Informatics for Integrating Biology and the Bedside

i2b2 Software Architecture

Data Repository (CRC) Cell

Document Version: *1.5.1*

I2b2 Software Version: *1.5*

Table of Contents

ABSTRACT

This is a software architecture document for CRC (Clinical Research Chart) cell. It identifies and explains the important architectural elements. This document will serve the needs of stake holders to understand the system concepts, and give a brief summary of the use of the CRC message format.

1. OVERVIEW

The Clinical Research Chart (CRC) repository cell is one of the core cells in the i2b2 Hive. The CRC cell is designed with several requirements. The main requirements are:

1. It must be able to hold healthcare information from many different venues and allow it to be queried rapidly even if there are hundreds of millions of rows.
2. It must be easily combined with other project repositories to form large unified repositories.
3. Finally, it must allow objects to be stored that are present in the genomic data.

Currently information in the CRC cell is related to clinical data and hence it's also called Clinical Research Chart. For the remainder of this document, the terms **CRC** and **Data Repository Cell** will be used interchangeably to refer to the same cell.

The CRC is a data warehouse of patient's phenotype and genotype information. It is supported by a powerful metadata management module (the Ontology Cell). Currently the CRC handles concepts such as diagnoses, procedures, medications, and lab tests, but the structure of the table gives enough flexibility to expand this to include virtually any kind of observation. The presence of both genotype and phenotype information makes this cell a powerful tool for researchers.

All patient data present in the CRC is de-identified; the only exception is the patient notes from hospitals. These notes are stored in encrypted form, so only users enabled with an encryption key can view them.

1.1 CRC Definitions, Acronyms and Abbreviations

1.1.1 Patient Data Object (PDO):

This object mirrors the star schema database model of the data mart. It holds patient information such as clinical observations, demographics and provider data.

1.1.2 Setfinder Query:

Setfinder queries are used to create a set of patients that satisfy a criteria presented in the query. The setfinder query is composed of query constraints, a list of panels and its items.

1.1.3 Observation Fact:

Any observation made on a patient can be stored as fact information in CRC data mart. The user can fetch the fact information via the PDO queries.

1.2 User Role

The CRC determines when and how data is presented to a user based on their user roles, which are specified in the PM Cell. Each user will have at least two roles per user_ID and product_ID combination. These two roles can be further defined as a Data Protection role and a Hive Management role.

The data protection role establishes the detail of data the ser can see while the hive managment role defines their level of functionality the user has in a project.The following tables summarize the roles in a hierarchical order of least to most access.

Data Protection Track	
Role	**Access Description**
DATA_OBFSC	OBFSC = Obfuscated ▪ The user can see aggregated results that are obfuscated (example: patient count). ▪ The user is limited on the number of times they can run the same query within a specified time period. If the user exceeds the maximum number of times then their account will be locked and only the Admin user can unlock it.
DATA_AGG	AGG = Aggregated ▪ The user can see aggregated results like the patient count. ▪ The results are not obfuscated and the user is not limited to the number of times they can run the same query.
DATA_LDS	LDS = Limited Data Set ▪ The user can see all fields except for those that are encrypted. ▪ An example of an encrypted field is the *blob fields* in the *fact* and *dimension tables*.
DATA_DEID	DEID = De-identified Data ▪ The user can see all fields including those that are encrypted. ▪ An example of an encrypted field is the *blob fields* in the *fact* and *dimension tables*.
DATA_PROT	PROT = Protected ▪ The user can see all data, including the identified data that resides in the Identity Management Cell.

Hive Management Track	
Role	**Access Description**
USER	Can create queries and access them if he/she is the owner of the query.
MANAGER	Can create queries and can access queries created by different users within the project.
ADMIN	

***Note**: Further details regarding roles can be found in the PM_Design_Document.*

1.3 Security

Users can accesses the CRC with domain-id, project-id, user-id and password combination, which is authenticated through the Project Management Cell. The implementation detail of Project Management Cell is considered out-of scope to this system context.

***Note**: Further details regarding the implementation of the Project Management cell can be found in the PM_Install_Guide.*

1.4 Scope of the system

Some other participants, currently outside the scope of CRC, are:

- Project Management Cell
- Ontology Cell
- edu.harvard.i2b2.common

1.5 Assumptions/Constraints

- The data in the CRC data mart database will not have identified data. The exception to this are the patient notes stored inside "OBSERVATION_BLOB" which will be encrypted.
- The client will make "Patient Data Object Query/Request" in multiple requests if the input list(PatientSet or ObservationSet) is big.

1.6 Technical Platform

The technology used to build the product is as follows

- Java 2 Standard Edition 6.0
- Oracle Server 10g database
- SQLServer 2005
- Xerces2 XML parser
- JBoss Application server version 4.2.2 and higher
- Spring Web Framework 2.0
- Axis2.1 web service (SOAP/REST)

1.6.1 Transaction

The CRC system is transactional, leveraging the technical platform capabilities. The transaction management model of the J2EE platform will be reused intensively.

***Note**: In the current implementation, to support long running setfinder queries, transaction management will be manually turned off until the completion of the query.*

1.6.2 Security

The application must implement basic security behaviors:

- Authentication: Authenticate using the combination of domain id, project id, user name and a password.
- Authorization: Based on the user role, the user may access setfinder queries created by other users, view patient notes,etc..
- Confidentiality: Sensitive data must be encrypted (Patient Notes).
- Data integrity: Data sent across the network cannot be modified by a tier.
- Auditing: All queries and retrieval of patient data is stored for auditing purposes.
- User Lockout: Users with the role of DATA_OBFSC will be limited to the number of times they can run the same query in a project. Once they reach that limit their account will be locked out and they will not be able to run queries again until an administrator unlocks the account.

1.6.3 Persistence

Application uses the JDBC calls to persist data.

1.6.4 Reliability/Availability

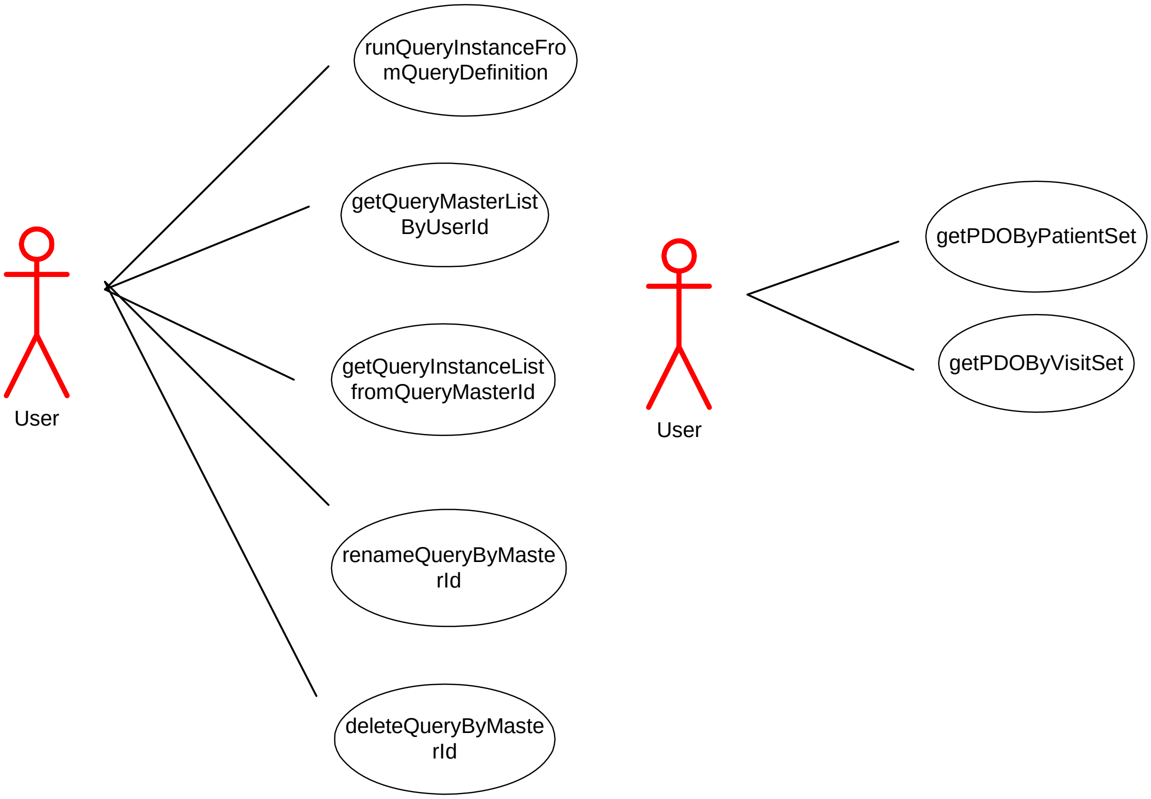
- The reliability/availability will be addressed through the J2EE platform
- Targeted availability is 16/7: 16 hours a day, 7 days a week
- The time left (8 hours) is reserved for any maintenance activities

1.6.5 Performance

- The user authentication with the project management cell must be under 10 seconds.
- The concept code lookup to the ontology cell must be under 10 seconds.

2. USE CASE

The diagram below depicts the common use cases a user can perform with the CRC cell.


runQueryInstanceFromQueryDefinition
getQueryMasterListByUserId
getQueryInstanceListfromQueryMasterId
renameQueryByMasterId
deleteQueryByMasterId
User
User
getPDOByPatientSet
getPDOByVisitSet

2.1 Use Case: Run a Query from Panel Definition

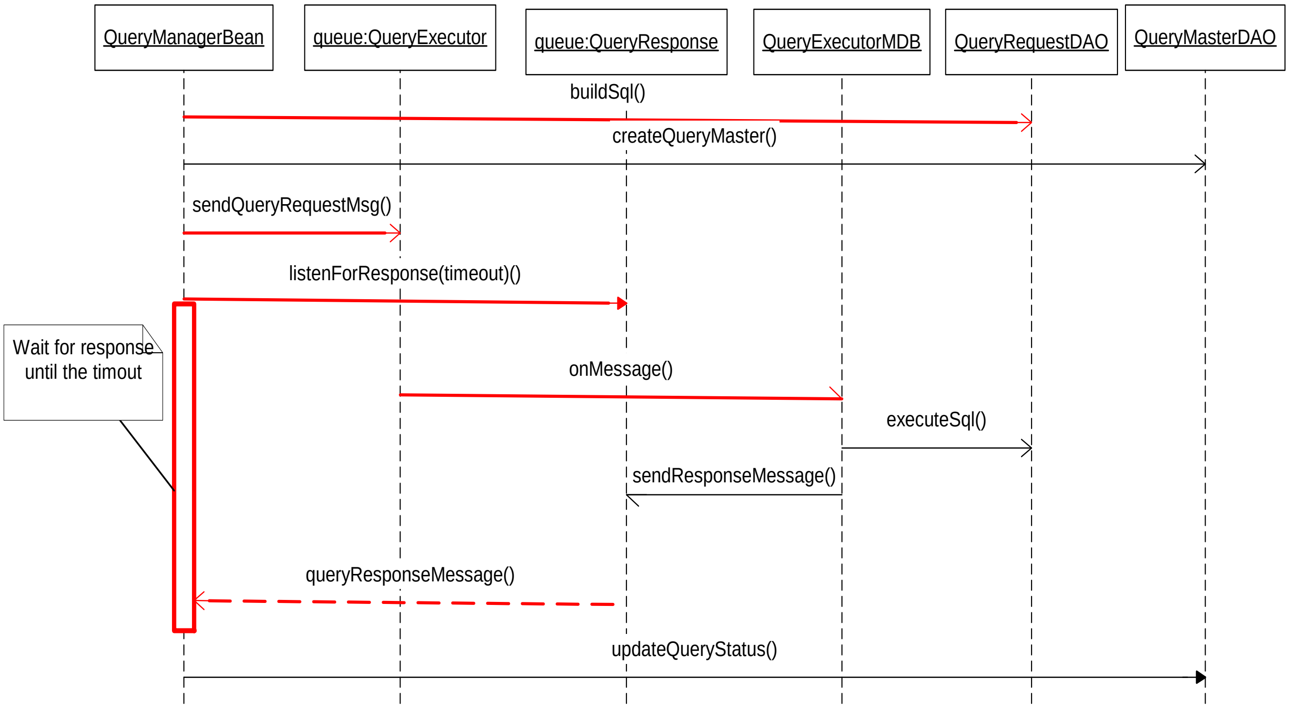
- Validate the user by calling the **Project Management Cell**.
- Select a data mart based on the combination of domain_id, project_id and user_id.
- Call the **Ontology Cell** with the item key and determine the dimension table to join with the fact table.
- Save the query panel definition and the generated SQL statements.
- Generate the list of output like the patient count, patient gender count, patient set, etc.
- To scale the application and to support long running SQL, the execution of SQL is handled inside a set of queues. At first the query SQL statements will

be executed inside a small job queue, if it didn't complete within a certain time period, then the jobs will be transferred to mid size job queue and then to large size job queue.

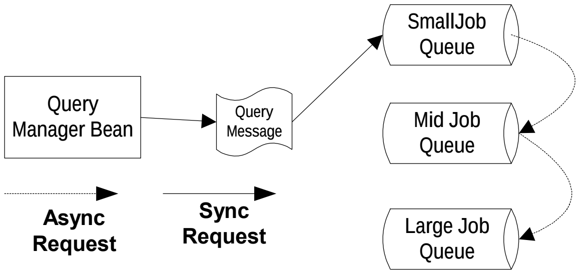
- If the SQL execution completes before the "result_waittime_ms" which is specified in the request, then the query results is passed in the response message, otherwise the status of the query is passed in the response message.

2.1.1 CRC Query execution using Queue Model:

2.1.1.1 SEQUENCE DIAGRAM



2.1.1.2 CONTEXT DIAGRAM



2.2 Use Case — Get PDO from PatientSet

- Validate the user via the Project Management Cell
- Select the data mart based on the domain_id, project_id and user_id.
- Call Ontology Cell with the item key and determine the dimension table to join with the fact table.
- Using the given patient set or Observation set, apply the Panel filters and return PDO.

3. ARCHITECTURE DESCRIPTION

As noted in “Documenting Software Architectures”[0], software architecture is a complex entity that cannot be described in a simple one-dimensional fashion. This document provides the description of the architecture as multiple views. Each view conveys the different attributes of the architecture.

1. Components and Connector View
    a. Client-Server Style

2. Module View
    b. Decomposition Style
    c. Uses Style

3. Data View
4. Deployment View

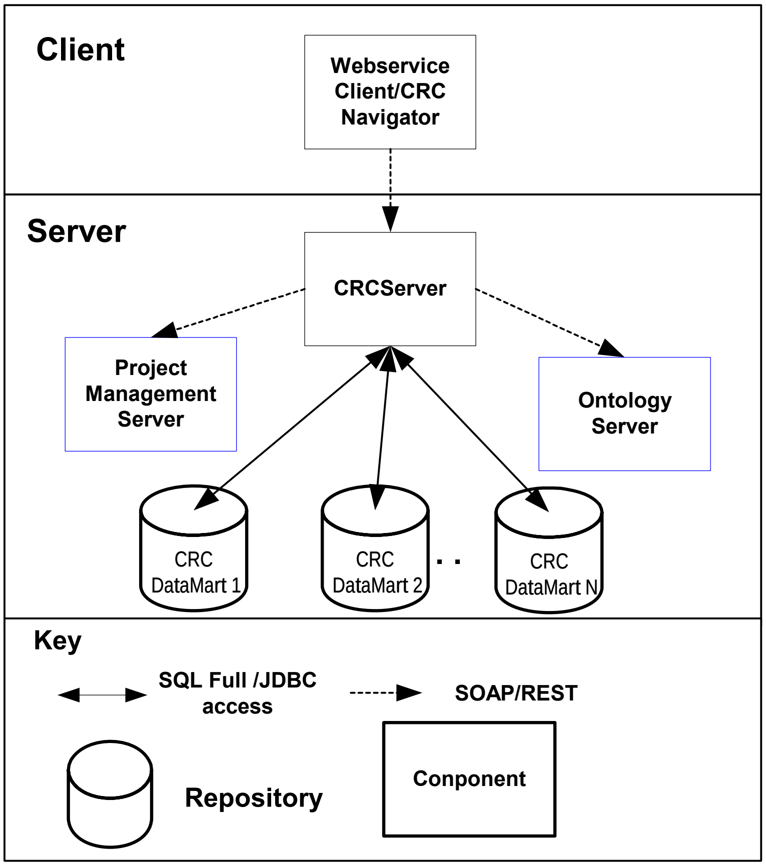
3.1 Components and Connector View

A **Component and Connector view** represents the runtime instances and the protocols of connection between the instances. The connectors represent the properties such as concurrency, protocols and information flows. The diagram shown in the *Primary Presentation* section represents the Component and Connector view for the multi-user installation. As seen in the diagram, component instances are shown in more detail with specific connectors drawn in different notations.

3.1.1 Client-Server View

The CRC system is represented using the C&C Client-Server view.

3.1.1.1 PRIMARY PRESENTATION



3.1.1.2 ELEMENT CATALOG

3.1.1.2.1 Elements and their Properties

The properties of CRC cell elements are:

- *Element name*: given in the following table
- *Type*: whether the element is a data repository, a data accessor, a communication method, a query, a client or a server component
- A *description* of the element

Element Name	Type	Description
Webservice Client	Client	Webservice client (i2b2 Workbench/Navigator) submits the requests to CRC Server components and renders response XML.
CRC Server	Server	Provides Web Service Interface for the CRC system. It supports both SOAP and REST protocols.

		It uses Project Management server to handle user authentication. It uses Ontology server to lookup the concepts metadata. Select the CRC data mart based on domain-id, project-id and user-id. It stores Setfinder query definition, query run instance and the corresponding query results. The user can then request Patient Data Object using the Setfinder results.
Project Management Server	Server	CRC cell uses the Project Management cell to authenticate the user. The CRC cell constructs PM Cell request message and makes a web service call to Project Management Cell.
Ontology Server	Server	CRC sends web service requests to the Ontology cell to get metadata information about an Observation fact's concepts.
CRC Datamart DB	Data Repository	This repository is mainly a data mart for patient's clinical observation information represented in star schema. The Server supports multiple data marts; the data marts are selected based on the domain_id, project_id and user_id combination. This database also holds CRC user's queries (setfinder query) information and its results like patient sets, etc.
Full SQL	Query Connector	SQL query used as a connector between the CRC System and the CRC Datamart DB.
Web Service	Request Connector	SOAP or REST request used to communicate with the external system.

3.1.1.3 RELATIONS AND THEIR PROPERTIES

The relation of this C&C view is *attachment*, dictating how components and connectors are attached to each other. The relations are as shown in the primary presentation section; there are no additional ones.

3.1.1.4 DESIGN RATIONALE, CONSTRAINTS

N-tier Architecture

The client-server style depicts the n-tier architecture that separates presentation layer from business logic and data access layer; thus providing for a high degree of portability through the application of the principle of Separation of Concerns.

3.2 Module View type

The module view shows how the system is decomposed into implementation units and how the functionality is allocated to these units. The layers show how modules are encapsulated and structured. The layers represent the “allowed-to-use” relation.

The following sections describe the module view using Decomposition and Uses Style.

3.2.1 Decomposition Style

The Decomposition view presents the functionality in terms of manageable work pieces. They can be further decomposed to present higher level of details. The decomposition view identifies modules and breaks them down into sub-modules and so on, till a desired level of granularity is achieved. The “Uses” style shows the relationships between modules and sub-modules. This view is very helpful for implementation, integration and testing the system.

3.2.1.1 PRIMARY PRESENTATION

System	Segment
CRC	Setfinder Manager
	PDO Manager

3.2.1.2 ELEMENT CATALOG

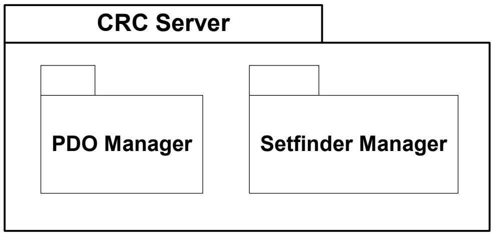
3.2.1.2.1 Elements and their properties

Element Name	Type	Description
Setfinder Manager	Subsystem	This subsystem manages user’s Setfinder queries. Keep tracks of query information like query definition, its SQL, owner of query, etc. Also the results of query like the patient set, visit set, etc is stored.
PDO Manager	Subsystem	This manages both plain and table Patient Data object queries.

3.2.1.3 RELATIONS AND THEIR PROPERTIES

The subsystem elements form the ***is-part*** of relation with the overall CRC system.

3.2.1.4 CONTEXT DIAGRAM



3.2.2 Uses Style

3.2.2.1 PRIMARY PRESENTATION

System	Segment
CRC	CRC Module
Setfinder Manager Subsystem	Setfinder Web Service
	Setfinder EJB
	Setfinder DAO
	edu.harvard.i2b2.common
PDO Manager Subsystem	PDO Web Service
	PDO EJB
	PDO DAO
	edu.harvard.i2b2.common

3.2.2.2 ELEMENT CATALOG

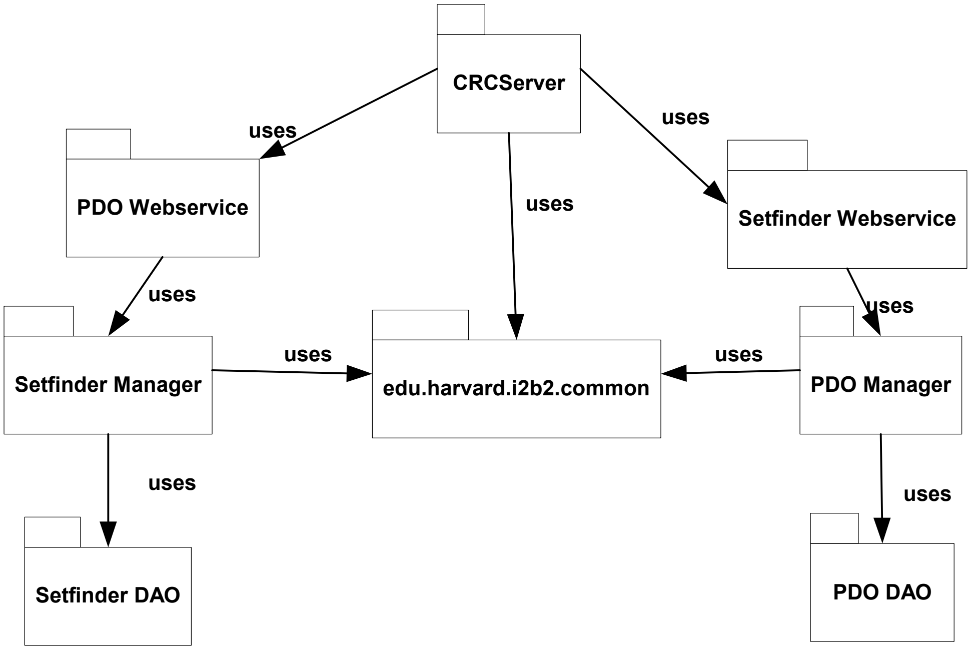
3.2.2.2.1 Elements and their properties

Element Name	Type	Description
CRC Module	Module	User Login Module authenticates through PIN Server System with user id and PIN.
Setfinder Webservice	Module	Provides web service interface to Setfinder operations.
Setfinder EJB	Module	Delegates Setfinder requests to DAO layer to perform database operations.
Setfinder DAO	Module	Supports operation like create query master, delete query, saving query definition and its results.
PDO Webservice	Module	Provides web service interface for PDO requests.
PDO EJB	Module	Module to delegate PDO requests to corresponding PDO and to build PDO response message.
PDO DAO	Module	Module to query database based on PDO requests.
edu.harvard.i2b2.common	Module	This module provides utility classes to handle JAXB, JNDI, etc.
Persistence Service	Module	Provides SQL interface to database.

3.2.2.3 RELATIONS AND THEIR PROPERTIES

The modules in this style follow a ***depends-on*** relation.

3.2.2.4 CONTEXT DIAGRAM



3.3 Mappings of Styles

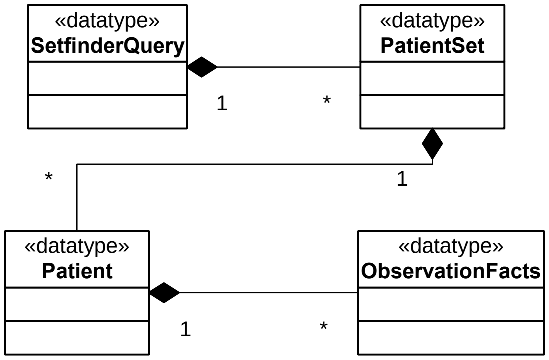
The following table is a mapping between the elements in the Component & Connector Client-Server view shown in section 4, and the Modules Uses view and Decomposition view shown in sections 5 and 6.

The relationship shown is *is-implemented-by*, i.e. the elements from the C&C view shown at the top of the table are implemented by any selected elements from the Modules views, denoted by an "X" in the corresponding cell.

	CRC Server	PM Server	Ontology Server	CRC Data Mart DB
CRC Service	X	X		
Setfinder Webservice	X			
PDO Webservice	X			
SetFinderEJB	X			
PDOEJB	X		X	
SetFinderDAO	X			X
PDODAO	X			
Persistence Service				X

4. DATA VIEW

The key data elements related to the CRC system are:

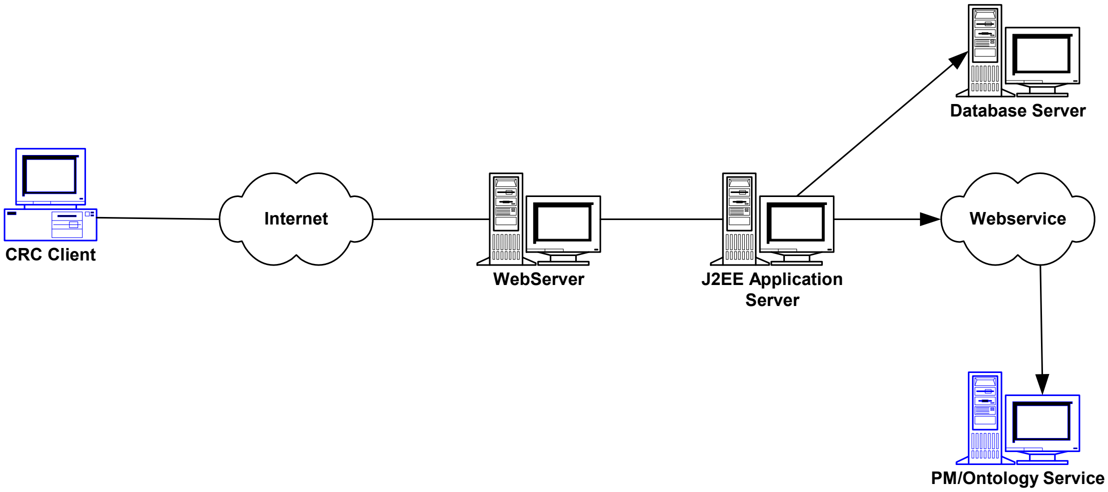


4.1 Volumes

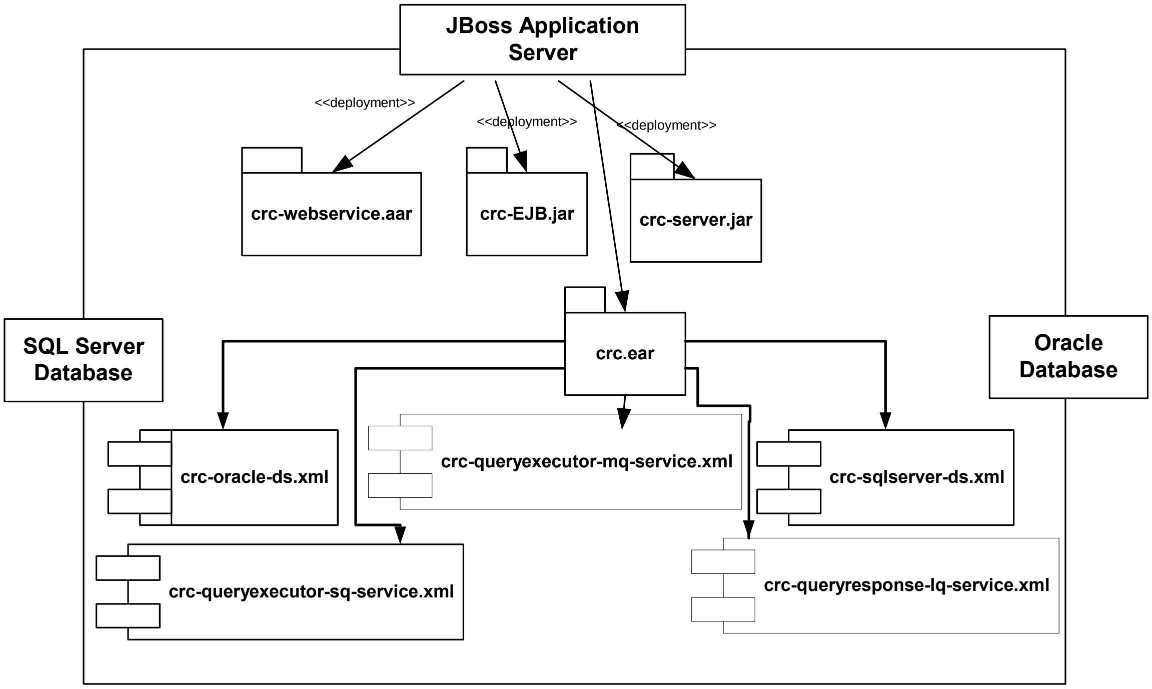
- Estimated new setfinder query : 100 a day, with peaks in the morning
- Average PatientSet size 100,000
- CRC registered individual user : about 150

5. DEPLOYMENT VIEW

5.1 Global Overview



5.2 Detailed deployment model



REFERENCES

Clements, P., Bachmann, F., Bass, L., Garlan, D., Ivers, J., Little, R., Nord, R. and Stafford, J., (2003). Documenting Software architectures – Views and Beyond. Addison Wesley, Boston, MA.

The “4+1” view model of software architecture, Philippe Kruchten, November 1995, http://www3.software.ibm.com/ibmdl/pub/software/rational/web/whitepapers/2003/Pbk4p1.pdf

Object Management Group UML 2.0 Specification - http://www.omg.org/technology/documents/formal/uml.htm

i2b2 (Informatics for Integrating Biology and the Bedside) https://www.i2b2.org/resrcs/hive.html