TASK 1: Figure out best you can what the problem is.

Total revenue according to **AIQ_User_Summary** file:

Query : select sum(`total_spend`) as `AIQTotal` from `AIQ_User_Summary` ;

Result : $ 29,162,044

Total revenue according to **Johnny_User_Summary** file:

Query : select sum(`total_spend`) as `JohnnyTotal` from `Johnny_User_Summary` ;

Result : $ 33,90,4525

Difference between these values is **$ 4,742,481**

As Next step, I want to check if all the user_ids and batch_ids in delta files were up to date compared to **Johnny_User_Summary** table. This can tell us if some of the user_ids are missing in delta files given to AIQ.

```
104
105 select first.`user_id`,first.`batch_id`, second.`user_id`,second.`batch_id`
106 from (select * from `Johnny_User_Summary`) first
107
108 inner join
109
110 (select user_id, max(batch_id) as batch_id from
111 (select *,"delta0" as file_name from delta0
112 union all
113 select *,"delta1" as file_name from delta1
114 union all
115 select *,"delta2" as file_name from delta2
116 union all
117 select *,"delta3" as file_name from delta3
118 union all
119 select *,"delta4" as file_name from delta4) union_query group by user_id) second
120
121 on first.`user_id` = second.`user_id` and first.batch_id = second.batch_id ;
122
```

Query Favorites Query History

user_id	batch_id	user_id	batch_id
5988	5731	5988	5731
5989	298	5989	298
5982	1489	5982	1489
5983	4623	5983	4623
5980	3496	5980	3496
5981	7180	5981	7180
5986	1281	5986	1281
5987	3682	5987	3682
5984	1950	5984	1950
5985	949	5985	949
6970	406	6970	406
6796	5132	6796	5132
6797	1295	6797	1295
6794	7017	6794	7017

No errors; 10000 rows affected, taking 45.9 ms

We can see that all 10,000 users and their updated **batch_ids** were present in delta files given to AIQ.

Next, Let's check if there are any user_ids in **AIQ_User_Summary** table whose latest **batch_ids** are greater than **Johnny_User_Summary** table.

```
select Jsum.`user_id`, Jsum.`batch_id` as J_batchID, Jsum.`total_spend`, ASum.`batch_id` as A_batchID,
ASum.`total_spend`
from `AIQ_User_Summary` as ASum
join `Johnny_User_Summary` as JSum
on ASum.`user_id` = JSum.`user_id`
where ASum.`batch_id` > JSum.`batch_id`;
```

Result : ***0 rows***

Therefore, there are no user_ids with higher batch_ids in AIQ table than Johnny Summary table.

Now, Let's check if vice-versa is true i.e if there are any user_ids in **Johnny_User_Summary** table whose batch_ids greater than **AIQ_User_Summary** table.

```
select Jsum.`user_id`, Jsum.`batch_id` as J_batchID, Jsum.`total_spend`, ASum.`batch_id` as A_batchID,
ASum.`total_spend`
from `AIQ_User_Summary` as ASum
join `Johnny_User_Summary` as JSum
on ASum.`user_id` = JSum.`user_id`
where JSum.`batch_id` > ASum.`batch_id`;
```

Result : ***1841 rows***

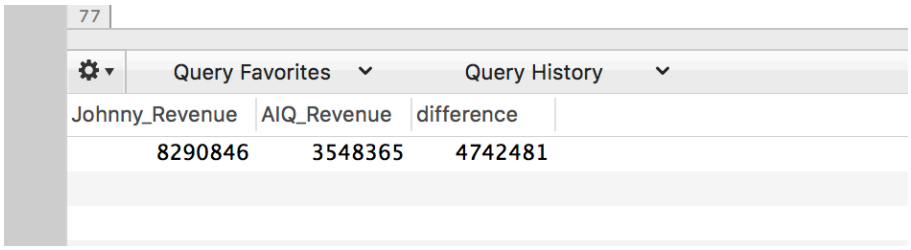
user_id	J_batchID	total_spend	A_batchID	total_spend
5988	5731	4208	3164	3500
5983	4623	4537	4271	3061
6796	5132	4417	997	1062
6792	5038	4271	3349	3583
6790	5267	4770	619	1912
6791	5893	4230	3156	3553
6798	4631	4537	3558	3160
272	5828	4835	194	1459
276	5560	4179	1324	1054
279	4541	4210	680	1604
9252	4957	4365	1293	239
3519	4909	4751	4366	3562
3511	5946	4654	1732	1557
3516	4886	4025	2828	2217
3514	5642	4857	2143	2078
2688	5353	4889	1247	842
2687	4589	4840	1711	453

We can clearly see that these **1841 rows** have different batch ids which can be the reason why we got different total revenue amounts.

As next step, I want to check if the difference between the sum of these two amounts is same as difference between AIQ and Johnny total revenue values.

```
select sum(Jsum.`total_spend`) as Johnny_Revenue, sum(ASum.`total_spend`) as AIQ_Revenue,
sum(Jsum.`total_spend`) - sum(ASum.`total_spend`) as `difference`
from `AIQ_User_Summary` as ASum
join `Johnny_User_Summary` as JSum
on ASum.`user_id` = JSum.`user_id`
where ASum.`batch_id` <> JSum.`batch_id`;
```

Result :



This difference **$ 4,742,481** is exactly the same as we found out in the first step.

So far we found the rows which are causing the problem. As next step, I want to check if these rows are coming from any particular delta file or all of them.

Query & Result :

In the above picture we can see that all the rows which are causing the problem are from **delta3.csv** file.

But, we should note that there are some rows in **delta3.csv** which got ingested correctly.

To be precise,

Query : select count(*) from delta3;

Result : 2473

So **batch_ids** of **these 1841 users** in **delta3.csv** file are causing the difference in total revenue values**.**