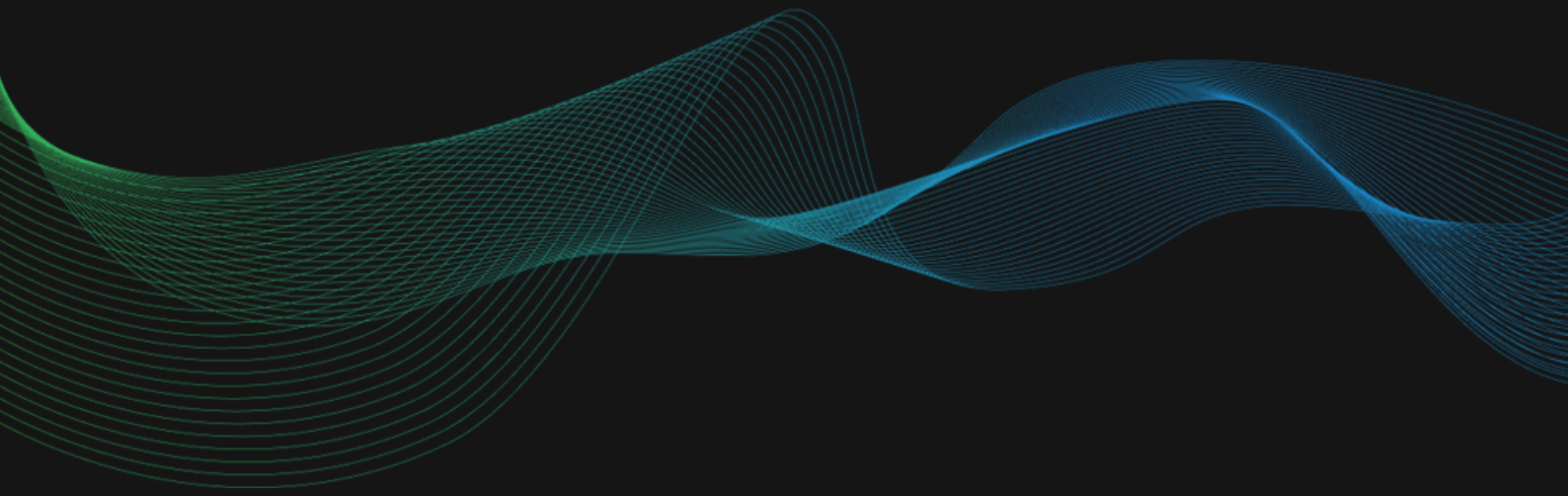
(WHITEPAPER)



# DATA PLATFORMS FOR A DATA-CENTRIC FORCE

(RAFT)

Together, we'll make waves

# TABLE OF CONTENTS

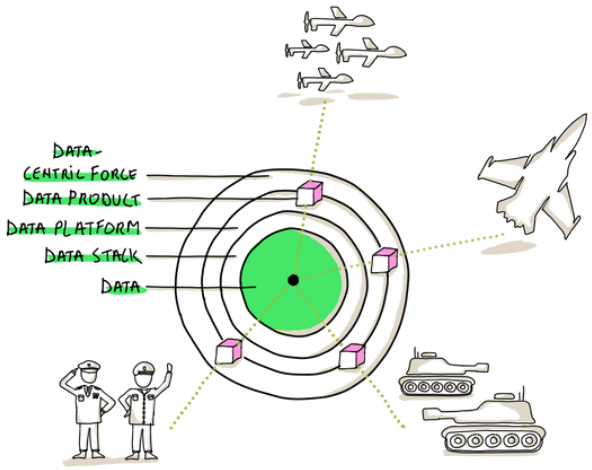
# INTRODUCTION

Modern warfare demands robust data platforms that enable warfighters to effectively use data for mission success. These platforms are integral to the broader strategy of transforming the military into a data-centric force

The global defense ecosystem is in the midst of a Revolution in Military Affairs (RMA), marked by the transition away from traditional armaments as the centerpiece of warfare to the strategic use of data. This change hinges not just on the existence of data but on the vast volumes now accessible and the computational advancements that allow for its strategic deployment. The contemporary battlefield demands warriors adept in the nuances of data, its management, and protection, much like soldiers and their weapons in the past.

Modern warfare requires soldiers to be skilled in understanding data's strengths, weaknesses, and potential applications, echoing the deep bond between a Marine and their rifle, as highlighted in the Rifleman's Creed. As warfare evolves, the mastery of data utilization will likely determine the victor in future conflicts, underscoring the need for data-driven strategies and the profound integration of data awareness in military training and operations.

This whitepaper outlines the transition of the military into a **data-centric force** through the development and use of **data products**—specifically tailored data subsets designed for decision-making. It explores how these data products are supported by **data platforms** equipped to handle large volumes of data efficiently. A robust software architecture, referred to as the **data stack**, forms the foundation of these platforms, ensuring they are capable of addressing the distinct challenges faced during military missions.

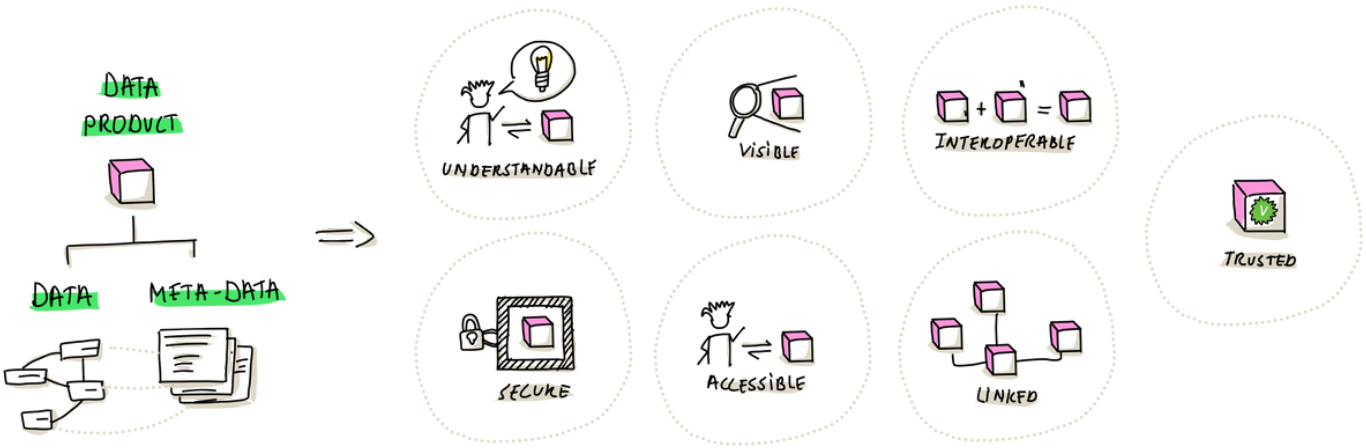


# {DATA PRODUCTS}

Data products are essential in building a data-centric force. By productizing data, they create a series of trusted and understandable sources that facilitate data-driven decision-making with confidence.

Data products are crucial for decision-making, consisting of meticulously curated data subsets. These are not merely aggregations of data but are precisely crafted to include vital information and metadata, transforming raw data into actionable insights. The design of these products focuses on facilitating informed decisions by presenting essential data in a clear and comprehensible format.

These data products are **trusted** and **secured** sources, designed to be highly **discoverable**, **accessible**, and **understandable**, complete with extensive **metadata** and documentation to ensure they are **user-friendly**. They are indispensable tools for various users, including warfighters, military leaders, and strategists, aiding them in navigating complex decision-making scenarios.

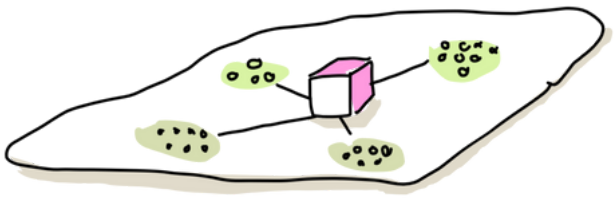


# DATA PRODUCT EXAMPLES

Artifacts like PowerPoint briefings are not considered as data products. This section will introduce three example data products to clarify this concept within the context of a data-centric force.

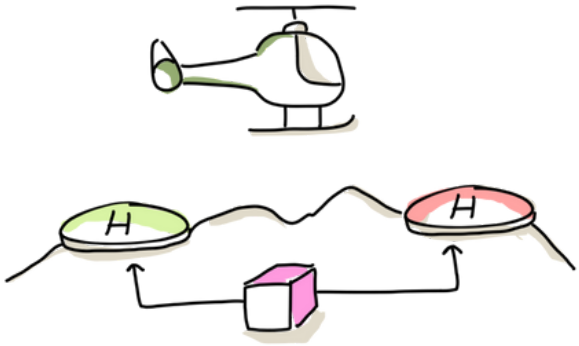
## Command and Control

This data product encompasses crucial information used for strategic planning and ongoing assessment of the current military situation. It aggregates various data points, including troop movements, enemy positions, and real-time communications, which are essential for making timely and informed command decisions. By synthesizing this information, command centers can better coordinate their responses and adapt to changing battlefield conditions, enhancing the effectiveness of military operations.



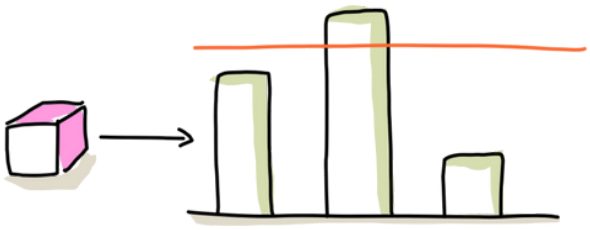
## Movement and Maneuver

The Movement and Maneuver data product provides detailed and precise data to identify optimal helicopter landing zones within a theater of operations. This includes geographical data, terrain analysis, weather conditions, and threat assessments to ensure safe and strategic placement of aerial assets. Utilizing advanced algorithms and real-time updates, this data product helps in planning routes and maneuvers that minimize risk and maximize tactical advantage for ground and air units.




## Sustainment

The Sustainment data product is designed to streamline logistical operations by providing comprehensive data on the necessary supplies for various military branches or specific operational sequences. It integrates inventory levels, consumption rates, and resupply timelines to ensure that all units are adequately provisioned without overstocking or shortages. This data is critical for maintaining operational readiness and efficiency, particularly during extended deployments or in remote locations where logistical challenges are most acute.



# DATA PRODUCT STRUCTURE

Data products are composed of datasets paired with various types of metadata. This structure allows data products to make data **Visible**, **Accessible**, **Understandable**, **Linked**, **Trusted**, **Interoperable**, and **Secure**, thereby enhancing their utility and reliability for decision-making processes.

### BASIC INFORMATION

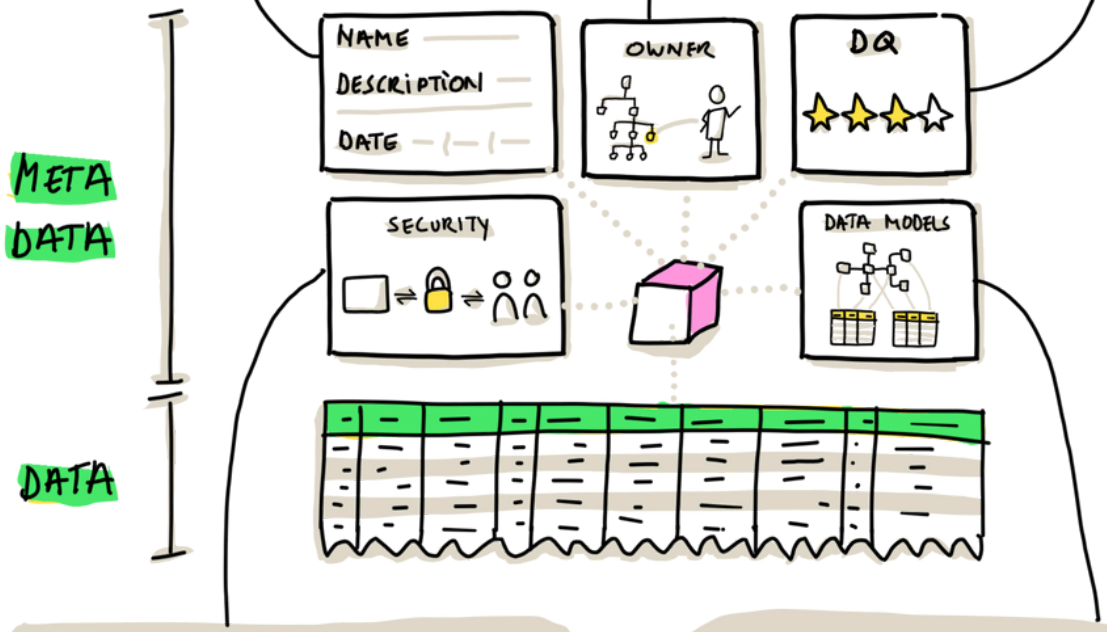
Including the name, description, and creation date of the product.

### DATA PRODUCT OWNERSHIP

Details about the owner of the product, their organizational role and contact information

### DATA QUALITY METRICS

Metrics that assess the data's accuracy and reliability.




### SECURITY SPECIFICATIONS

Information on the data's sensitivity, usage policies, and retention policies, often described as ‘time to live’.

### DATA MODELS

Describing how business entities (the logical model) align with the data table structures (the physical data model).

# DATA FLOW

This section outlines the typical journey of data as it transitions from raw data at its source to a fully formed data product. The process involves several key steps, each crucial for refining the data into a usable and valuable resource

Data products result from a data flow process where raw data is transformed into curated datasets, and necessary metadata is gathered and integrated with the dataset to create a data product. We consider a scenario with various data sources such as UAVs, tanks, and weather stations.

## **Step 1:** Raw Data

The data sources in our scenario continuously supply raw data, including location, resource levels, and environmental conditions.

## **Step 2:** Individual Products

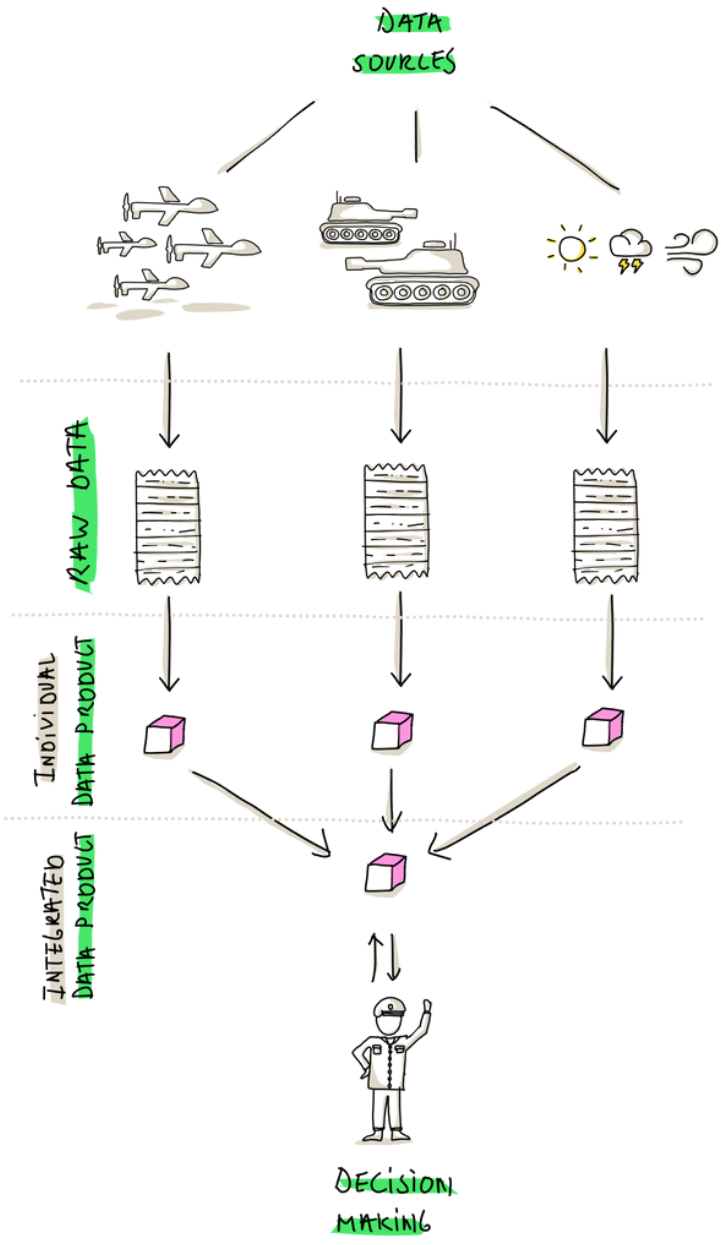
The raw data from each source is processed into individual data products.

## **Step 3:** Integrated Products

The interoperability of these initial data products facilitates the creation of a new, integrated data product.

## **Step 4:** Decision Making

This comprehensive data product aids in decision-making by providing a clear overview of both own and allied unit positions, which enhances coordination and strategic planning.



# CONCLUSIONS

In this chapter, we have introduced data products as fundamental components for establishing a data-centric force. We began by defining what constitutes a data product and provided several examples within a warfighting context to illustrate their practical applications. We then delved into the structure of data products, detailing how data transitions from its raw form at the source to a fully developed data product.

The chapters that follow will explore various data product management strategies, such as centralized and data mesh approaches, which are prevalent in data-centric forces. Subsequently, we will examine the data platforms and underlying data stacks necessary to support a data product-driven strategy within a data-centric force.

# {DATA-CENTRIC FORCE}

Large organizations, like the U.S. army, are divided into domains. A data-centric force adopts a data strategy that ranges from centralized to decentralized data mesh approaches to manage data products across these domains.

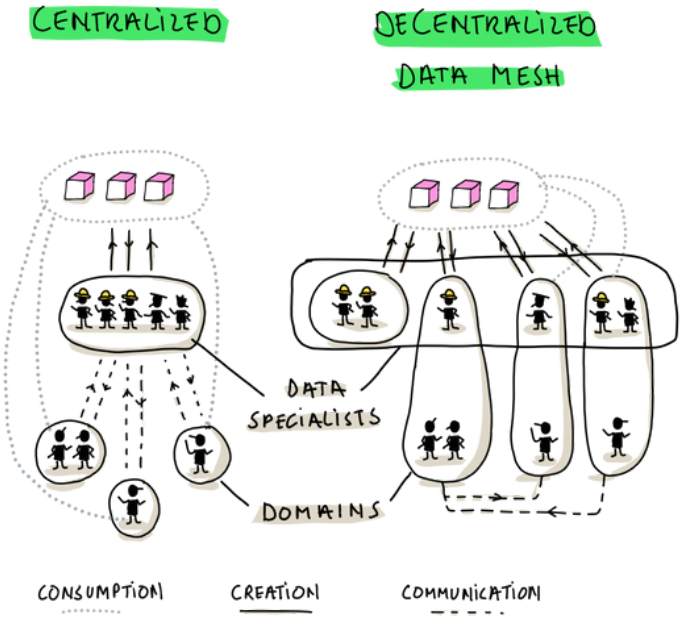
## Domains

Larger organizations are often segmented into various business domains that align with distinct organizational units, each centered around a specific business purpose. For instance, a bank might be divided into domains such as investment, insurance, and retail, with further subdivisions like life, non-life, or loans within these domains.

Similarly, the U.S. army organizes itself into domains based on warfighting functions, with personnel trained for specific roles within these areas. For example, the Warfighter Mission domain includes subdomains such as Mission Command, Fires, Intelligence, Movement and Maneuver, Protection, and Sustainment, each addressing different aspects of military operations.

## Data Strategies

A data-centric force can facilitate the management of data products within its domains using centralized, decentralized, or hybrid strategy. The centralized method centralizes all data-related tasks within a single team, whereas the decentralized, or data mesh approach, allocates the responsibility for creating and maintaining data products to the individual business domains. This means each domain operates with its own specialists rather than relying on a central data team.



# DATA STRATEGIES

Data-centric forces will manage data products, govern data, and align technology with domain expertise differently, depending on whether they adopt a centralized or decentralized data strategy.

| | **Centralized Strategy** | **Decentralized '*Data Mesh*' Strategy** |
|---|---|---|
| **Data Product Management** | A central team handles all data product creation and maintenance. | Each domain has the autonomy and expertise to develop and manage its own data products. |
| **Domain Expertise** | The central data team's understanding of specific business domains varies, necessitating significant effort to relay and clarify business requirements from each domain. | Domain expertise is closely integrated with data expertise within each business domain, fostering efficient communication and quicker time to market for data products. |
| **Data Governance** | A single central data governance team assumes full responsibility for data governance. | A federated data governance model grants data governance responsibilities to each business domain, with a data governance board ensuring alignment across all domains. |
| **Data Platform Management** | Regardless of the approach, the building, maintaining, and extending of the data platform software are managed by a data platform team. This platform serves as the foundational infrastructure for creating data products within both centralized and decentralized approaches. | |

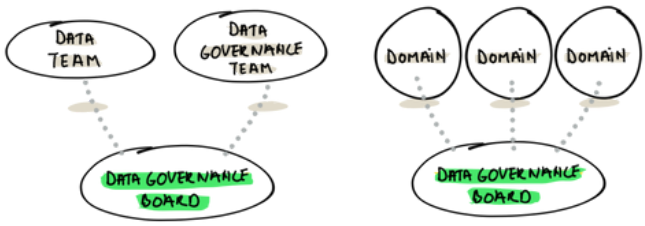
*Table 1 - The key differences between centralized and decentralized strategies.*

# DATA GOVERNANCE

Data governance within a data-centric force involves establishing policies, procedures, and standards to manage, secure, and utilize data effectively across all organizational domains.

## Data Governance Board

In a data-centric force, establishing a robust data governance framework is critical for defining roles, responsibilities, ownership, and accountability, maintaining high data quality, and accurately cataloging data by location and sensitivity. Such practices enhance data management and support organizational goals.

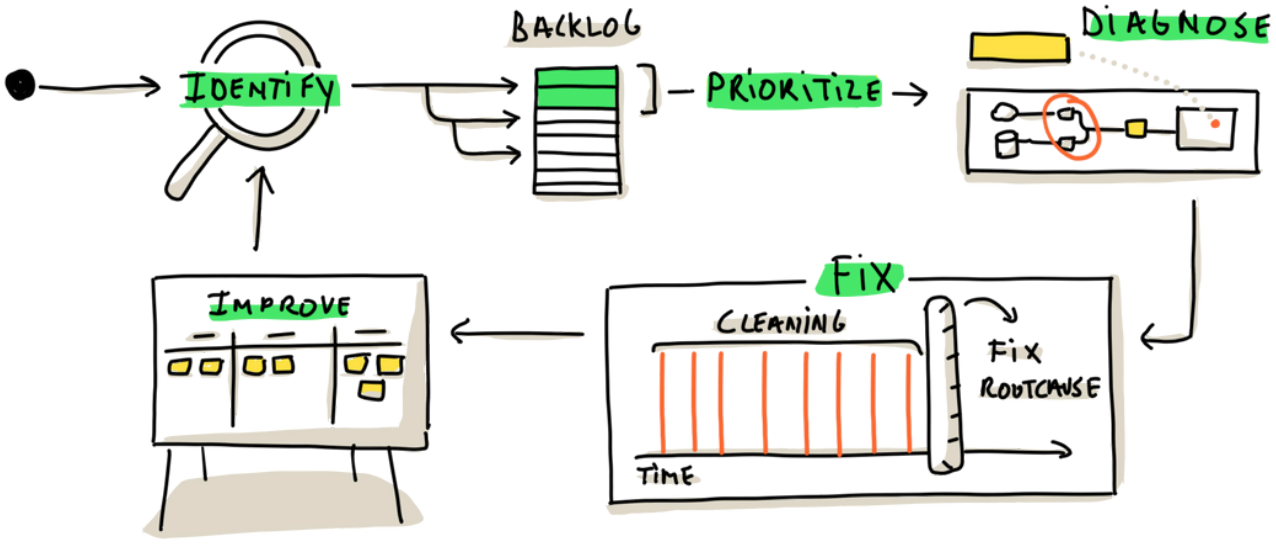



The "data governance board" plays a key role in ensuring consistency across governance models. In a data mesh, it harmonizes individual domain practices within a federated approach to align with broader organizational policies. In a centralized approach, the board collaborates with a central data governance team to set and enforce data standards, streamlining data management and policy adherence.

## Data Quality

A vital component of data governance is managing data quality (DQ), which focuses on ensuring the accuracy, completeness, and reliability of data. The management process for data quality unfolds through several crucial stages:

1. **Identifying** anomalies and violations against data quality standards.
2. **Prioritizing** data quality issues in a backlog.
3. **Diagnosing** the root causes of these data quality issues.
4. **Fixing** the root cause and cleaning the data.
5. Continuously **Improving** the data quality process.



# CONCLUSIONS

In this chapter, we explored the nuances of data product management within a data-centric force, considering both centralized and decentralized data mesh strategies. While the focus was primarily on these two approaches, it is important to note that hybrid approaches also exist, which blend elements from both centralized and decentralized strategies. Additionally, we delved into crucial aspects of data governance and data quality, highlighting their significance across different approaches.

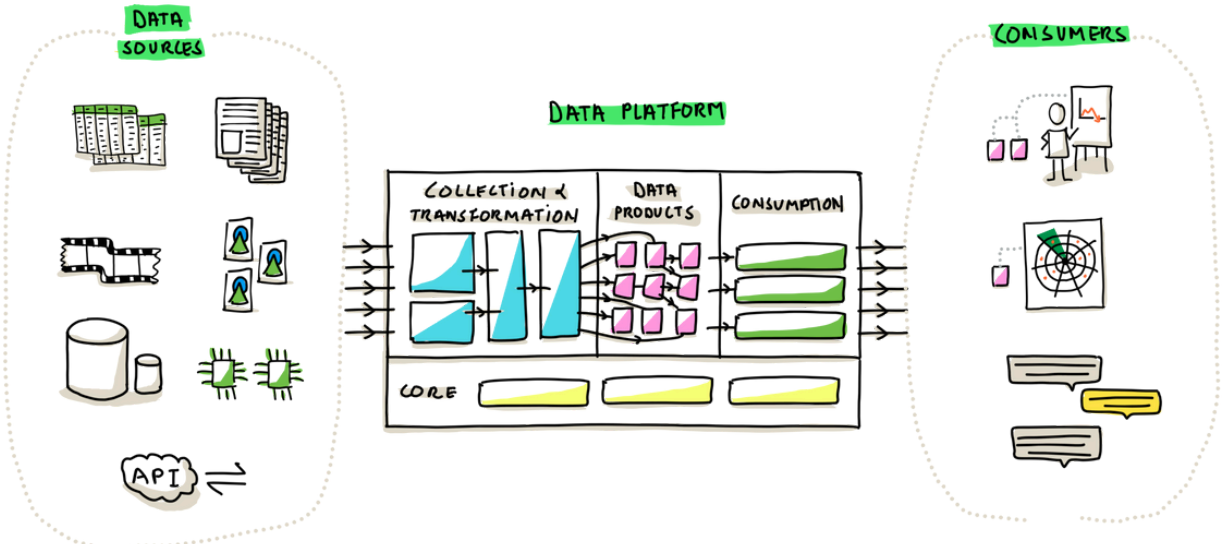
Choosing an effective data strategy involves more than just selecting an organizational structure; it also requires the right software infrastructure to support this strategy. This infrastructure is embodied in what we term a data platform. In the following chapter, we will examine the common characteristics of data platforms and demonstrate how these platforms can be tailored to support any data-centric strategy, enhancing the operational capabilities of the organization.

# {DATA PLATFORM}

A data platform is the suite of software tools that a data-centric force requires to implement its data strategy. Depending on whether the chosen strategy is centralized or decentralized, the focus and features of the data platform will vary.

Data platforms are pivotal for data-centric forces, as they provide the essential tools and functionalities required for data-driven decision-making. Comprising various software components, a data platform handles a wide range of tasks—from the initial **collection** of raw data to its **transformation** into valuable data products, and ultimately, enabling these products to be **consumed** and utilized by domain experts.

These platforms ensure the integrity of **data products**, making them easily discoverable, accessible, and understandable. This, in turn, supports crucial decision-making processes and enhances AI-driven analytics. At the heart of a data platform is a **core layer** that provides critical services, including computing, storage, and a suite of tools aimed at bolstering security, privacy, networking, and data access.



# DATA PLATFORM STRATEGIES

A data platform can support data-centric forces using either centralized or decentralized (data mesh) strategies. This section explores both types of data platforms in detail.

## Centralized Data Platform

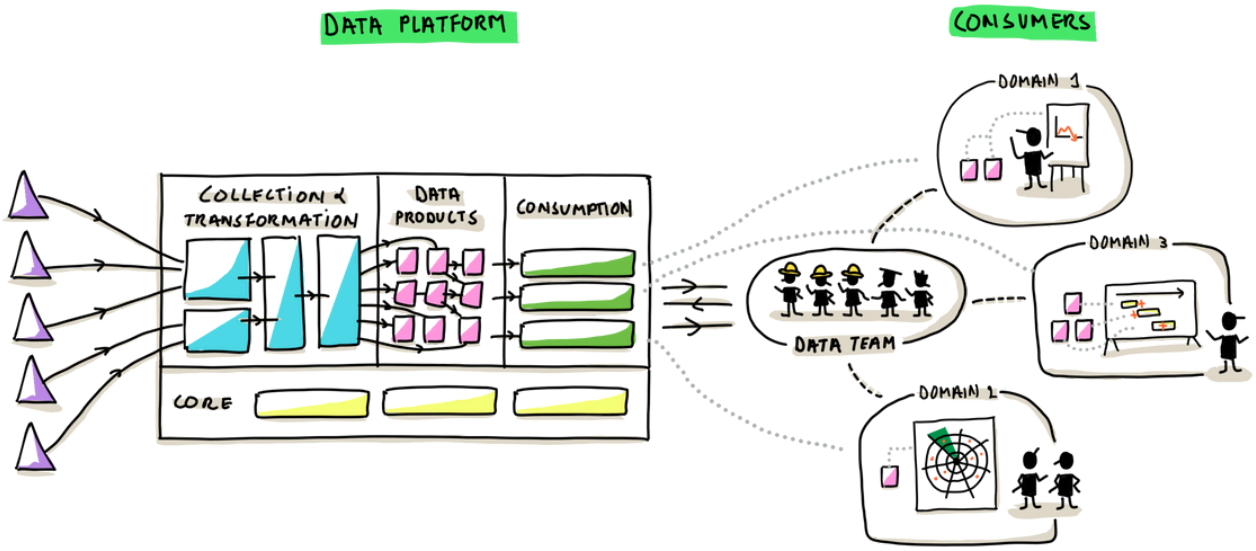
Data-centric forces using a centralized data strategy rely on one central data team to engineer the data platform and to manage data products. Data platforms support these teams by providing two capabilities:

1. **Development tools** to engineer the data platform so that it keeps supporting the force's data-centric goals. For example, configuring the platform to handle tasks like collecting real-time drone positioning information.
2. **Data product development tools** to create, maintain and manage data products, such as converting this raw drone positioning data into ready-to-use data products. Once development is complete, domain experts can utilize these data products, enhancing their effectiveness in the field.

In this approach, the central data team needs to be staffed with the needed domain and technical subject matter expertise to fulfill the data needs of every individual domain.

## Decentralized Data Platform

In a data mesh organization, individual data domains take charge of creating and maintaining their own data products. The data platform supports these domains by providing essential components such as data collection, transformation, and product creation as "services." This setup equips domain teams with self-service tools necessary for managing their data products efficiently. Additionally, the data platform typically includes services that allow domains to register their data products, making these products discoverable and accessible to other teams within the organization.

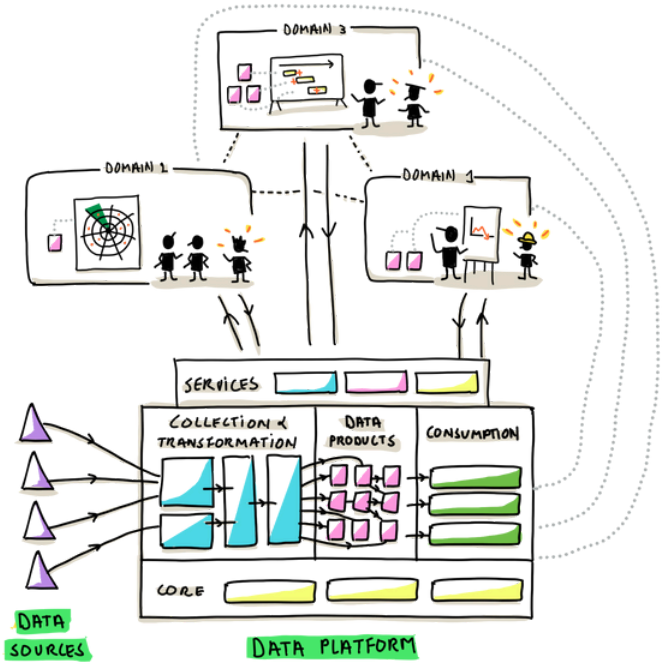



*The major components of a **centralized data platform***

# DATA PLATFORM STRATEGIES

A data platform can support data-centric forces using either centralized or decentralized (data mesh) strategies. This section explores both types of data platforms in detail.

The figure below illustrates how a data mesh platform provides self-service tools to all individual data domains.



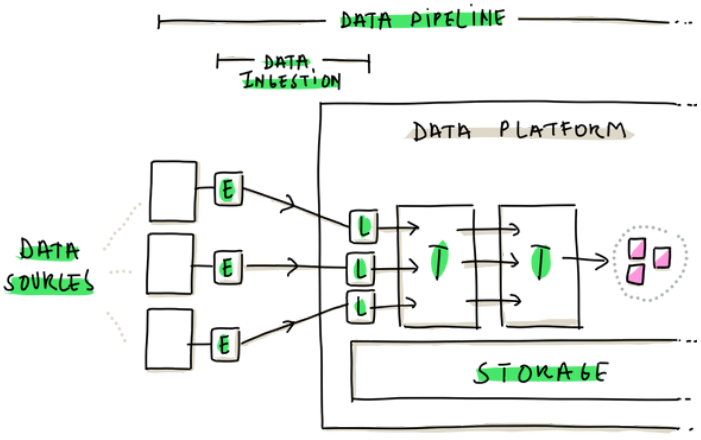

*The major components of a **decentralized data platform***

# DATA PLATFORM FEATURES

This section discusses the most common features of a data platform: data collection and transformation, data product management, data product consumption and the data platform core.

## Data Collection and Transformation

Data collection and transformation are fundamental processes managed by data pipelines, which are engineered to ingest raw data into the data platform and convert it into actionable information. The initial phase, known as Extract Load (EL), involves pulling data from various sources and loading it onto the platform. Data platforms are typically equipped with EL modules that cater to a wide range of data sources, including data streams, relational database management systems (RDBMs), sensors, files, APIs, and social media channels.




Following ingestion, the data undergoes transformation to convert it into more meaningful information. This stage may involve aggregating data, integrating data from different sources, restructuring, and filtering data to improve its utility and relevance. These transformations can be performed using a variety of tools, such as low-code platforms, SQL, or scripting languages like Python.

The effectiveness of data pipelines greatly depends on the platform's data storage and compute capabilities. Efficient storage coupled with sufficient compute power is essential for sustaining the platform's performance. For instance, scenarios requiring near real-time data access necessitate high-performance storage solutions, while situations where daily data refreshes suffice can utilize more cost-effective storage options. Moreover, to manage high volumes of data efficiently, it's crucial to maintain a good balance between storage capabilities and computing power, ensuring that data can be processed within a reasonable timeframe.

## Data Product Management

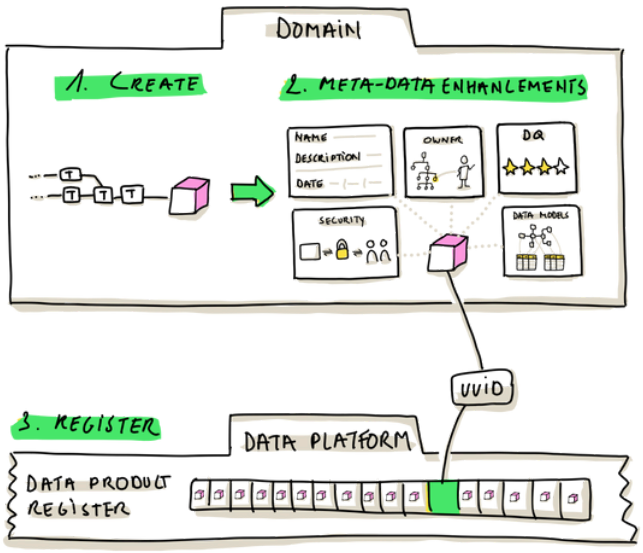
The production flow of data products encompasses three primary phases:

- **Creation**: Transforming raw data, once collected into the platform, into valuable information.
- **Meta-Data**: Adding crucial metadata to the data product to enhance its discoverability, accessibility, usability, connectivity, security, and trustworthiness across different systems.
- **Registration**: Assigning a unique identifier (UUID) through a data product register, making it easily findable for future consumers.

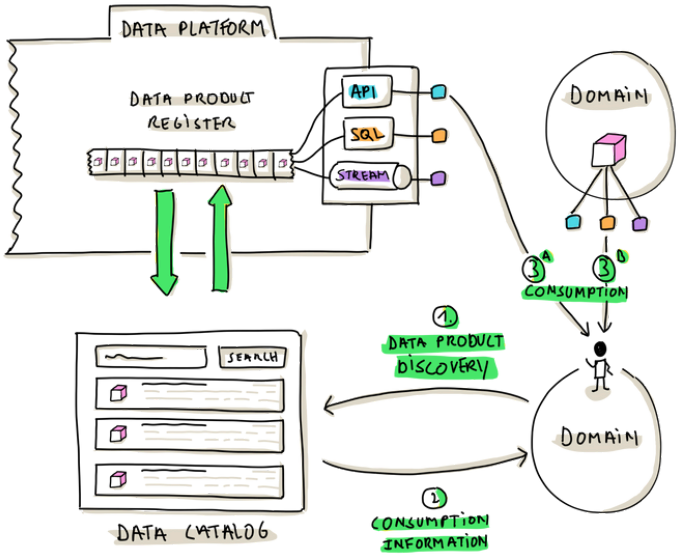
The figure on the next page illustrates the flow of these three phases in a data platform.

# DATA PLATFORM FEATURES

This section discusses the most common features of a data platform: data collection and transformation, data product management, data product consumption and the data platform core.




*Data product **management***




*Data product **consumption***

## Data Product Consumption

The data platform equips consumers with a data catalog to easily discover data products that match their needs, using metadata and a detailed product registry to evaluate product suitability. Once a product is selected, consumers receive essential information on consumption methods, including SQL endpoints, API details, or data stream specifics. Subsequently, effective consumption of the data product occurs, with the platform directing users to the data's physical storage location, which is accessible through both centralized and decentralized methods.

## Data Platform Core

Core services form the backbone of a data platform, ensuring efficient, secure, and reliable operations. These include:

- **Monitoring**: Oversees data pipeline stability and performance, swiftly addressing issues to maintain integrity.
- **Orchestration**: Manages the timing of data processes, like sequencing ingestion before transformation.
- **Infrastructure Management**: Maintains the platform's (virtualized) infrastructure, including cloud, networks and hardware.
- **Security Architecture**: Sets protocols for user authentication and data access, ensuring secure operations.

# CONCLUSIONS

In this chapter, we explored the critical role of data platforms within data-centric forces, detailing the essential functionalities these platforms provide. We covered how they collect and transform data, manage data products, facilitate data consumption, and offer crucial administrative and configuration features.

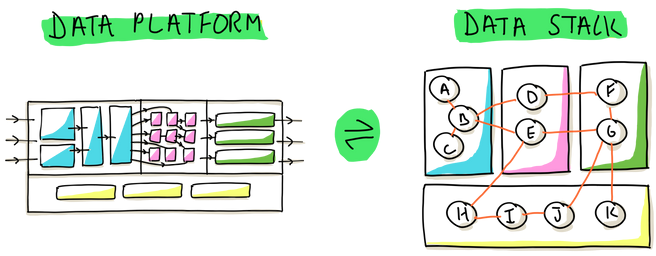
In the next chapter, we will delve deeper into the practical implementation of a data platform for a data-centric force, focusing specifically on the deployment of a **data stack**.

# DATA STACK

A data stack implements the architecture that your data platform requires. This chapter describes a reference data stack suitable for a data-centric force.

A Data Stack is the practical realization of a data platform, implementing the features detailed earlier such as data collection and transformation, product management, consumption, and core services. It consists of a software architecture that includes a suite of software components and their orchestration, forming the backbone of the data platform.

The data stack translates the conceptual design of the data platform into actual functionality. Data stacks combine a variety of technologies, including open-source, cloud services, and proprietary solutions, creating a flexible and reliable environment for effective data management.




In this chapter, we focus on the challenges inherent in constructing a data stack suitable for a data-centric military force. We begin by outlining these challenges, highlighting the complexities involved in designing and deploying a robust data stack. Following this, we present a reference architecture that has been developed based on extensive experience and lessons learned from numerous projects undertaken by Raft for a variety of clients. This reference architecture is designed to effectively address the challenges previously discussed. We will delve into the specific implementation choices made in developing each feature of the data platform, as introduced in the previous chapter, providing a comprehensive overview of the practical aspects of building a data stack.

# DATA STACK CHALLENGES

Implementing a data stack suitable for a data-centric military force presents several substantial challenges. This section provides a non-exhaustive overview of the most common issues encountered.

## Network Resilience

In Disrupted, Disconnected, Intermittent, and Low-bandwidth (DDIL) environments, network conditions can vary greatly. It is critical that the data platform is resilient, maintaining functionality across these varying conditions and compatible with a wide range of military network protocols to ensure robust data exchange.

## Time-Sensitive Data

Military operations often require that data be streamed in real-time due to the time-sensitive nature of military information. This necessitates not only real-time capabilities but also a system that can prioritize data streams, ensuring that critical information is transmitted promptly and prioritized over less urgent data.

## Fault Tolerance

The data platform must be capable of remaining operational even if parts of the software or even entire data centers experience failures.

## Security and Access Control

Military data is highly sensitive, requiring stringent security measures for data sharing within and between different security domains. The data stack must support complex querying and access controls that cater to various sensitivity levels and classifications, ensuring that users have appropriate access based on their roles and permissions.

## Data Quality Assurance

Decision-making must be based on high-quality data. Therefore, the stack needs mechanisms that allow warfighters to provide feedback on the accuracy and quality of the data products they use, helping to maintain and improve data standards.

## Data Product Accessibility and Discovery

The data stack must support diverse data consumption patterns, such as SQL endpoints, message brokers, APIs, and event- or demand-driven access. Additionally, efficient discovery of data products is crucial and should be supported by an extensive data catalog within the stack, enabling users to easily find and access the data they need.

# REFERENCE DATA STACK

Our reference data stack incorporates all data platform functionalities while addressing the specific challenges faced in the military domain.

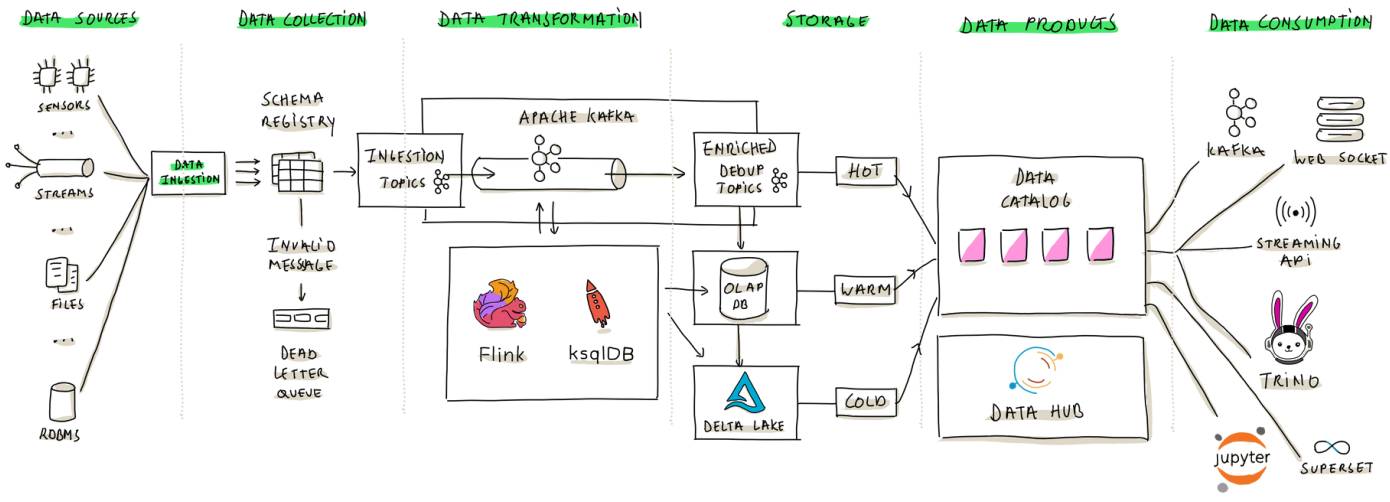
The reference data stack has been developed based on extensive experience and lessons learned from numerous projects undertaken by Raft for a variety of clients. It is architected to provide a robust suite of modular open system technologies and infrastructure, designed to securely manage data across every domain within a data-driven force. This architecture equips data domain teams with comprehensive functionalities including Data Management, Data Visualization, Data Analytics, Data Storage, and the capability to consume, transform, and enrich data sources as needed.

## Data Collection

The data stack adopts a streaming-first approach with Apache Kafka, which is central to the real-time events handling. Kafka serves as a highly available stream processing platform that organizes data across different producers and consumer applications into topics—virtual groups or logs that ensure messages are maintained in a logical sequence. This structure allows seamless data sending and receiving between Kafka producers and consumers.

Each ingestion pipeline within our data stack is designed to extract data from various source systems, including sensor data, RDBMS, streaming data, and files, integrating them into the platform's Kafka ecosystem. During this process, data is validated in real-time against a schema registry. Non-compliant data is directed to Dead Letter queues for further analysis, serving as an initial quality control gate.

The data stack is engineered to interface with both streaming and non-streaming data sources. For instance, to incorporate data from an OLTP RDBMS system, we utilize the Change Data Capture (CDC) feature of the open-source Debezium tool. This method transforms all database change events, such as INSERTs or UPDATEs, into streaming data, thereby guaranteeing that any alterations in the OLTP systems are promptly accessible for further processing.

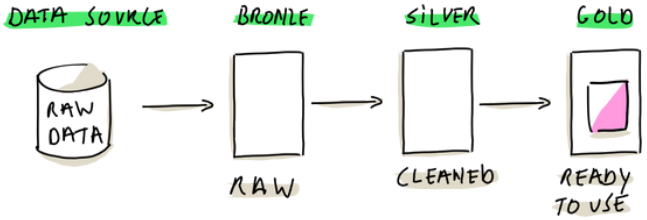


# REFERENCE DATA STACK

Our reference data stack incorporates all data platform functionalities while addressing the specific challenges faced in the military domain.

## Data Transformation

Real-time data transformation is executed using Apache Flink or KsqlDB, which operates on top of Apache Kafka. This setup allows for the consumption of ingestion topics, transformation of the underlying data into more refined information necessary for downstream processes, and subsequently producing this as new topics on the Kafka server. This continuous transformation pipeline ensures that data products are constantly updated and available for deployment.




The data transformation pipelines within the data stack utilize a three-tiered medallion architecture approach. Initially, raw data is collected and stored in the bronze layer. It then undergoes various processes such as matching, merging, conforming, and cleansing in the silver layer. Finally, the refined data products are stored in the gold layer, ready for use.

## Data Storage

Data storage is strategically organized into three distinct tiers—Hot, Warm, and Cold—each defined by specific data policies and employing suitable storage technologies:

- **Hot Tier**: For data where immediacy is crucial (less than 10 minutes old), and is required to be accessible within milliseconds. This data is stored within the Kafka ecosystem to meet high-performance needs.
- **Warm Tier**: Consists of data less than two weeks old, with a requirement for availability within seconds. This tier typically utilizes OLAP databases due to their efficient handling of slightly less urgent data.
- **Cold Tier**: For data older than two weeks, where accessibility within seconds (less than 10 seconds) is sufficient. Cold data is stored using object or data lake storage solutions like AWS S3, MinIO, or Azure Blob Storage, utilizing open formats such as Delta or Parquet to optimize performance and scalability.

## Data Products

Data products combine valuable datasets with required metadata, facilitating immediate use. The data stack supports this integration through a sophisticated data catalog, based on the OSS DataHub platform, that offers extensive data querying and management capabilities across all stored data. This catalog, accessible via a self-service browser interface, enables data domain experts to manage metadata for the data products they are responsible for. The catalog's centralized nature ensures that data products are discoverable enterprise-wide, subject to strict security controls that enforce data visibility based on user roles and attributes.

# REFERENCE DATA STACK

Our reference data stack incorporates all data platform functionalities while addressing the specific challenges faced in the military domain.

## Data Consumption

The data stack supports a variety of mechanisms for data consumption, catering to different operational needs:

- **Message Brokers**: Providing a client library in multiple programming languages like Java, Python, and C#, these brokers facilitate the consumption of streaming data, equipped with built-in logging, monitoring, and observability features.
- **RESTful APIs**: Offering more than 30 API endpoints that comply with the OpenAPI spec, these APIs allow secure access to various data products, integrating seamlessly with identity and access management systems to ensure data is accessible based on user roles and attributes.
- **Websockets**: A bi-directional websocket endpoint delivers data to consumer domain teams with minimal latency, ideal for real-time applications requiring instant data updates.
- **SQL Querying (JDBC)**: Provides exploratory access to data products, enabling SQL queries directly against the platform's databases. Access controls and threshold settings ensure efficient management of query loads.
- **Analytical Notebooks**: Jupyter notebooks offer an interactive environment for more detailed data analysis and modeling, supporting complex analytical tasks.
- **Data Visualization**: Utilizing tools like Apache Superset, the platform enables domain teams to create and provision dynamic, interactive dashboards and visualizations in a low-code environment. These visual tools help in transforming raw data into actionable insights, enhancing the decision-making process across different levels of the organization.

This comprehensive suite of data consumption tools ensures that all data products are accessible, actionable, and valuable to a wide array of end-users, from tactical operators to strategic decision-makers.

## Data Platform Core

At the core of the data stack is a robust platform layer, deployed on Kubernetes, which supports a wide range of functionalities essential for managing the unique challenges of a data-driven force. This core platform layer is meticulously designed to ensure that the data stack is flexible, loosely-coupled, scalable, and responsive across various operational environments. Importantly, this layer allows for versatile application of the data platform, enabling it to function effectively in both centralized and decentralized data mesh approaches. This adaptability is critical in accommodating diverse operational strategies and data management practices within the force, ensuring seamless integration and efficient data handling.

# REFERENCE DATA STACK

Our reference data stack incorporates all data platform functionalities while addressing the specific challenges faced in the military domain.

This paragraph discusses the several key functionalities of this core layer.

### Fault Tolerance

Built upon the principles of the Reactive Manifesto, the data stack is designed to be highly resilient and responsive even in the face of system failures. This includes:

- **Reactive**: The system quickly responds to changes and challenges, ensuring continuous operation and user interaction.
- **Resilient**: By replicating data and isolating components, the system maintains responsiveness even during failures, minimizing downtime and data loss.
- **Elastic**: The platform can scale resources up or down based on demand, ensuring efficient resource use without bottlenecks.
- **Message-Driven**: Utilizes asynchronous message-passing to enhance system integration and failure management, promoting better data flow and system health.

### Network Resilience

The platform ensures robust data synchronization and accessibility across distributed nodes, whether in cloud, on-premise, or edge environments:

- **Continuous Operation**: Operates seamlessly at the edge, utilizing local data caching and processing to function independently of central systems when network connectivity is compromised.
- **Prioritized Data Sync**: Automatically syncs data across nodes once connectivity is restored, employing intelligent tagging and prioritization to manage data flows efficiently.

### Orchestration

Employs Apache Airflow to manage and coordinate complex data workflows across multiple simultaneous jobs, enhancing data pipeline efficiency and reliability:

- **Directed Acyclic Graphs (DAGs)**: Each pipeline is defined as a DAG, allowing for flexible, reliable scheduling and execution of tasks.
- **Quality Checks**: Incorporates automated data quality checks within workflows to ensure the integrity and accuracy of data products.

### Security and Access Control

Integrates advanced security frameworks to protect data integrity and comply with stringent access requirements:

- **Integrated ICAM Solutions**: Works with Identity, Credential, and Access Management solutions like Keycloak to enforce secure access and authentication.
- **Data Encryption**: Implements strong encryption protocols for data at rest and in transit, safeguarding data against unauthorized access.

# REFERENCE DATA STACK

Our reference data stack incorporates all data platform functionalities while addressing the specific challenges faced in the military domain.

- **Zero Trust Architecture**: Applies a Zero Trust framework using Open Policy Agent (OPA), ensuring data access is strictly verified at every point, not just at perimeter.

### Machine Learning Integration

Supporting advanced machine learning capabilities to enhance predictive analytics and decision-making processes:

- Feature Store: Utilizes a dedicated feature store (Feast) to manage and serve model features consistently across training and production environments.
- Model Training and Deployment: Leverages Kubeflow on Kubernetes to streamline ML workflows from development to deployment, enhancing model accuracy and performance.

# CONCLUSIONS

This chapter provided an exploration of the challenges associated with developing a data stack tailored for a data-centric military force, followed by the presentation of a reference architecture designed to meet these challenges effectively. We dissected the architecture, examining the rationale behind each implementation choice made to support the functionalities of the data platform introduced earlier. The comprehensive detailing of this architecture, drawn from Raft's extensive experience across various projects, serves as a valuable blueprint for constructing robust data stacks.

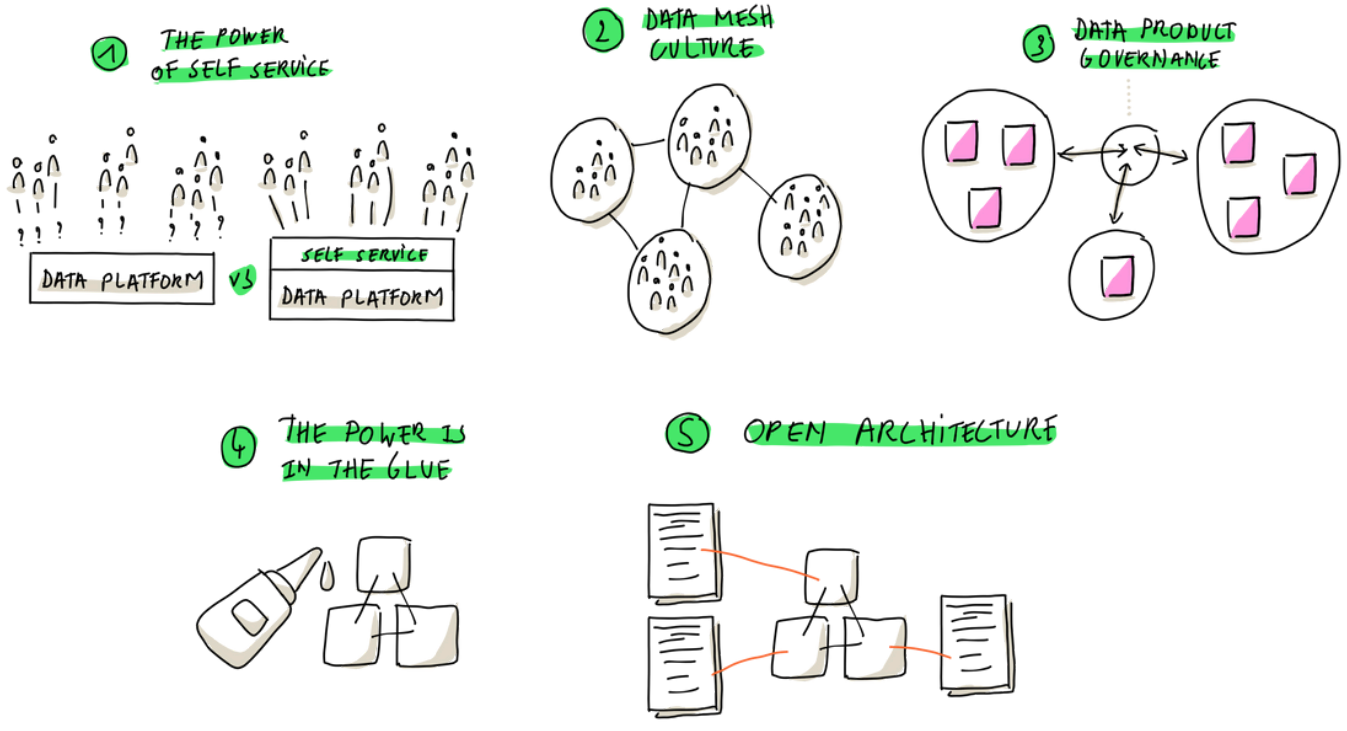
This reference data stack can serve as a source of inspiration for those looking to build or enhance their own data stacks in support of military or similarly complex operations, demonstrating how strategic planning and insights from past implementations can lead to the development of a highly functional and adaptable data management environment. This guidance is instrumental in helping organizations navigate the intricacies of data-centric system design, ensuring that their data architectures are both resilient and capable of meeting the dynamic demands of modern operations.

# {RECOMMENDATIONS}

Are you planning to build your own Data Stack? Consider our **five key recommendations** as you embark on a data project aimed at developing a data-centric force.

As we conclude this white paper, we offer five recommendations for organizations considering the development of a data platform stack tailored to a data-centric operation.

These recommendations are grounded in the substantial experience Raft has accumulated through its work with numerous organizations.



# 5 RECOMMENDATIONS

Are you planning to build your own Data Stack? Consider our **five key recommendations** as you embark on a data project aimed at developing a data-centric force.

## 1 - The Power of Self Service

Technical expertise is a scarce resource, and the demand for data subject matter experts (SMEs) to enhance and expand data platforms often exceeds supply. The most effective solution to this challenge is to invest in self-service platforms. Such platforms empower non-technical SMEs to create and manage data products, facilitating broader usability across different domains without constant reliance on highly specialized staff.

## 2 - Data Mesh Cultural Changes

The concept of Data Mesh is potent yet complex, encompassing both organizational structure (referred to as 'a decentralized approach' in this paper) and architectural pattern. It's crucial to recognize that implementing a data mesh architecture requires a significant cultural shift towards decentralized operations. Similarly, a data mesh organizational model hinges on the support of a compliant data platform. Balancing technical enhancements with these necessary cultural transformations is key to the success of your data platform initiatives.

## 3 - Data Product Governance

In the context of a data mesh, where the risk of creating isolated data silos is increased, data products become essential. They integrate local datasets with necessary metadata, enabling their use across various domains in appropriate contexts. Implementing a federated data governance approach, diligently monitored and enforced, is critical to ensure that data products effectively prevent the formation of silos and facilitate broad data utility.

## 4 - The Power is in the Glue

Data stacks integrate a diverse array of technologies—open-source, cloud services, and proprietary solutions—to create a versatile and robust data management environment. The selection of suitable technology components tailored to specific needs is the first critical design decision. The second, equally important task is the integration of these components into a cohesive data stack. Past projects have shown that the expertise required to effectively 'glue' these components together should not be underestimated. Mastery in integrating diverse technologies can lead to a data stack that is greater than the sum of its parts.

## 5 - Open Architecture

When building or selecting a data platform, it's vital to consider the risks of vendor lock-in. Opting for best-in-breed open-source software is an effective strategy to maximize flexibility and maintain transparency, which is crucial for detecting and mitigating security vulnerabilities. Furthermore, open-source components allow for deep audits of the technology and support from a robust ecosystem and community, enhancing both security and innovation.

# CONCLUSIONS

The recommendations provided in this chapter offer strategic guidance to help you build a data stack that not only meets your immediate needs but also prepares your organization to effectively manage and utilize data for improved decision-making and strategic advantage. As you embark on or advance your journey towards becoming a data-centric force, remember that support is readily available.

Should you need assistance at any point, feel free to contact Raft. We are committed to aiding your efforts by sharing our extensive expertise in data stack development and implementation. Allow us to help you unlock the full potential of your data to realize your strategic goals.

# [ABOUT RAFT]

We are *Team Raft*. Together, we're building the sharpest digital solutions and solving the most complex problems, while having fun.

info@teamraft.com
p: 703.570.4820
11800 Sunrise Valley Dr. #400
Reston, VA 20191