

GOTO **AMSTERDAM 2024**

WHAT'S THE BEST BIG DATA ARCHITECTURE FOR YOU?

#GOTOams

What's the best Big Data Architecture for you?

Christoph Windheuser

Databricks

June 12 2024

GOTO Amsterdam 2024

About me

- PhD in Machine Learning at ENST (Paris France), Carnegie Mellon University (Pittsburgh US) and Waseda University (Tokyo Japan)
- Different positions at SAP and Capgemini
- 4 years at ThoughtWorks, implementing Data Mesh Architectures with customers
- 2 years at Databricks, Sr. Manager Professional Services Central EMEA



The views and opinions on this talk are mine and not that of Databricks.

Agenda

- Why Big Data and Big Data Architectures (BDA)?
- What is a BDA and what are the requirements?
- A brief history of Data Architectures
- Different BDAs:
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- Outlook: The Future of BDA
- Summary: 4 things to remember

Agenda

- **Why Big Data and Big Data Architectures (BDA)?**
- What is a BDA and what are the requirements?
- A brief history of Data Architectures
- Different BDAs:
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- Outlook: The Future of BDA
- Summary: 4 things to remember

Why "Big Data"?

Data Driven Companies are more successful!

What is "Big Data"?

"Big Data" = "All Data"

(acc. James Serra, Deciphering Data Architectures, O'Reilly)

- Structured (ex. tables)
- Semi-Structured (ex. XML, JSON)
- Unstructured (ex. text, pictures, audio, video)
- Batch Data
- Streaming Data

Data- driven companies are:

23x

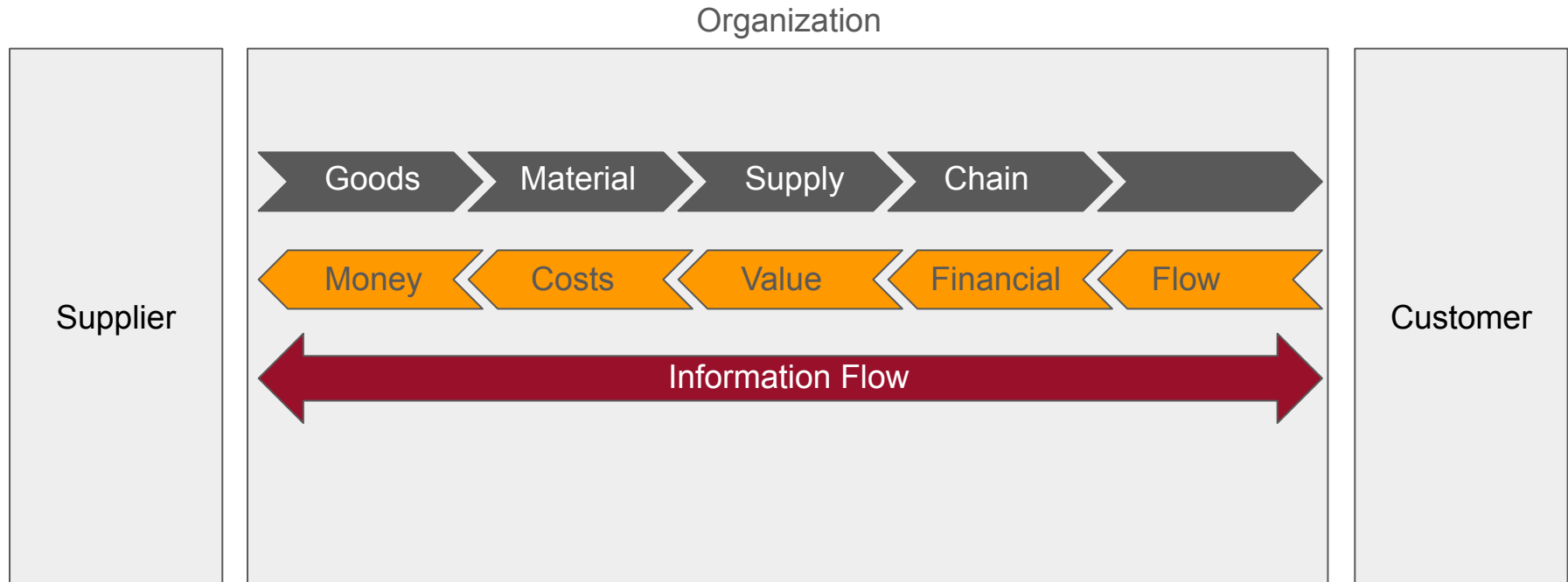
**more likely to grow
customer base**

19x

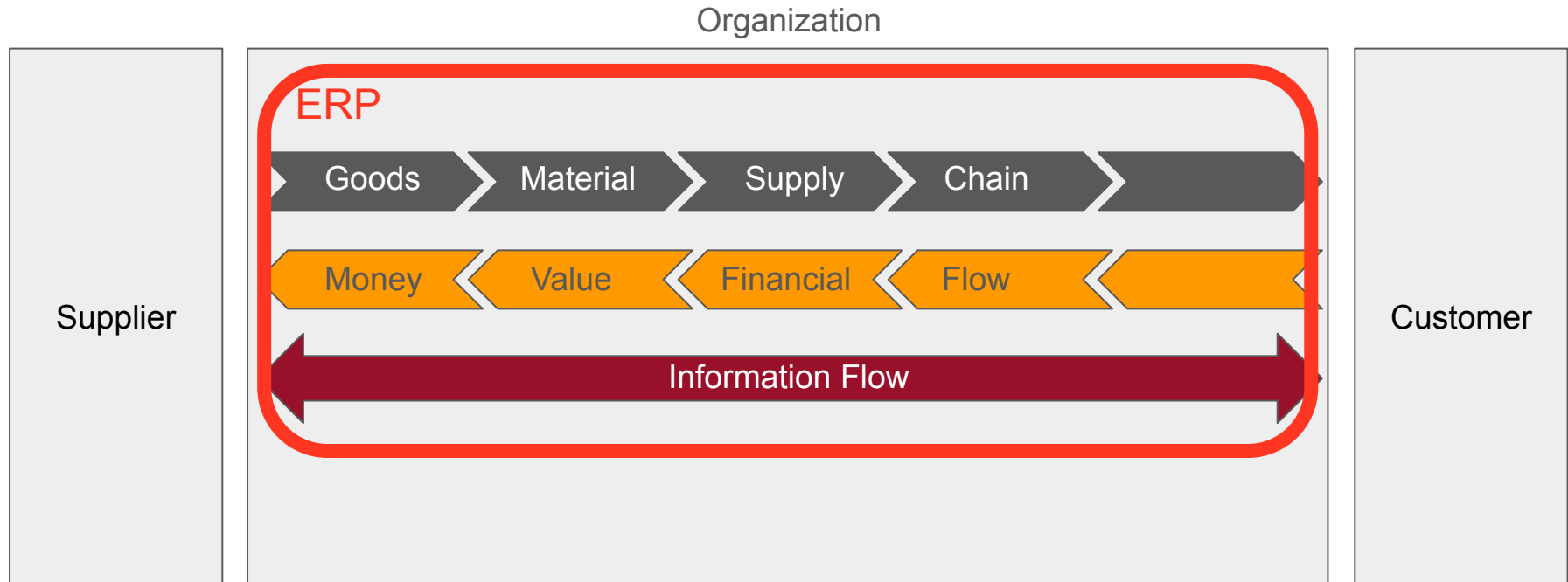
more likely to be profitable

Sources: Forrester, Forbes

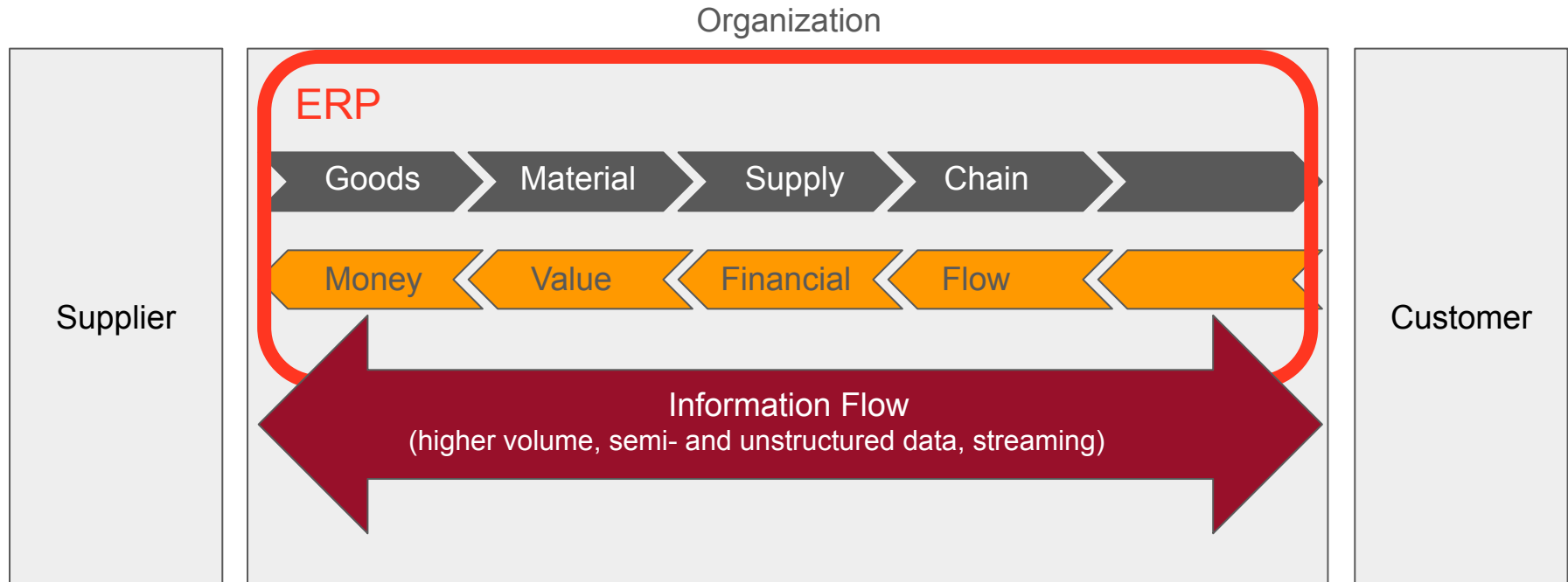
Why do we need a "Big Data Architecture" (BDA)?



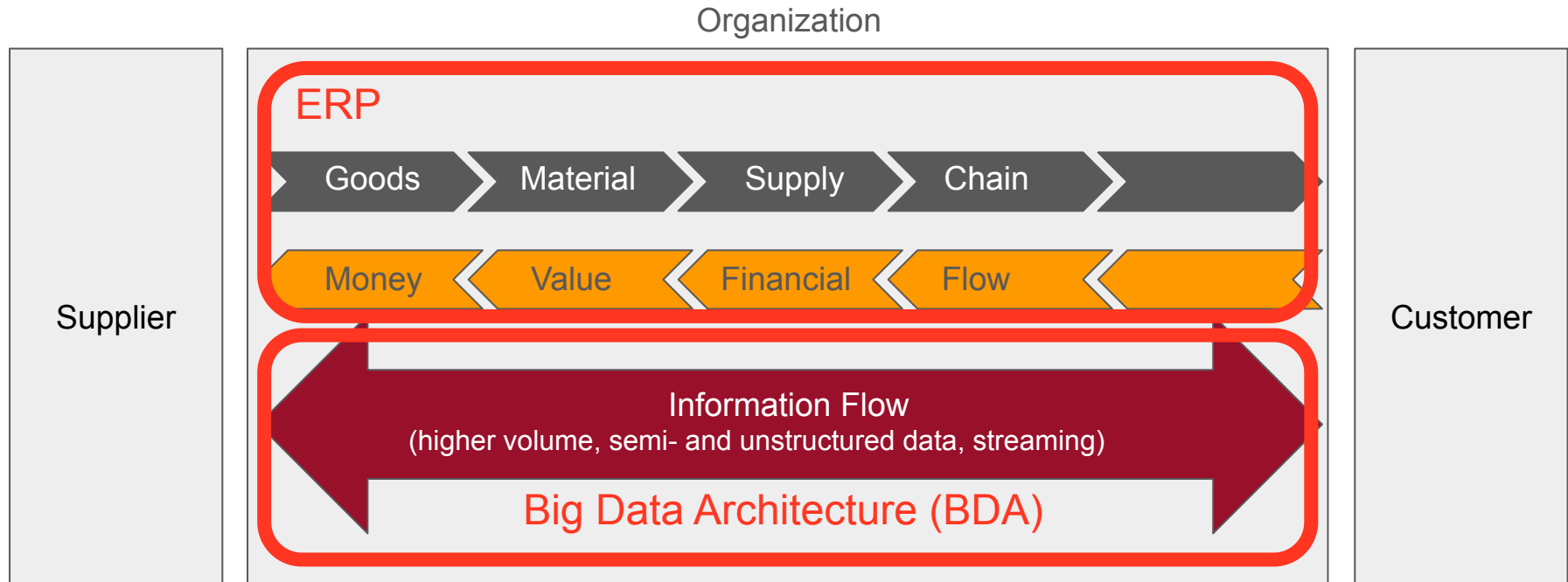
Why do we need a "Big Data Architecture" (BDA)?



Why do we need a "Big Data Architecture" (BDA)?



Why do we need a "Big Data Architecture" (BDA)?



Agenda

- Why Big Data and Big Data Architectures (BDA)?
- **What is a BDA and what are the requirements?**
- A brief history of Data Architectures
- Different BDAs:
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- Outlook: The Future of BDA
- Summary: 4 things to remember

What is a BDA and what are the Requirements?

A Big Data Architecture must provide the *right information at the right time at the right place for persons and applications, regardless of the type and size* of the data.

1. "Unlimited" storage
2. "Unlimited" compute
3. Capable to work with all kind of data (structured, semi- and unstructured)
4. Capable to support all use cases (analytics, streaming, DS, ML, GenAI, etc.)
5. Provides information via UI and API
6. Data must be trustworthy
7. Data must be discoverable
8. Enforces Data Government rules
9. Supports the company's data organization
10. Enables non-IT people to work with data (*data democratization*)

Agenda

- Why Big Data and Big Data Architectures (BDA)?
- What is a BDA and what are the requirements?
- **A brief history of Data Architectures**
- Different BDAs:
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- Outlook: The Future of BDA
- Summary: 4 things to remember

Important Milestones of Data Architectures



Data stored
on tape or
disks

- 1950



Edgar F.
Codd pro-
posed the
Relational
Database
Model, SQL
was invented

- 1970



RDBMS

- IBM
- ORACLE
- MS SQL Server



on-prem
Data
Warehouses

- Terradata
- ORACLE
- SAP BW



Cloud-based
data lakes

- AWS (2006)
- GCP (2008)
- Azure (2010)
- People talking about "Big Data"



Massive
Parallel
Compute

- MapReduce (2003 - Google)
- Hadoop (2006 - Yahoo)
- Apache Spark (2009 - Berkeley)



The Modern
Data Stack

- Data Mesh (2019)
- Lakehouse (2020)

Agenda

- Why Big Data and Big Data Architectures (BDA)?
- What is a BDA and what are the requirements?
- A brief history of Data Architectures
- **Different BDAs:**
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- Outlook: The Future of BDA
- Summary: 4 things to remember

The Modern Data Stack

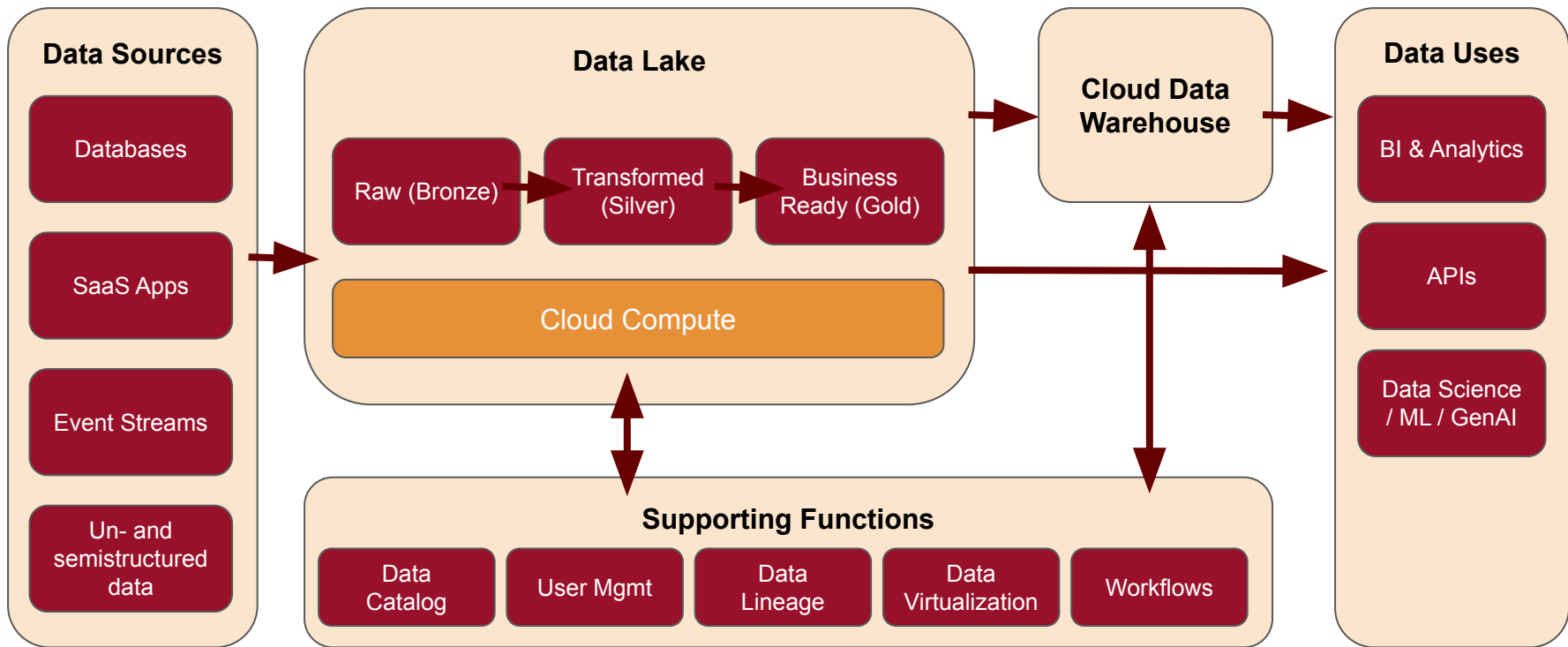
What is a "Modern Data Stack"?

Definition: ***A Modern Data Stack (MDS) is a collection of cloud-based tools and technologies used to gather, store, process, and analyze data in a scalable, efficient, and cost-effective way.***

(altexsoft <https://www.altexsoft.com/blog/modern-data-stack/>)

- Cloud-based
- Best-of-Breed
- Scalable
- Separation of Storage and Compute
- Always has a Data Lake
- Can have a Cloud-based Data Warehouse

Modern Data Stack Architecture



Characteristics of a Modern Data Stack

Pros

- Scalable
- Cost effective
(consumption based)
- Universal, works with all kind of data (tabular, semi- and unstructured data)
- Supports all kind of use cases (BI & Analytics, ML, Streaming, etc.)

Cons

- Complex
- A lot of different skills needed to build and maintain it
- Separation of Data Lake and Data Warehouse can cause problems (synchronization, data staleness, no central security and data governance)

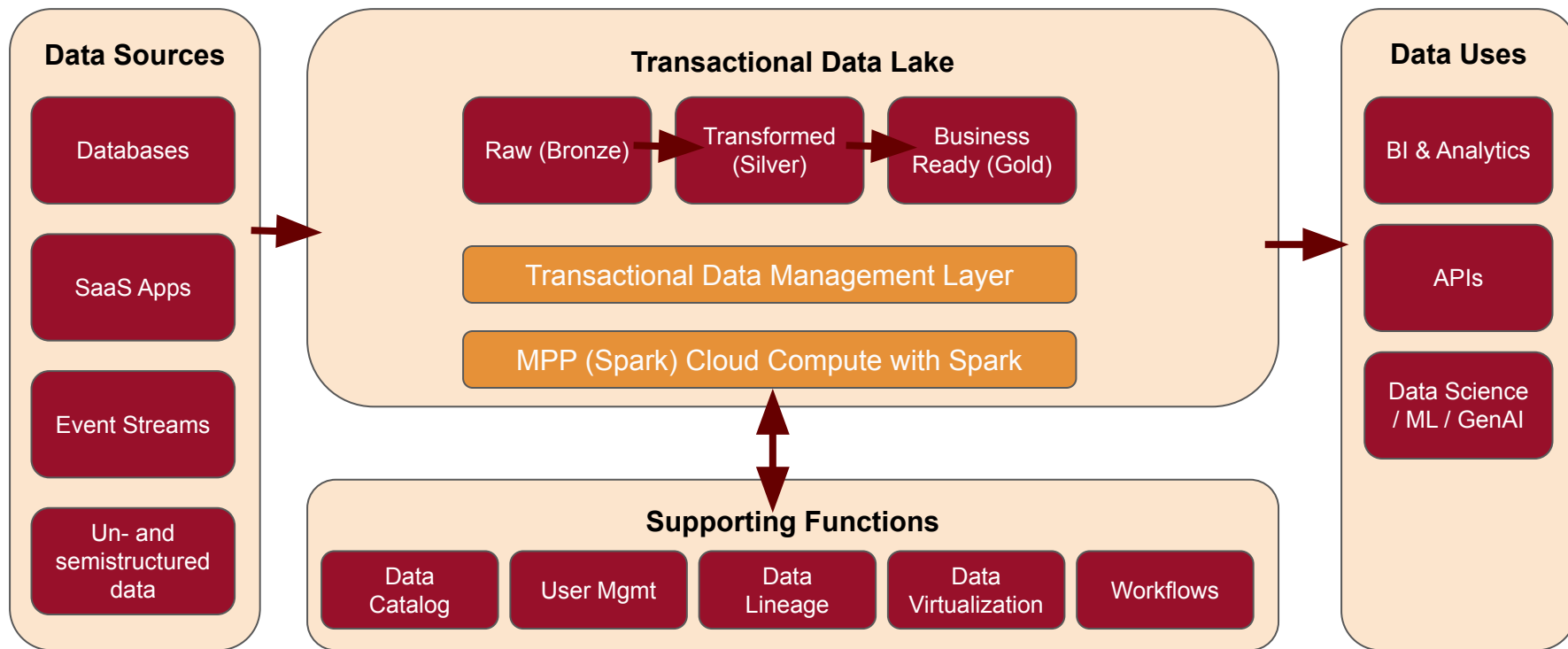
Data Lakehouse

What is a Lakehouse?

A "Lake-House" is a combination of a Data Lake and a Data Warehouse. The concept was invented by Databricks in 2020.

- Put a data management layer on top of a Data Lake with **transactional safety** (ACID properties = Atomicity, Consistency, Isolation, Durability) like Delta Lake, Apache Hudi or Apache Iceberg to get a ***Transactional Data Lake***.
- Use a cloud-based ***massive parallel processing*** framework like Spark SQL to get unlimited compute scalability and SQL performance comparable with state-of-the-art data warehouses.

Lakehouse Architecture



MPP = Massive Parallel Processing

Characteristics of a Lakehouse

Pros

- Simpler architecture compared to the Modern Data Stack
- All Data in one Data Lake makes Data Governance and Metadata Management much easier
- Can process all kinds of data
- Out-of-the-box solutions available

Cons

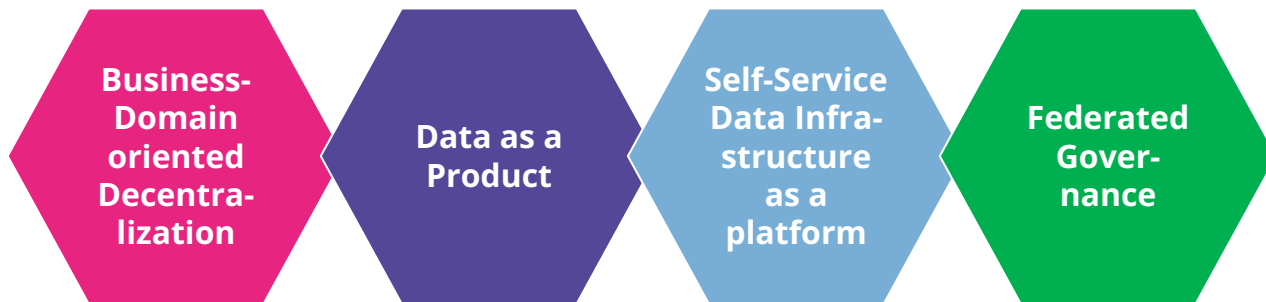
- Technology led approach, does not provide guidance to a organizational design (in particular a distributed approach for large organisations)

Data Mesh

What is a Data Mesh?

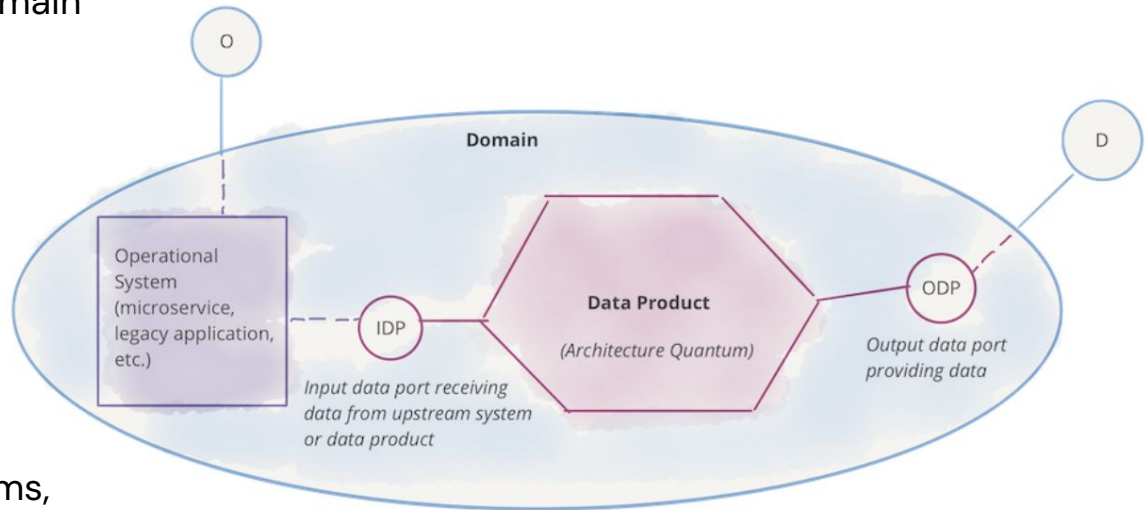
Developed 2019 by Zhamak Dehghani during her time at ThoughtWorks.

- More an organizational approach than a Big Data Architecture pattern.
- Decentralized Data organization
- Data is treated as a product



Data as a Product

- Decentralized managed by a Domain Data Product Owner
- Discoverable
- Addressable
- Secure (governed by access control)
- Trustworthy
- Interoperable
- Polyglot data ports (event streams, batch files, relational tables, graphs, etc.)



Characteristics of a Data Mesh

Pros

- Decentralized approach organized by Business Domains makes it interesting for large organizations
- IT or Analytics team is not longer a bottleneck
- Data as a Product is moving data ownership to business domains

Cons

- Data Mesh does not say anything on how to implement it (technology agnostic)
- No out-of-the-box solutions from vendors available
- Implementing a Data Mesh is foremost a business transformation project and less a technical IT project

Agenda

- Why Big Data and Big Data Architectures (BDA)?
 - What is a BDA and what are the requirements?
 - A brief history of Data Architectures
 - Different BDAs:
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- **Outlook: The Future of BDA**
 - Summary: 4 things to remember

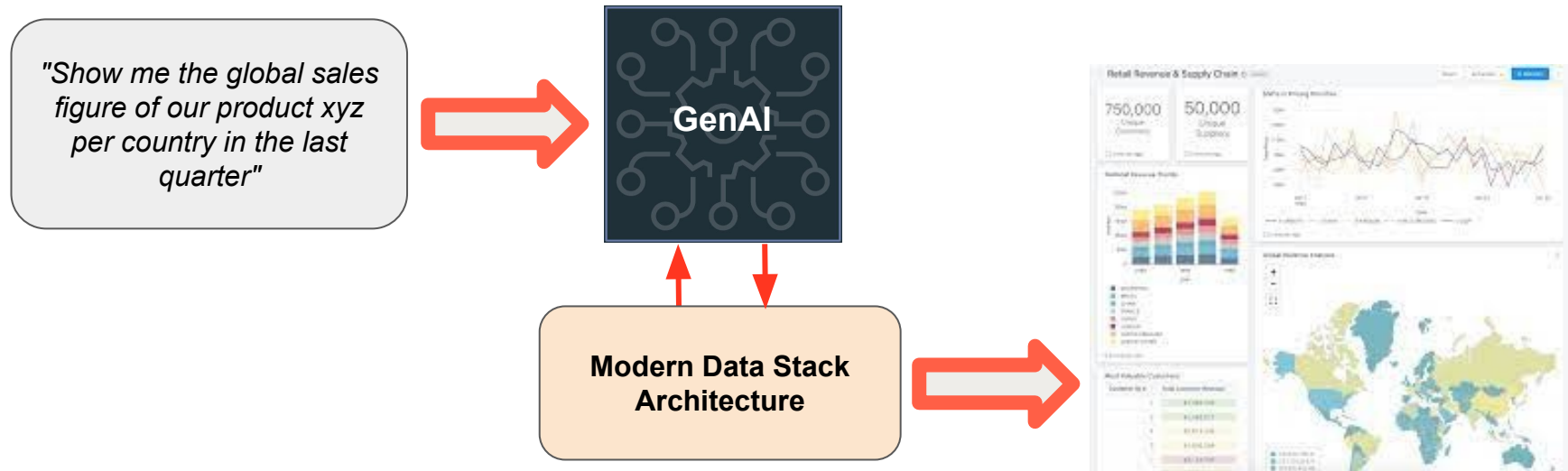
The Future of BDA: AI Capabilities will transform BDAs

Integrate AI capabilities into the Big Data Platform

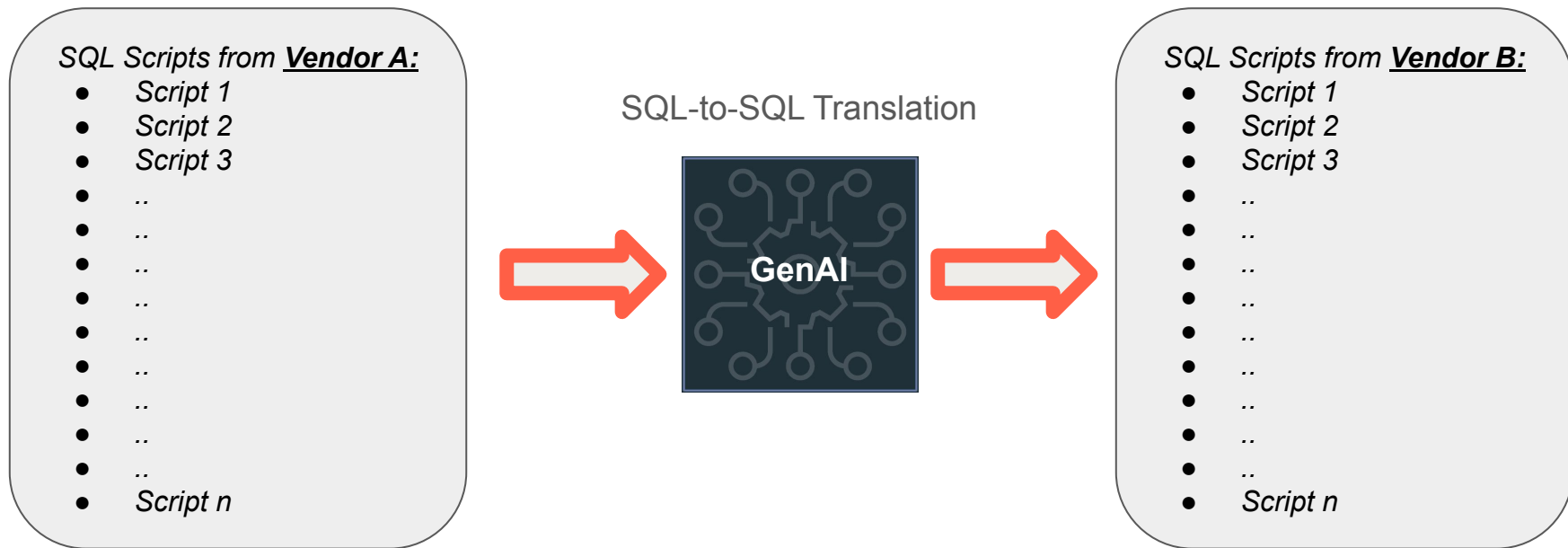
The ultimate goal is a BDA that **deeply understands the organization's data** by continuously analyzing incoming and outgoing data and training their GenAI models on it.

Use Case Example: Natural Language to SQL translation

NL-to-SQL Translation



Use Case Example: Translation of SQL dialects for DB migrations



Use Case Example: Platform Performance Optimization

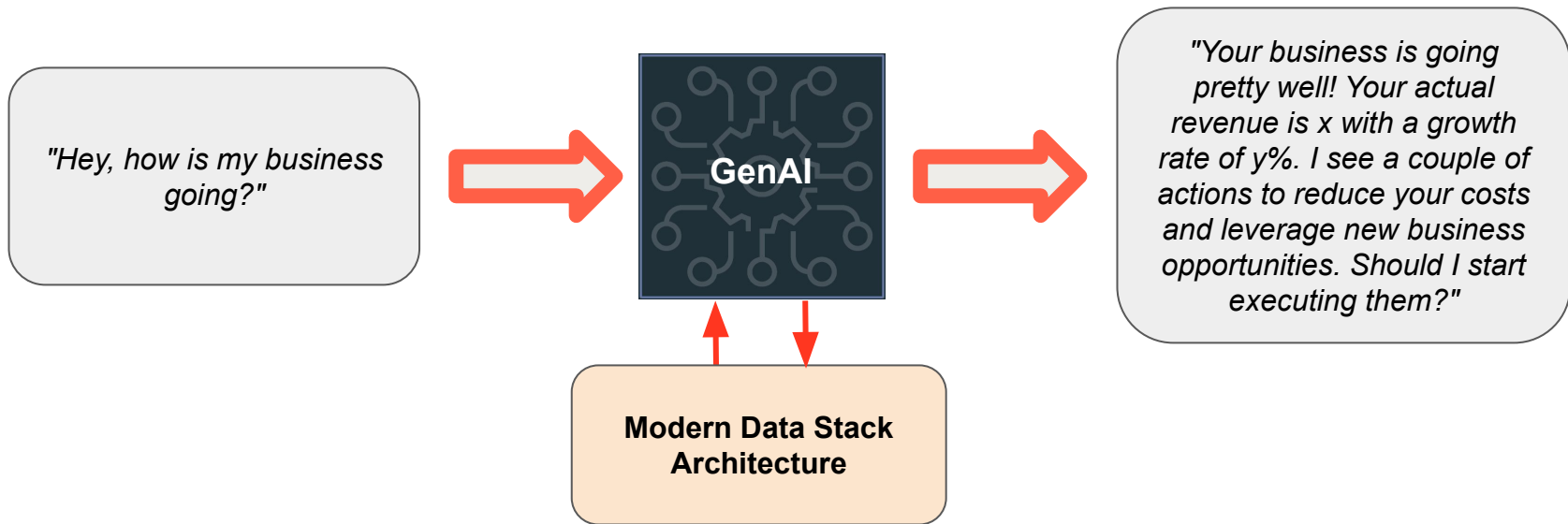
Analyzing Data
Optimization of Parameters



**Modern Data Stack
Architecture**

Optimal Performance and
Efficiency

The Vision: GenAI runs my company



Agenda

- Why Big Data and Big Data Architectures (BDA)?
- What is a BDA and what are the requirements?
- A brief history of Data Architectures
- Different BDAs:
 - The Modern Data Stack
 - Data Lakehouse
 - Data Mesh
- Outlook: The Future of BDA
- **Summary: 4 things to remember**

Summary: 4 things to remember

1. The Modern Data Stack is the state-of-the-art BDA
2. The Lakehouse is Modern Data Stack that excels with simplicity and versatility
3. Data Mesh is an organizational approach to distribute data ownership by data products along business domains
4. AI is radically changing the functionality and usage of BDA

THANK YOU!