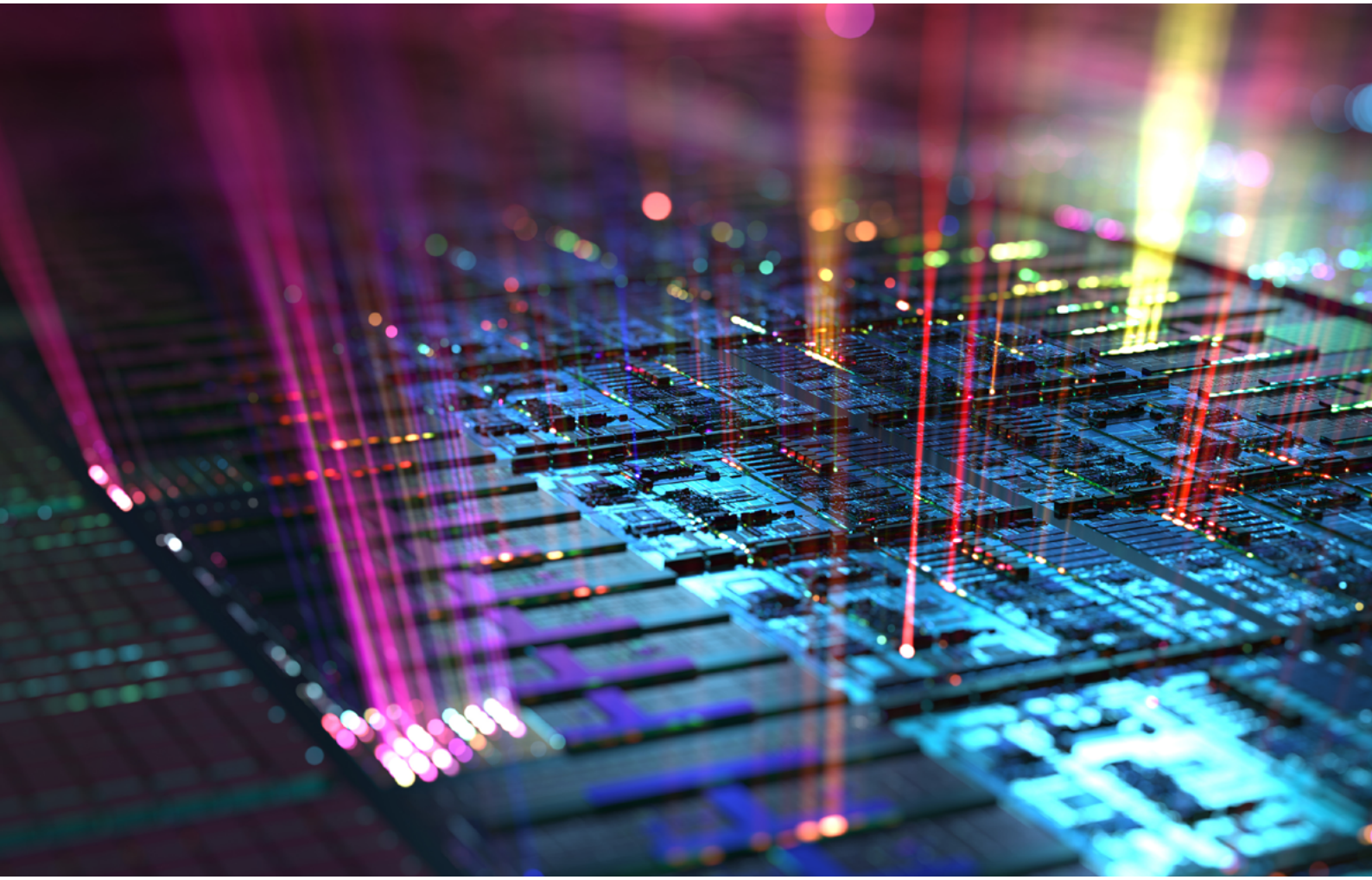
DXC TECHNOLOGY



Building a data marketplace by using Databricks

Data marketplace

Arun Khandelwal, Enterprise Data Solution Architect,
Databricks Partner Solution Architect Champion
at DXC Luxoft

Table of **contents**

Abstract: Building a data marketplace using Databricks

The explosion of data across industries has created immense opportunities for organizations to capitalize on data as a strategic asset. A data marketplace is a centralized platform where organizations can securely buy, sell, exchange and share data, fostering innovation, collaboration, and new revenue streams. By leveraging Databricks, an industry-leading cloud-based data platform, organizations can build scalable, governed, secure data marketplaces that facilitate seamless data transactions.

This white paper outlines the key steps in building a data marketplace using Databricks. Databricks' powerful data processing, advanced analytics, ML capabilities and collaborative environment make it the ideal foundation for a data marketplace. It enables organizations to easily manage large datasets, ensure data governance and compliance and provide a seamless user experience for data providers and consumers. By integrating Databricks with other cloud services and leveraging its data intelligent platform for data ingestion, processing, storage and sharing, organizations can unlock the full potential of their data assets, enabling innovation and creating new business opportunities in a data-driven economy.

Introduction

The digital era has ushered in an unprecedented surge in data generation, fundamentally reshaping the global economy. Today, data has evolved into a critical asset capable of driving innovation, fueling growth and enhancing decision making. A data marketplace — a platform that facilitates buying, selling and sharing data — offers organizations a powerful opportunity to tap into this resource, unlocking value in entirely new ways.

Data marketplaces provide a s**ecure and governed environment** for data exchange, opening up new revenue streams, promoting cross-industry collaboration and accelerating product development. These platforms allow organizations to **monetize their data assets**, gain valuable insights and make data-driven decisions that result in competitive advantages.

This white paper explores how Databricks can be the ideal foundation for building and managing a data marketplace. Leveraging Databricks' Lakehouse architecture, advanced analytics and robust data governance capabilities using Unity Catalog, organizations can create a secure, future-ready platform for seamless data exchange and monetization, ensuring they're well-equipped for the evolving data economy.

Overview: Data marketplace

Organizations are increasingly realizing the immense value of data as a strategic asset. A data marketplace, a centralized platform that enables the buying, selling and exchange of data, offers a powerful solution for unlocking data's full potential. By providing a secure, governed and scalable environment for data transactions, these marketplaces open new avenues for innovation, collaboration and economic growth.

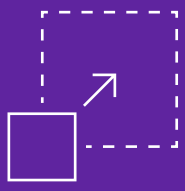
The success of a data marketplace hinges on a strong technological foundation capable of efficiently managing and facilitating data exchange. Databricks, with its advanced analytics engine, Lakehouse Architecture and robust data governance capabilities via Unity Catalog, is an ideal platform for building and managing a data marketplace. It empowers organizations to handle vast data volumes, safeguard sensitive information and deliver actionable insights to users.

By following Databricks' best practices for building a data marketplace, organizations can seize growing opportunities and stay ahead in the competitive landscape.

Key characteristics of a data marketplace

Secure and governed

Ensures data exchange occurs within a highly secure and compliant framework, protecting sensitive information while adhering to privacy regulations like GDPR or CCPA.



Scalable

Scales data operations from small datasets to petabytes of information, supporting thousands of users and complex data operations simultaneously.

Curated data

Available data is curated (i.e., preprocessed, organized and enriched) to ensure it's high quality and ready for consumption.

Key components of a Databricks-powered data marketplace

Data producers: These entities (businesses, research institutions or individuals) upload and share their datasets. They might also curate and enhance data to make it more valuable.

Data consumers: Data consumers purchase or subscribe to datasets. They might analyze the data for insights, build ML models or integrate it into their systems for decision making.

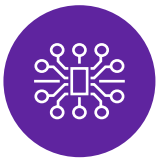
Marketplace governance: Using Databricks' Unity Catalog, administrators manage access to datasets, enforce compliance and ensure proper data usage with features like audit logs, role-based access control (RBAC) and data lineage tracking.

Data-sharing and access: With Delta Sharing, data producers can securely share data with consumers, even across different organizations or cloud environments. This data can be accessed in real-time without needing to duplicate or move datasets, ensuring data integrity.

Pricing and monetization: Marketplace operators can implement flexible pricing models, including subscriptions, one-time purchases or usage-based pricing. This enables data producers to monetize their datasets effectively while providing consumers with cost-effective data access.

Key features of a data marketplace built on Databricks

Centralized data hub: Acts as a single platform where data providers can share datasets and consumers can discover and subscribe to relevant data sources.



Scalable data processing: Leverages Databricks' distributed processing engine to handle both batch and real-time data at scale.

Delta Sharing for seamless data exchange: Delta Sharing facilitates secure, open sharing of data across different Databricks workspaces and cloud environments.

Data governance and security: Enforces strict governance through role-based access control, data lineage and compliance features, ensuring that shared data is secure and trustworthy.

Advanced analytics and ML: Enables consumers to explore, analyze and build ML models on the marketplace data using Databricks' data intelligent platform.

.

Why a data marketplace is critical for organizations

Monetize data assets: Organizations can unlock the economic value of their data by selling or sharing it with partners or third parties, creating new revenue streams.

Democratize data access: Break down data silos and make valuable data available to various departments, business units or external stakeholders in a governed manner.

Accelerate innovation: By providing ready access to high-quality data, organizations can accelerate innovation, enabling faster insights and reducing the time needed to develop new products or services.

Enhanced decision making: Data consumers, including business analysts and data scientists, can access rich datasets to perform advanced analytics, derive actionable insights and make data-driven decisions.

Key use cases for a data marketplace

A data marketplace provides a platform for buying, selling and sharing data in a secure, scalable and governed environment. Organizations can use a data marketplace to unlock new revenue streams, drive innovation and facilitate collaboration across industries. Below are some key use cases for a data marketplace:

1. Monetizing data assets

- **Selling proprietary data:** By selling access to proprietary data, a retail company can provide manufacturers looking to understand market trends with consumer purchasing data.
- **Data as-a-Service (DaaS):** Companies can package and sell real-time data or datasets on a subscription basis, turning data into a recurring revenue stream.

2. Data sharing and collaboration

- **Cross-industry collaboration:** Different organizations or industries can securely share data to foster collaboration. For instance, healthcare providers could share anonymized data with pharmaceutical companies to enhance research and development.
- **Partnerships and ecosystems:** Companies in the same ecosystem (e.g., automotive, financial services) can exchange data to create value-added services, such as enhanced customer experiences or improved operational efficiency.

3. Enabling AI and ML

- **Access to rich data for AI Models:** Data marketplaces provide AI/ML practitioners with access to diverse, high-quality datasets that can enhance model training and predictive accuracy.
- **Accelerating AI innovation:** Organizations can purchase curated datasets to accelerate ML model development, reducing the time and cost required to build AI-driven solutions.

4. Data enrichment and augmentation

- **Augmenting internal data:** Organizations can purchase external datasets from the marketplace to enrich their internal data. For instance, a company can buy weather data to enhance its demand forecasting models.
- **Customer insights:** Marketers can buy demographic, social, or behavioral data to gain deeper insights into their target audience and improve segmentation and targeting efforts.

5. Regulatory and compliance data exchange

- **Sharing regulatory-compliant data:** Industries like finance and healthcare can exchange data under strict regulatory frameworks (e.g., GDPR, HIPAA), ensuring compliance while gaining value from external data sources.
- **Regulatory reporting:** Financial institutions can buy datasets for compliance reporting, such as credit risk data or market trends that help them adhere to government regulations.

6. Supply chain optimization

- **Logistics and supply chain data:** Companies can buy and share data related to logistics, transportation and demand forecasting to improve their supply chain efficiency. For example, sharing supplier performance data can help buyers make more informed decisions.

7. Real-time data exchange

- **Trading market data:** Financial services firms can leverage real-time data exchanges to access and share stock prices, trading volumes and market sentiment data for faster decision making.
- **IoT and sensor data:** Industries such as manufacturing or smart cities can trade real-time IoT data on machinery performance, weather patterns or traffic flow, enabling immediate operational insights.

8. Market research and competitive analysis

- **Market trends:** Companies can purchase industry-specific data to analyze market trends, competitor performance or consumer sentiment to drive strategic decisions.
- **Competitive benchmarking:** Access to competitor data from the marketplace can help businesses benchmark their performance and identify areas for improvement.

9. Public sector and open data collaboration

- **Open data sharing:** Governments can use data marketplaces to share open data with the public, startups and businesses for transparency and to encourage innovation.
- **Public/Private partnerships:** Public organizations can collaborate with private entities to solve societal challenges, such as environmental monitoring or smart-city development, by sharing and analyzing relevant datasets.

10. Healthcare and life sciences innovation

- **Research and clinical trials:** Data marketplaces can facilitate the exchange of clinical trial data, research findings and patient outcomes across healthcare institutions to accelerate medical research and innovation.

- **Pharma data sharing:** Pharmaceutical companies can share data on drug efficacy, patient outcomes and adverse events to improve drug development processes and personalized medicine initiatives.

11. Financial market data sharing

- Financial institutions can share real-time and historical market data feeds with hedge funds, traders and analysts. Real-time data processing and predictive analytics models can help consumers make timely decisions.

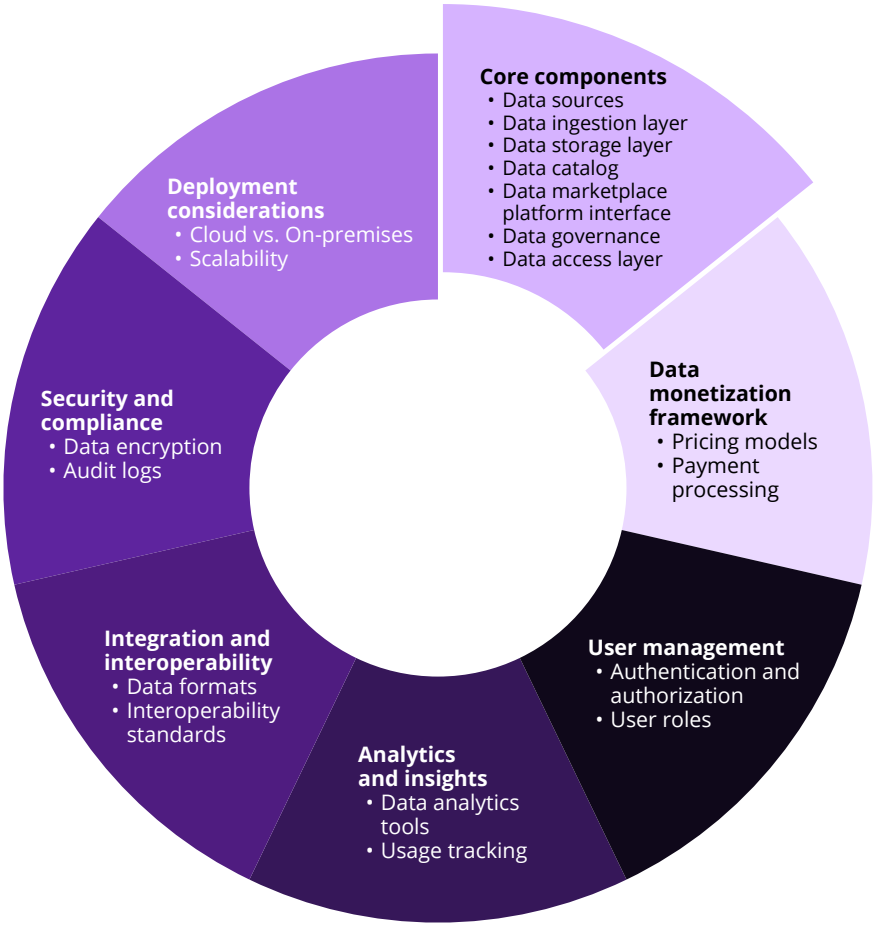
12. Supply chain data marketplace

- Manufacturers and logistics companies can share supply chain data (e.g., inventory levels and demand forecasts) with partners to optimize operations and reduce inefficiencies.

13. Retail data monetization

- Retailers can sell anonymized purchase data or customer insights to third-party vendors, helping them understand consumer behavior and trends.

Data marketplace reference architecture



Step-by-step guide to building a data marketplace using Databricks

1. Data ingestion and integration layer

The first step is to enable data providers to ingest and integrate datasets into the marketplace. This can include both batch and real-time data ingestion from various sources.

Components:

- **Databricks Auto Loader:** Automates the ingestion of raw datasets from various cloud storage systems (e.g., AWS S3, Azure Blob Storage, Google Cloud Storage)
- **Streaming data via Kafka:** Use **Apache Kafka** or **Azure Event Hubs** to integrate real-time data streams for events like social media activity, sensor data, or market feeds
- **Third-party data sources:** Integrate with third-party APIs (e.g., social, financial, or web APIs) to fetch external datasets

Key functions:

- Collect data from structured, semi-structured, and unstructured sources
- Support both batch processing and streaming ingestion pipelines for real-time data use cases
- Automatically detect and manage evolving data schemas to simplify ingestion

2. Data storage layer (Delta Lake)

Once the data is ingested, it must be stored in a reliable, scalable format that supports versioning, auditing, and ACID transactions. This ensures that data consumers can trust the quality and integrity of the datasets they access.

Components:

- **Delta Lake:** Provides the foundation for reliable and scalable storage. It supports ACID transactions, schema enforcement, and time-travel capabilities to ensure data integrity
  - **Bronze layer:** Raw data as ingested, providing an immutable source of truth
  - **Silver layer:** Cleaned and enriched data, ready for consumption
  - **Gold layer:** Aggregated, curated datasets tailored to specific business use cases

Key functions:

- **ACID** compliance to guarantee data consistency across ingestion and processing
- **Time travel** to access historical versions of datasets for auditing and compliance
- **Schema enforcement** and evolution to maintain data quality as data schemas change over time

3. Data governance and cataloging layer (Unity Catalog)

Data governance is a crucial aspect of a data marketplace, especially when dealing with external partners and consumers. Unity Catalog allows you to manage access permissions, maintain data lineage, and ensure compliance with industry regulations.

Components:

- **Unity Catalog:** A centralized governance solution allowing fine-grained access control, cataloging, and data lineage tracking
- **Role-based access control (RBAC):** Ensures only authorized users have access to specific datasets, preventing unauthorized access
- **Data lineage and audit logs:** This system tracks every transformation, access request and change to datasets, ensuring complete transparency for compliance and auditing purposes

Key functions:

- Provide role-based access controls (RBAC) to manage who can view, query and modify datasets
- Track and visualize data lineage to understand data flows, transformations and dependencies
- Enforce data privacy and security policies to ensure compliance with regulations such as **GDPR**, **HIPAA**, or **CCPA**

4. Data processing and transformation layer

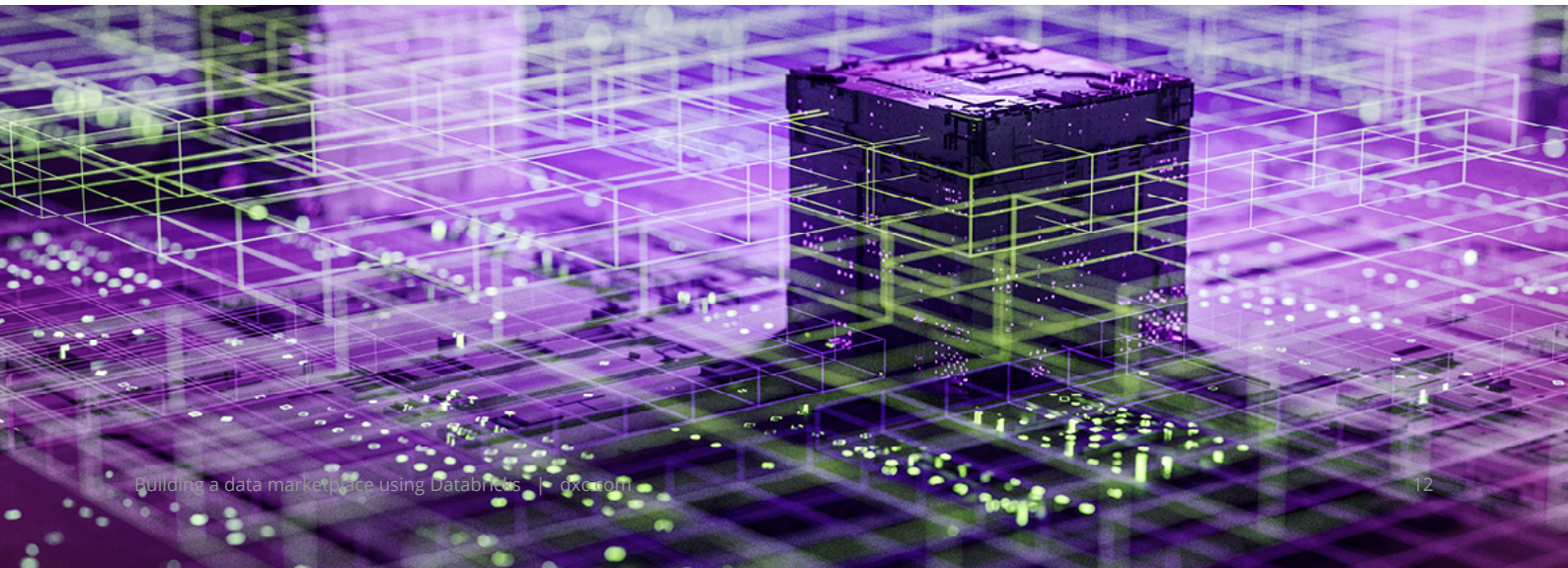
Before datasets are made available in the marketplace, they need to be cleaned, transformed and enriched for usability. The data processing layer allows for the large-scale transformation, enrichment and processing of raw data into high-value datasets.

Components:

- **Apache Spark on Databricks:** High-performance, distributed data processing for transforming and preparing data for consumption
- **Databricks workflows:** Orchestrate complex ETL pipelines to transform and clean data as it moves through the marketplace
- **Delta Live Tables (DLT):** Automate continuous data transformation for real-time processing and ensure data is always up to date for consumers

Key functions:

- Perform large-scale transformations, filtering and enrichment of datasets to prepare them for consumption
- Manage and optimize ETL pipelines to ensure data consistency and quality
- Enable real-time data processing for time-sensitive datasets such as market feeds, IoT data or streaming analytics



5. Data marketplace access and sharing layer (Delta Sharing)

A core feature of a data marketplace is enabling secure data sharing between different users and organizations. With Delta Sharing, Databricks enables the open and secure sharing of live data across cloud environments and between organizations.

Components:

- **Delta Sharing:** A secure, open protocol for sharing live data between different Databricks workspaces or external organizations without data replication
- **REST APIs for data access:** APIs allow data consumers to access datasets programmatically through APIs, enabling easy integration into external systems, models or applications

Key functions:

- Share live data securely with partners, customers and third-party organizations
- Consumers can subscribe to datasets and receive real time or near-real time updates without complex data replication
- Set up data subscriptions for consumers to automate access to newly updated datasets

6. Data consumption and analytics layer

Once the data is shared or purchased, consumers need powerful tools to analyze, visualize and derive insights from it. Databricks enables these with its built-in analytics and ML capabilities

Components:

- **Databricks SQL warehouse:** Allow users to interrogate datasets with standard SQL queries and generate insights from marketplace data
- **Power BI/Tableau integration:** Seamless integration with popular BI tools for data visualization and dashboarding
- **ML (MLflow):** Data scientists can use Databricks' **MLflow** to experiment, track and deploy ML models on top of the data available in the marketplace
- **Lakehouse federation**

Key functions:

- Run ad-hoc queries to explore and analyze datasets directly from the marketplace
- Create dashboards, reports and visualizations for business insights and decision making
- Build and deploy ML models for predictive analytics, recommendation engines and real-time predictions based on marketplace data

7. Security and compliance layer

The marketplace must be built with a focus on security, privacy and compliance to protect sensitive or private data. Databricks provides a comprehensive suite of security features to ensure the platform complies with industry standards.

Components:

- **Encryption (in transit and at rest):** All data is encrypted while being processed and stored in cloud environments
- **Unity Catalog:** Provides centralized security and auditing, ensuring that data access is restricted and monitored based on role-based permissions
- **Audit trails and logging:** Track all access, changes and data usage for full transparency and regulatory compliance

Key functions:

- Protect sensitive data with encryption and secure access protocols
- Provide comprehensive auditing and logging for compliance and governance
- Ensure data sharing adheres to global privacy standards such as GDPR, CCPA, etc.

8. Monetization and marketplace operations layer

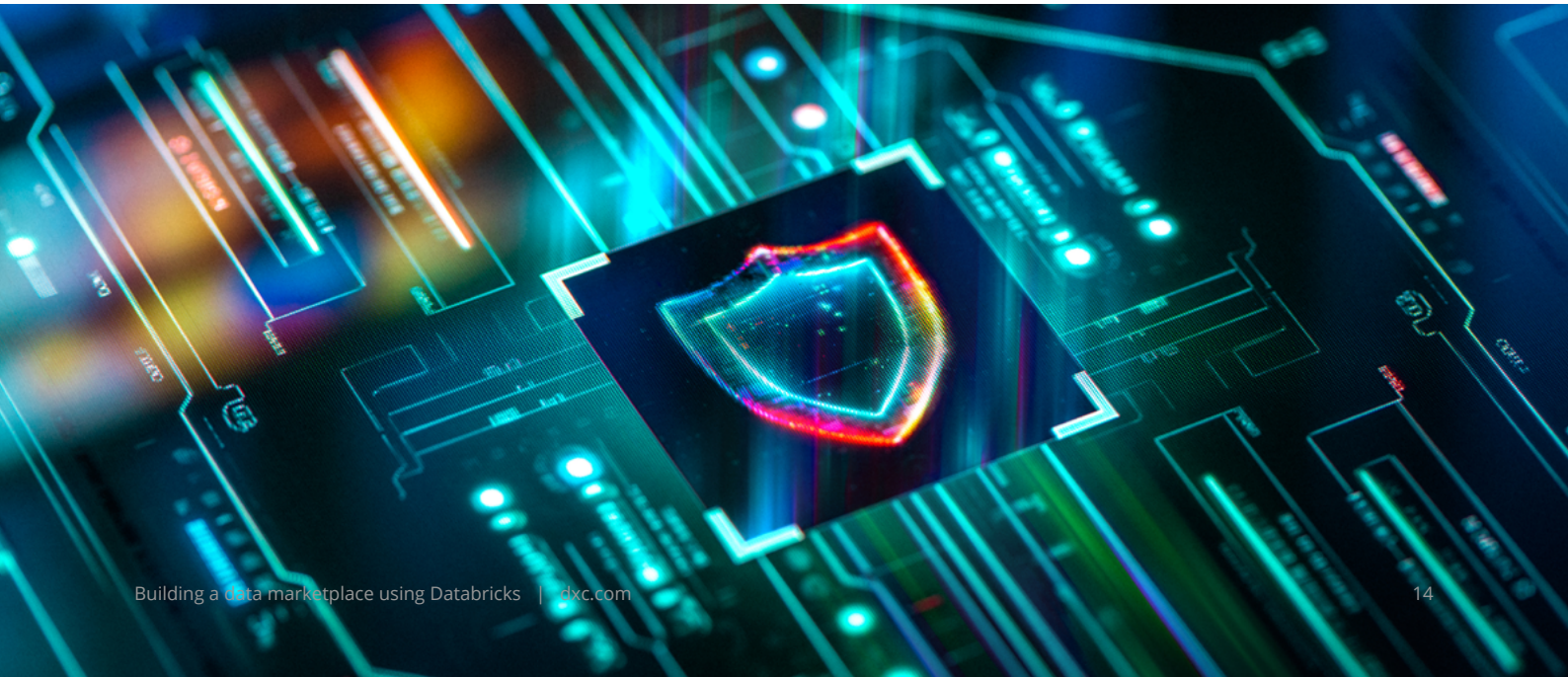
This layer supports subscription models, pricing and usage tracking for organizations looking to monetize their datasets.

Components:

- **Subscription management:** Allows data providers to offer their data under various pricing models (e.g., pay-per-use, subscription-based access)
- **Usage tracking and billing:** This service tracks data usage by consumers to generate billing information and charge users based on their access or consumption

Key functions:

- Offer flexible pricing and access models (one-time, subscription or usage-based) for marketplace datasets
- Monitor data consumption and generate reports for usage and billing
- Support multitenant environments where different organizations can contribute and consume datasets



Best practices for building a data marketplace on Databricks

1. Leverage Delta Lake for data consistency

- **Use Delta Lake:** Databricks' Delta Lake ensures ACID transactions, scalable metadata handling and data versioning. This is crucial for building a reliable data marketplace where users trust the data they are consuming

- **Data quality and lineage:** Enable features like time travel and schema enforcement to maintain data integrity and integrate with Databricks' built-in lineage tools for traceability

2. Implement robust data governance

- **Unity Catalog for data governance:** Use Databricks Unity Catalog to create a centralized data catalog with fine-grained access control and lineage tracking. Unity Catalog simplifies permissions management across different data assets, making securing and managing marketplace data easier

- **Tagging and metadata management:** Tag datasets with metadata like ownership, classification, sensitivity and relevance to make it easier for users to find and use the correct data

- **Data masking and row-level security:** Implement data masking and row-level security to protect sensitive data while ensuring authorized users can access relevant data

3. Create data products

- **Data as-a-product:** Treat data assets as products, ensuring that datasets are curated, documented and meet quality standards before being published on the marketplace

- **Standardized datasets:** Create standardized datasets that can be shared across different teams or organizations. Use schema enforcement to ensure consistency across data products

4. Establish clear data access and sharing policies

- **Fine-grained access control:** Ensure access controls are in place to prevent unauthorized access. Use role-based access controls (RBAC) and attribute-based access controls (ABAC) to manage permissions efficiently

- **Data sharing via Databricks Delta Sharing:** Leverage Delta Sharing to share data within your organization and with external partners securely. Delta Sharing enables the secure exchange of live data without replication

5. Data discovery and self-service analytics

- **Enable self-service:** The marketplace should empower users to discover and access data autonomously. Implement a searchable interface for users to find relevant datasets

- **Interactive querying with SQL Analytics:** Leverage Databricks SQL analytics to provide users with self-service access to the data marketplace, enabling them to explore datasets and run queries directly within the platform

- **Notebooks and APIs for advanced users:** Allow data scientists and engineers to use Databricks notebooks for custom analysis and APIs for programmatic access to the data

6. Scalability and performance optimization

- **Optimize workloads with auto-scaling:** Ensure the Databricks clusters are set up to manage large and variable workloads efficiently. This ensures that your marketplace can scale based on demand

- **Optimize data storage:** Use best practices like **Z-ordering** and **data partitioning** for optimal query performance, especially for large datasets

- **Caching:** Use Databricks' caching capabilities to speed up frequently accessed datasets and improve user query performance

7. Data monetization and usage tracking

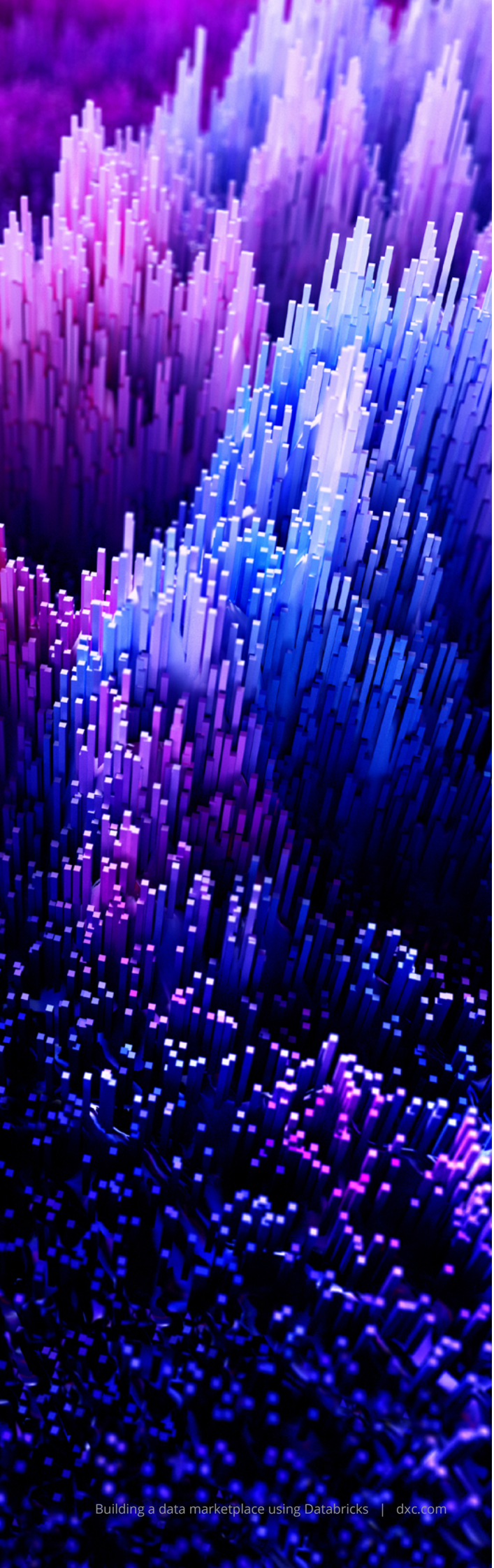
- **Track data usage:** Implement mechanisms to track how datasets are used within the marketplace. Unity Catalog provides audit logs that help track access and usage patterns

- **Monetize data:** If applicable, define a system to monetize data assets, enabling departments or external consumers to access valuable datasets on a subscription or pay-per-use basis

8. Collaboration and data-sharing culture

- **Promote cross-department collaboration:** Encourage teams to contribute high-quality datasets to the marketplace, fostering a culture of collaboration across departments or even external partners

- **Enable data sharing with external partners:** Use secure data-sharing features like Delta Sharing for external collaborations while maintaining strict security controls and auditability



9. Automation and continuous integration

- **Automated data pipelines:** Automate ETL/ELT pipelines to ensure data is continuously updated and available. Use Databricks jobs and workflows to schedule and automate these processes

- **CI/CD for data models:** Establish CI/CD pipelines for managing data transformations and ML models within the marketplace

10. Ensure compliance with regulations

- **Data privacy and GDPR:** Implement measures to comply with data protection regulations such as GDPR, HIPAA, etc. Use features like data anonymization, pseudonymization and audit logging to ensure compliance

- **Audit and logging:** Ensure all data access, transformations and sharing activities are logged for compliance and transparency

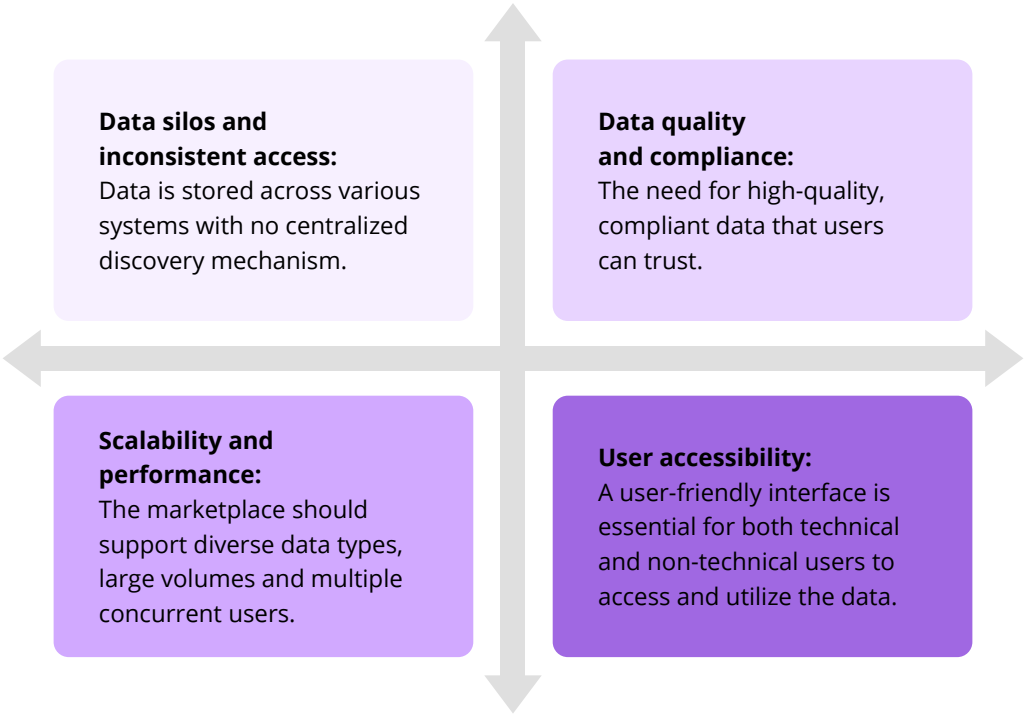
11. Monitor and maintain data health

- **Continuous monitoring:** Use Databricks' monitoring features, like structured streaming metrics, to track the health of your data pipelines and prevent disruptions

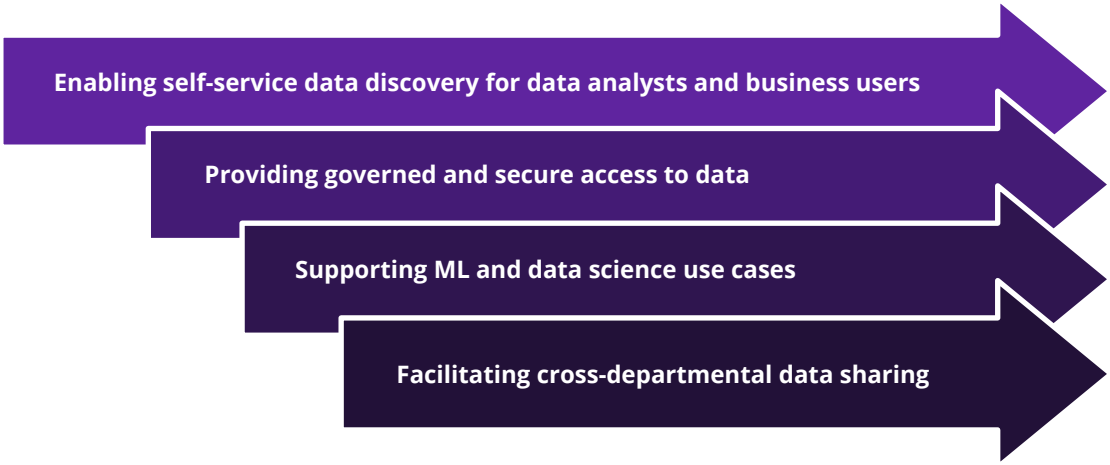
- **Data validation:** Set up data quality checks using libraries like Great Expectations to ensure that only clean, accurate, validated data is made available on the marketplace

Planning and implementing a real-world data marketplace

Challenge: To create a centralized data marketplace, providing a single platform for data discovery, access, analysis and monetization. The solution must break down data silos, improve data accessibility, ensure governance, foster a data-driven culture and enable new revenue streams.



Use cases covered

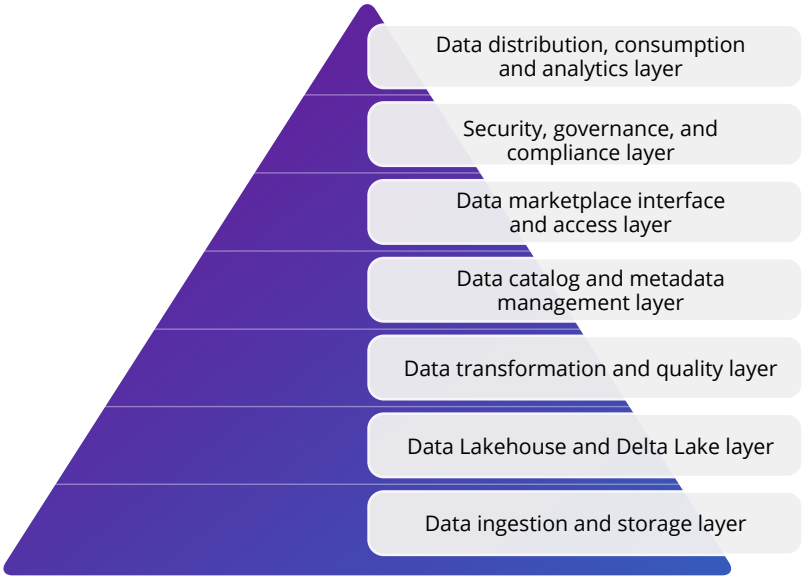


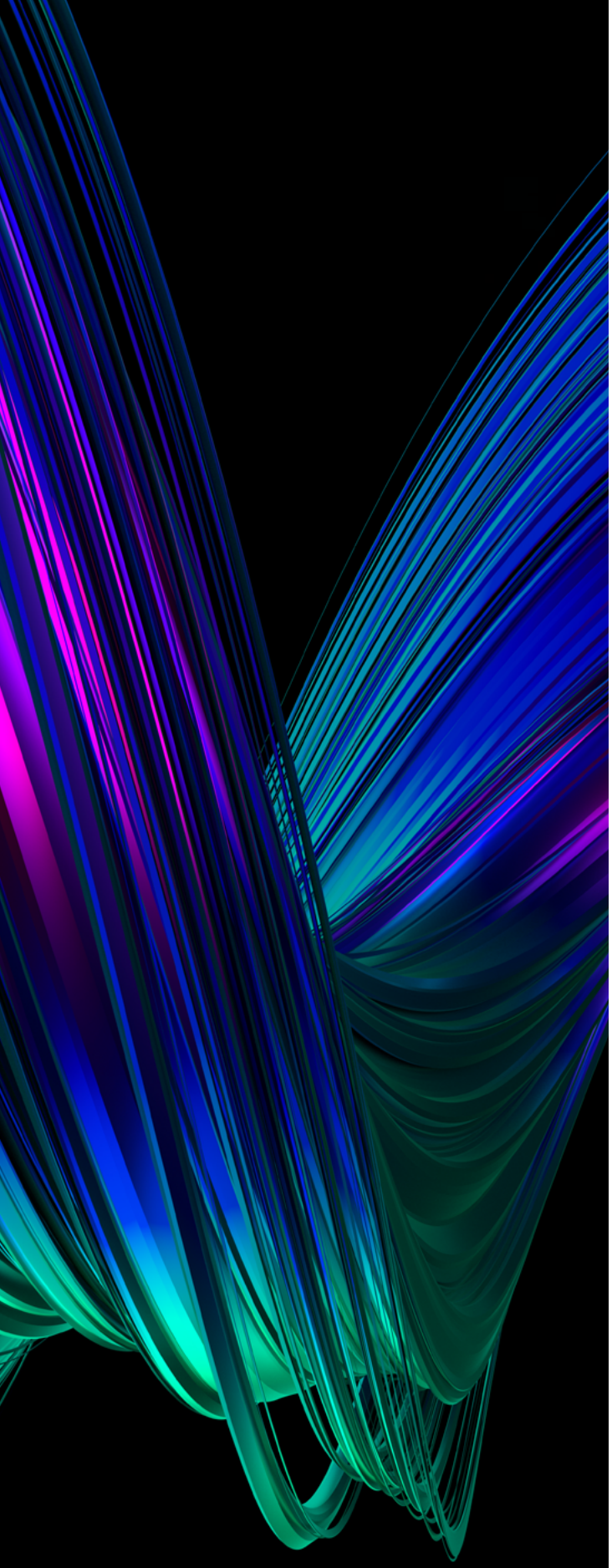
Implementation phases

Phase 1: **Data ingestion and storage setup**	• Configure data sources and set up Databricks Autoloader for data ingestion	• Store ingested data in Delta Lake bronze tables
Phase 2: **Transformation and quality layer**	• Set up Delta Live Tables to implement ETL workflows	• Apply data transformations and enforce quality checks
Phase 3: **Metadata and cataloging**	• Implement Unity Catalog for data cataloging, tagging, and lineage tracking	• Define access controls and classifications for each dataset
Phase 4: **Marketplace interface development**	• Create a user-facing portal for data discovery and access requests	• Provide preview, profiling, and quality score features
Phase 5: **Governance and security**	• Implement security controls, masking, and audit logging for data compliance	• Establish data governance policies for data ownership and usage
Phase 6: **Consumption enablement**	• Set up Databricks SQL, APIs, and BI tool integration	• Enable analytics and data science teams to access the data marketplace

Solution architecture overview

The architecture for a Databricks-powered data marketplace comprises the following layers:





Key benefits of Databricks data marketplace architecture

- **Data intelligent platform**: Unity Catalog and Databricks' Lakehouse architecture provide a single platform for data storage, transformation and analysis to reduce complexity
- **Scalable and performant**: The Lakehouse design allows the marketplace to scale seamlessly, supporting large volumes of data and high concurrency
- **Comprehensive governance**: Unity Catalog provides fine-grained access control, audit logging and data lineage, supporting regulatory compliance
- **Data quality assurance**: Delta Live Tables and data quality expectations help maintain high-quality data in the marketplace
- **Enhanced collaboration**: Data scientists, analysts and business users can work on the same platform, fostering cross-functional collaboration

What role does GenAI play in the data marketplace solution?

GenAI enhances the capabilities and value of data marketplace solutions by improving data accessibility, streamlining workflows and generating valuable insights. Here's how GenAI contributes to a data marketplace solution:

- **Data enrichment and synthesis**
  GenAI can enrich datasets by generating synthetic data that mimic real-world patterns without compromising privacy or security. This helps to augment existing data offerings and create new datasets that can be shared on the marketplace. For example, AI models can generate additional data points, fill in missing values or create simulated datasets that represent scenarios where real data is scarce or hard to collect

- **Personalized data recommendations**
  GenAI algorithms can analyze user behavior and preferences within the data marketplace and generate personalized data product recommendations. This helps consumers quickly find the most relevant datasets, models or solutions for their needs, improving the customer experience and engagement. GenAI can optimize product visibility and placement for providers, ensuring their offerings are exposed to the right audience

- **AI-generated data insights**
  Once a dataset is accessed, GenAI can automatically analyze the data and generate actionable insights. This could include identifying trends, anomalies or potential use cases. These AI-generated insights can be packaged and shared within the marketplace as value-added services, offering customers access to data, analysis and decision making

- **Data and model creation assistance**
  GenAI can assist data scientists and analysts by automating the process of creating, refining and testing ML models. In the context of a data marketplace, providers can leverage GenAI to accelerate the creation of high-quality models and solutions for consumers. This helps to bridge the gap between data availability and actionable insights, as consumers can access raw data and pre-trained AI models that can be deployed quickly

- **Enhanced search and discovery**
  GenAI-powered search tools can significantly improve how users discover market data and models. Through natural language processing (NLP), users can query the marketplace more intuitively (e.g., asking questions or describing data needs in natural language). GenAI can analyze these queries and return the most relevant results, allowing more efficient discovery of data products and solutions

- **Data governance and compliance automation**
  GenAI can also help automate data governance, compliance checks and quality assessments within a data marketplace. By analyzing datasets, GenAI can identify compliance issues, inconsistencies or potential privacy concerns, ensuring that the shared data adheres to required regulations. This helps build trust among consumers and provider, ensuring that data sharing is secure and ethical

- **Cost optimization and efficiency**
  GenAI can be used to optimize resource allocation for both consumers and providers. By predicting usage patterns and analyzing demand for specific datasets or models, GenAI can suggest pricing strategies, resource allocation and data-sharing methods that minimize costs and maximize value for all

User queries in the Databricks Marketplace

"How can I attract new customers and effectively monetize my listings on the Databricks Marketplace?"

We recognize that data product sales often require a high degree of customization and cannot always fit into a one-size-fits-all pricing structure. Consequently, all commercial transactions on the Databricks Marketplace are conducted directly between you and the consumer.

To help you reach potential customers, we recommend using the newly launched provider analytics dashboard. It enables you to monitor consumer engagement and track which listings attract the most attention. By analyzing key dashboard metrics, you can pinpoint high-potential leads and initiate tailored discussions around licensing and sales.

This data-driven approach supports a flexible and customized sales process and more effective optimization of listings and customer needs, leading to stronger engagement and improved commercialization opportunities.

What are the consumer benefits of using Databricks Marketplace?

Databricks Marketplace offers a diverse range of products, including datasets, ML models, notebooks, solution accelerators and, soon, applications. This wide variety enables consumers to leverage an array of analytics and AI initiatives to unlock the full potential of their data. With over 2,500 listings from 250+ providers, the marketplace offers unparalleled choice and accessibility.

One key benefit is its open approach, which allows consumers to access data without being restricted to the Databricks platform. This eliminates vendor lock-in, enabling users to access their preferred tools and platforms for data analysis and integration.

The marketplace is built on Delta Sharing, simplifying data access by removing the need for complex ETL processes and avoiding costly data replication. Consumers can seamlessly govern datasets from the marketplace alongside their existing Lakehouse data using Unity Catalog for consistent and streamlined data management. This integrated approach reduces the operational overhead, allowing customers to focus on extracting insights and driving value from their data instead of managing its complexities.

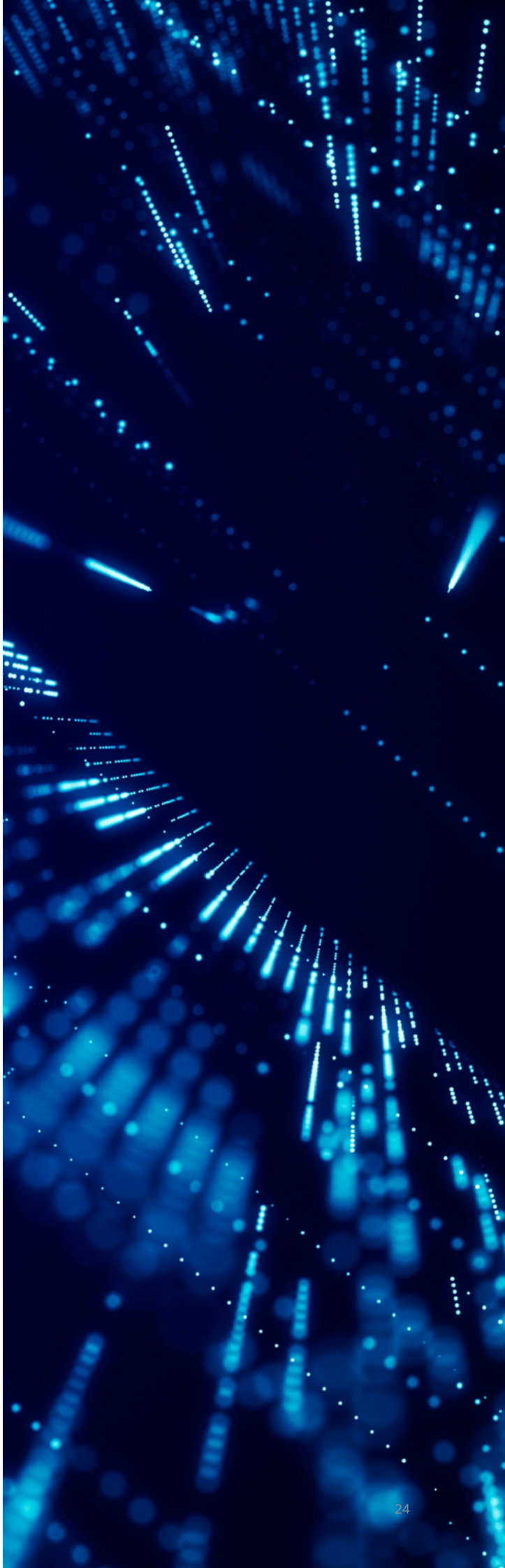
What are the benefits of offering my data products on Databricks Marketplace as a provider?

Key benefits of using Databricks Marketplace as a provider:

- **Maximum reach across platforms, regions and clouds**
  Databricks Marketplace enables you to reach a broader audience, unlike other marketplaces that often tie providers and recipients to proprietary systems. It allows data products to be shared across multiple platforms, cloud environments and regions without adopting a proprietary platform. This open approach increases the visibility and reach of your data products and enhances your ability to attract customers

- **Reduced total cost of ownership (TCO)**
  One of the standout features of Databricks Marketplace is the Delta Sharing protocol. This open-source standard allows data providers to share live, up-to-date data directly from their Databricks environment without complex data replication processes. This simplifies the data-sharing workflow and reduces the costs associated with data storage and duplication, resulting in a lower TCO

- **New revenue streams with AI models**
  In addition to sharing datasets, Databricks Marketplace enables providers to list AI models. By making your ML models available, you can tap into new revenue streams and provide customers with valuable pre-trained models to accelerate their AI initiatives. This opens additional business opportunities beyond traditional datasets



How does Databricks Marketplace stand out from other data marketplaces?

Databricks Marketplace stands out because of its unique combination of openness, flexibility and diverse offerings tailored to modern analytics and AI.

Key differentiators:

- **Powered by Delta Sharing**
  At its core, Databricks Marketplace leverages Delta Sharing, an open-source protocol that enables seamless and secure data sharing. This eliminates vendor lock-in, allowing users to benefit from the flexibility of open standards and the ability to integrate data with their preferred tools and platforms

- **A diverse range of offerings**
  Unlike traditional marketplaces that focus solely on structured data, Databricks Marketplace provides a variety of assets to support a broader range of use cases:

  - **Datasets (tables)**: Providers like Crisp offer high-quality, ready-to-use structured data.

  - **Volumes (unstructured data)**: Access datasets from providers like Shutterstock, including images, videos and audio files for multimedia analytics.

  - **AI models**: Providers like AI21 Labs offer pre-trained models that accelerate ML initiatives.

  - **Solution accelerators**: Resources like those from Datavant or Live Ramp provide prebuilt templates and tools to jumpstart specific analytics workflows.

  - **Notebooks**: Providers like Kythera Labs offer preconfigured notebooks for hands-on data exploration and modeling.

Cross-platform and multi-cloud accessibility
With Databricks-to-Open (D2O) Sharing, data can be shared across any platform or cloud infrastructure. This means consumers don't need to be on Databricks to access data, significantly expanding the reach for providers while ensuring flexibility for consumers.

•Enhanced integration and governance
For Databricks users, the marketplace integrates seamlessly with Unity Catalog, providing unified data governance across both marketplace datasets and existing Lakehouse data. This reduces operational complexity, ensuring secure, consistent access controls and metadata management.

•Support for analytics tools
Databricks Marketplace allows datasets and resources to be accessed through popular analytics tools such as Power BI and Apache Spark, ensuring compatibility with diverse workflows and technical ecosystems.

Why it matters

These features make Databricks Marketplace a versatile and powerful platform that caters to a wide range of data and analytics needs. Whether you're looking to enhance AI models, explore multimedia datasets or leverage prebuilt accelerators, the marketplace is designed to meet diverse requirements while promoting openness and ease of use.

How do I know which data product best fits my needs, and what evaluation criteria should I use?

The first step is to define your analytical objectives and data requirements. Understanding the specific problem you're trying to solve, the type of analysis you intend to perform, and any particular data attributes you need will guide your search. Once you have a clear vision, use the marketplace's robust filtering tools to narrow your options based on factors like product type, provider, category, model task and listing type.

For a comprehensive evaluation, take full advantage of Databricks Marketplace's resources, such as prebuilt notebooks and sample datasets. These tools allow you to conduct exploratory data analysis (EDA), helping you better understand how the data aligns with your objectives. For instance, many datasets come with accompanying notebooks that include step-by-step instructions and sample queries that allow you to explore the data's structure, assess its quality and test its suitability.

Take the HealthVerity Claims Sample Patient Dataset, for instance. This product is accompanied by a detailed notebook that outlines key use cases and demonstrates how the data can be used for healthcare analytics. By running the sample queries provided within the notebook, technical users can evaluate the dataset's schema, identify data quality issues and assess its applicability for their specific analytical tasks, such as predictive modeling or trend analysis.

By using these resources, you can make a better-informed decision about the best data product, in terms of technical compatibility and analytical value.

How do I become a provider and list my data and AI product assets on Databricks Marketplace?

There are two main ways to become a Databricks Marketplace provider, depending on your needs for visibility and accessibility.

1. Public listings through the Databricks data partner program

If you're looking to share your data and AI products with a broad audience, apply to the Databricks data partner program. Once accepted, you can access the provider console within your Unity Catalog-enabled Databricks workspace. This console allows you to create, manage and publish public listings, making your products available to a wide range of consumers. Public listing gives you maximum exposure, making it an ideal choice for providers aiming to scale their reach and grow their customer base.

2. Private exchange for select recipients

Providers who wish to share data exclusively with select Databricks recipients —across different accounts or internally — can become private exchange providers. If you're a marketplace admin, you can sign up through the provider console and set up a private exchange. This option allows you to share your assets securely with a specific audience, offering more control over who can access them. It's ideal for businesses that require tighter security or have specialized data products intended for a limited group of consumers.

Next steps

Follow the guidelines to upload and configure your listings, ensuring that you include relevant metadata, descriptions and licensing information. After your assets are listed, you can manage updates, track consumer engagement and monitor usage through the provider console.

By offering both public and private sharing options, Databricks Marketplace allows you to tailor your approach to data sharing and commercialization, whether aiming for broad exposure or needing more control over access.

Conclusion

A data marketplace built on Databricks represents a transformative shift in how organizations collaborate, share, and monetize their data. By providing a data intelligent platform that integrates data ingestion, scalable storage, advanced governance, analytics, and secure sharing, Databricks empowers organizations to unlock the full potential of their data assets.

With Databricks, businesses can seamlessly create a marketplace where data producers and consumers can engage in a trusted, transparent environment. Whether your goal is to generate new revenue streams by monetizing data, accelerate innovation through collaborative data sharing, or build a more data-driven organization, Databricks offers the robust infrastructure and comprehensive tools needed to realize these ambitions.

The scalability, security, and flexibility that Databricks provides make it an ideal solution for creating a future-proof data marketplace that can adapt to evolving business needs, regulatory requirements, and technological advancements. Ultimately, Databricks not only facilitates data-driven decision-making but also transforms data into a valuable economic asset, driving long-term growth and competitive advantage.

About **the author**

Arun Khandelwal
Enterprise Data Solution Architect,
Databricks Partner Solution Architect Champion
at DXC Luxoft

Arun has over 22 years of experience in data architecture and governance, analytics, engineering and modernization. Arun's expertise include building systems that leverage master data, reference data, microservices, SOA architectures, web APIs, data warehousing and optimization to support growth and business transformation. He re-architects, redesigns and migrates legacy systems, driving operational efficiency and scalability. His strategic vision and technical know-how make him a trusted leader in data-driven solutions and enterprise architecture.

About DXC Technology

DXC Technology (NYSE: DXC) helps global companies run their mission-critical systems and operations while modernizing IT, optimizing data architectures, and ensuring security and scalability across public, private and hybrid clouds. The world's largest companies and public sector organizations trust DXC to deploy services to drive new levels of performance, competitiveness, and customer experience across their IT estates.

Learn more about how we deliver excellence for our customers and colleagues at **DXC.com**.