SMR - Product and Technical Features, and Software Detailed Design

- Architecturally Significant Requirements
- Product and Technical Features
  - Logical Components and Data Flow Diagram
    - Data Access Module (DAM)
    - Data Ingest Module (DIM)
      - Benefits of the RDL
    - Source-specific Integration Module (SIM)
      - Benefits of the SDS
  - Deliver
- Technical Features and Product Capabilities
  - The SMR Technical Features and Product Capabilities to the HHS 2020 SMR implementation/requirements mapping
- SMR - Software Detailed Design
  - Design Considerations
    - Goals and Guidelines
  - Development Methods and Contingencies
  - Architectural Strategies
  - Architecture Design
  - System Design
  - Software Detailed Design

Architecturally Significant Requirements

<table>
<thead>
<tr>
<th>Key</th>
<th>Summary</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIINST-3791</td>
<td>BR-SMR-024</td>
<td>The SMR shall have a Claim data domain.</td>
</tr>
<tr>
<td>SIINST-3783</td>
<td>BR-SMR-020</td>
<td>The SMR shall have a Client data domain.</td>
</tr>
<tr>
<td>SIINST-3368</td>
<td>BR-SMR-001</td>
<td>The SMR shall access data from the source system(s).</td>
</tr>
<tr>
<td>SIINST-3367</td>
<td>BR-SMR-002</td>
<td>The SMR shall ingest data received from source system(s) into the Raw Data Lake (RDL).</td>
</tr>
<tr>
<td>SIINST-3366</td>
<td>BR-SMR-006</td>
<td>The SMR data that are not confidential but should not be made public are classified as restricted</td>
</tr>
<tr>
<td>SIINST-3365</td>
<td>BR-SMR-005</td>
<td>The SMR data used to identify an individual shall be classified into Personally Identifiable Information (PII).</td>
</tr>
<tr>
<td>SIINST-3364</td>
<td>BR-SMR-008</td>
<td>The SMR data classified as having moderate sensitivity shall be characterized into Internal data usage group.</td>
</tr>
<tr>
<td>SIINST-3363</td>
<td>BR-SMR-007</td>
<td>The SMR data classified as having high sensitivity shall be characterized into confidential data usage group.</td>
</tr>
<tr>
<td>SIINST-3362</td>
<td>BR-SMR-012</td>
<td>The SMR shall support the migration of data from SMR to the HHS2020 Modules.</td>
</tr>
<tr>
<td>SIINST-3360</td>
<td>BR-SMR-015</td>
<td>The SMR shall have a Managed Care Organization data domain.</td>
</tr>
<tr>
<td>SIINST-3359</td>
<td>BR-SMR-017</td>
<td>The SMR data with health information of an individual that is individually identifiable shall be classified as Protected Health Information (PHI).</td>
</tr>
<tr>
<td>SIINST-3358</td>
<td>BR-SMR-021</td>
<td>The SMR shall classify Adoption care data as PII++ (Adoption Care) that will be accessible only to specific individuals</td>
</tr>
<tr>
<td>SIINST-3356</td>
<td>BR-SMR-016</td>
<td>The SMR shall have a Third Party Liability data domain.</td>
</tr>
</tbody>
</table>
Product and Technical Features

The SMR is a data repository that integrates different legacy data sources identified as data providers and exposes the transformed and standardized (but still source-specific) data to the new Medicaid Management Information System Replacement (MMISR) modules. This source-specific data, which will be in common data model format, also includes the data used to populate the Master Data Management (MDM) of the Integration Platform (IP). The SMR consumes copies of entire databases, files, and other types of extracted data to measure and improve data quality. The data will then be made available in an approved format (schema) to systems inside of the HHS 2020 Enterprise.

Logical Components and Data Flow Diagram

The SMR is built on the platform of MarkLogic, a multi-model NoSQL database. The SMR has the following components that will be independently built, tested, deployed, maintained, and used repeatedly for new sets of data:

- Data Access Module (01: DAM)
- Data Ingest Module (02: DIM)
- Source-specific Integration Module (03: SIM)
Data Access Module (DAM)

The DAM accesses the data at the source using a standard set of processes and technologies that are based on the type of the data source being accessed. These include Relational Database Management System (RDBMS) systems, NoSQL middle-tier databases, web Application Programming Interface (API) enabled integrating systems, and file-based data sources. The accessed data is then stored as Extensible Markup Language (XML), or delimited Comma-separated Values (CSV) files, within the DAM.

The DAM supports the following access patterns:

- File-Based Access (Batch) – to access the file-based data sources
- Extract Transform and Load (ETL) Based Access (Batch) – to support the RDBMS data sources
- Representational State Transfer (REST)-Based Access (Real-time) – to support the web API-enabled integrating systems

The DAM will be implemented using the following tools: Oracle Managed File Transfer (MFT), Oracle Data Integrator (ODI), and custom-built Java RESTful Web Services.

Data Ingest Module (DIM)

The DIM transports raw format data from the DAM into the RDL. The DIM supports real-time ingestion of data into the SMR database as well as batch ingest of bulk-data.

The DIM supports the following ingest patterns:

- RESTful Web Services are used for real-time data ingest
- Batch ingest of bulk data
- Change Data Capture (CDC) of incremental data

The DIM will be implemented using the following ingestion tools:

- MarkLogic Content Pump (MLCP)
- Content Reprocessing in Bulk (CoRB)
- Data Movement Software Development Kit (SDK)

Benefits of the RDL

NoSQL databases handle high data volumes, provide extreme performance, and maintain an integrated view of the information in the database. The key benefits of using the NoSQL based RDL are provided in the table below.

Table 3: Benefits of the RDL

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not tied to a single structure</td>
<td>Relational databases, such as Omnicaid, represent and store data in a structure (tables with columns and rows). The non-relational RDL does not have a fixed schema for the data format, which allows the RDL to load any data in any format and interpret the data later.</td>
</tr>
<tr>
<td>Serves as integration point</td>
<td>The RDL stores the enterprise data enabling the MDM programs to create a materialized view of the data by integrating related data from disparate sources.</td>
</tr>
<tr>
<td>Accepts data in any format</td>
<td>Collects data from disparate sources, without formatting or processing the data to fit into a specified structure, thereby: • Easing the burden on trading partners. • Minimizing data integrity risk as the RDL stores data in its entirety and without any changes to its structure.</td>
</tr>
<tr>
<td>Single source for all raw HHS 2020 data and all modifications to that data over time</td>
<td>The RDL: • Provides audit trail for data sources and updates. • Supports repeated access to the original data and its modifications. • Provides input for data scientists who identify data-driven innovations.</td>
</tr>
<tr>
<td>Horizontal scalability across commodity technology</td>
<td>By design, the NoSQL database spreads data across an arbitrary number of servers, and automatically balances data and query loads across servers, independent of the composition of the server pool. Therefore, the RDL can easily: • Scale horizontally, and support data replication across commodity technology (instead of scaling up, requiring specialized servers and software). • Ensure high availability and disaster recovery.</td>
</tr>
<tr>
<td>Timeliness of data</td>
<td>The SMR can stream information directly into the RDL, providing an almost real-time view of the data.</td>
</tr>
</tbody>
</table>

Source-specific Integration Module (SIM)
The SIM refers to the materialization of heterogeneous data into a common data model. Materialization will standardize and transform the data, bringing it up to industry-accepted standards. SIM materializes the documents from RDL into the Standardized Data Store (SDS).

The following are the objectives of the SIM:

- Standardization of dates and other fields
- Enrichment of data with additional information gathered from other data sources
- Create indexes on selected data attributes for faster searching
- Leveraging of semantic triples/relationships
- De-normalizing multiple data sources into one document
- Management of historical data

The SIM is implemented within the SMR database.

Benefits of the SDS

The table below highlights the features and benefits of the SDS.

**Table 7: Benefits of SDS**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consolidates and materializes data from the HHS 2020 enterprise data source systems.</td>
<td>Provides a source of truth for the MMISR BPO systems and the enterprise MDM.</td>
</tr>
<tr>
<td>Cleanses the data across the HHS 2020 enterprise data source systems.</td>
<td>Centralized data replaces the need for point-to-point connections and costly maintenance of individual system-to-system interfaces. Data cleansing increases data accuracy across the enterprise.</td>
</tr>
<tr>
<td>Facilitates merging volumes of structured and unstructured content into a coherent, single view.</td>
<td>Provides the ability to extract relevant data from both structured, such as RDBMS, XML and JSON, and unstructured, such as Portable Document Format (PDF) and Microsoft Excel spreadsheet (XLS), sources.</td>
</tr>
<tr>
<td>Provides an audit trail, documenting an activity performed on a record.</td>
<td>Provides accountability – who, when, where and how a document was accessed or updated. Helps resolve conflicts in data quality. Helps reconstruct what happened before and during an event.</td>
</tr>
<tr>
<td>Provides tools and workflows to perform the data improvement activities both manually and automatically.</td>
<td>Cleanses data collected from disparate sources and resolves data quality issues. Provides tools for data stewards and data owners to analyze the source data in a holistic way.</td>
</tr>
<tr>
<td>Provides Create, retrieve, update, and delete (CRUD) operations on documents and metadata. Enables document search functions, semantic graphs, and query lexicon values, using query formats, including:</td>
<td>Supports the SOA architecture.</td>
</tr>
<tr>
<td>• String query</td>
<td></td>
</tr>
<tr>
<td>• Structured query</td>
<td></td>
</tr>
<tr>
<td>• Combined query</td>
<td></td>
</tr>
<tr>
<td>• XQuery</td>
<td></td>
</tr>
<tr>
<td>• SPARQL2 query</td>
<td></td>
</tr>
</tbody>
</table>

The SDS improves the data quality – when materializing data, the SDS may encounter data quality issues such as data inconsistency, data accuracy, and data completeness. The data cleansing standardizes document information and includes:

- Correcting data elements to align with external regulatory standards, for example, ensuring that the SDS stores street address, city, state, and postal code information in a uniform and consistent format.
- Using consistent date formats for all date related data.
- Ensuring the SDS follows the standard HHS 2020 enterprise common data model.
- Resolving or flagging data redundancies by consolidating the duplicate records into a single record.
- Enforce data governance rules, and to flag non-conforming records. The following are some examples of data governance rules applied in the SMR:
  - Proper document structure
  - Correct data types
  - Allowed data ranges and allowed values
  - Mandatory fields
Deliver

This module exposes the data in the SMR for processing and consumption in the form of web services, data files, and secured direct access.

This module supports the following deliver patterns:

- Bulk data migration as files
- SQL client-based Migration
- RESTful Web Services-based real-time data migration

The SMR will support the above patterns by leveraging the capabilities of the following tools:

- MLCP (for bulk file extracts)
- Oracle MFT (to transfer bulk-extracts)
- MarkLogic REST Server (for real-time system access)

Technical Features and Product Capabilities

- Robust, and repeatable product
- Capable of processing high volumes of data
- Standardizes disparate data from multiple sources into a given data structure
- Data loads are fault tolerant, picks up from where it left
- Detects and removes duplicate data
- Highly configurable
- Compatible with various data models including the HHS 2020 enterprise data model
- Supports variety of data formats including relational, NoSQL, text, web, structured, and unstructured
- Exports data in the native format
- Supports role based data access for data analysis, and administration
- Uses commodity cluster computing techniques, and low-cost data storage to achieve scalability
- Integrates with industry standard analytical and reporting tools

The SMR Technical Features and Product Capabilities to the HHS 2020 SMR implementation/requirements mapping

<table>
<thead>
<tr>
<th>SMR Technical Capability</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust, and repeatable product</td>
<td>SMR is built to repeat the data standardization process and integrate data from multiple business entities into a single document for variety of data domains. And the process shall be repeatable. As documented on the following page - SMR Data Delivery, the SMR has been configured to repeat the data materialization process for different data domains, over and again. The data has been successfully delivered to the DS vendor as indicated in the attached page.</td>
</tr>
<tr>
<td>Capable of processing high volumes of data</td>
<td>The Omnicaid production dump has been processed through SMR for various data domains and the counts have been verified using test cases. As shown in the SMR Data Delivery, the SMR has ingested and materialized high volumes of data for different data sets from Omnicaid source system.</td>
</tr>
<tr>
<td>Standardizes disparate data from multiple sources into a given data structure</td>
<td>SMR has the capability to configure the standardization process for multiple data sources. However, so far, only Omnicaid source has been configured and tested. Here’s a few test cases -</td>
</tr>
<tr>
<td>Data loads are fault tolerant, picks up from where it left</td>
<td>The SMR does recover from the system state loss. This has been observed during the system failures caused by CPU-overheat. In such cases, the SMR picks up from the last commit, and continues with the process. It also waits, and retries three times before shutting itself down. If this is observed again, we shall capture the log files and upload here.</td>
</tr>
</tbody>
</table>
Detects and removes duplicate data
The SMR soft-deletes the duplicate rows of data. This feature has been tested for all the data domains. Here are a few test cases -

<table>
<thead>
<tr>
<th>Test Case</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIINST-531</td>
<td>Getting issue details...</td>
</tr>
<tr>
<td>SIINST-542</td>
<td>Getting issue details...</td>
</tr>
</tbody>
</table>

Highly configurable
The SMR is configuration based and provides many configurable parameters. All the available configurations are discussed in the following section of SMR design:

SMR and Data Conversion Software Detailed Design#Properties-File

Compatible with various data models including the HHS 2020 enterprise data model
The SMR's compatibility with the HHS 2020 enterprise data models is demonstrated in all the SIM test cases, for all data domains, and the SMR Data Delivery page clearly documents this.

Link to the SI Physical Data Models (PDM).

Supports variety of data formats including relational, NoSQL, text, web, structured, and unstructured
Not Applicable. All the HHS 2020 data sources are relational in nature.

Exports data in the native format
Though native format data exports are supported, the HHS 2020 standard XML feature has been implemented and tested.

Here are a few test cases -

<table>
<thead>
<tr>
<th>Test Case</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIINST-444</td>
<td>Getting issue details...</td>
</tr>
<tr>
<td>SIINST-1840</td>
<td>Getting issue details...</td>
</tr>
</tbody>
</table>

Supports role based data access for data analysis, and administration
The SMR is configured for specific roles, and users. This feature has been tested with the following test cases -

<table>
<thead>
<tr>
<th>Test Case</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIINST-3517</td>
<td>Getting issue details...</td>
</tr>
<tr>
<td>SIINST-3518</td>
<td>Getting issue details...</td>
</tr>
</tbody>
</table>

Uses commodity cluster computing techniques, and low-cost data storage to achieve scalability
Scalability is demonstrated through adding more nodes to the SMR cluster as shown in the below confluence page -

SMR/MDM - QAT Environment - Server and Application Configuration
SMR - UAT Environment - Server and Application Configuration

Integrates with industry standard analytical and reporting tools
Not Applicable.

---

**SMR - Software Detailed Design**

**Design Considerations**

The following are the design considerations factored in the production of this design document. It explains the design goals, factors influencing design decisions and the overall design approach.

**Goals and Guidelines**

The MITA framework and CMS seven conditions and standards primarily drive the design of the SI Platform, and in turn, the SMR. Some of these key drivers are mentioned below as the goals and guidelines:

- **Modular Approach** – The SMR architecture enables a highly available, horizontally scalable solution to easily expand the infrastructure needs, if the system load increases in the future. This modular approach is intended to create a framework aligned with MITA version 3.0, which supports NM HSD’s goal of operating Medicaid functions at a MITA maturity level 4 in all business and technical areas.

- **Compliance with federal standards and State guidelines** – The SMR solution will comply with the CMS Seven Conditions and Standards, promote the use of industry standards for information.
exchange, and provide a seamless business services environment for users. The system also complies with MITA 3.0 requirements and standards and well as the CMS EA. The NM HSD supplied Enterprise Reference Architecture and the Technical Management Strategy documents the design guidelines and best practices to help in Architecture Governance. The Architecture Review Board (ARB) and Data Management Group act as governance bodies to review and evaluate design choices presented by SI as part of the continuous evolution of the SI Platform component design.

- **Tools-Driven** – The system design will maximize the use of readily available MarkLogic system while developing the data migration repository. The use of COTS products is preferred to minimize custom implementations.
- **Reusability** – The SMR system will promote reuse of IT assets within and outside HHS organizations and systems.

### Development Methods and Contingencies

The SI Contractor’s design approach takes into consideration all the requirements, standards, and factors influencing the design and supports an iterative evolution of this complex SI Platform. A component driven design of SI Platform that is driven by functional and non-functional and best practices is at the heart of providing inter-module communication platform. This component design at a high level is captured in the System Design Document and acts as the overall blueprint for an iterative evolution of the individual components. These components are refined for gradual evolutionary changes involving configuration and/or customization, which are driven by newly discovered business requirements. The healthy eco-system of tools like Jama, Jira, and other Software Development Life Cycle (SDLC) aiding tools help in this evolution. The SMR component design artifacts follow SDD as well as NM HSD/SMA supplied Enterprise Reference Architecture/Technical Management Strategy documents.

Figure below explains the overall design approach.

### Architectural Strategies

The following are the design decisions and/or strategies that affect the overall organization of the components of the SI Platform. These strategies provide insight into the key abstractions and mechanisms used in the system architecture.

The SI Platform encompasses the core infrastructure that will enable migration from the existing MMIS, communications across the MMISR solution as well as HHS 2020 participating modules, secure access to data and processes, functionality to support MMIS operations, data transfer, and data integrity.

The following are the software architecture guiding principles and architecture decisions that will be adhered to while designing and implementing the SI Platform.

#### Adherence to Standards:

- Build a standards-based SI Platform to integrate disparate systems and support different integration models (batch, real-time, etc.).
- The Architecture follows standards like MITA, National Information Exchange Model (NIEM), Health Insurance Portability and Accountability Act (HIPAA) and Health Level Seven International (HL7).
- Message Standardization - The solution will implement standard canonical message exchange data models between the legacy and new MMISR modules using NIEM, National Human Services Interoperability Architecture (NHSIA) and Federal Health Information Model (FHIM) models and message standards.

#### COTS Products:

- SI Platform prefers to use the COTS software and configure them in place of custom code where ever possible unless required to support the required solution. The complete list of software, tools (including COTS) and libraries required to build the SI platform can be found in Appendix G.
- The decision to implement a COTS solution for implementing SI platform enhances the ability to provide reliable, proven solution.

#### Security:

- Integrated security and privacy – The solution will comply with MARS-E specified security controls and privacy.
- The Architecture will meet federal and State security compliance requirements.

#### Performance:

- Augment traditional integration solutions with real-time integrations to reduce latency, improve decision-making, and responsiveness.
- Adaptable, extensible, and scalable – The SMR Platform will implement a modular and loosely coupled fashion to accommodate future expansion in response to evolving MMISR business requirements.
- Performance Scalability – The system design will meet performance and scalability needs.

#### Interoperability:

- The Architecture will follow SOA design principles to ensure seamless functionality between the SI Platform and other entities.
- The architecture will support and interoperability of discrete services across multiple systems.

#### Governance:

- Build a centralized governance mechanism to analyze, monitor the usage of the platform and maintain a record of resource levels and consumption within the solution.

### Architecture Design
The software architecture of the SI Platform includes the Enterprise Service Bus (ESB) for loosely coupled and reusable, vendor-, product-, and technology-agnostic, easily discoverable and interoperable services integration with rest of the modules of HHS 2020 enterprise solution and external systems and the Enterprise Shared Systems by imposing security and integrity. It also encompasses the components responsible for migrating data from the legacy systems to the enterprise data repository (SMR) and managing the Master Data (MDM). The SI Platform ESB architecture is consistent with HSD's HHS 2020 enterprise architecture and includes best practices regarding an efficient and sustainable approach to implementing the State’s 2020 vision.

The SI platform leverages many of the COTS products such as MarkLogic to implement the SMR system. MarkLogic provides a NoSQL database into which large volumes of data of any schema can be ingested, normalized, de-duplicated, harmonized, and mastered into a single source of truth, which can be accessed by stakeholders for data migration, data cleansing, and analytics. The combination of these technologies facilitates the system integration, collection, and management of data NM HSD Medicaid operations.

The data integration is achieved in the SMR and the MDM solutions.

The SMR will have the following logical components:

- The SMR core framework that is a repeatable, fault-tolerant, high-available, NoSQL database based data standardization and transformation process.
- The Raw Data Lake (RDL) that stores the legacy systems’ data in their raw format.
- The Standardized Data Store (SDS), which stores the enterprise MMISR data, standardized to the common data models.
- The Metadata Repository, where the metadata from all of the data sources and the SMR is stored, cataloged, and version controlled.
- The standardized reference code set repository, where the enterprise code sets are defined, and enhanced with the industry Medicaid

Details of the Enterprise Information Architecture has been provided in Section 4.4 of the System Design Document.

**System Design**

The SMR System Design is provided in the following page: SMR - Database Design.

**Software Detailed Design**

The detailed design for SMR components is provided in the following pages:

- SMR - Database Design
- SMR - Data Conversion Software Design
- SMR - SIM - Custom Functions
- SMR - Database Administration
- Data De-Identification Design