

Documenting ETL Requirements



Stacia Varga

CONSULTANT, INSTRUCTOR, AUTHOR

@_StaciaV_ datainspirations.com



Overview



Requirements summary

Standards

Job-level details



Summary

This document describes the ETL requirements for

- What type of project?
- What type of processes?
- How frequently?
- What type of data?



Summary

This document describes the ETL requirements for the Adventure Works Data Warehouse project to support sales analysis.



What type of project?



Summary

This document describes the ETL requirements for the Adventure Works Data Warehouse project to support sales analysis. The ETL processes for this iteration of the project must perform one historical extraction and daily incremental extractions

What type of processes?



Summary


This document describes the ETL requirements for the Adventure Works Data Warehouse project to support sales analysis. The ETL processes for this iteration of the project must perform one historical extraction and daily incremental extractions

How frequently?



Summary

This document describes the ETL requirements for the Adventure Works Data Warehouse project to support sales analysis. The ETL processes for this iteration of the project must perform one historical extraction and daily incremental extractions of **product, customer, and sales data**.



What type of data?



Summary

This document describes the ETL requirements for the Adventure Works Data Warehouse project to support sales analysis. The ETL processes for this iteration of the project must perform one historical extraction and daily incremental extractions of product, customer, and sales data **for loading into a dimensional model.**

**What type of
processes?**



Summary

This document describes the ETL requirements for the Adventure Works Data Warehouse project to support sales analysis. The ETL processes for this iteration of the project must perform one historical extraction and daily incremental extractions of product, customer, and sales data for loading into a dimensional model.



Standards



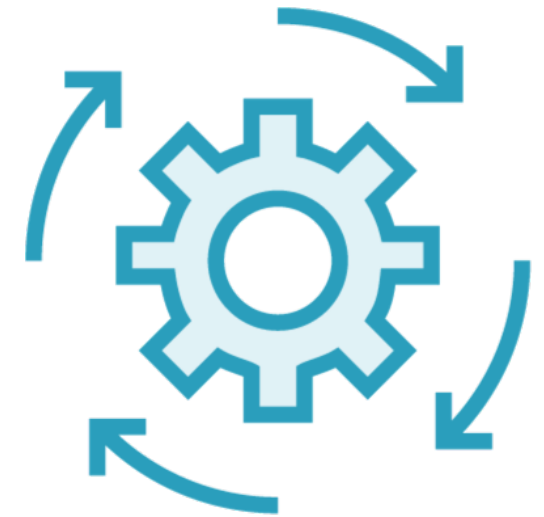
ETL architecture



Auditing and
logging
framework



System
availability
requirements

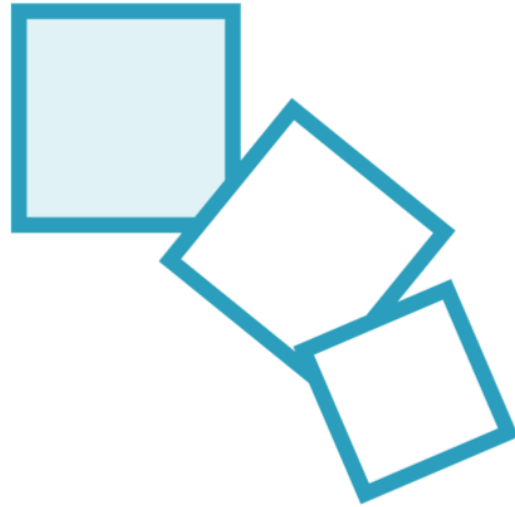


Process
standards

Standards



Extraction
framework



Slowly changing
dimension
handling



Business rule
validation
strategy



Notification
requirements

ETL Architecture

Data extraction

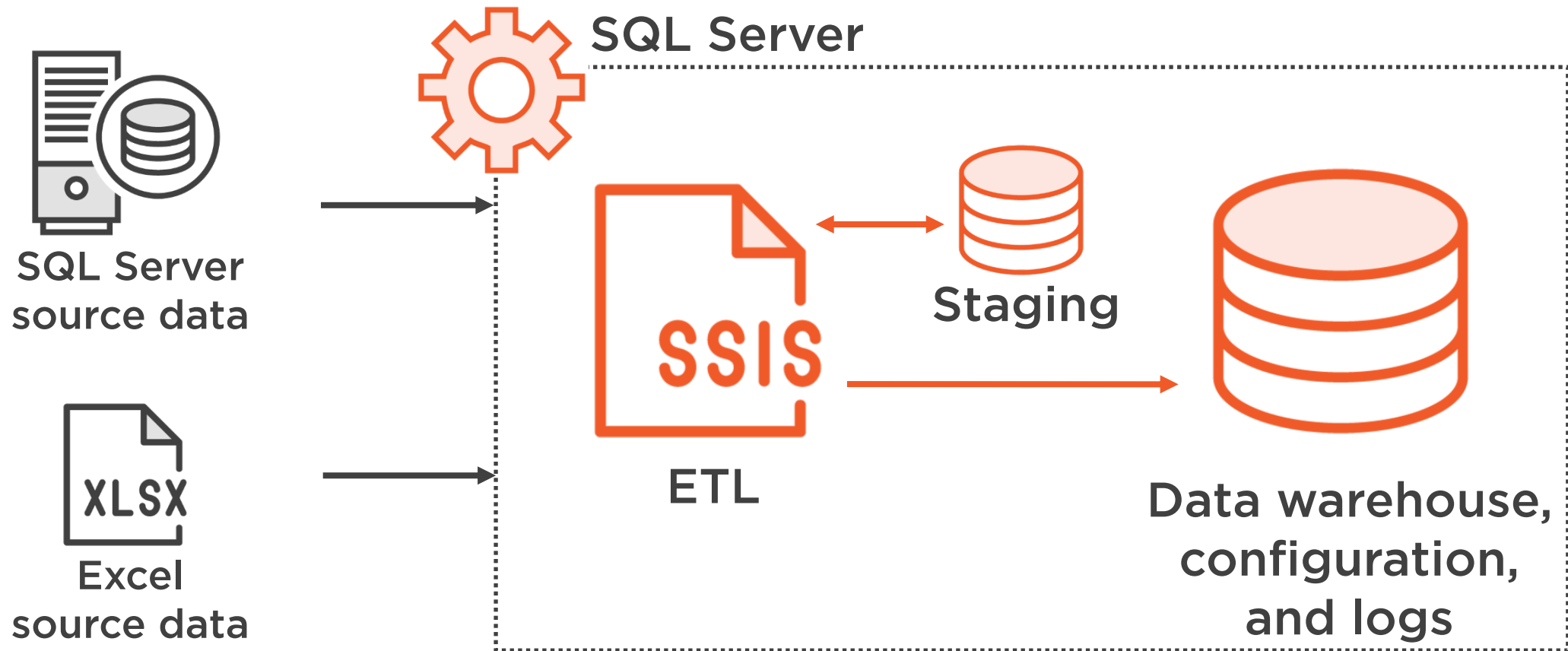
Transformation processing

Data load

Process management



ETL Architecture



Auditing and Logging Mechanism



Call stored procedure
Variable value change raises
event



SSISDB

Internal tables

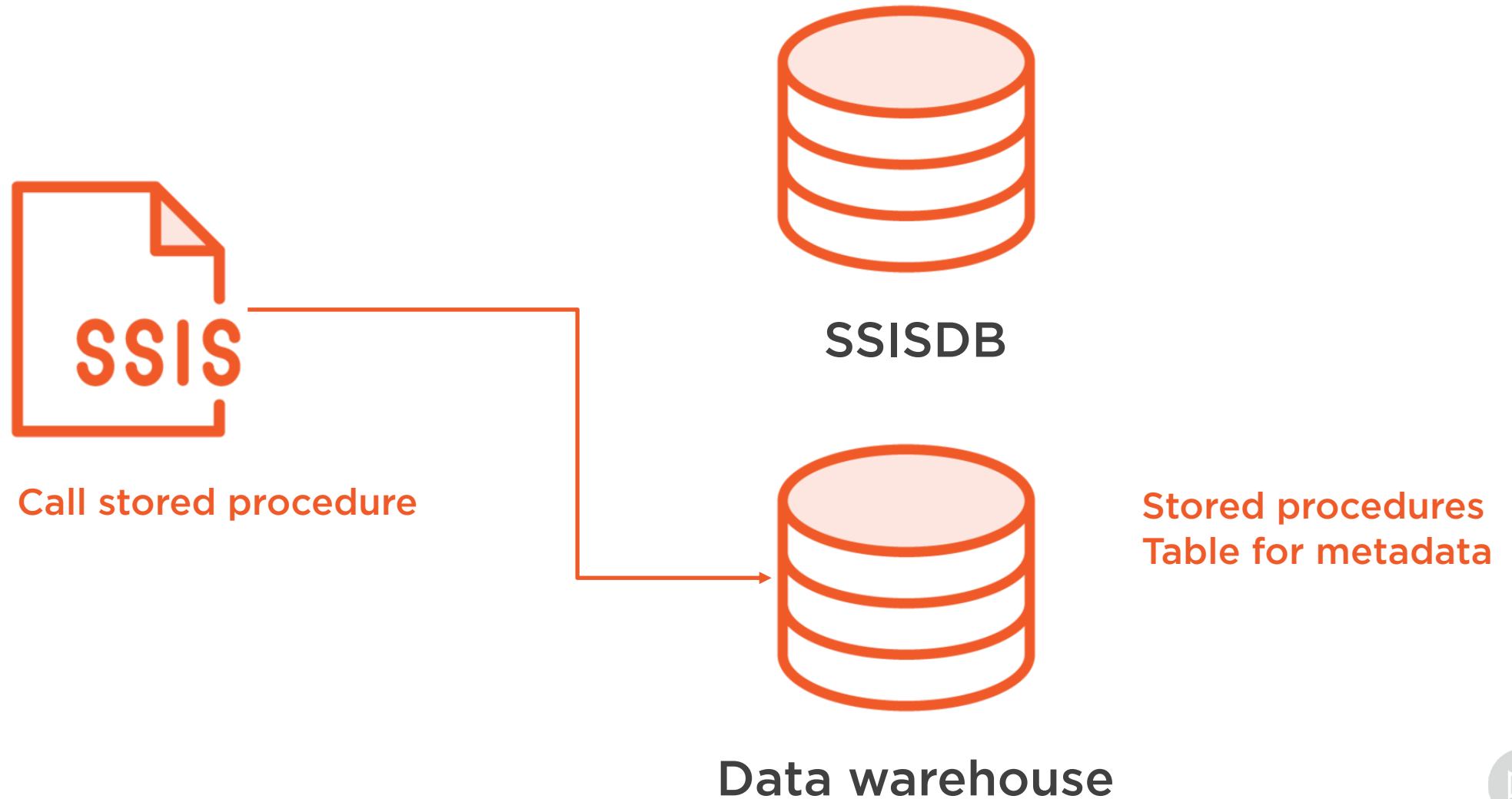


Data warehouse

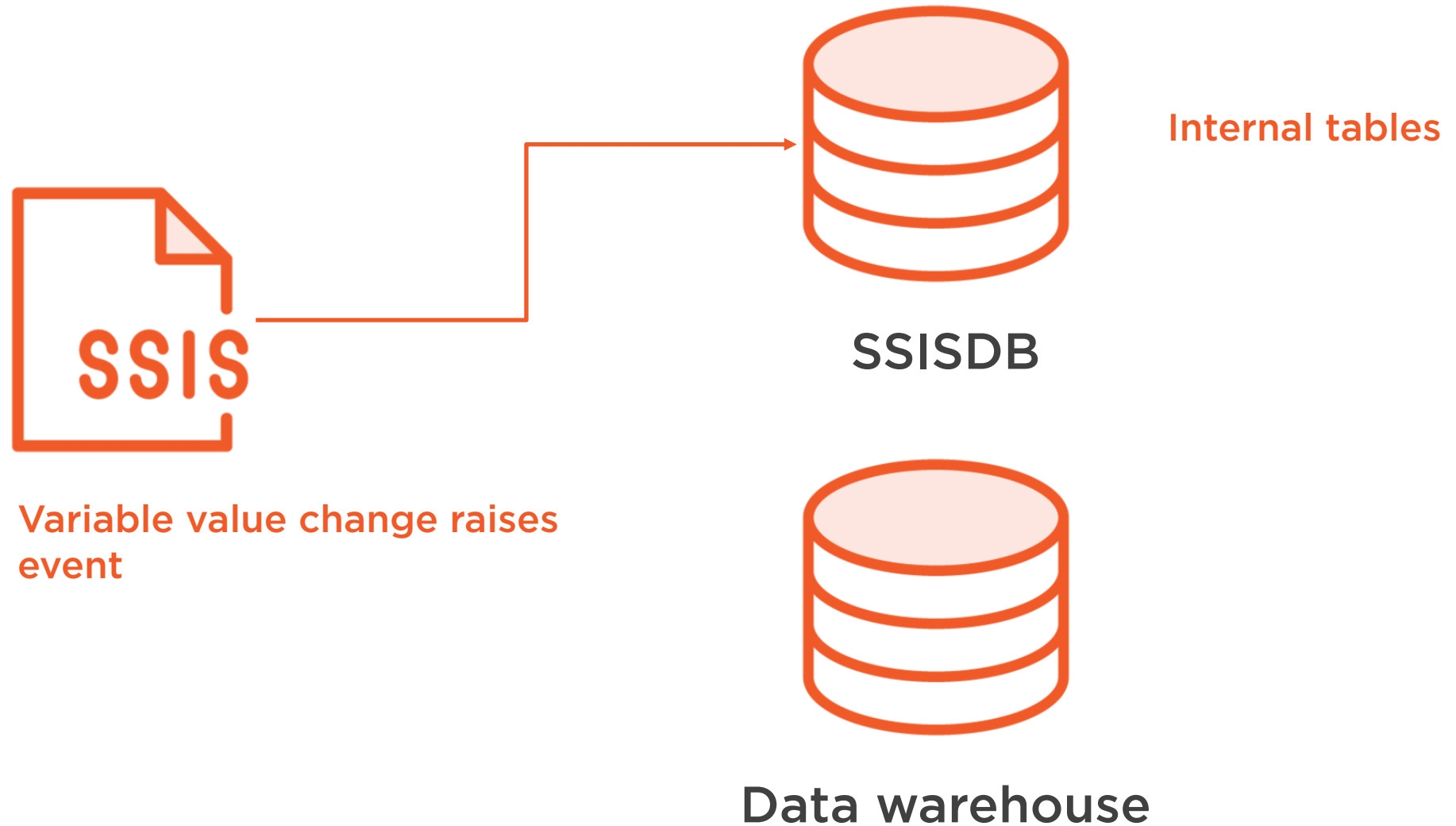
Stored procedures
Table for metadata



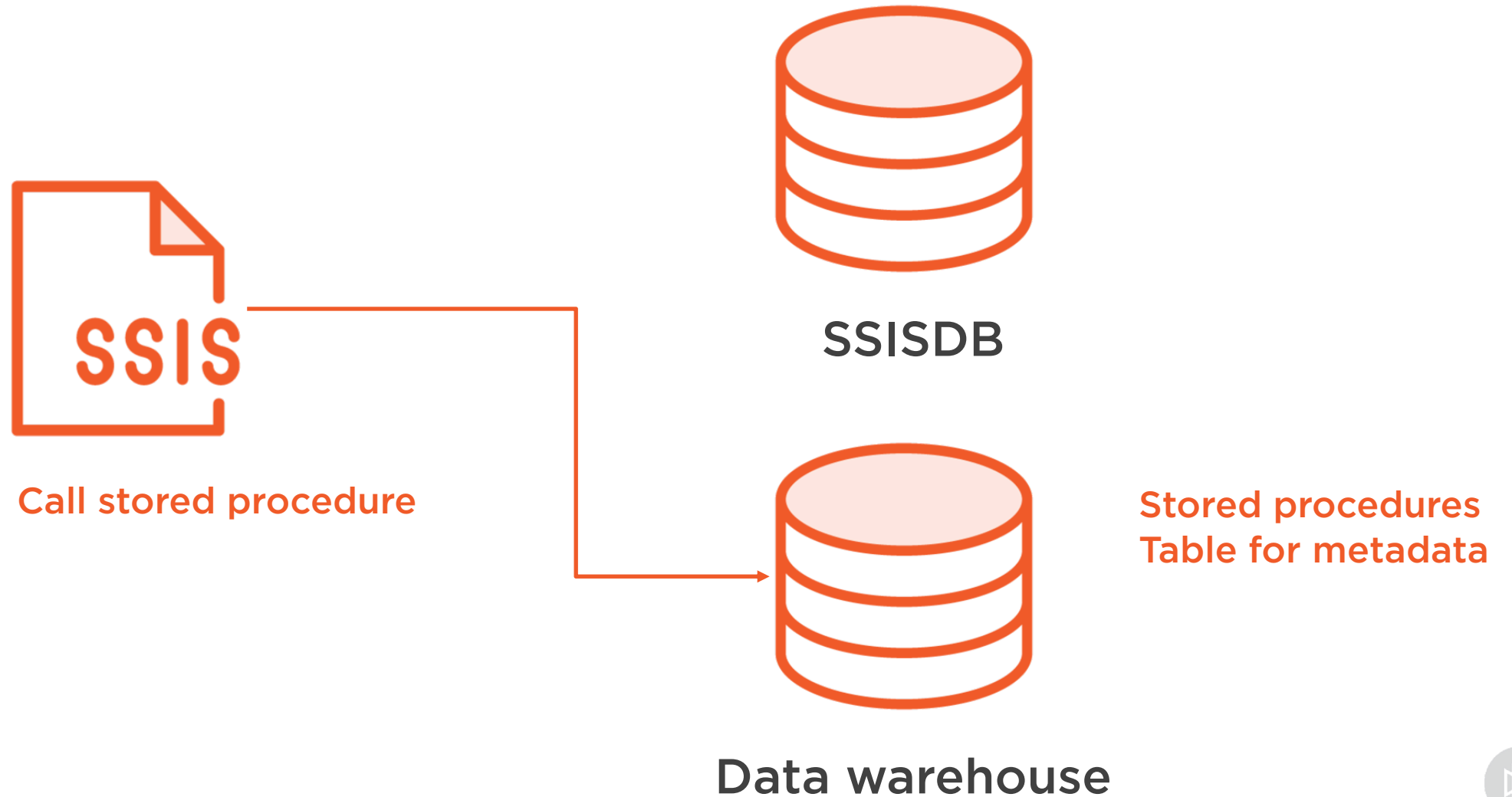
Auditing and Logging Mechanism



Auditing and Logging Mechanism



Auditing and Logging Mechanism



System Availability Requirements



Are there any limits on user access to the target data?



What is the schedule for processing operations?



What procedures are in place if source systems or services are not available?



What should happen if processing operations fail?



System Availability Requirements



Queries can be closed during processing



Processing window opens between 10 pm – 5 am 7 days per week



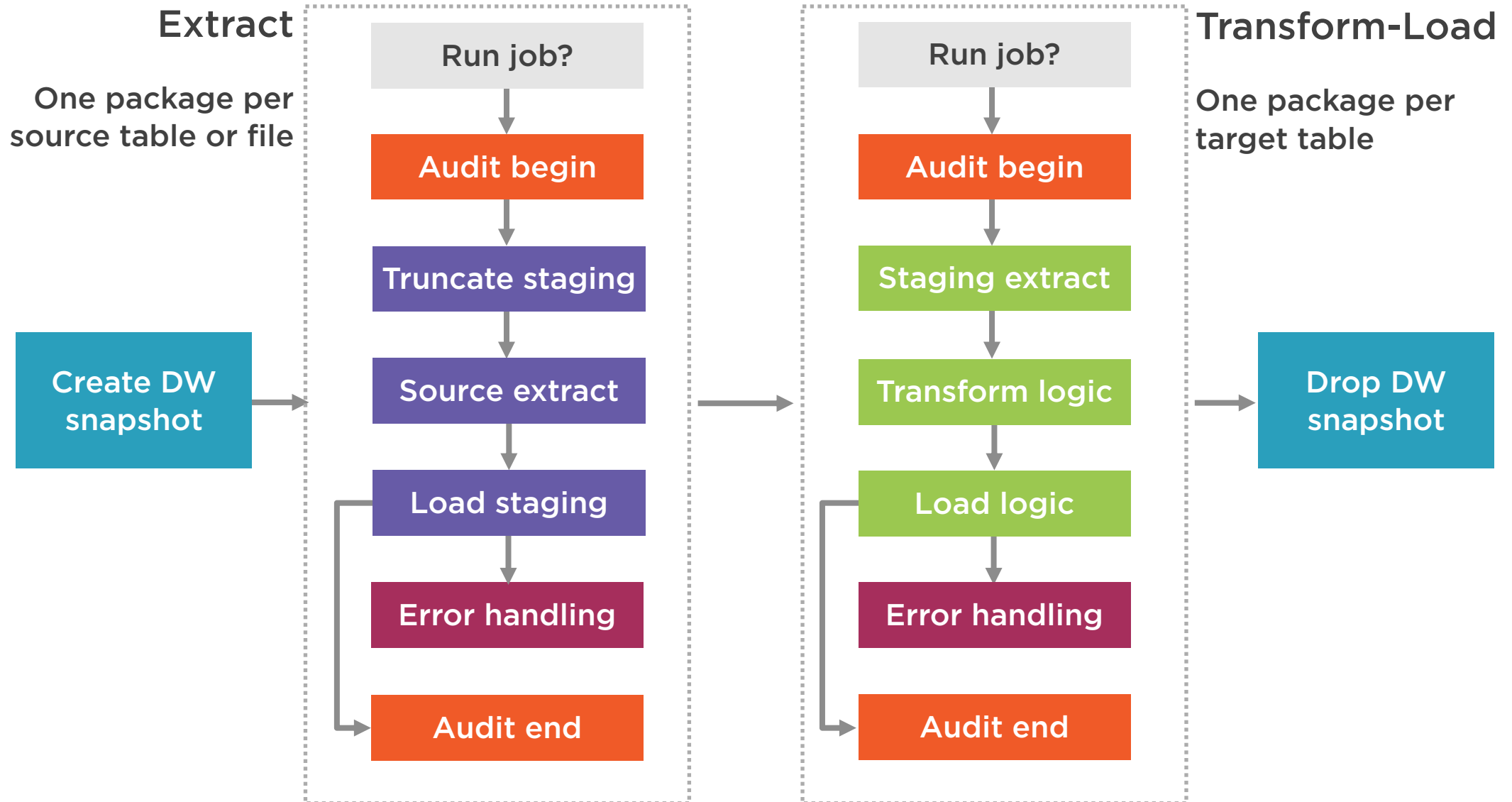
System administrators will monitor and respond to service outages affecting server components and source systems



SSIS jobs must support restarts

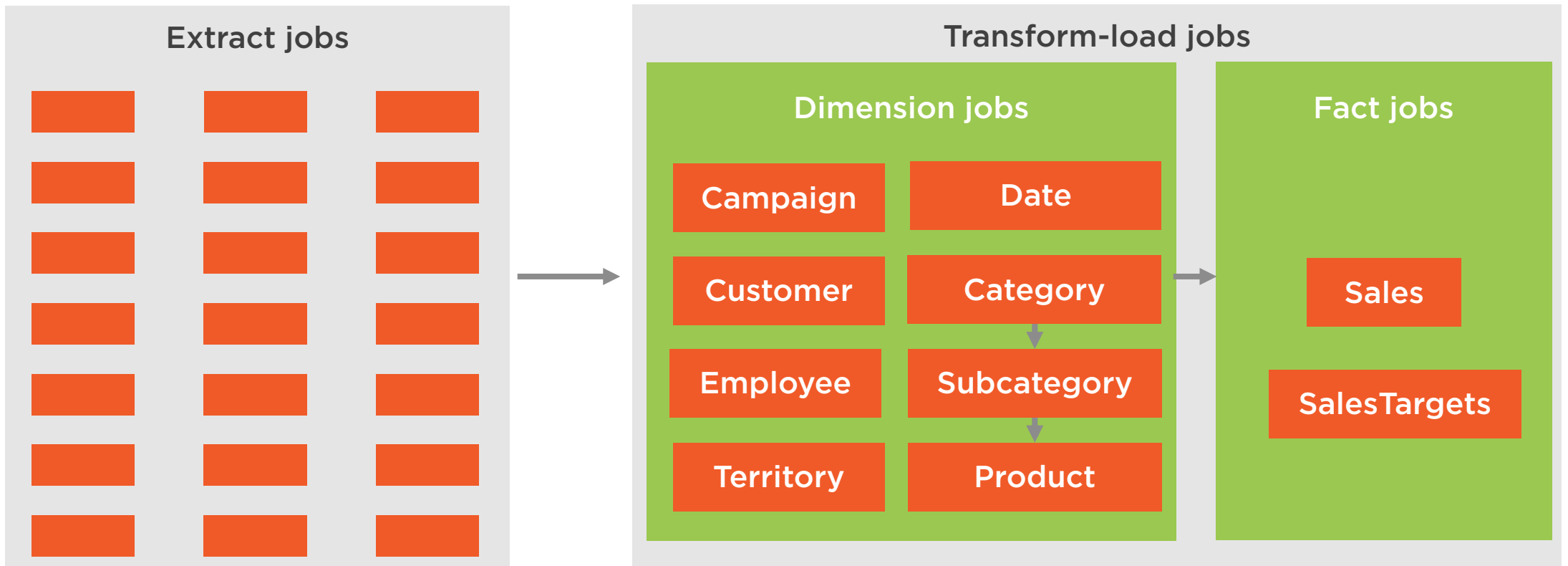


Process Standards

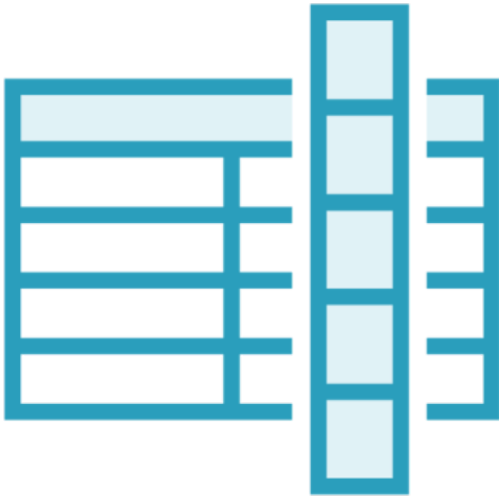


Workflow Orchestration

Data warehouse jobs



Error Handling



Column errors



Structure errors

Error Handling



Halt the current process, notify an operator, apply manual fix to data, and restart the process



Send error row to a separate location for evaluation, continue processing “good” rows, notify an operator, and process error rows separately after fix applied



Apply temporary fix to data, tag row, insert or update with “good” rows, notify an operator, and run process later with permanent fix



Extraction Framework

Staging area

Reduce potential for resource contention if transform-load fails

Table size strategy

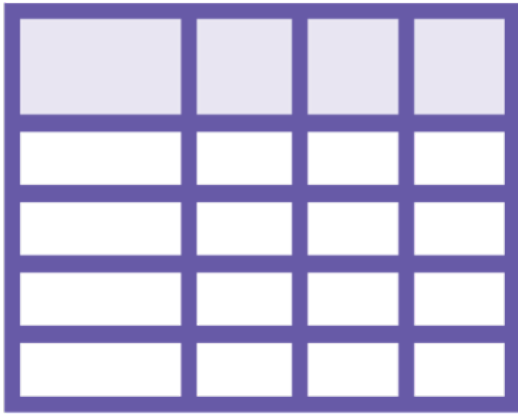
Consider whether data volume affects data extraction impact

Deleted rows

Identify method for finding rows deleted from source tables



Extraction Framework



Staging area

One table per source table or file



Table size strategy

Small: all data
Large: new/changed data



Deleted rows

Not supported in source system - no action required

Slowly Changing Dimension Handling

Type 1

Dimension
Key
Column1
Column2
...

Type 2

Dimension
Key
Column1
Column2
...
IsCurrent
StartDate
EndDate
...



Slowly Changing Dimension Handling: Type 1

Before

CustomerKey	LastName	FirstName
11006	Martin	Chloe

After

CustomerKey	LastName	FirstName
11006	Alvarez	Chloe

Update record

No history preserved



Slowly Changing Dimension Handling: Type 2

Before

ProductKey	Product Name	ListPrice	StartDate	EndDate
466	Half-finger gloves, L	23.5481	2012-07-01	

History preserved



Slowly Changing Dimension Handling: Type 2

After

ProductKey	Product Name	ListPrice	StartDate	EndDate
466	Half-finger gloves, L	23.5481	2012-07-01	2013-07-01
467	Half-finger gloves, L	24.49	2013-07-01	

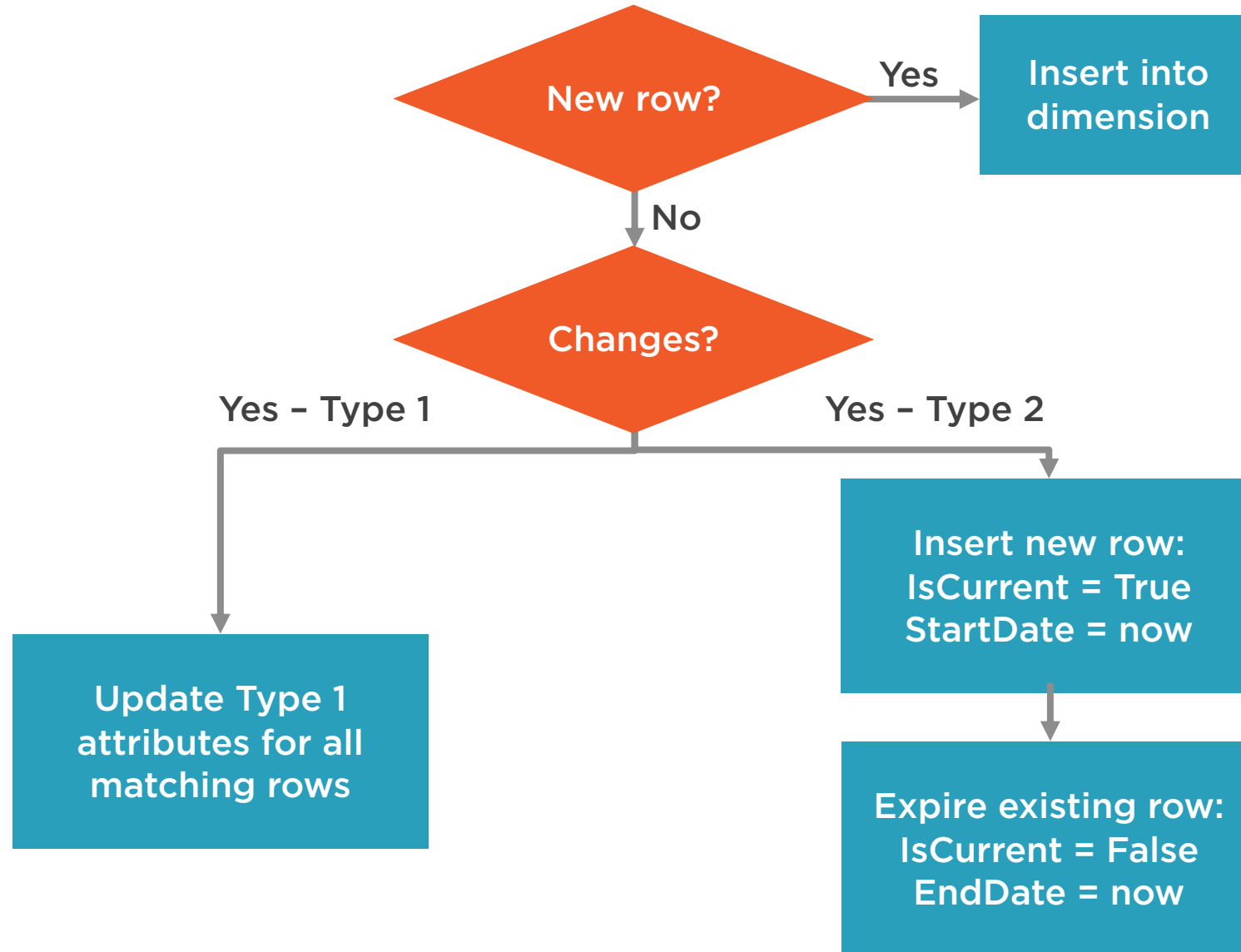
History preserved

Expire old record

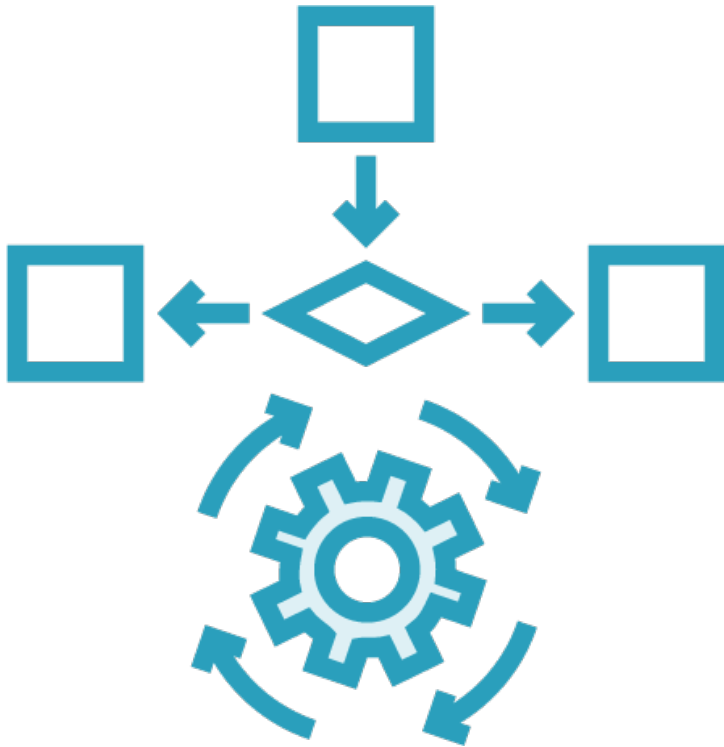
Insert new record



Slowly Changing Dimension Handling



Business Rule Validation Strategy

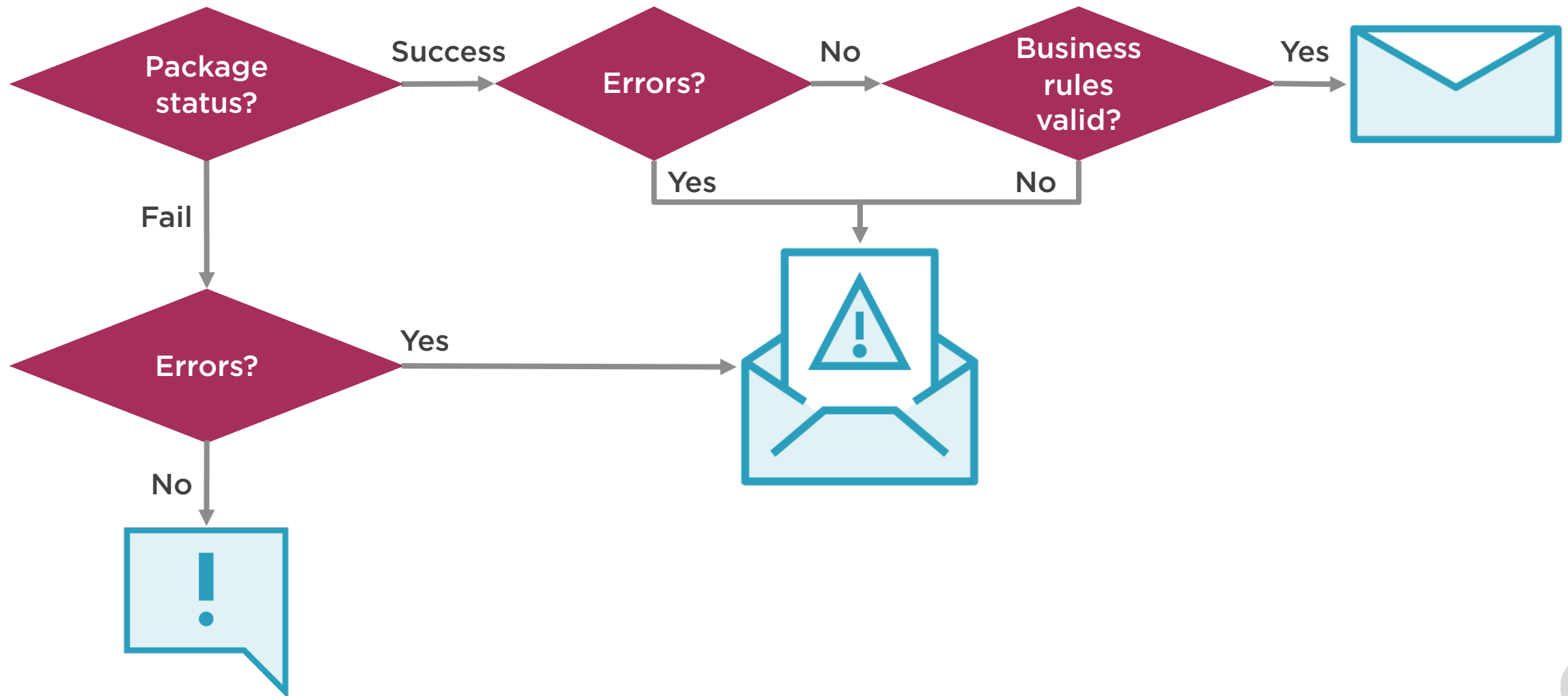


Accept as is?

Accept with default values?

Reject?

Notification Requirements



Job-level Details



Job diagram highlighting tables, known issues, and transformations



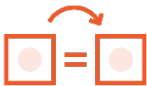
Load frequency for target table



Estimated data volumes for initial and incremental loads



Location of source and target data dictionaries



Location of source-to-target mapping details



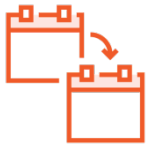
Job-level Details



Location of data profiling documentation



Slowly changing dimension handling by column



Late arriving data handling

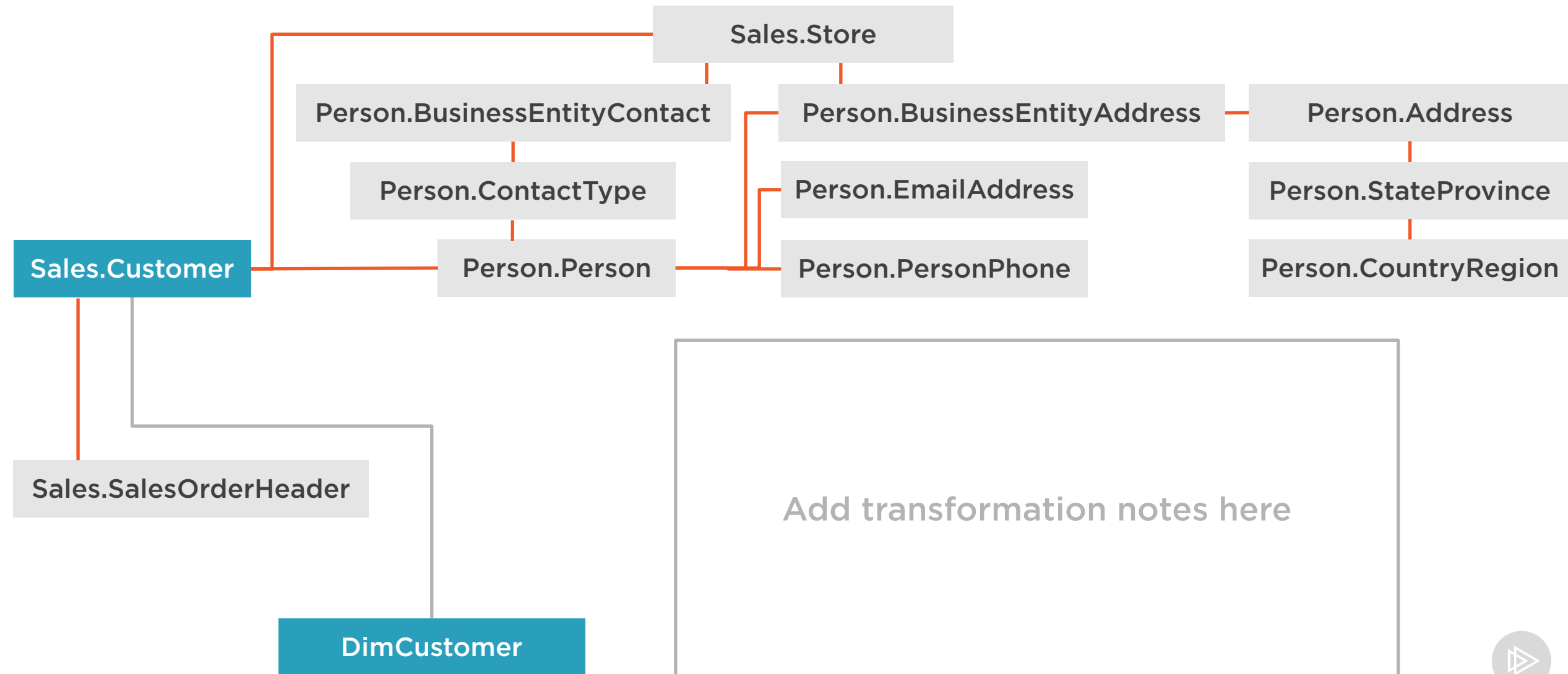


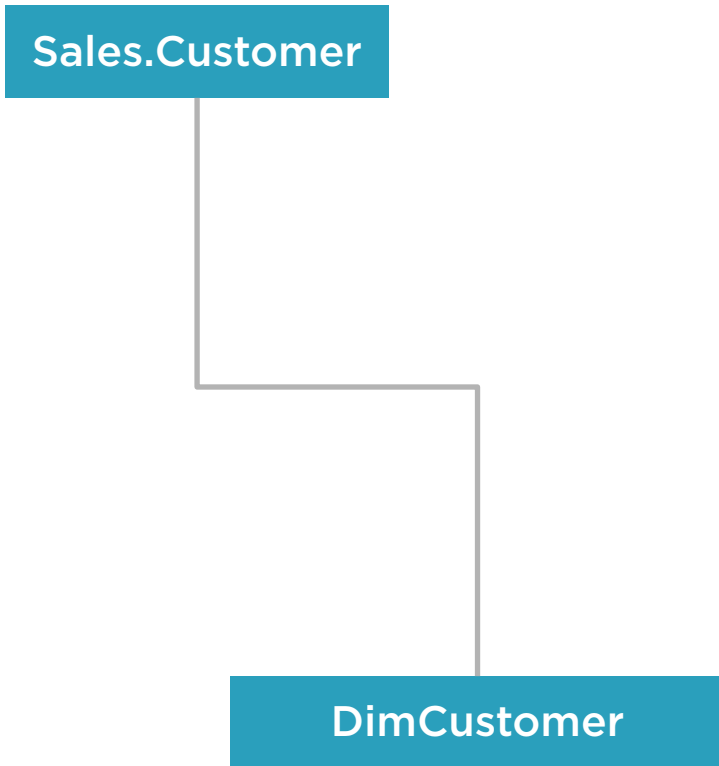
Dependencies



Deviations from standards

Job Diagram





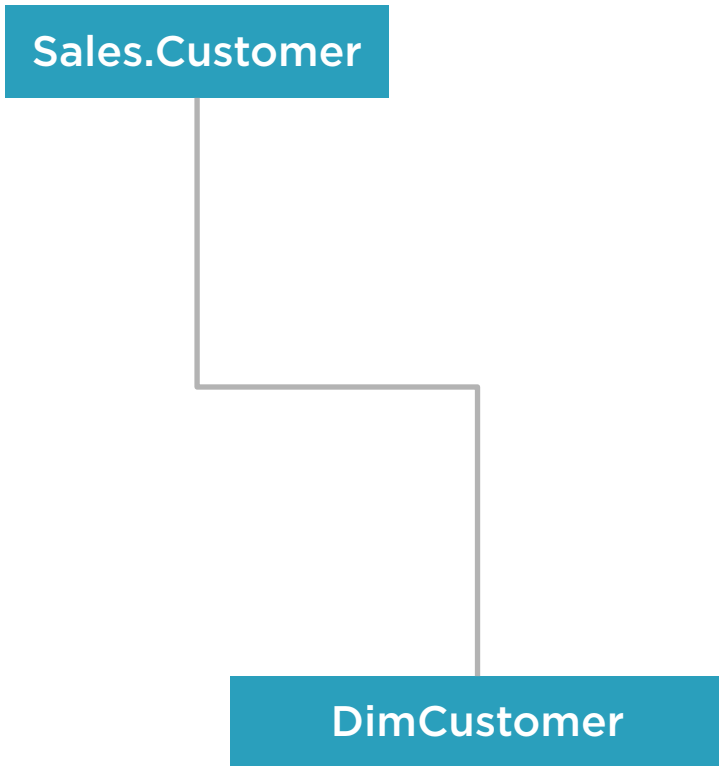
Build customer dimension row from Sales.Customer

Perform lookups to related tables for address, phone, etc.

Use contact type Purchasing Manager or Owner to get name, phone, email for store

Derive channel: if StoreID is null "Online" else "Retailer"





Derive FirstOrderDate

- After initial fact load
- After insert of new customer row

Derive LastOrderDate

- After initial or incremental fact load

SCD1: Names, Email, Phone, Addresses

SCD2: City, State, Country, Postal Code



Data Dictionaries

Sales.Customer			
Contains information about individual customers and retail stores having one or more sales			
Column	Data type	Nullability	Description
CustomerID	int	Not null	Primary key for customer rows
PersonID	int	Null	Unique number to identify an individual customer Foreign key to Person.Person
StoreID	int	Null	Unique number to identify a store Foreign key to Sales.Store
TerritoryID	int	Null	ID of territory in which customer is located Foreign key to SalesTerritory.TerritoryID
AccountNumber	varchar(10)	Not null	Unique number to identify customer, computed by prefixing CustomerID with "AW" and leading zeros
rowguid	uniqueidentifier	Not null	Unique identifier for row
ModifiedDate	datetime	Not null	Date and time of last update to row



Source-to-target Mapping



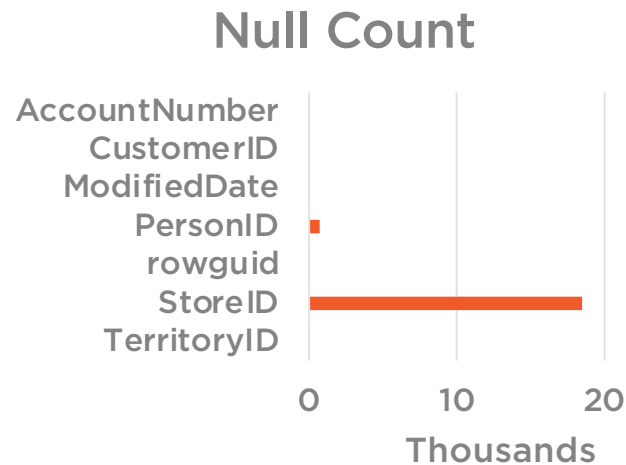
Source Table	Column	ETL Rules	Target Column	Key?	Nullable?	SCD Type
		Identity insert	CustomerKey	PK	N	
Sales.Customer	CustomerID		CustomerID		N	
		If StoreID is null then Online else Retailer	Channel		N	1
Person.Person	Title	Lookup PersonID using CustomerID joined to Person.Person. BusinessEntityID	CustomerTitle		Y	1
Person.Person	FirstName	“	FirstName		Y	1
...						



Data Profile Reports

Column	Min	Max
CustomerID	1	30118
PersonID	291	20777
StoreID	292	2051
TerritoryID	1	10

Value ranges



Completeness

Column	Distinct Values
AccountNumber	19,820
CustomerID	19,820
PersonID	19,119
StoreID	701
TerritoryID	10

Uniqueness



Late-arriving Data Handling

Dimension members

Insert placeholder row
and update columns
later

Dimension Type 2 updates

Update affected fact
table rows

Facts

Identify method for
finding rows deleted
from source tables



Summary



Requirements summary

Standards

- Architecture
- Strategies

Job-level details

- Diagrams
- Data structure and content
- Dependencies
- Special cases