Take Home Assignment

Introduction

Thank you for your interest in the Data Engineer role within the IT – AI/ML Team at SHI International Corp. As part of the interview process, we would like you to complete a take home assignment. After you have completed the assignment, we will review the deliverables and schedule a follow-up technical interview, if needed.

Scenario

You are a Data Engineer. You have been asked to create an ETL process to ingest new or updated product data from multiple sources. The product owner for this project has obtained sample datasets, described below. You have been told that unique pairs of Manufacturer and SKU identify unique product records, and that the ingested data will be heavily queried by business applications. What will you do?

Instructions

- Write a design specification document that summarizes the problem, assumptions, solution, alternative solutions, and open questions.
    - Design should address data pipeline, data storage, data integrity, scalability, and performance.
    - Include diagrams as needed.
- Create a repository on GitHub and write the code for your solution using standard coding practices.
    - Code should be written in Python.
    - Platforms and/or packages that you decide are fit for purpose may be used.
    - Coding assistants may not be used.
    - Include instructions for environment setup and running the code.
- Deliverables
    - Design specification document (PDF format)
    - Link to GitHub repo
    - Logs from successfully run code (PDF format)
    - Data summary report (PDF format)
        - Number of unique product records
        - Number of unique manufacturers
        - Number of unique categories
        - Number of unique price records
        - Number of unique distributors
        - Number of product records per manufacturer
        - Number of product records per category
        - Number of price records per distributor

Datasets

- products.json.gz: Product metadata from a third-party data curator's API. Updated hundreds of times per day.
- jillsjunk.csv.gz: Product price and quantity from distributor Jill's Junk. Jill's Junk is one of many distributors. Products can overlap among distributors. Updated dozens of times per day.
- samsstuff.csv.gz: Product price and quantity from distributor Sam's Stuff. Sam's Stuff is one of many distributors. Products can overlap among distributors. Updated dozens of times per day.

Sample from products.json.gz


```
[
  {
    "Category": "All Beauty",
    "Title": "Acca Kappa Professional Pro Hair Brush, Round, Boar Bristle/Nylon, Medium",
    "Details": {
      "Shape": "Round",
      "Unit Count": "1 Count",
      "Hair Type": "Drying Hair",
      "Item Weight": "0.08 Pounds",
      "Is Discontinued By Manufacturer": "No",
      "Product Dimensions": "6 x 5 x 4 inches; 1.33 Ounces",
      "Department": "Bath & Shower"
    },
    "Manufacturer": "Acca Kappa Professional",
    "SKU": "769898008207 885491870314 798813082411",
    "UpdatedOnUTC": "2020-08-07T23:20:35"
  },
  {
    "Category": "All Beauty",
    "Title": "YAYAFAIRY Ponytail Extension,Long Curly Ponytail Hair Extension Ash Blonde Mix Bleach Blonde Clip in Ponytail for
women",
    "Details": {
      "Color": "#Mixed Ash Blonde Mix Bleach Blonde",
      "Material": "Synthetic",
      "Extension Length": "22 Inches",
      "Style": "Mixing Color",
      "Package Dimensions": "10.83 x 6.69 x 0.87 inches; 3.53 Ounces"
    },
    "Manufacturer": "YAYAFAIRY",
    "SKU": "733424968716",
    "UpdatedOnUTC": "2023-12-10T23:22:05"
  },
  {
    "Category": "All Beauty",
    "Title": "SCENTW Short Afro Kinky Curly Synthetic Wigs For Black and african american Women Heat Resistant Full Party Wig with
Free Cap(Black)",
    "Details": {
      "Material": "Synthetic",
      "Hair Type": "Curly",
      "Style": "Curly",
      "Special Feature": "Heat Resistant",
      "Is Discontinued By Manufacturer": "No",
      "Package Dimensions": "9.06 x 6.69 x 2.17 inches; 9.92 Ounces"
    },
    "Manufacturer": "SCENTW",
    "SKU": "755320495417",
    "UpdatedOnUTC": "2022-04-17T12:51:38"
  },
  {
    "Category": "All Beauty",
    "Title": "Dorall Collection Lion Heart Eau de Toilette Spray for Men, 3.4 Ounce",
    "Details": {
      "Item Form": "Spray",
      "Item Volume": "3.4 Fluid Ounces",
      "Age Range (Description)": "Adult",
      "Style": "Modern",
      "Product Dimensions": "5 x 9 x 7 inches; 12 Pounds"
    },
    "Manufacturer": "Camrose Trading Inc. DBA Fragrance Express - DROPSHIP",
    "SKU": "843706025652",
    "UpdatedOnUTC": "2024-04-24T06:10:49"
  },
  {
    "Category": "All Beauty",
    "Title": "Angel Violet By Thierry Mugler For Women. Eau De Parfum Refill 1.7-Ounces",
    "Details": {
      "Is Discontinued By Manufacturer": "No",
      "Package Dimensions": "1.5 x 1.5 x 1.5 inches; 0.01 Ounces"
    },
    "Manufacturer": "Thierry Mugler",
    "SKU": "158271",
    "UpdatedOnUTC": "2023-05-07T21:12:00"
  }
]
```

Sample from jillsjunk.csv.gz

```
Manufacturer,SKU,Price,Quantity
Stehlen,714937183476,182.44476529754644,282
Old Varsity Brand,2608MICEA00,27.871807958555596,391
Cobella,US-B-0064,27.0128688616202,5
Jack Nicklaus,JNWF8027,20.448063682232306,156
Sperry Top Sider Mens Accessories,5H150,16.16519465681872,875
```

Sample from samsstuff.csv.gz

```
Manufacturer,SKU,Price,Quantity
Beewin,MHH077,21.759755738351785,3
Sea Gull Lighting,65061BLE-962,156.67449440002576,3
Alatino,AL095G,49.98610122662185,994
ROADFAR,103539-5231-1538273081,48.96344833627786,28
AUTEX,194741,31.43787152180632,233
```

Citation

This is a very small, slightly transformed, and slightly fabricated sample of data from https://amazon-reviews-2023.github.io/.

```
@article{hou2024bridging,
  title={Bridging Language and Items for Retrieval and Recommendation},
  author={Hou, Yupeng and Li, Jiacheng and He, Zhankui and Yan, An and Chen, Xiusi and McAuley, Julian},
  journal={arXiv preprint arXiv:2403.03952},
  year={2024}
}
```