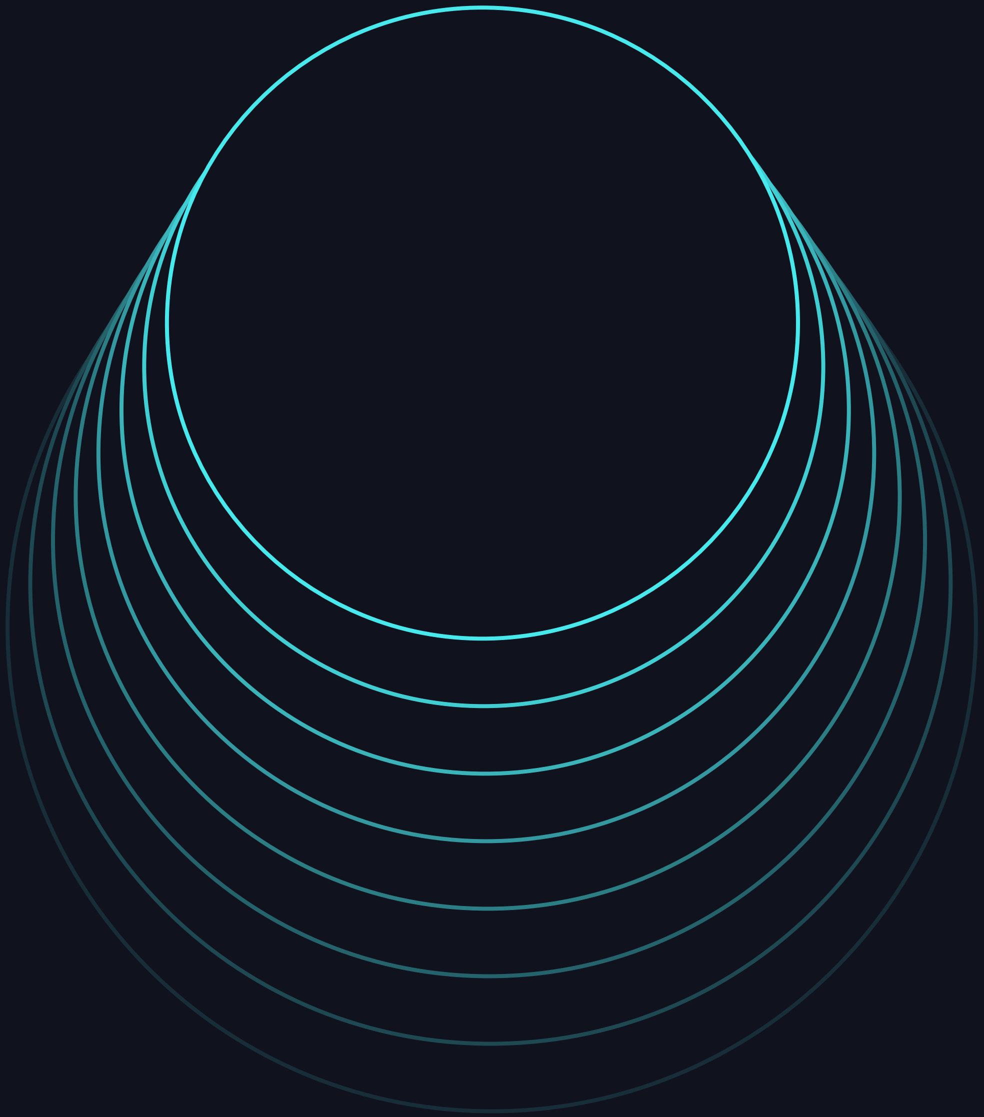
DATA+AI SUMMIT
BY databricks

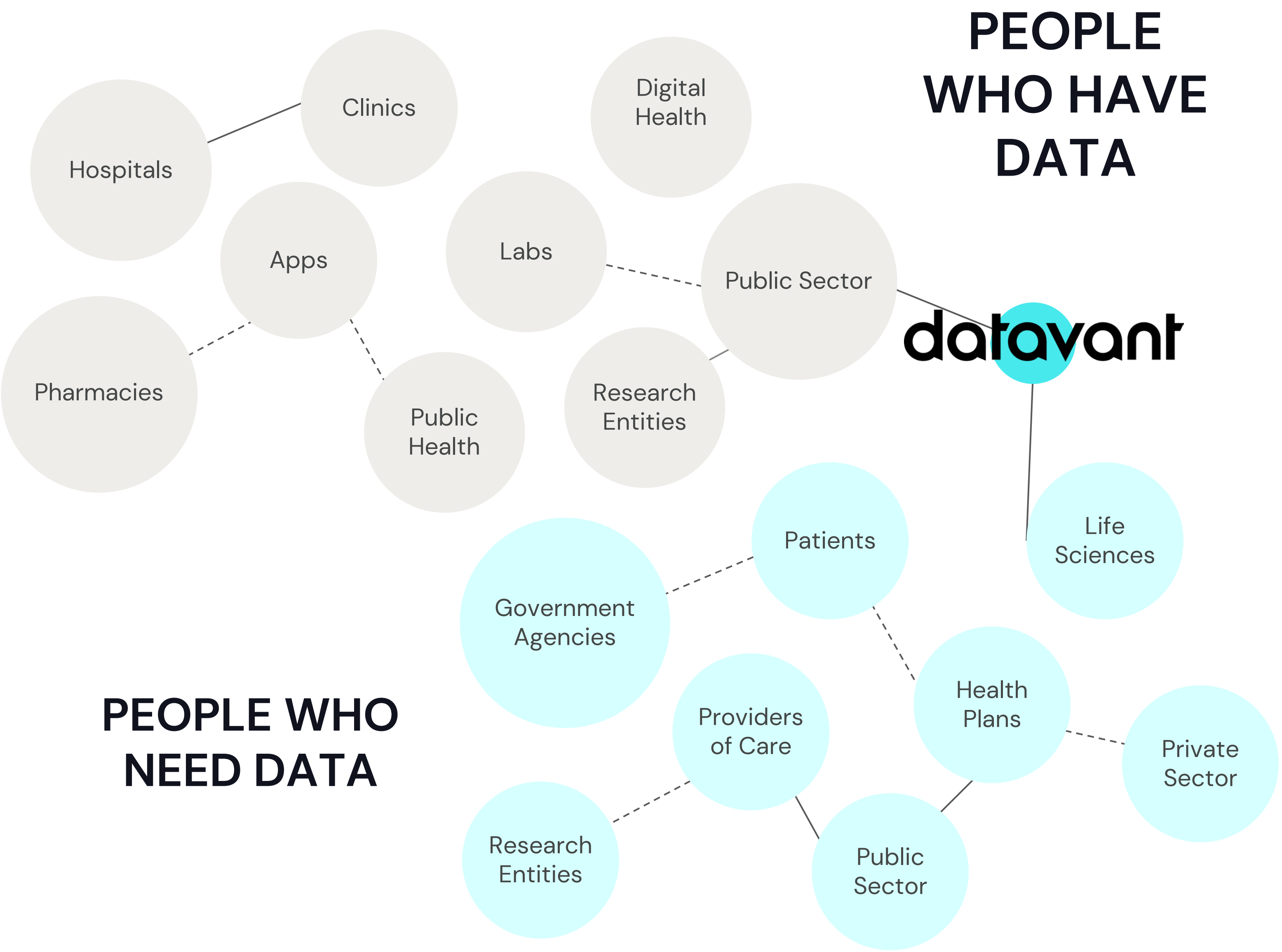
Sparking real world data connectivity: Datavant + Databricks Clean Rooms



Garrett Little, Customer Success, Datavant
Mark Lee, Industry Solutions Lead, Databricks
June 11th 2024

Datavant is solving the broken data supply chain issue

Datavant helps connect people who **have** data securely share it with people who **need** data



A dysfunctional supply chain hurts everybody

Health Care Providers

Develop policy without population data

~~Life Sciences~~

Cannot access fit-for-purpose RWD

~~Other Private Sector~~

Inability to access the right data to inform decisions

~~Health Plans~~

Struggle to optimize funding

~~Public Sector~~

Develop policy without population data

~~Patients~~

Suffer from a system with slowed innovation and insufficient care

Health data needs to move **securely, compliantly, and frictionlessly** from where it sits, to where it needs to be

- → From any source
- → In any format
- → To solve every use case

datavant

Bringing Data Logistics to Healthcare

datavant + databricks

In 2023 Datavant partnered with Databricks to bring patient data connectivity solutions to where the data resides

Growing Marketplace Ecosystem

AnalyticsIQ · AccuWeather · BlockFills · Bloomberg Government · BMLL · CATALINA · BrightQuery

CFTC COMMODITY FUTURES TRADING COMMISSION · Circana. · Collectors Data Store · Commerce Signals A TransUnion Company · CoreLogic · divirod · crisp.

equiduct · experian. · Facteus · FiscalNote · F SQ · LSEG · GMG

zoominfo · IQVIA · John Snow LABS · kumo · KYTHERA · /LiveRamp · Lend-Rx TECHNOLOGY

MORNINGSTAR · narrative · Nasdaq · OECD · Orbital Insight · Porch GROUP MEDIA · People Data Labs

PredictHQ · rearc · Ribbon · RS Metrics · S&P Global Market Intelligence · SOLEADIFY · sharethis

Sucali · tégo · TriNetX · VAISALA · veraset · WMBA LIMITED · VERTICAL KNOWLEDGE

Carto · DEFINITIVE HEALTHCARE · Habu · LexisNexis · Polk S&P Global Mobility · SIX · WEATHER SOURCE

Databricks Marketplace

170+
providers

1,700+
listings

100K+
visitors

80% QoQ
growth in active connections

Datavant's purpose-built solution on Databricks maximizes cloud investments

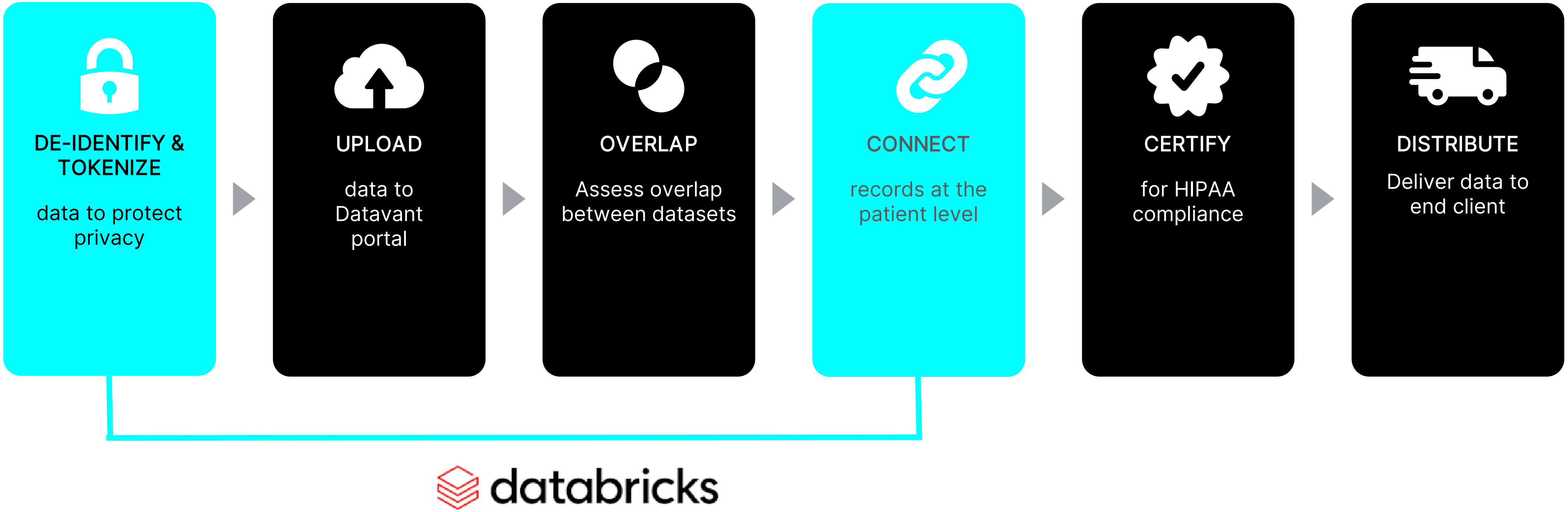
Improve data processing

Connect to Real World Data

Leverage privacy enabled analytics solutions

End-to-end solution on your cloud environment

Datavant solutions now natively usable within Databricks



DE-IDENTIFY & TOKENIZE

data to protect privacy

UPLOAD

data to Datavant portal

OVERLAP

Assess overlap between datasets

CONNECT

records at the patient level

CERTIFY

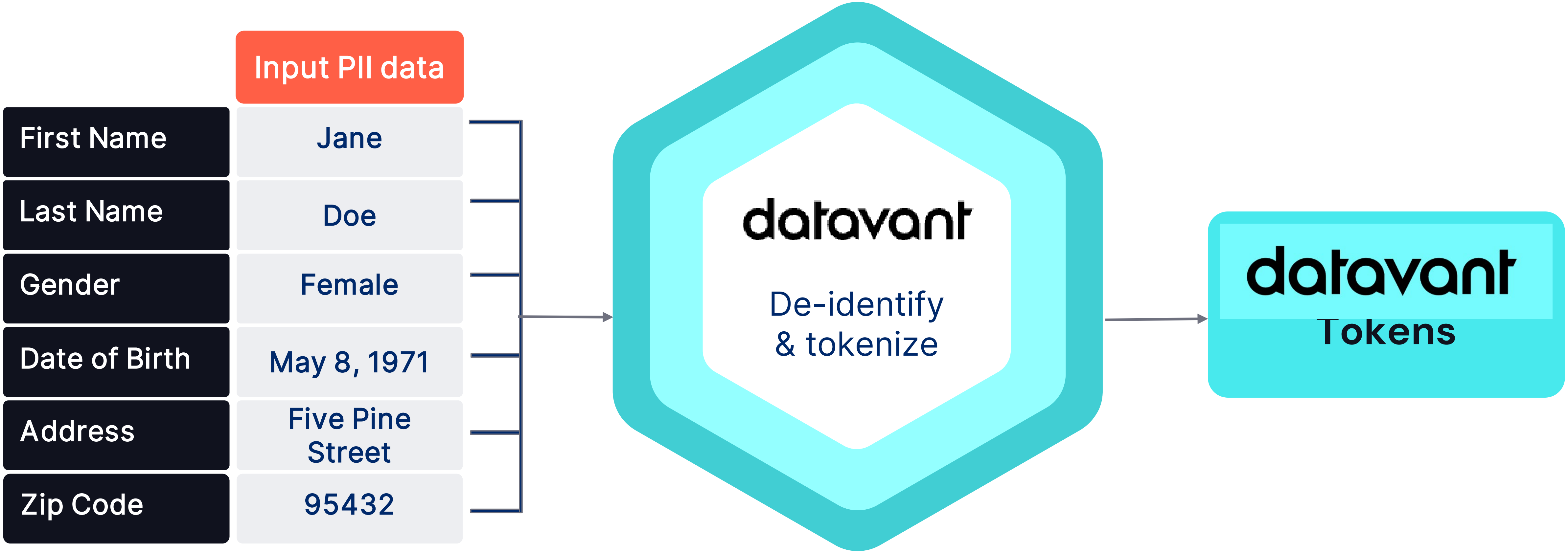
for HIPAA compliance

DISTRIBUTE

Deliver data to end client

databricks

Datavant's tokenization and privacy preserving connect keys integrated with Databricks



	Input PII data
First Name	Jane
Last Name	Doe
Gender	Female
Date of Birth	May 8, 1971
Address	Five Pine Street
Zip Code	95432

datavant

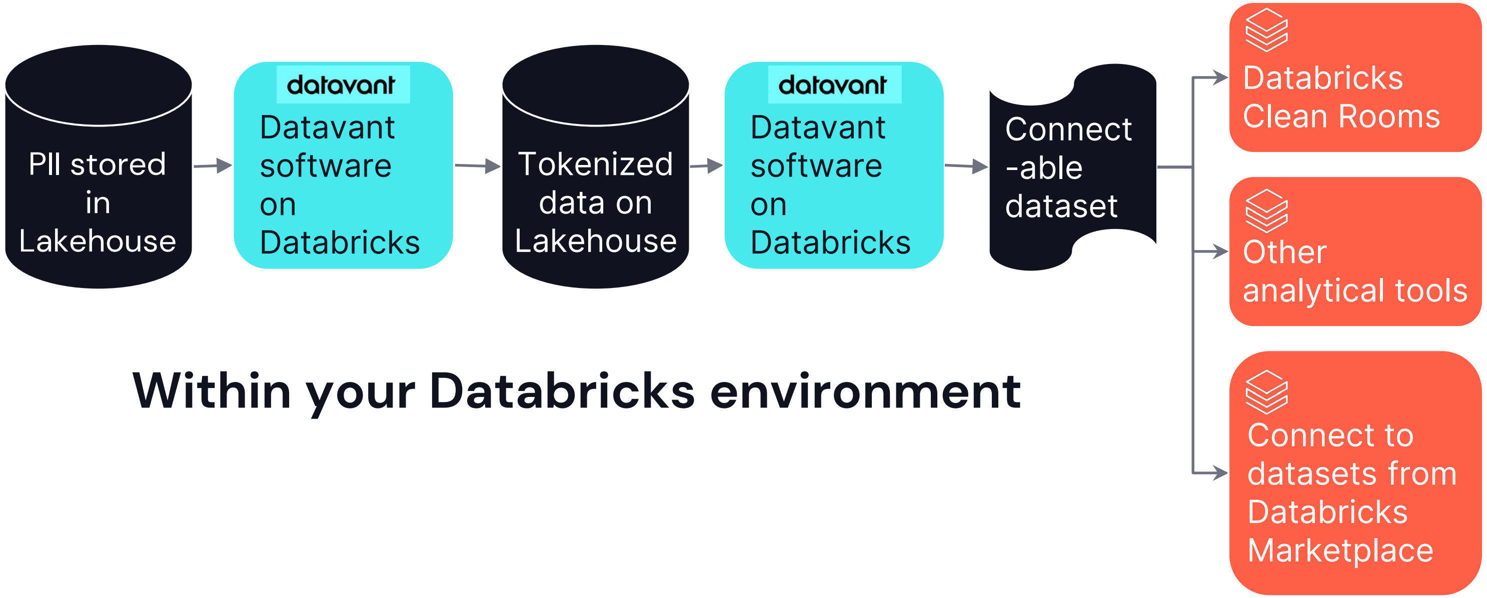
De-identify & tokenize

datavant

Tokens

databricks

Datavant on Databricks simplifies data processing



PII stored in Lakehouse
datavant
Datavant software on Databricks
Tokenized data on Lakehouse
datavant
Datavant software on Databricks
Connect -able dataset
Databricks Clean Rooms
Other analytical tools
Connect to datasets from Databricks Marketplace

Within your Databricks environment

- ✓ Data stays in your Databricks Lakehouse environment
- ✓ Tokenized data can be connected to internal and external datasets
- ✓ Data can be used in Databricks analytics tools, including Clean Rooms

CODE SAMPLE

Sample Cleanroom Code

PYTHON

```python
"""
Getting datasets in a cleanroom
"""

from pyspark.sql.functions import col, to_date

# Read in data assets from the cleanroom
dfA = (spark.read.format("delta").table("collaborator_a.cleanroom.a_tokens")
           .withColumn("DOB", to_date(col("DOB"), "yyyy/MM/dd"))
           .withColumn("ServiceDate", to_date("ServiceDate", "yyyy/MM/dd")))

dfB = (spark.read.format("delta").table("ml_demos.datavant.cleanroom_b_tokens")
           .withColumn("DOB", to_date(col("DOB"), "yyyy/MM/dd"))
           .withColumn("ServiceDate", to_date("ServiceDate", "yyyy/MM/dd")))
```

CODE SAMPLE

Sample Cleanroom Code

PYTHON

```python
"""
Join the two datasets along tokens and get patient collisions
"""

from pyspark.sql.functions import col

# join the two datasets together using token_1 and token_2
overlapping_rows = (dfA.join(dfB, on = ["token_1", "token_2", "ServiceDate", "DOB", "Diagnosis"], how = "inner"))


# print the count of identified patients
print(f"Number of identified patients : {overlapping_rows.count()}")
```

Case study: Veritas

Veritas was having challenges using Datavant on prem

- The Datavant on prem product was taking Veritas 24+ hours to tokenize, transform, and send
- Every hour was a delay in delivering data assets to data buyers

They tested the Datavant on Databricks integration

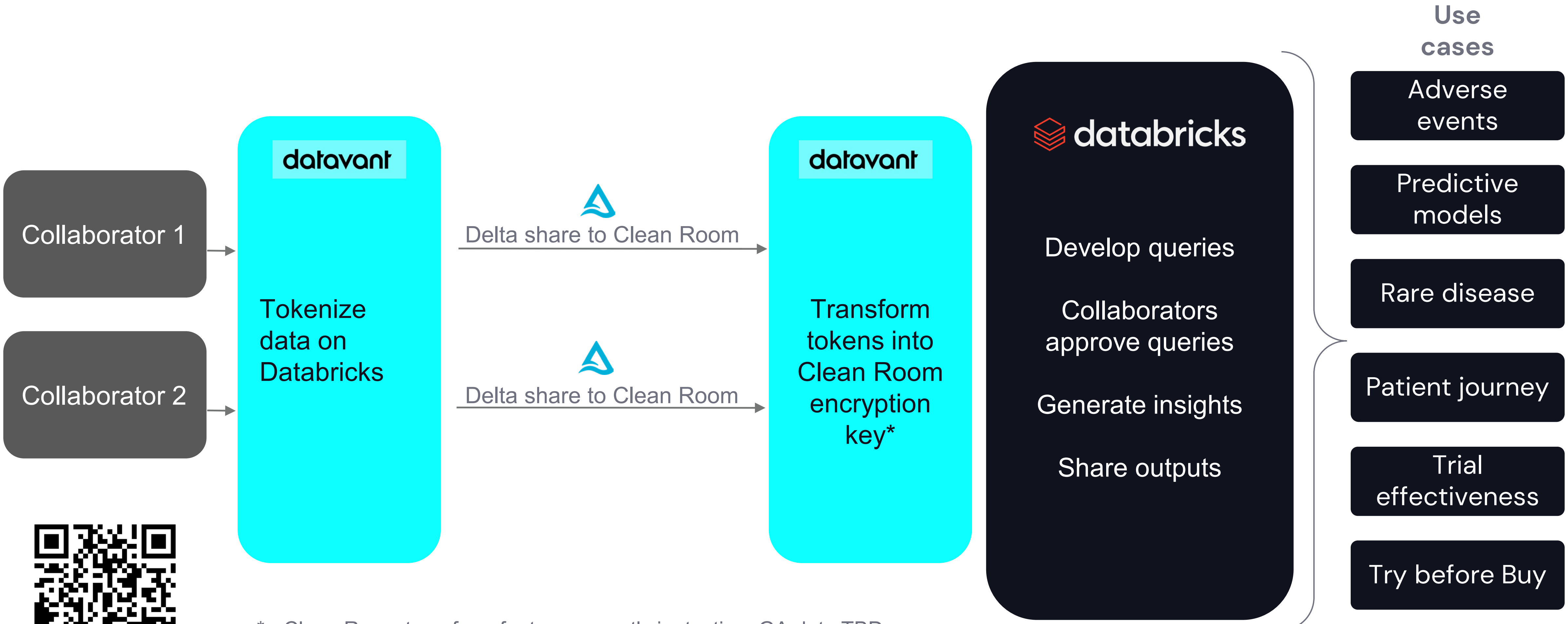
- Veritas became a pilot customer of the Datavant on Databricks product
- They tested workloads of 40-80M patient datasets stored on Databricks

Veritas saw significant speed improvements

- Veritas realized the power of Databricks' Spark technology
- Veritas achieved 4x time savings using the native application
- 80 million records were tokenized in 20 minutes, transformations in 5, and delivery in 15 minutes

RESULTS: ~4x time savings ● ~45% FTE time reduction ● ~75% CPU cost reduction

Databricks and Datavant are enabling a new way to analyze data in healthcare



Collaborator 1

Collaborator 2

datavant

Tokenize data on Databricks

Delta share to Clean Room

Delta share to Clean Room

datavant

Transform tokens into Clean Room encryption key*

databricks

Develop queries

Collaborators approve queries

Generate insights

Share outputs

Use cases

- Adverse events
- Predictive models
- Rare disease
- Patient journey
- Trial effectiveness
- Try before Buy

* - Clean Room transform feature currently in testing, GA date TBD

Our partnership enables new use cases for healthcare and life sciences

Opening the ecosystem for new data sources who want stricter controls on how their data is used

Enriching proprietary data with RWD for analytics on patient outcomes

Accelerating time to market for new interventions or drugs

Better understanding of the patient journey and intervention effectiveness

When Would This Make Sense for You?

If you have...

- High data volume
- Continuous data processing / frequent transformation
- Your team is spending time on upgrading / managing CLI
- You want to simplify your data pipelines for tokenization/linking
- You want to save money (no promises but possibility!)

datavant

Protect.
Connect.
Deliver.