

Question 1: Role-play [10 marks]

A. Background:

You are a data scientist in the data research team at XYZ Limited.

The flagship product of XYZ is a data platform providing real-time and historical financial data of cryptocurrencies, e.g., bitcoin. There are two kinds of users of the data platform: (1) paid subscribers: who can access any real-time data and all historical data of 1,000 cryptocurrencies; and (2) free users: who can only access data of the past 7 days for 10 cryptocurrencies and cannot access real-time data of the recent hour.

At the moment, there are 1,000 paid subscribers and 10,000 registered free users. The maximum number of concurrent users of the data platform is 500.

B. Current situation:

XYZ's current database system has been MySQL (on a single server machine) since the company was founded in Sep 2012. Recently, the CEO is discussing a partnership with ABC Limited, which is a financial consulting firm with a large customer base. To facilitate the partnership, ABC requires XYZ to provide social sentiment data – analyzing whether people on social networks are bullish or bearish on each cryptocurrency.

If the partnership is successful, XYZ's platform will also be used by ABC's customers. That would mean the number of concurrent users will increase to 10,000.

ABC also commented that MySQL is outdated technology and should be replaced. The CEO is not from a technical background and is not able to reply to ABC. Thus, the CEO requested the data research team to submit a proposal to review and revamp the existing data management system if appropriate. The proposal should of course address the need for the coming social sentiment feature.

C. Your task:

The head of the data research team has appointed you to study the technical part of the proposal. Other aspects, like budget and schedule, will be handled by your colleagues. Your head does not have a preferred solution. She will adopt your idea as long as it makes sense to her. Write the technical part of the proposal that includes at least the following parts:

1. Your proposed data management solution
 - a. How the existing data platform can be migrated to use your proposed solution
 - b. How the new feature of social network sentiment is supported by your proposed solution

2. Technical justifications for using your proposed system
 - a. Make sure you have also addressed all the concerns / comments from your colleagues (See Section D)
3. Good alternative approaches that you have considered
4. Reasons why the alternatives are not used

Note:

1. You should propose only ONE solution. Your solution should be specific, e.g., don't say you use a relational database management system, but give the actual system name in this case, e.g., MySQL.
2. Your solution should include implementation information like what data to store and where data are stored etc.
3. The primary reader of the proposal is your team head who is very technical. Make sure your justifications are technically sound and clear.
4. Your proposal may be later read by your CEO or other parties who are not technical. Make sure your content can be understood by a layman.
5. This is a formal proposal. Use a proper format for your proposal.

D. Information and views from your colleagues:

Your colleagues have given you more information about the current situation and their views for your reference. **Note that their opinions may not be the best.** Please use your own judgement to design your solution.

1. CEO

ABC wants to have sentiment data updated every hour. Similar to our existing data platform, historical sentiment data should be provided.

2. CTO

I have a concern about the risk of migration. If the current system is working well now, we should try to make use of the current system as much as possible.

3. Head of IT support

The current server (one machine) can only handle at most 1,000 concurrent users. You need to think about how to handle 10,000 concurrent users in the future.

4. Head of Sales

Some customers ask for extending the functions of our current platform. For example, adding more data attributes like a 5-day moving average. Adding these attributes will definitely increase the competitiveness of our platform.

5. Head of Data Research

I suggest we only work on the text part of messages on social networks and only in English. We ignore images and videos. Even if this is the case, we are talking about around 50GB of raw data per day. There are at least 5 years of social network data available. We can always download historical data from social networks at any time. However, the download speed is slow, around 50GB per hour.

6. Data scientist for NLP modelling

The text on social networks is often non-standard English. There are many typos too. I don't know if the NLP model will work well.

E. Database schema and sample data

The current data storage in MySQL has two tables. One table keeps real-time data. The other table keeps historical data. As our platform only provides hourly versions or daily versions for historical data, the current data storage is around 200GB.

Table 1: Realtime

Schema: (symbol: varchar(20), open: float, high: float, low: float, close: float, volume: float, quoteAssetVolume: float, numOfTrades: int, takerBaseVolume: float, takerQuoteVolume: float)

Sample data:

symbol	open	high	low	close	volume	quoteAssetVolume	numOfTrades	takerBaseVolume	takerQuoteVolume
BTCUSD	19360.390000	19365.320000	19320.730000	19329.720000	3.964925e+03	7.669239e+07	117304	1.979235e+03	3.828514e+07
ETHUSD	1346.350000	1347.110000	1342.650000	1343.610000	8.702463e+03	1.170066e+07	13390	4.043193e+03	5.435564e+06
FTMUSD	0.204600	0.204800	0.202400	0.203100	1.691242e+06	3.441736e+05	976	5.226340e+05	1.063287e+05
BNBUSD	273.700000	274.200000	273.100000	274.100000	4.534640e+03	1.241037e+06	3037	2.789261e+03	7.634511e+05
SOLUSD	28.450000	28.470000	28.230000	28.300000	5.145080e+04	1.458039e+06	1814	1.602215e+04	4.541063e+05
XRPUSD	0.456400	0.456800	0.452200	0.452900	9.439841e+06	4.288438e+06	3824	2.924272e+06	1.327629e+06
AVAXUSD	15.790000	15.810000	15.710000	15.750000	1.530210e+04	2.411120e+05	659	4.607450e+03	7.262861e+04
WAVESUSD	3.149000	3.151000	3.129000	3.137000	2.985449e+04	9.369833e+04	363	1.080588e+04	3.390359e+04
GALAUSD	0.033170	0.033180	0.032980	0.033050	7.201565e+06	2.381533e+05	764	2.471119e+06	8.172822e+04
CVPUSD	0.403100	0.403300	0.400200	0.400800	8.997980e+04	3.615620e+04	531	2.200140e+04	8.851406e+03
GRTUSD	0.080200	0.080200	0.079300	0.079600	8.035650e+05	6.405959e+04	233	3.615100e+05	2.882049e+04
DOTUSD	5.930000	5.940000	5.900000	5.920000	1.322765e+05	7.826993e+05	650	5.264047e+04	3.116123e+05
RUNEUSD	1.447000	1.447000	1.438000	1.442000	1.232339e+05	1.776310e+05	543	6.630090e+04	9.553267e+04
POLSUSD	0.430000	0.431000	0.430000	0.430000	1.467500e+03	6.322981e+02	11	1.273100e+03	5.487061e+02
SANDUSD	0.737700	0.738200	0.732600	0.735300	5.186180e+05	3.815443e+05	1096	2.343390e+05	1.723998e+05
SHIBUSD	0.000010	0.000010	0.000010	0.000010	2.988778e+10	2.965343e+05	697	1.148699e+10	1.139841e+05
ADAUSD	0.360400	0.360800	0.357100	0.357300	2.484028e+06	8.905731e+05	1838	1.084367e+06	3.887014e+05
JASMYUSD	0.004503	0.004509	0.004459	0.004471	6.957430e+07	3.115330e+05	1623	3.106419e+07	1.390748e+05
NEARUSD	2.937000	2.940000	2.912000	2.919000	1.731424e+05	5.057547e+05	1251	6.095510e+04	1.780564e+05
ATOMUSD	11.711000	11.724000	11.636000	11.650000	2.144239e+04	2.506514e+05	945	8.609810e+03	1.006314e+05
EGLDUSD	57.880000	57.950000	57.680000	57.860000	3.082440e+03	1.783071e+05	627	1.412580e+03	8.171770e+04
MATICUSD	0.899400	0.901900	0.890900	0.893100	3.298120e+06	2.956076e+06	5994	1.607961e+06	1.440091e+06
MANAUSD	0.605400	0.605700	0.601500	0.603200	3.232100e+05	1.950418e+05	754	1.553600e+05	9.373816e+04
QGNUSD	0.137100	0.137300	0.136100	0.136700	2.382780e+05	3.259827e+04	148	9.541000e+04	1.305348e+04
TRXUSD	0.061600	0.061630	0.061290	0.061310	1.961557e+07	1.205276e+06	2603	7.855071e+06	4.829485e+05

Table 2: Historical

Schema: (date: datetime, open: float, high: float, low: float, close: float, volume: float, quoteAssetVolume: float, numOfTrades: int, takerBaseVolume: float, takerQuoteVolume: float, symbol: varchar(20))

Sample data:

date	open	high	low	close	volume	quoteAssetVolume	numOfTrades	takerBaseVolume	takerQuoteVolume	symbol
2022-10-13 20:00:00	18754.09	19030.00	18190.00	18385.30	46815.95336	8.630864e+08	820136	22850.44627	4.213806e+08	BTCUSD
2022-10-13 21:00:00	18385.30	18479.96	18273.00	18432.14	29411.02329	5.399413e+08	504684	14663.27683	2.692024e+08	BTCUSD
2022-10-13 22:00:00	18430.95	18496.65	18386.97	18415.00	24250.43976	4.469228e+08	376597	12062.05619	2.223056e+08	BTCUSD
2022-10-13 23:00:00	18415.65	19049.99	18403.65	18958.59	45452.83697	8.544386e+08	687742	23502.16082	4.417294e+08	BTCUSD
2022-10-14 00:00:00	18958.59	19197.61	18895.08	19086.39	28162.97885	5.360659e+08	492864	14120.48820	2.688064e+08	BTCUSD
2022-10-14 01:00:00	19084.62	19228.00	19065.89	19157.90	19234.24066	3.682498e+08	343544	9680.63236	1.853480e+08	BTCUSD
2022-10-14 02:00:00	19157.90	19513.79	19142.07	19422.01	24185.47275	4.664837e+08	406443	12384.17983	2.388920e+08	BTCUSD
2022-10-14 03:00:00	19420.14	19470.00	19311.36	19369.94	20938.94370	4.058176e+08	365063	10750.22118	2.083524e+08	BTCUSD
2022-10-15 03:00:00	19213.84	19269.06	19106.16	19168.18	14038.45535	2.695487e+08	259493	6789.09253	1.303659e+08	BTCUSD
2022-10-15 04:00:00	19168.18	19217.64	19100.01	19175.65	6801.72861	1.304439e+08	149346	3327.83633	6.382852e+07	BTCUSD
2022-10-15 05:00:00	19174.43	19207.15	19070.37	19107.94	5330.36918	1.021237e+08	117436	2584.75501	4.952794e+07	BTCUSD
2022-10-15 06:00:00	19107.94	19166.41	19086.99	19159.83	5854.10575	1.119886e+08	121563	2992.04497	5.723802e+07	BTCUSD
2022-10-15 07:00:00	19159.83	19197.38	19141.28	19176.93	4375.24549	8.388242e+07	93911	2195.97620	4.210260e+07	BTCUSD
2022-10-15 08:00:00	19176.93	19227.68	19171.87	19200.86	4978.42383	9.558172e+07	125079	2560.06638	4.915150e+07	BTCUSD

F. Marking criteria

Item	Description	Marks
Feasibility	Does your solution work? Have you described your solution clearly?	2.5
Justifications	Have you provided enough justifications? Have you addressed all colleagues' concerns and comments? Are your justifications correct and logical?	4.5
Alternatives	Have you considered alternatives in the proposal? Do they work? Are your justifications about why your proposal is better logical and sound?	2
Presentation	Is your proposal clear and properly organized?	1

