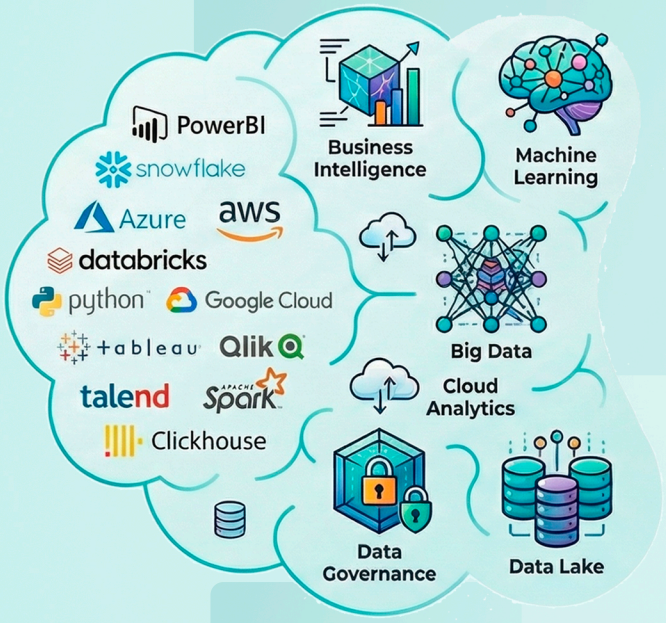
Cursos de

Data Analytics

TECNOLOGÍAS Y ÁREAS DE ESPECIALIZACIÓN



BENEFICIOS CLAVE DE LA FORMACIÓN

FORMACIÓN 100% PRÁCTICA Y APLICADA

Cursos basados en casos de uso reales con entrega de código y plantillas propias.

MODALIDAD FLEXIBLE CON TODO INCLUIDO

Elige entre presencial u online, con todo el software e infraestructura cloud incluidos.

+100 ORGANIZACIONES Y +1.500 PERSONAS FORMADAS

Experiencia probada y confianza depositada por empresas y profesionales del sector.

stratebi

Cursos de:

Azure, Spark, Python, Databricks, AI, PowerBI, Pentaho, Talend, AWS, Vertica, Clickhouse, LinceBI, Kafka, Business Intelligence, Data Lake, Snowflake, Machine Learning, Augmented Analytics, Big Data, Cloud Analytics, Data Governance, Sports Analytics... y muchos más.

- Para grupos y empresas
- Con tutorías y ejercicios prácticos el mes posterior para poner en práctica lo aprendido
- Todo el software e infraestructura cloud incluido en la formación para la realización de ejercicios
- Subvencionable por Fundae
- Documentación incluida
- Certificado de realización
- Totalmente práctico
- Incluyendo información de novedades, tendencias, costes de las herramientas, integración entre tecnologías y stacks...
- Basado en casos de usos reales
- Presencial/Online 100% con profesor
- Más de 100 organizaciones y 1.500 personas formadas
- Entrega de código, plantillas y desarrollos propios

Consigue tu camiseta:

Curso de

Introducción a PowerBI

PowerBI

Aprende a crear dashboards e informes con la mejor herramienta de analytics del mercado.

Temario

1. Introducción al concepto de Business Intelligence

2. Análisis de fuentes de datos

3. Introducción a Power BI

- Entorno de trabajo para Power BI Desktop
- Componentes: Power BI Desktop y Power BI Servicio Cloud
- Tareas: Conectar, integrar, modelar y visualizar
- Cuadros de mando (Paneles) e informes
- Más funcionalidades del entorno Power BI
- Paquetes de contenido y aplicaciones

4. Conectar

- Editor de consultas de Power BI
- Extracción de datos: Extracción vs Direct Query
- Conectar datos alojados en diferentes orígenes
- Realizar transformaciones básicas sobre los datos en la consulta
- Enlazar datos desde la consulta

5. Modelar

- Entorno de trabajo para modelar con Power BI
- Introducción al modelado tabular con Power BI
- Tablas y relaciones
- Introducción a fórmulas DAX
- Columnas calculadas y medidas
- Tablas calculadas
- Fórmulas DAX de inteligencia de tiempo (YTD, PreviousQuarter...)

6. Visualizar

- Entorno de trabajo para creación de gráficos con Power BI
- Trabajar con distintos tipos de gráficos
- Formatos para gráficos e informes
- Importación de visualizaciones extra desde el Office Store

7. Conectividad y Colaboración

8. Power BI Mobile (Alertas, notificaciones, favoritos)

Stratebi Dashboard Demos

Papers Técnicos y Buenas Prácticas incluidas con el Curso

Tips Vol.1 | Tips Vol.2 | Tips Vol.3 | Tips Vol.4

- Integracion SAP - PowerBI
- PowerBI Trucos (Vol I)
- PowerBI Trucos (Vol II)
- PowerBI + Synapse Analytics (paso a paso)
- 30 Consejos y Buenas Prácticas para hacer un proyecto de Power BI con éxito
- Cómo crear diseños de Dashboards espectaculares con PowerBI
- Videotutorial: Trabajando con Python en Power BI
- Aplicación PowerBi Turismo
- Aplicación PowerBI Financiera I
- Bravo para PowerBI
- Aplicación PowerBI Financiera II
- Aplicación PowerBI eCommerce
- Aplicación PowerBI Salud
- Aplicación PowerBI Smart City
- Aplicación PowerBI Energía
- Aplicación PowerBI Sports Analytics
- Power BI Premium Utilization and Metrics
- PowerBI Embedded: Funcionamiento y costes
- Guía para integrar Power BI con Microsoft Dynamics y Salesforce
- SQL Server Profiler para Power BI
- Como usar Report Analyzer en PowerBI, para mejorar el rendimiento
- Power BI embebido en Jupyter Notebook
- Tabular Editor para Power BI: Videotutorial y manual en español
- Personaliza tus gráficas en Power BI con Charticulator y Deneb
- Comparativa PowerBI vs Amazon QuickSight

Curso de

PowerBI avanzado

PowerBI

Herramientas de análisis empresarial.

Temario

1. Lenguaje DAX

- Contextos de evaluación en DAX
- Medidas y columnas calculadas
- Creación de tablas
- Operaciones lógicas
- Operaciones matemáticas y estadísticas
- Operaciones de inteligencia de tiempo
- Operaciones con cadenas de texto
- Otras funciones DAX
- Seguridad a nivel de fila (RLS)
- Dax Quiz
- Ejercicios

2. Dataflows

- Introducción a Dataflows
- Conceptos de Dataflows
- ETL con Dataflows
- Buenas prácticas
- Lenguaje M
- Ejercicios

3. External Tools

- Tabular Editor
  - Introducción a Tabular Editor
  - Grupos calculados
  - Perspectivas
  - Avance Scripting
  - Ejercicios
- DAX Studio
- ALM Toolkit

4. Conceptos avanzados

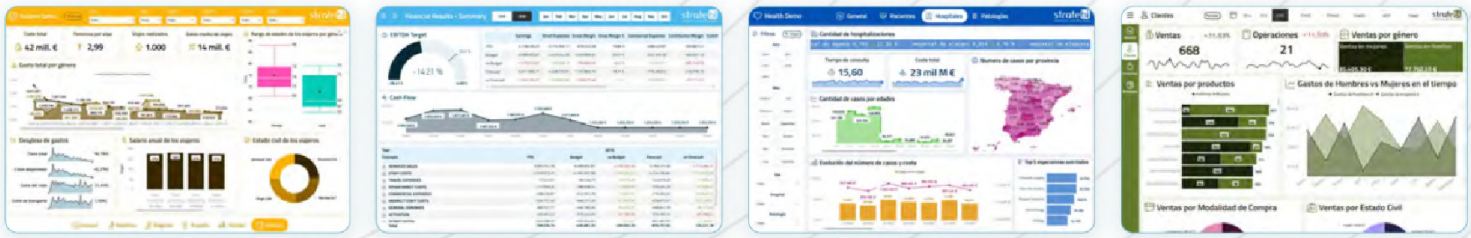
- Data Gateways
- Agregaciones
- Modelos duales
- Machine Learning en Power BI
- Python en Power BI
- R en Power BI
- Microsoft Teams con Power BI
- Power BI Report Builder

5. Buenas prácticas Microsoft Power BI

6. Ejercicios (Opcionales)

TABULAR EDITOR

python



Stratebi Dashboard Demos

Papers Técnicos y Buenas Prácticas incluidas con el Curso

Tips Vol.1 | Tips Vol.2 | Tips Vol.3 | Tips Vol.4

- Como usar emoticonos en PowerBI
- Buenas prácticas con Dataflows en Power BI
- Power Automate para Power BI: Cómo funciona
- ALM Toolkit para Power BI
- PowerBI Trucos (III)
- Os presentamos Goals in Power BI para hacer Scorecards
- Tutorial gratuito en español sobre Power BI Report Builder
- Conoce PowerBI Diagram View (Visual Data Prep): paso a paso
- Big Data para PowerBI
- Fútbol Analytics, lo que hay que saber
- Dashboard de medición de la calidad del aire en Madrid
- ¿Cómo funciona Microsoft Power BI? Videotutorial de introducción
- Cómo integrar Salesforce y PowerBI
- ¿Quieres crear aplicaciones empresariales usando PowerBI, PowerApps y Power Automate de forma conjunta?
- Power BI tip: Uso de parámetros what-if
- Videotutorial: Usando R para Machine Learning con PowerBI
- Las 50 claves para aprender y conocer PowerBI
- PowerBI: Arquitectura End to End
- Usando Python con PowerBI
- Todas las presentaciones del Workshop 'El Business Intelligence del Futuro'
- PowerBI + Open Source = Sports Analytics
- Comparativa de herramientas Business Intelligence
- Use Case Big Data "Dashboards with Hadoop and Power BI"
- Descarga Paper gratuito: Zero to beautiful (Data visualization)
- SAP connection tools for process automation: Microsoft, Pentaho, Talend (User Guide)

Curso de Apache HOP

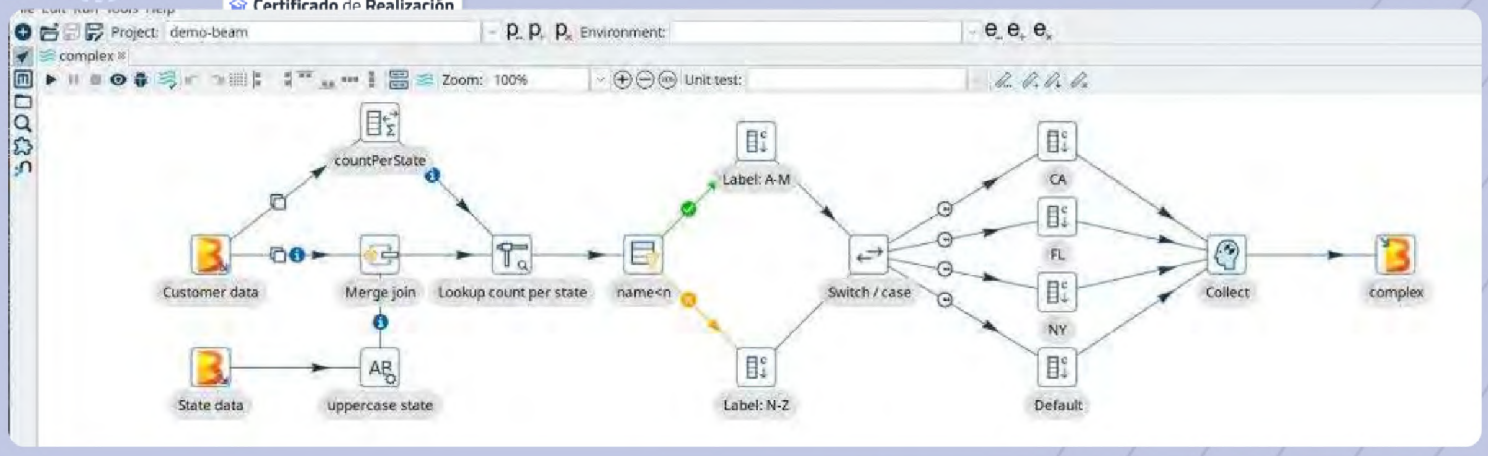
Curso Técnico: Apache Hop - Diseño y Gestión de Flujos ETL

Consultar para fechas y coste: info@stratebi.com

Duración de **4 jornadas (20 horas) | Distribuible** en **2 semanas** | Tanto **Presencial** como **Online** | **Totalmente práctico**

Plataforma Online preparada para **ejercicios** de los **asistentes** | **Set de datos y modelos** para los **ejercicios**

Certificado de Realización



Módulo 1: Introducción a Apache Hop

Temas

- ¿Qué es Apache Hop? Origen y evolución.
- Principios de Apache Hop: simplicidad, portabilidad y flexibilidad.
- Diferencias clave entre Apache Hop y Pentaho Data Integration (PDI).
- Componentes principales: Hop GUI, Hop Run y Hop Server.

Ejercicio práctico

- Instalar Apache Hop en un entorno local y explorar Hop GUI.
- Realizar una comparación funcional entre Apache Hop y PDI mediante tareas simples como creación de transformaciones básicas.

Módulo 2: Conceptos Básicos de Apache Hop

Temas

- Introducción a los Pipelines (transformaciones) y Workflows (procesos).
- Organización del trabajo con Proyectos y Entornos en Hop.
- Arquitectura sin código: diseño visual de ETL con Hop GUI.

Ejercicio práctico

- Crear un proyecto en Apache Hop, definir un entorno y desarrollar un pipeline básico para leer datos desde un archivo CSV.

Módulo 3: Diseño de Pipelines

Temas

- Conexiones a bases de datos (MySQL, PostgreSQL, etc.).
- Fuentes de datos comunes: archivos planos, bases de datos, APIs.
- Transformaciones clave:
  - Filtros y uniones.
  - Conversión de datos y cálculo de campos derivados.
  - Agregación y ordenación de datos.
- Manejo de variables y parámetros en los pipelines.

Ejercicio práctico

- Diseñar un pipeline que:
  - Ingesta datos desde un archivo CSV y una base de datos SQL.
  - Realiza transformaciones (filtros y cálculos).
  - Escribe los resultados en un archivo JSON.

Módulo 4: Workflows en Apache Hop

Temas

- Diferencias entre Workflows y Pipelines.
- Componentes de un Workflow:
  - Tareas (Tasks) y Condiciones.
- Automatización de procesos ETL.
- Integración de Workflows con Pipelines.

Ejercicio práctico

- Crear un Workflow que coordine:
  - La ejecución de un pipeline de ETL.
  - El envío de una notificación por correo al finalizar.

Módulo 5: Apache Hop Server y Automatización

Temas

- Configuración de Apache Hop Server.
- Ejecución remota de Pipelines y Workflows.
- Programación y orquestación de tareas con Apache Hop.

Ejercicio práctico

- Configurar un servidor Apache Hop y programar un Workflow para que se ejecute diariamente.

Módulo 6: Apache Hop y Big Data

Temas

- Integración de Apache Hop con ecosistemas de Big Data:
  - Apache Hadoop.
  - Apache Spark.
- Procesamiento de grandes volúmenes de datos con Apache Hop.

Ejercicio práctico

- Diseñar un pipeline que procese datos desde un clúster Hadoop y almacene los resultados en una base de datos SQL.

Módulo 7: Comparación Apache Hop vs Pentaho Data Integration (PDI)

Temas

- Diferencias clave en arquitectura y enfoque:
  - Proyectos y entornos en Hop vs Variables en PDI.
  - Modularidad y extensibilidad de Hop.
- Rendimiento y compatibilidad con entornos modernos (Kubernetes, Docker).
- Migración de transformaciones y jobs de PDI a Hop.

Ejercicio práctico

- Migrar un pipeline desarrollado en PDI a Apache Hop y comparar su funcionalidad y rendimiento.

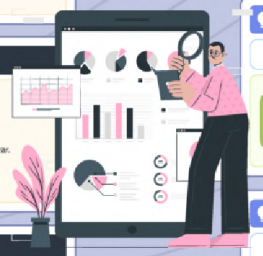
Módulo 8: Pruebas, Depuración y Buenas Prácticas

Temas

- Introducción a los Pipelines (transformaciones) y Workflows (procesos).
- Organización del trabajo con Proyectos y Entornos en Hop.
- Arquitectura sin código: diseño visual de ETL con Hop GUI.

Ejercicio práctico

- Crear un proyecto en Apache Hop, definir un entorno y desarrollar un pipeline básico para leer datos desde un archivo CSV.



Módulo 9: Extensibilidad y Personalización

Temas

- Extender Apache Hop con plugins personalizados.
- Creación de nuevas transformaciones y tareas.
- Integración con sistemas externos mediante scripts en Python y JavaScript.

Ejercicio práctico

- Crear un plugin personalizado para leer datos desde una API REST y procesarlos en un Pipeline.

Módulo 10: Proyecto Final

Objetivo

- Diseñar una solución ETL completa con Apache Hop que integre datos de múltiples fuentes, realice transformaciones complejas y genere un informe automatizado.

Ejercicio práctico

- Implementar un proyecto que:
  - Cargue datos desde diferentes fuentes (CSV, SQL y API).
  - Procese y transforme los datos.
  - Genere un archivo final para ser visualizado en una herramienta externa como Power BI, Tableau, Superset o LinceBI.

Curso de Data Lake Open Source

Curso Técnico: Creación de un Data Lake Open Source

Consultar para fechas y coste: info@stratebi.com

Duración de **5 jornadas (25 horas)** | **Distribuible** en **2 semanas** | Tanto **Presencial** como **Online** | **Totalmente práctico**

Plataforma Online preparada para **ejercicios** de los **asistentes** | **Set de datos y modelos** para los **ejercicios**

Con este temario, los participantes adquirirán un conocimiento técnico sólido y aplicable para diseñar, implementar y gestionar un Data Lake Open Source usando tecnologías modernas y escalables.

Módulo 1: Fundamentos del Data Lake

Temas

- Conceptos básicos. ¿Qué es un Data Lake? ¿Cómo se diferencia de un Data Warehouse?
- Beneficios del uso de tecnologías Open Source en un Data Lake.
- Revisión de la arquitectura general de un Data Lake Open Source.

Ejercicio práctico

- Analizar la arquitectura propuesta e identificar cada componente y su rol en el sistema.

Módulo 2: Configuración del Entorno Inicial

HAPROXY · træfik

Temas

- Instalación de un servidor base en Ubuntu Server 22.04.
- Configuración de Traefik y HAProxy como balanceadores de carga y proxy inverso.
- Seguridad básica: configuración de UFW, Fail2Ban y certificados SSL con Let's Encrypt.

Ejercicio práctico

- Instalar Traefik y HAProxy en una máquina virtual y configurarlos para balancear tráfico hacia un servicio simulado.

Módulo 3: Almacenamiento Centralizado con MinIO

MINIO

Temas

- Introducción a MinIO como sistema de almacenamiento distribuido compatible con S3.
- Instalación y configuración de un clúster de MinIO.
- Gestión de datos no estructurados en el Data Lake.

Ejercicio práctico

- Configurar un clúster de MinIO y cargar un conjunto de datos (archivos JSON o CSV).

Módulo 4: Implementación del Data Warehouse con ClickHouse

ClickHouse · MINIO

Temas

- Introducción a ClickHouse: características y casos de uso.
- Configuración de un clúster de ClickHouse para análisis de datos a gran escala.
- Consultas SQL avanzadas en ClickHouse.

Ejercicio práctico

- Instalar y configurar un clúster de ClickHouse con múltiples nodos. Cargar datos desde MinIO y realizar consultas analíticas.

Módulo 5: Procesamiento en Batch, Real-Time y Machine Learning con Apache Spark

ClickHouse · MINIO · Spark

Temas

- Fundamentos de Apache Spark y su integración con un Data Lake.
- Instalación de un clúster de Spark (Master y Workers).
- Procesamiento en tiempo real y batch: Spark Streaming.
- Introducción al uso de Spark para tareas de Machine Learning.

Ejercicio práctico

- Desarrollar un pipeline en Spark que procese datos almacenados en MinIO y los transforme para su análisis en ClickHouse.

Módulo 6: Orquestación de Flujos de Trabajo con Apache Airflow

Airflow · MINIO · ClickHouse · Spark

Temas

- Introducción a Apache Airflow y su uso en la orquestación de flujos de trabajo.
- Creación y ejecución de DAGs (Directed Acyclic Graphs).
- Integración de Airflow con MinIO, Spark y ClickHouse.

Ejercicio práctico

- Diseñar y ejecutar un DAG en Airflow que ingesta datos desde MinIO, procese con Spark y almacene resultados en ClickHouse.

Módulo 7: Indexación y Búsqueda con Elasticsearch

kibana · elasticsearch

Temas

- Configuración de un clúster de Elasticsearch para indexación de datos.
- Introducción a Kibana: creación de visualizaciones interactivas.
- Integración de Elasticsearch con otros componentes del Data Lake.

Ejercicio práctico

- Indexar un conjunto de datos en Elasticsearch y crear visualizaciones en Kibana.

Módulo 8: Gobernanza de Datos con DataHub

ClickHouse · DataHub · MINIO

Temas

- Introducción a la gobernanza de datos y el linaje.
- Instalación y configuración de DataHub como catálogo de datos.
- Integración de DataHub con MinIO y ClickHouse.

Ejercicio práctico

- Catalogar un conjunto de datos en DataHub y documentar su linaje desde la ingesta hasta el análisis.

Módulo 9: Exploración y Análisis de Datos con JupyterHub

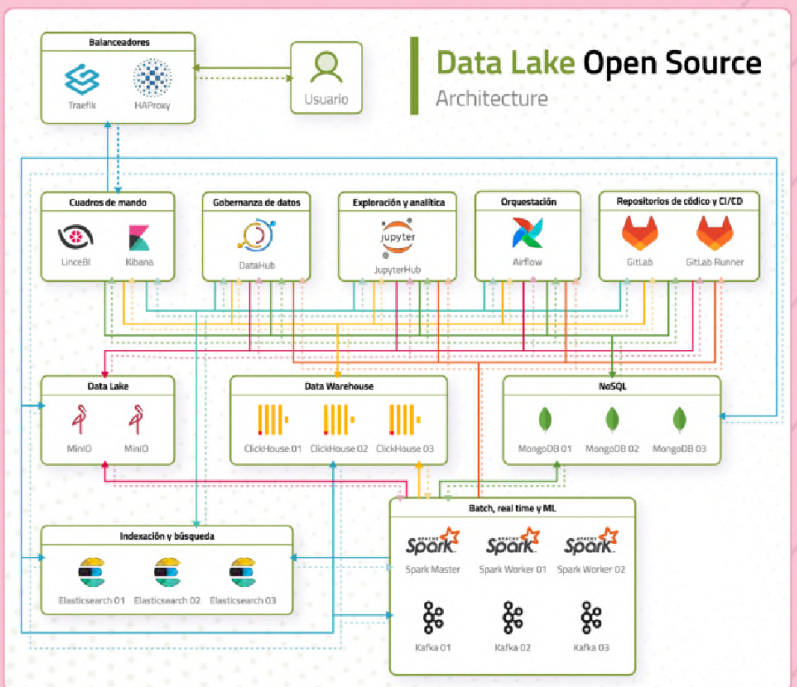
jupyterhub · MINIO · ClickHouse

Temas

- Configuración de JupyterHub para análisis colaborativo.
- Uso de Python para conectar JupyterHub con MinIO y ClickHouse.
- Exploración interactiva y visualización de datos.

Ejercicio práctico

- Crear un notebook en JupyterHub que conecte con MinIO y realice análisis exploratorio en ClickHouse.



Módulo 10: Creación de Dashboards e Informes con LinceBI y Kibana

LinceBI · kibana · ClickHouse

Temas

- Introducción a LinceBI como herramienta de inteligencia de negocio Open Source.
- Creación de cuadros de mando con LinceBI y Kibana.
- Publicación de dashboards interactivos para los usuarios.

Ejercicio práctico

- Diseñar un dashboard en LinceBI que combine datos de ClickHouse y otras fuentes de datos.

Módulo 11: Gestión del Código y CI/CD con GitLab

GitLab · GitLab RUNNER · Airflow

Temas

- Uso de GitLab para gestión de código y versionado.
- Configuración de GitLab Runner para pipelines de CI/CD.
- Despliegue automatizado de componentes del Data Lake.

Ejercicio práctico

- Configurar un repositorio en GitLab y crear un pipeline de CI/CD para desplegar un DAG en Airflow.

Módulo 12: Procesamiento en Tiempo Real con Apache Kafka

kafka · ClickHouse · Spark

Temas

- Introducción a Apache Kafka como sistema de mensajería en tiempo real.
- Configuración de un clúster de Kafka con múltiples brokers.
- Integración de Kafka con Spark para procesamiento en tiempo real.

Ejercicio práctico

- Configurar un flujo de datos en tiempo real utilizando Kafka y procesarlo con Spark para almacenamiento en ClickHouse.

Módulo 13: Seguridad y Monitoreo

Prometheus · Grafana · Spark · ClickHouse

Temas

- Configuración de permisos y roles en los distintos componentes del Data Lake.
- Uso de herramientas de monitoreo como Prometheus y Grafana.
- Estrategias para asegurar la alta disponibilidad y la recuperación ante fallos.

Ejercicio práctico

- Implementar un sistema de monitoreo para el clúster de Spark y ClickHouse, configurando alertas en caso de fallos.

Módulo 14: Proyecto Final

Airflow · MINIO · Spark · ClickHouse · kafka · elasticsearch · jupyterhub

Objetivo

- Construir un Data Lake funcional integrando todos los componentes del curso.
- Implementar un caso de uso real que abarque desde la ingesta, procesamiento, análisis e indexación hasta la creación de dashboards.

Ejercicio práctico

- Diseñar e implementar una solución completa utilizando MinIO, Spark, ClickHouse, Kafka, Elasticsearch, Airflow y JupyterHub.

Cursos de databricks | Microsoft Fabric | snowflake

snowflake

PARTNER | stratebi Analytics and Big Data

Capítulo 1 — Introducción a Snowflake y su Arquitectura

1. ¿Qué es Snowflake? Historia y evolución del Data Cloud
2. Arquitectura multi-clúster en la nube: separación de almacenamiento y cómputo
3. Diferencias entre Snowflake en **AWS, Azure y Google Cloud**
4. Conceptos clave: Virtual Warehouses, Databases y Schemas
5. Seguridad y gobernanza en Snowflake: roles, permisos y cifrado de datos
6. **Ejercicio:** Crear una cuenta en Snowflake, configurar un **Virtual Warehouse** y cargar datos en una tabla

Capítulo 2 — Almacenamiento y Gestión de Datos en Snowflake

1. Tipos de almacenamiento en **Snowflake**: Databases, Tables, Stages y External Tables
2. **Snowflake File Formats:** Ingesta de datos desde CSV, Parquet y JSON
3. **Ingesta de datos en streaming** con Snowpipe y Auto Ingest
4. **Time Travel y Fail-safe:** Recuperación de datos históricos
5. Uso de **Zero-Copy Cloning** y **Data Sharing** para colaboración entre cuentas
6. **Ejercicio:** Crear una tabla, cargar datos desde un Stage, restaurar datos con Time Travel

Capítulo 3 — Transformación y Procesamiento de Datos en Snowflake

1. Uso de **Snowflake SQL** para transformación de datos
2. **Streams & Tasks:** Automatización de pipelines de datos
3. **Materialized Views y Clustering** para optimización de consultas
4. **Stored Procedures y User-Defined Functions (UDFs)** en SQL y Python
5. Uso de **Dynamic Tables** para mantener datos actualizados sin ETL complejo
6. Ejercicio: Crear un **Stream** y un **Task** para capturar cambios en datos en tiempo real

Capítulo 4 — Integración de Snowflake con Herramientas Externas

1. **Conectores y API:** Snowflake Connector for Python, JDBC y ODBC
2. **Integración con Apache Spark y Databricks**
3. **Uso de Snowflake con herramientas de BI:** Power BI, Tableau y Looker
4. **Snowpark:** Desarrollo de código en Python y Java dentro de Snowflake
5. **Uso de Snowflake con Data Lakes:** External Tables y Iceberg Tables
6. **Ejercicio:** Conectar Snowflake con Python y ejecutar consultas con el Snowflake Connector

Capítulo 5 — Machine Learning e Inteligencia Artificial con Snowflake

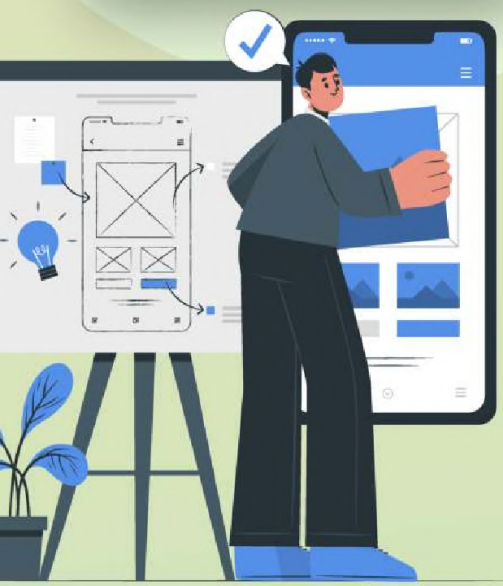
1. Introducción a **Snowflake Cortex AI:** Modelos preentrenados y funciones de IA
2. **Uso de Snowpark ML** para entrenar modelos dentro de Snowflake
3. Integración con **Hugging Face y TensorFlow**
4. **Feature Engineering** en Snowflake con funciones avanzadas de SQL y Python
5. **MLOps en Snowflake:** Entrenamiento, versionado y despliegue de modelos
6. **Ejercicio:** Crear y ejecutar un **modelo de Machine Learning** dentro de Snowflake con Snowpark ML

Capítulo 6 — Integración de Snowflake con Herramientas Externas

1. **Data Sharing:** Compartición de datos en tiempo real sin duplicaciones
2. **Snowflake Marketplace:** Acceso y publicación de datasets
3. **Data Governance:** Uso de **Object Tagging** y **Data Classification**
4. **Masking Policies y Row Access Policies:** Protección de datos sensibles
5. **Auditoría y Monitoreo** con Query History y Access History
6. **Ejercicio:** Configurar un **Data Share** entre cuentas de Snowflake y aplicar una **Masking Policy**

Capítulo 7 — Optimización del Rendimiento y Costos en Snowflake

1. Estrategias de optimización de consultas: **Micro-partitioning** y **Clustering**
2. **Query Performance Tuning:** Uso de Query Profile para análisis de ejecución
3. **Resource Monitors:** Control de costos y límites de cómputo
4. **Auto-suspend y Auto-scale:** Ajuste automático de Virtual Warehouses
5. **Optimización de almacenamiento** con Time Travel y Retention Policies
6. **Ejercicio:** Analizar y optimizar una consulta con **Query Profile**



databricks

PARTNER | stratebi Analytics and Big Data

Capítulo 1 — Introducción a Databricks y su Arquitectura

1. ¿Qué es Databricks? Historia y evolución
2. Arquitectura de Databricks: Lakehouse, Apache Spark y Mlflow
3. Integración con **AWS, Azure y Google Cloud**
4. Seguridad y administración de accesos en Databricks
5. Clústeres en Databricks: configuración, tipos y optimización
6. El motor optimizado de ejecución en Databricks
7. **Ejercicio:** Configurar un entorno Databricks y ejecutar consultas en un Notebook

Capítulo 2 — Almacenamiento y Manejo de Datos en Databricks

1. Almacenamiento en Databricks: DBFS, Delta Lake y external tables
2. Uso de **Auto Loader** para ingesta de datos en streaming
3. Administración de formatos de datos: CSV, Parquet, Delta y Avro
4. Optimización con **Delta Lake**: Compaction, Z-Ordering y Vacuum
5. Integración con Data Warehouses y Data Lakes externos
6. Uso de **Unity Catalog** para gobernanza de datos
7. **Ejercicio:** Crear y optimizar una tabla Delta con **Auto Loader**

Capítulo 3 — Transformación de Datos con PySpark y Apache Spark en Databricks

1. Introducción a **Apache Spark** en Databricks
2. Uso de **PySpark:** creación de DataFrames y transformación de datos
3. Aplicación de **Spark SQL** en Databricks
4. Optimización de consultas con **Catalyst Optimizer** y **Tungsten**
5. Introducción a **Photon Engine** y **Genie** para procesamiento optimizado
6. Creación de Jobs y Pipelines ETL con Databricks Workflows
7. **Ejercicio:** Construir una pipeline ETL con **PySpark** y **Delta Lake**

Capítulo 4 — Machine Learning y MLOps con Databricks

1. Introducción a **MLflow** y su integración con Databricks
2. Gestión del ciclo de vida del modelo con **MLflow Tracking**
3. Uso de **AutoML** en Databricks para entrenar modelos sin código
4. Implementación de modelos en producción con **Model Serving**
5. Optimización de modelos con **Hyperopt** y **MLflow**
6. Integración con librerías de IA como **Hugging Face** y **TensorFlow**
7. **Ejercicio:** Entrenar, registrar y desplegar un modelo con **MLflow**

Capítulo 5 — Inteligencia Artificial Generativa y LLMs en Databricks

1. Integración de Databricks con **Modelos de Lenguaje (LLMs)**
2. Uso de **Databricks Mosaic AI** para entrenamiento de modelos
3. Optimización de modelos de IA con **GPUs y clústeres distribuidos**
4. Creación de aplicaciones con modelos preentrenados
5. Fine-tuning de modelos con **Hugging Face** en Databricks
6. Inferencia en tiempo real con **Model Serving**
7. **Ejercicio:** Implementar un chatbot con un modelo de **IA generativa** en **Databricks**

Capítulo 6 — Visualización de Datos y Dashboards en Databricks

1. Uso de **Databricks SQL** para consultas interactivas
2. Creación de **Dashboards** y Visualizaciones en Databricks
3. Integración con **Power BI, Tableau y Looker**
4. Uso de **Python** y **Matplotlib** para visualización avanzada
5. Creación de informes dinámicos con **Plotly** y **Seaborn**
6. Optimización de queries en **Databricks SQL** para análisis en tiempo real
7. Ejercicio: Construcción de un **Dashboard interactivo** en Databricks SQL

Capítulo 7 — Automatización, Orquestación en Databricks

1. Introducción a **Databricks Workflows** y **Jobs**
2. Creación de **pipelines de datos automatizados**
3. Uso de **Delta Live Tables** para ETL en tiempo real
4. Programación de tareas con **Apache Airflow** en **Databricks**
5. Uso de **Genie** para ejecución de consultas y workflows optimizados
6. Monitoreo y troubleshooting de tareas en Databricks
7. **Ejercicio:** Implementar una pipeline ETL automatizada con **Delta Live Tables**

Capítulo 8 — Seguridad, Gobernanza y Optimización de Costos en Databricks

1. Principios de **seguridad en Databricks:** autenticación y control de accesos
2. Administración de permisos y roles con **Unity Catalog**
3. Auditoría y monitoreo de logs en Databricks
4. Estrategias de **optimización de costos** en **Databricks**
5. Uso de **Databricks SQL Warehouses** para control de costes
6. Mejores prácticas de **Data Governance** en entornos multi-nube
7. **Ejercicio:** Implementar controles de seguridad y monitoreo en un workspace Databricks

Microsoft Fabric

Gold Partner | stratebi Analytics and Big Data

Capítulo 1 — Introducción a Microsoft Fabric

Objetivo: Comprender los fundamentos de Microsoft Fabric y su diferencia con la arquitectura anterior de Microsoft.

1. **¿Qué es Microsoft Fabric?**
   - Plataforma unificada de analítica de datos en la nube.
   - Integración de servicios como Power BI, Synapse, Data Factory, etc.
2. **Componentes principales de Microsoft Fabric**
   - OneLake, Data Engineering, Data Science, Data Warehouse, Real-Time Analytics, Power BI, etc.
3. **Beneficios clave de una plataforma unificada**
   - Reducción de costos, integración sin fricciones, seguridad centralizada.
4. **Comparación con la arquitectura anterior de Microsoft**
   - Diferencias con Azure Synapse, Data Lake, Power BI separado, etc.
5. **Casos de uso en empresas y sectores clave**
   - Finanzas, retail, manufactura, salud, etc.
6. **Primeros pasos: Creación de un entorno de trabajo en Fabric**
   - Configuración en el portal de Microsoft Fabric.
   - Creación de un espacio de trabajo.
7. **Ejercicio práctico:**
   - Crear un espacio de trabajo en Microsoft Fabric y explorar su interfaz.

Capítulo 2 — OneLake y la unificación del almacenamiento

Objetivo: Entender cómo OneLake actúa como repositorio central de datos.

1. **¿Qué es OneLake?**
   - Un lago de datos basado en Delta Lake.
2. **Estructura jerárquica y organización de datos**
   - Archivos organizados en carpetas, áreas de trabajo y dominios.
3. **Accesos directos: Integración con otras fuentes de datos**
   - Conectando Amazon S3, Google Cloud Storage, y Azure Data Lake.
4. **Comparación con Azure Data Lake Storage Gen2**
   - Ventajas de OneLake: indexación optimizada, formato Delta por defecto.
5. **Buenas prácticas en la gestión de datos en OneLake**
   - Uso de particionamiento y Z-Ordering para eficiencia.
6. **Monitorización y seguridad en OneLake**
   - Implementación de Microsoft Purview para gobernanza.
7. **Ejercicio práctico:**
   - Subir datos a OneLake y acceder a ellos desde PowerBI

Capítulo 3 — Motores de procesamiento en Microsoft Fabric

Objetivo: Entender cómo funciona la computación distribuida en Fabric.

1. **Motores disponibles en Microsoft Fabric**
   - Apache Spark, SQL Engine, KQL Engine.
2. **Configuración y gestión de entornos de procesamiento**
   - Crear un entorno Spark, configurar nodos de ejecución.
3. **Optimización del rendimiento**
   - Uso de particionamiento, caching y autoescalado.
4. **Comparación con Synapse Analytics y otros servicios previos**
   - Mejoras en tiempos de ejecución y facilidad de integración.
5. **Integración con Notebooks y Data Pipelines**
   - Ejecución de código en Python, R y SQL.
6. **Manejo de datos con Apache Spark en Fabric**
   - Cargar y transformar datos con Spark DataFrames.
7. **Ejercicio práctico:**
   - Crear un Notebook en Fabric y realizar transformación de datos con PySpark.

Capítulo 4 — Direct Lake vs Import y DirectQuery

Objetivo: Explorar los modos de acceso a datos en Fabric.

1. **Introducción a Direct Lake**
   - Acceso en tiempo real a datos sin necesidad de carga.
2. **Ventajas sobre Import y DirectQuery**
   - Reducción de latencia, mejor rendimiento.
3. **Configuración del modo Direct Lake en Power BI**
   - Creación de un modelo de datos conectado a OneLake.
4. **Limitaciones y consideraciones**
   - Soporte para tipos de datos específicos.
5. **Comparación con arquitecturas anteriores**
   - Evolución desde modelos Import y DirectQuery.
6. **Uso de caché y optimización de consultas**
   - Configuración de almacenamiento en memoria.
7. **Ejercicio práctico:**
   - Crear un informe en Power BI con Direct Lake.

Capítulo 5 — Administración, gobernanza y seguridad en Microsoft Fabric

Objetivo: Implementar controles de seguridad y gobernanza.

1. **Políticas de acceso en Fabric**
   - Uso de Microsoft Purview para clasificación de datos.
2. **Gestión de dominios y subdominios**
   - Organización de espacios de trabajo en entornos empresariales.
3. **Implementación de etiquetas de confidencialidad**
   - Protección de datos sensibles con Purview.
4. **Comparación con modelos de seguridad anteriores**
   - Diferencias con Azure RBAC y modelos tradicionales.
5. **Auditoría y monitoreo de actividades en Fabric**
   - Implementación de alertas y logs.
6. **Respaldo y recuperación de datos en OneLake**
   - Estrategias de disaster recovery.
7. **Ejercicio práctico:**
   - Configurar permisos y auditoría en un espacio de trabajo.

Capítulo 6 — Integración de IA en Microsoft Fabric

Objetivo: Aplicar modelos de IA en Fabric.

1. **Servicios de IA disponibles en Fabric**
   - Modelos pre-entrenados, integración con Azure OpenAI.
2. **Uso de modelos de lenguaje y análisis predictivo**
   - Implementación de modelos ML en Fabric.
3. **Automatización con Notebooks y SparkML**
   - Crear y entrenar modelos con SparkML.lib.
4. **Comparación con Azure Machine Learning**
   - Diferencias clave y casos de uso.
5. **Implementación de pipelines de datos para IA**
   - Extracción, transformación y carga de datos en Fabric.
6. **Visualización de predicciones en Power BI**
   - Conexión de resultados con dashboards interactivos.
7. **Ejercicio práctico:**
   - Entrenar un modelo de predicción con SparkML en Fabric.

Capítulo 7 — Novedades recientes en Microsoft Fabric

Objetivo: Explorar las últimas actualizaciones de Fabric.

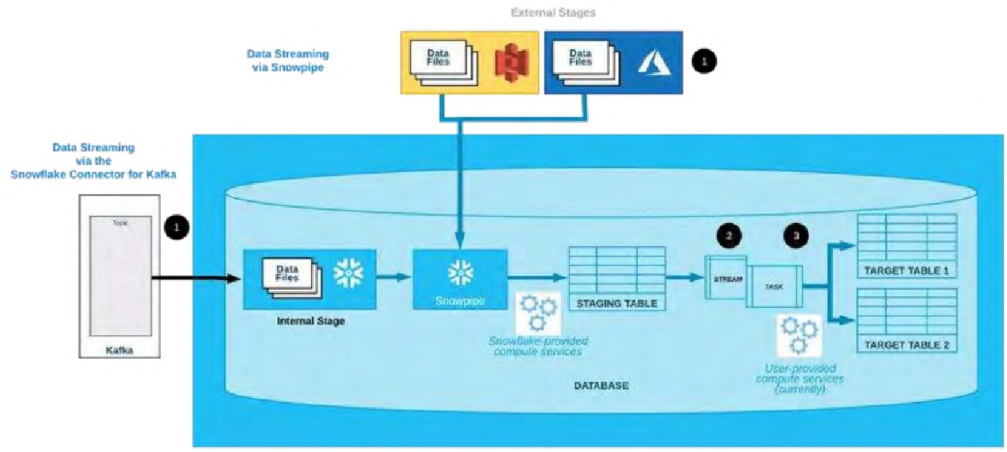
1. **Supervisión del espacio de trabajo (versión preliminar)**
2. **Uso compartido de datos externos en OneLake**
3. **Implementación de la API de GraphQL en Fabric**
4. **Actualizaciones en el catálogo de OneLake**
5. **Disponibilidad general del SDK de Fabric**
6. **Mejoras en la gestión de dominios**
7. **Ejercicio práctico:**
   - Explorar las nuevas funcionalidades en el portal de Fabric.

Capítulo 8 — Proyecto final de implementación

Objetivo: Explorar los modos de acceso a datos en Fabric.

1. **Definición del caso de negocio**
2. **Ingesta de datos en OneLake**
3. **Transformación con Spark y SQL**
4. **Construcción de modelos analíticos**
5. **Creación de dashboards en Power BI**
6. **Implementación de seguridad y gobernanza**
7. **Entrega del proyecto y optimización final**

Curso de Snowflake



Capítulo 1

Introducción a Snowflake y su Arquitectura

Google Cloud · Azure · aws

1. ¿Qué es Snowflake? Historia y evolución del Data Cloud
2. Arquitectura multi-clúster en la nube: separación de almacenamiento y cómputo
3. Diferencias entre Snowflake en **AWS, Azure y Google Cloud**
4. Conceptos clave: Virtual Warehouses, Databases y Schemas
5. Seguridad y gobernanza en Snowflake: roles, permisos y cifrado de datos
6. **Ejercicio:** Crear una cuenta en Snowflake, configurar un **Virtual Warehouse** y cargar datos en una tabla

Capítulo 2

Almacenamiento y Gestión de Datos en Snowflake

snowflake · snowpipe

1. Tipos de almacenamiento en **Snowflake:** Databases, Tables, Stages y External Tables
2. **Snowflake File Formats:** Ingesta de datos desde CSV, Parquet y JSON
3. **Ingesta de datos en streaming** con Snowpipe y Auto-Ingest
4. **Time Travel y Fail-safe:** Recuperación de datos históricos
5. Uso de **Zero-Copy Cloning** y **Data Sharing** para colaboración entre cuentas
6. **Ejercicio:** Crear una tabla, cargar datos desde un Stage, restaurar datos con Time Travel

Capítulo 3

Transformación y Procesamiento de Datos en Snowflake

snowflake · python

1. Uso de **Snowflake SQL** para transformación de datos
2. **Streams & Tasks:** Automatización de pipelines de datos
3. **Materialized Views** y **Clustering** para optimización de consultas
4. **Stored Procedures** y **User-Defined Functions (UDFs)** en SQL y Python
5. Uso de **Dynamic Tables** para mantener datos actualizados sin ETL complejo
6. Ejercicio: Crear un **Stream** y un **Task** para capturar cambios en datos en tiempo real

Capítulo 4

Integración de Snowflake con Herramientas Externas

databricks · Spark · PowerBI · tableau · Looker

1. **Conectores y API:** Snowflake Connector for Python, JDBC y ODBC
2. **Integración con Apache Spark y Databricks**
3. **Uso de Snowflake con herramientas de BI:** Power BI, Tableau y Looker
4. **Snowpark:** Desarrollo de código en Python y Java dentro de Snowflake
5. **Uso de Snowflake con Data Lakes:** External Tables y Iceberg Tables
6. **Ejercicio:** Conectar Snowflake con Python y ejecutar consultas con el Snowflake Connector

Capítulo 5

Machine Learning e Inteligencia Artificial con Snowflake

TensorFlow

1. Introducción a **Snowflake Cortex AI:** Modelos preentrenados y funciones de IA
2. **Uso de Snowpark ML** para entrenar modelos dentro de Snowflake
3. Integración con **Hugging Face y TensorFlow**
4. **Feature Engineering** en Snowflake con funciones avanzadas de SQL y Python
5. **MLOps en Snowflake:** Entrenamiento, versionado y despliegue de modelos
6. **Ejercicio:** Crear y ejecutar un **modelo de Machine Learning** dentro de Snowflake con Snowpark ML

Capítulo 6

Integración de Snowflake con Herramientas Externas

snowflake

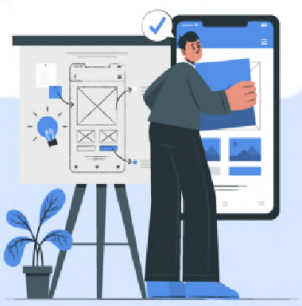
1. **Data Sharing:** Compartición de datos en tiempo real sin duplicaciones
2. **Snowflake Marketplace:** Acceso y publicación de datasets
3. **Data Governance:** Uso de **Object Tagging y Data Classification**
4. **Masking Policies y Row Access Policies:** Protección de datos sensibles
5. **Auditoría y Monitoreo** con Query History y Access History
6. **Ejercicio:** Configurar un **Data Share** entre cuentas de Snowflake y aplicar una **Masking Policy**

Capítulo 7

Optimización del Rendimiento y Costos en Snowflake

snowflake

1. Estrategias de optimización de consultas: **Micro-partitioning y Clustering**
2. **Query Performance Tuning:** Uso de Query Profile para análisis de ejecución
3. **Resource Monitors:** Control de costos y límites de cómputo
4. **Auto-suspend y Auto-scale:** Ajuste automático de Virtual Warehouses
5. **Optimización de almacenamiento** con Time Travel y Retention Policies
6. **Ejercicio:** Analizar y optimizar una consulta con **Query Profile**



PARTNER · snowflake | stratebi Analytics and Big Data

Curso de Databricks

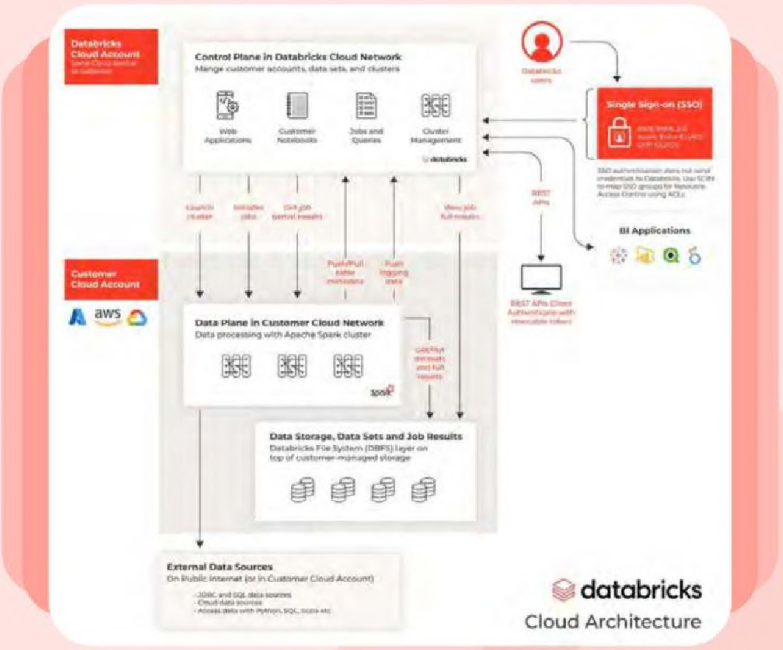
databricks

Capítulo 1

Introducción a Databricks y su Arquitectura

databricks · Spark · Google Cloud · Azure · aws

1. ¿Qué es Databricks? Historia y evolución
2. Arquitectura de Databricks: Lakehouse, Apache Spark y MLflow
3. Integración con AWS, Azure y Google Cloud
4. Seguridad y administración de accesos en Databricks
5. Clústeres en Databricks: configuración, tipos y optimización
6. El motor optimizado de ejecución en Databricks
7. **Ejercicio:** Configurar un entorno Databricks y ejecutar consultas en un Notebook



Capítulo 2

Almacenamiento y Manejo de Datos en Databricks

databricks · DELTA LAKE · Avro · Parquet

1. Almacenamiento en Databricks: DBFS, Delta Lake y external tables
2. Uso de **Auto Loader** para ingesta de datos en streaming
3. Administración de formatos de datos: CSV, Parquet, Delta y Avro
4. Optimización con **Delta Lake:** Compaction, Z-Ordering y Vacuum
5. Integración con Data Warehouses y Data Lakes externos
6. Uso de **Unity Catalog** para gobernanza de datos
7. **Ejercicio:** Crear y optimizar una tabla Delta con **Auto Loader**

Capítulo 3

Transformación de Datos con PySpark y Apache Spark en Databricks

databricks · PySpark · Spark · TUNGSTEN AUTOMATION · photon · DELTA LAKE

1. Introducción a **Apache Spark** en Databricks
2. Uso de **PySpark:** creación de DataFrames y transformación de datos
3. Aplicación de **Spark SQL** en Databricks
4. Optimización de consultas con **Catalyst Optimizer** y **Tungsten**
5. Introducción a **Photon Engine** y **Genie** para procesamiento optimizado
6. Creación de Jobs y Pipelines ETL con Databricks Workflows
7. **Ejercicio:** Construir una pipeline ETL con **PySpark** y **Delta Lake**

Capítulo 4

Machine Learning y MLOps con Databricks

databricks · mlflow · TensorFlow

1. Introducción a **MLflow** y su integración con Databricks
2. Gestión del ciclo de vida del modelo con **MLflow Tracking**
3. Uso de **AutoML** en Databricks para entrenar modelos sin código
4. Implementación de modelos en producción con **Model Serving**
5. Optimización de modelos con **Hyperopt y MLflow**
6. Integración con librerías de IA como **Hugging Face y TensorFlow**
7. **Ejercicio:** Entrenar, registrar y desplegar un modelo con **MLflow**

Capítulo 5

Inteligencia Artificial Generativa y LLMs en Databricks

databricks · mosaicML

1. Integración de Databricks con **Modelos de Lenguaje (LLMs)**
2. Uso de **Databricks Mosaic AI** para entrenamiento de modelos
3. Optimización de modelos de IA con **GPU y clústeres distribuidos**
4. Creación de aplicaciones con modelos preentrenados
5. Fine-tuning de modelos con **Hugging Face** en Databricks
6. Inferencia en tiempo real con **Model Serving**
7. **Ejercicio:** Implementar un chatbot con un modelo de **IA generativa** en **Databricks**

Capítulo 6

Visualización de Datos y Dashboards en Databricks

databricks · PowerBI · tableau · Looker · matplotlib · plotly · seaborn

1. Uso de **Databricks SQL** para consultas interactivas
2. Creación de **Dashboards** y Visualizaciones en Databricks
3. Integración con **Power BI, Tableau y Looker**
4. Uso de **Python** y **Matplotlib** para visualización avanzada
5. Creación de informes dinámicos con **Plotly** y **Seaborn**
6. Optimización de queries en **Databricks SQL** para análisis en tiempo real
7. Ejercicio: Construcción de un **Dashboard interactivo** en Databricks SQL

Capítulo 7

Automatización, Orquestación en Databricks

databricks · Airflow · DELTA LAKE

1. Introducción a **Databricks Workflows** y **Jobs**
2. Creación de **pipelines de datos automatizados**
3. Uso de **Delta Live Tables** para ETL en tiempo real
4. Programación de tareas con **Apache Airflow** en **Databricks**
5. Uso de **Genie** para ejecución de consultas y workflows optimizados
6. Monitoreo y troubleshooting de tareas en Databricks
7. **Ejercicio:** Implementar una pipeline ETL automatizada con **Delta Live Tables**

Capítulo 8

Seguridad, Gobernanza y Optimización de Costos en Databricks

databricks

1. Principios de **seguridad en Databricks:** autenticación y control de accesos
2. Administración de permisos y roles con **Unity Catalog**
3. Auditoría y monitoreo de logs en Databricks
4. Estrategias de **optimización de costos** en **Databricks**
5. Uso de **Databricks SQL Warehouses** para control de costos
6. Mejores prácticas de **Data Governance** en entornos multi-nube
7. **Ejercicio:** Implementar controles de seguridad y monitoreo en un workspace Databricks

Curso de Microsoft Fabric

Gold Partner

stratebi Analytics and Big Data | Microsoft

1. Introducción a Microsoft Fabric

Objetivo: Comprender los fundamentos de Microsoft Fabric y su diferencia con la arquitectura anterior de Microsoft.

Microsoft Fabric · PowerBI · Azure Synapse Analytics · Azure Data Factory

1. **¿Qué es Microsoft Fabric?**
   - Plataforma unificada de analítica de datos en la nube.
   - Integración de servicios como Power BI, Synapse, Data Factory, etc.
2. **Componentes principales de Microsoft Fabric**
   - OneLake, Data Engineering, Data Science, Data Warehouse, Real-Time Analytics, Power BI, etc.
3. **Beneficios clave de una plataforma unificada**
   - Reducción de costos, integración sin fricciones, seguridad centralizada.
4. **Comparación con la arquitectura anterior de Microsoft**
   - Diferencias con Azure Synapse, Data Lake, Power BI separado, etc.
5. **Casos de uso en empresas y sectores clave**
   - Finanzas, retail, manufactura, salud, etc.
6. **Primeros pasos: Creación de un entorno de trabajo en Fabric**
   - Configuración en el portal de Microsoft Fabric.
   - Creación de un espacio de trabajo.
7. **Ejercicio práctico:**
   - Crear un espacio de trabajo en Microsoft Fabric y explorar su interfaz.

2. OneLake y la unificación del almacenamiento

Objetivo: Entender cómo OneLake actúa como repositorio central de datos.

OneLake · Delta Lake · Amazon S3 · Google Cloud · Azure Data Lake · Purview · PowerBI

1. **¿Qué es OneLake?**
   - Un lago de datos basado en Delta Lake.
2. **Estructura jerárquica y organización de datos**
   - Archivos organizados en carpetas, áreas de trabajo y dominios.
3. **Accesos directos: Integración con otras fuentes de datos**
   - Conectando Amazon S3, Google Cloud Storage, y Azure Data Lake.
4. **Comparación con Azure Data Lake Storage Gen2**
   - Ventajas de OneLake: Indexación optimizada, formato Delta por defecto.
5. **Buenas prácticas en la gestión de datos en OneLake**
   - Uso de particionamiento y Z-Ordering para eficiencia.
6. **Monitorización y seguridad en OneLake**
   - Implementación de Microsoft Purview para gobernanza.
7. **Ejercicio práctico:**
   - Subir datos a OneLake y acceder a ellos desde Power BI.

3. Motores de procesamiento en Microsoft Fabric

Objetivo: Entender cómo funciona la computación distribuida en Fabric.

Microsoft Fabric · Apache Spark · Azure Synapse Analytics · Python

1. **Motores disponibles en Microsoft Fabric**
   - Apache Spark, SQL Engine, KQL Engine.
2. **Configuración y gestión de entornos de procesamiento**
   - Crear un entorno Spark, configurar nodos de ejecución.
3. **Optimización del rendimiento**
   - Uso de particionamiento, caching y autoescalado.
4. **Comparación con Synapse Analytics y otros servicios previos**
   - Mejoras en tiempos de ejecución y facilidad de integración.
5. **Integración con Notebooks y Data Pipelines**
   - Ejecución de código en Python, R y SQL.
6. **Manejo de datos con Apache Spark en Fabric**
   - Cargar y transformar datos con Spark DataFrames.
7. **Ejercicio práctico:**
   - Crear un Notebook en Fabric y realizar transformación de datos con PySpark.

4. Direct Lake vs Import y DirectQuery

Objetivo: Explorar los modos de acceso a datos en Fabric.

Microsoft Fabric · PowerBI

1. **Introducción a Direct Lake**
   - Acceso en tiempo real a datos sin necesidad de carga.
2. **Ventajas sobre Import y DirectQuery**
   - Reducción de latencia, mayor rendimiento.
3. **Configuración del modo Direct Lake en Power BI**
   - Creación de un modelo de datos conectado a OneLake.
4. **Limitaciones y consideraciones**
   - Soporte para tipos de datos específicos.
5. **Comparación con arquitecturas anteriores**
   - Evolución desde modelos Import y DirectQuery.
6. **Uso de caché y optimización de consultas**
   - Configuración de almacenamiento en memoria.
7. **Ejercicio práctico:**
   - Crear un informe en Power BI con Direct Lake.

5. Administración, gobernanza y seguridad en Microsoft Fabric

Objetivo: Implementar controles de seguridad y gobernanza.

Microsoft Fabric · Purview · Azure · OneLake

1. **Políticas de acceso en Fabric**
   - Uso de Microsoft Purview para clasificación de datos.
2. **Gestión de dominios y subdominios**
   - Organización de espacios de trabajo en entornos empresariales.
3. **Implementación de etiquetas de confidencialidad**
   - Protección de datos sensibles con Purview.
4. **Comparación con modelos de seguridad anteriores**
   - Diferencias con Azure RBAC y modelos tradicionales.
5. **Auditoría y monitoreo de actividades en Fabric**
   - Implementación de alertas y logs.
6. **Respaldo y recuperación de datos en OneLake**
   - Estrategias de disaster recovery.
7. **Ejercicio práctico:**
   - Configurar permisos y auditoría en un espacio de trabajo.

6. Integración de IA en Microsoft Fabric

Objetivo: Aplicar modelos de IA en Fabric.

Microsoft Fabric · Spark ML · PowerBI

1. **Servicios de IA disponibles en Fabric**
   - Modelos pre-entrenados, integración con Azure OpenAI.
2. **Uso de modelos de lenguaje y análisis predictivo**
   - Implementación de modelos ML en Fabric.
3. **Automatización con Notebooks y SparkML**
   - Crear y entrenar modelos con Spark MLlib.
4. **Comparación con Azure Machine Learning**
   - Diferencias clave y casos de uso.
5. **Implementación de pipelines de datos para IA**
   - Extracción, transformación y carga de datos en Fabric.
6. **Visualización de predicciones en Power BI**
   - Conexión de resultados con dashboards interactivos.
7. **Ejercicio práctico:**
   - Entrenar un modelo de predicción con SparkML en Fabric.

7. Novedades recientes en Microsoft Fabric

Objetivo: Explorar las últimas actualizaciones de Fabric.

Microsoft Fabric · GraphQL · OneLake

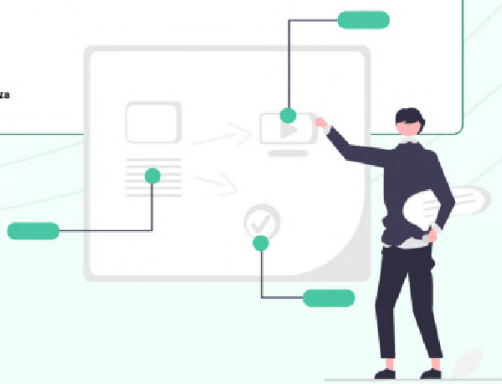
1. **Supervisión del espacio de trabajo (versión preliminar)**
2. **Uso compartido de datos externos en OneLake**
3. **Implementación de la API de GraphQL en Fabric**
4. **Actualizaciones en el catálogo de OneLake**
5. **Disponibilidad general del SDK de Fabric**
6. **Mejoras en la gestión de dominios**
7. **Ejercicio práctico:**
   - Explorar las nuevas funcionalidades en el portal de Fabric.

8. Proyecto final de implementación

Objetivo: Explorar los modos de acceso a datos en Fabric.

OneLake · PowerBI · Spark

1. **Definición del caso de negocio**
2. **Ingesta de datos en OneLake**
3. **Transformación con Spark y SQL**
4. **Construcción de modelos analíticos**
5. **Creación de dashboards en Power BI**
6. **Implementación de seguridad y gobernanza**
7. **Entrega del proyecto y optimización final**



stratebi Analytics and Big Data

Curso de Data Scientist

Avanzado y práctico

Temario del primer bloque

01 pandas python

Introducción a Snowflake y su Arquitectura

- Tipos de datos
- Lectura de ficheros tabulares
- Filtrado y Ordenación
- Funciones para cadenas de texto
- Agrupación y funciones de agregación
- Visualización de gráficas

02 Selenium BeautifulSoup

Webscraping (BeautifulSoup, Selenium)

- Scraping HTML con BeautifulSoup
- Scraping JavaScript con Selenium (Amazon)
- Conexión a APIS (Twitter, Idealista, DGT, ...)
- Lectura de ficheros XML y JSON

03 spaCy

Procesamiento de Lenguaje Natural (NTLK, Spacy)

- Conceptos: Corpus, Bolsa de Palabras, Normalización, Tokenización, Eliminación de Sufijos/Prefijos y palabras vacías y Lematización
- Etiquetado de Palabras (POS Parts-of-Speech)
- Análisis de sentimientos
- Clasificación de textos y detección de temas
- Reconocimiento de Entidades nominales

Temario del segundo bloque

01 learn

Machine Learning con Scikit-Learn

- Tipos de datos
- Lectura de ficheros tabulares
- Filtrado y Ordenación
- Funciones para cadenas de texto
- Agrupación y funciones de agregación
- Visualización de gráficas

02 Keras

Redes Neuronales y Deep Learning con Keras

- Perceptrón Multicapa
- Entrenamiento de Redes
- Tipos de redes
- Deep Learning

03 ultralytics YOLO python

Computer vision

- Procesamiento de imágenes
- Detección de Bordes
- Realidad Aumentada
- Machine Learning y Visión por computador
- Detección de Caras
- Deep Learning y modelos pre-entrenados YOLO y Caffe

04 Azure

Machine Learning en Azure

- Servicios Machine Learning en la nube

Curso de IA y ChatGPT

Avanzado y práctico

Temario | ChatGPT

01 Introducción a la Inteligencia Artificial y ChatGPT

- Definición y conceptos básicos de la Inteligencia Artificial.
- Historia y evolución de la IA.
- Introducción a los sistemas de conversación y el procesamiento del lenguaje natural (NLP).

02 Fundamentos de la Programación y Matemáticas para la IA

- Determinar cuándo es necesario o recomendable Databricks
- Comparación frente al uso de Dataflows (Data Factory o PowerBI), Azure Functions, Azure SQL, etc.
- Presentación y casos de uso reales

03 Machine Learning y Aprendizaje Profundo

- Introducción al Machine Learning.
- Redes neuronales artificiales y su funcionamiento.
- Entrenamiento de modelos de aprendizaje profundo.

04 Procesamiento del Lenguaje Natural (NLP)

- Preprocesamiento de texto.
- Modelos de lenguaje y embeddings de palabras.
- Técnicas avanzadas de NLP como NamedEntityRecognition (NER) y SentimentAnalysis.

05 Modelos de Generación de Lenguaje

- Arquitecturas de modelos generativos.
- Entrenamiento de modelos generativos de lenguaje.
- Evaluación de modelos y métricas de calidad.

06 Desarrollo de Chatbots con ChatGPT

- Introducción a los chatbots y su aplicación en la industria.
- Configuración y puesta en marcha de un chatbot con ChatGPT.
- Mejora y personalización de chatbots para tareas específicas.

07 Ética y Responsabilidad en la Inteligencia Artificial

- Problemas éticos y sesgos en los sistemas de IA.
- Directrices de ética en la IA y regulaciones.
- Estrategias para abordar problemas éticos en el desarrollo de chatbots.

08 Implementación y Despliegue de Sistemas de Conversación

- Integración de chatbots en sitios web y aplicaciones.
- Consideraciones de seguridad en la implementación de chatbots.
- Monitoreo y mantenimiento de chatbots en producción.

09 Casos de Uso y Aplicaciones en la Industria

- Estudios de caso de éxito en la implementación de chatbots.
- Aplicaciones en sectores como atención al cliente, salud, e-commerce, etc.
- Tendencias actuales y futuras en la IA y chatbots.

10 Trabajo Práctico y Proyecto Final

- Desarrollo de un proyecto práctico de chatbot.
- Implementación de funcionalidades avanzadas.
- Presentación y evaluación de proyectos por parte de expertos en IA.

stratebi Analytics and Big Data

Curso de Azure Data Factory

Integra datos a escala empresarial

Temario | Azure | Azure Data Factory

01
Introducción a Azure Data Factory

02
Principales características

- Pipelines, Datasets y Dataflows
- Resumen de conectores a fuentes y destinos de datos
- Integration Runtime y Linked Services

03
Creación de pipelines para el movimiento de datos y la orquestación de procesos usando servicios de Azure

- Movimiento de datos con Copy Data
- Invocación de procesos en Databricks y otros servicios de Azure

04
Monitorización de procesos

05
Modelar

- Overview de actividades
- Soporte para ejecuciones condicionales, uso de variables y parámetros

06
Movimiento de ficheros sobre Azure Data Lake

07
Creación de Dataflows para la transformación de datos

08
Creación de Dataflows para la transformación de datos

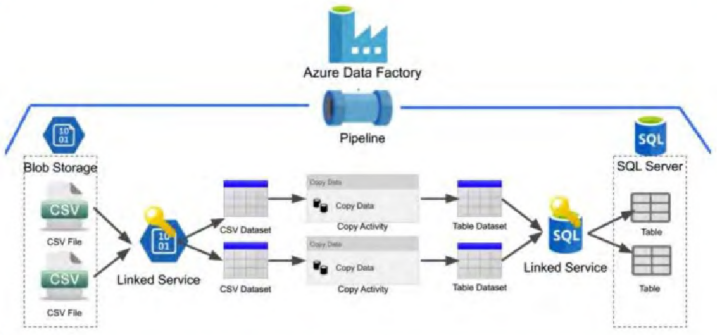
09
Conceptos clave para la automatización y desarrollo

- Creación de triggers
- Integración con repositorios GIT y procesos CI/CD entre múltiples entornos

10
Presentación y estudio de arquitecturas usadas en proyectos reales utilizando Azure Data Factory

11
Test de nivel

Final
Ejercicios y ejemplos



Curso de Pentaho Open BI

Domina la integración, modelado y análisis de datos con herramientas líderes en Business Intelligence.

Temario

pentaho

Módulo 01

Fundamentos de Business Intelligence y Data Warehousing

1. Introducción a BI: importancia y beneficios en la toma de decisiones.
2. Evolución del BI: de los sistemas DSS y EIS a las soluciones modernas.
3. Procesos clave en BI: análisis, integración, transformación y visualización.
4. Conceptos fundamentales de Data Warehousing (DWH) y su aplicación.
5. Diferencias entre Data Warehouse (DW) y Data Marts.
6. Modelos de datos en BI: Star Schema, Snowflake y Starflake.
7. Estrategias de implementación de DW: enfoque Top-Down vs. Bottom-Up.
8. Herramientas Open Source en BI: ventajas del software libre en análisis de datos.

Módulo 02

Integración y Transformación de Datos con Pentaho PDI

1. Introducción a Pentaho Data Integration (PDI): funciones y características.
2. Instalación y configuración de PDI: herramientas clave (Spoon, Kitchen, Pan, Carte).
3. Creación de procesos ETL: extracción, transformación y carga de datos.
4. Conexión y manejo de bases de datos con JDBC en Pentaho.
5. Modelado y transformación de datos: uso de repositorios y metadatos.
6. Ejercicios prácticos: carga de un Data Warehouse con dimensiones y hechos.
7. Configuración de variables en Spoon para entornos dinámicos.
8. Estrategias de depuración y optimización de flujos ETL en PDI.

Módulo 03

Modelado y Análisis de Datos con Pentaho Schema Workbench (PSW)

1. Introducción a Mondrian y modelado OLAP en Pentaho.
2. Capas de Mondrian: presentación, dimensional, estrella y almacenamiento.
3. Diseño y creación de cubos OLAP en Pentaho Schema Workbench (PSW).
4. Definición de dimensiones, jerarquías y niveles para análisis de datos.
5. Creación y optimización de medidas y métricas de negocio.
6. Consultas avanzadas con MDX (Multidimensional Expressions). Uso de STPivot
7. Ejercicios prácticos de modelado de cubos y consultas multidimensionales.
8. Integración de modelos OLAP con herramientas de reporting y dashboards.

Módulo 04

Administración y Gestión de Bases de Datos para BI

1. Introducción a MySQL como base de datos para Business Intelligence.
2. Instalación y configuración de MySQL Workbench.
3. Gestión y optimización de bases de datos relacionales en BI.
4. Uso de JDBC para la conexión entre Pentaho y MySQL.
5. Administración de consultas SQL en entornos analíticos.
6. Seguridad y buenas prácticas en la gestión de bases de datos.
7. Integración de MySQL con herramientas de procesamiento de datos.
8. Ejercicios prácticos sobre modelado y consulta de datos en MySQL.

Módulo 05

Creación de Dashboards y Reportes con Pentaho

1. Instalación y configuración de Pentaho Business Analytics.
2. Uso de Pentaho User Console (PUC) para gestión de reportes.
3. Creación de informes dinámicos y análisis en tiempo real.
4. Uso de herramientas avanzadas como STReport y STDashboard.
5. Personalización de dashboards con gráficos y visualizaciones interactivas.
6. Diseño de reportes OLAP con análisis Drill-down y Pivot.
7. Publicación de informes y distribución en entornos empresariales.
8. Optimización del rendimiento y estrategias de visualización de datos.

Módulo 06

Optimización y Pre-Agrupación de Datos en BI

1. Introducción a Pentaho Aggregation Designer
2. Creación y gestión de agregaciones de datos para mejorar el rendimiento.
3. Modelado de bases de datos multidimensionales optimizadas.
4. Generación de estructuras preagregadas para consultas rápidas.
5. Publicación y actualización de esquemas optimizados.
6. Configuración avanzada de fuentes de datos en Pentaho BI Server.
7. Técnicas de mejora en tiempos de respuesta de reportes OLAP.
8. Casos prácticos de optimización en entornos empresariales.

Módulo 07

Proyecto Final y Aplicaciones Prácticas en BI

1. Definición del caso de estudio: aplicación real en BI.
2. Diseño de la arquitectura del proyecto: integración de herramientas y datos.
3. Implementación del proceso ETL con Pentaho PDI.
4. Modelado y análisis de datos con Pentaho Schema Workbench.
5. Creación de dashboards y reportes con Pentaho y LinceBI.
6. Optimización y escalabilidad del sistema BI.
7. Presentación del proyecto final con validación de resultados.
8. Conclusiones, mejores prácticas y certificación.

Incluye

Material de estudio, ejercicios prácticos, acceso a herramientas y certificación.

stratebi
Analytics and Big Data

Curso de Azure Devops

Conjunto de servicios de desarrollo

Temario | Azure DevOps

01

Introducción a Devops

02

Creación de Pipelines de Release para despliegue de infraestructura

- Pasos o plantillas disponibles en Devops
- Plantillas ARM
- Pasos con Azure CLI/ Power Shell e introducción a la sintaxis
- Parametrización y creación de etapas por entorno
- Demos de despliegue de infraestructura en 1 y 2 entornos
- Data Factory + Data Bricks + Blob Storage

03

Repositorios

- Creación de repositorios GIT en Azure Devops
- Asociación de repositorios a componentes de infraestructura
- Demostración con Data Factory
- Introducción al uso de GIT en Azure Devops para el desarrollo colaborativo
- Ejercicio con Data Factory
- Demo con Databricks

04

Integración y despliegue continuo

- Creación de pipelines de compilación (Build)
- Creación de pipelines de despliegue
- Demo de compilación y despliegue de código Data Factory / Databricks

stratebi Analytics and Big Data

Curso de Microsoft Synapse

Data integration, enterprise data warehousing y big data analytics

Temario | Azure Synapse

01
Introducción a Business Intelligence y Data Warehouse

02
Introducción a procesos de Ingesta de datos, ETLs...

03
Modelado de Datos

04
Introducción a Azure Synapse

- Principales Características
- ETL
- Notebooks
- Integración con ADLS Gen 2
- SQL pools
- Casos de uso
- Principales Características ETL con Synapse
- Pipelines, Datasets y Dataflows
- Resumen de conectores a fuentes y destinos de datos
- Integration Runtime y Linked Services
- Ejercicio toma de contacto con Synapse y primer pipeline
- Descripción de las principales Activities
- Creación de Dataflows para la transformación de datos
- Conceptos clave para la automatización y desarrollo
- Creación de triggers
- Integración con repositorios GIT y procesos CICD entre múltiples entonos (DEV/PRE/PRO)
- Integration Runtime y Linked Services
- Monitorización de pipelines
- Ejercicio end to end ETL con Synapse
- Diferencias con Azure Data Factory

05
Apache Spark en Synapse (Introducción)

- Spark pools
- Notebooks en Spark
  - ¿Qué es un notebook?
  - Lenguajes soportados y tipos de celdas
  - Ejercicio introductorio: "Hola notebooks!"

Apache Spark | Azure Synapse

06
Introducción a Synapse SQL

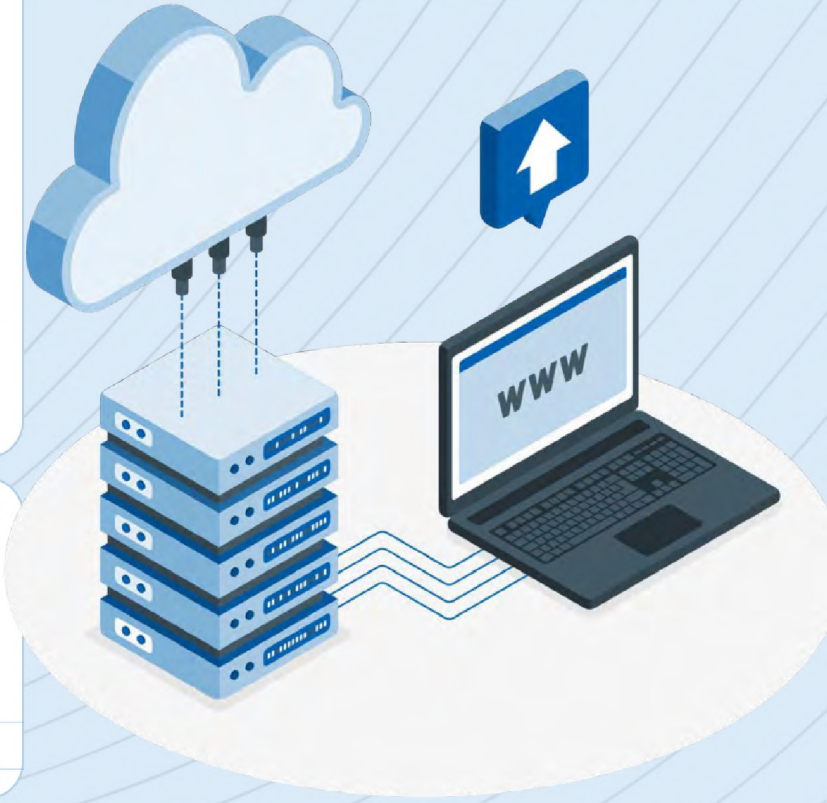
07
Integración con Power BI

- Problemas éticos y sesgos en los sistemas de IA.
- Directrices de ética en la IA y regulaciones.
- Estrategias para abordar problemas éticos en el desarrollo de chatbots.

08
Integración con otros servicios de Azure

- Azure Machine Learning
- Azure Purview

Azure | PowerBI



Curso de Cloud Analytics

Tecnologías de análisis en la nube

Temario (Parte 1)

01 Introducción

Explorar las opciones de computación y almacenamiento para cargas de trabajo de ingeniería de datos
- Introducción a Azure Synapse Analytics
- Describir Azure Databricks
- Introducción al almacenamiento de Azure Data Lake • Describir la arquitectura de Delta Lake
- Trabajar con flujos de datos mediante Azure Stream Analytics
- Combinar la transmisión y el procesamiento por lotes con una única pipeline
- Organizar el data lake en niveles de transformación de archivos
- Indexación de almacenamiento de data lake para la aceleración de consultas y cargas de trabajo

Diseño e implementación de la capa de servicio
- Diseñar un esquema multidimensional para optimizar las cargas de trabajo analíticas
- Transformación sin código a escala con Azure Data Factory
- Rellenar dimensiones que cambian lentamente en las pipelines de Azure Synapse Analytics
- Diseñar un esquema en estrella para cargas de trabajo analíticas
- Rellenar dimensiones que cambian lentamente con Azure Data Factory y mapeo de flujos de datos

Consideraciones de ingeniería de datos para archivos fuente
- Diseñar un almacén de datos moderno con Azure Synapse Analytics
- Proteger un almacén de datos en Azure Synapse Analytics
- Administrar archivos en un data lake de Azure
- Protección de archivos almacenados en un data lake de Azure

02 Data Engineering

Ejecutar consultas interactivas con grupos de SQL sin servidor de Azure Synapse Analytics
- Explorar las capacidades de los grupos SQL sin servidor de Azure Synapse
- Consultar datos en el lake mediante grupos de SQL sin servidor de Azure Synapse
- Crear objetos de metadatos en grupos SQL sin servidor de Azure Synapse
- Proteger los datos y administrar a los usuarios en los grupos SQL sin servidor de Azure Synapse
- Consultar datos de Parquet con grupos SQL sin servidor
- Crear tablas externas para archivos Parquet y CSV • Crear vistas con grupos de SQL sin servidor
- Acceso seguro a los datos en un data lake cuando se utilizan grupos de SQL sin servidor
- Configurar la seguridad del data lake mediante el control de acceso basado en roles (RBAC) y la lista de control de acceso

Explorar, transformar y cargar datos en el almacén de datos usando Apache Spark
- Comprender la ingeniería de big data con Apache Spark en Azure Synapse Analytics
- Ingestar datos con los cuadernos de Apache Spark en Azure Synapse Analytics
- Transformar datos con DataFrames en Apache Spark Pools en Azure Synapse Analytics
- Integrar grupos de SQL y Apache Spark en Azure Synapse Analytics
- Realizar exploración de datos en Synapse Studio Ingestar datos con cuadernos Spark en Azure Synapse Analytics
- Transformar datos con DataFrames en grupos de Spark en Azure Synapse Analytics
- Integrar grupos de SQL y Spark en Azure Synapse Analytics

Exploración y transformación de datos en Azure Databricks
- Describir Azure Databricks
- Leer y escribir datos en Azure Databricks
- Trabajar con DataFrames en Azure Databricks
- Trabajar con métodos avanzados de DataFrames en Azure Databricks ç
- Usar DataFrames en Azure Databricks para explorar y filtrar datos
- Almacenar en caché un DataFrame para consultas posteriores más rápidas
- Eliminar datos duplicados
- Manipular valores de fecha / hora
- Eliminar y cambiar el nombre de las columnas DataFrame
- Agregar datos almacenados en un DataFrame

Ingesta y carga datos en el almacén de datos
- Utilizar las mejores prácticas de carga de datos en Azure Synapse Analytics
- Ingestión a escala de petabytes con Azure Data Factory
- Realizar la ingestión a escala de petabytes con Azure Synapse Pipelines
- Importar datos con PolyBase y COPY usando T-SQL
- Utilizar las mejores prácticas de carga de datos en Azure Synapse Analytics

Transformar datos con Azure Data Factory o Azure Synapse Pipelines
- Integración de datos con Azure Data Factory o Azure Synapse Pipelines
- Transformación sin código a escala con Azure Data Factory o Azure Synapse Pipelines
- Ejecutar transformaciones sin código a escala con Azure Synapse Pipelines
- Crear un pipeline de datos para importar archivos CSV con formato deficiente
- Crear flujos de datos de mapeo

03 Cómo compilar un Almacén de Datos

Soporte del procesamiento analítico transaccional híbrido (HTAP) con Azure Synapse Link
- Diseñar procesamiento transaccional y analítico híbrido con Azure Synapse Analytics
- Configurar Azure Synapse Link con Azure Cosmos DB • Consultar Azure Cosmos DB con grupos de Apache Spark
- Consultar Azure Cosmos DB con grupos de SQL sin servidor
- Configurar Azure Synapse Link con Azure Cosmos DB
- Consultar Azure Cosmos DB con Apache Spark para Synapse Analytics
- Consultar Azure Cosmos DB con un grupo de SQL sin servidor para Azure Synapse Analytics

Seguridad de un extremo a otro con Azure Synapse Analytics
- Proteger un almacén de datos en Azure Synapse Analytics
- Configurar y administrar secretos en Azure Key Vault Implementar controles de cumplimiento para datos confidenciales
- Asegurar la infraestructura de soporte de Azure Synapse Analytics
- Asegurar el área de trabajo de Azure Synapse Analytics y los servicios administrados
- Proteger los datos del área de trabajo de Azure Synapse Analytics

Analizar y optimizar el almacenamiento del data wharehouse
- Analizar y optimizar el almacenamiento del data wharehouse de datos en Azure Synapse Analytics
- Comprobar si hay datos sesgados y uso de espacio
- Comprender los detalles de almacenamiento de la tienda de columnas
- Estudiar el impacto de las vistas materializadas
- Explorar las reglas para operaciones mínimamente registradas

Procesamiento de transmisión en tiempo real con Stream Analytics
- Habilitar la mensajería confiable para aplicaciones de Big Data con Azure Event Hubs
- Trabajar con flujos de datos mediante Azure Stream Analytics
- Ingesta flujos de datos con Azure Stream Analytics
- Utilizar Stream Analytics para procesar datos en tiempo real de Event Hubs
- Utilizar las funciones de ventana de Stream Analytics para crear agregados y resultados en Synapse Analytics
- Escalar el trabajo de Azure Stream Analytics para aumentar el rendimiento mediante la partición
- Repartir la entrada de flujo para optimizar la paralelización

Crear una solución de procesamiento de transmisión con Event Hubs y Azure Databricks
- Procesar datos de streaming con transmisión estructurada de Azure Databricks
- Explorar las características y usos clave de la transmisión estructurada
- Transmitir datos desde un archivo y escribirlos en un sistema de archivos distribuido
- Utilizar ventanas deslizantes para agregar fragmentos de datos en lugar de todos los datos
- Aplicar marcas de agua para eliminar datos obsoletos
- Conectarse a transmisiones de lectura y escritura de Event Hubs

Transformar datos con Azure Data Factory o Azure Synapse Pipelines
- Integración de datos con Azure Data Factory o Azure Synapse Pipelines
- Transformación sin código a escala con Azure Data Factory o Azure Synapse Pipelines
- Ejecutar transformaciones sin código a escala con Azure Synapse Pipelines
- Crear un pipeline de datos para importar archivos CSV con formato deficiente
- Crear flujos de datos de mapeo

Generar informes mediante la integración de Power BI con Azure Synpase Analytics
- Crear informes con Power BI utilizando su integración con Azure Synapse Analytics
- Integrar un área de trabajo de Azure Synapse y Power BI
- Optimizar la integración con Power BI
- Mejorar el rendimiento de las consultas con vistas materializadas y almacenamiento en caché de conjuntos de resultados
- Visualizar datos con SQL sin servidor y crear un informe de Power BI

Realizar procesos integrados de aprendizaje automático en Azure Synapse Analytics
- Utilizar el proceso de aprendizaje automático integrado en Azure Synapse Analytics
- Crear un servicio vinculado de Azure Machine Learning
- Activar un experimento de Auto ML con datos de una tabla Spark
- Enriquecer los datos utilizando modelos entrenados
- Ofrecer resultados de predicción con Power BI

Curso de Dashboards & Scorecards

Integra datos a escala empresarial

Temario

pentaho

Primer Bloque

Introducción y arquitectura

- El cuadro de mando en Business Analytics
- CDE en la Arquitectura Pentaho
- Diseño de Cuadros de mando con Ctools
- Arquitectura Ctools
- Arquitectura CDF, CDE, CDA
- Instalación y configuración
- Ejercicios

Segundo Bloque

Datasources - Orígenes de datos

- Fuentes de Datos
  - Introducción a CDA
  - Introducción a KTR
  - Introducción a WAQR
  - Introducción a OLAP Mondrian
- CDA with SQL
- CDA with MDX
- CDA with MQL
- CDA with Xaction Result Set
- Ejercicios

Tercer Bloque

Look & Feel I

- Estructura general de Layout en Pentaho CDE
- Frameworks: Introducción
  - Framework Blueprint
  - Framework Bootstrap
- Layout Mobile
- Conceptos básicos: HTML
- Conceptos básicos: Javascript
- Conceptos básicos: jQuery Conceptos básicos: Css
- Ejercicios

Cuarto Bloque

Visualización I

- Gráficos CCC
- Selectores
- Parametrización y Dependencias

stratebi
Analytics and Big Data

Curso de Cloud Analytics

Tecnologías de análisis en la nube

Temario (Parte 2) | databricks | Azure Synapse Analytics | Azure Stream Analytics | Azure Data Factory | Azure Synapse | amazon EMR | aws | snowflake

04

Cloud Computing

Fundamentos

- Qué es cloud computing.
- Diferentes tipos de cloud computing.
- Modelos básicos en la nube.
- Componentes de la nube.
- Hardware Cloud.
- Virtualización.
- Cloud storage.
- Grid Computing.
- Computing transaccional.
- Software Cloud.
- SaaS.
- Disponibilidad On-Demand.
- Pago por uso.
- SOA y la nube.
- Modelos de nubes.
- Seguridad, Auditoria y Cumplimiento en la Nube.
- Plataformas varias.

Amazon Elastic MapReduce - EMR

- Información general acerca de los grandes datos y Apache Hadoop.
- Soluciones AWS en un ecosistema Bigdata.
- Beneficios de Amazon EMR.
- Arquitectura de Amazon EMR. Utilización de Amazon EMR.
- Inicio y utilización y configuracion de un clúster de Amazon EMR.
- Marcos de programación de alto nivel de Apache Hadoop.
- Marcos de Programacion para Amazon EMR: Hive.
- Pig Streaming.
- Utilización de Hive para análisis promocionales.
- Hue para Anazon EMR.
- Analisis integrados en memoria con SPARK en Amazon EMR.

Amazon Warehouse - Amazon RedShift

- Opciones de almacenamiento de datos de AWS.
- Utilización de DynamoDB con Amazon EMR.
- Información general acerca de Amazon Redshift y los grandes datos.
- Utilización de Amazon Redshift para grandes datos Visualización y orquestación de grandes datos.
- Utilización de Tableau Desktop o de la inteligencia empresarial de Jaspersoft para visualizar grandes datos.
- Recursos y componentes de Amazon Redshift.
- Lanzamiento de un clúster de Amazon Redshift.
- Revisión de las estrategias de almacenamiento de datos.
- Identificación de requisitos y orígenes de datos.
- Diseño de almacén de datos.
- Carga de datos en el almacén de datos.
- Escritura de consultas y ajuste de rendimiento.
- Mantenimiento del almacén de datos.
- Análisis y visualización de datos.

Real time streaming data - Amazon Kinesis

- Transmision Datos en tiempo real.
- Kinesis Data Analytics.
- Amazon Kinesis Data Firehos; registrar, transformar y cargar transmisiones de datos en almacenes de datos de AWS para realizar análisis en tiempo real con herramientas de inteligencia empresarial existentes.
- Amazon Kinesis Data Streams: Crear aplicaciones personalizadas en tiempo real que procesen transmisiones de datos con marcos de procesamiento de transmisiones conocidos.
- Amazon Kinesis Video Streams: Transmisión segura de videos desde dispositivos conectados a AWS para análisis, aprendizaje automático y otros procesos.
- Amazon DynamoDB, Amazon Quicksight, Amazon Athena.

05

Snowflake

- SnowPro™ Core Certification Overview
- Snowflake Overview and Architecture
- Snowflake Virtual Warehouses
- Snowflake Storage and Protection
- Data Movement (Loading and Unloading)
- Snowflake Account and Security
- Snowflake Performance and Tuning

Curso de Web Scraping

Curso: Integración de fuentes externas y web scraping - Extracción de datos desde la web

Temario | SCRAPIAS | BeautifulSoup

01

Introducción al Web Scraping

- ¿Qué es el web scraping?
- Datos de un sitio web
- Estructura
- Selectores HTML/XPath
- Selectores XML

02

Scraping HTML

- BeautifulSoup
- Introducción y manejo de excepciones
- Tipos de objetos en BeautifulSoup
- Funciones find y find_all
- Recorrido de Árboles
- Expresiones Regulares
- Expresiones Lambda

03

Scraping JavaScript

- Selenium
- Instalar y configurar Chromedriver
- Selectores
- Espera de elementos HTML
- Ejecución de JavaScript
- Funciones del WebDriver
- Navegación
- Acciones
- Ejercicios (DGT, Portales públicos de Ayuda e incentivos)

04

Consumo de APIs

- APIs con Requests
- Peticiones GET, POST, PUT, DELETE
- Autorización y cabecera
- Parseo de JSON
- API de Twitter y consumo servicios Open Data

05

Manejo y lectura de JSON y XML

- JSON
  - Serializar JSON
  - Deserializar JSON
  - Deserialización de estructuras complejas
  - Introducción a librerías jsonpickle y json
  - Conversión entre objetos Python y JSON (diccionarios, listas, tuplas,)
- XML
  - Introducción a XML etree
  - Parseo de ficheros XML
  - Funciones de lectura (find, findall, findtext,iter)

06

Creación de caso de uso práctico accediendo a fuentes externas

07

SCRAPIAS

- SAP
- Salesforce
- Other ERPs. CRMs
- 3rd apps
- APIs
- Web Scraping
- Google Analytics
- Open Data
- IoT Data
- Conectores

stratebi Analytics and Big Data

talend | # Curso de Introducción a Talend

Curso: Integración de fuentes externas y web scraping - Extracción de datos desde la web

Temario talend

01
Introducción a Talend Open Studio

- Sistemas Transaccional vs DW
- Estructura de una solución BI
- ¿Qué es una ETL?
- Flujo de un proyecto ETL
- Talend Open Studio

02
Trabajando con Talend Open Studio

- Secciones TOS
- Instalación del Entorno
- Gestión de Proyectos

03
Componentes Entrada/Salida

- Creación de Metadatos
- Creación de Trabajos
- Transformación entre ficheros
- Conversión de modelos de datos
- Enriquecer los datos con búsquedas
- Validación de datos

04
Trabajo con bases de datos relacionales

- Metadatos de base de datos
- Lectura de base de datos
- Escribiendo en base de datos
- Sincronizar Bases de datos
- Modificación de datos
- Búsquedas Dinámicas
- Visualizar Queries

05
Gestión de ficheros

- Manipulación de ficheros
- Operaciones con ficheros
- Uso de ficheros temporales
- Almacenamiento de datos intermedios en memoria
- Operaciones FTP

06
Contextos y variables globales

- Contextos y variables globales
- Tipos de Variables
- Tipos de Contextos

07
Transformaciones de datos

- Filtrado de datos Ordenaciones
- Sumas y agregaciones
- Normalización
- Extracciones en ficheros delimitados
- Búsquedas y reemplazos
- Muestreo de filas

08
Orquestación de trabajos

- Subjobs
- Definición de flujos lógicos
- División de trabajos en Subjobs
- Iteraciones y bucles
- Separar y fusionar flujos
- Flujos entre Jobs

09
Funcionalidades de desarrollo

- Manejo básico de errores
- Estadísticas y logs
- Versionado de trabajos

stratebi Analytics and Big Data

Curso de Microsoft Power Automate

Descubre la manera más sencilla de automatizar tus procesos digitales

Temario | Power Automate | Microsoft Flow

01

Introducción a Power Automate

- Conceptos básicos (Microsoft Flow)
- Idioma, interfaz
- Propiedades del flujo y lanzadores de Triggers

02

Conceptos de Power Automate

- Introducción al Módulo
- Credenciales en cajas de acciones
- Renombrar flujo, caja de acción y comentarios
- Flow checker y test
- Expresiones y variables
- Definición de caso práctico

03

Trabajando con Power Automate

- ~~Creando Formas y Lista~~
- ~~Configuración y testeo de flujo~~
- ~~Flujos programables~~
- ~~Ejercicios~~
  - ~~Conversión de archivos~~
  - ~~Tareas de aprobación~~
- ~~Revisión de plantillas y tareas~~
- ~~Establecer notificaciones~~
- ~~Creando aplicaciones de Power Apps usando Power Automate~~
- ~~Uso de plantillas~~
- ~~Desarrollo de caso práctico~~
- ~~Buenas prácticas~~

stratebi
Analytics and Big Data

Curso de Microsoft Power Apps

Construye aplicaciones profesionales de forma sencilla

Temario | Power Apps | PowerBI | Microsoft Teams

01

Empezando con Power Apps

- Entorno en Microsoft 365
- ¿Qué son las Power Apps?
- Elementos que incluyen
- Compilando aplicaciones
- Buscando y creando Apps
- Ejemplos de Apps

02

Trabajando con Power Apps

- La plataforma de desarrollo
- Automatizar Procesos
- Conectar con datos
- Integración con sistemas
- Creación de App en base a plantillas de Power App
- Personalizaciones de Apps
- Elementos
  - Elementos insertables
  - Botones
  - Formas
  - Controles
  - Gráficos
  - Embeber en PowerBI
  - Control con cámara
- Licencias
- Ejemplo y Casos de uso

03

Ejecución de Power Apps

- Tipos de Apps:
  - Canvas Apps
  - Model Driven Apps
  - Template Apps
- Ejecución de una aplicación en explorador web
- Ejecución de una aplicación en dispositivo móvil
- Ejecución de una aplicación controlada por modelos en dispositivo móvil
- Incorporación y ejecución de una aplicación en Microsoft Teams
- Buscar y ejecutar una aplicación en AppSource
- Buenas prácticas con Power Apps

odoo | Curso de Odoo CRM/ERP

Extraer, manipular y visualizar información comercial

Temario odoo

01
Introducción al concepto CRM/ERP

02
Odoo Community Edition

03
Instalación

- Preparación del sistema
- BBDD PostgreSQL
- Proxy inverso nginx
- Seguridad

04
Administración

- Usuarios y grupos
- Aplicaciones
- Reglas y dominios
- Interfaz

05
CRM/VENTAS

- Producción y servicios
- Presupuestos y pedidos, firma online
- Iniciativas y oportunidades
- Tableros

06
Facturación/contabilidad

- Localización
- Facturación
- Pagos manuales y pagos online
- Diarios
- Apuntes contables

07
Compras

- Solicitudes de presupuestos, pedidos de compra
- Compras automáticas
- Pagos
- Workflows de validación de compras

08
Web y eCommerce

- Diseñador web
- Catálogo online
- Formas de pago
- Seguimiento de pedidos y formas de envío

09
Inventario

- Almacenes
- Ubicaciones y Rutas
- Inventario, lotes y números de serie
- Operaciones

10
Odoo Analytics

- Modelos de datos OLAP de Odoo Analytics
- Modelos de Reporting de Odoo Analytics
- Indicadores por módulos y cubos virtuales
- Permisos de acceso a elementos
- Grupos de seguridad y compatibilidad con mondrian Sttools:
- STPivot
- Vistas OLAP por defecto
- Nuevas vistas OLAP a medida
- STReport
- Reports por defecto
- Nuevos Reports a medida
- Exportación y programación de envío por email
- STDashboard
- Dashboards por defecto
- Nuevos Dashboards a medida

OLAP odoo

Curso de Python

Visualiza tus decisiones

Temario | python | Visual Studio

Bloque 1
Temario

- ¿Qué es Python? Conceptos básicos.
- Formas de ejecutar un programa en Python
- Variables, tipos de datos y conversiones.
- Operadores: Aritméticos, asignación, lógicos y otros.
- Cadenas de texto (funciones de cadenas y formatos de impresión)
- Entornos de desarrollo: Visual Studio
  - Creación de entornos virtuales.
  - Instalar módulos usando el comando "pip"
- Colecciones de datos (listas, tuplas, conjuntos y diccionarios)

Bloque 1
Ejercicios

- Configuración de entornos de desarrollo
- Ejercicio 1: Ejercicios sobre conceptos básicos: variables, operadores, cadenas, colecciones...

Bloque 2
Temario

- Estructuras de control: If/Else, For, While, Range, Break and Continue...
- Funciones
- Funciones Lambda
- Clases, Objetos y herencia de clases,...
- Invocación de programas (argumentos, configuración,...)
- Tratamiento de errores/excepciones

Bloque 2
Ejercicios

- Ejercicio 2: Creación de un programa completo con Python.
- Ejercicio 3. Generador de contraseñas

Bloque 3
Temario

- Función Map
- Uso de Math
- Funciones de fecha y hora, expresiones regulares
- Entrada/Salida: Manejo de ficheros: Texto, Binarios, XML y JSON

Bloque 3
Ejercicios

- Ejercicio 4: Gestión de archivos por lotes
- Ejercicio 5: Funciones Map y Filter

Bloque 4
Temario

- Request (peticiones a API Rest)
- Lectura/Escritura de datos desde base de datos (PostgreSQL)
- Introducción al tratamiento y análisis de datos con Pandas
- Otras librerías: Numpy

Bloque 4
Ejercicios

- Ejercicio 6: Integración vía API.
- Ejercicio 7: Tratamiento de logs con Pandas Hacer en proyecto de código

Curso de Qlik Sense

Informes y dashboards complejos

Temario

Qlik

01 Introducción a Qlik Sense

- Contexto actual de las plataformas de Business Intelligence
- Plataformas dentro del ecosistemas Qlik
- Fases de un proyecto de Implantación en Qlik Sense

02 Tipos de licencias de acceso

- Qlik Sense Business
- Qlik Sense Enterprise

03 Acceso a la plataforma Qlik Sense SAAS

- Qlik Sense SAAS
- Qlik Sense Windows

04 Navegación por la interfaz del modelo de igualdad en Qlik Sense

- Centro de control
- Vista general
- Opciones de navegación desde una hoja de análisis
- Sheet view
- Área de selección de variables para el análisis de datos
- Menú contextual de los gráficos interactivos
- Descripciones de la hoja de análisis

05 Principales tipos de gráficos

- KPIs
- Gráficos de tarta
- Gráficos de barras
- Gráfico combinado
- Diagrama de dispersión
- Tabla dinámica
- Gráfico de líneas
- Diagrama de distribución
- Diagrama de dispersión geográfica
- Diagrama Mekko
- Otros tipos de gráficos

06 Storytelling: gestión de presentaciones

- Generación de presentaciones
- Repositorio de screenshots
- Acceso a presentaciones live

07 Elementos Maestros y eventos

- Master items
- Sheet actions

08 Arquitectura Qlik Sense

- Data Manager
- Sheet View

09 Gestión de la carga de información

- Carga por fichero
- Carga de datos
- Sección Insights
- Generación de conexiones con Base de Datos

08 Secciones principales de las hojas de análisis

- Botones de navegación
- Logos
- Área de filtros
- KPIs
- Dashboards de análisis

09 Tips and tricks

- Test data generation
- Scripting
- Dual navigation

10 Casos de uso principales

- Casos de uso Ventas
- Casos de uso RRHH
- Casos de uso Finanzas

11 Caso práctico

- Definición estructural de las hojas de análisis
- Dashboard principal
- Dashboards secundarios
- Capacidades de navegación
- Introducción de eventos
- Generación de gráficos de seguimiento y control

stratebi
Analytics and Big Data

Curso de Pentaho Data Integration

ETL: Extrae, transforma y carga tus datos

Temario

pentaho

Aprende a extraer, transformar y cargar datos con una de las herramientas ETL más potentes del mercado, optimizando flujos de datos y automatizando procesos empresariales

Incluye: Material de estudio, ejercicios prácticos, acceso a infraestructura cloud y certificación oficial.

Módulo 1

Fundamentos de Pentaho Data Integration (PDI)

1. Introducción a los procesos ETL y su importancia en Business Intelligence.
2. ¿Qué es Pentaho Data Integration (PDI) y cómo se usa?
3. Instalación y configuración de PDI en distintos entornos.
4. Herramientas clave de PDI: Spoon, Pan, Kitchen y Carte.
5. Diferencias entre Transformations y Jobs y cuándo usarlas.
6. Configuración de variables de entorno y gestión de parámetros.
7. Uso de repositorios en PDI: archivos locales vs. bases de datos.
8. Conexión con bases de datos mediante JDBC y JNDI.
9. Ejercicio práctico: creación de una transformación básica en Spoon.

Módulo 2

Extracción de Datos desde Diferentes Fuentes

1. Tipos de fuentes de datos en procesos ETL: estructurados y no estructurados.
2. Importación de datos desde archivos CSV, Excel, JSON y XML.
3. Conexión con bases de datos relacionales (MySQL, PostgreSQL, SQL Server).
4. Integración con APIs REST y servicios web para la extracción de datos.
5. Uso de Pentaho Data Services para consultas dinámicas.
6. Extracción de datos en tiempo real desde Kafka y sistemas en streaming.
7. Conexión con sistemas NoSQL (MongoDB, Cassandra, Hadoop, Spark).
8. Estrategias de optimización en la extracción de datos.
9. Ejercicio práctico: integración de múltiples fuentes de datos en un solo flujo.

Módulo 3

Transformación y Limpieza de Datos

1. Normalización y estandarización de datos en PDI.
2. Aplicación de reglas de negocio en transformaciones de datos.
3. Manejo de datos faltantes, duplicados y valores nulos.
4. Uso de los pasos de transformación en Spoon: Filter, Join, Split, Replace.
5. Validación de datos mediante Expresiones Regulares (RegEx).
6. Transformaciones avanzadas con JavaScript y fórmulas.
7. Implementación de cálculos y agregaciones en PDI.
8. Auditoría de datos y estrategias de control de calidad.
9. Ejercicio práctico: limpieza y transformación de un dataset con errores.

Módulo 4

Carga y Almacenamiento de Datos

1. Tipos de almacenamiento de datos: relacional, NoSQL y Data Warehousing.
2. Inserción, actualización y eliminación de datos en bases de datos SQL.
3. Uso de Bulk Loading para optimizar cargas masivas de datos.
4. Estrategias de particionamiento y almacenamiento eficiente.
5. Implementación de Slowly Changing Dimensions (SCD) en PDI.
6. Exportación de datos a diferentes formatos: CSV, JSON, XML, Parquet.
7. Integración con Data Lakes y arquitecturas modernas.
8. Optimización del rendimiento en cargas ETL a gran escala.
9. Ejercicio práctico: carga de datos optimizada en una base de datos.

Módulo 5

Automatización y Orquestación de Procesos ETL

1. Uso de Jobs en PDI para la ejecución de flujos de datos.
2. Planificación y calendarización de procesos con Pentaho Scheduler.
3. Integración con herramientas de automatización: Airflow, Jenkins, Crontab.
4. Configuración de alertas y monitoreo de errores en ETL.
5. Control de flujo con condiciones, loops y variables dinámicas.
6. Procesamiento en paralelo para mejorar el rendimiento de ETL.
7. Integración de Pentaho con Apache Kafka y procesamiento en streaming.
8. Orquestación de flujos en entornos cloud y distribuidos.
9. Ejercicio práctico: creación de un flujo ETL automatizado con Jobs.

Módulo 6

Seguridad, Administración y Buenas Prácticas

1. Administración de usuarios y permisos en Pentaho BI Server.
2. Configuración de entornos de desarrollo, prueba y producción.
3. Implementación de logs y auditoría en procesos ETL.
4. Control de versiones en Git para flujos ETL.
5. Estrategias de recuperación ante errores y fallos en ETL.
6. Mantenimiento y optimización de procesos ETL en producción.
7. Integración de PDI con entornos Cloud (AWS, Azure, GCP).
8. Mejores prácticas en el desarrollo de procesos ETL escalables.
9. Ejercicio práctico: implementación de seguridad y logs en un flujo ETL.

Módulo 7

Proyecto Final y Aplicaciones Reales

1. Definición del caso de estudio: aplicación real en ETL.
2. Diseño de la arquitectura del proyecto: herramientas y modelos.
3. Implementación de procesos de extracción con Pentaho PDI.
4. Transformación y limpieza de datos según requerimientos de negocio.
5. Carga de datos optimizada en bases de datos SQL y Data Warehouses.
6. Automatización y monitoreo del flujo ETL.
7. Optimización y escalabilidad del sistema ETL.
8. Presentación del proyecto final y validación de resultados.
9. Conclusiones, mejores prácticas y certificación del curso.

stratebi
Analytics and Big Data

talend | Curso de Talend avanzado

Suite para data integration, data management y data governance

Temario talend

01
Monitorización Avanzada

- Creación de estructuras
- Monitorización de inserciones, actualizaciones, borrados, rechazo del flujo
- Obtención mensajes generados
- Seguimientos de updates frente a un benchmark
- Assertive Condition para la ejecución de un Job
- Medición tiempo de ejecución de un subjob

02
Usando Java en Talend

- Inclusión de código Java en Jobs
- Creación de rutinas de código
- Importación librerías Java

03
Data Integration & Data Quality

- Introducción a Data Quality
- Data Quality & BI
- Open Source & Data Quality

04
Talend Data Preparation

- Instalación del entorno
- Conceptos principales
- Operaciones sobre datos
- Definición de datasets y preparaciones
- Exportación de resultados

05
Talend Data Quality

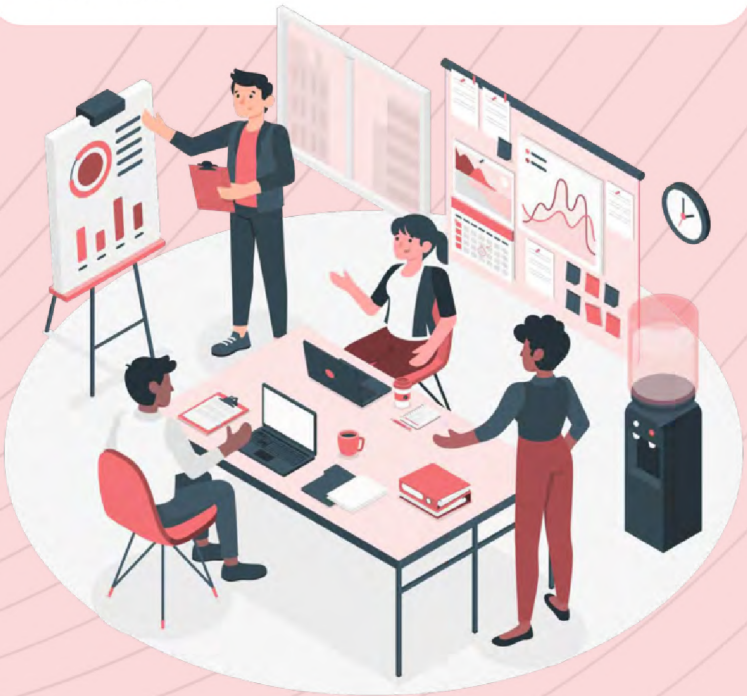
- Data Quality & Data Profiling
- Metodología para Profiling
- Métodos analíticos usados en DP
- Repositorio Data Profiling
- Instalación del entorno y características
- Análisis de BBDD, esquemas y tablas
- Correlación, redundancia y dependencia funcional de variables
- Patrones, análisis combinado y definición de reglas
- Visualización Gráfica

07
Talend MDM

- Qué es Master Data Management (MDM)
- Governance (Gestión)
- Subject Area (Area de Negocio)
- Master Data (Datos Maestros)
- Arquitectura de MDM
- MDM Governance (Gestión)
- Data Quality (Calidad de Datos) y MDM
- Roles y Responsabilidades de MDM
- Tecnología MDM
- Acciones de MDM
- MDM Open Studio

08
Talend ESB

- Introducción
- Instalación del entorno Componentes ESB
- Xpath
- Tratamiento de XML
- SOAP Web Service
- Rest Service
- Apache Active MQ



Curso de Sports Analytics

Procesamiento de datos deportivos

Temario

python

01
Introducción al Sports Analytics

- Casos de uso
- Ejemplos reales

02
Introducción al lenguaje de programación Python *

- Conceptos básicos del lenguaje: sintaxis, arrays, listas, diccionarios, matplotlib, etc.
- Ejercicios prácticos

03
Obtención del dato

- Proveedores de datos: Wyscout, Instat, OPTA, Statsbomb, mediacoach, etc.
- Eventing y tracking
- Páginas web de referencia: Sofascore, basketball-reference, fbref, etc.
- Webscraping
- Otros: GPS, herramientas de video análisis, iOT, etc.

04
Almacenamiento del dato

- Introducción al SQL
- Bases de datos relacionales
- Bases de datos no relacionales
- Big Data
- Ejercicios prácticos

05
Procesos ETL

- Introducción al ETL
- Procesos ETL con PDI y Talend
- Modelado: modelo en estrella
- Datawarehousing
- Ejercicios prácticos

06
Explotación del dato con Microsoft Power BI

- Ejercicios prácticos

PowerBI

07
Analítica Avanzada

- Análisis avanzado con Python
- IA
- Machine Learning
- Computer visión
- Ejercicios prácticos

08
Ejercicio final

stratebi
Analytics and Big Data

VERTICA | Curso de Vertica Database

Base de datos analítica

Temario VERTICA

01
Architecture Overview

- Vertica Analytics Platform
- Projections
- Query Execution
- Transactions and Locking
- Hybrid Data Store
- Lab Exercice

02
Projection Design

- Projection Fundamentals
- Projection and Table Properties
- Database Designer
- Manual Projection Design
- Projection Management
- Lab Exercice

03
Advanced Projection Design

- Manual Projection Design
- Join Operations
- Group By Operations

04
Logical Design & Security

- Schemas
- Tables and Views
- Constraints
- Users
- Roles
- External Procedures
- Lab Exercice

05
Resource Management

- Resource Manager
- Resource Pool Parameters
- Resource Pools
- User Defined Resource Pools
- Monitoring Resource Pools
- Lab Exercice

06
Resource Management II

- Backup and Restore
- Copy Vertica Database
- Online Recovery

07
Loading Data

- Load Data via INSERT, COPY and MERCE
- Flex Tables
- Connectors and User Defined Loads
- Post Loading Tasks
- Lab Exercises

stratebi Analytics and Big Data

Curso de Introducción al BI

Visualiza tus decisiones

Temario

01

- ¿Qué es la Inteligencia de Negocio? Conceptos Básicos. ¿Dónde y cómo se almacenan los datos?
- ¿Qué tipos de análisis puedo realizar?
- ¿Cómo me puedo beneficiar del uso del BI en mi trabajo diario?
- Convertir 'datos' en 'información'
- Eliminar información 'redundante'
- Areas y departamentos de mejora.
- ¿Qué es un Data Warehouse?
- ¿Qué es un Cuadro de Mando?
- ¿Qué es un Scorecard?
- Valoración de la información de mi empresa
- Optimización del ROI
- Cuales son los Indicadores claves de un negocio.
- Principales Soluciones Business Intelligence
- ¿Cómo obtener ventaja competitiva?
- Casos y ejemplos prácticos

Diccionario de Arquitecturas de Datos

Data Centric
Data Visualization
databricks
Dark Data
Lakehouse
Data Lake
Data Governance
talend
amazon REDSHIFT
Data Mashup
kafka
Cloud Analytics
Azure
Inmon Architecture
Data Pipelines
Data Silos
Data Sources
Data Hub
Data Warehouse
snowflake
Data Vault
APACHE Spark
Business Intelligence
Data Mesh
Data Frame
Data Mart
Data Fabric
Kimball Architecture

Curso de Experto Data engineer

Analytics en funcionamiento

Temario

01 Introducción

- Por qué Scala
- Por qué Spark
- Por qué Scala y Spark
- Scala y Spark dentro de entorno Hadoop: Importancia e Integración
  - Sistema de archivos distribuidos HDFS
  - Motor de procesamiento Map Reduce
  - Gestor de procesos: YARN
  - Sistema de mensajería distribuido para Big Data: Apache Kafka
  - Log streaming: Apache Flume
- Ejemplo de arquitecturas Big Data que usen Scala y Spark

Caso teórico/práctico: Diseño de una arquitectura para la detección de fraude en seguros en tiempo real.

02 Introducción a Scala

- Scala y la necesidad de paralelizar todo *"Single-core performance is running out of steam, and you need to parallelize everything"* (Martin Odersky, creador de Scala)
- Conceptos básicos de Scala
- Tipos de datos
- Estructuras de control
- Conclusiones
- Ejercicios prácticos:
- Tipos de datos, Colecciones y Estructuras de control en Scala

Caso práctico de procesamiento de datos con Scala: (Limpieza, filtrado, agregación)

03 Spark

- Introducción
- ¿Debo usar Scala, Python o Java para programar en Spark? Scala en Spark
- Introducción al Shell de Spark para Scala
- Concepto y creación del Spark Context (SC).
- Los conjuntos elásticos de datos distribuidos (RDDs).
- Operaciones sobre RDDs: Transformaciones y Acciones.
- Programación de funciones para RDD's
- Caché y persistencia de RDD's
- Trabajo con pares Clave-Valor (Key-Value Pair RDD)
- Carga y almacenamiento desde HDFS (Sistema de archivos distribuido de Hadoop)

03

- Ejercicios prácticos:
  - Sobre cada uno de los puntos anteriores
  - Caso práctico
  - Caso práctico* de procesamiento de datos con Scala: (Limpieza, filtrado, agregación).
- Otros lenguajes en Spark (Introducción):
  - SQL
  - R
- Ejecución en clúster Hadoop con programas Spark.
- Ejercicio práctico:
  - Ejecución en el clúster de Axa del programa del caso práctico desarrollado en el ejercicio anterior. (K, iii).

Caso práctico de procesamiento de datos con Scala: (Limpieza, filtrado, agregación)

04 Big Data Streaming 1: Apache Kafka

- Introducción a Apache Kafka
- Arquitectura
- Topics
- Productores y Consumidores.
- Kafka y Apache Zookeper
- Flujo de datos en Kafka
- Ejercicio práctico:
  - Estudio de una implementación de Kafka para lectura de datos desde Wikipedia.

05 Big Data Streaming 2: Spark Streaming

- Arquitectura y abstracción
- Transformaciones y Operaciones Streaming
- Fuentes de entrada
- Tolerancia a fallos
- Rendimiento
- Ejercicio práctico:
  - Diseño e implementación de un programa en Spark Streaming para el procesamiento de los datos de Wikipedia en Streaming usando la implementación proporcionada.

Caso práctico de procesamiento de datos con Scala: (Limpieza, filtrado, agregación)

Curso de Introducción al Big Data

Soluciones BI Open source StrateBI

Temario

01
Introducción al Big Data

- Directrices principales en las que se basa el Big Data.
- Visión histórica e introducción al contexto del Big Data a través de ejemplos intuitivos.
- Cómo afecta Big Data a los negocios.
- La relación entre Big Data, Business Intelligence & Data Science

02
Arquitecturas Big Data

- Introducción y clasificación a las diferentes arquitecturas y sistemas Big Data disponibles en el mercado
- Estudio de profundidad del entorno **Hadoop:** HDFS, Map Reduce, YARN, análisis de la pila de herramientas disponibles sobre HDFS y Map Reduce (Hive, Pig...), introducción a las distribuciones de Hadoop, etc.
- Estudio de las principales soluciones **NoSQL: Cassandra, MongoDB, ...**
- Introducción a las **bases de datos analíticas: HPVertica y MonetDB**
- Consideraciones para la elección de una arquitectura Big Data
- Ejemplos prácticos y visión de futuro sobre estas bases de datos:
  - Instalación de una **distribución de Hadoop** de un solo nodo para la realización de pruebas
  - Introducción a la gestión de un **clúster Hadoop**
  - Introducción al uso del sistema de archivos **HDFS**

03
Obtención y movimiento de datos

- Estudio de los principales tipos de fuentes de datos actuales
  - **Datos estructurados, semi estructurados y no estructurados**
  - **Batch y streaming**
- Análisis de las principales herramientas disponibles para la adquisición y movimiento de datos:
  - **Pentaho Data Integration: Carga, transformación y extracción de datos de cualquier naturaleza** desde fuentes de datos hacía HDFS y viceversa.
  - **Sqoop:** Carga y extracción de **datos relacionales** (SGBDR->HDFS, HDFS->SGBDR) en batch.
  - **Flume:** Carga y transformación de datos en **tiempo real**
- **Ejercicios** con las herramientas anteriores basados en un caso de estudio para la obtención de datos de logs, redes sociales...

04
Procesamiento del Big Data

- Análisis de los requerimientos temporales del análisis **(oportunidad del análisis)**
- Introducción a las principales herramientas para el procesamiento y análisis del Big Data
  - Herramientas sobre MapReduce: **Pig, Hive**
  - Herramientas que no usan Map Reduce: **Spark, Spark Streaming, Storm...**
- Ejercicio basado en un caso de estudio para el procesamiento de datos de logs, redes sociales...

05
Casos de Estudio

- Análisis de casos de estudio del mercado: Sistema de recomendación de Amazon, análisis de datos de sensores en empresas de transporte, análisis de clics en páginas web...
- Análisis de casos de estudios basados en nuestra amplia experiencia en el desarrollo de proyectos Big Data

stratebi Analytics and Big Data

Curso de Machine Learning

Analytics en funcionamiento

Temario

01

Introducción al Machine Learning

- Técnicas
  - Clasificación
  - Regresión
- Colesterina
- Preprocesamiento y Reducción dimensional
- Selección de atributos
- Evaluación del rendimiento
  - Matrices de confusión
  - Principales KPIs R2, MAE, MSE

02

Regresión (Predicción de valores continuos)

- Principales algoritmos
  - Ordinal Least Squares
  - Ridge Regression
  - Laso Regression
  - Elastic Net
- Ejemplos

03

Clasificación (Identificación de la categoría a la que pertenece un objeto)

- Principales algoritmos
  - Logistic Regression
  - Support Vector Machines
  - KNearest Neighbors
  - Decision Trees
- Random Forest
- Multi-layer Perceptron
- Ejemplos

04

Clustering (Agrupación de objetos similares en conjuntos)

- Principales Algoritmos
  - KMeans
  - Spectral Clustering
  - DBSCAN
- Ejemplos

Curso de **Data** Google Cloud Platform

1 | Fundamentos de la Plataforma de Datos en GCP

Arquitectura de referencia de una plataforma de datos moderna

Desafíos de arquitectura
Escalabilidad y elasticidad
Georredundancia y continuidad de servicio
Rendimiento y optimización de costos (finops)
Seguridad y aislamiento entre entornos

Organización de folders, projects y regiones en GCP

2 | Almacenamiento de Datos

Estructuración de zonas bronze, silver y gold

Cloud Storage
Buckets, clases de almacenamiento y control de versiones
Buenas prácticas de seguridad, organización y particionamiento

BigQuery
Conceptos: dataset, table (internal/external), partitioning y clustering
Carga de datos desde GCS
Optimización de consultas y control de costos (slots)

3 | Ingesta y Streaming de Datos

Pub/Sub
Creación de topics y subscriptions
Patrones de publicación y suscripción
Uso de Pub/Sub para telemetría e integración IoT

Dataflow
Procesamiento streaming y batch
Integración con Pub/Sub, BigQuery y Cloud Storage
Python vs Java

Datastream
CDC: Replicación de datos desde bases transaccionales
Configuración de streams y destinos

4 | Procesamiento y Transformación de Datos

BigQuery (como motor de transformación SQL)
Creación de vistas, materialized views y tablas derivadas
Partitioned y clustered tables
SQL avanzado para limpieza y modelado

Dataform
Orquestación de transformaciones SQL dentro de BigQuery
Control de dependencias y versionado de modelos
Integración con repositorios Git

5 | Gestión y Versionado de Artefactos

Artifact Registry
Almacenamiento de componentes ETL, scripts y modelos
Versionado y despliegue controlado
Integración con Dataflow, Cloud Run o Composer

6 | Análisis y Visualización

Modelado dimensional (Hecho y dimensiones)

Looker / Looker Studio
Conexión directa con BigQuery
Creación de dashboards y modelos semánticos

7 | Machine Learning e Inteligencia Artificial

Vertex AI
Entrenamiento y despliegue de modelos

Looker (preview)

BigQuery ML
Creación de modelos predictivos directamente desde SQL
Evaluación, predicción y monitoreo de resultados

Agent Space

8 | Gobernanza, Seguridad y Observabilidad

Cloud Monitoring & Logging
Supervisión de pipelines y recursos GCP
Alertas, métricas y paneles de observación

Dataplex
Gobernanza de datos a escala: políticas, metadatos y linaje
Gestión de zonas de datos y catálogos unificados

Gobierno de BigQuery
Gestión de permisos y roles por dataset y tabla
Data masking
Etiquetado, clasificación y políticas de retención

Curso de Data

1 | Fundamentos del Data Governance (basado en DAMA-DMBOK)

Objetivo: Obtener y afianzar los principios básicos de un sistema de Data Governance

1. Principios y objetivos del Data Governance según DAMA.
2. Dominios del DMBOK: Data Quality, Metadata, Security, Architecture, Stewardship.
3. Roles y responsabilidades: Owner, Steward, Custodian, Consumer. Políticas y marcos de gobierno de datos.
4. Modelo organizativo y comité de gobierno.
5. Mapeo entre gobernanza y madurez organizacional (DMM).
6. Indicadores de éxito y métricas de gobernanza (KDGIs).

Ejercicio Diseñar un Data Governance Operating Model para un ayuntamiento o empresa ficticia.

2 | Estrategia y alineación con el negocio

Objetivo: Entender y enfocar correctamente como sacar provecho en tu negocio

1. Relación entre Data Governance y Data Strategy.
2. Priorización de dominios de datos según valor de negocio.
3. Marco de decisión: top-down vs. bottom-up governance.
4. Identificación de Critical Data Elements (CDEs) y Key Data Assets.
5. Creación de Business Glossaries y definición de Data Domains.
6. Enfoque ROI: medir beneficios tangibles de la gobernanza.
7. Alineación con ESG, compliance, y gobierno corporativo.

Ejercicio Mapear los procesos de negocio de una organización y asociar sus CDEs principales.

3 | Gestión técnica de metadatos y linaje

Objetivo: Comprender sobre el correco flujo de uso de metadatos

1. Tipos de metadatos: técnicos, de negocio y operacionales.
2. Herramientas de catalogación: Microsoft Purview, DataHub, Collibra, Alation.
3. Automatización del linaje: extracción desde ETL/ELT, Power BI, Databricks, etc.
4. Estandarización con OpenMetadata y Egeria.
5. APIs y conectores para ingesta automática de metadatos.
6. Integración con arquitecturas Data Lakehouse (OneLake, Delta Lake, Unity Catalog).
7. Publicación y gobierno colaborativo de metadatos en portales internos.

Ejercicio Implementar un catálogo con DataHub y configurar linaje entre datasets de Snowflake y Power BI.

4 | Data Quality Management

Objetivo: Comprender sobre el correco flujo de uso de metadatos

1. Definición DAMA de calidad de datos: precisión, completitud, consistencia, validez y puntualidad.
2. KPIs y reglas de calidad: diseño y automatización.
3. Validación y profiling de datos (Great Expectations, Deequ, Soda).
4. Procesos de remediación y control de errores.
5. Integración de calidad en pipelines ETL/ELT (Azure Data Factory, dbt, Databricks).
6. Monitoreo continuo de calidad con alertas automáticas.
7. Dashboards de Data Quality en Power BI o Grafana.

Ejercicio Implementar un catálogo con DataHub y configurar linaje entre datasets de Snowflake y Power BI.

5 | Seguridad, privacidad y cumplimiento normativo

Objetivo: Mantener segura tu infraestructura de datos

1. Data Governance y Data Security: diferencias y sinergias.
2. Control de acceso basado en roles y atributos (RBAC, ABAC).
3. Catalogación sensible y clasificación automática en Microsoft Purview.
4. Cifrado, anonimización y pseudonimización de datos.
5. Cumplimiento con GDPR, LOPDGDD y ISO/IEC 38505.
6. Auditoría y trazabilidad de accesos y cambios.
7. Integración de gobernanza con Identity Providers (Azure AD, Okta).

Ejercicio Configurar políticas de sensibilidad y clasificación automática con Purview.

6 | Implementación tecnológica del gobierno de datos

Objetivo: Conseguir que tu gobierno de datos tenga sinergias con tus herramientas

1. Modelos de referencia híbridos: Purview + DataHub + Unity Catalog.
2. Diseño de arquitectura de gobierno: Metadata Hub & Federated Catalogs.
3. Integración con Data Fabric / Data Mesh.
4. Orquestación de políticas y flujos con Airflow, Synapse, Databricks Workflows.
5. APIs de gobernanza y automatización (Purview REST API, DataHub GraphQL).
6. Versionado de metadatos y automatización DevOps.
7. Comparativa: Purview vs. Collibra vs. Unity Catalog.

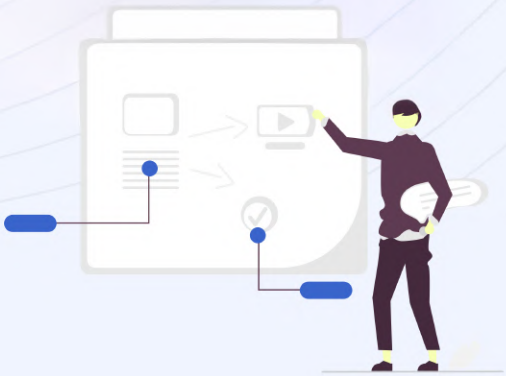
Ejercicio Desplegar un metadata lake unificado con Purview y Unity Catalog.

7 | Gobierno operativo, evolución y casos reales

Objetivo: Afianza conocimientos y lleva a cabo un proyecto completo

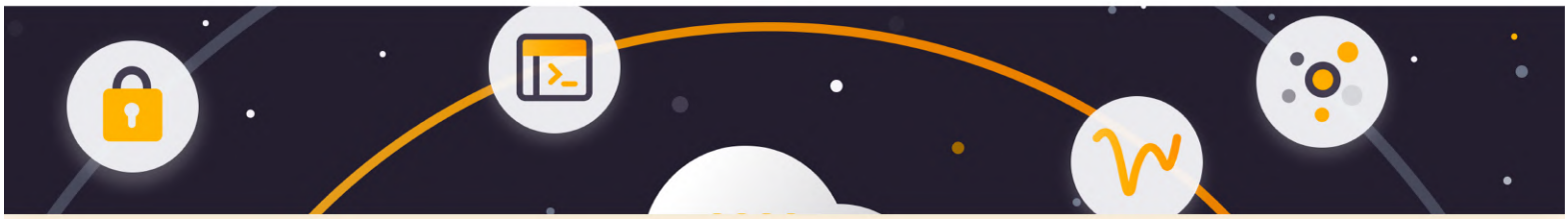
1. Ciclo de vida del Data Governance Program.
2. Evaluación de madurez con el Data Management Maturity Model (DMM).
3. Data Stewardship operativo y herramientas de colaboración.
4. Mecanismos de escalabilidad: Data Domains y Ownership descentralizado.
5. Monitoreo de KPIs y reportes de cumplimiento.
6. Nuevas tendencias: Data Products, Active Metadata, AI-assisted Governance.
7. Casos de uso: banca, administración pública, salud, retail.

Ejercicio Construir un caso práctico de Data Governance integrando Purview (catálogo), Unity Catalog (control de acceso) y DataHub (colaboración).



stratebi
Analytics and Big Data

Curso de ClickHouse



1. Introducción y arquitectura interna de ClickHouse

Objetivo: Conocer los distintos tipos de bases de datos y sus fundamentos para aplicarlos a una BD ClickHouse

1. Conceptos fundamentales de bases de datos columnar vs. row-based.
2. Arquitectura distribuida de ClickHouse: nodos, shards y réplicas.
3. El motor de almacenamiento MergeTree y sus variantes.
4. Procesamiento vectorizado y compresión de datos.
5. Modelo de concurrencia y ejecución de consultas paralelas.
6. Análisis de rendimiento I/O y uso eficiente del CPU.
7. Configuración inicial del servidor (config.xml, users.xml).

Ejercicio Desplegar un nodo ClickHouse y ejecutar consultas de benchmarking con clickhouse-benchmark.

2. Lenguaje SQL en ClickHouse y funciones avanzadas

Objetivo: Explorar en profundidad el uso de SQL en ClickHouse

1. Extensiones SQL específicas de ClickHouse.
2. Funciones de agregación, ventanas y arrays.
3. Tablas externas y funciones remote(), cluster().
4. Subconsultas distribuidas y WITH CTEs.
5. Funciones de ordenamiento, hashing y bloom filters.
6. Manipulación de tipos complejos: JSON, Map, Tuple.
7. Materialized Views y AggregatingMergeTree.

Ejercicio Crear una vista materializada que consolide métricas en tiempo real desde una tabla de logs.

3. Modelado y optimización del almacenamiento

Objetivo: Comprender y crear estructuras de datos optimizadas y eficientes

1. Estrategias de particionamiento y ordenación (PARTITION BY, ORDER BY).
2. Motores MergeTree, ReplacingMergeTree, SummingMergeTree, CollapsingMergeTree.
3. Indexación secundaria y uso de data skipping indexes.
4. Compresión de columnas (LZ4, ZSTD, Brotli).
5. TTLs para limpieza automática de datos.
6. Caching de resultados y uso de join_use_nulls, max_threads.
7. Optimización de consultas masivas y planificación de merges.

Ejercicio Crear un esquema particionado por mes y comparar rendimiento con distintos códecs de compresión.

4. Alta disponibilidad, replicación y clustering

Objetivo: Asegurar la disponibilidad de datos y monitorización de los mismos.

1. Configuración de clústeres multi-nodo (cluster.xml).
2. Replicación asíncrona y coordinada con Zookeeper.
3. Balanceo de carga y tolerancia a fallos.
4. Configuración de Distributed tables y ReplicatedMergeTree.
5. Backups y restauración con clickhouse-backup.
6. Monitoreo de replicación (system.replication_queue).
7. Despliegue de clusters en contenedores (Docker Compose / Kubernetes).

Ejercicio Implementar un clúster de 3 nodos con replicación y probar la recuperación de fallos.

5. Integraciones y ecosistema

Objetivo: Dominar las distintas herramientas que enriquecen ClickHouse

1. Integración con Kafka mediante KafkaEngine.
2. Ingesta desde S3 / GCS / Azure Blob con s3() y URL engines.
3. Conectores con Spark, Airflow y dbt.
4. Uso de ClickHouse con Python (clickhouse-driver) y Go (clickhouse-go).
5. Exportación de resultados a Power BI, Grafana y Superset.
6. ClickHouse Cloud y conexiones seguras HTTPS/TLS.
7. Sincronización con bases OLTP (MySQL, PostgreSQL) usando MaterializedMySQL.

Ejercicio Crear un pipeline de streaming Kafka → ClickHouse → Grafana en tiempo real.

6. Seguridad, gobernanza y mantenimiento

Objetivo: Tener un buen control de acceso y asegurar un correcto uso a futuro

1. Control de acceso basado en roles (RBAC).
2. Cifrado en tránsito y en reposo.
3. Auditoría de consultas (system.query_log).
4. Gestión de cuotas y límites (max_memory_usage, max_concurrent_queries).
5. Optimización de recursos para entornos multicliente.
6. Diagnóstico de rendimiento con system.parts y system.merges.
7. Estrategias de actualización y migración de versiones.

Ejercicio Configurar autenticación LDAP y políticas de seguridad básicas.

7. Novedades, roadmap y casos de uso avanzados

Objetivo: Profundizar con ejemplos prácticos de uso avanzado

1. ClickHouse Cloud y despliegues serverless (novedad).
2. Nuevas funciones vectoriales y soporte para modelos AI/ML.
3. Consultas federadas y mejoras en motores externos (PostgreSQLEngine, MySQLEngine).
4. Uso de Iceberg y Delta como fuentes externas.
5. Optimizaciones recientes en el planificador de consultas (2024-2025 releases).
6. Monitoreo avanzado con Prometheus + ClickHouse Keeper.
7. Casos de uso reales: analítica web, IoT, seguridad y fintech.

Ejercicio Diseñar una arquitectura ClickHouse completa para analítica de datos en streaming y dashboards BI.

Curso de Agentic AI

1 | Introducción a Agentic AI: del asistente al agente autónomo

Objetivo: Conocer las bases de Inteligencia Artificial, las diferentes variantes de uso y como usarla.

1.1. De la IA generativa a la IA Agéntica
1.2. Limitaciones actuales de ChatGPT, Claude, Gemini y Copilot
1.3. Conexión con sistemas internos: el salto real a la autonomía
1.4. Diferencias entre LLM vs Agentic AI: pensar, planificar, actuar

Ejercicio práctico

Comparar en tiempo real: "¿Cuántos envíos están retrasados?" usando un LLM aislado vs un agente conectado a una base de datos simulada.

ChatGPT · Claude AI · Copilot

2 | Arquitectura Cognitiva de un Agente (Agentic Loop de 7 pasos)

Objetivo: Conocer el flujo optimo de trabajo con IA y ponerlo en práctica.

2.1. Goal: fijación de objetivos humanos o automáticos
2.2. Perceive: ingesta de información desde APIs, bases y memoria
2.3. Reason: razonamiento sobre datos internos/externos
2.4. Plan, Action, Observe, Learn: ciclo continuo de mejora

Ejercicio práctico

Crear un flujo donde un agente recibe un objetivo ("Genera un informe mensual"), lo descompone en subtareas y explica su plan.

3 | Agentic Frameworks (Open Source y Enterprise)

Objetivo: Explorar los diferentes Frameworks y agentes

3.1. LangChain, CrewAI, AutoGen, Google ADK, OpenAI ADK
3.2. Agno Framework (el usado en el PDF)
3.3. Componentes: Tools, Knowledge, Models, Orchestration
3.4. Memory & Human-in-the-Loop

Ejercicio práctico

Levantar un agente simple con Agno o LangChain que consulte un CSV o PDF como "knowledge".

LangChain · crewai · AutoGen · Agno

4 | Diseño de Agentes: Autonomía, secuencias y equipos multi-agente

Objetivo: Ingesta de datos, examinarlos y tratarlos

4.1. Agentes individuales vs equipos de agentes
4.2. Secuencias: Sales Agent → Report Agent → Email Agent
4.3. Orquestación de workflows con planificación dinámica
4.4. Integración con APIs, CRMs, ERPs y bases SQL

Ejercicio práctico

Diseñar un "pipeline agéntico" donde tres agentes cooperan:
- Recuperar datos
- Analizarlos
- Enviar un email automático.

5 | Caso de Uso Real: Analista de Solicitudes de Crédito (del PDF)

Objetivo: Explorar tratado y validaciones avanzadas de datos.

5.1. Validación documental (DNI, nóminas, datos obligatorios)
5.2. Verificación cruzada entre solicitud y documentos
5.3. Validaciones normativas: edad, ratio de esfuerzo, errores formales
5.4. Recomendación final razonada: aprobar, rechazar o pedir datos

Ejercicio práctico

Implementar un agente que evalúe una solicitud de crédito simulada y emita una recomendación justificada con checklist.

6 | Data Privacy, Security & Governance (EU AI Act)

Objetivo: Conocer como las diferentes alternativas de IA tratan la seguridad y datos.

6.1. ¿Se usan los datos para entrenar modelos? Comparación: OpenAI, Azure OpenAI, Claude, Gemini
6.2. Políticas de retención: 0-retention, 30 días, control empresarial
6.3. Residencia de datos en UE: cuándo es obligatorio
6.4. EU AI Act: niveles de riesgo (prohibido, alto, limitado, mínimo)

Ejercicio práctico

Clasificar 5 casos de uso de tu empresa según la categoría del EU AI Act y definir medidas mínimas de cumplimiento.

OpenAI · Claude · Gemini

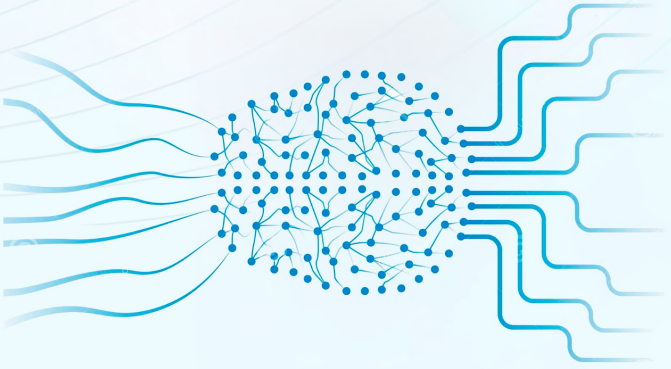
7 | Construcción de una Solución Empresarial Agentic (end-to-end)

Objetivo: Poner en práctica lo aprendido desde el inicio hasta el final.

7.1. Diseño de arquitectura: Agents + Tools + Knowledge
7.2. Implementación con APIs (REST) y memorias persistentes
7.3. Integración con MCP (Model Context Protocol) y monitoring
7.4. Próximos pasos: escalado, trazabilidad, auditoría y gobierno

Ejercicio práctico

Crear un blueprint de arquitectura Agentic AI para tu organización con:
- Agentes clave
- Flujos
- Herramientas integradas
- Puntos de control de gobernanza.



stratebi
Analytics and Big Data

Tecnología avanzada al servicio de tus datos.
Partners certificados:

Azure · aws · snowflake · PowerBI
databricks · LinceBI · talend
Qlik · VERTICA · ClickHouse · Google Cloud
Streamlit · KYLIGENCE · DAMA International
denodo · Microsoft Fabric · ANJANA DATA · dbt

Avda. de Brasil, 17 - Madrid
(+34) 917 88 34 10
www.stratebi.com

Emilio Arias

Professor of Master in Big Data and Business Intelligence

stratebi

Analytics and Big Data