
snowflake
databricks
THE DATA MASTERY TOUR
Start your journey as a grand master of data engineering and analytics in the cloud.
Wifi Network: Databricks
Wifi Password: masterytour

Welcome!

Wifi Network: Databricks
Wifi Password: masterytour

What's Next

1:30 PM	Registration
2:00 PM	Welcome to the Data Mastery Tour
2:10 PM	Achieve Data Mastery with the Latest Advances in Combining ETL, Data Warehousing and Machine Learning
2:40 PM	Data Mastery in Action -- Customer Stories
3:10 PM	Networking Break
3:40 PM	Implement a Successful Data Analytics and ML / AI Project with Databricks and Snowflake (Demo)
4:30 PM	Q&A/Networking Happy Hour

Pursuing the Promise of AI

- 80% of all companies are considering AI projects
- ~90% are investing in AI related technologies

CIO Survey - IDG , 2018

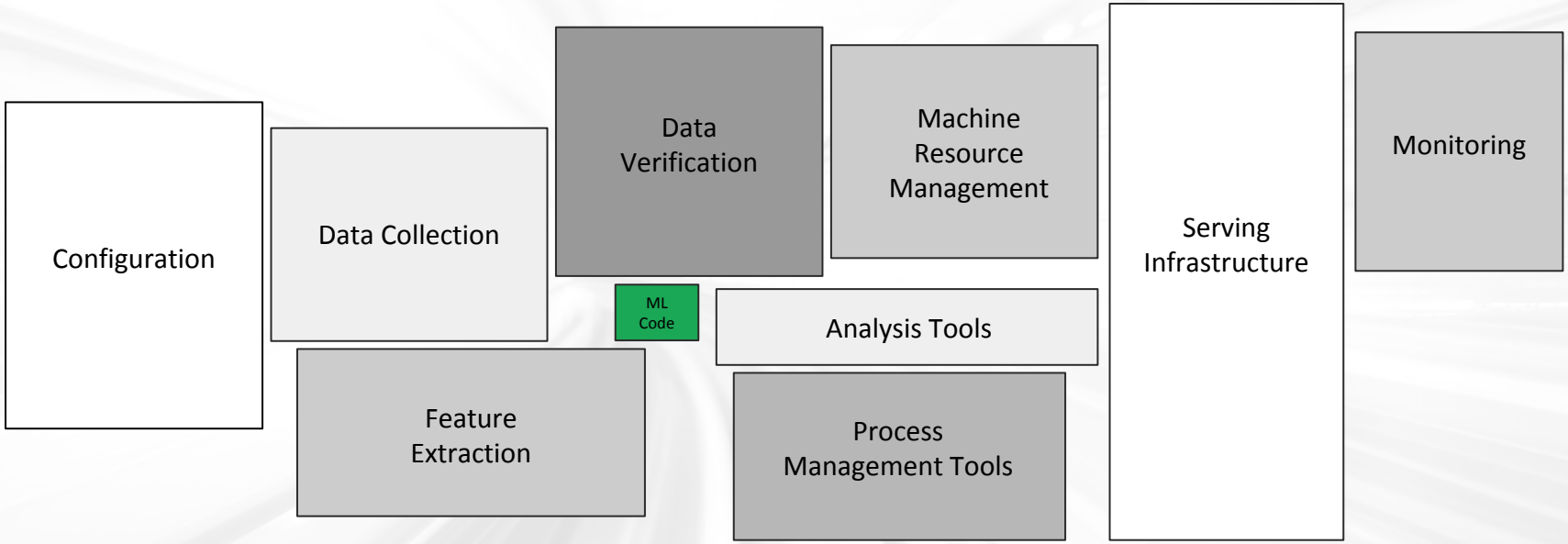
It’s Not About Just Algorithms

- Just 17% have moved an AI project into a core business area
- Only 1 in 3 projects are a success

CIO Survey, IDG, 2018

The Problem is with Data and Processes

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



Data Mastery Tour

Learn how to:

- Structure highly scalable data analytics pipelines
- Develop data strategies that bring data closer to ML and decision-making processes to make them more powerful and relevant
- Deliver results at scale with easy access to data by business-user dashboards or ML platforms

Does your data master you or do you master your data?

About Snowflake

Founded 2012 by database veterans, PhDs, with 80+ patents

First customers 2014, general availability 2015

Over 1000 customers today, growing rapidly

$923M+ in venture funding w/$3.5 B valuation

Data Warehouse Built for the Cloud

About Databricks

Founded 2013 at UC Berkeley Amp Lab

Over 2000 customers today, growing rapidly

First customers 2014, general availability 2015

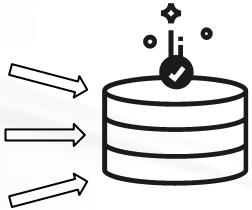
$498M+ in venture funding

DATA CHALLENGES

Data Challenges

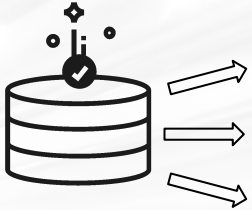
1. Messy data processes that don't provide reliable data for analytics
2. Slow onboarding of data and inability to handle data at scale limits powerful insights
3. Inability of data teams to collaborate slows innovation

Messy Data Processes

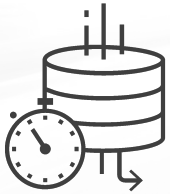


Structured/unstructured/batch /streaming

Dirty Data

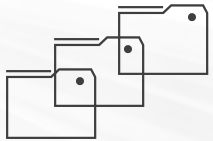


Accounting for Failures



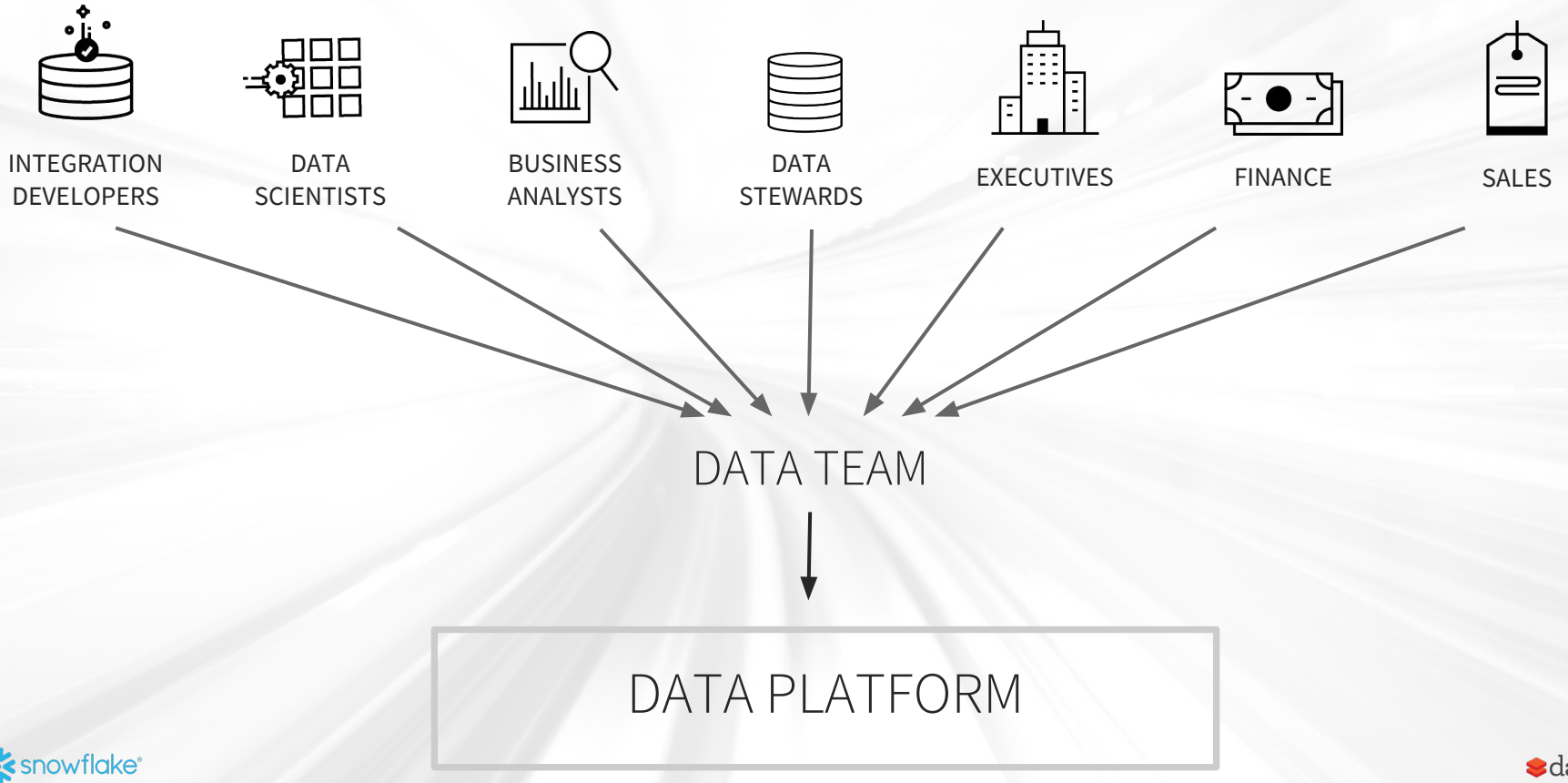
Where do I restart from?

Will I miss the SLA?

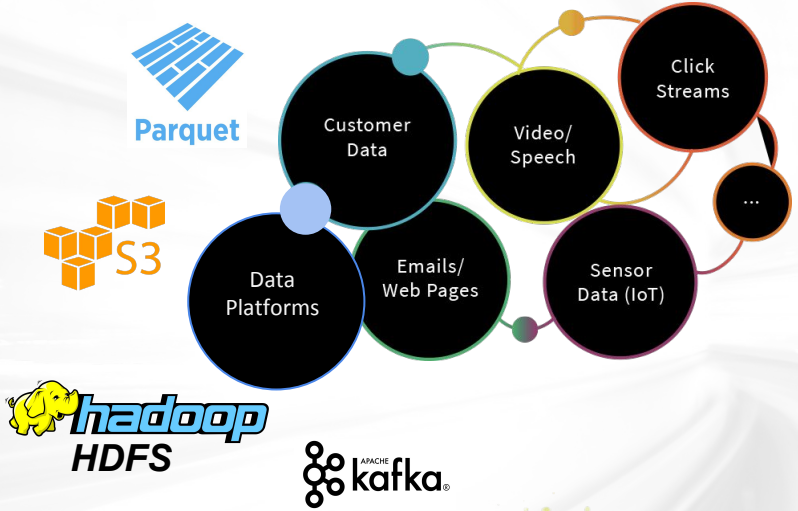


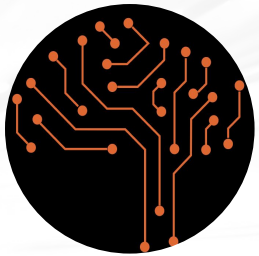
How do I automate this process?

The Problem of Scale - Bottlenecks



Data Teams are Siloed





DATA ENGINEERS

DATA SCIENTISTS

snowflake

databricks

The Path to Data Mastery

1

Reliable, performant data processing

2

Secure, scalable data warehouse service

3

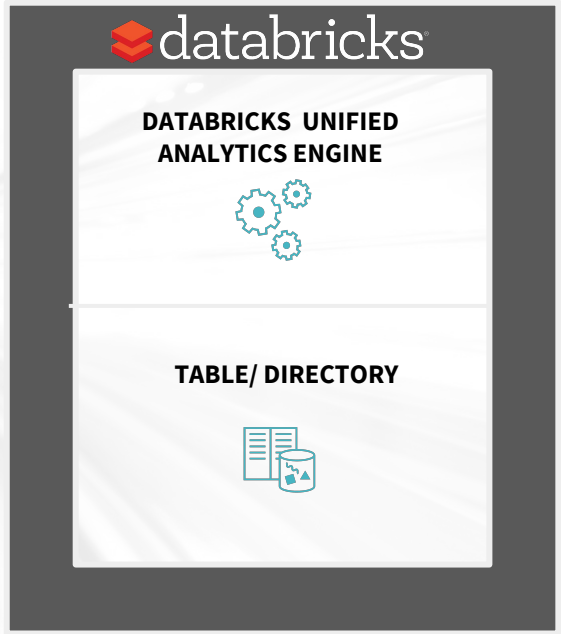
A collaborative machine learning platform

Reliable/Performant Data Processing

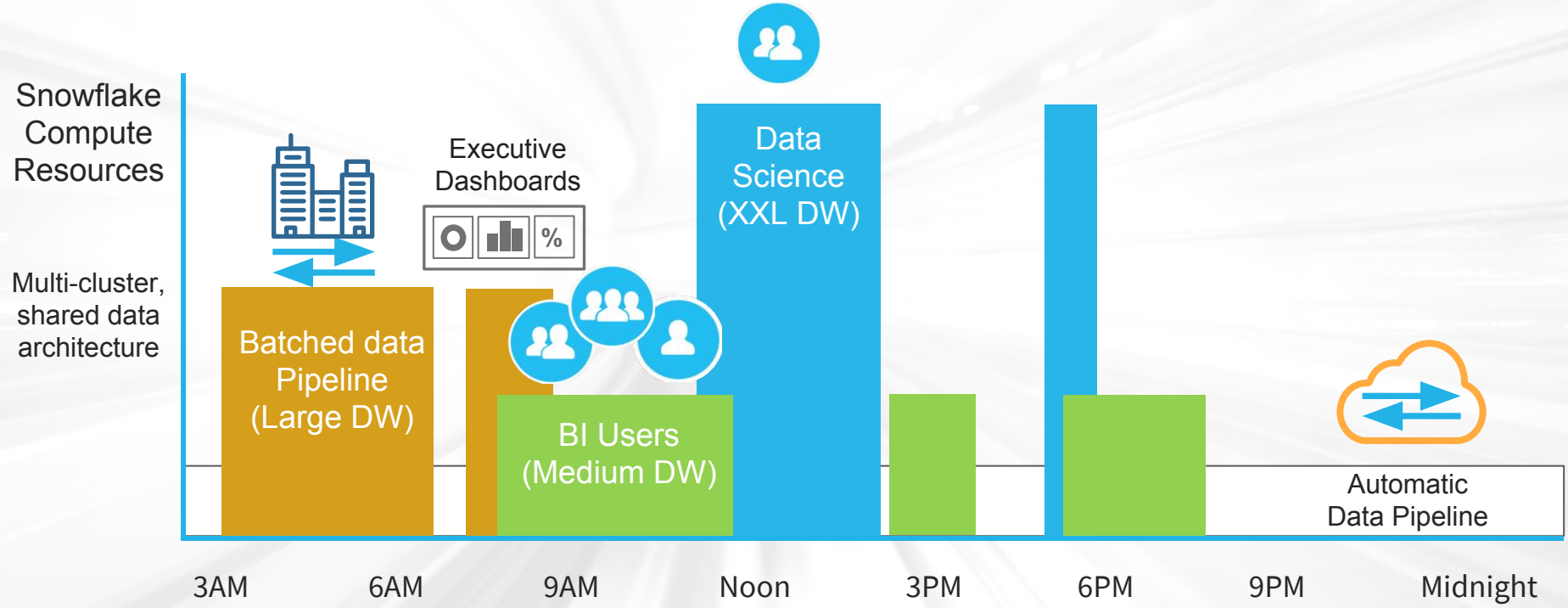
Fast streaming IOT data

CDC from traditional systems

GDPR deletes



Data Warehouse Concurrency and Workload Isolation



Customize for specific business needs
Never a fight for resources

A Collaborative Machine Learning Platform

Cmd 14

```
1 display(spark.sql("select * from products"))
```

(1) Spark Jobs

customer_id	card_number	checking_savings
237700	4427425867458998	chk
237701	4427422169569357	chk
237702	4427425979057208	sav
237703	4427424942295432	chk
237704	4427420913513382	chk
237705	4427425661330241	chk
237706	4427427764343261	chk
237707	4427429643437107	chk

Nauman Fakhar
1/31/2019, 10:04:01 PM

Jordan - can we enrich this dataset with CRM data from Salesforce? Need it for call center analyst dashboards & fraud models.

Jordan Martz
2/12/2019, 6:01:44 PM

Sure Nauman, let me source the sales force from API and add it to a Delta table for the customer 360 use case

Master Your Data

Raw Data
Batch + Streaming

databricks

Fast Reliable Data Pipelines
Built on Apache Spark

1

snowflake

Secure, Scalable, Cloud-built Data
Warehouse – Relational Processing

2

3

Machine Learning

databricks

tableau SOFTWARE

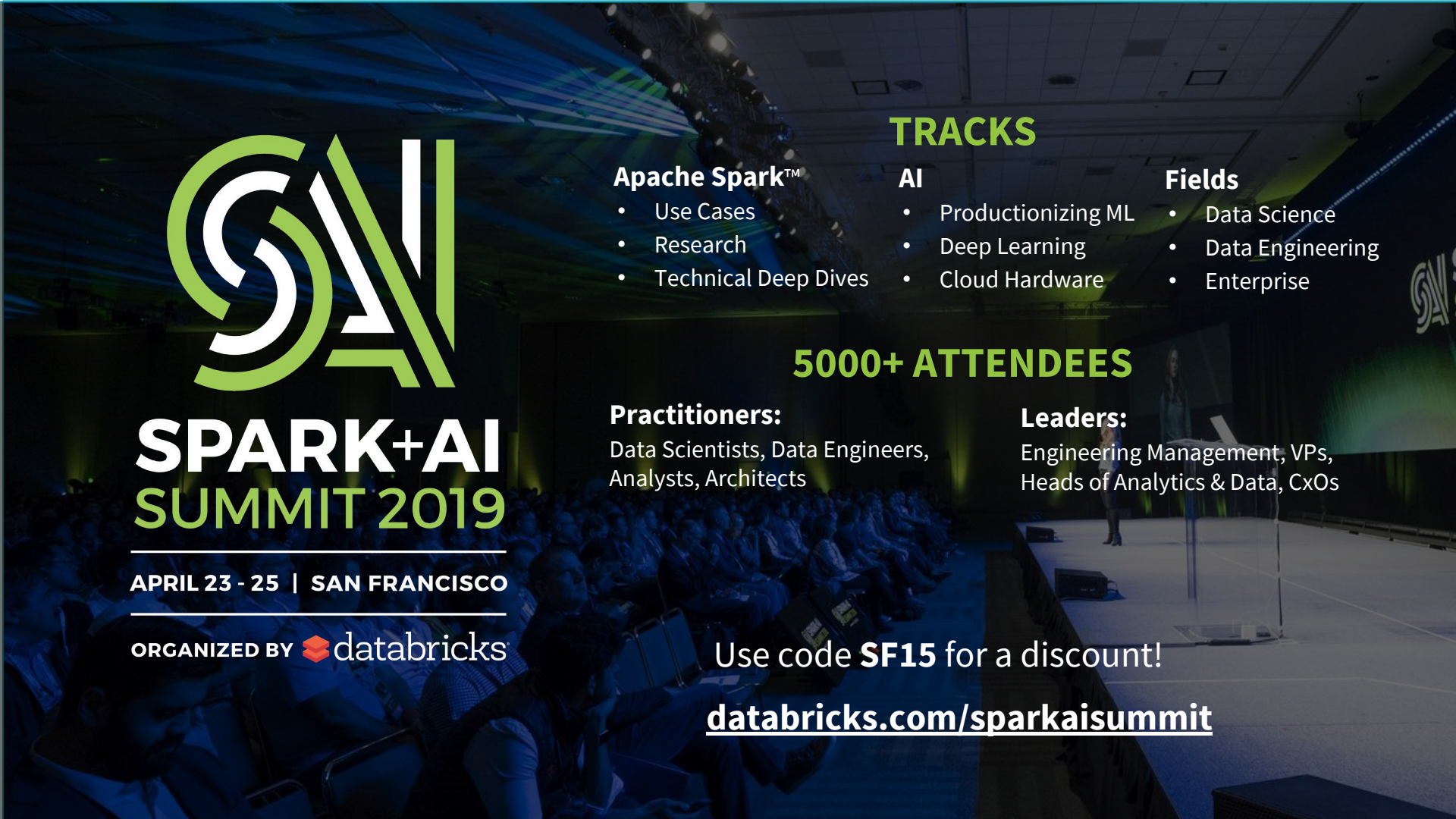
looker

BI Reporting
Dashboards

Joint Customers Mastering Their Data

Snowflake + Databricks =

powerful data pipelining, data access, and scalable workload support for the challenge of machine learning


SPARK+AI
SUMMIT 2019
APRIL 23 - 25 | SAN FRANCISCO
ORGANIZED BY databricks
TRACKS
Apache Spark™
• Use Cases
• Research
• Technical Deep Dives
AI
• Productionizing ML
• Deep Learning
• Cloud Hardware
Fields
• Data Science
• Data Engineering
• Enterprise
5000+ ATTENDEES
Practitioners:
Data Scientists, Data Engineers, Analysts, Architects
Leaders:
Engineering Management, VPs, Heads of Analytics & Data, CxOs
Use code SF15 for a discount!
databricks.com/sparkaisummit

$1,195 until the end of April! Full price is $1,995.

Data Mastery in Action

Customer Stories

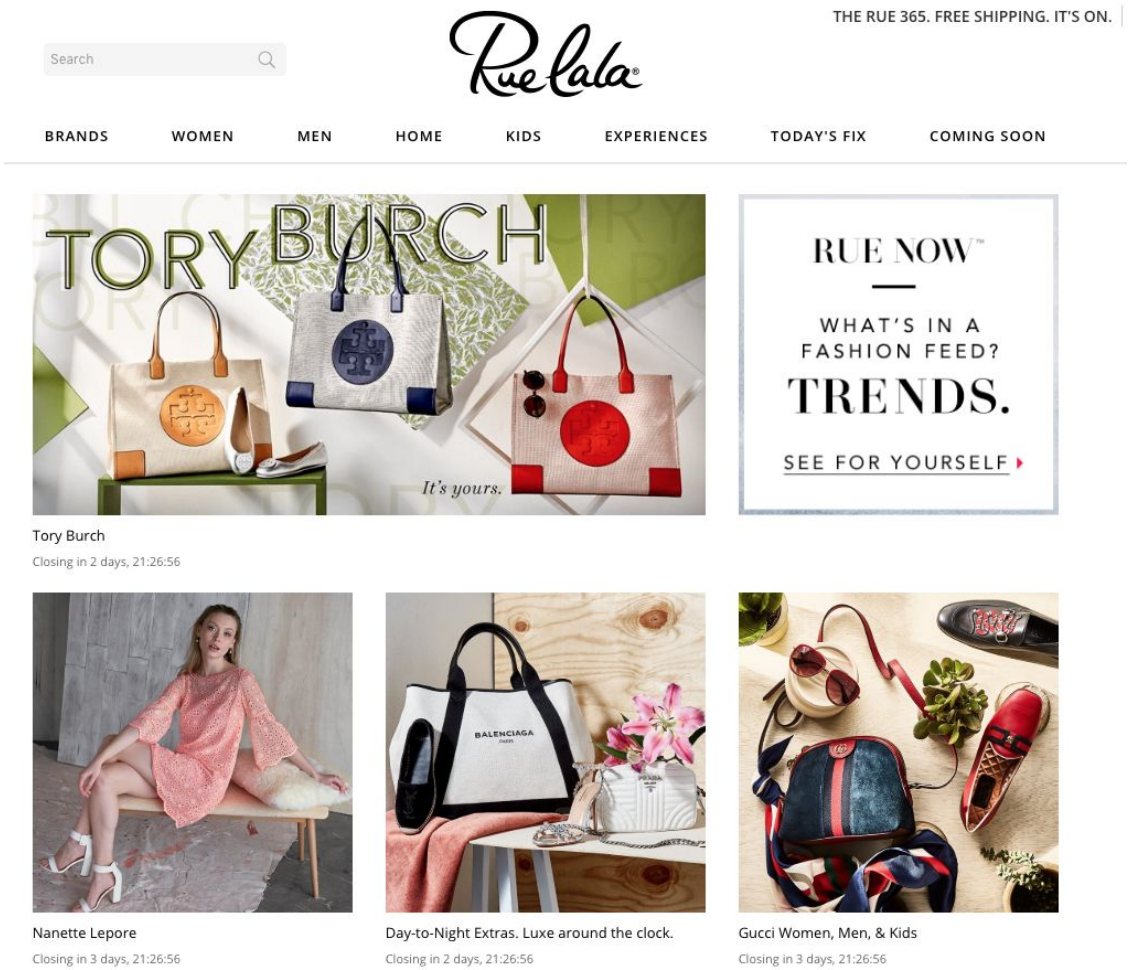
From Data to Personalization

Megan Wellons

mwellons@ruegiltgroupe.com

RUE GILT

GROUPE



Everything in "boutiques"

3 days

Email 11am, 3pm...

RUE GILT GROUPE

Tory Burch

Closing in 2 days, 21:26:15

Nanette Lepore

Closing in 3 days, 21:26:15

Day-to-Night Extras. Luxe around the clock.

Closing in 2 days, 21:26:15

Gucci Women, Men, & Kids

Closing in 3 days, 21:26:15

Looks for This Year's Top Travel Spots

Closing in 2 days, 01:26:15

Rue's One-Stop Glam Shop: La Mer to YSL

Closing in 2 days, 18:26:15

Nicole Benisti

Closing in 2 days, 21:26:15

Gucci Women, Men, & Kids

Closing in 3 days, 21:24:26

Men's Polos Featuring Lacoste. Can't beat classic.

Closing in 3 days, 21:24:26

Socks & More for His Top Drawer

Closing in 3 days, 21:24:26

$55 Denim: It's Moonlight Madness

Closing in 0 days, 01:24:26

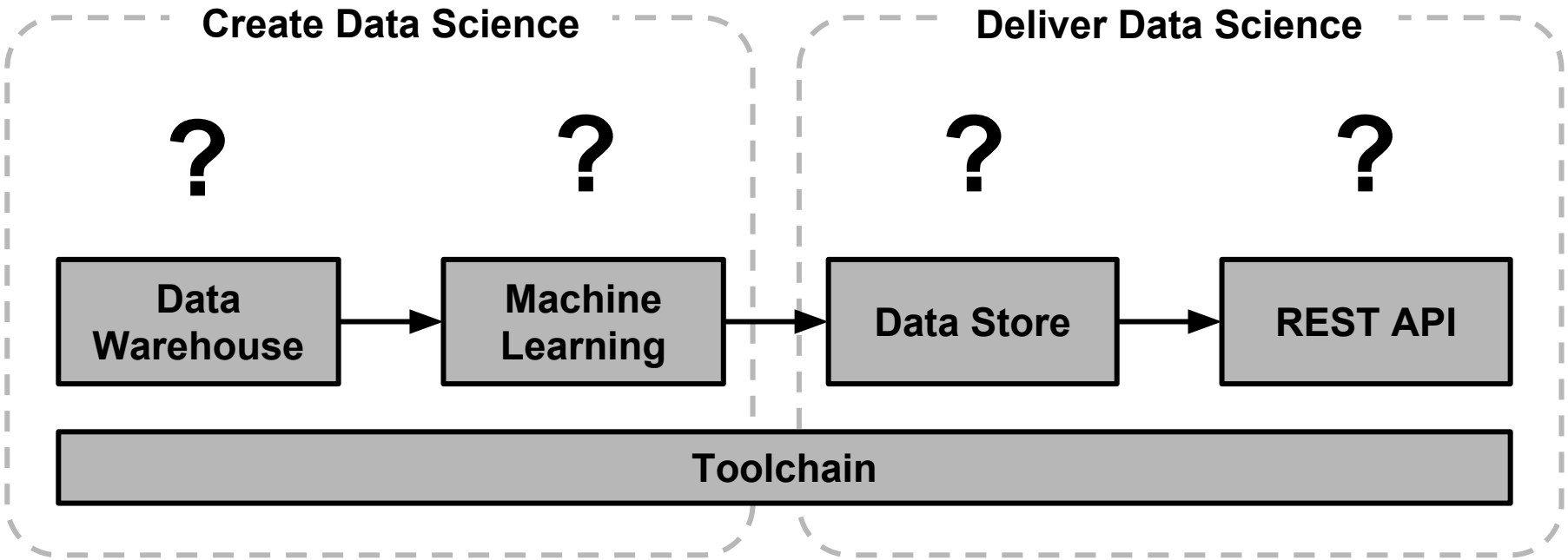
Diamond Watches for Her & Him. Talk about great timing.

Brioni & More Men's Tailored Musts: Hey, handsome.

JOE'S Jeans Women & Men

Closing in 1 day, 21:24:26

Data Science Pipeline in Concept



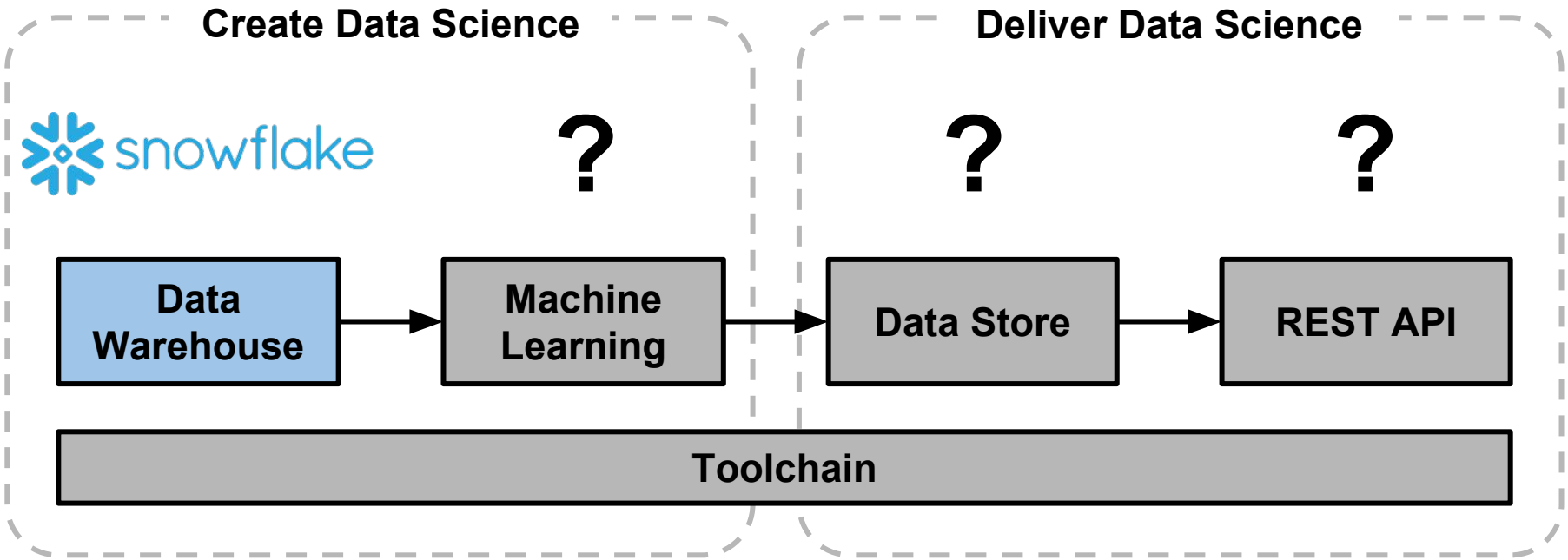
RUE GILT GROUPE

Challenge: Data Warehouse

Traditional Relational DWH	Snowflake Cloud DWH
Concurrency contention	Separate compute warehouses for dev/qa/prod/adhoc
DBAs spread thin	Fully managed solution
Painful/slow scaling	On-demand and instant scaling
Limited to structured data	Supports semi-structured too
Data volumes	Unlimited Storage

RUE GILT GROUPE

Solved: Data Warehouse



RUE GILT GROUPE

Challenge: Machine Learning Platform

Standalone Python	Databricks UAP
Single node	Managed Spark cluster of any size
ETL is also hard	Seamless integration Snowflake & Databricks
Complex path to production	Same SQL and notebooks for dev/QA/prod
Engineering & tools siloed	Unified software development + Data Science
DevOps is hard	Minimal cluster management

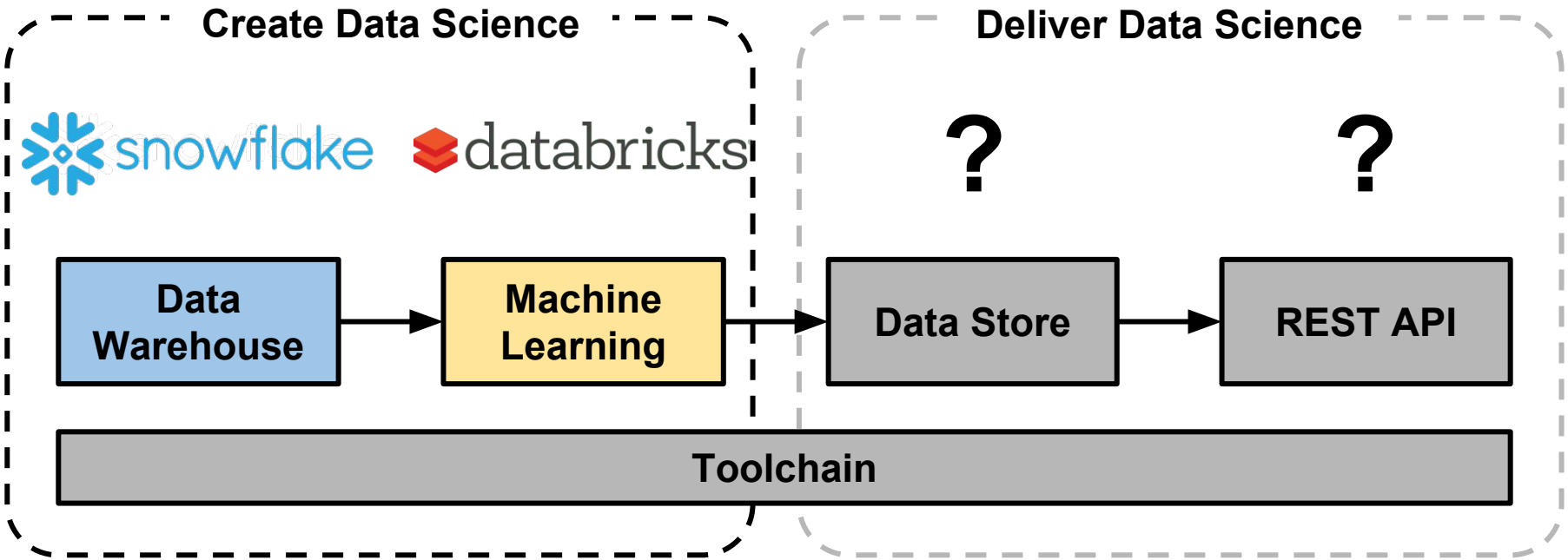
RUE GILT
GROUPE

Easy Setup using Snowflake Connector

```
val sourceData = spark
    .read
    .format("snowflake")
   .options(snowflakeOptions)
   .option("query", sourceQuery)
   .load()
```

RUE GILT
GROUPE

Solved: Machine Learning Platform

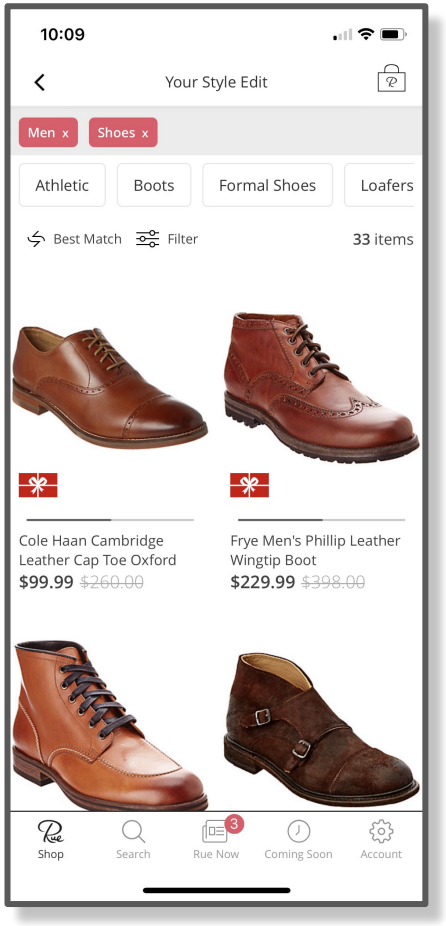


RUE GILT GROUPE

Member → Brand Recommendations

Boutiques

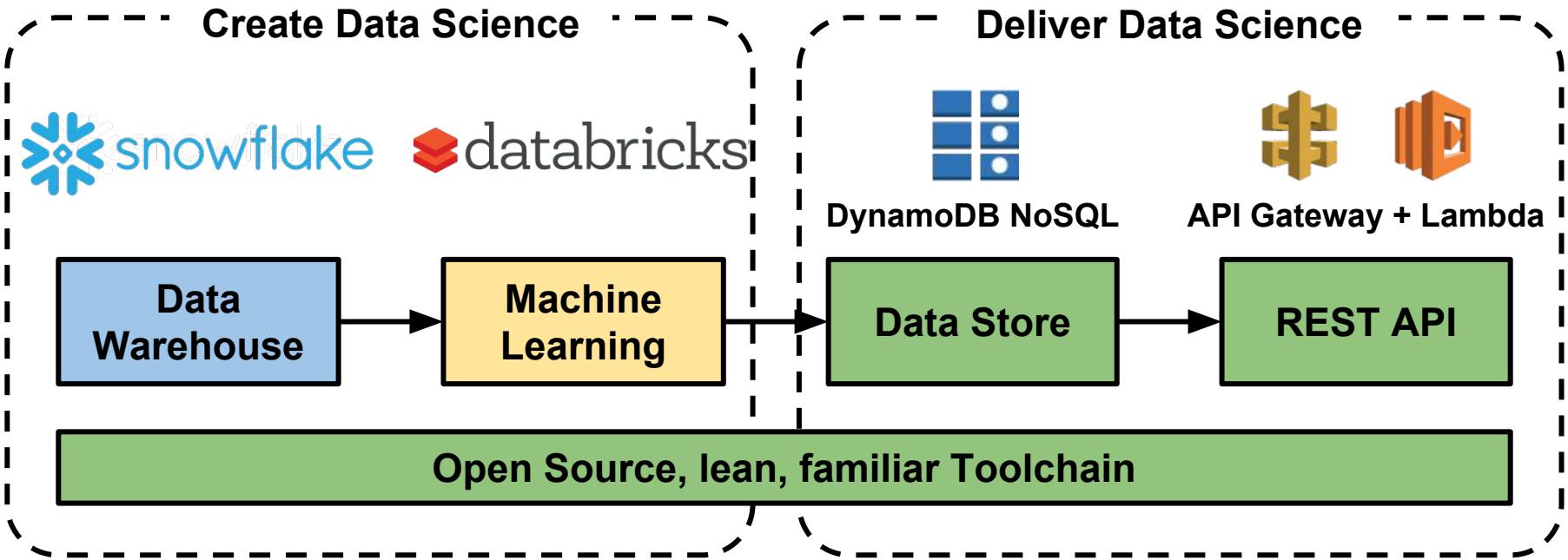
Detail



Architecture

RUE GILT
GROUPE

Solved: Here's Where we Ended Up



RUE GILT GROUPE

End state: No Compromises

	Requirements	Achieved
Fast to production	Simpler than legacy	2 people 2 months
DWH	Uptime, speed, scaling	Excellent SLAs, fast, elastic
ML platform	Turnaround is hours	Notebooks make it minutes
Architecture	“In the Cloud”	100% cloud-native
API latency	< 150 ms	~ 16 ms = 60 Hz
Uptime	99.95%	100%
Deployment downtime	< 10 mins	0

Snowflake + Databricks: Unifying Big Data and AI

- Built-in
  - All required jars and libraries are pre-loaded in the Databricks Unified Analytics Platform (and kept up to date)
  - Quick start notebooks and secure credentials using Databricks Secrets API
- Optimized
  - Automatic query pushdown optimizes performance without manual configuration
  - Native support for JSON and other semi-structured data
- Fully Managed and Elastic
  - Databricks provides full management of your clusters with auto-scaling / auto-termination
  - Snowflake automates all maintenance and tuning for optimal performance
  - Instantly scale up/down Databricks Spark clusters and Snowflake compute clusters

RUE GILT
GROUPE

You can have it all...

- **Integration complexity?**
  Dedicated Snowflake/Databricks connector
- **IT overhead?**
  100% managed cloud services
- **Data volume?**
  100’s of billions of rows—no problem
- **Custom data algorithms?**
  Databricks has everything we need out-of-the-box, and allows flexibility for customization

RUE GILT GROUPE

Data & Analytics at Devon Energy

NYSE: DVN
devonenergy.com

devon

About Devon Energy

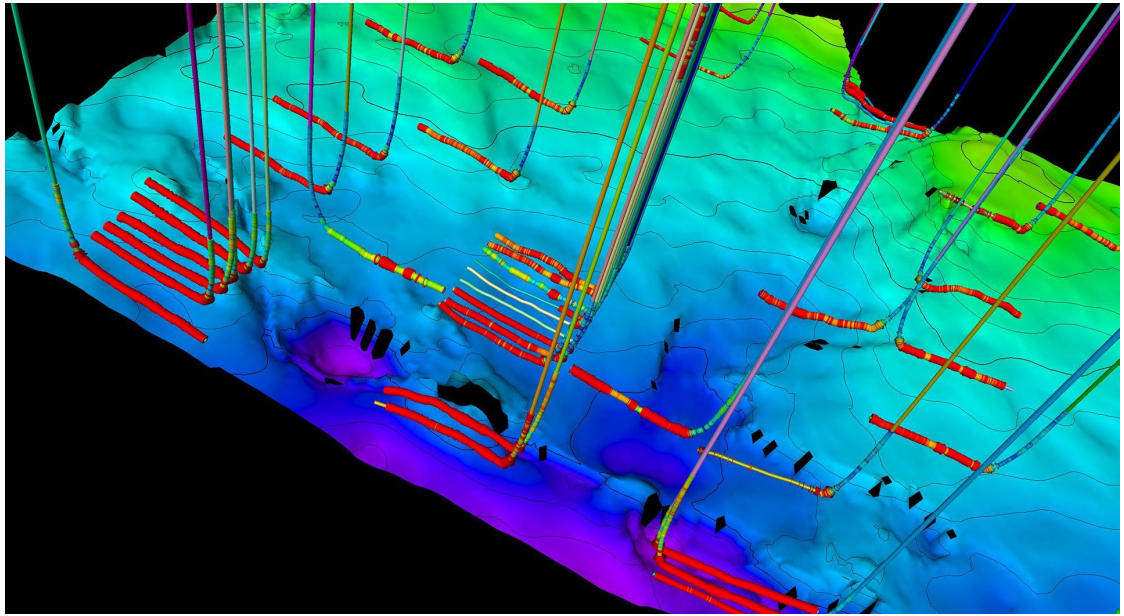
Devon Energy is a leading independent oil and natural gas exploration and production company.

- Over 3,000 employees
- $22 billion market cap
- Produce 541,000 Boe per day

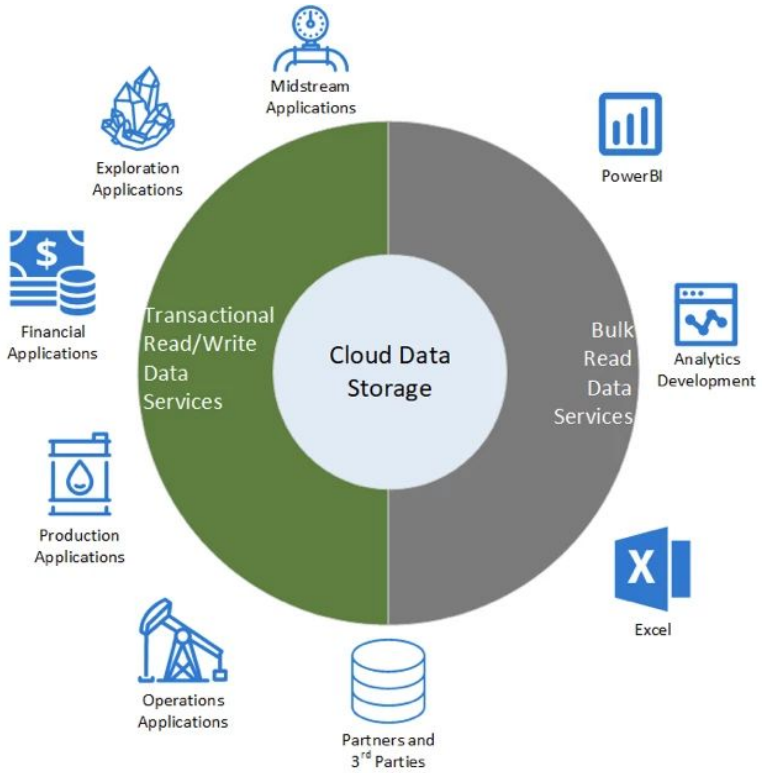
Business Challenges

As a producer of unconventional oil and gas reserves we face both technical and business challenges.

- Our wells wells can be more than 12,000 feet long, horizontally
- The oil and gas industry is under tremendous pressure be cost efficient
- How can we ramp up activity while maintaining technical excellence?



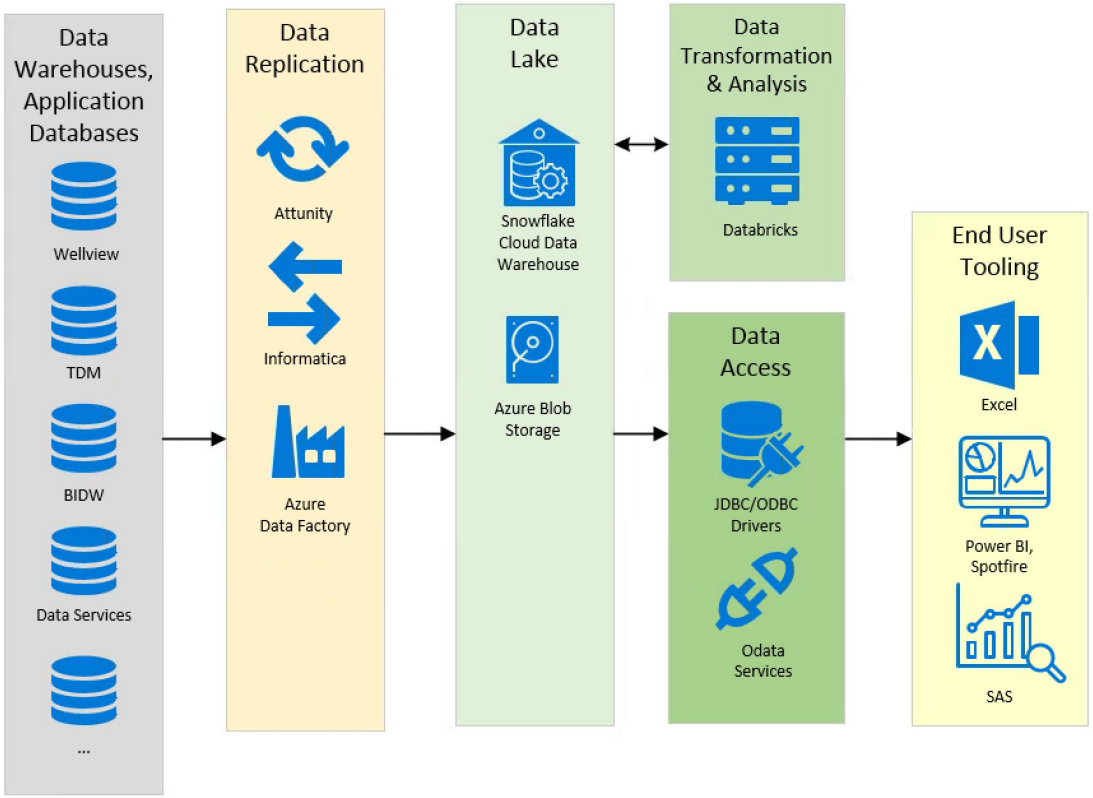
Data and Technology Driven



Data and technology are critical to our business.

We are building a Data Lake and associated integration services, the Data Hub, in order to accelerate time to insight and improve our agility.

Data Hub Architecture



The Data Hub reinvents our Data Warehouse and Integration landscape. This allows anyone to build their own data and analytics solutions and share insights.

Citizen Development

Our users are generally very technical, they understand the data and business context. Empowering them with a self-service platform makes everyone more productive.

Everyone in the company has access to Databricks and Snowflake and we offer training and assistance to anyone who is interested so they can get started.

We have over 12,000 tables registered in Databricks and Snowflake with data that's constantly updating.

Subsurface Analytics

We have massive volumes of geophysical data, users want to unlock the value from that data by analyzing it at scale.

Our primary petrophysical database contains information about sensor readings describing wells, it has over 100 billion records on various properties of our wells.

We can now do bulk analysis on all of our well logs to synthesize new information, look for correlations or connect it to production information.

Unified Platform

In the past our data storage, processing and machine learning toolchains were loosely connected, at best.

Moving up the value chain of data processing and analysis required switching tools at certain break points. Data engineering tasks and skillsets were disconnected from machine learning and analysis.

Our users can write basic queries and advance to machine learning and Python using Databricks with data stored in Snowflake without switching platforms or toolchains.

Improved Consistency

We are now doing Continuous Integration and Deployment with Databricks and Azure DevOps to manage the testing and promotion of data engineering pipelines.

Our previous toolsets didn't support code management or mature unit or integration testing approaches. Now, we have used a combination of out of the box and custom solutions to build automated pipelines.

We can automatically promote new enterprise objects and perform unit testing against reference datasets to ensure consistency or look for drift.

Document Processing

Many of our well sites are remote, with poor connectivity and many of our suppliers are small.

- Paper tickets are common
- PDF Invoices and documents are emailed
- Got 21.5 million invoices in 2017

Distributed Deep Learning

We have millions of invoices to use for training data. Processing and training deep learning algorithms would take too long without parallelism.

- We used Spark to call many instances of the Azure Cognitive OCR APIs to process the invoices
- Deep learning models developed in TensorFlow and distributed with Horovod were easy to scale on the Databricks ML Runtime
- One integrated pipeline for preprocessing and deep learning made iterating on the models and feature engineering faster
- New ML Runtime and AutoML make it easier to empower analysts and data scientists

smartsheet

Analytics and Insights at Smartsheet

SMAR
LISTED
NYSE

Paivand Jalalian & Christine Haggerty

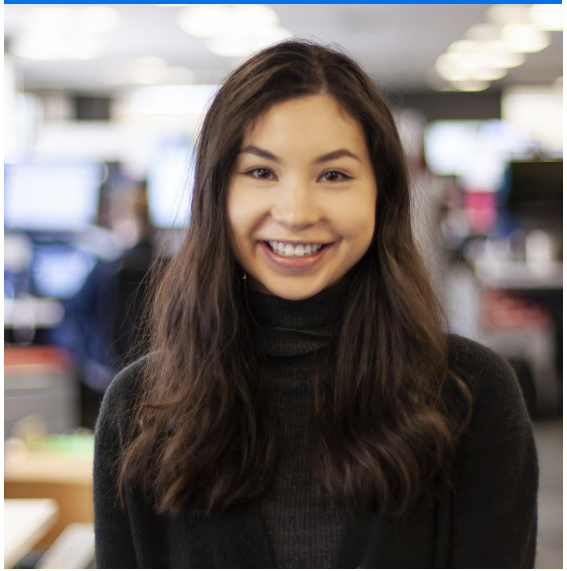
2/21/19

Agenda

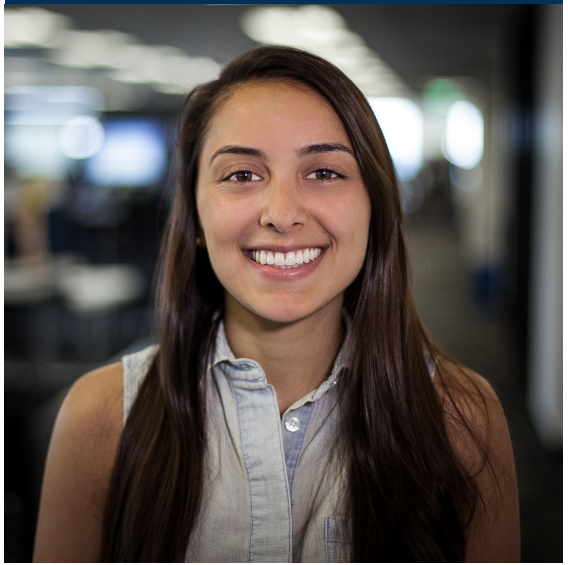
1. What is Smartsheet and why is data analytics important to us?
2. How do Snowlake and Databricks help us achieve our purpose?
3. What kind of impact do Snowflake and Databricks make?

Data Science at Smartsheet

Christine Haggerty
Data Scientist



Paivand Jalalian
Data Science Manager



smartsheet

What is Smartsheet?
Why is Data Analytics Important?

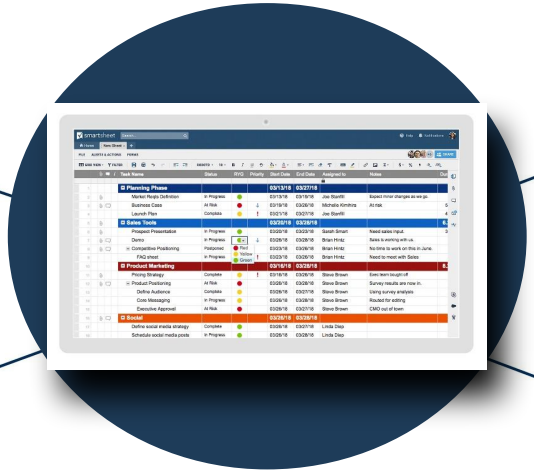
The Smartsheet Platform for Work Execution

Empowering organizations to **plan, capture, manage, automate, and report on work at scale**.

$47M
Q3 FY19 Revenue(1)

59%
YoY Revenue Growth (2)

77K+
Domain-Based Customers
(1),(4)



Capture
- Forms
- Connectors
- Integrations

Report
- Dashboards
- Portals
- Dynamic Reports

Automate
- Automations
- Converse.ai
- API
- Accelerators

Scale
- Control Center
- Security
- Auditability
- Compliance
- Accelerators

Plan & Manage
- Grids
- Projects
- Cards
- Calendars

Notes
1. As of October 31st, , 2018. Year-over-year revenue growth from Q3 FY18 and Q3 FY19.
4. We define domain-based customers as organizations with a unique email domain name such as @cisco. All other customers, which we designate as ISP customers, are typically small teams or individuals who register for our services with an email address hosted on a widely used domain such as @gmail, @outlook, or @yahoo.

smartsheet

One Platform, Many Uses

Project Management
- Project tracking
- Resource management
- Executive reporting
- Gantt charts

It & Operations
- Inventory / Assets
- System migration
- Issues triage
- Maintenance

Sales
- Sales pipeline
- Customer contacts
- Sales training
- Sales rep activities

Marketing
- Events
- Campaigns
- Website content
- Product launches

Company Management
- Company objectives
- Balanced scorecard
- Employee vacations
- Meeting action tracking

Product Development
- Development projects
- QA scenarios
- Production process
- Feature prioritization

Human Resources
- Candidate tracking
- New hire onboarding
- Exit processing
- Corporate calendar

Finance
- Contract process
- Quarterly reviews
- Corporate metrics
- Budget rollups

Specialty Solutions
- Store / branch communications
- Rental property maintenance
- Construction projects
- Client engagement management

smartsheet

Data analytics is not important. It’s imperative.

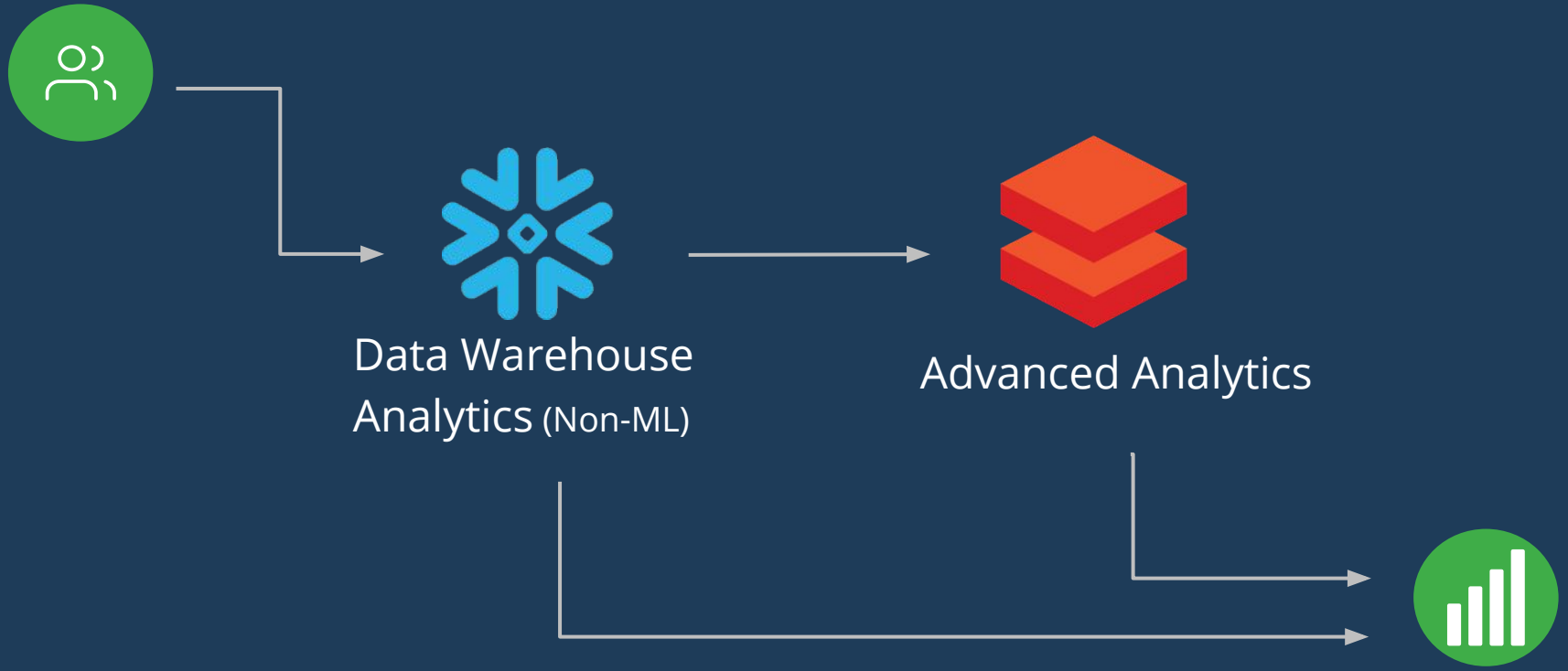
How do Snowlake and Databricks Help Us Achieve Our Purpose?

Data Platform Comparison

Differences in key features

	Legacy MySQL Platform (On-Prem)	Snowflake Platform (Cloud)
Replication & Data Latency	Easy & fast direct from app (~1 min)	Pipeline to S3 + Airflow (~5min)
Availability	Replica, constant maintenance	Distributed System
Easy Scalability	No - reaching limits of system	Yes
Elasticity	No - query tuning required	Yes (Minutes)
Ease of Use	MySQL - easy to learn	ANSI SQL - easy to learn
Occurence of table locks?	Frequently	Rare
Query large tables, ex. Aggregating 3B row table	Slow, Killed after running for 1.5 hours	Quick especially with adjustment of cluster, ~ 20 Minutes
Permissions	Simple based on DB and action	With views, as complex as needed
Syntax	Restricted to Mysql	ANSI sql, Java, + Connection to Databricks for ML, python, etc

Databricks for machine learning, Snowflake for everything else.



Key Benefits

Databricks + Snowflake together provides the unique ability to implement advanced analytics while maintaining structure and integrity of underlying data.

Snowflake
Platform ensures data structure and integrity

- Query speed (scaleable) + query large datasets
- Conditional Permissions
- Creation of views + copy DBs, schema's, tables with in seconds
- Un-drop tables
- Departmental usage w/ monitoring
- Connection to Tableau

Databricks
Flexibility

- Utilize different languages & packages
- Create UDFs & procedures (loops)
- Schedule jobs
- Easy Visualizations
- Intuitive UI/UX
- Share Notebooks
- Versioning via Git
- Allows self service via "Run" permissions

Use Cases and Impact

Use Cases

	Solution	Impact
Anomaly Detection	• Query 100M+ rows of telemetry data in Snowflake • Pivots, aggregations & visualizations in Databricks • Distribute Databricks dashboard to necessary parties	+ Results and insights derived quickly + Easy/fast distribution of data + Increase speed to action
Text Analytics of Unstructured Customer Comments	• Raw comment data stored in Snowflake • NLP model in Databricks Notebook (R) • Connector for end-to-end solution	+ Time savings human effort minimized + Consistency in categorizations + Ability to pull out patterns to derive insights

*The combination of **Snowflake** & **Databricks** has not only allowed us to finally **keep up** with the growing scale of our company but **get ahead**.*

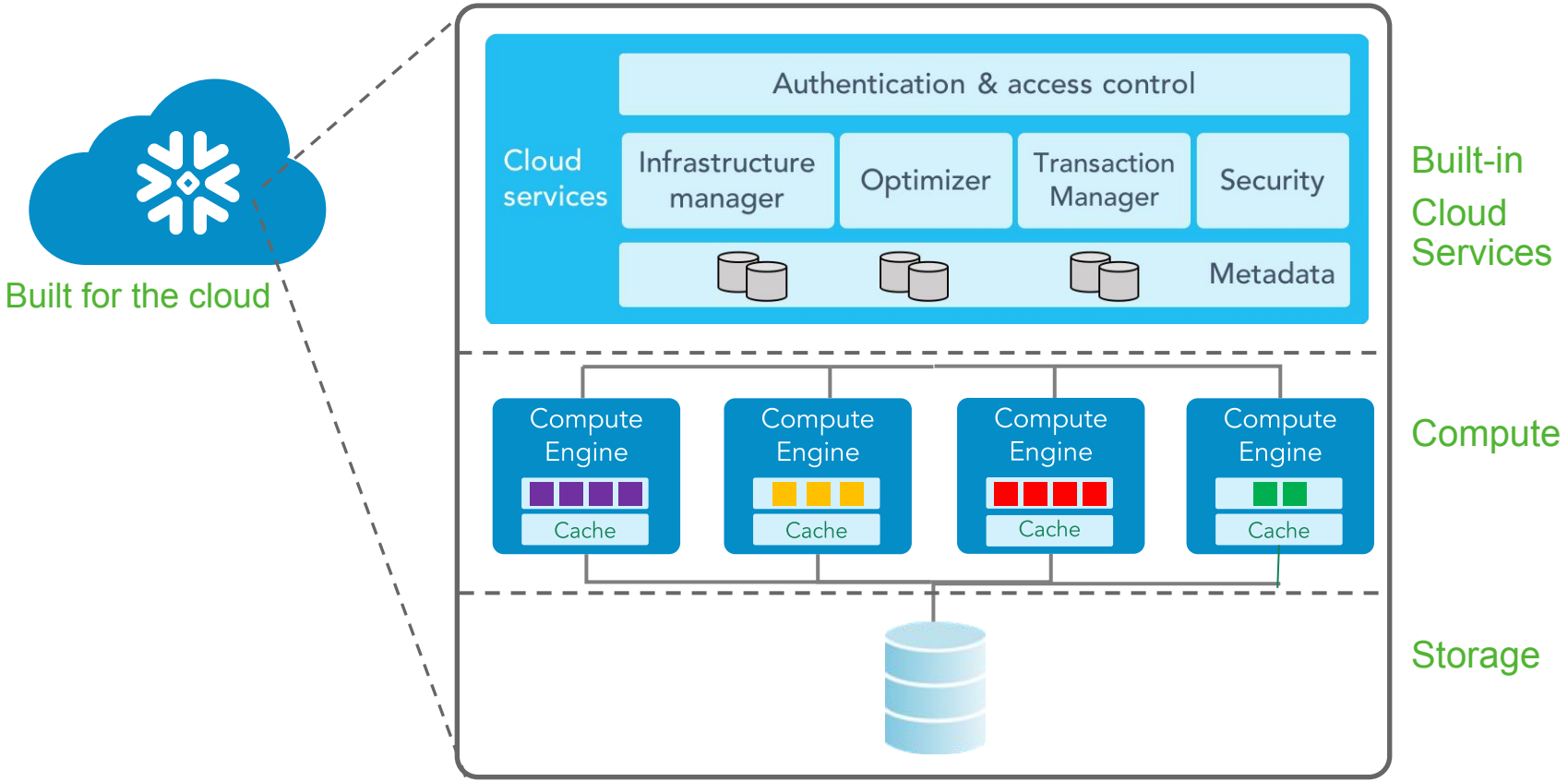
Questions?


smartsheet

What's Next

1:30 PM	Registration
2:00 PM	Welcome to the Data Mastery Tour
2:10 PM	Achieve Data Mastery with the Latest Advances in Combining ETL, Data Warehousing and Machine Learning
2:40 PM	Data Mastery in Action -- Customer Stories
3:10 PM	Networking Break
3:40 PM	Implement a Successful Data Analytics and ML / AI Project with Databricks and Snowflake (Demo)
4:30 PM	Q&A/Networking Happy Hour

Multi-cluster, Shared Data Architecture



Q&A - Call to Action

Databricks Trial

https://databricks.com/try-databricks

Snowflake Trial

https://trial.snowflake.com/


snowflake
databricks
THE DATA MASTERY TOUR
Thank you!