DEV DAY

ML Dev Day Workshop

Unifying Data Science and Data Engineering

WELCOME

- We will begin at 9:05am
- This event will be recorded
- You will receive an email with the Deck and recording tomorrow.
- Please feel free to use the Q&A tool for any questions
- Don’t forget to participate in our polls!

Introductions

- Cameron Kashani, cameron.kashani@databricks.com
  - Account Executive
- Lei Pan- lei.pan@databricks.com
  - Solutions Architect
- Aaron Binns - aaron.binns@databricks.com
  - Solutions Architect
- Don Warnes - don.warnes@databricks.com
  - AWS Alliance Manager

Additional Databricks Attendees: Raise your hand so you may introduce yourselves!

Introductions

- Brad Ito, Retina
  - CTO & Co-Founder, brad@retina.ai
- Mo Messidi, Retina
  - DataOps Engineer
- Jonathan Corners - Pariveda
  - jonathan.corners@parivedasolutions.com
- Brian Edwards - Pariveda
  - brian.edwards@parivedasolutions.com

Poll

Agenda

9:00am Databricks Keynote, Cameron Kashani
9:45am Customer Presentation | Retina
10:15am Partner Presentation | Pariveda
10:30am Break
10:40am Data Engineering Interactive Demo & Best Practices: Preparing Data for Analytics | Aaron Binns
11:15am Data Science Interactive Demo & Best Practices: Model Training and Machine Learning | Lei Pan
11:50am Q&A
12:00pm Finish

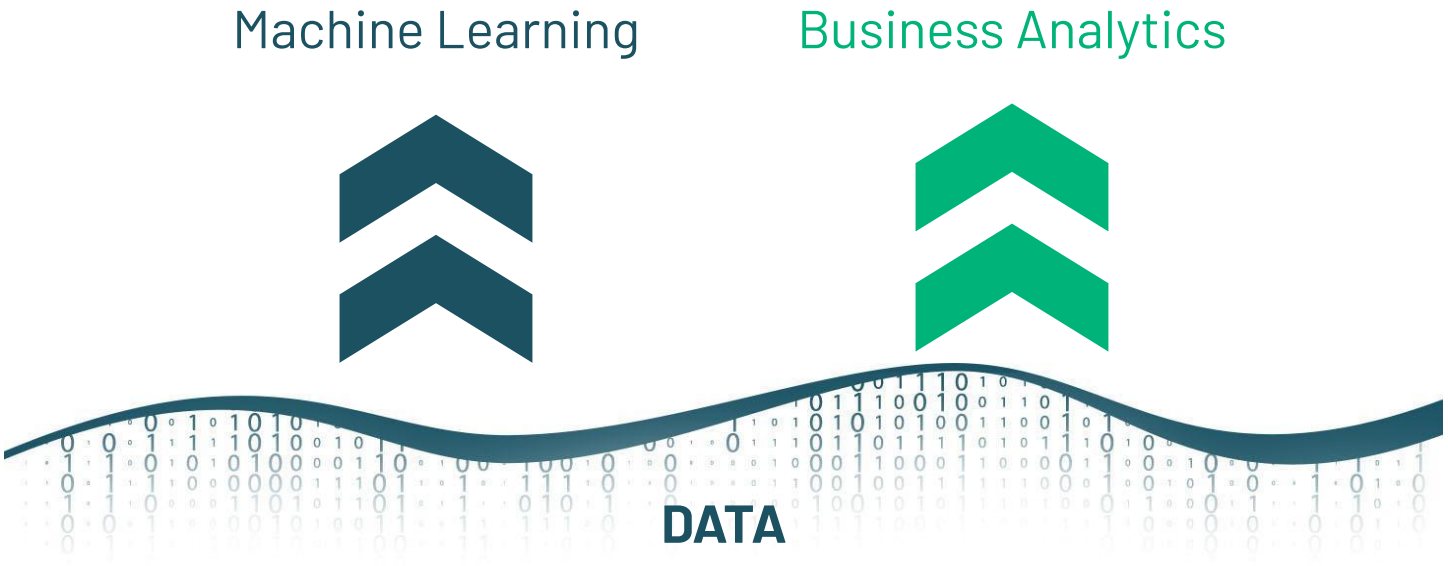
Unified Data Analytics Platform

Accelerating innovation across data science, data engineering, and business analytics

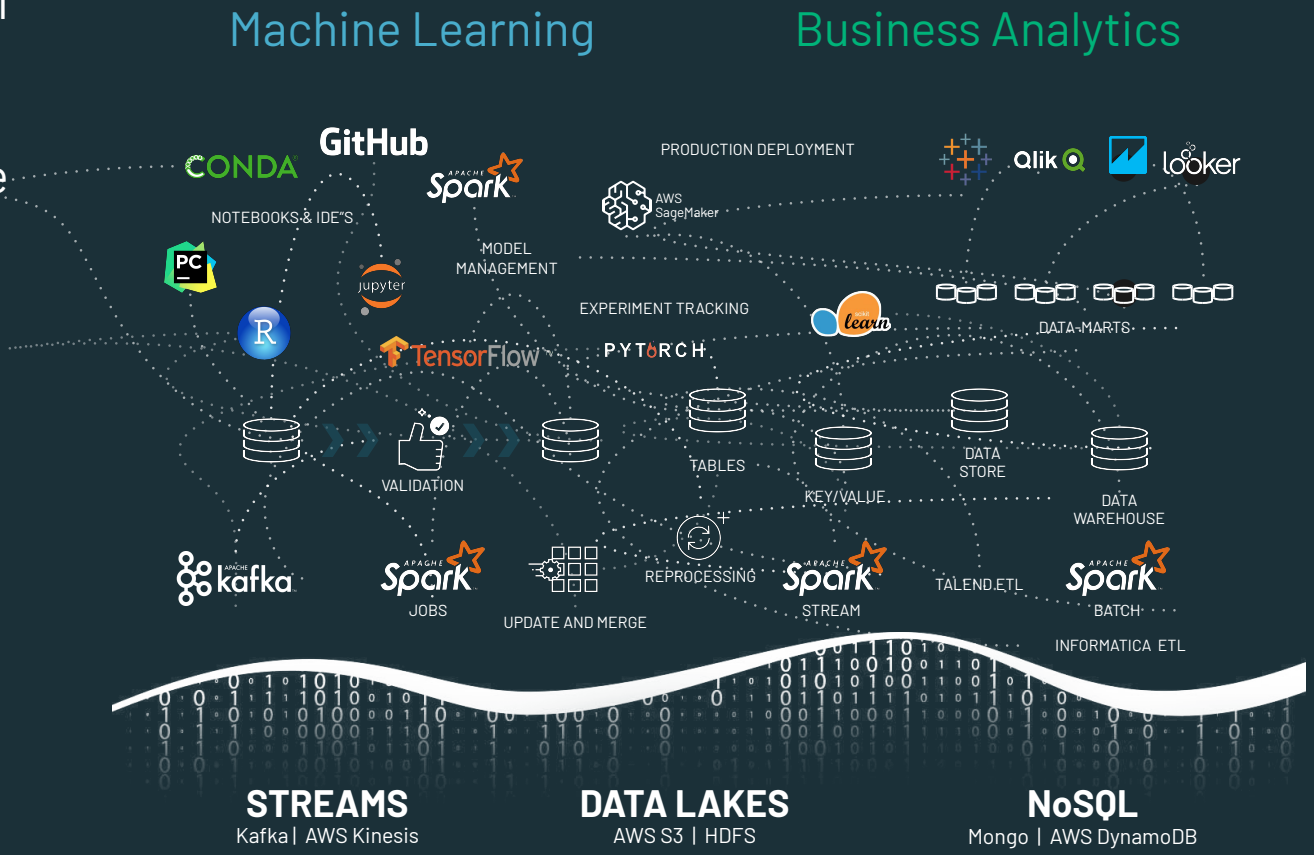
- Global company with 5,000 customers and 450+ partners
- Original creators of popular data and machine learning open source projects
- Empowering data teams to solve the world's toughest problems

Unlocking business value

Companies are combining massive data with ML and BI capabilities



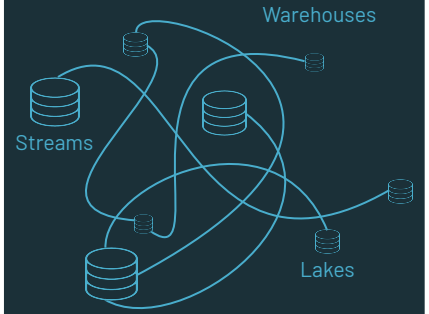
Most organizations fail to unlock business value due to data, technology and people silos



Unlocking business value: Four challenges

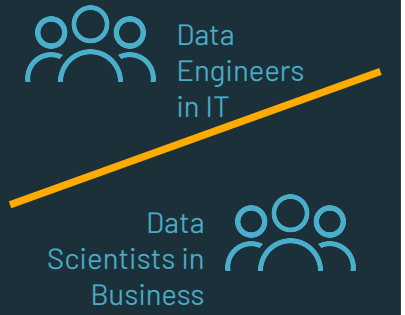
1

Data is messy, siloed and slow



2

ML is hard, Production is harder



3

BI is limited to a fraction of data

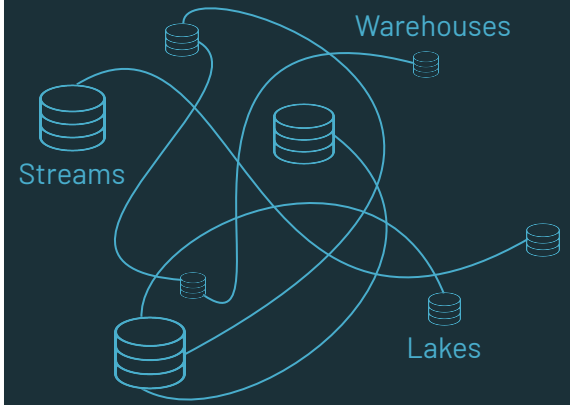
4

Lack of enterprise readiness

Make all your data ready for BI and ML

1

Data is messy, siloed and slow



Unified Data Service

Build open, reliable, fast data lakes with all your data

Big Data

Business Data

Applications

DELTA LAKE™

Open High Quality Fast

hadoop

Amazon S3

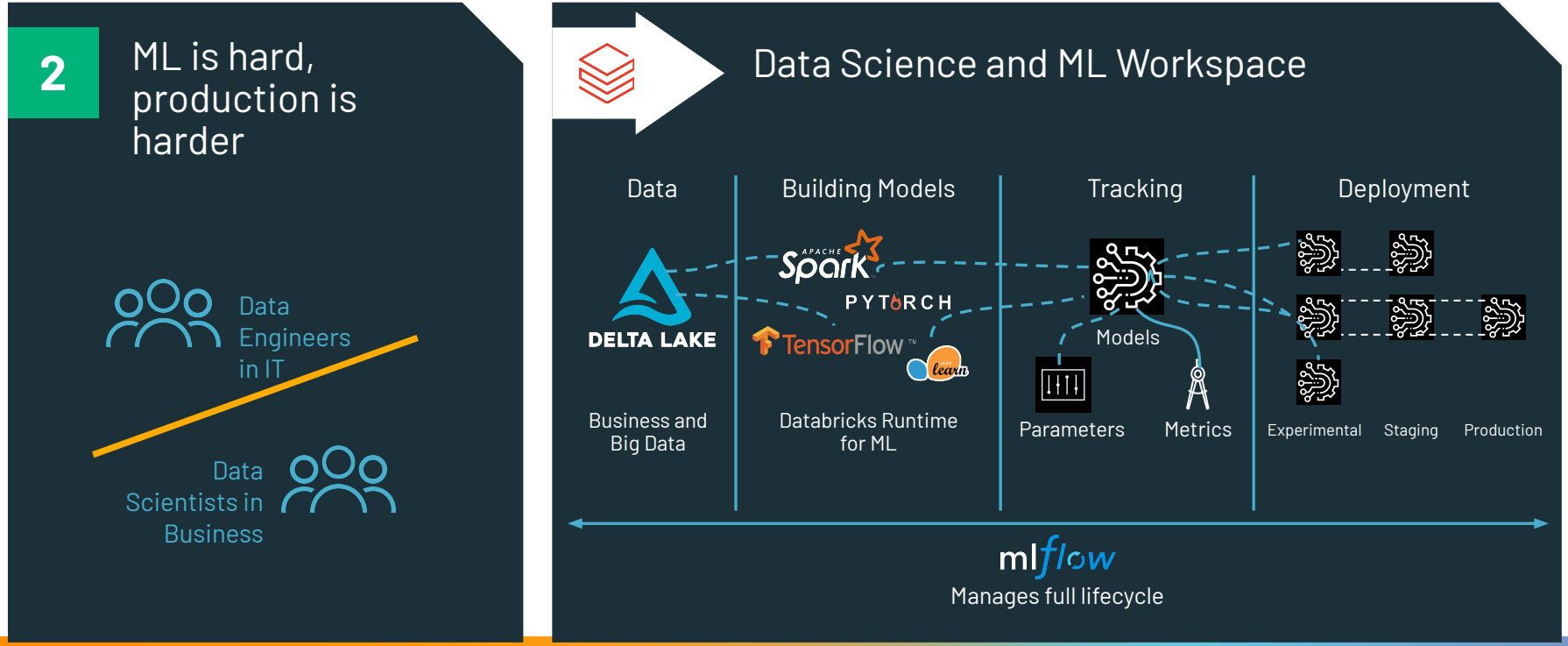
APACHE Spark

Unified Engine

BI Reporting

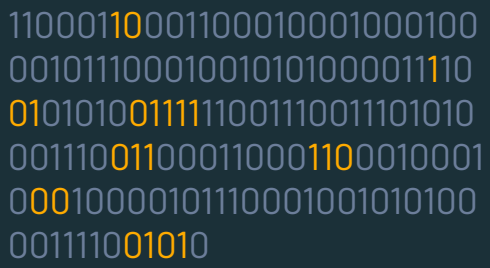
Machine Learning

Unify data and ML across the full lifecycle



Enable BI directly on all your source data

3 BI is limited to a fraction of data



BI Integrations for data lake

Reports | Dashboards | Ad hoc data science

DELTA LAKE

APACHE Spark

All your data with high quality and great performance

Applications: salesforce, Marketo, zendesk, Google Analytics

Data stores: mongoDB, ORACLE, MySQL, redis, kafka, cassandra

Files

Leverage cloud native platform for enterprise grade solution

4 Lack of enterprise readiness

- Fragmented security
- Poor reliability
- Disjointed governance

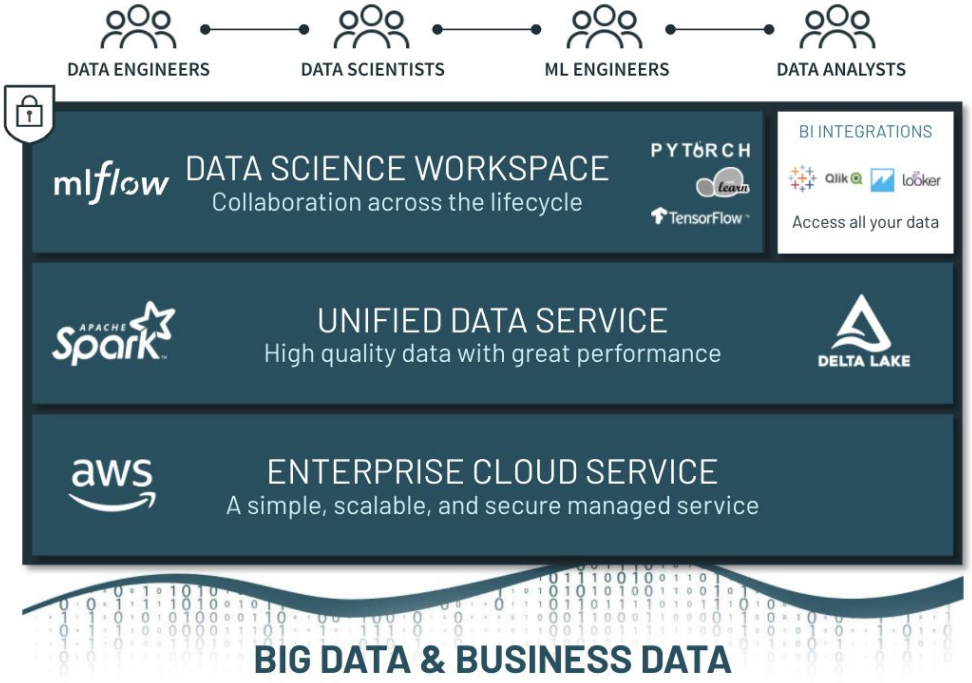
Databricks Enterprise Cloud Service

Databricks unified data analytics platform

Data Science, ML, and BI on one cloud platform

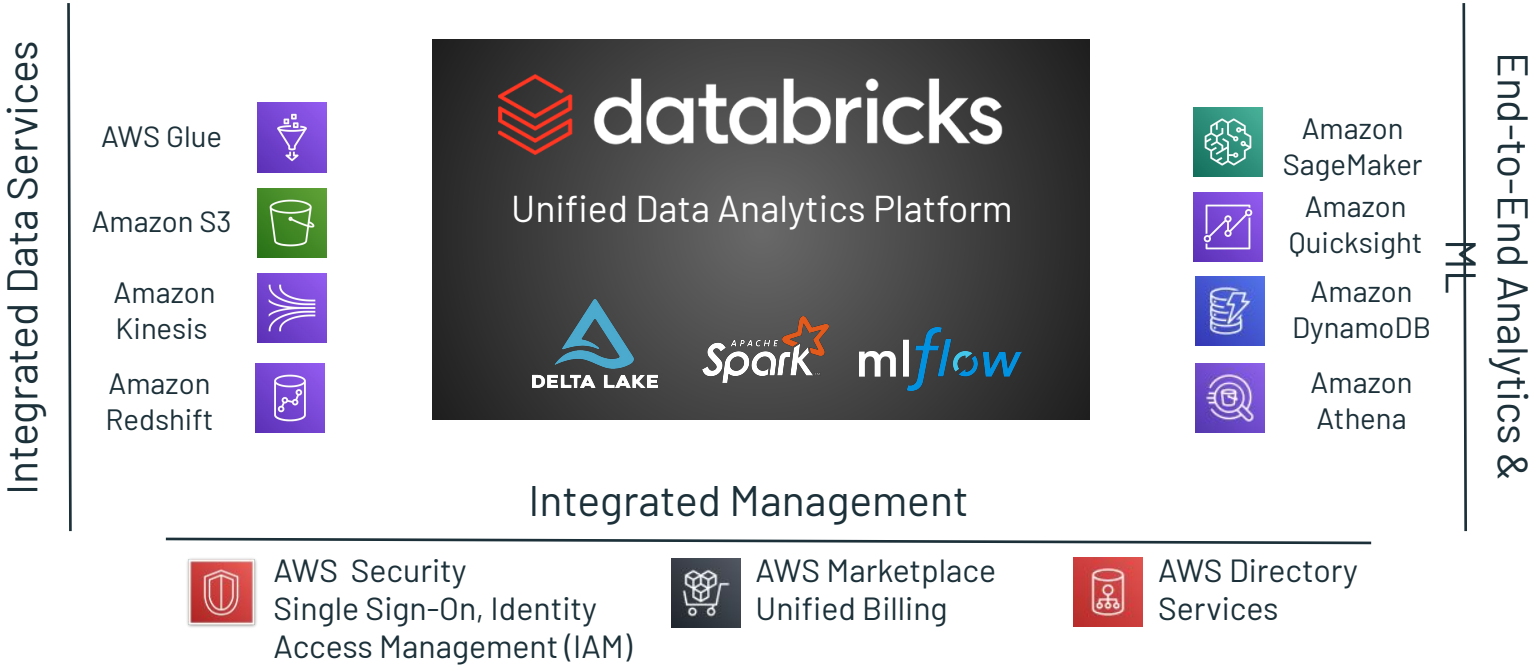
Access all business and big data in **open data lake**

Securely integrates with your **cloud ecosystem**



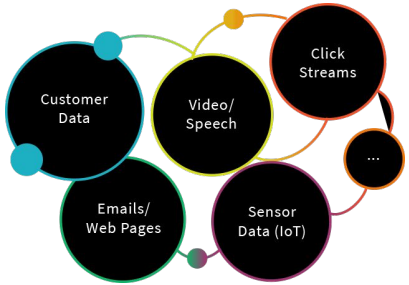
Databricks on AWS

Integrated with your AWS infrastructure



Our Business Value at- a-glance

Data engineering, management and transformation



LAKES

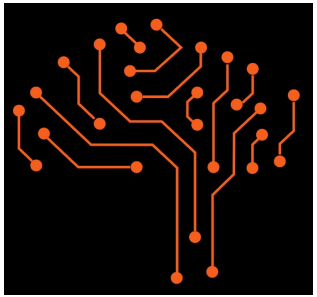
STREAMS

kafka

WAREHOUSES

NOSQL

Data Science, Machine Learning and Deep Learning



Studio

learn

XGBoost

mxnet

TensorFlow

Keras

PYTORCH

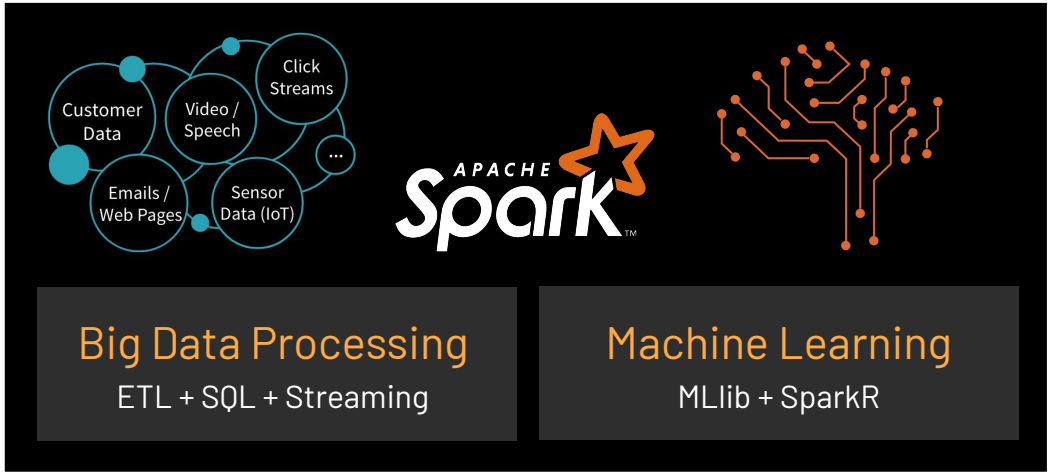
Data Science and Machine Learning Business Outcomes

- Recommendation & Personalization Engines
- Risk, Fraud, Intrusion Detection & Prevention
- Customer 360 Engagements , Ad Targeting
- Inventory Asset Optimization & Allocation
- Genomics & DNA Sequencing
- Predictive Maintenance, SCM Seasonal Costing
- Sentiment & Customer Churn Analysis
- Security Compliance & Intelligence

One Unified Data Analytics Platform // Guaranteed Expertise // Predictable Business Outcomes

Apache Spark: De-Facto Unified Data Analytics Engine

Uniquely combines Data & AI technologies




facebook

LinkedIn

hp

NETFLIX

Nasdaq

Tencent 腾讯

amazon.com

Bai du 百度

Bloomberg

Uber

mastercard

VISA

U.S. Immigration
and Customs
Enforcement

SIEMENS

NBC

viacom



THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

CERN


APACHE
Spark™
500,000
meetup members

In spite of Spark success, companies are still struggling with ML

Hardest Part of AI isn't AI, it's Data

"Hidden Technical Debt in Machine Learning Systems," Google NIPS 2015

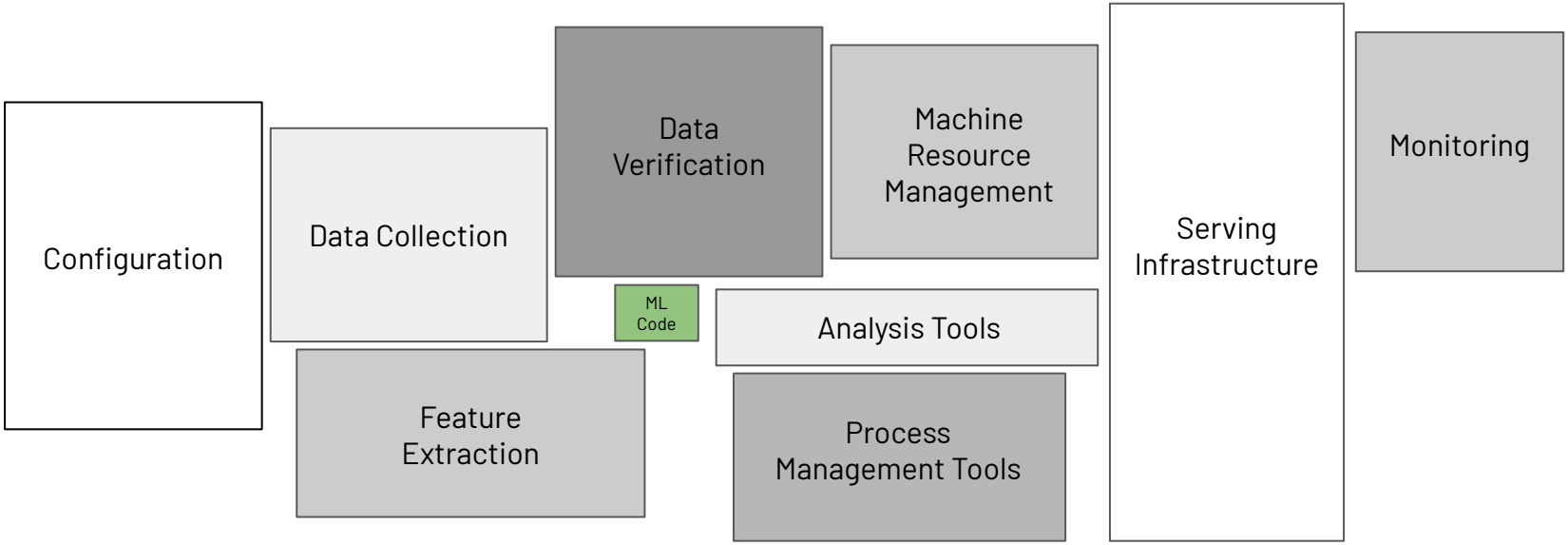
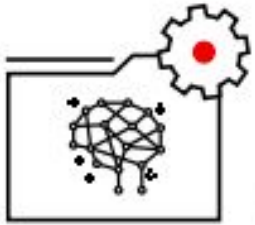


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

CIO Survey: AI Promise

CIO Survey: AI Dilemma



Just **17%**

have moved an AI project to production in a core business area

It takes an average of

6 months

to bring AI projects to deployment

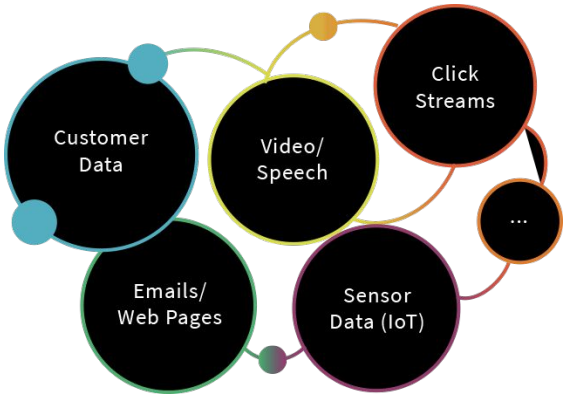
3 major challenges slowing ML Projects:

1. Data is not ready for AI
2. A Zoo of new ML Frameworks
3. Data Science & Engineering silos

Challenge 1

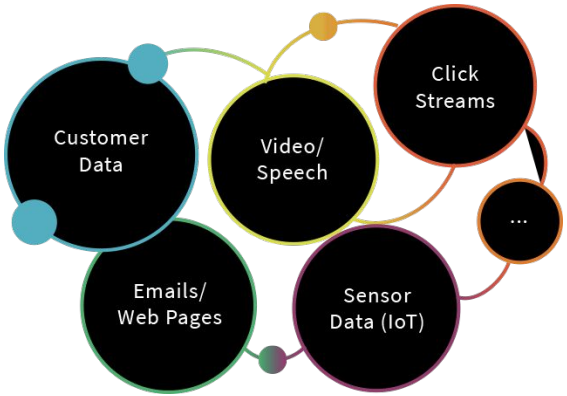
Data is not ready for AI

Enterprises have been spending millions of dollars getting data into data lakes



Data Lake

The aspiration is to do analytics & AI on all that data!



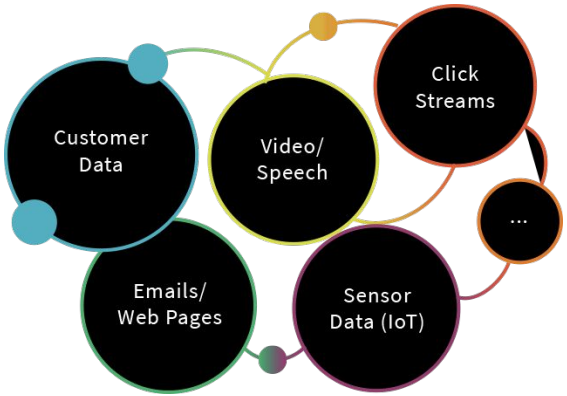
Data Lake

Analytics & AI

- **Recommendation Engines**
- **Risk, Fraud Detection**
- **IoT & Predictive Maintenance**
- **Genomics & DNA Sequencing**

But the data is not ready for analytics & AI

The **majority** of these projects are failing due to **unreliable data!**



Data Lake

Analytics & AI

- Recommendation Engines
- Risk, Fraud Detection
- IoT & Predictive Maintenance
- Genomics & DNA Sequencing

Delta Lake: makes data ready for analytics & AI

Reliability

Performance

Analytics & AI

- Recommendation Engines
- Risk, Fraud Detection
- IoT & Predictive Maintenance
- Genomics & DNA Sequencing

A New Standard for Building Data Lakes

Open Source and Open Format

Data Reliability and Quality

Compatible with Spark APIs

Delta Lake: Analytics Ready Data

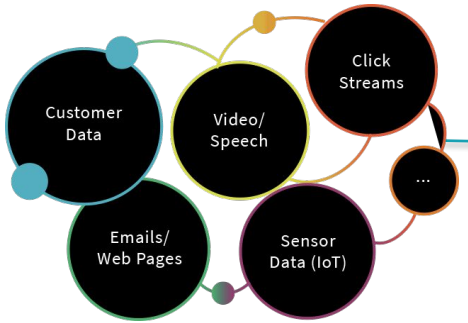
1. Data Reliability

Guarantee High Data Quality
Schema Enforcement & ACID Transactions

2. Query Performance

Very Fast at Scale
Unified Batch and Streaming
Indexing & Caching (10-100x Faster)

LOTS OF NEW DATA



DELTA LAKE

DATA LAKE

- Reporting
- Dashboards
- Alerting
- Machine Learning

Challenge 2

A Zoo of new ML Frameworks

Complexity - Zoo of ML frameworks

Machine Learning

Scikit-learn ,Spark MLlib, H2O, Mlpack, Mahout ...

Deep Learning

TensorFlow, Keras, Caffe, PyTorch, Theano, BigDL, SparkDL ...

Supporting Libraries

Python, R, Anaconda, Numpy, Scipy, Pandas, Matplotlib, PyViz ...

Serving and Monitoring

MLeap, TF Serving, Cassandra, Redis, TensorBoard ...

Databricks Runtime for ML

Ready to use clusters with built-in ML Frameworks

GPU support

Challenge 3

Data Scientists & Engineers are in silos

Data Scientists & Engineers are in Silos

1 Data Prep

Hard to make pipelines reliable

Apache Spark · Delta Lake

2 Build Model

Challenging to track and reproduce experiments

GitHub · Conda · TensorFlow · Excel · scikit-learn · XGBoost

3 Deploy Model

Have to ensure reliability, SLAs, and quality

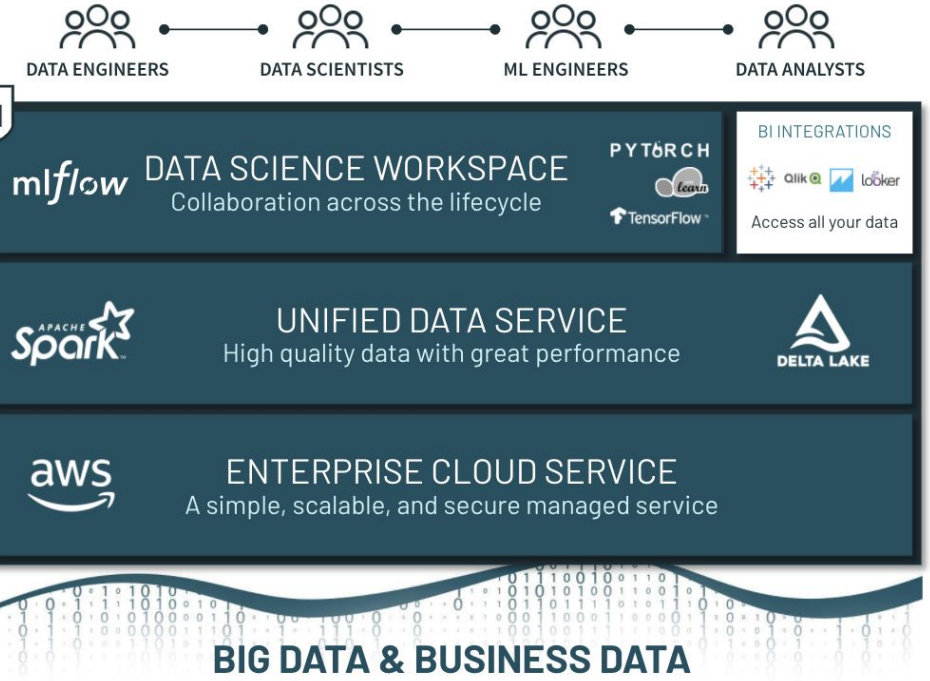
Amazon SageMaker · kubernetes · docker

Data Engineers

Data Scientists

Databricks Unified Data Analytics Platform

End-to-end ML platform that unifies people, processes and technologies



Collaborative workspace brings data scientists and data engineers together

3

Delta Lake Makes Data Lakes Ready for Data Science and ML

1

MLflow standardizes ML Lifecycle with experimentation, reproducibility & deployment

3

Runtime for ML provides ready to use clusters with built-in ML frameworks

2

Customer Presentation

RETINA

April 22, 2020

Data Platform Checklist

Brad Ito
CTO, Retina AI

RETINA

Introduction

- 4 startups as CTO
  - Email Marketing Platform
  - Performance Marketing
  - Crowdfunding Platform
  - Retina AI
- 10 Programming Languages
- Numerous data platforms
- Physics, MIT

RETINA

- Customer Lifetime Value experts
  - Early & accurate models
- Customer Intelligence Partner
- Clients include: Capital One, Chegg, Madison Reed, Dollar Shave Club

The Problem

Business Data Pain

- Unable to use data to drive business change
- Data more chore than delight
- Untrusted data
- Expensive and complex systems
    - Being "Big Data" is not enough

RETINA

Solution:
Data Platform Checklist

- See how you measure up with simple yes / no
- Broadly applicable
- Move from buzzwords to value
- Provide a "North Star" for data initiatives

RETINA

1. Is your data in one place?

- Accessible via a single platform to ease data discovery
- Single source of truth - trusted data insights
- Can be queried
  - quickly - at the speed of a question
  - without adverse effects (not OLTP, scaled resources)

2. Is your data self-service?

- Technical (Python / R / SQL) users can find / access data
  - can easily share transformed data
- Analysts can find / access data via BI Dashboard
- Applies: DataOps >> Data Engineering Requests
- Meta data available
  - Business context
  - Data lineage

3. Is your data checked for validity?

- Manual checks with business context
- Automated checks
  - Developed from real data quality problems
  - Manual intervention
- Data ownership

4. Is your data resilient?

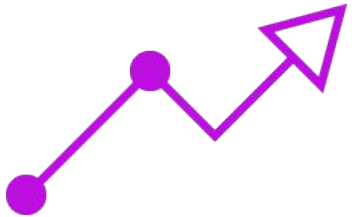
- Can you correct past data?
- Can you easily make small changes and undo them?
- Are your data transformations idempotent?
  - Event logging over state tracking
  - Universally unique ids
- Are data relationships explicit?
  - example: customer identity resolution

5. Are you optimized for cost / performance?

- Separate cloud compute from storage
- Schemas with appropriate data types
- Data structures
  - In-Memory
  - Columnar
- Data Storage
  - Compressed
  - De-duplicated

6. Is your data at the speed that you need it?

- Understand your data latency requirements
- Use streaming technologies for streaming data
- Schedule batch jobs appropriately
  - At least as fast as data sources
  - No faster than ability to consume data



7. Is your data platform connectable?

- Easily ingest from new data sources
  - example: Databricks partner integrations
- Easily use data with external systems
  - Queries -> Marketing Automation
  - Dashboards and Reports
  - Data Science and Machine Learning partners

Data Platform Checklist

1. Is your data in one place?
2. Is your data self-service?
3. Is your data checked for validity?
4. Is your data resilient?
5. Are you optimized for cost / performance?
6. Is your data available at the speed that you need it?
7. Is your data platform connectable?

Brad Ito
CTO, Retina AI
brad@retina.ai
www.retina.ai - retina.ai/blog

https://www.linkedin.com/in/brad-ito/

Partner Case Study

PARIVEDA

Modern Data Enterprise

Presented By:

Jonathan Corners and Brian Edwards

Hear Our Perspectives - https://parivedaperspectives.com/

PARIVEDA

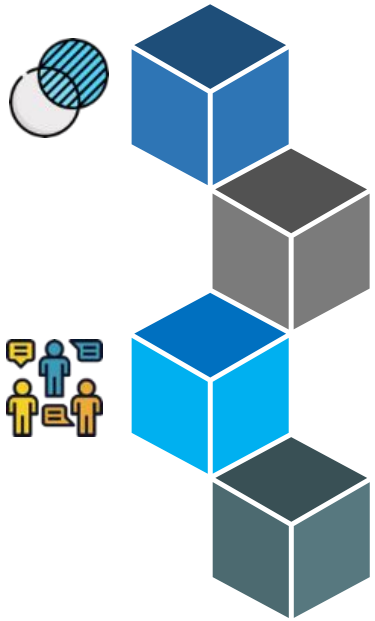
Pariveda Solutions Guiding Principles

TRANSPARENCY IS KEY TO A GOOD RELATIONSHIP

Be frank, give options, and enable our customers to do what makes sense in their context.

CO-CREATE VALUE WITH OUR CUSTOMERS

Giving and sharing is the best way to uncover potential.



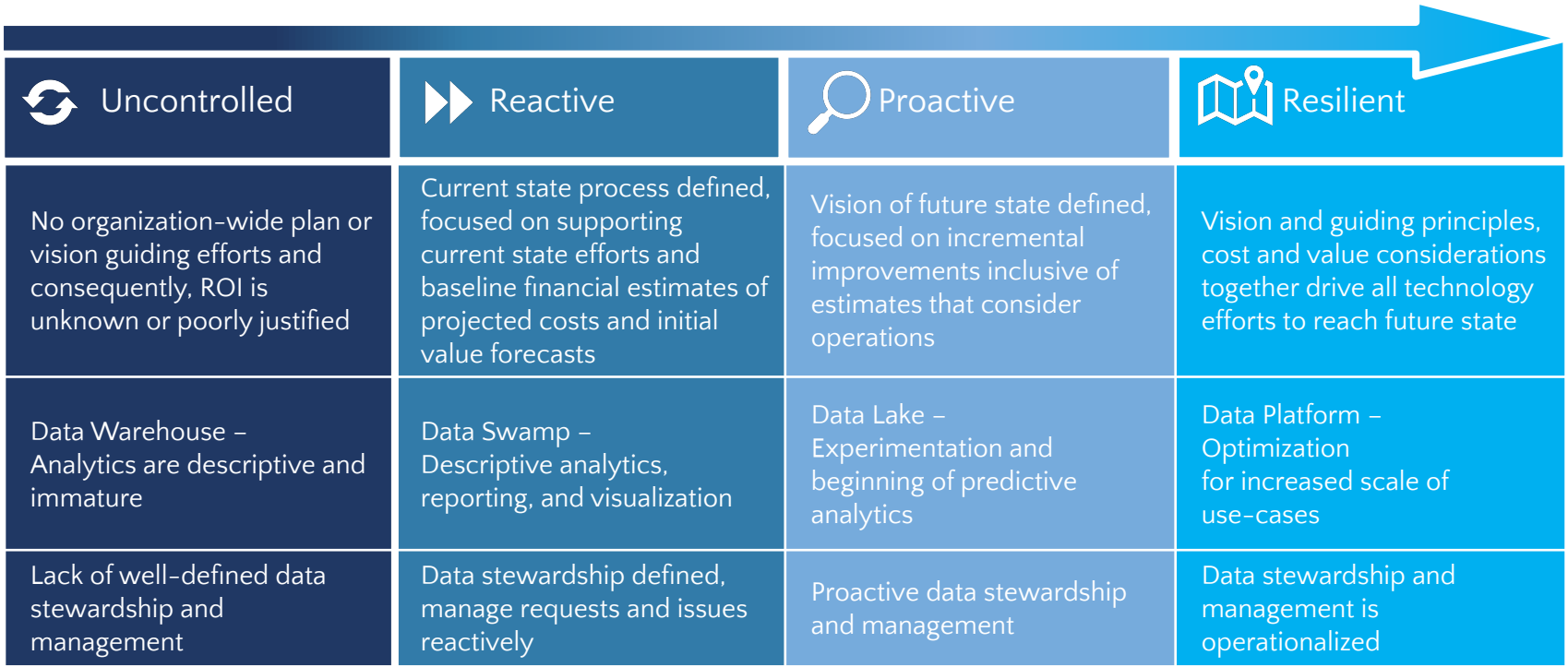
TECHNOLOGY IS AN ENABLER, NOT A SOLUTION

Apply the human picture when observing with a technical lens.

MAKE A DIFFERENCE

Take pride in the process and the product, but most of all strive for positive impact.

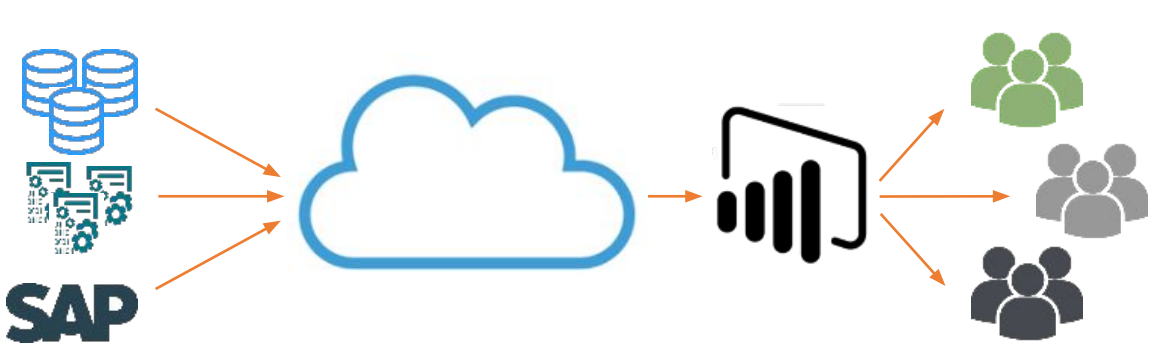
Pariveda's Modern Data Enterprise Maturity Model



The modern data enterprise is proactive and resilient with data driven insight!

PARIVEDA

Case Study: Data Driven Insights for Marketer of Consumer and Commercial Products



Situation

A Fortune 500 company that sells consumer packaged products is looking to accelerate insights by providing analysts with increased interaction and high levels of granularity on their datasets.

Data quality and data scale challenges made it difficult for the business to accurately forecast future sales and inventory needs.

The current approach was high cost for IT to maintain and difficult to integrate new data sources. A new scalable approach was required.

Solution

We worked with the client to developed a scalable, cloud based Digital Data Platform.

Using the cloud, we implemented an ingest, model, enhance, transform, deliver approach to processing large datasets and delivering data at its more granular level to 10,000 users. Automation of repeatable manual processes enabled the team to reallocate operational effort to building components that improved data insights.

As the platform matured, Databricks was introduced to enable advanced analytic scenarios at scale.

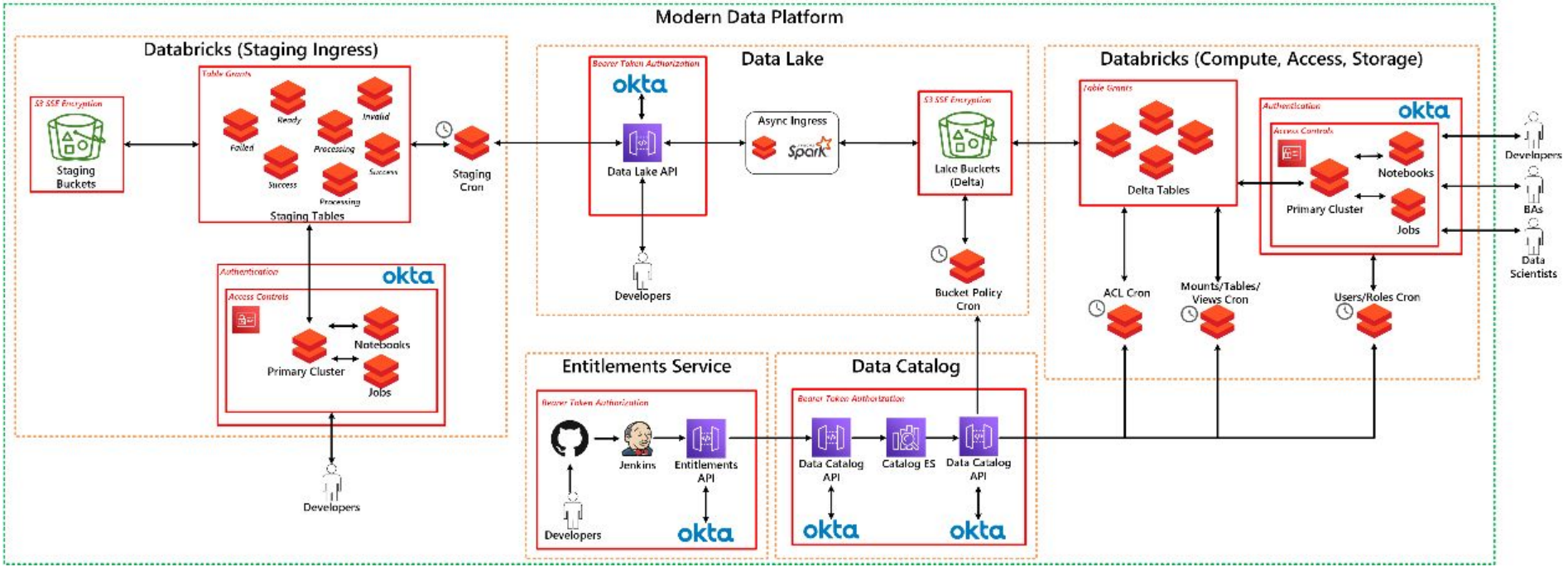
Value Provided

“Speed to insight” enabled executive decision makers to respond faster and more accurately to market changes.

Reduced IT effort on data quality issues by 25%. Reduced implementation and operational run costs by leveraging a patterned based approach for adding and processing datasets.

Data Science teams could identify and create new reports and models to make accurate predictions. Same day updates are now possible, allowing the ability to report on information across the entire enterprise.

Databricks simplifies and upskills multiple areas of the Modern Data Platform



DataForge - Component-Driven Data Platform

Ingest: Capture the data from the source (especially for ephemeral sources)

Model: Handle differences in input format for common data sets

Enhance: Add additional context, make each data record more usable

Transform: Combine and aggregate data sets to produce new data products that are ready for consumption

Deliver: Retrieve and send data sets / records to support data based applications and analytics

How Pariveda Can Help

How do I convince my boss of the value of modern data tools & practices?

Where should we start? What use cases have the best bang for the buck?

How do I turn my data swamp into a data lake

AI/ML Application: Participants & Outcomes

Participants: Product Leads, Business Unit Heads; CTOs & CDOs, BI/Analytics Leads

	Ideation	ML Proof of Concept	Production Integration
		Includes 'Ideation' plus an implementation of:	*Includes model from 'PoC' integrated into systems and processes:*
What Will You Learn?	What high-value decisions does your business make that ML-based predictions can help with? What data sources are available to support models? Can AI techniques encode unstructured data sources to enable new applications? How complex will the ML model development be for each idea? *Prioritized concept cards capturing high-value systems that can be enabled by Machine Learning*	What is the predictive accuracy of the ML model in for aiding in high value decisions? How can the model, running on the AWS Platform, be executed to make manual predictions? How will the model's predictions be integrated into business processes? *ML model accuracy report alongside plan and estimates for production application build*	How do we build/maintain/operate APIs & data pipelines for ML apps? How will business processes change to leverage predictions that the ML POC has made easily available? How can I measure the business value generated by the new predictive capability? *Realized business value from POC alongside plan to build Data Lake Platform and enable ML @ Scale*
Pricing	Free 1-2 Day Workshop[1]	Price based on complexity of POC	Price based on POC outcomes

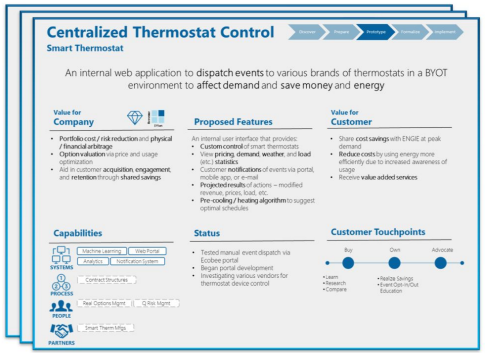
1. Best results achieved when workshop delivered over 2 days. Day one focused on ideation, day 2 on prioritization, with complexity analysis done offline between the days

Concept Cards

Pariveda's Concept Card minimizes the barriers and positions your organization to have meaningful conversations around new business ideas

CONTINUOUSLY COMMUNICATES STATUS AND PROGRESS

ENSURES GOVERNANCE IS ALIGNED THROUGHOUT



ALIGNS EVERYONE ON THE VALUE

INSTILLS QUICK AND ITERATIVE PROTOTYPING

PARIVEDA

For more information: jonathan.corners@parivedasolutions.com

Questions

Brian.Edwards@parivedasolutions.com
Jonathan.Corners@parivedasolutions.com

PARIVEDA


SPARK+AI SUMMIT
JUNE 22-26, 2020 | ORGANIZED BY databricks
THE VIRTUAL EVENT FOR DATA TEAMS
• Extended to 5 days with over 200 sessions
• 4x the pre-conference training
• Keynotes by visionaries and thought leaders
NOW FREE

Migrating Hadoop to Databricks Webinar

May 21, 2020 | 10am PT

https://bit.ly/HadoopToDB

Agenda

9:00am Databricks Keynote, Cameron Kashani
9:45am Customer Presentation | Retina
10:15am Partner Presentation | Pariveda
10:30am Break
10:40am Data Engineering Interactive Demo & Best Practices: Preparing Data for Analytics | Aaron Binns
11:15am Data Science Interactive Demo & Best Practices: Model Training and Machine Learning | Lei Pan
11:50am Q&A
12:00pm Finish

Break

To follow-along with our technical demos after the break, please have your Databricks Community Edition account created. No credit card needed.

https://databricks.com/signup/signup-community

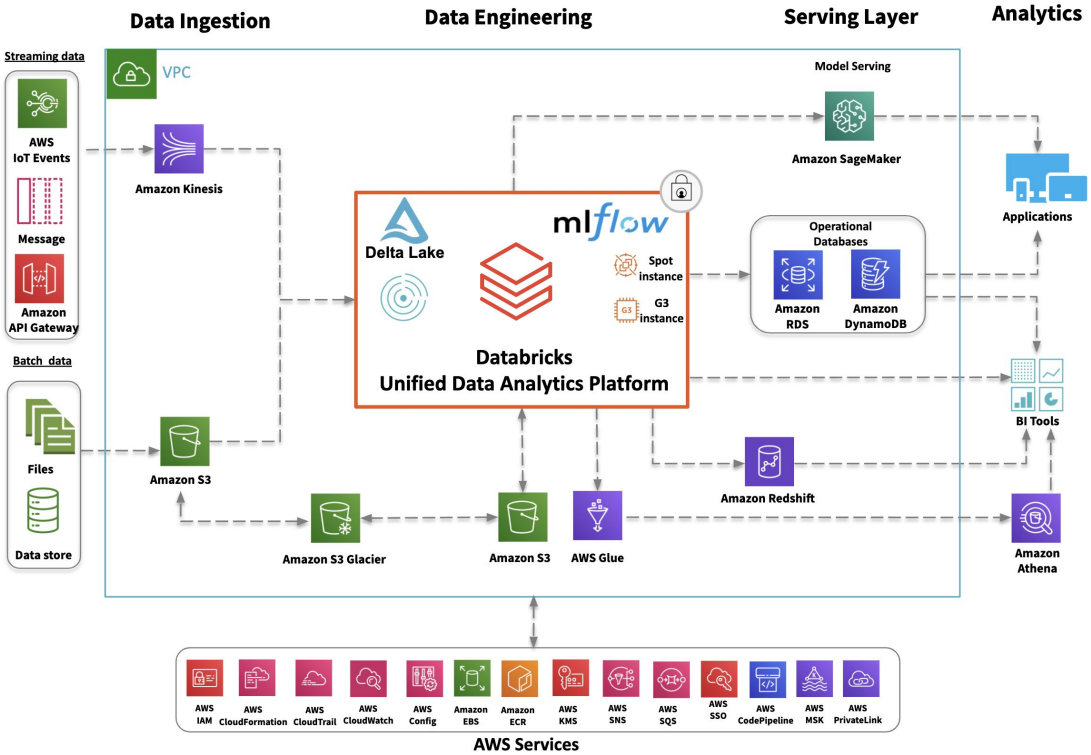
Demo Notebook links are in your reminder email.

AWS | Databricks Dev Day Quick Guide

https://dbricks.co/aws-devdays

Databricks Unified Data Analytics Technical Interactive Demo

Databricks & AWS data lake implementation

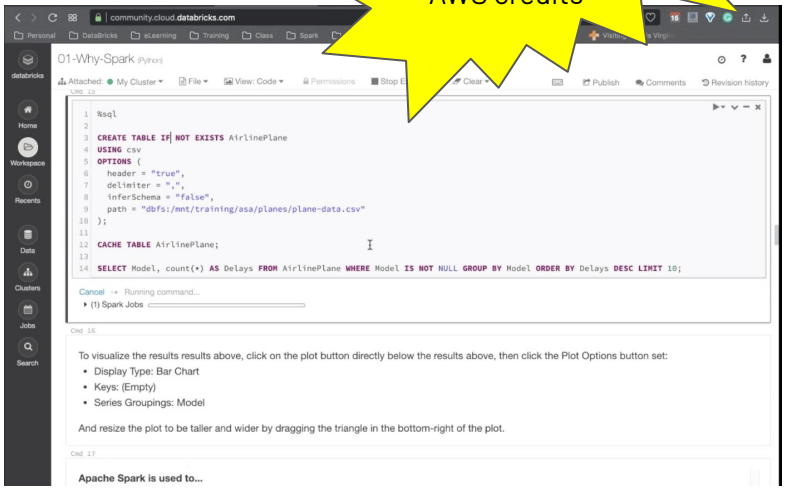


Free Training

Three virtual 2-hour training sessions using Databricks on AWS:

Getting Started with Apache Spark
Data Engineering and Streaming Analytics
Machine Learning

Speak to your Databricks team to get $50 in free AWS credits



Sign up now:

http://bit.ly/TrainingAWS

Post Event Survey

https://dbricks.co/mldevday

Thank you!