



BIG DATA Analytics

เส้นทางสู่การพัฒนา
ระบบวิเคราะห์ข้อมูลขององค์กร

ธรากรณ์ พรหมวิจิตร

สารบัญ

บทนำ : ขอต้อนรับสู่โลกของข้อมูล	11
แคปโตนึงจะเรียกว่า ‘บิก’	19
จากอดีตจนมาเป็น Big Data	19
ความสำคัญของ Big Data	23
ข้อมูลในยุคดิจิทัล	25
แหล่งที่มาของ Big Data	28
คุณลักษณะ 3V, 4V, 5V, 6V ของ Big Data	41
Gartner กับ 3V	42
IBM กับคำนิยาม 4V	44
Yun กับ 5V	44
Microsoft เติมเต็มเป็น 6V	44
บทบาทของ IoT กับ Big Data	47
ลิ้นแฉ่วเน็ตเวิร์กเกี่ยวข้องกับทุกสิ่ง	47
IoT เป็นโลกแห่งพลีเคชัน	49
IoT กับธุรกิจค้าปลีก	50
IoT กับหน่วยงานกลุ่มธนาคาร	51
Things ใน Smartphone	53
บ้านอัจฉริยะ	59
เมืองอัจฉริยะ (Smart City)	59
IoT ในงานอุตสาหกรรม หรือ IIOT (Industrial Internet of Things)	60
ฟาร์มอัจฉริยะ (Smart Farming)	62
อุปกรณ์อัจฉริยะ	63
IoT กับเทคโนโลยีสแต็ก (IoT Technology Stack)	68
IoT กับเทคโนโลยีการสื่อสาร	69
IoT กับคลาวด์แพลตฟอร์ม	71

ตัวอย่าง Industrial IoT ของ Google และ Kaa ท้ายบท	74 76
BDA = ML + CC	79
ความเป็นมาของเทคโนโลยี BDA	83
การเรียนรู้ด้วยตัวเองของคอมพิวเตอร์ (Machine Learning - ML)	87
การให้บริการคลาวด์ อีกหนึ่งของไอที (Cloud Computing - CC)	93
โมเดลให้บริการของคลาวด์	96
บริการคลาวด์รายใหญ่ของโลก	101
BDA กับ CC	109
BDA กับ ML	113
ตัวอย่าง : LinkedIn ใช้ BDA อย่างไร	115
ตัวอย่าง : ตลาดหลักทรัพย์แห่งประเทศไทยใช้แพลตฟอร์ม BDA ให้บริการที่รวดเร็ว	117
สรุป ML + CC -> BDA	119
เทคนิคฐานข้อมูลสำหรับ Big Data	121
ภาพรวมสูงของสถาปัตยกรรม Big Data	122
เราจัดการข้อมูลกับอย่างไร	123
ถ้าขนาดของข้อมูลที่ตีเป็นกบถ้ำ	129
การเก็บข้อมูลแบบมีโครงสร้าง	130
การเก็บข้อมูลแบบไม่มีโครงสร้าง	132
ฐานข้อมูลสำหรับ Big Data	133
ฐานข้อมูลแบบ NoSQL	134
สถาปัตยกรรมของ NoSQL	135
ทฤษฎีของ CAP Theorem	138
โมเดลสี่แบบของ NoSQL	140
ฐานข้อมูลแบบ In-Memory Data Fabric	155
ระบบ Data Storage แบบ In-memory	156
ระบบ Big Data Analytics (BDA) แบบ In-memory	157

สถาปัตยกรรมที่เก็บไฟล์แบบกระจาย (Distributed File System - DFS)	158
การเตรียมข้อมูลสำหรับ Big Data analytics (BDA)	162
ตัวอย่าง : ฐาน NoSQL ของ Microsoft Azure	165
ท้ายบท	167
เทคโนโลยีสำหรับ Big Data	169
อะไรคือ Batch Processing	171
จุดเปลี่ยนสำคัญเริ่มมาจาก Google File System (GFS)	172
ฮาดูป (Hadoop Distributing Computing - HDFS)	174
ทำไมต้องฮาดูป	175
กรอบการทำงานของฮาดูป (Hadoop Framework)	176
เป้าหมายการออกแบบของฮาดูป	177
สถาปัตยกรรม Master/Slave	178
คุณลักษณะของฮาดูป โดยอิงต้นแบบ	184
รูปแบบไฟล์ที่ใช้กับฮาดูป	186
การประมวลผลแบบขนานด้วย MapReduce	188
ตัวอย่างโปรแกรม MapReduce เพื่อคำนวณราคารวม	190
ตัวอย่างโปรแกรม MapReduce เพื่อนับคำ	194
YARN หน่วยปฏิบัติการของ Hadoop	196
ระบบนิเวศสำหรับ Big Data	199
กรอบการทำงานพื้นฐานระบบนิเวศของฮาดูป	200
ระบบเข้าถึงและเรียกใช้ข้อมูล (Data Access & Query)	202
ระบบจัดเก็บข้อมูล (Data Stores)	210
ระบบเชื่อมและรวมข้อมูล (Data Integration)	213
ระบบช่วยประสานงาน (Coordination and Orchestration)	217
โปรแกรมเรียนรู้ด้วยตัวเอง (Machine Learning)	221
ฮาดูปในเชิงพาณิชย์	223
ตัวอย่าง : บริษัท Nationwide ที่ใช้เทคโนโลยีด้าน BDA จาก IBM	228
ท้ายบท	233

การประมวลผลข้อมูลแบบเรียลไทม์เท่าเวลาจริง	235
เริ่มยุคโพรเซสซิงที่ถ่วงเร็ว	236
Event (เหตุการณ์)	238
Event Processing (การประมวลผลเหตุการณ์)	239
Data Stream or Event Stream (ข้อมูลสตรีม)	239
Data Stream Management (การจัดการข้อมูลสตรีม)	240
Data Stream Analytics (การวิเคราะห์ข้อมูลสตรีม)	240
Real-Time Analytics (การวิเคราะห์แบบเรียลไทม์เท่าเวลาจริง)	241
การประมวลผลข้อมูลสตรีมแบบเรียลไทม์เท่าเวลาจริง (Real-Time Stream Processing)	242
หลักการประมวลผลเหตุการณ์ที่ซับซ้อน (Complex Event Processing (CEP) Framework)	243
คุณลักษณะที่จำเป็นของ CEP	243
การทำงานของ Stream Processing ของ CEP	246
แพลตฟอร์ม Stream Analytics ในเชิงพาณิชย์	247
แพลตฟอร์ม Stream Analytics แบบ Open Source	248
SPARK	249
สถาปัตยกรรมของ Spark	252
Ecosystem ของ Spark	253
ความต่างของ Spark และ Hadoop	255
Storm	257
สถาปัตยกรรมของ Storm	257
Flink	260
สถาปัตยกรรมของ Flink	260
ตารางเปรียบเทียบ Spark, Storm และ Flink	263
การนำทางข้อมูลสำหรับระบบประมวลผลข้อมูลสตรีม	265
Message System	265
Kafka	266
สถาปัตยกรรมของ Kafka	267
Flume	268

สถาปัตยกรรมของ Flume	269
Big Data Stack และแพลตฟอร์ม	272
Google Cloud Services	273
Amazon Web Services	274
Microsoft Azure	276
Cloudera	277
แผนของไมโครซอฟท์ สำหรับบริการด้าน Big Data Analytics (BDA)	277
ตัวอย่าง : สถาปัตยกรรม Big Data ของ PayPal	283
ท้ายบท	285
การจัดการ Big Data ในองค์กรขนาดใหญ่	287
สถานการณ์ข้อมูลในองค์กรธุรกิจใหญ่	287
อินเทอร์เน็ต	289
อินเทอร์เน็ต	289
Data Warehouse	291
File Store	292
สถานการณ์ข้อมูลปัจจุบันขององค์กร Enterprise	292
Enterprise Data Lake คืออะไร	295
แนวคิดการสร้าง Enterprise Data Lake	297
ความต่างของ Data Lake และ Data Warehouse	300
แลมบ์ด้า (Lambda) สถาปัตยกรรมเพื่อผลักดัน Data Lake	301
Lambda คืออะไร	302
เลย์เออร์ของการรวบรวมข้อมูล	304
เลย์เออร์การจัดคิวงาน (Messaging Layer)	306
เลย์เออร์ส่งข้อมูลเข้า (Data Ingestion Layer)	306
เลย์เออร์การประมวลข้อมูลแบบแบตช์ (Batch Layer)	307
เลย์เออร์การประมวลข้อมูลแบบเรียลไทม์ (Speed Layer)	308
เลย์เออร์จ่ายข้อมูลพร้อมใช้ (Serving Layer)	308
เลย์เออร์ฐานเก็บข้อมูล (Data Storage Layer)	309
สรุป Lambda Architecture	309

ตัวอย่าง : SmartNews หรือสำนักข่าวหุ่นยนต์โดยใช้ Lambda Architecture ของ AWS	310
การออกแบบ	312
สถาปัตยกรรม Serverless Computing บนคลาวด์	317
ท้ายบท	321

จาก Big Data สู่ Smart Data	323
ความแตกต่างระหว่าง Data Mining และ Data Science	325
บทค้นพบความหมายจากข้อมูลความคิดเห็นของประชาชน	326
สถาปัตยกรรมของระบบค้นพบความหมายของข้อมูล (Data and Opinion Mining - DOM)	328
บทบาทของ MapReduce	329
การออกแบบระบบคิดเห็นภาษาไทยของประชาชน	329
การวิเคราะห์ความรู้สึก (Sentiment Analysis)	333
การจัดกลุ่มประโยคเพื่อหาข้อสรุป (Sentence Clustering-Based Summarization)	338
การวิเคราะห์ผู้มีอิทธิพลทางสังคม (Influencer Analysis)	343
แอปพลิเคชัน BDA ที่เกี่ยวกับ IoT	345
1. การรวบรวม Big Data จาก IoT	346
2. BDA สำหรับ IoT	348
3. IoT Application: การตรวจจับพฤติกรรมการจับรถของคนขับ	353
ใช้ GPS ในการตรวจจับพฤติกรรมการจับรถคนขับ	356
วิเคราะห์การใช้เว็บแอปพลิเคชันจากข้อมูล Logs ขององค์กรขนาดใหญ่	358
แนวทางการออกแบบ	360