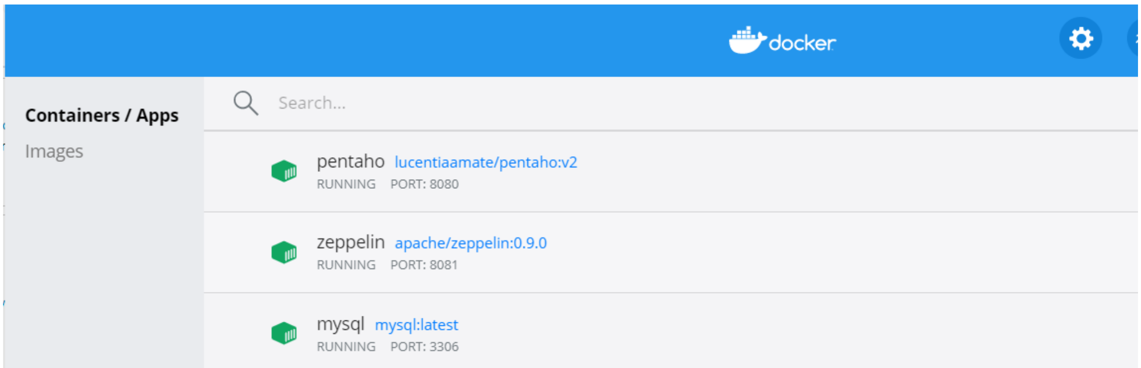
Task 4. Deployment and use of an analytic system.

In this task I am going to develop an analytic system using Pentaho server. To do that we either have to have the pentaho-server installed or the docker container given in the campus. I have chosen to go with the docker container.

This first thing we have to do is to have the docker container running at the port we want. I have chosen the 8080.

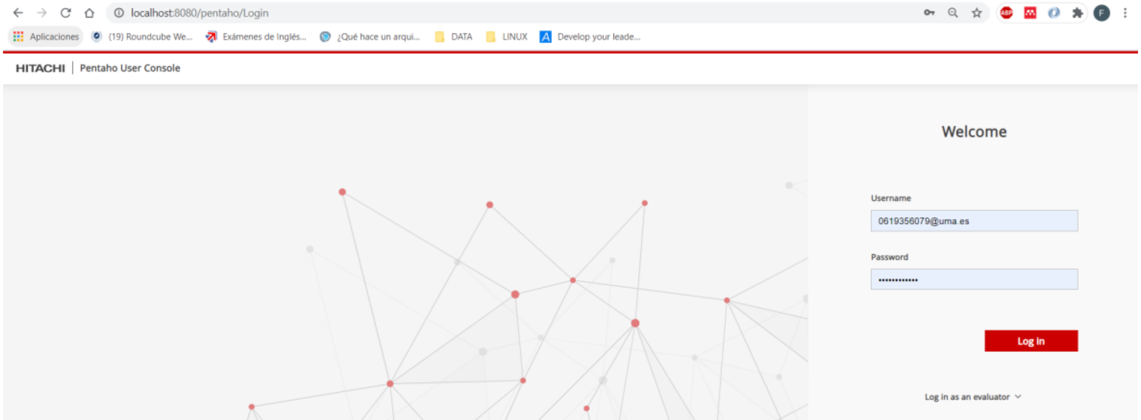


pentaho lucentiaamate/pentaho:v2
RUNNING PORT: 8080

zeppelin apache/zeppelin:0.9.0
RUNNING PORT: 8081

mysql mysql:latest
RUNNING PORT: 3306

When the docker is running we can access to Pentaho-server by going to localhost:8080/pentaho.

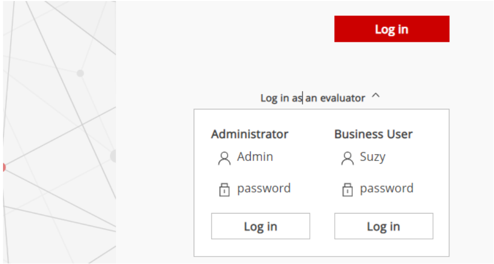


localhost:8080/pentaho/Login

Username: 0619356079@uma.es

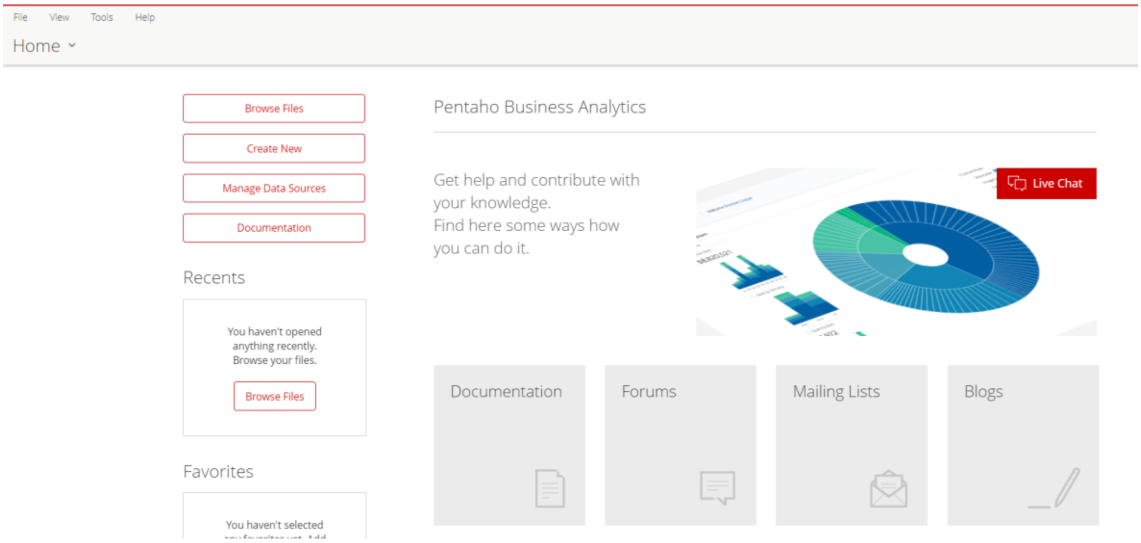
Password: ••••••••••••

Now we have to login as Admin.



Administrator	Business User
Admin	Suzy
password	password

Once logged in this window should appear.



Get help and contribute with your knowledge.
Find here some ways how you can do it.

You haven't opened anything recently. Browse your files.

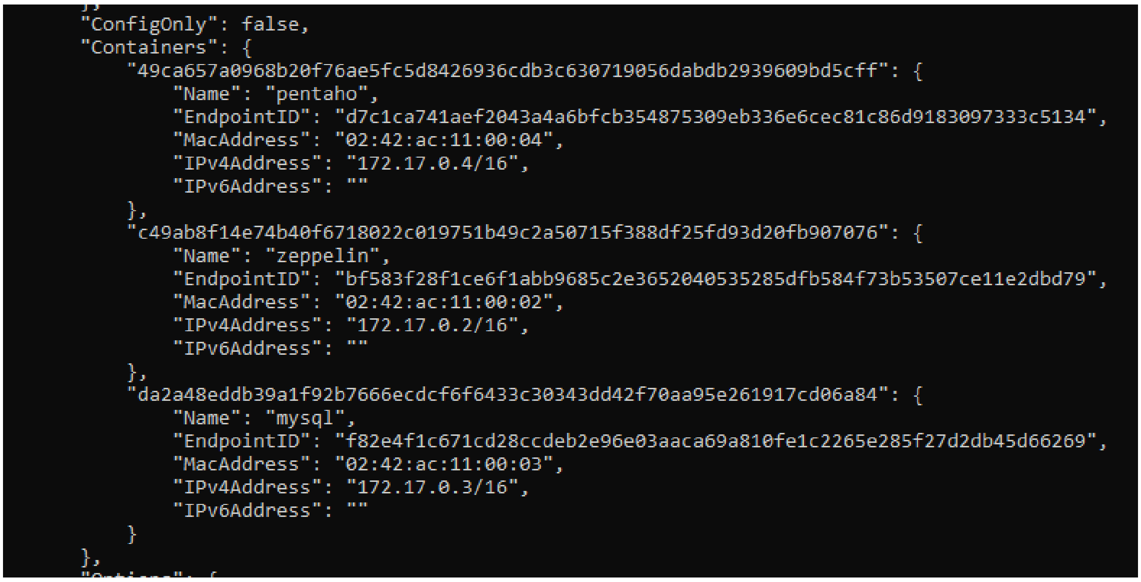
You haven't selected

The next step is to make the connection to the database. To make the connection to the thatabase we have to know that now we are not in our computer machine as we where in the other programs, now we are running inside pentaho-server docker machine.

To connect to Airlines database we have to know wich IP docker has given to the MySQL database. This is because now we are connecting two different machines.

To find out the MySQL machine IP we have to run the following command in Windows CMD.

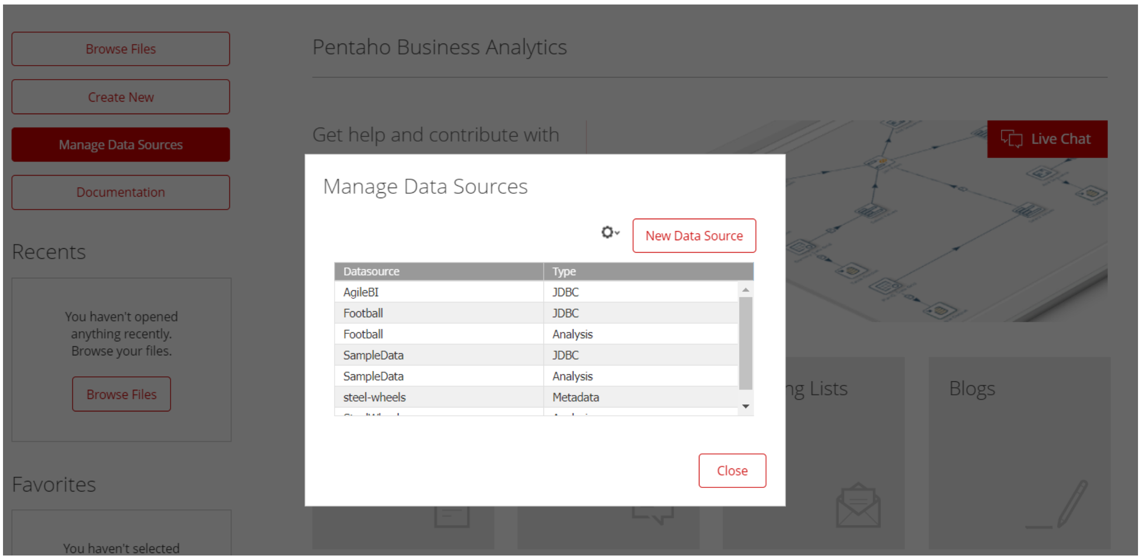
docker network inspect bridge



```
        },
        "ConfigOnly": false,
        "Containers": {
            "49ca657a0968b20f76ae5fc5d8426936cdb3c630719056dabdb2939609bd5cff": {
                "Name": "pentaho",
                "EndpointID": "d7c1ca741aef2043a4a6bfcb354875309eb336e6cec81c86d9183097333c5134",
                "MacAddress": "02:42:ac:11:00:04",
                "IPv4Address": "172.17.0.4/16",
                "IPv6Address": ""
            },
            "c49ab8f14e74b40f6718022c019751b49c2a50715f388df25fd93d20fb907076": {
                "Name": "zeppelin",
                "EndpointID": "bf583f28f1ce6f1abb9685c2e3652040535285dfb584f73b53507ce11e2dbd79",
                "MacAddress": "02:42:ac:11:00:02",
                "IPv4Address": "172.17.0.2/16",
                "IPv6Address": ""
            },
            "da2a48eddb39a1f92b7666ecdcf6f6433c30343dd42f70aa95e261917cd06a84": {
                "Name": "mysql",
                "EndpointID": "f82e4f1c671cd28ccdeb2e96e03aaca69a810fe1c2265e285f27d2db45d66269",
                "MacAddress": "02:42:ac:11:00:03",
                "IPv4Address": "172.17.0.3/16",
                "IPv6Address": ""
            }
        },
        "Options": {
```

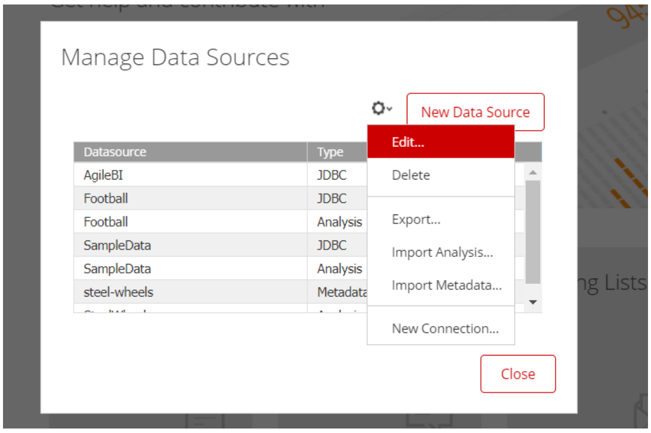
This command shows the bridges IP for each docker container running. We can see that MySQL IP is 172.17.0.3.

Having this clear we have to click on Manage Data Source in the Pentaho server window.



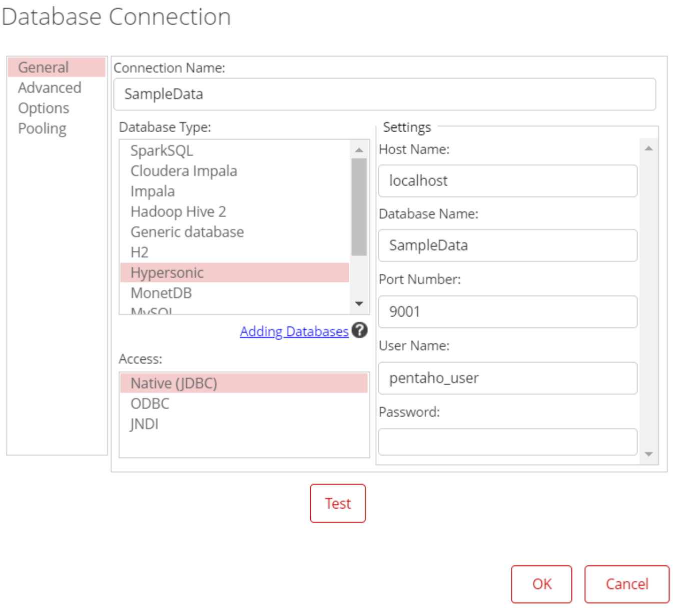
Datasource	Type
AgileBI	JDBC
Football	JDBC
Football	Analysis
SampleData	JDBC
SampleData	Analysis
steel-wheels	Metadata

And after that on edit



Datasource	Type
AgileBI	JDBC
Football	JDBC
Football	Analysis
SampleData	JDBC
SampleData	Analysis
steel-wheels	Metadata

When we click on edit this window should appear.



Connection Name: SampleData

Database Type: Hypersonic

Access: Native (JDBC)

Settings

Host Name: localhost

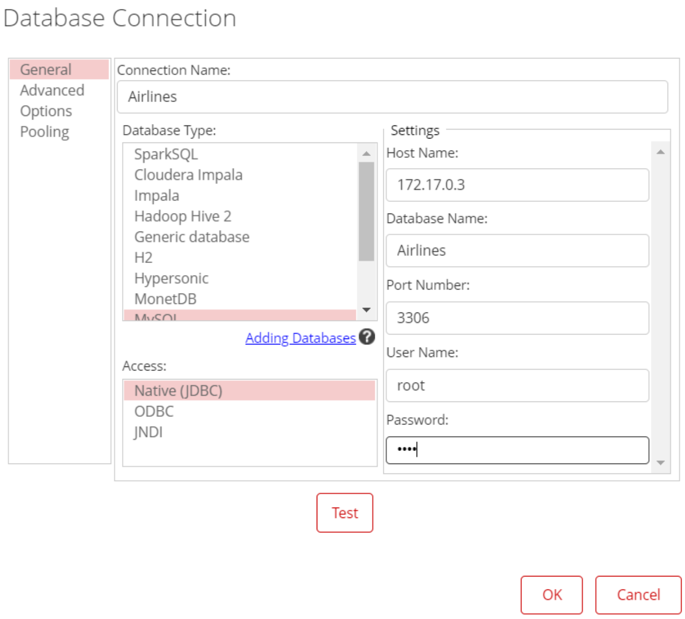
Database Name: SampleData

Port Number: 9001

User Name: pentaho_user

Password:

In here we have to set the Airlines database information in order to connect to it.



Database Connection

Connection Name: Airlines

Database Type: SparkSQL, Cloudera Impala, Impala, Hadoop Hive 2, Generic database, H2, Hypersonic, MonetDB, MySQL

Access: Native (JDBC), ODBC, JNDI

Settings

Host Name: 172.17.0.3

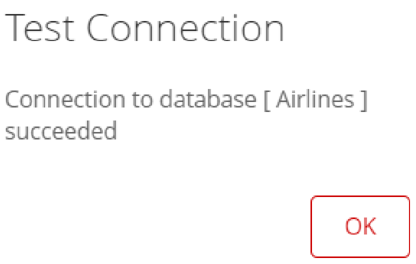
Database Name: Airlines

Port Number: 3306

User Name: root

Password: ••••

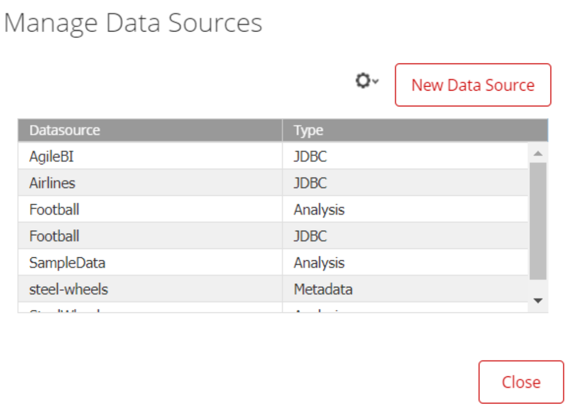
And now we click on ‘Test’. If the connection is successful, this windows should appear.



Test Connection

Connection to database [Airlines] succeeded

We click in ‘Ok’ and the connection should be made.

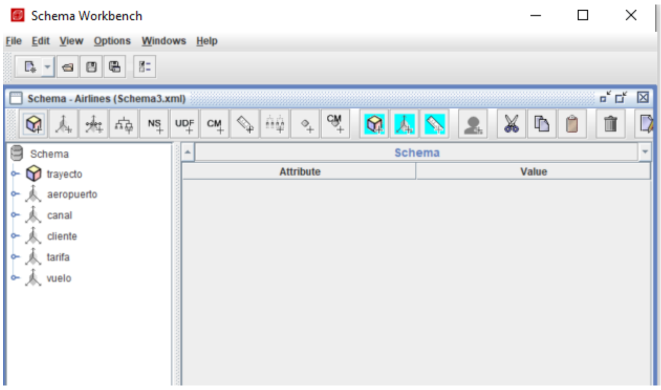


Manage Data Sources

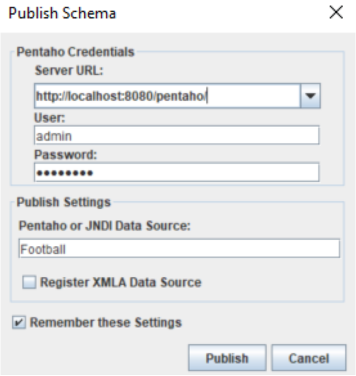
Datasource	Type
AgileBI	JDBC
Airlines	JDBC
Football	Analysis
Football	JDBC
SampleData	Analysis
steel-wheels	Metadata

The next step is to upload the multidimensional schema we did in schema-workbench to the Pentaho server.

To do that we have to open the schema-workbench program and upload it.



With the schema opened we have to click on File -> publish and put the pentaho login information.



Pentaho Credentials

Server URL: http://localhost:8080/pentaho/

User: admin

Password: ••••••••

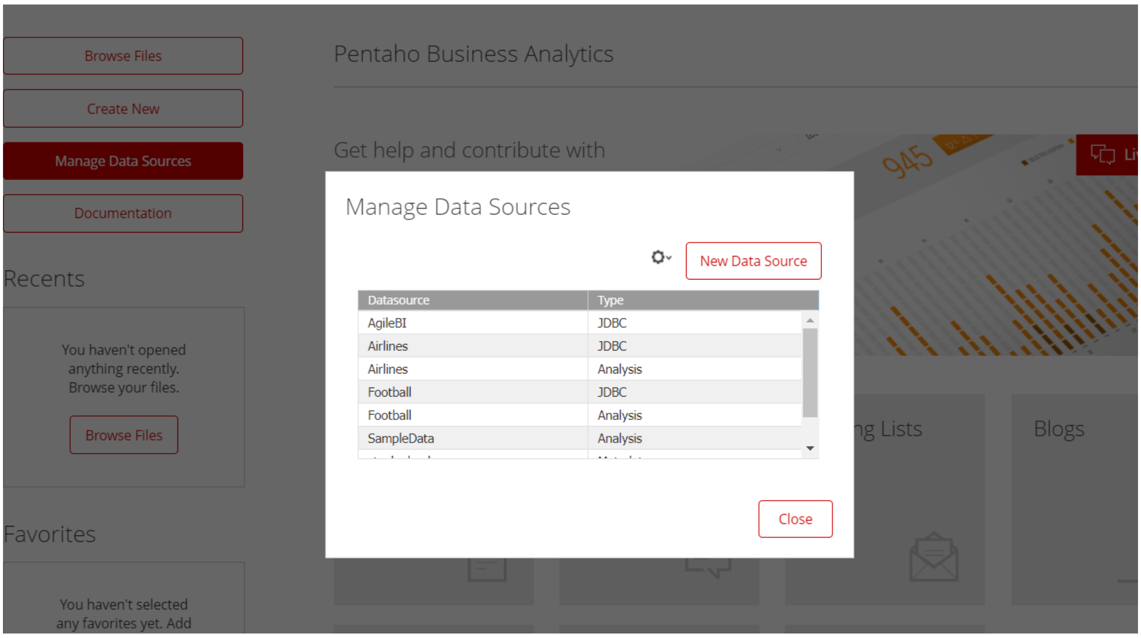
Publish Settings

Pentaho or JNDI Data Source: Football

□ Register XMLA Data Source

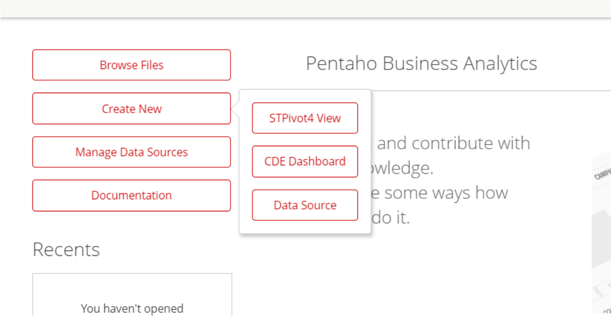
☑ Remember these Settings

Now we click on publish and it should now appear in the Pentaho server window.

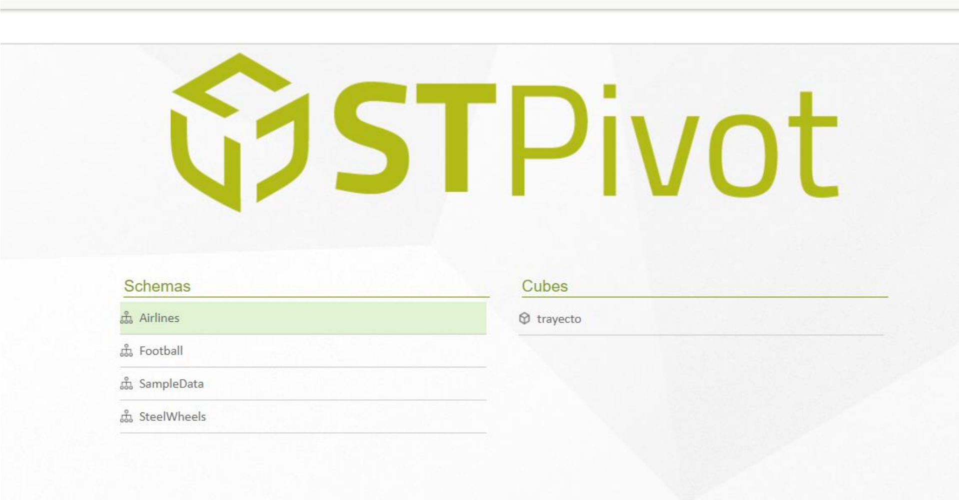

Browse Files
Create New
Manage Data Sources
Documentation
Pentaho Business Analytics
Get help and contribute with
Recents
You haven't opened anything recently. Browse your files.
Browse Files
Favorites
You haven't selected any favorites yet. Add
ng Lists
Blogs
945
Manage Data Sources
New Data Source
Close

Datasource	Type
AgileBI	JDBC
Airlines	JDBC
Airlines	Analysis
Football	JDBC
Football	Analysis
SampleData	Analysis

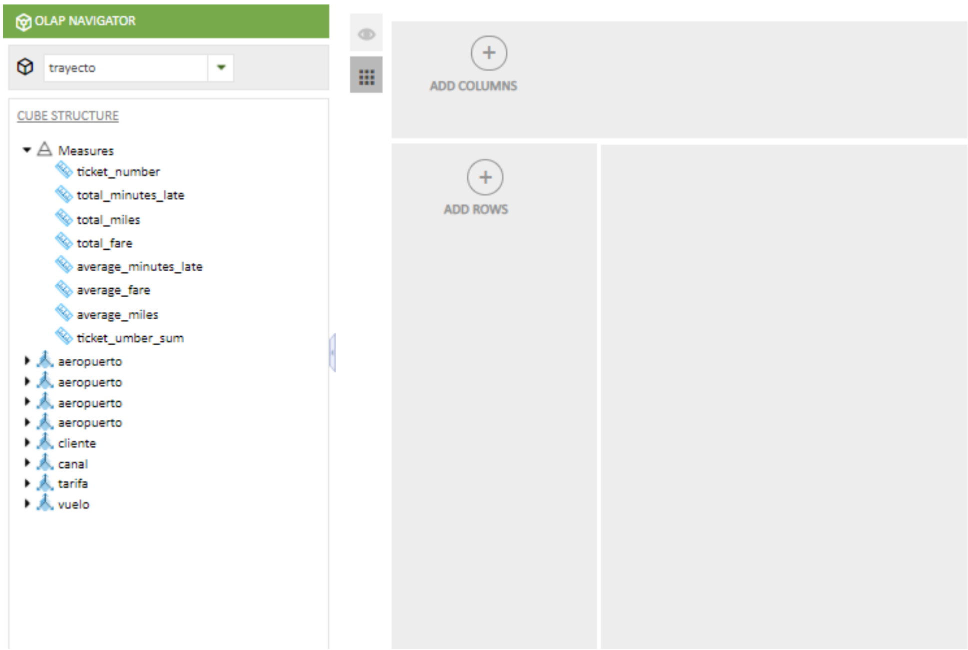
Now that we have the database connection and the multidimensional design uploaded we have to create a new STPivot 4 view.



After that we have to click on Airlines -> trayecto.



Now we have our multidimensional scheme ready to be analyzed.

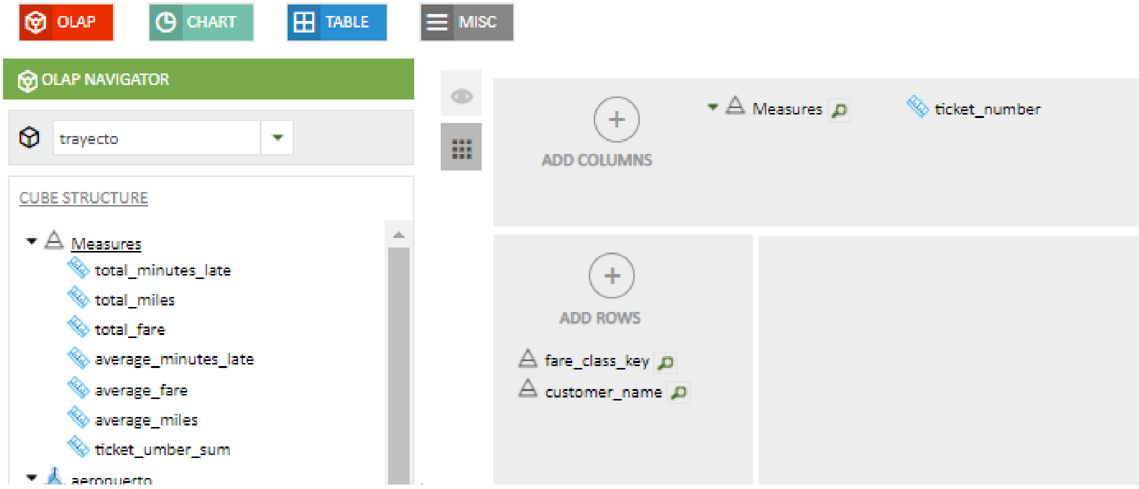


If we check the OLAP navigator we can see the dimensions and the measures, we designed in the second task. Now we can think of the questions we want to answer and start our analysis.

If we were to announce any type of discount we could be interested in the name of the clients who buy more each type of fare so the first question is:

1- Which client buys more each kind of fare?

In order to answer that question, we would have to analyze the ‘customer_name’ from the client dimension and the ‘fare_class_key’ from the fare dimension and analyze it with the ticket_number measure.



trayecto

CUBE STRUCTURE

Measures
total_minutes_late

total_miles

total_fare

average_minutes_late

average_fare

average_miles

ticket_umber_sum

aeropuerto

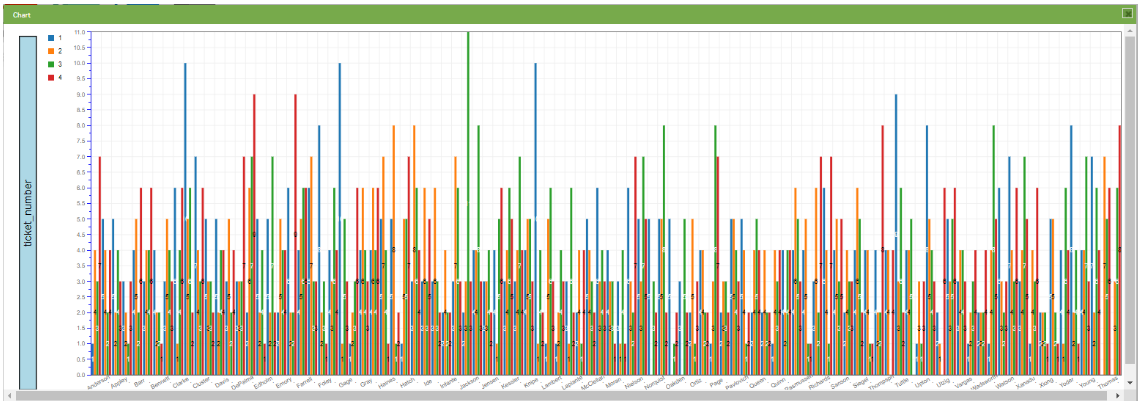
Measures

ticket_number

fare_class_key

customer_name

If we show this result in a line bar we could see which bar is higher.



If we go by clicking on the bars we could see that for the 1 type of fare the clients who bought more flights are Clewett, Gage and Knutsen with 10 tickets each.

For the second kind of fare we have Hale and Halton with 8 tickets each.

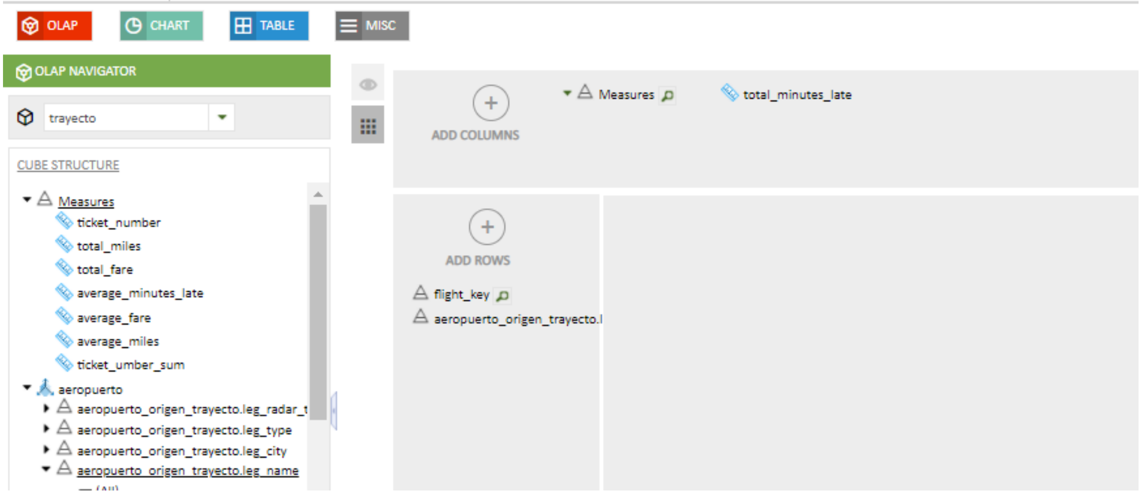
For the third kinf of fare we have a clear winner who is Jackson with 11 tickets.

And for the fourth kind of fare we have Dods and Erickson with 9 tickets each.

Another question could be which airport has had a bigger delay at the different flights.

2- Which airport suffers the bigger delay in every flight?

To answer this question we would have to have in mind the airport name, the flight key and the total_minutes_delay measure.



trayecto

CUBE STRUCTURE

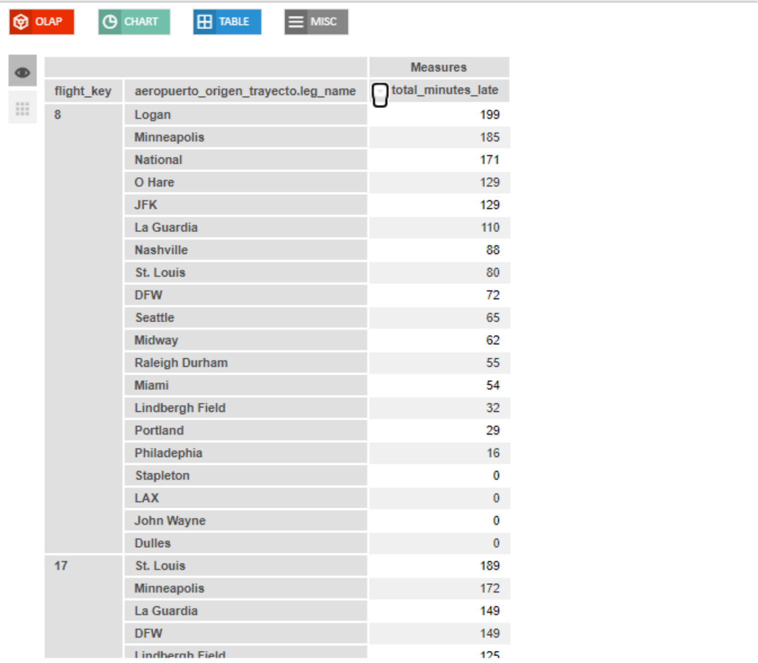
- Measures
  - ticket_number
  - total_miles
  - total_fare
  - average_minutes_late
  - average_fare
  - average_miles
  - ticket_umber_sum
- aeropuerto
  - aeropuerto_origen_trayecto.leg_radar_t
  - aeropuerto_origen_trayecto.leg_type
  - aeropuerto_origen_trayecto.leg_city
  - aeropuerto_origen_trayecto.leg_name

Measures total_minutes_late

flight_key

aeropuerto_origen_trayecto.l

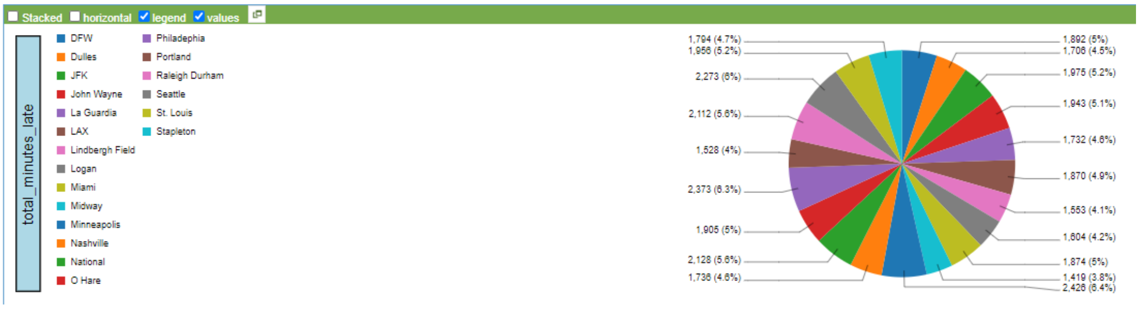
We get a table like this:



		Measures
flight_key	aeropuerto_origen_trayecto.leg_name	total_minutes_late
8	Logan	199
	Minneapolis	185
	National	171
	O Hare	129
	JFK	129
	La Guardia	110
	Nashville	88
	St. Louis	80
	DFW	72
	Seattle	65
	Midway	62
	Raleigh Durham	55
	Miami	54
	Lindbergh Field	32
	Portland	29
	Philadephia	16
	Stapleton	0
	LAX	0
	John Wayne	0
	Dulles	0
17	St. Louis	189
	Minneapolis	172
	La Guardia	149
	DFW	149
	Lindbergh Field	125

For example, the flight key 8 has had a total of 199 minutes delay in Logan airport.

We could also pol a chart that tell us which airport has the bigger delay.

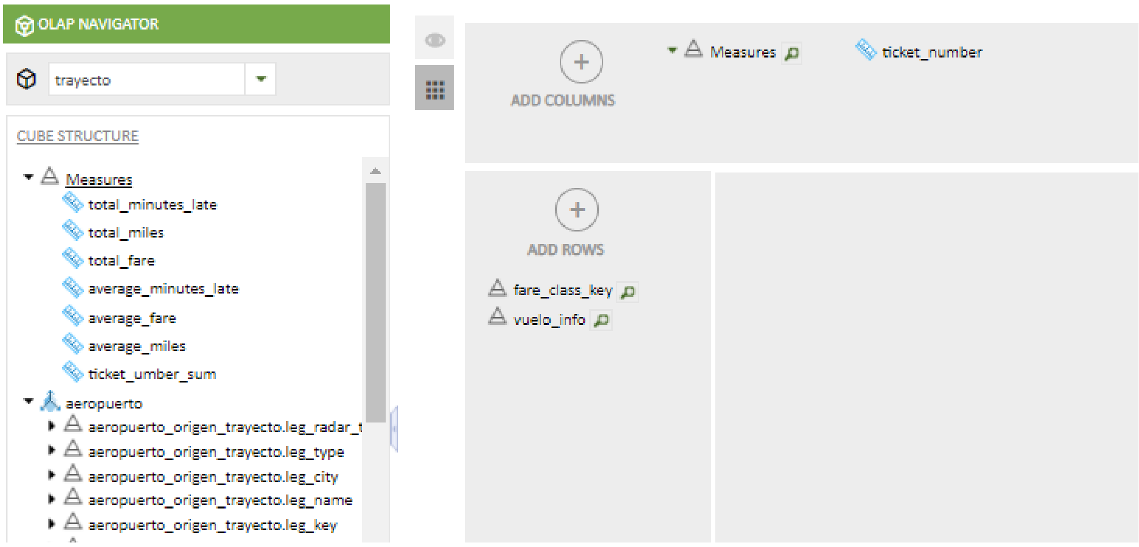


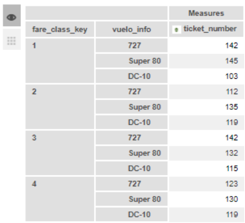
We can see that Minneapolis airport is the one with the bigger delay.

Another question could be if the kind of plane determines the kind of fare each customer buys.

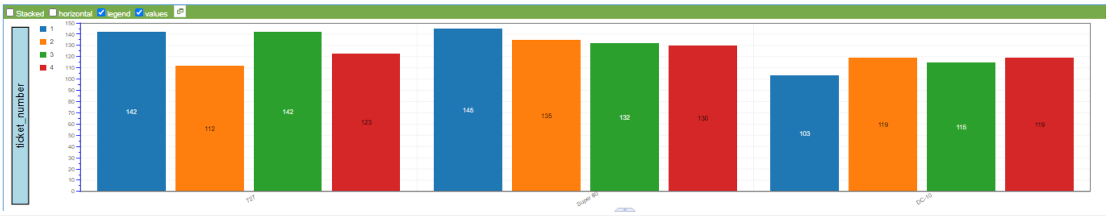
3- ¿Does the type of the plane determine the kind of fare the client buys?

To answer that question, we are going to use the dimensions fare and flight and the measure ticket_number.

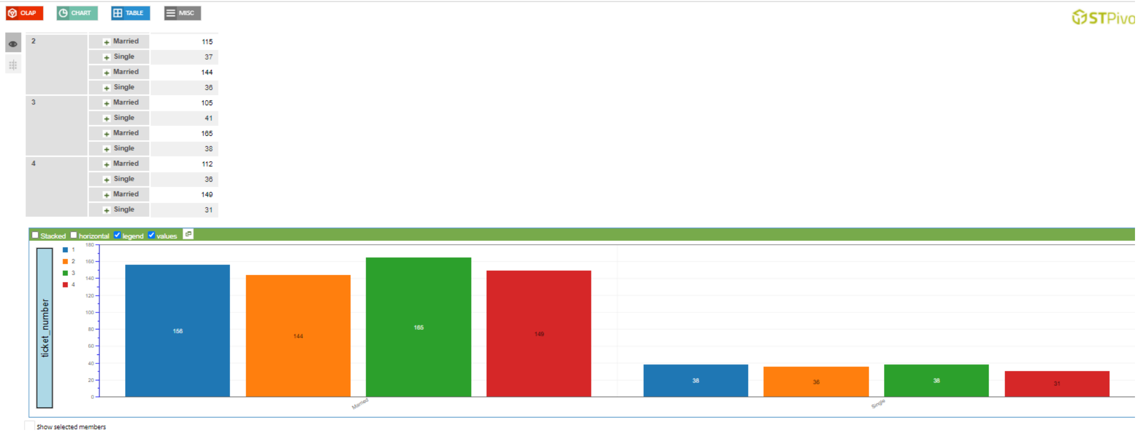




		Measures
fare_class_key	vuelo_info	ticket_number
1	727	142
	Super 80	145
	DC-10	103
2	727	112
	Super 80	135
	DC-10	119
3	727	142
	Super 80	132
	DC-10	115
4	727	123
	Super 80	130
	DC-10	119



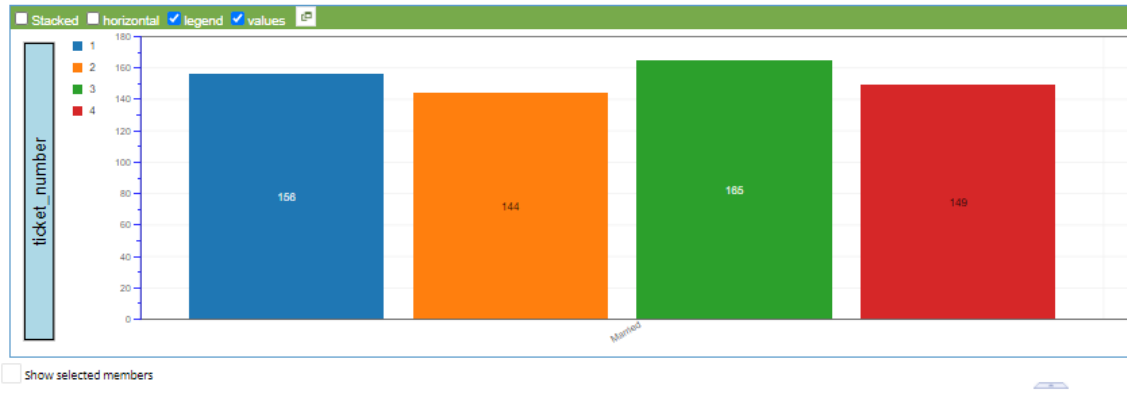
4- For the las question we want to know if the martial state of the customer has anything to do with how frequently they flight or if the martial state has anything to do with how frequently they flight.



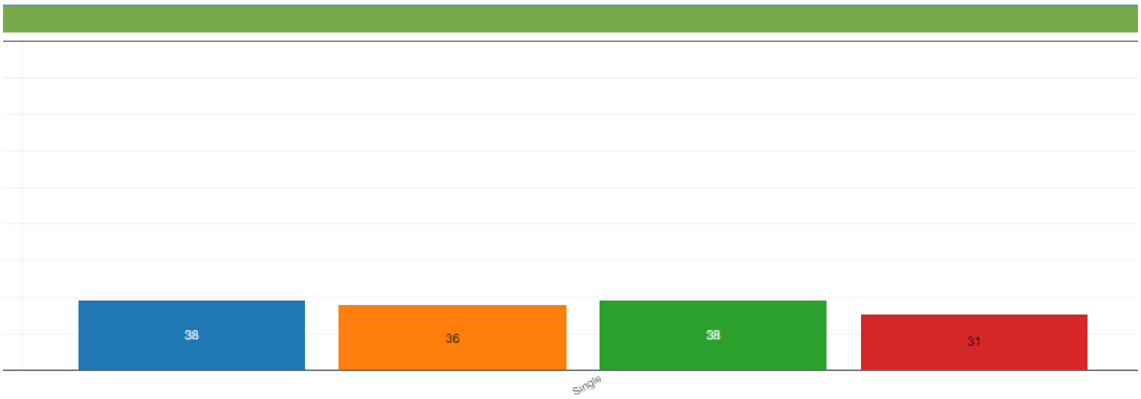
2	+ Married	115
	+ Single	37
	+ Married	144
	+ Single	36
3	+ Married	105
	+ Single	41
	+ Married	165
	+ Single	38
4	+ Married	112
	+ Single	36
	+ Married	149
	+ Single	31

□ Stacked □ horizontal ☑ legend ☑ values

□ Show selected members


□ Stacked □ horizontal ☑ legend ☑ values

□ Show selected members



We can clearly see that the married people travel much more than the single ones.