BEYOND.2022

ThoughtSpot

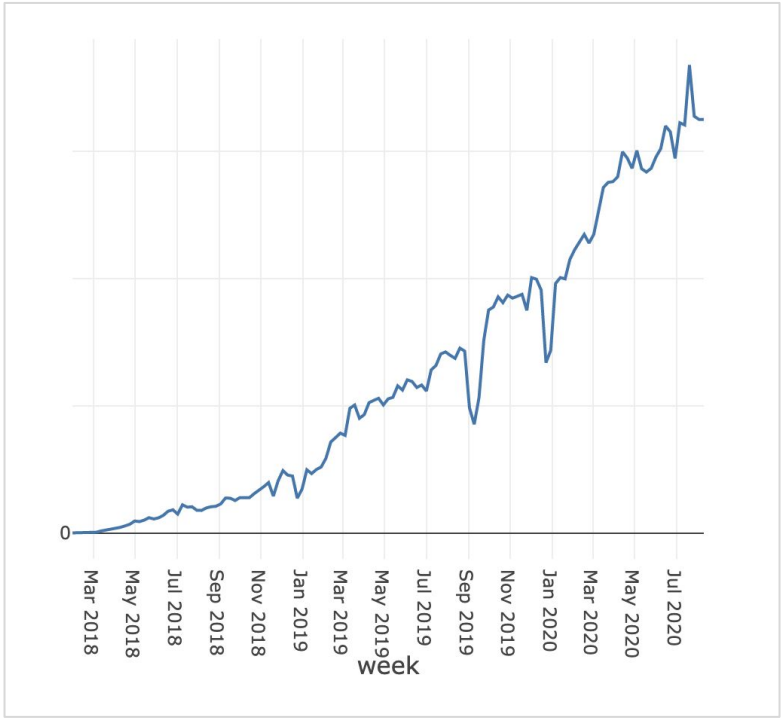
databricks

Data Lakehouse: The best of both worlds

Franco Patano
Lead Product Specialist
Databricks

Why build the Lakehouse ?

What we heard from our customers



SQL Usage Growth on Databricks (2018-2020)

1. Delta Lake enabled robust data management on the data lake.
2. Customers increasingly use SQL to directly query data lake data.
3. All the data is going to the lake. Only a portion gets into the data warehouse.

"We do not want another data warehouse"

-our customers

Delivering Analytics on the Lakehouse

Our Vision

- One source of truth for all your data
- World-class performance with data lake economics
- Data, analytics, and AI in one place

The Challenges Ahead

"Data lakes have no structure - they're just big data dumps - so it is impossible to turn it into a data warehouse."

"These two worlds are very different - one is based on files, the other on tables. It doesn't make any sense to blend them."

"It takes a very very long time to build a real data warehouse. People have been doing this for decades, ***you can't do this overnight****."*

"It will be impossible to leverage an open format like Parquet and get world-class performance out of it."

"Can you even layout the data automatically as you ingest it? In Spark, you have to specify partitioning, etc… but not in a data warehouse, how will you handle this ?"

"Can you even get concurrency out of a platform like this (for like 1,000's of users)? How would you handle concurrency on a framework built for batch analytics?"

"Can you even get existing BI and analytics tools to work with this open format and deliver data warehousing performance?"

"Data Scientists love the Databricks notebooks, but they are not suited for the majority of analysts"

"You can't get great performance out of open source tools"

"The data is not in this data mart, or the data warehouse"

- many smart people

Analyst Dilemma

Where is the data we need?

Business Partner

I need data on this new widget we put into production

Analyst

I don't see that data in the warehouse, let me check

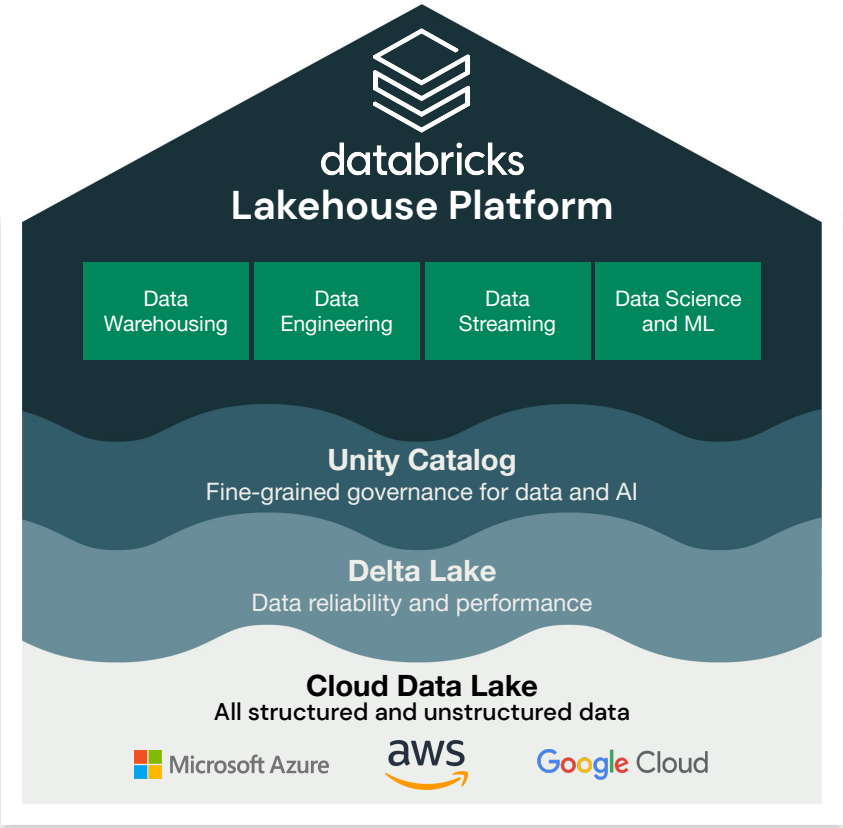
Data Warehouse Team

That data was not warehoused, it's in the data lake

Data Lake Team

We have the data in the data lake, but I can't give you access. How about an Extract?

Databricks Lakehouse Platform



Simple

Unify your data warehousing and AI use cases on a single platform

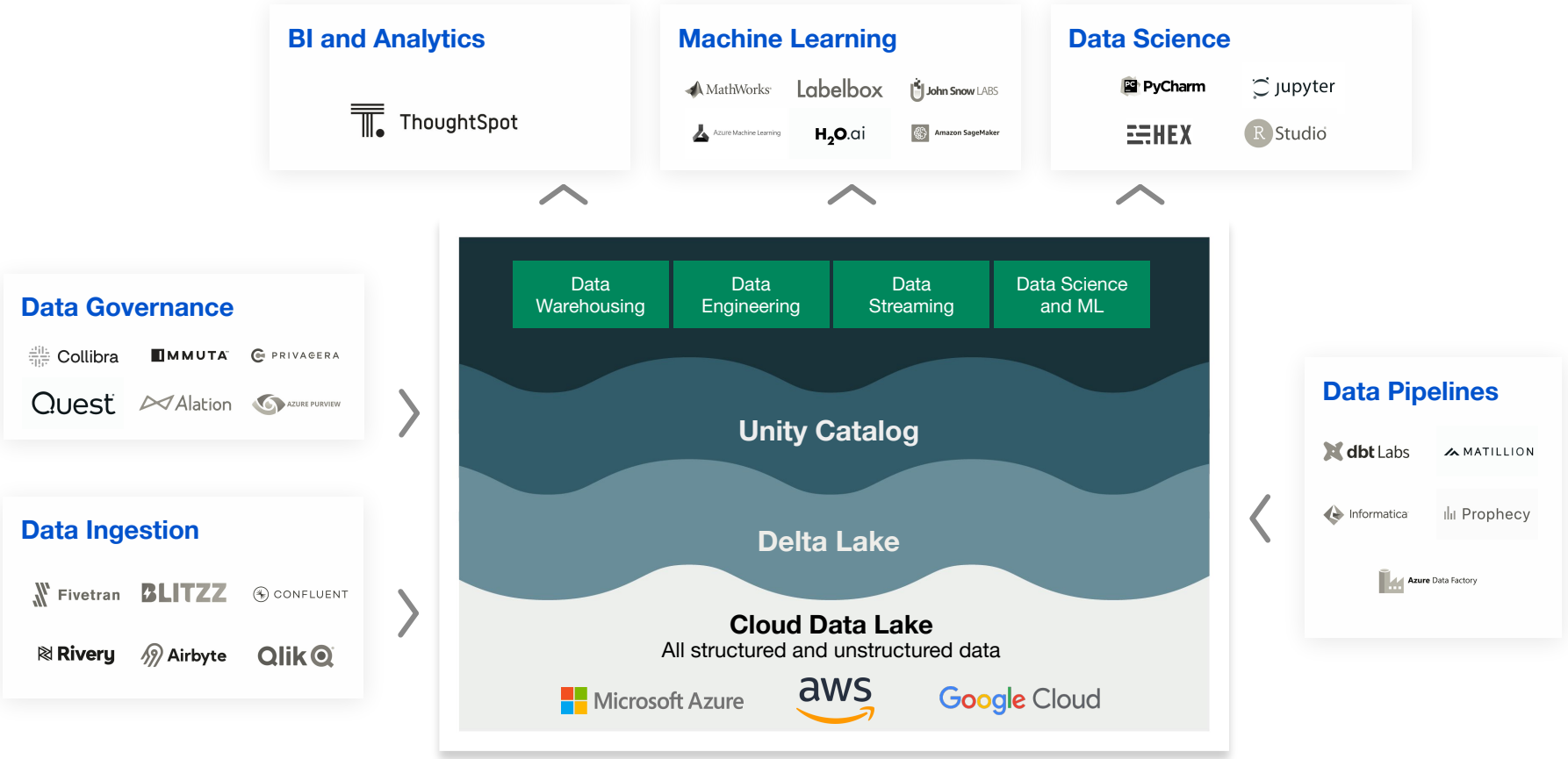
Open

Built on open source and open standards

Multi-cloud

One consistent data platform across clouds

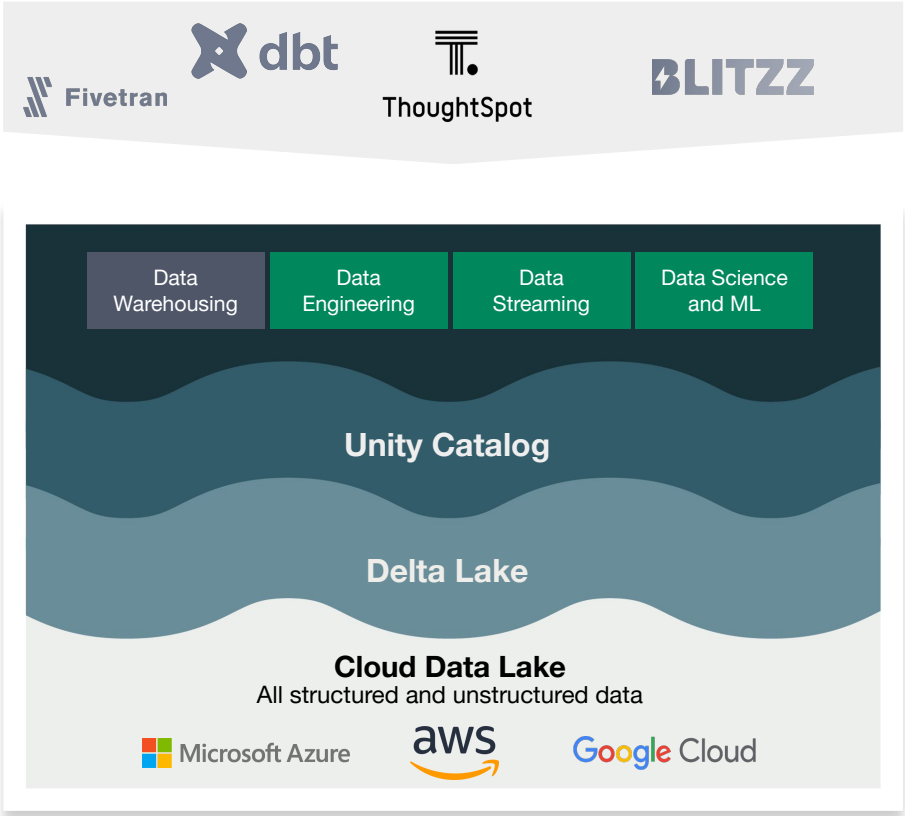
Databricks thrives within your modern data stack



Databricks SQL

Data Warehousing on the Lakehouse

- Analytics on all your data with your tools of choice
- Simplified administration and fine-grained governance
- In-place, lightning-fast analytics on data lake data



Analyst Journey

What can we do with the data?

 **Business Partner**

 **Analyst**

I need data on this new widget we put into production

Here you go, with Lakehouse, we can do analytics on all the data!

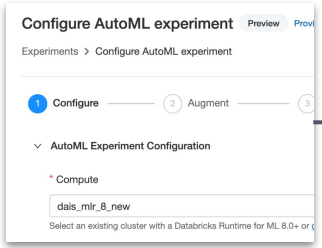
Great! Can you classify these actions for me real quick?

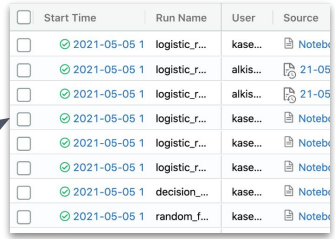
Uh?

Databricks AutoML

A glass-box solution that empowers data teams without taking away control

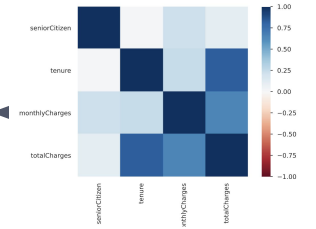
UI and API to start AutoML training





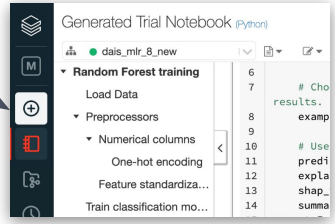
MLflow experiment
Auto-created MLflow Experiment to track models and metrics

Easily deploy to Model Registry



Data exploration notebook
Generated notebook with feature summary statistics and distributions

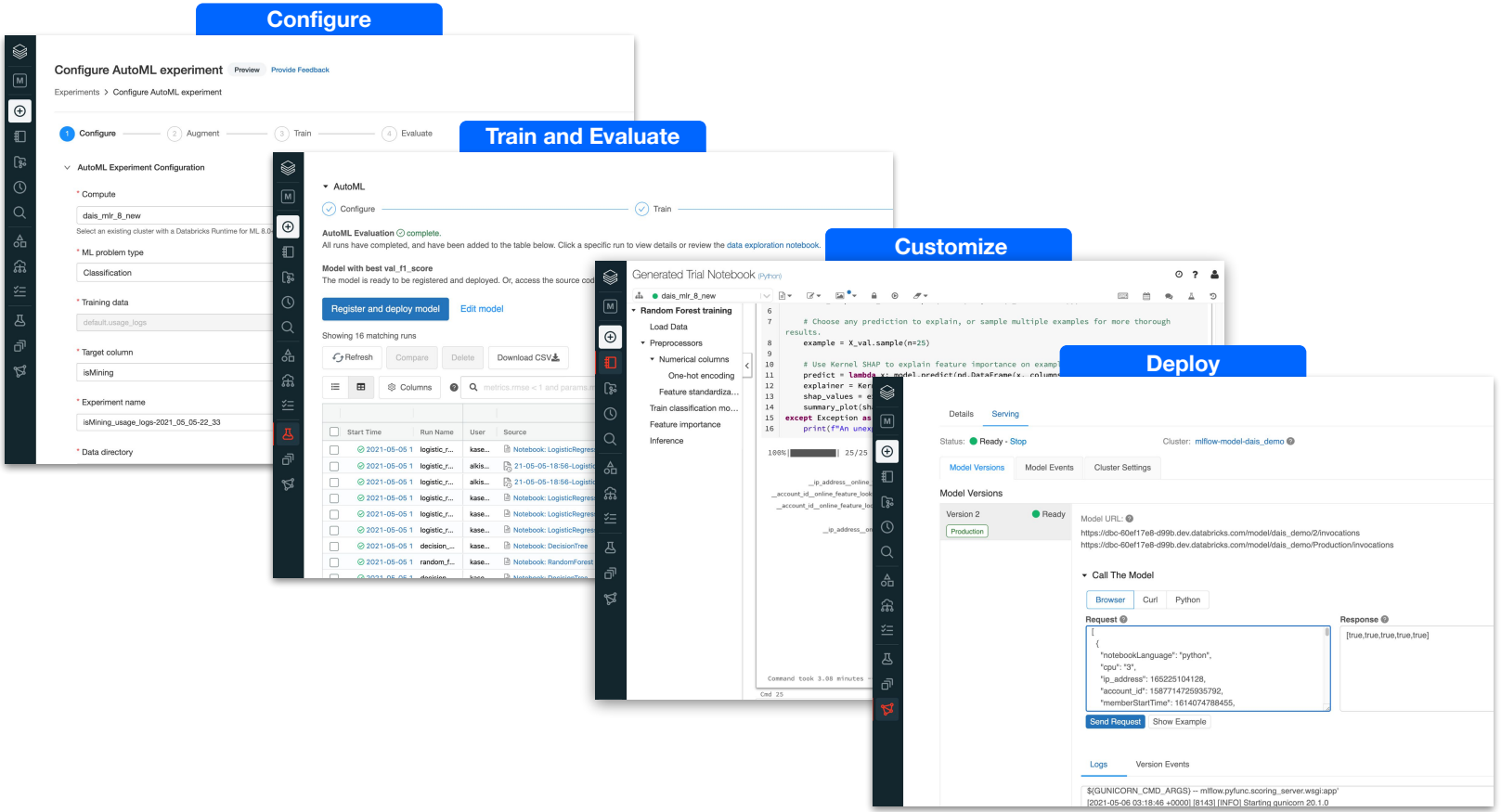
Understand and debug data quality and preprocessing



Reproducible trial notebooks
Generated notebooks with source code for every model

Iterate further on models from AutoML, adding your expertise

“Glass-Box” AutoML with a UI



Databricks AutoML Supported Features

Problem Types	Models / Tuning	Features	Tracking	Evaluation	Deployment
Classification	scikit-learn	Numerical	mlflow	Shap	Batch Scoring
Regression	XGBoost	Categorical	Metrics, Parameters	mlflow	Model Serving
	HYPEROPT	Timestamp	Artifacts, Models		

On Roadmap

Time Series Forecasting	TensorFlow	Text

Analyst Saves the Day

Serving business solutions every day

Business Partner

Analyst

I need data on this new widget we put into production

Here you go, with Lakehouse, we can do analytics on all the data!

Great! Can you classify these actions for me real quick?

Yeah, with Databricks AutoML I can do that with a couple of clicks!

BEYOND.2022

ThoughtSpot

Time to Science

ThoughtSpot for Pharmaceutical Research

Gregory M. Troup PhD

Sr. Principal Scientist
Merck Research Labs
Pharmaceutical Sciences & Clinical Supplies

Merck & Company Inc. Rahway NJ USA

So, what is Merck Pharmaceutical Sciences doing?

Drug Product Development/Clinical Plant Modernization

State of the art product development facilities

- IT systems
- Operational model
- Small scale, portable equipment

What does it feature?

- Designed for flexibility
- Modular Wall Systems/ POD suites
- Product facing staff model, self service oriented
- New mindsets around data.

Current State of Scientific Data: Unstructured & Essentially Single Use

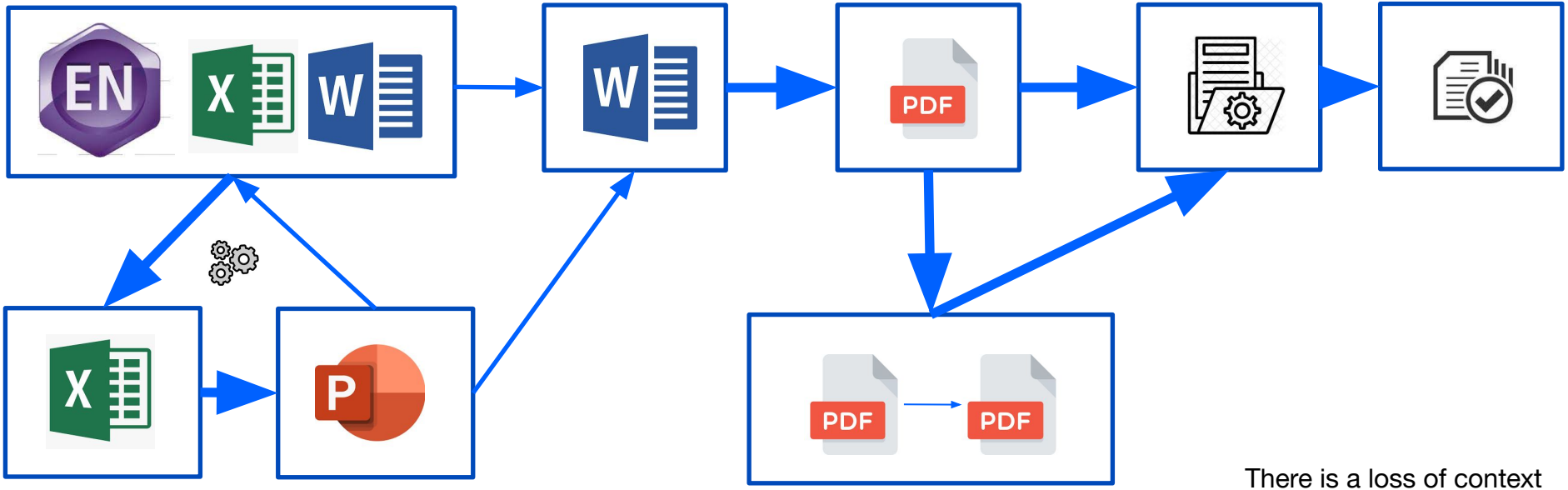
Document experiments & results in ELN, Sign, 2nd Sign, SPR

Reference ELN in a business documents

Convert to PDF, and sign the business document

Store in a doc repository

Approved Product filling



Spreadsheet that the scientists use to use pool data for analysis and plotting

Slide deck that the scientist uses to share results with the team

Business doc used as reference in another business document

There is a loss of context as our traverses the primary use arc

Where we are moving to

Current State

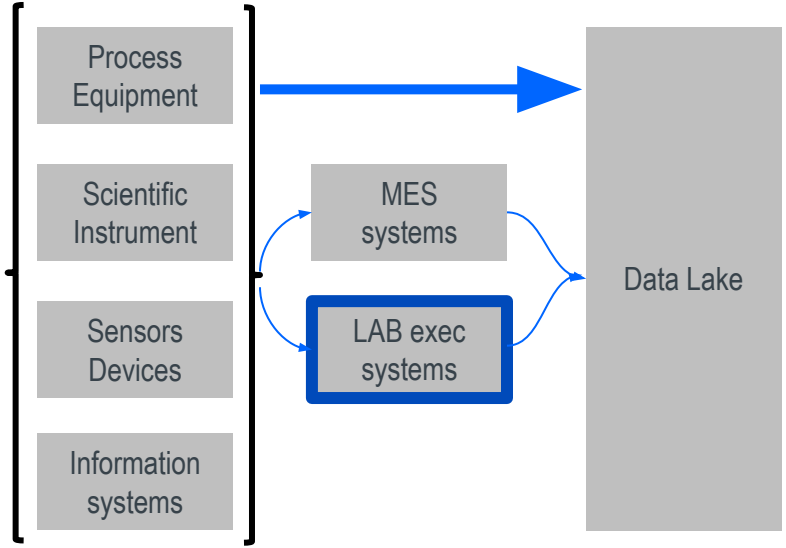
- Assembly of data sets for analysis is a manual process relies on , personal connections, emailing files, extensive file gymnastics
- Loss of context as our data traverses the “single use” arc.
- We have some desktop tools for shaping and analyzing data sets, but they are static
- No way to operationalize analytics, analysis is static

Future State

- Easy, intuitive searching, and assembling data sets for analysis.
- We have full data lineage to accompany our data collections, our analytics, and our models.
- In addition to our desktop tools, data science workbench and cloud-based data science tools for query, shaping, analytics, and modeling.
- Our analytics and models are now dynamic and easily deployed into routine use.

Data Lens | Data Acquisition | Context Acquisition | Store

IT Systems Lens

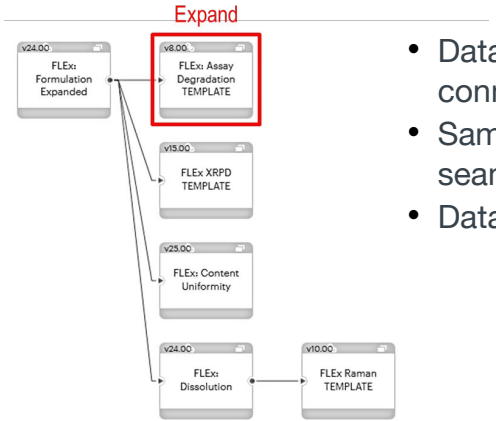


A new approach: Separate documentation of experimental exaction and analysis

Automate the collection of the complete record of the experiment, the who, what, how, where and when.

Make the user experience intuitive, think Panera Bread Kiosk intuitive

1. Internal solution (a combination of technology that we already had in house) to address the gap in the LAB execution space.
2. IoT platform that manages our equipment, instruments and sensors as IoT assets.
3. MES project to bring electronic batch records to development and clinical supplies.



- Data between labs can be connected
- Sample information is seamlessly passed downstream
- Data Acquisition integration

User Lens

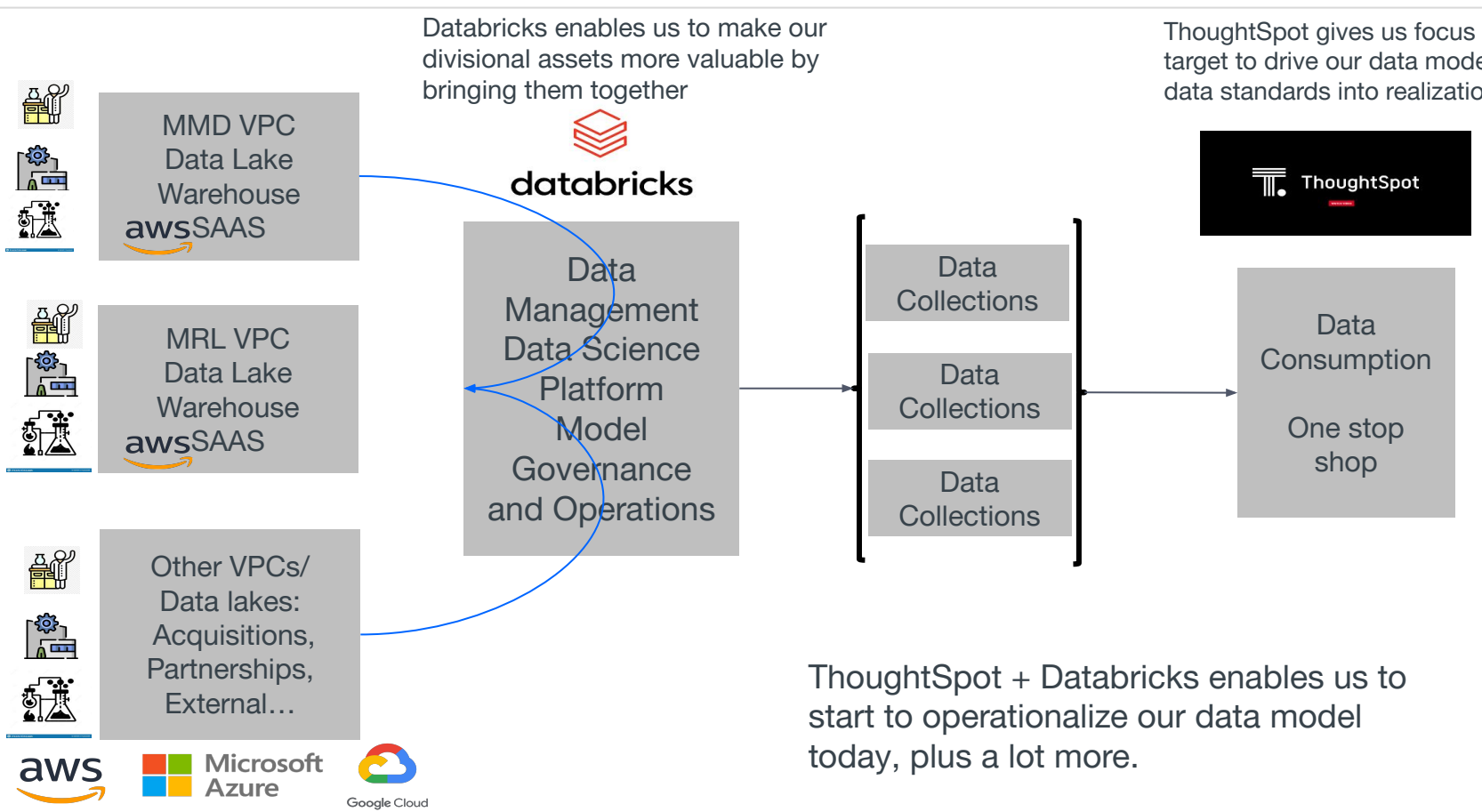
Adaptability to manage the complexity of our field instruments

Flexibility & User Experience to meet the business needs

GMP Store of official records: native and cloud operable, fed with complete data

MERCK INVENTING FOR LIFE

Bringing Together Data Platforms: Accelerating Data Use



Value Perspectives:

“I need real time aggregated data for the process I am currently running in one easy to set up view that I can see from anywhere in the room.”

“I can easily set up any combination of room, equipment, data that I need to displayed for run time viewing on our devices.”

“ I have easy access to historical data, so makes it easier for me to plan and execute new work.”

“ I can instantly see the results generated by my product team members; we now can focus our energy on planning the next steps.”

ThoughtSpot

“I can quickly configure a data visualization that shows all of the relevant data for my equipment/process subject matter expert group.”

“Tech transfer risks are reduced because we can now quickly find development data related to the product we are working on and similar products.”

“As a data scientist, I now have easy access to training and validation data, and I have a fast path to deployment.”

“Our ML model outputs are now easily accessible right along with our experimental data.”

A New Mindset is needed

Everyone is an Analyst!

- Using the best tools for the job, breaking the monolithic system trap
- Real time scientific inquiry, the end of slide decks as primary tool to sharing results.
- Ask and answer questions on the fly
- The second use, easy to find data sets for cross “fill in the blank” analysis (blank = program/product/material/equipment/site/…)
- Creating an environment where our scientists can learn faster and execute efficiently. More knowledge = less risk.