FROM DATA WAREHOUSES TO DATA MESH:

THE EVOLUTION OF DATA PLATFORMS

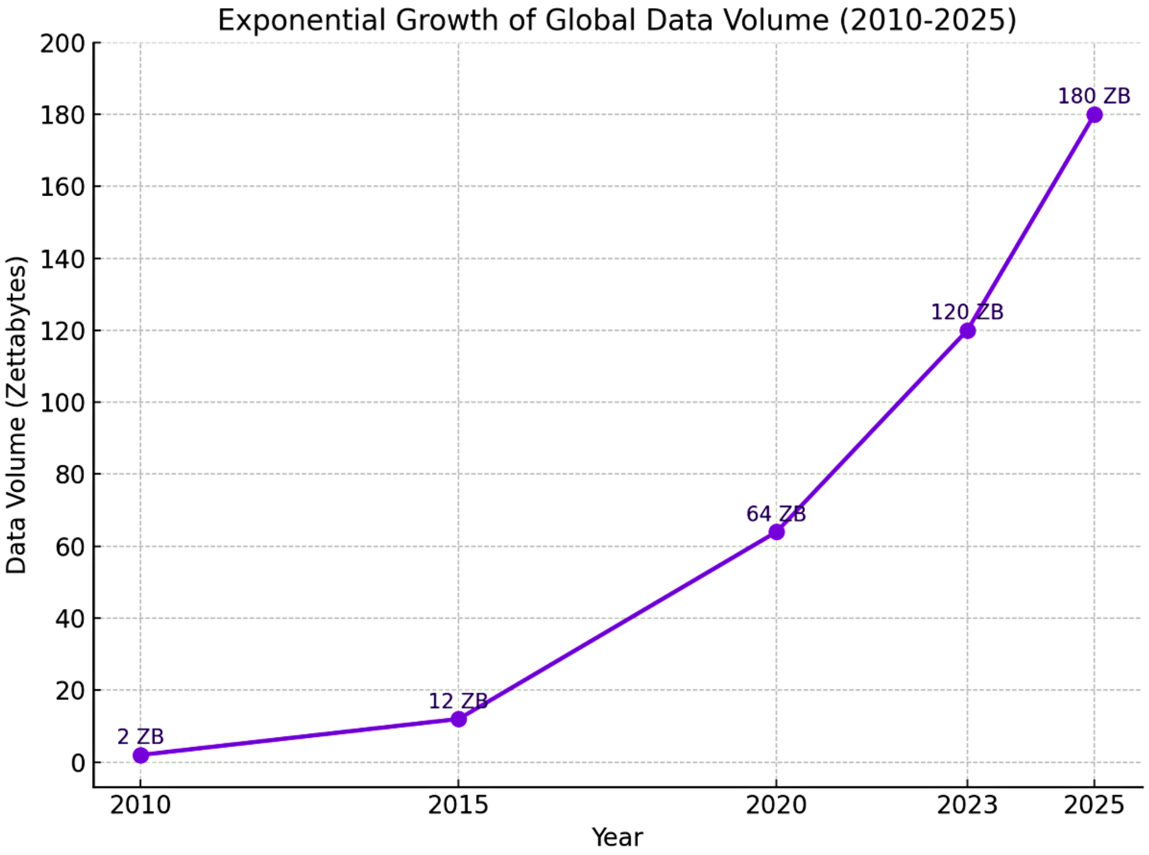
SYNTIO

INTRODUCTION

- The ability to analyze and act on data has been the defining factor for success in large enterprises
- Data analysis has always been at the heart of decision-making in large enterprises, from optimizing supply chains to understanding customer behavior
- As businesses grew, so did the complexity of their data. This required innovations in how data was collected, stored, and analyzed

CHALLENGES OF DATA

- The importance of data in decision-making has always been evident, but the scale and complexity of that data have grown in ways few could have predicted
- **3Vs of data**:
    - **Volume**, with massive datasets from sources like IoT and social media.
    - **Velocity**, as businesses demand real-time insights.
    - **Variety**, from structured databases to unstructured formats like video and logs



TYPES OF DATA ANALYTICS

PAST

Descriptive Analytics: Understanding the Past

Objective: Answers the question, *"What happened?"*

Focus: Aggregating and summarizing historical data into reports and dashboards.

NOW

Real-Time Analytics: Monitoring the Present

Objective: Answers the question, *"What is happening right now?"*

Focus: Processing data streams in real-time to identify current trends, anomalies, or events as they happen.

FUTURE

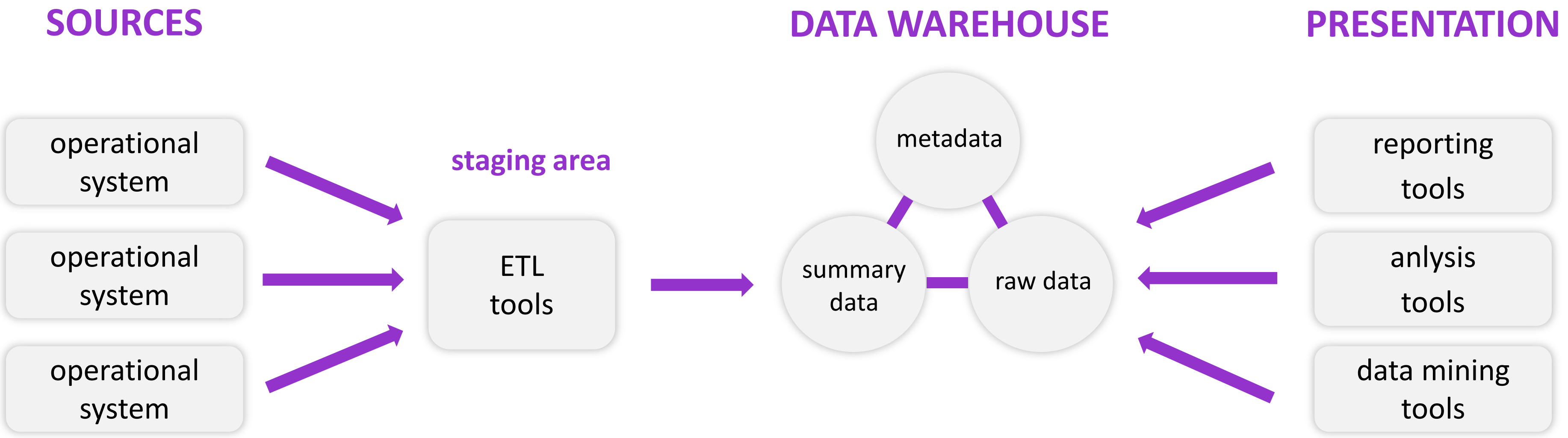
Predictive Analytics: Shaping the Future

Objective: Answers the question, *"What is likely to happen?"*

Focus: Using historical and real-time data to forecast trends, risks, and opportunities.

EARLY DAYS OF DATA ANALYSIS

- The 1980s and 1990s marked a turning point with the rise of Data Warehousing
- Warehousing introduced the concept of Extract, Transform, Load (ETL), a process that standardized and cleaned raw data for reporting
- Business Intelligence (BI) tools made data accessible to decision-makers
- Reports and dashboards provided insights into historical performance, enabling better strategic planning

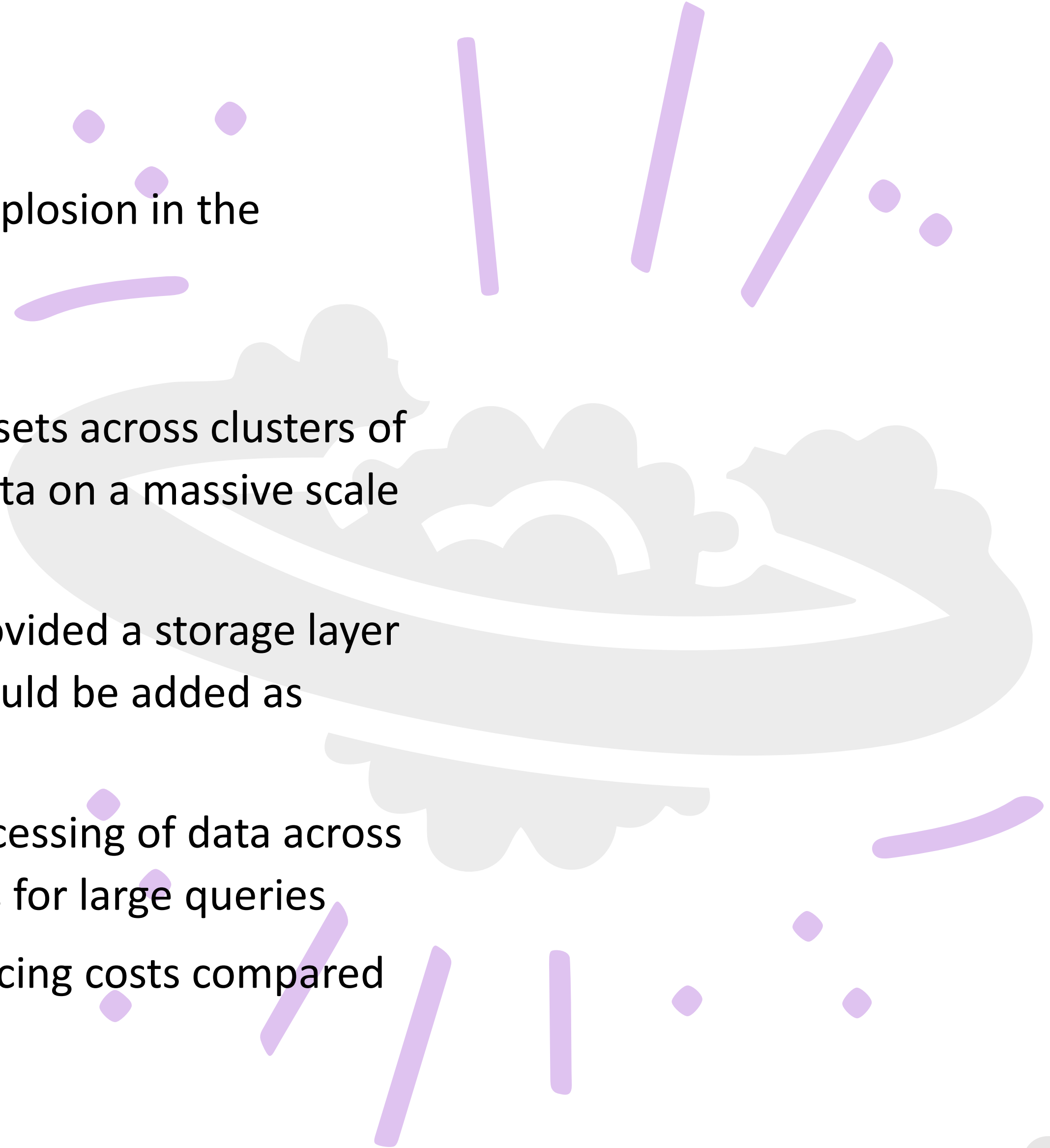


PEAK OF DATA WAREHOUSEING

- **High Performance and Scalability**
  - Traditional data warehouses reached their peak with powerful systems capable of handling **petabytes of structured data**.
  - These systems provided **fast query performance** and **advanced optimization** techniques to support large-scale analytics.
- **Integrated Solutions**
  - Combined **hardware** and **software** to maximize efficiency and performance for complex, analytical queries.
  - Data consolidation became easier, offering a **single source of truth** for decision-makers.
- **Limitations**
  - **High costs** and complex maintenance.
  - **Rigid schema designs** that lacked flexibility and agility.
  - Primarily optimized for **structured data**, struggling with newer data formats and real-time processing needs.
- **Catalyst for Change**
  - These limitations paved the way for newer technologies like **Big Data** systems (e.g., Hadoop, Spark) that could handle **greater variety, velocity, and volume** of data.

BIG DATA ERA

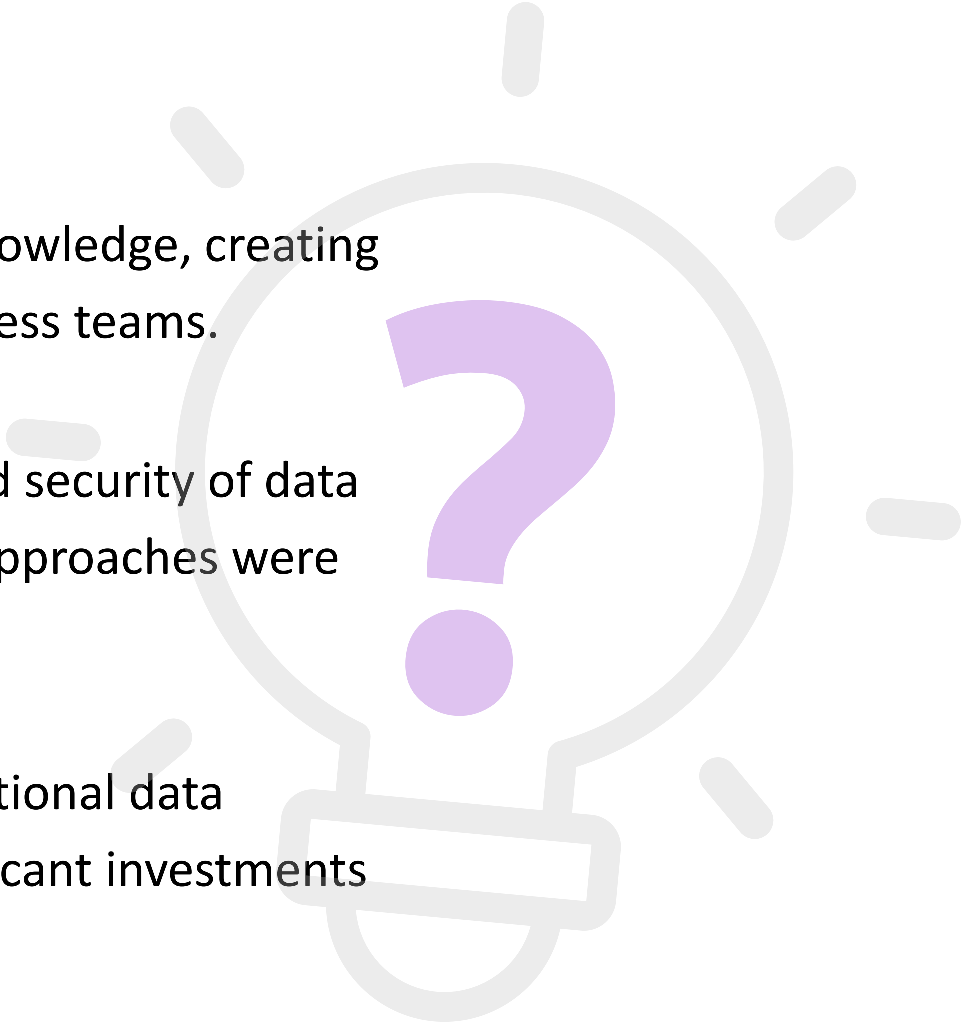
- In the late 2000s organizations were confronted with an explosion in the volume, velocity, and variety of data
- **The Emergence of Hadoop**
  - it allowed for the **distributed processing** of large datasets across clusters of computers, making it possible to store and process data on a massive scale at a fraction of the cost of traditional systems
  - Hadoop's **HDFS (Hadoop Distributed File System)** provided a storage layer that could scale horizontally, meaning more nodes could be added as needed to accommodate growing datasets
  - The **MapReduce** framework enabled the parallel processing of data across multiple machines, vastly improving processing times for large queries
  - built to run on commodity hardware, drastically reducing costs compared to expensive, traditional data warehousing solutions



BIG DATA ERA

- **Data Lakes** allowed organizations to store **structured, semi-structured, and unstructured data** in a single repository, without the need for predefined schema, providing greater flexibility.
- Spark addressed the limitations of Hadoop's **MapReduce**, offering significantly faster data processing by leveraging **in-memory computing**. This was a game-changer for data science and real-time analytics.

THE CHALLENGES OF BIG DATA



- **Complexity and Skills Gap:**
  Technologies like Hadoop and Spark required specialized knowledge, creating a gap between the capabilities of IT departments and business teams.
- **Data Governance and Quality:**
  As data volumes grew, ensuring the quality, consistency, and security of data became increasingly difficult. Traditional data governance approaches were not well suited to handle the scale and diversity of Big Data.
- **Integration with Existing Systems:**
  Integrating Big Data tools with legacy systems, such as traditional data warehouses or operational databases, often required significant investments in time and resources.

THE SHIFT TO CLOUD

- Before the cloud, organizations faced numerous challenges in scaling their data infrastructure, such as:
  - **High upfront costs** for hardware and storage
  - **Complexity in scaling** as data volumes increased
  - **Difficulty in maintaining** and securing infrastructure
- Cloud providers addressed these issues by offering **on-demand** computing resources and **managed services**, dramatically simplifying the process of scaling data infrastructure
- Key Features:
  - Scalability
  - Centralized Data Lakes and Warehouses
  - Advanced Analytics and AI Integration
  - Reduced Infrastructure Management
  - Global Accessibility and Collaboration

CENTRALIZED CLOUD ANALYTICS

ADVANTAGES:

- **Cost Efficiency:**

  Cloud services offer a **pay-as-you-go** model, allowing organizations to avoid large upfront capital expenditures and only pay for what they use.

- **Faster Time to Value:**

  With the cloud, companies can rapidly deploy new analytics solutions and start deriving insights quickly. The availability of pre-built **data pipelines** and **AI models** reduces the time required to set up complex systems.

- **Data Democratization:**

  Cloud platforms make it easier to provide employees across different departments with access to data and insights. The use of **self-service BI tools** democratized data access, empowering non-technical business users to create reports and analyze data independently.

CHALLENGES

- **Data Silos and Integration:**

  While the cloud offers centralized analytics, many companies still face issues with data being siloed across different cloud platforms, business units, or regions. **Data integration** between various sources often requires additional tools and management efforts.

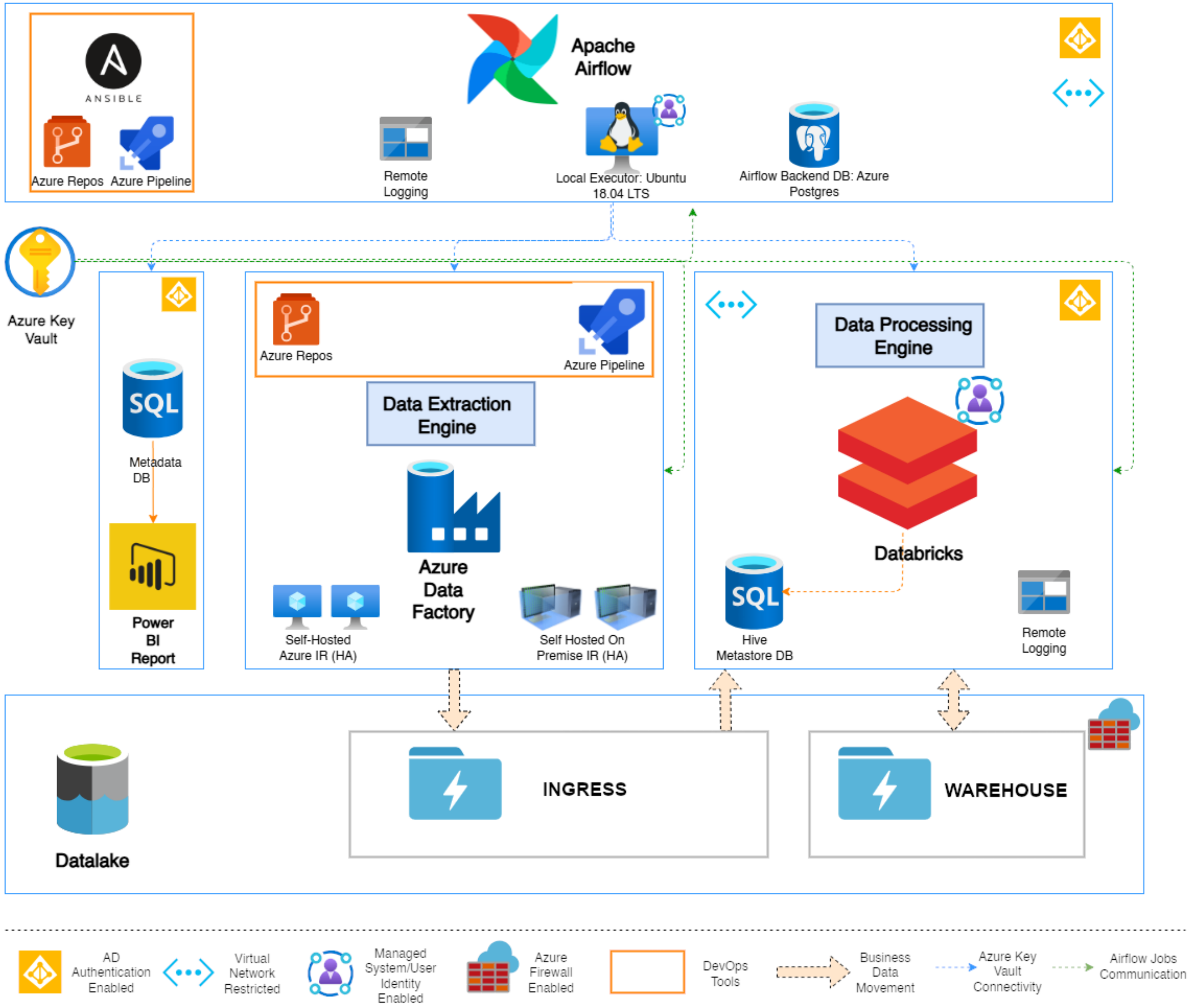
- **Data Governance and Security:**

  As enterprises migrate more data to the cloud, the **security** and **governance** of sensitive data become a pressing concern. Protecting against breaches, ensuring compliance with regulations (e.g., GDPR), and implementing strong data privacy controls are key challenges in the cloud era.

- **Cloud Cost Management:**

  The flexibility of cloud platforms meant that costs could quickly spiral if not carefully managed. **Unexpected costs** from over-provisioning or underestimating usage are a common pitfall, requiring organizations to implement sophisticated cost monitoring and optimization strategies.

CENTRALIZED CLOUD ANALYTICS



DATA MESH AND DECENTRALIZATION

- **CENTRALIZED DATA SYSTEMS** like data lakes and warehouses, while powerful, often fail to meet the diverse needs of large enterprises with complex data requirements.

  The centralization of data management leads to **data bottlenecks**, **siloed teams**, and challenges around **data quality**, **ownership**, and **scalability**

Data Mesh emerged as a way to decentralize and distribute data management, focusing on domain-specific data ownership and domain-driven design

DATA MESH KEY PRINCIPLES

DATA AS A PRODUCT:

Data should be treated as a product, owned and managed by specific teams (domains) rather than a central data team. Each domain is responsible for the quality, accessibility, and lifecycle of the data it produces.

DOMAIN-ORIENTED OWNERSHIP:

Instead of a single centralized team maintaining all the data, **domain teams** that understand the data best (e.g., finance, marketing, operations) take ownership. These teams are responsible for **producing, storing**, and **making their data available** in a usable, consumable format.

DECENTRALIZED DATA ARCHITECTURE:

Data Mesh promotes a decentralized architecture, with each domain team managing its own data pipelines, data storage, and analytics. While these domains can leverage shared infrastructure and technologies, the key is that the **data infrastructure** and **data governance** are domain-specific, not centralized.

INTEROPERABILITY AND DATA DISCOVERABILITY:

Data Mesh requires strong **interoperability** standards, ensuring that data produced by different domains is easily discoverable and usable by others. This is supported by **metadata management, data catalogs**, and standardized APIs that allow domains to share their data without losing control over it.

BENEFITS OF DATA MESH

SCALABILITY AND FLEXIBILITY

IMPROVED DATA QUALITY AND RELEVANCE

FASTER TIME TO INSIGHT

AVOIDANCE OF DATA BOTTLENECKS

CHALLENGES OF DATA MESH

COMPLEX IMPLEMENTATION

Implementing a Data Mesh architecture requires significant organizational change. It demands that organizations rethink their data culture, with an emphasis on cross-functional collaboration between domain teams, IT, and data engineers.

NEED FOR SKILLED TEAMS

Data Mesh relies on empowered domain teams, meaning that these teams must have the skills and resources to manage their own data products. This requires data literacy and technical expertise within each business unit.

DATA INTEGRATION

With decentralization, integrating data from multiple domains can become a challenge. Establishing common data standards, APIs, and data-sharing protocols is crucial to ensure seamless data flow between domains while maintaining autonomy.

GOVERNANCE AND CONSISTENCY

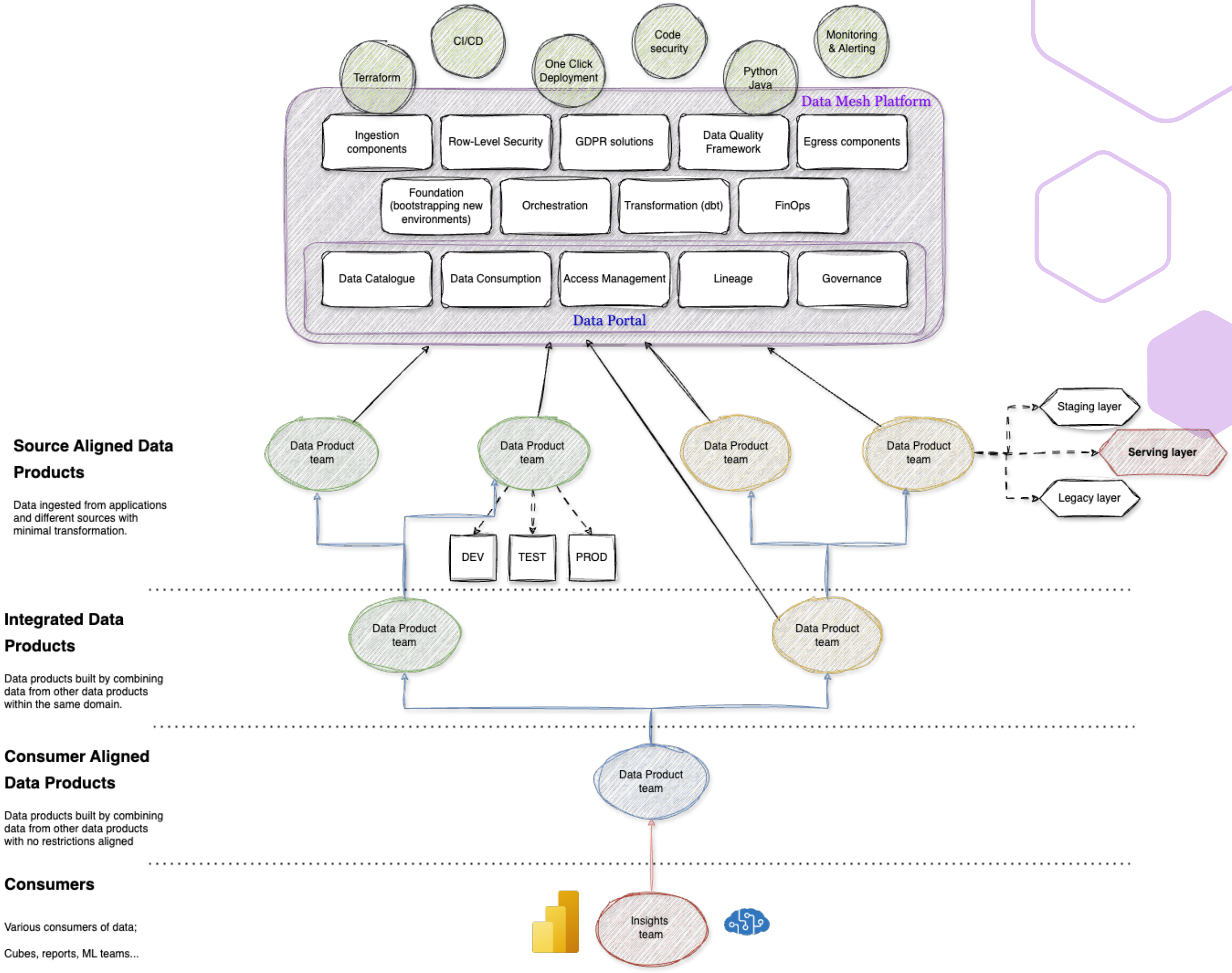
While decentralization offers flexibility, it also requires robust governance mechanisms to ensure consistency, security, and regulatory compliance across a distributed system. Without proper oversight, there is a risk of data fragmentation and inconsistency.

GOVERNANCE IN DATA MESH

- One of the core challenges of Data Mesh is ensuring **consistent data governance** across distributed domains. In a decentralized system, data governance must be designed to accommodate:
- **Data Security**: Each domain must ensure that its data is secured, with appropriate access controls.
- **Data Quality**: Establishing standards and processes for data quality is essential, as each domain is responsible for maintaining its own data pipelines and outputs.
- **Compliance**: Adherence to regulatory frameworks (e.g., GDPR) must be managed at the domain level to ensure that data is handled properly, while still enabling decentralized control.
- A strong **metadata layer** and the use of **data catalogs** are essential in making data discoverable, ensuring compliance, and providing visibility across domains. This helps to avoid data silos and ensures that data from different domains can be integrated when necessary.

DATA MESH PLATFORM EXAMPLE ARCHITECTURE

- Standardized bootstrapping for new Data Products
- ETL done in a standardized way
- Governance rules and checks are ingrained in every process
- Centralized services for Orchestration, Data Quality
- Data Portal – one-stop shop with integrated Data Catalogue, Access Management, Lineage, DQ, and Governance



Terraform

CI/CD

One Click Deployment

Code security

Python Java

Monitoring & Alerting

Data Mesh Platform

Ingestion components

Row-Level Security

GDPR solutions

Data Quality Framework

Egress components

Foundation (bootstrapping new environments)

Orchestration

Transformation (dbt)

FinOps

Data Catalogue

Data Consumption

Access Management

Lineage

Governance

Data Portal

Source Aligned Data Products

Data ingested from applications and different sources with minimal transformation.

Data Product team

Data Product team

Data Product team

Data Product team

Staging layer

Serving layer

Legacy layer

DEV

TEST

PROD

Integrated Data Products

Data products built by combining data from other data products within the same domain.

Data Product team

Data Product team

Consumer Aligned Data Products

Data products built by combining data from other data products with no restrictions aligned

Data Product team

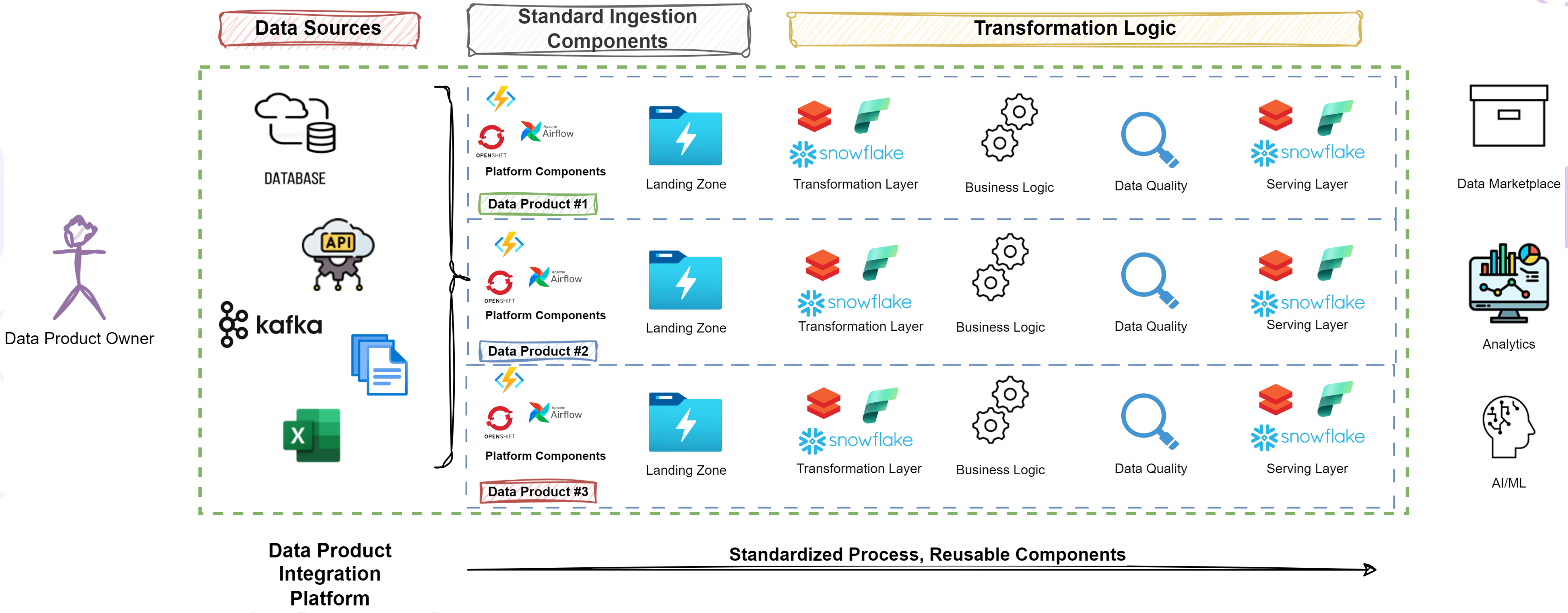
Consumers

Various consumers of data;

Cubes, reports, ML teams...

Insights team

DATA MESH PLATFORM EXAMPLE ARCHITECTURE



HOW TO CHOOSE BETWEEN CENTRALIZED CLOUD ANALYTICS AND DECENTRALIZED DATA MESH?

When to choose Centralized Cloud Analytics:

- Smaller or Less Complex Data Ecosystem
- Need for Centralized Control and Compliance
- Fewer Dependencies Across Teams
- Established Analytics and BI Practices
- Cost Efficiency at Scale

When to choose Data Mesh:

- Large, Complex Organizations with Multiple Domains
- Need for Agility and Speed
- Increasing Volume and Variety of Data
- Growing Need for Real-Time Data Processing
- Increased Focus on Data Products and Collaboration
- Long-Term Growth and Scalability

KEY DIFFERENCES

Organizational Size and Complexity:

Small to medium-sized organizations with less complexity may find centralized cloud analytics sufficient for their needs, while larger organizations with multiple departments and high volumes of data are better suited for Data Mesh.

Data Governance and Compliance Needs:

If centralized control, security, and compliance are paramount, a centralized cloud analytics model may be more appropriate, as it simplifies governance through a single, controlled system.

Organizational Size and Complexity:

Data Mesh is ideal for organizations that need to move quickly, innovate, and allow teams to independently manage their data, while centralized cloud analytics is better for those seeking standardized insights and streamlined operations.

Cost and Resource Availability:

The centralized cloud model is often more cost-effective for smaller teams, as it requires fewer specialized domain teams and relies on the efficiencies of the cloud. Data Mesh requires more investment in domain-specific expertise and governance mechanisms, so it may be costlier to implement initially.

HOW TO CHOOSE BETWEEN CENTRALIZED CLOUD ANALYTICS AND DECENTRALIZED DATA MESH?

- Companies need to carefully evaluate their unique needs, goals, and the maturity of their data infrastructure

- In the **Centralized Cloud Analytics** approach, all data is managed, stored, and analyzed centrally, often within a **cloud-based data warehouse** or **data lake**. This model relies on a centralized team to manage data integration, governance, and security. Major cloud platforms provide the infrastructure and services to support centralized analytics.

- The **Data Mesh** approach decentralizes data ownership, where individual **domain teams** (e.g., finance, marketing, operations) are responsible for managing their own data as a product. This decentralized approach is driven by **domain-oriented ownership** and **cross-functional collaboration**.

Thank you!

SYNTIO