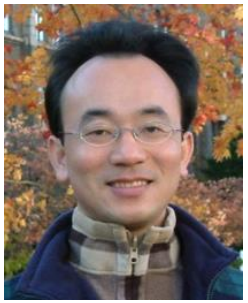
Tamkang University
淡江大學

Tamkang University

Big Data Mining
巨量資料探勘

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

1042DM02
MI4 (M2244) (3094)
Tue, 3, 4 (10:10-12:00) (B216)



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/

2016-02-23

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2016/02/16	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
2	2016/02/23	巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統 (Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)
3	2016/03/01	關連分析 (Association Analysis)
4	2016/03/08	分類與預測 (Classification and Prediction)
5	2016/03/15	分群分析 (Cluster Analysis)
6	2016/03/22	個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis – K-Means using SAS EM)
7	2016/03/29	個案分析與實作二 (SAS EM 關連分析)：Case Study 2 (Association Analysis using SAS EM)

課程大綱 (Syllabus)

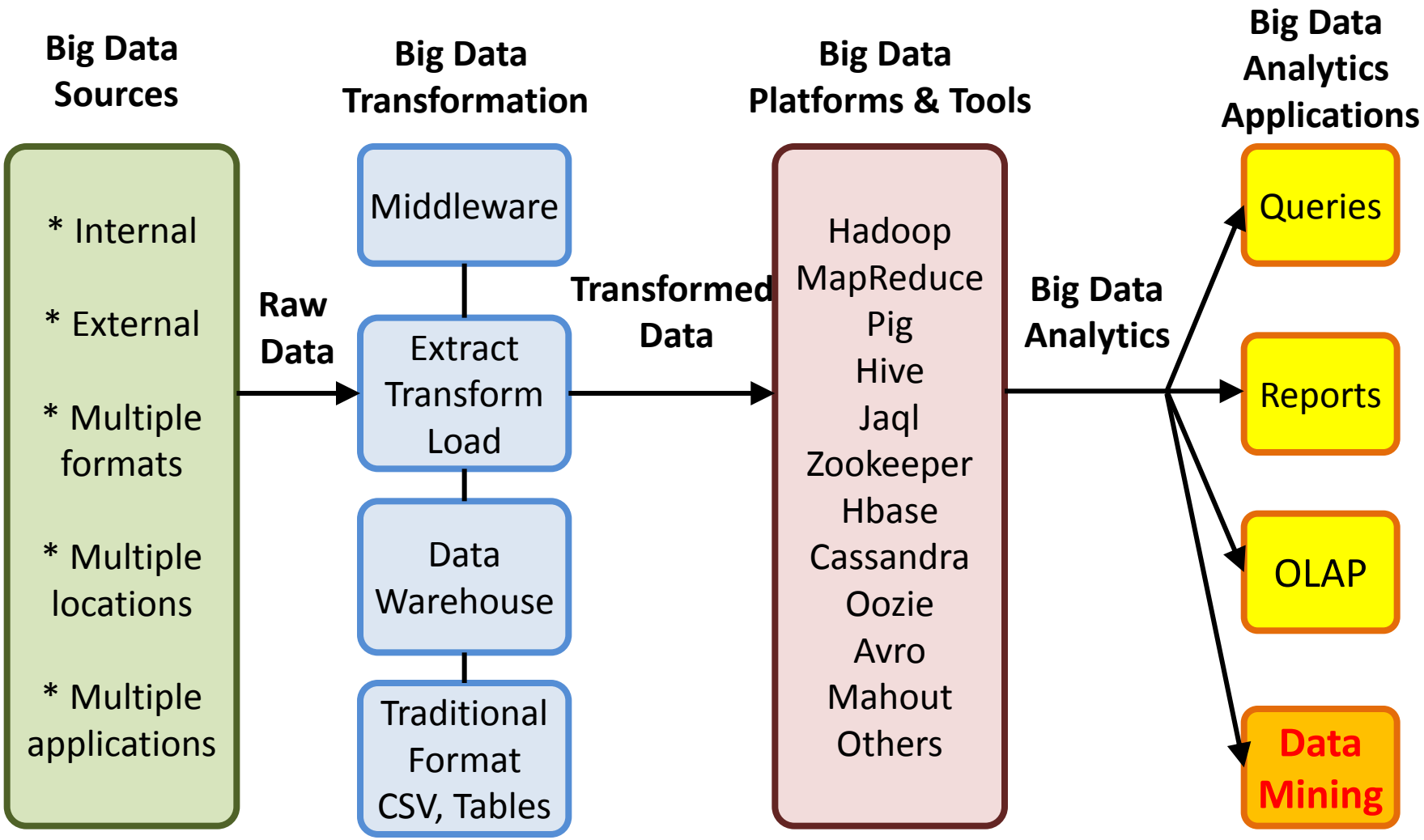
週次 (Week)	日期 (Date)	內容 (Subject/Topics)
8	2016/04/05	教學行政觀摩日 (Off-campus study)
9	2016/04/12	期中報告 (Midterm Project Presentation)
10	2016/04/19	期中考試週 (Midterm Exam)
11	2016/04/26	個案分析與實作三 (SAS EM 決策樹、模型評估)：Case Study 3 (Decision Tree, Model Evaluation using SAS EM)
12	2016/05/03	個案分析與實作四 (SAS EM 迴歸分析、類神經網路)：Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)
13	2016/05/10	Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)
14	2016/05/17	期末報告 (Final Project Presentation)
15	2016/05/24	畢業班考試 (Final Exam)

2016/02/23

巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統

(Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)

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

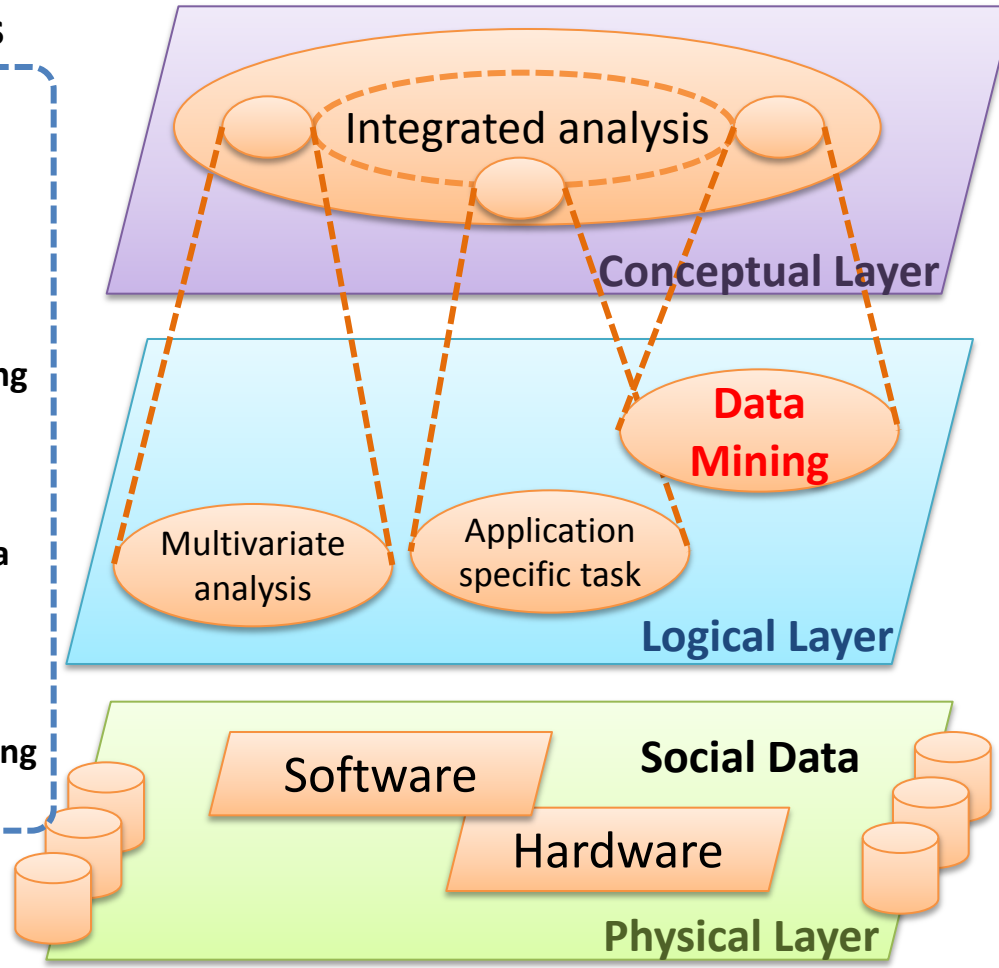
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

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**

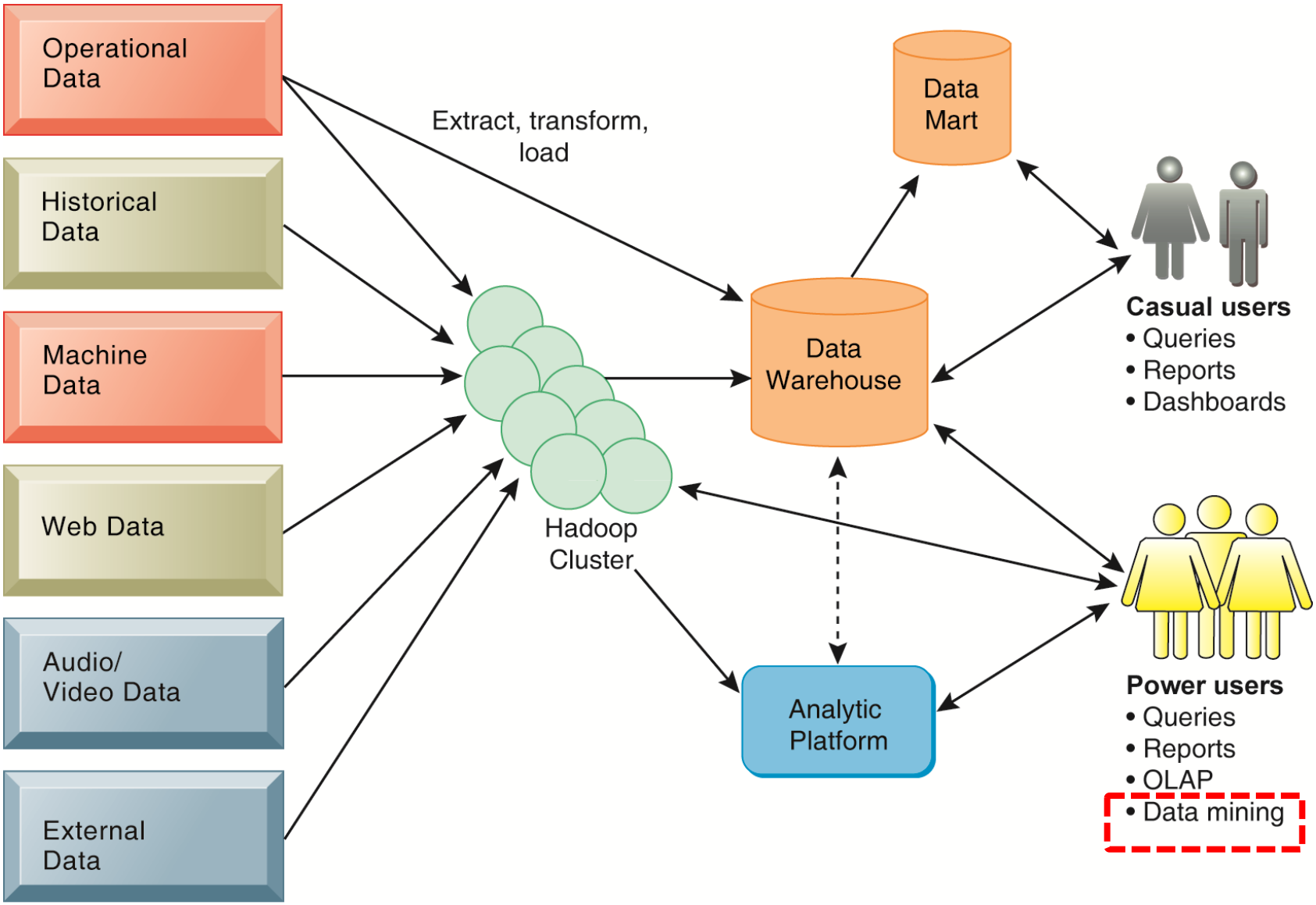


Analysts

- **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

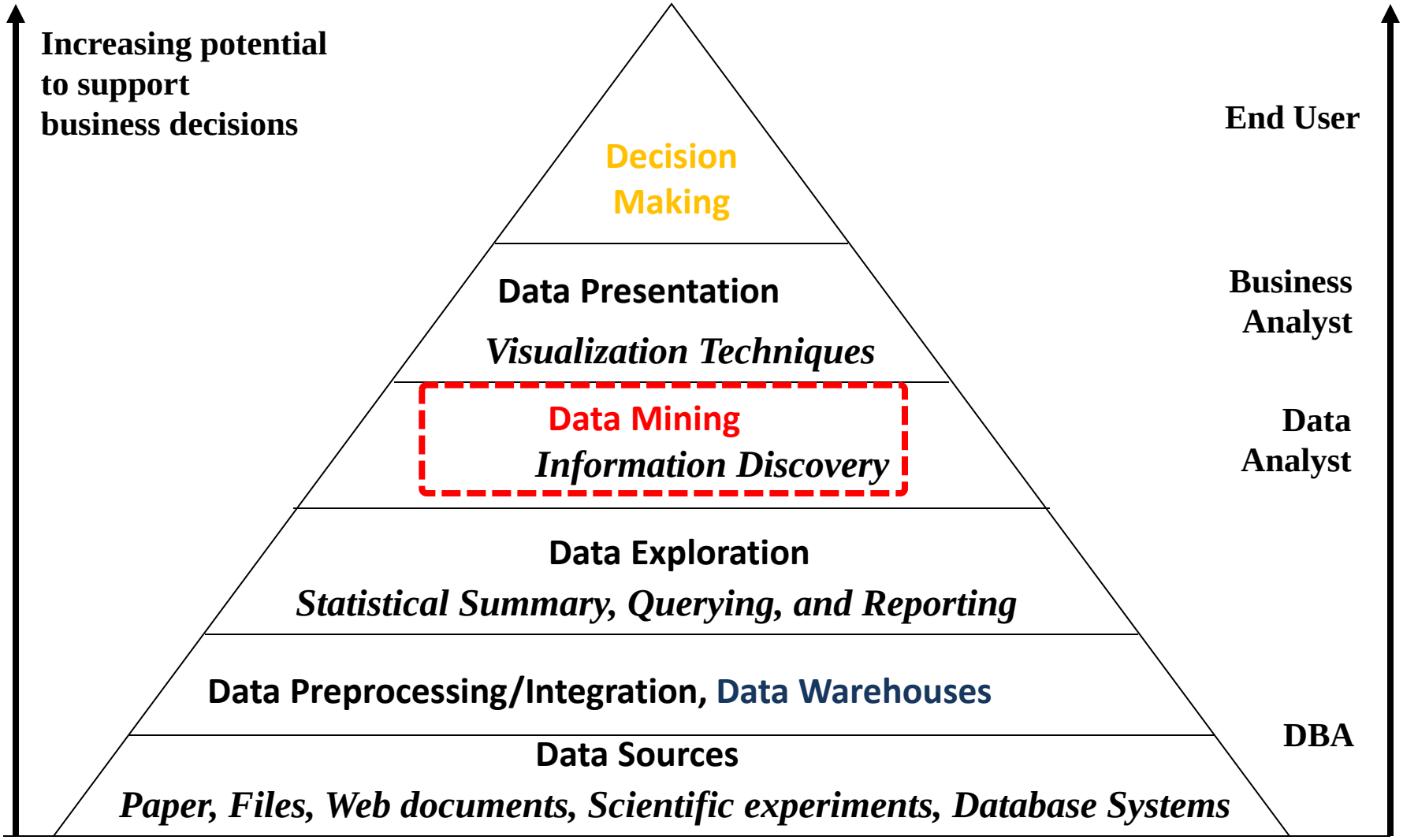
Business Intelligence (BI) Infrastructure



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

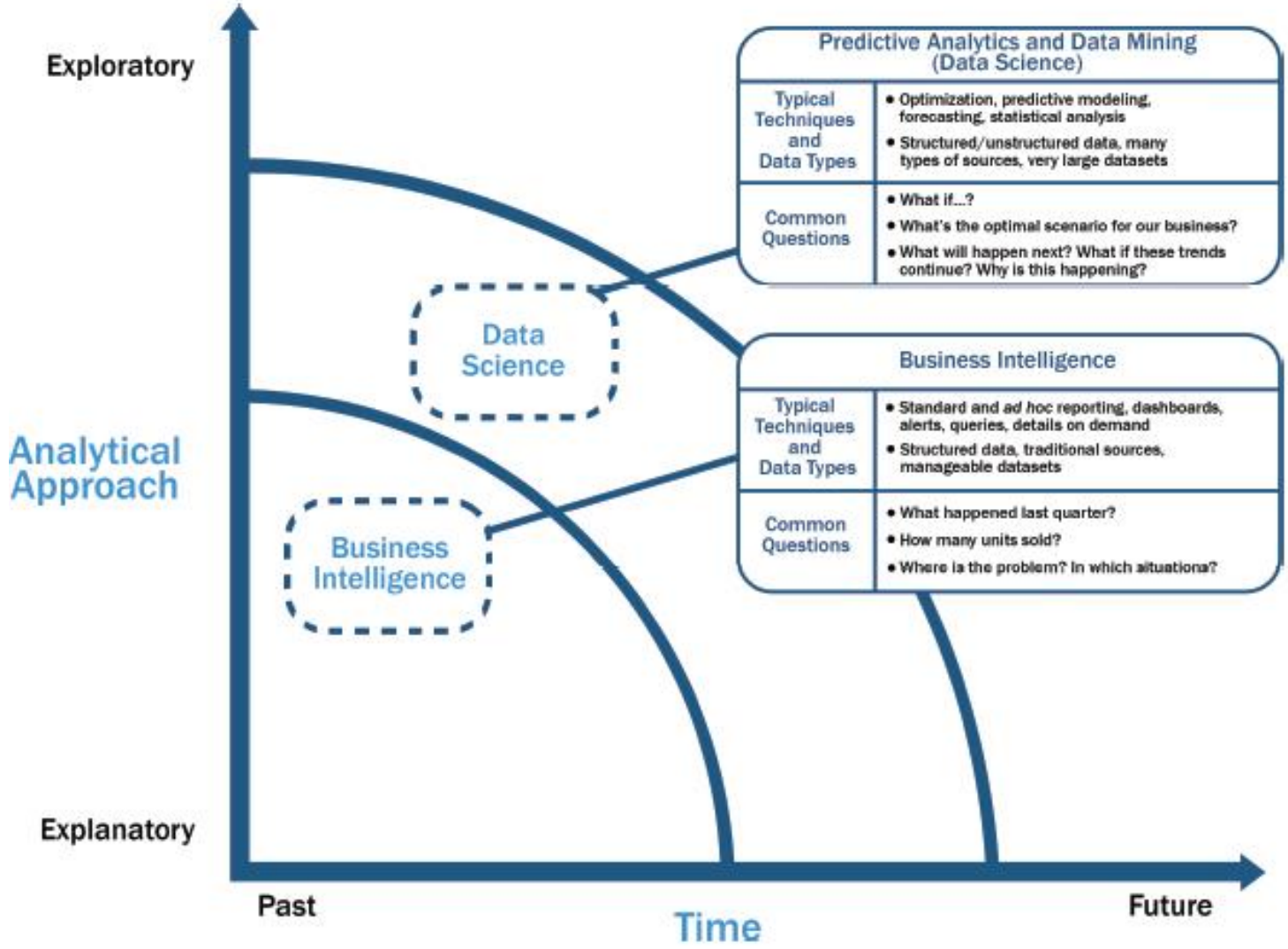


Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

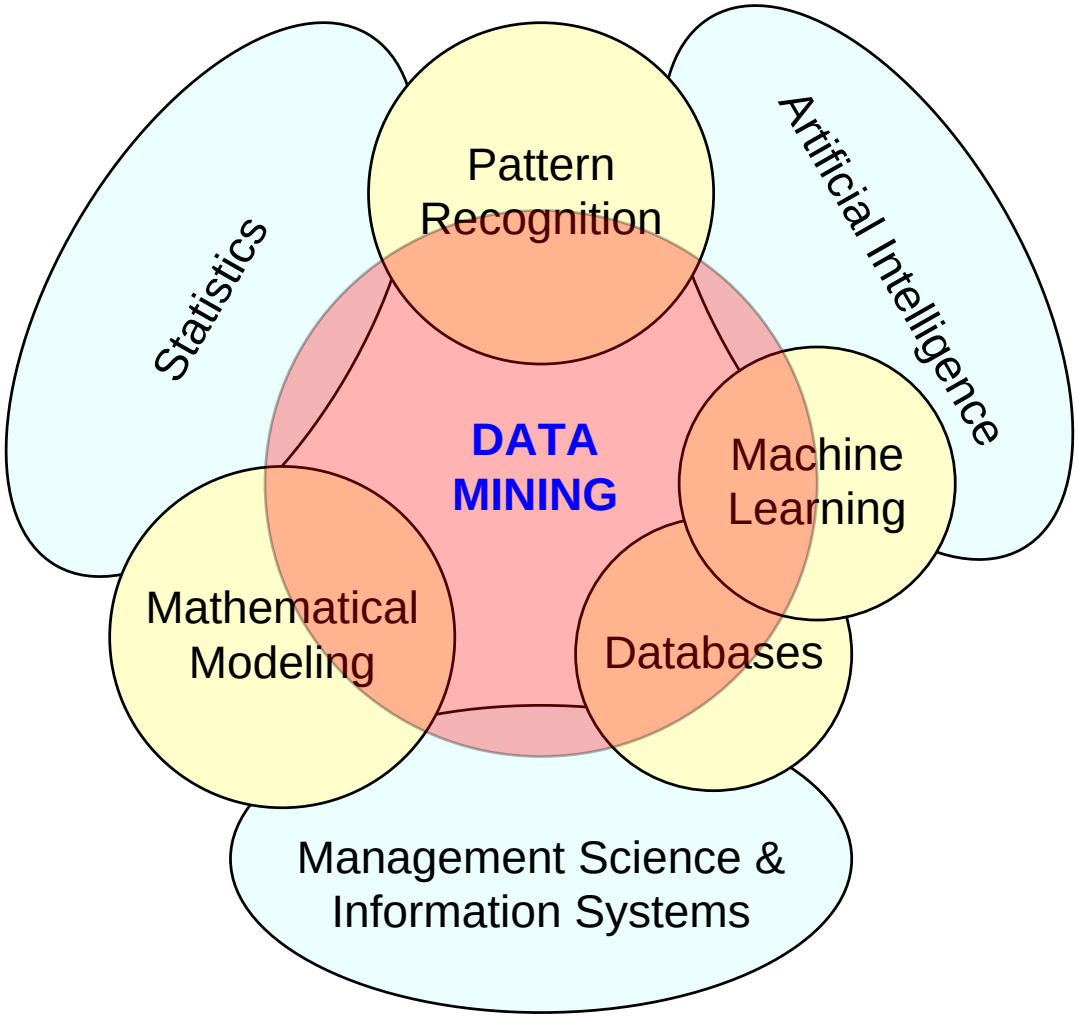
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Science and Business Intelligence



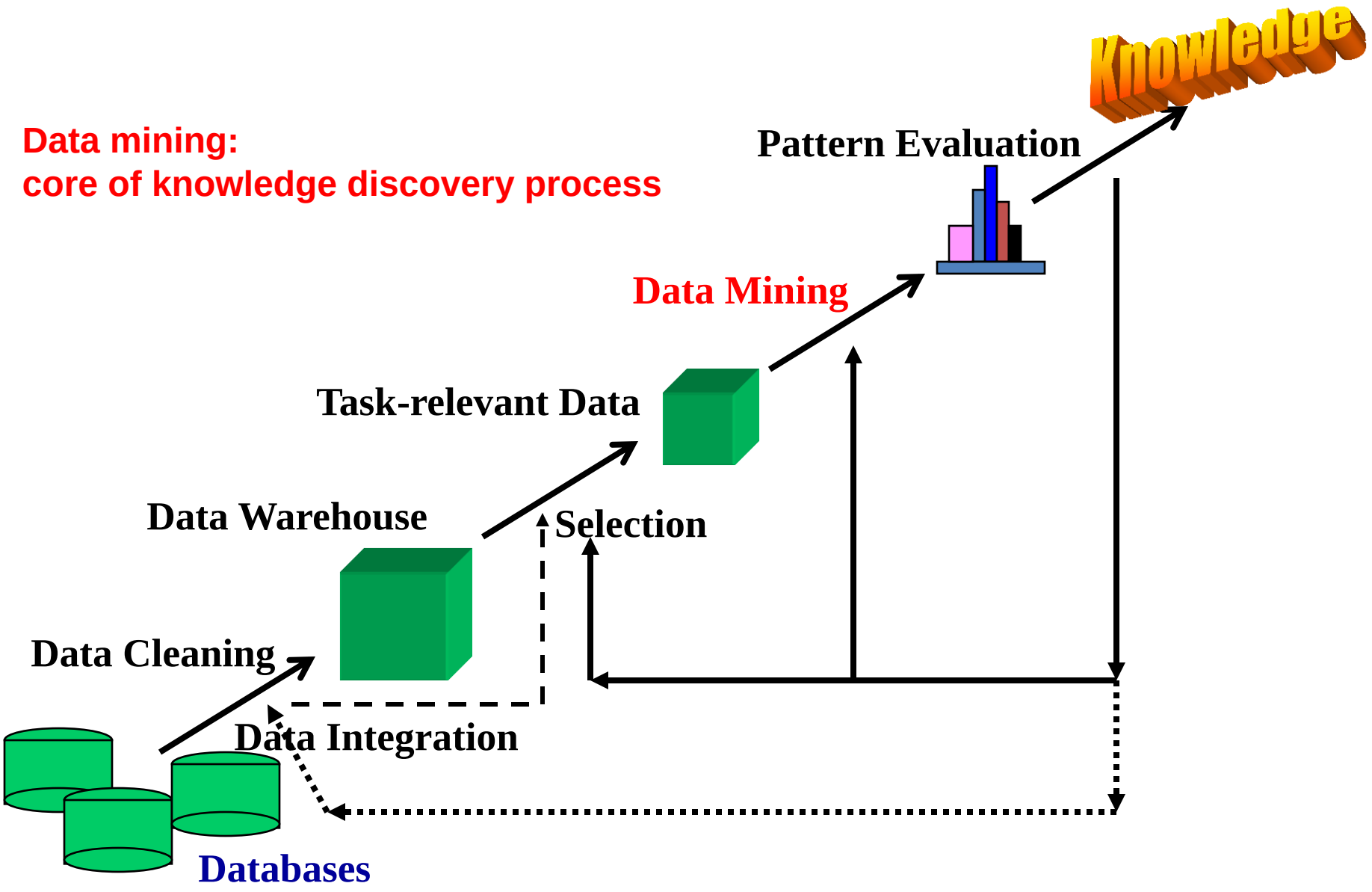
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Mining at the Intersection of Many Disciplines



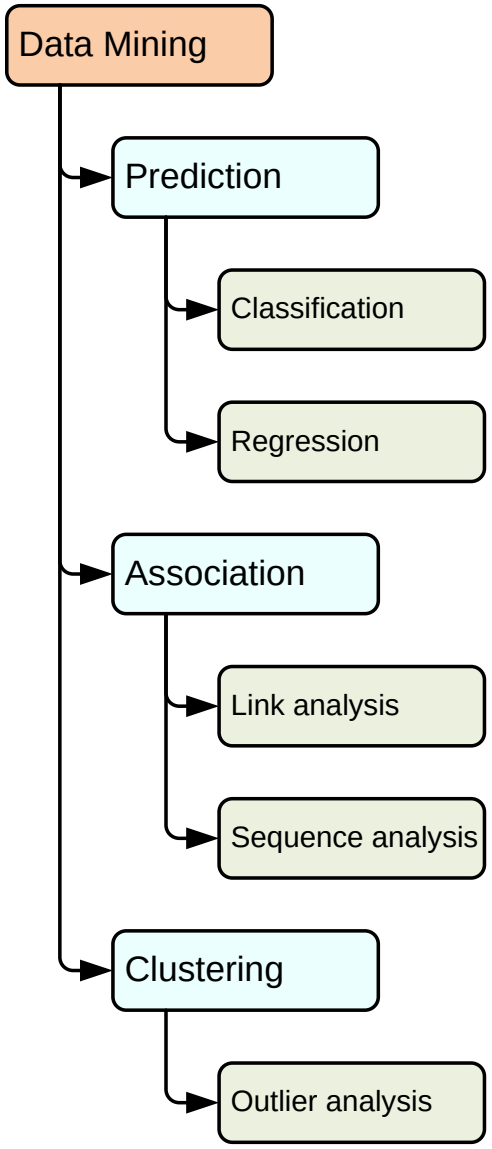
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Knowledge Discovery (KDD) Process



Source: Han & Kamber (2006)

A Taxonomy for Data Mining Tasks



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

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

Deep Learning

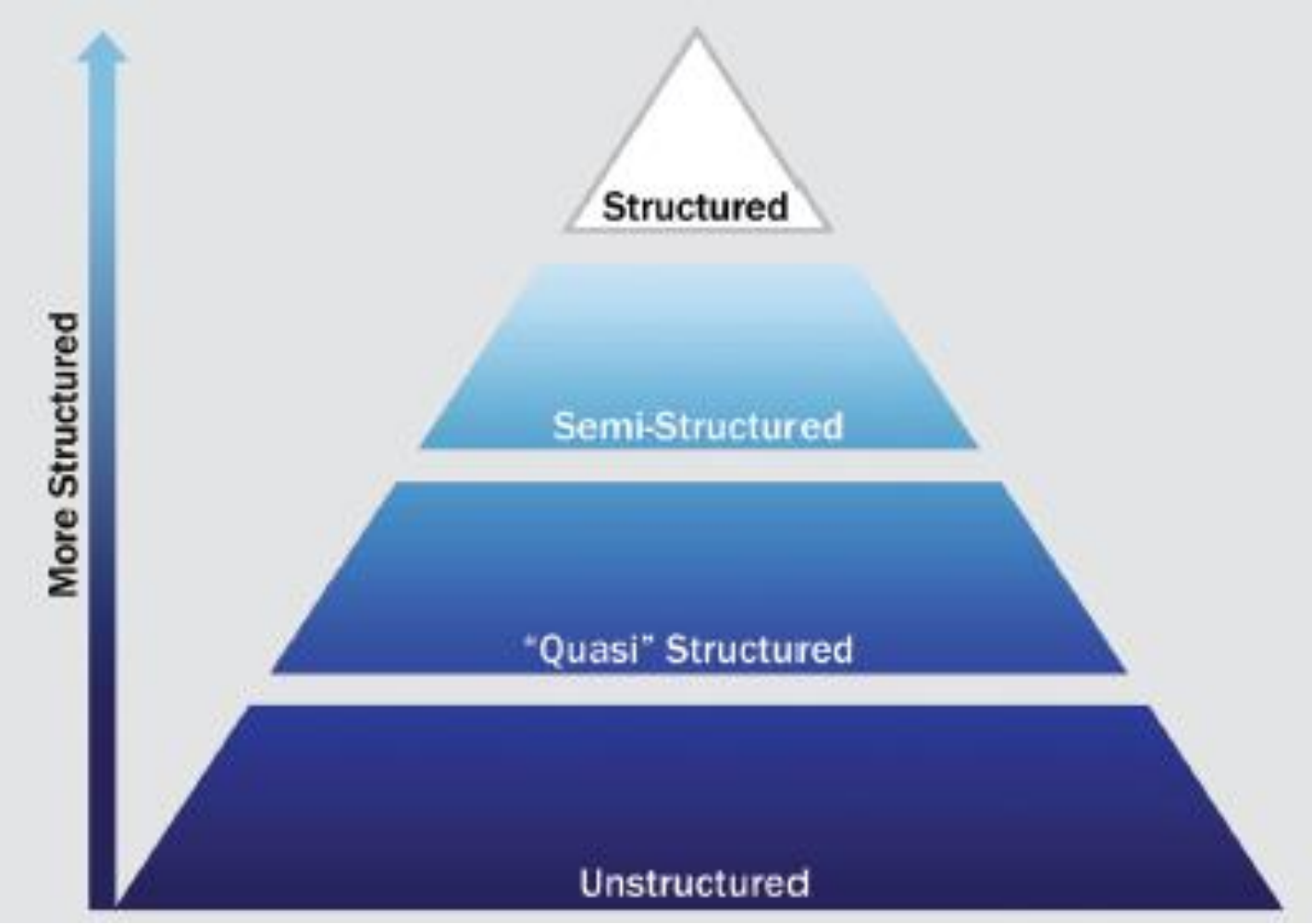
Intelligence from Big Data

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

Big Data

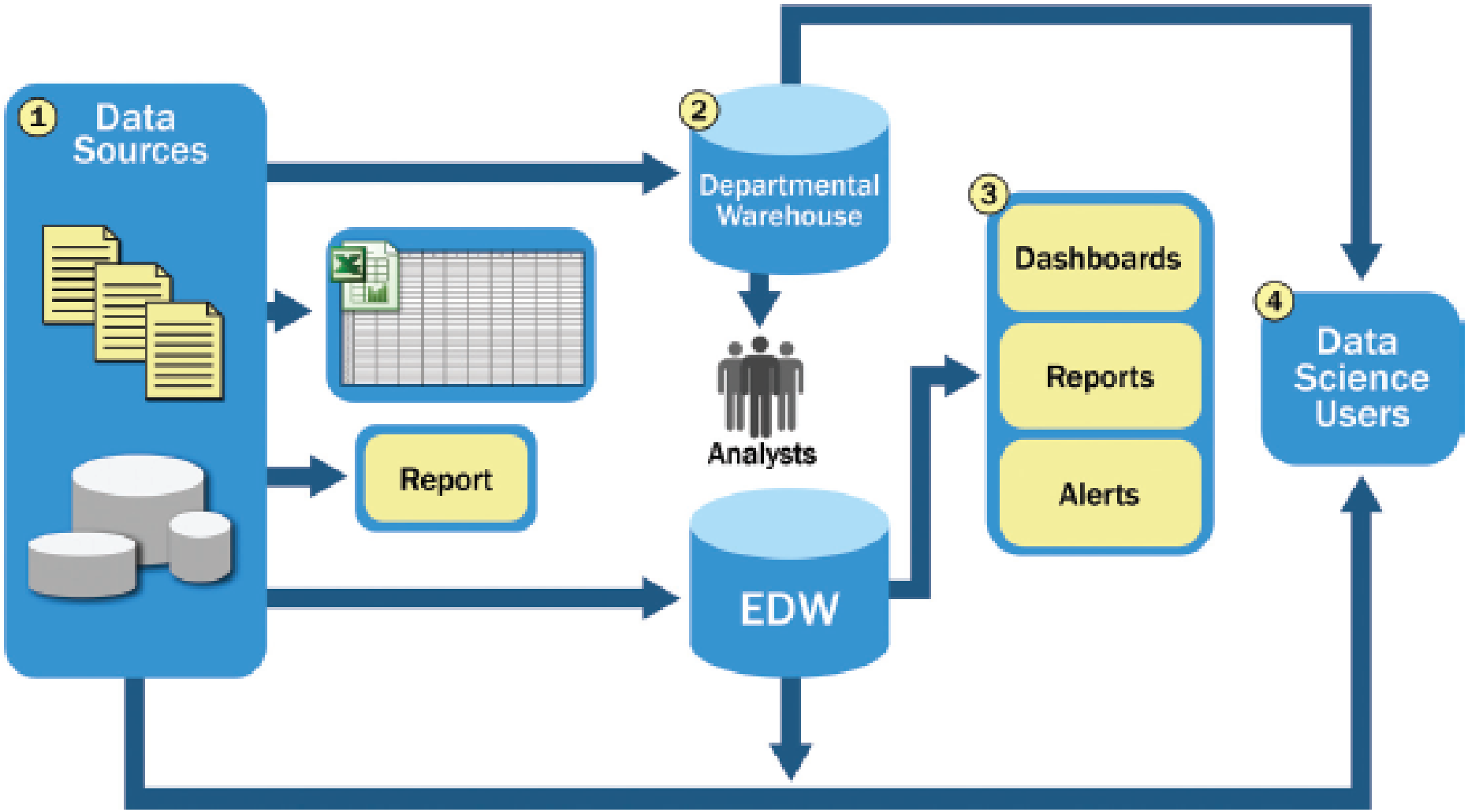
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Big Data Growth is increasingly unstructured



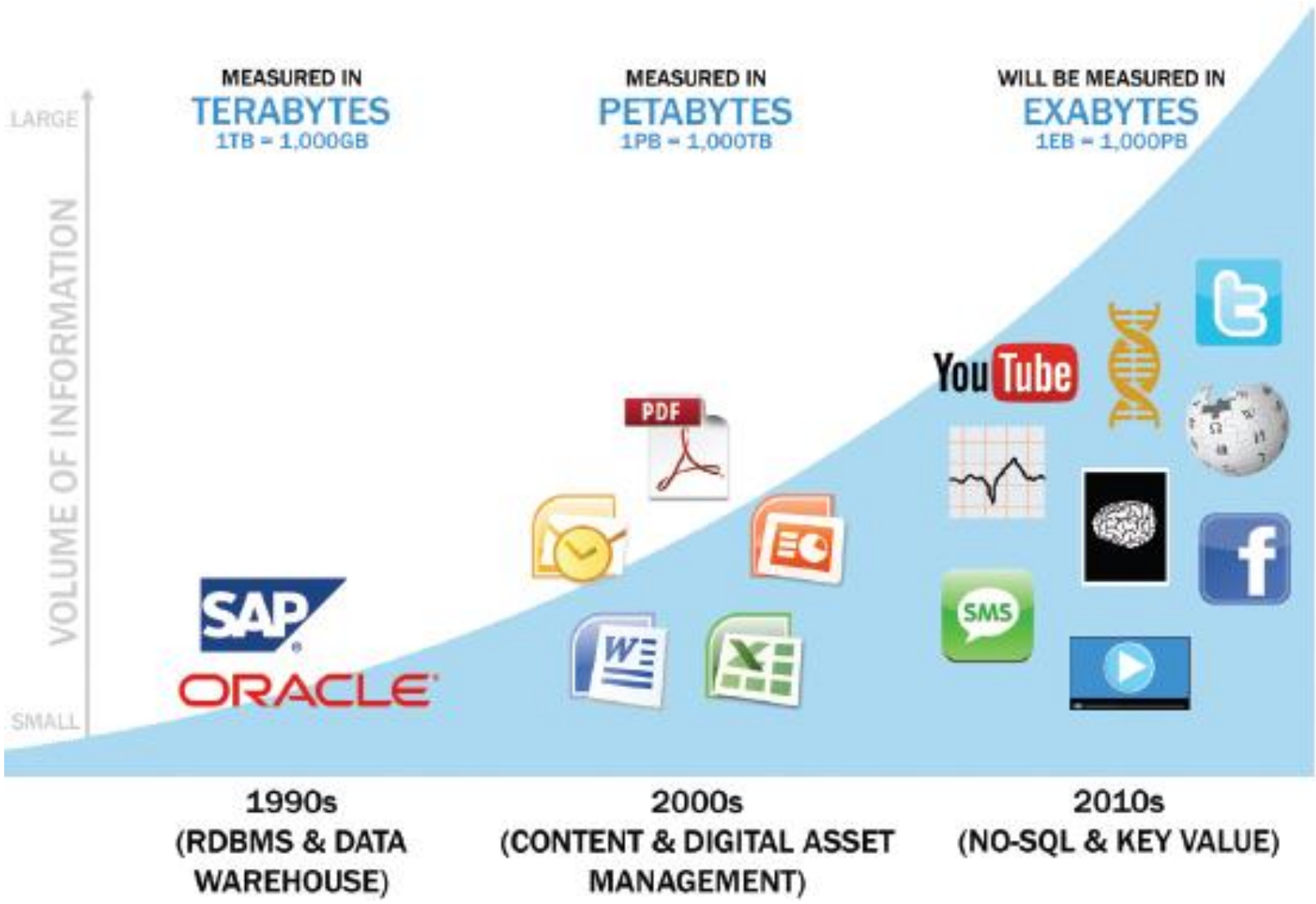
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Typical Analytic Architecture



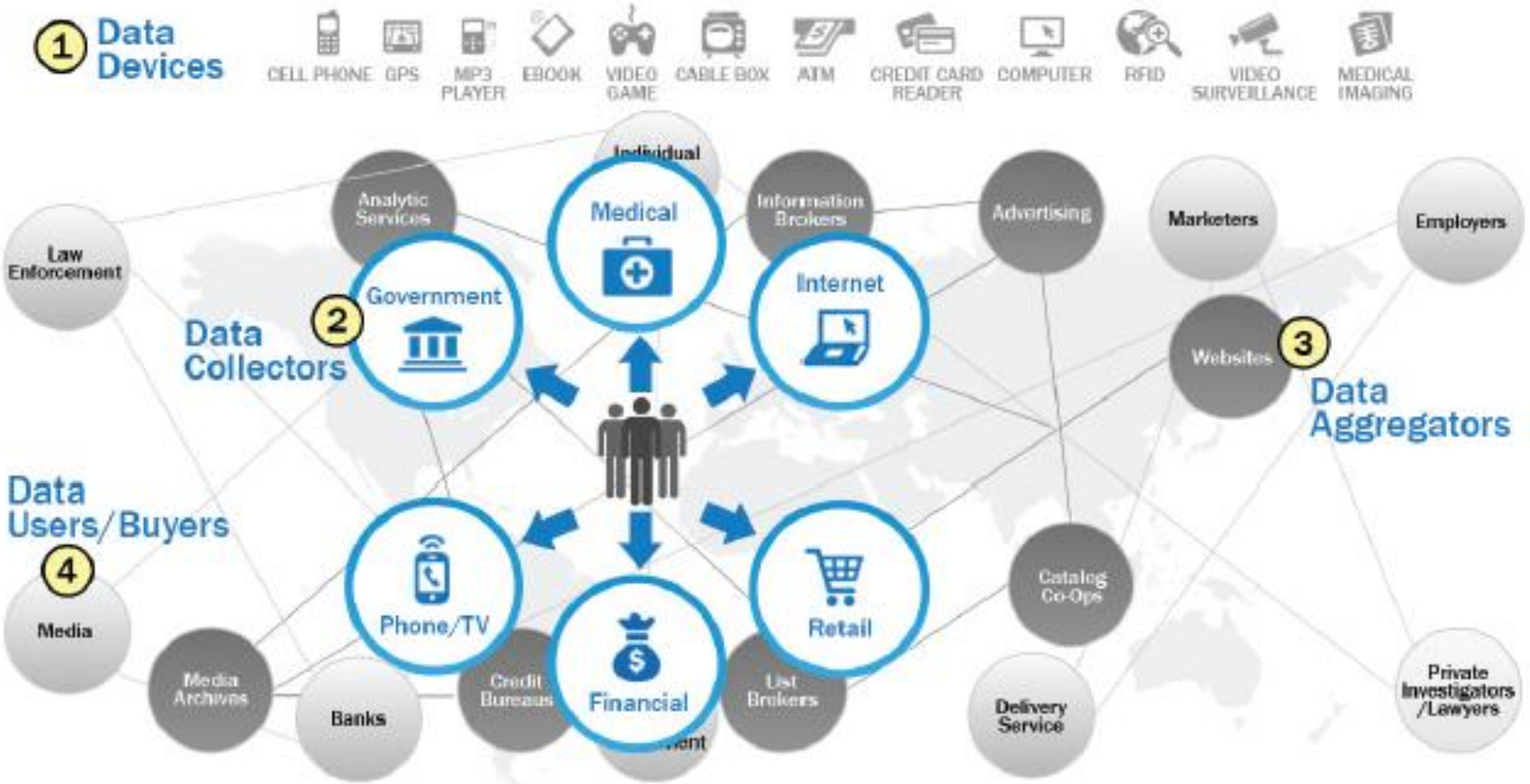
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Evolution and the Rise of Big Data Sources



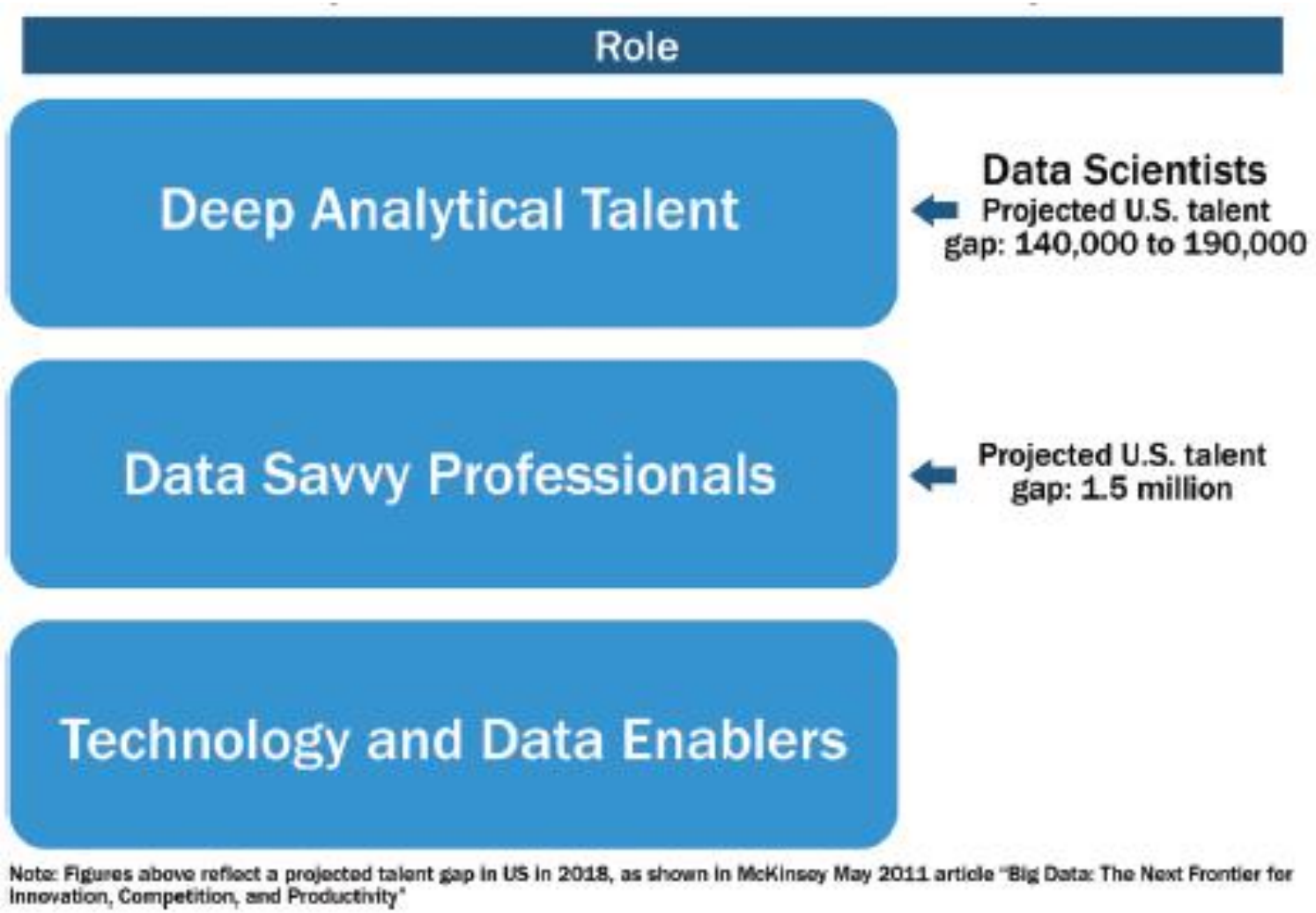
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Emerging Big Data Ecosystem



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

Key Roles for the New Big Data Ecosystem

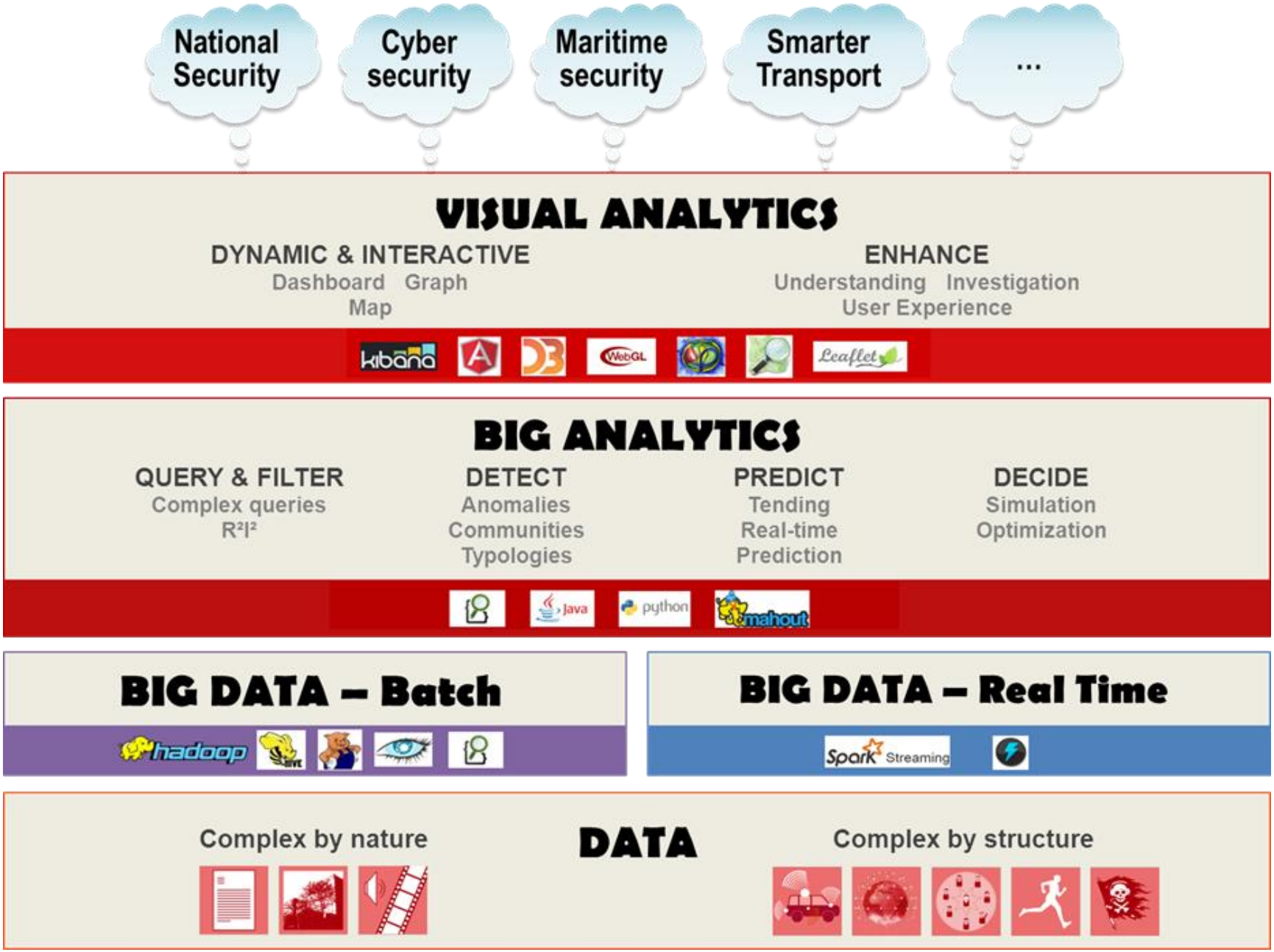


Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Profile of a Data Scientist

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

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

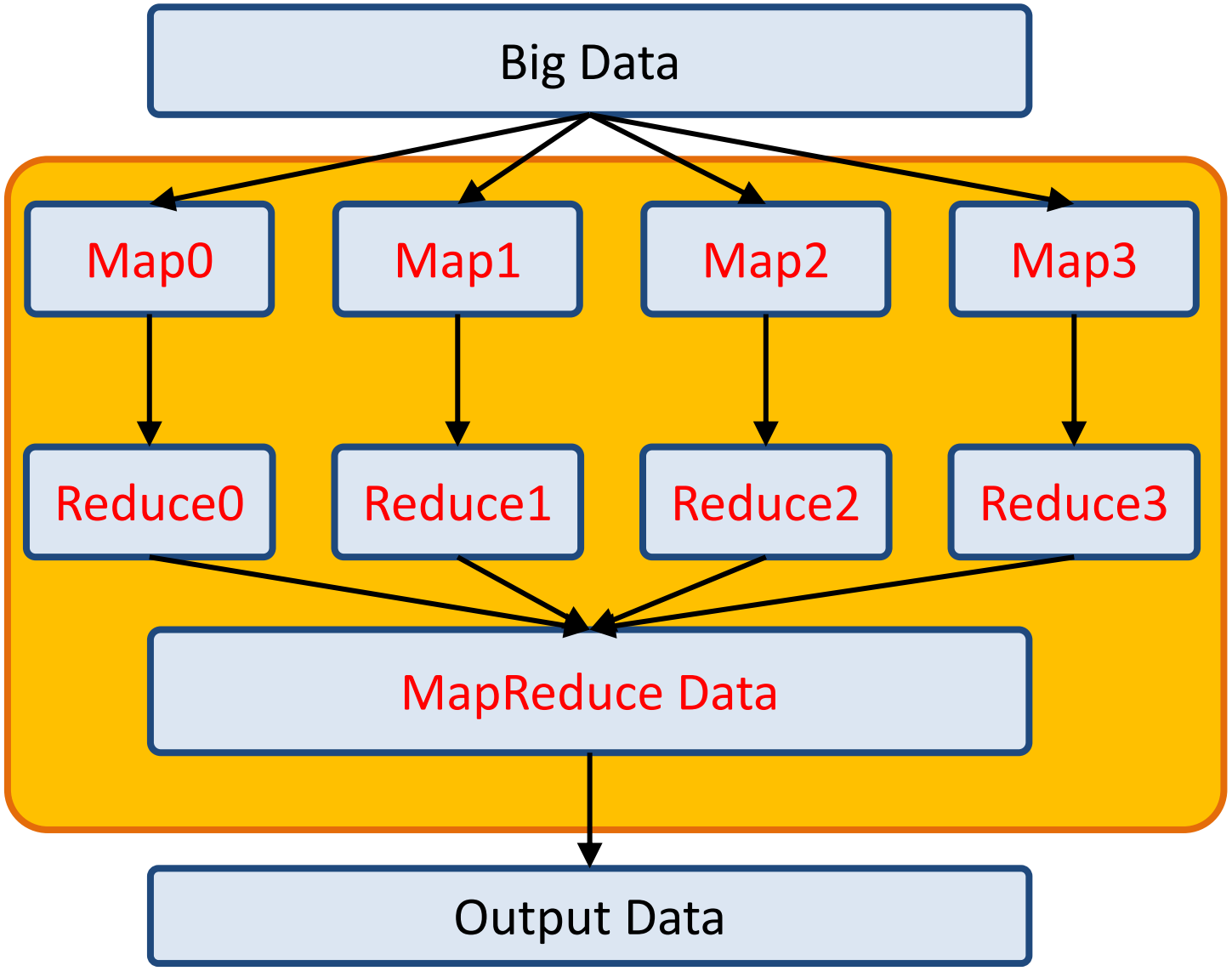


Source: https://www.thalesgroup.com/en/worldwide/big-data/big-data-big-analytics-visual-analytics-what-does-it-all-mean

MapReduce Paradigm

Map

Reduce



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

Source: http://hadoop.apache.org/

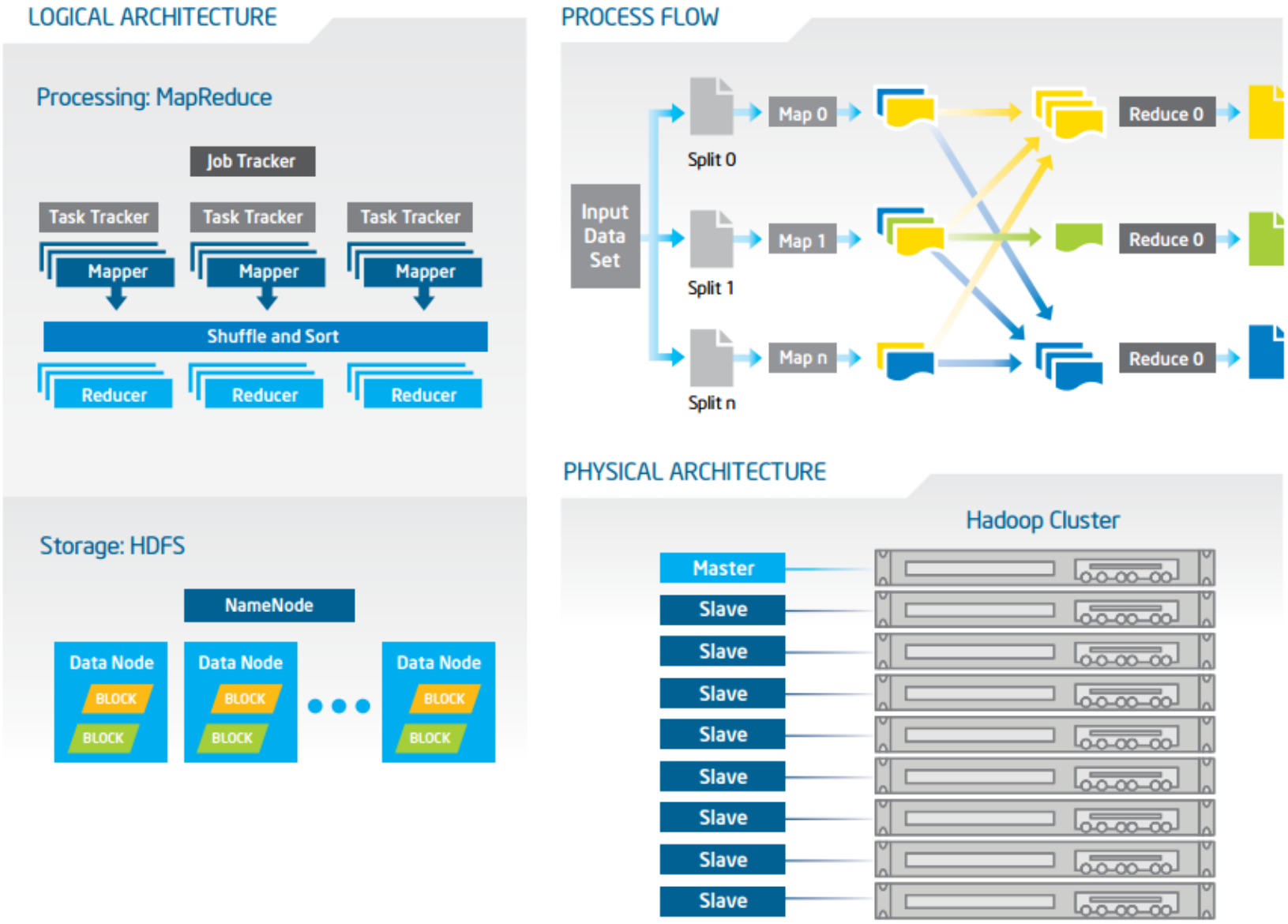
MapReduce **Processing**

+

HDFS **Storage**

Source: http://hadoop.apache.org/

Big Data with Hadoop Architecture

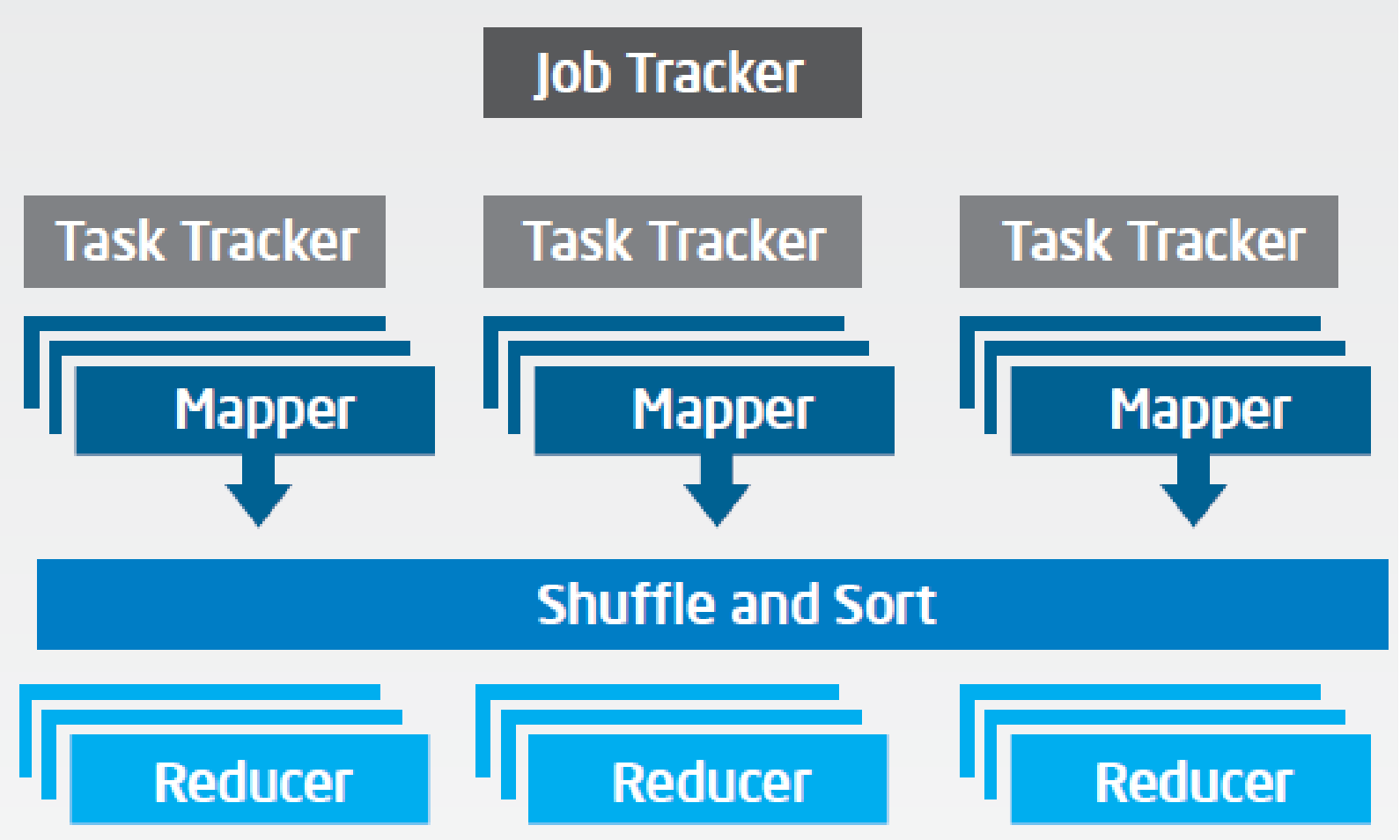


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

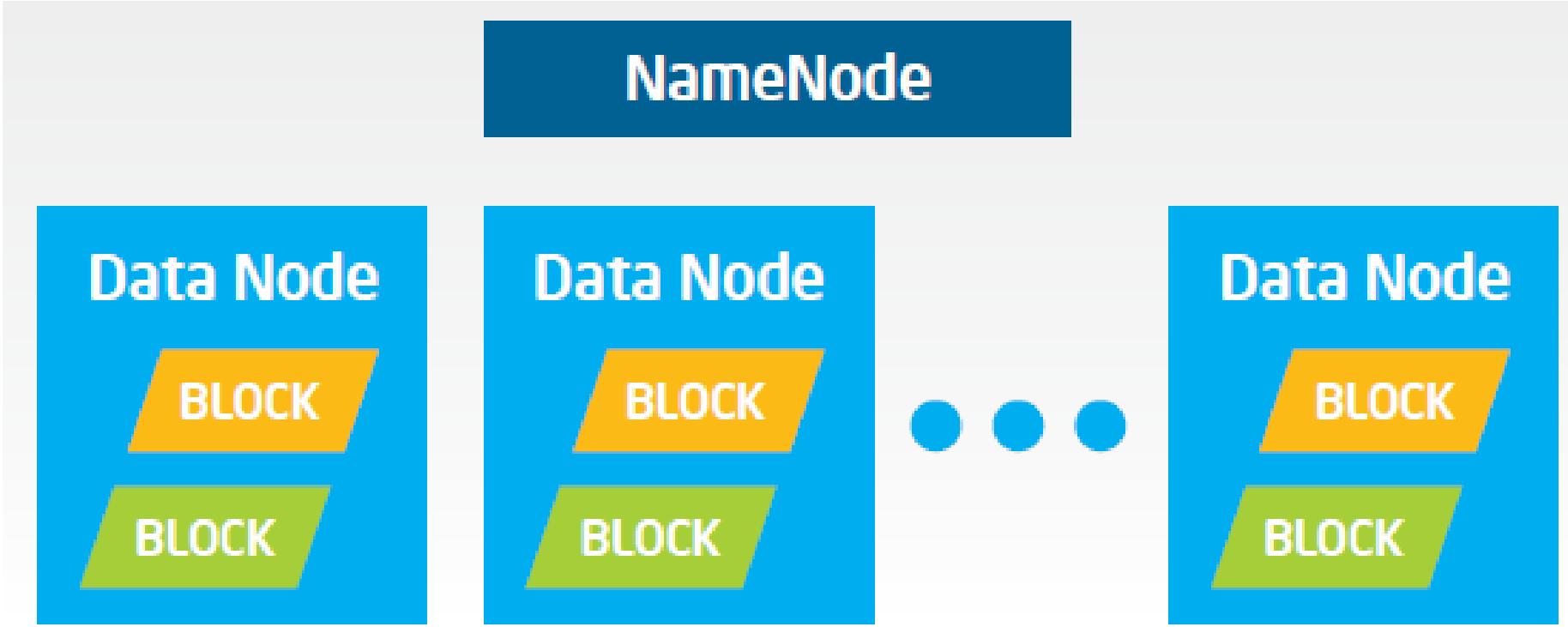


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

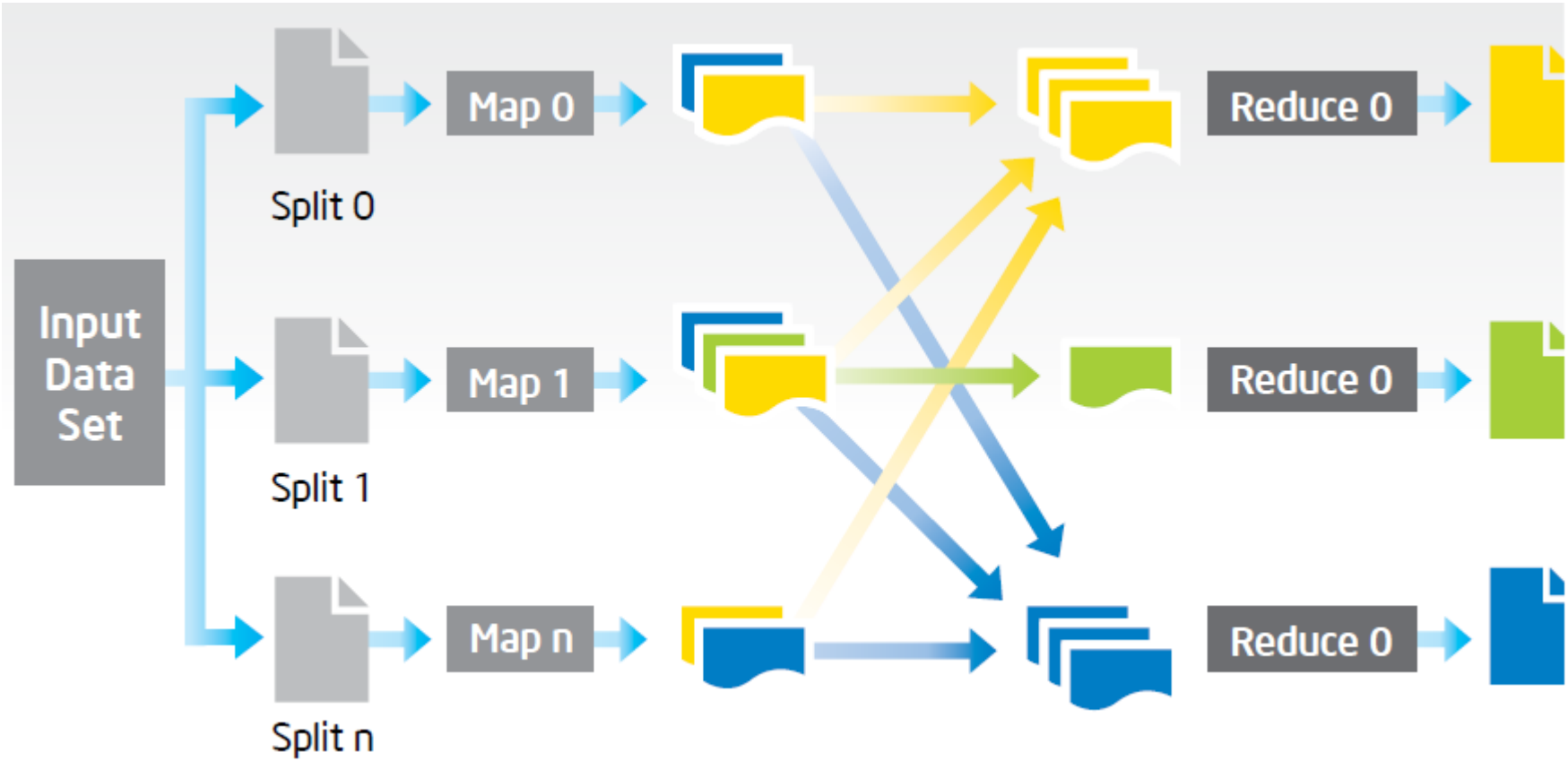
Logical Architecture

Storage: HDFS



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

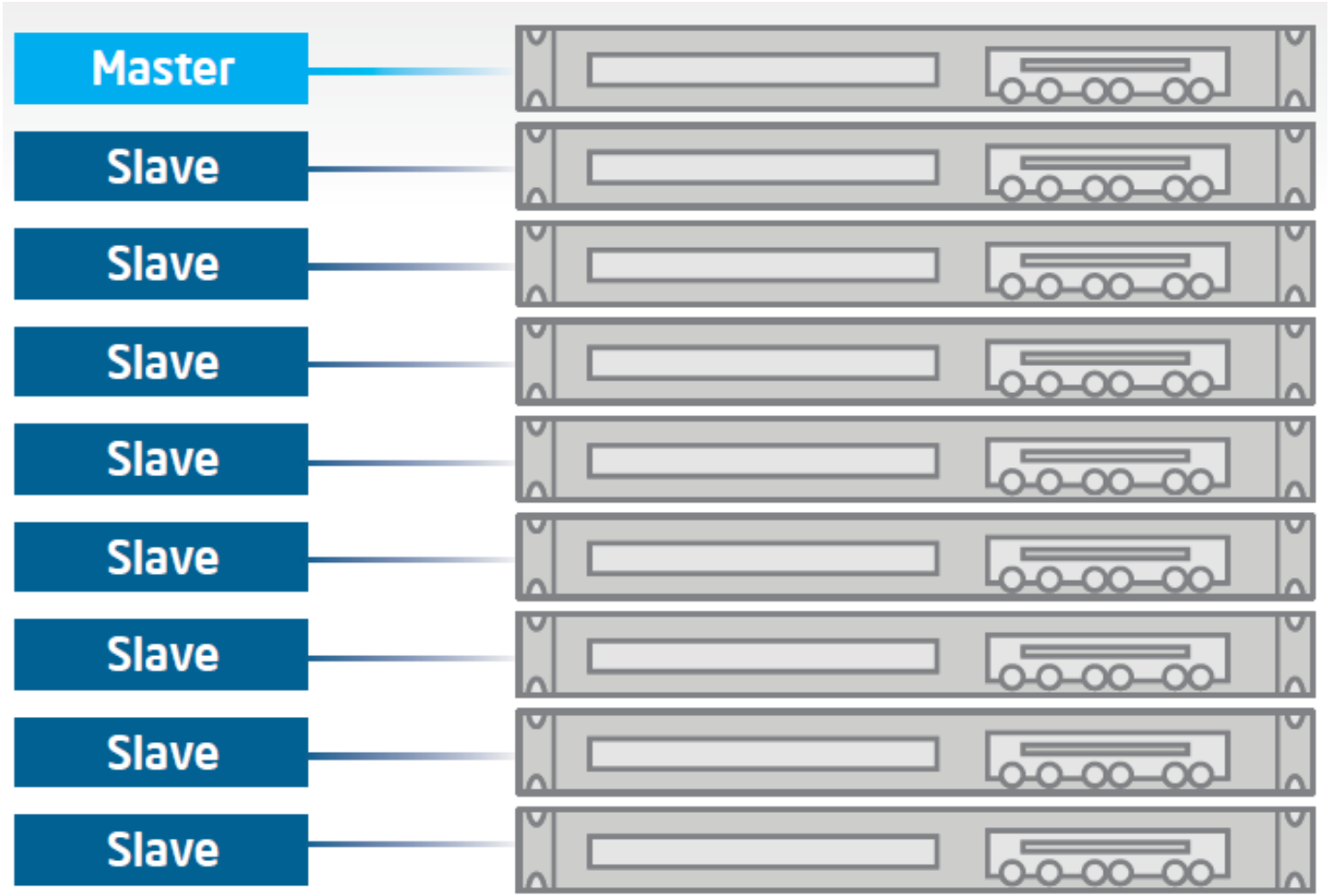
Big Data with Hadoop Architecture
Process Flow



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

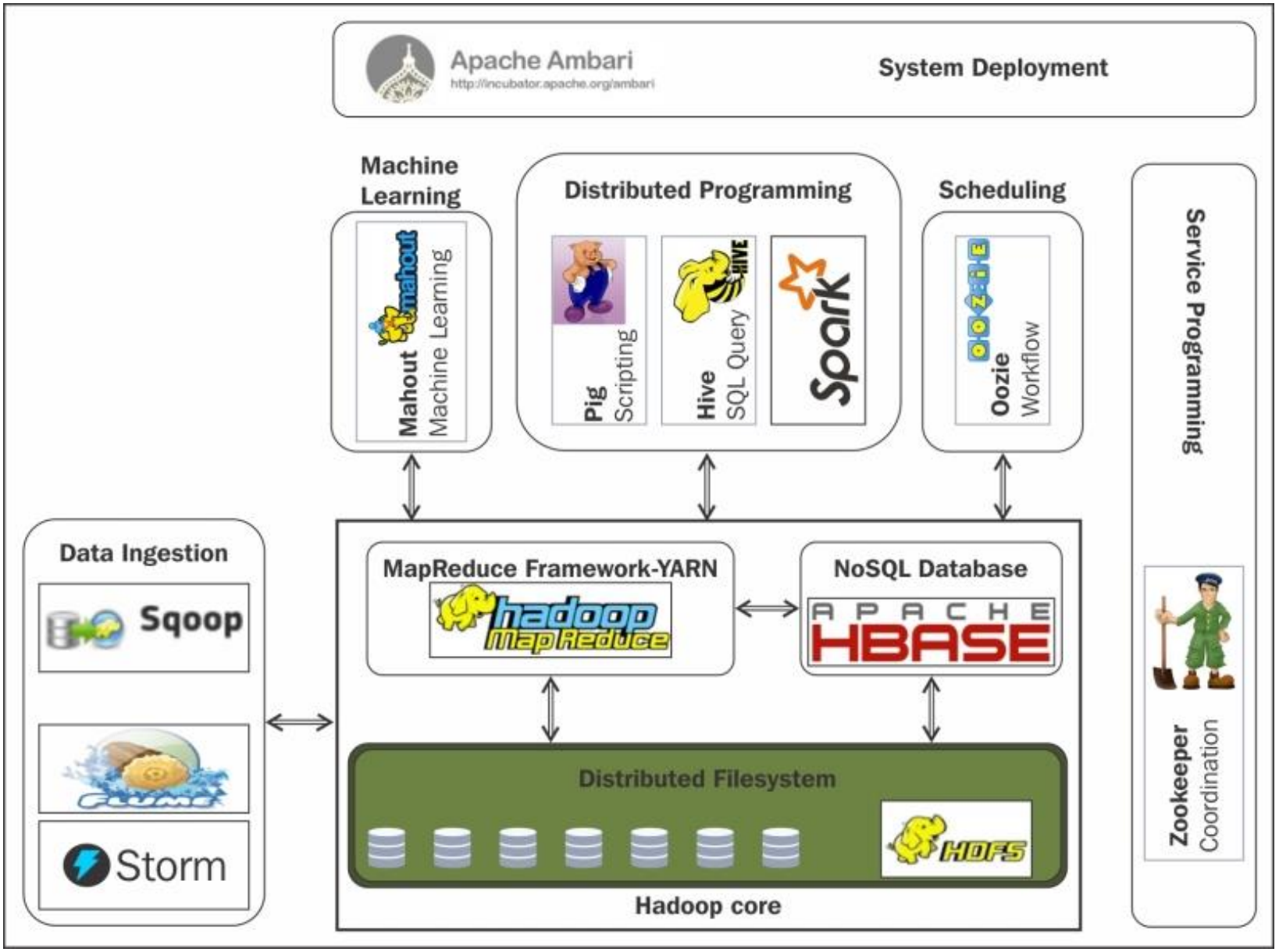
Big Data with Hadoop Architecture

Hadoop Cluster



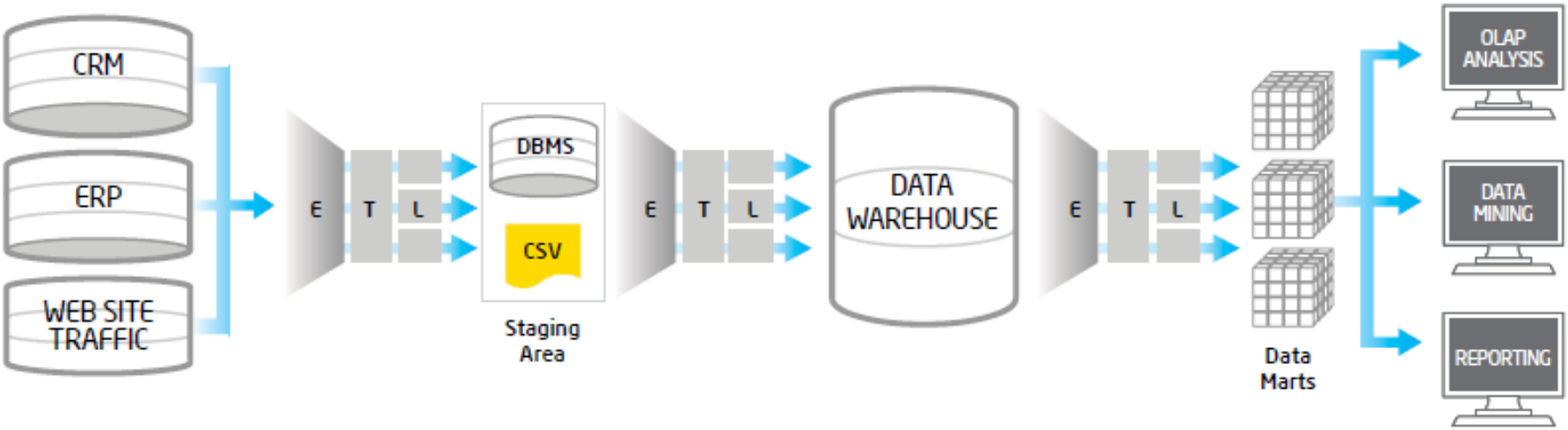
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Hadoop Ecosystem



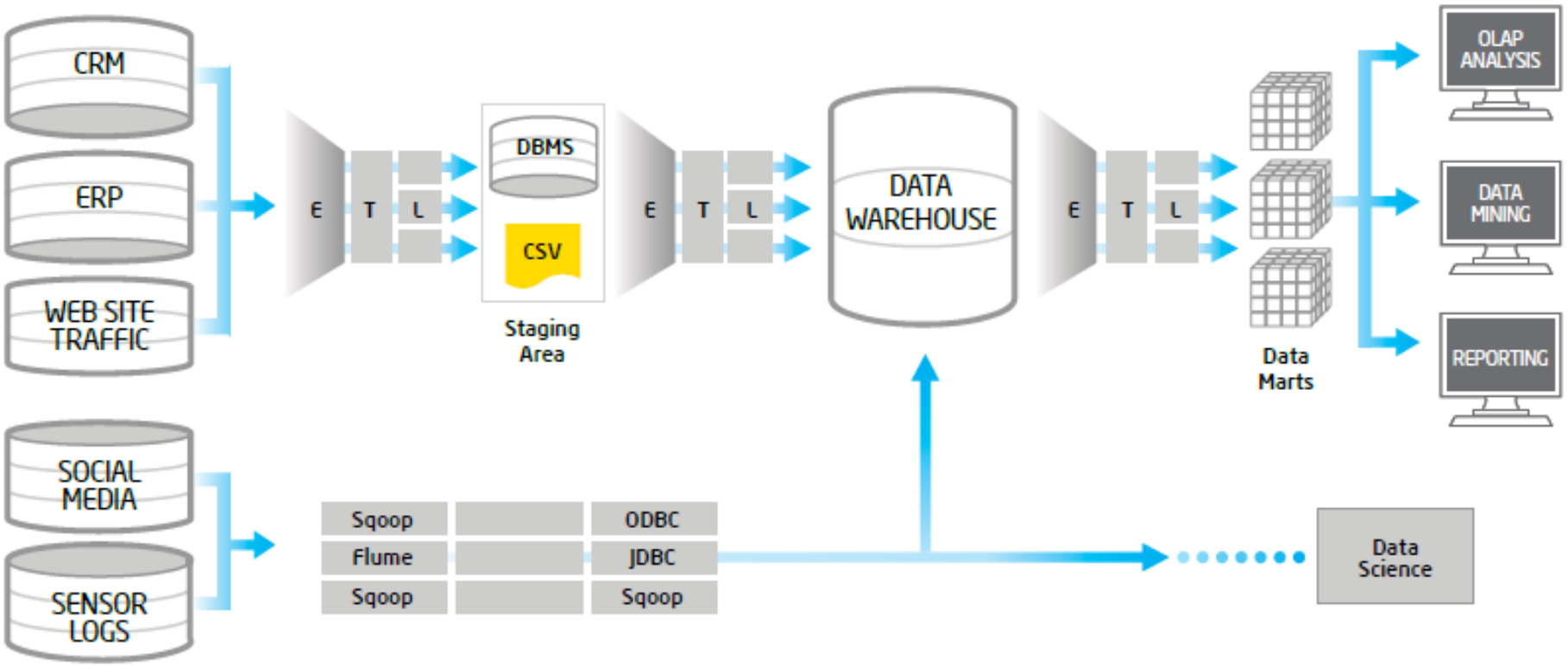
Source: Shiva Achari (2015), Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing

Traditional ETL Architecture



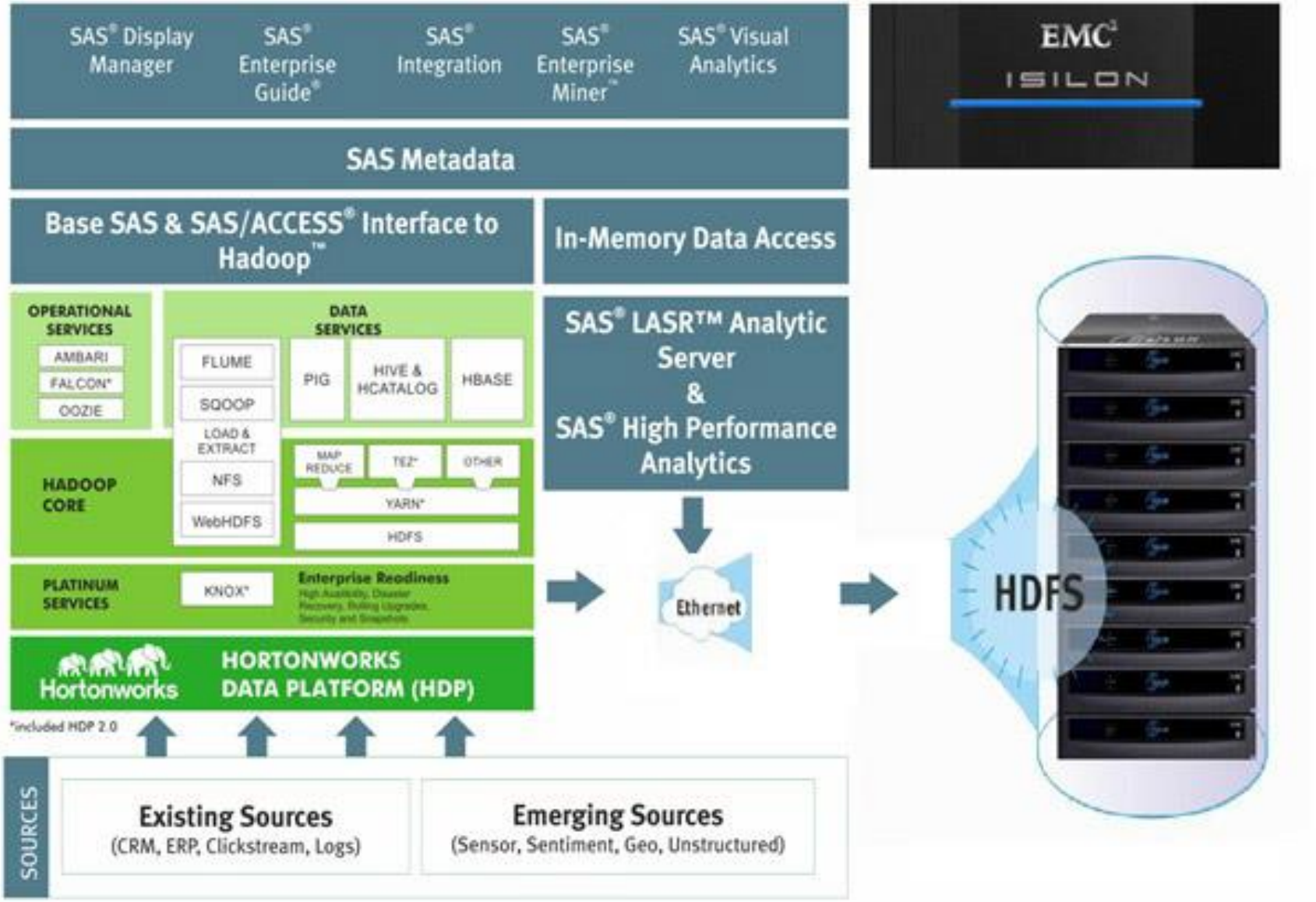
Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Offload ETL with Hadoop (Big Data Architecture)



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

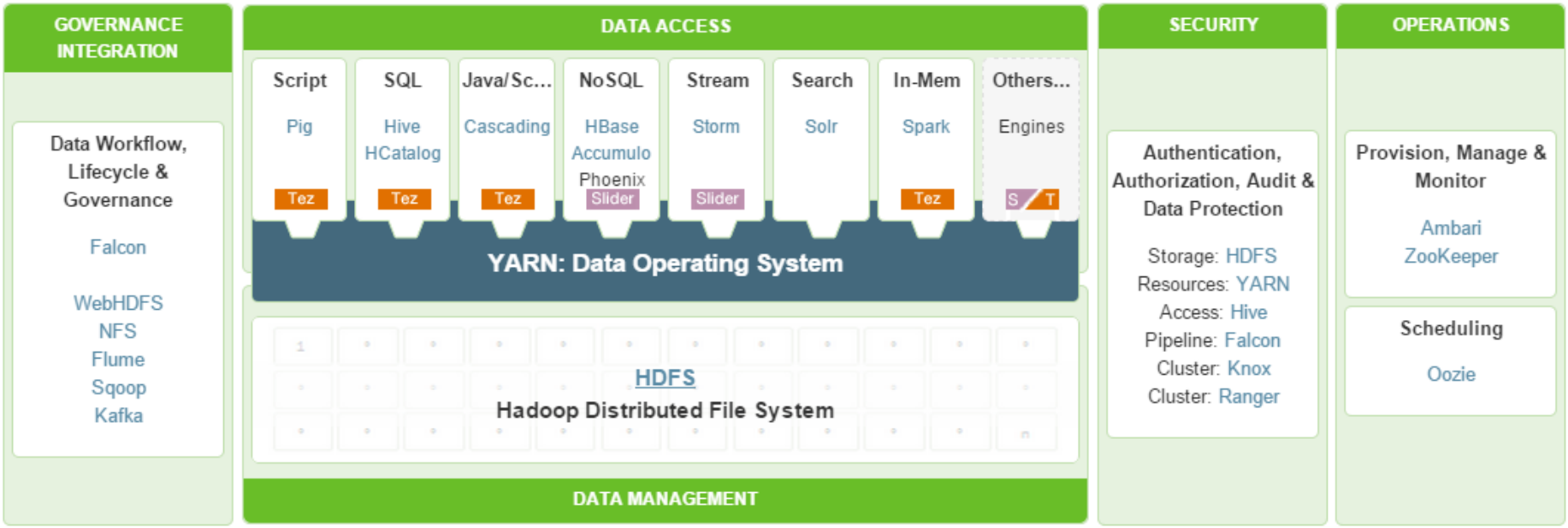
Big Data Solution



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

HDP

A Complete Enterprise Hadoop Data Platform



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

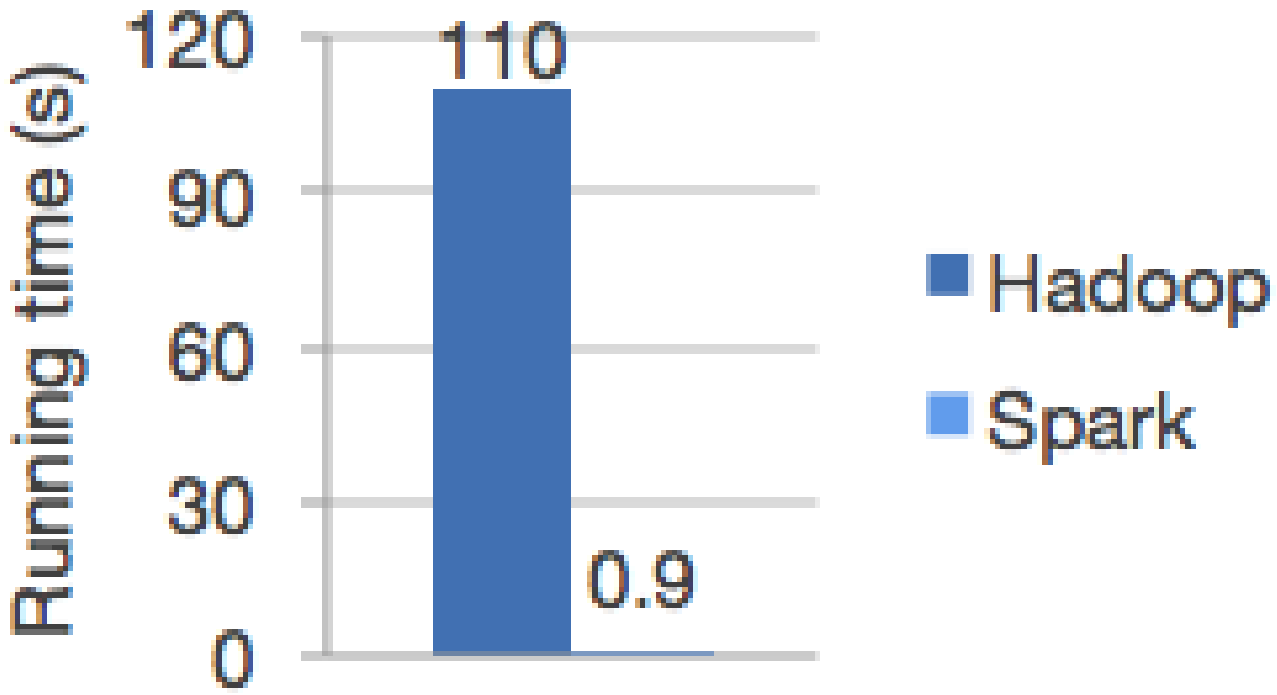
Lightning-fast cluster computing

Apache Spark

is a fast and general engine for large-scale data processing.

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

Logistic regression in Hadoop and Spark



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

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

Ease of Use

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

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

Word count in Spark's Python API

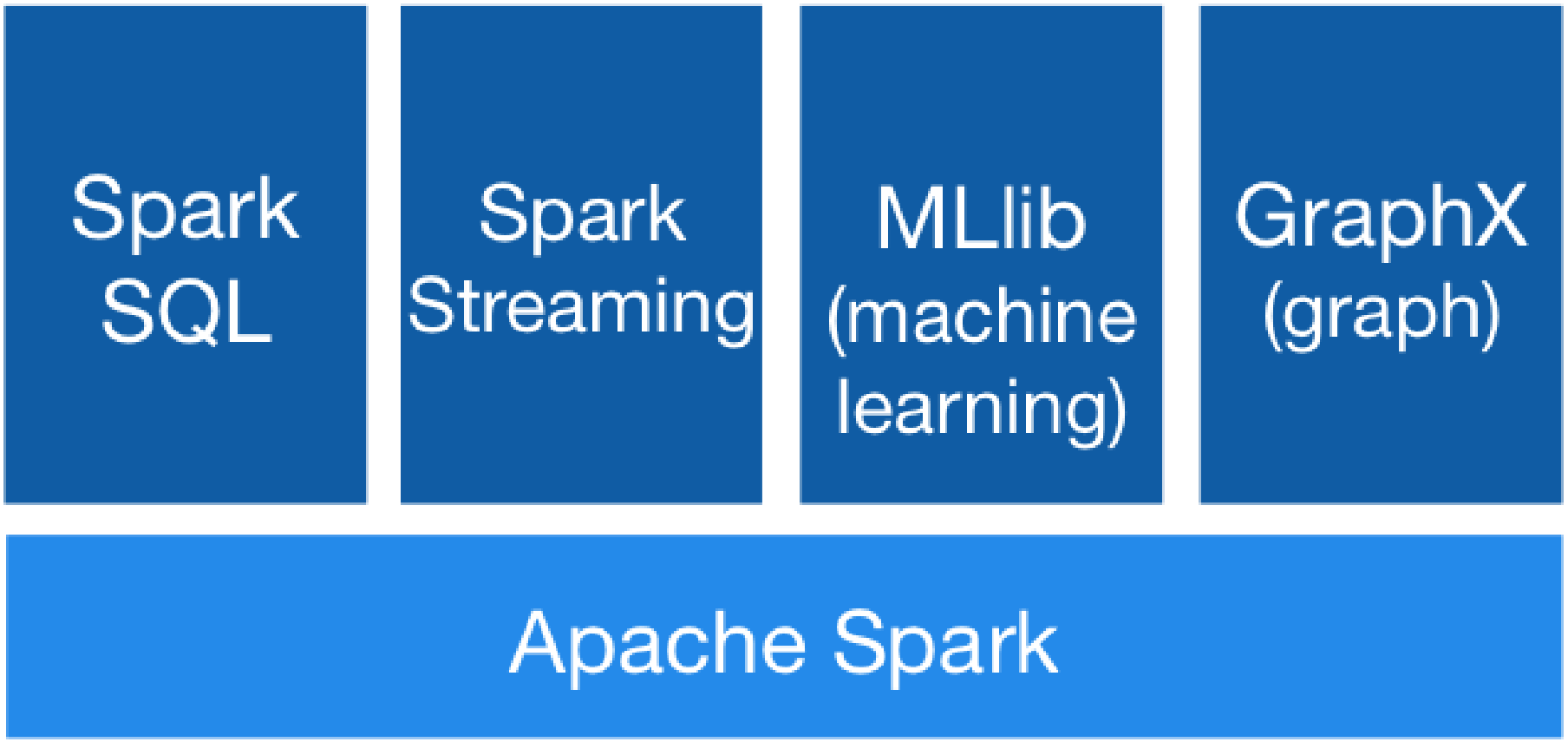
```
text_file = spark.textFile("hdfs://...")

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

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

Spark and Hadoop

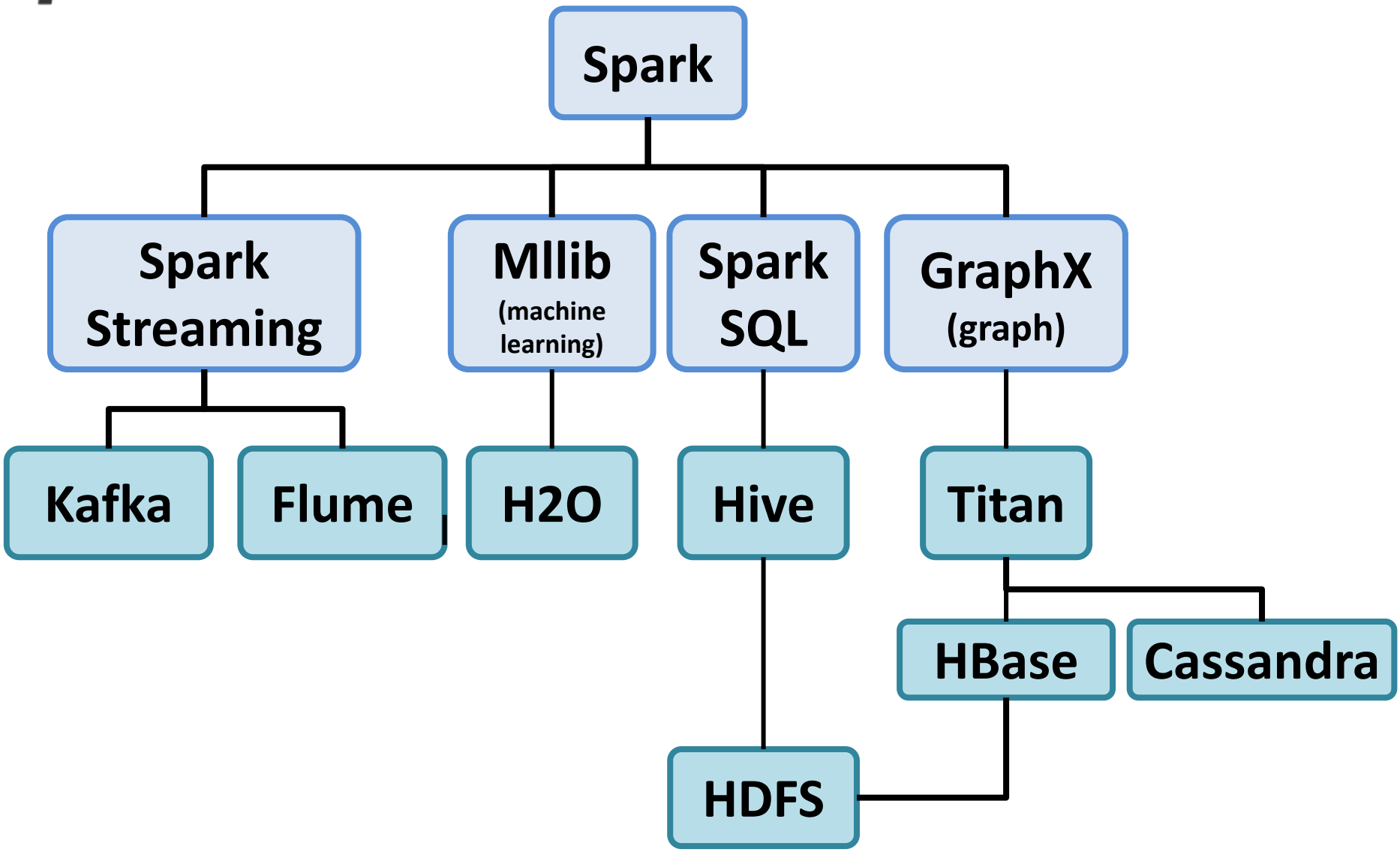
Source: http://spark.apache.org/

Spark Ecosystem



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

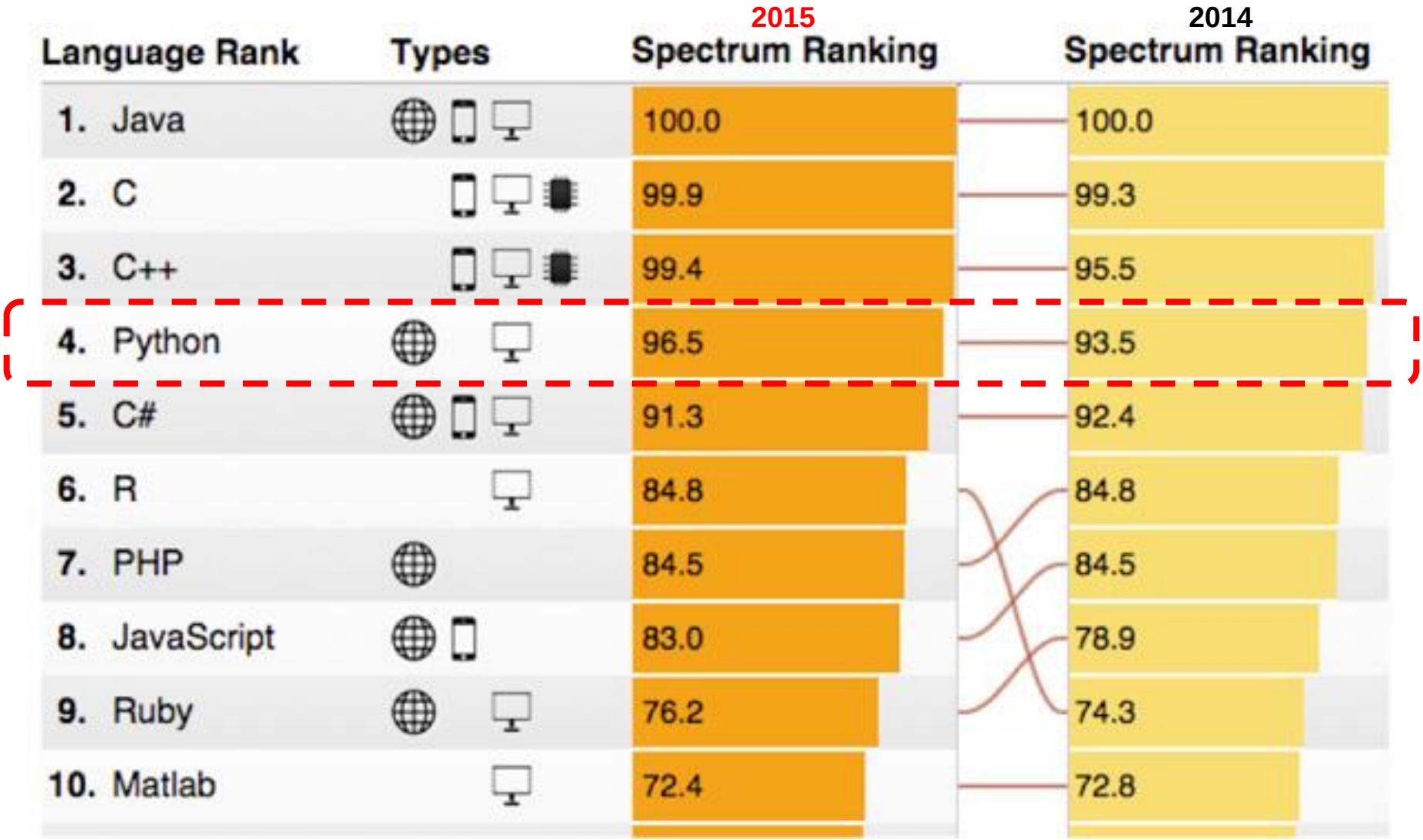
Spark Ecosystem



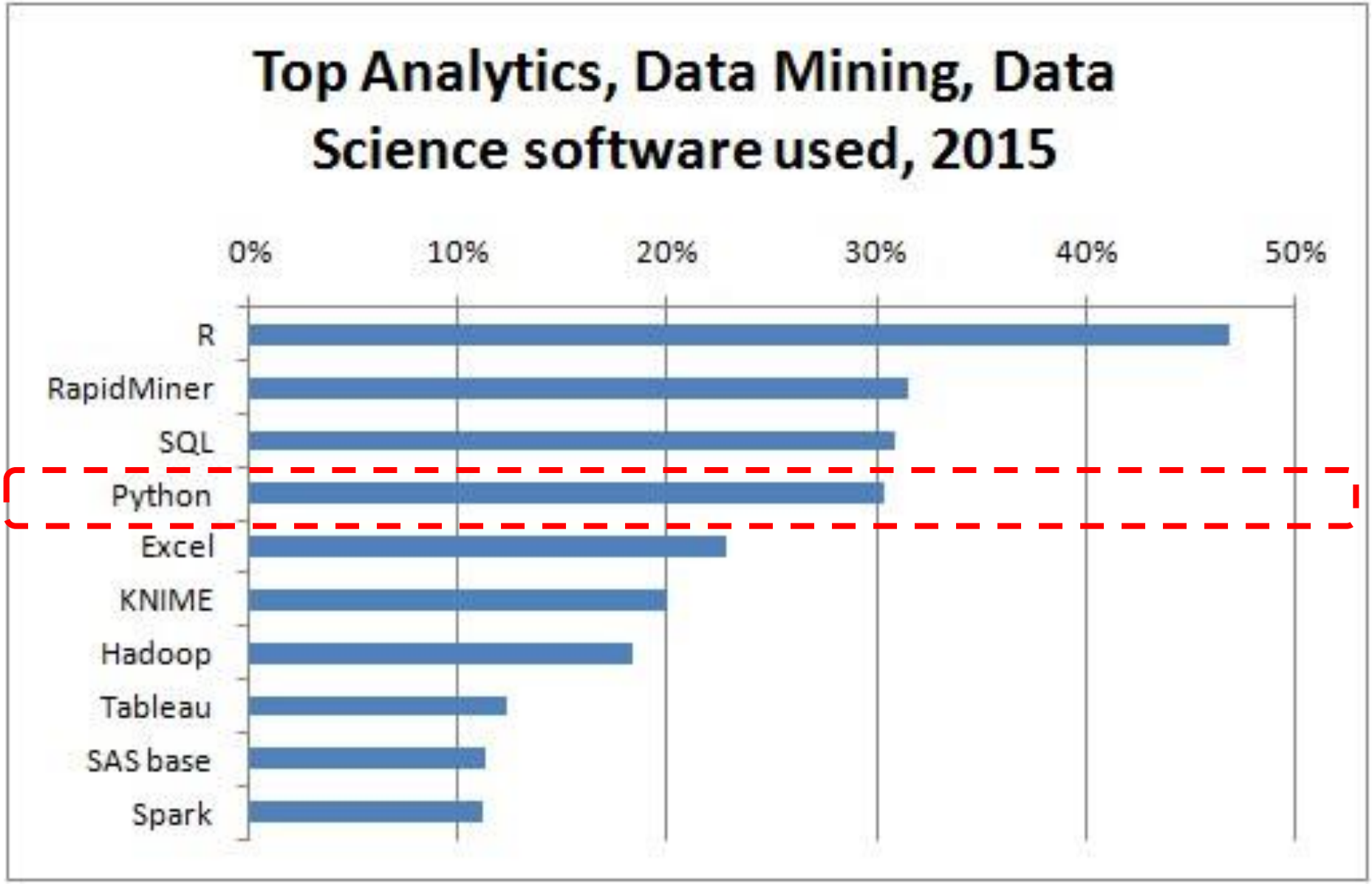
Source: Mike Frampton (2015), Mastering Apache Spark, Packt Publishing

Python for Big Data Analytics

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



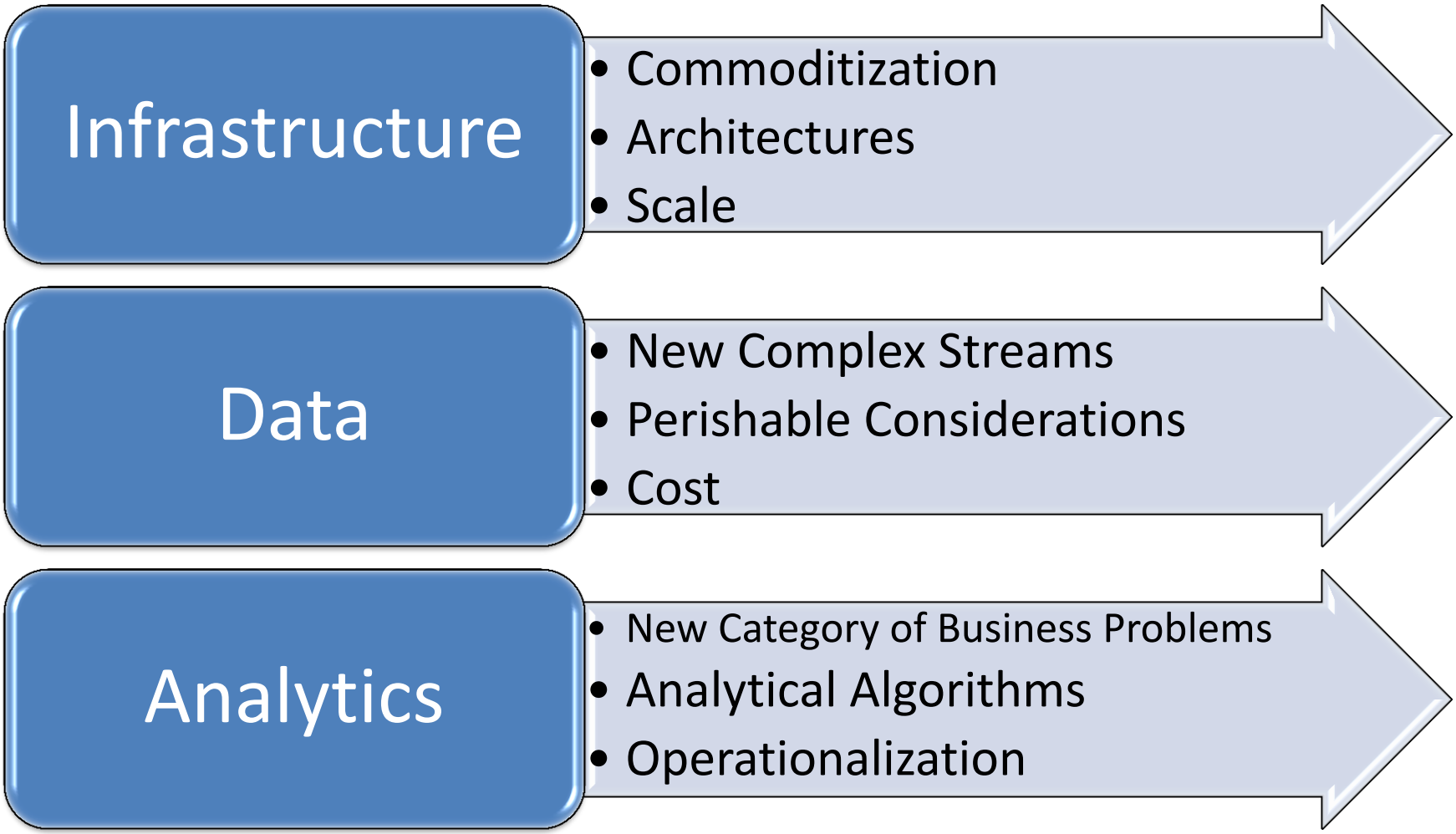
Source: http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages



Source: http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

SAS & Hadoop

Modern Reality



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

ADVANCED ANALYTICS

TEXT ANALYTICS

Finding treasures in unstructured data like social media or survey tools that could uncover insights about consumer sentiment

FORECASTING

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

OPTIMIZATION

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

DATA MINING

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

STATISTICS

Source: Deepak Ramanathan (2014)

Trends in Analytics

Complex Business Problems Are Driving Analytics Innovation

Speed Will Be Of Essence

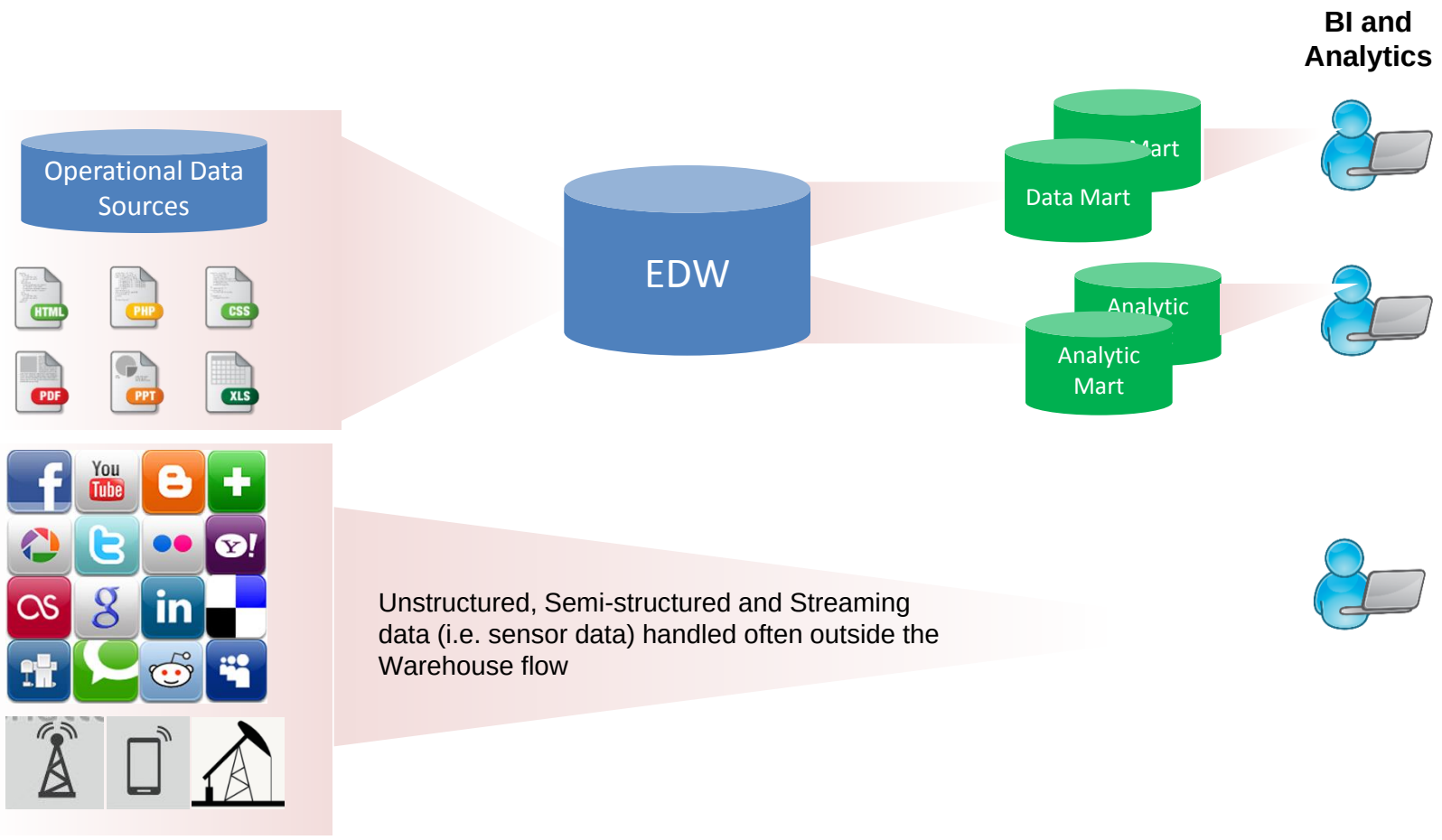
Leverage Analytics To Unlock The Information Contained In Unstructured Data

Operationalizing Analytics

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

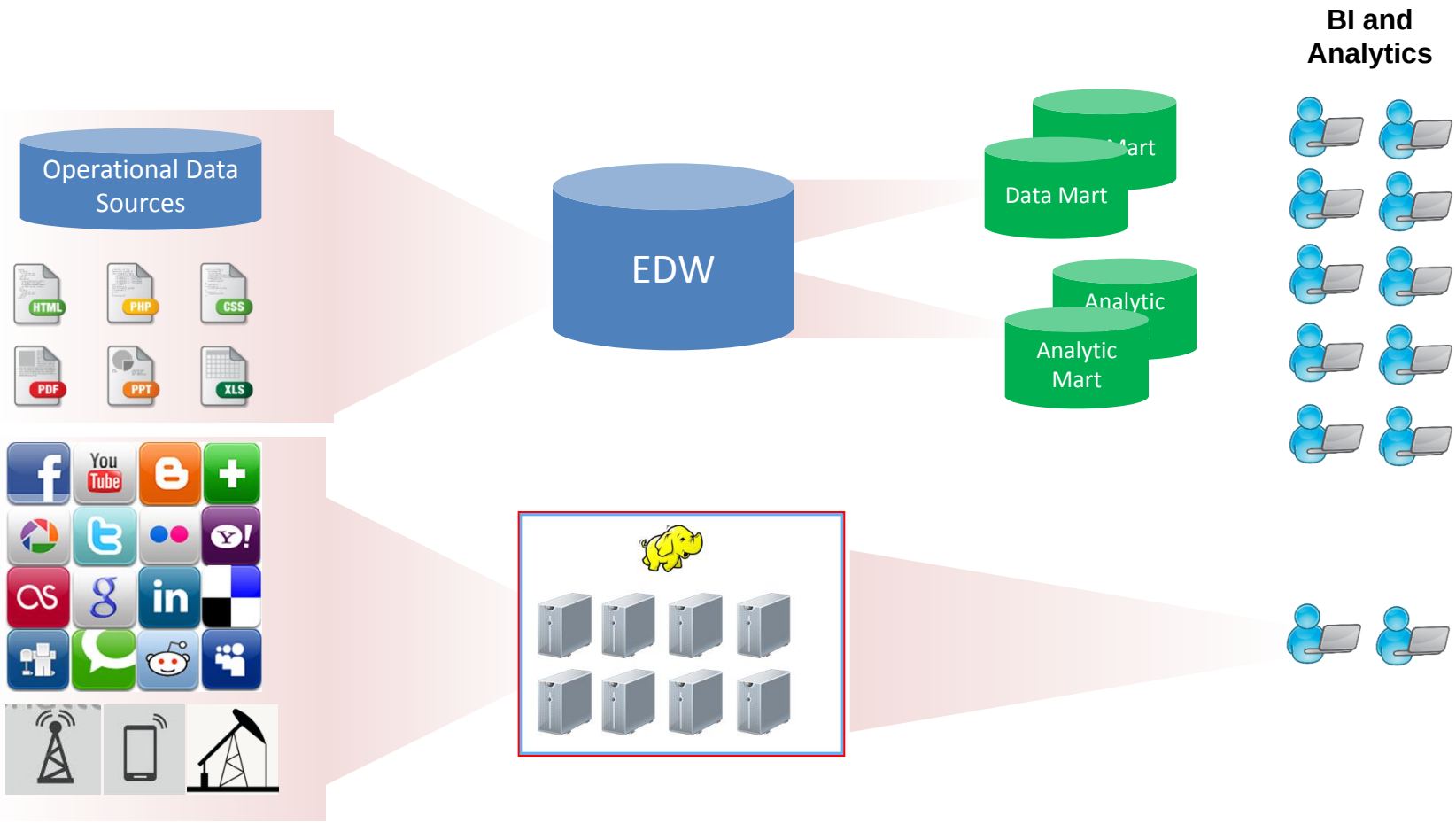
Architectures of Big Data Analytics

Traditional Analytics



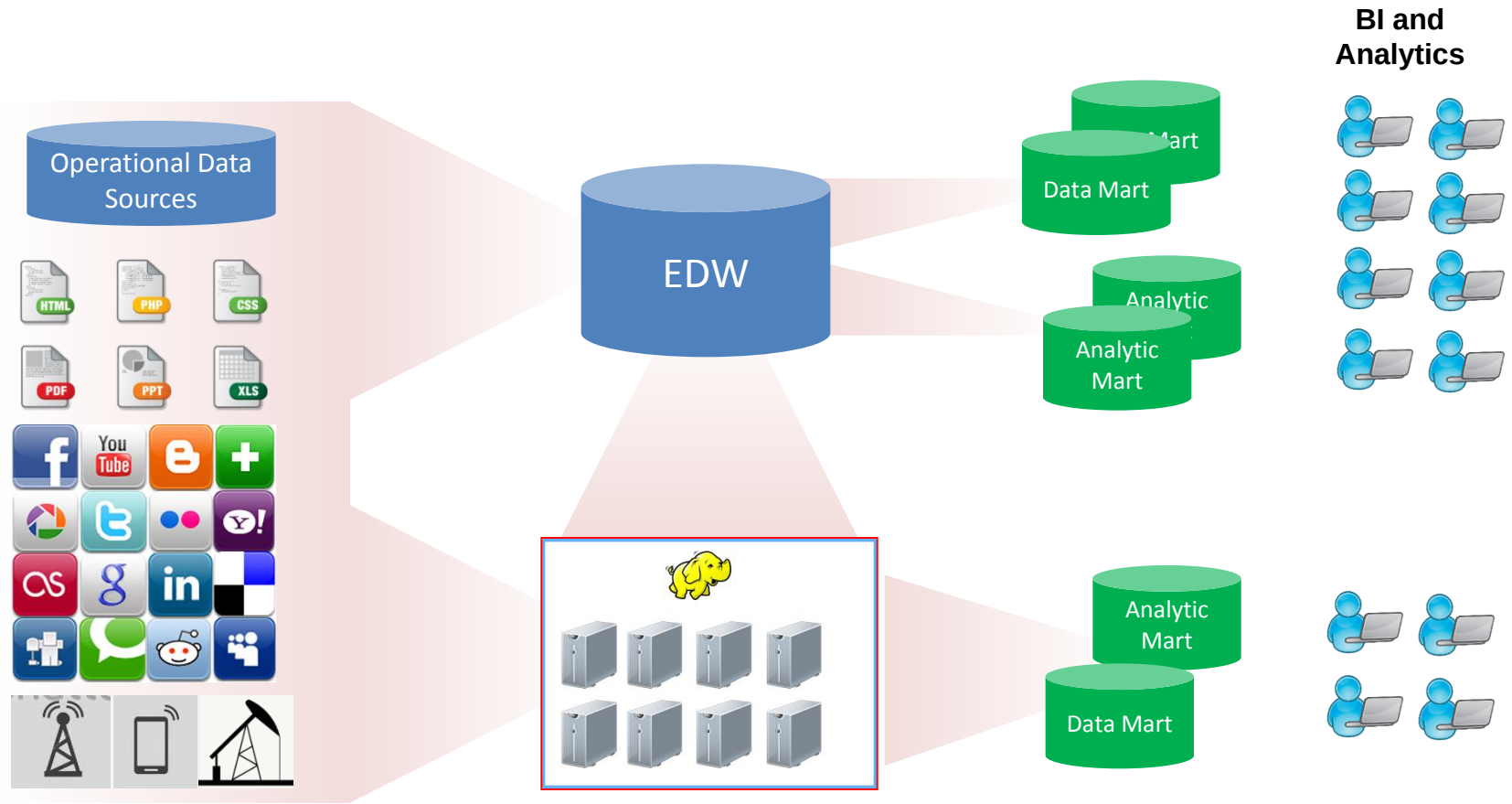
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

Hadoop as a “new data” Store



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

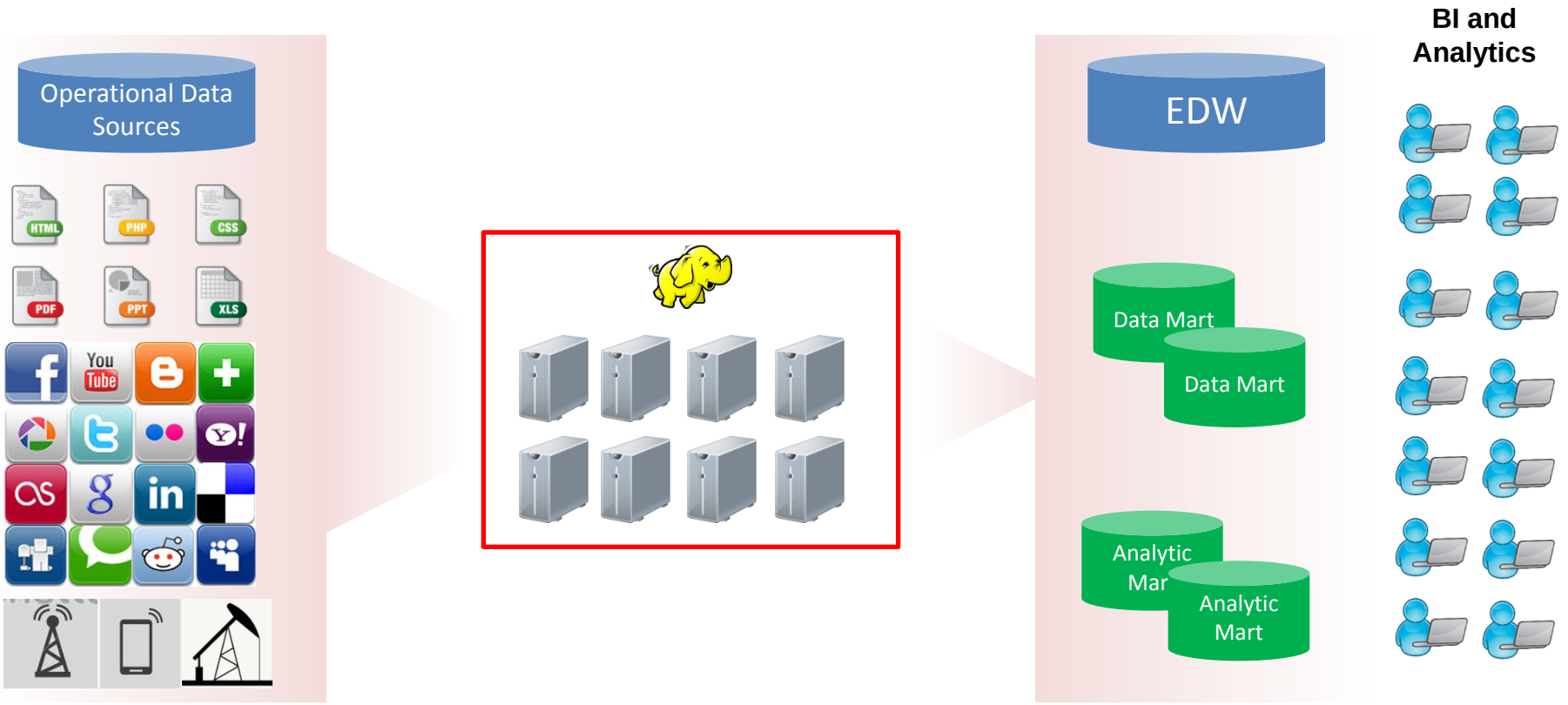
Hadoop as an additional input to the EDW



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

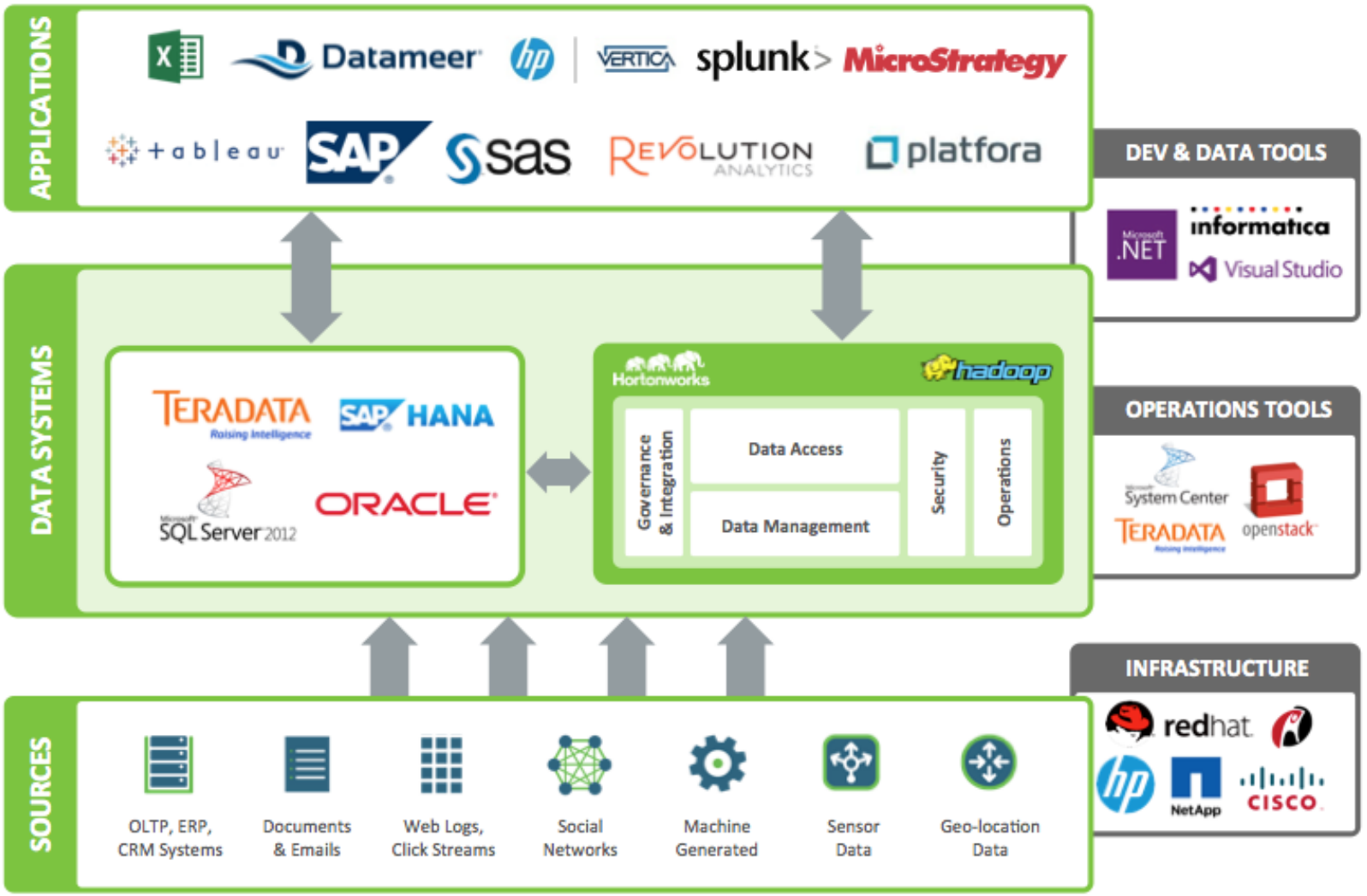
Hadoop Data Platform As a "staging Layer" as part of a "data Lake"

– Downstream stores could be Hadoop, data appliances or an RDBMS



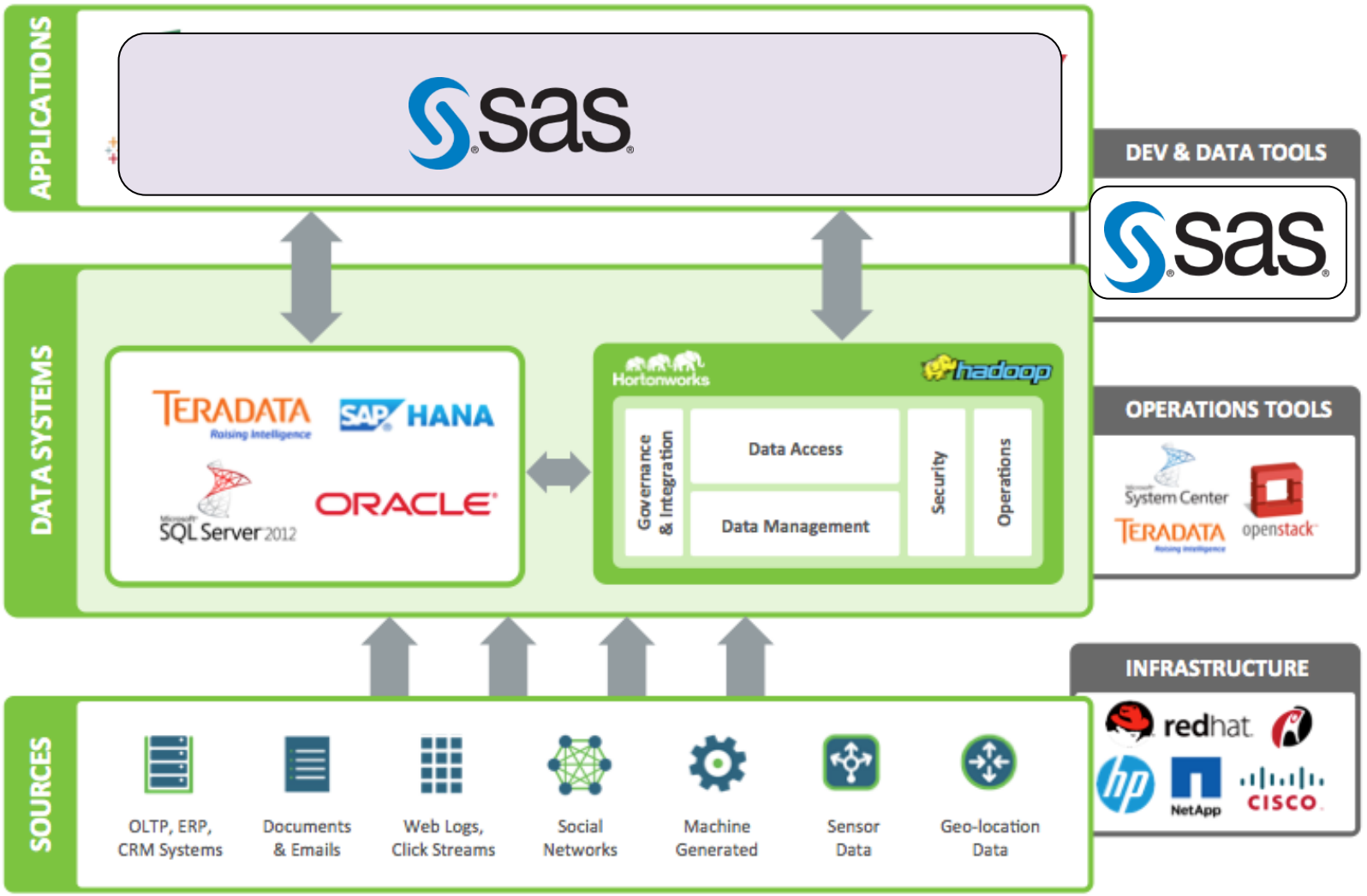
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



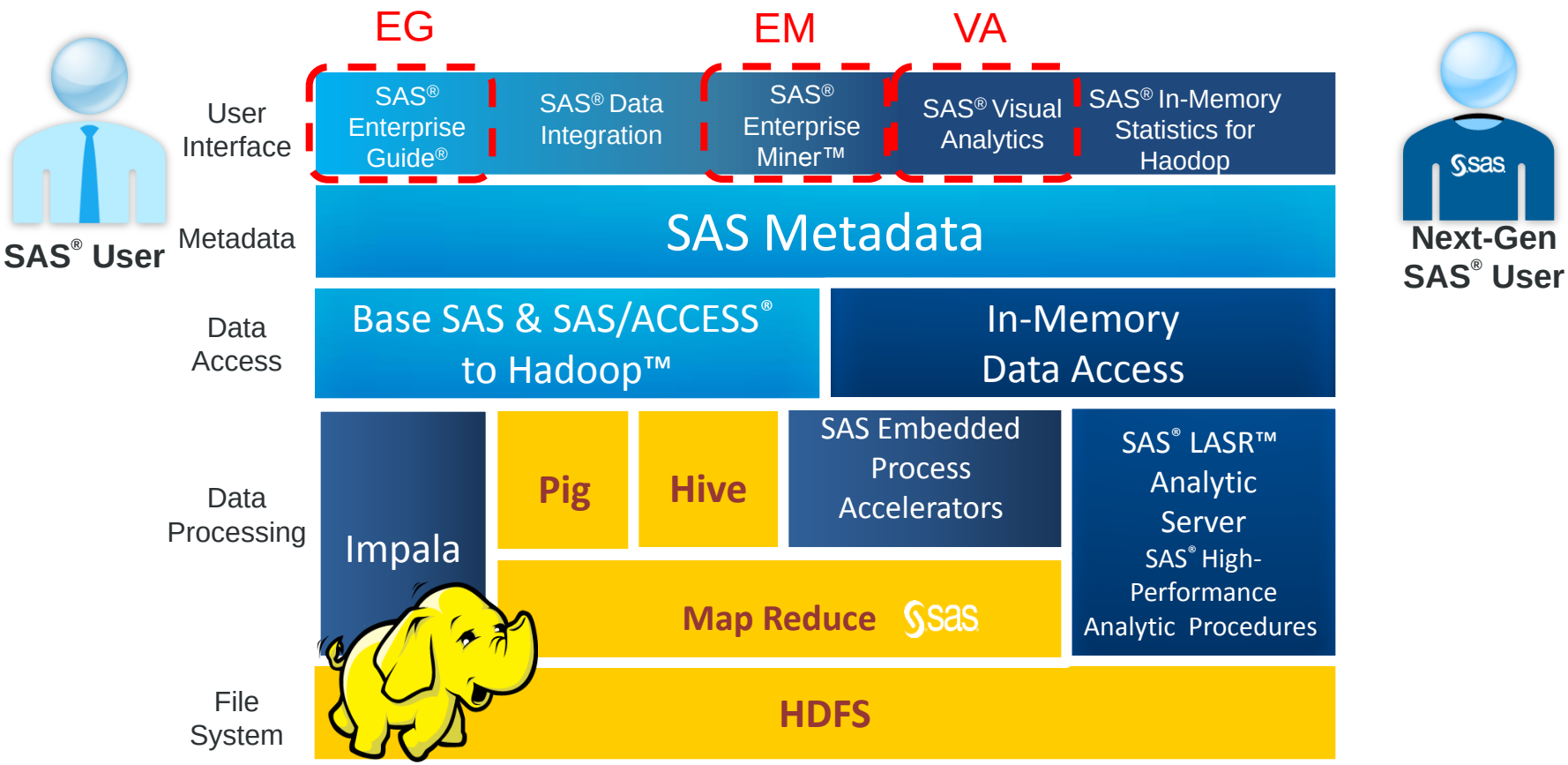
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



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

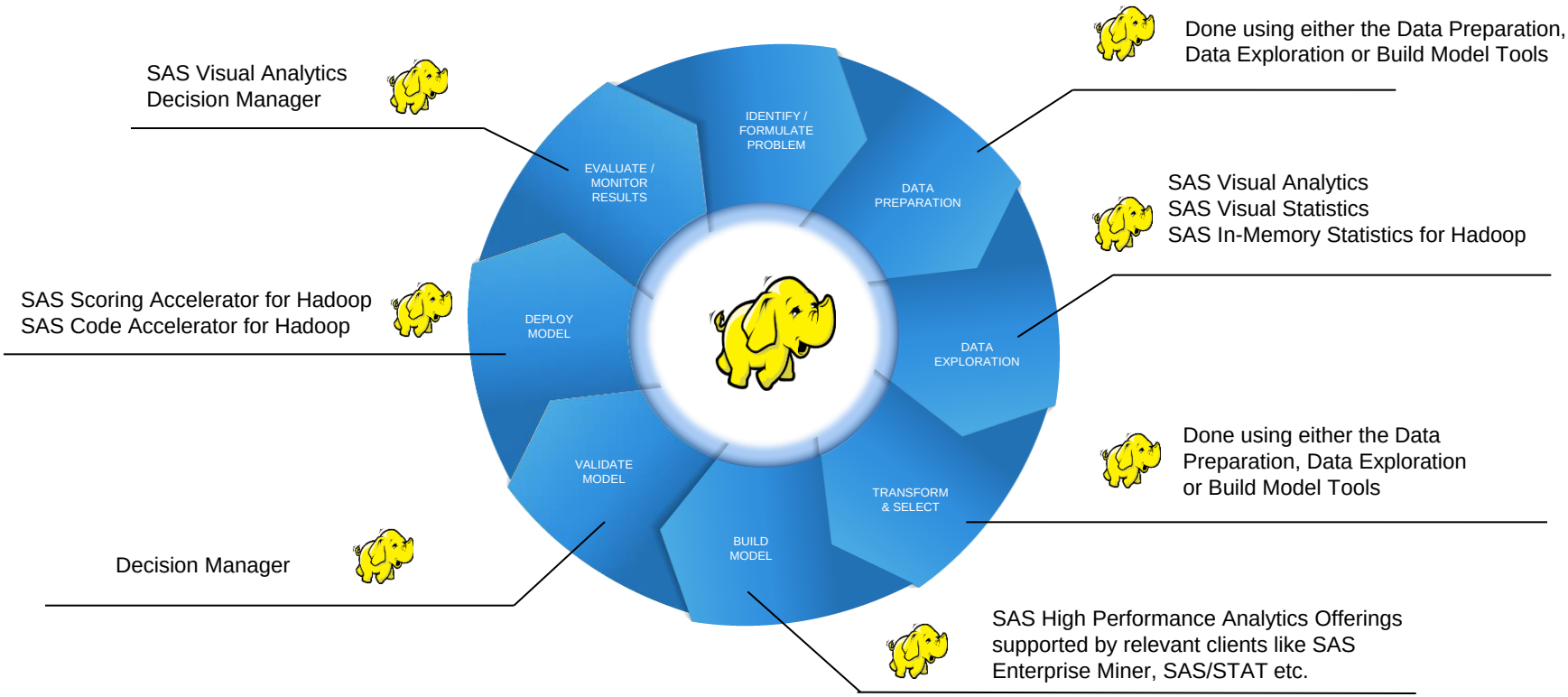
SAS® Within the HADOOP ECOSYSTEM



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

SAS enables the entire lifecycle around HADOOP

SAS enableS the entire lifecycle around HADOOP



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

SAS® VISUAL ANALYTICS

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

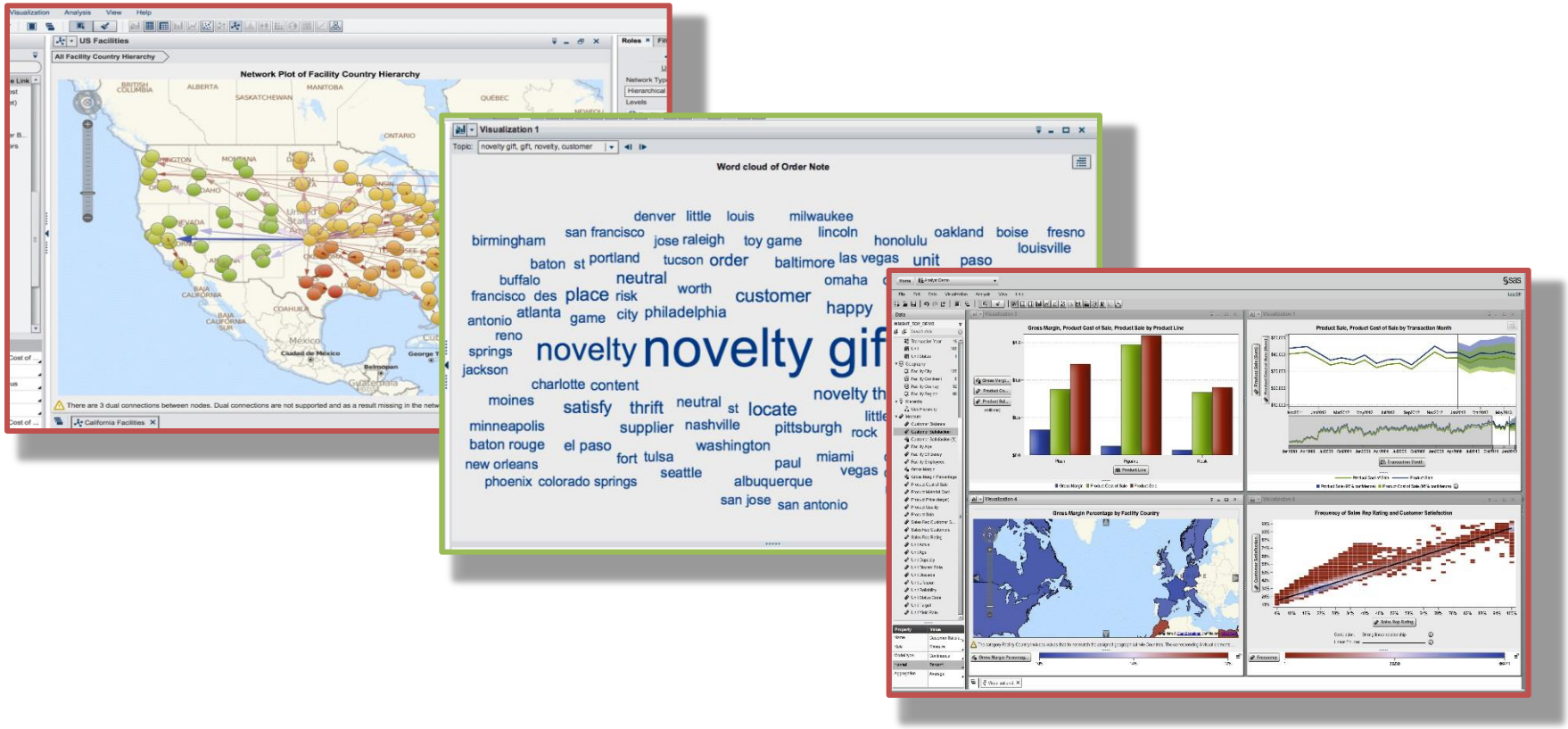
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS® VISUAL ANALYTICS

Example: text analysis gives you insight to customer experience and opinion

Analytics applied to text provides real MEANING

Visualization



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

References

- EMC Education Services (2015),
  Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley
- Shiva Achari (2015),
  Hadoop Essentials - Tackling the Challenges of Big Data with Hadoop, Packt Publishing
- Mike Frampton (2015),
  Mastering Apache Spark, Packt Publishing
- Deepak Ramanathan (2014),
  SAS Modernization architectures - Big Data Analytics,
  http://www.slideshare.net/deepakramanathan/sas-modernization-architectures-big-data-analytics