

**Nº Expediente: PA1/2025**

**PLIEGO DE PRESCRIPCIONES TÉCNICAS DEL CONTRATO MIXTO DE SUMINISTRO DESTINADO A LA ADQUISICIÓN DE LICENCIAS DE PRODUCTOS DE SOFTWARE PARA EL PROCESAMIENTO DE LENGUAJE NATURAL, EN FORMATO DE SUSCRIPCIÓN, JUNTO CON LOS SERVICIOS ASOCIADOS DE INSTALACIÓN Y SOPORTE, CON EL OBJETIVO DE DESARROLLAR CAPACIDADES ADICIONALES E INTEGRARLAS EN LA PLATAFORMA TECNOLÓGICA DEL ESPACIO DE DATOS SANITARIOS DE USO SECUNDARIO DE LA COMUNIDAD DE MADRID CON FINES DE INVESTIGACIÓN E INNOVACIÓN “HEALTHDATA@MAD-R&I”, EN CONSONANCIA CON EL ESPACIO EUROPEO DE DATOS EN SALUD, Y CONFORME A LA CONVOCATORIA DEL MINISTERIO DE TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA, EN EL MARCO DEL PLAN DE RECUPERACIÓN, TRANSFORMACIÓN Y RESILIENCIA – NEXT GENERATION EU (PROGRAMA ESPACIOS DE DATOS SECTORIALES), MEDIANTE PROCEDIMIENTO ABIERTO SIMPLIFICADO Y TRAMITACIÓN URGENTE.**

### **Primero. Justificación y Objeto del Contrato**

La presente contratación tiene como objetivo contribuir al fortalecimiento y ampliación de la capacidad de análisis de datos sanitarios para uso secundario con fines de investigación e innovación en salud mediante el suministro de licencias, y la implementación de los servicios asociados a su instalación y soporte, en el entorno tecnológico del *Data Lake* de HealthData@MAD-R&I y su convergencia con el espacio europeo de datos en salud. El objetivo final es mejorar la capacidad para utilizar datos existentes y nuevos para la toma de decisiones en tiempo real y el desarrollo de modelos predictivos en el ámbito de la salud.

El proyecto HealthData@MAD-R&I está financiado, en el marco de la convocatoria para la concesión de ayudas, en el ámbito de la digitalización, para la transformación digital de los sectores productivos estratégicos mediante la creación de demostradores y casos de uso de Espacios de Compartición de Datos, por el Ministerio para la Transformación Digital y de la Función Pública y por la Unión Europea a través de los recursos financieros derivados del Plan de Recuperación, Transformación y Resiliencia-Next Generation EU (Programa Espacios de Datos Sectoriales).

El proyecto HealthData@MAD-R&I tiene como objetivo general definir, desarrollar, validar y documentar casos de uso relacionados con la investigación y la innovación que permitan validar los servicios del Espacio de datos sanitarios electrónicos de uso secundario. En el marco del mismo, se han previsto cuatro casos de uso en diferentes áreas de investigación que sirven como demostradores del potencial científico-tecnológico y comercial que los datos de salud aportan a la economía del dato en el sector sanitario:

- Caso de uso 1: Optimización de las derivaciones de sujetos con enfermedades reumáticas y musculoesqueléticas entre atención primaria y reumatología
- Caso de uso 2: Trayectorias asistenciales de las mujeres largas supervivientes de cáncer de mama SURBCANMADRID
- Caso de uso 3: Hospitalización No Programada (MAD-HNP.eSTRATA)
- Caso de uso 4: Modelo explicativo en vida real sobre la efectividad de las estatinas en la reducción de eventos cardiovasculares y la mortalidad en la población anciana sin antecedentes de enfermedad cardiovascular de la Comunidad de Madrid.

Para lo anterior, se precisa ampliar las capacidades de la Plataforma HealthData@MAD-R&I que dará soporte a los casos de uso, por lo que resulta necesario la adquisición de licencias que cubran todas las necesidades y requerimiento técnicos.

Por todo lo cual, se precisa el suministro consistente en la adquisición de licencias y la implementación de los servicios asociados en el entorno tecnológico del *Data Lake* de HealthData@MAD-R&I para fortalecer y ampliar la capacidad de análisis de datos sanitarios, con el objetivo final de mejorar la capacidad para utilizar datos existentes y nuevos para la toma de decisiones en tiempo real y el desarrollo de modelos predictivos en el ámbito de la salud. La plataforma HealthData@MAD-R&I se alinea con las directrices y especificaciones técnicas para facilitar el uso ininterrumpido de los datos sanitarios en toda Europa en el marco del Espacio Europeo de Datos Sanitarios (EHDS).

En concreto, el objetivo de este contrato es implantar una solución de Procesamiento de Lenguaje Natural para extraer valor de los datos sanitarios existentes en la organización para tomar decisiones informadas en el ámbito de la salud, incluyendo:

- Facilitar la anonimización de datos de carácter personal en documentación clínica
- Desarrollar herramientas de Detección de Entidades (*Named Entity Recognition* – NER – por sus siglas en inglés)
- Codificar a terminología estándar las entidades reconocidas

Dicha solución se integrará con la plataforma *Cloudera Data Platform*, la cual centraliza datos y herramientas para iniciativas de analítica descriptiva y avanzada por el órgano competente en materia de Digitalización de la Comunidad de Madrid.

Las características técnicas principales de los programas y prestaciones a contratar se especifican en el punto tercero del presente pliego.

## **Segundo. Entorno Tecnológico Actual**

HealthDataMADR&I formará parte del Espacio de Datos de la Comunidad de Madrid, que actualmente utiliza la plataforma *Cloudera Data Platform*, con los componentes que se describen en el punto tercero del presente pliego. Dicha plataforma centraliza datos y herramientas para iniciativas de analítica descriptiva y avanzada.

## **Tercero. Características de las prestaciones**

Con respecto a las licencias objeto del contrato, se admiten programas:

- Puestos a disposición en modalidad de nube.
- Para su instalación en infraestructura local.
- En cualquier modalidad de puesta a disposición.

### **1. Requisitos funcionales de los programas a suministrar**

La Comunidad de Madrid, dentro del Plan de Salud Digital de la Consejería de Digitalización, tiene en curso varias iniciativas que persiguen la definición, diseño y desarrollo de nuevos sistemas analíticos con modelos predictivos que permitan extraer valor de los datos existentes,

sanitarios y no sanitarios, del Servicio Madrileño de Salud; como por ejemplo: optimizar la gestión global y local de la capacidad, alerta temprana de variabilidad en la práctica clínica, predicción, seguimiento y actuación en el ámbito de la cronicidad, detección temprana de desajustes y alertas epidemiológicas, todo ello para su aplicación en la toma de decisiones.

Actualmente, y dentro del proyecto Espacio de datos de salud de uso secundario con fines de investigación científica e innovación HealthData@MAD-R&I se precisa gestionar y procesar un gran volumen de datos.

Esencialmente, para gestionar y procesar volúmenes ingentes de datos y para detectar patrones en los mismos, se utilizan tecnologías denominadas como *Big Data*, las cuales son capaces de utilizar dichos datos para extraer conclusiones de valor.

Actualmente se dispone de un repositorio central a la fecha, con más de 13.000 millones de registros (*Data Lake*) que se sustenta en una infraestructura de *Big Data* basada en Cloudera (Data Platform private Cloud (CDP)) para el tratamiento de información, en el CPD Central de la DGSD Athene@ en el Hospital 12 de octubre. Está estructurado modularmente para que los procesos de alimentación e ingesta, así como de explotación de datos, permitan interconexión a otros subsistemas de análisis de datos. Además, este *Data Lake* permitirá la interoperabilidad de informaciones con la Historia Clínica del Sistema Nacional de Salud, proyecto que promueve el Ministerio de Sanidad para mejorar el acceso a la información.

Por tanto, la actual infraestructura está compuesta de varios nodos que disponen implementados los distintos roles en nodos diferenciados.

#### Nodos de proceso para el plano de Control

Estos nodos tienen implementado los roles del servicio de kubernetes y el Cloudera Experiences (Cloudera Data Engineering (CDE)/Cloudera Machine Learning (CML)/Cloudera Data Warehouse (CDW)), el plano de control y el catálogo de datos de Cloudera.

#### Nodos Master de Cloudera o equivalente

Estos nodos tienen implementados los roles HDFS, YARM, HBASE, sentry server, StateStore server y servidor de catálogo, IMPALA y Cloudera Manager.

#### Nodos Worker y/o Almacenamiento

Estos nodos tienen implementados para el almacenamiento con los roles HDFS Data node, YARM node manager, HBASE región server, IMPALA Daemon, Search worker daemons y Kudu Tablet.

La solución objeto de este contrato debe permitir el uso de técnicas avanzadas de Procesamiento de Lenguaje Natural para identificar y extraer entidades relevantes, anonimizar datos personales procedentes de las Historias Clínicas Electrónicas y codificar la información a terminologías sanitarias estándar. Asimismo, se integrará de manera transparente con el clúster de Cloudera, aprovechando las capacidades de la plataforma para la ingesta, procesamiento, análisis y disponibilización de grandes volúmenes de datos, garantizando así la escalabilidad y rendimiento necesarios para cubrir distintos casos de uso del proyecto HealthData@MAD-R&I y, por tanto, asimismo, en el ámbito del órgano competente en materia de Digitalización de la Comunidad de Madrid.

Se precisa que las librerías puedan ejecutarse sobre Spark 3.X y se permita la reutilización, el entrenamiento y la combinación de modelos de IA para tareas como el reconocimiento de

entidades, la clasificación de texto, la corrección ortográfica y gramatical, la respuesta a preguntas y la extracción de conocimiento.

Se debe permitir la preparación de datos para las soluciones RAG LLM, incluyendo la división de documentos, la limpieza, el enriquecimiento de metadatos, el resumen y el cálculo de incrustaciones.

Se solicita que el producto licenciado cumpla con:

- Extracción de información
- Clasificación de documentos
- Desambiguación de entidades
- Análisis contextual
- Puntuación del riesgo del paciente
- Ofuscación de datos
- Coherencia de nombres
- Coherencia de género
- Grupo de edad
- Formato
- Coherencia Gramática clínica
- Detector profundo de frases
- Corrección ortográfica médica
- Mapeo terminológico
- Aprendizaje Zero-Shot
- Entidades por pregunta
- Relaciones por pregunta
- Clasificación por Prompt
- Extracción de datos relativos

## **2. Servicios de instalación y soporte de los programas a suministrar**

El adjudicatario realizará todas las tareas de instalación de las librerías en la plataforma y las tareas de soporte necesarias para garantizar el correcto funcionamiento de la herramienta y de la actividad diaria de la plataforma HealthData@MAD-R&I, garantizando los objetivos establecidos en el presente pliego.

El adjudicatario ejecutará el contrato de forma remota desde sus propias oficinas, de acuerdo con las políticas de seguridad establecidas por el órgano contratante, así como el órgano competente en materia de Digitalización de la Comunidad de Madrid.

El adjudicatario deberá tener el aval del fabricante de software para comercializar el producto y la gestión del soporte y mantenimiento especializado.

Los recursos humanos a incorporar al contrato, aunque los licitadores podrán incorporar un número mayor de recursos si lo consideran conveniente. A continuación, se presenta la estimación de los recursos humanos necesarios para el desarrollo del contrato, de forma que los operadores económicos puedan disponer de margen suficiente para configurar su organización con la mayor eficiencia y ajustar de forma óptima su oferta.

Servicios / tareas a desarrollar	Perfil necesario	Número de recursos necesarios
Soporte extendido sobre licencias de la plataforma	Científico de Datos	1
Servicio de instalación y despliegue	Ingeniero de Datos	1

Ambos perfiles (Ingeniero/a de Datos y Científico de Datos) deben cumplir con dos tipos de requisitos:

**1. Titulación académica:**

- **Ingeniero/a de Datos:** Grado universitario de Nivel 2 (MECES 2) en Ingeniería o titulaciones técnicas.
- **Científico de Datos:** Grado universitario de Nivel 3 (MECES 3) en Biotecnología, Ingeniería u otras titulaciones técnicas.

**2. Certificaciones técnicas del fabricante:**

- Además, para ambos perfiles se requieren certificaciones del fabricante que acrediten formación técnica en el producto.
  - *Generative AI for Data Scientists*
  - *Medical Language Models for Data Scientists*

**3. Cobertura de la garantía extendida del adjudicatario**

La garantía extendida que debe prestar el adjudicatario durante todo el periodo de vigencia de las licencias se rige por:

- Soporte de nivel 1 y nivel 2 prestado por el adjudicatario a petición del organismo destinatario, en los términos descritos en el PPT.
- Soporte del adjudicatario al organismo para el acceso a la garantía del fabricante (acceso al soporte de nivel 3).
- Soporte a la instalación de actualizaciones.

Horario de contacto: Asistencia Empresarial durante 8 horas, 5 días a la semana (8x5).

**Quinto. Plazo de entrega de las licencias**

El plazo máximo de entrega será 10 días naturales contados a partir de la fecha de inicio de ejecución del contrato. No se admiten entregas parciales.

**Sexto. Plazo de ejecución del contrato**

El contrato tendrá una duración estimada de once (11) meses, puesto que se prevé su formalización el 1 de julio de 2025. No obstante, con independencia de la fecha efectiva de

formalización, el contrato deberá finalizar, en todo caso, el 15 de junio de 2026, por lo que la duración definitiva del mismo se ajustará a dicho plazo límite. La eventual reducción del plazo efectivo de ejecución no conllevará una disminución del importe de adjudicación, puesto que el presupuesto del contrato se ha calculado en función del cumplimiento de los resultados esperados, con independencia del tiempo requerido para su consecución.

### **Séptimo. Coordinación del contrato**

A efectos de lograr una correcta ejecución del contrato se designa como coordinador del mismo a la responsable del proyecto de HealthData@MAD-R&I en la Fundación para la Investigación e Innovación Biosanitaria de Atención Primaria (FIIBAP).

En Madrid, a fecha de última firma electrónica  
La Presidencia del Patronato

Fdo.: Ana Isabel González González  
Jefe de Proyecto.