

Capacitación de analítica avanzada

Científicos de datos e ingenieros de datos

Introducción al programa



Agenda

1

Descripción del curso

4

Cronograma

2

Objetivos del curso

5

Prueba de reconocimiento
- *HackerRank*

3

Estructura del programa

6

Equipo e instructores

Descripción del curso

Desarrollamos un curso a la medida para el equipo de Alicorp

Homologación de conocimientos y visión global

Este curso propone una introducción a los fundamentos de programación, algoritmia, nuevas tecnologías y teoría estadística y de modelación necesaria para **crear soluciones de analítica avanzada en los roles de ingeniero y científico de datos.**

Teoría y práctica

Los participantes reforzarán sus **fundamentos teóricos** de esta disciplina y a través de **sesiones prácticas** en cada módulo podrán ver la integración de los conocimientos en un ambiente aplicado a la solución de problemas en la industria.

Resolución de problemas y enfoque en resultados

El programa apalanza la experiencia de McKinsey & Company en la creación de soluciones de analítica avanzada en conjunto con los perfiles diversos de los instructores para crear una experiencia de **aprendizaje basado en la solución de problemas reales de la industria y las aplicaciones en Alicorp.**

Recomendación

El programa está diseñado para incluir conceptos de ingeniería y ciencia de forma cohesiva.

Por esta razón recomendamos que tanto científicos como ingenieros de datos participen de todo el curso simultáneamente con el fin de crear un perfil **full-stack en datos.**

Nuestro objetivo es generar un lenguaje común y una mutua comprensión de las problemáticas de ambos perfiles para facilitar sinergias y aumentar la productividad en su colaboración.

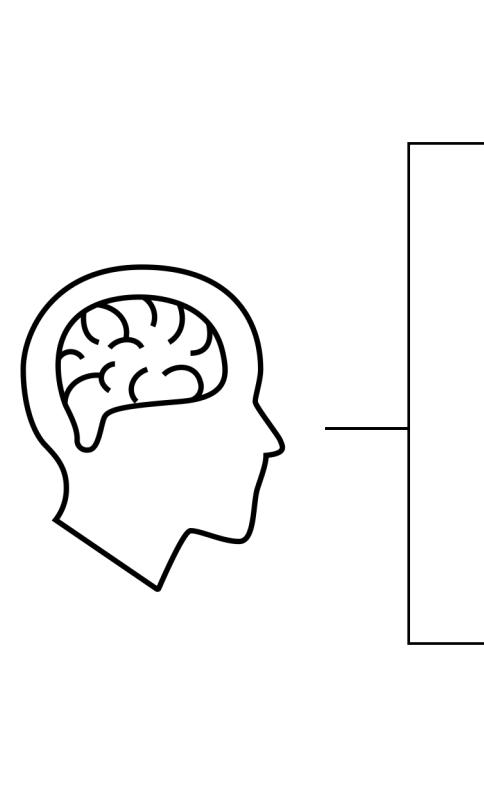
Objetivos del curso

Diseñamos el programa para cumplir con los objetivos futuros de desarrollo de Alicorp

Objetivos generales

Al final del curso se espera que los estudiantes sean capaces de desarrollar soluciones de analítica avanzada ya sea en el rol de ingeniero de datos o científicos de datos. De esta forma, se espera puedan:

- Configurar el ambiente de desarrollo
- Desarrollar código de producción usando metodologías de desarrollo de software
- Crear soluciones para la implementación modelos
- Entender los principios teóricos de probabilidad y estadística inherentes a los modelos de ML, para entender sus limitaciones y casos de aplicación



Científico de datos

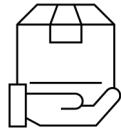
- Crear análisis descriptivos usando conceptos estadísticos avanzados y visualizaciones para la exploración de datos
- Crear algoritmos de aprendizaje automático en tareas de predictivas o no supervisados
- Utilizar técnicas de modelación alternativas como aprendizaje profundo o modelación de series temporales

Ingeniero de datos

- Configurar ambientes de producción usando herramientas cloud
- Crear pipelines de tratamiento y transformación de datos usando kedro
- Crear soluciones con contenedores para replicar soluciones en diferentes ambientes

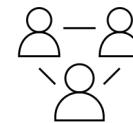
Estructura del programa

Organizamos el currículo para ser flexible y modular apoyado en 3 pilares



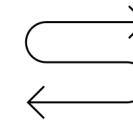
Logística

El curso está diseñado para ser impartido ya sea en formato presencial o virtual. El programa se estructura en módulos de aprendizaje que combinan formación teórica y ejercicios prácticos. Al final de cada modulo se hace una evaluación o mini proyecto realizado por los estudiantes de forma independiente.



Sistema de seguimiento

- Al final de cada modulo se hará un mini-proyecto para reforzar los conocimientos vistos
- En el caso del último modulo se hará una *datathon* que utiliza integralmente todo el conocimiento del curso para desarrollar una solución E2E a un problema real de **Alicorp**
- Al inicio y final del curso se hará una prueba general de los conocimientos usando la plataforma *HackerRank*



Estructura del curso

Es curso proporciona una introducción a los conocimientos y herramientas básicas crear proyectos de analítica avanzada:

- Consta de **10 módulos** incluyendo uno inicial como introducción y uno final como datathon
- Tiene una duración aproximada de **4 meses** en 12 sesiones

Profundización

Se puede profundizar los conocimientos de este curso en un siguiente nivel (avanzado)

Cronograma (1/2)

- Intervenciones teóricas
- Intervenciones prácticas en modo demo/tutorial

#	Modulo	Área	Tipo	Sesión
1	Introducción	Intro IA y ML	●	Introducción a la inteligencia artificial y el aprendizaje automático
2.2	Fundamentos para el análisis de datos	Coding	●	Sistema <i>Unix</i> , terminal y bash
2.1	Fundamentos para el análisis de datos	Coding	●	Configuración del ambiente de desarrollo: brew, apt, conda & pip
2.3	Fundamentos para el análisis de datos	Coding	●	Fundamentos de Python
2.4	Fundamentos para el análisis de datos	Coding	●	Fundamentos de Python
2.5	Fundamentos para el análisis de datos	Data engineering	●	Fundamentos de Pandas para el análisis de datos
				Mini-proyecto
3.1	Teoría de probabilidad y estadística	Teoría matemática	●	Probabilidad y estadística: distribuciones principales, teorema de Bayes, ley de números grandes, teorema del límite central
3.2	Teoría de probabilidad y estadística	Visualización de datos	●	Análisis descriptivos con matplotlib, seaborn y plotly: matrices de correlación, boxplots, cuartiles, visualizaciones interactivas
				Mini-proyecto
4.1	Optimización y simulación	Optimización	●	Visión general de teoría: LP, MILP, combinatorial opt., no- linear, non-deterministic, teoría de control, teoría de grafos
4.2	Optimización y simulación	Simulación	●	Simulación: MonteCarlo, Agnet-based, Cadenas Markov
4.3	Optimización y simulación	Optimización y simulación	●	ORtools: Toolkit de optimización
				Mini-proyecto
5.1	Modelación 1: aprendizaje automático	Machine Learning y AI	●	Principios de aprendizaje supervisado: regresion lineal, logistica, arboles, random forest, SVM,
5.2	Modelación 1: aprendizaje automático	Machine Learning y AI	●	Principios de aprendizaje supervisado: ensambles
5.3	Modelación 1: aprendizaje automático	Machine Learning y AI	●	Principios de aprendizaje no supervisado: reducción de dimensionalidad, clustering
5.4	Modelación 1: aprendizaje automático	Machine Learning y AI	●	Métricas de performance
5.5	Modelación 1: aprendizaje automático	Machine Learning y AI	●	ScikitLearn: toolkit de aprendizaje “shallow”
5.6	Modelación 1: aprendizaje automático	Machine Learning y AI	●	Estrategias de entrenamiento: balanceando set de data, validaciones cruzadas, búsqueda grid, selección de características
5.7	Modelación 1: aprendizaje automático	Machine Learning y AI	●	Training toolkit de estrategias y ejemplos
				Mini-proyecto

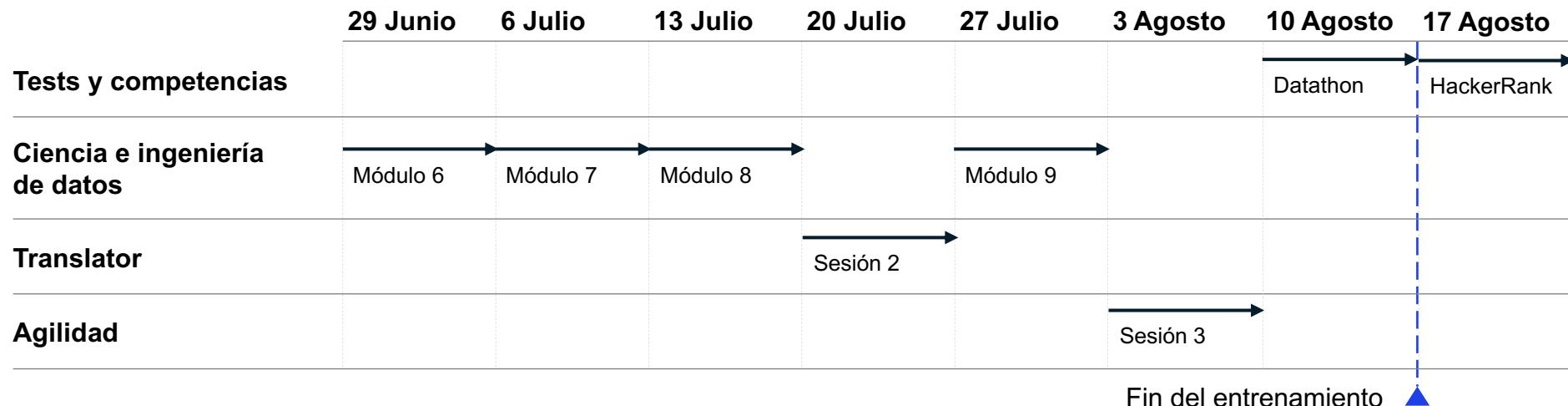
Cronograma (2/2)

#	Modulo	Área	Tipo	Sesión
6.1	Modelación 2: Deep Learning	Machine Learning y AI	●	Principios de aprendizaje supervisado: nets neurales + DL + RL
6.2	Modelación 2: Deep Learning	Machine Learning y AI	●	TensorFlow+Keras: toolkit de deep learning Mini-proyecto
7.1	Producción y back end	Data engineering	●	Despliegue e implementación de herramientas: flask, aws
7.2	Producción y back end	Coding	●	Editores: Command line, Pycharm, Visual Studio Code
7.3	Producción y back end	Coding	●	Principios de GitHub Mini-proyecto
8.1	Bases de datos	Data engineering	●	Visión general de manejo de bases de datos: relacional (SQL), no-relacional (MongoDB)
8.2	Bases de datos	Data engineering	●	Data pipelines: Kedro y Airflow Mini-proyecto
9.1	Infracstuctura	Data engineering	●	cloud computing: AWS, GCP, Azure
9.2	Infracstuctura	Data engineering	●	Cloud computing: AWS, GCP, Azure Mini-proyecto
10.1	Challenge	Datathon	●	Aplicar teoría a un problema práctico en un ambiente colaborativo
10.2	Challenge	Datathon	●	Presentar resultados al comité y obtener retroalimentación
10.3	Challenge	Datathon	●	Desarrollar un skill test a través de HackerRank para evaluar el punto de partida y evolución y que podrá ser utilizado para recruiting posteriormente Datathon

● Intervenciones teóricas

● Intervenciones prácticas en modo demo/tutorial

Calendario completo



1. Con algunas excepciones con previo aviso



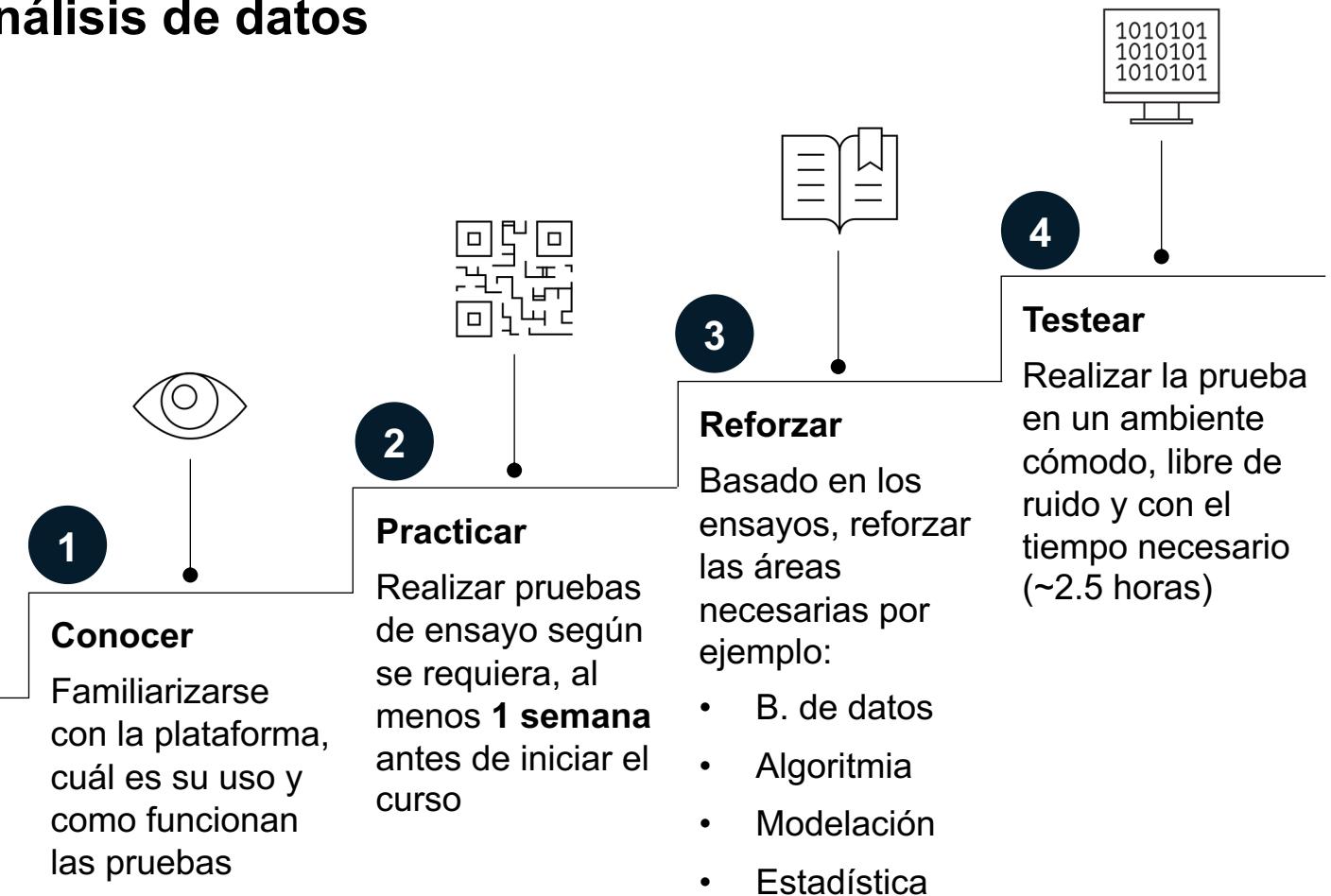
Prueba de reconocimiento – *HackerRank* (1/2)

Prueba inicial de *HackerRank* para análisis de datos

El curso inicia con la realización de una prueba inicial en los conceptos de programación, estadística y modelación usados para el trabajo con datos. Esta prueba nos dará un **punto de referencia** para nuestro iniciar nuestro curso.

Usaremos esta prueba para adaptar el curso y fortalecer los conocimientos del equipo.

Recomendamos el siguiente enfoque para realizar la prueba



Prueba de reconocimiento – *HackerRank* (2/2)

Paso

Conocer



Practicar



Reforzar



Testear

Descripción

Ingresa a la plataforma <https://www.hackerrank.com> y crea una cuenta para desarrolladores. Conoce la plataforma, sus preguntas de práctica y los proyectos y competencias **para ciencia de datos**

Realiza preguntas de prueba en **codificación, estadística, bases de datos y proyectos/competencias de modelación**. Cuando estés listo puedes hacer el test de prueba: <https://bit.ly/2K3yizW>

Refuerza los conocimientos que deseas, usando los tutoriales disponibles en la plataforma o fuentes externas.

Cada participante recibirá en su correo de **Alicorp**, un link habilitado por 1 semana para hacer la prueba (~3 horas).

Equipo

Equipo core



Pepe Cafferata
Partner
Sao Paulo



Jorge Grieve
Associate Partner,
Lima



Melissa Fitts
SEM
Bogotá

Expertos



Amalia Toro
EM, Agility en AA
Bogotá



Álvaro Fuentes
Data Science Specialist
Bogotá



Simón Tamayo
Data Science Specialist
Bogotá



Jonas Kemper
Data Engineer
Berlin



Erick Translateur
Data Science Fellow
Bogotá



Iván Torroledo
Data Science Fellow
Bogotá

Ivan Torroledo

Data Scientist, Bogotá



Ivan cuenta con varios años de experiencia aplicando herramientas de aprendizaje automático y aprendizaje profundo para finanzas y ciberseguridad.

También es experto en computación de alto rendimiento usando MPI, openMP, CUDA y lenguajes de programación de bajo nivel como Fortran y C para modelos de alta dimensión y programación distribuida.

En los últimos años, ha participado en proyectos de aprendizaje profundo para analizar datos no estructurados como imágenes o textos usando técnicas de procesamiento de lenguaje natural.

Experiencia profesional

Antes de unirse a QuantumBlack, Ivan trabajó en el banco central de Colombia aplicando herramientas HPC y modelos estadísticos al análisis de la política monetaria. También trabajó en una compañía de ciberseguridad aplicando herramientas de aprendizaje automático y aprendizaje profundo para desarrollar modelos predictivos.

Proyectos anteriores

Experiencia en varios sectores de la industria, incluyendo bienes de consumo, servicios financieros y ciberseguridad.

Sector financiero - Análisis macroeconómico y detección de fraudes.

Ciberseguridad: detección de anomalías, biométrica conductual, phishing y detección de malware.

Bienes de consumo - Segmentación de clientes utilizando aprendizaje no supervisado.

Antecedentes

Educación

Ivan tiene un BSc en Economía y un BSc en Física de la Universidad de los Andes, Colombia

Simon Tamayo

Data Scientist Specialist, Bogotá



Simon es especialista en data science en la oficina de Bogotá de -QuantumBlack de McKinsey & Company. Su trabajo se enfoca en la optimización industrial y aplicaciones de inteligencia artificial en manufactura y logística.

Antes de unirse a la Firma, fue profesor universitario en París, Francia, en los campos de Aprendizaje Automático e Investigación de Operaciones.

Ha publicado dos libros académicos sobre Optimización y Machine Learning, y más de 30 artículos científicos.

Experiencia relevante

Proyectos recientes

Creación de un modelo predictivo de crecimiento animal basado en las variables endógenas de cada ciclo productivo y datos exógenos ambientales. Incluida la creación del pipeline de procesamiento de datos y su implementación en la nube del cliente. Caja de herramientas: Python, Kedro, SkLearn, Pandas, TensorFlow, GCP.

Crear un modelo de recomendación para alimentar animales, capaz de reproducir mejores prácticas de la industria y optimizar el crecimiento. Caja de herramientas: Python, Kedro, SkLearn, Pandas, TensorFlow, GCP.

Optimizar un problema de asignación de oferta-demanda de gran tamaño. Dado 3 inputs: (1) capacidad de los proveedores; (2) demanda de clientes; y (3) costo de pedido de cada proveedor a cada cliente; encontrar las cantidades que cada cliente debe pedir de cada proveedor, con el fin de minimizar el costo total, mientras se suministra la cantidad máxima de demanda. Caja de herramientas: Python, SkLearn, Pandas, ORtools.

Antecedentes

Educación

Doctorado en Ciencias de la Computación, Université de Lorraine (Francia, 2011)

Maestría en Ingeniería Industrial, Ecole Nationale d'Ingénieurs (Francia, 2007)

Licenciatura en Ingeniería Mecánica, Universidad Eafit (Colombia, 2006)

Habla español, francés, inglés e italiano

Erick Translateur

Data Scientist, Bogotá



Erick Translateur es un data scientist de la oficina de Bogotá y se unió a la firma en el 2019. Previamente trabajó ~5 años en Quantil, una consultora de matemáticas aplicadas en Colombia. Se especializa en modelos de machine learning para el sistema financiero, cuantificación de riesgos y la valoración de derivados. También cuenta experiencia en el uso de modelos analíticos en diferentes industrias como la agroindustria y política.

Experiencia relevante

Desarrolló modelos de analítica avanzada para el sector agropecuario (Perú y Ecuador)

Experiencia previa

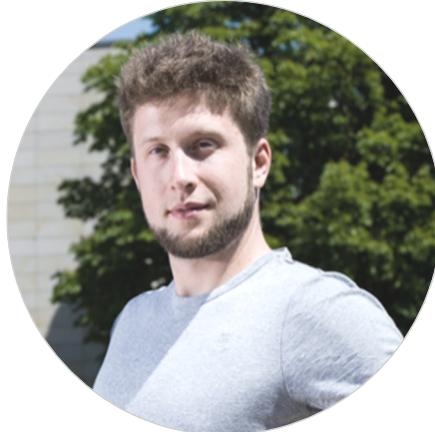
- Coordinó y desarrolló la selección del esquema de garantías de la primera subasta de energía renovable de Colombia.
- Dirigió proyectos relacionados al uso de la máquina learning para la predicción de financiera en el sector de pensiones en Colombia.
- Fue parte del desarrollo de estrategias de coberturas para empresas de petróleo y gas y construcción en Colombia. Se especializó en el análisis y optimización de coberturas de refinerías y oleoductos.
- Tiene experiencia en el desarrollo de algoritmos de comercio basado en máquina aprendiendo y estadística no paramétrica para la predicción de la tasa de cambio y los servicios colombianos.
- Tiene experiencia en riesgo de mercado y valoración de derivados en proyectos y bancarias de Colombia.
- Ha desarrollado investigación y trabajado en el uso de la máquina learning para la optimización de portafolios y detección de fraude.

Educación

- Erick es Economista y Máster en Economía de la Universidad de los Andes (Bogotá, 2017) donde obtuvo grado Cum Laude en ambos títulos. Habla fluido Español e Inglés.

Jonas Kemper

Data Engineer, McKinsey Digital Labs



Jonas es un ingeniero de Big Data con 5 años de experiencia trabajando en ambientes de start-ups, grandes creadores de emprendimientos y como Ingeniero de Investigación en las universidades de mayor rango.

Cuenta con experiencia automatizando procesos tediosos y diseñando arquitecturas de soluciones analíticas y de infraestructura de datos de última generación. Por lo general, esto implica adaptar a la medida los pipelines de datos complejos, como también desarrollar diseños innovadores y centrados en el usuario para informar las decisiones críticas en las operaciones diarias. Sus proyectos anteriores abarcaron una amplia gama de tecnologías futuras como inteligencia artificial cognitiva, realidad virtual, IoT y sistemas blockchain.

Expertise digital: Ingeniería de datos (++ Python, SQL ++ Java/Scala), infraestructura de procesamiento de datos, marcos Agile, diseño de producto, desarrollo backend, DevOps (Ci/CD, Gitlab, AWS, GCP, Terraform, Docker)

Experiencia reciente de McKinsey

Ingeniero de Datos, Digital Business Building en el Reino Unido - 2018

- Lideró el diseño y la implementación del almacén de datos y el pipeline analítico. Foco en diseño de AWS e implementación de arquitectura de micro-servicios con una API-gateway central para asegurar una escalabilidad sin esfuerzos de la app web de cara al cliente

Ingeniero de Datos, Scaling Analytics Solutions in Germany - 2018

- Creó de la estrategia para escalar prototipos de analítica dispersa en un importante player farmacéutico para una adopción a nivel de toda la organización lista para la producción.

Experiencia laboral previa

Analítica Patya, CTO/Co-Fundador

- Conceptualización, desarrollo y distribución de solución integrada de limpieza de datos
- Consultoría de ciencia de datos (ej. segmentación de clientes para marca online grande)

BCG Digital Ventures, Ingeniería de Software Full-Stack

- Creación de prototipos de hardware y desarrollo de firmware
- Infraestructura de datos y análisis de datos de sensores del sistema IoT

Antecedentes

Educación

Licenciatura en Informática e Interacción Humano-Computadora| LMU Munich

Maestría en ingeniería de sistemas de TI| Hasso-Plattner-Institute

Ingeniero de software de investigación| MIT media lab, UCSD Design Lab, SU