

| Business Template  **Staging Layer vs 3NF** |
| --- |
|  |

### *Purpose of the Staging Area*

The staging area in a data warehouse is an intermediate storage layer where raw data from multiple source systems is initially loaded *before* further processing. The main purposes include:

1. **Landing / Ingestion**

* It is the first place data lands after extraction from source systems.
* Supports loading large volumes of raw, unprocessed data.

2️. **Data Cleansing**

* Source systems are often *dirty* or *inconsistent*.
* The staging area is where initial data validation, error detection, and basic cleaning occur.

3️. **Transformation Preparation**

* Raw data often comes in different formats, structures, or coding schemes.
* In staging, data can be standardized and reformatted in preparation for transformation.

4️. **Auditing & Debugging**

* Keeping the raw snapshot in staging allows you to *trace* and *audit* what was loaded from the source.
* If a downstream load fails, you don't have to go back to the live source system to re-extract it.

5️. **Handling Late-Arriving Data**

* The transcript describes late-arriving facts and dimensions (e.g. sales data, roaming calls).
* Staging helps manage these by reloading recent periods (like the last 3 months) to ensure completeness.

6️. **Decoupling Source Systems from DW**

* The staging area insulates the source systems from frequent re-extracts, reducing load on operational systems.

### 

### 

### *Why you can't load directly into the 3NF layer*

**Short answer**: Because the raw source data is *not ready* or *clean* enough for the highly structured 3NF layer.

1️. **Source Systems Are Dirty / Inconsistent**

* The 3NF layer enforces strong consistency and integrity—data *must* be clean before it lands there.

2️. **Conformance of Data**

* Different source systems might describe the same thing differently (e.g. product IDs, customer codes).
* Staging is needed to *conform* these into a single, unified dimension.

3️. **Transformation Complexity**

* The transformation needed to get data from raw source to 3NF (joining, deduplication, SCD handling, business rules) is too big to do “on the fly” during load.
* Staging gives you a working area to break the process into manageable steps.

4️. **Audit & Error Recovery**

* If you skip staging and a load into 3NF fails halfway, you have no copy of the raw data to restart from.
* Staging ensures repeatable, reliable ETL.

5️. **Handling Late-Arriving Data**

* Staging enables reloading and reconciliation for these late updates.

6️. **Source Data Volume and Shape**

* Source extracts can be huge, denormalized, or in very different shapes than the normalized 3NF model.
* Staging is where these are reshaped to fit.