

Enpresa-Garapenerako Euskal Agentzia
Agencia Vasca de Desarrollo Empresarial
Business Development Basque Agency

www.spri.eus



SPRI-AGENCIA VASCA DE DESARROLLO EMPRESARIAL

PLIEGO PRESCRIPCIONES TECNICAS

PLATAFORMA DE VIRTUALIZACIÓN DE DATOS

SUMARIO

1. OBJETO DEL CONTRATO	3
2. ESPECIFICACIONES TÉCNICAS	5
2.1. CONECTIVIDAD	5
2.1.1. Tipos de conexiones e integraciones	5
2.1.2. Fuentes de datos	6
2.2. DATOS.....	7
2.2.1. Modelado	7
2.2.2. Entrega	9
2.2.3. Catálogo.....	10
2.2.4. Semántica	11
2.2.5. Caché	12
2.3. SEGURIDAD Y GOBERNANZA.....	13
2.4. OPTIMIZACIÓN Y RENDIMIENTO	15
2.5. CLOUD	17
2.6. PROCESAMIENTO POR LOTES	18
2.7. FUNCIONES DEVOPS.....	18
2.8. DESPLIEGUES	19
3. OTRAS ESPECIFICACIONES	20

1. OBJETO DEL CONTRATO

SPRI-Agencia Vasca de Desarrollo Empresarial (en adelante, “**SPRI**”), ente público de derecho privado dependiente del Departamento de Desarrollo Económico, Sostenibilidad y Medioambiente, centra su actividad en la gestión de diferentes tipos de instrumentos de apoyo a las empresas de la Comunidad Autónoma de Euskadi, con el objetivo de contribuir a la mejora de la competitividad de estas y al desarrollo económico de la región. Con carácter general, estos instrumentos suelen materializarse en forma de programas de ayudas o servicios dirigidos a diferentes ámbitos empresariales tales como el Emprendimiento, la Digitalización, el I+D, la Innovación, o el desarrollo industrial entre otros.

Para el desarrollo de esta actividad, SPRI hace uso de diferentes aplicaciones y servicios que se apoyan en diferentes volúmenes de datos. Estos repositorios de datos actualmente se encuentran descentralizados dificultando el acceso a los mismos e imposibilitando disponer de un único punto lógico de acceso a la información.

El objeto del contrato es la implantación de una plataforma de virtualización de datos que nos permita eliminar la barrera que actualmente se le presenta a SPRI en cuanto a la heterogeneidad de las fuentes de datos. A través de esta virtualización se buscará proporcionar un acceso homogéneo, unificado y gobernado a los datos internos o externos satisfaciendo de esta manera las necesidades operacionales, informacionales y técnicas presentes y futuras.

SPRI, dentro de las necesidades que desea cubrir con la virtualización de datos busca proporcionar a todas aquellas aplicaciones/servicios que así lo deseen un único punto lógico que simplifique su accesibilidad y disposición y además les permita evitar conexiones punto a punto complejas y difíciles de mantener. Cumplido ese objetivo se conseguiría desacoplar, y siempre que se requiera, a los consumidores de datos de las fuentes en las que estos se almacenan.

Esta capa de virtualización de datos deberá permitir generar múltiples vistas de negocio sobre los datos físicos para adaptarse a los requisitos específicos de cada usuario y/o aplicación/servicio, es decir, con convenciones de nombrado, formato y esquema específicos. A la hora de disponibilizar estos datos también será necesario que la capa de virtualización ofrezca diferentes soluciones de entrega para poder satisfacer las necesidades de las aplicaciones y servicios que así lo requieran (por ejemplo, mediante accesos JDBC/ODBC, REST/JSON, etc.). De esta forma SPRI podrá disponer de diferentes visiones del dato sin necesidad de duplicar los repositorios y sin necesidad de copiarlos a un repositorio centralizado y optimizando de esta forma sus recursos.

Con esta virtualización de datos, SPRI no solo buscará dotar de nuevas funcionalidades de acceso a los datos a través de sus aplicaciones, servicios y usuarios. También se conseguirá crear una capa de desacoplamiento para facilitar la

evolución de sus bases de datos, aplicaciones y servicios. Esto se quiere conseguir apantallando la capa de datos del resto de sistemas (permitiendo, por ejemplo, realizar una migración de base de datos de una infraestructura manteniendo la misma cadena de conexión y sin necesidad de parar los servicios asociados a esta base de datos).

Apoyándose en lo anteriormente comentado SPRI buscará dotarse de una plataforma de virtualización que pueda dar cobertura a los siguientes objetivos identificados:

- A. Abstracción.** Todos los servicios y aplicaciones que lo requieran deberán poder acceder a los datos a través de un único punto lógico de acceso. Este punto desvinculará la complejidad de las diferentes tecnologías y modelos de los consumidores.
- B. Capa semántica unificada.** Los usuarios designados en la herramienta deberán de poder definir las entidades de negocio representando una definición común para todo SPRI. Estas entidades, además, deberán poder ser expuestas como servicios de datos para que las aplicaciones y usuarios de servicios puedan consumir.
- C. Federación de datos.** La plataforma deberá permitir combinar datos de diferentes orígenes para que el acceso a los mismos se haga a través de la capa de virtualización de datos y sin implicar la necesidad de copiarlos a un repositorio centralizado.
- D. Procesamiento.** La plataforma deberá aprovechar al máximo las capacidades de procesamiento de los orígenes de datos (ejecutar agregaciones sobre grandes volúmenes de datos de manera eficiente, ejecutar consultas sobre datos alojados en la nube, etc.). Además, la plataforma de virtualización deberá minimizar el tránsito de datos a través de la red a la hora de conformar la capa de datos virtualizados.
- E. Gestión de cambios.** Los cambios en las fuentes de datos o en las tecnologías subyacentes se podrán acomodar fácilmente en esta capa, minimizando el impacto sobre las aplicaciones consumidoras. De esta forma se dotará a SPRI de la posibilidad de evolucionar sus sistemas con nuevos componentes tecnológicos con un impacto mínimo.
- F. Capa de seguridad y gobernanza unificada.** La plataforma debería poder definir, administrar y hacer cumplir reglas de seguridad y gobernanza de datos desde un solo punto, y además poder aplicarlas al modelo canónico (tipos, elementos, atributos y relaciones que representan las entidades de negocios) de la organización.
- G. Facilidad de uso.** La solución deberá ser fácil de usar. El proceso de conectarse a las fuentes de datos, combinarlas para crear entidades de negocio y publicarlas

a través de múltiples protocolos, se deberá de poder hacer de modo gráfico. Además, para ayudar a los usuarios de negocio, se deberá disponer de un catálogo de metadatos que permita facilitar la gestión de acceso a los mismos.

- H. Flexibilidad.** La plataforma deberá ser flexible para dar cabida a nuevos modelos y adaptarse de modo ágil a las necesidades de negocio.
- I. Rendimiento.** La plataforma deberá proporcionar rendimiento optimo mediante la combinación de técnicas de optimización avanzadas, posibilitar el uso de cachés. Además, deberá tener un motor de optimización basado en costes para seleccionar automáticamente, en tiempo de ejecución, la mejor estrategia entre todas las posibles, sin depender de configuración ad-hoc.
- J. Características orientadas al usuario de negocio.** La plataforma deberá ser capaz de proporcionar de forma ágil y sencilla a cada consumidor de datos de la organización la información que necesita, en los formatos que necesita y garantizando al mismo tiempo la coherencia, la gobernanza, la calidad y la seguridad de los datos.

2. ESPECIFICACIONES TÉCNICAS

2.1. CONECTIVIDAD

Actualmente SPRI dispone de un sistema heterogéneo de plataformas sobre las que residen los datos con los que opera diariamente (bases de datos, sistemas de ficheros, recursos compartidos, aplicaciones, servicios, etc.). Esta dispersión de los activos de datos dentro de la empresa en ocasiones provoca que se tengan que replicar datos, copiar, etc. Para poder abordar esta problemática se buscará que la plataforma de virtualización sea capaz de:

- A. Proporcionar una capa de abstracción completa de todas las fuentes de datos.** Creando un único punto de acceso desde el cual acceder a cualquier conjunto de datos y de forma que se evite la copia/replicación de los datos y sus derivados mantenimientos.
- B. Permitir la conexión a todos los activos de datos de SPRI.** Además de posibilitar ampliar la lista de conectores de fabrica con conectores personalizados (ad-hoc).
- C. Utilizar conectores que maximicen las capacidades de procesamiento.**

2.1.1. Tipos de conexiones e integraciones

A continuación, se detallan las integraciones que se deberán cubrir a través de la plataforma de virtualización de datos:

- A. Integración nativa** utilizando controladores y APIs proporcionados por otro fabricantes o servicios.
- B. Integración nativa con fuentes de datos en la nube** (Snowflake, Redshift, BigQuery, Presto, Synapse, etc.). Además, deberá posibilitar realizar las conexiones tanto de forma directa o a través de la creación de Cachés.
- C. SDK para el desarrollo** de conectores personalizados.
- D. Uso de APIs nativas de las fuentes de datos** para poder obtener e integrar descripciones de columnas y otros metadatos, así como para la obtención de estadísticas e índices que puedan ser utilizados para mejorar el rendimiento. Además, la plataforma deberá utilizar APIs nativas de carga (BULK data load APIs) de las fuentes de datos (tanto en lectura como en escritura en el caso de la cache) para proporcionar alto rendimiento.
- E. Integración con diferentes tipos de fuentes de datos**, incluidas bases de datos relacionales, bases de datos paralelas y distribuidas, bases de datos multidimensionales (SAP BW, etc.), almacenes de datos en la nube (Amazon, Azure, Redshift, Snowflake, etc.), aplicaciones empaquetadas (SAP, NAVISION, SAGE, etc.), SaaS (Salesforce, etc.), servicios web (SOAP, REST, OData), hojas de cálculo Excel, XML, JSON, ficheros planos, ficheros de log, colas JMS, fuentes en streaming (Kafka, etc.), bases de datos NoSQL (Cassandra, MongoDB, etc.) y el ecosistema Hadoop al completo (HDFS, Parquet, AVRO, HBase, HIVE, SPARK).

2.1.2. Fuentes de datos

A la hora de definir las fuentes de datos a las que se desea conectar la plataforma de virtualización SPRI busca dar soporte tanto a aquellas fuentes que ya se encuentran dentro de la infraestructura actual de la organización como a aquellas que puedan surgir como consecuencia de la evolución de sus sistemas de información. Por lo tanto, se buscará dar soporte, entre otros, a los siguientes tipos de fuentes de datos:

- A. Bases de datos relacionales:** Microsoft SQL Server, ORACLE, DB2, MySQL y PostgreSQL, etc.
- B. Bases de datos NoSQL:** MongoDB, Cassandra y Neo4j.
- C. Bases de datos en memoria:** Redis, Oracle In-Memory, etc.
- D. On Premise DataLakes:** Hive, SparkSQL, Impala, y Presto.

- E. Cloud DataWarehouse/DataLakes:** Redshift, Snowflake, AWS RDS, Databricks Delta, Azure Synapsis, Google BigQuery.
- F. Ficheros delimitados y estructurados:** ficheros de Logs, Excel, XML y JSON.
- G. Ficheros en almacenes de objetos distribuidos:** HDFS, AWS S3, Azure ADLS y Google Cloud Storage.
- H. APIs SaaS:** Google Analytics, ServiceNow, Microsoft Dynamics 365 y Twitter.
- I. APIs Web Services (SOAP, REST, OData)**
- J. Fuentes Streaming/IOT:** Kafka, JMS, etc.
- K. LDAP y Active Directory.**

2.2. DATOS

2.2.1. Modelado

La plataforma de virtualización de datos deberá proporcionar la creación de vistas integradas que permitan combinar los datos de cualquier fuente. Además, y debido a las necesidades de SPRI estas fuentes podrán ser fuentes que se encuentren tanto en la infraestructura de SPRI como en segundas ubicaciones como pueden ser nubes, servicios de terceros, etc.

Para el modelado de datos se proporcionará un entorno amigable en el que los usuarios de la herramienta puedan diseñar sus modelos de datos al igual que hace Microsoft con Microsoft SQL Server Management Studio.

SPRI busca en la plataforma cubrir los siguientes requisitos en cuanto a modelado de datos:

- A. La plataforma deberá proporcionar una interfaz de usuario común** para el diseño y desarrollo. Se esperan interfaces gráficas de usuario basadas en contexto para todas las fases de diseño, incluyendo conectividad de datos, transformación, calidad de datos, optimización de consultas, análisis de impacto, etc.
- B. La interfaz deberá de disponer de diferentes menús y herramientas:** Drag and Drop, autocompletar, ayuda en línea, validación.
- C. La interfaz deberá de disponer de herramientas de depuración.** De esta forma se asegurará un diseño seguro.

- D. Las IU deberá permitir a los usuarios diseñar completamente una capa de datos virtual.** El proceso de conectarse a fuentes de datos, combinarlas para crear vistas derivadas y publicar vistas deberá manejarse a través de un conjunto gráfico de menús, asistentes y acciones Drag and Drop.

Los usuarios deberán poder usar diferentes tipos de vistas derivadas. Estas vistas consistirán en especificar combinaciones/transformaciones entre la vista base y otras vistas derivadas a través de operaciones de Join, Intersection, Union, Minus y otras operaciones de álgebra relacional.

La definición de las vistas derivadas se deberá poder hacer gráficamente desde la herramienta de diseño o utilizando un lenguaje de consulta de virtualización compatible con ANSI SQL.

- E. Los permisos asociados al diseño de la capa de datos** se deberán de poder gestionar a través de la definición de usuarios y roles (ya sea de forma manual o a través del directorio activo de SPRI).
- F. Se proporcionará un entorno de desarrollo 100% web.**
- G. Sintaxis compatible con SQL.** Se proporcionará una alternativa de secuencia de comandos a la interfaz gráfica para el modelado de datos. Para ello deberán de poder utilizarse sentencias DDL y DML.
- H. Soporte nativo para tipos de datos complejos.** Este soporte deberá permitir modelar estructuras jerárquicas como APIs de servicios web.
- I. Tipología de datos.** La plataforma deberá de poder dar soporte a diferentes tipos de datos. Se deberán de soportar tanto tipos básicos (Integer, Long, Float, Double, Boolean, DateTime, XML, Bynary, BLOB, etc.) como tipos complejos (matrices, registros y enumeraciones, datos geoespaciales, etc.).
- J. Metadatos.** La plataforma de virtualización deberá permitir el descubrimiento automático de metadatos y esquemas sobre las fuentes de datos a las que se conecte. La información de los metadatos extraída de los orígenes de datos deberá incluir el tipo de dato en origen, las claves primarias y foráneas, las restricciones NOT NULL, los índices y estadísticas de campo (número de valores únicos por columna, número de nulos, número total de filas, tamaño promedio de columna, valores mínimos y máximos por columna, etc.).
- K. Representación nativa de datos.** Se soportará la representación nativa de tipos de datos compuestos (por ejemplo, datos jerárquicos de XML, servicios web, bases de datos orientadas a documentos, bases de datos no relacionales, etc.). Esto permitirá modelar estructuras jerárquicas como las que, por ejemplo, devuelven las APIs.

- L. Soporte para valores de texto grandes y valores binarios.** Soportará valores de texto grandes como parámetros de entrada o salida. Los valores binarios podrán incluir imágenes, documentos PDF, documentos de Microsoft Office, documentos de texto, etc.
- M. Integración con herramientas de modelado de terceros.** La plataforma debe ser capaz de importar modelos de datos de referencia/estándares de industria y definiciones de esquemas de XSD, WSDL, archivos XML y de herramientas de modelado de terceros (como podrían ser CA ERwin, Embarcadero ER/Studio o, IBM InfoSphere Data Architect).
- N. Librerías de funciones de transformación de datos.** Se espera que la plataforma pueda realizar este tipo de funciones: aritméticas, trigonométricas, procesamiento de textos y fechas, conversiones de tipos, funciones estadísticas y de cifrado, transformación de columna a fila y de fila a columna, condicionales, XML, XPath/XQuery, XSLT, similitud textual, búsquedas y mapeos, operadores de Nest/Flatten para manipulación de datos jerárquicos, expresiones regulares, etc.
- O. Soporte de funciones geoespaciales.** La plataforma de virtualización de datos deberá dar soporte integral para datos geoespaciales de tal forma que se puedan exponer a herramientas geoespaciales de visualización.
- P. Asignación de etiquetas (“tags”) a vistas y columnas.** Se deberá dar la posibilidad de usar etiquetas que ayuden a describir las vistas y columnas. Esto permitirá encontrarlas navegando o mediante búsquedas y servirán para clasificar los elementos del catálogo de acuerdo con criterios que se puedan definir dentro de SPRI (vistas públicas, privadas, etc.).
- Q. Integración nativa con sistemas de control de versiones.** La plataforma deberá de poder integrarse con herramientas como Subversion, Team Foundation Server (TFS) o GIT para desarrollo colaborativo.

2.2.2. Entrega

Dentro de los servicios que pretende cubrir SPRI con respecto a la plataforma de virtualización de datos se encuentra la entrega de los datos virtualizados a las aplicaciones y servicios desarrollados Ad-Hoc para SPRI (aplicaciones de escritorio, servicios, APIs, etc.) como para otros productos de mercado (PowerBI, QlikView, Navision, Salesforce, etc.).

Para poder satisfacer este tipo de necesidades se establecen una serie de requisitos generales que pretenden cubrir las necesidades actuales y futuras con las que SPRI se enfrenta o prevé enfrentarse como motivo de la evolución de sus productos y servicios.

- A. Múltiples interfaces de entrega de datos.** La plataforma deberá de permitir la entrega de datos sincrónica a través de interfaces:
- a. SQL (JDBC, ODBC y ADO.NET)
 - b. API (SOAP, REST representado como XML, JSON, HTML o RSS)
 - c. OData
 - d. GraphQL Web Services.
- B. Entrega asincrónica a través de colas de mensajes.** Como por ejemplo podrían ser JMS o Apache KAFKA.
- C. Creación y despliegue de APIs.** La plataforma deberá permitir crear y publicar APIs sin necesidad de llevar a cabo tareas de desarrollo (Zero-Code). Además, se deberá incluir soporte para SOAP, REST, OData 2 y OData 4.
- Dentro de la funcionalidad de creación y despliegue de APIs SPRI buscará poder generar la documentación de estas utilizando OpenAPI (Swagger).
- D. Entrega de datos para aplicaciones.** Se buscará que la plataforma de virtualización de datos pueda ser utilizada con aplicaciones Business Intelligence y Analítica de Datos. Además de las actualmente utilizadas por SPRI (PowerBI, QlikView, SQL Server Reporting Services, Excel)
- E. Soporte integrado para GraphQL.** Permitirá que las vistas virtuales puedan invocarse vía GraphQL. Además, la plataforma debería de poder encargarse de la combinación de datos desde múltiples puntos.

2.2.3. Catálogo

Junto con el punto anterior, y el cual, hacía referencia a un aspecto técnico, SPRI buscará acercarse a los usuarios de negocio para proporcionarles capacidades de descubrimiento de datos y autoservicio. De esta forma se reducirán las interacciones y dependencias con el Área de Tecnologías de la Información y Comunicaciones y les permitirá explorar tanto metadatos como datos en una única herramienta web y de forma autónoma.

Para lograr el objetivo marcado en la funcionalidad de autoservicio de datos la plataforma deberá cubrir los siguientes aspectos:

- A. Interfaz 100% web.** Se proporcionará una web y esta deberá de integrar un catálogo de datos que permitirá a los usuarios de SPRI (ya sean de negocio o técnicos) buscar y explorar los conjuntos de datos disponibles.
- B. Catálogo sincronizado.** El catálogo deberá estar sincronizado con la plataforma de datos, es decir, cualquier modificación que se haga en una vista/consulta

deberá trasladarse de forma automática e inmediata al catálogo de datos. De esta forma aseguraremos que el servicio sea un fiel reflejo de la plataforma.

- C. Estadísticas.** La plataforma deberá mostrar estadísticas de uso de los conjuntos de datos publicados en el catálogo de datos. También se espera que la plataforma sea capaz de sugerir conjuntos de datos basados en las estadísticas.
- D. Funciones colaborativas.** Los usuarios deberán de poder respaldar conjuntos de datos o registrar comentarios o advertencias sobre ellos. Esto ayudara a contextualizar el uso del conjunto de datos, comprender mejor como los experimentan otros usuarios, solucionar posibles incongruencias, etc.

2.2.4. Semántica

Dentro del área semántica SPRI quiere disponer de una plataforma que le permita modelar, analizar los cambios y visualizar el linaje de todos los datos que se agreguen a esta. También se buscará que la plataforma pueda integrarse con herramientas de terceros para complementar la capa semántica que se pueda ofrecer en la plataforma.

Las capacidades más destacables es que SPRI espera conseguir de la plataforma de virtualización en cuanto a la semántica de los datos serían las siguientes:

- A. Modelado semántico.** La herramienta que se le disponga a SPRI debería de permitir el modelo semántico. Este modelo semántico deberá ser fácilmente gestionado a través de un asistente gráfico.
- B. Agregación semántica.** La plataforma deberá permitir agregar información semántica a los modelos (vistas, tablas, columnas, etc.) utilizando etiquetas. Esta información semántica ayudará a SPRI a establecer políticas y controles generales de acceso (por ejemplo, aplicadas a todas las vistas/columnas etiquetadas con una determinada etiqueta).
- C. Análisis gráfico.** La herramienta proporcionará capacidades de análisis gráfico completo de "linaje de datos" y "uso de datos". El linaje esperado mostrará este en ambos sentidos (desde los orígenes de datos a las vistas de negocio y viceversa).
- D. Detección de cambios en las fuentes.** Para que se puedan soportar los escenarios donde los esquemas de origen cambian dentro de las áreas operativas de SPRI la herramienta deberá proporcionar funciones de actualización de los orígenes de datos y análisis de impacto de cambios de modo que se pueda regenerar el modelo virtual.
- E. Integración con herramientas.** Posibilidad de integración entre soluciones de gestión de metadatos. Como actualmente SPRI no dispone de ninguna de estas

herramientas le gustaría maximizar la compatibilidad con entre herramientas de metadatos como Collibra DGC (Collibra Data Governance Center), IBM IGC (Information Governance Catalog), IMM (Informatica Metadata Manager), Informatica EDC (Enterprise Data Catalog), Talend Data Catalog, etc.

2.2.5. Caché

Actualmente SPRI no dispone de un área informacional del que realizar aquellas tareas que no requieran el acceso a los datos de forma inmediata (Real-Time). Es por eso por lo que se ha identificado la necesidad, y de cara a que las áreas de datos operativas no se vean impactadas, de disponer de la capacidad de cachear determinada información. Utilizando la caché de la información lo que se buscará es reducir el impacto sobre las áreas de datos operativas para aquellas tareas que no requieran de un acceso en tiempo real, como puede ser, por ejemplo: informes business Intelligence, analítica descriptiva/predictiva, machine learning, etc.

Para conseguir estos objetivos las características que se desean cubrir con la plataforma de virtualización son las siguientes:

- A. Replicación selectiva.** Mediante la funcionalidad que permita la replicación selectiva de datos por parte de los perfiles técnicos de la plataforma se espera dar respuesta a los diversos casos de uso que se puedan presentar o ya estén presentes dentro de SPRI. Además, se espera que esta replicación pueda ofrecer diferentes tipologías de caché: incremental, parcial, con TTL, etc.
 - B. Soporte de almacenamiento.** Se espera que la plataforma disponga de un soporte versátil para el almacenamiento y que además este soporte sea ajeno a los orígenes de datos. El propósito será la mejora del rendimiento y el aislamiento de los orígenes de datos, pudiendo evitar así acceder a los orígenes de datos, y que, en el caso de ser de áreas operativas, podrían estar sobrecargados. También, gracias a este soporte, desde SPRI se podría dimensionar este soporte de almacenamiento de cara a mejorar el rendimiento.
 - C. Replicación de datos (ETL, Data Fabric).** Mediante la replicación en tablas externas para poder acceder a los datos, cuando, por ejemplo, se quiere aislar la zona operativa o el acceso directo a la fuente no da un rendimiento adecuado. SPRI buscará que la plataforma pueda cargar fácilmente los datos en una zona “landing” (que no tiene por qué ser en el mismo lugar que la caché) haciendo que la transición sea completamente fluida.
- Esta replicación de datos también será útil para almacenar resultados intermedios en tablas de bases de datos externas.
- D. Rendimiento.** Con el uso de caché de datos se espera mejorar el rendimiento recopilando directamente los datos de está caché. Utilizando está caché de la capa de virtualización de datos se reducirán los tiempos y se aumentará el

rendimiento. De esta forma se podrán programar operaciones complejas que de otra forma retrasarían o degradarían el acceso a los datos (tiempos de espera, sobrecarga de los sistemas, etc.).

E. Múltiples tecnologías. La solución no debe proporcionar un único (o un número muy limitado) soporte de caché, debe soportar el almacenamiento en caché distribuido y flexible basado en cualquiera de estas tecnologías:

- Bases de datos relacionales: Oracle, SQL Server, DB2 o MySQL.
- Bases de datos paralelas: Teradata, Netezza, Vertica.
- Bases de datos en memoria: SAP HANA, Oracle TimesTen u Oracle 12c IM.
- Bases de datos en la nube: AWS Redshift o Azure SQL.
- Sistemas de procesamiento masivo paralelos (MPP): Spark, Hive, Impala o Presto.

F. Ubicación física. Se desea que el cacheado no se haga en propia plataforma, sino en soportes externos configurables. Esta estrategia aprovechará la ubicación conjunta para reducir las combinaciones entre los datos almacenados en caché y los datos no almacenados en caché que se almacenan en el mismo sistema (siempre que sea posible).

2.3. SEGURIDAD Y GOBERNANZA

A. Capa unificada de gobernanza y seguridad. La plataforma permitirá definir, administrar y hacer cumplir las reglas de seguridad y gobierno de datos desde un único punto. Además, dichas políticas se definirán a través de un modelo canónico de entidades de negocio.

B. Soporte completo de gobernanza y seguridad. Este soporte incluirá un enmascaramiento de datos avanzado y un control de acceso basado en roles (RBAC). Este soporte estará disponible para todos los activos de datos de la organización a través de un único punto de control y administración.

C. Único punto lógico de acceso al dato. La virtualización de datos ofrecerá un único punto de acceso lógico que le permitirá a SPRI evitar las conexiones punto a punto de las aplicaciones consumidoras a las fuentes de información.

A medida que se vaya extendiendo el uso de la plataforma de virtualización como punto único de acceso a los datos de la organización SPRI buscará utilizar esta plataforma como herramienta para hacer cumplir las restricciones de seguridad de acceso que se pueden definir en términos del modelo canónico con una granularidad muy fina (p. ej., Acceso a "Proyecto", "Programa de Ayuda", "Importe", etc.).

- D. Acceso de las aplicaciones consumidoras.** La plataforma de virtualización deberá asegurar el acceso de las aplicaciones consumidoras mediante conexiones y protocolos de seguridad tales como SSL/TLS, HTTPS o SFTP.
- E. Autorización y autenticación.** La plataforma deberá admitir mecanismos de autorización y autenticación integrados de usuarios/roles con permisos a nivel de esquema (bases de datos, tablas, vistas, etc.) y permisos a nivel de datos (filas/columnas de una vista, filas/columnas de una tabla, etc.).
- F. Autenticación delegada.** La plataforma deberá de poder delegar la autenticación a un servidor LDAP o Active Directory externo, de forma que no sea necesario definir usuarios en el sistema de administración del que pueda disponer la plataforma. El sistema de autenticación de SPRI proporcionará los roles de usuarios y a su vez la plataforma usará estos para restringir el acceso a cualquier base de datos o vista dentro del servidor de virtualización de datos.
- G. Soporte Single sign-on (SSO).** Donde se buscará que la plataforma de virtualización pueda soportar OAuth, OpenID, Kerberos y SAML.
- H. Soporte de autenticación de doble factor.** La plataforma deberá soportar la autenticación de doble factor de cara a cumplir con las normativas de seguridad que se pudieran definir parte de SPRI.
- I. Soporte para autorizaciones basadas en usuarios y/o roles (RBAC).**
- J. Control de acceso granular.** SPRI buscará a través de la plataforma restringir los accesos de forma granular y llegando hasta el nivel de poder restringir estos accesos en filas y columnas.

La plataforma de virtualización podrá imponer permisos estrictos y detallados de usuario y basados en roles para todos y cada uno de los elementos definidos en ella. Las políticas de autorización podrán implementar la seguridad de filas y columnas. Estas políticas se deberán de poder especificar utilizando alguno de las siguientes combinaciones:

- Fila (aplicando una condición de selección).
- Columna.
- Fila y Columna (aplicando una condición de selección a las filas).

Estas autorizaciones se deberán de evaluar antes de la ejecución de cada consulta para determinar si el usuario tiene permiso para ver los resultados o no. Los datos no autorizados se podrán filtrar (es decir, eliminados de los resultados de la consulta) o enmascarar (es decir, se podrá verificar su existencia, pero no serán visibles para el usuario).

- K. Políticas de seguridad globales basadas en semántica.** Para simplificar SPRI buscará que la plataforma sea capaz de gestionar políticas de seguridad basadas

en semántica (por ejemplo, utilizando etiquetas). Por ejemplo, se deberá de poder definir una política global para que solo los usuarios con un determinado rol puedan consultar las vistas que tienen la etiqueta "SUBVENCIÓN". También, por ejemplo, se deberá de poder definir una política global que haga uso de enmascaramiento para todas las columnas que contengan una determinada etiqueta ("CONFIDENCIAL", "PRIVADO" o similar).

2.4. OPTIMIZACIÓN Y RENDIMIENTO

A la vez que SPRI desea disponibilizar un entorno centralizado de datos a nivel organizativo también buscará que tanto para la conformación de este entorno como para su explotación los recursos estén lo más optimizado posible de cara a reducir el impacto en sus sistemas operativos, los mantenimientos, el número de incidencias derivadas del bajo rendimiento, una experiencia de usuario optima, etc.

Para ello a la plataforma de virtualización de datos se le exigirán una serie de requisitos a fin de cumplir con estos requerimientos.

- A. Optimización de consultas.** Se precisará que la plataforma ofrezca técnicas de optimización de consultas de cara a maximizar el rendimiento. La plataforma deberá optimizar la delegación de consultas en las fuentes de datos sin necesidad de que previamente se tengan que trasladar los datos a la plataforma de virtualización o copiarlos a un segundo repositorio.
- B. Optimización basada en costes.** De tal forma que la plataforma sea capaz de seleccionar el mejor plan de ejecución para cada consulta, y sin la necesidad de depender de análisis y desarrollo manual por parte del personal de SPRI.
- C. Aceleración de consultas basadas en IA.** Se buscará que la plataforma disponga de técnicas inteligentes de aceleración de consultas, basadas en IA, para posibles escenarios analíticos complejos que a futuro se puedan dar dentro de SPRI.
- D. Tráfico de datos.** SPRI considera críticas las capacidades de la plataforma en base a reducir el tráfico de red y la ejecución optima de las consultas en origen. La plataforma de virtualización deberá priorizar la ejecución en el propio origen de datos y minimizar el trasiego de datos a la capa de virtualización (siempre que sea posible). Además, combinando estas técnicas con otras funcionalidades anteriormente destacadas como pueden ser los diferentes tipos de cacheado de la información se pretende conseguir un uso óptimo de recursos y un mínimo impacto sobre los orígenes de datos operativos actuales.
- E. Planes de ejecución.** La plataforma deberá permitir consultas ad-hoc y a su vez será necesario optimizar la ejecución de estas consultas con planes de ejecución dinámicos. De esta forma, y dado que normalmente solo los

usuarios técnicos tienen nociones sobre optimización, esta funcionalidad servirá para monitorizar o identificar posibles problemáticas que se puedan dar en las consultas.

F. Reglas estáticas y optimización basada en costes. La plataforma deberá aplicar tanto técnicas basadas en reglas estáticas (heurísticas) como técnicas de optimización basadas en costes. Además, para que el optimizador basado en costes de la plataforma pueda seleccionar automáticamente el plan de ejecución que mejor convenga se deberían de tener en cuenta factores como:

- Datos estadísticos como el tamaño de las tablas, el número de campos, el tamaño de cada campo, valores mínimo y máximo, número de valores diferentes, etc.
- Las capacidades específicas del origen de datos como puede ser el tipo de origen de datos (bases de datos, ficheros, documentos, etc.), el modo (fuentes de solo lectura o con limitación en la cantidad de datos que puede almacenar) o el tamaño (como pudieran ser el número de nodos de una fuente MPP).

G. Aceleración de consultas en escenarios analíticos. La plataforma debería incluir soporte para datos agregados y summaries e integración con MPP (Multi Parallel Processing) para poder ejecutar consultas sobre grandes volúmenes de datos.

H. Delegación de operaciones de agregación. La plataforma debería situar las operaciones “group by” por debajo de operaciones “join” o “unión” para así poder reducir el número de registros a recuperar.

I. Uniones Particionadas (Partition Pruning). Cuando los datos de una vista de unión se distribuyen en varias fuentes de datos y una consulta solo necesita datos de una fuente de la unión, el optimizador de la plataforma de virtualización deberá eliminar las fuentes que no participarán en la consulta.

J. Técnicas de movimiento de datos. Esta técnica se utilizará en situaciones en las que un pequeño conjunto de datos necesita combinarse con un gran conjunto de datos en otra fuente. El optimizador de la plataforma deberá de poder mover automáticamente los datos de la fuente con el conjunto de datos más pequeño a una tabla temporal en la fuente de datos con el conjunto de datos más grande. De esta forma se podría delegar la ejecución de la consulta completa en un único repositorio obteniendo mejoras de rendimiento significativas.

K. Aceleración inteligente de consultas. El optimizador deberá lograr mejoras significativas en el rendimiento utilizando de manera inteligente resúmenes (summaries) previamente en su plataforma (resultado de consultas

obtenidas/creadas en cualquier otro proceso o para el proceso en cuestión) de datos transaccionales de cara a poder optimizar consultas posteriores.

- L. Resúmenes.** Los resúmenes deberán de poder ser creados por el administrador bajo las sugerencias que le pueda hacer la plataforma de forma automática. Al analizar los diferentes patrones de consulta en un escenario dado y el rendimiento de consultas pasadas, el sistema aplicará un modelo de Machine Learning para identificar resúmenes adecuados (que ayudarán a acelerar consultas futuras). Los resúmenes se podrán almacenar en la caché o en una fuente de datos externa.
 - M. Datos de monitorización.** Se utilizarán datos de monitorización (por ejemplo, métricas de uso, tiempos de ejecución, distribución y volúmenes de datos, etc.) para alimentar el motor de recomendaciones. Este motor se deberá utilizar para ayudar a los administradores a ajustar el rendimiento. Se utilizará inteligencia artificial para analizar las consultas anteriores y recomendar nuevos “summaries” para acelerar las consultas.
 - N. Procesamiento en paralelo.** La plataforma de virtualización debe incluir integración nativa con motores de procesamiento masivo en paralelo (MPP) para la aceleración de consultas en escenarios complejos de Big Data. La plataforma debería poder llevar a cabo el procesamiento paralelo de operaciones delegándolas siempre que fuera necesario en un clúster que haga uso de procesamiento en paralelo que pueden ser aquellos clústeres basados en tecnologías como Spark, Databricks, Impala o Presto.
- Además, será necesario que se proporcione almacenamiento en caché para un acceso rápido a los datos usados recientemente dentro de las tecnologías comentadas (Spark, Impala, Presto, etc.).
- O. MPP Integrado.** El mecanismo de aceleración de consultas MPP debe estar integrado en el modelo de optimización basado en costes de la solución. La plataforma tomará la decisión final sobre si ejecutar el procesamiento final en memoria o delegar este procesamiento al clúster.

2.5. CLOUD

SPRI cuenta con soluciones de arquitectura de aplicaciones y productos en la nube. Por lo tanto, la plataforma de virtualización deberá contar con un soporte completo e integrado para fuentes nativas en la nube (fuentes de datos, cachés o destinos de replicación selectiva).

Bien sea como parte de despliegues de soluciones híbridas o como catalizador de planes de “movimiento a la nube”, la plataforma de virtualización deberá de ser compatible e integrarse con los ecosistemas de los principales proveedores de nube,

en concreto, AWS, Azure y Google Cloud Platform. Las fuentes de datos propias de entornos nube, que en muchos casos se ofrecen en modalidad SaaS, deberán poder ser utilizadas por la herramienta de virtualización para diferentes propósitos: como origen de datos escalables, como sistema de caché de la solución o como destino de materialización selectiva (tablas temporales, summaries, etc.)

2.6. PROCESAMIENTO POR LOTES

SPRI cuenta con un gran volumen de información recabada mediante aplicaciones, productos y fuentes externas. En muchos casos el procesamiento de esta información puede realizarse de forma asíncrona. Por lo tanto, la plataforma de virtualización deberá proporcionar capacidades para la ejecución planificada de tareas habituales como carga en caché, replicación selectiva, monitorización de cambios en las fuentes, etc.

La plataforma deberá de proporcionar un programador de tareas para ejecutar tareas batch. Los usos típicos que se podrían dar serían la exportación de datos a bases de datos o a archivos (ficheros delimitados, CSV, ficheros planos, MS Excel, índices, etc.), cargas de caché y tareas de administración periódicas (recopilación de estadísticas de datos, datos tareas de actualización de caché, buscar cambios en las fuentes, etc.), etc.

2.7. FUNCIONES DEVOPS

DevOps promueve la colaboración, la integración, la visibilidad y la transparencia entre equipos de desarrollo con el resto de los equipos que realizan operaciones IT. SPRI cuenta con una infraestructura DevOps para el despliegue de aplicaciones, y, por lo tanto, la plataforma de virtualización deberá disponer de un sistema centralizado para la gestión de despliegue de solución completo, en los entornos de desarrollo y producción tanto en on-premise y nube como en entornos híbridos.

Además, será preciso que la plataforma de virtualización ofrezca soporte para despliegues automatizados en proveedores cloud, incluyendo, pero no limitado a, AWS, Azure y GCP. También se precisará que la plataforma disponga de gestión automática y monitorización integrada de clústeres, balanceadores de carga, migración entre entornos, autoescalado, etc.

Dentro del ámbito DevOps la monitorización de la plataforma debería poder ser realizada en tiempo real y de forma integrada sobre todos sus componentes, además esta deberá de poder auditar todas las consultas, parametrizaciones y actividades realizadas en la plataforma.

Otras de las necesidades que se buscan dentro de la plataforma son las de poder realizar la una gestión centralizada de despliegues complejos (por ejemplo, con múltiples clústeres y en alta disponibilidad) y una gestión centralizada de actualizaciones, claves de licencia, parámetros específicos de diferentes entornos (por ejemplo, parámetros de conexión a los orígenes de datos), monitorización, registros de log y auditoría.

2.8. DESPLIEGUES

- A. Infraestructuras.** La plataforma de virtualización deberá soportar despliegues en las siguientes infraestructuras:
- a. Infraestructuras on-premise.
 - b. Infraestructuras de nube pública o nube privada y en los principales hyperscalers como son AWS, Azure o GCP.
 - c. Infraestructuras híbridas (on-premise + nube).
- B. Despliegues distribuidos.** La plataforma deberá de ofrecer soporte para para despliegues distribuidos geográficamente (bien por motivos de rendimiento o bien por motivos de cumplimiento de regulación).
- C. Contenedores.** La plataforma deberá de ofrecer la posibilidad de poder desplegarse haciendo uso de contenedores (utilizando, por ejemplo, contenedores Docker).

2.9. LICENCIAMIENTO

- A. Plataforma.** Dado que SPRI cuenta con soluciones en nube y on-premise, la licencia deberá de ser independiente al tipo de despliegue (on-premise, nube o híbrido) y no se deberá incurrir en la conversión de licencias ni costes añadidos por cambio de plataforma.
- B. Dimensionamiento/Costes.** SPRI, y al tratarse de un ente público de derecho privado dependiente del Departamento de Desarrollo Económico, aporta servicio público y busca que la plataforma de virtualización de datos no se encuentre restringida ni incurra en costes derivados del número de consumidores conectados, del número de conexiones a orígenes de datos (i.e. número de fuentes) o del volumen de datos procesados por el motor de virtualización.

3. OTRAS ESPECIFICACIONES

Otras especificaciones que aportarían un gran valor añadido a SPRI en cuanto a las funcionalidades de la plataforma de virtualización de datos son las siguientes:

- A. Solución escalable (horizontalmente, verticalmente).
- B. Alta disponibilidad.
- C. Sistema de copias de seguridad y recuperación antes desastres.
- D. Cifrado de metadatos.
- E. UI para diferentes roles (administradores, desarrolladores, administradores de datos, usuarios comerciales, etc.)
- F. Solución integrada.
- G. Soporte multilinguaje.
- H. Integración a través del empleo de estándares abiertos
- I. Motor MPP integrado: El proveedor deberá proporcionar soporte de este motor incluido en la licencia de soporte del motor de virtualización.