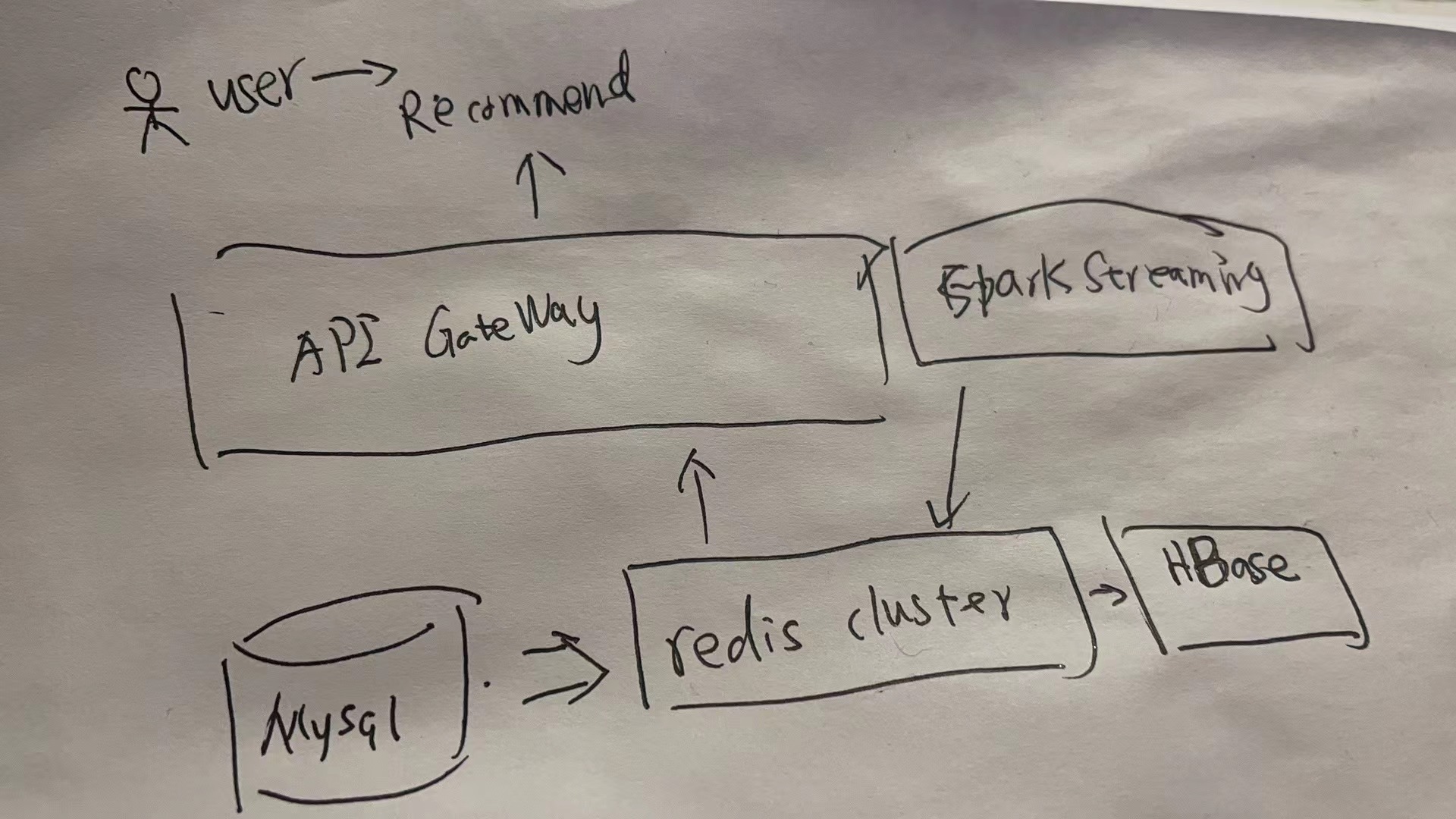
Q1

A .

User browsing data table

We must set expiration date: the expiration time of hot data needs to be set at least a few days (such as 2 days), but considering the process of execution failure and retry of big data tasks, it is necessary to reserve 2 days of task retry and data repair time, so hot data expiration time Set to 4 days. So when the user has not browsed a new product within 4 days, the browsing record viewed by the user is directly queried and displayed from Hbase

B.



C.

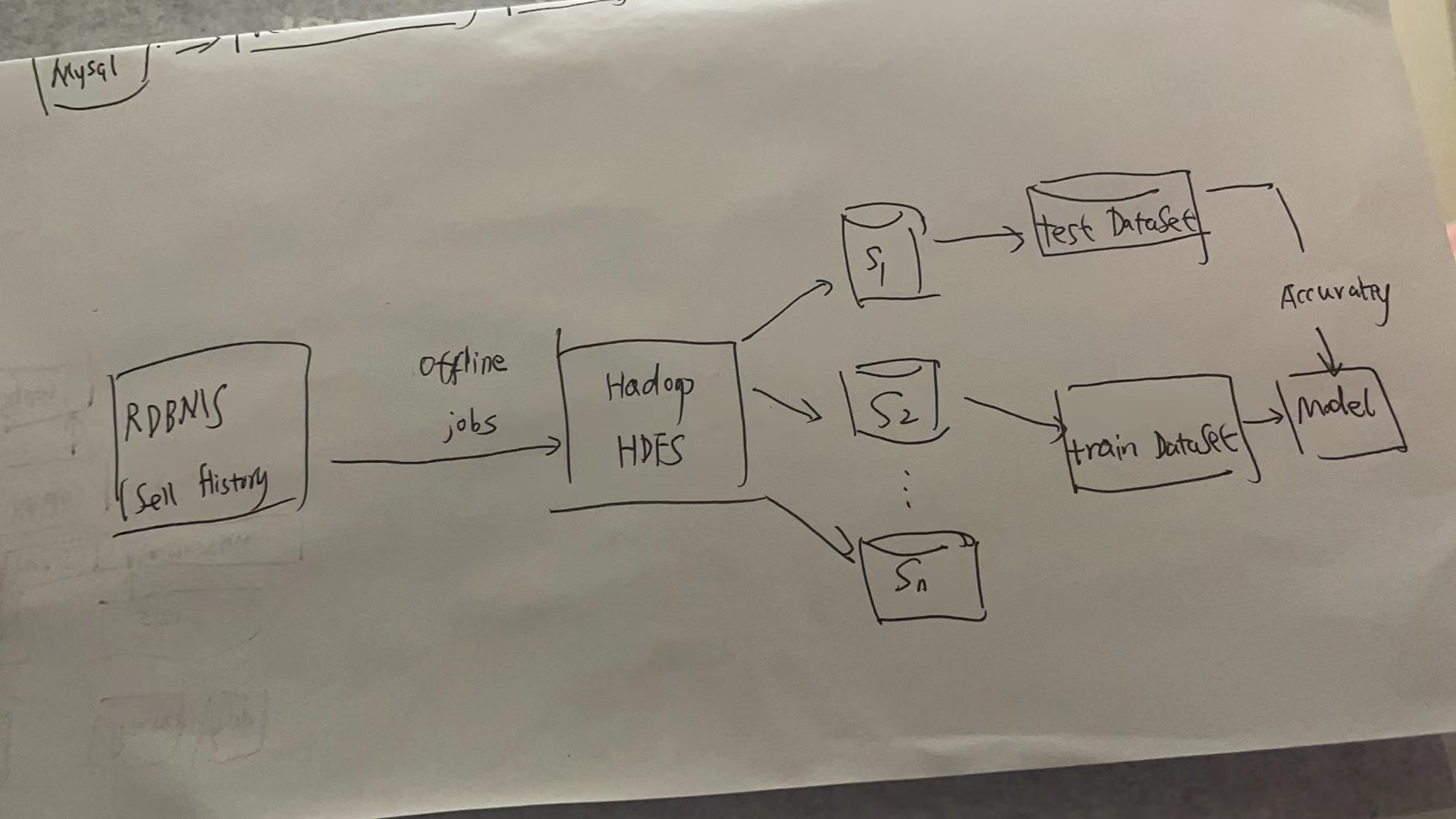
Recommend a user's behavior during this session

Each user recommends based on different behaviors, not a general model

We need key-value storage and map-reduce funacion to relation db (like HBase)

And streaming anlyis tool (like spark streaming) like picture in B.

Q2



We need hadoop(hddfs ) to store the offline dataset , we need batches jobs to export data to hadoop from RDBMS, we need Spark(MLLib) to to machine learning model .

Q3

a.

I choose to use:

Menu data

User review data

As The reasonable collocation of Menu-item is a rule to follow: the customer's eating habits, the meat and the taste of the dishes, some dishes are related to each other, and some dishes are in opposition or competition (negative correlation).

If the rules of customer ordering can be discovered through data miming, then customers’ tastes can be quickly identified.

b.

Not all the learned items.

But the ones what related in that sessions.

C

Association rules should be offline: because a person's tastes can change at will, but not abruptly. I think all offline data can still be considered.

d

It can be easily repaired: we reduce the relative support of item set I to meet the predefined minimum support threshold.

Let this value be lower than the probability range of these niche restaurants to appear in the recommendation.

Q4

A.

Single value decomposition and Clustering method can help:

　　 The single value decomposition algorithm uses the single value decomposition principle of the matrix to decompose the user-product matrix, thereby reducing the dimension of the matrix and extracting the main information.There are many clustering algorithms, such as the clustering algorithm for users using K-means. K-means algorithm is currently the most common fast clustering algorithm

B.

Very intuitive, you can use many numpy/pandas libraries to call directly

Disadvantage: use too many memory, If the memory is not enough, we can save a topK for each item.

c

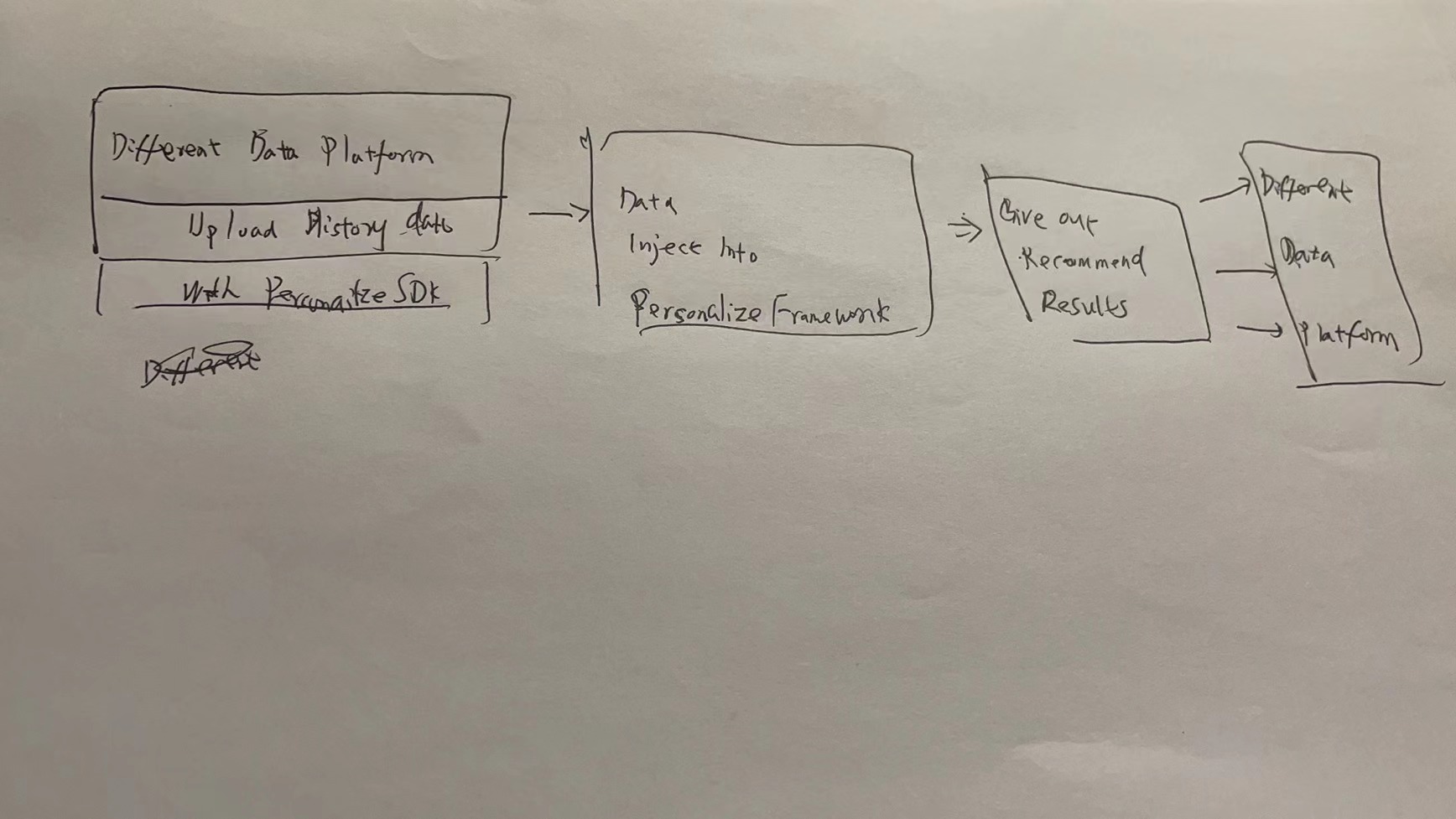
I think it’s ok:

the goal of hybrid recommendation systemis to build a hybrid system that can combine the advantages of different algorithms and models and overcome the shortcomings of them.

Q5

A.

I think Amazon Personalize is the most cost-effective and easiest framework for verifying product prototypes. Its operating requirements is very simple: First of all, we need to provide data, and the most important thing is to have (user-project) interaction data. In addition, it is best to provide data on the items to be recommended and user data

B

Q6

a.

Models built on the basis of steady growth over the past years are facing market changes that have never been experienced. The "knowledge" learned from historical data is no longer applicable to the current market environment, and the prediction results of the model begin to deviate from the actual transaction price in the market. The structural changes(circuit breaker period ) in the market during the new crown also made the model deviate

Act1：we gather and labels the new data , that could reflect new situations during circuit breaker period. The adjustments brought about by these changes in market demand must be done manually by us and market tests are used to confirm that the adjustments are effective.

Adjust the weight of each attribute in the algorithm to make it more dependent on recent data

Act2 : change algorithm not rely on history data too much, like Collaborative recommendation， not consider time-series too much.

B.

Because our data is mainly based on domestic training, cultural differences are very important for recommendation. Some countries are very small and can only obtain very little user data. Training recommendation algorithms on such sparse data will lead to noisy results, because the model It is difficult to identify clear personalization patterns from the data.

In this global context, two important signals may be language and location. They hope that their model can understand not only the location where the user logs in, but also various aspects of the restaurant, such as the source of the video, the language in which it is located, and the popular location.