Uncle Juan's arepas (a traditional Colombian dish) are considered to be the best in the world for the people from his hometown *Pueblo Pirelli*. He has 10 stores in the city spread over three zones, each selling more than 500 arepas a day. The arepas are produced at the store's kitchen in a three-phase cooking process, where the ingredients arrive daily in the morning to ensure the freshness of the final product. Thanks to Juanito (uncle Juan's son), the cooking process is completely tracked by a RDBMS installed on a small server in each one of the stores. Additionally, the ingredients purchases and inventory as well as arepas sales are tracked in a different centralized DWH system at uncle Juan's basement.

Due to Juanito's effort, the workers on each store have been able to create different applications that rely on the DBs to support their job, like dashboards to keep an eye on cooking metrics, daily reports on the quality of the produced arepas, or weekly summaries of the cooking failures. Back home, Maria (uncle Juan's wife) has also created other applications based on DWH data to support decision making like dashboards with the month-to-date sales information or daily reports on the existing inventory at different aggregation levels (store/zone/city).

Uncle Juan has just become the winner of an important funding program that will allow him to expand to other cities all over the region. In order to better prepare for such an expansion he has decided to take the company into the data-driven world and push to cloud-based technologies. You have been hired as part of the data engineering team that will support this transformation process. Below you will find the tasks you are expected to perform. Note that points are independent from each other, you don't need to complete one to start working on the other.

1. After some discussions with Juanito, it was decided that the project will start with the creation of a unified data platform on the cloud with the purpose of being the single point of access to all the data from the company. The goal is to come up with a proposal for taking the data into the cloud.
    a. You are asked to focus on the data storage components only. Which type of systems would you use and why (distributed filesystem, object storage, RDBMS, No-SQL, DWH, etc.)? How would you provide and manage access to the Maria and the kitchen workers?
    b. You are now discussing an architecture proposal for the whole system that starts from the source systems up to the storage components you defined before. What constraints or requirements would you take into account while evaluating the proposal? What types of proposals would you avoid? You might give examples of a high level architecture while writing your answers. Remember to justify them.

Consider the following requirements while working on your answers:

- ○ The impact on the existing applications developed by the stores' workers and Maria should be minimal.
- ○ A team of data scientists will be joining the company to start working on models to support the decision making processes, so don't forget about their potential needs!

2. It's been a year since the company transformation started and the data science projects don't stop coming! You have been assigned to one of the latest incoming proposals. The goal is to create a predictive maintenance product: a system that, based on cooking metrics, tells kitchen workers when to stop a machine for maintenance actions. Note that to prevent production to stop completely each kitchen has at least two machines for each cooking phase.
    a. Since there are lots of kitchens, machines and types of arepas, the first prototype will be developed on data regarding a specific combination of those, namely kitchen *k1*, machine *m1* (associated to cooking phase 1),

and arepa type *a1*. Your first task is to prepare the training dataset for the model. You have at your disposal the following datasets:

- Cooking metrics: a table containing cooking metrics retrieved by each machine associated to phase 1 for each cooked batch.
- Batch registry: a table containing the main information of each produced batch.
- Faulty intervals: a table containing for each machine associated to phase 1 the intervals in which metrics data are note reliable.

Your program is expected to receive as input start and end times (in the format *YYYY-MM-DDTHH:MM:SS*), and produce a dataset with the hourly averaged metrics for kitchen *k1* and machine *m1* associated to arepa type *a1* in the specified time interval. Keep in mind that faulty data must be filtered out.

Write your code in Python using preferably PySpark or Pandas. We expect a repository with the code needed to generate the training dataset; remember to follow development and data engineering best practices. Note that we would like not only to have a look at it but also to be able to run it, so documentation is well appreciated.

b. Suppose now that the prototype gave good results and training must be scaled to consider all the machines in all the kitchens associated to phase 1, but reduced to a list of arepa types given in a text file (whose location is provided as additional input to your program). Briefly describe how your code would change (if you think it will) and why.

c. Training is complete now and it is time to put the model in production! Machines maintenance actions are decided once a day at the beginning of the day. Sketch a high level architecture of the system. How would you present the results to the users? Which are the main components and how would they interact? What about CI/CD? Assume that there is one model for each arepa type that take as input the cooking metrics from the last 7 days.

Feel free to make any assumptions that you consider necessary while answering, but don't forget to write them down.