

# DATASET IOT PROJECT

El objetivo de este documento es describir e identificar todos aquellos datos utilizados en el proyecto, así como mostrar en detalla como se desengranan en diversos campos que proporcionan la información relevante de cada Dataset participe en este desarrollo.

Los campos definidos a continuación son:

- Dataset: nombre de cada dataset usado en el proyecto.
- Fuente: origen de los datos empleados
- Descripción: explicación del formato y la naturaleza del dataset
- Descripción variables: detalle de las variables que definen cada dataset

Dataset	Fuente	Descripción	Descripción variables
User	<p>Generación de datos aleatorios mediante código python</p> <pre>def generatedata():      data={}     data["user_id"] = generate_user_id()     data["user_name"] = fake.name()     data["user_phone_number"] = generate_phone_number()     data["user_email"] = fake.email()     data["userinit_lat"] = str(random.uniform(39.4, 39.5))     data["userinit_lng"] = str(random.uniform(-8.4, -8.3))     data["userfinal_lat"] = str(random.uniform(39.4, 39.5))     data["userfinal_lng"] = str(random.uniform(-8.4, -8.3))     data["payment_method"] = random.choice(['Credit card', 'Paypal', 'Cash'])</pre>	<p>json generado a través de un diccionario vacío ("data") que se va completando a medida que la función "generatedata ()" crea aleatoriamente para cada cliente las variables descritas en el apartado siguiente</p>	<p><b>user_id:</b> id cliente <b>user_name:</b> nombre del cliente <b>user_phone_number:</b> teléfono del cliente <b>user_email:</b> correo electrónico del cliente <b>userinit_lat*:</b> punto (grados) de latitud donde se localiza el cliente inicialmente <b>userinit_lng*:</b> punto (grados) de longitud donde se localiza el cliente inicialmente <b>userfinal_lat*:</b> punto (grados) de latitud hacia donde se dirige el cliente <b>userfinal_lng*:</b> punto (grados) de longitud hacia donde se dirige el cliente <b>payment_method:</b> tipo de pago</p>

Taxi	<p>Generación de datos aleatorios mediante código python</p> <pre>def generateData():     data={}     data["taxi_id"] = generate_user_id()     data["taxi_phone_number"] = generate_phone_number()     data["taxi_lat"] = str(random.uniform(39.4, 39.5))     data["taxi_lng"] = str(random.uniform(-8.4, -8.3))     data["taxibase_fare"] = 4.00     data["taxikm_fare"] = 1.09</pre>	<p>json generado a través de un diccionario vacío ("data") que se va completando a medida que la función "generatedata ()" crea aleatoriamente para cada taxi las variables descritas en el apartado siguiente</p>	<p><b>taxi_id</b>: id cliente  <b>taxi_phone_number</b>: teléfono del cliente  <b>taxi_lat*</b>: punto (grados) de latitud donde se encuentra el taxi  <b>taxi_lng*</b>: punto (grados) de longitud donde se encuentra el taxi</p>
Big Query	<p>Depuración, clasificación y transformación de datos en apachebeam y carga en tablas big query de los datos resultantes</p>	<p>Diccionario de datos que alberga toda la información reunida creada por los generadores "User" y "Taxi". En este diccionario se recopilan datos nuevos tras los procesos de DataFlow</p>	<p><b>total_distance</b> : Metros totales del viaje  <b>taxibase_fare</b>: Precio mínimo por viaje  <b>taxikm_fare</b> : Precio por kilómetro recorrido  <b>transaction_amount</b> : Suma del precio mínimo y el coste de los kilómetros recorridos</p>