Reto de innovación tecnológica: MARCO DE TRABAJO DE REFERENCIA PARA EL DESARROLLO Y EXPLOTACIÓN DE APLICACIONES DEL PLAN DE IMPULSO DE LAS TECNOLOGÍAS DEL LENGUAJE (MTR)

CONTEXTO

El Plan de Impulso de las Tecnologías del Lenguaje prevé en su eje III la creación de plataformas comunes de procesamiento de lenguaje natural y traducción automática para su utilización en el desarrollo de aplicaciones que contribuyan a dotar a los ciudadanos de servicios públicos avanzados (eje IV).

Para ello se necesita un marco de trabajo de referencia (en adelante, MTR) para la puesta en marcha y mantenimiento de los servicios asociados a los diferentes proyectos que compongan el Plan de Impulso de las Tecnologías del Lenguaje.

El MTR debe facilitar la creación de las citadas plataformas y facilitar el acceso a recursos lingüísticos y corpus documentales necesarios para llevar a cabo los proyectos faro del Plan de Impulso de Tecnologías del Lenguaje. Así mismo, el MTR debe facilitar el desarrollo de componentes desde entornos ajenos a la plataforma y la escalabilidad de las aplicaciones construidas con esos componentes.

El MTR debe posibilitar también la instanciación de las aplicaciones en diferentes centros de proceso de datos permitiendo su ejecución en diferentes entornos dentro de la Administración, tanto General como Autonómica. Por esta razón el MTR debe ser independiente de la infraestructura de hardware subyacente. Esta característica facilitará el desarrollo de aplicaciones incluso siendo el entorno de desarrollo no necesariamente escalable ni de altas prestaciones.

El MTR debe proporcionar tanto capacidades de cálculo intensivo y como capacidades de servicio continuado de alta concurrencia. Se prevé que ambas capacidades se den de forma diferenciada en entornos de supercomputación (Red Española de Supercomputación –en adelante RES-) y en el seno de la Administración respectivamente.

El MTR constituirá un entorno de trabajo para que los desarrolladores de aplicaciones y administradores de las mismas puedan desarrollar y probar componentes interoperables, ensamblarlos para crear aplicaciones, probarlas y ejecutarlas dentro de un entorno con orquestación y escalabilidad automática, y con altas capacidades de almacenamiento y procesamiento. Cada desarrollo realizado se deberá ceñir a los requisitos solicitados por el MTR. La plataforma ofrecerá una serie de servicios y recursos comunes que las aplicaciones podrán utilizar. Se requiere que la creación de componentes de escalado elástico sea ágil y eficiente en el uso de los recursos.

Los datos de la Administración deben ser tratados en un entorno aislado y seguro con las garantías marcadas por la legislación (protección de datos, confidencialidad, propiedad

intelectual, …). El acceso a dichos datos puede ser compartido entre distintas aplicaciones. Los productos de su procesamiento también pueden ser compartidos entre aplicaciones.

La independencia del entorno físico, la necesidad de aislamiento o compartición configurable entre aplicaciones y datos, la agilidad y flexibilidad en el crecimiento del servicio, así como la capacidad de portabilidad de las aplicaciones serán características esenciales de este marco.

Este MTR debe proporcionar al Plan de Impulso de las Tecnologías del Lenguaje una plataforma y modelo de desarrollo e implantación de soluciones que faciliten lo siguiente:

- Aprovechamiento de capacidades de computación y almacenamiento de altas prestaciones con una excelente conectividad entre nodos y hacia Internet (capacidades proporcionadas por la Red Española de Supercomputación –en adelante RES-).
- Nodos orientados a la investigación y el desarrollo en el ámbito del procesamiento de lenguaje natural y la traducción automática en el marco del Plan.
- Creación de un modelo de servicio dentro de la Administración en sus distintos ámbitos territoriales. Dichos servicios son competencia de los distintos órganos de la Administración y habrán de implementarse en coordinación con las correspondientes Subdirecciones Generales de Informática y Telecomunicaciones (SGTIC).
- Aceleración de las fases de prototipado y definición del proyecto, así como de la fase de desarrollo de los proyectos faro del Plan. También debe permitir acelerar el desarrollo de las plataformas de procesamiento de lenguaje natural y traducción automática previstas en el Plan.
- Facilitar la participación de la comunidad investigadora y de las PyMEs innovadoras en la creación de componentes de los proyectos y plataformas del Plan.

En el MTR se prevé que trabajen al menos los siguientes perfiles de usuarios:

- Administrador del sistema. Es un usuario interno encargado de que todo el MTR funcione correctamente y de dar servicio a las aplicaciones en él contenidas. Llevará a cabo las tareas de administración (instalación, aumento o disminución de nodos o recursos físicos), monitorización del MTR (revisión de ficheros de log, recepción de alertas) y soporte (resolución y de problemas y soporte al resto de usuarios). Pero no será responsable de la monitorización y soporte sobre las aplicaciones desplegadas sobre el MTR.
- Integrador de aplicaciones. Es un usuario interno con la responsabilidad de gestionar los componentes básicos y los recursos lingüísticos que se depositan en los repositorios comunes. Tendrá conocimiento de la arquitectura del marco de trabajo y sus capacidades y colaborará con el arquitecto de aplicaciones a la hora de crear las recetas que construyen las aplicaciones a partir de los componentes básicos y parametrizar correctamente sus condiciones de escalabilidad.
- Desarrolladores. Son los usuarios externos que desarrollan los componentes específicos dentro del ámbito del Procesamiento del Lenguaje Natural. Estos serán construidos según unas reglas básicas de interoperabilidad proporcionadas por la

plataforma de tal manera que su código fuente pueda dar lugar a componentes completamente funcionales. Disponen de conocimiento muy específico de su ámbito de trabajo, pero no disponen de conocimiento profundo de los procedimientos de escalabilidad de sistemas, por lo que uno de los objetivos del marco de trabajo es permitir/habilitar que sus componentes puedan usarse a escala industrial.

- Arquitectos. Son usuarios externos que tienen conocimiento de la aplicación que quieren construir a partir de los componentes aportados al sistema por los desarrolladores. Su función respecto al sistema será la de construir las recetas que generan los clusters que ejecutan la aplicación final. Esta tarea requiere de un conocimiento intensivo de las capacidades del marco de trabajo, por lo que serán apoyados en ella por los integradores de aplicaciones. También deberán gestionar con ellos la inclusión en el repositorio de componentes de aquellos servicios necesarios para las aplicaciones que no se encuentren disponibles, como pueden ser bases de datos o servidores de aplicaciones.
- Administrador de aplicaciones. Serán usuarios externos que supervisen el correcto funcionamiento de las aplicaciones una vez se encuentran en marcha. Existirá al menos uno por aplicación desplegada y harán las tareas cotidianas de administración de un sistema: supervisión de ficheros de log, detección de problemas en el sistema, etc.
- Usuarios finales: son los beneficiarios finales de las funcionalidades que ofrezcan las aplicaciones. Los habrá de diversos perfiles, a su vez.

El proceso de construcción e integración de una aplicación en el MTR del Plan de Impulso de las Tecnologías del Lenguaje comenzará cuando el Arquitecto de aplicaciones solicite la inclusión de un nuevo proyecto en la plataforma. Este proyecto podría ser uno completamente nuevo, pendiente de desarrollar, o uno ya existente, al que simplemente se le quiere dar forma para que pueda ejecutar dentro de la plataforma y aprovechar así sus capacidades.

En las primeras fases del proyecto, el Arquitecto trabajará conjuntamente con el Integrador de aplicaciones, que es quien realmente conoce cómo se debe construir una aplicación en este entorno. En este trabajo previo se deberán definir los componentes básicos que tendrá la aplicación, los recursos lingüísticos y en general, los paquetes de datos que utilizarán, y los componentes estándar de los que hará uso. Si la plataforma no ofreciera alguno de ellos, el usuario integrador le propondrá alternativas viables, pudiendo aceptar la inclusión de nuevos componentes de no existir una alternativa.

Una vez acordada cómo será la aplicación, los desarrolladores comenzarán a construir los componentes básicos que la compondrán. Hay que tener en cuenta que dependiendo el volumen de la aplicación, en sistemas pequeños el rol de arquitecto y desarrollador podría recaer sobre la misma persona. El equipo de desarrollo generará código fuente que adjuntará al sistema por medio del control de versiones, y a través del sistema de integración continua se construirán los contenedores correspondientes. En esta tarea, el desarrollador podría requerir de la ayuda del integrador de aplicaciones para que le asesore sobre cómo debe estructurar el código para que la generación de los componentes básicos sea correcta.

Una vez que los desarrolladores han incluido en el sistema sus componentes y estos se han construido bien, el siguiente paso sería el de la construcción de la aplicación. Esto será tarea de nuevo en colaboración del arquitecto de la aplicación y el integrador. Uno conoce los pormenores del nuevo sistema y otro como encajar aplicaciones en la plataforma.

El objetivo de esta fase será construir las recetas que generan los clusters de la aplicación a partir de los componentes básicos que han creado los desarrolladores. En este punto se incluirán también los componentes estándar necesarios, los paquetes de datos y recursos lingüísticos que necesitará el sistema en producción. También se definirá el comportamiento que debe tener la aplicación en cuanto a la escalabilidad y elasticidad. Todas estas informaciones se incorporarán en la receta de creación, que una vez añadida a la plataforma, desplegará los clusters para dar servicio.

El sistema, una vez se encuentra correctamente instalado en producción, debe ser administrado para verificar su correcto funcionamiento. Esta será la tarea del Administrador de aplicaciones. Debería al menos haber un usuario administrador por cada aplicación y entre sus atribuciones deben estar la supervisión de ficheros de log y la actuación en caso de caída del sistema. Aunque la aplicación se encuentre funcionando dentro del marco de trabajo, los detalles de cada sistema en marcha solo los conocen los miembros del equipo que desarrolló la aplicación, por lo que son ellos los que deben verificar su correcto funcionamiento, a menos que los problemas sean originados por la propia plataforma de ejecución.

En este caso se requerirá la intervención del Administrador del sistema. Éste será un usuario interno que supervisa que el marco de trabajo se encuentra operativo y dando servicio. Supervisará sus ficheros de log y comprobará el estado de los servidores físicos sobre los que se sustenta la arquitectura desplegada.

RETO TECNOLÓGICO:

El reto tecnológico es desarrollar un MTR que satisfaga simultáneamente los siguientes requisitos:

1. **Interoperabilidad:**

Ha de facilitar el desarrollo independiente de componentes que habrán de ser interoperables para garantizar tanto su ensamblaje en flujos de trabajo como su reutilización.

2. **Despliegue independiente de infraestructura y orquestación eficiente:**

Ha de facilitar el despliegue de las aplicaciones en diferentes infraestructuras. Por tanto, las aplicaciones han de ser independientes del hardware, del sistema operativo e incluso permitir despliegues híbridos con nubes privadas o públicas. Asimismo, ha de garantizar una orquestación eficiente y automática de los componentes de las aplicaciones.

3. **Escalabilidad automática ágil y eficiente:**

Ha de asignar automáticamente los recursos disponibles a las aplicaciones y sus componentes de manera ágil y eficiente, conforme a prioridades, consumos máximos, rendimiento, etc. Esta

capacidad de escalar automáticamente será tanto vertical como horizontal. Adicionalmente será capaz de aprovechar también recursos de GPU.

4. Integración y despliegue continuo (CI/CD):

Para el desarrollo de los componentes, los desarrolladores contarán con un repositorio de código y un entorno de integración continua de tal manera que el producto que aportarán en este caso será el código fuente del componente desarrollado y los scripts que definen la manera de compilarlo y generar sus ejecutables.

Una vez aportado el código, la plataforma de manera automática le pasará una serie de pruebas que validarán en la medida de lo posible su adecuación al entorno. Si el resultado de estas pruebas es el esperado, se generará su componente asociado, y éste se depositará en el registro de contenedores de componentes, quedando disponible para ser incorporado a una aplicación.

5. Doble ámbito de ejecución servicio/computación intensiva:

El MTR ofrecerá dos ámbitos de ejecución diferenciados (con un entorno de producción cada uno) en función de la naturaleza de las aplicaciones que contiene. El ámbito de servicio estará especializado en dar acceso a aplicaciones a las que entran usuarios de forma concurrente y con alta disponibilidad. Por el contrario, el ámbito de computación intensiva se especializará en cálculo intensivo concurrente (por lo que es la solución idónea para llevar a cabo el entrenamiento de modelos de procesamiento de lenguaje natural y de traducción automática). Las aplicaciones podrán combinar ambos ámbitos.

6. Otros requisitos:

El MTR ha de proporcionar trazabilidad y auditoría, contabilidad, aislamiento y compartición configurables de datos y procesos, seguridad y cumplir los requisitos legales que se exigen a las aplicaciones y datos de la Administración.

Estos requisitos se han de concretar en los siguientes casos de uso:

a) Indexación documental:

El MTR ha de indexar los corpus que alimenten las plataformas.

b) Procesamiento de lenguaje natural (PLN):

El MTR ha de implementar flujos de trabajo de PLN basados en UIMA y TextServer.

c) Topic models:

El MTR ha de implementar flujos de trabajo de topic model basados en LDA y algoritmos relacionados.

Y opcionalmente:

d) Machine learning:

El MTR ha de implementar flujos de trabajo de machine learning basados en Theano y Tensorflow.

e) Grafos:

El MTR ha de implementar flujos de trabajo para crear, visualizar y explotar grafos.

INSTRUCCIONES PARA PARTICIPAR:

Siguiendo las normas establecidas en Resolución de 20 de abril de 2017, de la Secretaria de Estado para la Sociedad de la Información y la Agenda Digital por la que se convoca una consulta preliminar al mercado previa al procedimiento de Compra Pública de Innovación en el Marco del Plan de Impulso de las Tecnologías del Lenguaje, para participar en la consulta preliminar al mercado sobre este reto se deberá cumplimentar este formulario **y acompañarlo de la documentación que explique con detalle la solución propuesta**:

1. Envío de propuestas:

Para la presentación de las propuestas, se deberá rellenar este formulario, validarlo y remitirlo por correo electrónico a la siguiente dirección: ConsultaPreliminarMercadoPlanTL@minetad.es

2. Validación de propuestas previa al envío:

El formulario se puede guardar parcialmente cumplimentado. No obstante, antes de enviarlo por correo electrónico deberá comprobar mediante el botón “Validar”, de la página 13 del formulario, que ha rellenado todos los campos obligatorios.

3. Tamaño del correo:

Se deberá acompañar el formulario con la documentación complementaria que explique con detalle la solución propuesta (el tamaño no deberá superar 15 MB al ser enviado por correo electrónico).

4. Identificador de propuestas:

Cada propuesta será identificada mediante la dirección de correo electrónico desde la que se remitió. Deberá coincidir con la que se indique en el apartado 3.3 de la sección A del formulario.

5. Nombre del fichero del formulario:

Se solicita que para el nombre del fichero del formulario se emplee la nomenclatura FormularioCPM_usuario.pdf, donde usuario se deberá sustituir por el nombre de usuario de la cuenta de correo electrónico del remitente (texto a la izquierda de la arroba @). Ejemplo: Si el remitente es usuario1@minetad.es, el fichero del formulario se llamará FormularioCPM_usuario1.pdf

6. Nuevas propuestas:

En caso de que un mismo participante envíe varias propuestas, deberá emplear diferentes cuentas de correo electrónico para diferenciarlas.

7. Actualización de propuestas anteriores:

Se podrán enviar sucesivas versiones de una propuesta, siempre que se mantenga la misma dirección de correo electrónico. Cada propuesta enviada sustituirá completamente a la anterior enviada con dicha dirección de correo. Por ello, la nueva propuesta deberá incluir todo lo que se considere que sigue siendo válido de las anteriores.

8. Resolución de dudas:

Se podrán consultar enviando un correo electrónico a la siguiente dirección: DudasConsultaMercadoPlanTL@minetad.es.