

# Social Computing and Big Data Analytics

## 社群運算與大數據分析

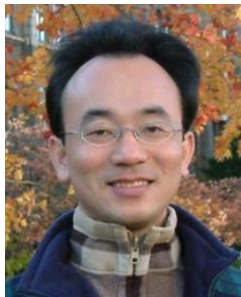
### Course Orientation for Social Computing and Big Data Analytics

(社群運算與大數據分析課程介紹)

1042SCBDA01

MIS MBA (M2226) (8628)

Wed, 8,9, (15:10-17:00) (Q201)



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/>

2016-02-17



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

# 淡江大學104學年度第2學期 課程教學計畫表

Spring 2016 (2016.02 - 2016.06)

- 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：Q201 (傳播館)

# 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         | 2016/02/17 | Course Orientation for Social Computing and Big Data Analytics<br>(社群運算與大數據分析課程介紹)   |
| 2         | 2016/02/24 | Data Science and Big Data Analytics:<br>Discovering, Analyzing, Visualizing and Presenting Data<br>(資料科學與大數據分析：<br>探索、分析、視覺化與呈現資料) |
| 3         | 2016/03/02 | Fundamental Big Data: MapReduce Paradigm,<br>Hadoop and Spark Ecosystem<br>(大數據基礎：MapReduce典範、<br>Hadoop與Spark生態系統)                |

# 課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
4	2016/03/09	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	2016/03/16	Big Data Analytics with Numpy in Python (Python Numpy 大數據分析)
6	2016/03/23	Finance Big Data Analytics with Pandas in Python (Python Pandas 財務大數據分析)
7	2016/03/30	Text Mining Techniques and Natural Language Processing (文字探勘分析技術與自然語言處理)
8	2016/04/06	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	2016/04/13	Social Media Marketing Analytics (社群媒體行銷分析)
10	2016/04/20	期中報告 (Midterm Project Report)
11	2016/04/27	Deep Learning with Theano and Keras in Python (Python Theano 和 Keras 深度學習)
12	2016/05/04	Deep Learning with Google TensorFlow (Google TensorFlow 深度學習)
13	2016/05/11	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	2016/05/18	Social Network Analysis (社會網絡分析)
15	2016/05/25	Measurements of Social Network (社會網絡量測)
16	2016/06/01	Tools of Social Network Analysis (社會網絡分析工具)
17	2016/06/08	Final Project Presentation I (期末報告 I)
18	2016/06/15	Final Project Presentation II (期末報告 II)

**2016/02/24**

**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)**

**2016/03/02**

**Fundamental Big Data:**

**MapReduce Paradigm,**

**Hadoop and Spark Ecosystem**

**(大數據基礎：**

**MapReduce典範、**

**Hadoop與Spark生態系統)**

**2016/03/09**

**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)**

**2016/03/23**

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

**(Python Pandas**

**財務大數據分析)**

2016/04/27

**Deep Learning with  
Theano and Keras  
in Python**

**(Python Theano  
和 Keras 深度學習)**

2016/05/04

**Deep Learning**

**with**

**Google TensorFlow**

**(Google TensorFlow**

**深度學習)**



# 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
- 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. Wes McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012
5. Michael Heydt , Mastering Pandas for Finance, Packt Publishing, 2015
6. Michael Heydt, Learning Pandas - Python Data Discovery and Analysis Made Easy, Packt Publishing, 2015
7. Yves Hilpisch, Derivatives Analytics with Python: Data Analysis, Models, Simulation, Calibration and Hedging, Wiley, 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

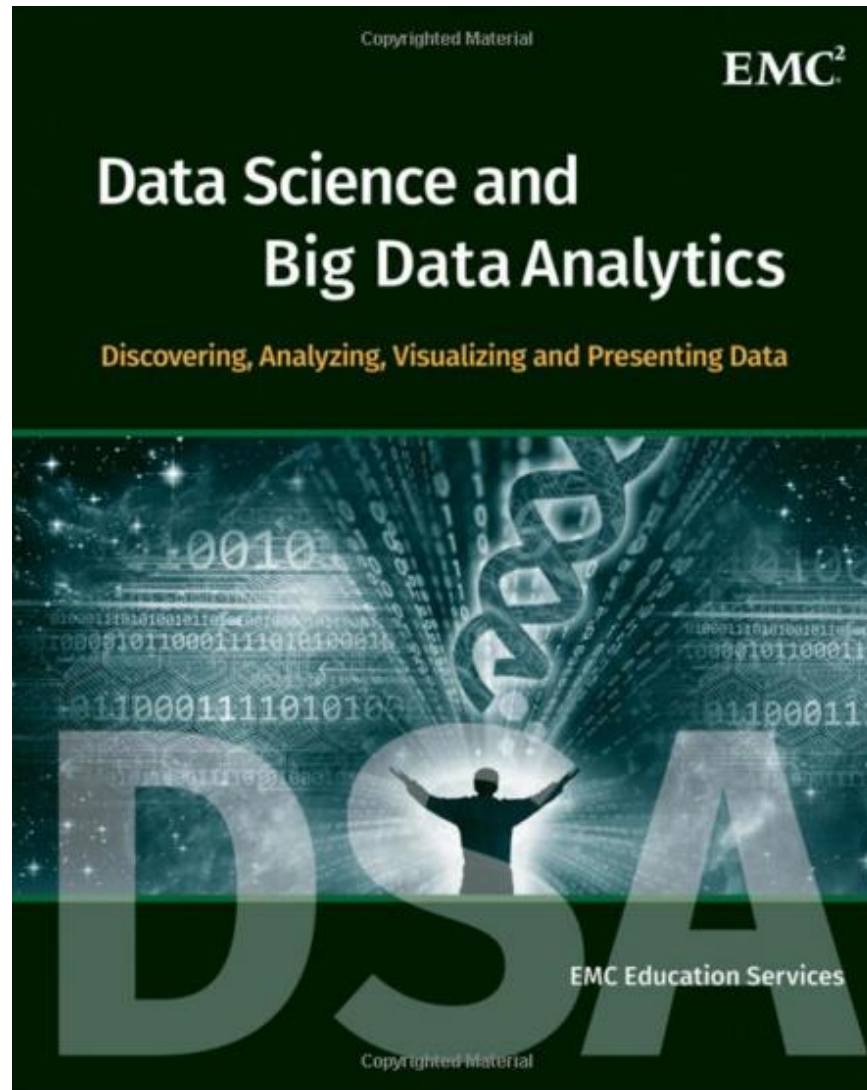
- 作業篇數
  - 3篇
- 學期成績計算方式
  - 期中評量：30 %
  - 期末評量：30 %
  - 其他（課堂參與及報告討論表現）：40 %

# Team Term Project

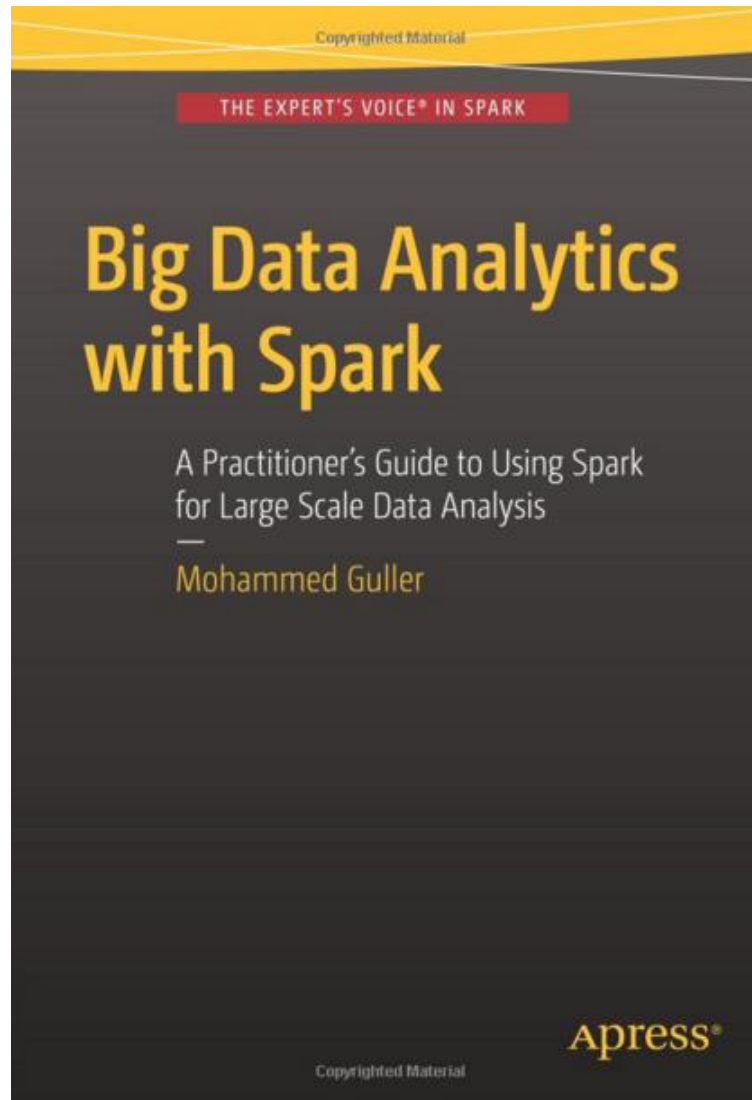
- Term Project Topics
  - Big Data Analytics
  - Social Computing
  - Big Data mining
  - Web and Text mining
  - Business Intelligence
- 3-4 人為一組
  - 分組名單於 2016/02/24 (三) 課程下課時繳交
  - 由班代統一收集協調分組名單

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

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

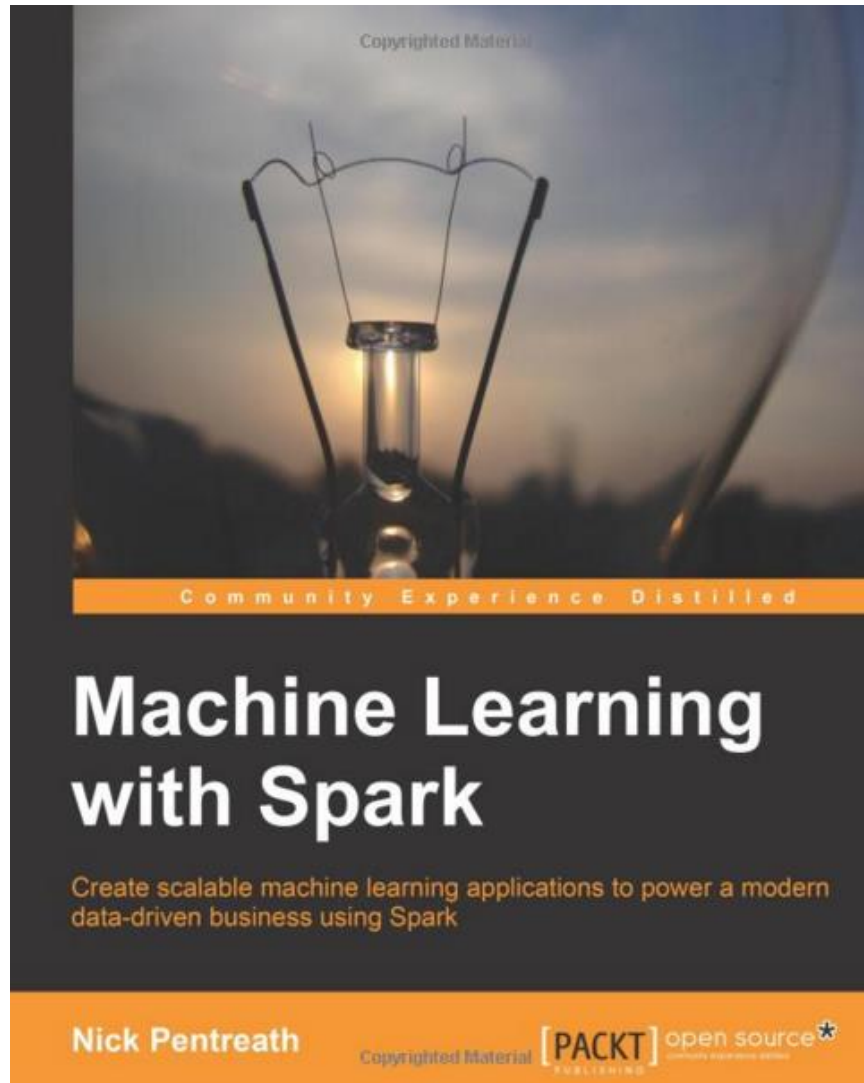


**Mohammed Guller,**  
**Big Data Analytics with Spark:**  
**A Practitioner's Guide to Using Spark for Large Scale Data Analysis,**  
**Apress, 2015**

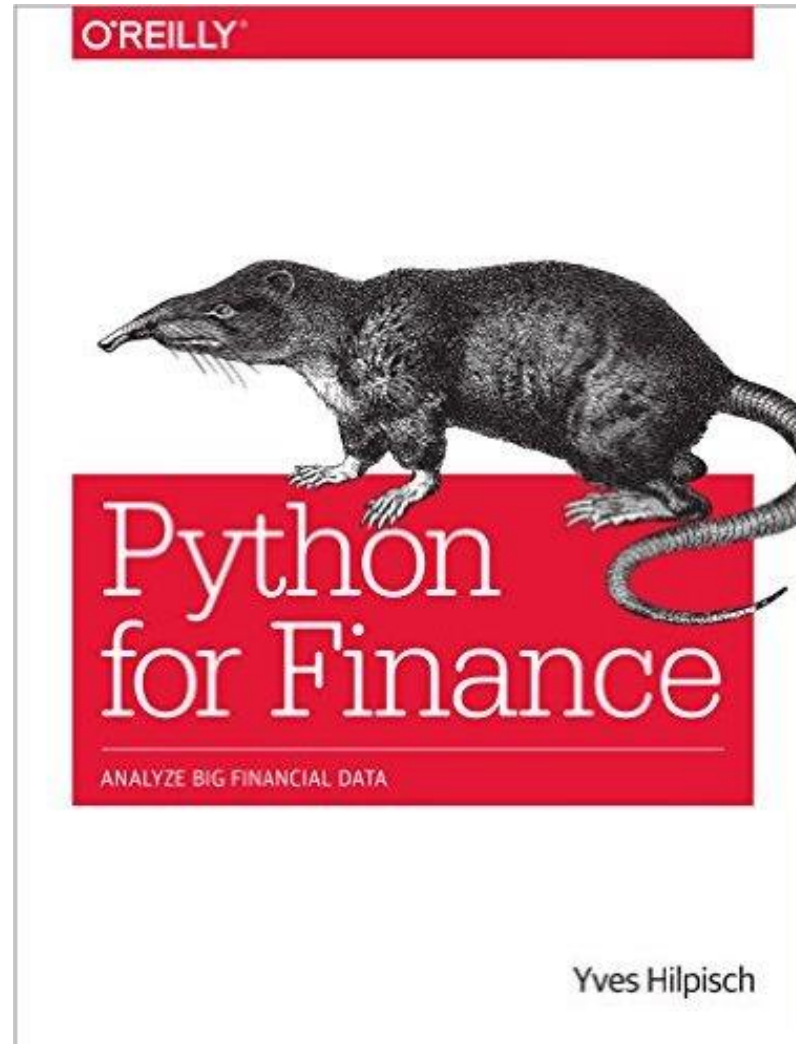




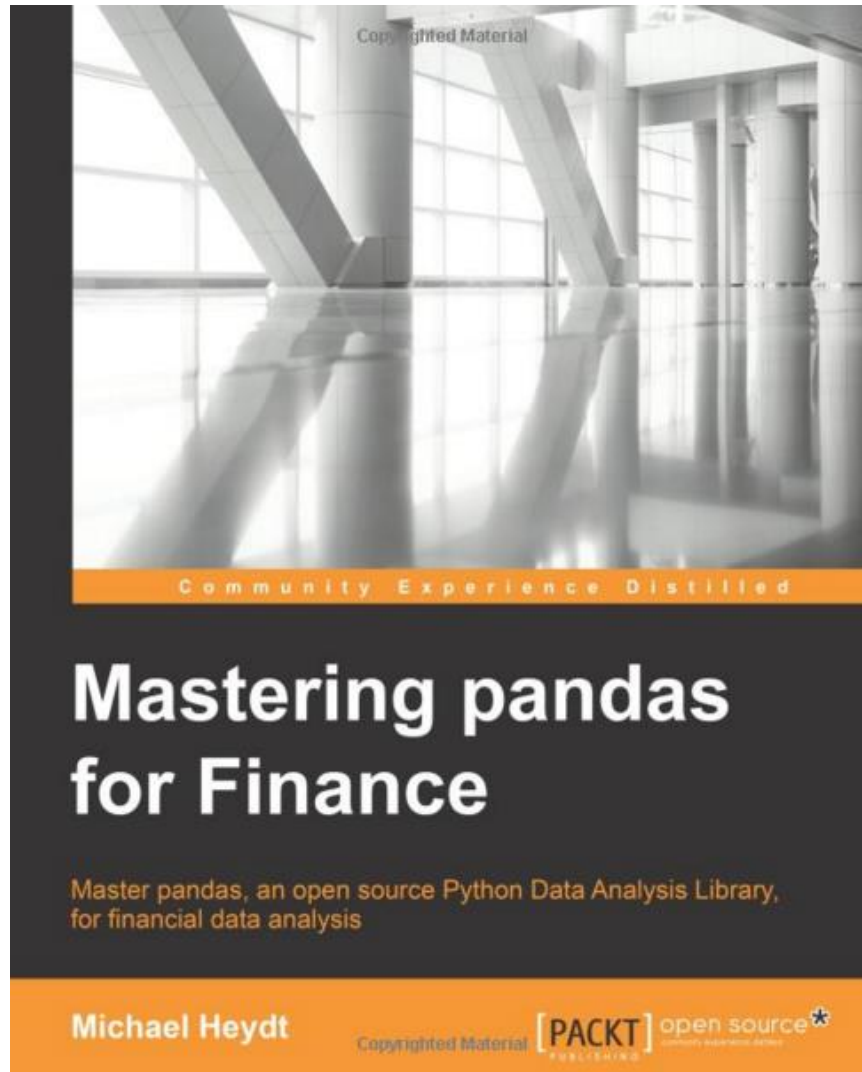
**Nick Pentreath,**  
**Machine Learning with Spark –**  
**Tackle Big Data with Powerful Spark Machine Learning Algorithms,**  
**Packt Publishing, 2015**



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



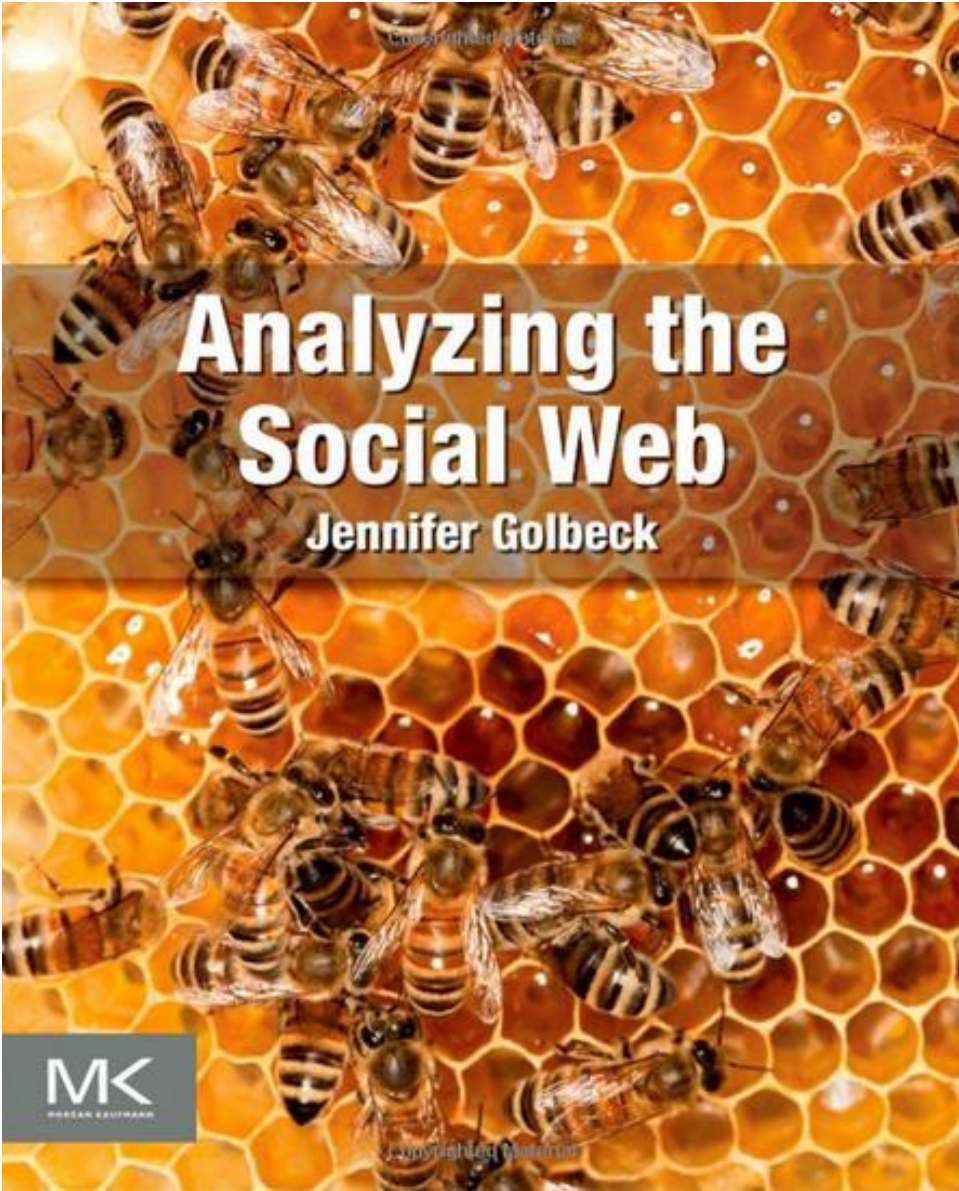
# Michael Heydt , Mastering Pandas for Finance, Packt Publishing, 2015



**Business Insights**  
**with**  
**Social Analytics**

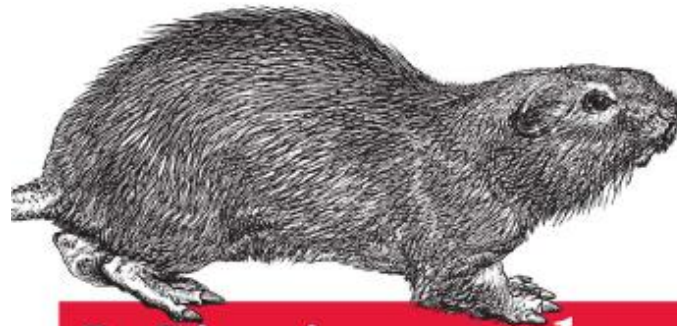
# Analyzing the Social Web: Social Network Analysis

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*



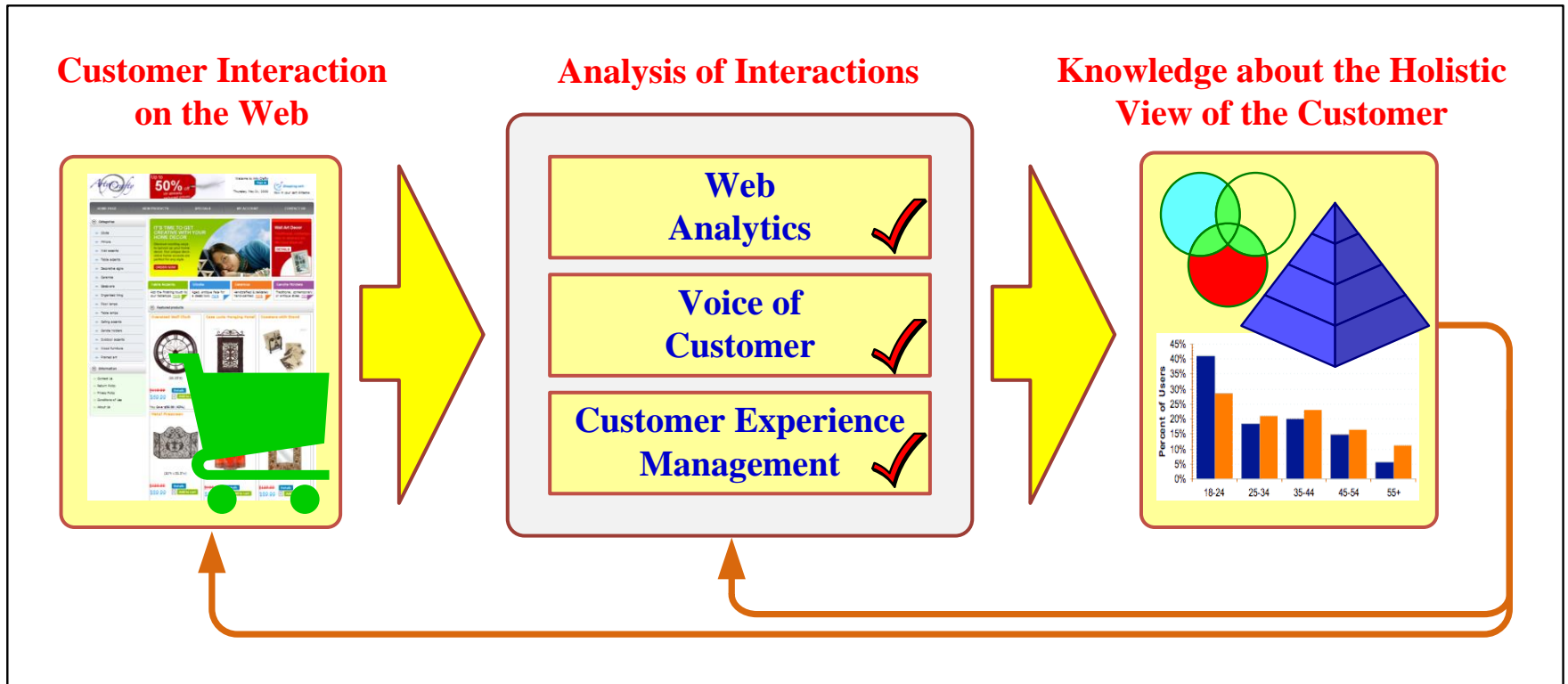
Mining the  
Social Web

O'REILLY®

*Matthew A. Russell*

# Web Mining Success Stories

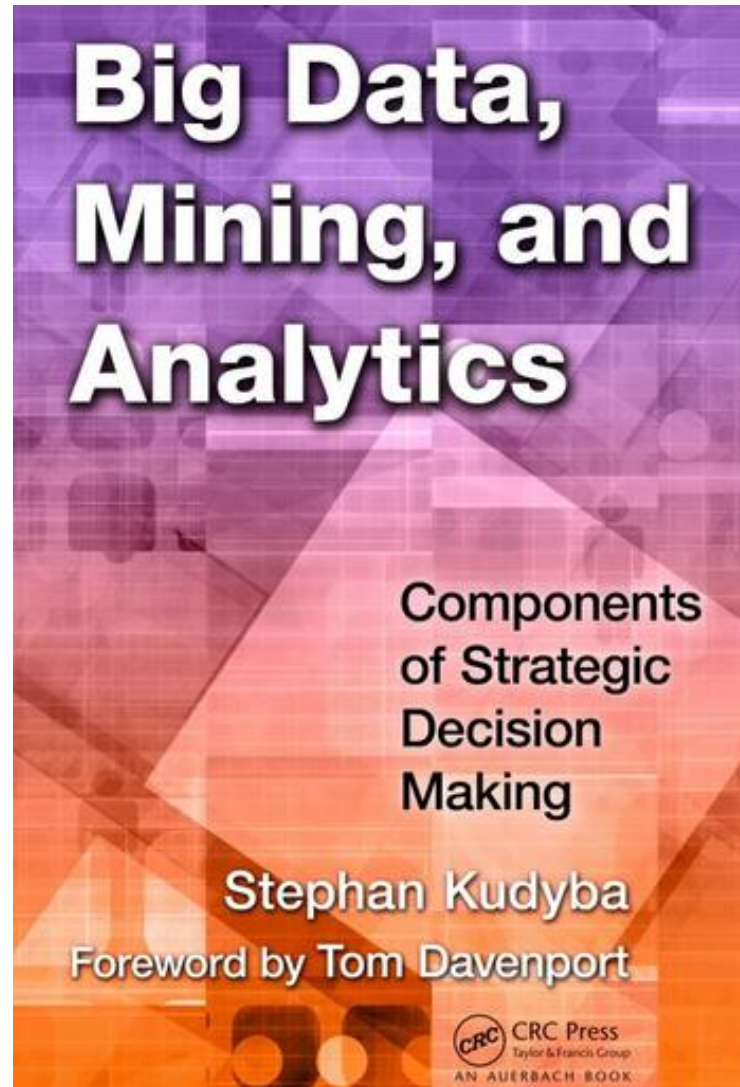
- Amazon.com, Ask.com, Scholastic.com, ...
- Website Optimization Ecosystem



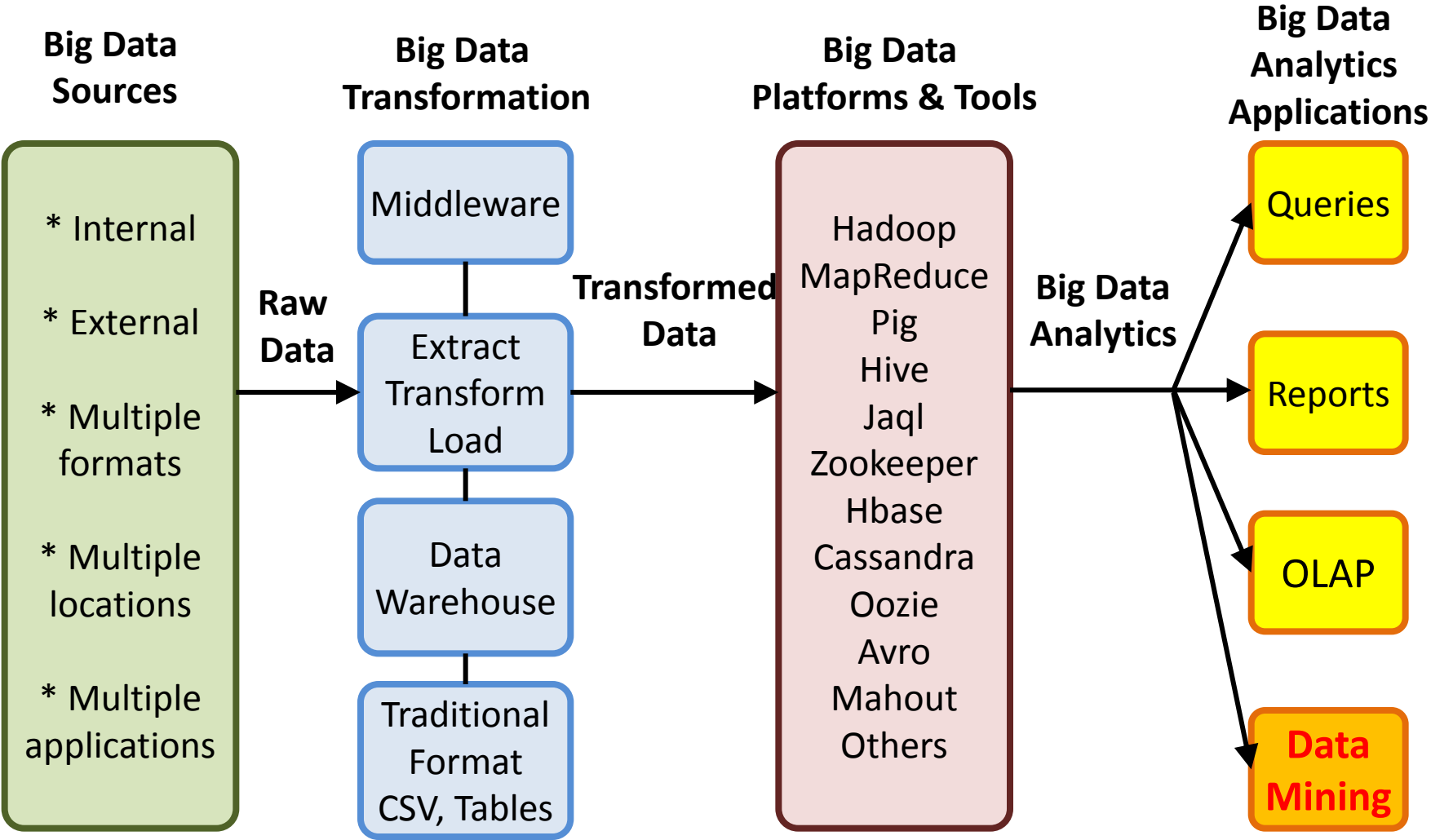


**Big Data**  
**Analytics**  
and  
**Data Mining**

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



# Architecture of Big Data Analytics



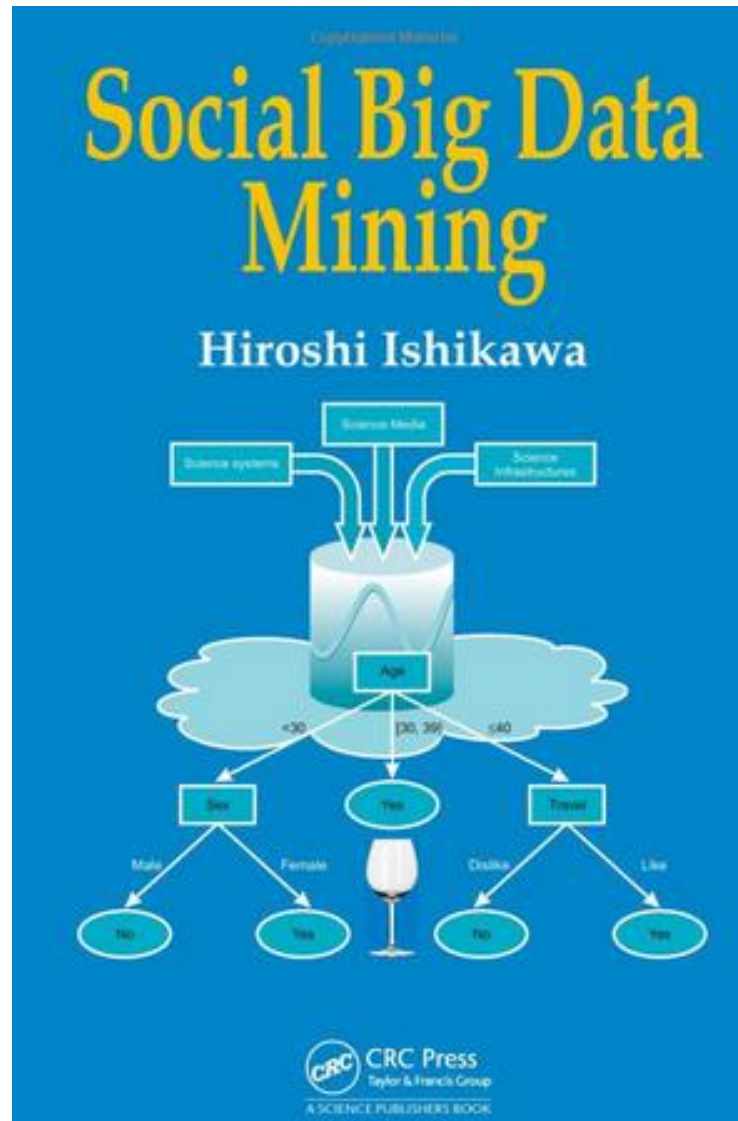
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

# Architecture of Big Data Analytics



# Social Big Data Mining

(Hiroshi Ishikawa, 2015)

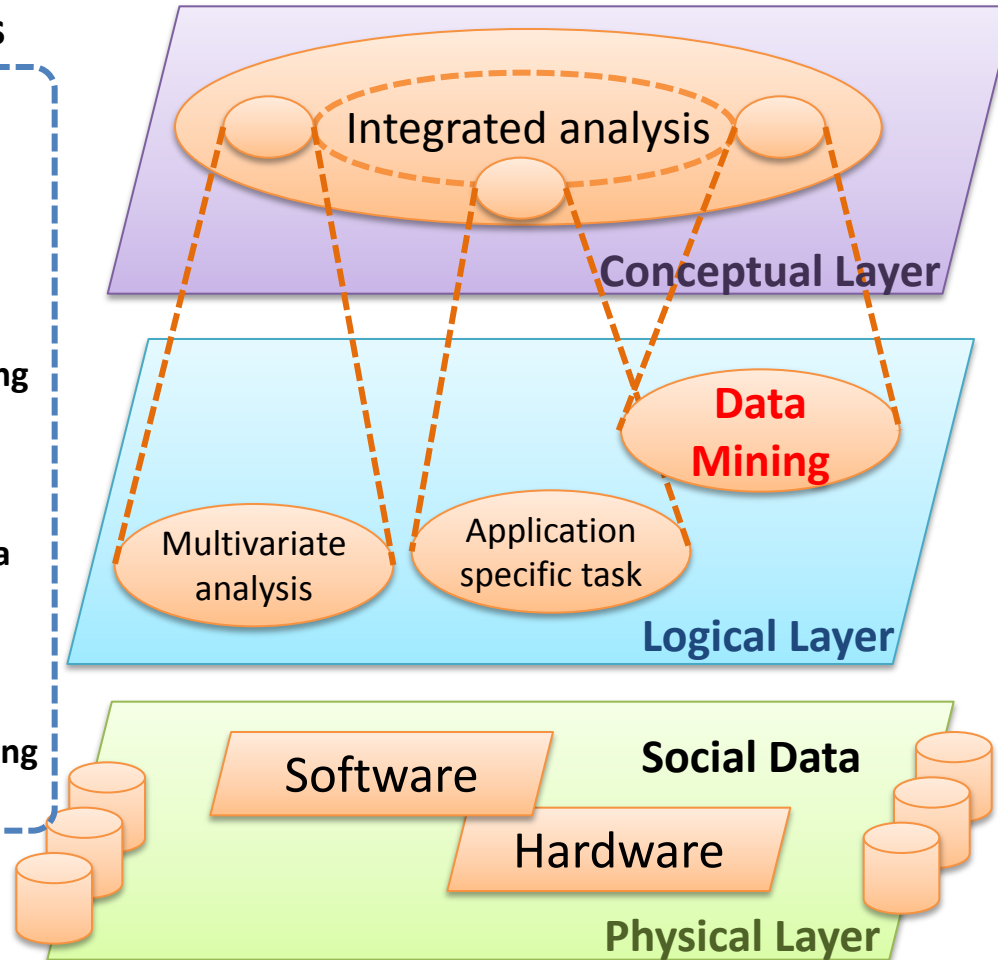


# Architecture for Social Big Data Mining

(Hiroshi Ishikawa, 2015)

## Enabling Technologies

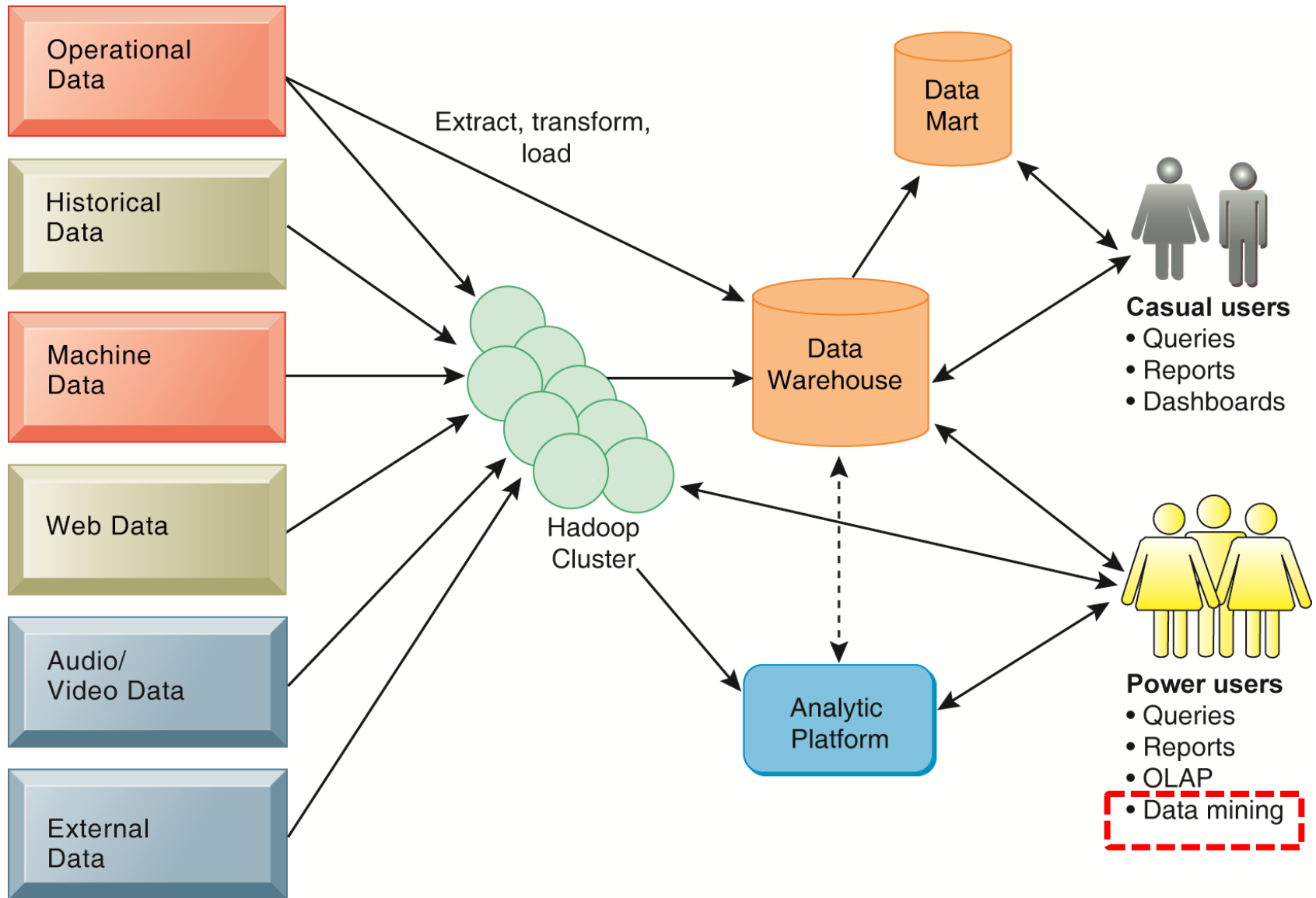
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distributed processing



## Analysts

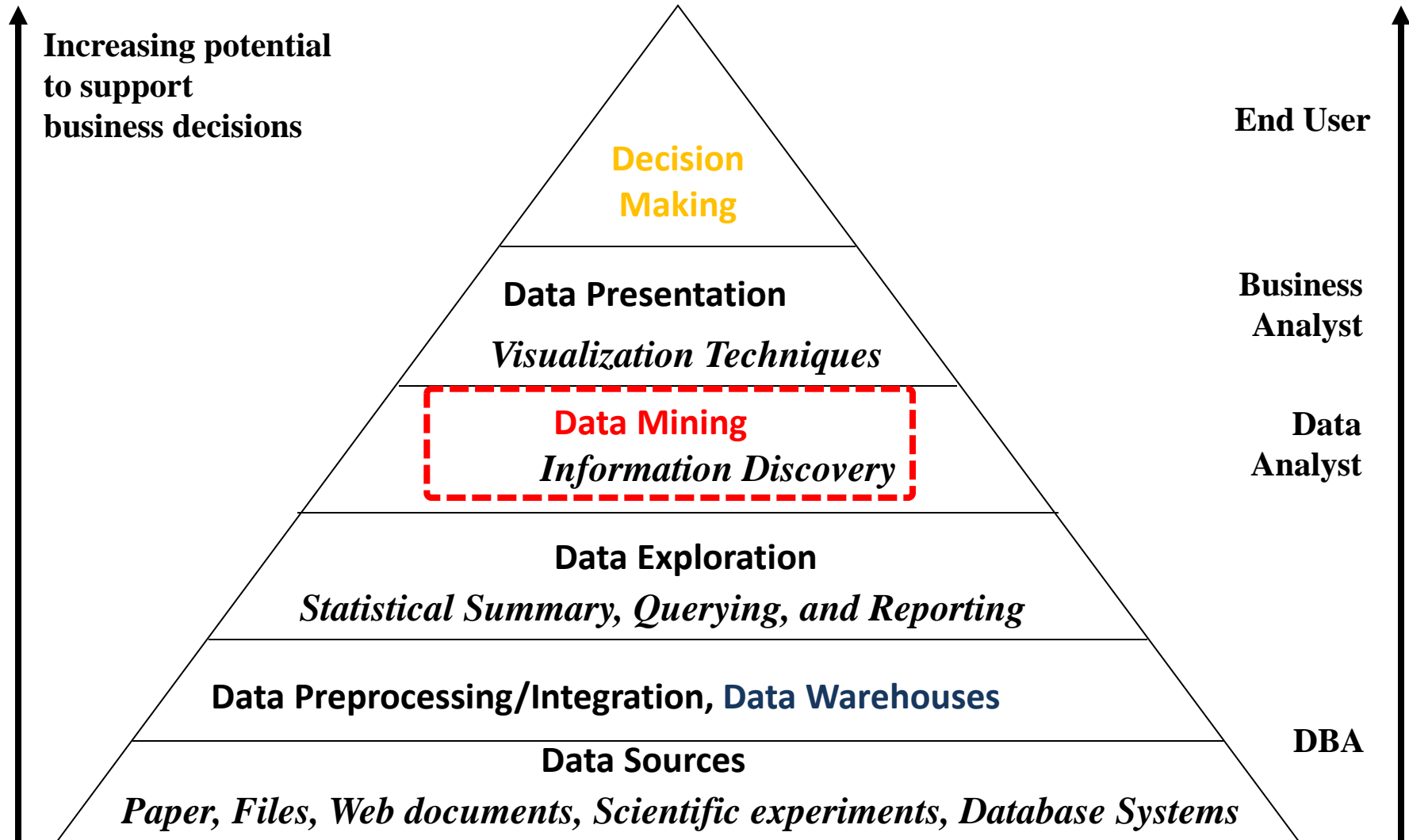
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

# Business Intelligence (BI) Infrastructure



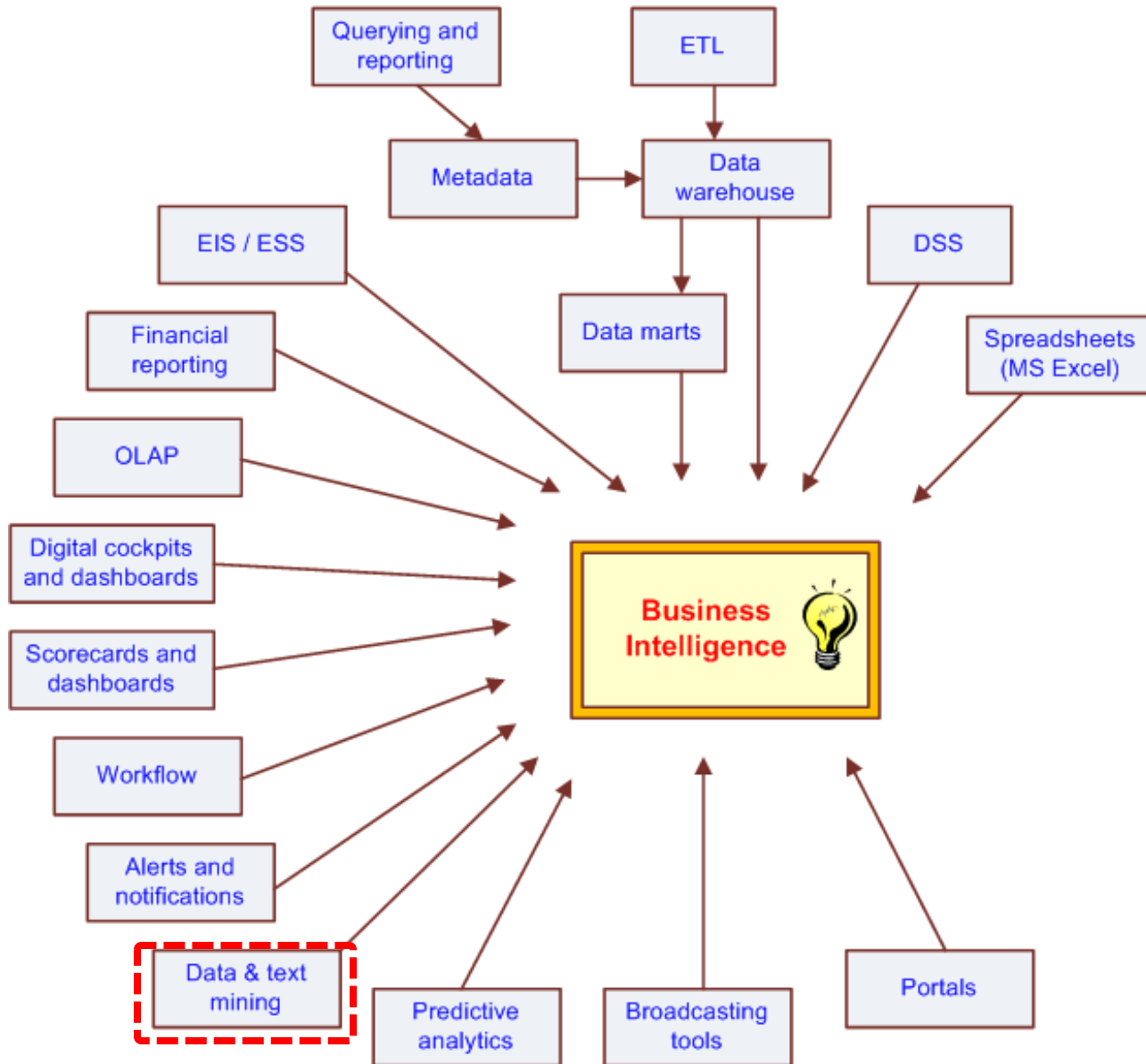
# Data Warehouse

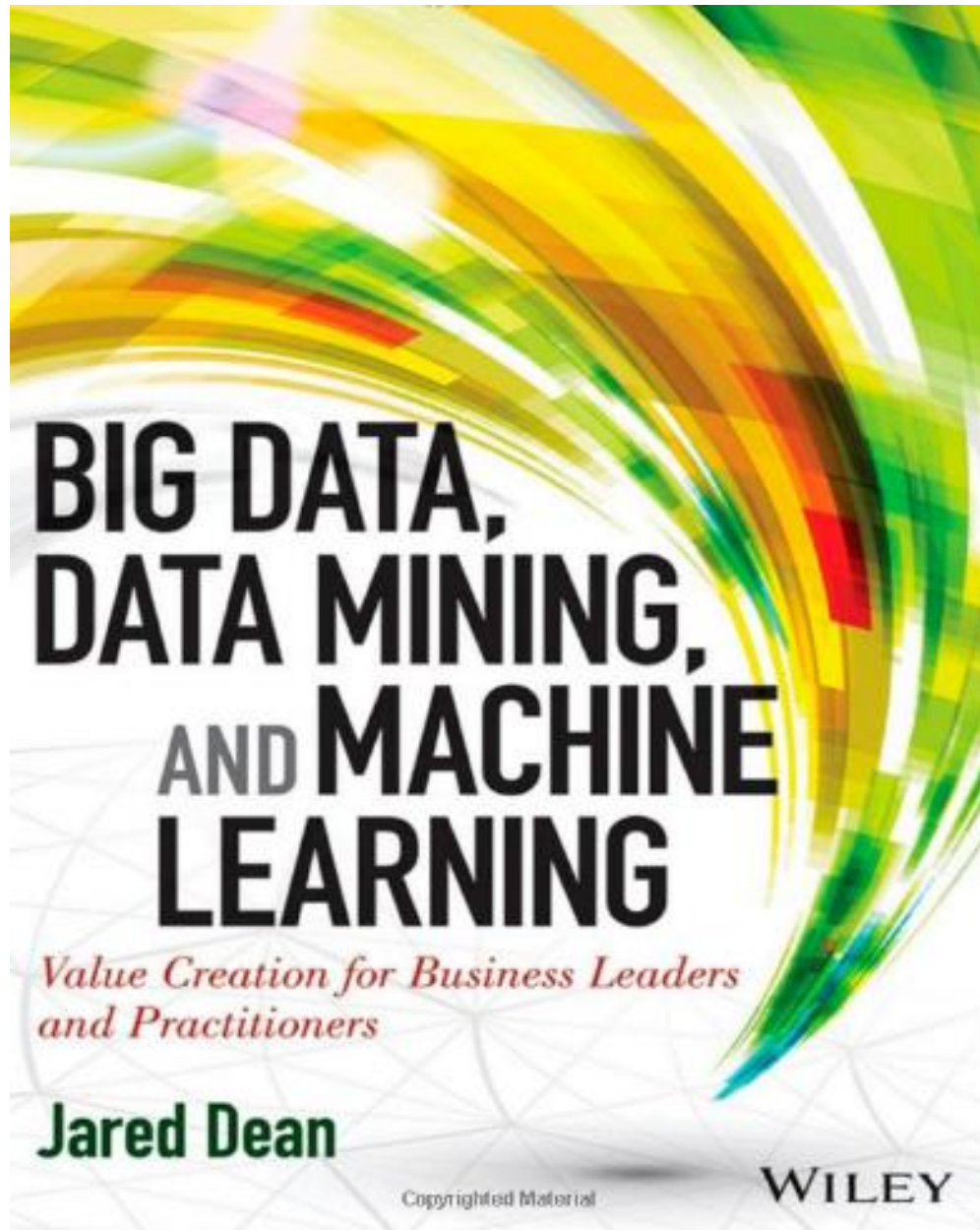
## Data Mining and Business Intelligence





# The Evolution of BI Capabilities

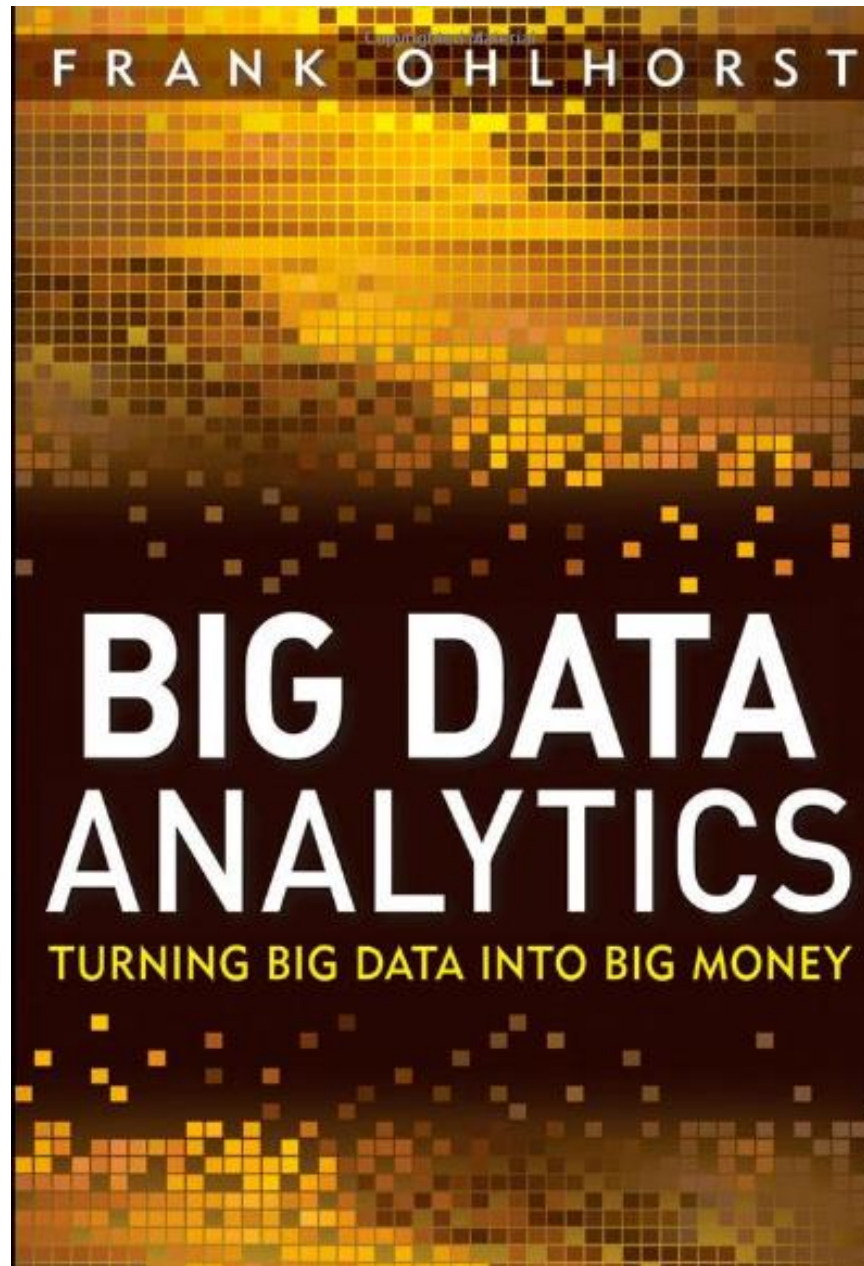




# Deep Learning

## 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



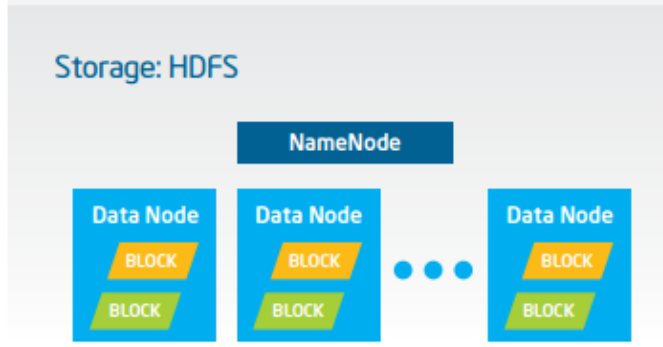
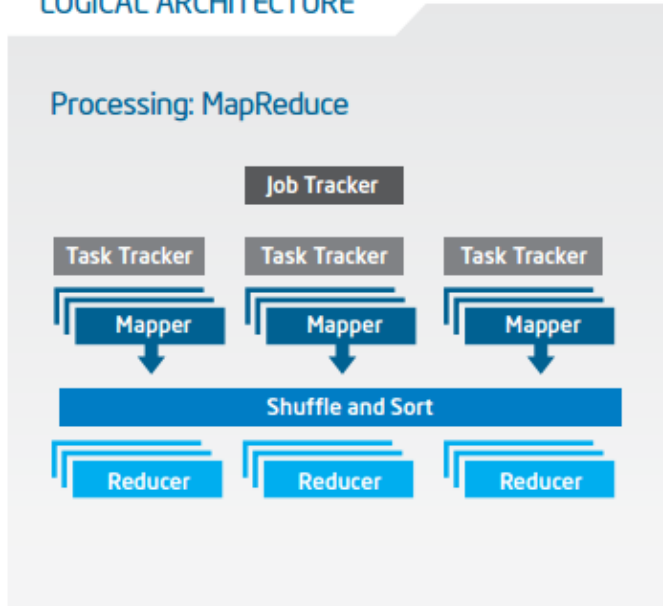
## DATA

Complex by structure

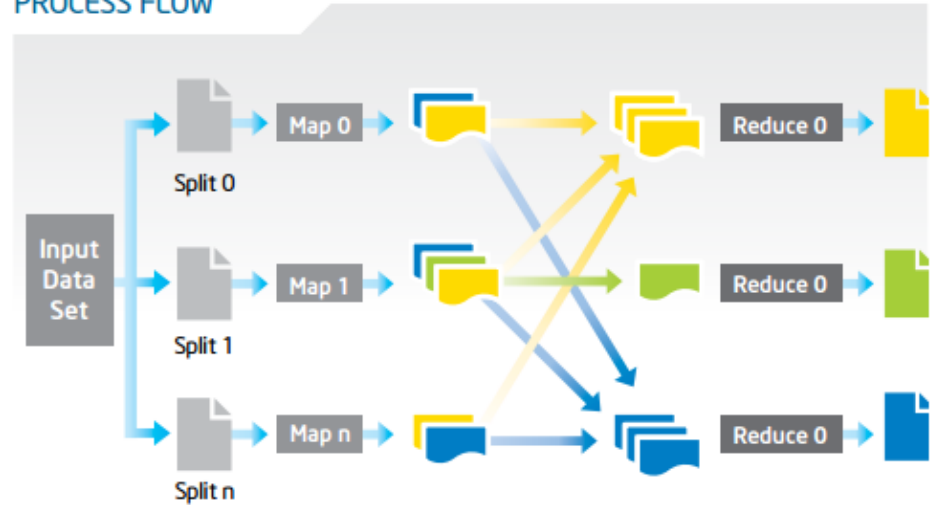


# Big Data with Hadoop Architecture

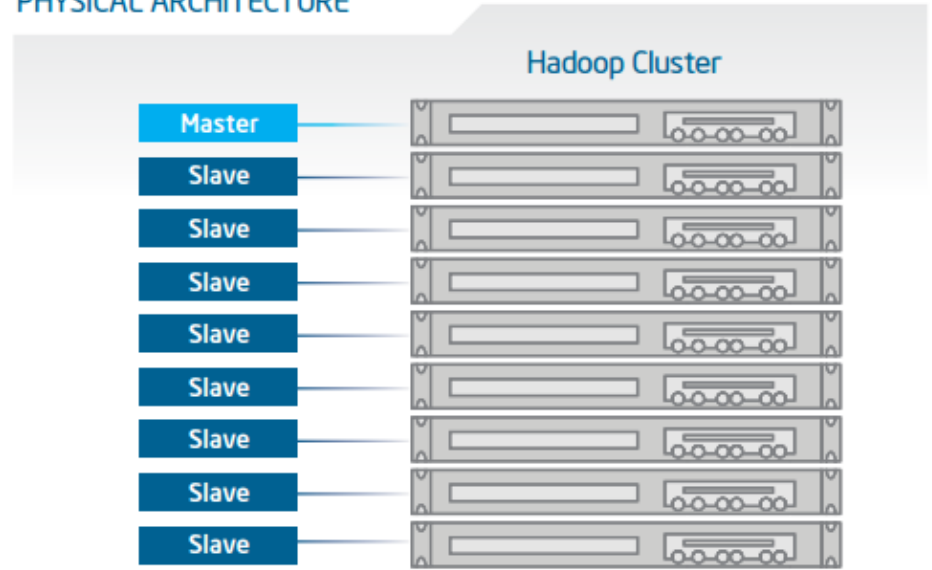
## LOGICAL ARCHITECTURE



## PROCESS FLOW



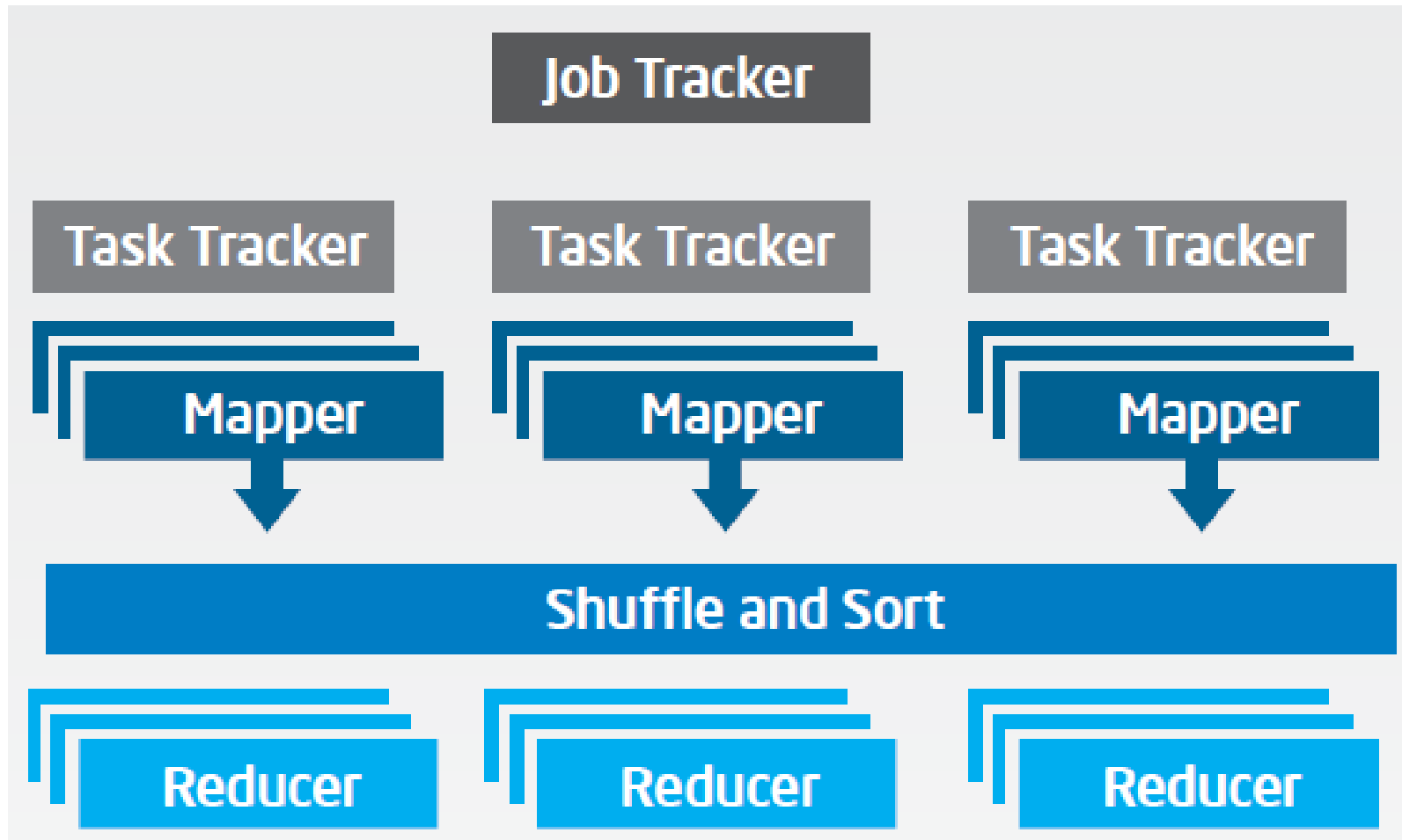
## PHYSICAL ARCHITECTURE



# Big Data with Hadoop Architecture

## Logical Architecture

### Processing: MapReduce

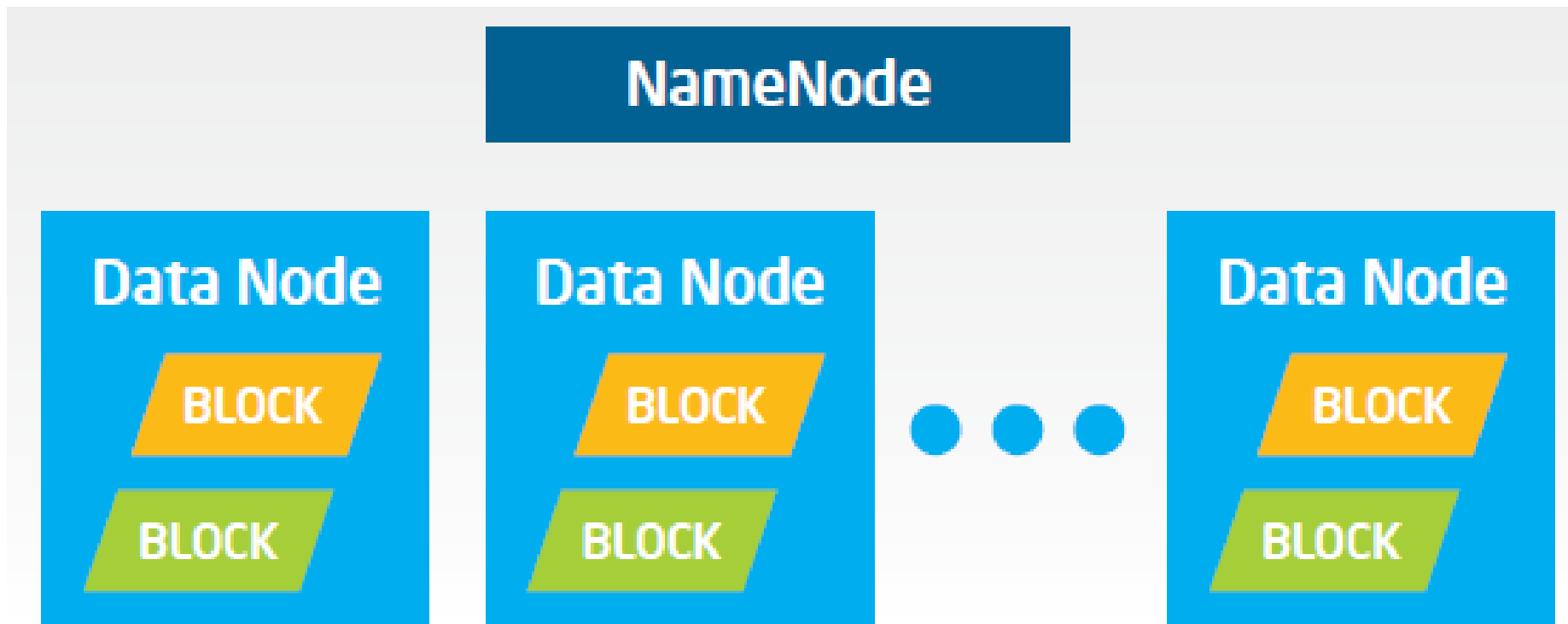




# Big Data with Hadoop Architecture

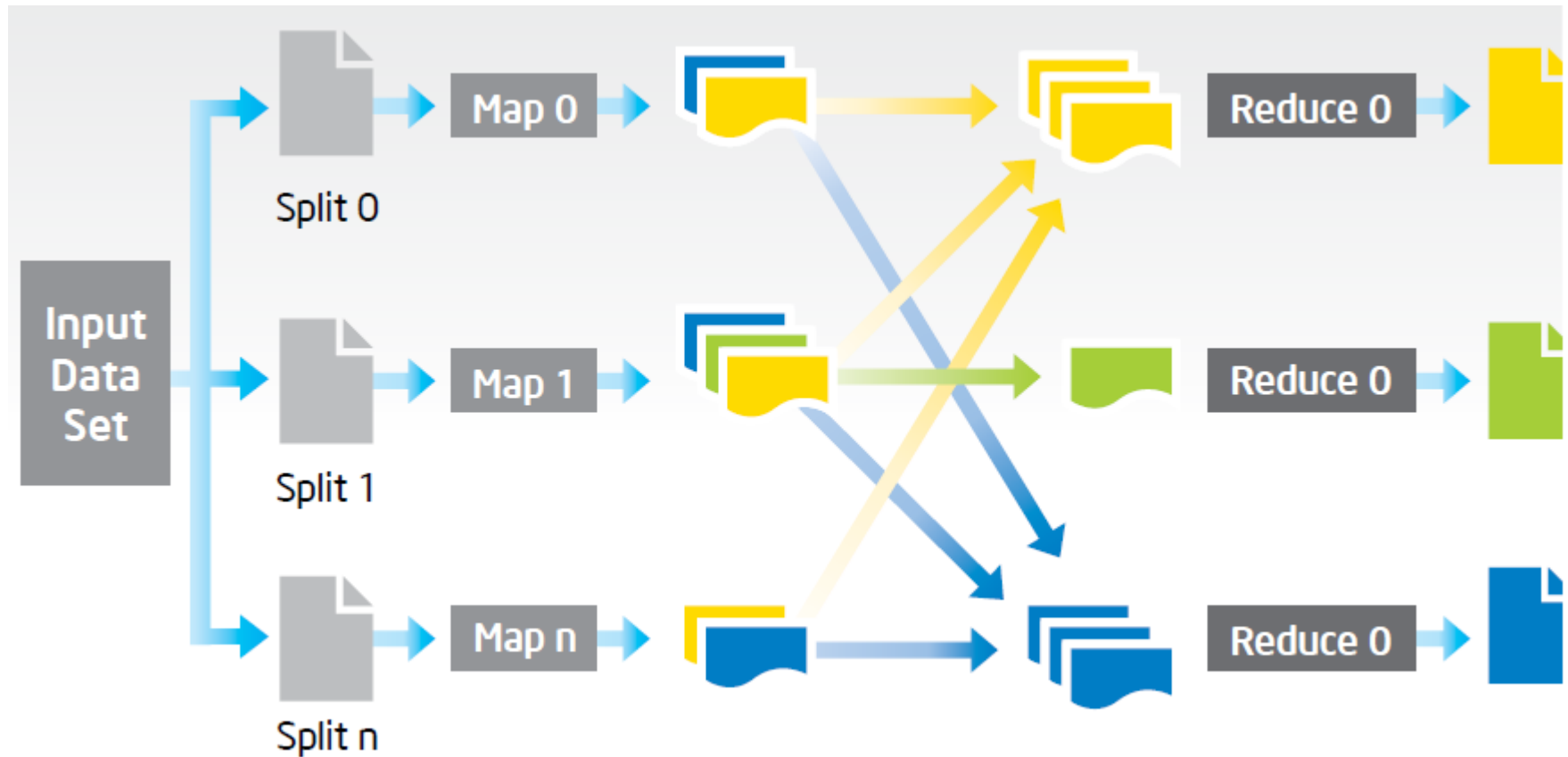
## Logical Architecture

Storage: HDFS



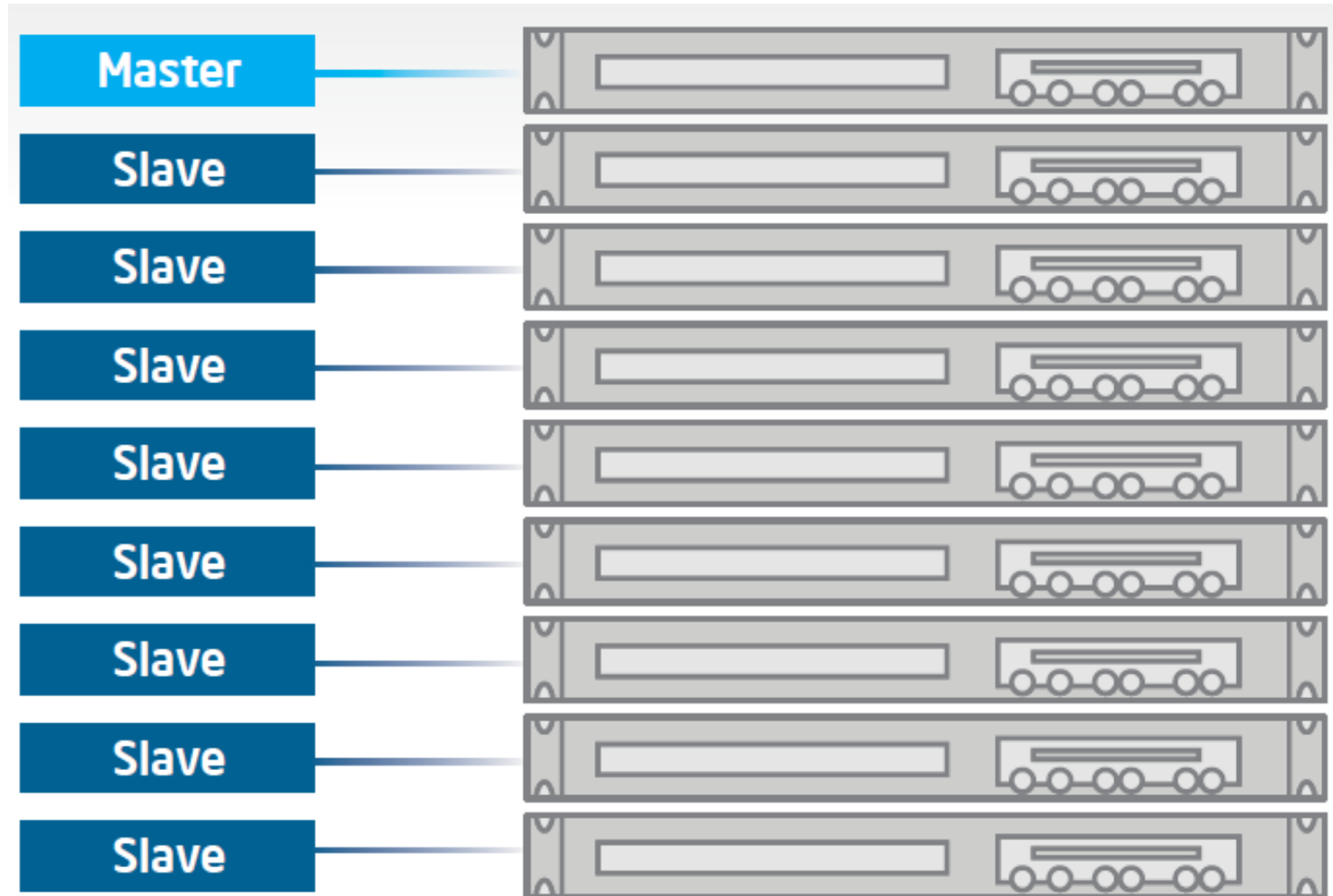
# Big Data with Hadoop Architecture

## Process Flow

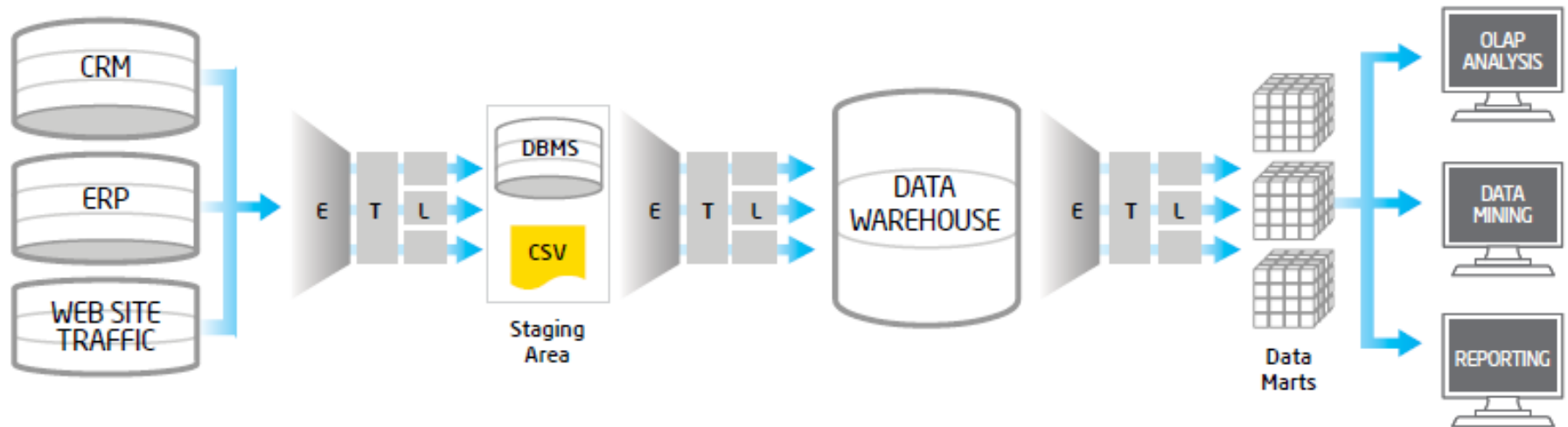


# Big Data with Hadoop Architecture

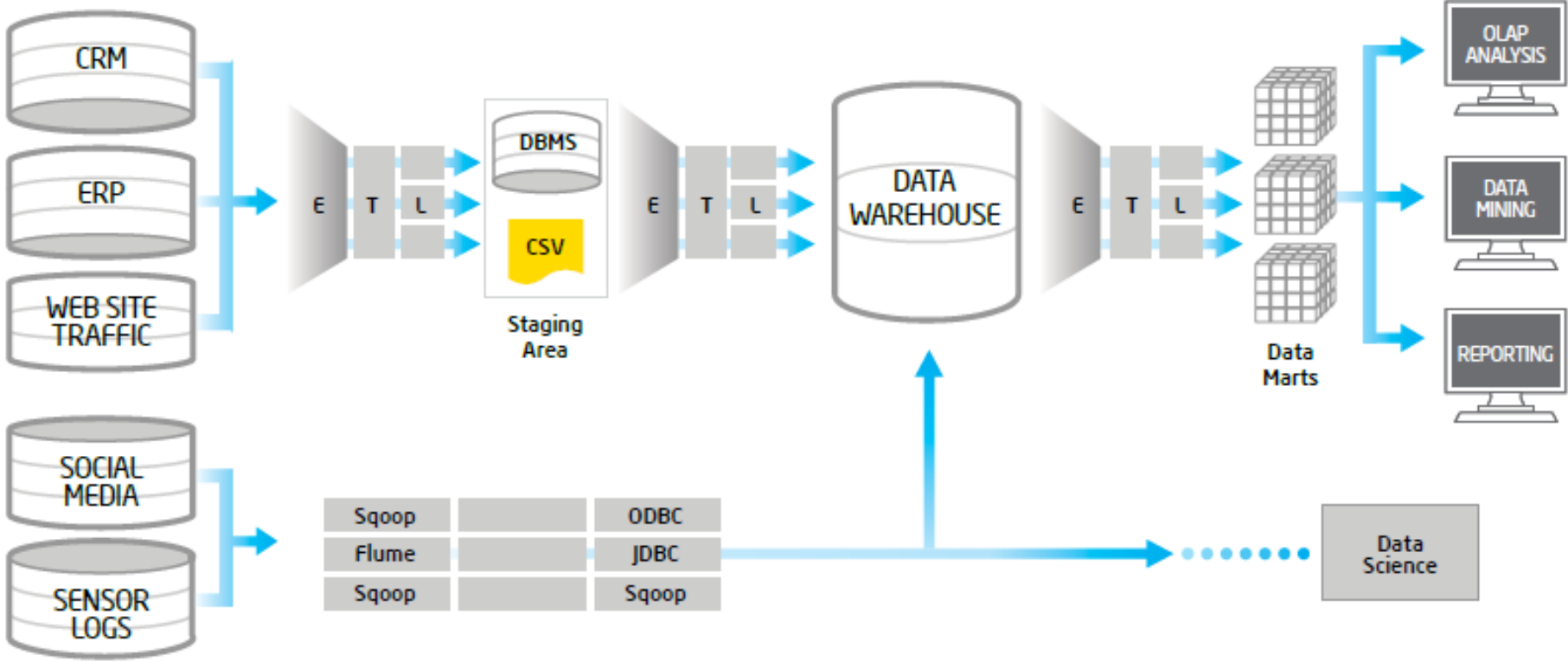
## Hadoop Cluster



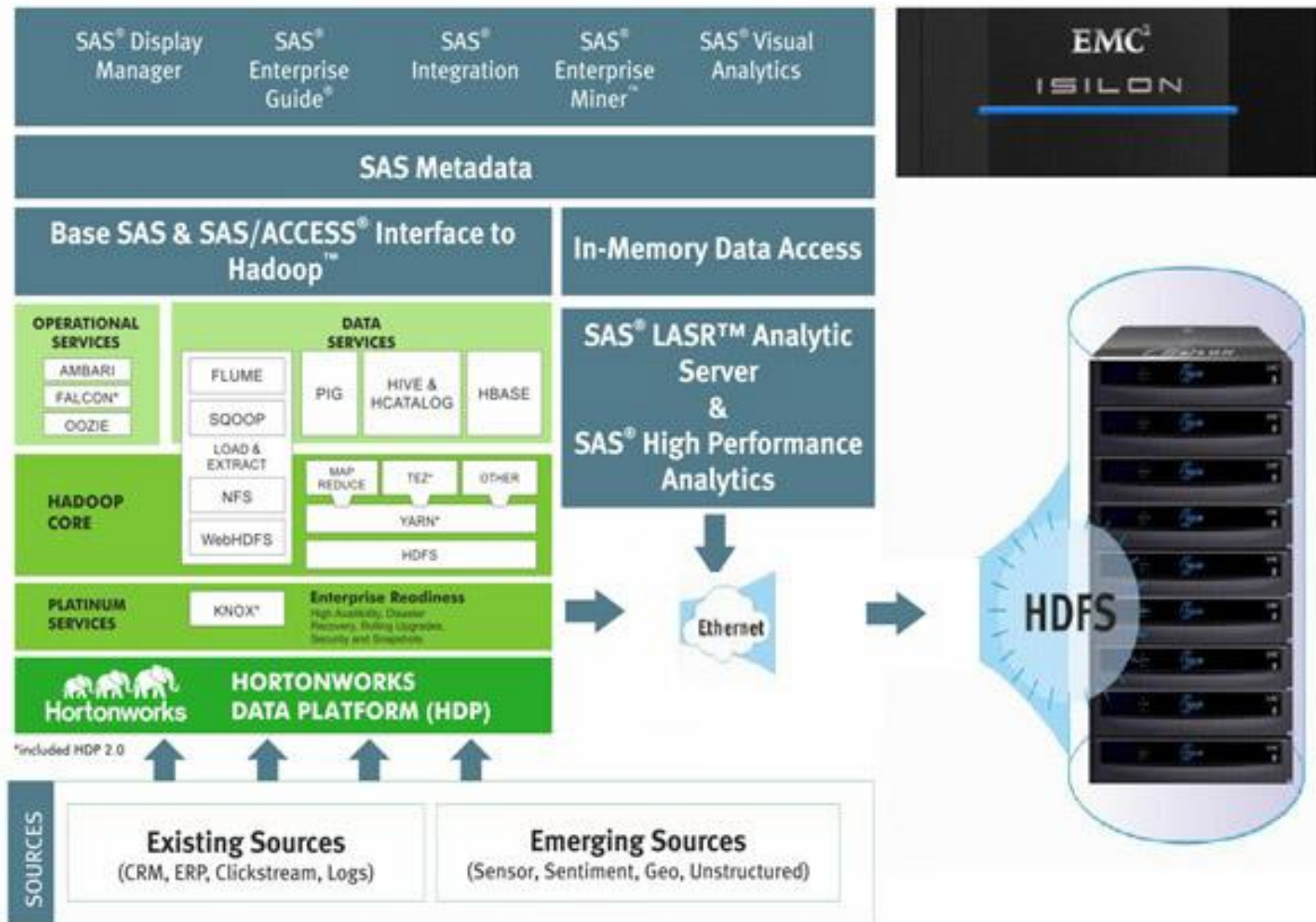
# Traditional ETL Architecture



# Offload ETL with Hadoop (Big Data Architecture)

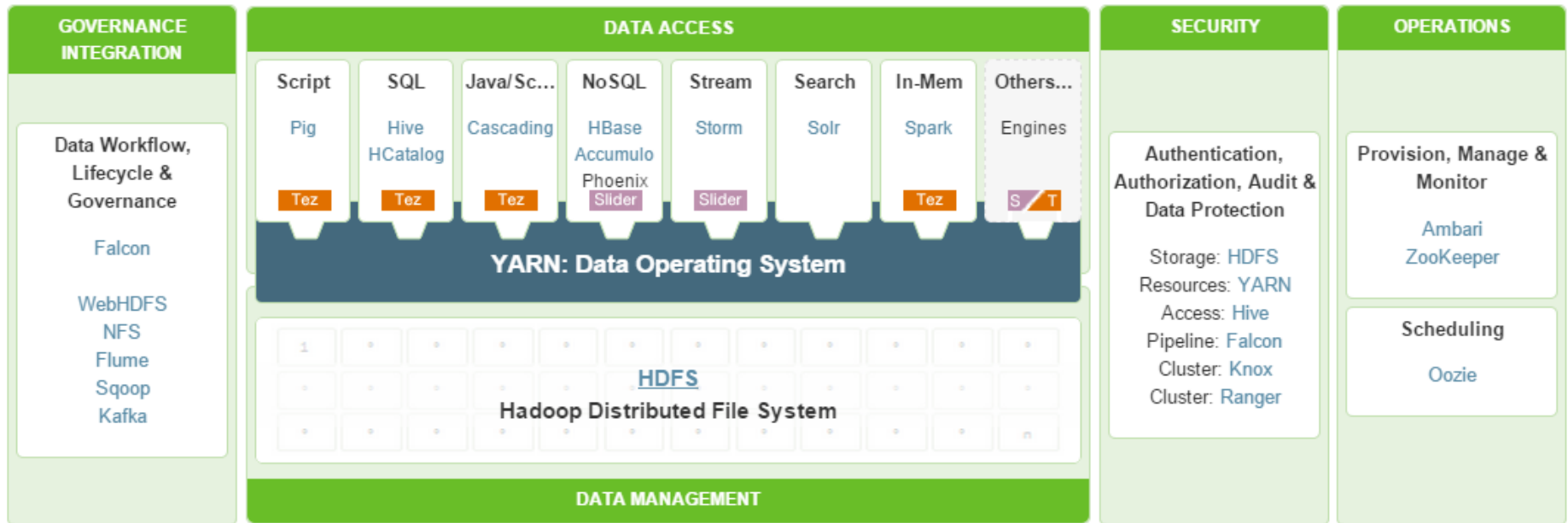


# Big Data Solution



# HDP

## A Complete Enterprise Hadoop Data Platform

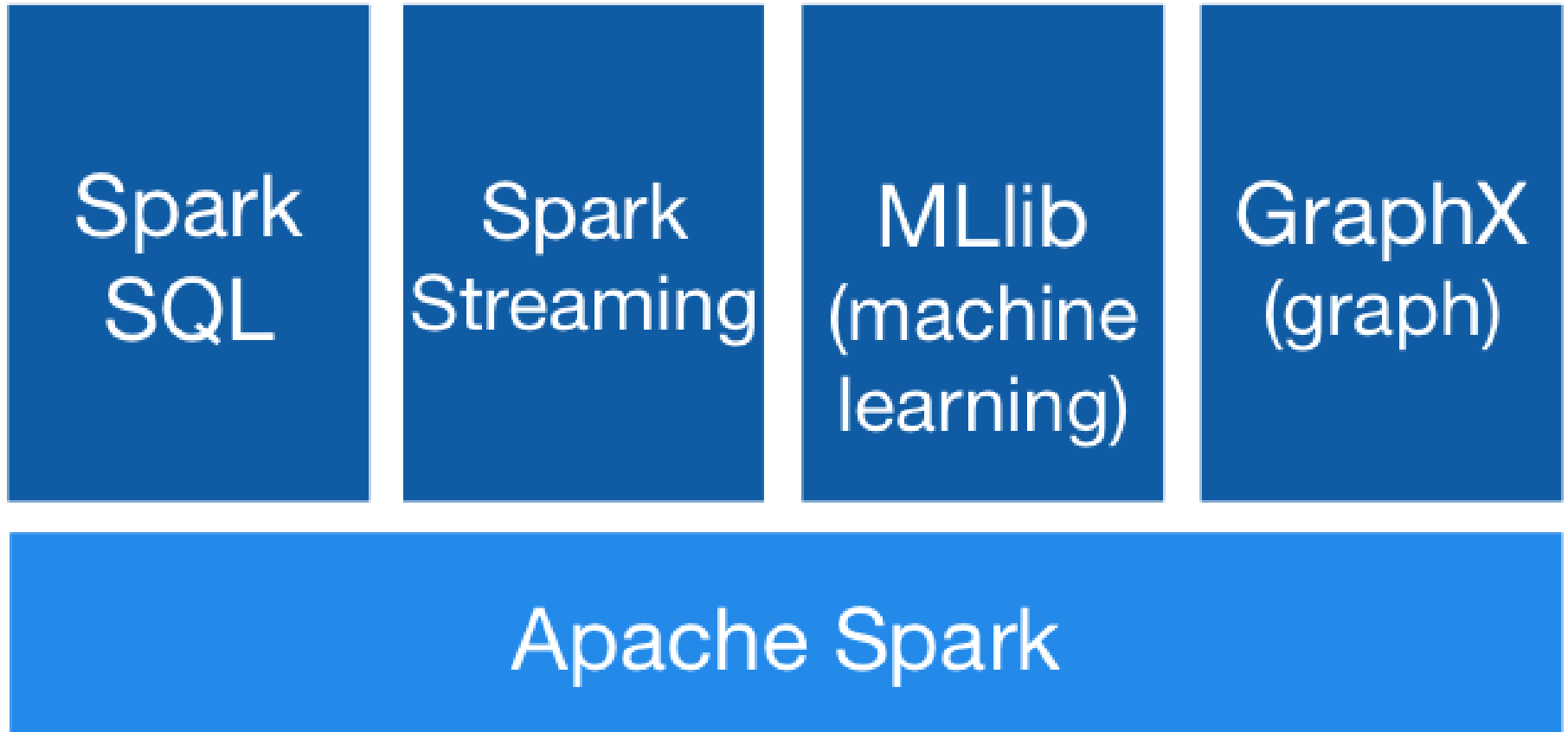


# Spark and Hadoop















# Spark Ecosystem

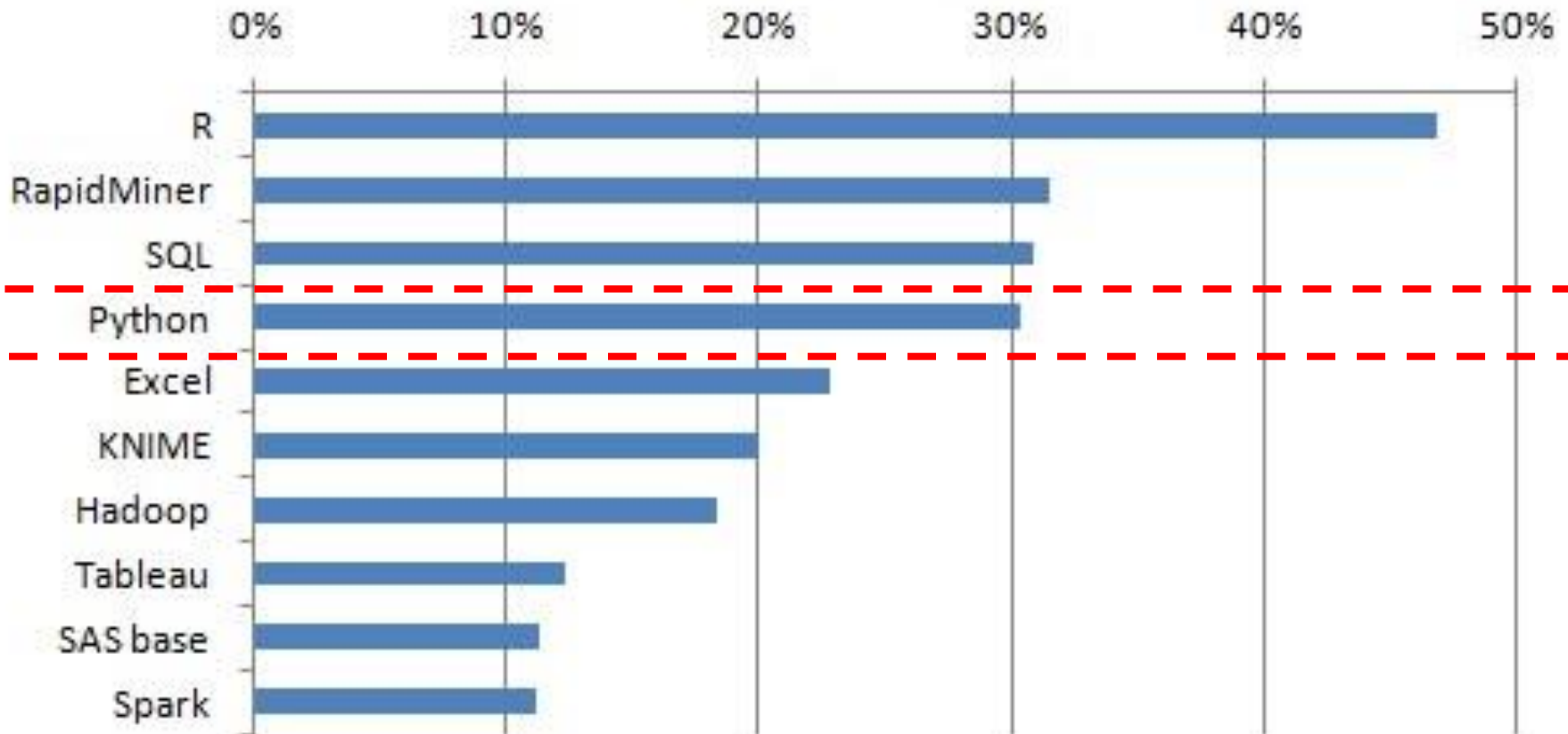


# Python for Big Data Analytics

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

Language Rank	Types	2015 Spectrum Ranking	2014 Spectrum Ranking
1. Java		100.0	100.0
2. C		99.9	99.3
3. C++		99.4	95.5
4. Python		96.5	93.5
5. C#		91.3	92.4
6. R		84.8	84.8
7. PHP		84.5	84.5
8. JavaScript		83.0	78.9
9. Ruby		76.2	74.3
10. Matlab		72.4	72.8

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



# 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*

**W**hen 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 第五屆大數據資料科學家競賽

## 文字分析與數位行銷大賽



最新消息

大賽起源

活動辦法

我要報名

常見問題



第五屆 大數據資料科學家競賽

## 文字分析 & 數位行銷大賽

資料科學家  
校園競賽系列

挑戰  
**\$300,000**  
高額總獎金

SAS 與  
玉山銀行  
優先招募與面試

我要報名

### 文字分析 & 數位行銷大賽

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!

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

# SAS第五屆大數據資料科學家競賽

## 文字分析與數位行銷大賽

### S 活動時間與地點

1. 報名日期：2016年2月22日（一）至2016年3月18日（五）額滿為止
2. 起跑說明會：2016年3月25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)：2016年3月26日（六）至3月30日（三），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)

【台中場】地點：逢甲大學（台中市西屯區文華路100號）

時間：上午九點至下午五點，每梯次共計一天

日期	名額
2016年3月26日（星期六）	120 人
2015年3月27日（星期日）	120 人

【台北場】地點：台北大學 民生校區（台北市中山區民生東路三段67號）

時間：上午九點至下午五點，每梯次共計一天台北場

日期	名額
2016年3月26日（星期六）@東吳大學城中校區	120 人
2016年3月27日（星期日）	120 人



# 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/>

