CDW Guide

**Introduction & Policies**



Business Intelligence Service Line

U.S. Department of Veteran Affairs

September 2016

**Contents**

[Introduction 1](#_Toc361660365)

[What is the Data Warehouse? 2](#_Toc361660366)

[Who are Data Warehouse Customers? 2](#_Toc361660367)

[Customer Responsibilities 2](#_Toc361660368)

[Server Resources 3](#_Toc361660369)

[Other Resources 3](#_Toc361660370)

[Authorizations 4](#_Toc361660371)

[Data Warehouse Policies and Guidelines 5](#_Toc361660372)

[Policies Introduction 5](#_Toc361660373)

[Policy Exceptions 6](#_Toc361660374)

[Maintenance Windows 6](#_Toc361660375)

[Queries 6](#_Toc361660376)

[New Data Releases 7](#_Toc361660377)

[Database Storage: xDWWork and Project 7](#_Toc361660378)

[Projects 9](#_Toc361660379)

[Data Movement 10](#_Toc361660380)

[Use of CDW-Raw Data 11](#_Toc361660381)

# 

# Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This document is part of a series of CDW Guide documents.

CDW Guide: Introduction and Policies *(current document)*  
CDW Guide: Workgroups and Development

CDW Guide: Obtain Report Builder Software

CDW Guide: Using Page Level Compression  
CDW Guide: Query Best Practices

## What is the Data Warehouse?

The Business Intelligence Service Line (BISL) is the section of the Office of Information and Technology (OI&T) that maintains a set of servers and other IT resources that provide clinical and other data to VA for analytical purposes. Each server contains a national data set (CDW), a regional data set (RDW) or a VISN data set (VDW). The servers contain the same data model (same table structures), are loaded by the same processes and the regional and national environments have the same permissions model. Therefore the term “Corporate Data Warehouse” is often used to mean the system overall as the metadata and other documentation apply to all environments.

The data maintained on the servers are in large relational databases organized by logical data domain, e.g. pharmacy, lab chemistry, etc. The data comes from the source systems (mostly Vista) without “scrubbing”, filtering and without business logic applied. The data is not optimized for any particular use but rather it is fairly normalized to allow its use by varied types of queries for analyses and reports.

## Who are Data Warehouse Customers?

For the context of this document Data Warehouse (DW) customers are individuals that have direct permissions on one or more of the national or regional data warehouse servers. (Those using VDW servers could also take advantage of this document).

Customers are authorized by either National Data Systems (NDS) or through a local authority (LSV).

DW customers are data analysts, researchers, report and software developers and anyone using the SQL language (or other query methods) to query a database directly. It is assumed that the customers have sufficient technical and business data knowledge to use the large data sets that are not yet “packaged”. This document does not discuss the reports, datamarts, analyses or data sets generated from the DW. The consumers of those products are not the intended audience of this document.

Researchers have different permissions than operations personnel. Therefore some policies and best practices may be handled differently.

## Customer Responsibilities

Customers are expected to read this document and adhere to the policies in it. Customers are expected to read the other documents in this series and adhere to the practices in them.

Customers are expected to obtain the technical understanding required to write queries that do not violate security/privacy agreements or significantly usurp the DW resources. Customers are expected to adhere to the principals in the CDW Guide: Query Best Practices document. While some training is available through BISL, other training may be necessary. Customers are responsible for assuring that their queries do not unduly tax the resources of the database server.

In most cases customers should operate in the context of a workgroup (see CDW Guide: Workgroups and Development). Customers are encouraged to query the data on the Data Warehouse server and store small result tables in their workgroup database, rather than transfer the data to another server. To accommodate larger result tables, customers may request additional resources through BaseCamp. A request for an exception to CDW policy may also be required.

Data transfers from the Data Warehouse environment must be done in accordance with NDS policy. Under no circumstances should data be transferred to a laptop or desktop computer. If the data must be moved to another database, all appropriate VA rules must be followed, such as only storing the data on a secure OI&T server behind the VA firewall.

## Server Environments

BISL maintains CDW databases for the VA enterprise. The majority of the data is health data from VHA. The data is routinely extracted from local medical centers (VistA) and other data sources, transformed into a single integrated data model and distributed to several access points for data consumption. The CDW servers which contain the data and resources to support use of the data are organized into enclaves (about half dozen at time of this writing). Each enclave contains one or more servers and is intended for a particular customer group. They all use the same architectures for data, databases, security, distribution, support, etc., but may have some variations that are specific to the intended audience. All enclaves have at least one SQL Server Database Engine server. The resources/services that BISL offers (ancillary to the data itself) is provided to a workgroup and each workgroup operates in the context of an enclave. Customers must know the enclave in which they will operate. Following is a list of the existing Enclaves:

* **GPE (General Purpose Enclave)** Most people would choose this one unless they belong to one of the specific customer groups below. Their supervisors should be able to advise them. The main SQL Server is VHACDWA01.
* **BIE (BI Enclave)** This one is reserved for staff of OABI.
* **VINCI** All resources for research projects are provisioned in this enclave
* **FRE (Field Reporting Enclave)** This is intended for the wide range of activities by the staffs at the District, VISN and facility levels. However it contains the full complement of national data. At the time of this writing Region 4 and its VISNs are in the process of migrating their work to this enclave.
* **NDE (Non-CDW Data Enclave)** Used by groups that use only their own data, this enclave does not have a full copy of CDW data.

**Additional Server Environments (not considered enclaves)**

The Regional (RDW) and VISN (VDW) environments are intended for use by only Region and VISN staffs. They contain only the relevant subsets of the entire national data set. Types of SQL Servers

* Production. Most servers are this type. This data is refreshed nightly. Data has been modeled, meaning it has been structured to allow flexibility in querying for any purpose and to have consistent naming and patterns to aid in usability. The data model is published at [Metadata](https://vaww.cdw.va.gov/metadata/default.aspx). There are several production servers each having a copy of the full data warehouse.
* CDW Raw. One server, VHACDWA06, which contains a set of data extracts that are not integrated with Production data. More information can be found at [Metadata for Raw](https://vaww.cdw.va.gov/bisl/Database/SitePages/Raw%20Extractor.aspx). CDW Raw contains data domains that are not (yet) on the production servers. With a couple of notable exceptions (e.g. DSS) the data is subject to removal based on resources available, current priorities and if that data is put into Production.  Developers should be aware of this eventuality before spending the effort to develop software that uses this as a data source. Not all BISL services are offered for this server.
* Static. These are copies of the full production data warehouse as it was on a day in early October (a few days after of the fiscal year ends). There are copies taken at the end of FY2012, FY2013, FY2014 and FY2015. They each contain all data up to that day. They will never be refreshed past that date. CDW keeps a copy of these static data sets as they can be useful for analysis and research.
* Development. These servers do not necessarily have a full copy of the data warehouse nor is it automatically refreshed every night. Rather they have sampled data, enough for a Workgroup to use as they do database development work.

## Other Resources

Customers should access the database with Microsoft SQL Server 2014 Client Tools for interactive sessions. SQL Server Management Studio (SSMS) typically provides the fastest access. However, you can use other software, if it has an ODBC driver for SQL Server. If working in the VINCI environment, only the tools provided can be used.

|  |  |
| --- | --- |
| **Resource** | **Description** |
| [CDW SharePoint site](https://vaww.cdw.va.gov/Pages/CDWHome.aspx) | Announcements, status, support, special interest groups, BISL Tech Teams |
| [CDW Metadata](https://vaww.cdw.va.gov/metadata/default.aspx) | Includes diagrams, data dictionary |
| SQL Server 2014 Client Tools | Recommended. Contact your local IT support. |

## Authorizations

There are two methods for gaining authorization to access data directly in the DW described in [How to Access the CDW](https://vaww.cdw.va.gov/support/Shared%20Documents/How%20to%20access%20the%20CDW_Jan2015.docx) .

There are differences in what data sets each of these mechanisms authorizes.

1. National authorization through National Data Systems (NDS). This type of authorization is relevant for access to CDW and RDW environments only.
2. Local authorization (LSV) (relevant for CDW, RDW and VDW environments). The LSV method authorizes access to all data from one or more stations (VHA medical centers).

The following table shows from which database a customer will operate, depending on the environment in which they are working and the authorization they have. However, all the DW views have the same structure; therefore a query should work the same in all the databases.

|  |  |  |
| --- | --- | --- |
| **Environment** | **National Authorization (NDS)** | **Local Authorization (LSV)** |
| **National(CDW)** | CDWWork | LSV |
| **Regional (RDW)** | RDWWork | LSV |
| **VISN (VDW)** | n/a | LSV |



# Data Warehouse Policies and Guidelines\_\_\_\_\_\_\_\_\_\_\_

## Policies Introduction

The VA Data Warehouse, typically called CDW, is a set of IT resources shared across the VA enterprise. The Business Intelligence Service Line (BISL) within the Office of Information and Technology is charged with its development, management and protection. The resources include servers, memory, disk storage, disk access (I/O), CPU usage and network bandwidth. Additionally, human resources are required to maintain these assets, to provide security and to support customers. All these resources at various times are limiting.

The CDW architecture reflects a self-service approach to allow customers the greatest flexibility. Operation customers are given direct access to data to construct queries and analytical data sets most appropriate to their requirements. Research customers are granted permissions to a data set (sometimes large) specifically prepared for that research project by VINCI staff.

In the short history of CDW, its use has been adopted quickly, now supporting thousands of customers. The scope of its use has also expanded, now supporting many software applications. Database sizes continue to grow as BISL staff add more data content and more years of data.

As a result of these factors, there is increasing pressure on the IT resources underlying the system. These pressures and the open architecture require customers be informed and conscientious about making good use of the limited and shared resources.

The policies in this document are to minimize security risks, protect IT resources and optimize the system’s use for all customers. Guidelines are general recommendations made to customers so they can easily adhere to policy. The Best Practices sections give more details and are more prescriptive in outlining the steps of an organized, efficient and secure process.

## Policy Exceptions

Customers can request a one-time exception from specific policies.  For example, if a customer believes it is necessary for their query to run longer than the time allowed by policy, they can request and present their justification.  They should do so by sending an email to [CDWProjectSupport@va.gov](mailto:CDWProjectSupport@va.gov) that includes the following information.

* Justification for exceeding these guidelines
* User / App or ETL account that the query or process will be run under
* Exact name of the program used to run the query or process
* DateTime window of when the query or process is expected to start and finish
* Contact name, email and phone number

## Maintenance Windows

The hours during which customers should plan their work on the Data Warehouse servers are Monday through Friday 8:00 AM – 6:00 PM, for the time zone of the server on which they are working. Outside these times, other processes of data loading, backups and database maintenance must take precedence. Weekends in particular are heavily used for maintenance operations.

Policy: If customers’ queries interfere with critical BISL processes they will be terminated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Region 1 (District 04 & 05) | Region 2 (District 03) | Region 3 (District 02) | Region 4 (District 01) | National (CDW) |
| Pacific time | Central time | Eastern time | Eastern time | Central time |

## Queries

The main factor governing how long it takes a query to finish is usually the time required to read from or write to disk storage. If one query takes 30 minutes to run, it means that during those 30 minutes other people’s queries will run slower. Badly designed queries can be very costly to overall performance for everyone.

If customers need to extract a large portion of a fact table, they should review the [Incremental Extracts Guide](https://vaww.cdw.va.gov/support/Shared%20Documents/Best%20Practices/CDW%20Guide%20to%20Incremental%20Extracts.docx) located on the [CDW Support SharePoint portal](https://vaww.cdw.va.gov/support/SitePages/CDWSupportHome.aspx) for help with implementation of an incremental extract. To move data from one server to another it is recommended to use SQL Server Integration Services (use of SSIS server can be requested from BaseCamp).

**Guideline A:** Customers should view the SQL Server Estimated Query Plan before actually running any query to get an estimated number of rows that will be returned by a query and the overall Query Cost.

* Policy 1a: Queries or processes running for more than 5 hours are subject to being automatic termination without notice.  Most processes should complete in much less time.
* Policy 2a: Queries that are estimated to return more than 500,000,000 rows are subject to being automatically terminated without notice.
* Policy 3a: Queries that are estimated to return more than 250,000 rows to a GUI query tool are subject to automatic termination without notice.  It is not reasonable that a person can read that many rows. This would include tools such as SSMS, Toad, MS Office tools such as Word & Excel but also to include PowerBI, PowerPivot, PowerView and Tableau.
* Policy 4a: Queries on an xDWWork fact table that have no WHERE clause or other effective means of filtering are subject to automatic termination without notice
* Policy 5a: Queries that are obviously malformed (e.g. incorrect join criteria) are subject to being automatically terminated without notice.
* Policy 6a: Connections idle for longer than 48 hours are subject to being automatically terminated without notice.
* Policy 7a: Queries joining two or more tables via linked servers without using SQL PassThru are subject to being automatically terminated without notice.
* Policy 8a: Queries with a Query Cost over 300K are subject to being automatically terminated without notice.  A Query Cost of over usually 300K indicates there is a problem with the query itself or the data being joined.
* Policy 9a: Connections idle for longer than 5 hours with an OPEN Transaction are subject to automatic termination without notice.
* Policy 10a: Connections that are idle for more than 5 hours and have more than 10GB of tempdb space reserved are subject to automatic termination. TempDB is a shared and limited resource that must be cleaned up in a timely manner.
* Policy 11a: Connections that are active and have more than 400GB of tempdb space reserved are subject to automatic termination.  Allocations of this size are an indication there is a problem with the process. TempDB is a shared and limited resource that must be cleaned up in a timely manner.

## New Data Releases

**Guideline B:** As a best practice, Microsoft recommends database developers to always create a set of data source views on which to base their applications, rather than query the source system objects directly. BISL recommends customers follow this practice. It gives customers greater protection against changes on the source system. Should a change occur to a view in xDWWork or SPV, it might be sufficient that the database developer only make a change to their view, rather than fixing potentially multiple references to the CDW view in their source code.

Policy 1b: BISL will attempt to give four weeks’ notice of any changes to data objects that could cause customers’ processes to stop working. Changes are most frequently field names and data types. Adding fields to an existing view would not be considered a change that would break existing queries and would not necessarily prompt an advance notice.

## Projects

**Guideline C:** Software developers should read and adhere to CDW Guide: Workgroups and Development.

Policy 1c: BISL staff will work with the customer to guide them to an approach that will meet the customer’s need, maintain security and optimize use of the resources. It is expected that, working together, such a solution can always be found. However, ultimately the BISL staff have authority to revoke an individual’s permissions should they deem it necessary to protect the resource or they believe there is a usage that is outside the security or privacy policies or is otherwise inappropriate.

**Guideline D:** Schemas are the means that Microsoft provides for managing permissions. Using schemas for the purpose of managing/organizing objects inevitably leads to conflicts. BISL recommends that projects should minimize the number of schemas in the project database and instead use naming conventions on the objects to organize. For example use a prefix separated by an underscore “\_” instead of a different schema.

Policy 1d: There are several independent security clearances that a project can have. A project must have at least one security clearance. These clearances correspond directly to the NDS authorization categories for individuals.

* If a project is cleared for **Contains CDW Basic** then all the members of the project must have been authorized by NDS for Basic Read Access. If the project does not have this clearance then all project members agree not to expose any CDW data through any project resources, including database, file share, application, even to other project members.
* If a project is cleared for **Contains Patient SSN** then all the members must have been authorized by NDS for Privileged Read Access. If the project does not have this clearance then all project members agree not to expose SPatient data through any project resources, including database, file share, application, even to other project members.
* If a project is cleared for **Contains Staff SSN** then all the members must have been authorized by NDS for Staff Real SSN. If the project does not have this clearance then all project members agree not to expose SStaff data through any project resources, including database, file share, application, even to other project members.
* If a project is cleared for **Contains SAS** then all the members must have been authorized by NDS for Med SAS data. If the project does not have this clearance then all project members agree not to expose data derived from the SAS files through any project resources, including database, file share, application, even to other project members.
* If a project is cleared for **Contains DSS** then all the members must have been authorized by NDS for DSS data, prepared by DSO. If the project does not have this clearance all project members agree not to expose DSS data through any project resources, including database, file share, application, even to other project members.
* If a project is cleared for **Contains Vital Status** then all the members must have been authorized by NDS for Vital Status. This is because part of that data comes from Centers for Medicare and Medicaid and they have their own rules for access. If the project does not have this clearance then all project members agree not to expose VitalStatus data through any project resources, including database, file share, application, even to other project members.

Policy 2d: App utility accounts are for use by software applications developed by the project members, subject to the security review by NDS.

**Guideline E:** The customer can use sp\_spaceused2 (see above) to discover the disk storage used by the project database.

* Policy 1e: Project databases have a default maximum of 80 GB disk storage. Customers must provide justification to exceed this limit.

**Guideline F:** When project members create schemas they should make the owner to be u<ProjectName>, not the default owner (themselves). See CDW Guide: Workgroups and Development.

* Policy 1f: Project members are only permitted to create schemas owned by the project user u<ProjectName> or by their own login (not recommended). Schemas having other owners such as dbo or u<ProjectName>\_Own are created by BISL staff for special purposes (see CDW Guide: Projects and Development)
* Policy 2f: The permissions granted to each App account are constrained to one location, the App schema in the project database (see CDW Guide: Workgroups and Development).

## Data Movement

Sometimes it is necessary to move data between databases or servers.

**Guideline G:** If the customer wants to periodically copy a set of records to another server or to otherwise maintain a “datamart”- type data set, they should consider using an incremental extract method rather than generating data sets that are largely redundant with the previous data set. This can be done with SPV views. It requires more setup work and expertise but is significantly more efficient than extracting and replacing large volumes of data on the target database.

* Policy 1g: ETL utility accounts, if authorized, are for use to move data between databases or servers under the close management of a project member. Under no circumstances are they to be used for data access by applications executed by non-members of the project. Violation of this policy will result in immediate revoking of the account.
* Policy 2g: The use of linked servers is not supported in the Data Warehouse. Joining across servers on huge tables is not feasible and can easily prevent all customers from performing their work. To simply move data from one server to another we recommend using SQL Server Integration Services from Microsoft.

## Use of CDW-Raw Data

**Guideline H:** Raw data can be used for data exploration and for analysis. Automated processes developed against CDW-Raw data (e.g. reports) should be avoided unless the customer is willing to accept that the effort is for a temporary benefit, depending on the domain.

* Policy 1h: Data tables on the CDW-Raw server are refreshed with various frequencies, depending on the domain. With some exceptions, notably DSS and Surgery, data extracts to the CDW-Raw server may be discontinued or delayed, depending on current priorities, available resources or data becoming available in the Production environment.

**Guideline I:** While use of CDW-Raw data by itself can be useful, attempting to match it to production data sets is difficult at best due to the following challenges

* Timing. CDW-Raw data is days to months old. Data edits, additions and deletions on the Vista source systems after an extract will not appear in the CDW-Raw system until the next extract but will be reflected in the Production data, usually the next day.
* SSN as key field (DSS only). The CDW-Raw data is based on SSN (or scrambled SSN) as a unique identifier. This is a poor identifier because it is sometimes incorrect. When corrected in the Vista source this results in what appears as missing records in Production or in CDW-Raw.
* Data movement. Because CDW-Raw data and Production data exist on separate servers there is always a step required to bring them together on the same server. For large tables this can be daunting.
* Difficult joins. Because CDW-Raw data does not have surrogate keys (integer-based SIDs) and the tables are minimally indexed, joins across large tables can be prohibitively complex.

Policy 1i: If a data domain on the CDW-Raw server is developed as a Production domain, the data will be eventually removed from the CDW-Raw environment. The structure of the data will always be different in Production than it was in CDW-Raw.