DATA+AI SUMMIT BY databricks

CLOUD MIGRATION AND COST OPTIMIZATION IN THE DATABRICKS LAKEHOUSE

Thiago Barcellos – Head of Data at Allos
June 2024

ABOUT ME

FIRST, A BRIEF INTRODUCTION ABOUT ME

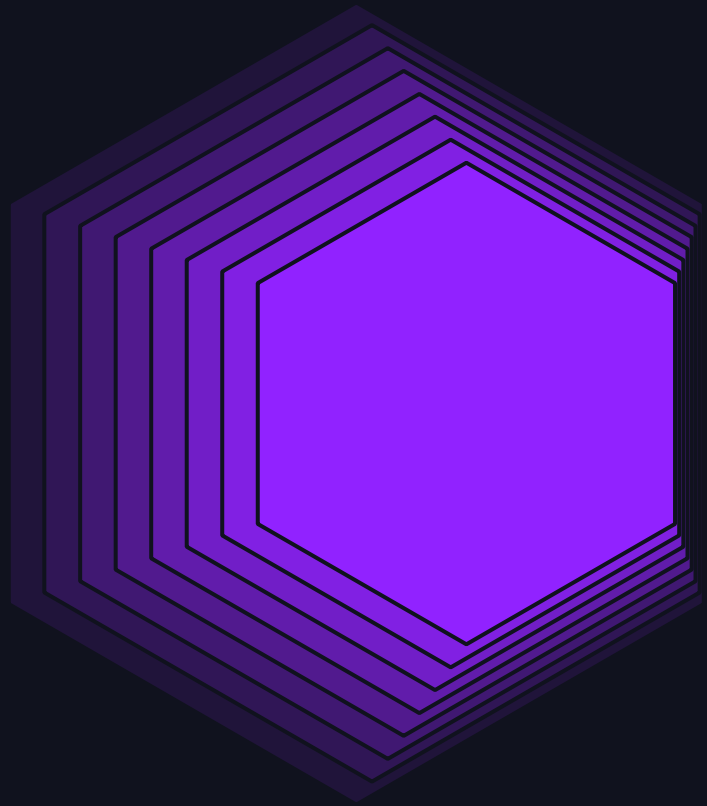
- Currently leading Allos' Data-Driven Journey with a team of over 40 professionals.
- Over 22 years of experience in Technology, with at least 17 years dedicated to implementing medium and large-scale data solutions, both in Brazil and other countries.
- Solid experience in structuring enterprise data strategy and assembling high-performance data teams from early stages.

Agenda

- About Allos
- Pre-Migration Context
- Migration Plan & Strategy
- Migration Implementation
- Results and Cost Optmization
- Current Data Environment
- Looking Ahead to the Future

DATA+AI SUMMIT BY databricks

ABOUT ALLOS



Thiago Barcellos
June/2024

ABOUT ALLOS

58
SHOPPING MALLS

+54 MN
VISITS PER MONTH

+2,5 MM
GROSS LEASABLE AREA

+15 THOUSAND
STORES

+16 MN
UNIQUE VISITORS

R$ 40 Bi
TOTAL SALES

Segment’s most innovative platform for entertainment, lifestyle, services and shopping and the largest shopping mall manager in Latin America.

ALLOS

INNOVATION

OUR ESSENCE

Biggest Phygital Plataform

in Latin America

Deeply understand consumers

Influence throughout the journey

INNOVATION

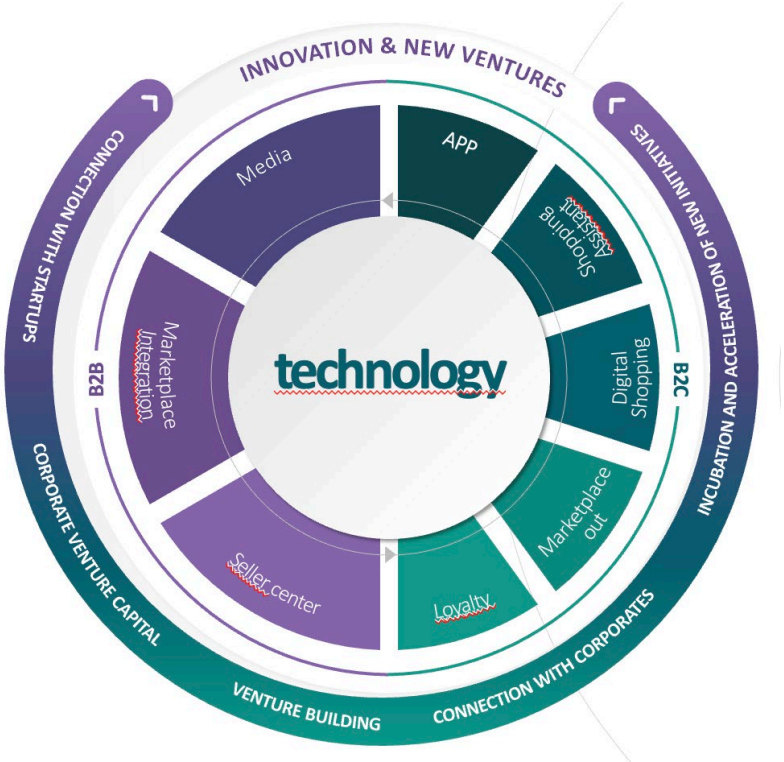
PLATAFORM ALLOSTECH

Goals

01.
Increase recurrency and share of life

02.
Monetize consumer knowledge

03.
Be a profitable platform

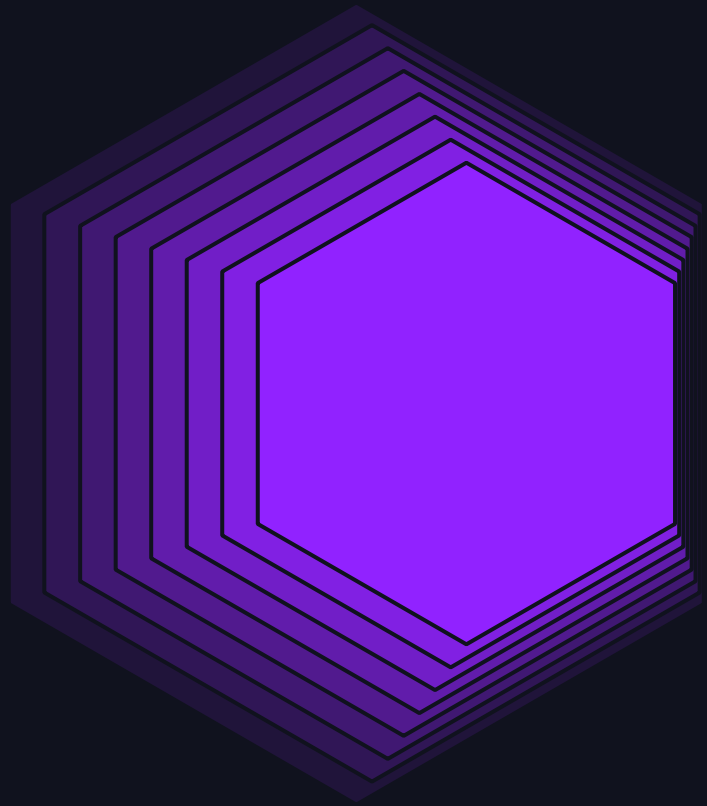


Strategic Pillars

- Relationship
- Phygital Solutions
- Monetization

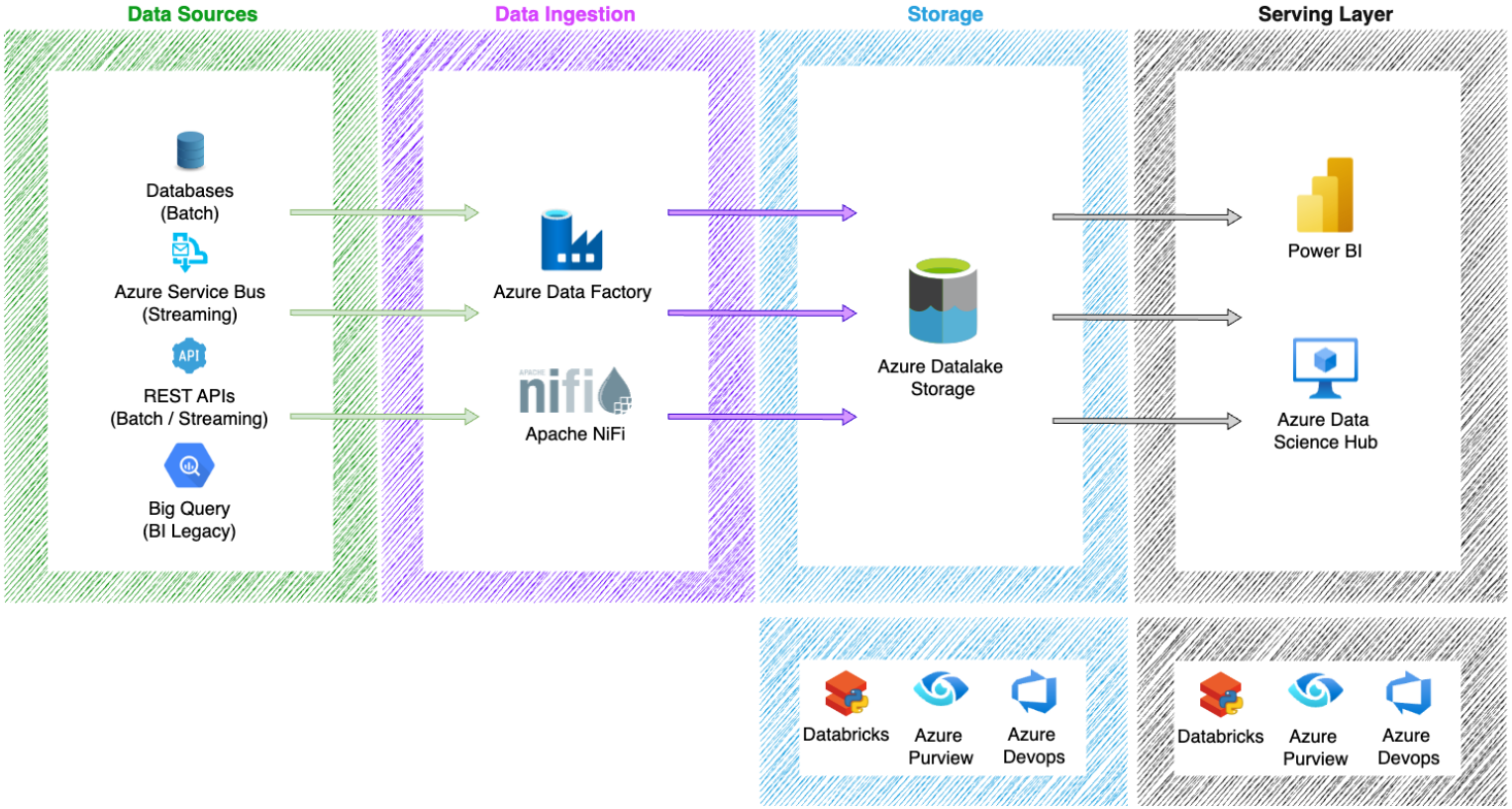
DATA+AI SUMMIT BY databricks

PRE-MIGRATION CONTEXT



Thiago Barcellos
June/2024

PREVIOUS ENVIRONMENT



MAIN PAINS AND ISSUES

The need for change was immediate.

01. Lack of standardization and Governance Processes

- Governance practices were lacking, resulting in unmonitored and uncontrolled data environments.
- No well-defined processes for testing and deployment, leading to frequent errors and inefficiencies.
- Absence of standardization across processes and tools creating inconsistencies and reliability issues.

02. Performance and Scalability Issues

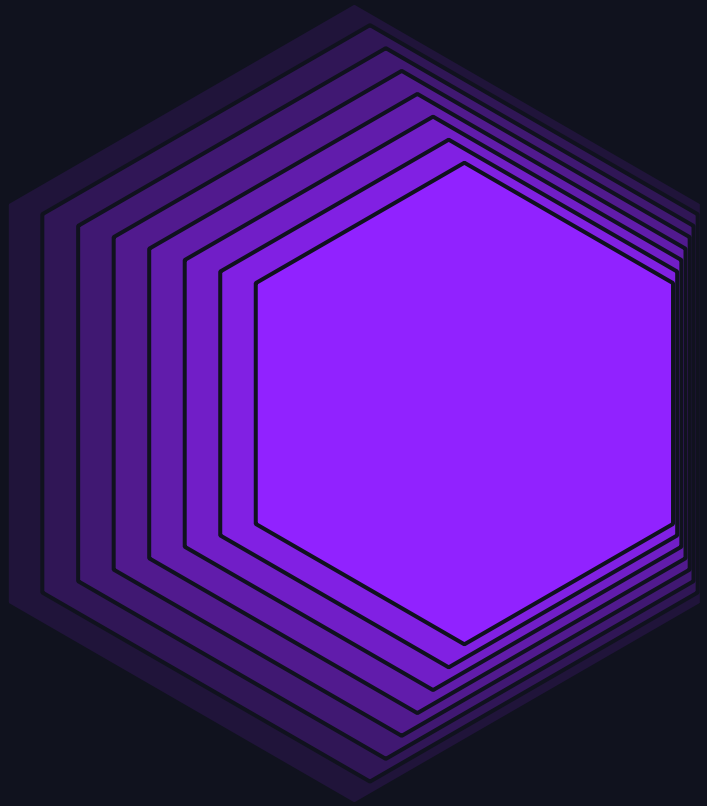
- The existing architecture was not designed to handle the volume and type of data we were operating with
- Performance bottlenecks were common due to the architecture's limitations.
- Scalability issues hindered our ability to efficiently manage growing data needs.

03. Cost and Efficiency Concerns

- The costs associated with Azure Data Factory were not well-managed, leading to budget overruns.
- Poorly developed code requiring us to use more computational capacity than necessary, increasing operational costs.
- Multiple dashboards containing business rules within Power BI itself, preventing reuse and leading to reliability issues.

DATA+AI SUMMIT BY databricks

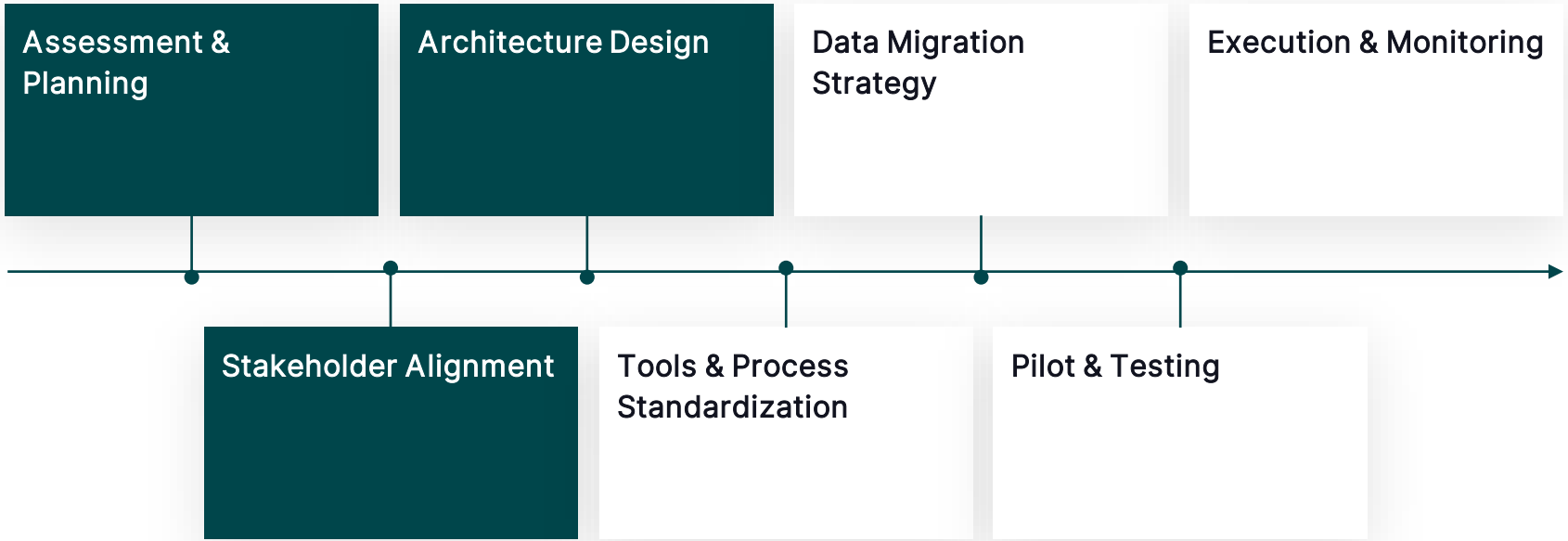
MIGRATION PLAN & STRATEGY



Thiago Barcellos
June/2024

MIGRATION PLAN

Steps for a sucessfull migration



THE SELECTION PROCESS

Why migrate to GCP?

Price

Performance

Productivity

SOME COLLATERAL EFFECTS

Changing the technical stack

databricks

Unity Catalog x Purview

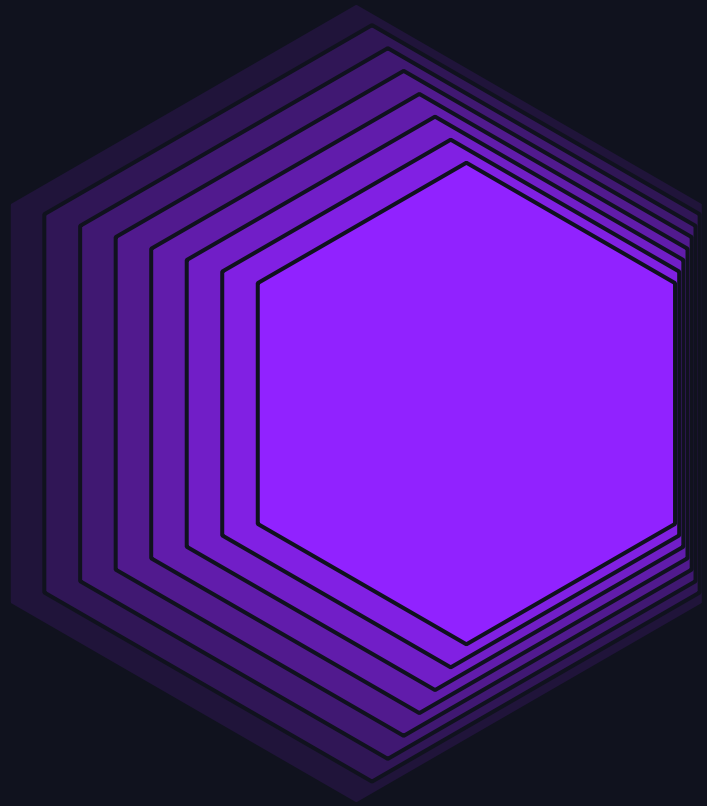
- Strong focus on data governance and compliance with detailed lineage tracking
- Native and Simplified Integration with Databricks
- Granular Access Control
- Advanced Metadata Management
- Performance and Scalability

Aiflow (Composer) x Data Factory

- Advanced Orchestration and Scalability
- Cost and Implementation Flexibility
- Flexibility and Customization
- Version Control and Integration with Git
- Active Community and Support

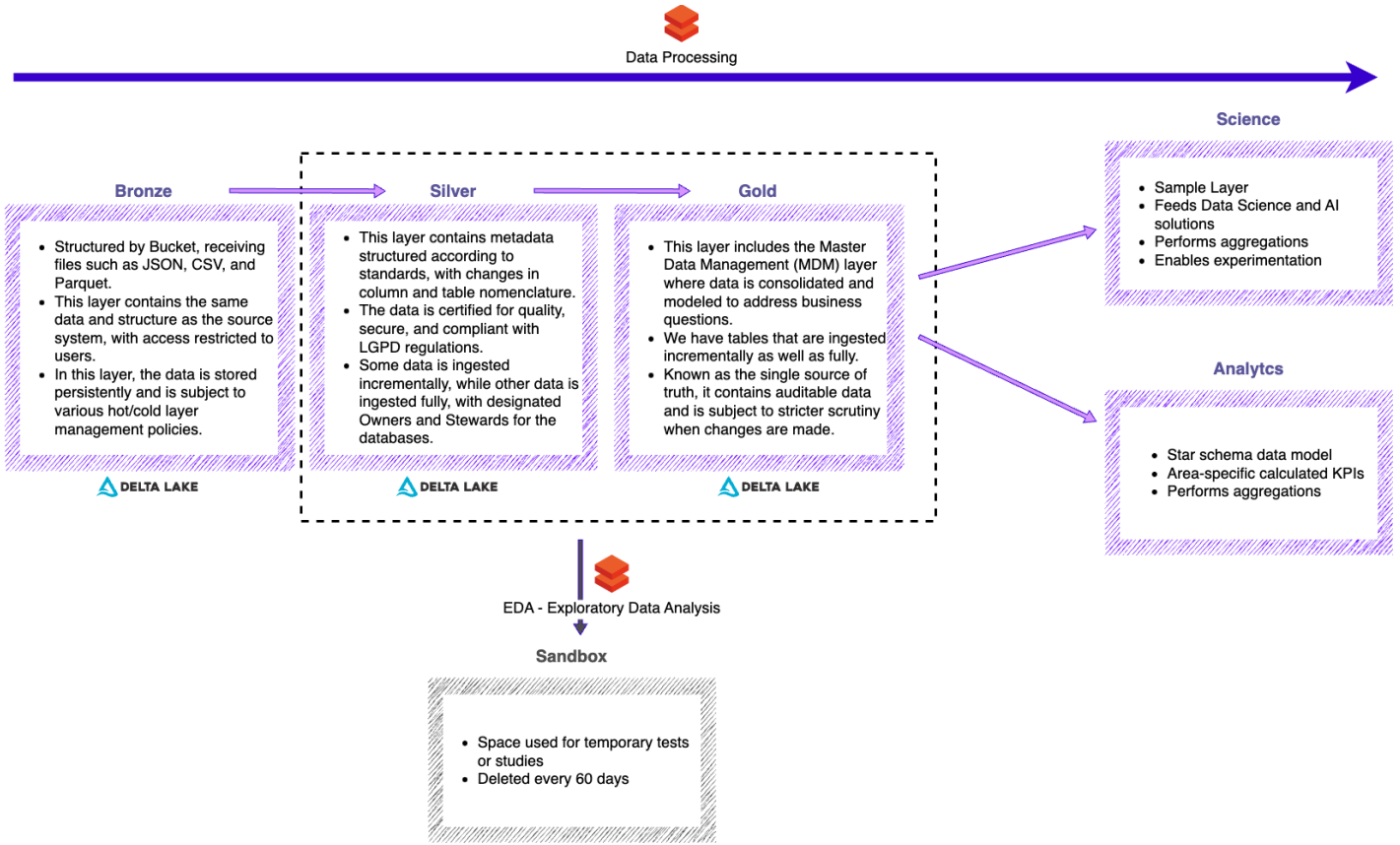
DATA+AI SUMMIT BY databricks

MIGRATION IMPLEMENTATION



Thiago Barcellos
June/2024

ALLOS LAKEHOUSE



EXECUTING THE MIGRATION

Some of the main actions

Code & Configuration

- Review and removal of business logic from within Power BI
- Review and restructuring of all ETL code
- Resizing of Databricks clusters
- Design of a brand new medallion architecture

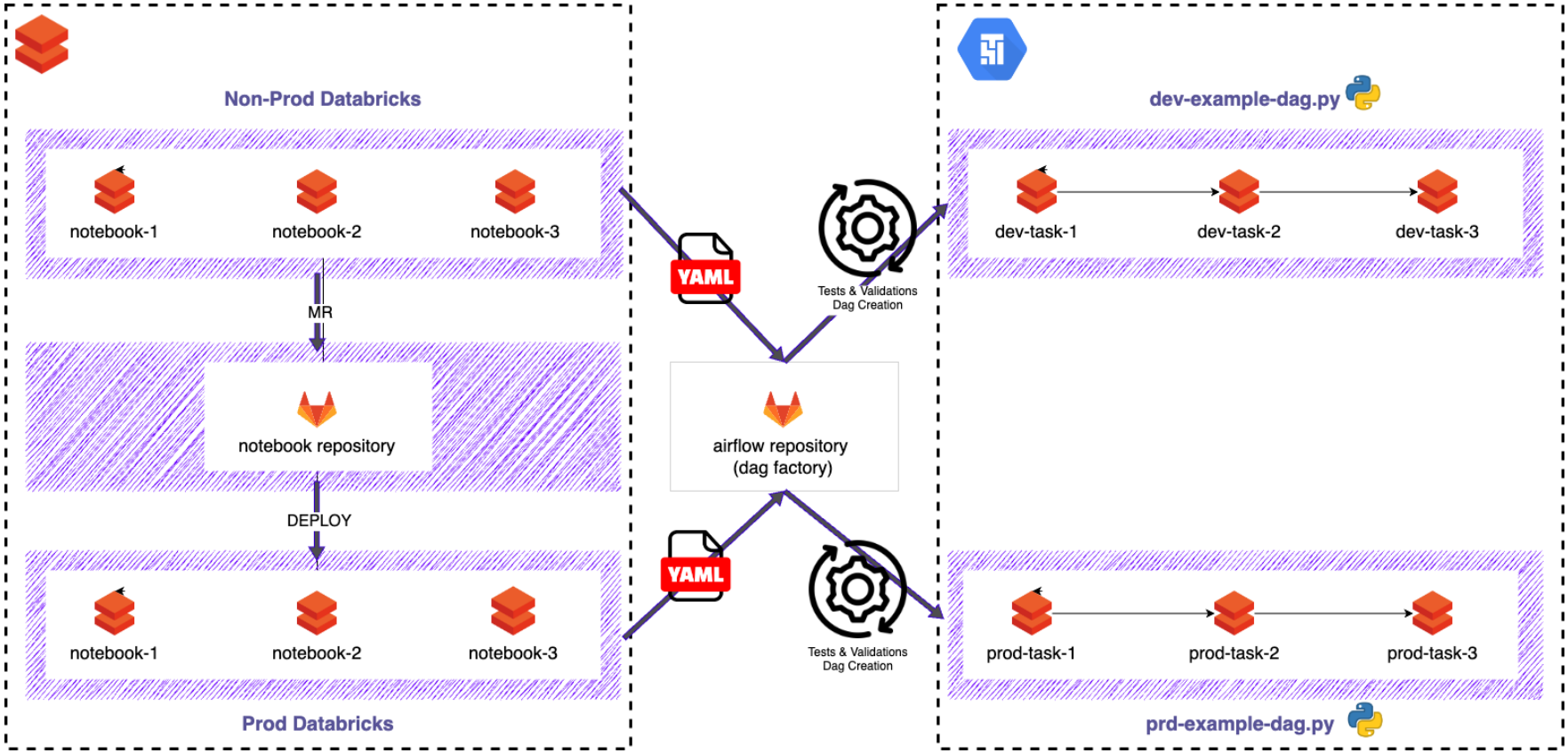
Policies and Standards

- Review of backup routines
- Tagging of clusters
- Creation of cold storage policies
- Establishment of development standards

New Implementations

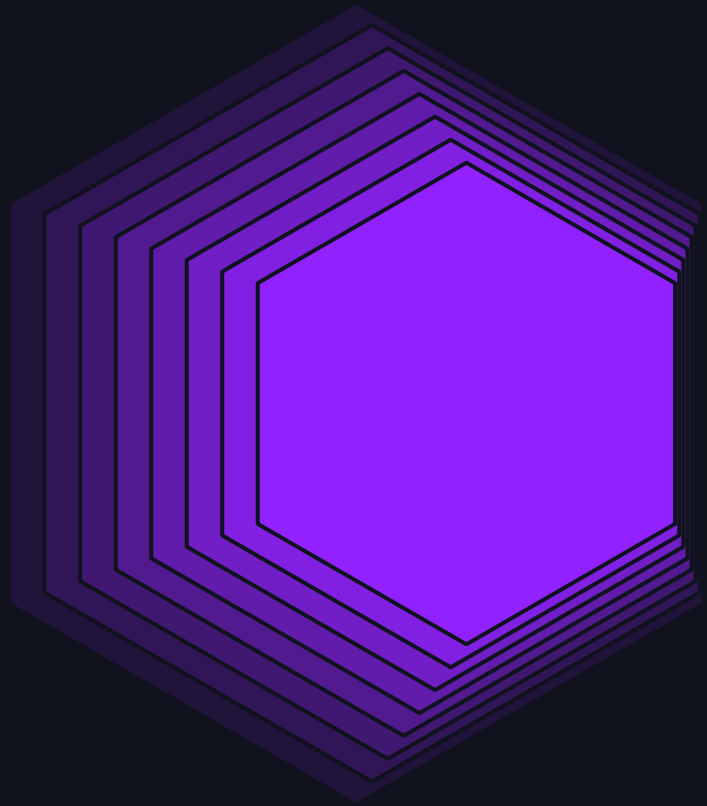
- Implementation of Unity Catalog
- Establishment of CI/CD routines using GitLab
- Incremental data collection and implementation of the Lakehouse strategy

GITLAB CI/CD



DATA+AI SUMMIT BY databricks

RESULTS & COST OPTIMIZATION



Thiago Barcellos
June/2024

MIGRATION RESULTS

MONTHLY COSTS

- Considering just the data tools
- Before Migration: **US$ 48,000/mo**
- Currently: **US$ 8,000/mo**

TIMELINE

- The migration process took **6 months**.
- The team consisted of **5 Data Analysts, 6 Data Engineers** and **1 Tech Lead**.

BENEFITS FROM MIGRATION

Data Governance at a new level

Enhanced Data Governance and Compliance

- Unity Catalog provides a unified interface for managing data access policies, ensuring consistent enforcement of governance across all data assets.
- Data lineage offers complete visibility into data transformations and movements, facilitating audits and compliance with regulatory requirements.

Improved Data Quality and Reliabilty

- Track the origin, transformation, and destination of data, enabling better understanding and trust in data quality.
- Quickly identify and resolve data issues by understanding the data flow and dependencies, leading to more reliable and accurate data.

Increased Productivity and Collaboration

- Simplify the process of granting and managing data access, reducing administrative overhead and improving productivity.
- Enable data scientists, analysts, and engineers to collaborate more effectively by providing a clear understanding of data sources and transformations through detailed lineage information.

BENEFITS FROM MIGRATION

Building a Lakehouse plataform with Delta Lake

Cost Efficiency and Simplified Data Management

- Combines the cost-efficiency of data lakes with the performance benefits of data warehouses, optimizing storage costs while providing robust data processing capabilities.
- Delta Lake's time travel feature allows users to access and query previous versions of the data, simplifying data auditing, troubleshooting, and rollback operations.

Optimized Performance and Scalability

- Delta Lake optimizes query performance through advanced file management techniques which accelerate data retrieval and processing times.
- Supports both batch and streaming data processing, providing a flexible, high-performance environment for real-time analytics and historical data analysis.

Enhanced Data Reliability and Consistency

- Delta Lake supports ACID transactions, ensuring reliable and consistent data operations. This means data integrity is maintained even during concurrent reads and writes, crucial for mission-critical applications.
- Enforces data schema to ensure data quality and allows schema evolution to accommodate changes over time

BENEFITS FROM MIGRATION

Standardization, collaboration, and quality in continuous improvement

Improved Collaboration

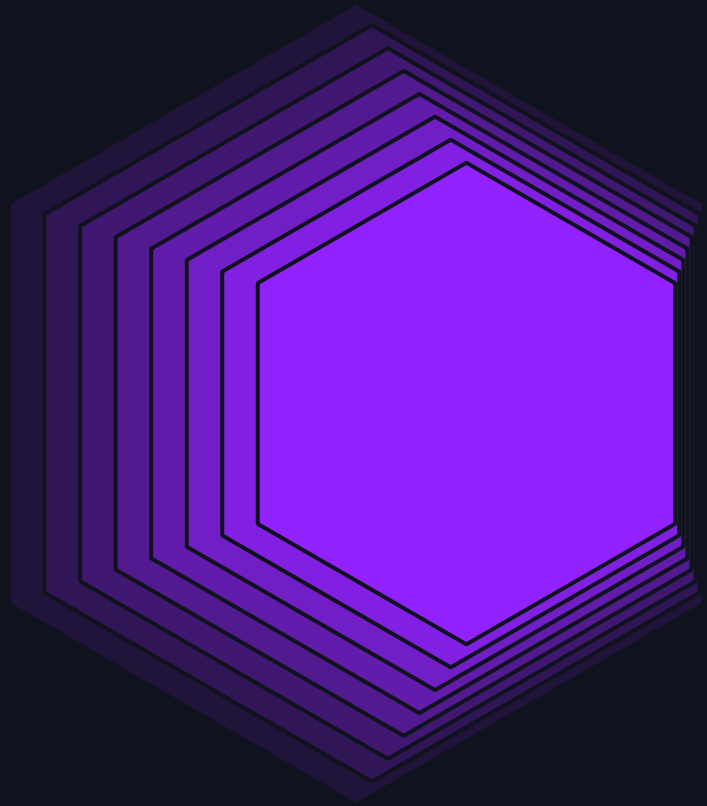
- Unified platform for all data team to collaborate seamlessly.
- GitLab CI/CD integration fosters better version control and collaborative development.

Streamlined CI/CD Pipeline

- Automated testing and deployment processes using GitLab, ensuring higher reliability and faster delivery of solutions.
- Consistent and repeatable workflows, reducing manual errors.

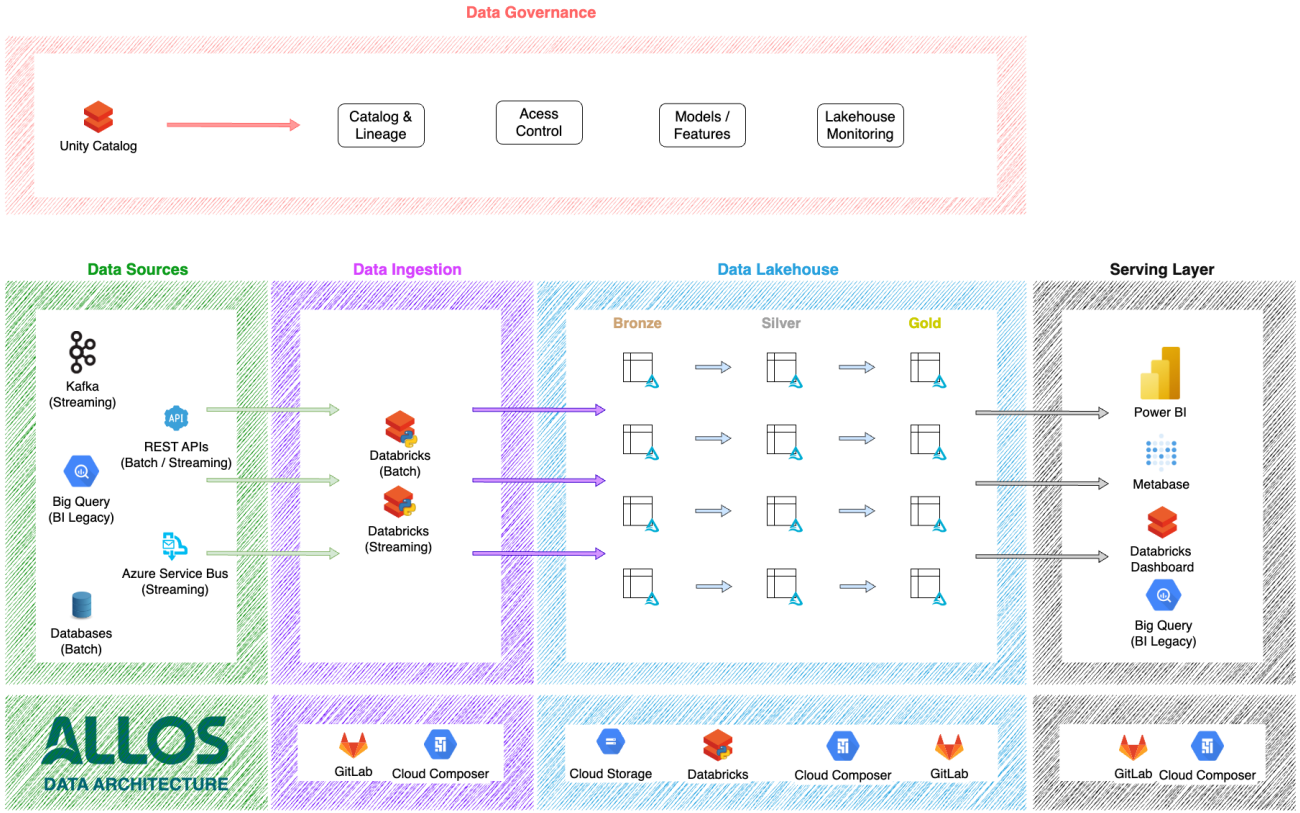
DATA+AI SUMMIT BY databricks

CURRENT DATA ENVIRONMENT



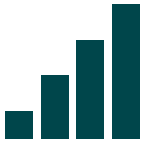
Thiago Barcellos
June/2024

CURRENT ARCHITECTURE



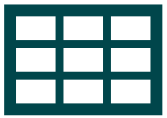
BIG NUMBERS

Considering Digital Environment

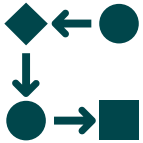


83 TB
Data

74
Catalogs



1273
Tables



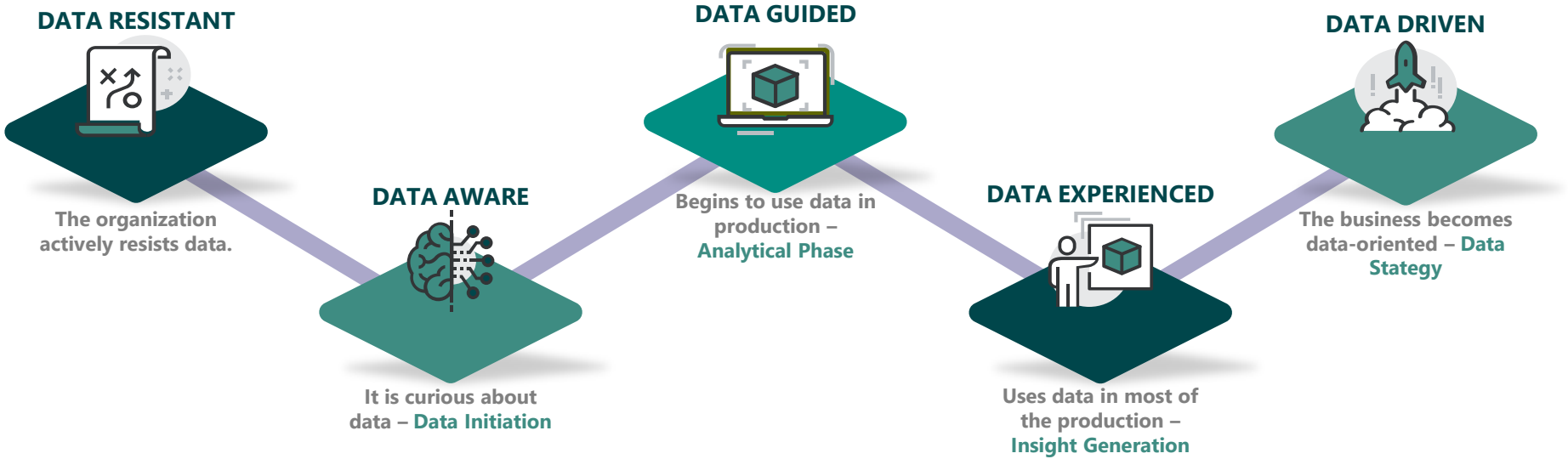
39
DAGs

1341
Reports

915
Users

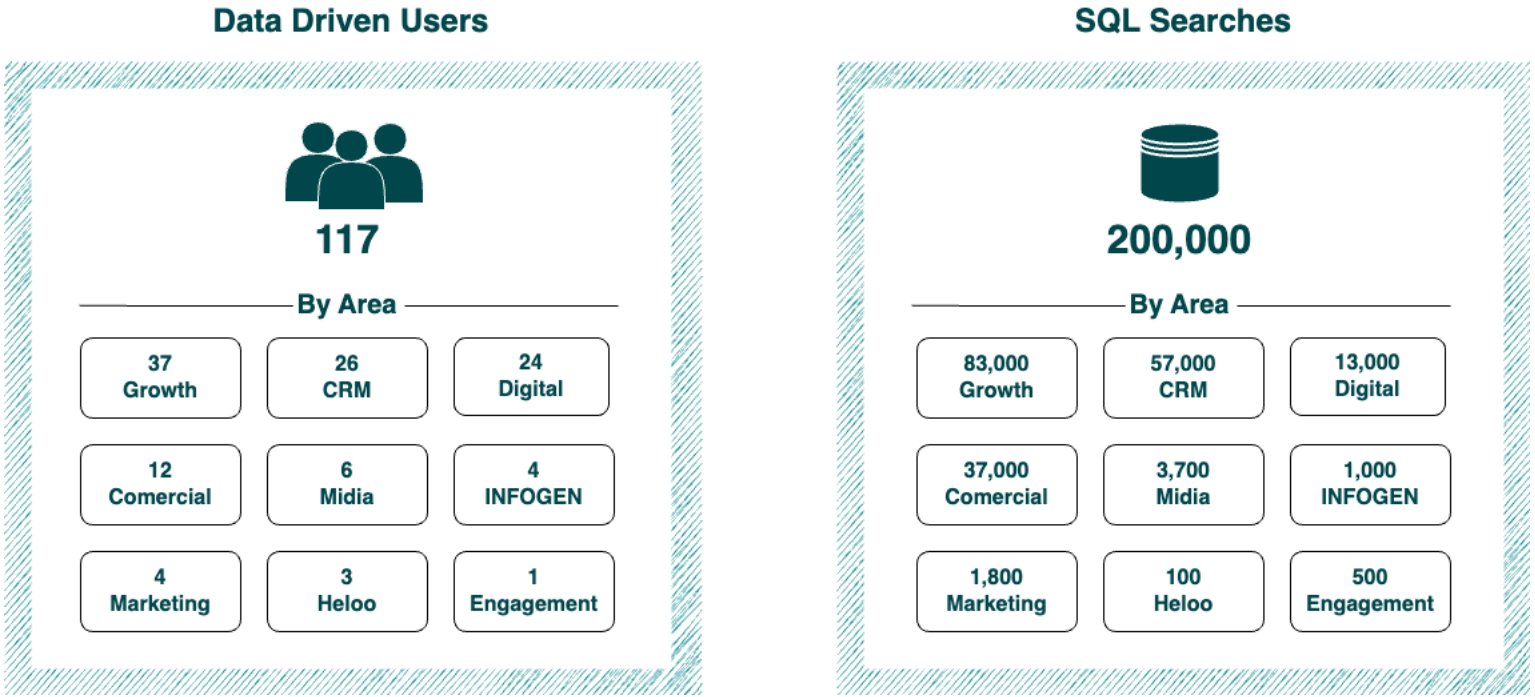
DATA GOVERNANCE MATURITY JOURNEY

With Databricks as one of the main technological levers in this journey



DATA DEMOCRATIZATION

Empowering users using Databricks SQL



* May/2024

CONCLUSIONS

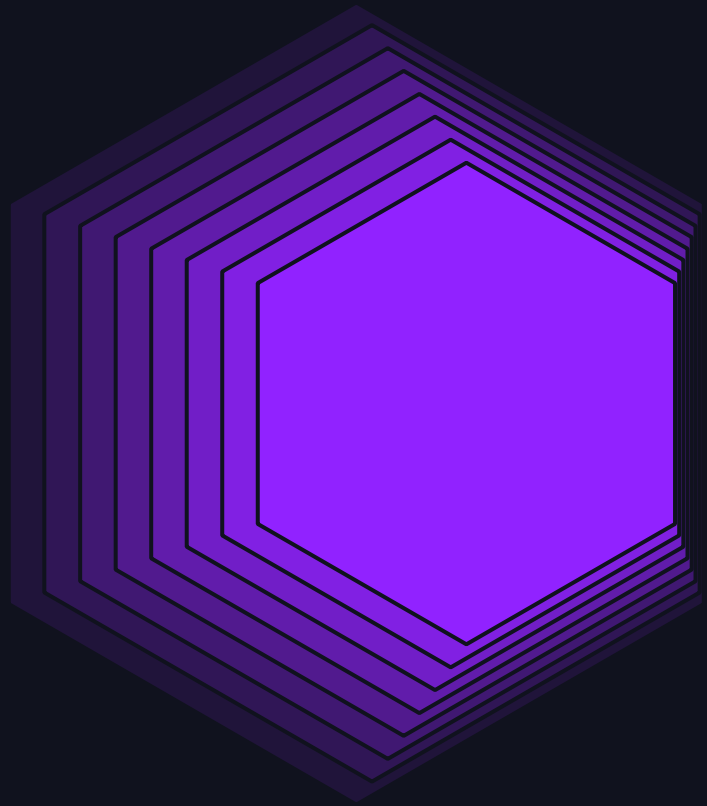
- Do a properly assessment is more important than just choose tool A or B. Avoid guiding the selection process just to include technologies that are currently trending. Resist the "hype of the moment.".

- Communication with all stakeholders, especially business stakeholders, a key success factor for large-scale migration projects

- It was crucial for the project to have the support of a partner like Databricks, who was closely involved throughout the journey and continuously supports the improvement of our team. Additionally, they offer an extensive network of partners and a large, active community of professionals."

DATA+AI SUMMIT BY databricks

LOOKING AHEAD TO THE FUTURE

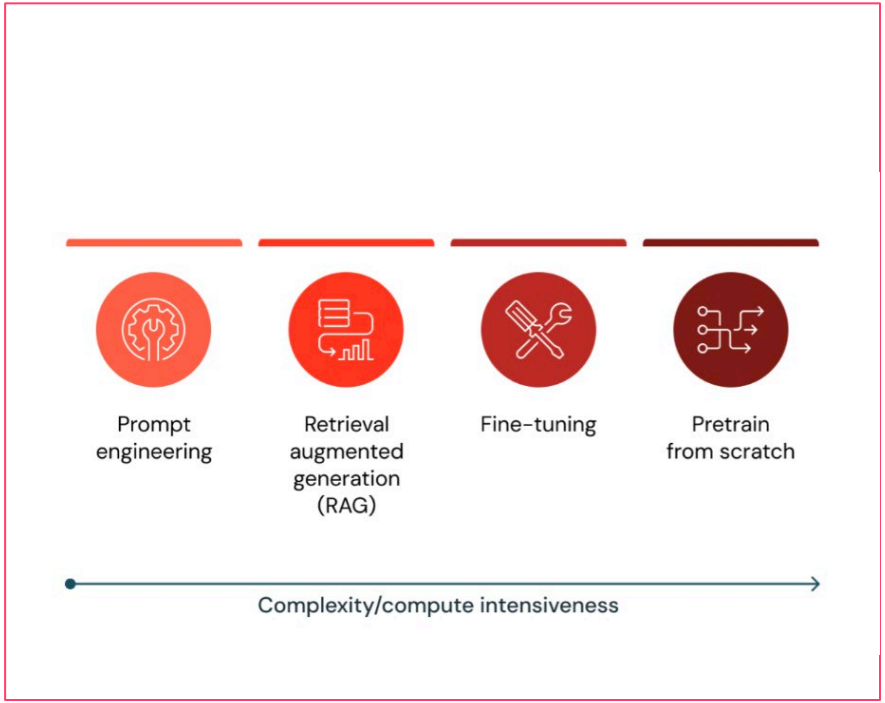


Thiago Barcellos
June/2024

NEXT CHALLENGES

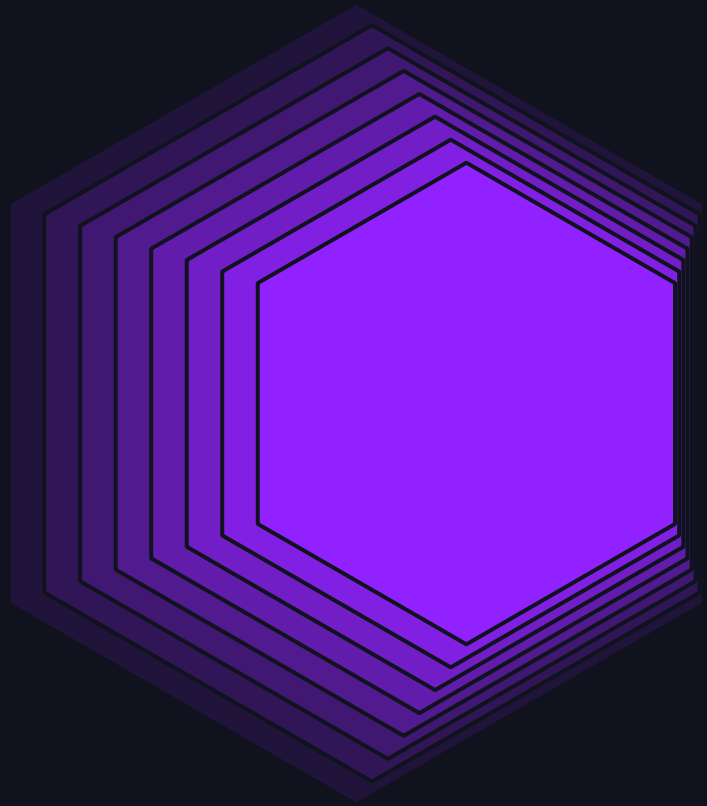
Driving Allos to the future

- Adopt LLM models into the company's decision-making process, driving the use of data and artificial intelligence to enhance the experience of customers and stores.
- Continue increasing efficiency and scale, in line with our mission to be the leading providers of innovation and technology for the shopping center ecosystem



DATA+AI SUMMIT BY databricks

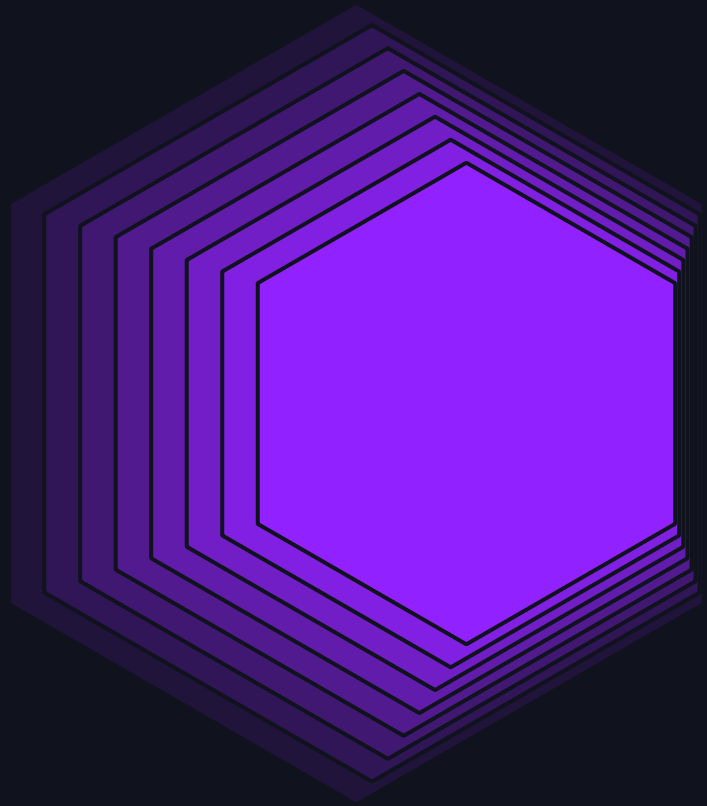
Q&A



Thiago Barcellos
June/2024

DATA+AI SUMMIT BY databricks

THANK YOU!



Thiago Barcellos
June/2024