عنوان: **تعریف انبار داده Data Warehouse**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۲/۱۵ ۱۰:۳۰
آدرس: www.dotnettips.info
گروه‌ها: data mining, data warehouse

در این مقاله در ادامه‌ی مطلبی که تحت عنوان «آموزش مفاهیم Data Warehouse» توسط آقای شاه قلی منتشر شده بود، به بررسی بیشتر مفهوم انبار داده (Data Warehouse) پرداخته می‌شود.

مقدمه

در سازمان ها، داده‌ها و اطلاعات معمولاً به دو شکل در سیستم‌ها پیاده سازی می‌گردد:

- **سیستم‌های عملیاتی OLTP:**

این سیستم‌ها باعث می‌گردند تا چرخ کسب و کار بگردد. وجود این سیستم‌ها سبب می‌شود تا داده‌های مربوط به کسب و کار، به بانک اطلاعاتی وارد شوند. این سیستم‌ها عموماً:

o به دلیل کوتاهی عملیات دارای سرعت قابل توجهی می‌باشند.
o محیطی جهت ورود داده‌ها می‌باشند.
o معمولاً اپراتورها، استفاده کننده‌های آن هستند.

- **سیستم‌های اطلاعاتی DSS، DW/BI ، OLAP :**

این سیستم‌ها باعث می‌گردند تا چرخش کسب و کار را بنگرید. فلسفه بکارگیری این سیستم‌ها در سازمان این است که اطلاعات مورد نیاز مدیران، از درون داده‌های سیستم‌های عملیاتی موجود، استخراج گردد. این سیستم‌ها عموماً:

o به دلیل آنالیز حجم انبوهی از داده ها، معمولاً کندتر از سیستم‌های عملیاتی می‌باشند.
o محیطی جهت تولید گزارشات تحلیلی و آماری می‌باشند.
o معمولاً مدیران و تصمیم گیرندگان سازمان ها، استفاده کنندگان آن می‌باشند.

سیستم‌های عملیاتی در جامعه ما سابقه بیشتری داشته و متخصصین فناوری اطلاعات عموماً با طراحی و تولید چنین سیستم هایی آشنایی کافی دارند. متاسفانه جایگاه سیستم‌های اطلاعاتی در جامعه ما کمتر شناخته شده و متخصصین فناوری اطلاعات بندرت با مفاهیم و نحوه پیاده سازی آن آشنایی دارند.
این نکته حائز اهمیت است که سیستم‌های اطلاعاتی یک سیستم یا محصول نیستند که بتوان آنها را خریداری کرد. بلکه یک راهبرد (Solution, Approach) هستند و در حقیقت هر راهبردی مربوط به یک نوع کسب و کار (Business) و یا سازمان می‌باشد و نمی‌توان فرمول واحدی را برای حتی سازمان‌های مشابه، ارائه نمود.

گارتنر در ابتدای سال 2011 گزارشی را منتشر کرده که نشان میدهد بازار BI با 9.7 % رشد، ارزشی بالغ بر 10.8 بیلیون دلار داشته، ولی متاسفانه پروژه‌های آن به طور متوسط با 75% شکست مواجه شده است. در حالیکه 4 سال پیش، این رقم حدود 50% بود. این موسسه BI را پنجمین اولویت مدیران IT ذکر کرده است.

مفاهیم و مباحث مربوط به Data Warehouse به اواسط دهه 1980 برمی گردد، به زمانی که IBM تحقیقاتی را در این زمینه شروع کرد و نتیجه آنرا «Information Warehouse» نامید و هنوز هم در برخی منابع از این واژه بجای Data Warehouse استفاده می‌شود. از این پس برای راحتی از اختصار DW بجای Data Warehouse استفاده می‌شود. انبارهای داده جهت رفع نیاز رو به رشد مدیریت داده‌ها و اطلاعات سازمانی که توسط پایگاه‌های داده سیستم‌های عملیاتی غیر ممکن بود، ساخته شدند.

انبار داده به مجموعه ای از داده‌ها گفته می‌شود که از منابع مختلف اطلاعاتی سازمان جمع آوری، دسته بندی و ذخیره می‌شود. در واقع یک انبار داده مخزن اصلی کلیه داده‌های حال و گذشته یک سازمان می‌باشد که برای همیشه جهت انجام عملیات گزارش گیری و آنالیز در دسترس مدیران می‌باشد. انباره‌های داده حاوی داده هایی هستند که به مرور زمان از سیستم‌های عملیاتی آنلاین سازمان، استخراج می‌شوند. بنابراین سوابق کلیه اطلاعات و یا بخش عظیمی از آنها را می‌توان در انباره داده‌ها مشاهده نمود.
از آنجائیکه انجام عملیات آماری و گزارشات پیچیده دارای بار کاری بسیار سنگینی برای سرورهای پایگاه داده می‌باشند، وجود انبار داده سبب می‌گردد که این گونه عملیات تاثیری بر فعالیت برنامه‌های کاربردی سازمان نداشته باشد.

همانگونه که پایگاه داده سیستم‌های عملیاتی سازمان (برنامه‌های کاربردی) به گونه ای طراحی می‌شوند که انجام تغییر، حذف و اضافه داده به سرعت صورت پذیرد، در مقابل انبار داده‌ها دارای معماری ویژه ای می‌باشند که موجب تسریع انجام عملیات آماری و گزارش گیری می‌شود. در حقیقت می‌توان اینگونه بیان نمود که انباره داده یک مخزن فعال و هوشمند از اطلاعات است که قادر است اطلاعات را از محیط‌های گوناگون جمع آوری و مدیریت کرده و نهایتا پخش نماید و در صورت لزوم نیز سیاست‌های تجاری را روی آنها اجرا نماید.

:Bill Inmon

او را پدر DW می‌نامند، از دیدگاه او DW هسته مرکزی چیزی است که او آنرا CIF اختصار (Corporate Information Factory) می‌نامد، که پایه و اساس BI بر مبنای آن قرار دارد. وی از طرفداران Top-Down Design می‌باشد که معتقد است در زمان طراحی باید با دیدی سازمانی، CIF را مدل سازی، ولی بصورت دپارتمانی پیاده سازی کرد (Think Globally, Implement Locally). در این نوع طراحی از DW به Data Mart خواهیم رسید.

:Ralph Kimball Ph.D

به نظر وی DW چیزی نیست جز یک کپی از داده‌های عملیاتی که به طرز خاصی برای گزارشات و تحلیل‌های آماری، آماده و ساختمند شده است. به بیان دیگر DW سیستمی است جهت استخراج، پالایش، تطبیق و تحویل اطلاعات منابع داده ای به یک بانک اطلاعاتی Dimensional و اجرای Query و گزارشات آماری و تحلیلی برای اهداف تصمیم گیری و استراتژیک سازمان.
وی معرفی کننده یکی از اساسی‌ترین مفاهیم طراحی یعنی Dimensional Modeling است؛ ماحصل چنین ایده ای، اساس شکل گیری مدلی است که امروزه کارشناسان آنرا به نام Cube می‌شناسند. وی از طرفداران Bottom-Up Design است که در این نگرش از Data Mart به DW می‌رسیم. این روش به نظر عملی‌تر از روشی می‌باشد که به یکباره DW جامع و کامل برای اهداف سازمانی طراحی و پیاده سازی گردد.

تعریف انبار داده :
W.H.Inmon پدر DW آنرا چنین تعریف می‌کند:

```
The Data Warehouse is a collection of
```

Integrated

```
,
```

Subject-Oriented

```
databases designed to support the DSS function, where each unit of data is
```

Non-Volatile

```
and
```

relevant

```
to some moment in
```

Time

از تعریف فوق دو مورد دیگر نیز به طور ضمنی استنباط می‌شود:

o انبار داده به طور فیزیکی، کاملاً جدا از سایر سیستم‌های عملیاتی است.
o داده‌های DW مجموعه ای Aggregated و Atomic از داده‌های تراکنش‌های سیستم‌های عملیاتی است که سوای کاربرد آنها در سیستم‌های عملیاتی، برای مقاصد مدیریتی نیز استفاده خواهد شد.

به بیان دیگر DW راهبردی است که دسترسی آسان به اطلاعات درست (Right Information)، در زمانی درست (Right Time) ، به کاربران درست (Right Users)، را فراهم می‌آورد تا «تصمیم گیری سازمانی» قابل انجام باشد. DW صرفاً یک محصول نرم افزاری و

یا سخت افزاری نیست که بتوان آنرا خریداری نمود بلکه فراتر از آن و در حقیقت یک محیط پردازشی می‌باشد که کاربران می‌توانند از درون آن اطلاعات مورد نیاز خود را بیابند.
DW اطلاعات خود را از سایر بانک‌های اطلاعاتی از نوع OLTP و یا سایر DWهای لایه پایین‌تر و به صورت دسته ای (Batch) و یا انبوه (Bulk Loading) جمع آوری می‌کند. یک DW به صورت سنتی باید شامل داده‌های Historic سازمان باشد و می‌توان اینگونه بیان نمود که در DW هرچه داده‌های قدیمی‌تری موجود باشد، اعتبار تحلیل‌های آماری سیستم افزایش خواهد یافت.

داده‌های سیستم عملیاتی را نمی‌توان بلافاصله درون بانک اطلاعاتی DW لود نمود، چنین داده هایی باید آماده سازی، پالایش و همگون گردند تا شرایط لود در DW را داشته باشند. حداقل کاری که انتظار داریم یک DW در مورد داده‌ها برای ما برآورده سازد شامل موارد زیر است:

- o استخراج داده‌ها از منابع مختلف (مبدإ)
- o تبدیل داده‌ها به فرمتی یکسان
- o لود داده‌ها به جداول مربوطه (مقصد)

با هر با اجرای پروسه فوق یکی از سه مورد زیر، بسته به نیاز طراحی و محدودیت‌های تکنولوژی رخ خواهد داد:

- o تمام داده‌ها در DW با داده‌های جدید جایگزین خواهند گردید(Full Load, Initial Load, Full Refresh).
- o داده‌های جدید به داده‌های موجود اضافه خواهند گردید Incremental Load (Inserted data).
- o نسخه جدیدی از داده‌های کنونی به سیستم اضافه خواهند گردید Incremental Load (Updated data).

ویژگی‌های داده‌های درون DW

داده‌های DW از نگاه Inmon دارای 4 ویژگی اصلی زیر هستند:

o فقط خواندنی (Non-Volatile):

هیچ رکوردی و یا داده ای Update نخواهد شد و صرفاً رکوردهایی که محتوای مقادیر جدید داده‌ها هستند، به سیستم اضافه خواهند شد.

o موضوع گرا (Subject-Oriented):

منظور از «موضوع» پایه‌های اساسی یک کسب و کار هستند، به شکلی که با حذف یکی از این پایه ها، شاید ماهیت آن کسب و کار از ریشه دگرگون شود. برای مثال موضوعاتی چون «مشتری» و یا «بیمه نامه» برای شرکت‌های بیمه.

o جامع (Integrated):

باید تمامی کدهایی که در سیستم‌های عملیاتی وجود دارند و معانی یکسانی دارند، برای مثال کد جنسیت، در DW به یک روش ذخیره و نمایش داده شوند.

o زمانگرا (Time Variant):

هر رکورد باید حاوی فیلد و یا کلیدی باشد که نمایانگر این باشد که این رکورد در چه زمانی ایجاد، استخراج و ذخیره شده است. از آنجا که داده‌های درون سیستم‌های عملیاتی آخرین و به روزترین داده هر سیستم میباشد، نیازی به وجود چنین عنصری در سیستم‌های OLTP احساس نمی‌گردد، ولی چون در DW تمام داده‌های نسخ قدیمی داده‌های سیستم‌های عملیاتی موجود می‌باشد، باید حتماً مشخص گردد که هر داده ای در سیستم‌های عملیاتی در چه زمانی، چه مقادیری داشته است. این عنصر زمانی کمک می‌کند تا بتوانیم:

- o گذشته را آنالیز کنیم.
- o اطلاعات مربوط به حال حاضر را بدست آوریم.
- o آینده را پیش بینی کنیم.

منبع: کتاب آقای خشایار جام سحر با عنوان بانک داده تجمیعی

Comparison Kimball vs. Inmon Inmon

```
Continuous & Discrete Dimension Management
```

```
Define data management via dates in your data
```

Continuous time

When is a record active

Start and end dates

Discrete time

A point in time

Snapshot

Kimball

Slowly Changing Dimension Management

Define data management via versioning

Type I

Change record as required

No History

Type II

Manage all changes

History is recorded

Type III

Some history is parallel

Limit to defined history

Inmon	Kimball
Subject-Oriented Integrated Non-Volatile Time-Variant	Business-Process-Oriented Stresses Dimensional Model, Not E-R
Top-Down	Bottom-Up and Evolutionary
Integration Achieved via an Assumed Enterprise Data Model	Integration Achieved via Conformed Dimensions
Characterizes Data marts as Aggregates	Star Schemas Enforce Query Semantics

	Inmon	Kimball
Overall approach	Top-down	Bottom-up
Architectural structure	Enterprise-wide DW feeds departmental DBs	Data marts model a business process;enterprise is achieved with conformed dims
Complexity of method	Quite complex	Fairly simple
Data orientation	Subject or data driven	Process oriented
Tools	Traditional ERDs and DIS	Dimensional modeling; departs from traditional relational modeling
End user accessibility	Low	High
Timeframe	Continuous & Discrete	Slowly Changing
Methods	Timestamps	Dimension keys