



Architecting a Modern Data Platform

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

 /in/wmcknight

www.mcknightcg.com

(214) 514-1444



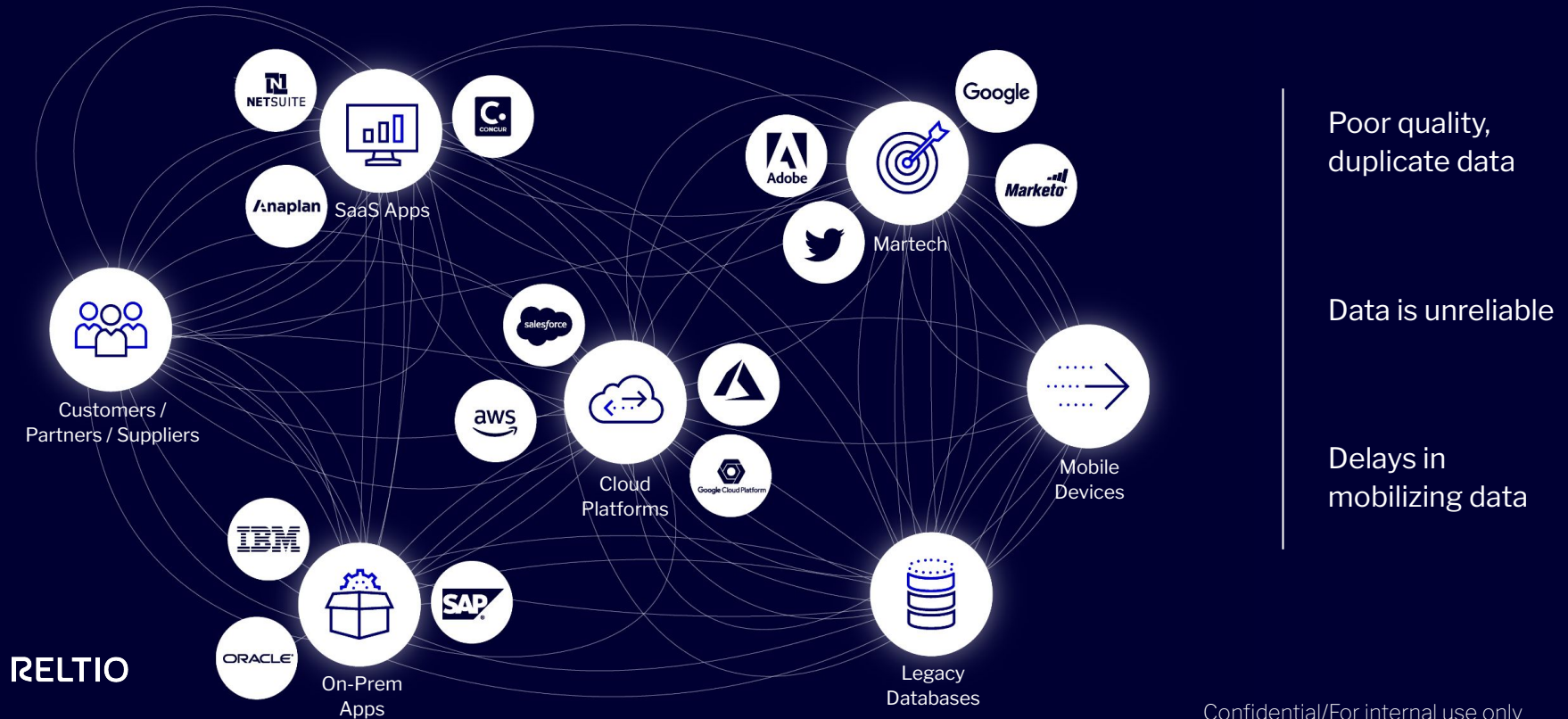
Rank	Name	Company	Role
1	William McKnight	McKnight Consulting Group	President
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

RELTIO

The background features a dynamic, abstract pattern of dots. The dots are arranged in a series of overlapping, wavy bands that create a sense of depth and movement. The color palette is primarily blue, with a gradient from light to dark, and a prominent yellow-gold color that highlights the central text. The overall effect is modern and data-oriented.

Architecting a Modern Data Platform


Siloed, fragmented data + failed digital outcomes = double blow to future growth



Data Unification & Management:

Core data as a singular, interoperable asset

Data Sources

-  Data Lake
-  Data Warehouse
-  Legacy Systems
-  SaaS Apps
-  On-prem Apps
-  3rd Party Data

Intelligence and automation

Next best actions

Propensity Models

Risk intelligence

Unification



Account



Contact



Product



Supplier



Assets



Location

Metadata, Quality, Integration

Data catalog

Data quality

Data integration

Reference data mgmt

Data Consumers

-  GPT / LLMs
-  Applications
-  BI, Analytics & Reporting
-  Digital Automation

Reltio unifies, manages, and mobilizes your core data

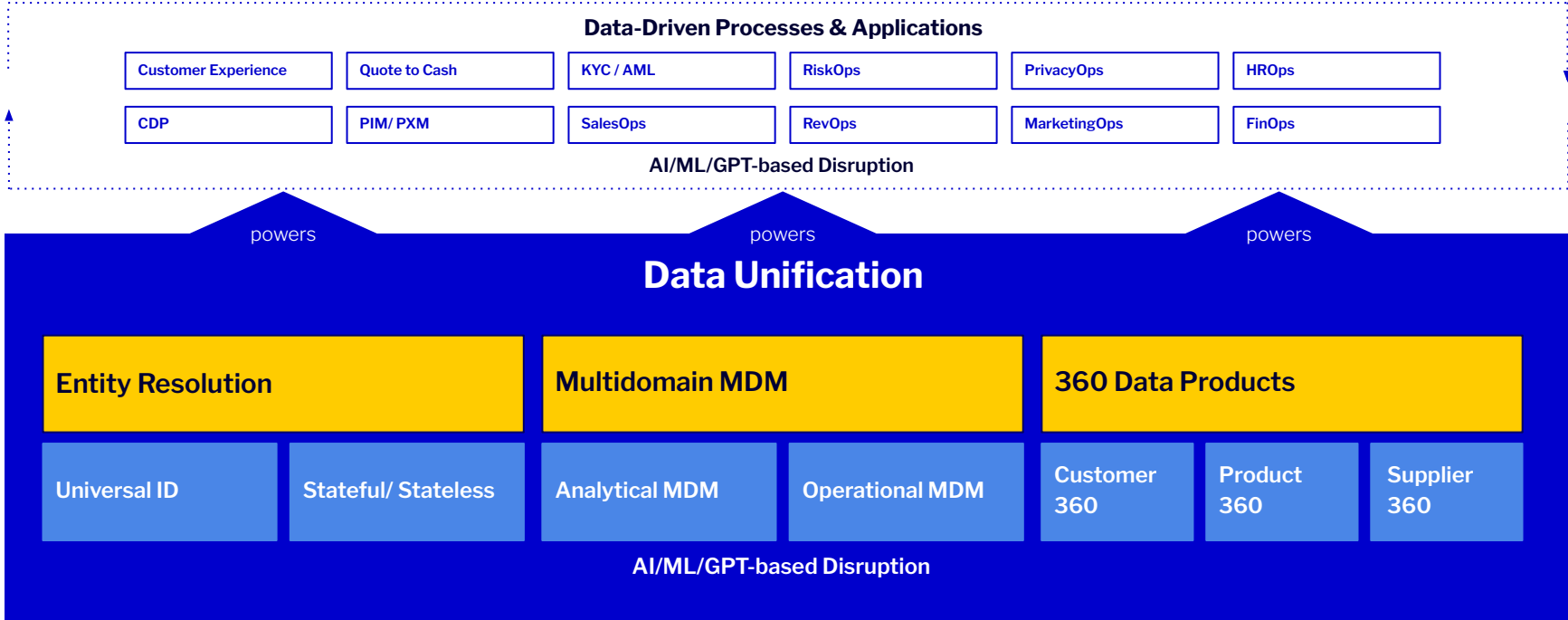


Trusted,
interoperable data

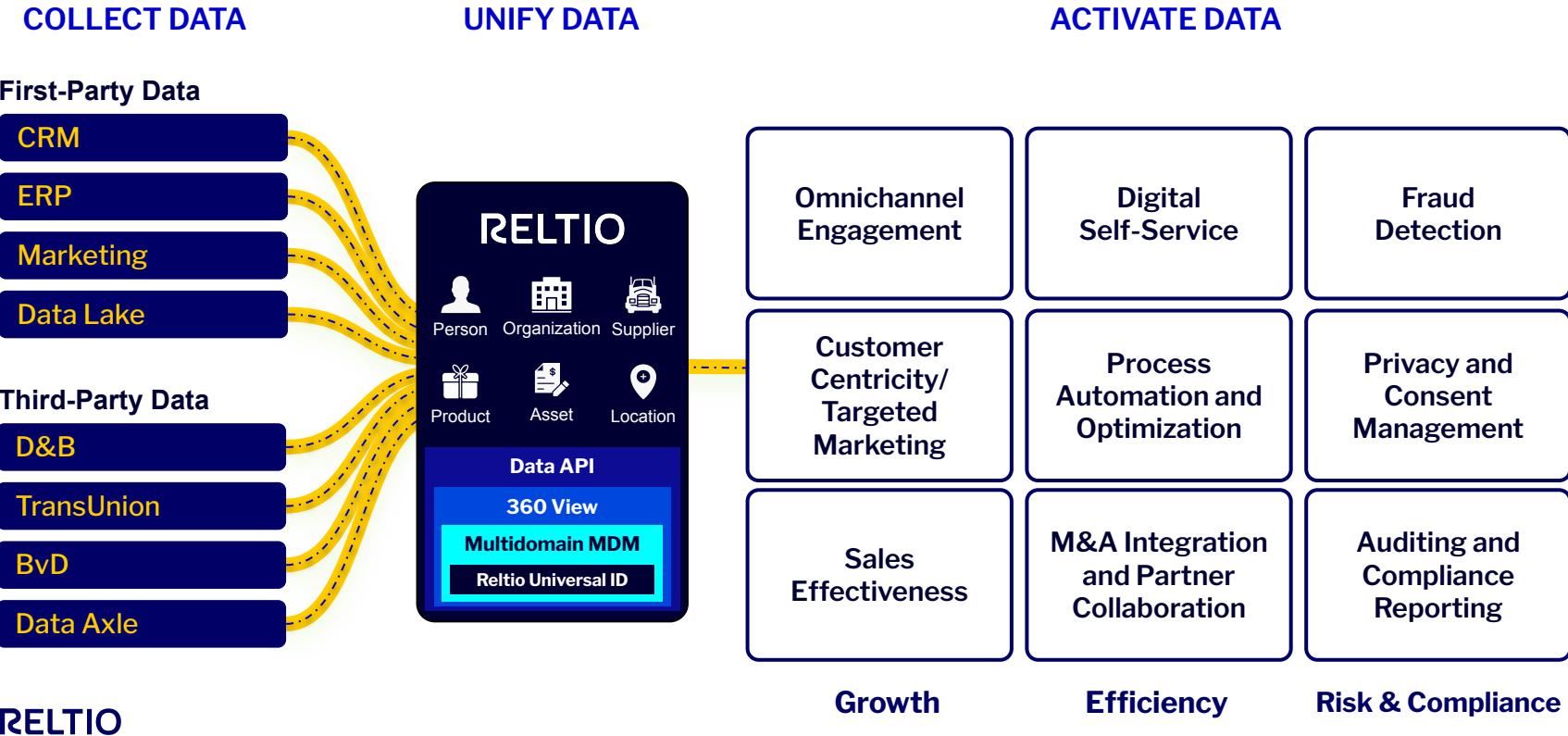
Real-time,
always on

Secure, scalable,
and flexible

Reltio meets organizations where they are in their data unification journey



Trusted, reusable data assets to active across the enterprise



Thank you

DATA DRIVEN

MODERN DATA MANAGEMENT CONFERENCE

October 7 – 9, 2024

Orlando, FL

datadriven24.com

RELTIO

McKnight Consulting Group Partial Technology Implementation Expertise

Big/Analytic/Vector/Mixed Data Management



Data Movement and APIs



Data Management



Operational/Transactional Data Management





Modern Data Platforms

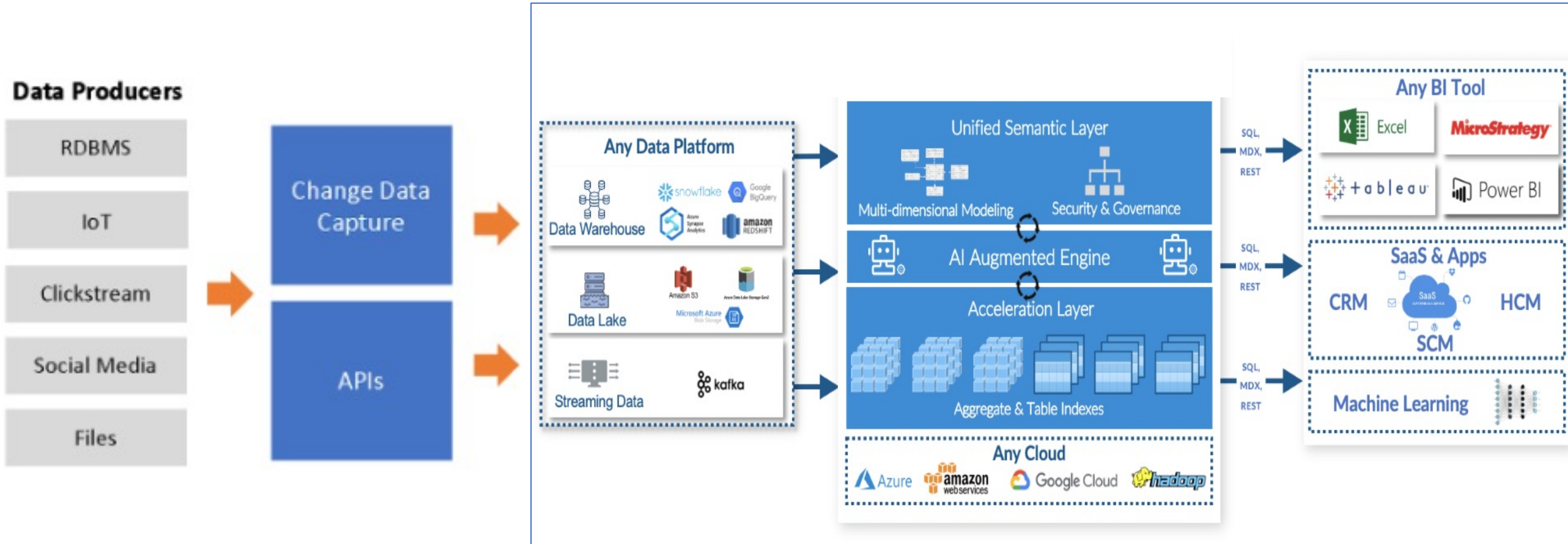


Data Platform Decisions

- OLTP vs OLAP
 - Operational vs Analytical Data
- Relational or Object storage
- Batch vs Stream
 - Lambda Architecture
- Big Data vs Not Big Data
- SMP and MPP
- Polyglot Persistence
 - Multi-Model Databases
- Single Vendor vs Best-of-Breed
- Departmental vs Enterprise-Wide
- Open Source vs Closed Source



Modern Data Platform







Modern Data Platforms - Sample Components



Category				
Data Warehouse Compute	Azure Synapse	Amazon Redshift ra3.4xlarge	Google BigQuery Annual Slots	Snowflake
Storage	Azure Synapse SQL Pool	Amazon Redshift Managed Storage	Google BigQuery Active Storage	Snowflake
Data Integration	Azure Data Factory	AWS Glue	Google Dataflow Batch	Talend Cloud Data Integration
Streaming	Azure Stream Analytics	Amazon Kinesis	Google Dataflow Streaming	Kafka Confluent Cloud
Data Exploration	Azure Synapse	Amazon Redshift Spectrum	Google BigQuery On-Demand	Snowflake
Data Lake	Azure HDInsight	Amazon EMR	Google Dataproc	Cloudera Data Hub + S3
Business Intelligence	Power BI Professional	Amazon Quicksight	Google BigQuery BI Engine	Tableau
Data Science and Machine Learning	Azure Machine Learning	Amazon SageMaker	Google BigQuery ML	Amazon SageMaker
Identity Management	Azure Active Directory P1	Amazon IAM	Google Cloud IAM	Amazon IAM
Data Catalog	Azure Purview	AWS Glue Data Catalog	Google Data Catalog	Alation Data Catalog

Data Warehouse Compute

- **Core of the Analytics Stack:** Dedicated compute represents the data warehouse itself, the heart of the analytics stack.
- **Separate Architecture:** Modern cloud data warehouses require separate compute and storage architecture.
- **Independent Scaling:** Scaling compute and storage independently is an industry standard, allowing for optimized resource allocation.
- **Cost Component:** This section focuses on the costs associated with running the compute portion of the data warehouse.

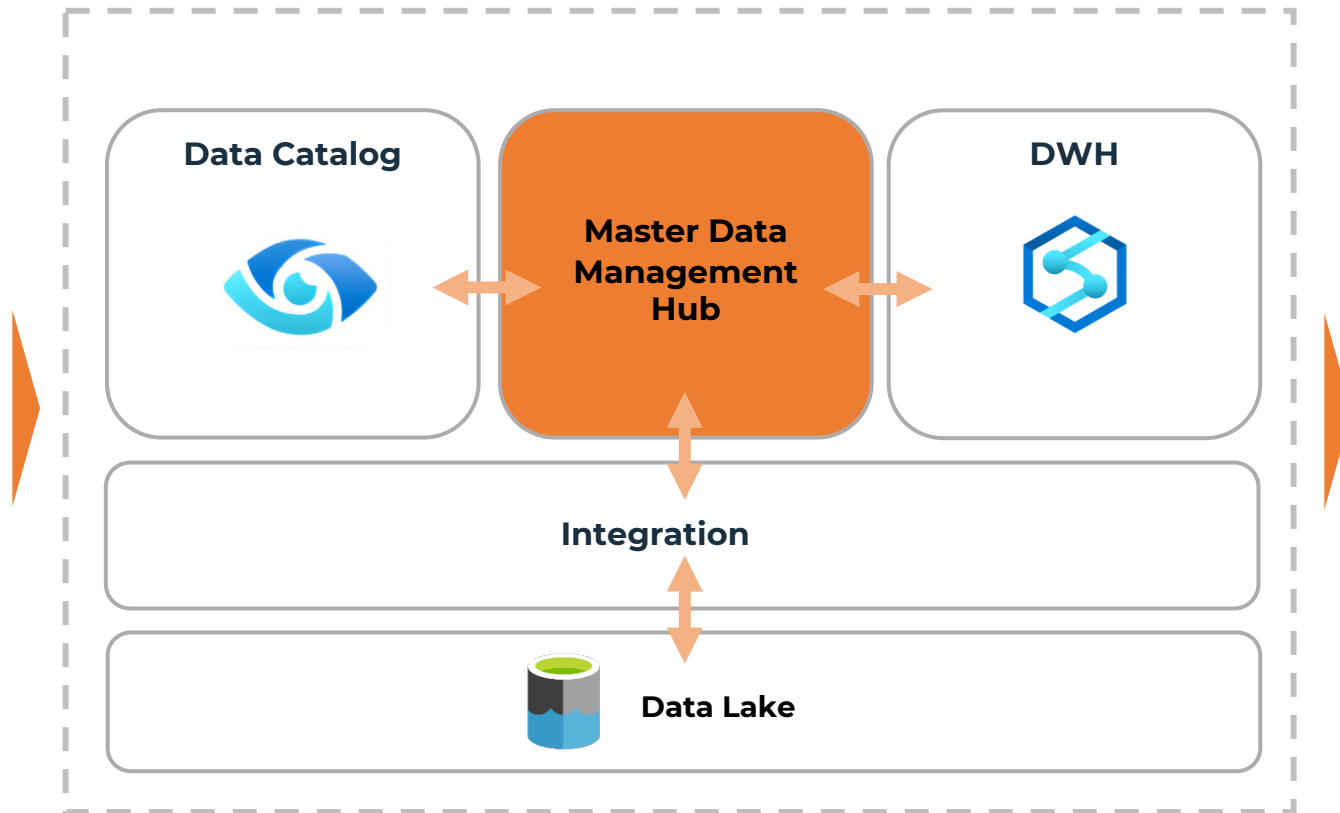
	<i>Vendor Offering</i>	<i>Pricing Used</i>
	Azure Synapse Analytics Workspace	Pay as you go (\$1.20/hour per 100 DWU) ²
	Amazon Redshift RA3	1-year commitment all-upfront (\$8.61 effective hourly) ³
	Google BigQuery	Annual slot commitment (\$1,700 per 100 slots) ⁴
	Snowflake Computing	Enterprise+ (\$4.00 per hour per credit) ⁵

Storage

- **On-premises storage:** Provides direct control but requires hardware management and maintenance. (e.g., HDDs, SSDs)
- **Cloud storage:** Scalable and cost-effective, offering various deployment models (public, private, hybrid)
- **Network-attached storage (NAS):** Centralized storage accessible across a network, ideal for file sharing
- **Storage area network (SAN):** High-performance storage for mission-critical applications



Master Data Management



Enterprise Subject Areas





- Customer
- Employee
- Partner
- Patient
- Supplier
- Product
- Bill of Materials
- Assets
- Equipment
- Media
- Geography
- Citizen
- Agencies
- Branches
- Facilities
- Franchises
- Stores
- Account
- Certifications
- Contracts
- Financials
- Policies
- Weather

Data Integration

- ETL vs ELT
- Reverse ETL
- **Azure Data Factory (ADF):** We considered integration runtime pricing and Data Integration Unit (DIU) utilization. Pricing details for DIU compute power are not publicly available.
- **AWS Glue:** We based costs on Data Processing Units (DPU) with compute power of 4 vCPU and 16 GB of memory per DPU.
- **Google Dataflow:** Costs are based on worker-hours, with a default worker offering 1 vCPU and 3.75 GB memory.
- **Snowflake:** Requires a third-party solution like Talend Cloud Data Integration, priced per user per year, with additional cloud vendor charges for virtual machines (VMs) to run Talend.

<https://tinyurl.com/McKnightDI>



Vendor Offering	Pricing Used
 Azure Data Factory (ADF)	\$0.25 per DIU-hour + \$1.00 per 1,000 activity runs
 AWS Glue	\$0.44 per DPU-hour
 Google Dataflow (Batch)	\$0.0828 per worker-hour
 Talend Cloud Data Integration	\$12,000 per user per year + compute (Azure VM E16a v4 at \$1.008 per hour)

Streaming

- **Emerging Software Category:** Data streaming is a new approach for processing data in real-time.
- **Apache Kafka and Managed Kafka Services:** Numerous vendors offer Kafka platforms and cloud-based services for easier deployment and management.
- **Stream Processing Ecosystem:** A wide range of complementary stream processing engines like Apache Flink and SaaS solutions have emerged to handle the processed data streams.
- **Competitive Landscape:** Technologies like Pulsar and Redpanda are vying for market share alongside established solutions.



Data Science and Machine Learning



Focus: Predictive and prescriptive analytics using machine learning and AI techniques.



Benefits: Streamline processes for efficiency and transparency.

Free up data scientists from manual tasks.

Support the entire data science lifecycle:

- Data preparation
- Model training
- Feature engineering
- Testing
- Deployment



Integration: Integrate with common frameworks for faster development.



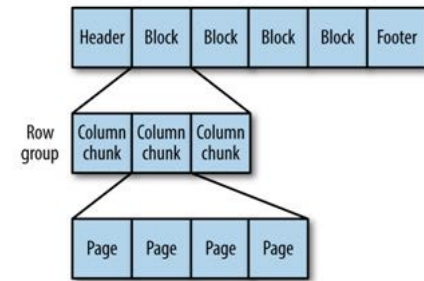
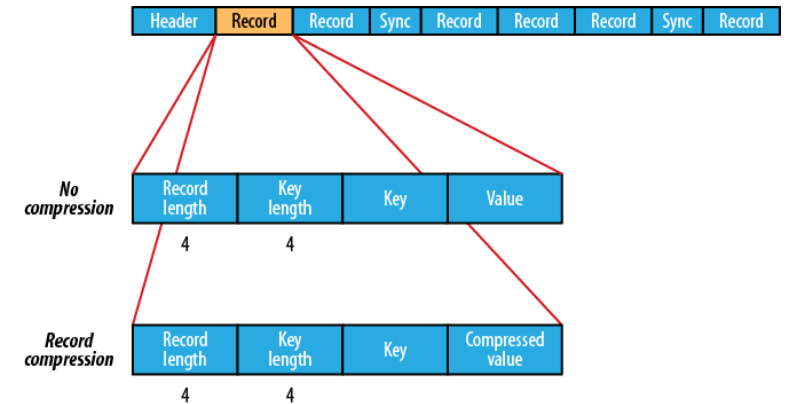
Security and Governance: Adhere to enterprise data governance and security standards.







Scalability: Enable organizations to scale data science efforts for growing AI needs.

Data Lake

- Common & centralized storage for the enterprise
- No defined data model into which the data is formed
- No relationships between the datasets
- Historical data retention
- All data formats
- For big data
- Analytical processing
- Data scientists and analysts
- Less governance/quality than data warehouse
 - Focus: Ingestion



	Vendor Offering	Pricing Used
	Azure Synapse Serverless	\$5 per TB-scanned
	Amazon Redshift Spectrum	\$5 per TB-scanned + compute (\$8.61 effective hourly) ²²
	Google BigQuery	\$5 per TB-scanned (On demand rate) ²³
	Snowflake	Enterprise+ (\$4.00 per hour per credit) ²⁴

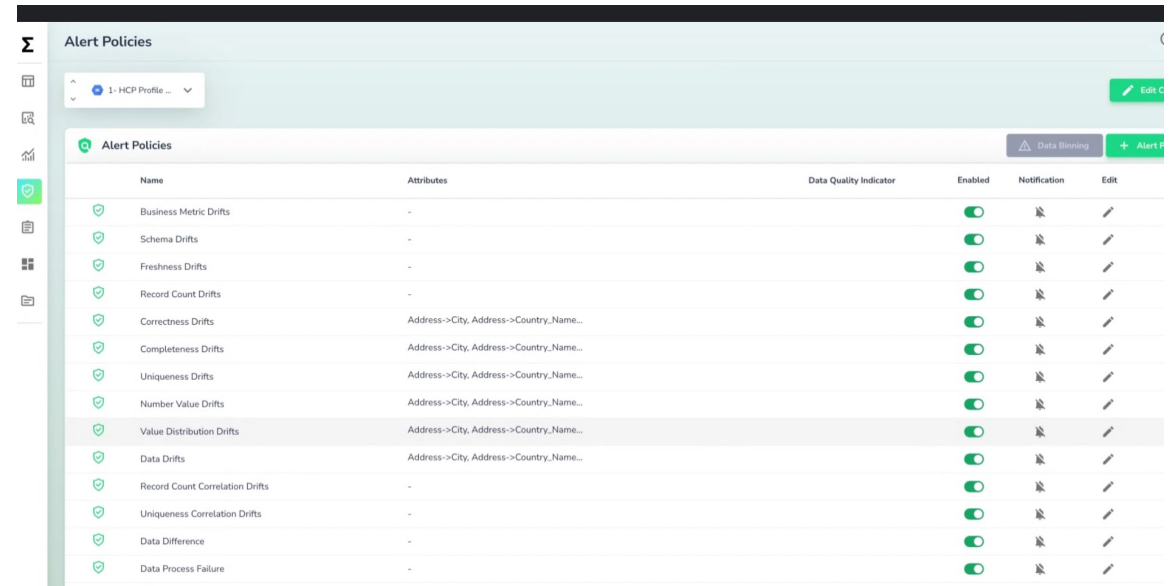
Data Governance

- Data Catalog
- DevOps
- MLOps
- Security information and event management



Data Observability

- Improved data quality (discoverable, available, usable, governable, high-quality)
- Faster troubleshooting and issue resolution
- Reduced data downtime and increased reliability
- Enhanced ability to leverage data for business goals
- Data observability is broader than data quality, encompassing data in motion and at rest, while traditional data quality focuses on data at rest.
- Relevant for data pipelines, repositories, and various deployment environments (distributed, edge, on-premises, hybrid, multi/polycLOUD)
- Leverages automation, AIOps, predictive analytics, and knowledge representation for core functionalities.



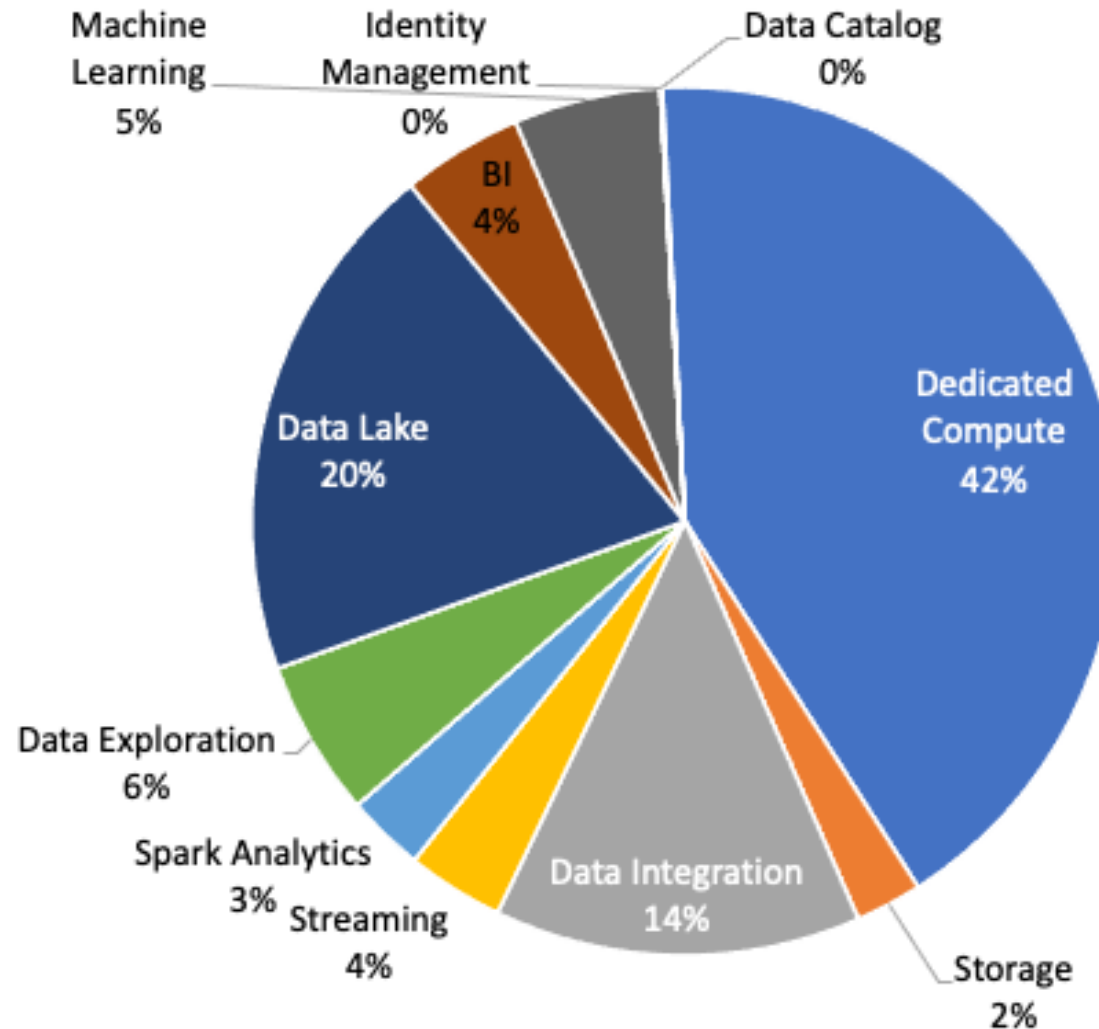
The screenshot displays the 'Alert Policies' configuration page. At the top, there is a dropdown menu for '1- HCP Profile ...' and an 'Edit' button. Below this is a table of alert policies. The table has columns for Name, Attributes, Data Quality Indicator, Enabled, Notification, and Edit. The 'Value Distribution Drifts' policy is highlighted in grey.

Name	Attributes	Data Quality Indicator	Enabled	Notification	Edit
Business Metric Drifts	-		<input checked="" type="checkbox"/>		
Schema Drifts	-		<input checked="" type="checkbox"/>		
Freshness Drifts	-		<input checked="" type="checkbox"/>		
Record Count Drifts	-		<input checked="" type="checkbox"/>		
Correctness Drifts	Address->City, Address->Country_Name...		<input checked="" type="checkbox"/>		
Completeness Drifts	Address->City, Address->Country_Name...		<input checked="" type="checkbox"/>		
Uniqueness Drifts	Address->City, Address->Country_Name...		<input checked="" type="checkbox"/>		
Number Value Drifts	Address->City, Address->Country_Name...		<input checked="" type="checkbox"/>		
Value Distribution Drifts	Address->City, Address->Country_Name...		<input checked="" type="checkbox"/>		
Data Drifts	Address->City, Address->Country_Name...		<input checked="" type="checkbox"/>		
Record Count Correlation Drifts	-		<input checked="" type="checkbox"/>		
Uniqueness Correlation Drifts	-		<input checked="" type="checkbox"/>		
Data Difference	-		<input checked="" type="checkbox"/>		
Data Process Failure	-		<input checked="" type="checkbox"/>		



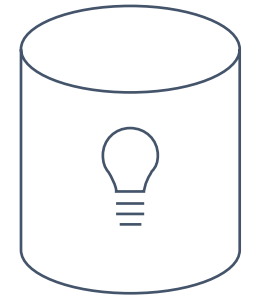
TCO of a Modern Data Platform

Sample Stack Cost Breakout



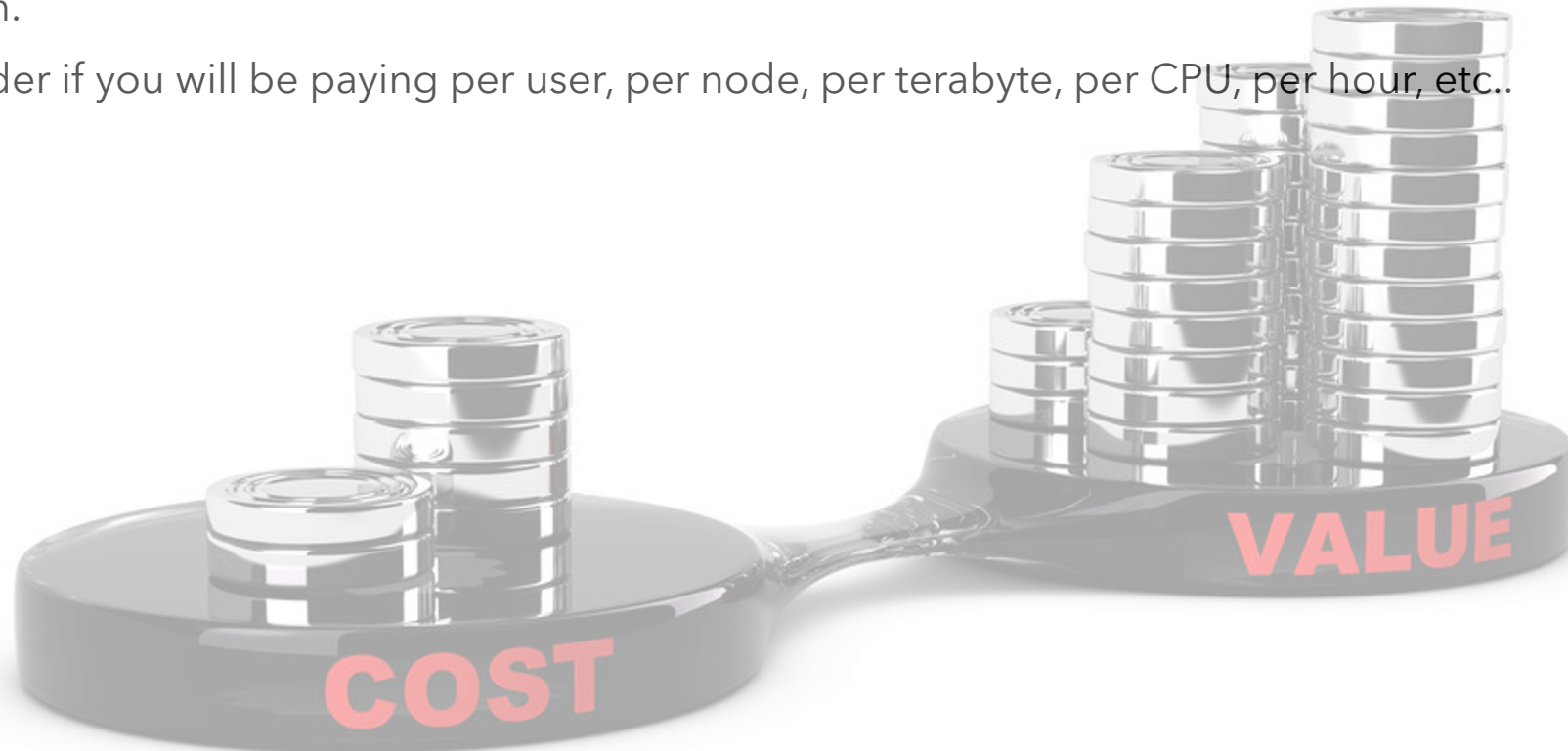
Cost Predictability and Transparency

- The cost profile options for cloud databases are straightforward if you accept the defaults for simple workload or proof-of-concept (POC) environments
- Initial entry costs and inadequately scoped environments can artificially lower expectations of the true costs of jumping into a cloud data warehouse environment.
- For some, you pay for compute resources as a function of **time**, but you also choose the hourly rate based on certain **enterprise features** you need.
- With some platforms, you pay for **bytes processed** and the underlying architecture is unknown. The environment is scaled automatically without affecting price. There is also a cost-per-hour flat rate where you would need to calculate how long it would take to run your queries to completion to predict costs.
- Customers need to analyze current workloads, performance, and concurrency and project those into realistic pricing in alternative platforms.



Cost Consciousness and Licensing Structure

- Be on the lookout for cost optimizations like not paying when the system is idle, compression to save storage costs, and moving or isolating workloads to avoid contention.
- Look for the ability to directly operate on compact open file formats Parquet and ORC
- Also, costs can spin out of control if you have to pay a separate license for each deployment option or each machine learning algorithm.
- Finally, also consider if you will be paying per user, per node, per terabyte, per CPU, per hour, etc..

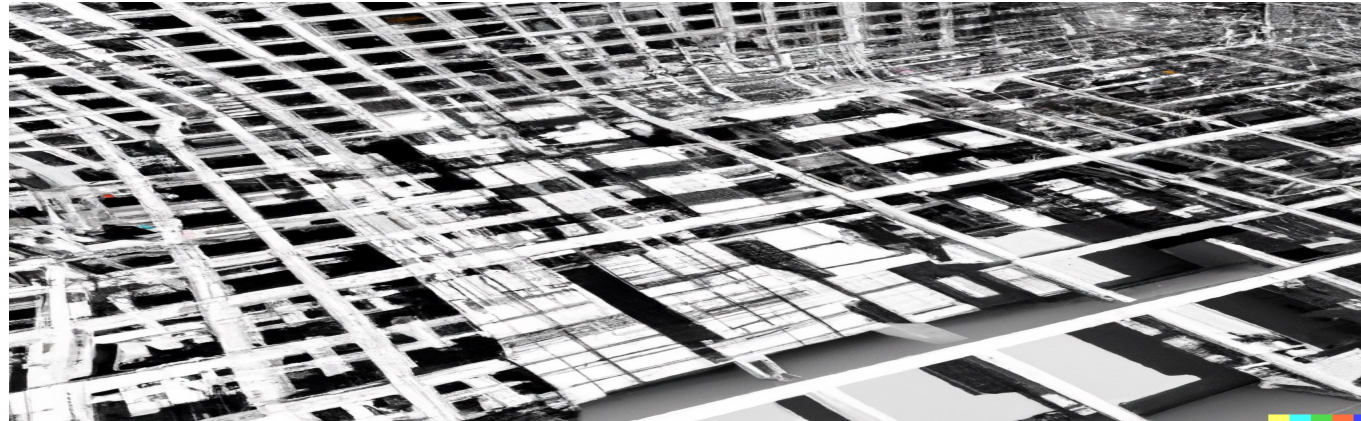




Distributed Data Architecture Patterns

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

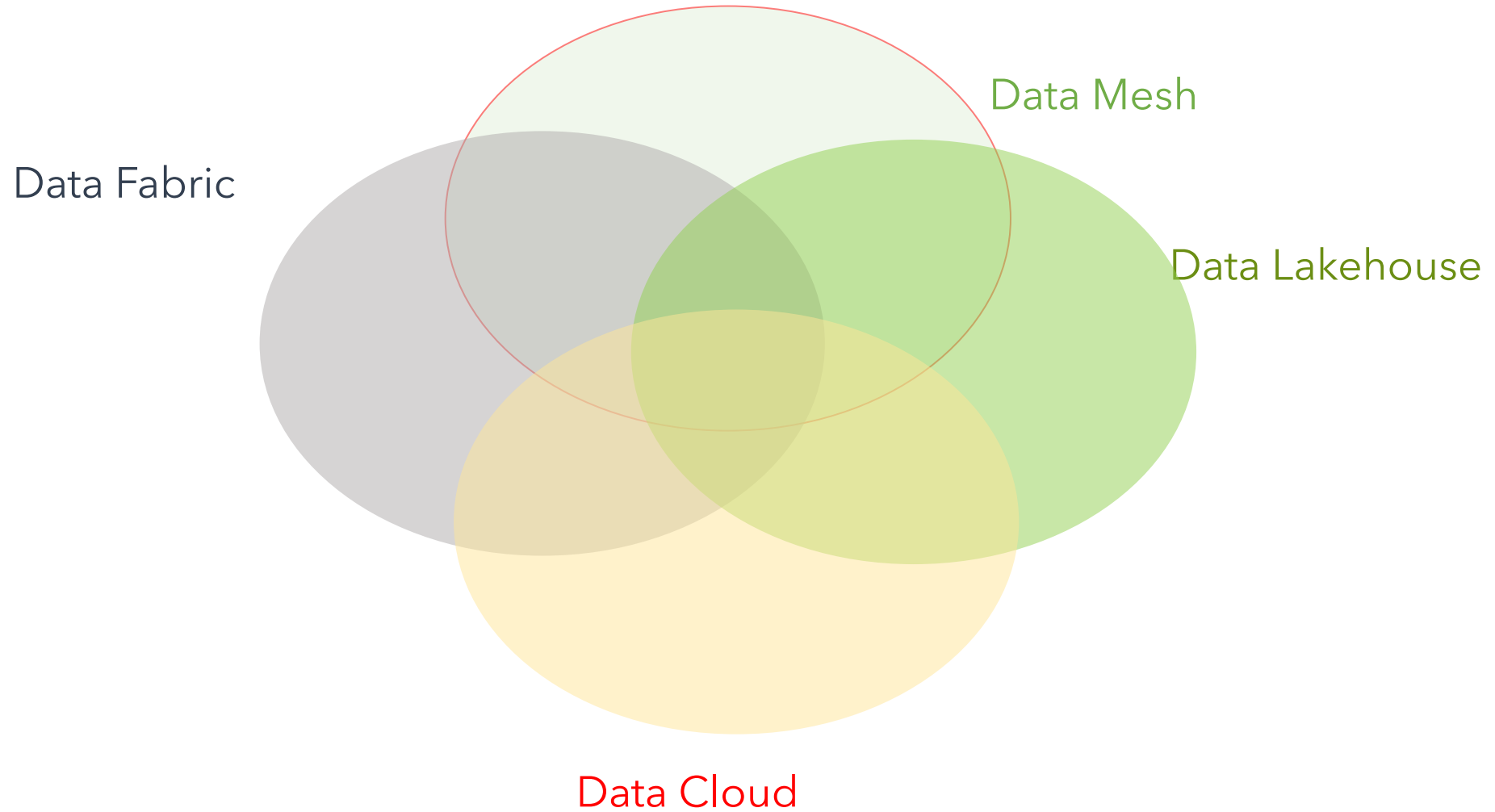
Pros

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

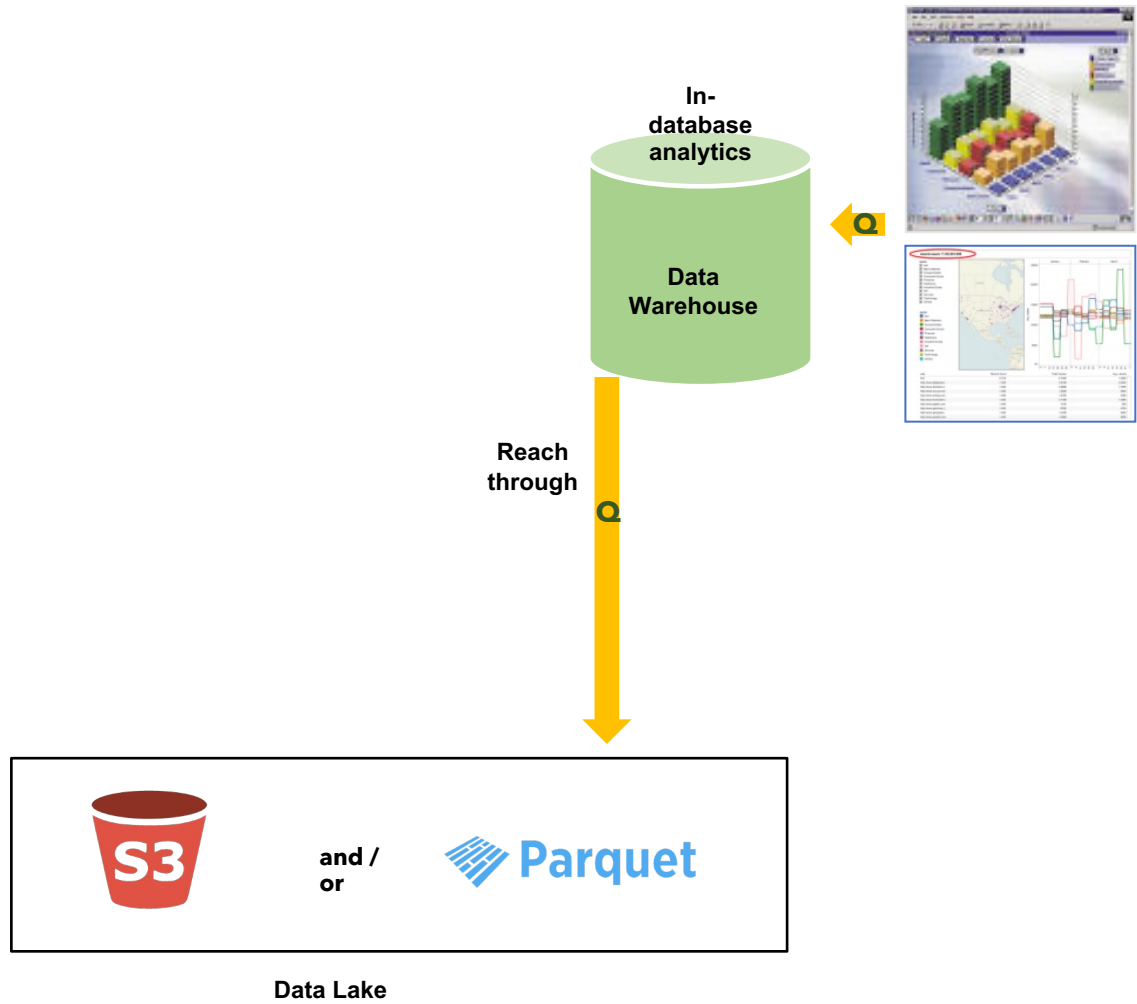
Cons

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

These are not Mutually Exclusive



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

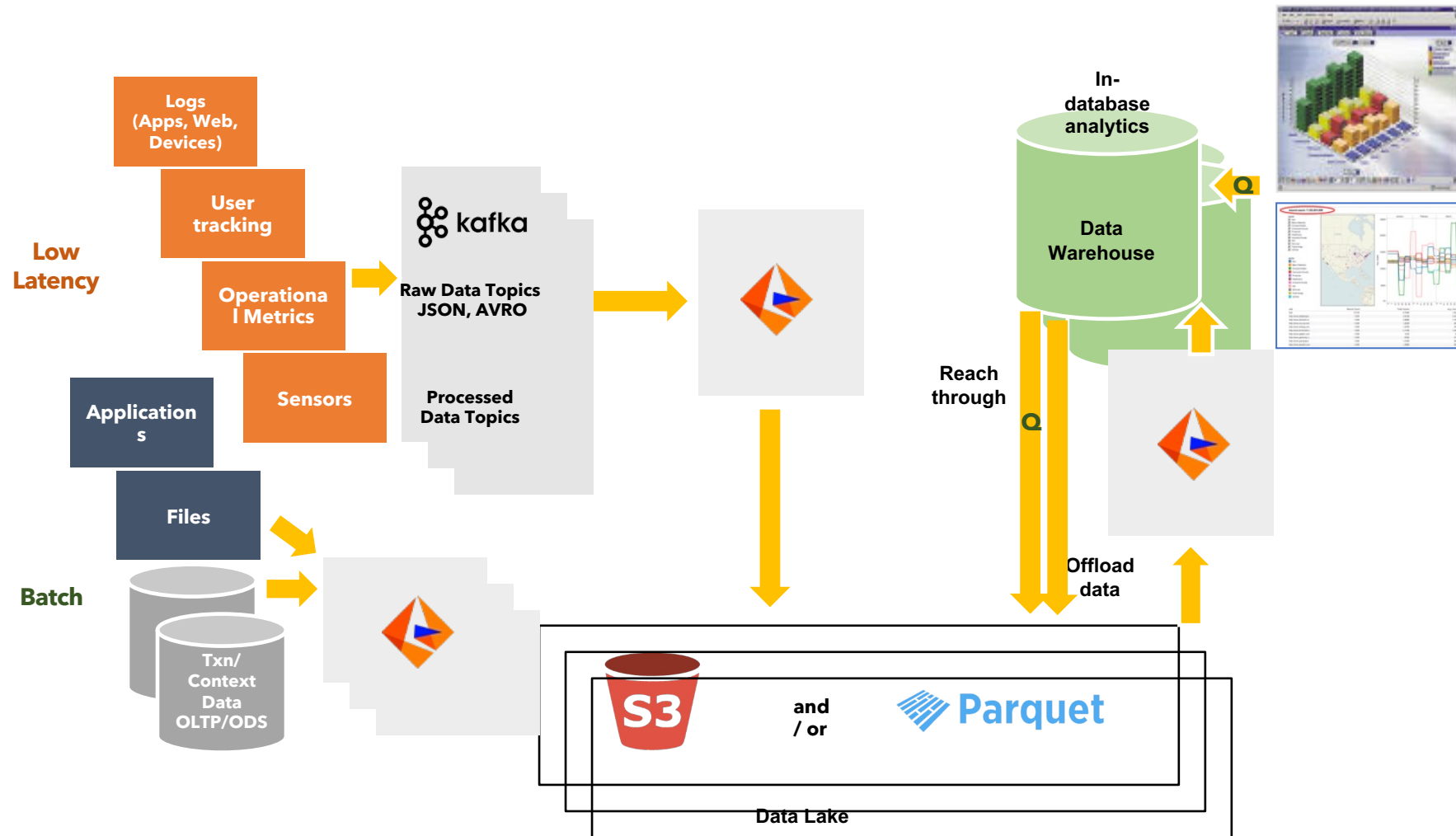
- Administration Management
- Better Organization of Information
- Simplified Rules and Regulations
- More Cost-Efficiency

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



Data Mesh



Data Mesh Principles

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

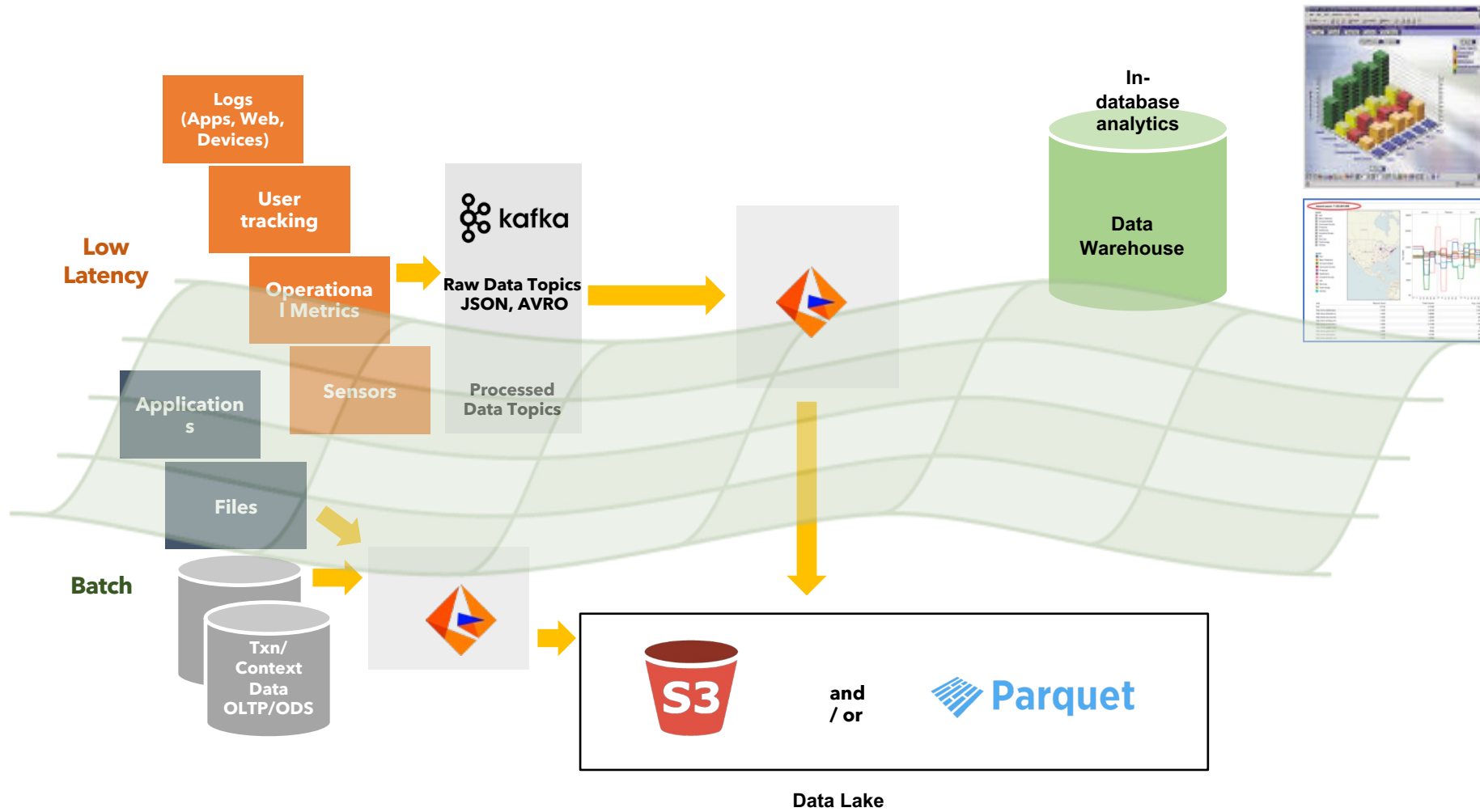


Benefits of a Data Mesh

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



Data Fabric



Data Fabric Principles

Intelligent and Automated

Unification of Disparate Data Systems

Access to Integrated Enterprise Data

Scale Efficiently

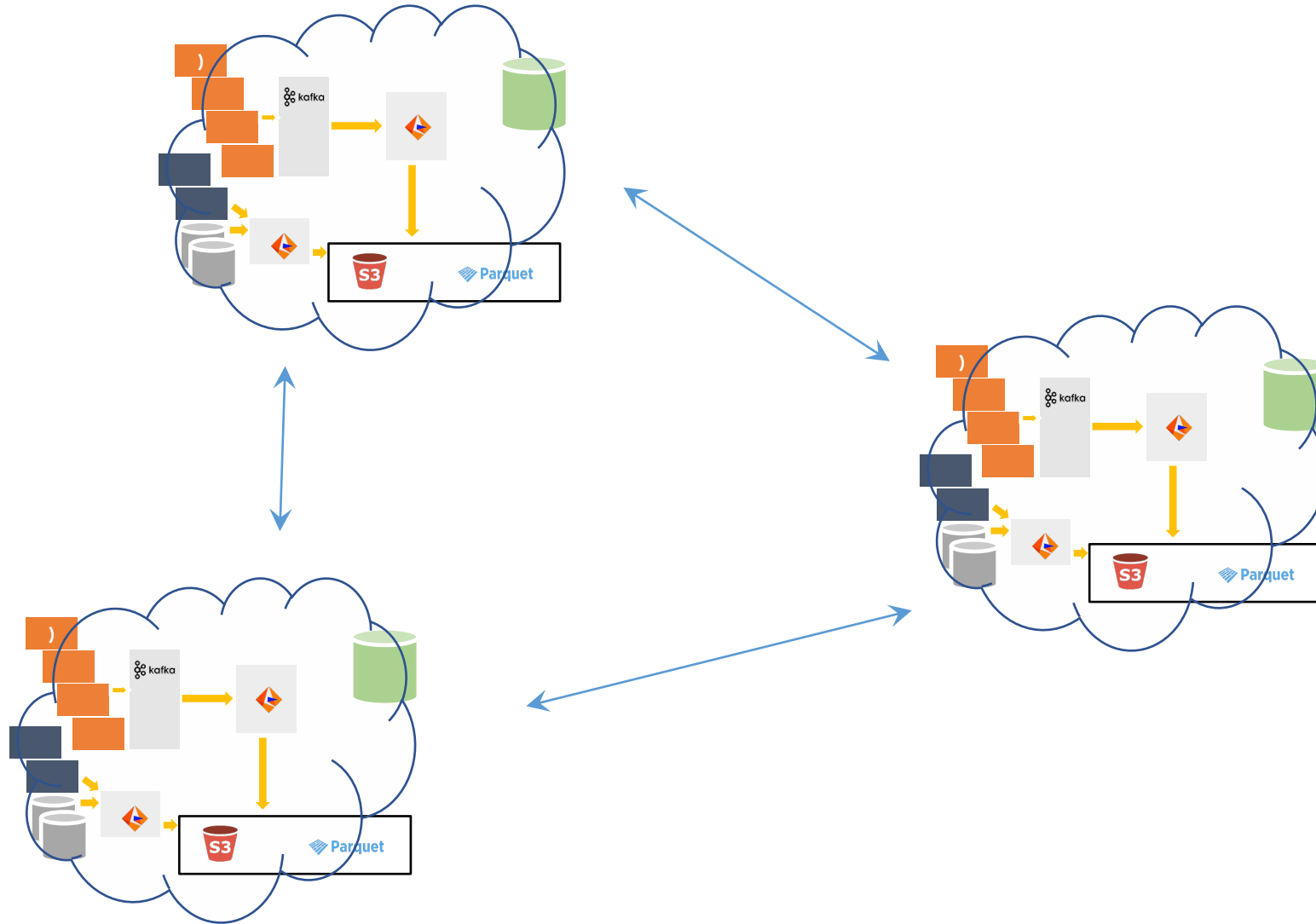
Multi-Cloud Awareness

Benefits of a Data Fabric

- Integrated Intelligence
- Data Democratization
- Improved Data Security
- Universal access to all data sources
- Standardized data format
- Reduced complexity for end users
- Improved data integration and standardization
- Enables access to data from various sources



Data Cloud



Summary

- There are many lens to view data platform decisions
- Components include data warehouse compute, storage, master data management, data integration, streaming, data science/machine learning, data lake, data governance and data observability
- For most companies, all are essential, and most are product decisions
- Data architecture can easily make or break a company
- TCO, ROI, cost predictability and transparency are important
- Data lakehouse, data mesh, data fabric and data cloud are all valuable and not mutually exclusive





Architecting a Modern Data Platform

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

 /in/wmcknight

www.mcknightcg.com
(214) 514-1444

