. Key MDM Concepts

Understanding these core pillars helps you speak the same language as business stakeholders:

- **The Golden Record (Single Source of Truth):** The core "master" version of an entity (e.g., a specific Taxi Zone) that has been de-duplicated, cleansed, and validated.
- **Data Domains:** Logical groupings of master data. In your case: **Location**, **Partner/Vendor**, and **Product/Tariff**.
- **Data Stewardship:** The "human" side of MDM. While you build the algorithm, a **Data Steward** (e.g., a member of the TLC Operations team) is the one who manually resolves "suspect" matches that the algorithm flags.
- **Survivorship:** The automated ruleset that decides which source system "wins" when data conflicts. (e.g., "If two systems provide a Zone Name, always trust the GIS Department's version first.")

---

2. Strategic Business Alignment

MDM is expensive to build, so it must be justified by business value. For the NYC Taxi dataset, the alignment looks like this:

Business Objective	MDM Contribution	ROI Impact
Revenue Integrity	Ensuring **Rate Codes** are standard across all vendors.	Prevents fare leakage and avoids lawsuits over illegal surcharges.
Operational Reporting	Standardized **Taxi Zones** for all 10M+ monthly trips.	Allows planners to identify "dead zones" accurately to optimize driver distribution.
Vendor Accountability	Unified **Vendor Master** with contact and contract data.	Reduces time to resolve data outages from days to minutes.

Export to Sheets

---

3. The MDM Maturity Framework (Crawl, Walk, Run)

You don't need a million-dollar software suite to start. Most organizations follow this path:

1. **Registry Style (Crawl):** You keep data in original systems but use a central "Index" (your RDS table) to link them together.
2. **Consolidated Style (Walk):** You pull data from various systems into your central RDS to create the Golden Record for reporting. **(This is where your current project sits.)**
3. **Centralized/Transactional Style (Run):** Systems must check with the MDM RDS *before* they can create a new record. This prevents duplicates from ever entering the ecosystem.

---

4. Implementation Checklist for Business Buy-In

To move from "concepts" to "production," you need three non-technical components:

- **The Data Governance Council:** A monthly meeting with the Business Owners you identified earlier to approve changes to Rate Codes or Zones.
- **SLA (Service Level Agreement):** A promise to the business on data quality (e.g., "99.9% of all trips will map to a valid Master Zone ID").
- **Change Management:** A process to notify downstream analysts when a Master record changes (e.g., if a Zone ID is retired).

What is MDM and why it matters to the business?

What is MDM?

Master Data Management is a business-led program that ensures an organization's shared data (like customers, products, or taxi zones) is accurate, consistent, and available across the entire enterprise.

It is the process of taking data from different "silos" (different apps or departments), cleaning it, and creating one **Golden Record** that everyone agrees is the truth.

The "Before and After" MDM

- **Before MDM:** You ask three different departments "Which taxi zone earned the most revenue?" and you get **three different answers** because each department uses a different version of the "Zone List."
- **After MDM:** Every department pulls from the same **Golden Record** table in RDS. There is only one answer.

Why It Matters to the Business

Businesses don't invest in MDM just to have "clean tables"—they do it because "dirty data" has a high financial and operational cost.

1. The "One Version of the Truth" (Decision Making)

If the **Rate Code** master data is inconsistent, a CEO might see a report saying airport trips are declining, while the Finance team sees them increasing. MDM eliminates these "data brawls" in meetings.

- **Business Impact:** Faster, more confident strategic decisions.

2. Operational Efficiency (Cost Savings)

Imagine if a **Vendor** changes their name or contact info. Without MDM, an admin might have to update that info in 10 different systems manually. With MDM, you update it once, and it "pushes" to everywhere else.

- **Business Impact:** Reduced manual labor and "data repair" costs.

3. Regulatory Compliance & Risk

In the NYC Taxi world, the TLC is a regulator. If the **Taxi Zone** data is wrong, the city might collect the wrong amount of "Congestion Surcharge" tax. This leads to audits, fines, and bad PR.

- **Business Impact:** Lower legal risk and guaranteed compliance with city laws.

4. Improved Customer/Passenger Experience

If the master data for **Payment Types** is messy, a passenger might be charged twice or have a "Dispute" recorded incorrectly.

- **Business Impact:** Higher trust in the service and fewer support tickets.

The "Business Justification" Summary

If you were pitching MDM to the TLC Commissioner, you would say:

"MDM is the foundation that ensures every dollar we track, every zone we analyze, and every vendor we regulate is based on the same high-quality information. **We aren't just managing data; we are managing the integrity of the city's transportation policy.**"

Master data domains: Customer, Product, Location, Vendor

For each domain: Who cares? What decisions depend on it? What breaks if it's wrong?

1. Customer Domain

In the NYC Taxi context, this usually refers to "Credit Card Hash/ID" or "Corporate Accounts."

- **Who cares?** Marketing, Customer Loyalty teams, and the Finance Department.
- **Decisions depending on it:** Whom do we offer discounts to? How do we identify "Power Users"? What is the Customer Lifetime Value (CLV)?
- **What breaks if it's wrong?**
    - **Privacy Violations:** You accidentally link Trip A to the wrong Person B.
    - **Customer Friction:** A regular rider doesn't get their "10th ride free" perk because their ID was duplicated as two different people.

---

2. Product Domain

*In the Taxi dataset, the "Product" is the **Service Type** (Yellow Taxi, Green Taxi, For-Hire-Vehicle) or the **Rate Code** (Standard, Airport, Negotiation).*

- **Who cares?** Product Managers, Policy Makers, and Pricing Analysts.
- **Decisions depending on it:** Do we need to increase the price of JFK trips? Is the "Green Taxi" program failing to serve the outer boroughs?
- **What breaks if it's wrong?**
    - **Financial Leakage:** You undercharge for a premium service because the "Product" was mapped incorrectly to a cheaper rate.
    - **Strategy Failure:** The city invests in Green Taxis based on data that was actually mixed with Yellow Taxi data.

---

3. Location Domain

*This is your **Taxi Zones** (Boroughs, Neighborhoods, Service Zones).*

- **Who cares?** City Planners, Dispatchers, and Drivers.
- **Decisions depending on it:** Where should we build new taxi stands? Where is the "Congestion Fee" applied? Which areas are "underserved"?
- **What breaks if it's wrong?**

    - **Legal/Tax Risk:** A driver is charged a "Manhattan Congestion Fee" for a drop-off that actually happened in Brooklyn.
    - **Operational Chaos:** Drivers are sent to the wrong "Hot Zone" because the heatmap logic is based on faulty location master data.

---

4. Vendor Domain

*These are the **Technology Providers** (Verifone, CMT) and **Fleet Owners**.*

- **Who cares?** Procurement, IT Operations, and Legal/Licensing.
- **Decisions depending on it:** Which vendor has the most technical downtime? Which contract should we renew? Which vendor is failing to report data accurately?
- **What breaks if it's wrong?**
    - **Audit Failure:** You cannot prove to a regulator which company was responsible for a data gap.
    - **Payment Delays:** The city sends a payout to "Vendor A" when it should have gone to "Vendor B" due to a name change that wasn't updated in the master list.

Summary Table for Stakeholders

Domain	Primary Stakeholder	Worst Case Scenario
Customer	Marketing/CRM	Legal Lawsuits & Bad Brand Rep
Product	Finance/Policy	Revenue Loss & Wrong Pricing
Location	City Planning	Traffic Gridlock & Unfair Taxing
Vendor	Procurement/IT	Contract Breaches & Data Gaps

MDM implementation styles with real-world scenarios:

Here are the four standard MDM implementation styles applied to the NYC Taxi (TLC) ecosystem:

---

1. Registry Style

Concept: The "Index" approach. You don't move the data from the silos. Instead, you create a central list of IDs that links records together.

- **Real-World Scenario:** The TLC keeps a central registry that says "Vendor ID 1" in the Finance system is the same as "Vendor ID 1" in the GPS tracking system.

- **Best for:** Large organizations where departments refuse to change their local systems.
- **The "Golden Record":** It doesn't actually exist in one place; it is "assembled" on the fly when needed.

2. Consolidated Style

Concept: The "Reporting Hub." Data is pulled from various silos into a central RDS (like your project) to create a single Golden Record for analytics.[1]

- **Real-World Scenario:** You pull "Taxi Zone" names from the GIS team and "Zone Surcharge Rates" from the Legal team into one Master Table. Analysts use this table for their monthly reports instead of the raw silo data.
- **Best for:** Business Intelligence (BI) and reporting.
- **The "Golden Record":** Lives in your central RDS, but changes made there **do not** flow back to the source systems.

---

3. Coexistence Style

Concept: The "Bi-Directional" approach. The Golden Record is created in the central hub, but any updates made to the Golden Record are pushed back to the original silos.[2]

- **Real-World Scenario:** A **Vendor** changes their official business address. You update it in the MDM Hub. The Hub then automatically updates the Licensing Department's database AND the Payment Department's database.
- **Best for:** Maintaining consistency across the whole company while letting departments keep their own software.
- **The "Golden Record":** Lives in the hub and acts as the "Master Copy" that synchronizes everyone else.

---

4. Centralized (Transaction) Style

Concept: The "Dictatorship" approach. Data is created, stored, and managed **only** in the MDM hub. If a department wants to add a new "Rate Code," they must do it directly in the MDM system.

- **Real-World Scenario:** The TLC creates a new **Rate Code 7** (e.g., "Holiday Surcharge"). No vendor is allowed to use it until it is first defined and published in the Central Master Table. The vendors then "subscribe" to this table to get the update.
- **Best for:** High-security and highly regulated data where mistakes are unacceptable.
- **The "Golden Record":** The **only** record that exists.

Comparison of Styles

Style	Difficulty	Data Consistency	Best Use Case
Registry	Low	Low (Lookup only)	Quick wins with limited budget.
Consolidated	Medium	Medium (Reporting)	**Your current NYC Taxi project.**
Coexistence	High	High (Synchronized)	Enterprise-wide digital transformation.
Centralized	Very High	Total (Single Source)	Critical financial or legal definitions.

MDM Governance Model:

Governance defines the rules, but the **People** make it work. Here is how those roles and responsibilities break down for your specific taxi domains.

---

1. The MDM Governance Matrix

Role	Responsibility	For the NYC Taxi Project...
Data Owner	**Accountable.** Defines the business rules and takes the "blame" if data is legally wrong.	**TLC Commissioner / Head of Policy.** (e.g., Decides that the JFK rate is exactly $70).
Data Steward	**Operational.** Manages the "Golden Record." Resolves duplicates and approves new Rate Codes.	**TLC Business Analyst.** (e.g., Manually checks if "Zone 264" is valid before you add it to RDS).
Data Custodian	**Technical.** You! You build the pipelines, the RDS tables, and the matching algorithms.	**Data Engineer.** (e.g., Writes the SQL and Python that moves data from silos to the hub).
Data Consumer	**Usage.** Queries the Golden Records to build insights or train models.	**Data Scientists / Traffic Planners.** (e.g., Person building the "Taxi Demand Prediction" model).

---

2. Responsibilities in Action: The "New Zone" Scenario

Imagine a new taxi zone is created. Here is how your roles interact:

1. **Data Owner:** Signs a legal document authorizing a new neighborhood zone for congestion pricing.

2. **Data Steward:** Enters the new Zone Name and Borough into a "Staging" table and double-checks for typos.
3. **Data Custodian (You):** Runs the ETL script that promotes that record into the **Golden Record** table and assigns it a version_id. You ensure the database index is updated for performance.
4. **Data Consumer:** Refreshes their PowerBI dashboard to show the new zone's trip volume.

Hands-on:

Design golden record strategy for taxi zones with survivorship rules documented

Taxi Zone Survivorship Matrix

This table acts as your governance documentation. It defines the "Tie-breaker" logic for when sources conflict.

Field	Primary Source (High Trust)	Secondary Source	Survivorship Rule	Logic
Location ID	TLC Official List	System Internal ID	**Static / Immutable**	This is the anchor. If the ID changes, it's a new record, not an update.
Zone Name	City Planning / GIS	Dispatch App	**Manual Overwrite**	Official names (e.g., "Stuyvesant Town") win over app shorthands (e.g., "Stuy Town").
Borough	GIS Mapping	User Input	**Validation Logic**	The Borough is determined by the geographic coordinates, regardless of what the user types.
Service Zone	TLC Business Rules	Dispatch App	**Most Recent**	Use the value from the source with the latest updated_at timestamp.