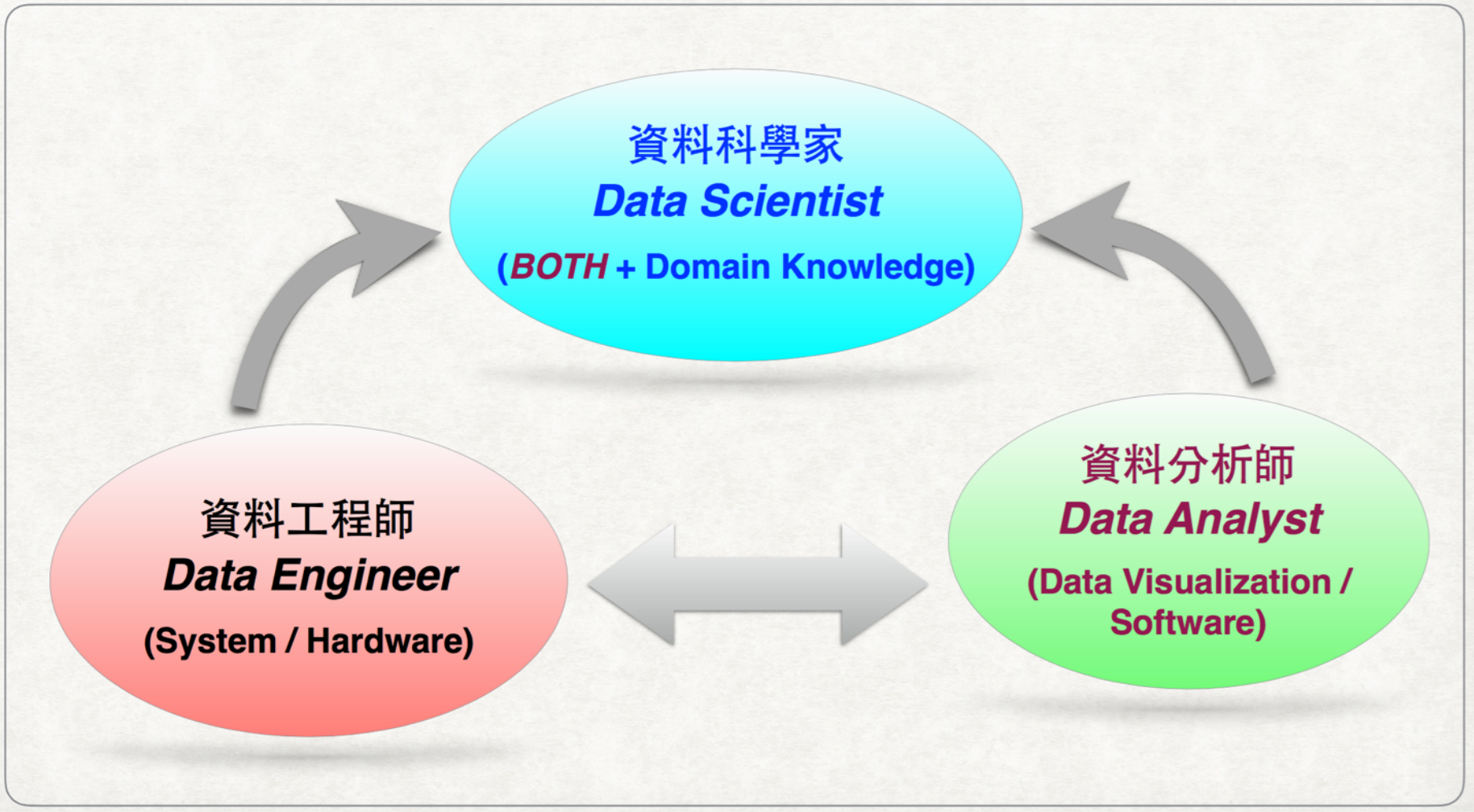
數據工程VS數據分析－
淺談論譠分組構想

Yung-Chuan Lee
2017.02

為何要分組？

- 透過**分組**，以**母雞帶小雞**方式，使新學員得以快速上手、並藉此**培育種子講師**及**專案領導**。
- 數據應用牽涉眾多領域，**一人之力難以全部精通**；希望**結合眾人之力**，**共同學習**、快速破關。
- 論譠學員來自各領域，各有所長；透過分組學習和交流，**貢獻所長**、**各取所需**。

New Careers ☞ Data Scientist, Data Analyst & Data Engineer



資料工程師 VS 資料分析師

- **資料工程師**
  - 平台(Hadoop Ecosystem / Spark)規劃
  - 平台安裝及設定
  - 系統調教、備份、安全性及災難復原
- **資料分析師**
  - 數據分析、機率、統計
  - 演算法選擇
  - 資料視覺化

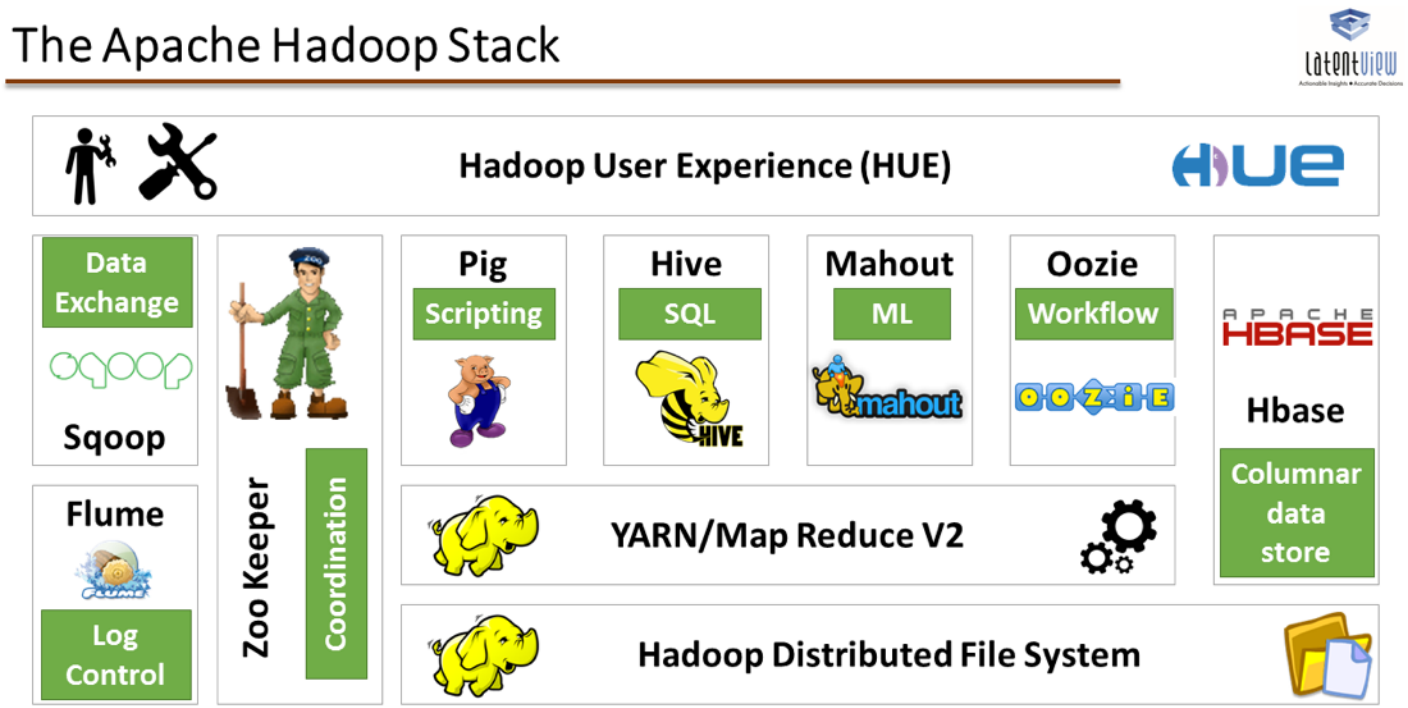
資料工程師養成三步曲－技能

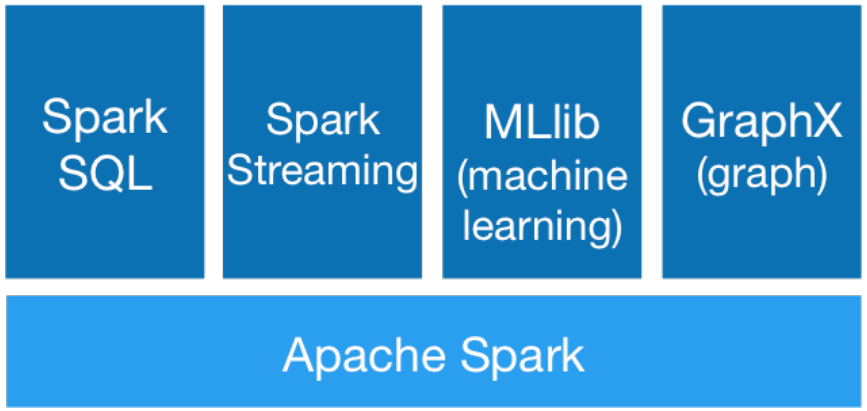
- **入門**
  - Hadoop及Spark單節點 / 多節點安裝
  - HDFS及Spark-shell操作
  - 基本的平台錯誤排除
- **初階**
  - 熟悉Hadoop Ecosystem成員用途及安裝
  - 熟悉Spark Stack Library用途及設定
  - 針對應用情境可提出Solution
- **進階**
  - 有能力進行軟硬體規劃、安裝及設定，以滿足SLA
  - 有能力進行效能調教、並建置備份、安全性及災難復原的機制

資料工程師養成三步曲－軟體元件

- 入門
  - Hadoop Core(HDFS / MapReduce)
  - Spark Shell
- 初階
  - Hadoop Ecosystem
  - Spark Stack Library
- 進階
  - HDSF HA、HDFS Federation、Zookeeper
  - Spark＋Mesos＋Akka＋Cassandra＋Kafka(SMACK)

Hadoop / Spark生態系統





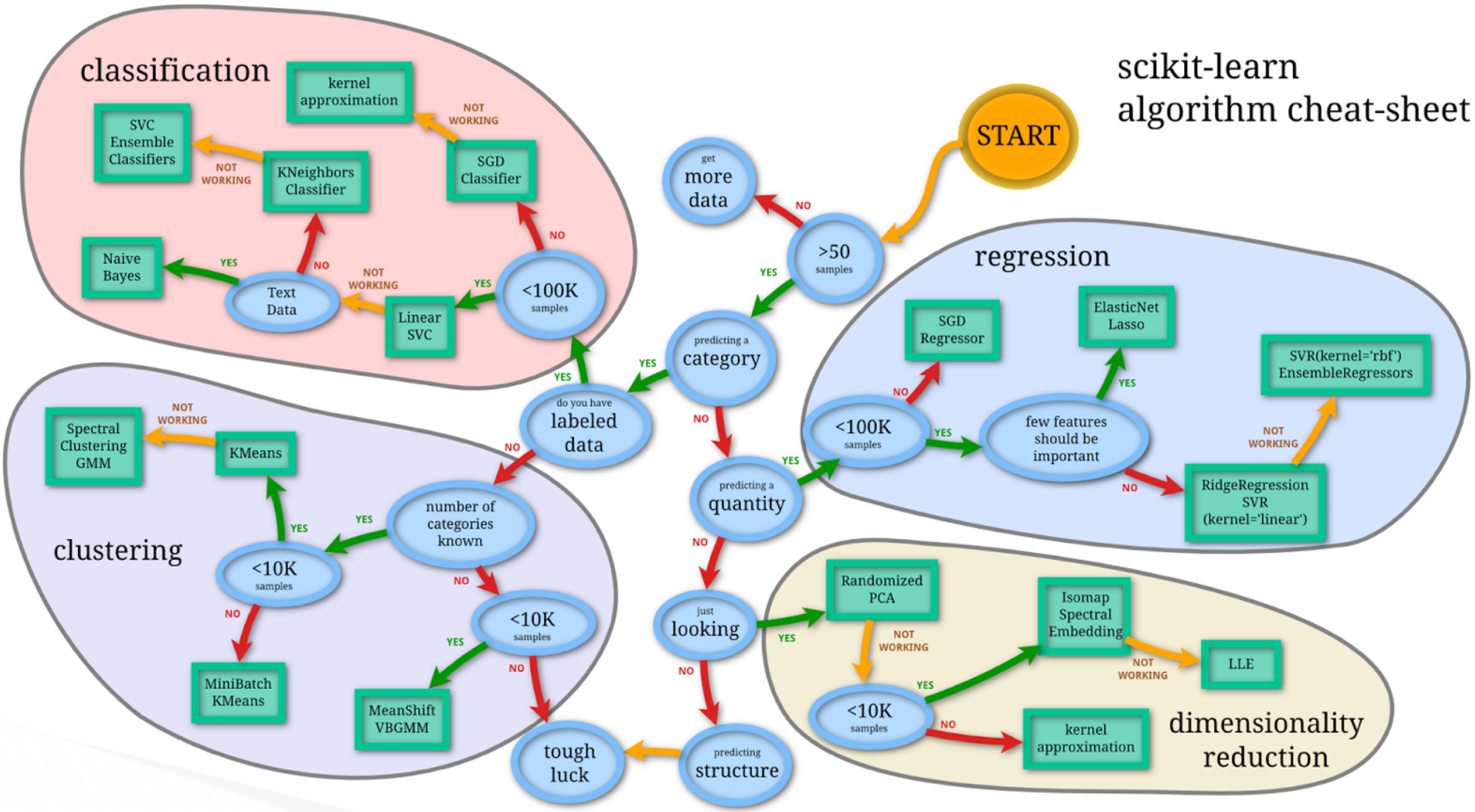
資料分析師養成三步曲－技能

- 入門
  - 機率與統計
  - 簡易分析工具
- 初階
  - 資料視覺化
  - 機器學習演算法
  - 進階分析工作
- 進階
  - deep learning
  - 建構自己的演算法

資料分析師養成三步曲－軟體元件

- **入門**
  - Excel
- **初階**
  - R＋Shiny
  - SparkR、Spark MLlib
- **進階**
  - Kaggle競賽，成為kaggler(Top500－Top50)

資料分析常用演算法


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
<100K samples
few features should be important
RidgeRegression SVR (kernel='linear')
clustering
Spectral Clustering GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
<10K samples
MeanShift VBGMM
dimensionality reduction
Randomized PCA
Isomap Spectral Embedding
LLE
<10K samples
kernel approximation
YES
NO
NOT WORKING

論讀分組規劃

- **Data-Engineer Training Group**
  - 資料工程師訓練團隊
  - 簡稱第一組
- **Data-Analyst Training Group**
  - 資料分析師訓練團隊
  - 簡稱第二組
- **Big-Data Tech Group**
  - 大數據技術團隊
  - 簡稱第三組

組別任務說明－第一組

- **Data-Engineer Training Group(資料工程師訓練團隊)**
  - **養成以下技能，並規劃課程、撰寫教材、成為種子講師**
    - 基礎Hadoop / Spark 平台架設及除錯
    - Hadoop Ecosystem成員熟悉及建置
    - Spark＋Hadoop叢集配置、Spark Submit使用
    - 案例實作
    - 證照：Cloudera Certified Administrator for Apache Hadoop(CCAH)

組別任務說明－第二組

- **Data-Analyst Training Group（資料分析師訓練團隊）**
  - **養成以下技能，並規劃課程、撰寫教材、成為種子講師**
    - R語言相關技術
      - 資料視覺化、資料處理、機率與統計等套件、函式和語法
    - Spark DataFrame / SparkR / Spark MLlib等函式庫使用
    - 參加Kaggle競賽，以成為Kaggle Top 500為目標
    - 證照：CCA Data Analyst、CCA Spark and Hadoop Developer

組別任務說明－第三組

- **Big-Data Tech Group（大數據技術團隊）**
  - **執行以下任務**
    - 追蹤最新技術
    - 支援以上兩組的技術更新
    - 內部講師訓練
    - 證照：CCP Data Engineer、CCP Data Scientist

日期 / 時間	活動地點	研習/活動主題	備 註
2017/2/7 (二) 19:00 ~ 21:10	國立高雄第一科技大學 (NKFUST) 創新育成中心	1. 活動：19:00 ~ 19:20 新春團拜 2. 演講：19:20 ~ 19:45 「數據工程 vs. 數據分析－淺談論壇分組構想」 [講者]：李泳泉 (Yungchuan Lee) Q&A：19:45 ~ 20:00 3. 研習：20:10 ~ 21:10 「R 與統計分析 — 基礎」 [講師]：胡中興	免費研習與社團活動
2017/2/15 (三) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「R 與統計分析 — 進階」 [講師]：胡中興	免費研習活動
2017/2/18 (六) 9:00 ~ 17:00 & 2017/2/18 (日) 9:00 ~ 17:00	(待定)	2/18 & 2/19 兩天 (14小時) 短期研習 [主題]：“R 與 機器學習原理與實務” [講師]：胡中興	收費研習活動 (請留意社團正式公告)
2017/2/22 (三) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「R 與統計分析 — 應用」 [講師]：胡中興	免費研習活動
2017/3/3 (五) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「Scala 程式設計 - 1」 [講師]：胡中興	免費研習活動
2017/3/7 (二) 19:00 ~ 21:00	(待定)	< 基礎研習 >：19:00 ~ 21:00 「Hadoop 2.7.0 單節點叢集架設」 [講師]：社團講師群 (待定)	免費研習活動 (歡迎初學者參加)

日期 / 時間	活動地點	研習/活動主題	備 註
2017/3/10 (五) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「Scala 程式設計 - 2」 [講師]：胡中興	免費研習活動
2017/3/14 (二) 19:00 ~ 21:00	(待定)	< 基礎研習 >：19:00 ~ 21:00 「Hadoop 2.7.0 多節點叢集架設」 [講師]：社團講師群 (待定)	免費研習活動 (歡迎初學者參加)
2017/3/17 (五) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「Scala 程式設計 - 3」 [講師]：胡中興	免費研習活動
2017/3/21 (二) 19:00 ~ 21:00	(待定)	< 基礎研習 >：19:00 ~ 21:00 「Spark 2.1.0/Hadoop 2.7.0平台架設」 [講師]：社團講師群 (待定)	免費研習活動 (歡迎初學者參加)
2017/3/24 (五) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「Scala 程式設計 - 4」 [講師]：胡中興	免費研習活動
2017/3/28 (二) 19:00 ~ 21:00	(待定)	< 基礎研習 >：19:00 ~ 21:00 「R 程式設計 - 1」 [講師]：社團講師群 (待定)	免費研習活動 (歡迎初學者參加)
2017/3/31 (五) 19:00 ~ 21:00	(待定)	研習：19:00 ~ 21:00 「Spark Core 核心技術介紹」 [講師]：胡中興	免費研習活動

- 2 月份 研習活動重點：「R 與統計分析」、「R 與機器學習」
- 3 月份 研習活動重點：「Scala 程式設計」、「Spark Core 核心技術」

  3 月份 初學者研習活動：「Spark & Hadoop 平台架設」、「R 程式設計入門」

各組基本配備...

- 資料工程組
  - 可施作的Lab平台，要準備較佳的硬體配備
    - 4核心、16GB以上RAM的電腦
- 資料分析組
  - ？？？
- 大量的時間、精力及熱情。。。

意見討論
決定分組
推舉組長

參考連結

- 人人都可成為資料科學大師！一整年的網路自學清單就在這了
- Cloudera 證照介紹