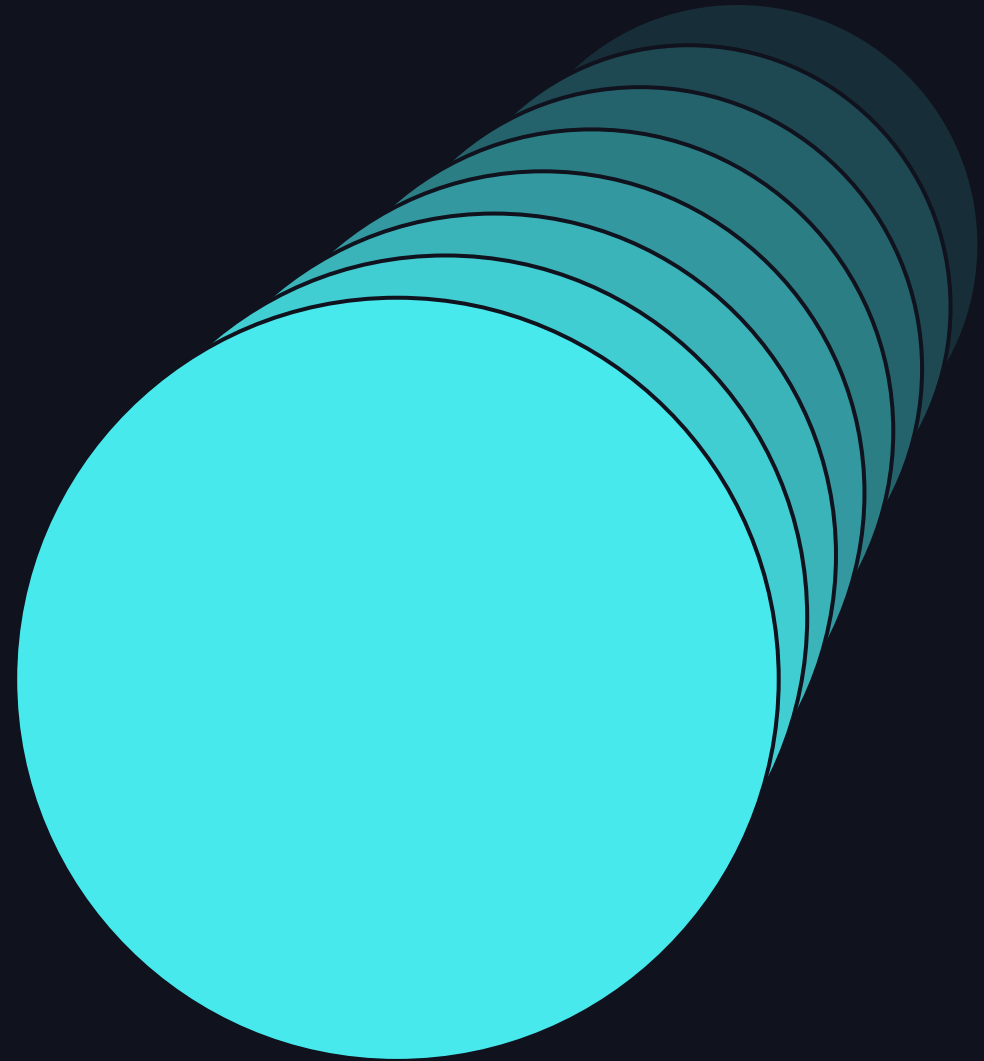


# BUILDING A TRUSTED DATA FOUNDATION for AI on DATABRICKS



---

Sharad Kumar, Field CTO Data - Qlik  
June 11, 2024





Best-in-class data integration, data quality, analytics, AI and machine learning.



50+ offices  
around the world



~3,600  
employees



12 state of the art  
development centers across North  
America, Europe and Asia



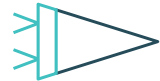
>\$1B invested  
in innovation



>40,000  
customers



1,850  
partners



235,000+  
community  
members

### Extensive Data Foundation Portfolio



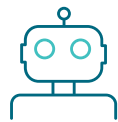
Data  
Integration



Data  
Quality



Analytics



AI & Machine  
Learning

### Gartner® Magic Quadrant Leadership



Data Integration Tools

7 consecutive years



Data Quality Solutions

4 consecutive years



Analytics and Business  
Intelligence Platforms  
13 consecutive years

### Leading companies count on Qlik & Talend



AIRBUS



Lenovo



Domino's



ebay

ABInBev

Deloitte.

ABB

TRAVELERS

NOVARTIS

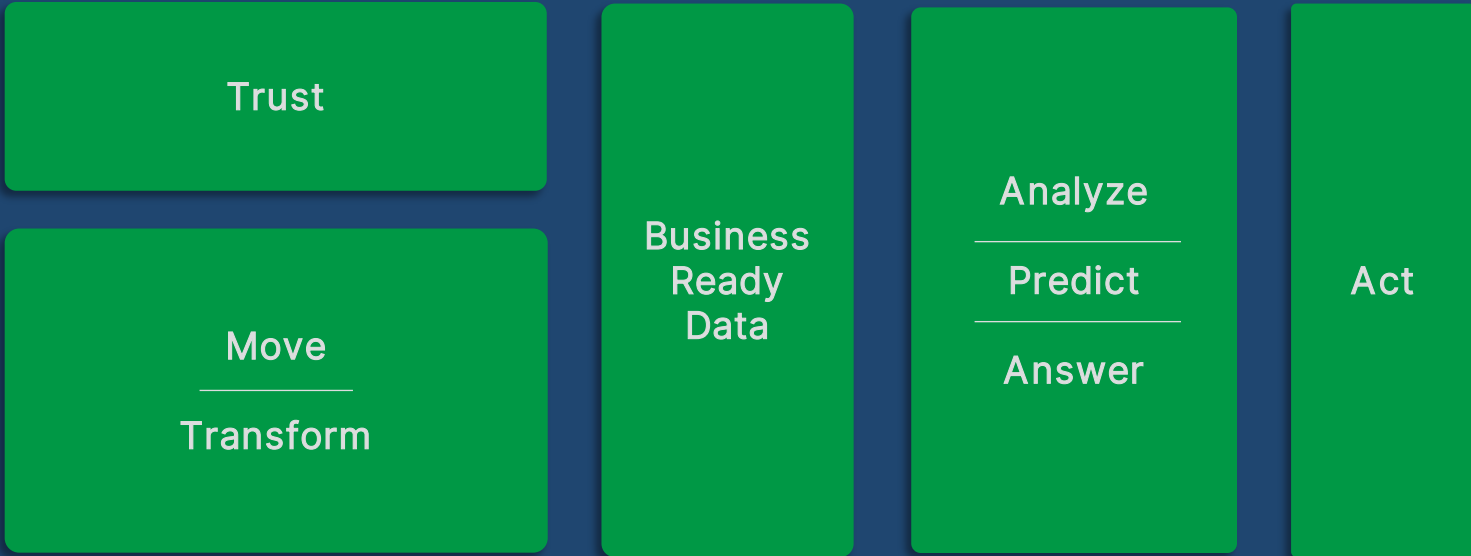
### Supporting 500+ Charities and NGOs impacting the world





Best-in-class data, analytics,  
and AI to drive better outcomes

DATA



OUTCOME



cloudera

databricks

Google

Azure

snowflake



# Qlik Talend

DATA

Trust

Move  
Transform

Business  
Ready  
Data

Analyze  
Predict  
Answer

Act

OUTCOME

Create a trusted data foundation to maximize the transformative value of all your data.

# Qlik Talend Cloud

200+ Connectors

DATA

SaaS, SAP, Mainframe, Databases, Files, Messages

Data Quality  
Observability  
Lineage

Data Classification  
Data Protection

Metadata  
Glossary

Data Products

Trust

Move  
Transform

Business  
Ready  
Data

AI

- Vector DB
- Multi LLM
- Multi Vector Store
- Databricks AI Functions

Ingestion

Bulk, CDC,  
Streaming,  
APIs

Transformation

Multi-pattern: ELT, ETL  
No code/ pro code  
AI Assistant

# Qlik and Databricks

Better together

- 150%+ Growth YoY
- 300+ joint customers
- Joint Solutions:
  - Lakehouse for Manufacturing
  - SAP and Mainframe Solutions
  - Cloud & Hybrid Data Lakes
- Boost your data lake or lakehouse ROI with universal, real-time data ingestion
  - Analytics-ready data delivery
  - Virtually any source
- Eliminate labor-intensive tasks
  - Automatically generate Spark SQL transformations
- Reduce risk with trusted, enterprise-grade data
  - Profile and cleanse data
  - Secure, fully governed, self-service catalog for all data, not just in Databricks

2023 Partner of the Year-Data Integration

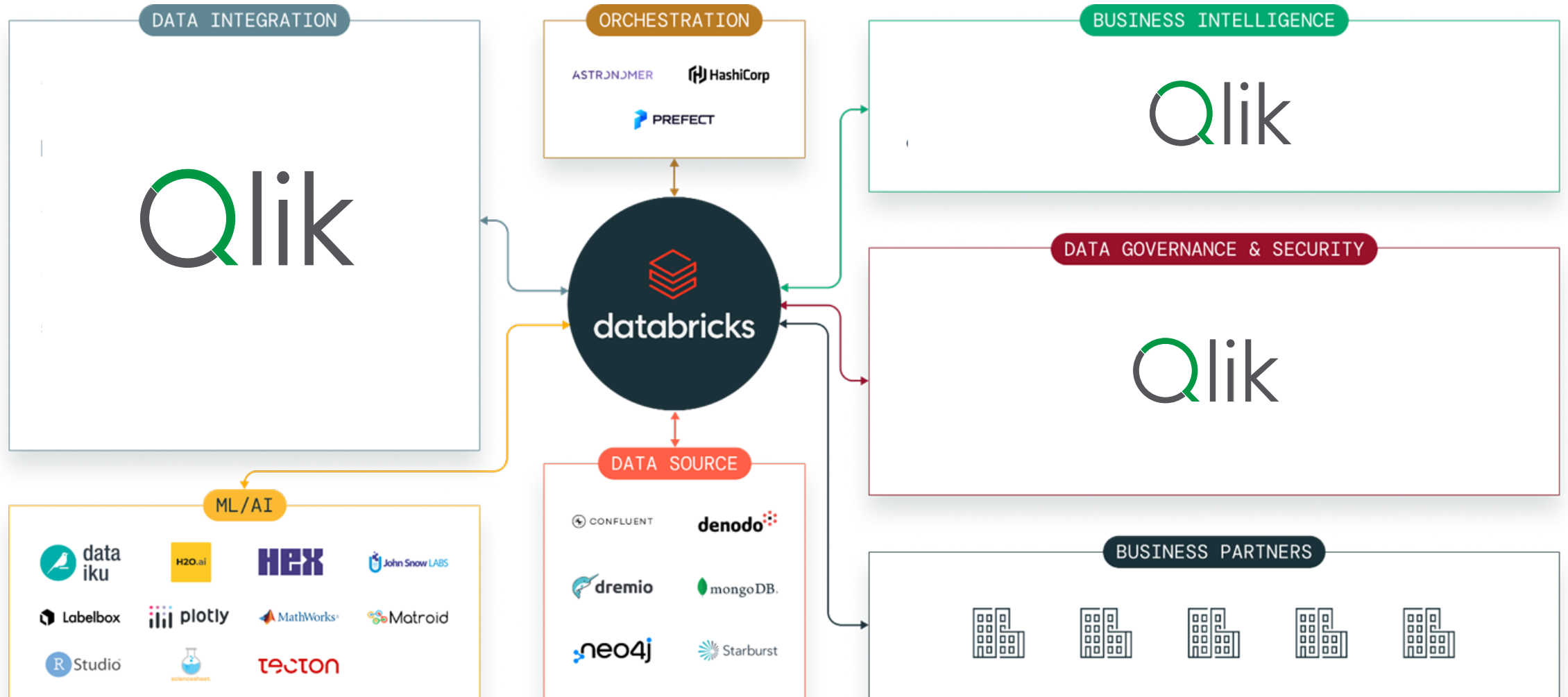


## JOINT CUSTOMERS



# The Databricks data ecosystem

Qlik is the only partner that checks the Integration, Analytics, and Governance boxes



# From Business Intelligence to Generative AI

## Market Evolution

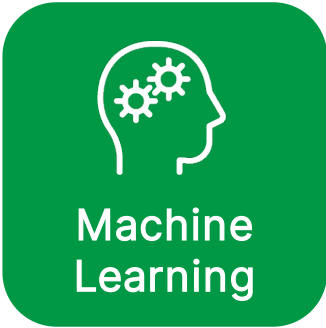
Descriptive



Business Intelligence

Examples  
Reports  
Dashboards  
Visualizations

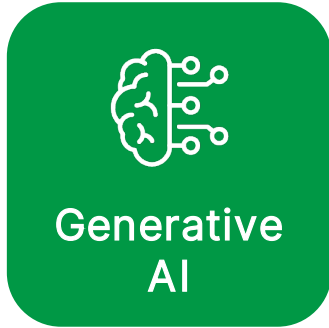
Predictive



Machine Learning

Examples  
Image Recognition  
Next Best Action  
Churn Prediction

Generative



Generative AI

Examples  
Text Generation  
Creative Copilots  
Chatbots

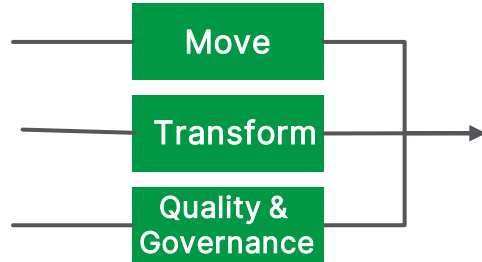


# From Business Intelligence to Generative AI

## Market Evolution



Business Intelligence



Data Warehouse



Data Marts

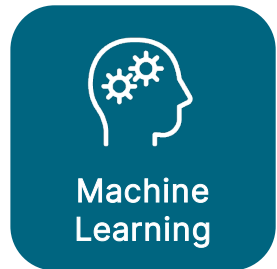


Data Engineer

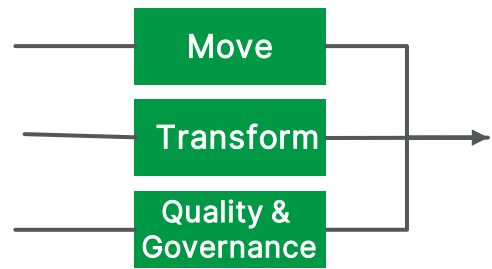
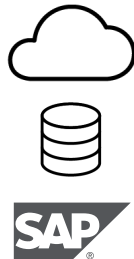


BI Engineer

Examples  
Reports  
Dashboards  
Visualizations



Machine Learning



Data Lakehouse



Feature Tables



Data Engineer

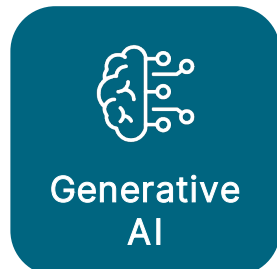


Data Scientist

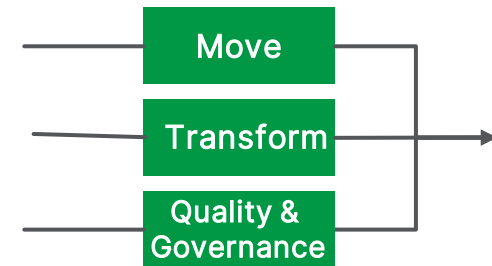


Analytics Engineer

Examples  
Image Recognition  
Next Best Action  
Churn Prediction



Generative AI



Enterprise Corpus



Vector Embeddings



Data Engineer



Data Scientist



Analytics Engineer



App Developer

Examples  
Text Generation  
Creative Copilots  
Chatbots

# 6 Critical Principles for building a Trusted Data Foundation



Diverse

Unbiased across silos



Timely

Up to date and real-time



Accurate

Reliable and trustworthy



Secure

Protected from unauthorized use



Discoverable

Easier to find and understand

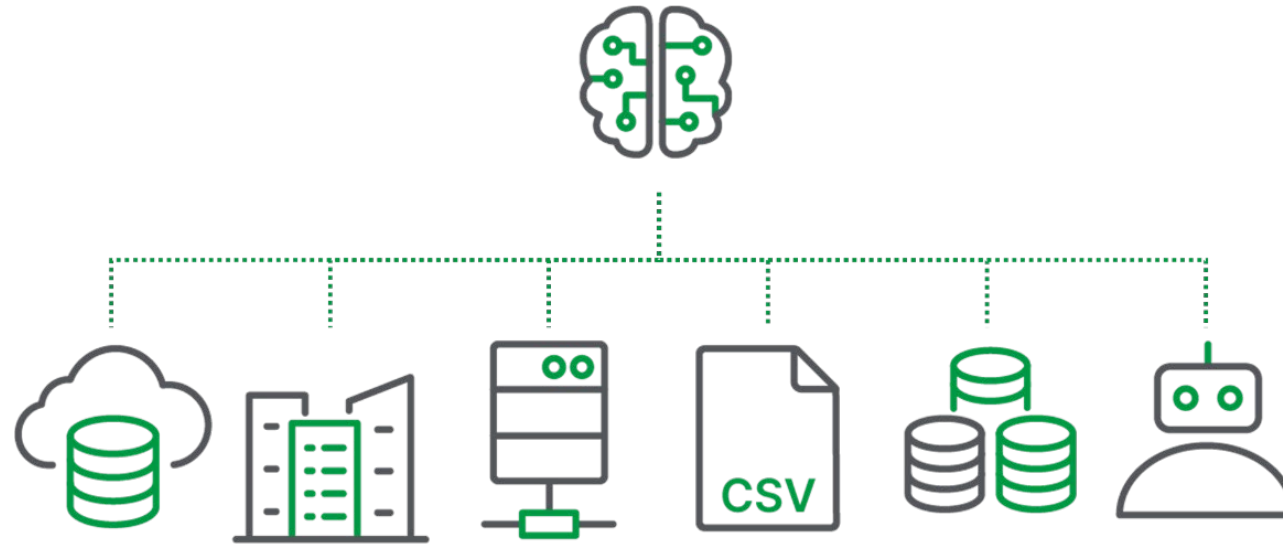
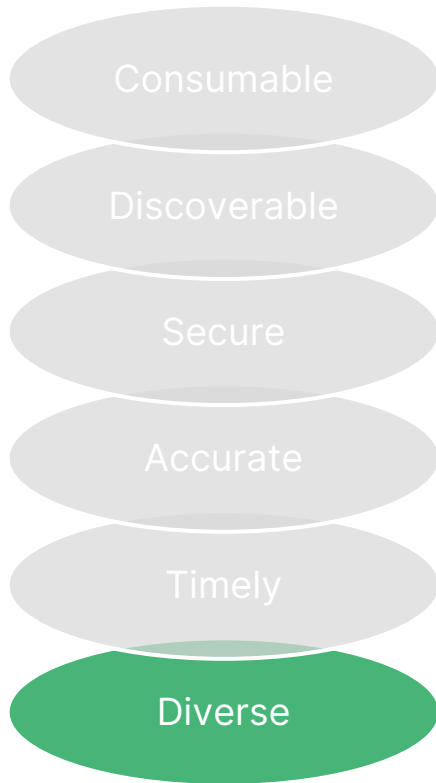


Consumable

In a form that Analytics can consume

# Date Needs to be Diverse

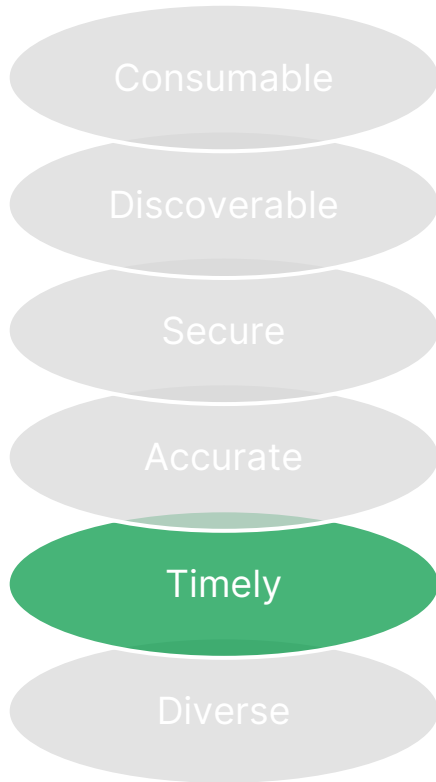
Remove bias in AI systems



Data acquisition capabilities from a wide range of sources and using variety of interface methods – Structured, Unstructured, Semi-Structured

# Data Needs to be Timely

Enable real-time decision making



Change Data Capture

Locating and recording high velocity database changes and instantly sending those updates to a system or process downstream



Stream Data Capture

Capture of data emitted at high volume in a continuous, incremental manner with the goal of low-latency processing

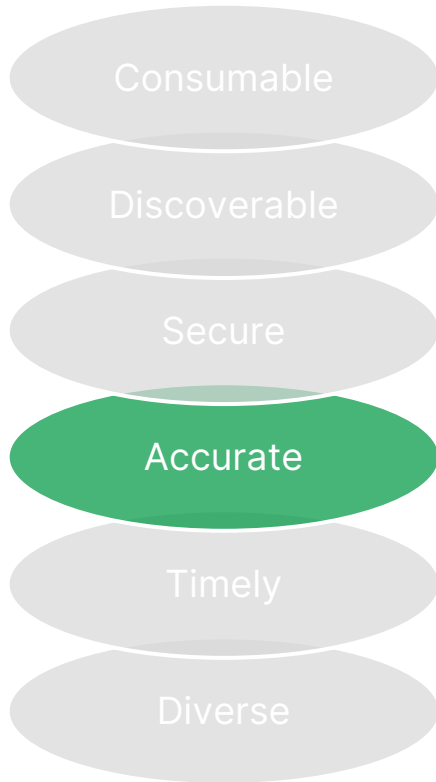


Continuous Data Processing

Instantaneous update of downstream data stores (operational and analytical) for latest data

# Data Needs to be Accurate

Ensure reliability and trustworthiness in AI



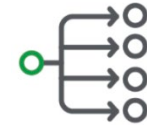
Data Profiling

Gain deeper understanding of data by examining, analyzing and creating useful summaries of data



Data Quality & Observability

Monitor overall health of data and ensure data is accurate, reliable & complete

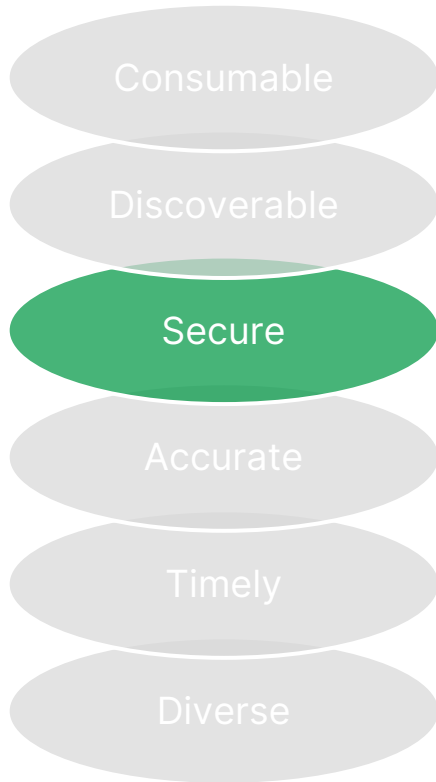


Data Lineage & Impact Analysis

Trace origin & flow of data and assess impact of change

# Data Needs to be Secure

Safeguard reputational value of AI



Data Classification

Automatic detection & classification of data into different categories according to sensitivity



Data Protection

Protect access to sensitive data by applying techniques of masking, tokenization and access control

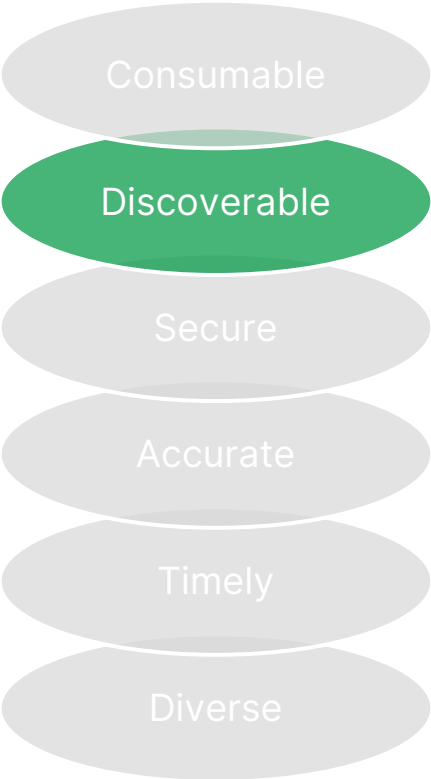


Data Security

Carefully control movement of data outside of private network

# Data Needs to be Discoverable

Ensure more precise predictions



Semantic Typing

Detect and understand data meaning to provide more context



Business Glossary

Describe data in business terms to provide clarity, consistency and productivity

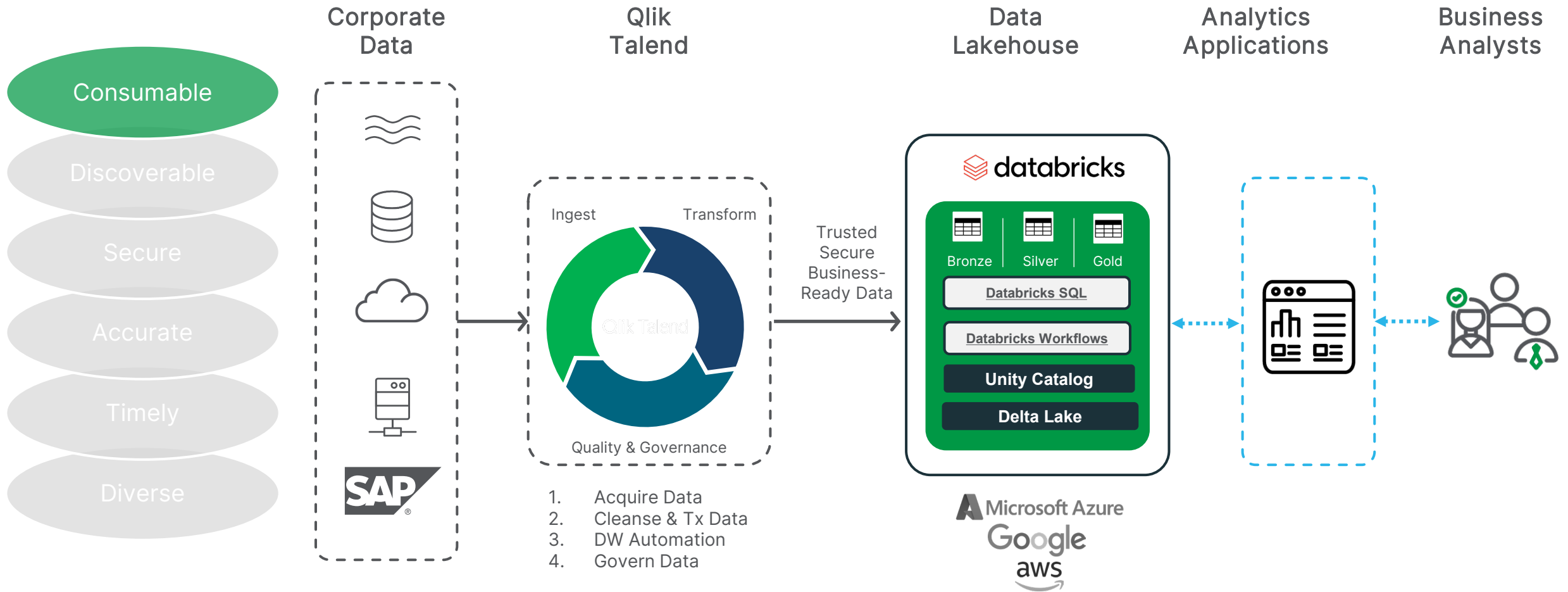


Data Catalog

Index and organize metadata to make data findable and usable

# Data needs to be Consumable

## Qlik Talend Data Pipelines for BI





# Qlik Talend Cloud Data Integration

## Modern Data Pipelines

### Data Ingestion

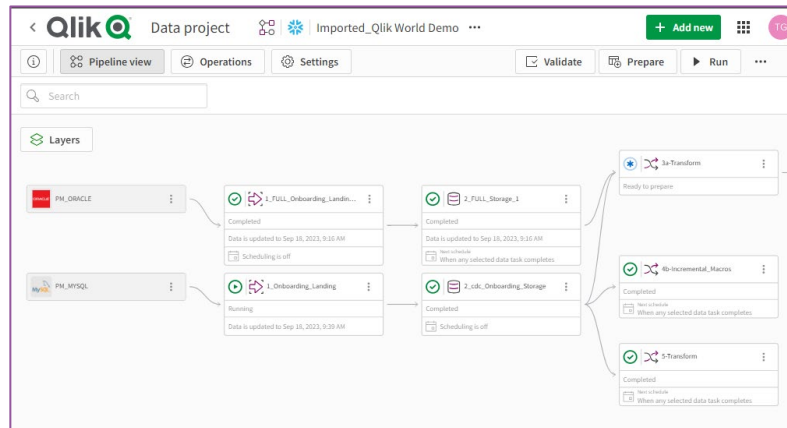
Comprehensive data sources:  
RDBMS, Mainframe, SAP, and SaaS

IBM DB2 for iSeries	Adobe Analytics (Preview)
IBM DB2 for LUW	Cerner Millennium (Preview)
IBM DB2 for z/OS	Coupa (Preview)
Microsoft SQL Server (log based)	Epic on FHIR (Preview)
Microsoft SQL Server (Microsoft CDC based)	Facebook Ads (Preview)
MySQL	Facebook Pages (Preview)
OData	Google Ads (Preview)
Oracle	Google Analytics 4 (Preview)

- Log-based near real-time CDC  
Incremental API integration for SaaS
- Automated Ingestion framework  
manages type 1 and type 2 datasets

### Native Transformation

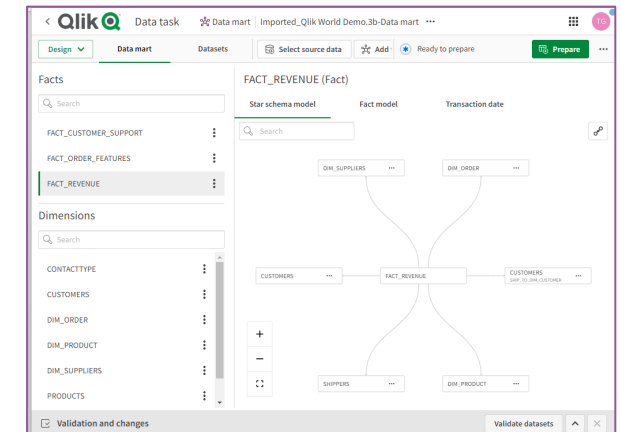
Curate and transform data with  
push-down ELT transformation  
features



- Multi-modal design experience
- AI augmented transform creation
- Fit for purpose transformations

### Automated Patterns

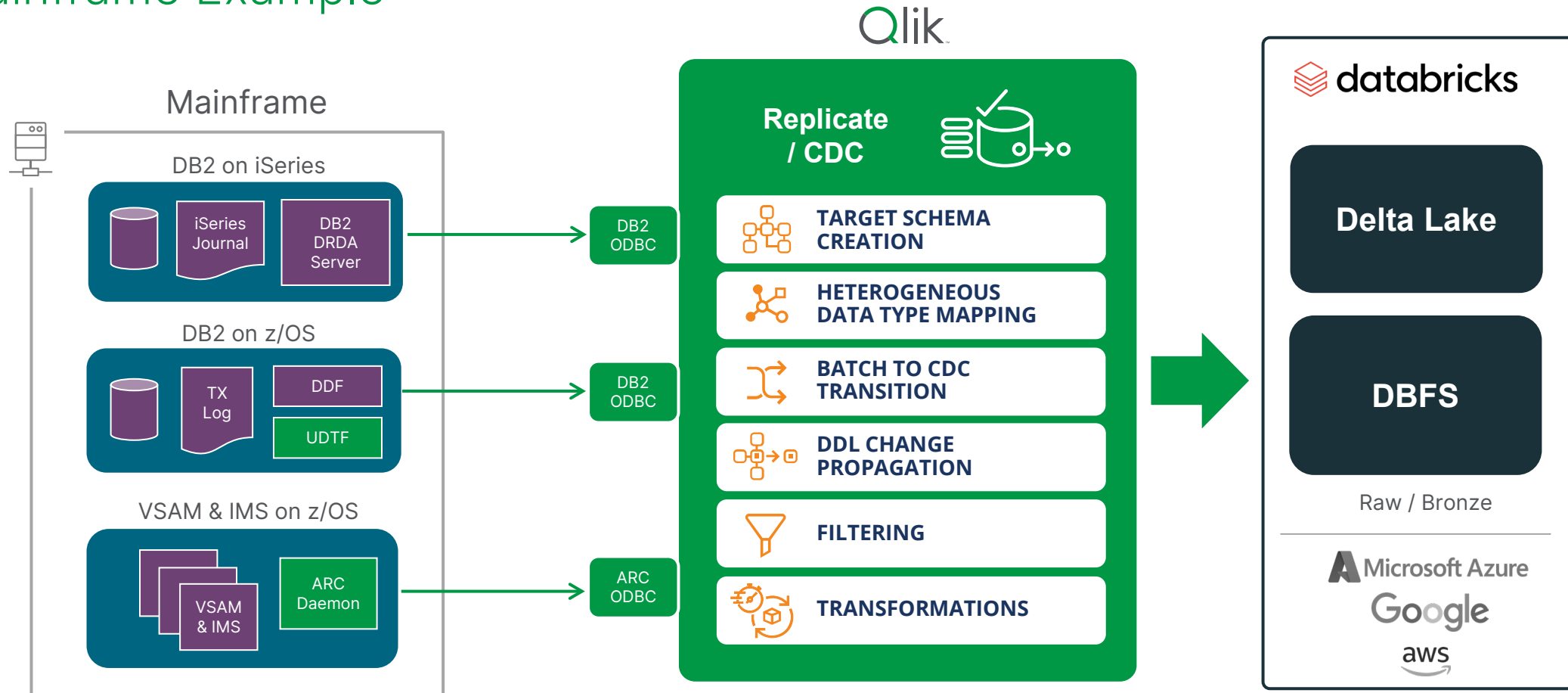
Reduce time to value with  
automated DDL and ELT



- Automated DDL and ELT code generation
- Model-driven Data Mart Automation

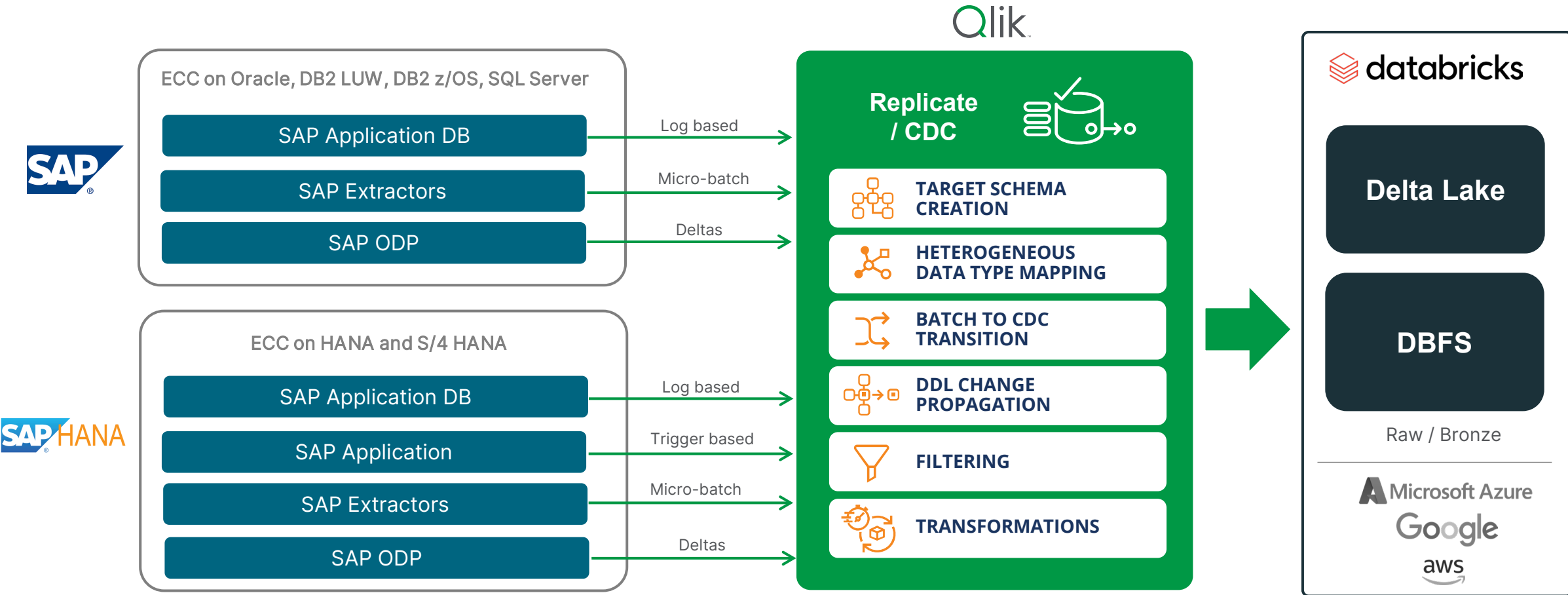
# Qlik Replication – Unlock Data and Hydrate Delta Lake

## Mainframe Example



# Qlik Replication – Unlock Data and Hydrate Delta Lake

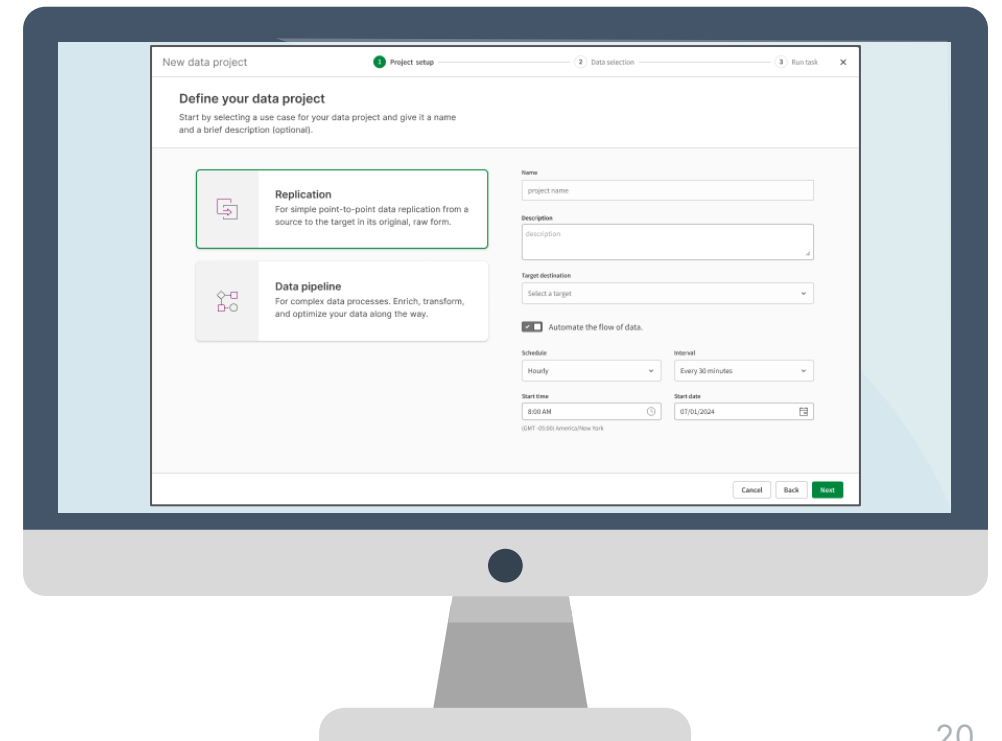
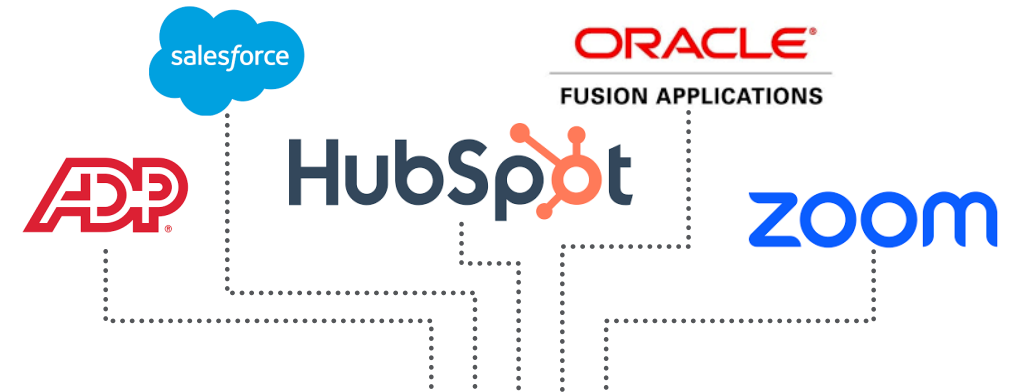
## SAP Example



# Qlik Cloud Data Integration – SaaS Sources

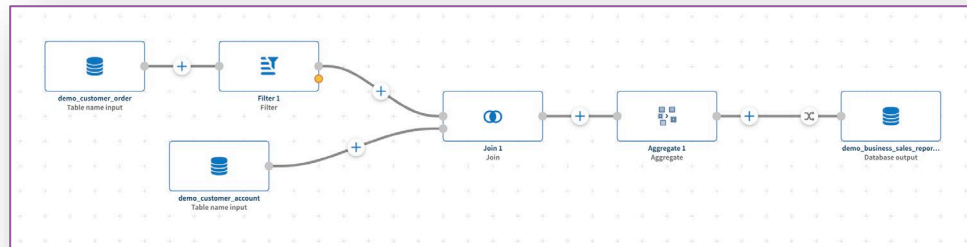
## Ubiquitous Connectivity

- Data Ingest from Cloud Sources
  - **200+ connectors** including: ADP, HubSpot, Oracle Fusion, Anaplan, Zoom, etc.
  - Full and incremental load options
- Cloud-managed Infrastructure
  - 100% Qlik-managed option - no gateways & zero driver management
  - Integrated with all data pipeline components
  - **Optional client-managed data gateway** for sources and targets behind corporate firewall or in Private Cloud configurations
- Simplified Data Ingestion User Experience
  - Simple UI for Data Analyst and Data Scientist personas
  - Advanced UI and tuning options available for Data Architects and Data Engineers



# Powerful Data Transformation

With Push-down Spark and SQL to Databricks Lakehouse



## Natural Language Query

Show all product names and the number of customers having an order on each product.

## Generated SQL Request

```
SELECT product_name,  
       COUNT(DISTINCT customer_id) AS number_of_customers  
FROM Products  
INNER JOIN Order_Items ON Products.product_id = Order_Items.product_id  
INNER JOIN Orders ON Order_Items.order_id = Orders.order_id  
GROUP BY product_name  
ORDER BY number_of_customers DESC
```

PUSHDOWN  
<Spark>

PUSHDOWN  
<SQL>



 databricks

- No-code/low-code design experience with live data preview
- Automated SQL data mart generator, transformation flow designer, or custom SQL
- AI co-pilot to accelerate high-code development approach (prompt-to-SQL)
- AI augmented, in-pipeline Data Quality assessment and remediation
- Model-first automated ELT to complement existing data-first workflow
- Column-level lineage and pipeline observability

# No-Code Transformation Flow: Drag and Drop

Graphical Designer

Generated SQL

The screenshot displays the Qlik No-Code Transformation Flow interface for a process named 'TRF\_DELIVERY\_CALCULATIONS'. The interface is divided into several sections:

- Processors:** A vertical sidebar on the left contains various data processing operators such as Join, Union, Aggregate, Filter, Incremental filter, Select columns, Remove columns, Math, Numbers, and Strings.
- Flow Canvas:** The central workspace shows a graphical flow of data. It starts with a 'Get ORDER Join' processor, followed by an 'LKU Ship Rate Join' processor (highlighted with a green box), and then a 'Select columns 1 Select columns' processor. The flow is connected by lines representing data paths.
- Configuration Panel:** On the right, a panel for the 'LKU Ship Rate Join' processor is visible. It shows the 'Join type' set to 'Left outer join' and 'Conditions' for a 'SHIPVIA SHIPPER' join, with 'Left key\*' as 'SHIPVIA' and 'Right key\*' as 'SHIPPER'. There is a checkbox for 'Keep both left and right keys in the output' which is currently checked.
- Preview and SQL:** At the bottom, there is a 'Preview - LKU Ship Rate' section with a toggle for 'SQL' (checked) and 'Data preview'. Below this is a text area containing the generated SQL statement.

```
1 SELECT *
2 FROM (SELECT *
3 FROM (SELECT "SHIPMENT_ID", "HANDLER_ID", "REQUIREDDATE", "SHIPPEDDATE", "SHIPVIA", "FREIGHT", "SHIPPING_DELAY_DAYS", "LATE_DAYS"
4 FROM "A_QLIK_CONNECT"."public_employee_data"."DELIVERY_CALCS") AS "t"
5 INNER JOIN (SELECT "ODID", "ORDERID", "PRODUCTID", "UNITPRICE", "QUANTITY", "DISCOUNT", "CATEGORY"
6 FROM "A_QLIK_CONNECT"."public_employee_data"."ORDER_DETAILS") AS "t0" ON "t"."SHIPMENT_ID" = "t0"."ORDERID") AS "t1"
7 LEFT JOIN (SELECT MOD("ID_SHIPPING_ZONES", 3) AS "SHIPPER", AVG("PRICE") AS "AVG_RATE", MIN("ID_SHIPPING_ZONES") AS "ID_SHIPPING_ZONES"
8 FROM "A_QLIK_CONNECT"."shopify_storage"."SHIPPING_RATES"
9 GROUP BY MOD("ID_SHIPPING_ZONES", 3)) AS "t4" ON "t1"."SHIPVIA" = "t4"."SHIPPER"
```

# AI Co-Pilot: NL Prompt to SQL Transformation

Prompt   SQL

generate a dataset of RAW materials only including the material number, base unit, different stock and inventory levels (including the reorder level) and calculate the outstanding (undelivered) quantity of goods on order and goods delivered

**NL Transformation Prompt**

Leverages LLM Integration

This feature uses generative AI. ⓘ

Generate

Cancel   **OK**

Edit SQL transformation: Inventory\_Management\_Data

Add a SQL transformation to create a table or view based on a custom SQL SELECT query. The query can include, for example, selecting columns, filtering records, and joining tables.

```
1 SELECT
2
3 ${MARA}.MATNR AS "Material Number",
4 ${MARA}.MEINS AS "Base Unit",
5 ${MARD}.LABST AS "Stock Level",
6 ${MARD}.INSME AS "Inventory Level",
7 ${MARD}.LMINB AS "Reorder Level",
8 SUM(${EKPO}.MENGE) - SUM(${EKPO}.ABMNG) AS "Outstanding Quantity"
9 FROM ${MARA}
10 JOIN ${MARD} ON ${MARA}.MANDT = ${MARD}.MANDT AND ${MARA}.MATNR = ${MARD}.MATNR
11 LEFT JOIN ${EKPO} ON ${MARA}.MANDT = ${EKPO}.MANDT AND ${MARA}.MATNR = ${EKPO}.MATNR
12 WHERE ${MARA}.MTART = 'ROH'
13 GROUP BY ${MARA}.MATNR, ${MARA}.MEINS, ${MARD}.LABST, ${MARD}.INSME, ${MARD}.LMINB;
14
```

**Generated Transformation SQL Code**

Prompt   SQL

```
SELECT
${MARA}.MATNR AS "Material Number",
${MARA}.MEINS AS "Base Unit",
${MARD}.LABST AS "Stock Level",
${MARD}.INSME AS "Inventory Level",
${MARD}.LMINB AS "Reorder Level",
SUM(${EKPO}.MENGE) - SUM(${EKPO}.ABMNG) AS "Outstanding Quantity"
FROM ${MARA}
JOIN ${MARD} ON ${MARA}.MANDT = ${MARD}.MANDT AND ${MARA}.MATNR = ${MARD}.MATNR
LEFT JOIN ${EKPO} ON ${MARA}.MANDT = ${EKPO}.MANDT AND ${MARA}.MATNR = ${EKPO}.MATNR
WHERE ${MARA}.MTART = 'ROH'
GROUP BY ${MARA}.MATNR, ${MARA}.MEINS, ${MARD}.LABST, ${MARD}.INSME, ${MARD}.LMINB;
```

This feature uses generative AI. ⓘ  
Help us improve by rating this result: ⤴ ⤵

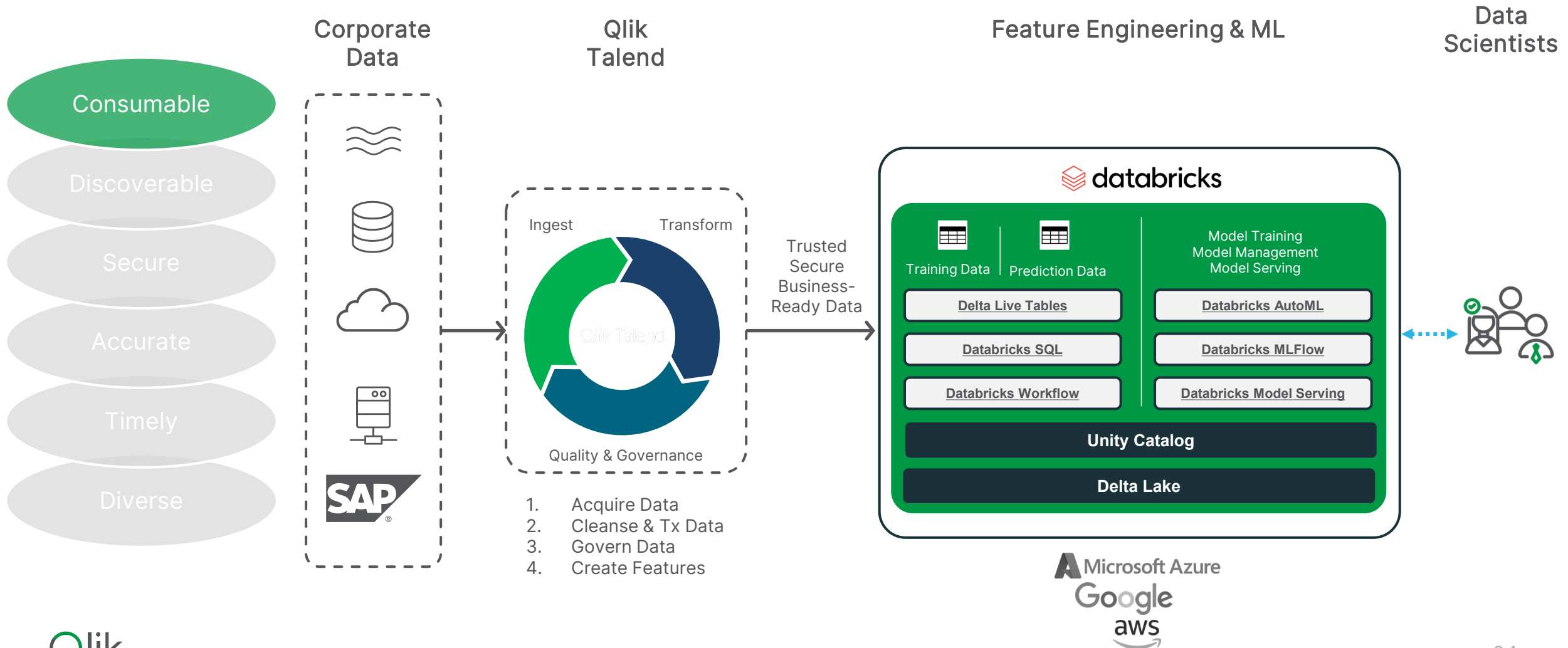
Extract parameters   Collapse   SQL Assistant   Apply   Edit prompt

Cancel   **OK**

⚠ Your work will be saved but not applied. You need to click "Extract parameters" and "Describe table" to complete the transformation.

# Data needs to be Consumable

## Qlik Talend Data Pipelines for ML





# The Rise of GenAI

Generative AI or GenAI is artificial intelligence that leverages natural language processing and machine learning to create new content, such as text, videos, audio, and images. The underlying models of GenAI have been trained on extensive web data, enabling them to learn patterns and structures from this data and produce new content with similar attributes.

Generative AI will add between **\$2.6<sup>1</sup>** and **\$4.4<sup>1</sup>** trillion in annual value to the global economy

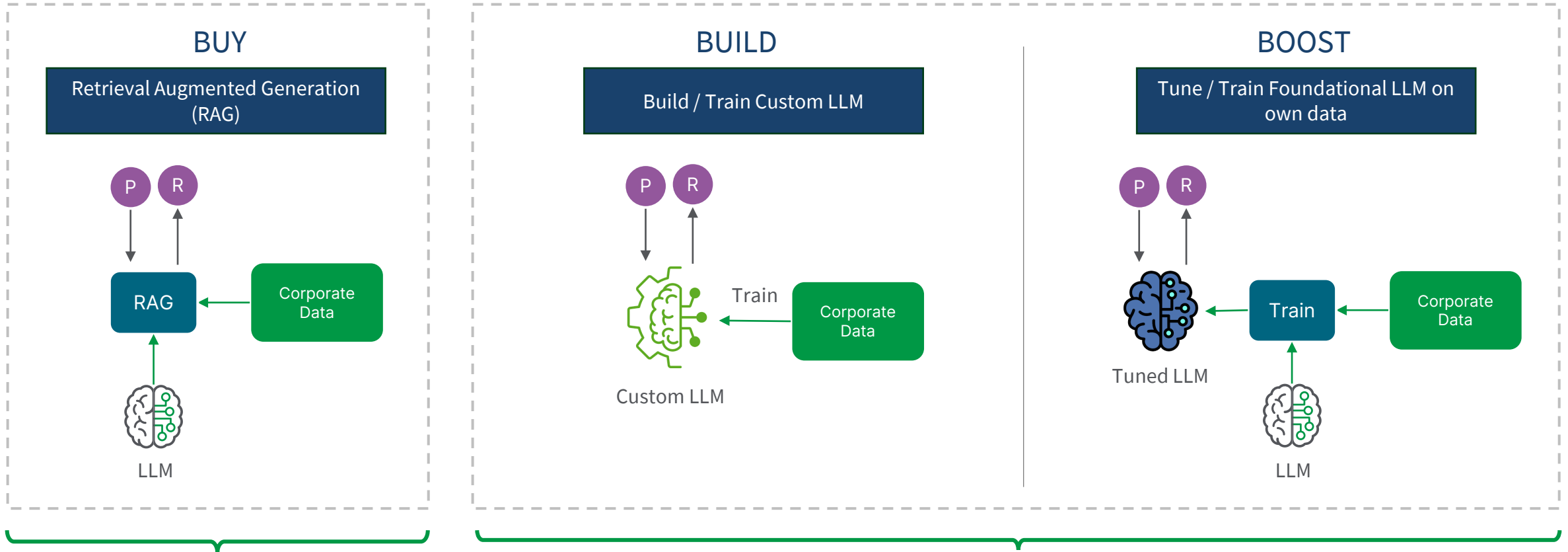
By 2025, generative AI will be a workforce partner for **90%<sup>2</sup>** of companies globally

By 2026, over **80%<sup>3</sup>** of enterprises will have used Gen AI APIs or models to deploy GenAI-enabled apps in production

Spending on GenAI Solutions Will Reach **\$143B<sup>4</sup>** in 2027 with a Five-Year Compound Annual Growth Rate of **73.3%<sup>4</sup>**

# Harness the power of GenAI

## Grounding LLMs with corporate data



+ Domain contextual data at inference time

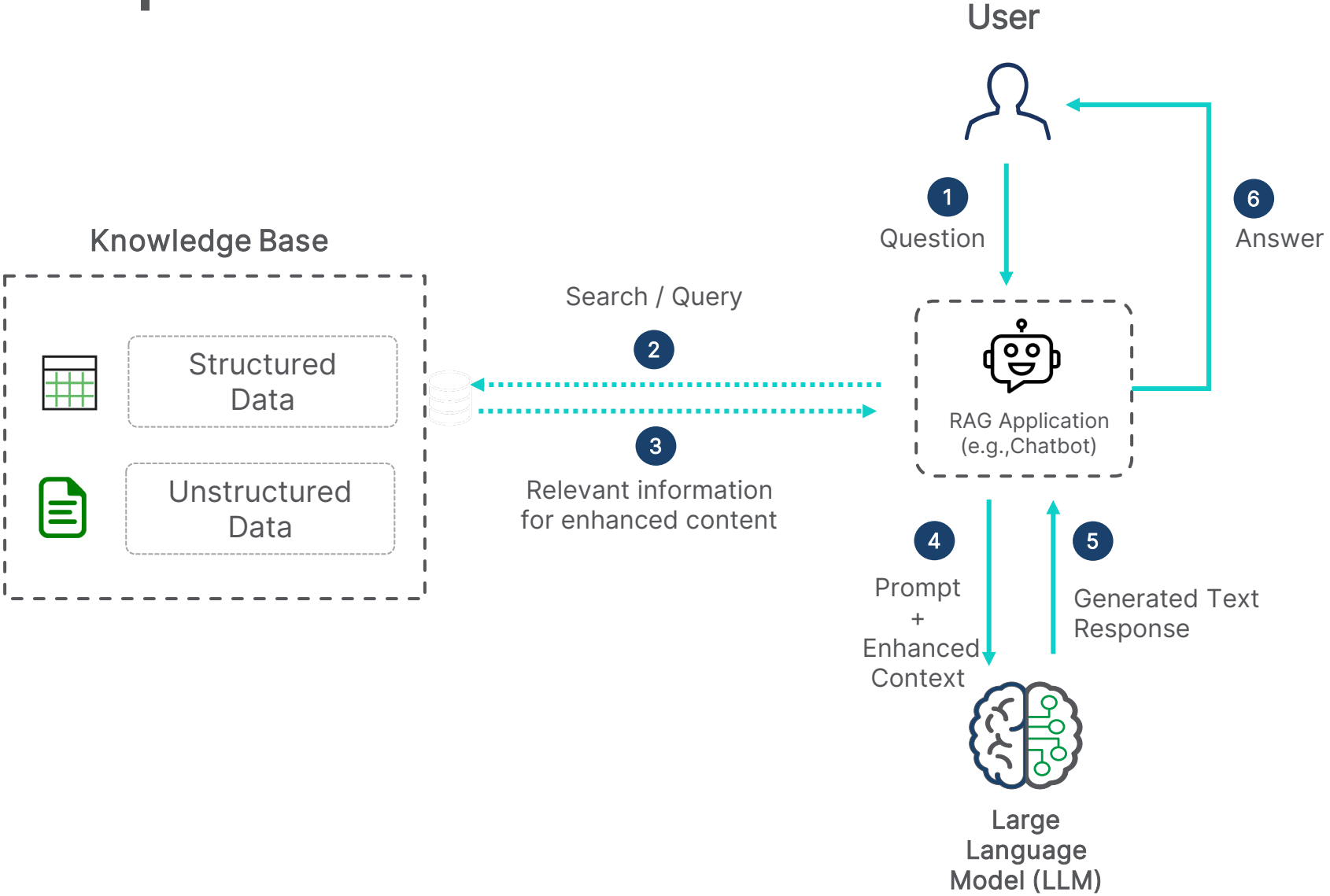
+

- Domain / task specific
- Better performance

-

- Costly
- Requires specialized expertise
- Sizable infrastructure needed

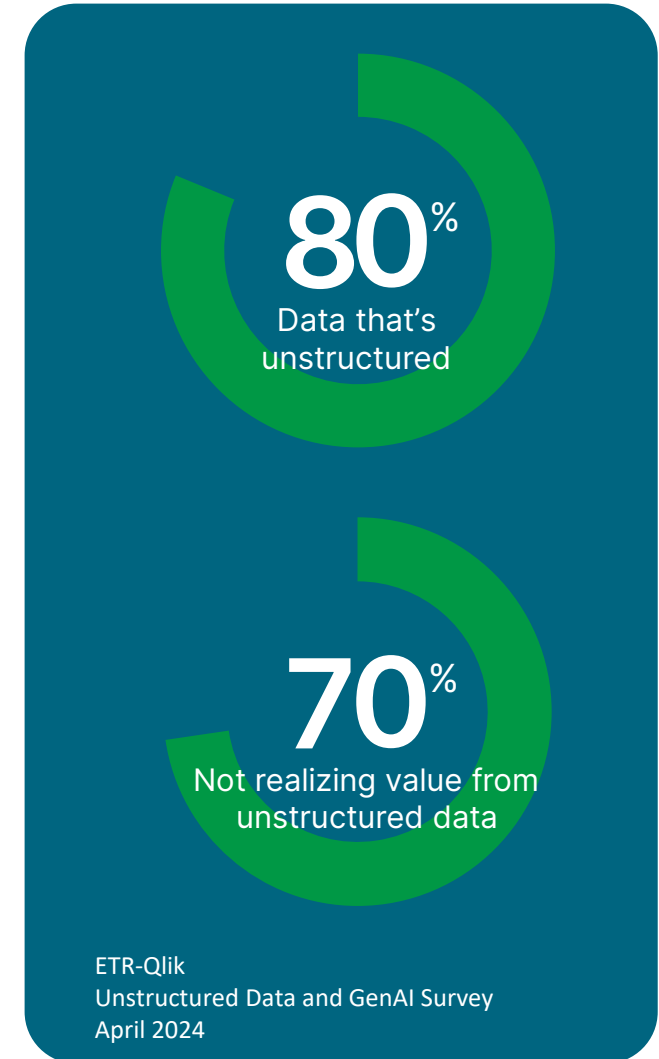
# RAG Explained



# Huge untapped potential in unstructured data

- Information contained in *unstructured data* is often under-utilized or overlooked and un-governed
- Decisions and actions *are being taken* without the best knowledge, and there are consequences that add up
- The good news is that these resources *usually exist*, in knowledge libraries, document repositories, and online content

But how do you get answers into the hands of those who need them?



# Traditional search doesn't work

There's a big difference between a search result and an answer

## Traditional (Built-In) Search

- X List of possible results**  
Returns a list of possible sources for an answer, but users have to manually investigate
- X Single source of content**  
Built-in solutions can only access content contained in their system
- X Static ranking and keyword search**  
Keyword searches ranked according to pre-defined algorithms, limiting relevance

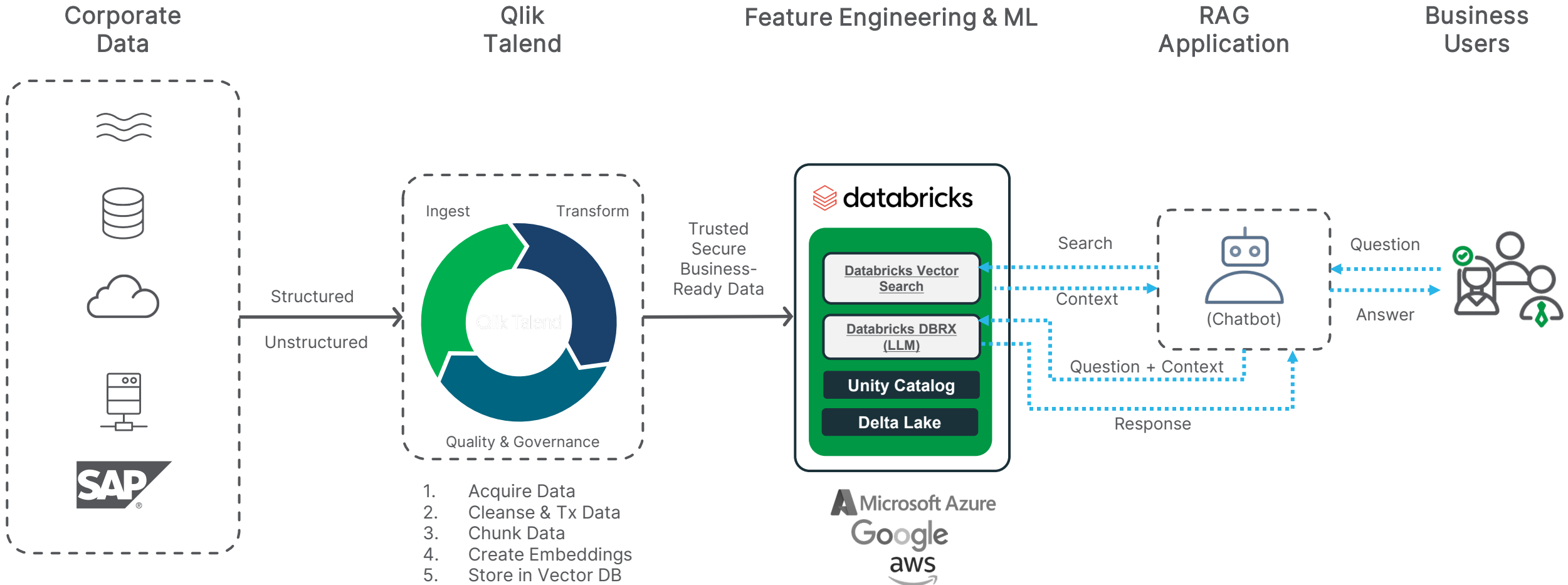
A better way

## Generative AI (and RAG)

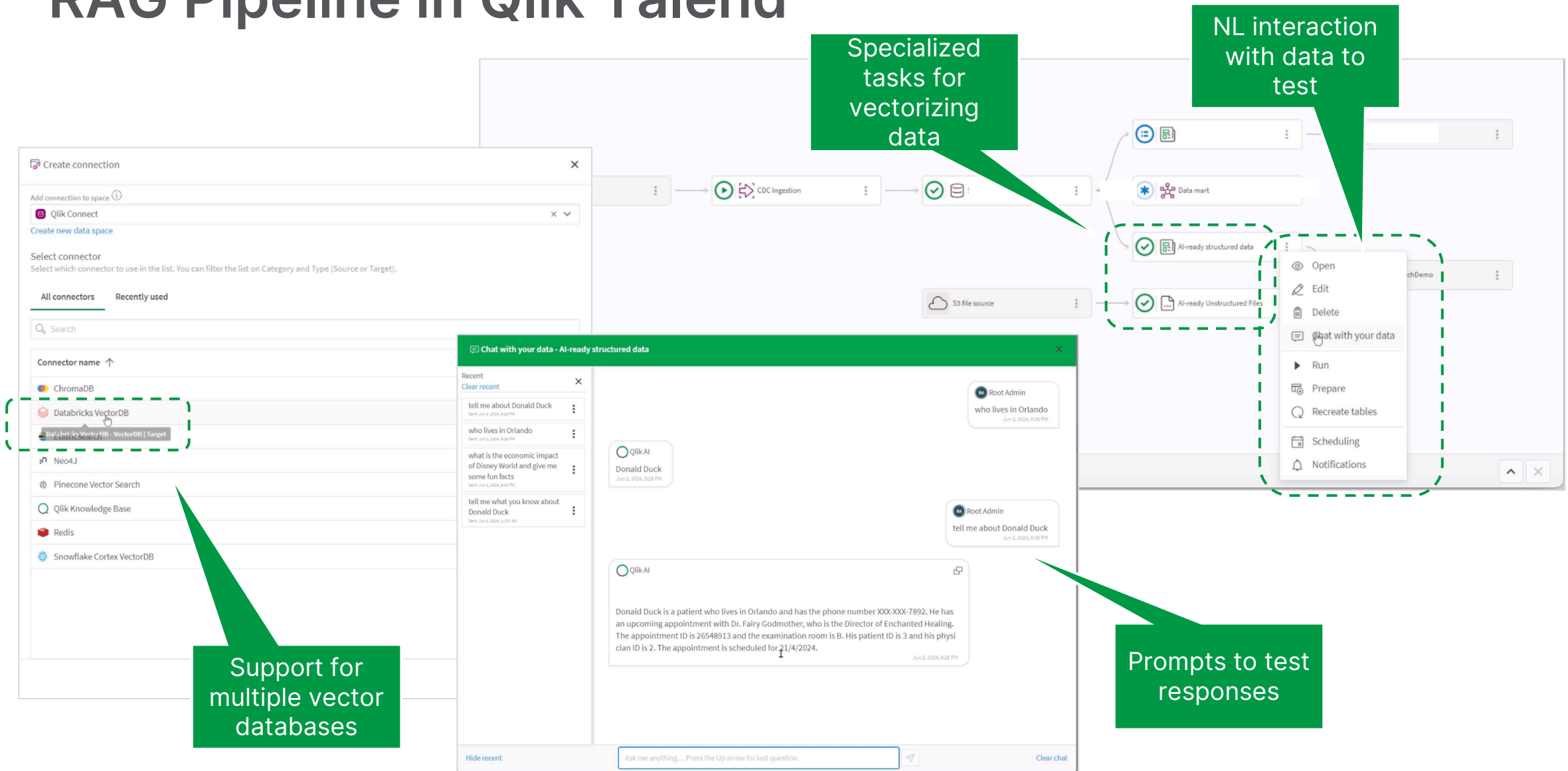
- ✓ Personalized, relevant answers**  
Provides a personalized answer to your question, based on sources
- ✓ Multiple domain specific sources**  
Can access content from a variety of different sources, carefully curated
- ✓ Advanced Semantic Search**  
Relevant, repeatable, explainable information specific to user questions

# Data needs to be Consumable

## Qlik Talend Data Pipelines for RAG Applications

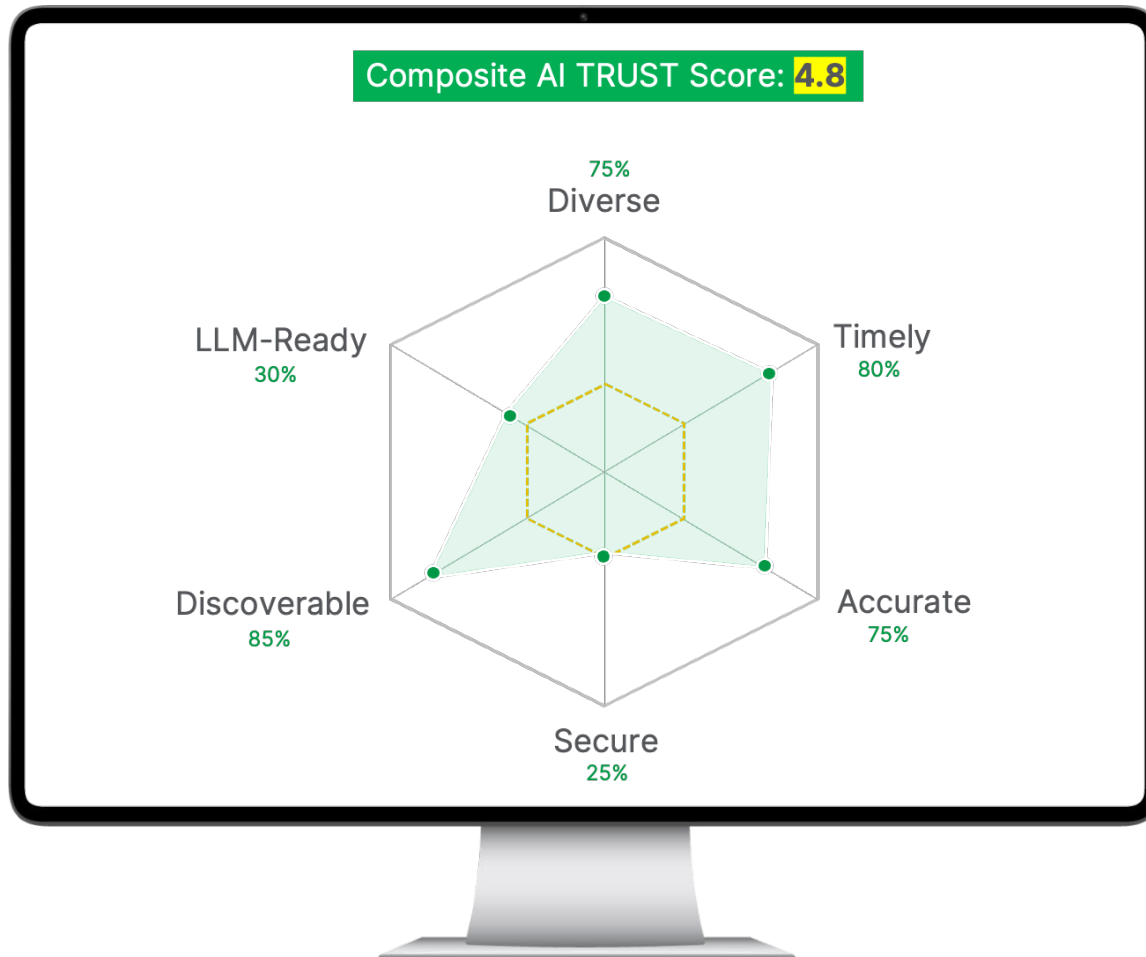


# RAG Pipeline in Qlik Talend



# Qlik Talend Trust Score for AI

## Providing Confidence for Data Consumers



- **AI Trust Score** is a global readiness indicator that aggregates multiple metrics into a single, understandable score.
- **Signals 'AI-readiness'** Calculated from six principles of AI-ready data
  - Diversity
  - Timeliness
  - Accuracy
  - Security
  - Discoverability
  - ML/LLM Consume-ability



## Customer Story



“We’re seeing more real-time data in J.B. Hunt 360, which gives shippers and carriers up-to-the-minute information on how they are performing.”

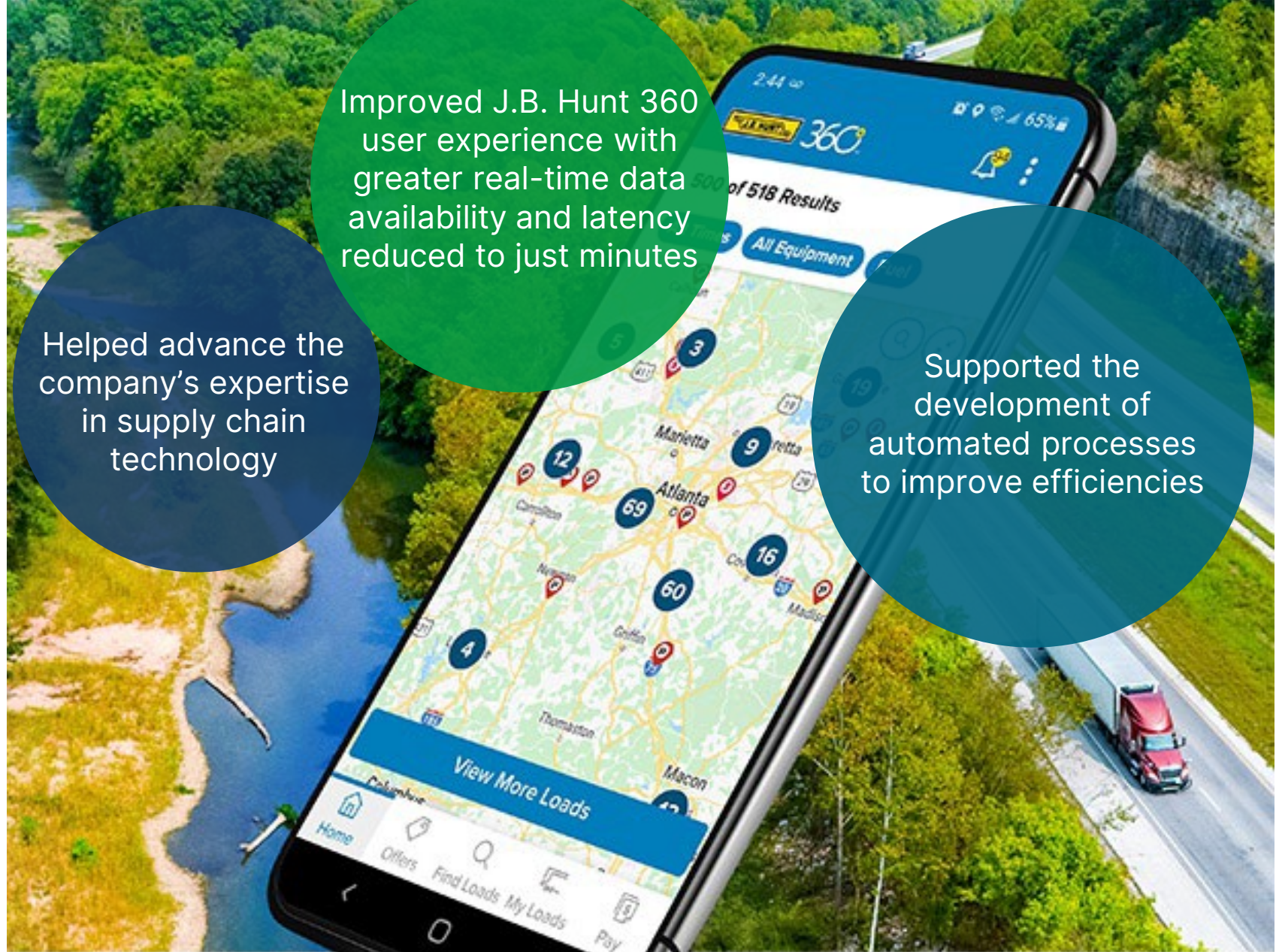


CDC Streaming  
Data Catalog  
Data Warehouse

Helped advance the company's expertise in supply chain technology

Improved J.B. Hunt 360 user experience with greater real-time data availability and latency reduced to just minutes

Supported the development of automated processes to improve efficiencies





## Customer Story



“Databricks and Qlik, provide Mercedes with flexibility and cloud computing power to run artificial intelligence (AI) and analytics at global scale”

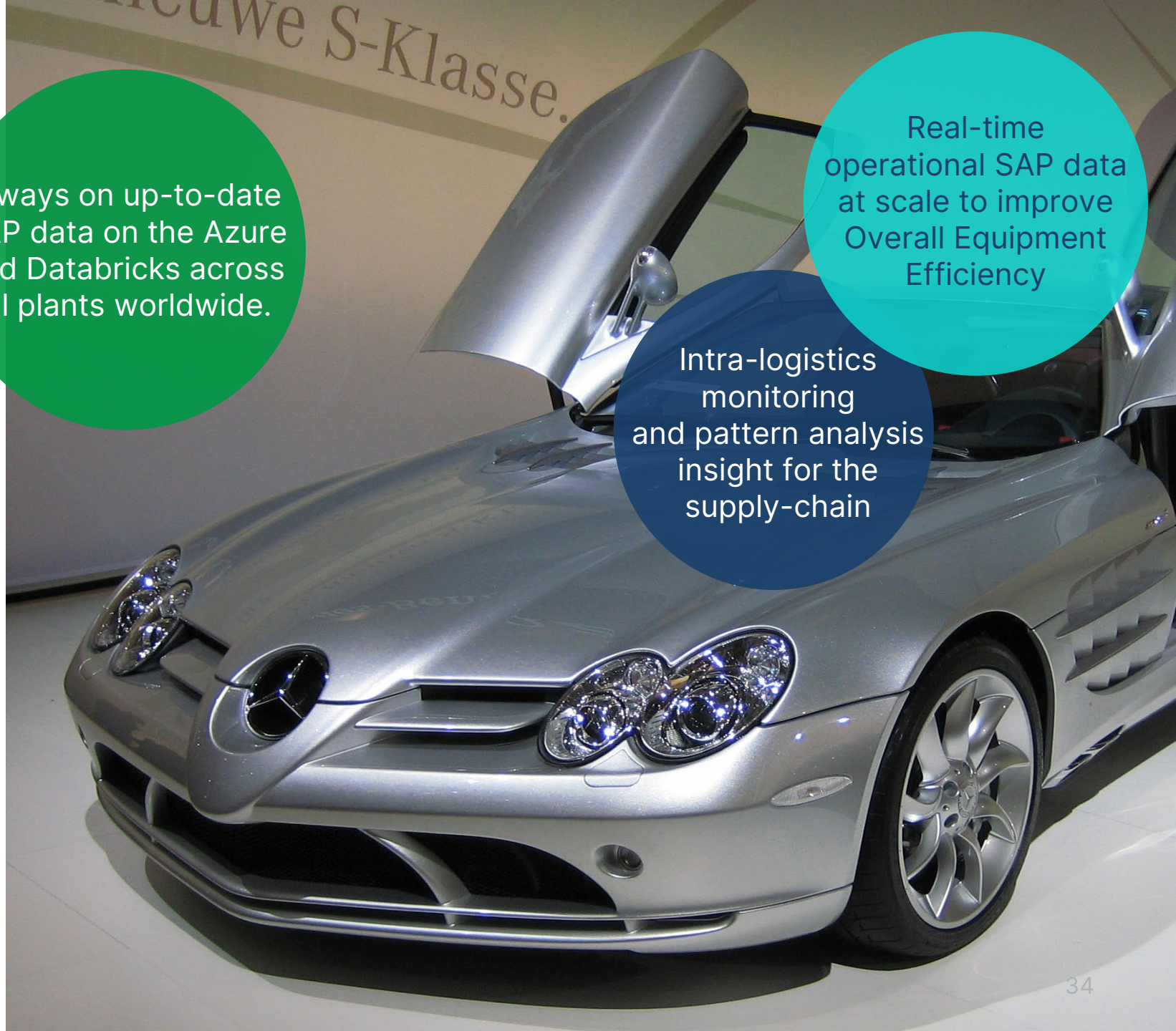
Always on up-to-date SAP data on the Azure and Databricks across all plants worldwide.

Real-time operational SAP data at scale to improve Overall Equipment Efficiency

Intra-logistics monitoring and pattern analysis insight for the supply-chain



CDC Streaming  
Data Catalog  
Data Warehouse





# TEXTRON

 SPECIALIZED VEHICLES

“Databricks Data Intelligence Platform and Qlik are the backbone of our new data architecture supporting SAP S/4HANA. These solutions are driving easier data access and improved reliability across our business.”



CDC Streaming  
Data Catalog  
Data Warehouse



**LOWERED COMPUTING COSTS**  
Implementing best practices to maximize operational efficiency and optimize business performance.

**MINUTES TO REFRESH DATA**  
By connecting Qlik Replicate to new data every few minutes, TSV is making it available to users in near-real time.

**25% REDUCTION**  
Less time on running & maintaining data environment. Staff focus on other high-value priorities.



## Customer Story



“How do you scale up analytics without blowing a hole in your technology budget? For us, the clear answer was to run all our workloads on Databricks Data Intelligence Platform and replicate our data in near real-time with Qlik.”

Enhanced retail insights with 80-90% improvement in runtime of retail analytics.

Replicated ERP data in near real-time enabling valuable insights for sales and customer service teams.

Migrated to Databricks Data Intelligence Platform, which significantly increased processing speeds and allowed for the unification of data from global sources.

Qlik.



CDC Streaming  
Data Catalog  
Data Warehouse



# SUMMARY – 6 Critical Principles for Data Readiness

1 **Diverse**

Unbiased across silos

2 **Timely**

Up to date and real-time

3 **Accurate**

Reliable and trustworthy

4 **Secure**

Protected from unauthorized use

5 **Discoverable**

Easier to find and understand

6 **Consumable**

In a form that Analytics can consume

## TRUSTED DATA FOUNDATION

 + 



THANK YOU !

Sharad Kumar  
[/in/sharad-kumar67](#)