Data Cleaning in Shell

A. Learning Objective

- Mengaplikasikan perintah bash untuk pemrosesan data
- Mengaplikasikan command csvkit untuk pemrosesan data
- Melakukan Cleaning Data di Shell
- Memanfaatkan Git Version Control sebagai repository

B. Background:

Anda diminta untuk membersihkan [data traffic ecommerce] di tempat anda bekerja. Data yang diberikan adalah data bulan oktober dan november tahun 2019. Atasan anda ingin melakukan analisis produk yang terjual. Anda diminta untuk membersihkan datanya, dan menyisakan data-data yang relevan. Berikut gambaran awal data yang diberikan:

a	event_time	event_type	product_id	category_id
8	2019-10-01 00:00:10	view	28,719,074	2,053,013,565,480,109,009
17	2019-10-01 00:00:18	view	10,900,029	2,053,013,555,069,845,885
91	2019-10-01 00:01:17	view	1,004,856	2,053,013,555,631,882,655
101	2019-10-01 00:01:29	view	1,801,551	2,053,013,554,415,534,427
103	2019-10-01 00:01:30	view	3,900,930	2,053,013,552,326,770,905
105	2019-10-01 00:01:30	view	1,005,115	2,053,013,555,631,882,655
113	2019-10-01 00:01:36	view	1,003,549	2,053,013,555,631,882,655
114	2019-10-01 00:01:36	view	1,004,720	2,053,013,555,631,882,655
133	2019-10-01 00:01:48	view	28,714,158	2,053,013,565,413,000,141

category_code	brand	price	user_id	user_session
apparel.shoes.keds	baden	102.71	520,571,932	ac1cd4e5-a3ce-4224-a2d7-ff660a105
appliances.kitchen.mixer	bosch	58.95	519,528,062	901b9e3c-3f8f-4147-a442-c25d5c5ed
electronics.smartphone	samsung	130.76	515,757,896	4938043e-e50f-44ad-944d-958d04df6
electronics.video.tv	haier	463.31	515,511,944	d63ef339-2d6a-411a-95bd-d58c77ae6
appliances.environment.water_heater	teploross	90.32	555,444,559	98b88fa0-d8fa-4b9d-8a71-3dd403afa
electronics.smartphone	apple	975.57	514,218,020	d7c4761f-de75-454b-9164-177db5e53
electronics.smartphone	samsung	344.64	546,089,729	796f2f69-c266-4202-a583-b87394db8
electronics.smartphone	huawei	130.65	555,357,251	a4184ca6-5dcb-40e3-b709-dde079c84
apparel.shoes.slipons	caprice	41.19	535,055,930	9f5785c3-2c85-4e0e-8173-0ae452772

Dari data di atas anda diminta:

1. **Menggabungkan kedua data** ke dalam 1 file.
2. **Menyeleksi kolom yang relevan** untuk analisis produk.
3. **Melakukan filtering** untuk mendapatkan **aktivitas pembelian saja**.
4. **Melakukan splitting data kategori produk dan nama product** pada kolom category code

   Note: category produk berada di posisi kata pertama dan nama produk di posisi kata terakhir dari kolom category_code

Ekspektasi hasil pengolahan:

event_time	event_type	product_id	category_id	brand	price	category	product_name
2019-10-01 02:20:11	purchase	4,804,055	2,053,013,554,658,804,075	apple	189.91	electronics	headphone
2019-10-01 02:23:05	purchase	4,802,157	2,053,013,554,658,804,075	jbl	20.57	electronics	headphone
2019-10-01 02:28:09	purchase	1,004,856	2,053,013,555,631,882,655	samsung	130.76	electronics	smartphone
2019-10-01 02:28:40	purchase	1,005,104	2,053,013,555,631,882,655	apple	975.57	electronics	smartphone
2019-10-01 02:29:32	purchase	8,800,448	2,053,013,555,573,162,395	nokia	43.66	electronics	telephone
2019-10-01 02:31:18	purchase	12,400,121	2,053,013,556,252,639,687	alteco	45.82	construction	drill
2019-10-01 02:36:12	purchase	1,004,833	2,053,013,555,631,882,655	samsung	174.76	electronics	smartphone
2019-10-01 02:37:40	purchase	1,005,105	2,053,013,555,631,882,655	apple	1,415.48	electronics	smartphone
2019-10-01 02:38:51	purchase	1,004,776	2,053,013,555,631,882,655	xiaomi	185.08	electronics	smartphone

Tools:

- Bash
- Csvkit
- Git

Panduan Pengerjaan:

1. Download data ke dalam folder kerja
2. Buatlah file bash yang dapat melakukan pemrosesan data seperti yang telah dijelaskan diatas.
3. Coba jalankan program pada terminal.
   ***Note**: ini sebagai contoh saja, silahkan buat perencanaan alur program masing-masing dengan alur yang dibuat sendiri*

C. Waktu Pengerjaan

Project ini dikerjakan selama 2 Minggu, deadline tugas dapat dilihat di keterangan tugas.

D. Submission:

- Buatlah sebuah **repository di github** anda
- Simpan hasil pengerjaan anda ke dalam repository tersebut
- Tambahkan deskripsi project dengan mengikuti template **README.md**:
  - Tujuan Pengerjaan Project
  - Detail / deskripsi task
    Contoh
    1. membuat syntax x, syntax x melakukan proses a
    2. dst
  - Cara Running / Penggunaan Program
  - Saran Perbaikan
- Salin tautan repository tersebut dan **inputkan ke dalam form submission**

E. Penilaian:

- Keberhasilan Pemrosesan Data (30)
  - Penggabungan Data
  - Penyeleksian Kolom
  - Filtering Data
  - Splitting Kolom

F. Forum Diskusi Project di Discourse

JIka Anda memiliki pertanyaan silahkan memanfaatkan forum diskusi project Shell dapat anda akses pada Discourse

G. Validasi Hasil

Hasil Word Count

```
faizal@MSI:/mnt/c/Users/faiza/Documents/shell/ecommerce behaviour/tugas_4$ cat data/data_clean.csv | wc
  71416  214246 6859389
```

Jalankan Kode Berikut:

cat data/data_clean.csv | grep electronics | grep smartphone| awk -F ',' '{print $5}'| sort | uniq -c | sort -nr

```
faizal@MSI:/mnt/c/Users/faiza/Documents/shell/ecommerce behaviour/tugas_4$ cat data/data_clean.csv | grep electronics | grep smartphone| awk -F ',' '{print
}'| sort | uniq -c | sort -nr
  18300 samsung
  14403 apple
   5104 xiaomi
   2698 huawei
   1511 oppo
    225 vivo
    206 meizu
     84 honor
     70 nokia
     66 sony
     60 oneplus
     56 tp-link
     31 prestigio
     19 zte
     18 inoi
     14 tecno
     14 lg
     10 doogee
      9 haier
      8 htc
      8 bq
      7 umi
      5 google
      3 nubia
      3 jinga
      3 gionee
      2 fly
      2 blackberry
      1 texet
```