Universidad Autónoma de Aguascalientes
Propuesta de curso de tecnologías Cloud Computing y Big Data

Elaboró: Luis Eduardo Bautista Villalpando, PhD.

Objetivo: Introducción, aplicación y administración de tecnologías emergentes en computo distribuido relacionados a Cloud Computing y Big Data como servicios sobre demanda, vitalización y análisis de datos.

1. Introducción a Cloud Computing
   1.1. ¿Qué es Cloud Computing?
   1.2. Servicios en Cloud Computing (SaaS, PaaS, IaaS, NaaS)
   1.3. Cloud Computing y Big Data
   1.4. Introducción a la plataforma Hadoop
   1.5. Sub proyectos Hadoop (Avro, Cassandra, Chukwa, HBase, Hive, Mahout, Pig, ZooKeeper)
   1.6. El ecosistema Hadoop (Distribuciones - Cloudera, HortonWorks, IBM, etc.)
   1.7. Sistema de archivos distribuidos Hadoop (HDFS)
   1.8. Arquitectura del HDFS
   1.9. Mecanismo de replicación del HDFS
2. Framework para el desarrollo de aplicaciones MapReduce
   2.1 Arquitectura de aplicaciones MapReduce (Job tracker and Task tracker)
   2.2 Modelo de programación functional
   2.3 Mappers y Reducers
   2.4 Etapas de ejecución del Framework
   2.5 Particionadores y combinarores
   2.6 Tipos y formatos de datos
3. Sistema de coordinación de aplicaciones distribuidas ZooKeeper
   3.1 ¿Qué es ZooKeeper?
   3.2 Mecanismo de prevención y restablecimiento de fallas en aplicaciones distribuidas
   3.3 Arquitectura de ZooKeeper
   3.4 Modelo de datos
   3.5 Interface de programación de aplicaciones (API)
   3.6 Bloqueo y ACL´s con ZooKeeper
4. HBase, sistema de Base de Datos para Big Data
   4.1 RDBMS y almacenamiento para Big Data
   4.2 Vistas conceptuales y físicas
   4.3 Estructura de tablas (renglones, columnas, familias y celdas)
   4.4 Modelo de datos para Operaciones
   4.5 Esquemas (llaves, versiones, tipos de datos soportados)

4.6 Joins y queries alternativos
4.7 Integración de HBase y MapReduce

Recursos:

- Hadoop, The Definitive Guide, O'Reilly / Yahoo Press, by Tom White
- HBase, The Definitive Guide, O'Reilly, by Lars George
- Data-intensive Text Processing with MapReduce, Morgan & Claypool, by Jimmy Lin and Chris Dyer (http://www.umiacs.umd.edu/~jimmylin/)
- Hadoop Project: http://hadoop.apache.org/
- Cloudera Hadoop Distribution: https://ccp.cloudera.com/display/SUPPORT/Downloads

Papers:

- Jeffrey Dean and Sanjay Ghemawat. (2004) MapReduce: Simplified Data Processing on Large Clusters. Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI 2004), pages 137-150.
- Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. (2003) The Google File System. Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP-03), pages 29-43.
- Jeffrey Dean and Sanjay Ghemawat. (2010) MapReduce: A Flexible Data Processing Tool. Communications of the ACM, 53(1):72-77.
- Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. (2006) Bigtable: A Distributed Storage System for Structured Data. Proceedings of the 7th Symposium on Operating System Design and Implementation (OSDI 2006), pages 205-218.