MCKNIGHT CONSULTING GROUP

Competitive Analytic Architectures: Comparing the Data Mesh, Data Fabric, Data Lakehouse, and Data Cloud

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

A 2 time Inc. 5000 Company

linkedin.com/in/wmcknight

www.mcknightcg.com

(214) 514-1444

13th April 2023

Data Integration for Everyone and Everywhere

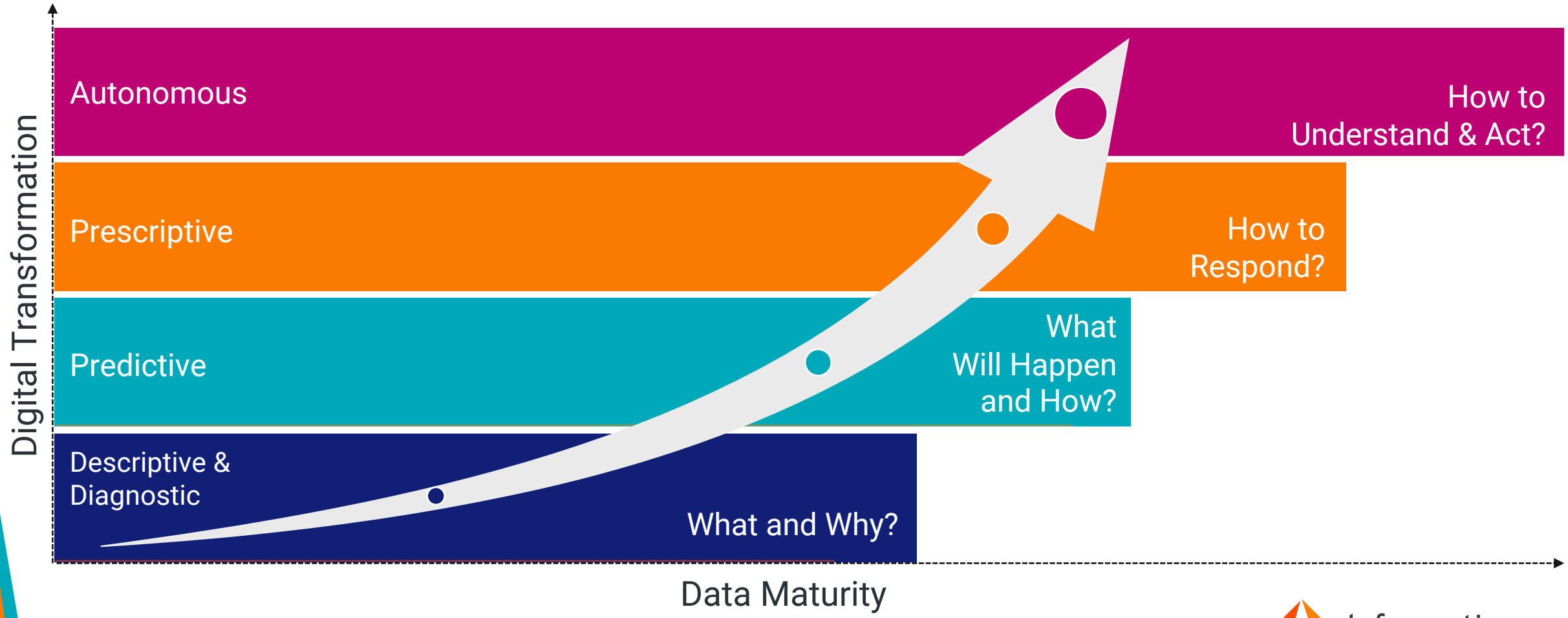
Vivin Nath

Director, Product Management

Informatica

How Businesses Want to Transform With the Power Of Data

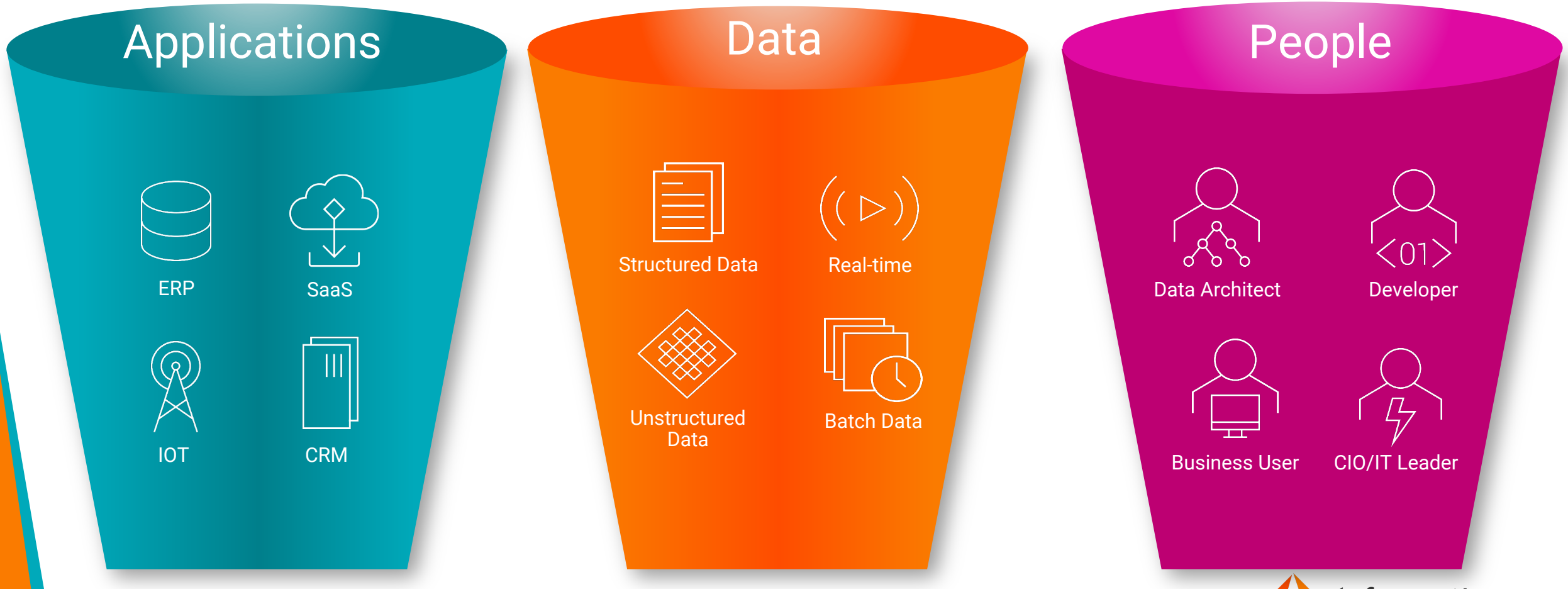
From Descriptive to Autonomous



Informatica

Only 20% of the companies are progressing in the Digital Transformation Journey

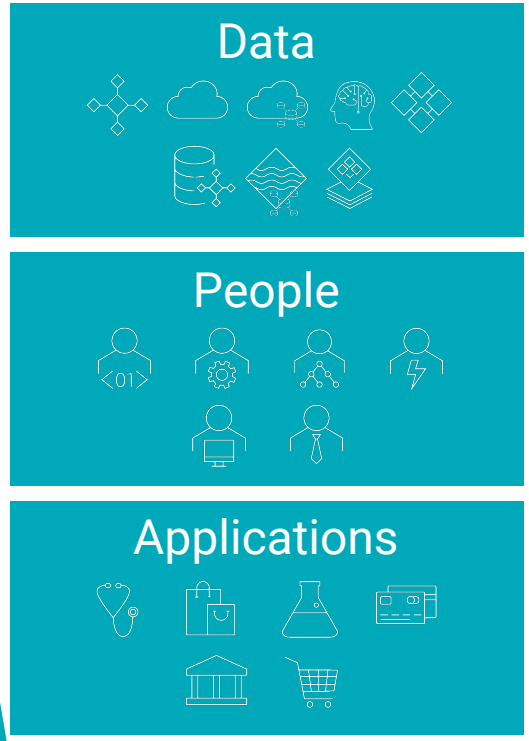
Silos in Applications, Data, and People are Creating Friction in the Organization



Informatica

Easy, Efficient, cost-Effective

Data Integration for Everyone and Everywhere



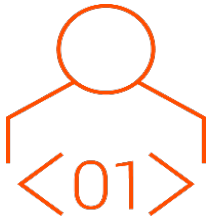
Informatica

Data Integration for Everyone

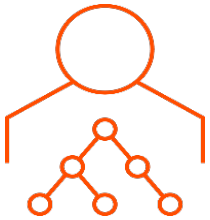
Any Data, Any Type, Any Volume, Any Latency, Any Use Case

Data/Business Analyst

Data/IT Ops

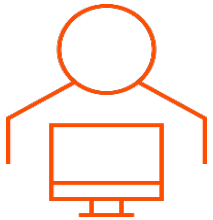


ETL Developer



Data/Software Architect

Data Engineer



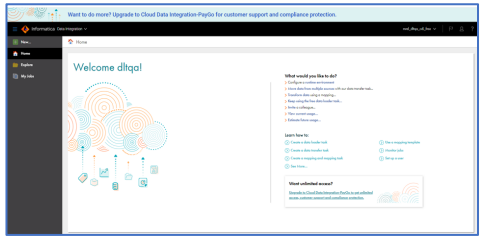
Data Scientist

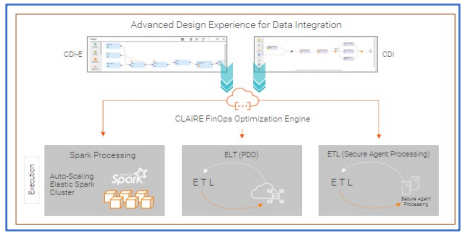
CDI-Free/PayGo

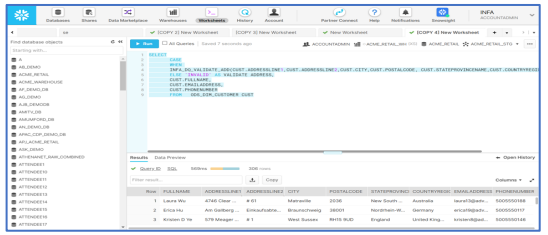
Advanced Data Integration

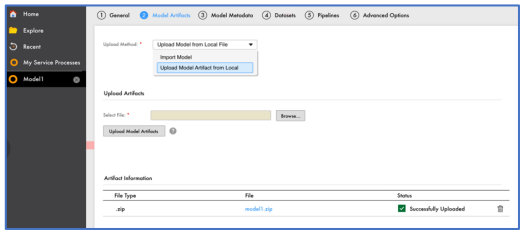
INFACore

ModelServe



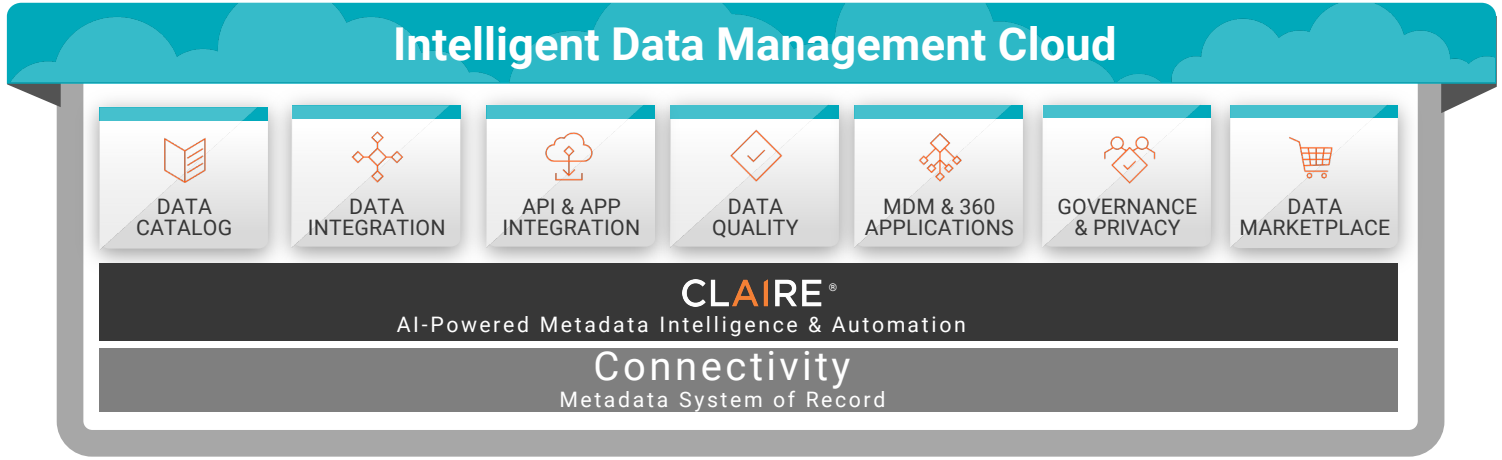




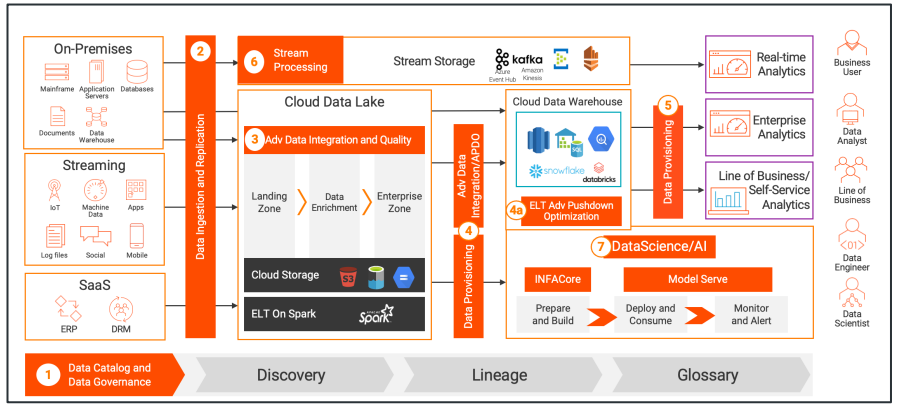


Informatica

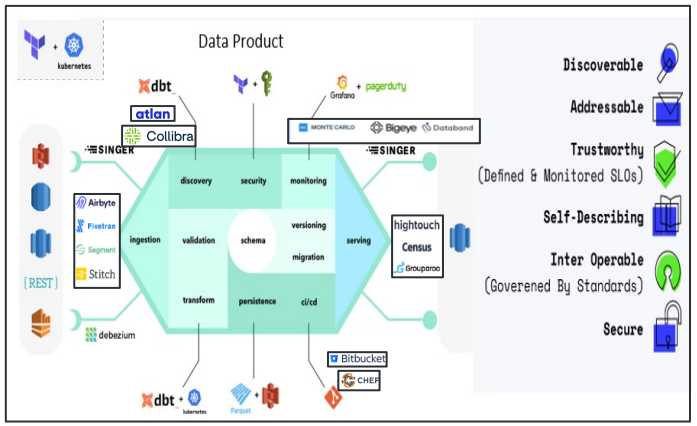
One Platform For All Data Architecture Patterns



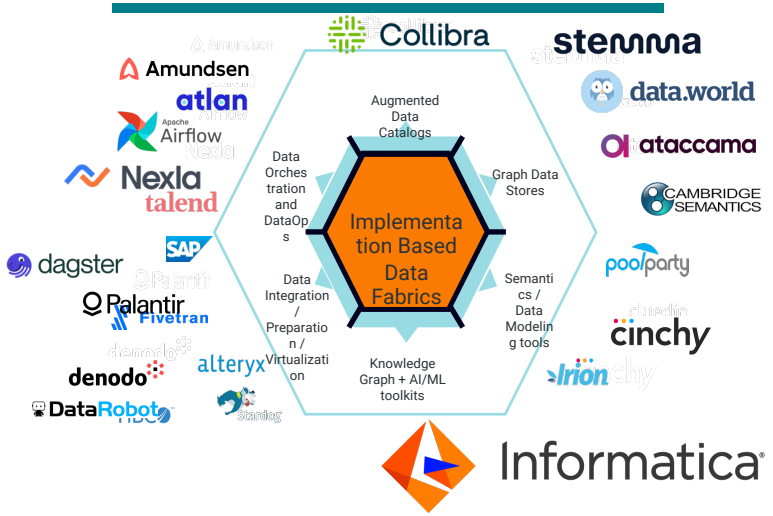
Lakehouse



Data Mesh

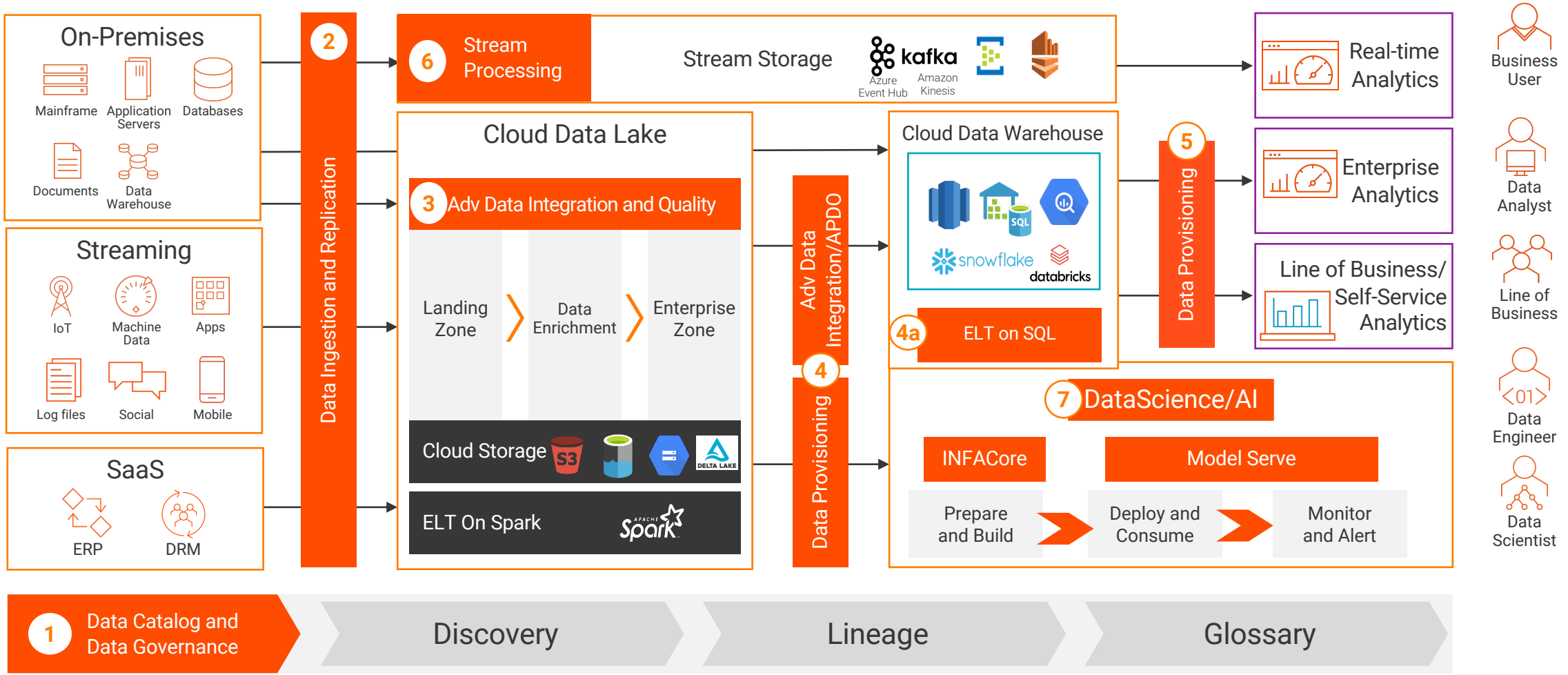


Data Fabric



Informatica

Lakehouse Architecture with IDMC



IDMC Products

Informatica

Learn More

- Free 30-day trial of Cloud Data Integration
- Research Paper: Elastic Cloud Service for Data Engineering
- White paper: 5 Reasons to go Serverless to Achieve Your Cloud Data Integration Need
- Datasheet: Informatica ModelServe - Put AI into Action
- Datasheet: Informatica INFACore - Open, Embeddable and Extensible headless data management

Thank you

vnath@informatica.com

CONSUMER PRODUCTS/RETAIL

FINANCIAL

INSURANCE/HEALTHCARE

PUBLISHING

OTHER

GOVERNMENT AND UTILITIES

EDUCATION

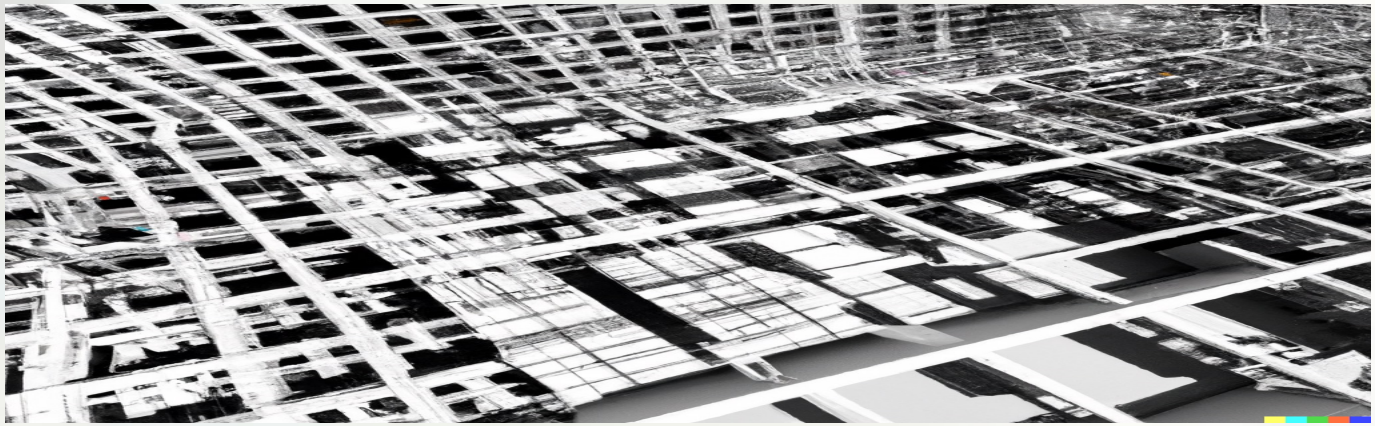
PHARMACEUTICAL

TELECOMMUNICATIONS

MCKNIGHT CONSULTING GROUP

Distributed Data Architecture Patterns

- The data lake architecture has shortcomings that lead to unfulfilled promises at scale
  - Monolithic, Centralized
  - Coupled pipeline decomposition
  - Hyper specialized ownership



Pros and Cons of Following Architectural Patterns

Pros

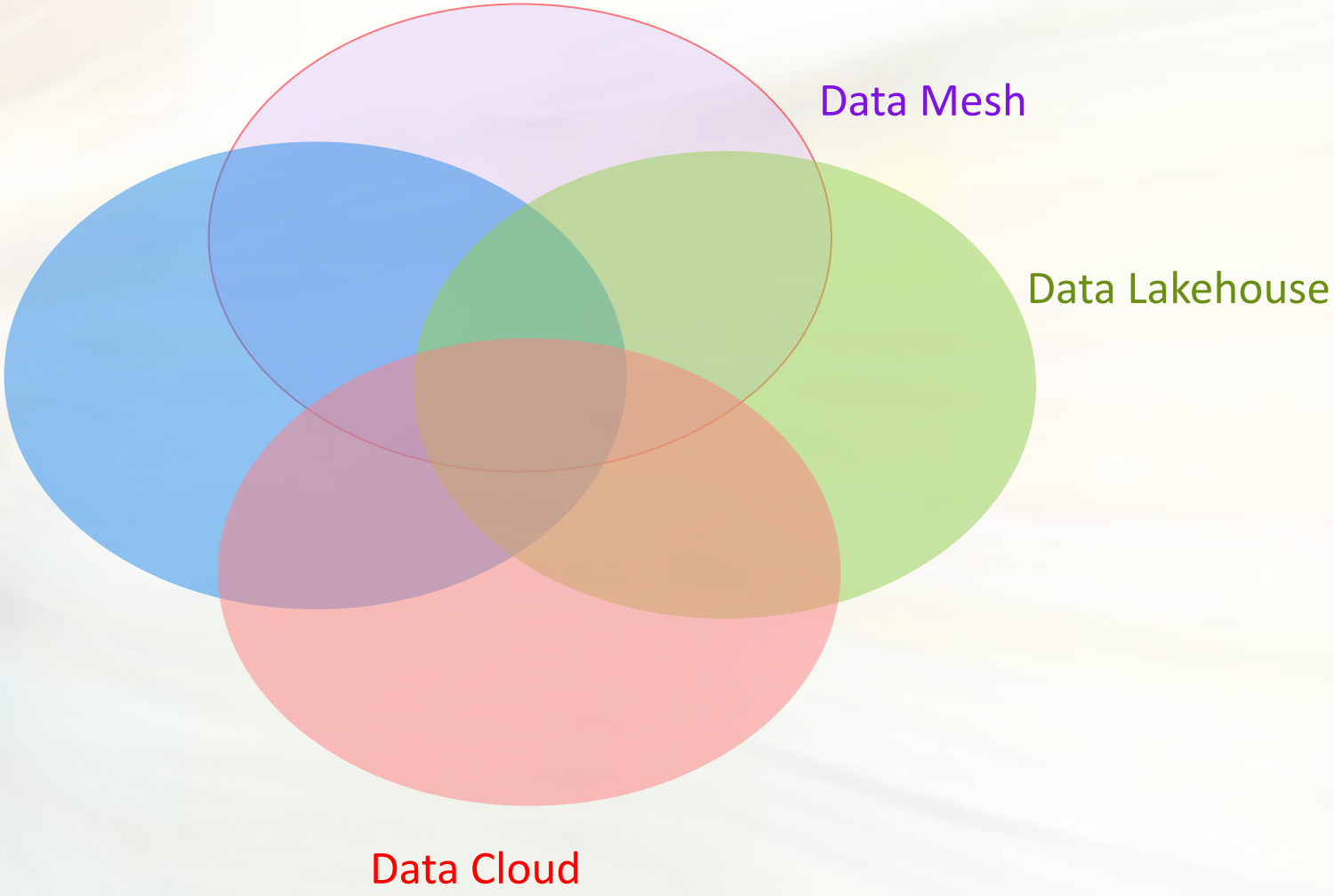
- Theoretically, it's science and has been validated
- Decisions addressed you were unaware of
- Understandable

Cons

- Can lose focus on the business priorities
- May not be right for you
- Can take longer for adherence

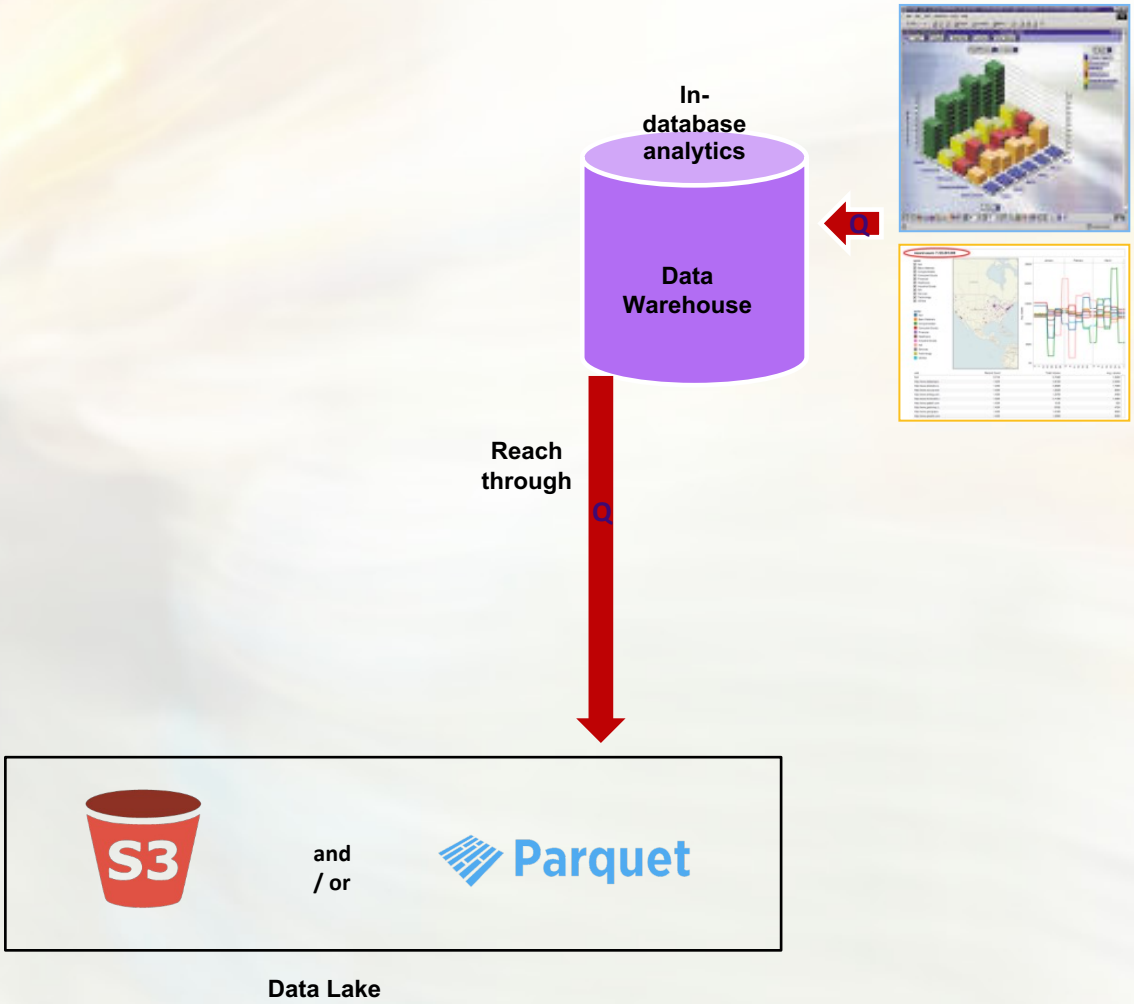
MCKNIGHT CONSULTING GROUP

These are not Mutually Exclusive



MCKNIGHT CONSULTING GROUP

Data Lakehouse



Data Lakehouse Principles

- Managing Data
- Formats that can be Accessed Easily
- Adaptable Storage
- Facilitating the Continuous Flow of Data
- Handling Varied Tasks

Benefits of a Data Lakehouse

- Administration Management
- Better Organization of Information
- Simplified Rules and Regulations
- More Cost-Efficiency

Redshift Spectrum (External Tables)

- Spectrum resides on dedicated Redshift servers that are independent of your cluster
- Scales independently and automatically
- Uses an external data catalog
  - AWS Glue, Athena data catalog, or your own Apache Hive metastore
- Can't perform UPDATE or DELETE operations
  - But can INSERT

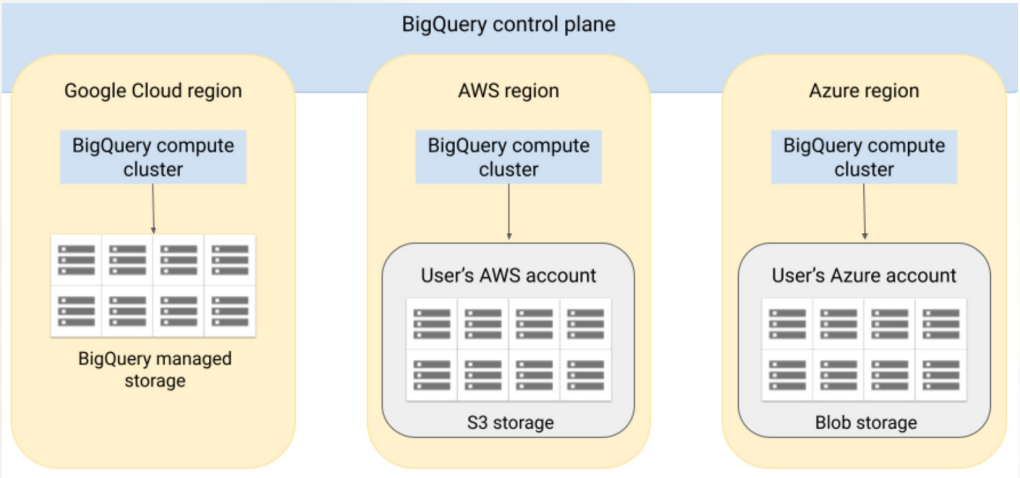
Snowflake External Tables

- Schema on read
  - If an error occurs, it skips to the next file, but still returns rows found in the current file up until the error occurred.
- Recommended 16MB – 256MB file sizes (256-512MB for Parquet)
- Delta Lake support
- Workflow:
  - CREATE STAGE > CREATE EXTERNAL TABLE > Create cloud object storage event notification > Automatic refresh

BigQuery Omni – External Tables

- Run BQ analytics on data stored in Amazon S3, Azure Blob Storage or Google Cloud Storage
- Uses BigLake tables
- Extends architecture by running BigQuery compute in the other cloud
  - No data egress charges (data doesn't move)

- Limitations:
  - 20GB per query result
  - 1TB per day
  - Only in AWS US East 1 and Az East US 2



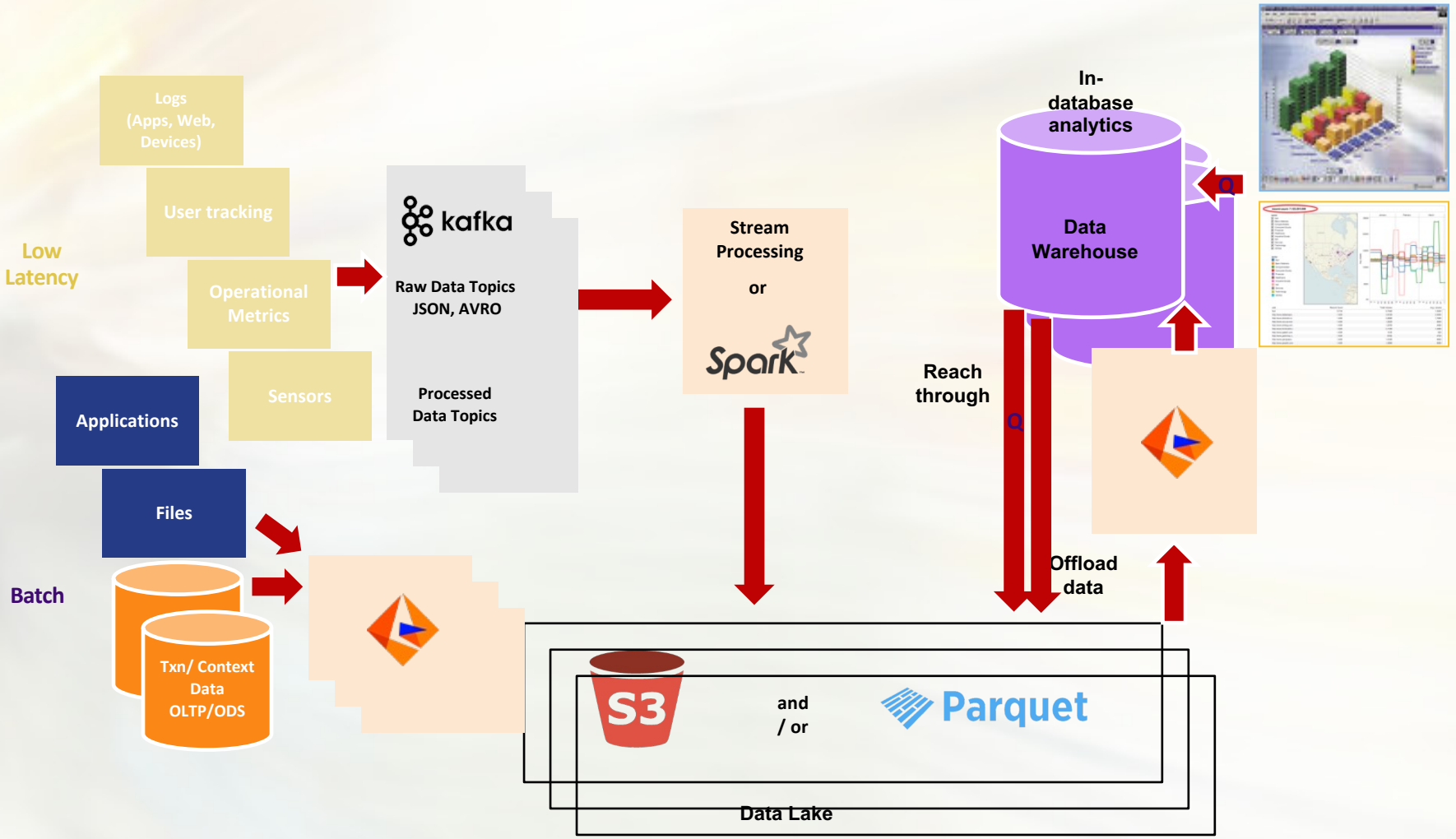
Synapse External Data Sources

- Native (new) vs. Hadoop (old)
- Hadoop requires:
  - Master key
  - Database scoped credential
  - External data source
  - External file format
  - External table
- Native uses OPENROWSET

	Hadoop	Native
Dedicated	Yes	Preview
Serverless	No	Yes
Folder partition elimination	No	Yes
File elimination	No	Yes
Custom folder paths	No	Yes*
Column mapping	By position	By name
CETAS	Yes	Yes

*except for Delta Lake tables

Data Mesh



MCKNIGHT CONSULTING GROUP

Data Mesh Principles

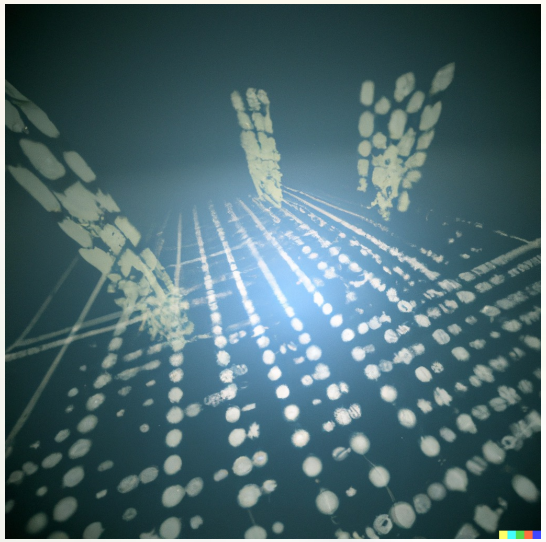
- Domain Ownership
- Data as a Product
- Self-Service Data
- Federated Governance

Benefits of a Data Mesh

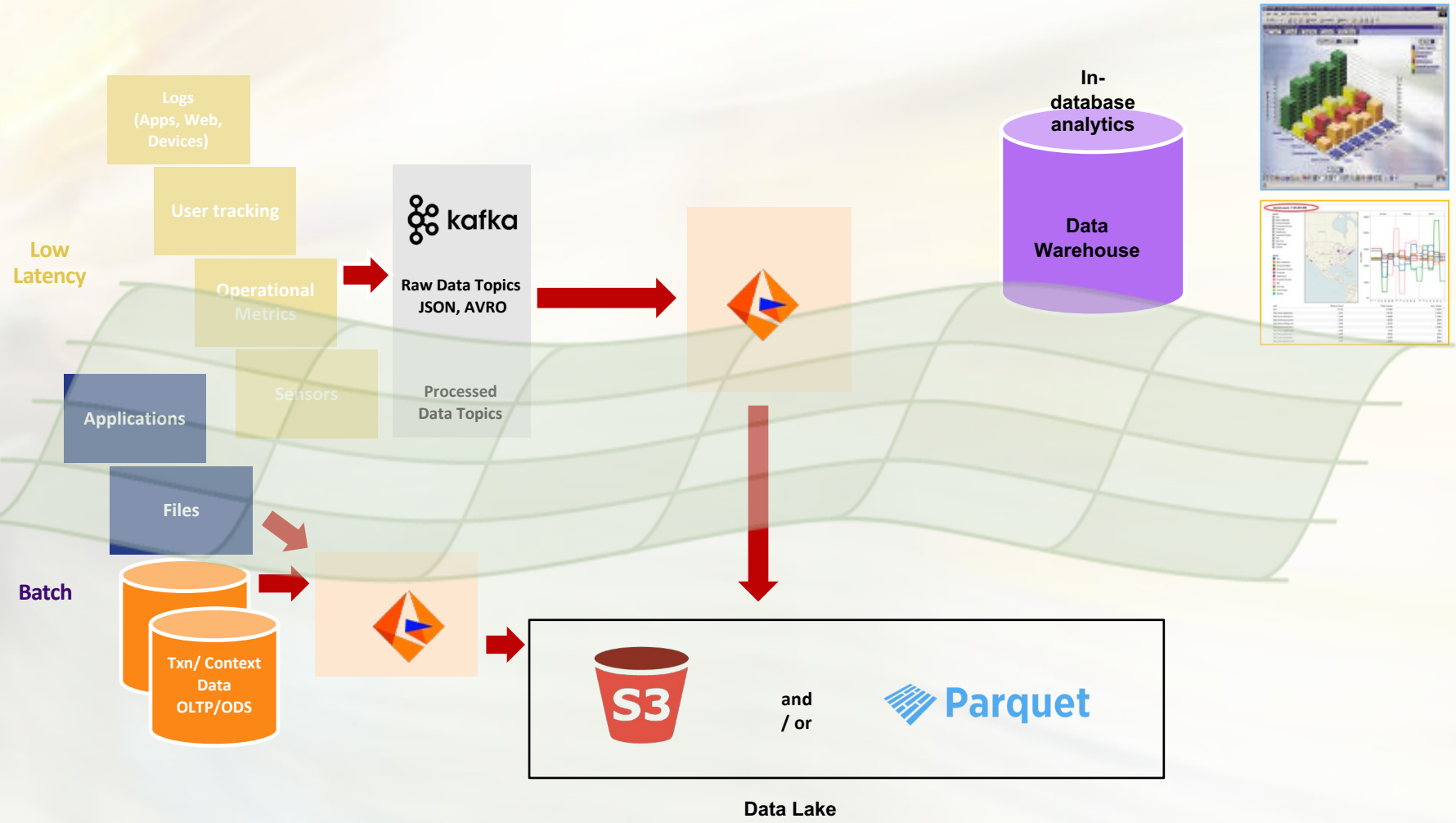
- Democratization of Data
- Cost-saving Measures
- Reduced Technical Debt
- Collaboration
- Safety and Adherence

Use Cases for Data Mesh

- Business Intelligence Dashboards
- Customer Experiences
- Machine Learning



Data Fabric


Logs
(Apps, Web,
Devices)
User tracking
Low
Latency
Operational
Metrics
Sensors
Applications
Files
Batch
Txn/ Context
Data
OLTP/ODS
kafka
Raw Data Topics
JSON, AVRO
Processed
Data Topics
In-
database
analytics
Data
Warehouse
S3
and
/ or
Parquet
Data Lake

Data Fabric Principles

- Intelligent and Automated
- Unification of Disparate Data Systems
- Access to Integrated Enterprise Data
- Scale Efficiently
- Multi-Cloud Awareness

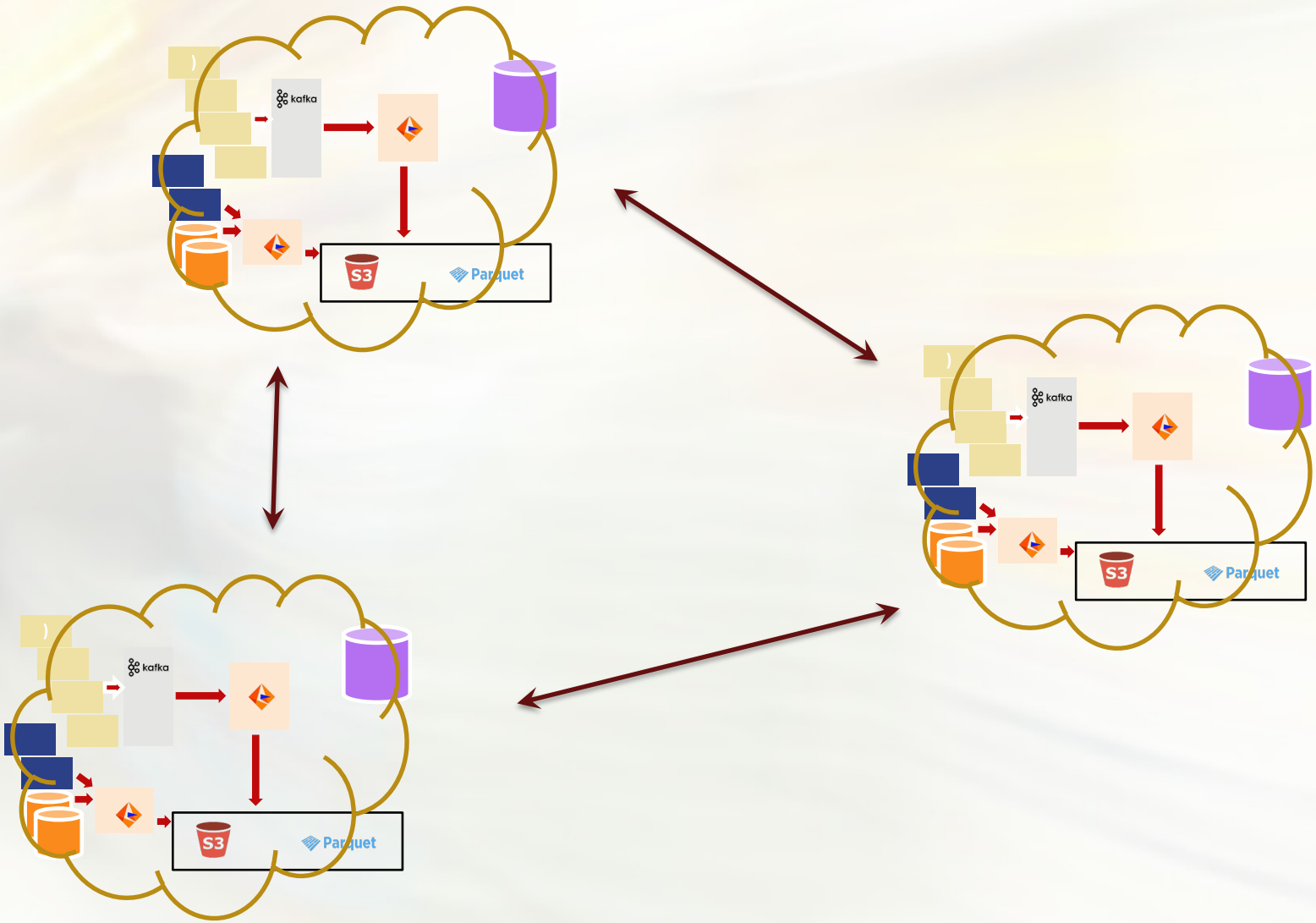
Benefits of a Data Fabric

- Integrated Intelligence
- Data Democratization
- Improved Data Security

Use Cases for Data Fabric

- Fraud Detection
- Preventative Maintenance
- Data Discovery
- Customer Profiling
- Risk Modeling

Data Cloud



MCKNIGHT CONSULTING GROUP

Summary

- The Distributed Data Architecture Patterns are not Mutually Exclusive
- The Data Lakehouse is all about Drill-Through Pathing
- The Data Mesh architecture decentralizes and decouples components by business domain
- The Data Fabric provides common shared services, connectivity, and application portability making more automation possible through patterns in metadata
- The Data Cloud allows organizations to unify and connect to a single copy of all of their data and external data

MCKNIGHT CONSULTING GROUP

Competitive Analytic Architectures: Comparing the Data Mesh, Data Fabric, Data Lakehouse, and Data Cloud

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

A 2 time Inc. 5000 Company

linkedin.com/in/wmcknight

www.mcknightcg.com

(214) 514-1444