Unified Analytics

Unifying Data Pipelines and Machine Learning with Apache Spark

Microsoft

databricks

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

ABOUT BLUEGRANITE

your data & analytics experts

Microsoft partner delivering Business Intelligence, Advanced Analytics, and Data Management solutions on Microsoft's Azure, SQL Server and Power BI platforms

DATA

ACQUISITION
ETL / ELT, Batch, Streaming

PREPARATION
Enrichment, Governance, Master

PROVISIONING
Data Lakes, Data Warehouses

ANALYTICS

MODELING
Business Logic, Cubes

CONSUMPTION
Monitoring, Reporting, Exploration

ADVANCED
Data Mining, Predictive, ML

DISCOVER

Explore innovative ideas to understand value and build a foundation for success beyond requirements to possibilities

CREATE

Enable users and help address key pain points, realize gains, uncover insights

REALIZE

Extend ROI and harvest value from your investments by aligning culture

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

www.blue-granite.com
Phone: 877.817.0736 | e-mail: sales@blue-granite.com

Microsoft Partner
Microsoft
Gold Data Analytics
Gold Cloud Platform
Gold Data Platform

BlueGranite Azure Databricks Workshop

Meet the team – Denver CO – Aug 2019

Josh Fennessy

Principal Architect

Grand Rapids, MI

Andy Lathrop

Principal Data Scientist
Denver, CO

Josh Smarrella

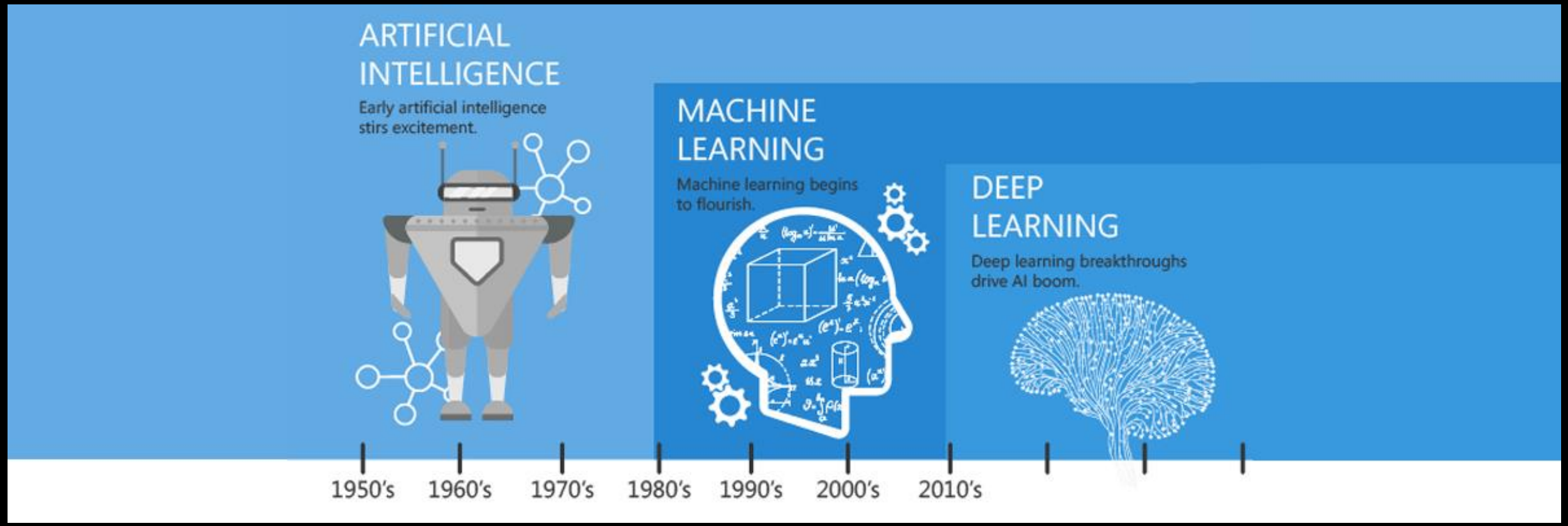
Account Leader
Indianapolis, IN

“Paradigm Shifts”: Fundamental Change In Our View Of How Things Work

Paradigm shifts require major architectural changes or business model changes.

Latest AI projects across industries

AI Evolution



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

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

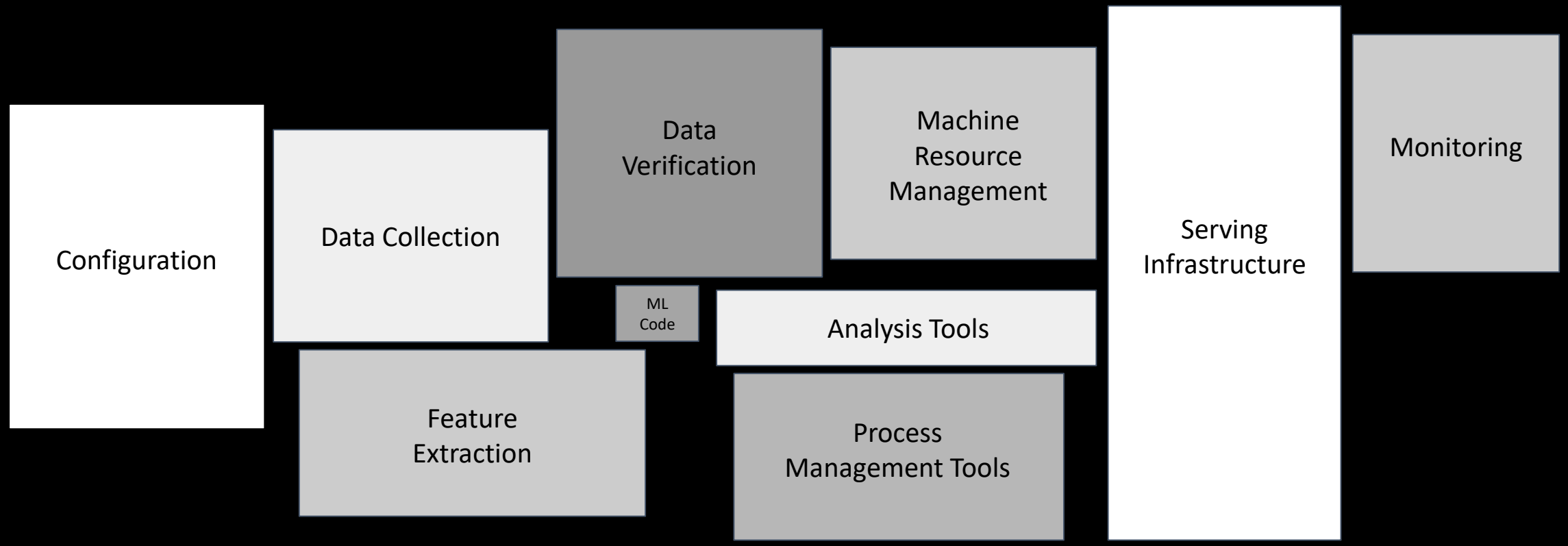
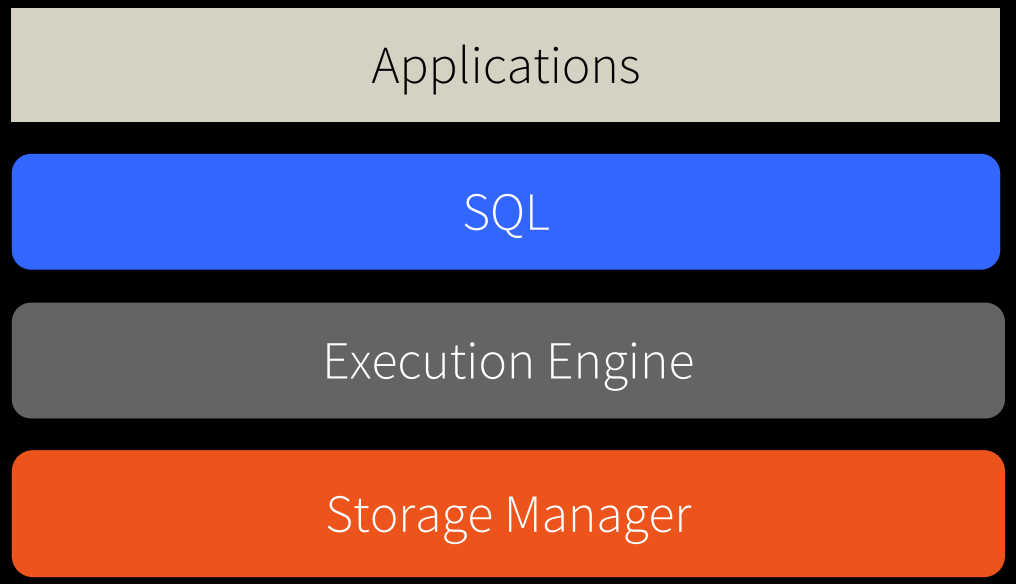


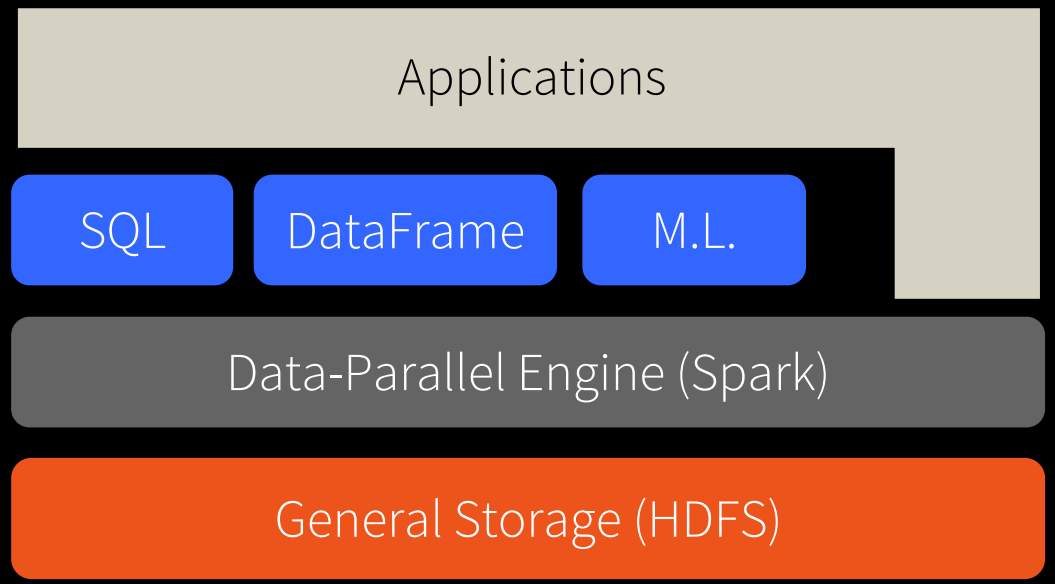
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

Relational Databases



One way (SQL) in/out and
data must be structured

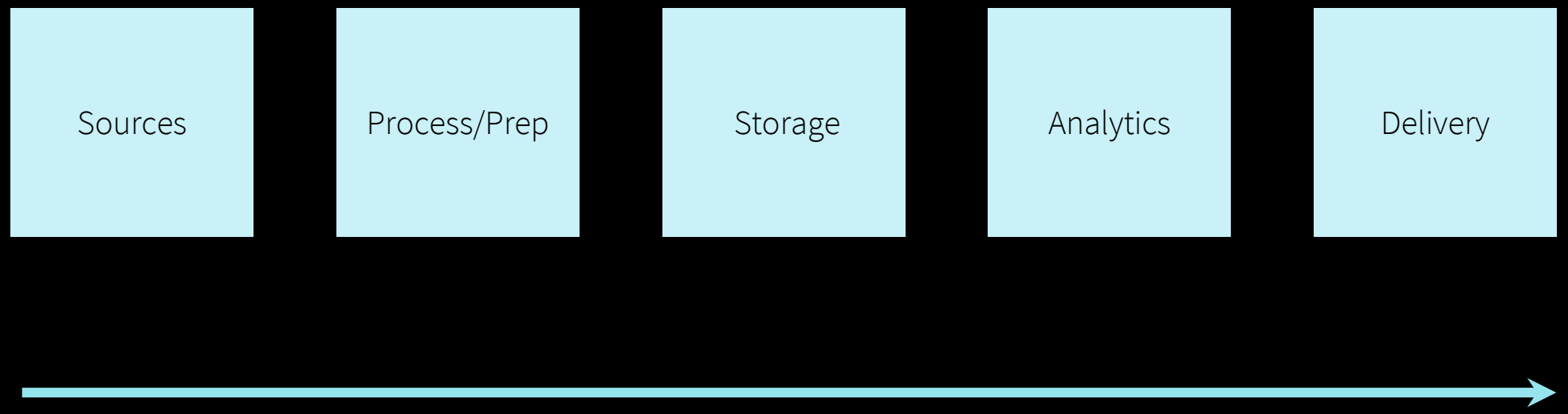
Big Data Stack



Not only SQL
Structured, semi-structured, unstructured data

Focusing on analytics…

Components of a Data Platform:



Data Sources	Process Prep	Storage	Analytics	Delivery
Streaming Event Hub IoT Hub Kinesis Kafka Log files ...	**ETL** Spark Data Factory Mapping Flow ...	**Data Lake** Delta Parquet ...	**Model Training** MLlib MLRuntime	**Production ML** Azure ML ML Flow ...
Batch MySQL Postgres SQL Server CSV files ...		**DW** Azure DW Azure DB	**Ad-hoc & Reporting** SQL DataFrames	**Viz & Dashboard** notebooks dashboards Power BI

Integrated Data Science Environment (**notebooks, RStudio, Data Robot**)

Workflow Orchestration (**jobs, MLflow**, airflow, …)

Data Governance and Security (**OpsSec, ACL**, …)

Dr. Lester Mackey

NETFLIX Prize

anonymized movie rating dataset

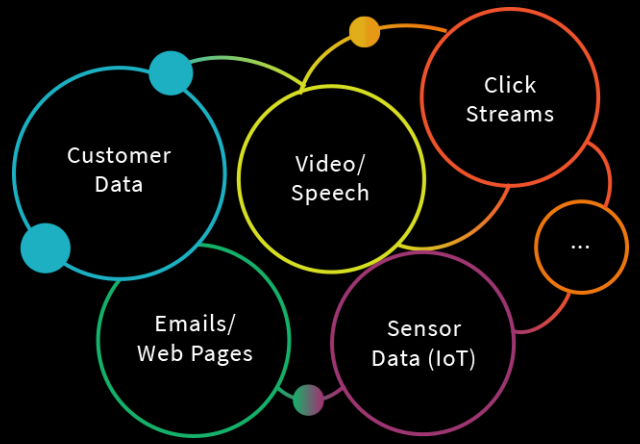
best recommendation algorithm

$1m

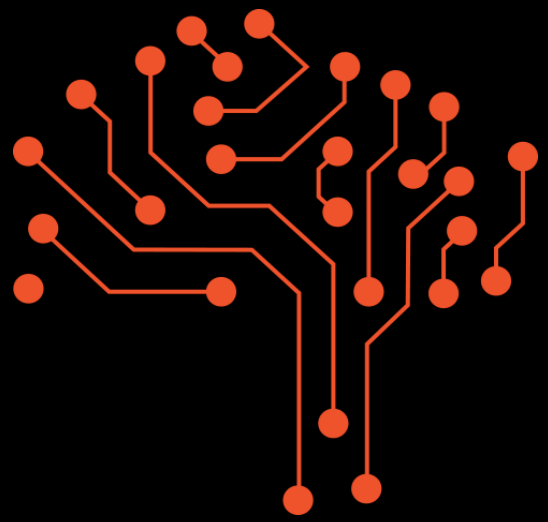
Dr. Lester Mackey

Dr. Matei Zaharia

The first unified analytics engine in 600 lines of code…



APACHE Spark™



Big Data

Machine Learning

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update | Download

Leaderboard

tied for best score

20 mins late

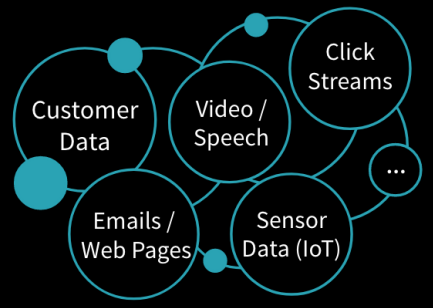
Showing Test Score. Click here to show quiz score

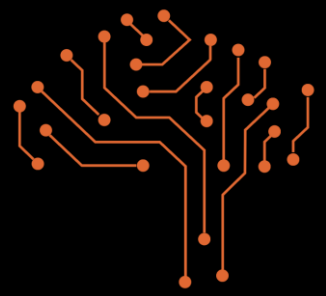
Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

Apache Spark: De-Facto Unified Analytics Engine

Uniquely combines Data & AI technologies





Big Data Processing

ETL + SQL + Streaming

Machine Learning

MLlib + SparkR

500,000

meetup members

In spite of Spark's success, companies are still struggling with Analytics and ML

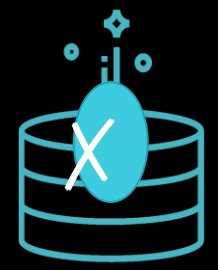
3 challenges for Analytics and ML Projects:

1. Data is not Ready for Analytics
2. A Zoo of ML and AI Frameworks
3. Data Science & Engineering silos

Challenge 1

Data is not ready for Analytics

Data reliability challenges with data lakes



Failed production jobs leave data in corrupt state requiring tedious recovery

Lack of schema enforcement creates inconsistent and low quality data

Lack of consistency makes it almost impossible to mix appends and reads, batch and streaming

A New Standard for Building Data Lakes

Open Source and Open Format

Data Reliability and Quality

Compatible with Spark APIs

Delta Lake ensures data reliability

- Full ACID Transactions
- Unified Streaming and Batch
- Schema Enforcement
- Time Travel/Data Snapshots
- Native Support for UPDATE/DELETE/MERGE

Log Structured Storage

Changes to the table are stored as *ordered, atomic* units called commits

```
000000.json
000001.json
…
```

```
Add 1.parquet
Add 2.parquet
```

```
Remove 1.parquet
Remove 2.parquet
Add 3.parquet
```

Checkpoints as Data

Large tables can contain millions of files. How do we scale the metadata? Use Spark.

```
Add 1.parquet
Add 2.parquet
Remove 1.parquet
Remove 2.parquet
Add 3.parquet
```

Get Started with Delta using Spark APIs

Instead of **parquet**...

```
CREATE TABLE ...
USING parquet
...

dataframe
    .write
    .format("parquet")
    .save("/data")
```

... simply say **delta**

```
CREATE TABLE ...
USING delta
…

dataframe
    .write
    .format("delta")
    .save("/data")
```

Challenge 2

A Zoo of new ML Frameworks

Complexity - Zoo of ML frameworks

Machine Learning

Scikit-learn,Spark MLlib, H20, 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, Azure ML, Cassandra, Redis, TensorBoard
...

Databricks Runtime for ML

Ready to use clusters with built-in ML Frameworks

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

kubernetes · docker

Data Engineers

Data Scientists

Databricks MLflow: Unifies Data Scientists & Engineers

1 Data Prep

Build reliable data pipelines
Track the datasets

Delta Lake

2 Build Model

Track Experiments
Reproduce experiments

MLflow Tracking
Databricks Runtime for ML
Delta Lake Time Travel

3 Deploy Model

Deploy models in production, track their quality

MLflow Models

Data Engineers

Data Scientists

databricks

Databricks Unified Analytics End-to-End for Analytics and ML

Interactive Demo

Introduction to Azure Databricks

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

Customer Case Study

Global Retail Marketing Firm

Microsoft Azure databricks

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

Company Profile

- 30,000 team members in 100 offices worldwide
- 16,000 front-line team members keep items on the shelf and accurately priced
- 600 Analysts and Data Scientists drive insights from syndicated data
- Clients are many of the largest grocery and CPG companies in the world

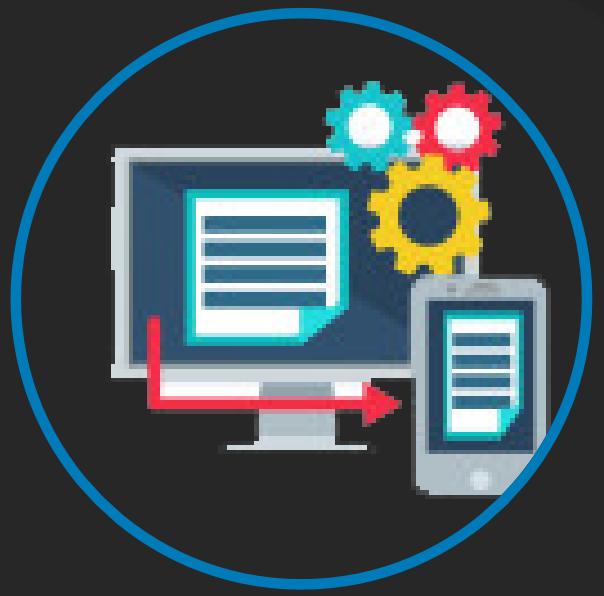
The Data

Syndicated Data

Retail Syndicated Data is purchased and merged with client point-of-sale data to build actionable insights for Clients and Retailers

Retail Insights

With deep industry experience and a strong front-line army of marketers, our Customer can provide valuable, actionable insights to their Clients



Self-service

In addition to pre-designed reports, our customer needs to provide the ability for their Clients to self-serve custom analytics

Challenges

Data Size

- Up to 50GB per client daily and growing
- ETL processes extending beyond job windows

Data Format

- Source data for a single data set in many formats
- Formats change often
- Requires Manual Processing

Data Restatements

- Data from previous days, weeks, months will need to be restated
- Manual data processes make this incredibly difficult
- Reports can be confusing for Clients due to static nature

Manual Data Management

- Data is generated, aggregated, pivoted in Excel and Access
- Manually prepped data is then ingested through automated processes
- Delays are common and custom requests are not supported

Onboarding New Clients

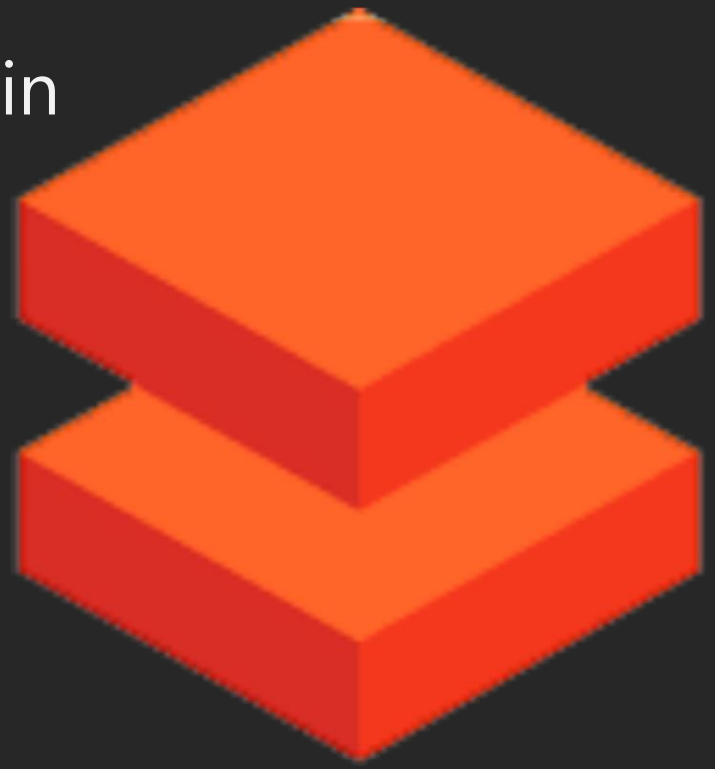
- Due to manual processing onboarding new clients takes weeks to months
- Data validation errors are common for first few cycles
- Data Security is also complex due to manual processing

Requirements

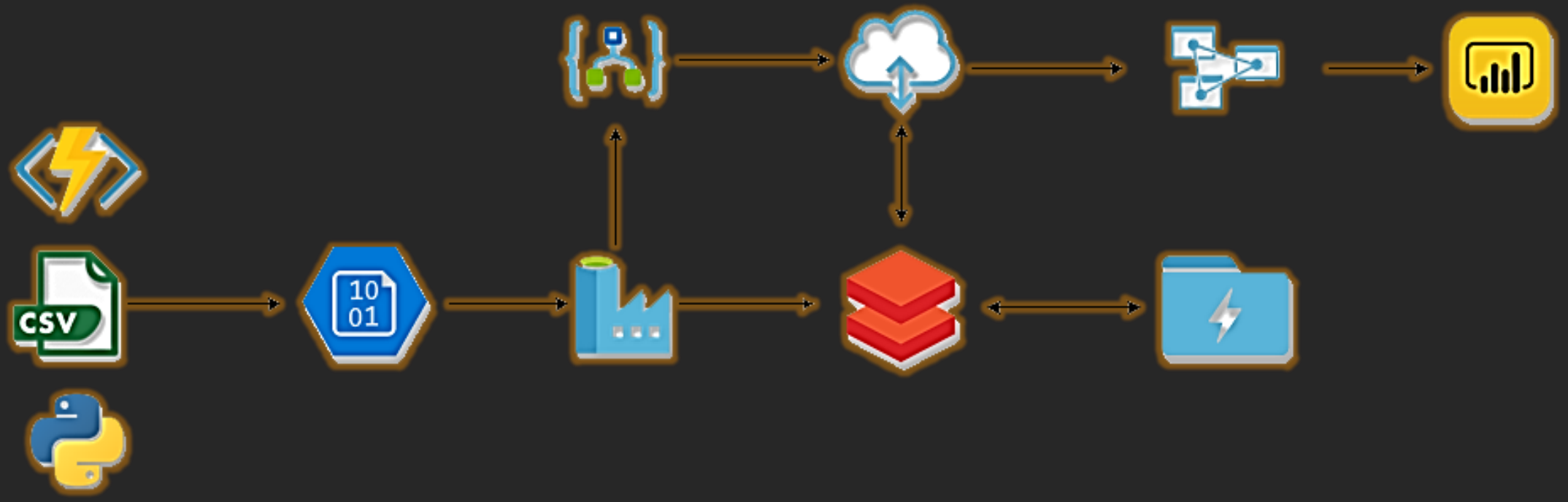
- Modern functional interface for data access and compatibility with modern reporting and visualization tools
- Integrated security and authentication
- Ability to build Metadata driven code set
  - Unable to custom build ETL/data processing for each client
- Cost Effective
- Cloud First

Why Azure Databricks?

- It’s Cloud Only
- It’s Cost Effective
- Integrated AD authentication and RBAC built in
- Flexible programing API/Python support
- Scalability
- Ease of platform management
- Integration with Azure DevOps/Github
- Integration with other Azure tools



Solution Architecture



Results

- Onboard first customer within 6 weeks of completion. 2 other followed within another 3 weeks.
- Data mapping now automated and only requires small Metadata input to work with varying data formats
- Data processing window cut by at least 4x
- Reports now generated before they are due (instead of after)
- New report development is no longer held up with data management issues
- Databricks named as organization-wide ETL tool for future projects

Lessons Learned

- Don’t lock in early platform decisions
  - Build PoCs/experiments to make sure the proposed architecture is correct
  - This solution evolved over 2 years and the tools have changed as new features are released, etc.
- Don’t shy from Preview features
- Build DevOps in early on
- Maximize collaboration in Notebooks

Hands On Lab

Unified Analytics with
Databricks and Delta Lake

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

Machine Learning Development is Complex

ML Lifecycle and Challenges

mlflow
An open source platform for the machine learning lifecycle

Raw Data → ETL → Featurize → Train → Score/Serve Batch + Realtime → Monitor Alert, Debug

Production Logs

Update Features

Retrain

Zoo of Ecosystem Frameworks

Tuning | Deploy | Model Mgmt

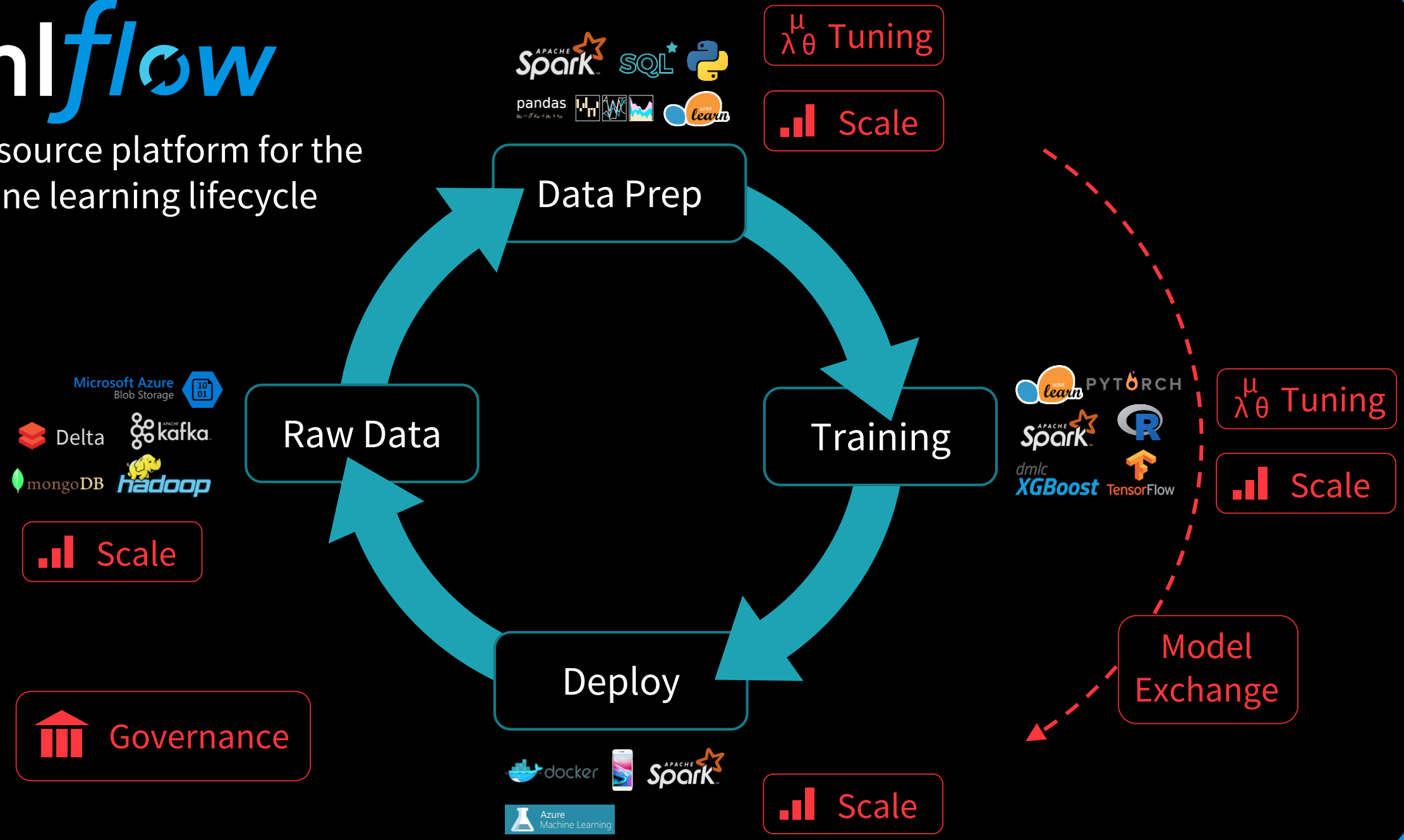
Collaboration | Scale | Governance

Feature Repository | Experiment Tracking | AutoML, Hyper-p. search | Remote Cloud Execution | Project Mgmt (scale teams) | Model Exchange | A/B Testing | CI/CD/Jenkins push to prod | Orchestration (Airflow, Jobs) | Lifecycle mgmt. | Data Drift | Model Drift

databricks

mlflow

An open source platform for the machine learning lifecycle



+ mlflow

Vision: Make it painless for a single user to go from raw data to production ML without leaving Databricks

MLflow Components

mlflow Tracking

Record and query experiments: code, data, config, results

mlflow Projects

Packaging format for reproducible runs on any platform

mlflow Models

General model format that supports diverse deployment tools

Hands On Lab

Supercharged AI with
Databricks and MLFlow

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

3 challenges for Analytics and ML Projects:

1. Data is not Ready for Analytics
2. A Zoo of ML and AI Frameworks
3. Data Science & Engineering silos

Databricks Unified Analytics brings it all together

1 2 3

Databricks Unified Analytics Platform

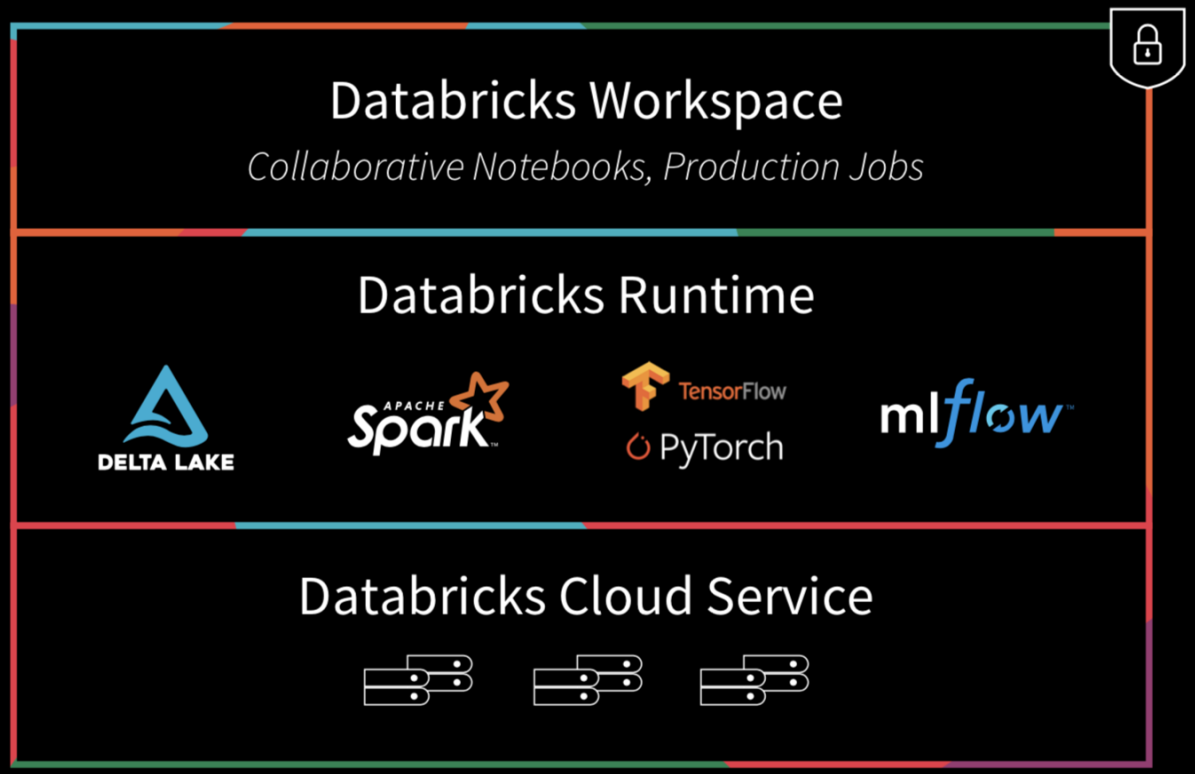
End-to-end ML platform that unifies people, processes and technologies

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

What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure

Best of Databricks

Best of Microsoft

- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage, ADLS, Event Hub, Azure Monitor, Azure DevOps, Azure Data Factory etc.)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

BLUEGRANITE
DATA • INSIGHTS • ANALYTICS

CASE STUDIES | BLOG | EVENTS | RESOURCES

SOLUTIONS | INDUSTRIES | ABOUT | CONTACT US |

DATABRICKS RECOGNIZES BLUEGRANITE AS 2019 PARTNER OF THE YEAR

SEE WHY WE WON

AZURE DATABRICKS SERVICES

Azure Databricks Workshop

In this 1-day, on-site workshop, up to 10 attendees from your organization will receive a detailed overview of Azure Databricks, giving them an understanding of where and how it fits into the Azure Data Platform.

LEARN MORE

Azure Databricks Proof of Concept

In this 1-2 week Proof of Concept, BlueGranite experts will work with your organization to demonstrate Azure Databricks capabilities for data engineering or data science, depending on the needs of your team.

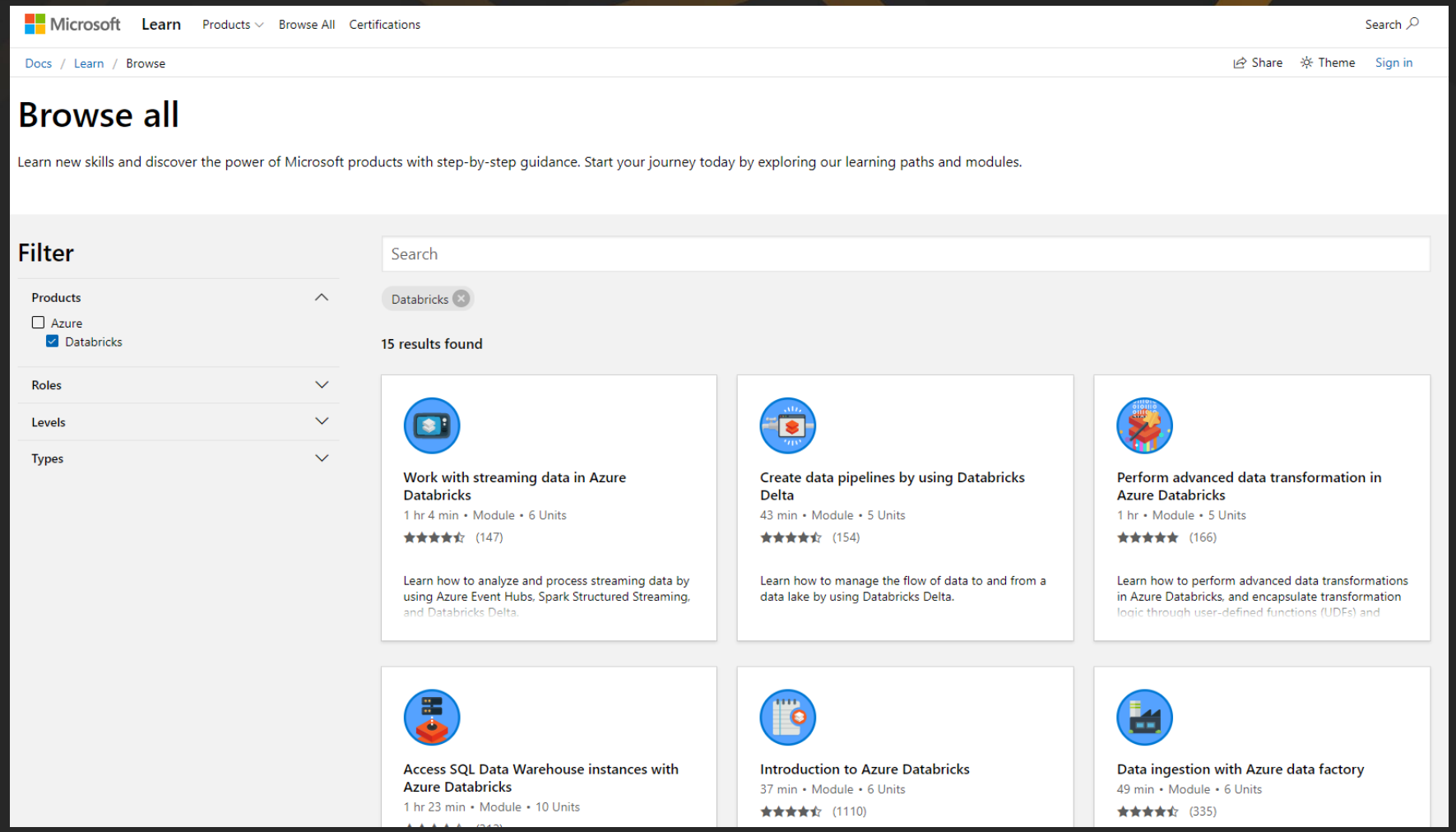
LEARN MORE

Azure Databricks & Spark

These all day sessions are co-sponsored by BlueGranite and Microsoft. We introduce a deep dive experience into Azure Databricks.Check out our Event's Page to learn more about future dates and locations across the US.

VIEW UPCOMING EVENTS

https://www.blue-granite.com/databricks



https://docs.microsoft.com/en-us/learn/browse/?products=azure-databricks