

## หลักสูตร Hadoop Ecosystem for Big Data

### รายละเอียดหลักสูตร

หลักสูตรนี้มุ่งหมายให้ผู้เรียนเข้าใจกระบวนการทำงานของระบบ Hadoop และสามารถใช้งานรวมถึงเข้าใจเครื่องมือแต่ละตัวและประยุกต์ใช้งานซอฟต์แวร์เหล่านั้นได้ ในเนื้อหาเป็นการศึกษาส่วนประกอบหลัก ๆ ของ Hadoop ไล่ไปที่ละส่วน ตั้งแต่ส่วนของระบบไฟล์แบบกระจายที่เรียกว่า Hadoop Distributed File System (HDFS) การประมวลผลข้อมูลด้วย MapReduce รวมถึงซอฟต์แวร์แวดล้อม (Hadoop Ecosystem) ที่มาทำงานบนระบบ MapReduce อย่าง Pig และ Hive เพื่อใช้จัดการกับข้อมูลในรูปภาษาสคริปต์ และภาษาในลักษณะ SQL ตามลำดับ นอกจากนี้ยังได้หัดใช้ Sqoop เพื่อเชื่อมต่อกับซอฟต์แวร์ฐานข้อมูล (DBMS) รวมถึงการใช้งาน Hue, OOZIE, Hbase, Phoenix และ Spark และการเชื่อมต่อกับ BI Tools เพื่อเรียกใช้ข้อมูลใน Hadoop โดยผู้เรียนจะได้ศึกษาไปที่ละขั้น

ระยะเวลา 18 ชั่วโมง (3 วัน)

### ความรู้พื้นฐาน

ผู้เข้าอบรมควรมีความสามารถในการใช้งานคำสั่ง SQL พื้นฐาน

### สิ่งที่ผู้เข้าอบรมต้องเตรียม

ผู้เข้าอบรมต้องเตรียมเครื่องโน้ตบุ๊กของตนเองมาในวันอบรม ติดตั้งโปรแกรม SSH

### ซอฟต์แวร์ที่ใช้การอบรม

บริษัทจะเตรียม Hadoop Cluster ที่ติดตั้งด้วยซอฟต์แวร์ Hortonworks และ Hue พร้อมเชื่อมต่อบริบยืนยันตัวตนผ่าน LDAP ไว้ใช้ในการอบรม

## เนื้อหาหลักสูตร

### วันที่ 1

- แนะนำ Big Data ในภาพรวม
- เข้าใจการทำงานและรู้จักองค์ประกอบของ Hadoop
- แนะนำ Hadoop, Hortonworks Data Platform
- การใช้งาน HDFS
  - การใช้คำสั่ง hadoop การจัดการไฟล์ในระบบ HDFS
- การใช้งาน MapReduce2 (Yarn)
  - 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
  - การคิวรีข้อมูลที่จัดเก็บบน JSON File
  - รู้จักกับรูปแบบการจัดเก็บข้อมูลอื่น ๆ บน Hive
  - กรณีศึกษาตัวอย่างการใช้งานจริง
- รู้จักกับ Sqoop เครื่องมือที่ใช้เชื่อมต่อกับ JDBC เพื่อนำเข้าข้อมูลจากฐานข้อมูล
  - การนำเข้าข้อมูลจาก DBMS สู่ HDFS และ Hive
  - การสร้างเก็บรหัสผ่านฐานข้อมูลเพื่อความสะดวกในการตั้งเวลา
- รู้จักกับ Hue Web Interface
  - การใช้งาน Hue UI
  - การแสดงผลข้อมูลแผนภูมิแบบต่าง ๆ
  - การนำเข้าข้อมูลผ่าน Hue เพื่อสร้างเป็น Hive Table

**วันที่ 2**

- รู้จักกับ Oozie
  - การสร้าง Workflow
  - การตั้งเวลาให้ workflow ทำงาน (schedule)
- การใช้งาน ODBC & JDBC ในการเชื่อมต่อ Hive
  - การเชื่อมต่อผ่าน JDBC ด้วยโปรแกรม DBeaver
  - การเรียกใช้งานข้อมูลผ่าน BI Tools (PowerBI Desktop)
- รู้จักกับ Spark
  - การใช้งาน Spark ผ่านภาษา python (pyspark)
  - ตัวอย่างการใช้งาน Spark ML ด้วยการรัน K-mean (Classification)
  - ตัวอย่างการใช้งาน Spark SQL
  - การใช้งาน spark ผ่าน หน้าเว็บ Zeppelin

**วันที่ 3**

- รู้จักกับ HBase และ Phoenix
  - การใช้งานคำสั่งพื้นฐาน HBase Shell
  - การใช้งานคำสั่ง SQL ผ่าน Apache Phoenix
- รู้จักกับ Kafka และใช้งาน
- รู้จักและทดลองใช้งาน WebHDFS API
- การออกแบบระบบ Hadoop
  - การออกแบบระบบที่เหมาะสม
  - สถาปัตยกรรมฮาร์ดแวร์
  - คุณสมบัติและหน้าที่การทำงานของโหนดแต่ละประเภท
  - กรณีศึกษา