



Pliego de Prescripciones Técnicas para la Implantación de una infraestructura Cloud que permita la interoperabilidad de los datos de la Dirección General de Estadística de la Comunidad Autónoma de Madrid

Objeto del contrato	2
Contenido del trabajo	2
Elaboración de documento de Análisis Funcional y Diseño de arquitectura	2
Elección de tecnologías e infraestructura de la arquitectura.....	3
Proveedor Cloud y elección de suministros IaaS o PaaS.....	3
Período de servicio activo con el presupuesto asignado.....	3
Elaboración de documento de diseño de proceso de ingesta de información.....	4
Herramientas de ingesta de información.....	4
Herramientas para la monitorización y tracking histórico de las ingestas	4
Herramientas de transformación de información	5
Herramienta para el desarrollo del catálogo de datos interno	5
Herramienta para la disponibilidad externa de la información	5
Herramienta de Business Intelligence.....	6
Ecosistema de proyectos de Data Science y Machine Learning	6
Elementos de gestión del software y elementos de despliegue.....	6
Organización y seguimiento del proyecto	7
Composición del equipo del proyecto	7
Dependencia funcional del personal	7
Metodología de seguimiento del proyecto y comunicación con el personal involucrado de la Dirección General de Estadística.....	8
Recursos materiales del proyecto	8
Lugar de ejecución del proyecto	8
Plazo de ejecución	8
Presupuesto	8
Pago de la prestación	8
Secreto estadístico	8

Objeto del contrato

El objeto de este pliego es establecer las condiciones técnicas para la contratación, mediante procedimiento abierto, del proyecto titulado: “Implantación de una infraestructura Cloud que permita la interoperabilidad de los datos de la Dirección General de Estadística de la Comunidad Autónoma de Madrid”.

El proyecto consiste en plantear, diseñar e implantar una arquitectura basada en infraestructura Cloud compuesta de distintos suministros que permita a la DGE de la CAM la ingesta, almacenamiento, procesamiento de información procedente de distintas fuentes, tanto de terceros que suministran datos para la elaboración de informes como de fuentes de datos de elaboración propia. Estos datos deberán estar estructurados, relacionados, normalizados y gobernados sobre dicha infraestructura y, adicionalmente, se establecerán distintas herramientas para consumir esta información de acuerdo a los distintos perfiles técnicos que accederán a las mismas. En este sentido, se tendrá que tener en cuenta cómo accederán a los datos profesionales con las siguientes características y a través de qué características:

- Analistas de datos, acostumbrados a utilizar herramientas tipo Excel para integrar manualmente las fuentes de información con la que trabajan y elaboran informes a partir de las conclusiones que son capaces de extraer.
- Analistas de Business Intelligence, con un perfil más técnico en cuanto a la extracción de datos desde bases de datos SQL y capaces de definir métricas y crear cuadros de mando.
- Data Scientists, con capacidad de análisis de datos más profunda, con capacidades de programación en lenguajes Open Source

Quedan fuera del alcance del presente pliego de prescripciones técnicas todos aquellos suministros profesionales asociados con la ingesta o explotación de la información, quedando exclusivamente ceñido al diseño e implantación de la plataforma y las pruebas de acceso entre los distintos suministros que se planteen en la misma. Queda fuera también del ámbito del proyecto cualquier tipo de tarea de gestión del cambio o formación al personal de la DGE de la CAM.

Contenido del trabajo

Elaboración de documento de Análisis Funcional y Diseño de arquitectura

Fase de recopilación de información y análisis para elaborar, por un lado, un listado de requisitos funcionales detallados que debe satisfacer la infraestructura planteada. El potencial adjudicatario deberá proponer una metodología para crear este listado de

requisitos finales, más concretos y específicos que los recogidos en la presente propuesta, a partir del entendimiento del contexto de la DGE de la CAM. Será a partir de esta definición funcional a partir de la cual se desarrollará el diseño final de la plataforma, indicando en todo momento qué suministros o características de los mismos impactan positivamente en el cumplimiento de cada uno de los requisitos establecidos.

Aunque pueda haber variaciones en el diseño final de acuerdo con estos requisitos, en la propuesta enviada como contestación al presente pliego de prescripciones técnicas debe quedar establecido, al menos, los siguientes puntos:

- Proveedor de infraestructura cloud que soporte la plataforma
- Suministros de *Infrastructure as a Service* o *Platform as a Service* propuestos por el proveedor y finalidad concreta de cada uno de estos suministros.
- Herramientas Open Source que se propongan para las distintas tareas y en qué componentes de la infraestructura se utilizarán. Se valorará muy positivamente el uso de este tipo de herramientas, cualificando también aquellas tecnologías con un nivel de madurez suficiente.
- Herramientas propietarias que se propongan en los distintos procesos, costes anuales de las mismas y justificación frente alternativas Open Source.
- Estimación de costes anuales de cada componente propuesto.

Elección de tecnologías e infraestructura de la arquitectura

Proveedor Cloud y elección de suministros IaaS o PaaS

La propuesta debe contener una justificación de las motivaciones de la elección del proveedor de suministros Cloud, así como la experiencia pasada en proyectos de características similares en este tipo de infraestructuras. Dentro de dicha plataforma, el proveedor indicará todos los suministros que propone contratar dentro de la misma, así como el dimensionamiento y posibles entornos de desarrollo, pre y producción.

Período de servicio activo con el presupuesto asignado

El proveedor indicará en la propuesta el coste directo de la contratación de los suministros con el proveedor Cloud y, en base a ello y al presupuesto del proyecto, la duración del servicio con la plataforma activa. Deberá indicar también la forma de contratación y pagos de los suministros en los que aplique (reserva de instancias, pagos anticipados, etc.) y el descuento en presupuesto o el beneficio en términos del período de actividad del servicio que esto supone.

Elaboración de documento de diseño de proceso de ingesta de información

Un punto esencial de la propuesta es establecer las líneas generales de cómo será el proceso de ingesta de la información en la plataforma, así como de la definición de roles que participarán en el mismo. La DGE de la CAM utiliza tanto fuentes de datos que elabora de manera interna como fuentes de información procedentes de terceros (otras consejerías de la Comunidad de Madrid, Instituto Nacional de Estadística, Padrón, ...) por lo que, desde el punto de vista de los datos, el problema de gobernanza tiene como base la variedad de la información (formatos, fuentes, normalización y estandarización) más que otros potenciales problemas como el volumen de los datos. El proveedor indicará en la propuesta tanto los pasos fundamentales del proceso como los roles y herramientas asociados en cada punto. El proveedor debe tener en cuenta que dentro de este proceso también se incluye la propia gestión de datos derivados dentro de la propia DGE de la CAM, esto es, debe cubrir la posibilidad de que cualquier dato generado sea susceptible de ser compartido a través de la plataforma con el resto de la organización.

Herramientas de ingesta de información

El proveedor indicará en la propuesta las herramientas de ingesta de la información desde los distintos posibles orígenes a la capa inicial propuesta en la arquitectura de datos y, precisamente, los componentes de la arquitectura que conecta. Sin carácter limitativo, la solución debe satisfacer, al menos, los siguientes casos de uso:

- Extracción de la información desde orígenes relacionales
- Extracción de la información desde orígenes de gestión manual: por ejemplo, debe contemplarse la posibilidad de un proceso de ingesta de información desde un servidor FTP donde se sube información manualmente.
- Extracción de la información desde APIs de terceros
- Tratamiento de datos en cualquier formato: CSV, TXT, JSON, Excel, Parquet, XML, GeoJSON, KML, SHP, ...

Herramientas para la monitorización y tracking histórico de las ingestas

El proveedor propondrá, dentro de la herramienta anterior o como otro servicio, el mecanismo por el cual se establecerá un tracking histórico de las ingestas que contenga, al menos, la siguiente información:

- Datos subidos a través de los distintos procesos.
- Fechas de ejecución de los procesos.
- Status de las subidas.

- Mecanismo de relanzamiento de los procesos de ingesta sobre datos disponibles.

Herramientas de transformación de información

Puesto que en la arquitectura propuesta se pueden proponer distintas capas de almacenamiento de información (raw, staging, master, etc.) se deberá especificar en la propuesta las herramientas y suministros de transformación de datos entre las distintas capas, así como los aspectos generales diferenciales de la información que se albergue entre ellas, esto es, niveles de transformación y estandarización entre las mismas.

Herramienta para el desarrollo del catálogo de datos interno

Debido a la gran variedad de información que se maneja dentro de la DGE de la CAM, así como la propia información que genera, debe ser gobernada, documentada y accesible a través de un catálogo de datos interno que, al menos, permita realizar las siguientes acciones:

- Listar todas las tablas contenidas en la plataforma
- Etiquetar las tablas contenidas en la plataforma
- Desglosar todos los campos de las tablas
- Documentar tanto la descripción de la tabla como su frecuencia de actualización y los campos que la contienen. Esta documentación puede generarse directamente en la herramienta o como meta-información almacenada en la propia infraestructura que sea consumido por el catálogo de datos.
- Buscar a partir de texto libre y obtener como resultado el listado de tablas y/o campos que encajen con la búsqueda.

Herramienta para la disponibilidad externa de la información

Ya sea para público en general o para compartición con otras administraciones o consejerías de la Comunidad de Madrid, el trabajo realizado puede dar lugar a que la totalidad o parte de los datos incluido en la plataforma sean susceptibles de ser compartidos. Para ello, el proveedor deberá proponer las herramientas que considere adecuadas y señalar en qué elementos de la arquitectura impacta. Pueden ser herramientas de mercado o software libre tipo CKAN.

Un punto obligatorio en toda la propuesta es que, ya sea en este punto de la arquitectura o en otros, como directamente las capas de consumo del repositorio de

información, se tiene que albergar metadata sobre las tablas y datasets generados que sean compatible con el formato Statistical Data and Metadata Exchange (SMDX), un estándar para la compartición de datos estadísticos. Esto facilitará, por un lado, el entendimiento entre los propios datos de la DGE de la CAM y ayudará a la interconexión con otras plataformas de datos estadísticos de administraciones distintas del estado.

Herramienta de Business Intelligence

En la actualidad, la DGE de la CAM posee licencias de Microsoft Power BI para la creación de cuadros de mando que, en la actualidad, los analistas técnicos utilizan directamente contra datos que tienen en ficheros Excel y CSV. De acuerdo al ecosistema y al planteamiento de plataforma de interoperabilidad de datos planteada, el proveedor propondrá la mejor manera de conectar la herramienta de Business Intelligence existente a los datos de la plataforma para su correcta explotación, indicando cómo se llevarán a cabo distintos casos de uso como la creación de dashboards compartidos internamente o dashboards que puedan ser publicados y ofrecidos a través de los distintos portales y disponibles para toda la población en general. En caso de que se considere la opción de otra herramienta de Business Intelligence, el proveedor deberá indicar la motivación y el impacto en los costes que ello puede conllevar.

Ecosistema de proyectos de Data Science y Machine Learning

Dentro de la arquitectura planteada, el proveedor definirá los componentes que permitirán desarrollar sobre los mismos, en el futuro, proyectos de Data Science y Machine Learning sobre los que se instalarán herramientas que suelen ser usadas en este tipo de proyectos:

- RStudio: entorno para el desarrollo de proyectos de análisis de datos basados en el lenguaje de programación R.
- JupyterLab: entorno para el desarrollo de proyectos de análisis de datos basados en el lenguaje de programación Python.

Elementos de gestión del software y elementos de despliegue

Asimismo, dentro de la arquitectura, deben aparecer elementos que faciliten la aplicación de buenas prácticas de desarrollo, que servirán en el futuro tanto para mejorar los procesos de implementación de ingestas de datos como proyectos de Data Science. En particular, debe aparecer:

- Sistema de control de versiones de código.
- Repositorio de imágenes de Docker.
- Entorno para despliegue a partir de imágenes Docker y Kubernetes (Rancher, Portainer, ...)
- Sistema de gestión de tickets para que, en el futuro, las peticiones de información o los reportes de incidencias se canalicen a través de esta plataforma.

Organización y seguimiento del proyecto

Composición del equipo del proyecto

En la propuesta, el proveedor deberá indicar:

- Perfiles profesionales asociados con tareas de gestión del proyecto y elaboración de entregables (figuras tipo Project Manager)
- Perfiles profesionales técnicos asociados con la ejecución de las tareas necesarias para llevar a cabo la implantación de la plataforma.

El proveedor adjuntará un Curriculum Vitae ciego para cada uno de los perfiles que indique en la propuesta. El proveedor se compromete a asegurar que el personal tiene las capacidades técnicas y conocimientos necesarios para la correcta ejecución del proyecto en tiempo y forma. Se valorará la experiencia individual de los miembros en proyectos similares.

Dependencia funcional del personal

La empresa adjudicataria deberá aportar el personal preciso para cumplir las obligaciones derivadas de la ejecución de este contrato. Dicho personal dependerá exclusivamente del adjudicatario, por cuanto este tendrá todos los derechos y deberes inherentes a su calidad de empresario y deberá cumplir con las disposiciones vigentes en materia laboral, de seguridad social y de seguridad e higiene en el trabajo, sin que en ningún caso pueda alegarse derecho alguno por dicho personal en relación con la Administración contratante, ni exigirse a esta responsabilidades de cualquier clase como consecuencia de las obligaciones existentes entre el adjudicatario y sus empleados, aún en el supuesto de que los despidos o medidas que se adopte se basen en el incumplimiento, interpretación o resolución de este contrato.

Metodología de seguimiento del proyecto y comunicación con el personal involucrado de la Dirección General de Estadística

Para asegurar el correcto desarrollo de acuerdo con las condiciones de las Prescripciones Técnicas, este se llevará a cabo bajo el control y supervisión del Subdirector General de Estadística de la Dirección General de Estadística y, a quien el Director de Proyecto designado por el adjudicatario informará del estado del proyecto.

Adicionalmente, el proveedor propondrá una metodología de seguimiento con el personal técnico involucrado en el proyecto y un modelo de comunicación a distintos niveles, con el fin de garantizar la agilidad operativa del proyecto y el reporte de los resultados parciales y totales del mismo.

Recursos materiales del proyecto

El proveedor se compromete a proporcionar al equipo destinado todos los medios materiales necesarios para el correcto desarrollo del proyecto.

Lugar de ejecución del proyecto

El proyecto se podrá ejecutar de manera remota desde las oficinas del proveedor, realizando las reuniones de seguimiento de forma telemática o presencial según las circunstancias y según lo que requiera la DGE de la CAM.

Plazo de ejecución

El plazo de ejecución del proyecto es de 1 mes y se iniciará el 1 de octubre de 2020.

Presupuesto

El presupuesto del proyecto es de 248.582,40 €, impuestos incluidos.

Pago de la prestación

Pago único previa entrega y recepción de conformidad de la totalidad del suministro. El pago se efectuará mediante la presentación de la factura debidamente conformada por el Responsable del contrato.

Secreto estadístico

La empresa adjudicataria queda sujeta al secreto estadístico en lo que respecta a toda la información que pase por sus manos por motivo de esta operación estadística (Ley de Estadística de la Comunidad de Madrid; art. 15 y siguientes), con las obligaciones que se derivan para todo el personal. Todo el personal implicado en los trabajos deberá conocer estas circunstancias y firmar un documento en el que afirme conocer y tener voluntad de cumplir estos extremos. Dichos documentos serán entregados a la Dirección General de Estadística.



DIRECCIÓN GENERAL DE ESTADÍSTICA
VICEPRESIDENCIA, CONSEJERÍA DE DEPORTES, TRANSPARENCIA
Y PORTAVOCÍA DEL GOBIERNO

En Madrid, a fecha de la firma
LA DIRECTORA GENERAL DE ESTADÍSTICA

Firmado digitalmente por: DE LA FUENTE CORRALES MARÍA
Fecha: 2020.08.31 13:34