Apelidos, nome Apellidos, nombre: Jonatan Enes Alvarez	DNI:	e-mail:	Teléfono de contacto:
Título: *Diseño e implementación de un entorno web para el despliegue de aplicaciones MapReduce en sistemas HPC.*			**Mención cursada:** Enxeñaría de Computadores

Resumen / Resumo:

En los últimos años el uso de las tecnologías Big Data, entre las que se incluyen las aplicaciones MapReduce, ha sufrido un importante incremento en el ámbito empresarial ante la necesidad de explotar el creciente volumen de datos para obtener un beneficio o para, directa o indirectamente, mejorar la competitividad de la empresa y su eficiencia. Esto se pone en evidencia al ver modelos de negocio como los de Google o Facebook, cuyo actual objetivo es recabar la mayor cantidad de información posible. Pero no solo estas grandes empresas de servicios TIC están interesadas en los datos. Comenzando por las grandes corporaciones, cada vez es más común que muchos negocios tengan interés en almacenar datos para su posterior análisis. De esta forma la información pasa a tener un valor y a ser un recurso más de la empresa que ya no solo es necesario para su funcionamiento interno, sino también para su supervivencia futura y posicionamiento en el mercado.

Pero las tecnologías Big Data, a pesar de tener ya más de una década de historia a sus espaldas, han demostrado no ser fáciles de implantar en una empresa, tanto por la heterogeneidad de sus funciones básicas, como por la complejidad técnica de sus implementaciones. Esto provoca que una empresa interesada en usar Big Data tenga que hacer una inversión inicial muy grande, que a su vez conlleva un riesgo también grande. A menudo este riesgo puede provocar una pérdida de interés y el posterior descarte de un proyecto, o de llevarse a cabo, tener una mala experiencia. Este TFG surgió para explorar dichos problemas y las tecnologías Big Data en conjunto, al considerar que el uso de estas tecnologías va a ser obligatorio y necesario para empresas en un futuro cercano, pero también puede ser necesario que haya sistemas intermedios que faciliten el proceso de iniciación.

Partiendo del muy amplio mundo que es Big Data y de la gran cantidad de tecnologías que lo forman, nos hemos centrado en el uso de Hadoop como framework a desplegar y en MapReduce como aplicación y paradigma para la explotación de los datos. La elección de estas dos tecnologías parte de su popularidad, además de ya haber sido probadas y desplegadas por múltiples empresas en sistemas reales usados en producción. Con dichas tecnologías en mente, nuestro principal objetivo en este proyecto es doble y claramente diferenciado.

Por una parte queremos simplificar en gran medida la ejecución de aplicaciones MapReduce para que así usuarios inexpertos puedan tener una primera experiencia con estas nuevas tecnologías, sin tener que conocer los temas más técnicos. Se considera que este usuario tiene un interés, pero carece de los conocimientos específicos para desplegar, administrar y usar la infraestructura subyacente. Por ello, a este usuario le ofrecemos:

- **Interfaz web gráfica de usuario**: Dicha interfaz se ha diseñado para que sea muy simple y "responsive" para el usuario, de tal forma que siempre esté informado de su actividad pasada, de sus tareas en ejecución y de las opciones que tiene. Esta interfaz web será todo lo que el usuario vea, sin tener que preocuparse de lo que sucede "por debajo". De igual forma, el uso de una interfaz web nos permite que la aplicación se use desde varios sistemas operativos o de forma remota.

Por otro lado, pensando en aquellos administradores que tienen a su cargo la gestión de recursos de la empresa, hemos pensado en cómo hacerles más fácil el despliegue o soporte de un framework Hadoop. A estos administradores les ofrecemos:

- **Soporte para clúster Hadoop nativo**: Desde la interfaz web se podrá usar uno o varios clusters Hadoop ya desplegados. Estos clusters no están restringidos a ninguna versión concreta ni a ningún tipo de infraestructura para poder ser usados, pudiendo variar en el tiempo, desde distintas versiones al uso de una infraestructura "local" o "cloud".

- **Soporte para planificadores de recursos**: De igual forma, se puede usar una variedad de planificadores de recursos* para la ejecución de tareas. Con esta funcionalidad abrimos también la puerta a áreas como la Computación de Altas Prestaciones (HPC, High Performance Computing), típicamente orientada a casos de uso específicos y a un ámbito más científico o de I+D.

Estas dos funcionalidades que le proporcionamos a los administradores les permite hacer un uso de infraestructura ya desplegada sin que además ninguna de las dos requiera realizar modificaciones invasivas sobre el sistema existente, lo que ofrece un bajo riesgo, agiliza el despliegue y facilita una prueba rápida de la tecnología. Por otro lado el usuario no se tendrá que preocupar de qué tipo de clúster usar, viéndolos dentro de la interfaz como recursos expuestos y disponibles.

* Planificadores que ofrezcan una implementación del API DRMAA. E.g., SGE, SLURM, Torque/PBS o HTCondor, entre otros.

Posibles aplicaciones / Posibles aplicacións:

- **Ejecución simplificada de aplicaciones MapReduce:** Mediante el uso de la interfaz web ofrecemos un servicio orientado tanto a PYMES, como a ámbitos académicos y entornos científicos, con el que es posible ejecutar aplicaciones MapReduce de una manera muy sencilla sobre una infraestructura hardware ya desplegada. Los usos del paradigma MapReduce pueden ser muy variados y dependen fuertemente de la naturaleza de los datos y del caso de uso concreto.
- **Proyectos piloto centrados en Hadoop:** Si se hace uso de un planificador de recursos, es posible el despliegue de tareas sobre clusters Hadoop que pueden ser totalmente distintos, permitiendo de esta manera que usuarios avanzados o administradores prueben distintas versiones o configuraciones, sin requerir modificar el planificador en sí o la infraestructura subyacente.

Etapas para su desarrollo futuro / Etapas para o seu desenvolvemento futuro:

El trabajo y conocimiento adquiridos con este TFG están siendo directamente aplicados en un proyecto piloto para el despliegue de un servicio Hadoop [1] en el Centro de Supercomputación de Galicia (CESGA), con fecha prevista de finalización y puesta en producción para Diciembre de 2015. El principal objetivo de este piloto es ofrecer a los usuarios una interfaz amigable para su primera experiencia con las tecnologías Big Data. Las principales funcionalidades y etapas de desarrollo futuro que se han detectado, en las cuales ya estoy trabajando desde el pasado mes de Julio, una vez presentado mi TFG, son:

- **Ampliar el rango de infraestructuras subyacentes usables** por la interfaz, añadiendo a las dos ya soportadas (planificador de recursos y Hadoop nativo), otras como servicios en la nube (OpenNebula) u otras tecnologías de virtualización (Docker). Esto añade nuevos escenarios en los que la aplicación se puede desplegar dentro de los recursos internos de una empresa, como por ejemplo sobre un sistema OpenNebula desplegado en un servicio de IaaS externo a modo de nube corporativa.
- **Ofrecer una parametrización más avanzada** de las tareas. Manteniendo la interfaz simplificada, a algunos usuarios les puede interesar configurar sus tareas para que sean ejecutadas con unos parámetros específicos dentro del clúster Hadoop, de forma acorde con los datos a analizar o al diseño interno de la aplicación MapReduce.
- **Integrar una gestión básica de los datos.** Teniendo en cuenta que algunos usuarios se inician en el uso de estas tecnologías haciendo pruebas sobre conjuntos pequeños de datos, puede ser interesante que la interfaz no solo ofrezca un control sobre tareas sino también una gestión básica de los datos (exportar, visualizar...).

[1] http://www.cesga.es/es/ver_nova/idnoticia/5374

Imágenes representativas: