Architectural Description - Risk Assessment

The purpose of this document is to assist organizations who are undertaking a risk assessment process, which involves DataSHIELD with Armadillo or Opal. The document will describe a reference deployment of DataSHIELD including Armadillo or Opal, and other related components. The document will also describe the security purpose and features of the components deployed in the reference deployment.

The intent is that this reference deployment can be used to examine the issues which need to be explored and addressed during the rise assessment process.

Reference Deployment

This description of a reference deployment of DataSHIELD assumes that the intention is to permit active disclosure protected data analysis from external organizations, and to have the risk of disclosure mitigated by the use of DataSHIELD. In these circumstances active disclosure protection analysis means the data analysts can request analytical operations be performed on the data, but the results of such analysis are summary statistics which are not disclosive, and conform to legal, ethical and governance constraints which are being upheld by the organization. To be clear, at no point will the data analyst have direct access to the data, or a copy of the data.

DataSHIELD Background

The active disclosure protection is achieved by utilizing DataSHIELD, which is essentially a set of R functions collected into R packages. The DataSHIELD packages are split into client-side and server-side packages. The DataSHIELD client-side packages are used by the data analysts and the DataSHIELD server-site packages are deployed within the organization's Armadillo or Opal server, through which analytical operations can be performed.

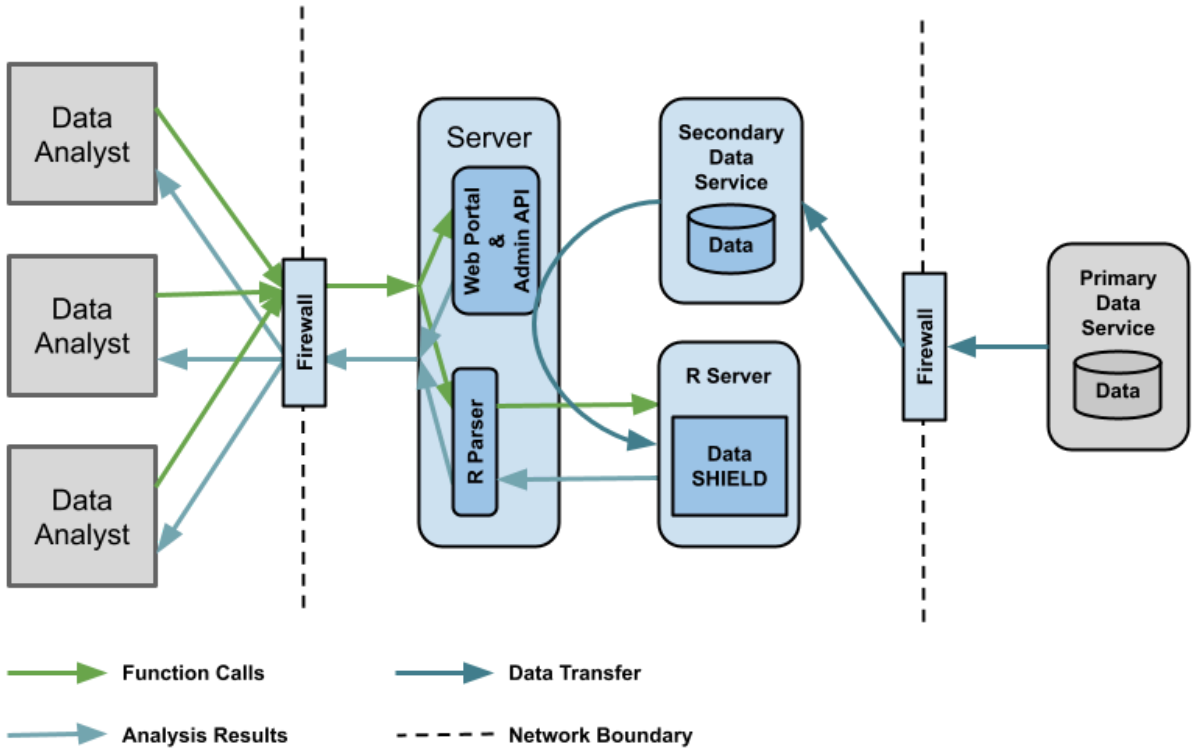
The DataSHIELD client-side R packages provide 97 functions, such as 'ds.mean' and 'ds.glm', to permit data analysis to perform analysis on multiple data sets (each possibly hosted by different organizations). DataSHIELD client-side functions are also responsible for collating the summary statistics for each data set, into a single collated set of summary statistics which is provided to the data analyst.

The DataSHIELD server-side R packages provide 101 functions, such as 'meanDS', 'glmDS1', and 'glmDS2' to perform analysis on the data. The server-side functions are split into two types: assign and aggregate functions. The assign functions of DataSHIELD are operations which don't return any analytical results, but assigns the results to a specified variable within the server. The aggregate functions of DataSHIELD are operations which return summary statistics to the client-side. The level of disclosure of the aggregate functions can be configured using control parameters, for example, the minimal number of values which can be combined to create a mean value. All functions could possibly return a non-disclosive error response, if the function is unsuccessful.

This discussion assumes the DataSHIELD deployment isn't using the new Resourcer capabilities provided by the latest versions of Armadillo and Opal.

Reference Deployment Structure

This reference deployment is split over three network regions connected by firewall and reverse proxies (not shown), which restrict traffic flows between the network regions. The firewall between the data analysts and the Server (either Armadillo or Opal) will be configured to only allow ‘https’ network traffic from data analysts nodes to a single port on the Server, response will be routed back to the data analysts via the firewall. Another firewall between the primary data service and the secondary data service is used to ensure that access can not be gained to the network region holding the primary data service from the network region containing the secondary data service. The purpose of this firewall is to permit updates to the primary data service to be copied to the secondary data service.



The purpose of the components and interactions within the reference deployment are:

- Data Analyst: the node from which data analysts make analysis request to the Server, and receive analysis results;
- External Firewall: the firewall between the data analysis and Server;
- Internal Firewall: the firewall between primary and secondary data service;
- Analysis Requests: the request sent by the data analysis to DataSHIELD, via the Server;
- Analysis Results: the non-disclosive results send from DataSHIELD to the data analysis,via the Server;
- Server (either Armadillo or Opal): the Server manages: authentication and authorization of sessions and demultiplexing of requests to the R Parser or Web Portal. User authentication, by the Server, can be performed via password or certificate, and supports 2-factor authentication;
- Web Portal (Opal only): within the Server provides a web based administrative interface to Opal server;
- Admin API (Armadillo or Opal): within the Server provides a REST based administrative interface to Server;
- R Parser: the R parser checks the analysis requests only contain valid requests, and in particular only permitted functions;

- Primary Data Service: this data service contains the primary copy of the data to be analysed;
- Secondary Data Service: the data service, which could be provided by MySQL or mongodb, contains the secondary copy of the data to be analysed. This version of the data will be copied into the R server to be analysed;
- R Server: the R server is an execution environment, one per user, within which analysis requests are executed;
- DataSHIELD: an R package, within the R server, which provides data analytic functions.