

Cleaning Compendium

What was wrong in each file, the rule we applied, and exactly which rows changed.
Deterministic rules only: no estimates, no imputation, no simulation.

Version 3.0 • 22 October 2026 • Input: the raw files described in document 1, plus Binzagr's mapping files of 22 October
Output: folder **02_Cleaned** - 19 tables (26 parquet files, each with a CSV copy), DATA_DICTIONARY.csv, SMOKE_TEST_RESULTS.md, audit files and the scripts that produced them.
New in version 3: a monthly head count straight from Binzagr's user list (counts only, no ids or names) and smoke test M09.
Version 2 applied Binzagr's transaction mapping, WMS user list, conversion factors and pallet-id sample (section 3.7).

4,604,674	4,539,533	65,141	47 / 49
WMS rows received	WMS rows after cleaning	rows removed: proven duplicates only	smoke tests pass (2 warnings, explained)

1. Rules we followed

- **Deterministic only.** Every change is a proven fix of an export defect, a type conversion, or a flag. Nothing is estimated, imputed or smoothed here; those steps live in 03_Simulated and are labelled there.
- **Remove only what is proven redundant.** Transaction rows removed: only the 65,141 WMS rows proven, at raw-text level, to be second copies. In the master lists, 4 duplicate payment-term rows collapse into their codes and 7 damaged product rows are replaced by the 155 records they contained.
- **Null only provable placeholders.** 01/01/1900 dates and quantities that are really pallet ids are set to null, always with a flag.
- **Flag, don't fix.** Suspicious but possible values (negative quantities, identical rows, very long tasks, zero-value sales) are kept and flagged.
- **Lineage on every row.** Each cleaned row carries src_file + src_line (CSV) or src_row (Excel), so any value can be traced back to the raw file.
- **Data minimisation.** Personal and identifying columns (contacts, tax and bank ids) are dropped from the customer table; employee names from the WMS user list are never stored (only the payroll number, the key for attendance). The raw files keep them.
- **Client mappings as received.** Binzagr's own mapping files are attached exactly as sent; where they contradict themselves both values are kept and flagged, never chosen silently.

Contents

1. Rules we followed	1
2. Summary by dataset	3
3. Jeddah WMS transaction log	4
3.1 Month-boundary duplicates (W1)	4
3.2 Placeholder dates by month (W2, W4)	5
3.3 Client code for billing events (W10)	5
3.4 The 12 quantities that were pallet ids (W3)	5
3.5 Negative quantities by transaction type (W7, kept)	5
3.6 Columns and tables derived from the log (no estimates)	6
3.7 Binzagr's mappings, received 22 October (W14-W17)	6
4. Product master and hierarchy	7
4.1 Example of rebuilt records	7
4.2 Hierarchy table	7
5. Customer master	8
5.1 Columns dropped	8
6. Riyadh sales and returns	8
7. Small master tables	9
8. Drinks and Lipton subset	9
9. Smoke tests	10
10. How to trace any value back to the raw file	11

2. Summary by dataset

Dataset	Raw rows	Clean rows	Rows removed or replaced	Values set to null / filled	Main flags (rows)	Clean table
Jeddah WMS log (19 -> 59 columns)	4,604,674	4,539,533	65,141 duplicates removed	end time 37,694; qty 12; production 185,054; expiry 186,579 nulled. Client filled on 280,844 billing rows	pallet id lossy 2,860,057 (58 recovered from the sample); identical 58,012; negative qty 5,658; expiry = production 12,043	jeddah_wms_transactions (8 monthly files)
Riyadh sales lines (11 -> 26 columns)	325,544	325,544	none	none	identical 176; zero-value sellable 12; document reused 39	riyadh_sales_lines
Riyadh return lines (11 -> 26 columns)	7,170	7,170	none	none	FOC with amount 8; zero-value sellable 2	riyadh_return_lines
Customer master (119 -> 43 columns)	18,809	18,983	none; 174 subset-only customers added	none	credit-limit placeholder 10,755; negative balance 184; no coordinates 751	dim_customer
Product master (67 -> 60 columns)	23,497	23,645	7 damaged rows replaced by 155 rebuilt records	exact category key rebuilt for 23,496	category name ambiguous: L5 10,035, L6 18,206	dim_product
Product hierarchy (11 -> 15 columns)	4,285	4,285	none	none	codes shared after truncation 1,399	dim_product_hierarchy
UOM conversion (3 -> 4 columns)	45,891	45,891	none	none	-	dim_uom_conversion
Salesman master (54 -> 51 columns)	119	119	none	none	-	dim_salesman
Payment terms (2 -> 8 columns)	28	24	4 duplicate rows collapsed	none	ZC12 unparsed	dim_payment_terms
Binzagr transaction mapping (22 Oct) (7 -> 14 columns)	88	88	none	none	code 080 listed twice; code 100 description differs from the log	ref_wms_tran_mapping
Binzagr WMS user list (22 Oct) (9 -> 12 columns)	32,920	1,936	kept only the 242 Jeddah badges x 8 months	names removed	badge re-used for another payroll number: 1; scanned while inactive: 3 badges	ref_wms_user_month; ref_wms_user_headcount_month (counts)
Binzagr conversion factors (22 Oct) (3 -> 18 columns)	36,859	36,851	6 identical repeats collapsed	factors of 1 conflicting item set to null	no pallet factor 1,155; no case factor 163; pallet < case 154; 99999 placeholder 7	dim_item_conversion
(new) calendar, badges, people, sessions, transaction types	-	365 / 387 / 338 / 24,409 + 23,946 / 69	-	-	session > 12 h: 513	6 new tables (sections 3.6, 3.7 and 7)

Totals were preserved wherever rows were kept: sales and returns rows, net amounts and quantities equal the raw files exactly (SAR 40,409,418.71 and SAR 1,678,644.91), and WMS quantity totals per transaction type equal the raw totals for all 69 types (smoke tests C01, C02, J10).

3. Jeddah WMS transaction log

8 monthly CSV files, 4,604,674 rows, 19 columns. All start and end timestamps parsed (0 failures). The ledger lists every problem found, the rule applied and the rows affected; the tables after it show the detail.

#	Problem found	Rule applied	Rows	Where to see the rows
W1	Each monthly file also contains the whole first day of the next month.	For each boundary day the earlier copy was compared with the next file (multiset of whole rows, then raw text). The earlier copy is always a strict subset, so it was dropped.	65,141 removed	audit/jeddah_removed_boundary_rows.csv.gz lists every src_file + src_line
W2	End time 01/01/1900 12:00:00 AM (placeholder) on RoS Shipment Audit rows, which also caused negative durations.	end_ts = null, flag_end_sentinel = true. No negative duration remains.	37,694 (39,027 before W1)	flag_end_sentinel; e.g. February file line 690504
W3	TRAN_QTY holds a pallet id (2.10004E+11) on LP Audit rows.	qty = null, flag_qty_corrupt = true.	12	flag_qty_corrupt; all 12 lines in table 3.4
W4	Production / expiry date 01/01/1900 (placeholder) on pallet-level events.	Date = null, flag set.	185,054 / 186,579	flag_production_sentinel / flag_expiration_sentinel
W5	Pallet ids (HU_ID) in scientific notation: last digits destroyed by Excel. Binzagr's 22 Oct sample proves the loss: 12-digit ids kept to their first 6 digits (the site code).	Kept as text and flagged (hu_id_status); the 58 sample rows get their real id; the rest cannot be repaired (re-export requested). HU_ID_2 is not carried: Binzagr confirmed it is unused.	2,860,057 (58 recovered)	hu_id_status, hu_id
W6	Rows identical in all 19 columns inside one file (pallet-level events whose pallet ids were destroyed).	Kept: they are different pallets. Group size recorded.	58,012	n_identical_rows_in_file > 1
W7	Negative quantities.	Kept and flagged: adjustments, scrap and count corrections are negative by nature.	5,658	flag_qty_negative (table 3.5)
W8	Expiry before / equal to production date.	Flagged only.	1 / 12,043	flag_exp_before_prod / flag_exp_equals_prod
W9	Tasks left open for hours (very long durations).	Kept; filter with dur_s when modelling task time.	3,585 over 1 h (39 over 8 h, max 185 h)	dur_s
W10	Billing events exported without a client code.	Client filled only when exactly one client appears in the same-second scan set of the same badge and site (or, failing that, among scans of the same pallet). Otherwise left null and labelled.	280,844 filled; 19,900 left null	client_code_source (table 3.3)
W11	Text "NULL" and blank strings.	Converted to real nulls in every column.	all columns	-
W12	File names misleading ("Feburary"; each file holds a day of the next month).	Month taken from the timestamp, never from the file name.	all	month
W13	Badge id case variants.	Upper-cased copy for timelines; original kept.	0 variants found	employee_id_norm
W14	Which activities are billed was unknown.	Binzagr's transaction mapping attached to every row: case / pallet, type of movement, cost / revenue.	all 69 types	movement_type, cost_or_revenue
W15	Badges are not people (45 people use 2+ badges).	Payroll number and WMS role from Binzagr's user list, month by month; badge sessions of one person merged into person sessions (6 h rule).	93.9% of rows mapped	person_id, wms_role, person_session_id
W16	Quantities are in mixed units of sale.	Units per case / pallet from Binzagr's conversion file; qty itself unchanged.	99.6% of item rows	qty_cases, qty_pallets
W17	End dates 01/01/1900 (W2): could they be "standardised to 2026"?	No: all are RoS Shipment Audit with time 00:00 - no end was ever recorded. Left null (test M08).	37,694	flag_end_sentinel

3.1 Month-boundary duplicates (W1)

Boundary day	Rows in earlier file	Rows in own file	Earlier rows not found in own file	Removed from
2026-03-01	7,854	8,585	0	Jeddah Feburary 2026.csv
2026-04-01	10,403	11,371	0	Jeddah March 2026.csv
2026-05-01	936	991	0	Jeddah April 2026.csv

Boundary day	Rows in earlier file	Rows in own file	Earlier rows not found in own file	Removed from
2026-06-01	9,808	10,278	0	Jeddah May 2026.csv
2026-07-01	10,570	11,069	0	Jeddah June 2026.csv
2026-08-01	12,002	12,306	0	Jeddah July 2026.csv
2026-09-01	13,568	14,269	0	Jeddah August 2026.csv
Total	65,141		0	65,141 rows

The September file has no overlap because there is no October file. Smoke tests J03 and J04 re-checked every removed row against the raw text of the next file.

3.2 Placeholder dates by month (W2, W4)

Month	End time nulled	Production date nulled	Expiry date nulled
2026-02	5,069	23,572	23,581
2026-03	5,090	16,077	16,111
2026-04	5,642	17,861	18,094
2026-05	4,139	16,956	16,974
2026-06	5,172	13,501	13,597
2026-07	4,318	34,720	35,305
2026-08	3,906	30,169	30,408
2026-09	4,358	32,198	32,509
Total	37,694	185,054	186,579

End-time placeholders occur only on RoS Shipment Audit. Production / expiry placeholders are mostly on billing, LP audit and Close Load events, which do not refer to one lot.

3.3 Client code for billing events (W10)

Billing event	Rows	Unique client in same-second scans	Same pallet	Coverage	Ambiguous	No match
Loading Per Pallet (Billing)	151,799	140,635	1,444	93.6%	9,604	116
LP Audit Per Pallet (Billing)	85,001	81,557	584	96.6%	2,821	39
Consolidation Per Pallet (Billing)	63,944	55,878	746	88.6%	7,268	52

Re-verified independently on a 2,000-row sample (smoke test J12b). The order-number prefix was not used: it matches the client only 48% of the time because it identifies the order-owning sales org, not the stock owner.

3.4 The 12 quantities that were pallet ids (W3)

Raw file and line	Raw file and line	Raw file and line
Jeddah July 2026.csv line 514304	Jeddah July 2026.csv line 466809	Jeddah July 2026.csv line 509309
Jeddah July 2026.csv line 517682	Jeddah August 2026.csv line 552526	Jeddah August 2026.csv line 493962
Jeddah August 2026.csv line 495143	Jeddah August 2026.csv line 502269	Jeddah August 2026.csv line 461607
Jeddah August 2026.csv line 505086	Jeddah September 2026.csv line 407920	Jeddah September 2026.csv line 480485

3.5 Negative quantities by transaction type (W7, kept)

Transaction type	Raw rows with negative qty
Inventory Adjust by Location	2,852
Cycle Count Approval	1,294
Adjust Receipt of Shipment	1,237
Scrap	227
Underload Adjustment	16
Cycle Count Check	14
Underload Issue	13
Close Load	4
Picking (Put)	1

Counts before removing the boundary duplicates; after cleaning 5,658 rows keep the flag.

3.6 Columns and tables derived from the log (no estimates)

Derived item	Definition
start_ts, end_ts, dur_s	Date + 12-hour time parsed into timestamps; duration = end - start in seconds (null when end is a placeholder).
tran_category, tran_leg, is_billing_event	Activity family and leg (pick / put / receipt / billing / single) from the description text.
client_code_source	source, linked_same_second_unique, linked_same_second_same_pallet, unresolved_ambiguous, unresolved_no_match.
employee_id_norm, badge_type, badge_suffix	Upper-cased id; standard (letters + 6 digits), nonstandard, or system (SAPImport, Admin, ShortPick); trailing 6 digits of standard ids (home site).
control_prefix; item_client, item_code, item_kind	First 4 digits of the order number; client and item parts of ITEM_NUMBER; sku or pallet asset.
secs_since_prev_event, idle_secs_before	Time since the same badge's previous scan start / end (negative idle = overlapping scans).
session_id	A new work session starts after more than 6 h without a scan by the badge (trough of the gap histogram). Sensitivity: 4 h -> 26,298 sessions, 6 h -> 24,409, 8 h -> 23,889.
flag_fast_wh_switch	Badge appears at another site less than 2 minutes after its previous scan: 728 rows.
jeddah_badge_sessions	24,409 sessions (513 longer than 12 h, 209 longer than 16 h): start, end, scans, sites, span.
jeddah_badges	387 badges: 236 standard, 148 nonstandard, 3 system. 38 pairs that differ only by a "WA.AD." prefix are flagged (possible_alias_group), never merged.
dim_wms_transaction_type	69 types with family, leg, billing flag, row count, zero-duration share, median duration and Binzagr's classification.
flag_same_second_as_prev	Scan logged in the same second as the badge's previous scan (batch posting).

3.7 Binzagr's mappings, received 22 October (W14-W17)

Badr sent four files after the 21 October data walkthrough. They are attached as received; contradictions inside them are flagged for Binzagr, not resolved by us.

File	What it gives	Coverage in the Jeddah log	Issues flagged for Binzagr
Trans_Log_Mapping.xlsx (88 rows)	Case / pallet, grouping 1 and 3, type of movement (Basic Service / Added Value / In House / Duplicate), cost / revenue	all 69 codes; revenue codes in the log: 100, 153, 172, 183, 204, 206, 350, 391, 400, 410	code 080 listed twice (In House / Duplicate); code 100 "Picked Volume" vs log "Item Import"; 18 codes never in the log (110-920: billing volumes?)
Users_Mapping.xlsx (32,920 rows, monthly snapshots)	Badge -> payroll number, WMS menu level (role), status; names dropped	242 of 387 badges = 93.9% of scans; 338 people (45 with 2+ badges)	145 logins not in the list (WA.AD.nnn, plain numbers, HM1008 ...); badge MAA210007 re-used; 3 badges scanned while inactive
Conversion_Factor.xlsx (36,859 rows)	Units of sale per pallet (PAL) and per case (CS)	99.6% of item rows; matched on client-item (419,100 rows) or item code (3,682,839)	1,155 items without PAL, 163 without CS, 154 with PAL < CS, 1 item with two PAL values
Sample_of_HU_ID.xlsx (100 rows)	The same WMS columns with the pallet id kept as text (cycle counts, Sep 2026)	all 100 rows matched a cleaned row exactly; 58 had been destroyed in our export	proves the loss: e.g. 210004132239 exported as 2.10004E+11

Person sessions	Value
Badge sessions -> person sessions	24,409 -> 23,946
Person sessions using 2+ badges	445
... with two badges active at the same time	239 (258 h) - badge-sharing signal
Rows by Binzagr type of movement	Duplicate 2,131,861, Basic Service 1,421,852, Added Value 822,299, In House 159,325, Duplicate In House 4,196

Head count straight from the list (new in version 3). ref_wms_user_headcount_month counts the people (distinct payroll numbers) with an active login on the snapshot of the 1st of each month, by WMS menu level, for the 9 sites in the log and for all Jeddah sites. Counts only: no ids or names. Smoke test M09 re-derives every count from the raw snapshots. These are people registered in the WMS, not people on shift: fewer scan in a month (document 3, section 3.4b).

Month	Active people, 9 log sites	Tally clerk menu	Storekeeper menu	Active people, all Jeddah sites
2026-02	274	201	83	290
2026-03	272	198	87	288
2026-04	275	201	87	292

Month	Active people, 9 log sites	Tally clerk menu	Storekeeper menu	Active people, all Jeddah sites
2026-05	276	201	88	293
2026-06	261	190	84	278
2026-07	263	190	87	280
2026-08	263	189	86	280
2026-09	265	187	91	282

A person can hold logins at several menu levels, so the menu columns do not add up to the total.

4. Product master and hierarchy

#	Problem found	Rule applied	Rows	Where to see them
P1	A stray quote character made the export glue product records into single cells: 7 Excel rows (224, 513, 525, 654, 669, 1184, 1185) contain 155 products, and their text spilled into other columns.	Each damaged cell was split back into records using the tab structure (a new record starts at a line beginning "HQ" + tab); quote characters stripped. Every rebuilt code and description was found again inside the raw cell (test P04).	7 rows -> 155 records	record_source = rebuilt_from_merged_cell; src_row; rebuilt_seq
P2	PRDCAT_KEY (18 digits) stored as a number, so Excel kept 15 significant digits (last 3 lost).	Exact key = first 10 stored digits + MATGRP_CD (8 digits). The rule holds for 100% of keys and is confirmed directly by the 149 keys that survived as text (tests P03a/P03b).	23,496 rebuilt; 149 exact as stored	prdcatt_key_exact, prdcatt_key_method
P3	Hierarchy codes of levels 5-6 truncated the same way, so different categories share one stored code.	Codes are exact on the product side (from the key). Names attached only when the truncated code is unique; otherwise left empty and labelled, never guessed.	L5 names ambiguous 10,035; L6 18,206	l5_desc_status, l6_desc_status
P4	25 columns empty in every valid row (e.g. SHELF_LIFE, BRAND_HIER1-3, HQ_DEFINE1-5).	Dropped from the clean table (raw keeps them).	25 columns	stats_products.json
P5	12 legacy alphanumeric product codes.	Kept as text; prd_cd_is_numeric = false.	12 products	prd_cd_is_numeric

Row accounting: 23,497 raw rows - 7 damaged rows + 155 rebuilt records = **23,645 products**, exactly the product list inside the sales file, with identical codes (test P01). Before the rebuild 0.3% of sales lines pointed to a product missing from the master; after it, none.

4.1 Example of rebuilt records

Product code	Raw Excel row	Position inside the damaged cell
201217	224	1
201218	224	2
201219	224	3
201220	224	4
201221	224	5

4.2 Hierarchy table

Level	Nodes	Codes exact?	Rows sharing a truncated code
1	158	yes	0
2	519	yes	0
3	587	yes	0
4	618	yes	0
5	916	no	333
6	1,487	no	1,066

Dropped empty columns: RSC_PRD, SHELF_LIFE, BASEPRD_CD, FGRP_CD, DISTBASE_CD, BRAND_HIER1, BRAND_HIER2, BRAND_HIER3, PTAXGRP_CD, STAXGRP_CD, PL_UOM_CD, LY_UOM_CD, HQ_DEFINE1, HQ_DEFINE2, HQ_DEFINE3, HQ_DEFINE4, HQ_DEFINE5, BE_IND, BUSINESS_GRP, POSM_CAT, CRATE_IND, MY_TAX_CD, MY_TAX_CD_IN, SEQ_NO, CHECK_LICENSE.

5. Customer master

#	Problem found	Rule applied	Rows / columns
C1	119 columns, most of them unused template fields or personal data.	Dropped 82 columns: 51 empty, 23 with one constant value, 15 personal / identifying (7 of these are also empty).	119 -> 37 source columns
C2	174 customers appear only in the Drinks and Lipton subset.	Added with record_source = lipton_subset_customer_master_only.	+174 rows
C3	Credit limit 9,999,999 = "no limit" placeholder.	Kept; flag_credit_limit_sentinel.	10,755
C4	Negative outstanding balance (customer in credit).	Kept; flag_outstanding_negative.	184
C5	Missing coordinates.	geo_status = missing (all others inside Saudi Arabia).	751
C6	Credit / cash customer type.	has_credit_terms derived: every Q customer has a payment term, no C customer does (100%).	11,068 with terms
C7	4 customer codes appear under two distributor codes.	Kept; the key is DIST_CD + CUST_CD (unique, test C05).	4 codes

5.1 Columns dropped

Reason	Columns
Empty in every row (51)	DIST_PRICEGRP_CD, BILL_METHOD, SUBDIST_IND, KEYACCREG_CD, ADD_CONT_NO, CUST_TAXNO, CUST_MUNICIPALNO, REMARKS, BANK_CD, BANK_BRANCH, BANK_ACCNO, OPENING_BAL, OPENING_BAL_DT, CUSTCLASS_UPLFLAG, CUSTCLASS_DWLFLAG, CUSTCLASS_UPDDT, USER_DEFINE2, USER_DEFINE3, USER_DEFINE4, USER_DEFINE5, HQ_DEFINE1, HQ_DEFINE2, HQ_DEFINE3, HQ_DEFINE4, HQ_DEFINE5, CUST_BARCODE, SEASON_START_DT, SEASON_END_DT, NEW_CUST_CD, UPLDW_STATUS, SVR_UPLDW_DT, CUST_GRP_DISC_CD, CUST_ACC, PREVIOUS_CUST_HIER3, NO_OF_SKU, CHECK_LICENSE, ATTR3_CD, ATTR4_CD, PAN_NO, LINK_CODE, ADDR_1_CD, ADDR_2_CD, ADDR_3_CD, ADDR_4_CD, ADDR_5_CD, ADDR_6_CD, ADDR_POSTAL_CD, STATE_CD, SUBDIST_SUBDIST_CD, TERTIARY_SUBDIST_CD, IMAGE_REC_IND
One constant value (23)	COMMUNICATION, RELATIONSHIP, CUST_TAX, CUSTCLASS_STATUS, DKA_IND, INV_COUNT, CUST_SEASON, SYNCOPERATION, SYNCFLAG, VISION_STORE, SH_REPLN, DIGITAL_IND, ADDR_6, RSS_IND, SMS_NOTIFICATION, EMAIL_NOTIFICATION, GST_SCHEME, TERTIARY_IND, PROMO_IND, VMI_TRANSIT_STK, PRINT_TERTIARY_ORDER_IND, SPECIAL_PRC, OUTLET_LOC_IND
Personal / identifying (15)	CONT_PR, CONT_NO, CONT_NO_EXT, ADD_CONT_NO, MOBILE_NO, CONT_FAXNO, CONT_EMAIL, CUST_REGNO, CUST_TAXNO, CUST_TAXREGNO, CUST_MUNICIPALNO, BANK_CD, BANK_BRANCH, BANK_ACCNO, PAN_NO

The cleaned customer table keeps 37 source columns plus src_row, record_source, has_credit_terms, the two flags and geo_status (43 columns).

6. Riyadh sales and returns

#	Item	Rule applied	Sales rows	Return rows
S1	Rows, amounts, quantities	No row removed or changed; totals identical to raw (tests C01, C02).	325,544	7,170
S2	PRD_DESC is an Excel formula	Its cached text result is stored as prd_desc.	all	all
S3	No line number in the source	doc_line_seq = position of the line within its document in the raw sheet.	all	all
S4	Quantities in mixed units	qty_base_ea = qty x CONV_FACTOR from the UOM sheet (test C03).	all	all
S5	Price per each	unit_net_price_per_ea = net_amt / qty_base_ea for paid lines (null on FOC).	305,863	6,834
S6	Sellable line with 0 amount	Flagged (flag_sellable_zero_amount).	12	2
S7	FOC line with an amount	Flagged (flag_foc_nonzero_amount).	0	8
S8	Identical lines in the sheet	Kept (no line id exists to tell them apart); group size recorded.	176	0
S9	Document number reused for another customer and date	Flagged (flag_doc_multi_customer, flag_doc_multi_date): 11 invoices.	39	0
S10	Joins to customer and salesman masters	customer_in_master, salesman_in_master (test C04).	100%	100%

Example rows (Excel row numbers): zero-value sellable sales lines 23891, 23925, 24016, 24575, 24617; reused invoice lines from row 105962; FOC returns with an amount 3070, 946, 2847, 2855, 2858.

7. Small master tables

Table	What was done	Result
UOM conversion	Codes typed as text (12 legacy alphanumeric codes kept); lineage added.	45,891 rows
Salesman master	5 empty columns dropped (BU_CD, USER_CD, SVR_UPLDW_DT, TEAM_CD, CHILD); is_active = SLSMAN_STATUS 1.	119 reps, 70 active
Payment terms	28 rows -> 24 codes: ZC09, ZC10, ZC18 and ZC23 were listed twice with whitespace differences. Each description parsed into net_days, anchor (end of month / invoice date / immediate) and early-payment discount; usage counted.	24 codes; ZC12 unparsed (no description)
Calendar (new)	Every day of 2026 with weekday, ISO week, Friday / Saturday flags and the Hijri (Umm al-Qura) date. Anchors checked: 1 Ramadan 1447 = 18 Feb, Eid al-Fitr = 20 Mar, Eid al-Adha = 27 May (test C08).	365 days

Parsed payment terms

Code	Description	Net days	Anchor	Discount terms	Customers
ZC01	Cash Cheque	0	immediate		0
ZC02	15 Days from End of Month	15	end_of_month		158
ZC03	20 Days from End of Month	20	end_of_month		0
ZC04	30 Days from End of Month	30	end_of_month		10,459
ZC05	45 Days from End of Month	45	end_of_month		0
ZC06	60 Days from End of Month	60	end_of_month		446
ZC07	4 days 2% cash disc 15days net from month end	15	end_of_month	2% if paid within 4 days	0
ZC08	30 days 2% disc 60 days net from month end	60	end_of_month	2% if paid within 30 days	0
ZC09	15 Days from Invoice / DN date	15	invoice_date		1
ZC10	4 days 1% disc 60 days net from month end	60	end_of_month	1% if paid within 4 days	0
ZC11	3 Days from Invoice date	3	invoice_date		0
ZC12	ZC12	-			0
ZC13	30 Days from Invoice date	30	invoice_date		4
ZC14	90 Days from End of Month	90	end_of_month		0
ZC15	20 Days from Invoice date	20	invoice_date		0
ZC16	15D 3% 30D 2% 60D 1% Over60net		unspecified	3% <=15d, 2% <=30d, 1% <=60d, net after 60d	0
ZC17	75 Days from End of Month	75	end_of_month		0
ZC18	30 Days from Invoice / DN date	30	invoice_date		0
ZC19	66 Days from Month End	66	end_of_month		0
ZC20	2 Days from Invoice date	2	invoice_date		0
ZC21	60 Days from Invoice date	60	invoice_date		0
ZC22	180 Days from Invoice date	180	invoice_date		0
ZC23	45 Days from Invoice date	45	invoice_date		0
ZC24	15D 3% 30D 2% 60D 1% Over60net		unspecified	3% <=15d, 2% <=30d, 1% <=60d, net after 60d	0

8. Drinks and Lipton subset

The first, smaller send was compared row by row with the main files (test C06): all 80,960 sales lines, 981 return lines and 2,812 unit conversions are contained in the main files, all its salesmen and (after the rebuild) all its products exist in the main masters. It was therefore not appended, which would have double-counted. The only new information, 174 customers, was added to the customer table.

9. Smoke tests

Each test re-derives a result from the raw files (raw CSV text, raw Excel cells, Binzagr's mapping files) instead of re-using the pipeline code, in a separate process per stage. Result: 47 pass, 0 fail, 2 warnings.

ID	Test	Result	Detail
J00	Raw CSV lines all have exactly 19 comma-separated fields (no quoting)	PASS	lines with other widths: 0
J01	Raw rows = clean rows + removed boundary rows (total and per file)	PASS	raw 4,604,674 = clean 4,539,533 + removed 65,141; per-file problems
J02	Clean + removed lineage covers every raw line exactly once	PASS	overlap 0, raw lines with no lineage 0, duplicate clean lineage 0
J03	Every removed row is dated day 1 of the next month	PASS	
J04	Every removed row exists (as identical raw text) in the next month file	PASS	removed lines not found in next file: 0
J05	Every clean row belongs to its file month	PASS	
J06	Field-level spot check: 3,000 random clean rows vs raw CSV text	PASS	mismatches: 0
J07	end_ts is null exactly where the 01/01/1900 placeholder was found	PASS	sentinel rows: 37,694
J08	No negative durations remain	PASS	min dur_s = 0.0
J09	qty is null exactly where the raw qty was an Excel-corrupted ID	PASS	corrupt qty rows: 12
J10	Quantity totals per transaction type equal raw totals	PASS	types checked 69; differences:
J11	Item-number prefix equals client code on every source row	PASS	rows checked 4,101,952
J12a	Linked client codes only appear on billing rows	PASS	linked rows 280,844
J12b	Linked client codes re-verified independently (2,000 sample)	PASS	failures: 0
J13a	Sessions table matches transaction session ids	PASS	sessions 24,409
J13b	Each session = one badge; gaps between sessions > 6h	WARN	session pairs with gap <= 6h: 1
J14	Badge table covers every badge and every event	PASS	badges 387
J15	Transaction-type dictionary covers every type and row	PASS	types 69
J16	Mandatory columns have no nulls	PASS	none
M01	Every transaction type in the cleaned log is in Binzagr's mapping file	PASS	types 69; missing
M02	Movement type and cost / revenue on every row equal the mapping file	PASS	types with differences:
M03	Mapping file consistent with itself and with the log descriptions	WARN	code 080 listed 2 times (Duplicate In House); code 100 log 'item import' vs mapping 'Picked Volume'
M04	Case / pallet factors and qty in cases re-derived from the raw conversion file (20,000 sample); case factor on >= 99% of item rows	PASS	factors match True; qty_cases match True; rows with a case factor 99.63%
M05	Badge -> payroll links exist in Binzagr's raw user list; unmapped logins are truly absent; no WMS user names stored	PASS	links checked 242; bad; unmapped present in list; name leaks
M09	User-list head count (active people per month at the log's sites) re-derived from the raw snapshots; table holds counts only	PASS	active people [274, 272, 275, 276, 261, 263, 263, 265]; differences
M06	Each badge session sits in exactly one person session, one person per person session, events reconcile, person sessions > 6h apart	PASS	person sessions 23,946 from 24,409 badge sessions; min gap 6.0 h
M07	Pallet-id sample: every matched row exported as the 12-digit id rounded to 6 significant digits (digits 7-12 lost); recovered rows carry the text id	PASS	sample rows 100; recovered 58; destroyed in the export 2,860,057 rows; e.g. 2.1E+12 <- 2100004013429
M08	1900 end dates are all RoS Shipment Audit (914) with time 00:00 - no end time was ever recorded (nothing to standardise)	PASS	raw rows 39,027; types ['914']; Binzagr movement type Duplicate
P01	Product count equals the independent product list in the Sales file (23,645) and codes match	PASS	dim 23,645 vs sales-file list 23,645; codes only in dim 0, only in list 0
P02	prd_cd unique	PASS	
P03a	Exact key is 18 digits and digits 11-18 equal MATGRP_CD	PASS	products with key 23,645 / 23,645
P03b	Exact key truncates back to the value Excel stored (independent re-read)	PASS	checked 23,490; bad 0
P04	Every rebuilt product (code + description) appears in the raw merged cell text it came from	PASS	rebuilt 155, verified 155
P05	Row accounting: raw rows - damaged rows + rebuilt records = final	PASS	23497 - 7 + 155 = 23645
P06	Hierarchy levels 1-4 codes all exist in the hierarchy table with descriptions	PASS	
P07	Every product code used in sales/returns exists in dim_product	PASS	missing 0
P08	Spot check 1,000 products vs raw Excel (code, status, description)	PASS	mismatches 0

ID	Test	Result	Detail
C01	riyadh_sales_lines: rows, net amount and qty equal raw	PASS	rows 325,544/325,544; NET_AMT 40,409,418.71/40,409,418.71; qty 1,461,861/1,461,861
C01b	riyadh_sales_lines: 1,000-row spot check vs raw Excel	PASS	mismatches 0
C02	riyadh_return_lines: rows, net amount and qty equal raw	PASS	rows 7,170/7,170; NET_AMT 1,678,644.91/1,678,644.91; qty 47,257/47,257
C02b	riyadh_return_lines: 1,000-row spot check vs raw Excel	PASS	mismatches 0
C03	qty_base_ea = qty x conversion factor from the UOM table	PASS	
C04	Every sales/return line joins to a customer and a salesman	PASS	missing customers 0, salesmen 0
C05	Customers unique on (dist_cd, cust_cd), 18,809 + 174 rows, no contact/bank/tax-ID columns	PASS	rows 18,983
C06	Lipton files proven to be subsets (sales, returns, UOM) - not duplicated	PASS	
C07	Payment terms: 24 unique codes and every code used by customers is defined	PASS	
C08	Hijri calendar anchors (1 Ramadan, 1 Shawwal, 10 Dhul-Hijjah 1447, 1 Muharram 1448, Umm al-Qura)	PASS	none
F01	Every parquet table has a CSV twin with the same row count	PASS	tables 26; mismatches
F02	No file larger than 29 MB (transfer limit)	PASS	none

The two warnings. J13b is a single LP Audit task opened at 14:40 and closed at 09:26 the next day (18.8 h): its end overlaps the badge's next shift; sessions are split on gaps between scans, so the two shifts correctly stay separate. M03 lists two inconsistencies inside Binzagr's mapping file that only Binzagr can resolve (code 080 listed twice; code 100 described differently from the log).

10. How to trace any value back to the raw file

- **WMS rows:** src_file + src_line give the physical line in the raw CSV (header = line 1). Removed duplicates are listed with the same two fields in audit/jeddah_removed_boundary_rows.csv.gz.
- **Excel rows:** src_row is the Excel row number (header = row 1). Rebuilt products also carry rebuilt_seq (position inside the damaged cell).
- **Flags** keep the reason for every changed or suspicious value next to it; nothing is changed silently.
- **Scripts** (02_Cleaned/scripts): clean_mappings.py, clean_jeddah.py, clean_products.py, clean_commercial.py, smoke_tests.py, data_dictionary.py, build_cleaning_report.py. Re-running them on the raw files reproduces the folder.