MEMORIA JUSTIFICATIVA

SUMINISTRO DE UNA LICENCIA DE USO DE LA HERRAMIENTA PENTAHO BUSSINESS ANALYTICS (PBA) PARA LA CONSEJERÍA DE AGRICULTURA, GANADERÍA, PESCA Y DESARROLLO SOSTENIBLE. EXPTE N.º: CONTR 2019/790546

1. INTRODUCCION.-

La Junta de Andalucía, consciente de que la reducción de emisiones es uno de los mayores retos a los que nos enfrentamos, inició hace una década diversas iniciativas en la lucha contra el cambio climático.

Por acuerdo de Consejo de Gobierno de 3 de septiembre de 2002, la Junta de Andalucía aprobó la Estrategia Andaluza ante el Cambio Climático, como aportación de nuestra Comunidad Autónoma a la Estrategia Española de Cambio Climático y Energía Limpia.

Dentro de esa Estrategia, el 5 de junio de 2007 se aprobó mediante acuerdo de Consejo de Gobierno, el Plan Andaluz de Acción por el Clima 2007-2012 (PAAC): Programa de Mitigación, que a través de 140 medidas estructura la política regional de reducción de las emisiones de gases de efecto invernadero durante el periodo de compromiso del Protocolo de Kioto.

Mediante Acuerdo de Consejo de Gobierno de 3 de agosto de 2010 fue aprobado el Programa Andaluz de Adaptación al Cambio Climático y, recientemente, el Programa de Comunicación, por Acuerdo de Consejo de Gobierno de 31 de enero de 2012.

Las emisiones de gases de efecto invernadero de un país quedan dividas en dos grandes grupos, las emisiones de dióxido de carbono de las instalaciones industriales afectadas por el Régimen de Comercio de Derechos de Emisión, y el resto de emisiones, que constituyen el denominado sector difuso, en el que están incluidas las emisiones procedentes de algunas actividades industriales, el sector agrario, el sector residencial, institucional y empresarial, el tratamiento de residuos y el sector del transporte.

Dentro de este sector difuso, los gobiernos locales tienen competencias directas en la mitigación de emisiones de GEI, en sectores como el transporte y la movilidad, el planeamiento urbano, la gestión energética municipal, la gestión de residuos o la sensibilización ciudadana.

Además, un diagnóstico inicial y continuado de las emisiones es esencial para diseñar, aplicar y realizar un seguimiento de políticas de acción municipales contra el cambio climático que contribuya a alcanzar el objetivo nacional.

Por estos motivos, la entonces Consejería de Medio Ambiente de la Junta de Andalucía puso a disposición de los 770 municipios andaluces en el año 2010 la herramienta Huella de Carbono de los municipios andaluces, para el cálculo de emisiones difusas a nivel municipal. Esta herramienta,

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 1 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	

que se inició en formato Excel, ha sido objeto de mejoras sustanciales en el año 2011, pasando a ser una Aplicación web accesible desde el Portal Web de la Consejería de Medio Ambiente, que calcula las emisiones del periodo 2000-2008 para todos los municipios andaluces.

2. OBJETO.-

2.1. Objeto del contrato:

El objeto del presente contrato disponer de una licencia de uso de la herramienta Pentaho Bussiness Analytics (PBA), por un periodo de 4 meses para testar la solución de Big Data Analytics en la Huella de Carbono Municipal.

Dicha herramienta permite afrontar de forma sistemática las siguientes fases dentro del proyecto huella de carbono municipal en el área ambiental de la Consejería:

1. Extracción de datos de todas aquellas fuentes que puedan estar relacionadas con la actividad de la Huella de Carbono de, sean de interés para la Dirección y las distintas áreas de gestión.
2. Industrializar los procesos complejos de datos mediante orquestación automática.
3. Correlación de fuentes diversas para la toma informada de decisiones a través de métricas complejas que no están disponibles en la actualidad.

2.2. Requisitos técnicos:

La solución propuesta deberá incluir una plataforma analítica extremo a extremo que incluya las siguientes funcionalidades integradas, no permitiéndose la combinación de distintas herramientas, que complicarían la correcta implantación y gobernanza de la plataforma:

- Portal analítico al que se conecten usuarios para crear y consumir cuadros de mando, informes auto-servicio y descubrimiento de datos mediante cubos multidimensionales interactivos sobre motor OLAP.
- Modelado de datos auto-servicio y automático, para simplificar los procesos de definición de modelos analíticos.
- Integración, preparación, orquestación de datos y procesos, que permita la correlación de múltiples fuentes de datos de diversas características, así como el procesamiento necesario para presentar visualmente métricas avanzadas de la actividad sanitaria.

Los siguientes son requisitos que la solución propuesta deberá cumplir de forma obligatoria:

- La plataforma deberá estar basada en software Open Source y deberá proporcionarse soporte por parte del fabricante del producto, de forma que se disponga de gestión de incidencias, portal de soporte oficial y contacto telefónico en español.
- Deberá incluir funcionalidades de ETL, preparación de datos, calidad de datos, perfilado de datos y correlación de múltiples fuentes de datos. Dada la naturaleza de los entornos de Big

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 2 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	

Data y los posibles requisitos de integración futuros, la plataforma deberá incluir, al menos, conectores para poder extraer datos y actualizar diversas fuentes.

- Deberá proporcionar un entorno gráfico de trabajo para integración y preparación de datos con tecnología arrastrar y soltar con iconos, de forma que se pueden definir procesos complejos sobre datos de una manera rápida y sencilla, que sean auto-descriptivos y fácilmente revisables y mantenibles sin necesidad de tener un conocimiento profundo del ecosistema. Este entorno gráfico no podrá estar basado en una aplicación web, debido al riesgo que una herramienta de estas características puede suponer al ser accedido a través de un navegador web.
- Deberá incluir de serie y sin coste adicional, múltiples pasos de transformación para conectarse con las fuentes indicadas en el segundo punto de este listado, así como para realizar preparación, enriquecimiento y calidad de datos. Estos pasos deberán encapsular las propiedades necesarias para su ejecución y no requerirán de uso de código, salvo aquellos pasos cuyo objetivo sea específicamente la ejecución de scripts.
- Deberán existir pasos genéricos que permitan desarrollar acciones más particulares y también pasos específicos para simplificar los procesos. En particular, deberán incluirse pasos de transformación con la siguiente funcionalidad:
  - Expresiones regulares.
  - Fuzzy match para correlación de texto libre.
  - Procesamiento en streaming con ventanas por registro y tiempo.
  - Serialización de ficheros.
  - HL7 input y output.
  - Control de Checksums: CRC-32, Adler-32, MD5, SHA-1.
  - Control de flujo mediante espera de registros, condiciones anidadas y comparación con registros anteriores y posteriores en tiempo real.
- Deberá disponer de tablas virtuales cargadas en memoria, de forma que se pueda correlacionar distintas tablas en tiempo de ejecución sin necesidad de cargar los datos en una BBDD intermedia. Esto permitirá una correlación automática, eficiente y rápida de fuentes diversas, como por ejemplo, BBDD, servicios web y ficheros en distintos formatos.
- Deberá incluir pasos de transformación para ejecutar scripts que ya estén desarrollados. Al menos, se deberá poder ejecutar código java, javascript, SQL, Hive SQL, Spark SQL, procedimiento almacenado, Shell, Regex, Spark, Pig, R, Python.
- Deberá proporcionar funcionalidad de orquestación de procesos, de forma que puedan ejecutarse transformaciones con lógica de control del tipo: cuándo deben ejecutarse, qué debe ejecutarse a continuación, qué acción o acciones deben ejecutarse en caso de error, etc. Por tanto, se podrán automatizar los procesos de integración y transformación sin necesidad de una herramienta de orquestación externa, aunque adicionalmente, deberá también ser posible realizar la orquestación desde herramientas tipo Control-M. Por ello, la plataforma deberá disponer de línea de comandos que permite la ejecución de transformaciones.
- Deberá permitir añadir funcionalidad mediante desarrollo de nuevos pasos de transformación, de forma que se puedan cumplir todas las necesidades de integración, aun

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 3 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	

cuando la plataforma no disponga de una funcionalidad incluida de serie.

- Deberá proporcionar orquestación de procesos de data science para análisis predictivo dentro de la propia plataforma, incluyendo el entrenamiento, testing y scoring de modelos predictivos y usando librerías estándar de machine learning como R, Weka o Python Scikit-learn. Estos procesos predictivos podrán formar parte de un flujo completo de orquestación junto con la extracción, preparación, correlación de datos, de forma que se tenga control sobre todas las operaciones de datos del entorno analítico y optimice, así, la productividad de los científicos y analistas de datos.
- Deberá proporcionar gestión de modelos predictivos para una correcta industrialización de Machine Learning, de forma que se pueda garantizar el rendimiento de los modelos con el paso del tiempo. Deberá poder automatizar el entrenamiento, evaluación y puesta en producción de los modelos predictivos mediante un mecanismo campeón/contendiente, asegurando que los modelos se pueden evaluar periódicamente y de forma automática, con distintos algoritmos y distintas plataformas de data science.
- Deberá integrarse de forma nativa con Hadoop y Spark, de forma que los procesos se puedan ejecutar en el cluster de Hadoop/Spark sin generar el código correspondiente. Únicamente se enviarán metadatos en formato xml a YARN, y este ejecutará los .jar desplegados en el cluster, correspondientes a los pasos de transformación utilizados. De esta forma, se asegurará la escalabilidad futura, no siendo necesaria la ejecución de procesos en un servidor dedicado para la integración de datos.
- Deberá abstraer la ejecución de procesos en Spark, de forma que los pasos de transformación puedan ejecutarse indistintamente en el servidor propio de la plataforma o en el cluster de Spark. Adicionalmente, deberá poder decidirse en tiempo de ejecución dónde se ejecutará el proceso. De esta forma, la experiencia de trabajo con Spark será gráfica y no se requerirá desarrollar código, ni pasos específicos para ejecución en Spark y esto simplificará notablemente el paso de los procesos a producción y las actualizaciones a nuevas versiones.
- Deberá permitir la correlación de múltiples fuentes de datos en una misma transformación, incluyendo BBDD relacionales, BBDD analíticas, ficheros en filesystem Windows y Linux, ficheros en HDFS, Hive, MongoDB, Cassandra, servicios web https/S3 y fuentes IoT a través de MQTT.
- Deberá permitir procesamiento micro-batch nativo mediante integración con Kafka y ventanas de hasta 1ms. Para ello, deberá disponer de paso productor de Kafka, así como de un paso consumidor de Kafka, que escuche de forma continua una serie de tópicos y sea capaz de procesar los mensajes con ventanas de tiempo o tamaño. Los datos consumidos desde Kafka podrán ser combinados y correlacionados con otras fuentes de datos.
- Deberá soportar virtualización de datos, de forma que se puedan analizar datos directamente desde la fuente, sin necesidad de cargar los datos en un almacenamiento destino y abstrayendo las características de la fuente al consumidor.
    - Deberán poder publicarse modelos analíticos de forma automática, como un paso más en un proceso de transformación. Esto acelerará los tiempos de preparación de datos y el usuario analítico podrá conectarse directamente y consumir el modelo, sin necesidad de

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 4 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	

crearlo previamente de forma manual.

- Deberá soportar la paralelización de procesos mediante tecnología multithread.
- Deberá permitir Data as a Service, pudiendo ofrecer el resultado de cualquier paso de una transformación a un consumidor externo, a través de una conexión JDBC.
- Deberá incluir visualización analítica en todas las fases, desde la extracción y preparación de datos, hasta el consumo en el portal analítico, de forma que en cualquier paso de una transformación puedan consumirse los datos mediante, al menos, cubos multidimensionales, gráfica de barras, de líneas, de burbujas y geo-mapas. Esta visualización en tiempo de desarrollo de plujos deberá ser nativa, sin necesidad de abrir una herramienta externa.
- Deberá permitir definir plantillas de transformación y preparación de datos que reciban metadatos en tiempo de ejecución, de forma que se puedan ejecutar numerosos procesos sobre fuentes de datos diversas, sin necesidad de definir una transformación específica para cada uno de los procesos. Esta funcionalidad supone un ahorro muy importante de tiempo en el desarrollo de procesos de carga.
- Deberá proporcionar Linaje de datos en formato estándar graphML, de forma que pueda consumirse con cualquier herramienta que soporte este formato.
- Deberá incluir herramienta de debugging en tiempo real, por paso de transformación, por número de registros, por condiciones con valores de registro y operadores.
- Deberá incluir pasos de transformación para realizar Slowly Changing Dimension nativo, de tipo 1 y 2.
- Deberá proporcionar un repositorio de metadatos en el que resida la lógica de negocio y los modelos analíticos. Este repositorio dispondrá de gestión de versiones nativo y adicionalmente, deberá poder integrarse con GIT. Dicho repositorio podrá descansar sobre, al menos, las siguientes BBDD: MySQL, Oracle, PostgreSQL, MS SQL Server.
- Deberá incluir un portal analítico que permita el consumo auto-servicio de cuadros de mando, informes y cubos multidimensionales, gráficas de barras, líneas, geo-mapas, mapas de calor y sunburst. Además, deberá permitir la colaboración mediante compartición de contenidos.
- Deberá permitir la incrustación de elementos visuales analíticos y procesos de integración en aplicaciones existentes, pudiendo tener comunicación bidireccional entre la plataforma analítica y la aplicación anfitriona y pudiendo incluso estar embebido el código completo en la propia aplicación.
- Deberá disponer de elementos visuales analíticos, como cuadros de mando, informes, cubos y gráficos, en modo auto-servicio. No obstante, también deberá permitir el desarrollo de cuadros de mando a medida, según especificaciones muy concretas. Estos cuadros de mando no deberán requerir más de dos semanas para su generación.
    - Deberá poder incorporar librerías visuales externas con el fin de hacer frente a un gran número de requerimientos visuales.

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 5 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	

- Deberá integrarse con LDAP y Directorio Activo para autenticación con la herramienta y control de acceso sobre contenidos. Además, deberá soportar integración con Kerberos para autenticación con Hadoop, así como Apache Ranger y Apache Sentry, para la autorización de acceso sobre los datos en el cluster de Hadoop.
- Deberá disponer de tres niveles de multi-tenancy:
  - Data multi-tenancy: permite aplicar reglas personalizadas de seguridad y negocio para controlar el acceso a los datos.
  - Content multi-tenancy: separa el contenido, como reportes y carpetas, en diferentes tenants.
  - UI multi-tenancy: presentando distintos estilos de la interfaz para diferentes tenants.

3. IMPORTE.-

El importe correspondiente a esta contratación será de 14.999,00 €, lo que comporta un IVA de 3.149,79 € y un importe total de **18.148,79 €** IVA incluido.

4. PARTIDA PRESUPUESTARIA Y DISTRIBUCIÓN DE ANUALIDADES.-

1300010000 G/44B/ 61905/ 00 01 1995000850

Se realizará con la siguiente anualidad:

Año 2019: 18.148,79 euros (21% IVA INCLUIDO)

5. PLAZOS DEL CONTRATO.-

Plazo de ejecución: 1 mes
Plazo de garantía: 4 meses

6. RESPONSABLE DEL CONTRATO.-

María Peña Mate Aranda, Jefa del Servicio de Informática

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 6 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	

7. OTROS DATOS RELATIVOS AL CONTRATO.-

Tipo de contrato: Suministro
Empresa contratista: GESTIÓN DE PROYECTOS DE INFORMÁTICA Y COMUNICACIONES, S.L. (GPIC), NIF: B-91798256 , domicilio Avda. Blas Infante n.º 6, puerta 10 B, 41011 Sevilla.

CPV: 48218000-9: " Paquetes de Software de gestión de licencias ".

LA JEFE DE SERVICIO DE INFORMÁTICA

Fdo.: M.ª de la Peña Mate Aranda

LA DIRECTORA GENERAL,

Fdo.: María López Sanchís

MARIA LOPEZ SANCHIS		20/12/2019	PÁGINA 7 / 7
VERIFICACIÓN	NJyGw05wcKu3ZLb0b1JhVOOxYAjvY6	https://ws050.juntadeandalucia.es/verificarFirma	