Mobile User and App Analytics in China

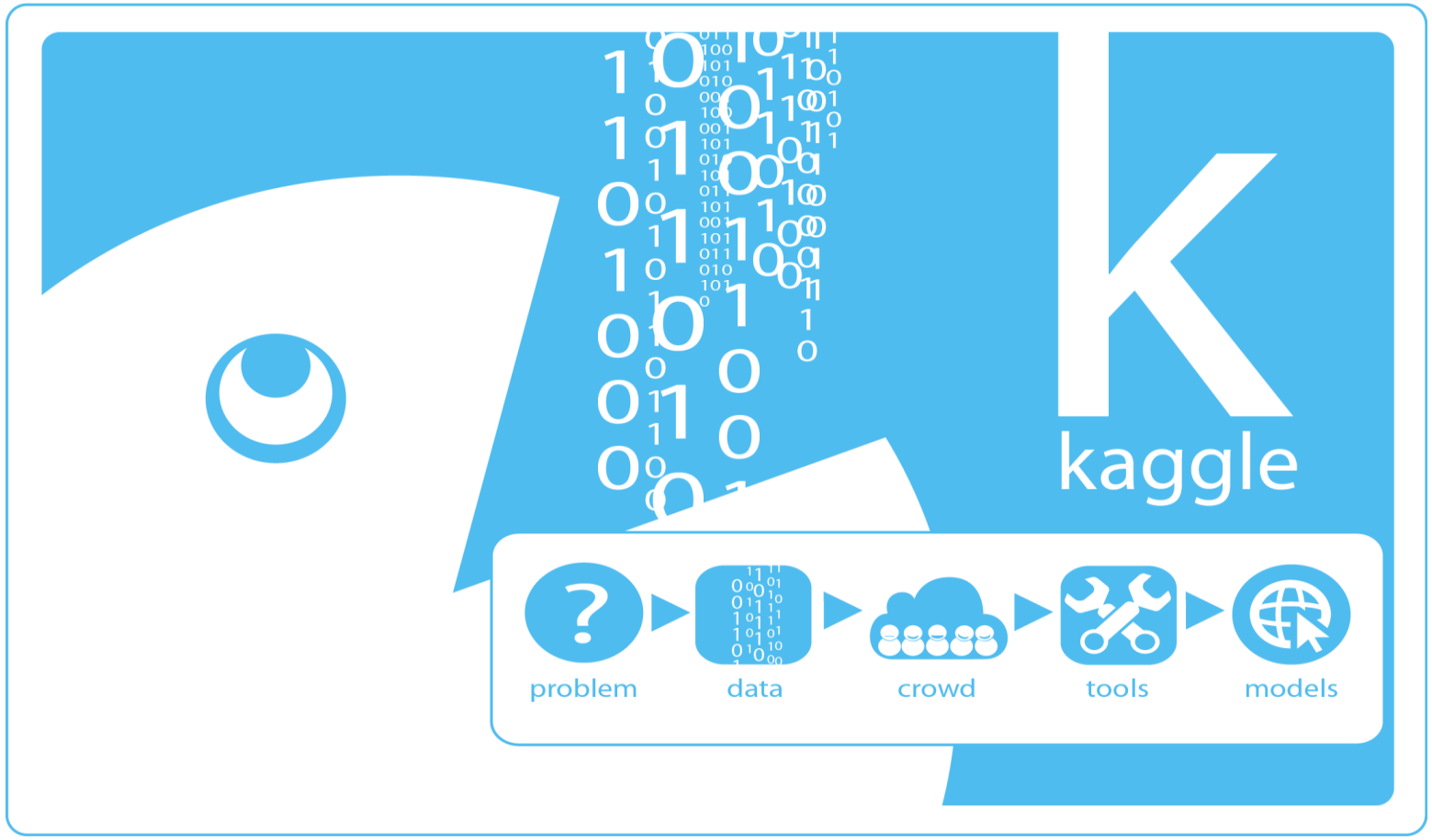
TEAM APACHE HADOOP, IMC INSTITUTE

30 JULY 2016

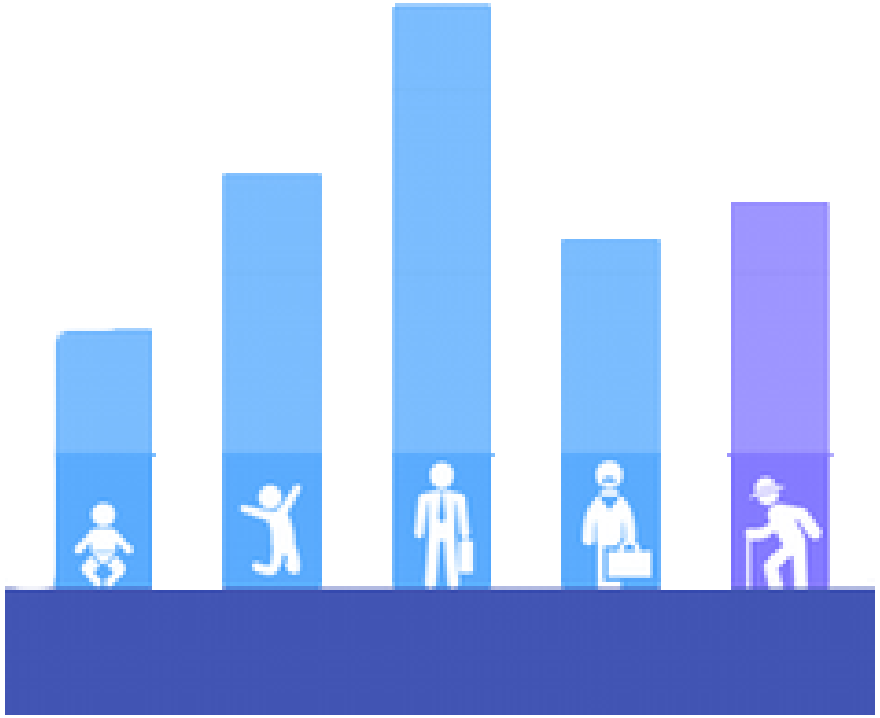
IMC Institute: Apache Hadoop Team Logo

Logo Credit: Agile Thailand 2016

ความเป็นมาของโจทย์: แนะนำ Kaggle

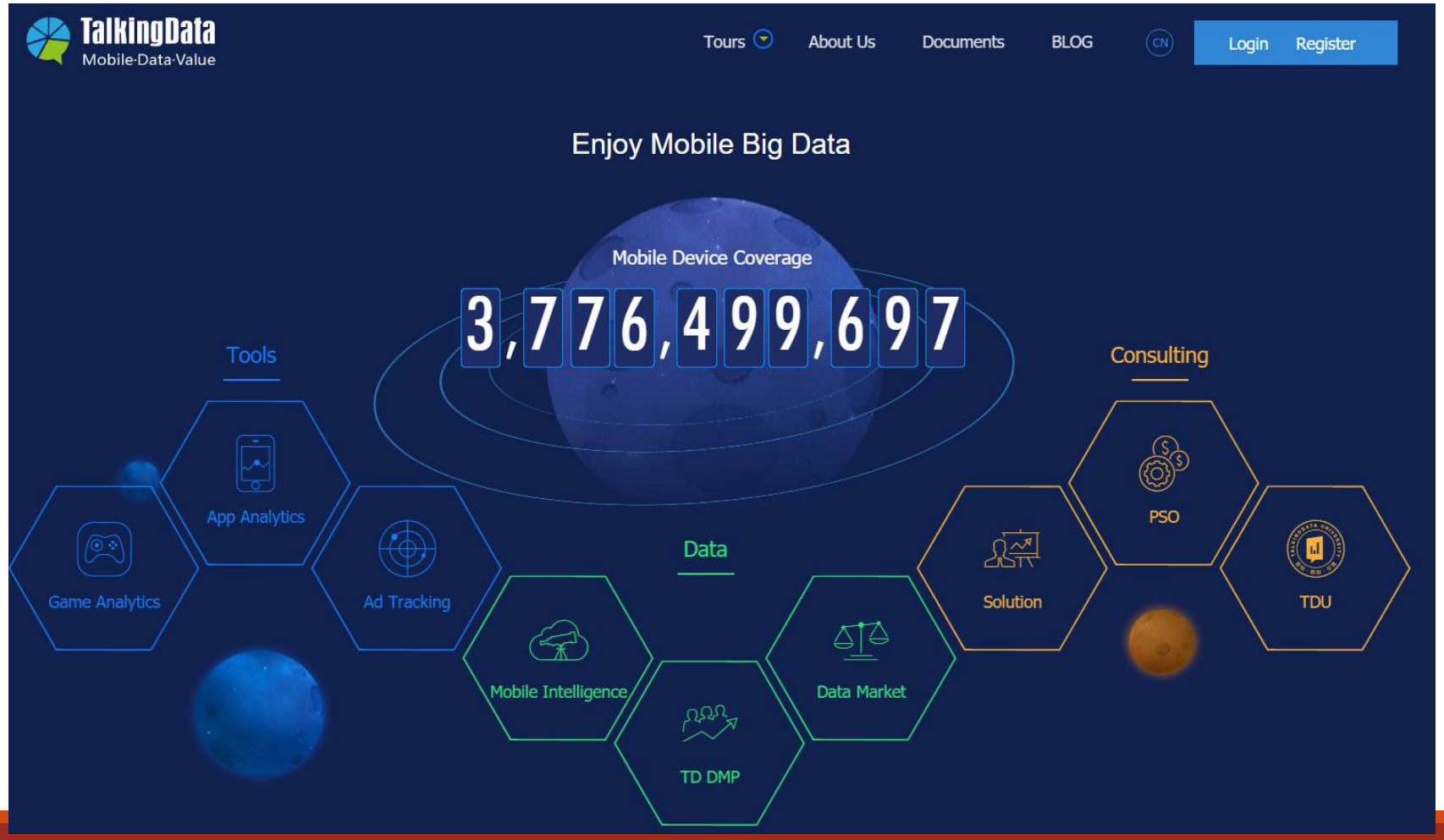


TalkingData คือบริษัทอะไร?



*"**TalkingData** เป็น**แพลตฟอร์มของบุคคลที่สาม**ข้อมูลมือถือที่ใหญ่ที่สุดของประเทศจีน ทางบริษัทเข้าใจว่าทางเลือกในชีวิตประจำวันและพฤติกรรมของผู้ใช้มือถือผลักดันให้พวกเราสร้างคุณค่าต่างๆได้ ปัจจุบันบริษัท TalkingData กำลังมองหา**ประโยชน์จากฐานข้อมูลพฤติกรรมผู้ใช้มือถือจากกว่า 70% ของ 500 ล้านโทรศัพท์มือถือที่ใช้งานในชีวิตประจำวัน**ในประเทศจีน**เพื่อช่วยให้ลูกค้าของตนเข้าใจและมีปฏิสัมพันธ์กับผู้ใช้ของพวกเขา"*

TalkingData: ข้อมูลขนาดใหญ่ที่เข้ามาในแต่ละวัน



TalkingData Website: https://www.talkingdata.com/

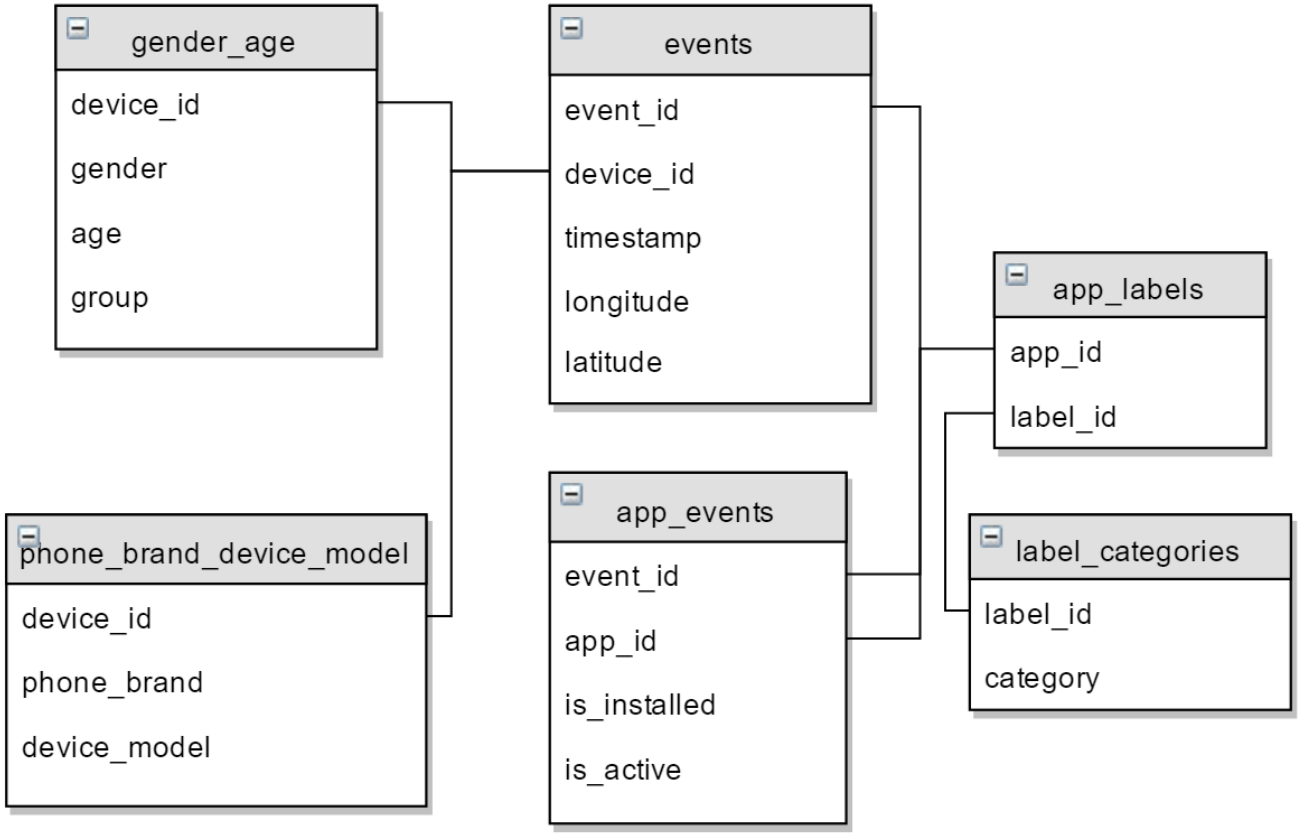
โจทย์ปัญหา

- พฤติกรรมการใช้แอพพลิเคชั่นของผู้ใช้มือถือ
  - แอพพลิเคชั่นประเภทใดได้รับความนิยมมากที่สุด
  - ผู้ใช้มือถือนิยมใช้แอพพลิเคชั่นในช่วงใดของวันและวันใดบ้างในแต่ละอาทิตย์
  - จำนวนผู้ใช้มือถือแบ่งตามเพศและอายุกลุ่มใดมากที่สุดที่ปรากฏในชุดข้อมูล
- แบรนด์โทรศัพท์มือถือใดกำลังครองตลาดอยู่ในประเทศจีน
- รุ่นโทรศัพท์มือถือใดกำลังครองตลาดอยู่ในประเทศจีน
- ความสัมพันธ์ระหว่างจำนวนแอพพลิเคชั่นในแต่ละประเภทของแอพพลิเคชั่น
- เราจะมีวิธีอย่างไรบ้างในการคาดเดากลุ่มผู้ใช้มือถือตามการเข้าใช้งานแอพพลิเคชั่น
- เราจะมีวิธีอย่างไรบ้างในการคาดเดาอัตราการใช้งานของผู้ใช้มือถือ

จุดประสงค์ของโปรเจ็ค

- เรียนรู้การใช้ **AWS & Microsoft Azure** เพื่อสร้าง Instances การทำงานแบบ Single Node & Cluster (Lecture: อ.ธนชาติ)
- ทราบถึงความสำคัญของ Big Data และวิธีการรับมือข้อมูลขนาดใหญ่
- การใช้ **Hadoop** เพื่อเก็บข้อมูลเข้า HDFS รวมไปถึงการดึงข้อมูลโดยใช้ภาษา SQL ผ่านเครื่องมือ **Hive Impala และ SparkSQL**
- เรียนรู้การใช้ **Mass Analytics Tools** เพื่อการวิเคราะห์ข้อมูล แปลงจากข้อมูลเป็น Knowledge/Discovery (Lecture: อ.โกเมธ)
- ทดลองการใช้ **Machine Learning for Business** แก้ปัญหาเชิงธุรกิจ
- สร้าง **Web-based and Interactive Visualization ด้วยภาษา Javascript** เพื่อสวยงามและสะดวกต่อผู้ใช้บริการ (Lecture: อ.ชินวิทย์)

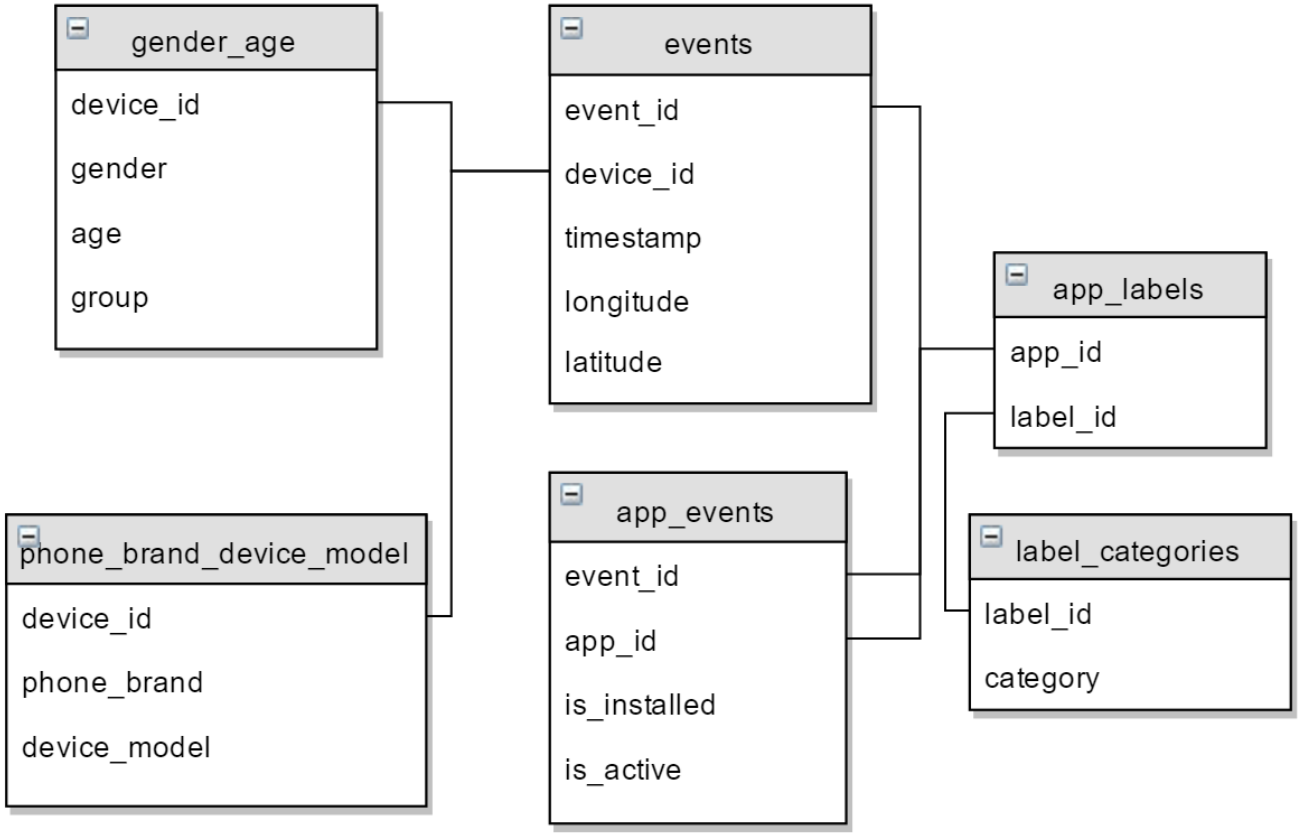
คำอธิบายชุดข้อมูล TalkingData on “Kaggle”



ข้อมูล Gender_age ประกอบด้วย 4 สดมภ์ 74,645 แถว มีคำอธิบายตัวแปรดังนี้

- Device_id คือ หมายเลข (นิรนาม สำหรับข้อมูลชุดนี้) ที่เป็นเฉพาะของผู้ใช้แอพพลิเคชั่น
- Gender คือ เพศของผู้ใช้แอพพลิเคชั่น
- Age คือ อายุของผู้ใช้แอพพลิเคชั่น
- Group คือ การจัดกลุ่มอายุของผู้ใช้ของแอพพลิเคชั่น ซึ่งทาง TalkingData จัดไว้ให้แล้ว

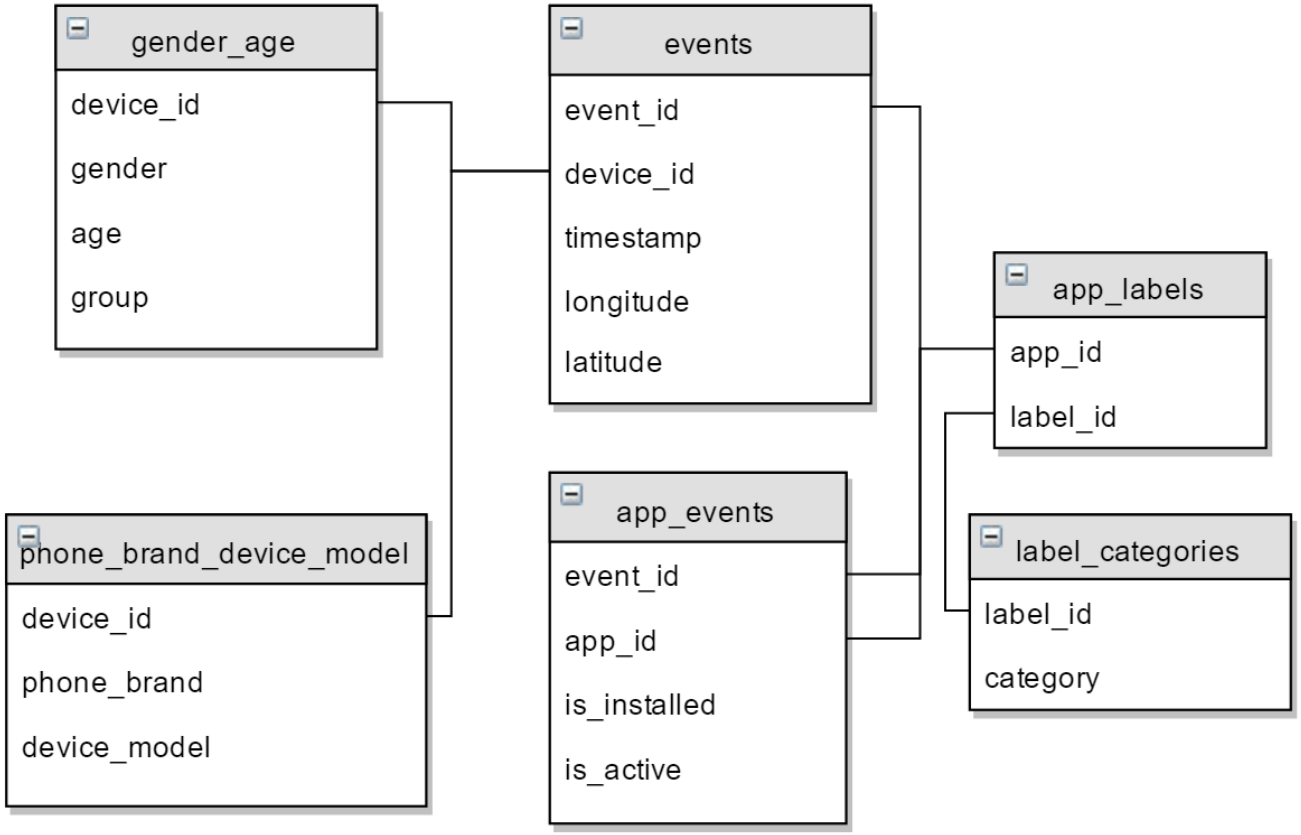
คำอธิบายชุดข้อมูล TalkingData on “Kaggle”



ข้อมูล Phone Brand Device Model ประกอบด้วย 3 สดมภ์ 187,245 แถว มีคำอธิบายตัวแปรดังนี้

- Device_id คือ หมายเลข (นิรนาม สำหรับข้อมูลชุดนี้) ที่เป็นเฉพาะของผู้ใช้แอพพลิเคชั่น สดมภ์นี้สามารถรวมกับ Gender_age ได้
- Phone_brand คือ แบรนด์ของโทรศัพท์ผู้ใช้ (ในประเทศจีนเท่านั้น) เช่น 三星 (Samsung) 美图 (meitu) และ 酷珀 (kupo) เป็นต้น
- Device_model คือ รุ่นของโทรศัพท์ผู้ใช้ (ในประเทศจีนเท่านั้น) เช่น 红米,Galaxy S4, 时尚手机 และ Galaxy Note 2 เป็นต้น

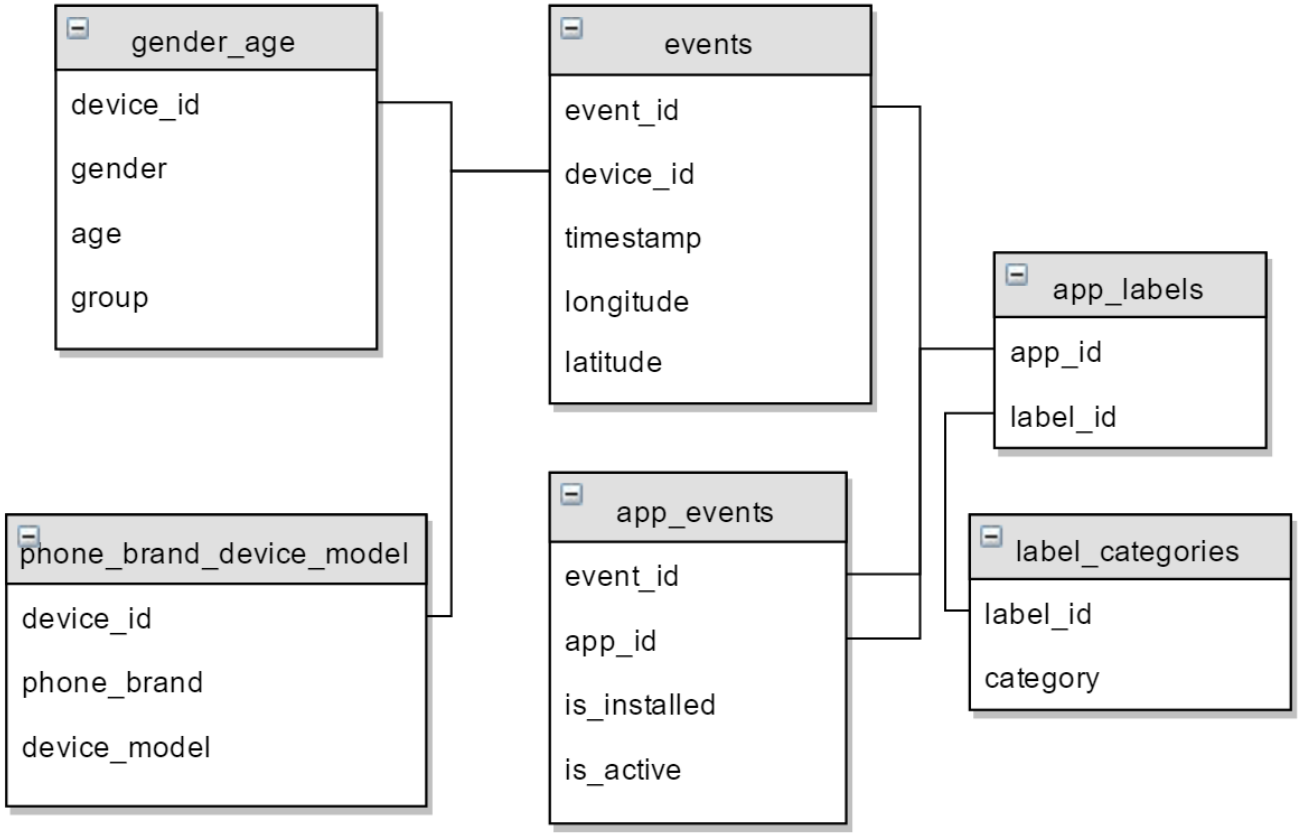
คำอธิบายชุดข้อมูล TalkingData on “Kaggle”



ข้อมูล Events ประกอบด้วย 5 สดมภ์ 3,252,950 แถว มีคำอธิบายตัวแปรดังนี้

- Event_id คือ รหัสการเกิดของเหตุการณ์การใช้แอพพลิเคชั่น
- Device_id คือ หมายเลข (นิรนาม สำหรับข้อมูลชุดนี้) ที่เป็นเฉพาะของผู้ใช้แอพพลิเคชั่น สดมภ์นี้สามารถรวมกับ Gender_age ได้
- Timestamp คือ วันและเวลาของการเข้าใช้งานแอพพลิเคชั่น
- Longitude คือ ลองจิจูดที่ TalkingData เก็บข้อมูลไว้จากการใช้แอพพลิเคชั่นของผู้ใช้งาน
- Latitude คือ ละติจูดที่ TalkingData เก็บข้อมูลไว้จากการใช้แอพพลิเคชั่นของผู้ใช้งาน

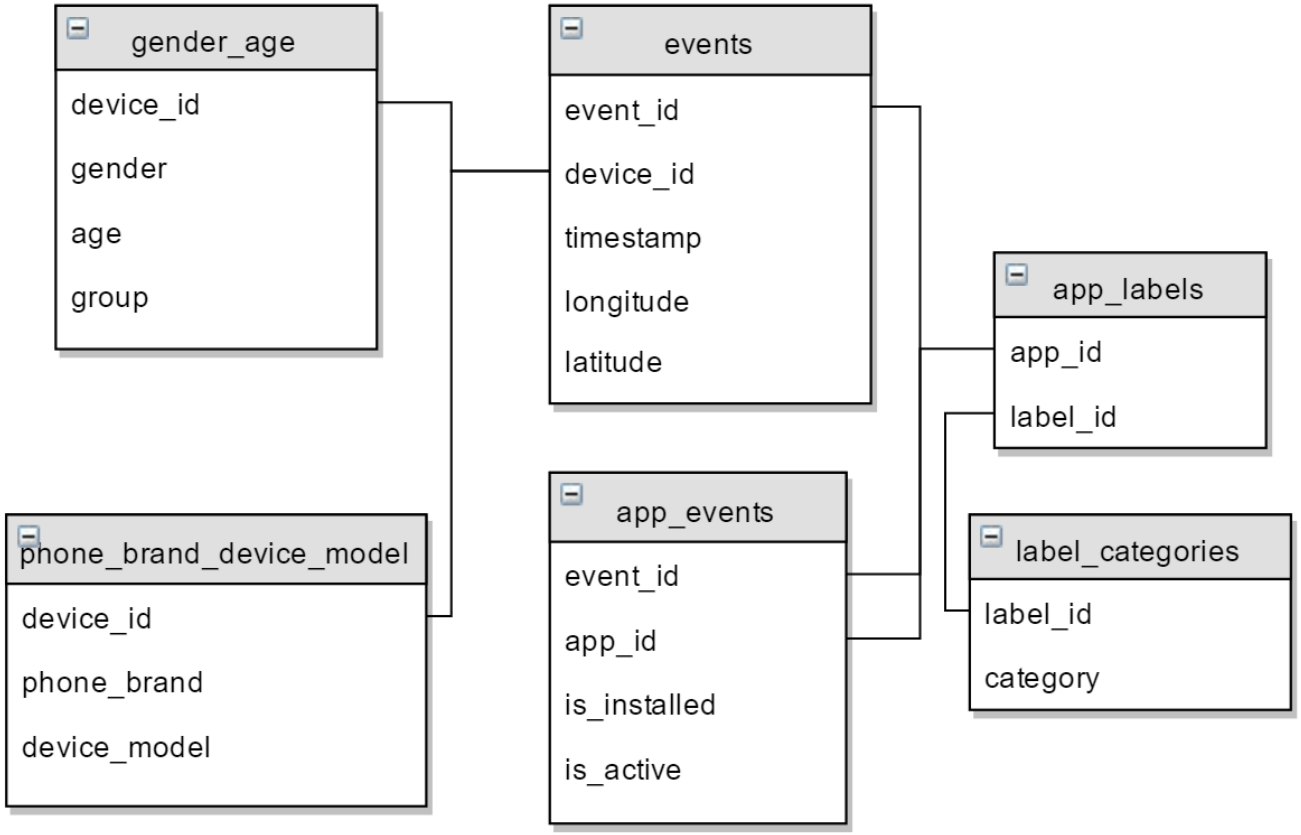
คำอธิบายชุดข้อมูล TalkingData on “Kaggle”



ข้อมูล App Events ประกอบด้วย 4 สดมภ์ 32,473,067 แถว
มีคำอธิบายตัวแปรดังนี้

- Event_id คือ รหัสการเกิดของเหตุการณ์การใช้แอพพลิเคชั่น สดมภ์นี้สามารถรวมกับ Events ได้
- App_id คือ รหัสเฉพาะของแอพพลิเคชั่นนั้นๆ
- Is_installed คือ แอพพลิเคชั่นได้รับการติดตั้งหรือไม่ (1 คือใช่ 0 คือไม่ใช่)
- Is_active คือ แอพพลิเคชั่นยังคง active อยู่หรือไม่จากการเก็บข้อมูลของ TalkingData ณ เวลานั้น (1 คือใช่ 0 คือไม่ใช่)

คำอธิบายชุดข้อมูล TalkingData on “Kaggle”



ข้อมูล App Labels ประกอบด้วย 2 สดมภ์ 459,943 แถว

มีคำอธิบายตัวแปรดังนี้

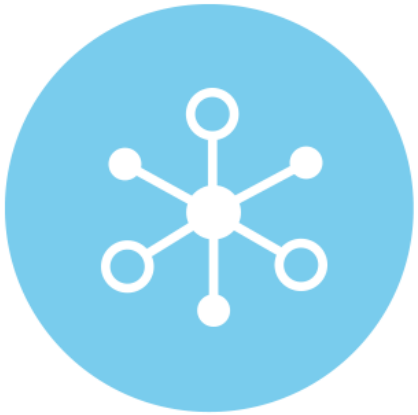
- App_id คือ รหัสเฉพาะของแอพพลิเคชั่นนั้นๆ สดมภ์นี้สามารถรวมกับ App Events ได้
- Label_id คือ รหัสลาเบลเพื่อระบุประเภทของแอพพลิเคชั่น

ข้อมูล Label_category ประกอบด้วย 2 สดมภ์ 930 แถว

มีคำอธิบายตัวแปรดังนี้

- Label_id คือ รหัสลาเบลเพื่อระบุประเภทของแอพพลิเคชั่น สดมภ์นี้สามารถรวมกับ App Labels ได้
- Category คือ หมวดหมู่ของแอพพลิเคชั่น เช่น game-Game themes, game-Art Style, Internet Banking และ Romance เป็นต้น

Team Dynamism and Battle Plan



Integrate

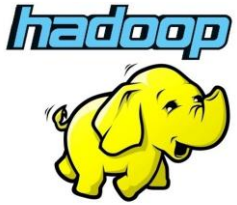
Analyze

Visualize

Mass Analytics Tools

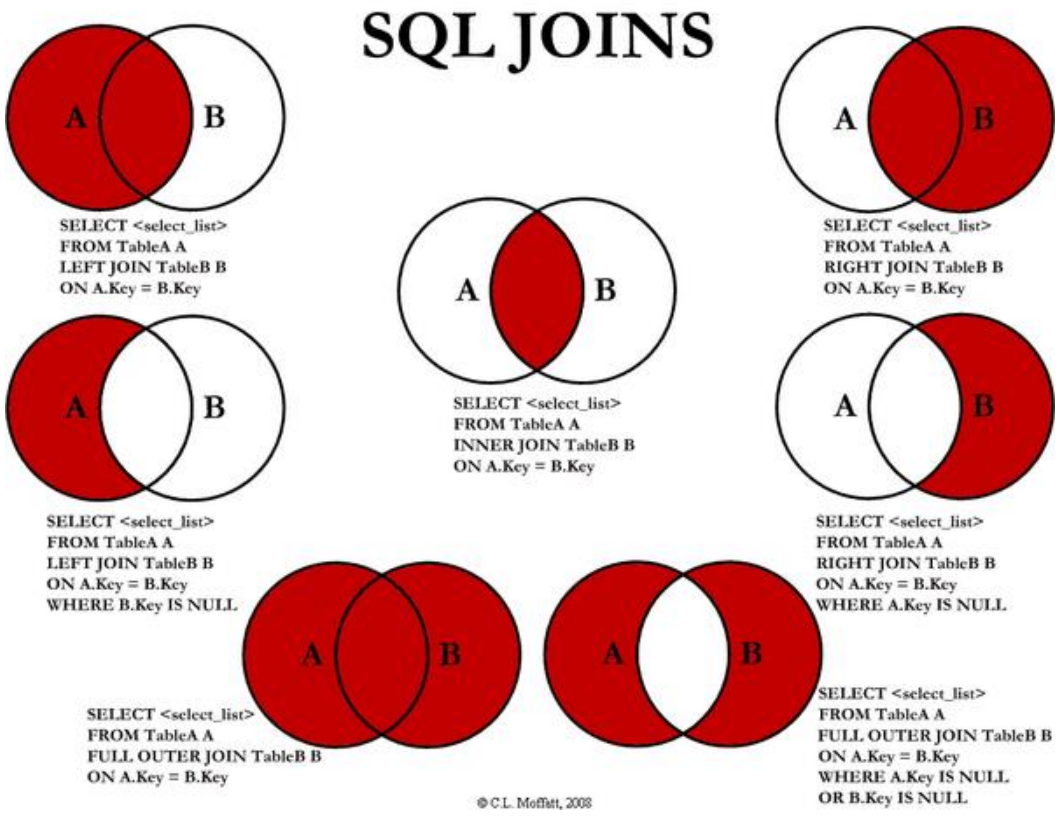
Analytics Tools

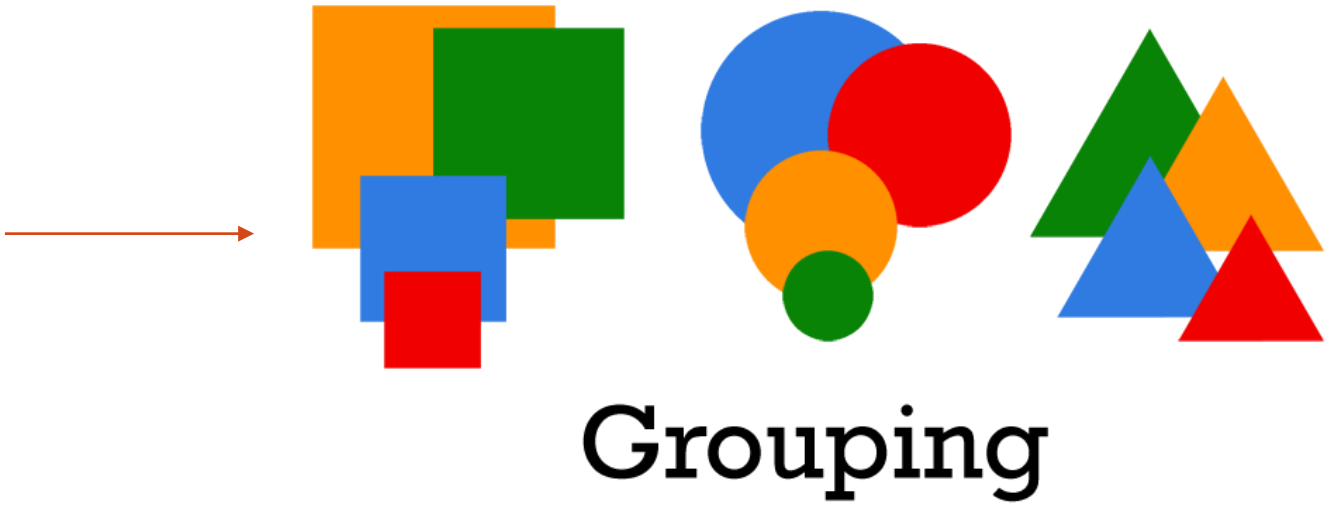
Cloud Computing & Server Management



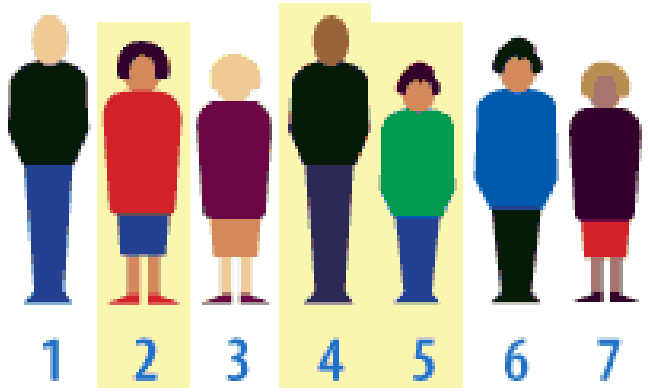
Visualization Tools

Data Cleansing



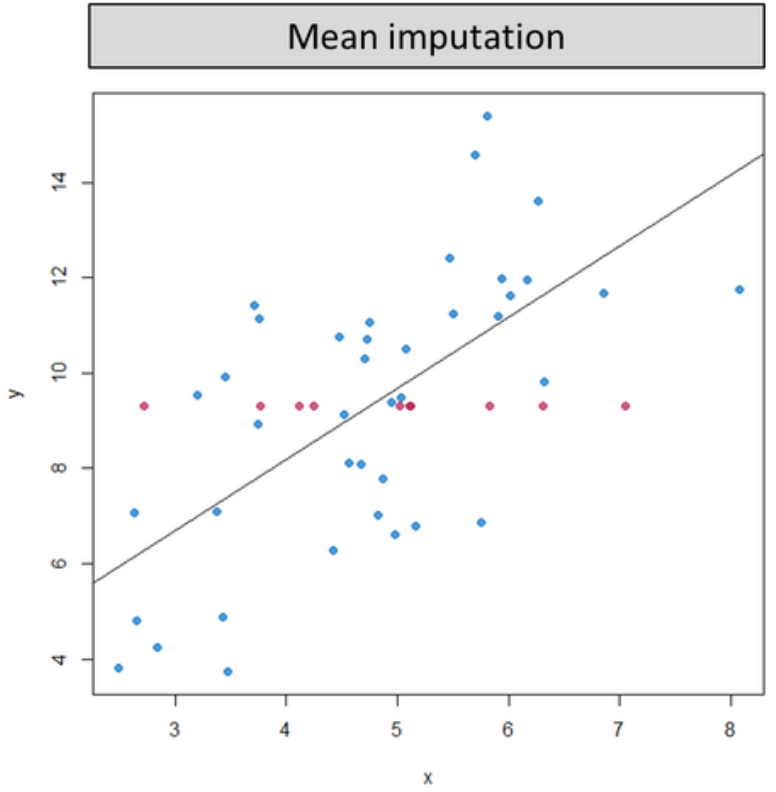


Data Cleansing



Assign Numbers, Auto-Generate Random Selections

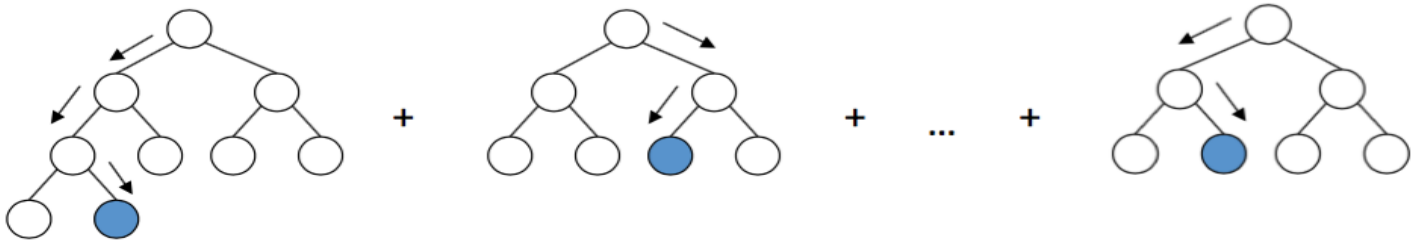
Vs.

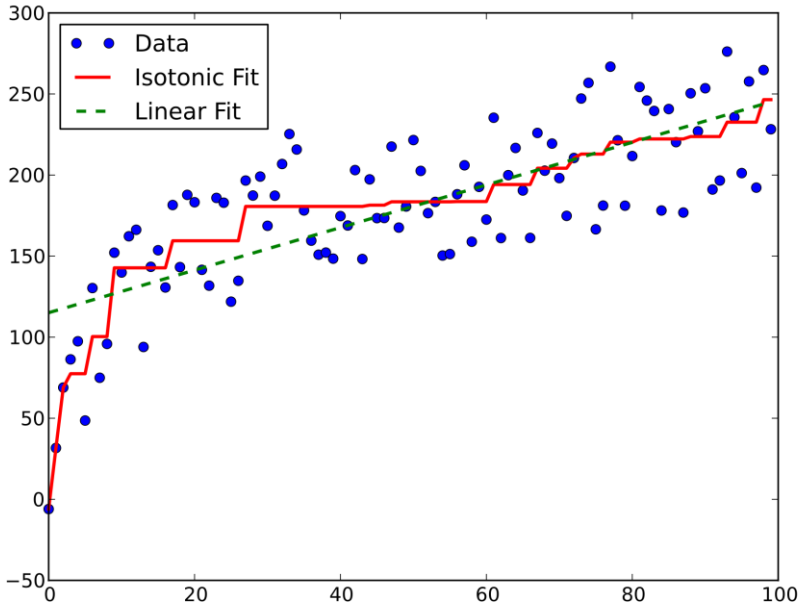


การนำเสนอการวิเคราะห์ข้อมูล

- Descriptive Analytics
  - Basic Visualization
  - Spatial Visualization
  - Network Visualization

- Predictive Analytics
  - Classification Algorithms (Gradient Boosting)
  - Regression Algorithms (Multivariate Linear)





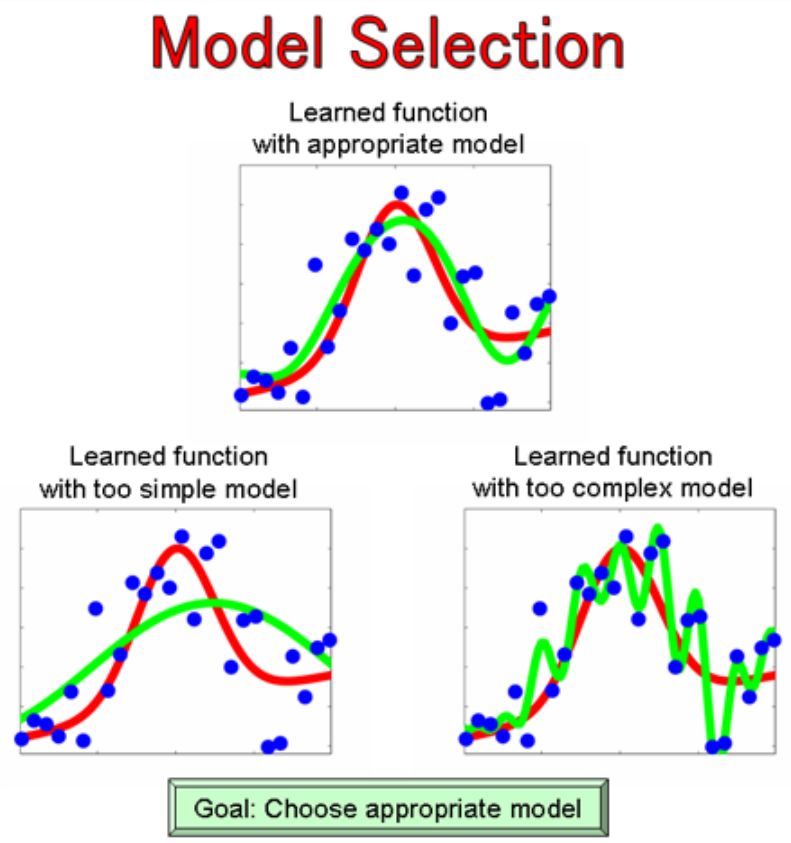
Data Insight

- เข้าใจกระบวนการเก็บข้อมูลของบริษัทโทรคมนาคมมากขึ้น อาจเป็นประโยชน์ต่อบริษัทในประเทศไทยหากต้องการวิเคราะห์ลูกค้าในรูปแบบที่คล้ายกันกับโจทย์นี้
- เข้าใจพฤติกรรมของผู้ใช้งานแอพพลิเคชั่นว่า ต้องการแอพพลิเคชั่นประเภทใด ใช้ช่วงเวลาใดของวันและช่วงอาทิตย์ จากการวิเคราะห์พบว่า คนเข้าใช้มือถือในเวลา 11:00 am. และ 11:00 pm. มากที่สุดและคนเข้าใช้วันอังคารมากที่สุด จากกราฟเส้นของเวลาการใช้ตามอาทิตย์ ข้อมูลดังกล่าวเป็นประโยชน์ต่อนักพัฒนาแอพพลิเคชั่นและนักการตลาดทั่วโลกในการตอบสนอง Demand ของผู้ใช้
- แบรนด์โทรศัพท์ยอดนิยม 3 อันดับแรกได้แก่ 小米, 三星, และ 华为 และโมเดลโทรศัพท์ 3 อันดับแรกได้แก่ 红米note, MI 3, และ MI 2S
- จากการวิเคราะห์แผนที่ของผู้ใช้งานแอพพลิเคชั่นทำให้สามารถ Traceback สถานที่การใช้งานของผู้ใช้แอพพลิเคชั่นในแต่ละกลุ่มตามเพศและอายุ รวมไปถึงแบรนด์โทรศัพท์และรุ่นโทรศัพท์มือถือ
- การทดสอบโมเดล Classification พบว่าปัจจัยที่สำคัญได้แก่จำนวนการลงแอพพลิเคชั่น จำนวนการใช้แอพพลิเคชั่น จำนวนเหตุการณ์การเข้าใช้ แบรนด์โทรศัพท์มือถือ และโมเดลโทรศัพท์มือถือ
- การทดสอบโมเดล Regression พบว่าปัจจัยสำคัญได้แก่ อายุ เพศ จำนวนเหตุการณ์การเข้าใช้ แบรนด์โทรศัพท์มือถือ และโมเดลโทรศัพท์มือถือ

What's Next?


Apache Zeppelin


Fast!
Statistics!
Scalable
Spark
+
Packages
Interactive Shell
Plots


Model Selection
Learned function
with appropriate model
Learned function
with too simple model
Learned function
with too complex model
Goal: Choose appropriate model


Scala


Spark
python
Fast Data Applications

Thank you!
Time for Q & A!