Databricks
Configuración del entorno

Tanto para esta PEC como para las posteriores usaremos los servicios cloud de databricks[1]. Para ello nos daremos de alta de forma gratuita en su plataforma como usuarios de la versión "Community Edition"[2]. Podéis usar cualquier cuenta de correo electrónico.

La versión "Community Edition"' nos dará acceso a un mini-cluster de 6 Gigas de RAM y nos permitirá usar su servicio de notebooks especialmente adaptados al framework de Spark[3].

Antes de poder trabajar con los notebooks, hemos de instalar una librería de Python que usaremos para comprobar que los ejercicios se realizan de forma correcta. Realizar esto es muy sencillo. Solo hay que seguir los siguientes pasos:

1. Pulsar en el icono de workspace y abrir el menú contextual.
2. Indicar *Pypi package name = spark_mooc_meta*
3. Instalar libreria
4. Hacer click en el checkbox - *attach automatically to all clusters*

Una vez hayamos finalizado el registro e instalado la librería de auto-corrección, pulsaremos en el botón *workspace* ubicado en la barra lateral izquierda, nos desplazaremos hasta nuestro espacio personal y allí crearemos una carpeta para ubicar nuestros archivos. La carpeta puede llamarse como queramos, por ejemplo "Análisis-Big-Data".

Para crear una carpeta, pulsaremos botón derecho en nuestro espacio personal e indicaremos "Create" → "Folder".

Una vez dentro de la nueva carpeta importaremos el notebook de la sesión. El notebook está contenido en un fichero con extensión ".ipynb"' adjunto a las distintas actividades del curso. Para hacer esto usaremos botón derecho "Import" y arrastraremos el fichero .ipynb deseado hasta el espacio destinado a esta funcionalidad.

[1] https://databricks.com
[2] https://databricks.com/try-databricks
[3] https://databricks.com/product/databricks

Posteriormente hemos de asignar un clúster para su ejecución. Esto se realiza pulsando sobre icono “detached” y creando un nuevo clúster. Para esta práctica podemos usar tanto un cluster con Spark 1.6 como Spark 2.0. Por estabilidad es recomendable usar Spark 1.6.2. El servidor creado lo podremos usar en todas las PEC, no es necesario crear un nuevo servidor con cada una de las PEC.

Luego podremos ir ejecutando cada parte de la PEC usando los botones “play” que encontraremos en la parte superior izquierda de cada una de las celdas de código. El “shortcut” para esta funcionalidad es “mayúsculas + intro”. Encontrareis un listado completo de los atajos de teclado pulsando el icono teclado de la parte superior derecha de la pantalla.

Es muy importante recordar que el orden de ejecución es básico y que si dejáis la PEC a medias deberéis ejecutar las celdas anteriores para poder continuar trabajando otro día.