datahistorian

You call it Data Lake; we call it Data Historian

Naghman Waheed – Data Platforms Lead
Brian Arnold – Data Platforms Architect

May-24-2018

MONSANTO

Naghman Waheed

Data Platforms Lead

- 25+ year career at Monsanto.
- Data Warehousing, Business Intelligence, Data Architecture, Cloud Engineering.
- Data solutions spanning key business functions such as Supply Chain, Manufacturing, Order-To-Cash, Finance and Procurement.

Brian Arnold

Data Platforms Architect

- 10 year career in IT, 6 years in Big Data
- Software Development, Functional Programming, Streaming, Big Data, Cloud Engineering
- Ecommerce, Recommendation Engines

MONSANTO
INFORMATION TECHNOLOGY

Monsanto - Who are we?

- Bringing a broad range of solutions to help nourish our growing world
- Headquartered in Saint Louis, Missouri
- >20,000 employees in 66 countries
- A global company with >50% employees based outside of the United States
- One of the 25 World's Best Multinational Workplaces by Great Place to Work Institute

FORTUNE WORLD'S MOST ADMIRED COMPANIES 2014

Science 2015 TOP EMPLOYER

DiversityInc TOP 50 2015 COMPANIES FOR DIVERSITY

CR's 100 Best Corporate Citizens 2015

DIGITAL EDGE25 AWARDS 2016

BEST PLACES TO WORK 2015 for LGBT Equality 100% CORPORATE EQUALITY INDEX

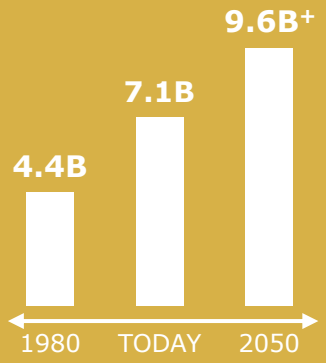
MIT Technology Review 50 SMARTEST COMPANIES

MONSANTO
INFORMATION TECHNOLOGY

Solving real challenges in agriculture industry

Rising Population

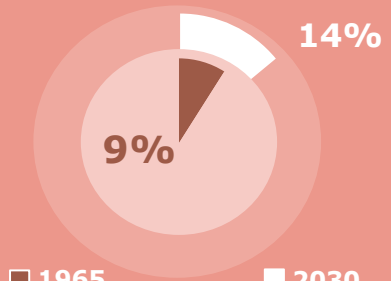
Growing enough for a growing world



Global Population

Changing Economies and Diets

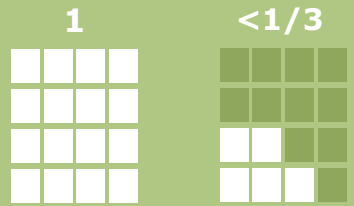
A growing global middle class is choosing animal protein – meat, eggs, and dairy – as a larger part of their diet



Dietary Percentage of Protein

Limited Farmland

Farmers will need to produce enough food with fewer resources to support our world population



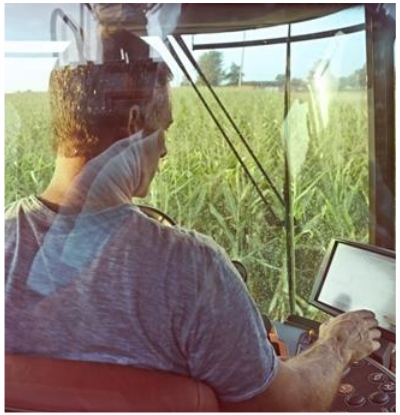
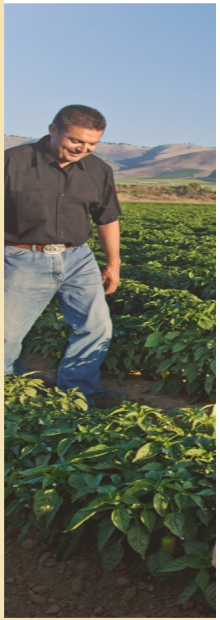
Acres per Person

Changing Climate

Farmers are impacted by climate change in many ways:

- WATER AVAILABILITY ISSUES
- INCREASINGLY UNPREDICTABLE WEATHER
- INSECT RANGE EXPANSION
- WEED PRESSURE CHANGES
- CROP DISEASE INCREASES
- PLANTING ZONE SHIFTS

MONSANTO
INFORMATION TECHNOLOGY

Our Solutions for Sustainable Agriculture

Our toolkit includes:

- Plant Breeding
- Biotechnology
- Crop Protection
- Precision Agriculture

MONSANTO
INFORMATION TECHNOLOGY

Key Technology Trends In Agriculture

Economies of Data Science at Scale

A typical farm is generating 20GB of unique field data every year

Computing unit costs have gone down by 1,000x in last 10 years

Low-cost Observation Technology /IoT

Connected sensors on tractors, combines, and in fields has increased over 1000x in the last 10 years

The cost of the average digital sensor had dropped more than half over that time

Mobile Device Proliferation among Growers

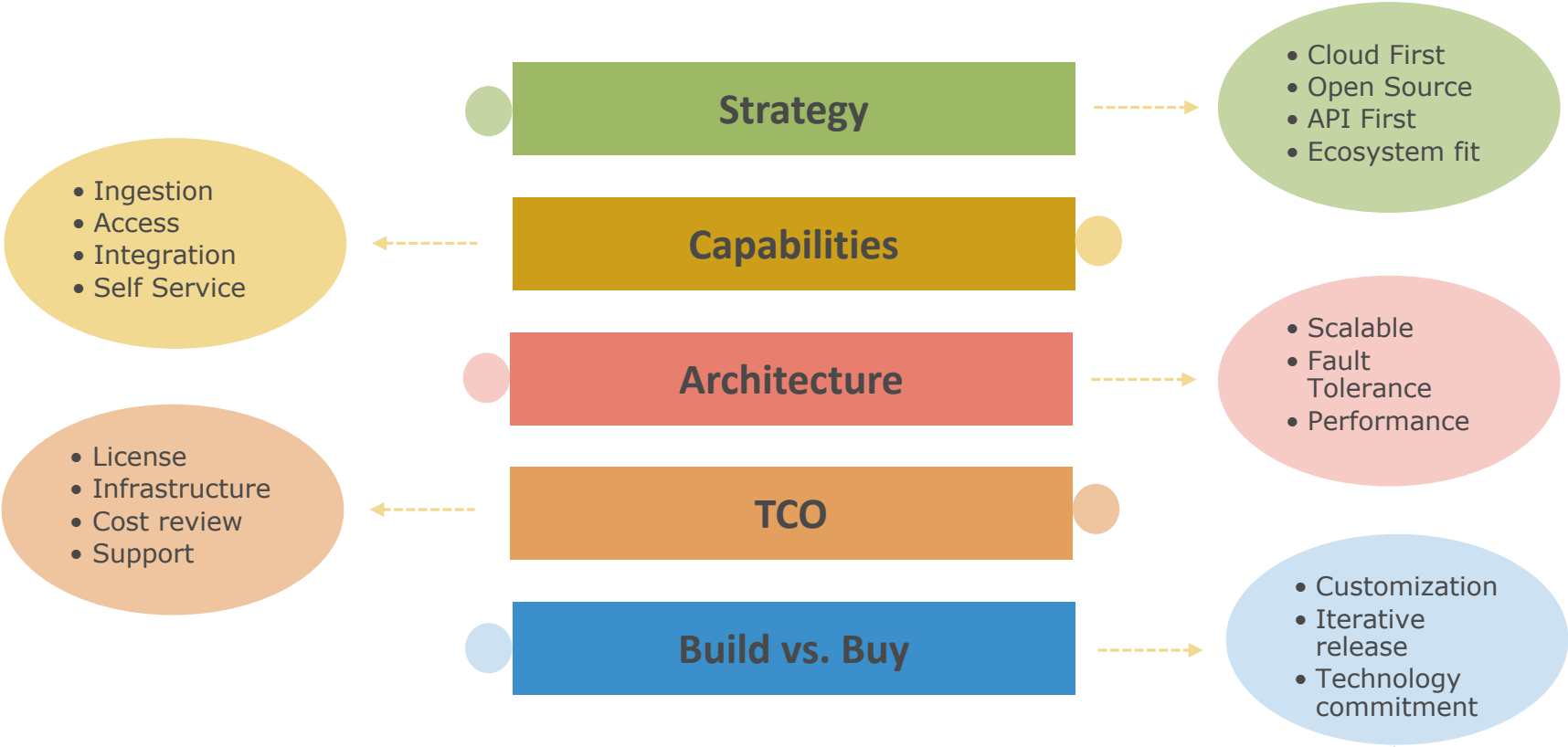
94% of US farmers own a mobile phone or a smartphone

Compared to less than 10& 10 years ago

Source : Gartner Technology Trends 2015

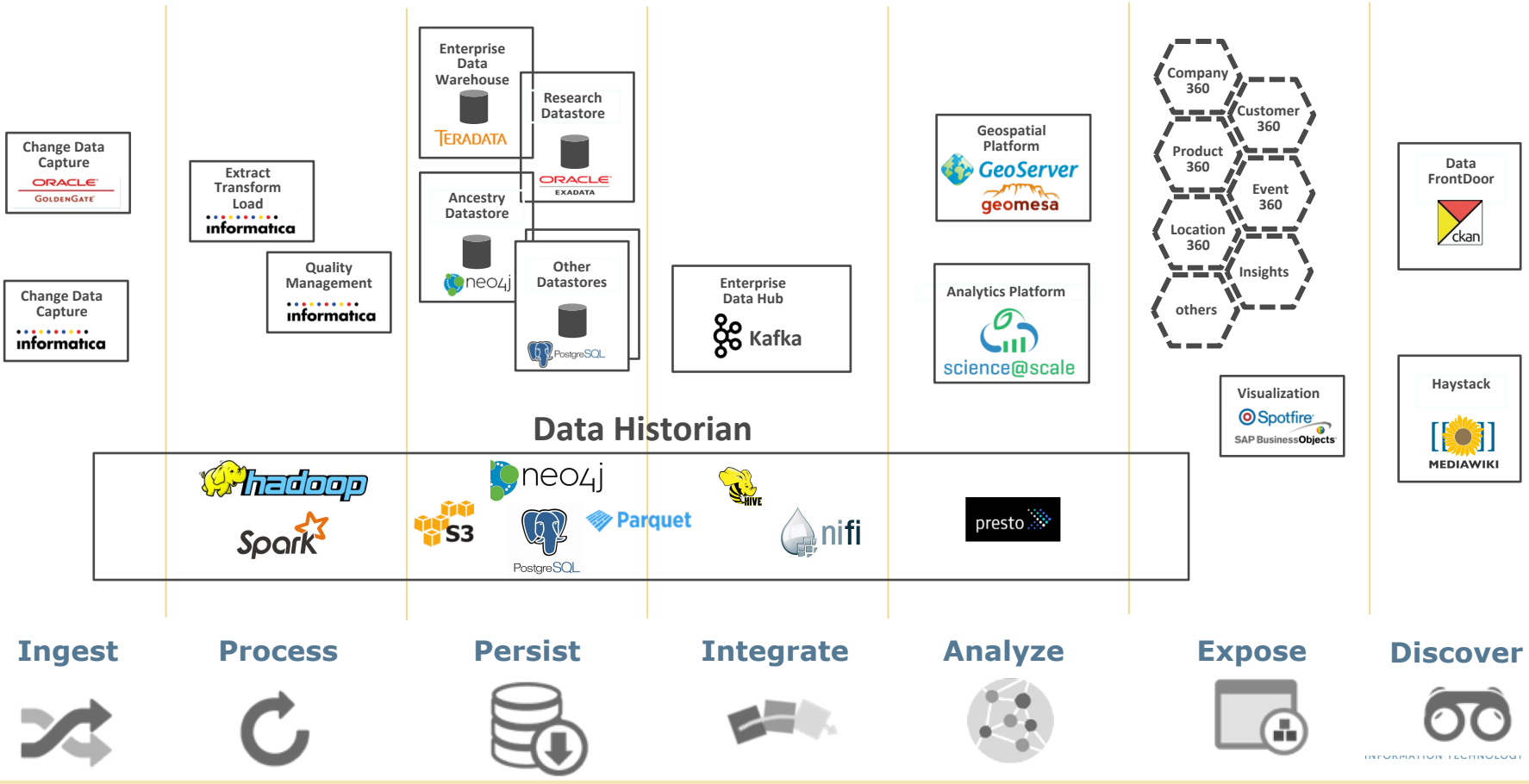
MONSANTO
INFORMATION TECHNOLOGY

Why Data Historian?

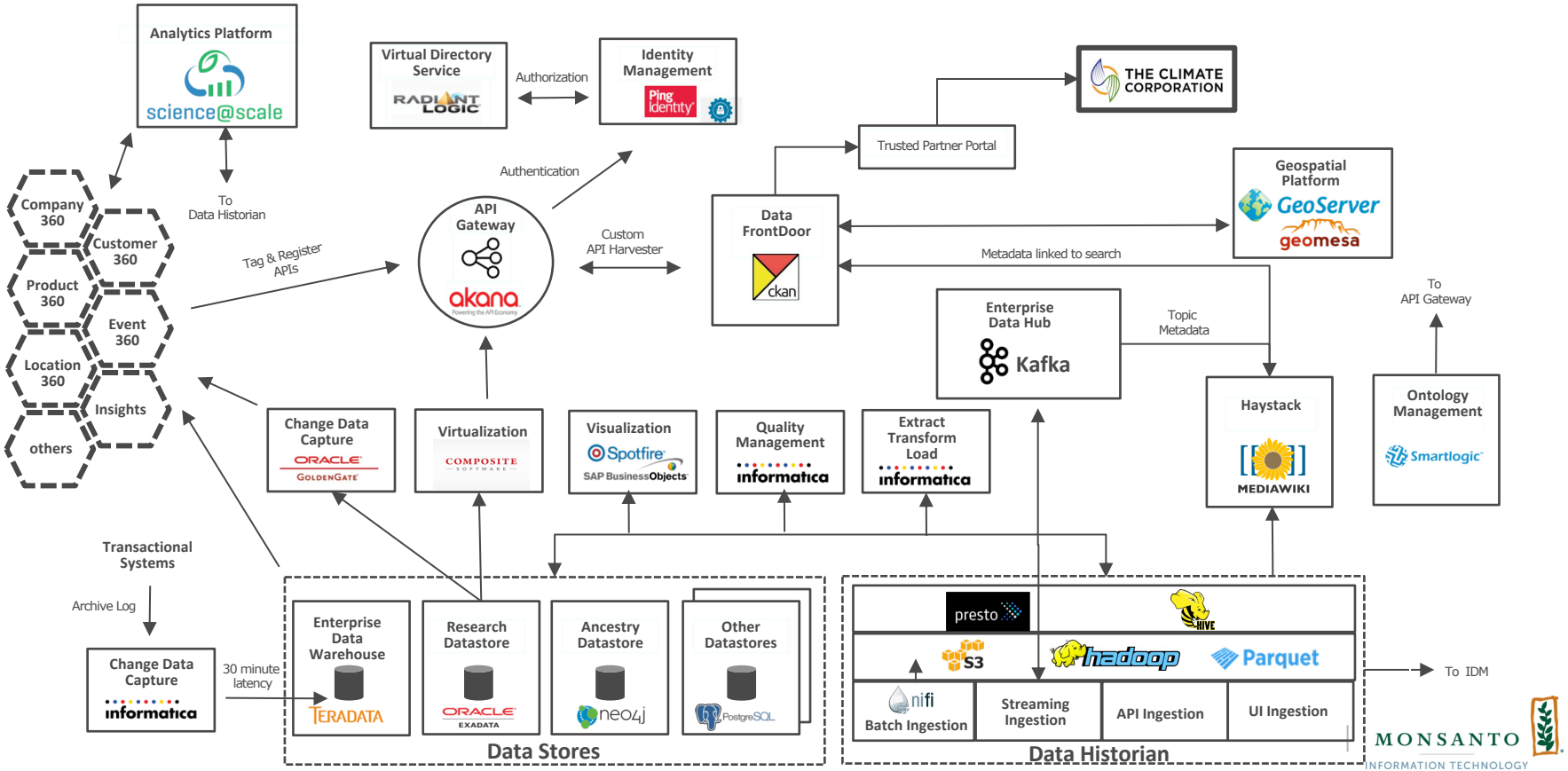


MONSANTO
INFORMATION TECHNOLOGY

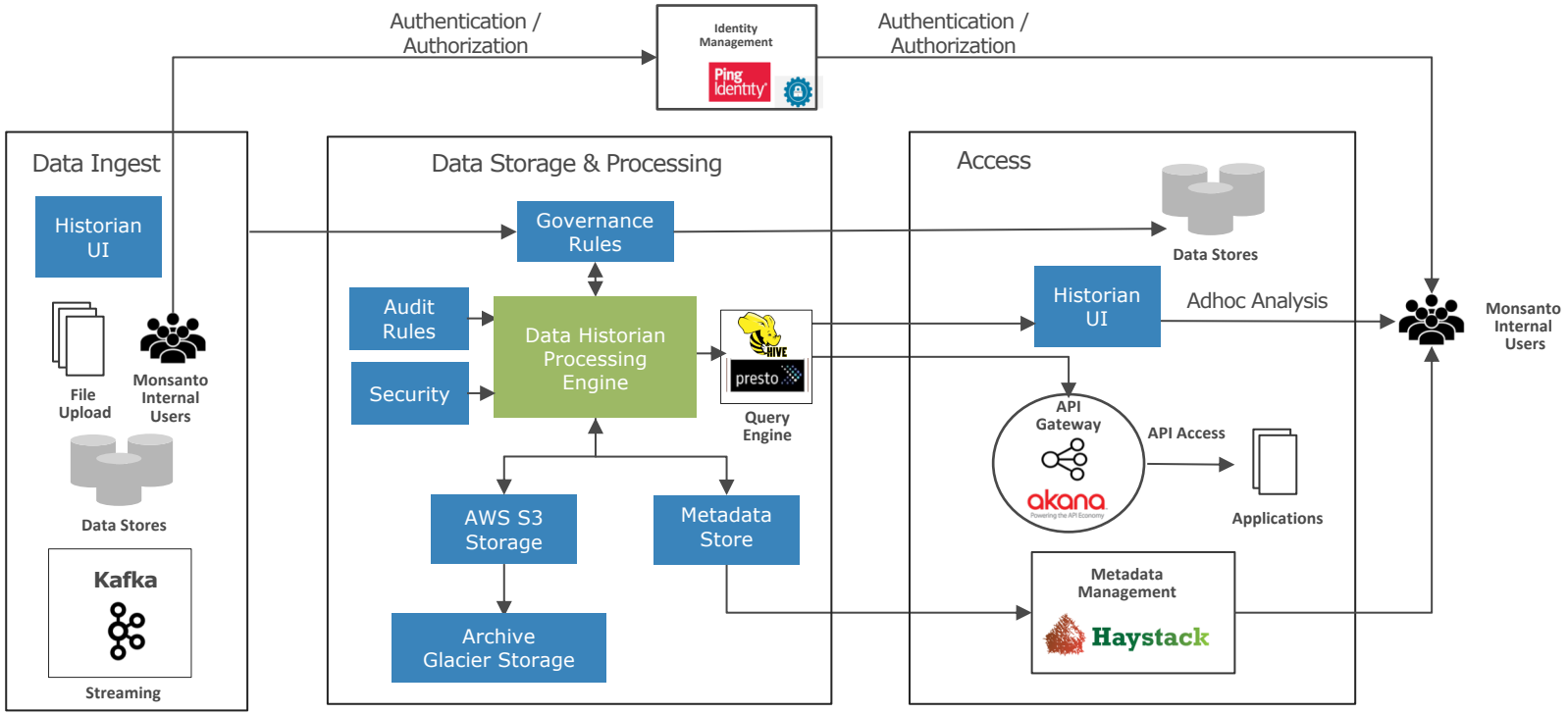
Data Strategy



Data Platforms Ecosystem



Data Historian - Reference Architecture

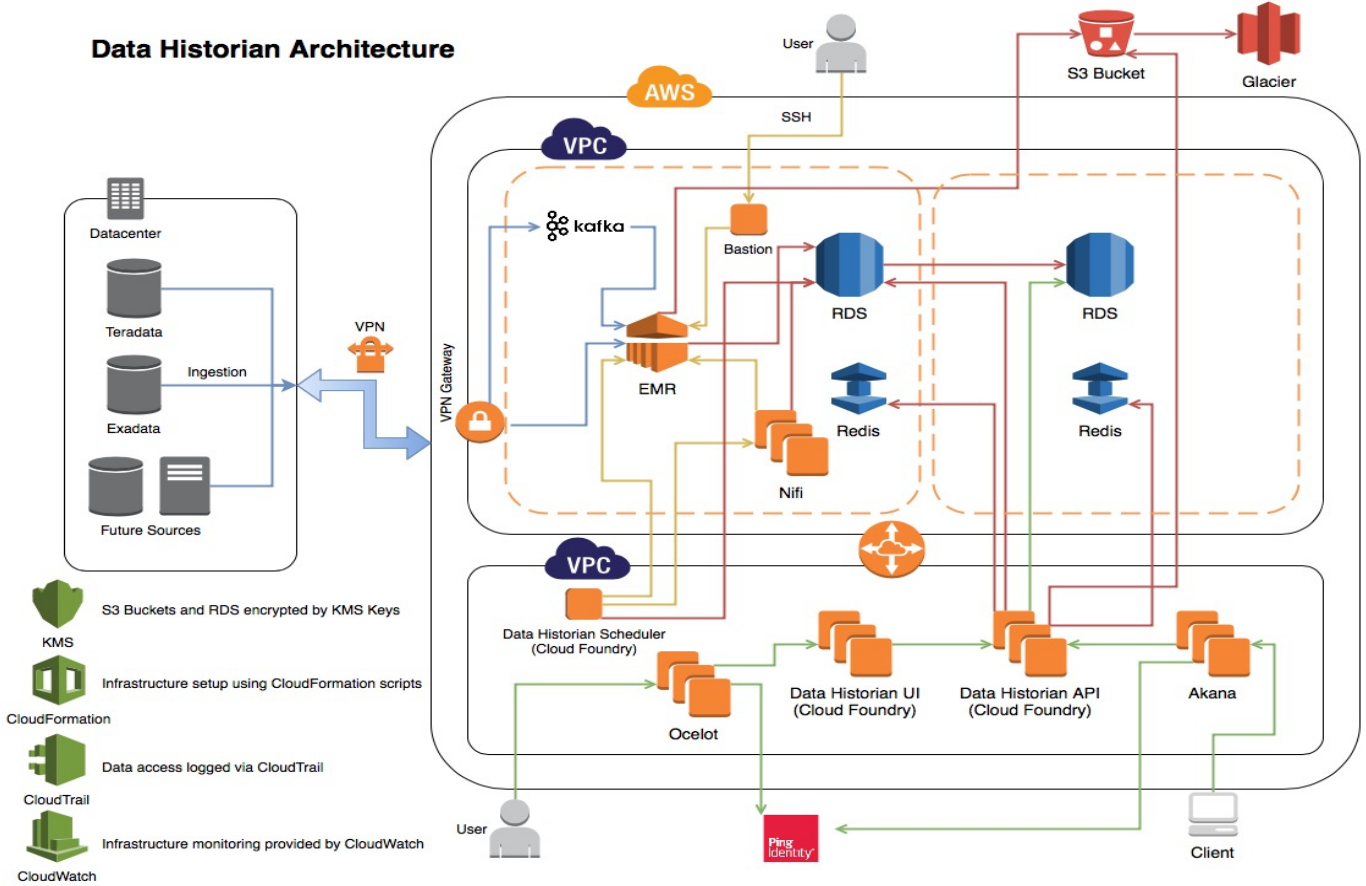


MONSANTO
INFORMATION TECHNOLOGY

Data Historian – Technology

MONSANTO
INFORMATION TECHNOLOGY

AWS Data Historian Architecture



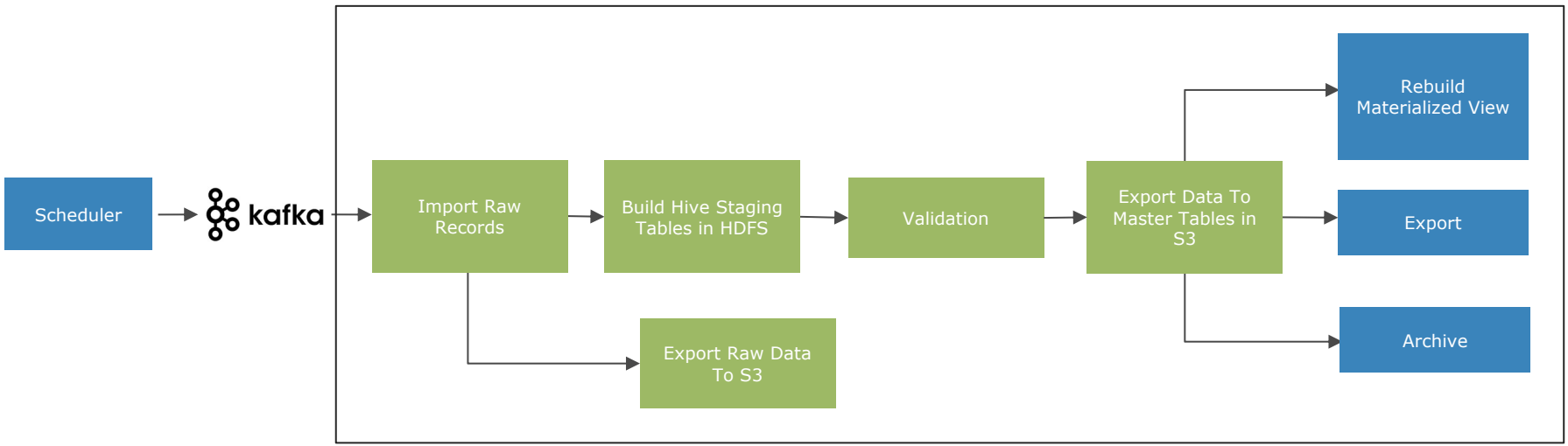
MONSANTO
INFORMATION TECHNOLOGY

Ingest

- Batch imports from RDBMS
  - Full, delta, merge
- Streaming from Kafka (Datahub)
- File ingestion through API and Data Historian UI
  - Users can append files to existing datasets as well

MONSANTO
INFORMATION TECHNOLOGY

Ingestion Process



MONSANTO
INFORMATION TECHNOLOGY

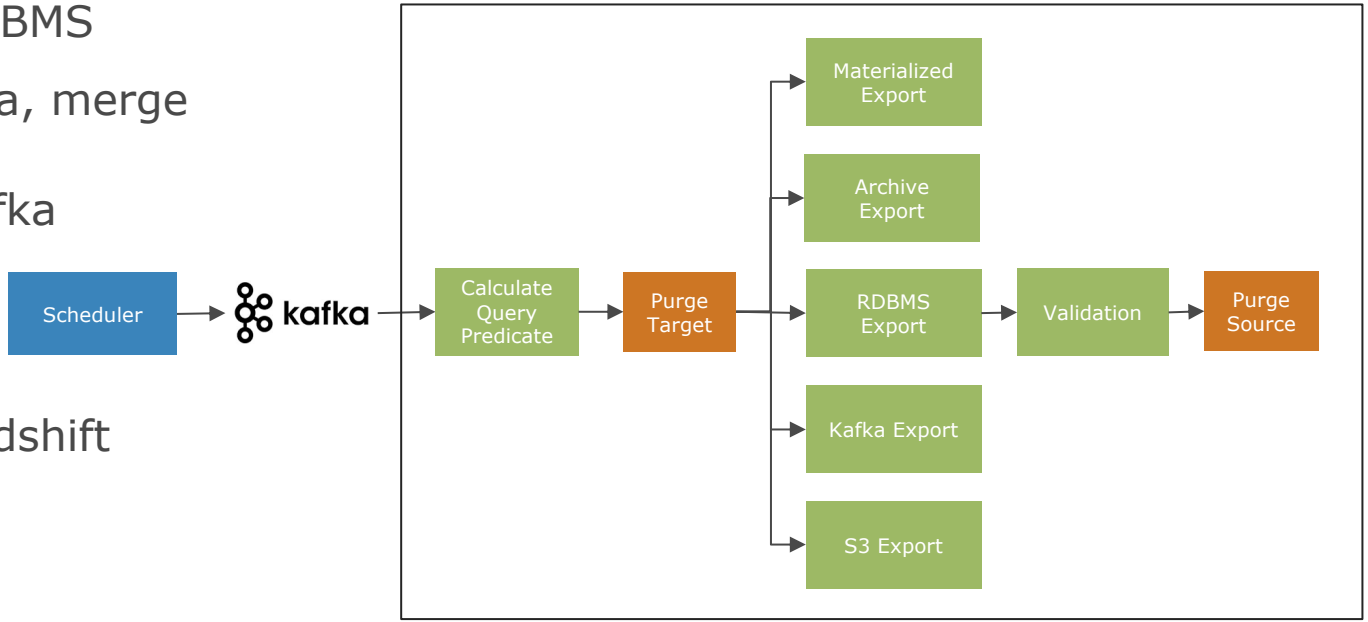
Metadata

- Required fields
  - Name, Description, Source, Publisher, etc.
- Optional fields
  - Tags, custom fields
- Forwarded to our metadata platform (Haystack)
  - Metadata objects pushed through Kafka (Datahub)

MONSANTO
INFORMATION TECHNOLOGY

Exports

- Export to RDBMS
  - Full, delta, merge
- Export to Kafka
- Export to Redshift
- Export to S3
- Materialized Export



MONSANTO
INFORMATION TECHNOLOGY

Governance

- Archive & Retention
- Automated Compliance Checks
- Security
- Permissions

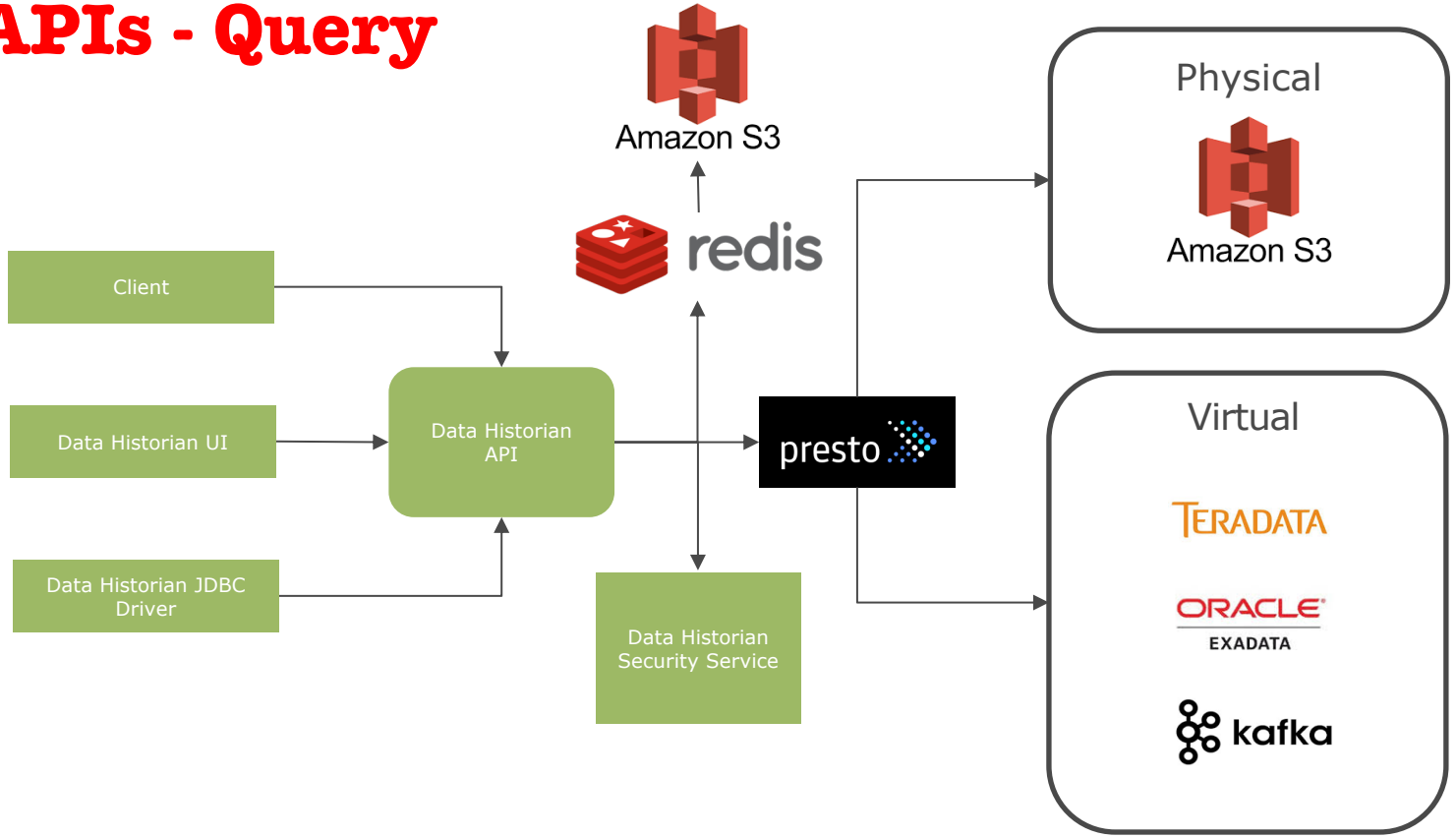
MONSANTO
INFORMATION TECHNOLOGY

APIs & Integration

- Get/List/Put Datasets
- Get/Put Dataset Metadata
- Get Dataset Status
- Query
- SDKs
  - Java, Scala, R, Python

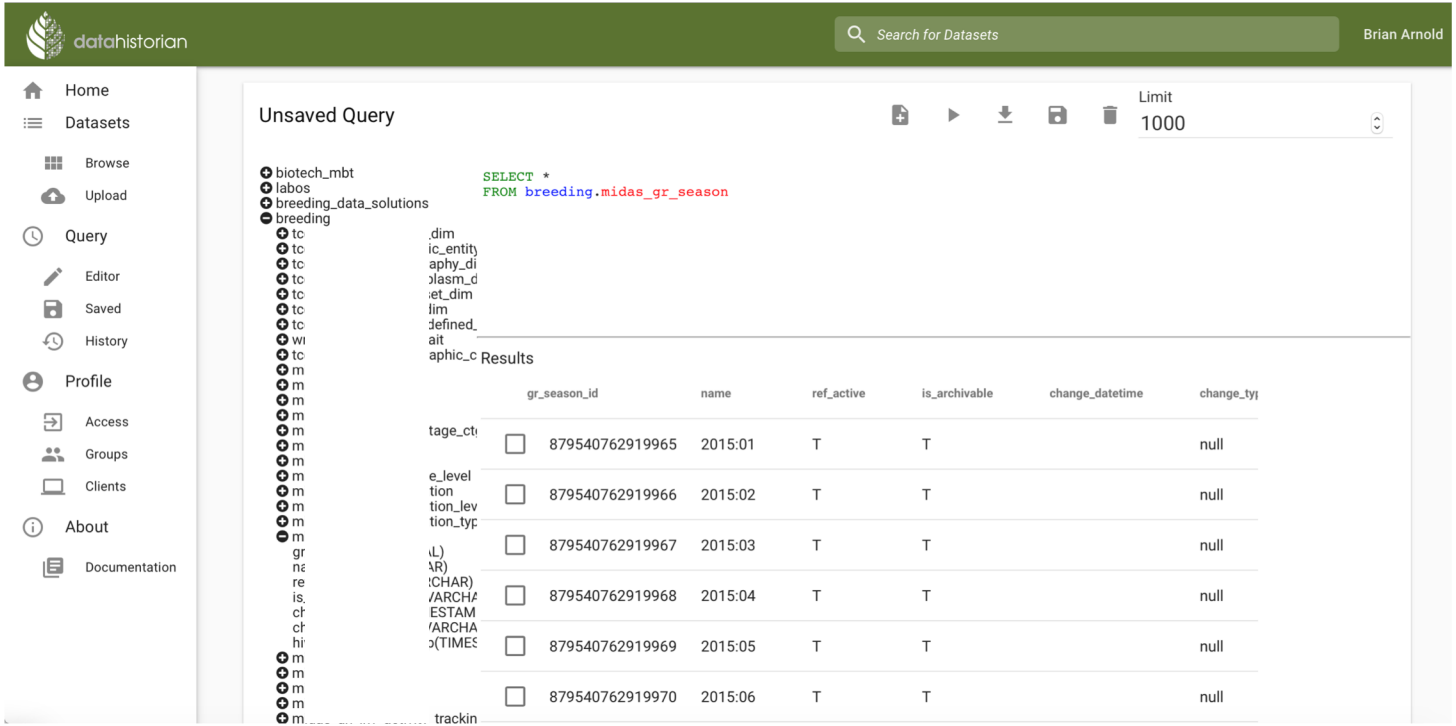
MONSANTO
INFORMATION TECHNOLOGY

APIs - Query



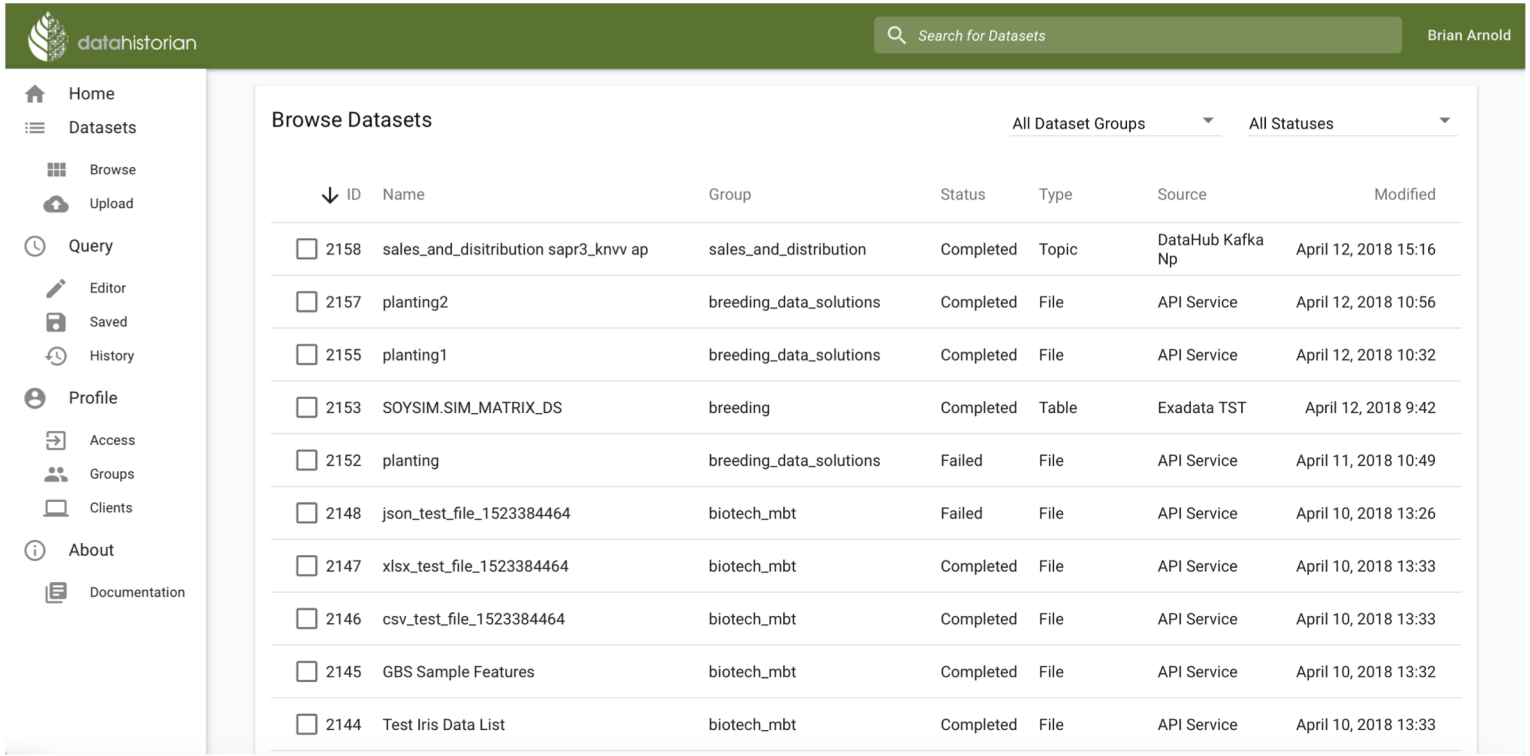
MONSANTO
INFORMATION TECHNOLOGY

Data Historian UI - Query Interface



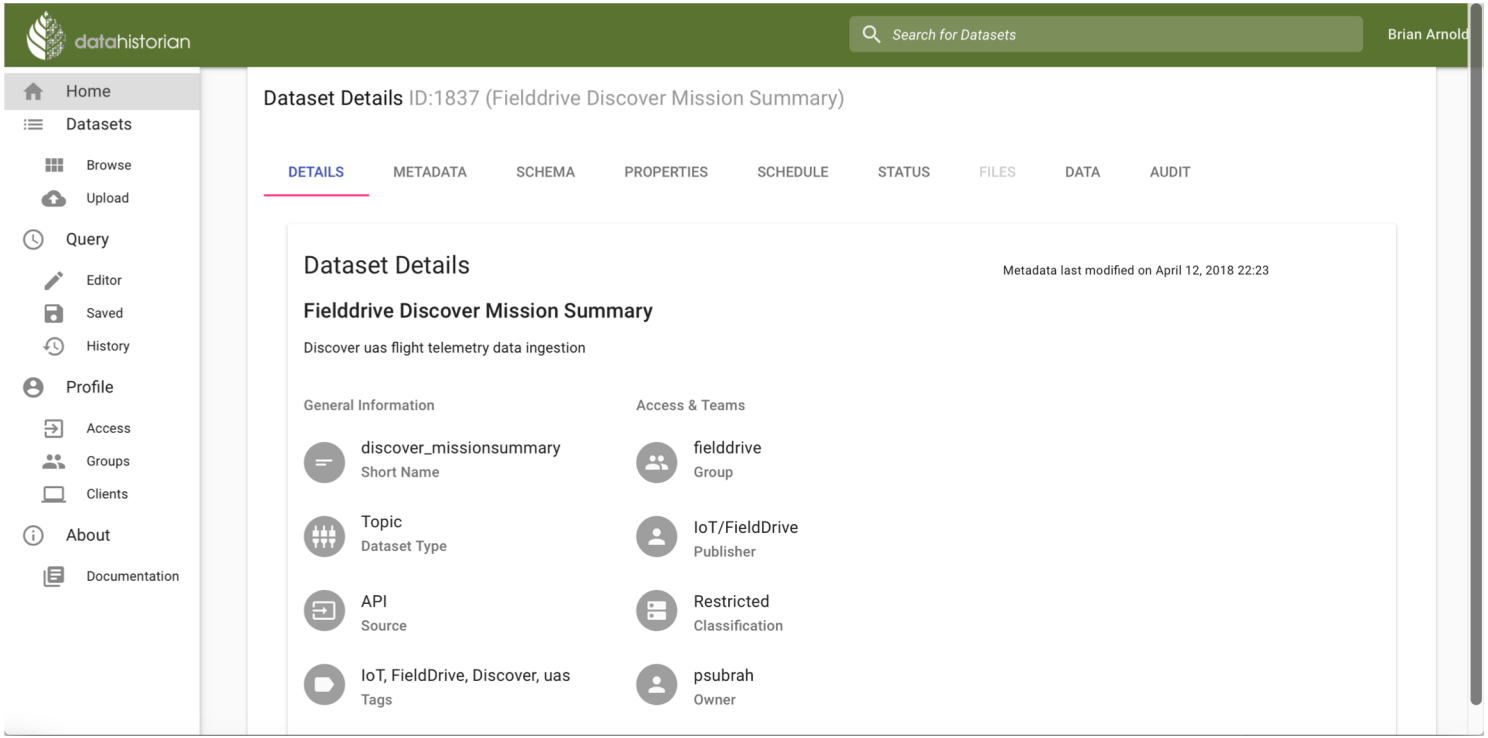
MONSANTO
INFORMATION TECHNOLOGY

Data Historian UI – Browse Datasets



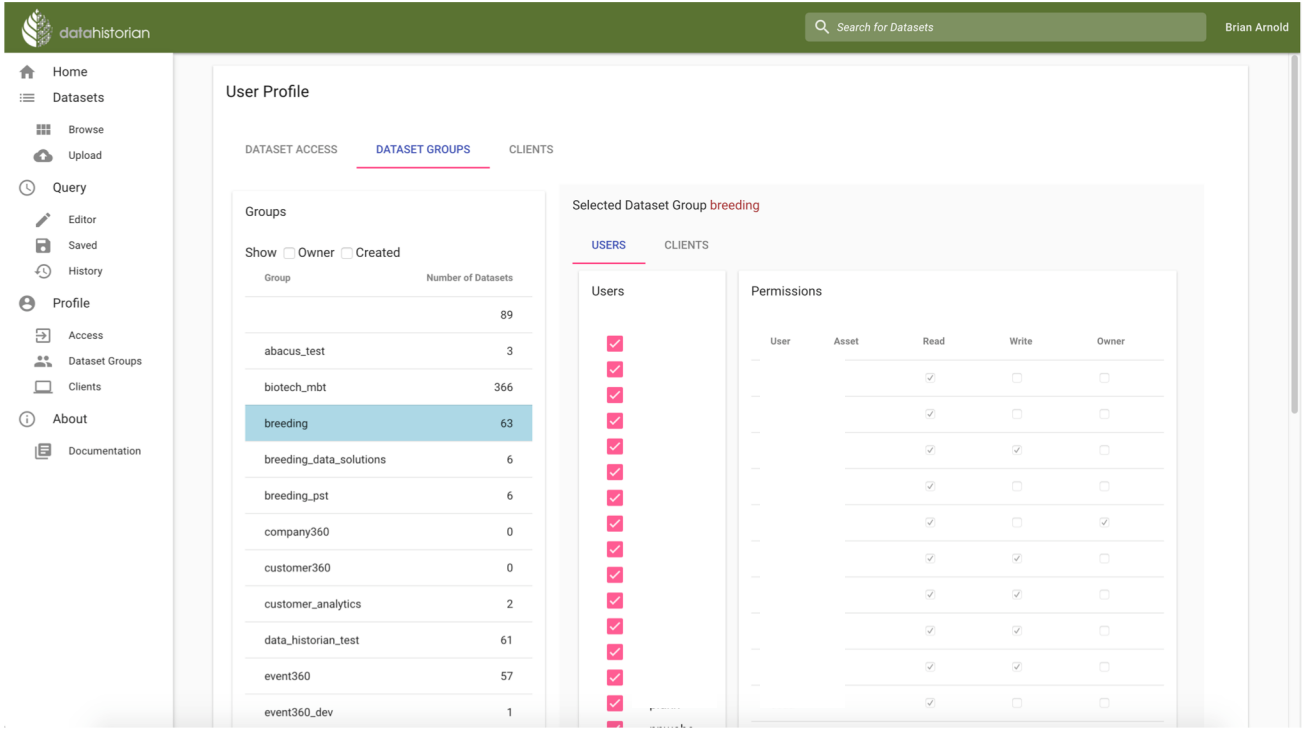
MONSANTO
INFORMATION TECHNOLOGY

Data Historian UI – Dataset Details



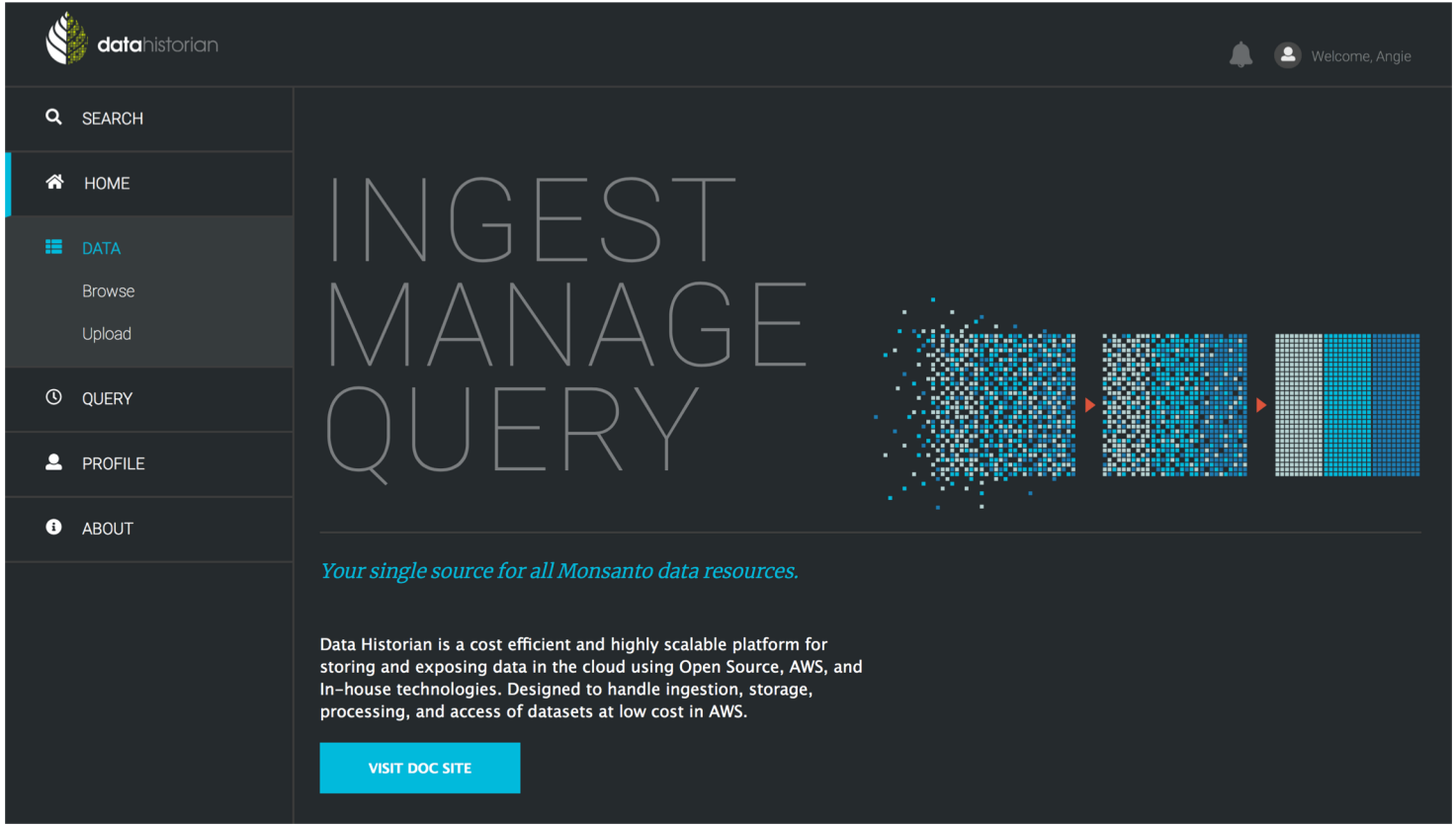
MONSANTO
INFORMATION TECHNOLOGY

Data Historian UI – Permissions Management



MONSANTO
INFORMATION TECHNOLOGY

Data Historian UI - Future



MONSANTO
INFORMATION TECHNOLOGY

Highlights

- v1.0 production release 16 months ago
- 164 active datasets in prod
- 10TB of data in prod
- >1,000+ query requests per day
- Early Adopters :
  - Internal Security Office, Research & Development
- Early Majority :
  - IoT, Data Assets, Supply Chain
- Late Majority :
  - Finance, Commercial, HR, Other

Lessons Learned

- Open Source
  - Flexibility
  - Learning Curve
  - Specialized Skill Set
- Cloud – AWS
  - Agility
  - Security
- Support
- Resource Staffing

Questions?