



Especialización en

DATABRICKS ASSOCIATE

Programa de Certificación (Data Engineer Associate)

Online 100% en vivo



FORMACIÓN PARA LOS NUEVOS LÍDERES DE DATA & AI

DMC INSTITUTE, según el Estudio de Adopción de la Ciencia de Datos y Brecha de Talento 2025 de Apoyo Consultoría, es la institución no universitaria

Nº1

más valorada por

**Chief Data Officers y
empleadores del Perú**

para la formación en

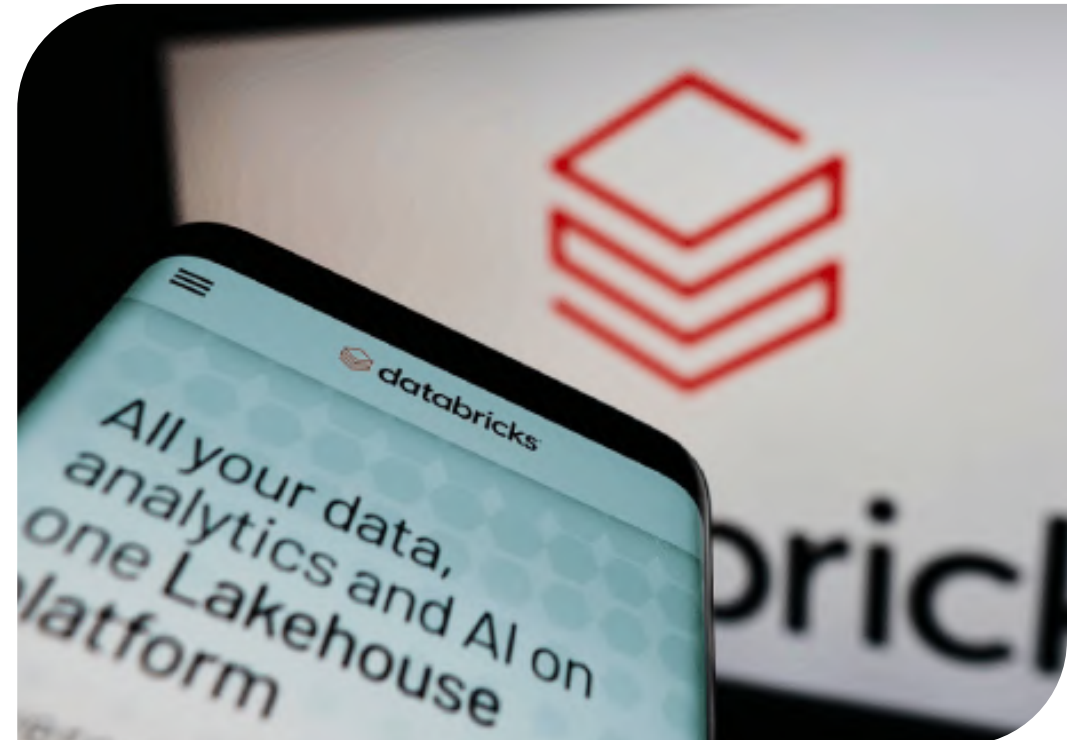
CIENCIA DE DATOS

Fuente: Estudio de Adopción de la
Ciencia de Datos y Brecha de Talento 2025.
Apoyo Consultoría

Presentación

¿Sabías que las arquitecturas modernas de datos ya no se construyen solo con ETL, sino con plataformas Lakehouse capaces de escalar, automatizar y gobernar datos de forma eficiente?

La Especialización en Databricks Associate de DMC Institute ha sido diseñada para que domines Databricks sobre Microsoft Azure desde un enfoque práctico y productivo, alineado al rol real de un Data Engineer y a los estándares que exige la certificación oficial Databricks Data Engineer Associate.



Sobre esta Especialización

14

sesiones

56

horas
académicas

13

proyectos para
tu portafolio

¿Cómo impulsamos tu carrera?

- Sesiones 80% **enfocadas en la práctica.**
- Enfoque en **Casos Aplicados a Negocio**, enfrentando los retos del mercado.
- Énfasis en **habilidades técnicas y blandas.**
- **Mentoría especializada** con docentes expertos.
- Acompañamiento **constante.**



¿Porqué estudiar esta Especialización?

Desarrolla competencias técnicas avanzadas para diseñar, construir y operar arquitecturas de datos modernas en Databricks sobre Azure, aplicando automatización, buenas prácticas de ingeniería de datos y gobierno, alineadas a entornos reales y a la certificación Databricks Data Engineer Associate.



Arquitecturas Lakehouse listas para producción

Diseña e implementa pipelines de datos completos usando Databricks y Delta Lake, aplicando arquitectura Medallion, ingesta incremental con Auto Loader, control de versiones, auditoría y recuperación ante fallos con Time Travel.



Automatización, CI/CD y operación real de pipelines

Aprende a orquestar y desplegar notebooks, jobs y pipelines declarativos usando Databricks Workflows y Asset Bundles, integrándolos con flujos CI/CD para entornos DEV, UAT y PRD.



Gobierno, seguridad y escalabilidad de datos

Configura modelos de gobierno con Unity Catalog, control de accesos, linaje de datos, calidad y federación, asegurando cumplimiento, trazabilidad y consumo seguro de la información en organizaciones reales.



Objetivo de la Especialización

- Usa Databricks para construir un Datalake en ADLS Gen2, cargar datasets multiformato y aplicar arquitectura Medallion con ETLs automatizados y time-travel.
- Empaqueta y despliega ETLs (notebooks, scripts y assets) mediante CI/CD, y configura datos federados manteniendo el linaje con Unity Catalog.
- Conoce la estructura del examen Databricks Data Engineer Associate: tipos de preguntas, inscripción y buenas prácticas.

Objetivo Final

Emplea Databricks en el entorno de Microsoft Azure para configurar ingesta, procesamiento, orquestación, automatización y gobierno de datos en un Datalake ADLS Gen2.

¿A quién está dirigido?

1. Data Engineers y Data Analysts con enfoque en ingeniería que ya trabajan con datos y buscan especializarse en Databricks para diseñar pipelines escalables, automatizados y gobernados en entornos cloud.
2. Profesionales de datos que desean certificar sus habilidades y prepararse específicamente para rendir la certificación Databricks Data Engineer Associate, fortaleciendo su perfil técnico y empleabilidad.
3. Arquitectos de datos, Analytics Engineers o Developers que interactúan con plataformas de datos y necesitan comprender la operación, automatización y gobierno de arquitecturas Lakehouse.
4. Profesionales de BI o Analytics que buscan evolucionar a roles de Data Engineering, adquiriendo una base sólida en procesamiento, orquestación y gobierno de datos en Databricks.



¿Cuáles son los requisitos?



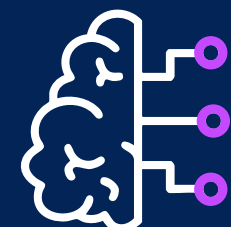
Conocimientos / Habilidades

- Manejo de conceptos básicos de ingeniería de datos.
- Manejo de Azure a nivel básico.



Experiencia Laboral

- (Ninguna).



Tecnológicos

- Contar con una laptop o computadora de escritorio con sistema operativo y disponibilidad de micrófono y cámara web. Contar con una Cuenta de Ms. Azure.

Herramienta



Microsoft
Azure



Databricks



Malla Curricular

I. Databricks Fundamentals & Data Ingestion

1. Fundamentos de Databricks

- Control plane vs. data plane y responsabilidades de cada uno.
- Integración con servicios de Azure: ADLS Gen2, Azure Key Vault, Azure Active Directory.
- Databricks Runtime: versiones LTS, compatibilidad con Spark, administración de librerías.

2. Componentes principales y flujo de trabajo

- Workspace y organización de artefactos: folders, notebooks, repos.
- Clusters: tipos (all-purpose, job, serverless), autoscaling, spot/preemptibles, políticas.
- Repos: integración con Git (GitHub/Azure DevOps), flujos de trabajo con ramas y PR.
- Jobs: programación, parámetros, permisos, mejores prácticas de operación.

3. Arquitectura Lakehouse y valor diferencial

- Lago de datos transaccional con Delta Lake (ACID).
- Beneficios frente a DW tradicionales: costos, elasticidad, gobernanza unificada, simplificación de ETL.

4. Proyecto 1. Configuración básica de un entorno productivo en Databricks

5. Conectividad y seguridad de acceso a datos

- Azure AD y Managed Identities: conceptos y patrones de autenticación.
- Montaje de ADLS Gen2: dbutils.fs.mount vs. acceso directo con ABFS.
- Gestión de secretos con Key Vault y Databricks Secrets.

6. Exploración de datos y formatos de archivo

- Lectura de CSV, JSON, Avro y Parquet con PySpark y SQL; inferencia vs. esquema explícito.
- Definición de esquemas con StructType/StructField y validaciones de tipos.
- Perfilado inicial: conteos, estadísticas, distribución de valores, detección de outliers.

7. Persistencia y reutilización

- Tablas gestionadas vs. externas; ubicaciones en UC.
- Vistas temporales y permanentes; caching y cacheTable.
- Buenas prácticas de naming y organización de objetos en Catálogo/Esquema.

8. Proyecto 2. Configuración de accesos y carga desde múltiples formatos

II. Databricks for Data Processing

9. Conceptos de ingesta incremental

- Batch vs. streaming; cuándo aplicar cada enfoque.
- Arquitectura de Auto Loader: list-based vs. file notification, implicancias.
- Esquema de metadatos y control de archivos procesados.

10. Configuración práctica de Auto Loader

- Propiedades: cloudFiles.format, cloudFiles.schemaLocation, cloudFiles.maxFilesPerTrigger, rescuedData-Column.
- Checkpoints, idempotencia, schema evolution y manejo de cambios de tipo.
- Implementación en PySpark/SQL; patrones de directorio y particionamiento.

11. Buenas prácticas de operación

- Re-procesos controlados, control de duplicados, cuarentena de archivos corruptos.
- Observabilidad: logs de ingesta, métricas y alertas.
- Costos y rendimiento: tuning de triggers y tamaño de micro-batches.

12. Proyecto 1. Configuración de Auto Loader para detección de nuevos archivos en el Lakehouse y carga incremental

13. Principios de diseño de Bronze (Raw Zone)

- Schema-on-read y preservación de fidelidad de origen.
- Particionamiento por tiempo/entidad; jerarquías de directorio recomendadas.
- Estrategias de tamaño de archivo (small files problem) y compaction.

14. Metadatos y auditoria

- Columnas técnicas: ingestion_timestamp, source_system, source_file, batch_id.
- Tablas de control: ingestion_log, audit_events, correlación con jobs.
- Trazabilidad desde origen a tabla (linaje básico).

15. Estándares operativos

- Naming conventions por dominio/esquema/tabla.
- Catálogo y esquema en Unity Catalog; ubicaciones externas vs. gestionadas.
- Políticas de retención y archivado de archivos fuente.

16. Proyecto 2. Implementación de capa Bronze con una tabla Delta con particiones por fecha, columnas técnicas y log de ingesta

Malla Curricular

17. Limpieza y estandarización

- Reglas de calidad: nullability, dominios válidos, trimming/normalización de strings.
- Conversión de tipos y estandarización de fechas/zonas horarias.
- Deduplicación basada en llaves naturales y ventanas.

18. Integración y modelado operacional

- Joins multi-fuente y resolución de conflictos.
- Generación de llaves sustitutas (surrogate keys) y mapeos de referencia.
- Enriquecimiento con dimensiones maestras.

19. Validación y contratos de datos

- Asserts y expectations (p.ej., con Delta constraints o librerías de validación).
- Mecanismos de rechazo y cuarentena de registros.
- Versionamiento de esquemas Silver.

20. Proyecto 3. Implementación de capa Silver y un pipeline PySpark para poblarlo aplicando limpieza de datos y deduplicación.

21. Modelado analítico

- Patrones de consumo: tablas wide vs. modelos dimensionales (hechos/dimensiones).
- Cálculo de KPIs, ventanas analíticas, periodizaciones y acumulados.
- Materialización de vistas y manejo de actualizaciones incrementales.

22. Rendimiento y consumo

- Particionamiento y clustering Z-Order en tablas de consulta.
- Gestión de índices estadísticos y OPTIMIZE programado.
- Publicación para BI: convenciones de nombres, estabilidad de esquemas.

23. Gobierno de la capa semántica

- Catálogo y permisos por dominio/rol.
- Control de cambios y compatibilidad hacia consumidores.
- Trazabilidad desde Gold hacia Silver/Bronze.

24. Proyecto 4. Implementación de la capa Gold y un pipeline para poblarlo aplicación funciones de agregación y cálculo.

25. Operaciones ACID, DDL/DML

- CREATE/REPLACE TABLE, ALTER TABLE, propiedades Delta.
- INSERT, UPDATE, DELETE, MER-GE INTO con condiciones.
- Constraints y columnas generadas.

26. Versionado y mantenimiento

- Time travel, DESCRIBE HISTORY, políticas de retención.
- OPTIMIZE, ZORDER BY, VACUUM seguro y consideraciones de compliance.
- Change Data Feed (CDF): habilitación y usos en incrementales.

27. Buenas prácticas

- Esquemas explícitos, operaciones idempotentes, manejo de concurrencia.
- Estrategias de compaction y tamaño de archivos objetivo.
- Observabilidad del estado de tablas Delta.

28. Proyecto 5. Simulación de flujo de correcciones con operaciones DML y uso de Time-travel para recuperar un estado previo

III. Databricks for Orchestration & Automatization (CI/CD)

29. Conceptos y componentes

- LakeFlow Connect, LakeFlow Jobs y Pipelines declarativos.
- Declarative ETL: diseño basado en intención (YAML).
- Recursos, tareas, dependencias y orquestación.

30. Definición y despliegue

- Estructura YAML: sources, transforms, targets, policies.
- Parametrización, variables por entorno y secretos.
- Validación, prueba local y despliegue en entornos.

31. Operación y monitoreo

- Ejecución programada, retries, SLAs y alertas.
- Métricas, logs y trazabilidad de runs.
- Integración con DAB/CI-CD para promoción entre entornos.

32. Proyecto 1. Implementación de Pipelines declarativos para el Data Warehouse.

33. Modelado de workflows

- Tipos de tareas: Notebook, SQL, Python, JAR, DBT, Delta Live Tables.
- Dependencias entre tareas (DAG), parámetros y salidas.
- Job clusters vs. all-purpose clusters, permisos y isolation.

34. Operación resiliente

- Políticas de reintento, "repair runs", manejo de fallos y compensaciones.
- Alertas, notificaciones y webhooks.
- Integración con repos y versiones de código.

Malla Curricular

35. Costos y seguridad

- Dimensionamiento de tareas, paralelismo y costos.
- Control de acceso a jobs y a credenciales.
- Auditoría de ejecuciones y logs centralizados.

36. Proyecto 2. Automatización de notebooks y notificación vía correo

37. Fundamentos de DAB

- Estructura de bundles: bundle.yml, assets, targets, variables.
- Environments: DEV, UAT, PRD; configuraciones diferenciales.
- Referencia de recursos: jobs, pipelines, clusters, permisos.

38. Integración con CI/CD

- Azure DevOps/GitHub Actions: pipelines de build y deploy.
- Estrategia de ramas, PRs, revisiones y aprobaciones.
- Versionado semántico y promoción controlada.

39. Seguridad y gobernanza en despliegues

- Manejo de secretos y referencias a Key Vault.
- Políticas de acceso a recursos y auditoría de despliegues.
- Rollback y estrategias de recuperación.

40. Proyecto 3. Empaquetado de notebooks, scripts y assets y despliegue con Pipelines DevOps.

41. Diagnostico con Spark UI

- Stages, tasks, shuffles, spills; lectura de DAGs.
- Métricas clave: tiempo por stage, I/O, GC, skew.
- Identificación de cuellos de botella (join, agregaciones, UDFs).

42. Técnicas de optimización

- AQE: coalescing, skew join optimization, repartition adaptativo.
- Join strategies: broadcast/hash/sort-merge; hints y tamaños umbral.
- Particionamiento y tamaño de archivos; OPTIMIZE y ZORDER.

43. Gestión de memoria y costos

- Caching/persist con criterio; evitar cache de datasets masivos volátiles.
- Dimensionamiento de clusters: núcleos, memoria, autoscaling, spot.
- FinOps: medir costo por job y oportunidad de ahorro.

44. Proyecto 4. Optimización con Spark UI mediante identificación de cuellos de botella, particionamiento inteligente y broadcast joins.

IV. Data Governance & Quality

45. Modelo de gobierno

- Metastore, Catalogs, Schemas, Tables/Views; objetos y jerarquía.
- Modelos de permisos: GRANT/REVOKE, roles y grupos.
- Data lineage y system tables para auditoría.

46. Integración con CI/CD Seguridad avanzada

- Data masking y row filters; políticas de acceso por columna/fila.
- External locations y storage credentials.
- Auditoría de accesos.

47. Buenas prácticas de gobernanza

- Naming estándar por dominio y entorno.
- Separación de duties: data owners, stewards, producers/consumers.
- Catálogos por dominio y zonas (bronze/silver/gold) con permisos específicos.

48. Proyecto 1. Configuración de permisos granulares y validación del linaje con Unity Catalog.

49. Delta Sharing

- Proveedores y receptores; shares, schemas, tables.
- Interno vs. externo; autenticación, permisos y revocación.
- Consideraciones de costos y seguridad en el intercambio.

50. Federación de datos: Lakehouse Federation

- Conectores a fuentes externas; consultas federadas.
- Casos de uso: minimizar replicación, acceso en tiempo real.
- Limitaciones, latencias esperadas y patrones de caché.

51. Operación y control

- Monitoreo del consumo, auditoría de accesos y cumplimiento.
- Versionado de contratos de datos y compatibilidad hacia clientes.
- Estrategias de gobierno multi-organización.

52. Proyecto 2. Consumo de Datasets compartidos, manejo de Delta Share y Recipients

V. Simulacro Databricks Data Engineer Associate

- Simulacro de Examen Oficial.
- Repaso de solucionario.
- Tips y recomendaciones finales para rendir el Examen Oficial.

Certificación DMC INSTITUTE

Por aprobación de la Especialización en Databricks Associate, por un total de 56 horas académicas.

DMC | INSTITUTE

CERTIFICADO

SE OTORGA A

Nombres Y Apellidos

Por haber aprobado la Especialización en

DATABRICKS ASSOCIATE

CLASES EN VIVO

Realizado del 00 de mes al 00 de mes,
completando un total de 56 horas académicas



MBA Jonny Chambi
GERENTE GENERAL



MBA Dionicio Velásquez
DIRECTOR ACADÉMICO

Mes del 2026



000000

METODOLOGÍA LEARNING AGILE

Prepárate para nuevos retos, aprende con nuestro método de primer nivel y lánzate a conquistar el mundo.



¿Por qué elegirnos?

+16

Más de 16 años de experiencia.

+300

Más de 300 empresas asesoradas en Perú, Ecuador y Bolivia.

35k

35 mil profesionales capacitados en más de 20 países de América Latina.



Propuesta integra en formación en Data & AI.

+150

Más de 150 docentes expertos de Latinoamérica, España y Estados Unidos.



Comunidad más grande en Data & AI con beneficios exclusivos: Networking, empleabilidad, habilidades blandas.



Excelente nivel de servicio.



Nuestros Partners

CertiProf® | Partner

Google Partners



Estas empresas confían en nosotros



BBVA



ANTAMINA



PROM PERÚ



SCOTIABANK



PACÍFICO
SEGUROS



SUNAT



CAJA
HUANCAYO



BUENAVENTURA



PRONABEC



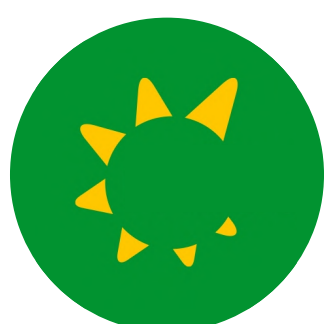
CAJA
AREQUIPA



RIMAC



BCRP



MIBANCO



MAPFRE



ONCOSALUD



LOS ANDES

Métodos de pago

J&J DATA MINING CONSULTING S.A.C.
RUC: 20520972740

1. Depósito en cuenta BCP

- Corriente soles BCP: **193-225-1181-0-01**
- CCI BCP: **00219300225118100116**
- Corriente BCP dólares: **193-2318515-1-52**
- CCI BCP dólares: **002-193-002318515152-11**

2. Depósito en cuenta BBVA

- Ahorros BBVA soles: **0011-0177-02-00180473**
- CCI BBVA: **011-177-000200180473-37**

2. Pago Online

Generamos un link de pago online donde se acepta todas las tarjetas.

3. Pago con Yape

A nombre de J J Data
Mining Consulting Sac



4. Pago online por PayPal



06 CUOTAS SIN INTERESES pagando con:





Visita nuestra web

www.dmc.pe

Síguenos en:

