สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

การศึกษาเปรียบเทียบระบบการให้คำแนะนำบนฮาดูป

COMPARATIVE STUDY OF RECOMMENDATION SYSTEMS ON HADOOP

โดย

ณัฏฐินันท์ ยนต์ไชย

NATTINAN YONTCHAI

อาจารย์ที่ปรึกษา

ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์

วพ. 243/ก 2557

เลขหมู่.....
เลขทะเบียน 144192
วัน,เดือน,ปี 09 พ.ย. 2559

b.00264250
b.12817748X
i.....

รายงานนี้เป็นส่วนหนึ่งของวิชาการศึกษาอิสระ 2
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2557

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COMPARATIVE STUDY OF RECOMMENDATION SYSTEMS ON HADOOP

NATTINAN YONTCHAI

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS OF THE COURSE

INDEPENDENT STUDY 2

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/2014

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2015

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ใบรับรอง

การศึกษาอิสระ 2 (Independent Study 2)

เรื่อง

การศึกษาเปรียบเทียบระบบการให้คำแนะนำบนฮาดูป

Comparative Study of Recommendation Systems on Hadoop

นางสาวณัฏฐินันท์ ยนต์ไชย

รหัสประจำตัว 56606032

ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด

รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ

การศึกษาวิชา การศึกษาอิสระ 2 หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)

ภาคเรียนที่ 2 ปีการศึกษา 2557

.....อาจารย์ที่ปรึกษา

(ผศ.ดร.ภัทรชัย ลลิตโรจน์วงศ์)

.....กรรมการสอบ

(รศ.ดร. วรพจน์ กรีสุระเดช)

.....กรรมการสอบ

(ดร.กนกวรรณ อัจฉริยะชาญวณิช)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อ	การศึกษาเปรียบเทียบระบบการให้คำแนะนำบนฮาดูป
นักศึกษา	นางสาวณัฏฐินันท์ ยนต์ไชย
รหัสนักศึกษา	56606032
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	เทคโนโลยีระบบสารสนเทศ
ปีการศึกษา	2557
อาจารย์ที่ปรึกษา	ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์

บทคัดย่อ

การศึกษาอิสระนี้มีวัตถุประสงค์เพื่อศึกษาและพัฒนาระบบการให้คำแนะนำบนฮาดูป (Hadoop) ซึ่งฮาดูปเป็นเฟรมเวิร์กโอเพนซอร์ส ที่มีความสามารถในการวิเคราะห์และประมวลผลข้อมูลขนาดใหญ่ เนื่องจากกลุ่มข้อมูลที่ใช้ในการศึกษานี้เป็นข้อมูลขนาดใหญ่ จึงได้นำความสามารถของฮาดูปมาพัฒนาเป็นระบบการให้คำแนะนำ เพื่อให้รองรับและประมวลผลกับข้อมูลขนาดใหญ่ได้

นอกจากนั้น ในการศึกษานี้ยังได้ทำการศึกษาและเปรียบเทียบหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับการพัฒนาระบบการให้คำแนะนำที่จะนำมาใช้ประมวลผลข้อมูลขนาดใหญ่ ระหว่างอัลกอริทึมของการคัดกรองสิ่งของร่วม (Item-based Collaborative Filtering) และการคัดกรองผู้ใช้ร่วม (User-based Collaborative Filtering) ผลการศึกษาพบว่าอัลกอริทึมการคัดกรองสิ่งของร่วมเหมาะสมกับข้อมูลที่มีจำนวนสิ่งของน้อยกว่าจำนวนผู้ใช้ และอัลกอริทึมการคัดกรองผู้ใช้ร่วมเหมาะสมกับข้อมูลที่มีจำนวนผู้ใช้น้อยกว่าจำนวนสิ่งของ แต่โดยธรรมชาติของข้อมูลที่ใช้ประมวลผลร่วมกับระบบการให้คำแนะนำ มักจะเป็นข้อมูลที่มีจำนวนผู้ใช้มากกว่าสิ่งของเสมอ ดังนั้นการคัดกรองสิ่งของร่วมจึงเหมาะที่จะเป็นอัลกอริทึมของระบบการให้คำแนะนำบนฮาดูป โดยหวังว่าการศึกษานี้จะมีประโยชน์ต่อผู้ที่ต้องการประยุกต์ใช้ระบบการให้คำแนะนำกับข้อมูลขนาดใหญ่ได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Comparative Study of Recommendation Systems on Hadoop
Student	Miss. Nattinan Yontchai
Student ID.	56606032
Degree	Master of Science
Program	Information Technology
Major	Information System Technology
Academic Year	2014
Advisor	Asst.Prof.Dr. Pattarachai Lalitrojwong

ABSTRACT

This independent study focuses on Hadoop framework and the Hadoop Distributed File System which uses the MapReduce algorithm to manage the extensively large amounts of data to develop the recommendation systems, algorithm analyzing data for a particular problem to find the items a user is looking for and to produce a predicted likeliness score or a list of top N recommended items for a given user.

In this independent study uses several variants of techniques including user-based and item-based collaborative filtering of recommendation approaches. Finally, the comparison of the results were in consideration consisting of the several dataset of many millions ratings. The main purpose of this study is to be useful to those whom projects or interested readers.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

การศึกษาอิสระครั้งนี้จะประสบความสำเร็จไม่ได้ หากขาดความความอนุเคราะห์และความเมตตาจาก ผศ.ดร.ธนิศา นุ่มนนท์และ รศ.ดร.ธนชาติ นุ่มนนท์ ที่ให้ความช่วยเหลือและถ่ายทอดความรู้ให้ตลอดมาโดยที่ไม่เหน็ดเหนื่อยและย่อท้อกับลูกศิษย์ จึงใคร่อยากขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

ขอขอบพระคุณ ผศ.ดร.ภัทรชัย ลลิตโรจน์วงศ์ สำหรับทุกคำแนะนำ ทั้งเรื่องหลักการวิจัย หลักการนำเสนอและความรู้อีกมากมายที่ไม่สามารถหาได้จากในห้องเรียนและในหนังสือ

ขอขอบคุณครอบครัวที่น่ารักที่คอยสนับสนุนและให้กำลังใจตลอดมา

ณัฏฐินันท์ ยนต์ไชย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป (ต่อ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหาการวิจัย

ในยุคปัจจุบันข้อมูลที่ใช้งานอยู่บนเครือข่ายและอินเทอร์เน็ตมีปริมาณมากขึ้นและเติบโตอย่างรวดเร็ว ทั้งยังมีรูปแบบที่หลากหลาย จนทำให้เกิดคำนิยามของข้อมูลเหล่านี้ว่า "ข้อมูลขนาดใหญ่" (Big Data) ข้อมูลเหล่านี้นอกจากมีปริมาณมหาศาลแล้ว ยังมีรูปแบบที่หลากหลายชนิด ทั้งที่มีโครงสร้าง (Structured) กึ่งโครงสร้าง (Semi-Structured) และที่ไม่มีโครงสร้าง (Unstructured) ซึ่งไม่สามารถนำมาประมวลผลกับฐานข้อมูลในปัจจุบันได้ และนอกจากจะต้องใช้พื้นที่จำนวนมากในการเก็บข้อมูลแล้ว ข้อมูลเหล่านี้ก็ยังไม่สามารถนำมาประยุกต์ใช้เพื่อหาประโยชน์ได้มากนัก

ดังนั้น การศึกษาอิสระฉบับนี้จึงศึกษาวิธีการนำข้อมูลขนาดใหญ่มาใช้ประโยชน์ในเชิงธุรกิจ และเล็งเห็นว่าในยุคสมัยนี้ยังไม่นิยมนำข้อมูลขนาดใหญ่มาประมวลผลกับระบบการให้คำแนะนำ (Recommendation Systems) เนื่องจากระบบนี้ไม่มีความสามารถในการรองรับการขยายตัวของข้อมูล (Scalability) [1][2] ทั้งนี้ ระบบการให้คำแนะนำ คือ การแนะนำรายการสินค้าให้กับผู้ใช้โดยพยากรณ์หารายการสิ่งของที่ผู้ใช้น่าจะชื่นชอบมานำเสนอ ซึ่งดูจากประวัติความชื่นชอบสินค้าของผู้ใช้เอง ดังนั้น ถ้าข้อมูลที่ใช้ศึกษาเพื่อพยากรณ์นั้น มีปริมาณมากเท่าไร การพยากรณ์ก็จะยิ่งแม่นยำมากขึ้นเท่านั้น ในการศึกษานี้จึงได้มีวัตถุประสงค์ที่จะนำข้อมูลขนาดใหญ่มาประยุกต์ใช้กับระบบการให้คำแนะนำ แต่การที่จะพัฒนาระบบคำแนะนำให้สามารถประมวลผลกับข้อมูลขนาดใหญ่ได้นั้น จำเป็นต้องใช้ฐานข้อมูลที่มีความสามารถในการประมวลผลกับข้อมูลขนาดใหญ่ได้เป็นอย่างดี ในการศึกษานี้จึงได้พัฒนาระบบการให้คำแนะนำโดยใช้เฟรมเวิร์กโอเพนซอร์สอย่างฮาดูป (Hadoop) ที่มีระบบแฟ้มกระจายแบบฮาดูป (Hadoop Distributed File Systems) และกระบวนการแมปรีดิวซ์ (MapReduce) ที่ช่วยในการวิเคราะห์ข้อมูลขนาดใหญ่ได้ ซึ่งระบบการให้คำแนะนำนี้จะทำงานโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) และแบ่งออกเป็นสองอัลกอริทึมย่อย คือ การคัดกรองผู้ใช้ร่วม (User-based Collaborative Filtering) และการคัดกรองสิ่งของร่วม (Item-based Collaborative Filtering) ในการศึกษานี้จึงทำการทดลองเพื่อเปรียบเทียบและพิสูจน์ว่าอัลกอริทึมใดมีความสามารถและมีประสิทธิภาพที่ดีกว่ากันและเหมาะสมในการนำมาประยุกต์ใช้กับระบบการให้คำแนะนำบนฮาดูปที่สามารถประมวลผลกับข้อมูลขนาดใหญ่ได้ โดยมีกลุ่มข้อมูลขนาดใหญ่ที่ใช้ในการทดลองทั้งหมด 3 ชุดข้อมูลมาจากเว็บไซต์มูฟวี่เลนส์ [3] ทริปแอดไวเซอร์ [4] และเยลป์ [5]

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.2 วัตถุประสงค์ของการวิจัย

ในการศึกษาอิสระนี้มีวัตถุประสงค์เพื่อศึกษาและทดลองหาอัลกอริทึมที่เหมาะสมกับระบบการให้คำแนะนำบนฮาดูป โดยเปรียบเทียบระหว่างการคัดกรองสิ่งของร่วมและการคัดกรองผู้ใช้ร่วมโดยที่มีกลุ่มข้อมูลที่ใช้ในการทดลองเป็นข้อมูลขนาดใหญ่และระบบที่พัฒนาขึ้นมานั้นจะถูกพัฒนาโดยฮาดูปเฟรมเวิร์คที่มีความสามารถในการประมวลผลกับข้อมูลขนาดใหญ่

1.3 สมมติฐานการวิจัย

ระบบการให้คำแนะนำที่พัฒนาขึ้นโดยฮาดูปจะสามารถรองรับการทำงานกับข้อมูลขนาดใหญ่ได้ที่มีปริมาณมากได้ โดยเทคนิคการคัดกรองสิ่งของร่วมจะประมวลผลได้เร็วกว่าการคัดกรองแบบผู้ใช้ร่วม เนื่องจากจำนวนสิ่งของมีน้อยกว่าจำนวนของผู้ใช้ จึงทำให้ใช้เวลาน้อยกว่าในการหาผลลัพธ์ ในเรื่องของความแม่นยำในการประมวลผลนั้น เทคนิคการคัดกรองสิ่งของร่วมจะมีผลลัพธ์ที่ดีกว่าการคัดกรองแบบผู้ใช้ร่วม เนื่องจากค่าความผันผวนของผู้ใช้มีมากกว่าค่าความผันผวนของสิ่งของ

1.4 ขอบเขตของการวิจัย

1.4.1 กลุ่มตัวอย่างข้อมูลขนาดใหญ่ที่ใช้ในการวิจัย

1.4.1.1 ข้อมูลการให้คะแนนความนิยมของภาพยนตร์จากเว็บไซต์มูฟวี่เลนส์ แบ่งออกเป็น 3 กลุ่มข้อมูล คือ เล็ก กลาง ใหญ่ ดังนี้

I. ข้อมูลขนาดเล็ก ประกอบไปด้วยภาพยนตร์ 700 เรื่องและผู้ใช้จำนวน 1,000 ราย
II. ข้อมูลขนาดกลาง ประกอบไปด้วยภาพยนตร์ 10,000 เรื่องและผู้ใช้จำนวน 72,000 ราย
III. ข้อมูลขนาดใหญ่ ประกอบไปด้วยภาพยนตร์ 27,000 เรื่องและผู้ใช้จำนวน 230,000 ราย

1.4.1.2 ข้อมูลการให้คะแนนความนิยมของโรงแรมจากเว็บไซต์ทริปแอดไวเซอร์ แบ่งออกเป็น 3 กลุ่มข้อมูล คือ เล็ก กลาง ใหญ่ ดังนี้

I. ข้อมูลขนาดเล็ก ประกอบไปด้วยโรงแรม 2,500 แห่งและผู้ใช้จำนวน 1,000 ราย
II. ข้อมูลขนาดกลาง ประกอบไปด้วยโรงแรม 30,000 แห่งและผู้ใช้จำนวน 216,000 ราย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

III. ข้อมูลขนาดใหญ่ ประกอบไปด้วยโรงแรม 81,000 แห่งและผู้ใช้จำนวน 690,000 ราย

1.4.1.3 ข้อมูลการให้คะแนนความนิยมของร้านอาหารจากเว็บไซต์เยลป์ แบ่งออกเป็น 3 กลุ่มข้อมูล คือ เล็ก กลาง ใหญ่ ดังนี้

I. ข้อมูลขนาดเล็ก ประกอบไปด้วยร้านอาหาร 2,800 ร้านและผู้ใช้จำนวน 1,000 ราย

II. ข้อมูลขนาดกลาง ประกอบไปด้วยร้านอาหาร 40,000 ร้านและผู้ใช้จำนวน 288,000 ราย

III. ข้อมูลขนาดใหญ่ ประกอบไปด้วยร้านอาหาร 108,000 ร้านและผู้ใช้จำนวน 920,000 ราย

1.4.2 เครื่องมือที่ใช้ในการศึกษา

1.4.2.1 ระบบการให้คำแนะนำพัฒนาโดยใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมโดยแบ่งออกเป็น 2 อัลกอริทึม คือ การคัดกรองสิ่งของร่วม และการคัดกรองผู้ใช้ร่วม

1.4.2.2 พัฒนาโดยประยุกต์ใช้ไลบรารีของฮาดูป เช่น แมปรีดิวซ์ เพื่อให้สามารถประมวลผลร่วมกับข้อมูลขนาดใหญ่ได้

1.4.2.3 ระบบการให้คำแนะนำบนฮาดูปถูกพัฒนาขึ้นบนระบบคลาวด์ของเอมาซอนเว็บเซอร์วิส

1.4.2.4 ระบบปฏิบัติการที่ใช้ คือ ลินุกซ์ อูบันตู เซิร์ฟเวอร์ 14.04 แพลตฟอร์ม 64 บิต (Ubuntu Server 14.04 64 bit)

1.4.2.5 ภาษาที่ใช้ในการพัฒนา คือ จาวา (Java) และไพธอน (Python)

1.5 แนวทางการดำเนินงาน

1.5.1 ศึกษาความเป็นไปได้

ในการศึกษาอิสระนี้ได้นำข้อมูลขนาดใหญ่มาประมวลผลร่วมกับระบบการให้คำแนะนำที่มีข้อจำกัดในเรื่องการรองรับการขยายตัว (Scalability) ดังนั้น เพื่อศึกษาความเป็นไปได้จึงต้องทดสอบการสร้างระบบการให้คำแนะนำขึ้นโดยประยุกต์ใช้ฮาดูปเฟรมเวิร์กมาพัฒนาและทดสอบประมวลผลข้อมูลขนาดใหญ่ เริ่มจากข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดใหญ่ เมื่อผลลัพธ์เป็นที่น่าพอใจแล้วจึงเริ่มการทดลองเพื่อดำเนินตามวัตถุประสงค์ที่ตั้งไว้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.5.2 วางแผน พัฒนาและทดสอบผลลัพธ์

1) วางแผนการทำงาน กำหนดตัวแปร ขอบเขตการทำงานและระยะเวลาที่ใช้

2) ศึกษาการทำงานของฮาดูป หลักการทำงานและความต้องการของระบบ พร้อมทั้งศึกษาทฤษฎี องค์ประกอบต่าง ๆ และกระบวนการพัฒนาระบบการให้คำแนะนำ ทั้งเทคนิคการคัดกรองสิ่งของร่วมและการคัดกรองผู้ใช้ร่วม

3) ติดตั้งระบบปฏิบัติการลินุกซ์ อูบันตู เซิร์ฟเวอร์ บนเอมาซอนเว็บเซอร์วิสอีซีทู พร้อมทั้งติดตั้งฮาดูปและปรับตั้งพารามิเตอร์ต่าง ๆ ให้พร้อมกับการทำงาน

4) พัฒนาระบบการให้คำแนะนำโดยแบ่งสร้างเป็น 2 ระบบ คือ ระบบที่ใช้อัลกอริทึมการคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม หลังจากนั้น ทดสอบประมวลผลข้อมูลขนาดใหญ่ร่วมกับระบบการให้คำแนะนำทั้ง 2 ระบบที่สร้างขึ้น เพื่อศึกษาความเป็นไปได้

5) ดาวน์โหลดข้อมูลการให้คะแนนความนิยมเกี่ยวกับภาพยนตร์มาจากเว็บไซต์มูฟวี่เลนส์และเตรียมข้อมูลเหล่านั้นมาทำให้อยู่ในรูปแบบของแบบจำลองข้อมูลเพื่อนำไปเป็นข้อมูลนำเข้าของระบบ เมื่อข้อมูลผ่านเข้าไปประมวลผลในระบบการให้คำแนะนำแล้ว ให้บันทึกเวลาที่ใช้ในการประมวลและค่าการพยากรณ์ภาพยนตร์ที่น่าสนใจทั้ง 10 อันดับของผู้ใช้แต่ละรายเอาไว้

6) ดาวน์โหลดข้อมูลการให้คะแนนความนิยมเกี่ยวกับภาพยนตร์มาจากเว็บไซต์ทริปแอดไวเซอร์และเตรียมข้อมูลเหล่านั้นมาทำให้อยู่ในรูปแบบของแบบจำลองข้อมูลเพื่อนำไปเป็นข้อมูลนำเข้าของระบบ เมื่อข้อมูลผ่านเข้าไปประมวลผลในระบบการให้คำแนะนำแล้ว ให้บันทึกเวลาที่ใช้ในการประมวลและค่าการพยากรณ์โรงแรมที่น่าสนใจทั้ง 10 อันดับของผู้ใช้แต่ละรายเอาไว้

7) ดาวน์โหลดข้อมูลการให้คะแนนความนิยมเกี่ยวกับโรมแรมมาจากเว็บไซต์เยลป์และเตรียมข้อมูลเหล่านั้นมาทำให้อยู่ในรูปแบบของแบบจำลองข้อมูล เพื่อนำไปเป็นข้อมูลนำเข้าของระบบ เมื่อข้อมูลผ่านเข้าไปประมวลผลในระบบการให้คำแนะนำแล้ว ให้บันทึกเวลาที่ใช้ในการประมวลและค่าการพยากรณ์ร้านอาหารที่น่าสนใจทั้ง 10 อันดับของผู้ใช้แต่ละรายเอาไว้

8) นำข้อมูลการให้คะแนนความนิยมเกี่ยวกับภาพยนตร์มาอีกครั้ง ในรอบนี้ให้แยกผู้ใช้ออกมา 100 รายและดึงค่าคะแนนความนิยมที่ผู้ใช้แต่ละรายได้ให้คะแนนภาพยนตร์ต่าง ๆ ไว้ออกมาเป็นอัตราส่วน 10%, 15%, 20%, 25% และ 30% ตามลำดับ จากนั้นให้ส่งข้อมูลเหล่านี้เข้าไปประมวลผลอีกเป็นจำนวนชุดละครั้ง รวมเป็น 15 ครั้ง เนื่องจากมีข้อมูลทั้งหมด 3 ชุด คือ ข้อมูลขนาดเล็ก กลาง ใหญ่ เมื่อผลลัพธ์การประมวลผลออกมาแล้ว ให้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บันทึกผลลัพธ์นั้นเพื่อนำไปเปรียบเทียบกับผลลัพธ์จากข้อ 7 ที่บันทึกไว้ เพื่อดูอัตราส่วนความถูกต้องของการพยากรณ์ภาพยนตร์

9) นำข้อมูลการให้คะแนนความนิยมเกี่ยวกับโรงแรมมาอีกครั้ง ในรอบนี้ให้แยกผู้ใช้ออกมา 100 รายและดึงค่าคะแนนความนิยมที่ผู้ใช้แต่ละรายได้ให้คะแนนโรงแรมแต่ละแห่งไว้ออกมาเป็นอัตราส่วน 10%, 15%, 20%, 25% และ 30% ตามลำดับ จากนั้นให้ส่งข้อมูลเหล่านี้เข้าไปประมวลผลอีกเป็นจำนวนชุดละครั้ง รวมเป็น 15 ครั้ง เนื่องจากมีข้อมูลทั้งหมด 3 ชุด คือ ข้อมูลขนาดเล็ก กลาง ใหญ่ เมื่อผลลัพธ์การประมวลผลออกมาแล้ว ให้บันทึกผลลัพธ์นั้นเพื่อนำไปเปรียบเทียบกับผลลัพธ์จากข้อ 8 ที่บันทึกไว้ เพื่อดูอัตราส่วนความถูกต้องของการพยากรณ์โรงแรมที่น่าสนใจสำหรับผู้ใช้แต่ละราย

10) นำข้อมูลการให้คะแนนความนิยมเกี่ยวกับร้านอาหารมาอีกครั้ง ในรอบนี้ให้แยกผู้ใช้ออกมา 100 รายและดึงค่าคะแนนความนิยมที่ผู้ใช้แต่ละรายได้ให้คะแนนร้านอาหารแต่ละร้านไว้ออกมาเป็นอัตราส่วน 10%, 15%, 20%, 25% และ 30% ตามลำดับ จากนั้นให้ส่งข้อมูลเหล่านี้เข้าไปประมวลผลอีกเป็นจำนวนชุดละครั้ง รวมเป็น 15 ครั้ง เนื่องจากมีข้อมูลทั้งหมด 3 ชุด คือ ข้อมูลขนาดเล็ก กลาง ใหญ่ เมื่อผลลัพธ์การประมวลผลออกมาแล้ว ให้บันทึกผลลัพธ์นั้นเพื่อนำไปเปรียบเทียบกับผลลัพธ์จากข้อ 9 ที่บันทึกไว้ เพื่อดูอัตราส่วนความถูกต้องของการพยากรณ์ร้านอาหาร

11) นำผลลัพธ์ที่ได้มาบันทึกและวิเคราะห์ผลลัพธ์ พร้อมทั้งสรุปผลการทดลองและเปรียบเทียบ

1.6 ประโยชน์ที่ได้รับจากงานวิจัย

การศึกษาอิสระนี้ได้ทดลองพัฒนาระบบการให้คำแนะนำบนฮาดูปที่สามารถประมวลผลกับข้อมูลขนาดใหญ่ได้ ซึ่งข้ามข้อจำกัดของระบบการให้คำแนะนำในอดีตที่ไม่สามารถรองรับข้อมูลที่มีปริมาณมากได้ ดังนั้น ในงานวิจัยนี้หวังว่าจะมีประโยชน์ไม่มากก็น้อยสำหรับผู้ที่สนใจนำข้อมูลขนาดใหญ่มาประมวลผลบนระบบการให้คำแนะนำบนฮาดูป การยิ่งมีข้อมูลที่ใช้ประมวลผลในระบบมาก ค่าของการพยากรณ์ก็จะยิ่งแม่นยำมากขึ้น ซึ่งเปรียบได้กับหลักการตามธรรมชาติ เช่น เด็กนักเรียนยิ่งอ่านหนังสือมากเท่าใด คะแนนสอบก็จะยิ่งดีมากขึ้นเท่านั้น เช่นเดียวกันกับระบบการให้คำแนะนำนี้ ยิ่งข้อมูลที่ถูกป้อนให้ระบบได้เรียนรู้มากเท่าใด ค่าของการพยากรณ์ก็จะดีขึ้นเท่านั้น ดังนั้น การศึกษาและวิจัยนี้จึงอาจจะมีประโยชน์ต่อผู้ที่สนใจไม่มากก็น้อย

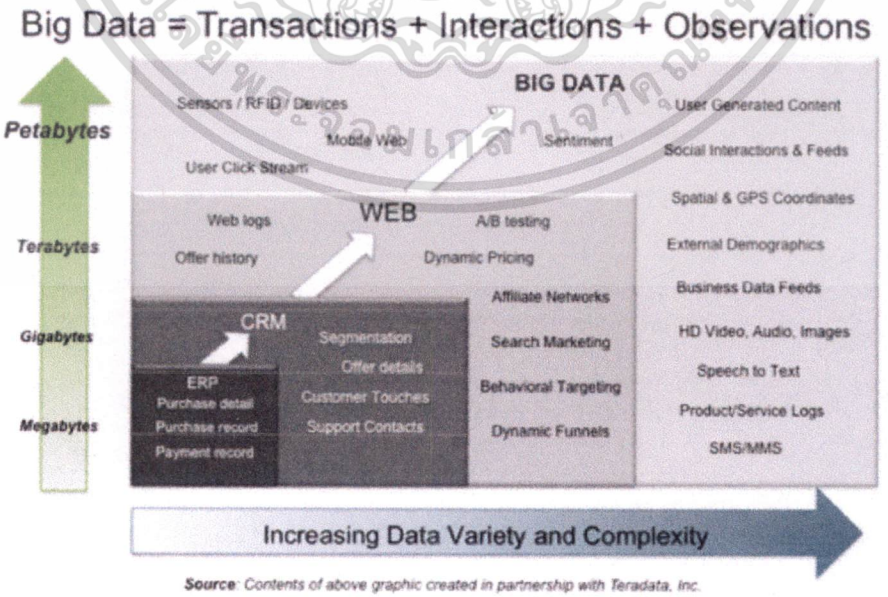
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องในการศึกษาอิสระฉบับนี้ ซึ่งประกอบไปด้วยเรื่องราวของข้อมูลขนาดใหญ่ (Big Data) ระบบคลาวด์ (Cloud Systems) ฮาดูป (Hadoop) ระบบการให้คำแนะนำ (Recommendation Systems) และงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดดังต่อไปนี้

2.1 ข้อมูลขนาดใหญ่

ข้อมูลขนาดใหญ่ หมายถึง ข้อมูลที่มีปริมาณมาก เพิ่มขึ้นอย่างรวดเร็วและมีหลากหลายรูปแบบ อาจจะประกอบไปด้วยข้อมูลที่มีโครงสร้าง (Structured data) ข้อมูลไร้โครงสร้าง (Unstructured data) และข้อมูลกึ่งโครงสร้าง (Semi-Structured data) สาเหตุที่ทำให้เกิดข้อมูลขนาดใหญ่ขึ้นมานั้น เกิดจากในปัจจุบันเทคโนโลยีมีความก้าวหน้าทำให้ทุกคนสามารถรับและส่งข้อมูลต่าง ๆ ได้ทั่วถึงกันทั้งโลกแค่มีอุปกรณ์อิเล็กทรอนิกส์ เช่น สมาร์ทโฟน แล็ปท็อปหรือคอมพิวเตอร์ส่วนบุคคลและมีเครือข่ายอินเทอร์เน็ตก็สามารถรับส่งข้อมูลต่าง ๆ ได้ ไม่ว่าจะเป็นอีเมลการเผยแพร่ข่าวสาร รวมทั้งข้อมูลจากสังคมออนไลน์ ทำให้ผู้คนรับส่งและเผยแพร่ข้อมูลกันอยู่ตลอดเวลา จึงเป็นสาเหตุที่ก่อให้เกิดข้อมูลจำนวนมากในระบบ ซึ่งข้อมูลเหล่านั้นมีทั้งข้อความ ภาพ เสียง วิดีโอ ไฟล์ข้อมูลต่าง ๆ และไม่ได้มีการจัดเก็บอย่างเป็นระเบียบ ข้อมูลเหล่านั้นจะสิ้นเปลืองพื้นที่จัดเก็บโดยใช่เหตุถ้าเราไม่สามารถนำมาสร้างมูลค่าให้กับมันได้



รูปที่ 2.1 การเพิ่มขึ้นของข้อมูลขนาดใหญ่ [6]

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 สถาปัตยกรรมการให้บริการของคลาวด์

การประมวลผลแบบคลาวด์ (Cloud Computing) เป็นแนวคิดการประมวลผลเชิงเซิร์ฟเวอร์ไคลเอนต์ โดยมีแนวความคิด คือ ให้ทุกอย่างมาประมวลผลอยู่ที่เซิร์ฟเวอร์ส่วนกลางและให้เครื่องไคลเอนต์เป็นผู้รับส่งอินพุตและเอาต์พุตอยู่ปลายทาง โดยที่ผู้ใช้งานไม่รู้ว่าเซิร์ฟเวอร์อยู่ที่ใด ส่วนการติดต่อกันระหว่างผู้ใช้และคลาวด์นั้นจะติดต่อผ่านทางเครือข่ายหรืออินเทอร์เน็ต จุดเด่นของคลาวด์ คือ ความยืดหยุ่นและความคงทนถาวรของระบบ กล่าวคือ ความยืดหยุ่นหมายถึง ความสะดวกในการใช้งาน ไม่ว่าจะอยู่สถานที่ใดผู้ใช้ที่มีเครือข่ายหรืออินเทอร์เน็ตก็สามารถใช้งานคลาวด์ได้ หรือคำว่าความยืดหยุ่นในอีกแง่หนึ่ง คือ ความสามารถในการขยายหรือลดตัวของระบบตามการใช้งานจริง ทำให้ระบบมีความยืดหยุ่นรองรับการเปลี่ยนแปลงของจำนวนโหลดตามจำนวนผู้ใช้งานจริงได้ ส่วนคำว่าความคงทนถาวรของระบบ คือ ความมั่นคงของระบบ กล่าวคือ ระบบจะไม่มีทางล่มหรือมีโอกาสที่จะล่มได้น้อยมาก เพราะกลุ่มของคลาวด์เซิร์ฟเวอร์นั้นมีลักษณะที่ช่วยกันทำงานอยู่เป็นกลุ่มใหญ่ มีการกระจายโหลดการทำงานและสามารถทำงานทดแทนกันได้ในกรณีที่มีเครื่องใดเครื่องหนึ่งล้มเหลวไป

โดยทั่วไป สถาปัตยกรรมการให้บริการของคลาวด์แบ่งออกเป็น 3 รูปแบบ คือ การให้บริการเชิงโครงสร้าง หรือ IaaS (Infrastructure as a Service) การให้บริการเชิงแพลตฟอร์ม หรือ PaaS (Platform as a Service) และการให้บริการเชิงซอฟต์แวร์ หรือ SaaS (Software as a Service) [7]

ในการศึกษาอิสระนี้เลือกสร้างระบบทดลองโดยเลือกใช้บริการเชิงโครงสร้าง คือ การที่ผู้ให้บริการให้ผู้ใช้บริการเลือกใช้บริการได้ตั้งแต่ระดับของเซิร์ฟเวอร์เสมือนหรือที่เรียกว่าอินสแตนซ์ สามารถให้เรากำหนดได้ว่าต้องการหน่วยประมวลผล หน่วยความจำ พื้นที่ดิสก์เท่าใด รวมถึงดูแลการตั้งค่าเครือข่ายต่าง ๆ เองได้ สามารถติดตั้งระบบปฏิบัติการเองได้ และตั้งค่าปรับแต่งพารามิเตอร์ต่าง ๆ ได้เอง นั่นหมายถึงว่า ผู้ใช้งานจะได้เป็นเจ้าของเซิร์ฟเวอร์เสมือนตัวนั้น และต้องดูแลรองรับความเสี่ยงเองถึงขั้นของระบบปฏิบัติการและเครือข่าย เป็นต้น ในการศึกษาอิสระนี้เราเลือกใช้การให้บริการของเอมาซอนเว็บเซอร์วิส (Amazon Web Service) ซึ่งบริการนั้นมีชื่อว่า อีซีทู (EC2)

ยังมีการบริการอีกประเภทหนึ่ง คือ การให้บริการเชิงฮาดูป (Hadoop as a Service) เป็นบริการแบบพิเศษที่จะเห็นได้ว่าไม่ได้อยู่ในสถาปัตยกรรมของคลาวด์แบบทั่วไป เพราะการให้บริการนี้ถือเป็นกรณีพิเศษที่แยกออกมา แล้วแต่ว่าค่ายใดจะนำเสนอการให้บริการนี้ ในการศึกษาอิสระนี้เลือกใช้การให้บริการเชิงฮาดูปจากค่ายเอมาซอนเว็บเซอร์วิส โดยมีชื่อบริการว่า อีเอ็มอาร์ (EMR) ย่อมาจากคำว่า Elastic Map Reduce หลักการก็คือ ผู้ใช้งานไม่ต้องติดตั้งระบบปฏิบัติการเอง รวมถึงไม่ต้องติดตั้งฮาดูปและมาเฮาท์เอง ผู้ใช้งานแค่แจ้งความต้องการหลัก ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ไปว่าต้องการใช้ฮาดูปจำนวนกี่โหนด จากนั้นระบบจะสร้างขึ้นมาให้ ซึ่งต่างจากการให้บริการเชิงโครงสร้างที่ผู้ใช้งานต้องติดตั้งทุกอย่างเอง

2.3 ฮาดูป

ฮาดูป (Hadoop) เป็นเฟรมเวิร์กโอเพนซอร์สที่ถูกสร้างขึ้นเพื่อให้มีความสามารถในการจัดเก็บและประมวลผลกับข้อมูลขนาดใหญ่ได้ [8] สาเหตุที่ทำได้เพราะหลักการในการออกแบบฮาดูปออกแบบมาให้สอดคล้องกับการทำงานกับข้อมูลขนาดใหญ่ โดยเน้นเรื่องของหลักการกระจายการทำงาน ซึ่งก่อให้เกิดเป็นจุดเด่นของฮาดูป เพราะยิ่งมีการกระจายโหลดการทำงานมาก ก็ยิ่งทำให้การประมวลผลข้อมูลขนาดใหญ่ทำได้เร็วกว่าปกติ และยังช่วยเรื่องความทนทานของระบบด้วย (Fault Torrance) เพราะมีการช่วยกันทำงานอยู่ ดังนั้น หากมีคอมพิวเตอร์หรือเซิร์ฟเวอร์ตัวใดตัวหนึ่งทำงานล้มเหลว ก็ยังมีอีกหลายตัวที่ช่วยกันทำงานแทน หากเรามาพิจารณาองค์ประกอบของฮาดูปจะพบว่า ฮาดูปประกอบไปด้วย 2 องค์ประกอบหลัก คือ ระบบการกระจายแฟ้มข้อมูลแบบฮาดูป (Hadoop Distributed File Systems) และอัลกอริทึมการประมวลผลที่เรียกว่าแมปรีดิวซ์ (MapReduce) รายละเอียดของทั้งสององค์ประกอบมีดังนี้

2.3.1 ระบบการกระจายแฟ้มข้อมูลแบบฮาดูป

ระบบการกระจายแฟ้มข้อมูลแบบฮาดูป หรือ HDFS เป็นระบบการกระจายแฟ้มข้อมูลที่ถูกออกแบบมาให้มีความทนทานต่อความผิดพลาดสูง และจากที่มันถูกออกแบบมาให้มีการกระจายของไฟล์อยู่มาก ทำให้ประสิทธิภาพด้านความเร็วในการเข้าถึงข้อมูลมีมากขึ้นไปด้วย ดังนั้น จึงเหมาะสำหรับแอปพลิเคชันที่มีชุดของข้อมูลขนาดใหญ่ ข้อดีของระบบการกระจายแฟ้มข้อมูลแบบฮาดูป คือ เพิ่มความแข็งแกร่งของระบบ เช่น เมื่อมีฮาร์ดแวร์ตัวใดตัวหนึ่งทำงานผิดพลาดก็ยังมีฮาร์ดแวร์อีกหลายตัวช่วยกันทำงานแบบกระจายโหลด ทำให้ระบบล่มได้ยากและยังสามารถเข้าถึงข้อมูลได้เร็ว รองรับปริมาณของข้อมูลขนาดใหญ่ได้ดี ฮาดูปเรียกการทำงานเหล่านี้ว่า ฮาดูปคลัสเตอร์ (Hadoop Cluster)

ในกลุ่มของฮาดูปหรือฮาดูปคลัสเตอร์นั้น จะมีคอมพิวเตอร์เพียงแค่หนึ่งเครื่องที่ทำหน้าที่เป็นหัวหน้าหรือในศัพท์ของฮาดูปเรียกว่า เนมโหนด (Name Node) ทำหน้าที่เก็บข้อมูลประเภทเมทาดาทา (Metadata) และข้อมูลบันทึกการใช้งาน (Logs) ที่คอยเก็บข้อมูลว่าโหนดต่าง ๆ ในคลัสเตอร์เก็บชิ้นส่วนของข้อมูลใดอยู่บ้าง ส่วนคอมพิวเตอร์เครื่องอื่น ๆ จะทำหน้าที่เป็น เดทาโหนด (Data Node) แบ่งกันเก็บแต่ละชิ้นส่วนของข้อมูลขนาดใหญ่และกระจายโหลดการทำงาน

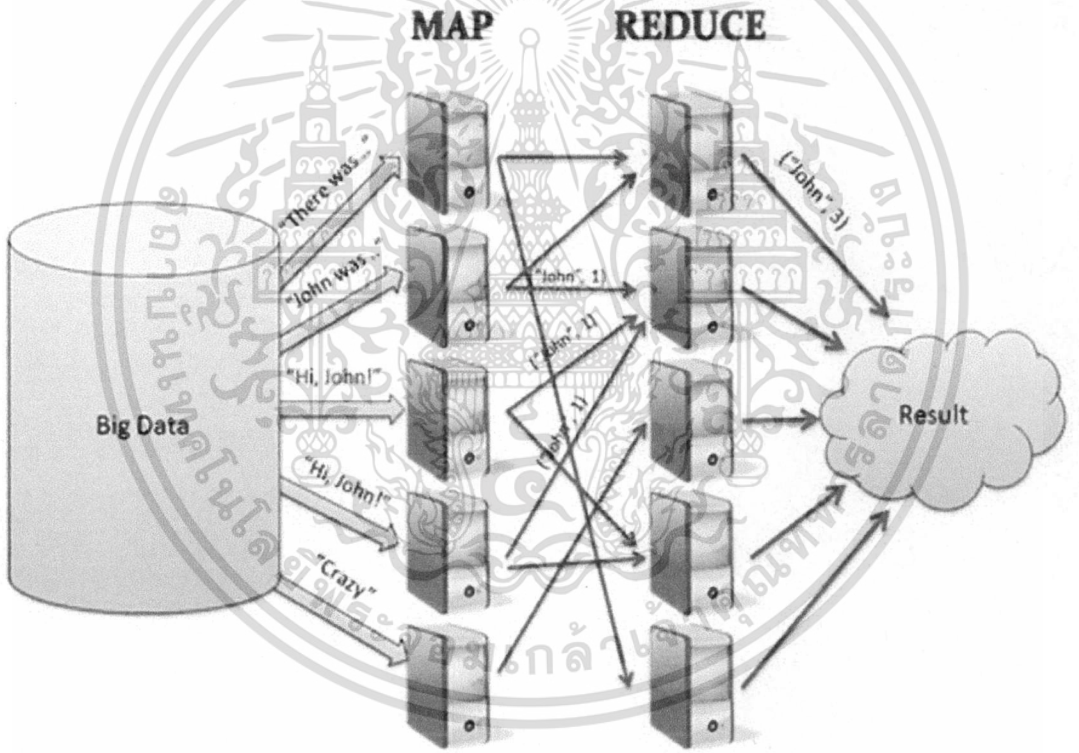
2.3.2 แมปรีดิวซ์

แมปรีดิวซ์เป็นเฟรมเวิร์กในการเขียนโปรแกรมของฮาดูปที่ช่วยในการประมวลผลข้อมูลขนาดใหญ่ ซึ่งใช้หลักการเดียวกับระบบการกระจายแฟ้มข้อมูลแบบฮาดูป [9] คือ อาศัยเครื่องคอมพิวเตอร์หลายเครื่องช่วยกันทำงาน แต่ในส่วนนี้จะเน้นอัลกอริทึมของการแมปและการรีดิวซ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เท่านั้น โดยการทำแมปหมายถึง การจับคู่ของ คีย์/แวลู (Key/Value) หรือหาคีย์เวิร์ด (Keyword) ว่าสิ่งที่เราต้องการในข้อมูลเหล่านี้ประกอบด้วยคีย์เวิร์ดอะไรบ้าง จากนั้น เมื่อช่วยกันหาจนครบแล้วจึงส่งผลลัพธ์นั้นไปให้กับรีดิวซ์เพื่อรวบรวมผลลัพธ์และแสดงผลออกทางเอาต์พุต

ในกลุ่มของฮาดูปหรือฮาดูปคลัสเตอร์จะมีคอมพิวเตอร์เพียงแค่หนึ่งเครื่องที่ทำหน้าที่เป็นหัวหน้าหรือในศัพท์ของฮาดูปเรียกว่า จ็อบแทร็กเกอร์ (Job Tracker) ซึ่งเป็นผู้ที่กระจายงานไปให้กับโหนดอื่น ๆ ที่เรียกว่า ทาสก์แทร็กเกอร์ (Task Tracker) เพื่อช่วยกันทำงาน โดยเริ่มต้นจากการที่ จ็อบแทร็กเกอร์จะแบ่งงานหรือข้อมูลชิ้นใหญ่ให้เป็นชิ้นขนาดเล็ก แล้วส่งให้กับทาสก์แทร็กเกอร์แต่ละตัว เพื่อประมวลผลแมปรีดิวซ์โดยการแมปหาคีย์แวลูและรวบรวมงานจากทุก ๆ โหนดในคลัสเตอร์เข้าเป็นชิ้นเดียวกันจนกลายเป็นเอาต์พุตที่ต้องการในที่สุด [10]



รูปที่ 2.2 หลักการทำงานของแมปรีดิวซ์

อย่างที่กล่าวมาข้างต้นจะเห็นว่า การทำงานของฮาดูปคลัสเตอร์จะประกอบไปด้วยมาสเตอร์และเวิร์กเกอร์ ในส่วนของระบบการกระจายแฟ้มข้อมูลแบบฮาดูป เครื่องที่ทำหน้าที่เป็นมาสเตอร์ จะมีชื่อว่า เนมโหนด ส่วนเครื่องที่ทำหน้าที่เป็นเวิร์กเกอร์ จะมีชื่อเรียกว่า เดทาโหนด และส่วนของแมปรีดิวซ์ เครื่องที่ทำหน้าที่เป็นมาสเตอร์จะชื่อว่า จ็อบแทร็กเกอร์ ส่วนเครื่องที่ทำ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หน้าที่เป็นเวิร์กเกอร์จะมีชื่อเรียกว่า ทาสก์แทร็กเกอร์ นอกจากนี้แล้วยังมีคอมพิวเตอร์อีกเครื่องหนึ่งที่ต้องทำหน้าที่เป็น เนมโหนดรอง (Secondary Namenode) เพื่อรองรับการทำงานแทนเครื่องเนมโหนดในกรณีที่ฮาร์ดแวร์ล้มเหลวหรือมีการทำงานที่ผิดพลาด [11]

นอกจากนั้นแล้ว การพัฒนาระบบการให้คำแนะนำในที่นี้ ยังได้ใช้ความสามารถของเฟรมเวิร์กโอเพนซอร์สอีกตัวหนึ่งที่สามารถทำงานร่วมกับฮาดูปได้ คือ มาเฮาท์ (Mahout) ซึ่งมีความสามารถด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และยังนำมาประมวลผลร่วมกับฮาดูปได้ [12] ความสามารถของมาเฮาท์นั้นแบ่งออกเป็น 3 เทคนิคหลัก คือ การให้คำแนะนำ (Recommendation) การจำแนกประเภท (Classification) และการจัดกลุ่มข้อมูล (Clustering) ในการศึกษาอิสระนี้เราจะใช้เทคนิคการให้คำแนะนำของมาเฮาท์มาประยุกต์ใช้ในการสร้างระบบ ซึ่งระบบการให้คำแนะนำ โดยใช้เทคนิคที่เรียกว่า การกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม หมายถึง วิธีแนะนำที่ประเมินค่าความคล้ายคลึงกันระหว่างผู้ใช้ปัจจุบันกับผู้ใช้คนอื่นในระบบ และเลือกที่จะแนะนำวัตถุที่ได้รับคะแนนความพึงพอใจจากผู้ใช้คนอื่นที่มีลักษณะคล้ายกับผู้ใช้ปัจจุบันมากที่สุด ซึ่งจะเอาข้อมูลตรงนี้ไปสร้างเป็นโปรไฟล์ของผู้ใช้แต่ละคน สำหรับการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ยังถูกแบ่งออกเป็นสองรูปแบบย่อย ๆ คือ การคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม

2.4 ระบบการให้คำแนะนำ

หลักการทำงานของระบบการให้คำแนะนำนั้นประกอบไปด้วย 4 ขั้นตอนหลักคือ การเตรียมข้อมูล การคำนวณค่าความคล้ายคลึงกัน การทำนายผลคำแนะนำ และการให้คำแนะนำ [13]

โดยภายในขั้นตอนการเตรียมข้อมูล คือ การเตรียมข้อมูลให้อยู่ในรูปแบบของแบบจำลองข้อมูล ซึ่งประกอบไปด้วยหมายเลขผู้ใช้ (User ID) หมายเลขสิ่งของ (Item ID) และคะแนนความชื่นชอบที่ผู้ใช้มีให้กับสิ่งของชิ้นนั้น (Rating) หลังจากนั้นจึงสร้างเมตริกซ์ระหว่างผู้ใช้และสิ่งของ เพื่อคำนวณขั้นต่อไป

ภายในขั้นตอนของการคำนวณหาค่าความคล้ายคลึง เป็นวิธีที่ใช้คำนวณความคล้ายคลึงของผู้ใช้กับกลุ่มผู้ใช้เป้าหมาย โดยนำคะแนนการให้เรทติ้งของผู้ใช้มาคำนวณและนำคะแนนเรทติ้งของผู้ใช้คนอื่น ๆ ที่ได้ให้คะแนนค่าความชอบกับชิ้นข้อมูลนั้นร่วมกัน ซึ่งเรียกว่า โคเรท (Co-rate)

กำหนดให้ Sim(t,c) คือ ค่าความคล้ายคลึงจากโคเรทระหว่างรายวิชาt กับ c และ Ru,t กับ Ru,c คือ ค่าคะแนนความนิยมที่ผู้เรียนu มีต่อ รายวิชา t และผู้เรียน u ต่อรายวิชา c แสดงดังสมการที่ 2.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$sim(t,c) = \frac{\sum_{u \in U}(R_{u,t} - \overline{R}_t)(R_{u,c} - \overline{R}_c)}{\sqrt{\sum_{u \in U}(R_{u,t} - \overline{R}_t)^2}\sqrt{\sum_{u \in U}(R_{u,c} - \overline{R}_c)^2}} \quad (2.1)$$

ในขั้นตอนการทำนาย เป็นการทำนายค่าความชอบของผู้ใช้ต่อข้อมูลชิ้นหนึ่ง โดยคำนวณจากความชอบและความคล้ายกันระหว่างชิ้นข้อมูล ซึ่งเซ็ตที่นำมาคำนวณนั้น จะเป็นเซ็ตของสมาชิกข้างเคียง (Neighborhood) ที่ถูกเลือกไว้ กำหนดให้ Pu,i คือ ค่าความพึงพอใจที่คาดว่าผู้ใช้เป้าหมาย u จะมีต่อชิ้นข้อมูล i แสดงดังสมการ

$$P_{u,i} = \frac{\sum all_{similar\_items}(sim_{i,k} * R_{u,k})}{\sum all_{similar\_items}(|sim_{i,k}|)} \quad (2.2)$$

ในขั้นตอนสุดท้าย คือ การแสดงรายการแนะนำ โดยมีหลักการคือ เรียงลำดับรายการที่มีผลคะแนนของการทำนายจากมากที่สุดไปยังน้อยที่สุด และแสดงผลลัพธ์การแนะนำตามจำนวนลำดับที่ต้องการแนะนำ

2.5 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าวิธีส่วนใหญ่ที่ใช้วัดประสิทธิภาพของอัลกอริทึมการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมนั้น ใช้วิธีการดึงข้อมูลบางส่วนในอินพุตของระบบออกและนำเข้าไปประมวลผลในระบบการให้คำแนะนำ หลังจากนั้นจึงนำผลลัพธ์ที่ได้มาเปรียบเทียบกับก่อนการดึงข้อมูลว่ามีความถูกต้องมากน้อยเพียงใด [14] ยังมีงานวิจัยบางส่วนที่ทำการเปรียบอัลกอริทึมการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วมมาแล้ว แต่ไม่ได้ทดสอบกับชุดข้อมูลขนาดใหญ่ [15] ดังนั้น ในการศึกษานี้จะทำการทดสอบระบบการให้คำแนะนำกับข้อมูลขนาดใหญ่เพื่อพิสูจน์ว่าอัลกอริทึมใดเหมาะสมกับการประมวลผลบนข้อมูลขนาดใหญ่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
การศึกษาและทดลอง

ในบทนี้จะนำเสนอในเรื่องของกระบวนการคิดและวิธีการทดลองในขั้นตอนต่าง ๆ โดยจะเริ่มตั้งแต่การติดตั้งฮาดูป การสร้างระบบการให้คำแนะนำ โดยภายใต้อัลกอริทึมของระบบการให้คำแนะนำจะประกอบไปด้วยการทำงานของแมปรีดิวซ์หลาย ๆ งานรวมกัน และทั้งหมดนี้จะต้องประมวลผล และทำงานอยู่บนระบบการกระจายแฟ้มข้อมูลแบบฮาดูปเท่านั้น

3.1 การออกแบบแผนการทดลอง

1) ทำการติดตั้งฮาดูปบนเอมาซอนเว็บเซอร์วิสอีซีทู และเขียนโปรแกรมทดสอบด้วยวิธีการแมปรีดิวซ์ เพื่อศึกษากระบวนการทำงานของอัลกอริทึมที่นำมาประยุกต์ใช้กับข้อมูลขนาดใหญ่ ซึ่งเป็นหนึ่งในจุดประสงค์ที่สำคัญของการศึกษาอิสระครั้งนี้

2) พัฒนาระบบการให้คำแนะนำ โดยแบ่งการพัฒนาออกเป็น 2 อัลกอริทึม คือ การคัดกรองผู้ใช้ร่วม และการคัดกรองสิ่งของร่วม

3) เตรียมกลุ่มข้อมูลที่ใช้ในการทดลอง โดยนำข้อมูลขนาดใหญ่มาจากเว็บไซต์มูฟวี่เลนส์ ทริปแอดไวเซอร์และเยลป์ โดยแบ่งข้อมูลเป็นอย่างละ 3 ชุดข้อมูล คือ ชุดข้อมูลขนาดเล็ก ชุดข้อมูลขนาดกลาง และชุดข้อมูลขนาดใหญ่ ตามลำดับ

4) นำข้อมูลเหล่านั้นเข้าไปประมวลผลในระบบการให้คำแนะนำที่สร้างขึ้นมา หลังจากนั้นให้บันทึกระยะเวลาในการประมวลผลและบันทึกผลลัพธ์จากการแนะนำข้อมูลเอาไว้

5) นำข้อมูลนำเข้าจากข้อ 3 มาแยกผู้ใช้ออกมา 100 ราย และลบข้อมูลการคะแนนของแต่ละผู้ใช้เหล่านั้นออกเป็น 10%, 15%, 20%, 25% และ 30% ตามลำดับ

6) นำข้อมูลจากข้อ 5 เข้าไปประมวลผลในระบบการให้คำแนะนำที่สร้างขึ้นมา และบันทึกผลลัพธ์จากการแนะนำข้อมูลเอาไว้

7) เปรียบเทียบผลลัพธ์จากข้อ 4 และข้อ 6 เพื่อดูอัตราส่วนความถูกต้องในการทำนาย

8) บันทึกผลลัพธ์ สรุปผลการทดลอง และวิเคราะห์ผลการเปรียบเทียบ

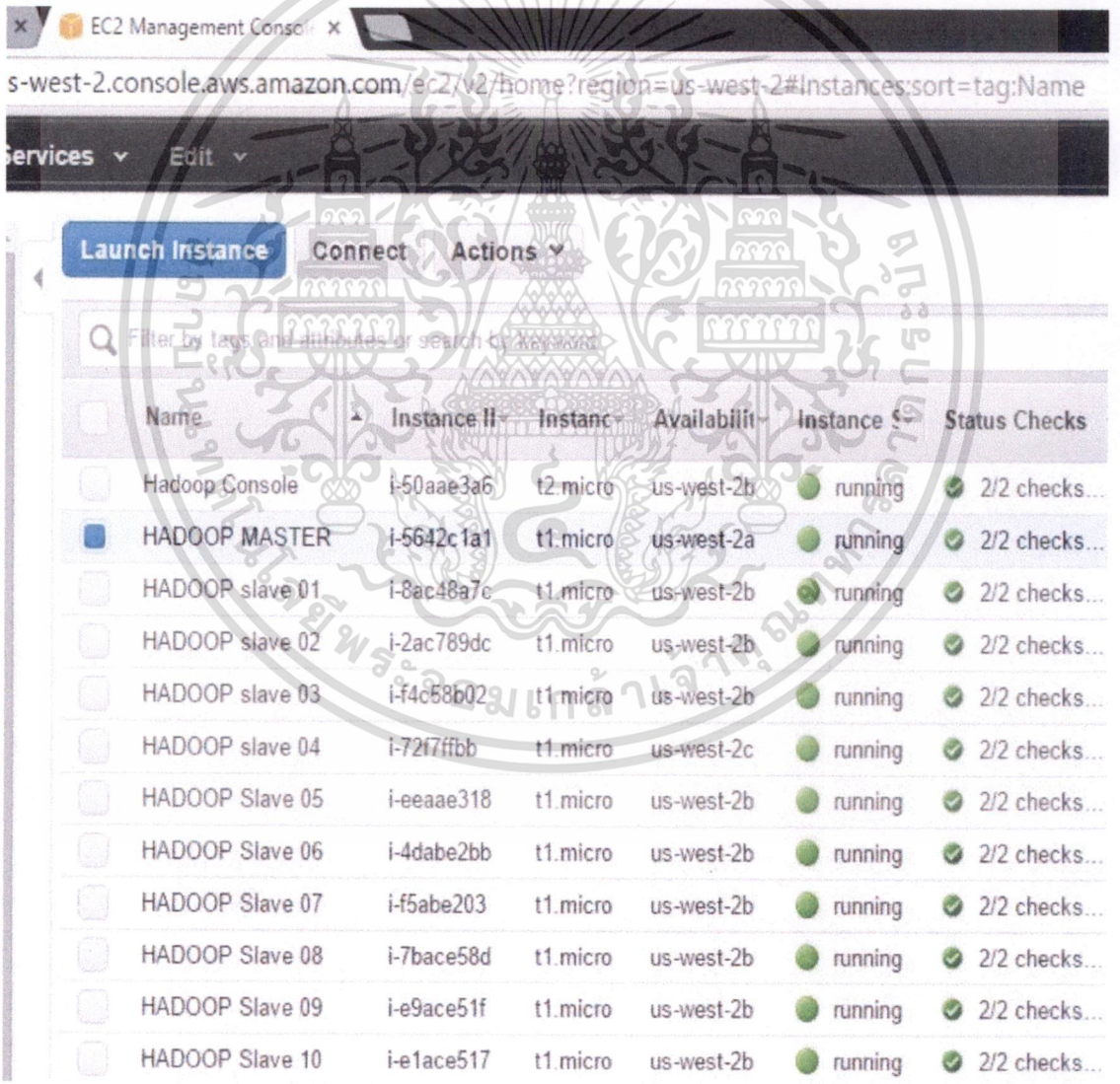
3.2 การเตรียมระบบบนคลาวด์

ในการศึกษาอิสระนี้เลือกใช้บริการของเอมาซอนเว็บเซอร์วิสอีซีทู โดยสร้างเซิร์ฟเวอร์เสมือนหรืออินสแตนซ์ขึ้นมาเพื่อใช้เป็นตัวทดลองและติดตั้งฮาดูป ในการทดลองนี้เลือกติดตั้งระบบปฏิบัติการบนเซิร์ฟเวอร์เสมือนเป็นอูบันตูเซิร์ฟเวอร์ เวอร์ชัน 14.04 ในการกำหนดค่าต่าง ๆ ระหว่างการสร้างเซิร์ฟเวอร์เสมือน ควรปรับตั้งค่าตามข้อกำหนดพื้นฐานของฮาดูป เช่น ในขั้นตอน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ของการตั้งค่ากลุ่มความปลอดภัย (Security Group) ต้องเปิดให้ใช้งาน Secure Shell (SSH) ได้ เป็นต้น

ในการศึกษาอิสระนี้สร้างเซิร์ฟเวอร์เสมือนขึ้นมาทั้งหมด 11 เซิร์ฟเวอร์ โดยที่มีเพียง 1 เซิร์ฟเวอร์เท่านั้นที่ทำหน้าที่เป็นมาสเตอร์โหนด อีก 10 เซิร์ฟเวอร์ทำหน้าที่เป็นเวิร์กเกอร์โหนด และอีก 1 เซิร์ฟเวอร์ทำหน้าที่เป็นคอนโซล ซึ่งทำหน้าที่เป็นตัวจัดการระบบ เช่น การเขียนโปรแกรมเพื่อพัฒนาระบบการให้คำแนะนำโดยใช้เครื่องมือพัฒนาระบบอีคลิปส์ (Eclipse IDE) เพื่อการจัดการฮาดูปผ่านเว็บเบราว์เซอร์ เพื่อการมอนิเตอร์ไฟล์ซิสเท็มส์ของฮาดูป หรือการอัพโหลด และดาวน์โหลดไฟล์จากระบบแฟ้มกระจายแบบฮาดูป (HDFS) เป็นต้น



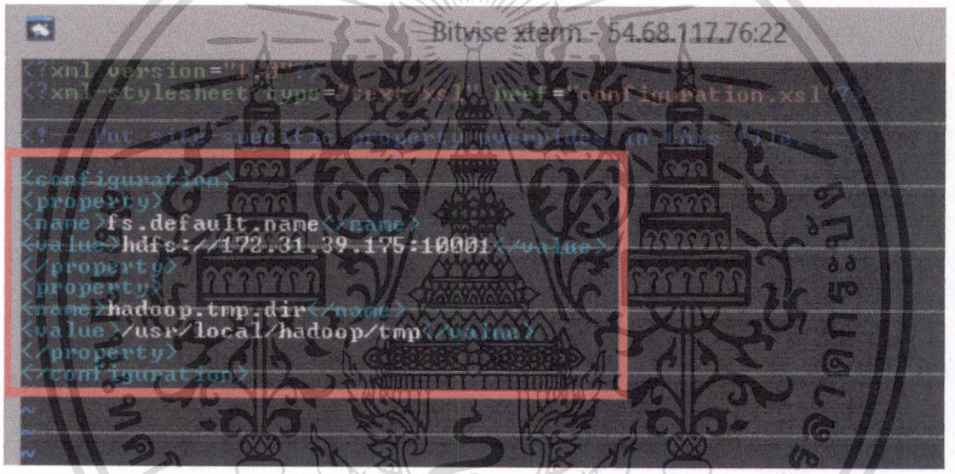
รูปที่ 3.1 ระบบฮาดูปคลัสเตอร์บนเอมาซอนเว็บเซอร์วิส

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

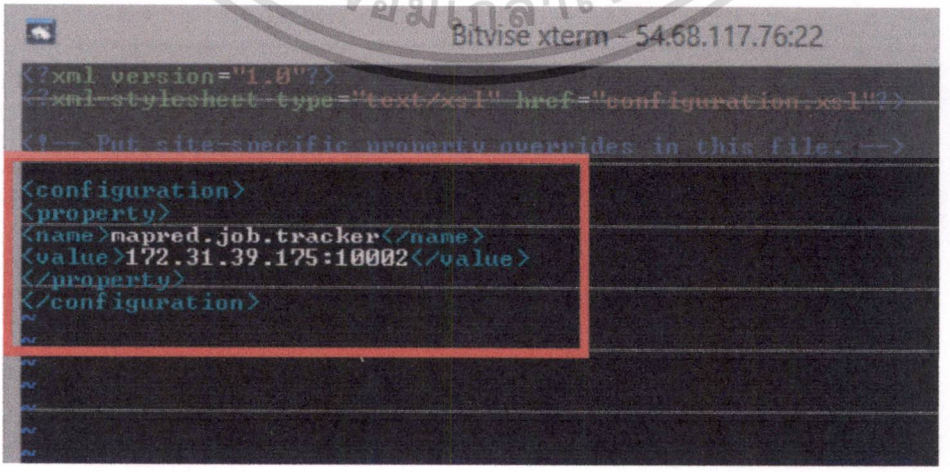
3.3 การติดตั้งฮาดูปและการทดสอบ

ก่อนติดตั้งฮาดูป ระบบควรเตรียมพร้อมในการรองรับการทำงานของฮาดูปให้เรียบร้อย เช่น การตั้งค่าไอพีแอดเดรส การตั้งค่า Secure Shell (SSH) การติดตั้งจาวา รวมถึงการอัพเดทแพทช์ของระบบปฏิบัติการ

สำหรับไอพีแอดเดรสบนระบบเอมาซอนเว็บเซอร์วิสอีซีทู จะพิเศษกว่าไอพีแอดเดรสของคอมพิวเตอร์หรือเซิร์ฟเวอร์ทั่วไปที่ไม่ได้อยู่บนระบบคลาวด์ เพราะเซิร์ฟเวอร์เสมือนที่อยู่บนระบบคลาวด์ เช่น เอมาซอนเว็บเซอร์วิสอีซีทู จะมีค่าไอพีแอดเดรสอยู่ 2 ประเภท คือ ไอพีสาธารณะ (Public IP) และไอพีส่วนตัว (Private IP) ซึ่งการนำค่าของไอพีแอดเดรสไปใส่ไว้ในคอนฟิกกูเรชันของฮาดูปนั้น ต้องใช้ไอพีส่วนตัวเท่านั้น มิเช่นนั้นแล้ว หากนำไอพีสาธารณะไปใส่ไว้ในคอนฟิกกูเรชันของฮาดูป จะทำให้ระบบฮาดูปไม่ทำงานและไม่สามารถเปิดขึ้นมาใช้งานได้

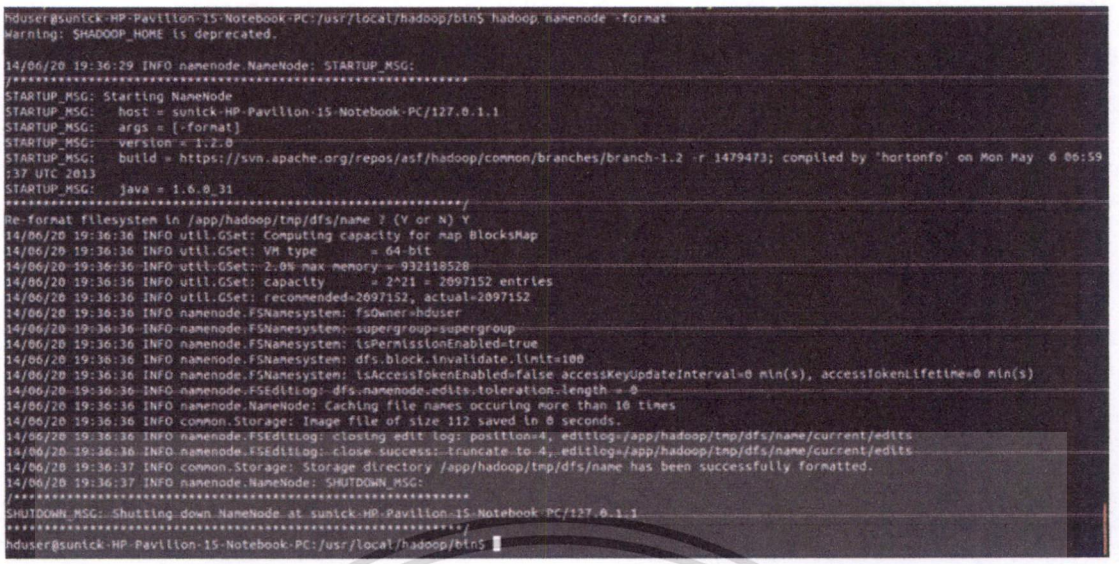


รูปที่ 3.2 ใช้ไอพีส่วนตัวในการตั้งค่าของระบบแฟ้มกระจายแบบฮาดูป



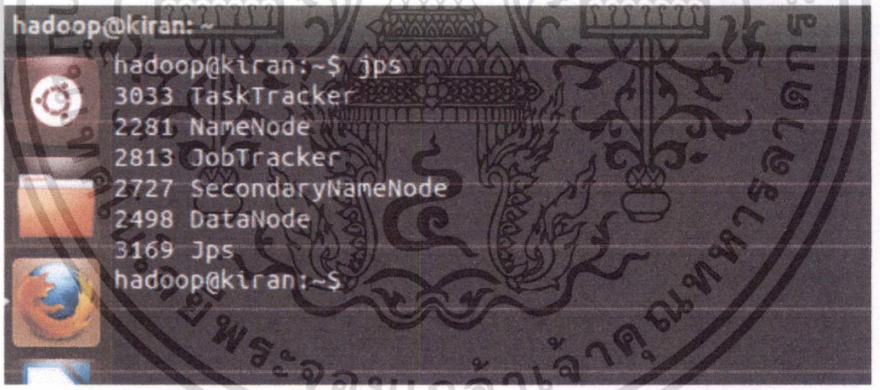
รูปที่ 3.3 ใช้ไอพีส่วนตัวในการตั้งค่าของจ็อบแทร็กเกอร์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.4 ล็อกที่แสดงบนหน้าจอหลังจากติดตั้งฮาดูปสำเร็จ

เมื่อติดตั้งฮาดูปเสร็จเรียบร้อยแล้วต้องตรวจสอบดูว่าเซอร์วิสของฮาดูปทำงานครบทุกตัวหรือไม่ โดยการใช้คำสั่งเจพีเอส (jps) ตามภาพด้านล่าง



รูปที่ 3.5 ตรวจสอบการทำงานของโปรเซสฮาดูป

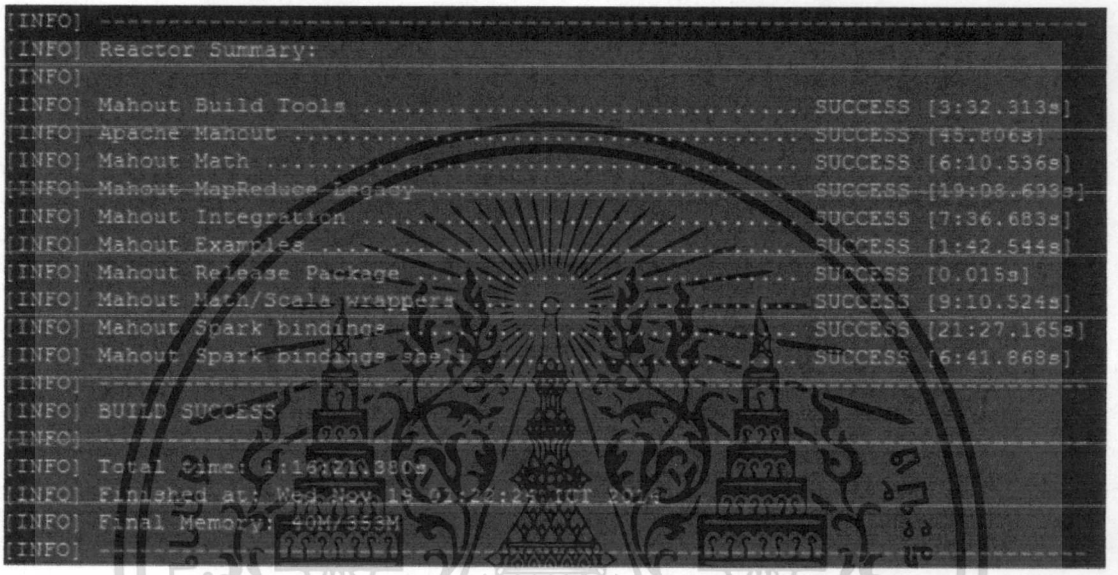
จากภาพจะพบว่า ถ้าฮาดูปทำงานได้เป็นปกตินั้น จะต้องมีเซอร์วิสทำงานอย่างน้อย 5 เซอร์วิส คือ ทาสก์แทร็กเกอร์ (TaskTracker) เนมโหนด (NameNode) จ็อบแทร็กเกอร์ (JobTracker) เซคันดารีเนมโหนด (SecondaryNameNode) และเดทาโหนด (DataNode)

หลังจากนั้นจึงติดตั้งมาเฮาท์ไลบรารี (Mahout Library) เพื่อช่วยในการพัฒนาระบบการให้คำแนะนำ ก่อนเริ่มต้นต้องเตรียมพร้อมเรื่องความต้องการของระบบ จำเป็นต้องมีชุดพัฒนาจาวา (Java JDK), มาเว็น (Maven) และซับเวอร์ชัน (Subversion) แล้วจึงจะเริ่มติดตั้งมาเฮาท์ได้ จากนั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ก่อนจะใช้งานมาเฮาท์จะต้องเปิดการทำงานของฮาดูปและตรวจสอบว่ามีโปรเซสทำงานอยู่ครบเรียบร้อย เนื่องด้วยการทดลองนี้จะใช้มาเฮาท์เพื่อฝึกการเรียนรู้ของเครื่องโดยประมวลผลข้อมูลขนาดใหญ่ที่ทำงานอยู่บนระบบการกระจายแฟ้มข้อมูลแบบฮาดูป

หลังจากการติดตั้งมาเฮาท์เสร็จสิ้น จอแสดงผลจะขึ้นรายละเอียดบนหน้าจอแสดงผลดังรูปที่ 3.6 เพื่อรายงานผลการติดตั้งว่าสำเร็จหรือล้มเหลว



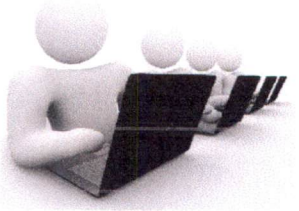
รูปที่ 3.6 รายงานผลการติดตั้งมาเฮาท์สำเร็จ

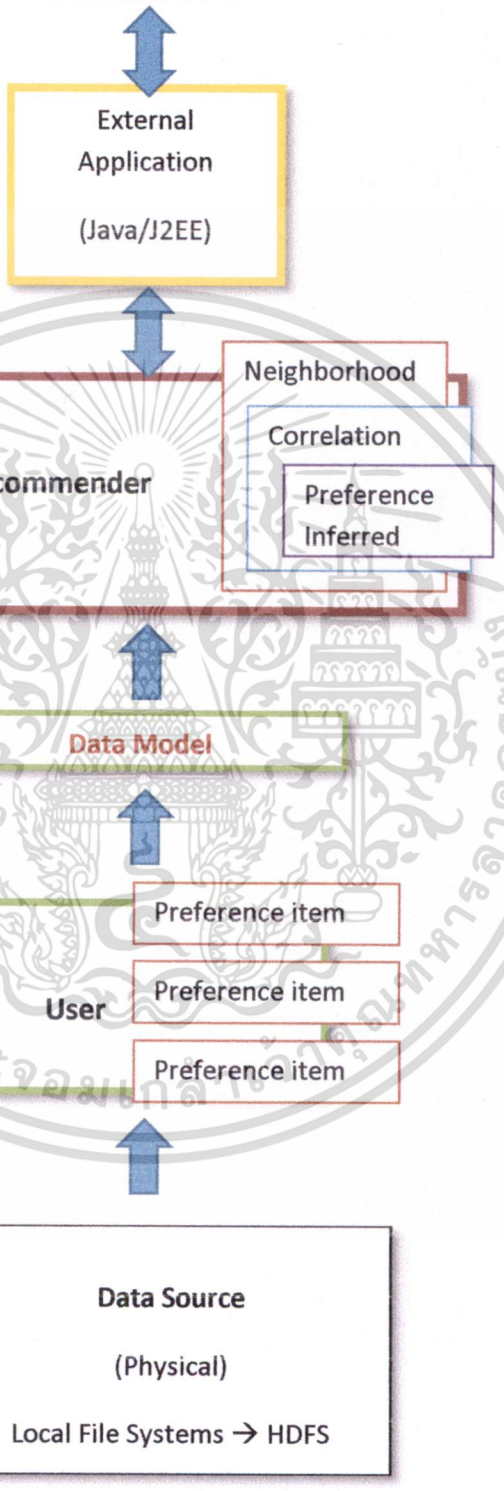
3.4 การพัฒนาระบบการให้คำแนะนำ

ในการศึกษาอิสระนี้ได้สร้างระบบแนะนำภาพยนตร์ที่น่าสนใจ ระบบแนะนำโรงแรม และระบบแนะนำร้านอาหาร ในการศึกษานี้เราใช้เทคนิคการกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ซึ่งแยกเป็น 2 อัลกอริทึม คือ การคัดกรองสิ่งของร่วมและการคัดกรองผู้ใช้ร่วม โดยเริ่มจากการนำเข้าข้อมูล กระบวนการเตรียมแบบจำลองข้อมูล การหาค่าความคล้ายคลึงระหว่างผู้ใช้และกลุ่มผู้ใช้เป้าหมาย การทำนายหรือการพยากรณ์ค่าความชอบของผู้ใช้ต่อข้อมูลชิ้นใดชิ้นหนึ่ง โดยพิจารณาจากความชอบและความคล้ายระหว่าง ชิ้นข้อมูลนั้นกับ ชิ้นข้อมูลอื่น ๆ ซึ่งจะนำเซตของสมาชิกข้างเคียงที่ถูกเลือกไว้นำมาคำนวณ และขั้นตอนสุดท้าย คือ การสร้างรายการแนะนำ โดยนำรายการที่มีคะแนนความชื่นชอบมาเรียงตามลำดับจากมากที่สุดไปยังน้อยที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง





รูปที่ 3.7 หลักการทำงานของระบบการให้คำแนะนำ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้
144192

จากรูปที่ 3.7 อธิบายถึงหลักการทำงานของระบบให้คำแนะนำที่ประยุกต์ใช้ไลบรารีของมาเฮาท์และประมวลผลบนฮาดูป ซึ่งมีขั้นตอนการทำงานหลัก ๆ คือ รับอินพุตมาในลักษณะข้อมูลดิบ จากนั้นจึงนำข้อมูลดิบเหล่านั้นมาสร้างเป็นแบบจำลองข้อมูล เพื่อให้พร้อมประมวลผล เมื่อพร้อมแล้วจึงนำแม่แบบข้อมูลเข้าไปประมวลผลในระบบ และรอผลลัพธ์การให้คำแนะนำที่แสดงออกมาเป็นเอาต์พุต

3.5 การประมวลผลข้อมูลขนาดใหญ่บนระบบการให้คำแนะนำบนฮาดูป

ในการนำข้อมูลขนาดใหญ่เข้าไปประมวลผลในระบบการให้คำแนะนำนั้น จะต้องมีการเตรียมข้อมูลเหล่านั้นให้อยู่ในรูปแบบของแบบจำลองข้อมูลก่อน โดยที่รูปแบบของแบบจำลองข้อมูลจะประกอบไปด้วย หมายเลขผู้ใช้ (User ID) หมายเลขสิ่งของ (Item ID) และคะแนนความชื่นชอบที่ผู้ใช้มีต่อสิ่งของชิ้นนั้น (Rating) มีลักษณะเช่น รูปที่ 3.8 ดังนี้

1	196	5	888205088
1	679	2	880037164
1	384	4	879877127
1	143	5	880474293
1	423	5	881107687
1	515	4	881103977
1	20	3	881171009
1	288	1	879667584
1	219	4	884112673
1	526	3	882141053
1	919	4	884920949

รูปที่ 3.8 ตัวอย่างแบบจำลองข้อมูลจากมูฟวี่เลนส์

ภาพตัวอย่างแบบจำลองข้อมูลจากมูฟวี่เลนส์ นั้นจะสังเกตว่าในคอลัมน์แรก คือ หมายเลขผู้ใช้ คอลัมน์ที่สอง คือ หมายเลขภาพยนตร์ คอลัมน์ที่สาม คือ ค่าคะแนนความชื่นชอบที่ผู้ใช้ให้กับภาพยนตร์เรื่องนั้นไว้และในคอลัมน์สุดท้ายเป็นไทม์แสตมป์ (Timestamp) คือ ตัวแปรที่ใช้เก็บบันทึกวันและเวลาที่ผู้ใช้รายนั้นได้ให้คะแนนภาพยนตร์ไว้ ซึ่งคอลัมน์สุดท้ายนี้ไม่ได้นำไปร่วมประมวลผลในระบบการให้คำแนะนำ

สำหรับข้อมูลขนาดใหญ่ที่เก็บคะแนนความชื่นชอบของผู้ใช้เกี่ยวกับโรงแรมและร้านอาหารที่นำมาจากทริปแอดไวเซอร์และเยลป์ จะเป็นข้อมูลที่ยังไม่ได้อยู่ในรูปแบบของแบบจำลองข้อมูล จึงต้องนำมาจัดเรียงข้อมูลให้อยู่ในรูปแบบของแบบจำลองข้อมูลก่อน ซึ่งต้องประกอบไปด้วย หมายเลขผู้ใช้ หมายเลขโรงแรมหรือร้านอาหาร และคะแนนความชื่นชอบที่ผู้ใช้มีต่อโรงแรมหรือร้านอาหาร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลักษณะการไหลของข้อมูลสำหรับระบบการให้คำแนะนำ สามารถแสดงออกมาได้ดังภาพที่ 3.9 ดังนี้

รูปที่ 3.9 การทำงานของระบบการให้คำแนะนำ

ยกตัวอย่างระบบการให้คำแนะนำภาพยนตร์ เริ่มจากการป้อนข้อมูลที่เป็นแบบจำลองข้อมูลเข้าไปในระบบ จากนั้นระบบจะไปคำนวณหาภาพยนตร์สิบอันดับแรกที่จะแนะนำให้กับผู้ใช้ โดยเรียงจากภาพยนตร์ที่มีคะแนนสูงสุดไปยังน้อยสุด คำว่าคะแนนในที่นี้ได้มาจากการพยากรณ์คะแนนของอัลกอริทึม ว่าจะให้ภาพยนตร์ใดได้คะแนนความชอบเท่าใด

ผลลัพธ์ที่ออกมาจากมีลักษณะเป็น หมายเลขผู้ใช้และรายชื่อภาพยนตร์สิบอันดับรวมทั้งคะแนนความชื่นชอบที่พยากรณ์ออกมา ดังภาพด้านล่าง

```
1       [1028:5.0,1009:5.0,926:5.0,879:5.0,866:5.0,845:5.0,813:5.0,762:5.0,748:5.0,742:5.0]
2       [748:5.0,546:5.0,568:5.0,288:5.0,25:5.0,531:5.0,527:5.0,523:5.0,518:5.0,516:5.0]
3       [652:5.0,732:5.0,124:5.0,382:5.0,231:5.0,539:5.0,285:5.0,531:5.0,4:5.0,419:5.0]
```

รูปที่ 3.10 ตัวอย่างผลลัพธ์ของการแนะนำภาพยนตร์สิบอันดับ

ตัวเลขในคอลัมน์แรกสุด คือ หมายเลขผู้ใช้ จากภาพตัวอย่างด้านบนจะแสดงเฉพาะผู้ใช้หมายเลข 1 ถึง 3 และในกรอบสีแดง คือ ตัวอย่างชุดข้อมูลที่หมายถึงหมายเลขภาพยนตร์ที่ 652 ได้ค่าพยากรณ์คะแนนความชื่นชอบเท่ากับ 5.0 ซึ่งถูกแนะนำมาเป็นอันดับที่ 1 เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สังเกตว่าผลลัพธ์ที่ออกมาจะมีลักษณะเป็นตัวเลข ซึ่งทำให้ยากต่อผู้อ่านในการเข้าใจว่าระบบแนะนำภาพยนตร์เรื่องใดมา จึงต้องสร้างโปรแกรมเล็ก ๆ เพื่อประมวลผลชุดตัวเลขเหล่านี้ให้แสดงออกมาในรูปตัวอักษร โดยดึงข้อมูลจากฐานข้อมูลในฮาดูปที่เก็บรายละเอียดเรื่องสินค้า

ตัวอย่างผลลัพธ์ที่ได้จากประมวลผลออกมาเป็นตัวอักษรผ่านระบบการให้คำแนะนำภาพยนตร์ที่น่าสนใจสำหรับผู้ใช้รายหนึ่ง ซึ่งแสดงออกมาทั้งสองอัลกอริทึมคือการคัดกรองสิ่งของร่วมและการคัดกรองผู้ใช้ร่วม

```
        Movies Recommendation Systems
      Using Item-based Collaborative Filtering


==========================================================
          Top Ten Recommended Movies
==========================================================
Saint, The (1997)
Indian Summer (1996)
Broken Arrow (1996)
Speed (1994)
Anastasia (1997)
People vs. Larry Flynt, The (1996)
Casablanca (1942)
Trainspotting (1996)
Courage Under Fire (1996)
Money Talks (1997)
==========================================================
```

รูปที่ 3.11 ผลลัพธ์ของระบบแนะนำภาพยนตร์โดยใช้อัลกอริทึมการคัดกรองสิ่งของร่วม

```
        Movies Recommendation Systems
      Using User-based Collaborative Filtering


==========================================================
          Top Ten Recommended Movies
==========================================================
Scream (1996)
Star Wars (1977)
Wedding Singer, The (1998)
Starship Troopers (1997)
Air Force One (1997)
Conspiracy Theory (1997)
Contact (1997)
Indiana Jones and the Last Crusade (1989)
Desperate Measures (1998)
Seven (Se7en) (1995)
==========================================================
```

รูปที่ 3.12 ผลลัพธ์ของระบบแนะนำภาพยนตร์โดยใช้อัลกอริทึมการคัดกรองผู้ใช้ร่วม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
HOTEL Recommendation Systems

==================================================
ITEM-based Collaborative Filtering
==================================================
Meriton Serviced Apartments Campbell Street, Sydney
Four Seasons Hotel, Sydney
Marriott Sydney Harbour at Circular Quay, Sydney
InterContinental, Sydney
The Langham, Melbourne
Oaks On William, Melbourne
Pullman Quay Grand Sydney Harbour, Sydney
Sir Stamford at Circular Quay Hotel, Sydney
Hilton Melbourne South Wharf, Melbourne
Crown Towers, Melbourne
==================================================
```

รูปที่ 3.13 ผลลัพธ์ของระบบแนะนำโรงแรมโดยใช้อัลกอริทึมการคัดกรองสิ่งของร่วม

```
HOTELs Recommendation Systems

==================================================
USER-based Collaborative Filtering
==================================================
Meriton Serviced Apartments Campbell Street, Sydney
Four Seasons Hotel, Sydney
Marriott Sydney Harbour at Circular Quay, Sydney
InterContinental, Sydney
The Langham, Melbourne
Ibis Styles Melbourne, The Victoria Hotel, Melbourne
Crown Metropol Melbourne, Melbourne
The Sebel Melbourne Docklands, Melbourne
Quest Docklands, Sydney
Pegasus Apart'Hotel, Sydney
==================================================
```

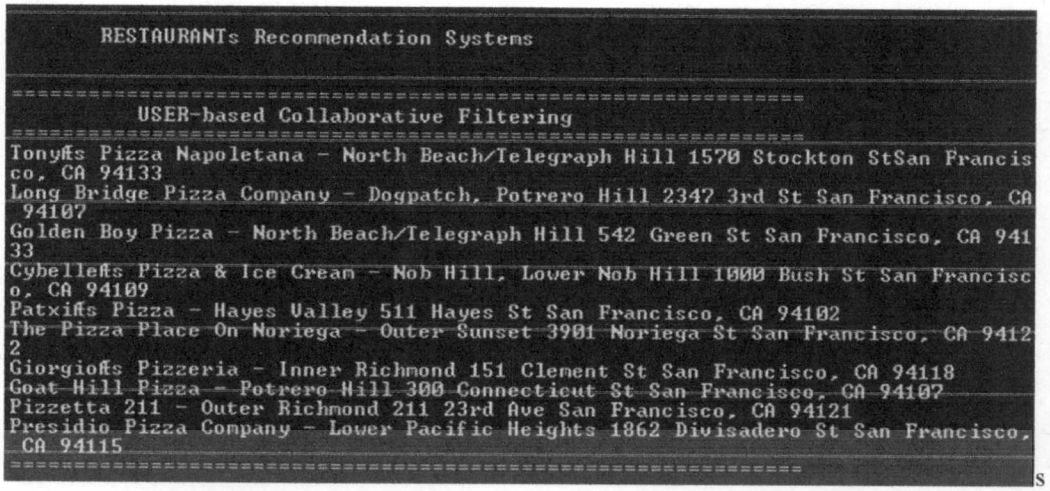
รูปที่ 3.14 ผลลัพธ์ของระบบแนะนำโรงแรมโดยใช้อัลกอริทึมการคัดกรองผู้ใช้ร่วม

```
RESTAURANTs Recommendation Systems

============================================================
ITEM-based Collaborative Filtering
============================================================
Mustafio's Pizza - Lower Nob Hill 1116 Polk St San Francisco, CA 94109
Tony's Pizza Napoletana - North Beach/Telegraph Hill 1570 Stockton StSan Francis
co, CA 94133
Little Star Pizza - Alamo Square 846 Divisadero St San Francisco, CA 94117
Long Bridge Pizza Company - Dogpatch, Potrero Hill 2347 3rd St San Francisco, CA
 94107
Golden Boy Pizza - North Beach/Telegraph Hill 542 Green St San Francisco, CA 941
33
Gialina Pizzeria - Glen Park 2842 Diamond St San Francisco, CA 94131
Marcello's Pizza - Castro 420 Castro St San Francisco, CA 94114
Escape From New York Pizza - The Haight 1737 Haight St San Francisco, CA 94117
Zero Zero - SoMa 826 Folsom St San Francisco, CA 94107
The Pizza Shop - Mission 3104 24th St San Francisco, CA 94110
============================================================
```

รูปที่ 3.15 ผลลัพธ์ของระบบแนะนำร้านอาหารโดยใช้อัลกอริทึมการคัดกรองสิ่งของร่วม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.16 ผลลัพธ์ของระบบแนะนำร้านอาหารโดยใช้อัลกอริทึมการคัดกรองผู้ใช้ร่วม

เมื่อได้ผลลัพธ์ออกมาทุกระบบแล้ว ให้เก็บข้อมูลเหล่านี้ไว้เป็นชุดข้อมูลตั้งต้น ซึ่งชุดข้อมูลตั้งต้นจะประกอบไปด้วยข้อมูลทั้งหมด 9 ชุด เนื่องจากมีข้อมูลทั้งหมด 3 ประเภท คือ ข้อมูลการให้คะแนนความนิยมของภาพยนตร์ โรมแรมและร้านอาหาร และแต่ละประเภทยังแยกย่อยเป็น 3 ชุดข้อมูล คือ ข้อมูลขนาดเล็ก ข้อมูลขนาดกลาง และข้อมูลขนาดใหญ่

หลังจากนั้นให้เตรียมชุดข้อมูลทดสอบ โดยมีวิธีการ คือ นำข้อมูลทั้งหมด 9 ชุดดังกล่าวมาดัดแปลงโดยการแยกผู้ใช้ออกมาชุดละ 100 รายและลบค่าคะแนนความนิยมที่ผู้ใช้แต่ละรายได้ให้คะแนนภาพยนตร์ ร้านอาหาร และโรงแรมของแต่ละชุดไว้ออกมาเป็นอัตราส่วน 10%, 15%, 20%, 25% และ 30% ตามลำดับ ยกตัวอย่างเช่น ผู้ใช้รายที่ 1 ของชุดข้อมูลขนาดเล็กสำหรับการให้คะแนนภาพยนตร์ เคยให้คะแนนความนิยมภาพยนตร์ไว้ 100 เรื่อง ให้ลบค่าคะแนนความนิยมออกจำนวน 10, 15, 20, 25 และ 30 เรื่อง ตามลำดับ ดังนั้นเราจะได้ข้อมูลทั้งหมด 5 ชุดสำหรับชุดข้อมูลขนาดเล็กสำหรับการให้คะแนนภาพยนตร์ ดังนั้น เนื่องจากเรามีชุดข้อมูลตั้งต้นจำนวน 9 ชุด เมื่อนำมาลบค่าคะแนนความนิยมออกทั้งหมดเป็น 5 อัตราส่วน จึงทำให้เกิดชุดข้อมูลใหม่ทั้งหมด 45 ชุด หลังจากนั้นจึงส่งข้อมูลใหม่ทั้งหมด 45 ชุด ที่ได้หลังจากการแก้ไขแล้วเข้าไปประมวลผลในระบบการให้คำแนะนำที่ใช้อัลกอริทึมการคัดกรองสิ่งของร่วม เมื่อผลลัพธ์ออกมาให้บันทึกไว้เป็นชุดข้อมูลทดสอบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

ผลการทดลอง

ในบทนี้จะแสดงผลการเปรียบเทียบอัลกอริทึมของระบบการให้คำแนะนำที่ประมวลผลอยู่บนฮาดูประหว่างการคัดกรองสิ่งของร่วม และการคัดกรองผู้ใช้ร่วม โดยมีเกณฑ์การเปรียบเทียบคือ ความเร็วในการประมวลผล และความแม่นยำในการทำนายระหว่าง 2 อัลกอริทึม

4.1 ผลการเปรียบเทียบความเร็วในการประมวลผลระหว่างเทคนิคการคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม

4.1.1 ข้อมูลที่ใช้ในการทดลอง

4.1.1.1 ข้อมูลการให้คะแนนความนิยมของภาพยนตร์จากเว็บไซต์มูฟวี่เลนส์ แบ่งออกเป็น 3 กลุ่มข้อมูล คือ เล็ก กลาง ใหญ่ ดังนี้

I. ข้อมูลขนาดเล็ก ประกอบไปด้วยภาพยนตร์ 700 เรื่องและผู้ใช้จำนวน 1,000 ราย

II. ข้อมูลขนาดกลาง ประกอบไปด้วยภาพยนตร์ 10,000 เรื่องและผู้ใช้จำนวน 72,000 ราย

III. ข้อมูลขนาดใหญ่ ประกอบไปด้วยภาพยนตร์ 27,000 เรื่องและผู้ใช้จำนวน 230,000 ราย

4.1.1.2 ข้อมูลการให้คะแนนความนิยมของโรงแรมจากเว็บไซต์ทริปแอดไวเซอร์ แบ่งออกเป็น 3 กลุ่มข้อมูล คือ เล็ก กลาง ใหญ่ ดังนี้

IV. ข้อมูลขนาดเล็ก ประกอบไปด้วยโรงแรม 2,500 แห่งและผู้ใช้จำนวน 1,000 ราย

V. ข้อมูลขนาดกลาง ประกอบไปด้วยโรงแรม 30,000 แห่งและผู้ใช้จำนวน 216,000 ราย

VI. ข้อมูลขนาดใหญ่ ประกอบไปด้วยโรงแรม 81,000 แห่งและผู้ใช้จำนวน 690,000 ราย

4.1.1.3 ข้อมูลการให้คะแนนความนิยมของร้านอาหารจากเว็บไซต์เยลป์ แบ่งออกเป็น 3 กลุ่มข้อมูล คือ เล็ก กลาง ใหญ่ ดังนี้

VII. ข้อมูลขนาดเล็ก ประกอบไปด้วยร้านอาหาร 2,800 ร้านและผู้ใช้จำนวน 1,000 ราย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

VIII. ข้อมูลขนาดกลาง ประกอบไปด้วยร้านอาหาร 40,000 ร้านและผู้ใช้จำนวน 288,000 ราย

IX. ข้อมูลขนาดใหญ่ ประกอบไปด้วยร้านอาหาร 108,000 ร้านและผู้ใช้จำนวน 920,000 ราย

4.1.2 เปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำภาพยนตร์

เปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำภาพยนตร์ระหว่างเทคนิคการคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม ซึ่งแบ่งการทดลองเป็น 3 รอบใหญ่ คือ ชุดข้อมูลขนาดเล็ก ขนาดกลางและขนาดใหญ่ ภายในแต่ละรอบใหญ่นั้นมีการประมวลผลทั้งสิ้น 10 รอบย่อย เพื่อหาค่าเฉลี่ยให้กับรอบใหญ่ ดังนี้

4.1.2.1 ข้อมูลขนาดเล็ก ประกอบด้วยภาพยนตร์ 700 เรื่องและผู้ใช้ 1,000 ราย

ตารางที่ 4.1 ความเร็วในการประมวลผลของชุดข้อมูลขนาดเล็กสำหรับระบบแนะนำภาพยนตร์

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	9.12	3.45
2	9.56	4.12
3	10.17	5.05
4	10.43	3.49
5	10.03	4.28
6	9.58	4.59
7	11.12	4.13
8	10.49	3.36
9	10.24	5.01
10	10.39	4.27
เฉลี่ย	10.11	4.18

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้ ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 10 นาที 11 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 4 นาที 18 วินาที

4.1.2.2 ข้อมูลขนาดกลาง ประกอบไปด้วยภาพยนตร์ 10,000 เรื่องและผู้ใช้จำนวน 72,000 ราย

ตารางที่ 4.2 ความเร็วในการประมวลผลของชุดข้อมูลขนาดกลางสำหรับระบบแนะนำภาพยนตร์

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	107.12	41.59
2	110.45	39.21
3	107.31	42.34
4	103.35	38.15
5	98.12	48.43
6	112.39	42.35
7	110.01	44.12
8	111.45	40.03
9	110.50	39.58
10	111.12	45.34
เฉลี่ย	108.18	42.11

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 108 นาที 18 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 42 นาที 11 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.2.3 ข้อมูลขนาดใหญ่ ประกอบไปด้วยภาพยนตร์ 27,000 เรื่องและผู้ใช้จำนวน 230,000 ราย

ตารางที่ 4.3 ความเร็วในการประมวลผลของชุดข้อมูลขนาดใหญ่สำหรับระบบแนะนำภาพยนตร์

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	845.55	121.02
2	911.02	130.13
3	889.14	127.57
4	851.43	122.54
5	831.32	119.46
6	893.28	128.01
7	900.19	129.35
8	882.34	126.15
9	918.09	131.07
10	893.17	128.59
เฉลี่ย	881.15	126.39

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 881 นาที 15 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 126 นาที 39 วินาที

4.1.3 เปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำโรงแรม

เปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำโรงแรมระหว่างเทคนิคการคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม ซึ่งแบ่งการทดลองเป็น 3 รอบใหญ่ คือ ชุดข้อมูลขนาดเล็ก ขนาดกลางและขนาดใหญ่ ภายในแต่ละรอบใหญ่นั้นมีการประมวลผลทั้งสิ้น 10 รอบย่อย เพื่อหาค่าเฉลี่ยให้กับรอบใหญ่ ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.3.1 ข้อมูลขนาดเล็ก ประกอบไปด้วยโรงแรม 2,500 แห่งและผู้ใช้จำนวน 1,000 ราย

ตารางที่ 4.4 ความเร็วในการประมวลผลของชุดข้อมูลขนาดเล็กสำหรับระบบแนะนำโรงแรม

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	10.45	15.12
2	11.56	17.04
3	9.17	16.58
4	11.03	23.03
5	10.58	16.48
6	12.24	20.08
7	11.12	15.29
8	10.49	14.49
9	10.24	19.08
10	10.39	14.09
เฉลี่ย	10.27	17.13

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 10 นาที 27 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 17 นาที 13 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.3.2 ข้อมูลขนาดกลาง ประกอบไปด้วยโรงแรม 30,000 แห่งและผู้ใช้จำนวน 216,000 ราย

ตารางที่ 4.5 ความเร็วในการประมวลผลของชุดข้อมูลขนาดกลางสำหรับระบบแนะนำโรงแรม

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	745.25	221.52
2	811.02	230.13
3	789.17	227.17
4	751.43	222.54
5	731.38	219.36
6	793.28	228.01
7	800.49	229.45
8	782.34	226.15
9	818.59	231.27
10	793.37	228.59
เฉลี่ย	781.59	226.42

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 781 นาที 59 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 226 นาที 42 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.3.3 ข้อมูลขนาดใหญ่ ประกอบไปด้วยโรงแรม 81,000 แห่งและผู้ใช้จำนวน 690,000 ราย

ตารางที่ 4.6 ความเร็วในการประมวลผลของชุดข้อมูลขนาดใหญ่สำหรับระบบแนะนำโรงแรม

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	2,400.15	656.11
2	2,408.58	649.43
3	2,398.15	660.58
4	2,424.13	653.32
5	2,401.21	655.18
6	2,410.43	640.51
7	2,456.10	668.13
8	2,387.27	679.05
9	2,427.01	663.18
10	2,480.12	649.04
เฉลี่ย	2,419.32	657.45

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 2,419 นาที 32 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ นาที 657 นาที 45 วินาที

4.1.4 เปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำร้านอาหาร

เปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำร้านอาหารระหว่างเทคนิคการคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม ซึ่งแบ่งการทดลองเป็น 3 รอบใหญ่ คือ ชุดข้อมูลขนาดเล็ก ขนาดกลางและขนาดใหญ่ ภายในแต่ละรอบใหญ่นั้นมีการประมวลผลทั้งสิ้น 10 รอบย่อย เพื่อหาค่าเฉลี่ยให้กับรอบใหญ่ ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.4.1 ข้อมูลขนาดเล็ก ประกอบไปด้วยร้านอาหาร 2,800 ร้านและผู้ใช้จำนวน 1,000 ราย

ตารางที่ 4.7 ความเร็วในการประมวลผลของชุดข้อมูลขนาดเล็กสำหรับระบบแนะนำร้านอาหาร

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	12.24	20.24
2	11.12	17.04
3	10.49	16.58
4	12.24	23.03
5	10.39	16.48
6	10.45	20.08
7	11.56	15.29
8	9.17	24.58
9	11.03	19.08
10	10.58	24.09
เฉลี่ย	10.53	19.55

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 10 นาที 53 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 19 นาที 55 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.4.2 ข้อมูลขนาดกลาง ประกอบไปด้วยร้านอาหาร 40,000 ร้านและผู้ใช้จำนวน 288,000 ราย

ตารางที่ 4.8 ความเร็วในการประมวลผลของชุดข้อมูลขนาดกลางสำหรับระบบแนะนำร้านอาหาร

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	831.38	319.36
2	893.28	328.01
3	900.49	329.45
4	899.34	342.17
5	882.34	326.15
6	918.59	331.27
7	893.37	328.59
8	845.25	321.52
9	911.02	330.13
10	889.17	327.17
เฉลี่ย	886.42	886.2

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 886 นาที 42 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 328 นาที 38 วินาที

4.1.4.3 ข้อมูลขนาดใหญ่ ประกอบไปด้วยร้านอาหาร 108,000 ร้านและผู้ใช้จำนวน 920,000 ราย

ตารางที่ 4.9 ความเร็วในการประมวลผลของชุดข้อมูลขนาดใหญ่สำหรับระบบแนะนำร้านอาหาร

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
1	3,413.50	751.18

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.9 (ต่อ)

ครั้งที่	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
2	3,420.10	748.23
3	3,456.10	768.13
4	3,387.27	779.05
5	3,427.01	763.18
6	3,480.12	749.04
7	3,398.15	760.58
8	3,424.13	753.32
9	3,401.21	755.18
10	3,410.43	740.51
เฉลี่ย	3,421.58	756.54

จากการทดลองประมวลผลเป็นจำนวน 10 รอบ ด้วยข้อมูลชุดเดียวกันบนระบบเดียวกัน และอยู่บนระบบคลาวด์ (Cloud systems) เดียวกัน เมื่อหาค่าเฉลี่ยแล้วได้ผลดังนี้

ผลเฉลี่ยของ User-based Collaborative Filtering เท่ากับ 3,421 นาที 58 วินาที

ผลเฉลี่ยของ Item-based Collaborative Filtering เท่ากับ 756 นาที 54 วินาที

4.1.5 สรุปผลการเปรียบเทียบความเร็วของทุกระบบ

ผลการเปรียบเทียบความเร็วในการประมวลผลของระบบแนะนำภาพยนตร์ ระบบแนะนำโรงแรมและระบบแนะนำร้านอาหาร แสดงดังตารางที่ 4.10 ดังนี้

ตารางที่ 4.10 ผลการเปรียบเทียบความเร็วในการประมวลผลของทุกระบบ

ระบบการให้คำแนะนำ	กลุ่มข้อมูล	User-based CF (นาที.วินาที)	Item-based CF (นาที.วินาที)
ระบบแนะนำภาพยนตร์	กลุ่มข้อมูลขนาดเล็ก	10.11	4.18
	กลุ่มข้อมูลขนาดกลาง	108.18	42.11
	กลุ่มข้อมูลขนาดใหญ่	881.15	126.39
ระบบแนะนำโรงแรม	กลุ่มข้อมูลขนาดเล็ก	10.27	17.13

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.10 (ต่อ)

	กลุ่มข้อมูลขนาดกลาง	781.59	226.42
	กลุ่มข้อมูลขนาดใหญ่	2,419.32	657.45
ระบบแนะนำร้านอาหาร	กลุ่มข้อมูลขนาดเล็ก	10.53	19.55
	กลุ่มข้อมูลขนาดกลาง	886.42	328.38
	กลุ่มข้อมูลขนาดใหญ่	3,421.58	756.54

ผลการเปรียบเทียบพบว่าความเร็วที่ใช้ในการประมวลผล พบว่า การคัดกรองผู้ใช้ร่วมจะใช้เวลาในการประมวลมากกว่าการคัดกรองสิ่งของร่วม ถ้าหากข้อมูลที่ใช้มีจำนวนผู้ใช้มากกว่าจำนวนของสิ่งของ และการคัดกรองสิ่งของร่วมจะใช้เวลาในการประมวลมากกว่าการคัดกรองผู้ใช้ร่วม ถ้าหากข้อมูลที่ใช้มีจำนวนสิ่งของมากกว่าจำนวนของผู้ใช้

โดยทั่วไปธรรมชาติของข้อมูลเกี่ยวกับการให้คะแนนความนิยมนั้น จะมีจำนวนผู้ใช้มากกว่าจำนวนสิ่งของ ดังนั้น จึงได้ข้อสรุปว่าระบบการให้คำแนะนำบนฮาดูปควรใช้อัลกอริทึมการคัดกรองสิ่งของร่วม

4.2 ผลการเปรียบเทียบความแม่นยำในการพยากรณ์ระหว่างการคัดกรองผู้ใช้ร่วมและการคัดกรองสิ่งของร่วม

4.2.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองครั้งนี้เป็นข้อมูลชุดเดียวกับข้อมูลที่ใช้ในหัวข้อ 4.1

4.2.2 วิธีการเปรียบเทียบ

วิธีการเปรียบเทียบความแม่นยำของการพยากรณ์ระหว่างเทคนิคการคัดกรองสิ่งของร่วมและการคัดกรองผู้ใช้ร่วมมีขั้นตอนดังนี้

1) นำข้อมูลทั้ง 9 ชุดเข้าไปประมวลผลตามปกติ รอจนได้ผลลัพธ์การแนะนำและบันทึกผลลัพธ์ชุดนี้ไว้ เป็นชุดข้อมูลตั้งต้น

2) แก้ไขข้อมูลในข้อมูลทั้ง 9 ชุดนั้น โดยการแยกผู้ใช้ออกมาชุดละ 100 รายและลบค่าคะแนนความนิยมที่ผู้ใช้แต่ละรายได้ให้คะแนนภาพยนตร์ ร้านอาหาร และโรงแรมของแต่ละชุดไว้ออกมาเป็นอัตราส่วน 10%, 15%, 20%, 25% และ 30% ตามลำดับ ยกตัวอย่างเช่น ผู้ใช้รายที่ 1 ของชุดข้อมูลขนาดเล็กสำหรับการให้คะแนนภาพยนตร์ เคยให้คะแนนความนิยมภาพยนตร์ไว้ 100 เรื่อง ให้ลบค่าคะแนนความนิยมออกจำนวน 10, 15,

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

20, 25 และ 30 เรื่อง ตามลำดับ ดังนั้นเราจะได้ข้อมูลทั้งหมด 5 ชุดสำหรับชุดข้อมูลขนาดเล็กสำหรับการให้คะแนนภาพยนตร์ ดังนั้น เนื่องจากเรามีชุดข้อมูลตั้งต้นจำนวน 9 ชุด เมื่อนำมาลบค่าคะแนนความนิยมออกทั้งหมดเป็น 5 อัตราส่วน จึงทำให้เกิดชุดข้อมูลใหม่ทั้งหมด 45 ชุด

3) ส่งข้อมูลใหม่ทั้งหมด 45 ชุด ที่ได้หลังจากการแก้ไขแล้วเข้าไปประมวลผลในระบบการให้คำแนะนำที่ใช้อัลกอริทึมการคัดกรองสิ่งของร่วม เมื่อผลลัพธ์ออกมาให้บันทึกไว้เป็นชุดข้อมูลทดสอบ

4) นำชุดข้อมูลทดสอบไปเปรียบเทียบกับชุดข้อมูลตั้งต้น เพื่อดูว่าสิ่งที่พยากรณ์ออกมาได้นั้น แตกต่างจากชุดข้อมูลตั้งต้นเท่าไร โดยวัดเป็นอัตราส่วนความถูกต้อง

5) นำข้อมูล 45 ชุดมาส่งเข้าไปประมวลผลในระบบการให้คำแนะนำที่ใช้อัลกอริทึมการคัดกรองผู้ใช้ร่วม เมื่อผลลัพธ์ออกมาให้บันทึกไว้เป็นชุดข้อมูลทดสอบ

6) หลังจากนั้นจึงนำผลลัพธ์ที่อยู่ในรูปแบบของอัตราส่วนความถูกต้องที่วัดเป็นเปอร์เซ็นต์มาเปรียบเทียบกันระหว่างการคัดกรองสิ่งของร่วมและการคัดกรองผู้ใช้ร่วมว่าอัลกอริทึมใดมีความแม่นยำในการพยากรณ์มากกว่ากัน

7) บักทึกผลการทดสอบ วิเคราะห์ผลการเปรียบเทียบและสรุปผลการทดลอง

4.2.3 ผลลัพธ์การเปรียบเทียบความแม่นยำในการพยากรณ์

ผลการวัดความแม่นยำในการพยากรณ์ได้มาจากการเปรียบเทียบผลลัพธ์การให้คำแนะนำระหว่างชุดข้อมูลตั้งต้นและชุดข้อมูลทดสอบว่ามีอัตราส่วนความถูกต้องมากน้อยเพียงใด

ตารางที่ 4.11 ผลลัพธ์การเปรียบเทียบชุดข้อมูลตั้งต้นและชุดข้อมูลทดสอบ

อัตราส่วนที่ดึงข้อมูลการให้คะแนนของผู้ใช้ออก		อัตราอัตราส่วนความถูกต้องระหว่างชุดข้อมูลตั้งต้นและชุดข้อมูลทดสอบ		ความแตกต่างของผลลัพธ์
		User-based Collaborative Filtering	Item-based Collaborative Filtering	
ระบบแนะนำภาพยนตร์	10 %	80.70 %	83.30 %	2.60 %
	15 %	73.90 %	77.80 %	3.90 %
	20 %	61.30 %	65.70 %	4.40 %

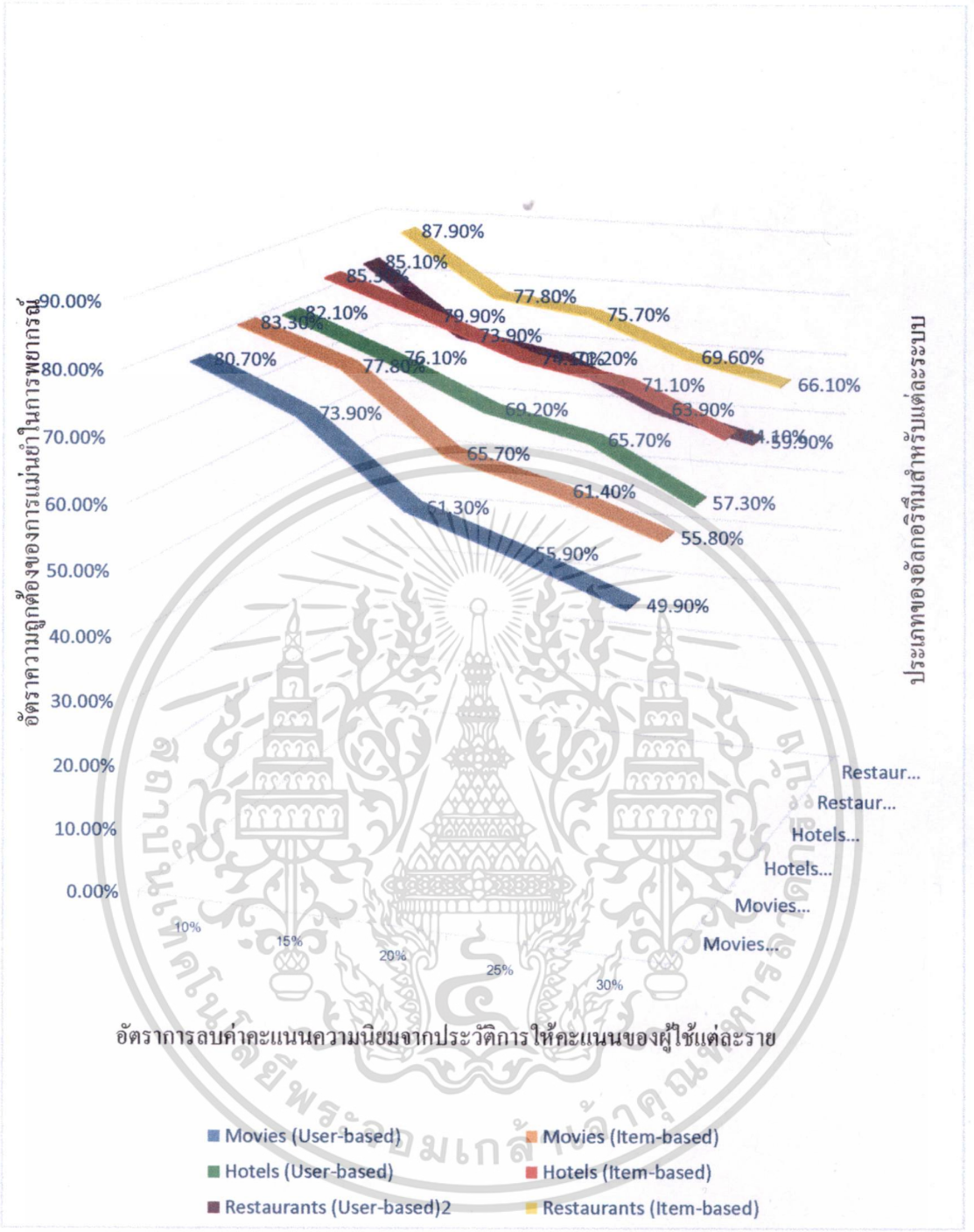
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.11 (ต่อ)

อัตราส่วนที่ดึงข้อมูลการให้คะแนนของผู้ใช้ออก		อัตราอัตราส่วนความถูกต้องระหว่างชุดข้อมูลตั้งต้นและชุดข้อมูลทดสอบ		ความแตกต่างของผลลัพธ์
		User-based Collaborative Filtering	Item-based Collaborative Filtering	
ระบบแนะนำภาพยนตร์	25 %	55.90 %	61.40 %	5.50 %
	30 %	49.90 %	55.80 %	5.90 %
ระบบแนะนำโรงแรม	10 %	82.10 %	85.30 %	3.20 %
	15 %	76.10 %	79.90 %	3.80 %
	20 %	69.20 %	74.10 %	4.90 %
	25 %	65.70 %	71.10 %	5.40 %
	30 %	57.30 %	64.10 %	6.80 %
ระบบแนะนำร้านอาหาร	10 %	85.10 %	87.90 %	2.80 %
	15 %	73.90 %	77.80 %	3.90 %
	20 %	71.20 %	75.70 %	4.50 %
	25 %	63.90 %	69.60 %	5.70 %
	30 %	59.90 %	66.10 %	6.20 %

ผลลัพธ์ที่ได้จากการเปรียบเทียบความแม่นยำในการทำนายระหว่างชุดข้อมูลตั้งต้นและชุดข้อมูลทดสอบพบว่า อัลกอริทึมการคัดกรองสิ่งของร่วมจะมีอัตราความถูกต้องมากกว่าการคัดกรองผู้ใช้ร่วม และจากการทดลองโดยเพิ่มอัตราส่วนที่ดึงข้อมูลการให้คะแนนของผู้ใช้ออกมากขึ้นก็พบว่า ความแตกต่างของผลลัพธ์ระหว่าง 2 อัลกอริทึมมีค่ามากขึ้นตามลำดับ ดังที่แสดงในตารางที่ 4.11 และแสดงผลลัพธ์ในรูปแบบของกราฟ ดังรูปที่ 4.1 ต่อไปนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.1 กราฟผลลัพธ์อัตราส่วนความแม่นยำในการพยากรณ์ของแต่ละอัลกอริทึม

จากกราฟจะเห็นว่ายิ่งมีอัตราการดึงข้อมูลการให้คะแนนของผู้ใช้ออกมากเท่าไร การพยากรณ์ก็จะมีความแม่นยำน้อยลงตามลำดับ ดังนั้นลักษณะของกราฟจึงดิ่งลงตามอัตราการดึงข้อมูลการให้คะแนนของผู้ใช้ออกที่มากขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3 สรุป

ผลลัพธ์ที่ได้จากการเปรียบเทียบความแม่นยำในการทำนายระหว่างชุดข้อมูลตั้งต้นและชุดข้อมูลทดสอบพบว่ามีความใกล้เคียงกัน โดยที่เทคนิคการคัดกรองสิ่งของร่วมจะมีอัตราความถูกต้องมากกว่าการคัดกรองผู้ใช้ร่วมเล็กน้อย ส่วนในเรื่องของความเร็วในการประมวลผลนั้น การคัดกรองผู้ใช้ร่วมจะใช้เวลาในการประมวลมากกว่าการคัดกรองสิ่งของร่วม ถ้าหากข้อมูลที่ใช้มีจำนวนผู้ใช้มากกว่าจำนวนของสิ่งของ และการคัดกรองสิ่งของร่วมจะใช้เวลาในการประมวลมากกว่าการคัดกรองผู้ใช้ร่วม ถ้าหากข้อมูลที่ใช้มีจำนวนสิ่งของมากกว่าจำนวนของผู้ใช้

ดังนั้นข้อสรุปของการศึกษานี้ คือ ระบบการให้คำแนะนำบนฮาดูปควรใช้อัลกอริทึมการคัดกรองสิ่งของร่วม เนื่องจากโดยทั่วไปข้อมูลเกี่ยวกับการให้คะแนนความนิยมนั้น จะมีจำนวนผู้ใช้มากกว่าจำนวนสิ่งของ ดังนั้น การใช้อัลกอริทึมการคัดกรองสิ่งของร่วมจึงเหมาะสมเพราะใช้เวลาประมวลผลได้เร็วกว่า ในขณะที่อัตราความถูกต้องในการพยากรณ์ก็มีมากกว่าเช่นกัน

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
บทสรุป

5.1 สรุปผลการศึกษา

ในการศึกษานี้ได้สร้างระบบการให้คำแนะนำขึ้นมา 3 ระบบ คือ ระบบแนะนำภาพยนตร์ ระบบแนะนำโรงแรม และระบบแนะนำร้านอาหาร และเปรียบเทียบอัลกอริทึมของระบบการให้คำแนะนำ ซึ่งพบว่า อัลกอริทึมการคัดกรองสิ่งของร่วมให้ผลลัพธ์ที่ดีกว่าการคัดกรองผู้ใช้ร่วม ถ้าหากข้อมูลที่ใช้มีจำนวนผู้ใช้มากกว่าจำนวนของสิ่งของ และยังมีอัตราในการพยากรณ์ที่แม่นยำกว่า ซึ่งได้ทดลองกับชุดข้อมูลหลายขนาดแตกต่างกัน ตั้งแต่ชุดข้อมูลขนาดเล็กที่มีขนาดของผู้ใช้และสิ่งของเป็นจำนวนหลักพันไปจนถึงชุดข้อมูลขนาดใหญ่ที่มีจำนวนถึงหลักแสนและได้ทดสอบกับชุดข้อมูลที่มีอัตราการดึงข้อมูลออกเพื่อทดสอบในอัตราส่วนที่แตกต่างกันเป็นอัตรา 10%, 15%, 20%, 25% และ 30% ตามลำดับ ซึ่งก็ได้ผลลัพธ์ไปในทางเดียวกัน คือ การคัดกรองสิ่งของร่วมมีอัตราการพยากรณ์ที่แม่นยำกว่า โดยทั่วไปธรรมชาติของข้อมูลเกี่ยวกับการให้คะแนนความนิยมนั้นจะมีจำนวนผู้ใช้มากกว่าจำนวนสิ่งของ เนื่องจากการเพิ่มขึ้นของผู้ใช้เกิดขึ้นได้ง่ายกกว่าการเพิ่มขึ้นของสิ่งของ ทำให้จำนวนของผู้ใช้มักจะมากกว่าจำนวนของสิ่งของเสมอ ดังนั้น ข้อสรุปของการศึกษานี้ คือ ระบบการให้คำแนะนำบนฮาดูปควรใช้อัลกอริทึมการคัดกรองสิ่งของร่วม

5.2 ปัญหาและอุปสรรคในการทดลอง

ในการศึกษาและทดลองนี้เน้นเรื่องของขนาดข้อมูลเป็นอย่างมาก เพราะต้องการประมวลผลข้อมูลขนาดใหญ่บนระบบการให้คำแนะนำ นั่นหมายความว่า ต้องมีทรัพยากรที่ใช้ในการทดลองเพียงพอที่จะรองรับข้อมูลขนาดใหญ่ได้ มิฉะนั้นแล้วการประมวลผลอาจจะไม่สำเร็จ การติดตั้งระบบเพื่อทำการทดสอบจึงเป็นเรื่องที่สำคัญและต้องหาวิธีในการหลีกเลี่ยงค่าใช้จ่ายที่ไม่จำเป็น เช่น การทดลองบนคลาวด์ จะมีการคิดค่าใช้จ่ายตามจริง ใช้เท่าไร จ่ายเท่านั้น ดังนั้น การทดลองบนคลาวด์อาจจะเสียค่าใช้จ่ายได้มากเช่นกัน นอกจากว่าเป็นการวิจัยที่มีงบลงทุน เช่น มีบริษัทพร้อมจะลงทุนการประมวลผลข้อมูลขนาดใหญ่เพื่อผลประโยชน์ทางธุรกิจ ลักษณะนี้จะสามารถลงทุนได้เพราะจะมีผลตอบแทนในอนาคตแน่นอน

5.3 ข้อเสนอแนะ

ผู้ที่จะมาศึกษาหรือวิจัยงานนี้ต่อ นอกจากควรจะเข้าใจในหลักการทำงานของแมปรีดิวซ์และระบบการให้คำแนะนำเป็นอย่างดีเพื่อที่จะเขียนโปรแกรมได้แล้วนั้น ยังต้องมองความสัมพันธ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในข้อมูลขนาดใหญ่ให้ออกและวิเคราะห์ให้แตกฉานว่า อะไรคือสิ่งที่ต้องการรู้และสิ่งที่ต้องการนำมาใช้วิเคราะห์คืออะไร จึงจะนำข้อมูลขนาดใหญ่ที่สนใจมาวิเคราะห์ได้อย่างมีประสิทธิภาพและได้คำตอบที่ต้องการ

เทคโนโลยีในสมัยนี้ก้าวกระโดดอย่างรวดเร็ว ระยะห่างของคาบเวลาที่มีการคิดค้นเทคโนโลยีใหม่สั้นลงเรื่อย ๆ อีกไม่นานก็จะมีเทคโนโลยีใหม่ที่ดีกว่าและมีประสิทธิภาพมากกว่าตามมา เทคโนโลยีฮาดูปและมาเฮาท์อาจจะล้าสมัยไปแล้วก็เป็นได้ ซึ่งเป็นเรื่องที่ไม่อาจคาดคะเนได้ ดังนั้น การติดตามเทคโนโลยีใหม่อย่างเสมอจะทำให้การศึกษาและวิจัยน่าสนใจและมีประโยชน์มากขึ้นเรื่อย ๆ

5.4 สรุปโครงงาน

ความท้าทายและความน่าค้นหาของการศึกษาอิสระฉบับนี้ คือ การได้ใช้เทคโนโลยีที่ก้าวไปทันยุคสมัย เช่น การประมวลผลข้อมูลขนาดใหญ่โดยใช้ฮาดูป ปัจจุบันโลกของเราถือเป็นยุคแห่งข้อมูล ซึ่งเป็นยุคถัดมาจากยุคอุตสาหกรรมและยุคเกษตรกรรม ในเมื่อยุคนี้ได้ขึ้นชื่อว่าเป็น ยุคแห่งข้อมูล แน่นอนอยู่แล้วว่าจะต้องมีจำนวนปริมาณข้อมูลอย่างมหาศาล ย้อนไปในสมัยยุคเกษตรกรรม สิ่งที่มีค่าที่สุด คือ ที่ดิน ใครได้ครอบครองที่ดินมากก็กลายเป็นผู้มั่งคั่งหรือเป็นชนชั้นบรรดาศักดิ์ ไม่ต่างกับในสมัยนี้ที่ “ข้อมูลขนาดใหญ่” ถือเป็นสิ่งมีค่า ใครที่เป็นผู้ครองครองถือว่าเป็นกลุ่มที่เป็นผู้นำ เพราะเมื่อมีข้อมูลในมือมาก ย่อมวิเคราะห์คาดการณ์ทั้งปัจจุบันและอนาคตได้อย่างแม่นยำ ถ้าหากรู้เหตุการณ์ที่จะเกิดขึ้นในอนาคตแล้ว ก็ไม่ยากที่จะก้าวนำได้ก่อนผู้อื่น ในการศึกษาและวิจัยนี้จึงได้นำเครื่องมือฮาดูป ที่สามารถวิเคราะห์ข้อมูลขนาดใหญ่มาทดลองและสร้างระบบที่มีความแม่นยำในการแนะนำรายการสินค้าให้กับผู้ใช้ การยิ่งมีข้อมูลจำนวนมากและมีเครื่องมือการวิเคราะห์ที่มีประสิทธิภาพ การคาดการณ์ก็จะยิ่งแม่นยำมากขึ้น ดังนั้น การศึกษานี้จึงเป็นประโยชน์และสามารถตอบโจทย์ผู้ที่ต้องการเป็นผู้นำแห่งยุคได้อย่างมาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

[1] Takacs, G. Pilaszy, I. Nemeth, B. and Tikk, D. 2009. "Scalable Collaborative Filtering Approaches for Large Recommender Systems." **Journal of Machine Learning Research.** 10(1): 623-656.

[2] Pagare, R. and Patil, S. 2013. "Study of Collaborative Filtering Recommendation Algorithm - Scalability Issue." **International Journal of Computer Applications.** 67(25): 0975-8887

[3] GroupLens. 2015. **Movie lens datasets.** [Online]
Available: https://grouplens.org/datasets/movielens.

[4] TripAdvisor LLC. 2015. **Hotels Rating datasets.** [Online]
Available: http://www.tripadvisor.com.

[5] Yelp Inc. 2015. **Dataset Challenge Academic Dataset.** [Online]
Available: https://www.yelp.com/dataset_challenge/dataset.

[6] IBM Corporation 1994, 2015. **Big Data Technology.** [Online]
Available: http://www.ibm.com/big-data/us/en.

[7] IBM Corporation 1994, 2015. **Cloud Architecture.** [Online]
Available: https://www.ibm.com/developerworks/community/cloud-architecture.

[8] Srinath, P. and Thilina, G. 2013. **Hadoop MapReduce Cookbook.** Birmingham: Packt Publishing.

[9] Alex, H. 2012. **Hadoop in Practice.** New York, NY: Manning Publications.

[10] Boris, L. Kevin, T. Smith and Alexey, Y. 2013. **Professional Hadoop Solutions.** Indianapolis, IN: John Wiley & Sons, Inc.

[11] Donald, M. and Adam, S. 2012. **MapReduce Design Pattern.** Shelter Island, NY: O'Reilly Media, Inc.

[12] Sean, O. Robin, A. Ted, D. and Ellen, F. 2012. **Mahout in Action.** New York, NY: Manning Publications.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม (ต่อ)

[13] Konstan, J. Miller, B. Maltz, D. Herlocker, J. Gordon, L. and Riedl, J. 1997. "GroupLens: Applying Collaborative Filtering to Usenet News." **Communications of the ACM.** 40(3):77-87.

[14] Hofmann, T. 2004. "Latent semantic models for collaborative filtering." **ACM Trans. Info. Syst.** 22(1):89–115

[15] Deshpande, M. and Karypis, G. 2004. "Item-based top-n recommendation algorithms." **ACM Trans. Inf. Syst.**, 22(1):143–177

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

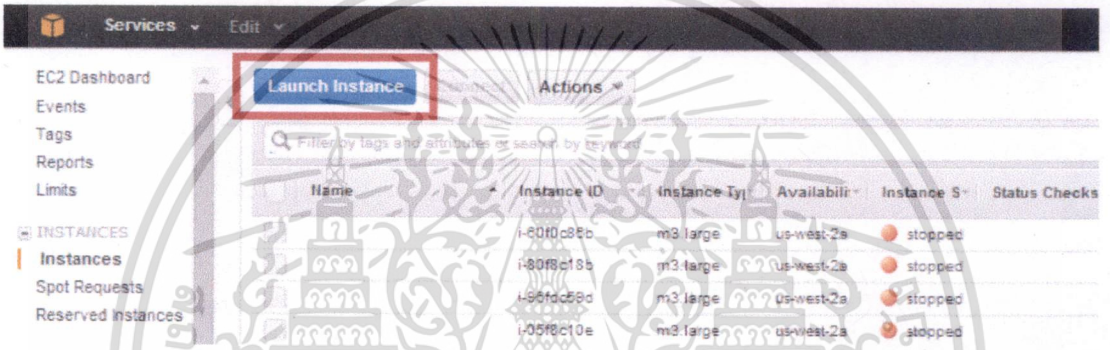
ภาคผนวก ก.

การติดตั้งฮาดูปบนเอมาซอนเว็บเซอร์วิสอีซีทู

ในภาคผนวก ก. นี้จะแสดงขั้นตอนการสร้างเซิร์ฟเวอร์เสมือน หรืออินสแตนซ์ บนเอมาซอนเว็บเซอร์วิสอีซีทู มีขั้นตอนดังนี้

1. ไปที่เว็บไซต์ https://console.aws.amazon.com/ec2/

2. คลิก Lunch Instance



รูปที่ ก.1 คลิกสร้างอินสแตนซ์

3. เลือกอิมเมจของเอมาซอนแมชชีน (AMI) เป็นอูบันตูเซิร์ฟเวอร์ เวอร์ชัน 14.04 ดังรูปที่ ก.2

รูปที่ ก.2 เลือกระบบปฏิบัติการของอินสแตนซ์

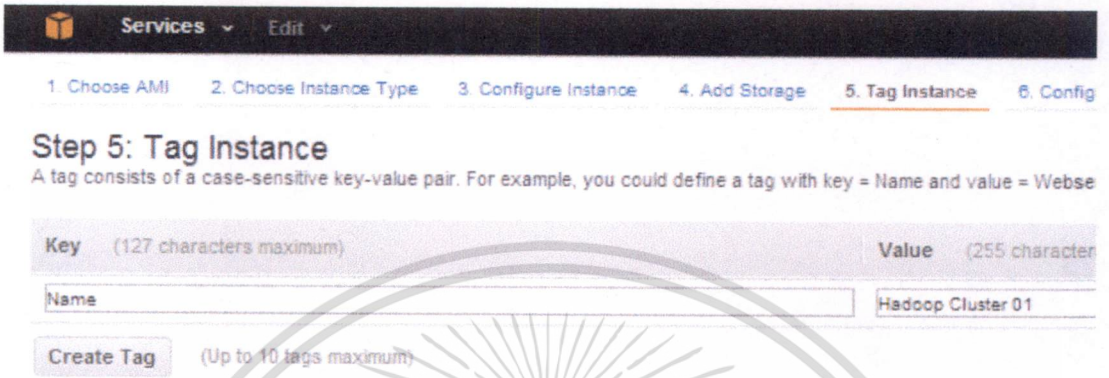
4. เลือกชนิดของอินสแตนซ์เป็น m3.large และคลิก Configure Instance Details

5. ในหน้าของ Configure Instance Details ให้เลือกการตั้งค่าโดยปริยาย (Default Settings) จากนั้นให้คลิกต่อไป (Next) Add Storage

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

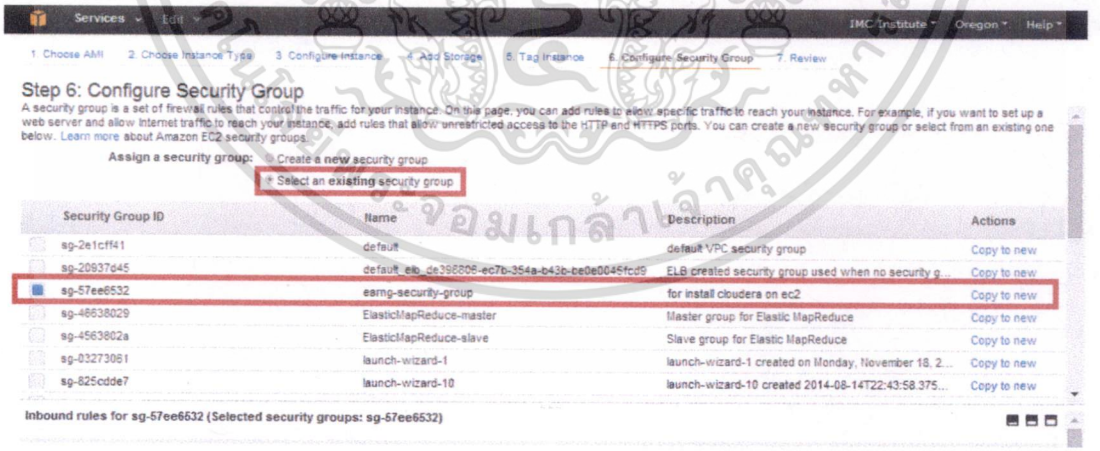
6. ในหน้าเพิ่มสตอเรจ (Add Storage) ให้เลือกการตั้งค่าโดยปริยาย จากนั้น ให้คลิกต่อไป (Next) Tag Instance

7. ในช่อง Tag Instance ใส่ชื่อว่า Hadoop Cluster 01 คลิกต่อไป (Next)



รูปที่ ก.3 ตั้งชื่อของอินสแตนซ์

8. ในหน้า Configure Security Group เลือก an existing security group เลือก Security Group Name: earng-security-group จากนั้นให้คลิก Review and Launch

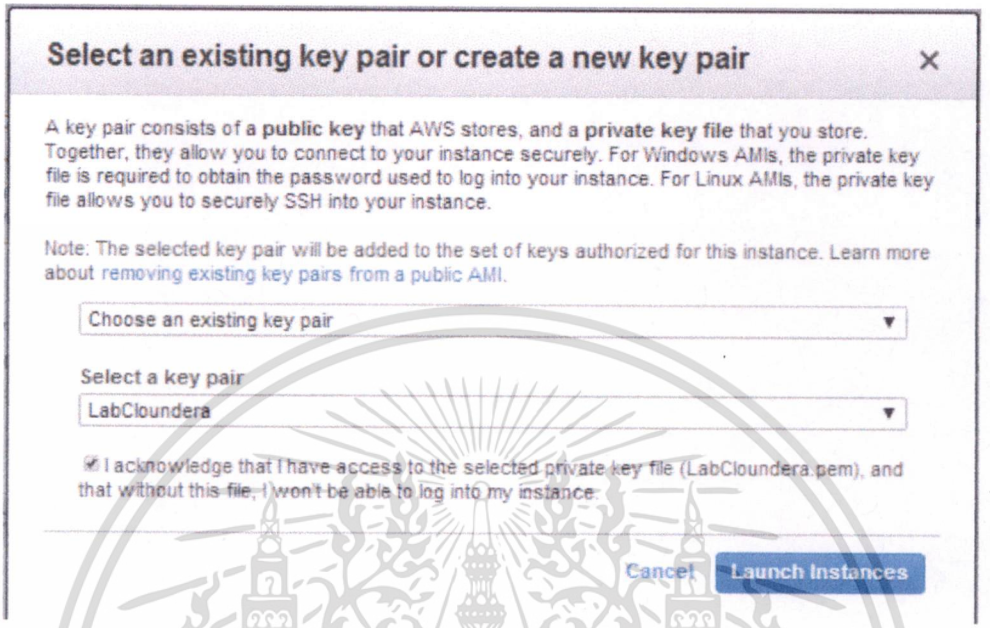


รูปที่ ก.4 กำหนดค่าของกลุ่มความปลอดภัย

9. ในหน้า Review Instance Launch ให้คลิก Click Launch

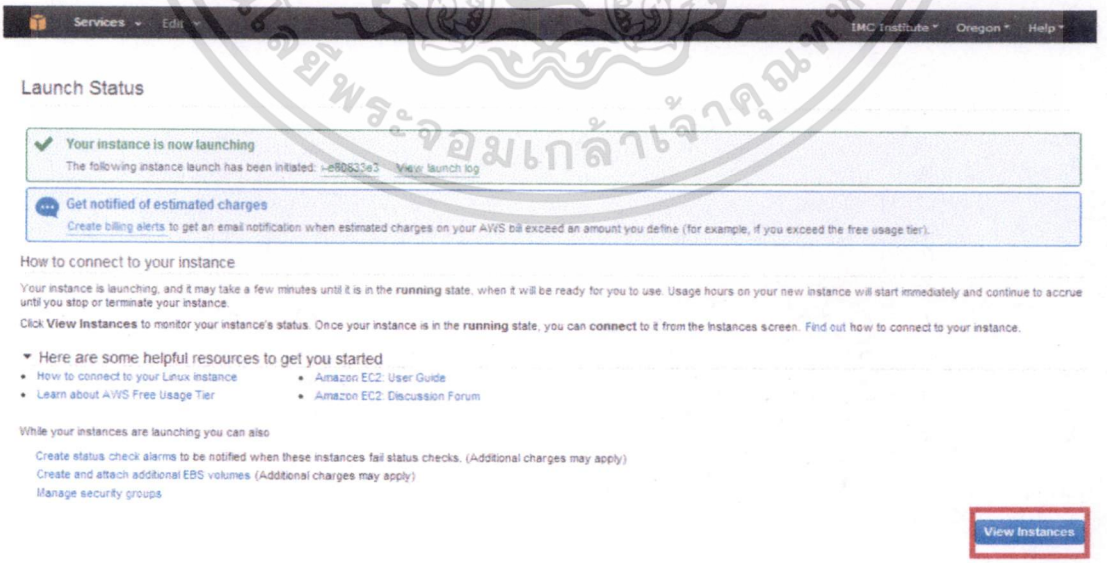
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

10. เลือก an existing key pair เลือก LabCloudera ทำเครื่องหมายถูกในช่อง acknowledge จากนั้นคลิก Launch Instances



รูปที่ ก.5 เลือกคีย์แพร์เพื่อใช้เป็นกุญแจในการเข้าถึงอินสแตนซ์

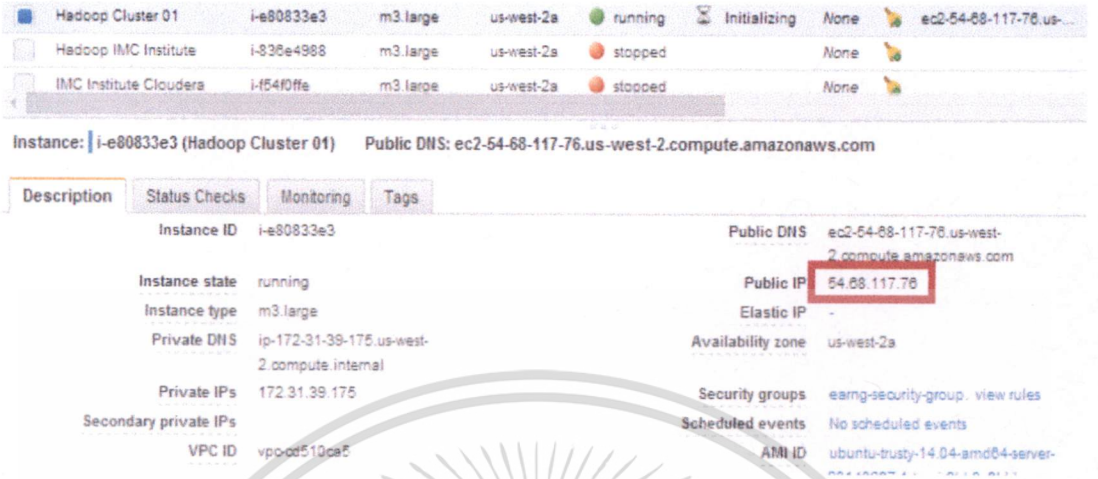
11. คลิก View Instances.



รูปที่ ก.6 ผลลัพธ์การสร้างอินสแตนซ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

12. เปิดเทอร์มินัล (Terminal) ไปยังไอพีสาธารณะที่ปรากฏอยู่ในรายละเอียดของอินสแตนซ์

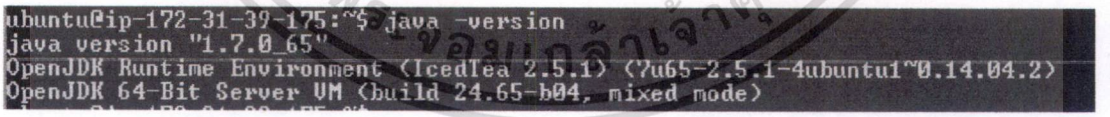


รูปที่ ก.7 เก็บค่าไอพีสาธารณะไว้เพื่อนำไปเข้าถึงอินสแตนซ์

13. เมื่อเข้ามาในหน้าเทอร์มินัลแล้ว ให้พิมพ์คำสั่งดังนี้

```
sudo apt-get update
ssh-keygen
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ssh 172.31.39.175 > Enter yes
exit
sudo apt-get install openjdk-7-jdk
```

14. พิมพ์คำสั่ง java –version เพื่อตรวจสอบเวอร์ชันของจาวา ดังรูปที่ ก.8 ดังภาพ



รูปที่ ก.8 ตรวจดูเวอร์ชันของจาวา

15. พิมพ์คำสั่งดังต่อไปนี้

```
wget http://mirror.issp.co.th/apache/hadoop/common/h
adoop-1.2.1/hadoop-1.2.1.tar.gz
tar -xvzf hadoop-1.2.1.tar.gz
sudo cp -r hadoop-1.2.1 /usr/local/hadoop
sudo vi $HOME/.bashrc
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

16. พิมพ์การตั้งค่าตามข้อความด้านล่างต่อไปนี้

```
export HADOOP_PREFIX=/usr/local/hadoop
export PATH=$PATH:$HADOOP_PREFIX/bin
```

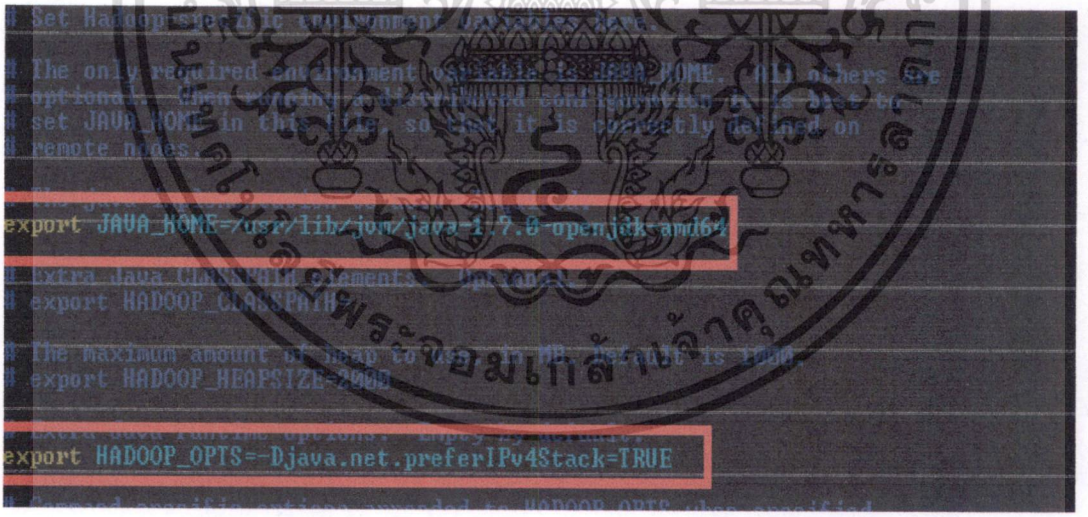
รูปที่ ก.9 การตั้งค่าในแบชไฟล์

17. พิมพ์คำสั่งดังต่อไปนี้

```
exec bash
sudo vi /usr/local/hadoop/conf/hadoop-env.sh
```

18. พิมพ์การตั้งค่าตามข้อความด้านล่างต่อไปนี้

```
export JAVA_HOME=/usr/lib/jvm/java-1.7.0-openjdk-
amd64
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=TRUE
```



รูปที่ ก.10 การตั้งค่าสิ่งแวดล้อมของฮาดูป

19. พิมพ์คำสั่งดังต่อไปนี้

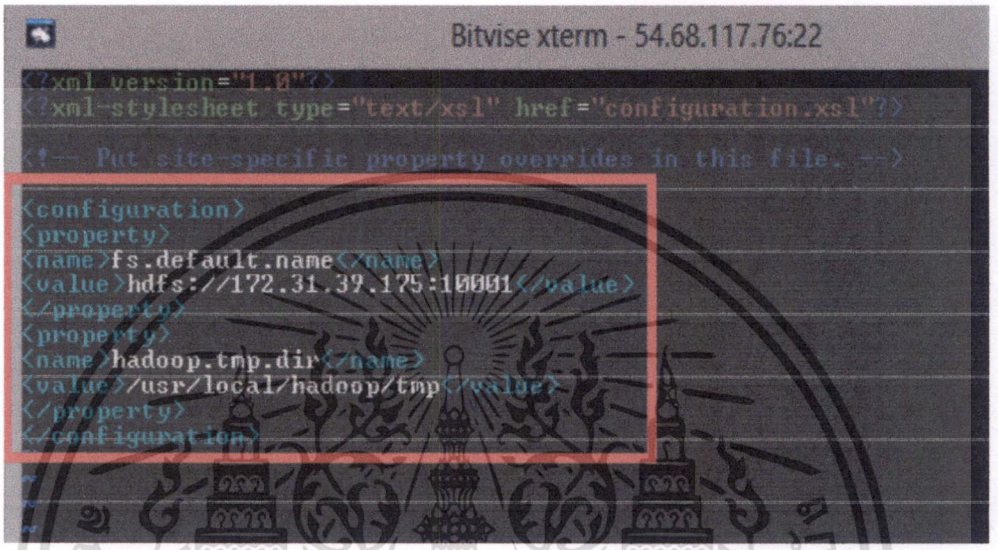
```
sudo vi /usr/local/hadoop/conf/core-site.xml
```

20. ใส่ไอพีส่วนตัวสำหรับเนมโหนด ตามตัวอย่างด้านล่างต่อไปนี้

```
<property>
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
<name>fs.default.name</name>
<value>hdfs://172.31.33.165:10001</value>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/usr/local/hadoop/tmp</value>
</property>
```



รูปที่ ก.11 การกำหนดค่าในคอนฟิกไฟล์ของฮาดูป

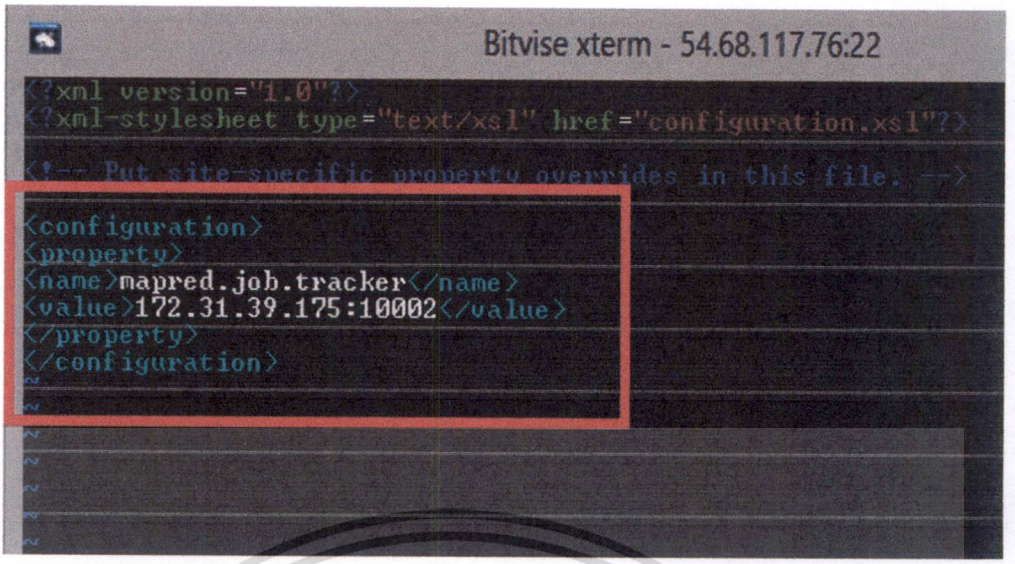
21. พิมพ์คำสั่งดังต่อไปนี้

```
sudo vi /usr/local/hadoop/conf/mapred-site.xml
```

22. ใส่ไอพีส่วนตัวสำหรับจ็อบแทร็กเกอร์ ตามตัวอย่างด้านล่างต่อไปนี้

```
<property>
<name>mapred.job.tracker</name>
<value>172.31.33.165:10002</value>
</property>
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.12 การตั้งค่าฮาดูปจ็อบแท็กเกอร์

23. พิมพ์คำสั่งดังต่อไปนี้

```
sudo vi /usr/local/hadoop/conf/hdfs-site.xml
```

24. พิมพ์การตั้งค่าตามข้อความด้านล่างต่อไปนี้

```
<property>
 <name>dfs.replication</name>
 <value>10</value>
</property>
```

25. พิมพ์คำสั่งดังต่อไปนี้

```
sudo mkdir /usr/local/hadoop/tmp
sudo chown ubuntu /usr/local/hadoop
sudo chown ubuntu /usr/local/hadoop/tmp
hadoop namenode -format
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
ubuntu@ip-172-31-39-175:~$ hadoop namenode -format
14/09/16 17:59:54 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = ip-172-31-39-175.us-west-2.compute.internal/172.31.39.175
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 1.2.1
STARTUP_MSG:   build = https://svn.apache.org/repos/asf/hadoop/common/branches/b
ranch-1.2 -r 1503152; compiled by 'mattf' on Mon Jul 22 15:23:09 PDT 2013
STARTUP_MSG:   java = 1.7.0_65
************************************************************/
14/09/16 17:59:54 INFO util.GSet: Computing capacity for map BlocksMap
14/09/16 17:59:54 INFO util.GSet: VM type       = 64-bit
14/09/16 17:59:54 INFO util.GSet: 2.0% max memory = 932184064
14/09/16 17:59:54 INFO util.GSet: capacity      = 2^21 = 2097152 entries
14/09/16 17:59:54 INFO util.GSet: recommended=2097152, actual=2097152
14/09/16 17:59:55 INFO namenode.FSNamesystem: fsOwner=ubuntu
14/09/16 17:59:55 INFO namenode.FSNamesystem: supergroup=supergroup
14/09/16 17:59:55 INFO namenode.FSNamesystem: isPermissionEnabled=true
14/09/16 17:59:55 INFO namenode.FSNamesystem: dfs.block.invalidate.limit=100
14/09/16 17:59:55 INFO namenode.FSNamesystem: isAccessTokenEnabled=false accessK
eyUpdateInterval=0 min(s), accessTokenLifetime=0 min(s)
14/09/16 17:59:55 INFO namenode.FSEditLog: dfs.namenode.edits.toleration.length
= 0
14/09/16 17:59:55 INFO namenode.NameNode: Caching file names occuring more than
10 times
14/09/16 17:59:55 INFO common.Storage: Image file /usr/local/hadoop/tmp/dfs/name
/current/fsimage of size 112 bytes saved in 0 seconds.
14/09/16 17:59:55 INFO namenode.FSEditLog: closing edit log: position=4, editlog
=/usr/local/hadoop/tmp/dfs/name/current/edits
14/09/16 17:59:55 INFO namenode.FSEditLog: close success: truncate to 4, editlog
=/usr/local/hadoop/tmp/dfs/name/current/edits
14/09/16 17:59:55 INFO common.Storage: Storage directory /usr/local/hadoop/tmp/d
fs/name has been successfully formatted.
14/09/16 17:59:55 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-39-175.us-west-2.compute.inter
nal/172.31.39.175
************************************************************/
ubuntu@ip-172-31-39-175:~$
```

รูปที่ ก.13 ล็อกแสดงผลการติดตั้งฮาดูปสำเร็จ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

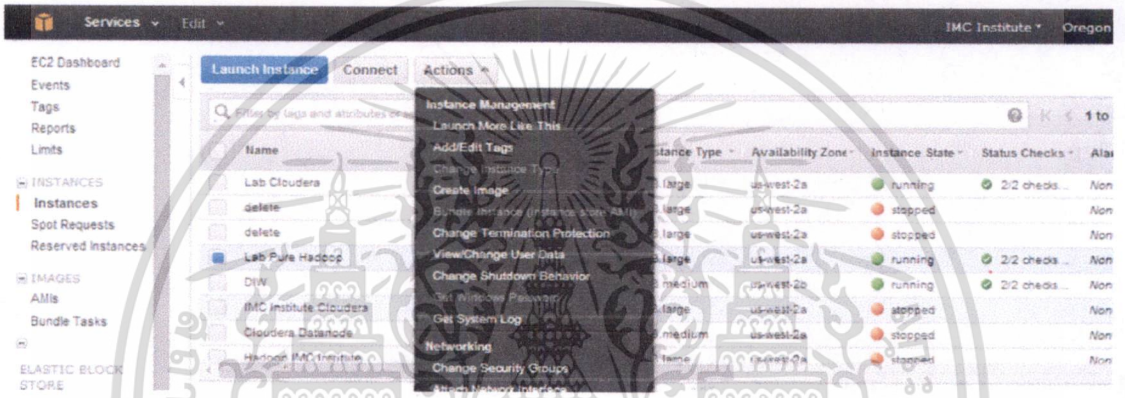
ภาคผนวก ข.

วิธีการสร้างฮาดูปคลัสเตอร์เอมาซอนเว็บเซอร์วิสอีซีทู

ในภาคผนวก ข. นี้แสดงวิธีการสร้างกลุ่มของฮาดูปคลัสเตอร์บนเอมาซอนเว็บเซอร์วิสอีซีทู มีขั้นตอนดังนี้

1. ไปที่เมนู EC2 Instance

2. คลิก Hadoop cluster 01 เลือก Actions เลือก Create Image



รูปที่ ข.1 คลิกที่สร้างอิมเมจ

3. คลิกเมนู AMI tab ในหน้าต่างฝั่งซ้าย คลิก Image คลิก Launch

รูปที่ ข.2 คลิกสร้างอินสแตนซ์จากอิมเมจ

จะพบกับอินสแตนซ์ที่สร้างใหม่ดังรูปที่ ข.3 ดังภาพ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Name	Instance ID	Instance Ty	Availabili	Instance S	Status Checks	Alarm Stat	Public DNS	Public IP
Hadoop Cluster 01	i-e80833e3	m3.large	us-west-2a	running	2/2 checks...	None	ec2-54-68-117-76.us-...	54.68.117.7
Hadoop Cluster 02	i-42e9d249	m3.large	us-west-2a	running	Initializing	None	ec2-54-69-27-58.us-w...	54.69.27.58
Hadoop Cluster 03	i-82e9d289	m3.large	us-west-2a	running	Initializing	None	ec2-54-68-194-85.us-...	54.68.194.8
Hadoop Cluster 04	i-8de9d296	m3.large	us-west-2a	pending	Initializing	None		

รูปที่ ข.3 ดูผลลัพธ์การสร้างอินสแตนซ์จากอิมเมจที่สร้างไว้

6. เปิดเทอร์มินัลไปที่มาสเตอร์โหนด แล้วพิมพ์คำสั่งดังต่อไปนี้

```
sudo vi /usr/local/hadoop/conf/masters
```

7. ใส่หมายเลขไอพีส่วนตัวของเนมโหนดในไฟล์คอนฟิกกูเรชันนี้ แล้วบันทึกผล

8. พิมพ์คำสั่งดังต่อไปนี้

```
sudo vi /usr/local/hadoop/conf/slaves
```

9. ใส่หมายเลขไอพีส่วนตัวของเดทาโหนดในไฟล์คอนฟิกกูเรชันนี้ แล้วบันทึกผล

10. พิมพ์คำสั่งดังต่อไปนี้

```
ssh-copy-id -i $HOME/.ssh/id_rsa.pub
ubuntu@172.31.33.6
yes
ssh 172.31.33.6
exit
```

11. ทำซ้ำข้อ 6 ถึงข้อ 9 สำหรับเวิร์กเกอร์โหนดทุกเครื่อง

12. เริ่มต้นการทำงานของฮาดูปด้วยการพิมพ์คำสั่ง start-all.sh

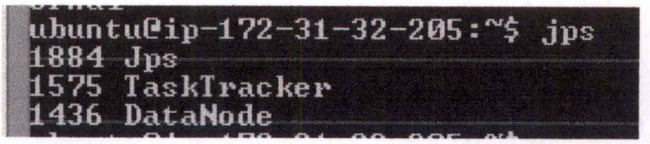
13. พิมพ์คำสั่ง jps ที่มาสเตอร์โหนด เพื่อตรวจสอบการทำงานของโปรเซสฮาดูป จะได้ผลลัพธ์ ดังรูปที่ ข.4

```
ubuntu@ip-172-31-33-165:~$ jps
1659 JobTracker
1581 SecondaryNameNode
1835 Jps
1367 NameNode
```

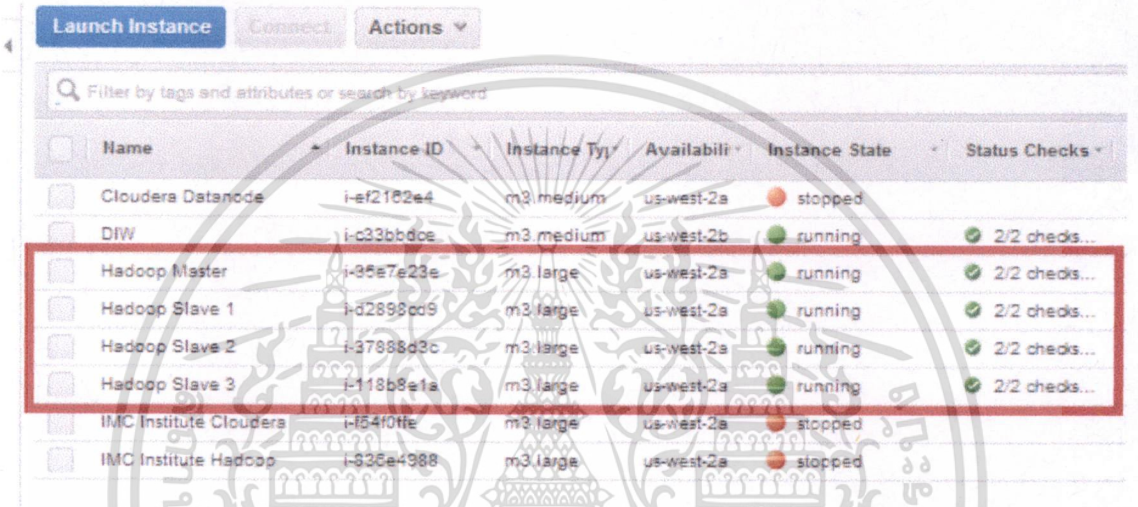
รูปที่ ข.4 ดูเซอร์วิสที่ทำงานอยู่บนมาสเตอร์โหนด

14. พิมพ์คำสั่ง jps ที่เวิร์กเกอร์โหนดทุกเครื่อง เพื่อตรวจสอบการทำงานของโปรเซสฮาดูป จะได้ผลลัพธ์ ดังรูปที่ ข.5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ข.5 ดูเซอร์วิสที่ทำงานอยู่บนเวิร์กเกอร์โหนด



รูปที่ ข.6 ดูผลลัพธ์ของฮาดูปคลัสเตอร์บนหน้าจอแสดงผลของเอมาซอนเว็บเซอร์วิสอีซีทู

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ค.

การติดตั้งมาเฮาท์ไลบรารีบนฮาดูป

การติดตั้งมาเฮาท์บนฮาดูปนั้น จำเป็นต้องติดตั้งฮาดูปให้เรียบร้อยเสียก่อน นอกจากนั้นแล้วต้องติดตั้งมาเว็น และซับเวอร์ชัน ดังนั้น ในภาคผนวกนี้จะเริ่มติดตั้งแต่มาเว็น ซับเวอร์ชัน และมาเฮาท์ตามลำดับ

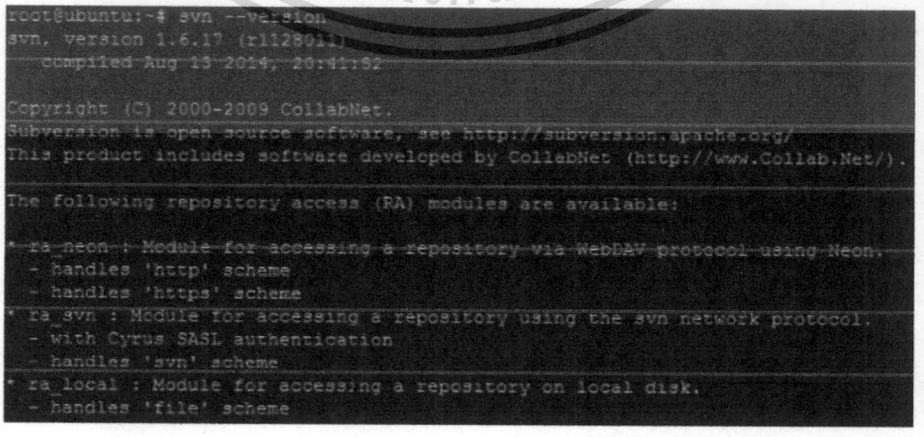
การติดตั้งมาเว็น

1. เปิดเทอร์มินัลขึ้นมาและพิมพ์คำสั่ง `sudo apt-get install maven`
2. ตรวจสอบว่าติดตั้งสำเร็จหรือไม่ ด้วยการพิมพ์คำสั่ง `mvn -v` ต้องขึ้นรายละเอียดดังนี้

รูปที่ ค.1 ตรวจสอบผลลัพธ์การติดตั้งมาเว็น

การติดตั้งซับเวอร์ชัน

1. เปิดเทอร์มินัลขึ้นมาและพิมพ์คำสั่ง `sudo apt-get install subversion`
2. ตรวจสอบหลังจากลงเสร็จด้วยคำสั่ง `svn -version` จะต้องขึ้นรายละเอียดดังภาพ



รูปที่ ค.2 ดูผลลัพธ์การติดตั้งซับเวอร์ชันสำเร็จ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การติดตั้งมาเฮาท์

1. เปิดเทอร์มินัลไปยังไดเรกทอรีที่เราต้องการติดตั้งมาเฮาท์ และพิมพ์คำสั่งดังนี้

```
cd /home/hadoop
mkdir mahout
cd mahout
```

2. พิมพ์คำสั่ง ดังต่อไปนี้

```
svn co http://svn.apache.org/repos/asf/mahout/trunk
cd trunk
mvn install
```

3. ถ้าจบด้วยหน้าจอลักษณะเช่นนี้ แสดงว่าติดตั้งมาเฮาท์เสร็จเรียบร้อยแล้ว

รูปที่ ค.3 ผลลัพธ์การติดตั้งมาเฮาท์สำเร็จ

จากนั้นให้แก้ไขไฟล์ ~/.bash_profile ดังนี้

```
export HADOOP_CONF_DIR=$HADOOP_HOME/conf
export MAHOUT_HOME=/opt/mahout
export PATH=$PATH:$MAHOUT_HOME/bin
export HADOOP_PREFIX=/usr/local/hadoop
export HADOOP_PREFIX=/usr/local/hadoop
export PATH=$PATH:$HADOOP_PREFIX/bin
export HADOOP_CONF_DIR= HADOOP_PREFIX/conf
export MAHOUT_HOME=/opt/mahout
export PATH=$PATH:$MAHOUT_HOME/bin
export CLASSPATH=$MAHOUT_HOME/lib:$CLASSPATH
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. สั่งให้แบช (bash) ทำงานด้วยการพิมพ์คำสั่ง `exec bash`

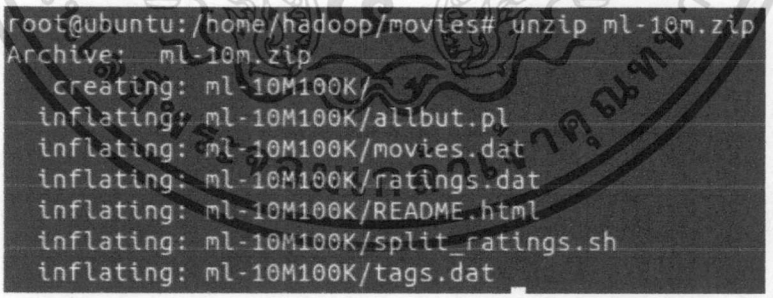
5. ตรวจสอบการทำงานของแบช โดยการพิมพ์คำสั่ง `echo $MAHOUT_HOME` ผลลัพธ์ในหน้าจอแสดงผลจะต้องขึ้นว่า /opt/mahout

6. เปิดเทอร์มินัลของฮาดูปเพื่อสร้างโฟลเดอร์ในการเก็บข้อมูลที่จะดาวน์โหลดมา จากนั้นให้พิมพคำสั่ง ดังนี้

```
mkdir /home/hadoop/movie
wget http://files.grouplens.org/datasets/movielens/m
l-10m.zip
```

รูปที่ ค.4 ผลลัพธ์การดาวน์โหลดข้อมูลจากมูฟวี่เลนส์

จากนั้นให้แตกไฟล์ zip ออกมาโดยใช้คำสั่ง `unzip ml-10m.zip`



รูปที่ ค.5 การพิมพ์คำสั่งแตกไฟล์ของข้อมูลมูฟวี่เลนส์

เมื่อเข้ามาตรวจสอบไฟล์ด้านในดูจะพบว่ามีไฟล์อยู่ 6 ไฟล์ คือ allbut.pl, movies.dat, rating.dat, README.html, split_ratings.sh และ tags.dat

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
root@ubuntu:/home/hadoop/movies# pwd
/home/hadoop/movies
root@ubuntu:/home/hadoop/movies# ls
ml-10M100K  ml-10m.zip
root@ubuntu:/home/hadoop/movies# cd ml-10M100K/
root@ubuntu:/home/hadoop/movies/ml-10M100K# ls
allbut.pl  movies.dat  ratings.dat  README.html  split_ratings.sh  tags.dat
root@ubuntu:/home/hadoop/movies/ml-10M100K#
```

รูปที่ ค.6 รายละเอียดภายในของไฟล์มูฟวี่เลนส์

ไฟล์ movies.dat ประกอบไปด้วยหมายเลขภาพยนตร์ ชื่อเรื่อง ปีที่ฉาย และหมวดหมู่ภาพยนตร์

```
root@ubuntu:/home/hadoop/movies/ml-10M100K# tail movies.dat
65006::Impulse (2008)::Mystery|Thriller
65011::Zona Zamfirova (2002)::Comedy|Drama
65025::Double Dynamite (1951)::Comedy|Musical
65027::Death Kiss, The (1933)::Comedy|Mystery
65037::Ben X (2007)::Drama
65088::Bedtime Stories (2008)::Adventure|Children|Comedy
65091::Manhattan Melodrama (1934)::Crime|Drama|Romance
65126::Choke (2008)::Comedy|Drama
65130::Revolutionary Road (2008)::Drama|Romance
65133::Blackadder Back & Forth (1999)::Comedy
```

รูปที่ ค.7 รายละเอียดของไฟล์ movies.dat

```
root@ubuntu:/home/hadoop/movies/ml-10M100K# tail ratings.dat
71567::1984::1::912580553
71567::1985::1::912580553
71567::1986::1::912580553
71567::2012::3::912580722
71567::2028::5::912580344
71567::2107::1::912580553
71567::2126::2::912649143
71567::2294::5::912577968
71567::2338::2::912578016
71567::2384::2::912578173
```

รูปที่ ค.8 รายละเอียดของไฟล์ ratings.dat

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
root@ubuntu:/home/hadoop/movies/ml-10M100K# tail tags.dat
71534::30701::Orson Welles::1196647472
71536::920::a good romantic film::1162110221
71536::1219::Alfred Hitchcock::1162126283
71536::3996::a old story of fight::1162110266
71556::1302::Baseball::1188263570
71556::1377::Gothic::1188263571
71556::2424::chick flick::1188263606
71556::3033::comedy::1188263626
71556::3081::Gothic::1188263565
71556::7438::Western::1188263589
```

รูปที่ ค.9 รายละเอียดของไฟล์ tags.dat

```
root@ubuntu:/home/hadoop/movies/ml-10M100K# cat split_ratings.sh
#!/bin/sh

RATINGS_COUNT=`wc -l ratings.dat | cut -d ' ' -f 1`
echo "ratings count: $RATINGS_COUNT"
SET_SIZE=`expr $RATINGS_COUNT / 5`
echo "set size: $SET_SIZE"
REMAINDER=`expr $RATINGS_COUNT % 5`
echo "remainder: $REMAINDER"

for i in 1 2 3 4 5
  do
    head -`expr $i \* $SET_SIZE` ratings.dat | tail -$SET_SIZE > r$i.test

    head -`expr \( $i - 1 \) \* $SET_SIZE` ratings.dat > r$i.train
    tail -`expr \( 5 - $i \) \* $SET_SIZE` ratings.dat >> r$i.train

    if [ $i -eq 5 ]; then
      tail -$REMAINDER ratings.dat >> r5.test
    else
      tail -$REMAINDER ratings.dat >> r$i.train
    fi

    echo "r$i.test created. `wc -l r$i.test | cut -d " " -f 1` lines."
    echo "r$i.train created. `wc -l r$i.train | cut -d " " -f 1` lines."
done

./allbut.pl ra 1 10 0 ratings.dat
echo "ra.test created. `wc -l ra.test | cut -d " " -f 1` lines."
echo "ra.train created. `wc -l ra.train | cut -d " " -f 1` lines."

./allbut.pl rb 11 20 0 ratings.dat
echo "rb.test created. `wc -l rb.test | cut -d " " -f 1` lines."
echo "rb.train created. `wc -l rb.train | cut -d " " -f 1` lines."
```

รูปที่ ค.10 รายละเอียดของไฟล์ split_rating.sh

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ง.

ตัวอย่างข้อมูลนำเข้า

ในภาคผนวก ง. แสดงตัวอย่างข้อมูลนำเข้าที่ส่งเข้าไปประมวลผลในระบบการให้คำแนะนำ โดยจะแสดงเป็นแบบจำลองข้อมูล (Data model) ในตารางฝั่งซ้ายเป็นข้อมูลนำเข้า (Input data) ที่เมื่อประมวลผลออกมาแล้วจะได้เป็นชุดข้อมูลตั้งต้นและในตารางฝั่งขวาเป็นข้อมูลนำเข้าที่มีการดึงคะแนนความนิยมที่ผู้ใช้ได้ให้คะแนนไว้ออกเป็นจำนวน 20% ของสิ่งของทั้งหมด ซึ่งเมื่อประมวลผลออกมาแล้วจะได้เป็นชุดข้อมูลทดสอบ

ข้อมูลตัวอย่างที่แสดงในภาคผนวก ง. นี้ แสดงเฉพาะผู้ใช้จำนวน 5 รายแรกเท่านั้น

ตารางที่ ง.1 ตัวอย่างข้อมูลนำเข้า

Input 1 (original)		
User ID	Movie ID	Rating
1	61	4
1	189	3
1	33	4
1	160	4
1	20	4
1	202	5
1	171	5
1	265	4
1	155	2
1	117	3
1	47	4
1	222	4
1	253	5
1	113	5
1	227	4
1	17	3
1	90	4

Input 2 (testing)		
User ID	Movie ID	Rating
1	61	0
1	189	0
1	33	0
1	160	0
1	20	0
1	202	0
1	171	0
1	265	0
1	155	0
1	117	0
1	47	0
1	222	0
1	253	0
1	113	0
1	227	0
1	17	0
1	90	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	266	1
1	121	4
1	114	5
1	132	4
1	74	1
1	134	4
1	98	4
1	186	4
1	221	5
1	84	4
1	31	3
1	70	3
1	60	5
1	177	5
1	27	2
1	260	1
1	145	2
1	174	5
1	159	3
1	82	5
1	56	4
1	272	3
1	80	4
1	229	4
1	140	1
1	225	2
1	235	5
1	120	1
1	125	3

Input 2 (testing)		
User ID	Movie ID	Rating
1	266	0
1	121	0
1	114	0
1	132	0
1	74	0
1	134	0
1	98	0
1	186	0
1	221	0
1	84	0
1	31	0
1	70	0
1	60	0
1	177	0
1	27	0
1	260	0
1	145	0
1	174	0
1	159	0
1	82	0
1	56	0
1	272	0
1	80	0
1	229	0
1	140	0
1	225	0
1	235	0
1	120	0
1	125	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	215	3
1	6	5
1	104	1
1	49	3
1	206	4
1	76	4
1	72	4
1	185	4
1	96	5
1	213	2
1	233	2
1	258	5
1	81	5
1	78	1
1	212	4
1	143	1
1	151	4
1	51	4
1	175	5
1	107	4
1	218	3
1	209	4
1	259	1
1	108	5
1	262	3
1	12	5
1	14	5
1	97	3
1	44	5

Input 2 (testing)		
User ID	Movie ID	Rating
1	215	0
1	6	0
1	104	0
1	49	3
1	206	4
1	76	4
1	72	4
1	185	4
1	96	5
1	213	2
1	233	2
1	258	5
1	81	5
1	78	1
1	212	4
1	143	1
1	151	4
1	51	4
1	175	5
1	107	4
1	218	3
1	209	4
1	259	1
1	108	5
1	262	3
1	12	5
1	14	5
1	97	3
1	44	5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	53	3
1	163	4
1	210	4
1	184	4
1	157	4
1	201	3
1	150	5
1	183	5
1	248	4
1	208	5
1	128	4
1	242	5
1	148	2
1	112	1
1	193	4
1	264	2
1	219	1
1	232	3
1	236	4
1	252	2
1	200	3
1	180	3
1	250	4
1	85	3
1	91	5
1	10	3
1	254	1
1	129	5
1	241	4

Input 2 (testing)		
User ID	Movie ID	Rating
1	53	3
1	163	4
1	210	4
1	184	4
1	157	4
1	201	3
1	150	5
1	183	5
1	248	4
1	208	5
1	128	4
1	242	5
1	148	2
1	112	1
1	193	4
1	264	2
1	219	1
1	232	3
1	236	4
1	252	2
1	200	3
1	180	3
1	250	4
1	85	3
1	91	5
1	10	3
1	254	1
1	129	5
1	241	4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	130	3
1	255	2
1	103	1
1	118	3
1	54	3
1	267	4
1	24	3
1	86	5
1	196	5
1	39	4
1	164	3
1	230	4
1	36	2
1	23	4
1	224	5
1	73	3
1	67	3
1	65	4
1	190	5
1	100	5
1	226	3
1	243	1
1	154	5
1	214	4
1	161	4
1	62	3
1	188	3
1	102	2
1	69	3

Input 2 (testing)		
User ID	Movie ID	Rating
1	130	3
1	255	2
1	103	1
1	118	3
1	54	3
1	267	4
1	24	3
1	86	5
1	196	5
1	39	4
1	164	3
1	230	4
1	36	2
1	23	4
1	224	5
1	73	3
1	67	3
1	65	4
1	190	5
1	100	5
1	226	3
1	243	1
1	154	5
1	214	4
1	161	4
1	62	3
1	188	3
1	102	2
1	69	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	170	5
1	38	3
1	9	5
1	246	5
1	22	4
1	21	1
1	179	3
1	187	4
1	135	4
1	68	4
1	146	4
1	176	5
1	166	5
1	138	1
1	247	1
1	89	5
1	2	3
1	30	3
1	63	2
1	249	4
1	269	5
1	32	5
1	141	3
1	211	3
1	40	3
1	270	5
1	133	4
1	239	4
1	194	4

Input 2 (testing)		
User ID	Movie ID	Rating
1	170	5
1	38	3
1	9	5
1	246	5
1	22	4
1	21	1
1	179	3
1	187	4
1	135	4
1	68	4
1	146	4
1	176	5
1	166	5
1	138	1
1	247	1
1	89	5
1	2	3
1	30	3
1	63	2
1	249	4
1	269	5
1	32	5
1	141	3
1	211	3
1	40	3
1	270	5
1	133	4
1	239	4
1	194	4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	256	4
1	220	3
1	93	5
1	8	1
1	205	3
1	234	4
1	105	2
1	147	3
1	99	3
1	1	5
1	197	5
1	173	5
1	75	4
1	268	5
1	34	2
1	144	4
1	271	2
1	119	5
1	26	3
1	158	3
1	37	2
1	181	5
1	136	3
1	257	4
1	237	2
1	131	1
1	109	5
1	182	4
1	71	3

Input 2 (testing)		
User ID	Movie ID	Rating
1	256	4
1	220	3
1	93	5
1	8	1
1	205	3
1	234	4
1	105	2
1	147	3
1	99	3
1	1	5
1	197	5
1	173	5
1	75	4
1	268	5
1	34	2
1	144	4
1	271	2
1	119	5
1	26	3
1	158	3
1	37	2
1	181	5
1	136	3
1	257	4
1	237	2
1	131	1
1	109	5
1	182	4
1	71	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	223	5
1	46	4
1	169	5
1	41	2
1	162	4
1	110	1
1	66	4
1	77	4
1	199	4
1	57	5
1	50	5
1	192	4
1	178	5
1	5	3
1	87	5
1	238	4
1	156	4
1	106	4
1	167	2
1	115	5
1	11	2
1	245	2
1	35	1
1	137	5
1	127	5
1	16	5
1	79	4
1	261	1
1	45	5

Input 2 (testing)		
User ID	Movie ID	Rating
1	223	5
1	46	4
1	169	5
1	41	2
1	162	4
1	110	1
1	66	4
1	77	4
1	199	4
1	57	5
1	50	5
1	192	4
1	178	5
1	5	3
1	87	5
1	238	4
1	156	4
1	106	4
1	167	2
1	115	5
1	11	2
1	245	2
1	35	1
1	137	5
1	127	5
1	16	5
1	79	4
1	261	1
1	45	5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	48	5
1	25	4
1	251	4
1	195	5
1	153	3
1	101	2
1	168	5
1	123	4
1	191	5
1	4	3
1	263	1
1	203	4
1	55	5
1	42	5
1	139	3
1	240	3
1	7	4
1	149	2
1	43	4
1	165	5
1	116	3
1	198	5
1	124	5
1	95	4
1	217	3
1	58	4
1	142	2
1	216	5
1	126	2

Input 2 (testing)		
User ID	Movie ID	Rating
1	48	5
1	25	4
1	251	4
1	195	5
1	153	3
1	101	2
1	168	5
1	123	4
1	191	5
1	4	3
1	263	1
1	203	4
1	55	5
1	42	5
1	139	3
1	240	3
1	7	4
1	149	2
1	43	4
1	165	5
1	116	3
1	198	5
1	124	5
1	95	4
1	217	3
1	58	4
1	142	2
1	216	5
1	126	2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
1	83	3
1	231	1
1	204	5
1	3	4
1	207	5
1	244	2
1	19	5
1	29	1
1	18	4
1	59	5
1	15	5
1	111	5
1	52	4
1	88	4
1	13	5
1	28	4
1	172	5
1	122	3
1	152	5
1	94	2
2	292	4
2	251	5
2	50	5
2	314	1
2	297	4
2	290	3
2	312	3
2	281	3

Input 2 (testing)		
User ID	Movie ID	Rating
1	83	3
1	231	1
1	204	5
1	3	4
1	207	5
1	244	2
1	19	5
1	29	1
1	18	4
1	59	5
1	15	5
1	111	5
1	52	4
1	88	4
1	13	5
1	28	4
1	172	5
1	122	3
1	152	5
1	94	2
2	292	0
2	251	0
2	50	0
2	314	0
2	297	0
2	290	0
2	312	0
2	281	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
2	13	4
2	280	3
2	303	4
2	308	3
2	307	3
2	257	4
2	316	5
2	315	1
2	301	4
2	313	5
2	279	4
2	299	4
2	298	3
2	19	3
2	277	4
2	282	4
2	111	4
2	258	3
2	295	4
2	242	5
2	283	5
2	276	4
2	1	4
2	305	3
2	14	4
2	287	3
2	291	3
2	293	4
2	294	1

Input 2 (testing)		
User ID	Movie ID	Rating
2	13	0
2	280	0
2	303	0
2	308	0
2	307	3
2	257	4
2	316	5
2	315	1
2	301	4
2	313	5
2	279	4
2	299	4
2	298	3
2	19	3
2	277	4
2	282	4
2	111	4
2	258	3
2	295	4
2	242	5
2	283	5
2	276	4
2	1	4
2	305	3
2	14	4
2	287	3
2	291	3
2	293	4
2	294	1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
2	310	4
2	309	1
2	306	4
2	25	4
2	273	4
2	10	2
2	311	5
2	269	4
2	255	4
2	284	4
2	274	3
2	237	4
2	300	4
2	100	5
2	127	5
2	285	5
2	289	3
2	304	4
2	272	5
2	278	3
2	288	3
2	286	4
2	275	5
2	302	5
2	296	3
3	335	1
3	245	1
3	337	1

Input 2 (testing)		
User ID	Movie ID	Rating
2	310	4
2	309	1
2	306	4
2	25	4
2	273	4
2	10	2
2	311	5
2	269	4
2	255	4
2	284	4
2	274	3
2	237	4
2	300	4
2	100	5
2	127	5
2	285	5
2	289	3
2	304	4
2	272	5
2	278	3
2	288	3
2	286	4
2	275	5
2	302	5
2	296	3
3	335	0
3	245	0
3	337	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
3	343	3
3	323	2
3	331	4
3	294	2
3	332	1
3	328	5
3	334	3
3	350	3
3	341	1
3	318	4
3	300	2
3	345	3
3	299	3
3	324	2
3	348	4
3	351	3
3	330	2
3	327	4
3	307	3
3	272	2
3	354	3
3	264	2
3	349	3
3	321	5
3	260	4
3	268	3
3	288	2
3	355	3
3	320	5

Input 2 (testing)		
User ID	Movie ID	Rating
3	343	0
3	323	0
3	331	0
3	294	0
3	332	0
3	328	0
3	334	0
3	350	0
3	341	1
3	318	4
3	300	2
3	345	3
3	299	3
3	324	2
3	348	4
3	351	3
3	330	2
3	327	4
3	307	3
3	272	2
3	354	3
3	264	2
3	349	3
3	321	5
3	260	4
3	268	3
3	288	2
3	355	3
3	320	5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
3	258	2
3	339	3
3	342	4
3	303	3
3	329	4
3	317	2
3	181	4
3	338	2
3	302	2
3	322	3
3	352	2
3	271	3
3	333	2
3	344	4
3	326	2
3	319	2
3	325	1
3	347	5
3	336	1
3	353	1
3	340	5
3	346	5
4	264	3
4	303	5
4	361	5
4	357	4
4	260	4
4	356	3

Input 2 (testing)		
User ID	Movie ID	Rating
3	258	2
3	339	3
3	342	4
3	303	3
3	329	4
3	317	2
3	181	4
3	338	2
3	302	2
3	322	3
3	352	2
3	271	3
3	333	2
3	344	4
3	326	2
3	319	2
3	325	1
3	347	5
3	336	1
3	353	1
3	340	5
3	346	5
4	264	0
4	303	0
4	361	0
4	357	0
4	260	0
4	356	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
4	294	5
4	288	4
4	50	5
4	354	5
4	271	4
4	300	5
4	328	3
4	258	5
4	210	3
4	329	5
4	11	4
4	327	5
4	324	5
4	359	5
4	362	5
4	358	2
4	360	5
4	301	5
5	2	3
5	17	4
5	439	1
5	225	2
5	110	1
5	454	1
5	424	1
5	1	4
5	363	3
5	98	3

Input 2 (testing)		
User ID	Movie ID	Rating
4	294	5
4	288	4
4	50	5
4	354	5
4	271	4
4	300	5
4	328	3
4	258	5
4	210	3
4	329	5
4	11	4
4	327	5
4	324	5
4	359	5
4	362	5
4	358	2
4	360	5
4	301	5
5	2	0
5	17	0
5	439	0
5	225	0
5	110	0
5	454	0
5	424	0
5	1	0
5	363	0
5	98	0

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
5	102	3
5	211	4
5	382	5
5	376	2
5	62	4
5	231	2
5	377	1
5	407	3
5	24	4
5	423	4
5	394	2
5	384	3
5	267	4
5	445	3
5	167	2
5	426	3
5	222	4
5	453	1
5	403	3
5	173	4
5	241	1
5	234	2
5	154	3
5	436	5
5	42	5
5	422	4
5	139	3
5	40	4
5	90	3

Input 2 (testing)		
User ID	Movie ID	Rating
5	102	0
5	211	0
5	382	0
5	376	0
5	62	0
5	231	0
5	377	0
5	407	0
5	24	0
5	423	0
5	394	0
5	384	0
5	267	0
5	445	0
5	167	0
5	426	0
5	222	0
5	453	0
5	403	0
5	173	0
5	241	0
5	234	0
5	154	0
5	436	0
5	42	0
5	422	4
5	139	3
5	40	4
5	90	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
5	153	5
5	94	3
5	389	1
5	411	1
5	109	5
5	230	3
5	388	2
5	227	4
5	397	2
5	444	2
5	402	1
5	100	5
5	143	3
5	370	1
5	176	3
5	441	1
5	69	1
5	417	3
5	79	3
5	418	3
5	429	3
5	385	4
5	372	3
5	421	1
5	144	3
5	243	1
5	185	3
5	393	2
5	413	3

Input 2 (testing)		
User ID	Movie ID	Rating
5	153	5
5	94	3
5	389	1
5	411	1
5	109	5
5	230	3
5	388	2
5	227	4
5	397	2
5	444	2
5	402	1
5	100	5
5	143	3
5	370	1
5	176	3
5	441	1
5	69	1
5	417	3
5	79	3
5	418	3
5	429	3
5	385	4
5	372	3
5	421	1
5	144	3
5	243	1
5	185	3
5	393	2
5	413	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
5	89	5
5	400	1
5	80	2
5	433	5
5	219	3
5	428	5
5	259	1
5	457	1
5	410	1
5	369	1
5	435	4
5	214	3
5	364	1
5	209	5
5	391	4
5	379	3
5	381	1
5	99	3
5	21	3
5	427	3
5	451	1
5	430	5
5	135	4
5	449	2
5	226	3
5	401	5
5	405	3
5	443	4
5	257	5

Input 2 (testing)		
User ID	Movie ID	Rating
5	89	5
5	400	1
5	80	2
5	433	5
5	219	3
5	428	5
5	259	1
5	457	1
5	410	1
5	369	1
5	435	4
5	214	3
5	364	1
5	209	5
5	391	4
5	379	3
5	381	1
5	99	3
5	21	3
5	427	3
5	451	1
5	430	5
5	135	4
5	449	2
5	226	3
5	401	5
5	405	3
5	443	4
5	257	5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
5	25	3
5	442	1
5	186	5
5	432	4
5	194	4
5	229	2
5	183	4
5	440	1
5	29	4
5	378	1
5	392	2
5	390	5
5	101	5
5	208	4
5	399	3
5	456	1
5	168	3
5	50	4
5	408	5
5	416	1
5	162	1
5	95	4
5	452	1
5	395	2
5	172	5
5	414	3
5	70	4
5	437	1
5	151	3

Input 2 (testing)		
User ID	Movie ID	Rating
5	25	3
5	442	1
5	186	5
5	432	4
5	194	4
5	229	2
5	183	4
5	440	1
5	29	4
5	378	1
5	392	2
5	390	5
5	101	5
5	208	4
5	399	3
5	456	1
5	168	3
5	50	4
5	408	5
5	416	1
5	162	1
5	95	4
5	452	1
5	395	2
5	172	5
5	414	3
5	70	4
5	437	1
5	151	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
5	415	1
5	371	1
5	396	5
5	398	2
5	204	4
5	216	1
5	420	3
5	438	1
5	387	3
5	447	3
5	228	5
5	250	3
5	145	1
5	409	2
5	386	2
5	450	1
5	235	4
5	169	5
5	431	3
5	121	4
5	374	3
5	365	1
5	189	5
5	446	4
5	406	1
5	239	4
5	425	2
5	66	1
5	366	3

Input 2 (testing)		
User ID	Movie ID	Rating
5	415	1
5	371	1
5	396	5
5	398	2
5	204	4
5	216	1
5	420	3
5	438	1
5	387	3
5	447	3
5	228	5
5	250	3
5	145	1
5	409	2
5	386	2
5	450	1
5	235	4
5	169	5
5	431	3
5	121	4
5	374	3
5	365	1
5	189	5
5	446	4
5	406	1
5	239	4
5	425	2
5	66	1
5	366	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ง. 1 (ต่อ)

Input 1 (original)		
User ID	Movie ID	Rating
5	455	4
5	367	3
5	434	5
5	63	1
5	181	5
5	412	3
5	163	5
5	404	2
5	383	3
5	233	4
5	200	2
5	380	3
5	105	3
5	210	3
5	448	2
5	419	3
5	375	3
5	373	3
5	368	1
5	174	5

Input 2 (testing)		
User ID	Movie ID	Rating
5	455	4
5	367	3
5	434	5
5	63	1
5	181	5
5	412	3
5	163	5
5	404	2
5	383	3
5	233	4
5	200	2
5	380	3
5	105	3
5	210	3
5	448	2
5	419	3
5	375	3
5	373	3
5	368	1
5	174	5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวณัฏฐินันท์ ยนต์ไชย
วันเกิด	15 ธันวาคม 2532
สถานที่เกิด	กรุงเทพมหานคร
สถานที่พัก	เขตประเวศ กรุงเทพมหานคร
วุฒิการศึกษาระดับปริญญาตรี	สาขาวิศวกรรมคอมพิวเตอร์
	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ประสบการณ์การทำงาน	
พ.ศ.2554-ปัจจุบัน	ตำแหน่งวิศวกรชำนาญการพิเศษ
	บริษัท เทอร์ราไบท์ เน็ท โซลูชั่น จำกัด (มหาชน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวณัฏฐินันท์ ยนต์ไชย
วันเกิด	15 ธันวาคม 2532
สถานที่เกิด	กรุงเทพมหานคร
สถานที่พัก	เขตประเวศ กรุงเทพมหานคร
วุฒิการศึกษาระดับปริญญาตรี	สาขาวิศวกรรมคอมพิวเตอร์
	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ประสบการณ์การทำงาน	
พ.ศ.2554-ปัจจุบัน	ตำแหน่งวิศวกรชำนาญการพิเศษ
	บริษัท เทอร์ราไบท์ เน็ท โซลูชั่น จำกัด (มหาชน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้