Big Data Mining
巨量資料探勘
巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統
(Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)

Min-Yuh Day
戴敏育
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2016-02-23
週次 (Week)  日期 (Date)  內容 (Subject/Topics)
1 2016/02/16  巨量資料探勘課程介紹  
   (Course Orientation for Big Data Mining)
2 2016/02/23  巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統  
   (Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)
3 2016/03/01  關連分析 (Association Analysis)
4 2016/03/08  分類與預測 (Classification and Prediction)
5 2016/03/15  分群分析 (Cluster Analysis)
6 2016/03/22  個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis – K-Means using SAS EM)
7 2016/03/29  個案分析與實作二 (SAS EM 關連分析)：Case Study 2 (Association Analysis using SAS EM)
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2016/04/05</td>
<td>教學行政觀摩日 (Off-campus study)</td>
</tr>
<tr>
<td>9</td>
<td>2016/04/12</td>
<td>期中報告 (Midterm Project Presentation)</td>
</tr>
<tr>
<td>10</td>
<td>2016/04/19</td>
<td>期中考試週 (Midterm Exam)</td>
</tr>
<tr>
<td>11</td>
<td>2016/04/26</td>
<td>個案分析與實作三 (SAS EM 決策樹、模型評估)：Case Study 3 (Decision Tree, Model Evaluation using SAS EM)</td>
</tr>
<tr>
<td>12</td>
<td>2016/05/03</td>
<td>個案分析與實作四 (SAS EM 迴歸分析、類神經網路)：Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)</td>
</tr>
<tr>
<td>13</td>
<td>2016/05/10</td>
<td>Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)</td>
</tr>
<tr>
<td>14</td>
<td>2016/05/17</td>
<td>期末報告 (Final Project Presentation)</td>
</tr>
<tr>
<td>15</td>
<td>2016/05/24</td>
<td>畢業班考試 (Final Exam)</td>
</tr>
</tbody>
</table>
巨量資料基礎：
MapReduce典範、
Hadoop與Spark生態系統
(Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)
Architecture of Big Data Analytics

Big Data Sources
- * Internal
- * External
- * Multiple formats
- * Multiple locations
- * Multiple applications

Big Data Transformation

Big Data Platforms & Tools
- Queries
- Reports
- OLAP

Data Mining

Big Data Analytics Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Analysts
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Business Intelligence (BI) Infrastructure

- Operational Data
- Historical Data
- Machine Data
- Web Data
- Audio/Video Data
- External Data

Extract, transform, load

Data Warehouse

Hadoop Cluster

Data Mart

Casual users
- Queries
- Reports
- Dashboards

Power users
- Queries
- Reports
- OLAP
- Data mining

Data Warehouse

Data Mining and Business Intelligence

Increasing potential to support business decisions

- Decision Making
- Data Presentation
- Visualization Techniques
- Data Mining
- Information Discovery
- Data Exploration
- Statistical Summary, Querying, and Reporting
- Data Preprocessing/Integration, Data Warehouses
- Data Sources
  - Paper, Files, Web documents, Scientific experiments, Database Systems

End User
- Business Analyst
- Data Analyst
- DBA

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
The Evolution of BI Capabilities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Science and Business Intelligence

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Data Mining at the Intersection of Many Disciplines

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Knowledge Discovery (KDD) Process

Data mining: core of knowledge discovery process

Data Cleaning

Data Integration

Data Warehouse

Task-relevant Data

Selection

Data Mining

Pattern Evaluation

Knowledge

Source: Han & Kamber (2006)
### A Taxonomy for Data Mining Tasks

<table>
<thead>
<tr>
<th>Data Mining</th>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>Supervised</td>
<td>Classification and Regression Trees, ANN, SVM, Genetic Algorithms</td>
</tr>
<tr>
<td>Classification</td>
<td>Supervised</td>
<td>Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms</td>
</tr>
<tr>
<td>Regression</td>
<td>Supervised</td>
<td>Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM</td>
</tr>
<tr>
<td>Association</td>
<td>Unsupervised</td>
<td>Apriory, OneR, ZeroR, Eclat</td>
</tr>
<tr>
<td>Link analysis</td>
<td>Unsupervised</td>
<td>Expectation Maximization, Apriory Algorithm, Graph-based Matching</td>
</tr>
<tr>
<td>Sequence analysis</td>
<td>Unsupervised</td>
<td>Apriory Algorithm, FP-Growth technique</td>
</tr>
<tr>
<td>Clustering</td>
<td>Unsupervised</td>
<td>K-means, ANN/SOM</td>
</tr>
<tr>
<td>Outlier analysis</td>
<td>Unsupervised</td>
<td>K-means, Expectation Maximization (EM)</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Deep Learning
Intelligence from Big Data

Source: https://www.vlab.org/events/deep-learning/
Big Data

- Mobile Sensors
- Social Media
- Video Surveillance
- Video Rendering
- Smart Grids
- Geophysical Exploration
- Medical Imaging
- Gene Sequencing

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Big Data Growth is increasingly unstructured
Typical Analytic Architecture

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Data Evolution and the Rise of Big Data Sources

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Emerging Big Data Ecosystem

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Key Roles for the New Big Data Ecosystem

- **Deep Analytical Talent**
  - Projected U.S. talent gap: 140,000 to 190,000

- **Data Savvy Professionals**
  - Projected U.S. talent gap: 1.5 million

- **Technology and Data Enablers**

*Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article “Big Data: The Next Frontier for Innovation, Competition, and Productivity”* 

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
Profile of a Data Scientist

• Quantitative
  – mathematics or statistics
• Technical
  – software engineering, machine learning, and programming skills
• Skeptical mind-set and critical thinking
• Curious and creative
• Communicative and collaborative

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
VISUAL ANALYTICS

DYNAMIC & INTERACTIVE
Dashboard
Graph
Map

ENHANCE
Understanding
Investigation
User Experience

BIG ANALYTICS

QUERY & FILTER
Complex queries
R21²

DETECT
Anomalies
Communities
Typologies

PREDICT
Trending
Real-time
Prediction

DECIDE
Simulation
Optimization

BIG DATA – Batch

BIG DATA – Real Time

Complex by nature

DATA

Complex by structure

MapReduce Paradigm

Big Data

Map0 → Reduce0
Map1 → Reduce1
Map2 → Reduce2
Map3 → Reduce3

MapReduce Data

Output Data
The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

Source: http://hadoop.apache.org/
Big Data with Hadoop Architecture

**LOGICAL ARCHITECTURE**

**Processing: MapReduce**
- Job Tracker
- Task Tracker
  - Mapper
  - Mapper
  - Mapper
- Shuffle and Sort
  - Reducer
  - Reducer
  - Reducer

**Storage: HDFS**
- NameNode
- Data Node
  - Block
  - Block
- Data Node
  - Block
  - Block

**PROCESS FLOW**
- Input Data Set
  - Split 0
    - Map 0
    - Reduce 0
  - Split 1
    - Map 1
    - Reduce 0
- Split n
  - Map n
  - Reduce 0

**PHYSICAL ARCHITECTURE**
- Hadoop Cluster
  - Master
  - Slave
  - Slave
  - Slave
  - Slave
  - Slave
  - Slave
  - Slave

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

Big Data with Hadoop Architecture

Logical Architecture

Storage: HDFS

Big Data with Hadoop Architecture

Process Flow

Big Data with Hadoop Architecture

Hadoop Cluster

Hadoop Ecosystem

Source: Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing
Traditional ETL Architecture

Offload ETL with Hadoop (Big Data Architecture)

HDP
A Complete Enterprise Hadoop Data Platform

Source: http://hortonworks.com/hdp/
Apache Spark is a fast and general engine for large-scale data processing.

Source: http://spark.apache.org/
Logistic regression in Hadoop and Spark

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Source: http://spark.apache.org/
Ease of Use

• Write applications quickly in Java, Scala, Python, R.

Source: http://spark.apache.org/
Word count in Spark's Python API

text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a+b)

Source: http://spark.apache.org/
Spark and Hadoop

Source: http://spark.apache.org/
Spark Ecosystem

Spark SQL
Spark Streaming
MLlib (machine learning)
GraphX (graph)

Apache Spark

Source: http://spark.apache.org/
Spark Ecosystem

- Spark
  - Spark Streaming
    - Kafka
    - Flume
  - Mlib (machine learning)
  - Spark SQL
  - GraphX (graph)
    - Titan
    - HBase
    - Cassandra
  - H2O
  - Hive
  - HDFS

Source: Mike Frampton (2015), Mastering Apache Spark, Packt Publishing
Python for Big Data Analytics

(The column on the left is the 2015 ranking; the column on the right is the 2014 ranking for comparison)

<table>
<thead>
<tr>
<th>Language Rank</th>
<th>Types</th>
<th>Spectrum Ranking</th>
<th>2015</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Java</td>
<td>![Network Symbol] ![Phone] ![Desktop]</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>2. C</td>
<td>![Phone] ![Desktop] ![Server]</td>
<td>99.9</td>
<td>99.3</td>
<td></td>
</tr>
<tr>
<td>3. C++</td>
<td>![Network Symbol] ![Phone] ![Desktop]</td>
<td>99.4</td>
<td>95.5</td>
<td></td>
</tr>
<tr>
<td>4. Python</td>
<td>![Network Symbol] ![Phone] ![Desktop]</td>
<td>96.5</td>
<td>93.5</td>
<td></td>
</tr>
<tr>
<td>5. C#</td>
<td>![Network Symbol] ![Phone] ![Desktop]</td>
<td>91.3</td>
<td>92.4</td>
<td></td>
</tr>
<tr>
<td>6. R</td>
<td>![Desktop]</td>
<td>84.8</td>
<td>84.8</td>
<td></td>
</tr>
<tr>
<td>7. PHP</td>
<td>![Network Symbol] ![Phone]</td>
<td>84.5</td>
<td>84.5</td>
<td></td>
</tr>
<tr>
<td>8. JavaScript</td>
<td>![Network Symbol] ![Phone]</td>
<td>83.0</td>
<td>78.9</td>
<td></td>
</tr>
<tr>
<td>9. Ruby</td>
<td>![Network Symbol] ![Desktop]</td>
<td>76.2</td>
<td>74.3</td>
<td></td>
</tr>
<tr>
<td>10. Matlab</td>
<td>![Desktop]</td>
<td>72.4</td>
<td>72.8</td>
<td></td>
</tr>
</tbody>
</table>

Top Analytics, Data Mining, Data Science software used, 2015

- R
- RapidMiner
- SQL
- Python
- Excel
- KNIME
- Hadoop
- Tableau
- SAS base
- Spark

SAS & Hadoop
Modern Reality

Infrastructure
- Commoditization
- Architectures
- Scale

Data
- New Complex Streams
- Perishable Considerations
- Cost

Analytics
- New Category of Business Problems
- Analytical Algorithms
- Operationalization

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
Finding treasures in unstructured data like social media or survey tools that could uncover insights about consumer sentiment.

Leveraging historical data to drive better insight into decision-making for the future.

Mine transaction databases for data of spending patterns that indicate a stolen card.

Analyze massive amounts of data in order to accurately identify areas likely to produce the most profitable results.
Trends in Analytics

Complex Business Problems Are Driving Analytics Innovation

Speed Will Be Of Essence

Leverage Analytics To Unlock The Information Contained In Unstructured Data

Operationalizing Analytics

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
Architectures of Big Data Analytics
Traditional Analytics

Unstructured, Semi-structured and Streaming data (i.e. sensor data) handled often outside the Warehouse flow

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
Hadoop as a “new data” Store
Hadoop as an additional input to the EDW

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
Hadoop Data Platform As a “staging Layer” as part of a “data Lake”
– Downstream stores could be Hadoop, data appliances or an RDBMS

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
SAS Big data Strategy – SAS areas

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
SAS Big data Strategy – SAS areas

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
SAS® Within the HADOOP ECOSYSTEM

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
SAS enables the entire lifecycle around HADOOP

SAS enables the entire lifecycle around HADOOP

Done using either the Data Preparation, Data Exploration or Build Model Tools

SAS Visual Analytics
SAS Visual Statistics
SAS In-Memory Statistics for Hadoop

SAS High Performance Analytics Offerings supported by relevant clients like SAS Enterprise Miner, SAS/STAT etc.

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
SAS® VISUAL ANALYTICS

A Single solution for Data Discovery, Visualization, analytics and reporting

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
Example: text analysis gives you insight to customer experience and opinion

Analytics applied to text provides real MEANING

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
Visualization

Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics
References

• EMC Education Services (2015),
  Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley

• Shiva Achari (2015),
  Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing

• Mike Frampton (2015),
  Mastering Apache Spark, Packt Publishing

• Deepak Ramanathan (2014),
  SAS Modernization architectures - Big Data Analytics,
  http://www.slideshare.net/deepakramanathan/sas-modernization-architectures-big-data-analytics