



PRIME MINISTER'S DEPARTMENT
DEPARTMENT OF STATISTICS MALAYSIA

STATISTICAL, ECONOMIC AND SOCIAL
RESEARCH AND TRAINING CENTRE
FOR ISLAMIC COUNTRIES (SESRIC)
TRAINING COURSE ON BIG DATA
APPLICATIONS ON PRICE STATISTICS



20 OCT



2016 - 2030

PRICE INTELLIGENCE Data Management



16th JUNE 2022 (THURSDAY)



1430 - 1530 (MYT)



ONLINE



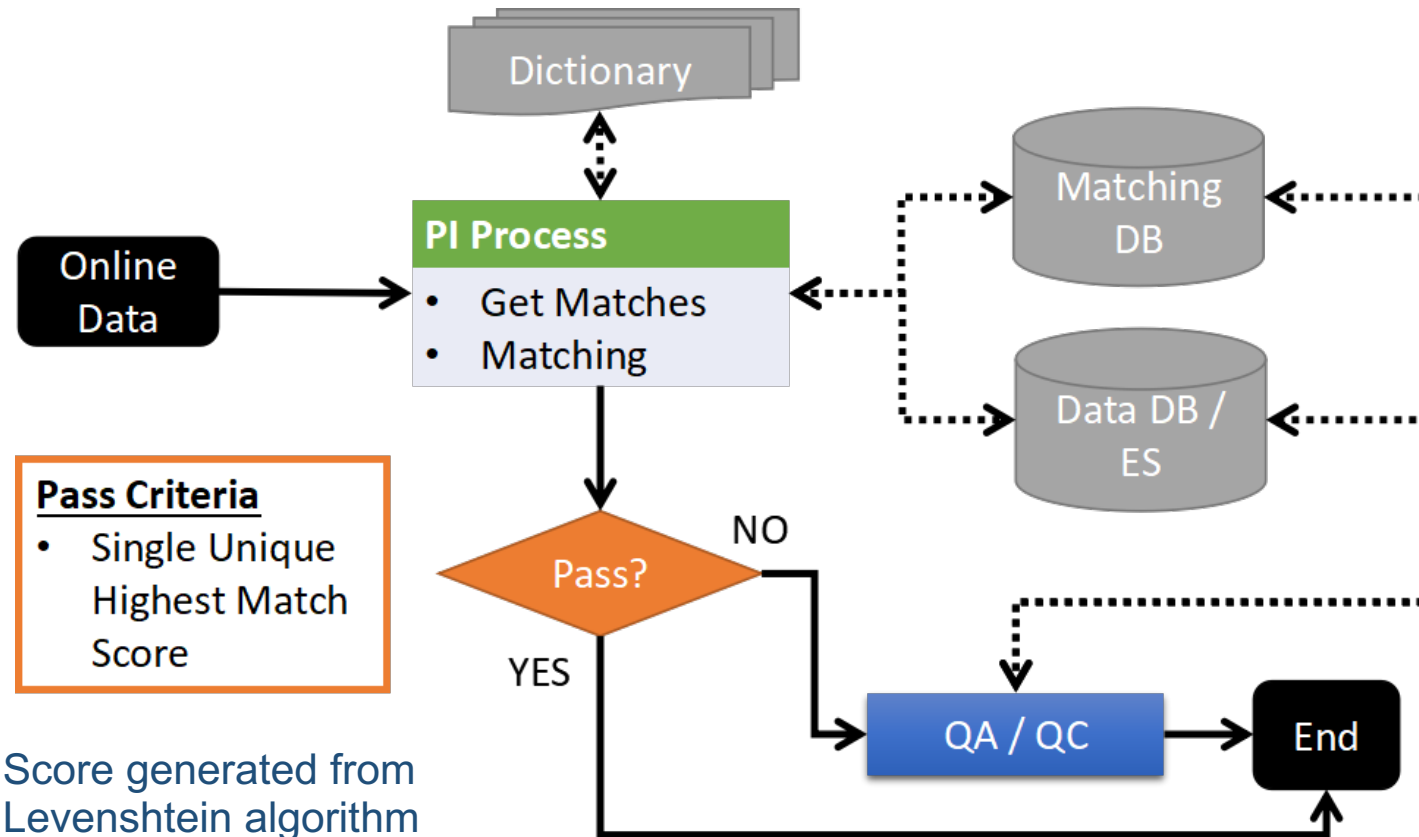
   
StatsMalaysia
www.DOSM.gov.my



Wan Ahmad Ridhuan Wan Jaafar
Department of Statistics Malaysia



General PI Architecture



1. Online Data

Online data crawled feed to PI Process

2. PI Process

Perform data matching process.

3. Matching DB

Store list of matched results for QA / QC

4. Data DB / ES

Store successful / accepted match data

5. QA / QC

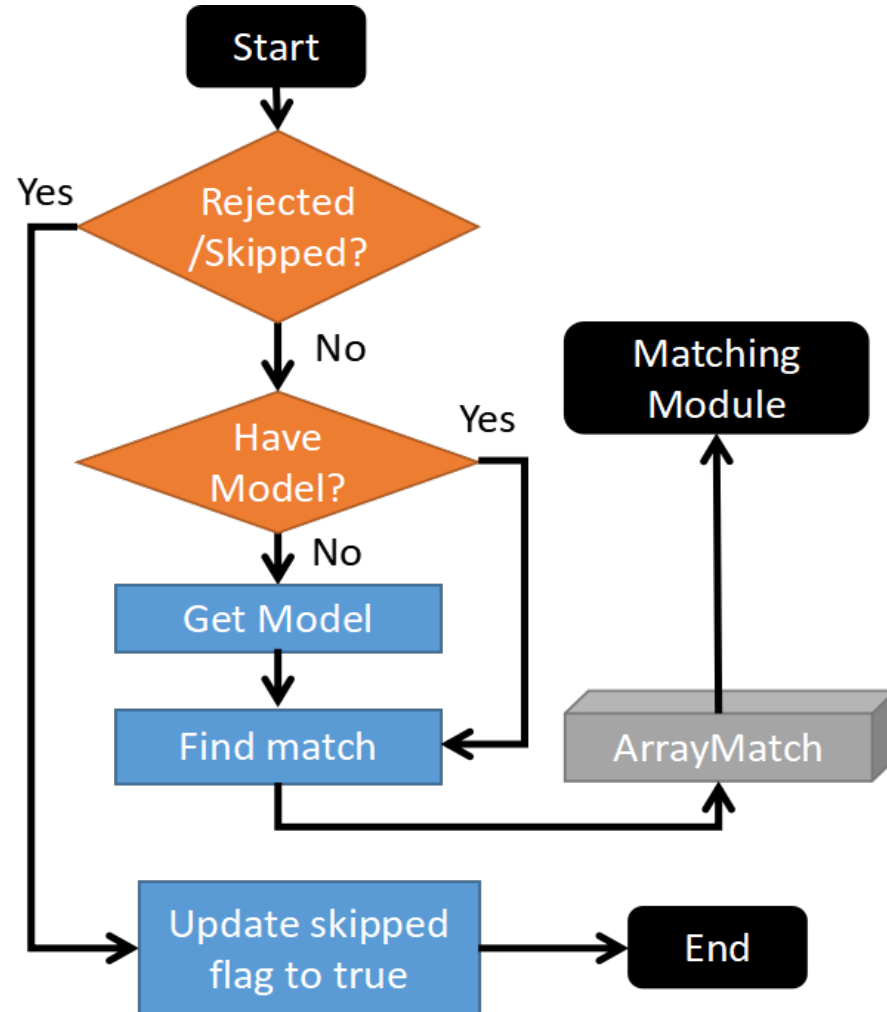
Post match verification process which require user action to determine match correctness and improve dictionary data



7 COICOP Categories are covered for PI exercise

PI Category	COICOP Category
Groceries	Food and Non-Alcoholic Beverages
Clothing	Clothing and Footwear
Appliances	Furnishing, Household Equipment and Routine Household Maintenance
Vehicles	Transport
Electronics	Communication Recreation Services and Culture
Wearables	Miscellaneous Goods and Services

GET MATCHES - VEHICLE



Matching process for vehicle are based on multiple input data fields

COICOP field to be populated	Input data field	Dictionary data field
COICOP4 COICOP6 category_group	● menu_type	● menu_type
	● make	● COICOP4Name
	● model	● COICOP6Name
	● condition	● condition
COICOP7	● engine_capacity	● engine_capacity
	● transmission	● transmission
	● variant	● variant

INPUT DATA FIELDS - VEHICLE

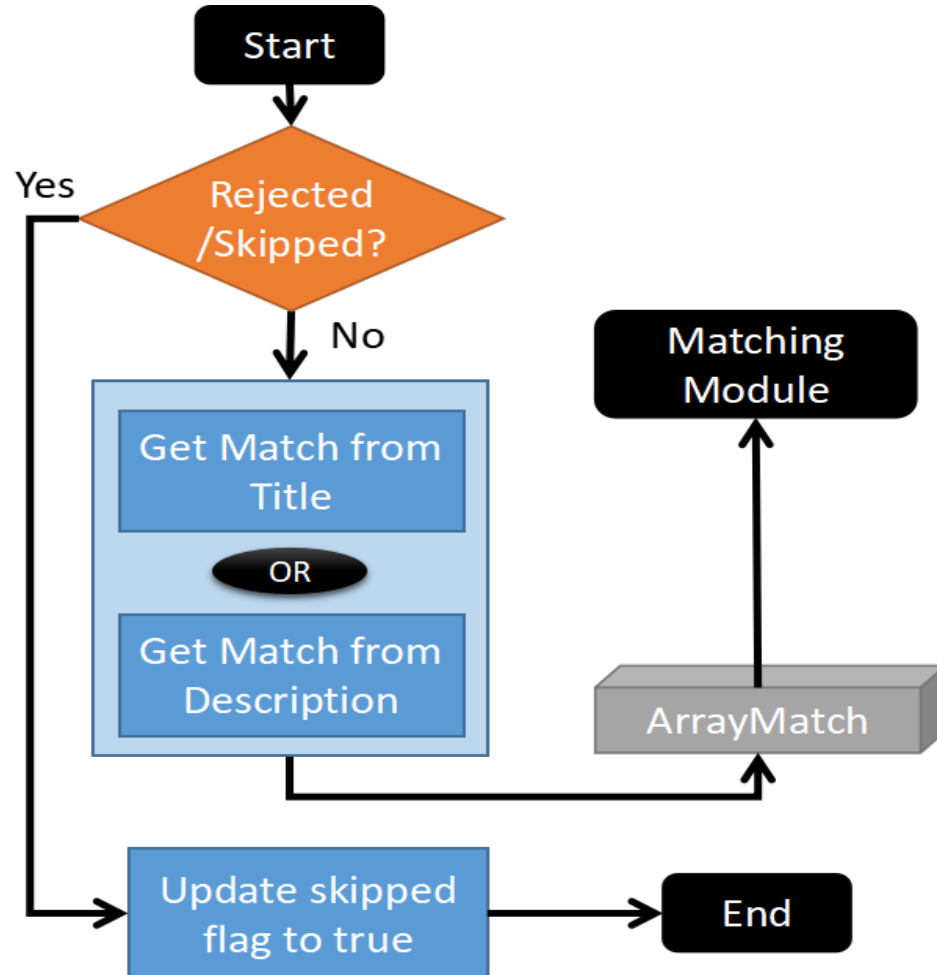


Accepted with exact match – given COICOP 7D code

coicop4Code	coicop4Name	coicop6Code	coicop6Name	origin	type	condition	menu_type	coicop7	engine_capacity	transmission	variant	category_group
0711	Proton	071101	Saga	National	Car	New	for sale	0711014	1300		CVT	L/D
0711	Proton	071101	Preve	National	Car	New	for sale	0711015	1600		CVT	L/D
0711	Proton	071101	Suprima S	National	Car	New	for sale					L/D
0711	Proton	071101	Iriz	National	Car	New	for sale					L/D
0711	Proton	071101	Persona	National	Car	New	for sale					L/D
0711	Proton	071101	Waja	National	Car	New	for sale					L/D
0711	Proton	071101	Satria	National	Car	New	for sale					L/D
0711	Proton	071101	Wira	National	Car	New	for sale					L/D
0711	Proton	071101	Iswara	National	Car	New	for sale					L/D
0711	Proton	071101	Perdana	National	Car	New	for sale					L/D
0711	Proton	071101	Satria Neo	National	Car	New	for sale					L/D
0711	Proton	071101	Gen2	National	Car	New	for sale					L/D
0711	Proton	071101	Savvy	National	Car	New	for sale					L/D
0711	Proton	071101	Inspira	National	Car	New	for sale					L/D
0711	Proton	071101	Satria GTI	National	Car	New	for sale					L/D
0711	Proton	071101	Juara	National	Car	New	for sale					L/D
0711	Proton	071101	Ertiga	National	Car	New	for sale					L/D
0711	Proton	071101	Putra	National	Car	New	for sale					L/D
0711	Proton	071101	Tiara	National	Car	New	for sale					L/D
0711	Proton	071105	Exora	National	MPV	New	for sale					L/D
0711	Proton	071197	Arena	National	Others		for sale					L/D
0711	Proton	071102	Saga	National	Car	Used	for sale	0711021	1300	Auto		L/D

Meet requirements of COICOP 6D but doesn't match at specification level (COICOP 7D)

GET MATCHES - CLOTHING

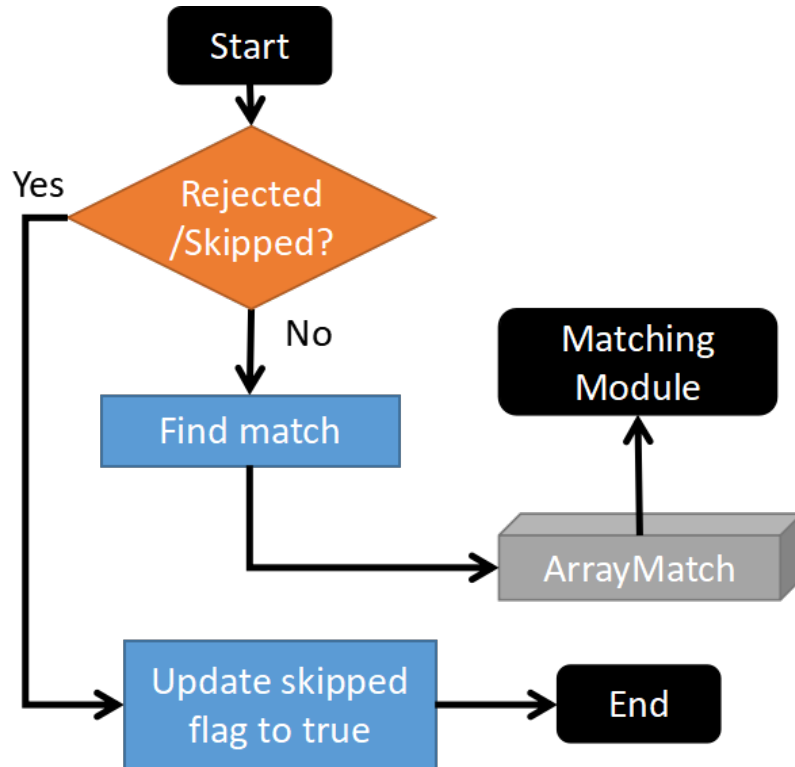


Matching process for clothing are based on multiple input data fields

COICOP field to be populated	Input data field	Dictionary data field
COICOP4	● Title	● COICOP6_Name
COICOP6	● Description	● Variation
category_group	● Condition	
COICOP7	● Title	● Criteria1
	● Description	● Criteria2
	● Gender	● Criteria3
		● Criteria4
		● Criteria5



Matching process for groceries are based on multiple input data fields



COICOP field to be populated	Input data field	Dictionary data field
COICOP4	● Title	● COICOP6_Name
COICOP6	● Description	● Variation
category_group		
COICOP7	● Title	● Criteria1
	● Description	● Criteria2
	● Brand	● Criteria3
	● Unit	● Criteria4
	● Variation	● Criteria5

INPUT DATA FIELDS - GROCERIES



coicop4	coicop4name	coicop6	coicop6name	Variation	coicop7	criteria1	criteria2	criteria3	criteria4
0131	Fresh Fish	013101	Kembong, Rumahan, Kembung		013101			1kg	
0131	Fresh Fish	013102	Tenggiri, Kow Yee		0131021	Batang	Potong, Cut, Hiris, Slice	1kg	
0131	Fresh Fish	013102	Tenggiri, Kow Yee		0131022	Papan		1kg	
0131	Fresh Fish	013102	Tenggiri, Kow Yee		0131023	Batang		1kg	
0131	Fresh Fish	013103	Cencaru, Ngam Mei Chee, Geronggong		013103			1kg	
0131	Fresh Fish	013104	Bawal Hitam, Duai Hitam		0131041			1kg	Local, Tempatan
0131	Fresh Fish	013104	Bawal Hitam, Duai Hitam		0131042			1kg	Thailand
0131	Fresh Fish	013107	Merah, Hoong Yee		0131071		Potong, Cut, Hiris, Slice	1kg	
0131	Fresh Fish	013107	Merah, Hoong Yee		0131072			1kg	
0131	Fresh Fish	013115	Pari, Stingray		013115			1kg	
0131	Fresh Fish	013118	Tongkol, Kayu, Aya, Simbak, Bakulan, Lengkumis		0131181	Putih, White		1kg	
0131	Fresh Fish	013118	Tongkol, Kayu, Aya, Simbak, Bakulan, Lengkumis		0131182	Hitam, Black		1kg	
0131	Fresh Fish	013124	Kerisi		013124			1kg	
0131	Fresh Fish	013128	Selayang, Basung		013128			1kg	
0131	Fresh Fish	013139	Siakap, Selongsong		0131391	Hitam, Black		1kg	
0131	Fresh Fish	013139	Siakap, Selongsong		0131392	Putih, White		1kg	



Rule based classification with keywords and variation of the products information determine level of MCOICOP code matching – either not match, partial match or exact match.

Keywords and variation consist of information such brand name and other information that makes the product meet requirement at COICOP 6D



PI Dictionary

☐ Only show Inactive

Classifier	4D Code	6D Code	Description	Sample Keyword	Sample Variation
wearables	1213	121321	Lampin pakai buang untuk bayi/ Diapers for baby	Baby Diapers Diapers Disposable Diapers Drypers Genki Goon Huggies Mamypoko Pampers Whoopee	Classic Diapers Drynights Night Pants Petpet Super Tape Ultra Wee Wee Dry
wearables	1213	121321	Lampin pakai buang untuk bayi/ Diapers for baby	Baby Diapers Disposable Diapers Genki Goon Huggies Lampin Pakai Buang Bayi Mamypoko Pampers Tesco Loves Whoopee	Baby Diapers Dry Drynights Drypant Extra Pants Tape Wee Wee Dry Wwd
wearables	1213	121322	Syampu/perapi/ Shampoo/conditioner	COND1 Clairel PANTENE Perapi SHIM SHIP SSK Sunsilk Chinese Herbal Repairing Syampu	

Edit data

General

Classifier *

Category Group *

4D Code *

4D Code Description *

4D Code Filter

6D Code Details

6D Code *

6D Code Description *

6D Code Keyword *

Variation 1

Variation 2

Variation 3

Item Code Details



 View 0102032 Data

General

Classifier

groceries

Category Group

TTL/ND - Tidak Tahan Lama / Non-Durable

4D Code

0102

4D Code Description

Tea, Cocoa, etc

4D Code Filter

NOT0102

6D Code details

6D Code

010203

6D Code Description

Serbuk teh/uncang/daun/ Tea dust/bags/leaves

6D Code Keyword

Tea, Teh, CAMOMILE, PU ER, BOH, LIPTON

Variation 1

Serbuk Teh, Tea Dust, , Daun Teh, Tea Leaves

Variation 2

NOT010203

Variation 3

Variation 3

Item Code Details

Item Code

0102032

Brand

Boh

Variant/Model/Type

Cameron Highlands

Unit of Measurement

weight100gram

Additional Criteria 1

Criteria 4

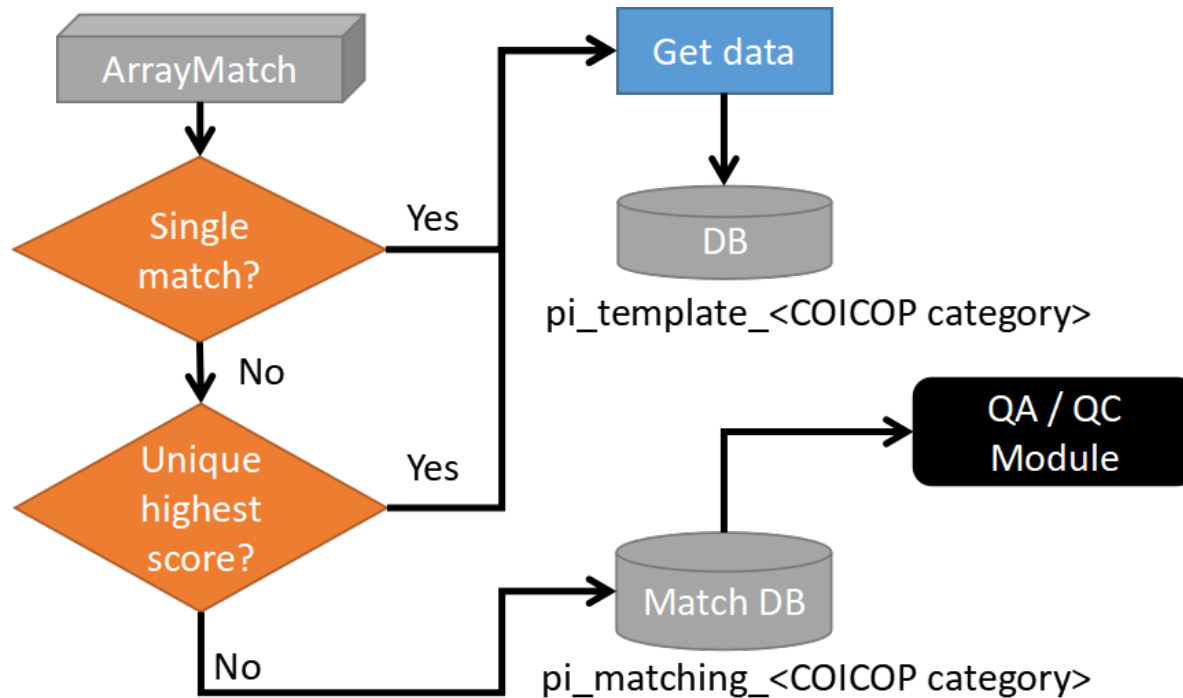
Additional Criteria 2

Criteria 5

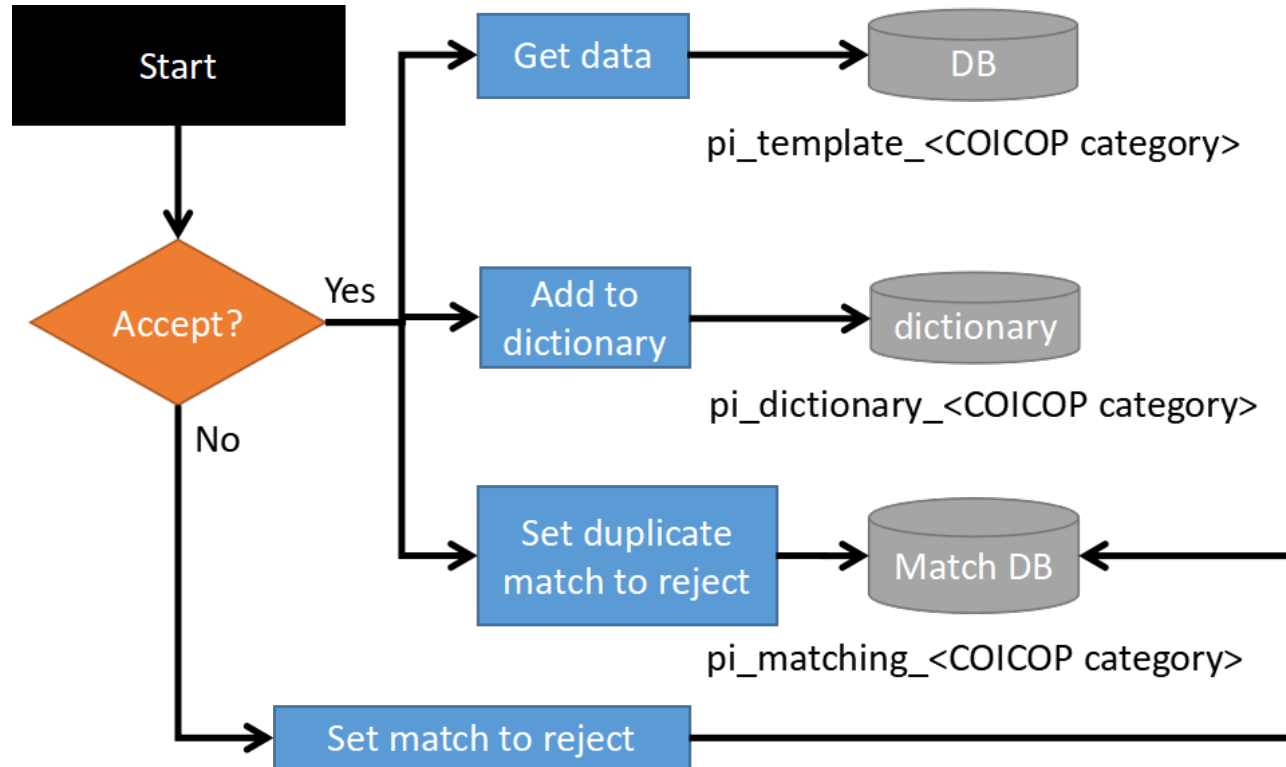
Specific information to be filled in the dictionary

- Brand
- Variant/Model/Type
- Unit of measurement
- Any additional criteria

Any result meet all the requirement above will be tag as exact match and filled with MCOICOP 7D



No	Process	Description
1	ArrayMatch	Temporary storage for matching results
2	Single Match?	Check if there is only 1 match for particular input data
3	Unique highest score?	Check for a single record with highest matching score if there are multiple matches for a single input data
4	Get data	Get remaining necessary dictionary data if only 1 match is passed from previous process
5	DB	Store successful data to db
6	Match DB	Store match results for QA and QC



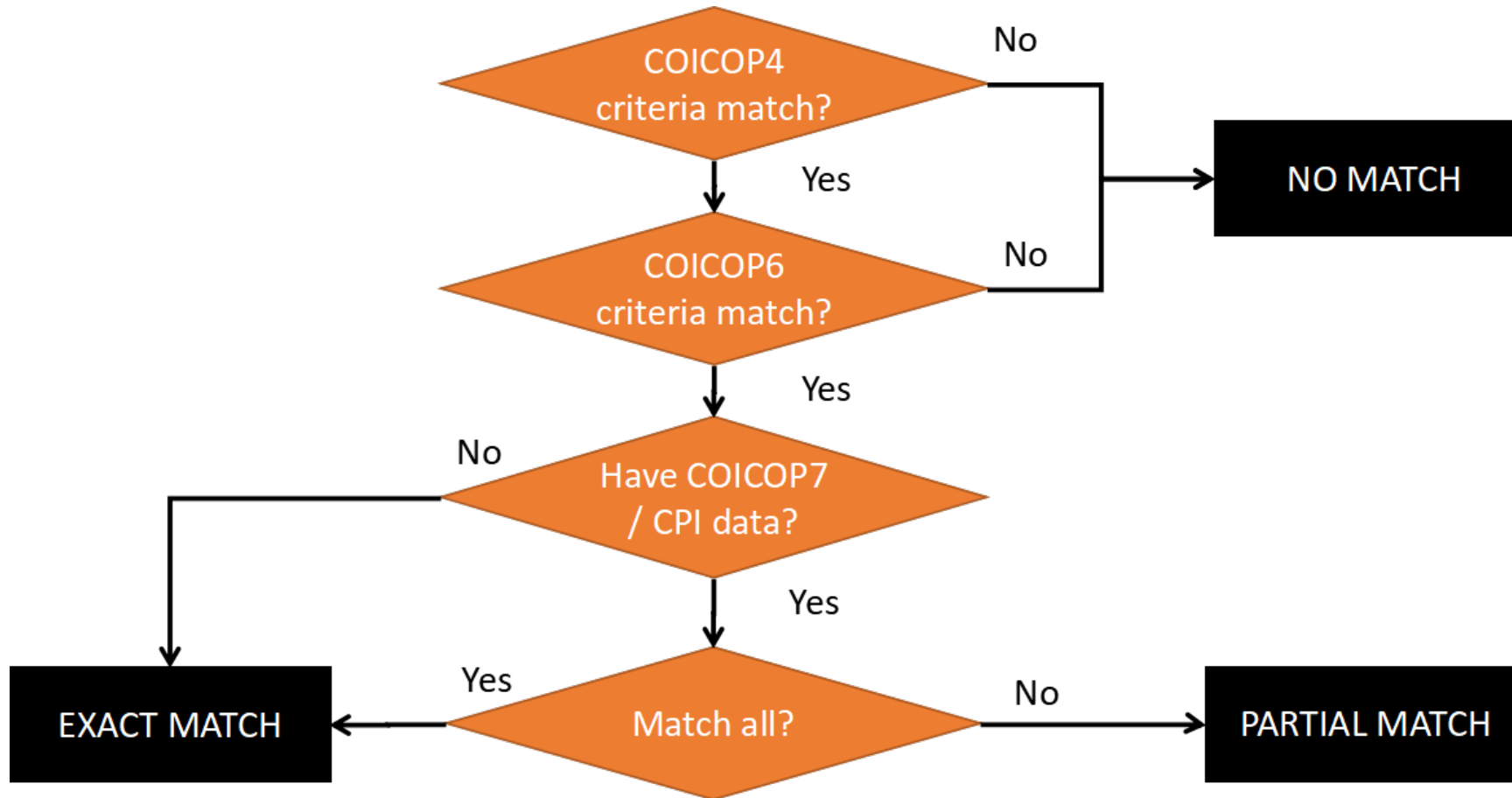
No	Process	Description
1	Accept	User decide on the correctness of the match
2a	Get Data	Store accepted match to DB and discard rejected match
2b	Add to dictionary	Add accepted match to dictionary for improvement
2c	Set duplicate match to reject	Set qaqc_flag of matched records with same input id to rejected and update to match DB
3	Set match to reject	Set qaqc_flag to rejected and update back to match DB



PI QAQC

☐	Source Id	Source	Source Data	Status
☐	000000bc3ebf153ee6b0	mudah	<i>i</i> huawei mate 7 original	Accepted
☐	00000126cc17485b9e38	mudah	<i>i</i> peplum blouse saloma undefined	Accepted
☐	000001a74e14baf47941	mudah	<i>i</i> panasonic washing machine 6 kg - very clean	Accepted
☐	000005f438a8a718b97a	mudah	<i>i</i> g-shock shirt undefined	Accepted
☐	00000f86d4088b306240	mudah	<i>i</i> camera ixus 145	Accepted
☐	000013a96bd3cae00575	mudah	<i>i</i> water filter cuckoo purifier city centre y9qc	Accepted
☐	00001484679e9368a9a2	mudah	<i>i</i> ikan bilis premium johas gred aa	Accepted
☐	000014c933343616cd14	mudah	<i>i</i> hardisk pc 160gb sata	Accepted

MATCHING STATUS



MATCHING STATUS



Offer



[Darlie Double Action Mint Fluoride Toothpaste 250g](#)

Now RM9.60, Before RM11.29, Save RM1.69
Promotion valid until 05/04/2017

[See more in this category >](#)

RM 9.60/each **RM 9.60**

Same specifications but
different unit quantity

Partially matched Items

Offer



[Darlie Double Action Mint Fluoride Toothpaste 175g](#)

Now RM7.47, Before RM8.79, Save RM1.32
Promotion valid until 05/04/2017

[See more in this category >](#)

RM 7.47/each **RM 7.47**

Same specifications
and unit quantity

Exact matched Items

DARLIE Darlie Enamel Protect Strong Mint 200G

Be the first to review this product

Brand: DARLIE | More Personal Care from DARLIE



- Strengthens tooth enamel, your smile's first line of defence
- Protects teeth by reducing decalcification
- Added fluoride to strengthen teeth and prevent cavities

SELLER PROMOTIONS

Use Code "10LAZWATSONS" to enjoy RM10 OFF with minimum spend of RM 50 on all products... [more](#)

RM 8.80
Before RM11.29
You save 22%

ADD TO CART

Different specifications
and unit quantity

Partially matched Items

Offer



[Gillette Foamy Lemon Lime Shaving Foam 175g](#)

Now RM11.60, Before RM13.65, Save RM2.05
Promotion valid until 05/04/2017

[See more in this category >](#)

RM 11.60/each **RM 11.60**

Item not included in CPI
Basket

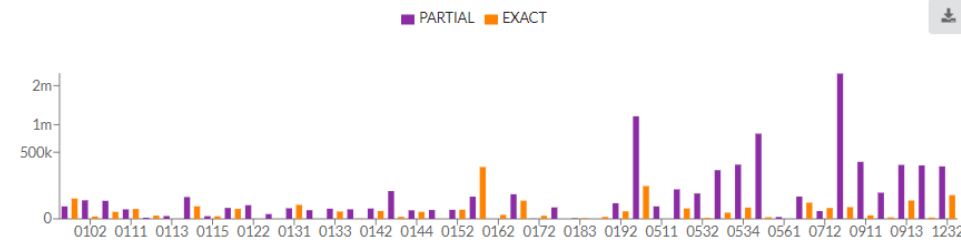
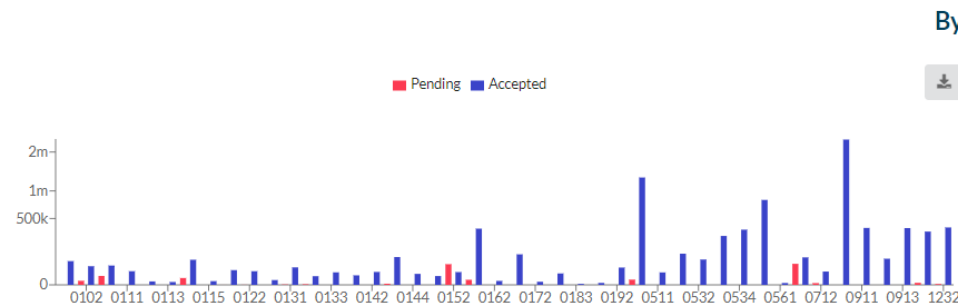
Non-matched Items

Each exact matched items and partially matched items from online price scraping, can be used as an additional information in CPI

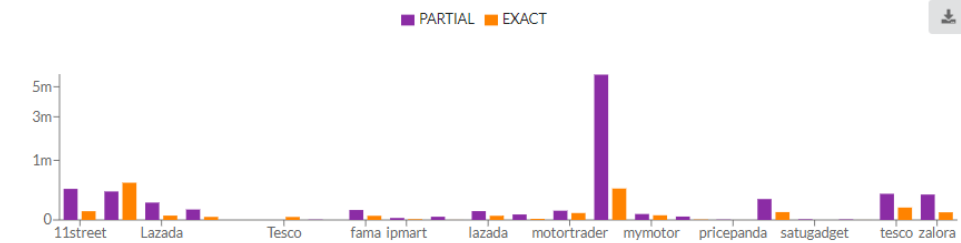
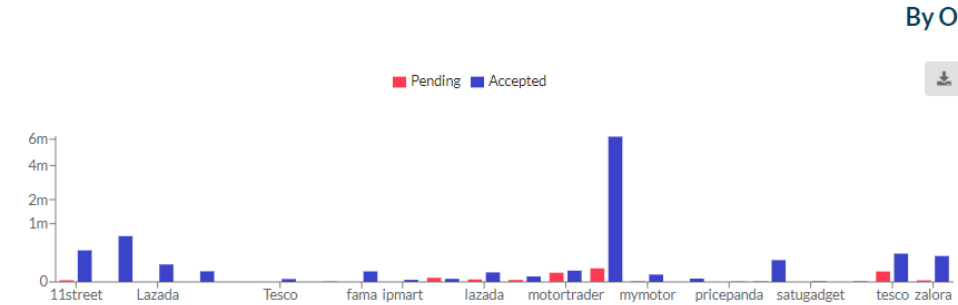
- As a price movement proxy and reference for current CPI price collection
- As a benchmark in determining new product specification in CPI basket of goods



PI Dictionary – Classification at 4 Digit level of COICOP



Pending/Accepted refers to the matching result at the 4 Digit COICOP level



Partially/exact refers to the matching result with MCOICOP

11.3% exact match
88.7% partially match

The word most often appear on the item description

highly mesin singapore japan peti blouse basuh galaxy myset apple korea oppo latest incredible newest price paradise products sony refrigerator sleeve size dispenser machine 5s 6s merah over
plus xiaomi note **iphone** swap cuckoo white purifier baju shirt items pro air quartz penapis watch panasonic water huawei redmi toshiba trends

*Information based on 231 item specifications updates in PI Dictionary

PI Dictionary – Text matching scoring

Specification that meet requirement of items in CPI basket of goods with high scoring value will return exact match, mapped to 7 digits item specification MCOICOP code

Item Specification	6D Code	Item Code	Score	Match Type
 huggies dry diapers m 6-11 kg 72 pieces each	121321	1213212	75.14	EXACT
 drypers wee wee dry disposable diapers m 6-11kg 74pcs each	121321	1213211	76.59	EXACT
 huggies dry diapers m 6-11 kg 72 pieces each	121321	1213212	75.14	EXACT
 huggies dry diapers m 6-11 kg 72 pieces each	121321	1213212	75.14	EXACT
 drypers wee wee dry disposable diapers m 6-11kg 74pcs each	121321	1213211	76.59	EXACT
 drypers wee wee dry disposable diapers m 6-11kg 74pcs each	121321	1213211	76.59	EXACT

Meanwhile, broad specification items has only 6 digits item code to be matched with online item specification. Items is considered matching as long as the item is the same regardless of brand name, units etc

Item Specification	6D Code	Item Code	Score	Match Type
 drypers wee wee dry disposable diapers xxi 15+kg 40pcs each	121321		67.12	PARTIAL
 huggies dry pants diapers l 9-14kg 50pcs each	121321		64.44	PARTIAL
 tena value diapers, medium 8 pack x 12s	121321		50.86	PARTIAL



PI Dictionary – Rule based classification



PI Dictionary

Search...

☐ Only show Inactive

Page Size

+

Classifier	4D Code ▲	6D Code	Description	Sample Keyword	Sample Variation	Item Code	Status
wearables	1213	121321	Lampin pakai buang untuk bayi/ Diapers for baby	Baby Diapers Diapers Disposable Diapers Drypers Genki Goon Huggies Mamypoko Pampers Whoopee	Classic Diapers Drynights Night Pants Petpet Super Tape Ultra Wee Wee Dry	1213211	✓
wearables	1213	121321	Lampin pakai buang untuk bayi/ Diapers for baby	Baby Diapers Disposable Diapers Genki Goon Huggies Lampin Pakai Buang Bayi Mamypoko Pampers Tesco Loves Whoopee	Baby Diapers Dry Drynights Drypant Extra Pants Tape Wee Wee Dry Wwd	1213212	✓
wearables	1213	121322	Syampu/perapi/ Shampoo/conditioner	CONDI Clairol PANTENE Perapi SHM SHP SSK Sunsilk Chinese Herbal Repairing Syampu		1213223	✓

PI Dictionary has been built in order to get the best match to MCOICOP 6/7 digits based on description of the items using text matching scoring

Dictionary will be updated regularly based on basket of goods in CPI



1. Incomplete information

levis checked western usa
levis western w vintage shirt
levis big e orange tab

1. Title field only have brand name without item type.

Process will need to get item type from description field of the record

jules shirt - new-original-jm02
wrangler western shirt
gap shirt blue black size m
t-shirt asics size l
tshirt custom kawasaki ninja zx6r - b24s

2. Title field does not contain sufficient information to conclude the correct item type.

A list of matches is created first before deciding on the correct match. Dictionary is populated with multiple variation of keyword / tag e.g tshirt and t-shirt.



2. Unneeded information

title	price	month	url	description
carhartt	80		http://www	Carhartt PantMade In USAAll Sizes Are AvailablePrice can be negotiable for serious buyers. I have a bulk quantity of Leather Jackets, Branded Shirts, Leather Bags, Branded Bags, Levis Ori Jeans,vintage t-shirts, leather pant etc. My prices are really reasonable. If you want you can also visit my warehouse or i can give you free delivery (only for Selangor and KL area) The rest i can deliver it via Poslaju. Those who are really interested feel free to contact me at:0060176229313, 011391533531wari356@gmail.com

Item owner use the description field to advertise other products available in their store such as Jacket, Shirt, Jeans, and Pants.

a match score and match status is introduced to help determine the correct match for the record. An additional QA/QC module will allow user to select the correct match

DATA SET GENERATOR



1

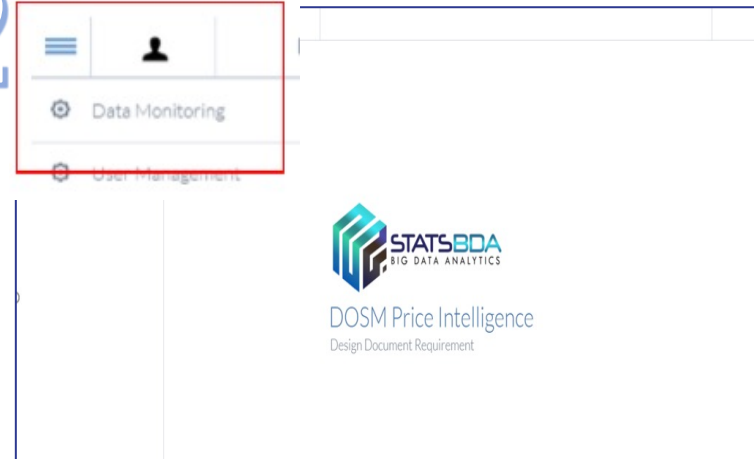
DOSM Price Intelligence

username

password

Login

2



3

PRICE INTELLIGENCE

CSV Generator

Source Category: ECOMMERCE

Source: Tesco

Item:

Item is Required

Input

Field Filter: Title

Time Range From:

From date is required.

To:

To date is required.

☐ Daily Update

4

PRICE INTELLIGENCE

CSV Impala Upload

DirectD

Fama

Iproperty

KL Sogo

Lazada

MAS

MASApr2019

MASApr2020_D11

MASApr2020_D12

MASApr2020_D20

MASApr2020_D30

MASDec2019

MASFeb2019

MASFeb2020

MASJan2019

MASJan2020

MASJuly2019

Generate New Data Create Folder

Data Properties

Title: MASApr2020_D30

Date Modified: 2020-05-01 10:14:34.0

Size: 0.20776844024658203

Status: 100%

Path: ~\hdfs\csv\MAS\MASApr2020_D30\MASApr2020_D30.csv

100%

Delete Download

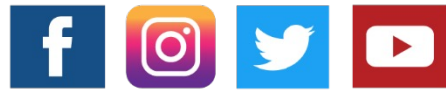
5

1	dates	site	flight_class	date_flight	flight_code	arrive	price	from	depart	dest
2	20200413	malaysiaairlines	business_smart	20200420	mh2	5:25	16,747	kuala lum	23:00	london lhr
3	20200413	malaysiaairlines	economy_flex	20200420	mh2	5:25	6,196	kuala lum	23:00	london lhr
4	20200413	malaysiaairlines	economy_promo	20200420	mh2	5:25	0	kuala lum	23:00	london lhr
5	20200413	malaysiaairlines	economy_smart	20200420	mh2	5:25	0	kuala lum	23:00	london lhr
6	20200413	malaysiaairlines	business_smart	20200420	mh3010	10:00	0	kota kinat	6:10	tawau twu
7	20200413	malaysiaairlines	economy_flex	20200420	mh3010	10:00	348	kota kinat	6:10	tawau twu
8	20200413	malaysiaairlines	economy_promo	20200420	mh3010	10:00	0	kota kinat	6:10	tawau twu
9	20200413	malaysiaairlines	economy_smart	20200420	mh3010	10:00	0	kota kinat	6:10	tawau twu
10	20200413	malaysiaairlines	business_smart	20200420	mh3012	14:50	0	kota kinat	12:00	tawau twu
11	20200413	malaysiaairlines	economy_flex	20200420	mh3012	14:50	503	kota kinat	12:00	tawau twu
12	20200413	malaysiaairlines	economy_promo	20200420	mh3012	14:50	0	kota kinat	12:00	tawau twu
13	20200413	malaysiaairlines	economy_smart	20200420	mh3012	14:50	0	kota kinat	12:00	tawau twu
14	20200413	malaysiaairlines	business_smart	20200420	mh3014	15:05	0	kota kinat	9:30	tawau twu
15	20200413	malaysiaairlines	economy_flex	20200420	mh3014	15:05	348	kota kinat	9:30	tawau twu
16	20200413	malaysiaairlines	economy_promo	20200420	mh3014	15:05	0	kota kinat	9:30	tawau twu
17	20200413	malaysiaairlines	economy_smart	20200420	mh3014	15:05	0	kota kinat	9:30	tawau twu
18	20200413	malaysiaairlines	business_smart	20200420	mh3016	15:05	0	kota kinat	12:15	tawau twu
19	20200413	malaysiaairlines	economy_flex	20200420	mh3016	15:05	348	kota kinat	12:15	tawau twu
20	20200413	malaysiaairlines	economy_promo	20200420	mh3016	15:05	0	kota kinat	12:15	tawau twu
21	20200413	malaysiaairlines	economy_smart	20200420	mh3016	15:05	0	kota kinat	12:15	tawau twu
22	20200413	malaysiaairlines	business_smart	20200420	mh3041	7:05	0	kota kinat	6:30	labuan ibu
23	20200413	malaysiaairlines	economy_flex	20200420	mh3041	7:05	190	kota kinat	6:30	labuan ibu
24	20200413	malaysiaairlines	economy_promo	20200420	mh3041	7:05	0	kota kinat	6:30	labuan ibu
25	20200413	malaysiaairlines	economy_smart	20200420	mh3041	7:05	0	kota kinat	6:30	labuan ibu
26	20200413	malaysiaairlines	business_smart	20200421	mh3042	7:40	0	labuan ibu	7:05	kota kinabalu bki
27	20200413	malaysiaairlines	economy_flex	20200421	mh3042	7:40	190	labuan ibu	7:05	kota kinabalu bki
28	20200413	malaysiaairlines	economy_promo	20200421	mh3042	7:40	0	labuan ibu	7:05	kota kinabalu bki
29	20200413	malaysiaairlines	economy_smart	20200421	mh3042	7:40	0	labuan ibu	7:05	kota kinabalu bki
30	20200413	malaysiaairlines	business_smart	20200420	mh3043	9:00	0	kota kinat	8:25	labuan ibu
31	20200413	malaysiaairlines	economy_flex	20200420	mh3043	9:00	190	kota kinat	8:25	labuan ibu
32	20200413	malaysiaairlines	economy_promo	20200420	mh3043	9:00	0	kota kinat	8:25	labuan ibu
33	20200413	malaysiaairlines	economy_smart	20200420	mh3043	9:00	0	kota kinat	8:25	labuan ibu
34	20200413	malaysiaairlines	business_smart	20200421	mh3044	16:10	0	labuan ibu	15:35	kota kinabalu bki
35	20200413	malaysiaairlines	economy_flex	20200421	mh3044	16:10	190	labuan ibu	15:35	kota kinabalu bki
36	20200413	malaysiaairlines	economy_promo	20200421	mh3044	16:10	0	labuan ibu	15:35	kota kinabalu bki
37	20200413	malaysiaairlines	economy_smart	20200421	mh3044	16:10	0	labuan ibu	15:35	kota kinabalu bki

Download Data

1. Click Data Monitoring on the top left of the screen
2. Click Folder Name and Click file name
3. Status of the generated data will be displayed.
Click Download to download to local PC

THANK YOU



StatsMalaysia

www.DOSM.gov.my



2022



2012-2022



20 OKT



2016-2030

