บริษัท คลัสเตอร์คิท จำกัด
91 ซ.ริมคลองชักพระ ถนนบางขุนนนท์ แขวงบางขุนนนท์ เขตบางกอกน้อย กทม.10700
Tel. 0 2881 3800 Fax. 0 2424 7603
Website: http://www.clusterkit.co.th/

หลักสูตร Hadoop Bootcamp V.2

รายละเอียดหลักสูตร

หลักสูตรนี้มุ่งหมายให้ผู้เรียนเข้าใจกระบวนการทำงานของระบบ Hadoop และสามารถติดตั้งระบบ Hadoop Cluster เพื่อใช้งานรวมถึงเข้าใจเครื่องมือแต่ละตัวและประยุกต์ใช้งานซอฟต์แวร์เหล่านั้นได้ ในเนื้อหาเป็นการลงมือปฏิบัติคอนฟิกเครื่องเซิร์ฟเวอร์คลัสเตอร์ให้ทำงานร่วมกัน และศึกษาส่วนประกอบหลัก ๆ ของ Hadoop ไล่ไปทีละส่วน ตั้งแต่ส่วนของระบบไฟล์แบบกระจายที่เรียกว่า Hadoop Distributed File System (HDFS) การประมวลผลข้อมูลด้วย MapReduce รวมถึงซอฟต์แวร์แวดล้อมที่มาทำงานบนระบบ MapReduce อย่าง Pig และ Hive เพื่อใช้จัดการกับข้อมูลในรูปภาษาสคริปต์ และภาษาในลักษณะ SQL ตามลำดับ นอกจากนั้นยังได้หัดใช้ Sqoop เพื่อเชื่อมต่อกับซอฟต์แวร์ฐานข้อมูล (DBMS) รวมถึงการติดตั้งและใช้งาน Hue, impala และ spark ผู้เรียนจะได้ศึกษาไปทีละขึ้น รวมถึงจะได้เรียนรู้คำสั่งจำเป็นต่อการดูแลระบบ การอ่านและวิเคราะห์ Log File

ระยะเวลา 18 ชั่วโมง (3 วัน)

ความรู้พื้นฐาน

ผู้เข้าอบรมควรมีความสามารถในการใช้งานคำสั่งลีนุกซ์ (Linux) พื้นฐาน และ SQL พื้นฐาน

ซอฟต์แวร์ที่ใช้

- Cloudera Hadoop (CDH6) or Hortonworks Data Platform (HDP3)
- JDK-1.8
- CentOS-7 x86_64
- VirtualBox (ทีมงานคลัสเตอร์ตคิทจะเตรียม VirtualBox Image ที่ติดตั้ง Linux CentOS-7 ไว้ให้)
- OpenLandscape Cloud (ผู้เรียนจะได้ใช้งานคลาวด์จำนวน 6 VMs มีหน่วยความจำขนาด 1x16GB และ 5x4GB ตลอด 3 วัน)

เนื้อหาหลักสูตร

บริษัท คลัสเตอร์คิท จำกัด
91 ซ.ริมคลองชักพระ ถนนบางขุนนนท์ แขวงบางขุนนนท์ เขตบางกอกน้อย กทม.10700
Tel. 0 2881 3800 Fax. 0 2424 7603
Website: http://www.clusterkit.co.th/

วันที่ 1

- แนะนำ Big Data ในภาพรวม
- เข้าใจการทำงานและรู้จักองค์ประกอบของ Hadoop
- แนะนำ Cloudera Hadoop และ Hortonworks Data Platform
- การติดตั้ง JDK
- การปรับแต่งระบบลีนุกซ์เพื่อเตรียมติดตั้ง Hadoop แบบคลัสเตอร์
    - การสร้าง ssh key และวางคีย์เพื่อสร้างสภาพแวดล้อมแบบ Single Sign On
    - การปรับแต่งไฟล์วอลล์เพื่อความปลอดภัย
    - การกำหนดค่าไฟล์ /etc/hosts
    - การปิด selinux
- ติดตั้งและใช้งาน HDFS (ติดตั้งบนสภาพแวดล้อม 3 เครื่อง ประกอบด้วย 1 Name Node และ 2 Data Node)
    - การออกแบบระบบ HDFS
    - รู้จักหลักการทำงานของ HDFS และการใช้งาน HDFS
    - รู้จักกับค่าคอนฟิกกูเรชั่นที่เกี่ยวข้อง
    - การตรวจดูสถานะและใช้งานหน้าเว็บ HDFS
    - การใช้คำสั่ง hadoop การจัดการไฟล์ในระบบ HDFS
    - การตรวจดูสถานะ HDFS ผ่านคำสั่งที่เกี่ยวข้อง เช่น dfsadmin
    - การอ่าน Log File และการวิเคราะห์ปัญหาที่เกิดขึ้น
    - การจัดการบัญชีผู้ใช้งาน
- ติดตั้ง Hadoop ผ่าน Cloudera Manager หรือ Apache Ambari (Hortonwork) ผู้เรียนสามารถเลือกติดตั้งได้ โดยติดตั้งบน Cloud จำนวน 6 เครื่อง
    - ปรับแต่งระบบลีนุกซ์เพื่อเตรียมติดตั้ง Hadoop
    - ติดตั้งฐานข้อมูล MySQL และ MySQL JDBC
    - ติดตั้ง Parallel command เพื่อสั่งคำสั่งพร้อมกันที่เดียวหลายเครื่อง
    - ติดตั้ง Services ต่างๆ ผ่าน Cloudera Manager หรือ Apache Ambari

วันที่ 2

- การทำ High Availability (HA)
    - การทำ High Availability สำหรับ HDFS
    - การทำ High Availability สำหรับ YARN
- การใช้งาน HDFS
    - การใช้คำสั่ง hadoop การจัดการไฟล์ในระบบ HDFS
- การใช้งาน MapReduce2 (Yarn)
    - การรันโปรแกรมคำนวณค่า Pi ผ่าน MapReduce2
    - การคอมไพล์และรันโปรแกรม MapReduce
    - ตัวอย่างโปรแกรม WordCount
    - การ Monitor MapReduce Task
- การใช้งาน Pig
    - การเขียน Pig Script และรัน
- รู้จักกับ Hive เครื่องมือที่จะช่วยให้เราสามารถสั่ง SQL เพื่อทำ MapReduce ได้
    - การใช้งาน Hive ผ่านคำสั่ง SQL
    - การใช้งาน Hive ผ่านคำสั่ง hive และ beeline
    - เทคนิคการนำเข้าข้อมูล Hive
    - การคิวรี่ข้อมูลที่จัดเก็บบน JSON File
    - รู้จักกับรูปแบบการจัดเก็บข้อมูลอื่น ๆ บน Hive
    - กรณีศึกษาตัวอย่างการใช้งานจริง
    - การคอนฟิก Hive ODBC และทดลองใช้งานผ่าน Power BI
    - การเชื่อมต่อ Hive ผ่าน JDBC ด้วยโปรแกรม DBeaver
- รู้จักกับ Sqoop เครื่องมือที่ใช้เชื่อมต่อกับ JDBC เพื่อนำเข้าข้อมูลจากฐานข้อมูล
    - การติดตั้งและใช้งาน Sqoop
    - การนำเข้าข้อมูลจาก MySQL สู่ HDFS และ Hive
    - การนำออกข้อมูลจาก HDFS และ Hive สู่ MySQL
- Flume เครื่องมือในการดึงข้อมูลแบบ streaming

- ติดตั้ง Flume ผ่าน tar package
- ทดลองใช้งาน flume เพื่อดึงข้อมูล log data

วันที่ 3

- ติดตั้งและใช้งาน JDBC, ODBC สำหรับ Hive และ Impala
- รู้จักกับ Spark
  - ทดสอบการใช้งาน Spark ด้วยโปรแกรมหาค่า Pi
  - การใช้งาน Spark ผ่านภาษา python (pyspark)
  - ตัวอย่างการใช้งาน Spark ML ด้วยการรัน K-mean กับชุดข้อมูล Iris
- รู้จักกับ HBase
  - การใช้งานผ่าน command-line “hbase”
  - การเรียกข้อมูลผ่านหน้าเว็บ Hue
- รู้จักกับ Kafka และใช้งาน
- การใช้งาน WebHDFS API
- การปรับแต่งประสิทธิภาพที่สำคัญสำหรับการใช้งานจริง
- การออกแบบระบบที่เหมาะสม และกรณีศึกษา

เนื้อหาในส่วนที่แตกต่างกันระหว่าง Cloudera และ Hortonwork

Cloudera Distribution Hadoop (CDH)	Hortonwork Data Platform (HDP)
• การใช้งาน Hadoop ผ่านหน้าเว็บ Hue ○ รู้จักกับ Hue Web Interface ○ การใช้งาน Hue UI ○ การใช้งาน Hive บน Hue ○ การใช้งาน Sqoop บน Hue ○ การนำเข้าข้อมูล CSV และสร้างตารางผ่าน Hue ○ การใช้งาน MapMaker แสดงพิกัด Lat,	• รู้จักกับ Apache Zeppelin ○ การติดตั้งและใช้งาน Apache Zeppelin UI ○ การใช้งาน Hive บน Zeppelin ○ การใช้งาน Spark บน Zeppelin

Long บนแผนที่ • การใช้งาน Impala – SQL Query enging for Hadoop ○ การใช้งาน Impala บน Hue ○ ความแตกต่างระหว่าง Hive และ Impala ○ การใช้งานและการเชื่อมต่อ ODBC & JDBC • การใช้งาน OOZIE การทำ Workflow ○ หลักการทำงานของ workflow ○ การสร้าง workflow ผ่านหน้าเว็บ hue ○ การตั้งเวลาทำงานของ workflow	

การเตรียมเครื่องก่อนวันอบรม

ผู้เข้าอบรมต้องเตรียมเครื่องโน๊ตบุ๊คของตัวเอง โดยมีหน่วยความจำไม่น้อยกว่า 8GB และมีพื้นที่ว่าง (Disk space) ไม่น้อยกว่า 50GB สำหรับสร้าง VMs โดยในการอบรมจะใช้ซอฟต์แวร์ VirtualBox จำลองเครื่องและเปิดฟังก์ชั่น Virtualization ใน BIOS มาให้เรียบร้อยตาม คู่มือ