OLAP a Datamining

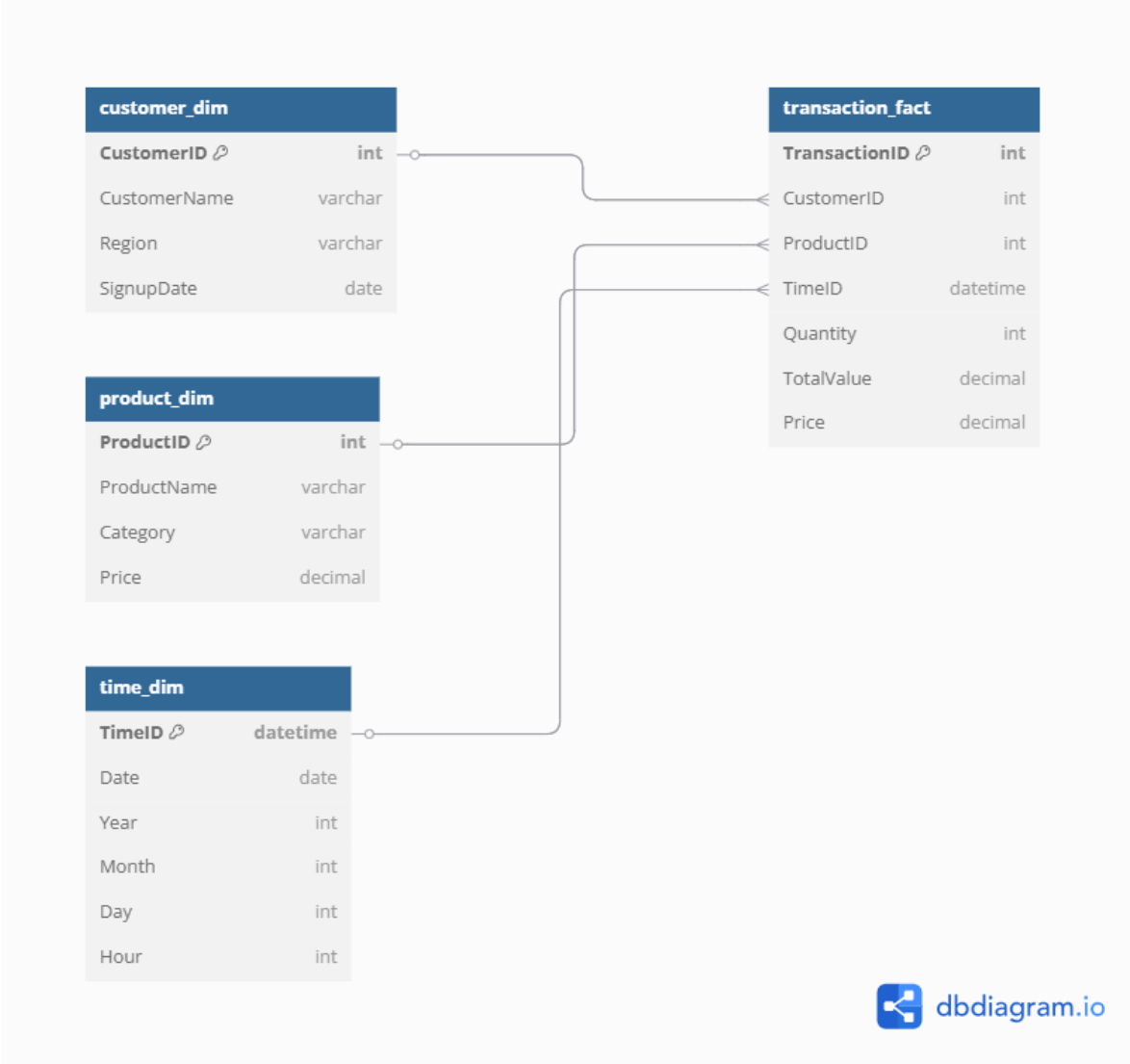
Michal Soukup

30. června 2025

Obsah

1 ERD

Zvolená data data [1]



2 DuckDB jako analytické řešení

DuckDB[2] je in-process analytická databáze určená pro OLAP[3] dotazy. Mezi její hlavní výhody patří:

- Integrace přímo v Pythonu, či jiném jazyce (nepotřebuje server)
- Vysoký výkon pro sloupcové operace

V projektu byla DuckDB využita pro spouštění SQL dotazů uložených v souboru `olap_queries.sql`. Dotazy byly spouštěny prostřednicvím Pythonu

3 Analýza výstupů OLAP dotazů

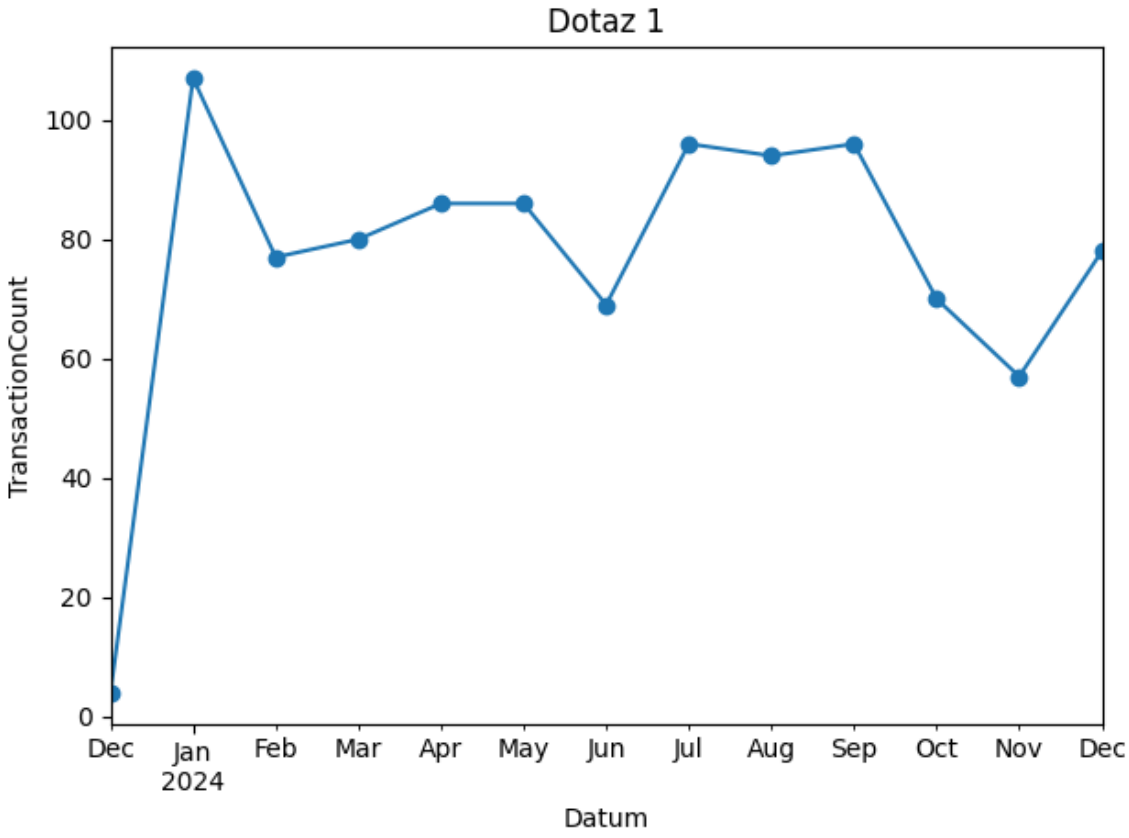
V této části jsou prezentovány výsledky čtyř vybraných analytických dotazů nad datovým skladem. Výsledky byly zpracovány pomocí nástroje DuckDB a uloženy jako textové i CSV výstupy.

3.1 Dotaz 1 – Počet transakcí za měsíc

```
SELECT t.Year, t.Month, COUNT(*) AS TransactionCount
FROM transaction_fact f
JOIN time_dim t ON f.TimeID = t.TimeID
GROUP BY t.Year, t.Month
ORDER BY t.Year, t.Month
```

Year	Month	TransactionCount
2023	12	4
2024	01	107
2024	02	77
2024	03	80
2024	04	86
2024	05	86
2024	06	69
2024	07	96
2024	08	94
2024	09	96
2024	10	70
2024	11	57
2024	12	78

Tabulka 1: Počet transakcí za jednotlivé měsíce

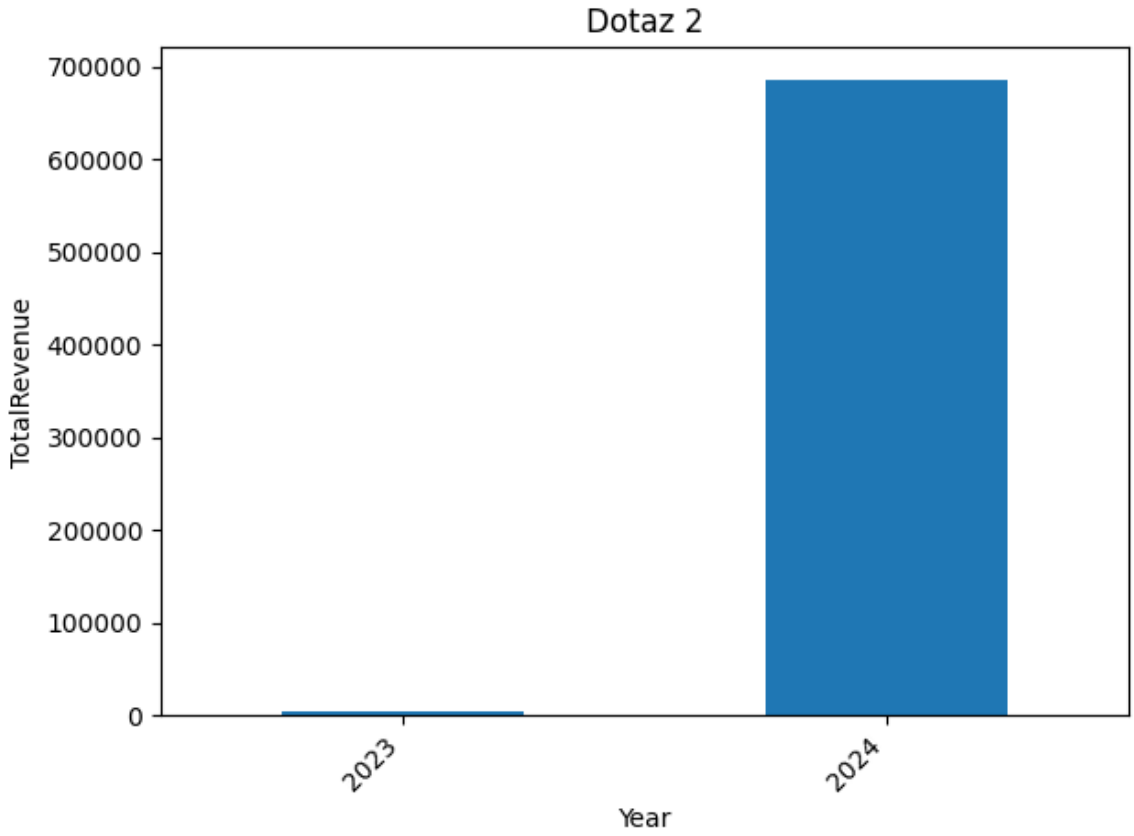


3.2 Dotaz 2 – Celkový obrat za jednotlivé roky

```
SELECT t.Year, SUM(f.TotalValue) AS TotalRevenue
FROM transaction_fact f
JOIN time_dim t ON f.TimeID = t.TimeID
GROUP BY t.Year
ORDER BY t.Year
```

Year	TotalRevenue
2023	3769.52
2024	686226.04

Tabulka 2: Celkový obrat za roky 2023 a 2024

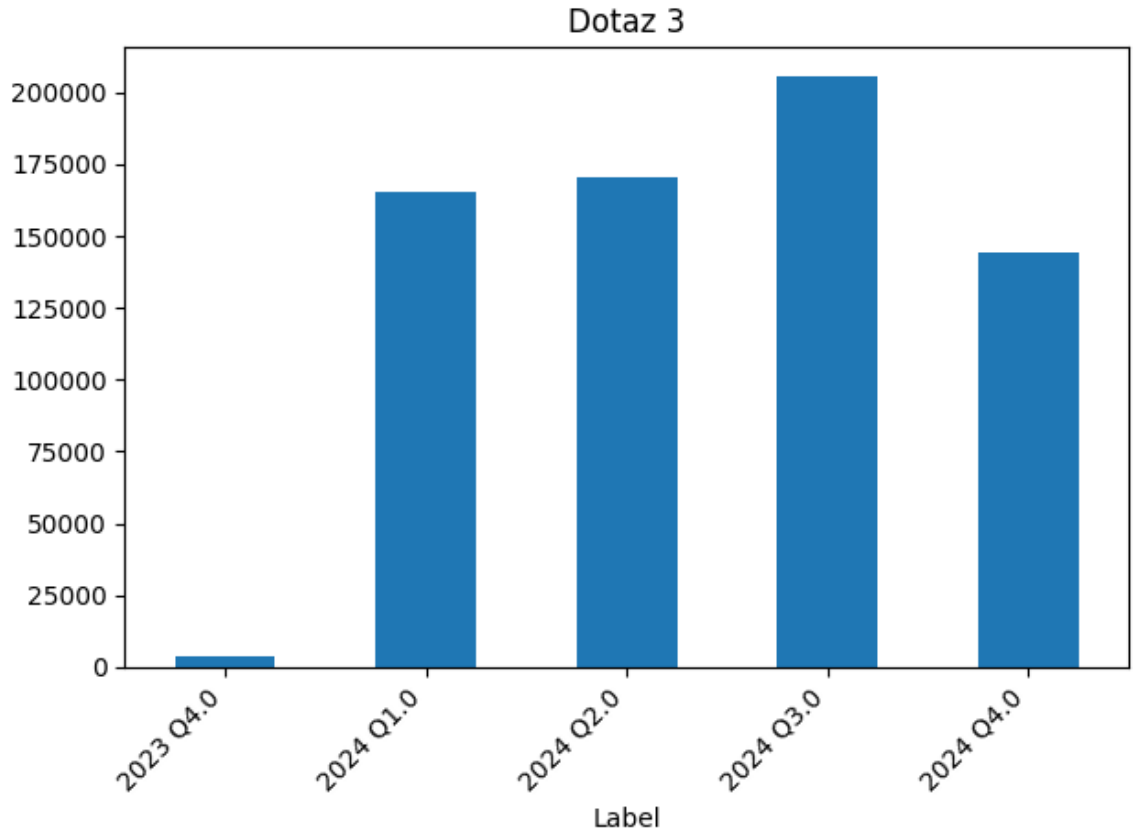


3.3 Dotaz 3 – Sezónnost: obrat podle kvartálu

```
SELECT
    t.Year,
    FLOOR((CAST(t.Month AS INTEGER) - 1) / 3) + 1 AS Quarter,
    SUM(f.TotalValue) AS QuarterlyRevenue
FROM transaction_fact f
JOIN time_dim t ON f.TimeID = t.TimeID
GROUP BY t.Year, Quarter
ORDER BY t.Year, Quarter
```

Year	Quarter	QuarterlyRevenue
2023	4	3769.52
2024	1	165664.39
2024	2	170817.98
2024	3	205406.88
2024	4	144336.79

Tabulka 3: Obrat podle kvartálů

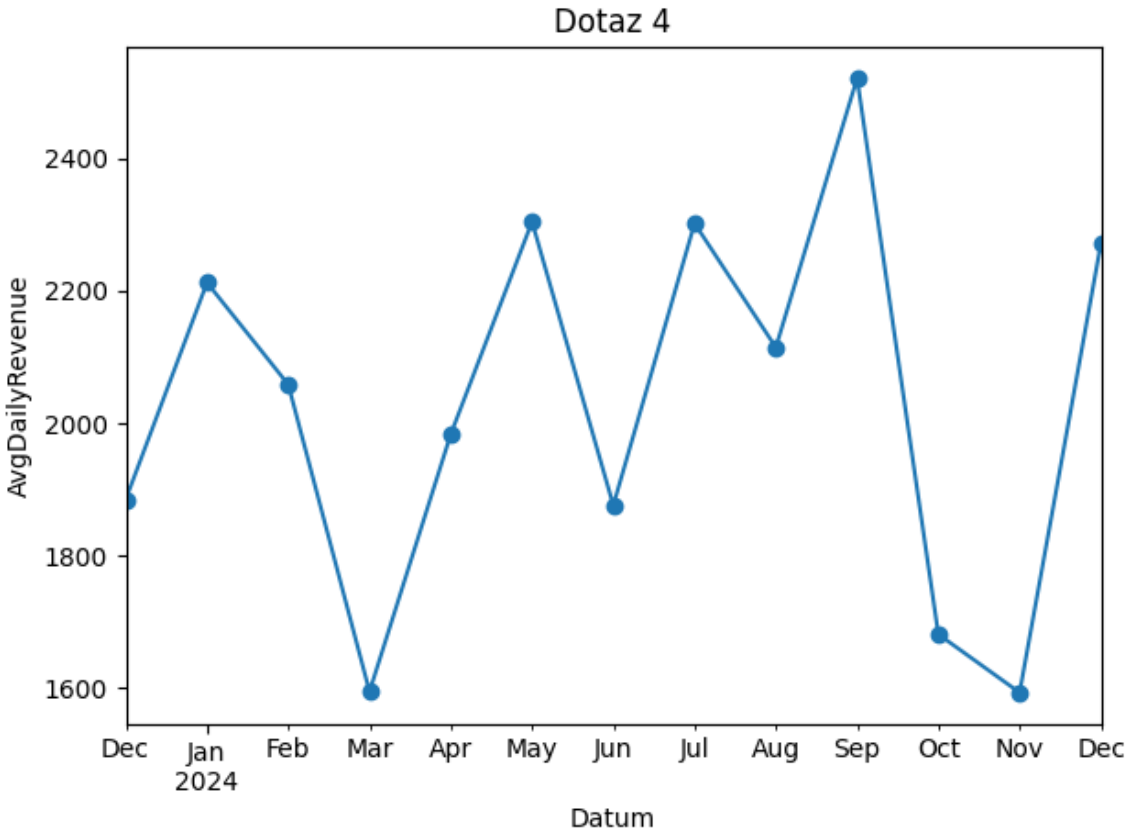


3.4 Dotaz 4 – Průměrná denní tržba v jednotlivých měsících

```
SELECT
    daily_data.Year,
    daily_data.Month,
    AVG(daily_data.DailyRevenue) AS AvgDailyRevenue
FROM (
    SELECT
        t.Year,
        t.Month,
        t.Day,
        SUM(f.TotalValue) AS DailyRevenue
    FROM transaction_fact f
    JOIN time_dim t ON f.TimeID = t.TimeID
    GROUP BY t.Year, t.Month, t.Day
) AS daily_data
GROUP BY daily_data.Year, daily_data.Month
ORDER BY daily_data.Year, daily_data.Month
```

Year	Month	AvgDailyRevenue
2023	12	1884.76
2024	01	2212.55
2024	02	2058.37
2024	03	1594.29
2024	04	1983.42
2024	05	2304.56
2024	06	1875.81
2024	07	2302.14
2024	08	2114.56
2024	09	2521.56
2024	10	1680.83
2024	11	1592.68
2024	12	2271.12

Tabulka 4: Průměrná denní tržba v jednotlivých měsících



4 Analýza klasifikačních modelů

Pro klasifikaci produktů podle jejich kategorie byly vyzkoušeny různé algoritmy strojového učení. Hodnocení probíhalo pomocí pětinásobné křížové validace a metrik přesnosti, recallu a F1 skóre. Výsledky ukázaly významné rozdíly mezi jednotlivými modely.

4.1 Random Forest

Model **Random Forest** dosáhl vynikajících výsledků s průměrnou přesností z křížové validace **94,7 %**. V testovací množině dosáhl celkové přesnosti **89 %**, přičemž jednotlivé třídy byly klasifikovány velmi vyrovnaně. Nejnižší recall měl pro kategorii `Clothing` (81 %).

Kategorie	**Precision**	**Recall**	**F1-score**
Books	0.85	0.95	0.90
Clothing	0.93	0.81	0.87
Electronics	0.92	0.88	0.90
Home Decor	0.87	0.89	0.88
Celkem	**0.89**	**0.89**	**0.89**

Tabulka 5: Výsledky klasifikace – Random Forest

4.2 Gradient Boosting

Model **Gradient Boosting** dosáhl dokonce průměrné přesnosti z křížové validace **99,9 %**. Na testovací množině dosáhl celkové přesnosti **96 %** a skvělých výsledků napříč všemi kategoriemi. Byl velmi přesný zejména u `Electronics` (precision 1.00) a `Home Decor` (f1-score 1.00).

Kategorie	**Precision**	**Recall**	**F1-score**
Books	0.92	1.00	0.96
Clothing	0.94	0.90	0.92
Electronics	1.00	0.95	0.97
Home Decor	1.00	1.00	1.00
Celkem	**0.97**	**0.96**	**0.96**

Tabulka 6: Výsledky klasifikace – Gradient Boosting

4.3 Decision Tree

Model **Decision Tree** vykazoval extrémně vysokou přesnost v rámci křížové validace i testovací množiny. S průměrnou přesností **99,9 %** a F1 skóre přesahujícím 0.97 u všech tříd dosáhl podobně skvělých výsledků jako Gradient Boosting.

Kategorie	Precision	Recall	F1-score
Books	0.96	1.00	0.98
Clothing	1.00	0.93	0.97
Electronics	0.99	1.00	1.00
Home Decor	1.00	1.00	1.00
Celkem	**0.99**	**0.98**	**0.98**

Tabulka 7: Výsledky klasifikace – Decision Tree

4.4 Shrnutí a interpretace

Z výše uvedených výsledků je patrné, že modely **Gradient Boosting** a **Decision Tree** dosáhly téměř bezchybné klasifikace. Přestože jejich přesnost působí výjimečně, je důležité zvážit riziko přeučení modelu (overfitting), zejména u modelu Decision Tree, který vykazoval téměř dokonalou shodu s trénovacími daty.

Model **Random Forest** poskytl rovněž velmi kvalitní predikce s výborným poměrem mezi přesností a generalizací. Lze jej považovat za vhodného kandidáta pro reálné nasazení, kde je žádoucí kompromis mezi přesností a robustností.

Na základě těchto výsledků lze usoudit, že data obsahují dostatek relevantních informací pro úspěšnou predikci kategorií produktů. To je cenné zejména při automatickém označování nových produktů, detekci nesprávných záznamů nebo návrhu cílených marketingových kampaní.

5 Závěr

Projekt demonstruje možnost použití DuckDB jako lehkého analytického nástroje a propojení OLAP dotazů s data mining pipeline. Kromě dobrých výsledků GradientBoosting a DecisionTree metod, byla vizualizace klíčová pro interpretaci sezónních trendů i predikční přesnosti modelů.

Odkazy

1. WAMBLES, Chad. *Ecommerce Transactions Dataset*. 2021. Dostupné také z: `https://www.kaggle.com/datasets/chadwambles/ecommerce-transactions`. Citováno 30. června 2025.
2. RAASVELDT, Mark; MÜHLEISEN, Hannes. *DuckDB*. 2019. Dostupné také z: `https://duckdb.org`. Citováno 30. června 2025.
3. WIKIPEDIA CONTRIBUTORS. *Online Analytical Processing*. 2024. Dostupné také z: `https://cs.wikipedia.org/wiki/Online_Analytical_Processing`. Citováno 30. června 2025.