BIG DATA

張晃崚 CCIE #13673
區域技術處
麟瑞科技

Google, Amazon, Facebook, Twitter

- 共通的成功要素：資料分析
- Google
  - 每個月 900 億筆網路搜尋、處理 600 PB 資料
  - Google Search 自動建議、自動修正
- Amazon
  - 產品推薦系統
- Twitter、Facebook
  - 您可能認識的人

Non-IT Company

- Walmart
  - 每小時 100 萬筆交易
  - 庫存與定價優化
  - BI 分析
  - 關聯式資料庫
- Google, Facebook
  - 非結構化資料
  - 資料多樣性

Another Big Data Examples

- Internet of Everything
- Smart Grid
- Smart City

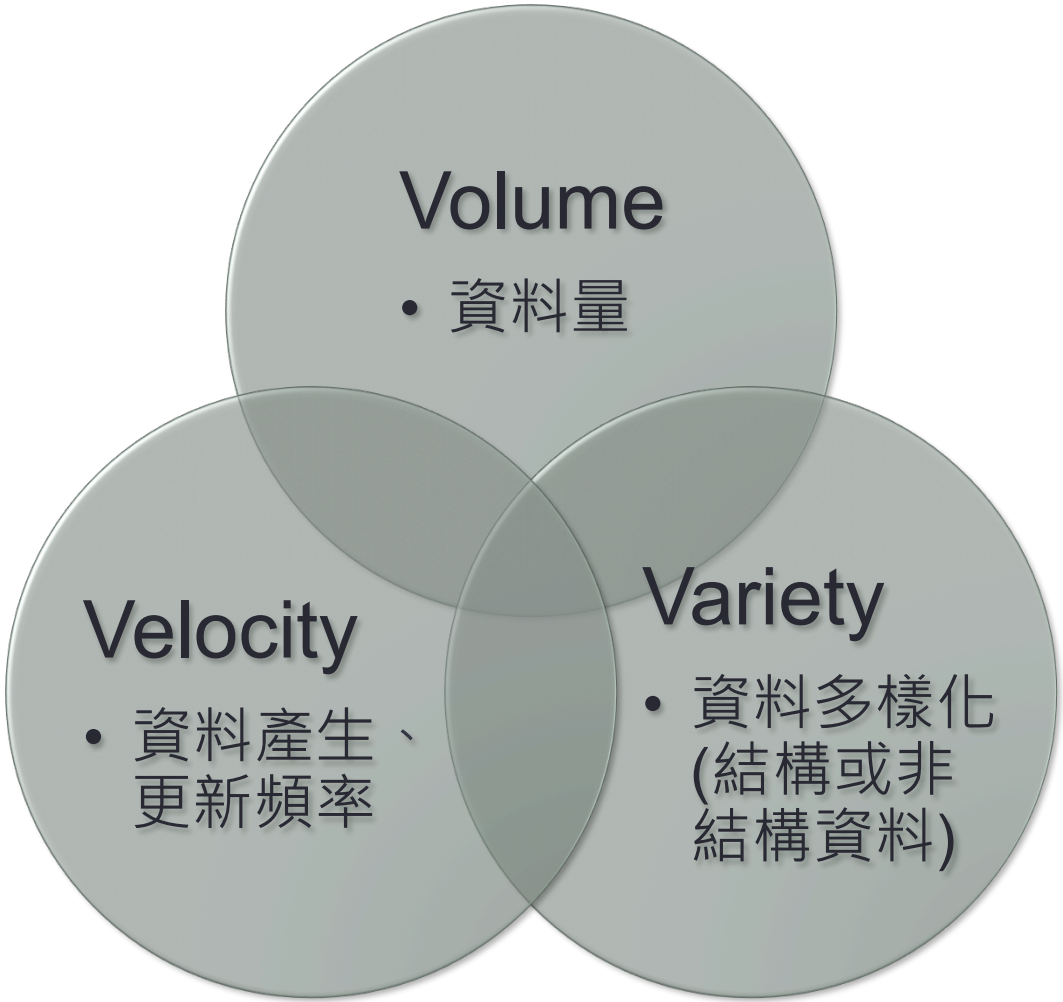
Data is the New Oil

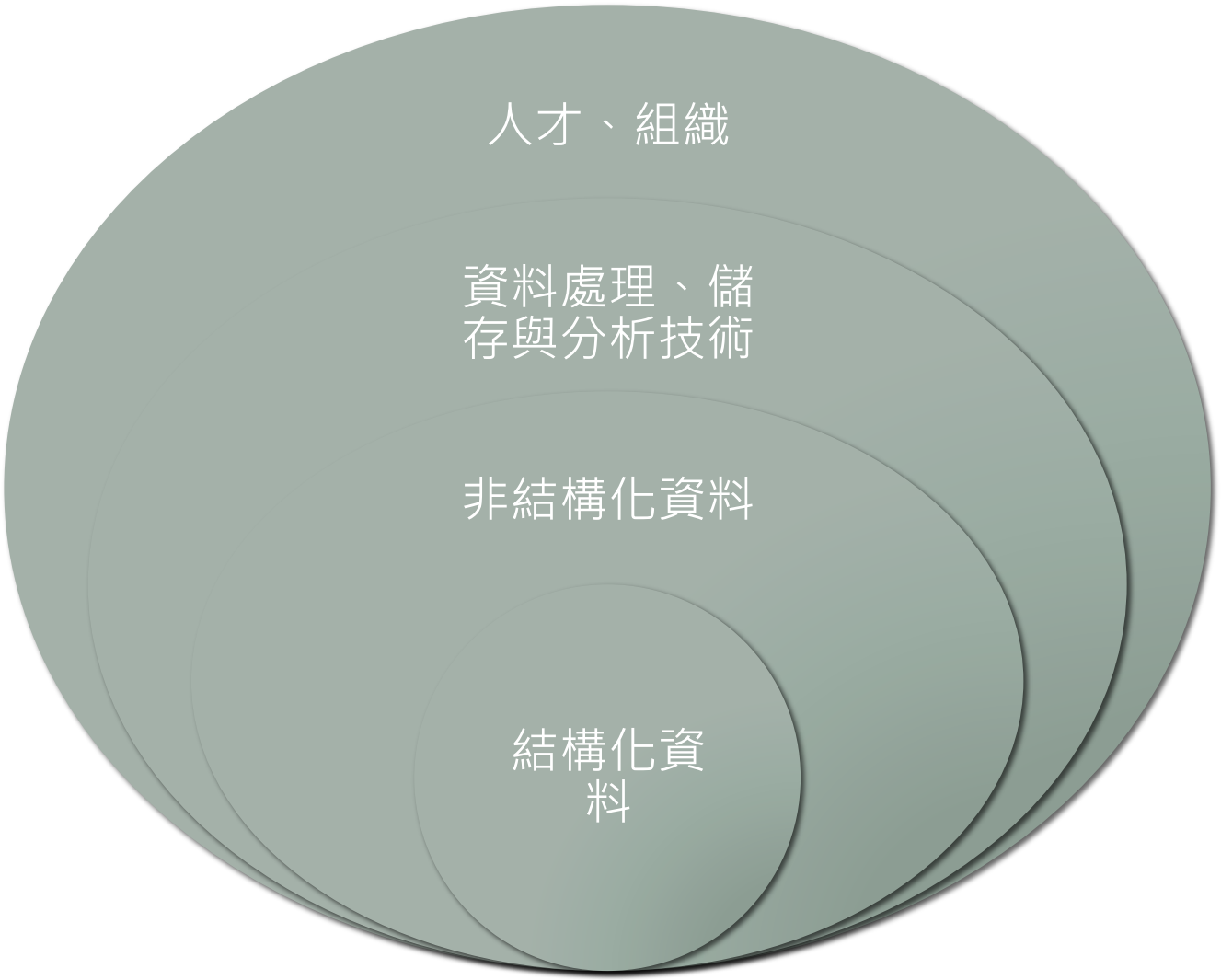
- Walmart 分析 social media 的文章來調整產品線與庫存
- 悠遊卡、NFC 追蹤消費記錄

資料洪流 Data Deluge

- Feb 2010 Economist: “The Data Deluge”
- Big Data:
  - 現有一般技術難以管理的大量資料群
  - 因資料量的增加，導致 Query 資料的回應時間超過容許範圍的大量資料。

巨量資料的特性




人才、組織
資料處理、儲存與分析技術
非結構化資料
結構化資料

Why Now?

- 巨量資料的民生化
- 軟硬體技術進化與提升
  - 電腦性價比提升
  - 儲存裝置價格下滑
- Hadoop 分散式處理技術問世
- 雲端服務的普及
  - EC2, S3, EMR

From Past to Future

- 產品賣出去了、一位客戶解約了
- 為什麼這個產品賣出去了？為什麼顧客流失了？
- 點擊串流
- 社群客戶關係管理
- O2O (Online to Offline)

Big Data Example

- Caesars Entertainment
- Disney World
- Esunbank

What is Hadoop?

- Google MapReduce
  - Hadoop MapReduce
- Google BigTable
  - HBase
- Google File System
  - Hadoop Distributed File System (HDFS)

Hadoop Distribution

- Cloudera
- HortonNetworks
- MapR
- IBM

NoSQL

- “No SQL” or “Not only SQL”?

	RDBMS	NoSQL
資料形態	結構化資料	非結構化資料
資料一致性	嚴格一致性	最終一致性
擴充性	Scale-up	Scale-out
伺服器	單一或較小叢集	分散式
耐故障性	成本高	成本低
查詢語言	SQL	NoSQL
資料量	較小	較大

Data Processing in Big Data

- ETL
  - Extract -> Transform -> Load
  - CRM/ERP -> Data Warehouse -> BI
  - HBase -> HIVE -> Mahout
- Data warehouse
  - Massively Parallel Processing
  - Shared Nothing
  - Compressed
  - Commodity hardware
  - Integrated appliance
  - Hadoop support

Real-time Data Processing

- RDBMS vs. Real-time Data Processing
- Process in RAM first
- Differential processing
- IBM InfoSphere Streams, Oracle CEP
- Facebook Data Freeway, Yahoo! S4, Twitter Storm, LinkedIn Kafka, Wal-Mart Muppet
  - Facebook process 9GB per second
- Smart city, smart grid

Other Technologies

- Machine learning
- Data mining
- Data clustering
- Neural network
- Recursive analysis
- Decision tree
- Relational analysis

Other Technologies

- Natural language processing
- Semantic search
- Link mining
- A/B test

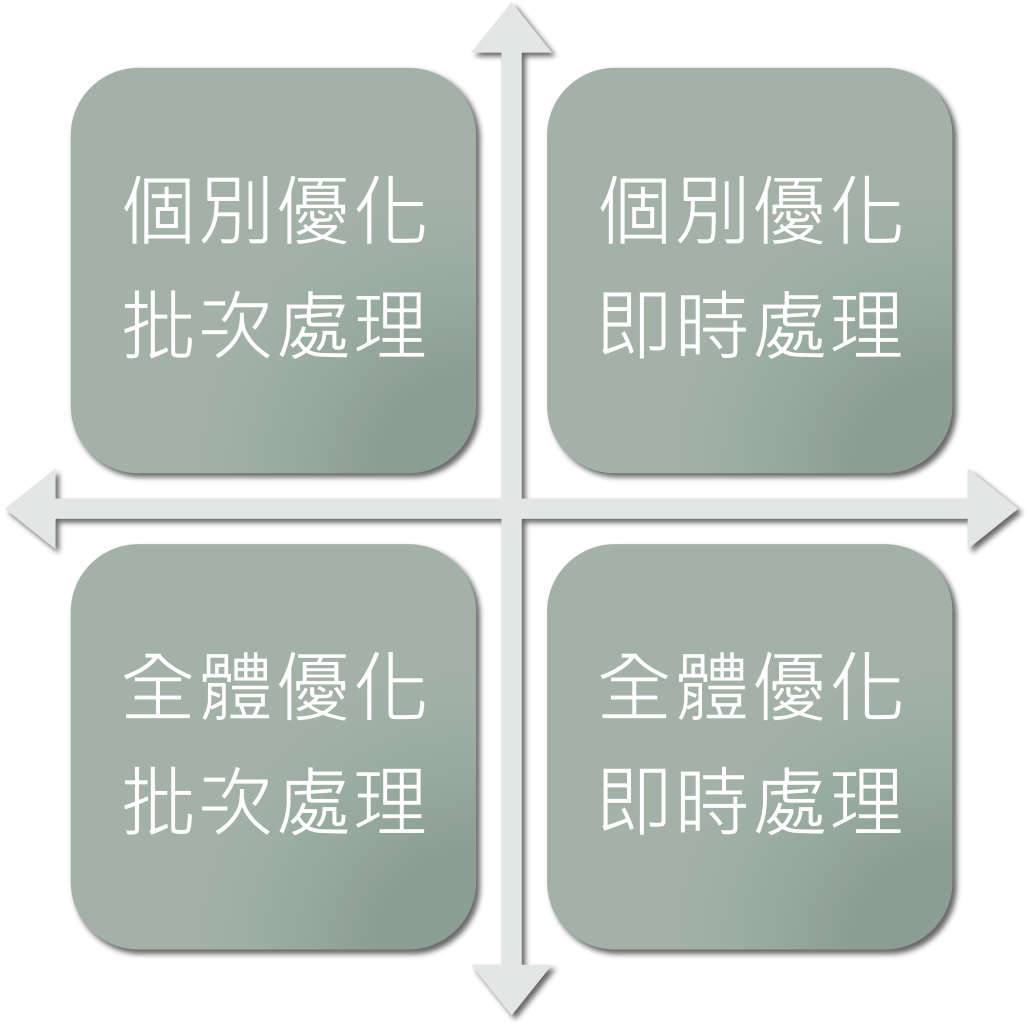
Big Data 應用企業

- eBay
- Zynga: viral coefficient, 3 click rule, first time user experience
- Centrica (UK): Smart Grid, in home display
- Catalina Marketing: 收銀台優待券

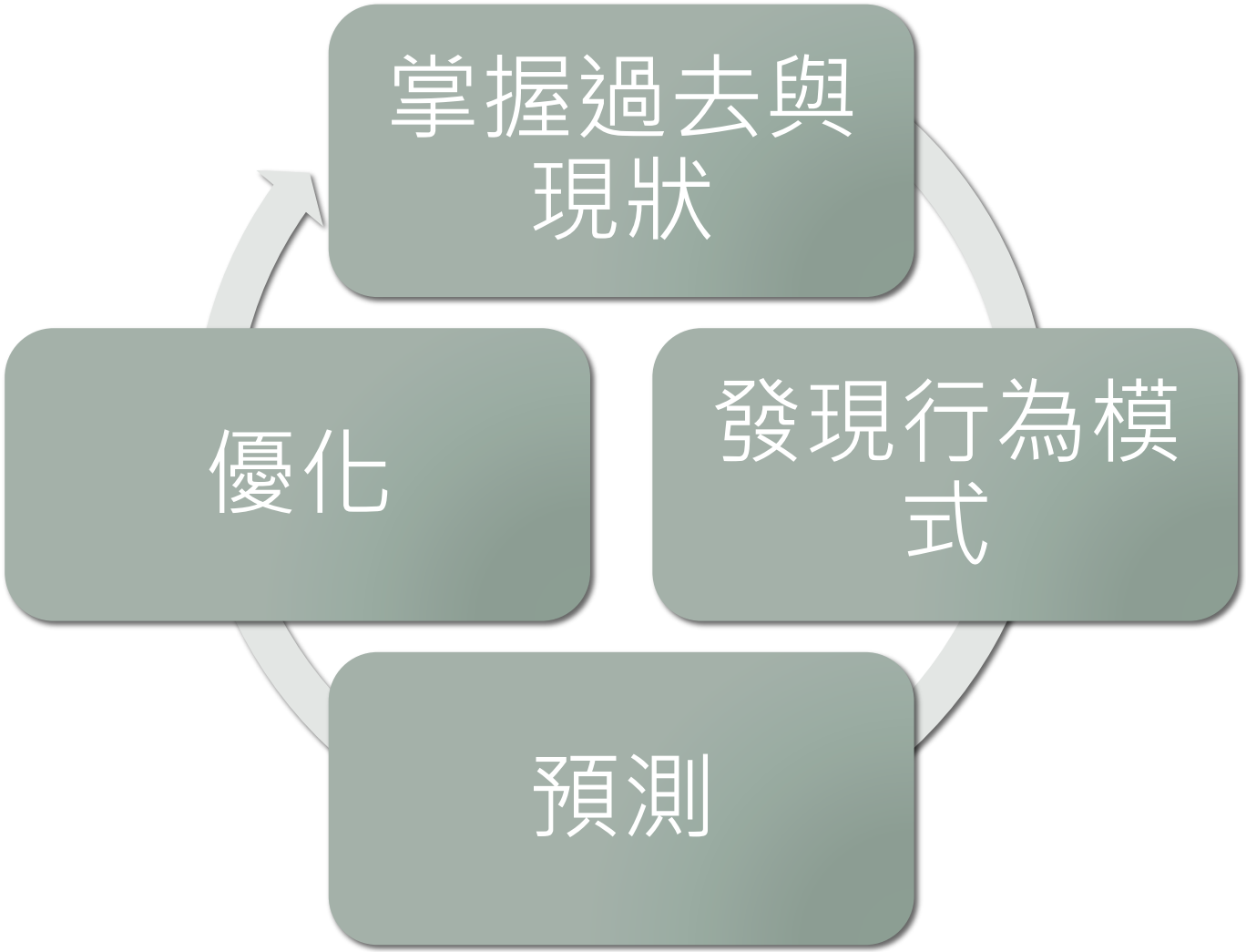
Big Data 應用模式

- 精準推薦商品
- 行為定位廣告
- 運用地點資訊的行銷
- 揪出盜刷
- 客戶流失分析
- 驗出異常、預測油資成本變化
- 改善服務
- 預測路況
- 預測感冒流行

Big Data 應用模式



Value of Big Data



Big Data Privacy

- Do Not Track
- Sand Box Browsing
- Personal information protect
- App/OS statistics feedback
- DuckDuckGo search engine
- Social media authorization

Summary

- Volume, Velocity, Variety
- Hadoop, NoSQL
- Data processing technology
- Data analytics
- Privacy issues