Tamkang University
淡江大學

Tamkang University

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)

1082DM01
MI4 (M2244) (2744)
Tue 3, 4 (10:10-12:00) (B218)

Min-Yuh Day
戴敏育
Associate Professor
副教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/

2020-03-03

淡江大學108學年度第2學期
課程教學計畫表
Spring 2020 (2020.03 - 2020.06)

- 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) (2744)
- 開課資料：選修 單學期 2 學分 (2 Credits, Elective)
- 上課時間：週二 3,4 (Tue 10:10-12:00)
- 上課教室： B218

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)
  - AI人工智慧與大數據分析 (Artificial Intelligence and Big Data Analytics)
  - 關連分析 (Association Analysis)
  - 分類與預測 (Classification and Prediction)
  - 分群分析 (Cluster Analysis)
  - 機器學習與深度學習 (Machine Learning and Deep Learning)
  - SAS企業資料採礦實務 (SAS Enterprise Miner)
  - 巨量資料探勘個案分析與實作

Course Introduction

- This course introduces the fundamental concepts and applications technology of big data mining.
- Topics include
  - Big Data Mining
  - Artificial Intelligence and Big Data Analytics
  - Association Analysis
  - Classification and Prediction
  - Cluster Analysis
  - Machine Learning and Deep Learning
  - Data Mining Using SAS Enterprise Miner (SAS EM)
  - Case Study and Implementation of Big Data Mining

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	2020/03/03	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
2	2020/03/10	AI人工智慧與大數據分析 (Artificial Intelligence and Big Data Analytics)
3	2020/03/17	分群分析 (Cluster Analysis)
4	2020/03/24	個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis - K-Means using SAS EM)
5	2020/03/31	關連分析 (Association Analysis)
6	2020/04/07	個案分析與實作二 (SAS EM 關連分析)：Case Study 2 (Association Analysis using SAS EM)
7	2020/04/14	分類與預測 (Classification and Prediction)
8	2020/04/21	期中報告 (Midterm Project Presentation)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容(Subject/Topics)
9	2020/04/28	期中考試週
10	2020/05/05	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)
11	2020/05/12	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)
12	2020/05/19	機器學習與深度學習 (Machine Learning and Deep Learning)
13	2020/05/26	期末報告 (Final Project Presentation)
14	2020/06/02	畢業考試週
15	2020/06/09	教師彈性補充教學

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
  - Learning Data Mining with Python - Second Edition, Robert Layton, Packt Publishing, 2017

作業與學期成績計算方式

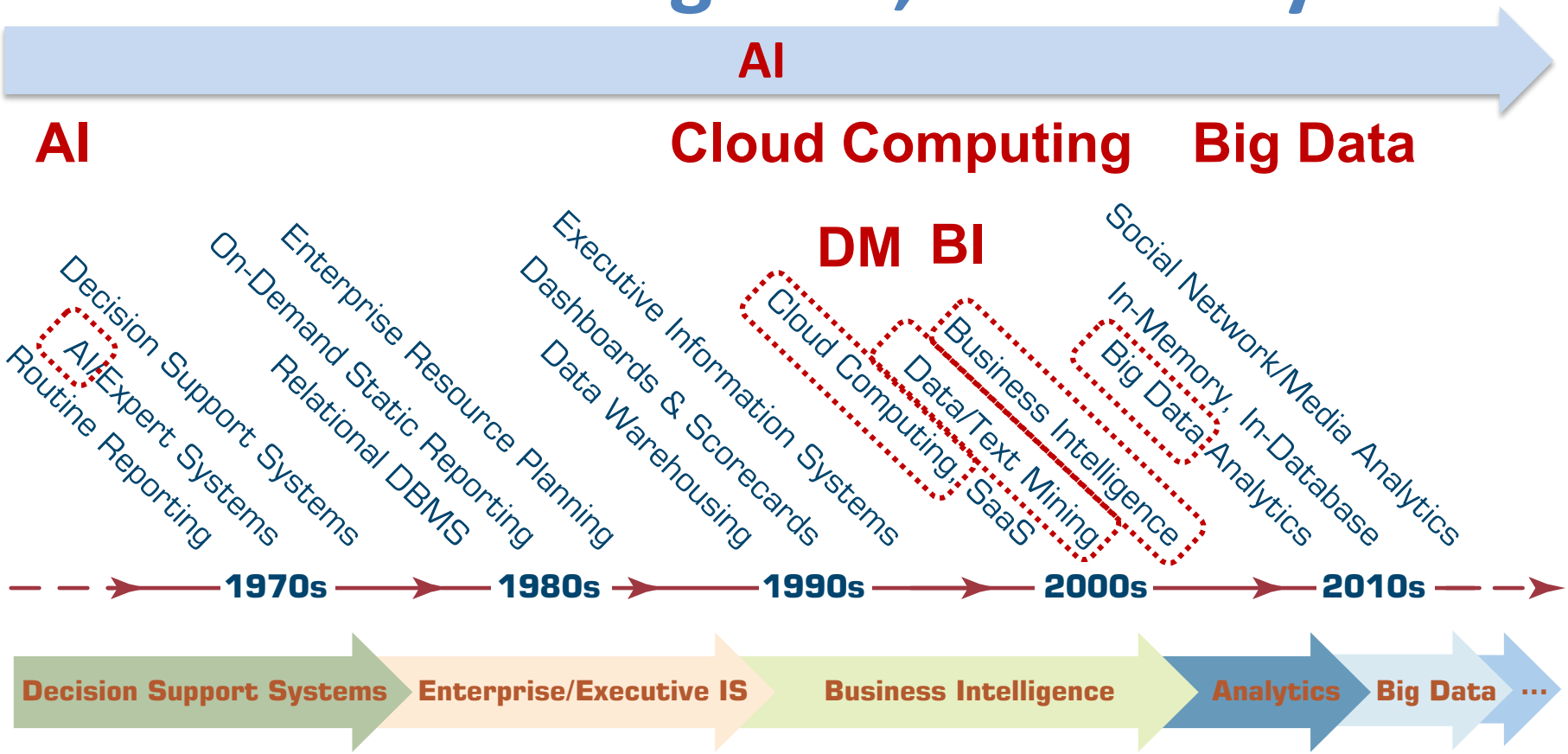
- 作業篇數
  - 3篇
- 學期成績計算方式
  - ☑期中評量：30 %
  - ☑期末評量：30 %
  - ☑其他（課堂參與及報告討論表現）：40 %

Team Term Project

- Term Project Topics
  - Big Data mining
  - Big Data Analytics
  - Business Intelligence
  - FinTech
- 3-4 人為一組
  - 分組名單於 2020/03/10 (二) 課程下課時繳交
  - 由班代統一收集協調分組名單

AI, Big Data, Cloud Computing

Evolution of Decision Support, Business Intelligence, and Analytics

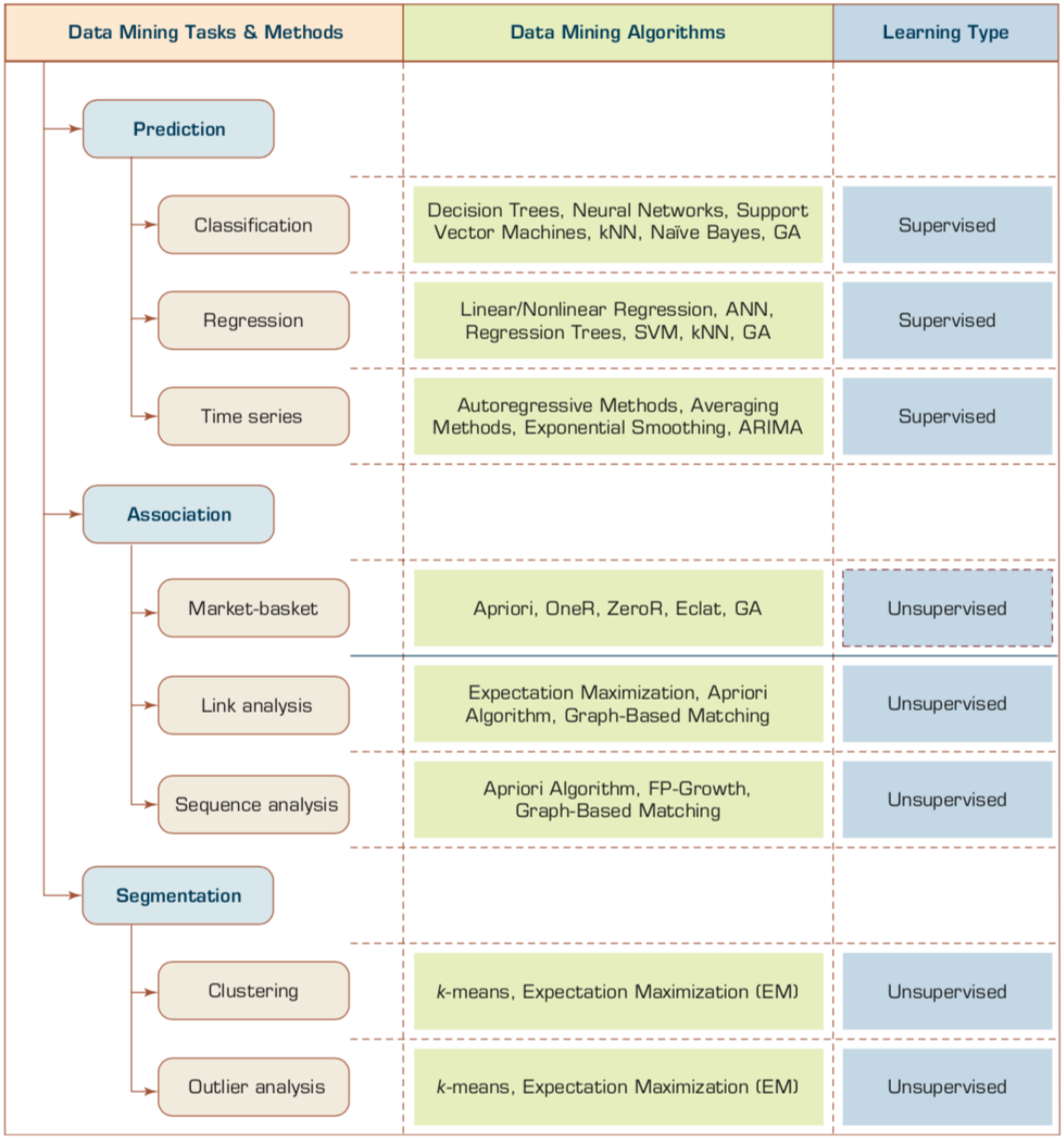


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017),
Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Data Mining
Is a Blend of Multiple Disciplines

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

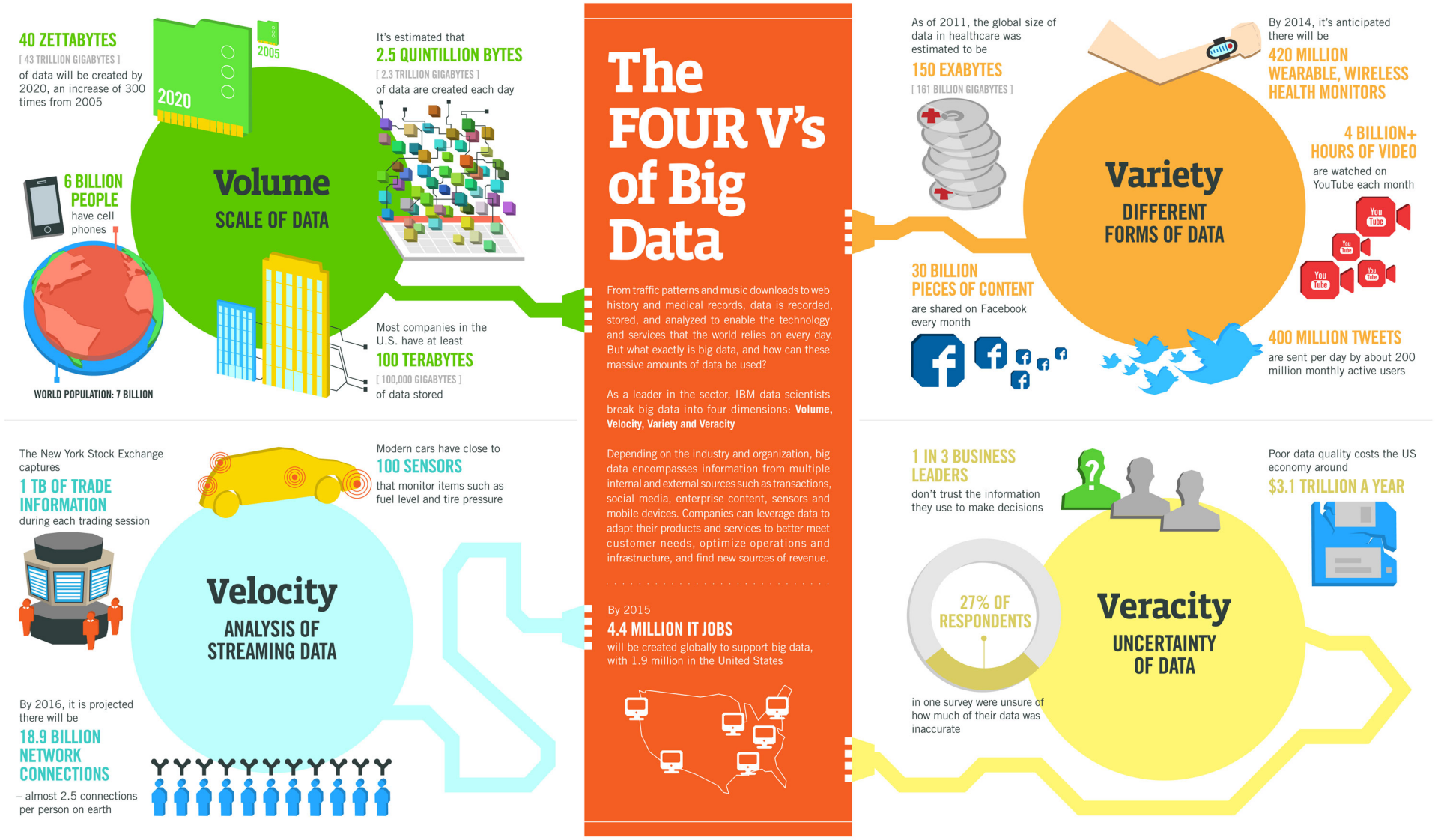
Data Mining Tasks & Methods



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Big Data Analytics and Data Mining

Big Data 4 V



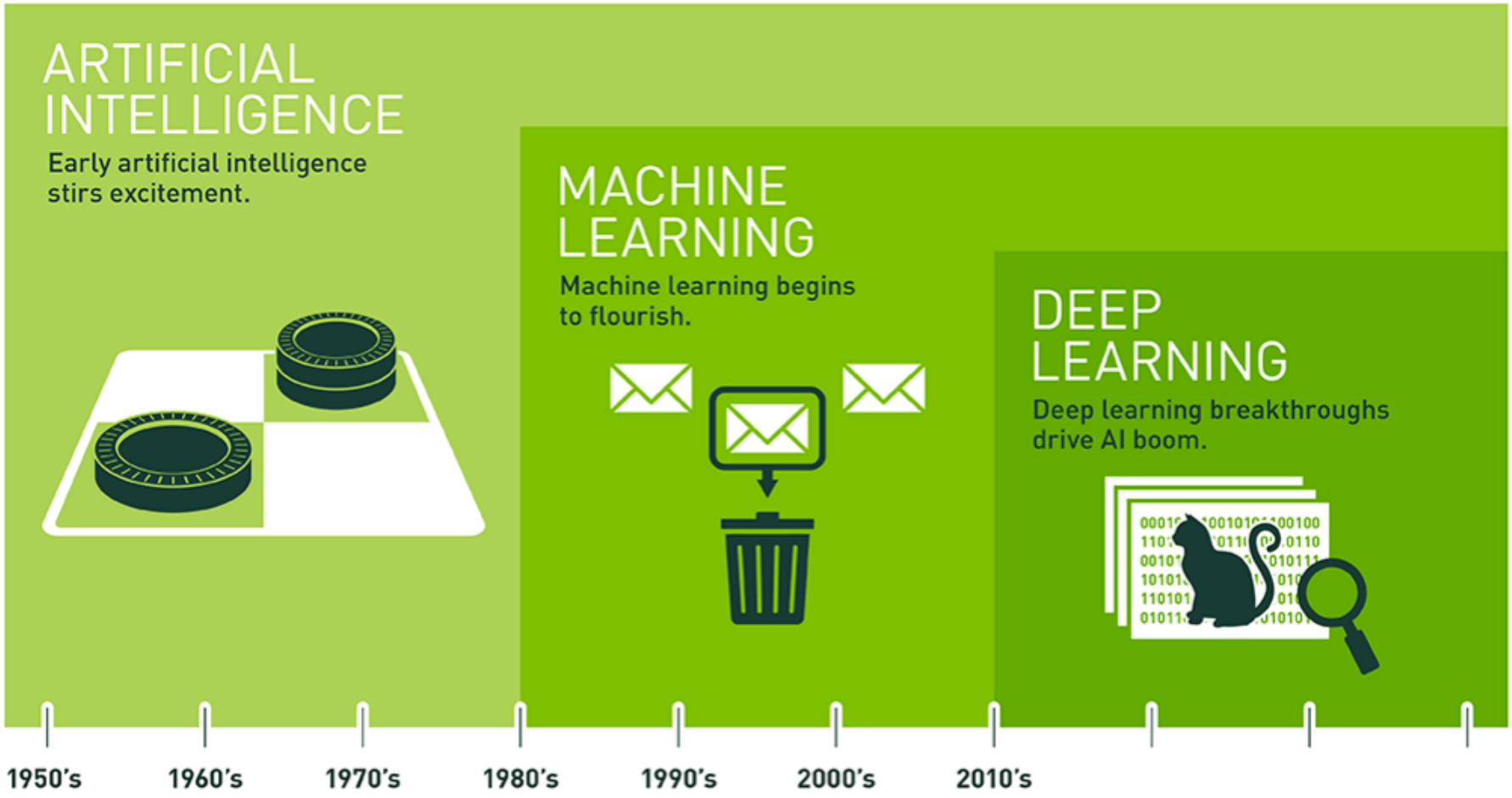
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Source: https://www-01.ibm.com/software/data/bigdata/

Value

Artificial Intelligence
Machine Learning & Deep Learning

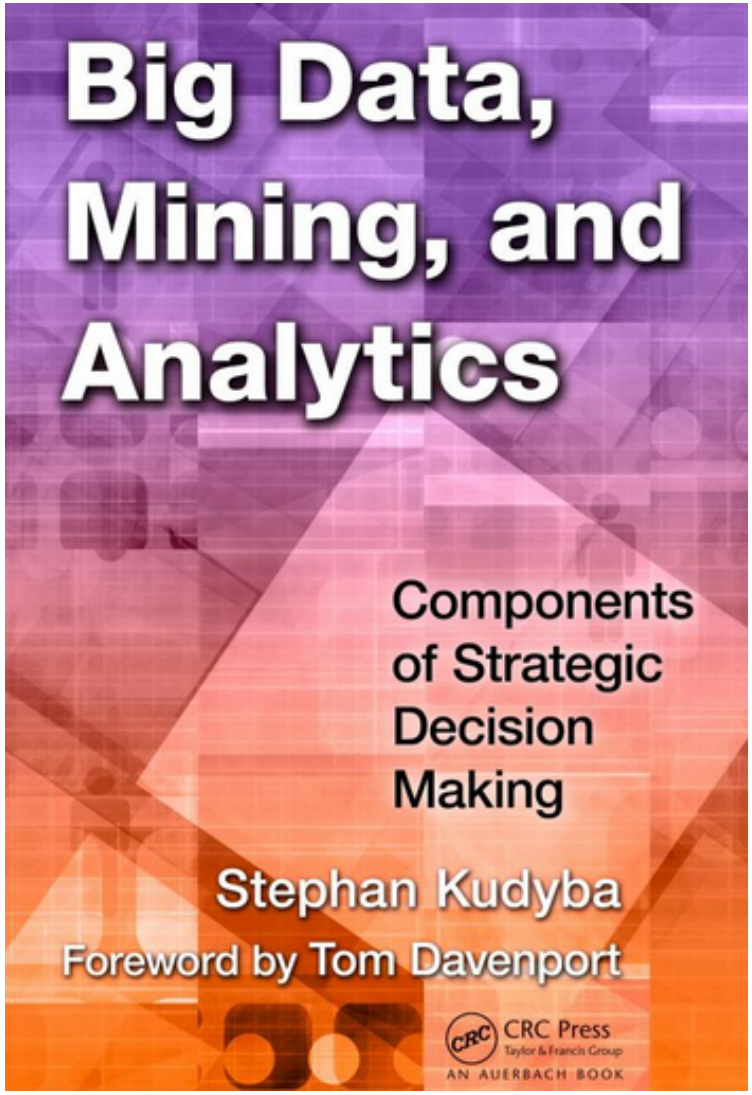


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Source: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

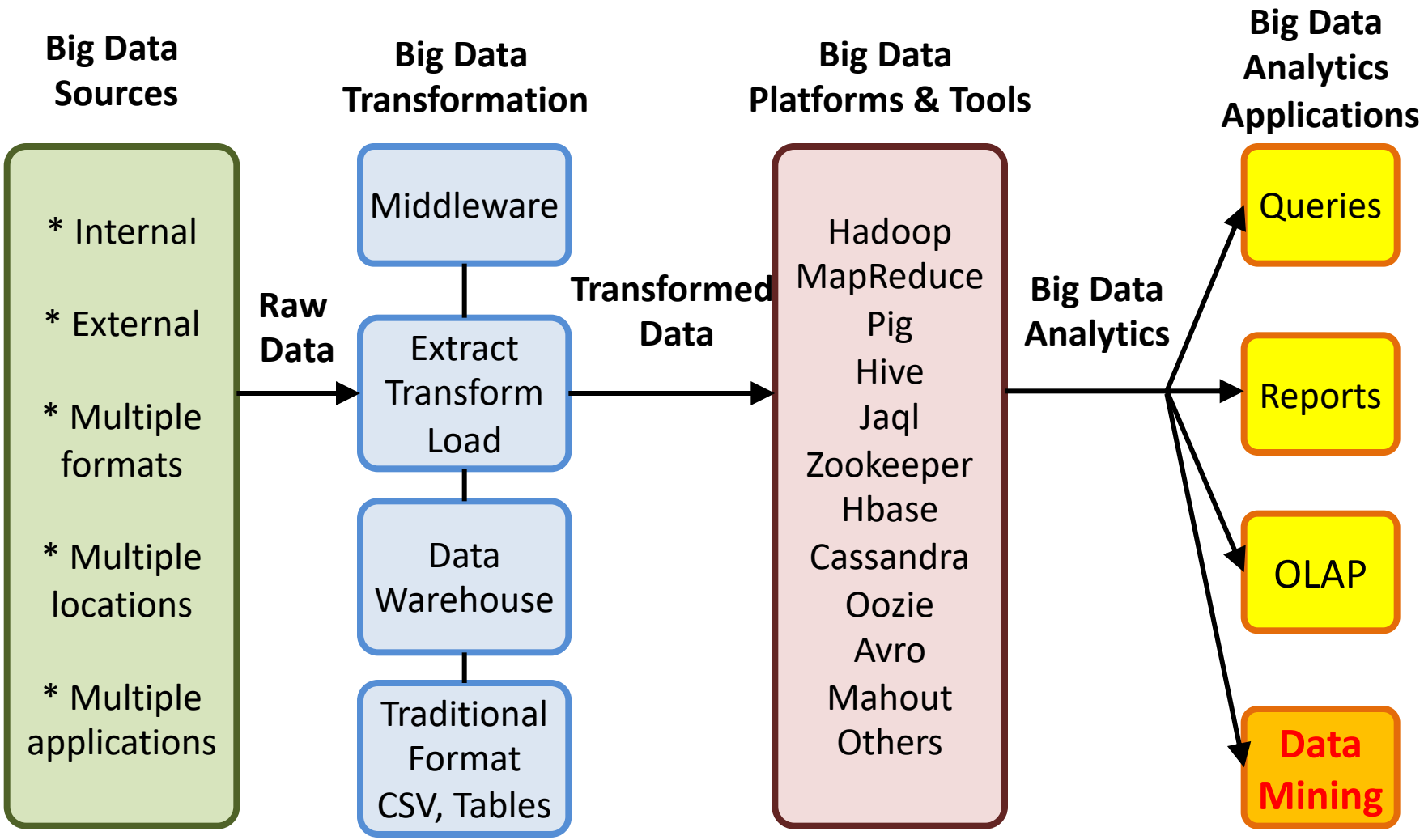
Stephan Kudyba (2014),

Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications



Source: http://www.amazon.com/gp/product/1466568704

Architecture of Big Data Analytics



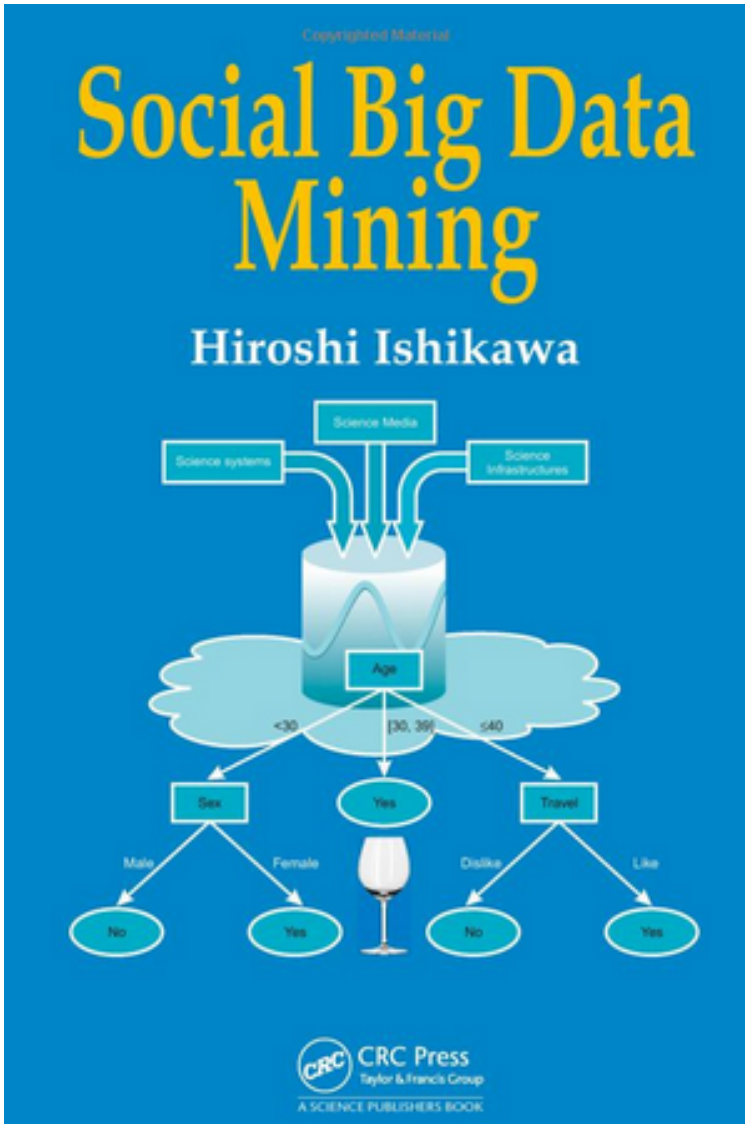
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

Architecture of Big Data Analytics

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

Social Big Data Mining

(Hiroshi Ishikawa, 2015)



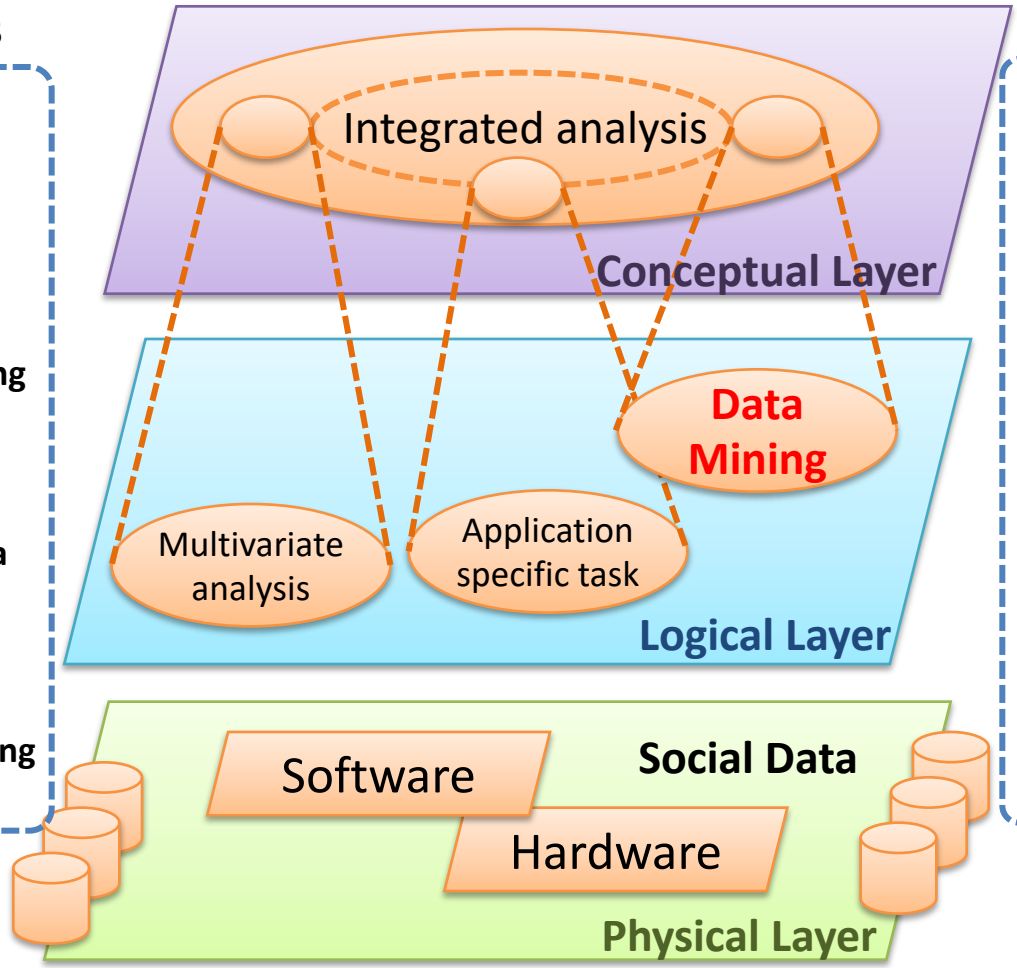
Source: http://www.amazon.com/Social-Data-Mining-Hiroshi-Ishikawa/dp/149871093X

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 distrusted processing

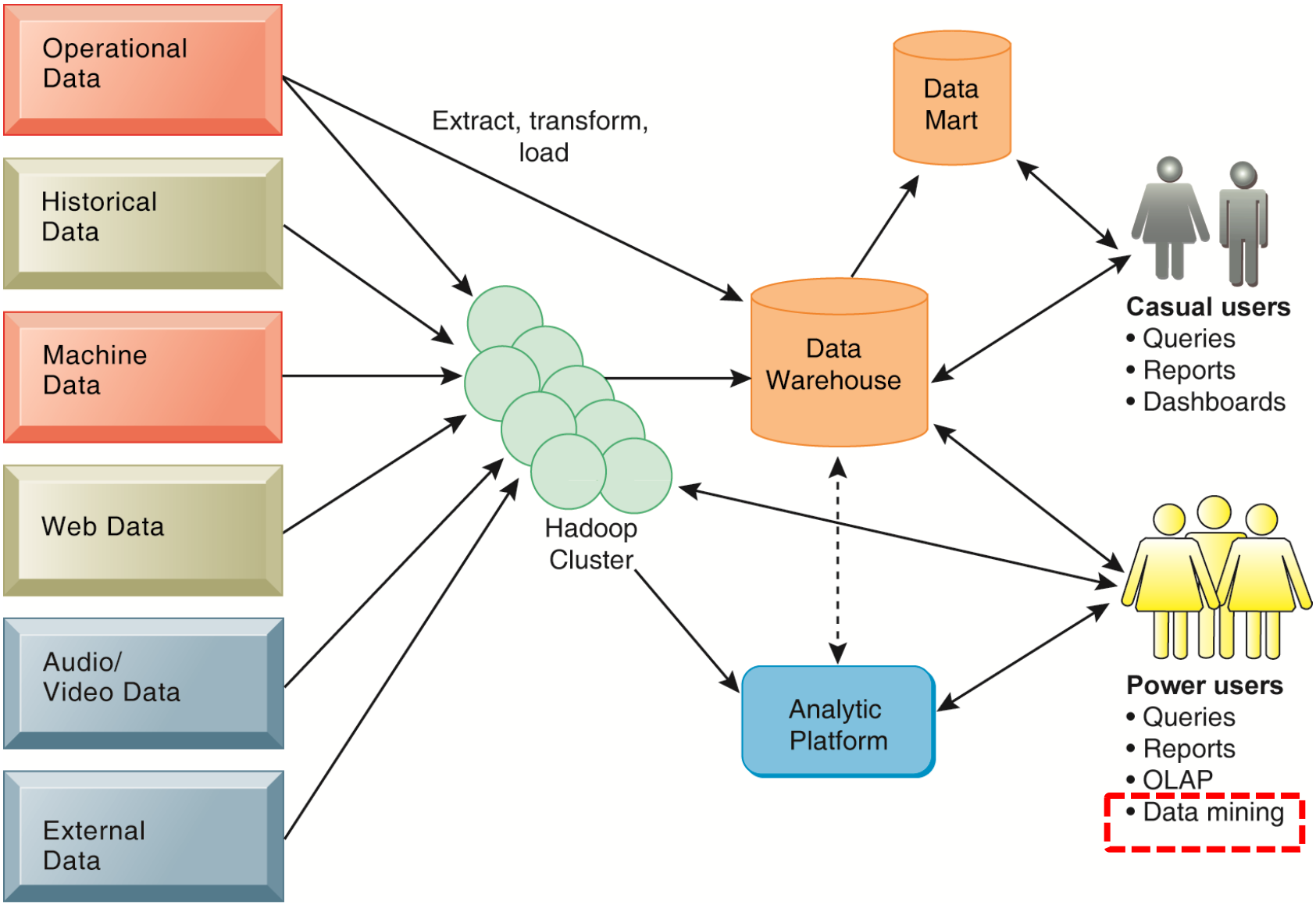


Analysts

- 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

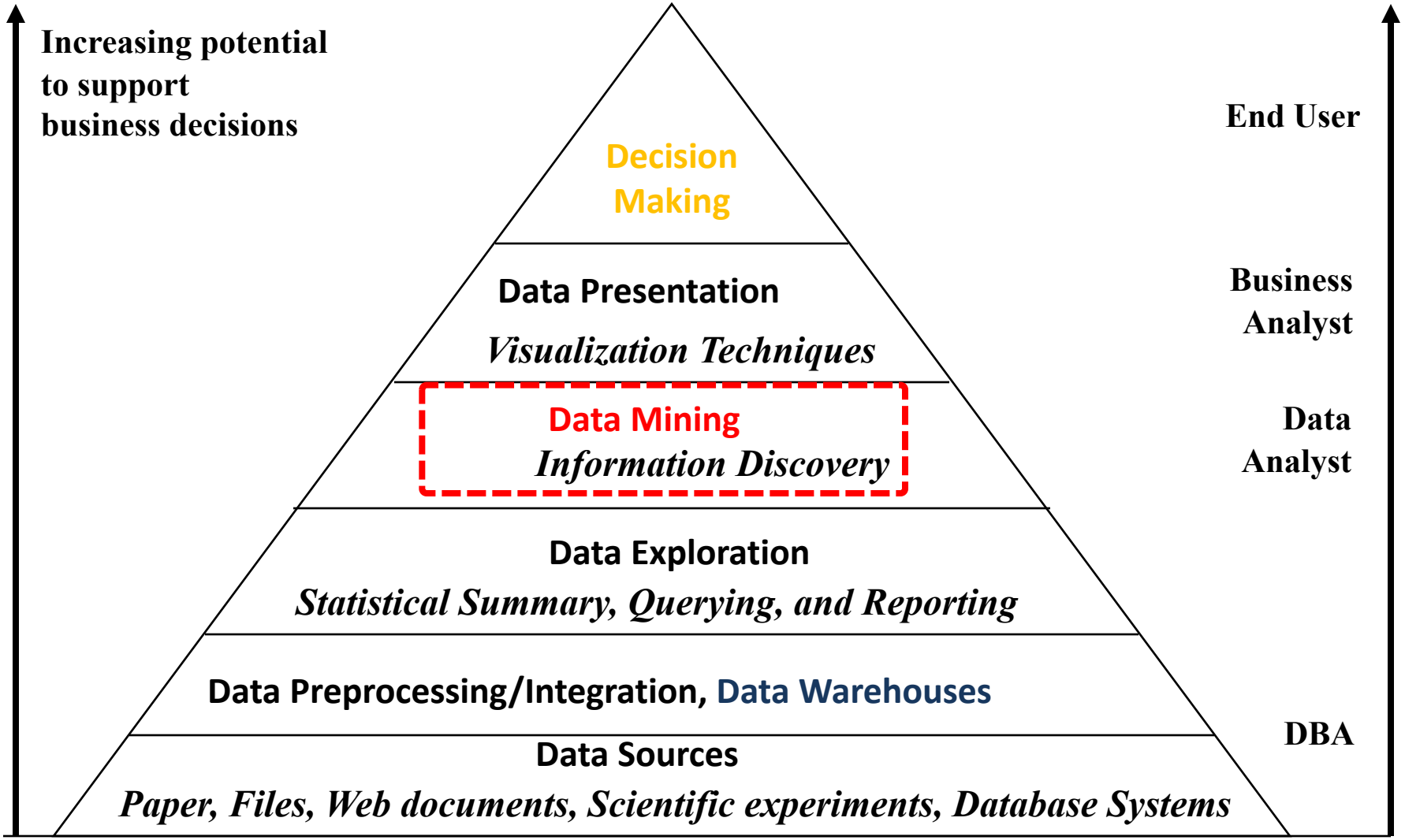
Business Intelligence (BI) Infrastructure



Source: Kenneth C. Laudon & Jane P. Laudon (2014), Management Information Systems: Managing the Digital Firm, Thirteenth Edition, Pearson.

Data Warehouse

Data Mining and Business Intelligence

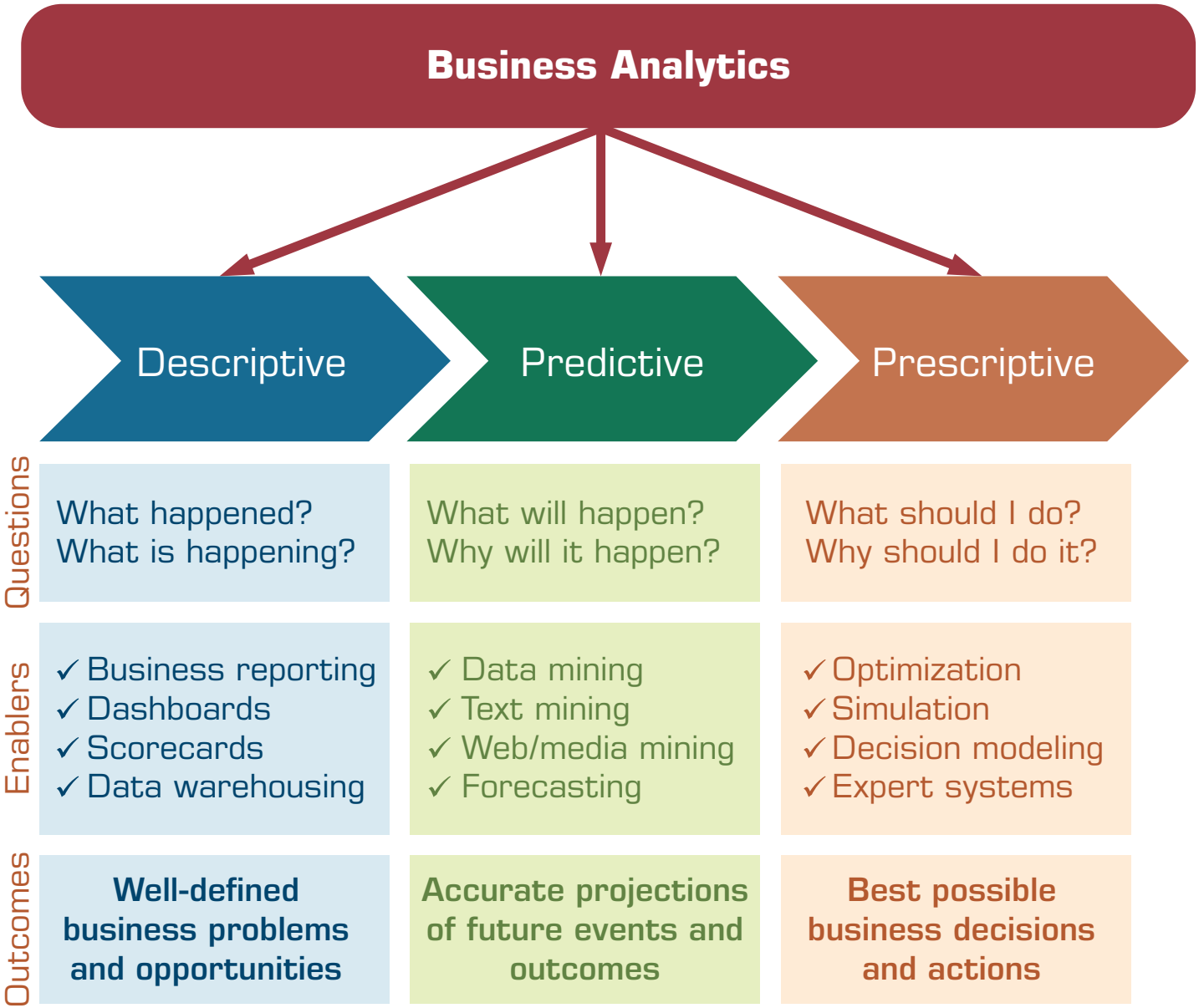


Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

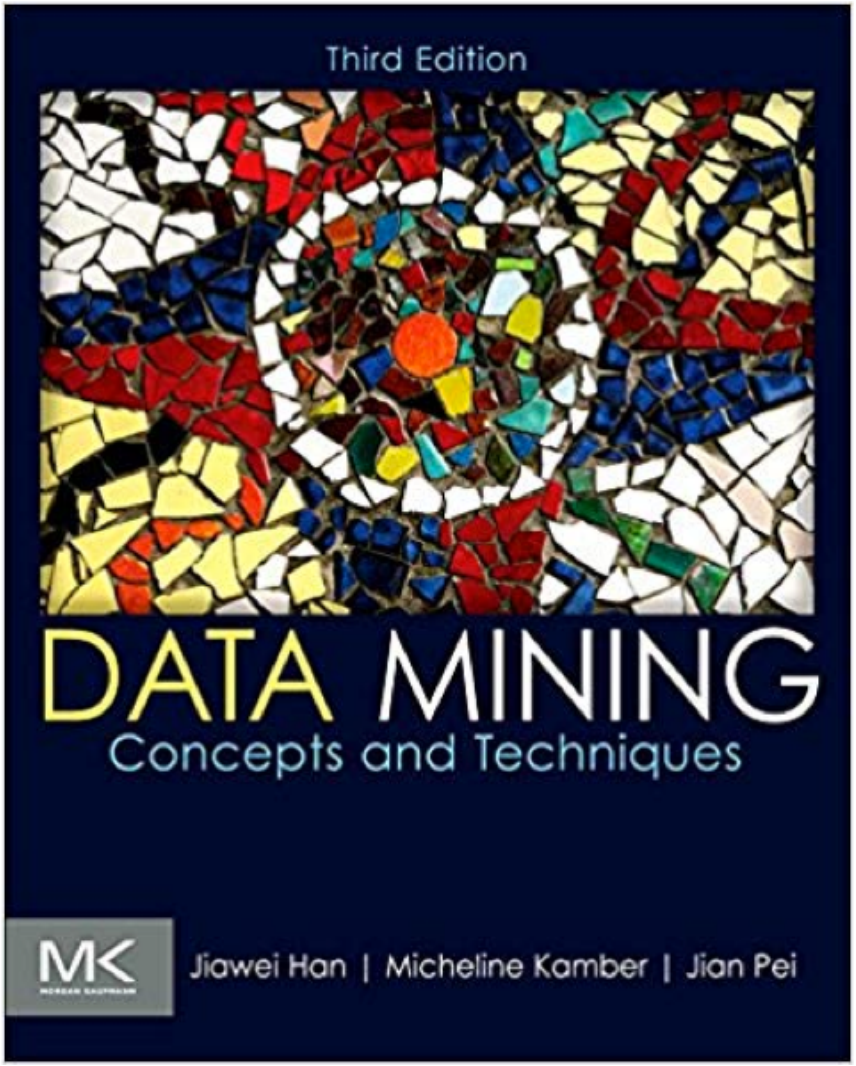
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Three Types of Analytics



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017),
Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

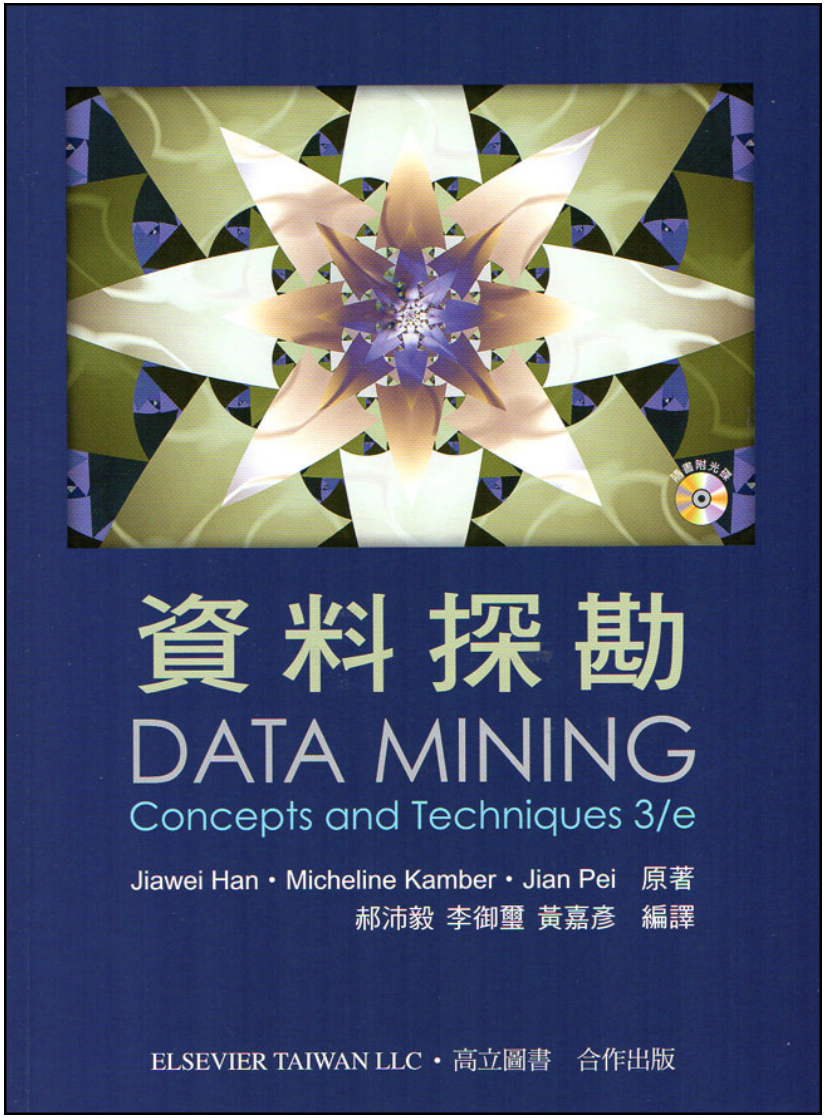
Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, MichelineKamber and Jian Pei, Morgan Kaufmann, 2011



Source: https://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/0123814790

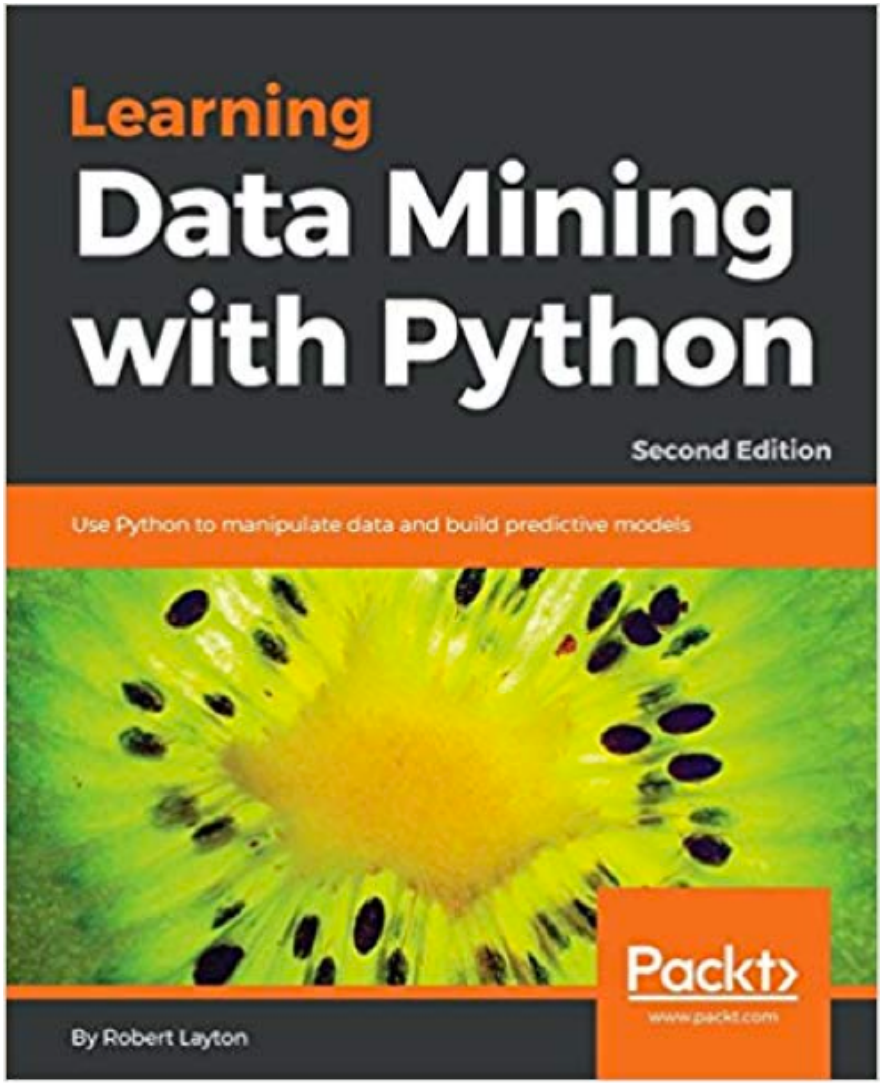
郝沛毅, 李御璽, 黃嘉彥 編譯, 資料探勘

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



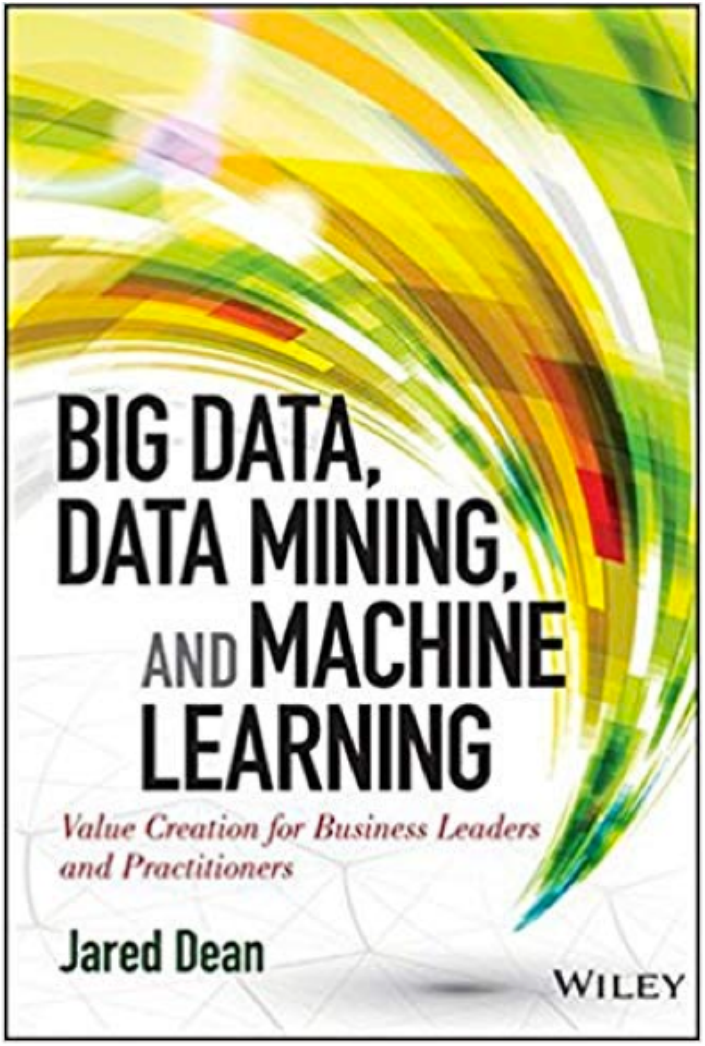
Source: http://www.books.com.tw/products/0010646676

Learning Data Mining with Python - Second Edition, Robert Layton, Packt Publishing, 2017



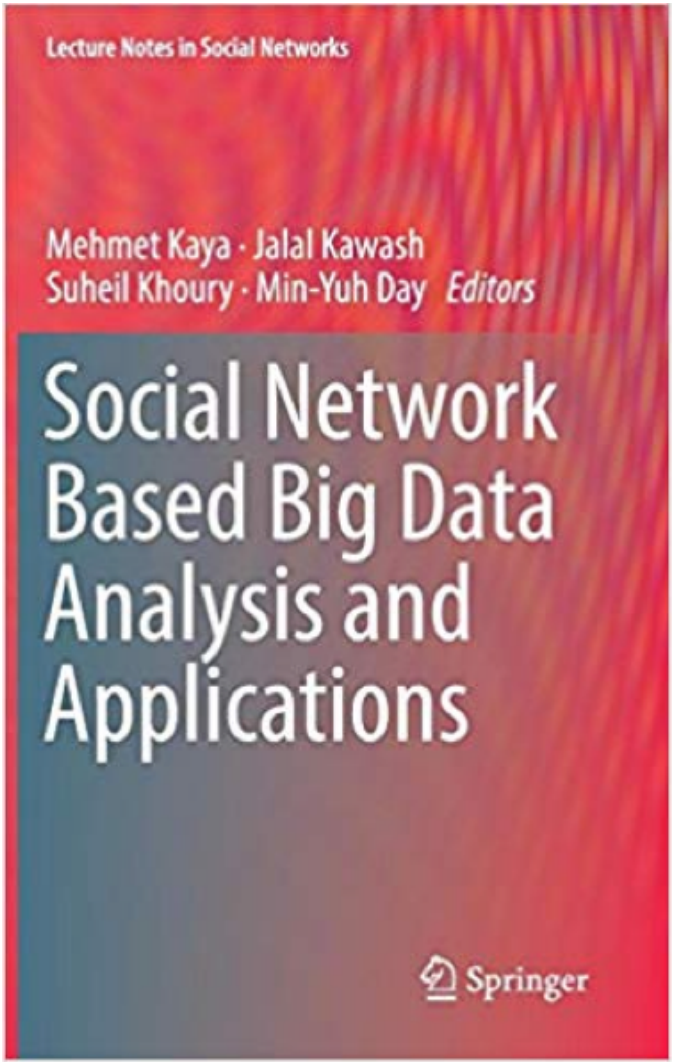
Source: https://www.amazon.com/Learning-Data-Mining-Python-Second/dp/1787126781

Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, Jared Dean, Wiley, 2014.



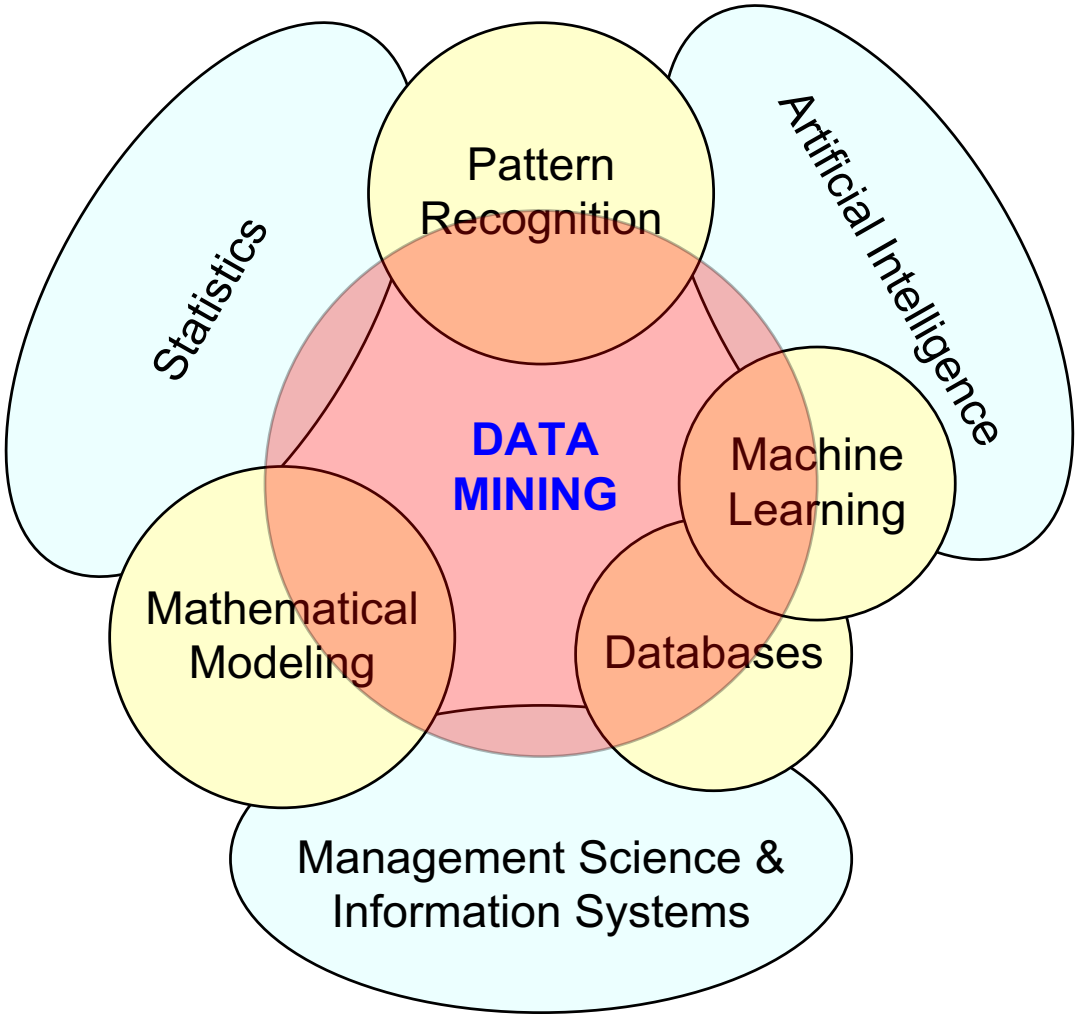
Source: https://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

Social Network Based Big Data Analysis and Applications, Lecture Notes in Social Networks, Mehmet Kaya, Jalal Kawash, Suheil Khoury, Min-Yuh Day, Springer International Publishing, 2018.



Source: https://www.amazon.com/Network-Analysis-Applications-Lecture-Networks/dp/3319781952

Data Mining at the Intersection of Many Disciplines



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

Data Mining:

Core Analytics Process

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Source: 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.

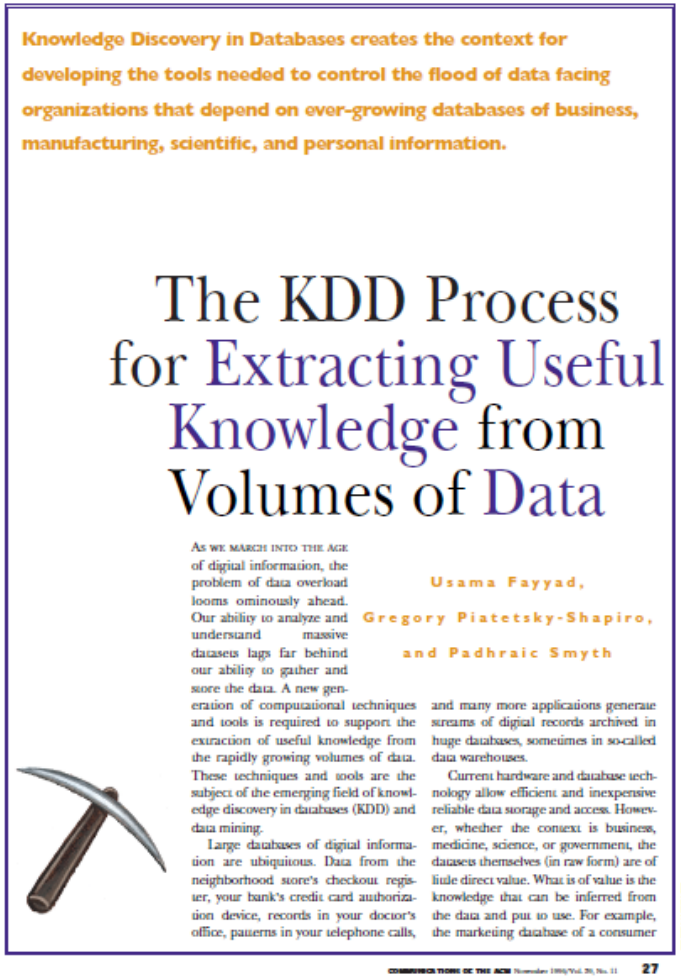
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

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

Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth

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.

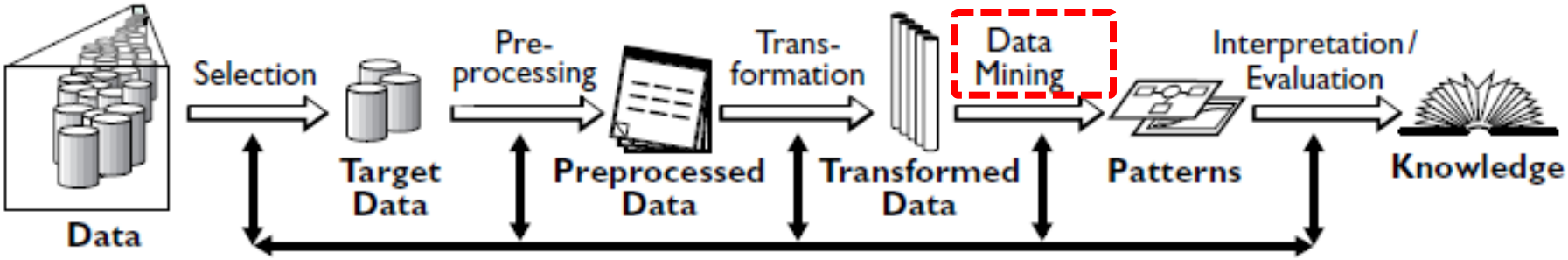
Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred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

TERRY WEEHER

Data Mining

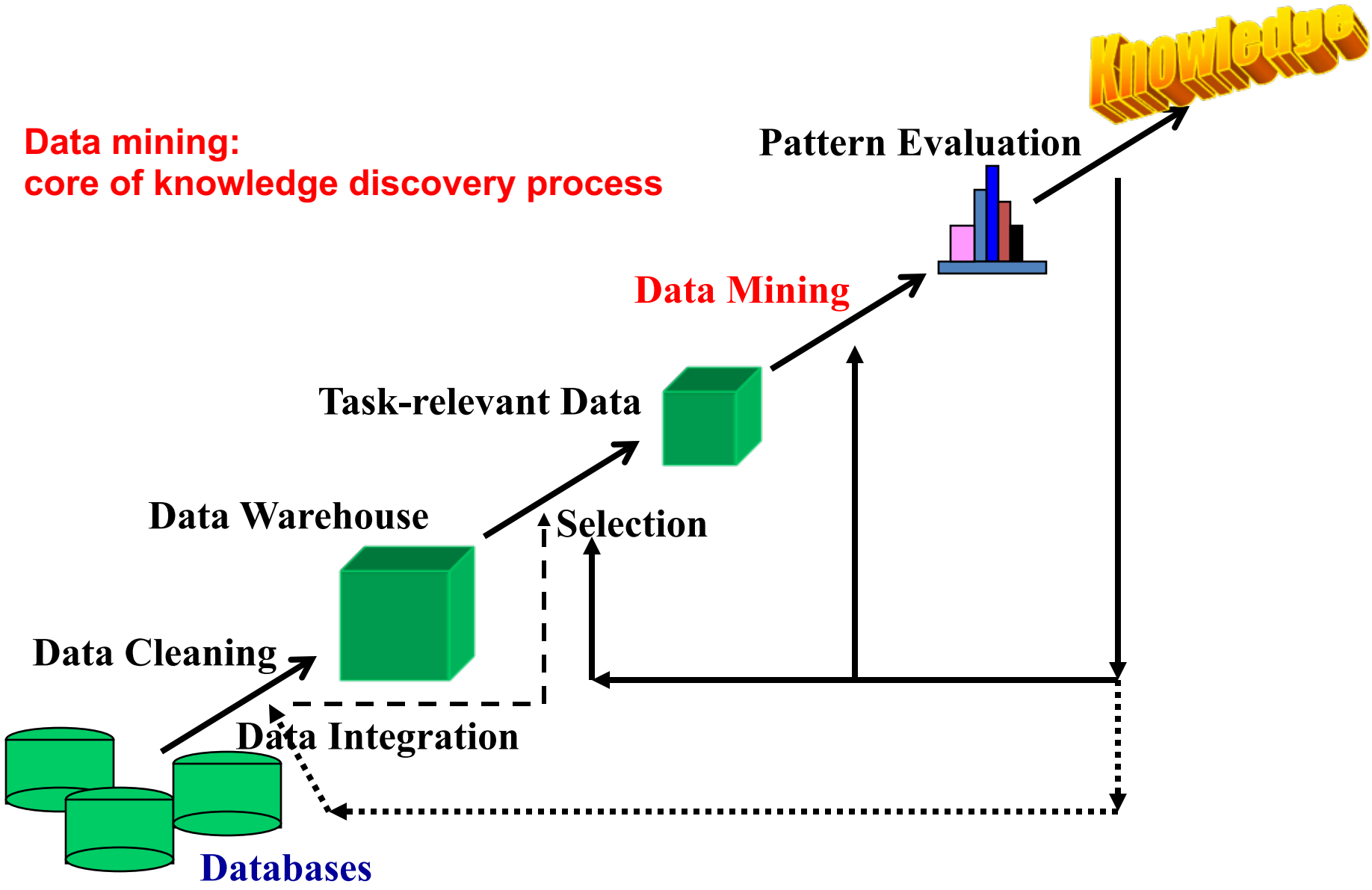
Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)



Source: 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.

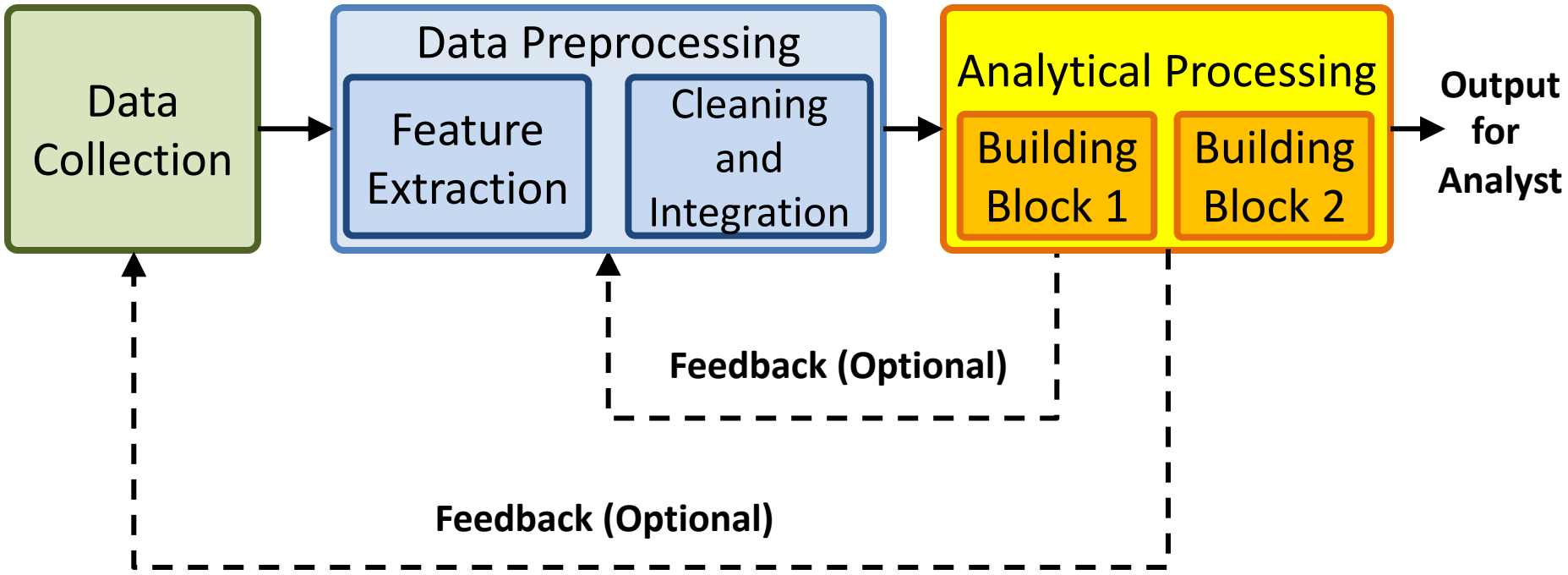
Knowledge Discovery (KDD) Process



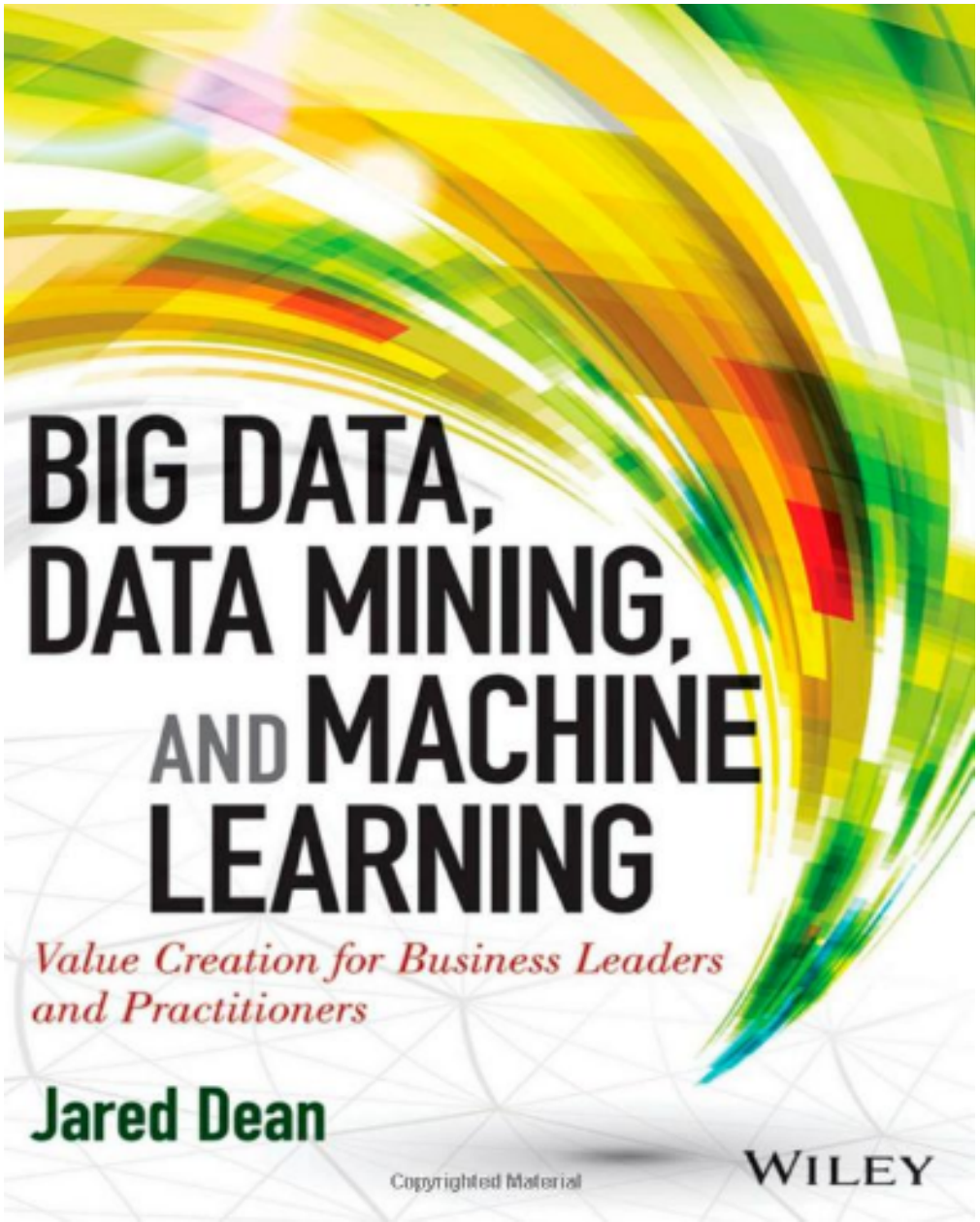
Source: Han & Kamber (2006)

Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

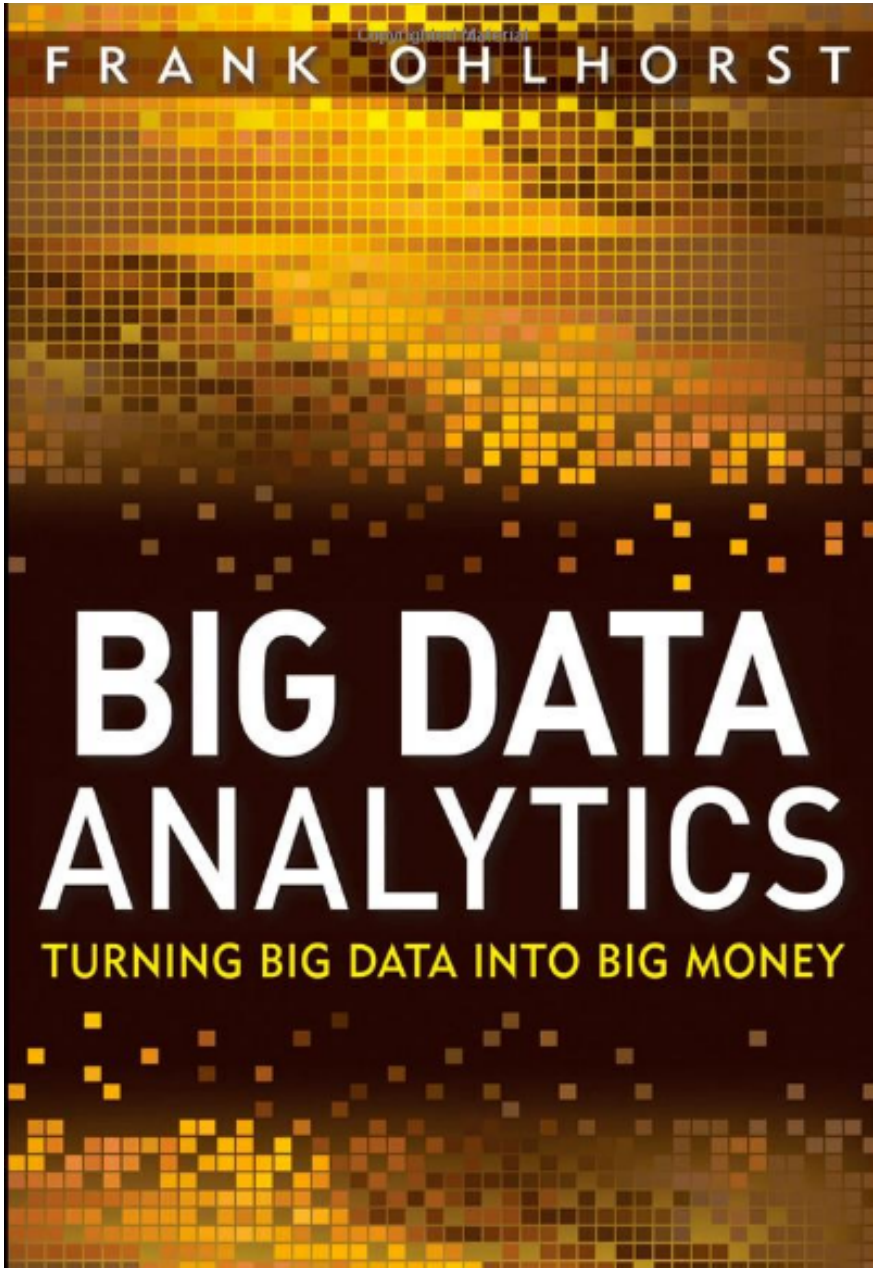


Source: http://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041

Deep Learning

Intelligence from Big Data

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



Source: http://www.amazon.com/Big-Data-Analytics-Turning-Money/dp/1118147596

Source: http://www.amazon.com/Big-Data-Revolution-Transform-Mayer-Schonberger/dp/B00D81X2YE

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
R²I²

DETECT
Anomalies
Communities
Typologies

PREDICT
Tending
Real-time
Prediction

DECIDE
Simulation
Optimization

BIG DATA – Batch

BIG DATA – Real Time

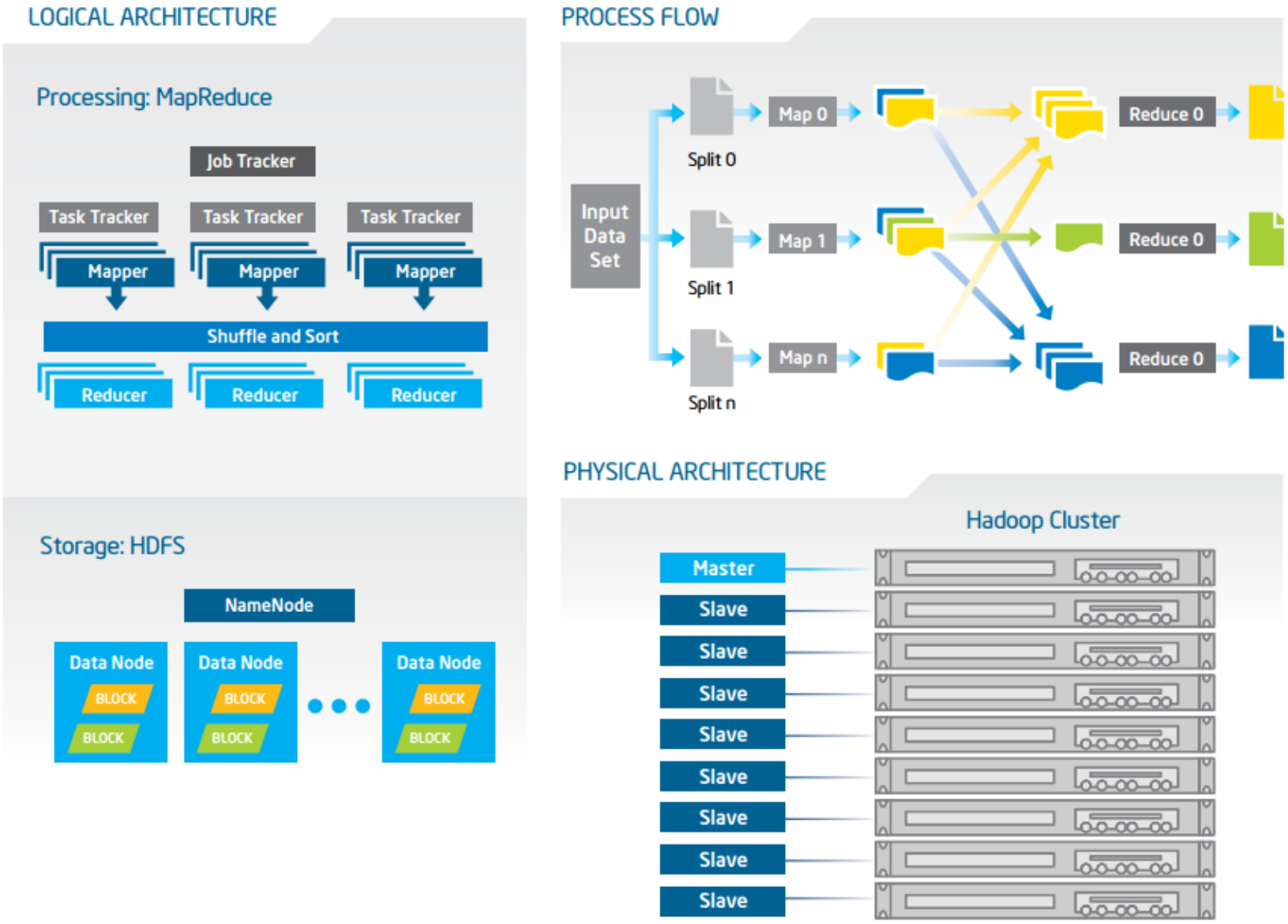
DATA

Complex by nature

Complex by structure

Source: https://www.thalesgroup.com/en/worldwide/big-data/big-data-big-analytics-visual-analytics-what-does-it-all-mean

Big Data with Hadoop Architecture

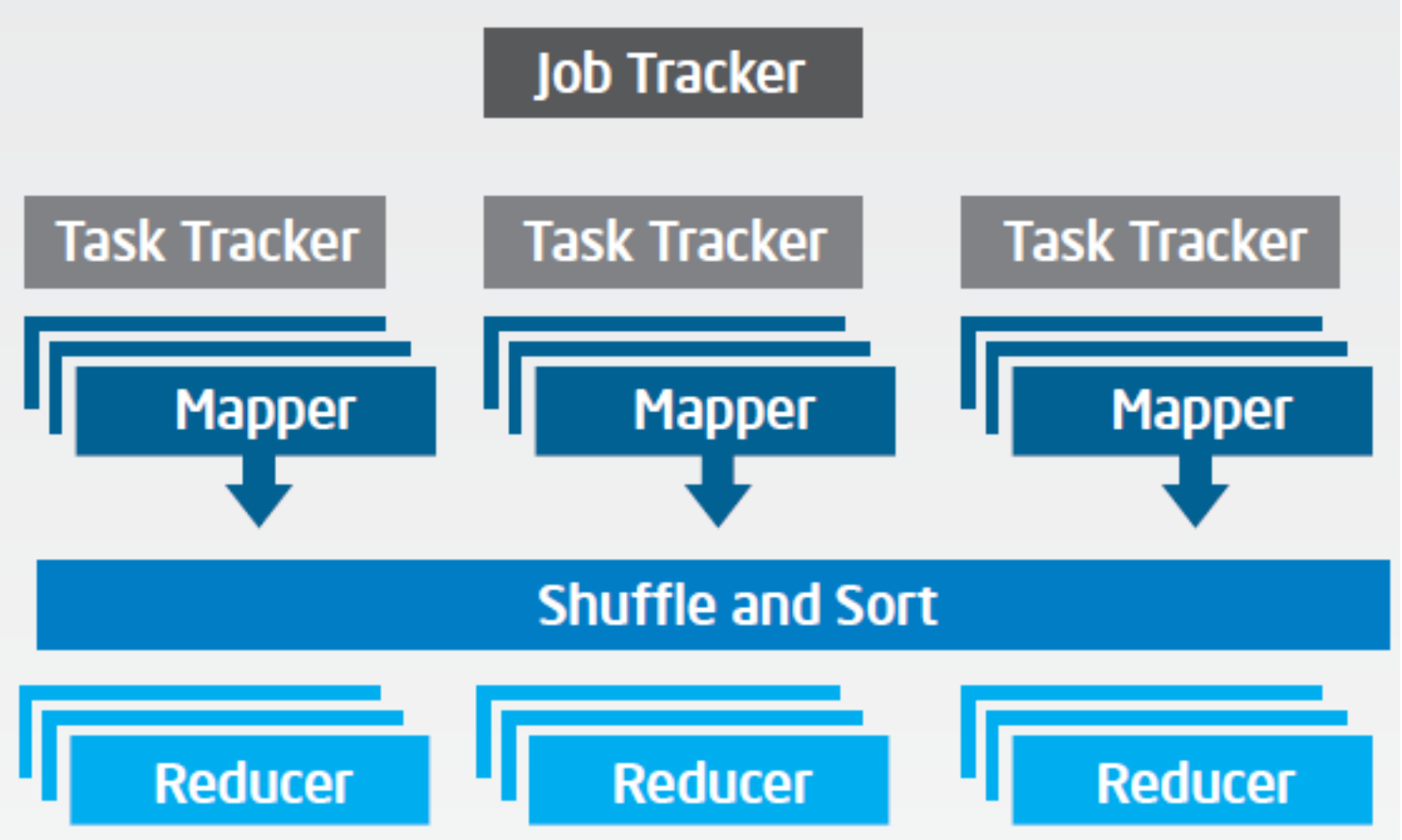


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

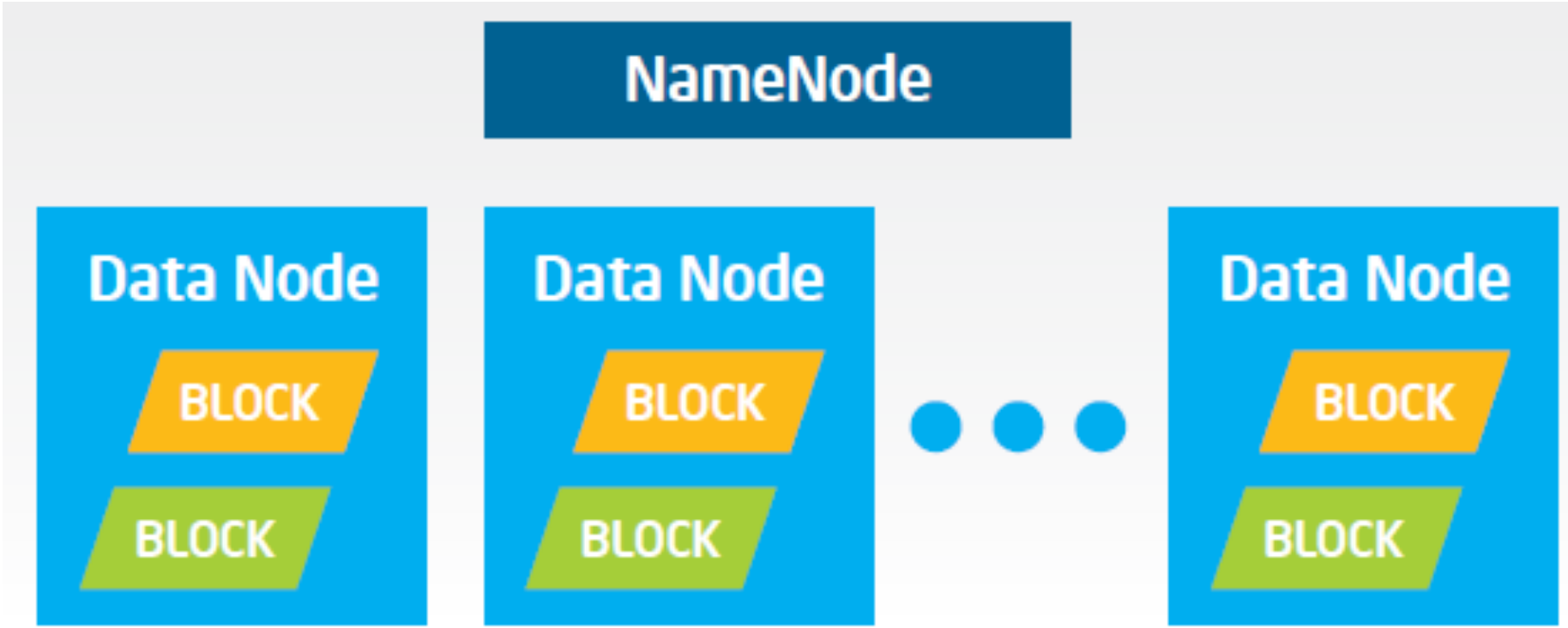


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

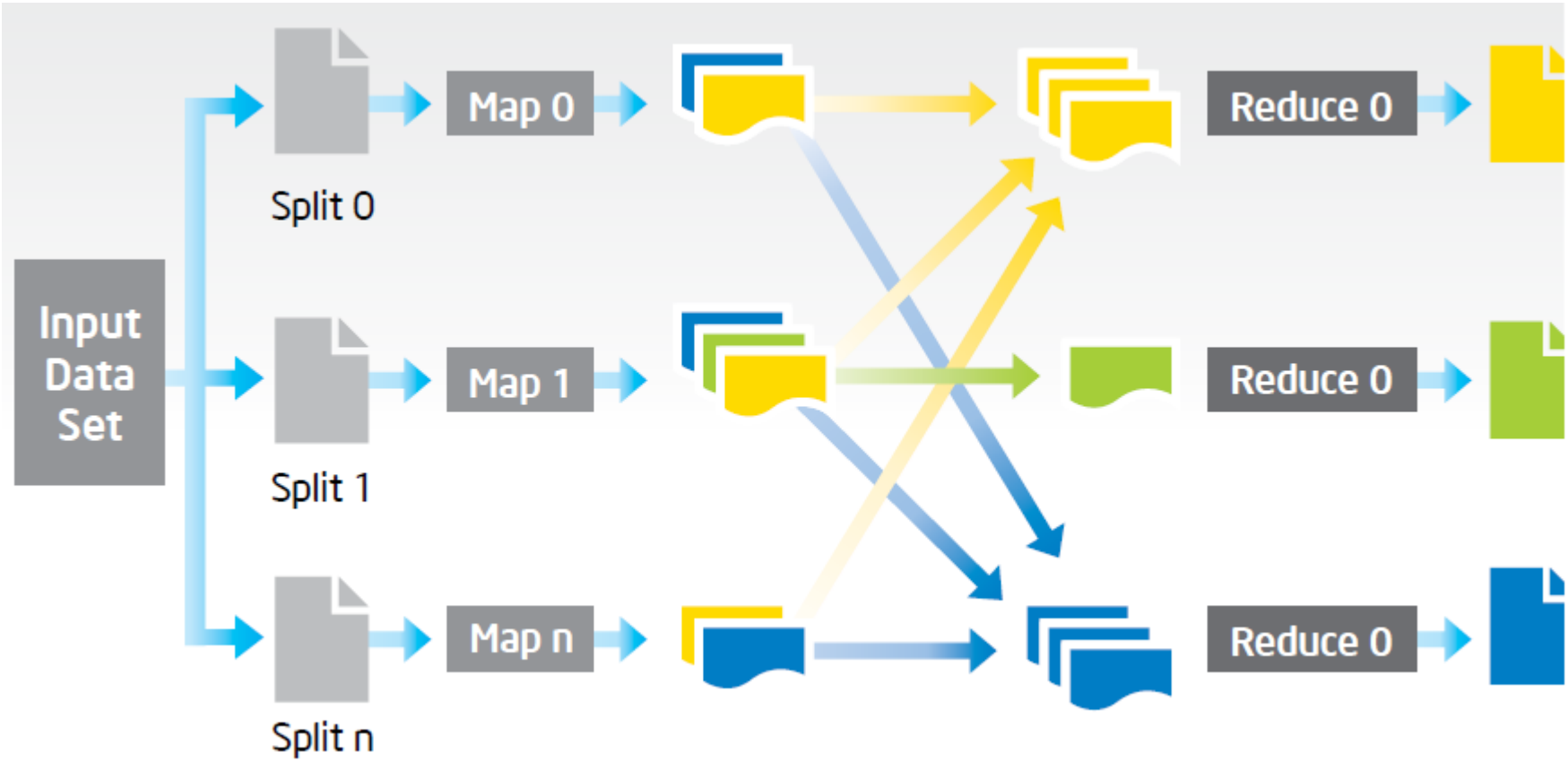
Storage: HDFS



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

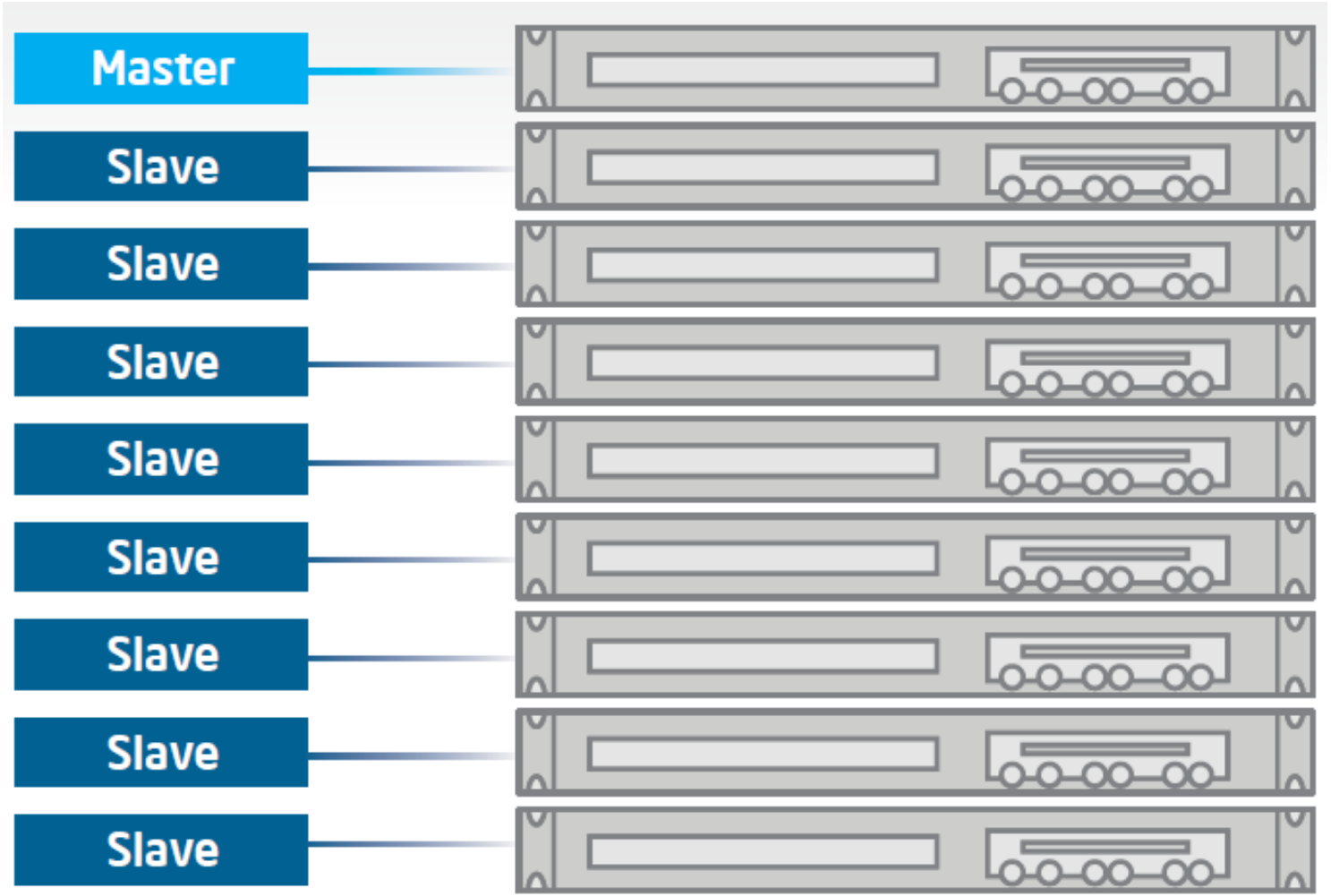
Process Flow



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

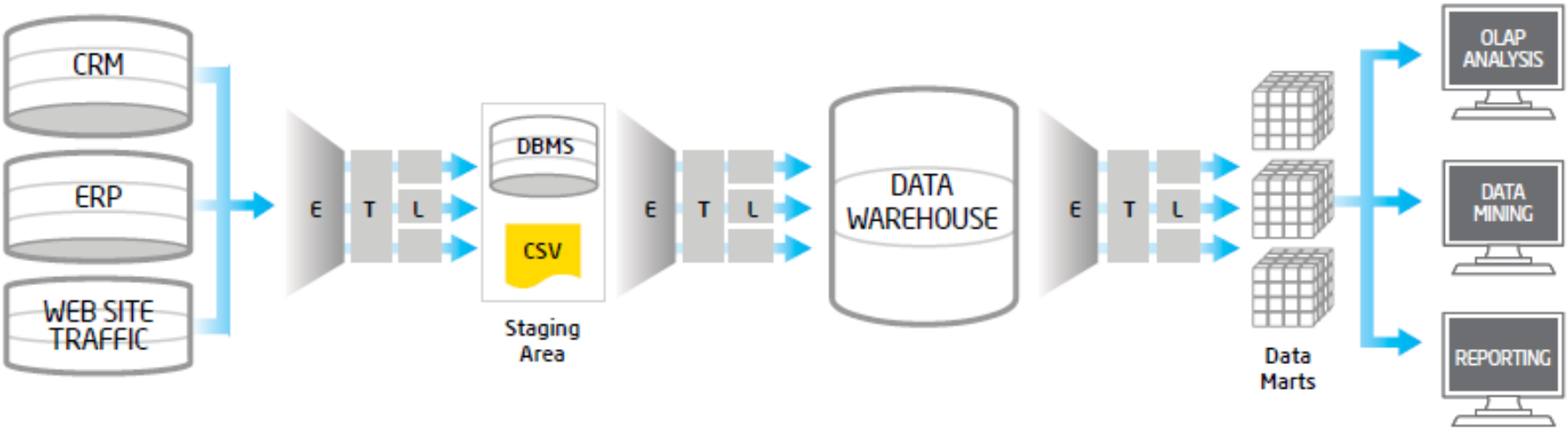
Big Data with Hadoop Architecture

Hadoop Cluster



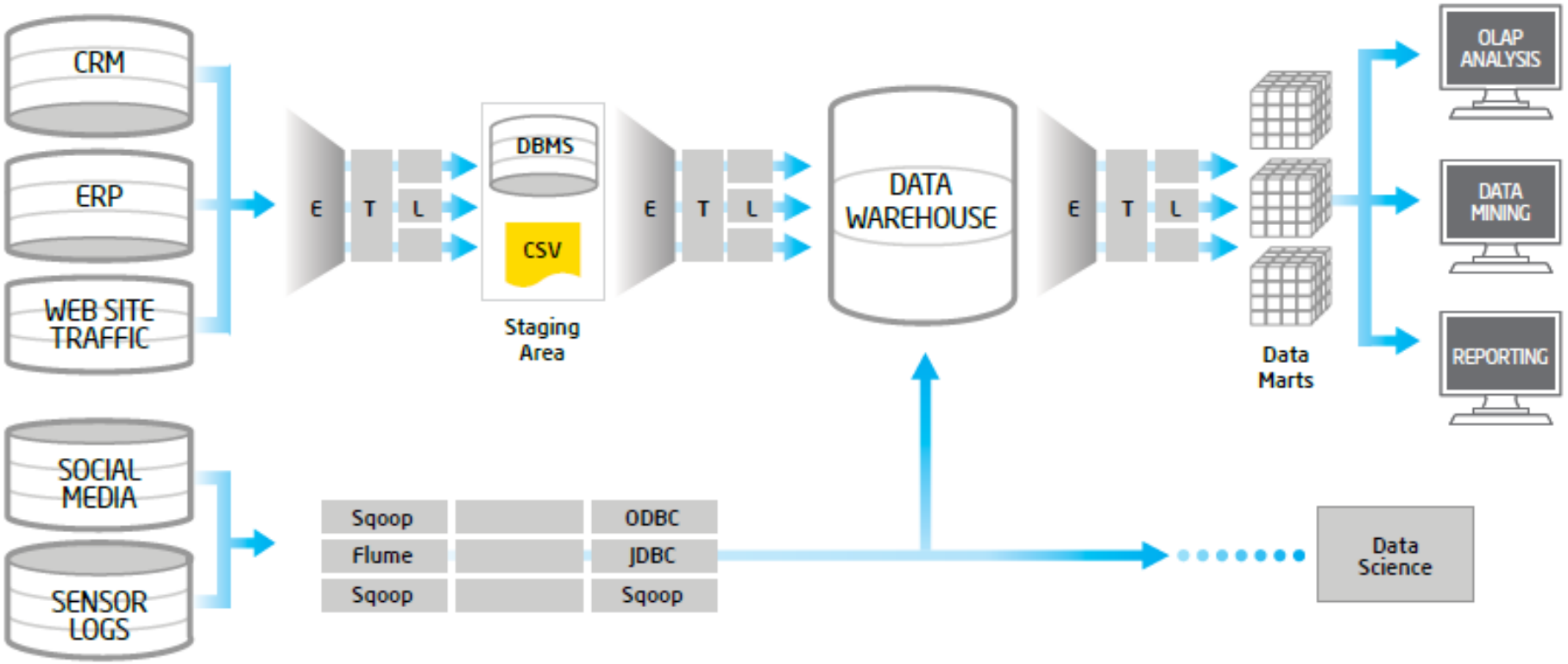
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Traditional ETL Architecture



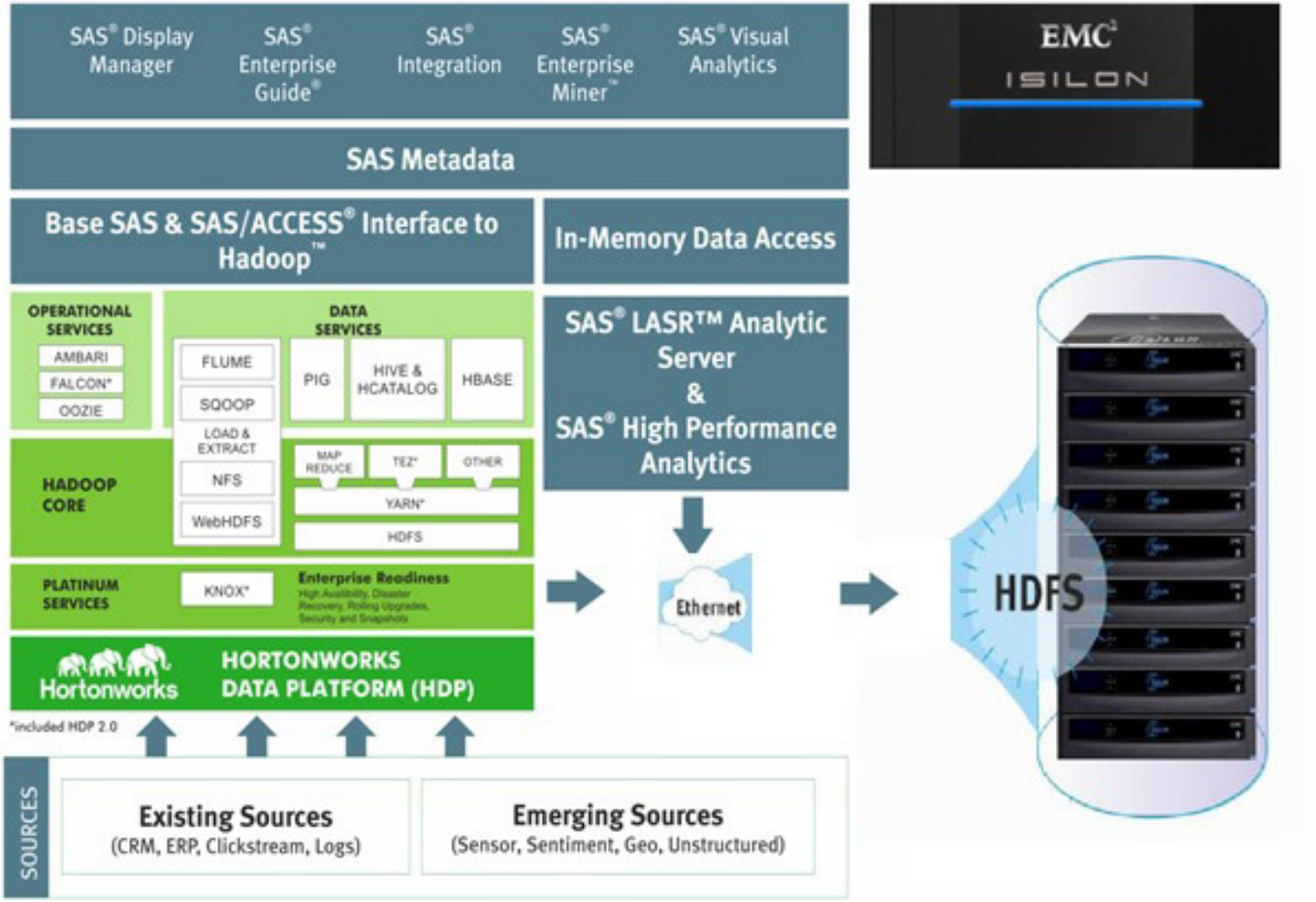
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Offload ETL with Hadoop (Big Data Architecture)



Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

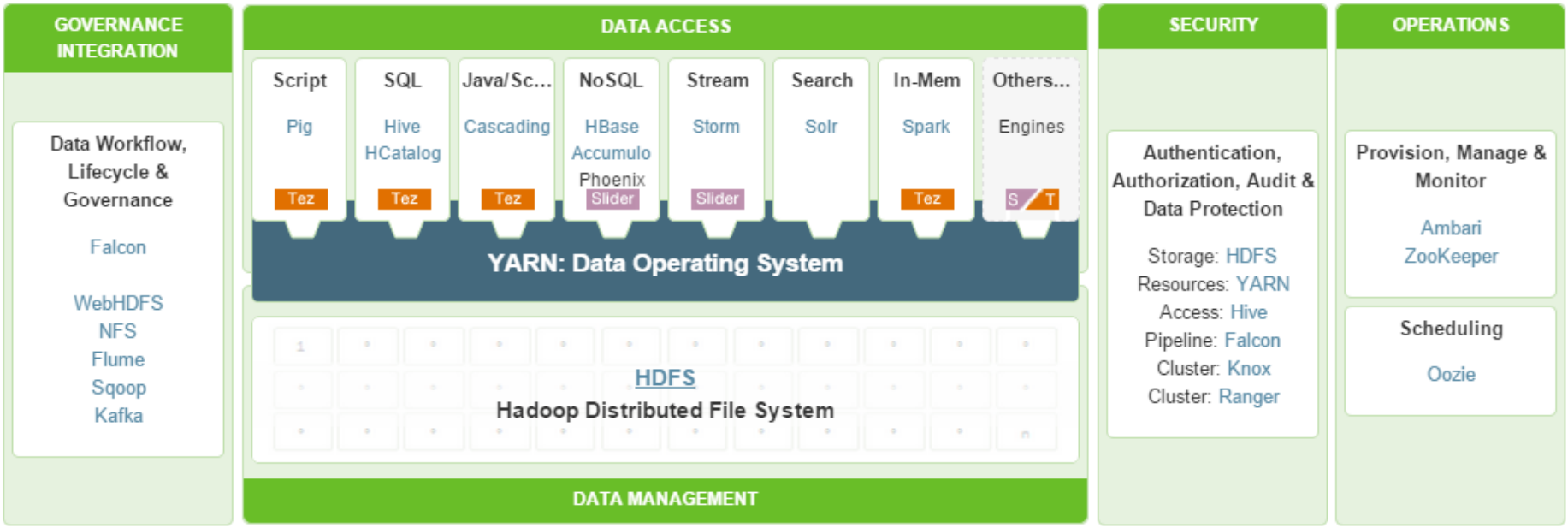
Big Data Solution



Source: http://www.newera-technologies.com/big-data-solution.html

HDP

A Complete Enterprise Hadoop Data Platform

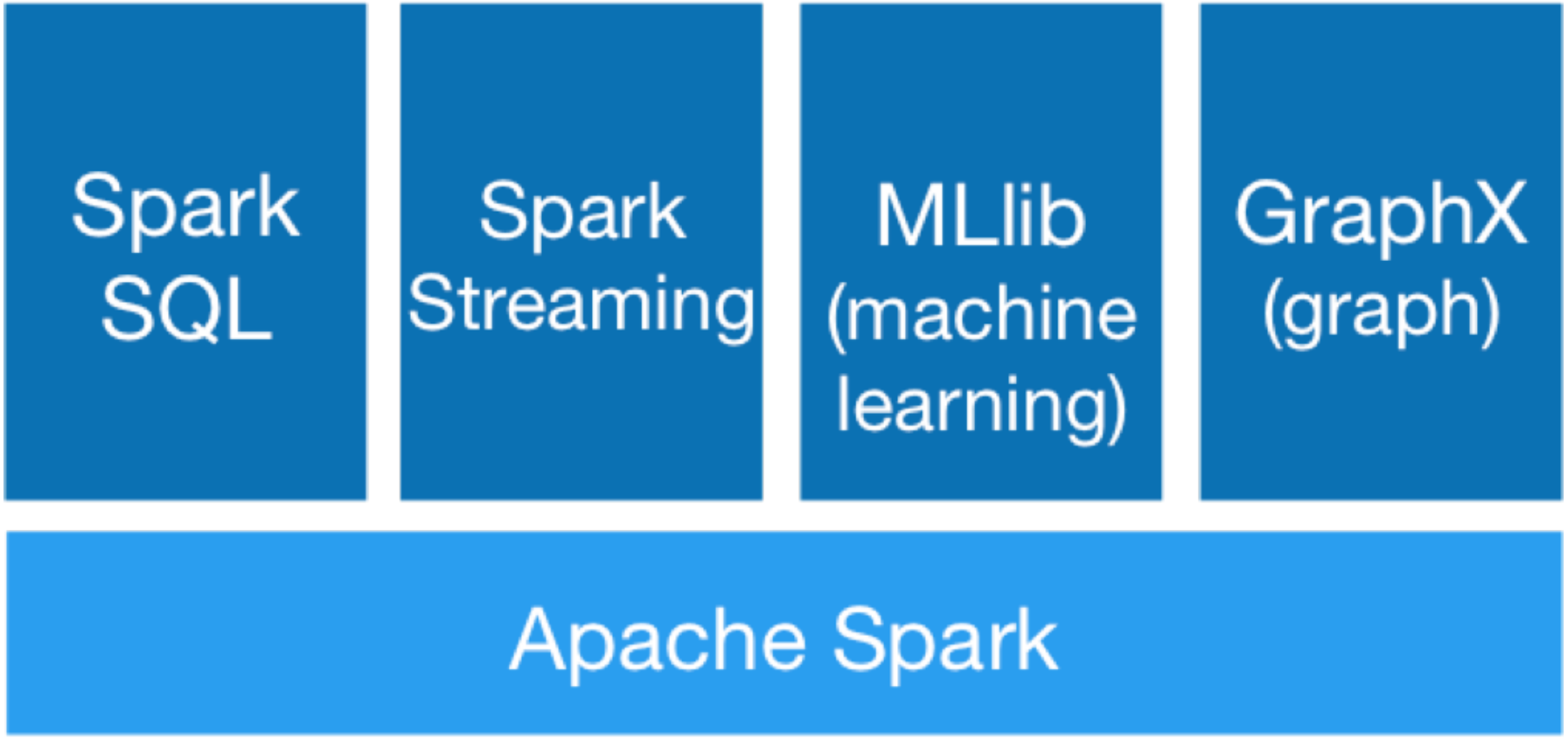


Source: http://hortonworks.com/hdp/

Spark and Hadoop

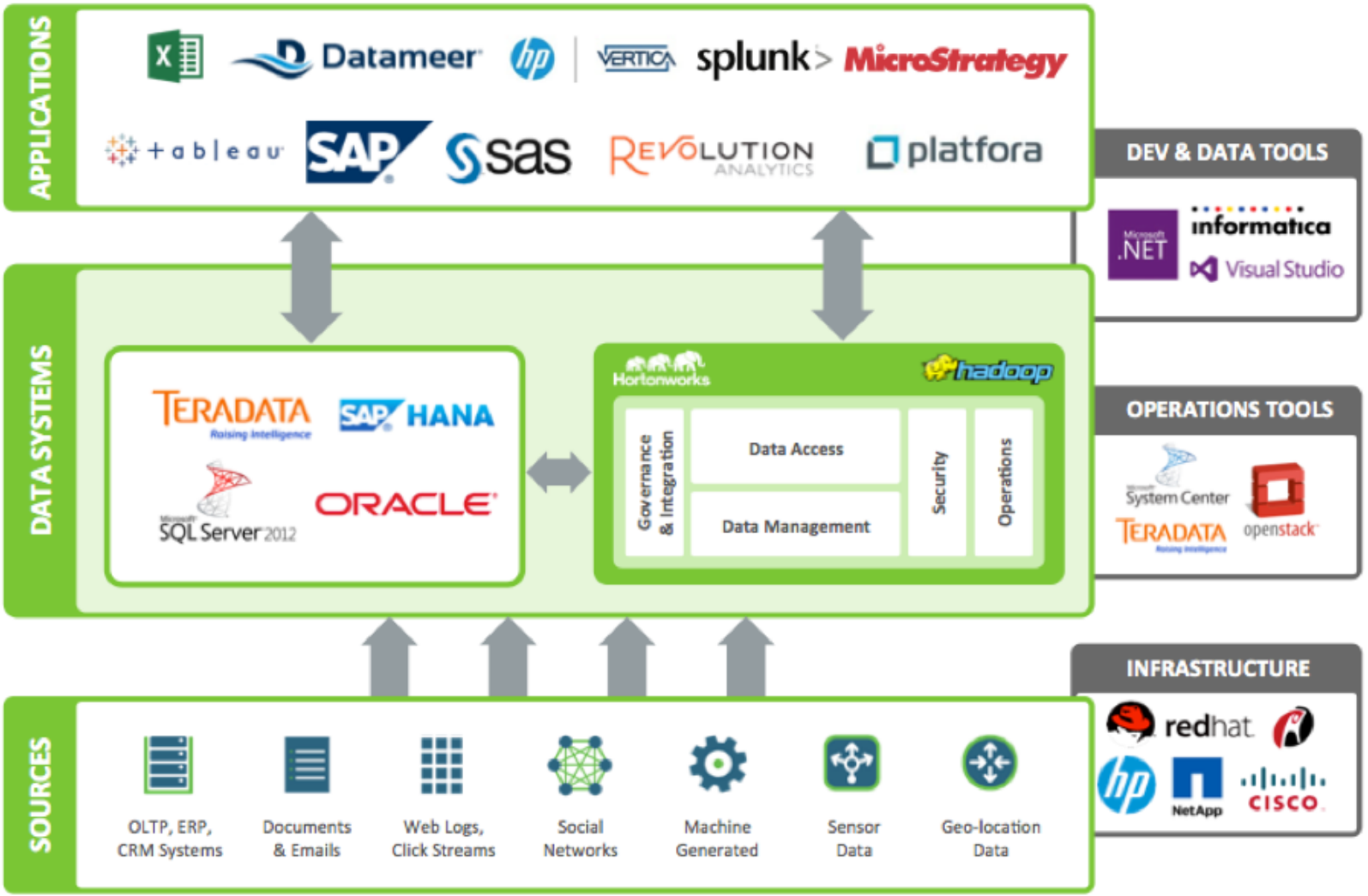
Source: http://spark.apache.org/

Spark Ecosystem



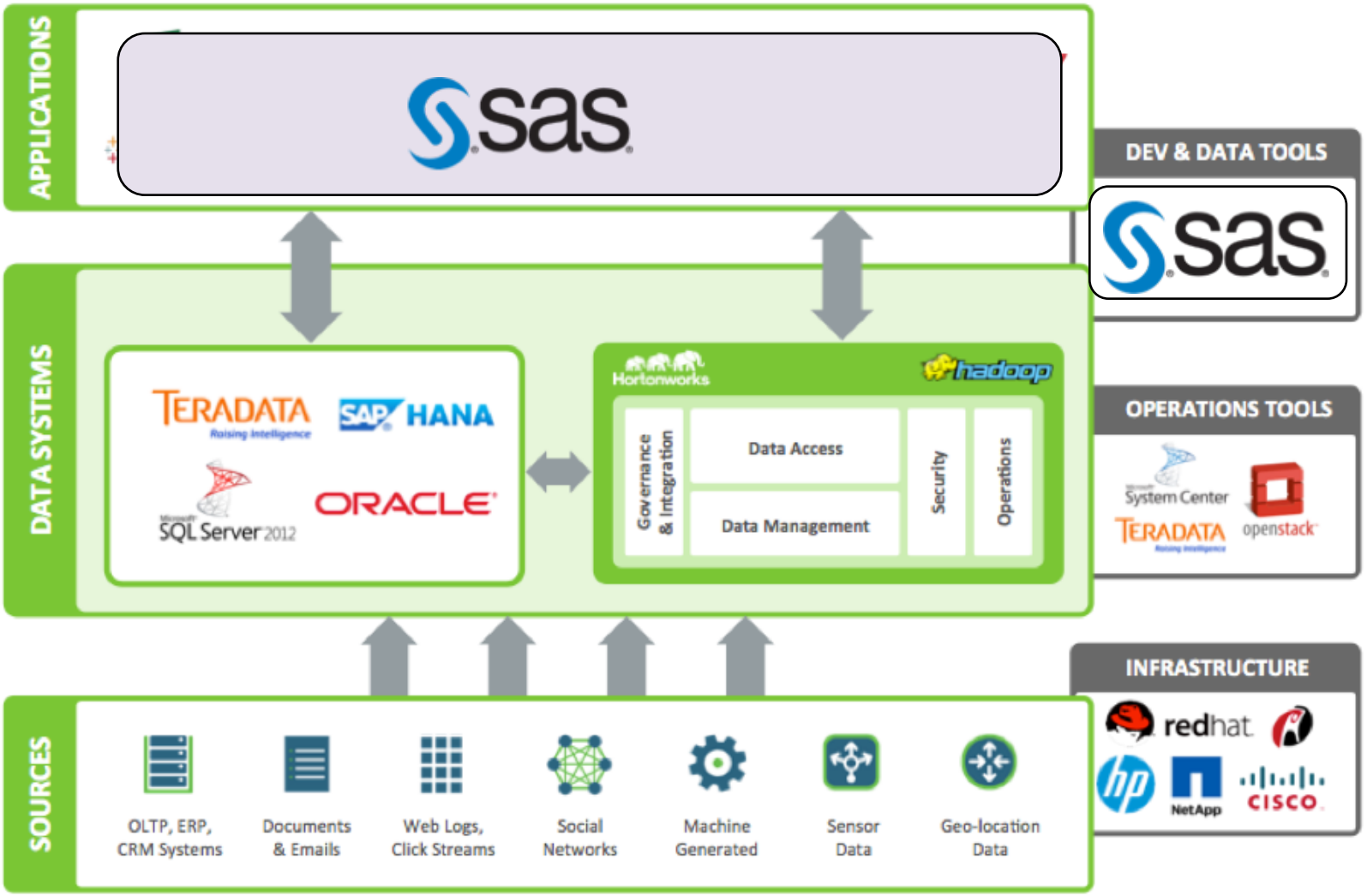
Source: http://spark.apache.org/

SAS Big data Strategy – SAS areas



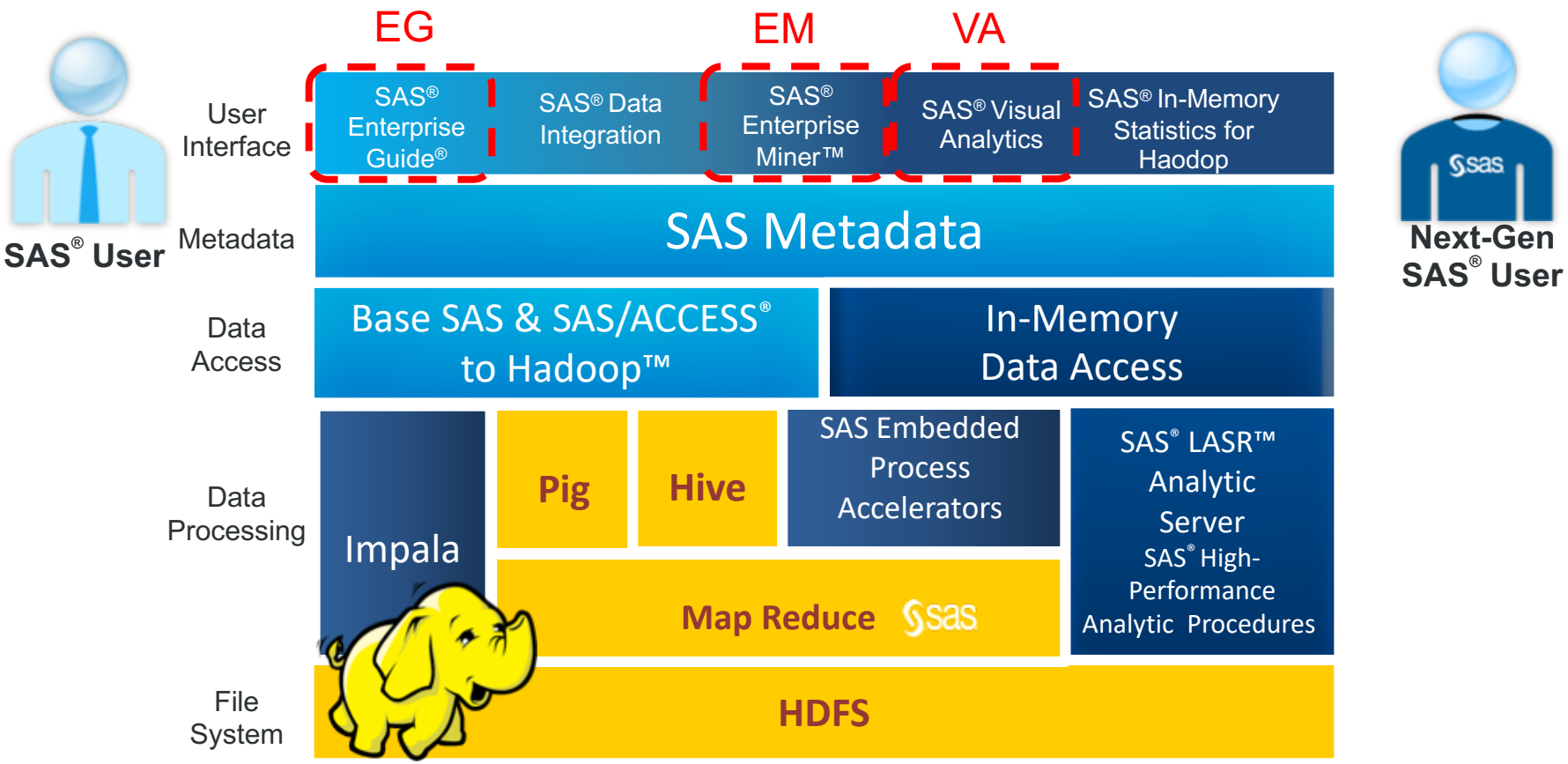
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

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

Business Intelligence Trends

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

Source: http://www.businessspectator.com.au/article/2013/1/22/technology/five-business-intelligence-trends-2013

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

Source: Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17

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

Source: Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17

Summary

- This course introduces the fundamental concepts and applications technology of big data mining.
- Topics include
  - Big Data Mining
  - Artificial Intelligence and Big Data Analytics
  - Association Analysis
  - Classification and Prediction
  - Cluster Analysis
  - Machine Learning and Deep Learning
  - Data Mining Using SAS Enterprise Miner (SAS EM)
  - Case Study and Implementation of Big Data Mining

Contact Information

戴敏育 博士 (Min-Yuh Day, Ph.D.)

副教授
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/