عنوان: **آموزش مفاهیم Data Warehouse**
نویسنده: اردلان شاه قلی
تاریخ: ۱۳۹۲/۱۰/۱۱ ۱۱:۰
آدرس: www.dotnettips.info
گروه‌ها: OLAP, MDX, SSAS, data warehouse

مفاهیم مقدماتی Data Warehouse :

OLTP (Online Transaction Processing) : سیستم‌هایی می‌باشند که برای اهداف اصلی سازمان استفاده می‌شوند و این سیستم‌ها کار پردازش و ذخیره کردن داده‌ها را در OLTP Database انجام می‌دهند. مانند تمامی سیستم‌های ERP,MIS,...

OLTP Database : پایگاه داده‌ی سیستم‌های OLTP می‌باشد. به طور معمول هر تراکنش کاربر در کمترین زمان ممکن برروی این سیستم‌ها ذخیره می‌گردد و در طول روز بارها دستورات (Insert/Update/Delete) برروی آنها انجام می‌شود. این پایگاه‌های داده، همان Main Data ها یا Source System ها می‌باشند.

ETL (extract, transform, and load) : مراحل انتقال داده از OLTP Database به پایگاه داده‌ی Stage می‌باشد. ETL سیستمی می‌باشد که توانایی اتصال به OLTP را دارد و اطلاعات را از OLTP واکشی می‌کند و به پایگاه داده‌ی Stage انتقال می‌دهد. سپس ETL داده‌ها را مجتمع (integrates) کرده و از Stage به DDS (Dimensional Data Source) انتقال می‌دهد .

Retrieves Data : عملیات واکشی داده‌ها طبق یک سری قوانین و قواعد می‌باشد .

برای انجام عملیات ETL دو روش وجود دارد

1. Data مجتمع (Integrate) و تمیز (Data cleansing) شود و در نهایت وارد Data Warehouse گردد.

2. Data وارد Data Warehouse گردد سپس مراحل مجتمع سازی و پاک سازی داده‌ها بر روی داده‌ها در خود Data Warehouse انجام گردد.

Consolidates Data : برخی شرکت‌ها داده‌های اصلی خودشان را در چندین پایگاه داده دارند. در این حالت برای انجام عملیات ETL باید داده‌ها تحکیم و مجتمع شوند و سپس در Data Warehouse ذخیره شوند.

به طور کلی موارد زیر در فرایند ETL در نظر گرفته می‌شود:

1. Data availability : برخی داده‌ها در یک سیستم وجود دارند ولی در سیستم دیگری وجود ندارند و یا تفاوت در نگهداری داده‌ها در سیستم‌های مختلف داریم. مثلا در یک سیستم آدرس در سه فیلد نگه داری می‌شود (کشور-شهر-آدرس) اما در سیستمی دیگر در دو فیلد(کشور-آدرس) نگه داری می‌شود. در این حالت باید ما در ETL راه کار هایی برای مجتمع کردن این موارد در نظر بگیریم.

2. Time ranges : در سیستم‌های مختلف امکان دارد بعدهای زمانی مختلف باشد . مثلا در یک سیستم بررسی‌ها در بازه‌ی ساعتی و در سیستم دیگر بررسی‌ها در بازه‌ی روزانه یا ماهانه باشد . بنابر این در تجمیع داده‌ها باید این مورد مد نظر گرفته شود.

3. Definitions : تعاریف در سیستم‌های مختلف می‌تواند متفاوت باشد. مثلا در یک سیستم، مبلغ کل فاکتور شامل مالیات می‌باشد ولی در سیستمی دیگر این مبلغ فاقد مالیات می‌باشد.

4. Conversion : در فرآیند ETL باید باز از قواعد موجود در سیستم‌های مختلف آگاهی داشته باشیم. مثلا در یک سیستم ممکن است دما را به صورت سانتیگراد و در دیگری فارنهایت نگه داری کنند.

5. Matching : باید بررسی لازم را انجام دهیم که کدام داده مرتبط با کدام سیستم می‌باشد. به عبارت دیگر کدام سیستم مالک داده می‌باشد و دقیقا داده‌ها در کدام سیستم معتبرتر می‌باشند. مثلا پرسنل، هم در سیستم حسابداری می‌باشند هم در سیستم پرسنلی؛ ولی معمولا داده‌های اصلی از سیستم پرسنلی می‌آیند.

Periodically : عملیات واکشی داده‌ها (Retrieves Data) و مجتمع سازی داده‌ها (Consolidates Data) در فرآیند ETL فقط یکبار اتفاق نمی‌افتد و این مراحل در بازه‌های زمانی خاص تکرار می‌گردند. این واکشی و انتقال داده‌ها می‌تواند در روز چند بار تکرار شود یا می‌تواند چند روز یک بار اجرا گردد و این بستگی دارد به سیاست موجود در Data Warehouse .

DDS (Dimensional Data Source) (Data Warehouse) : یک پایگاه داده از نوع نرمال شده (Normalized) یا بعدی (Dimensional) می‌باشد. که داده‌های مجتمع شده و تمیز شده سیستم‌های OLTP را در خود جای داده است. این پایگاه داده برای واکشی‌های سیستم‌های آنالیز داده مورد استفاده قرار می‌گیرد. ورود اطلاعات در Data Warehouse به صورت Batch می‌باشد و به هیچ عنوان مانند پایگاه داده‌های OLTP ویرایش داده‌ها به صورت Online و هر زمان که داده‌ها تغییر می‌کنند، صورت نمی‌گیرد. اطلاعات در Data Warehouse معمولا به صورت تجمیع شده روزانه، ماهانه، فصلی یا سالانه می‌باشد. DDS ها مجموعه ای از Dimensional Data Mart ها هستند. و عمدتا به صورت denormalized می‌باشند.

Dimensional Data Mart : مجموعه ای از جداول Fact , Dimension می‌باشند که در یک بیزینس خاص باهم در ارتباط و مشترک می‌باشند.

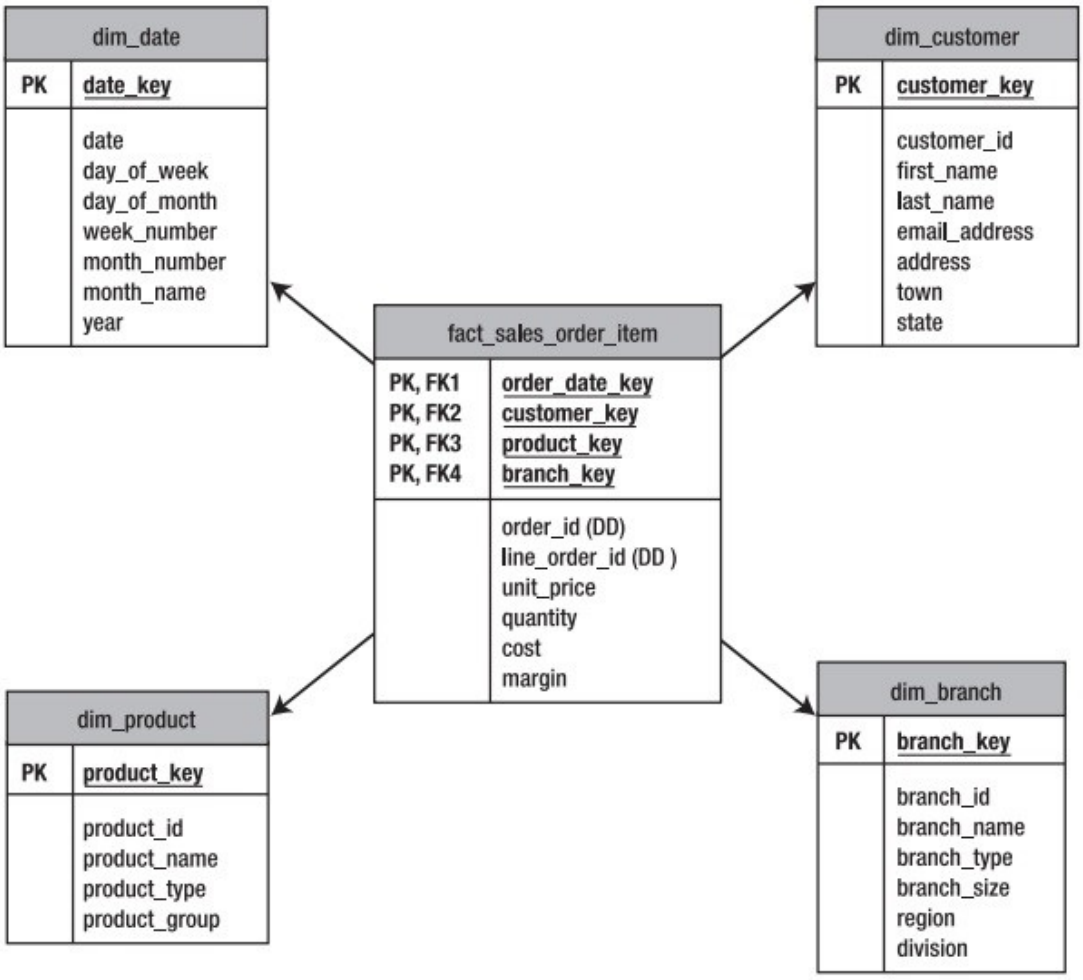
dimensional data store schemas : طراحی‌های مختلفی از جداول Fact , Dimension در DDS وجود دارد که عبارتند از

1. Star schema : ساده‌ترین روش پیاده سازی Data Warehouse

2. Snowflake : در این روش جداول Dimension کمی نرمال سازی بیشتری دارند. سیستم‌های آنالیز داده با این روش بهتر کار می‌کنند.

3. Galaxy schemas : طراحی در این روش بسیار سخت و پیچیده می‌باشد. با این وجود فرایند ETL در این طراحی ساده‌تر انجام می‌شود.

نمونه‌ی طراحی Star به صورت زیر می‌باشد :



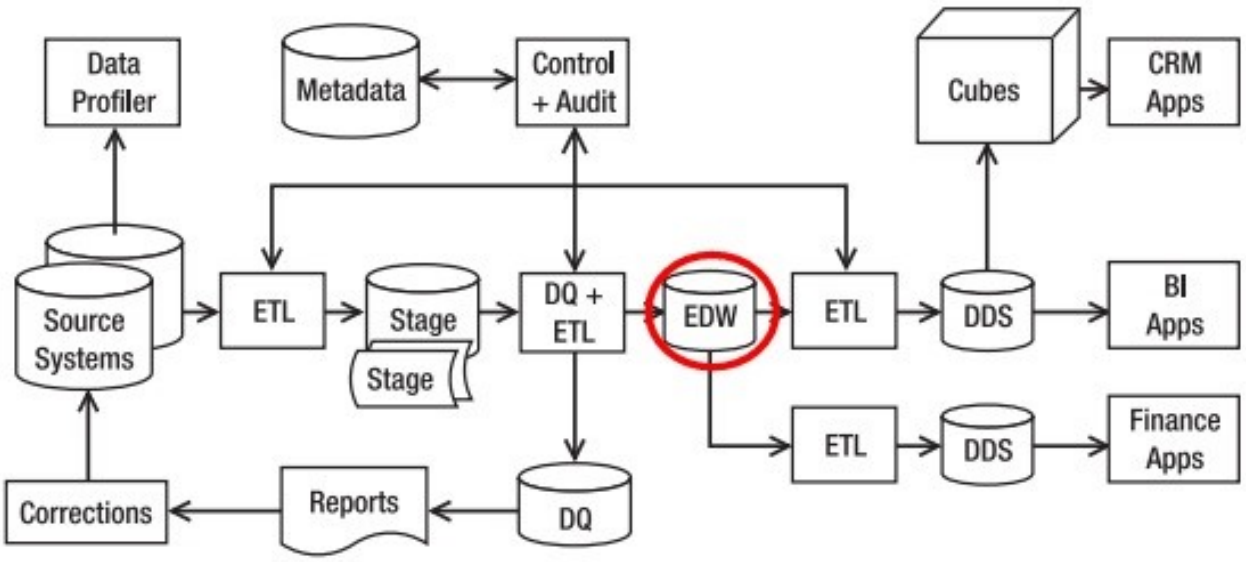
تفاوت‌های DDS و NDS :

1. در DDS ها هیچ گونه نرمال سازی خاصی انجام نمی‌دهیم و عملا تمامی جداول را دینرمال کرده ایم، در حالی که در NDS تمامی جداول تا سطح سوم و گاهی تا سطح پنجم نرمال شده اند.

2. سرعت واکشی و پردازش کوئری‌ها روی DDS خیلی بیشتر از NDS ها می‌باشد.

3. در صورتی که نیاز باشد Data Warehouse های خیلی بزرگ طراحی کنیم با حجم بسیار زیاد توصیه می‌شود از NDS ها استفاده شود در حالی که برای Data Warehouse های کوچک و متوسط بهتر است از DDS ها استفاده شود.

تصویر طراحی یک (EDS) Enterprise Data Source = NDS در زیر آمده است :



History : جداول Data Warehouse میتوانند در طول زمان بسیار بزرگ شوند و دارای تعداد رکورد زیادی گردند. اینکه حداکثر داده‌های چند سال را در Data Warehouse نگه داری کنیم بستگی به سیاست‌های سازمانی دارد که سیستم OLAP برای آن تهیه می‌گردد. استفاده کردن از table partitioning می‌تواند در جبران افزایش تعداد رکورد کمک زیادی به ما بکند.

(slowly changing dimension (SCD : سه روش برای نگه داری سابقه‌ی تغییرات در جداول Dimension وجود دارد.

1. SCD type 1 : هیچ گونه سابقه‌ی تغییراتی را نگه داری نمی‌کنیم

2. SCD type 2 : سابقه‌ی تغییرات در ردیف‌ها نگه داری می‌شود. در این روش هر ردیف، شماره ردیف قبلی را دارد و تعداد نا محدودی از تغییرات را نگه داری می‌کنیم.

3. SCD type 3 : سابقه‌ی تغییرات در ستون‌ها نگه داری می‌شوند و فقط ردیف جاری و آخرین تغییرات را نگه داری می‌کنیم.

Query : فقط ETL حق تغییرات در Data Warehouse را دارد و کاربر نمی‌تواند Data Warehouse را تغییر دهد. البته کاربران حق Query کردن از Data Warehouse را دارند.

دقت داشته باشید که کوئری‌های پیچیده در NDS ها بسیار کندتر از همان کوئری در DDS می‌باشد.

Business Intelligence : مجموعه ای از فعالیت‌ها که در یک سازمان برای شناخت بهتر وضعیت Business آن سازمان انجام می‌شود. نتایج BI کمک بسیاری برای تصمیم گیری‌های تکنیکی و استراتژیکی درون سازمان می‌کند. همچنین کمک به بهبود فرایندهای Business جاری می‌کند.

فعالیت‌های Business Intelligence در سه دسته بندی قرار می‌گیرند :

1. Reporting : گزارشاتی که از Data Warehouse گرفته می‌شود و به کاربر نمایش داده می‌شود و عمدتا این گزارشات به صورت tabular form می‌باشند.

2. OLAP : فعالیت‌های انجام شده روی MDB برای گرفتن گزارشات Drill-Down و ... می‌باشد.

3. Data mining : فرآیند واکشی و داده کاوی داده‌های درون سیستم می‌باشد، که منجر به کشف الگوها و رفتارها و ارتباطات

داده‌ها در سیستم می‌شود. توسط داده کاوی ما متوجه می‌شویم چرا برخی داده‌ها در سیستم تولید شده اند.

a. descriptive analytics : زمانی که از داده کاوی برای شرح وقایع گذشته و حال استفاده می‌شود.

b. predictive analytics : زمانی که از داده کاوی برای پیش بینی وقایع گذشته استفاده می‌شود.

Real time data warehouse : به DW هایی گفته می‌شود که در کمترین زمان، تغییرات OLTP را در خود خواهند داشت. امروزه این نوع DW ها تغییرات 5 دقیقه تا حداکثر 1 ساعت قبل را در خود دارند. برای دسترسی به چنین DW هایی دو راه زیر وجود دارد :

1. بر روی هر جدول، Trigger هایی باشد تا تغییرات را به DW انتقال دهد. (البته برای این منظور باید Business مربوط به ETL را در این تریگرها نوشت)

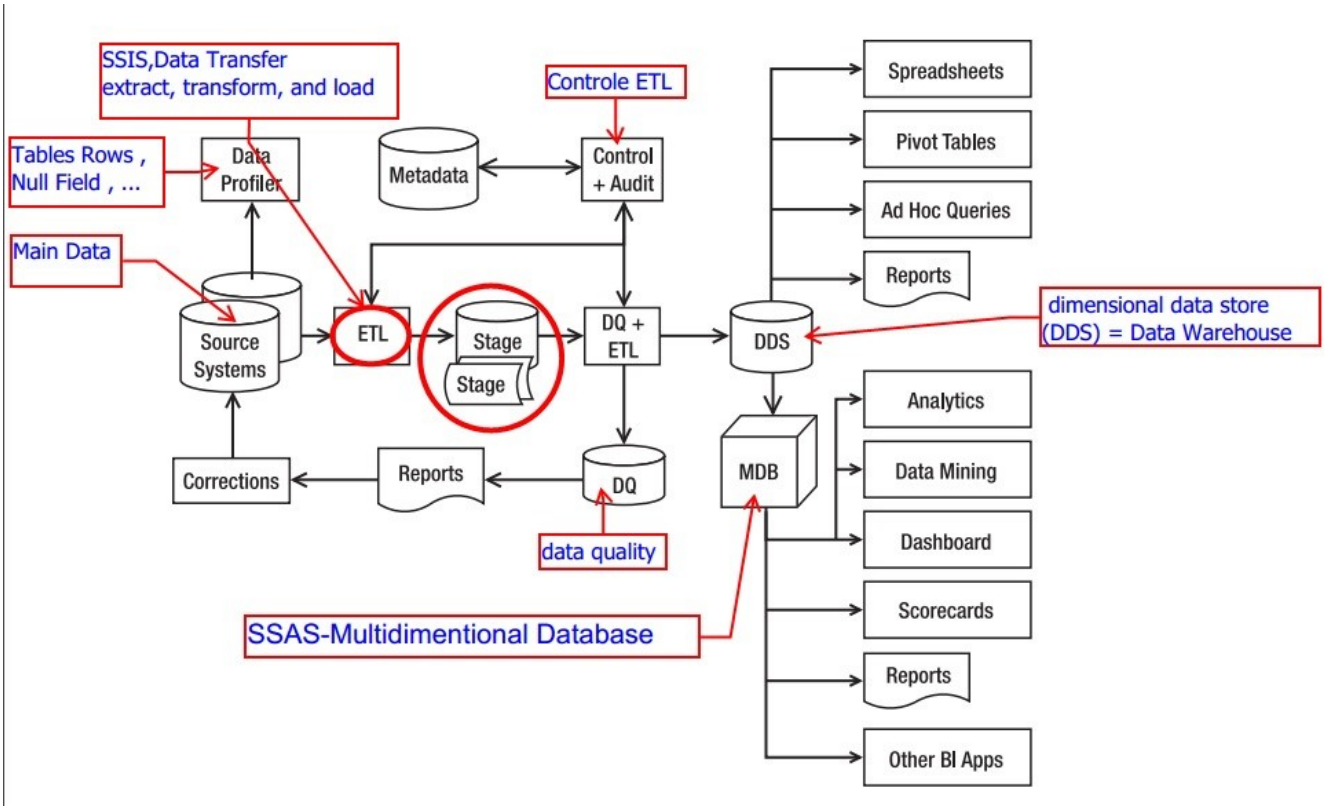
2. سورس برنامه‌های اصلی کاربر (OLTP) تغییر کند تا علاوه بر OLTP Database ها Data Warehouse را هم تغییر دهند.

روش‌های فوق بسیار روی سرعت و کارایی برنامه‌های اصلی تاثیر خواهند گذاشت.

NDS (Normalize Data Source) : در صورتی که طراحی Data Warehouse به صورت Dimensional نباشد و به صورت Normalize باشد، نوع Data Warehouse از نوع NDS می‌باشد.

روش ساخت MDB :

OLTP Database -> ETL -> Stage Database -> DDS (Dimensional Data Source = Data Warehouse) -> SSAS -> MDB



روش ساده‌تر ساخت Data Warehouse :

منظور از Source System همان OLTP Database ها می‌باشد.

به خاطر داشته باشید که Source System ها جزئی از Data Warehouse نمی‌باشند.

از کاربردهای Data Warehouse می‌توان به موارد زیر اشاره کرد

1. Data Mining

2. استفاده در گزارشات

3. تجمیع داده ها

Data Mining کمک به درک بهتر Business جاری در سازمان می‌کند. همچنین منجر به کشف دانش از درون داده‌ها می‌شود.

برای Data Mining می‌توانید از انواع پایگاه داده‌های موجود مانند رابطه ای ، سلسله مراتبی و چند بعدی استفاده کرد . حتا می‌توان از فایل‌های Excel , XML نیز استفاده کرد.

(Customer Relationship Management (CRM :

منظور از مشتری، مصرف کننده‌ی سرویسی است که سازمان شما ارایه می‌کند. یک سیستم CRM شامل تمامی برنامه ایی می‌باشد که تمام فعالیت‌های مشتری را پشتیبانی می‌کند.

(Operational Data Store (ODS :

این پایگاه داده به صورت رابطه ای و نرمال شده می‌باشد و شامل تمامی اطلاعات پایگاه داده ای OLTP می‌باشد که در این پایگاه داده مجتمع شده اند. تفاوت ODS با Data Warehouse در این می‌باشد که داده‌ها در ODS با هر Transaction به روز می‌شوند (سرعت بروز رسانی اطلاعات در ODS بالاتر از DW می‌باشد).

(Master Data Management (MDM :

در یک نگاه می‌توان داده‌ها را به دو دسته تقسیم کرد

1. transaction data

2. master data

transaction data : شامل داده ای transactional در سیستم‌های OLTP می‌باشد.

master data : توضیح دهنده‌ی Business جاری در سازمان می‌باشد.

برای تشخیص این دو نیاز است Business سازمان را به خوبی شناسایی نمایید. به عبارت دیگر رویدادهای Business ی همان transaction data می‌باشند و master data شامل پاسخ‌های این سوال‌ها می‌باشد. چه کسی، چه چیزی و کجا در مورد Business transaction .

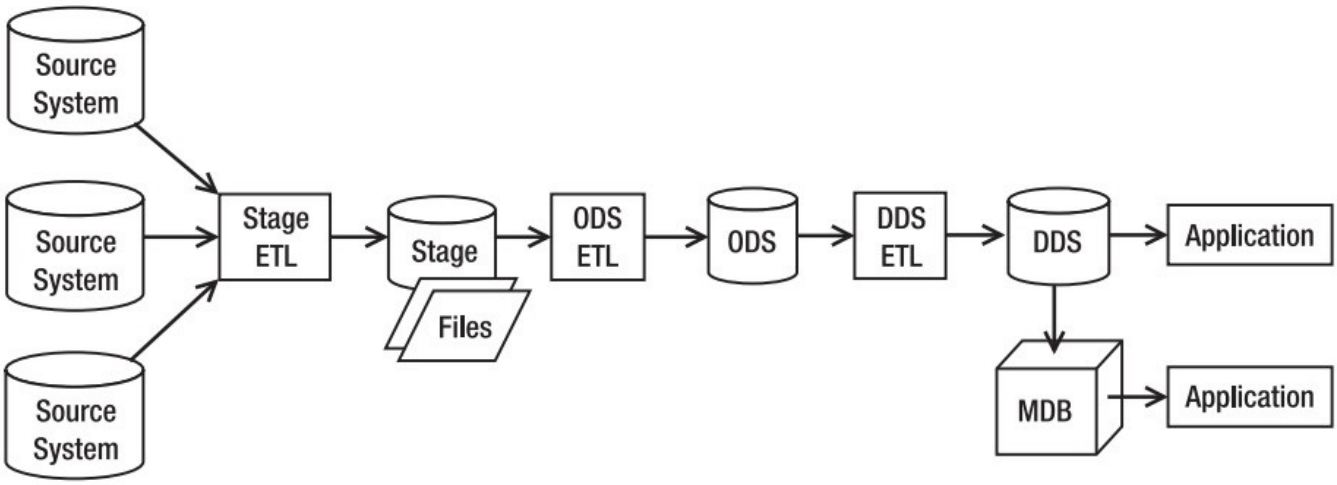
CDI (Customer data integration) : عبارت است از MDM در رابطه با مشتری داده ها. کار این قسمت عبارت است از واکشی، پاک سازی ، ذخیره سازی ، نگه داری و به اشتراک گذاشتن داده ای مشتری می‌باشد.

Unstructured Data : داده ای ذخیره شده در پایگاه داده ، structured Data می‌باشند و داده هایی مانند عکس و فیلم و صوت و ...

SOA (Service-Oriented Architecture) : یک متد ساخت برنامه می‌باشد که در این روش تمامی اجزا برنامه به صورت ماژول هایی دیده می‌شود که در آنها ارتباطات با دیگر سیستم‌ها به صورت سرویس می‌باشد و این زیر سیستم‌ها را می‌توان در پروژه‌های مختلف به کار برد.

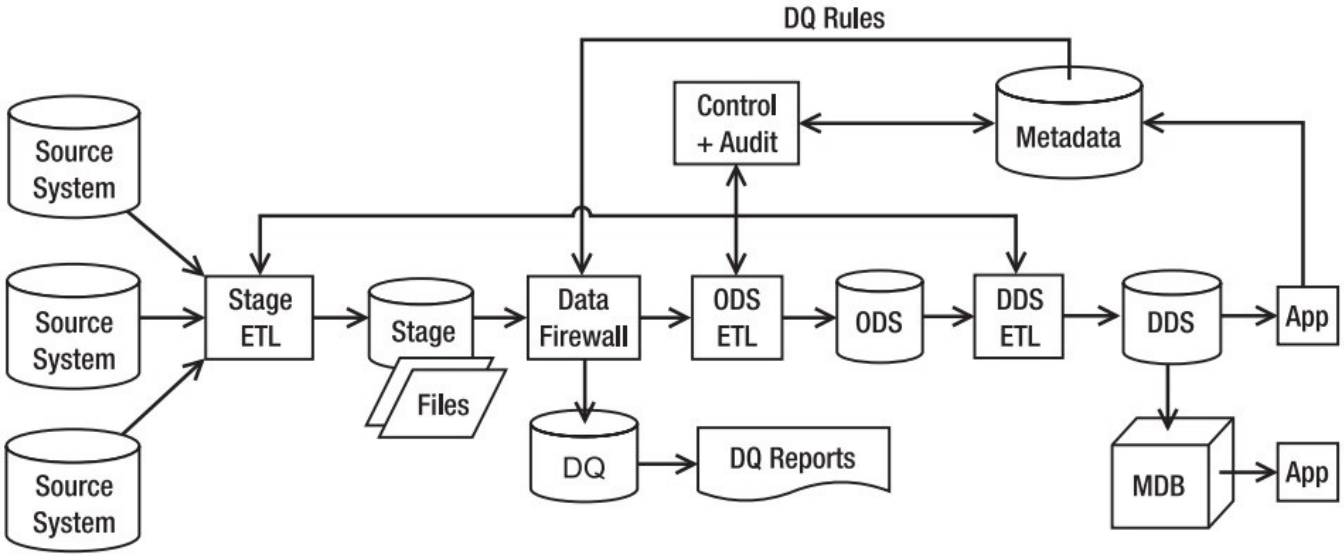
DW : Real-Time Data Warehouse هایی که توسط ETL به روز می‌شوند در هنگامی که یک Transaction روی OLTP اتفاق می‌افتد.

مراحل انتقال داده از OLTP Database به MDB به صورت زیر می‌باشد.



Data quality : مکانیسم اطمینان بخشی از این که در DW داده‌های مناسب و درست وارد می‌شوند. به عبارت دیگر DQ همان firewall برای DW در مقابل داده‌های نامناسب می‌باشد.

برای بهتر مشخص شدن مکان DQ شکل زیر را در نظر بگیرید



نحوه‌ی حرکت داده ای از OLTP به MDB اولین چیزی می‌باشد که شما باید به آن فکر کنید و برای آن روشی را انتخاب نمایید قبل از ساخت Data Warehouse .

چهار روش برای معماری انتقال اطلاعات از OLTP به DW وجود دارد (البته به عنوان نمونه و شما می‌توانید از روش‌های دیگر و طراحی‌های مختلف و ترکیبی نیز بهره ببرید)

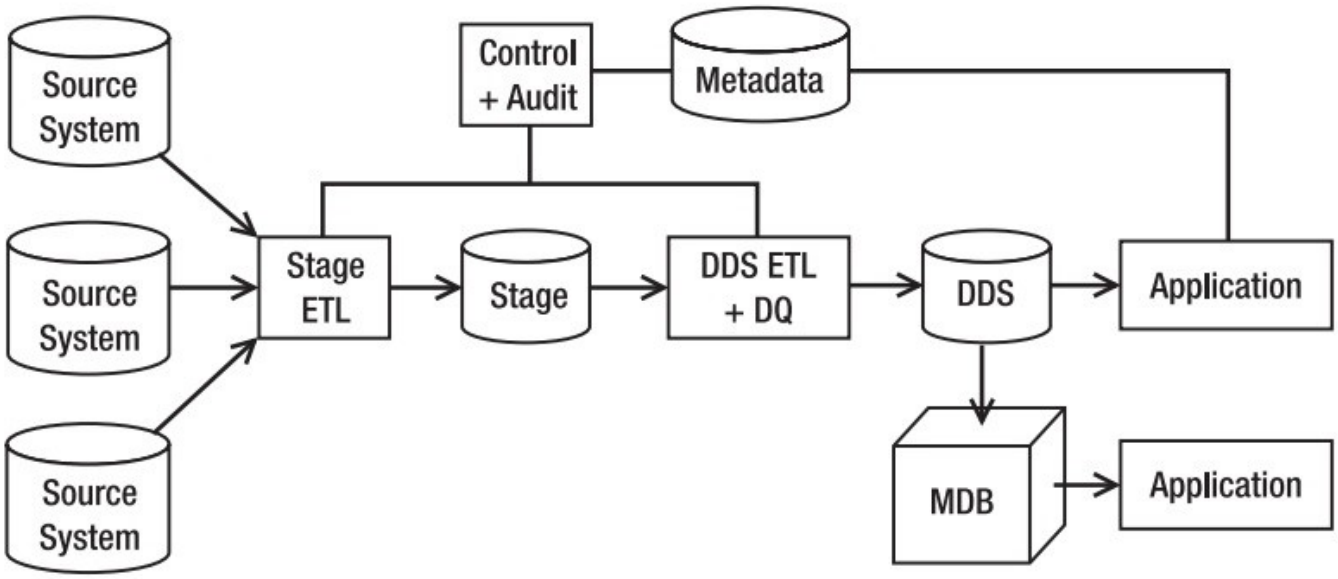
1. single DDS : در این روش فقط Stage , DDS وجود دارد.

2. NDS + DDS : در این روش علاوه بر Stage,DDS از NDS نیز استفاده می‌شود.

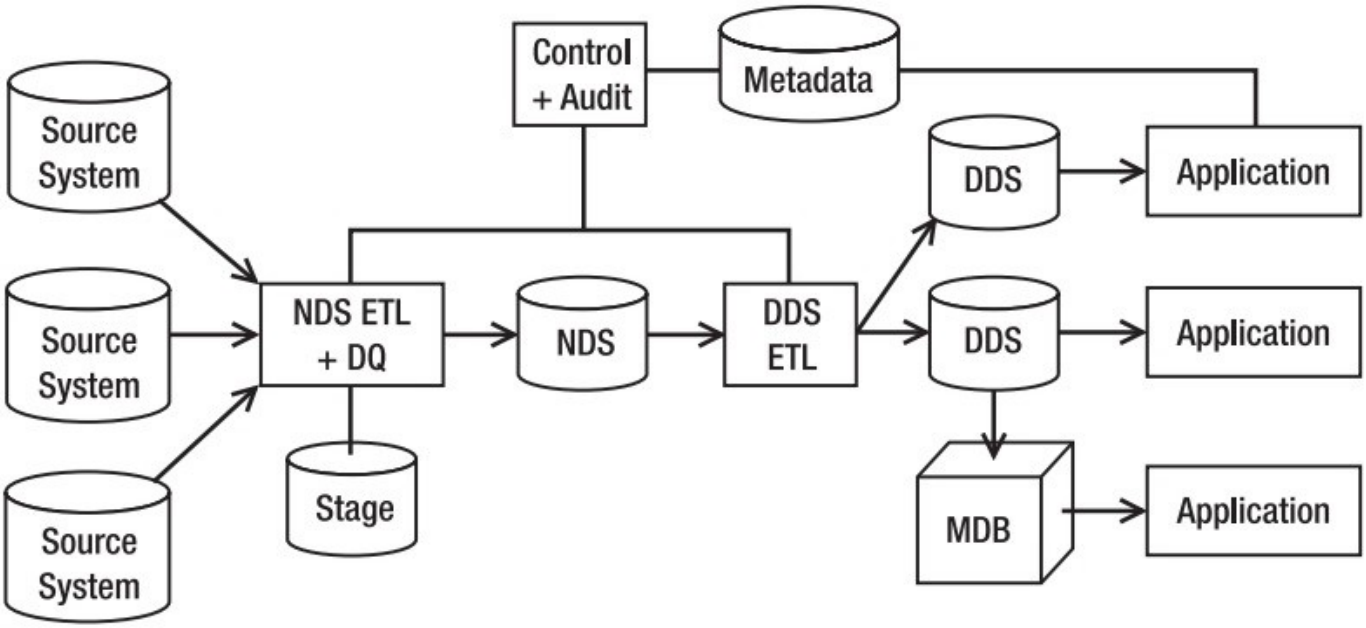
3. ODS + DDS : در این روش از Stage,ODS,DDS استفاده می‌گردد.

4. federated data warehouse (FDW) : استفاده از چندین DW که با هم تجمیع شده اند.

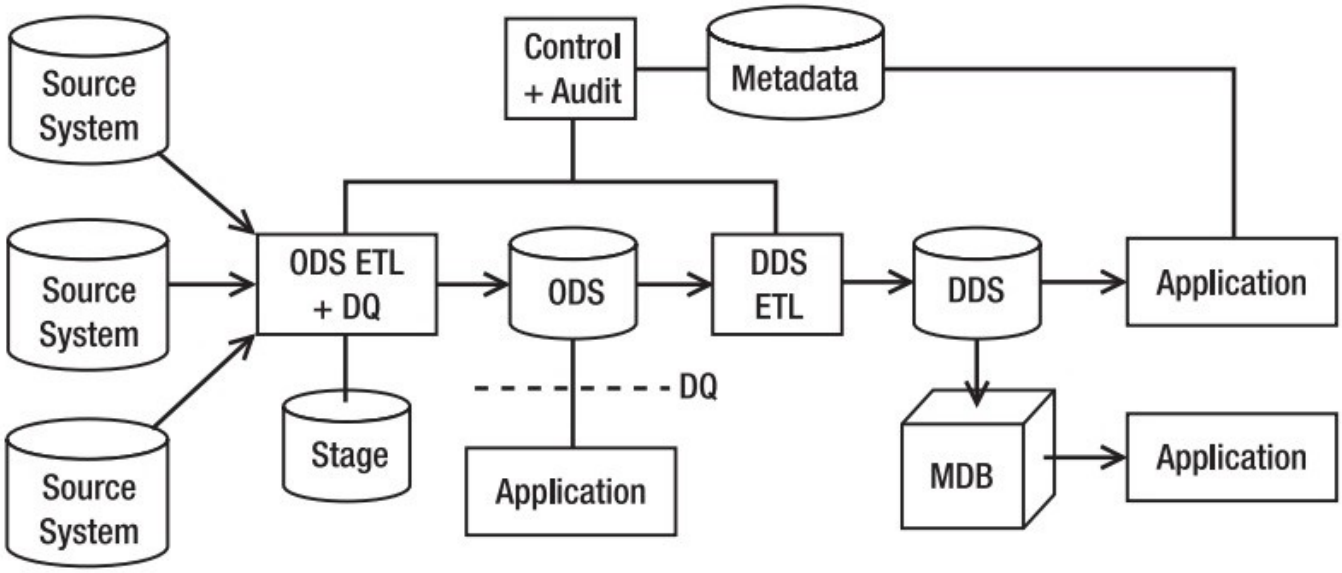
تصویر Single DDS :



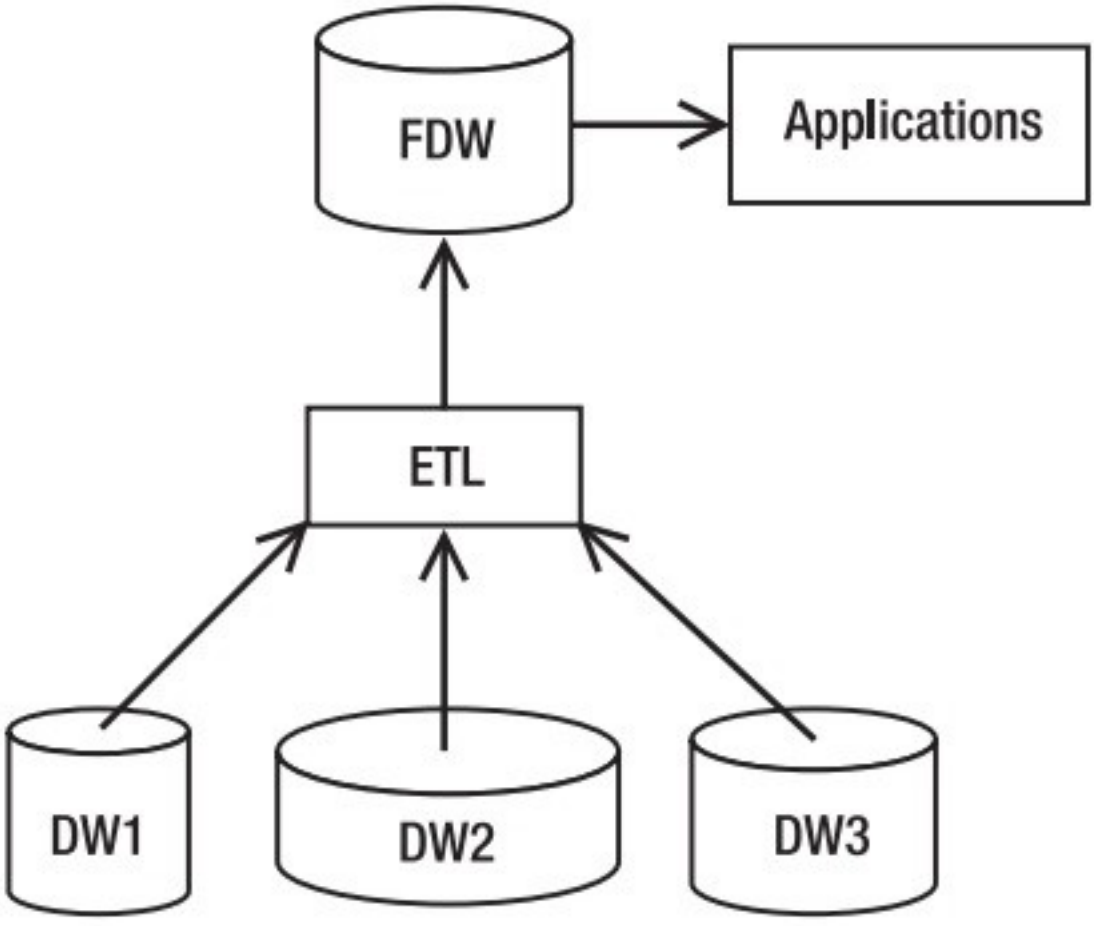
تصویر NDS + DDS :



تصویر ODS + DDS :



تصویر FDW (federated data warehouse) :



منبع : Building a Data Warehouse With Examples in SQL Server انتشارات Apress

نظرات خوانندگان

نویسنده: لیبرتاد
تاریخ: ۱۳۹۲/۱۰/۱۱ ۱۳:۳

بسیار عالی
آیا OLTP و DW می‌توانند بر روی یک سرور باشند مثلا" بر روی یک SQL Server باشند

نویسنده: اردلان شاه قلی
تاریخ: ۱۳۹۲/۱۰/۱۱ ۱۵:۴۵

قطعا امکان پذیر می‌باشد. اما توصیه می‌شود اینچنین نباشد. درضمن در نظر داشته باشید که در خیلی از موارد، DBMSهای سیستم‌های OLTP برروی SQL Server نمی‌باشند(مثلا سیستم حسابداری از پایگاه داده‌ی پارادوکس استفاده می‌کند در حالی که سیستم منابع انسانی دارای پایگاه داده‌ی اراکل می‌باشد و ...)