FACULTAD DE INFORMÁTICA

UNIVERSIDAD NACIONAL DE LA PLATA

Infraestructura software para el procesamiento de Big Data	**Profesor Responsable:** Dr. Remo Suppi **Carrera:** Doctorado en Ciencias Informáticas **Créditos:** **Duración:** 70 horas

OBJETIVOS GENERALES

Big Data se ha transformado en una palabra de moda ya que, tanto en centros de R+D+i como en las empresas/industrias/instituciones, hay cada vez más cantidad de datos y es necesario nuevas técnicas y herramientas para procesarlos y obtener información de ellos. La revolución digital/electrónica no solo ha causado un enorme volumen de datos sino también en la variedad (es decir, estructurados, semi-estructurados y no estructurados) y en la velocidad que se generan (millones de dispositivos y aplicaciones que generan datos enormes cada segundo).

Este enorme crecimiento de los datos ha llegado al límite de procesamiento/ almacenamiento de las infraestructuras de gestión de información existentes, que ha obligado a las empresas e instituciones a invertir en más en hardware y actualizaciones de bases de datos. Esta tendencia, que aparentemente soluciona el problema inicial, no es una solución real ya que el conjunto de datos sigue creciendo y por lo cual se entra en un ciclo de más hardware + almacenamiento y así indefinidamente. Además, la infraestructura tradicional no es eficiente debido a los altos costos, las limitaciones de escalabilidad (cuando se trata de petabytes) y la incompatibilidad de los sistemas de bases de datos relacionales con datos no estructurados.

Para abordar la enorme cantidad de datos en la web, Google en 2004, presentó un modelo de programación llamado Map-Reduce (MR) para realizar, en forma paralela, tareas de búsqueda sobre sus clústeres de servidores. Para lograr éste objetivo publicaron sus ideas sobre un sistema de archivos distribuidos (para tener los datos cerca de donde se tuvieran que procesar) en 2003 y luego el algoritmo de procesamiento Map-Reduce (en diciembre de 2004).

Basados en estas ideas, dos investigadores desarrollaron un sistema de archivos y un entorno de procesamiento que finalmente de denominó Hadoop. Este proyecto es actualmente un entorno de código abierto en la Apache Software Foundation y se ha convertido en el estándar de facto para almacenar, procesar y analizar cientos de terabytes o petabytes de datos. Es un proyecto de código abierto y disponible para la comunidad que lo desee utilizar incluso con fines comerciales.

Después del lanzamiento del entorno Hadoop, el Big Data Analytics se transformó en el gran reto tecnológico que permitía a las empresas/instituciones dar el salto estratégico de la retrospectiva a la prospectiva extrayendo valor de los grandes conjuntos de datos.

Si bien Map-Reduce ha sido durante años la metodología de desarrollo de software para el procesamiento distribuido por lotes (*batch*) de Big Data, este tipo de procesamiento no es la respuesta para todas las situaciones computacionales. Con el interés de buscar mejores prestaciones al Map-Reduce, los investigadores comenzaron a elaborar nuevas propuestas y entre las cuales destaca el proyecto Spark. Spark fue diseñado para ser un motor de procesamiento de propósito general para tareas interactivas, por lotes y de flujo donde su procesamiento se hace en memoria y es capaz de aprovechar los mismos recursos de procesamiento distribuidos que utiliza MapReduce bajo Hadoop.

Este curso presenta tanto actividades de formación teórica donde se analizarán los principales conceptos del Big Data, la infraestructura Cloud y el ecosistema Hadoop, como práctica donde el alumno adquirirá habilidades y competencias para instalar y explotar una arquitectura software para el procesamiento del Big Data sobre una infraestructura Cloud. El contenido práctico del curso se desarrollará sobre ejercicios guiados donde el alumno desplegará y trabajará con las máquinas virtuales preparadas a tal fin y aplicará los conceptos y realizará diferentes casos de estudio en el procesamiento del Big Data a través de Hadoop y Spark. El curso finalizará con una evaluación donde los alumnos deberán realizar un proyecto de desarrollo práctico sobre la infraestructura propuesta.

METODOLOGÍA Y MODALIDAD DE EVALUACIÓN

La metodología se basa en un conjunto de clases presenciales/virtuales combinadas con sesiones de prácticas para aplicar los conceptos teóricos y que así el alumno adquiera las competencias y habilidades sobre cada uno de los temas que forman parte del contenido de la asignatura. El trabajo se complementa con un proyecto que deberá desarrollar el alumno para cumplimentar las horas asignadas con soporte tutorizado por el profesor (*on-line*) y seguimiento a través del CV de la UNLP.
El despliegue práctico se realizará sobre una infraestructura virtualizada (parcialmente desplegada) en un cloud privado para mostrar los aspectos funcionales y donde los alumnos deberán ampliar y ejecutar diferentes casos de estudio.

La evaluación se realizará con un test de preguntas al final de las sesiones presenciales/virtuales para evaluar el grado de conocimientos del alumno (20%), el proyecto que deberá entregar el alumno al final de las horas programadas (70%) y la participación y aportaciones de calidad/excelencia durante el desarrollo de la asignatura (10%).

CONTENIDO

M1. Introducción a conceptos de Big Data y Cloud (IaaS). Modelo de programación Map-Reduce.

M2. Virtualización, Gestión y Administración de MV y entornos Linux.

M3. Ecosistema Hadoop. Arquitectura Hadoop. Despliegue de Infraestructura. Casos de estudio.

M4. Infraestructura Spark. Conceptos y estructuras de datos. Despliegue de una infraestructura Spark. Casos de estudio.

M5. Análisis y despliegue de una selección de herramientas complementarias del ecosistema Hadoop:

- **Ambari**: una herramienta basada en web para aprovisionar, administrar y monitorizar clústeres de Apache Hadoop.
- **HBase**: base de datos distribuida y escalable que admite el almacenamiento de datos estructurados para tablas grandes.
- **Pig**: lenguaje de flujo de datos de alto nivel y un marco de ejecución para computación paralela.
- **Submarine**: plataforma de IA unificada que permite ejecutar cargas de trabajo de Machine Learning y Deep Learning en un clúster distribuido.

BIBLIOGRAFIA

1. Big Data. Principles & Best Practices of Scalable Real-Time Data Systems. Nathan Marz & James Warren. Manning Plublications Co.
2. Big Data For Dummies. Judith Hurwitz, Alan Nugent, Dr. Fern Halper and Marcia Kaufman. John Wiley & Sons, Inc.
3. Big Data Fundamentals. Concepts, Drivers & Techniques. Thomas Erl, Wajid Khattak, and Paul Buhler. Prentice Hall.
4. Big Data Made Easy. Michael Frampton. Appres.
5. Big Data. Análisis de grandes volúmenes de datos en organizaciones. Luis Joyanes Aguilar. Alfaomega Grupo Editor.
6. Introducing Data Science. Big Data, ML, & more using Pyhton Tools. Davy Cielen. Arno Meysman, Mohamed Ali. Manning Plublications Co.
7. Handbook of BigDataBench (Version 3.1). A Big Data Benchmark Suite. Chunjie Luo, Wanling Gao, Zhen Jia, Rui Han, Jingwei Li, Xinlong Lin, Lei Wang, Yuqing Zhu, and Jianfeng Zhan.
8. The NIST Definition of Cloud Computing. P. Mell and T. Grance. National Institute of Standrs & Technology. Special Publication 800-145. 2011. http://dx.doi.org/10.6028/NIST.SP.800-145
9. Administración de sistemas GNU/ Linux (2016). Josep Jorba y Remo Suppi. OpenBook (CC). http://openaccess.uoc.edu/webapps/o2/handle/10609/60688
10. Administración avanzada del sistema operativo GNU/Linux (2016). Josep Jorba y Remo Suppi. OpenBook (CC). http://openaccess.uoc.edu/webapps/o2/handle/10609/60686
11. Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale. Tom White.
12. Data Analytics with Hadoop: An Introduction for Data Scientists. Benjamin Bengfort and Jenny Kim
13. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems. 2012. Adam Shook and Donald Miner. Hadoop For Dummies. Dirk deRoos.
14. Spark2nd IBM Limited Editionby Robert D. Schneider and Jeff Karmiol. https://www.ibm.com/downloads/cas/WEB4XBOR (promocionado por IBM).
15. Spark: The Definitive GuideBig Data Processing Made SimpleBill Chambers and Matei Zaharia. O'Reilly.
16. Getting Started with ApacheSparkInception to ProductionJames A. Scott.
17. Learning Apache Spark 2. Muhammad Asif Abbasi. On line: https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785885136
18. Pyspark DataFrame Operations - Basics | Pyspark DataFrames. https://datanoon.com/blog/pyspark_dataframe_operations/
19. Spark SQL, DataFrames and Datasets Guide. https://spark.apache.org/docs/3.1.1/sql-programming-guide.html