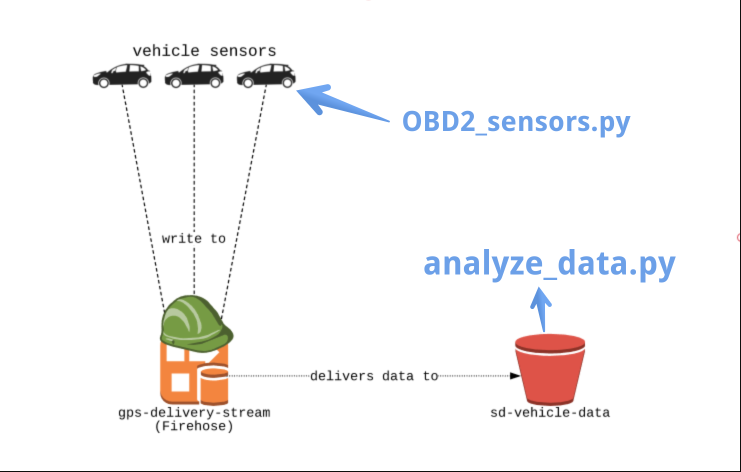
**Reading Firehose data**

In the past two exercises, you have created destination S3 bucket ("sd-vehicle-data") and a Firehose stream ("gps-delivery-stream"). You have written records to the stream which got delivered to the bucket.

Now it's time to read those records and do some basic analysis. If you're successful, you will be able to analyze top speeds reached by each vehicle. Having this information would prevent drivers from violating speed limits, and would likely save many lives.



**File to edit**

You will be editing the analyze\_data.py file in the editor.

**Slides**

Don't forget - you can always click on the **slides** icon in the top right to see the slides - they'll be pretty helpful!

**Instructions**

**100XP**

* List the objects that have been written to the S3 bucket.
* For every object, load it from S3.
* Load it into a dataframe, specifying a delimiter and column names.
* Execute analyze\_data.py by running python3 analyze\_data.py

Code:

# analyze\_data.py: Analyze written sensor data. EDIT HERE.

import \_setup, \_run\_deps, pandas as pd

firehose, s3, records = \_setup.ex\_vars

# List the objects that have been written to the S3 bucket

objects = s3.list\_objects(Bucket='sd-vehicle-data')['Contents']

# Create list for collecting dataframes from read files.

dfs = []

# For every object, load it from S3.

for obj in objects:

data\_file = s3.get\_object(Bucket='sd-vehicle-data', Key=obj['Key'])

# Load it into a dataframe, specifying a delimiter and column names

dfs.append(pd.read\_csv(data\_file['Body'],

delimiter = " ",

names=["record\_id", "timestamp", "vin", "lon", "lat", "speed"]))

# Concatenate the resulting dataframes.

data = pd.concat(dfs)

print(data.groupby(['vin'])['speed'].max())

repl:~/workspace$ cd /home/repl/workspace

repl:~/workspace$ python3 analyze\_data.py

vin

1FTEX1C8XAK855191 97

2LAXX1C8XAK844292 98

3FTEX1G5XAK844393 96

4FTEX4944AK844294 100

5FTEX1MAXAK844295 99

Name: speed, dtype: int64