Perceptive Reach

Integrated Reach Database System

(IRDS)

Data Analytics Sandbox Specification



**Department of Veterans Affairs**

Version 1.1

JANUARY 2015

Revision History

| Date | Version | Description | Author |
| --- | --- | --- | --- |
| 1/9/14 | 1.1 | Update per client comments | Shea Kelly |
| 12/12/14 | 1.0 | Final Review | Monica Mohler/Paul Bradley |
| 12/11/14 | .02 | Peer Review | Radina Ivanova |
| 12/10/14 | .01 | Initial Version | Shea Kelly |

**Table of Contents**

[1. Introduction 4](#_Toc409038820)

[1.1. Purpose 4](#_Toc409038821)

[1.2. References 4](#_Toc409038822)

[1.2.1. Documentation, Resources, and Repositories 4](#_Toc409038823)

[1.3. Acronyms 5](#_Toc409038824)

[2. System Description 6](#_Toc409038825)

[2.1. System 6](#_Toc409038826)

[2.2. Security 7](#_Toc409038827)

[3. Environment 8](#_Toc409038828)

[3.1. Hardware 8](#_Toc409038829)

[3.2. Software 9](#_Toc409038830)

[3.2.1. Software Configuration 10](#_Toc409038831)

[3.3. Interfaces 10](#_Toc409038832)

[3.3.1. External Interfaces 10](#_Toc409038833)

[3.3.2. External Data Sources 11](#_Toc409038834)

[4. Analytics Workflow 12](#_Toc409038835)

[4.1. Model Development / Applications 12](#_Toc409038836)

[4.2. System Data Flow 13](#_Toc409038837)

[4.2.1. Import Process 13](#_Toc409038838)

[4.2.2. Normalization/Standardization 13](#_Toc409038839)

[4.2.3. Database Tables 13](#_Toc409038840)

[4.2.4. Linking/Common Identifiers 13](#_Toc409038841)

[4.2.5. Risk Model Execution (SQL Server) 13](#_Toc409038842)

# Introduction

This document provides the general purpose of and specifications for the Data Analytics Sandbox (“Sandbox”). The Sandbox environment will be used to develop and test the IRDS solution, which includes data analytics functions, such as data importing, aggregation, and analysis from the Corporate Data Warehouse (CDW), Suicide Data Repository (SDR), and existing SAS datasets[[1]](#footnote-2) to produce Veteran suicide completion risk scores.

## Purpose

The purpose of this document is to describe the specifications needed to setup, configure, and house the Sandbox. The document’s intended audience is for IT Systems Administrators.

## References

### Documentation, Resources, and Repositories

This document uses the following resources and repositories to support its contents and claims:

* OneVA EA ETA, August 2012

<http://www.ea.oit.va.gov/EAOIT/docs/ETAComplianceFINAL_12_8_28.pdf>

* VA Directive 6221 Accessible Electronic and Information Technology, Directive/Handbook

<http://www.section508.va.gov/docs/Directive_6221.pdf>

* VA Handbook 6500 – Information Security Program  
  <http://vaww1.va.gov/vapubs/viewPublication.asp?Pub_ID=638&FType=2>
* VA Software Document Library (VDL)

<http://www.va.gov/vdl/>

* VA’s Strategic Plan Refresh FY2011–2015

<http://www.va.gov/op3/docs/strategicplanning/strategic_addendum_fy2011-2015.pdf>

* Perceptive Reach Wiki Page

<http://vacloud.us/groups/558/>

* Perceptive Reach GitHub Repository

<https://github.com/VHAINNOVATIONS/PerceptiveReach>

* Perceptive Reach Jira Repository

<https://opensourceehr.atlassian.net/secure/Dashboard.jspa?selectPageId=10600>

## Acronyms

**Table 1: Acronyms and Abbreviations**

| Acronym | Term |
| --- | --- |
| BIRLS | Beneficiary Identification Records Locator Subsystem |
| A&A | Assessment and Authorization |
| ATO | Authority to Operate |
| CDW | Corporate Data Warehouse |
| CIO | VA Chief Information Officer |
| CISO | OIS Chief Information Security Officer |
| DSS | Decision Support System |
| FISMA | Federal Information Security Management Act |
| FISMA | Federal Information Security Management Act |
| FOIA | Freedom of Information Act |
| GB | Gigabyte |
| GHZ | Gigahertz |
| GRC | The Governance, Risk and Compliance |
| HDD | Hard Disk Drive |
| HIPAA | Health Insurance Portability and Accountability Act |
| IRDS | Integrated Reach Database System |
| NDI | National Death Index |
| OCS | Office of Cyber Security |
| OS | Operating System |
| PACER | Public Access to Court Electronic Records |
| R | Statistical Computing |
| RAM | Random Access Memory |
| RPC | Remote Procedure Call |
| SDCD | State Death Certificate Data |
| SDR | Suicide Data Repository |
| SPAN | Suicide Prevention Application Networks |
| SPC | Suicide Prevention Coordinator |
| SPI | Management of Data Breaches Involving Sensitive Personal Information |
| SQL | Structured Query Language |
| SSIS | SQL Server Integration Services |
| SSMS | SQL Server Management Studio |
| TAC | Technology Acquisition Center |
| TRM | Technical Reference Model |
| VA | Department of Veterans Affairs |
| VACI | Veterans Affairs Center for Innovation |
| VCL | Veterans Crisis Line |
| VistA | Veterans Health Information Systems and Technology Architecture |
| VLER | Virtual Lifetime Electronic Record |

# System Description

## System

The Sandbox is a development environment which will house and enable the analysis of multiple integrated datasets using data analytic and modeling techniques and visualizations to identify individual veterans and populations with higher risk of suicide completion and provide proactive and secure results notifications to Veteran support services.

As shown in **Figure 1**, the Sandbox will be a development environment with the following components:

* **Reach Database:** An SQL database used to store relevant data from the CDW and SDR.
* **Data Analytics Platform and Dashboard:** An integrated collection of analytics and visualization tools, including a surveillance dashboard aimed at identifying at-risk Veterans.
* **Direct Messaging:** A method to produce and transmit a secure message regarding Veterans at high risk for suicide completion to authorized outreach and intervention service providers.



Figure 1: Conceptual Program Design

The underlying technology and data analytics and modeling platform will provide methods by which at-risk Veterans for suicide can be identified. Specifically, we propose a programmable and configurable solution that can be customized, enhanced and expanded over time as more data sources become available, clinical research identifies new risk factors, and statistical models for predicting suicide completion are updated. The main components of the data analytic solution are comprised of:

a) A method of identification and measurement of Veteran-specific risk factors

b) The development of a statistical suicide completion predictive model and

c) The computation of the suicide completion risk scores for individual veterans

This is a precursor to the design of an automated reporting model. This effort includes data analysis and predictive modeling, executed in collaboration with VA stakeholders and clinical subject matter experts and using the existing suicide completion risk model with modifications aimed at improving the out-of-sample predictive power. Harnessing the automated reporting model, information will be presented in a customizable national surveillance dashboard and transmitted to authorized officials and Veteran support services organizations via secure messaging (i.e. Direct Messaging).

## Security

The IRDS solution, including the functions within the data analytics sandbox, utilize Personally Identifiable Information (PII) and Protected Health Information (PHI). This data will be secured and protect per the following regulations and directives:

* Health Insurance Portability and Accountability Act (HIPAA) of 1996
* U.S.C. § 3541, “Federal Information Security Management Act (FISMA) of 2002”
* U.S.C. § 552a, as amended, “The Privacy Act of 1974”
* VA Directive 6500, “Managing Information Security Risk: VA Information Security Program,” September 20, 2012
* VA Handbook 6500, “Risk Management Framework for VA Information Systems – Tier 3: VA Information Security Program,” September 20, 2012
* VA Handbook 6500.1, “Electronic Media Sanitization,” March 22, 2010
* VA Handbook 6500.2, “Management of Data Breaches Involving Sensitive Personal Information (SPI)”, January 6, 2012
* VA Handbook 6500.3, “Assessment, Authorization, And Continuous Monitoring Of VA Information Systems,” February 3, 2014
* VA Handbook, 6500.5, “Incorporating Security and Privacy in System Development Lifecycle” March 22, 2010
* VA Handbook 6500.6, “Contract Security,” March 12, 2010
* VA Directive 6508, VA Privacy Impact Assessment, October 3, 2008
* VA Directive 6300, Records and Information Management, February 26, 2009
* VA Handbook, 6300.1, Records Management Procedures, March 24, 2010
* OMB Memorandum M-07-16, Safeguarding Against and Responding to the Breach of Personally Identifiable Information, May 22, 2007
* NIST SP 800-63-2, Electronic Authentication Guideline, August 2013

The system security will be maintained according to VA policy and will include obtaining an Authority to Operate (ATO). This will include the creation of security artifacts:

* Security Management Plan,
* System Security Plan,
* Privacy Impact Assessment,
* Configuration Management Plan,
* Disaster Recovery Plan,
* Risk Assessment and Security Certification Checklist.

The Governance, Risk and Compliance (GRC) tool (RiskVision) shall act as the management tool for the Assessment and Authorization (A&A) process, and systems shall be assessed in RiskVision by an Office of Cyber Security (OCS) representative for an accreditation recommendation to be submitted to the OIS Chief Information Security Officer (CISO) and VA Chief Information Officer (CIO) for final review and determination.

# Environment

## Hardware

Table 2: Hardware Requirements

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

## Software

Table 3: Software Requirements

| **Software Component** | **Requirement** | **Description** |
| --- | --- | --- |
| Operating System | Microsoft Windows 64 bit OS | Operating system with 64 bit processing power |
| Database | Microsoft SQL Server 2012 | SQL Server is a relational database management system |
| Database | SQL Server Integration Services (SSIS) | SSIS provides data migration task capabilities for SQL Server |
| Analytics Tool | R version 3.1.2 | R is an open source software package for statistical analysis, data mining, and visualization |
| Analytics Tool | RStudio version 0.98.1091 | RStudio is an open source, user interface for the R software package |
| Analytics Tool | KNIME version 2.10.4 | KNIME is an open source, data analytics, reporting and data integration software package |
| Javascript | AngularJS version 1.3.6 | AnularJS is an open source web application framework |
| Other | HTML5 | HTML5 is an open source web technology markup language |
| Other | CSS3 | CSS3 is a style sheet language used for HTML5 |
| Other | JQuery v 2.1.1 | JQuery is an open source cross-platform for writing JavaScript and HTML |
| Testing | Cucumber JVM 1.2.0 | Cucumber JVM is an open source computational programming testing tool |
| Testing | Jasmine 2.0 | Jasmine is an open source testing framework for JavaScript |

### Software Configuration

R will connect to SQL Server through an ODBC connection created using Microsoft ODBC Data Source Administrator. This will allow R to access data from the SQL Server database within the Sandbox.

## Interfaces

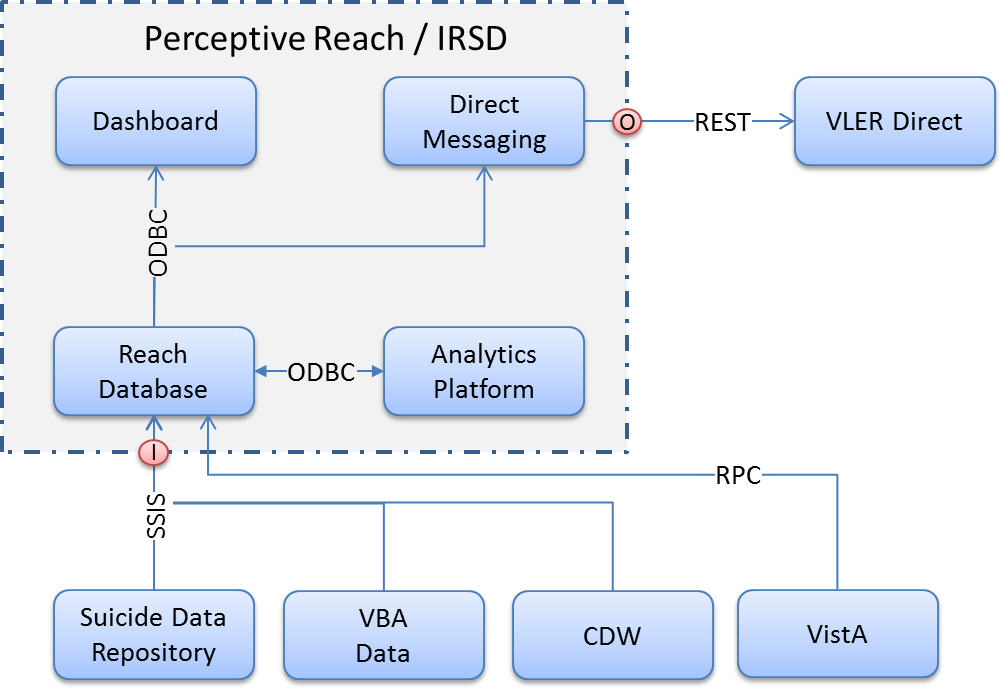


Figure 2: Example of IRDS interface design

### External Interfaces

The external interfaces of the IRDS solution are:

1. SDR SSIS Interface: SSIS will be the primary tool for importing data from SDR into the IRDS Reach database.
2. CDW SSIS Interface: The CDW Interface will be added upon further clarification of requirements.
3. VistA RPC Interface: 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.
4. VLER Direct Messaging: The Direct Messaging component will integrate into the existing VLER Direct Messaging architecture. The Direct Messaging will create unique messages based on new and existing data annotated in the IRDS.
5. IRDS Surveillance Dashboard: The dashboard will provide a browser based user interface featuring representations of the analyzed consolidated data sources. These are processed and organized into visualizations which will assist SPCs and other VA personnel in their duties.

### External Data Sources

The external data sources to be imported into Reach database of the IRDS have been identified as:

1. VA Suicide Data Repository (SDR) - The SDR contains VA and DoD suicide and mortality data. The data store is Microsoft SQL Server. Periodically data is imported into the system from

* Mortality search results from the National Death Index (NDI)
* State Death Certificate Data (SDCD)
* Veterans Crisis Line (VCL)
* Suicide Prevention Applications Network (SPAN)
* VA Beneficiary Identification Records Locator Subsystem (BIRLS) Death File

1. Corporate Data Warehouse - CDW warehouses VHA and VBA data is SQL Server format. It is assumed that VHA and VBA data will be imported into the IRDS Reach database. The plan is to connect to the CDW servers via a SQL connection and directly pull the data into the IRDS system.
2. Veterans Health Information Systems and Technology Architecture (VistA) RPC Interface (tentative) - The VA uses the Veterans Health Information Systems and Technology Architecture (Vista) system, for managing Veterans health data. Data may be imported into the IRDS system directly from VistA using RPC calls when immediate access to recently updated clinical records is retired. As VistA is modernized additional interface protocols may be used in place of RPCs. This is to be determined. -
3. Non VA data sources - The VA has expressed an interest in importing and analyzing data that is external to VA data sources. Some of the sources being considered are LexisNexis and PACER. The PwC requirements team is currently reviewing possibilities with the VA.

Data in the CDW and SDR are stored in SQL Server tables. SSIS import packages will be developed connect to the CDW and SDR, extract select data from the SQL tables, and copy the data to SQL tables located in the IRDS Reach Database. Appropriate transformations such as standardization, normalization, and cleaning will be made to the data as it is transferred into the Reach Database.

Data extracted from VistA is extracted using RPC calls. This data will also be copied into SQL tables located in the Reach Database.

# Analytics Workflow

## Model Development / Applications

Data will be pulled from the CDW and the SDR into the Sandbox and stored in the SQL database based upon the interfaces (see section 3.3.1.).

The data will imported from the SQL database into R allowing for variable creation and suicide completion predicative model development. The coefficients fitted during the model development will be stored as an SQL table.

On an ongoing basis and using SQL, these stored coefficients will be applied to updated data to produce suicide completions risk scores for individual Veterans.

Individual Veteran risk scores will be used for dashboarding and direct messaging purposes. **Figure 3** provides a visual overview of the data analytic model workflow within the Sandbox environment.



Figure 3: Example of data analytic model workflow

## System Data Flow

### Import Process

The import process for analytic model development will take place in two phases:

1. SQL Server will use SSIS to query data from CDW and SDR and save the data tables to be accessed by R.
2. R will then import the data from data tables for model development.

### Normalization/Standardization

For current suicide completion risk model recreation and development of its modifications, all data normalization and standardization will be performed in R. SQL Server will pull tables and fields with no normalization or standardization applied.

### Database Tables

The database tables will be queried from the CDW and the SDR; these tables will inherit the structure from their native source.

### Linking/Common Identifiers

The linking / common identifiers will be determined once data access has been granted and linkages can be determined.

### Risk Model Execution (SQL Server)

The suicide completion risk model supports the Perceptive Reach program by identifying individual Veterans at high risk for suicide completion. The application of the suicide completion risk model coefficients to produce individual Veteran suicide completion risk scores will be executed within the SQL Server. More specifically, the coefficients from the predictive suicide completion risk model will be saved to a table within the SQL Server and applied to the ongoing Veteran data in SQL. These risk scores will be imported into the dashboard and used to generate direct messages.

1. The existing SAS datasets shall be the datasets used to develop the initial suicide completion risk model. [↑](#footnote-ref-2)