

MATLAB EXPO

2021

云端数据工作流程：科学家和工程师的新晋之路

耿煜, 亚马逊云科技大中华区高性能计算产品经理
刘海伟, MathWorks 中国高级应用工程师



Agenda

	Time	Presenter
Introduction (why we are here)	5 minutes	刘海伟
AWS –工程数据上云之旅	6 minutes	耿煜
Databricks – 帮助所有研究者更好的使用云上数据	6 minutes	耿煜
MathWorks – 分析云端数据并在生产环境中运行分析	5 minutes	刘海伟
Q&A	3 minutes	

使用瞬态数据提升预测性维护性能

挑战

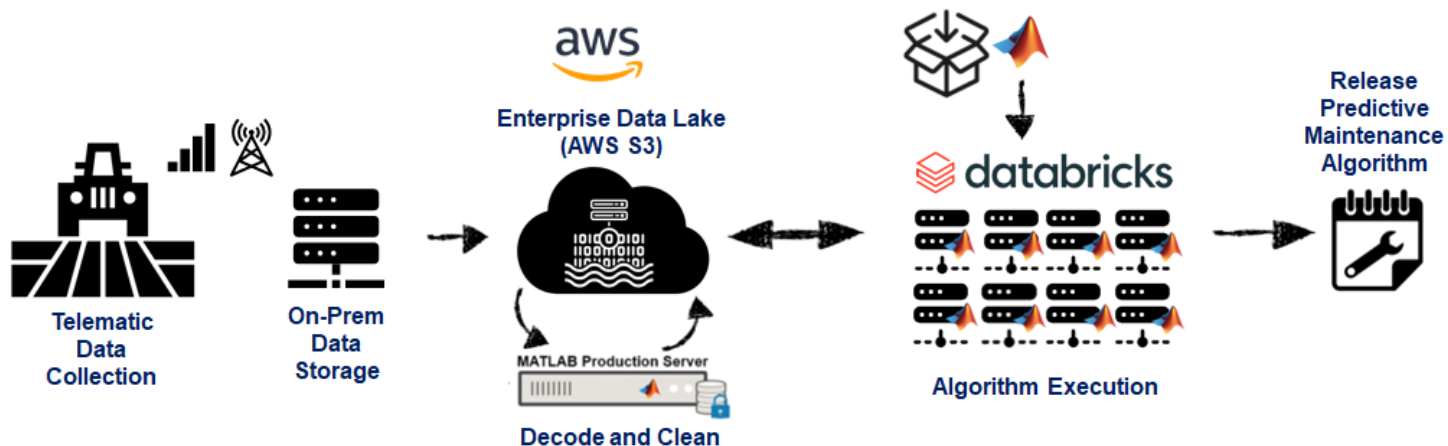
- Embedded controllers have limited computing power
- Engineers don't have easy access to data in Enterprise Data Lake

解决方案

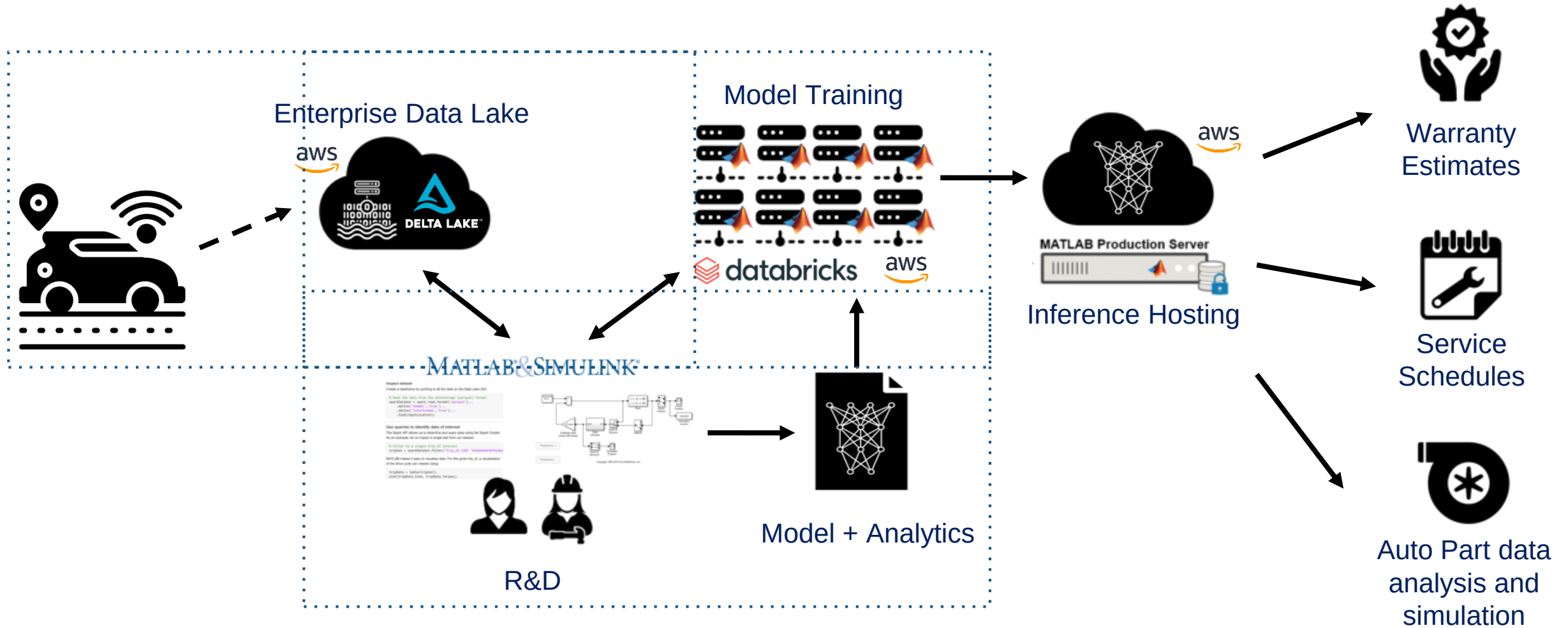
- Integrate MATLAB with cloud-based data storage and compute environment to access data and run analytics

结果

- Speed up algorithm development and deployment
- Domain experts can focus on developing algorithms



数据流架构





工程数据上云之旅

耿煜



MATLAB EXPO 2021

摆脱架构限制，加速科学计算

Amazon CloudWatch

Applications and Services

Automation and orchestration



AWS Batch



AWS ParallelCluster

NICE EnginFrame

Storage



Amazon EBS



Amazon FSx for
Lustre



Amazon EFS



Amazon S3

Compute



Amazon EC2 instances



Amazon EC2 Spot



AWS Auto Scaling

Visualization



NICE DCV



Amazon
AppStream 2.0

Networking

Enhanced networking

Placement groups

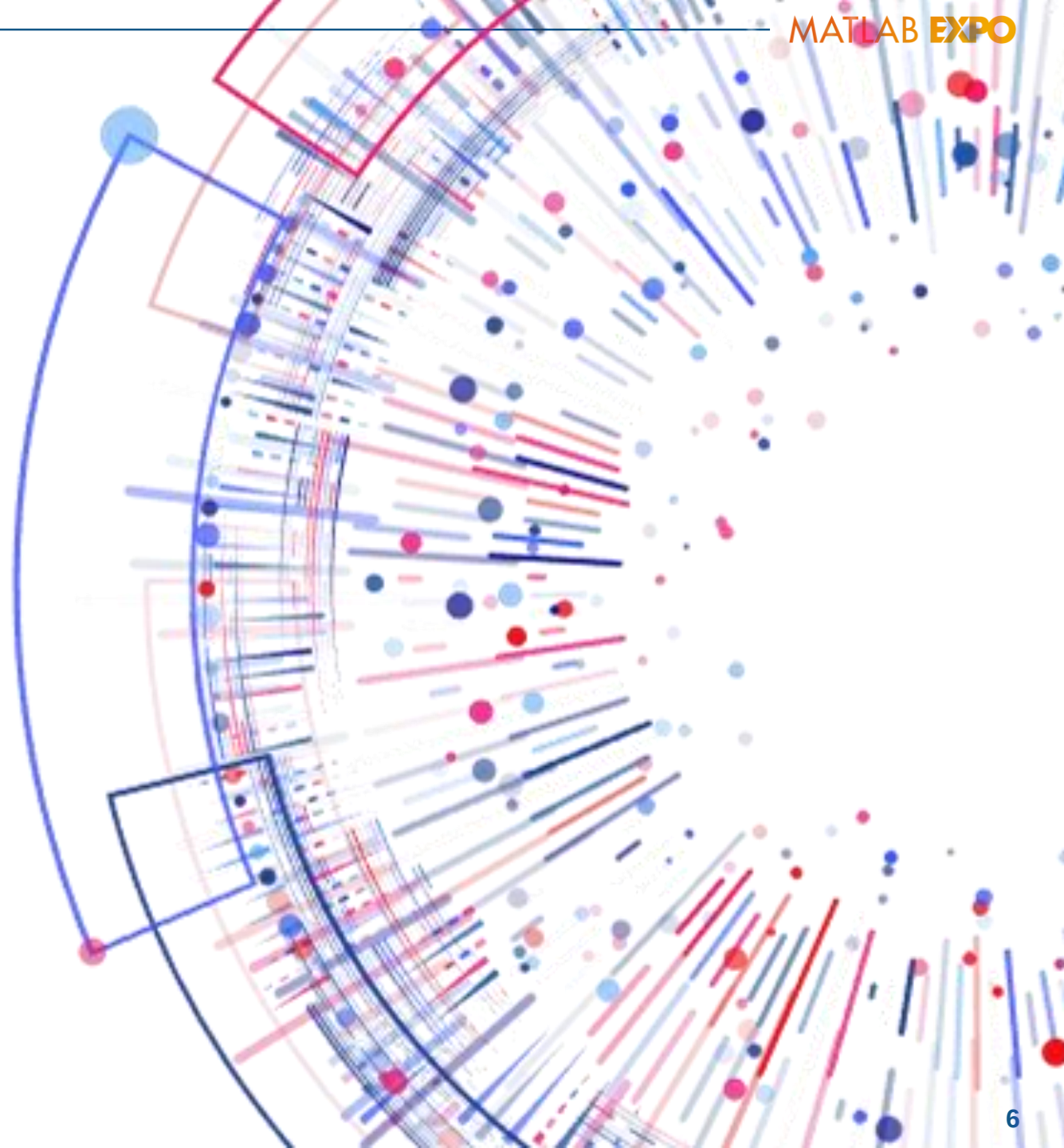
Elastic Fabric Adapter

Amazon IAM (Identity and Access Management)

AWS Budgets

现代架构的关键驱动力 – 数据

- Data volume
- Data variety
- Use case complexity
- Data access
- Breaking down silos



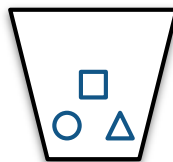
从研发的数据中发掘更多的价值



Growing
exponentially



From new
sources



Increasingly
diverse



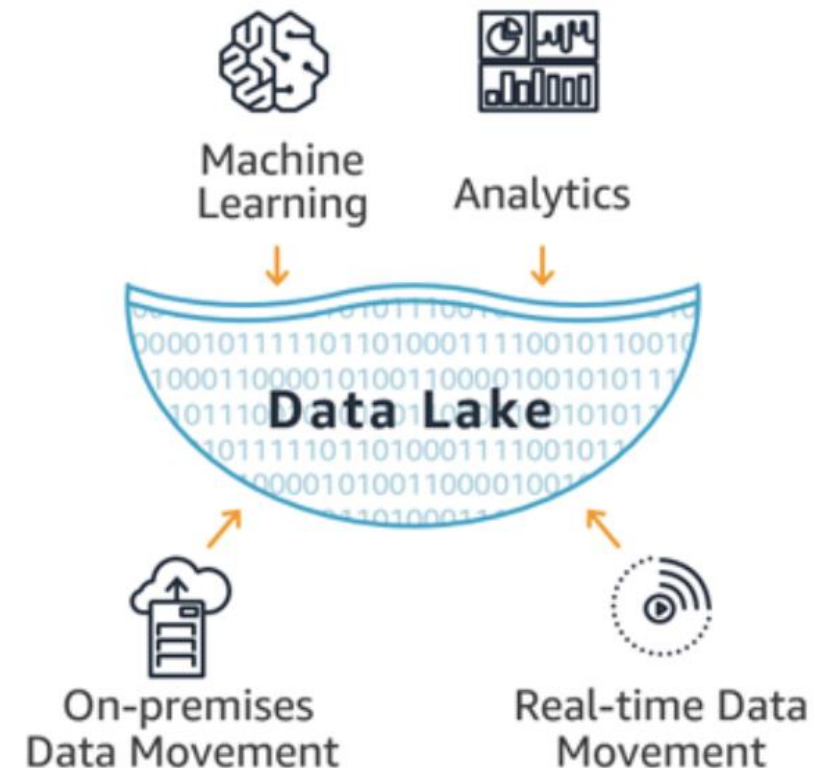
Used by
many people



Analyzed by
many applications

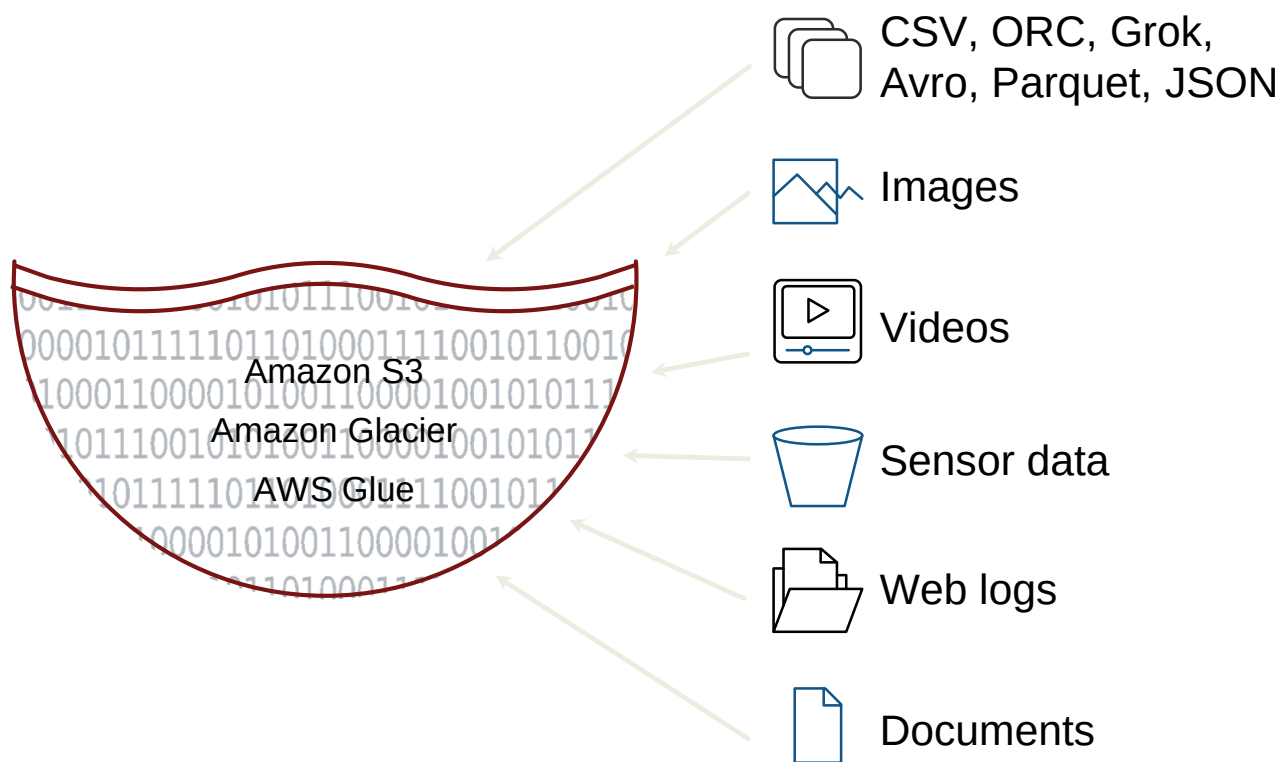
数据湖:让存储无界限

A data lake is an architectural approach that allows you to store massive amounts of data into a central repository, so it's readily available to be categorized, processed, analyzed, and consumed by diverse groups within an organization.



结合亚马逊S3对象存储的数据湖赋能科学计算

安全，高扩展性，持续可靠的，低延迟数据访问



Store data in the format you want

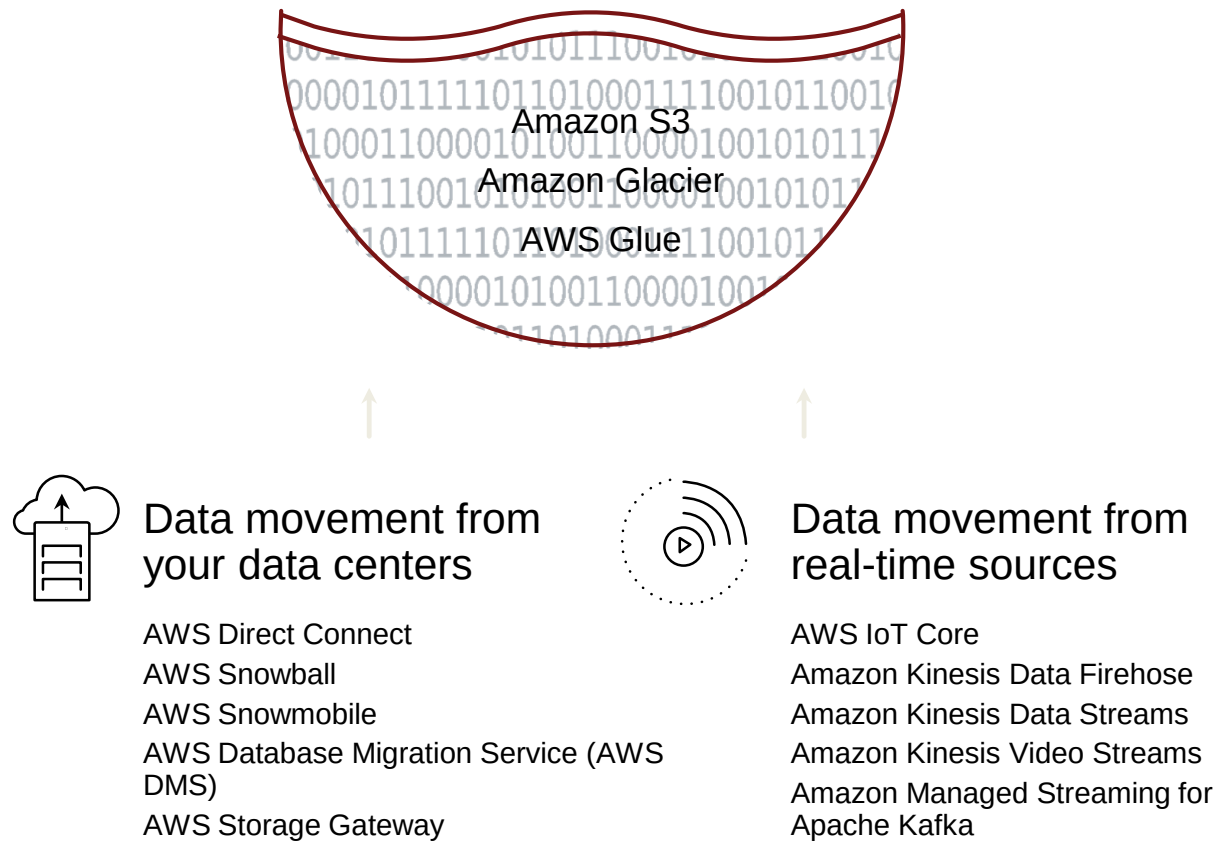
Built for 11 nines of durability

Three different forms of encryption

Run analytics and ML on data lake without data movement

Classify, report, and visualize data usage trends

如何将您的数据平滑高效的迁移上云



Data movement from on-premises datacenters

- Dedicated network connection
- Secure appliances
- Ruggedized shipping container
- Database migration
- Gateway that lets applications write to the cloud

Data movement from real-time sources

- Connect devices to AWS
- Real-time data streams
- Real-time video streams



帮助所有研究者更好的使用云上数据

耿煜

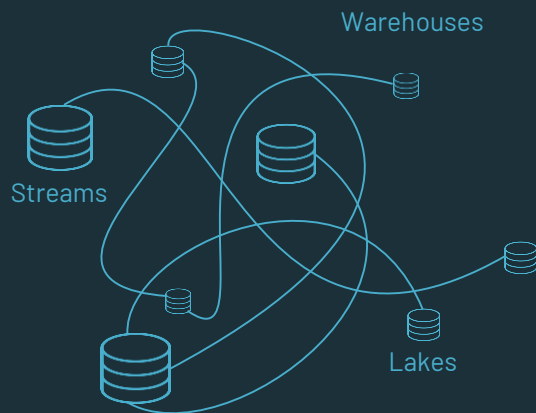


MATLAB EXPO 2021

解锁数据的商业价值面临四大挑战

1

Data is messy,
siload and slow



2

ML is hard,
Production is
harder



Data Engineers
in IT

Data Scientists
in Business



3

BI is limited to a
fraction of data

```
110001100011000100010001
000010111000100101010000
11100101010011111001110011
1010100011100110001100011
00010001000100001011100
01001010100001111001010
```

4

Lack of enterprise
readiness



Fragmented
security

Poor
reliability

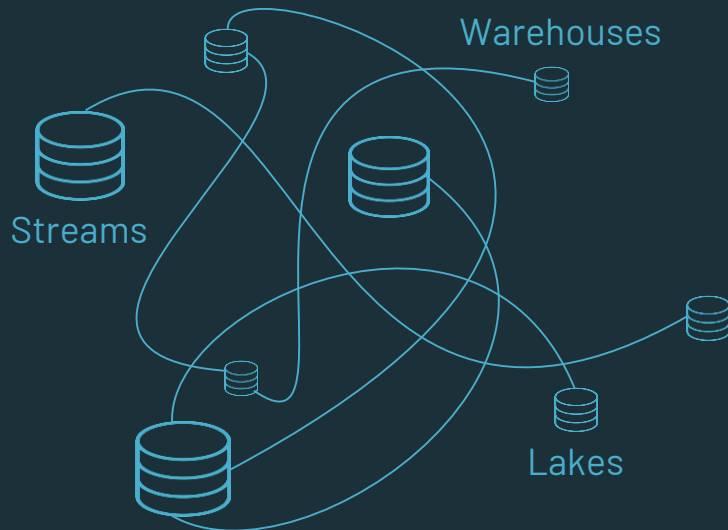


Disjointed
governance

让你的数据适合分析和机器学习

1

Data is messy,
siload 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

aws

Delta
Engine
with
**APACHE
Spark™**
API's

MATLAB &
Simulink

BI
Reporting

Machine
Learning

统一的数据和跨全生命周期的机器学习

2

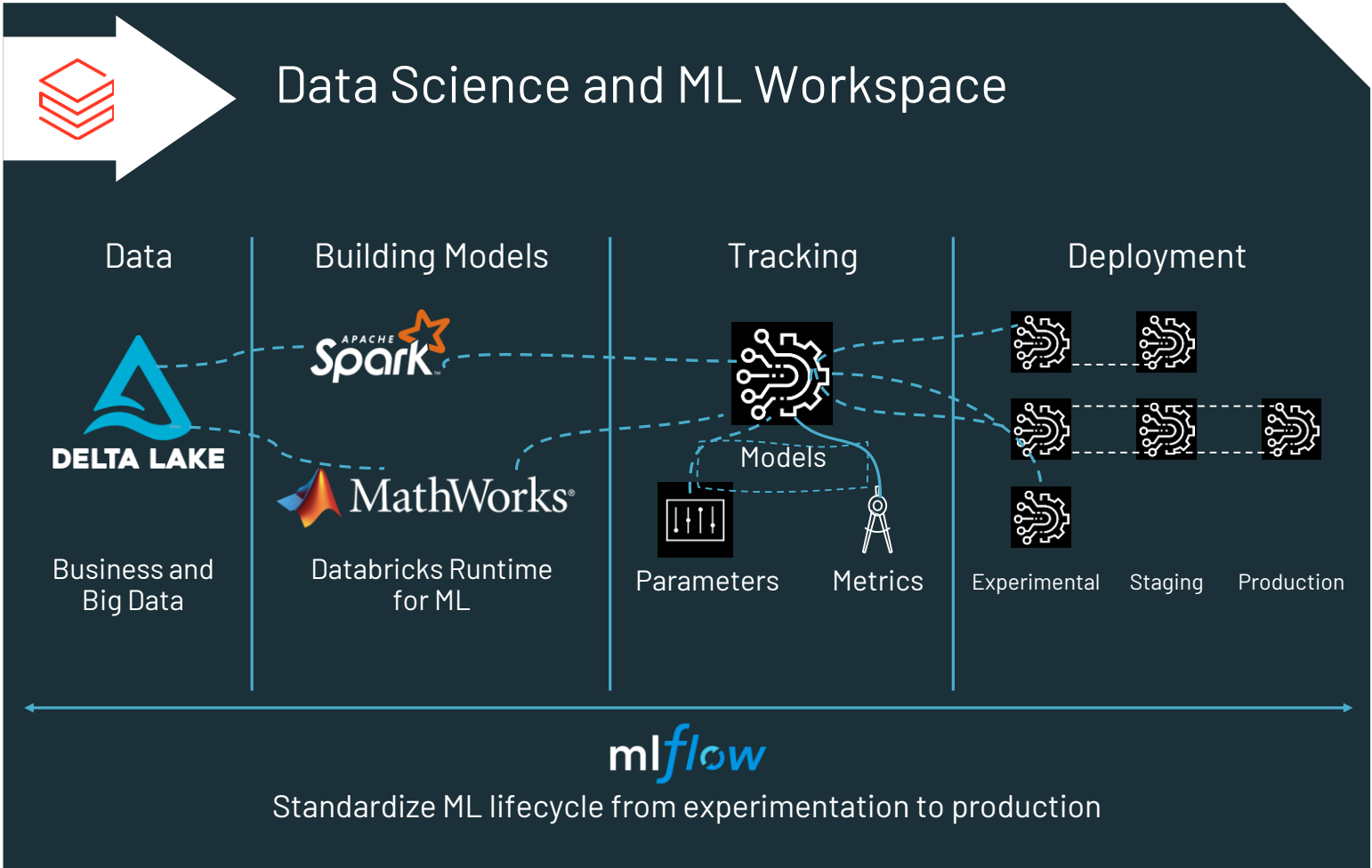
ML is hard,
production is
harder



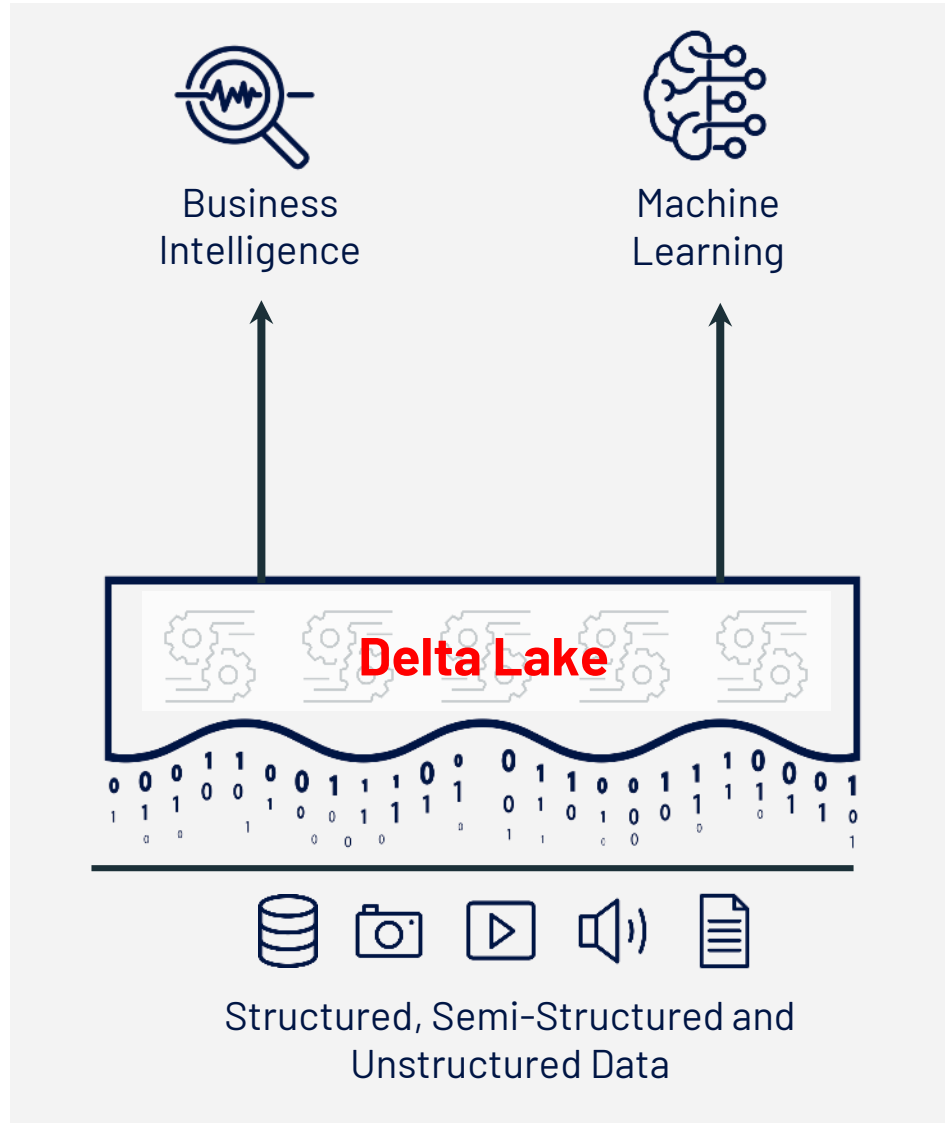
Data Engineers
in IT

Data Scientists
in Business



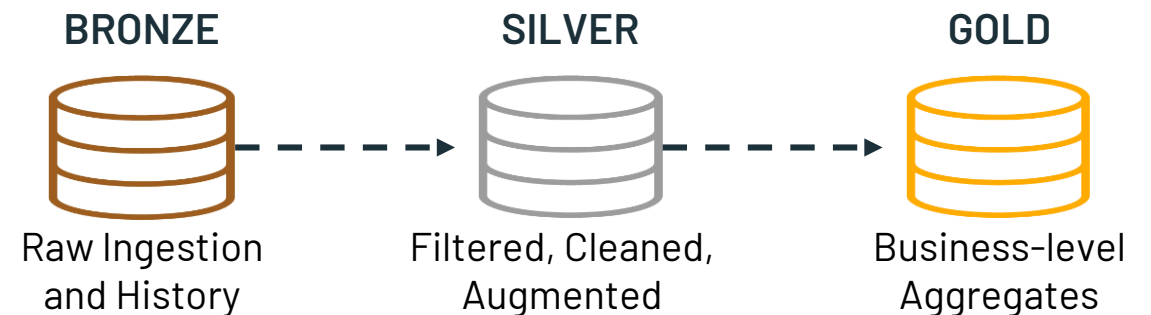


数据工程工作负载:为分析准备好数据

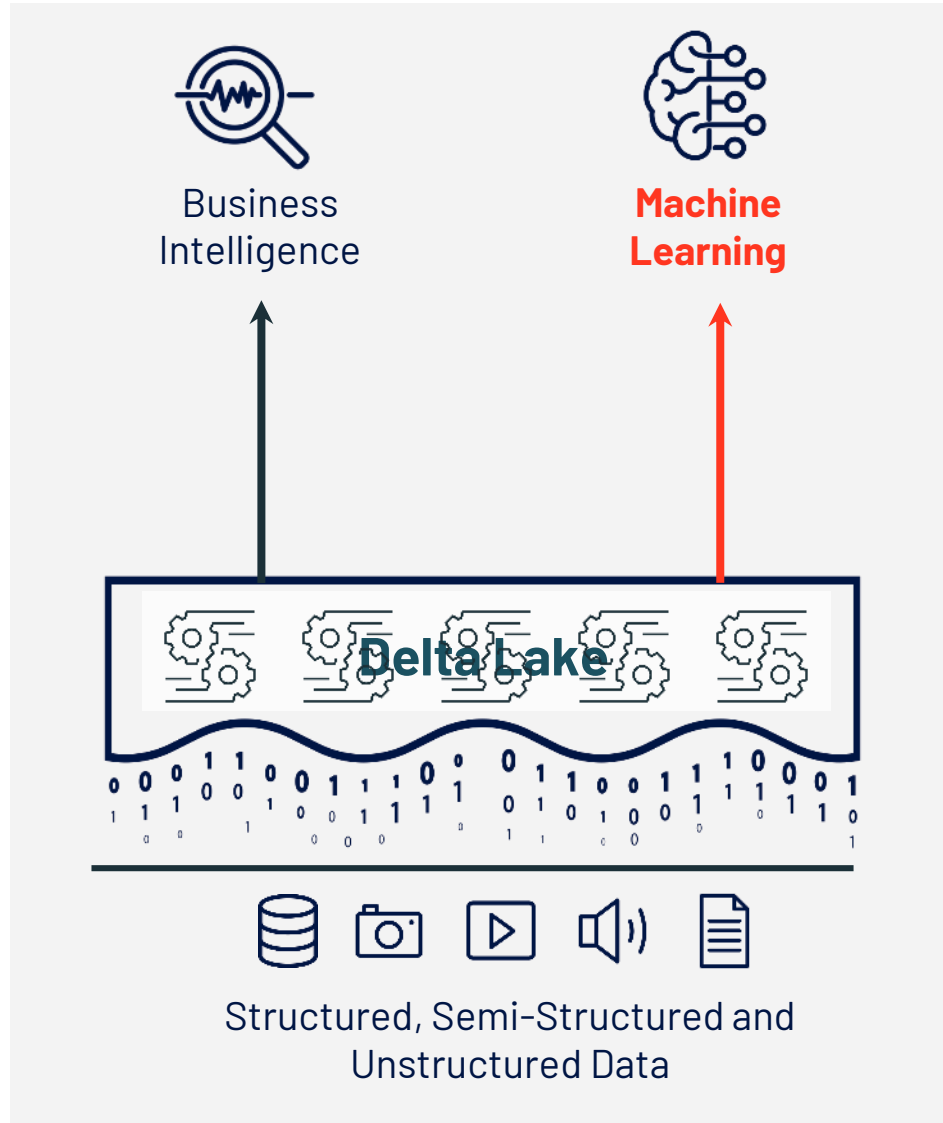


One Data Foundation for BI, Data Science & ML

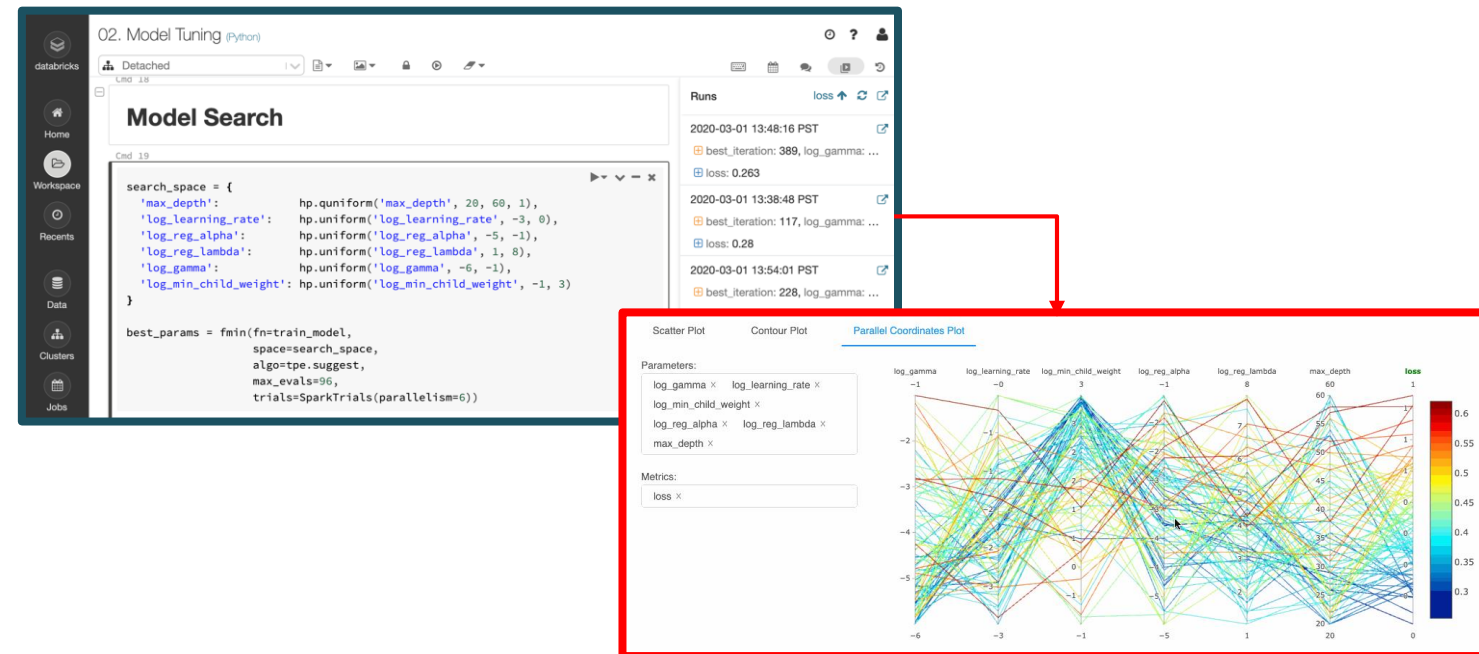
- Adds **reliability, performance, governance,** and **quality** to existing data lakes
- Based on **open data format** (Parquet)
- Simplifies data engineering with a **curated data lake** approach



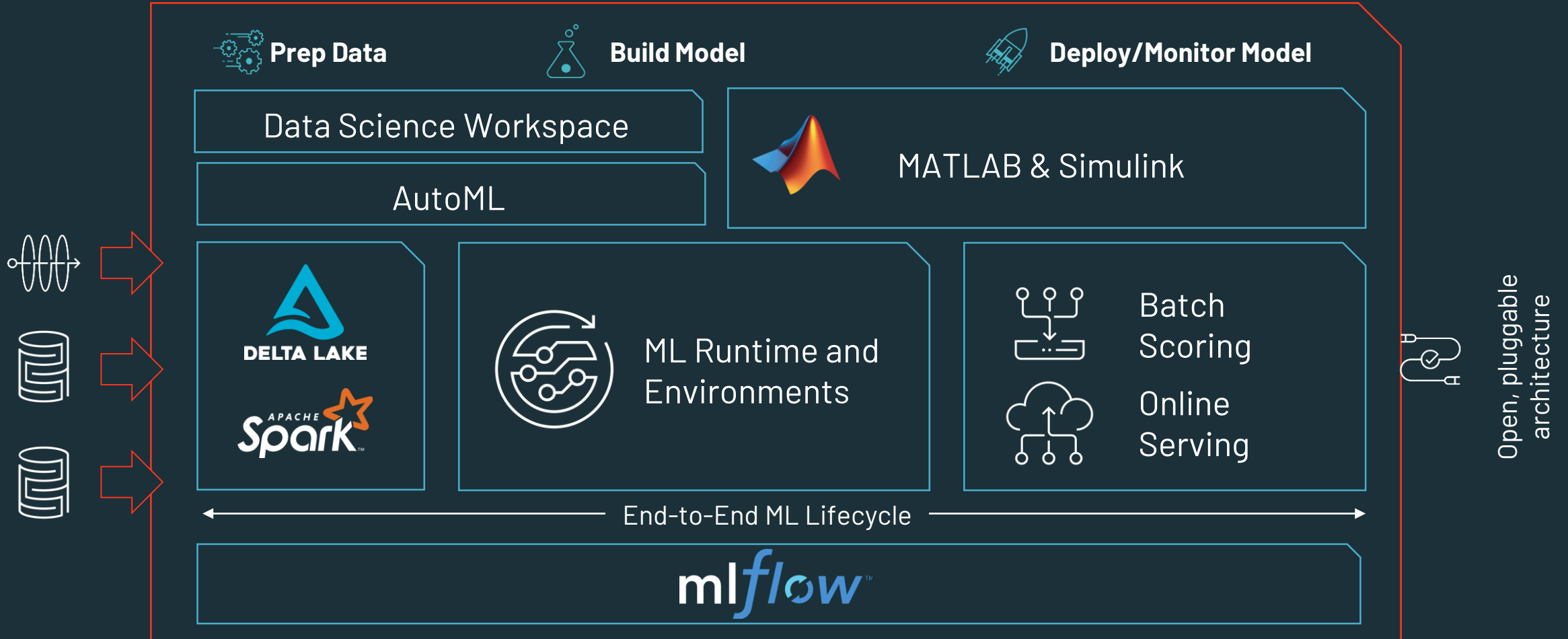
Machine Learning workloads on the curated Delta Lake



- Model registry, reproducibility, productionization
- Leverages Delta Lake for reproducibility



End-to-End Data Science and ML on databricks





分析云端数据并在生产环境中运行分析

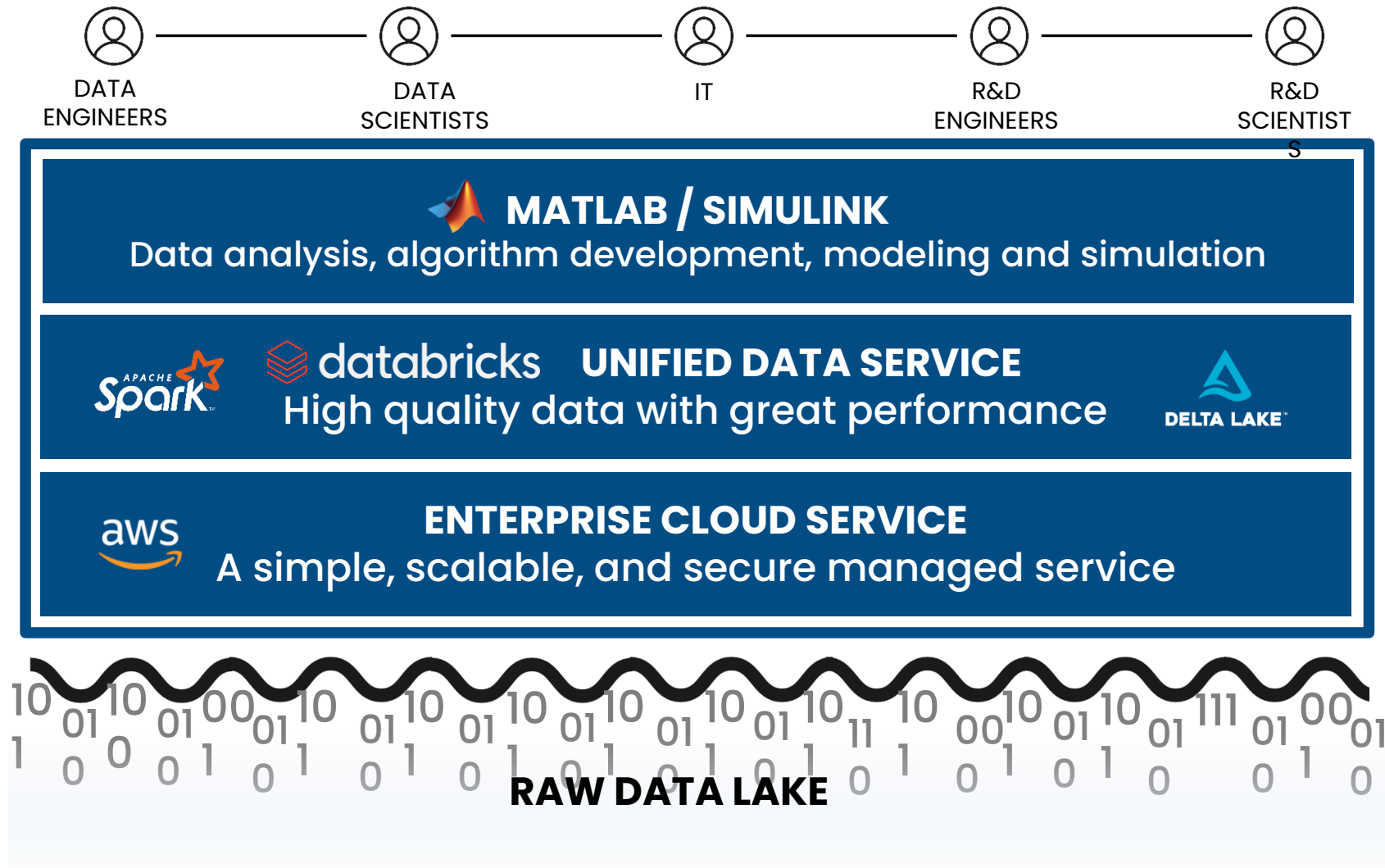
刘海伟



MATLAB EXPO 2021

平台集成

“The entire company on one data and compute platform in the cloud”



使用瞬态数据提升预测性维护性能

挑战

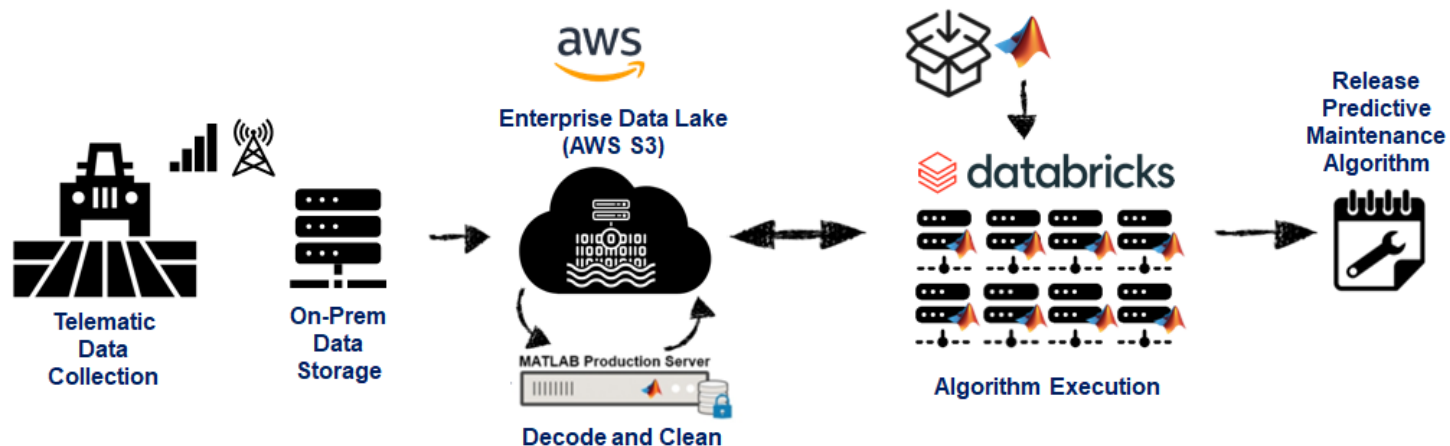
- Embedded controllers have limited computing power
- Engineers don't have easy access to data in Enterprise Data Lake

解决方案

- Integrate MATLAB with cloud-based data storage and compute environment to access data and run analytics

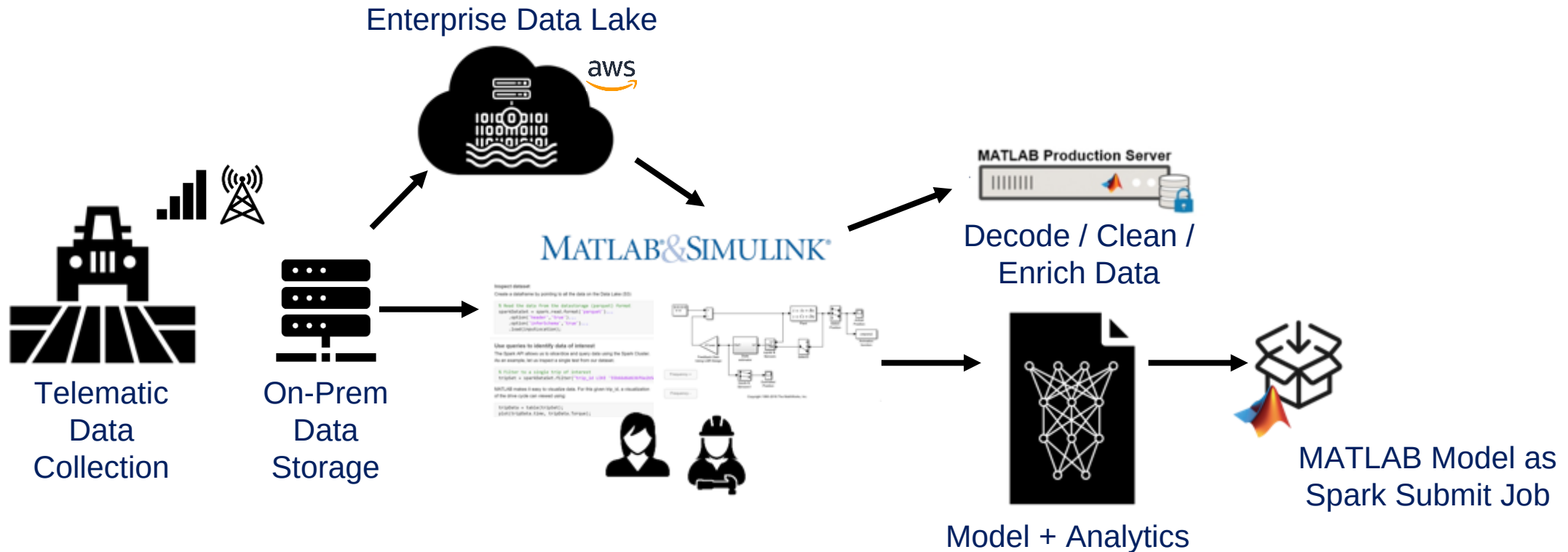
结果

- Speed up algorithm development and deployment
- Domain experts can focus on developing algorithms

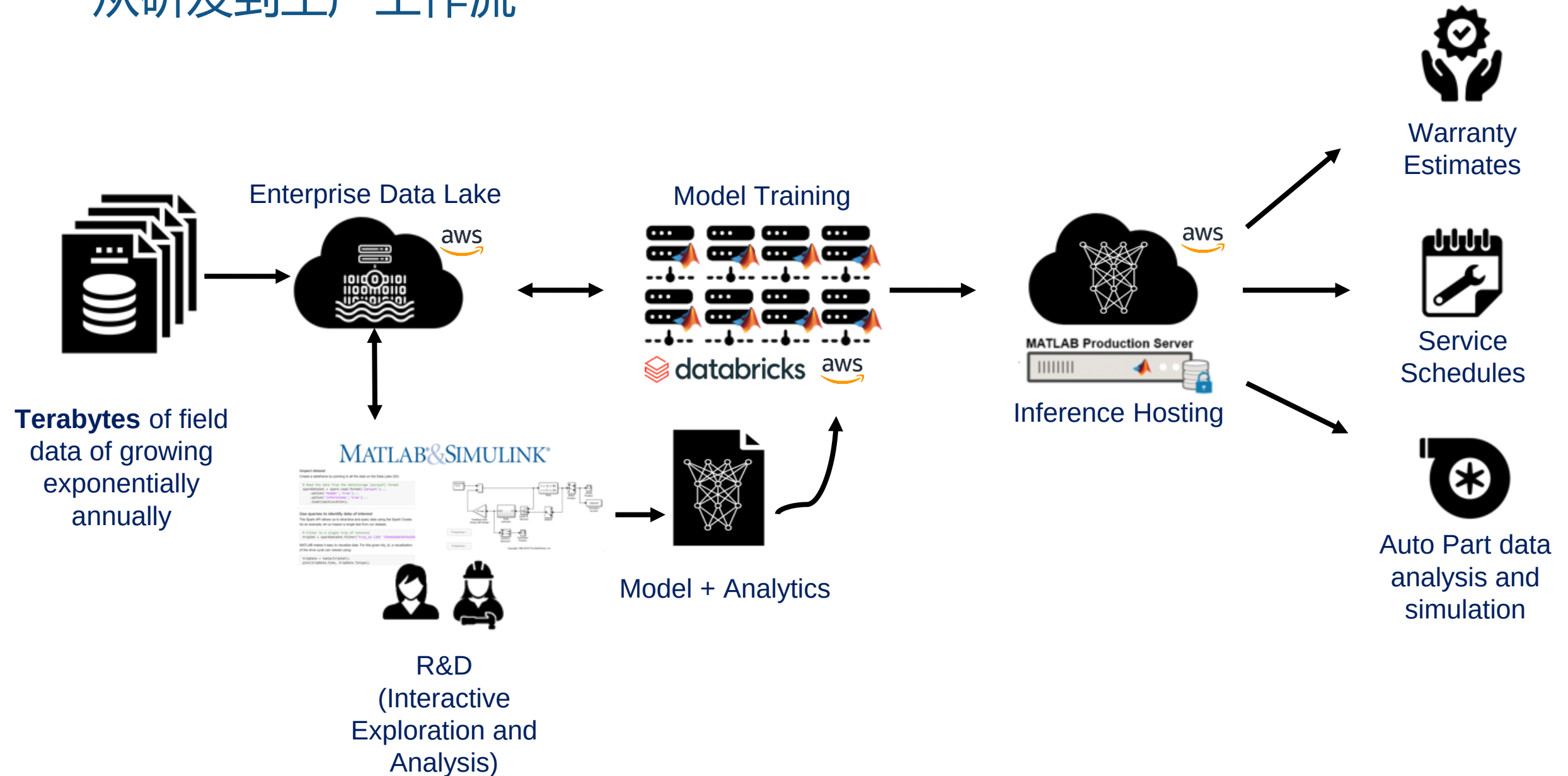


数据工程 workflows

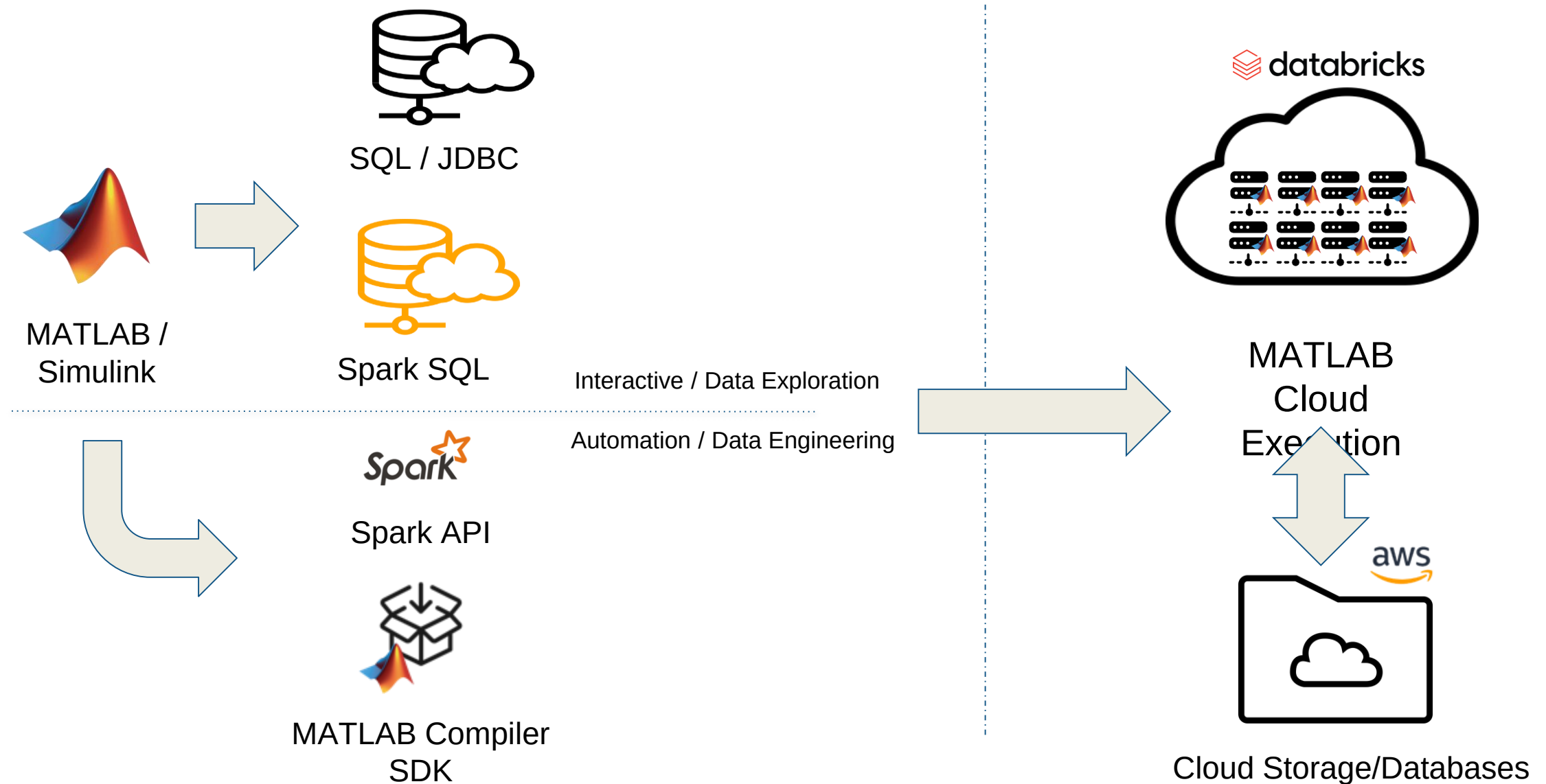
Helping engineers with data preparation algorithms



从研发到生产 workflows



Databricks workflow



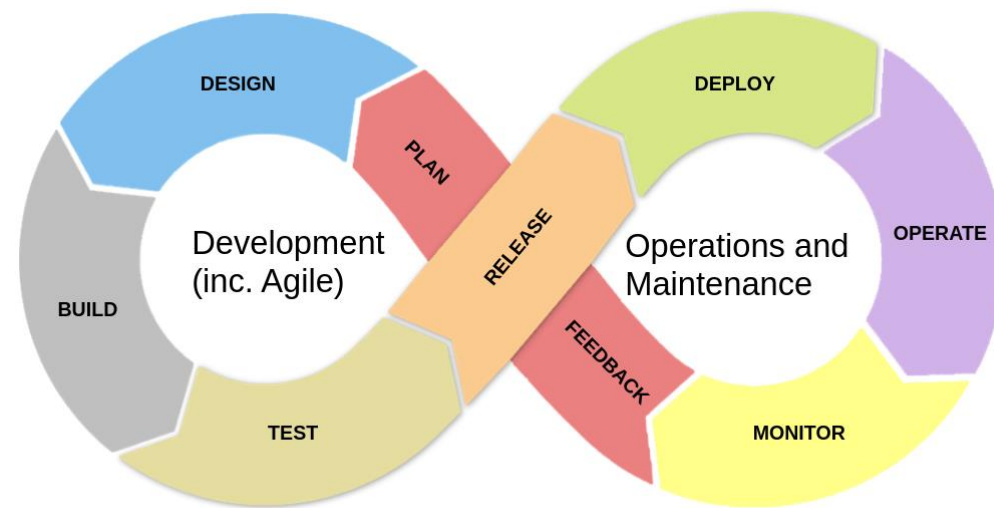
模型开发运维

针对工程师优化的 workflow 来满足自身运维需求

- Take machine learning and AI models from interactive development to production
- Build, Test and Deploy via automated CI/CD pipelines following best practices for security and scaling
- Monitor performance and refine designs quickly and painlessly
- Align with DevOps maturity best practices throughout the process

MATLAB
& SIMULINK

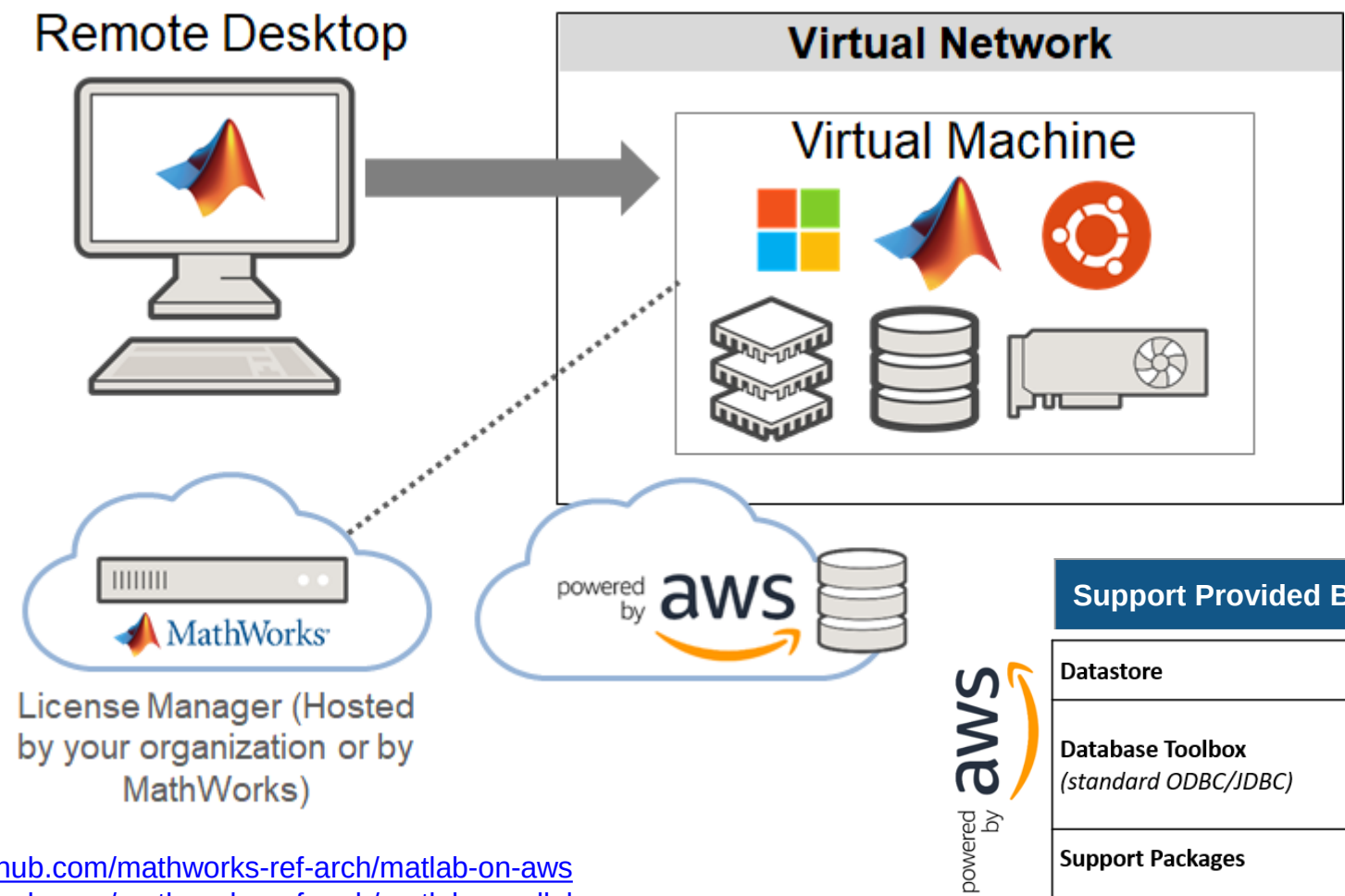
databricks



aws

MATLAB 针对AWS的集成架构

MATLAB Reference Architecture



Support Provided By:	Data Services		
Datastore	Amazon S3 <i>(read, out of memory data)</i>		
Database Toolbox <i>(standard ODBC/JDBC)</i>	Amazon Aurora	Amazon RDS for PostgreSQL/MySQL/MariaDB/Oracle/SQL Server	
Support Packages	Amazon S3 <i>(read/write/delete, encryption, access control)</i>	Amazon EFS <i>(NFS/Linux)</i>	Amazon Athena <i>(Query of S3 data)</i>

<https://github.com/mathworks-ref-arch/matlab-on-aws>
<https://github.com/mathworks-ref-arch/matlab-parallel-server-on-aws>
<https://github.com/mathworks-ref-arch/matlab-production-server-on-aws>

总结

- MATLAB提供了使用一流存储和计算平台的集成体验
- 使工程师能够在云中分析大规模的数据集
- 支持各种工作流以支持Model DevOps最佳实践
- 利用与AWS和Databricks的合作关系来支持您在云中使用MATLAB
- 如果你想了解更多关于如何为你的公司实现这一目标的信息，请与我们联系

MATLAB EXPO 2021

Thank you



© 2021 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.