

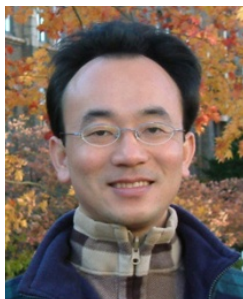
Big Data Mining 巨量資料探勘

Course Orientation for Big Data Mining (巨量資料探勘課程介紹)

1052DM01

MI4 (M2244) (3069)

Thu, 8, 9 (15:10-17:00) (B130)



Min-Yuh Day

戴敏育

Assistant Professor

專任助理教授

Dept. of Information Management, Tamkang University

淡江大學 資訊管理學系

<http://mail.tku.edu.tw/myday/>

2017-02-16



淡江大學105學年度第2學期

課程教學計畫表

Spring 2017 (2017.02 - 2017.06)

- 課程名稱：巨量資料探勘 (Big Data Mining)
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系級：資管四P (TLMXB4P) (M2244) (3094)
- 開課資料：選修 單學期 2 學分 (2 Credits, Elective)
- 上課時間：週四 8,9 (Thu 15:10-17:00)
- 上課教室：B130

課程簡介

- 本課程介紹巨量資料探勘 (Big Data Mining) 的基礎概念及應用技術。
- 課程內容包括
 - 巨量資料探勘 (Big Data Mining)
 - 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)
 - 關連分析 (Association Analysis)
 - 分類與預測 (Classification and Prediction)
 - 分群分析 (Cluster Analysis)
 - SAS企業資料採礦實務 (SAS EM)
 - 巨量資料探勘個案分析與實作
 - Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)

Course Introduction

- This course introduces the fundamental concepts and applications technology of big data mining.
- Topics include
 - Big Data Mining
 - Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
 - Association Analysis
 - Classification and Prediction
 - Cluster Analysis
 - Data Mining Using SAS Enterprise Miner (SAS EM)
 - Case Study and Implementation of Big Data Mining
 - Deep Learning with Google TensorFlow

課程目標 (Objective)

- 瞭解及應用 巨量資料探勘基本概念與技術。
- Understand and apply the fundamental concepts and technology of big data mining

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2017/02/16	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
2	2017/02/23	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	2017/03/02	關連分析 (Association Analysis)
4	2017/03/09	分類與預測 (Classification and Prediction)
5	2017/03/16	分群分析 (Cluster Analysis)
6	2017/03/23	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	2017/03/30	個案分析與實作二 (SAS EM 關連分析)： Case Study 2 (Association Analysis using SAS EM)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
8	2017/04/06	教學行政觀摩日 (Off-campus study)
9	2017/04/13	期中報告 (Midterm Project Presentation)
10	2017/04/20	期中考試週 (Midterm Exam)
11	2017/04/27	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)
12	2017/05/04	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)
13	2017/05/11	Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)
14	2017/05/18	期末報告 (Final Project Presentation)
15	2017/05/25	畢業班考試 (Final Exam)

教學方法與評量方法

- 教學方法
 - 講述、討論、賞析、模擬、實作、問題解決
- 評量方法
 - 紙筆測驗、實作、報告、上課表現

教材課本

- 教材課本

- 講義 (Slides)

- 資料採礦運用：以SAS Enterprise Miner為工具，
李淑娟，2015，SAS賽仕電腦軟體

- 參考書籍

- Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014
 - Data Science for Business: What you need to know about data mining and data-analytic thinking, Foster Provost and Tom Fawcett, O'Reilly, 2013
 - Applied Analytics Using SAS Enterprise Mining, Jim Georges, Jeff Thompson and Chip Wells, SAS, 2010
 - Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kaufmann, 2011

作業與學期成績計算方式

- 作業篇數
 - 3篇
- 學期成績計算方式
 - ☒ 期中評量：30 %
 - ☒ 期末評量：30 %
 - ☒ 其他（課堂參與及報告討論表現）：40 %

Team Term Project

- Term Project Topics
 - Big Data mining
 - Big Data Analytics
 - Social Computing
 - Business Intelligence
 - FinTech
- 3-4 人為一組
 - 分組名單於 2017/02/23 (四) 課程下課時繳交
 - 由班代統一收集協調分組名單

2017/02/23

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)**

2017/05/11

Google TensorFlow

深度學習

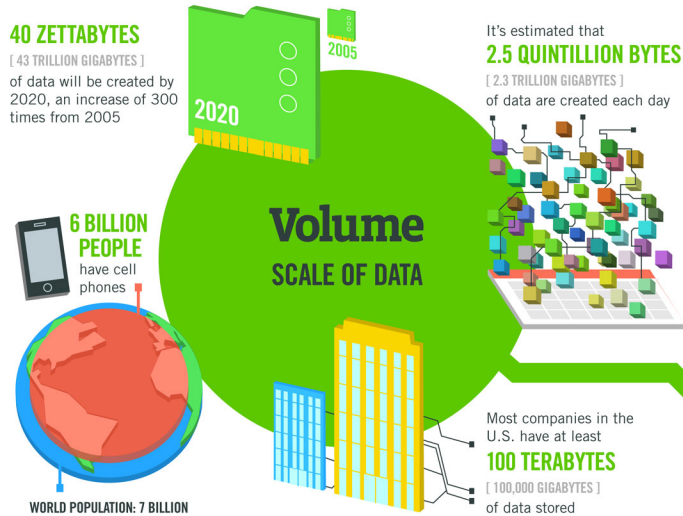
(Deep Learning

with

Google TensorFlow)

Big Data Analytics and Data Mining

Big Data 4 V



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data,
with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

Variety

DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
— almost 2.5 connections per person on earth

Modern cars have close to **100 SENSORS**
that monitor items such as fuel level and tire pressure

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Veracity

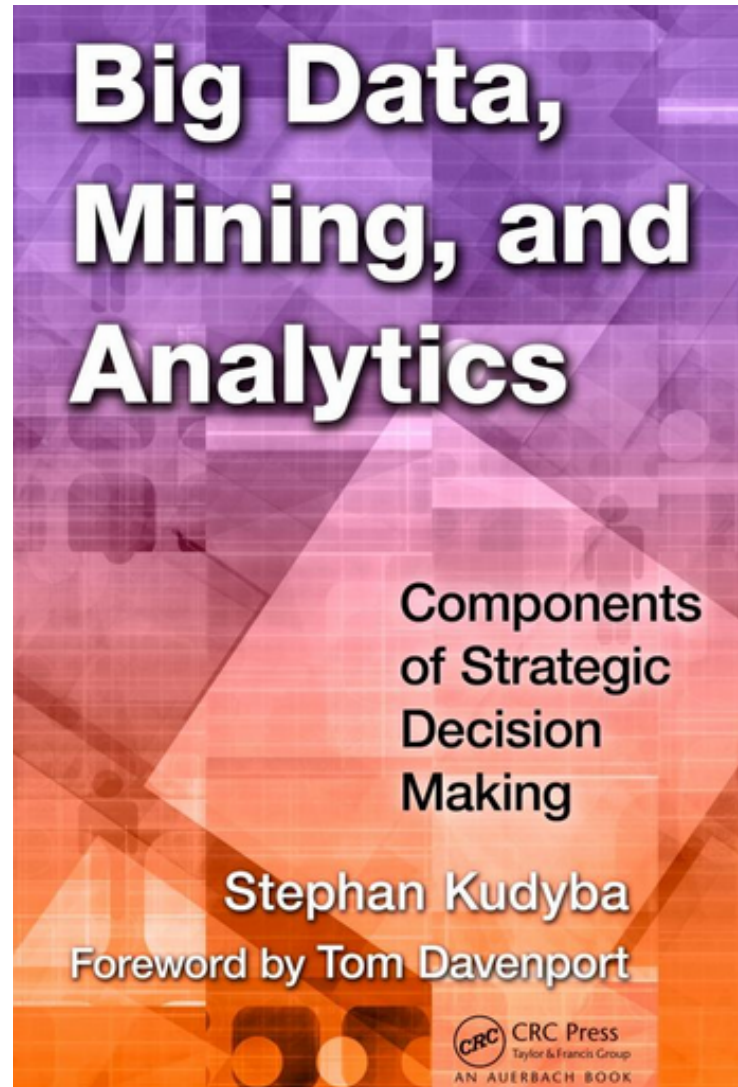
UNCERTAINTY OF DATA

Poor data quality costs the US economy around

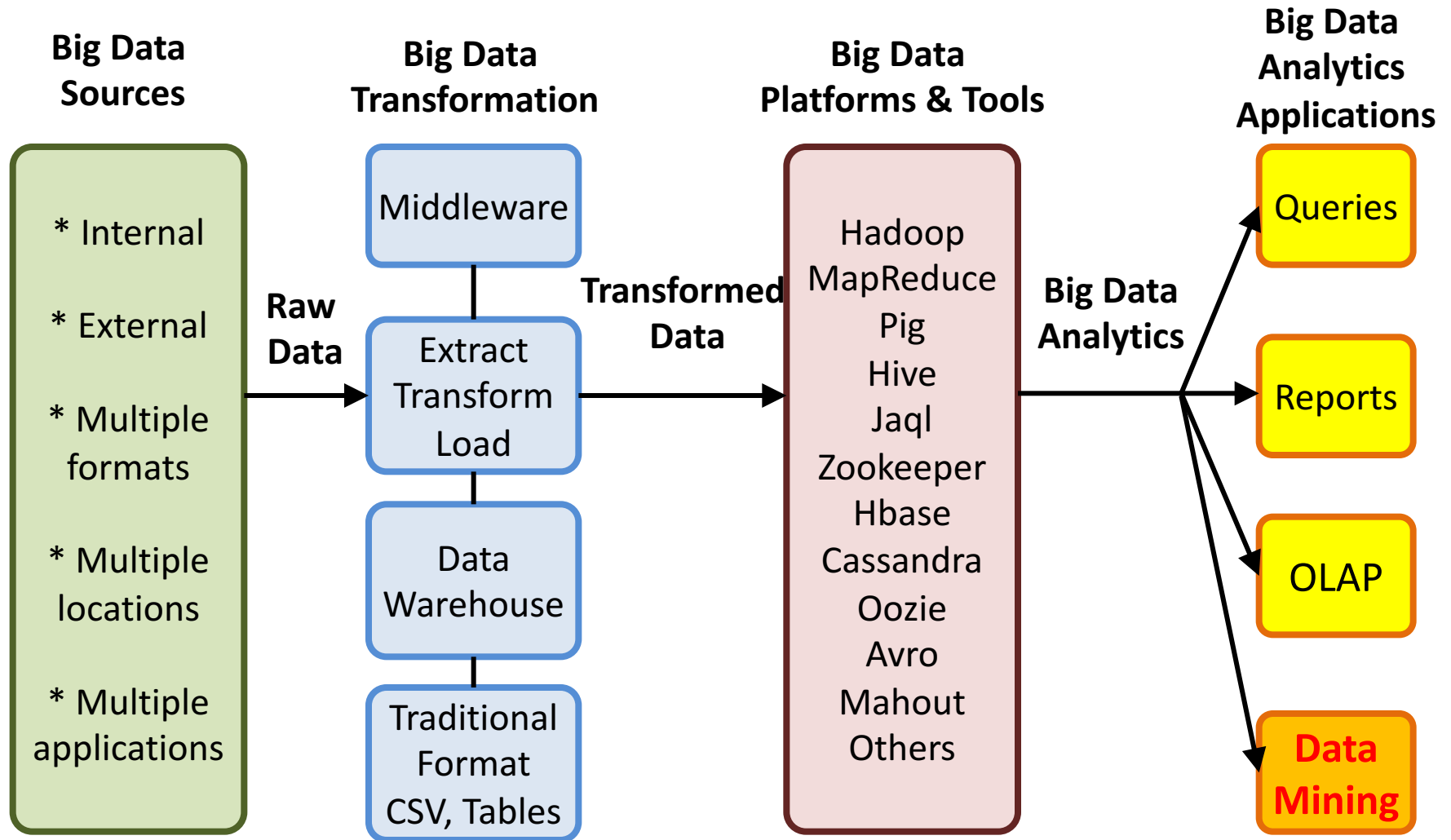
\$3.1 TRILLION A YEAR

Value

Stephan Kudyba (2014),
Big Data, Mining, and Analytics:
Components of Strategic Decision Making, Auerbach Publications



Architecture of Big Data Analytics

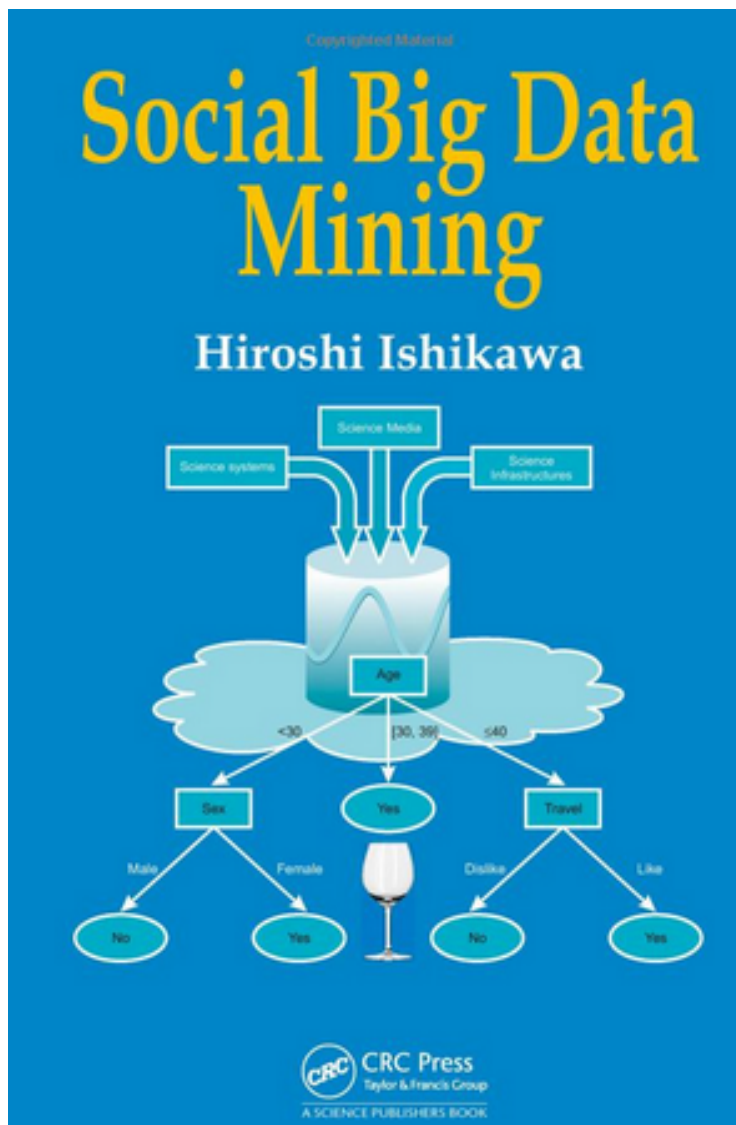


Architecture of Big Data Analytics



Social Big Data Mining

(Hiroshi Ishikawa, 2015)



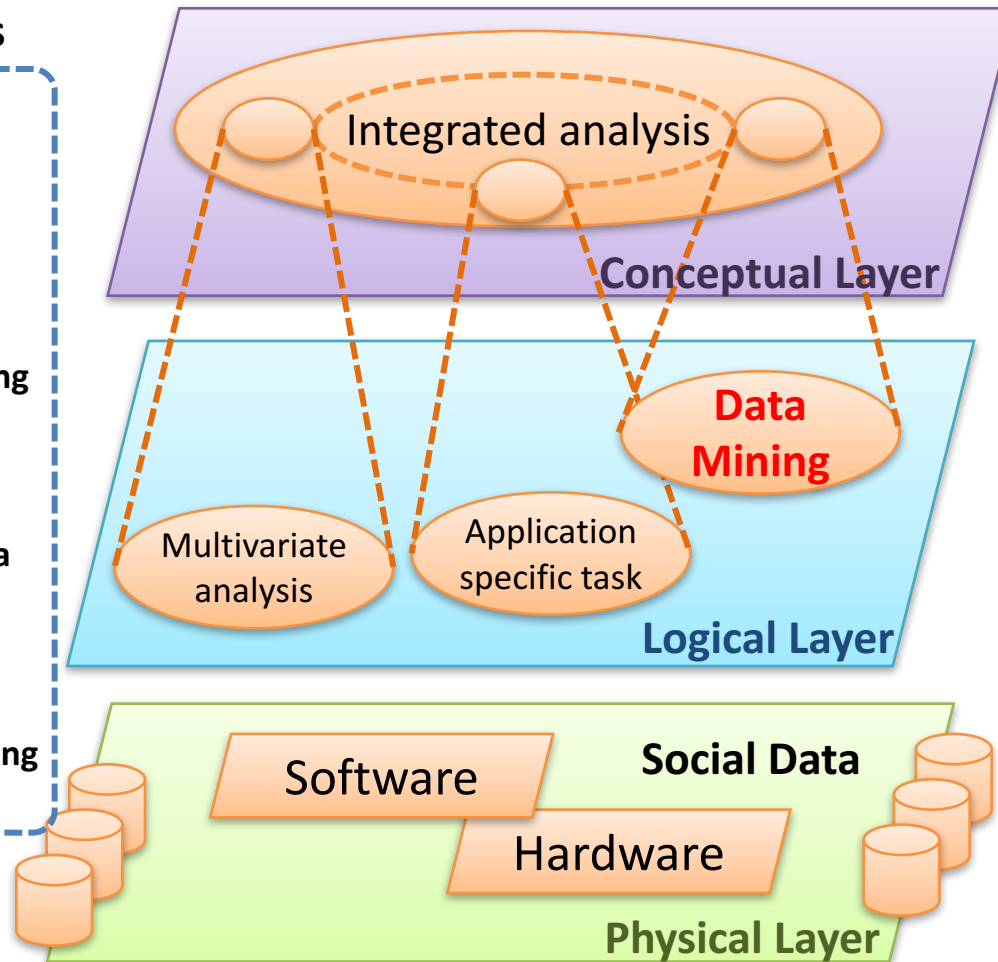
Source: <http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X>

Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Enabling Technologies

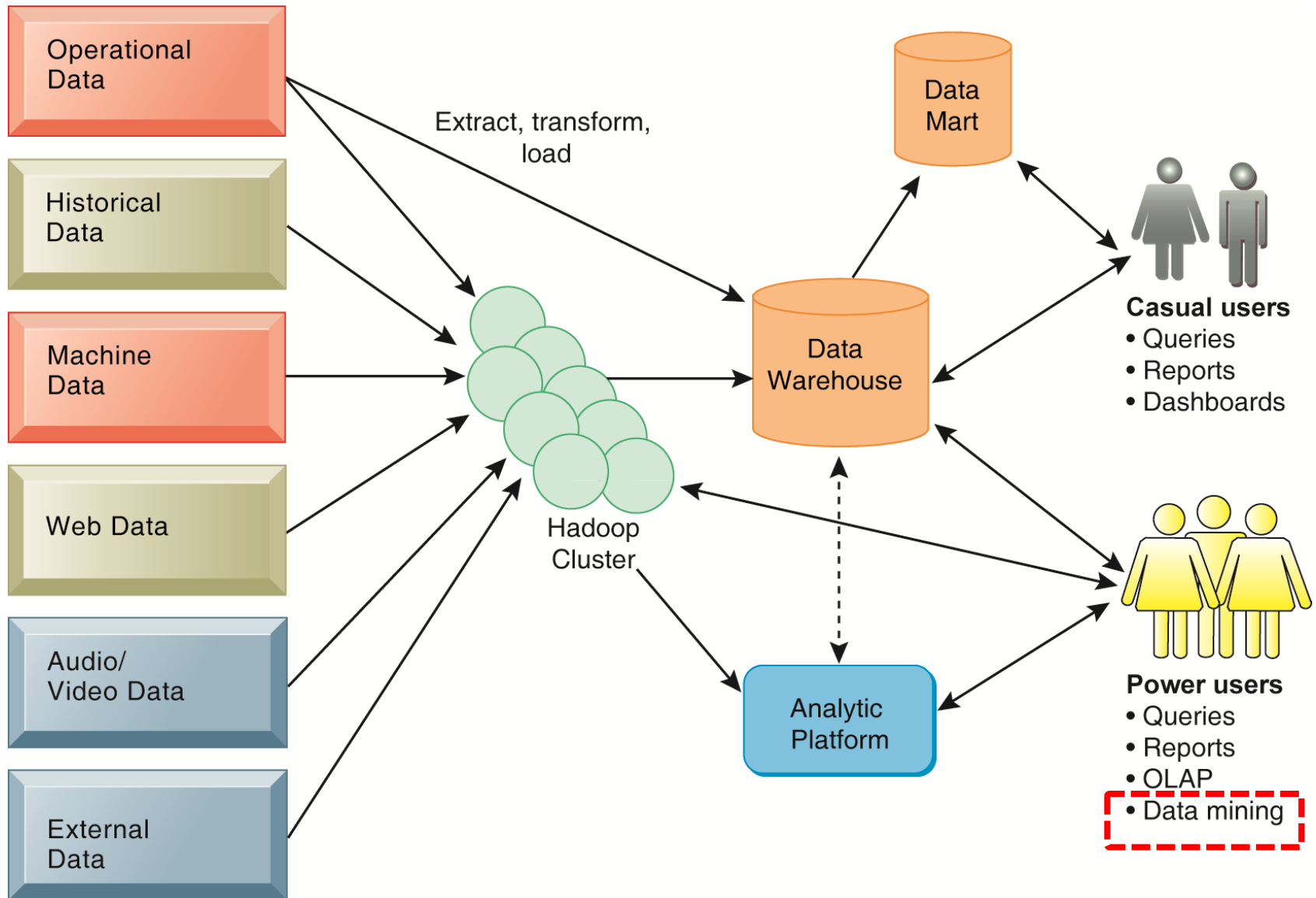
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distributed processing



Analysts

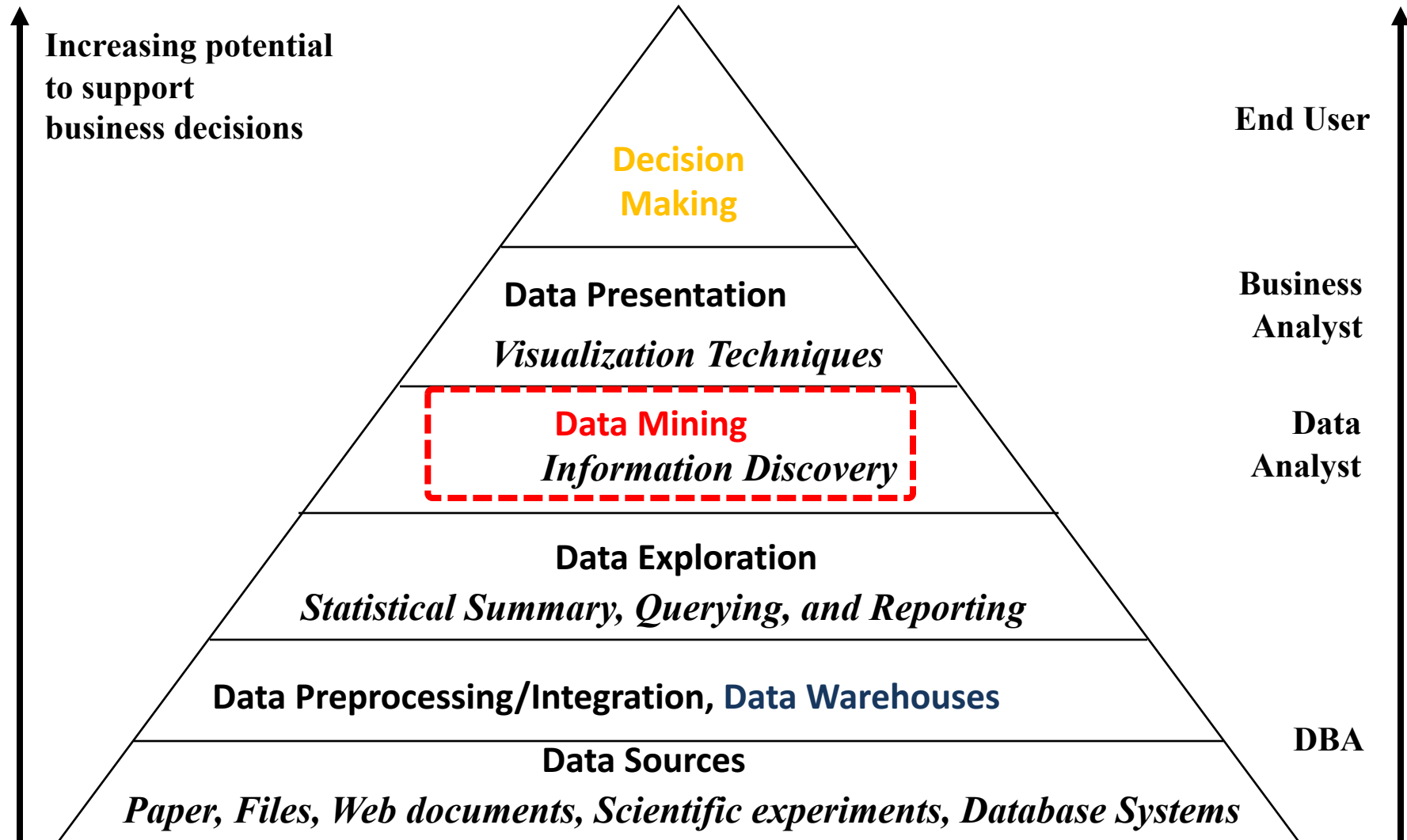
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Business Intelligence (BI) Infrastructure

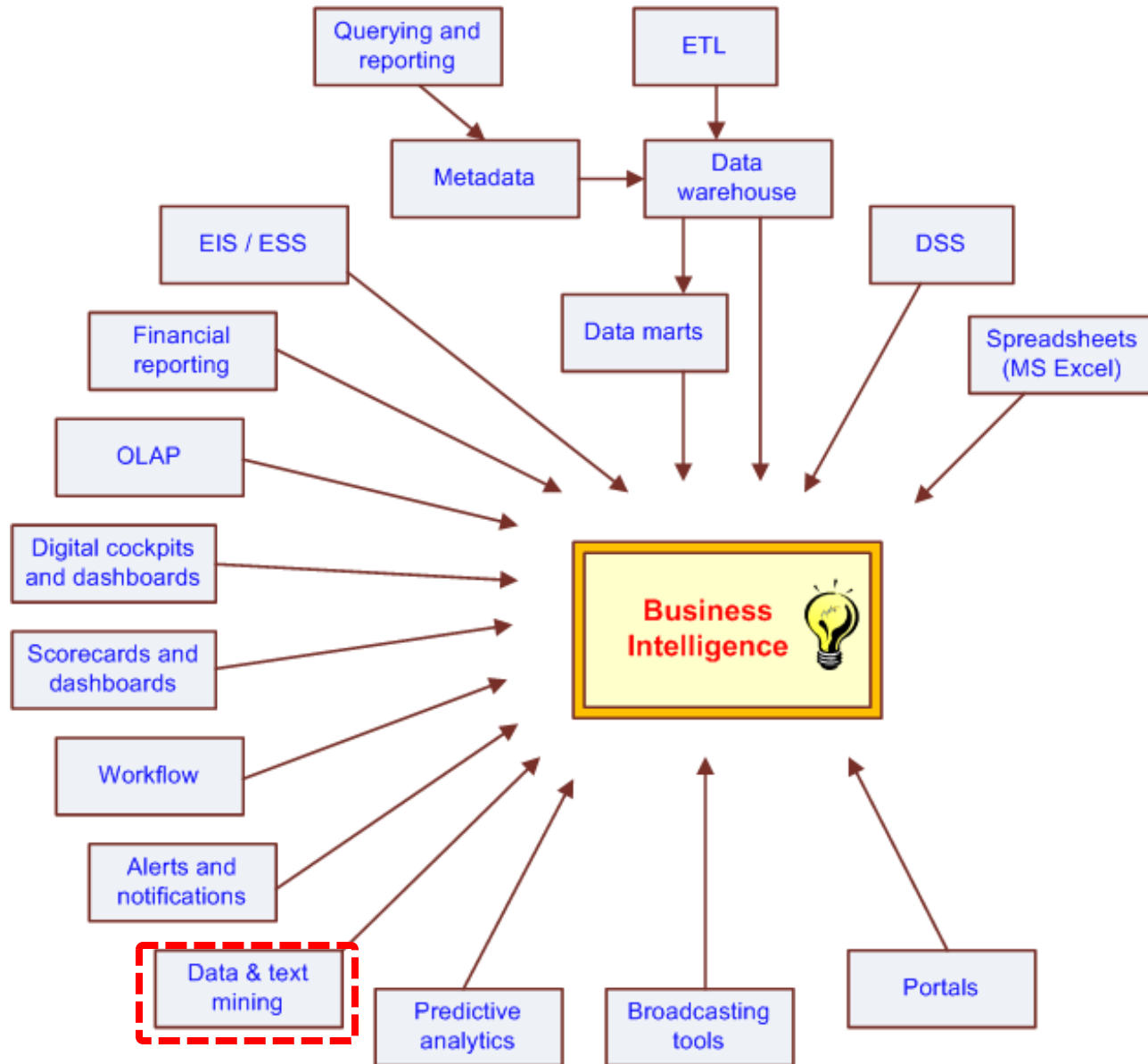


Data Warehouse

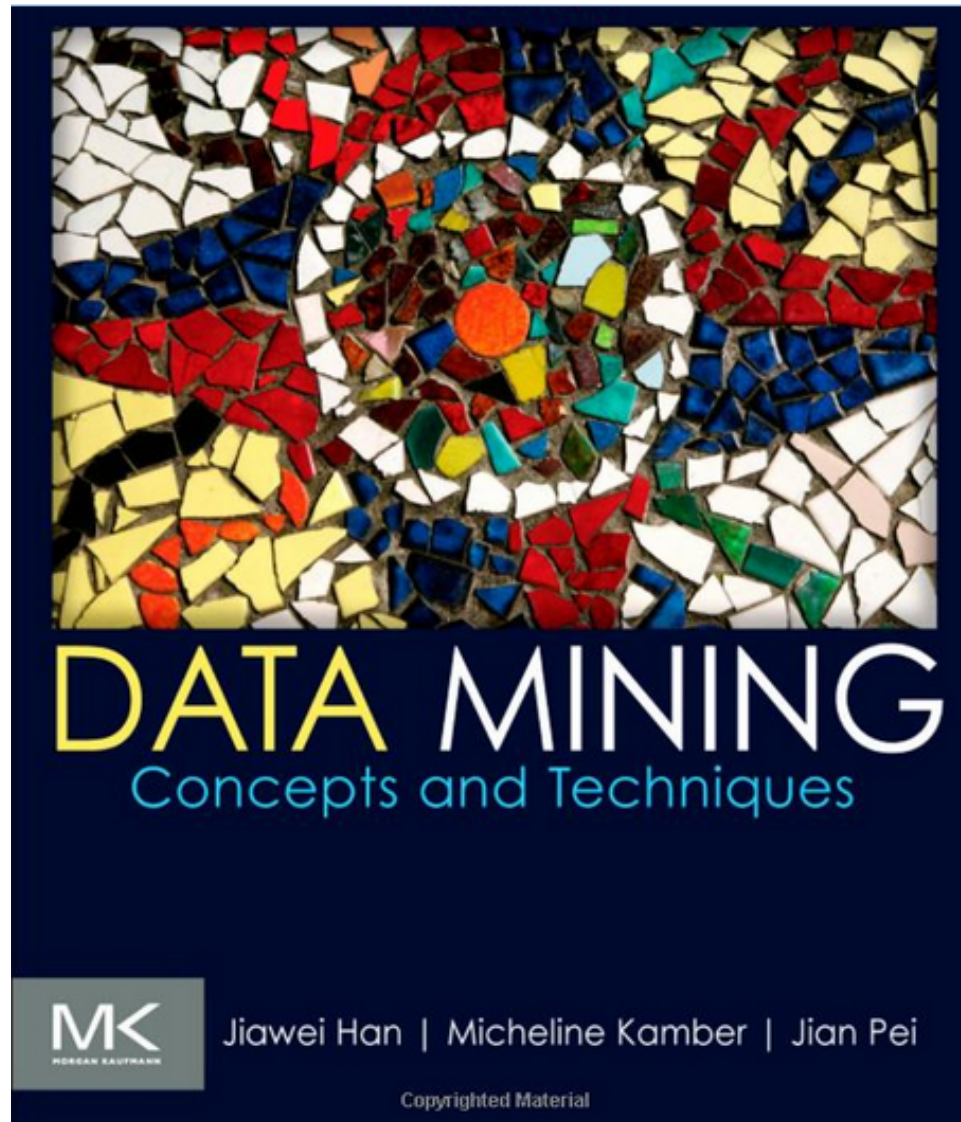
Data Mining and Business Intelligence



The Evolution of BI Capabilities

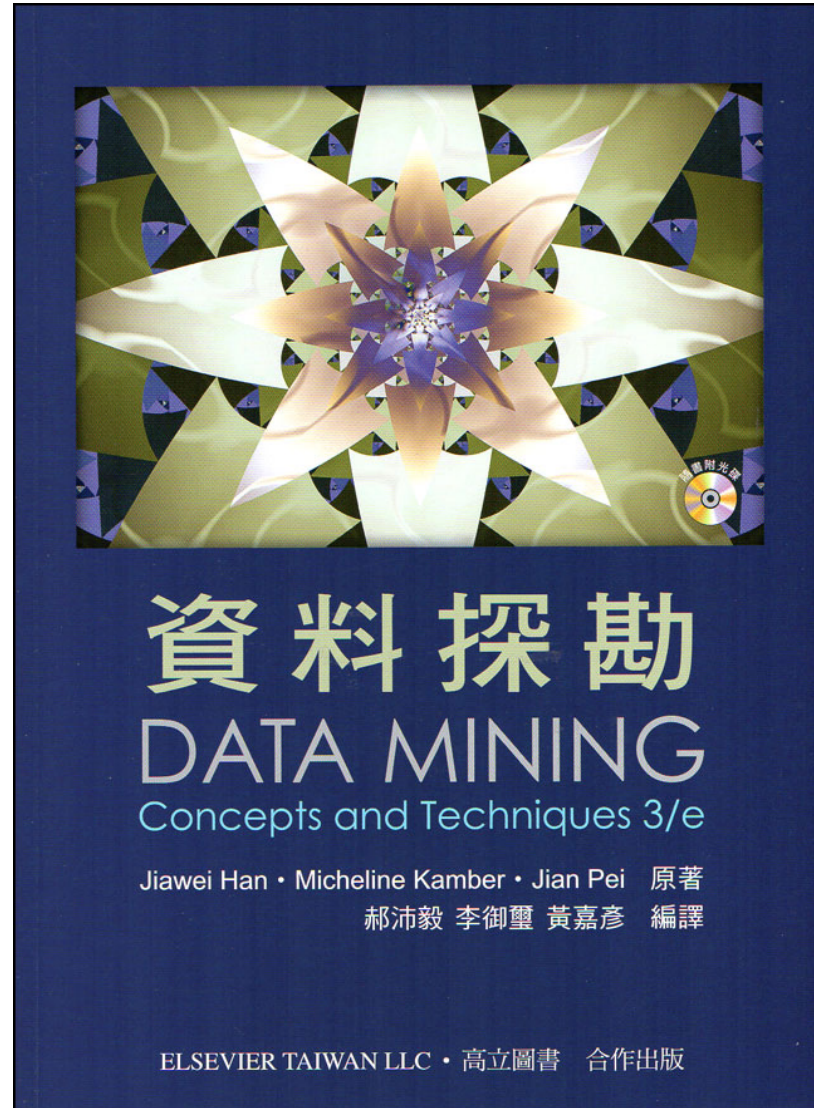


Data Mining

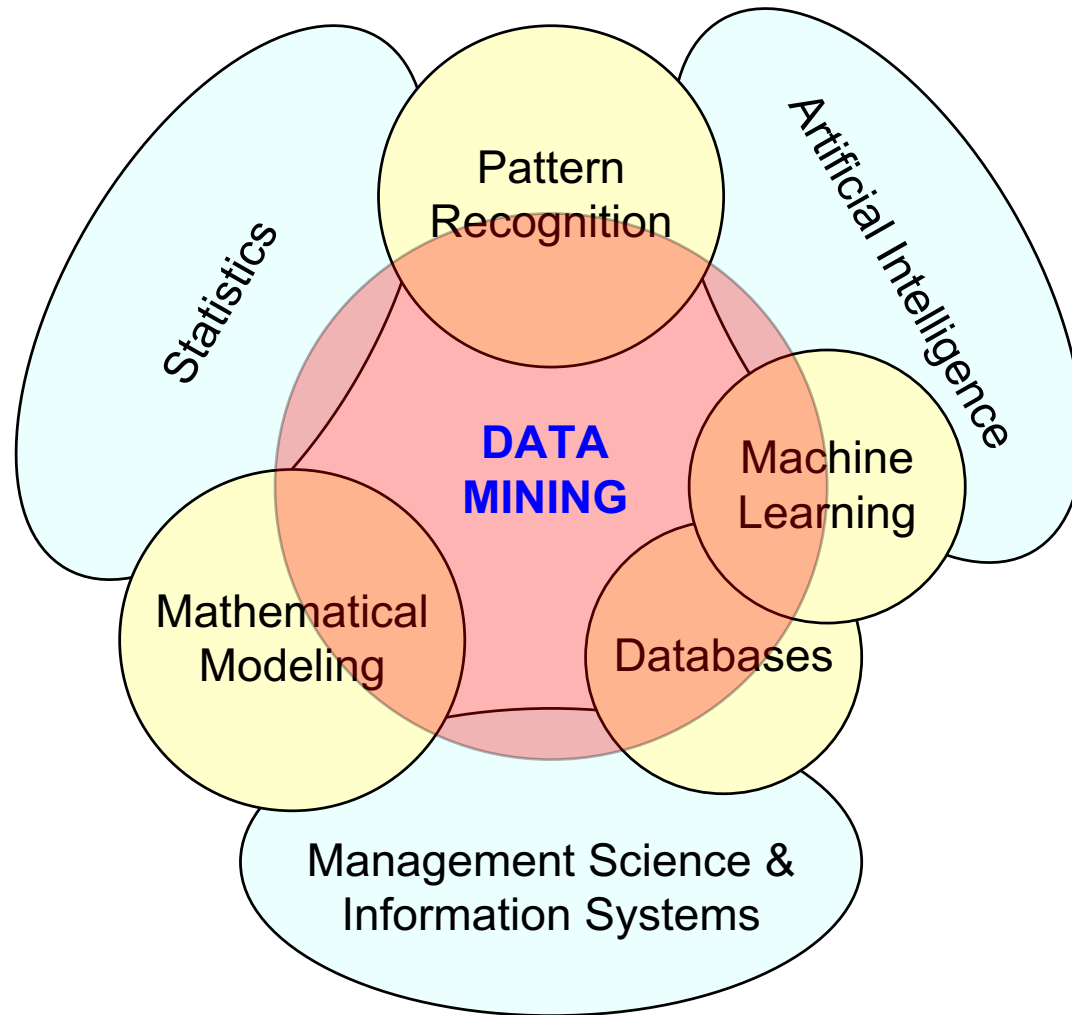


郝沛毅, 李御璽, 黃嘉彥 編譯, 資料探勘

(Jiawei Han, Micheline Kamber, Jian Pei, Data Mining - Concepts and Techniques 3/e),
高立圖書, 2014



Data Mining at the Intersection of Many Disciplines





Data Mining:

Core **Analytics** Process

The **KDD** Process for
Extracting Useful **Knowledge**
from Volumes of **Data**

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

The **KDD Process** for Extracting Useful **Knowledge** from Volumes of **Data**.

Communications of the ACM, 39(11), 27-34.

Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information.

The KDD Process for Extracting Useful Knowledge from Volumes of Data


AS WE MARCH INTO THE AGE of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls,

and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses.

Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer

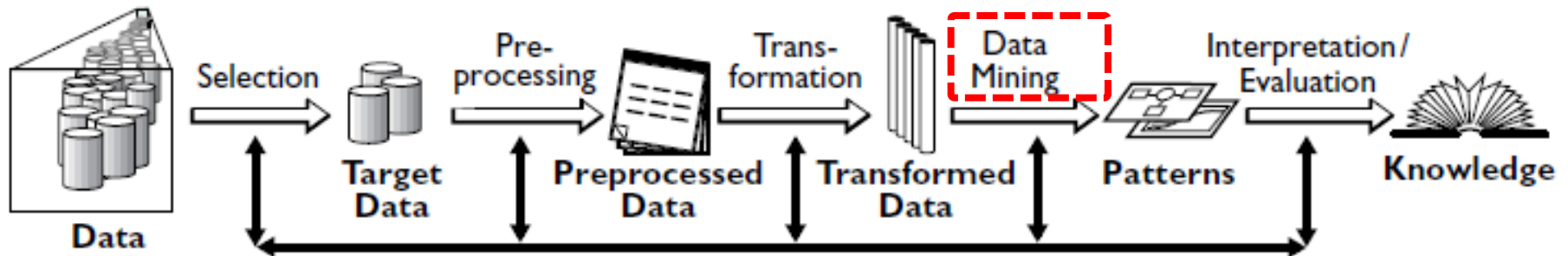
Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth



Data Mining

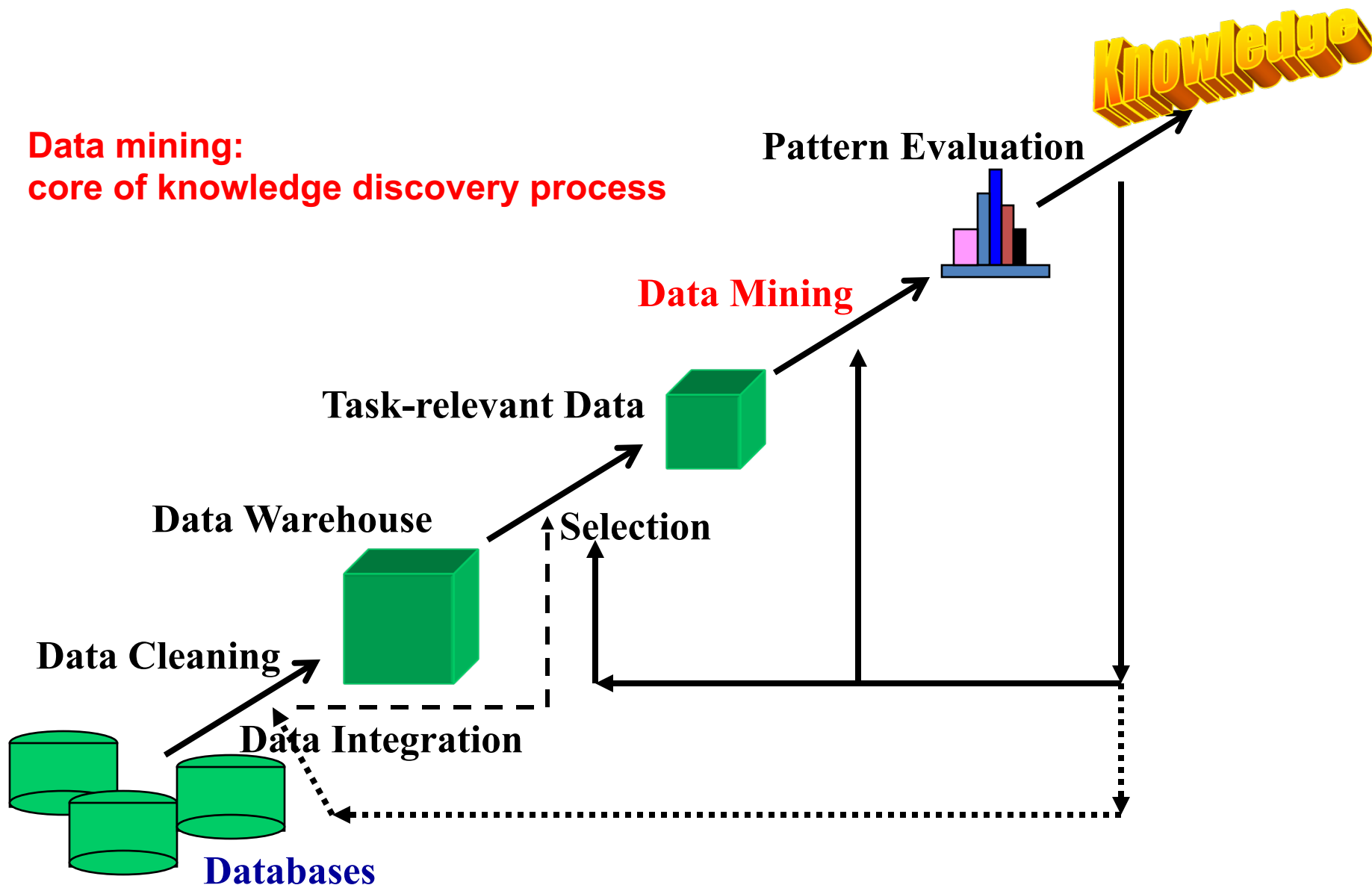
Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)



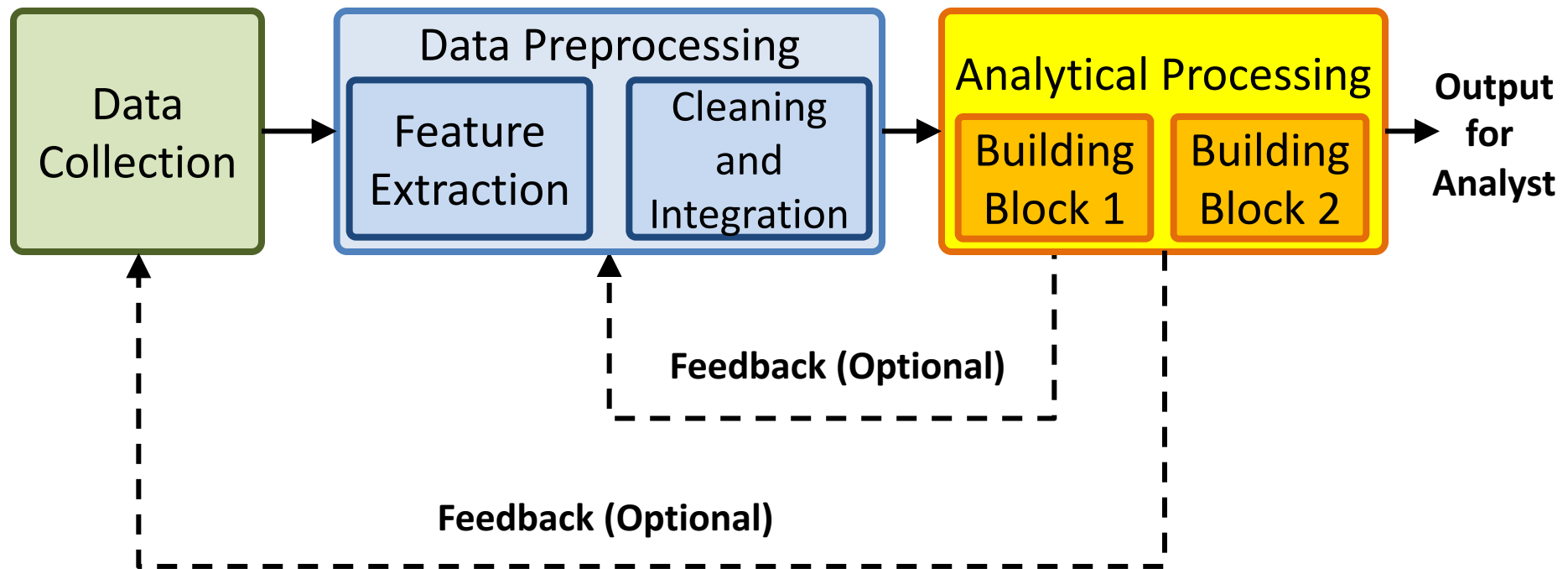
Knowledge Discovery (KDD) Process

Data mining:
core of knowledge discovery process

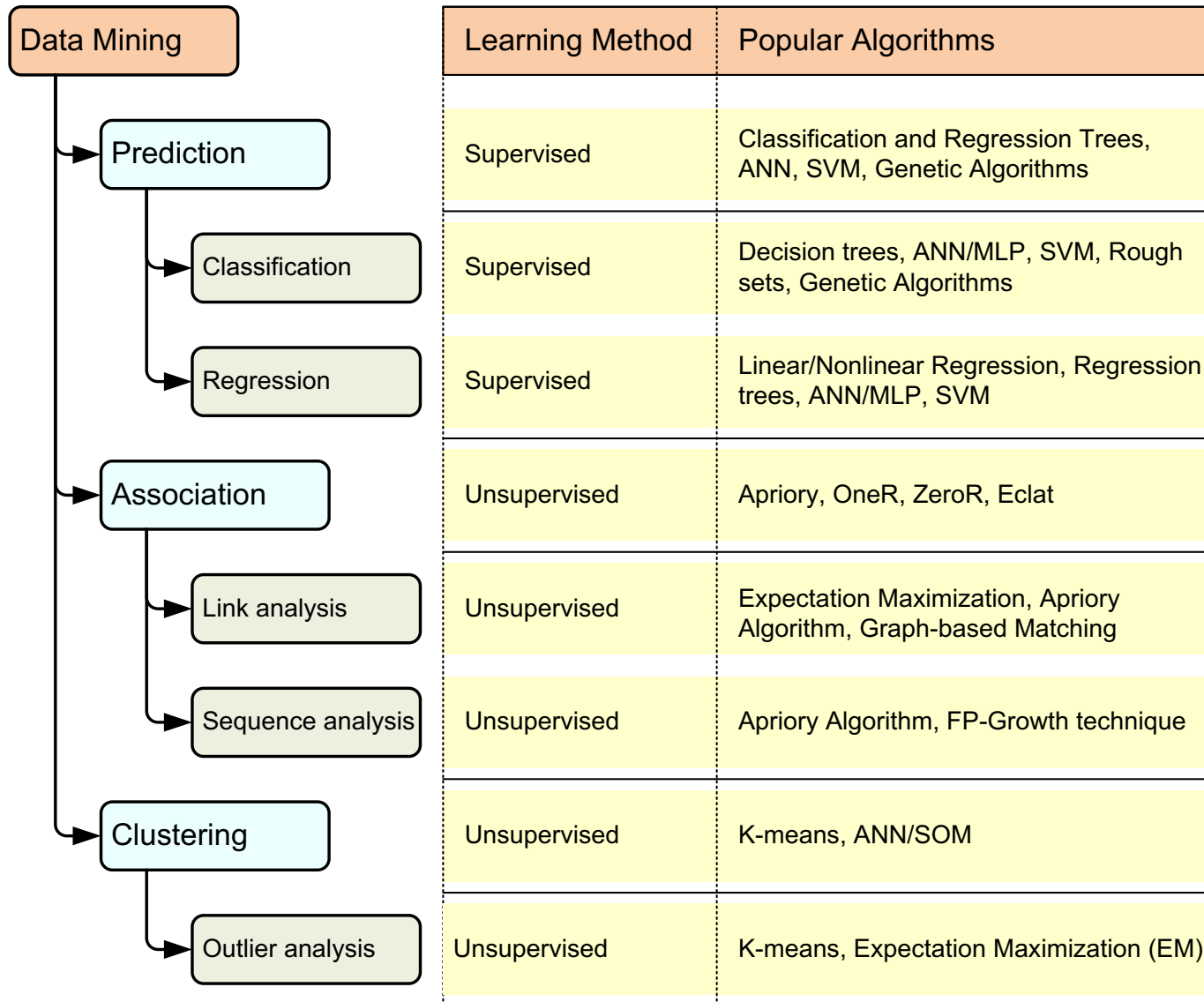


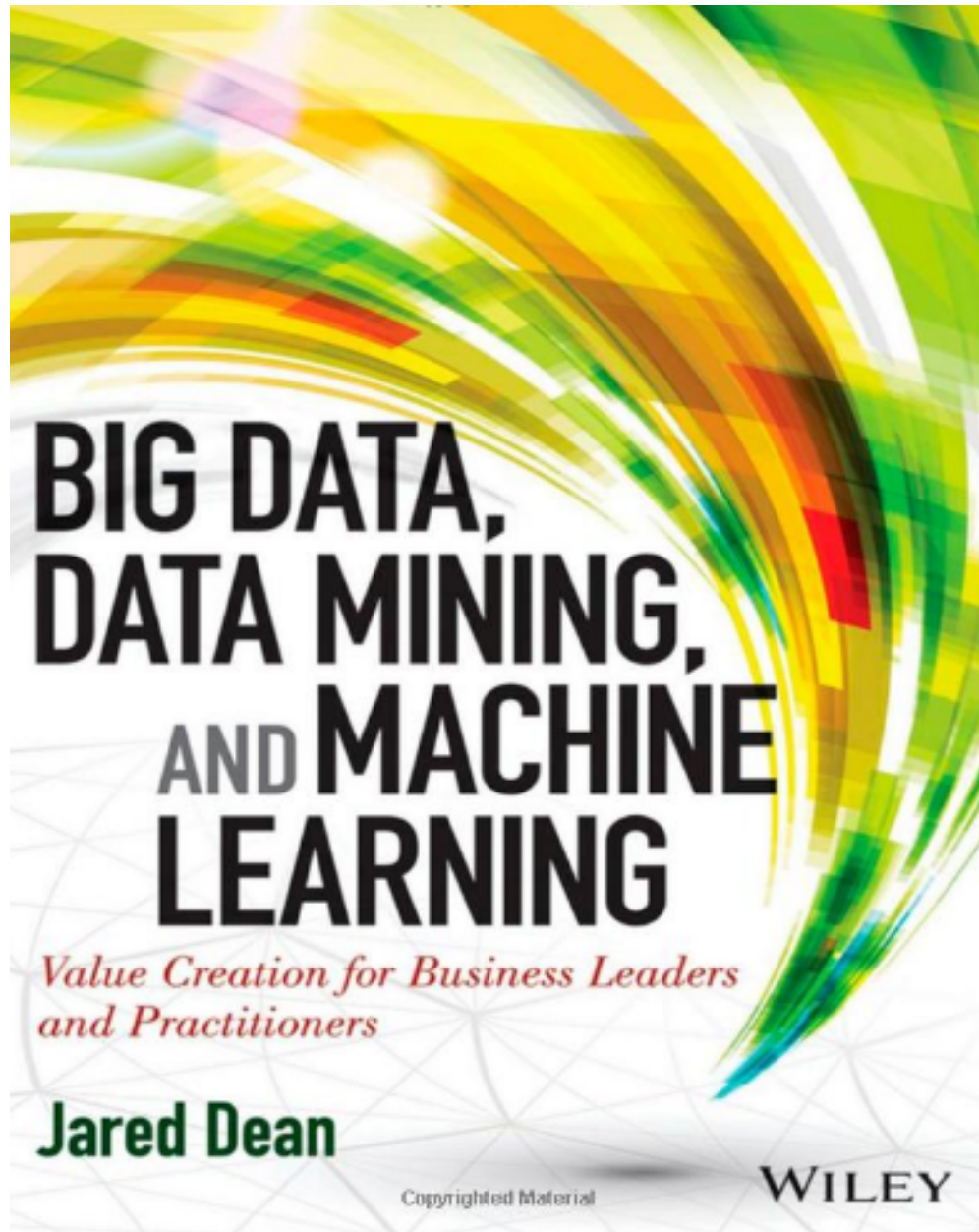
Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



A Taxonomy for Data Mining Tasks

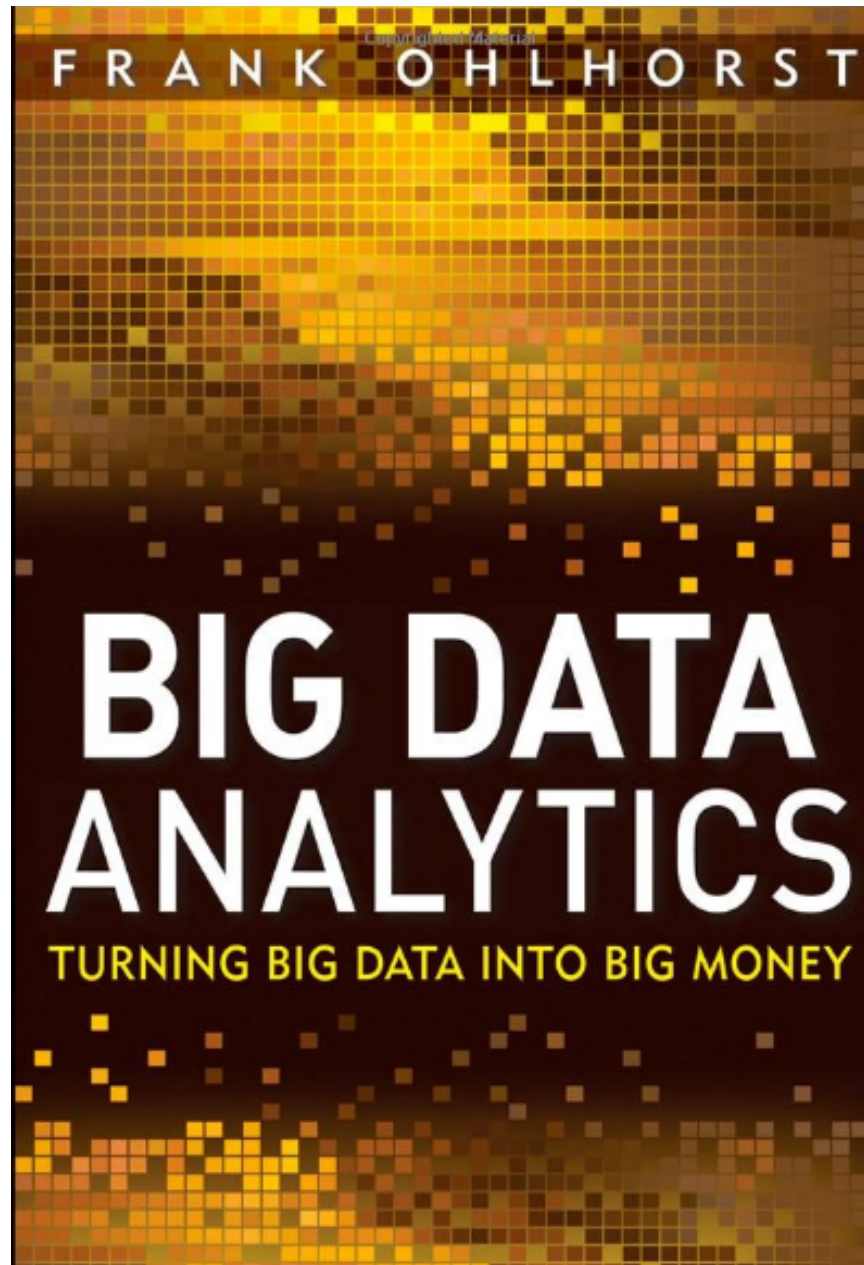




Deep Learning

Intelligence from Big Data







National
Security

Cyber
security

Maritime
security

Smarter
Transport

...

VISUAL ANALYTICS

DYNAMIC & INTERACTIVE

Dashboard Graph
Map

ENHANCE

Understanding Investigation
User Experience



BIG ANALYTICS

QUERY & FILTER

Complex queries
 R^2I^2

DETECT

Anomalies
Communities
Typologies

PREDICT

Trending
Real-time
Prediction

DECIDE

Simulation
Optimization



BIG DATA – Batch



BIG DATA – Real Time

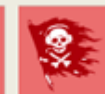


Complex by nature



DATA

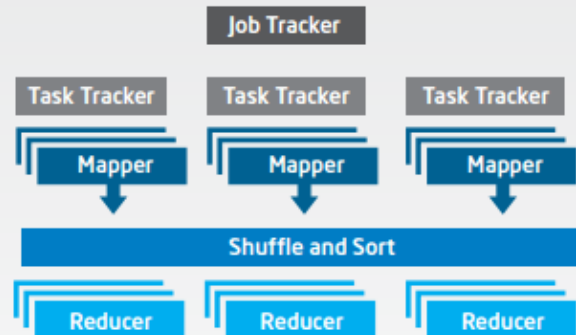
Complex by structure



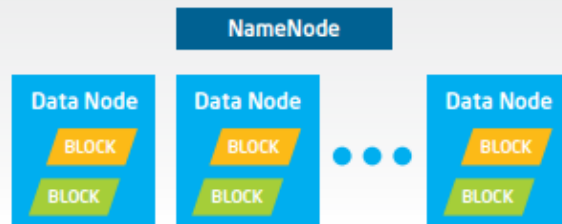
Big Data with Hadoop Architecture

LOGICAL ARCHITECTURE

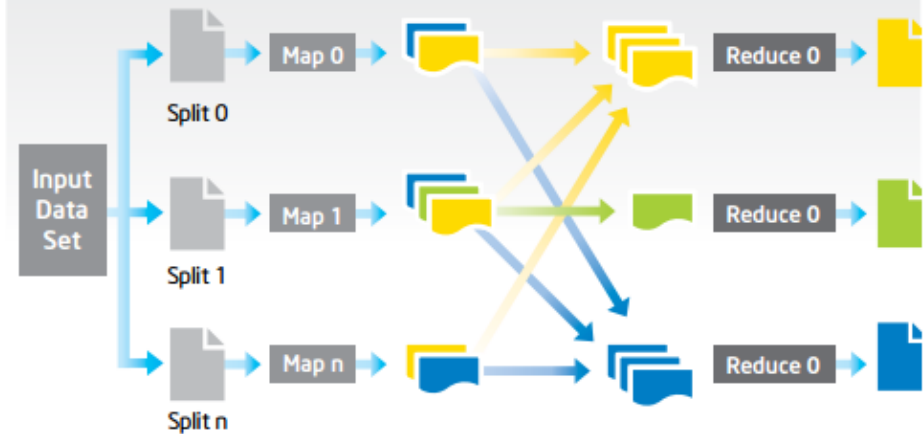
Processing: MapReduce



Storage: HDFS

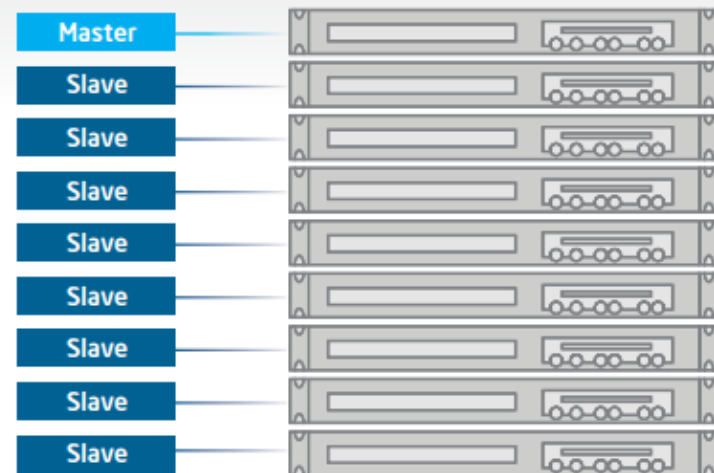


PROCESS FLOW



PHYSICAL ARCHITECTURE

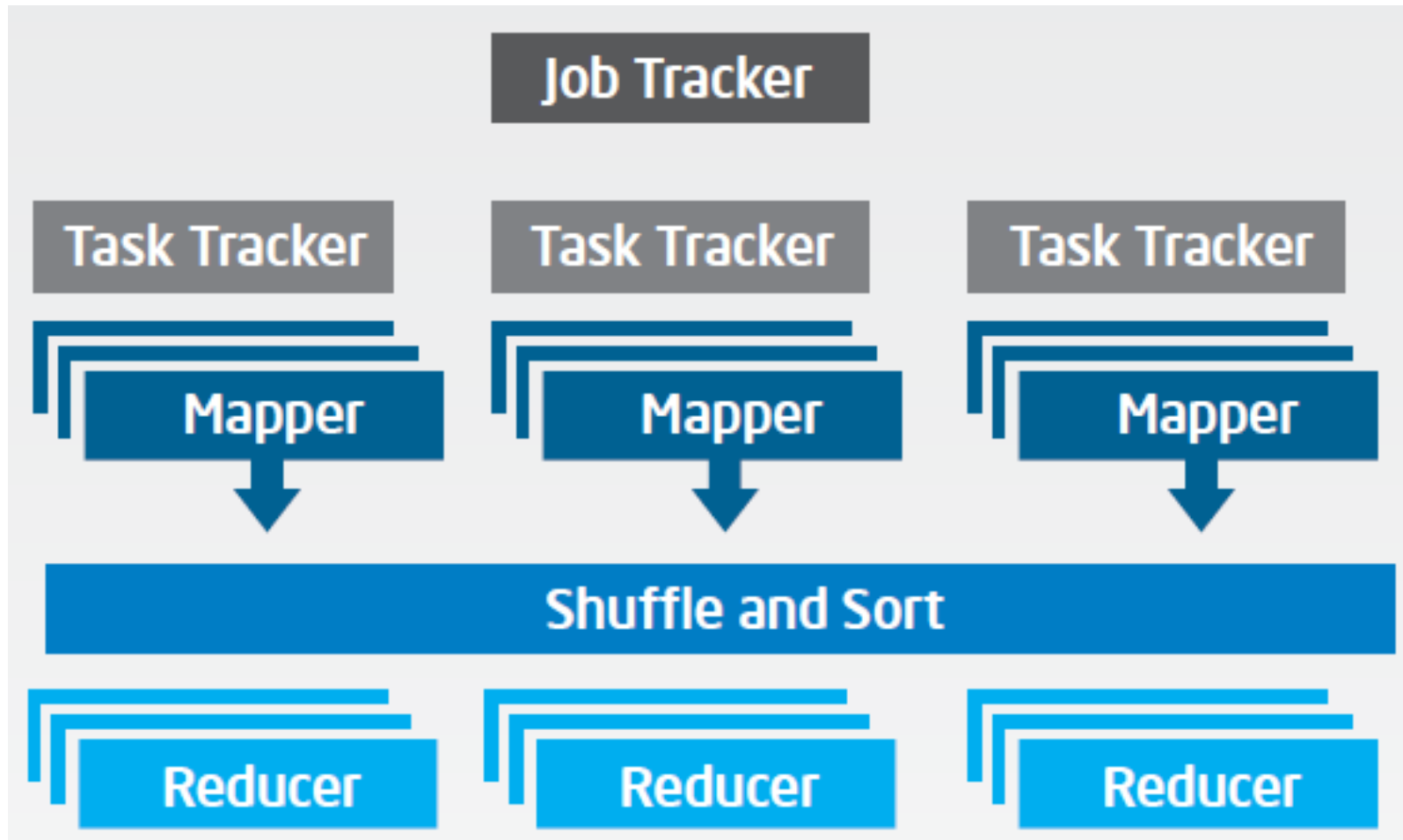
Hadoop Cluster



Big Data with Hadoop Architecture

Logical Architecture

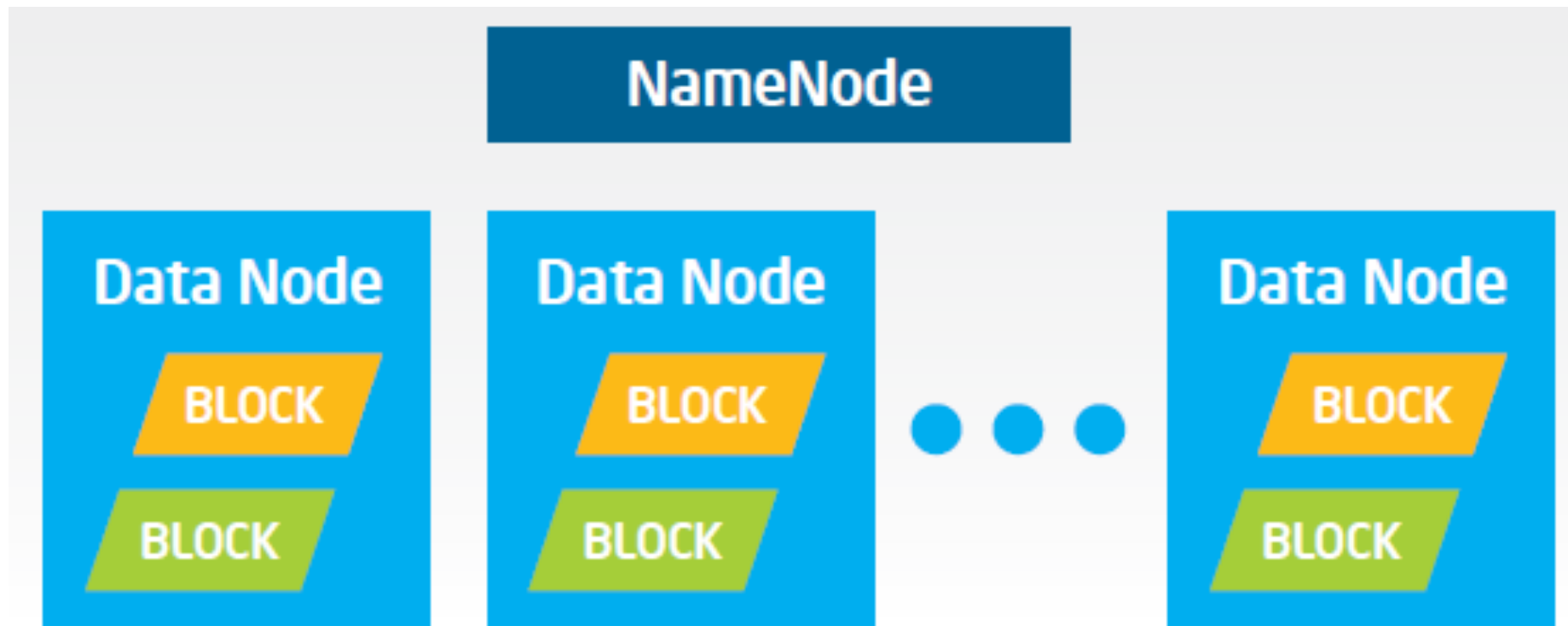
Processing: MapReduce



Big Data with Hadoop Architecture

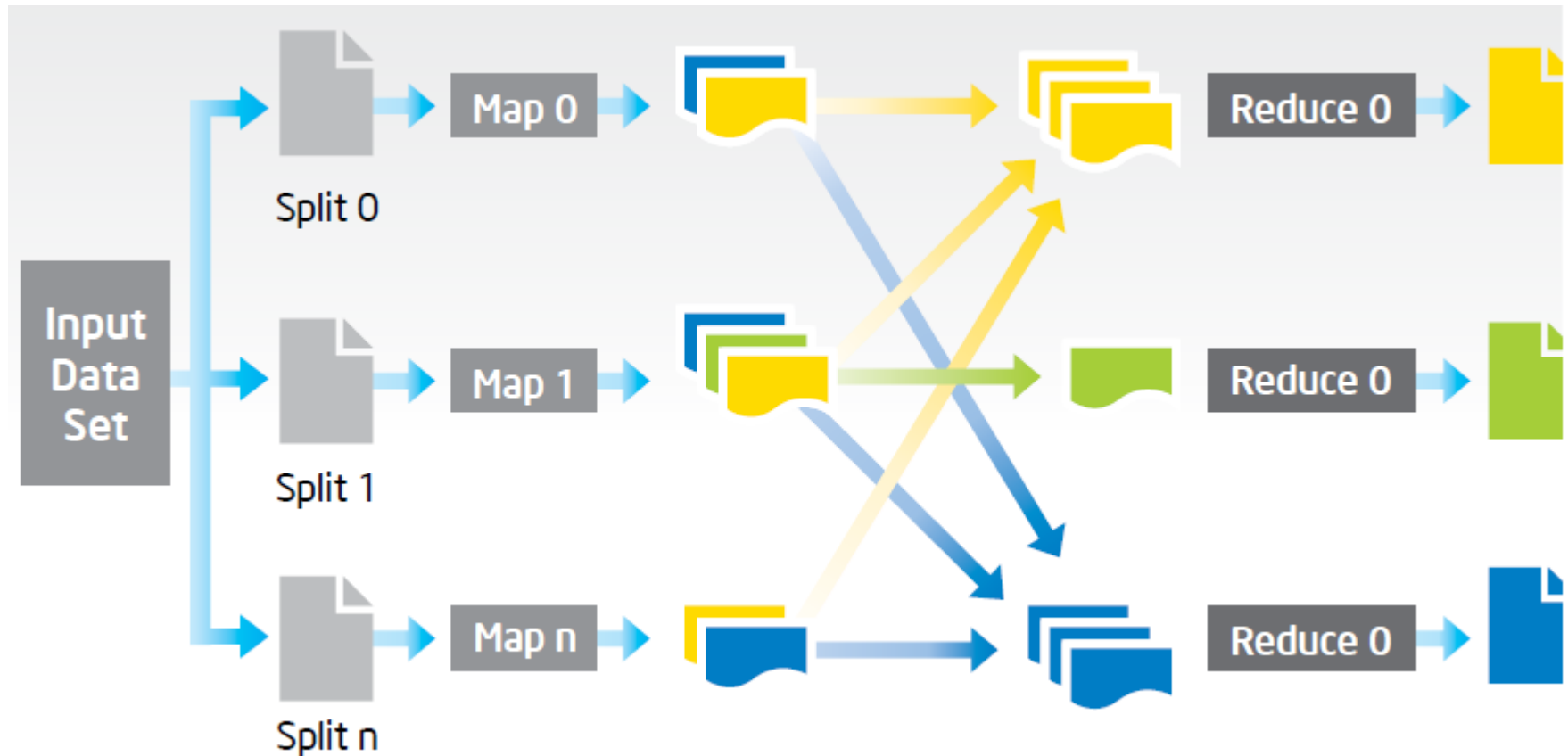
Logical Architecture

Storage: HDFS



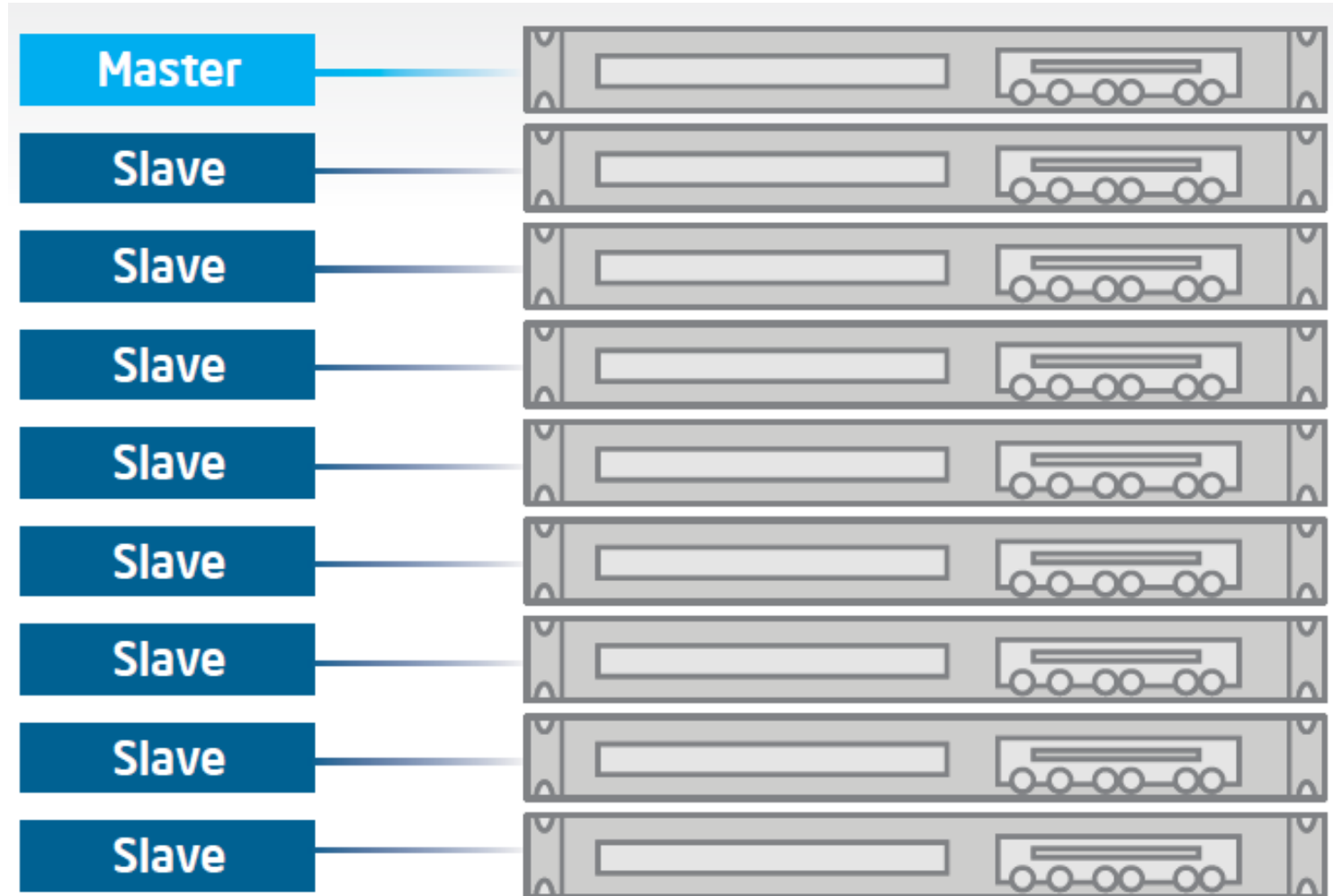
Big Data with Hadoop Architecture

Process Flow

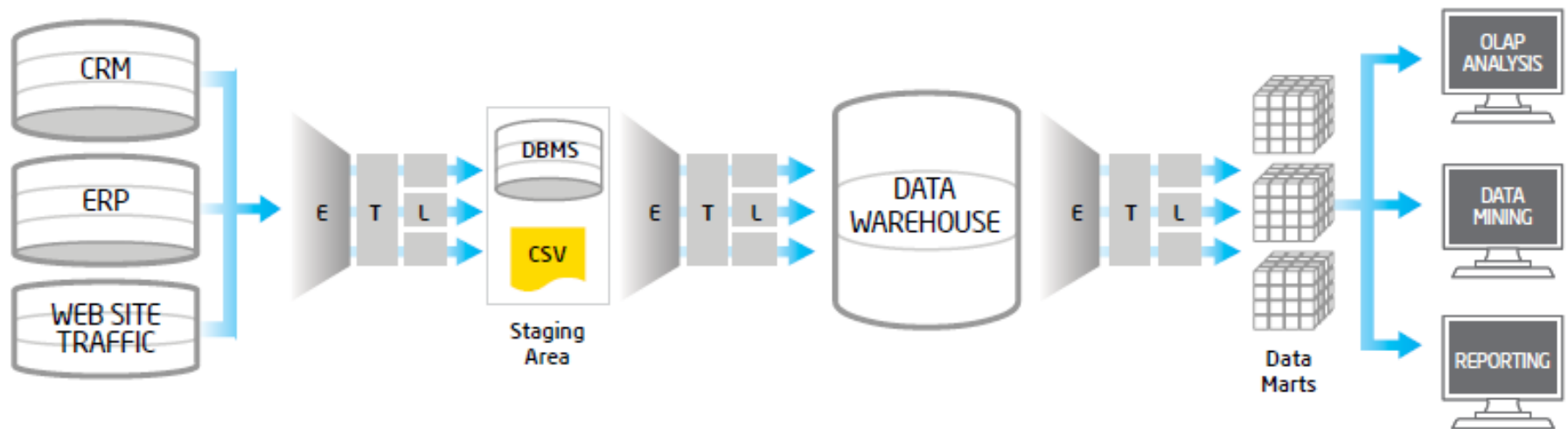


Big Data with Hadoop Architecture

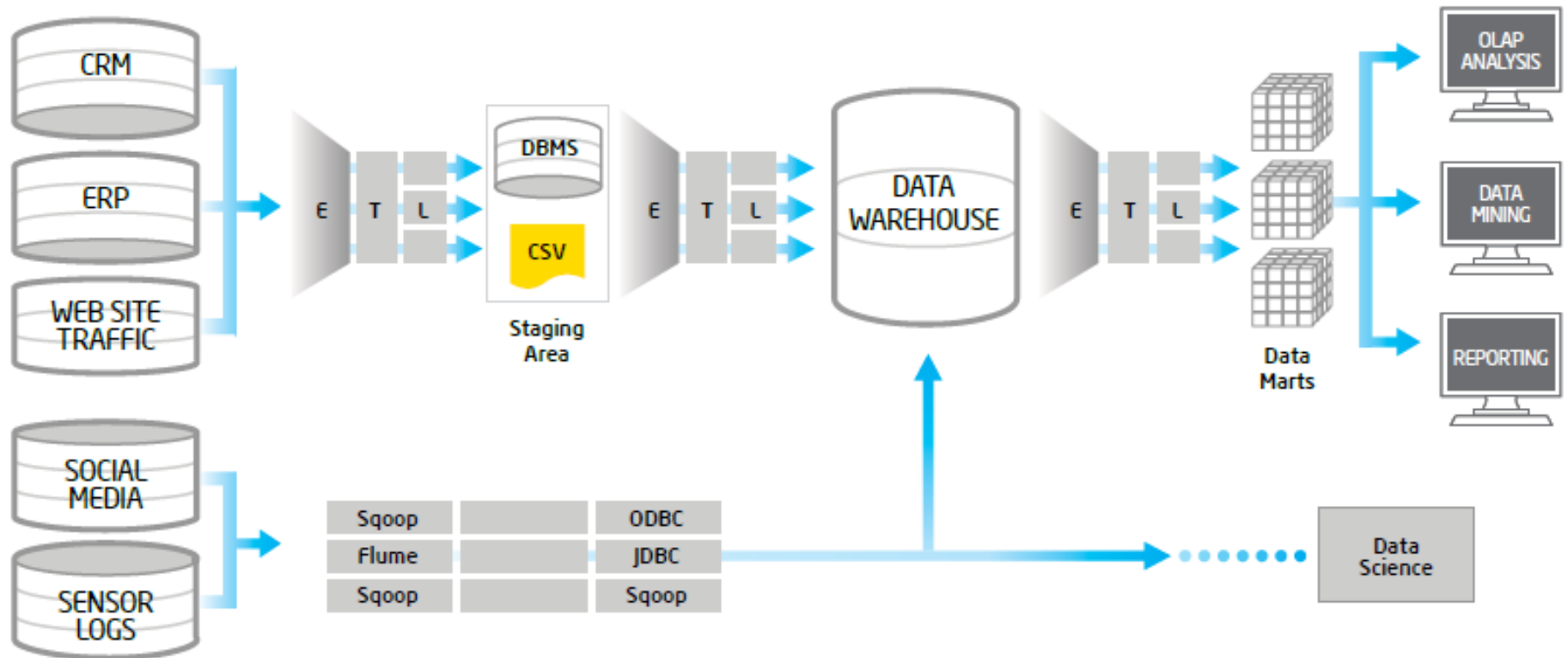
Hadoop Cluster



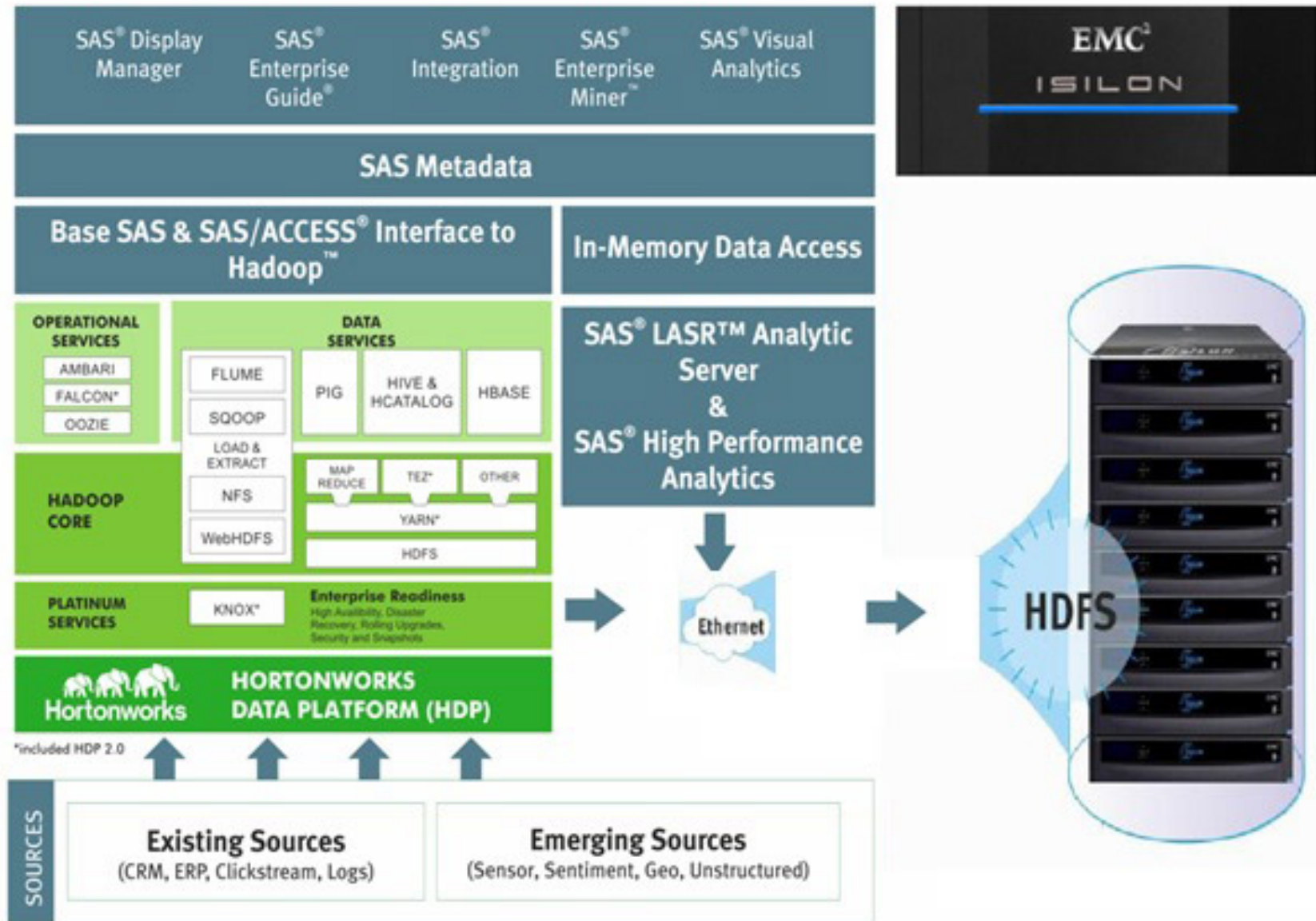
Traditional ETL Architecture



Offload ETL with Hadoop (Big Data Architecture)

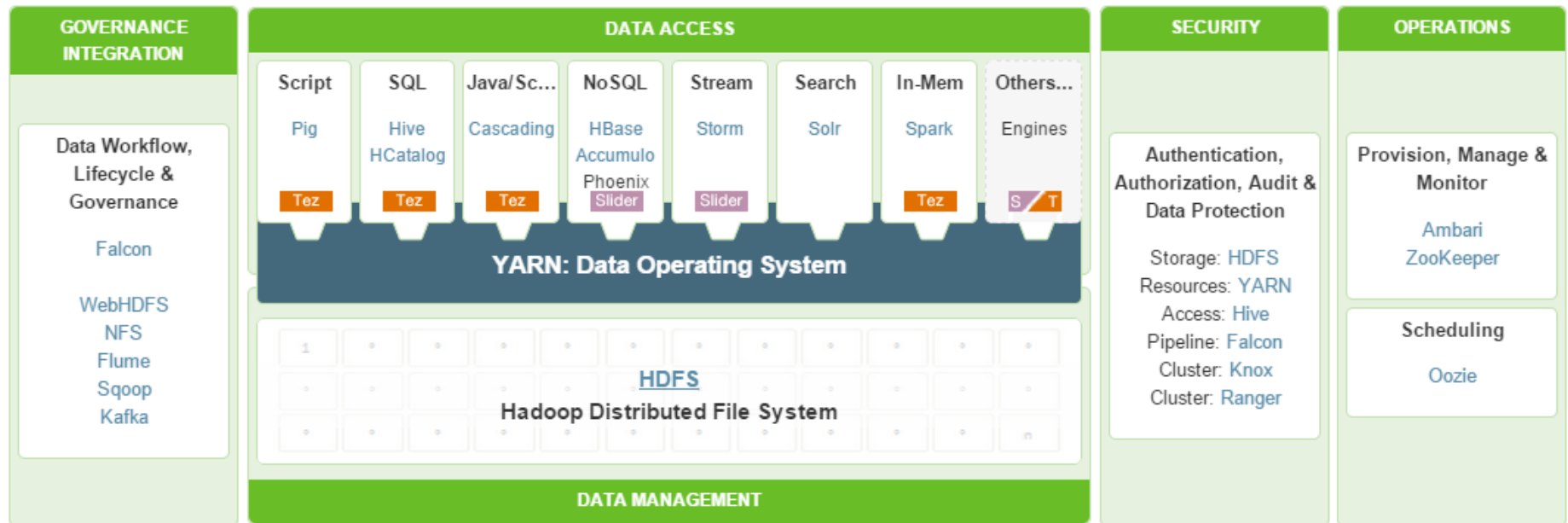


Big Data Solution



HDP

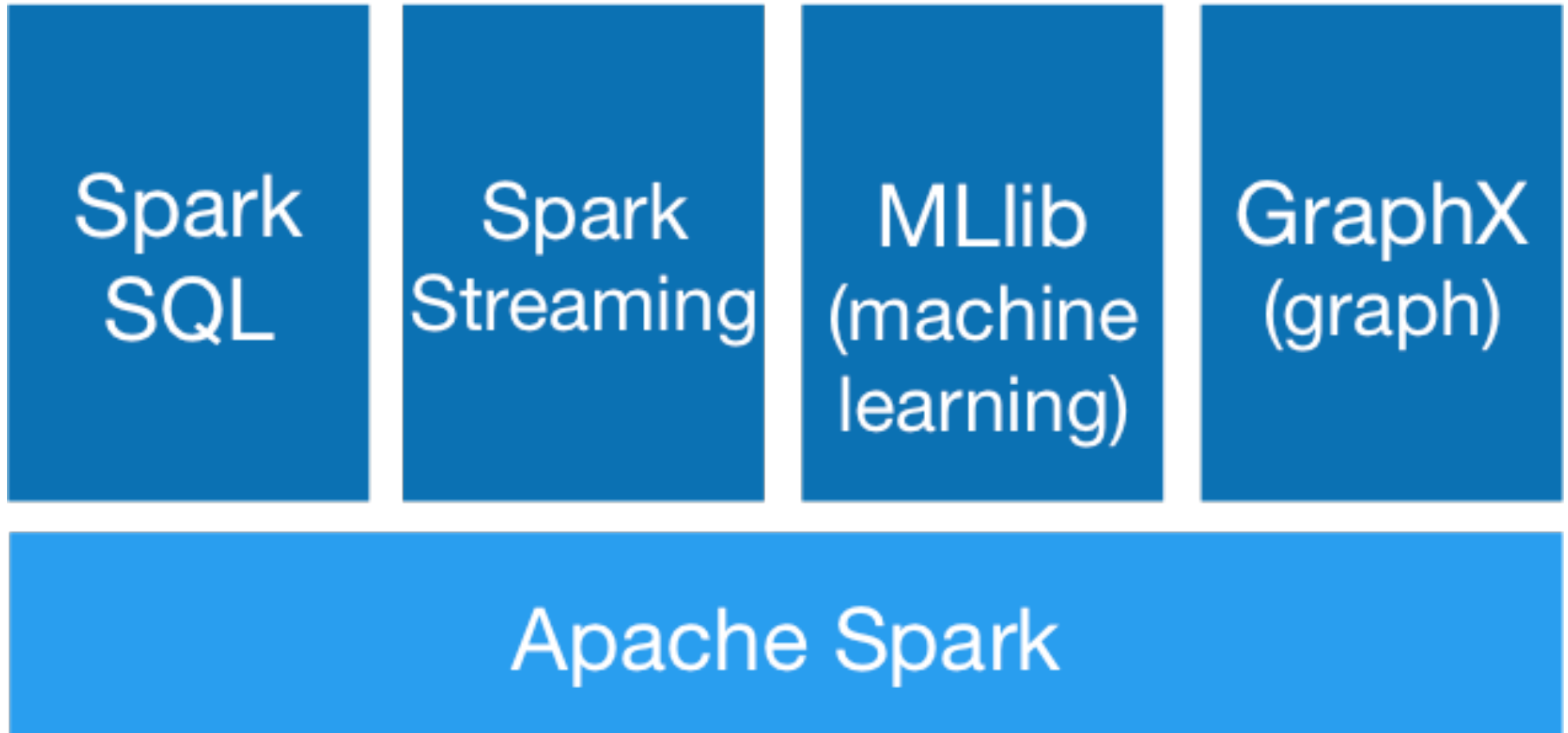
A Complete Enterprise Hadoop Data Platform



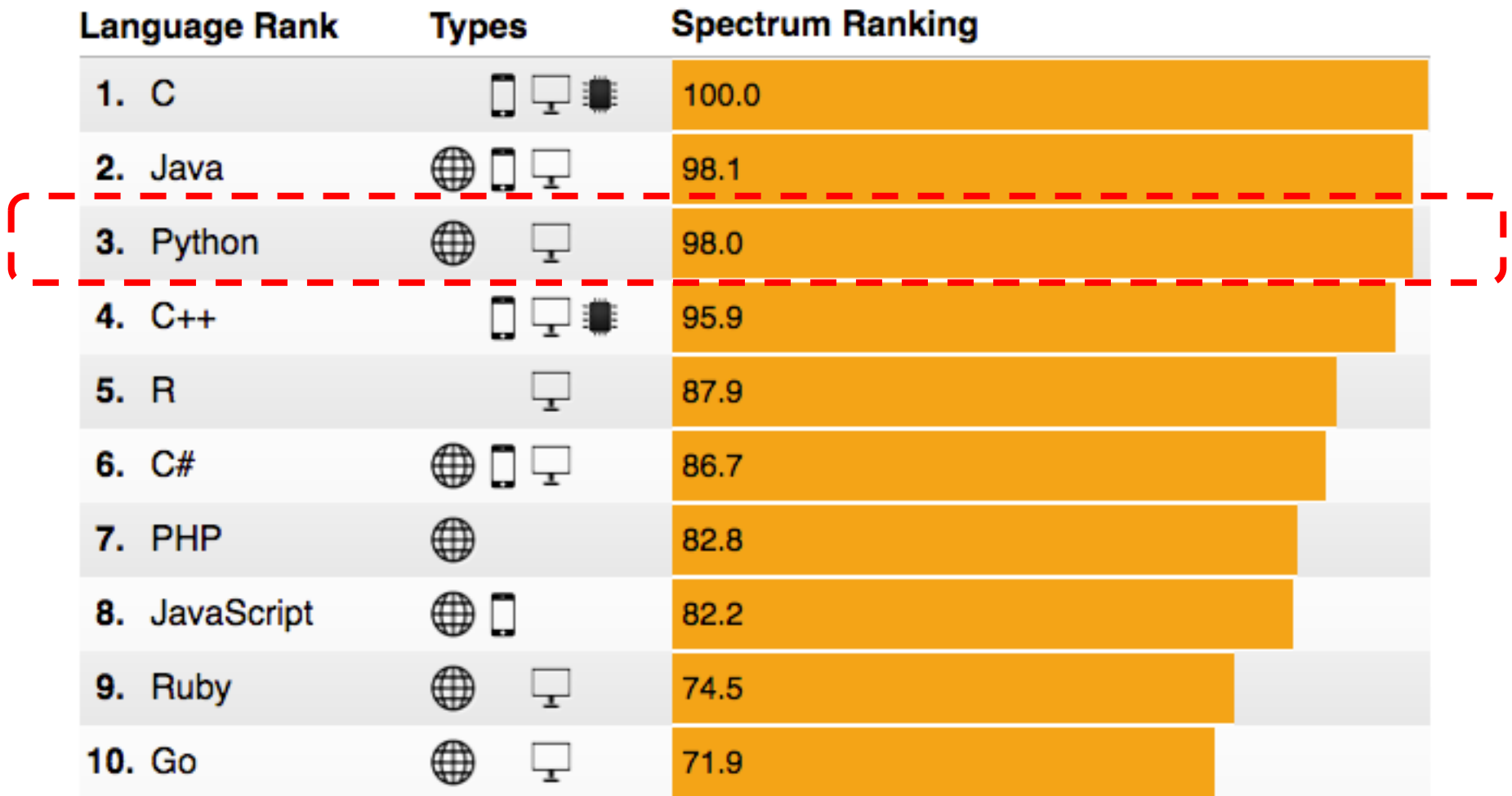
Spark and Hadoop



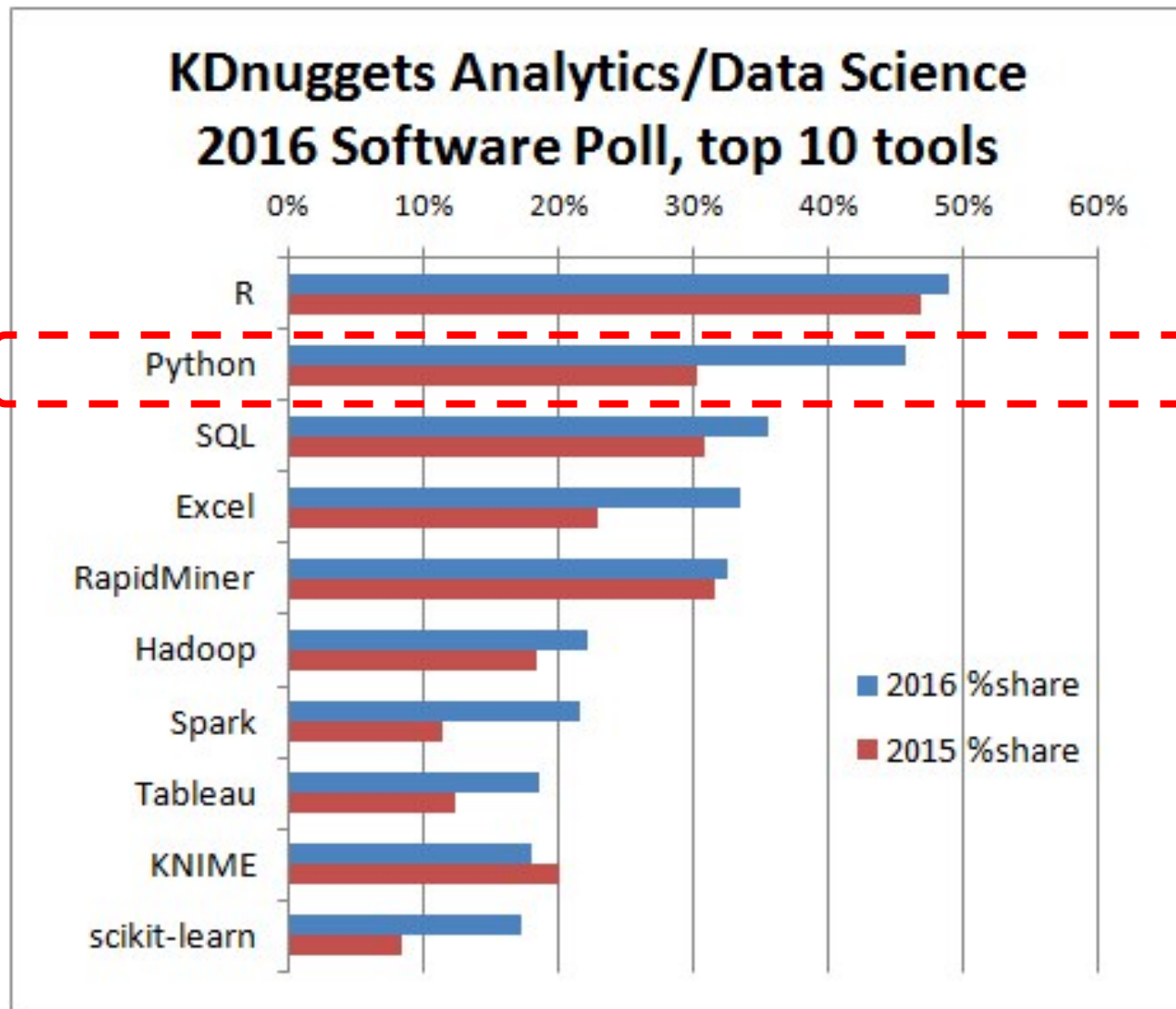
Spark Ecosystem



Python for Big Data Analytics

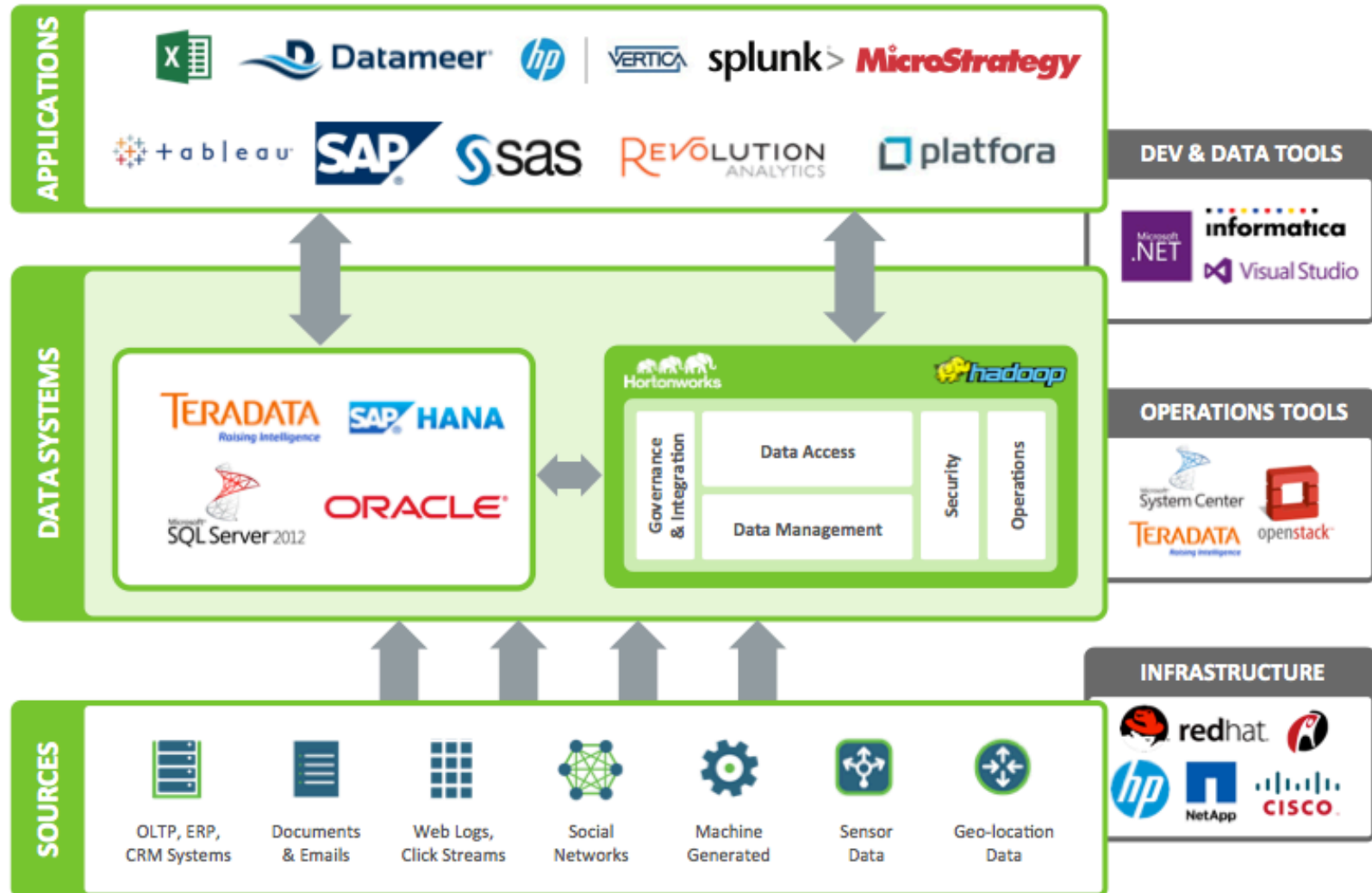


Python: Analytics and Data Science Software



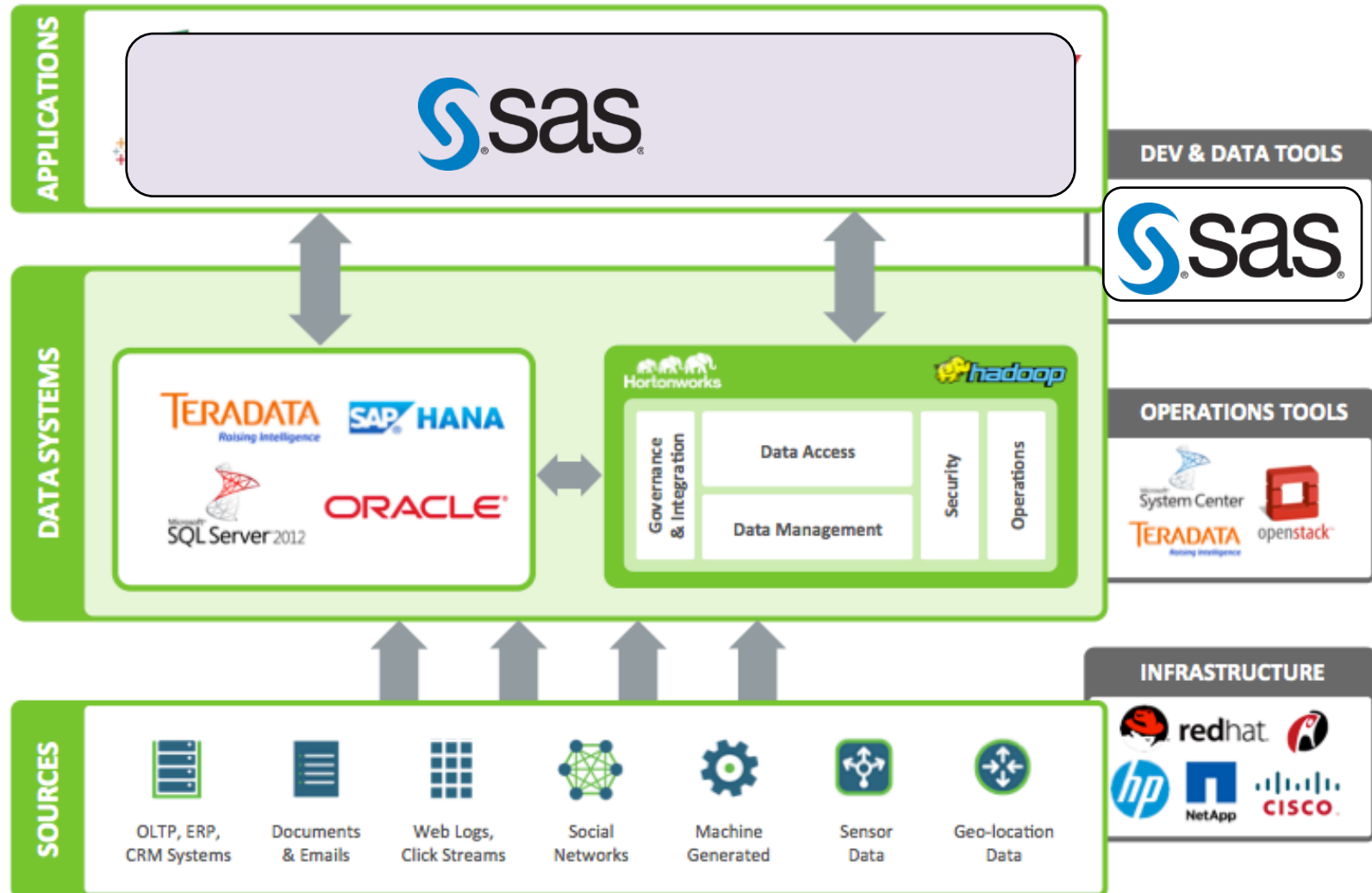
SAS Big data Strategy

– SAS areas

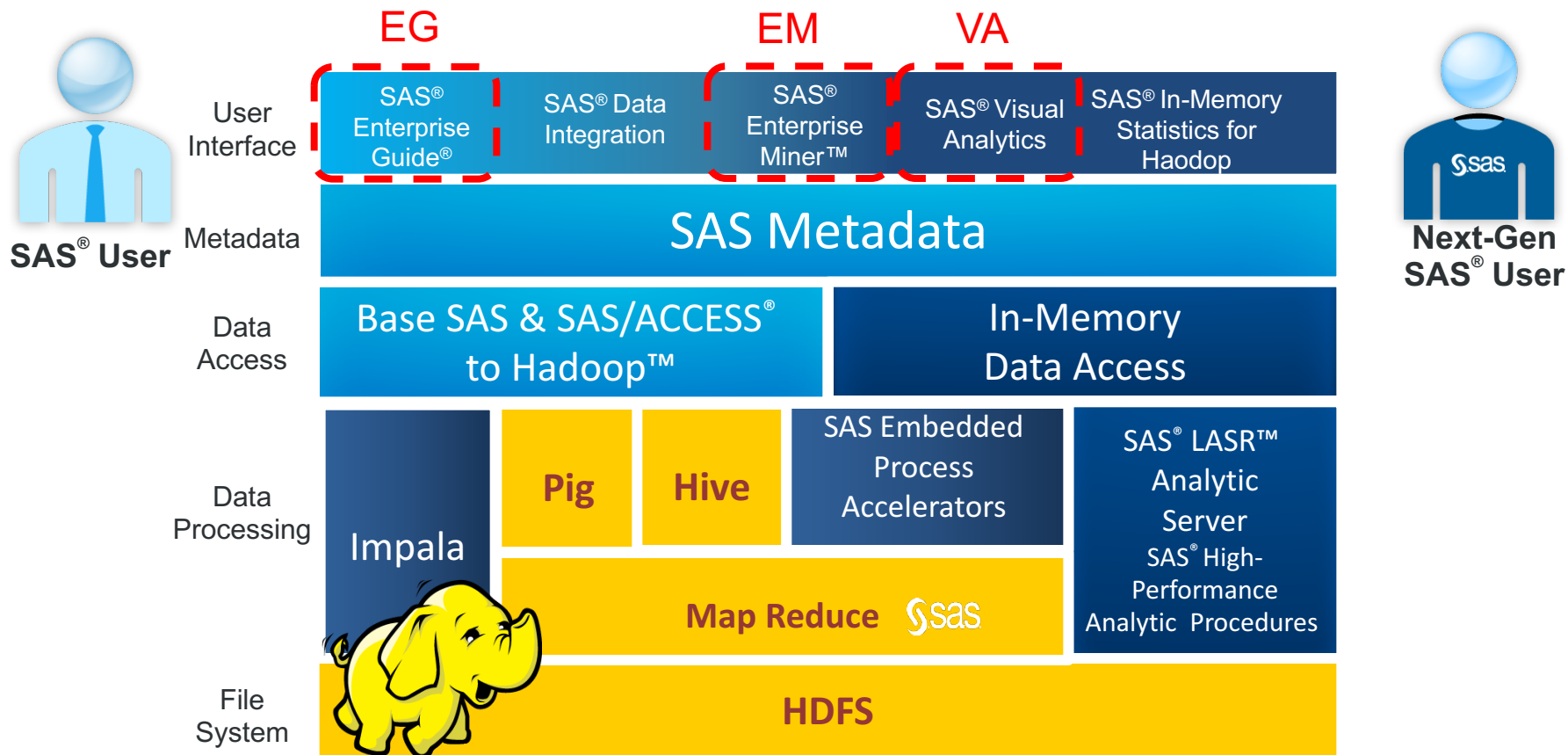


SAS Big data Strategy

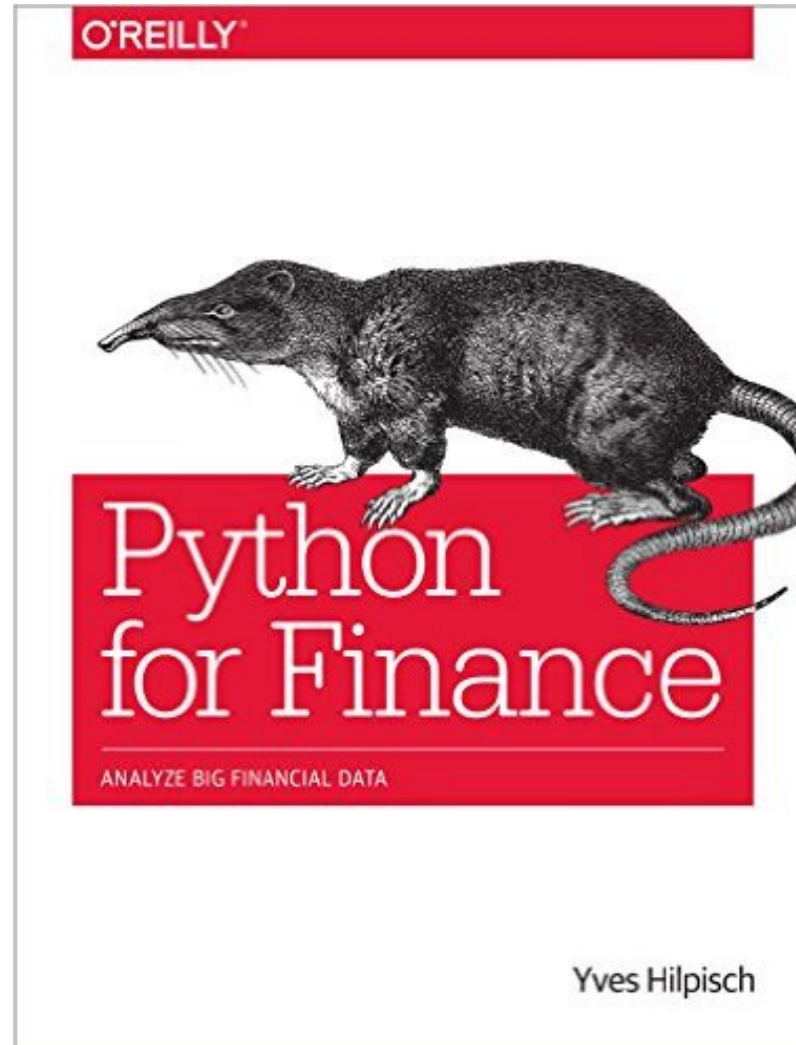
– SAS areas



SAS® Within the HADOOP ECOSYSTEM



Yves Hilpisch, Python for Finance: Analyze Big Financial Data, O'Reilly, 2014

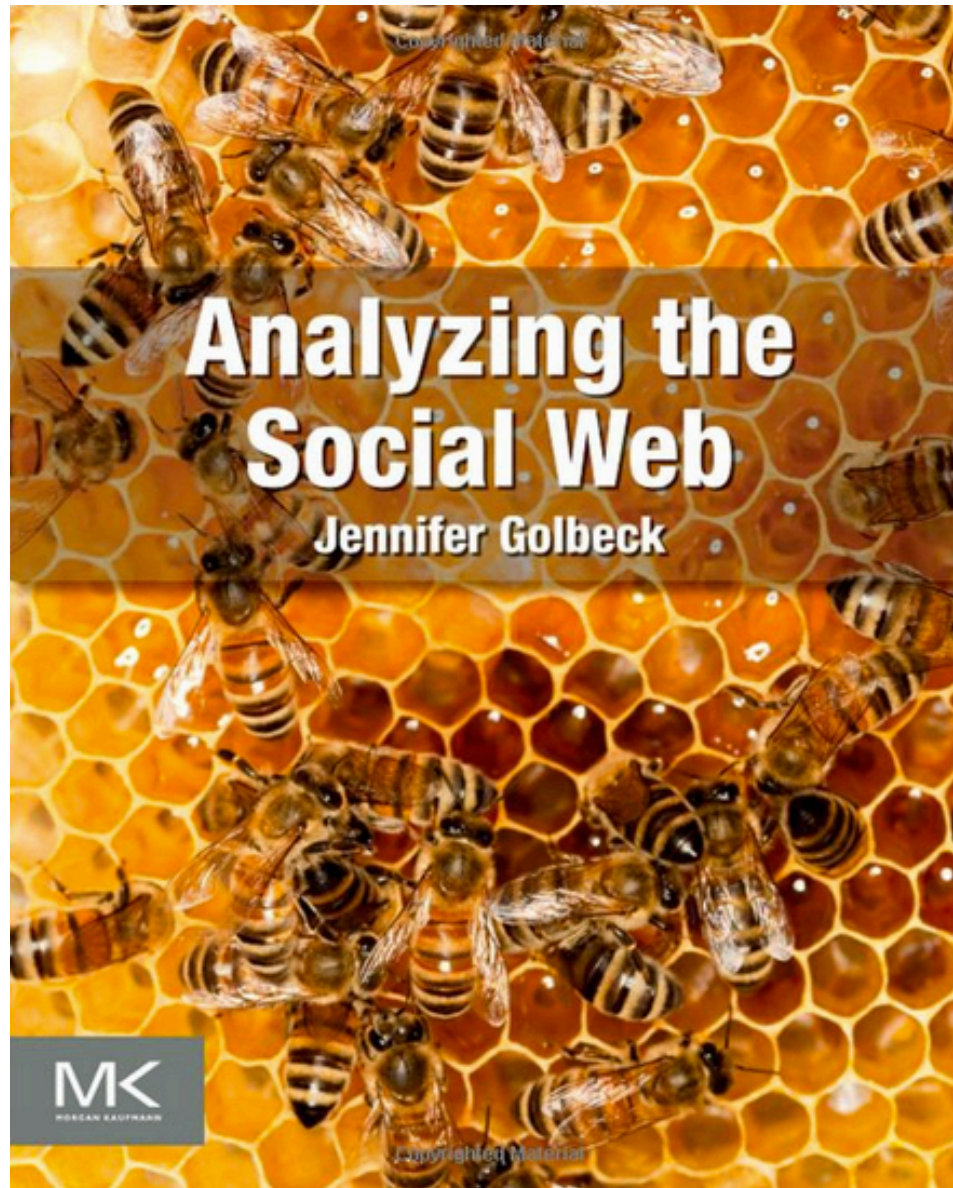


Business Insights with Social Analytics

Analyzing the Social Web:

Social Network Analysis

Jennifer Golbeck (2013), **Analyzing the Social Web**, Morgan Kaufmann



Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

*Analyzing Data from Facebook, Twitter, LinkedIn,
and Other Social Media Sites*

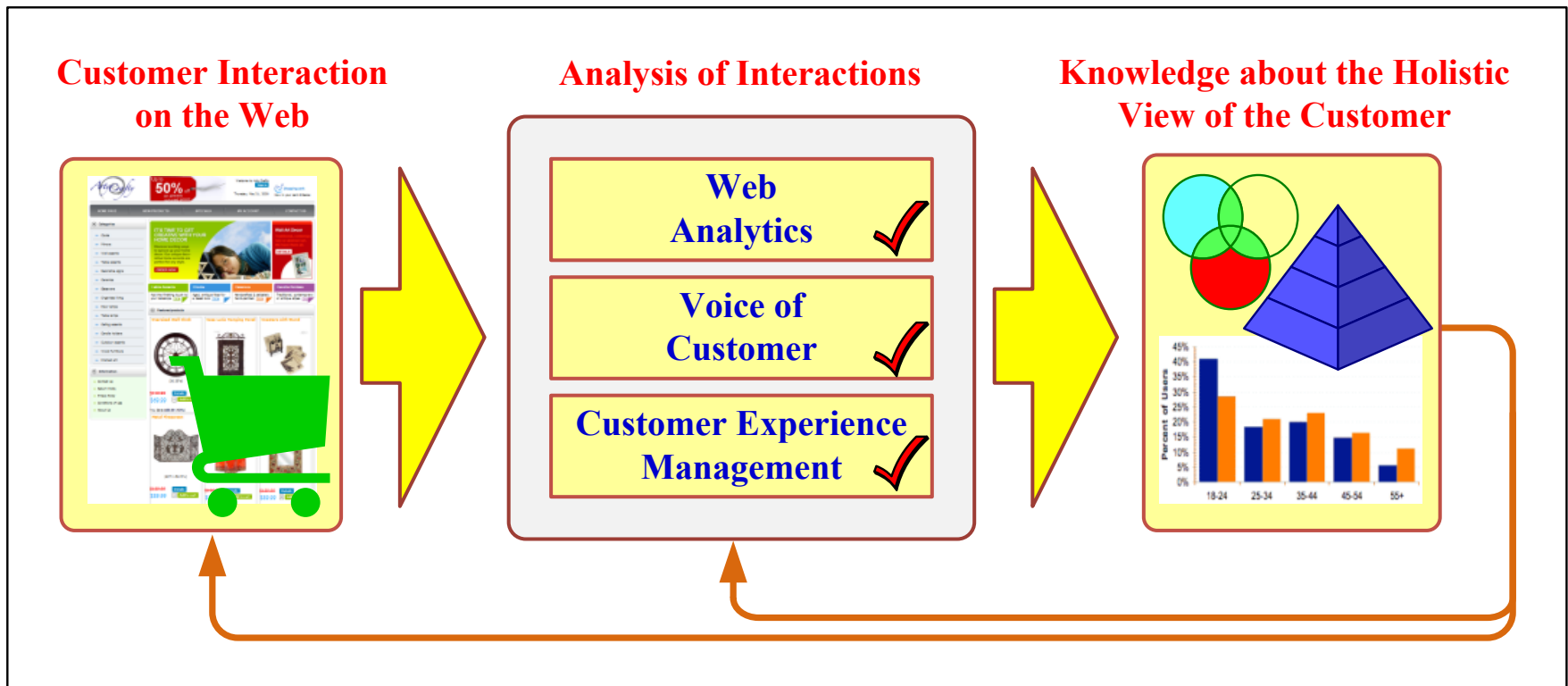


O'REILLY®

Matthew A. Russell

Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, ...
- Website Optimization Ecosystem



Business Intelligence Trends

1. **Agile** Information Management (IM)
2. **Cloud** Business Intelligence (BI)
3. **Mobile** Business Intelligence (BI)
4. **Analytics**
5. **Big Data**

Business Intelligence Trends: Computing and Service

- Cloud Computing and Service
- Mobile Computing and Service
- Social Computing and Service

Business Intelligence and Analytics

- Business Intelligence 2.0 (BI 2.0)
 - Web Intelligence
 - Web Analytics
 - Web 2.0
 - Social Networking and Microblogging sites
- Data Trends
 - Big Data
- Platform Technology Trends
 - Cloud computing platform

Business Intelligence and Analytics: Research Directions

1. Big Data Analytics

- Data analytics using Hadoop / MapReduce framework

2. Text Analytics

- From Information Extraction to Question Answering
- From Sentiment Analysis to Opinion Mining

3. Network Analysis

- Link mining
- Community Detection
- Social Recommendation

Harvard Business Review

HBR.ORG



OCTOBER 2012

REPRINT R1210C

SPOTLIGHT ON BIG DATA

Big Data: The Management Revolution

Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision-making culture.

by Andrew McAfee and Erik Brynjolfsson

Data Scientist:

The Sexiest Job of the 21st Century

**Meet the people who
can coax treasure out of
messy, unstructured data.**

*by Thomas H. Davenport
and D.J. Patil*

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

SAS第六屆大數據資料科學家競賽

FinTech預測未來挑戰賽

[最新消息](#)[大賽起源](#)[活動辦法](#)[我要報名](#)[常見問題](#)

SAS與玉山銀行

第六屆 大數據資料科學家競賽

校園競賽

FinTech

數據分析培訓專業課程資格

SAS與玉山銀行 優先面試與招募

挑戰 **\$300,000** 總獎金

預測未來 挑 · 戰 · 賽

主辦單位



THE
POWER
TO KNOW.



玉山銀行 E.SUN BANK

FinTech 預測未來挑戰賽

在這個巨量資料的時代，懂得巨量分析的專業人才「資料科學家」

(Data Scientist) 將成為未來炙手可熱的明日之星。SAS 希望學生以創意無限及發掘新商機的角度出發，搭配巨量資料分析實例主題，鼓勵全國大學以分組專案及簡報競賽方式，分析高達兩千萬筆的巨量資料，親身體驗巨量分析的神奇魔力！

早鳥報名・優惠方案

報名成功者，並於**2017年3月5日前匯款完畢**

即享有**八折早鳥報名優惠！**

(原報名費每隊1000元，早鳥優惠價每隊800元)

我要報名

<http://saschampion.com.tw/>

SAS 第六屆大數據資料科學家競賽

FinTech 預測未來挑戰賽

[最新消息](#)[大賽起源](#)[活動辦法](#)[我要報名](#)[常見問題](#)

活動時間與地點:

1. 報名日期：2017年2月20日（一）至2017年3月10日（五）額滿為止
2. 起跑說明會：2017年3月17日（五）下午六點半至八點半止（每組皆須指派隊員出席，須事先報名）
3. 玉山銀行玉山人力發展中心1樓 登峰廳（台北市中山區撫順街41巷13號1樓）
4. 初賽資料分析訓練課程(Enterprise Guide)：2017年3月20日（一）至 3月26日（日），
每梯次為期1天(每梯次名額有限，依名額順序額滿為止，活動執行單位將通知參賽者可參加場次)

初賽【EG個人能力檢測】：2017 年 4 月 22 日（六）下午一點半至四點止

入圍複賽公布日期：2017 年 4 月 26 日（三）

複賽密集實戰課程(SAS密集實戰課程)：2017 年 4 月 28 日（五）及 4 月 29 日（六）共2梯次，於台北大學資訊中心教室舉辦，每梯次為期1天，時間由主辦單位安排並通知，若該堂時間無法參與，請於收到通知後二天內提出相關證明，以利其他課程之安排與協調。

***備註：入圍複賽之隊伍方可參加**

複賽比賽日期：2017 年 5 月 01 日（一）～ 2017 年 5 月 19 日（五）下午五點止

入圍決賽公布日期：2017 年 6 月 2 日（五）下午五點

決賽日期：2017 年 6 月 9 日（五）賽仕電腦軟體股份有限公司（台北市中山區民生東路三段10號3樓）

公布得獎名單日期：2017 年 6 月 9 日（五）晚上九點

頒獎典禮：2017 年 6 月 27 日（二）

<http://saschampion.com.tw/detail.php>

The 13th NTCIR (2016 - 2017)

NTCIR (NII Testbeds and Community for Information access Research) Project



Publications/
Online Proceedings

Data/Tools

NTCIR CMS Site

Related URL's

Contact us

[NTCIR Home](#) > [NTCIR-13](#)

NTCIR 13

NTCIR-13 Conference

NEWS

NTCIR-13 Aims

Call for Task Proposals

How to Participate NEW

Task Participation NEW

Task Overview/Call for
Task Participation

User Agreement Forms
NEW

Organization

Important Dates

Contact Us

NTCIR 12

NTCIR-13

The 13th NTCIR (2016 - 2017)

Evaluation of Information Access Technologies

June 2016 - December 2017

Conference: December 5-8, 2017, NII, Tokyo, Japan

What's New

NEW December 16, 2016: [NTCIR-13 Task Registration is still possible in each tasks after December 15, 2016 \(final deadline updated on Dec. 22, 2016\)](#)

[Lifelog-2](#): until Jan 15, 2017 (for Phase I) and until Jun 15, 2017 (for Phase II)

[MedWeb](#): until March 31, 2017

[OpenLiveQ](#): until Feb 28, 2017

[QALab-3](#): until Jan 26, 2017 (for Phase-1) and until May 1, 2017 (for Phase-2) [[see detailed schedule here](#)].

[STC-2](#): until April 30, 2017

[AKG](#): until Dec 31, 2016 (for AM subtask) and until Jun 1, 2017 (for AKGG subtask)

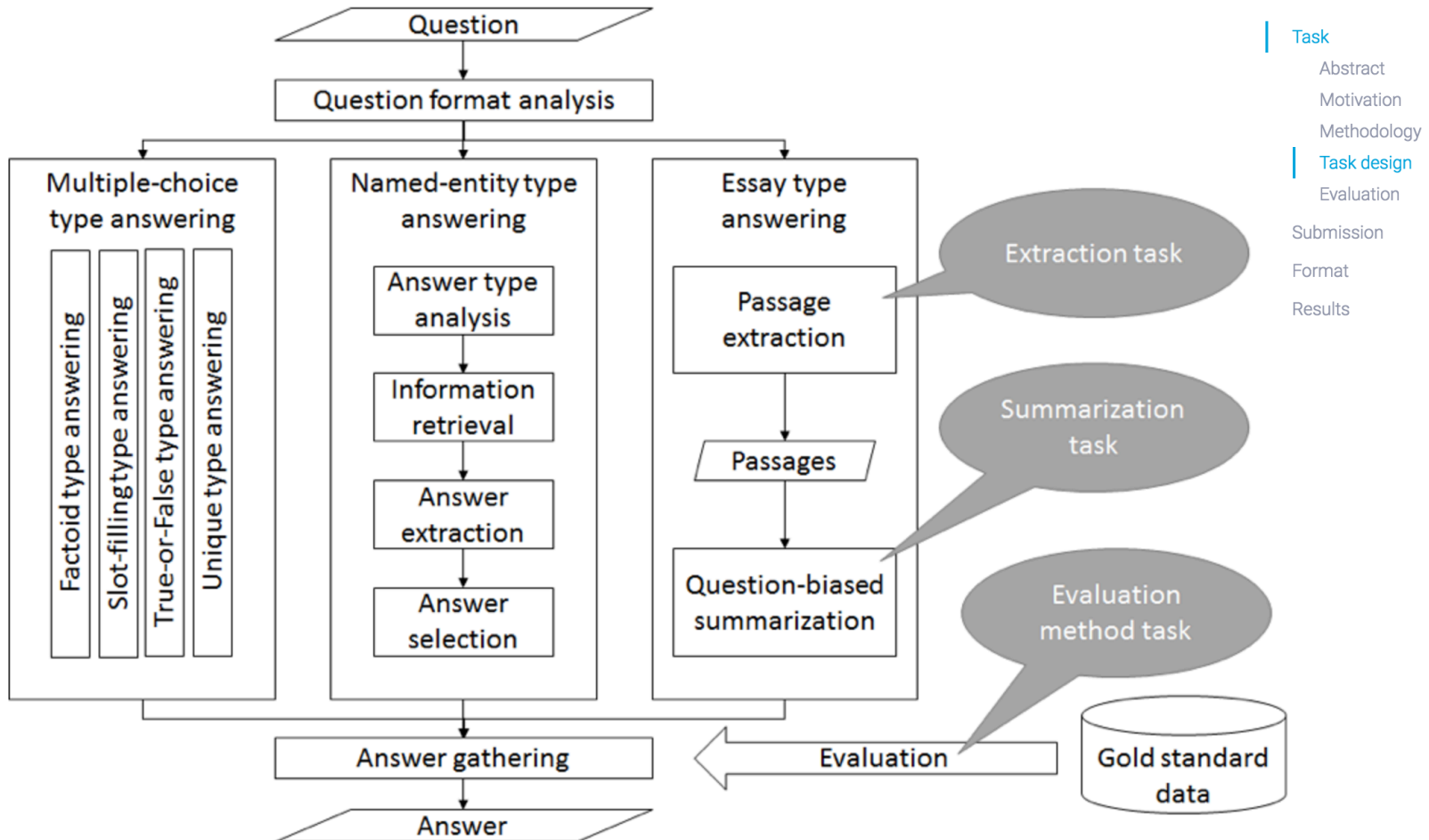
[ECA](#): until Dec 1, 2016 (The registration deadline was over)

[NAILS](#): until March 15, 2017

[WWW](#): until April 30, 2017

<http://research.nii.ac.jp/ntcir/ntcir-13/index.html>

NTCIR-13 QALab-3



Summary

- This course introduces the fundamental concepts and applications technology of big data mining.
- Topics include
 - Big Data Mining
 - Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
 - Association Analysis
 - Classification and Prediction
 - Cluster Analysis
 - Data Mining Using SAS Enterprise Miner (SAS EM)
 - Case Study and Implementation of Big Data Mining
 - Deep Learning with Google TensorFlow

Contact Information

戴敏育 博士 (Min-Yuh Day, Ph.D.)

專任助理教授

淡江大學 資訊管理學系

電話：02-26215656 #2846

傳真：02-26209737

研究室：B929

地址：25137 新北市淡水區英專路151號

Email：myday@mail.tku.edu.tw

網址：<http://mail.tku.edu.tw/myday/>

