Duración total: 160 horas con una intensidad horaria de 9 horas semanales durante 4.5 meses.
Formato del Programa: 100 % virtual y en vivo
Metodología: Aprendizaje teórico-práctico, con un proyecto práctico transversal y progresivo.
Panel Frente a Expertos: Los participantes desarrollarán un proyecto real de ingeniería de datos, integrando los conocimientos adquiridos en Big Data, automatización, procesamiento distribuido y despliegue en la nube.
Admisión: Mínimo de 290 / 480 puntos en la prueba de admisión.
Mentoría y feedback: Revisión individual y grupal de proyectos y código con expertos de la industria. Corrección de malas prácticas en ingeniería de datos.

Objetivo

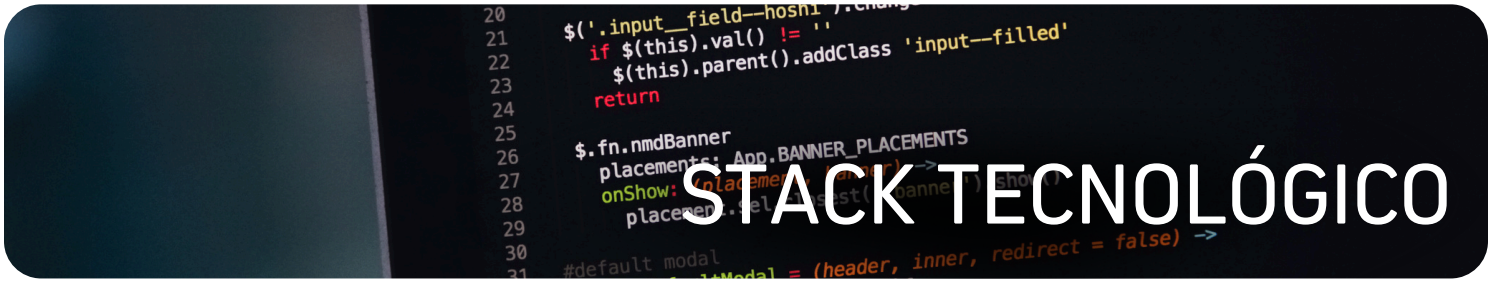
Formar ingenieros de datos capaces de diseñar, automatizar y escalar soluciones de datos modernas en la nube (AWS), con enfoque en pipelines distribuidos, modelado de datos, gobernanza, orquestación, tiempo real y visualización con QuickSight, al servicio de modelos de IA y decisiones estratégicas.

Perfil de Ingreso

Desarrolladores/as con mínimo 2 años de experiencia en desarrollo, manejo intermedio de bases de datos SQL y conocimientos básicos en Python, Git, APIs y cloud computing.

Perfil de salida

Al finalizar el programa, los participantes estarán en capacidad de desempeñarse como Ingenieros/as de Datos modernos, preparados para operar en equipos técnicos de alto desempeño, participando activamente en el ciclo completo de vida de los datos, desde la ingesta hasta la visualización y automatización en la nube.



- **Lenguajes**: Python, SQL, Bash
- **Bases de Datos**: PostgreSQL, MongoDB, DynamoDB
- **Procesamiento**: Apache Spark (PySpark), Kinesis, Kafka
- **Orquestación**: Apache Airflow, GitHub Actions, Prefect
- **Cloud**: AWS (S3, Redshift, Glue, Lambda, Lake Formation, DataZone), Terraform
- **Visualización**: Amazon QuickSight
- **Validación**: Great Expectations

Habilidades Adquiridas al Finalizar el Programa

- Comprender e implementar **arquitecturas de datos modernas** (data lakehouse, lambda, kappa).
- Aplicar metodologías de diseño de soluciones de datos (CRISP-DM, ASUM-DM).
- Diseñar pipelines escalables y desacoplados para entornos cloud y distribuidos.
- Modelar datos a nivel **conceptual, lógico y físico** para diferentes necesidades.
- Implementar esquemas **relacionales, NoSQL y multidimensionales** para analítica.
- Diseñar **bodegas de datos** y estructuras optimizadas para BI y reportes.
- Dominar el uso de **PostgreSQL, MongoDB y DynamoDB** con foco en escalabilidad.
- Entender estrategias de **particionamiento, sharding y clustering.**
- Implementar soluciones de almacenamiento en **data lakes** sobre AWS S3.
- Usar **PySpark** para transformar datos de manera distribuida.
- Aplicar técnicas de **optimización de jobs Spark** (Catalyst, Tungsten).
- Crear soluciones de **streaming en tiempo real** con Kinesis y Kafka.
- Aplicar políticas de **gobierno de datos** usando AWS Lake Formation y DataZone.
- Implementar **validaciones de calidad** con Great Expectations.
- Automatizar workflows con **Apache Airflow**, CI/CD y Terraform.
- Desplegar pipelines en AWS usando servicios como **S3, Glue, Lambda, Redshift.**
- Gestionar infraestructura como código con **Terraform.**
- Garantizar **seguridad, monitoreo y confidencialidad de datos** (IAM, CloudWatch, encriptación, anonimización).
- Crear dashboards interactivos con **Amazon QuickSight**, conectando modelos de datos BI.
- Diseñar reportes orientados a negocio con foco en métricas e indicadores clave.

Estructura del Programa

Módulo I – Fundamentos de Ingeniería de Datos y Cloud

- Rol del ingeniero de datos y su relación con IA
- Metodologías: CRISP-DM, ASUM-DM
- Gobierno de datos: calidad, linaje, privacidad
- Arquitecturas de datos: monolítica, distribuida, data lakehouse
- Patrones comunes: ingestion, ETL/ELT, lambda, kappa
- Fundamentos de almacenamiento cloud con AWS S3

Módulo II – Bases de Datos Relacionales y NoSQL

- SQL avanzado, normalización, modelado relacional
- MongoDB y DynamoDB
- Esquemas particionados, clustering, sharding
- Integración entre bases relacionales y NoSQL

Módulo III – Arquitectura Escalable y Data Lakes

- Hadoop y HDFS (bases conceptuales)
- Apache Hive e Impala
- AWS Glue y Glue Catalog
- Iceberg y Hudi como frameworks para lagos transaccionales
- Gobernanza: AWS LakeFormation, DataZone

Módulo III – Arquitectura Escalable y Data Lakes

- Hadoop y HDFS (bases conceptuales)
- Apache Hive e Impala
- AWS Glue y Glue Catalog
- Iceberg y Hudi como frameworks para lagos transaccionales
- Gobernanza: AWS LakeFormation, DataZone

Módulo IV Procesamiento Distribuido con Spark

- Apache Spark & PySpark
- DataFrames y RDDs
- Optimización con Catalyst y Tungsten
- Procesamiento batch vs. real-time
- Spark jobs en producción

Módulo V Procesamiento de Datos en Tiempo Real

- Fundamentos de streaming
- Apache Kafka vs. AWS Kinesis

- Productores y consumidores
- Casos de uso de tiempo real aplicados a IA
- Hands-on con pipelines en tiempo real

Módulo VI DataOps y Automatización

- Apache Airflow: DAGs, scheduling, logging
- CI/CD con GitHub Actions o Jenkins
- Validación de datos con Great Expectations
- Automatización de pipelines y tareas recurrentes

Módulo VII Modelado de Datos y Bodegas

- Modelado conceptual, lógico y físico
- Modelado multidimensional
- Diseño de Bodegas de Datos
- Conexión a BI con énfasis en dimensiones analíticas

Módulo VIII Cloud & DevOps en AWS

- AWS: S3, Glue, Redshift, Lambda
- Terraform: infraestructura como código
- Seguridad en AWS: IAM, políticas
- Observabilidad: CloudWatch, métricas
- Confidencialidad: encriptación, anonimización, protección de datos sensibles