Step 1: Dataset Analysis (Characteristics & Requirements)

Before classifying, we must look at the nature of the data and who is using it:

1. **Application Logs & IoT Data:** High velocity and variable. These require a "catch-all" storage that won't break when a developer adds a new log field.
2. **Sales & Marketing Results:** These are "The Truth." They drive financial decisions and require strict enforcement to ensure the numbers match every time.
3. **Ad-hoc Extracts:** These are temporary and unpredictable. They don't belong in a structured warehouse because they aren't "permanent" assets.
4. **Customer Master Data:** This is the "Anchor." It must be highly reliable and available to all other systems.

Step 2: Classification Table

Dataset	Storage Type	Zone (if Lake)	Schema Strategy	Ingestion Type	Reasoning
Application Logs	**Data Lake**	Raw	Schema-on-Read	Streaming	High volume/variety makes strict schema enforcement impossible at entry.
Daily Sales Summary	**Data Warehouse**	Gold	Schema-on-Write	Batch	Highly structured and used for BI dashboards where consistency is critical.

Raw IoT Sensor Data	**Data Lake**	Raw	Schema-on-Read	Streaming	Velocity is too high for complex validation; must be stored "as-is" for history.
Customer Master Data	**Data Warehouse**	Silver/Gold	Schema-on-Write	Batch	Reference data requires strict integrity to ensure joins across systems work.
Marketing Results	**Data Warehouse**	Gold	Schema-on-Write	Batch	Structured analytical data used for periodic ROI reporting.
Ad-hoc Extracts	**Data Lake**	Bronze/Sandbox	Schema-on-Read	Batch	Irregular structure and one-off use do not justify the cost of warehouse modeling.

Step 3: Review and Justification

Consistency Check

- **Logs vs. IoT:** Both are assigned to the **Data Lake/Raw** zone. This is consistent because both are "source-of-truth" telemetry data that is too volatile for a strict Warehouse schema during initial ingestion.

- **Sales vs. Marketing:** Both go to the **Data Warehouse**. This aligns with best practices for **OLAP (Analytical)** workloads where the structure is stable and query performance is the priority.

Edge Cases & Hybrid Approaches

- **The IoT Transition:** While IoT starts in the **Lake (Raw)** for real-time monitoring, it almost always flows into the **Warehouse (Gold)** after being aggregated into "Hourly Averages." This is a classic hybrid move.
- **Customer Data:** Some architects keep Customer data in the **Silver Lake** to allow Data Scientists to perform fuzzy matching before "promoting" it to the **Warehouse** for the general business.

Trade-offs

- **Lake Trade-off:** We gain **flexibility and low cost**, but we lose **query speed**. If an analyst tries to query 1TB of raw JSON logs, it will be significantly slower than querying a Warehouse table.
- **Warehouse Trade-off:** We gain **extreme speed and reliability**, but we lose **agility**. If the "Marketing Campaign" format changes, the dbt models and Warehouse schema must be updated before data can flow again.