**Tech Task Result Report – Huijing Wu**

**The url of my repo:**

<https://github.com/HuijingWu/tech-interview-task.git>

On top the notes in codes, I am gonna briefly summarize and explain the checks I conducted on the trades & users tables, quality checks on every column by order:

**For the trades table:**

**login\_hash:**

1. checked if the hash length is in the expected length

result: clear

1. checked if we have duplicate login\_hash values

result: we have a lot of duplicate login\_hash in the trades table.

actions taken:

look into one case, for example, for login\_hash = ‘756EAF7A606437A4294A0E2E806ACF6E’, we have 556 records sharing the same login\_hash, different tick\_hash and open\_time with huge date gap over 10 days. For cases like this, possible reasons might be: long-term inverstment strategies; intermittent trading activity; inconsistent trading behavior, but for significant time gap like this case, data anomalies, like the inaccuracies in the timestamps or other trading parameters that create the appearance of a significant time gap, might worth considering. Fruther actions need to be taken after understaing the way how the data has been collcted and the logic behind this column.

**ticket\_hash:**

1. checked if the hash length is in the exepected length

result:clear

1. checked if we have duplicate ticket\_hash in the trades table

result: clear

**server-hash:**

1. checked if the hash length is in the exepected length

result:clear

**symbol:**

1. checked if there is any unexpected characters in the symbol column:

result: two records with unexpected characters found

Unexpected characters found in symbol column: USD,CHF

Unexpected characters found in symbol column: USD,CHF

1. checked if we have trades between the same assets, say “USDUSD”

result:clear

**digits:**

1. checked if digits is with the reasonable range (2-5)

result : we have 46k rows with final prices of digits smaller than 2

actions: more information needed to identify the reason of the failure to capture more detailed prices, whether its because the final price is an int itself or we failed to capture more deatied info.

**cmd:**

1. checked if cmd has any other value apart from 0 and 1

result: clear

**volume:**

1. checked if volume is bigger than 0

result: clear

**open\_time:**

1. checked if the date column is of right datetime format

result: clear

1. checked if the trading open time is later than the close time or not

result: clear

**open\_price:**

1. checked if the open price is bigger than 0

result:clear

**close\_time:**

1. same as the open\_time

**contractsize:**

1. checked if the contract size is bigger than 0

result:clear

**For the users table:**

**login\_hash:**

1. checked if the hash length is expected

result:clear

1. checked if we have duplicate login\_hash in the user table

result: many duplicate login\_hash values found

actions: same as the login\_hash in the trades table

**server\_hash:**

1. checked if the hash length is expected

result: clear

**country\_hash:**

1. checked if the hash length is expected

result: clear

**currency:**

1. checked if we have any unusual currency units

result: two records of NZD found

**enable:**

1. checked if enable column has any values apart from 0 and 1

result: clear

**Other checks:**

1. checked all the records which can be found in one table but not the other

result: there are quite a lot login\_hash in the trades table which cannot be matched in the users table, vice versa.

actions: possible reasons for this: data synchronization problems; different data sources; data filtering or exclusion applied. More detailed information of how the data pipelines have been developed needed to locate the reason.