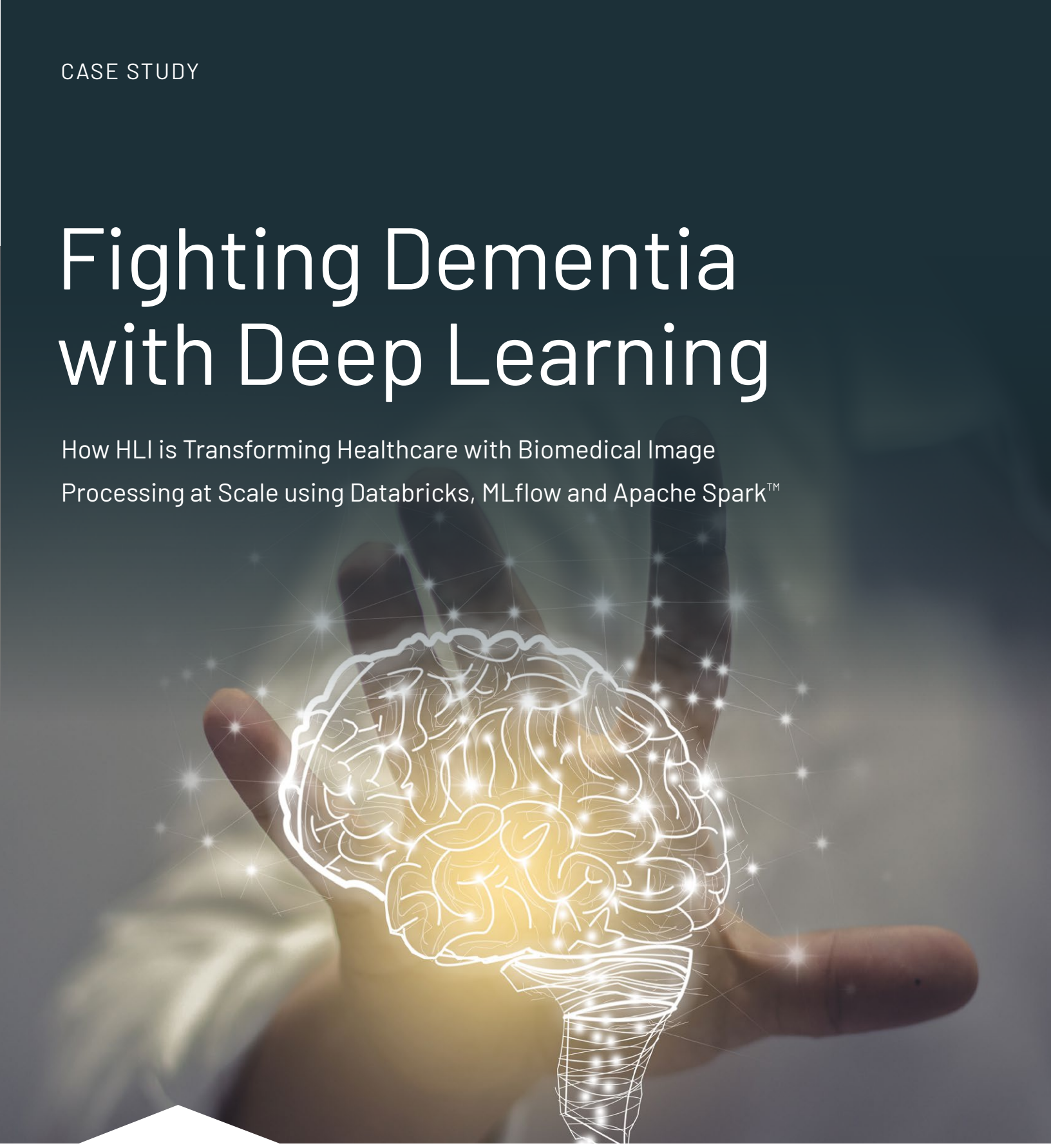
CASE STUDY

Fighting Dementia with Deep Learning

How HLI is Transforming Healthcare with Biomedical Image Processing at Scale using Databricks, MLflow and Apache Spark™



databricks

Summary

Human Longevity Inc. spoke during a live webinar on how they use Databricks Unified Data Analytics Platform to help predict a patient's likelihood for developing chronic disease. This is a summary of their talk.

The Challenges

- **Poor collaboration across bioinformatics, data science and research teams:** siloed data teams using disparate tools and inefficient workflows slowed data flow to machine learning
- **Poor collaboration across bioinformatics, data science and research teams:** complicated data pipelines that lacked integration with CI/CD, Data lake, and ill-suited for processing medical images
- **Massive volumes of unstructured data:** 60,000 3D MRI images totaling 12 terabytes of data with new images arriving weekly
- **Difficult to meet HIPAA standards:** hard to anonymize images to meet HIPAA requirements

The Databricks Solution

- **Simplified infrastructure management:** reduced operational costs through automated cluster management and cost management features such as autoscaling and spot instances
- **Collaborative workspaces:** interactive notebooks improve cross-team collaboration and data science creativity, allowing Human Longevity to greatly accelerate model prototyping for faster iteration
- **Simplified ML lifecycle:** managed MLflow simplifies the machine learning lifecycle
- **Reliable ETL at Scale:** an agile and efficient analytics pipelines that can handle medical images while retaining regulatory compliance

The Results

- **Faster ETL pipelines and shorter ETL development time:** Databricks allows Human Longevity to ETL 60 million files in less than 24 hours
- **Reduced Costs:** Human Longevity reduced the cost to ETL patient files by 2.5x
- **Higher Productivity:** fostered collaboration between data scientists, bioinformatics and research teams by providing a shared workspace with support for a broad set of languages and visualizations for different level users
- **Faster Deployment:** reduced deployment times from weeks to minutes as operations teams deployed models on disparate platforms
- **Accelerated discovery of new biomarkers:** enabled team to build a risk model that identifies dementia 8+ years prior to diagnosis empowering clinicians to provide early intervention

HUMAN LONGEVITY, INC.

BUSINESS USE CASE

Preventative Health Screens — predict a patient's individual risk for developing chronic conditions by applying deep learning to MRI brains scans, genetic profiles and EHR data

TECHNICAL USE CASES

- Build reliable and performant data pipelines to process terabytes of 3D medical images for downstream machine learning
- Scale machine learning/deep learning while maintaining regulatory compliance

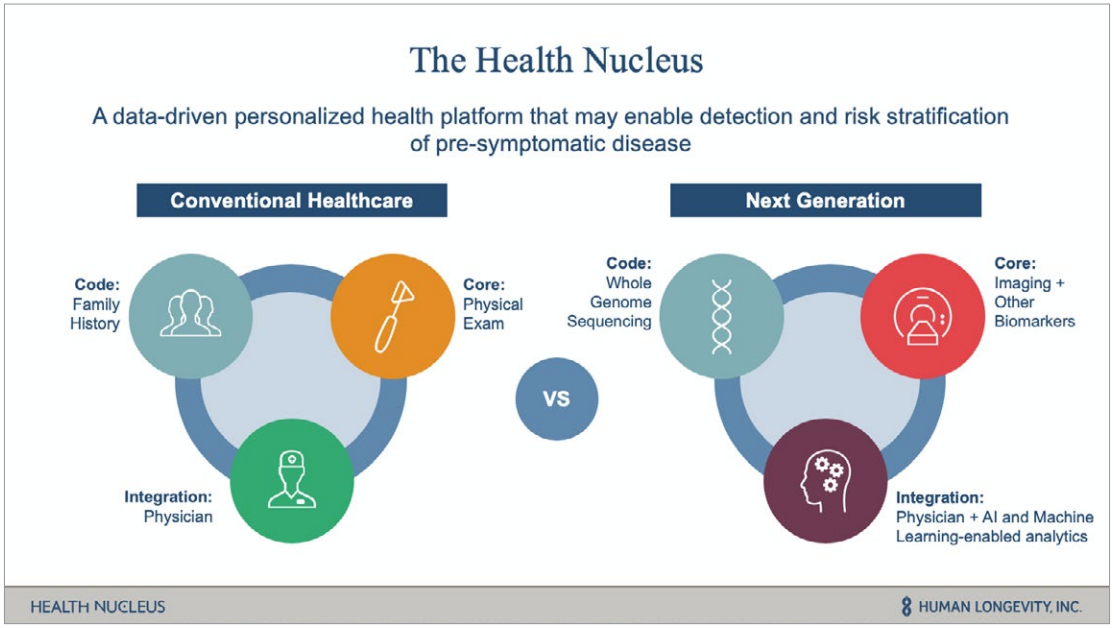
Overview

Knowledge is power. In medicine, knowledge can be the power to prevent disease and extend life.

Despite huge technological strides in medicine, conventional preventative healthcare relies on knowledge that's not so far removed from the pre-digital era. Physicians assess factors like family history, lifestyle, and the results of a simple physical exam (blood pressure, heart rate, temperature) to look for warning signs of incipient illness.

But many illnesses take years, even decades, to manifest symptoms detectable through traditional means. Because early detection often translates into better outcomes, a new generation of data-driven, personalized preventative healthcare aims to discover illnesses - and risk for illness - long before symptoms appear.

Leading-edge healthcare providers are augmenting the diagnostic expertise of physicians with insights from AI and machine learning, which draw on diverse data sources, such as whole-genome sequencing and images from MRIs to provide a more detailed picture of an individual's current health and risk for future illness.



Human Longevity, Inc. is one of these providers. Founded in 2013, the company offers a product called Health Nucleus, a personalized health platform for detecting and determining the risk of diseases like dementia.

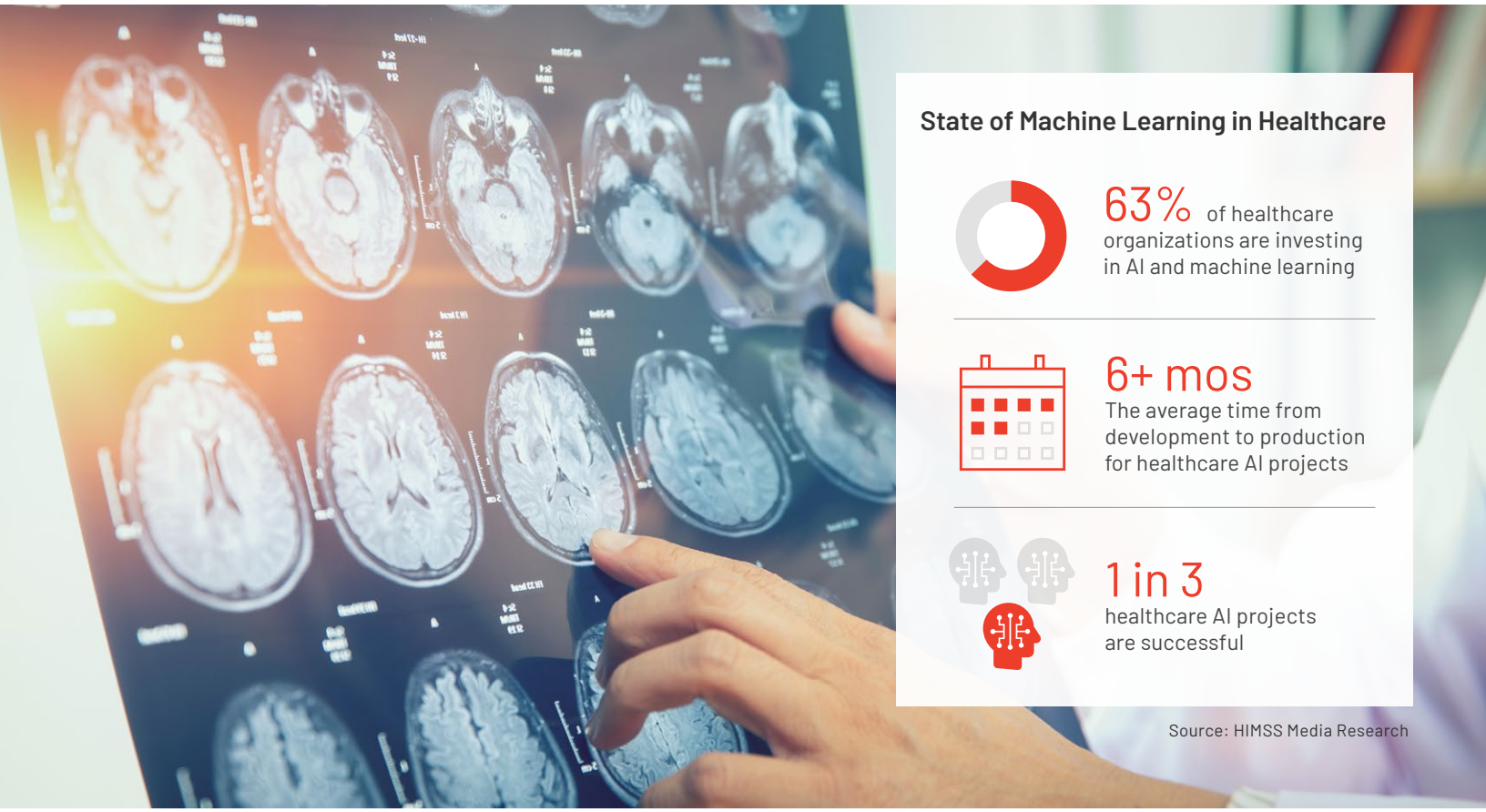
Tackling Dementia with Deep Learning

Dementia is an insidious condition that can begin with small changes in the brain 20 years before symptoms arise.

By the time memory loss and language problems become apparent, the damage to the brain is irreparable. But not everyone with increased risk for dementia has to ultimately suffer from the disease.

Researchers estimate one out of three cases of dementia could be prevented if at-risk individuals made lifestyle changes in midlife. Even if the disease can't be staved off entirely, early intervention can slow dementia's progression and reduce its severity.

To help patients better understand where their risk of dementia lies, Human Longevity, Inc uses deep learning pipelines to study thousands of MRIs for quantitative and qualitative analysis. When paired with genomic data, the final product is a report that empowers patients to better manage their health.



State of Machine Learning in Healthcare

63% of healthcare organizations are investing in AI and machine learning

6+ mos
The average time from development to production for healthcare AI projects

1 in 3
healthcare AI projects are successful

Source: HIMSS Media Research

Challenges Hampering Innovation

From the outset, Human Longevity, Inc.'s machine learning ambitions were big. However, there were a number of challenges that slowed their ability to ingest and build models at scale.

MASSIVE VOLUMES OF UNSTRUCTURED DATA

One of the biggest issues was sheer data volume. In order to deliver on their use cases, the company needed to analyze large volumes of unstructured data, with more than 60 million DICOM images totaling around 12 terabytes with new images arriving weekly. Given their legacy infrastructure and inefficient workflows, this proved to be a daunting task.

POOR CROSS-TEAM COLLABORATION

HLI also found it difficult to iterate on models with data science and engineering teams that were partitioned in their respective silos. A common enough problem for most organizations where data science teams were bottlenecked by the lack of data engineering resources, often spending too much of their own time on DevOps work managing and maintaining clusters.

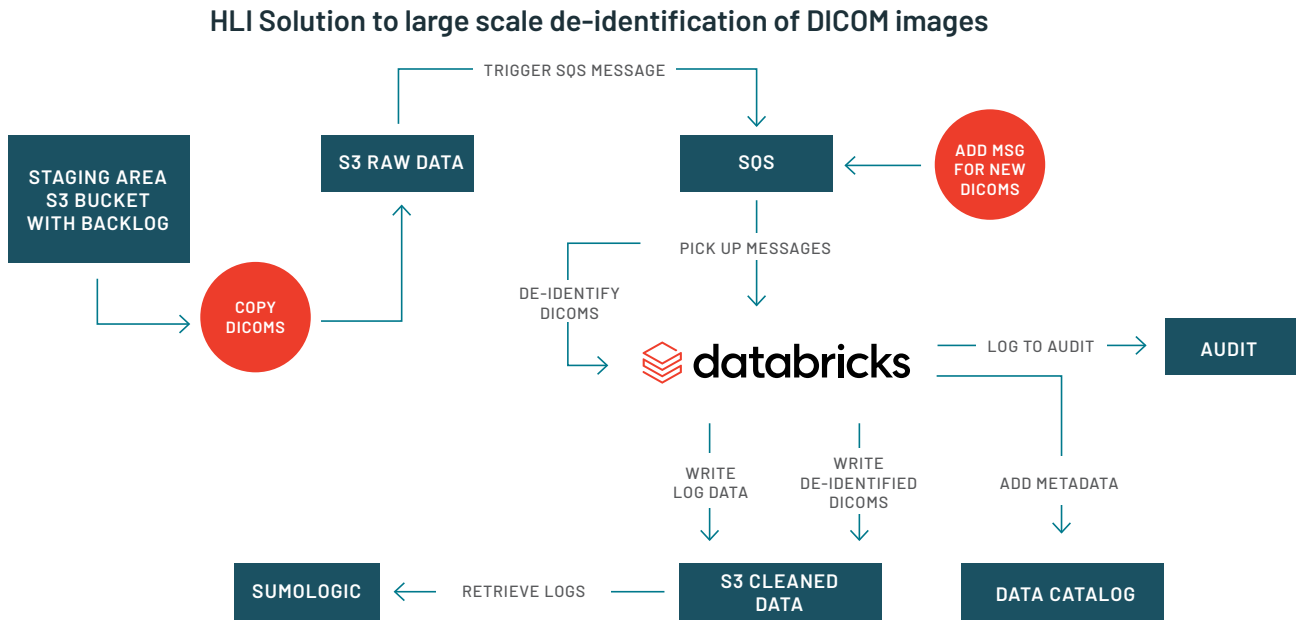
DIFFICULT TO MEET HIPAA STANDARDS

Within the healthcare industry, regulatory compliance is always at the forefront of any data strategy. HLI engineers were hamstrung with anonymizing the brain scans and remaining in compliance with rigorous HIPAA requirements.

INFRASTRUCTURE COMPLEXITY AND LIMITATIONS

Further complicating matters, HLI's infrastructure slowed their ability to build highly performant novel diagnostic pipelines. They realized that they needed to raise the engineering bar with core functionality such as logging issues, CI/CD integration, and integration with their data lake.

To overcome these challenges, HLI began to look for a standardized and unified platform that could help the company deliver on the promise of machine learning in the service of personalized medicine.



Saving Money With Simplified Infrastructure Management

Even with the economies of scale offered by the cloud, biomedical image processing at scale can rapidly become quite expensive. To help manage costs, HLI turned to automated cluster management from Databricks and AWS spot instances. Because autoscaling with Databricks allows HLI to precisely scale AWS clusters up or down when demand is high or utilization is low, clusters can be resized much more aggressively in response to actual load without killing tasks or recomputing intermediate results. This minimizes wasted compute resources without compromising cluster responsiveness or efficiency. Autoscaling combined with spot pricing on 125 M5.large EC2 instances allowed HLI's DevOps to run more smoothly and at a lower cost. With Databricks at the core of its infrastructure, HLI was able to ETL 60 million files in less than 24 hours while reducing costs by cost 2.5x.

"Databricks helps us solve all of our data engineering AND data science problems."

CHRISTINE SWISHER, DIRECTOR OF MACHINE LEARNING, HUMAN LONGEVITY, INC.

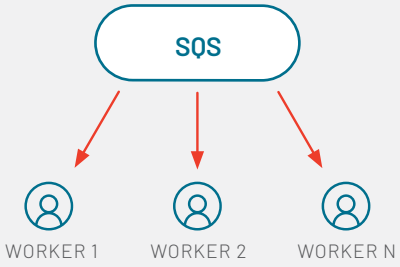
Building a Reliable and High-Performing Data Pipeline

HLI used Databricks Delta Lakes to efficiently ingest and prepare data for downstream machine learning. HLI's data is stored in S3, but it is now fed through an SQS messaging system into Databricks, which initiates ETL batch jobs to cleanse and prepare the data for downstream analytics.

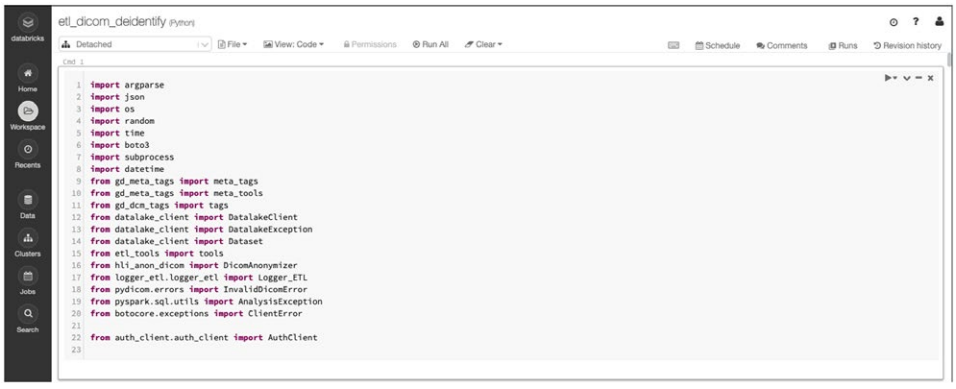
SPEED/LOW COST/SCALABILITY

- Automatic cluster with spot instances
- Up to 120 M5 large instances in autoscaling cluster
- Use boto3, rather than mounted buckets
- Distribute work evenly across workers

HLI is able to ETL 60+ million files within 24 hours.



Metadata is critical to DICOM specifically and biomedical processing at scale generally. Because Delta Lake leverages Spark's distributed processing power to handle metadata, Delta Lake was easily able to handle HLI's terabytes of files. HLI also stores a subset of metadata in a non-relational database optimized for big data to give researchers a standardized and speedy way to query images for study. This also offers an extra level of security, a consideration which is always important when dealing with medical data.

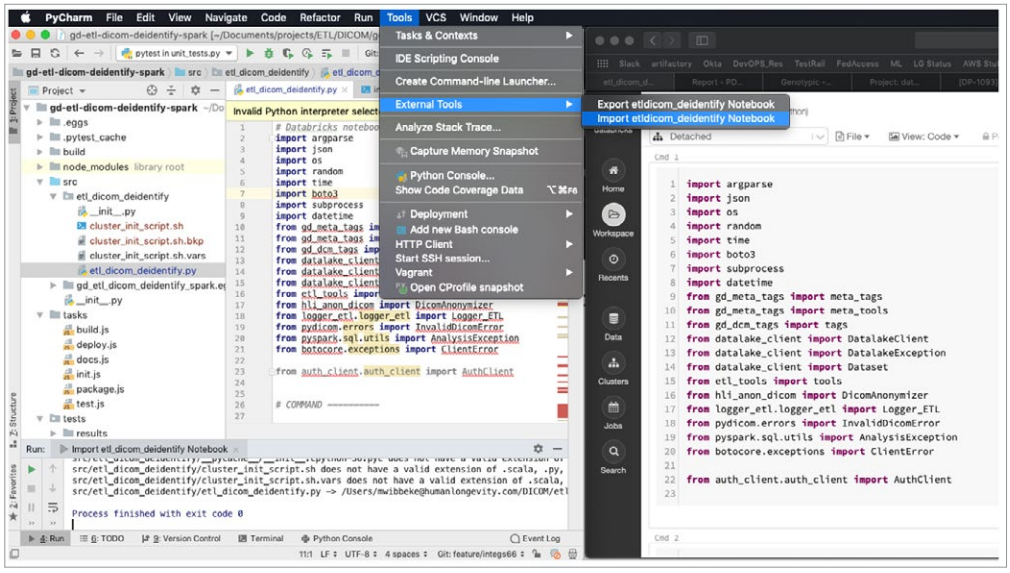


HLI developed their core logic in Databricks on an interactive cluster

HLI developed their core logic in Databricks on an interactive cluster, and an interactive workspace CLI made it simple to copy and paste code from the IDE to a Databricks notebook for quick and easy debugging. Exporting the clean code back into the IDE was just as painless. This integrated the new and improved pipeline with HLI's CI/CD process.

Fostering Collaboration and Managing the Model Lifecycle with Notebooks and MLflow

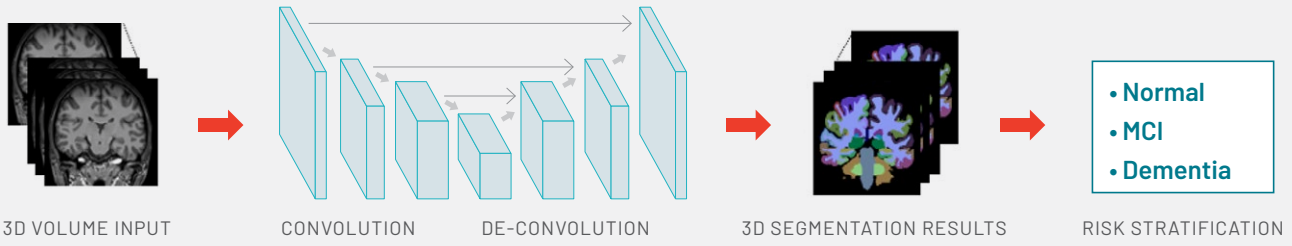
HLI data scientists used Databricks' interactive workspace to build models in their preferred scripting languages (such as R, Python, Scala, and SQL) and libraries (such as Tensorflow, Keras, Pytorch, scikit-learn, nltk ML, pandas, etc) in a shared notebook environment, and then seamlessly move those models to production with a single click.



Easily copy code from your IDE to a Databricks Notebook via Databricks IDE Integration

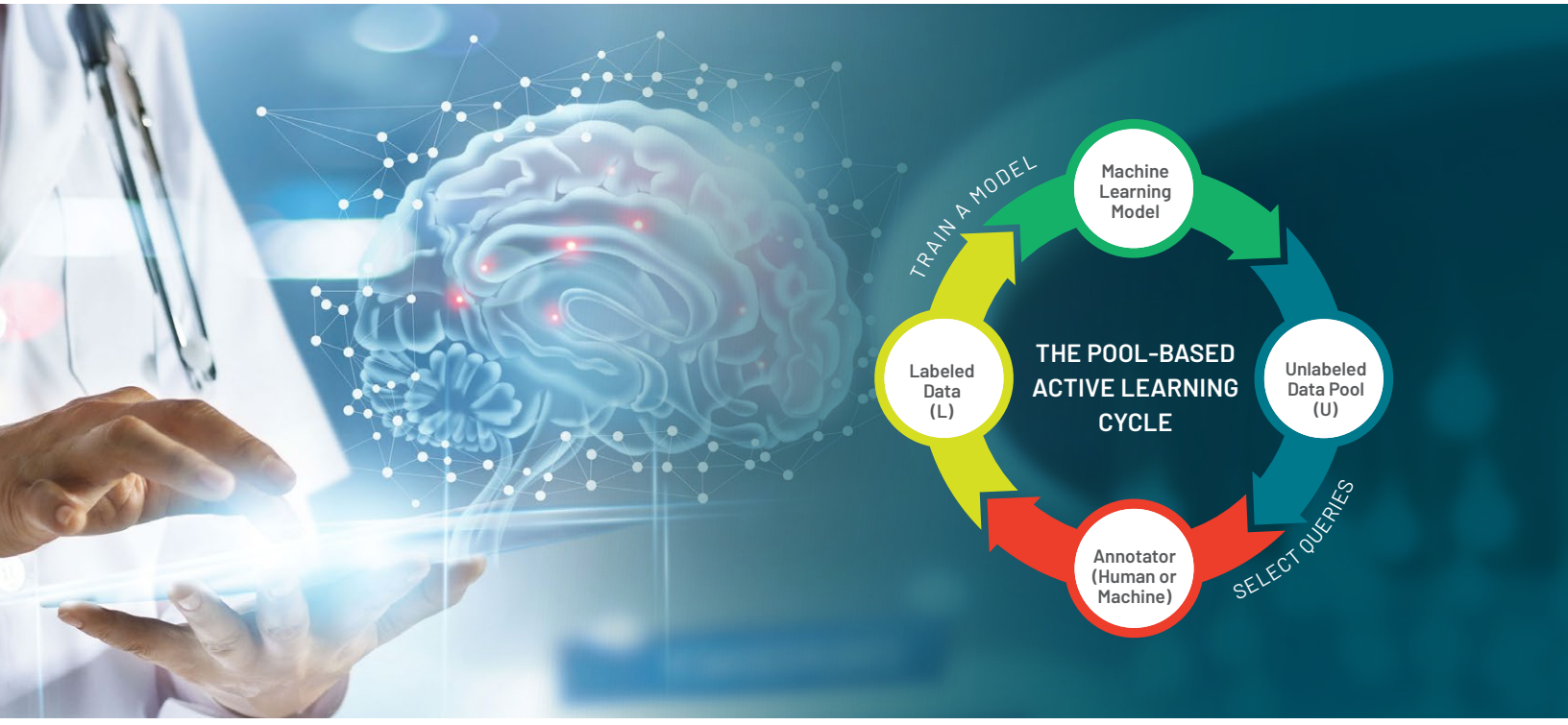
"Shorter ETL time and faster collaboration enabled our team to build our models faster."

MICHAEL WIBBEKE, DATA ENGINEER, HUMAN LONGEVITY, INC.



HLI developed a 3D fully convolutional neutral network for brain segmentation

Databricks MLflow, an open-source framework for managing the complete machine learning lifecycle, allowed HLI to iterate and share models across frameworks with ease.



The MLflow tracking feature offered HLI's data scientists a convenient location to store results and share the parameters for deep learning models, which allows for easier reproducibility while fostering better collaboration. Data scientists needing a fast method of testing ideas were able to associate a Databricks-hosted Notebooks enabled them to manage a particular MLflow project and visualize model performance — accelerating machine learning across the organization.

```
class MLFLow_Keras_Logger(Callback):
    """
    logs whatever is in the logs dict from keras that matches whats in the passed in end
    """
    def __init__(self, end_of_epoch_items=['loss']):
        # Call to super class
        super(MLFLow_Keras_Logger, self).__init__()
        self.items_to_log_end_epoch = end_of_epoch_items

    def on_epoch_end(self, epoch, logs={}):
        for value in self.items_to_log_end_epoch:
            log_metric(value, logs.get(value))
```

```
name: mlflowtest
channels:
  - defaults
dependencies:
  - numpy
  - pip:
      - mlflow
      - tensorflow==1.9.0
      - keras
```

```
name: mlflowtest
conda_env: conda.yaml
entry_points:
  main:
    parameters:
      debug: {type: str, default: "false"}
      run_name: {type: str, default: "CHANGE ME"}
      experiment: {type: str, default: "/Users/agr

    command: "python main.py"
```

MLflow is environment agnostic, allowing data teams to easily deploy models into production

Results

Accelerating Healthcare Innovation

FASTER ETL PIPELINES AND SHORTER ETL DEVELOPMENT TIME
Databricks allows Human Longevity to ETL 60 million files in less than 24 hours

FASTER DEPLOYMENT
Reduced deployment times from weeksto minutes as operations teams deployed models on disparate platforms

REDUCED COSTS
Human Longevity can ETL patient files at a cost of 50 cents per file

HIGHER PRODUCTIVITY
Fostered collaboration between data scientists by enabling different programming languages through a single interactive workspace

BETTER CODE
Databricks IDE integration makes debugging code easier and faster

By rapidly ingesting terabytes of biomedical imaging data, allowing teams to collaborate on a single platform, streamlining model development, and accelerating time to production, HLI broke new ground in the field of dementia research.

A unified data analytics platform improved the ability of AI to differentiate the disease's progression using newly identified biomarkers. More importantly from the perspective of HLI's patients, the company found that AI is as accurate as far more invasive diagnostic tests for dementia.

The immediate product is an extremely sophisticated medical report that patients can use to understand their risk for dementia and take measures to prevent or slow the disease's worst impacts.

But the biggest and most priceless returns will be decades from now when people who might have suffered from dementia are instead enjoying full and happy lives.

Learn More

ABOUT DATABRICKS

Databricks is the data and AI company. Thousands of organizations worldwide —including Showtime, Shell, Conde Nast and Regeneron — rely on Databricks' open and unified platform for data engineering, machine learning and analytics. Databricks is venture-backed and headquartered in San Francisco with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

EVALUATE DATABRICKS FOR YOURSELF

START YOUR FREE TRIAL

Contact us for a personalized demo databricks.com/contact

databricks