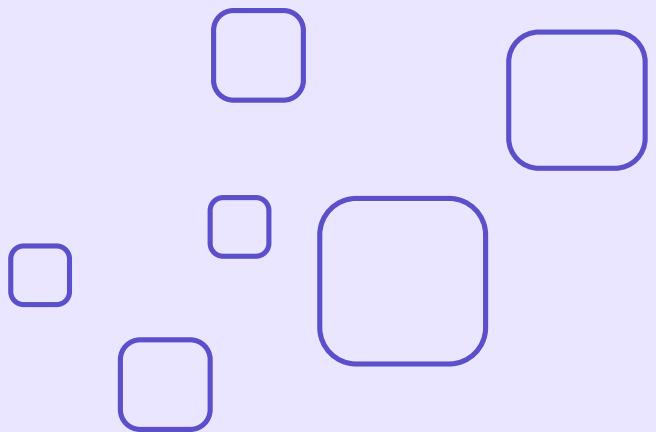


ROAD TO DATA ENGINEER 2.0 WORKBOOK



CHAPTER 2
*Data Cleansing with
Apache Spark*



ยินดีต้อนรับ สู่บทที่ 2

ในบทนี้ เราจะไปเรียนรู้พร้อมกันกับน้องໂຮດดี้ว่า:

- Data Cleansing หรือ การทำความสะอาดข้อมูล คืออะไร ทำไมต้องทำ
- Data Quality คืออะไร มีเครื่องมืออะไรบ้าง
- Exploratory Data Analysis (EDA) คืออะไร ทำอย่างไร มีเทคนิคอะไรบ้าง
- Data Anomalies แบบต่าง ๆ และวิธีแก้ไข
- Distributed Data Processing คืออะไร ทำงานอย่างไร
- Hadoop คืออะไร ทำงานอย่างไร
- Spark คืออะไร ทำงานอย่างไร



แล้วเจอกันในบทเรียนครับ / ค่ะ



แอดเพิร์ธ



แอดฟัน

สิ่งที่ต้องทำก่อน ทำ Workbook นี้

- ดูวีดิโอ "Class 3" ซึ่งจะสอน
 - Chapter 2 Data Cleaning
 - Workshop 2
- ก่อนเรียน Workshop 2 เปิดไฟล์ในบทเรียน "ไฟล์สำหรับเริ่มต้น Workshop 2"
- หลังเรียน Workshop 2 แล้ว หรือ เรียนแล้วพับปัญหาได้ สามารถเช็ค โค้ดที่ถูกต้องในบทเรียน "ไฟล์เฉลย Workshop 2" ได้
- หากพับปัญหาระหว่างเรียนก็อยู่นอก เหนือจากไฟล์เฉลย สามารถสอบถาม ได้ในบทเรียน "กระดานถาม-ตอบ Discussion Board" ทางกีบสอนจะรีบ มากตอบ

ด้าดูจบแล้ว
มาทำ Workbook นี้กันเลย >>



Data Cleansing (1/3)

Data Cleansing คืออะไร

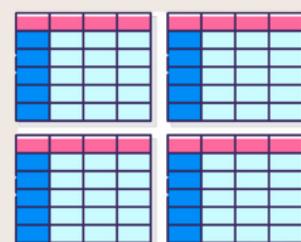
สาเหตุที่ต้องทำ Data Cleansing คือ

Data Cleansing (2/3)

Data Cleansing เป็นงานที่ยากและใช้เวลา เพราะ



Data Quality ก็ดี ควรจะประกอบด้วย



Data Scientist ใช้เวลา 60%
ในการคัดสิ่งข้อมูลเลขณะ

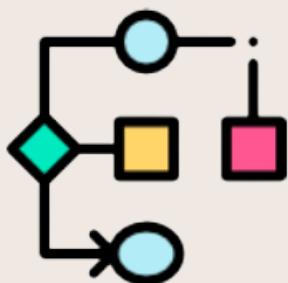


Data Cleansing (3/3)

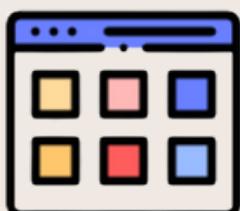
3 เครื่องมือที่ช่วยพัฒนา Data Quality ให้ดีขึ้น คือ



ช่วยพัฒนา Data Quality ด้วยการ...



ช่วยพัฒนา Data Quality ด้วยการ...



ช่วยพัฒนา Data Quality ด้วยการ...

EDA - Exploratory Data Analysis (1/4)



Data Profiling คือ

ตัวอย่างข้อมูลที่คำนวณในการทำ Data Profiling เช่น

Count หมายถึง...

Mean หมายถึง...

Std หมายถึง...

Min หมายถึง...

25% หมายถึง...

50% หมายถึง...

75% หมายถึง...

Max หมายถึง...

EDA - Exploratory Data Analysis (2/4)



ประเภทของ EDA แบ่งตาม...

1 2 3



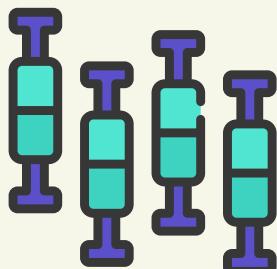
และ



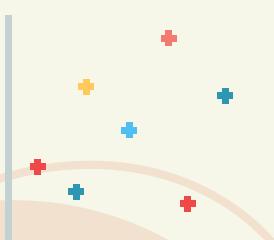
EDA แบบใช้กราฟฟิก ทำได้หลายแบบ คือ



แสดงข้อมูลเกี่ยวกับ...



แสดงข้อมูลเกี่ยวกับ...

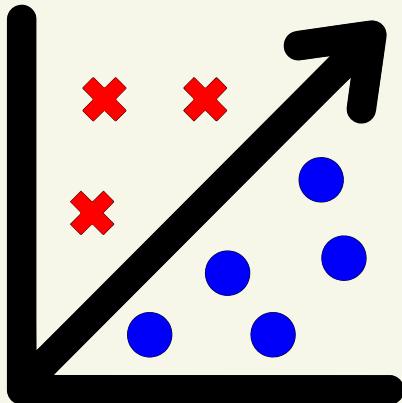


แสดงข้อมูลเกี่ยวกับ...

EDA - Exploratory Data Analysis (3/4)



Data Anomaly คือ



ประเภทของ Data Anomaly แบบต่าง ๆ ได้แก่

1. Syntactical Anomalies

ตัวอย่างข้อมูล...

วิธีแก้ไข...

2. Semantics Anomalies

ตัวอย่างข้อมูล...

วิธีแก้ไข...

3. Missing Values

หมายถึง...

วิธีแก้ไข...

4. Outliers

หมายถึง...

วิธีการเช็ค...

EDA - Exploratory Data Analysis (4/4)



Regular Expression ช่วยในการค้นหา Data Anomalies ประเภท

และ

Regular Expression มีในภาษาโปรแกรมมิ่งต่าง ๆ เช่น Python, R, Java, JavaScript หรือไม่

มี

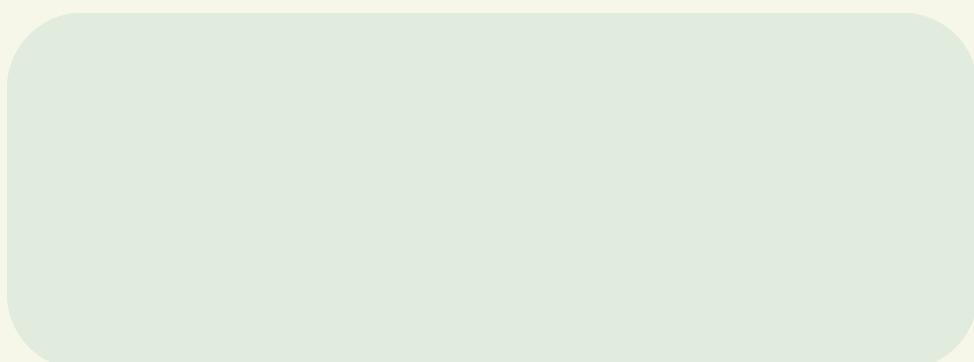
ไม่มี

Regular Expression ต่อไปนี้ มีความหมายว่า...

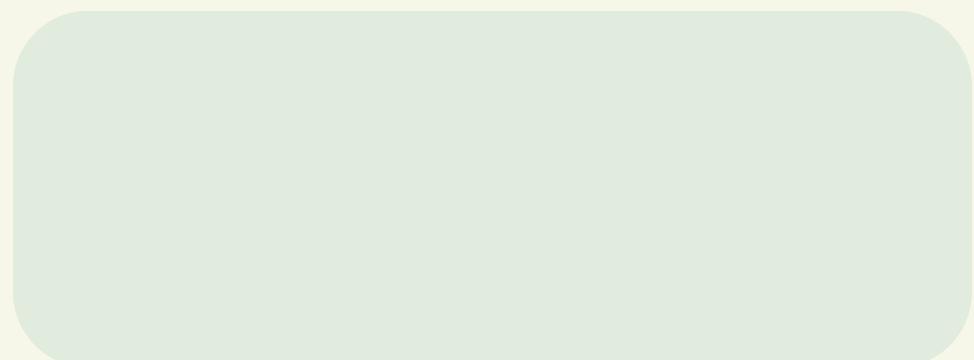


Tip: ตรวจสอบเช็คว่า Regular Express ไหน ทำงานยังไง ด้วยเว็บ regex101.com

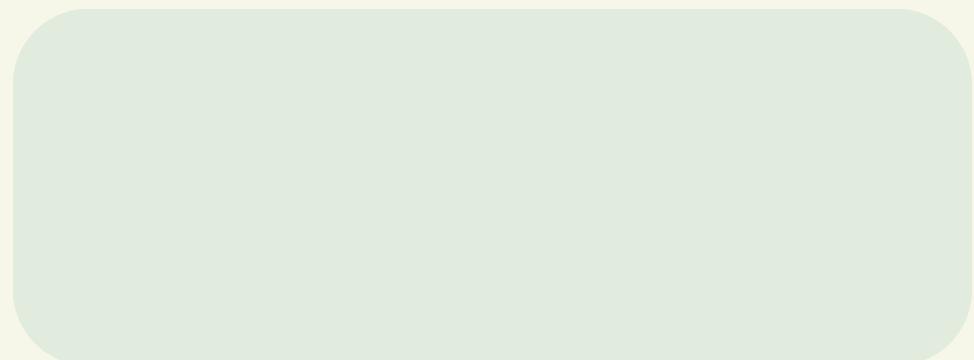
[A-Z]+



[a-zA-Z]+



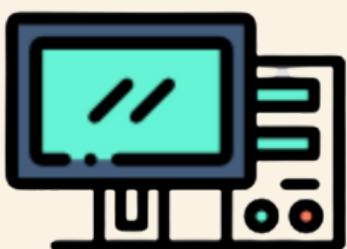
[a-zA-Z0-9]+



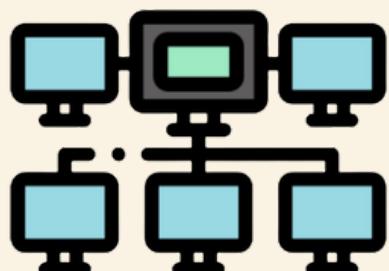
Distributed Data Processing (1/3)

Distributed Data Processing คือ...

Standalone (1 Node)



Cluster (2+ Nodes)



Hadoop คือ...

ข้อดีของ Hadoop คือ Commodity Hardware หมายถึง...



Distributed Data Processing (2/3)

วิธีการทำงานของ Hadoop คือ...

Input

วิธีการทำงาน...

Map Tasks

วิธีการทำงาน...

Reduce Tasks

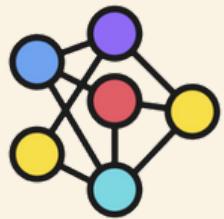
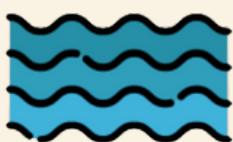
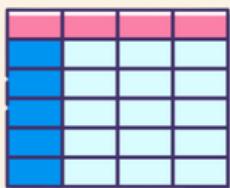
วิธีการทำงาน...

Output

วิธีการทำงาน...

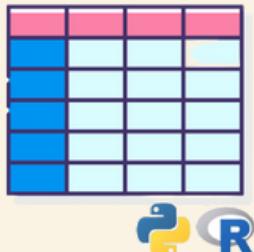
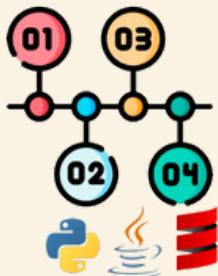
Apache Spark คือ...

ส่วนเสริมของ Spark ที่นำเสนอใน มีดังนี้...



Distributed Data Processing (3/3)

ประเภทข้อมูลใน Spark มีดังนี้



ข้อมูลแบบ RDD กับ DataFrame
แตกต่างกันอย่างไร

Transformation กับ Action ใน RDD
แตกต่างกันอย่างไร

Spark SQL ทำงานอย่างไร

Databricks คือ...



databricks



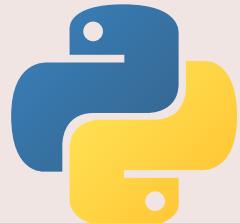
Workshop 2: Data Cleansing with Spark (1)

Pandas กับ Spark แตกต่างกันอย่างไร



คำสั่ง Spark สำหรับอ่านไฟล์ CSV คือ...

Package Python สำหรับสร้าง Chart ที่ Interactive ซึ่ว่า...



คำสั่ง Spark สำหรับแสดงข้อมูล คือ...

คำสั่ง Spark สำหรับการเลือกข้อมูลด้วย Regular Expression คือ...

ข้อมูลบน Spark สามารถแปลงเป็น Pandas DataFrame ได้

ใช้

ไม่ใช้

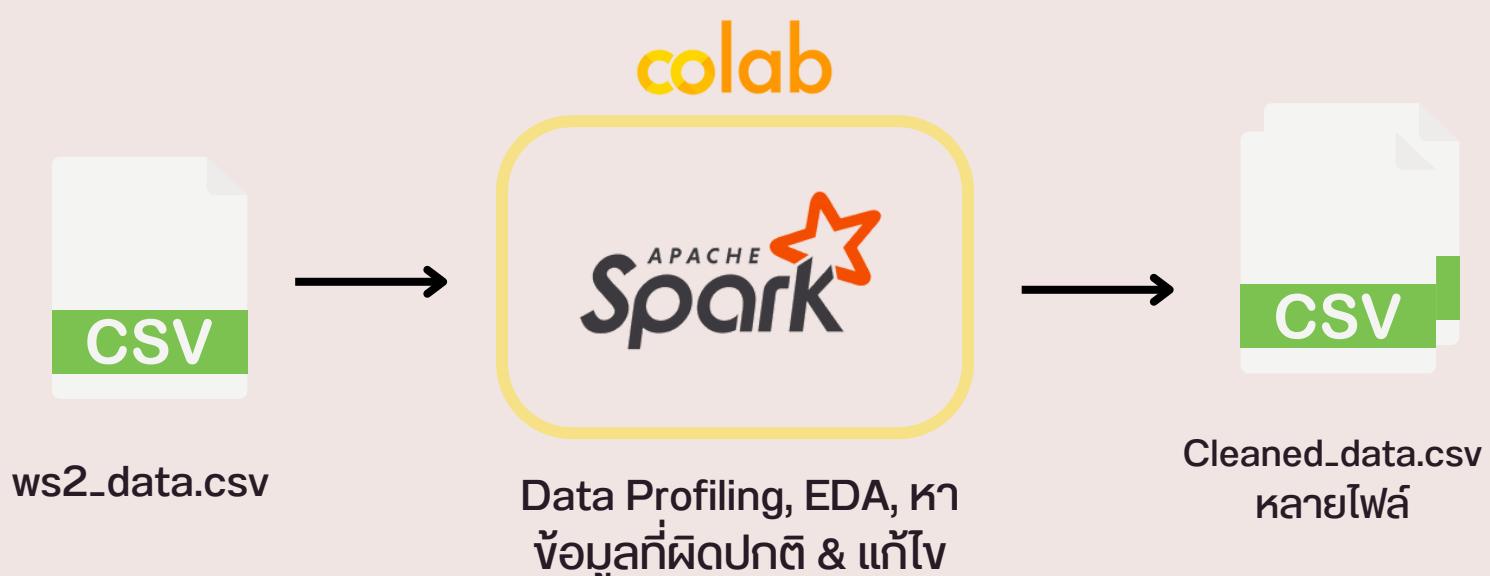




Workshop 2: Data Cleansing with Spark (2)

คำสั่ง coalesce ใน Spark ใช้สำหรับ...

สรุปสิ่งที่เราทำใน Workshop 2 คือ



ยินดีด้วย คุณจบบทที่ 2 แล้ว!



คุณพร้อมไปต่อบทที่ 3 เรื่อง Basic Cloud แล้ว
ในบทต่อไปเราจะเรียนรู้เรื่อง Public / Private / Hybrid Cloud Computing และวิธีเขียน Bash Script ที่ได้ใช้ในงานบ่อย



บทนี้หนักนิดนึง
บทหน้าซีลแล้ววว!

Recap สิ่งที่เราได้เรียนรู้ จากบทที่ 2:

- Data Cleansing หรือ การกำกับความสะอาดข้อมูล คืออะไร ทำไมต้องทำ
- Data Quality คืออะไร มีเครื่องมืออะไรบ้าง
- Exploratory Data Analysis (EDA) คืออะไร ทำอย่างไร มีเทคนิคอะไรบ้าง
- Data Anomalies แบบต่าง ๆ และวิธีแก้ไข
- Distributed Data Processing คืออะไร ทำงานอย่างไร
- Hadoop คืออะไร ทำงานอย่างไร
- Spark คืออะไร ทำงานอย่างไร