




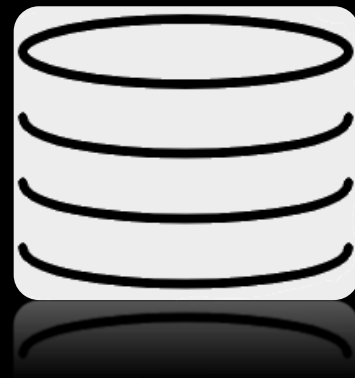
# Lambda architecture design using Azure Databricks for advanced analytics

# Agenda

- 1 **Introductions and use case**
- 2 **Who, what, and why of Spark and  databricks**
- 3 **Lambda Architecture: Design Principles**
- 4 **Demos: ATM Fraud, Azure Airlines**
- 5 **Wrap-Up: Solution Architectures**



Linked In	<a href="https://www.linkedin.com/in/lucas-feiock">https://www.linkedin.com/in/lucas-feiock</a>
Twitter	@sqlstack
Blog	<a href="https://sql-stack.com">https://sql-stack.com</a>
GitHub	<a href="https://github.com/sqlstack">https://github.com/sqlstack</a>
Email	<a href="mailto:sqlstack@outlook.com">sqlstack@outlook.com</a>
Work	<a href="mailto:Lucas.Feiock@kizan.com">Lucas.Feiock@kizan.com</a>





## VISION

Accelerate innovation by unifying data science, engineering and business

---

## PRODUCT

Unified Analytics Platform powered by Apache Spark

---

## WHO WE ARE

- Founded by the creators of Apache Spark
- Contributes **75%** of the open source code, **10x** more than any other company
- Trained **100k+** Spark users on the Databricks platform

# Who's Using it Today



Amgen adopted Databricks to migrate to a **Modern Big Data Platform on the Cloud**, enabling accelerated drug discovery that involves petabytes of data from multiple biotech field; using Databricks, researchers can use more data during drug development while easily collaborating across teams, leading to greater productivity and new insights



Shell adopted Databricks for **Predictive Supply Chain Analytics** to optimize their global inventory of spare parts – reducing legacy batch scoring jobs that took longer than 40 hours to less than 45 minutes – and save millions of dollars in operational efficiencies



Viacom adopted Databricks to increase customer retention and loyalty by analyzing petabytes of **Streaming Data** on video quality in real time to improve user experience; this is a machine learning use case on massive amounts of streaming data; the result was higher an increase in customer retention by 3.5x – 7.0x



HP adopted Databricks to support their **Consumer IoT** and **Machine Learning** use cases. HP's data scientists use sensor readings from > 20 M connected devices to create new marketing campaigns for HP Ink; using Databricks the data scientists eliminated the need for DevOps efforts and were able to focus on training more models.

# Databricks Customers Across Industries

## Financial Services



## Healthcare & Pharma



## Media & Entertainment



## Data & Analytics Services



## Technology



## Public Sector



## Retail & CPG



## Consumer Services




## Marketing & AdTech



## Energy & Industrial IoT



# Agenda

- 1 Introductions and use case
- 2 Who, what, and why of Spark and  databricks
- 3 Lambda Architecture: Design Principles
- 4 Demos: ATM Fraud, Azure Airlines
- 5 Wrap-Up: Solution Architectures



Open source data processing engine built around **speed**, **ease of use**, and **sophisticated analytics**

10-100x faster than  
MapReduce (Hadoop)

Easier to program  
Python, SQL, R, Java, Scala

APIs for SQL, machine learning,  
deep learning, streaming,  
graph

Storage agnostic, allowing  
federation & simple data  
access

More interactive  
data exploration

1000+ contributors  
across 250+ companies



# Apache Spark Ecosystem



Spark SQL +  
Data Frames

Streaming

MLlib  
*Machine Learning*

GraphX  
*Graph  
Computation*

Spark Core API

R

SQL

Python

Scala

Java

## Do you know Databricks?

Databricks makes building big data and AI applications simple, fast, easy, and collaborative with our **Unified Analytics Platform** powered by Apache Spark™ and built for cloud.

Data  
Engineering

Data  
Scientists

Business  
Analysts



Apache Spark™ provides a **single processing engine** for your big data and AI workloads including batch/ETL, streaming, SQL, graph, machine learning and deep learning workloads on petabytes of data on cloud data lakes. The result is higher productivity and faster time to insights and outcomes for your clients.

Databricks' founders are the original creators of Apache Spark™ and we have engineered our **platform as a service** for the cloud to improve elasticity, ease of use, performance, reliability, and cost-effectiveness compared to alternatives.

## Get more from Big Data & AI Projects

- Do More -  
Higher productivity  
without DevOps or cluster administration

- With More Data -  
At scale (volume and variety)  
with better cost/performance and elasticity

- More Reliably -  
Avoid project delays due to bugs/breaks  
and simplified data pipelines

- With a Lower TCO -  
Lower cloud and personnel costs  
and pay only for what you use

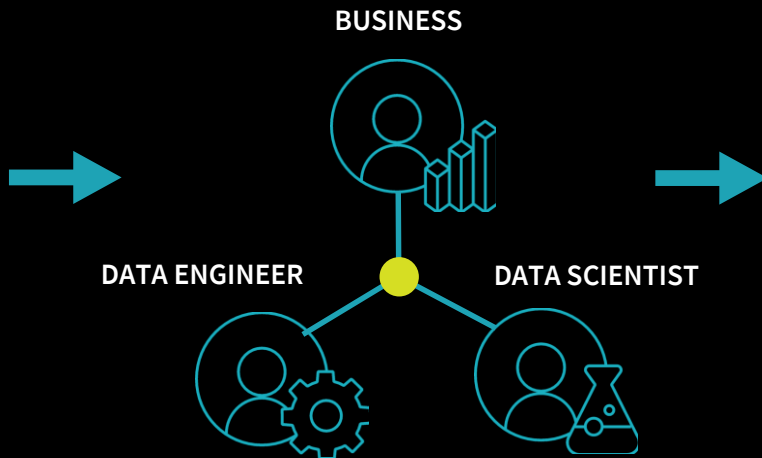
- More Secure -  
Satisfies industry security requirements  
(e.g. GDPR, HIPAA, and PCI)

- Enable AI & ML -  
Reach the potential of AI & ML use cases





# AI is a Game Changing Opportunity

## LOTS OF NEW DATA

Customer Data  
Click Streams  
Sensor data (IoT)  
Video/Speech  
...



## OPPORTUNITY

Fraud Detection   
Genome Sequencing   
Recommendation Engine   
Predictive Maintenance   
...

Machine Learning Requires  
**Collaborative Experimentation** on Big Data

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

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

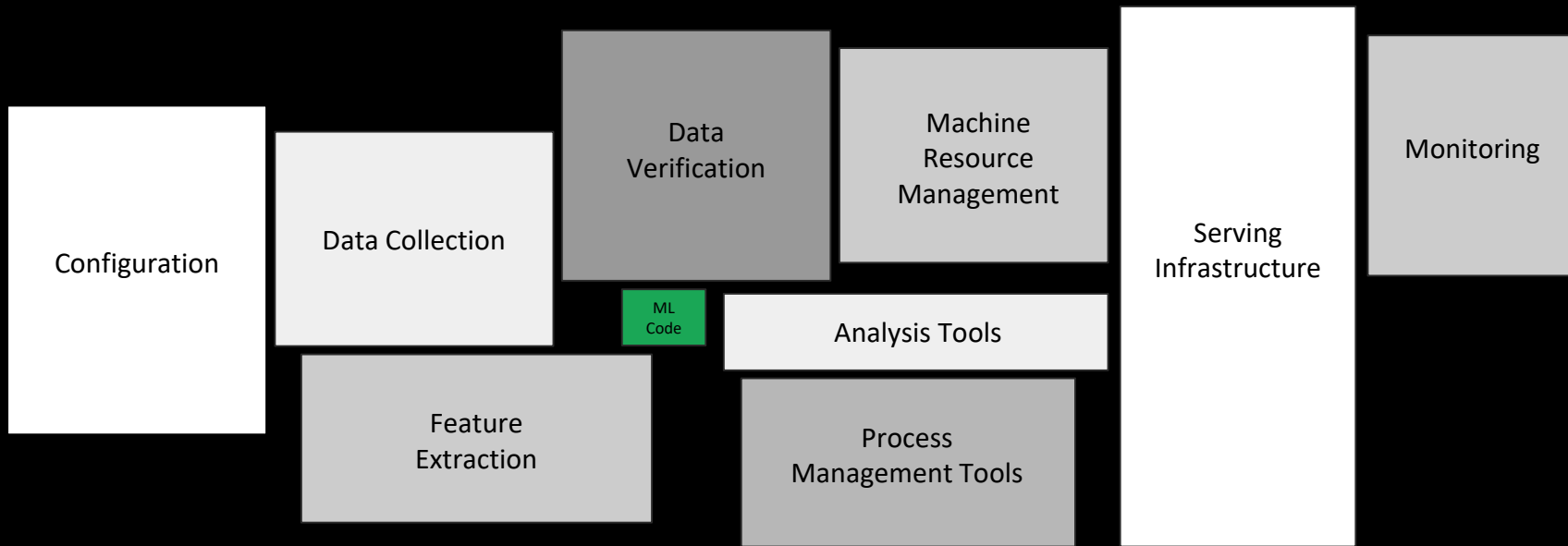


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

# Big Data and AI Complexity Slows Innovation

Lack of collaboration  
reduces data science  
productivity



Data pipelines produce  
stale data with poor  
performance

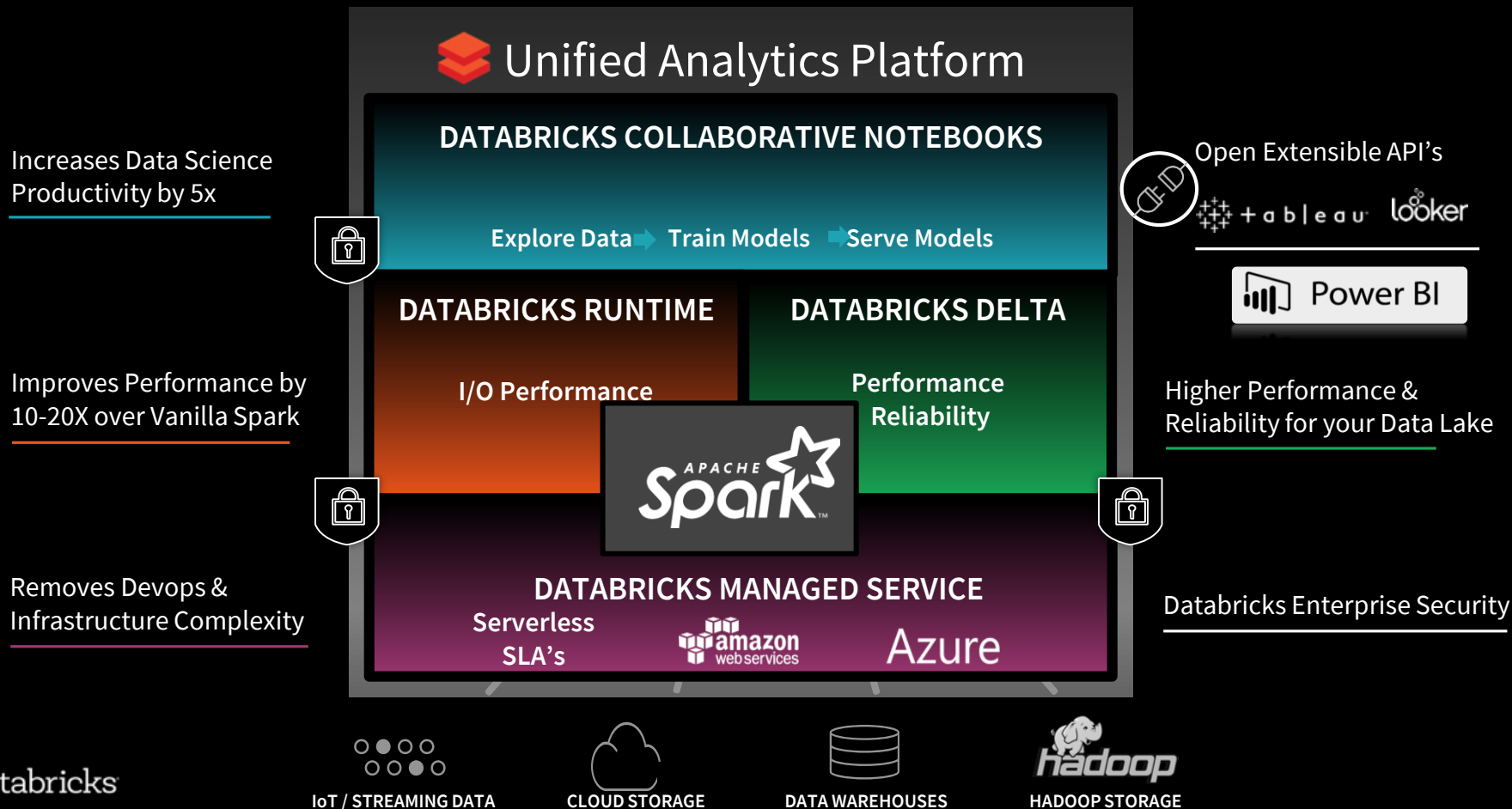


Multiple analytics engines  
increase complexity and  
slows innovation

Manual operations to  
manage big data  
infrastructure



# Accelerate Innovation with Databricks



# Agenda

- 1 Introductions and use case
- 2 Who, what, and why of Spark and  databricks
- 3 **Lambda Architecture: Design Principles**
- 4 **Demos: ATM Fraud, Azure Airlines**
- 5 **Wrap-Up: Solution Architectures**

# What is Lambda Architecture

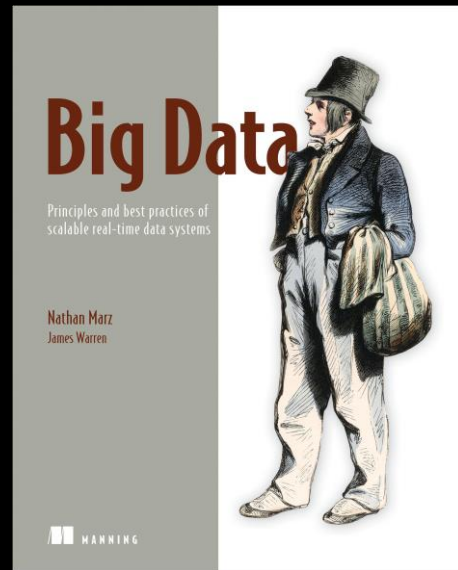
Data processing architecture

Generic, scalable, fault-tolerant

Low-latency reads, updates, ad-hoc queries

Nathan Marz - Apache Storm @ Twitter

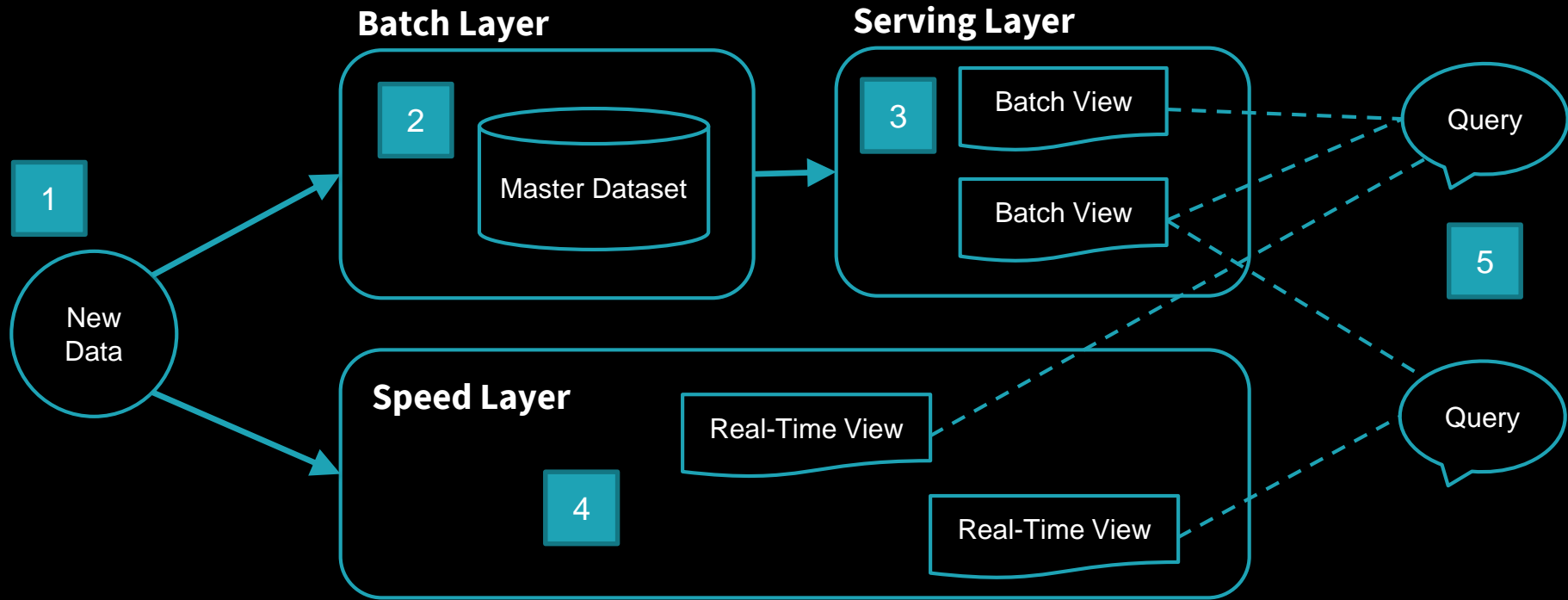
Principles and best practices of scalable real-time data systems





# Lambda Architecture

## Design Principles



# Lambda Architecture

## Real World Examples

Fault / Fraud detection

Manufacturing / Machine Logs / Robotics

Network / Security monitoring

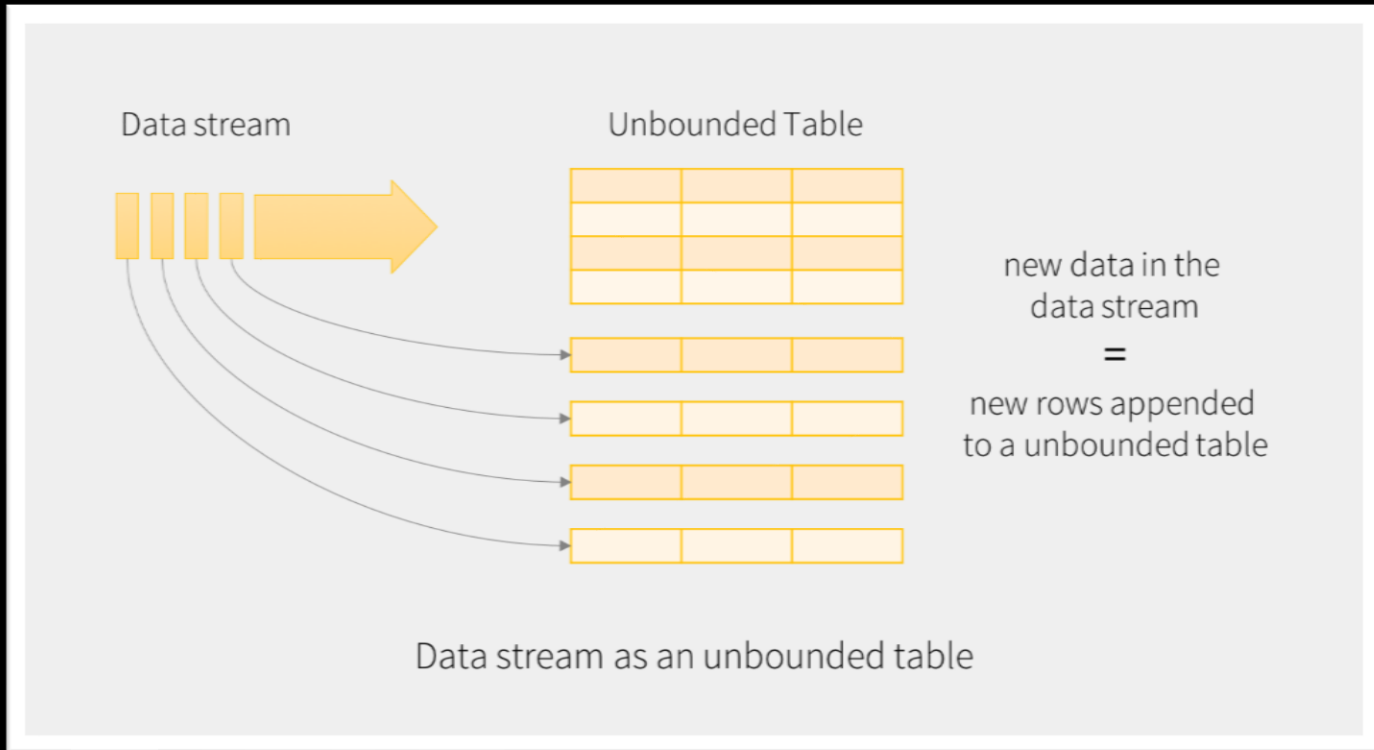
Digital Marketing / Websites Clicks / Telemetry

Portfolio Management / Algorithmic Trading

IOT / Connected devices

# Lambda Architecture

## Databricks Structured Streaming

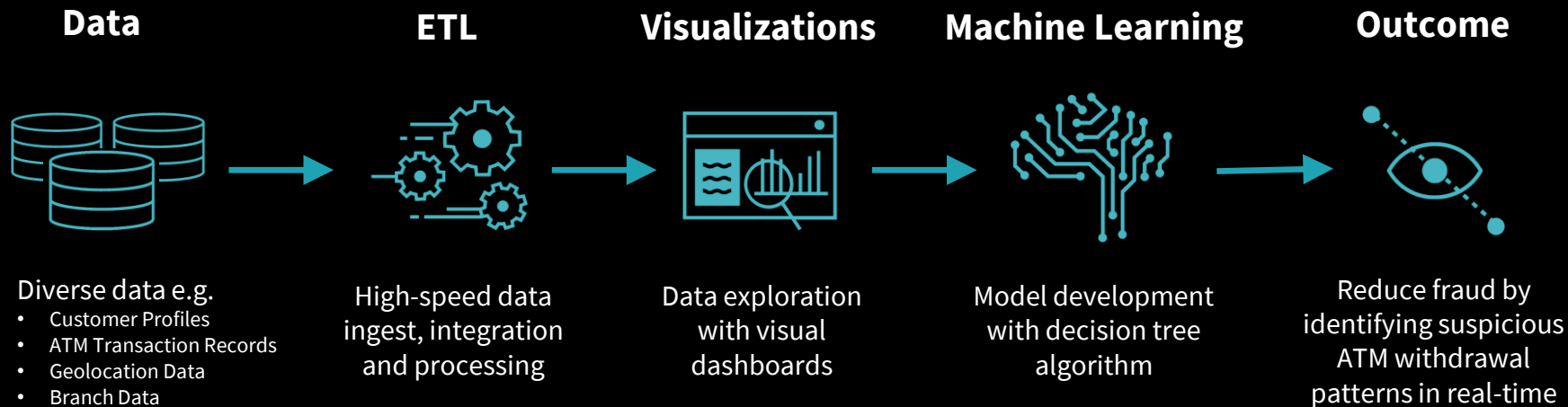


# Agenda

- 1 **Introductions and use case**
- 2 **Who, what, and why of Spark and  databricks**
- 3 **Lambda Architecture: Design Principles**
- 4 **Demos: ATM Fraud, Azure Airlines**
- 5 **Wrap-Up: Solution Architectures**

# ATM Fraud Detection

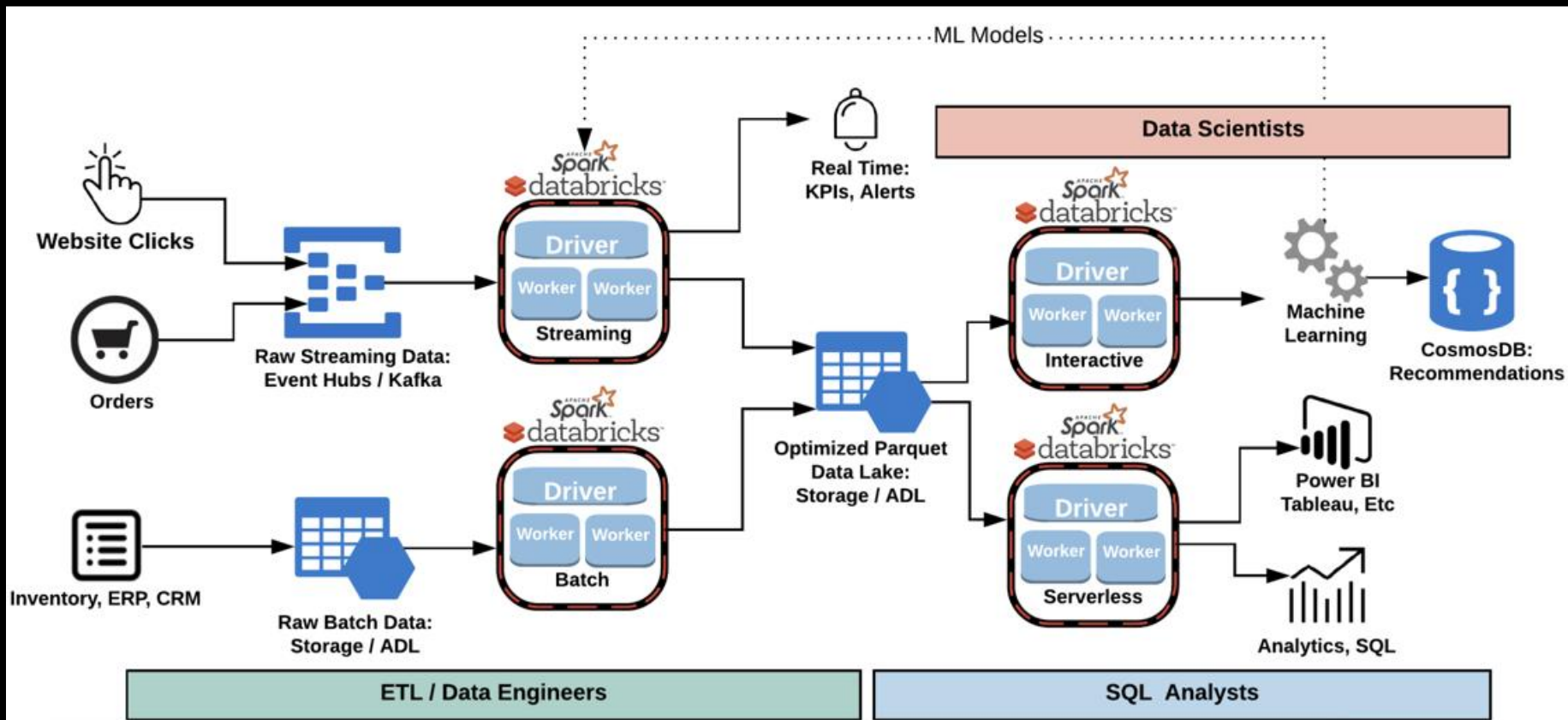
## Financial Services



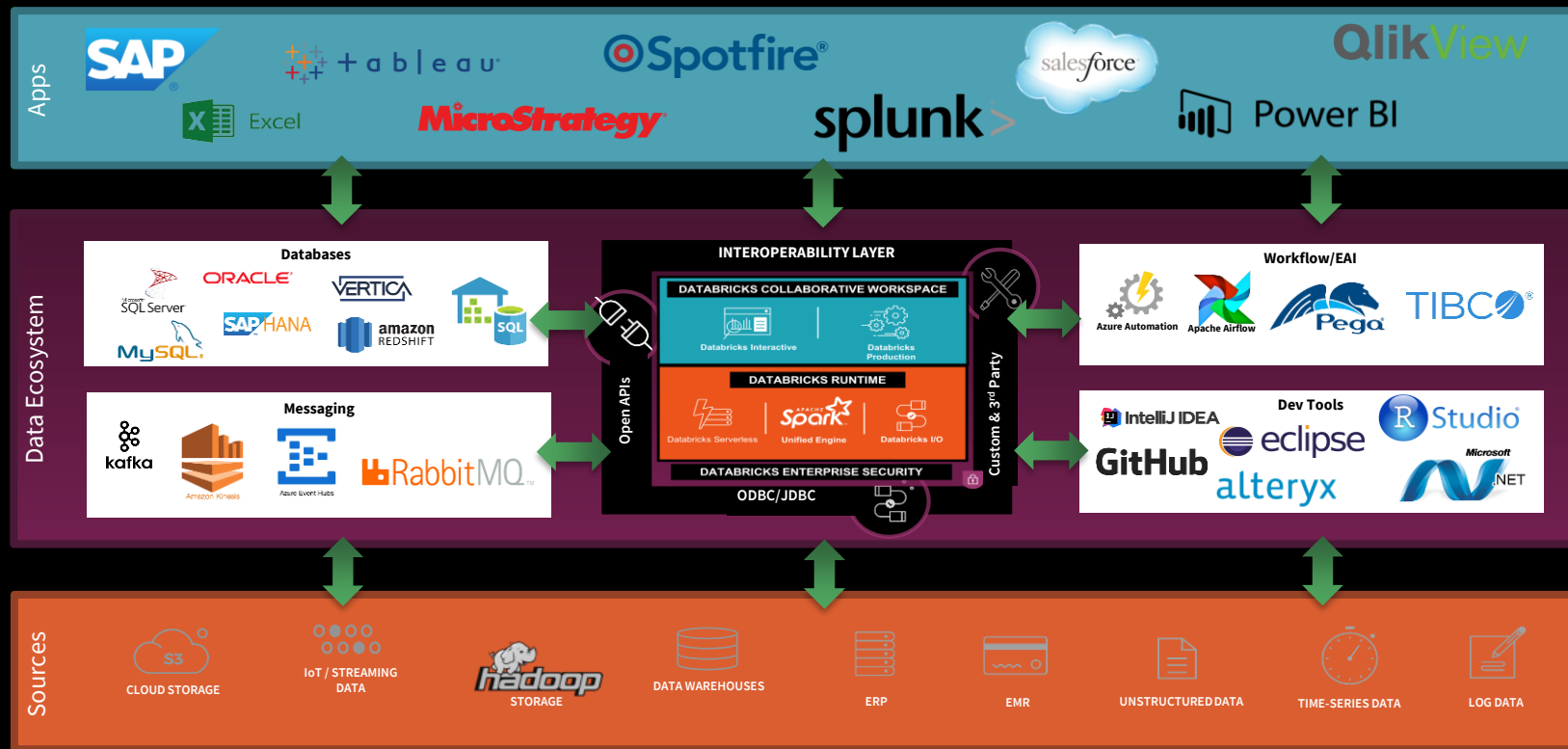
# Agenda

- 1 **Introductions and use case**
- 2 **Who, what, and why of Spark and  databricks**
- 3 **Lambda Architecture: Design Principles**
- 4 **Demos: ATM Fraud, Azure Airlines**
- 5 **Wrap-Up: Solution Architectures**

# Databricks 101 Architecture



# Databricks Ecosystem Reference Architecture





# Azure Airlines – Flight Demo



Flight  
Data  
Simulator



Event  
Hubs



Azure  
Databricks  
Structured  
Streaming



Power BI  
API  
Mapping



Thank you ! Questions?

# Main Links

<https://azure.microsoft.com/en-us/services/databricks/>

<https://spark.apache.org/>

<https://databricks.com/>

<https://databricks.com/spark/about>

<http://lambda-architecture.net/>

<http://spark.apache.org/powered-by.html>

<https://www.businesswire.com/news/home/20190131005243/en/Databricks-Named-Visionary-Consecutive-Year-Gartner-Magic>

# Databricks Community Edition

<https://databricks.com/product/faq/community-edition>

Free version of the cloud based platform

Hosted on Amazon Web Services

Uses a micro-cluster of one driver with 6 GB of memory

Contains training resources

Great way to get started learning about Apache Spark

# Blogs

<https://www.desertislesql.com/> - Ginger Grant

<https://databricks.com/blog>

<https://databricks.com/blog/category/engineering>

<https://databricks.com/blog/category/company>

<https://curatedsql.com/?s=spark>

# Videos

<https://databricks.com/resources/type/videos>

<https://www.youtube.com/user/TheApacheSpark/feed>

<https://www.youtube.com/channel/UC3q8O3Bh2Le8Rj1-Q-UUbA>

<https://databricks.com/resources/type/product-videos>

<https://sparkhub.databricks.com/videos/>

<https://www.youtube.com/watch?v=TJcEP6AX02U>

<https://databricks.com/sparkaisummit/north-america/sessions>

<https://databricks.com/session/jaws-data-warehouse-with-spark-sql>

<https://databricks.com/azure-databricks-demo>

<https://databricks.com/resources/type/customer-stories>

# Projects and Papers

<https://cs.stanford.edu/~matei/>

<https://amplab.cs.berkeley.edu/tag/spark/>

<https://spark.apache.org/research.html>

<https://databricks.com/resources/type/research-papers>

# Training and Certification

<https://databricks.com/training>

<https://docs.databricks.com/>

<https://legacy.gitbook.com/@jaceklaskowski>

<https://www.edureka.co/blog/spark-tutorial/>

<https://github.com/midomsft/DatabricksHOL>

<http://spark.apache.org/docs/latest/building-spark.html>

<https://www.coursera.org/specializations/big-data>

<https://www.coursera.org/specializations/scala>

<https://databricks.com/training/certified-spark-developer>



# Other projects

<https://www.microsoft.com/en-us/research/project/urban-computing/>

<https://www.microsoft.com/en-us/research/project/dryadling/>

<https://github.com/Microsoft/SmartHotel360-Backend>

<https://eng.uber.com/uber-big-data-platform/>

<https://github.com/Microsoft/Mobius>

# Dayton Gray Sort Record - 2014

<https://databricks.com/blog/2014/11/05/spark-officially-sets-a-new-record-in-large-scale-sorting.html>

<https://spark.apache.org/news/spark-wins-daytona-gray-sort-100tb-benchmark.html>

<http://sortbenchmark.org/>

<http://sortbenchmark.org/Spark2014.pdf>

# References and posts - 1

<https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/stream-processing-databricks>

<https://databricks.com/glossary/what-are-continuous-applications>

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

<https://docs.databricks.com/spark/latest/structured-streaming/index.html>

<https://databricks.com/blog/2016/07/28/continuous-applications-evolving-streaming-in-apache-spark-2-0.html>

# References and posts - 2

<https://databricks.com/blog/2018/05/03/benchmarking-apache-spark-on-a-single-node-machine.html>

<http://datastrophic.io/core-concepts-architecture-and-internals-of-apache-spark/>

<https://lenadroid.github.io/posts/connecting-spark-and-eventhubs.html>

<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/spark-data-exploration-modeling>

# References and posts - 3

<https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/spark-overview>

[https://github.com/mspnp/reference-architectures/tree/master/data/streaming\\_azuredatabricks](https://github.com/mspnp/reference-architectures/tree/master/data/streaming_azuredatabricks)

<https://azure.microsoft.com/en-us/blog/azure-databricks-industry-leading-analytics-platform-powered-by-apache-spark/>

<https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>

<https://towardsdatascience.com/sql-at-scale-with-apache-spark-sql-and-dataframes-concepts-architecture-and-examples-c567853a702f>

# References and posts - 4

<https://docs.microsoft.com/en-us/sql/big-data-cluster/big-data-cluster-overview?view=sqlallproducts-allversions>

<https://github.com/giulianorapoz/DatabricksStreamingPowerBI>

<https://databricks.com/blog/2017/01/19/real-time-streaming-etl-structured-streaming-apache-spark-2-1.html>

<https://databricks.com/blog/2016/05/23/apache-spark-as-a-compiler-joining-a-billion-rows-per-second-on-a-laptop.html>