Deloitte.

Takeda

databricks

Migrating from Databricks PVC to Databricks E2

Platform Migration Approach

MAKING AN
IMPACT THAT
MATTERS
since 1845
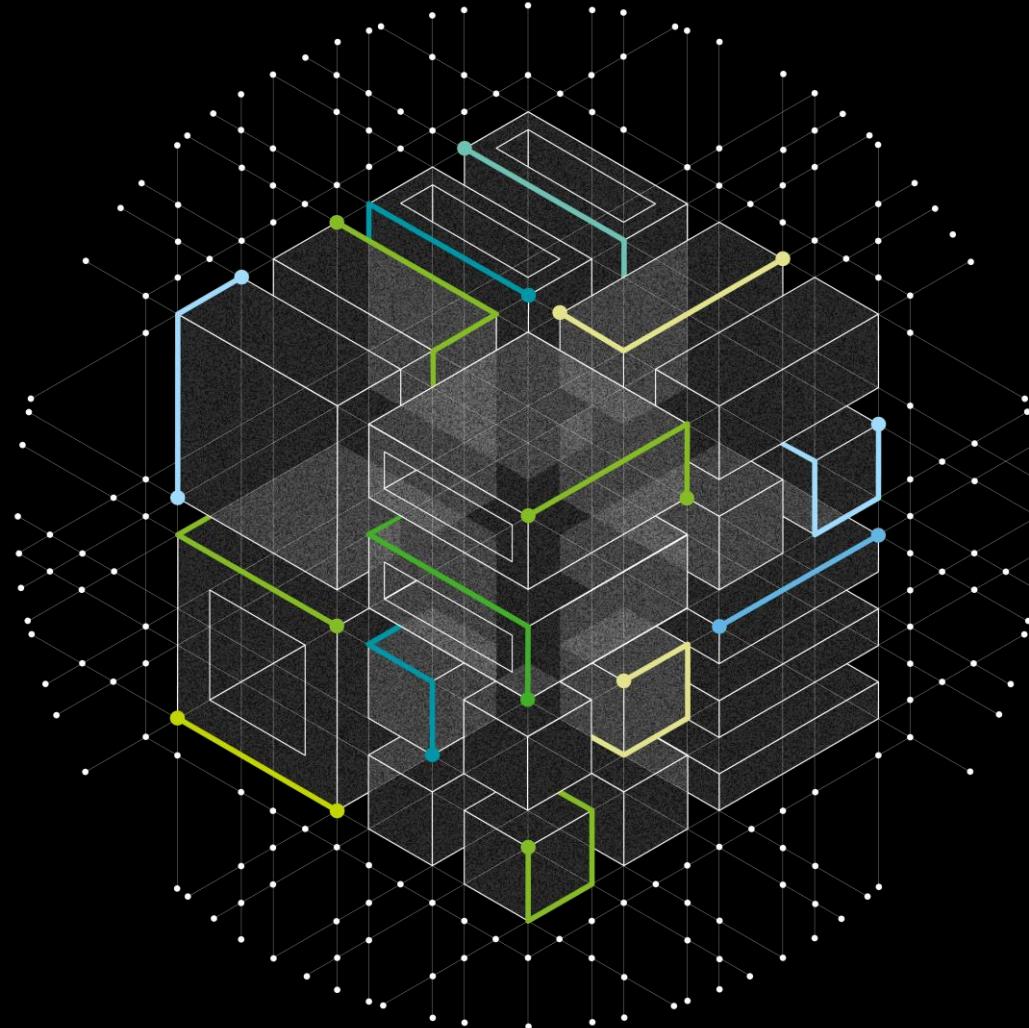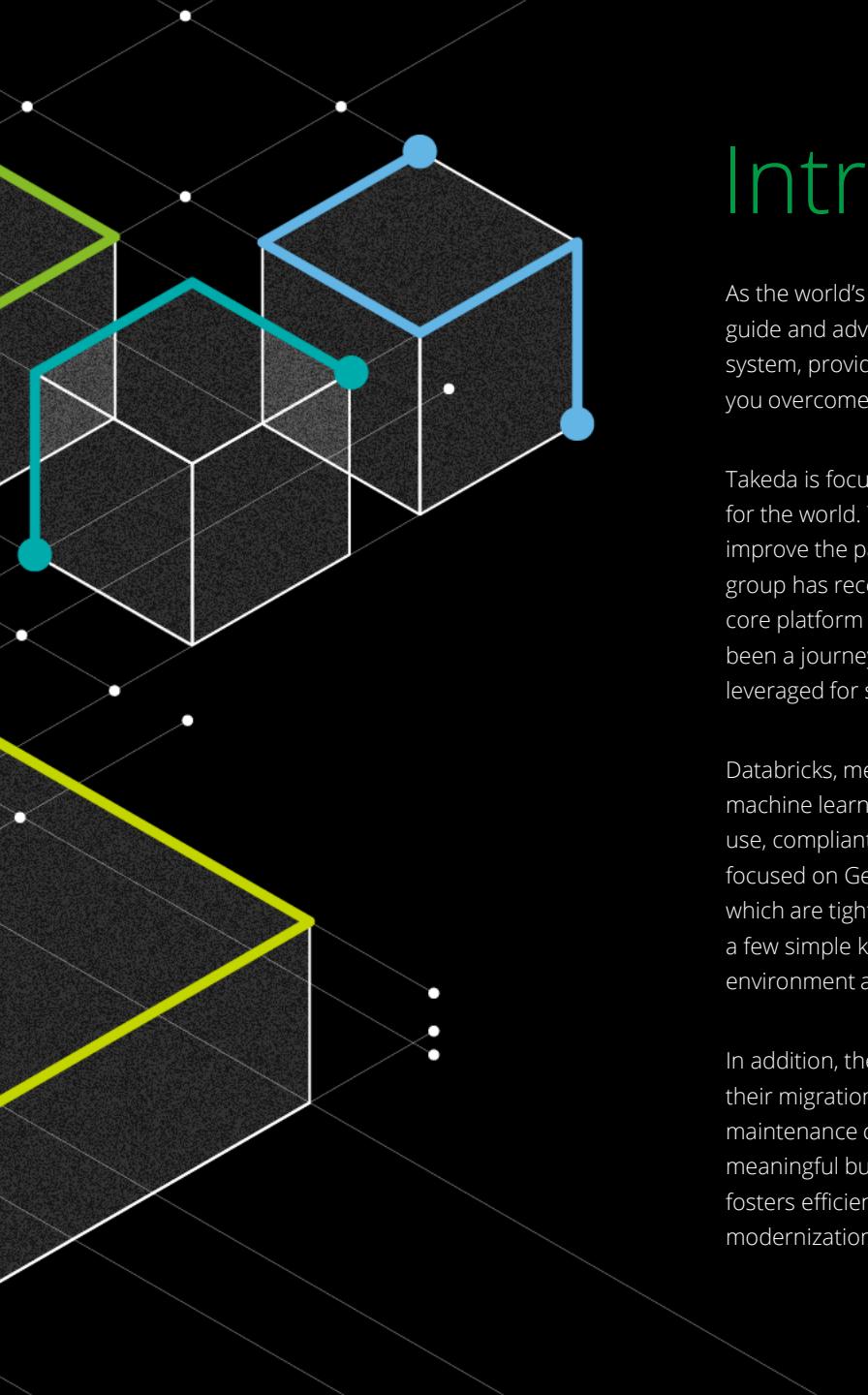
Introduction

As the world's leading consulting firm, Deloitte is uniquely positioned to guide and advise your team as you journey toward a modernized data system, providing expert insight and implementation know-how to help you overcome any roadblock.

Takeda is focused on creating better health for people and a brighter future for the world. Takeda has been focused on technology-driven innovation to improve the patient experience. Their EDAS (enterprise data and services) group has recently completed DBX platform upgrade to E2 which serves as core platform to support data and analytics use-cases. The program has been a journey with lots of learnings and best practices that can be leveraged for similar work.

Databricks, meanwhile, equips organizations with advanced analytics, AI, machine learning, and other emerging data technologies within its easy-to-use, compliant platform. Databricks also provides the latest innovations focused on Generative AI and Large Language Models, such as MosaicML, which are tightly integrated in their Databricks Data Intelligence Platform. In a few simple keystrokes, organizations can quickly set up a Databricks environment and avoid intensive build processes.

In addition, the platform's pricing model further enables teams to accelerate their migration journey without upfront investments in infrastructure and maintenance costs, allowing them to deliver actionable insights and drive meaningful business outcomes. Databricks is an end-to-end solution, which fosters efficiency, cost reduction, and multi-cloud capabilities—powering modernization and growth for teams across industries.

As you consider migration from Databricks PVC (Private Virtual Cloud) to Databricks E2 (Enterprise 2.0), this can present a set of unique challenges. This paper discusses the methodologies to be employed, the challenges, and their corresponding solutions during the migration process.

A large-scale platform migration requires careful understanding of the existing PVC infrastructure and its dependencies. These complexities could lead to difficulties in listing and cataloging the PVC resources and planning for a smooth migration to the E2 environment. To address these challenges, it is necessary to conduct an extensive analysis and documentation of the PVC environment. This involves creating an inventory of all existing resources, their configurations, and interdependencies while working with multiple business units (BU's), application owners, and stakeholders of an enterprise. A detailed impact assessment outlining application-level complexities should be performed as part of the discovery phase. The E2 environments should focus on architectural best practices to enable improved Data Access & Data Governance while preserving existing functionality and business continuity. Change management activities must be prioritized to keep business stakeholders and end users informed and address any technical and business challenges that may come their way throughout the migration process.

Additionally, thorough testing and validation must be planned and executed to ensure compatibility and functionality between the PVC and E2 environments are meeting the business expectations.

Drivers for Migration (1/2)

While PVC has served as a reliable and powerful platform, the need for migration from Databricks PVC to Databricks E2 (Enterprise 2.0) is driven by several factors.

1

End of Support

PVC is considered a legacy version of Databricks, and it is approaching the end of its support lifecycle. As a result, organizations need to transition to a newer version, such as Databricks E2, to ensure they receive continued support, security updates, and bug fixes.

2

Lower Costs and Total Cost of Ownership (TCO)

Databricks E2 offers cost advantages and a lower TCO compared to PVC. E2 provides optimized resource allocation, scalability, and performance, resulting in reduced infrastructure and operational costs. By migrating to E2, organizations can benefit from better cost efficiency and allocate resources more effectively.

3

Access to Latest Features and Improvements

Databricks E2 brings a host of new features and improvements that includes Unity Catalog for managing and organizing data assets, Serverless SQL for on-demand SQL queries, SQL Warehouses for optimized data warehousing, Delta Sharing for secure data collaboration, Generative AI and more. Migrating to E2 enables organizations to leverage these advanced capabilities and stay ahead in their data processing and analytics endeavors.

4

Enhanced Efficiency for Data Warehouse Workloads

Databricks E2 is specifically designed to enhance efficiency for data warehouse workloads, including business intelligence (BI) and SQL tasks. It provides improved query performance, optimization techniques, and caching mechanisms, enabling faster and more efficient data analysis. Organizations migrating to E2 can benefit from enhanced data processing speed and improved productivity in their data warehouse workflows.

Drivers for Migration (2/2)

While PVC has served as a reliable and powerful platform, the need for migration from Databricks PVC to Databricks E2 (Enterprise 2.0) is driven by several factors.

5

AutoML and Feature Store

Databricks E2 offers AutoML capabilities for classification, regression, and forecasting tasks. It simplifies the process of developing and deploying machine learning models by automating various stages. Additionally, E2 provides a feature store and Generative AI capabilities, allowing organizations to centralize and manage their feature data effectively. By migrating to E2, organizations can leverage these advanced machine learning capabilities and streamline their data science workflows.

6

SaaS Service with Expert Support

E2 is a fully managed SaaS service with 24/7 monitoring and expert support provided by Databricks, ensuring organizations receive the necessary assistance and guidance for their data operations.

7

Simplified Managed Stream

E2 simplifies the process of building managed streaming pipelines, making it easier for organizations to create end-to-end streaming data workflows and process real-time data efficiently

8

Roadmap for Budgeting, Governance & Lineage

E2 has a near-term roadmap that includes upcoming features for budgeting, data governance, and data lineage, enabling organizations to align their data operations with future initiatives and benefit from enhanced budget control, data compliance, and traceability.

Migration Approach

The Databricks E2 migration process is divided into distinct stages, each playing a crucial role in the success of migration workloads.

Migration Approach

PVC to E2 migration process is split into 6 different stages

MIGRATION STAGES

Assessment

Architecture Evaluation

Platform Setup

Workload Migration

Data Migration

Hypercare Support

Assessment

Assessment

Architecture Evaluation

Platform Setup

Workload Migration

Data Migration

Hypercare Support

MIGRATION STAGES
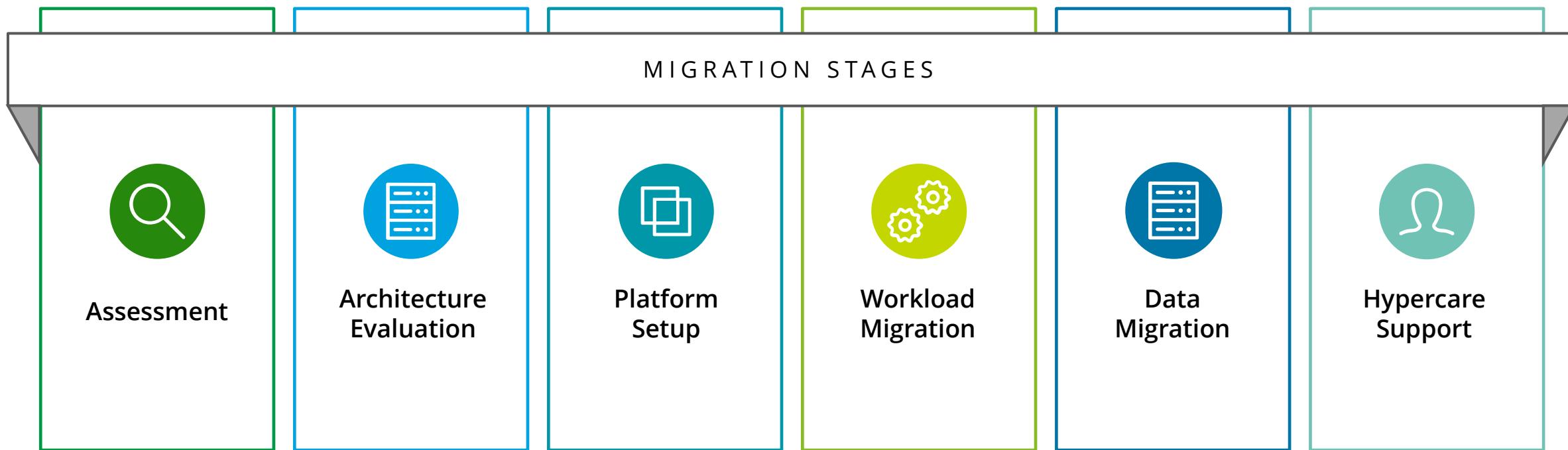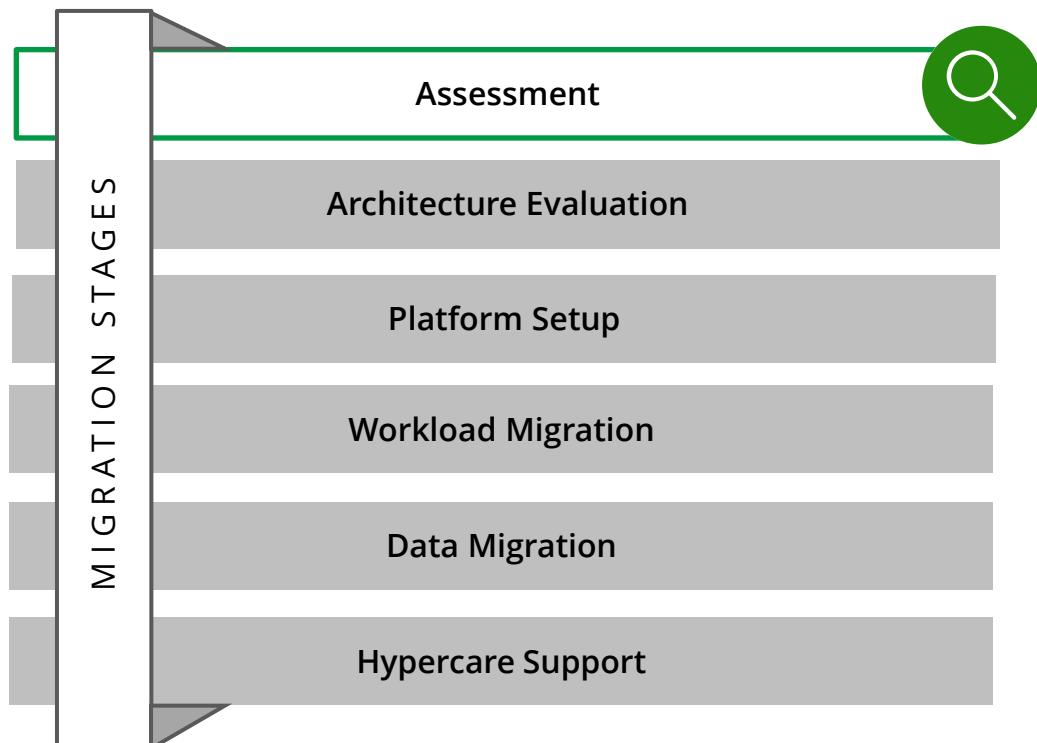
- Detailed inventory can be captured using Databricks APIs, tools, scripts in discovery stage. Inventory includes list of clusters, jobs, user/shared workspaces, notebooks, ACLs, schemas, tables, service accounts, etc.
- Dependency analysis shall be performed to enlist all the platforms/services interconnected with Databricks (like BI tools, external databases, orchestration tools, data ingestion pipelines, ETL tools). Requirements for connectivity of these services in E2 need to be captured
- The E2 Upgrade Team initiates migration process by identifying the owners of the workloads to be migrated to Databricks E2
- Collaborative discussions take place between the business units (BUs) and the E2 Upgrade Team to define the scope of the migration, ensuring a clear understanding of which workloads will be included
- Databricks solution needs to be designed for cluster, user, group and notebook migration

Architecture Evaluation

MIGRATION STAGES

- Assessment
- Architecture Evaluation
- Platform Setup
- Workload Migration
- Data Migration
- Hypercare Support
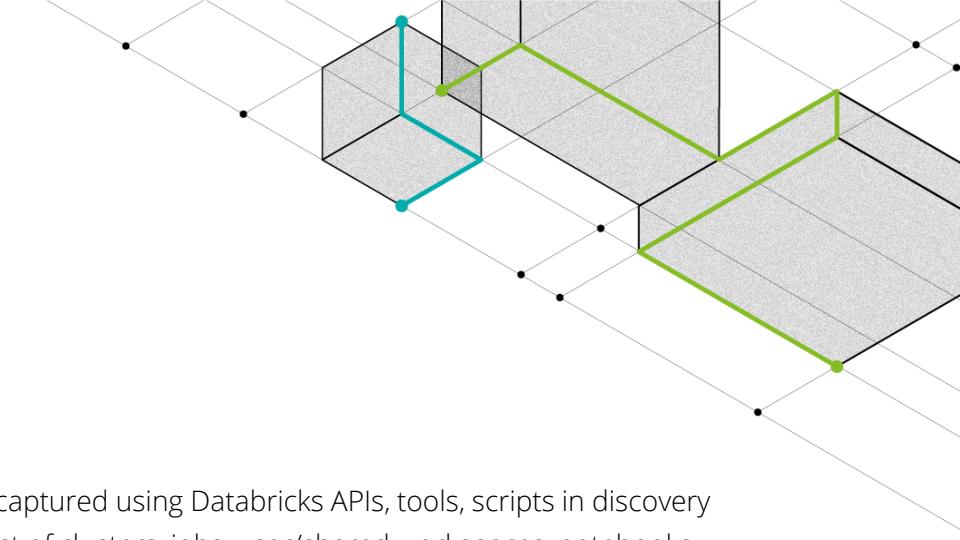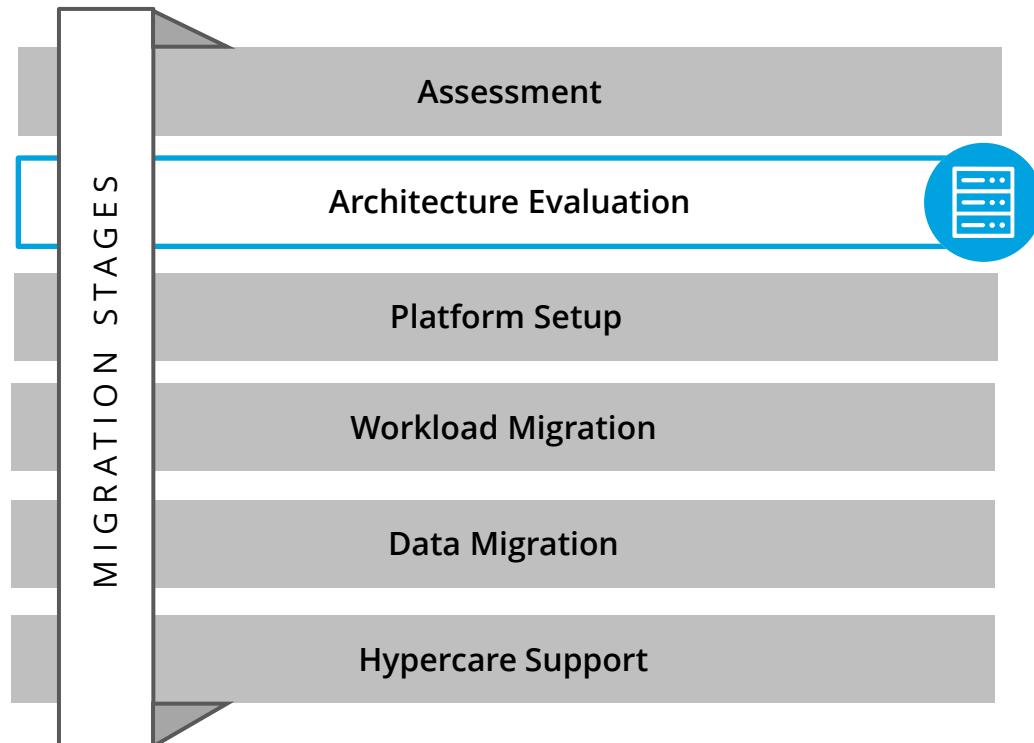
- **Workspace:** To fully leverage all the benefits of E2, and support future growth and manageability, workspace layout is carefully architected. Number of workspaces should be minimized and created when separation is necessary for compliance, isolation, or geographical constraints (single workspace vs. workspaces based on environment vs. workspaces based on line of business)

- **SSO/SCIM:** Single sign-on integration and user groups, based on personas and access levels are to be defined and implemented

- **Private Link:** To ensure enterprise compliance, cloud-native private-only connectivity for both front-end and back-end interfaces of Databricks workspaces is evaluated and implemented, if required

- **Governance:** Best practices around naming, tagging, chargeback, CICD, deployment, etc., are defined

- **Catalog:** Unity catalog in E2 provides centralized access control, auditing, lineage, and data discovery capabilities. As part of upgrade, if migrating to unity catalog, key stakeholders should be identified to develop the migration plan (discovery of assets, setting up unity catalog, planning the migration, executing, and retiring hive metastore)

Platform Setup

MIGRATION STAGES

- Assessment
- Architecture Evaluation
- Platform Setup
- Workload Migration
- Data Migration
- Hypercare Support
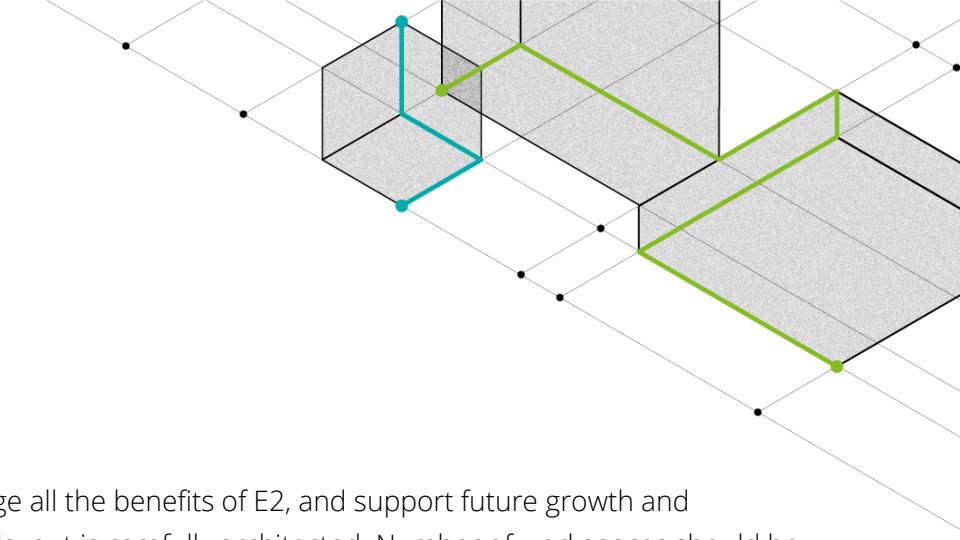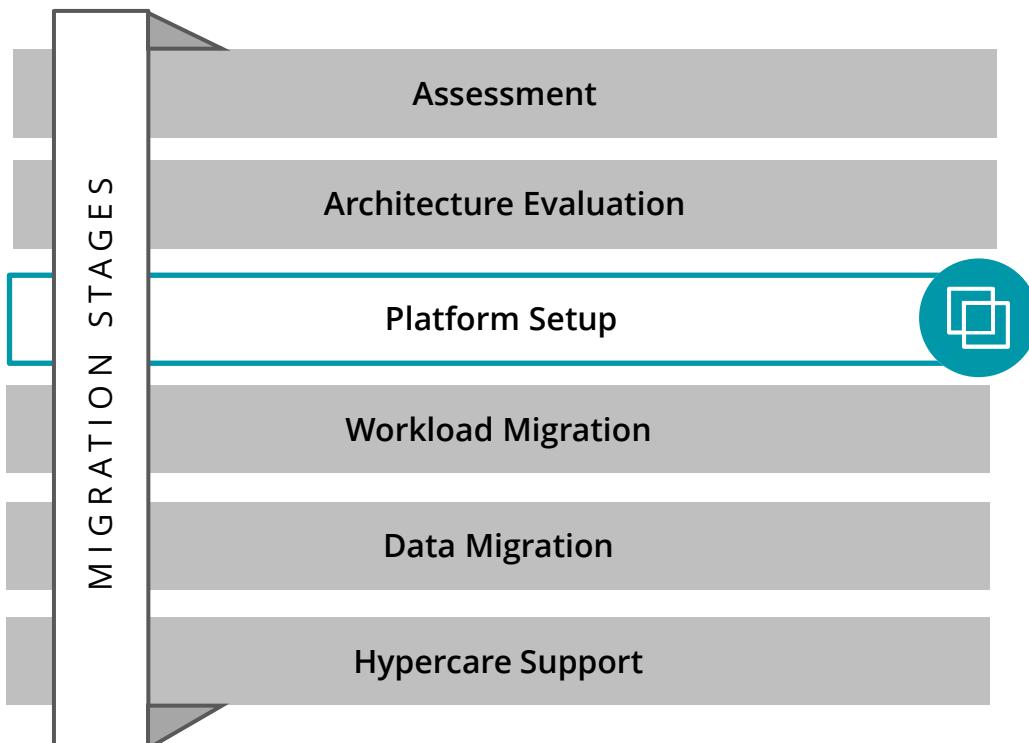
- **Architecture:** E2 environments are architected based on Databricks best practices (for example, distinct workspaces per region/environment) and addressing all the shortfalls/pain points of PVC environments (for example, right sizing the subnets based on projected growth). Based on enterprise' set-up, E2 workspaces can be configured in the existing AWS accounts of PVC or in new dedicated AWS accounts with cross-account access configurations

- **AWS Infra Build:** The foundational AWS infrastructure is to be constructed following DevOps best practices, including Infrastructure as Code (IaC), continuous integration and continuous deployment (CI/CD) pipelines, and code repositories for version control and code management. Leveraging various AWS services such as CodeCommit, CodeBuild, CodeDeploy, CodePipeline, CloudFormation, and more, ensures efficient and streamlined development and deployment processes

- **Databricks Workspace Build:** Multiple Databricks environments are configured utilizing Terraform (IaC), a configuration management tool, to capture the desired configuration for the Databricks E2 environment, including infrastructure setup and resource configurations

- **Business Continuity:** E2 environments can be created as parallel environments accessing the common datalake. This enables PVC and E2 environments to be functional at the same time, accessing the same data while the migration is underway

Workload Migration

MIGRATION STAGES

- Assessment
- Architecture Evaluation
- Platform Setup
- **Workload Migration**
- Data Migration
- Hypercare Support
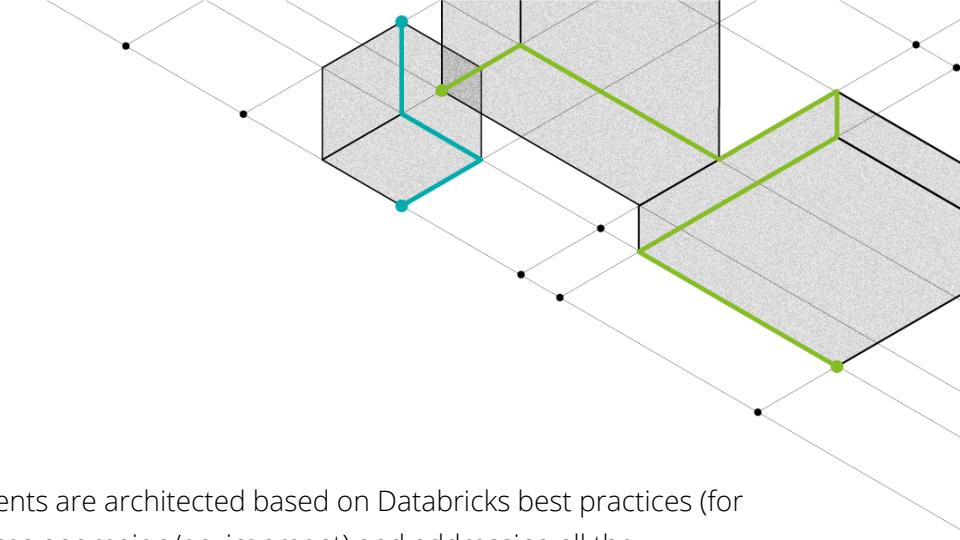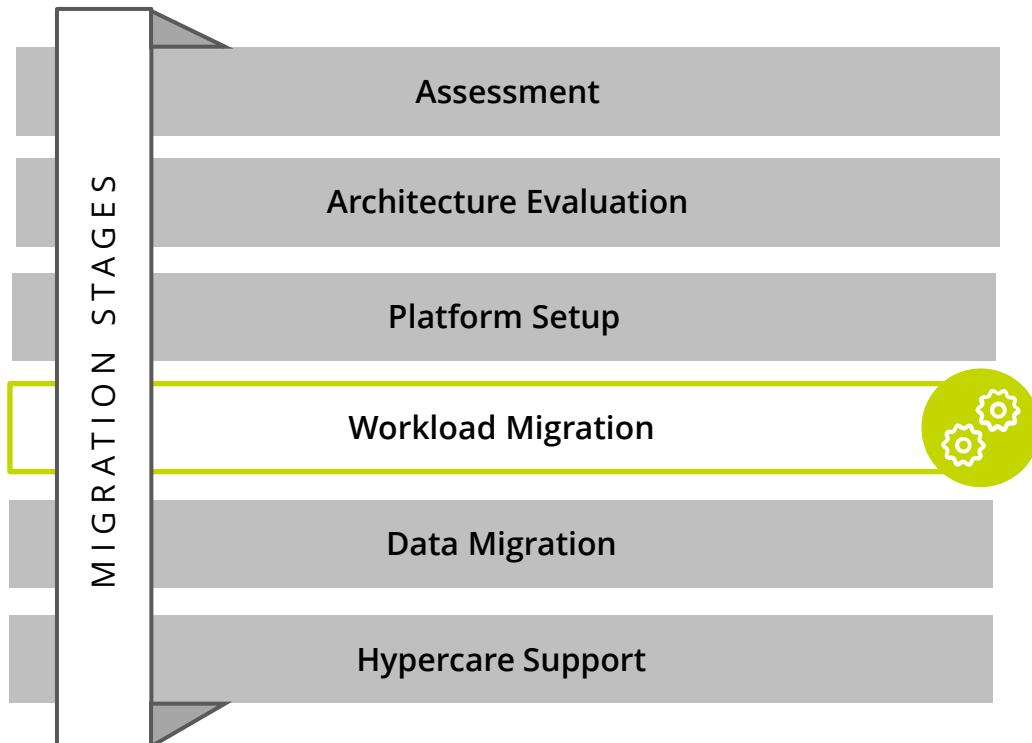
- The Terraform configurations need to be modified and thoroughly tested to ensure they accurately reflect the intended state of the Databricks E2 environment Point of Contact (POC) from BUs and the Migration Unit validate and confirm that all necessary inventory has been successfully migrated and is functioning as expected
- Close coordination is established with the workflow orchestration and application teams to update the job connections, ensuring seamless integration with the new Databricks E2 environment
- A sign-off shall be obtained for the future state environment, indicating that the migration has been completed successfully. A dedicated hyper care team provides support during this phase, addressing any post-migration issues that may arise

Workload Migration Contd.

MIGRATION STAGES

- Assessment
- Architecture Evaluation
- Platform Setup
- **Workload Migration**
- Data Migration
- Hypercare Support
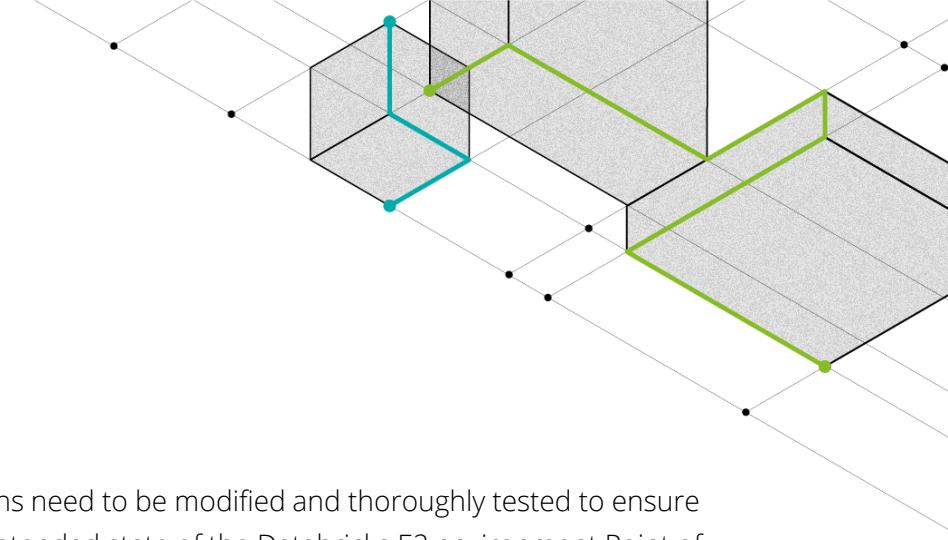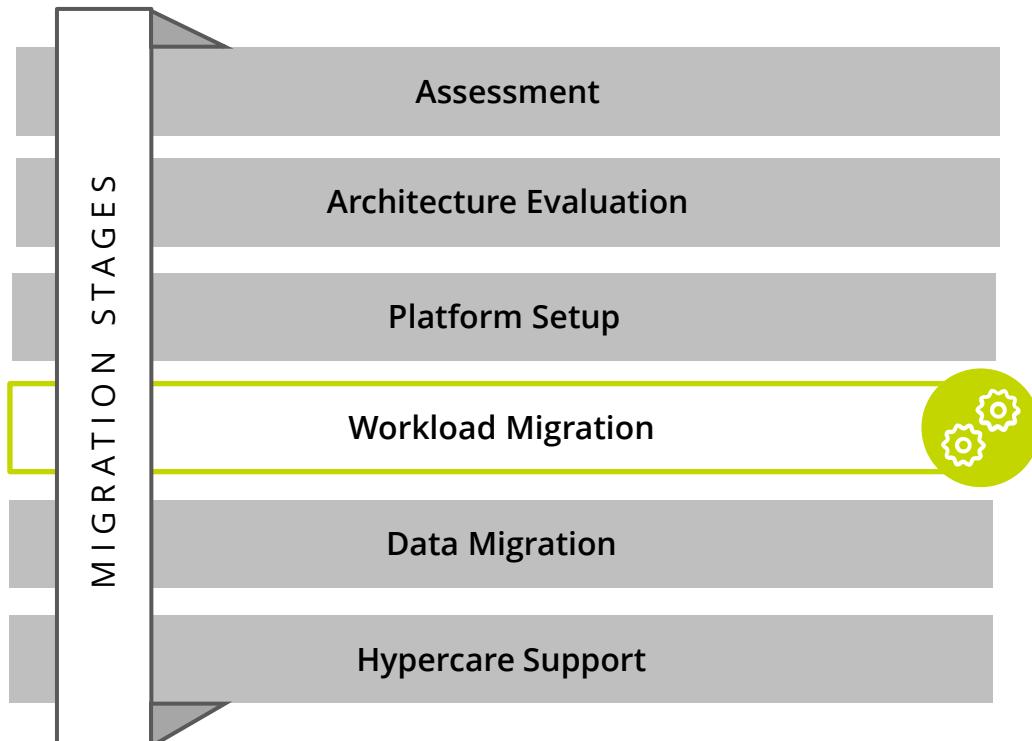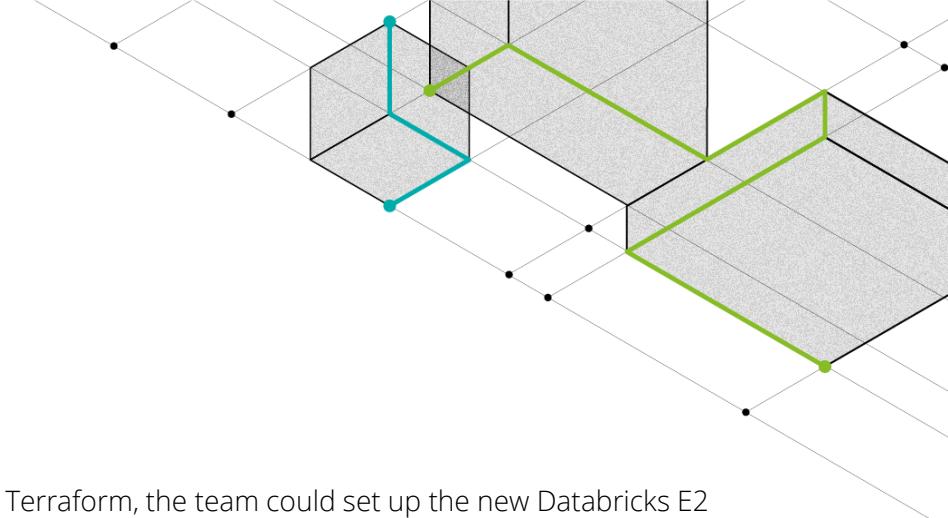
- **Version Control:** By utilizing Terraform, the team could set up the new Databricks E2 infrastructure while ensuring version control of the code used to establish the required resources. This meant that any modifications or updates made to the infrastructure could be efficiently tracked, providing a comprehensive history of changes.
- **CI/CD Pipelines:** Adopting Terraform for building the resources offers the opportunity to develop and integrate CI/CD pipelines in the future. This approach would facilitate a streamlined and automated workflow for deploying, testing, and releasing infrastructure changes. With CI/CD pipelines, managing the Databricks environment becomes more efficient and reliable.
- **Comprehensive Inventory:** Terraform allows meticulous tracking of all the objects that needs to be migrated. By utilizing the tool, the team could gain a clear and comprehensive view of the resources and their dependencies involved in the migration. This ensures that no critical components were overlooked or left behind during the transition from Databricks PVC to Databricks E2.

Data Migration

MIGRATION STAGES

- Assessment
- Architecture Evaluation
- Platform Setup
- Workload Migration
- **Data Migration**
- Hypercare Support
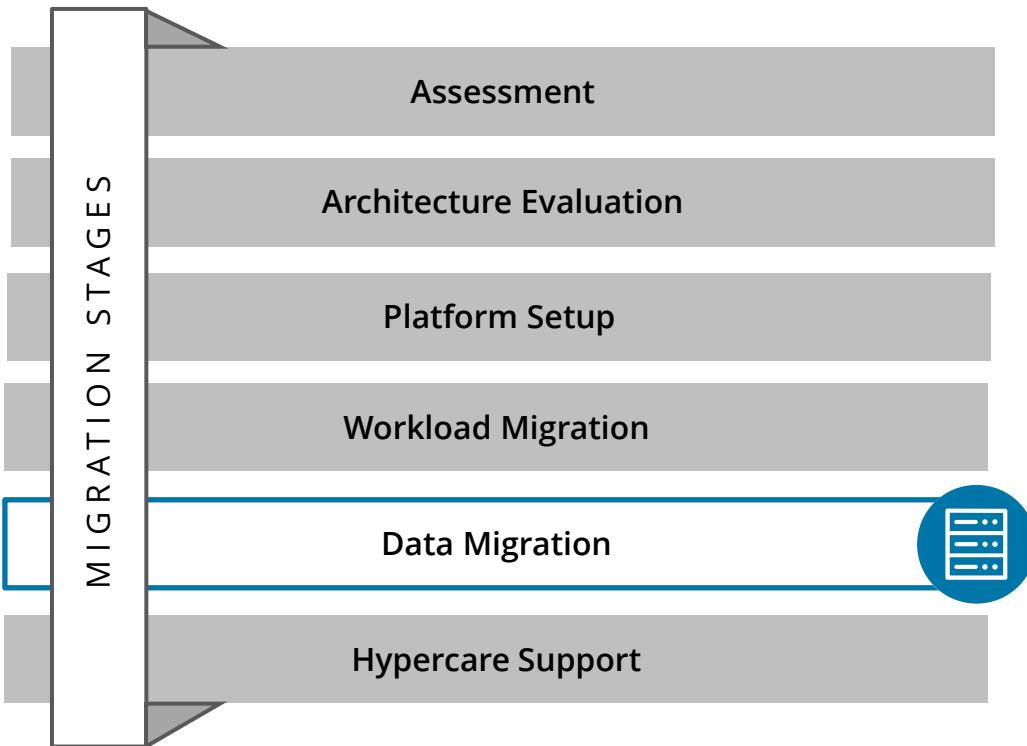
As part of upgrade to E2, it is preferred not to migrate the data layer and only upgrade the compute, to expedite the upgrade process. Migration from HMS to unity catalog should be considered post the upgrade is complete.

Customers can choose to create a parallel E2 compute environment within the existing AWS account or in its own dedicated new AWS account. As part of this process, customers can leave the data layer intact in its existing location. The data layer comprises of elements such as S3 buckets, tables, and schemas. Below picture shows the setup with shared data layer.

Databricks PVC   Data Layer   Databricks E2

Clusters   Notebooks   Jobs

Databricks Workspace

Hypercare Support

MIGRATION STAGES
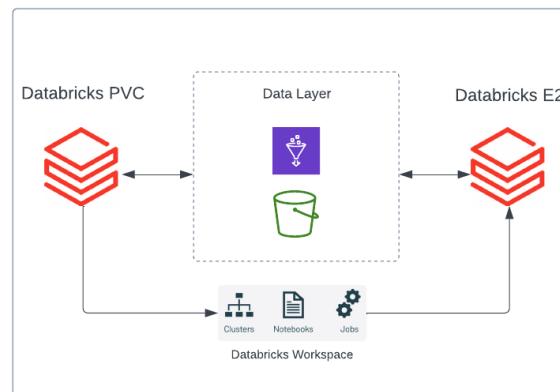
- Assessment
- Architecture Evaluation
- Platform Setup
- Workload Migration
- Data Migration
- Hypercare Support
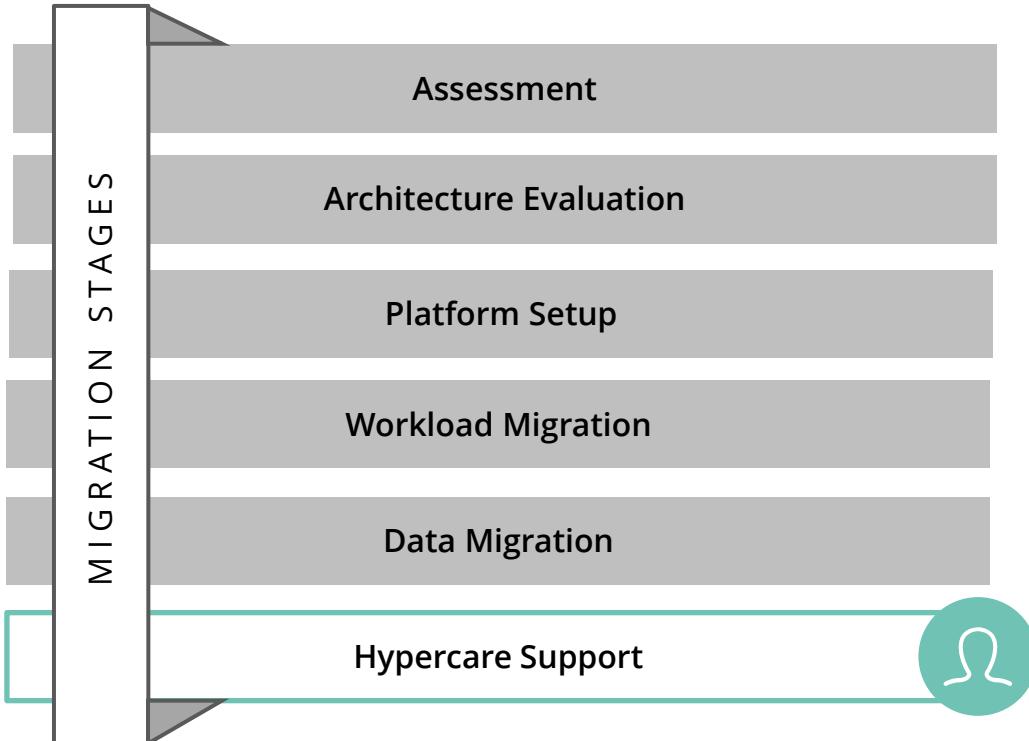
To ensure a smooth Databricks E2 migration, having a dedicated hypercare support team in place is crucial to swiftly address any potential issues during the project implementation. It's advisable for the current platform operation team to maintain their focus on managing service requests related to PVC instances. They can triage tickets and, based on categories, route all E2-related tickets to the E2 hypercare team. This allows the dedicated team to effectively handle issues arising from E2 instances.

As each Business Function or business unit undergoes migration to the E2 environment, consider transitioning the E2 hypercare support responsibilities to the current platform support team. While this process seems straightforward, it demands robust collaboration skills, particularly in cases involving multiple vendor teams, to ensure its success.

Fusion Support Team During Migration
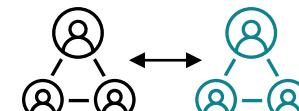
PVC Tickets   E2 Tickets

Fusion Support Team Post Migration to E2

Handover Support

Hypercare Support Contd.

MIGRATION STAGES

- Assessment
- Architecture Evaluation
- Platform Setup
- Workload Migration
- Data Migration
- Hypercare Support
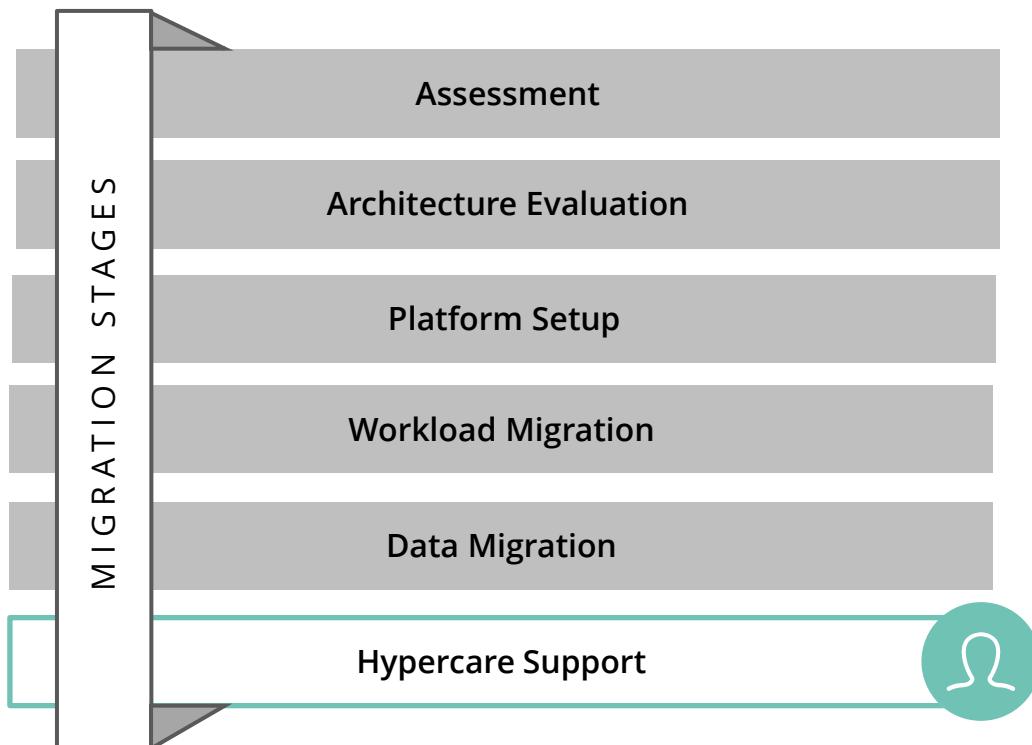
1. **Pre-Cutover Sync**: Begin with a synchronization process that includes notebooks, workflows, and ML experiments. This ensures seamless pausing of jobs/workflows in PVC and their simultaneous resumption in E2 environments.

2. **SDLC-Aligned Workflow Transition**: Follow the SDLC phases (Dev-Test-Prod) when shifting workflows from PVC to E2. This phased approach guarantees a smooth overall transition and allows support teams sufficient time to respond to any unexpected events.

3. **Re-Sync Strategy**: Implement re-syncs for notebooks in scenarios where lower environment cutover activities are finished, and updates occur in the PVC production environment just before the cutover. Minimize such occurrences through effective planning and collaboration with application teams.

4. **Cutover & Decommission Protocol**: After the hypercare phase is completed, plan to decommission PVC workloads one month later. This involves retiring clusters and jobs, marking the conclusion of the migration process, and ensuring full transition of workloads to Databricks E2.

5. **Collaborative Execution**: Ensure close collaboration among the E2 Upgrade Team, Business Units (BUs), Proof-of-Concept (POC) teams, Tidal, application teams, and the hypercare team throughout each migration phase. This collaborative effort guarantees a smooth and successful transition to Databricks E2.
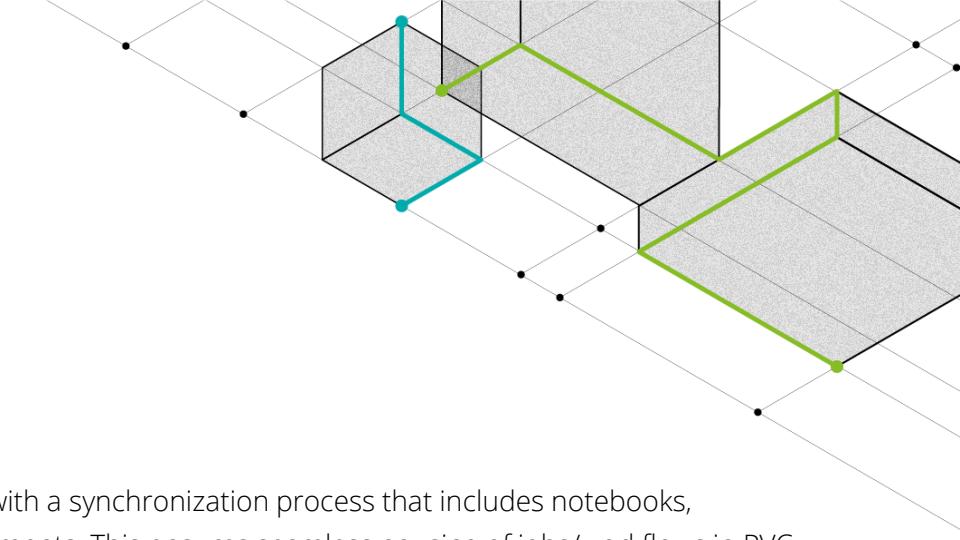
Common Challenges

During the migration effort from PVC to E2, one can run into several challenges, if not planned accordingly. This section includes a few of the most common ones, along with steps for mitigating.

1. Enhanced Security and Restrictive Access in E2

During the migration process from Databricks PVC to Databricks E2, different challenges pertaining to the access of schemas/tables could be encountered

- **Design decision to enhance security:** Upgrading to E2 provides the opportunity to enhance data access policies and implement governed access control mechanisms. Gaps in least privileged access mechanisms of existing PVC , if discovered, can be addressed while upgrading to E2. This could impact applications and user experience.

- **Absence of access privileges:** As part of E2 migration, access permissions on certain schemas and data sources can be missed due to gaps in inventory during requirements and solutioning phase.

Security

Restrictive Access
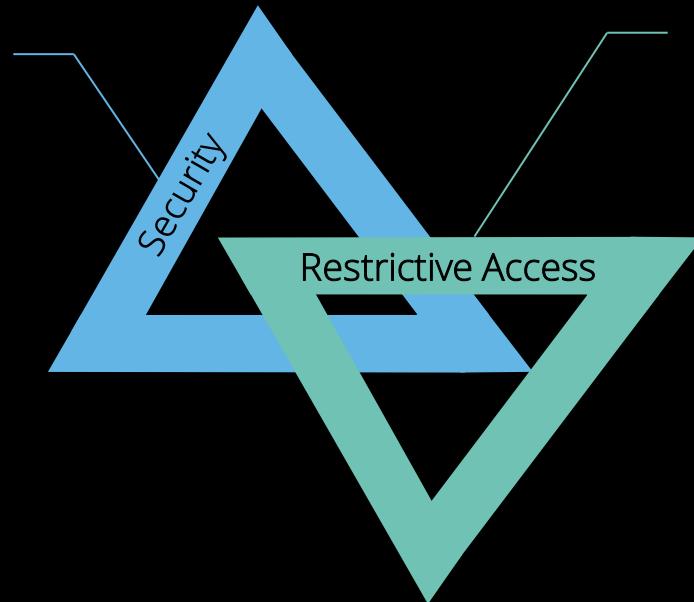
- **Access requests beyond authorized scope:** Certain users, based on previous experience with more permissive PVC environments having access to schemas and tables outside of their designated authorization level, could object to newer security enhancements.

- During the migration process, all such issues need to be identified and addressed. This requires close coordination with the application teams to ensure they are able to continue to achieve their business goals, while having the least access/privileges possible. Any security enhancements should be explained and communicated to application teams before the upgrade.

2. AWS Cross Account Access

In the process of upgrading to Databricks E2, encountering cross-account issues is a possibility, particularly if the E2 compute plane is established in a new AWS account, distinct from PVC

- **IAM roles:** IAM roles pose a significant consideration. If E2 operates within a new AWS account, all existing IAM roles used by clusters and jobs must be recreated within this new account. This demands identifying precise permissions for each role, adhering to the least privileged access model. This task can be intricate, especially if the existing roles were not initially designed following similar standards.

- **Sharing AWS resources:** Data stored in S3 and other critical AWS resources like SNS, Secrets Manager, Lambda, KMS, RDS, etc., must be shared to ensure accessibility from the new AWS account. This effort could be extensive, especially if the inventory of resources is incomplete or inadequately identified

- **Encryption:** Encryption can also emerge as a challenge. Cross-account access becomes unfeasible if data buckets (S3) and other resources (such as secrets and SNS topics) are encrypted using default AWS managed keys. To enable cross-account access, these data objects and resources necessitate re-encryption with shareable customer-managed KMS keys.

To proactively address these challenges, it is imperative to implement specific measures. This includes enforcing precise data replication rules with appropriate KMS keys, ensuring accurate encryption of AWS resources and buckets, and establishing streamlined processes for key updates, testing, and replication between environments. By effectively tackling these engineering challenges, we can ensure data integrity and secure access in both the PVC and E2 environments
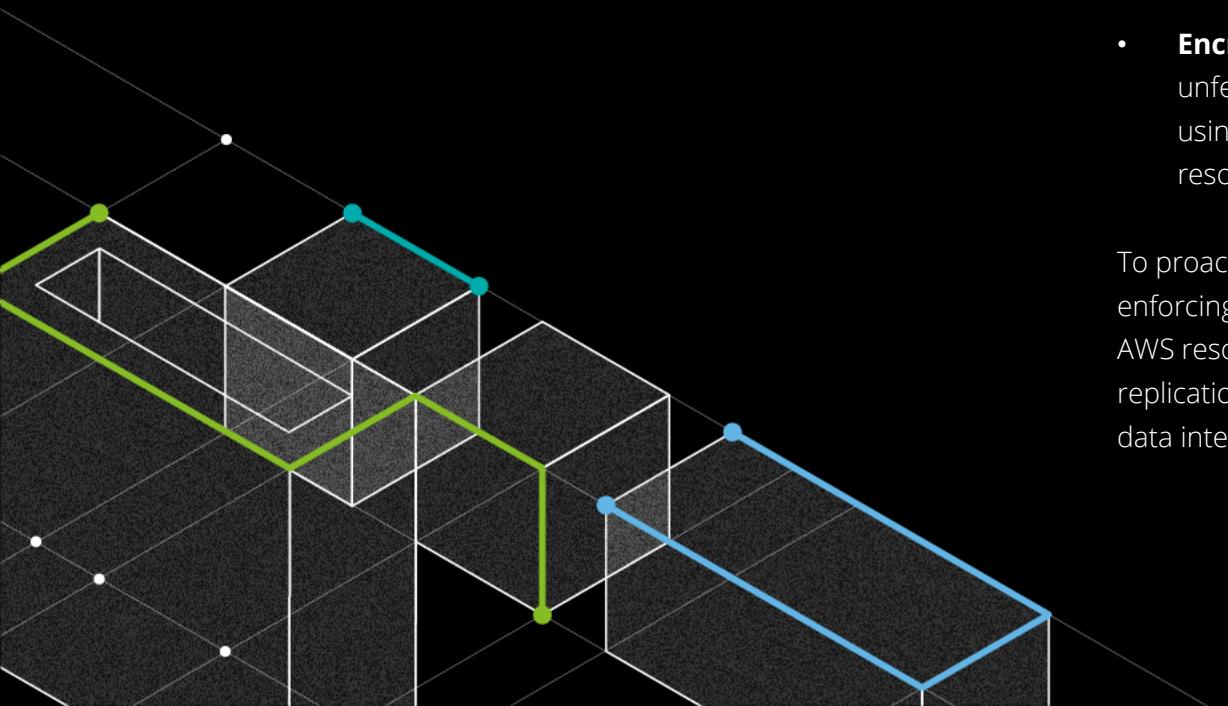
3. User Training and Adoption

The following issues might require attention to facilitate the user training and adoption of new Databricks best practices

- **Changes in access control type (cluster):** As part of upgrade to E2, users could face challenges, if access control on clusters is changed, i.e., from non-ACL to ACL-enabled. Few adjustments shall be necessary for users to align their processes and workflows with the new standards. This involve transitioning to job clusters, refraining from using dbutils, restricting access to boto3, and adopting alternative approaches as per the new guidelines.

- **New environment and Spark upgrade:** With the introduction of a new environment and Spark upgrade, users might need to make certain code refinements to ensure compatibility. This includes making minor code changes to accommodate the differences between the PVC and E2 versions.

- **New features**: As E2 introduces new features and services like SQL warehouses, unity catalog, etc., it is essential to provide users with necessary resources and comprehensive training.

By prioritizing user education and providing the necessary resources and assistance, we can effectively promote the adoption of new Databricks best practices, enhance user productivity, and ensure a seamless transition to the upgraded environment
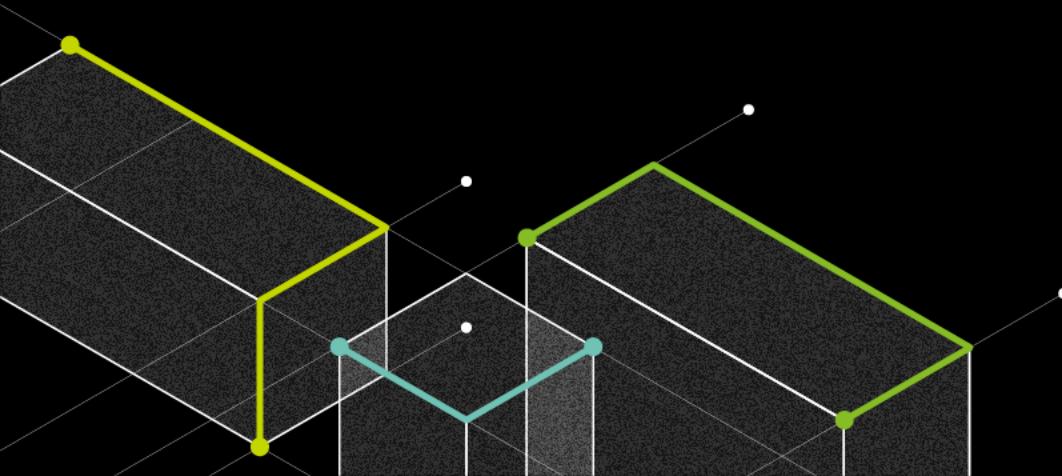
4. Support Process Post E2 Release

To help ensure efficient support post-release of the E2 environment, we recognized the importance of defining a clear process. The following areas require attention

- **Connection updates to third-party applications (Power BI, Tableau, Qlik, INFA, etc.,):** Users may need to make connection updates to third-party applications such as Power BI, Tableau, Qlik, INFA, and other tools. This process often involves requesting service tokens, credentials, and other related information through tickets. To address this, a standard process like directing end users to coordinate with relevant tool teams for troubleshooting configuration issues, can help streamlining the support process.

- **Job migration and reconfiguration on orchestration tools:** As users transition to the E2 environment, they need to make changes in all the environments and collaborate with the tool teams for migration. However, users may face challenges either due to lack of readiness or resource constraints to perform these changes themselves.

By clearly defining the support process, we can streamline the resolution of issues related to connection updates for third-party applications and facilitate the smooth migration and reconfiguration of jobs for orchestration tools. This approach will help enhance user experience, optimize resource allocation, and help ensure efficient troubleshooting and support in the post-release phase of the E2 environment.

5. Ownership of E2 Resources

One of the potential challenges could be the ownership assignment of Databricks resources when created using Terraform. In the E2 account, the ownership of resources by default, would be set to the individual who executed the Terraform code. This could pose potential issues if the resources become inaccessible when the person/owner leaves the organization

- To address this concern, implement multiple Terraform Databricks providers at both the individual workspace level and the Account level. This implementation ensures the ownership of the created resources to a Service Account rather than being tied to the individual that executes the code.

- By adopting this approach, one can mitigate the risk of resource inaccessibility and establish a more sustainable ownership model for Databricks resources.
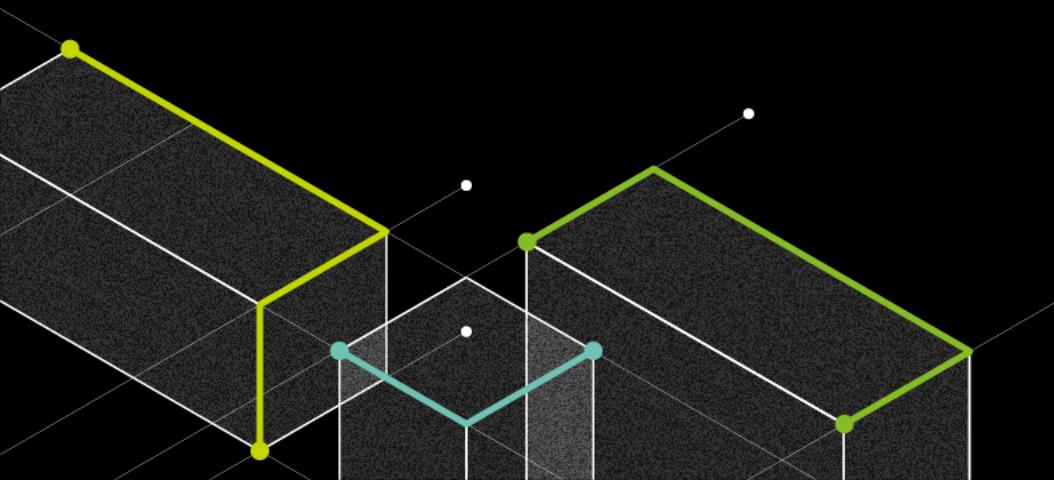
Best Practices

Running a large-scale platform migration project such as transitioning from Databricks PVC to Databricks E2, is a complex endeavor that necessitates meticulous planning, close collaboration with various stakeholders, and adherence to best practices.

Project Planning & Collaboration

During the migration process from Databricks PVC to Databricks E2, different challenges pertaining to the access of schemas/tables could be encountered

- Begin by clearly defining the project's scope, objectives, and timelines. Understand the reasons behind the migration, such as the need for improved performance, scalability, or cost optimization

- Identify the specific components, workloads, and applications that will be migrated. This may involve assessing the existing Databricks PVC environment, understanding its dependencies, and determining the required configurations in Databricks E2

Project Planning

Team Collaboration
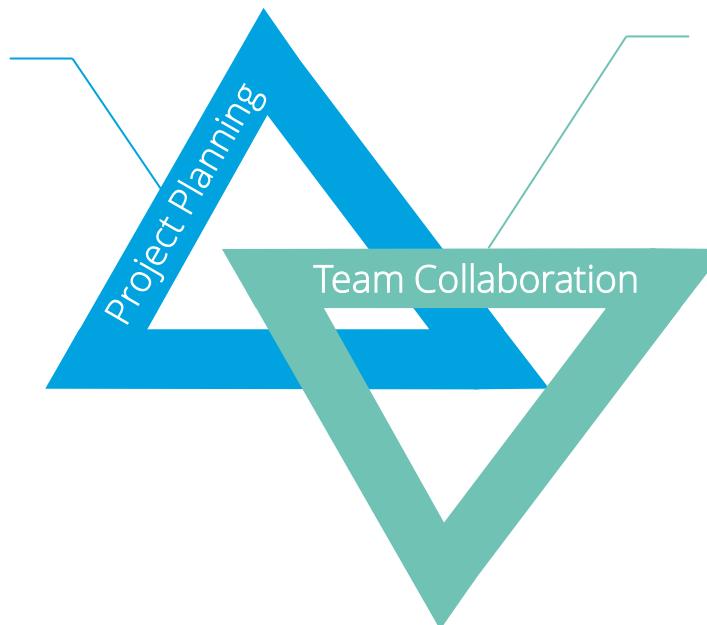
- Engage a cross-functional team comprising technical experts, data engineers, data scientists, and other relevant stakeholders. Ensure their active participation throughout the migration project

- Assign responsibilities and roles to team members, establishing clear communication channels and reporting structures

- Conduct regular meetings and status updates to monitor progress, address concerns, and ensure everyone is aligned
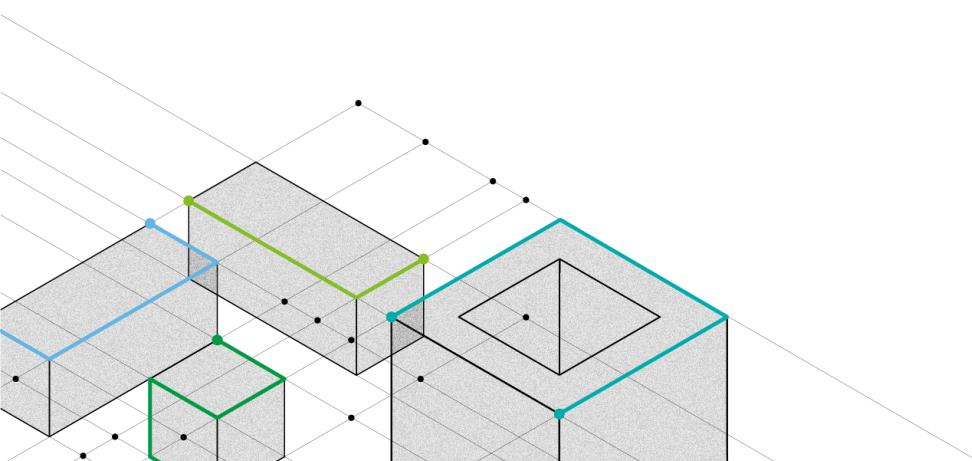
Stakeholder Communication & Engagement

Having a well-developed and comprehensive communication strategy is crucial when it comes to working effectively with the business units during the migration process. The goal is to ensure that the business units fully understand and embrace the migration methodology and timeline. In order to achieve this, we employed a range of communication channels to establish constant and meaningful engagement with the business teams.

	Scoping	Release & Test	Sign-Off	Decommission
Key Messages	Technical team confirms clusters and cluster owners to be migrated to new environment in wave approach Technical team has to meet with impacted teams to discuss solutioning design and provision of access	Clusters have been released to impacted teams Teams must confirm access and perform user testing in the new Databricks environment	Cluster owners must reference end user sign-off checklist and playbook to ensure their new Databricks environments are functioning as expected	Must migrate from old to new environment to be on supported version of Databricks Former Databricks version decommission date Timeline view of decommission process
Vehicles	• Live calls • Scoping closeout emails (by Wave) • Teams, Yammer, SharePoint • All Hands & Office Hours	• Cluster Release & Test emails (by team) • Follow-up emails and Teams messages as needed • All Hands & Office Hours	• Follow-up emails & Teams messages • Live calls • All Hands & Office Hours	• Decommission reminder emails • Teams, SharePoint • Live calls
Audience	Databricks Cluster Owners	Databricks Cluster Owners	Databricks Cluster Owners	Databricks Cluster Owners

Use available channels of communication to ensue business engagement kept alive, consistent, and effective. This helps in building strong relationships, gain buy-in, and success plan for business units through the migration process
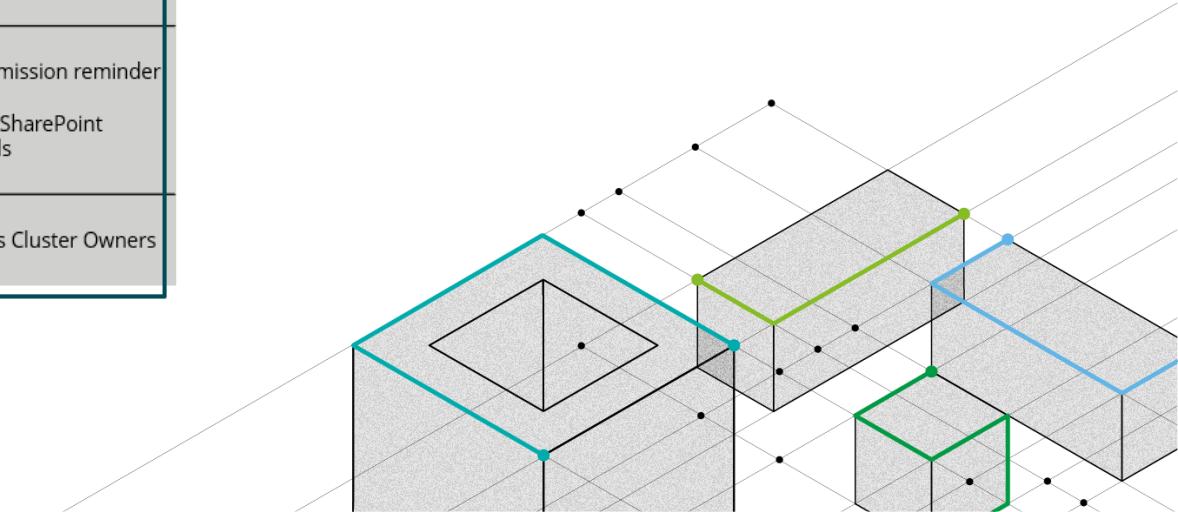
Environment and Configuration Setup

- Establish the Databricks E2 environment and configure it to meet the specific needs of the organization. This may involve defining cluster configurations, libraries, data sources, access controls, and security policies

- Migrate existing PVC environments, notebooks, libraries, and jobs to Databricks E2. Develop a comprehensive strategy for data migration, ensuring data integrity and minimal disruption to ongoing operations

- As standard leading practice, the workspace deployment should be done via IaC (Example, Terraform)

Automation

Migration from Databricks PVC to E2 requires a high level of automation rigor. Often overlooked in large projects, automation plays a crucial role in enhancing performance and efficiency, especially when it comes to rebuilding processes from scratch.

Third Party Application

- Scripts that can automatically update the connection settings, endpoints, and credentials used by third party applications like PowerBI, Informatica, Datameer, etc. to connect to Databricks E2

- Use automation tool to orchestrate the repointing process across multiple instances or environments, ensuring a streamlined and consistent approach

User/Group Management

- Utilize APIs to automate the creation, modification, and deletion of user accounts and groups in Databricks E2

- Automate the process for the assignment of appropriate roles and permissions to users and groups in Databricks E2

Job Migration

- Automate the migration of jobs (e.g., Spark jobs, scheduled notebooks) from the Databricks PVC environment to Databricks E2

- Automation for extraction of job configurations, dependencies, and code from the PVC environment and automated deployment to E2 environment

Grants on schemas and tables

- Automate the migration of grants (permissions) on schemas and tables from the Databricks PVC environment to Databricks E2

- Use automation scripts that could extract the existing grant configurations from the PVC environment and apply them automatically to the corresponding schemas and tables in the E2 environment

Service principal creation and token generation

- Automate the creation of service principles and token in Databricks E2 to enable secure authentication and access control

- Create automation script that could programmatically create service principles and its associated credentials
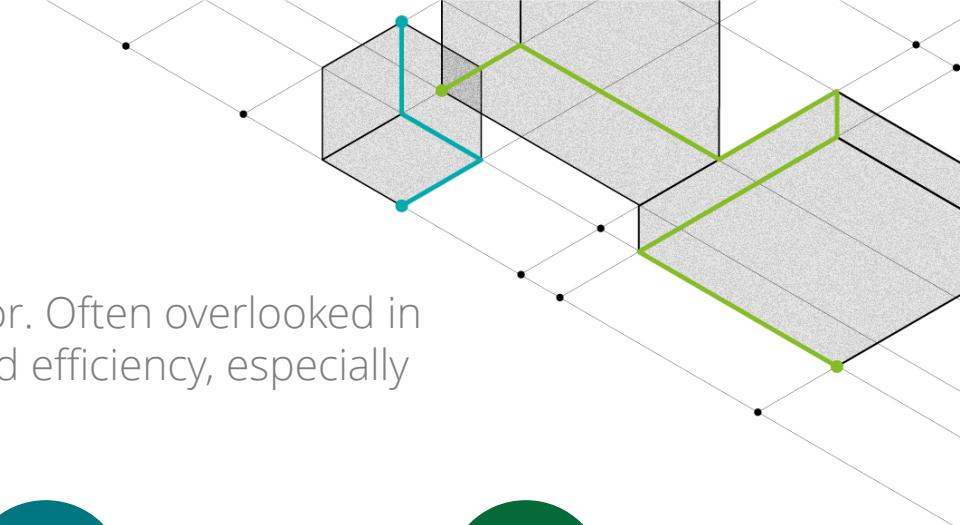
Change Management

During the migration of Databricks PVC to Databricks E2, several change management steps need to be implemented, including the use of GitHub for version control and the incorporation of JIRA tickets for effective change tracking.

1

Planning and scoping

o Define the migration project scope, identifying the specific components and functionalities to be migrated.

o Create JIRA tickets to capture and track the tasks related to the migration project.

o Involve relevant stakeholders and assign them to the corresponding JIRA tickets.

2

Documentation and version control

o Set up a repository on GitHub to store all the code related to the migration process.

o Ensure that all code changes, configurations, and scripts are version-controlled using Git.

o Utilize JIRA integration with GitHub to link relevant JIRA tickets with corresponding code changes, facilitating traceability.

3

Code migration and testing:

o Create a branch in the GitHub repository dedicated to the migration process, with the branch name associated with the corresponding JIRA ticket.

o Implement the required code and configuration changes, making commits and pull requests in GitHub linked to the associated JIRA ticket.

o Perform thorough testing of the migrated components, documenting the testing activities and results within the corresponding JIRA tickets.

4

Change review and approval:

o Establish a formal change management process, leveraging JIRA's workflow capabilities, to ensure thorough review and approval of code changes.

o Conduct code reviews within the team.

o Obtain approvals ensuring that all changes are appropriately reviewed before proceeding with the migration.

Using SQL Warehouse for BI Loads

Using SQL Warehouse for BI load in Databricks E2 offers several benefits

✓ Optimized Data Warehousing: SQL Warehouse in Databricks E2 is specifically designed for efficient data warehousing and analytics. It provides optimized query performance, query optimization techniques, and caching mechanisms, resulting in faster and more efficient processing of large datasets commonly used in BI workloads.

✓ Scalability and Performance: SQL Warehouse in Databricks E2 offers scalable compute and storage resources. It can handle large volumes of data and support concurrent queries from multiple users, ensuring high performance and responsiveness for BI load requirements.

✓ Unified Analytics Platform: Databricks E2 provides a unified analytics platform that integrates with SQL Warehouse. This integration allows BI users to seamlessly leverage the power of SQL queries alongside other advanced analytics capabilities provided by Databricks, such as machine learning and streaming data processing. It enables organizations to consolidate their BI and advanced analytics workloads into a single platform, reducing complexity and enhancing productivity.

✓ Serverless SQL: Databricks E2 introduces Serverless SQL, which enables on-demand querying without the need to manage or provision dedicated clusters. With Serverless SQL, BI users can execute ad-hoc SQL queries directly against SQL Warehouse, making it easy to explore and analyze data without upfront infrastructure provisioning or management overhead.
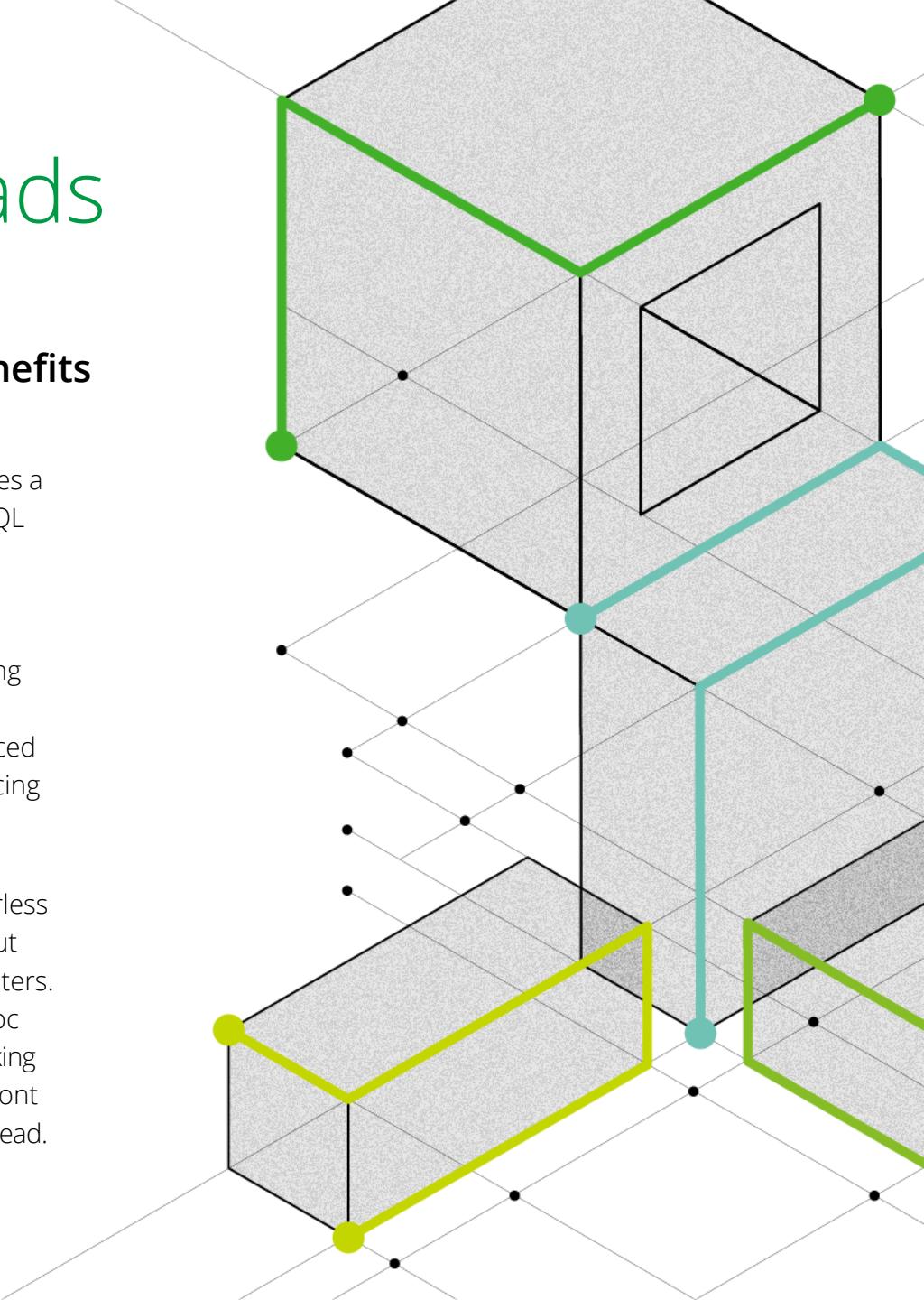
Value Delivered

In addition to the new features, services, and architectural benefits which Databricks E2 provides, implementation of best practices during the migration helped in realizing value across various domains:

1. Lower TCO and lower maintenance costs – As compared to Databricks PVC, in Databricks E2, the control plane is hosted and managed by Databricks, reducing the infrastructure costs and operational costs associated with applying updates, patches, and other maintenance activities. Databricks E2 is a SaaS based fully managed service with 24/7 monitoring and technical support

2. Better consumption transparency – Using well-defined cluster policies and tagging strategies enables better visibility into resource consumption at the application level. This further enables the implementation of a robust chargeback model across the enterprise

3. Secure and self-managed user access - Better access control mechanism via integration with identity access management systems like Okta, enabling self-service persona-based access and eliminating operational overhead on support teams

4. Cluster re-architecture based on user personas (Data Analyst, Data Engineer, Data Scientist, App Admin) – Architected cluster type, cluster configuration, access control, and permissions based on defined personas to streamline specific workloads and use-cases

5. SQL warehouses – Enabling new Databricks E2 features like SQL warehouses for faster, easier, cost-efficient, and controlled integration with visualization tools like Power BI, Qlik Sense, Tableau, etc., (BI Infrastructure)

6. Infrastructure as Code (Terraform) – Leveraging IaC best practices using terraform to build and configure Databricks E2 workspaces. Terraform is used for faster and easier migration of workloads (clusters, notebooks, user groups, jobs, etc.,). Implemented code repositories and pipelines for building consistent and repeatable workloads across environments.
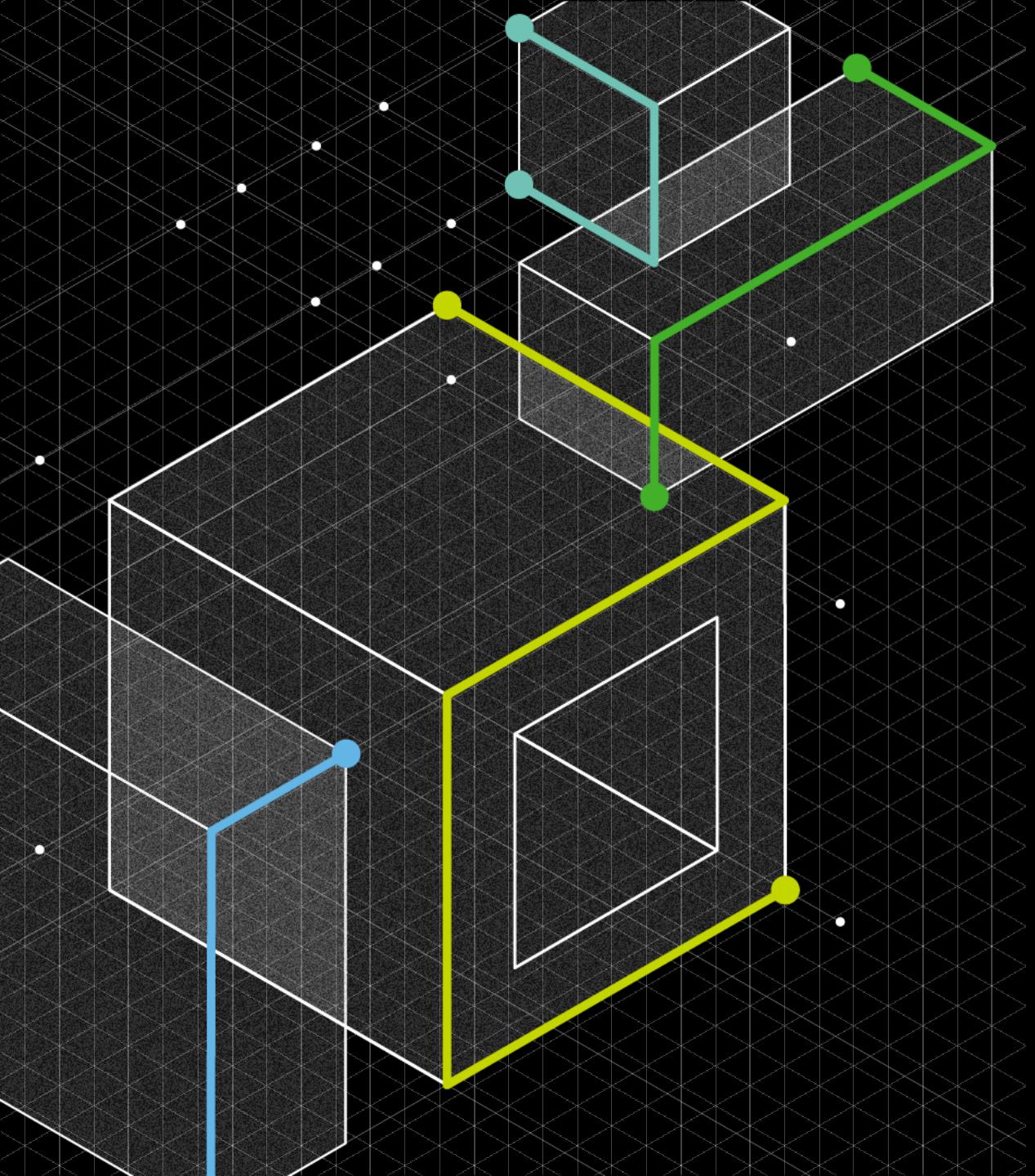
Conclusion

To conclude, migrating from Databricks PVC to E2 proves to offer several business benefits, including improved performance, improved platform stability, and cost savings. Furthermore, Databricks E2 demonstrates the ability to have new features, including SQL warehouse, Delta Live Tables, and Delta Data Sharing. When enabled, these will help drive business innovation.

The Databricks version upgrade from PVC to E2 is not a minor upgrade and Deloitte is equipped to guide and support you through this transition. As part of the migration, it will require level of code retrofitting, Spark version compatibility checks, performance evaluation with different applications across different networking tenants and thorough planning around regression testing. During the journey, all the different stakeholders from different consumers of the platform (ad-hoc users, IT development team, solution support team, business internal team and IT leadership) must be onboarded with a well-planned change management strategy. Doing a regular check-in on the efficiency, process improvement and features are some best practices to be followed during the execution cycle. Decommission is the big topic that must be planned to ensure a sufficient cool off period is given between the decommission stages. Overall migration to E2 could be challenging if things are not planned well in advance with a proper change management process.

Meet the team | Takeda

Barbara Latulippe
Barbara is Chief Data Officer with Takeda Pharmaceuticals U.S.A., Inc.

Jim Lee
Jim is Data Science Product Manager (IT) with Takeda Pharmaceuticals U.S.A., Inc.

Ross Hoffner
Ross is Head of Data Platform Architecture and Engineering (Enterprise Data) with Takeda Pharmaceuticals U.S.A., Inc.

Shyam Dadala
Shyam is Enterprise Solutions Architect with Takeda Pharmaceuticals U.S.A., Inc.

Meet the team | Deloitte

Vikranth Gudala

Vikranth is Consulting Managing Director with Deloitte Consulting LLP AI & Data Engineering offering

Rajinder Sobti

Rajinder is Managing Director with Deloitte Consulting India AI & Data Engineering offering

Kashinath Yadav

Kashinath is Manager with Deloitte Consulting LLP AI & Data Engineering offering

Aman Bedi

Aman is Specialist Leader with Deloitte Consulting LLP CBO offering

Gautam Sarup

Gautam is Senior Manager with Deloitte Consulting India AI & Data Engineering offering

Mani Kandasamy

Mani is a Technology Fellow with Deloitte Consulting LLP AI & Data Engineering Offering

Deloitte.

Thank you