

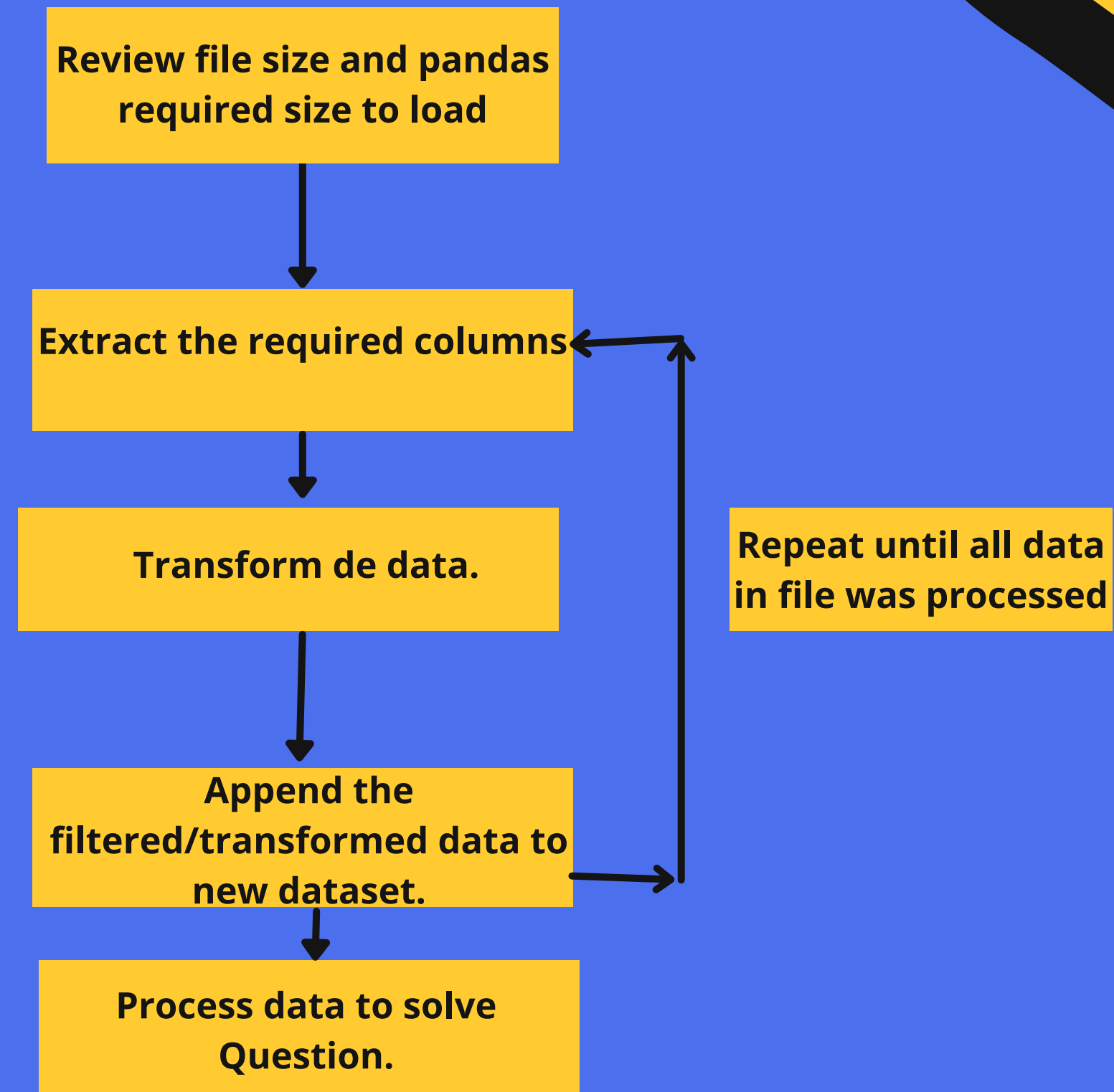
# Data Engineering Bootcamp Challenge

**Carla Garcia**

Report



# General Workflow



# Main Tools used



**Pandas**



**Chunk segmentation**



**grouping and aggregate**




**Map**

# # Q5 What are the lessons learned from this excersice?

The main problem faced in this task is to identify the valuable information to answer the questions and process the data in an smart segmented way to administrate the hardware resources that are available.

Another lesson learn is that always we need to Divide to conquer the solution. Trying to load the entire dataset causes instability in the system and computational crashes.

Also making a leverage for this can get imposible if its not designed to be modular.






## Q6 Can you identify other ways to approach this problem? Explain.

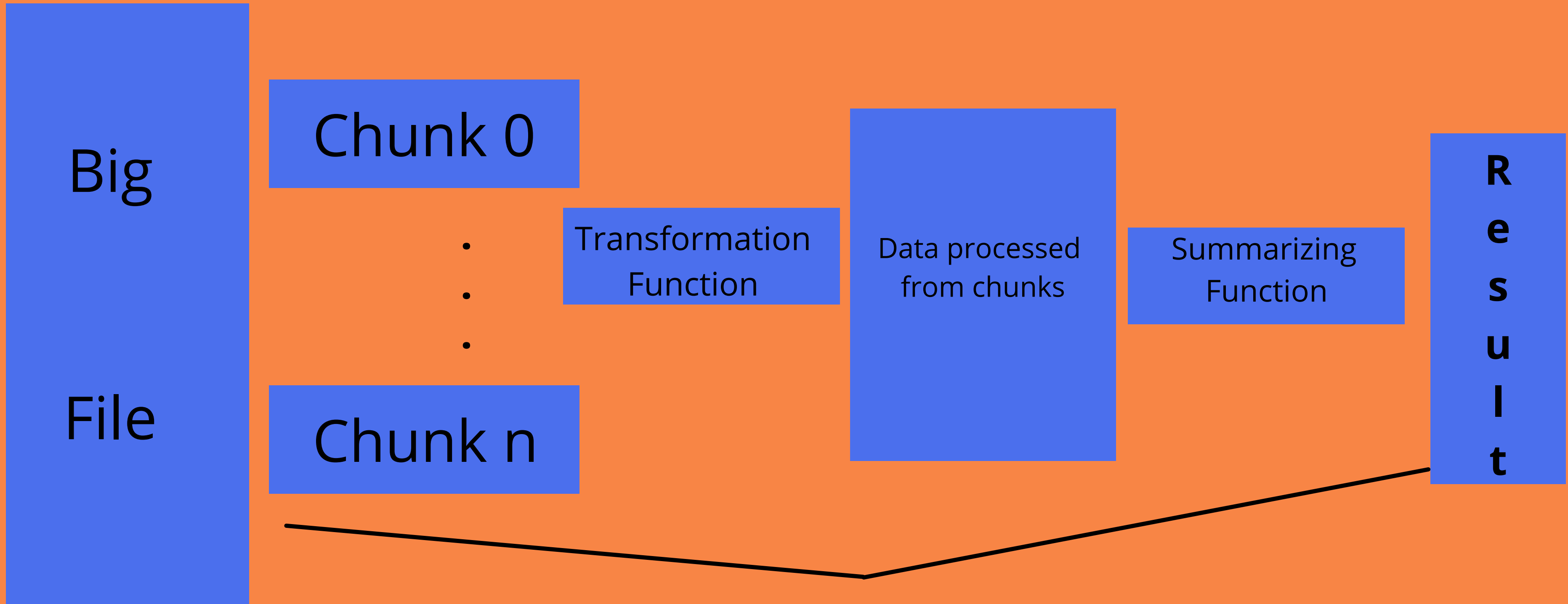
One way I would have wanted to explore to solve this problem is to move the entire file in to a Database.

A Database server is not required SQLite package can be used in this exercise to maintain all processing locally, but a database will give the flexibility to run queries also it can be run as a batch.

Also exploring parallelism packages like Dask can make more efficient the calculations.

If the core functions get designed in modular and vectorized way thinking. In how to make the extraction of the information of interest and the combination of the data in the chunks to reach a conclusion or result. If this step is well done moving to database, parallelism or pipelines in the clouds becomes smoother.





This can be addressed using a DB or making parallel computing.