Data Warehouse Optimization – report

1. Aim of the laboratory

The aim of the task is to show the issues related to the various physical models of cubes as well as aggregation design.

2. Preliminary Assumptions

Size of the database (data warehouse): 89.65 MB
Number of fact tuples in the database (data warehouse): 1196262

Testing environment:

Hardware:

- Intel® Xeon® CPU E5-2680 v4 processor
- 32GB of RAM

Software:

- Windows 10
- SQL Server Management Studio 20 (opened during testing)
- SQL Server Profiler 20 (opened during testing)
- Visual Studio 2022

3. Testing

Testing query execution times for different models, with and without defined aggregations. Testing cube processing times in the same testing settings.

Queries description

Brief description of the queries:

1. What profit was earned during each month from May 2019 to September 2019?

```
select {[Start Date].[Hierarchy].[Year].[2019].[2].[May]:[Start
Date].[Hierarchy].[Year].[2019].[3].[September]} on columns,
{[Measures].[Profit]} on rows
from Warehouse;
```

2. In how many tours was the number of participants not maximal?

```
select {[Measures].[Number of tours]} on columns, {[Tour].[Was
Group Full].[False]} on rows
from Warehouse;
```

3. In how many cases was there a big "negative" difference between the transportation rating and the transportation price in the last year?

```
WITH MEMBER [Measures].[No of cases] AS
'SUM(UNION([Tour].[Transportation Rating].[Low],
[Tour].[Transportation Rating].[Very low]), [Measures].[Number of
tours])'
SELECT [Transportation Group].[Price Category].[High] ON COLUMNS,
[Measures].[No of cases] ON ROWS
FROM Warehouse
WHERE [Start Date].[Year].[2019];
```

Description of testing procedure

Processing times of cube and queries for MOLAP with and without aggregations (outliers are marked in grey), in milliseconds.

MOLAP							
Cube processing		Query 1		Query 2		Query 3	
Aggr	No aggr	Aggr	No aggr	Aggr	No aggr	Aggr	No aggr
8684	7887	43	120	20	75	57	70
8666	7809	41	131	16	75	54	59
8724	7719	43	130	20	80	61	56
8613	7830	31	123	19	71	55	59
8662	7726	39	122	24	73	59	59
8586	7872	46	132	24	77	64	46
8545	7809	38	138	21	64	51	59
8667	7838	38	127	22	69	60	65
8665	7961	42	134	22	67	61	64
8652	7930	40	123	20	73	59	56

Processing times of cube and queries for ROLAP with and without aggregations (outliers are marked in grey), in milliseconds.

ROLAP							
Cube processing		Query 1		Query 2		Query 3	
Aggr	No aggr	Aggr	No aggr	Aggr	No aggr	Aggr	No aggr
2543	2754	429	418	262	249	220	242

2538	2602	340	377	263	246	248	246
2579	2682	399	414	267	251	222	232
2621	2647	368	364	239	252	242	237
2537	2482	416	403	253	251	238	231
2551	2546	390	407	244	252	246	242
2708	2554	361	395	250	248	233	233
2519	2444	392	403	249	254	229	229
2524	2515	423	384	253	254	247	239
2671	2565	357	405	247	253	242	224

Processing times of cube and queries for HOLAP with and without aggregations (outliers are marked in grey), in milliseconds.

HOLAP							
Cube processing		Query 1		Query 2		Query 3	
Aggr	No aggr	Aggr	No aggr	Aggr	No aggr	Aggr	No aggr
2619	2597	32	433	22	248	223	237
2564	2757	44	405	17	244	222	220
2696	2600	51	395	19	253	234	221
2630	2678	39	403	21	239	200	225
2620	2679	38	397	21	253	245	227
2611	2559	50	403	21	238	242	221
2573	2703	38	387	21	242	241	228
2573	2450	41	400	21	243	239	233
2561	2711	43	426	27	244	215	224
2620	2530	33	415	17	244	239	220

Comparison between MOLAP, ROLAP and HOLAP

Mean and standard deviation values for measurements (outliers were not excluded).

	MOLAP		ROLAP		HOLAP	
	Aggr	No aggr	Aggr	No aggr	Aggr	No aggr
Delay	-	-	-	-	-	-
Querying speed (mean) [ms]	40.1	128	387.5	397	40.9	406.4
	20.8	72.4	252.7	251	20.7	244.8
	58.1	59.3	236.7	235.5	230	225.6
Querying speed (standard deviation) [ms]	4.1	5.9	30.2	17.1	6.3	14.2
	2.4	4.8	8.9	2.6	2.8	5.1
	3.9	6.4	10.3	6.9	14.6	5.8
Processing time (mean) [ms]	8646.4	7838.1	2579.1	2579.1	2606.7	2626.4

Processing time (standard deviation) [ms]	51.5	78.6	65.3	94.3	41.1	95.5
Total size [MB]	112.87	89.64	63.51	63.52	82.01	63.52

Clearing cache

Cache had been cleared after every query execution.

Outliers

Outliers were defined as data points that are not within [μ ± 2σ] bound (excluded from the calculations when are not within [μ ± 3σ] bound).

Aggregates description

Aggregations on Tour dimension:

- Was group full
- Transportation rating

Cube Objects	Default	Full	None	Unrestricted
City	◉	○	○	○
Status	◉	○	○	○
Was Group Full	○	◉	○	○
Hotel Rating	◉	○	○	○
Beach Rating	◉	○	○	○
Transportation Rating	○	◉	○	○
Overall Rating	◉	○	○	○

Aggregations on Date dimension:

- Year
- Month

Cube Objects	Default	Full	None	Unrestricted
Start Date	5	2	0	0
Date ID	◉	○	○	○
Year	○	◉	○	○
Season	◉	○	○	○
Month	○	◉	○	○
Month No	◉	○	○	○
Day	◉	○	○	○
Quarter	◉	○	○	○

4. Discussion

Theoretical assumptions

	MOLAP	HOLAP	ROLAP
Querying time	Short	Moderate (short with well-designed aggregations)	Long
Processing time	Long	Moderate (if no aggregations are designed, it will be short)	Short
Total size	Big (size of the measure group is much smaller if no aggregations are designed for them)	Moderate	Small

Querying time

	MOLAP	ROLAP	HOLAP
	No aggr	No aggr	No aggr
Querying speed (mean) [ms]	128	397	406.4
	72.4	251	244.8
	59.3	235.5	225.6

The obtained results are consistent with the theoretical assumptions. Reading from external database greatly impacts the querying speed in the ROLAP and HOLAP (without aggregations) models. MOLAP model contains a copy of the fact table, so querying speed is much faster than in the other models.

Processing time

	MOLAP	ROLAP	HOLAP
	No aggr	No aggr	No aggr
Processing time (mean) [ms]	7838.1	2579.1	2626.4

The obtained results are consistent with the theoretical assumptions. Locating the data in the analytical database greatly impacts the cube processing time in the MOLAP model. Processing time of the ROLAP and HOLAP cubes is short because data is extracted from the external database “on demand”.

Aggregations

	MOLAP		ROLAP		HOLAP	
	Aggr	No aggr	Aggr	No aggr	Aggr	No aggr
Querying speed (mean) [ms]	40.1	128	387.5	397	40.9	406.4
	20.8	72.4	252.7	251	20.7	244.8
	58.1	59.3	236.7	235.5	230	225.6
Processing time (mean) [ms]	8646.4	7838.1	2579.1	2579.1	2606.7	2626.4

The obtained results are consistent with the theoretical assumptions. Aggregations are calculated during the processing of cube; that's why processing time with precalculated aggregations can be greater than without.

There is no obvious dependency between the aggregations and querying speed. However, in most cases it seems that precalculated aggregations speed up the querying. Results are different for each model (querying speed gains almost nothing from having the precalculated aggregations in the ROLAP model, while in the other models there is quite a big difference). No aggregates were precalculated for Query 3.

Total size

	MOLAP	ROLAP	HOLAP
	No aggr	No aggr	No aggr
Total size [MB]	89.64	63.52	63.52

The obtained results are consistent with the theoretical assumptions. The need to store cube schema, mappings, all data, and precalculated aggregations in the MOLAP model greatly impacts the size of analytical database. In the ROLAP model only cube metadata and mappings are stored in OLAP server, resulting in smaller size of analytical database. In the HOLAP model only aggregations are found in the analytical database; that's why when aggregates are defined, the size of analytical database is bigger than that of the ROLAP.

Conclusion

MOLAP model has long querying time but short processing time with big analytical database size.

ROLAP model has short querying time but long processing time with small analytical database size.

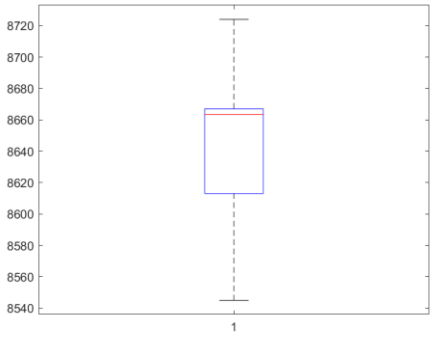
Querying time, processing time, as well as size of analytical database in HOLAP model depend on defined aggregates. The presence of well-designed aggregates reduces querying time but increases processing time and analytical database size.

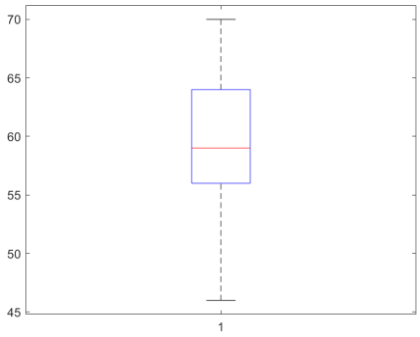
Generally, defining aggregates allows to speed up querying time for each model.

Appendix I. Box plots of the measurements

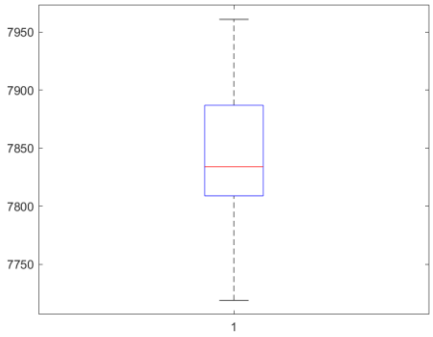
MOLAP

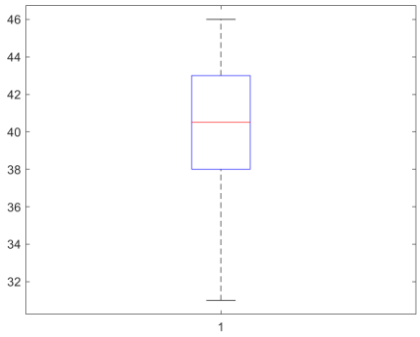
Cube processing with and without aggregations.



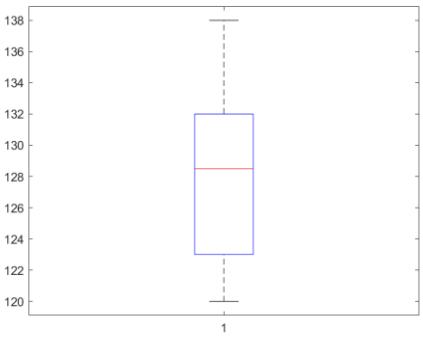


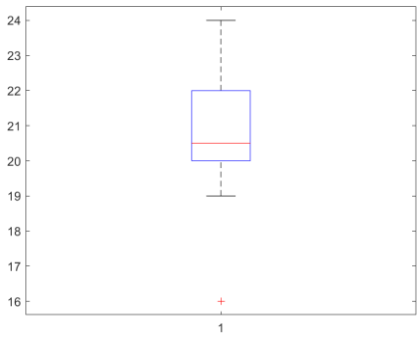
Query 1 with and without aggregations.



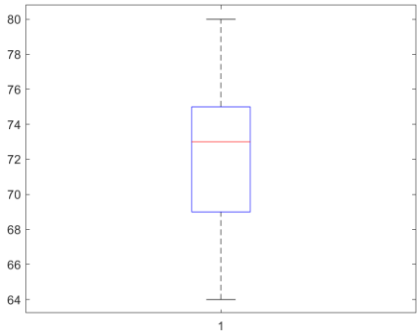


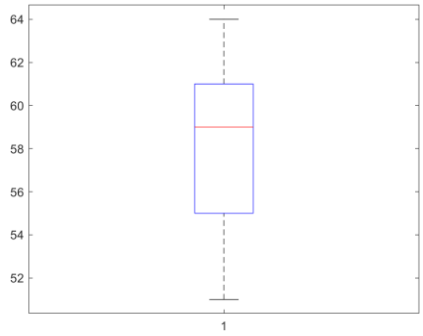
Query 2 with and without aggregations.





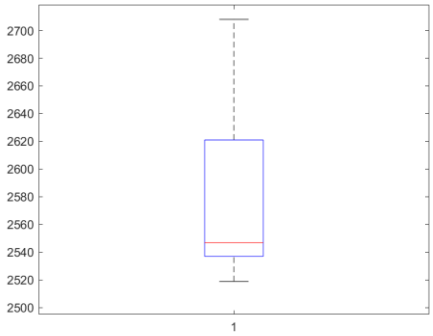
Query 3 with and without aggregations.

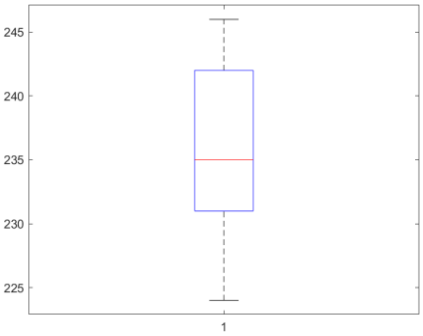




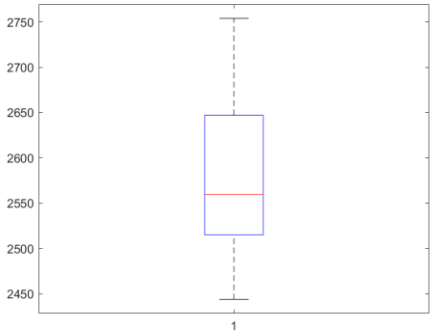
ROLAP

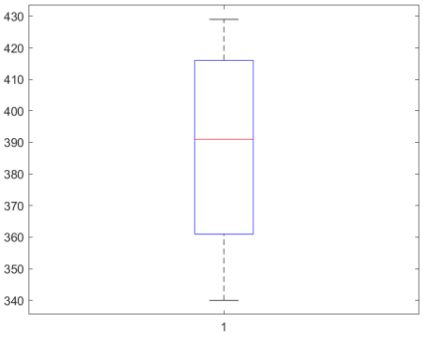
Cube processing with and without aggregations.



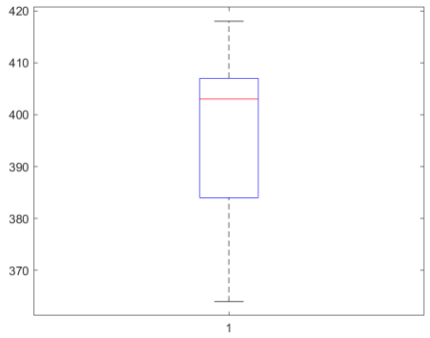


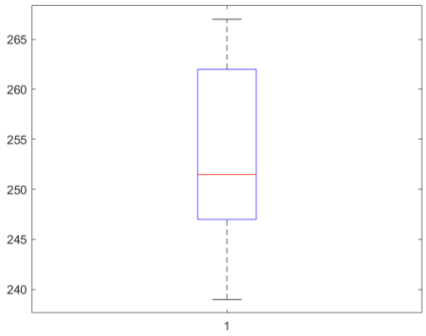
Query 1 with and without aggregations.



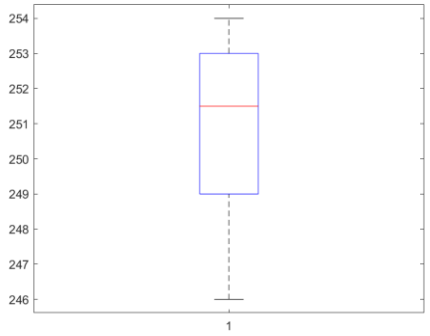


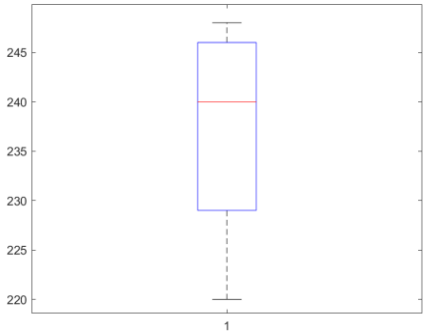
Query 2 with and without aggregations.





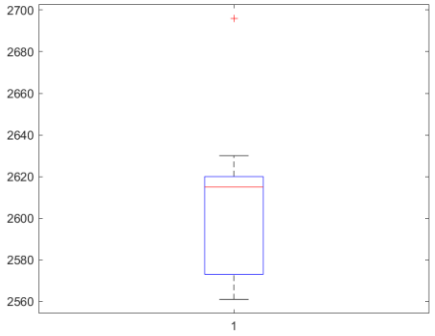
Query 3 with and without aggregations.

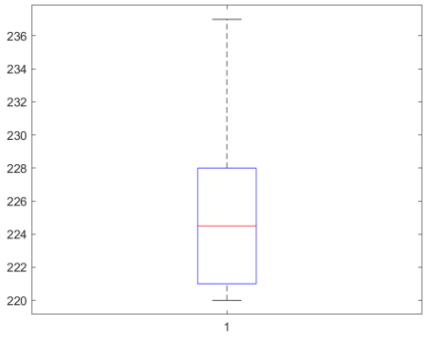




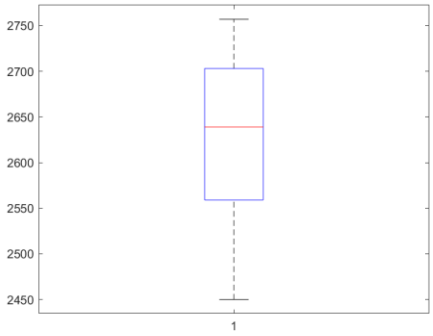
HOLAP

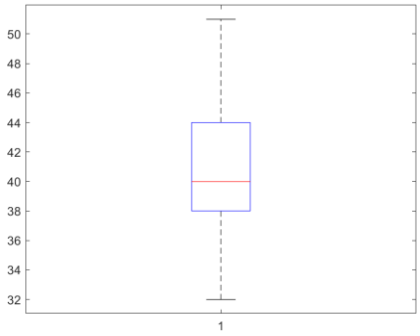
Cube processing with and without aggregations.



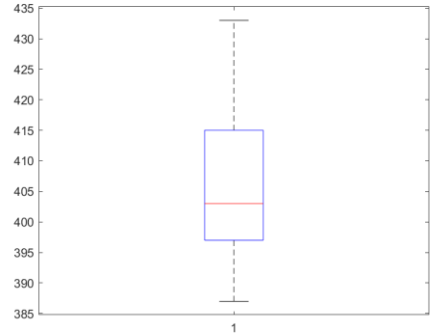


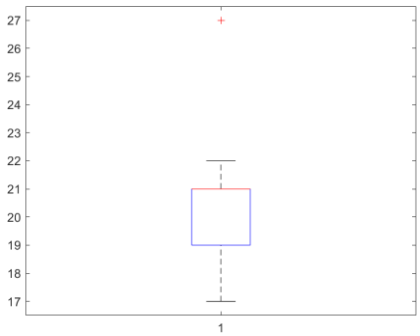
Query 1 with and without aggregations.



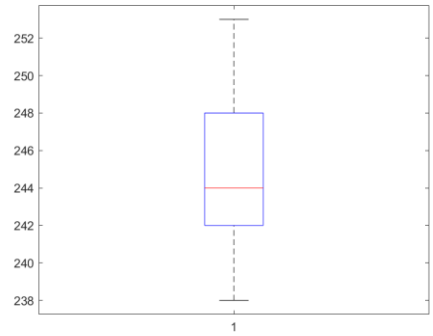


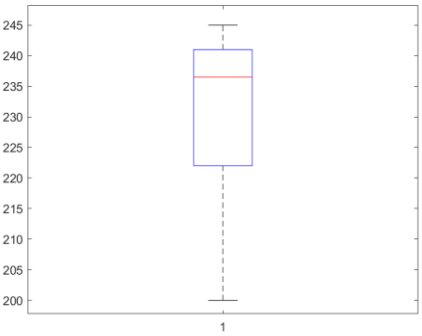
Query 2 with and without aggregations.





Query 3 with and without aggregations.





Appendix 2. T-tests

T-tests calculated for the MOLAP vs ROLAP data points with the 5% significance level.

Null hypothesis: the data in x – y comes from a normal distribution with a mean equal to zero and unknown variance.
Alternative hypothesis: ... with a mean not equal to zero ....

	Was the null hypothesis rejected		p-value	
	Aggr	No aggr	Aggr	No aggr
Cube processing	Yes	Yes	4.1688e-16	2.0816e-17
Query 1	Yes	Yes	4.1688e-16	3.1361e-11
Query 2	Yes	Yes	7.2376e-12	8.5329e-14
Query 3	Yes	Yes	6.0364e-15	1.7223e-12