**FA23: DATA-225 Sec 11 - Db Systems for Analytics**

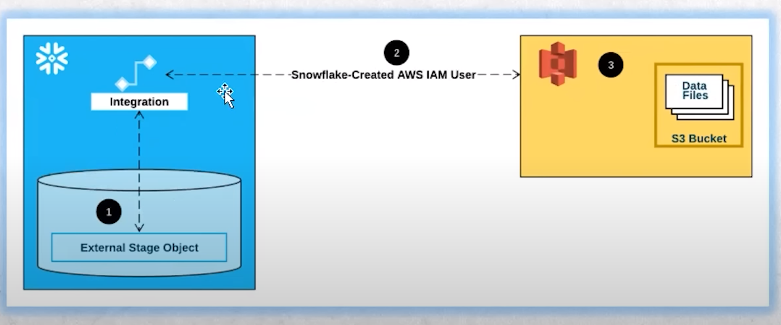
**Homework - - 4**

**Name:- Prayag Nikul Purani**

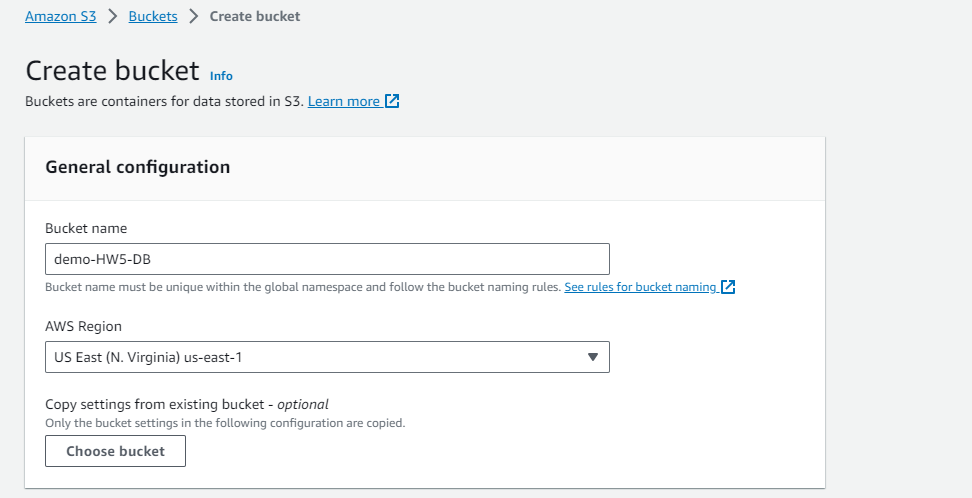
**SJSU Id:- 017416737**

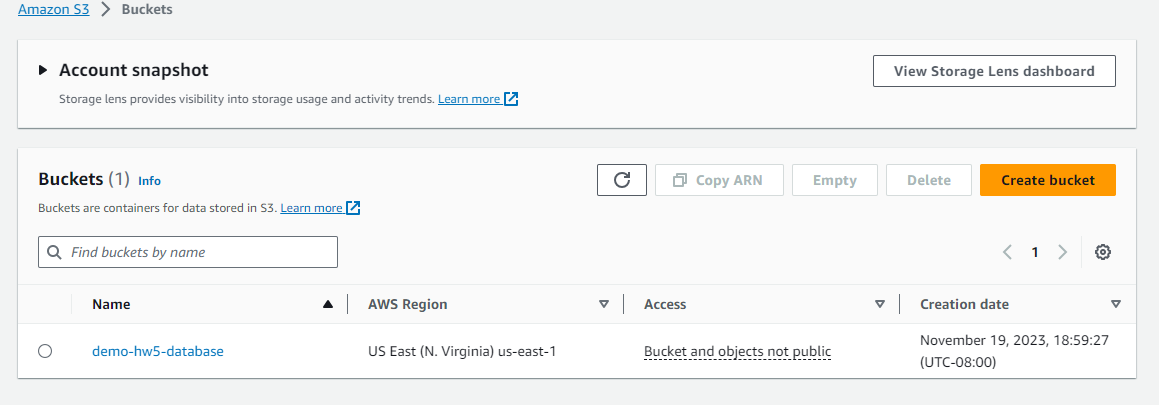
**Question 1 a: Snowflakes data-ware house**

**The data-set is taken from the** [**https://brightdata.com/cp/datasets/browse?id=hl\_1876193f**](https://brightdata.com/cp/datasets/browse?id=hl_1876193f) **which real-estate data set of from Poland**

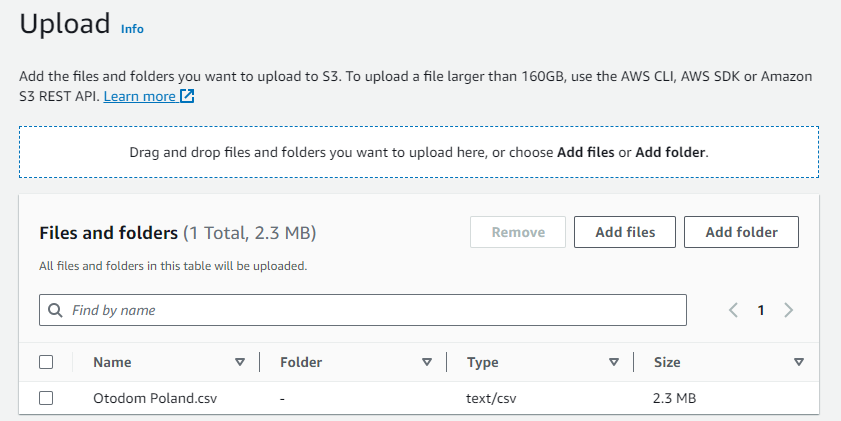


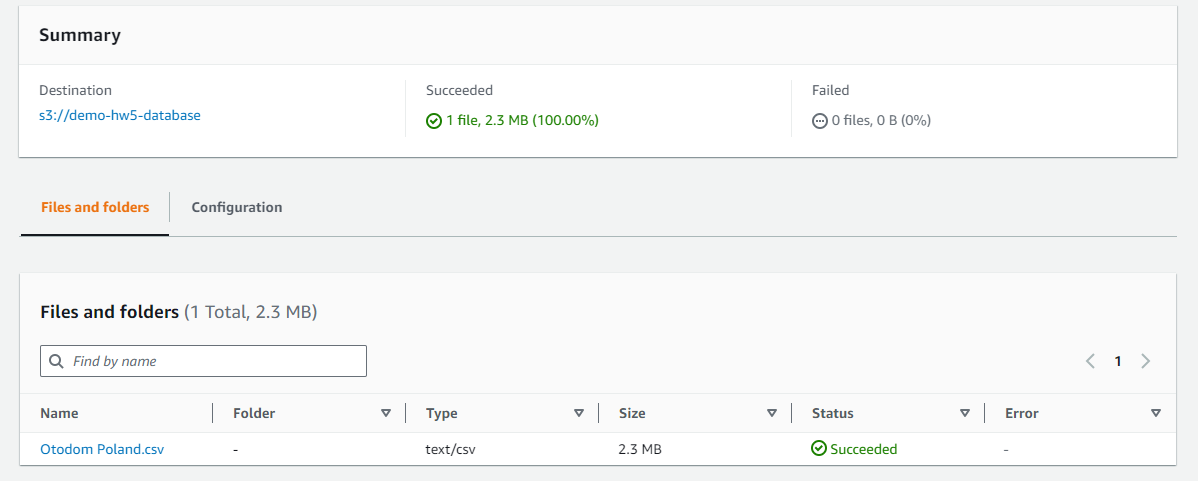
**Step 1: Creating bucket in AWS S3**

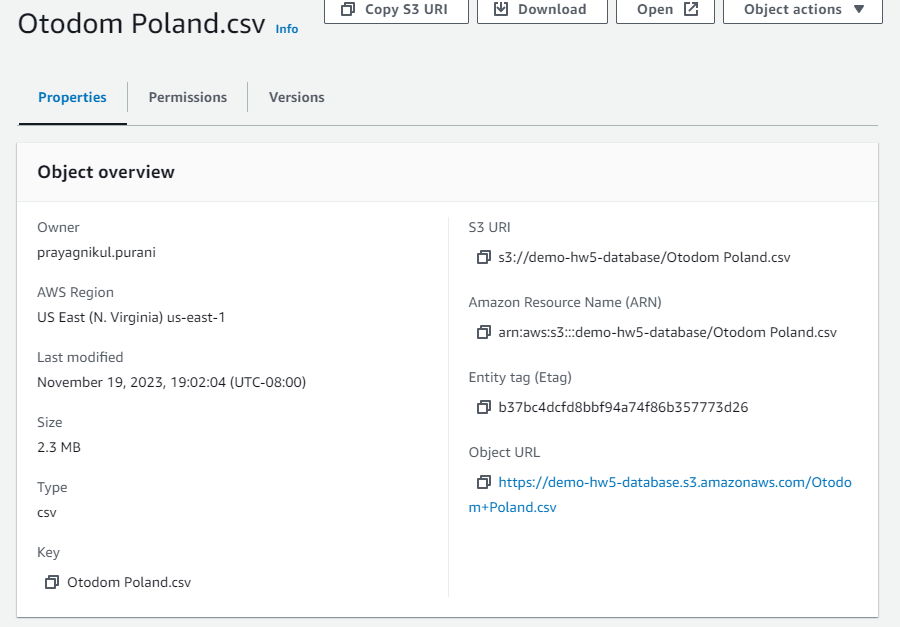




**Step 2: Uploading the file**







S3 Bucket Name: s3://demo-hw5-database/Otodom Poland.csv

**Step 3: Creating a Role ARN**

* **Polices**

{

"Version": "2012-10-17",

"Statement": [

{

"Sid": "Statement1",

"Effect": "Allow",

"Action": [

"s3:PutObject",

"s3:GetObject",

"s3:GetObjectVersion",

"s3:DeleteObject",

"s3:DeleteObjectVersion"

],

"Resource": [

"arn:aws:s3:::demo-hw5-database/\*"

]

},

{

"Sid": "Statement2",

"Effect": "Allow",

"Action": [

"s3:ListBucket",

"s3:GetBucketLocation"

],

"Resource": [

"arn:aws:s3:::demo-hw5-database"

],

"Condition": {

"StringLike": {

"s3:prefix": "\*"

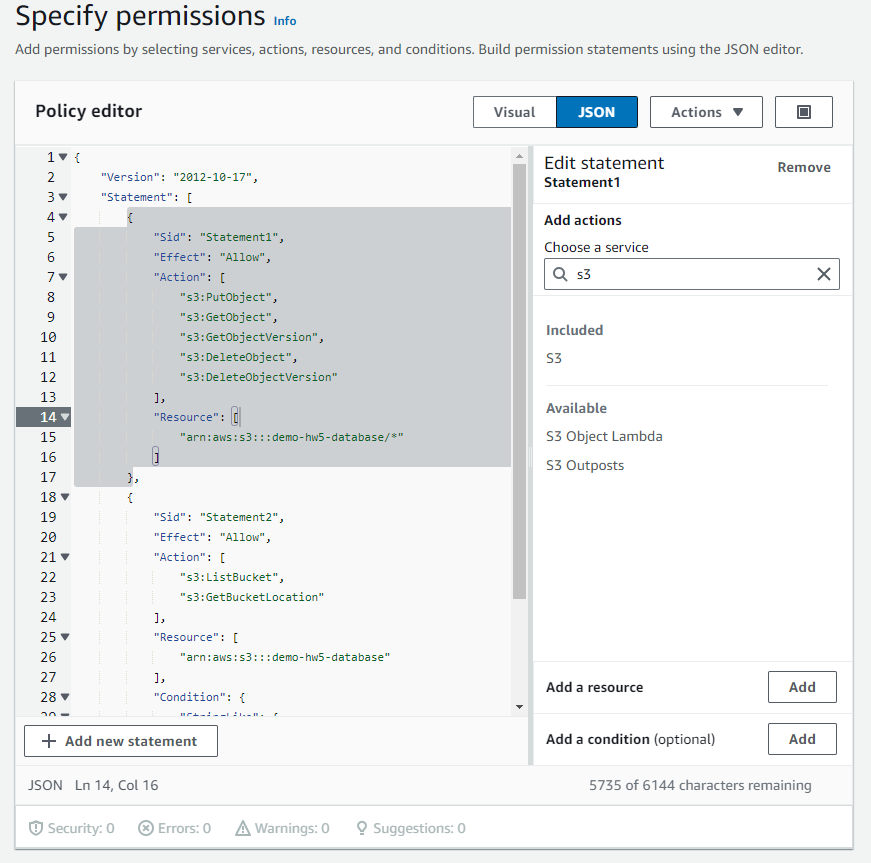
}

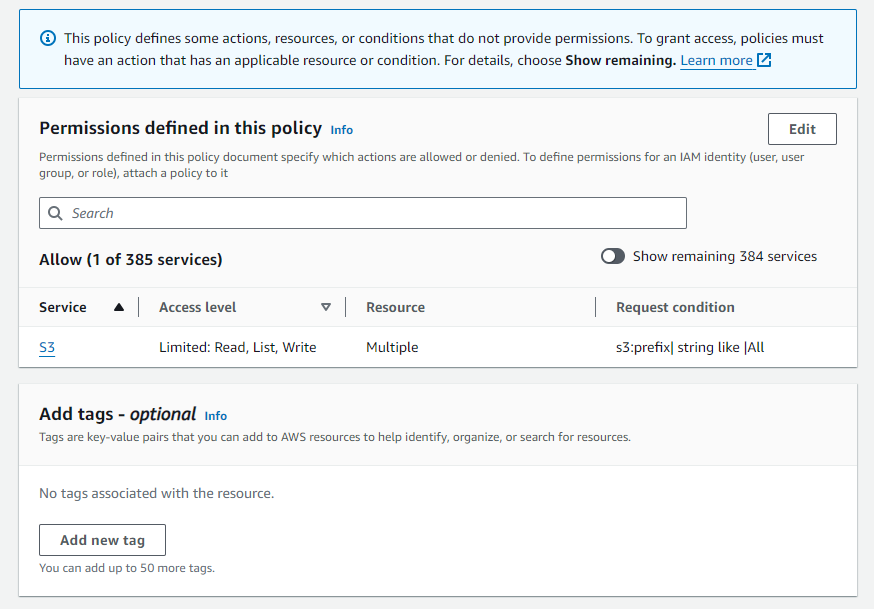
}

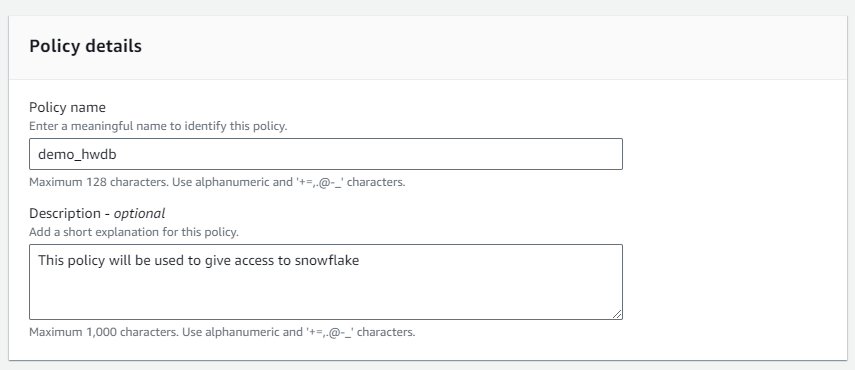
}

]

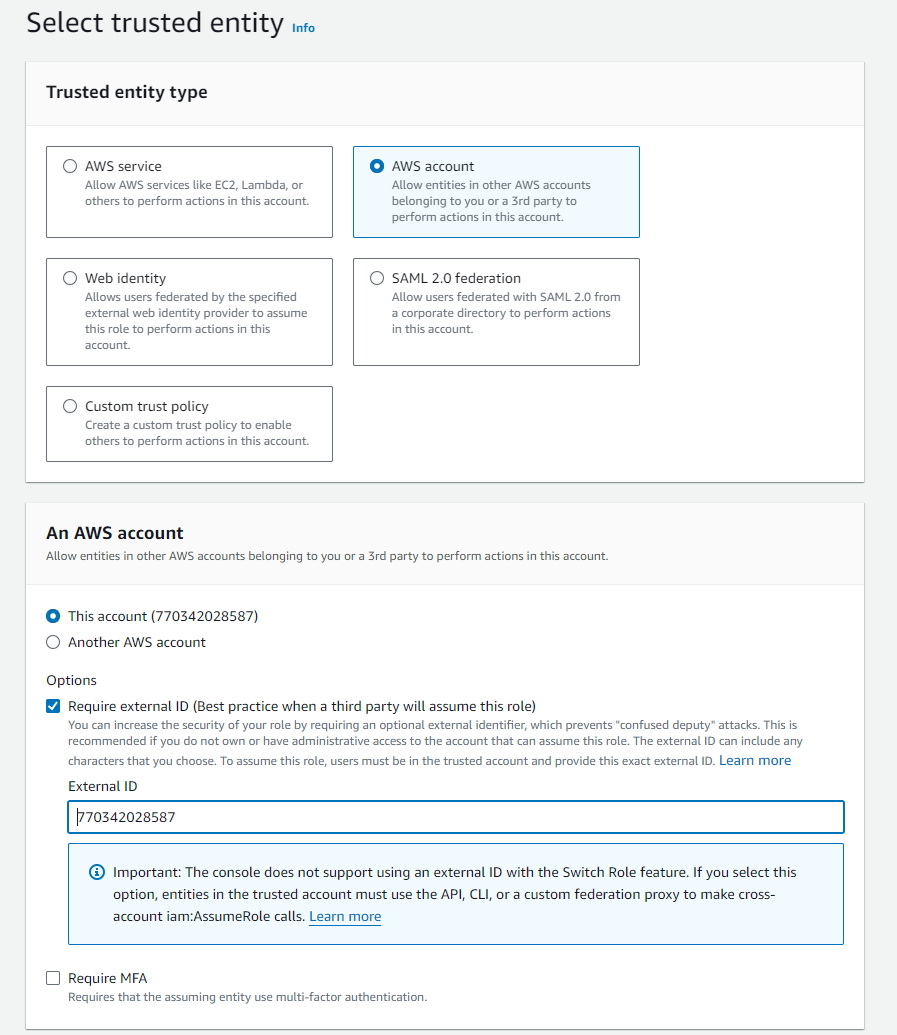
}

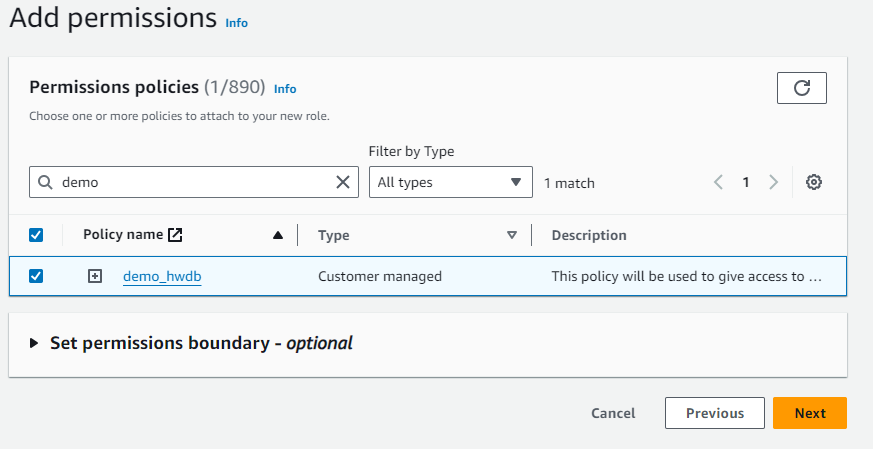


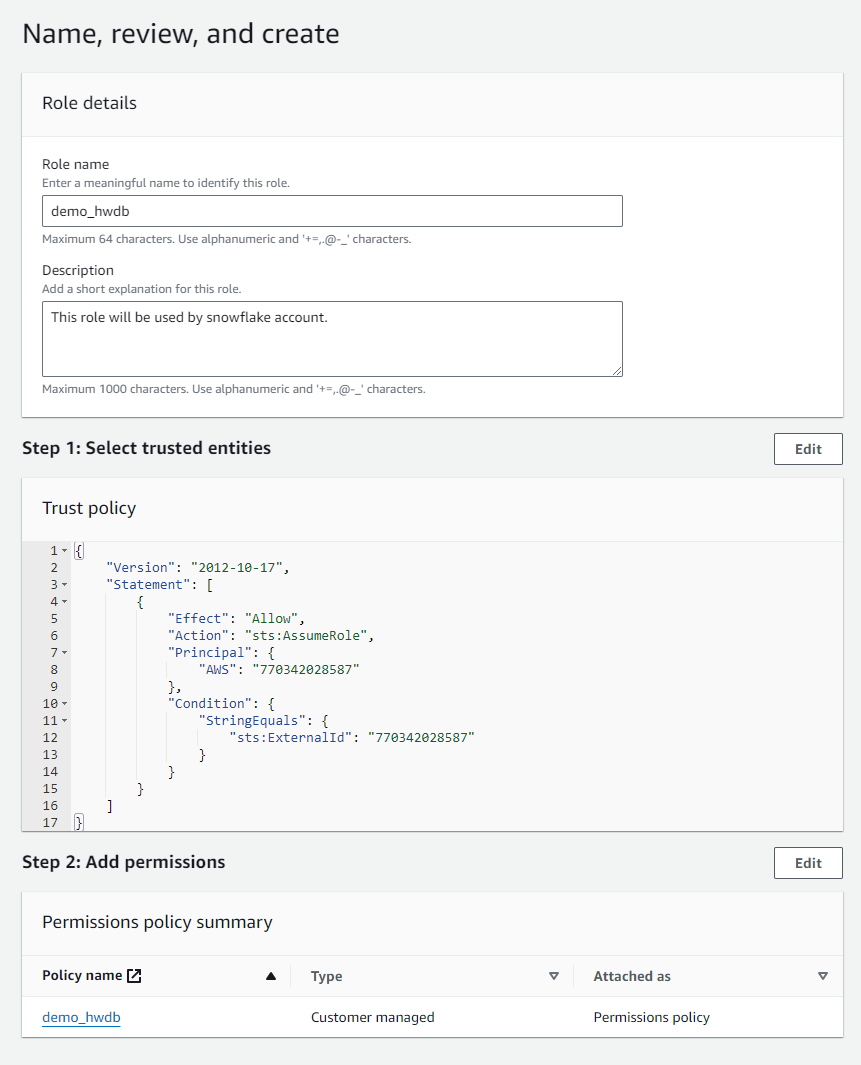


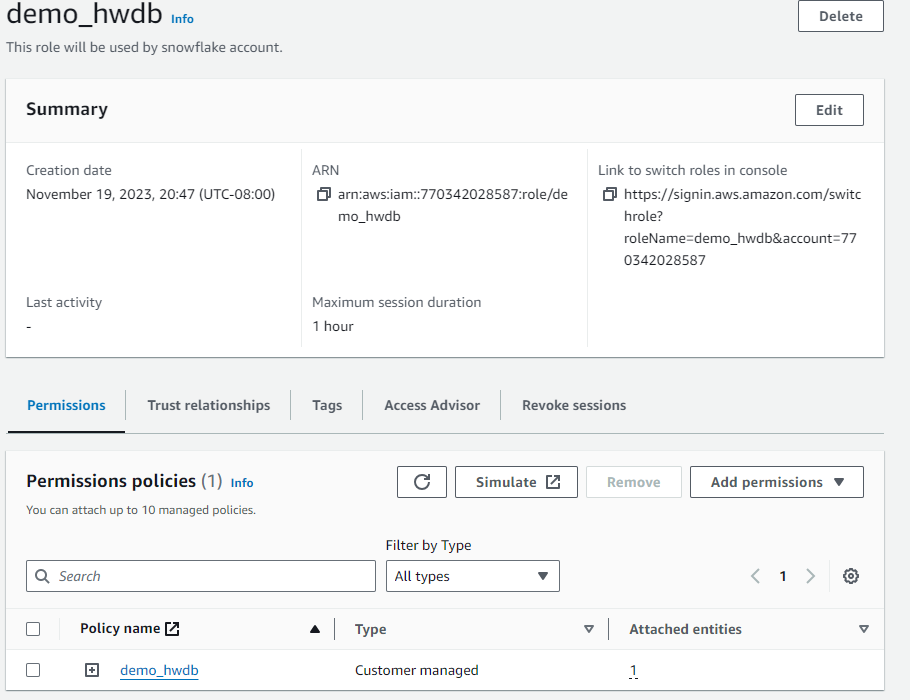


* **Roles**







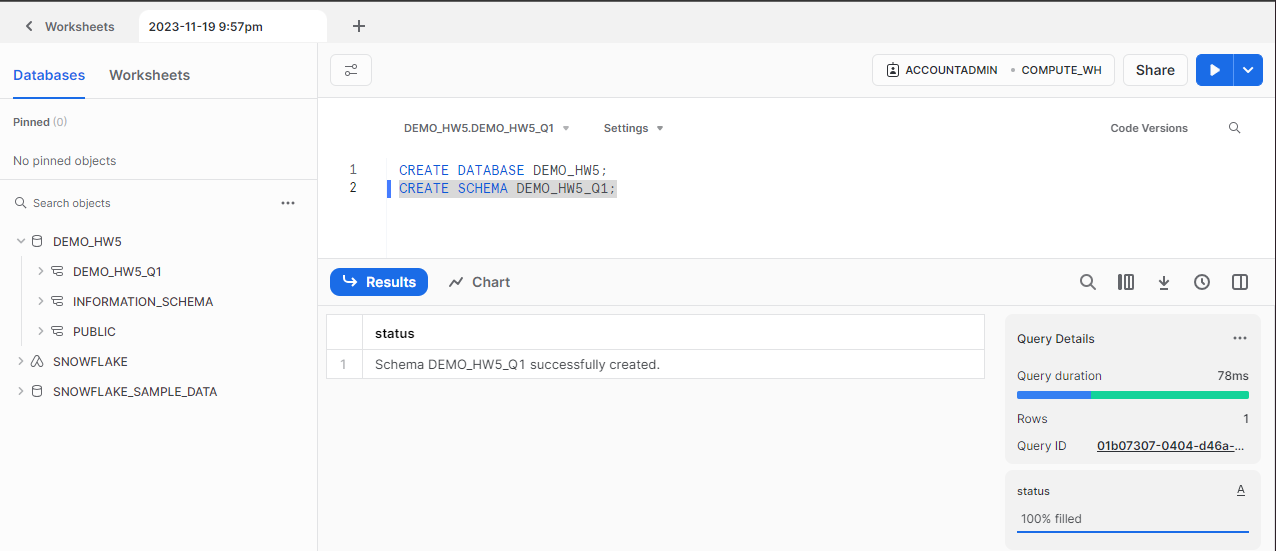


Role ARN: arn:aws:iam::770342028587:role/demo\_hwdb

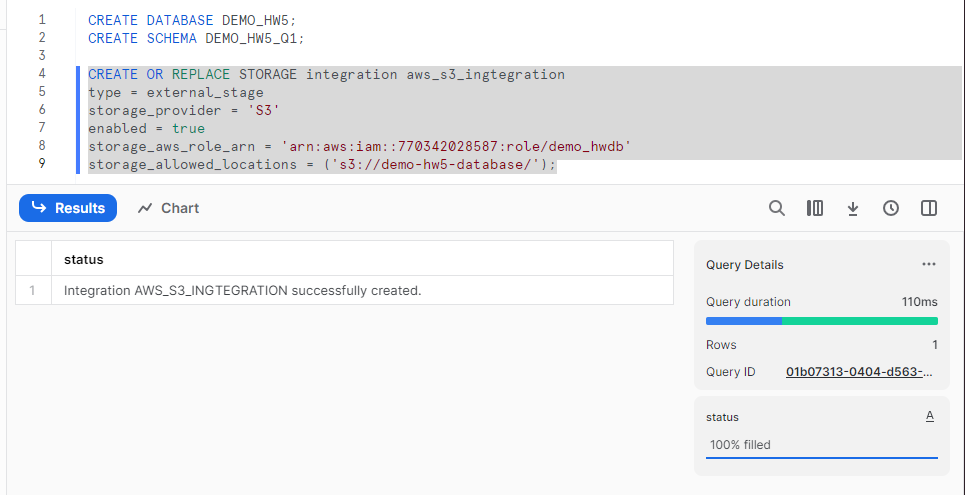
Account no: 770342028587

**Everything is done from AWS the only thing left is to do the snowflake part**

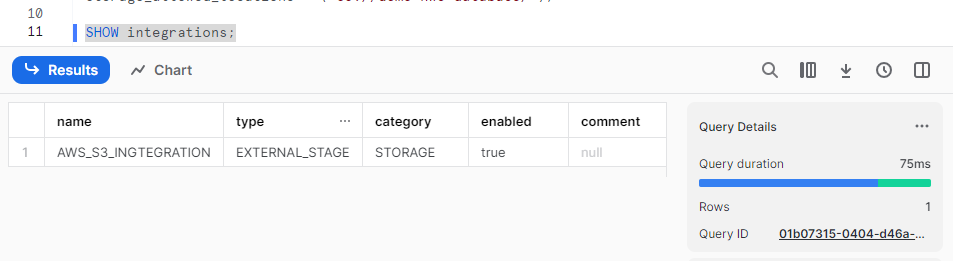
**Step 4: Creating the database and schema in snowflake.**

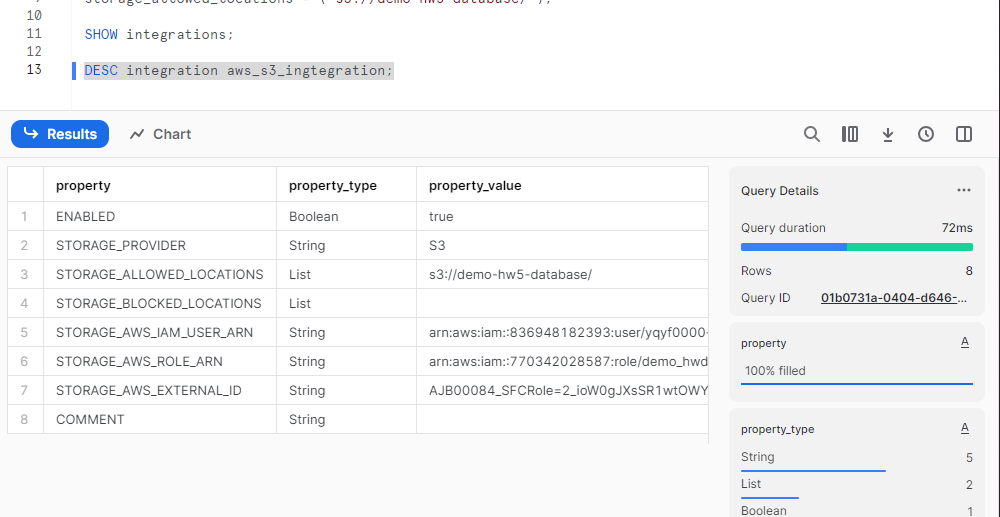
****

**Step 5: Integration of snowflake and AWS**

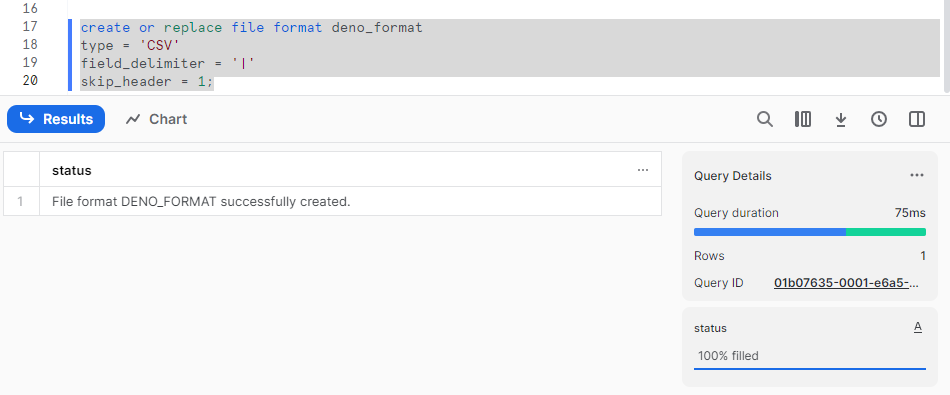


Integration with aws is successful.

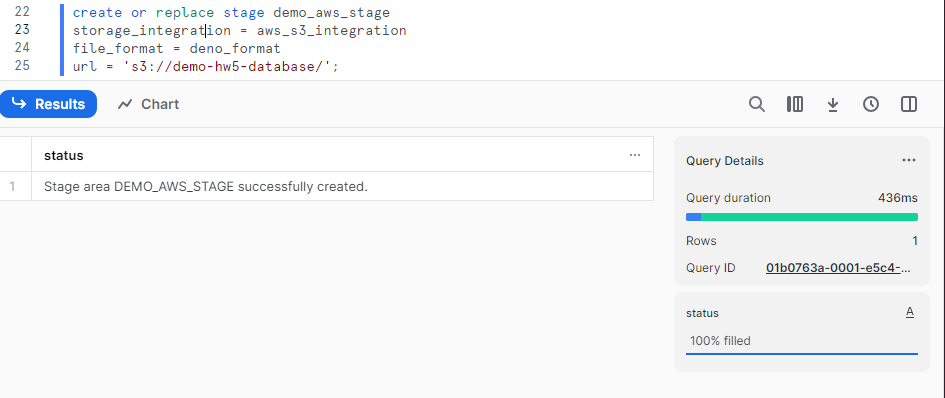


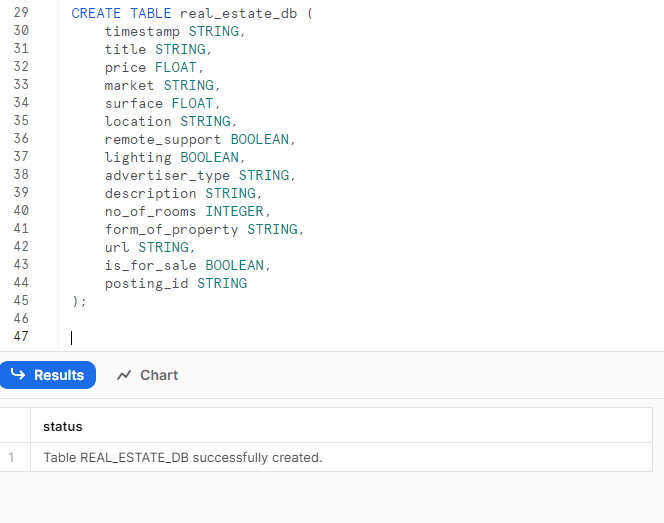


**Creating the file format**

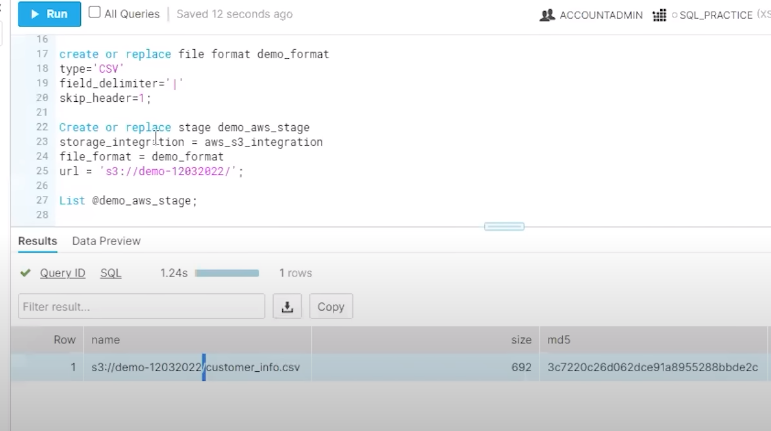


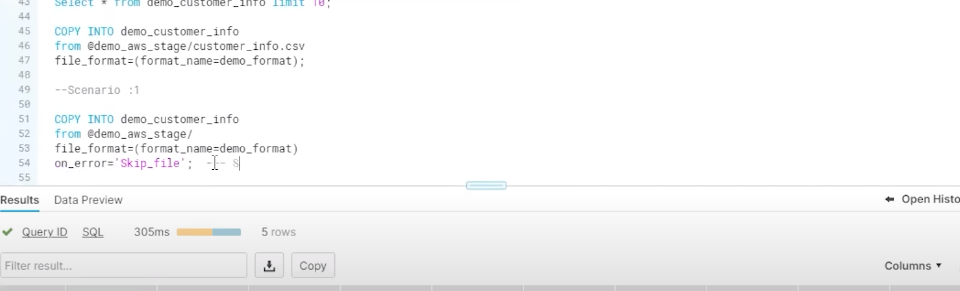
**Creating the stage:**



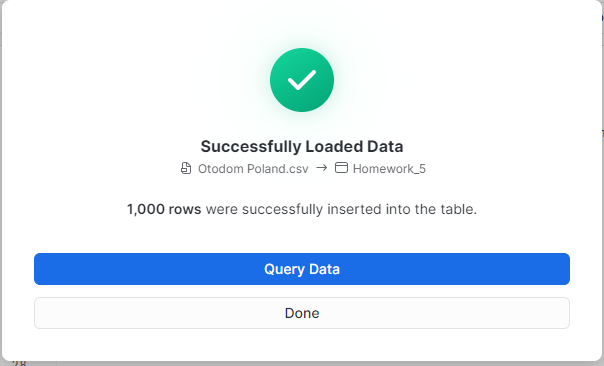
****

**Now we will check whether the stage is connected or not.**

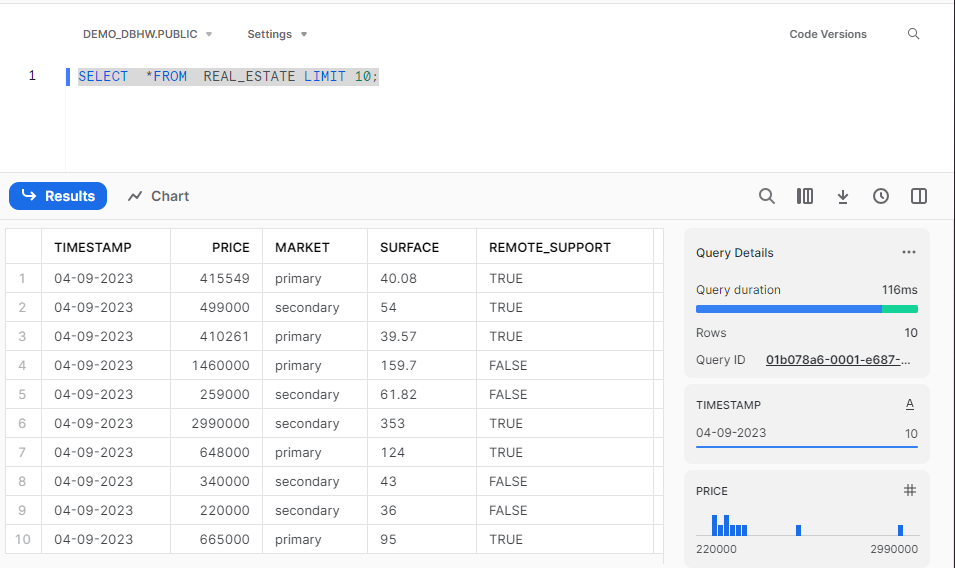
****

****

**Importing the data**

****

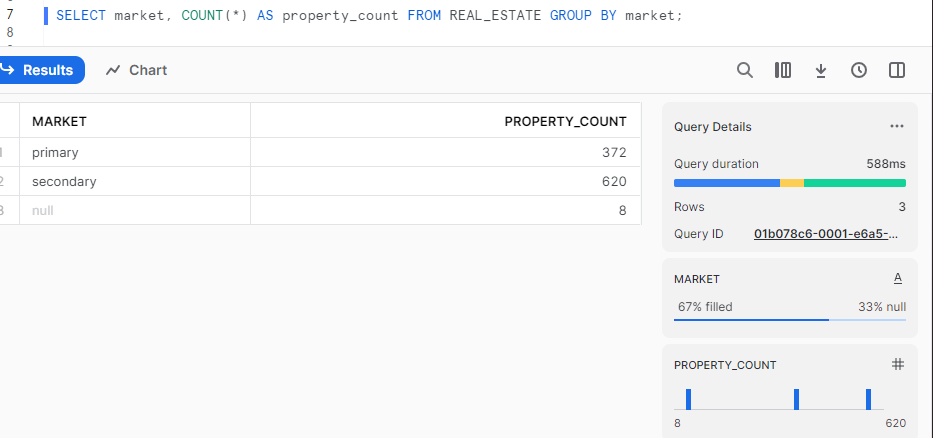
Know we will do some query on the data to get insights from the data.

****

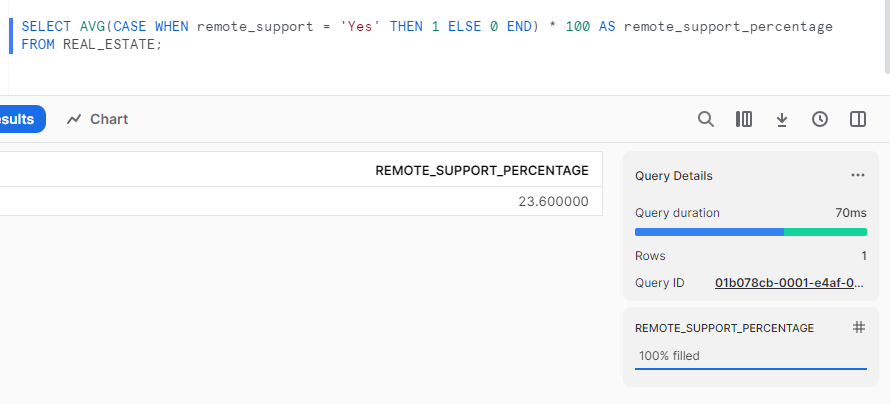
Query 1



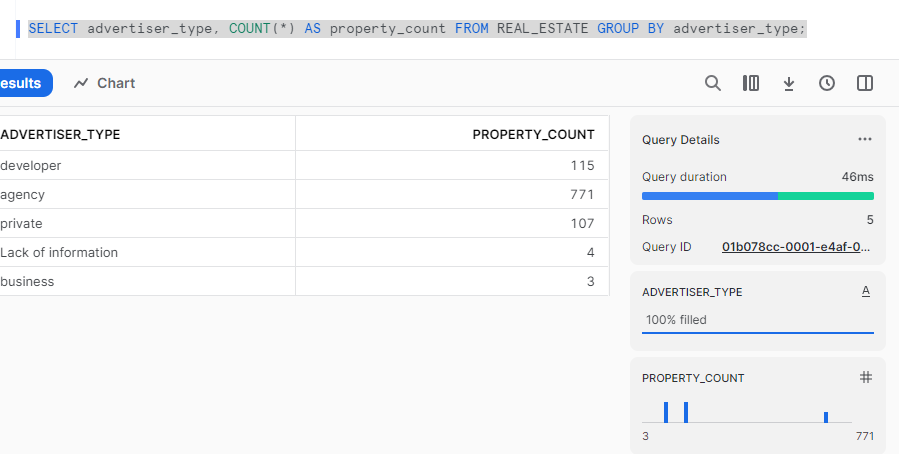
Query 2



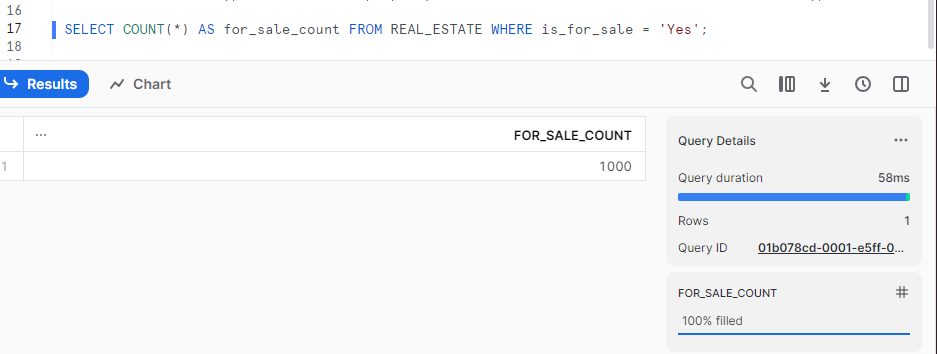
Query 3



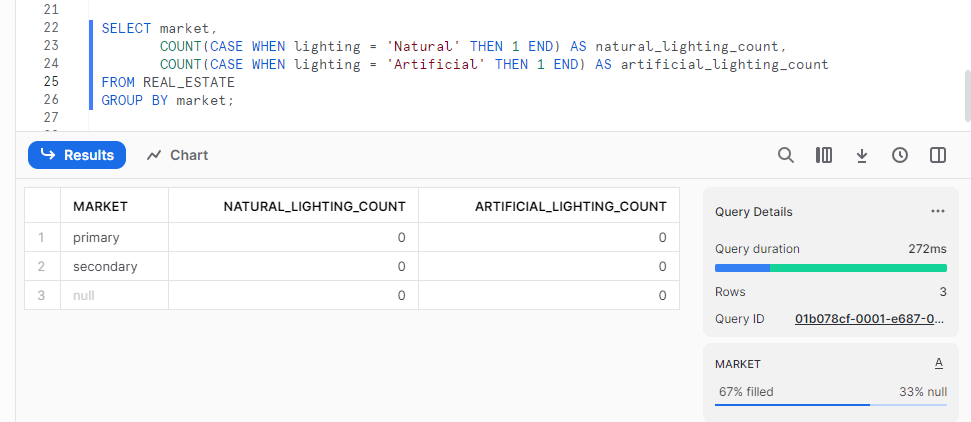
Query 4



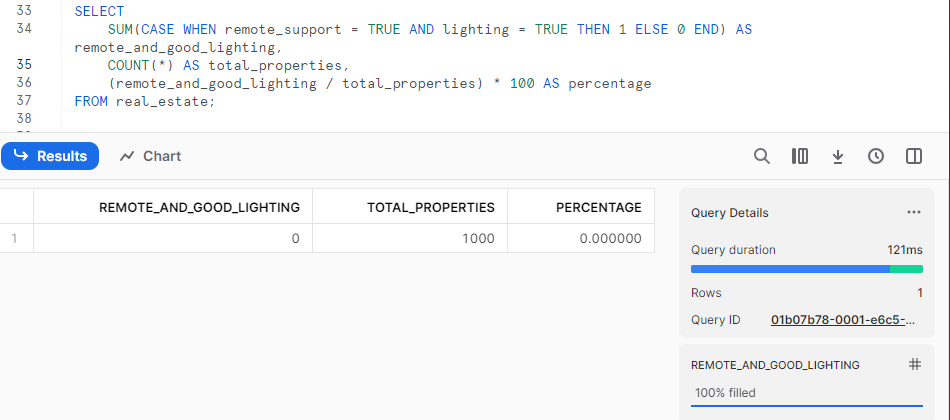
Query 5



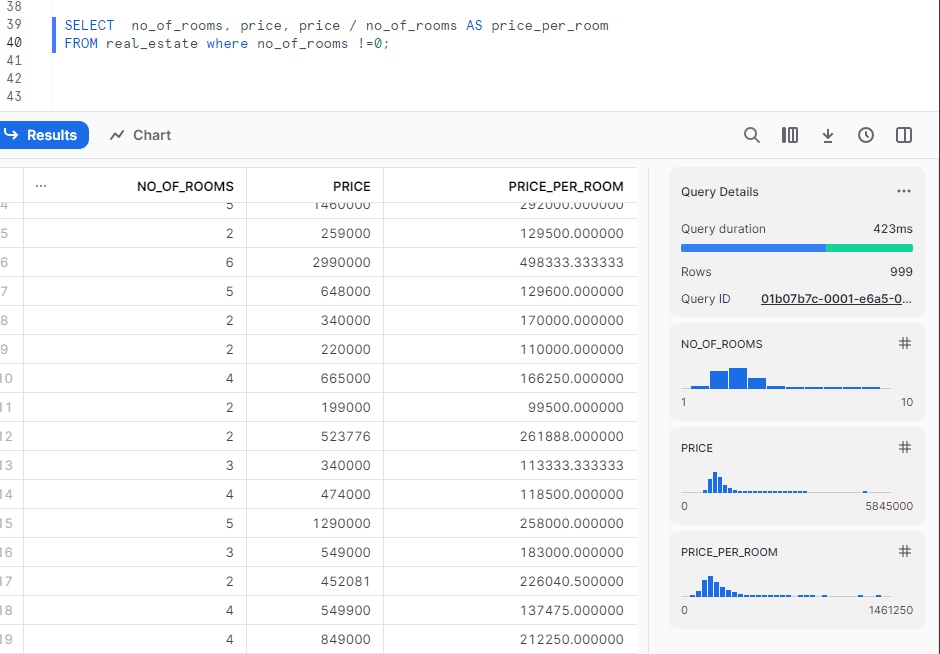
Query 6



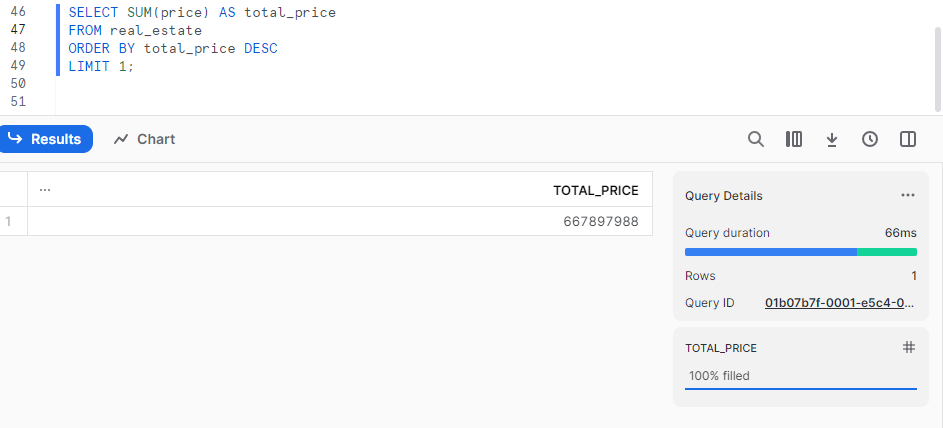
Query 6



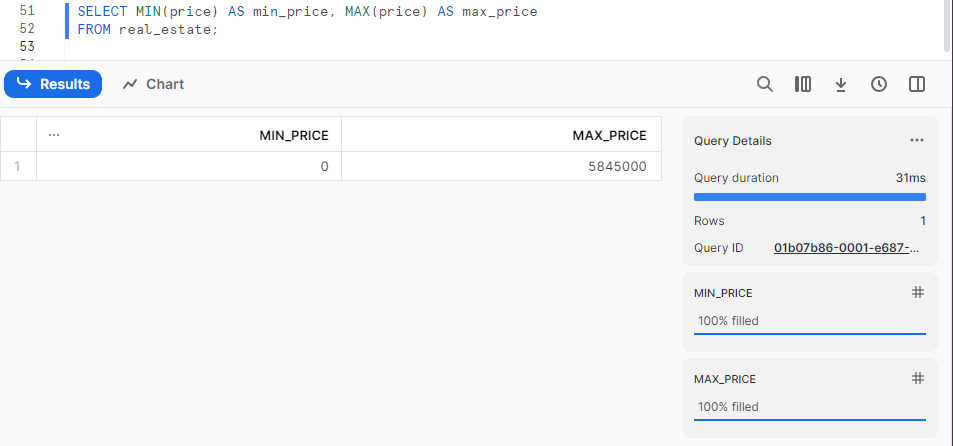
Query 7



Query 8



Query 9



**CODE:**

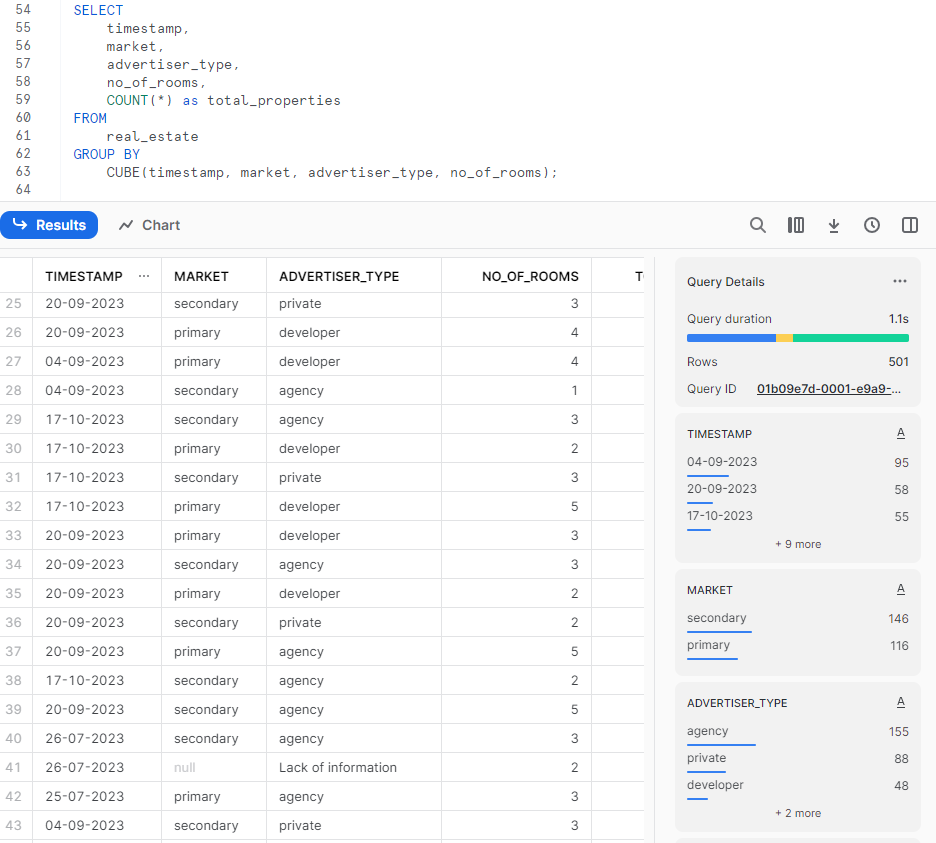
I have added the worksheet from the snowflake in which all the code are present as we can’t download the worksheet from snowflake in free version so, I have copied all the code to the notepad and have upload it’s attached with files

**Question 1 b:**

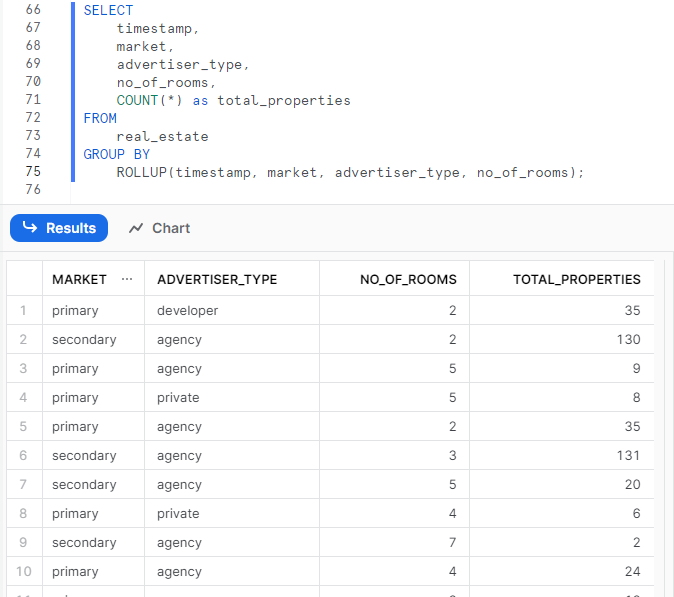
**Medium link:-**

[**https://medium.com/@prayagpurani11/first-experience-with-snowflakes-397d065579d3**](https://medium.com/@prayagpurani11/first-experience-with-snowflakes-397d065579d3)

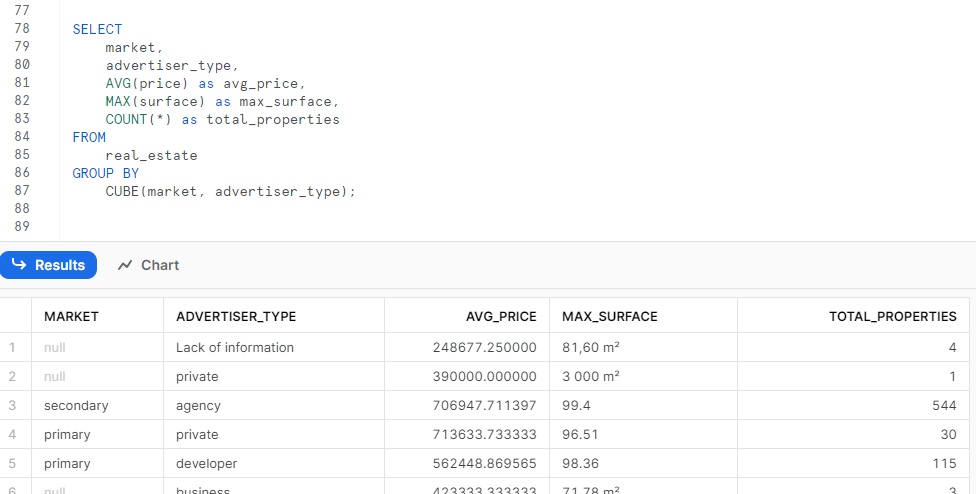
**Query 1:**

****

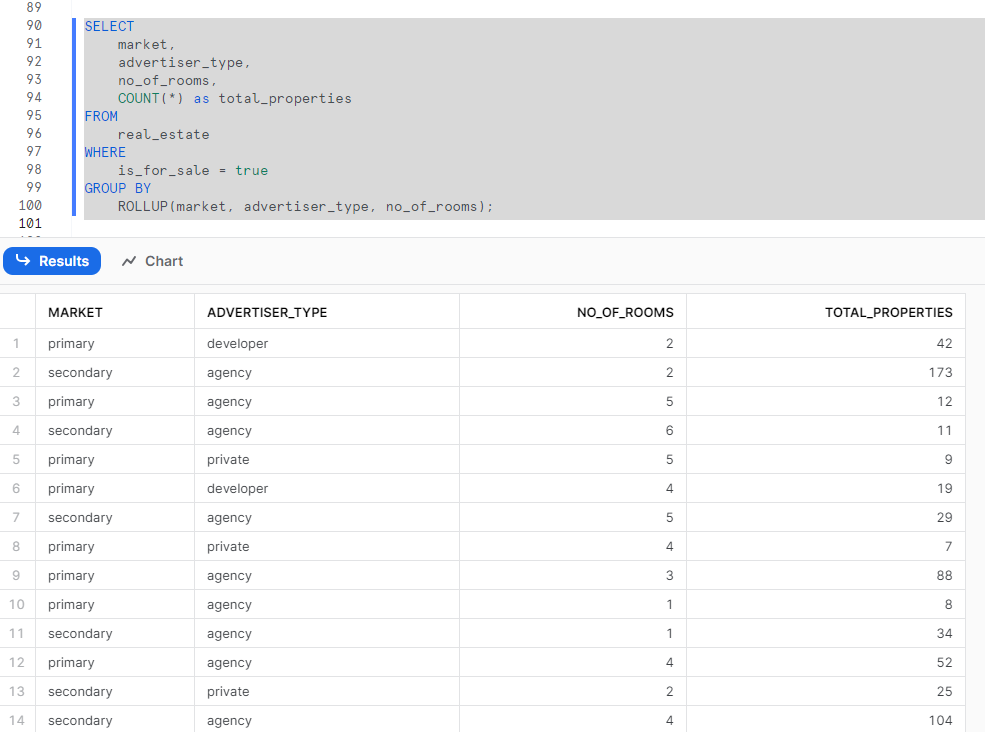
**Query 2:**

****

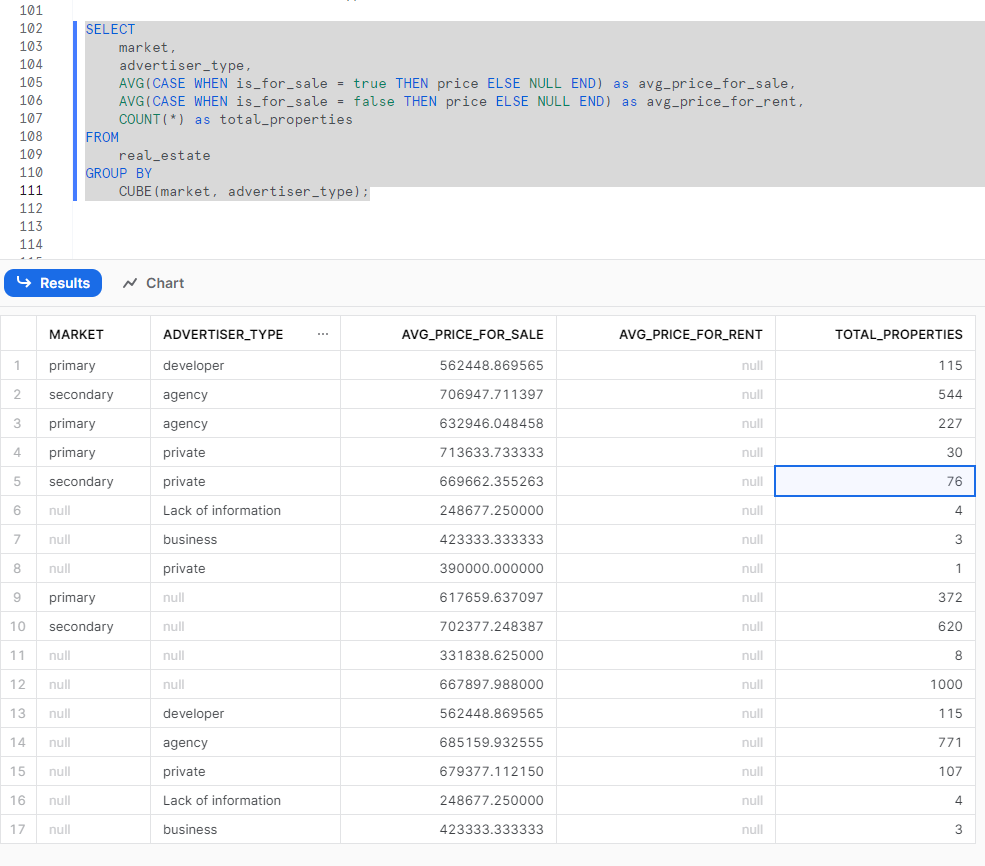
**Query 3:**

****

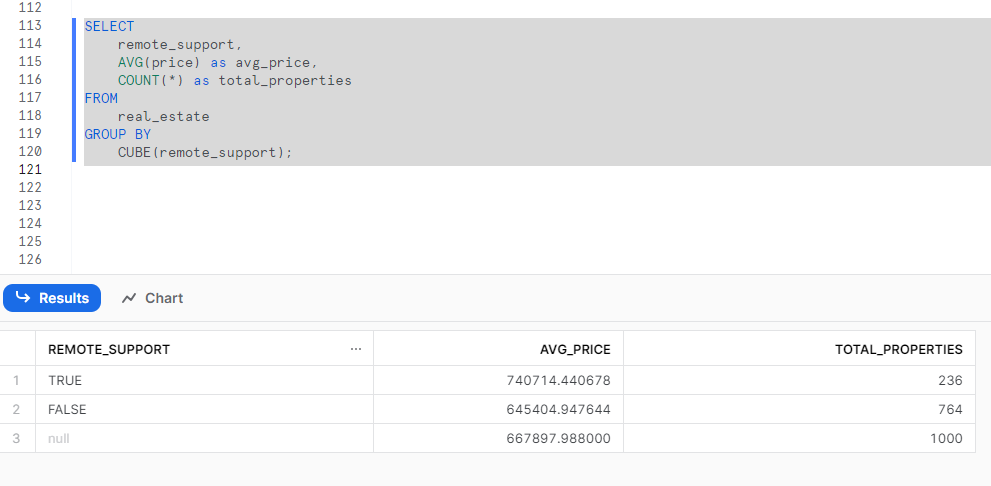
**Query 4:**

****

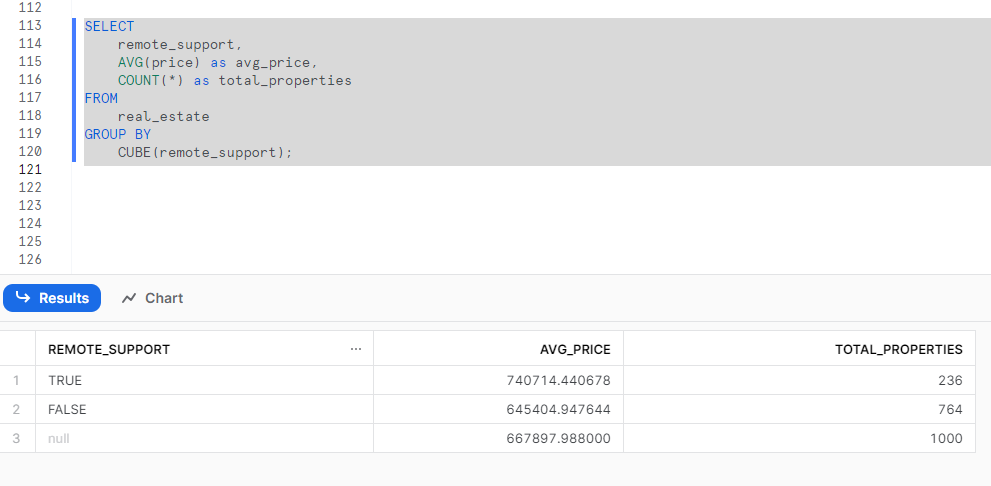
**Query 5:**

****

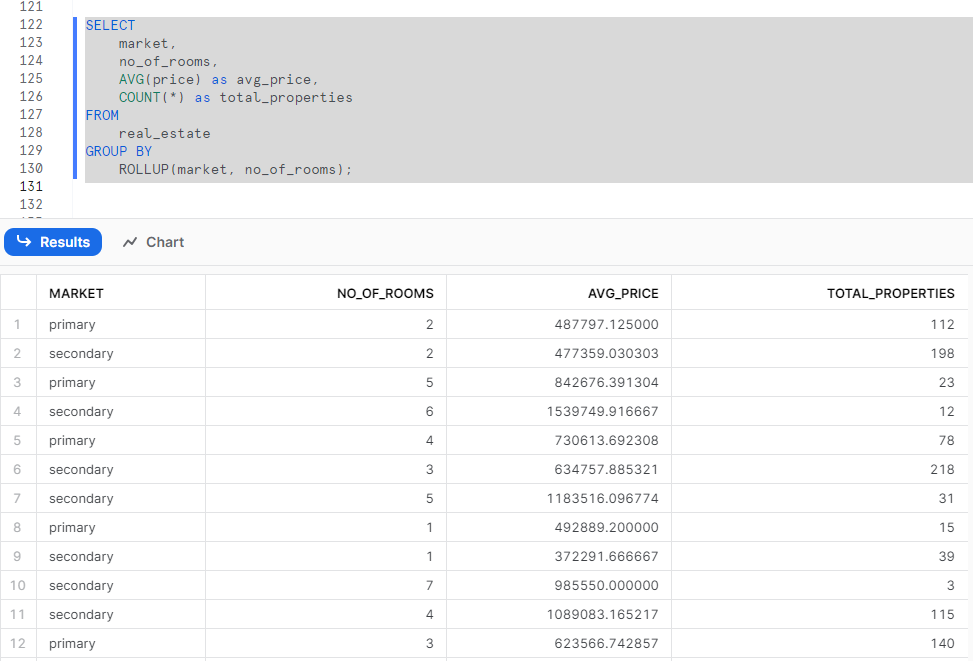
**Query 6:**

****

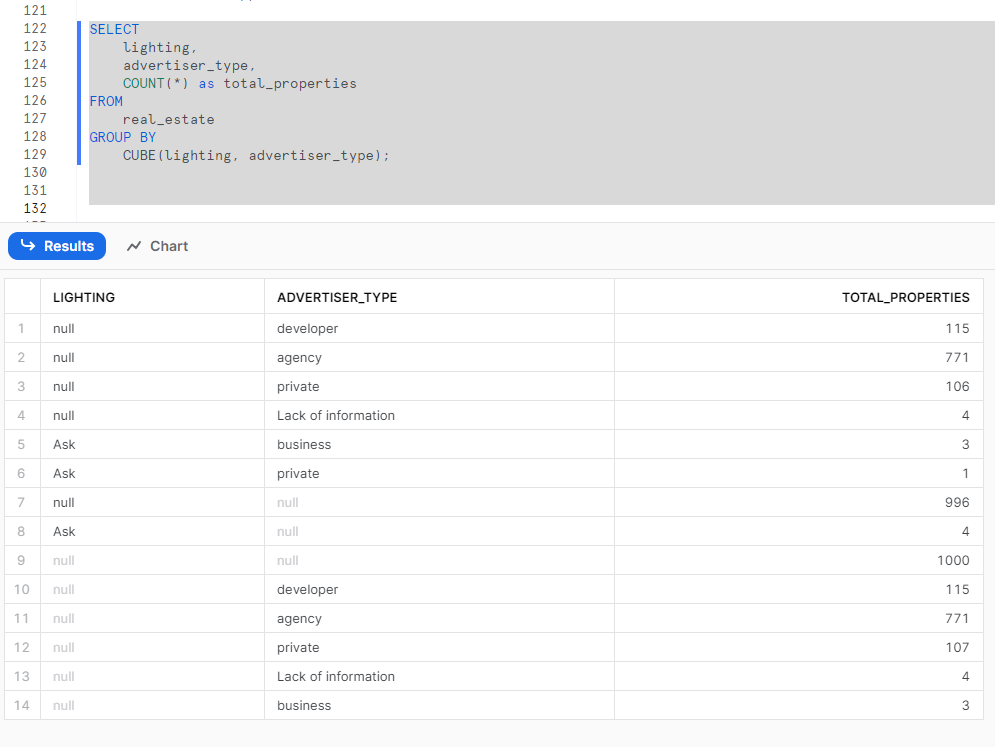
**Query 7:**

****

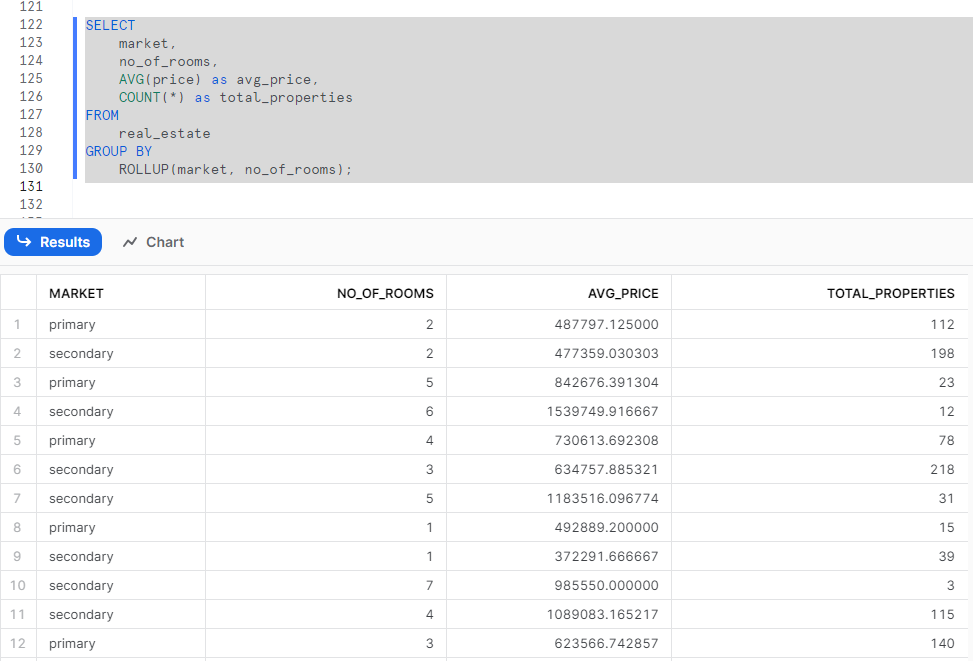
**Query 8:**

****

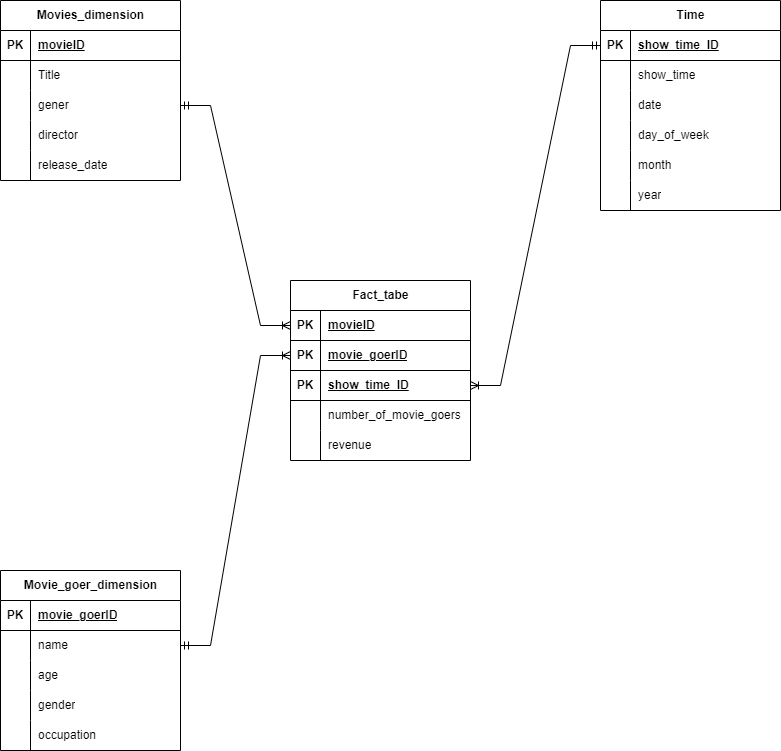
**Query 9:**

****

**Query 10:**

****

**Question 2 A:**

****

Within this star schema:

The measurements (number of movie goers and revenue) and foreign keys referencing the dimensions are contained in the Fact\_table table.

Descriptive information on movies, moviegoers, and time is contained in the Dimension\_Movie, Dimension\_MovieGoer, and Dimension\_Time tables, respectively.

The foreign key connections build the linkages between the fact and dimension tables**.**

**Question 2 b:**

SELECT

m.name, SUM(f.revenue) as total\_revenue

FROM

Fact\_table f

JOIN

Movies\_dimension m ON f.movie\_id = m.movie\_id

JOIN

Time t ON f.show\_time\_id = t.show\_time\_id

WHERE

t.year = 2020

GROUP BY

m.name;

**Justification:**

The Fact\_table database has the total income, while the Movies\_dimension table contains the name of the film. Using movie\_id as the common key, it combines the fact table with the movie dimension table. Using show\_time\_id as the common key, it also links the fact table with the time dimension table. The WHERE clause restricts the entries that appear in the results to those from 2020. The results are grouped by movie name under the GROUP BY clause. The end result is the total amount of money made from the sale of tickets for every film in 2020.

**Question 2 c:**

SELECT

movie, SUM(ts.ticket\_price) as total\_revenue

FROM

ticket\_sales ts

JOIN

Movie\_dimension m ON ts.movie = m.title

WHERE

ts.year = 2020

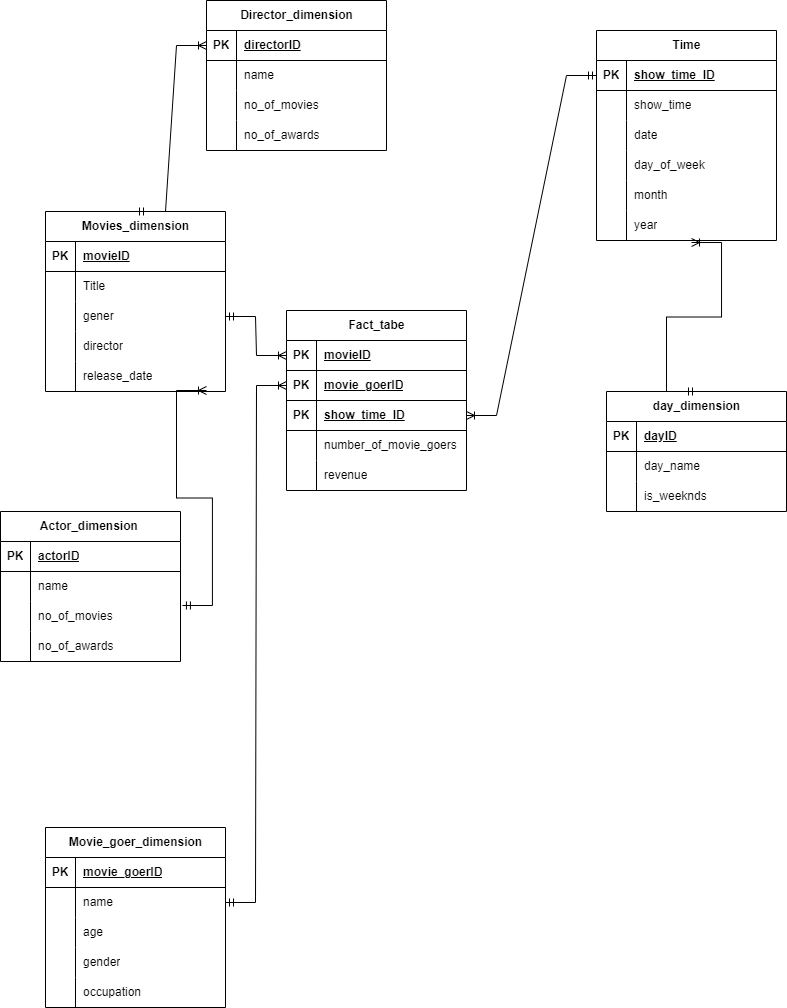
GROUP BY

movie;

**Justification:**

The total income field in the query is set to the sum of the ticket prices, and the movie column is taken from the ticket sales table. The WHERE clause restricts the entries that appear in the results to those from 2020. The movie column is used to group the results in the GROUP BY clause. The end result is the total amount of money made from the sale of tickets for every film in 2020.

**Question 2 d:**

****

**Query 1:**

SELECT

mg.occupation,

SUM(f.number\_of\_movie\_goers) as total\_movie\_goers

FROM

Fact\_MovieGoerTime f

JOIN

Dimension\_MovieGoer mg ON f.movie\_goer\_id = mg.movie\_goer\_id

GROUP BY

mg.occupation

ORDER BY

total\_movie\_goers DESC;

**Query 2:**

SELECT

m.genre,

SUM(f.number\_of\_movie\_goers) as total\_movie\_goers

FROM

Fact\_MovieGoerTime f

JOIN

Dimension\_Movie m ON f.movie\_id = m.movie\_id

GROUP BY

m.genre

ORDER BY

total\_movie\_goers DESC;

**Question 3 a:**

Because columnar databases are so good at performing analytical tasks, they have become more and more common in the business, particularly when used to construct data warehouse solutions. The following five points highlight the importance of adopting columnar databases in data warehousing:

1. **Performance Improvement for Analytic Queries:**

Remark: Data is stored in columnar databases column-wise as opposed to row-wise. For analytical queries including aggregations, filtering, and other actions on a subset of columns, this architecture is especially helpful.As an illustration, the columnar database Vertica is renowned for its capacity to greatly improve the efficiency of intricate analytical queries. Vertica has been widely used in sectors like banking and telecoms where quick and effective data analysis is essential.

1. **Compression Techniques for Storage Efficiency:**

Observation: Columnar databases frequently use compression strategies designed especially for columns, which results in significant storage reductions. This is so that column values in a database may be compressed more successfully because they often have comparable data types.

Columnar storage with compression is used by the well-known cloud-based data warehousing system, Amazon Redshift. This is advantageous for enterprises with variable data quantities, such as e-commerce and online shopping, since it enables firms to handle and analyze enormous datasets more economically.

1. **Scalability for Growing Data Volumes:**

Remark: Columnar databases are ideal for managing massive and expanding data volumes because of their horizontal scalability architecture. For businesses handling large volumes of data, including e-commerce, healthcare, and online advertising, scalability is essential.

Columnar storage is used by Google BigQuery, a serverless, highly scalable data warehouse. As their data expands, this enables enterprises to smoothly scale their analytical workloads. BigQuery is used by businesses in industries such as internet advertising to do real-time analytics on large datasets.

1. **Optimized for Aggregations and Reporting:**

Observation: Columnar databases are best suited for reporting and aggregations, which are common components of analytical tasks. During query execution, the ability to read only the columns that are required speeds up the retrieval and processing of data.

An example of a cloud-based data warehousing platform that employs a columnar storage architecture with several clusters is Snowflake. Because of its nature, Snowflake is ideally suited for sectors like retail, where complicated data aggregations and reporting are standard needs.

1. **Integration with Analytical Tools and Ecosystem:**

Observation: Columnar databases are easy to connect with popular ecosystems and analytical tools. The deployment of analytics and business intelligence (BI) systems is made easier by this compatibility.

For instance: Online gaming and other businesses employ distributed NoSQL columnar databases like Apache Cassandra. Because of its good integration with analytical tools, businesses may examine user behaviour and enhance the game experience. High-performance and scalable applications require distributed storage, which meets the needs of columnar storage for analytical queries.

In conclusion, the industry's use of columnar databases for data warehousing has shown to be advantageous in terms of enhancing analytical query performance, attaining storage efficiency, guaranteeing scalability, enhancing reporting and aggregations, and simplifying interaction with analytical tools. These findings demonstrate the applicability and efficiency of columnar databases in meeting the intricate data analysis requirements of several businesses.

**Question 3 b:**

Because OLAP (Online Analytical Processing) technologies allow users to study multidimensional data interactively, they are essential to business intelligence. These are five key points about the different OLAP tools that are utilized in the market, such as Oracle OLAP, Microsoft SSAS (SQL Server Analysis Services), and IBM Cognos.

1. Multidimensional Modeling and Analysis Capabilities
2. Performance Optimization through Aggregation and Caching
3. Integration with BI Platforms for Seamless Reporting
4. Support for Both MOLAP and ROLAP Architectures
5. Advanced Analytical Features and Predictive Analytics Integration

**IBM Congos:**

A popular tool for corporate intelligence and performance management, IBM Cognos offers a range of applications for scorecarding, reporting, analysis, and event and metric monitoring. Based on the analysis of IBM Cognos, the following five key findings are presented, along with particular instances from sectors where the product has had a notable influence.

1. Comprehensive Business Intelligence Suite
2. Integration with Data Warehousing Solutions
3. Self-Service Analytics Empowering Business Users
4. Mobile Business Intelligence for On-the-Go Access
5. AI-Powered Insights and Predictive Analytics

**Microsoft SSAS:**

Online analytical processing (OLAP) and data modeling may be accomplished with Microsoft SQL Server Analysis Services (SSAS), a multidimensional and data mining tool. These are five key takeaways from Microsoft SSAS, along with particular instances from sectors where it has had a notable influence:

1. Tabular and multidimensional models
2. Attachment to Microsoft BI Stack
3. Analytics Predictive and Data Mining
4. Optimizing performance and scalability
5. Data governance and security based on roles

In conclusion, Microsoft SSAS provides a flexible range of functionalities, such as support for both multidimensional.

**Oracle OLAP:**

A feature of the Oracle Database called Oracle OLAP (Online Analytical Processing) offers sophisticated analytical tools for multidimensional data processing. Here are five key insights about Oracle OLAP, along with concrete instances from sectors where it has had a notable influence:

1. Oracle Database Integration
2. Support for Cubes with Multiple Dimensions
3. Features for Advanced Aggregation and Calculation
4. Flexibility in Handling Big Datasets
5. Combining Business Intelligence Tools with Integration

All things considered, Oracle OLAP offers strong analytical capabilities, smooth connection with the Oracle Database, multidimensional cube support, sophisticated calculation functions, scalability for big datasets, and business intelligence tool compatibility. Because of these characteristics, Oracle OLAP is a useful tool for businesses in a variety of sectors that want to extract important insights from their multidimensional data.

**References:**

<https://www.youtube.com/watch?v=NXaRxwax9qA>

<https://docs.aws.amazon.com/AmazonS3/latest/userguide/creating-bucket.html>

<https://docs.aws.amazon.com/quickstarts/latest/s3backup/step-1-create-bucket.html>

<https://docs.aws.amazon.com/AmazonS3/latest/API/API_CreateBucket.html>

<https://www.youtube.com/watch?v=_DRklnnJbig>

<https://www.youtube.com/watch?v=Q_2CIivxVVs>

<https://community.claris.com/en/s/question/0D50H00006dtMgASAU/upload-file-from-container-to-s3>

<https://community.claris.com/en/s/question/0D50H00006dtMgASAU/upload-file-from-container-to-s3>

<https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies.html>

<https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles.html>

<https://www.strongdm.com/blog/aws-iam-roles-vs-policies>

<https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction_access-management.html>

<https://docs.aws.amazon.com/IAM/latest/UserGuide/reference_identifiers.html>

<https://www.arubanetworks.com/techdocs/central/2.5.1/content/gateways/vgw/vgw-orch-aws-crte-role-arn.htm>

<https://docs.panther.com/cloud-scanning/cloud-resource-attributes/aws/iam-role>

<https://snowflakewiki.medium.com/connecting-snowflake-to-aws-ef7b6de1d6aa>

<https://docs.snowflake.com/en/user-guide/data-load-s3-config>

<https://docs.snowflake.com/en/user-guide/data-load-s3-config-storage-integration>

<https://www.youtube.com/watch?v=m5-78NkUzZA>

<https://www.ibm.com/products/cognos-analytics?utm_content=SRCWW&p1=Search&p4=43700050328201997&p5=e&gclid=CjwKCAiAvJarBhA1EiwAGgZl0Ax0FNUNwQO0nNXGdESkZmB0r4lT9QMbTx5aHXurX2nHMDryYdW7WhoC2TgQAvD_BwE&gclsrc=aw.ds>

<https://www.ibm.com/products/cognos-analytics>

<https://en.wikipedia.org/wiki/IBM_Cognos_Analytics>

<https://en.wikipedia.org/wiki/Cognos>

<https://www.solarwinds.com/sql-sentry/use-cases/ssas-performance-monitoring?CMP=KNC-TAD-GGL-SW_NA_X_PP_CPC_LD_EN_PRODB_TIN-XDB&17913394546~139205942626_g_c_ssas-e~614098017743~~p72469533976&s_kwcid=AL!11508!3!614098017743!e!!g!!ssas&ds_cid=71700000098707476&ds_agid=58700007958078445&network=g&device=c&keyword=Ssas&matchtype=e&creative=614098017743&feeditemid=&gad_source=1&gclid=CjwKCAiAvJarBhA1EiwAGgZl0NP8UUiukfgVU8l69U3Cp0xIK5xnL2IAvRvoVn5cWu3CwQt4hZ6rthoCO0IQAvD_BwE&gclsrc=aw.ds>

<https://learn.microsoft.com/en-us/analysis-services/ssas-overview?view=asallproducts-allversions>

<https://en.wikipedia.org/wiki/Microsoft_Analysis_Services>

<https://learn.microsoft.com/en-us/analysis-services/analysis-services-overview?view=asallproducts-allversions>

<https://www.oracle.com/database/technologies/olap.html>

<https://www.oracle.com/database/technologies/olap.html#:~:text=Oracle%20OLAP%20is%20a%20world,speed%20of%20thought%20response%20times>.

<https://www.techtarget.com/searchdatamanagement/definition/OLAP#:~:text=OLAP%20(online%20analytical%20processing)%20is,from%20different%20points%20of%20view>.

<https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Oracle.Options.OLAP.html>

<https://www.dba-oracle.com/t_olap_dimensions_cubes.htm>