Designing SSIS Integration Solutions

PREPARING FOR THE ETL DESIGN PROCESS

Stacia Varga
CONSULTANT – INSTRUCTOR - AUTHOR

@_StaciaV_ datainspirations.com

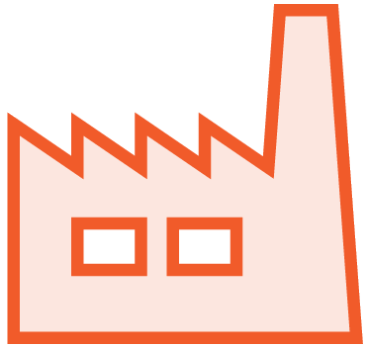
Overview

Project introduction

Requirements

Data profiling

Project Introduction



Company overview

Project plan

Team

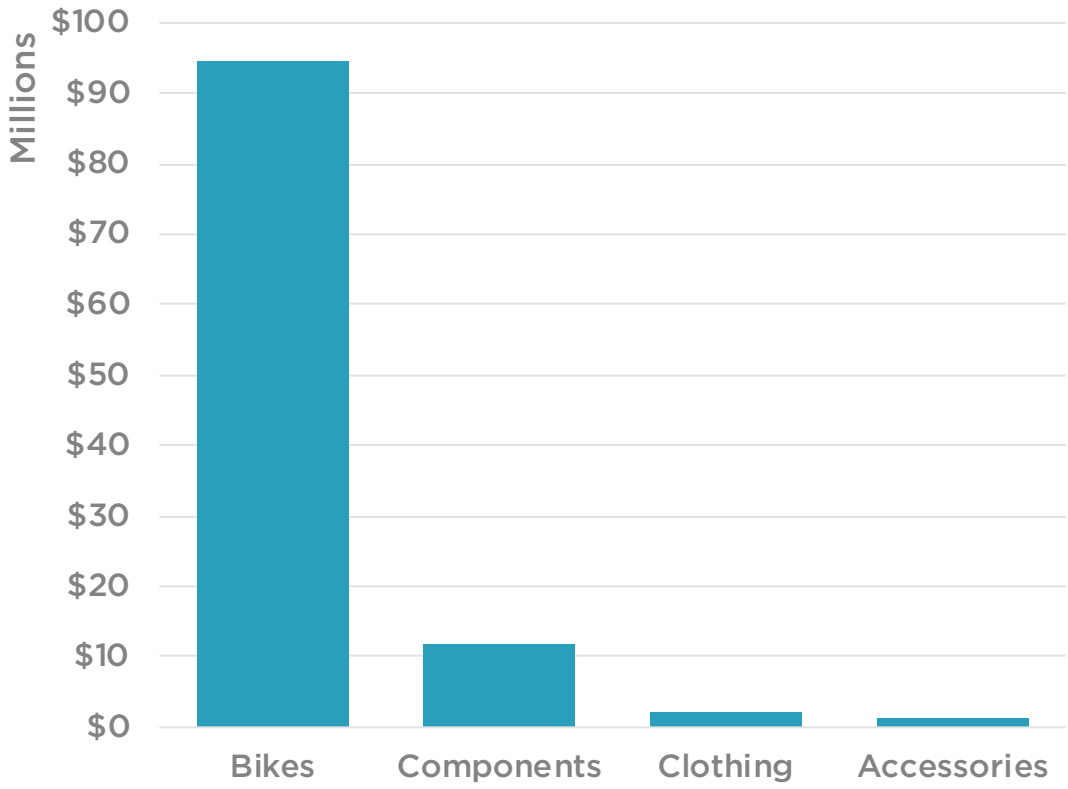
Adventure Works

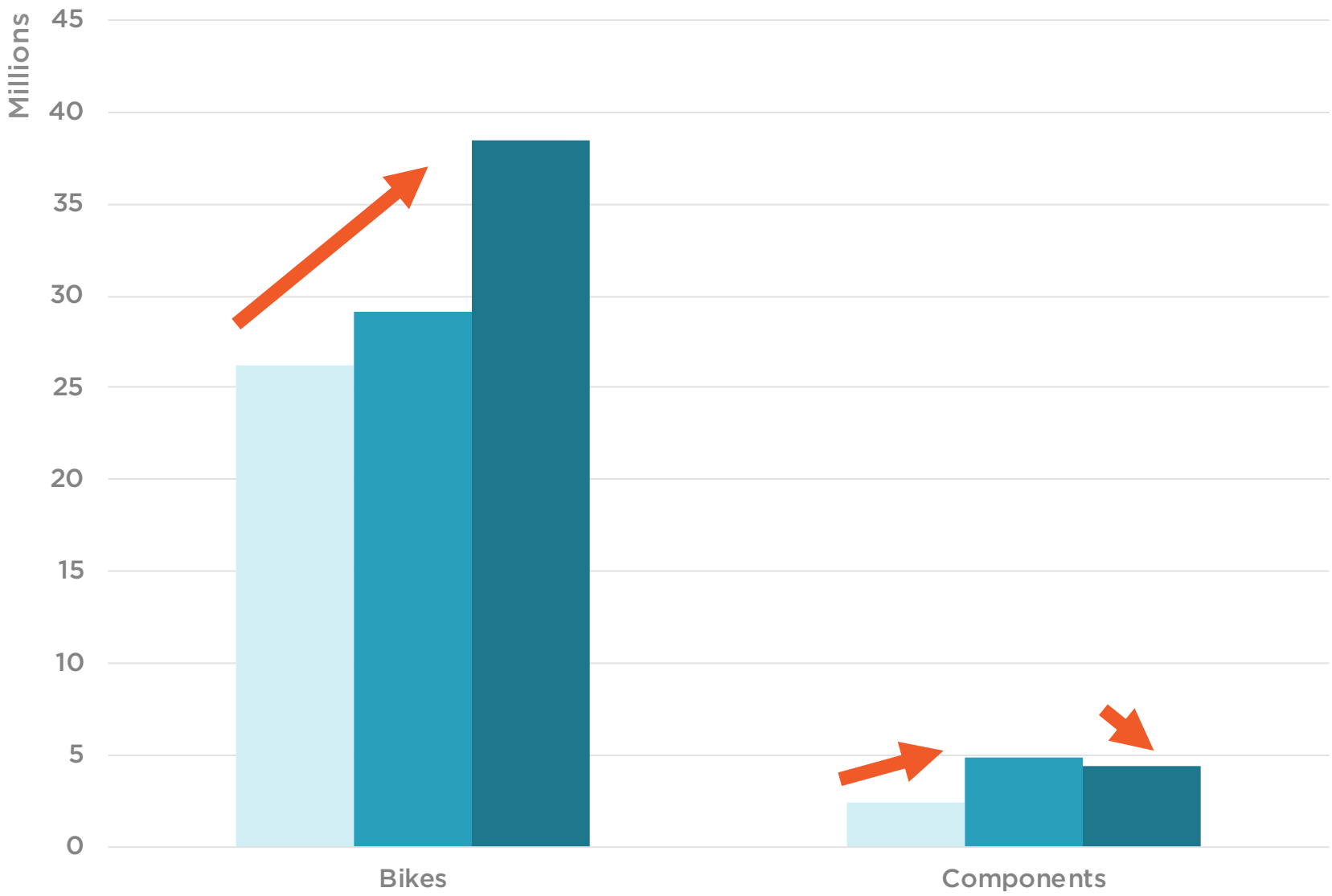
World-wide sales of bicycles and accessories

Wholesale sales to bicycle shops and sporting goods stores

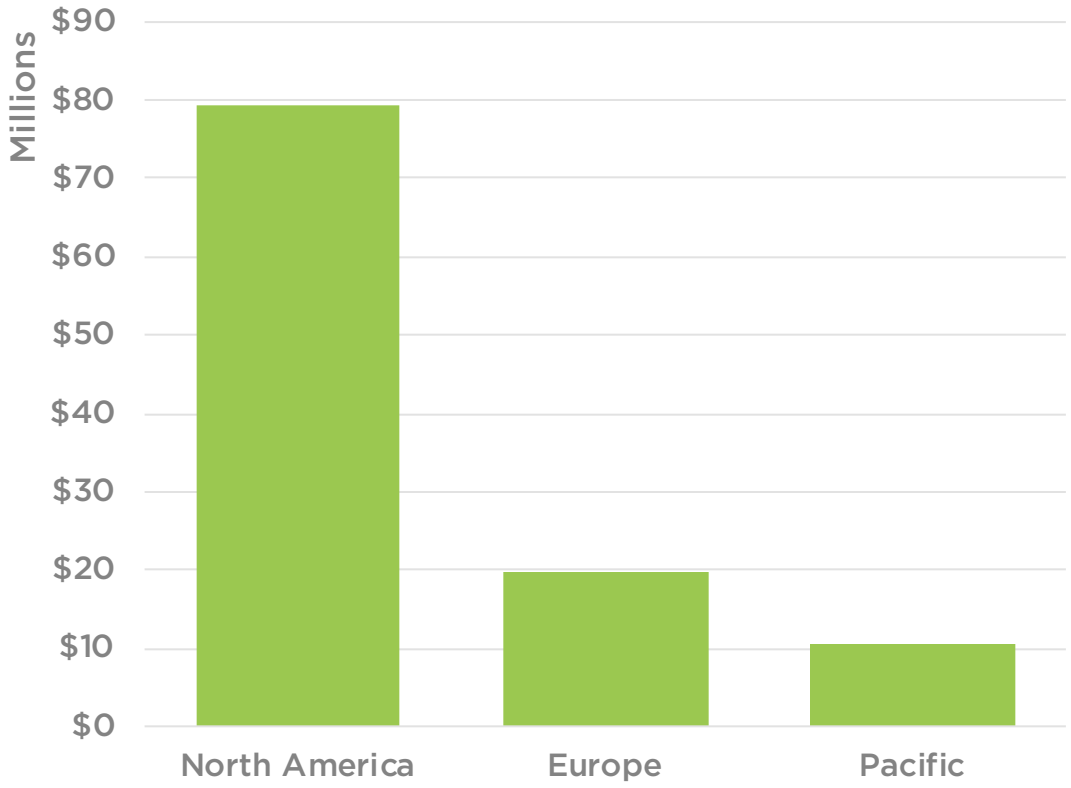
Online sales direct to consumers

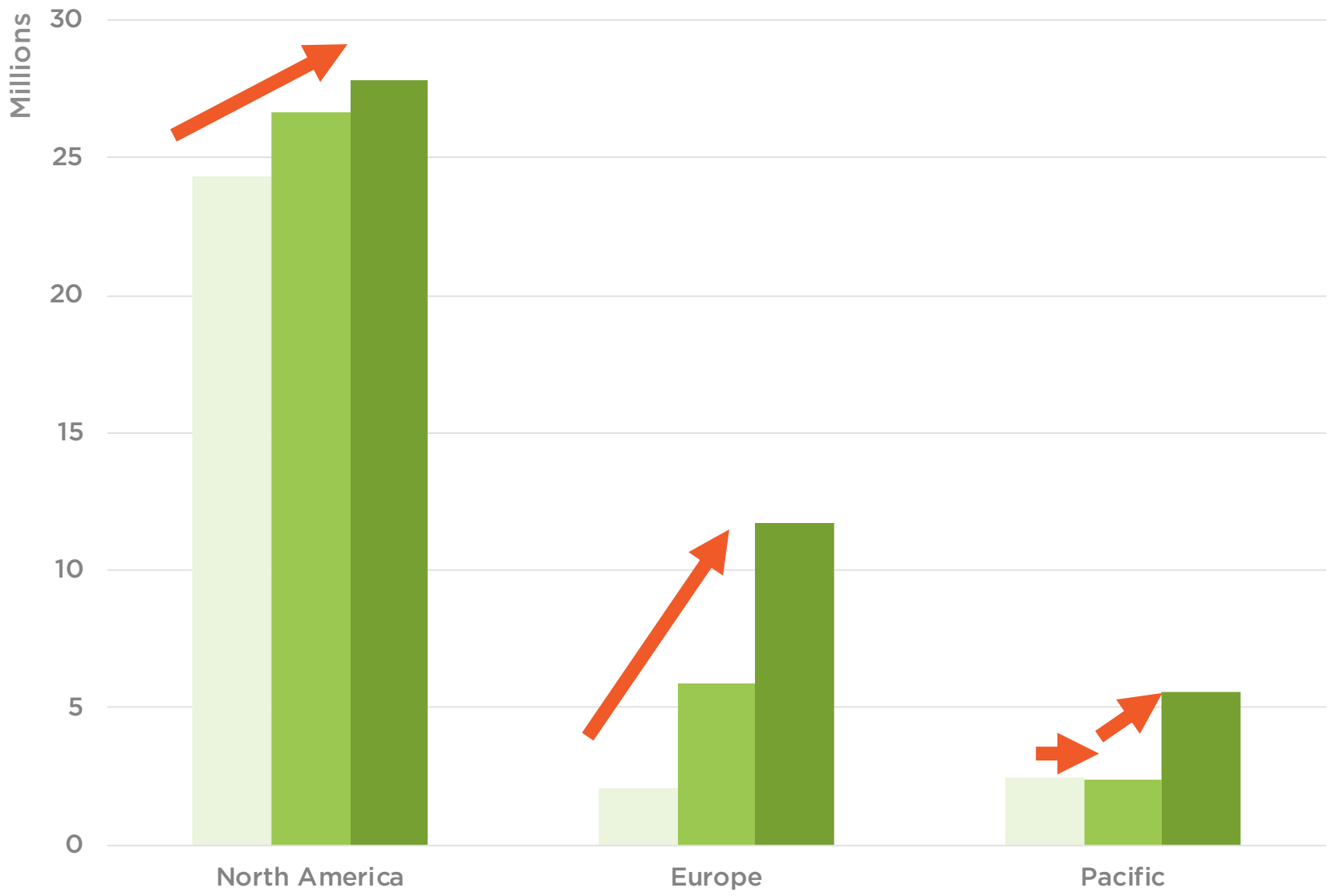
Category	Sales	Sales %
Bikes	$94,651,173	86.2%
Components	$11,802,593	10.7%
Clothing	$2,120,543	1.9%
Accessories	$1,272,073	1.2%
Total	**$109,846,381**	**100.0%**



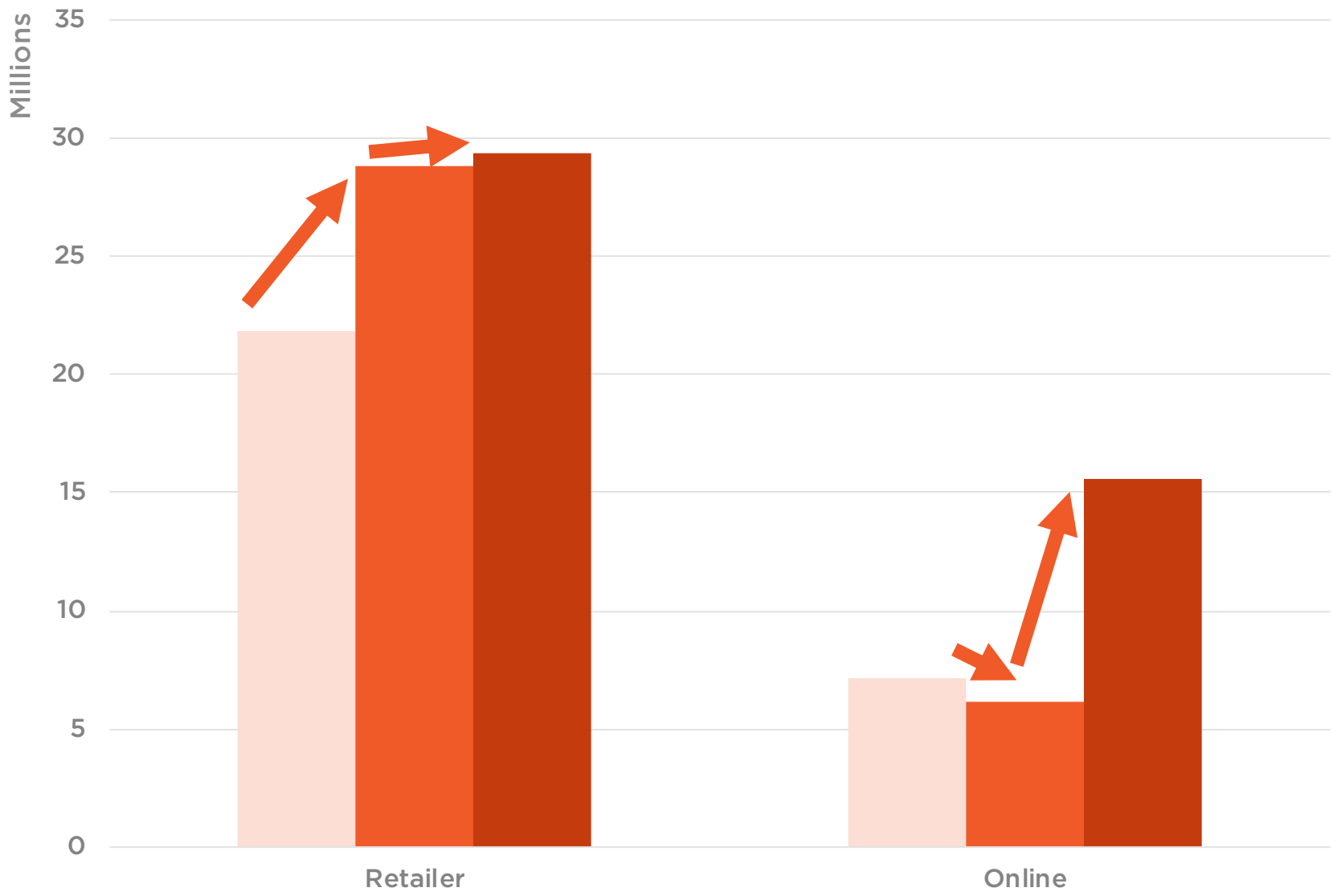

Millions
45
40
35
30
25
20
15
10
5
0
Bikes
Components

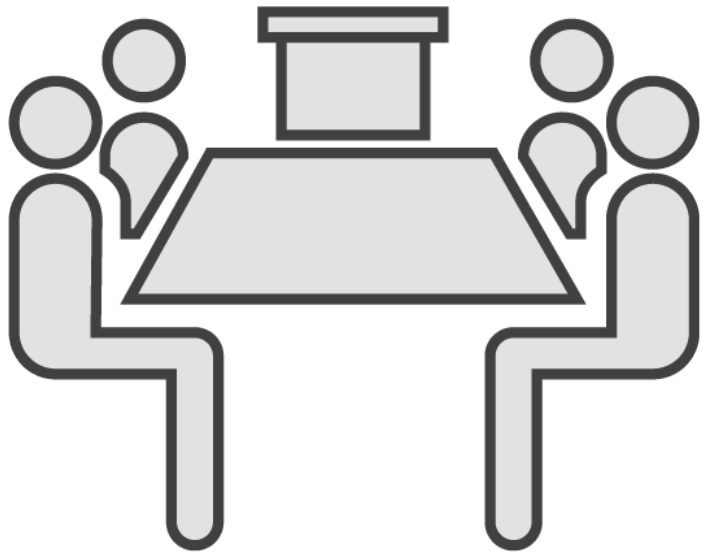
Territory Group	Sales	Sales %
North America	$79,353,361	72.2%
Europe	$19,837,684	18.1%
Pacific	$10,655,336	9.7%
Total	**$109,846,381**	**100.0%**




Millions
30
25
20
15
10
5
0
North America
Europe
Pacific

Channel Sales	Sales	Sales %
Retailer	$80,487,704	73.3%
Online	$29,358,677	26.7%
Total	$109,846,381	100.0%


Millions
35
30
25
20
15
10
5
0
Retailer
Online



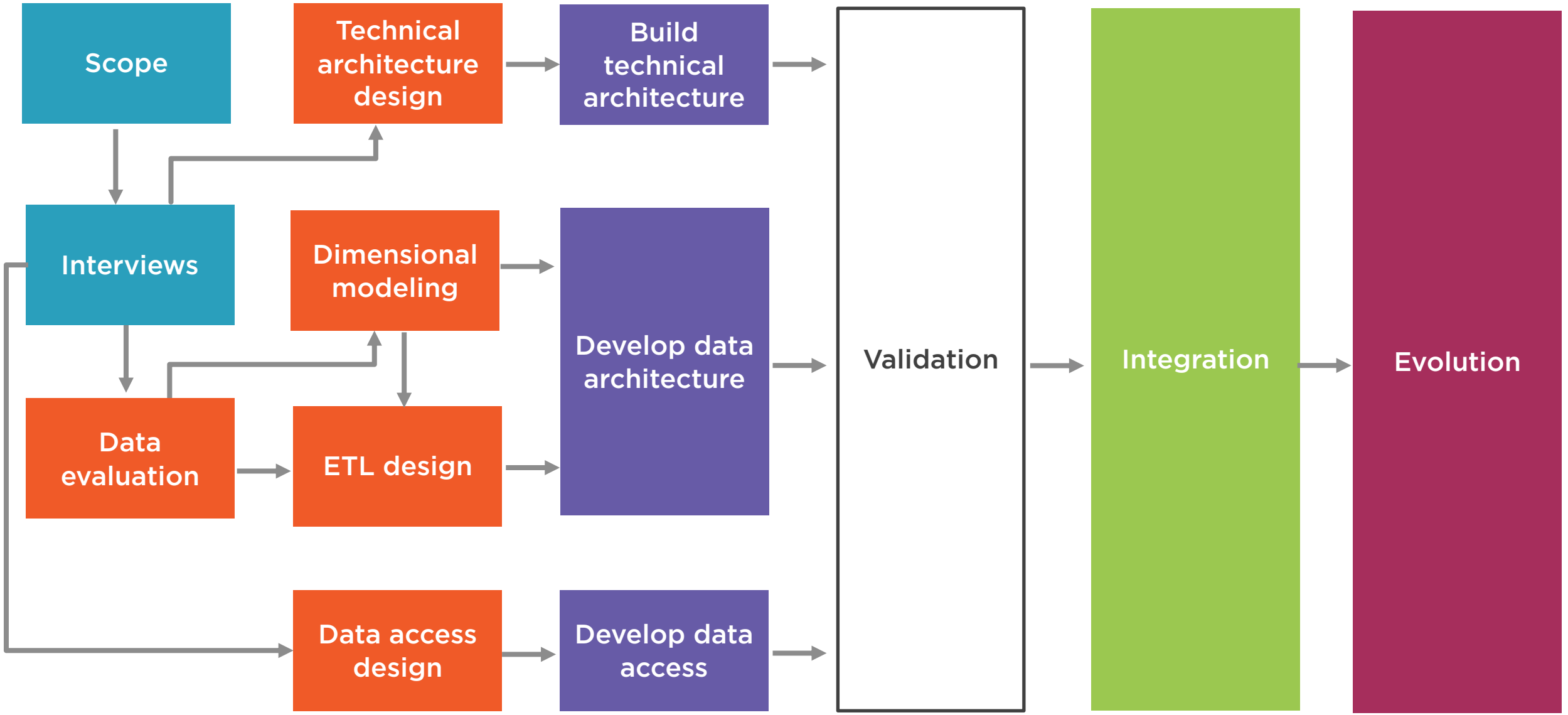
How can we see FASTER what's happening now?

How is business changing over time?

How can we analyze the data more effectively?

How can we discover what's driving the positive trends?

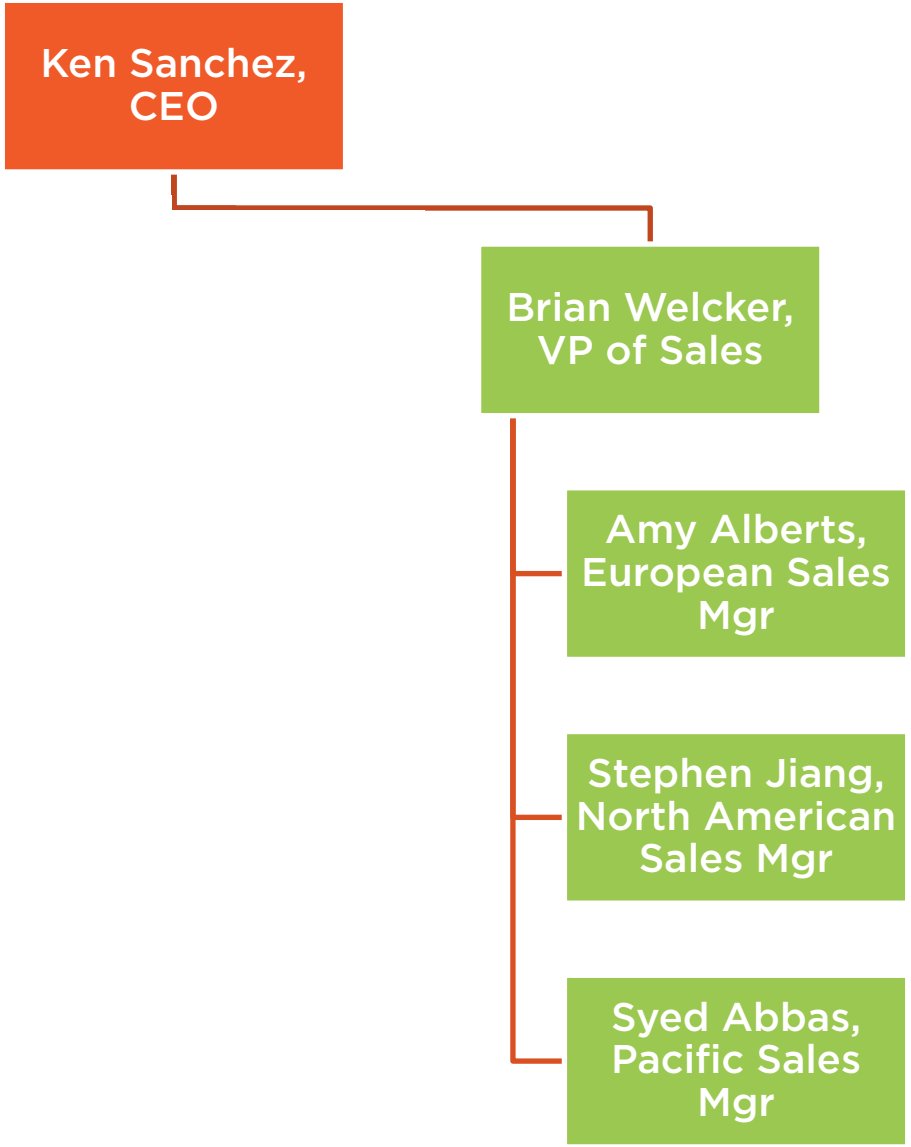
How can we stop or reverse negative trends?


Scope
Technical architecture design
Build technical architecture
Interviews
Dimensional modeling
Develop data architecture
Data evaluation
ETL design
Validation
Integration
Evolution
Data access design
Develop data access

Ken Sanchez,
CEO

No consistency in pulling data from transaction system

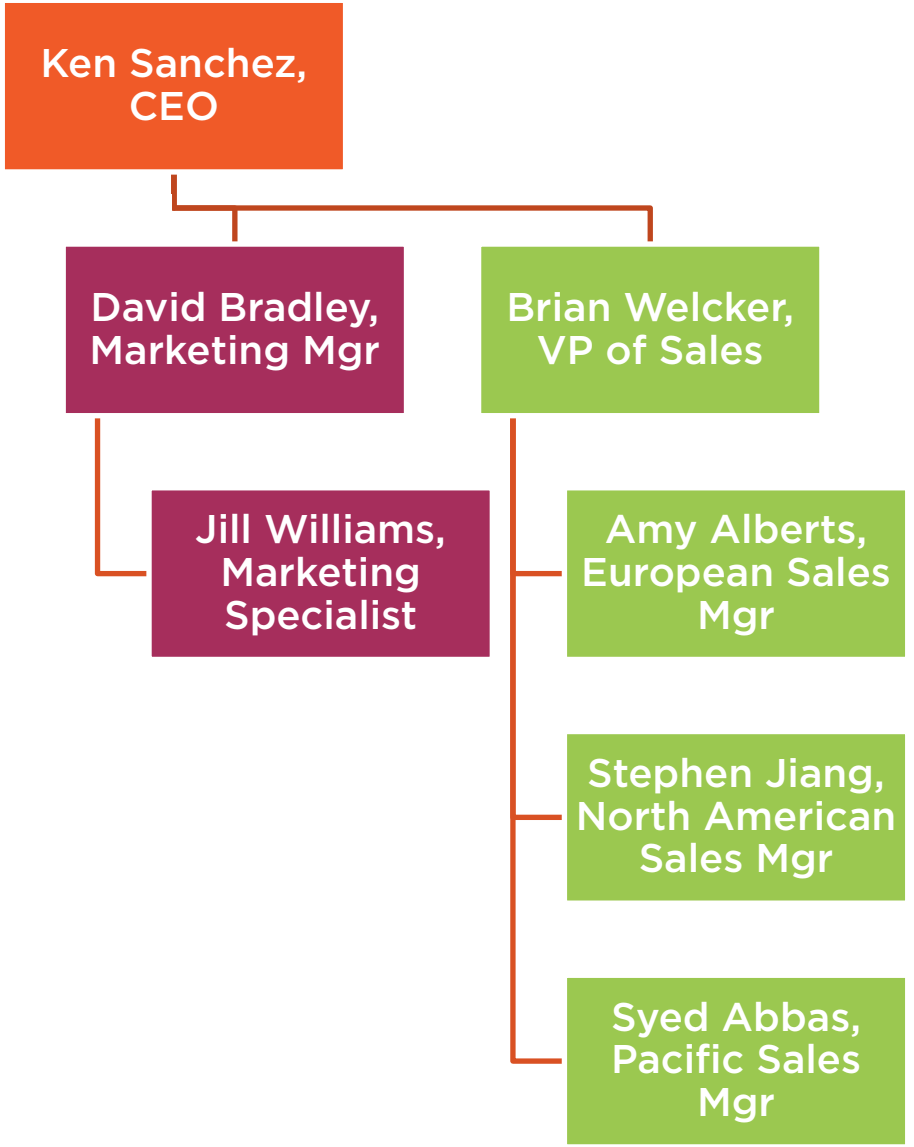
Difficulty making comparisons and spotting trends over time


Ken Sanchez, CEO
Brian Welcker, VP of Sales
Amy Alberts, European Sales Mgr
Stephen Jiang, North American Sales Mgr
Syed Abbas, Pacific Sales Mgr

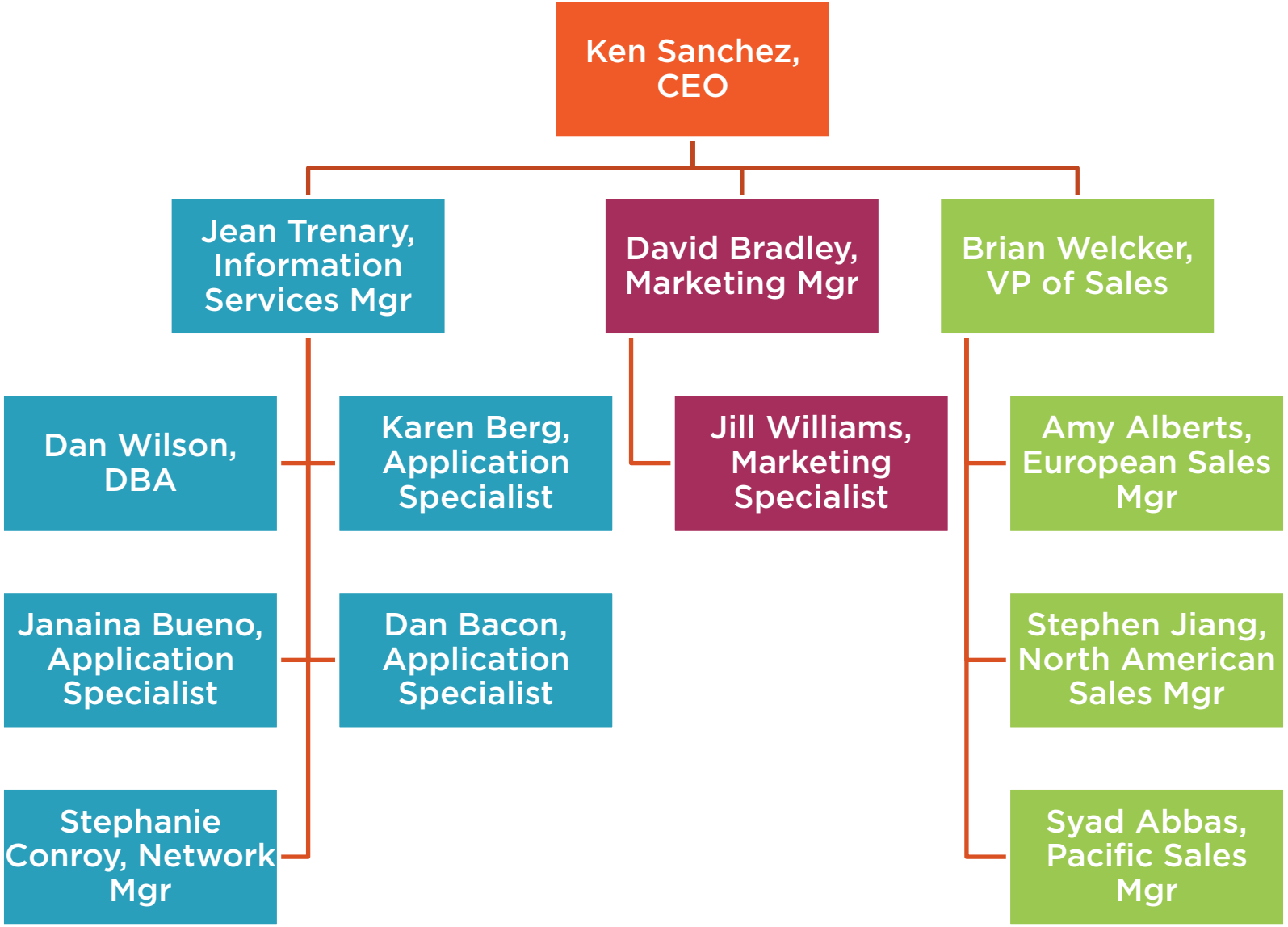
New report requests take too long

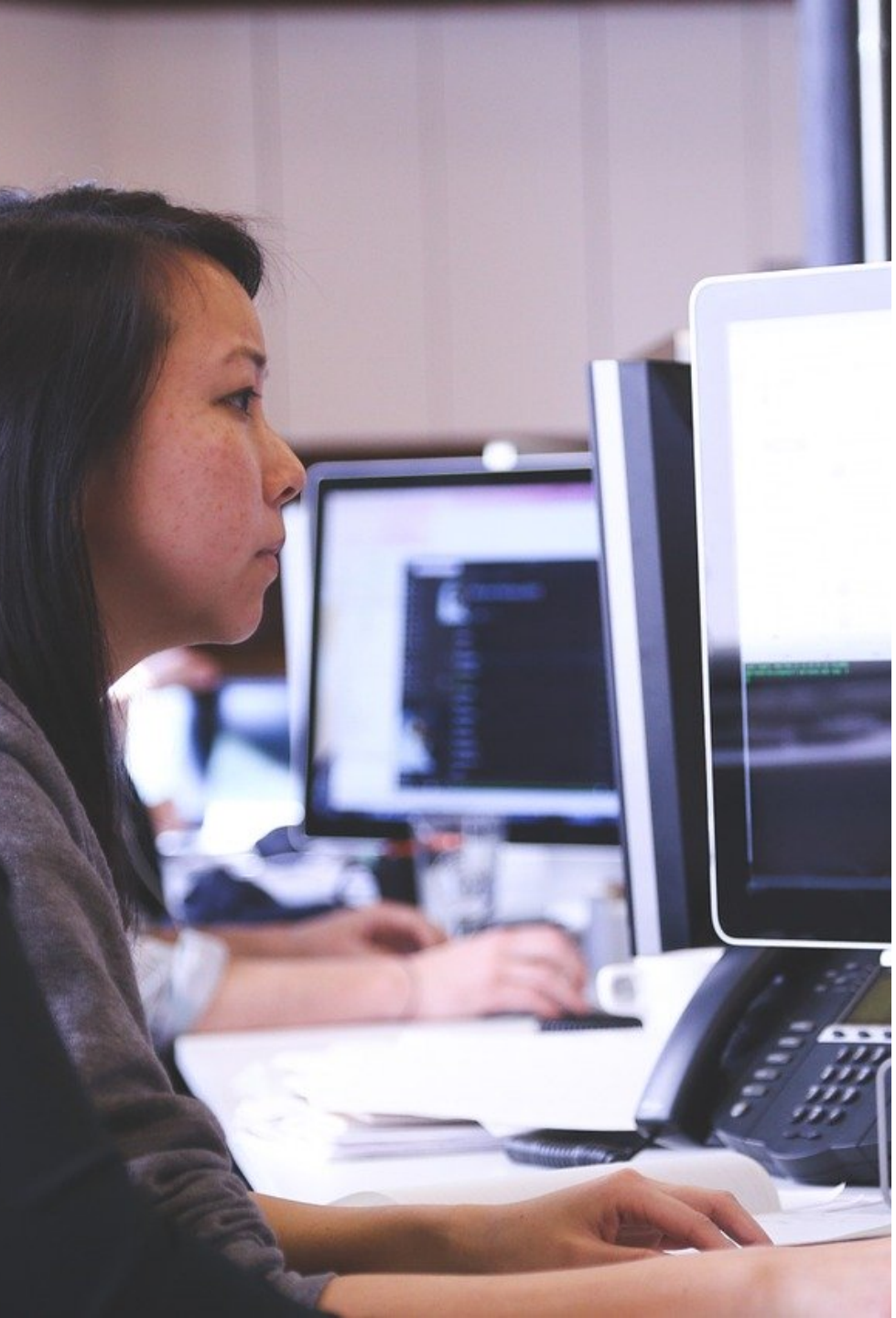
Existing reports run too slowly

Unable to respond quickly to sales trends or to changing customer behaviors


Ken Sanchez, CEO
David Bradley, Marketing Mgr
Brian Welcker, VP of Sales
Jill Williams, Marketing Specialist
Amy Alberts, European Sales Mgr
Stephen Jiang, North American Sales Mgr
Syed Abbas, Pacific Sales Mgr

Measuring effectiveness of campaigns requires too much manual processing


Ken Sanchez, CEO
Jean Trenary, Information Services Mgr
David Bradley, Marketing Mgr
Brian Welcker, VP of Sales
Dan Wilson, DBA
Karen Berg, Application Specialist
Jill Williams, Marketing Specialist
Amy Alberts, European Sales Mgr
Janaina Bueno, Application Specialist
Dan Bacon, Application Specialist
Stephen Jiang, North American Sales Mgr
Stephanie Conroy, Network Mgr
Syad Abbas, Pacific Sales Mgr



Many one-off requests for new reports come faster than we can manage

Queries against the transaction system are creating too much resource contention

Requirements

Interviews of management and subject matter experts

Summarization of findings

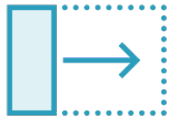
Prioritization of opportunities

Project AWDW Requirements Summary

Design and develop data warehouse to support sales and marketing

Executives, sales and marketing managers, and their staff will access information

Benefits include consistent reporting available on-demand quickly



Infrastructure will expand in the future to support more subject areas

Scope

Scope definition

Three years of historical sales data by customer, date, region, and product

Microsoft data platform: SQL Server, SSIS, SSAS, SSRS, Power BI

Scope exclusions

Manufacturing, finance, human resources, and customer support

Third-party data reporting and analysis tools

Business Requirements

Sales planning

Growth trends

Customer trends

Regional trends

Marketing analysis

Sales support

Auditing and Logging Requirements

Required	Not required
Proof of complete ETL workflow at job level only	Regulatory compliance
Enable tracing data back to source and ETL load operation	Fiduciary compliance
Confirmation that data was loaded as expected	Archival of source data
Initial retention period: 365 days	

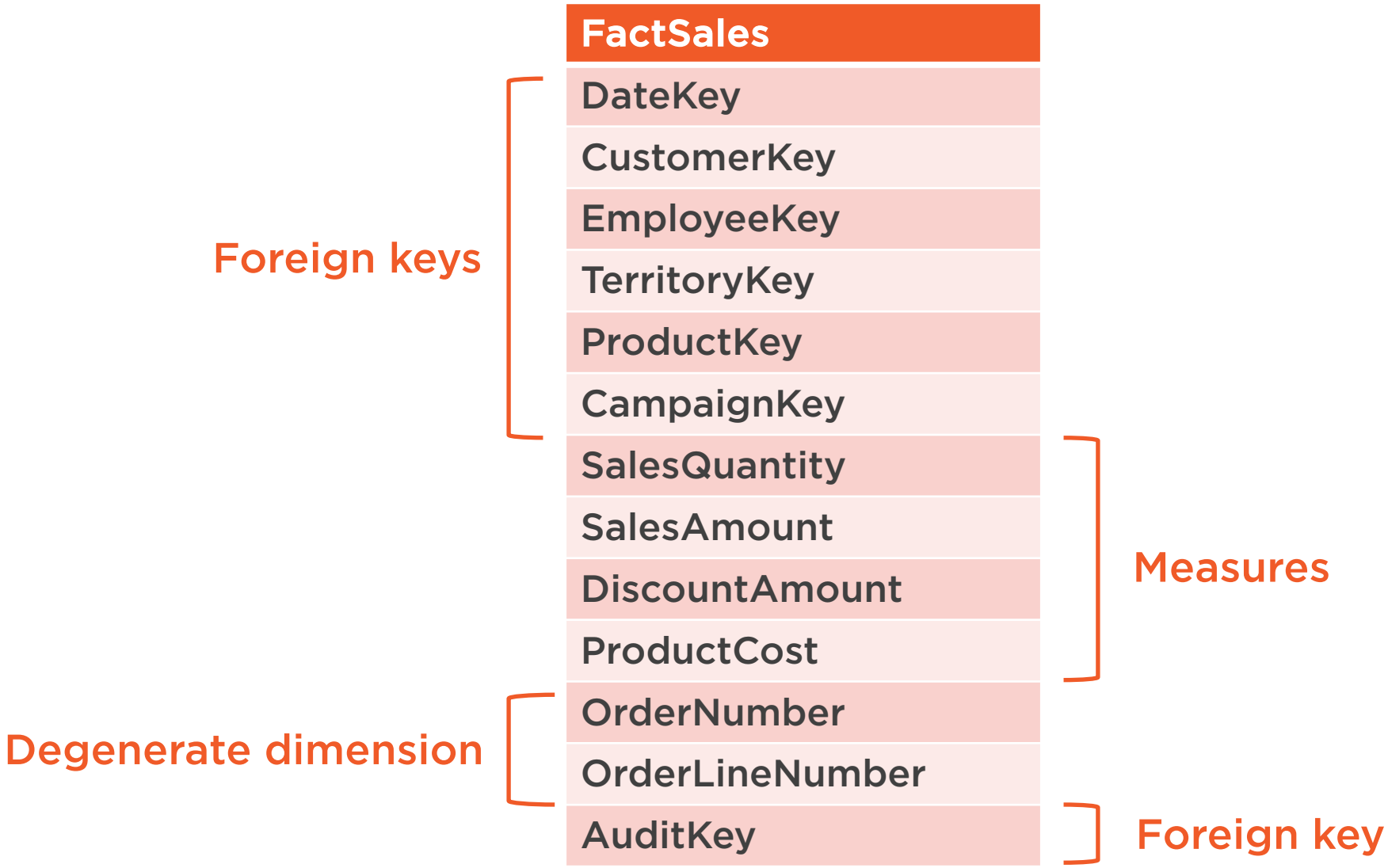
Data Latency Requirements

Nightly batch loads from source to start

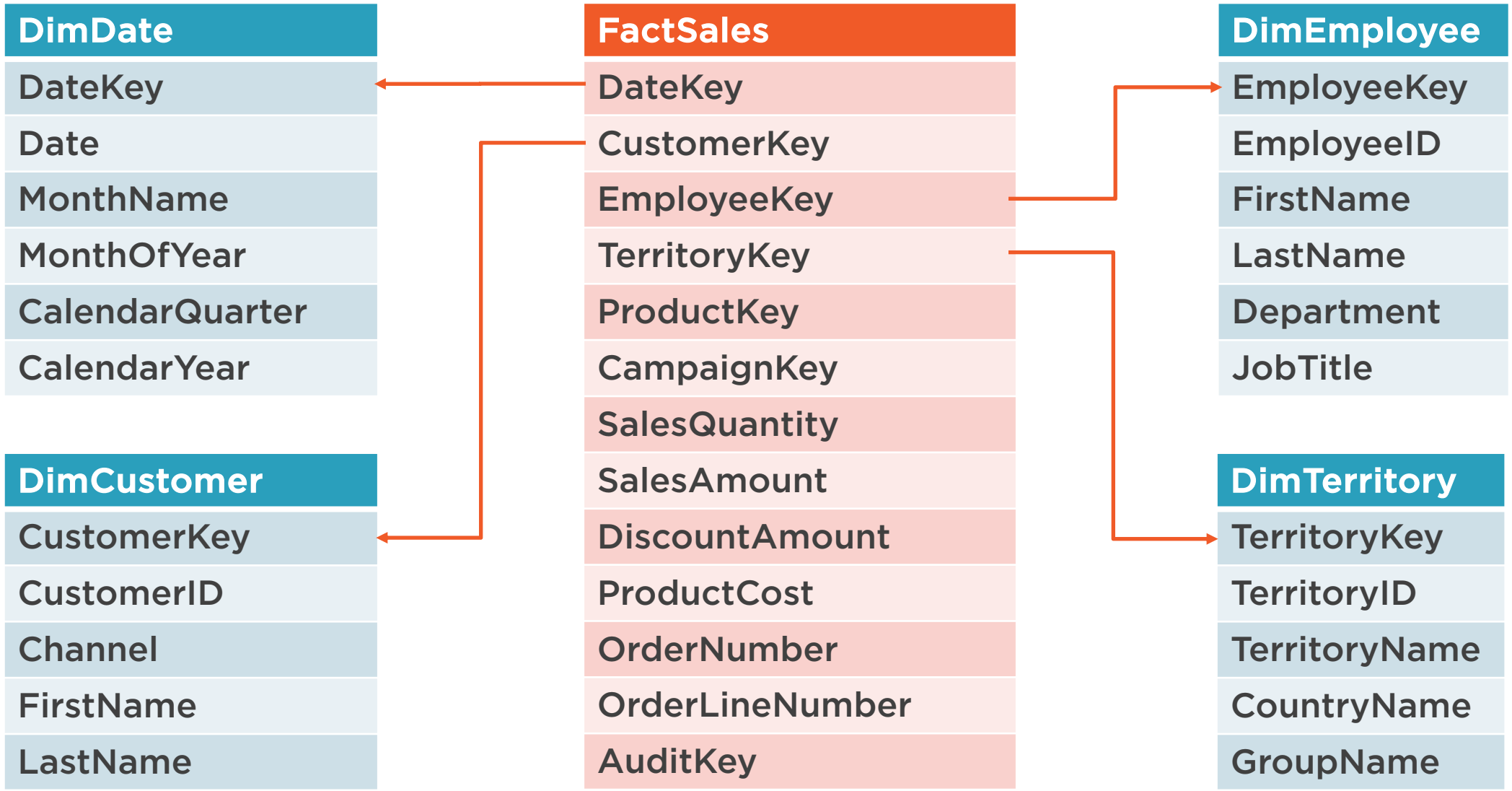
Eventually support four-hour window

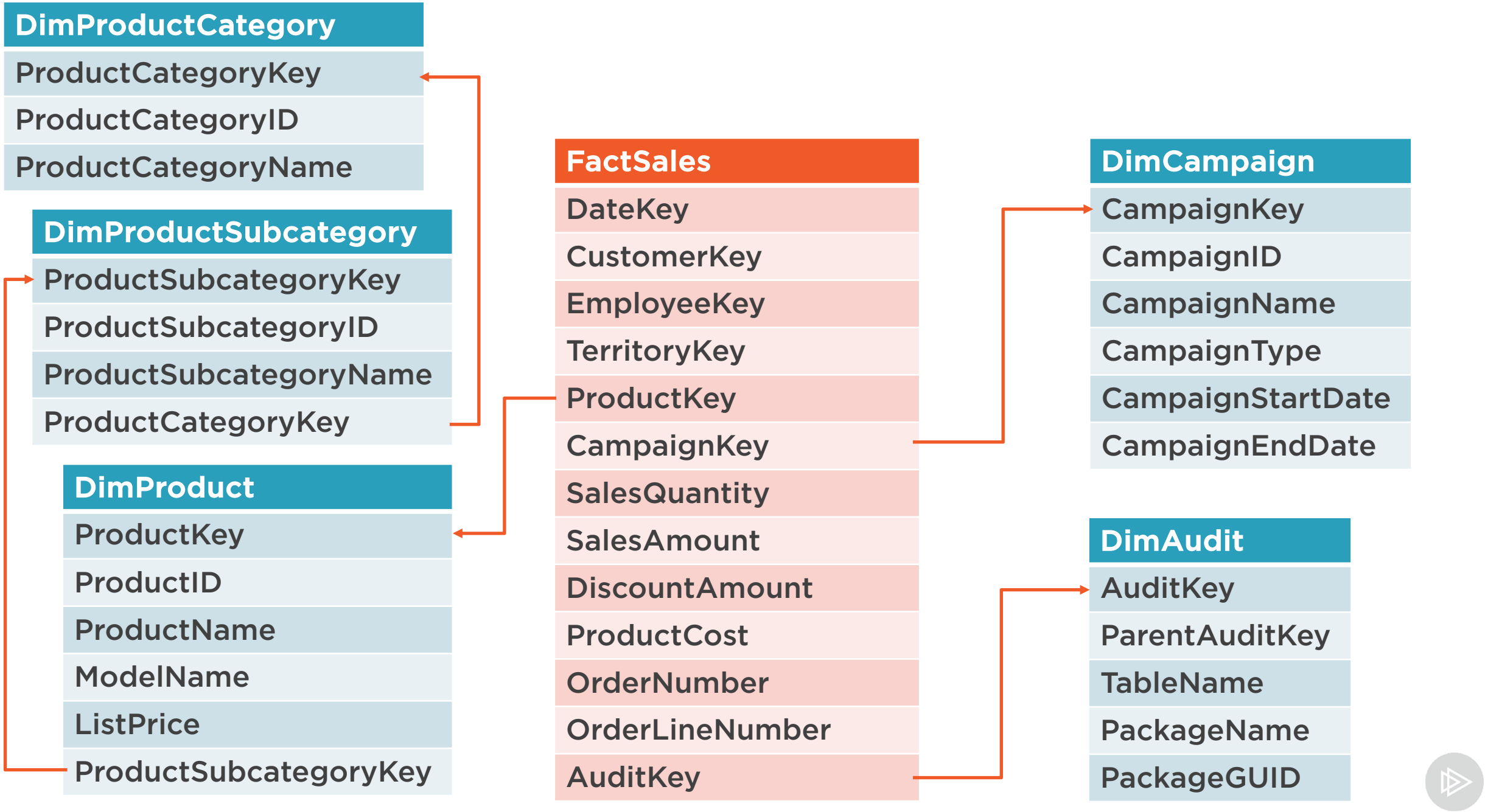
No need for real-time streaming data

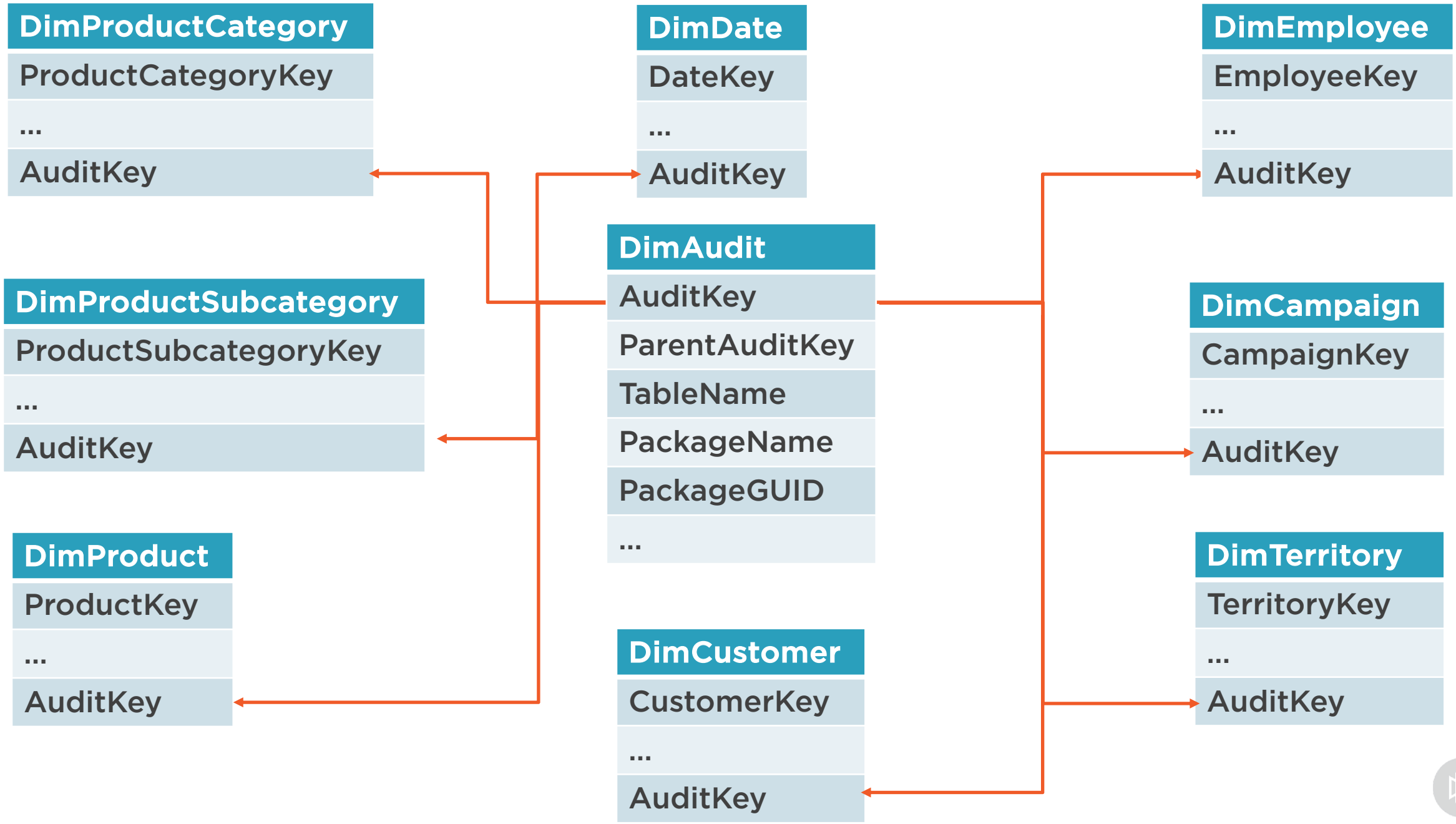
Dimensional Model

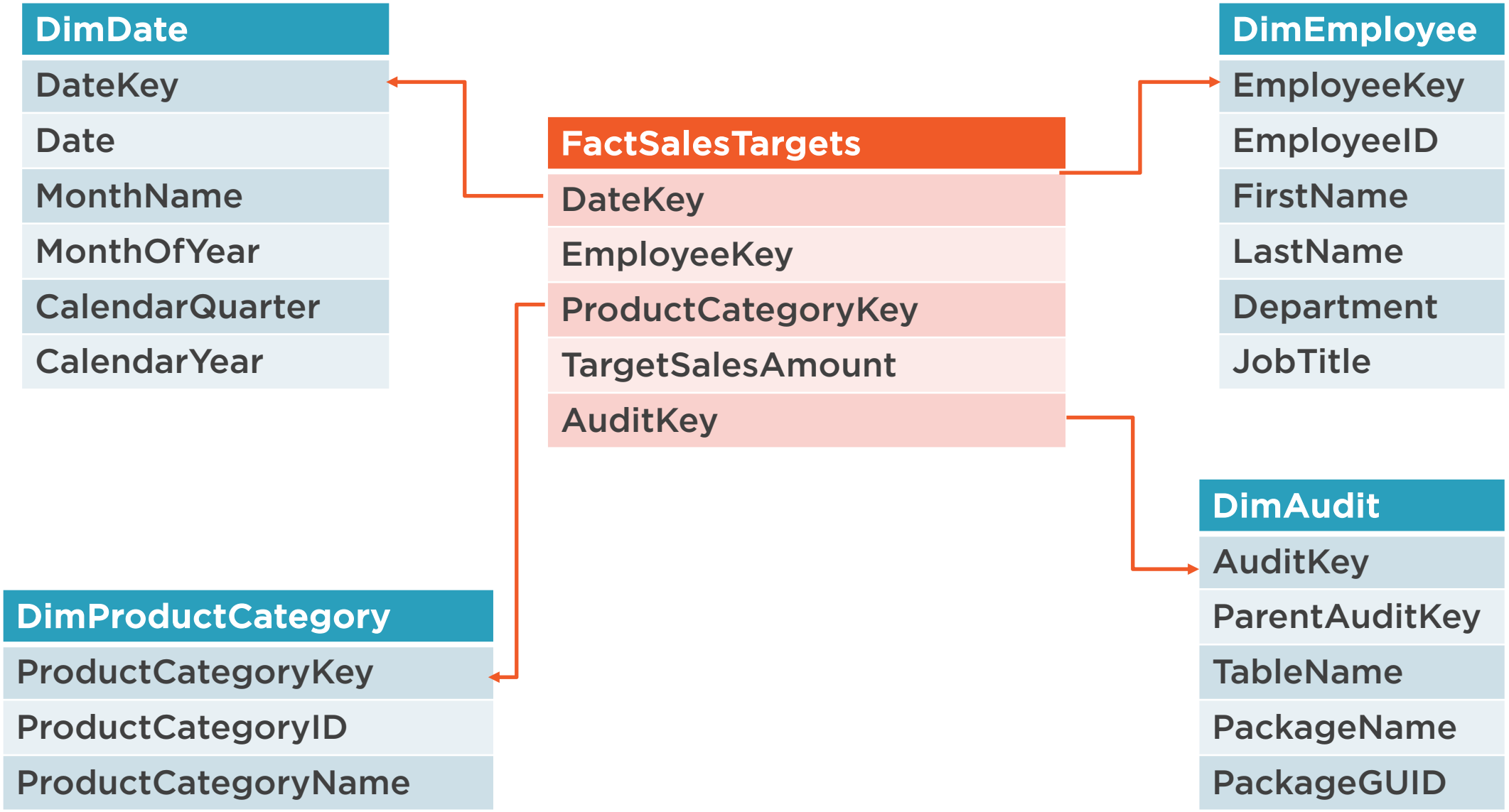


Dimensional Model




DimProductCategory
ProductCategoryKey
ProductCategoryID
ProductCategoryName
DimProductSubcategory
ProductSubcategoryKey
ProductSubcategoryID
ProductSubcategoryName
ProductCategoryKey
DimProduct
ProductKey
ProductID
ProductName
ModelName
ListPrice
ProductSubcategoryKey
FactSales
DateKey
CustomerKey
EmployeeKey
TerritoryKey
ProductKey
CampaignKey
SalesQuantity
SalesAmount
DiscountAmount
ProductCost
OrderNumber
OrderLineNumber
AuditKey
DimCampaign
CampaignKey
CampaignID
CampaignName
CampaignType
CampaignStartDate
CampaignEndDate
DimAudit
AuditKey
ParentAuditKey
TableName
PackageName
PackageGUID


DimProductCategory
ProductCategoryKey
...
AuditKey
DimDate
DateKey
...
AuditKey
DimEmployee
EmployeeKey
...
AuditKey
DimProductSubcategory
ProductSubcategoryKey
...
AuditKey
DimAudit
AuditKey
ParentAuditKey
TableName
PackageName
PackageGUID
...
DimCampaign
CampaignKey
...
AuditKey
DimProduct
ProductKey
...
AuditKey
DimTerritory
TerritoryKey
...
AuditKey
DimCustomer
CustomerKey
...
AuditKey


DimDate
DateKey
Date
MonthName
MonthOfYear
CalendarQuarter
CalendarYear
FactSalesTargets
DateKey
EmployeeKey
ProductCategoryKey
TargetSalesAmount
AuditKey
DimEmployee
EmployeeKey
EmployeeID
FirstName
LastName
Department
JobTitle
DimProductCategory
ProductCategoryKey
ProductCategoryID
ProductCategoryName
DimAudit
AuditKey
ParentAuditKey
TableName
PackageName
PackageGUID

Data Profiling

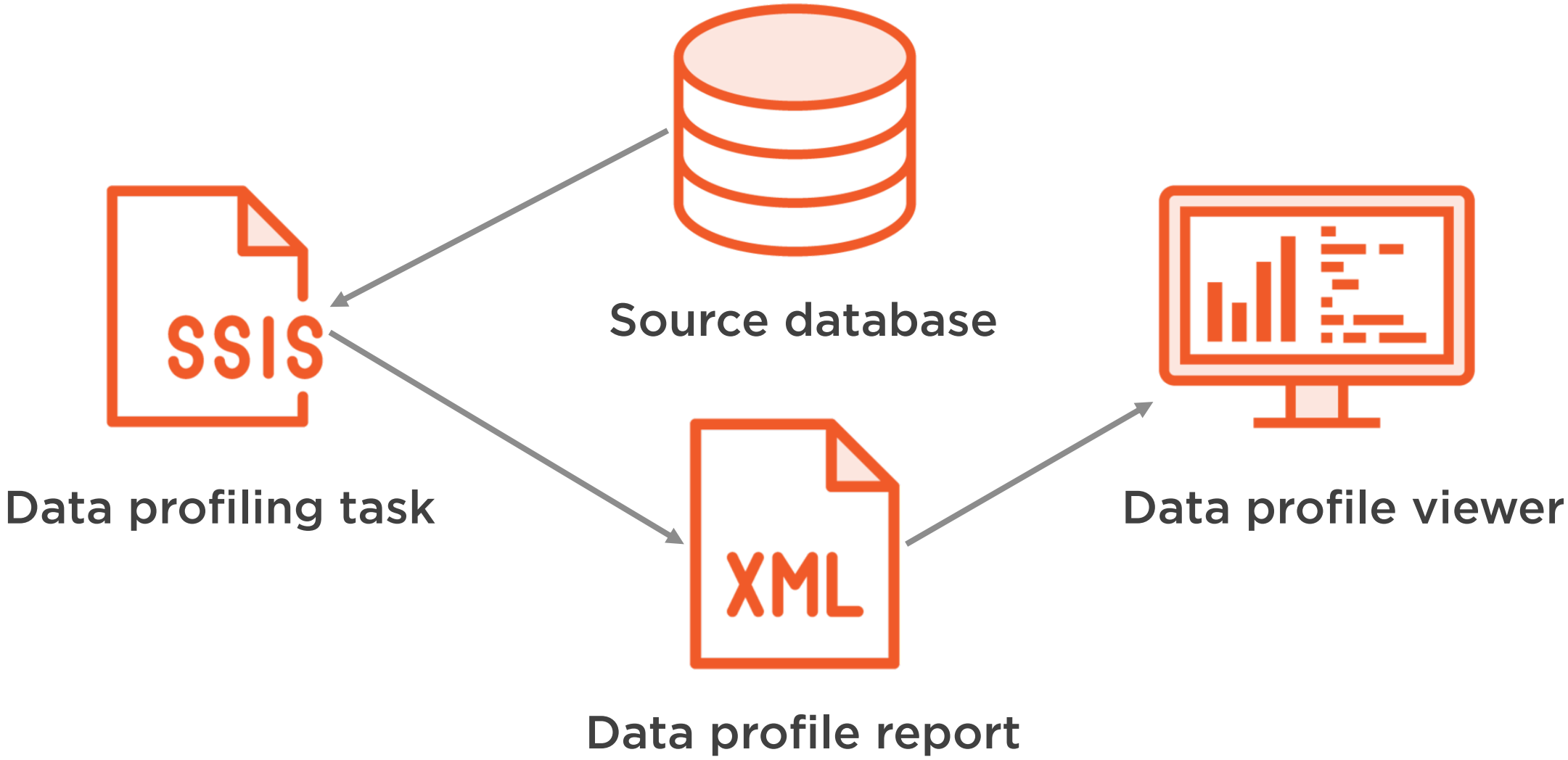
Data structure

Data availability

Potential problems in content

Transformation requirements

SSIS Data Profiling Task



SSIS Data Profiling Task

Candidate key: Search for a column containing unique values for each row in a table

Column statistics: Show minimum, maximum, mean, and standard deviation values for numeric and date/time data types

Column null ratio: Display percentage of null values in a column

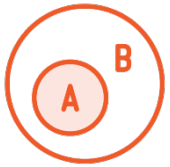
Column pattern: Report regular expressions that describe patterns in values of a column

SSIS Data Profiling Task

Column length distribution: Display distinct lengths of column values and percentage of rows for each length

Column value distribution: Show distinct values in a column and percentage of rows for each value

Functional dependency: Report extent to which values in one column depend on values in another column or discover violations



Value inclusion: Confirm that values in one column are found in another specified column

Additional Data Profiling Details

Table	Row Count	Column Count
Address	19,614	9
BusinessEntityAddress	19,614	5
CountryRegion	238	3
Department	16	4
EmailAddress	19,972	5
Employee	290	16
EmployeeDepartmentHistory	296	6
Person	19,972	13
PersonPhone	19,972	4
Product	504	25

Additional Data Profiling Details

Table	Column	Data Type	Max Length
Address	AddressID	int	10
	AddressLine1	nvarchar	60
	AddressLine2	nvarchar	60
	City	nvarchar	4
	StateProvinceID	int	10
	PostalCode	nvarchar	15
	SpatialLocation	geography	
	Rowguid	uniqueidentifier	16
	ModifiedDate	datetime	8

Summary

Project introduction

- Company overview
- Project plan
- Team

Requirements

- Summary
- Scope
- Business requirements
- Auditing and logging requirements
- Data latency requirements
- Dimensional model

Data profiling

Resources

Pluralsight courses

- Best Practices for Requirements Gathering
- Introduction to Data Warehousing and Business Intelligence
- Dimensional Modeling on the Microsoft SQL Server Platform

Data profiling task and viewer

- Integration Services Fundamentals
- https://tinyurl.com/dataprofilingtask