



**Build. Unify. Scale.**

WIFI SSID:SparkAISummit | Password: UnifiedAnalytics

ORGANIZED BY  
 databricks



SPARK+AI  
SUMMIT 2019

# Databricks + Snowflake: Catalyzing Data and AI

Garren Staubli  
Solutions Architect

[garren@databricks.com](mailto:garren@databricks.com) | @gstaubli

**#UnifiedAnalytics #SparkAISummit**  
**Slides & Resources: [garrens.com/DataSnowCat](https://garrens.com/DataSnowCat)**

# Agenda

Introductions

Scenario

Challenges

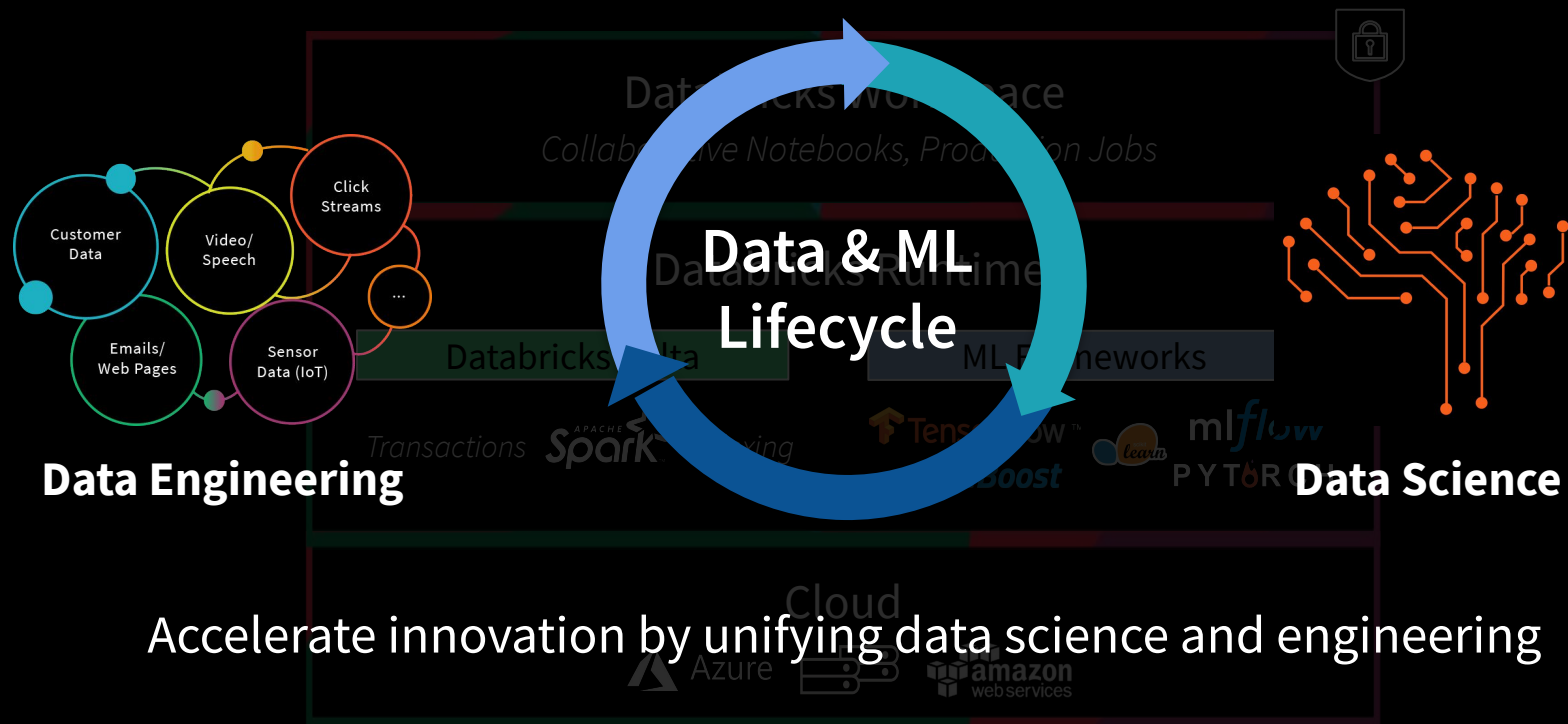
Solutions

Demo

# Introductions - Me



# Introductions - Databricks



# Introductions - Snowflake

What is Snowflake?



**All-new SQL data warehouse**  
*No legacy code or constraints*

**Delivered as a service**  
*Infrastructure, resiliency, optimization built in*

**Designed for the cloud**  
*Running in Amazon Web Services*

 snowflake

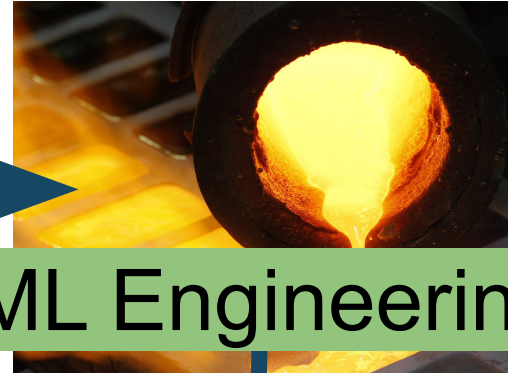
35

Forget Oil. Data is worth  
more than Gold

# Scenario



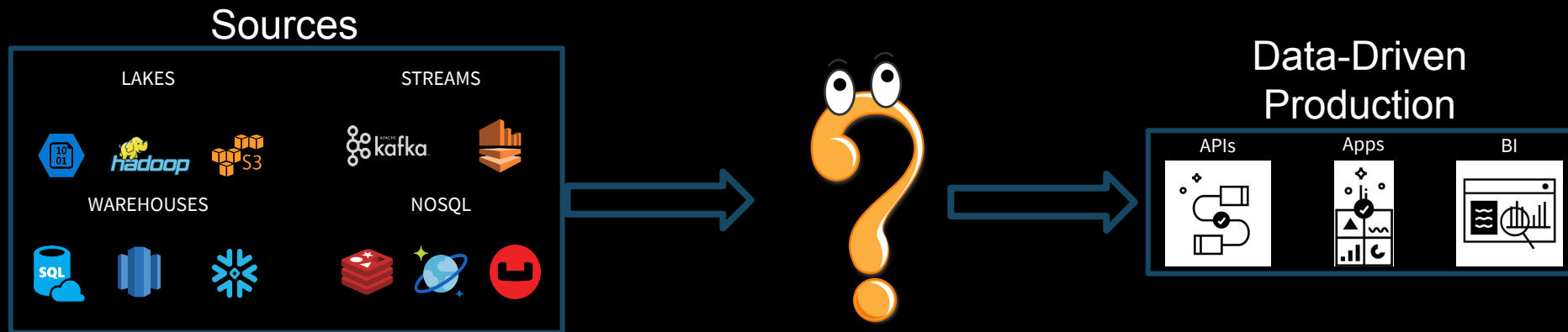
# Scenario - Annotated



# Scenario - Reality



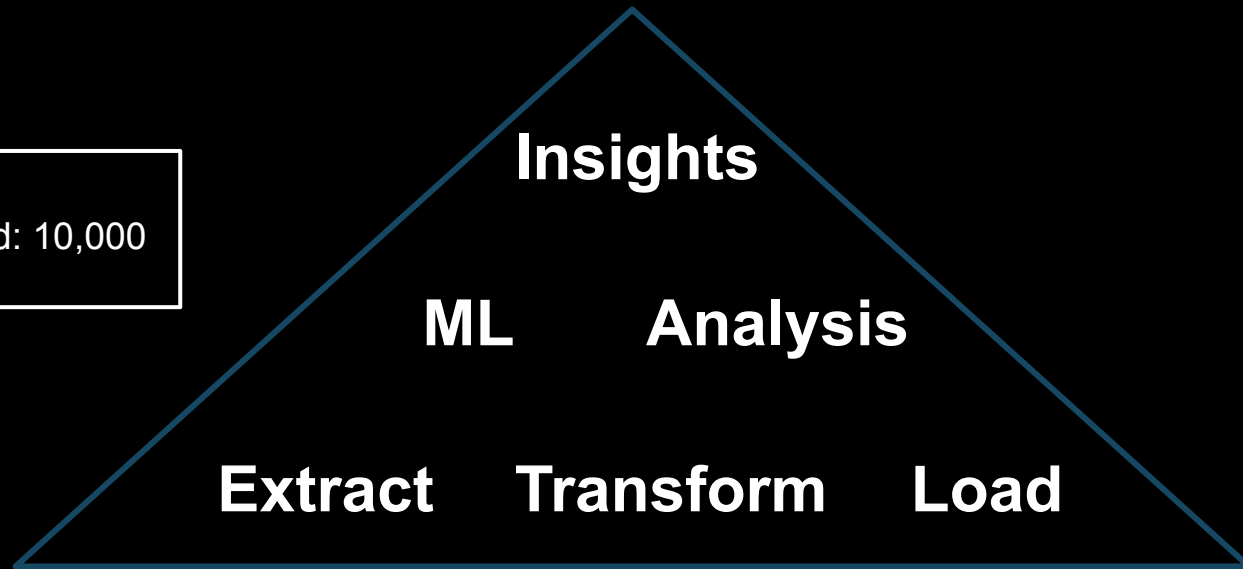
# Challenges



# Challenges - Reality



Partitions: 20  
Rows per second: 10,000  
Format: JSON



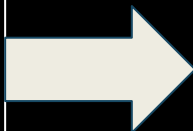
# Challenges & Solutions - ETL



Partitions: 20  
Rows per second: 10,000  
Format: JSON



**Sources**  
**Syntax**  
**Scale**  
**Languages**  
**Performance**  
**Expressiveness**



**Flat, RDBMS, Streams, etc**  
**Unified batch & stream APIs**  
**Autoscaling with usage**  
**Python, Scala, SQL, R & Java**  
**JVM w/ optimization**  
**Multilevel APIs (SQL + RDDs)**

# Challenges & Solutions - ETL



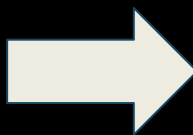
Partitions: 20  
Rows per second: 10,000  
Format: JSON



**DELTA LAKE™**

**Malformed Records**  
**Errors**  
**Changing Fields**  
**Writes**

- **Performance**
- **Semantics**
- **Reliability**



**Ignore/infer + log records**  
**Handle + retry w/ checkpoint**  
**Schema Evolution**

**Partitioned + optimized files**  
**Exactly once**  
**ACID transactions**

# Challenges & Solutions - ML

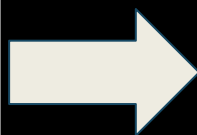


Partitions: 20  
Rows per second: 10,000  
Format: JSON



**Data Access**  
**Syntax**  
**Collaboration**  
**Models**

- Iteration
- Reproducibility
- Deployment



**Apache Spark + Delta**  
**Koalas**  
**Databricks Notebooks**  
**mlflow**

- Tracking
- Projects
- Models

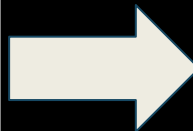
# Challenges & Solutions - Analysis



Partitions: 20  
Rows per second: 10,000  
Format: JSON



**Time to value**  
**Intermittent demand**  
**Language**  
**Common Tooling**  
**Ease of Use**  
**Cost control**



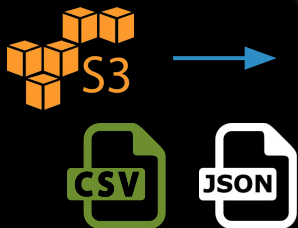
**Interactive queries**  
**Instant Scaling**  
**SQL**  
**Tableau, PowerBI, etc**  
**Optimized DWaaS**  
**Decoupled storage + compute**



# Final Solution Architecture



Partitions: 20  
Rows per second: 10,000  
Format: JSON



mlflow

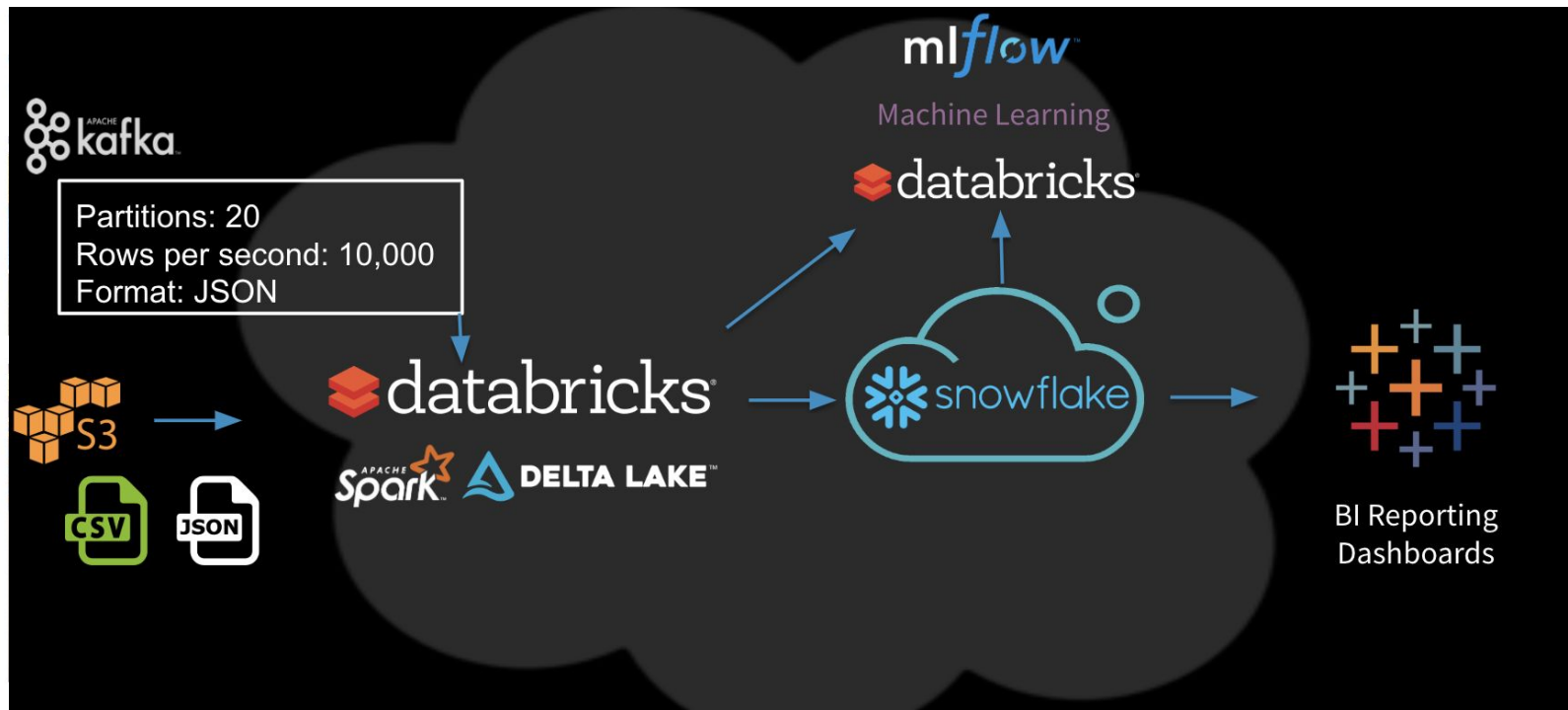
Machine Learning



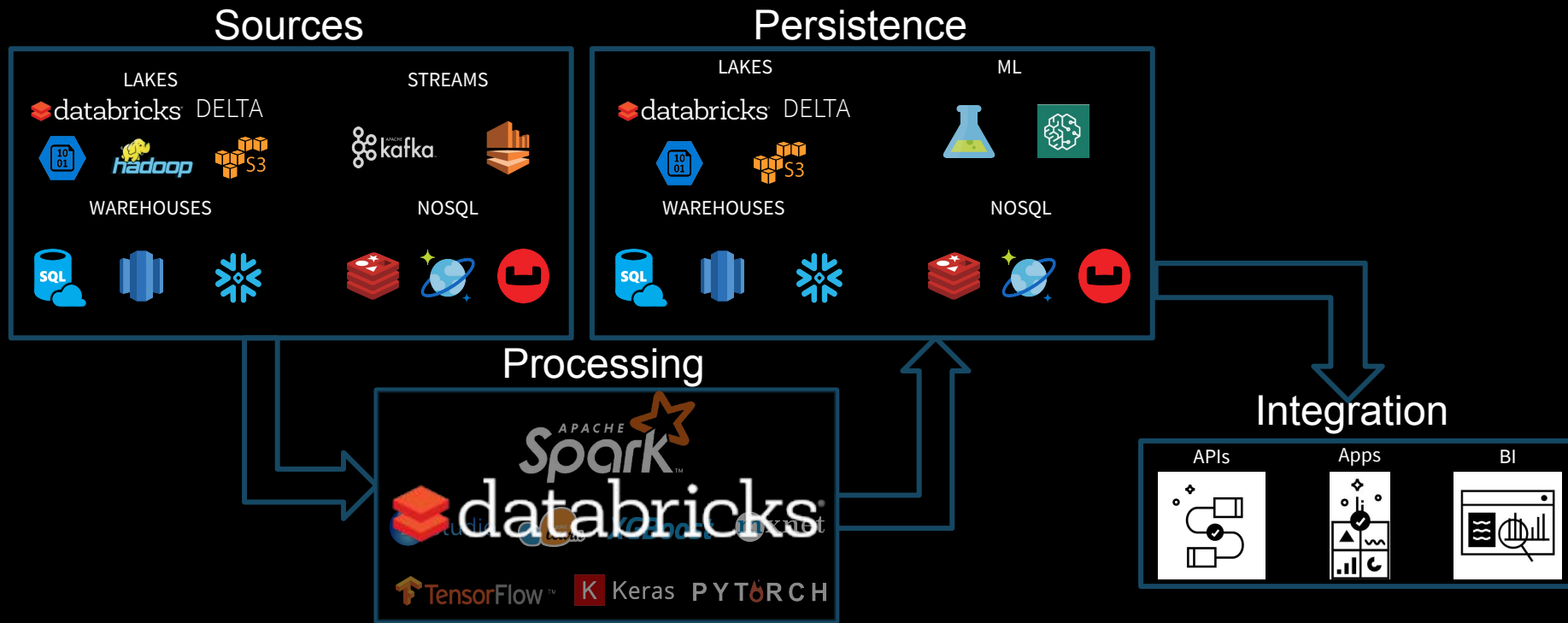
BI Reporting  
Dashboards

# Demo

# Review



# Solution





SPARK+AI  
SUMMIT 2019

08:13

T-Mobile LTE

99%

99%

08:13

T-Mobile LTE

08:13

T-Mobile LTE

99%

99%

08:13

T-Mobile LTE

<

Databricks + Snowflake: Catalyzing...

<

Rate this session

<

16:40-17:20

Rate this session

DETAILS

DISCUSSION

Please take time to provide feedback for this session.

the unparalleled  
distributed big d  
tool to the table,  
every technolog  
Kafka) to databa  
stores (e.g. S3) &

Key Takeaways:  
How Databricks  
Why they're so p  
How Databricks  
catalyze analytic

SPEAKERS



Garren S  
Databricks

FORMS

Rate this session  
Help make the Summit better

>

5 being the highest.

1

2

3

4

5

Add to my schedule?

Add now >

<

>

↺

...



SPARK+AI



SPARK+AI  
SUMMIT 2019

A word cloud of "Thank You" in various languages and scripts, including English, Spanish, French, German, Italian, Japanese, Chinese, Korean, Thai, and Hindi. The words are arranged in a circular pattern, with "Thank You" being the largest and most prominent. Other visible words include: GRACIAS, DANKE, MERCI, SALAMAT, TAKK, GRAZAS, MAHALO, ARIGATO, EUXCHARISTO, and many others. The background is dark with faint "iStock by Getty Images" watermarks.