Carrera Data Science

Labs

Proyecto Grupal

#soyhenry

Proyecto Grupal

LABS: Proyecto Grupal -Temáticas

Taxis NYC & Weather

Datos históricos de viajes en taxis de la ciudad de Nueva York y una API del clima
Diversos KPIs y una serie de correlaciones entre viajes y clima

KPIs

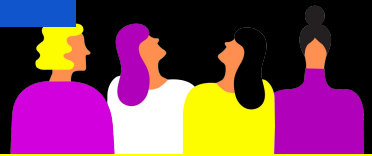
- Días de la semana con más viajes
- Barrios con mayor participación
- Correlación entre frío/calor y viajes
- Analytics sobre viajes/pasajeros/montos

Olist

Datos históricos de compras y envíos de una de las empresas más grandes de E-commerce de Brasil
Diversos KPIs y distintas correlaciones entre compras y estadios pandémicos

KPIs

- Performance del delivery
- Feedback de los productos y clientes
- Plcos de ventas
- Meses con mejor revenue



LABS: Proyecto Grupal
Objetivo Final

HENRY

Data Ingest

Dado una cantidad de datasets y API, poder obtener la estructura y datos

- Docker
- Python (pandas, numpy)
- MinIO Local, S3 compatible object-storage

Data Lake Storage

Almacenar los datos con un mínimo de limpieza y normalización

- Docker
- Python (pandas, numpy)
- Nifi

Data process

Mediante distintas técnicas y algoritmos vamos a proceder a actualizar nuestro sistema de almacenamiento de dato estructurado

- Docker
- Python (pandas, numpy)
- Airflow
- SQL

Data Warehouse ⭐

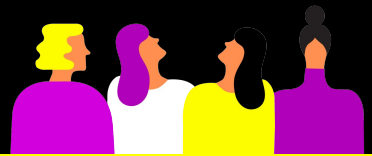
Sistema de almacenamiento de datos estructurados, sobre el cual la organización va a obtener sus datos para la toma de decisiones

- Docker
- Python (pandas, numpy)
- Airflow
- SQL

Data Analytics

Mediante reportes y visualizaciones vamos a facilitar la toma de decisiones

- Python (pandas, numpy)
- Airflow
- SQL
- PowerBI



LABS: Proyecto Grupal
Cronograma

HENRY

	W1 - Data Ingest 💾			W2 - Data Process 💻			W3 - Data Analytics 📈			W4 - Demo Final 🏆		
	Daily	Weekly	Demo	Daily	Weekly	Demo	Daily	Weekly	Demo	Daily	Weekly	Demo
Lunes	✅	✖	✖	✅	✖	✖	✅	✖	✖	✅	✖	✖
Martes	✅	✖	✖	✅	✖	✖	✅	✖	✖	✅	✖	✖
Miercoles	✅	✖	✖	✅	✖	✖	✅	✖	✖	✖	✖	✅
Jueves	✅	✖	✖	✅	✖	✖	✅	✖	✖	✖	✖	✅
Viernes	✖	✅	✅	✖	✅	✅	✖	✅	✅	✖	✖	✅
Objetivo	**Entender el alcance del proyecto y los datasets propuestos. Diseñar una solución y entregables.**			**Ingesta total de datos en un Data Lake local Diseño y creación del DW Creación de los Pipelines que alimentan el DW**			**Diseño y creación de reportes y visualizaciones KPIs a destacar Distintos niveles de presentación para distintas audiencias**			🎉 🚀		

LABS: Proyecto Grupal
Hitos y baseline

HENRY

Semana #1	Semana #2	Semana #3	Semana #4
Puesta en marcha el proyecto	**Trabajando los datos**	**Etapa de Analytics**	**Retoques finales y presentación**
• Kickoff del proyecto • Entendimiento de las necesidades • Documentar alcance, objetivo y entregables	• Creación del DW • Reglas de negocio aplicadas • Automatizar el DW	• Reportes • Storytelling • Ajustes necesarios al modelo	• Preparar demo por equipo • Entregable final • Documentacion

LABS: Proyecto Grupal
Hitos - Semana #1

HENRY

Semana #1

Puesta en marcha el proyecto

- Kickoff del proyecto
- Entendimiento de las necesidades
- Documentar alcance, objetivo y entregables

1. Entendimiento de la situación actual
2. Objetivos
3. Alcance
4. Fuera de alcance
5. Solución propuesta - Incluir Stack tecnológico
6. Metodología de trabajo
7. Diseño detallado – Entregables
8. Equipo de trabajo – Roles y responsabilidades
9. Cronograma general

LABS: Proyecto Grupal
Hitos - Semana #2

HENRY

Semana #2

Trabajando los datos

- Creación del DW
- Reglas de negocio aplicadas
- Automatizar el DW

1. Diseño adecuado del Modelo
2. Documentación
3. Pipelines para alimentar el DW
4. Automatización
5. Validación de datos

LABS: Proyecto Grupal
Hitos - Semana #3

HENRY

Semana #3

Etapa de Analytics

- Creación del DW
- Reglas de negocio aplicadas
- Automatizar el DW

1. Diseño de Reportes/Dashboards
2. Documentación
3. Pipelines para alimentar el DW
4. Automatización
5. Validación de datos

LABS: Proyecto Grupal
Hitos - Semana #4

HENRY

Semana #4

Retoques finales y presentación

- Preparar demo por equipo
- Entregable final
- Documentacion

1.Prepara la demo, visualización efectiva

2. Documentación

3. Probar todo el proceso antes!!!

LABS: Proyecto Grupal
Baseline esperado

HENRY

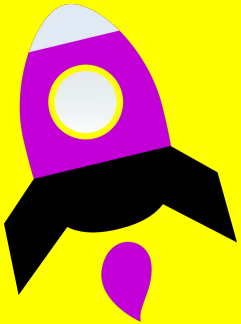
Semana #1	Semana #2	Semana #3	Semana #4
Puesta en marcha del proyecto y definiciones iniciales: • Al menos 4 KPIs • Tecnologías a usar • Documento de alcance del proyecto	**Trabajando los datos** • Datawarehouse automatizado con carga inicial. Al menos 2 tablas de hechos y 5 dimensionales	**Etapa de Analytics** • Carga incremental • Dashboard y reportes	**Retoques finales y presentación** • La presentación debe estar dirigida a la dirección de la Compañía • Storytelling
PLUS	PLUS	PLUS	PLUS
• Incrementar número de KPIs • Planificación y estimación de esfuerzos. Diagrama Gantt.	• Uso de herramientas Big Data como HDFS, Hive, Spark y/o motores No-SQL	• Implementar modelo de Machine Learning	• Implementar un reporte con visualización geográfica

LABS: Proyecto Grupal
Docker para trabajar el PF

https://github.com/sercasti/datalaketools

Alternativa https://github.com/Marcel-Jan/docker-hadoop-spark

Q&A



#soyhenry

Muchas Gracias

#soyhenry