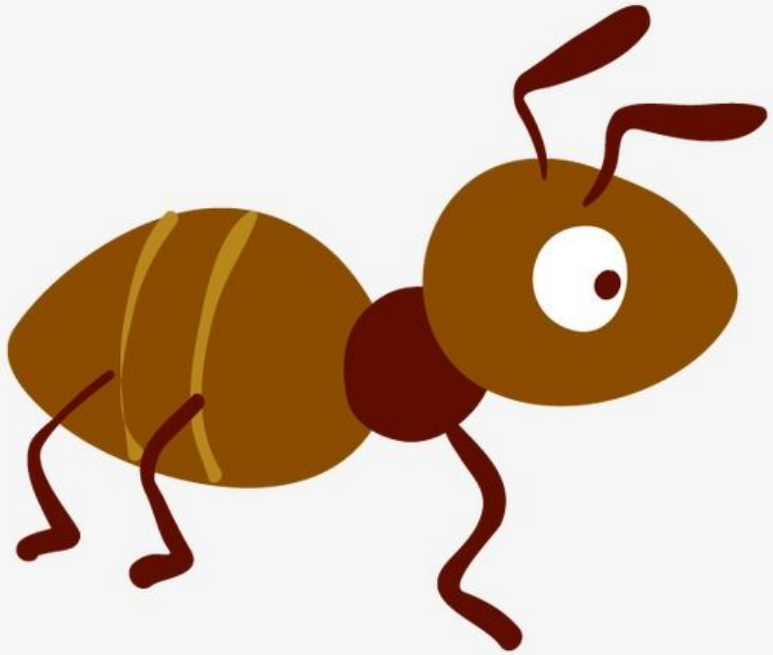
Welcome to the data

Data Analyst Progress

6%

Tangarana



OUR DATASET:

users

payments

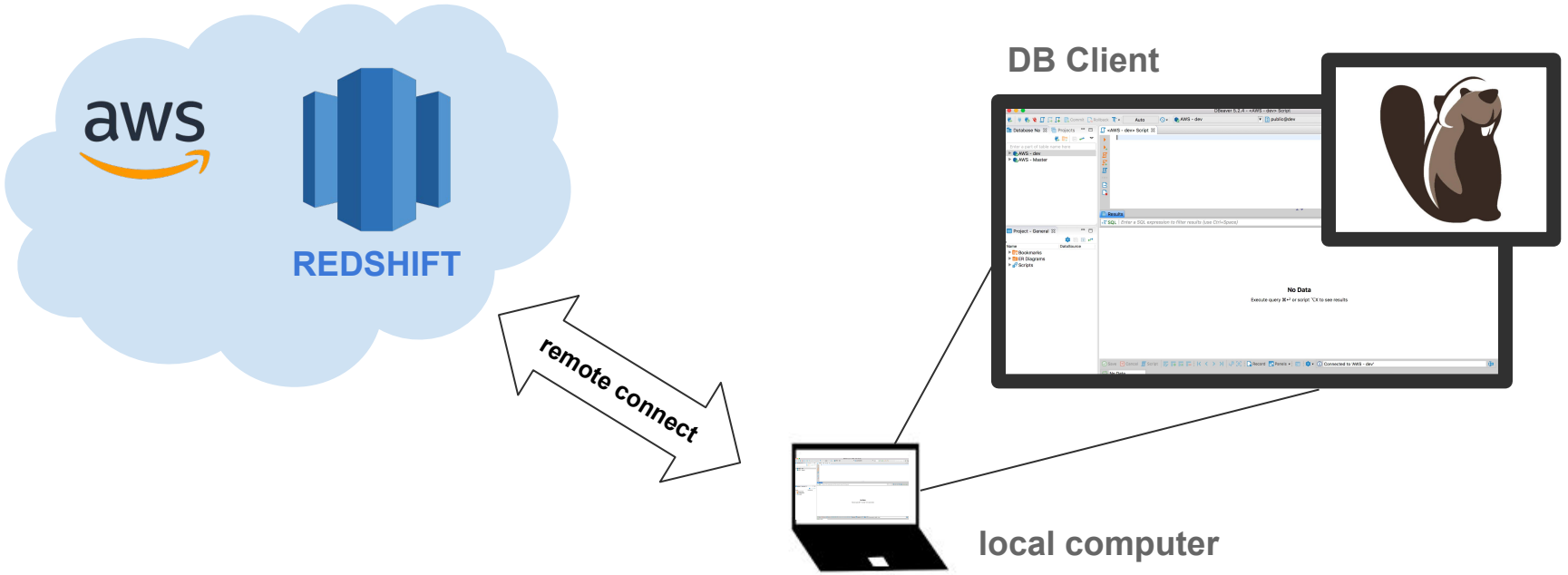
2017

*we will work as if we were Jan 1st, 2018

What is a database?

Data WareHouse

Connect to **REDSHIFT**



Host:
redshift-cluster-1.cvqknyv7jbo2.eu-west-1.redshift.amazonaws.com

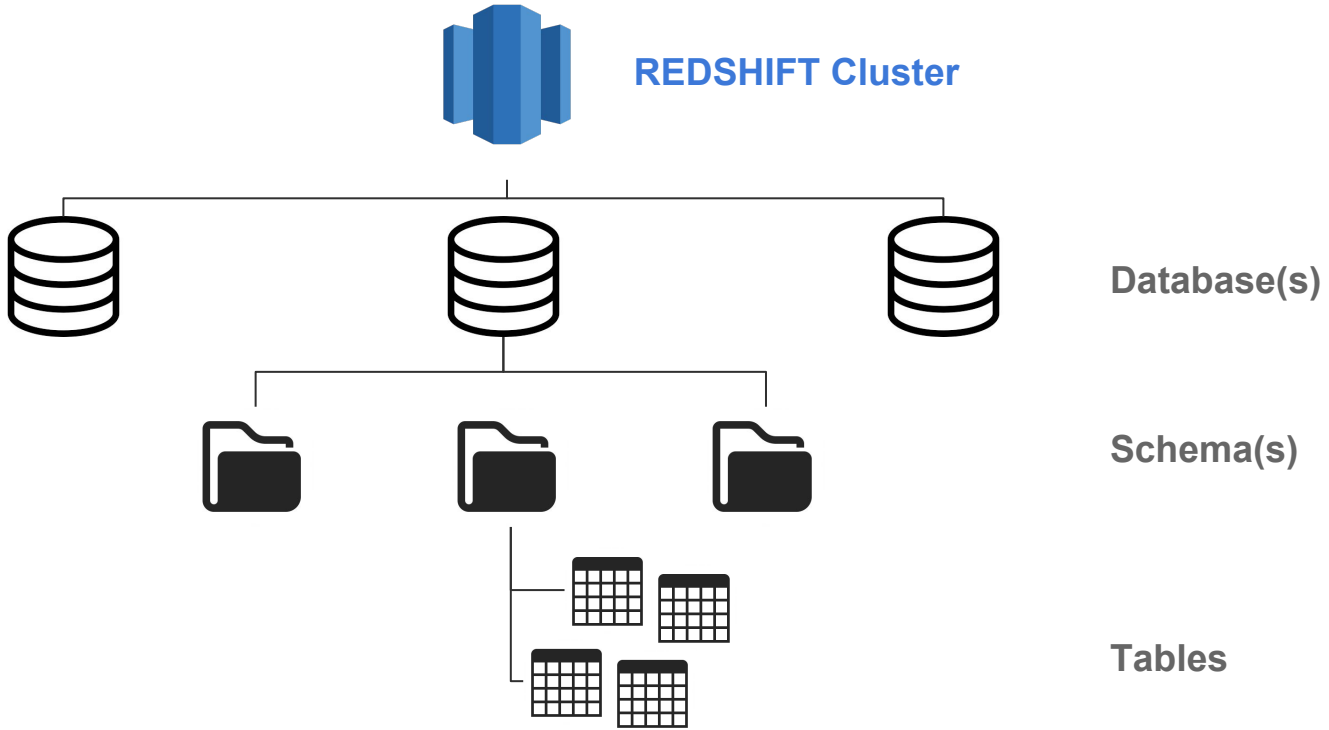
Port: ***5439***

Database: ***dev***

User: ***student***

Password: ***

REDSHIFT data structure



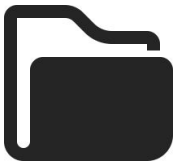
Assignment 1.

- Read about multiple REDSHIFT alternatives.
- Spot their advantages and disadvantages.
- Present your findings.

José María Manso

CHIEF FINANCIAL OFFICER

Tangarana Data



SCHEMA: **tangarana**

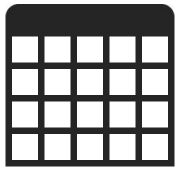


TABLE: **tangarana.users**

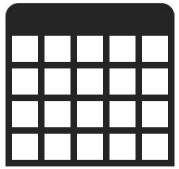


TABLE: **tangarana.payment_transactions**

SELECT ___ FROM ___ WHERE ___ LIMIT ___ ;

$$\text{Average Ticket} = \frac{\sum \text{amount}}{\text{\# of transactions}}$$

Refund

Classroom Assignment

Task:

- **Jose María** is working on a business plan for year 2018. He needs some basic statistics from 2017. These statistics will be used as inputs for the business plan. You receive an e-mail with questions from José María that you should answer ASAP. The email reads:

  “Hi, I need to know the payments we have in France, how much revenue we did and the ticket. The same for Italy. Which one performs better? Waiting for your response. Best. JM.

- You should answer the e-mail looking at the data. **You don’t have means to ask for clarifications.**

Assignment 2. Statistical test.

Task:

- **José María** is not very familiar with statistics however he knows the concept of statistical significance. He writes to you another e-mail:

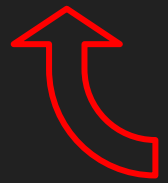
  “ Hi, Thanks for the numbers. Just one doubt: is the difference between France and Italy meaningful? Let me know. Best. JM.

- You should answer the question by writing a brief summary.

Hint: *“Meaningful”* can be interpreted as *“statistically significant”*.

Why do we aggregate the data?

SELECT ___ FROM ___ WHERE ___ LIMIT ___ ;



count(____)

```
SELECT _____ , count( _____ )
FROM ____
WHERE _____
GROUP BY ____
LIMIT ____  ;
```

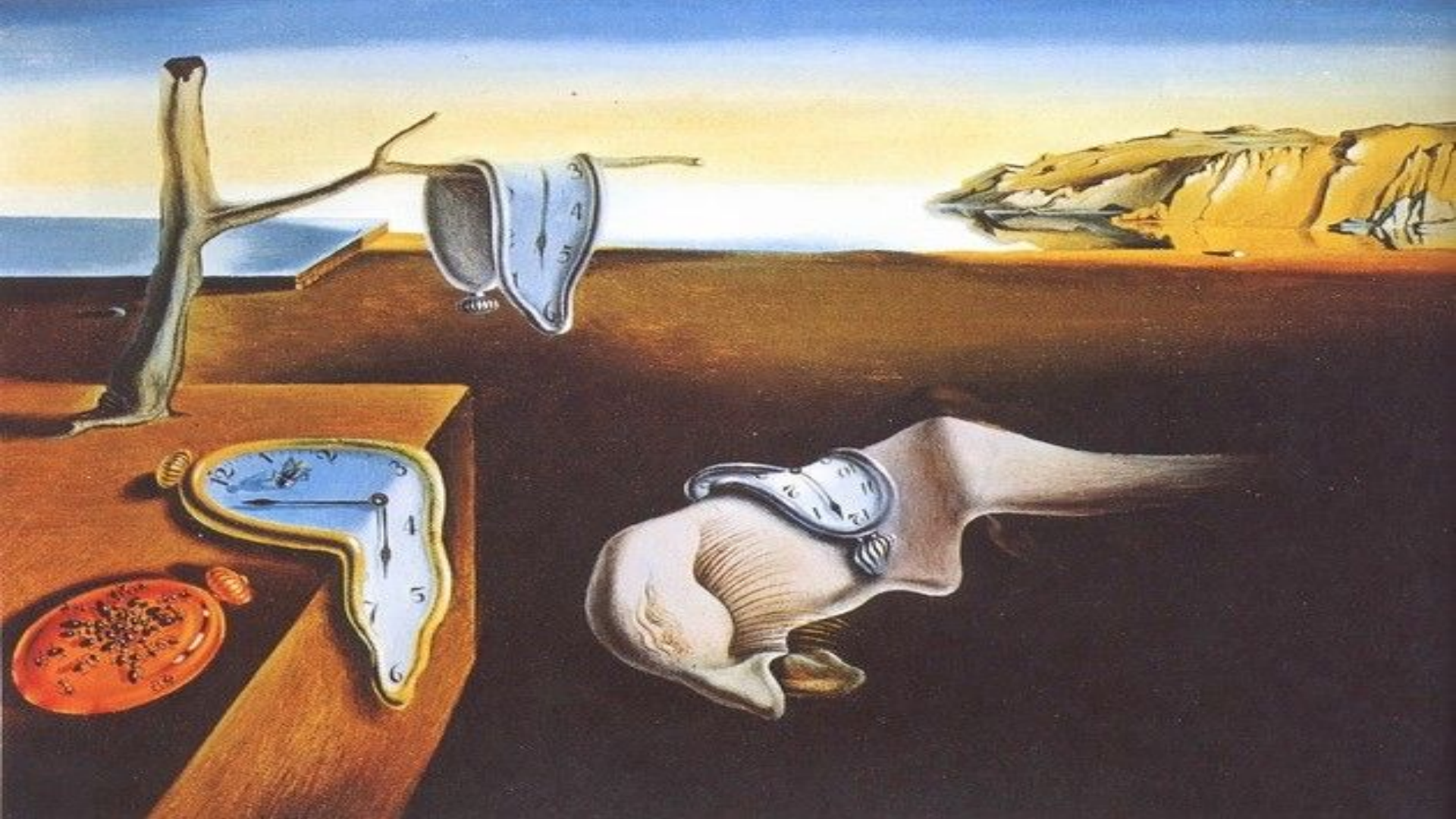
Assignment 2b: Aggregate Data

Task:

- Since your insights are helpful to **Jose María**, he keeps asking for more. Once again he writes to you an e-mail:

  “ Hi, Thanks for your help One more thing: what are our top countries in sales (which currency?). Also, do you consider how many invalid transactions we have? What is the refund rate? Let me know, once you get the data. Best. JM.

- You should answer the e-mail looking at the data. In case of doubt, you can discuss with your peer**.**



442088460

Timestamp: 1542450180

Representation:
2018-11-17 10:23:00.0000

Date: 2018-11-17

Also:
2018-11-17 00:00:00.0000

SELECT date (‘2018-11-17 10:23:00.0000‘)

results in: 2018-11-17 00:00:00.0000

Month: 2018-11

Useful representation:

2018-11-01

(2018-11-01 00:00:00.0000)

Classroom Assignment

Task:

- **Jose María** has data on payments. He wants to know more however. He writes to you an e-mail:

  “Hi, Could you check how many users we acquire every month? Rough estimation for monthly average Q1, 2017? Best. JM.

- You should answer the e-mail looking at the data. In case of doubt, you can discuss with your peer**.**

Assignment 3. Date functions.

Task:

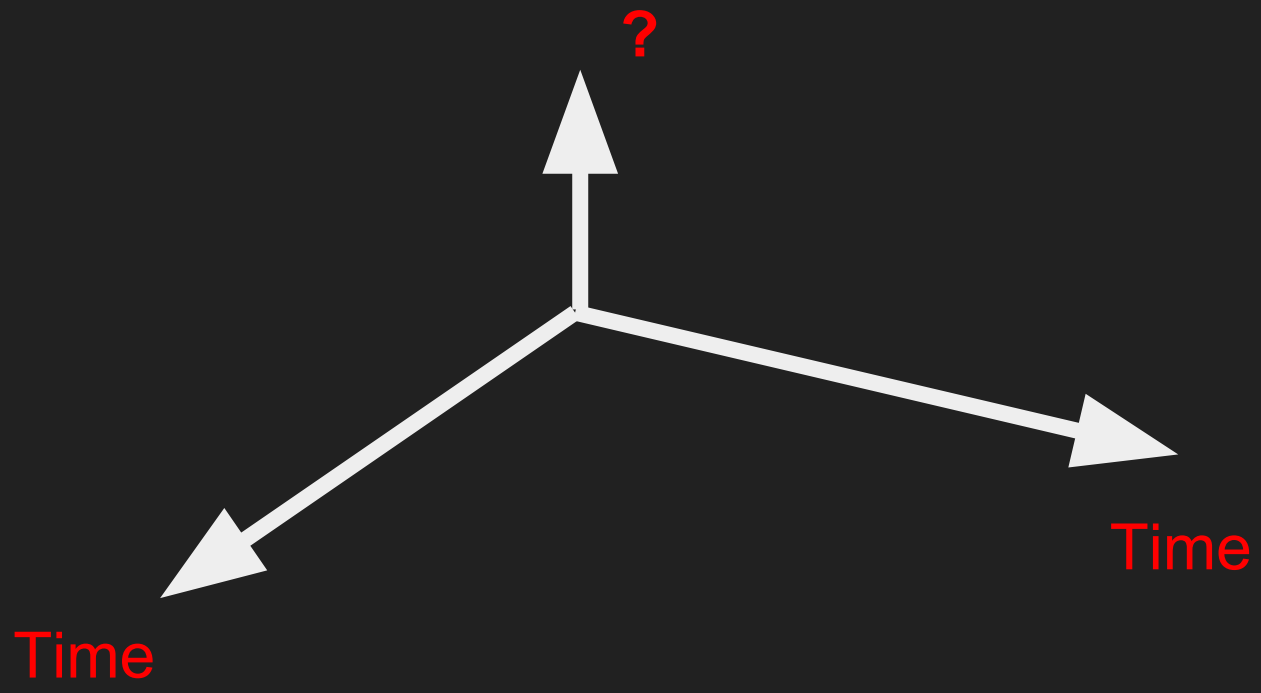
- You foresee **José María's** questions flood. In order to respond fast, you need to be familiar with date and time manipulation. There are multiple functions that ease any analysis.

- Review Redshift documentation and get familiar with most common:

  DATEADD, DATETRUNC, DATEPART, DATEDIFF, etc.

- https://docs.aws.amazon.com/redshift/latest/dg/Date_functions_header.html

Two timelines?



Conversion Rate

Average
Revenue
Per
User

Assignment 4. Performance tracking - Part 1.

Task:

- **José María** leaves your with a challenge. In 2018 you will be responsible for reporting. He suggest the following format to track performance month over month:

Date	Acquisition	Conv. 1D	Conv. 3D	...	Conv. 30D	CR 1D	CR 3D	...	CR 30D	Rev 1D	Rev 3D	...	Rev 30D	ARPU 1D	ARPU 3D	...	ARPU 30D
2017-01	1000	10	12	...	100	1 %	1.2 %	...	10%	250 €	310 €	...	2700 €	0.25 €	0.31	...	2.7 €
2017-02	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

- Your objective is to write the queries, so that you can build similar report. Once you have the queries, you should run them and create the report. You should save your queries in GitHub, so that you can easily reproduce the analysis every month.

Assignment 4. Performance tracking - Part 2.

Task:

- **José María** tires to estimate how much money will the company earn in Q1, 2018 from the users acquired in 2017.
- Your will help him. Your objective is to analyze historical data, time to conversion and provide some estimates.
- You should write a brief summary that justify your estimates. You should store your report in GitHub, so that you can easily track any changes.

Hint:

- Linear extrapolation is a simple estimation method. Consider there might be a relations between short and long term conversions.

Assignment 4. Performance tracking - Part 3.

Task:

- Estimates that goes beyond Q1 2018 are imprecise if you consider only the users from 2017.
- **José María** is aware of that, so he rather wants to analyze seasonality curve.
- Your task is to build monthly seasonal curve for 2017 (ARPU), so that it can be used for predictions in 2018.
- Store your results in GitHub, so that your peers can benefit from your work.

Assignment 4. Performance tracking - Part 4.

Task:

- Tangarana works with two type of business: B2B and B2C. 2018 estimations are meant be be splitted by type of business. Moreover, UTM source for users is key for decision making.

- You anticipate **José María's** next questions, so that, you add to your previous report: business type and UTM Source dimensions.

- As usual you store your analysis in GitHub.