












Receipts DQ –

```
select receipt_id, count(*) from receipts_fact
group by receipt_id
having count(*) > 1
```

Data Output	Messages	Notifications
        		
receipt_id [PK] character varying (700) 	count bigint 	

```
ALTER TABLE receipts_fact
ADD CONSTRAINT pk_receipts_fact PRIMARY KEY (receipt_id);
```

```
select * from receipts_fact
where "rewardsReceiptStatus" = 'FINISHED'
```

125

where "rewardsReceiptStatus" = 'FINISHED'

Data Output

Messages

Notifications

<

There are some NULLS in bonusPoints even when status is Finished and pointsawardeddate is also present. Is this behaviour expected?

```
select * from receipts_fact
where "rewardsReceiptStatus" = 'REJECTED' and "pointsAwardedDate" is not null
```

125 `where "rewardsReceiptStatus" = 'REJECTED' and "pointsAwardedDate" is not null`

	receiptId [PK] character varying (700)	userId character varying (700)	rewardsReceiptStatus character varying (500)	bonusPointsEarned double precision	bonusPointsEarnedReason character varying (2000)	pointsEarned double precision
1	5f9c74f70a7214ad07000037	5f9c74f7c88c1415cbddb839	REJECTED	750	Receipt number 1 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	7
2	5fa8d5730a7214adc30001c3	5fa8d573770b516dda4f4d21	REJECTED	750	Receipt number 1 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	7
3	5fb1bd510a7214ad890000e3	5fb1bd519c09d811e5211cef	REJECTED	750	Receipt number 1 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	7
4	5fb2da580a7214ad89000114	5fb2da582357ec11d4ecc403	REJECTED	750	Receipt number 1 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	7

Is the above behavior expected? 4 records where the status is rejected still points awarded date is populated.

```
select integration_id, count(*) from receipts_line_item
group by integration_id
having count(*) > 1
```

	integration_id [PK] character varying (500)	count bigint
--	--	-----------------

```
ALTER TABLE receipts_line_item
ADD CONSTRAINT pk_integration_id PRIMARY KEY (integration_id),
ADD CONSTRAINT fk_receipt_id FOREIGN KEY (receipt_id) REFERENCES
receipts_fact(receipt_id);
```

```
select * from receipts_fact
where receipt_id not in (select receipt_id from receipts_line_item) and "rewardsReceiptStatus"
= 'FINISHED'
```

	receiptId [PK] character varying (700)	userId character varying (700)	rewardsReceiptStatus character varying (500)	bonusPointsEarned double precision	bonusPointsEarnedReason character varying (2000)	pointsEarned double precision
1	6009eb000a7214ada2000003	6009e60450b3311194385009	FINISHED	250	Receipt number 3 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	250
2	600aff160a720f053500000c	600afb2a7d983a124e9aded0	FINISHED	500	Receipt number 2 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	500

We observe 2 records where rewards have been given but the line item table does not have those receipt_ids

```

SELECT 'user_table ids' AS u_ids, COUNT(distinct user_id) AS count
FROM users
WHERE user_id NOT IN (SELECT user_id FROM receipts_fact)
UNION
SELECT 'receipt_fact ids' AS u_ids, COUNT(distinct user_id) AS count
FROM receipts_fact
WHERE user_id NOT IN (SELECT user_id FROM users);

```

	u_ids text	count bigint
1	receipt_fact ids	117
2	user_table ids	71

This is a clear violation for our data model which needs to be handled firstly. There are some user_ids present in receipts_fact table which are not in the users_dim table. Hence this leads to failure of referential integrity.

Also, let's analyze Main fact v/s line level details for one receipt_id –

```
select * from receipts_fact where receipt_id = '5ff1e1b60a7214ada100055c'
```

	bonusPointsEarnedReason character varying (2000)	pointsEarned double precision	purchasedItemCount double precision	totalSpent double precision	createDate timestamp witho
1	50 Receipt number 5 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a...	8850	10	290	2021-01-03 10:2

```
select * from receipts_line_item where receipt_id = '5ff1e1b60a7214ada100055c'
```

		barcode character varying (30)	description character varying (2000)	finalPrice double precision	itemPrice double precision	needsFetchReview boolean	preventTargetGapPoints boolean	quantityPurchased double precision	userFI charac
1	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
2	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
3	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
4	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
5	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
6	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
7	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
8	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
9	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]
10	fdof1	034100573065	MILLER LITE 24 PACK 12OZ CAN	29	29	[null]	[null]		1 [null]

For this receipt_id it's consistent for the quantityPurchased (sum =10) and totalSpent = (price*10) = 290

Thus we can observe, that fact table is aggregated and line_item table contains granular/line level details for the receipt_id

But,

```
with t as (SELECT rl.receipt_id,
  COALESCE(SUM(rl."quantityPurchased"), 0) AS total_quantity_rl,
  rf."purchasedItemCount" AS total_purchased_rf,
  CASE
    WHEN COALESCE(SUM(rl."quantityPurchased"), 0) = rf."purchasedItemCount" THEN
'Consistent'
    ELSE 'Inconsistent'
  END AS consistency_status
FROM receipts_line_item rl
INNER JOIN receipts_fact rf ON rl.receipt_id = rf.receipt_id
GROUP BY rl.receipt_id, rf."purchasedItemCount")
```

```
select consistency_status, count(*)
from t
group by consistency_status
```

	consistency_status text	count bigint
1	Consistent	586
2	Inconsistent	93

We observe that around 93 records are Inconsistent, w.r.t non matching between fact and line level fields.

On further analysis, I found that its due to Nulls as well as counts mismatch.

```
ALTER TABLE receipts_line_item
ADD CONSTRAINT pk_integration_id PRIMARY KEY (integration_id),
ADD CONSTRAINT fk_receipt_id FOREIGN KEY (receipt_id) REFERENCES
receipts_fact(receipt_id);
```

Also,

```
select * from receipts_fact rf
inner join receipts_line_item rl on rf.receipt_id = rl.receipt_id
where rf."rewardsReceiptStatus" = 'FINISHED' and rl."pointsNotAwardedReason" is Not NULL
```

	userFlaggedPrice double precision	userFlaggedQuantity double precision	needsFetchReviewReason character varying (200)	pointsNotAwardedReason character varying (2000)	rewardsGroup character varying (500)	rewardsProductPartnerId character varying (500)	us ch
1	10	1	USER_FLAGGED	Action not allowed for user and CPG	DORITOS SPICY SWEET CHILI SINGLE SERVE	5332f5f3e4b03c9a25efd0ba	D
2	[null]	[null]	[null]	Action not allowed for user and CPG	DORITOS SPICY SWEET CHILI SINGLE SERVE	5332f5f3e4b03c9a25efd0ba	[n]
3	[null]	[null]	[null]	Action not allowed for user and CPG	OLD EL PASO BEANS & PEPPERS	5332f5f3e4b03c9a25efd0ae	[n]
4	[null]	[null]	POINTS_GREATER_THAN_THRESHOLD	Action not allowed for user and CPG	ANNIE'S HOMEGROWN MULTI-SERVING MAC & CHEESE	5332f5f3e4b03c9a25efd0ae	[n]
5	10	1	[null]	Action not allowed for user and CPG	DORITOS SPICY SWEET CHILI SINGLE SERVE	5332f5f3e4b03c9a25efd0ba	[n]
6	[null]	[null]	[null]	Action not allowed for user and CPG	PEPSI 12 OZ 12 PACK	5332f5f3e4b03c9a25efd0ba	[n]
7	[null]	[null]	[null]	Action not allowed for user and CPG	DORITOS NACHO CHEESE MULTI SERVE	5332f5f3e4b03c9a25efd0ba	[n]
8	[null]	[null]	[null]	Action not allowed for user and CPG	GOODBELLY BARS	5332f5f3e4b03c9a25efd0ae	[n]
9	[null]	[null]	POINTS_GREATER_THAN_THRESHOLD	Action not allowed for user and CPG	ANNIE'S HOMEGROWN MULTI-SERVING MAC & CHEESE	5332f5f3e4b03c9a25efd0ae	[n]
10	[null]	[null]	[null]	Action not allowed for user and CPG	DORITOS SPICY SWEET CHILI SINGLE SERVE	5332f5f3e4b03c9a25efd0ba	[n]

Need to understand context for pointsNotAwardedReason field.

```
select * from receipts_line_item
where barcode = '4011'
```

<div><div><div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div></div></div>									
	<div><div><div></div><div></div><div></div></div><div>id</div><div>varying (500)</div></div>	<div><div><div></div><div></div><div></div></div><div>partnerItemId</div><div>integer</div></div>	<div><div><div></div><div></div><div></div></div><div>pointsPayerId</div><div>character varying (500)</div></div>	<div><div><div></div><div></div><div></div></div><div>barcode</div><div>character varying (30)</div></div>	<div><div><div></div><div></div><div></div></div><div>description</div><div>character varying (2000)</div></div>	<div><div><div></div><div></div><div></div></div><div>finalPrice</div><div>double precision</div></div>	<div><div><div></div><div></div><div></div></div><div>itemPrice</div><div>double precision</div></div>	<div><div><div></div><div></div><div></div></div><div>needsFetchReview</div><div>boolean</div></div>	<div><div><div></div><div></div><div></div></div><div>preventTargetGapPoints</div><div>boolean</div></div>
1	0a720f0523000575	1	[null]	4011	ITEM NOT FOUND	26	26	false	true
2	0a720f052300056b	1	[null]	4011	ITEM NOT FOUND	1	1	[null]	[null]
3	0a7214ada100056f	1	[null]	4011	ITEM NOT FOUND	28	28	false	true
4	0a7214ada1000561	1	[null]	4011	ITEM NOT FOUND	1	1	[null]	[null]
5	0a7214ada1000566	1	5332f5f4b03c9a25efd0ba	4011	ITEM NOT FOUND	3.25	3.25	false	true
6	0a7214ada100056e	1	[null]	4011	ITEM NOT FOUND	20	20	false	true
7	0a7214ada100056b	1	[null]	4011	ITEM NOT FOUND	20	20	false	true
8	0a720f052300056c	1	[null]	4011	ITEM NOT FOUND	1	1	[null]	[null]
9	0a7214ada100055a	1	[null]	4011	ITEM NOT FOUND	1	1	[null]	[null]
10	0a7214ada1000564	1	[null]	4011	ITEM NOT FOUND	1	1	false	[null]

How to handle around 170 records for ITEM NOT FOUND. Is it expected?