

Social Computing and Big Data Analytics

社群運算與大數據分析

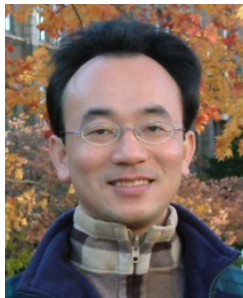
Course Orientation for Social Computing and Big Data Analytics

(社群運算與大數據分析課程介紹)

1052SCBDA01

MIS MBA (M2226) (8606)

Wed, 8,9, (15:10-17:00) (B505)



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/>

2017-02-15



Social Computing
and
Big Data Analytics
(社群運算
與
大數據分析)

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：社群運算與大數據分析
(Social Computing and Big Data Analytics)
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系級：資管所碩士班(TLMXM1A)
- 開課資料：選修 單學期 2 學分 (2 Credits, Elective)
- 上課時間：週三 8,9 (Wed 15:10-17:00)
- 上課教室：B505 (商管大樓)

課程簡介

- 本課程介紹社群運算與大數據分析的基本概念及研究議題。
- 課程內容包括
 - 資料科學與大數據分析：探索、分析、視覺化與呈現資料
 - 大數據基礎：MapReduce典範、Hadoop與Spark生態系統
 - 大數據處理平台SMACK: Spark, Mesos, Akka, Cassandra and Kafka
 - Python Pandas財務大數據分析
 - 文字探勘分析技術與自然語言處理
 - 社群媒體行銷分析
 - 深度學習 (Deep Learning) 社群媒體情感分析
 - Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)
 - 社會網絡分析、量測、工具

Course Introduction

- This course introduces the **fundamental concepts** and **research issues** of **social computing** and **big data analytics**.
- Topics include
 - **Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data**
 - Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
 - Big Data Processing Platforms with SMACK: Spark, Mesos, Akka, Cassandra and Kafka
 - Big Data Analytics with Numpy in Python
 - **Finance Big Data Analytics with Pandas in Python**
 - Text Mining Techniques and Natural Language Processing
 - Social Media Marketing Analytics
 - **Deep Learning with Theano and Keras in Python**
 - **Deep Learning with Google TensorFlow**
 - **Sentiment Analysis on Social Media with Deep Learning**
 - **Social Network Analysis, Measurements, and Tools**

課程目標 (Objective)

- 瞭解及應用社群運算與大數據分析基本概念與研究議題。

(Understand and apply the fundamental concepts and research issues of Social Computing and Big Data Analytics.)

- 進行社群運算與大數據分析相關之資訊管理研究。

(Conduct information systems research in the context of Social Computing and Big Data Analytics.)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2017/02/15	Course Orientation for Social Computing and Big Data Analytics (社群運算與大數據分析課程介紹)
2	2017/02/22	Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data (資料科學與大數據分析： 探索、分析、視覺化與呈現資料)
3	2017/03/01	Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem (大數據基礎：MapReduce典範、 Hadoop與Spark生態系統)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
4	2017/03/08	Big Data Processing Platforms with SMACK: Spark, Mesos, Akka, Cassandra and Kafka (大數據處理平台SMACK： Spark, Mesos, Akka, Cassandra, Kafka)
5	2017/03/15	Big Data Analytics with Numpy in Python (Python Numpy 大數據分析)
6	2017/03/22	Finance Big Data Analytics with Pandas in Python (Python Pandas 財務大數據分析)
7	2017/03/29	Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理)
8	2017/04/05	Off-campus study (教學行政觀摩日)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
9	2017/04/12	Social Media Marketing Analytics (社群媒體行銷分析)
10	2017/04/19	期中報告 (Midterm Project Report)
11	2017/04/26	Deep Learning with Theano and Keras in Python (Python Theano 和 Keras 深度學習)
12	2017/05/03	Deep Learning with Google TensorFlow (Google TensorFlow 深度學習)
13	2017/05/10	Sentiment Analysis on Social Media with Deep Learning (深度學習社群媒體情感分析)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
14	2017/05/17	Social Network Analysis (社會網絡分析)
15	2017/05/24	Measurements of Social Network (社會網絡量測)
16	2017/05/31	Tools of Social Network Analysis (社會網絡分析工具)
17	2017/06/07	Final Project Presentation I (期末報告 I)
18	2017/06/14	Final Project Presentation II (期末報告 II)

2017/02/22

Data Science and

Big Data Analytics:

Discovering, Analyzing,

Visualizing and Presenting Data

(資料科學與大數據分析：

探索、分析、

視覺化與呈現資料)

2017/03/01

Fundamental Big Data:

MapReduce Paradigm,

Hadoop and Spark Ecosystem

(大數據基礎：

MapReduce典範、

Hadoop與Spark生態系統)

2017/03/08

**Big Data Processing Platforms
with SMACK:**

**Spark, Mesos, Akka,
Cassandra and Kafka**

(大數據處理平台 SMACK :

**Spark, Mesos, Akka,
Cassandra, Kafka)**

2017/03/22

**Finance Big Data Analytics
with Pandas in Python**

(Python Pandas

財務大數據分析)

2017/04/26

Deep Learning with Theano and Keras in Python

(Python Theano
和 Keras 深度學習)

2017/05/03

Deep Learning

with

Google TensorFlow

(Google TensorFlow

深度學習)

2017/05/27

Social Network Analysis

(社會網絡分析)

教學方法與評量方法

- 教學方法

- 講述、討論、
賞析、模擬、
問題解決、實作

- 評量方法

- 實作、報告、上課表現

教材課本

- 教材課本
 - 講義 (Slides)
 - 社群運算與大數據分析相關個案與論文
(Cases and Papers related to Social Computing and Big Data Analytics)

參考書籍

1. EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015
2. Mohammed Guller, Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis, Apress, 2015
3. Nick Pentreath, Machine Learning with Spark - Tackle Big Data with Powerful Spark Machine Learning Algorithms, Packt Publishing, 2015
4. Raul Estrada and Isaac Ruiz, Big Data SMACK: A Guide to Apache Spark, Mesos, Akka, Cassandra, and Kafka, Apress, 2016
5. Wes McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012
6. Michael Heydt , Mastering Pandas for Finance, Packt Publishing, 2015
7. Michael Heydt, Learning Pandas - Python Data Discovery and Analysis Made Easy, Packt Publishing, 2015
8. Yves Hilpisch, Python for Finance: Analyze Big Financial Data, O'Reilly Media, 2014
9. James Ma Weiming, Mastering Python for Finance, Packt Publishing, 2015
10. Fabio Nelli, Python Data Analytics: Data Analysis and Science using PANDAs, matplotlib and the Python Programming Language, Apress, 2015

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 Analytics
 - Social Computing
 - Big Data mining
 - Business Intelligence
 - FinTech
- 3-4 人為一組
 - 分組名單於 2017/02/22 (三) 課程下課時繳交
 - 由班代統一收集協調分組名單

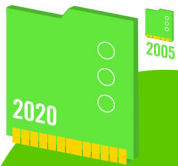
Social Computing
and
Big Data Analytics
(社群運算
與
大數據分析)

Big Data 4 V

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005



Volume SCALE OF DATA



It's estimated that **2.5 QUINTILLION BYTES**
[2.3 TRILLION GIGABYTES]
of data are created each day



Most companies in the U.S. have at least **100 TERABYTES**
[100,000 GIGABYTES]
of data stored



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]

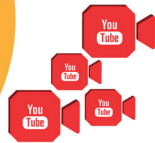


30 BILLION PIECES OF CONTENT are shared on Facebook every month



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



Variety DIFFERENT FORMS OF DATA

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



Velocity ANALYSIS OF STREAMING DATA



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

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



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



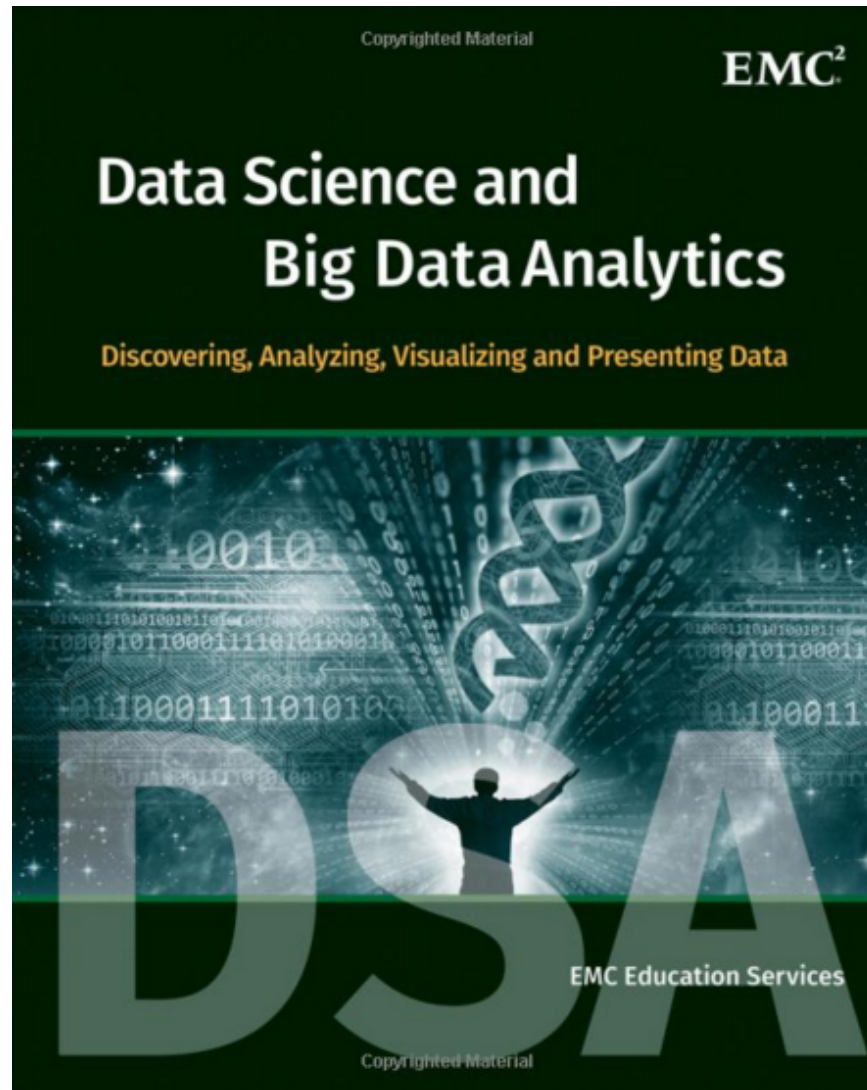
27% OF RESPONDENTS

Veracity UNCERTAINTY OF DATA

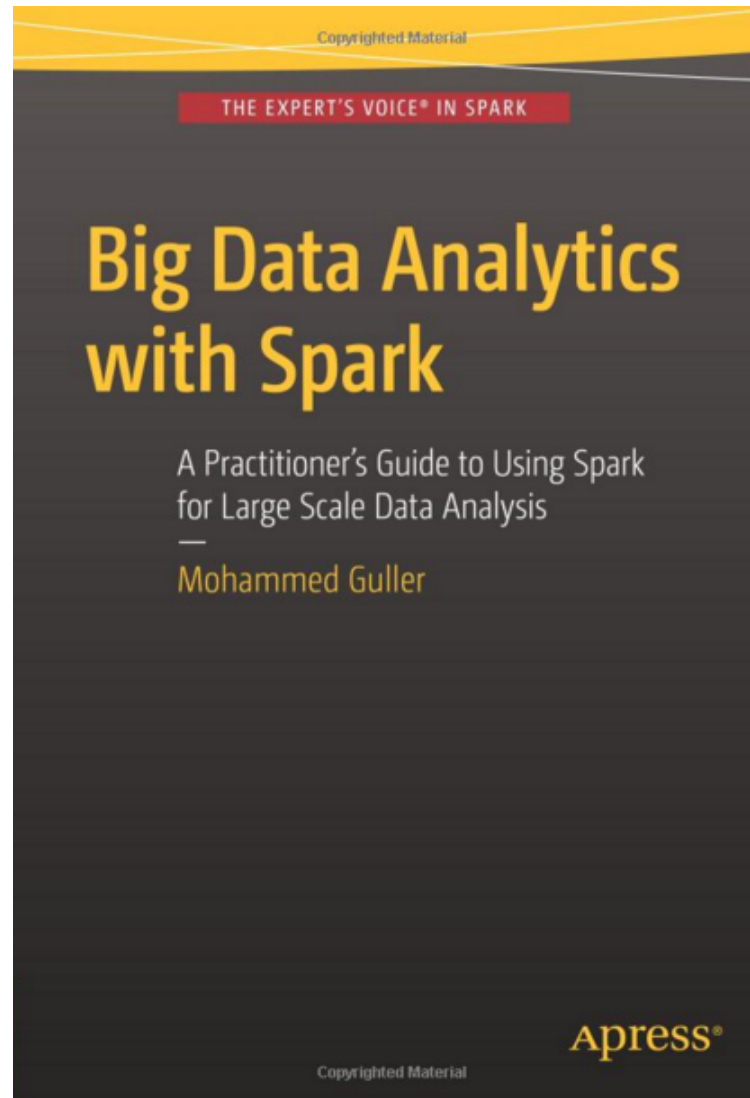
in one survey were unsure of how much of their data was inaccurate

value

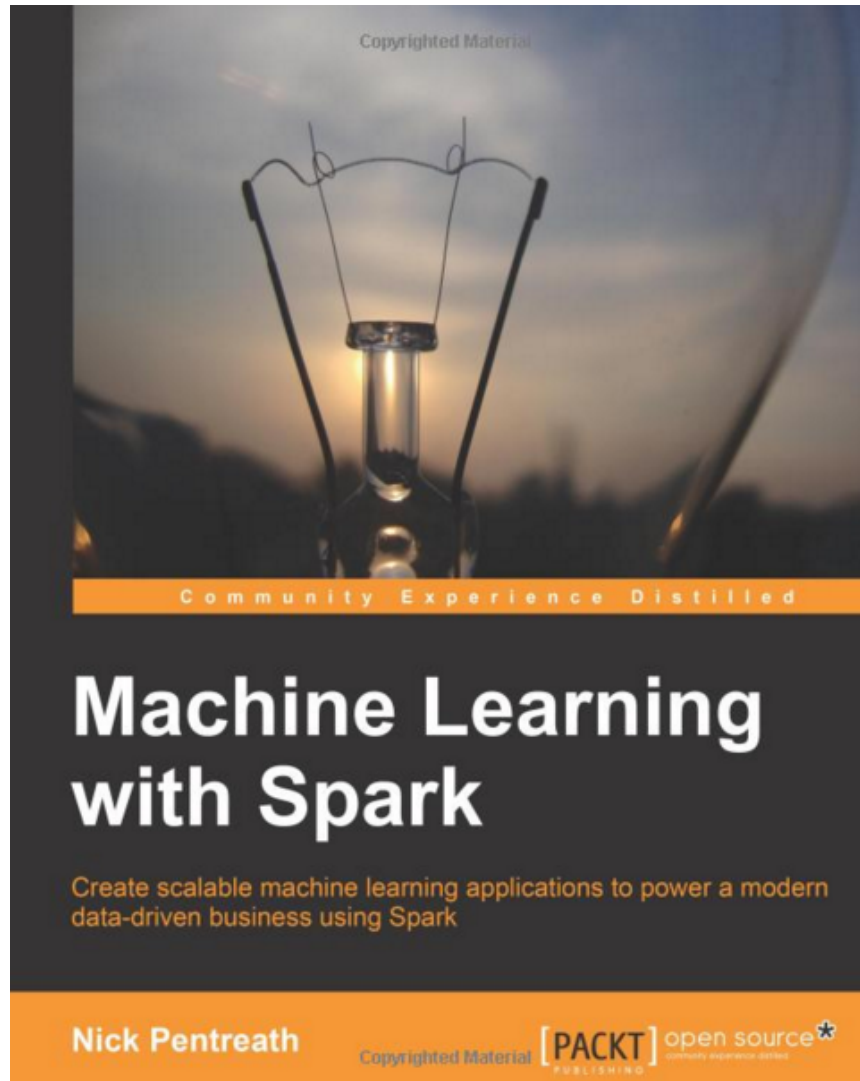
**EMC Education Services,
Data Science and Big Data Analytics:
Discovering, Analyzing, Visualizing and Presenting Data,
Wiley, 2015**



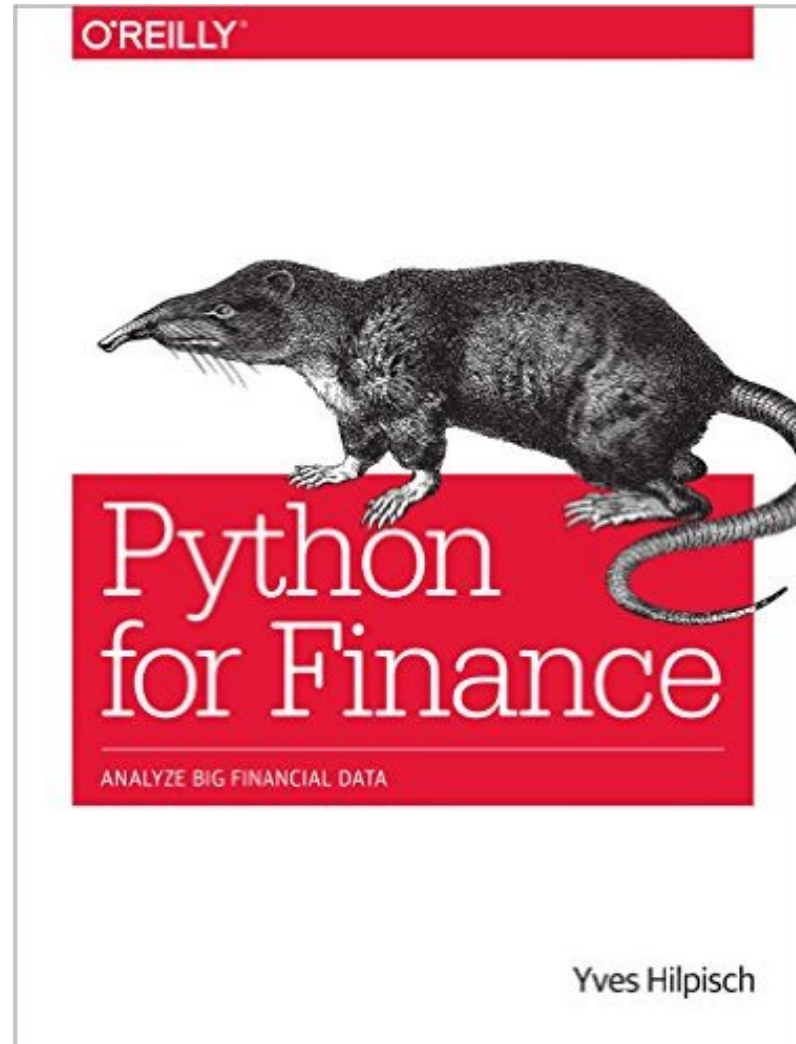
Mohammed Guller,
Big Data Analytics with Spark:
A Practitioner's Guide to Using Spark for Large Scale Data Analysis,
Apress, 2015



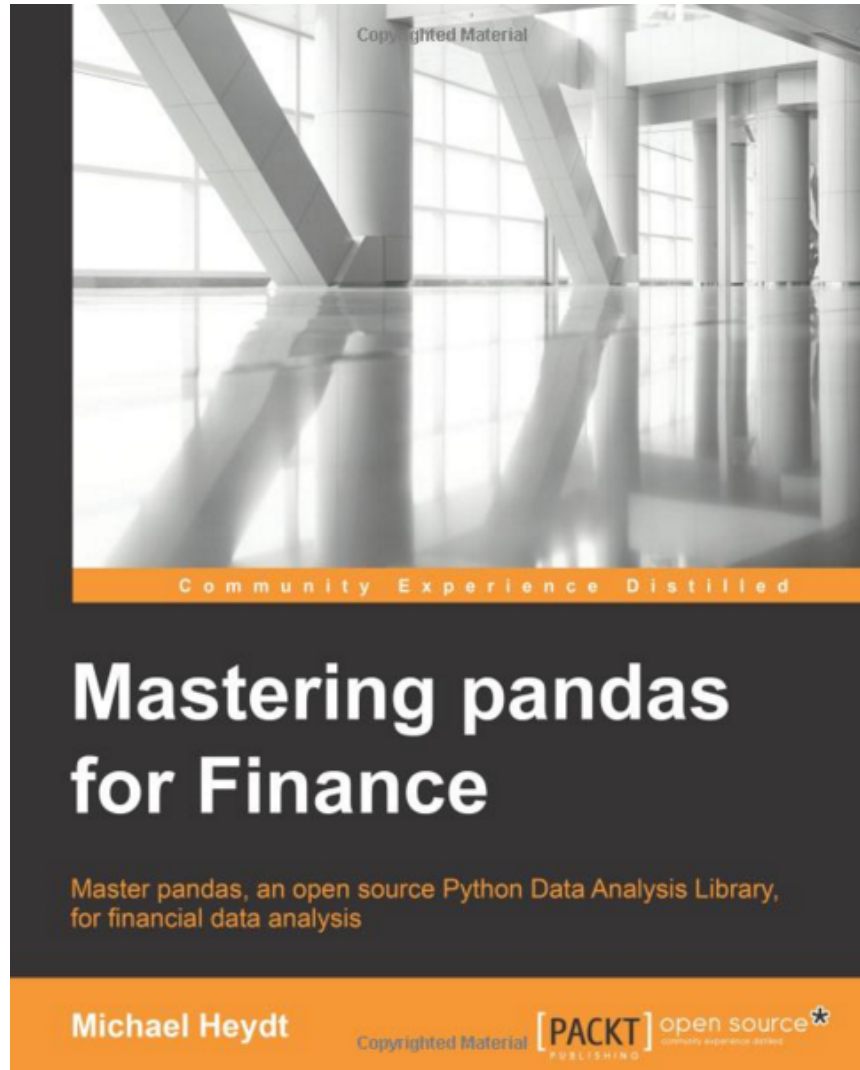
Nick Pentreath,
Machine Learning with Spark –
Tackle Big Data with Powerful Spark Machine Learning Algorithms,
Packt Publishing, 2015



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



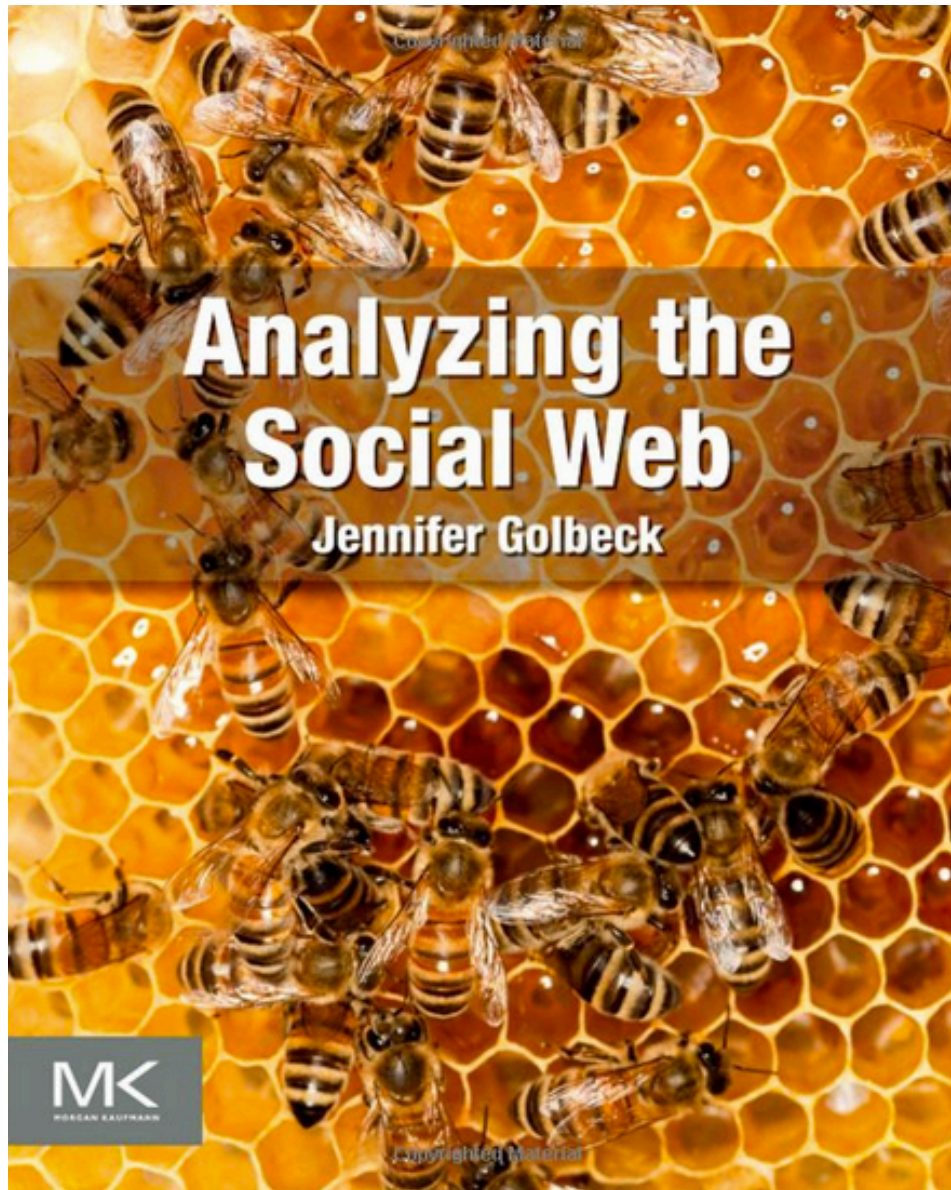
Michael Heydt , Mastering Pandas for Finance, Packt Publishing, 2015



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*



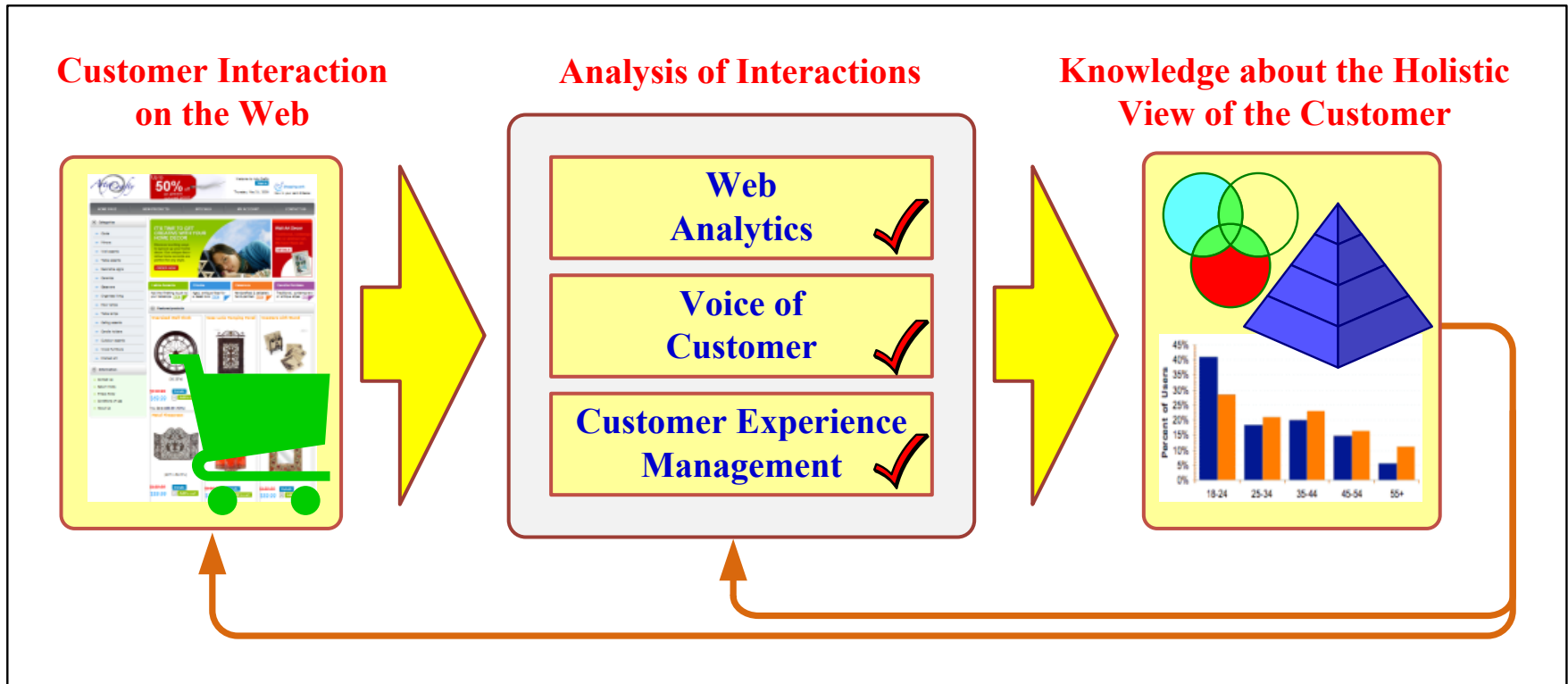
Mining the
Social Web

O'REILLY®

Matthew A. Russell

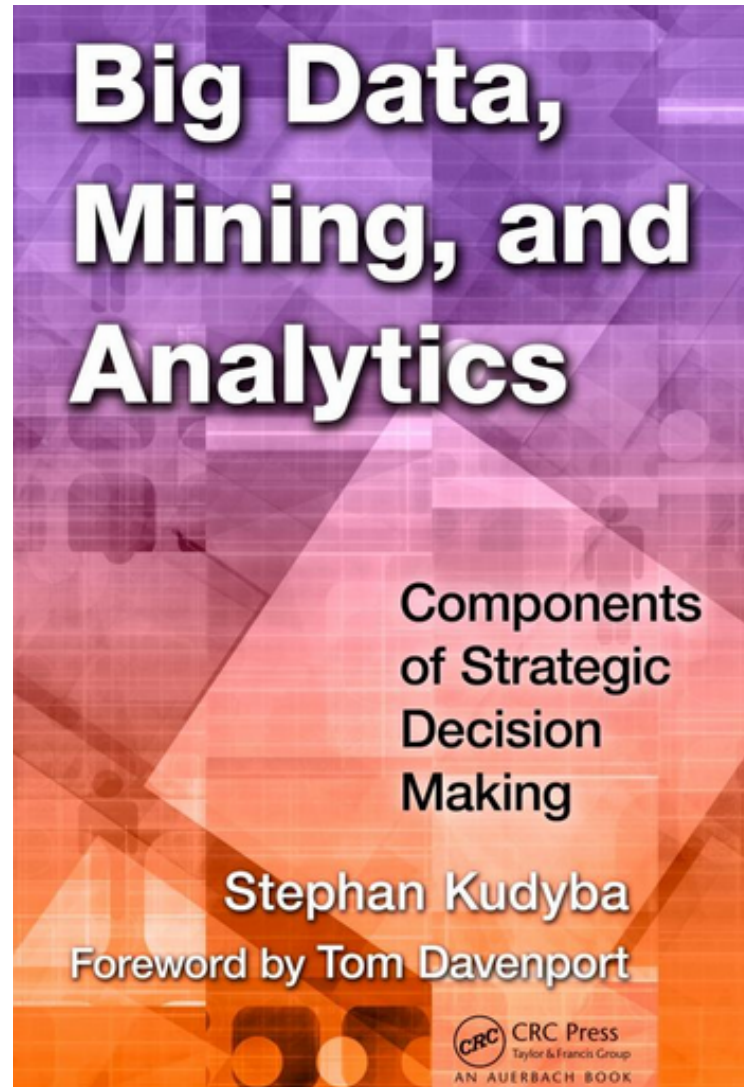
Web Mining Success Stories

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

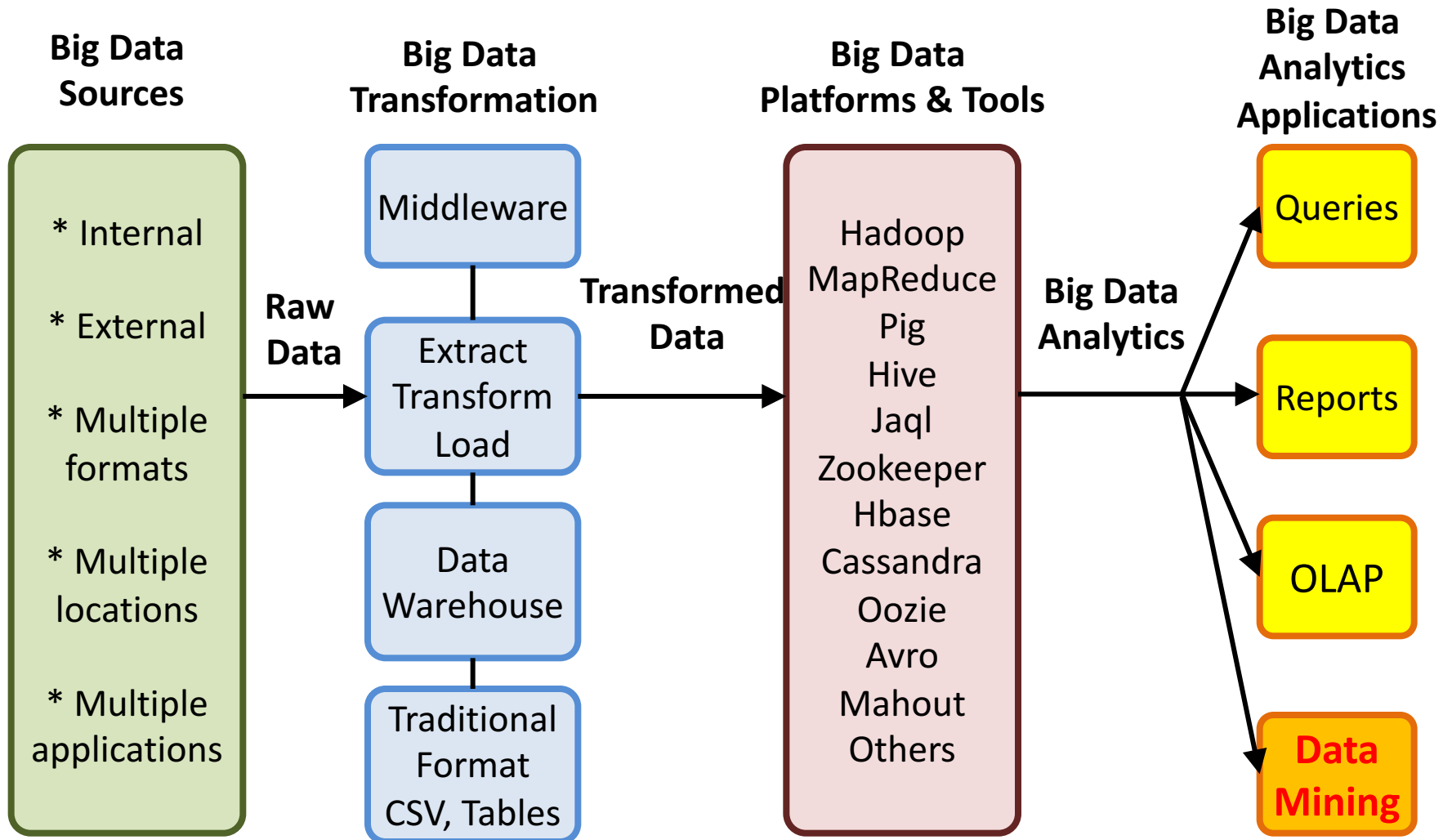


Big Data
Analytics
and
Data Mining

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



Architecture of Big Data Analytics



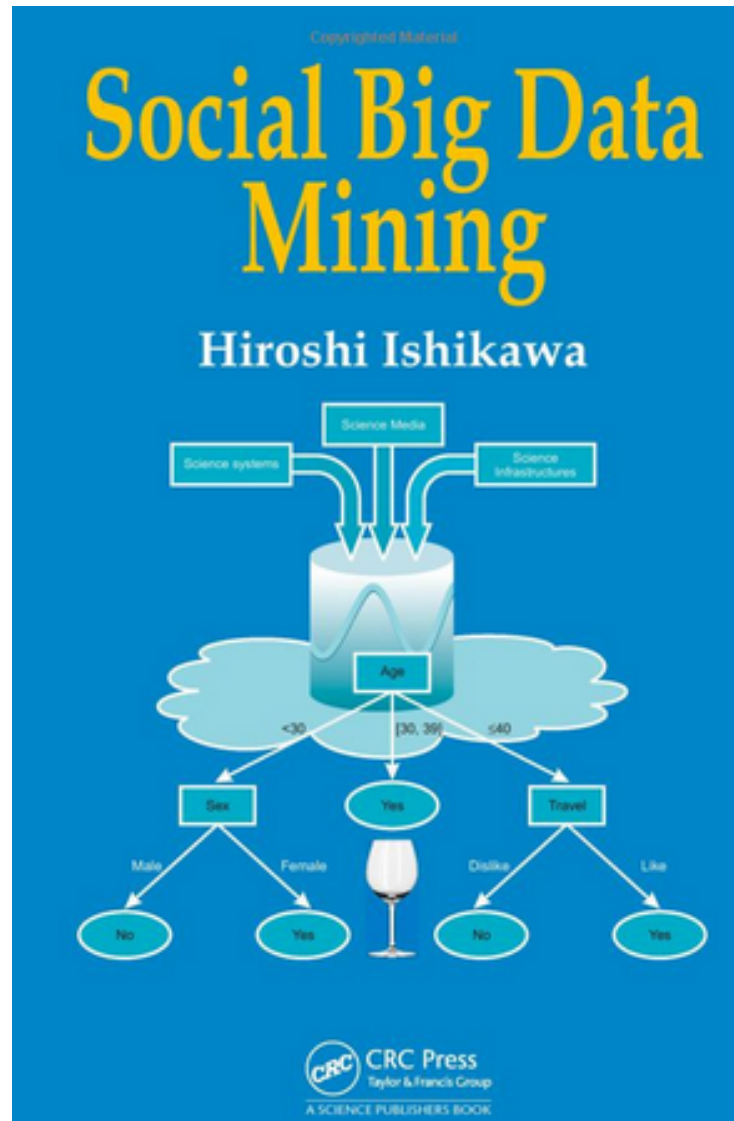
Architecture of Big Data Analytics



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

Social Big Data Mining

(Hiroshi Ishikawa, 2015)

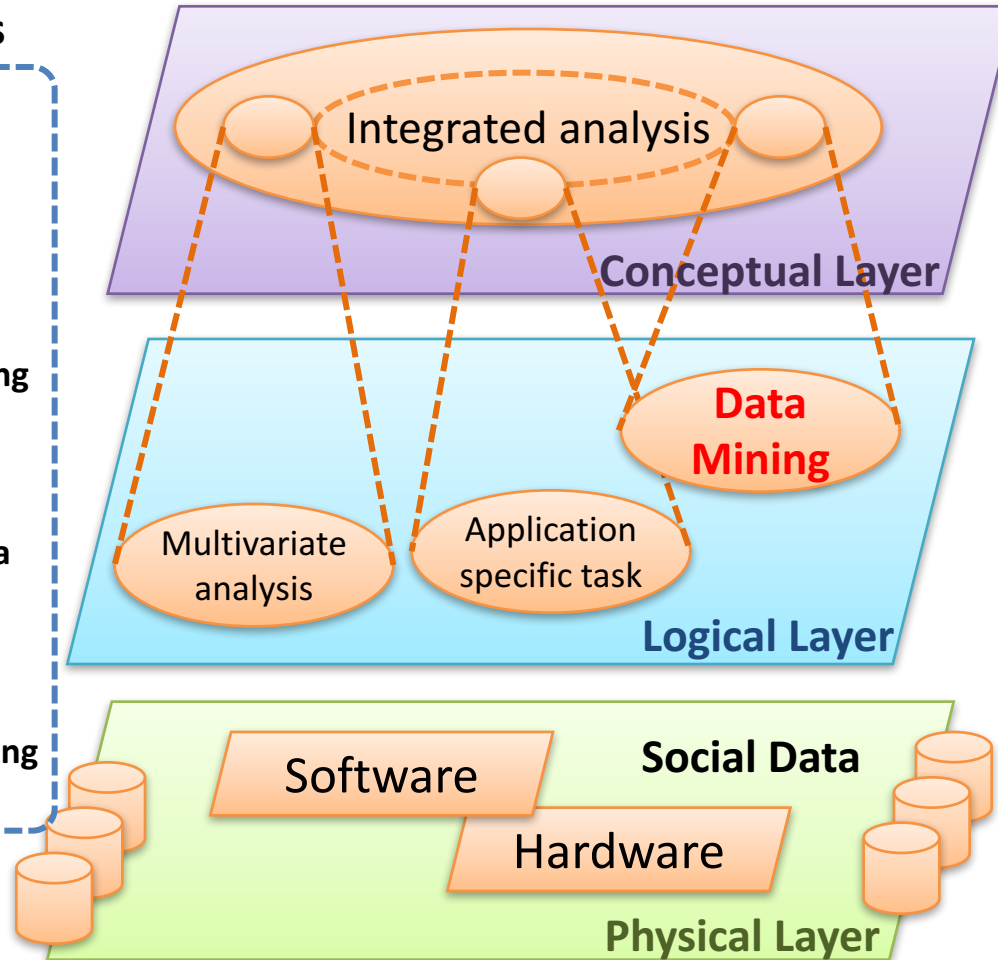


Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Enabling Technologies

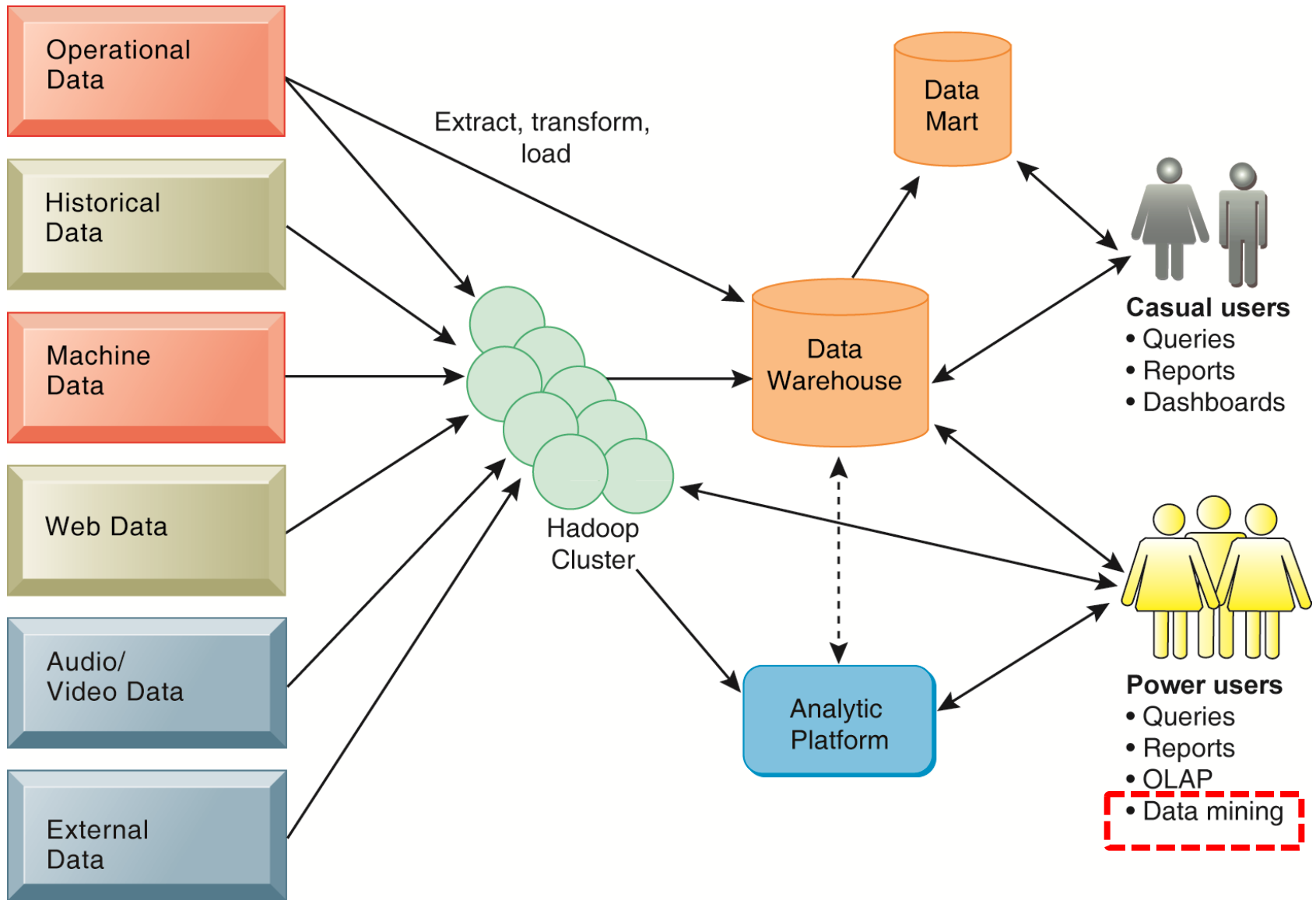
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distributed processing



Analysts

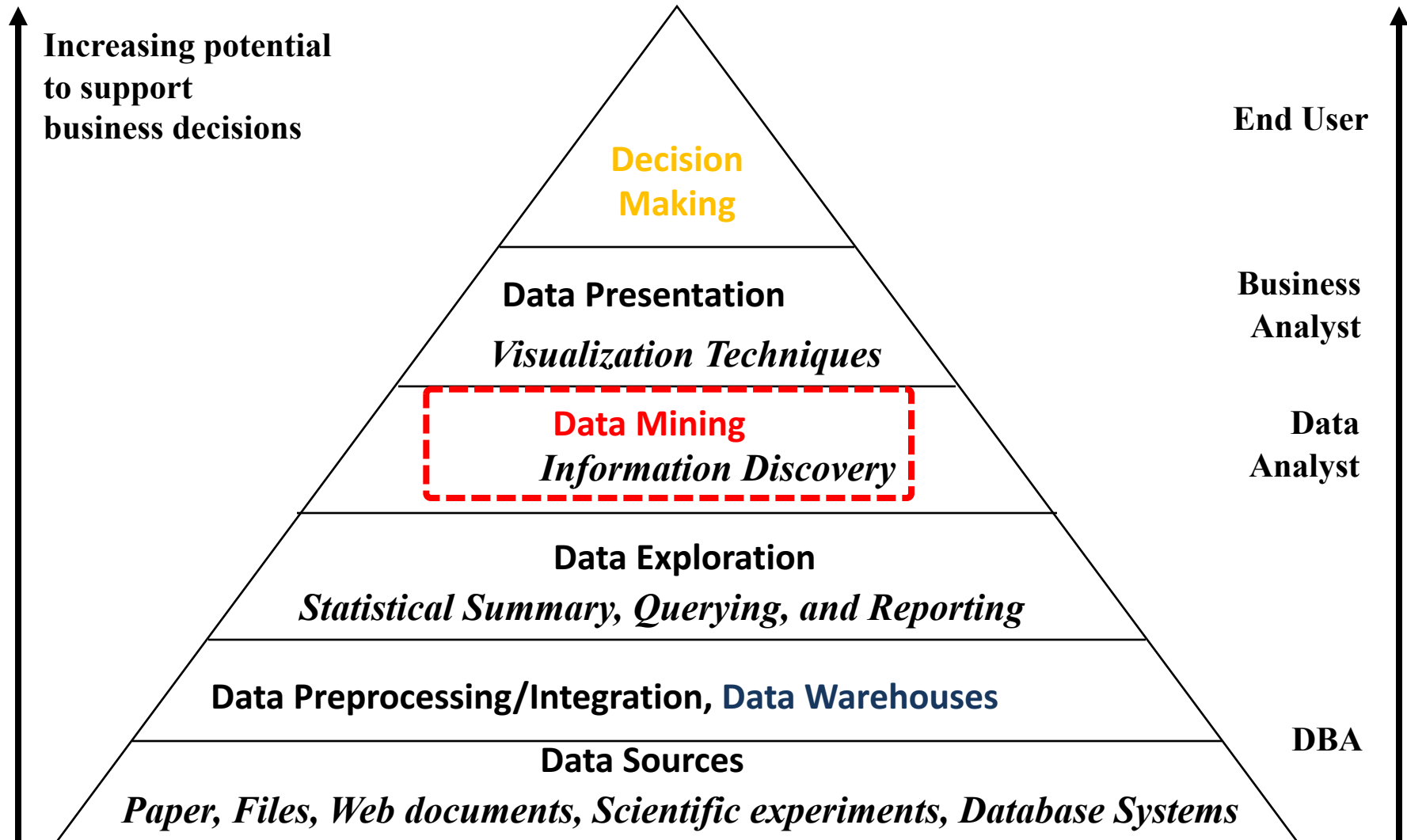
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Business Intelligence (BI) Infrastructure

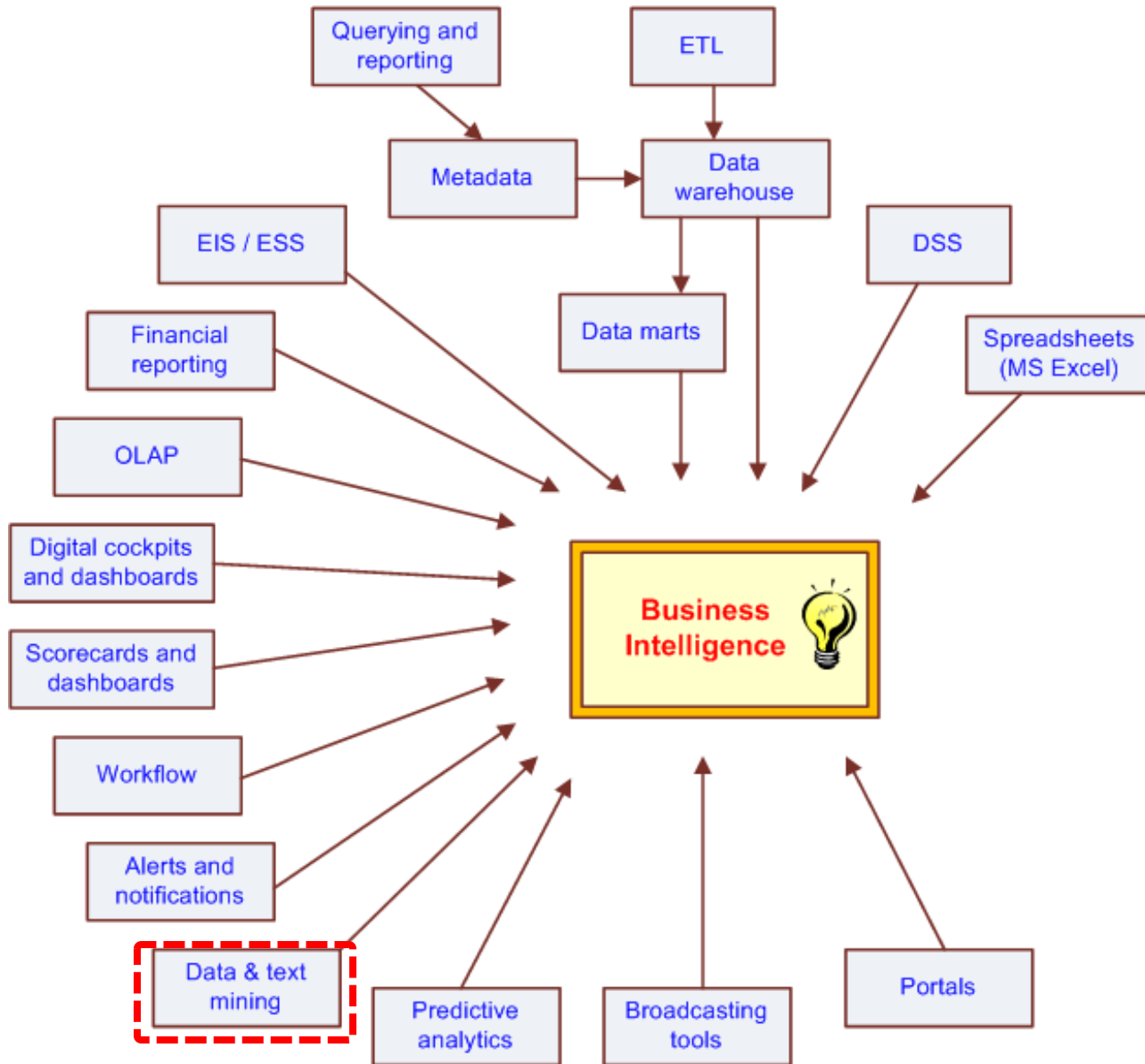


Data Warehouse

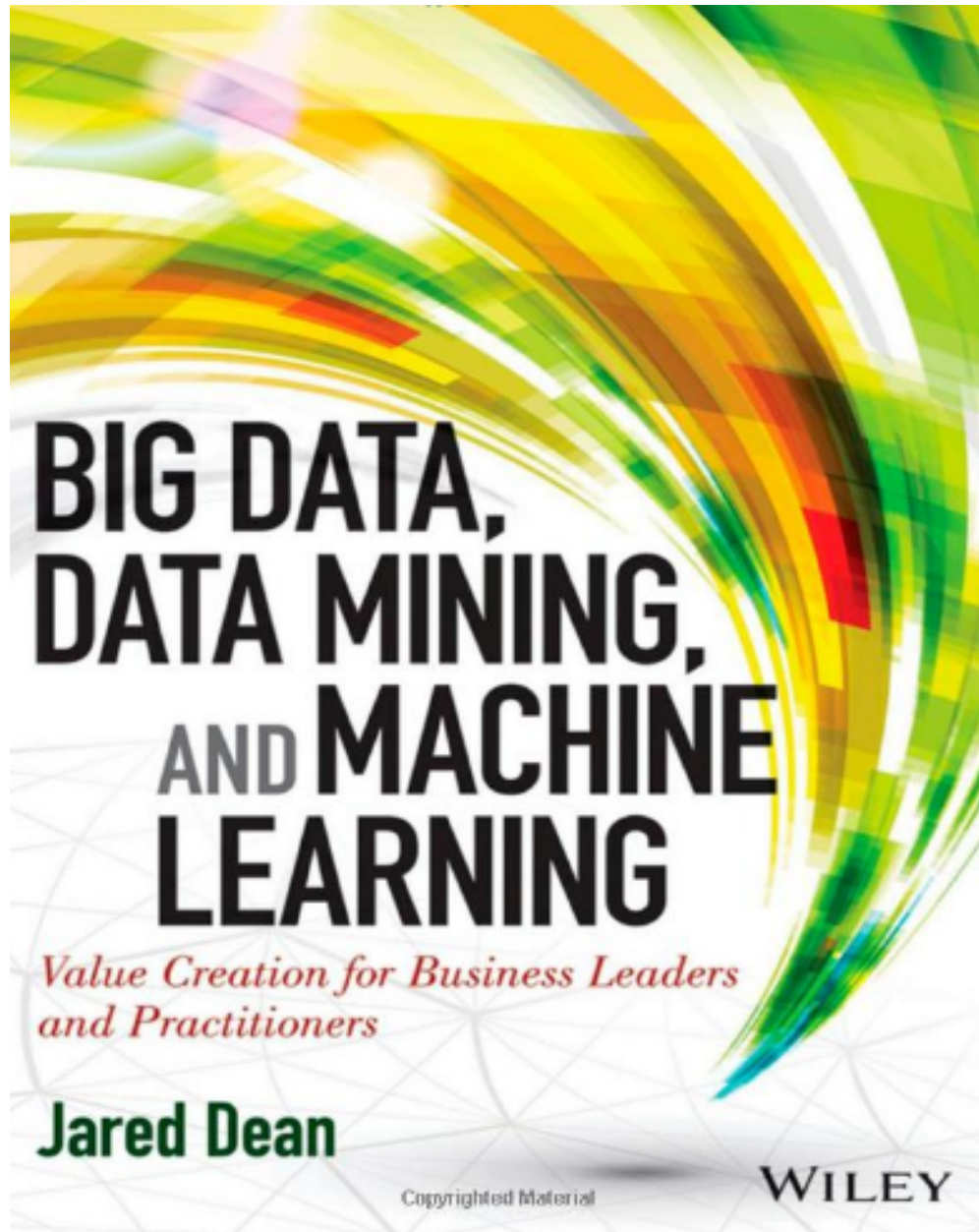
Data Mining and Business Intelligence



The Evolution of BI Capabilities



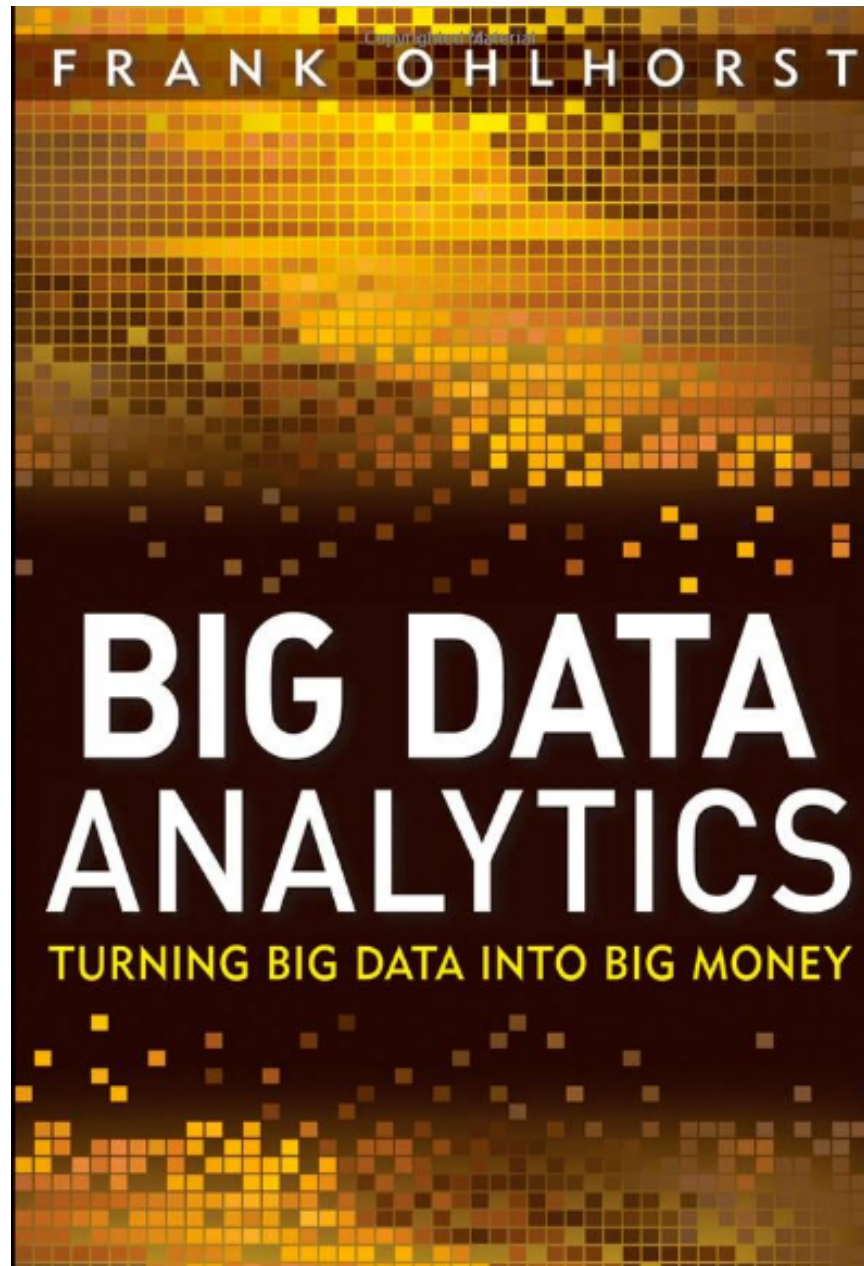
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems



Deep Learning

Intelligence from Big Data









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

Tending
Real-time
Prediction

DECIDE

Simulation
Optimization



BIG DATA – Batch



BIG DATA – Real Time



Complex by nature



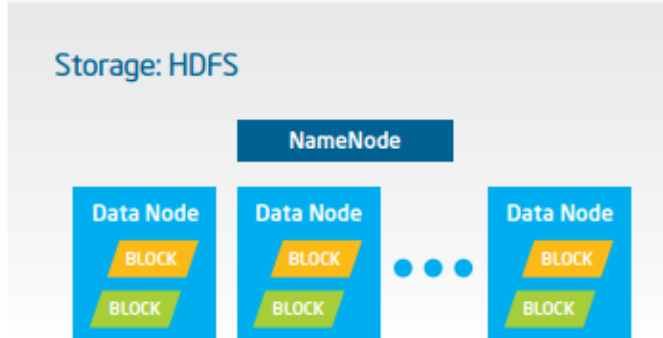
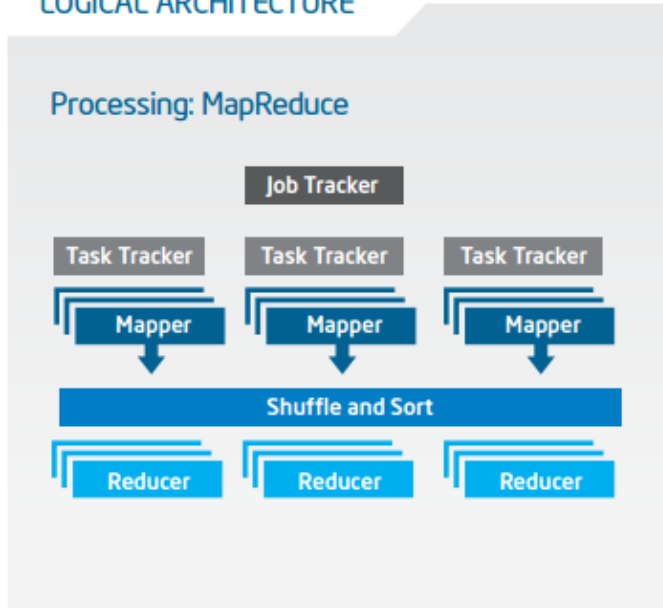
DATA

Complex by structure

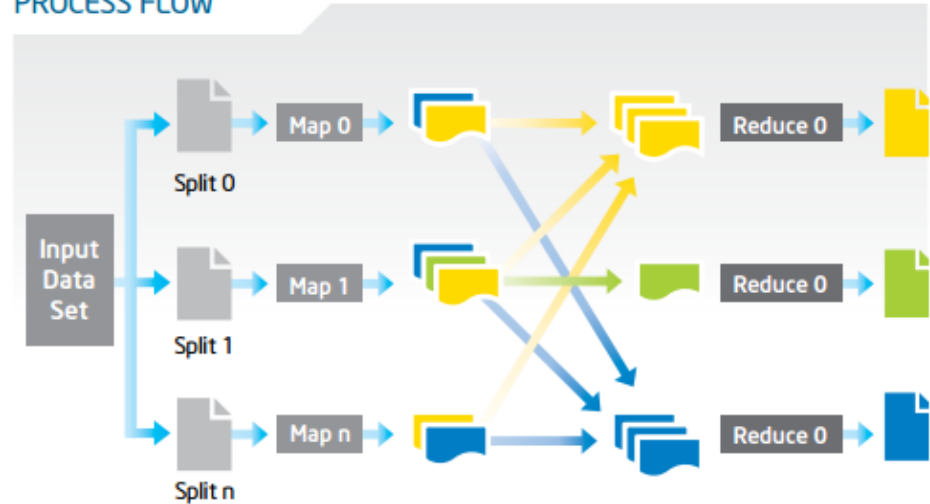


Big Data with Hadoop Architecture

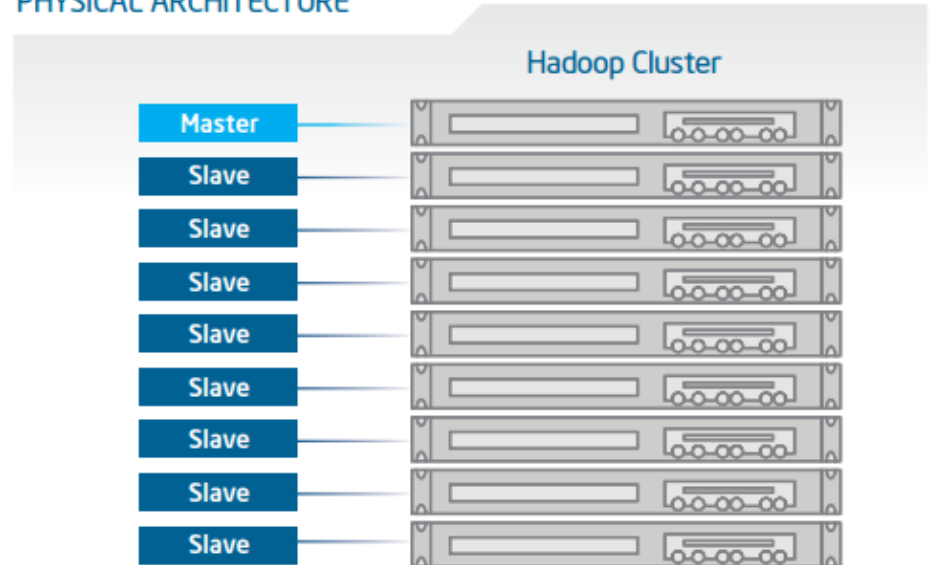
LOGICAL ARCHITECTURE



PROCESS FLOW



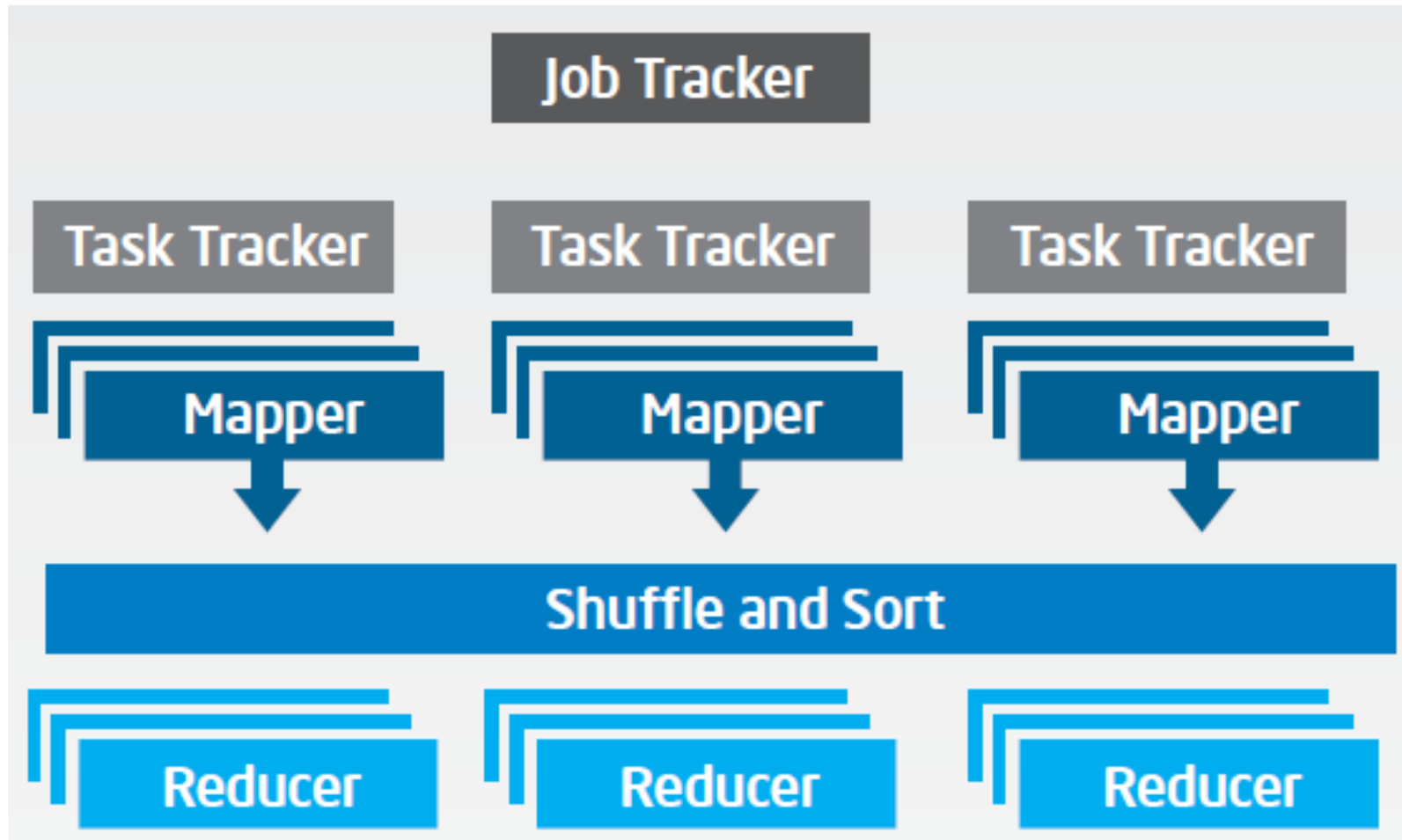
PHYSICAL ARCHITECTURE



Big Data with Hadoop Architecture

Logical Architecture

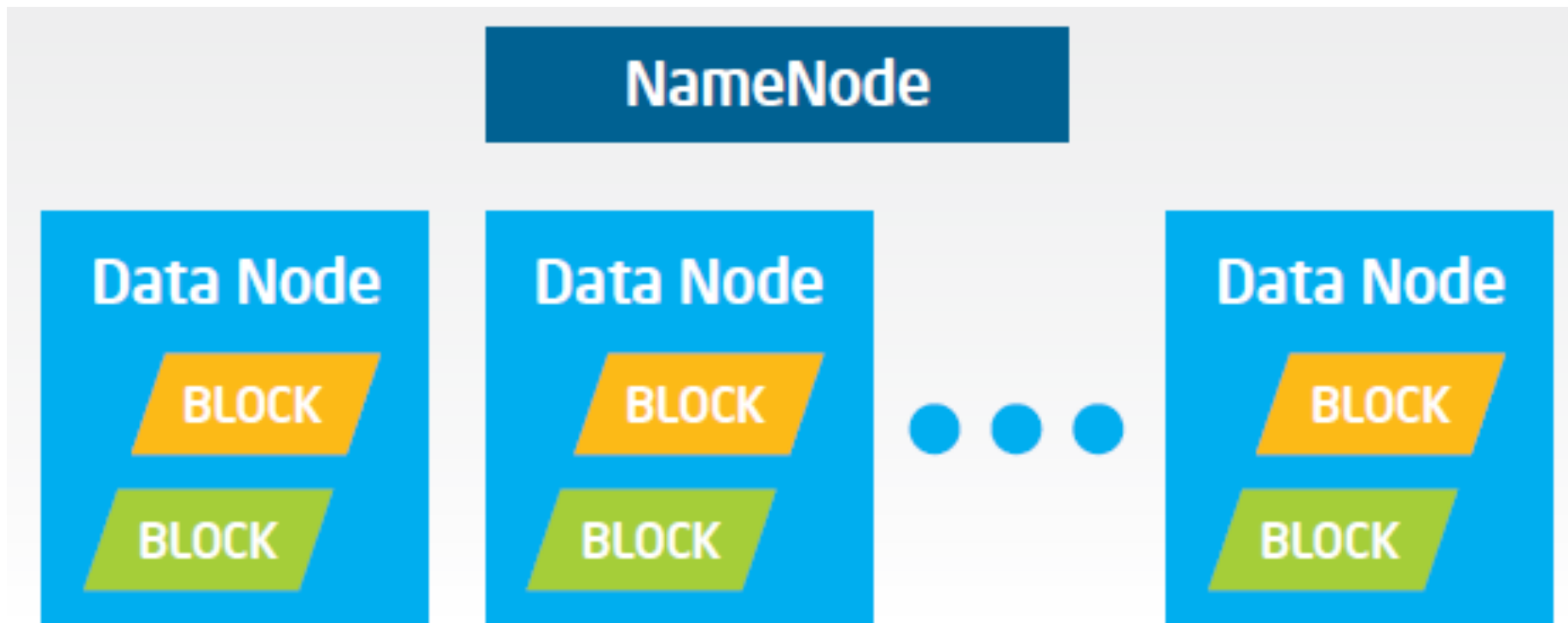
Processing: MapReduce



Big Data with Hadoop Architecture

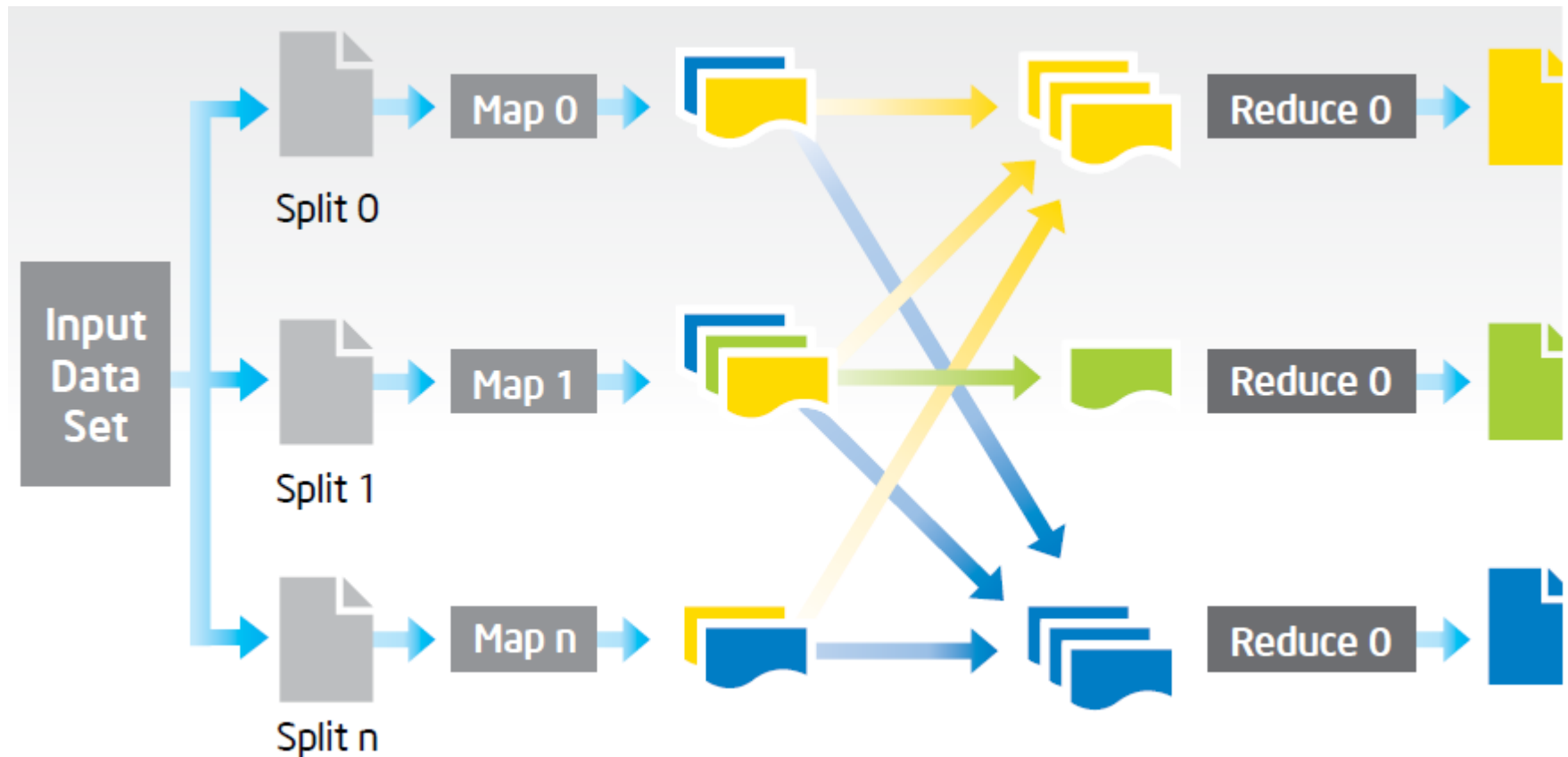
Logical Architecture

Storage: HDFS



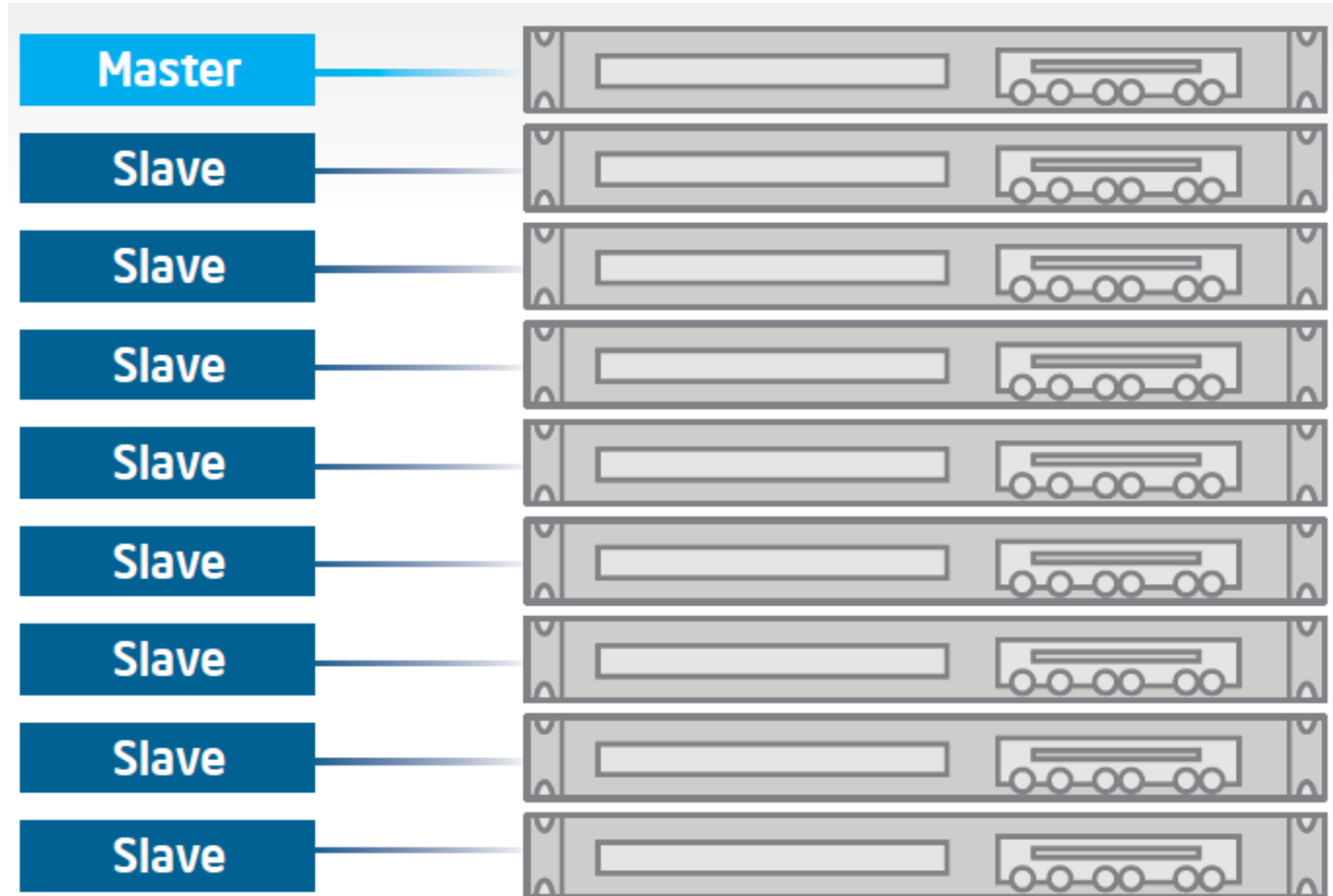
Big Data with Hadoop Architecture

Process Flow

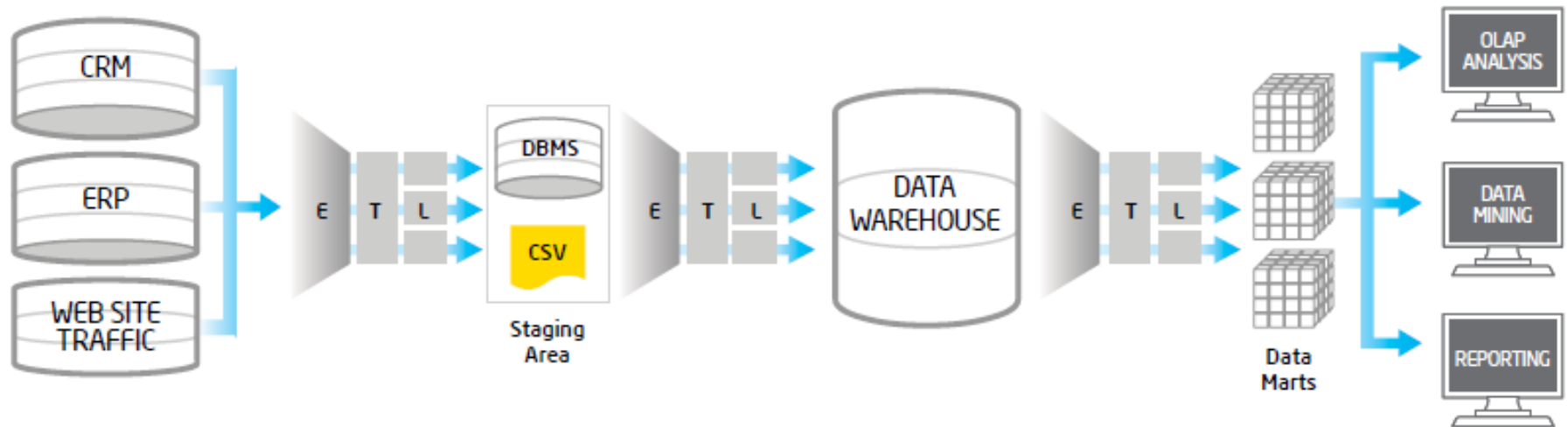


Big Data with Hadoop Architecture

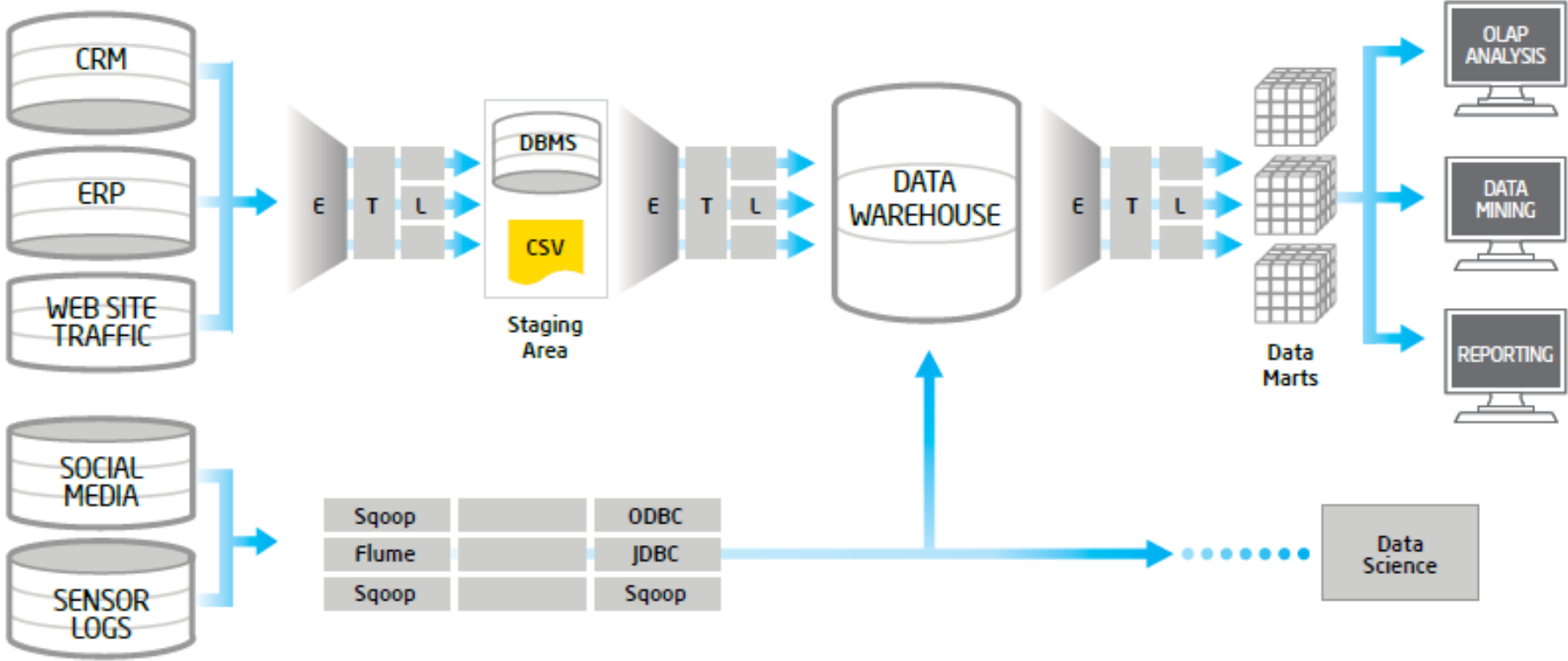
Hadoop Cluster



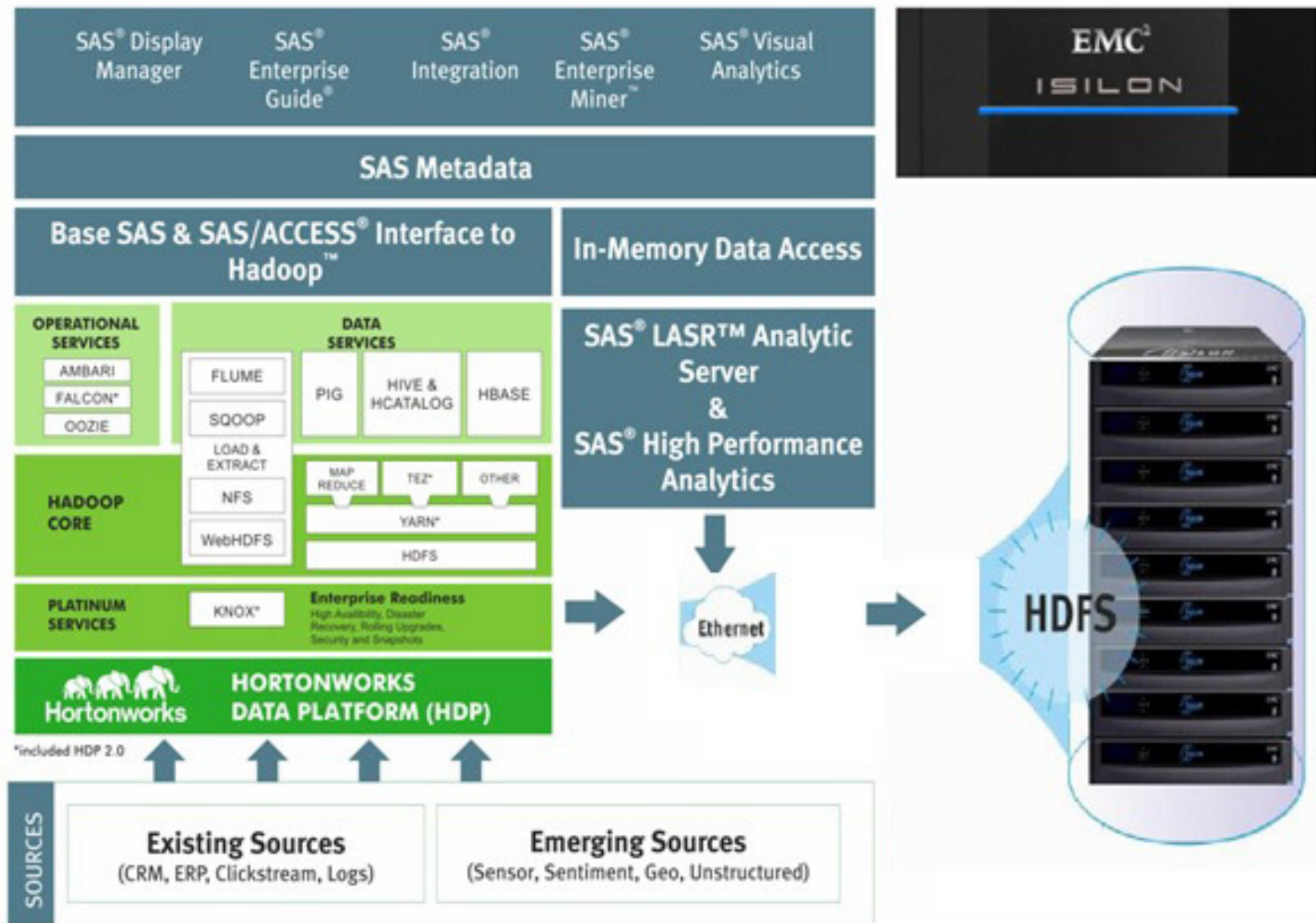
Traditional ETL Architecture



Offload ETL with Hadoop (Big Data Architecture)

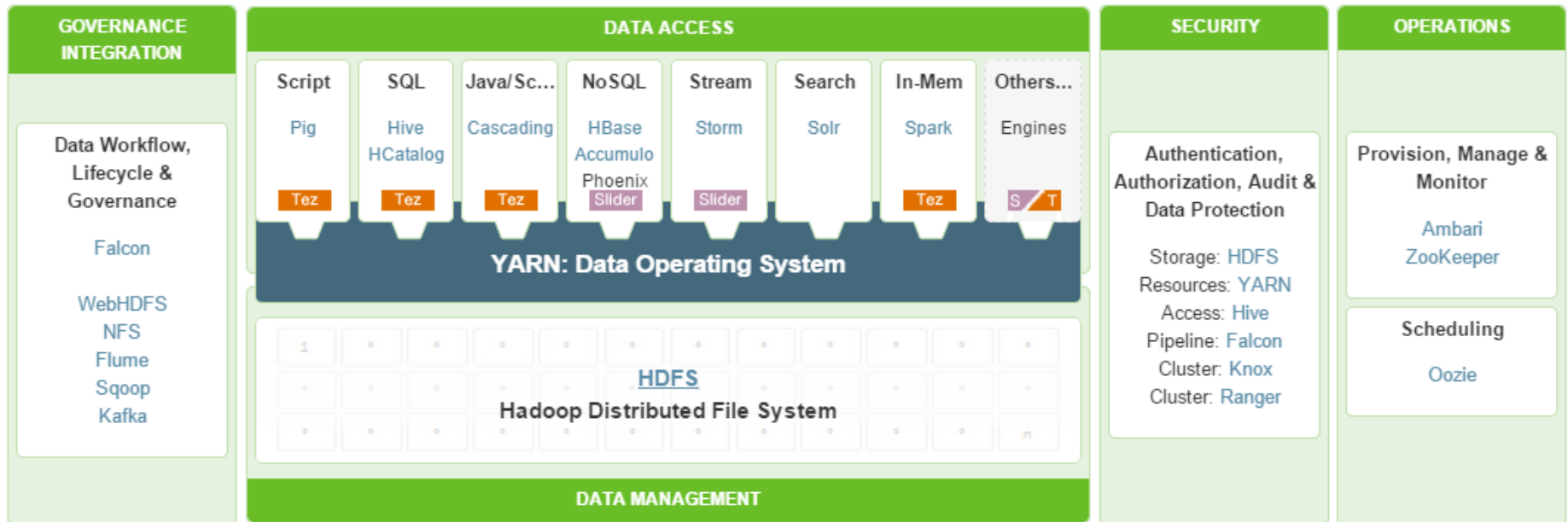


Big Data Solution



HDP

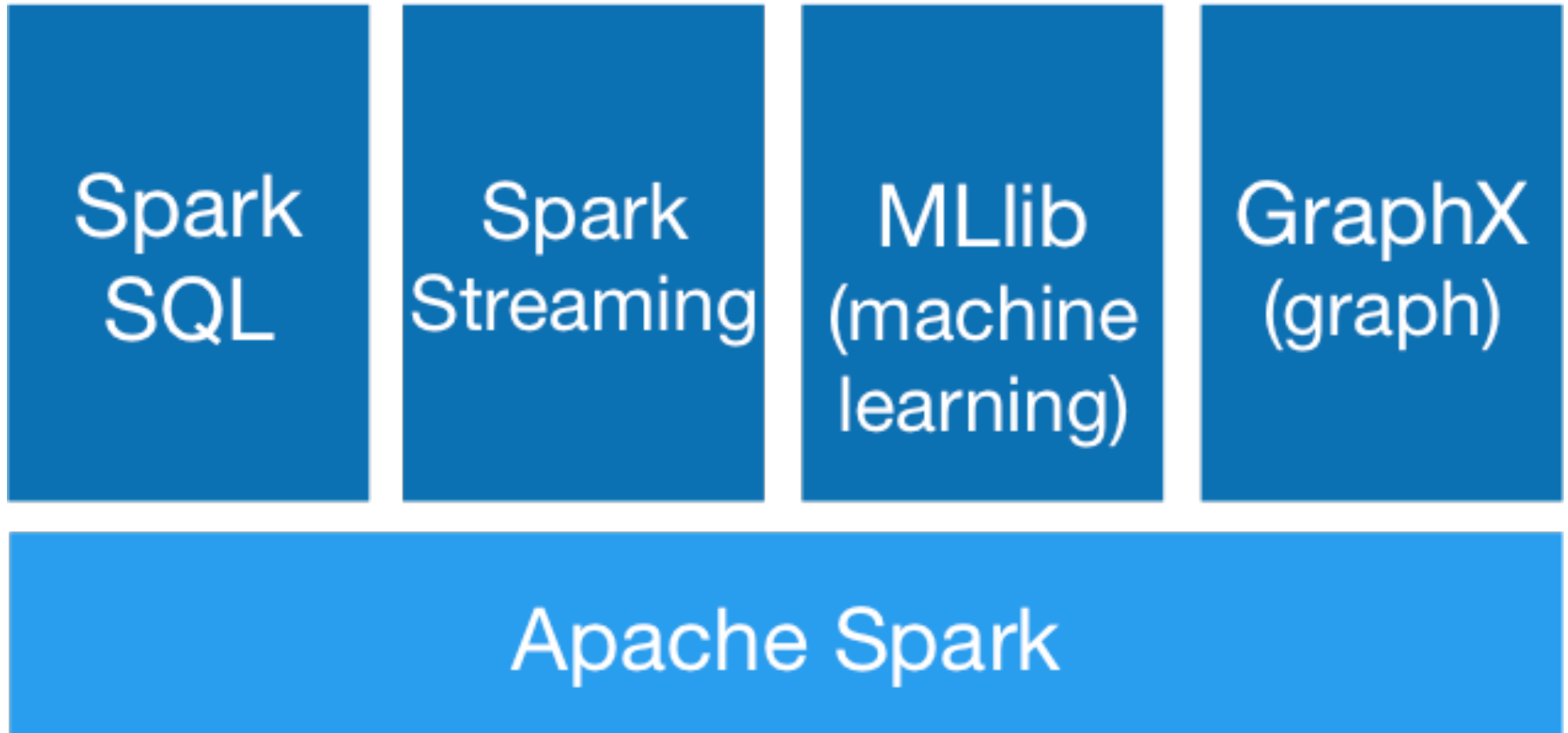
A Complete Enterprise Hadoop Data Platform





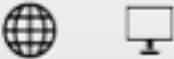







Spark and Hadoop



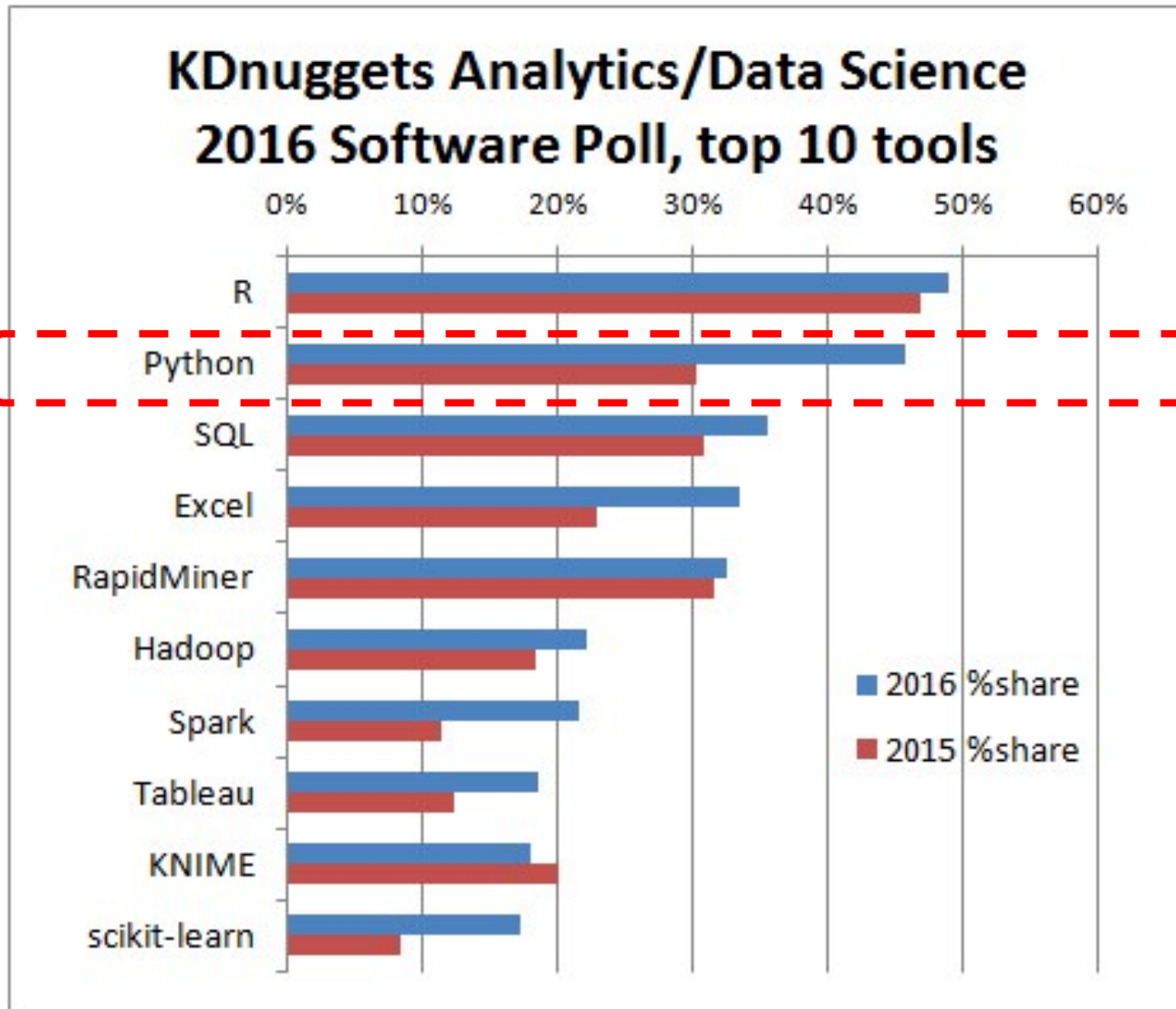
Spark Ecosystem



Python for Big Data Analytics

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

Python: Analytics and Data Science Software



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

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



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)

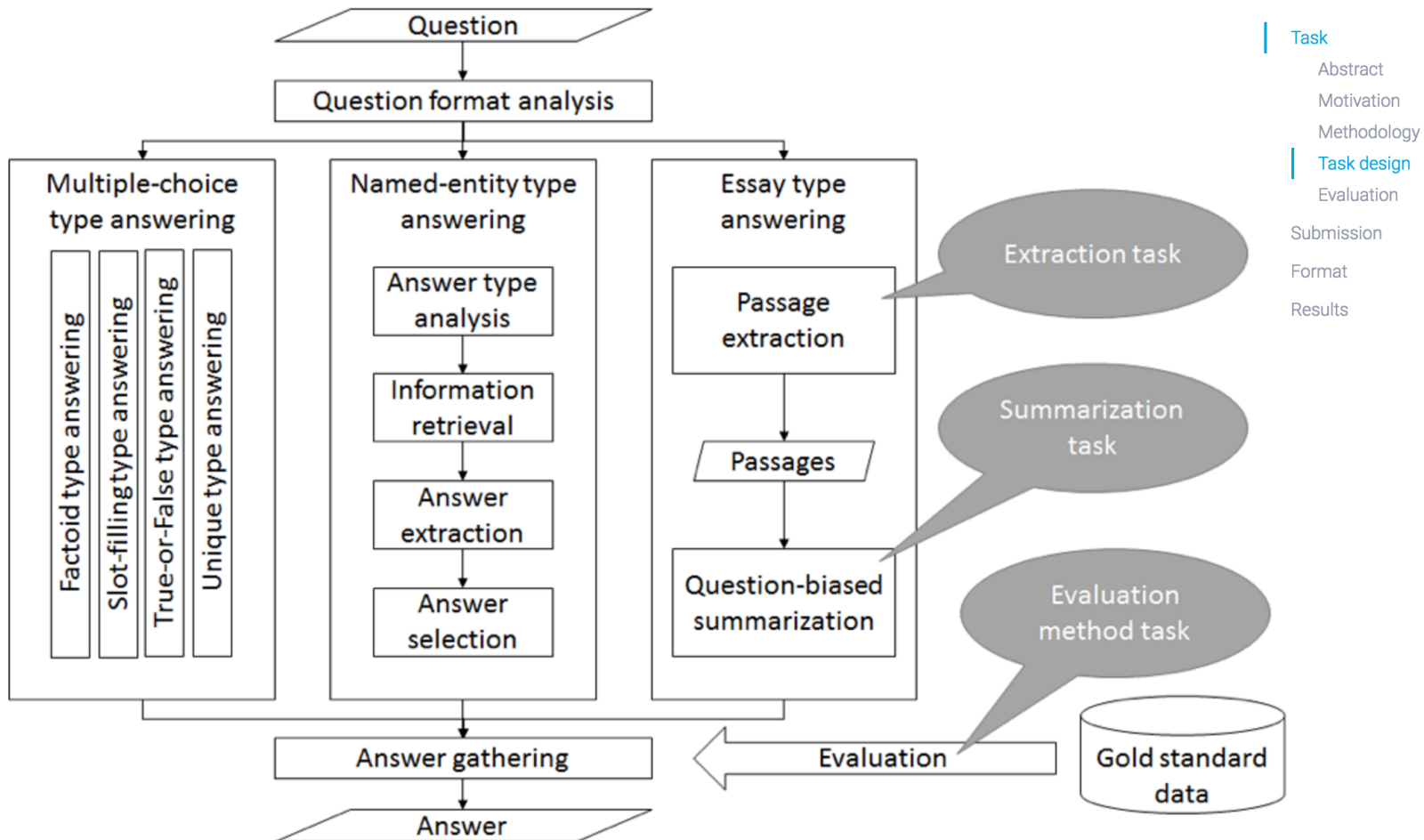
[ECA](#): until Dec 1, 2016 (The registration deadline was over)

[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 **research issues** of **social computing** and **big data analytics**.
- Topics include
 - **Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data**
 - Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
 - Big Data Processing Platforms with SMACK: Spark, Mesos, Akka, Cassandra and Kafka
 - Big Data Analytics with Numpy in Python
 - **Finance Big Data Analytics with Pandas in Python**
 - Text Mining Techniques and Natural Language Processing
 - Social Media Marketing Analytics
 - **Deep Learning with Theano and Keras in Python**
 - **Deep Learning with Google TensorFlow**
 - **Sentiment Analysis on Social Media with Deep Learning**
 - **Social Network Analysis, Measurements, and Tools**

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/>

