

  
 Success Story

# Migrating and generalising a forecasting framework from Hadoop to Azure Databricks

---

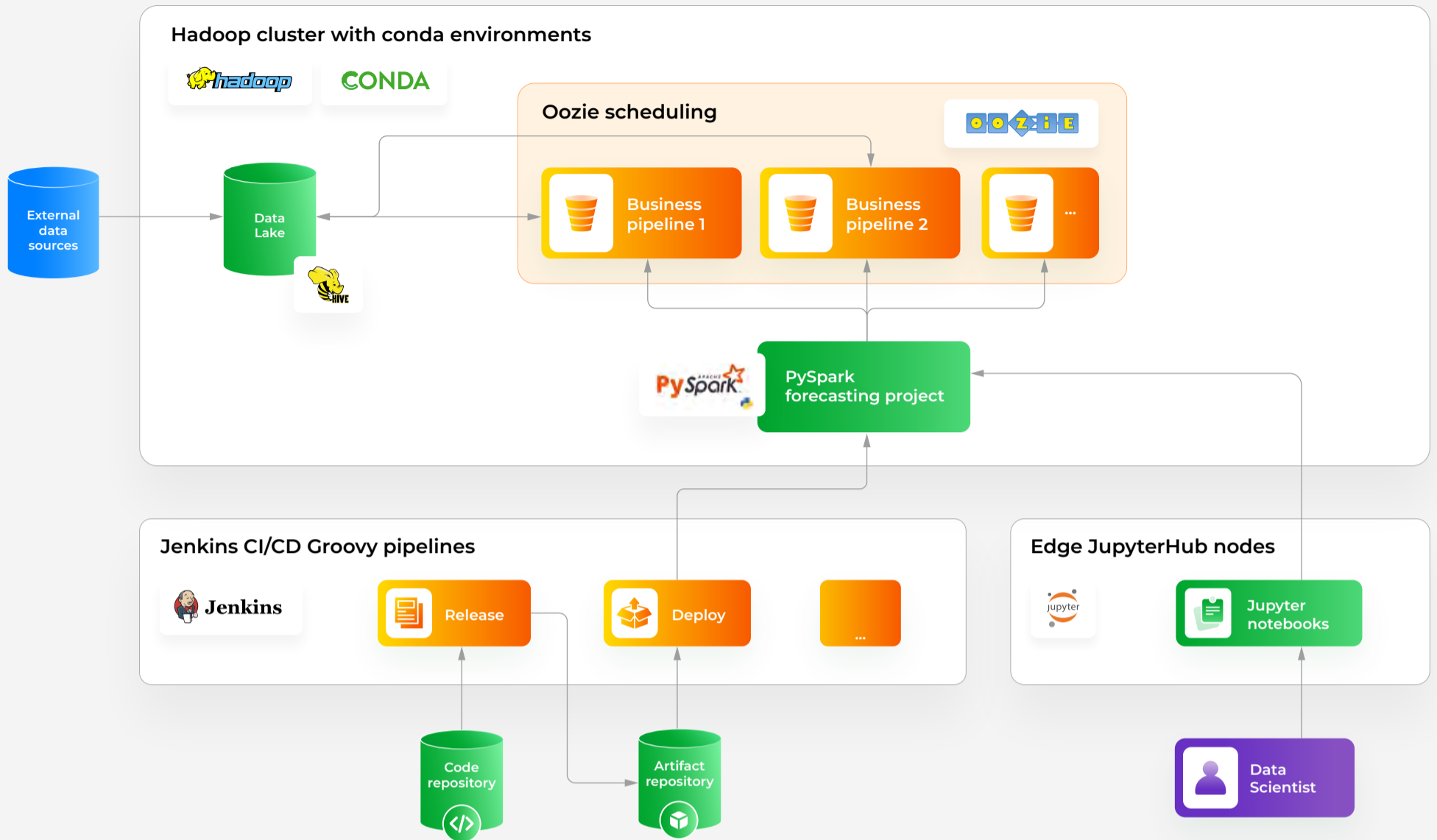
**Our client, a leading UK retailer,** aimed to extend its ML Forecasting Framework to other subsidiaries within its group network. The objective was to transition the extensive platform from Hadoop to a dedicated Azure Databricks environment, including all library dependencies and datasets.

However, lacking the necessary expertise and resources posed a challenge in refactoring the framework for the new environment. [VirtusLab](#), having previously developed the [forecasting framework](#) and being a partner of Databricks, emerged as the ideal choice to assist in the migration. Upon completion, the revamped framework resulted in the initiation of five distinct projects within the client's network.

 **The Challenge**

The forecasting framework achieved significant usability within the client's organisation. Because of its success, the retailer decided to migrate the platform to other organisations within the group. The migration aimed to enhance support for day-to-day forecasting and research activities related to mobile products, including phones, subscriptions, and accessories.

## The old architecture



The scalable framework itself was a large project, written in pySpark on a legacy Hadoop platform. It encompassed:

1. **A pySpark project** with tens of data processing pipelines.
2. **A set of machine learning models** featuring custom configuration and hyperparameter tuning code, running in parallel through Spark User Defined Functions (UDFs).
3. **Jenkins CI/CD pipelines** implemented in Groovy for automation.
4. **Management of the Python** dependency environment using Conda.
5. **Jupyter notebooks** provided for research purposes, leveraging the forecasting framework code.

Our client aimed to migrate this large scale project to a completely new Azure Databricks environment. This endeavour demanded to:

- **Revamp the framework** for use in a different environment, moving from Hadoop to Azure Databricks.
- **Make** the framework domain and company agnostic.
- **Move** selected data pipelines and all libraries to Azure Databricks.



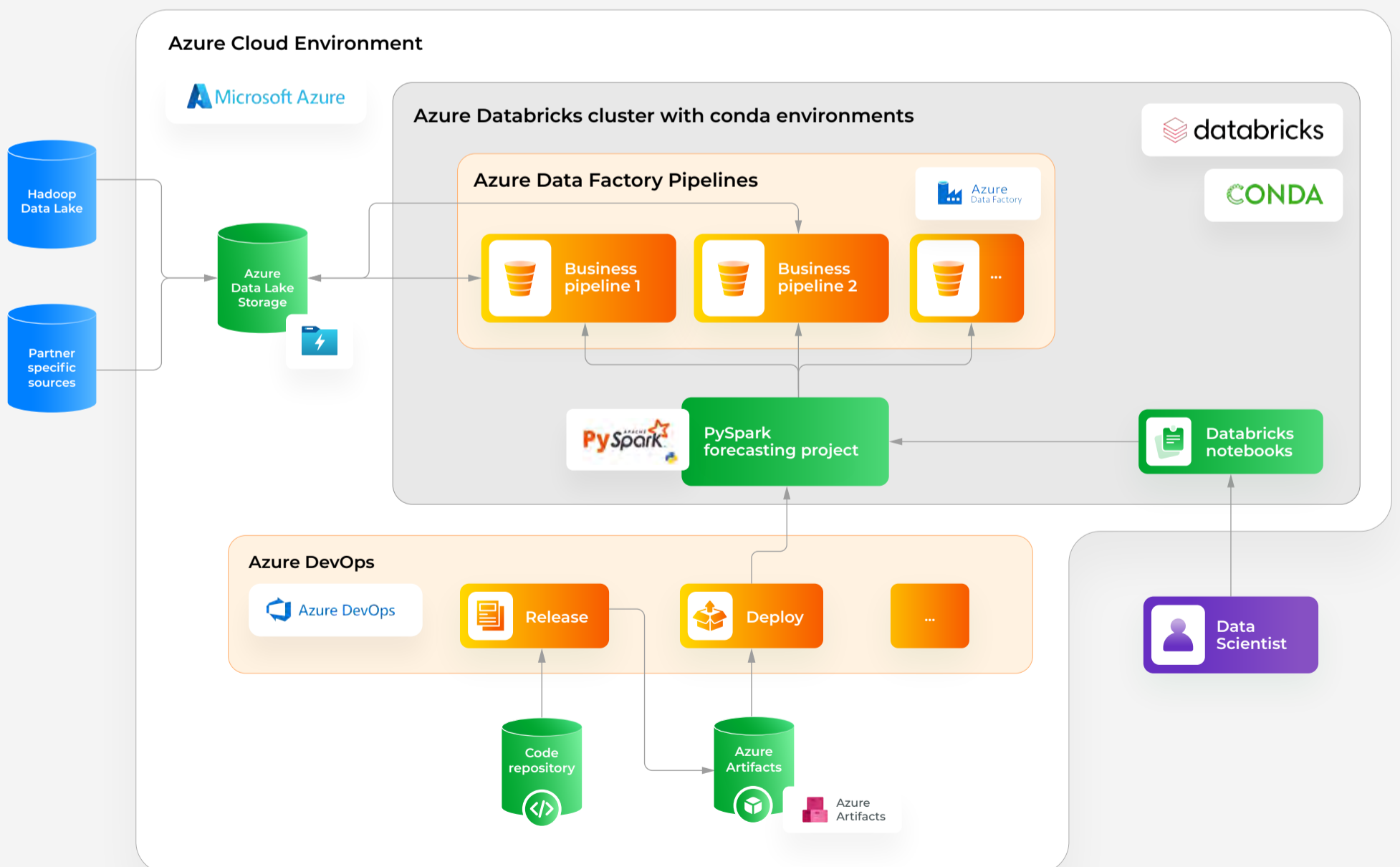
- **Enable merging central Hadoop** cluster datasets with new custom datasets from external sources.
- **Migrate Jenkins CI/CD pipelines** to Azure DevOps.
- **Empower Data Scientists** to perform research using notebooks in the new environment, accessing the code as a library.
- **Ensure strong security measures** to protect access to code, data, and artifacts in the new environment.

Relying on trust as the foundation of our partnership, our client reached out to VirtusLab for assistance.

## The solution

VirtusLab refactored the framework to suit new organisations and incorporated the option to integrate additional data sources in two major steps: **Refactoring the code** for migration and **preparing the infrastructure** for seamless framework execution.

### The new Azure Databricks architecture



## Code preparation

We enhanced the versatility of the forecasting framework by removing domain-dependent code elements like column and table names, as well as specific config settings:

- Removing scheduler-dependent code
  - Revamping the definition of defining clusters to make them applicable in any environment

This made it adaptable for implementation in various organisations. We also extracted Hadoop-specific components from the code, facilitating its execution in Azure Databricks, and other environments. For instance, we extracted the Oozie workflow generation used for Hadoop deployments.

## Infrastructure preparation

We helped to set up the infrastructure on Azure Databricks to enable the smooth execution of the framework. This involved:

- Creating new Databricks Dev and Prod compute clusters with preinstalled environments
- Automating updates for the forecasting framework and Python dependencies through Conda
- Migrating all CI/CD pipelines to Azure DevOps while hosting the framework as Azure Artifacts
- Integrating the preinstalled Forecasting Framework into Azure Data Factory for new projects
- Enabling the use of the framework for research via notebooks
- Implementing regular data exports from the Hadoop cluster to the new Azure Cloud
- Employing Azure Key Vault for secure secrets management.

### ★ The results

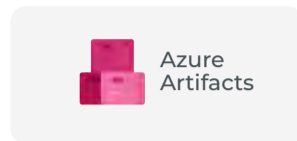
VirtusLab deployed the generalised forecasting framework in both Hadoop and Azure Databricks. Our client's subsidiaries used the framework within six months, following its successful restructuring and implementation in the new Azure Databricks environment. They also gained:

- **Customised Forecast Generation** – Regularly generated forecasts utilizing bespoke ML and statistical models tailored for the new domains, incorporating their individual models.
- **Migration of Engineering Best Practices** – Successfully transitioned all best practices such as CI/CD, tests, code reviews, and the creation of separate DEV and PROD environments to the new ecosystems and teams.
- **Immediate Project Implementation** – Promptly implemented five distinct, new projects in the updated environment using the migrated framework.



## The tech stack

### / Cloud environment:



### / Languages and dependencies:



### / Frameworks:



# About VirtusLab

At VirtusLab, we aim to lead in software technology, working consistently to enhance efficiency. Our profound commitment to research and development and a dedicated focus on emerging trends and inspirations fuels an innovative culture. This ethos precisely guides advancing our cutting-edge solutions, inviting collaboration to expand the boundaries of software technology collectively. We welcome you to be a part of this transformative journey.

[Let's connect](#)

## Contact Details

[info@virtuslab.com](mailto:info@virtuslab.com)

**POLAND**

**Kraków Headquarters**

Virtus Lab Sp. z o.o.  
ul. Szlak 49  
31-153 Kraków

**GERMANY**

**Berlin Office**

+49 30 52014256  
VirtusLab GmbH  
Potsdamer Platz 10  
10785 Berlin

**UNITED KINGDOM**

**London Office**

+44 (0)20 4577 1051  
Virtuslab Ltd.  
40 Bank Street HQ3  
London E14 5NR