Programa de Especialización Analítica

DATA ENGINEER

BROCHURE 2024

PRESENTACIÓN

El avance tecnológico ha promovido el crecimiento exponencial de los datos almacenados por las empresas, que es lo que hoy se denomina Big Data, que a su vez ha motivado el surgimiento de nuevas técnicas y herramientas de recolección, organización y almacenamiento para tales volúmenes de información. Ahora, el reto está en proveer soluciones que transporten y democraticen la data en toda la organización, y constituyan un soporte confiable.

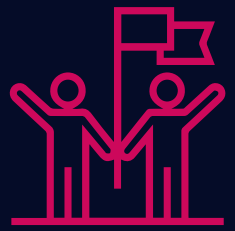
Por ello DMC Perú, presenta el **Programa de Especialización: Data Engineer** donde aprenderás el diseño e implementación de soluciones ETL desde diversas fuentes hacia repositorios como Data-Warehouses, Data-Lakes, entre otros; empleando lenguajes de programación líderes en el mercado como Python y Transact-SQL, contemplando en el proceso el paradigma y las herramientas para datos masivos, y complementos importantes como Web Scraping, Real-Time Data, DataOps, entre otros.

Pre-Requisitos
• Experiencia en desarrollo de software, y conocimientos de programación y estructura de bases de datos.

	Inicio: 14/03/24 - Fin: 24/09/24
	188 horas académicas
	Martes y jueves 7:30pm a 10:30pm

OBJETIVO GENERAL

Aprende a diseñar e implementar soluciones ETL desde diversas fuentes hacia repositorios como Data-Warehouses, Data-Lakes, entre otros; empleando lenguajes de programación como Python y Transact-SQL, contemplando en el proceso y las herramientas para datos masivos, y complementos como Web Scraping, Real-Time Data, DataOps, entre otros.



OBJETIVOS ESPECÍFICOS

- **Fundamentals of Data Engineering:** Aprende a utilizar Python y Transact-SQL para para el diseño de soluciones ETL.
- **Big Data Specialization:** Aprende sobre los fundamentos de Big Data, su arquitectura, herramientas y los lenguajes disponibles para manipulación de datos para entornos On-premise como cloud para soluciones de ingeniería de datos.
- **Tools for Data Engineer:** Aprende a emplear herramientas y frameworks como Web Scraping para la automatización de extracción de datos, Data Visualization para comprender los ETL desde la vista del usuario final; así como DataOps, Data Integration y Orchestration para la automatización de los flujos de trabajo.

OBJETIVO MODULAR

- **Python for Data Engineering:** Aprende a emplear el lenguaje Python para la implementación de soluciones ETL, destinadas al transporte y procesamiento de datos entre diversas fuentes, empleando en el proceso la librería Pandas, los DataFrames y sus diversos métodos.
- **SQL for Data Engineering:** Aprende a emplear el lenguaje T-SQL para la implementación de Scripts que cumplan la función de ETL entre bases de datos locales como externas.
- **Workshop Bases de datos No-SQL:** Aprende sobre la estructura de las bases de datos No-SQL y como realizar acciones CRUD sobre ellas empleando el motor MongoDB y código pre-elaborado.
- **Big Data Processing:** Aprende sobre el almacenamiento distribuido y a utilizar diversas herramientas para la implementación de Data-Lakes y para el tratamiento de datos a gran escala en modos Batch y Real-Time, como son Apache Hive, Apache Spark, Databricks, y Apache Kafka.
- **Cloud Data Engineering:** Aprende a utilizar los principales servicios de los proveedores cloud líderes en el mercado como Azure, GCP y AWS, para el diseño e implementación de ETL básicos.
- **Data Visualization:** Aprende a realizar el tratamiento de datos y la presentación de los mismos desde la perspectiva del usuario final, empleando Power BI y Power Query, con el propósito de alinear los entregables de los pipelines de datos.
- **Web Scraping con Python:** Aprende a emplear el lenguaje Python para la implementación de soluciones que permitan automatizar la etapa de extracción de datos desde una web, mediante diversos métodos y técnicas disponibles.
- **DataOps:** Aprende sobre DataOps, la adaptación de DevOps para entornos de datos, para la automatización de los flujos de trabajo entre los equipos de Datos y Operaciones, para lo cual emplearás herramientas como Jenkings y GitHub.
- **Data Integration & Orchestration:** Aprende a automatizar tus pipelines de datos mediante la herramienta AirFlow, como una forma de garantizar la entrega oportuna de los datos sobre todo en entornos grandes.

Dirigido a

Profesionales de ingeniería de sistemas, informática, desarrollo de software y similares interesados en incursionar en soluciones de ingeniería de datos e implementación de pipelines. Áreas de inteligencia de negocios o ciencias de datos interesados en conocer tecnologías y herramientas que soporten y automaticen los flujos de datos para sus proyectos habituales. Áreas de ingeniería de datos, interesados en actualizar su tool-set con herramientas y técnicas de vanguardia en este campo de los datos.

CARACTERÍSTICAS

Clases en Vivo

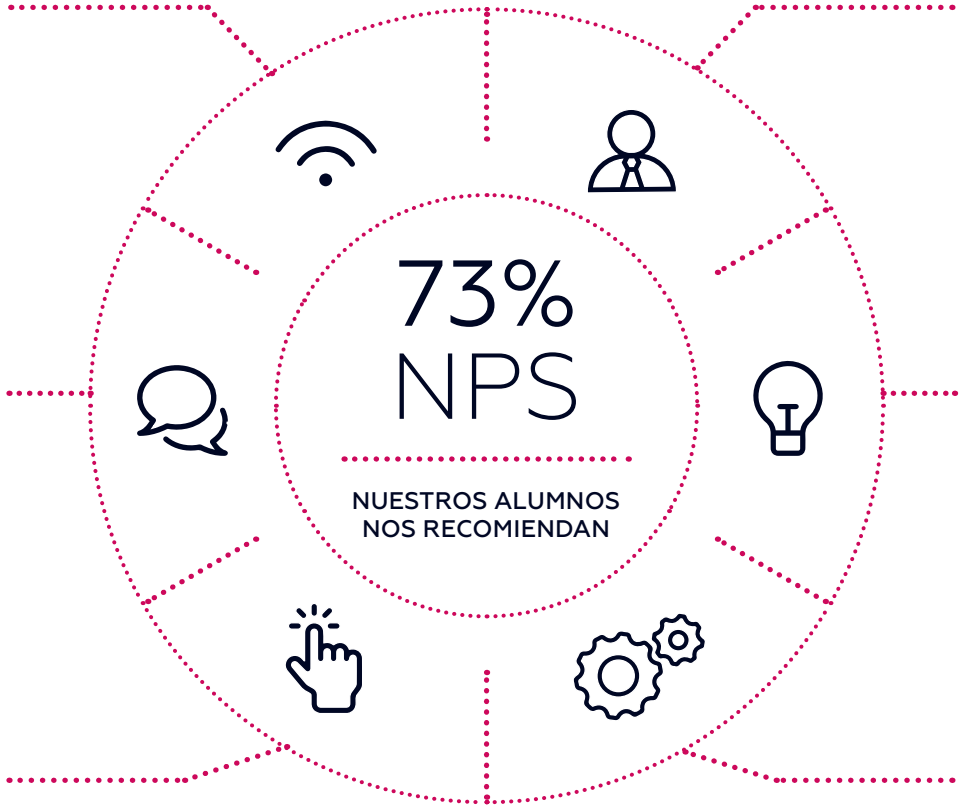
El 100% de las clases que se desarrollan en el programa son en vivo.

Asesoría Académica

Resuelve tus dudas con el asesor académico en línea

Plataforma E-Learning

Accede en cualquier momento a materiales complementarios: lecturas, videos, tutoriales, clases grabadas y más.



Aprende haciendo

Desarrolla casos con datos reales, incluso puedes proponer casos de tu propio sector.

Proyecto integrador

Pondrás a prueba tus ideas para convertirlas en soluciones analíticas.

Soporte técnico

Asistencia técnica permanente y acceso a máquinas virtuales de ser necesario.

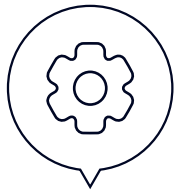
MALLA CURRICULAR

Taller de marca personal y empleabilidad
4 horas académicas

Python for Data Engineering
24 horas académicas

SQL for Data Engineering
24 horas académicas

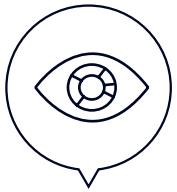
Workshop: Bases de datos No-SQL
4 horas académicas



Big Data Processing
48 horas académicas

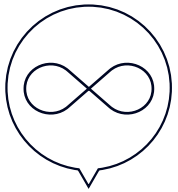
Cloud Data Engineering
12 horas académicas

SOFT SKILLS | FUNDAMENTALS OF DATA ENGINEERING | BIG DATA SPECIALIZATION



Data Visualization Fundamentals
12 horas académicas

Web Scraping con Python
12 horas académicas



DataOps
20 horas académicas

Data Integration & Orchestration
16 horas académicas

Proyecto Integrador
12 horas académicas

TOOLS FOR DATA ENGINEER

PROGRAMA

1

Python for Data Engineering

Emplea Python para soluciones ETL, para el transporte y procesamiento de datos entre fuentes con librería Pandas, DataFrames y más.

Fundamentos de ETL con Python
- ETL. Definición y herramientas.
- Herramientas Python para ETL.
- Python y sus entornos de ejecución.

Introducción a Python
- Manejo de excepciones e instrucciones.
- Tipos de datos en Python.
- Creación de Programas en Python.
- Interactuando con el OS.
- API. Definición y librerías para extraer datos.
- Taller: Consulta de datos desde un API.

Object Relational Mapper
- ORM. Definición, ventajas de su uso.
- Tipos de ORM en Python.
- SQLAlchemy. Definición y características.
- Taller: Creación de un Engine con SQLAlchemy.
- Taller: Conexión a base de datos con SQLAlchemy.

Pandas. Series y Dataframes
- Pandas. Definición, carga en Python.
- Pandas Series. Características y uso de vectores.
- Operaciones con Series. Búsquedas, Slicing, operaciones aritméticas, tipos de datos.
- Pandas DataFrames. Características y uso de
- DataFrames. Diferencias respecto a Series.
- Operación con DataFrames. Creación, descripción, visualización.
- Operaciones de agrupación. Agrupaciones directas y por Agregación simple y múltiple (varios campos).
- Guardar DataFrames en archivos planos (Json y CSV) y base de datos (MySQL).
- Taller: Carga de datos de un API, procesamiento, y descarga en una base de datos.
- Taller: Carga de datos desde un archivo plano, procesamiento y descarga en una base de datos.

2

SQL for Data Engineering

Emplea T-SQL para la implementación de Scripts que cumplan la función de ETL entre bases de datos locales como externas.

El Lenguaje Transact-SQL
- SQL y T-SQL. Definición, diferencias.
- Lenguaje de definición de datos (DDL). Definición, alcance y comandos asociados (create, alter, drop).
- Lenguaje de manipulación de datos (DML). Definición, alcance y estructura del comando SELECT...FROM...
- Consultas básicas, uso de SELECT...FROM...
- Consultas condicionales, uso de WHERE y operadores lógicos.
- Consultas de agregación, uso de GROUP BY, COUNT, MAX, MIN, SUM, AVG,
- Pivoteo de tablas, uso de PIVOT.
- Consultas multi-tabla. Uso del comando JOIN y variantes (LEFT, RIGHT, FULL)
- Operadores de conjunto, uso de UNION, INTERSECT,EXCEPT.
- Taller: Extracción de datos desde una base de datos local con comandos T-SQL.

Transact-SQL Avanzado
- Transformación y operación de columnas, uso de operadores aritméticos, funciones de fechas, funciones de textos, uso de IIF, ISNULL, NULLIF.
- Filtrado avanzado, uso de IN, ANY AND SOME, ALL, EXISTS.
- Conversión de tipos de datos, uso de CAST, CONVERT,FORMAT, PARSE.
- Encapsulamiento de consultas en Procedimientos almacenados. Uso de estructuras condicionales y bucles.
- Carga de datos externos, uso de Linked Servers, OpenQuery y OpenRowSet.
- Taller: Transformación de datos externos con comandos T-SQL.

Tópicos Avanzados sobre Tablas
- Tablas derivadas y tablas temporales como alternativas para pre-procesamiento de datos.
- Instrucciones tipo Query jerárquica y correlaciónales.
- Manejo de expresiones de tablas (CROSS APPLY, OUTER APPLY).
- Uso de cursores para copia y transferencia de datos.
- Automatización de scripts con Jobs (schedule).
- Taller: Diseño de un script ETL con lenguaje T-SQL (De BD a BD)

3

Workshop Bases de datos No-SQL

Estructura de las bases de datos No-SQL y como realizar acciones CRUD sobre ellas con MongoDB y código pre-elaborado.

Fundamentos de No-SQL

- Bases de datos NoSQL. Definición y tipos.
- MongoDB. Definición, métodos y tipos de datos que soporta.
- Terminología y Conceptos (comparativo SQL y MongoDB).

Operaciones con MongoDB

- Operaciones CRUD en MongoDB.
- Creación de documentos, uso de INSERTONE() e INSERTMANY().
- Uso de diversas fuentes de datos: documentos definidos previamente, array de documentos, archivos JSON.
- Lectura de documentos, uso FIND().
- Actualización de documentos, uso de UPDATEONE(), UPDATEMANY(),
- REPLACEONE().
- Eliminación de documentos, uso de DELETEONE(), DELETEMANY(),
- REMOVE().
- Consulta de documentos.
- Consultas simples, uso de FIND(), FINDONE().
  Consultas avanzadas, comparación, Cadenas, Existencia.
- Consultas con operaciones lógicas, uso de Y ($AND), O ($OR),
- Negación ($NOT), Dos expresiones ($NOR).
- Consulta de arrays, uso de DOT NOTATION
  Consultas en subdocumentos.
  Uso de Cursores.

4

Big Data Processing

Aprende a implementar Data-Lakes, y modos Batch y Real-Time, como son Apache Hive, Apache Spark, Databricks, y Apache Kafka.

Introducción a Big Data

- Big Data. Definición, filosofía, las Vs.
- Big Data como marco de trabajo.
- Arquitectura conceptual.
- Componentes tecnológicos disponibles.
- Arquitectura moderna de datos.

Almacenamiento distribuido con Apache Hadoop

- Tecnologías Open-source para Big Data.
- Fundamentos de Apache Hadoop.
- Almacenamiento distribuido en HDFS.
- Taller: Procesamiento de datos con Apache Hive.
- Diferencias de Map Reduce vs Tez vs Apache Spark.
- Datalake. Definición y arquitectura (capas).
- Taller: Poblamiento de un Datalake con Apache Tez, Apache Hive y HDFS.

Procesamiento distribuido con Apache Spark

- Introducción a Spark.
- Funciones con PySpark.
- Extracción y Transformación de datos.
- Dataframes y RDDs.
- Funciones de Apache Spark.
- Tunning en Apache Spark.

Ingeniería de datos con Databricks

- ¿Qué es Databricks?
- Databricks Community vs Azure Databricks.
- Conociendo la interfaz de Databricks.
- Data lakehouse vs Datalake.
- Iceberg vs Delta lake vs Apache Hudi.
- Taller: Creación de una Lakehouse con arquitectura Medallion con Delta Lake.

Real-time Fundamentals

- Real-time en datos. Definición, casos de aplicación, diferencias respecto a Batch.
- Arquitectura Publicador – Suscriptor.
- Tecnología para carga Real-Time en Big Data.
- Taller: Uso y configuración de Apache Kafka para carga de datos Real-Time.
- Apache Spark & Kafka setup enviroments.
- Taller: Integración de Apache Spark con Apache Kafka para procesamiento de datos Real-time.

5

Cloud Data Engineering

Utiliza los servicios cloud líderes en el mercado como Azure, GCP y AWS, para el diseño e implementación de ETL básicos.

Fundamentos de Cloud Computing

- ¿Qué es computación en la nube?
- Conceptos de virtualización, Uso bajo demanda, despligue escalable.
- Ventajas del cloud computing.
- Regiones y zonas de disponibilidad.
- Tipos de nubes.
- IAAS, PAAS y SAAS.

Introducción a la Ingeniería de datos con AWS

- Introducción a la Consola de AWS.
- Principales servicios de data en AWS. S3, EMR, Cloud Funtions, IAM, Redshift, Athena,etc.
- Arquitectura de datos en AWS.
- Taller: Diseño e implementación de un ETL básico con AWS.

Introducción a la Ingeniería de datos con Azure

- Introducción a la Consola de Azure.
- Principales servicios de data en Azure. Blob Storage, Data factory, Databricks, Synapse, Event Hub, Azure SQL.
- Arquitectura de datos en AZURE.
- Taller: Diseño e implementación de un ETL básico con Azure.

Introducción a la Ingeniería de datos con GCP

- Introducción a la Consola de Azure.
- Servicios principales servicios de data en GCP. Cloud functions, Cloud Storage, Bigquery, Dataproc, data fusion, composer.
- Arquitectura de datos en GCP.
- Taller: Diseño e implementación de un ETL básico con GCP.

6

Data Visualization

Conoce el tratamiento de datos y su presentación como usuario final, con Power BI y Power Query, alineando los entregables de los pipelines de datos.

- Inteligencia de negocios y herramientas de visualización.
- Power BI for Big data. Conexión a fuente Cloud.
- Transformación de datos con Power Query.
- Taller: Conexión y transformación de fuente de datos con Power BI y Power Query.

7

Web Scraping con Python

Automatización en la etapa de extracción de datos desde la web.

- Web Scraping. Definición, debate legal-ético.
- Métodos de extracción de datos.
- Técnicas de web scraping. Estático, Requests, Dinámico.
- Taller: Scraping a un sitio web y almacenamiento de su contenido.

8

DataOps

Aprende sobre DataOps, la adaptación de DevOps para entornos de datos, para la automatizar workflows con Jenkings y GitHub.

Fundamentos de DataOps

- DataOps. Definición y características.
- Devops vs DataOps.
- Conceptos asociados: Continuos-delivery y Continuos-integrations.
- Servicios de automatización de despliegue: Jenkins, Azure Devops, Github Actions, Gitlab CI).
- Fundamentos de Infraestructura como código. Definición y características de Terrraform.

Git and GitHub

El control de versiones. Definición y características.

- Git. Definición, principales comandos.
- GitHub. Definición.
- Configuración de llaves SSH.
- Directorio de trabajo.
- Commit.
- Fusionar y el comando Merge.
- Ramas y el comando Branch.
- Taller: Creación de un repositorio en GitHub.
-

Jenkins

Jenkins. Definición, características, configuración e instalación básica.
Configuración de un Job.
Plugin de Jenkins.
Conexión a GitHub.
Taller: Despliegue de código automatizado.

9

Data Integration & Orchestration

Automatiza tus pipelines de datos mediante AirFlow, para la entrega de los datos en entornos grandes.

Apache Airflow

- Apache Airflow. Definición.
- DAG (Direct-Acyclic-Graph). Definición y casos de uso.
- Uso de Scheduler.
- Task and Operator.
- Taller: Implementado un DAG.
- Bash Operator,Python Operator y Apache Spark Operators.
- Taller: Procesamiento de datos con Apache Airflow.

PROYECTO INTEGRADOR

Desarrolla una solución de ingeniería de datos (ETL) con tecnología específica, acorde a las necesidades y requerimientos de un negocio con una problemática real.

EVALUACIÓN

+ Evaluaciones parciales: 40%
+ Trabajo final: 60%
+ Nota mínima aprobatoria: 14
+ Asistencia mínima: 80%

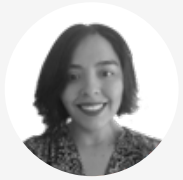
CERTIFICACIÓN

Por haber aprobado el Programa en Data Engineer. 188 horas académicas.

+Big Data Professional Certificate (BDPC).
+DevOps Essentials Professional Certificate (DEPC).
Emitido por CertiProf

* Todos los certificados se emitirán de forma digital luego de concluido el programa.

PLANA DOCENTE



SHEILLA LA ROSA

Educadora y
Terapeuta gestalt

Independiente

ANGEL TINTAYA

Senior Data
Engineer

number8

JUAN SALINAS

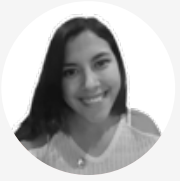
Senior Data
Engineer

encora

FRANCIS DE LA CRUZ

Big Data Architect

/bluetab



MILAGROS VILLEGAS

Business Intelligence
Consultant

Globant

TONY TRUJILLO

Data Architect

DATHA

*En caso de contingencias podría cambiar alguno de los docentes por otro profesional de similar perfil.

Buscamos liderar la transformación de las empresas.

+14 Años de Experiencia

Desde el año 2009 capacitamos con técnicas de análisis de datos a profesionales de diferentes empresas y sectores.

+18K Profesionales Capacitados

Nuestros alumnos inscritos pertenecen a las mejores compañías del medio y amplifican con nosotros su red de contactos especializada.

+300 Empresas Asesoradas

Las empresas top del mercado buscan nuestra asesoría y les brindamos soluciones analíticas ad hoc.

DIIC Perú

+100 Expertos en Analítica

Nuestra plana docente ocupa los cargos más importantes en las áreas analíticas de todos los sectores.

+50 Capacitaciones Especializadas

Contamos con una variedad de líneas temáticas y niveles de especialización.

+5 Big Data Analytics Summit

Organizamos el evento más grande del Perú, con los mejores ponentes internacionales.

BIG DATA

Formando profesionales mediante la analítica de los datos.


Perú