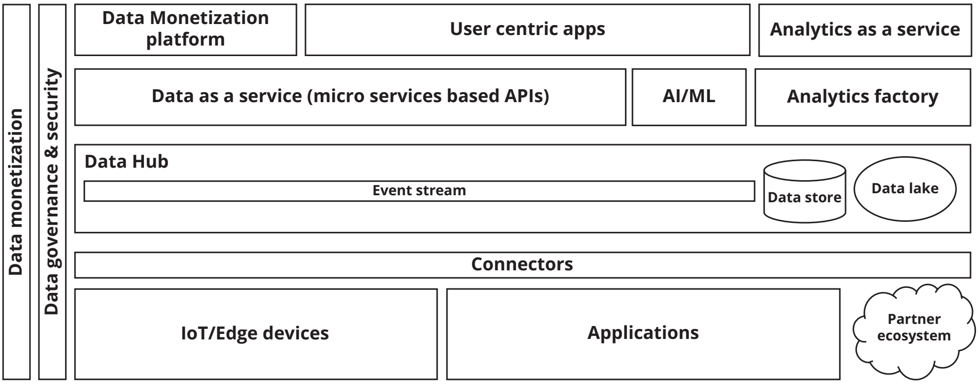

Figure 1-1: Blueprint of Data Centric Architecture

The core components of our Data Centric Architecture are a **Data Hub**, a capability for **data governance**, and state-of-the-art **security mechanisms**.

Data Centric Architecture consists of the following components:

- ✔ **Data Hub:** The Data Hub represents a data virtualization layer in the sense of a digital twin. This hub is connected to the central data management component, including metadata management, access control, data protection, and the ability to share data in a wider ecosystem. All data in the Data Centric Architecture flows through the Data Hub.
- ✔ **Streaming Technology:** The most important element within the Data Hub is a streaming technology for exchanging data between the different data sources and data consumers. All data is sourced as an event stream, even files from applications.
- ✔ **Connector Layer:** The connector layer offers a wide variety of connectors to enable the data sourcing from any source, starting from native event sources such as sensors in the IoT to legacy applications such as ERP systems. The connectors feed this data as an event stream to build the digital twin, allow easy transformation and enable stream analytics.
- ✔ **Data storage and data lake:** Besides streaming capabilities, the Data Hub offers data storage and data lake capabilities. This can include a relational, graph, NoSQL and other databases, file stores, warehouses, and archives. It has the transformation capabilities to transform the data between different data layers and formats.

- ✔ **Data governance:** The fundamental building blocks of the data governance component are the metadata management and the data catalog components, combined with an access management system that works in conjunction with Active Directory and enterprise identity management to control access to specific data streams and tables based on classification and tagging of data.

- ✔ **Analytics Factory and AI/ML repositories:** In the capability analytics factory tools and environments are provided to data scientists for analytics. A central repository of AI and ML components enables fast and appropriate learning on the data. The access can be offered to a broader user group such as management and business matter experts via analytics as a service interface.

- ✔ **Microservice layer:** Data from the Data Hub can be consumed by any application on the network and by microservices in the microservice layer. Access control management controls authorization and is based on the meta information and specific rights assigned. The microservice layer uses access to transformed data streams and offers micro UIs or APIs on this data within the company or the wider ecosystem.

- ✔ **User-centric apps:** The micro UIs can be combined in user-centric apps to perform a role-specific task and give data access to allow data contribution. This enlarges the value and quality of data in the hub.

- ✔ **Data monetization:** Finally, the Data Centric Architecture can be enhanced by data monetization capabilities to exchange the information and data in the broader ecosystem with precise controls, ensuring data sovereignty and the ability to monetize.

The architecture approach is open-source compatible, as we believe in the value of the community approach.

Key Feature Highlights

In summary, Data Centric Architecture provides the following key features:

- ✔ **User-centric apps:** It enables user-focused and decentralized app development based on microservices.

- ✔ **Analytics as a Service:** It leverages the value of data and information in the organization using AI/ML capabilities and containerized data management.

- **Data governance:** It enables centralized data governance through transparent data flows, access management, traceability, and service management processes.

- **Unified interface:** It leverages a unified technology for data access and a microservice approach, which eliminates the duplication of the same/ similar functions.

- **Digital twin:** It enables a digital twin of information as the Data Hub of the organization.

Pain Relievers through Data Centric Architecture

Data Centric Architecture delivers fast and high value to the business and the organization, while counteracting the pain points of traditional enterprise architecture approaches.

The key pain relievers are:

- **Data catalog:** The catalog enables transparency about data and metadata and is the entry point for defined data access.

- **Strong data governance and protection**. A core component is access management, monitoring, and control of data throughout its lifecycle in the Data Hub.

- **Cost reduction and higher flexibility:** This is achieved by using one interface per (new) application with metadata design.

- **A publish once and subscribe by many approach instead of point-to-point interfaces:** Instead of getting the data from at least two vendors, you select the data from the catalog and subscribe. Where approvals are necessary, you talk to data governance.

- **Effective application of advanced analytics/AI enabled by an availability of digital twins in the Data Hub:** The dream of the data scientists comes true. They can get an overview of all data in the organization and correlate ex-post and real-time.

- **Data Hub is based on event bus – event-centric data exchange:** Thus, all data is in the same format, independent of whether individual subsets come from your different IoT devices or application data sources.

- ✔ **Full scalability** and seamless integration of the event bus from edge to core/cloud. You even get the single namespace and management capabilities from edge to cloud with the right tools – in multi-tenancy.
- ✔ **Enablement of role-centric UIs**. Instead of requesting your workers on the street, tarmac, or shopfloor to use multiple applications for data entry, you enable them to do the same as in private life. That means, you provide them with a simple and role-centric UI for their job, through which they can receive all relevant data and information and accept or complete tasks.

Through these pain relievers, Data Centric Architecture delivers significant value to the business.

Business Value

Data Centric Architecture not only lowers your IT costs but also delivers direct value to the business by:

- ✔ **Introducing new business models:** It empowers your organization to build new digital business models through data monetization and artificial intelligence.
- ✔ **Reducing the operating costs:** It enables control and reduction of operational effort by reducing complexity and data duplication between allocations and workloads, and provides the ability to consolidate your application landscape and the variety of integration interfaces/technologies.
- ✔ **Giving real-time insights:** It enables the organization to derive insights from data across all departments at the time of data generation through AI and analytics dedicated frameworks.
- ✔ **Providing an end-to-end view:** It enables transparent management of data from creation to archiving.
- ✔ **Granting data sovereignty:** It protects the business from risks by taking control of data access, data processing, and required processes, and it ensures full control over data and compliance (for example, GDPR rules).

Now that you understand the business value of Data Centric Architecture and how it counteracts the pain points of traditional enterprise architecture approaches, it is now time to dive deeper into the actual realization of Data Centric Architecture.

IN THIS CHAPTER

Basic principles for DCA

Security principles for DCA

Operations principles for DCA

Chapter 2

Realizing my Data Centric Architecture: Where to Start?

The following chapter takes an in-depth look at eight design principles for DCA. If you are interested in a high-level overview only, you should move directly to Chapter 3.

The Key Principles to Start With

How do you implement data centricity in your legacy environments, and what do you need to consider to avoid ending up with a proof of concept (PoC) that doesn't go live? And what do you have to do to ensure your architecture can keep up when it gets to a higher maturity level?

We at Hewlett Packard Enterprise like to work agile and incrementally. Instead of PoCs, we think and work with minimum viable products (MVP). And to work agile, it is good to have some guiding principles. We will describe some of them in the following pages. All of them are essential, but we recommend that you start with a few, then grow with increasing loads and maturity, in order to define additional principles to cope with the specific needs of your business and environment.

The following principles 1 to 9 are the must-haves to full scalable Data Centric Architecture!

Basic Principles for DCA

The following sections describe basic design principles which are applicable to any Data Centric Architecture.

Design Principle 1

»All data need to be exchanged via the Data Hub as an event stream!«

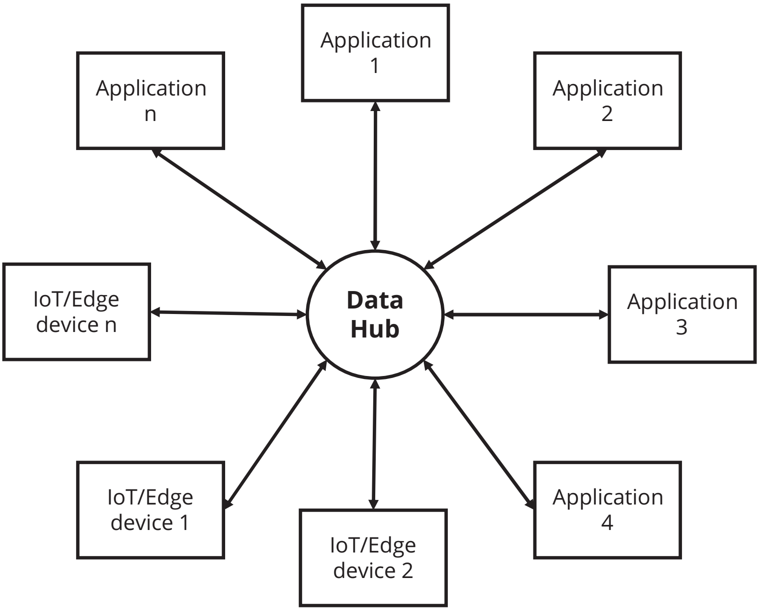

Figure 2-1: Design Principle 1

Let us start with the difference from the connection with traditional enterprise application integration. When connecting the data sources into the Data Hub, all reasonable data need to be pushed into the event stream as indicated in the above figure.

For an application, this is the case for all data from its logical data model. In the case of an IoT device, all status information is collected in the device. In any case the data shall be pushed in the source format of the application, no transformation should be done on the application level. There are different options to source these data, such as MQTT for IoT or Kafka Connect, CDC, Nifi, and several other technologies for applications. The closer the sourcing is to the primary data model of the producer, the faster and cheaper is the integration.

Simultaneously, the Data Hub incrementally becomes the »digital twin« of all enterprise/ecosystem data.

Design Principle 2

»Connect once – publish all data via the Data Hub event stream.«

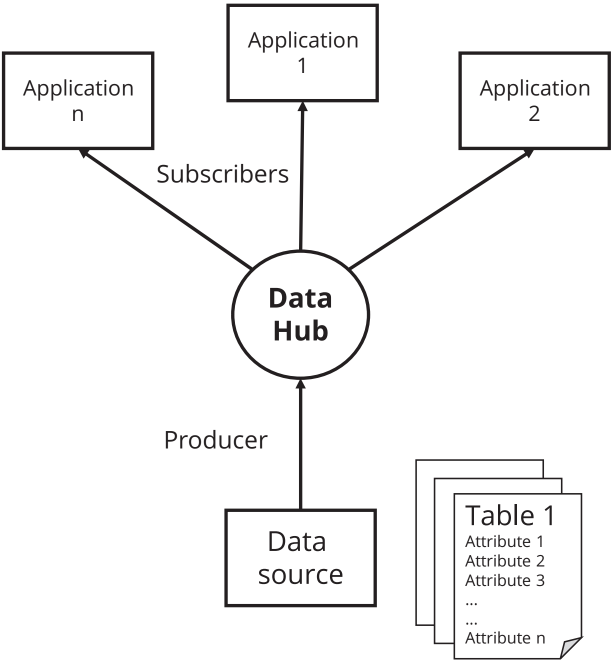

Figure 2-2: Design Principle 2

This principle is the hardest to achieve in architecture development. Most organizations and IT departments are trained over the years to deliver data only with a purpose. This is why there are these wonderful hairball architectures, which limit us from leveraging the value of our data. But time has changed, and certain technologies have been brought into play that enable data exchange and protection on a scale that was not thinkable before. In general, every relevant data is any log from IoT, and any change in data from the logical data model in applications needs to be transferred via the hub.

Your first and major challenge on your road to data-centricity is the »connect once principle«. If you cannot convince your organization within a given time frame to go in this direction, forget about Data Centric Architecture!

This principle is most important, as you do not want to re-adjust the interfaces with each new request for additional information. If you limit the data push into the Data Hub,

✔ you limit the cost savings on the interface side and

✔ you reduce the value of the simple-to-subscribe data pool.

The more different consumers subscribe to the Data Hub, the more the profitability increases.

Moreover, the connect once principle enables fast learning on the data, the ease to access for new digital services, and the generation of the digital twin.

Design Principle 3

»All producers need to publish the data and metadata information.«

As soon as you have overcome establishing the connect once principle, you can start to create value. Ensure that all published data is documented in a central metadata management tool with all relevant meta information.

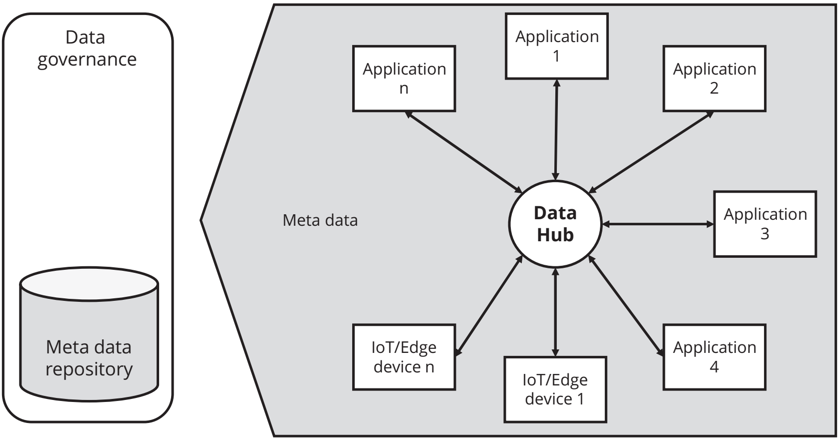

Figure 2-3: Design Principle 3

You can start small and grow step by step.

- ✔ Start with name, semantic and classification.
- ✔ Correlate data/identify real master sources.
- ✔ Add additional metadata.

Let us begin with the data name, semantic, and maybe the classification (public data (C1), internal data (C2), confidential data (C3), and secret data (C4)). This step will already help a lot to increase data security. You know the source and have some meta information. In most cases, you will see that many people will start to see the value because more and more data sources have been connected.

The next level will come step by step – and will depend on your environment. Maybe you start with correlating the data to recognize where the real master source of your data is (which again makes good metadata to be documented), or you add other meta information such as personal identifiable information (PII), data automatically or manually collected, real-time information, periodically gathered information, trust level, GDPR information requests, GDPR deletion requests, keeping period, etc.

These metadata allow access to specific topics and data in the Data Hub. Only consumers who are entitled to read PII data will get access to topics with PII data. Only topics that are allowed to read secret data of a specific domain will access these data.

How can you realize this? To get the ball rolling fast every application or data source owner is documenting the data pushed into the Data Hub with each »producer«, at least for some metadata such as semantic, classification, and/or PII relevance.

Security Principles for DCA

While creating value with data-centricity we should not forget about security. Let's take a closer look.

Design Principle 4

»All data need to be encrypted.«

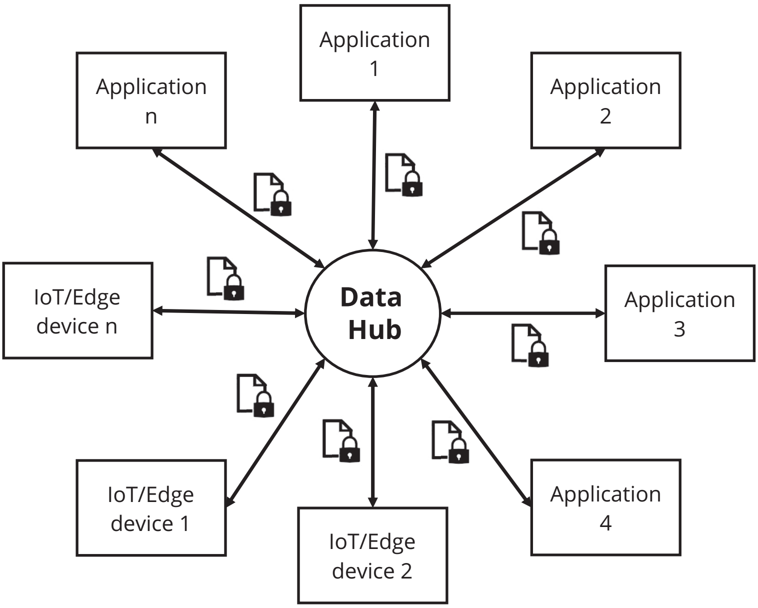

Figure 2-4: Design Principle 4

Having metadata in place, you come to the next level of quality and protection. As you want to generate the »digital twin« of your organization, it is mission-critical to gain the trust of the data owners, stewards and organization. It is like giving money to a bank: Would you trust a bank that stores your money in a standard locker? Certainly not.

You want to see that your values are best protected – in a safe. Ideally with access only for you or for people you trust. There is no difference between you and the owners of data.

So, if you want to convince the data owner to trust you – and your Data Centric Architecture – you have to ensure that your architecture delivers the »Fort Knox« capability for data he is entrusting you. No one should be able to access the data who is not entitled to by the data owner.

Therefore, the DCA ensures that all data are encrypted in transfer – ideally in rest as well. The key will only be delivered to the consumer, who is entitled to receive it. No one managing the Data Hub or developing on it will read the content, only the subscribers who get the key. Using the DCA approach you can enable DevOps operations on the Data Hub without allowing them to read the plain text of the data. This is more than what is possible in most applications.

The decryption key will only be given to data, where the data classification and the relevant metadata tags you defined for your organization are set. As long as there is no metadata provided and approved, the classification level will be »unclassified«; let us call it »C0«. This means that no decryption key will be given to anyone. The data can still be streamed into the Data Hub, but no one can subscribe and read them. This allows the producer, data stewards, and data owners to set the metadata while data are already streamed. The data are protected in the »bank« Data Hub.

Design Principle 5

»Decryption keys are available only for managed classified data.«

Let us look further into our Fort Knox approach. The Data Centric Architecture embraces a zero-trust approach. We assume that everyone wants to steal our »oil of the future« – data and knowledge. We protect everything from everyone. Then we start to give access to the ones we trust and are allowed to work with our data.

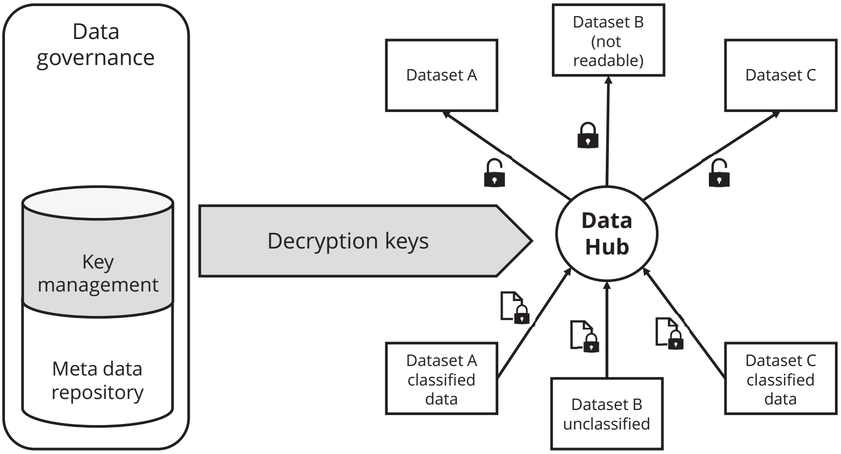

Figure 2-5: Design Principle 5

In the governance capability, as illustrated above, there are functions for metadata management and access control to manage the access to data. It is like the guard at the entry, combined with the key of the bank director and our personal code for our locker and account. We enable access only under clearly defined conditions and checks.

A minimum viable product, for example, can combine the metadata information in the data catalog with role-based access of systems and manage the access on a topic level, depending on metadata. This can be enhanced by integrating identity and access management (IAM) and/or Active Directory (AD) information in the next maturity stage. It can be used for direct access via microservices on a personal level.

A potential use case could be the need to ensure that you can only access the European HR date if you are situated in Europe. As a vision, you could enable access to secret data only from specific devices and from given regions in the world. But that is another story.

Operations Principles for DCA

You know what to build and how to harden it, but how do you actually work with Data Centric Architecture? Let's find out.

Design Principle 6

»Access to topics is managed by Data Hubs governance capability – for each security classification, there is an extra topic.«

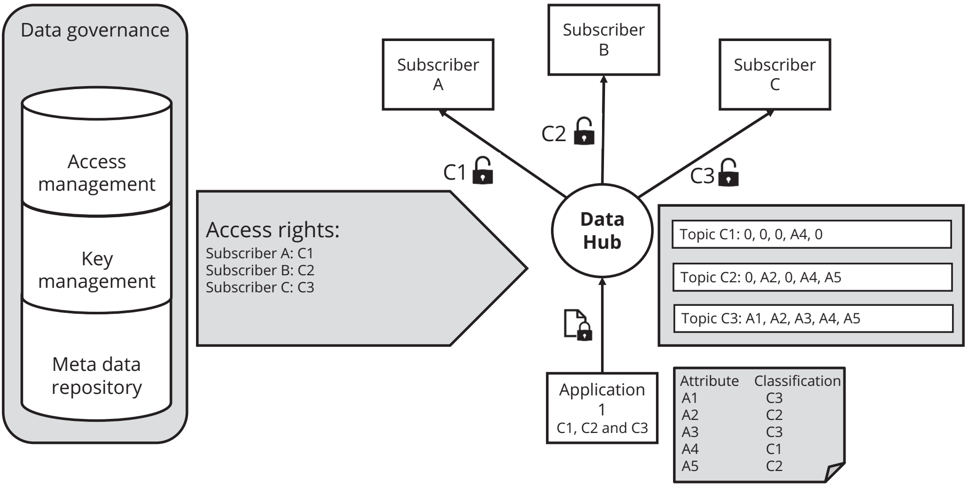

Figure 2-6: Design Principle 6

In general, we want to make our data accessible on an object level, for example for someone who wants to access product information. But not all data should be visible for everyone. Some attributes of the product data are internal, some are confidential, some could be secret. There might be a lot more metadata that would be relevant, as said before. But, let us start small: for the moment, we only want to classify our access control by the security classification (feel free to develop it further for your specific needs). An illustration of the following data classifications can be found in Figure 2-6.

You intend to facilitate access to data and ensure that subscribers who are allowed to read only internal data (C2) are not getting access to confidential data (C3). The fastest way for that depends on the classification needs of subscribers. For example, you could produce parallel topics with the same schema.

The Data Hub transformation logic ensures that in the stream with C2 data, all fields with attributes with a higher classification are shown as zero, i.e. the schema is not changed but confidential entries to the payload are obscured.

Design Principle 7

»Each interface needs to be monitored and be part of the CMDB.«

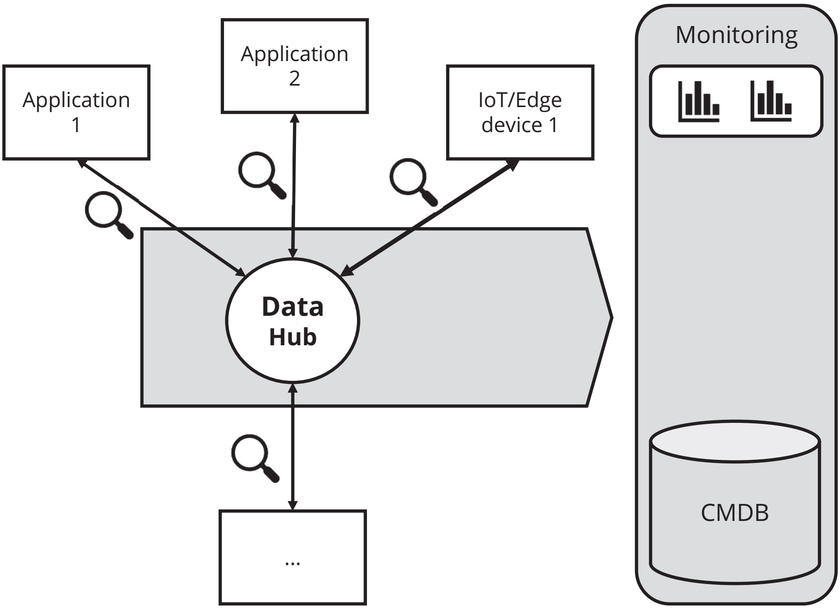

Figure 2-7: Design Principle 7

As the Data Hub will become a major component in the communication between intelligent devices and all applications, highest availability is required. Platforms like Apache Kafka and HPE Data Fabric (formerly MapR) are built for these requirements. If you can use connectors that are embedded in Kafka, like Kafka Connect or Kafka Streams, you are lucky. They can leverage the size and high availability of the stream.

But for sure, you will get into the situation where the producer or consumer is not yet ready to stream the data. RestAPIs or just CSV files might be available instead. You do not want to lose the momentum you have generated in the organization. The management has decided to support you – even the connect once principle was bought. There needs to be a way to get any data source into the hub.

And there is one! There is no data source that cannot be connected to a stream. Excellent and easy fast starters include tools like Nifi. So you are always able to succeed. It is essential to keep in mind that these tools need to have sufficient built-in redundancy, as it does not come for free like with native tools. So always consider the redundancy you need!

Even if designed to run on redundant nodes, and even for Kafka, there are situations where an interface may not be acting as it should. The reason can be on the connectivity side, or a change in the sourcing or consuming application itself.

We learn from best practices that it is crucial to have every single interface as a configuration item (CI) in the configuration management database (CMDB). You can monitor and open tickets not just to the hub but to a dedicated interface. Besides, we recommend you to monitor each interface related to specific key performance indicators, which need to be assigned to them in alignment with the owner of the connected application. In a higher maturity stage, you can generate these thresholds with machine learning (ML) or artificial intelligence (AI) to leverage the knowledge of daily operations.

Design Principle 8

»Interfaces need to be manageable/controlled.«

IT landscapes are changing continuously and fast. Therefore Data Centric Architecture has to be adaptive to change and facilitate to support flexibility. Loose coupling, we call it. The producers might change their data model, but the end-to-end experience must not suffer.

For this reason, we at HPE have chosen AVRO schemas and the schema registry to support schema evolution. Nevertheless, the producing and consuming systems might end up struggling for some unforeseen reason, and it is important to prevent the overall ecosystem from suffering a failure in one system.

One aspect is the change of producing or consuming connector interfaces. The changes should happen smoothly and under control. Moreover, it is essential that interactions can be done in daily operational situations with a simple click from first- and second-level operations support. Therefore, you need to build in basic functionalities such as »stop« and »restart«, and make them available to the service desk and second-level operations.

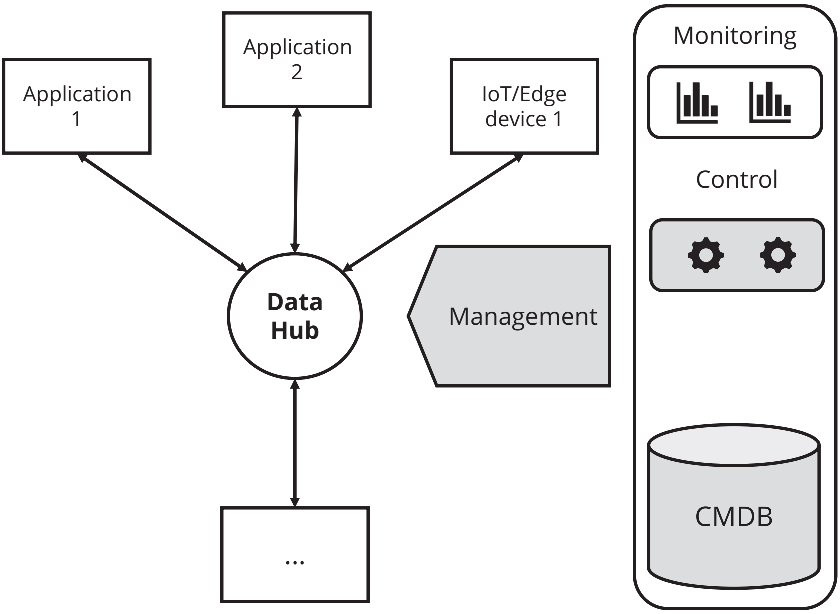

Figure 2-8: Design Principle 8

If you are making changes to the application landscape, it may be useful to send only a sample of events through the Data Hub before restarting a connection, so that the test can be run without much clean-up of exceptions after the fact. Therefore, in a higher maturity stage, the interfaces should support the basic restart function and send a selected number of events before opening to regular operations.

This will enable the operations and DevOps teams to react fast – and it generates the basis on which the automation can be built quickly. The reaction to failure patterns and exceptions can be trained to the system and a more and more autonomous environment can be created.

IN THIS CHAPTER

- Data sources, connectors, and the Data Hub
- Data governance and security
- Data as a Service
- User-centric apps
- Analytics Factory, AI/ML/NLP, and Analytics as a Service
- Data monetization

Chapter 3

Components: What Are the Key Building Blocks of DCA?

Having understood the key principles of DCA, let us look into the ten building blocks of DCA, which you can find in Figure 3-1.

Component 1: Data Sources

Data sources for the Data Centric Architecture could be anything that delivers data – independent of whether these are applications, log files of applications or servers, IoT sources like doors, cars, intelligent clothes – or social media and others.

Everything that is consumed in the organization or ecosystem is a relevant data source for the digital twin, represented in the Data Hub.

You need to define and manage which data will be transferred via the streaming bus of the hub. In the case of applications, the logical data model of the source application and any changes to this data could be your starting point. In the case of log files, the starting point would be every log file of an application or device. For IoT sources, the starting point would be all status information of any sensor and actor.

This sounds like overkill. But the point is, when you start with the Data Hub, you cannot know which information will be relevant for a machine learning algorithm. Similarly, you do not have time for a lengthy process for data access once you need the data. It needs to be readily available, and you need to rely on the source application or device vendor.

Once data is in the stream, specific deletion tools triggered by the data governance delete the information after a given time if they are not sourced from consumers. The use of deletion tools has a couple of advantages:

- ✔ They ensure that data protection is observed and regulations are complied with.
- ✔ They allow continuous control of the functionality of a connector (and, if used correctly, of the source application).
- ✔ They facilitate the extension of data subscription (just change the deletion process).
- ✔ They enable the use of machine learning for more insights, using all available data.

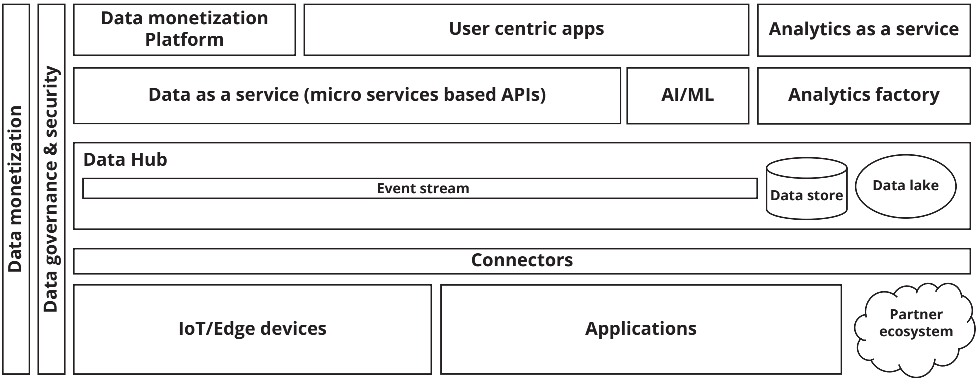

Figure 3-1: Components of Data Centric Architecture

Component 2: Connectors

Connectors can be of any type. Using Kafka you would use KConnect, Change Data Capture (CDC), KStreams, KSQL, MQTT, Nifi, or others. The possibilities are plentiful, and they should not be a showstopper to get the relevant data into the Hub. Besides, there are specially licensed connectors available for nearly every application in the market.

You would certainly intend to get as much data as a real-time event stream into the hub as possible. So KConnect is an excellent start. Nevertheless, in most environments, REST APIs are available and easy to get data from as a quick start. In this case, Hewlett Packard Enterprise have solved nearly every challenge with Nifi. You could even access web services with NiFi.

Regarding the connector technologies, for transformation and learning purposes it is essential to start with a limited set of connector types (Kafka Connect, CDC, NiFi, MQTT, etc.). This way,

✔ you learn fast,

✔ you increase the capability of the connectors (monitoring, rollback, etc.), and

✔ you facilitate initial DevOps operations.

There should be no excuse not to connect to the Data Hub. Therefore, the lean-agile architecture team will always allow a temporary exception, even for proprietary connectors, to get the data quickly. This is where the value is generated.

In the long run, the architecture aims to enable near real-time streams of all data sources to get the maximum value out of the digital twin in the Data Hub.

Component 3: Data Hub

The Data Hub is not a monolithic data store. As has been shown, it comprises different types of storage, depending on the specific needs in the environment. There will be at least a graph database, a data lake, and some NoSQL storage beside the streaming service. These stores will be in multiple instances and maybe globally distributed, but the central knowledge and metadata management are controlled within an organization and federated in a wider ecosystem.

At the heart of the Data Hub are the event streams. The topics are connected to a schema registry and use Avro schemas to document the formats. This enables schema evolution within the hub. If a producing data source's format is changed, this could be done on the fly, under the condition that there is no format change but only an extension of the schema with new columns. The overall data exchange will continue and will be based on the latest information in the schema registry.

Within the Data Hub, the data will be managed in separated namespaces, as can be seen in Figure 3-2.

There is one namespace for the source data, where all producers publish their data. The other one is the consumer namespace, where the data is being streamed according to the enterprise data model – or microdata domain model.

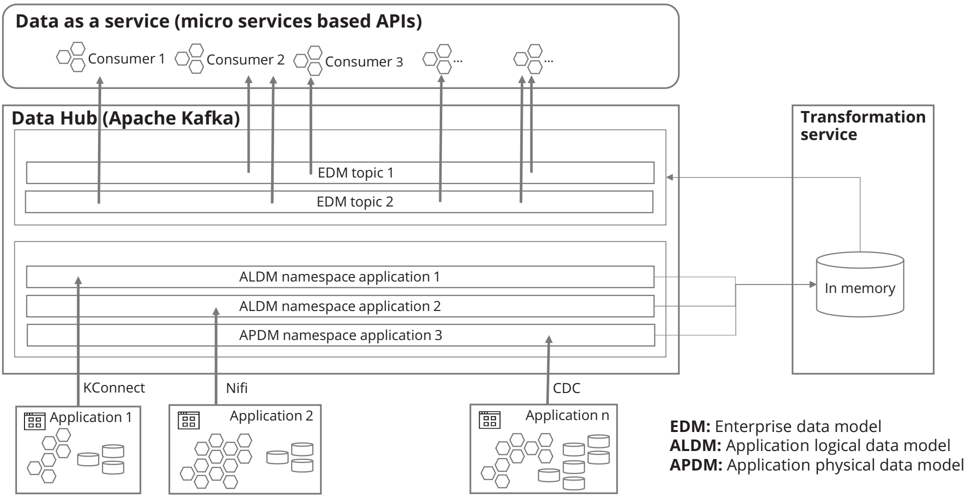
Data as a service (micro services based APIs)
Consumer 1
Consumer 2
Consumer 3
...
...
Data Hub (Apache Kafka)
EDM topic 1
EDM topic 2
ALDM namespace application 1
ALDM namespace application 2
APDM namespace application 3
Transformation service
In memory
KConnect
Nifi
CDC
Application 1
Application 2
Application n
EDM: Enterprise data model
ALDM: Application logical data model
APDM: Application physical data model

Figure 3-2: Namespaces in the Data Hub

We recommend that you go with the source model as close as possible to the logical data model. This allows you to reduce the transformation effort and alignment needs. Moreover, it brings you to a better insight into what data are where and how they are structured and named.

For Hub management, you will need some connections to the active directory, PKI environment, and components from the data governance side.

Depending on use, you can realize the streaming part with Kafka or the HPE Data Fabric (former MapR), which supports Edge to Cloud and multi-tenancy (see Figure 3-3).

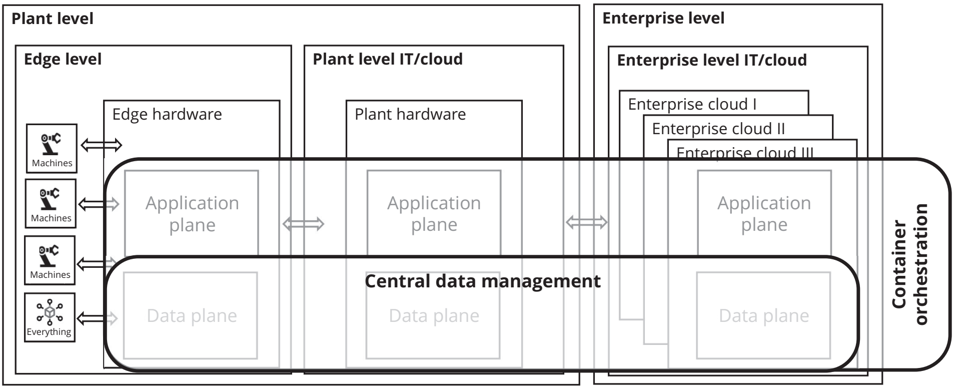

Figure 3-3: Edge to Cloud Data Hub

Component 4: Data Governance and Security

Core components are

- a data catalogue,
- a metadata management system, and
- access control.

In the open-source community, there is a strong value provider in Apache Atlas. More and more licensed products are being introduced to the market that extend Kafka with friendlier user interfaces, prebuild process and role models, and a lot of other functions. But in the core, Apache Atlas is still a good fit for the metadata management and data catalogue to start – the APIs help to connect to most tailored products. In Atlas, you can manage all data that has been published, transformed, and sourced via the Data Hub. This includes not only the streaming data but also data in the data stores or data lake. The metadata information helps to document the source information, the data catalogue, semantics, classifications, PPI relevance, trust level, accuracy, and use case-specific data.

Atlas brings the ability

- to manage microdata domains as well as monitor and control data lineage,
- to track the changes on the source data (schema evolution), as you can connect Atlas with hooks into the Kafka Schema Registry,
- to follow up the changes with workflow services or external data catalogue systems for operational management and metadata management.

Apache Ranger is a very good fit for role-based access management, as it supports access control on Apache Kafka and a lot of other components, based on metadata tagging in Apache Atlas. In combination with an identity provider (IDP), you can manage the access of systems to topics. For overall monitoring, especially logs, Elastic Search (ELK) combined with Kibana is a good start. For metrics, we recommend to use Prometheus and Grafana.

Component 5: Data as a Service

Data as a Service is realized by microservices. You can develop them in any language you may find appropriate. They connect to the hub and send their information as events back into the Data Hub. According to the principle, there are no data that are not shared via the Data Hub.

Recently some products with automated app engines, model libraries, and low code realization of apps have become available and have proven their value. You might want to analyze and select one as a recommended platform for microservice development. In particular, a model library with the ability to send back the built models into your architecture repository or service management system will improve the manageability and speed up the learning curve of your organization.

Do not underestimate: If your users and internal customers are enabled to realize fast, easy, and reliable apps, they will love your Data Centric Architecture even more.

Component 6: User-Centric Apps

In the best case, you can build user-centric apps out of the micro UIs of the microservices. Generally, low code platforms offer an excellent set to build fast and good UIs. Instead of the user having to follow complex and overloaded application workflows, the applications are specifically built to match the users' needs, offering an easy-to-use and intuitive user interface with only the functionality that is needed for the specific user (see Figure 3-4).

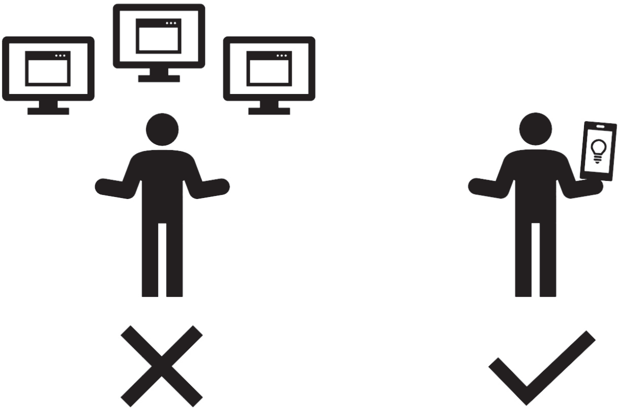

Figure 3-4: User-centric UI

Component 7: Analytics Factory

Within the analytics capability, you can choose different tools to offer your users for data analysis. Ultimately, it's clear that data scientists will use whatever is available and test different views. Therefore, you need to differentiate

- a toolbox for scientists, which is open and wide-ranging, and
- a limited predefined set of tools to be used by management, data governance, marketing, and other users with less IT expertise.

You need to manage user access rights to the data through the central data governance component.

For containerization, there are many good tools within the container environment, such as the HPE Ezmeral Container Platform (formerly BlueData), Dataiku, Spark, TensorFlow, etc.

Based on the huge amount of data in the Data Hub and the ability to steer the keeping period by tuning the deletion tools you can implement new algorithms easily. These algorithms will help you to identify interdependencies that a human being is not able to highlight.

Component 8: AI/ML and NLP

This building block contains all tools and methods you will use as a proprietary repository in your organization. The knowledge of which type of models helps most in which scenario will become a competitive advantage for your organization and the trained algorithms for your specific use cases.

Component 9: Analytics as a Service

The more you enable your organization to learn from the data, the better the organization will become – and the more you can show value generation with Data Centric Architecture.

Select easy-to-use tools for the start. The appetite to learn more and expand your analytics capabilities will come with use.

We have seen several cases where a start with Qlik, PowerBI, or Tableau was the low entry into more relevant analytics tools.

Component 10: Data Monetization

Data monetization is about exchanging and monetizing the value of your data. It is the enablement of the GAIA-X idea of data sovereignty. You can use blockchain (private for closed user groups or public for an open ecosystem) or other tools.

In Europe, we would recommend following the GAIA-X principles and approach.

In the following chapter, the transformation journey to data-centricity will be discussed in detail.

IN THIS CHAPTER

Building a solid foundation for a successful transformation

Key success factors in building Data Centric Architecture

Finding the right approach for transformation and incremental change

Chapter 4
Your Transformation Journey to Data-Centricity

You have seen the need for Data Centric Architecture, you looked into the key principles you should apply if you do not want to end in a deadlock, and you got to know the key building blocks of Data Centric Architecture.

But how do you get there? Let us now look into the right approach and the potential ways to succeed.

Like in all transformations, there is a need for a clear commitment to the organization's digital transformation. It is not the IT who decides. The data-centric approach has to be supported by the business: »Do we want to participate in the digital age: yes or no?«

As the fundament of everything, data-centricity needs a clear business strategy, which relies on data and information, and pushes into the direction of digitalization. Data has to become the shared value of the organization.

When you begin with the transformation, it is essential that you base it on specific business needs and true value creators to succeed in the competition. These business needs depend on the specific situation of the company. It can be

- the need to save costs because there are too many overlapping applications,
- the urgent need for replacement, or
- the wish to offer a new digital product.

In the end, the first step needs to be based on a compelling story. The success will then lead to more and fast-following demands.

But the basis of everything is: The organization must have a clear strategy to go for data-centricity!

Five Key Pillars to Build Data Centric Architecture

There are five key pillars, on which Data Centric Architecture should be built. They are shown in Figure 4-1 below and discussed in detail in the following.

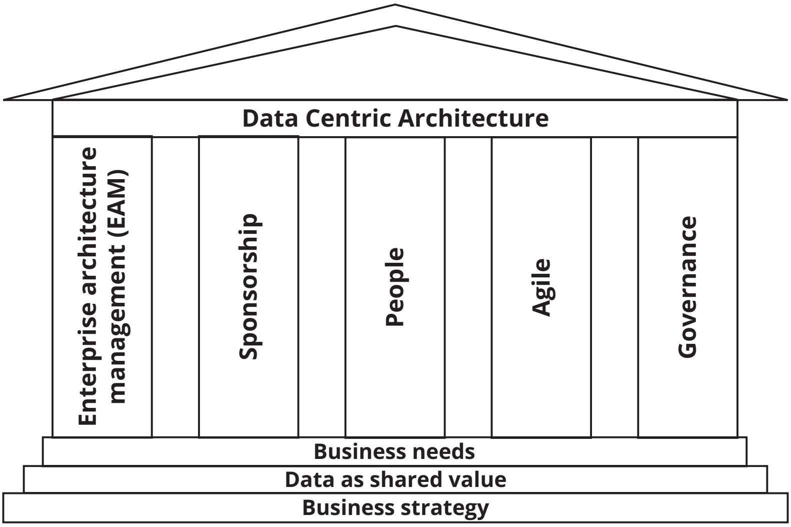

Figure 4-1: Framework to approach the implementation of DCA

Enterprise Architecture Management

The first pillar is a lean-agile thinking enterprise architecture management, empowered by business, as it has to guide the incremental steps to a truly Data Centric Architecture and business.

Agile EAM means

- ✔ **Concentrating on design principles** and best practices rather than books of standards.
- ✔ **Establishing an architecture that needs to be easy and understandable** so that architecture management can be democratized within the organization. Everyone who applies the principles should understand them.
- ✔ **Getting rid of unnecessary principles**. A regular reflection, whether a specific principle is still needed, must happen. Principles should only be present where there is real value in having them.

Any deviation from these architecture principles should be discussed, approved, and assigned with an end date or at least a review date. Some good practices go as far as saying that the origination project must provide the funding to get architecture compliance in place, for example when the deviation from the standard is driven by time constraints.

Sponsorship

The second pillar is top management sponsorship. Like in all transformations, the management has to support these processes not only in the beginning but also during the whole life cycle.

Therefore, incremental steps with continuous value generation have to be planned. Good practices show:

- ✔ Executing a two- to three-month cycle of incremental minimum viable products is best to ensure management sponsorship.
- ✔ Anchoring the architecture design principles in the project portfolio management and as a basis for any purchasing decision speeds up progress to a truly digital enterprise.
- ✔ Sending requests for proposal (RFP) only to vendors that have accepted the design principles and support the change process.
- ✔ Paying invoices only for deliveries with a signed architecture approval.

People

People are the third and central pillar holding up Data Centric Architecture. The most valid method to support the building of user and business-oriented Data Centric Architecture is the empowerment of people, ideally in combination with the introduction of scaled lean-agile methods.

As the market environments and needs are changing and the complexity of integrating the legacy is often not predictable, scaled lean-agile methods are very helpful, if not mission-critical.

Empowered, intelligent people

- ✔ will request clear and understandable guidelines and principles,
- ✔ will challenge the guidelines, and
- ✔ will apply and further develop the principles as soon as they understand their value.

Empowered intelligent people and intelligent lean agile enterprise architecture management are catalysts for one another – and will result in resilient and future-proof architectures.

Agility

The fourth value is again related to people and architecture: It is to be agile.

Agility empowers the employees, closes the gap between business and delivery. In the best case, an introduction of scaled-agile anchors DevOps in the organization for continuous value delivery and at the same time supports the agile portfolio management with a strong position of EAM in the decision process.

Governance

Finally, there is the column of governance, which completes the five pillars of transition to Data Centric Architecture.

Clear principles for architecture development must be followed and managed throughout the organization. Data-centricity must become the DNA of the organization in development and business.

The principle of »no excuse« is central to governance if you want to develop new solutions, services, interfaces, etc. Any data exchange must be centralized, documented, secured, and monitored. Cybersecurity plays a significant role in this regard.

These five pillars are an essential basis for Data Centric Architecture. And with these pillars, Data Centric Architecture is the right approach to respond to the market needs and enable true digitalization. Based on this architecture, real data analytics, industry 4.0, and digital service delivery are possible.

Finding the right Approach

It is important to build the transformation on specific business needs. This means that you have to find the pain point in your organization to start the journey and to establish Data Centric Architecture that enables your data-driven organization.

Several approaches are possible:

- **Start with intrinsic motivation:** In the best case, but still very seldom, the organization has already recognized that they want to be part of digital business and join the group of companies monetizing their data by sharing them in a wider ecosystem.
- **Start small, grow large:** If this is not yet the case, data can first be shared to a smaller extent. For example, the production sites want to use and learn from their machines' data. They want to get real-time insights and optimize the overall production. The data-centric approach can start from here and later scale in the organization.
- **Start at the top, then expand:** Another approach that we have often seen is that the management needs to access data more efficiently and learn about their organization's data sources and data. Here the Data Centric Architecture can use the momentum and connect the relevant sources via the streaming part of the Data Hub. As soon as the first data are in, more consumers will want to subscribe.

A partner of choice should be the data governance and enterprise architecture teams of your organization, as they will get real-time insights into the real world of applications and data.

- **Data governance as your partner in crime:** The approach could come as a first design principle to connect via a streaming technology or with a central approach for the next-generation integration layer – to break the silos. Data governance will be your partner to ensure that all application owners have to publish their logical data models into the data catalogue and provide metadata.
- **Focusing on compliance and data privacy:** GDPR compliance is a good use case for Data Centric Architecture as well. Central data streaming, the data catalogue, and even more the real-time adaptation to changing data feed helps to hold track of the GDPR relevant changes and facilitates the use cases of "information request" and "deletion request".

As an ISP or telecom operator, an excellent start is to monitor all your network devices, components and services used. The correlation of data from different data silos, such as facility management, cooling, fans, service usage, location, and active communication devices, will bring you soon new and relevant insights and enables a journey to predictive maintenance and autonomous operations.

As you can see with these few examples, there are multiple strategies to become a data-driven organization by installing Data Centric Architecture.

Incremental Change

It is obvious that you need an incremental approach. There is no big bang, and the world is rescued. You cannot describe all needs you want to cover in your organization. The world is changing continuously – and it would be far too costly if you first build the platform and then start to use it.

You need to make your own journey into the data-driven business. Therefore, start with small steps, but every step should deliver value.

Even the first proof of concept (PoC) should be built in a way that you can continue using it and bringing it into production. Environmental awareness means respect for any resources. So do not spend money on a PoC that you cannot reuse afterwards. Figure 4-2 describes a best-practice approach.

- The PoC should be your first minimum viable product (MVP), and it should contain the streaming part of the Hub and first connectors, which stream data into the Hub. Together with this, the transformation of the source data and the subscription by a microservice can be implemented. Low code apps are a possible good fast start here as well.

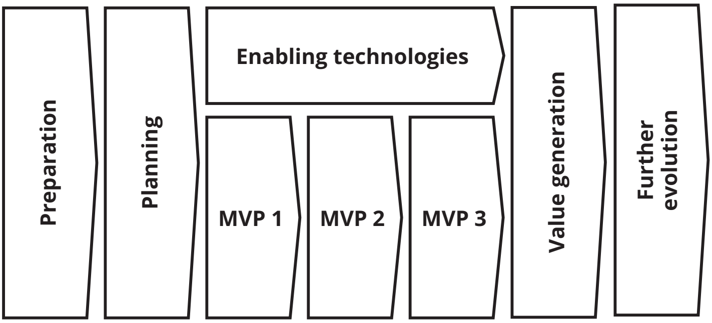

Figure 4-2: Incremental approach

- ✔ The next level would be to add more data sources and microservices. Additionally, it would be wise to add metadata management capabilities and get the first metadata and learnings.
- ✔ In the next step, learning from the metadata could be done using basic analytics tools. Moreover, access management should be established to show that the protection of the access can be managed based on the metadata.

Further steps depend very much on the specific starting point you have chosen. If you already have a good number of data in, the Data Hub could be enhanced with store and lake functionalities. Maybe you want to show the value of data in general by building a data cube for a specific purpose to show the value that you could deliver if you enhance these capabilities.

On the other hand, your organization's priority might be to enable the majority of workers, who might not have a digital workplace with user- and role-centric apps. In this case you would provide those workers with relevant information in a simple app, and commit all relevant task, shift plans and holidays via this tool.

Your organization urgently needs to connect a lot of IoT devices and learn the real-time information they are providing.

Real-time and ex-post analytics will be the steps you want to prioritize. Further analytics services and connections into MES or even ERP might bring value.

In some cases, your organization might want to exchange information between different applications or integrate data from partners – the Data Centric Architecture with stream, transformation, microservices, data catalogue, and access management will help to perform this task more easily than with numerous point-to-point-solutions.

IN THIS CHAPTER

Fake news about Data Centric Architecture exposed

Truths about Data Centric Architecture

Chapter 5
Ten Myths about Data Centric Architecture

Data Centric Architecture is the young side of tech companies. Like any new architecture, movement, or trend, it has to deal with reluctance and negation on one side and myths and hype on the other.

Here are some common myths about Data Centric Architecture that you may face, depending on the situation:

1. **Data Centric Architecture is only for IoT:** The reality is that streaming platforms like Kafka were born around pure streaming, but they can be the event bus for any information to be shared – even CSV files.

2. **Data Centric Architecture can be one method for the data exchange, but direct API calls should be done outside of it:** If you bypass DCA, you will not get into a position to build the digital twin. The interface savings will not be as rapid as planned, and the use of the persistence of the streamed data will not be as good as possible. Moreover, the source applications will not be overloaded with unnecessary calls.

3. **For Data Centric Architecture you do not need to have a data schema:** Using just the technical capability to stream and store data without a schema will not leverage the value of your data. Schema evolution and metadata management will be impossible.

4. **DCA needs no data access control:** On the contrary, metadata management and data life cycle management are preconditions for relevant data-centric management. If you do not know the classification, quality, and sourcing of your data, you will not be able to protect and manage the relevant and trustful exchange.

5. **As streaming technologies are high performing, you do not need to bother about service monitoring:** Not only do you need to monitor the streaming technology even if operated in a high-performance cluster, but even more, you need to monitor each interface with relevant KPIs as your Data Centric Architecture will be the nervous system of your IT. The functionality of each interface might work technology-wise, but a mismatch in processing or the missing delivery from source systems need to be tracked.

6. **DCA Kafka or MapR streams are good for video distribution:** Event streaming is suitable for nearly everything – but not for things like video streaming. There are specific technologies available. Suppose you are doing video streaming as a core concept. In that case, you might integrate the DCA with the video streaming and analysis technologies to exchange events in real time with many data subscribers.

7. **Data Centric Architecture is only for on-prem:** Data Centric Architecture works in any environment – on edge, core, and in the cloud. If you want to stretch with a single namespace from edge to core to cloud, you should look for technologies other than Kafka, such as HPE Data Fabric.

8. **Data Centric Architecture makes application databases redundant:** DCA is not intended to replace the application database. It helps build a digital twin of all data exchange in the company or eco-system and integrates existing applications.

9. **Data Centric Architecture makes application interfaces obsolete with user-centric UIs:** The user and role-centric UIs enable fast access to data and agile development of new services on data, including insights for managers, employees, and partners. It does not replace a targeted ERP user interface or a management console of a network management center.

10. **Data Centric Architecture is only for greenfield:** Data Centric Architecture can work on a greenfield site, but the strength is even more apparent when building on a brownfield site with many legacy applications to cope with for the next few years.

IN THIS CHAPTER

Services HPE offers you to introduce Data Centric Architecture in your company

Chapter 6
Ten Ways HPE Can Help You Get Started

You now have a pretty good understanding of Data Centric Architecture so go ahead and build your own! But of course, you do not have to do it on your own since Hewlett Packard Enterprise (HPE) is here to help you.

HPE has broad experience in realizing a Data Centric Architecture and is offering a wide set of services around introduction, realization, and operations.

1. The first step for you can be a **strategic workshop** using the HPE methodology of the Digital Next Advisory and Digital Journey Map. Together with the management, we develop your approach to digital transformation and action areas. In further workshops we can select specific options and create a more detailed action plan, or help to prioritize suitable options.

2. As a technical starting point HPE can offer a **Data Centric Architecture Readiness Assessment** and develop an **architecture vision** for your organization.

3. Suppose you have already started your digital journey and want to establish Data Centric Architecture as a core enabler. In that case HPE can help you develop a **target architecture** and a corresponding roadmap for your option jointly and can consult you on the advantages of different tools and approaches.

4. Starting with a PoC until ready for operations, HPE can be your partner to **realize Data Centric Architecture** with a global delivery team to create the solution design, deliver the configuration of the Data Hub and data governance components.

5. HPE can configure and **deliver connectors** of all types.

6. HPE can help and prepare for operational readiness with operation concepts and manuals for the hub and all relevant components of the Data Centric Architecture.

7. HPE can support and enable the realization of the **microservice layer** with low code components and enable **user-centric UIs**.

8. Moreover, HPE offers comprehensive and in-depth knowledge to establish the **analytics capability** and MLOps.

9. Data Centric Architecture can be delivered in projects or provided **as a service** in the cloud, **from edge to core to cloud**, regional or global in scale, single-tenant or multi-tenant capable.

10. HPE can design, deliver, and operate all underlying platforms and technologies on demand as a project or as a service.

You are now ready for your own data-centric transformation journey!

- ✔ Contact an HPE sales rep to learn how you can start on your journey today.
- ✔ Find additional resources:

  Lightboard session on YouTube

  Podcast episodes on Spotify, Apple Podcasts, Google Podcasts and Stitcher
- ✔ Visit the HPE website

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.

The Next Generation of Enterprise Architecture

HPE Data Centric Architecture (DCA) sets the stage for data-driven digital transformation and data sovereignty. The proven architectural approach decouples your data from the heterogeneous application environment, enabling you to put your data at the core of your business. It will empower your organization to reduce operating costs using real-time insights, while introducing new digital business models that unleash the value of your data.
This book will allow you to understand the fundamentals of DCA, how to apply them to transform your traditional enterprise architecture, and how HPE can help your organization to embark on this new architectural paradigm.

Inside, you will...

- understand the business value of DCA
- learn about its key architectural components
- take an in-depth look at essential design principles
- gain insights for your transformation journey to data-centricity
- find out how HPE can help you get started.

About Hewlett Packard Enterprise

HPE is a global, edge-to cloud company built to transform your business. How? By helping you connect, protect, analyze, and act on all your data and applications wherever they live, so you can turn insights into outcomes at real-time.

Cover image: © Hewlett Packard Enterprise

WILEY

for dummies®