样卷

——————————————————————————————————————

Q1

a.

Is not enough for completely identifying all possible traffic congestion inducing patterns .

Car-to-car distance

Traffic flow detection at entrances and exits

Without knowing the distance between the cars, the obstruction caused by collisions between cars cannot be avoided;

I don’t know the flow rate of the entrance and exit. There may be a lot of cars coming at one time or a lot of cars going out at the exit at once, and the maximum throughput of the entrance and exit will be exceeded instantaneously.

b.

We don’t know the realtime of car positions in the system.

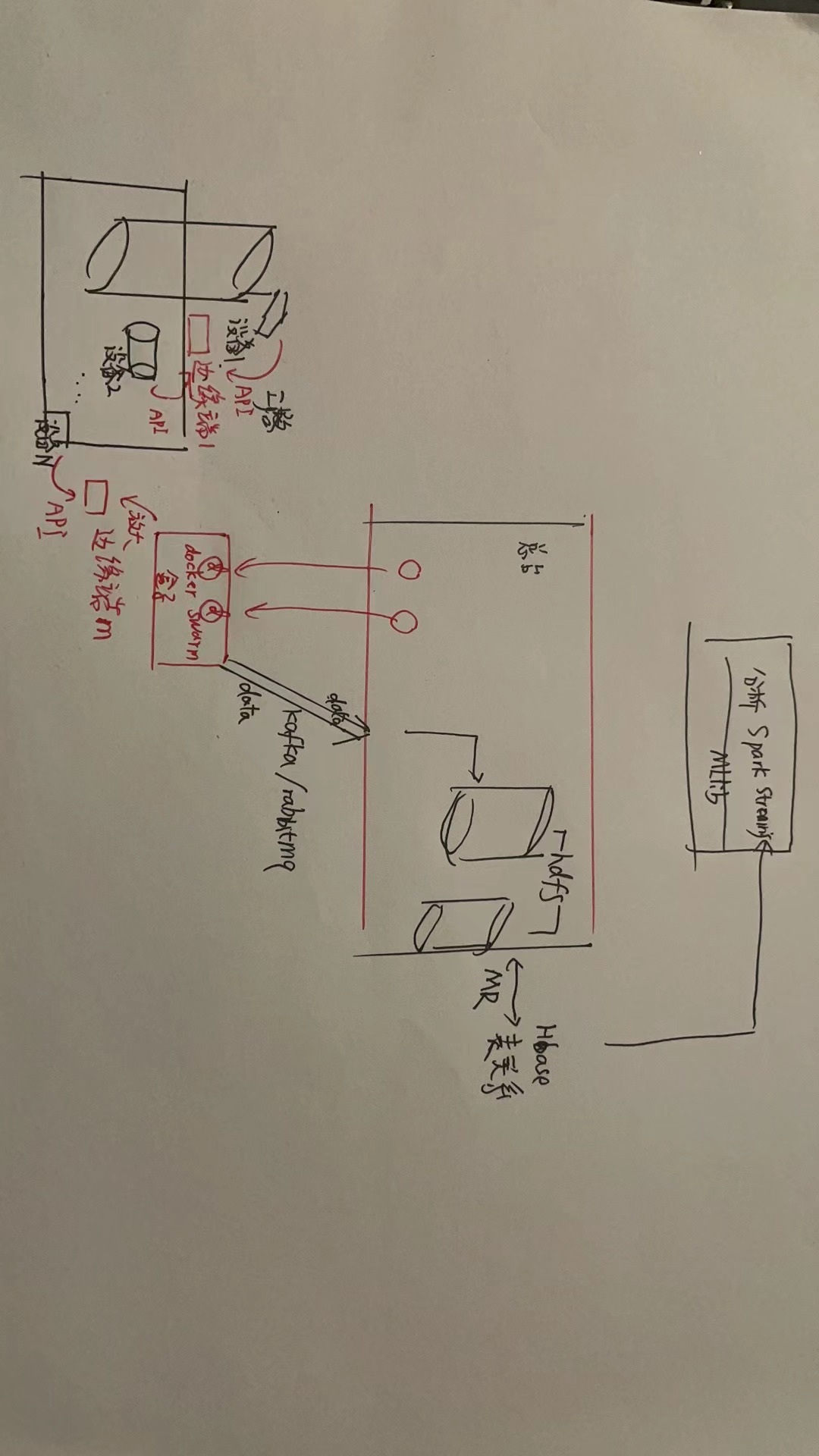
We need collect data from:

Sensors from cars( maybe sensor in car / phones of driver ) to know the location of car in real time.

Use intelligent camera of entrance and exit to detect traffic flow.

Q2

a.



We can use camera with arm processor, that decode video and upload directly by rabbitmq/kafka into amazon S3.

On the use query-in-place analytics tools, we can use spark streaming and spark MLLib to analysis the accidents situations. Or you can use Amazon SNS to analysis video on S3.

(参考04 Introducing Apache Spark Framework - Sep 2021)

b.

We chose ORC and Parquest

ORC AVRO Parquet

Support both structured and un-structured data 1 1 1

Have high compression ratio very good ok good

Support scalability very good ok good

Support scheme modification/evolution 1 1 1

Storage efficiency. read write fast read

(参考10 Data Formats - Sep 2021.pdf

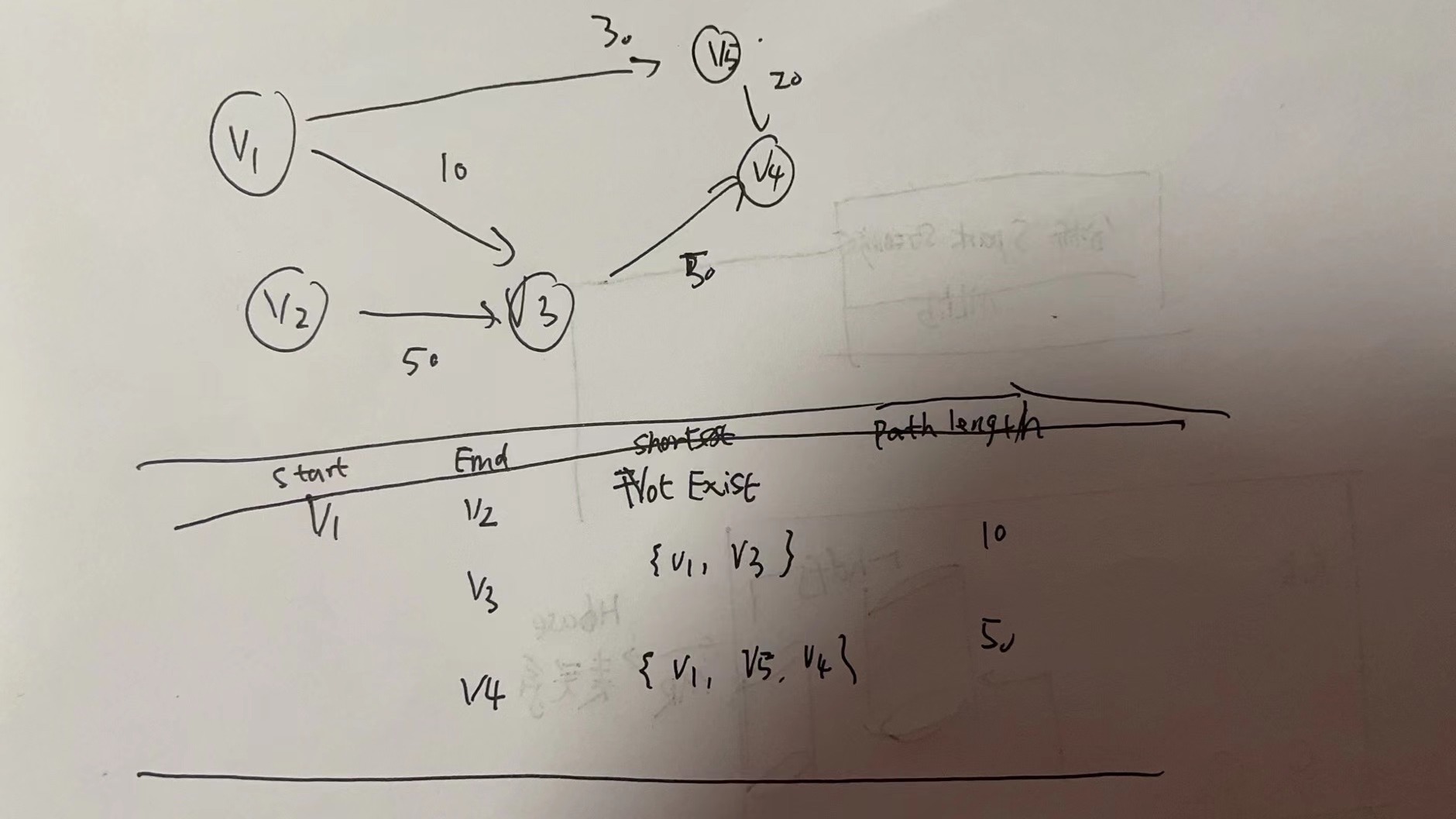
<https://yx91490.github.io/bigdata/hadoop/columnar_storage_parquet_orc.html>

）

Q3

By installing smart cameras at the interval between two road tests, we can build connectivity between various intersections. In addition, the distance between different intersections has long been known (because the distance is fixed).

We can build a transportation network, and then update the connectivity between network nodes in real time, save the adjacency list, and run the dijkstra algorithm from the start point to the end point on this basis to plan the best path in real time.

As for the recommended route, in addition to the shortest distance above, we can also collect the preferences of usual users: according to the distance, the difference between the usual route changes, etc., it can be personalized.

（参考day3-RCS-hybrid-and-commercial）

Q4

a.

We should alert RDBMS to nosql database layer.

Video type files are not suitable for storing in relational databases: it is not convenient to read or write.

When video files are put into redis/mongodb and other key-value pairs database, it can be more convenient to read.

（参考01 Module Introduction - Sep 2021中nosql部分）

b.

The smart camera detects the motorcycle rider who is not wearing the helmet correctly and can send a warning;

Only the camera can detect vehicles with abnormal speed (such as suddenly exceeding the speed of the warning line) or vehicles with too low speed, you can send a warning

The data source :

The analysis from camera to get the real time speed of vehicle.

We use deep learning to detect Helmet

We use deep learning to detect seat belt wearing state

Feature:

Car /motorbycle speed

helmet wearing stay

Belt waring state

（看机器学习相关部分 spark MLib之类）

C.

No, because the dataset is collected , it means it belong to the past, not present.

We need to do analysis of the dataset to use it:

1)We need to clean the dataset, get rid to record that missing value too much

2) we need to transfer time\_interval to time span to hours and days, and use history data to predict today ,we need to consider the time range.

3) we need to add more external data source(like weather data)

4) we construct model that Fitting the positive correlation of travel date and time, weather and location signal heat.

Then we can use the data.

（参考看整个processing big data for analytics，很难说哪一部分）

试卷

——————————————————————————————————————

Q1

a.

Transfer to was kinesis. Be can do :

Time series analysis, real-time dashboards, and real-time alerts and notifications

b.

Now current data architecture is lack of realtime analytics.

1. clean and preprocessing data
2. Real time streaming process and analyst , like spark streaming
3. Give feed-back in realtime

C.

AWS kinesis

Spark Streaming

D.

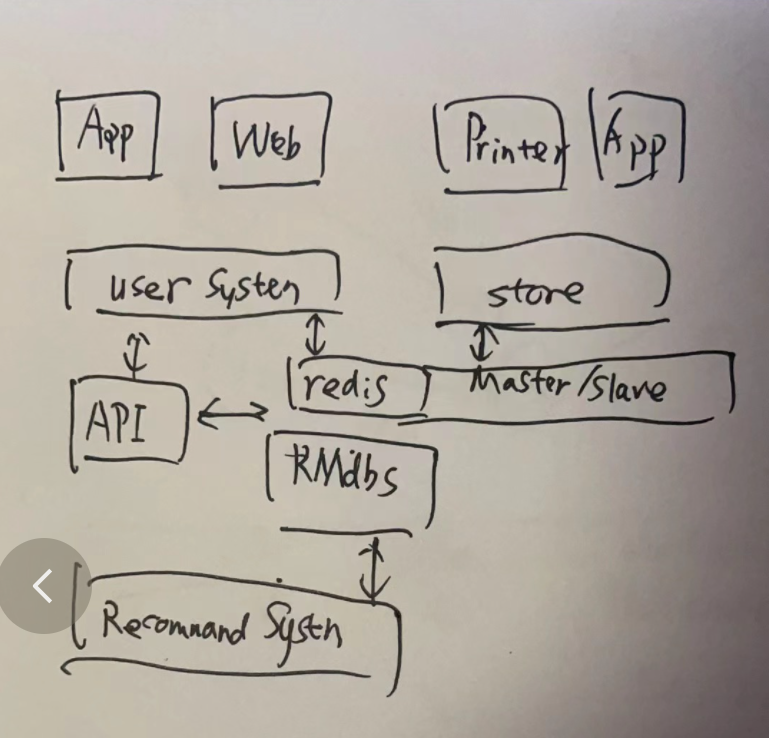
We can analyze the functions that users are mainly interested in and stay for the longest time.

We can also analyze where the clicks are missing, more likely to be errors, etc.

(参考12 Analytics using Spark Machine Learning - Sep 2021)

Q2

A.



As a whole, there is a situation where multiple logics are coupled together and split into services. Each service focuses on doing one thing independently, and then we do application-level fault tolerance. Up to now, we are doing fault tolerance in multiple computer rooms.

For the cache, we use Redis

Recommendation systems These offline training systems pull data from the database and cache them locally for offline training, then generate recommendation results and provide them to the api as the result for fan h

b.

We can load historical data into the haddop（**hdfs** cluster）, analyze the correlation between the location of each buyer and the food purchase, save the analysis result in the cache **redis**, and then provide it to the front-end app through the API, so as to obtain better purchase retention Rate and shorter delivery time (you can notify stores in different regions for different stocking)

（看推荐系统相关章节）

Q3

A.

Perform community structure analysis for each social tag network at the site level(same user’s session have buy same cust\_id Food\_id ), cluster social tags belonging to the same subject area to form a community structure, and then use centrality analysis methods in social network analysis (such as degree indicators, feature vectors) Indicator), find important social tags in each community to generate social tag usage documents, and then match all social tag usage documents with each social tag network of user u to form multiple local interests of user u View, the fine-grained user interest model.

(参考4 managing business analytics)

B.

(1) Trajectory matching on the road network: According to the starting point of the user's order and the location of the store, a trajectory matching is constructed

(2) Real-time road condition model: Accurately model the traffic conditions on the road network to reflect the true driving speed of the road.

(3) Empirical road condition model: Analyze historical trajectory data to obtain historical traffic conditions on the road network to predict future road conditions. Different vehicles give different weights and speeds

(4) Path planning algorithm: Comprehensive use of offline index and online search algorithm to achieve efficient path planning

(参考4 managing business analytics和推荐系统）

Q4

A.

Our food delivery system has to realize the integrated intelligent green coverage of dispatchers, vehicles, goods, scenes, warehouses, and urban terminal distribution.

The vehicle path planning algorithm has been applied to a number of services. In the vehicle distribution link, the number of vehicles used and the distance of the vehicle are reduced, and the electric exchange box transport vehicle can improve the efficiency of multi-frequency transportation from warehouse to site and meet the demand for multi-frequency delivery.

In the picking link inside the warehouse, the walking distance of pickers is reduced. Subverting the previous scenario of "people looking for goods", intelligent order sorting technology and dynamic positioning technology directly realize pre-sorting and deliver goods directly to couriers, realizing mobile outlets and reducing picking time in the intelligent warehousing link.

In addition, the vehicle path planning algorithm can also help the takeaway delivery staff to plan the delivery route, reducing the turnover from the front-end order to the end-end goods delivery, and the delivery is directly based on the route. Thereby improving customer experience and greatly reducing distribution costs.

B.

The time from the time the dinner browses to the actual start of ordering a food

The total time dinner to finish order

The time the diner have to wait tilll get a dilvery boy notification to his app

c.

Naive Bayesian machine learning can be used to predict how long it will take to deliver meals, based on previous orders at the same location and similar

(参考12 Analytics using Spark Machine Learning - Sep 2021 和Module 07 Model Aggregation\_Oct2021\_v1\_student ）

d.

For the case where the normal range is not set, we can use unsupervised machine learning to run a clustering algorithm on all electricity usage data. Then find outliers, and then analyze the data of the outliers, whether it is a real abnormal situation, or caused by the extreme situation of too many or too few customers.

In the case of setting the normal range, we can use various regression methods to compare the data with the threshold of the normal situation. For the part that exceeds the threshold, we will conduct targeted data analysis, and at the same time conduct correlation analysis between the number of customers and the time required for the dishes in the menu to see if it is a reasonable range

（参考13 Machine Learning on Spark - Sep 2021和其他机器学习部分）