# VOD Datawarehouse (DWH) Design

In order to create as broad a basis for reporting from various sources there may be no filtering in the DWH on data, up till the reporting layer. To achieve this, the DWH is composed of 4 layers. In the first 3 layers data from the various sources is collected and monitored. To be able to monitor and log changes the data is not being combined.

In the first 3 layers the data is stored based on the source. Every source has its own section in the DWH named a Schema. This way the source can easily be identified.

## Layer 1

Layer 1 is the **Source layer.** The tables have the prefix SRC. In this layer data from text files and other databases than SQL Server databases is imported. There is no filtering and data is not changed in any way. For example a number that appears as text in a text file is imported as text. This way no data can be lost.

## Layer 2

Layer 2 is the **Staging layer**. The tables have the prefix STG. In this layer data from the SRC layer is imported, while the data is converted to the correct data type. Numbers stored as text in the SRC layer are converted to numbers (for example integer). Data that does not conform to the data type specifications is filtered. For example an empty line without ID or a numeric value containing alphanumeric characters. There is no contents check on the data, so data that is compliant with the data type, but has values that shouldn’t exist in the application, is included.

## Layer 3

Layer 3 is the Data Warehouse layer. The tables have the prefix DWH. In this layer the latest version of the data is stored including all changes that have occurred to this data. Historical information is stored in a table with the same name but preceded by the prefix *audit\_*. Data going from the STG layer to the DWH layer is being processed, taking into account data already present. From the STG layer to the DWH layer only those data-columns are taken that are being used in the reports including any data needed to process that data. There is no filtering on the contents of the data itself. Data columns that are not being used are left out completely to maintain the DWH at a manageable size. For instance because the corresponding fields in SIMS are not being used,

## Layer 4

Layer 4 is the DataMart layer. The tables have the prefix DMT. In this layer the data is transformed to direct reportable data. The data from the DWH layer is combined and calculated so data that belongs together or should be in the same report is collected into a single table or table group. In this layer there will be several tables that are a direct representation of a report. Filtering on this data will occur on the reporting level, which means that faulty data (for instance data entered wrongly by a user) will be in this layer.

# Sources of the DWH

Detailed information about which data(fields) come (s) from which sources is documented in VOD\_DWH3.vsd (Microsoft Visio 2010)

In the DWH the following sources are being used:

1. SIMS
2. VODORA
3. CDS
4. TVA files
5. VAF VOD Archive Files
6. VAF Audio information files
7. ADI files

SIMS and VODORA are SQL Server databases and don’t need data-type correction. Their contents are directly copied to the Staging layer. The other sources are copied to the Source layer first.

## SIMS

Storer Information Management System.  
This is the application that contains the information about Licensing and Scheduling. Most of the information about these subjects comes from SIMS  
SIMS use a SQL Server database that can be read directly by the DWH. Data conversion is not required. The tables that have actually been used in reports so far are the only ones collected from SIMS. Additional tables can be added. Of these tables only the columns that contain data that has been used are being imported into the DWH, including all associated data. There is no filtering on the data level itself.

## VODORA

VOD Order Application.  
This is the application that contains ordering(ordered) and delivery(received) information. VODORA uses SIMS as a basis for its information.  
VODORA uses a SQL Server database. From this database 1 table is being retrieved at the moment. On this table there is no filtering being applied whatsoever.

## CDS

Content Delivery System.  
CDS is the system that sends all content to the clients of Chellomedia.  
CDS uses a PostGre SQL database that is not completely accessible to the DWH. CDS produces an XML export file on regular intervals that is being imported by the DWH. 1 table is accessible and this one is being copied to the SRC layer of the DWH. There is no filtering on columns or data.

## TVA

TV Anywhere.  
These are XML files that are being delivered by external parties, containing information about what is on their platform.   
The TVA files cannot be imported into the DWH directly. They are first being processed by a script that runs in FA (File Automation) which filters out the relevant data and creates a more readable format for the DWH. The FA script can be adjusted to extract more information from the TVA file as required.

## VAF VOD Archive Files

These are the “physical” files that are being stored on disk in the archive. These include Movies, Trailers, Posters and Boxcovers.  
The application FileCrawler reads all archives and creates a list of all files found. This list is imported into the SRC layer of the DWH directly by FileCrawler. The STG layer filters out all files that do not have the correct file extension. ('.ts','.png','.jpg','md5')

## VAF Audio information files

These are XML files being created by the application MediaInfo. Every movie or trailer created is checked by MediaInfo and information about the videofile and audiotracks is stored in an XML file that can be imported by the DWH.

## ADI files

ADI files contain information that is being sent by SIMS or IBMS to CDS.  
SIMS creates an export file that contains all relevant information about a movie and additional relevant data needed to send a movie title to a client successfully. This export file is an ADI file. This file is first being processed by FA that splits the file in 2 files: An ADI offer file and an ADI content file. Both files can be imported into the DWH.

# Datawarehouse buildup and construction.

Every source in the DWH has its own Schema (NameSpace) that holds all objects (tables, views, stored procedures, functions) related to this source. The Schema dbo holds all obeject that are general or multi-schema. The DataMart layer is in the dbo schema to facilitate reporting.

## Naming conventions

Every object in the DWH has a specific name, based upon its function or location in the data flow.

* Stored procedures always start with the prefix usp (user stored procedure)
* Stored procedures that import XML files are called: usp\_Import<Source>XML
* Stored procedures that read data from one layer and fill the table in the next layer have the name: usp\_<TargetTable>\_Fill
* Views start with the prefix
  + vw for views on source tables. Source tables are tables in Source databases that can be connected to directly by the DWH
  + vw\_STG for views on Staging tables
  + vw\_DWH for views on DWH tables
  + vw\_FLAGS for views on DWH tables that hold Flags for SIMS. Every type of Flag collection has its own view
* Tables start with the prefix
  + SRC for Source tables. These hold source data without any conversion.
  + STG for Staging tables. These hold data with corrected data types.
  + DWH for DataWareHouse tables. These hold filtered data on the column level (less columns) to filter out unused columns.
  + Audit\_DWH. These hold auditing data for the DWH tables. Every Insert, Update or Delete is registered.
  + DMT for Data Mart Tables. These hold all reporting data. This data is combined, enriched and filtered.

## Execute sequence

The Stored Procedure usp\_ ProcessData runs the following procedures:

* usp\_SRC\_Fill runs all stored procedures that fill SRC tables, including Import usp’s.
* usp\_STG\_Fill runs all stored procedures that fill STG tables.
* usp\_DWH\_Fill runs all stored procedures that fill DWH tables.
* usp\_TVA\_Fill runs all stored procedures that fill TVA tables.
* usp\_DMT\_Fill runs all stored procedures that fill DMT tables.

# Filters before the DWH Layer

TVA. usp\_STG\_TVA\_Attributes\_Fill  
WHERE [AudioLanguage] IS NOT NULL

AND [Prid] IN (SELECT DISTINCT [provider\_id] COLLATE Database\_default

FROM [SIMS].[vw\_pitch\_contract])

The Audiolanguage must have a value and the provider Id must exist in SIMS. If the Provider ID does not exist in SIMS, the value is (probably) from another vendor than ChelloMedia.

VAF. usp\_STG\_VODArchiveFiles\_Fill  
WHERE [FileExtension] IN ('.ts','.png','.jpg','md5')

The file extension must be a specific value, other files are not monitored. If a new file extension is required, it should be added here.

<>

# Overview of common fields throughout the various sources

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Workflow | SIMS | VODARA | VAF | ADI | CDS | TVA |
| Asset ID | Asset\_id\_prefix + asset\_reference |  |  | Asset\_ID | assetid | Id |
|  | Clearance\_id | ClSimsId |  |  |  |  |
| FileName | File\_name (media\_detail) |  | FileName | MovieFile  PosterFile  TrailerFile  BoxCoverFile | filename |  |
| Component Language | Ac3audio.language  Caption.language |  | TrackLanguage | Audio\_Type  Languages  trailerAudio\_Type  TrailerLanguages  Subtitle\_Languages |  | AudioLanguage  CaptionLanguage |
| Viewing Period  Maximum Viewing Length | Checkout\_length |  |  | Maximum\_Viewing\_Length |  | ViewingLength |
| License Start | Start\_date (contract\_clearance\_ window) | ClStartDate |  |  |  |  |
| License End | End\_date (contract\_clearance\_ window) | ClendDate |  |  |  |  |
| Version major | Version\_major (pitch\_title) |  |  | Version\_Major | Version\_major |  |
| Version minor | Version\_minor (pitch\_title) |  |  | Version\_Minor | Version\_minor |  |
| Clearance State | Cleacance\_status | ClStatus |  |  |  |  |
| Territory | Territory | TeName |  |  |  |  |
| Title name | Name (title / name\_sets) | TiName |  | Title\_Name | Assetname |  |
| Series Name | Name (series / name\_sets) | TiSeries |  |  |  |  |
| Distributor | Name (contract / Company) | CoProvider |  |  |  |  |
| Clearance on date | Cleared\_on | ClClearenceDate |  |  |  |  |
| Orion | Orion Clearance / Excluded |  |  |  |  |  |
| SD/HD | Is HD (Flags\_TitleHD) |  |  | HD |  |  |
| Box office category | Box office category |  |  | Category |  |  |
| Language Rights | Language Rights |  |  |  |  |  |
| Genre | Type\_code\_name / TitleGenre |  |  | Genre |  |  |
| COD | COD |  |  |  |  |  |
| Contract | Name (contract) | CoName |  |  |  |  |
| IBMS/SIMS (Source?) |  |  |  |  |  |  |
| Customer | Custom Flag (tbd) |  |  |  |  |  |
| Media Name / ID | Name (media) |  |  |  |  |  |
| Components | Media\_ac3audio  Media\_caption |  | VODTrackInfo |  |  | TVA\_Attributes |
| Media created on date |  |  | DateCreated  DateModified |  |  |  |
| Ordered date |  | ClForstOrderenOn  ClMostRecentOrderedOn |  |  |  |  |
| Media Type | Name (media\_type) |  |  |  |  |  |
| Received Date |  | ClSourceDeliveredOn |  |  |  |  |
| Platform |  |  |  |  | Locationname |  |
| All media complete date |  |  |  |  |  |  |
| Schedule Start Date | Start\_date (pitch\_title\_window) |  |  | Licensing\_Window\_Start |  |  |
| Schedule End Date | End\_date (pitch\_title\_window) |  |  | Licensing\_Window\_End |  |  |
| Adult Flag | Adult\_flag (titles)  Adult\_flag (contract) |  |  | Advisories (18) |  |  |
| Pitch contract | Name (pitch\_contract) |  |  |  |  |  |
| Pitch Status | Status (pitch) |  |  |  |  |  |
| Price | distribution\_price |  |  | Suggested\_price |  |  |
| FA State |  |  |  |  |  |  |
| Delivery Date |  |  |  |  | Finished |  |
| provider | provider |  |  | Provider |  | scope |
|  |  |  |  |  |  |  |