WESTMINSTER UNIVERSITY

Data Capstone Pipeline Report

SAP Backorder & WOC: Data, ETL, and Next Steps

Purpose: Documents the data pipeline, master tables, and BRD metrics built for the capstone project.

Scope: data/ (clean + processed), src/data/ (ETL + BRD), notebooks/ (EDA next)

Status: Pipeline operational: 6 master tables produced

Overview: Data & Pipeline Structure

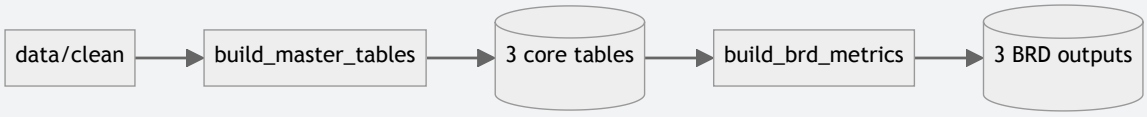
What I Built: A two-phase pipeline that turns SAP ERP clean CSVs into master tables and BRD-aligned metrics (backorders, weeks of coverage). Aligned with EnableCV BRD logic.

Directory Structure:

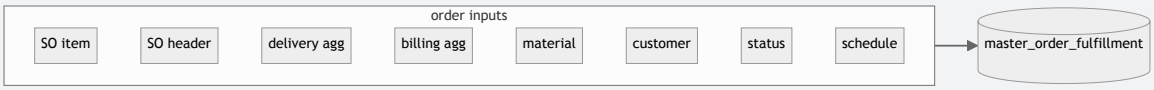
- `data/clean/main/` : Transactional tables (orders, delivery, billing, inventory, PO)
- `data/clean/supporting/` : Reference tables (plant, company_code, sales_org); separated so I join only when needed
- `data/processed/` : Master tables + BRD outputs (6 CSV files)
- `src/data/` : build_master_tables.py, build_brd_metrics.py, run_pipeline.py
- `notebooks/` : EDA and modeling (01_eda_targets.ipynb next)

How Processed Tables Were Built

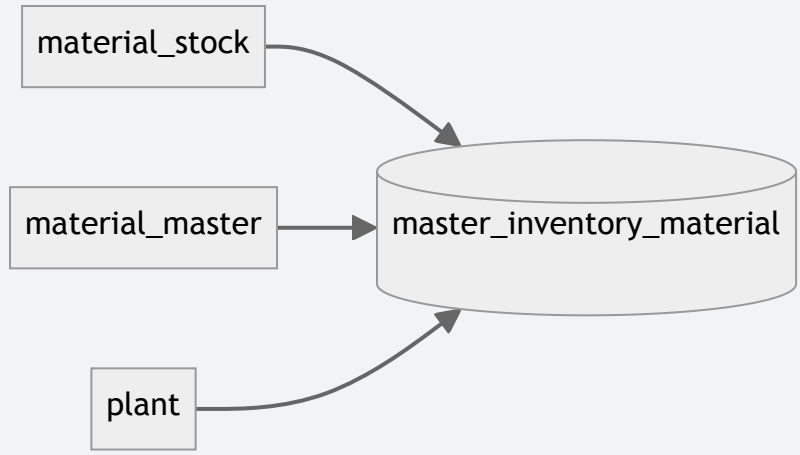
High-level flow:



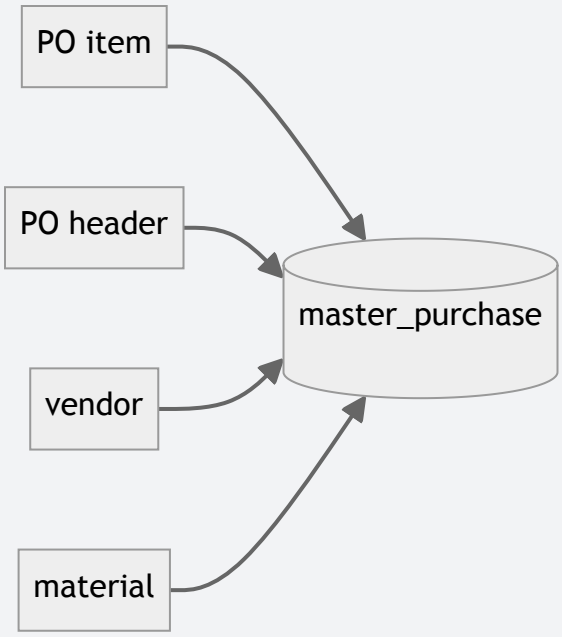
Flow 1: master_order_fulfillment



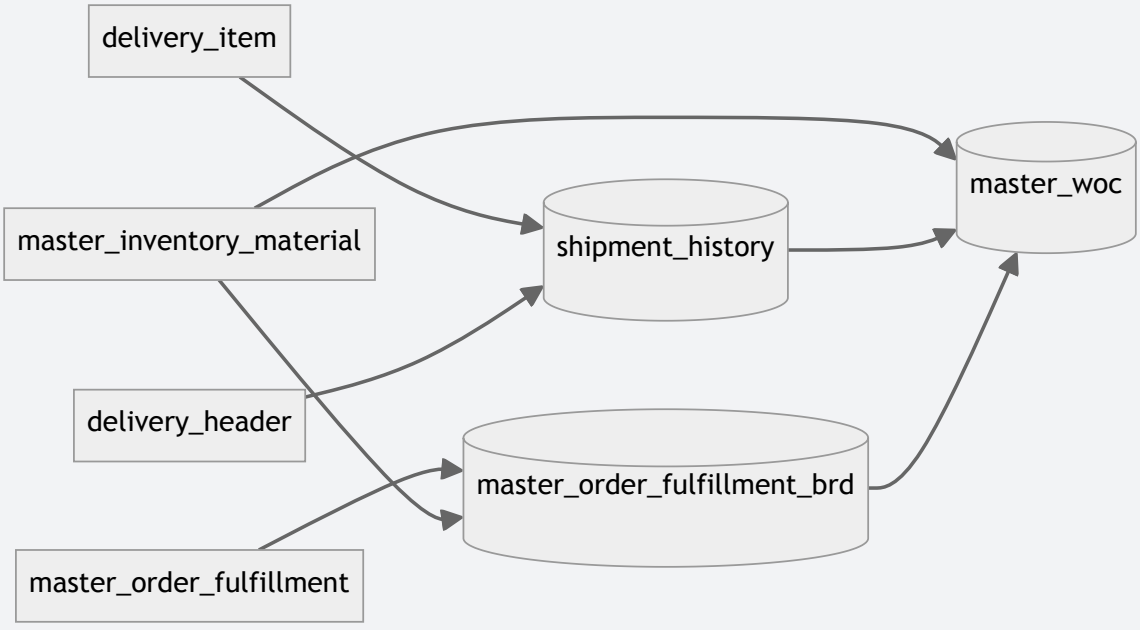
Flow 2: master_inventory_material



Flow 3: master_purchase



Flow 4: BRD metrics



Pipeline Output Tables

Table	Rows	Description
`master_order_fulfillment`	~52k	Order-to-cash: SO + delivery + billing + material + customer
`master_order_fulfillment_brd`	~52k	Same + outstanding_qty, saleable_inventory, backorder_units/amount, aging
`master_inventory_material`	~66k	Inventory by material/plant: unrestricted_stock, blocked, etc.
`master_purchase`	~862k	Purchase orders: PO + vendor + material
`shipment_history`	~12k	Material × week shipments for AWD (rolling 24 weeks)
`master_woc`	~26k	Weeks of Coverage: SI, net_available, AWD, WOC, woc_low_flag

Table Relationships: Keys & Joins

Normalized keys: All join keys are normalized (string, trimmed, nulls → empty). For delivery/billing joins to SO, we use *canonical document/item keys* so `4500000051` and `51` match.

Primary Keys (Grain)

Table	Primary Key (Grain)
`master_order_fulfillment`	`client_id`, `sales_document_number`, `item_number`: one row per SO item
`master_inventory_material`	`client_id`, `material_number`, `plant_code`: one row per material/plant
`master_purchase`	`client_id`, `purchase_order_number`, `purchase_order_item_number`: one row per PO item
`master_order_fulfillment_brd`	Same as master_order_fulfillment
`shipment_history`	`client_id`, `material_number`, `plant_code`, `shipment_week`
`master_woc`	`client_id`, `material_number`, `plant_code`

Join Logic by Table

master_order_fulfillment

Base: `sales_order_item`. Joined with:

- **sales_order_header** on `client_id`, `sales_document_number`: customer, dates, value
- **delivery_item** (agg) on canonical `client_id`, `_doc`, `_item`: reference_document_number → _doc, reference_item_number → _item
- **billing_document_item** (agg) on canonical `client_id`, `_doc`, `_item`: aubel → _doc, aupos → _item
- **sales_order_schedule_line** on `client_id`, `sales_document_number`, `item_number`
- **material_master**, **material_description** on `client_id`, `material_number`
- **customer_master** on `client_id`, `customer_number`
- **sales_document_status_item** on `client_id`, `sales_document_number`, `item_number`

master_inventory_material

Base: `material_stock`. Joined with:

- **material_master** on `client_id`, `material_number`
- **material_description** on `client_id`, `material_number`
- **plant** on `client_id`, `plant_code`

master_purchase

Base: `purchase_order_item`. Joined with:

- **purchase_order_header** on `client_id`, `purchase_order_number`
- **vendor_master** on `client_id`, `vendor_number`
- **material_master** on `client_id`, `material_number`

master_order_fulfillment_brd

Adds BRD metrics to `master_order_fulfillment`. Joins:

- **master_inventory_material** (agg by material) on `client_id`, `material_number` → `saleable_inventory` (sum of unrestricted_stock)

shipment_history

Built from `delivery_item` + `delivery_header`:

- **delivery_item** + **delivery_header** on `client_id`, `sales_document_number` → goods_issue_date
- Aggregated by `client_id`, `material_number`, `plant_code`, `shipment_week`

master_woc

Combines three sources on `client_id`, `material_number`, `plant_code`:

- **master_inventory_material** → saleable_inventory (unrestricted_stock)
- **master_order_fulfillment_brd** (open lines) → open_order_qty (outstanding_qty)
- **shipment_history** → AWD (rolling 24-week average)
- net_available = SI – open_order_qty; WOC = net_available ÷ AWD

Pipeline Flow

Step 1: Core Master Tables

`src/data/build_master_tables.py`

Reads from `data/clean/main/` → Produces: master_order_fulfillment, master_inventory_material, master_purchase

Step 2: BRD Metrics

`src/data/build_brd_metrics.py`

Reads core tables + delivery data → Produces: master_order_fulfillment_BRD, shipment_history, master_WOC

Next Step: EDA & Targets

notebooks/01_eda_targets.ipynb: Exploratory analysis on master tables, target definition (backorder/overstock), data quality checks, visualizations.

Inputs: master_order_fulfillment_brd, master_inventory_material, master_woc, shipment_history

Glossary

Term	Definition
AWD	Average Weekly Demand: rolling 24-week average of units shipped
BRD	Business Requirements Document

Term	Definition
CSV	Comma-Separated Values: flat file format
EDA	Exploratory Data Analysis
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
PO	Purchase Order
SAP	Systems, Applications, and Products (ERP software)
SI	Saleable Inventory: unrestricted stock available to fulfill orders
SO	Sales Order
agg	Aggregated: summed/grouped by key columns
sales_org	Sales Organization: SAP organizational unit
WOC	Weeks of Coverage: net available inventory ÷ AWD