StudentIDs:	Boran Gobel, Aslan Aghakishiyev (251560)	Score:
Class:	15:15 Friday	
Date:	16.06.2023	

This mini project assignment consists of 2 tasks. You should focus on providing complete solution, however if you are unable to solve a particular step, try to give at least a partial solution or provide justification for the reason for the lack of a solution. Please note that quality and scope of the provided solution affects the grade.

Mini-project 3 – Implementation

The focus of this stage is on implementation tasks: implementation of the ETL process using SQL Server Integration Services (SSIS) and implementation of a multidimensional cube (OLAP Cube) using SQL Server Analysis Services (SSAS). You will utilise the artifacts prepared in the previous stage, where you have managed to design a detailed multidimensional model schema (star/snowflake) and an overall ETL process map.

The draft of the map serves as the blueprint for the ETL process. Now your task is to implement the process. Bear in mind that due to limited resources (mostly time) we focus on a simplified task of implementing an exemplar ETL solution, rather than building a complete, more general, ETL system. You have already specified all needed data movement, cleaning, and conformation tasks. Now it is time to act. As such, in Task 1 you need to implement the planned pipeline using SSIS (if you prefer to use a different ETL tool, or set of tools, please consult your selection with the teacher – but generally it is possible and encouraged).

Task 1 – ETL PROCESS – details:

Before the start, please divide your dataset into historical load data (data that is initially loaded into your data mart) and current load data batches (at least, two data batches that can be later loaded incrementally). Please prepare appropriate implementation of the historical load (remember to create needed target data structures) and incremental loads (remember to automate the entire process, keep metadata of your runs). Obviously, both loads require proper handling of data cleaning and conforming tasks – as planned in your ETL map.

If you create a solution that can handle only historical load of data your final score for this part of the MP is highly limited, you can get a passing grade, but it is limited to 65%. Moreover, solutions which do not introduce two separate solutions (understood, as two completely different systems that are triggered separately) to handle historical and incremental loads are preferred.

Task 1 – ETL SYSTEM – Solutions:

ETL implementation – using a tool of your choice (suggested SSIS and MS SQL Server, remember about staging the data):

1. Divide your data into three batches (first batch can be, and should be , significantly larger)
2. Historical load of data – with the first batch
    a. Define target data structures – required dimension tables and a fact table
    b. Perform extraction of needed data from the source system. Please consider limiting the extraction to small portions of data (like 10000 rows), you'll that this is a required approach for larger datasets (required for **>85%**).
    c. Perform basic anomaly detection and resolution transformations. Define a basic set of data-quality screens and implement them (and further run against extracted data). Please note that a simple data-quality screen can check for null values and mark the rows as requiring attention. If you are using SSIS, please note that more advanced screens can be implemented using external scripts. In short, define an approach and clean the identified anomalies
    d. Perform needed data transformation - creation of assumed multi-dimensional scheme - star, snowflake, or fact constellation:
        i. Create dimension tables. Remember to integrate and standardize the needed information, to make names and values verbose, to introduce surrogate keys, to handle SCDs, please also include basic sanity checks against the completeness and correctness of data, e.g., row counts. Consider implementing automated tests (required for **>85%**).
        ii. Create fact table. Remember to integrate and standardize the needed information. Please also include basic sanity checks against the completeness and correctness of data, e.g., row counts. Consider implementing automated tests (required for **>85%**).
    e. Finally load the prepared data into target tables.
        i. Define proper relations between dimensions and fact tables.

   f. Use proper logging mechanisms and proper error handling (required for **>85%**). Remember to capture all erroneous data.

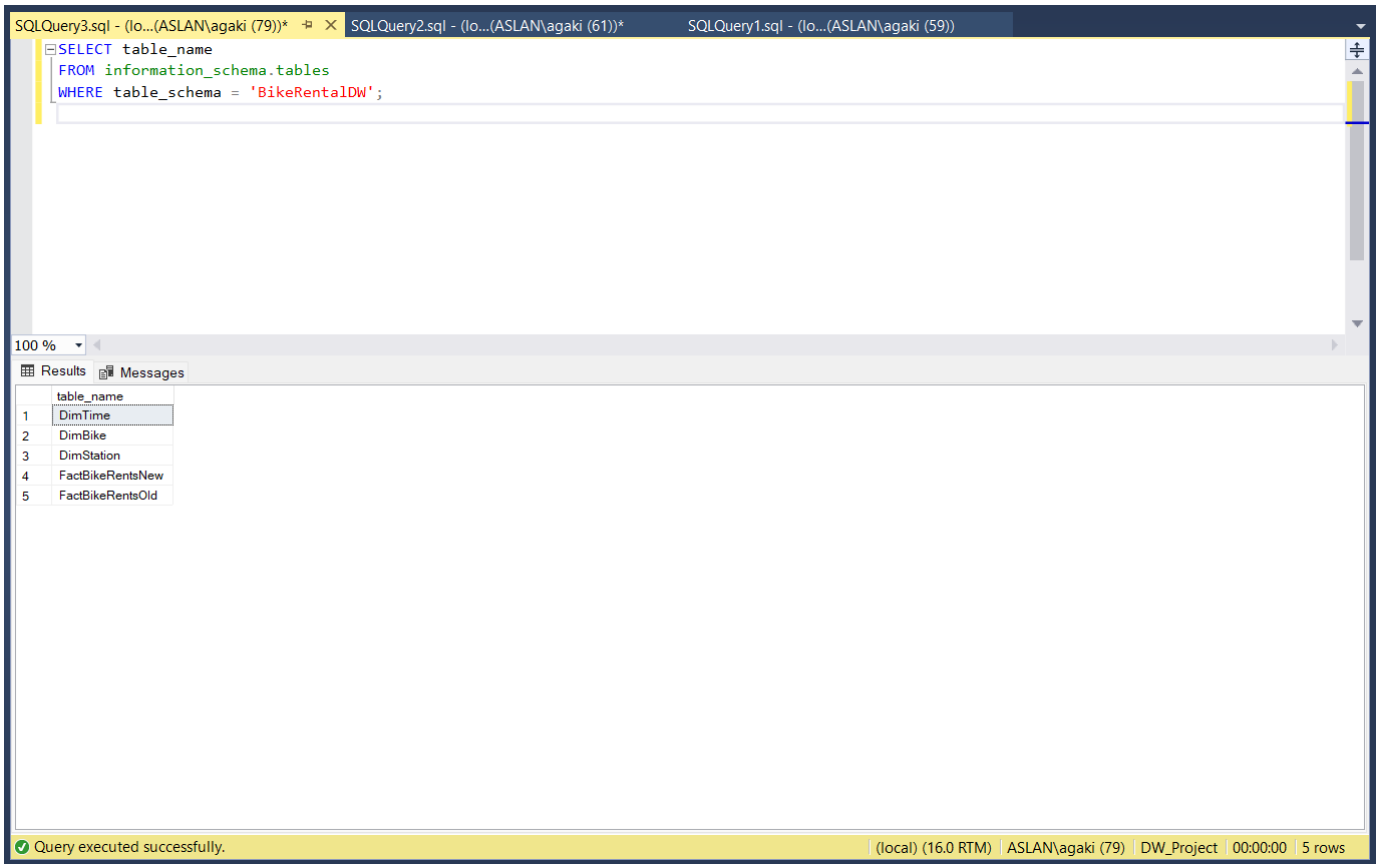
3. Incremental load of data – with the second and third batch
   a. Identify required extraction approach and implement it – remember to utilise a mechanism to capture only new/modified data (required for **>65%**). Perform extraction of needed data from the source system
   b. Perform basic anomaly detection and resolution transformations. Define a basic set of data-quality screens and implement them (and further run against extracted data). Please note that a simple data-quality screen can check for null values and mark the rows as requiring attention. If you are using SSIS, please note that more advanced screens can be implemented using external scripts. Consider implementing automated quality tests (required for >**50%)**. In short, define an approach and clean the identified anomalies.
   c. Perform needed data transformation - creation of assumed multi-dimensional scheme - star, snowflake, or fact constellation. Create dimension and fact tables. Remember Implement a mechanism to update only new/modified data (required for **>65%**)
   d. Finally load the prepared data into target tables.
   e. Use proper logging mechanisms and proper error handling (required for **>85%**). Remember to capture all erroneous data.

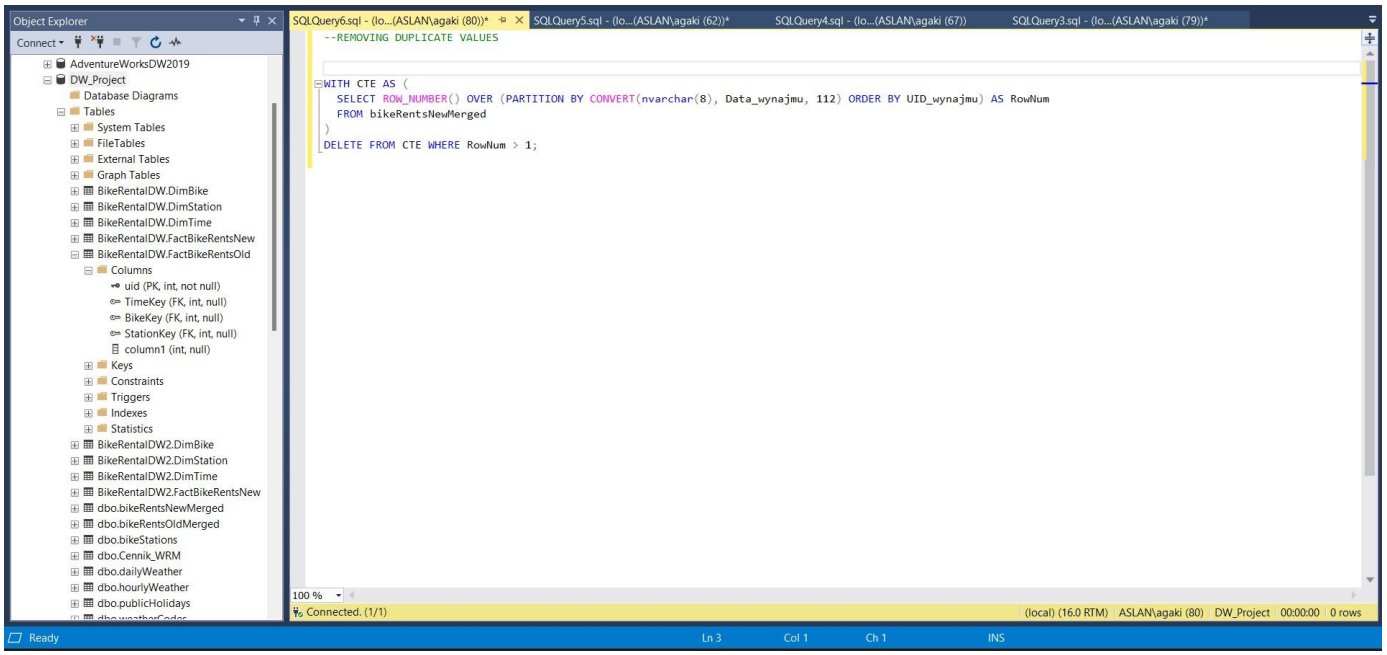
As the final solution, please upload your source files. In the final report provide a concise description of your implementation in accordance with the following points:

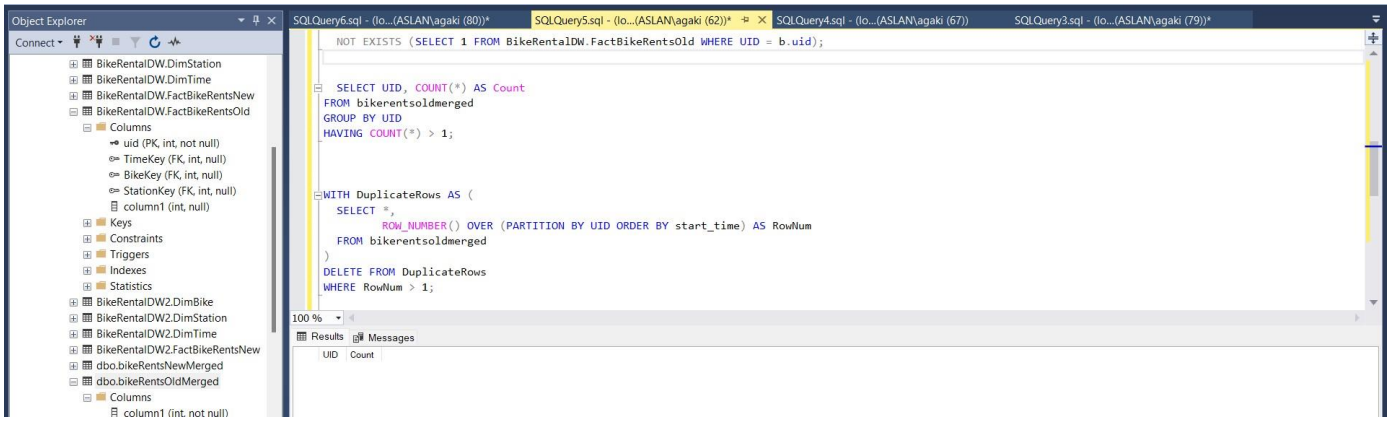
1.1 Target database creation script

Please specify

Target database we created divided into 2 batches, 1 historical and 1 incremental batch. BikeRentsNewmerge is used for historical batch, meanwhile BikeRentsOldmerges is for incremental batch. Schema created for this process named as BikeRentalDW.

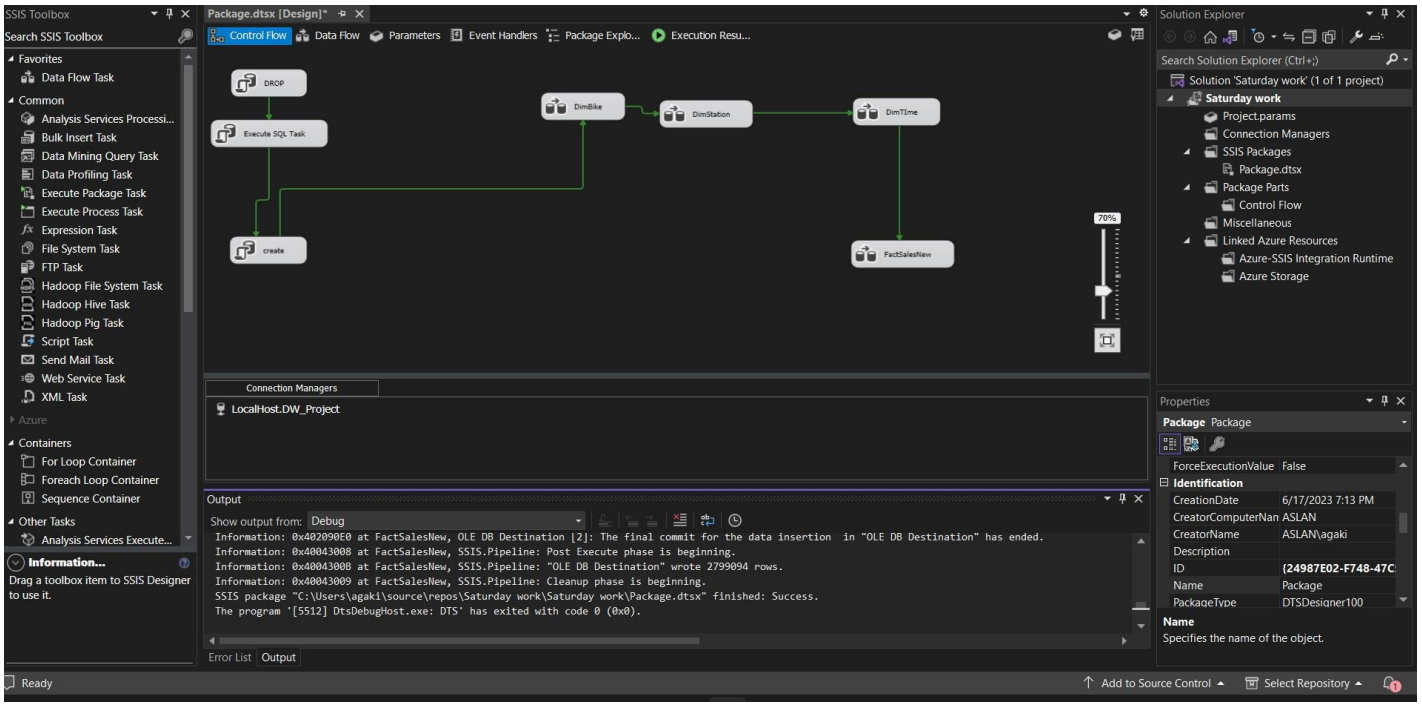






1.2 ETL Processes

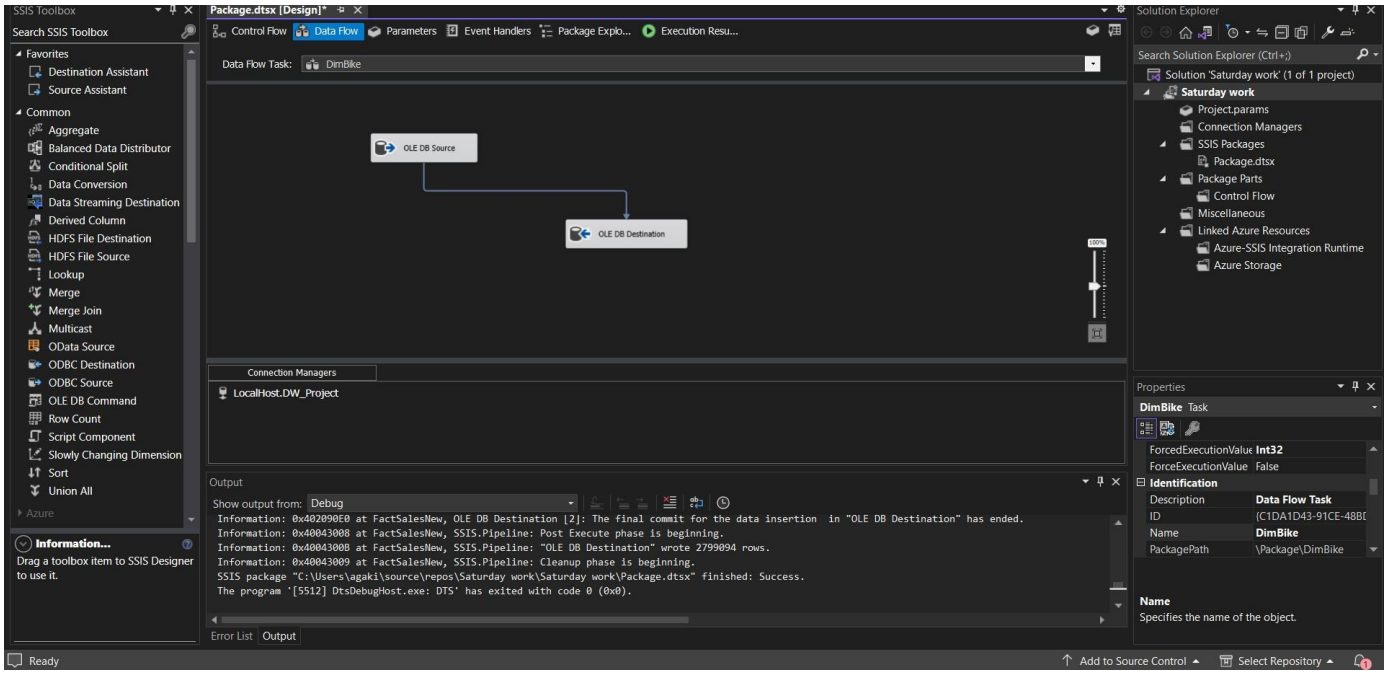
Please specify stages of your approach and provided detailed description of each stage, provide details for each target table.

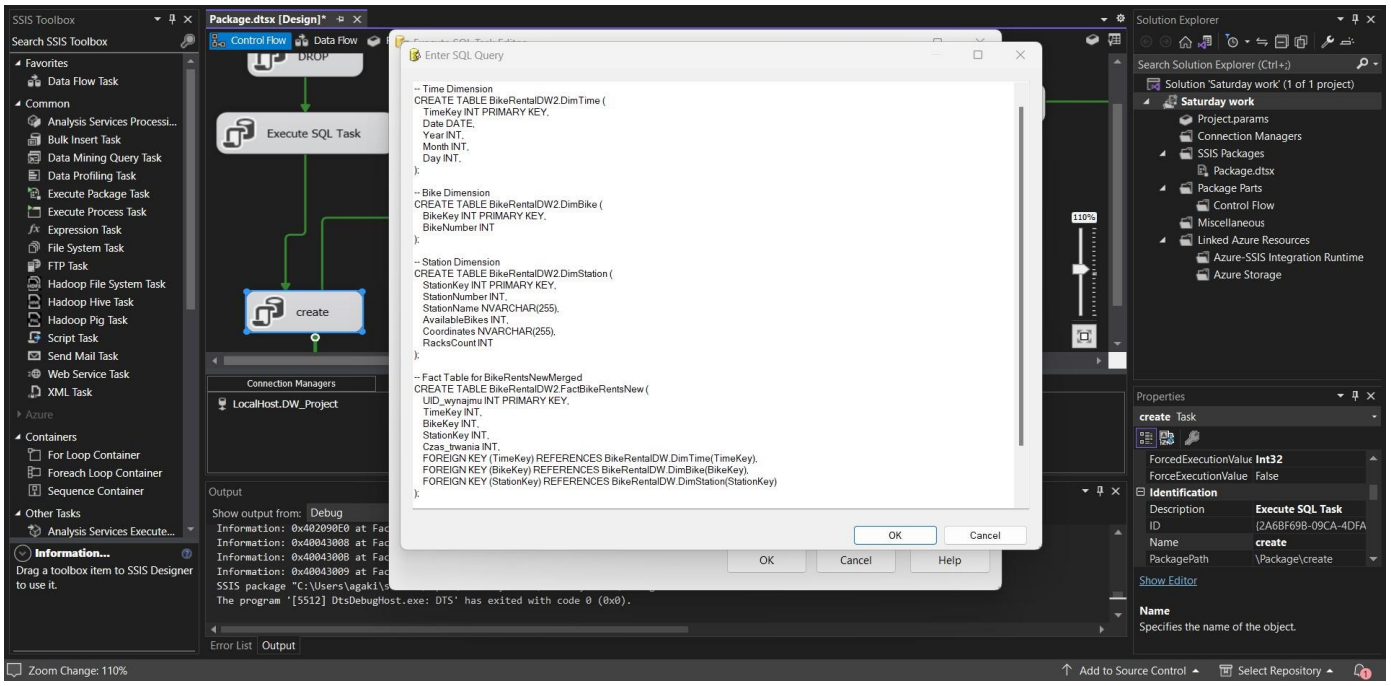


1.2.1 General approach

Please specify stages of your approach and provided detailed description of each stage.

...

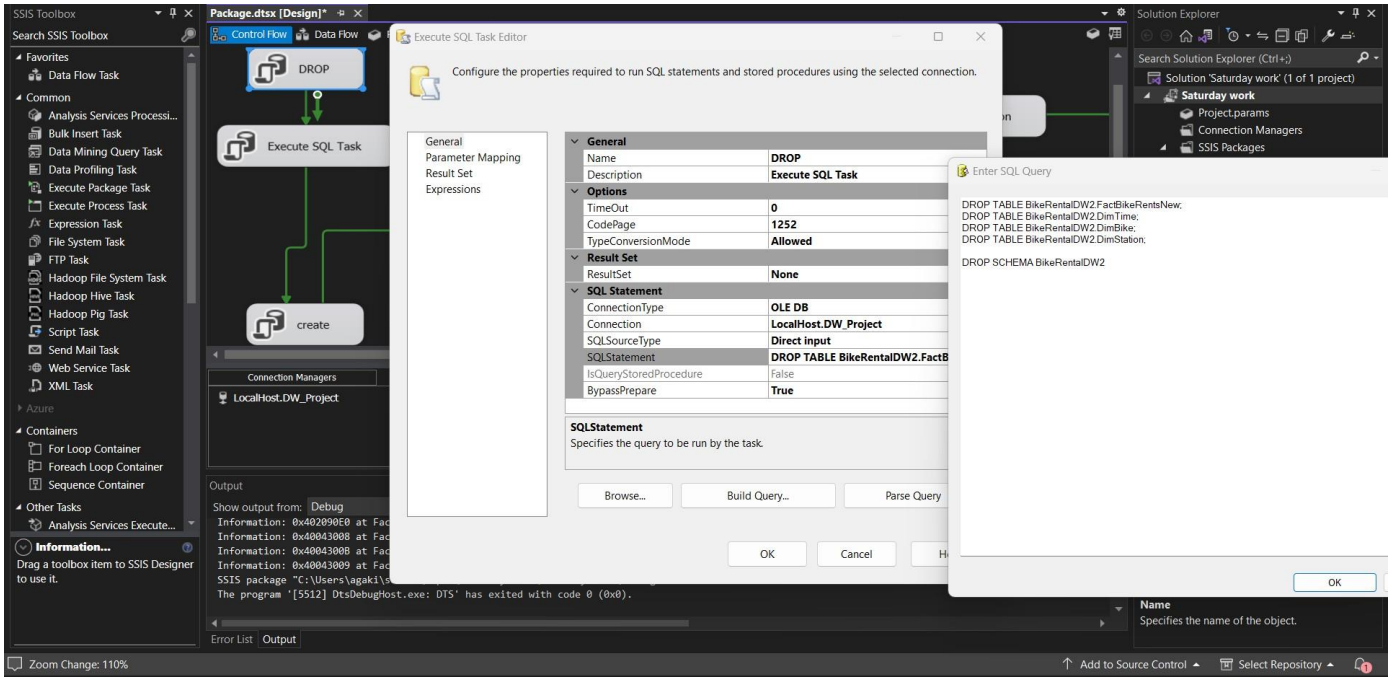




1.2.2 Dimension [X]

For each dimension [X], please specify stages of your approach and provided detailed description of each stage.

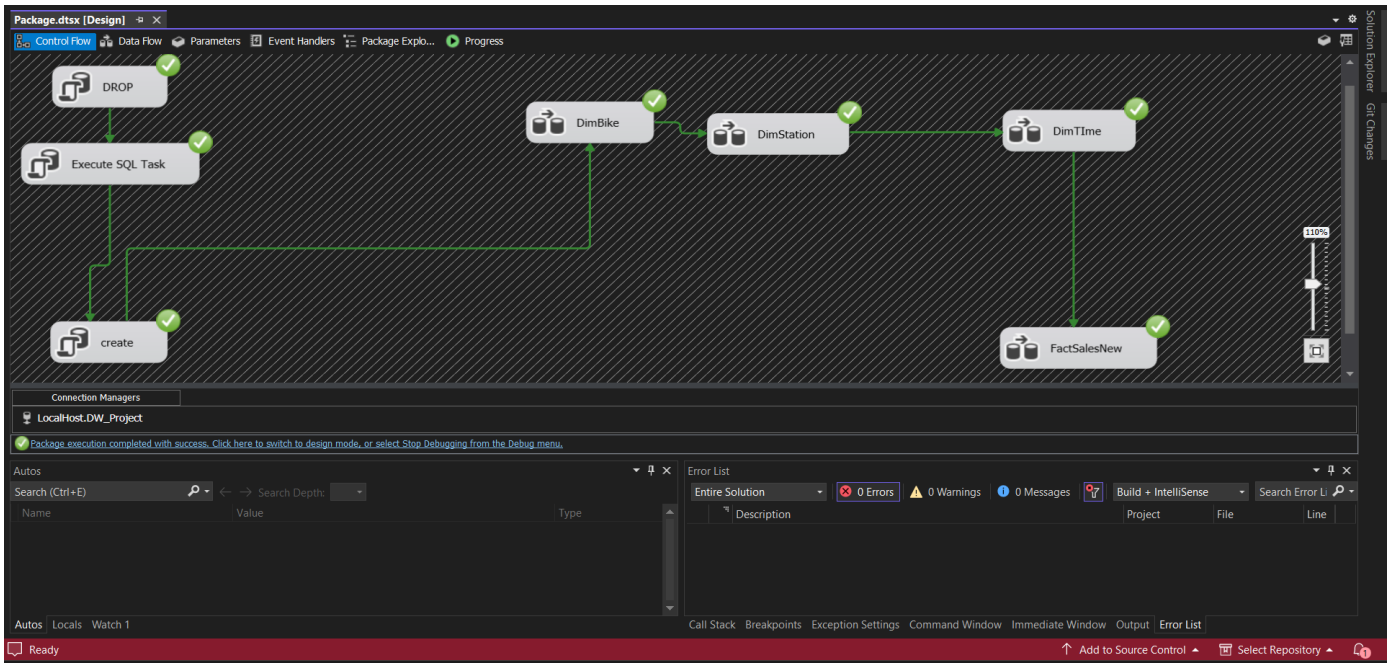
...

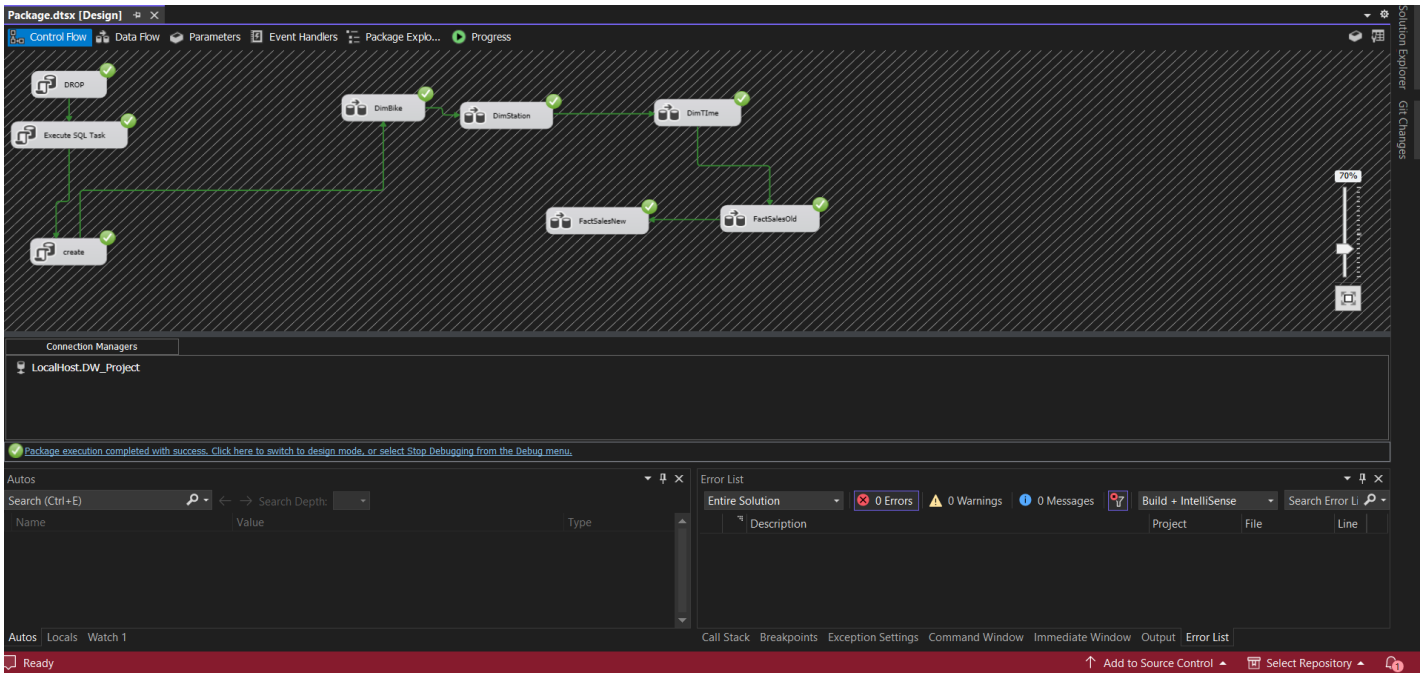


1.2.3 FACT [X]

For each fact [X], please specify stages of your approach and provided detailed description of each stage.

...





Task 2 – CUBE model – details:

Please prepare the model of the ETL process, in accordance with the below specification and discuss the solution with the lecturer.

General note: ***The complexity and completeness of the created cube affects the grade. During ongoing lab assignments and during lecture we have tackled different properties and settings related to proper OLAP cube definition and creation – try to use them and try to use them wisely.***

Task 2 – CUBE in SSAS – Solutions:

1. Prepare a cube based on prepared data from the previous task.
    a. Remember to define needed dimensions, attributes, attribute relations, hierarchies, etc. Remember about proper ordering of members within the dimensions – e.g., January comes before February
    b. Define needed measures and aggregation functions, try including at least one calculated measure

c. Additionally define (*required for 85%*): at least a single perspective, at least one KPI and proper aggregations (cube materialisations)

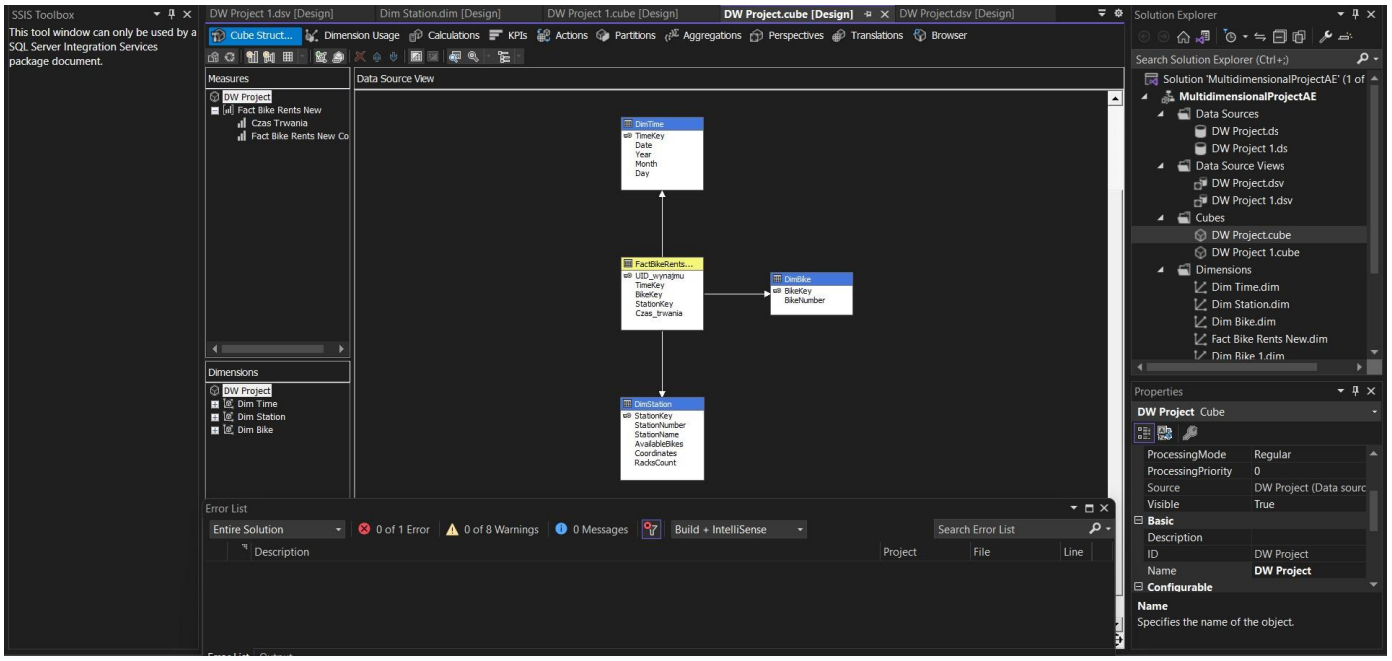
2. Process and deploy your cube in your local SSAS instance
   a. Be able to show and document your results in cube's browser

As the final solution, please upload your source files. In the final report provide a concise description of your implementation in accordance with the following points:

2.1 Final cube structure

Prepare a screenshot with your final cube structure (from SSAS).

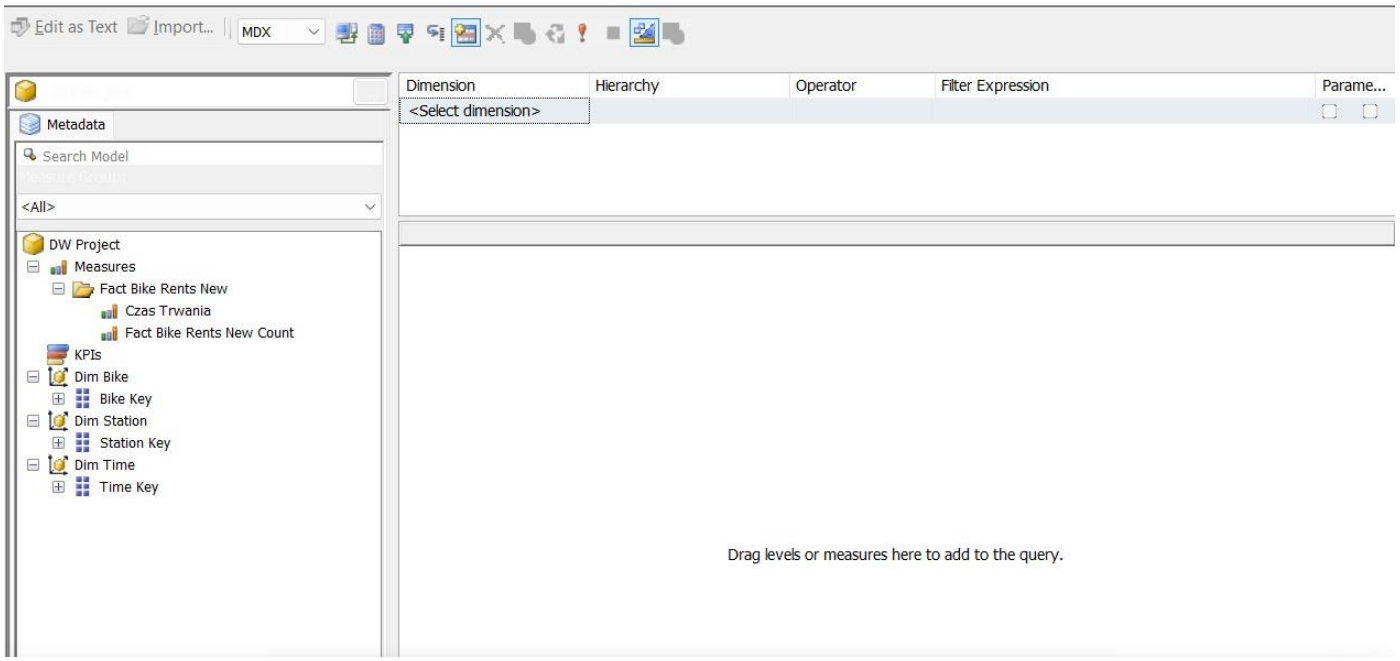
…



2.2 Measures

Prepare a short description and a screenshot for all measures (include information about formatting string and aggregation function).

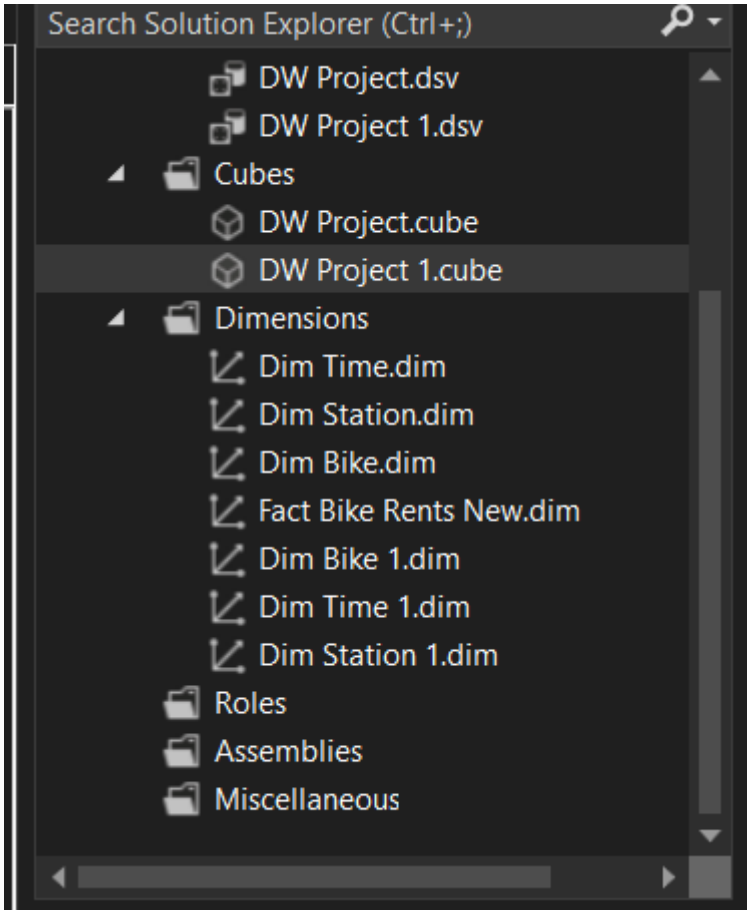
…

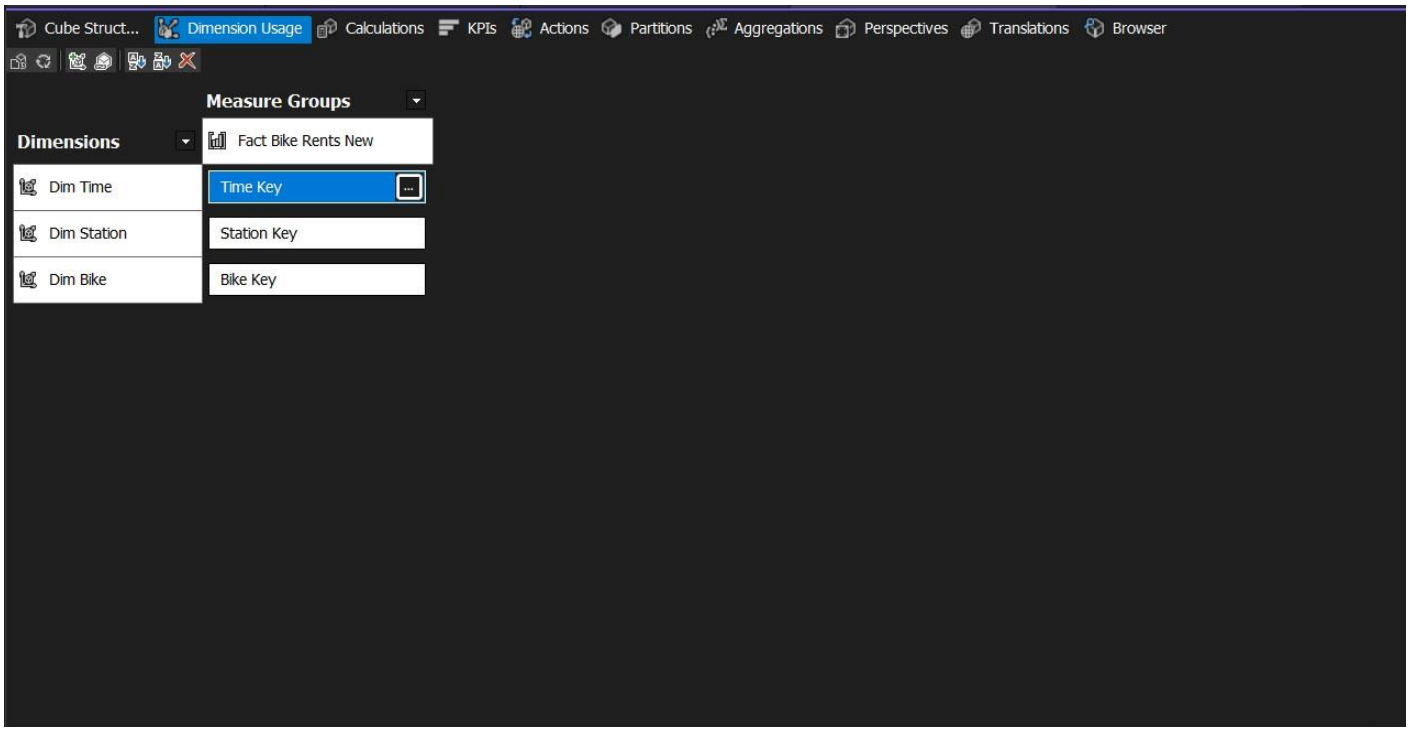


2.3 Dimensions

Prepare a short description and a screenshot for all dimension structures. Remember to include information about hierarchies, attribute relations, and additional processing within the cube (e.g., grouping, ordering, key definitions).

…

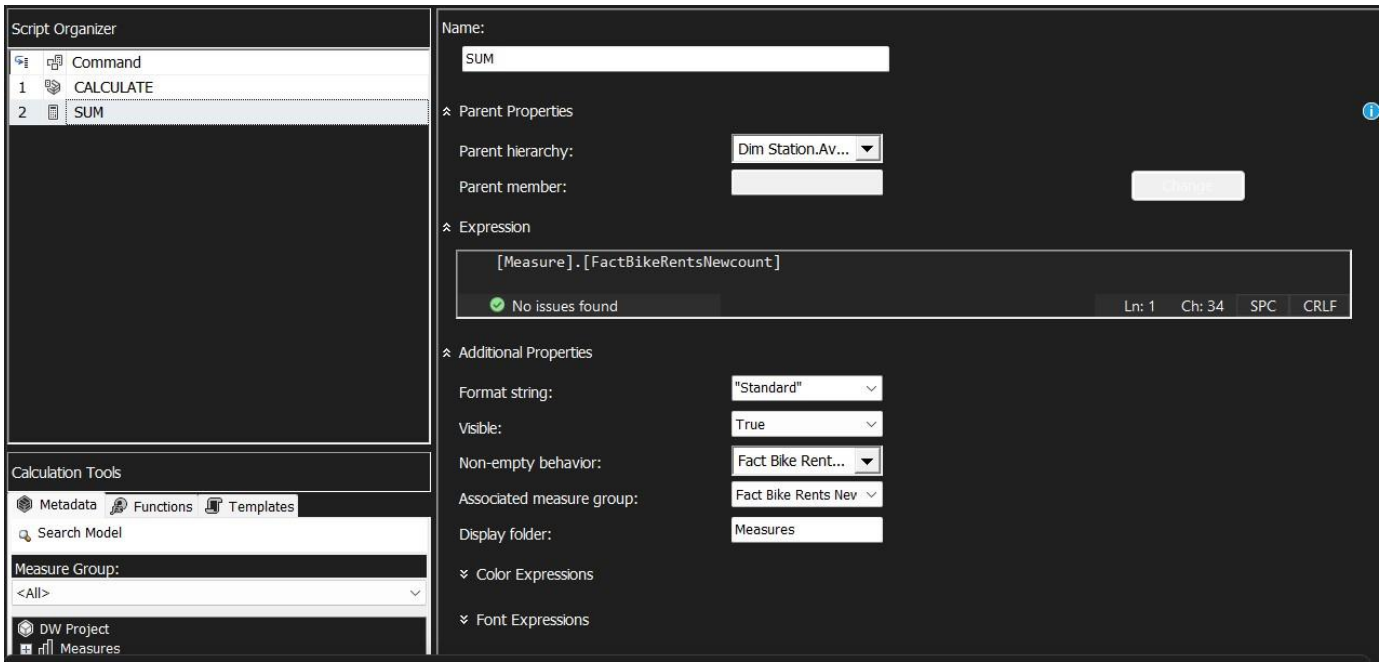




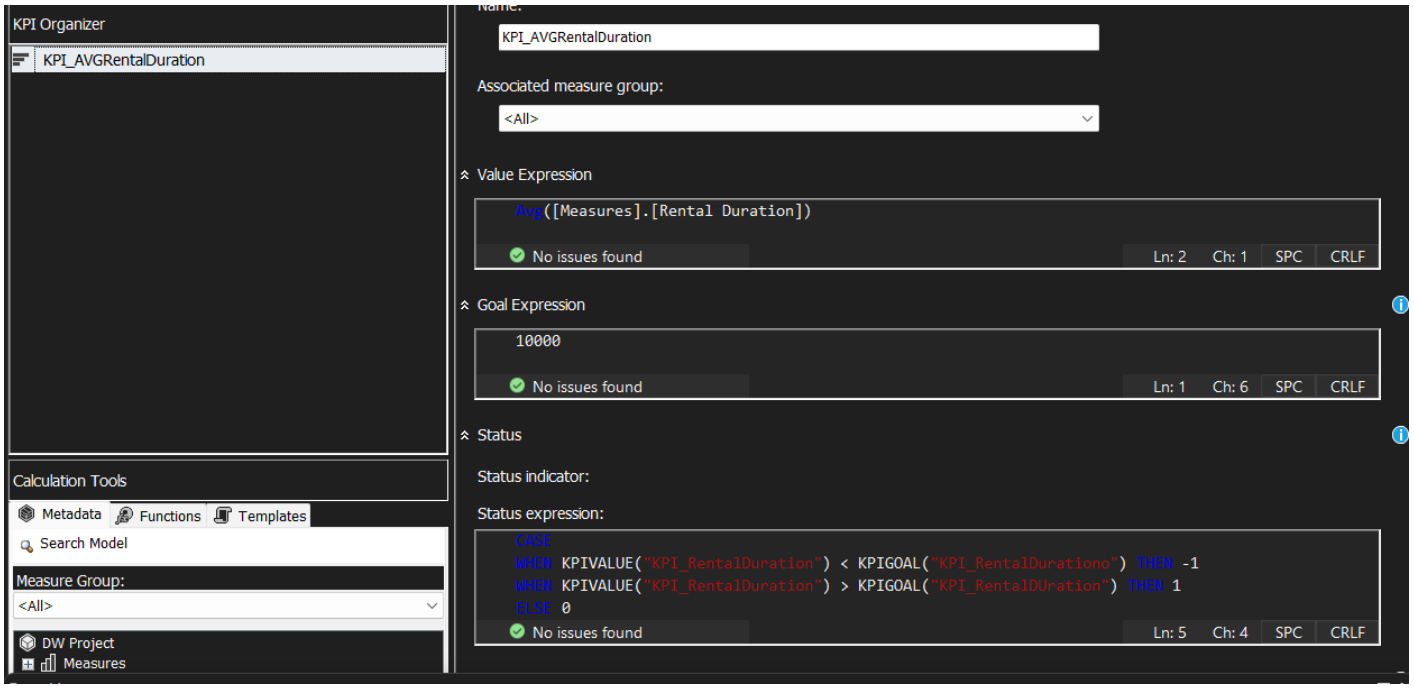
2.4 Cube details

Prepare a short description and a screenshot for all additional mechanisms utilised, e.g., calculations, KPIs, aggregations, partitions, perspectives.

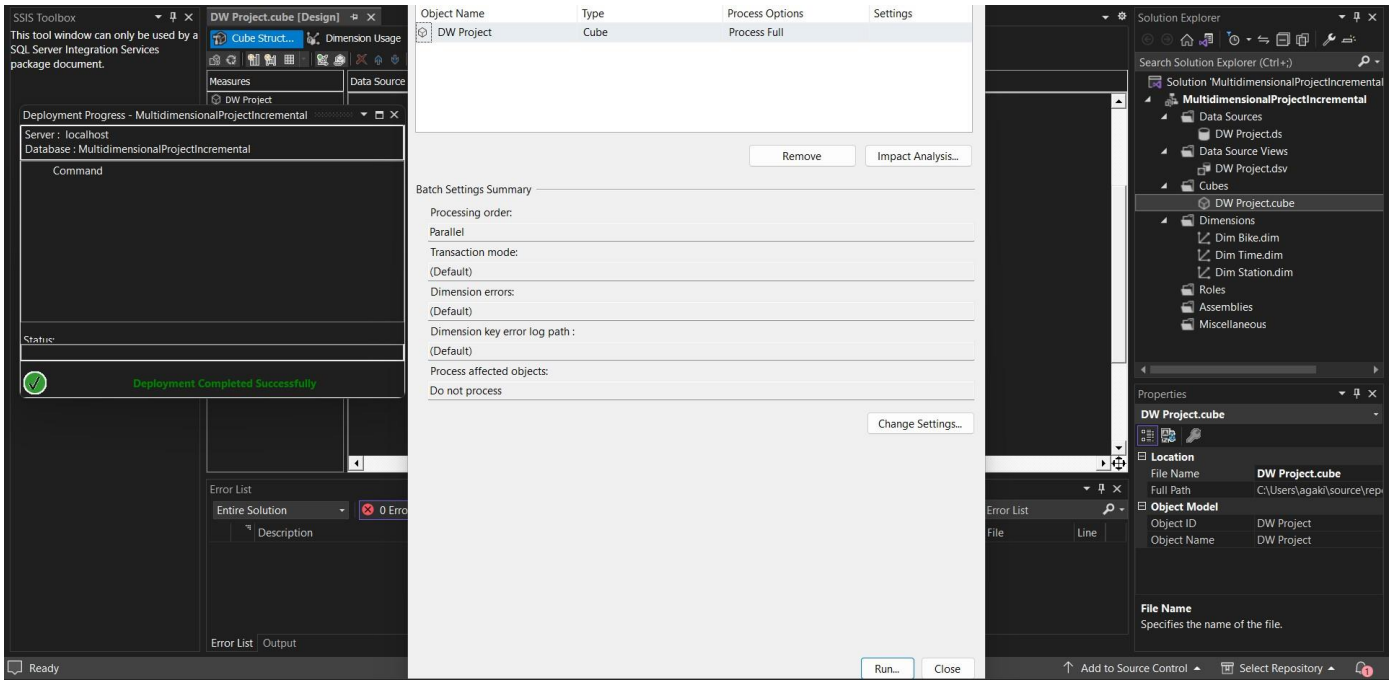
…

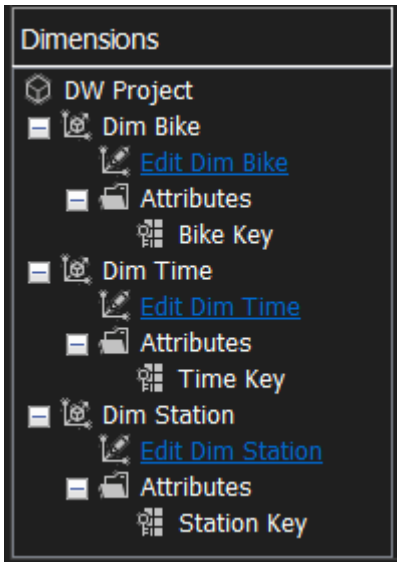


Data Warehouses - Report MP - 3



Cube Incremental





2.5 Process results

Prepare a screenshot from successful completion of cube processing and deployment process, and a screenshot from cube browser (with an exemplar query of your choice – make sure to capture cube's structure).

…

General Conclusions:

Use this section to provide your general conclusions, please share basic information about encountered problems and sketch your approach to handle them:

…

Remarks:

- *A report without final conclusions will not be checked and results in a negative score!*
- *You should use MS SQL SERVER 20XX, SQL Server Integration Services (SSIS) and SQL Server Analysis Services (SSAS); if needed you can use additional scripts or software to develop the ETL solution (usage of dedicated libraries or tools is encouraged)*