Northwind BI solution

Project Scope

|  |  |  |
| --- | --- | --- |
| High business value |  | * See sales with and without discounts. * See average discount by customers and geography. * See sales on holidays, weekdays, and weekends. * See sales by salespersons and employee geography. |
| Low business value |  |  |
|  | Low feasibility | High feasibility |

# Required BI Software

|  |  |
| --- | --- |
| Software | Rationale |
| SQL Server Database Engine | * The CEO wants a consistent view of all business data, and a centralized data warehouse in SQL Server would provide this. |
| SQL Server Integration Services | * The business data required for analysis and reporting is currently spread across a range of data sources. Integration Services will provide an ETL platform to populate and refresh the data warehouse. |
| SQL Server Master Data Services | * The CEO has complained about inconsistent data, which could potentially be caused by a lack of central data management for key business entities. |
| SQL Server Data Quality Services | * The specialists have difficulty ensuring that their analysis of sales data is accurate because of quality issues in the data. Data Quality Services could be used to cleanse records and improve the consistency, thereby reducing the inaccuracy of the sales analysis. |
| Power BI Report Server | * The database administrators team needs to receive email notifications about errors in the ETL process. * Data stewards need to be notified of errors and raw values in data cleansing processes. |

# BI Topology

## Single BI server



* SQL Server DB Engine
  + Data Warehouse
  + Staging database
  + SSIS catalog
  + Reporting Services catalog
  + DQS catalog
  + MDS catalog
* SQL Server Integration Services
* Power BI Report Server
* Data Quality Services
* Master Data Services

SRVBI



CLIENT PC

* Microsoft Excel
  + Power Query
  + Power Pivot
  + Power View
* Report Builder
* Power BI Desktop
* Microsoft Edge

Notes:

* Minimal server hardware and software license requirements, but the server would require significant memory, CPU, and disk resources for all but the most lightweight BI workloads.
* The range of different workload types on the server would make it difficult to specify and configure hardware resources appropriately.
* The server could be clustered to provide high availability.

## Dedicated report server



* SQL Server DB Engine
  + Data Warehouse
  + Staging database
  + SSIS catalog
  + DQS catalog
  + MDS catalog
* SQL Server Integration Services
* Data Quality Services
* Master Data Services

SRVDWH



CLIENT PC

* Microsoft Excel
  + Power Query
  + Power Pivot
  + Power View
* Report Builder
* Power BI Desktop
* Microsoft Edge



* SQL Server DB Engine
  + Reporting Services catalog
* Power BI Report Server

SRVPBIRS

Notes:

* The reporting and DWH workloads are full separated.
* Using a dedicated data warehouse server makes it easier to manage the ETL and data warehouse workloads.
* Additional report servers could be added to scale out reporting.
* One or both servers could be clustered to provide high availability.

## Dedicated ETL server



* SQL Server DB Engine
  + Data Warehouse
  + Staging database

SRVDWH



CLIENT PC

* Microsoft Excel
  + Power Query
  + Power Pivot
  + Power View
* Report Builder
* Power BI Desktop
* Microsoft Edge



* SQL Server DB Engine
  + Reporting Services catalog
* Power BI Report Server

SRVPBIRS



* SQL Server DB Engine
  + SSIS catalog
  + DQS catalog
  + MDS catalog
* SQL Server Integration Services
* Data Quality Services
* Master Data Services

SRVETL

Notes:

* The DWH, reporting and ETL workloads are full separated.
* The staging database could be hosted on the ETL server or data warehouse server depending on workloads.
* SRVDWH, SRVPBIRS, and SRVETL could be clustered to provide high availability.
* Additional report servers and ETL could be added to scale out.

## Dedicated master data server



* SQL Server DB Engine
  + Data Warehouse
  + Staging database

SRVDWH



CLIENT PC

* Microsoft Excel
  + Power Query
  + Power Pivot
  + Power View
* Report Builder
* Power BI Desktop
* Microsoft Edge



* SQL Server DB Engine
  + Reporting Services catalog
* Power BI Report Server

SRVPBIRS



* SQL Server DB Engine
  + DQS catalog
  + MDS catalog
* Data Quality Services
* Master Data Services

SRVMDM



* SQL Server DB Engine
  + SSIS catalog
* SQL Server Integration Services

SRVETL

Notes:

* The DWH, reporting ETL and MDM workloads are full separated.
* The staging database could be hosted on the ETL server or data warehouse server depending on workloads.
* Servers could be clustered to provide high availability.
* Additional report servers and ETL could be added to scale out.

## Distributed report server



* SQL Server DB Engine
  + Data Warehouse
  + Staging database

SRVDWH



CLIENT PC

* Microsoft Excel
  + Power Query
  + Power Pivot
  + Power View
* Report Builder
* Power BI Desktop
* Microsoft Edge



* Power BI Report Server

SRVPBIRS



* SQL Server DB Engine
  + DQS catalog
  + MDS catalog
* Data Quality Services
* Master Data Services

SRVMDM



* SQL Server DB Engine
  + SSIS catalog
* SQL Server Integration Services

SRVETL



* SQL Server DB Engine
  + Reporting Services catalog

SRVPBIRSDB

Notes:

* The DWH, reporting ETL and MDM workloads are full separated.
* The staging database could be hosted on the ETL server or data warehouse server depending on workloads.
* The report server uses a separate database server to host the report catalog.
* Servers could be clustered to provide high availability.
* Additional report servers and ETL could be added to scale out.

# Planning Data Warehouse Hardware

[DW Hardware Spec.xlsx](https://dev.azure.com/zinykov/20767/_git/Northwind_BI_Solution?path=/Docs/DW%20Hardware%20Spec.xlsx&version=GBmaster&_a=contents)

# Identify business processes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Business processes** | **Dimensions** | | | |
| Date | Product | Customer | Employee |
| Orders | X | X | X | X |

Notes:

* There is no interest in shipment geography, so there is no reason to create dedicated geography dimension. Employee and Customer geography will be implemented as attributes.
* In Date dimension will be used opensource project [Производственный календарь в XML (xmlcalendar.ru)](http://xmlcalendar.ru/) for getting information about production calendar.

# Order dimensional model

Date

(Order, Required, Shipped)

Year

Quarter

Month

Day

Year

Quarter

Mon

ISO Week

Weekday

Year

Quarter

Month

Week

Weekday

Customer

Country

City

Customer

Contact name

Contact title

Phone

Fax

DayOfQuarterNumber

DayOfYearNumber

StartOfYear

EndOfYear

YearQuarter

StartOfQuarter

EndOfQuarter

YearMonth

StartOfMonth

EndOfMonth

StartOfWeek

EndOfWeek

Holiday

Workday type

Workday hours

Product

Category

Product

Employee

Country

City

Employee

Title

Title of courtesy

# Data source profiling

## Date

1. <!--

2. year - год на который сформирован календарь

3. lang - язык на котором представлены названия праздников

4. date - дата формирования xml-календаря в формате ГГГГ.ММ.ДД

5. -->

6. <calendar year="2014" lang="ru" date="2014.01.01">

7. <!--

8. id - идентификатор праздника

9. title - название праздника

10. -->

11. <holidays>

12. <holiday id="1" title="Новогодние каникулы" />

13. <holiday id="2" title="Рождество Христово" />

14. </holidays>

15. <!--

16. d (day) - день (формат ММ.ДД)

17. t (type) - тип дня: 1 - выходной день, 2 - рабочий и сокращенный (может быть использован для любого дня недели), 3 - рабочий день (суббота/воскресенье)

18. h (holiday) - номер праздника (ссылка на атрибут id тэга holiday)

19. f (from) - дата с которой был перенесен выходной день

20. суббота и воскресенье считаются выходными, если нет тегов day с атрибутом t=2 и t=3 за этот день

21. -->

22. <days>

23. <day d="01.01" t="1" h="1" />

24. <day d="01.02" t="1" h="1" />

25. <day d="01.03" t="1" h="1" />

26. <day d="02.22" t="1" f="01.03" />

27. </days>

28. </calendar>

## Customer

Table 1. Northwind (упрощённая).dbo.Customers.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CustomerID | CompanyName | ContactName | ContactTitle | City | Country | Phone | Fax |
| ALFKI | Alfreds Futterkiste | Maria Anders | Sales Representative | Berlin | Germany | 030-0074321 | 030-0076545 |
| ANATR | Ana Trujillo Emparedados y helados | Ana Trujillo | Owner | Mexico D.F. | Mexico | (5) 555-4729 | (5) 555-3745 |
|  | SCD1 | SCD1 | SCD1 | SCD2 | SCD2 | SCD1 | SCD1 |
| Business key | Drill-thought detail | Drill-thought detail | Slicer | Hierarchy | | Drill-thought detail | Drill-thought detail |

## Product

Table 2. Northwind (упрощённая).dbo.Products.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ProductID | ProductName | SupplierID | CategoryID | UnitPrice |
| 1 | Chai | 1 | 1 | 18,00 |
| 2 | Chang | 1 | 1 | 19,00 |
|  | SCD1 |  |  |  |
| Business key | Drill-thought detail | Foreign key | Foreign key (Hierarchy) | Drill-thought detail |

Table 3. Northwind (упрощённая).dbo.Categories.

|  |  |  |
| --- | --- | --- |
| CategoryID | CategoryName | Description |
| 1 | Beverages | Soft drinks, coffees, teas, beers, and ales |
| 2 | Condiments | Sweet and savory sauces, relishes, spreads, and seasonings |
|  | SCD1 |  |
| Business key | Drill-thought detail | Drill-thought detail |

## Employee

Table 4. Northwind (упрощённая).dbo.Employees.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| EmployeeID | LastName | FirstName | Title | TitleOfCourtesy | City | Country |
| 1 | Davolio | Nancy | Sales Representative | Ms. | Seattle | USA |
| 2 | Fuller | Andrew | Vice President, Sales | Dr. | Tacoma | USA |
|  | SCD1 | SCD1 | SCD2 | SCD1 | SCD2 | SCD2 |
| Business key | Drill-thought detail | Drill-thought detail | Drill-thought detail | Drill-thought detail | Hierarchy | |

## Order

Table 5. Northwind (упрощённая).dbo.Orders

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| OrderID | CustomerID | EmployeeID | OrderDate | RequiredDate | ShippedDate | ShipCity | ShipCountry |
| 10248 | VINET | 5 | 1996-07-04 00:00:00.000 | 1996-08-01 00:00:00.000 | 1996-07-16 00:00:00.000 | Reims | France |
| 10249 | TOMSP | 6 | 1996-07-05 00:00:00.000 | 1996-08-16 00:00:00.000 | 1996-07-10 00:00:00.000 | Münster | Germany |
| Degenerate dimension | Dimension key | Dimension key | Dimension key | Dimension key | Dimension key |  |  |

Table 6. Northwind (упрощённая).dbo.Order Details.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| OrderID | ProductID | UnitPrice | Quantity | Discount |
| 10248 | 11 | 14,00 | 12 | 0 |
| 10248 | 42 | 9,80 | 10 | 0 |
| Degenerate dimension | Dimension key | Measure | Measure | Measure |

# DWH logical schema

## Dimension

### Customer

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| COLUMN NAME | COLUMN TYPE | SCD TYPE | DATA TYPE | NULL |
| CustomerKey | Surrogate key |  | INT | NOT |
| CustomerAlterKey | Business key |  | NVARCHAR(5) | NOT |
| Customer | Drill-thought detail | 1 | NVARCHAR(50) | NOT |
| ContactName | Drill-thought detail | 1 | NVARCHAR(50) |  |
| ContactTitle | Slicer | 1 | NVARCHAR(50) |  |
| Country | Hierarchy | 2 | NVARCHAR(25) |  |
| City | Hierarchy | 2 | NVARCHAR(25) |  |
| Phone | Drill-thought detail | 1 | NVARCHAR(30) |  |
| Fax | Drill-thought detail | 1 | NVARCHAR(30) |  |
| StartDate | Metadata |  | DATETIME2 | NOT |
| EndDate | Metadata |  | DATETIME2 |  |
| Current | Metadata |  | BIT | NOT |
| LineageKey | Metadata |  | INT | NOT |

### Date

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| COLUMN NAME | COLUMN TYPE | SCD TYPE | DATA TYPE | NULL |
| DateKey | Surrogate key |  | INT | NOT |
| AlterDateKey | Business key |  | DATE | NOT |
| DayOfMonth | Hierarchy |  | TINYINT | NOT |
| DayOfWeek | Hierarchy |  | NVARCHAR(5) | NOT |
| DayOfWeekNumber | Sort by column |  | TINYINT | NOT |
| DayOfQuarterNumber | Drill-thought detail |  | TINYINT | NOT |
| DayOfYearNumber | Drill-thought detail |  | SMALLINT | NOT |
| Year | Hierarchy |  | SMALLINT | NOT |
| StartOfYear | Drill-thought detail |  | DATE | NOT |
| EndOfYear | Drill-thought detail |  | DATE | NOT |
| Quarter | Hierarchy |  | NVARCHAR(5) | NOT |
| YearQuarter | Hierarchy |  | NVARCHAR(10) | NOT |
| YearQuarterNumber | Sort by column |  | INT | NOT |
| StartOfQuarter | Drill-thought detail |  | DATE | NOT |
| EndOfQuarter | Drill-thought detail |  | DATE | NOT |
| Month | Hierarchy |  | NVARCHAR(10) | NOT |
| Mon | Hierarchy |  | NVARCHAR(5) | NOT |
| MonthNumber | Sort by column |  | TINYINT | NOT |
| YearMonth | Hierarchy |  | NVARCHAR(10) | NOT |
| StartOfMonth | Drill-thought detail |  | DATE | NOT |
| EndOfMonth | Drill-thought detail |  | DATE | NOT |
| Week | Hierarchy |  | NVARCHAR(50) | NOT |
| IOSWeekNumber | Hierarchy |  | TINYINT | NOT |
| StartOfWeek | Drill-thought detail |  | DATE | NOT |
| EndOfWeek | Drill-thought detail |  | DATE | NOT |
| Holiday | Drill-thought detail |  | NVARCHAR(100) | NOT |
| WorkDayType | Slicer |  | NVARCHAR(25) | NOT |
| WorkDayHours | Measure |  | TINYINT | NOT |

### Employee

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| COLUMN NAME | COLUMN TYPE | SCD TYPE | DATA TYPE | NULL |
| EmployeeKey | Surrogate key |  | INT | NOT |
| EmployeeAlterKey | Business key |  | INT | NOT |
| Employee | Drill-thought detail | 1 | NVARCHAR(35) | NOT |
| Title | Drill-thought detail | 1 | NVARCHAR(30) |  |
| TitleOfCourtesy | Drill-thought detail | 1 | NVARCHAR(10) |  |
| City | Hierarchy | 2 | NVARCHAR(25) |  |
| Country | Hierarchy | 2 | NVARCHAR(25) |  |
| StartDate | Metadata |  | DATETIME2 | NOT |
| EndDate | Metadata |  | DATETIME2 |  |
| Current | Metadata |  | BIT | NOT |
| LineageKey | Metadata |  | INT | NOT |

### Product

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| COLUMN NAME | COLUMN TYPE | SCD TYPE | DATA TYPE | NULL |
| ProductKey | Surrogate key |  | INT | NOT |
| ProductAlterKey | Business key |  | INT | NOT |
| Product | Hierarchy | 1 | NVARCHAR(50) | NOT |
| Category | Hierarchy | 1 | NVARCHAR(50) | NOT |
| LineageKey | Metadata |  | INT | NOT |

## Fact

### Order

|  |  |  |  |
| --- | --- | --- | --- |
| COLUMN NAME | COLUMN TYPE | DATA TYPE | NULL |
| OrderKey | Degenerate dimension | INT | NOT |
| ProductKey | Dimension key | INT | NOT |
| CustomerKey | Dimension key | INT |  |
| EmployeeKey | Dimension key | INT |  |
| OrderDateKey | Dimension key | INT | NOT |
| RequiredDateKey | Dimension key | INT |  |
| ShippedDateKey | Dimension key | INT |  |
| UnitPrice | Measure | MONEY |  |
| Quantity | Measure | INT |  |
| Discount | Measure | MONEY |  |
| SalesAmount | Measure | MONEY |  |
| SalesAmountWithDiscount | Measure | MONEY |  |
| LineageKey | Metadata | INT |  |

# Physical Design

## Storage

|  |  |  |
| --- | --- | --- |
| Logical Drive | Filegroups | Notes |
| Disk1 | 1. PRIMARY 2. Default\_FG | * The system tables are in a dedicated filegroup to minimize fragmentation of data files. * The default filegroup is initialized with a single file of 100MB, which should be large enough for the tables it will contain for the foreseeable future. |
| Disk2 | Dimention\_Data | * Putting different tables used in the same join queries in different filegroups. This step will improve performance, because of parallel disk I/O searching for joined data. |
| Disk3 | Dimention\_Index | * Using different filegroups for heavily accessed tables and the nonclustered indexes will improve performance, because of parallel I/O if the files are located on different physical disks. |
| Disk4[[1]](#footnote-1) | Order\_Unkown\_Member\_Data | * This table is likely to grow extremely large, and so it is assigned to a dedicated logical drive to avoid competing for space. |
| Disk5 | Order\_1996\_Data | * As above |
| Disk6 | Order\_1997\_Data | * As above |
| Disk7 | Order\_Unkown\_Member\_Index | * Using different filegroups for heavily accessed tables and the nonclustered indexes will improve performance, because of parallel I/O if the files are located on different physical disks. |
| Disk8 | Order\_1996\_Index | * As above |
| Disk9 | Order\_1997\_Index | * As above |
| Disk10 | 1. Landing 2. Staging DB data files 3. Logs DB data files | * A separate staging database, landing filegroup and logs database is used, and its data files are stored on a dedicated logical disk to minimize the effects of fragmentation and I/O on the data warehouse data and activity. |
| Disk11 | TempDB data files | * TempDB is given its own logical disk to minimize the effect of fragmentation on the data warehouse while enabling fast I/O to a suitable pre-sized TempDB. |
| Disk12 | Database log files | * The logfiles for TempDB, the Staging database, the Logs database and the data warehouse are stored on this drive to isolate log activity from data I/O. |
| Disk13 | Backup volume | * This logical drive will be used for backup file storage |

## Partitioning

The **Order** table are partitioned on the **ShippedDateKey** column. This column was chosen because orders are not considered complete until shipped, so it is shippedto assume that the ship date is used as the cut off for each month’s orders.

Orders for the current month are partitioned into daily partitions, as this is likely to reflect the ETL load frequency. Orders from previous month are merged to form a single partition per month and Orders from previous years are merged to form a single partition per year, as this reduces management complexity.

## Indexes

The dimension tables each have a clustered index on the surrogate key and a non-clustered index on the alternate key, which includes SCD metadata columns where they exist. Additional non-clustered indexes are created on attribute columns that are likely to be queried frequently.

The fact table has clustered columnstore index to maximize query performance. Additional nonclustered indexes are created on dimension key columns to improve performance for joins.

## Compression

All row-stored indexes are compressed using PAGE.

Column-stored indexes compressed using COLUMNSTORE for current and previous year partitions and COLUMNSTORE\_ARCHIVE for archived data.

## Views

Each table has a matching view in the **Reports** schema. These views have user-friendly names and use the NOLOCK query hint to minimize locking.

# Designing an ETL Solution

[Source to target mapping.xlsx](https://dev.azure.com/zinykov/20767/_git/Northwind_BI_Solution?path=/Docs/Source%20to%20target%20mapping.xlsx&version=GBmaster&_a=contents)

## Customer

Northwind (упрощённая)

Customers

Filter on modified date

Clean data with DQS

Load and validate customers data with MDS

Filter on LastChgDateTime and succeeded validation

Add metadata columns

Lookup existing rows

Start audit

Insert new rows (generate surrogate key)

Update SCD1 rows

Customer

Update and Insert SCD2 rows (generate surrogate key)

End audit

DQS

MDS

Customer, ContactName, ContactTitle, Phone, Fax

## Date

xmlcalendar

.\ru

Clean data with DQS

Load and validate holidays data with MDS

Add unknown member

Start audit

Date

End audit

Get start and load dates from Landing.Orders

Populate Date dimension from 01-01 first year to 12-31 last +1 year

DQS

MDS

## Employees

Northwind (упрощённая)

Employees

Filter on modified date

Clean data with DQS

Load and validate employees data with MDS

Add metadata columns

Lookup existing rows

Start audit

Insert new rows (generate surrogate key)

Update SCD1 rows

Employee, TitleOfCourtesy

Employee

Update and Insert SCD2 rows (generate surrogate key)

End audit

Filter on LastChgDateTime and succeeded validation

DQS

MDS

## Product

Northwind (упрощённая)

Products

Filter on modified date

Clean data with DQS

Load and validate products & categories data with MDS

Add metadata columns

Lookup existing rows

Start audit

Insert new rows (generate surrogate key)

Update SCD1 rows

Product, Category

Product

End audit

Categories

Filter on modified date

Clean data with DQS

Filter on LastChgDateTime and succeeded validation

DQS

MDS

## Order

Northwind (упрощённая)

Orders

Create new partition

Create staging table

Load data to staging table

Start audit

Switch partition with unknown shipped date to fact table

Order

End audit

Order Details

Create filegroups if necessary

Get start load date for incremental load

Switch partition with new shipped date to fact table

Insert or update data for previous dates

Drop staging table

1. For every fact partition is needed dedicated disk. [↑](#footnote-ref-1)