Perceptive Reach

Integrated Reach Database System

(IRDS)

Database Design Specification



Department of Veterans Affairs

December 2014

Version 1.0

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 12/11/2014 | 1.0 | Final Review | Monica Mohler/Paul Bradley |
| 12/10/2014 | .02 | Peer Review | Radina Ivanova |
| 12/09/2014 | .01 | Initial Version | Bill Balshem |

Table of Contents

1. About this document 5

2. Introduction 5

2.1. Purpose 5

2.2. Scope, Approach and Methods 5

2.3. Acronyms and Abbreviations 6

3. System Overview 7

3.1. Business Process 8

3.2. System Information 10

3.2.1. Hardware Requirements 10

3.2.2. Support Software 10

3.3. Architecture 11

3.3.1. Software Architecture 11

3.3.2. Interfaces 12

3.3.3. Data Stores 12

4. Database Design Decisions 12

4.1. Assumptions 13

4.2. Issues 13

4.3. Constraints 13

5. Database Administrative Functions 13

5.1. Naming Conventions 13

5.2. Database Identification 13

5.3. Schema Information 14

5.3.1. Description 14

5.3.2. Logical Data Model 14

5.3.3. Physical Data Model 16

5.3.4. Data Dictionary 17

5.3.5. Veteran De-Duping Process 17

5.4. Denormalization 19

5.5. Performance Improvement 19

5.6. Storage 19

5.7. Recovery 19

6. Database Interfaces 19

6.1. Suicide Data Repository (SDR) 19

6.2. Perceptive Reach Dashboard 19

7. Data Access 20

7.1. Role Definitions (System) 20

7.2. Role Definitions (Dashboard) 21

8. Implementation Considerations 21

8.1. Large Objects 21

8.2. Partitioning 21

8.3. Error Processing 21

# About this document

This document is a “work in progress.” The database will be designed through a series of sprint cycles in collaboration with VA stakeholders and users. The document will be populated with content as the design evolves with each sprint, including all potential data options, configurability rules and options. The early versions of the document will focus on the conceptual design, key business requirements, and the primary end-user features related to the design. Future versions will add additional detail as it becomes available.

# Introduction

VA is seeking to expand suicide prevention to include upstream approaches, designed to reduce initiation or escalation of a risk factor. Upstream suicide interventions target individuals or groups who exhibit biological, psychological, or social risk factors that are more prominent among high-risk groups than among the larger population. Understanding the unique needs of our nation’s Veterans and the military culture as it relates to stigma and mental health is important for early intervention. The goal of the Integrated Reach Database System (IRDS) innovation is to promote the general health of the Veteran population and effectively intervene in issues before they escalate in crisis.

The IRDS solution innovates the current process of risk data collection, analysis, and use in effective intervention strategy. The solution will harness the power of large and diverse data stores to aggregate, analyze and identify risk onset as well as reveal previously unidentified at-risk individuals and populations as a holistic and integrated approach.

The IRDS innovation will serve to bolster the three major components of VHA’s Strategic Plan for Suicide Prevention: surveillance, risk and protective factors, and prevention interventions. The IRDS innovation will target antecedent events specific to Veteran populations prior to the onset of risk to mitigate the development of risk.

## Purpose

The purpose of the Database Design Specification is to give detailed documentation of the Data and Database components of the IRDS system.

## Scope, Approach and Methods

This document covers the following items:

* Database Properties and Configuration
* Database Conceptual Design
* Database Objects and Table Schemas
* External Data Sources and Interfaces

NOTE: This is a working document. As access to data sources becomes available and technical requirements are further defined, the document will be updated with details such as, database configuration, indexing and the types of data transformations that will take place when importing data from sources into the IRDS system.

## Acronyms and Abbreviations

**Table 1: Acronyms and Abbreviations**

| Acronym | Term |
| --- | --- |
| API | Application Programming Interface |
| CDW | Corporate Data Warehouse |
| DOB | Date of Birth |
| EDW | Enterprise Data Warehouse |
| ETL | Extract, Transform, Load |
| GB | Gigabyte |
| HDD | Hard Disk Drive |
| ICD | International Classification of Diseases |
| IRDS | Integrated Reach Database System |
| JSON | JavaScript Object Notation |
| KNIME | Konstanz Information Miner |
| MS | Microsoft Server |
| NDI | National Death Index |
| POC | Point of Contact |
| RAM | Random Access Memory |
| RPC | Remote Procedure Call |
| SDCD | State Death Certificate Data |
| SDR | Suicide Data Repository |
| SFTP | Secure File Transfer Protocol |
| SPAN | Suicide Prevention Applications Network |
| SQL | Structured Query Language |
| SSIS | SQL Server Integration Services |
| SSMS | SQL Server Management Studio |
| SSN | Social Security Number |
| TBD | To Be Determined |
| TBI | Traumatic Brain Injury |
| T-SQL | Transact-SQL |
| VA | Department of Veterans Affairs |
| VAMC | VA Medical Center |
| VBA | Veterans Benefits Administration |
| VCL | Veterans Crisis Line |
| VHA | Veterans Health Administration |
| VISN | Veterans Integrated Service Networks |
| VistA | Veterans Health Information Systems and Technology Architecture |
| VLER | Virtual Lifetime Electronic Record |
| VR&E | Vocational Rehabilitation and Employment |

# System Overview

The main components of the IRDS system are:

* **Reach Database -**  A SQL Server database on the IRDS server that stores data from the various VA sources imported into the system
* **Analytics Risk Model/Surveillance Model** – A program written in R that will run periodically to update a list of high risk factors associated with Veteran suicide. These factors will be persisted in one or more tables in the Reach database. A SQL process will run on a regular basis to monitor Veterans as being high risk for suicide based on factors as determined by the Analytics Risk model
* **Direct Messaging** -. As Veterans are identified by the surveillance model as being high risk, the VA staff will be notified via the direct messaging component of the IRDS system, which leverages the VLER solution adopted by the VA
* **Perceptive Reach Dashboard** – VA staff can login to the dashboard to read messages and review data associated with their regional or functional responsibility within the VA (i.e. region, facility)

## Business Process

1. Data sources will be imported into the IRDS system via SQL Server Integration Services (SSIS) import solutions. Each data source will have its own SSIS package (.dtsx). The import solution will transform the data and load it into the appropriate tables in the Reach database.
2. An R program will be run periodically (every year or so) on the production server, that uses the Reach data as input, to update the Risk Model. The results of the run are stored in one or more tables in the Reach database.
3. On a regular basis (daily, weekly) a SQL Server process will run that does surveillance against a list of Veterans tracked in the Reach database tables, against the variables in the Risk Model. A Risk score is calculated for each Veteran and that value is placed in the record for that Veteran in the ‘Veteran’ table, a master list of all Veterans imported into the IRDS system. All Veterans whose score is equal or above a threshold score will be defined as high risk for suicide.
4. For each Veteran identified as a high risk for suicide, the system will notify appropriate VA staff member(s) through a secure message, leveraging the VA Virtual Lifetime Electronic Record (VLER) architecture.
5. A staff member opens up their dashboard via a compatible web browser and a client side java component connects to a server side java component, which queries the Reach database for both specific and aggregate data regarding high risk Veterans at the users’ management level (region, state, VISN, VAMC). The query results are passed to the client browser and populated in the web page.

**Figure 1: Business Process Diagram**

## System Information

### Hardware Requirements

**Table 4: Hardware Requirements**

| Hardware Component | Requirements |
| --- | --- |
| Processor | Intel Xeon E5-2600 Family (2670 or 2690), 2.6GHZ |
| RAM/Memory | 64GB |
| Storage/HDD | 500GB |

NOTE: The IRDS is a virtualized solution; the table above is a list of resources required for the virtual solution.

### Support Software

**Table 5: Support Software**

| Product | Version | Purpose |
| --- | --- | --- |
| Microsoft (MS) Windows | 2008 R2 or later | Operating System |
| Microsoft (MS) SQL Server | 2012 | Database platform and data store |
| MS SQL Server Integration Services (SSIS) |  | Data import solution development tool Performs ETL work. |
| Transact SQL (T-SQL) |  | Query language native to SQL Server |
| Attachmate Reflection |  | SFTP software to download data files onto the IRDS server. |
| Python | 2.7 | Programming language. Used for Individual de-duping functionality |
| R | 3.1.2 | Programming language. Program will be developed in R to analyze Reach database data and develop Risk Model and store results in a SQL table. |
| Node JS | 0.10.33 | Tool to develop server side component of the dashboard. The dashboard sends requests to the Node JS, which queries the database and returns the result back to the client browser. |
| Knime | 2.10.4 | Data mining tool. VA staff can connect to the Reach database through Knime to do ad-hoc reporting. |

## Architecture

### Software Architecture

* **Data Imports** – SSIS will be the primary tool for importing external data sources into the IRDS Reach database. For a specific data import, an SSIS package will be developed to

1. Make a connection to the source (SQL table, text file, other)
2. Import the data into a staging area
3. Make the appropriate data transformations (cleaning, standardization)
4. Load the transformed data into the appropriate Reach data store tables

The execution of SSIS packages (.dtsx files) can be automated by scheduling them as a Windows process via SQL Server Agent.

* **Remote Procedure Calls (RPCs) –** The VA uses the Veterans Health Information Systems and Technology Architecture (Vista) system, for managing Veterans health data. Data will be imported into the IRDS system directly from VistA using RPC calls. VistA data is stored against a MUMPS back end, which uses text based files for data storage. For each set of VistA data imported into IRDS:

1. Either a custom RPC will be written (in M) or a currently existing one will be leveraged
2. An automated java process will execute the RPC and return the query results in text format
3. Those results will be stored in a flat file on the IRDS server to be imported into the reach database via an SSIS package

**Figure 2: IRDS Database Software Architecture**

### Interfaces

* **Corporate Data Warehouse (CDW)** – Data warehouse for VHA data. Records are stored in SQL Server tables.
* **Suicide Data Repository (SDR)** – Suicide and Mortality Data from four sources are stored on a centralized server in SQL Server tables. Those sources are: Mortality Search Results from the National Death Index (NDI), State Death Certificate Data (SDCD), Veterans Crisis Line (VCL), Suicide Prevention Applications Network (SPAN)
* **Veterans Information Systems And Technology (VISTA)** – Open source enterprise system used by VHA network. Data is stored in a MUMPS back end.
* **VA Enterprise Data Warehouse (EDW)** – Data warehouse that stores VBA data.

NOTE: At this time there is an attempt to include VBA data elements housed at EDW in the IRDS system. It is unclear whether access to the data will be available within timeframe set to develop the IRDS system.

* **Perceptive Reach Dashboard** – Users of the IRDS system will connect through a web based dashboard.
* **Analytics Sandbox** – VA users will be able to run ad-hoc reporting by connecting to the IRDS database using tools such as R or Knime.
* **Data Sources External to VA** – The IRDS system may potentially pull data from one or more outside sources such as LexisNexis or PACER.

### Data Stores

* All data stored in the IRDS system will be captured in SQL Server tables.
* In the event that some data will be accessed via a file format, such as FLAT files, those files will be uploaded to the IRDS server via SFTP and placed in a specified location on the IRDS server file system, to be processed for import.

# Database Design Decisions

The decision was made to use SQL Server as opposed to other data platforms such as MySQL or Microsoft Access because:

* Microsoft SQL Server is approved and is widely used within the VA
* Several of the systems from which data will be imported are stored in SQL Server tables
* The IRDS development team has expertise in SQL Server and has successfully developed and deployed the SDR system into production
* The platform is scalable enough to handle the size and performance requirements that are expected for this system

The initial database will be configured to the default configuration for a SQL server database. As technical requirements evolve, that setup will change to accommodate the updated specifications (i.e., breaking the database file into multiple files).

## Assumptions

There are currently no technical assumptions being made for the Database portion of the IRDS system.

## Issues

There are currently no issues for the Database portion of the IRDS system.

## Constraints

There are currently no constraints for the Database portion of the IRDS system.

# Database Administrative Functions

## Naming Conventions

**Table 6: Database Naming Conventions**

| Type | Guideline |
| --- | --- |
| Reference Tables | Begin with “Ref\_” (ex: Ref\_Gender) |
| Table names | Mountain style (ex: SuicideEvents) |
| Constraints | Begin with constraint abbreviation then underscore (foreign key example: fk\_VetID) |
| Unique Identifier | All tables will have an ID field of type integer – identity. |

## Database Identification

**Table 7: Database Identification**

| Element | Element Name | Meaning |
| --- | --- | --- |
| db\_name | Reach | Production/Master Database |
| db\_name | Reach\_Dev | Development database |
| db\_name | Reach\_Test | Test/Pre-production database |
| db\_path | D:\MSSQL11.MSSQLSERVER\MSSQL\DATA | The full path to the location where the database is stored on the system. |
| db\_file | Reach.mdf | Database filename |
| db\_log | Reach\_log.ldf | Database log file |

## Schema Information

### Description

The Reach database will contain the following schemas:

* .dbo – The MS SQL Server default schema, will contain data imported from VA data sources, such as the VA Suicide Data Repository (SDR). Reference data will be stored here as well, such as a list of VAMC and a list of ICD codes. The risk model output will be persisted into one or more tables in this schema as well.
* SSIS – Will contain tables created during source data import processes, developed using MS SQL Server Integration Services (SSIS). The records created by the last run per SSIS solution (.dtsx) will be persisted in the SSIS schema tables, leaving a footprint available for troubleshooting and debugging
* System – Tables used by the IRDS dashboard, such as a list of preferences for each dashboard user.
* Staging – This will be the staging area for data imported into IRDS before it is transformed by SSIS import packages and imported into tables described in the .dbo schema

### Logical Data Model

The Reach Database contains the following tables:

Data Tables

The Reach database will store source imported data on 2 levels:

* Veteran Demographics
* Veteran Case Data

For each Veteran, a master record will be created in the ‘Veterans’ table and a unique ID will be assigned , ‘ReachID’. The table will also contain basic demographic information for that individual, such as Name, SSN, DOB, Gender, etc. The table also contains a risk ‘Score’ field, which the surveillance component of the system will populate.

All other data elements imported for Veterans will be stored in case level tables. There will be a one to many relationship from the Veteran table to any case level table. For instance, any data that may be available from the SDR on previous suicide attempts for that individual will be stored in a suicide attempt table. For each suicide attempt that is documented for a Veteran, a ‘SuicideAttempt’ record will be created, including a ReachID column that will link the suicide Attempt record back to the Veteran.

Reference Tables

Examples are lists to be used for reporting and normalizing of the data such as a list VAMCs and a list of ICD Codes that contain a diagnosis description associated with each code.

High Risk Veterans View

The High Risk Veterans view will roll up data from the Veterans master list and case level data associated with each Veteran to show all vets that are identified as high risk with their risk score and the factors that contributed to that score. The dashboard will access this view to populate the dashboard screens.

Risk Factors

The Risk Factors determined by the IRDS Risk Model will be captured in the RiskFactors table(s). This table will be updated any time the Risk model is updated, by running the methodology encapsulated in an R program that exists on the IRDS server.

System Tables

Tables used by the dashboard such as User roles and Preferences (see Data Access).



**Figure 3: IRDS Reach Database Logical Data Model**

### Physical Data Model

At this time the following data elements have been identified to be imported into the Reach database. Their exact location within the VA data sources are being determined and access/documentation to those sources are in the process of being requested.

When the database table objects have been defined and created, a SQL Server database diagram will be inserted into this section.

General Veteran Demographics

* First Name
* Last Name
* Middle Name/Initial
* SSN
* DOB
* Race
* Gender
* Contact Information (address, phone)
* Emergency Contacts (name, relationship)

VHA Data

All data elements included in the current VA suicide risk model. For a complete list please refer to the Risk Model Data Dictionary.

VBA Data

* Presence of co-morbid psychiatric disorders
* Alcohol abuse/dependence (303. ICD-9)
* Substance abuse/dependence (304. ICD-9)
* Deployment history, and location of deployments
* History of TBI
* Marital status
* Financial status
* Homeless
* Chronic/terminal illness (non-pain)
* Foreclosure/bankruptcy
* % of service connected disability
* Employment status
* First notice of death
* Enrollment in VR&E
* Recency of divorce
* Legal history (domestic violence, arrests, assaults, other violent/drug offenses)
* History of violence (with or without legal charges)
* Does Veteran have beneficiaries
* History of motor vehicle accidents
* Beneficiary travel reimbursement qualification
* Medical diagnoses to include chronic pain conditions
* Revocation of driver’s license
* Homeownership

### Data Dictionary

The data dictionary will be completed as access and documentation to the data sources become available and requirements are further defined.

### Veteran De-Duping Process

The Reach data model will contain one master list of all Veterans imported into the IRDS system. As each Veteran is imported into the database through one of the data sources, a record will be created for that Veteran in the Veteran table and a unique ID will be assigned. It is possible that an Individual might be imported into the system through multiple sources. When this occurs, the multiple Veteran records created for the individual will be merged and all case level data will be linked to that one merged Veteran record. To achieve this, a de-duping process will be run after each data import is run.

This de-duping process will be contained in an SSIS package, which does the following:

1. Reviews the Veteran table for possible duplicates
2. Duplicate groups are determined (2 or more records that could be duplicates)
3. For each duplicate group a text file is created containing the pertinent demographic information (Name, SSN, DOB, Gender) for all records in that group
4. A python program processes these files and evaluates all the records in a group via a record de-duping algorithm
5. The python program returns the results to the SSIS process via another set of text files
6. The results from these files are placed in a temporary SQL table and the Veteran table is reconciled accordingly, by merging any records that were determined to be duplicates

**Figure 4: Veteran De-duping Overview**

## Denormalization

To be determined as technical requirements are gathered.

## Performance Improvement

To be determined as technical requirements are gathered.

## Storage

To be determined as technical requirements are gathered.

## Recovery

To be determined when location where system is housed is identified.

# Database Interfaces

## Suicide Data Repository (SDR)

**Table 8: SDR Interface Details**

|  |  |
| --- | --- |
| Interface | Details |
| Purpose | Data Repository for VA Suicide and Mortality data which will be imported into the Reach database |
| Characteristics | The SDR system runs on a Windows 2008 R2 machine with a SQL Server data store |
| Interface Architecture | Import solution will developed using SSIS |
| API and Error Conditions | Data will be imported using T-SQL to pull data directly from the SDR tables to the tables in the IRDS staging area |
| Security | A SQL connection will be made to the SDR server using Windows Authentication |

## Perceptive Reach Dashboard

**Table 9: Perceptive Reach Interface Details**

|  |  |
| --- | --- |
| Interface | Details |
| Purpose | IRDS users will access data via the dashboard which manages their access and presentation of that data |
| Characteristics | The Dashboard can be run a complaint web browser |
| Interface Architecture | Client Side is developed in Angular JS, the Server side is developed with Node JS which leverages Express JS |
| API and Error Conditions | The Browser will send requests to Server, which will query the database and return the query results to the browser in JSON format |
| Security | Users login to the dashboard. The dashboard queries settings for that user stored in the Reach database to manage data access |

NOTE: Details for CDW, EDW, VistA and the Analytics sandbox are still being gathered.

# Data Access

1. System users: System users will be given a Login (Server level) or User (Database level) account in SQL Server and granted the appropriate set of rights. System users will connect to the IRDS databases via a SQL Server connection with Windows Authentication, using a tool such as SQL Server Management Studio (SSMS), R or Knime.
2. Dashboard users: Dashboard users will connect to the Reach database indirectly by logging in via the Perceptive Reach dashboard. The dashboard will manage their data access via a set of SQL tables in the ‘System’ schema.

* User Roles – List of roles in the system
* Preferences – List of dashboard preferences
* User – Contains general information about dashboard users, as well as their user role and location
* User Preferences – Contains preferences for each user

When a user logs into the dashboard their presentation of/ and access to data is dictated by the user role and preferences.

## Role Definitions (System)

**Table 10: System Role Definitions**

|  |  |  |
| --- | --- | --- |
| Role-name | Account Type | Rights |
| Database Administrator | Login | sys\_admin |
| Tester | Login | db\_creator |
| Analytics | User | db\_datareader, db\_datawriter |
| Ad-Hoc Reporting | User | db\_datareader |

## Role Definitions (Dashboard)

**Table 11: Dashboard Role Definitions**

|  |  |
| --- | --- |
| User name | Description |
| VAMC level | TBD |
| VISN Level | TBD |
| Region Level | TBD |
| National Level | TBD |

# 

# Implementation Considerations

## Large Objects

Free form text data will be stored in fields with a data type of varchar(max). In SQL Server 2012, a field of type varchar(max) will allow up to 8000 characters. Any data elements that require larger capacity will be stored in fields of type Text.

## Partitioning

At this time there is no partitioning planned.

## Error Processing

After each data import is run, a completion report will be created, any errors incurred will be listed in the report. A database administrator will review the report and take the appropriate action which could include troubleshooting and/or data restoration.