DATA-DRIVEN SUSTAINABILITY

HOW TO:

DEVELOP A BLUEPRINT FOR A DATA-DRIVEN ENTERPRISE ARCHITECTURE

2ND EDITION

BY

LAILA FETTAH

RONALD MEIJER

JAN SCHRAVESANDE

THE BLUEPRINT FOR A RESILIENT DATA-DRIVEN ORGANISATION

IBM is committed to our clients and relations to be successful in the transformation of becoming a data-driven organisation. It is our ambition to be relevant as a knowledge partner for our clients and relations as well as to focus on our company's strategy: Hybrid Cloud and AI.
The strategy of IBM helps to accelerate the digital transformation of companies and be successful in building strategic platforms and eco-systems for better business outcomes.

This booklet is the result of focus, ambition and commitment from architects of our Dutch IBM Government team. Working together as a team on becoming a data-driven organisation gives an enormous amount of positive energy. Writing a book is not part of the daily activities of our architects nor having writing skills is a prerequisite to become an IBM architect.
Be aware! This is a 'read and work book', it is not an 'academic study and remember book'.
Hopefully you read this booklet with interest and pleasure for it is written with passion for the subject and the wish to help you in your journey in becoming a data-driven organisation.

Johan Heij
IBM Nederland

INTRODUCTION

In 2021 the Dutch Government CIO Office produced a strategy paper called “I-strategy Government”[1]. This strategy contained a very interesting topic on ‘informatiehuishouding', in our own words 'information housekeeping’. In English, we are talking about information governance and information management. Even though 'information housekeeping' does not officially exist, we use this term as it addresses exactly the message we want to convey. This topic triggered our curiosity and one year later it looks like it is almost the only thing that drives our activities, mindset, discussions and our idea’s. It certainly keeps us occupied!
We developed a ‘point of view’ (which is definitely not a vision), an opinion on this topic and we discussed this with many people within the Dutch Government, from policy and decision makers to information architects and technical girls and guys responsible for Enterprise Content Management and Information Housekeeping.
Nobody, well almost nobody, really likes housekeeping in the first place at all. Why should people like to do housekeeping in their information space just to have it nice and tidy? A well known English phrase is: "A place for everything and everything in its place"!

Well, it is our strong believe that good information housekeeping is the foundation and an absolute prerequisite for a data-driven organisation. This believe and insight has grown over the last period and our involvement in this dossier.
Our first scope was on unstructured data, the way it is used in government organisations, the way its life cycle is managed

[1] https://www.rijksoverheid.nl/documenten/beleidsnotas/2021/09/06/i-strategie-rijk-2021-2025

and how Government organisations can develop a better information housekeeping approach. In the mean time we realise the importance of having all data in scope (structured, semi-structured and unstructured) for information housekeeping and in becoming a data-driven organisation. Which, from all organisations in the world, Government is eminently a prime example!
In order to predict outcomes or prescribe activities, organisations must become information savvy and must develop new capabilities. Do not make the mistake believing that by upgrading your enterprise data warehouse by adopting AI technology will be the trick.
New technologies have emerged, new processes have emerged and new professions have emerged: they all help with the journey in becoming a data-driven organisation.

It is almost a buzzword today or even a long-lasting hype: “We want to become a data-driven organisation”. We would like to say: “of course you do”! It is almost the same as “I want a car with an engine”!
We're sure you don't want to be the Fred Flintstone from Bedrock.....

This booklet provides you with an overview of services and capabilities to help you build an enterprise architecture for a data-driven organisation. Besides this detailed overview it provides a set of best practices to help the IT organisation to avoid pitfalls.
Most published articles on data and AI are highlighting specific aspects of this domain. Aspects such as culture, complexity and best practices but mainly focussing on the information space and on machine learning, the cool stuff. But there is more to it than the information space alone.

This content will help you to create a holistic overview of the services, frameworks and approaches.
You can compare it with cooking: we describe both the recipe with ingredients as well as the approach on how to prepare the meal. The recipe contains the ingredients, these are the services we describe. We have described those quite detailed and we believe this makes it very useful as a reference.
The approach entails the preparation method; you cook, roast, fry or use other ways to prepare your meal. Those are the frameworks and approaches that we describe in the second part.
We do not describe the frameworks in detail, but we provide an overview and explain why those approaches are important to build your data-driven organisation. We refer to other articles if you need to know more about the approach or framework itself. By explaining those frameworks and approaches in the context of a data-driven architecture we trust that you will be able to select the ones that will fit your context and purpose and your organisation.

THIS IS OUR MISSION FOR YOU:
DEVELOP THIS BLUEPRINT, MAKE IT YOUR OWN, USE IT IN YOUR ENGAGEMENTS WITH YOUR STAKEHOLDERS.
MAKE THIS BLUEPRINT A LIVING ASSET FOR YOU AND YOUR ORGANISATION.

The idea is that you start with the first chapter, (that is also why we put it in as the first chapter), to give you an overview of the general idea of a data-driven organisation. You can consult the other chapters randomly as long as you don't forget to read these as well!
You need the complete story if you want to become a resilient data-driven organisation! We used the principle: write it down as

simple as possible. We realise that this may come at the cost of nuances or precision. However, we felt that understanding the big picture, the context and understanding the relationships between architectural domains will help organisations in their journey.
This booklet is intended for solution architects, enterprise architects, business- and IT consultants and strategy people. It is also valuable for CIO's, CTO's, CDO's and program managers.

With this booklet there is an associated large poster (A0) with a generic layout of the blueprint. We added a chapter that describes how you can use this poster to develop your own blueprint.

The final word in this introduction is about transparency. Business transparency (according to Forbes[2]) is the process of being open, honest, straightforward about various business operations. Transparent companies share information relating to performance, business revenue, internal processes, sourcing, pricing and business value. It is mandatory to inform all stakeholders with accurate information in order to be respected as a transparent and sustainable enterprise. You need to have accurate data in order to do this reporting.

PS Our writing style is Dutch - English. With other words: Dutch thoughts translated in English words. Hope you like it!

Enjoy!
Laila Fettah, Ronald Meijer, Jan Schravesande

[2] https://www.forbes.com/sites/mikekappel/2019/04/03/transparency-in-business-5-ways-to-build-trust/

THE SECOND EDITION

In this second edition we added some information about two topics based on the feedback we received.

First, we added some information about the usage of the booklet. We completed a number of workshops with different organisations and included the lessons learned in this second version of the booklet.

Second, we included more information on the topic of sustainability. Data and AI in particular are significant consumers of energy. We felt the responsibility to put more emphasis on this topic.

That feeling of responsibility went so far that we decided to write a second booklet about sustainability. That is also why we now call this Orange series part 1, as we are working on a part 2, albeit with a broader sustainability focus.

PS Even though we went through this book a million times to track spelling errors there are some people who have extraordinary skills in finding hidden errors, those have been corrected as well.

Any error that is being found in this second version is probably intentional.

Table of Contents

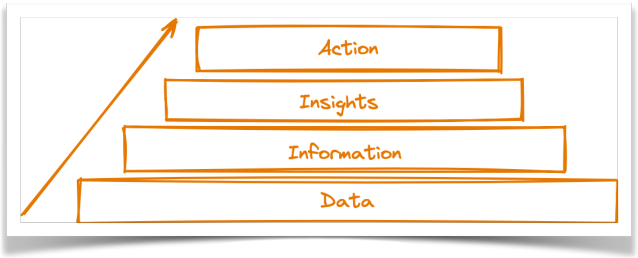
A DATA-DRIVEN ORGANISATION

Let's first talk definitions but let's not go into an academic dissertation. Why do organisations want to become data-driven?

To truly be a data-driven organisations means to live and breathe data and use the power of information in a predictive and prescriptive way. What is going to happen and what action can we take? Whether that means providing better citizen support when they reach out with questions on subsidies, to identify process inefficiencies and resolve those or ... to win Formula 1 races.

It all comes down to transforming the raw material into useable bits and pieces for insights for your business. This means it is vital to align your data strategy with your business strategy.

But to get there, we must step up in the pyramid and move up from data to information by adding context to the data. From information to insights by adding logic and intelligence and from insights we can generate possible actions.



Value of Data

As an example: how about an organisation that lives and breathes being data-driven and literally takes it to an extreme (speed that is): Formula 1 racing! During each race, 120 sensors on each car

generate 3 GB data, and 1,500 data points are generated each second[3]. They literally drive on data.
In the weekend of May 29, 2022, the importance of data and insight became paramount in the Grand Prix of Monaco. With Ferrari as the favourites in pole position, they should have made it to position 1 and 2 and Red Bull to 3 and 4. Ferrari should have collected 43 points and Red Bull 27.
With each point worth more than a million, Ferrari should have collected 16 million more than Red Bull. During the race, the weather changed and different tires were required.....

Based on the collected data, Red Bull calculated that the distance between the cars was just enough to have the tires of their two cars changed during one round.
However, Ferrari did not calculate this correctly and this resulted in confusion and frustration which led to a delay at the exchange. When they returned to the track, they were passed by both the Red Bulls. The end result was 40 points for Red Bull and 30 for Ferrari. So, a loss of 13 million and even worse, 13 points to the direct competitor as well.
The credits went to the data scientist who executed the scenarios, she literally gained millions of dollars that day.

A RESILIENT ORGANISATION

Why did we add the word resilient? As a reader you may think that, by definition, a data-driven organisation is resilient.
But we believe that is not the case! The risk of building a data-driven organisation based on todays and yesterday's situation can lead to wrong business decisions. If you are not able to easily adapt to these changes of business models and external events, your prediction of business outcome may fail. In the Formula 1, changes to cars could mean that as a data scientist you have to change your model swiftly in order to predict the right outcome.

[3] https://aws.amazon.com/solutions/case-studies/formula-one/

Therefore, you must be:

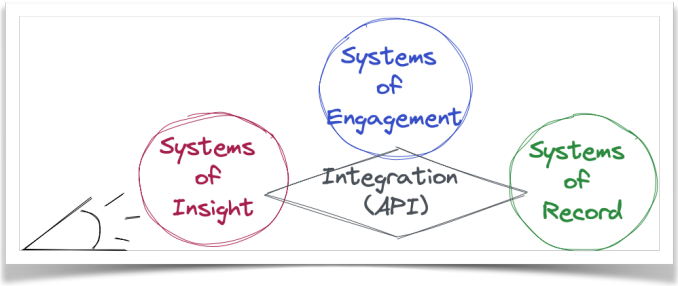
- Agile; new rules and algorithms must be developed and deployed in the existing or new business processes in short time frames.
- Secure and Compliant.
- Highly Available; systems must be available to access the data.
- Reliable; insight must be reliable and explainable. You do not want to unnecessarily impact healthy flows by false positives.
- Timely; insight must be available to you at the right moment in time. Not everything has to be actual, but you should have insights in the actual situation when necessary.
- Scalable; you must be able to scale up in cases of peaks and scale down in times of quietness (think about sustainability).

'THE SORTS OF SYSTEMS'

When looking at the landscape of systems we have to make a distinction between our core systems, our systems made for interaction like apps, web and portal and systems designed for intelligence and reporting.
Systems of Record are typically the core administration systems. For banks that would be the core banking system, for logistics it would be the product and ordering systems.
Often referred to by innovation enthusiasts as legacy. However it is still the core of the business operation!
The Systems of Engagement are those systems that interact with end users and employees. Systems of Insight are typical data warehouse and reporting systems. We connect these different systems through integration technology, preferably through API's.

This is a model created by IBM research more than a decade ago but it is very helpful to separate different types of systems, because they all have their own heartbeat, their own speed as

Gartner calls it: 'Bimodal IT[4]'. When we are talking about a data-driven organisation, our focus is on the Systems of Insight. But how does the architecture for such a system look like?



Different types of systems

In this booklet you will also find some critical remarks about the impact that the Systems of Insight may have on your IT strategy. It looks like the (technical) IT strategy is mostly driven by the Systems of Engagement, but this strategy does not always fit the strategy for the Systems of Insight. Systems of Engagement focus on interaction with users and have different requirements. Think about choices for hybrid cloud adoption and a "run anywhere" approach. Your data has different requirements!
It could mean that you must reconsider your strategy and do not follow all trends blindly. Our objective is to have you reflect on these topics before you end up following the hypes of the day.

The following is an attempt to highlight the differences in focus on the requirement for the “sorts of systems”. This is not black and white, on the contrary, you will find that the boundaries between the systems blur. Many of the terms used here are derived from the ISO/IEC 25010 standard for software product quality, but it is

[4] https://www.gartner.com/en/information-technology/glossary/bimodal

not about the exact terms. It should give you some more feeling about the focus on requirements[5].

	Systems of Record	***Systems of Engagement***	***Systems of Insight***
Positioning	Long lived systems. Systems of the truth.	Interaction and Interface with the SoR	Downstream systems, data derived from SoR and also SoE
Users	Expert users	Customers (and some other users)	Employees, business customers
Concepts	Business logic, rules,	Caching and Queuing, GDPR	Storage, ((No) SQL) Databases, Data Models, Data Product
Compliance	Legislation	Privacy, GDPR	Ethical and the ones mentioned before
Functional Suitability	Functional Correctness and Completeness	Functional Appropriate-ness (MVP)	Functional Appropriate-ness
Technology	Commercial of the Shelf (COTS): Customisations and Compiled code. Custom interface.	Portals: Multi-channel, multi-device. Scripting. API's, Open Source. Internet.	Dashboards: Data analysis, BI and Reporting. ETL and file transfer. Big Data up to Cognitive analytics

[5] https://iso25000.com/index.php/en/iso-25000-standards/iso-25010.

Performance Efficiency	Resource Utilisation, Latency and Throughput, Sustainability	Response time	Capacity
Security	Accountability	Authenticity and Non-repudiation	Confidentiality and Integrity
Maintain-ability	Reusability, Modularity, Analysability	Modifiability, Testability	Lineage
Usability	Learnability, Operability	User Error Protection, User Interface Aesthetics	Accessibility of data
Focus	Stability	Portability to support the user's device and agility to align with changing user demand	Data Volumes, data structures and data formats, data governance
Characteristics	Transactional sensitivity, commits and rollbacks.	From providing information to single user interaction.	Governance
High Availability	Data synchronisation for Stateful applications and transactions.	Application availability for Stateless applications	Asynchronous copies to rebuild the environment and
Disaster Recovery	RPO = 0 or near 0 (two phase commit), RTO = minutes to hours	RTO = 0, Master-Slave, concept of immutability.	High Availability is the solution for Disaster Recovery.
Reliability	Maturity, Recoverability	Fault Tolerance	Data Quality
Critical Resource	I/O and memory Bound	CPU and memory Bound	Storage and GPU bound
Scalability	Vertical	Horizontal	Vertical and Horizontal

Framework	Iterative, Ops	DevOps, SRE	ModelOps, SRE
Trends	Modernization. Move to packaged and standard products	Access to SoR via API's. Low Code.	AI, ML, data-driven.
Cloud Affinity	IaaS / SaaS (Private or Public)	PaaS (Public)	Data Fabric (Private)

As you can see there are differences in the requirements for SoR, SoI and SoE (sorts of systems). Do not treat them equally!
You have to realise that 80% of the software cost are related to quality aspects. To make your solution affordable, you have to balance between the quality aspects themselves and decide where to invest your money. As you can only spend it once!
And that one exception that you have in mind? That one that is very different? That exception proves the rule!
One of these exceptions we will discuss later: real-time analytics, a 'contradictio in terminis'.

WHY IS THIS BLUEPRINT RELEVANT FOR ME?

This is probably the first question that people ask us when introducing our booklet. Based on the discussions, feedback and workshops we believe there are several reasons why this booklet is useful for you:

Blueprint Relevancy

1. It is a holistic, high-level introduction to the topic of “data-driven”. It provides insights in relationships between the different services domains that we have identified. As a CTO/Program Manager it can help you asking the right questions and therefor challenging project managers and architects. It can support identifying interdependencies. It can also introduce you to topics which you have never considered so far. For some it will be something that provides details, for others it may bring it to a higher level, helicopter view, to become aware of the complex relationships between components that are required for a data-driven architecture.

2. You will find some statements that are a little bit “tongue in cheek”. These statements are related to topics that nobody would like to talk about because it makes the whole thing “more complicated”. Way of working, compliance, governance, demarcations of services, data security, just to name a few. We also make some statements questioning trends, we see that these are sometimes blindly followed. Not in your case of course! Our ambition is to bring these discussions above the table as we say in Dutch.

3. If you do not have a reference architecture (yet), you can use this as a starting point.

4. If you do have a reference architecture, you can use this as a validation. For each of the missing services you could ask yourself the question why you don’t need it. For those we did not identify, you can ask yourself the question wether you defined your services with the right level of granularity. You could use the idea of the booklet and A0 poster to describe your architecture and use that as a communication vehicle and we can definitely confirm it work and it resonates!

5. We have been asked if this architecture blueprint could be the magic accelerator for the development of a data-driven enterprise?
The easy answer would be yes of course! But to be honest that answer would be too simple. You need to comprehend the content of this booklet and bring it in context to the situation of your own organisation. And that is tough, because must decide how to use this content to make your (architectural) decisions. Perhaps even on topics you have not considered before! That means additional work but is worthwhile doing it for it creates the holistic view on data-driven where every organisation is craving for.

One thing that have been very clear to us, a big bang approach will not work, it is like boiling the ocean. That is why we have introduced the topics around use case driven, MVP's and agility. We believe that a business-driven use case/user scenario approach is the way to go forward, but you should avoid one-off solutions and that is where the blueprint will show its value. Two examples. The first is a text search function. Department A and B wants to have a set of services for searching data. Department A is a research department and would like to use it for their knowledge catalog and B would like to have it as an intelligent search for end users. From an architecture perspective you must bring this together because you like to reuse. Who will be the owner of such a service? Will it be one product? Are there any future requirements on the horizon to consider?

6. We also discovered that the blueprint could be very useful to define ownership. Which department is responsible for what services? From our perspective we have grouped them in logical functional domains. But that is probably not your reality, because when we developed this blueprint we had the advantage to start with a Greenfield situation. Anyway it can help the ownership discussion to make responsibilities clear. These responsibilities could go beyond your company boundaries, for example to discuss private and public cloud, managed services, outsourcing, etc.

7. Finally, we point you at some cultural aspects which are required to build a data-driven organisation. Most likely you will find different types of culture across the different domains. We would like you to be aware of these differences. Some differences need to be bridged, other differences you might want to embrace!

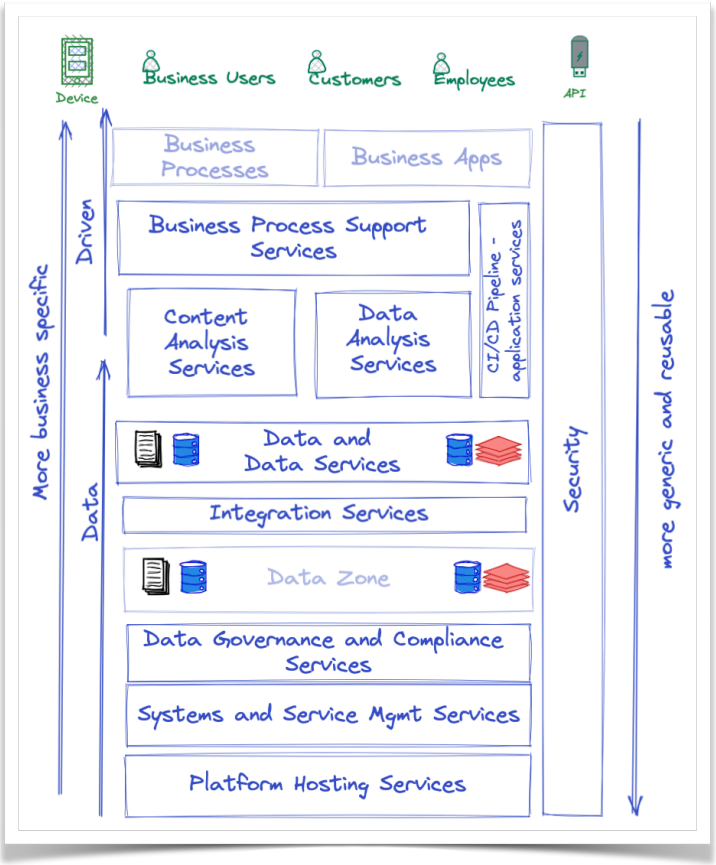
Besides describing how you can use it; it is probably also good to mention what you should not expect. It is not a reference architecture. A reference architecture should describe a set of

high-level requirements, it must contain architectural principles, it should detail the services to an individual service level.
It is also not a complete description about the frameworks that we point to. The frameworks that we refer to are described in more details somewhere else. What is important is that you are aware of the existence of the frameworks we refer to. They are extremely important to collaborate successfully and also address cultural aspects. DevOps, ModelsOps, AIOps and garage method are the frameworks that we refer to. Please investigate these, be open minded to these frameworks and where relevant apply those.

Finally, it is also not a solution architecture for a data platform. It is product agnostic, and the idea of a data platform is mentioned, however not always necessary to build a data-driven organisation.

ARCHITECTURE OVERVIEW

In our architecture overview we have defined a set of common services that are very useful for a data-driven organisation. We have had discussions whether we should build it from the bottom to the top (which seems the more logical approach from the standpoint of an architect) or from the top to the bottom? We chose the second approach (top to bottom) to lead with the business in mind but there is no good or bad choice in our honest opinion.



Architecture overview

1. ***Business Processes***. The Business Processes are the primary (and secondary) functions of an organisation. Each organisation has these functions to execute the business operation. Business processes vary by industry. It could be a summary report of sales transactions that were performed that month. It could be a list of tasks to be performed that day. Because these are specific business related services we do not discuss them in detail.

2. ***Business Apps.*** The Business Apps are the core systems of the organisation and support the primary business functions. These systems are often referred to as the Systems of Record and contain the operational business data. This is the source for the Systems of Insight.

3. ***Business Support Services.*** The Business Support Services are the tools that provide users the capabilities to perform their activities. Typical services for a data-driven organisation are services like rules management, case management and business process management. Rules use data or data models to provide input to a business process or a case in order to make informed decisions.

4. ***Content Analysis Services.*** We had long discussions whether we should combine Content and Data Analysis Services in one group: as from a service perspective overlap exists between those two. For example the anonymisation services. However we also identified services that apply only to the structured or the unstructured domain. For example a document summary service or a Question & Answering service is only applicable to unstructured data. A second reason is that the concepts (and therefor also the technology) for the implementation of these services are very different from each other.

5. ***Data Analysis Services***. What Content Analysis Services are for unstructured data are Data Analysis Services for structured data. Both the Content Analysis Services as well as the Data Analysis Services are considered the facade of a building for a data-driven organisation. It provides insight in data! From factual historical data to insights in what is happening now, as well as predictive outcomes and even prescriptive suggestions on how to act based on certain information. The reason for calling this the facade of the building is that Data and Content Analysis Services cannot operate well if the foundation of the building is not working properly. In practice, we see that the focus is on the facade while the complexity and the work is in the foundation: collect, understand, curate and classify the data. A typical role that we have assigned to the curator. The roles that make use of the data analysis services would typically be analysts and data scientists.

6. ***Data and Data Services***. What can you do without data? Nothing! The Data and Data Services are a collection of data, more data and much and much more data in any kind of format. Not all data have to be physically available on premise (see Data Zone), it could also be virtually retrieved from the cloud.

7. ***Integration Services.*** Access to data is provided by Integration Services. Quite a common approach is to use API's to integrate for example between a process and an information service. The scope of Integration Services is however much broader, it is a collection of services that facilitate any type of integration, from large, we mean really, really large files until remote processing of data, from traditional copying to making it accessible via views.

8. ***Data Zone.*** Data Zones are the locations where the data resides. This can be 'On-premises' in your own data centre or

data residing in another location, e.g. another companies data centre or data stored in the Cloud.

9. ***Data Governance & Compliance Services.*** Data governance helps you understand what data you have, where that data resides and how it can be used, ideally via self-service mechanisms. These services provide solutions and tools for data understanding, data usage and data quality. It also provides a fundament that supports collaboration between business and IT, and allows for information integration.

10. ***Systems and Service Management Services***. The hybrid cloud requires services for managing the operational aspects as well as the services aspects. Even though the focus is on managing containers, which is performed by Kubernetes, the management should also consider the existing enterprise platform, the Systems of Record.
    - Application Runtime. Whether we use existing applications or build our own, we see the future for data-driven organisations based on an agile approach using microservices running in containers.

11. ***Platform Hosting Services***. The data-driven organisation runs on cloud concepts. Most likely this will include a private cloud with capabilities to extend this to the public cloud. It should be hardware agnostic with sustainable technology with a focus on high utilisation and sharing expensive GPU servers for executing AI models.

12. ***CI/CD Pipeline***. The CI/CD Pipeline is the vehicle to produce (Micro)services. We need tools to store our code and configurations to develop our code, to build the environment, to test it, and ultimately after a successful test we have to deploy it into production. All these services are part of the CI/CD Pipeline.

13. ***Data Security Services.*** We hope we still have your attention. We felt a little bit awkward to put Security as the last Service. Now that you have an overview of the data-driven architecture you must feel the security's relevance.
    Some of the data you use is public data some of it is sensitive data. What kind of data is it and who can access it? Definitely worth a separate chapter! Please be aware that the focus here is on **DATA**. We do not address infra security like firewalls.

FRAMEWORKS

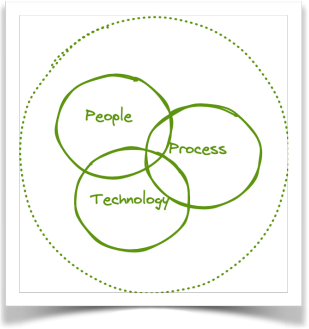
With all these services in place, we have the technology at hand to run a data-driven organisation. Technology is an important piece to solve the puzzle but without the right skilled people and an adopted culture, we will not be able to transform data into usable insights. The IT industry developed frameworks which addresses each aspect of the three dimensions People, Process and Technology. The following frameworks are useful when developing a data-driven organisation:

- ***Garage.*** The method to understand the need of the different stakeholders and work with users towards a minimum viable product.
- ***DevOps.*** A DevOps process approach is used to build applications and realise business processes identified by the Garage.
- ***ModelOps.*** The business processes will need insight into data. This could be based on AI models. This is a process to develop and train data models and deploy and monitor these in a production environment.
- ***AIOps.*** The whole operational environment must be holistically managed. This includes the container environments and also existing platforms.

Further on we will provide some more information about these frameworks.

PEOPLE

One of the complexities of a data-driven organisation is the broadness and deepness of experience and skills that are required.
On one hand we have the world of unstructured data, this typically requires the expertise of a librarian (nowadays Curator), considered by us as one of the oldest professions in the field of Information Housekeeping in the world. Think of the librarians of Alexandria who catalogued and classified all the papyrus rolls of this famous library! Organising unstructured data by classifying this data using metadata and applying ontologies and taxonomies is complex. The right choices are important to make it easier to classify, but also easier to find the data (the papyrus rolls).



People, process, technology

The approach the librarians from Alexandria used by separating the content from the index is from all times. Our current library system is still based on this principle (but also our current IT systems!). When you rent a book, movie, or music you search in the index and that refers to the position in the library.
The library is categorised in categories and media are organised accordingly. Changing categories or moving a library is an immense amount of work. Each individual item has value.

On the other hand of this spectrum, we have the data scientist. The value of structured data in the context of the data scientist is in the amount of the data together, not the individual records.

Applying AI techniques and models and running these against large amounts of data is completely at the other end of the spectrum.
We therefore believe that it is important to make a distinction between structured and unstructured data. Different professions, different tools lead to different processes. In contradiction to what sometimes has been advertised, it is our strong personal belief that these techniques will not be merged into one solution. The characteristics are very different.

CAN AI INCREASE THE QUALITY OF BUSINESS PROCESSES?

Artificial Intelligence is a key technology in a resilient data-driven organisation. When becoming a data-driven organisation it is no longer feasible to process and interpret data manually, whether in structured or unstructured form. The controlled appliance of AI will provide opportunities to increase the quality of decision making as long as the human is kept in the loop, and decisions can be traced back.
This is not an easy feat; it requires a proper fundament that allows insights in how an algorithm behaves, what training data was used, which operational data was used etc.
In other words, it supports the human with the rationales behind the data-driven decisions, but still allowing the human to overrule these decisions.

HYBRID CLOUD

This blueprint for a reference architecture is product and infrastructure deployment agnostic. However, during the next steps you have to make a choice in selecting your products. It is our believe that a resilient data-driven organisation should be cloud independent. So, avoid lock-in by using specific cloud provided products.
Usually cloud vendor agnostic solution takes a little bit more effort compared to using a cloud vendor native product off the shelf.

But take our word for it, it brings much more advantages:

- You are not dependent on the cloud provider; it also works in your own private cloud.
- It forces you to use more generic solutions, leading to “basic” appliance of the product, avoiding risks of technical debt.
- The AI part of a data-driven organisation is very hungry for data. Cloud egress (retrieve data from the cloud) costs could be substantial. Ingress (bring data to the cloud) is commonly free of charge. You must place the data where it fits the workload and not where the cloud provider wants you to place it.

By the way you have to insist on an explicit exit plan/strategy not because you have to but it is what you want!
There are three things we like to share with you that could be very useful as a best practice in your system when operating in a hybrid cloud environment:

1. The first is Cloud Service Management and Operations (CSMO)[6]. This is an add-on to ITIL for supporting the ways problems can be solved in a cloud native environment.

2. Second is the activities that must be performed for a platform. Specifically so called day-2 activities. Refer to a GitHub repository for valuable information on this: CASE/OCP-Day2-operations[7].

3. And third, there is a free book: ‘The Cloud Adoption Playbook’[8]. It is a little bit older already, but still contains a comprehensive overview of topics related to cloud adoption. You could consider that a complementary book to this one because it handles all the cloud related topics that we only mention here

[6] https://www.ibm.com/cloud/architecture/content/course/csmo-advocate/

[7] https://github.ibm.com/CASE/OCP-Day2-operations

[8] https://www.ibm.com/cloud/architecture/adoption/the-cloud-adoption-playbook/

but not handle in detail.

TAKE DATA REPATRIATION AS A SERIOUS WARNING

Data repatriation is a trend in the market where organisations return from their public cloud journey and start to do things locally again. The public cloud journey became a disappointment and mainly the cost of the cloud is why many went back to on-prem activities. The reason they returned was mainly caused by the Systems of Insight. Let us explain:

In the early days of cloud one of the promises was 'cloud is cheaper'! That argument became, to say it mildly debatable. In 2008, during the financial crisis demand went down and companies realised they could not scale down their IT cost because these where completely fixed and caused significant over-capacity. So, the drive was to create a flexible IT organisation that could adapt according to variable demand (cost).
The problem for some data intensive companies became the cost of using data in the public cloud. Companies went back to on premises IT services but as they experienced the advantages from the cloud, they adopted cloud technology as basis for these services. Vendors embraced this new trend and developed private cloud solutions to meet the new demand. Nutanix is one of the companies who successfully embraced this opportunity.

One of the cost factors in the public cloud are the data transport cost and storage cost on primary storage. We are not saying you should not put your data in the public cloud but you should be very cautious about where you do your processing and where you store your data. In the blueprint we provide a Distributed Processing Service that can do the processing on the location where the data resides and only returns the result. That could be a solution to resolve this problem.

REAL-TIME ANALYTICS

Something else that we would like to highlight is the possibility of performing real-time analytics. Real-time analytics can provide significant benefits for your company. Your company can take action at the moment the transaction takes place. That is data-driven to the extreme!
This works in two steps. The first step is to create a model based on historical data. The second step is to use that model and “inject” it at the moment the transaction takes place. For example a model is developed that detects a fraudulent credit card transaction. In the past that model was executed against a list of past transactions and you detected the fraudulent transactions after they occurred. But now, you would like to perform the check at the moment the transaction takes place, and you have the possibility to block this transaction. This is a manifestation of an optimal integration between the Systems of Record and Systems of Insight.
And now a warning!! We have to write something about one of our own products. Please forgive our enthusiasm and if you think you are better off with another solution for this real-time transactional challenge it is also fine of course. It is about the 'IBM Z' and 'LinuxONE'. To be more precise it is about the 'Telum' processor in those systems. 'IBM Z' is also known as the mainframe. 'LinuxONE' is the same hardware but it is a Linux server, comparable when you put Linux on a X86 server but in this case you can put a farm of servers on one 'LinuxONE' box. The processor combines AI capabilities and CPU onto one processor.

Imagine that transactions are handled by the CPU while at the same time there is parallel AI process checking if the transaction is valid. Realising that 80% of all financial transactions in the world are handled by these kind of servers, one should seriously consider this platform for real time analytics. There is another CPU with a similar concept which is the Mac M1 processor, it contains

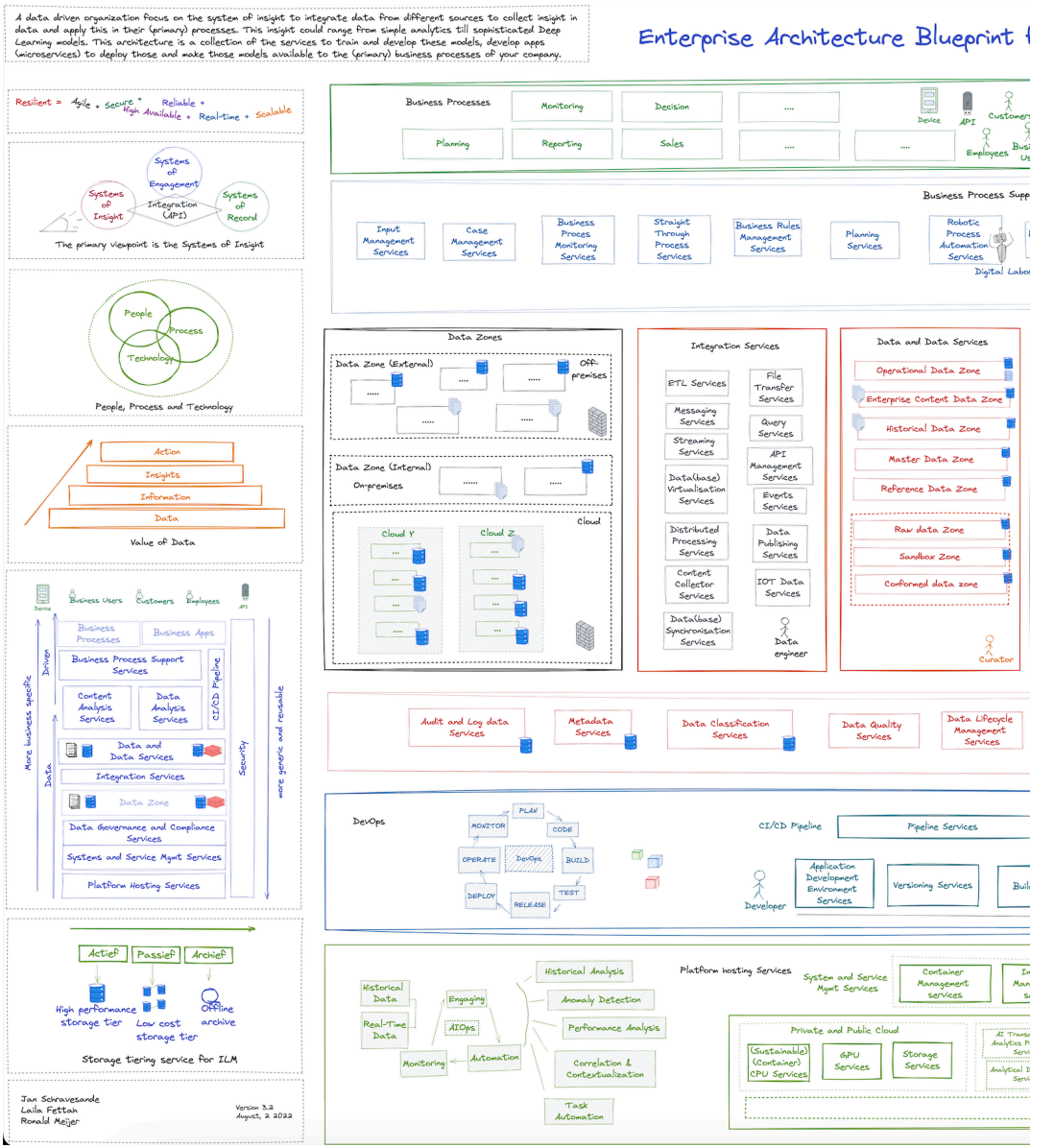
GPU and CPU on the same chip. But you don’t put your Mac as an enterprise server in a datacenter....
End of the commercial, back to the fundament of the idea!

ROADMAP AND PLANNING

When you are confronted with all these services, it could be quite overwhelming. Especially if your organisation is limited in size. But don't worry! The best approach is to define your most valuable use cases. Use the blueprint for the reference architecture to plot your uses cases. Using this approach, you will find your most important services. Develop those services first!

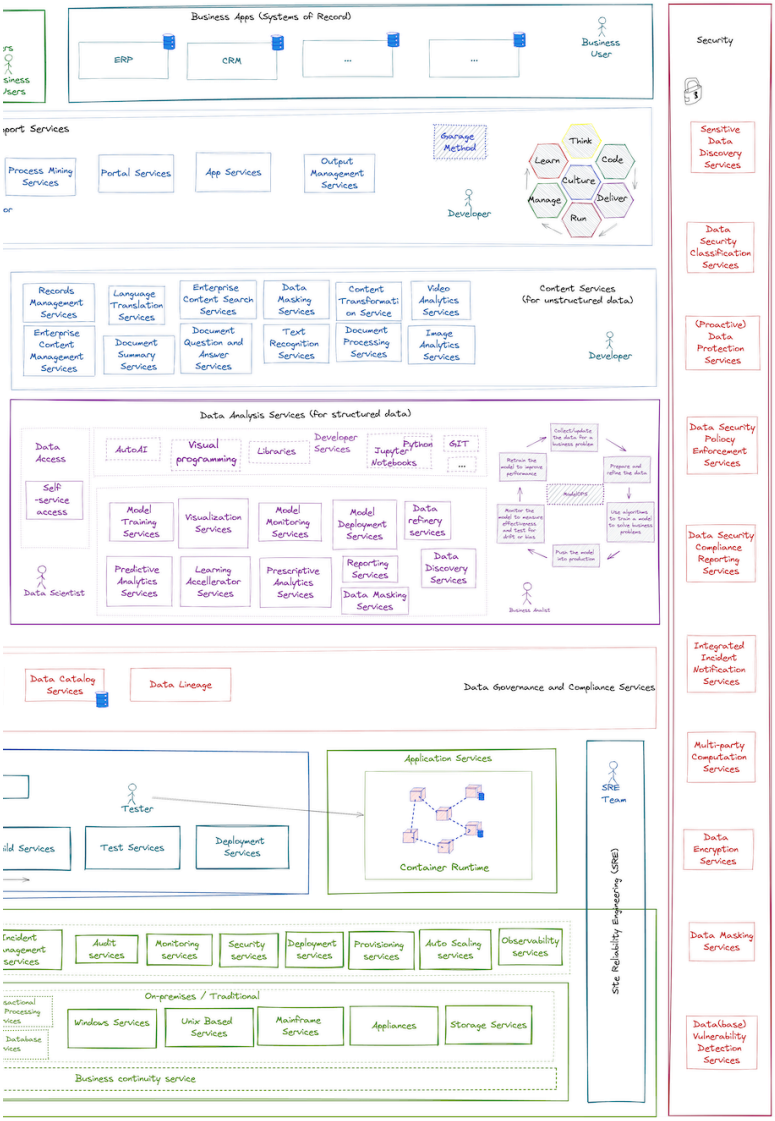
THE BLUEPRINT POSTER

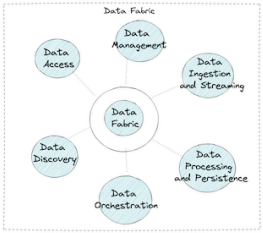
Forever ashamed about this white page, but we just could not make it work with the updates of the 2nd edition.

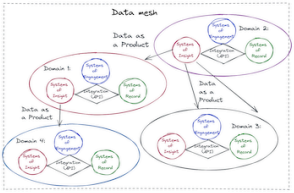

A data driven organization focus on the system of insight to integrate data from different sources to collect insight in data and apply this in their (primary) processes. This insight could range from simple analytics till sophisticated Deep Learning models. This architecture is a collection of the services to train and develop these models, develop apps (microservices) to deploy those and make those models available to the (primary) business processes of your company.
Enterprise Architecture Blueprint
Resilient = Agile + Secure + High Available + Reliable + Real-time + Scalable
Systems of Engagement
Systems of Insight
Integration (API)
Systems of Record
The primary viewpoint is the Systems of Insight
People
Process
Technology
People, Process and Technology
Action
Insights
Information
Data
Value of Data
Business Users
Customers
Employees
API
Business Processes
Business Apps
Business Process Support Services
Content Analysis Services
Data Analysis Services
CI/CD Pipeline
Data and Data Services
Integration Services
Data Zone
Data Governance and Compliance Services
Systems and Service Mgmt Services
Platform Hosting Services
Driven
Data
More business specific
Security
more generic and reusable
Actief
Passief
Archief
High performance storage tier
Low cost storage tier
Offline archive
Storage tiering service for ILM
Jan Schravesande
Laila Fettah
Ronald Meijer
Version 3.2
August, 2 2022
Business Processes
Monitoring
Decision
Planning
Reporting
Sales
Device
API
Employees
Business Process Supp
Input Management Services
Case Management Services
Business Proces Monitoring Services
Straight Through Process Services
Business Rules Management Services
Planning Services
Robotic Process Automation Services
Digital Labor
Data Zones
Data Zone (External)
Off-premises
Data Zone (Internal)
On-premises
Cloud
Cloud Y
Cloud Z
Integration Services
ETL Services
File Transfer Services
Messaging Services
Query Services
Streaming Services
API Management Services
Data(base) Virtualisation Services
Events Services
Distributed Processing Services
Data Publishing Services
Content Collector Services
IOT Data Services
Data(base) Synchronisation Services
Data engineer
Data and Data Services
Operational Data Zone
Enterprise Content Data Zone
Historical Data Zone
Master Data Zone
Reference Data Zone
Raw data Zone
Sandbox Zone
Conformed data zone
Curator
Audit and Log data Services
Metadata Services
Data Classification Services
Data Quality Services
Data Lifecycle Management Services
DevOps
PLAN
CODE
BUILD
TEST
RELEASE
DEPLOY
OPERATE
MONITOR
DevOps
CI/CD Pipeline
Pipeline Services
Developer
Application Development Environment Services
Versioning Services
Historical Data
Real-Time Data
Engaging
AIOps
Monitoring
Automation
Historical Analysis
Anomaly Detection
Performance Analysis
Correlation & Contextualization
Task Automation
Platform hosting Services
System and Service Mgmt Services
Container Management services
Private and Public Cloud
(Sustainable) (Container) CPU Services
GPU Services
Storage Services

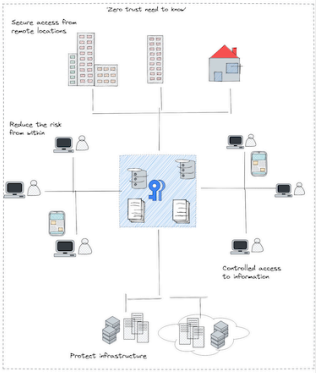
A0 Poster

for a Resilient Data-Driven Organisation









STRUCTURE OF THE BLUEPRINT

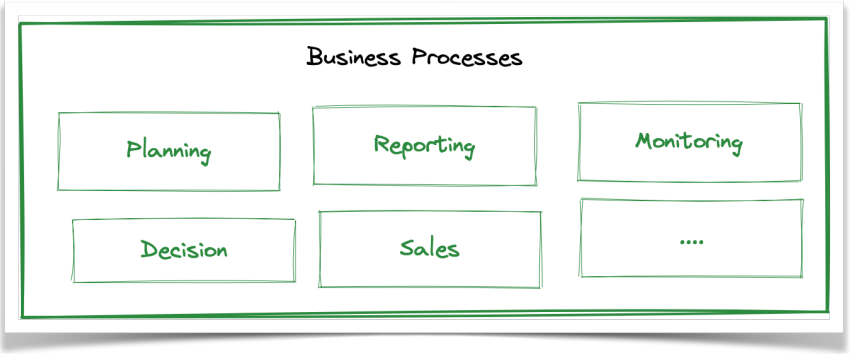
The blueprint is drawn on poster size (A0) format.
In order to create a good structure for the explanation of the different domains of the poster we go through the poster from top, middle layer and bottom layers (see architecture overview). We might as well have started bottom up, but we chose this top down approach in order to focus primarily on the use of data in the business processes! Technology aspects such as platform hosting services are definitely important as well but in this case not the focus starting point.
Each domain will be described as services or capabilities.

PERSONA'S IN THIS BLUEPRINT

 Business user	The business user is in the heart of the business operation. In a data-driven organisation, the business user relies on trustworthy insights.
 Developer	The developer is active in a number of domains. In business support services the developer develops solution in order to increase business productivity and the automation of business processes. The developer for integration services is more focussed on integration of services of internal as well as external sources.
 Data scientist	The data scientist is an expert role for data analysis. They have skills to solve complex problems and are able to predict business outcome and support business decision making.

 Curator	The data curator is responsible for maintaining and managing metadata. They catalog the data and make sure that data is used in the way it is supposed to be used. They bridge the gap between the world of Information technology and Data Science.
 Data engineer	Data engineers are technical specialists who are responsible for designing and maintaining the architecture of data systems. They are involved in the processes for data quality improvement. They are also responsible for the processes for modelling, mining, verification and acquisition.
 Tester	There are many different types of testers. In every context they test the product against functional, and non-functional requirements.
 Engineer	Engineers are responsible for the design, installation and maintenance of enterprise IT systems.
 SRE team	SRE teams are responsible for proactively building and implementing services that helps the IT organisation to become better at their job. Their skills vary from infrastructure skills, middleware skills as well as coding skills.

BUSINESS PROCESSES



Business Processes

Business processes are typically related to a specific business. Even though there are some general processes that are related to the support of the business, the core business processes are quite unique per industry type.

Primary business processes for example are:

Manufacturing:
- Procuring
- Producing

Retail:
- Assortment management
- Advertising / marketing

Travel & Transport:
- Planning
- Ticketing

Supporting business processes are known as COPAFIJTH or SCOPAFIJTH:

- Commerce
- Organisation
- Personnel
- Administration
- Finance
- Information
- Juridical aspects
- Technology
- Housing

We have to ask ourselves two important questions:

1. How can organisations feed their business processes with quality information to become more data-driven?
2. What kind of information is required to feed a certain process?

Work out your own data requirements by developing use cases and identify the (data)services for each of the use cases (see section 'Bring it forward').

Some ideas regarding your business processes for a data-driven organisation:

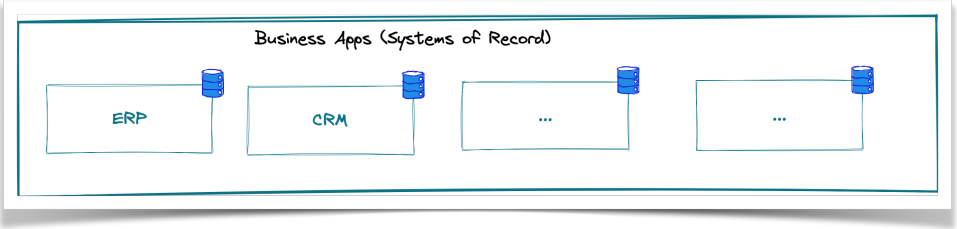
Decision Process	All organisations have to make decisions. For example, your organisation needs a new CTO. HR needs to decide who the best person for the job is. Let's not discuss the compliance part of the solution, but the process goes through all HR files and it comes up with the best person for the job. And guess what, yes it is you!

Planning Process	Planning is done in all kinds of organisations. Especially if you are in logistics. For example, you are a parcel deliverer and you want to know what the best route is for delivery. Another example, you are a police officer on duty and you would like to know the best surveillance route. Planning is everywhere. A good planning process can save time, money and even lives. Optimisation algorithms can optimise your planning. What is your most important planning process? What would you gain if that would be only 1% better? No, not only money!
Reporting Process	There will always be a need to generate reports for management that reflect the status of the organisation, for example monthly sales. To find a starting point for a use case you could ask yourself the question: "What are the most important reports in my organisation?". Can you enrich those with additional information that can help decision makers. For example you provide a monthly report on the sales of ice cream. Would it make sense to add information about the weather, flavours and the competition?
Sales Process	All organisations have a kind of a sales process. Whether you are a government organisation issuing identities or a grocery shop selling potatoes. You ‘sell/deliver’ something and there is probably a lot if insights that may be useful for you in your primary process like opportunities, sales, losses, prospects, future projections, etc.

Monitoring Process	The Monitoring Process uses the Business Process Monitoring Services to get insight in the status of internal and external processes. For example it can monitor stock and it alerts you if the amount of products drop below a threshold.

What are the three most important business processes in your organisation and how can you improve those by injecting data?

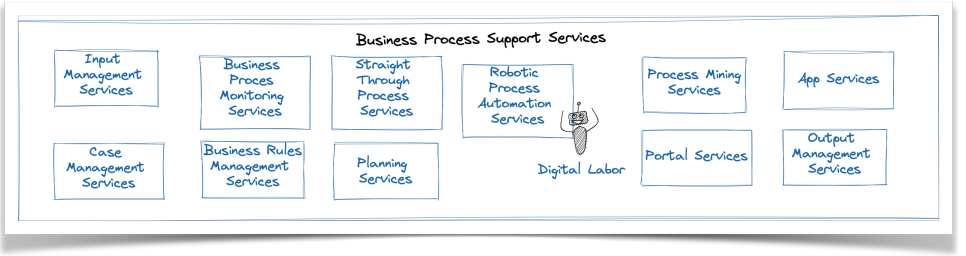
BUSINESS APPS (SOR)



Business Apps

The business apps are the core systems of the organisation and support the primary business functions. As mentioned before these systems are often the result of many years of development and subject to functional and technical changes. Therefore it can be a technical challenge to unlock the valuable data in these systems, especially when we want to use the (real-time) data in AI (Artificial Intelligence) processes. Many organisations use ERP software (SAP, Oracle, Microsoft) and hopefully they use it as COTS (commercial of the shelve), otherwise these organisations face the same challenge as the ones who developed their own systems. Upgrading to a new version of ERP is almost or even a bigger challenge when you build too much home grown functionality in a standard ERP.
We are not going into detail about business apps as they are organisation specific. In the A0 poster you can place your core systems in the business apps block.

BUSINESS PROCESS SUPPORT SERVICES



Business Process Support Services

Business Process Support Services are generic building block services available to create a business process service. Examples are rules management and process management. These services are not necessarily dedicated to the Systems of Insight we discussed before, but they are part of the data-driven organisation. A data-driven organisation is about 'data' and is about 'driven'. The focus for the Business Process Support Services is the “driven” part. It uses data and makes that available for the business processes.

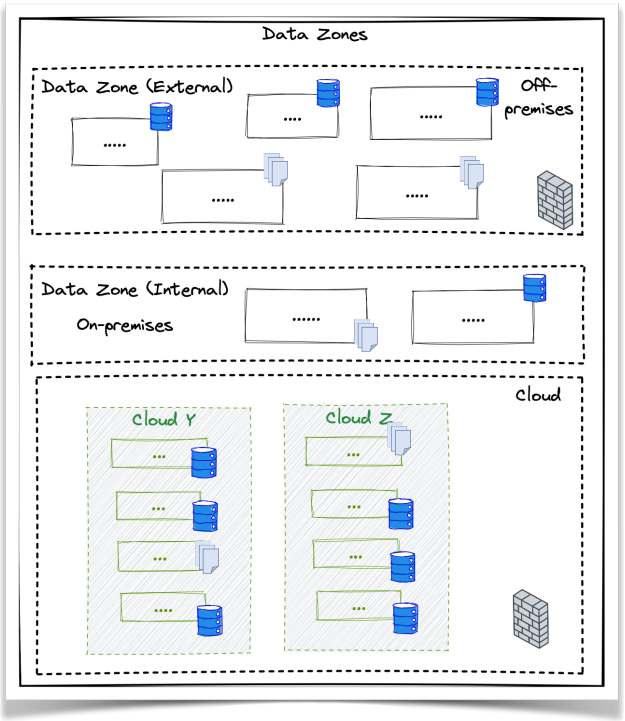
Case Management Services	Case Management is a specific implementation of process management. It is used when information is handled via an undefined sequence of tasks. The outcome of the current task determines what task to perform next. It collects information and insights along the way and it becomes a case. For example, if you committed a speeding offence a business rule is defined: when the offence is more than 50km, the offence will turn into a case, if it is less, it goes through Straight Through Processing that sends you a fine and maybe a link to a picture where and when the offence took place. Expensive pictures! PS A case not touched for years is called a “cold case”.
Straight Through Processing Services	Straight Through Processing Service or STP is the other implementation of business process management. STP handles a piece of information via a predefined sequence of tasks. The business process steps and sequences are predetermined. Companies try to have most of their processes handled via STP because this requires no or very limited user interaction. STP integrates with Case Management. As soon as the process can no longer be processed via the standard process, it will be handed over to the Case Management Services. And also the other way around. If a case becomes a standard process it flows into STP.

Business Rules Management Services	Business Rules Management Services are services that hold all kinds of business logic and rules. It is made available centrally and can be accessed by business processes using a simple interface, preferably an API. The major advantages are: 1. single source of truth and 2. the rules can be changed relatively easy, because they are not hard coded.
Process Mining Services	Process Mining Services measure and report about the flow of process activities and give advise to improve the flow by proposing a more optimal process.
Business Process Monitoring Services	While the mining service is looking at the *flow* of the processes, the Business Process Monitoring Services are focused on the actual *status* of the process. For example overdue requests, the length of a queue, average response times, etc.
Application Services (Apps)	Application Services are providing business services that can be accessed by any of the business support services. Apps are preferably deployed as microservices in a container runtime.
Robotic Process Automation Services	Robotic Process Automation Services (aka digital labour) automate a sequence of repeatable human interactions with a computer. It is a solution to gain productivity benefits quickly and automate *(boring?!)* repeatable tasks. In general RPA is more a short term solution because it is sensitive to modifications in for example a user interface. Therefore the preference is to move into a business process based solution with reusable services.

Input Management Services	Organisations have different input channels (phone, Internet, Mail) and different media (mp3, mp4, paper, xml). These channels and media provide different formats. Input management is the service to unify the incoming information in a digital format so that it can be stored in an IT system. That data can then be fed into the business processes. Input Management creates as much as possible structure around unstructured information and it usually stores the original input as the starting point for an audit trail. Input management uses many of the Content Services. For example Text Recognition Service or Auto Classification Service. A good Input Management Service will provide benefits throughout all your processes.
Planning Services	Planning Optimisation and Analytics Services are used for decision making. It is almost a world on its own. To be completely honest, it is a bit too complex for some of us. (We will not disclose for whom, because we feel that this may impact the rest of our career and we like what we are doing way too much (-:). In the past we had planning optimisation and analytics as two separate topics but more and more they tend to integrate. Planning analytics provides insight in actual status and forecast. Planning Optimisation is more focused on the prescriptive part, to improve your planning. It is also a domain where a lot of research still takes place. And here it comes, you haven't read it yet in our entire booklet: it could be the domain where Quantum computing could play a role.

Portal Services	Portal Services are primarily belonging to the Systems of Engagement. However we see dashboards and forms as a kind of portal. Dashboards are a way of presenting data and are therefore more related to the Systems of Insight. It uses the views in the virtualisation service (see Data Analysis Services). Forms are used by the Input Management Services to collect information in a more structured way.
Output Management Services	Output management takes the result of a business process and translates that into a consumable output format. In correspondence this could be a PDF, a message or an email. The output that is shared with the user is stored in a repository as part of audit information.

DATA ZONES



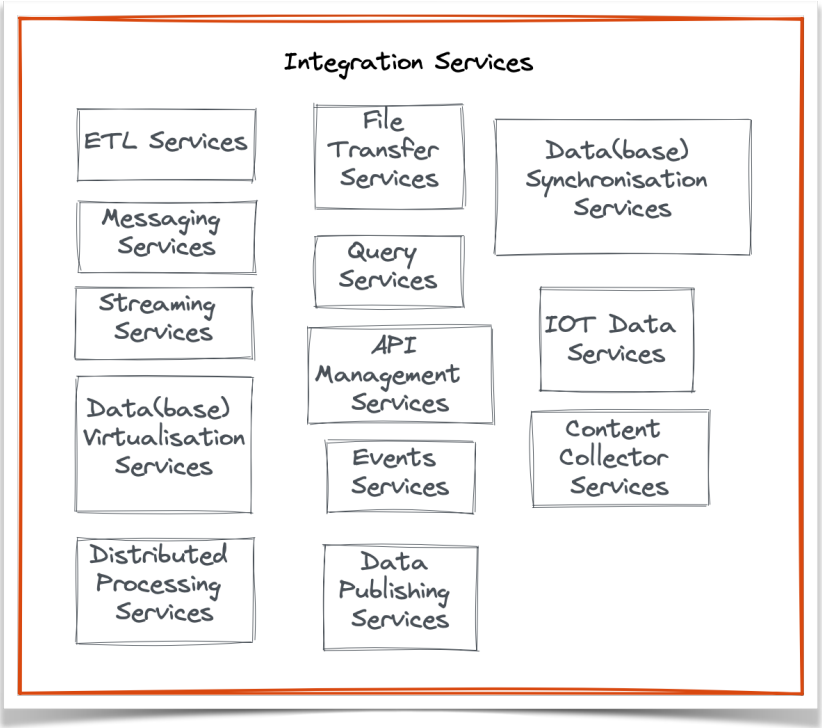
Data Zones

Data zones are the locations where the data resides. This can be ‘On-premises’ in your own data centre or in another location, e.g. a data centre from another company or in the Cloud.

Obviously data ’on premises’ is secured but external data is also secured. Who may access which data in which context is always a very important consideration in your architecture.

Data Zone (External)	With external data zones we define the data that is not part of your enterprise. It could be data used by you or data belonging to you and used by others. Characteristics of external data that must be considered are typically: latency, security, accuracy, transport protocols and meaning.
Data Zone (Internal)	The internal data zone contains the data within your enterprise. This could be inside or outside your domain. For domains refer to the chapter on Data Mesh. Important characteristics here are data ownership, data virtualisation and real-time analytics.
Cloud	Data from the cloud could be both external or internal data. Typical considerations you have with storing data in the cloud are your levels of confidentiality which will influence the required level of encryption. Encryption of data at rest and data in motion we assume to be generally available. But if data must be encrypted while processing you can consider homomorphic encryption and confidential computing. A second topic, is the cost of egress. Egress is when you take data out of the cloud. While in the Systems of Engagement or Systems of Record this is usually not an issue, the amounts of data can be large for the Systems of Insight. Therefore you have to consider where you perform your processing and where you store your data. Distributed Processing Services (see Integration Services) provide a service to process data remotely.

INTEGRATION SERVICES



Integration Services

Integration Services are a wide range of services to access data and make data available inside your company or make your data (services) available to others.

Integration could be one single API call or it could be a complex content delivery network (CDN) delivering streaming services. A data-driven organisation must eventually provide this wide set of integration and transportation services to retrieve data from

sources. This data may be be used by business processes to make certain decisions.

The following integration services can be required in a data-driven organisation. In the roadmap you decide what integration services are most important and start with those first.

API Management Services	API Management Services publish and manage your API's. API's are a preferred way of loose coupling between internal and external applications. Most often API's are using the REST protocol. The API gateway is one of the important components to enforce security and manage the load. API Management also collects usage statistics, which can be valuable input for the data scientist to get insight in user behaviour and interest.
Messaging Services	Asynchronous and reliable transport of messages using queues.
ETL Services	Extract, transform and load (large amounts of) structured data between databases.
Data(base) Virtualisation Services	Data can be accessed virtually by the Database Virtualisation Services. There is no need to copy the data to a local repository. The advantage is that the data is *always actual* and *no local resources* are required to store and synch data.
Query Services	Execute a (remote) query on a database to retrieve data for processing. SQL or Structured Query Language is an example of a query language, LDAP is another one you may be familiar with but you may have others as well.

File Transfer Services	File Transfer Services transport files from one location to another in a controlled way. Depending on the amount of data and the level of security you have different tools that fit your needs best.
Data Publishing Services	Data Publishing Services discloses, mostly analytical, data for others to access.
Distributed Processing Services	Distributed Processing Services perform (data) processing at the source. In this case there is no need to transport all the source data, only the results will be transported. This avoids egress cost or could be useful if analysis must be done on actual data, basically on the operational systems. Only limited products are available that can perform that without severe impact on the operational systems.
Events Services	Event Services or Event and Trigger Services detect events and may respond to those events by triggering a process, an activity or a task.

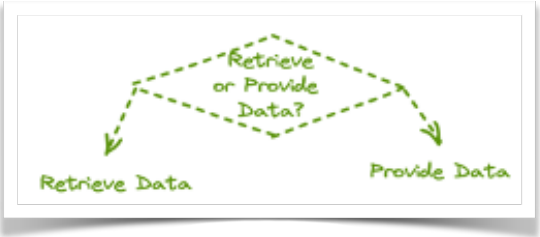
Data(base) Synchronisa-tion Services	Data(base) synchronisation keeps two or more data collections identical. The purpose is to always have a copy of your data in case one of the data sources becomes corrupted. The actuality of the copy is expressed in RPO. This stands for recovery point objective. *RPO=0* means no loss of data in case of an incident. *RPO=1 (Hour)* means you lose one hour of data at most. Change data capture is an example of synchronising two databases. This is done based on the journal data from the database. Change data capture has no impact on the performance of the database because it uses its log file. It is almost real-time (RPO almost 0). Real time synchronisation (RPO=0) is only possible if latency allows.
Content Collector Services	Content Collector Services bring unstructured data under the control of an Enterprise Content Management system. For example email data will be copied into the ECM database. From that moment the email will be controlled by the ECM system, the changes to the email database will be propagated. This is a way to bring your email system under governance of information life cycle management, which could be required for compliance: e.g. who was involved in making the decision at which point in time?

IoT Data Services	IoT Data Service is a service that makes the data from edge devices available for you. This could run from temperature sensor data until drone videos. IoT data is typical data that drives (instant) decisions. The amounts are large, the velocity high. In the example of F1 racing many sensors are used to generate 1200 data point/ sec. In the airplane industry, every airplane landing at the gate transmits 1 TByte of flight data. How about our cars being behind the wheel in the future?

SELECTING YOUR TRANSPORT OR ACCESS TECHNOLOGY

Depending on several characteristics of data you want to access, transport or provide, you can choose from a wide range of technologies to perform that task for you.

The decision table at the end of this chapter is an example to support you with the choice of technology.



Data direction decision

Different characteristics apply for data retrieval as well as for data provisioning.

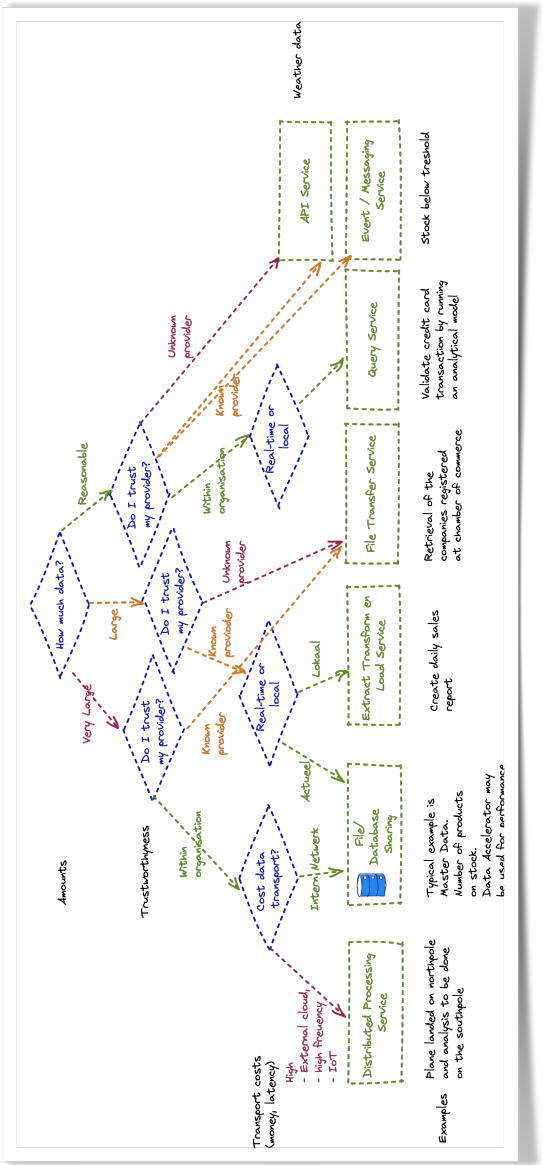
RETRIEVING DATA

For retrieving data you could for example follow this script:

- How trustworthy is my provider?
  - You trust him because it is in your own organisation

    - You know him and you have an agreement in place (authentication/authorisation)
    - You do not know the provider. It is for example open data.
- How much data do I expect to retrieve?
    - Very Large: > 100 GByte
    - Large: > 100 MByte
    - Reasonable: Max 100 MByte
- How expensive is this transport?
    - It could be expensive, because you pay a penalty of latency or financially: egress cost for a cloud provider.
    - It may be almost nothing, because it is on your own premises. You have the cost organising access to it and some energy consumption for your routers and switches.

The following is an example of a decision tree which should be part of your solution architecture (architectural decision)



Decision tree for data retrieval

PROVIDING DATA

When you provide data, the first question is: “why?” Immediately followed by: “to whom?”. There could be several reasons why you provide data to others:

1. It may be part of your business strategy. For example you provide weather data for free to the public but if companies are using that data they have to buy the data. You provide a list of available airplane seats in the hope that others will use that list to get those seats occupied. You have to balance between the cost of making data available and the benefits.
2. You have an obligation, want to be transparent or just for fun. If you are a government, financial institute or non-profit organisation you may have or may want to publish data to demonstrate transparency.
3. It is requested by your organisation or clients to achieve over-arching business goals. If you are part of a supply chain, you want to make your stock available to others so that they can plan production.

The second question is who is using it. There are two aspects to this question:

1. Is there a benefit in sharing the data?
2. What data are they allowed to see?

We leave it up to you to build your own decision tree for data provisioning. But not without some additional advise!

The idea of a data mesh (we address data mesh further on) may become useful.

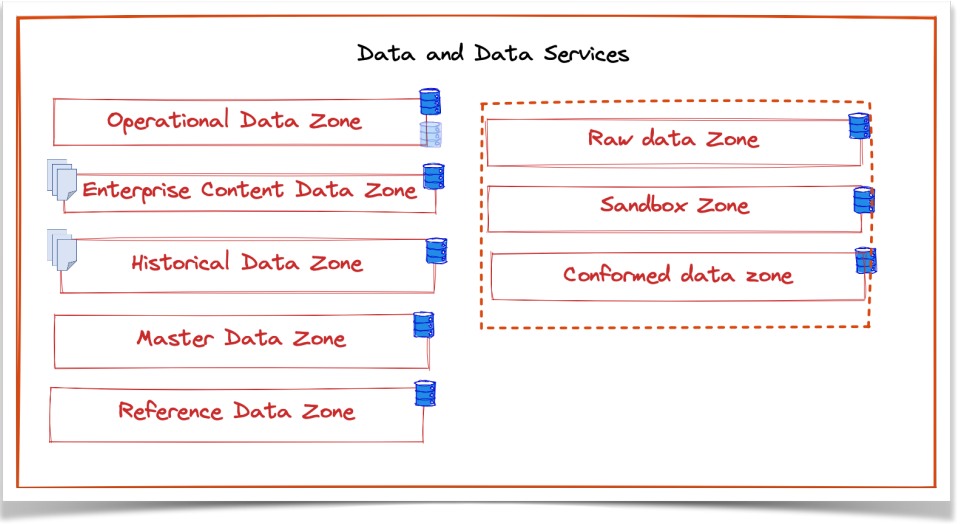
Even though you are probably not so much concerned about data placement from a latency perspective, if you make your data accessible to a large audience on a public cloud be aware of the

egress cost! The amount of data (=usage * size) may result in large costs and you have to consider the business case for doing so.

When you have a need to dive into this topic in more detail, please refer to an article on GitHub with an extensive article on data placement[9].

[9] https://pages.github.ibm.com/IBMAoT/i-hc_data_placement

DATA AND DATA SERVICES



Data and Data Services

Data and Data Services are in essence collections of all kinds of data. There are many different architecture patterns related to data warehousing, analytics, development and test environments, technical patterns for upgrades etc, etc. We do not cover all these patterns but we highlight the most relevant data zones in this context.

- From structured to unstructured,
- from operational to analytical,
- from master data to reference information,
- from real-time to historical information.

We obviously make a distinction between the different types of data and its functionality. For example a zone for data scientists, a

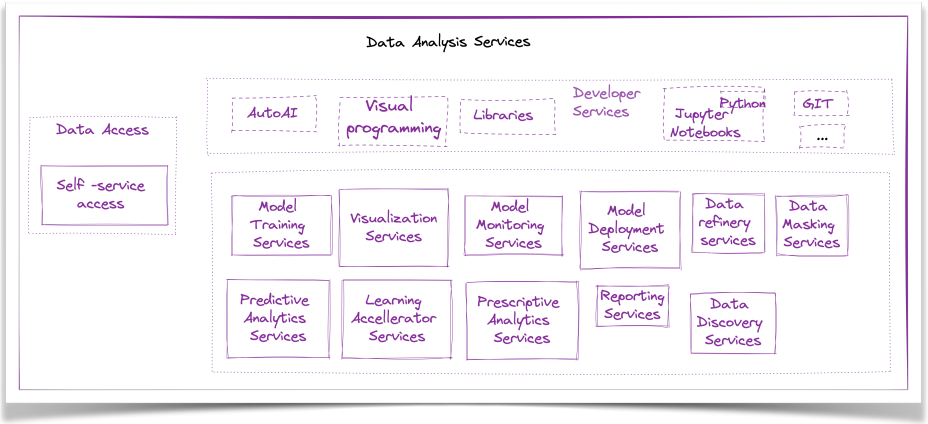
zone for BI, for reporting and of course a zone for operational data to be used in data-driven business processes. Transactional business data is positioned in the Systems of Record.
The next step is to determine the best technology supporting that specific data format. For example a reference table could be name value pairs that could be best be stored in a key-value database like Redis or etcd, while your master data could be stored in a relational database such as DB2 or PostgreSQL. Your governance process is responsible for defining which technologies will be implemented (and which not) in order to avoid a scattered landscape of too many data storage solutions.

Master Data Zone	Master Data is data related to basic registration. Consider the citizen registration as master data or the registration of vehicles. Preferable there is only one source of the master data. The so called "single source of truth". To prevent all kind of copies, master data will be made available through API's. The master data in our view belongs to Systems of Insight even though it will be used frequently by the Systems of Record.
Reference Data Zone	Reference Data is comparable with master data, however it is more static data. Country, language codes are examples. But also prices for energy, which could fluctuate by the hour are still considered to be reference data.

Enterprise Content Data Zone	Enterprise Content Data is unstructured or semi-structured data. Large amount of data like documents, emails, video's, voice recordings and images are all considered content. For systems to understand this data it uses metadata belonging to the content. An ECM Database consists of two parts, the metadata, used as an index and the content itself. AI is frequently used to interpret the content.
Operational Data Zone	The Operational Data zone contains the operational data of the Systems of Insight. This could be relational database, noSQL databases or even the etcd database of the Kubernetes environment where all config data is stored.
Historical Data Zone	Operational systems and enterprise content data systems will remove old historical information to make sure that the operational systems will continue to perform. While this data may be worthless for the operational systems, the value of historical data for the data scientist may be tremendous. Therefor data will be kept for analytical purposes of any kind. Often sensitive information like social security numbers are not required and will be anonymised.

Raw Data Zone	The Raw Data Zone contains copies of operational data to be used by the Systems of Insight. The copy is made to prevent operational impact on the Systems of Record. Most common format are relational databases, but it could also be other types of structured data, like NoSQL databases. This raw data zone can be used to feed data warehouses and data marts, these raw data copies are typically stored in a so called staging area (intermediate/ temporary results). But also for analytical exploration and discovery, including real-time analytics.
Sandbox Zone	The sandbox is a safe playground for the data scientist. He or she can do no harm and they can experiment with the data to their heart's content.
Conformed Data Zone	The conformed data zone contains curated data ready for usage. This could be a data warehouse, a data mart or the training dataset for a model that will be deployed into production and has undergone processing and manipulation.

DATA ANALYSIS SERVICES



Data Analysis Services

Data Analysis Services could be considered the core of the capabilities/services required by a data-driven organisation. It can provide deep insight into data. From reflecting on events in the past from historical data to predictive insights on what is going to happen.

AutoAI	AutoAI is a capability for citizen scientists or data scientist to start quickly with making models. It can thus be used by data scientists that do not have a programming background.

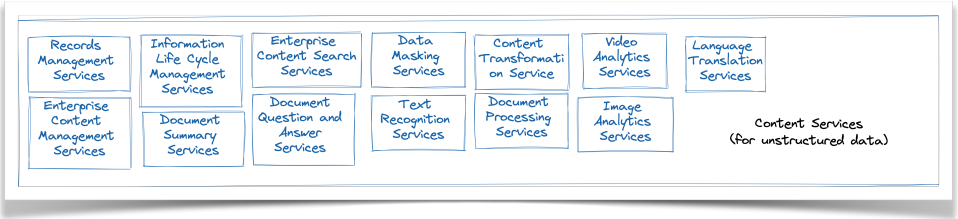
Visual Programming	Visual Programming is a tool that supports data scientist to visually create a data flow. Dragging sources and operations into a canvas will provide an easy way of modelling. Tools like SPSS Modeler are familiar visual data preparation and modelling tools for most data scientists.
Jupyter Notebooks	The typical tool for the data scientist to create models is using a document format (notebook). It can use multiple development languages, but the name Jupyter comes from three popular languages: Julie, Python and R.
Libraries	Analytical and statistical (open source) libraries that can be used by data scientists in their programs.
Python	The development language for the programmer data scientist. Another commonly used language is R. This is developed for statistical and data analytical purposes. We mention the most popular language, but this service could also have been called "analytical programming language services". Feel free to replace this with the languages used in your organisation.
GIT	GIT is a repository to store source data of any kind. Different versions can be kept in a GIT repository. In the cloud native environment it becomes custom to build applications from the source. This way you do not run into situations that the source code for a running module cannot be found. For trustworthy AI it is important to keep track of both the source applications, as well as the source data. GIT will support this form of documentation and compliance.

Data Masking Services	Data Masking for structured data means anonymisation or pseudonymisation of the data when displayed to audiences that do not have the rights to read it. Quite often used to comply with GDPR. It will show an asterisk instead of a banking account or it will generate a random banking account that does not exists.
Model Training Services	The service to train the model by feeding data into it.
Visualisation Services	Visualisations Services visualise the outcome of the models in a user friendly way. Different visualisations can convey different messages. Choosing the appropriate visualisation makes sure that the outcomes of a model can be quickly interpreted by a user or consumers from the business.
Model Monitoring Services	Monitor the deployed model by validating if the model is still operating within defined boundaries. It measures trust, drift, fairness, bias, etc.
Model Deployment Services	Support the model hand-over to production. This is experienced as a complex and time consuming activity but with the deployment service, model deployment into production can be automated.
Data Refinery Services	Data scientists have access to large amounts of data. Data refinery service supports the data scientist in making selections, transformations and manipulation of data that are required for a specific analysis.

Data Discovery Services	Data Discovery Services are intelligent components that discover the types of data by reading its content. It can recognise e.g. an email address and classify that as personal data. Data Discovery Services can identify hundreds of different data types and support the data operator to build a data catalog. This is used by the data engineer and the data curator to make the data available to users.
Predictive Analytics Services	It provides predictive insights like trends and anomalies based on the models. AIOps also uses these services to predict possible incidents by observing normal behaviour as well as anomaly detection.
Prescriptive Analytic Services	It provides what action to perform based on the insights generated by using the models. AIOps also uses this to take actions to prevent incidents to take place. In this way, systems can "heal" themselves, so called self-healing systems. Because the data-driven organisation will most likely have thousands of microservices deployed over time, these services become more and more important.
Learning Accelerator Services	Model generation based on large sets of data takes significant time and capacity. A learning accelerator combines all available (GPU) capacity offers the input to a batch scheduler to create a model and optimises in that way the usage of scarce (GPU) resources.
Reporting Services	The reporting service provides the capability to create daily, weekly, quarterly and monthly standard reports or ad-hoc reports and dashboards on request (Business Intelligence).

Self Service Access	This service allows for access to distributed data, by enabling data discovery and analytics either for business users or for more advanced users, regardless of where the data is stored or the underlying systems.

CONTENT ANALYSIS SERVICES



Content Analysis Services

What Data Analysis Services are for structured data is Content Analysis Services for unstructured data.
During our discussions with experts we gathered a good insight in the difference between the usage of content and structured data.

Structured	Unstructured
Good for predictions	Used for evidence
Prescriptive for determining next steps	Knowledge base
Reports and Business Intelligence	Content + Structure (Metadata)
Large volumes, small amounts	Large files

The following services have been identified:

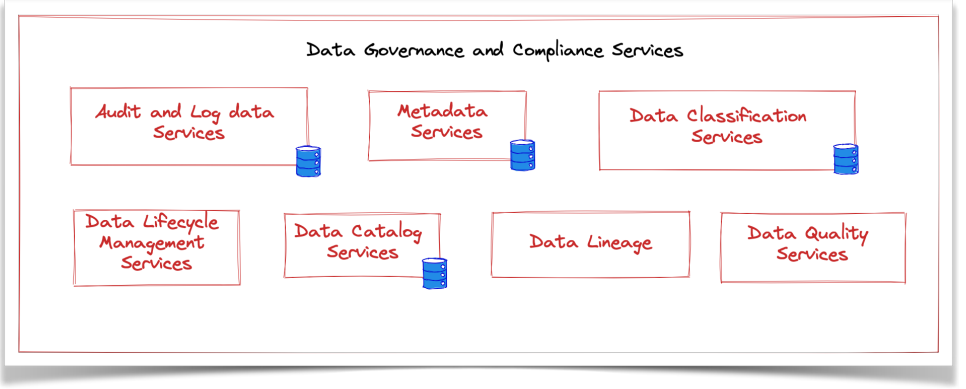
Text Recognition Services	Text Recognition Services recognise (handwritten) text. Often used by the Input Management Services.
Records Management Services	Records Management is the implementation of Information Life Cycle Management for unstructured data. Content could come under governance or records management. The moment the content is assigned a record, creation, usage, retention and deletion will be assigned to that content and be handled by records management.
Enterprise Content Search Services	Enterprise Content Search Services search through a federation of content repositories, both in metadata as well as in the content itself. For the content itself it will use services like Image and Video analytics services.
Enterprise Content Management Services	ECM is for unstructured data what a RDMS is for structured data. It is tightly related to Enterprise Content Data (Refer to Data and Data Services). It handles all kind of activities that are executed against the data, like viewing, basic search, linking metadata to content, etc.

Document Summary Services	The Document Summary Service gives you a summary of one or more documents. This could also include sentiment indications. The amount of good summary services is still limited. It seems to be quite difficult to extract the highlights from a document. Most of these types of services are available in the cloud. You have to verify if they comply with your security and GDPR standards. Wouldn't it be nice if a video summary can be made or a picture summary. I would like to have that for my holiday pictures! In the Document Summary Services the question always remains" "What is not in the summary that should have been there?".
Document Question and Answer Services	Document Question and Answer Services read and interpret the document so that you can ask it questions which then will be answered by the service. An example is GAAMA. Bored by reading all this stuff? Go to the Internet, search for "GAAMA is a Natural Language Understanding technology" go to the side, select the tab "custom" and load the PDF of this booklet (copy - paste) and ask the question: "who are the authors?" That's us! nice to meet you! Then try a less serious question like "what is a data-driven organisation". Wow it knows it! It even knows who invented Data Mesh! What makes this service useful is that it returns the certainty factor and the line where it found the answer to validate if the answer is correct.

Data Masking Services	Data Masking Services for unstructured data masks data based on the profile of the user. Examples are masking number plates. Replace cursing with beeps, replace names (GDPR) with roles. Black blocks are also a form of data masking in documents.
Content Transformation Services	Content Transformation Services can do the transformation from one data format into another. For example it transforms a speech file to text, a video file to pictures or a PDF to data. OCR could be seen as a specific form of Content Transformation.
Document Processing Services	The Document Processing Services take a document and interprets the context of that document and puts that context into your business process. For example you received an invoice, the Document Processing Services reads the invoice and recognises the amount to be paid, the sender, the amount of tax etc. and feeds this into your business process. Document processing service can be used by input management for documents. If parts of the form are in handwritten or "text in picture", text recognition services can be used to translate that input into machine readable characters.
Video Analytics Services	This service does the same for video as what the image analytics services do for images. For example you provide a dashcam video and it automatically extract the piece of video where the accident happens. Or it collects the faces of people using violence in a stadium.

Image Analytics Services	Using AI this service will analyse pictures. Image analysis involves processing an image into fundamental components to extract meaningful information. Image analysis can include tasks such as finding shapes, detecting edges, removing noise, counting objects, and calculating statistics for texture analysis or image quality. Depending on your context, you may want to split this into multiple services to increase the accuracy of the analysis. An example could be the "French car number plates recognition service". This would be an image service specifically trained for recognising French number plates. Typically in a data-driven organisation you will find many of those types of analytical services.
Language Translation Services	The Language Translation Services translate content from one language to another. That could be text, but also spoken records. What about translating the text in pictures? How would you translate those?
Auto Classification Services	The Auto Classification Services can read the content of your data and determine the metadata for that content and can also assign a security classification to your document, for example when it finds personal data like an email address in the document.

DATA GOVERNANCE AND COMPLIANCE SERVICES



Data Governance and Compliance Services

Data governance and compliance solutions are crucial for data-driven organisation as they help understand what data is available, where that data resides and how it can be used, ideally via self- service mechanisms. It fosters trust and provides lineage to trace back data sources, transformations and movement. Finally, they support compliance with corporate and regulatory requirements.

Just as a reminder, technical governance is never enough, we have said it before, an organisational culture often needs proper changes and support from leadership is necessary as well.

Data Lineage Services	Data Lineage Services are services that keep track of the data flow. It understands the flow of data from the source until it ends up in a report or an analysis and it also understands what happened to the data during that flow.
Data Catalog Services	The Data Catalog Services provide the definition of all the attributes and entities, this could be business or technical. It provides the relationship between those definitions and it also holds data about the users of those entities.
Data Quality Services	Data Quality Services detect anomalies, which help you to determine the quality of the data and improve the quality of the data. These services are also known as data profiling.
Audit and Log Data Services	In a data-driven environment decisions will be made by people and machines based on data and algorithms. Audit data explain on what basis decisions have been made. It also contains data on who or what made the decisions and provides full transparency in the processes. The log data describes who did what. In a separate chapter of this booklet we will spend some words on AI ethics and compliancy and why this is so important.
Metadata Services	Metadata Services provide definitions of ontologies, taxonomies with all its relationships. It is the representation of the organisation in a data structure. Where the data catalog describes the data itself, the metadata describes how the entities in the catalog relate.

Data Life Cycle Management	Records Management is the Data Lifecycle Management approach for unstructured data. However similar services do apply for structured data as well. These solutions are sometimes known as data archiving solutions, because that is most likely where the focus is. Cleansing your operational environments to keep them performant. While this may be part of your Systems of Record, the resulted archive most likely becomes part of the "Systems of Insight". And your data analyst thanks you for providing access!
Data Models	You have to govern your data models themselves as well. This is a description of the implementation of the database.
Data Classification Services	Data Classification Services assign a data class, this describes the type of data contained in data assets. This helps to describe the domain of the data values. This is both applicable to structured as unstructured data and can be previewed by catalog users.

AI TRUSTWORTHINESS HITS THE AGILE MANIFESTO BETWEEN THE EYES

Are you raising your eyebrows and thinking what do they mean with this strange title? Let us explain. The agile manifesto states: "***Working Product over Comprehensive Documentation***", it also states: "***Individuals and Interaction over Processes and Tools***".

Both statements are not a good idea when we are talking about AI Trustworthiness and EU AI regulations. And since there are only four statements in the agile manifesto, we are talking about 50% of the manifesto, that hurts!
So, from a trustworthiness perspective, it would make sense to review the manifesto carefully, because it can cost you up to 30

million euro or 6% of your global revenue when not complying with EU AI regulations, do we have your attention?

That is also the reason why we describe “governance and compliance” as a separate group of services.

Back to the EU regulation[10]: the purpose of the regulation is to ensure AI is used fairly. These regulations are applicable to any AI system used in the EU. They only apply to situations where AI is exposed to the public, for example a chatbot that advises you to cool down your hot, 240V power supplied hairdryer under the cold-water tap... Some applications of AI are not allowed at all within the EU, for example social scoring. Also, applications that apply face recognition are very sensitive, that is why companies like Microsoft and IBM discontinued these services, even though competitive advantages for companies to use these are very big.

These regulations are not applicable internally, so if you completely mess up your own organisation the EU does not care. It's on you. Therefore it may be a good idea to use the regulation as guidelines to protect your own organisation from serious missteps.

But before you become really scared, the regulations are not yet in force, which gives you some time to fix it, or better yet, start with a thorough approach by implementing the governance and compliance services that we have suggested and setup your governance in your organisation. That is also why we have spent some sentences on ModelOps and why we have the model monitoring service in place, to detect e.g. model drift.

[10] https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

We have defined five “pillars of trust” which we believe your AI solution must comply to[11]:

1. Fairness
2. Explainability
3. Robustness
4. Assurance
5. Governance

Be aware that there are lots of discussion on this topic and also a lot of good principles established that you can take over. Here are for example the three from IBM:

The purpose of AI is to augment human intelligence

The purpose of AI and cognitive systems developed and applied by IBM is to augment – *not replace* – human intelligence. Our technology is and will be designed to enhance and extend human capability and potential. At IBM, we believe AI should make ALL of us better at our jobs, and that the benefits of the AI era should touch the many, not just the elite few. To that end, we are investing in initiatives to help the global workforce gain the skills needed to work in partnership with these technologies.

Data and insights belong their creator

IBM clients’ data is their data, and their insights are theirs. Client data and the insights produced on IBM’s cloud or from IBM’s AI are owned by IBM’s clients. We believe that government data policies should be fair and equitable and prioritise openness.

New technology, including AI systems, must be transparent and explainable

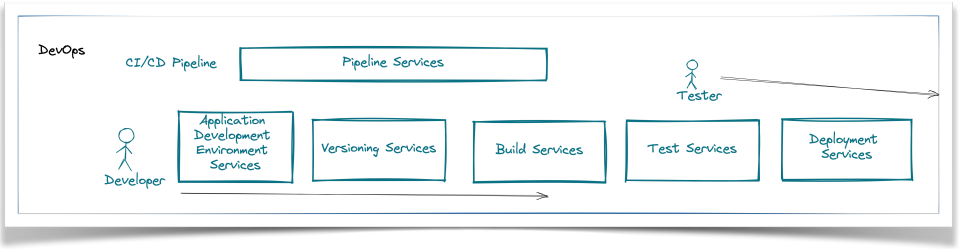
For the public to trust AI, it must be transparent. Technology companies must be clear about who trains their AI systems, what data was used in that training and, most importantly, what went into their algorithm’s recommendations. If we are to use AI to help make important decisions, it must be explainable.

---

[11] https://www.ibm.com/blogs/watson/2020/10/how-ibm-makes-ai-based-on-trust-fairness-and-explainability/

These are not statements that you just write down. These are statements that you should live through within your organisation.

CI/CD PIPELINE SERVICES



CI/CD Pipeline Services

A CI/CD pipeline is important to be able to develop and deploy our applications in an agile way.
There is so much information available about this topic, we only have listed the most important services. There is much more that you can include in your pipeline. There are however a couple of important guidelines:

- Make sure your pipeline is cloud independent.
- Don't force yourself to go for one pipeline, but develop a limited amount of pipelines for different type of systems. We call that "fit for purpose". For example, a CI/CD pipeline you developed for your Systems of Engagement is most likely different from the one you use for your Systems of Record.
- Consider future development that should be supported by your pipeline, like server-less.
- Include sustainability aspects: how efficient is the pipeline itself and the code it produces?
- Make sure the pipeline is secure. DevSecOps is raised as a concept to make sure that there are no vulnerabilities in your

code that you put into production and helps prevent costly redesigns. Test at least OWASP vulnerabilities[12].

- GitOps is on the rise as an interesting concept. It may be interesting to consider this concept because it forces you to think and act in a cloud native way[13], taking most out of automation. Even though we put automation and provisioning services in the infrastructure management layer these capabilities could be considered part of a CI/CD Pipeline as well.

Pipeline Service	The CI/CD pipeline service is the orchestration service that links specific services which together create an end to end process from code creation until deployment of a service in production. The amount of available services for composing the pipeline is extremely large. XebiaLabs created a nice map (called periodical table) of 'all' services related to their objectives (testing, versioning, deploying, ...). In our case we list only the major ones. We recommend that pipeline services and services executed by the pipeline should preferably be cloud agnostic. A second consideration is the movement towards server-less computing. And finally consider the usage of more lightweight, born in the cloud pipelines. If you start new development, it is worthwhile to look at tools like Tekton and GraalVM instead of good old Jenkins.

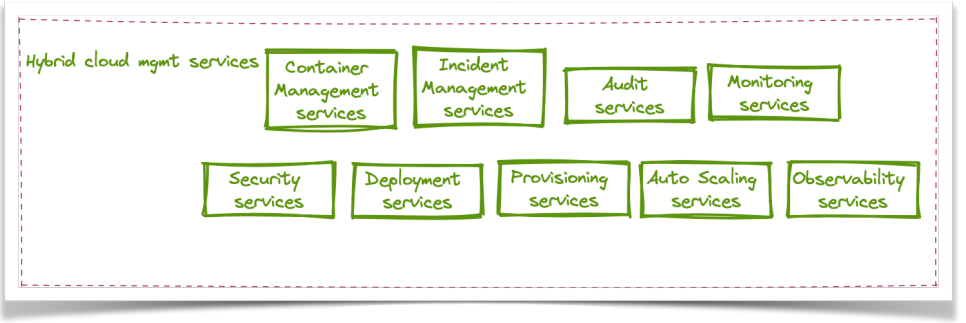
[12] https://owasp.org/www-community/vulnerabilities/

[13] http://ibm.biz/ProgModel

Application Development Environment (ADE) Services	The application development environment is the environment where an application developer can create and perform a unit test. Very important in this environment is that developers can iterate very quickly between changes to code and performing tests. Write and test code is a continuous cycle that happens many, many times per day. Performance is the key success factor for this environment. In a data-driven organisation developers will build services that either provide access to data, data models, algorithms and business rules to improve the business process.
Versioning Services	Version management could be very complex in an environment where many releases (branches) of an application are running. Version management should be integrated into the CI/CD pipeline. As soon as versions are "checked-in" into branches the CI/CD pipeline will start building, testing and deploying a new version of the code. An example is Git, it becomes a trend to run everything from there. GitOps maybe something to consider.
Build Services	The build services bring libraries and code together in a compiler to create an application (micro service) that can run in the operational environment.
Test Services	Test services are those services that can perform all kind of tests against the developed application. In Dev**Sec**Ops, security testing is paramount. Make sure you at least test on the OWASP Top 10 vulnerabilities.

Deployment Services	In the data-driven environment applications will be deployed in the container runtime environment. Usually that container environment is managed by Kubernetes. Kubernetes handles a lot of the deployment activities, for example it allocates your runtime, setups disk space and installs monitoring tools etc.

HYBRID CLOUD MANAGEMENT SERVICES



Hybrid cloud management services

A data-driven organisation has the ability to expose its data-services as microservices and run them on a container platform. This container environment must be managed. Due to the large amount of microservices and containers that must be managed, it becomes almost impossible for humans to keep track of the entire environment with hundreds of containers and thousands of connections. That is why AIOps is put in place and AIOps requires a broad set of tools. The hybrid cloud management services provide the set of monitoring, observability and systems management tooling. AI is used to predict incidents and prescribes solutions to resolve it. With focus on sustainability, over-allocation of resources is no longer acceptable and application resource management will be put in place to right size the environment.

Container Management Services	Kubernetes is the standard management service for managing containers. It provides scaling and self healing capabilities.

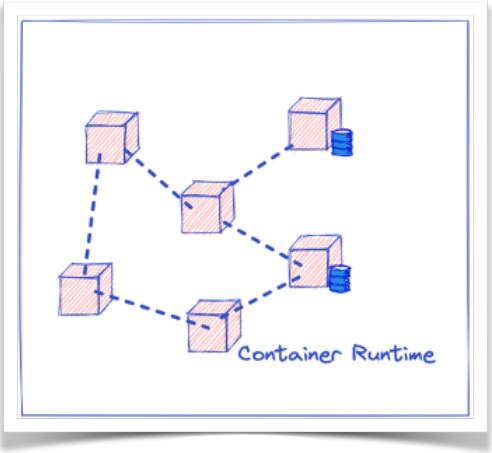
Incident Management Services	Register the incident and monitor the actions to resolve that incident from inception until it is resolved. The Incident Management Services are part of the Systems of Record, however as we observed, in a data-driven organisation the Systems of Insight becomes more and more part of the Systems of Record. The moment you use the SoI to control the SoR it essentially becomes part of it.
Audit Services	Audit Services collect lineage and audit data and create reports for compliancy.
Monitoring Services	Monitoring Services continuously report the status of hardware, operating system, middleware, network, virtualisation and applications and highlight problems or potential problems based on thresholds like CPU and memory and storage utilisation.
Security Services	Security Services is a collection of all kinds of security services for identity management, privileged access management, vulnerability testing, zero trust, malware, intrusion detection etc. Security Services are not worked out in detail because they apply to all your Systems. In a data-driven organisation special attention may be on data security, the most valuable asset in your organisation. Make it secure by design.
Deployment Services	The Deployment Services are the same as the services in the CI/CD pipeline. But not all applications are part of a pipeline, for example application you bought, middleware products, etc. Your architectural decisions should guide when to use the Deployment Services from the CI/CD Pipeline and when to use those in the Hosting Service.

Provisioning Services	This service is the core of "Infrastructure as Code". It has the ability to create infrastructure that can be used by applications. For example you execute a script to setup a new vLAN or you setup a script to create a new storage volume. There are several scripting tools like Chef, Puppet, Terraform but most widely and commonly accepted is Ansible. Cloud providers do have their own tools. It is a fundamental architectural decision to go for a generic tool or cloud specific tools. Our multi-hybrid cloud vision clearly votes for the first.
Auto Scaling Services	On one hand this is the ability to scale within the container environment but also in the traditional environment, for example VMs. Kubernetes can assign worker nodes and machines to clusters dynamically. In the VM world there are services available to scale up and down. Auto scaling becomes a key sustainability requirement.
Observability Services	Understand infrastructure and the relationships within the infrastructure. It determines the performance of your systems and can be used to find and analyse root causes of issues. This is also known as application performance monitoring.

APPLICATION RUNTIME

It may sound a bit obvious. You need an application runtime to run your applications. And why should this be a microservices architecture running in containers?

In February 2022, Randy Bean published an article in Harvard Business Review[14]. As an important solution for overcoming the obstacles his advice is: “Fail fast, learn faster”. A microservices based architecture supports this advice. A second reason why this runtime is very important is that it should be able to run anywhere.



Container runtime

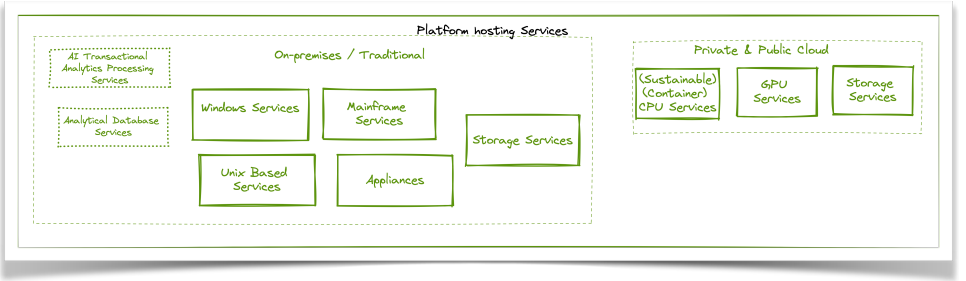
The applications should have the same runtime available when they run in the private cloud, the public cloud, when they run on bespoke technology but also when it runs on modern RISC based sustainable technology. It should even run at the edge, when the need is there to process the data at the edge, for example to avoid large egress costs or when bandwidth is simply not available.

So we believe that the selection of a runtime environment (e.g. AKS, Open Shift, GKE) is not that obvious. However bear in mind the runtime environment is an essential part of the blueprint for the reference architecture.

The choice of the runtime could be related with the CI/CD pipeline. The combination of the runtime environment and the CI/CD pipeline should support the process of “Fail fast, learn fast”. Later on we will highlight the importance of ModelOps to provide the same speed to create models.

[14] https://hbr.org/2022/02/why-becoming-a-data-driven-organization-is-so-hard

PLATFORM HOSTING SERVICES



Platform Hosting Services

The infrastructure provided is based on cloud. Not necessarily public cloud, but this could very well be a private cloud. For data-driven organisations it is very important to consider the potential cost of data egress and data ingress before you decide where to perform the data analysis. In general it is best to do the analytics at the location where your data is stored. Due to the agile way of working, large amounts of data may be copied from operational systems to the systems of insight.

Second topic to consider is the analytical infrastructural capabilities, in other words, the GPU processors. If you have a lot of work for the processors you want to “own” them. If the workload is fluctuating, GPU’s may be used in the cloud.

A third is the question if analytical capabilities are required directly on the operational data. An example here is a credit card transaction that takes place. To validate the transaction against analytical models, you will need these capabilities at your operational systems. From the perspective of a data-driven organisation, we believe that this could be a differentiator for your

company to better secure and validate your transactions while they are being executed.

Therefore we split the platform hosting service into three types of hosting services:

1. The private cloud. Many companies use this as the production environment to run the models. Where models are trained is dependent on the amounts and location of data. The Learning Accelerator Service is of indisputable value for optimising scheduling model training and can increase processor utilisation up till 95%.

2. The public cloud. This could be used as a development environment. Specifically for smaller companies that have limited models to train or do not have the scale nor people to manage their own environment. When you start building models and are expected to grow, it is recommended to use tools that are not only available in the public but also in the private cloud to avoid cloud lock-in.

3. The existing on-premise environment where you make capabilities available for real time integration of the Systems of Record and Systems of Insight for the reasons described earlier. Because the solutions are developed close to the processor hardware and/or the operational database management system, they have a proprietary character.

AI Transactional Analytics Processing Services	A platform service to handle analytics. It contains GPU type of capabilities. There are several solutions in the market. This service consists of several solutions that takes the best in the context of the processing to be done.

Business Continuity Services	Data-driven organisations become more and more dependent on the "Systems of Insight". This increases the need for qualitative aspects like high availability and disaster recovery. Different type of solutions require different HA/DR services.
Windows Services	For traditional (Windows based) data-driven workload that cannot run in containers yet.
Unix Based Services	For traditional Unix based workload, based on proprietary OS like AIX, HPUX.
Mainframe Services	Enterprise mainframe platform for high transactional workloads. In the chapter "Fast Food" we described the rationale of including a typical "Systems of Record" service in our "Systems of Insight" blueprint. The focus is on the real-time analytical capabilities and real-time insight in the operational system.
Database Accelerator	The Database Accelerator allows you to perform analytical queries on an online operational database. In most cases this would not be considered a good practice, but if there is a requirement to provide actual insights at any point in time, it is a service that does not impact the performance of the operational database.
GPU Services	GPU services work very well with neural networks and are therefore ideal for AI. The GPU services could be GPU only services or could be combined into a regular (CPU) server. GPU services are quite expensive services. The Learning Accelerator Services (see learning accelerator services) can optimise the workload.

(Sustainable) (Container) CPU Services	RISC or CICS based processors that can run your applications. This processing power could be available locally or in the cloud. Depending on how serious you are regarding sustainability, significant hardware improvements can be made for container based workloads running Linux on, for example, an ARM, RISC-V, OpenPower or Telum based processor architecture.
Storage Services	We combined storage into one simple block of storage services. You will have to differentiate between block, file and object storage as well as between active, passive and archive and finally backup storage. We have explained this in the following paragraph about storage. But for now, no worries, these storage services are managing it all for you, from (over-)provisioning to storage tiering.
Appliances	There are many appliances around, from Teradata platforms to Data Power integration.

STORAGE

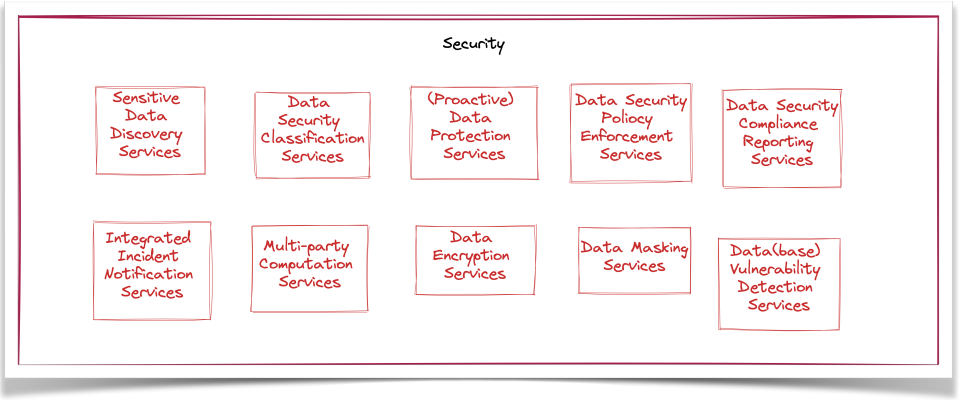
In our blueprint you found one “Storage Services” service. The systems of insight is a large data consumer, therefore, from that perspective it is worth to spend some words on storage.

First you have to distinguish between block, file and object storage. Second you have to distinguish by the storage medium like Flash, Tape, Disks. Thirdly you have to think about the access characteristics like active, passive and archive and finally you have to consider backup and high availability. And to make it more complex, there is a relationship between those aspects because you don’t use object storage on tape medium for active access.

Storage tiering is a concept that can transparently move your data between storage media. Other services are: storage virtualisation, storage compression and over provisioning.

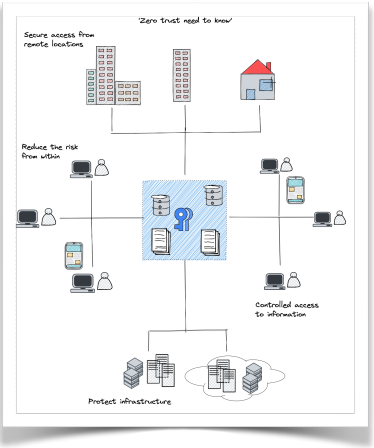
You may have the feeling that data storage is not very expensive, so why bother? And in a way that is correct. Storage has become very cheap if you store 1 GByte or even a couple of Terabytes that is no concern. And there is a huge factor of 1000 between those. When you multiply it again by 1000 it becomes 1 Petabytes and suddenly it becomes a factor, now only multiply this 1 PetaBytes by a factor of 10 and it becomes a topic in the management meeting and multiply that again by only a factor of ten and it becomes a boardroom discussion. Therefore, you need to define a storage strategy how to handle all that data. If your company has a hyper converged infrastructure strategy, that is the storage, CPU and network are all converged into one box, which is ideal for the Systems of Engagement. But you have to review if that fits your data intensive use cases as well.

DATA SECURITY SERVICES



Data Security Services

Often security is not an integral part of an architecture for data-driven organisations. However we think the more organisations are working with data as a vital instrument the more it needs to be protected from malicious influences. Therefore we feel data- driven organisations need to adopt a security architecture based on 'Zero trust, need to know' principles. A basic mistake we also identified, is that a good cyber security strategy is sufficient enough to protect your data as well.
As we all know the danger of attacks (modification, deletion, stealing or sharing of data) is very serious. The complexity of employees being able



Zero trust - need to know

to work from any location and from any device as a result of the Covid pandemic accelerated the need for advanced security architectures.

Traditionally applications were built with access control (who may use this application) and within the application delegation rules were applied in order to have control over who may see/access what information or who may initiate a process. With data being used in other processes and accessed in many other ways other solutions are required in order to have control over access to the enterprise data.
Security, Compliance and Governance are three topics that are different but related. With governance we make sure that we are in control, with security we implement that control and compliance provides us with the evidence. Let us not make it more difficult than this.

But.....at first we doubted whether we should go into the topic of security as it is complex and you can write a complete book on the subject. But we have decided we **must** at least address the topic of *data security*. Security is too important to ignore in a data-driven environment. Let's see what we discover when we open up the 'geode'[15]

In our 'writers' discussions we were confronted with some dilemmas that we would like to share with you. These dilemma's were also the reason why we hesitated on describing data security in more detail in the first place:

- Many of the Data Security Services are already available as services in other domains in our blueprint. Why should we describe them separately in data security services?

[15] A **geode**; is a geological secondary formation within sedimentary and volcanic rocks. From the outside it looks like a normal rock but geodes are hollow, vaguely spherical rocks, in which masses of mineral matter (which may include crystals) are secluded.

- We expect that many of our Data Security Services are already available in your organisation as an enterprise security service. Why should we describe them separately in data security services?

> *The importance of data security is sometimes poorly understood. You implement all infrastructure and application security measures, and it feels rock solid. But data security is like a 'geode'... When you open it, there is a whole new dimension to it. You discovered a whole new set of data security services to make your enterprise more safe and less vulnerable.*

For example, in our blueprint we defined Data Discovery Services, Classification Services, Masking Services etc. We discussed whether or not we should address these services again in the data security domain? Our initial conclusion was: **No....!** (For the reason mentioned above: complex and broad).
But on second thought we decided we must go into a little more detail for the same reason as why we made a distinction between structured and unstructured data services. From an architectural perspective they are the same services but the reality is that the technologies used for structured and unstructured data are quite different.

How good are your *general* Data Discovery Services compared to the *specific* Sensitive Data Discovery Services? The latter is specifically built for detecting sensitive data, so you might expect that these services are very advanced and are really good in understanding the sensitivity of your data and for example it can

map your data to the CIA Rating. CIA stands for Confidentiality, Integrity and Availability[16].
In the ideal world, Security Services would preferably be one set of services, but as you know we are not living (yet) in an ideal world. So maybe today, you need both sets of Discovery Services. You will find the Data Security Services we defined are always going a level deeper.
Instead of securing 'which system is accessing what database', we want to know and control more: 'who is accessing what data, how is he/she accessing this data and in which context is this data accessed, etc'. In both cases it is an access control service, but the level of control mechanisms are making the difference.

We defined this sometimes overlapping set of services to draw your attention to the different aspects of data security.
For each of those services ask yourself the question: is it addressed well enough as a service in another domain or is it more effective to implement **specific** data security services?

Where Security Services could be compared with a bouncer, some Data Security Services are like undercover agents, they walk around in your organisation, always alert for the person who behaves suspiciously.

P.S. They do not hesitate calling the bouncer to eliminate this person....!

In a recent study it was shown that the cost of data breaches is increasing[17].
And of course a data breach will not happen to your company... until..... it happens.
'Penny wise pound foolish' is the proverb to remember.

---

[16] https://www.fortinet.com/resources/cyberglossary/cia-triad

[17] https://www.ibm.com/security/data-breach

Sensitive Data Discovery Services	When data is stored there could be sensitive data included. Think about IBAN numbers, telephone numbers, social security number, (email) addresses, etc. Sensitive Data Discovery searches throughout the enterprise in structured and unstructured repositories for sensitive data. It can also correlate the sensitive data. This is very useful in the case of GDPR for example, to identify all the sensitive personal data that belongs to one person.
Data Security Classification Services	We have defined classification services already in other domains (see Data Governance and Compliance Services). The Data Security Classification Services are doing exactly the same the generic classification services but specifically from the perspective of data security and on a much more detailed level. It goes hand in hand with Data Discovery Services. Where Data Discovery Services identify sensitive data, the Data Security Classification Services define the class of the data (CIA as an example) .
(Proactive) Data Access Protection Services	These services protect your data from access by unauthorised users. It constantly monitors the queries done on the data and it compares this with the profiles of users in your authentication system. It can identify queries that are out of the ordinary (anomalies). Based on policies, you determine what actions must be taken.

Data Security Policy Enforcement Services	These services extend the Data Access Protection Services by validating the queries against policies. For example an HR employee accesses the payroll. That would not be a surprise, so the Data Access Protection Service is perfectly fine with this request. But what if this employee is performing this activity on Sunday midnight and request the salaries of all employees of the company? Policies could for example be set to maximise the number of employee salaries queried per day or on which day of the week this query can be executed.
Data Security Compliance Reporting Services	Financial organisations, governments, health care providers, retail organisations, all organisations have to deal with regulations. The Data Security Compliance Reporting Services support you with generating reports which can prove your organisation is compliant with the (security) regulations applicable for your industry.
Integrated Incident Notification Services	In your security environment you will have a Security Information and Event Management Service (SIEM). Integrated Incident Notification Services are services that collect your related data security incidents and provide one consolidated alert to your SIEM.

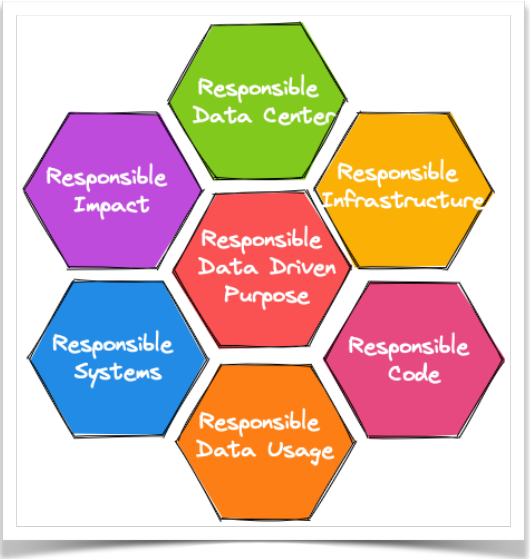
Secure Multi-party Computation Services	With secure multi-party computation (MPC) services you can perform data analysis across several sources without sharing the data. These are probably the only part of services not defined somewhere else, even though strongly related to (federated) distributed processing services. Encrypted data requires specific e.g. homomorphic encryption technology in order to use the data without the need to de-encrypt the data. For more information on homomorphic encryption see: https://www.ibm.com/security/services/homomorphic-encryption.
Data Encryption Services	Data Encryption Services enable you to encrypt your data and manage your encryption keys. That data encryption can be done on database level, on column, row or field level or on storage level. As an architect you will have to define a set of architectural decisions about encryption. For example 'Hold your own Key (HYOK)', 'Bring Your own Key (BYOK)' or 'Control Your own Key (CYOK)'. You have to think about key management, for example the usage of a HSM (Hardware Security Module). As you can see, this set of services require some thought!
Data Masking Services	Data Masking Services are those services that support the masking of data so this can only be viewed if permitted. These services can be called by the Data Access Protection Services.

Data Vulnerability Detection Services	Data Vulnerability Detection Services detect all kinds of internal and external threats. It detects weak passwords, excessive (DBA) logins and other unusual activities as well as missed database patches and misconfigured privileges. Detection of vulnerabilities may lead to adaption of the current security policies. E.g. this system may no longer be used in a production environment until all patches for security are applied.

SUSTAINABILITY

We **must** address the topic of sustainability, we cannot ignore it!

In this updated edition of our booklet, we decided to address sustainability as a separate topic. Not because it is a trend, or it would "sell" better (even though this booklet is for free) but because it is necessary. We did address several sustainability topics already, but we felt the necessity to highlight it explicitly.



Responsible Computing Framework

We also believe that resiliency cannot exist without being sustainable. If you want to build a resilient data-driven organisation, it has to be sustainable.

We took the Responsible Computing framework as a starting point. This framework will help us to focus on the people, process and technology topics in our context. Even though a data-driven organisation would be able to support the *United Nations Sustainability Development Goals (SDG)*, we think that is too big for our scope.

Responsible computing contains 6 domains:

1. Responsible Code
2. Responsible Infrastructure
3. Responsible Data Centers
4. Responsible Data Usage
5. Responsible Systems
6. Responsible Impact

All these domains are important, because data driven organisation are IT intensive organisations. We will discuss *how* data-driven organisation can take their responsibilities in each of those domains and can use scarce resources in a responsible way.

The first three domains are technical. The focus will be on optimised usage of scarce resources. The other three domains discuss the possibilities of applying data-driven architecture in such a way that it will positively contribute to the society.

RESPONSIBLE CODE

To clarify the importance of this framework for our topic is best understood when starting with the domain of coding.
It is also the topic where gain on energy consumption has the most impact. AI and Big Data are large consumers of power and commonly used development languages like Python and R are typical examples of very “*expensive*” languages (as far as CPU usage is considered). Python is an interpretative language and requires almost 80 times more CPU power than good old C. This means 80! times more energy compared to C (which is a compiled language) and 40 times more "*expensive"* compared to Java.
Using Python is unavoidable but when you use it, use it wisely.

So, a couple of considerations for developers:

1. Is python the best language to use or are there alternatives?
2. Is my code reviewed by a peer developer on efficiency?
3. Did I use best practice performance patterns?

With regards to code patterns please take a look of the work of Prof. Dr. Patrica Lago (University of Amsterdam) where she published several articles about software and sustainability (http://patricialago.nl).

RESPONSIBLE INFRASTRUCTURE

Code require IT infrastructure platforms to run. With regards to cloud we already mentioned that hybrid cloud requires network capabilities for transporting data. Distributing your bits and bytes across the network is not for free, it cost energy as well. In a data-driven organisation the amounts of data can be very, very large.

Besides the network, data is processed and especially for AI quite a lot of processing power is required. Intelligent usage of CPU and GPU workloads can save significantly on energy. In contrary to what you would expect in the name, the OpenPower consortium provides a technical solution that is not only open but also uses a different processor architecture, that is only 1/3th of energy compared to traditional processors in your data center.

And finally, storage. Data requires a lot of storage as well. Consider the usage of tape if possible, because that is a very sustainable medium. Also storage compression technology can be very useful to reduce the amount of storage required.

The second topic about sustainability is the usage of the infrastructure. We put services in place to increase platform utilisation, lower energy usage and increase the infrastructure life cycle. When building a new platform, these services could easily be integrated and decrease energy consumption by 70 to 90%. Or put the other way, you could use only 10-30% of the energy you are using now for your IT infrastructure. When you have an existing platform, please consider the implementation of these technologies during hardware life cycle management

RESPONSIBLE DATA CENTER

We put this at the latest because we have no idea if and how you can impact the choice of data center or (public) cloud provider. Request the datacenter provider for Power Usage Effectiveness (PUE) and Water Usage Effectiveness (WUE) as one of the selection criteria. Whether you can influence it or not that question you must ask! You have to make it clear that sustainability is important for you!

And now we are going looking to our data-driven organisation from a different angle: How can we positively use our project, our program, our organisation to give something to the society.

RESPONSIBLE DATA USAGE

In our data driven project we make sure that the data is secure. We respect the data of the client, citizen or other company. We clearly describe how we use it, and we only use it for the purpose for which the consent is given.

We know exactly where what data is and what it is. We have a strong governance in place.
We also know where the data is coming from and have lineage implemented.

RESPONSIBLE SYSTEMS

We are using the data in our systems and we are transparent in what we do. We implement strong ModelOps practices to avoid bias and drift and we save our models and the data.

RESPONSIBLE IMPACT

We build systems with the best intention for society. We make the world a little bit better with our system. We build a system that contains available homes and compare those with the list of people looking for one. We detect aggression on the street, and we make sure that police will be there in time to avoid escalation. How will the system that you build make a difference for society? Which of the SDG's is addressed by your project?

Besides these six domains there are also six principles:
Sustainability: Looking at those domains, we make sure we do things in a sustainable way. For these you can also refer to the UN SDGs or probably more specific ESG concerns.
Inclusivity: We build our system with respect for the users and in an ethical way.
Circularity: Can we reuse what we used before?
Openness: We are honest, also if it is a bit less green, better that than greenwashing!
Authenticity: Do we really take it seriously? Is it not only in our head, but is sustainability already in our heart, do you really care? Yes you, not the other one!
Accountability: You take ownership, yes you, what is your role in the data-driven organisation?

We would like to emphasis on one key stakeholder. And since we are architects, you can guess the role we take!

The (enterprise) architect is ***a key stakeholder*** when discussing sustainability. You must include sustainability as a non-functional requirement in your requirements, you must have at least one principle and one architecture decision that relates to sustainability. It's serious it is our planet, and it is not going into the right direction – yet.

We do understand it is not simple but when using this framework we became enthusiastic about the topic.

Implement the principle "Sustainability by Design".

FRAMEWORKS

We identified a number of different frameworks. Technology companies often define frameworks around specific technologies in order to identify the relation between people (users), processes and technology.

BUSOPS

This is not an official framework term, but we all know that enterprises have specific activities that make them unique. A supermarket has other activities than a clothing manufacturer or a travel organisation. We positioned this domain in the blueprint in order to link specific IT capabilities or services in context with the support of the primary (and secondary) business activities. By making the business activities unique, we can leverage generic support services.

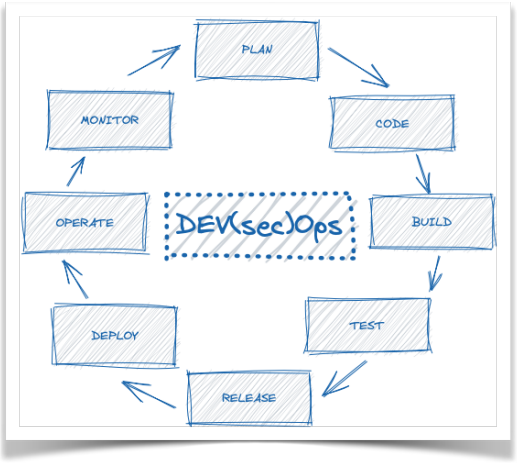
DEVOPS

We do not spend a lot of text on the topic of DevOps because there is so much good material available but there are a couple of things we would like to share with you in the context of a data-driven organisation.

We believe that the best approach for developing customer services in a ***resilient*** data- driven organisation is the DevOps approach, why?

We added the word resilient on purpose because you may think that resiliency is automatically build into your organisation as soon as you make it data-driven. To give you an example: in a customs department there is a constant focus on the detection of possible illegal material to be transported. They might find that ships from

trader A that travel via harbour X have an increased potential of illegal shipment. Because they now intercept the illegal loads, this situation does no longer apply so they have to remove this business rule and implement another and be vigilant on new routes being used

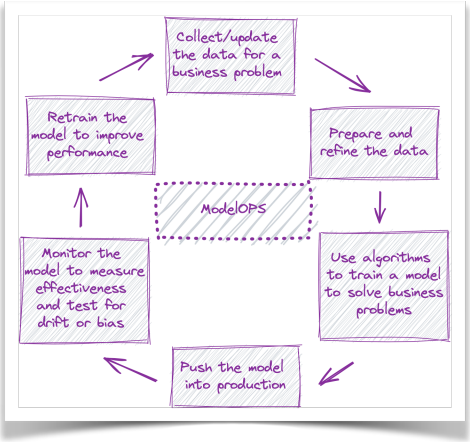


DevSecOps

So even though your organisation is data-driven, market disruptions, adaption of law, etc. require fast interaction and flexibility. The application that we develop must be delivered secure for various reasons, but one we would like to highlight here, is that the DevOps loop in itself can have many vulnerabilities due to the automated process for deployment. It introduces the risk that malicious code can be inserted when going into production.
As the loop in itself is straight forward: defining the minimum viable product, planning the sprints, developing the code and building the (micro)service, testing becomes important because that includes vulnerability tests as well. After testing you create the release and finally store that in the version management system, usually the start of a new branch. Then it will be deployed and from then on it can be used and will be monitored, ready to start a new cycle as soon as necessary (for monitoring please refer to AIOps).

MODELOPS

As indicated earlier in the document, "Fail fast, learn fast" is a fundamental success factor for a data-driven organisation. Developers have learned to apply agile principles when building applications. This is implemented in a DevOps process. The same principles should be applied for AI workloads to improve the adoption rate in operational environments and move to a higher rate of success from experimentation.



ModelOps

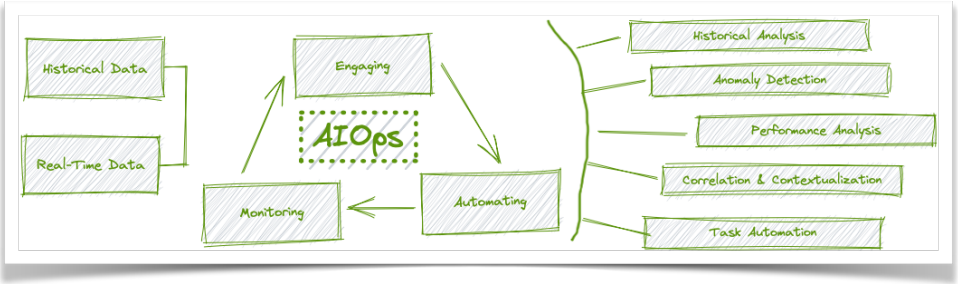
Furthermore adopting an approach enhances the trust and usability as it allows for audit-ability and transparency. By utilising ‘similar’ principles to DevOps, operationalisation can be improved and enhanced, leading to more successful production applications.

However, the work of a data- scientist is significantly different from that of an application developer. Creating models based on large amounts of data and working with AI and “unpredictable” outcomes is almost the opposite of developing code for explicitly defined functions that can be tested against specs and deliver consistent results.

The first questions to ask is what kind of business problem must be resolved and how that relates to the operation? What applications will be enhanced, what workflows will be touched? What can be agreed upon on quality measures for monitoring model performance?

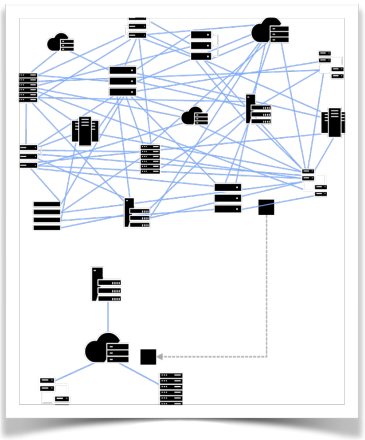
Practice learns that the interaction between business and data-scientist must be improved and needs to be more frequent to keep aligned on the quality. Developers utilise use cases to verify if the requirements have been understood. As soon as the business problem is clear, data scientists can collect the data. That is typically the next problem and that is why we have defined the data zone. That is where all the data- is available for the data scientist to work with. No endless discussions if data may be accessed, no endless delays to make data available. So a data scientist can spend time on his/her core task, the primary reason he or she comes out of bed every morning. Applying algorithms and training a model that provides predictive insights that will solve the business problems identified, and where applicable using prescriptive capabilities in our reference architecture. However this has an iterative nature as well, and can consist of multiple types of models, sub-models and different parameters.

ModelOps provides a systemic approach, a framework, that encompasses the end- to end pipeline of AI workloads, from training, testing and deployment of models. With the right technology in place it supports all pieces of an AI pipeline; datasets, data refinement, model definitions, trained models, applications and finally monitoring. Together this allows for lineage and as stated earlier, the ability to reproduce and audit a pipeline. Together this helps to apply the principle of failing fast and learning fast.



AIOps

AIOps is bringing AI and Operations together. Due to the increased complexity of an IT landscape and the development of a microservices architecture the number of services, interactions and dependencies exploded. It becomes impossible for humans to keep track of all these dependencies and understand the impact of a failing service on the business process. And let's be honest, also in the existing environment, documentation and Configuration Management Databases (CMDBs) are usually outdated and do not represent the actual situation. That is why the last years AI has been injected to recognise patterns and anomalies to predict error situations and prescribe solutions to resolve those incidents.



AIOps collects a lot of data by first building a topology of the infrastructure landscape. You could compare it with your CMDB where

relations between infrastructure and application are stored. AIOps observes the landscape and starts to build knowledge about the relationships within your environment. Then AIOps compares real-time data with historical data. The historical data contains both normal behaviour as well as incidents. When anomalies are detected it correlates these anomalies and can raise one or more incidents in your ticketing system. SRE teams use a feature as ChatOps to collaborate in resolving the incident. However, that is not the end of the cycle, because AIOps also performs prescriptive analysis. It can support you in problem resolution and even in problem prevention by executing scripts.

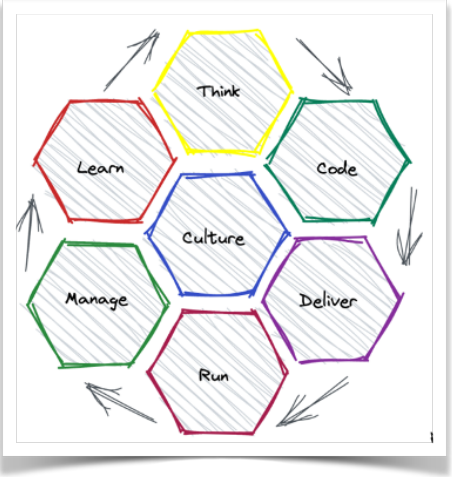
THE GARAGE

The garage is an approach that brings together the different processes in a pragmatic way. It is based on a multi-disciplinary team of business roles, designers, developers, data analysts, data-scientists, data- engineers, SRE's and architects who go from an idea to production in approximately 8 weeks.

> *"IBM Garage Is a Collection of Practices Woven Together as a Methodology To Bring all of IBM Together in Helping a Client Realize Business Value at the Fastest and Most Efficient Rate Possible."*
> *Dr. Mohamed El-Refai.*

The starting point is the "Think" phase. Based on the Design Thinking method, empathy maps are created, ideas are generated and a first scope is defined that provides value to the business: *a so called minimal viable product (MVP).*
During two week sprints the MVP is created with feedback-loops with the stakeholders. The MVP product will be installed on the application runtime and managed with the hybrid systems management services. As soon as the product is in production, users will provide feedback which can result into a new version of the product.

Even though this way of working is applied widely, it is not necessarily embedded very well in the entire organisation. Project managers are used to work with deliverables in timelines instead of two weeks sprints and backlogs, agile versus project management deliverables.



The garage method

Operations requires quality gates instead of SRE teams who achieve their service level agreements over time in an iterative fashion. From a security and compliance point of view mitigations need to be in place before going into production, instead of taking a risk based approach. According to a very well known agile coach:

"Agility should be understood at board level and they should agree ***and believe*** in it to make it successful".

More information on IBM Garage method can be found on the IBM Website[18] and in an article by Distinguished Engineer Dr. Mohamed El-Refai[19].

[18] https://www.ibm.com/garage/method/

[19] https://melrefai.medium.com/so-what-is-ibm-garage-564750a28da9

DATA MESH, THE ANSWER TO RESILIENCY?!

An important question is: “How to organise the availability of data for your business processes?”. There is a valuable approach handling this topic, this is called a ***Data Mesh***. The term Data Mesh is quite new and introduced by Zhamak Dehghani in 2019[20]. The idea is based on the theory of domain-driven design, a theory that is gaining popularity at the moment.

She highlights the difference between Systems of Record (SoR) and Systems of Insight (SoI). Even though she sees that these are coming together, she confirms the separation of these two. The Data Mesh is applicable to the Systems of Insight.

The data-driven paradigm must handle changes in processes, data proliferation and a diversity of use cases. That is why we added the term “***resilient***”. A Data Mesh can provide an answer to that requirement.

Traditional data warehouses were used to perform BI and reporting. The data warehouse was a centralised data warehouse and as soon as the data left the operational boundaries, it became a responsibility for the data warehouse team.

Then, when AI gained popularity, data analytics and data scientists entered the stage and they had different requirements. For example, raw operational data instead of aggregated and a continuous need for new data to feed the creativity to find meaningful insights; transformations in all kinds of ways,

20 https://martinfowler.com/articles/data-mesh-principles.html

aggregations, slices, new correlations etc. This requires a better understanding and availability of data, a reliable foundation that could be trusted, instead of several copies that exist and a lack of response times on data requests.

At the same time, the complexity of the data landscape increased, and even with data engineers as part of a team, they struggle to keep up with data definitions and providing access to what is needed. A centralised approach to managing data, given the large landscape, simply means that it is not realistic, leading to a lack of proper source and/or domain knowledge.

On top of this data quality can be an issue. Is the available data good enough for decision making? The quality aspects for decision making are much higher. Data you use for decision making should be accurate, secure, available, etc. Data used for decision making requires almost the same level of quality as the operational systems. The idea of a Data Mesh is to bring back the ownership of the data into the business domain to keep it manageable, by using a domain driven design. The technologies that we know from the data lake world are still applicable, albeit they are used differently, as the end-goal is supporting the domains. They become the owners of a piece of the data landscape and also have the responsibility of serving consumers from other domains within an organisation. Data is left where it resides if possible but made available in case of low overhead needs or to separate the operational environment from the analytical environment and needs.

THE IDEA OF A DATA MESH IS TO BRING BACK OWNERSHIP OF ANALYTICAL DATA TO THE BUSINESS DOMAIN

Four architecture principles have been defined by Dehghani:

1) *domain-oriented decentralised data ownership and architecture*
2) *data as a product*
3) *self-serve data infrastructure as a platform*
4) *federated computational governance.*

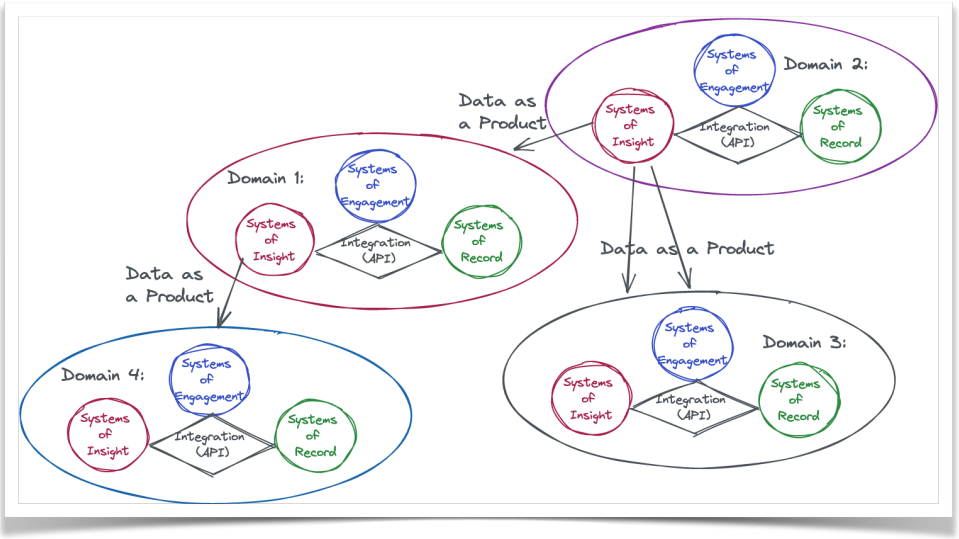
1) Domain-oriented decentralised data ownership and architecture

The idea is that data ownership is in the business domain. So instead of providing a copy of the data and as a result not knowing how and by whom that data will be used or if it is still accurate, the business domain becomes the owner, also for the analytical part. That means that within your organisation, you have a business domain responsible for its own systems: Systems of Record, Systems of Engagement and Systems of Insight. The business knowledge is concentrated in this domain.

2) Data as a product

So how do we deliver this data to the data scientists and data analysts? We provide data as products and we treat our data scientists and data analysts as users. This can be compared to the Systems of Engagement where we have clients and employees that use the Systems of Record. We deliver data in the format, with the performance, accuracy and availability that they require. We are aware of how they use it and we will keep them up to data on changes, additions etc. To realise data as a product this a combination of code, data, metadata and infrastructure.

But that is quite complex!



Data Mesh

That is why the third principle is being put in place:

3) Self-serve data infrastructure as a platform

It is too much for each domain to establish a complete platform for data provisioning, so best practice is to have one platform and provide that platform as a service: this is the data platform as a service. This also clearly separates the responsibilities for the data centre. Where traditionally data centres owned the Systems of Record, they should not own the data warehouses, but they should provide a platform as a service. For the Systems of Engagement that would be the (Container) Platform as a Service (SaaS or PaaS) and for the Systems of Insight that would be the Data Platform as a Service (DPaaS). A data fabric could be the implementation for this principle, as long as it supports the capability of supporting the domain oriented architecture, it should preferably provide a multi-tenancy organisational structure.

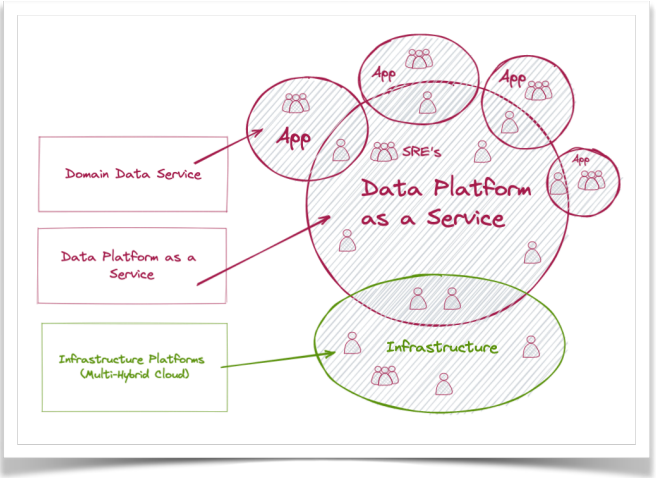
4) Federated computational governance

Finally you need a model to work together, both from an organisational perspective as well as from an infrastructure perspective. From an organisation perspective you would like to know how to provide what data to whom, what are the priorities, etc. From an infrastructure perspective you would like to agree upon platform standards, usage, cost allocation, etc. This is the fourth principle: federated computational governance. We think however that this must go beyond the computational aspect only. This is where the role of the enterprise architect is vital: understand the options and make the best choices.

To cut it short: Data Mesh could be the answer to Data Mess!

MANAGING YOUR DATA PLATFORM USING DEVOPS WITH SRE TEAMS

Now we have listed some Data Mesh principles we are taking it one step further. The question is: *"How do we establish an organisation which can provide a data platform as a service?"*. Even if you use a data platform in the cloud, it is still important to understand what capabilities and what support is provided, or better, what is ***not*** provided, because that is what you have to do yourself.
In an article[21] on uncontained.io, Red Hat provided an organisational model for managing a platform and its applications. In this article the focus is on an application platform, but the same could be applied to a data platform. Let's call it Data Platform as a Service.



Data platform

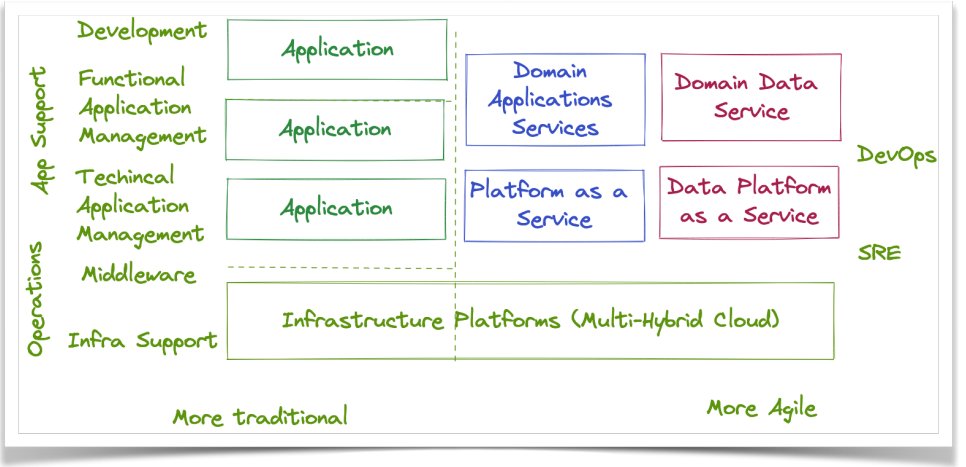
[21] http://uncontained.io/articles/openshift-and-the-org/

It is organised around DevOps principles. So you will find developers participating in the platform team as well as data engineers, curators and engineers.

Even though these different roles participate in this process, the key role is that of the Site Reliability Engineers (SRE's). SRE's are responsible for keeping the platform up and running and continuously improve platform quality. They depend on others to support them. These specialists could be more in the infrastructure area or in the data development area.

BRINGING THE DIFFERENT SYSTEMS TOGETHER

In the following picture you see on a high level how you could organise the different sorts of systems. As you can see, the Systems of Insight using a data platform can be compared with the Systems of Engagement using an application platform. You can follow that same model and you have to be aware that the Systems of Record may be (still) organised in a different way. Our advice: Don't solve it, work with it.



Different sorts of systems may have different organisational models.

DATA FABRIC

Over the years, there have been several approaches to work with the ever increasing amount of data. Nowadays data keeps growing faster and complexity is increasing. Coupled with internal organisation challenges such as a lack of control, a need for domain knowledge, an inability to understand data and therefore what data fits certain use cases, it has become a challenge to obtain the right information for decision making.

Over the years we have seen several shifts to address the challenges, starting from the invention of the data warehouse as an answer to performance issues, followed by Hadoop to solve scalability as it allowed for parallel processing, to Spark for in memory processing. Next, we moved to the concept of a Data Lake, which allowed for the grouping of data so valuable information could be extracted.

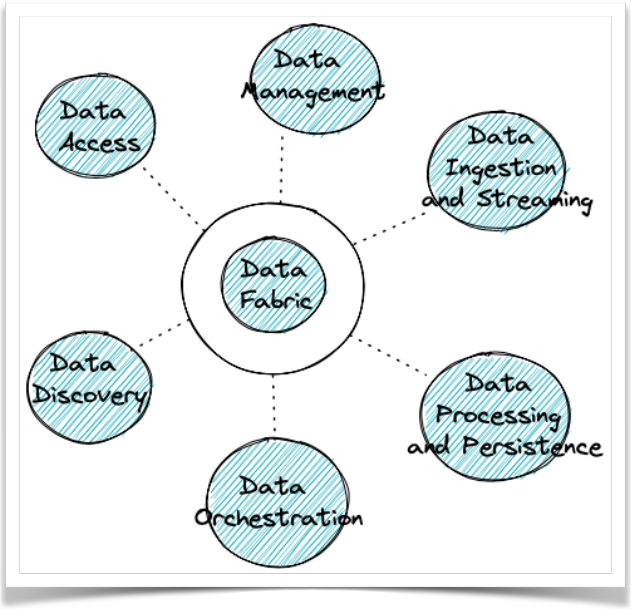
Technology innovation provided answers for these problems, but the challenges have kept growing due to the volume and complexity of data but also the hybrid/multi cloud world. But more importantly, coming back to the internal challenges, it did not solve these and organisations were still struggling with managing and providing access to the right data and users struggled whether data could be trusted at all by multiple copies, data quality issues etc. Research has shown that most data is not used within organisations[22] and that the majority is inhibited by data silos[23].

---

22 [1]Rethink Data: Put More of Your Business Data to Work – From Edge to Cloud (PDF, 8.3 MB, link resides outside ibm.com), Seagate Technology, July 2020

23 “The Total Economic Impact Of IBM Garage”, a commissioned study conducted by Forrester Consulting, October 2020 (link resides outside ibm.com)

A combination of technical capabilities available nowadays and lessons from the past years from organisational challenges has led to an approach, not one piece of technology, to support the above: ***a Data Fabric***.



Data Fabric

Gartner defines a Data Fabric as follows:

"a design concept that serves as an integrated layer (fabric) of data and connecting processes. A data fabric utilises continuous analytics over existing, discoverable and inferences metadata assets to support the design, deployment and utilisation of integrated and reusable data across all environments, including hybrid and multi-cloud platforms."[24]

[24] https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration

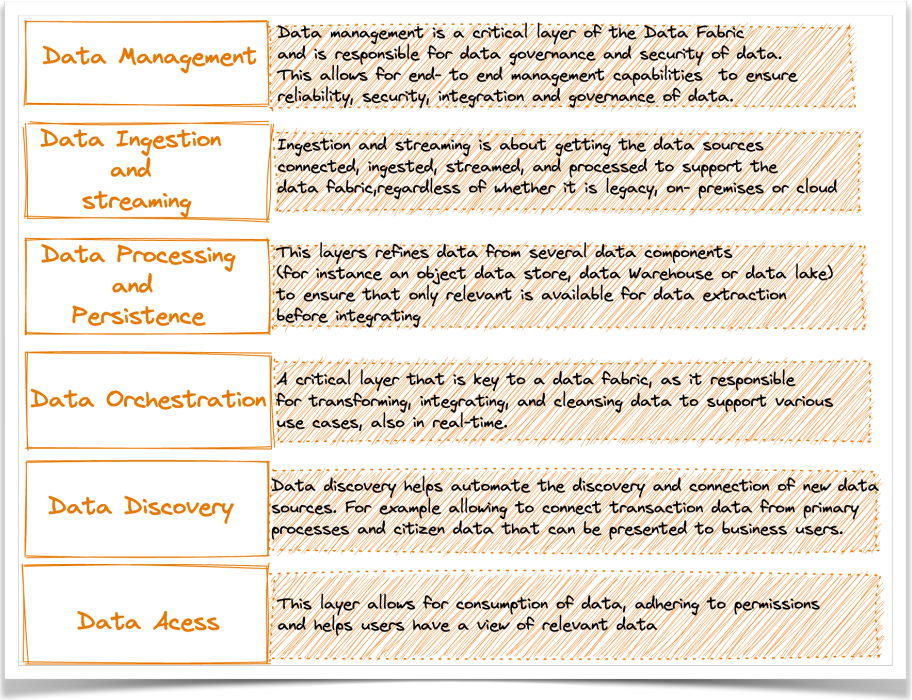
In other words, it can be described as a framework that weaves together various data pipelines and environments by using intelligence and automation with a semantic abstraction layer on top. This meant loosely coupled data to provide it where it is allowed and needed. Thus, bringing data across an enterprise together regardless of the location. In even simpler words, it connects the right data, at the right time, to the right people no matter the location by taking away the complexities of the underlying differing technologies.

To enable this, the semantic layer allows to unify the different silo's, embed governance and allow for autonomous search capabilities even for business users. The key factor here is metadata, and for it to work this means both technical as well as business metadata. This also allows for privacy and security measures. This is not to say that it is an easy feat. However, proper understanding of data and search capability from a self-service perspective provide the means to truly provide the capability to offer better services to citizens or increase competitiveness by better decisions.

Next to this semantic layer, data virtualisation is another key to the data fabric. It allows to connect to different sources and bringing those together without a traditional copy or move, by creating a virtual layer that is for instance used by analysts or data scientists. It does not mean it replaces ETL, each technology has its place depending on the use case and/or non-functional requirements.

Forrester has described a complete set of the key capabilities that are critical for any architecture and deployment[25]:

[25] Forrester: "Enterprise Data Fabric Enables DataOps"



Data Fabric key capabilities

In summary, it provides intelligent automation, leaves data where it lives, brings ML/AI to the data and provides access.

But what does this mean in relation to a data lake or a data mesh, that has been discussed in an earlier chapter? Quite simply: each has its place, a data lake could actually be part of a data fabric. In comparison a data fabric provides a semantic layer that provides a context of data (business terms) and allows data to be left where it resides.

In relation to a data mesh, this can be seen as a highly decentralised data architecture as part of a data fabric, in which domain driven design is key. They can actually benefit a lot from each other, as the service thinking by domain in a data mesh

provides the responsibility to own and manage their specific domain in such a way that it is accessible to users who are not part of the domain and they thus benefit from each other.

BRING IT FORWARD

By now, we have confronted you with many aspects that are required to build capabilities for a resilient data-driven organisation. You may agree with it, you may agree with most of it or you may totally disagree. In this last case, depending on your emotional status, you can throw his book away, burn it or give it to someone else and find your own way in developing an architecture for your data-driven organisation.

If the first two cases apply and we have touched some aspects to your interest, we invite you to read further and see how you can start to build your own blueprint for a data-driven enterprise architecture!

The best start is to select a business scenario, one that is hopefully not too complex, but use a scenario that will provide business value.
Secondly, define the use cases that implement this business scenario.

At this moment you can use the A0 size (841 x 1189) blueprint poster we supplied with this booklet. If you want to use a new poster (after you practiced with the original poster) you can download it from github.com:
https://github.com/datadrivenblueprint/downloads#readme

Here you will find the booklet as well as the poster in pdf format.

We are quite sure that you do not have an A0 printer at hand, please go to your repro department at the office or you can use print services available in the market.

Now you can identify the services which are required to implement your use cases. And this is the point where you can start using the garage method, but it is also the point where the problems start, because you will have many questions to be answered.

And these questions are crucial, as they set the standards for your future projects. Following is a list of questions that you probably will encounter:

1. I have selected a service and now I have to determine the implementation. Do I have this capability and can I reuse it in this context? If I have to acquire this capability, would that meet my future needs?
2. I have to develop a method, for example ModelOps, but I do not have the capabilities. Do I ask a vendor? Do I take shortcuts for this first trial?
3. I need data that I do not own, how do I get it?
4. What interface should I use? Best is to use file transfer, but I do have an API available now.
5. How do I integrate with the as-is situation of the current organisation, processes and technology?

GET TO WORK

You can now use the poster and write the use case on top of the poster. Give all identified services a unique number in the steps of the use case. By the end of this exercise you should have a complete overview of all services required for your use case. And please, please do not forget the underlying infrastructure services in order to make this use case enterprise ready. Remember you will have many stake holders, from business to strategy, architects and technical engineers. Make sure everybody recognises its own role in this blueprint.

ROOM FOR NOTES

Use case 1

.....

.....

.....

.....

Use case 2

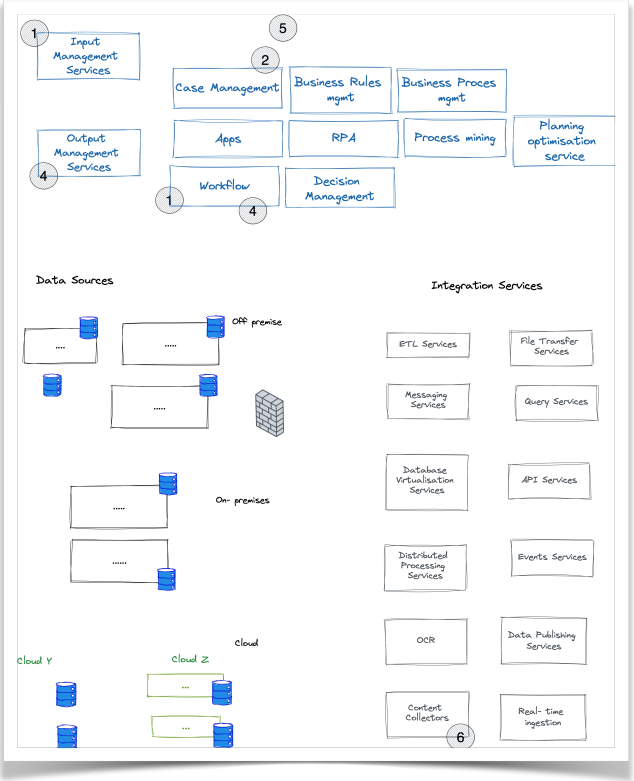
.....

.....

.....

.....

Here is an example how you can number the use case in your poster. In the appendix an example of a use case on the blueprint can be found.



Use case example

Now that you have this overview you can start thinking about the solution. Shortcuts are inevitable in this process. If you wait until you have developed everything you will achieve nothing. Think

big, act small really applies here. Sometimes the cloud may help you because complete platforms are available in the cloud.

If you are a data centre manager or architect, there is an opportunity for you: provide a data platform for your users. If you can add a business automation and application development and integration platform you provide a comprehensive set of services that is more than sufficient to build a resilient, data driven organisation. These three platforms are available in the market and when you deliver those, the developers can focus on the business requirements and the adaption of the organisation.

And if it all fails, at least we have played a tiny role to get your organisation to align more closely and get to know each other, while hopefully having enjoyed some good old cups of coffee.

EPILOGUE

There is so much to say about building a data driven organisation, you can fill your enterprise content management system with hundreds of books and thousands of articles. The appliance of AI in a data-driven organisation is unavoidable. We observe that many companies are following trends towards cloud adoption and/or microservices architecture based on their strategy for “Systems of Engagement” without realising the impact for the “Systems of Insight”: **they are not the same!**

A data-driven organisation in our perspective should dare to challenge trends and develop their own. Questions you should ask yourself:

- What is the cost of putting large amounts of data in the public cloud? What is the cost of using or migrating the data out of the cloud?
- Is agile the way of working for trustworthy AI?
- How do you explain ?“sustainability” if 25% of your servers are performing analytics
- Who takes the decisions? Your application developers, your data scientists or your governance organisation? Site Reliability Engineers, in name or in practice?
- Trust? Or Trust!
- Is data driven a choice or an obligation?
- What if I could perform analytics on my operational data?
- Do we want to be transparent or do we have to be?
- Until what extend do we want to be dependent on our “Systems of Insight”? What does that mean for the qualitative aspects? Are we prepared to pay the price?

There is not a right answer, but these are good questions to ask. And by now, at the end of this book we started to get to know each other a little bit, we know that you have other (and better) questions as well! That is good, because that means we have achieved our goal: provide you with a holistic overview of the impact of building a data driven organisation, think about the impact of technology and discuss the services that are required.

We enjoyed it, thanks for your interest!

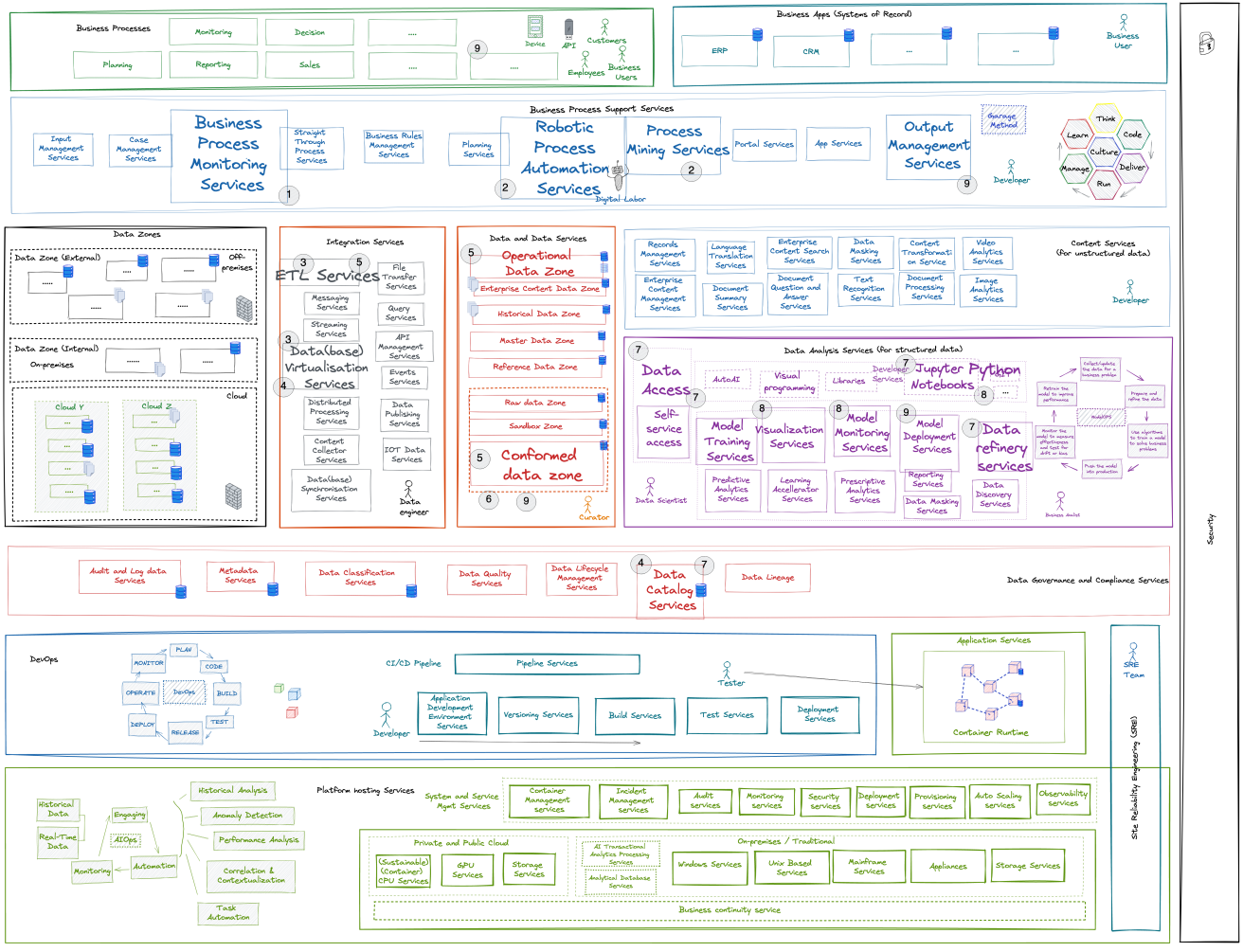
Laila, Jan, Ronald

APPENDIX

EXAMPLE

Describe the use case or scenario with steps:
Now put these steps to the services that are required, and link those, we have not included the infrastructure nor the security aspects, simply a matter of simplicity:

1. Process owner Marie receives a notification that the waiting times for requests are increasing which needs her attention.
2. She runs a simulation, based on the expected numbers ,and finds out that the number will only increase prompting her to take immediate action as her department will not be able to handle the number of requested within the required timelines. She reaches out to the privacy officer to request approval for the usage of a machine learning model of machine learning to implement a scoring function which will separate out special cases whereas regular non- risky cases can be processed directly. David, a data scientist is assigned to work on the project.
3. David, the data scientist cannot wait to start! He wants to build a solution that separates the low risk request from request that require personal attention. These low risk request can then be handled automatically!
4. He goes to the catalog and self-service portal to request the required data.
5. Meanwhile Marco, the data engineer receives a request to provide the data for David in the operational environment.
6. Marco makes the data available in the sandbox zone, masking private data and accessible for David and Marie.
7. David prepares and transforms the data before applying Machine Learning.
8. After a couple of iterations the model seems to return the proper results and he shares the outcome with Marie. They review the model as well as the outcome.
9. The model will be approved and operations will implement the model for production.


Business Processes
Monitoring
Decision
....
Planning
Reporting
Sales
....
....
9
Device
API
Customers
Employees
Business Users
Business Apps (Systems of Record)
ERP
CRM
...
...
Business User
Business Process Support Services
Input Management Services
Case Management Services
Business Process Monitoring Services
1
Straight Through Process Services
Business Rules Management Services
Planning Services
Robotic Process Automation Services
2
Digital Labor
Process Mining Services
2
Portal Services
App Services
Output Management Services
9
Orange Method
Developer
Think
Code
Deliver
Run
Manage
Learn
Culture
Data Zones
Data Zone (External)
Off-premises
Data Zone (Internal)
On-premises
Cloud
Cloud Y
Cloud Z
Integration Services
3
ETL Services
5
File Transfer Services
Messaging Services
Query Services
Streaming Services
3
Data(base) Virtualisation Services
API Management Services
4
Events Services
Distributed Processing Services
Data Publishing Services
Content Collector Services
IOT Data Services
Data(base) Synchronisation Services
Data engineer
Data and Data Services
5
Operational Data Zone
Enterprise Content Data Zone
Historical Data Zone
Master Data Zone
Reference Data Zone
Raw data Zone
Sandbox Zone
5
Conformed data zone
6
9
Curator
Content Services (for unstructured data)
Records Management Services
Language Translation Services
Enterprise Content Search Services
Data Masking Services
Content Transformation Service
Video Analytics Services
Enterprise Content Management Services
Document Summary Services
Document Question and Answer Services
Text Recognition Services
Document Processing Services
Image Analytics Services
Developer
Data Analysis Services (for structured data)
7
Data Access
AutoAI
Visual programming
Libraries
Developer Services
7
Jupyter Python Notebooks
8
...
7
Self-service access
8
Model Training Services
Visualization Services
8
Model Monitoring Services
9
Model Deployment Services
7
Data refinery services
Predictive Analytics Services
Learning Accelerator Services
Prescriptive Analytics Services
Reporting Services
Data Masking Services
Data Discovery Services
Data Scientist
Business Analyst
Audit and Log data Services
Metadata Services
Data Classification Services
Data Quality Services
Data Lifecycle Management Services
4
Data Catalog Services
7
Data Lineage
Data Governance and Compliance Services
DevOps
PLAN
CODE
BUILD
TEST
RELEASE
DEPLOY
OPERATE
MONITOR
DevOps
CI/CD Pipeline
Pipeline Services
Tester
Developer
Application Development Environment Services
Versioning Services
Build Services
Test Services
Deployment Services
Application Services
Container Runtime
SRE Team
Site Reliability Engineering (SRE)
Historical Data
Real-Time Data
Engaging
AIOps
Monitoring
Automation
Historical Analysis
Anomaly Detection
Performance Analysis
Correlation & Contextualization
Task Automation
Platform hosting Services
System and Service Mgmt Services
Container Management services
Incident Management services
Audit services
Monitoring services
Security services
Deployment services
Provisioning services
Auto Scaling services
Observability services
Private and Public Cloud
(Sustainable) (Container) CPU Services
GPU Services
Storage Services
AI Transactional Analytics Processing Services
Analytical Database Services
On-premises / Traditional
Windows Services
Unix Based Services
Mainframe Services
Appliances
Storage Services
Business continuity service
Security

GLOSSARY

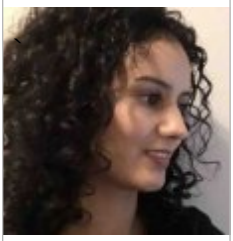
A0	Basic paperformat	The A-Format is a paper format used in countries that apply the metrics standard. A0 is the basic form of exactly 1m2. Width and Length are in the ratio square root of 2. If you fold this paper 5 times you have an A5, that is the format of this book.
AI	Artificial Intelligence	Computer Intelligence that goes beyond the standard IF THEN ELSE operations of a computer
AIX	Advanced Interactive eXecutive	Series of Unix operating system, developed by IBM. They are mostly used for enterprise servers.
AKS	Azure Kubernetes Services	Microsoft's hosted implementation of Kubernetes
API	A Programming Interface	A way applications can communicate to each other, a necessity so to say and allows companies to open up themselves to others.
CDN	Content Delivery Network	A network of servers that live a nomadic life while working together to provide high availability and performance
CDO	Chief Data Officer	The person that is responsible for everything related to data. A relative new role in organisations, became more popular with the introduction of GDPR, if you do not have one, appoint one.
CI/CD	Continuous Integration / Continuous Delivery	Practices of continuous integration and continuous delivery or deployment.
CIA	Confidentiality, Integrity and Availability	A model that guides policies for information security within an organisation, not one we would like/are allowed to joke about.
CIO	Chief Information Officer	The boss of IT, probably your boss.
CISC	Complex Instruction Set Computer	Minimises the amount of instructions per program. The opposite of RISC.
CMDB	Configuration Management Database	A database to keep track of the hardware and software implemented and their relationships.
COTS	Commercial off the Shelf	A product that you buy from a vendor and customize to make it usable in your own organization.
CPU	Central Processing Unit	The brain of the computer.

CSMO	Cloud Service Management and Operations	Not the popular Chief Social Media Officer, but a framework on how to run applications in the cloud.
CTO	Chief Technology Officer	The boss of technology strategy, including IT.
DBA	Database Administrator	The one responsible for making sure databases are designed, maintained and run as required. Oh, and makes them accessible...
ECM	Enterprise Content Management	Process of managing the lifecycle of content (from documents to video) and the accompanying technology.
ERP	Enterprise Resource Planning	Software used for day to day operations, such as transactions, procurement or accounting.
ETL	Extract, Transform, Load	A way to combine data from multiple sources and making It available in a single store.
EU	European Union	We hope that makes sense!
GDPR	Global Data Protection Regulation	Regulation that specifies how a company should handle personal data.
GIT	Whatever rocks your boat, but it could be referring to Torvald (the creator) himself	Open source software for distributed version control, in essence a stupid content tracker (so they said themselves).
GKE	Google Kubernetes Engine	Google's version of Kubernetes, as a managed service.
GPU	Graphical Processing Unit	The second brain of the computer. Meant to address the display, but the way this processor works, it is very useful to do complex calculations for AI as well.
HR	Human Resources	Basically the workforce of an organisation, although we refer to the management of the workforce.
I/O	Input and Output	The bits and bytes transferred between systems.
IaaS	Infrastructure as a Service	Providing or using Infrastructure as a service. No longer necessary to acquire your own hardware but just use it on demand.
IBAN	International Bank Account number	Internationally agreed system of identifying bank accounts across national borders.
IBM	International Business Machines	A great company (at least that is our opinion) which strategy is to develop AI and cloud solutions.
IT	Infrastructure Technology	With Infrastructure in this context we mean computer infrastructure like servers, storage, network, applications, operating systems, etc.
ITIL	Information technology infrastructure library	A set of best practices for IT service management and IT management, to deliver the best services.

LDAP	Lightweight Directory Access Protocol	An application protocol to allow access to resources in a computer network.
ML	Machine Learning	A discipline of AI that aims to teach computers how to learn and act without being explicitly programmed.
MVP	Minimum Viable Product	An initial version of a product that can be used in a business operation, and thus shared with customers. Which will iteratively be updated with the gathered feedback.
OCR	Optical Character Recognition	A technology that recognises text within a digital image. It is commonly used to recognise text in scanned documents and images.
OWASP	Open Web Application Security Project	International non-profit organisation dedicated to web application security. Its goal is helping website owners and security experts protect web applications from cyber attacks.
PaaS	Platform as a Service	Providing or using an entire platform as a service. This includes infrastructure, a bundle of related software and it's management services. Examples are a development platform service like ARO (Azure Red Hat Openshift) or it could also be a data platform.
PDF	Portable Document Format	We are sure you are familiar with the format.
PS	Post Script	It comes from Latin postscriptum, you write that at the end if you forgot something important and would like to highlight that. You can also ad PPS if you forgot another thing. But what should we say about a person that starts to forget so many things...
REST	Representational State Transfer	An architecture that provides a way to bring together decoupled components on the internet.
RISC	Reduced Instruction Set Computer	Reduce the amount of cycles per instruction. The opposite of CISC.
RPA	Robotic Process Automation	Automation technology that provides a digital robot who can take over repetitive manual tasks from us that are driving us nuts on a daily basis.
RPO	Recovery Point Objective	The maximum acceptable amount of data loss.
RTO	Recovery Time Objective	The maximum amount of time required to restore your system. Usually this includes also the potential administrative processes.
SaaS	Software as a Service	Providing and using software as a service.
SAP	Systems, Applications, Products	A German software company.

SIEM	Security Information and Event Management	Technology that supports threat detection, compliance and security incident management by aggregating and analysing information from different resources.
SoE	Systems of Engagement	Systems (applications) that engage with clients, employees and partners.
SoI	Systems of Insight	Systems that translate the data into insights. This is what this booklet is all about.
SoR	Systems of Record	A group of systems that holds the core records of the company. These could be customers, citizens, products, transactions, etc.
SQL	Structured Query Language	Good old SQL is a programming language to manage relational databases
SRE	Site Reliability Engineer	A role for a person that manages a computer platform rather than one system. This concept was introduced by Google to avoid linear scaling of support teams.

ABOUT THE AUTHORS

 Laila Fettah	Laila Fettah is an Open Group and IBM certified IT Specialist, who started in the world of statistics and has spent the first ten years of her career in finding what matters in data-driven projects. While working on efficiency projects in industrial sector, supporting marketing projects or unraveling value from unstructured documents, bridging technology and business has been her passion. This has driven her to her current architect role to bring those two together more holistically, with a nod to her past: “make it simple, but significant”.
 Ronald Meijer	Ronald Meijer is a certified IT Architect working at IBM and worked for several large organisations like Rabobank, ABN/AMRO, Shell, KLM, ING, Delta Lloyd, Aegon. He is a passionate teacher on TOGAF(R), Architecture Thinking and Microservices Modelling. Experiences range from business architecture, data architecture, application architecture and infrastructure architecture. He has a degree in computer science and speaker at public conferences. “A data-driven organisation is so interesting because it brings together AI based technologies, different processes like AIOps and ModelOps and culture, I like to unravel that complexity”

 Jan Schravesande	Jan Schravesande is an IBM and Open Group certified Enterprise Architect and has experience in different business domains such as Insurance, Supply Chain, Energy and Utilities and Government. The last ten years he has been a technical advisor for Dutch Government organisations. “I find it fascinating that organisations use their data to predict outcomes and to advise on actions in order to stay ahead of competition or to excel in the best service. Data has become so much more than an asset. It has become the keeper of all knowledge ready to reveal its secrets when the right key is used.....”

For reactions and / or inquiries please contact one of the authors (or all of them), or reach out for a good chat over a cup of coffee:

Go to for downloading this booklet in pdf format or a new version of the blueprint poster:
https://github.com/DataDrivenBlueprint/Downloads#readme

Or use the QR code:

Laila Fettah: lailafettah@nl.ibm.com
Ronald Meijer: meijerr@nl.ibm.com
Jan Schravesande: schravesande@nl.ibm.com

Dear reader,

This booklet is for those who want to know more about how to build capabilities for a data-driven organisation.
It is written by IBM practitioners who are very interested and passionate about this subject.

"*This topic triggered our curiosity and one year later it looks like it is almost the only thing that drives our activities, mindset, discussions and our idea's. It certainly keeps us occupied!*"

We turned this positive energy into something that we can share with other people who are interested in this subject.
The goal of this booklet is to help organisations making the first steps in becoming a data-driven organisation. Business decisions based on predictions and even prescriptive analysis in order to become or stay competitive in a market where almost everything is commoditised. That is the dream of many organisations!

How do you bridge the complex world of people working with data and the complex world of multi hybrid cloud operations and everything in between? We have no answer to that but we can help you to make the first steps and avoid making bad decisions you may regret later.

We are quite sure that we get feedback from the community out there and that feedback will be the basis for new releases of this booklet.
In the mean time we hope you read it with the same pleasure as us writing it.

Amsterdam,
December, 2022