SALFORD & CO.

HOME ABOUT US SERVICES TEST DRIVE CONTACT

web crawlers using carlist.my

Group B - Data Drillers

THEVAN RAJU A/L JEGANATH (A22EC0286)
MULYANI BINTI SARIPUDDIN (A22EC0223)
ALIATUL IZZAH BINTI JASMAN (A22EC0136)
MUHAMMAD ANAS BIN MOHD PIKRI (A21SC0464)

INTRODUCTION

PROJECT BACKGROUND

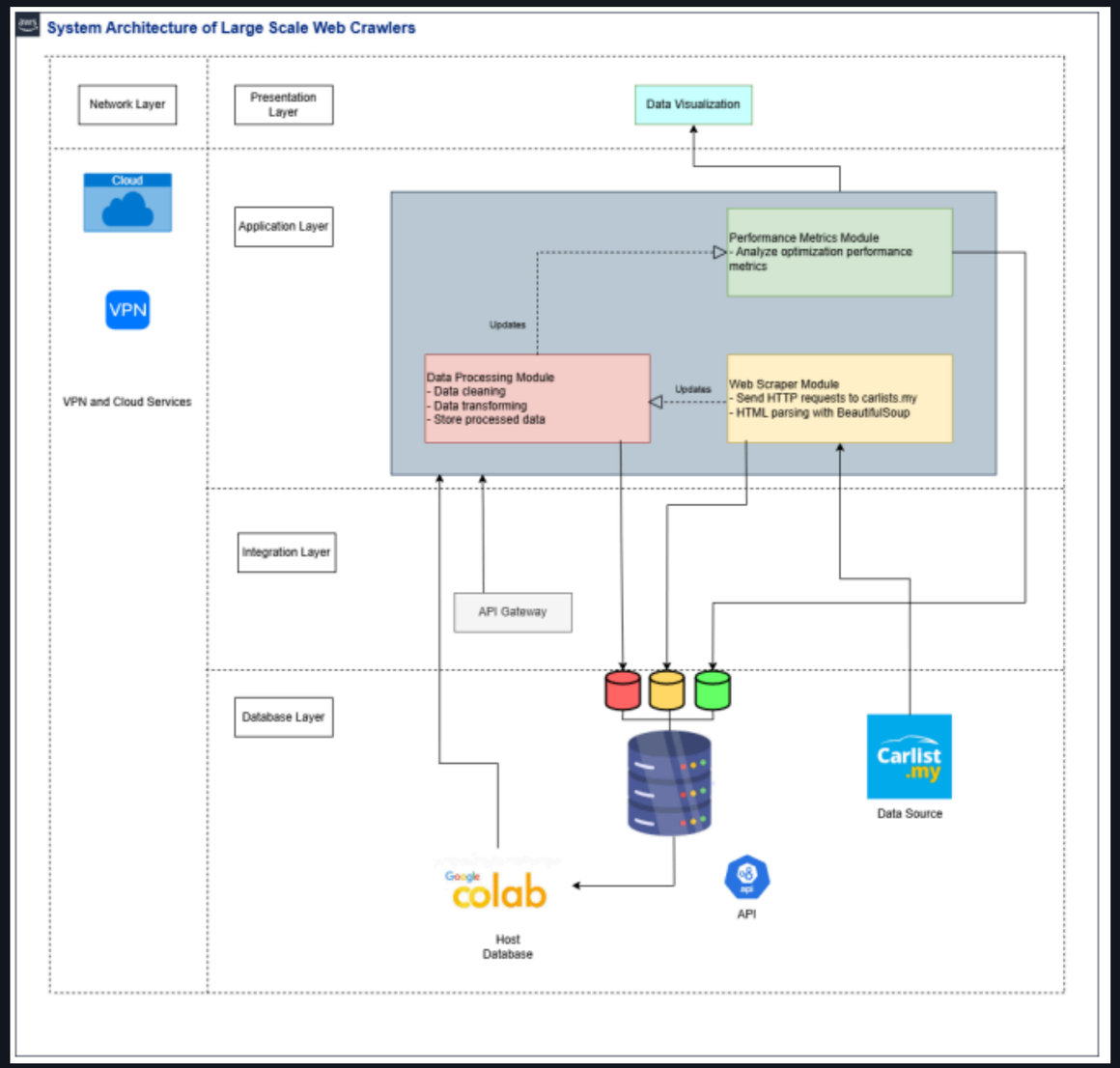
This project will address the application of High-Performance Computing (HPC) technologies to optimise information gathering from Carlists.my to learn and apply practical skills in multithreading, multiprocessing, optimising and distributed computing

PROJECT OBJECTIVES

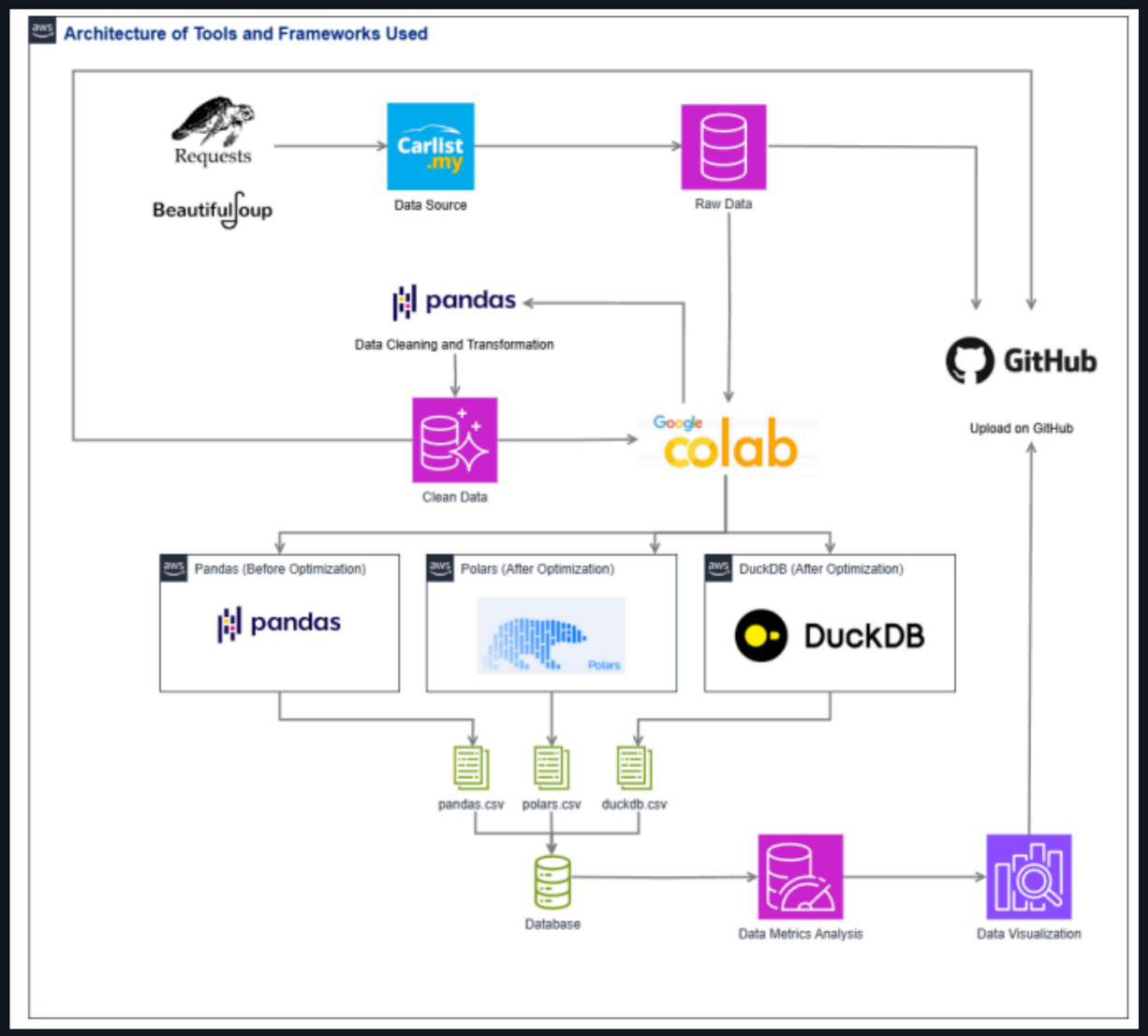
Data Collection
Data Processing
Performance Optimisation
Evaluation and Metrics
Comparison

SYSTEM DESIGN AND ARCHITECTURE

SYSTEM ARCHITECTURE



TOOLS AND FRAMEWORKS



TOOLS AND FRAMEWORKS USED

Google colab

Beautiful Soup

pandas

DuckDB

Polars

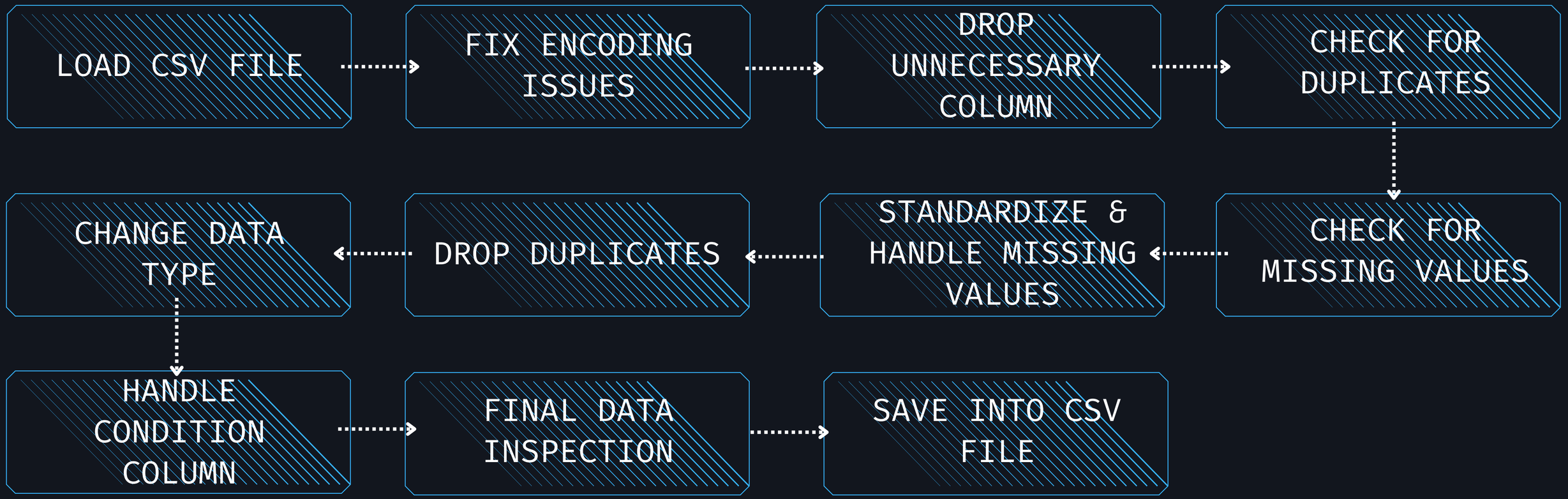
DATA COLLECTION & ENTITIES

Data	Description
car_name	The car's title or name on the listing is defensible
price	Asking price for the car
location	City/ locality where the car is sold
region	Malaysian state or territory
brand	Car manufacturer
model	Car model type
year	Manufacturing year
mileage	Driven distance specifically in kilometres (km)
fuel_type	Fuel type, such as petrol or diesel
color	The body colour of the car
body_type	Car type, such as hatchback, sedan, etc
seating_capacity	The seating number in the car
condition	Vehicle condition, such as used, new, etc
image	Vehicle image link to identify the car
description	Seller's description of the car
url	Full list link.

RAW DATA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Car Name	Price (MYR)	Currency	Location	Region	Brand	Model	Year	Mileage	Fuel Type	Color	Body Type	Seating Capacity	Condition	Image	Descriptio	URL	
2	Toyota Vellfire	420000	MYR	Subang Jaya	Selangor	Toyota	Vellfire	2023	2500	Petrol - Ur	Black	MPV	7	RefurbishedCo	https://im	2023 TOYO	https://www.carlist.	
3	Mazda CX-5	36900	MYR	Seri Kembangan	Selangor	Mazda	CX-5	2016	82500	Diesel	Grey	SUV	5	UsedConditior	https://im	2016 MAZI	https://www.carlist.	
4	MINI 3 Door	152000	MYR	TTDI	Kuala Lumpur	MINI	3 Door	2020	17500	Petrol - Ur	Black	Hatchback	4	RefurbishedCo	https://im	ðŸ˜ŽðŸ˜Ž A	https://www.carlist.	
5	Nissan X-Trail	59800	MYR	Johor Bahru	Johor	Nissan	X-Trail	2020	52500	Hybrid	White	SUV	5	UsedConditior	https://im	WhatsApp	https://www.carlist.	
6	Ford Focus	13900	MYR	Seri Kembangan	Selangor	Ford	Focus	2013	132500	Petrol - Ur	White	Sedan	5	UsedConditior	https://im	2013 FORD	https://www.carlist.	
7	Lexus RX300	261000	MYR	Subang Jaya	Selangor	Lexus	RX300	2021	17500	Petrol - Ur	White	SUV	5	RefurbishedCo	https://im	2021 LEXU	https://www.carlist.	
8	Audi A8	46900	MYR	Seri Kembangan	Selangor	Audi	A8	2013	97500	Petrol - Ur	Black	Sedan	4	UsedConditior	https://im	2013 AUDI	https://www.carlist.	
9	Land Rover Defender	419000	MYR	Subang Jaya	Selangor	Land Rove	Defender	2023	7500	Petrol - Ur	Bronze	SUV	5	RefurbishedCo	https://im	2023 LANI	https://www.carlist.	
10	Citroen Grand C4 SpaceTourer	60900	MYR	Seri Kembangan	Selangor	Citroen	Grand C4 S	2020	77500	Petrol - Ur	White	MPV	7	UsedConditior	https://im	2020 CITR	https://www.carlist.	
11	Audi Q7	49900	MYR	Seri Kembangan	Selangor	Audi	Q7	2014	112500	Petrol - Ur	White	SUV	7	UsedConditior	https://im	2014 AUDI	https://www.carlist.	
12	Porsche Cayenne	69900	MYR	Seri Kembangan	Selangor	Porsche	Cayenne	2011	127500	Diesel	Silver	SUV	5	UsedConditior	https://im	2011 PORS	https://www.carlist.	
13	Land Rover Defender	379000	MYR	Subang Jaya	Selangor	Land Rove	Defender	2022	17500	Petrol - Ur	Blue	SUV	7	RefurbishedCo	https://im	LAND ROV	https://www.carlist.	
14	Mazda Biante	60900	MYR	Seri Kembangan	Selangor	Mazda	Biante	2017	62500	Petrol - Ur	Silver	MPV	8	UsedConditior	https://im	2017 MAZI	https://www.carlist.	
15	Mercedes-Benz C200	188000	MYR	Ampang	Selangor	Mercedes-	C200	2020	12500	Petrol - Ur	Black	Sedan	5	RefurbishedCo	https://im	MODEL: M	https://www.carlist.	
16	Toyota Harrier	203000	MYR	Old Klang Road	Kuala Lumpur	Toyota	Harrier	2023	12500	Petrol - Ur	White	SUV	5	RefurbishedCo	https://im	2023 Toyo	https://www.carlist.	
17	Porsche Panamera	518000	MYR	Old Klang Road	Kuala Lumpur	Porsche	Panamera	2019	32500	Petrol - Ur	Silver	Hatchback	4	RefurbishedCo	https://im	2019 Porsc	https://www.carlist.	
18	Toyota Alphard	351000	MYR	Johor Bahru	Johor	Toyota	Alphard	2024	2500	Petrol - Ur	White	MPV	6	RefurbishedCo	https://im	* GENUIN	https://www.carlist.	
19	Lexus NX250	283000	MYR	Old Klang Road	Kuala Lumpur	Lexus	NX250	2023	12500	Petrol - Ur	White	SUV	5	RefurbishedCo	https://im	2023 Lexu	https://www.carlist.	
20	Mercedes-Benz G350	578000	MYR	Old Klang Road	Kuala Lumpur	Mercedes-	G350	2021	37500	Diesel	Black	SUV	5	RefurbishedCo	https://im	2021 Merc	https://www.carlist.	
21	Lexus RX300	261000	MYR	Subang Jaya	Selangor	Lexus	RX300	2021	27500	Petrol - Ur	Black	SUV	5	RefurbishedCo	https://im	2021 LEXU	https://www.carlist.	
22	Toyota Alphard	323000	MYR	Old Klang Road	Kuala Lumpur	Toyota	Alphard	2023	27500	Petrol - Ur	White	MPV	7	RefurbishedCo	https://im	2023 Toyo	https://www.carlist.	
23	MINI Clubman	213000	MYR	Old Klang Road	Kuala Lumpur	MINI	Clubman	2022	17500	Petrol - Ur	White	Wagon	5	RefurbishedCo	https://im	2022 MINI	https://www.carlist.	
24	Toyota Vellfire	395000	MYR	Old Klang Road	Kuala Lumpur	Toyota	Vellfire	2023	22500	Petrol - Ur	White	MPV	7	RefurbishedCo	https://im	2023 Toyo	https://www.carlist.	
25	Lexus IS300	237000	MYR	Old Klang Road	Kuala Lumpur	Lexus	IS300	2021	22500	Petrol - Ur	White	Sedan	5	RefurbishedCo	https://im	2021 Lexu	https://www.carlist.	
26	Volvo XC60	128900	MYR	Seri Kembangan	Selangor	Volvo	XC60	2020	102500	Petrol - Ur	White	SUV	5	UsedConditior	https://im	2020 VOLV	https://www.carlist.	
27	Mitsubishi Triton	19800	MYR	Kajang	Selangor	Mitsubishi	Triton	2010	152500	Diesel	White	Pickup Truc	5	UsedConditior	https://im	(WELCOM	https://www.carlist.	

DATA CLEANING



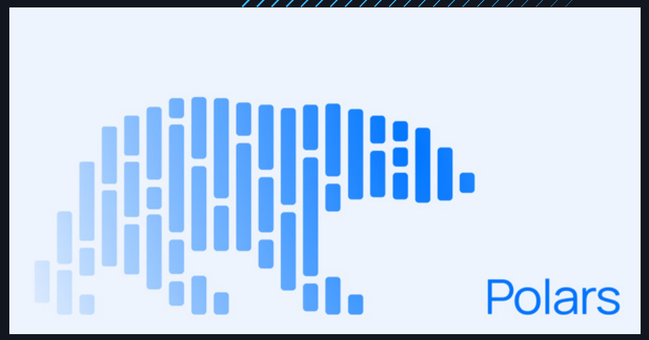
CLEANED DATA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	car name	price (myr)	location	region	brand	model	year	mileage	fuel type	color	body type	seating capacity	condition	image	description	url		
2	toyota vellfire	420000	subang jaya	selangor	toyota	vellfire	2023	2500	petrol - un	black	mpv	7	refurbishe	https://im	2023 toyo	https://www.carlist.my/cars-for-		
3	mazda cx-5	36900	seri kembangan	selangor	mazda	cx-5	2016	82500	diesel	grey	suv	5	used	https://im	2016 mazc	https://www.carlist.my/cars-for-		
4	mini 3 door	152000	ttdi	kuala lumpur	mini	3 door	2020	17500	petrol - un	black	hatchback	4	refurbishe	https://im	ðŸ˜Žð	https://www.carlist.my/cars-for-		
5	nissan x-trail	59800	johor bahru	johor	nissan	x-trail	2020	52500	hybrid	white	suv	5	used	https://im	whatsapp	https://www.carlist.my/cars-for-		
6	ford focus	13900	seri kembangan	selangor	ford	focus	2013	132500	petrol - un	white	sedan	5	used	https://im	2013 ford	https://www.carlist.my/cars-for-		
7	lexus rx300	261000	subang jaya	selangor	lexus	rx300	2021	17500	petrol - un	white	suv	5	refurbishe	https://im	2021 lexus	https://www.carlist.my/cars-for-		
8	audi a8	46900	seri kembangan	selangor	audi	a8	2013	97500	petrol - un	black	sedan	4	used	https://im	2013 audi	https://www.carlist.my/cars-for-		
9	land rover defender	419000	subang jaya	selangor	land rover	defender	2023	7500	petrol - un	bronze	suv	5	refurbishe	https://im	2023 land	https://www.carlist.my/cars-for-		
10	citroen grand c4 spacetourer	60900	seri kembangan	selangor	citroen	grand c4 sp	2020	77500	petrol - un	white	mpv	7	used	https://im	2020 citro	https://www.carlist.my/cars-for-		
11	audi q7	49900	seri kembangan	selangor	audi	q7	2014	112500	petrol - un	white	suv	7	used	https://im	2014 audi	https://www.carlist.my/cars-for-		
12	porsche cayenne	69900	seri kembangan	selangor	porsche	cayenne	2011	127500	diesel	silver	suv	5	used	https://im	2011 porsc	https://www.carlist.my/cars-for-		
13	land rover defender	379000	subang jaya	selangor	land rover	defender	2022	17500	petrol - un	blue	suv	7	refurbishe	https://im	land rover	https://www.carlist.my/cars-for-		
14	mazda biante	60900	seri kembangan	selangor	mazda	biante	2017	62500	petrol - un	silver	mpv	8	used	https://im	2017 mazc	https://www.carlist.my/cars-for-		
15	mercedes-benz c200	188000	ampang	selangor	mercedes-	c200	2020	12500	petrol - un	black	sedan	5	refurbishe	https://im	model: me	https://www.carlist.my/cars-for-		
16	toyota harrier	203000	old klang road	kuala lumpur	toyota	harrier	2023	12500	petrol - un	white	suv	5	refurbishe	https://im	2023 toyo	https://www.carlist.my/cars-for-		
17	porsche panamera	518000	old klang road	kuala lumpur	porsche	panamera	2019	32500	petrol - un	silver	hatchback	4	refurbishe	https://im	2019 porsc	https://www.carlist.my/cars-for-		
18	toyota alphard	351000	johor bahru	johor	toyota	alphard	2024	2500	petrol - un	white	mpv	6	refurbishe	https://im	* genuine	https://www.carlist.my/cars-for-		
19	lexus nx250	283000	old klang road	kuala lumpur	lexus	nx250	2023	12500	petrol - un	white	suv	5	refurbishe	https://im	2023 lexus	https://www.carlist.my/cars-for-		
20	mercedes-benz g350	578000	old klang road	kuala lumpur	mercedes-	g350	2021	37500	diesel	black	suv	5	refurbishe	https://im	2021 merc	https://www.carlist.my/cars-for-		
21	lexus rx300	261000	subang jaya	selangor	lexus	rx300	2021	27500	petrol - un	black	suv	5	refurbishe	https://im	2021 lexus	https://www.carlist.my/cars-for-		
22	toyota alphard	323000	old klang road	kuala lumpur	toyota	alphard	2023	27500	petrol - un	white	mpv	7	refurbishe	https://im	2023 toyo	https://www.carlist.my/cars-for-		
23	mini clubman	213000	old klang road	kuala lumpur	mini	clubman	2022	17500	petrol - un	white	wagon	5	refurbishe	https://im	2022 mini	https://www.carlist.my/cars-for-		
24	toyota vellfire	395000	old klang road	kuala lumpur	toyota	vellfire	2023	22500	petrol - un	white	mpv	7	refurbishe	https://im	2023 toyo	https://www.carlist.my/cars-for-		
25	lexus is300	237000	old klang road	kuala lumpur	lexus	is300	2021	22500	petrol - un	white	sedan	5	refurbishe	https://im	2021 lexus	https://www.carlist.my/cars-for-		
26	volvo xc60	128900	seri kembangan	selangor	volvo	xc60	2020	102500	petrol - un	white	suv	5	used	https://im	2020 volvc	https://www.carlist.my/cars-for-		
27	mitsubishi triton	19800	kajang	selangor	mitsubishi	triton	2010	152500	diesel	white	pickup truc	5	used	https://im	(welcome	https://www.carlist.my/cars-for-		
28	isuzu d-max	19800	kajang	selangor	isuzu	d-max	2011	142500	diesel	black	pickup truc	5	used	https://im	(welcome	https://www.carlist.my/cars-for-		

DATA OPTIMIZATION

POLARS

POLARS IS AN OPEN-SOURCE LIBRARY FOR DATA MANIPULATION, KNOWN FOR BEING ONE OF THE FASTEST DATA PROCESSING SOLUTIONS ON A SINGLE MACHINE. IT FEATURES A WELL-STRUCTURED, TYPED API THAT IS BOTH EXPRESSIVE AND EASY TO USE.



PANDAS

PANDAS IS A POWERFUL, OPEN-SOURCE DATA ANALYSIS AND MANIPULATION LIBRARY FOR PYTHON. IT PROVIDES EASY-TO-USE DATA STRUCTURES AND TOOLS FOR WORKING WITH STRUCTURED DATA.

DUCKDB

DUCKDB SERVES AS A POWERFUL ENGINE FOR ANALYZING AND TRANSFORMING DATA RESIDING IN VARIOUS FORMATS LIKE CSV, PARQUET, JSON, OR DATABASES SUCH AS POSTGRESQL AND SQLITE. IT CAN BE USED TRANSIENTLY FOR DATA MANIPULATION OR TO CREATE PERSISTENT TABLES FOR ANALYTICAL QUERIES.

COMPARISON

TOTAL PROCESSING TIME (SECONDS)

Optimization Stage	Run 1	Run 2	Run 3	Average
Pandas Optimization	5.35	3.49	3.44	4.09
DuckDB Optimization	0.45	0.89	0.43	0.59
Polars Optimization	0.43	0.26	0.61	0.43

CPU USAGE (%)

Optimization Stage	Run 1	Run 2	Run 3	Average
Pandas Optimization	14.67	16.67	3.6	11.65
DuckDB Optimization	19.8	50	46.2	38.67
Polars Optimization	83.1	36.2	14.8	44.7

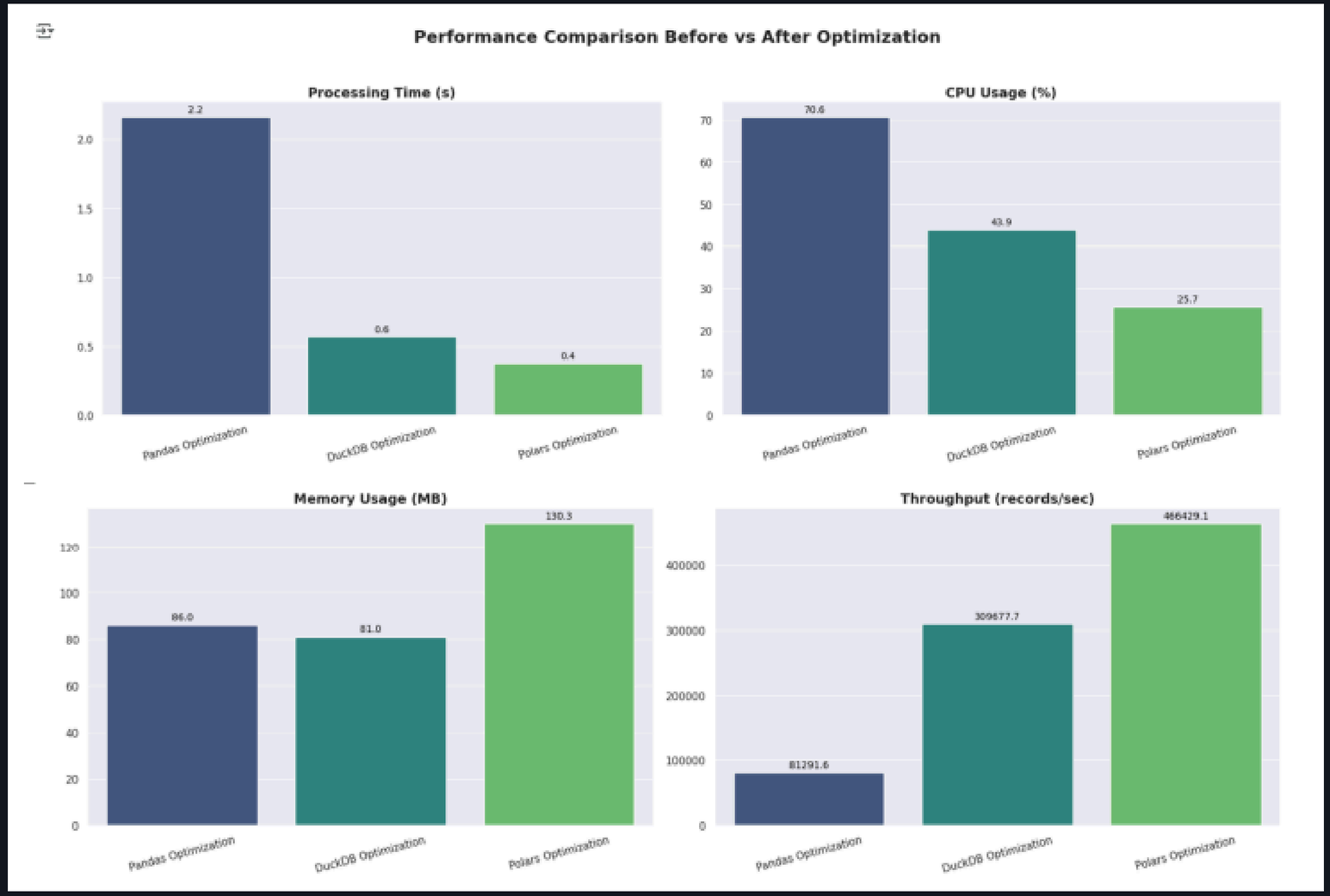
MEMORY USAGE (MB)

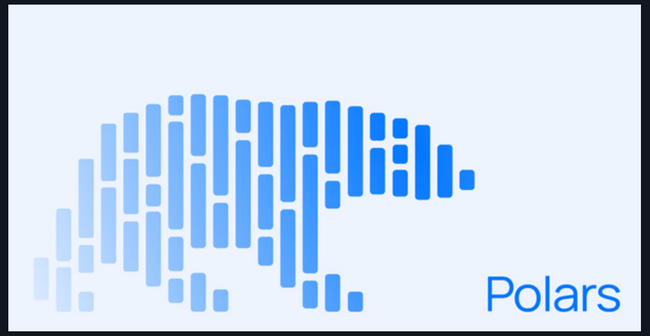
Optimization Stage	Run 1	Run 2	Run 3	Average
Pandas Optimization	102.34	119.98	125.8	116.04
DuckDB Optimization	65.79	98.83	81.64	82.09
Polars Optimization	28.63	3.81	129.63	54.02

THROUGHPUT

Optimization Stage	Run 1	Run 2	Run 3	Average
Pandas Optimization	32827.06	50356.60	51003.49	44729.0
DuckDB Optimization	390337.6	196502.90	406446.43	331095.6
Polars Optimization	412467.7	676460.87	289688.39	459539.0

VISUALIZATION





CHALLENGES AND LIMITATIONS

LONG DATA SRAPING DURATION DURING EARLY PHASE

FINDING SUITAIBLE OPTIMIZATION LIBRARY

SUDDEN CHANGES AND RESTRICTIONS BY CARLISTS.MY

THANK YOU

...

+123-456-7890

www.reallygreatsite.com

hello@reallygreatsite.com

123 Anywhere St., Any City, ST 12345