



**PROGRAMA FORMATIVO DE LA ESPECIALIDAD FORMATIVA  
BIG DATA PARA INGENIERÍAS  
IFCT165PO**

**PROGRAMAS DE FORMACIÓN DIRIGIDOS PRIORITARIAMENTE A TRABAJADORES OCUPADOS**

**Noviembre 2018**

**PROGRAMA DE LA ESPECIALIDAD FORMATIVA:  
BIG DATA PARA INGENIERÍAS**

---

**DATOS GENERALES DE LA ESPECIALIDAD FORMATIVA**

**1. Familia Profesional:** INFORMÁTICA Y COMUNICACIONES

**Área Profesional:** SISTEMAS Y TELEMÁTICA

**2. Denominación:** BIG DATA PARA INGENIERÍAS

**3. Código:** **IFCT165PO**

**4. Objetivo General:** Conocer las tecnologías disponibles para realizar estrategias de Big Data para Ingenierías, realizar un desarrollo con Spark y Hadoop y analizar datos con Pig Hive e Impala.

**5. Número de participantes:** Según normativa, el número máximo de participantes en modalidad presencial es de 30.

**6. Duración:**

Horas totales: 80

Modalidad: Indistinta

Distribución de horas:

Presencial:..... 80

Teleformación:..... 80

**7. Requisitos mínimos de espacios, instalaciones y equipamiento:**

7.1 Espacio formativo:

AULA POLIVALENTE:

El aula contará con las instalaciones y equipos de trabajo suficientes para el desarrollo de la acción formativa.

- Superficie: El aula deberá contar con un mínimo de 2m<sup>2</sup> por alumno.
- Iluminación: luz natural y artificial que cumpla los niveles mínimos preceptivos.
- Ventilación: Climatización apropiada.
- Acondicionamiento eléctrico de acuerdo a las Normas Electrotécnicas de Baja Tensión y otras normas de aplicación.
- Aseos y servicios higiénicos sanitarios en número adecuado.
- Condiciones higiénicas, acústicas y de habitabilidad y seguridad, exigidas por la legislación vigente.
- Adaptabilidad: en el caso de que la formación se dirija a personas con discapacidad dispondrá de las adaptaciones y los ajustes razonables para asegurar la participación en condiciones de igualdad.
- PRL: cumple con los requisitos exigidos en materia de prevención de riesgos laborales

Cada espacio estará equipado con mobiliario docente adecuado al número de alumnos, así mismo constará de las instalaciones y equipos de trabajo suficientes para el desarrollo del curso.

7.2 Equipamientos:

Se contará con todos los medios y materiales necesarios para el correcto desarrollo formativo.

- Pizarra.
- Rotafolios.
- Material de aula.
- Medios audiovisuales.
- Mesa y silla para formador/a.
- Mesas y sillas para alumnos/as.
- Hardware y Software necesarios para la impartición de la formación.
- Conexión a Internet.

Se entregará a los participantes los manuales y el material didáctico necesarios para el adecuado desarrollo de la acción formativa

Las instalaciones y equipamientos deberán cumplir con la normativa industrial e higiénico sanitaria correspondiente y responderán a medidas de accesibilidad universal y seguridad de los participantes. En el caso de que la formación se dirija a personas con discapacidad se realizarán las adaptaciones y los ajustes razonables para asegurar su participación en condiciones de igualdad.

7.2.1 En el caso de formación en modalidad teleformación, se requiere el siguiente equipamiento:

### **Requisitos técnicos de la plataforma de teleformación y del contenido virtual de aprendizaje para especialidades formativas no dirigidas a la obtención de certificados de profesionalidad en la modalidad de teleformación**

#### *1. Requisitos técnicos de la plataforma de teleformación*

La plataforma de teleformación que se utilice para impartir acciones formativas no conducentes a la obtención de certificados de profesionalidad deberá reunir los siguientes requisitos técnicos:

- Compatibilidad con los estándares SCORM e IMS.
- Rendimiento, entendido como número de alumnos que soporte la plataforma, velocidad de respuesta del servidor a los usuarios, y tiempo de carga de las páginas Web o de descarga de archivos, que permita:
- Soportar un número de alumnos equivalente al número total de participantes en las acciones formativas que esté impartiendo el centro o entidad de formación, garantizando un hospedaje mínimo igual al total del alumnado de dichas acciones, considerando un número de usuarios concurrentes del 40% de ese alumnado.
- Disponer de la capacidad de transferencia necesaria para que no se produzca efecto retardo en la comunicación audiovisual en tiempo real, debiendo tener el servidor en el que se aloja la plataforma un ancho de banda mínimo de 100Mbps, suficiente en bajada y subida.
- Funcionamiento 24 horas al día, los 7 días de la semana.
- Compatibilidad tecnológica y posibilidades de integración con cualquier infraestructura informática o sistema operativo, base de datos, navegador de Internet de entre los más usuales o servidor web, debiendo ser posible utilizar las funciones de la plataforma con complementos (plug-in) y visualizadores compatibles. Si se requiriese la instalación adicional de algún soporte para funcionalidades avanzadas, la plataforma debe facilitar el acceso al mismo sin coste.
- Integración de herramientas y recursos necesarios para gestionar, administrar, organizar, diseñar, impartir y evaluar acciones formativas a través de Internet, disponiendo, específicamente, de las siguientes:

Herramientas que faciliten la colaboración y la comunicación entre todos los alumnos, tanto de carácter asíncrono (foros, tablones, correo, listas, etc.), como síncrono, (sistema de mensajería, chat, videoconferencia, etc.).

Herramientas de desarrollo, gestión e integración de contenidos.

Herramientas de seguimiento formativo, control del progreso del alumnado y evaluación del aprendizaje.

Herramientas de administración y gestión del alumnado y de la acción formativa.

- Disponer del desarrollo informático a través del cual el Servicio Público de Empleo de la Administración Competente, de manera automática, realice el seguimiento y control de las acciones formativas impartidas, conforme al modelo de datos y protocolo de transmisión establecidos en el anexo II y en la página web de dicho organismo, a fin de auditar la actividad de los centros y entidades de formación y evaluar la calidad de las acciones formativas.

Para poder realizar tal seguimiento, el Servicio Público de Empleo de la Administración Competente, con la periodicidad que determine, se conectará automáticamente con las plataformas de teleformación, por lo que las mismas deberán contar con los desarrollos informáticos que posibiliten tales acciones de seguimiento (protocolo de conexión SOAP).

Sin perjuicio de lo anterior, y de cara al seguimiento puntual de las acciones formativas de certificado de profesionalidad que se impartan, será preceptivo proporcionar al Servicio Público de Empleo de la Administración Competente una dirección (con sus correspondientes credenciales) de acceso a la plataforma, con permiso de administrador, pero sin posibilidad de modificar datos.

- Niveles de accesibilidad e interactividad que como mínimo cumplan las prioridades 1 y 2 de la Norma UNE 139803:2012 o posteriores actualizaciones, según lo estipulado en el Capítulo III del Real Decreto 1494/2007, de 12 de noviembre.

- El servidor de la plataforma de teleformación ha de cumplir con los requisitos establecidos en la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal, por lo que el responsable de dicha plataforma ha de identificar la localización física del servidor y el cumplimiento de lo establecido sobre transferencias internacionales de datos en los artículos 33 y 34 de dicha Ley Orgánica y en el Título VI del Reglamento de desarrollo de la misma, aprobado por Real Decreto 1720/2007, de 21 de diciembre.

- Incluir la imagen institucional del Servicio Público de Empleo de la Administración Competente y de las entidades que él designe, con las pautas de imagen corporativa que se establezcan.

- Disponibilidad de un servicio de atención a usuarios que proporcione soporte técnico y mantenga la infraestructura tecnológica y que, de forma estructurada y centralizada, atienda y resuelva las consultas e incidencias técnicas del alumnado. El servicio, que deberá estar disponible para el alumnado desde el inicio hasta la finalización de la acción formativa, deberá mantener un horario de funcionamiento de mañana y de tarde, tendrá que ser accesible mediante teléfono y mensajería electrónica y no podrá superar un tiempo de demora en la respuesta superior a 2 días laborables.

## 2. Requisitos técnicos del contenido virtual de aprendizaje

Para garantizar la calidad del proceso de aprendizaje del alumnado, el contenido virtual de aprendizaje de las especialidades formativas no dirigidas a la obtención de certificados de profesionalidad mantendrá una estructura y funcionalidad homogénea, cumpliendo los siguientes requisitos:

- Como mínimo, ser los establecidos en el correspondiente programa formativo que conste en el fichero de especialidades formativas previsto en el artículo 20.3 del Real Decreto 395/2007, de 23 de marzo y esté asociado a la especialidad formativa para la que se solicita inscripción.
- Estar referidos tanto a los conocimientos como a las destrezas prácticas y habilidades recogidas en los objetivos de aprendizaje de los citados programas formativos, de manera que en su conjunto permitan conseguir los resultados de aprendizaje previstos.
- Organizarse a través de índices, mapas, tablas de contenido, esquemas, epígrafes o titulares de fácil discriminación y secuenciarse pedagógicamente de tal manera que permitan su comprensión y retención.
- No ser meramente informativos, promoviendo su aplicación práctica a través de actividades de aprendizaje (autoevaluables o valoradas por el tutor-formador) relevantes para la práctica profesional, que sirvan para verificar el progreso del aprendizaje del alumnado, hacer un seguimiento de sus dificultades de aprendizaje y prestarle el apoyo adecuado.
- No ser exclusivamente textuales, incluyendo variados recursos (necesarios y relevantes), tanto estáticos como interactivos (imágenes, gráficos, audio, video, animaciones, enlaces, simulaciones, artículos, foro, chat, etc.). de forma periódica.
- Poder ser ampliados o complementados mediante diferentes recursos adicionales a los que el alumnado pueda acceder y consultar a voluntad.
- Dar lugar a resúmenes o síntesis y a glosarios que identifiquen y definan los términos o vocablos básicos, relevantes o claves para la comprensión de los aprendizajes.
- Evaluar su adquisición durante o a la finalización de la acción formativa a través de actividades de evaluación (ejercicios, preguntas, trabajos, problemas, casos, pruebas, etc.), que permitan medir el rendimiento o desempeño del alumnado.

## 8. Requisitos necesarios para el ejercicio profesional:

(Este epígrafe sólo se cumplimentará si existen requisitos legales para el ejercicio de la profesión)

## 9. Requisitos oficiales de los centros:

(Este epígrafe sólo se cumplimentará si para la impartición de la formación existe algún requisito de homologación / autorización del centro por parte de otra administración competente.)

## 10. CONTENIDOS FORMATIVOS:

### 1. INTRODUCCIÓN

#### 1.1. ¿Qué es Big Data?

#### 1.2. Paradigmas de procesamiento en Big Data

#### 1.3. Las 8 V de Big Data (Volumen, Volatilidad, Variedad, Valor, Velocidad, Variabilidad, Veracidad, Validez).

### 2. BATCH PROCESSING

#### 2.1. MapReduce

##### 2.1.1. Entorno MapReduce

##### 2.1.2. Función Map y función Reduce

##### 2.1.3. Flujo de datos

##### 2.1.4. Características de MapReduce

##### 2.1.5. Uso de MapReduce

##### 2.1.6. Ventajas e inconvenientes de Map Reduce

##### 2.1.7. Ejercicios y ejemplos con MapReduce

#### 2.2. Hadoop

##### 2.2.1. Entorno Hadoop

##### 2.2.2. Almacenamiento: HDFS

##### 2.2.3. Características de HDFS

##### 2.2.3.1. Lectura y escritura de archivos

##### 2.2.3.2. Consideraciones acerca de memoria de NameNode

##### 2.2.3.3. Aspectos generales de seguridad en HDFS

##### 2.2.3.4. Empleo del Namenode Web UI

##### 2.2.3.5. Empleo del File Shell de Hadoop

##### 2.2.3.6. Procesamiento: MapReduce

##### 2.2.3.7. Ventajas e inconvenientes de Hadoop

#### 2.3. Apache Hadoop YARN

##### 2.3.1. Funciones de Framework computacionales

##### 2.3.2. YARN: El gestor de recursos del cluster

##### 2.3.3. Conceptos de Apache Spark

##### 2.3.4. Ejecución de Computational Frameworks en YARN

##### 2.3.5. Exploración de las aplicaciones de YARN Applications a través de la Web UIs y de Shell

#### 2.4. Agregación de los logs de YARN

##### 2.4.1. Configuración de Hadoop y registros de Daemon

- 2.4.2. Localizar configuraciones y aplicar cambios de configuración
- 2.4.3. Gestión de instancias de Role y añadir servicios
- 2.4.4. Configuración del servicio HDFS
- 2.4.5. Configuración de los logs de Hadoop Daemon
- 2.4.6. Configuración del servicio YARN
- 2.5. Obtención de datos en HDFS
  - 2.5.1. Ingestión de datos desde fuentes de recursos externos con Flume
  - 2.5.2. Ingestión de datos desde bases de datos relacionales con Sqoop
  - 2.5.3. REST Interfaces
  - 2.5.4. Buenas prácticas para la importación de datos
- 2.6. Planificación de un cluster Hadoop
  - 2.6.1. Consideraciones generales de planificación
  - 2.6.2. Elección correcta de Hardware
  - 2.6.3. Opciones de Virtualización
  - 2.6.4. Consideraciones de red
  - 2.6.5. Configuración de nodos
- 2.7. Instalación y configuración de Hive, Pig e Impala
- 2.8. Clientes Hadoop incluidos en Hue
  - 2.8.1. ¿Qué es un cliente de Hadoop?
  - 2.8.2. Instalación y configuración de clientes Hadoop
  - 2.8.3. Instalación y configuración de Hue
  - 2.8.4. Autorizaciones y autenticación Hue
- 2.9. Configuración avanzada de un cluster
  - 2.9.1. Parámetros avanzados de configuración
  - 2.9.2. Configuración de puertos Hadoop
  - 2.9.3. Configuración de HDFS para la organización en rack
  - 2.9.4. Configuración de HDFS para obtención de alta disponibilidad
- 2.10. Seguridad Hadoop
  - 2.10.1. ¿Por qué es importante la seguridad en Hadoop?
  - 2.10.2. Conceptos del sistema de seguridad de Hadoop
  - 2.10.3. Qué es Kerberos y cómo funciona
  - 2.10.4. Securitización de un clúster Hadoop Cluster con Kerberos
  - 2.10.5. Otros conceptos de seguridad
- 2.11. Gestión de recursos
  - 2.11.1. Configuración de cgroups con Static Service Pools
  - 2.11.2. El Fair Scheduler
  - 2.11.3. Configuración de Dynamic Resource Pools
  - 2.11.4. Configuraciones de CPU y memoria YARN
  - 2.11.5. Impala Query Scheduling
- 2.12. Mantenimiento de un cluster
  - 2.12.1. Chequeo del estado de HDFS
  - 2.12.2. Copia de datos entre clústers
  - 2.12.3. Añadir y eliminar de nodos en el clúster
  - 2.12.4. Rebalanceo del Cluster
  - 2.12.5. Directorio de Snapshots
  - 2.12.6. Actualización del clúster
- 2.13. Solución de problemas y monitorización de un cluster
  - 2.13.1. Sistema general de monitorización
  - 2.13.2. Monitorización de clústers Hadoop
  - 2.13.3. Solución de problemas habituales en el clúster de Hadoop
  - 2.13.4. Errores habituales en la configuración
- 3. CIENCIA DE DATOS
  - 3.1. Data Science
    - 3.1.1. Que hacen los data scientists, herramientas y procesos que utilizan
    - 3.1.2. Aplicación de lo aprendido en módulo 2: Uso de Hue
  - 3.2. Apache Spark
    - 3.2.1. Cómo trabaja Apache Spark y que capacidades nos ofrece
    - 3.2.2. Que formatos de ficheros populares puede usar Spark para almacenar datos
    - 3.2.3. Que lenguajes de programación puedes utilizar para trabajar con Spark
    - 3.2.4. Cómo empezar a utilizar PySpark y Sparklyr
    - 3.2.5. Cómo comparar PySpark y Sparklyr
  - 3.3. Machine Learning
    - 3.3.1. ¿Qué es machine learning?
    - 3.3.2. Algunos conceptos y términos importantes
    - 3.3.3. Diferentes tipos de algoritmos
    - 3.3.4. Librerías que se utilizan
  - 3.4. Apache Spark MLlib
    - 3.4.1. Que capacidades de machine learning nos proporciona MLlib
    - 3.4.2. Cómo crear, validar y utilizar modelos de machine learning con MLlib
    - 3.4.3. Ejecución de trabajos Apache Spark
    - 3.4.4. Cómo un trabajo de Spark se compone de una secuencia de transformaciones seguida de una acción

- 3.4.5. Cómo Spark utiliza la ejecución lenta
- 3.4.6. Cómo Spark divide los datos entre las particiones
- 3.4.7. Cómo ejecuta Spark operaciones limitadas y grandes
- 3.4.8. Cómo Spark ejecuta un trabajo en tareas y fases
- 4. DESARROLLO PARA SPARK Y HADOOP
  - 4.1. Datasets y Dataframes
  - 4.2. Operaciones en Dataframe
  - 4.3. Trabajar con Dataframes y Schemas
  - 4.4. Crear Dataframes a partir de Data Sources
  - 4.5. Guardar DataFrames en Data Sources
  - 4.6. DataFrame Schemas
  - 4.7. Rapidez y lentitud de ejecución
  - 4.8. Análisis de datos con consultas de DataFrame
    - 4.8.1. Consultar DataFrames con el empleo de expresiones de columna
    - 4.8.2. Agrupación y agregación de consultas
    - 4.8.3. Unión de DataFrames
  - 4.9. RDD
    - 4.9.1. Introducción RDD
    - 4.9.2. RDD Data Sources
    - 4.9.3. Creando y guardando RDDs
    - 4.9.4. Operaciones con RDDs
  - 4.10. Transformación de datos con RDDs
    - 4.10.1. Escritura y paso de funciones de transformación
    - 4.10.2. Ejecuciones de transformación
    - 4.10.3. Conversión entre RDDs y DataFrames
  - 4.11. Agregación de datos con Pair RDDs
    - 4.11.1. Key-Valué Pair RDDs
    - 4.11.2. Map-Reduce
    - 4.11.3. Otras operaciones Pair RDD
  - 4.12. Consulta y vistas de tablas con Spark SQL
    - 4.12.1. Datasets y DataFrames
    - 4.12.2. Creación de Datasets
    - 4.12.3. Ejecución y guardado de Datasets
    - 4.12.4. Operaciones de Dataset
  - 4.13. Creación, configuración y ejecución de aplicaciones Spark
    - 4.13.1. Creación de una aplicación Spark
    - 4.13.2. Compilar y ejecutar la aplicación
    - 4.13.3. Application Deployment Mode
    - 4.13.4. La interfaz Spark Application Web UI
    - 4.13.5. Configuración de las propiedades de la aplicación
  - 4.14. Procesamiento distribuido
    - 4.14.1. Apache Spark en un Clúster
    - 4.14.2. Particiones RDD
    - 4.14.3. Ejemplo: Particionamiento en consultas
    - 4.14.4. Etapas y Tareas
    - 4.14.5. Planificación de tareas de ejecución
  - 4.15. Persistencia de datos distribuidos
    - 4.15.1. Persistencia en Datasets y DataFrames
    - 4.15.2. Persistencia en niveles de almacenamiento
    - 4.15.3. Visualización de RDDs persistentes
  - 4.16. Patrones comunes al procesar datos con Spark
    - 4.16.1. Casos comunes de uso de Spark
    - 4.16.2. Algoritmos de iteración en Apache Spark
    - 4.16.3. Machine Learning
  - 4.17. Spark Streaming: Introducción a DStreams
    - 4.17.1. Vista general de Spark Streaming
    - 4.17.2. DStreams
    - 4.17.3. Desarrollo de aplicaciones en Streaming
  - 4.18. Spark Streaming: procesamiento de múltiples lotes
    - 4.18.1. Operaciones Multi-Batch
    - 4.18.2. Time Slicing
    - 4.18.3. Operaciones de estado
    - 4.18.4. Operaciones Sliding Window
    - 4.18.5. Vista previa: Streaming estructurado
  - 4.19. Apache Spark Streaming: Data Sources
    - 4.19.1. Vista general de Streaming Data Source
    - 4.19.2. Apache Flume y Apache Kafka Data Sources
    - 4.19.3. Ejemplo: uso de un Kafka Direct Data Source
- 5. ANÁLISIS DE DATOS
  - 5.1. Introducción a Pig
    - 5.1.1. ¿Qué es Pig?
    - 5.1.2. Características de Pig

- 5.1.3. Casos de empleo de Pig
- 5.1.4. Interacción con Pig
- 5.2. Análisis de datos básico con Pig
  - 5.2.1. Sintaxis Pig Latin
  - 5.2.2. Carga de datos
  - 5.2.3. Tipos simples de datos
  - 5.2.4. Definición de campos
  - 5.2.5. Datos de salida
  - 5.2.6. Vistas y esquemas
  - 5.2.7. Filtrado y ordenación de datos
  - 5.2.8. Funciones habituales
- 5.3. Procesado de datos complejos con Pig
  - 5.3.1. Formatos de almacenamiento
  - 5.3.2. Tipos de datos complejos y anidados
  - 5.3.3. Agrupaciones
  - 5.3.4. Funciones predefinidas para datos complejos
  - 5.3.5. Iteración de datos agrupados
- 5.4. Operaciones con multiconjuntos de datos con Pig
  - 5.4.1. Técnicas para combinar conjuntos de datos
  - 5.4.2. Unión de conjuntos de datos con Pig
  - 5.4.3. Conjunto de operaciones
  - 5.4.4. División de conjuntos de datos
- 5.5. Troubleshooting y optimización de Pig
  - 5.5.1. Troubleshooting en Pig
  - 5.5.2. Inicio de sesión
  - 5.5.3. Empleo de UI web Hadoop
  - 5.5.4. Muestreo de datos y depuración
  - 5.5.5. Visión general del rendimiento
  - 5.5.6. Comprensión del plan de ejecución
  - 5.5.7. Consejos para mejorar el rendimiento de Jobs en Pig
- 5.6. Introducción a Hive e Impala
  - 5.6.1. ¿Qué es Hive?
  - 5.6.2. ¿Qué es Impala?
  - 5.6.3. ¿Por qué utilizar Hive e Impala?
  - 5.6.4. Schema y almacenamiento de datos
  - 5.6.5. Comparación entre Hive y bases de datos tradicionales
  - 5.6.6. Casos de uso
- 5.7. Consultas con Hive e Impala
  - 5.7.1. Tablas y bases de datos
  - 5.7.2. Sintaxis básica en consultas Hive e Impala
  - 5.7.3. Tipos de datos
  - 5.7.4. Empleo de Hue para ejecutar consultas
  - 5.7.5. Empleo de Beeline (la Shell de Hive)
  - 5.7.6. Empleo de la Shell de Impala
- 5.8. Administración de datos
  - 5.8.1. Almacenamiento de datos
  - 5.8.2. Creación de bases de datos y tablas
  - 5.8.3. Carga de datos
  - 5.8.4. Alteración de bases de datos y tablas
  - 5.8.5. Simplificación de consultas con vistas
  - 5.8.6. Almacenamiento de resultados de consultas
- 5.9. Almacenamiento y datos de rendimiento
  - 5.9.1. Partición de tablas
  - 5.9.2. Carga de datos en tablas particionadas
  - 5.9.3. Cuándo utilizar el particionamiento
  - 5.9.4. Elección de formato de almacenamiento
  - 5.9.5. Gestión de metadatos
  - 5.9.6. Control de acceso a datos
- 5.10. Análisis de datos relacional con Hive e Impala
  - 5.10.1. Unión de conjuntos de datos
  - 5.10.2. Funciones predefinidas habituales
  - 5.10.3. Agregaciones y Windowing
- 5.11. Datos complejos con Hive e Impala
  - 5.11.1. Datos complejos con Hive
  - 5.11.2. Datos complejos con Impala
- 5.12. Análisis de texto con Hive e Impala
  - 5.12.1. Empleo de expresiones regulares
  - 5.12.2. Procesamiento de texto con SerDes en Hive
  - 5.12.3. Análisis de los sentimientos y N•Grams
- 5.13. Optimización Hive
  - 5.13.1. Rendimiento de las consultas
  - 5.13.2. Bucketing

- 5.13.3. Indexación de datos
- 5.13.4. Hive en Spark
- 5.14. Optimización de Impala
  - 5.14.1. Ejecución de consultas
  - 5.14.2. Mejorar el rendimiento de Impala
- 5.15. Extendiendo Hive e Impala
  - 5.15.1. Customizar SerDes y formatos de fichero en Hive
  - 5.15.2. Transformación de datos con Scripts personalizados en Hive
  - 5.15.3. Funciones definidas por el usuario
  - 5.15.4. Consultas parametrizadas
  - 5.15.5. Comparación entre MapReduce, Pig, Hive, Impala, y bases de datos relacionales. ¿Cuál elegir?