Arquitectura Big Data
Introducción de la asignatura

Máster en Big Data. Tecnología y Analítica Avanzada (MBD)

Felipe Cerezo

Escuela Técnica Superior de Ingeniería (ICAI)

Presentación

Felipe Cerezo

Senior Solution Architect
HPE (Hewlett Packard Enterprise)

8+ años en proyectos de Big Data

jfcerezo@icai.comillas.edu

30 horas de clase

Objetivos principales de la asignatura:

- Comprender cuales son los elementos arquitecturales básicos de las herramientas de Big Data (hardware y software)
- Comprender como se usan en entornos empresariales (on-premise, cloud y virtualizados)
- Conceptos de dimensionamiento

Examen (65% nota) preguntas cortas y algún ejercicio de dimensionamiento

Practicas obligatorias (35% nota) 3 ó 4 practicas
Estudio y presentación en clase de una herramienta de Big Data

Temario

1. Distributed systems
    1.1. Distributed processing concepts: Software clusters, Multithread, Multiprocess, High Availability
    1.2. Data replication
    1.3. Big Data Projects. Cycle of life. Professionals profiles.
2. Main Big Data Tools Architecture
    2.1. Paradigma Map-Reduce
    2.2. HDFS
    2.3. YARN and Map Reduce (tool)
    2.4. Kafka
    2.5. Main Non-Hadoop Tools
3. Commercial Big Data platforms & Distributions
    3.1. On-premise
    3.2. Virtualization. Containers.
    3.3. Cloud computing
4. Design of Big Data architectures
    4.1. Methodology of design and sizing of a cluster

Conocimientos previos

Inglés B2

Manejo de unix a nivel de usuario

Manejo de un editor de ficheros en unix (vi / nano)

Conexión y uso de ssh

Conceptos básicos de redes

- Direcciones IP / Subredes
- Puertos TCP
- Protocolos básicos y su uso (ssh, http, https, …)

Conceptos básicos de programación (no es necesario ningún lenguaje específico)

Conceptos básicos de arquitectura de ordenadores

Prácticas

Tenemos un entorno de servidores Linux con las herramientas de Big Data dentro de la universidad (Distribución antigua de Cloudera)

Tendréis usuarios individuales para cada uno de vosotros

De cara a facilitar vuestro trabajo podréis acceder en remoto a través de una VPN

Las conexiones y el trabajo se realizarán a través de ssh, vais a necesitar una cliente ssh en vuestros ordenadores

- MacOS: es un unix y por tanto podéis abrir un terminal y lanzar ssh
- Windows: necesitáis un cliente específico
  - recomendables: MobaXterm, Cygwin, Putty

Para problemas con el cluster: Felipe Cerezo

¿Qué es Big Data?

Procesamiento y almacenamiento de gran cantidad de datos…

… con hardware y herramientas adecuadas …

…para solucionar problemas y generar un beneficio para la compañía

Por ejemplo… red móvil de Orange

Procesamiento y almacenamiento de gran cantidad de datos…

Llamadas de voz y “datos”
2.500 Millones de registros diarios
1.5 Tb de datos planos
300K registros por segundo de procesamiento

… con hardware y herramientas adecuadas …

31 Servidores, total: 2448 cores, 11 Tb RAM, 1.5 Pb disco
Kafka + Storm + HDFS + Hive + Yarn + Elasticsearch

…para solucionar problemas y generar beneficio

Análisis agregados, a nivel de celda, a nivel de provincia, cliente, tarifa o roamers
Análisis evolutivos del uso de la red
Análisis de incidencias:
- un cliente en concreto
- una celda en concreto

Análisis de navegación de usuarios para ver la calidad de la red

Una arquitectura....

Fuentes de datos
- Estructurados
- No estructurados
- En streaming
- Internas / externas
- Periódicas / aperiódicas
- Video, audio, textos

Extracción
- Selección de fuentes
- Filtrado en origen
- Almacenamiento datos intermedios

Procesamiento
- Integración
- Limpieza
- Enriquecimiento,
- Anonimización
- Compresión
- Replicación de datos
- Agregaciones
- Carga en repositorios finales
- Homogeneización
- Filtrado
- Clasificación
- Autocompletado de valores
- Detección de outliers
- Detección de duplicados
- Comprobación restricciones

Análisis
- Modelos de machine learning
- Modelos predictivos
- Data mining
- Análisis geo-espaciales
- Modelos multidimensionales
- Análisis semántico
- Reconocimiento de patrones
- Generación de alertas

Visualización
- Generación de informes
- Publicación/envío de informes
- Cuadros de mandos
- Data discovery
- Consulta libre

Funcionalidades (Big Data Capabilities)

Usuarios finales
- Consumo de los datos
- Automáticos / manuales

Almacenamiento

Repositorios

Procesamiento

Análisis / Búsqueda

Soporte tecnológico (Big Data Framework)

Orquestación
- Control de procesos,
- Secuenciación
- Control de tiempos de respuesta,
- Garantías de ejecución

Monitorización
- Procesos
- Recursos hardware empleados/disponibles
- Alertas
- Calidad de los datos
- Tiempos de respuesta
- detección de fallos hardware y software

Gobierno del dato
- Repositorio de metadatos
- Calidad de la información,
- Seguridad
- Control de accesos,
- Trazabilidad del dato
- Golden records,
- Ciclo de vida del dato

… su implementación