Data Warehouse Optimization – report

1. Aim of the laboratory

The aim of the task is to show issues concerning various physical cube models and aggregation design.

2. Preliminary assumptions

Size of the database
(fact count): 1.770.248
(size): 784 Mb

Testing environment: Visual Studio 22, SQL Server Profiler

3. Testing

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

Queries:

3.1. SELECT NON EMPTY { [Measures].[Amount Of Passengers] } ON COLUMNS FROM (SELECT ({ [Dim Date].[Date Day Of Week].&[6] }) ON COLUMNS FROM [ETL Trains]) WHERE ([Dim Date].[Date Day Of Week].&[6])

3.2. SELECT NON EMPTY {[Measures].[Fact Real Section Count]} ON COLUMNS, NON EMPTY {TOPCOUNT(FILTER([Dim Event].[Event Type].[Event Type].ALLMEMBERS, [Dim Event].[Event Type].CURRENTMEMBER.MEMBER_CAPTION <> 'Unknown'), 1, [Measures].[Fact Real Section Count])}

3.3. WITH MEMBER [Measures].[Average Congestion Politechnika] AS AVG({[Start Station].[Station Name].&[Politechnika]}, [Measures].[AVG_Congestion]) MEMBER [Measures].[Average Congestion Raclawicka] AS AVG({[Start Station].[Station Name].&[Raclawicka]}, [Measures].[AVG_Congestion]) MEMBER [Measures].[Change in Average Congestion] AS [Measures].[Average Congestion Raclawicka] - [Measures].[Average Congestion Politechnika] SELECT {[Measures].[Average Congestion Politechnika], [Measures].[Average Congestion Raclawicka], [Measures].[Change in Average Congestion]} ON COLUMNS FROM [ETL Trains]

	MOLAP		ROLAP		HOLAP	
	Aggr.	No aggr.	Aggr.	No aggr.	Aggr.	No aggr.
Querying speed	14	59	1049	1021	103	2485
	17	174	21547	21336	31	21372
	16	126	2622	2591	86	2482
Processing time	10082	10344	3389	3426	3349	3659
Total size	48.54 Mb	48.47 Mb	23.26 Mb	23.26 Mb	23.32 Mb	23.32 Mb

4. Discussion

The processing time in MOLAP is significantly longer than in ROLAP and HOLAP due to storing the duplication data in the analytical database. On the other hand, MOLAP has the shortest querying times because of not requiring connection to the data warehouse and reading the information that process provides. Moreover, MOLAP distinguish itself from other models by greater total size, which mostly results from storing additional copy of data. What is more, values of querying speed in MOLAP and HOLAP are certainly lower with aggregations compared to without aggregations. It is because precomputed summaries allow for faster retrieval of aggregated data. It is also worth seeing that in our case querying time for the second query is extraordinary, because for aggregations: we collected value about three times smaller than the other ones and without aggregations: value over 9 times greater than the other ones. We assume, that this defect results from assigning inappropriate aggregation to the given query.