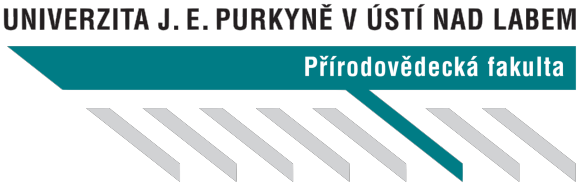
Univerzita Jana Evangelisty Purkyně
v Ústí nad Labem
Přírodovědecká fakulta



OLAP a ClickHouse

Seminární práce

Vypracoval: Martin Kopecký

Studijní program: Aplikovaná informatika
Studijní obor: Informační systémy

Ústí nad Labem 2024

Obsah

1. Výběr vhodného DBMS

Na návrh ostatních kolegů jsem si vybral ClickHouse. Při srovnání různých databázových systémů jsem dospěl k názoru, že výběr správného DBMS pro moje potřeby není zásadně důležitý. ClickHouse mi však umožnil úspěšně zvládnout téměř všechny zadané úkoly. Jediný úkol který jsem nebyl schopen udělat bylo udělání data miningu nad daty.

1.1 Instalace

ClickHouse lze hostovat přímo na uživatelově zařízení, což je podrobně popsáno v oficiální dokumentaci ClickHouse. Bohužel instalace na osobní zařízení neobsahuje grafické rozhraní tak jsem se rozhodl pro použití alternativy v podobě cloudového řešení přímo na stránkách ClickHouse.

1.2 Volba datasetu

Jako dataset jsem zvolil data ohledně cen bytů v oblasti Singaporů, které pocházejí z let 2012 až 2023. Datase obsahuje záznamy o jednotlivých bytech a to sice -

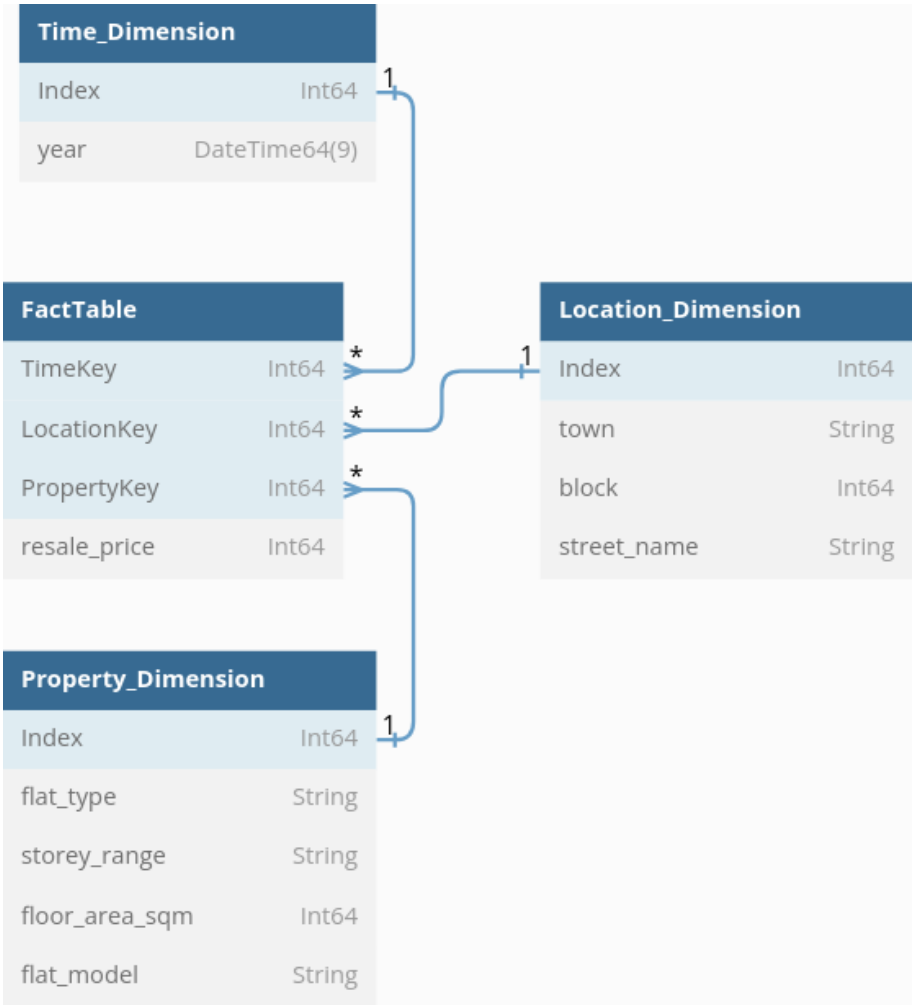
1. MONTH - Rok a měšíc zápisu
2. TOWN - Název města
3. FLAT_TYPE - Velikost bytu (2 ROOM, 3 ROOM, 4 ROOM atd.)
4. BLOCK - Čtvrť
5. STREET_NAME - Název ulice
6. STOREY_RANGE - Patra na kterých se byt nachází
7. FLOOR_AREA_SQM - Rozloha bytu
8. FLAT_MODEL - Typ či kategorie bytu (Imporved, New Generation atd.)
9. LEASE_COMMENCE_DATE - Datum počátku nájmu
10. REMAINING_LEASE - Čas do konce nájmu (nekonzistentní mezi soubory)

2. Čištění datasetu

Protože jsem měl štěstí při výběru datasetu, tak jsem měl již čistá data v datasetu.

2.1 Datová kostka

Při pohledu na data jsem se rozhodlo pro použití schématu hvězdy. Kdy jsem dataset rozdělil na jednotlivé dimenze. Nakonec jsem skončil u 3 dimenzí (Čas, Lokace, Majetek), které spolu svazuje klíčová tabulka.



2.2 Datová kostka

Po rozdělení datasetu do dimenzí vypadají jednotlivé dimenze zhruba takto

2.2.1 Location_Dimension

Location_Dimension

#	Index	town	block	street_name
1	1	ANG MO KIO	170	ANG MO KIO AVE 4
2	2	ANG MO KIO	174	ANG MO KIO AVE 4
3	3	ANG MO KIO	216	ANG MO KIO AVE 1
4	4	ANG MO KIO	215	ANG MO KIO AVE 1
5	5	ANG MO KIO	218	ANG MO KIO AVE 1

2.2.2 Time_Dimension

Time_Dimension

#	Index	year
1	1	2000-01-01 00:00:00.000000000
2	2	2000-01-01 00:00:00.000000000
3	3	2000-01-01 00:00:00.000000000
4	4	2000-01-01 00:00:00.000000000
5	5	2000-01-01 00:00:00.000000000

2.2.3 Property_Dimension

Property_Dimension

#	Index	flat_type	storey_range	floor_area_sqm	flat_model
1	1	3 ROOM	07 TO 09	69	Improved
2	2	3 ROOM	04 TO 06	61	Improved
3	3	3 ROOM	07 TO 09	73	New Generation
4	4	3 ROOM	07 TO 09	73	New Generation
5	5	3 ROOM	07 TO 09	67	New Generation

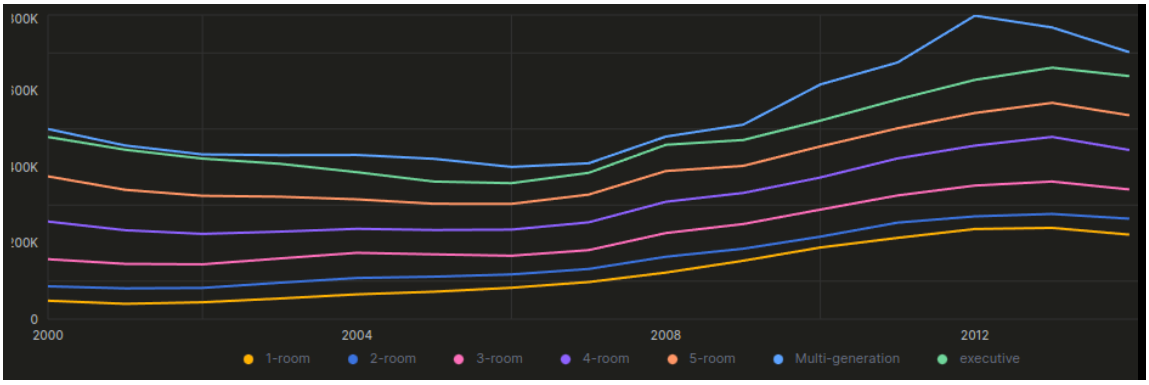
3. Jednotlivé řezy kostky

3.1 Ceny bytů napříč roky - rozděleny podle typu

```
SELECT
    toYear(Time_Dimension.year) as Rok,
    avgIf(resale_price, flat_type = '1 ROOM') as `1-room`,
    avgIf(resale_price, flat_type = '2 ROOM') as `2-room`,
    avgIf(resale_price, flat_type = '3 ROOM') as `3-room`,
    avgIf(resale_price, flat_type = '4 ROOM') as `4-room`,
    avgIf(resale_price, flat_type = '5 ROOM') as `5-room`,
    avgIf(resale_price, flat_type = 'MULTI-GENERATION')
        as `Multi-generation`,
    avgIf(resale_price, flat_type = 'EXECUTIVE')
        as `executive`
FROM
    FactTable
JOIN
    Property_Dimension ON
        Property_Dimension.Index = FactTable.PropertyKey
JOIN
    Time_Dimension ON
        Time_Dimension.Index = FactTable.TimeKey
GROUP BY
    Rok;
```

#	Rok	1-room	2-room	3-room	4-room	5-room	Multi-generation	executive
1	2000	48	86397.6618705036	157649.63752716736	256722.8454379005	375507.44337722694	500050	479088.62939549587
2	2001	40334.61538461538	80914.53703703704	145396.36860655033	233900.2003795563	340183.6247446374	456665.2173913043	445524.24054982816
3	2002	44427.27272727273	82158.20307692308	144188.71717454106	224250.82910830202	324427.59698387934	433185.71428571426	422225.57708254806

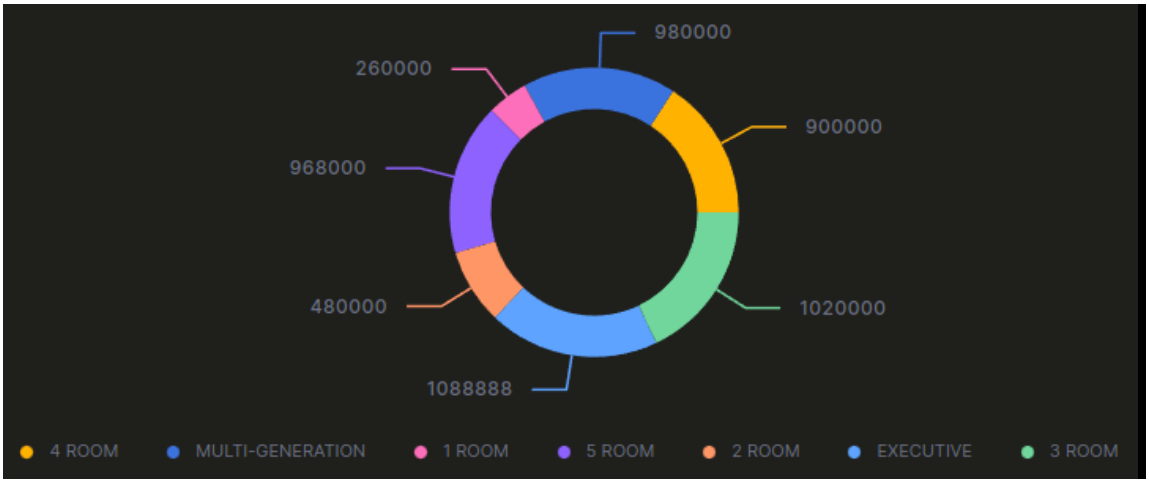
1-room - Float64



3.2 Nejvyšší cena bytu - rozděleno na typy

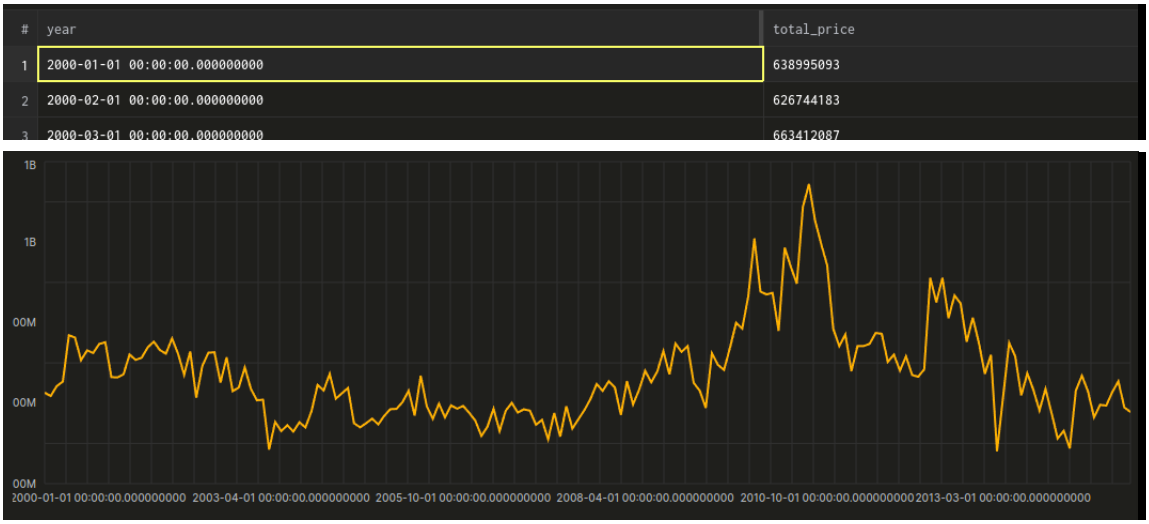
```
SELECT
    flat_type, MAX(resale_price)
        as highest_resale_price
FROM
    FactTable
JOIN
    Property_Dimension ON
        FactTable.PropertyKey = Property_Dimension.Index
GROUP BY
    flat_type;
```

#	flat_type	highest_resale_price
1	4 ROOM	900000
2	MULTI-GENERATION	980000
3	1 ROOM	260000



3.3 Průměrná cena bytů - na časové ose

```
SELECT
    Time_Dimension.year, sum(FactTable.resale_price)
        as total_price
FROM
    FactTable
JOIN
    Time_Dimension ON
        FactTable.TimeKey = Time_Dimension.Index
GROUP BY
    Time_Dimension.year
ORDER BY
    Time_Dimension.year;
```

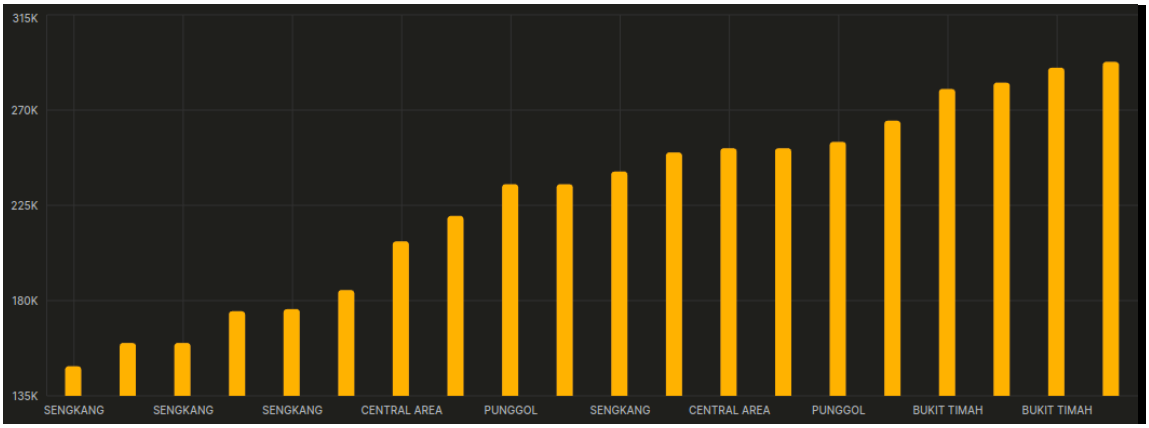


#	year	total_price
1	2000-01-01 00:00:00.000000000	638995093
2	2000-02-01 00:00:00.000000000	626744183
3	2000-03-01 00:00:00.000000000	663412087

3.4 20 nejvyšší cena bytů - rozděleno na města

```
SELECT
    td.year, ld.town, MAX(ft.resale_price)
        AS highest_resale_price
FROM
    FactTable AS ft
JOIN
    Location_Dimension AS ld ON ft.LocationKey = ld.Index
JOIN
    Time_Dimension AS td ON ft.TimeKey = td.Index
GROUP BY
    td.year, ld.town
ORDER BY
    highest_resale_price
LIMIT 20;
```

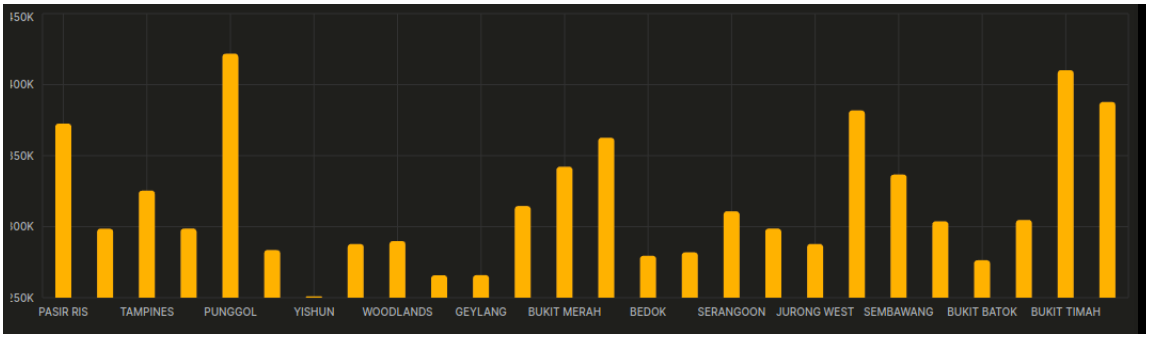
#	td.year	ld.town	highest_resale_price
1	2000-08-01 00:00:00.000000000	SENGKANG	149000
2	2000-11-01 00:00:00.000000000	SENGKANG	160000
3	2000-10-01 00:00:00.000000000	SENGKANG	160008



3.5 Průměrná cena bytů - rozděleno na města

```
SELECT
    l.town, AVG(f.resale_price) AS average_resale_price
FROM
    FactTable AS f
JOIN
    Location_Dimension AS l ON f.LocationKey = l.Index
GROUP BY
    l.town;
```

#	town	average_resale_price
1	PASIR RIS	372582.98076190474
2	TAMPINES	325410.64166521817
3	PUNGGOL	421770.3691547556
4	CENTRAL AREA	298647.3360861759



4. Data Mining

Z toho co jsem dohledal tak ClickHouse neumožňuje využívat data mining metod, tedy jsem se rozhodl tento krok vynechat a místo něj udělat 1 řez navíc.

5. Závěr

V rámci předmětu a vypracováním tohoto projektu, jsem pochopil využití OLAP databází. Dále jsem si zlepšil znalosti principů OLAP databází a uplatnil načerpané znalosti ze seminářů v praxi. Troufnu si říct, že toto není naposled co jsem usedal k OLAP databázím a v budoucím profesním životě se k ním jistě rád budu vracet. Nicméně stále nejsem dostatečně znalý OLAP systémů natolik, abych si troufl říct že je plně ovládám, proto se jistě v rámci samostudia ještě k tomuto tématu navrátím.

Seznam použité literatury

1. SY-RAHMADI. Resale HDB Flat Prices 2012 - 2023 [https://www.kaggle.com/datasets/syrahmadi/resale-hdb-flat-prices-2000-2022/data]. 30. prosinec 2023.
2. CLICKHOUSE. ClickHouse Docs. ClickHouse. 2024. Dostupné také z: https://clickhouse.com/docs/.