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)

1052DM02
MI4 (M2244) (3069)
Thu, 8, 9 (15:10-17:00) (B130)

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

2017-02-23

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2017/02/16	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
2	2017/02/23	巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統 (Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)
3	2017/03/02	關連分析 (Association Analysis)
4	2017/03/09	分類與預測 (Classification and Prediction)
5	2017/03/16	分群分析 (Cluster Analysis)
6	2017/03/23	個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis – K-Means using SAS EM)
7	2017/03/30	個案分析與實作二 (SAS EM 關連分析)：Case Study 2 (Association Analysis using SAS EM)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
8	2017/04/06	教學行政觀摩日 (Off-campus study)
9	2017/04/13	期中報告 (Midterm Project Presentation)
10	2017/04/20	期中考試週 (Midterm Exam)
11	2017/04/27	個案分析與實作三 (SAS EM 決策樹、模型評估)：Case Study 3 (Decision Tree, Model Evaluation using SAS EM)
12	2017/05/04	個案分析與實作四 (SAS EM 迴歸分析、類神經網路)：Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)
13	2017/05/11	Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)
14	2017/05/18	期末報告 (Final Project Presentation)
15	2017/05/25	畢業班考試 (Final Exam)

2017/02/23

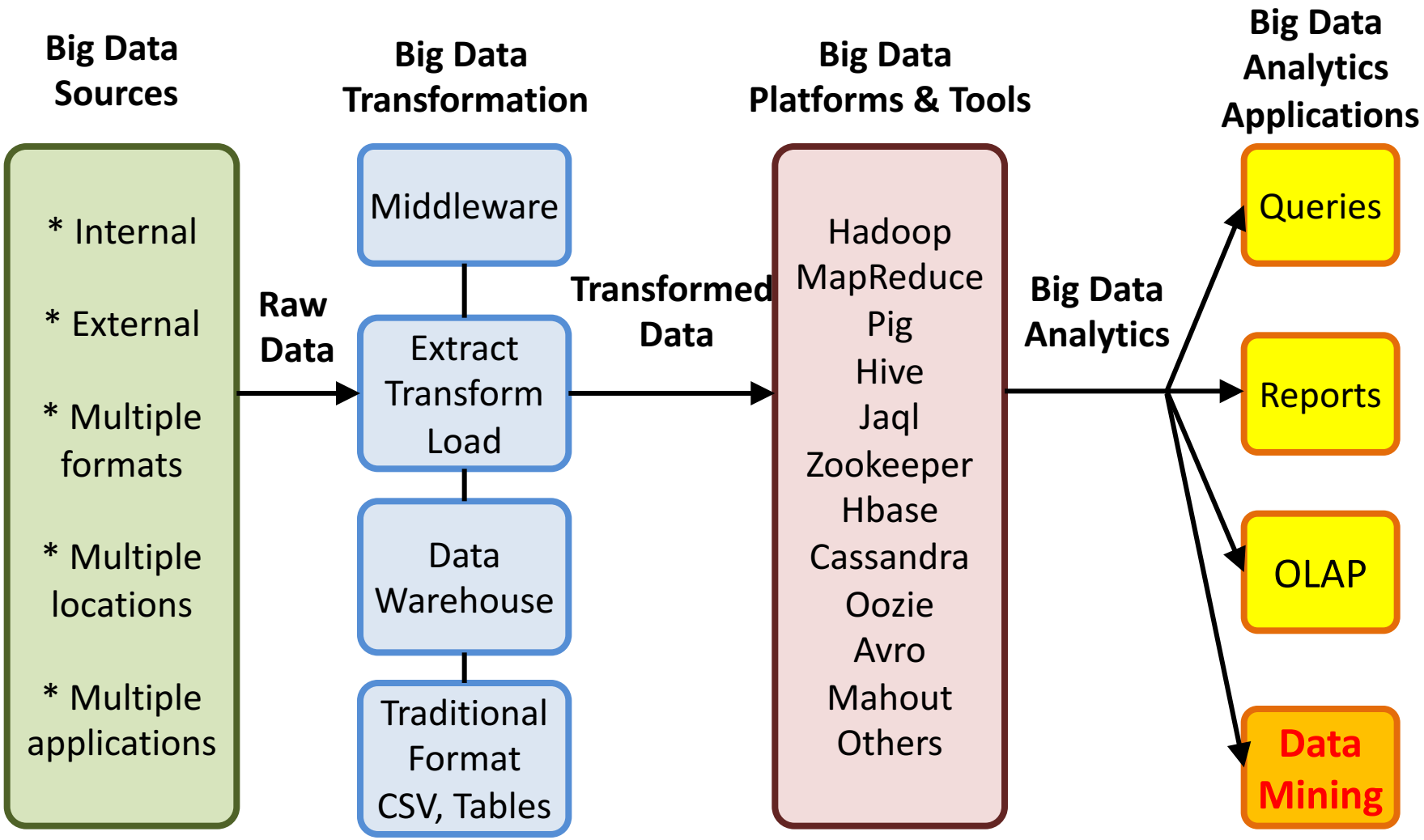
巨量資料基礎：
MapReduce典範、Hadoop與Spark生態系統

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

Big Data Analytics and Data Mining

Architectures 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

Architecture of Big Data Analytics

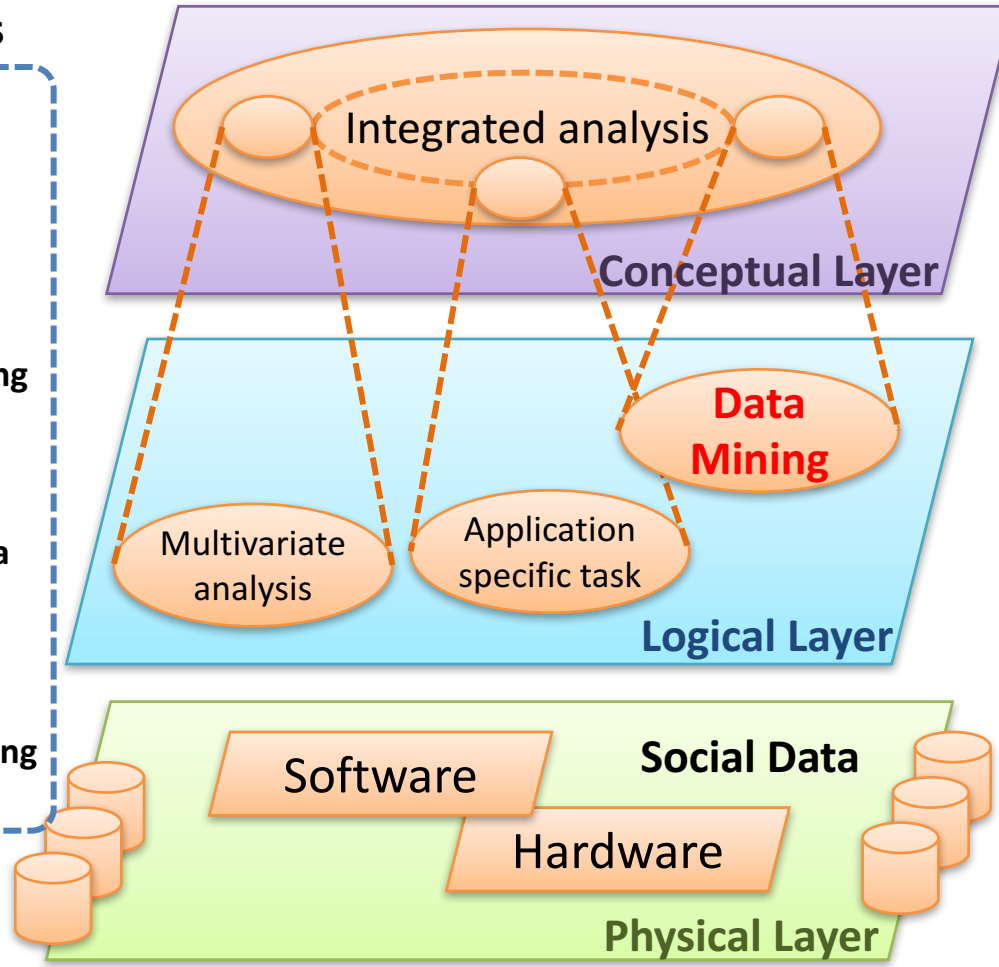
Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications

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**

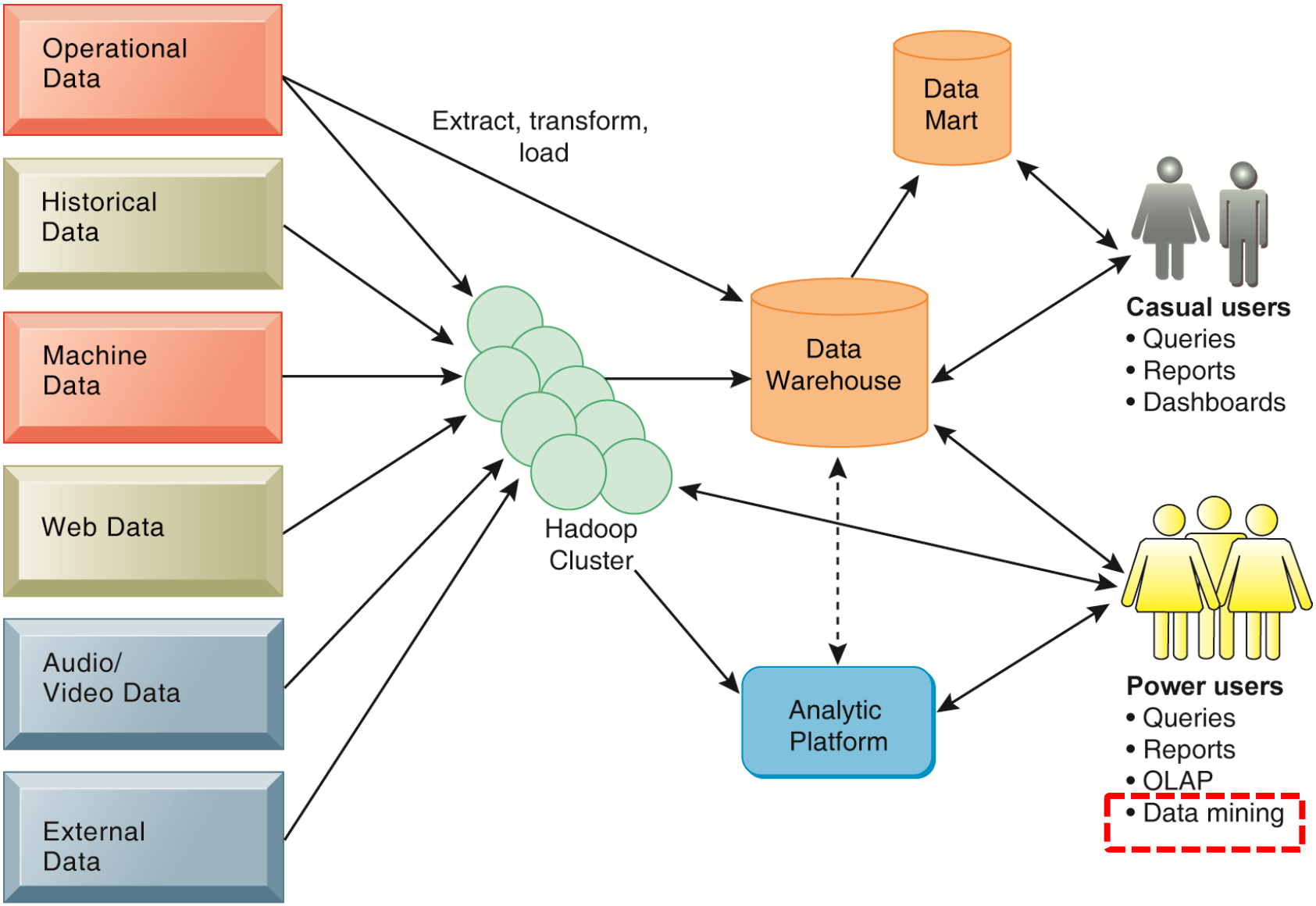


Analysts

- **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

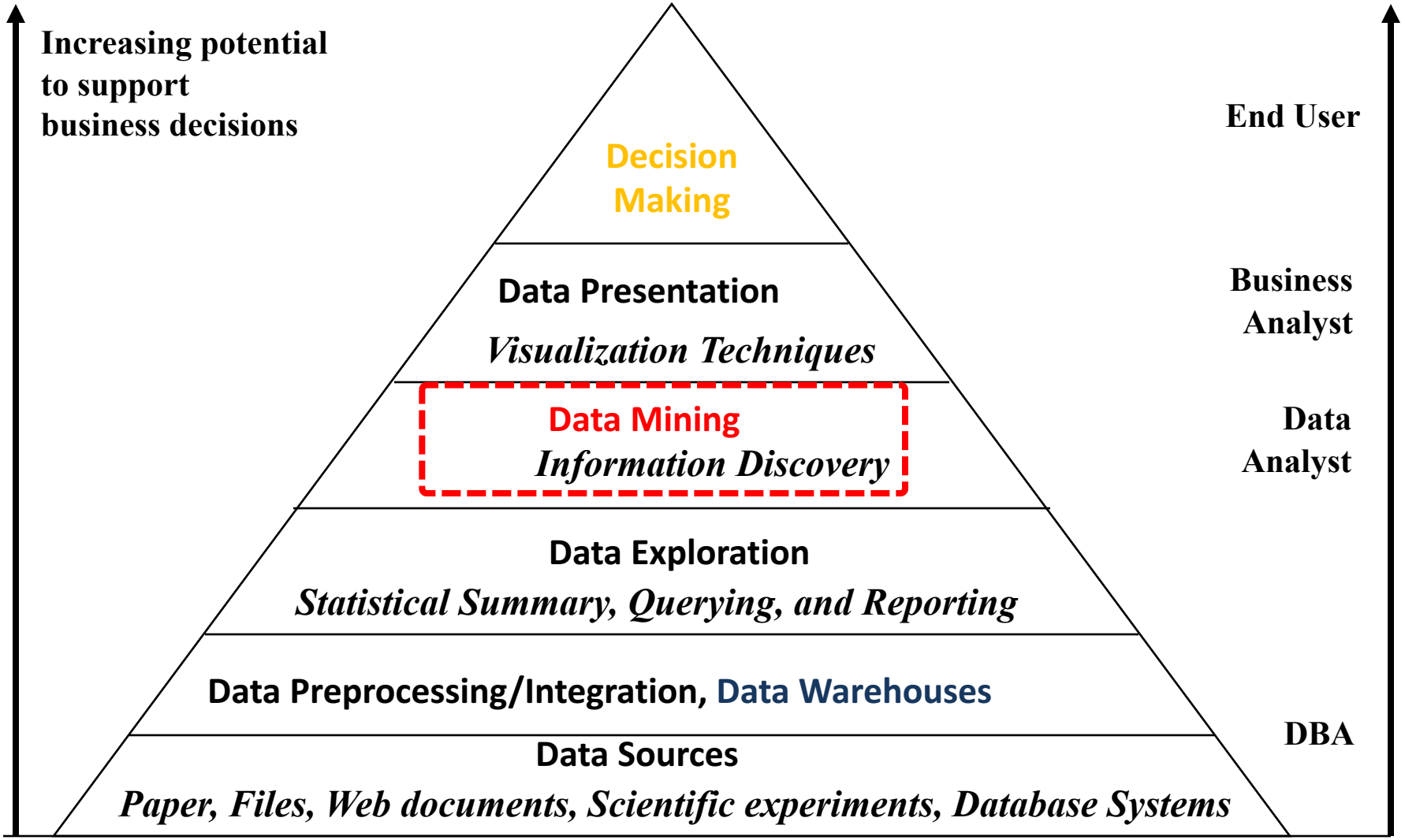
Business Intelligence (BI) Infrastructure



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

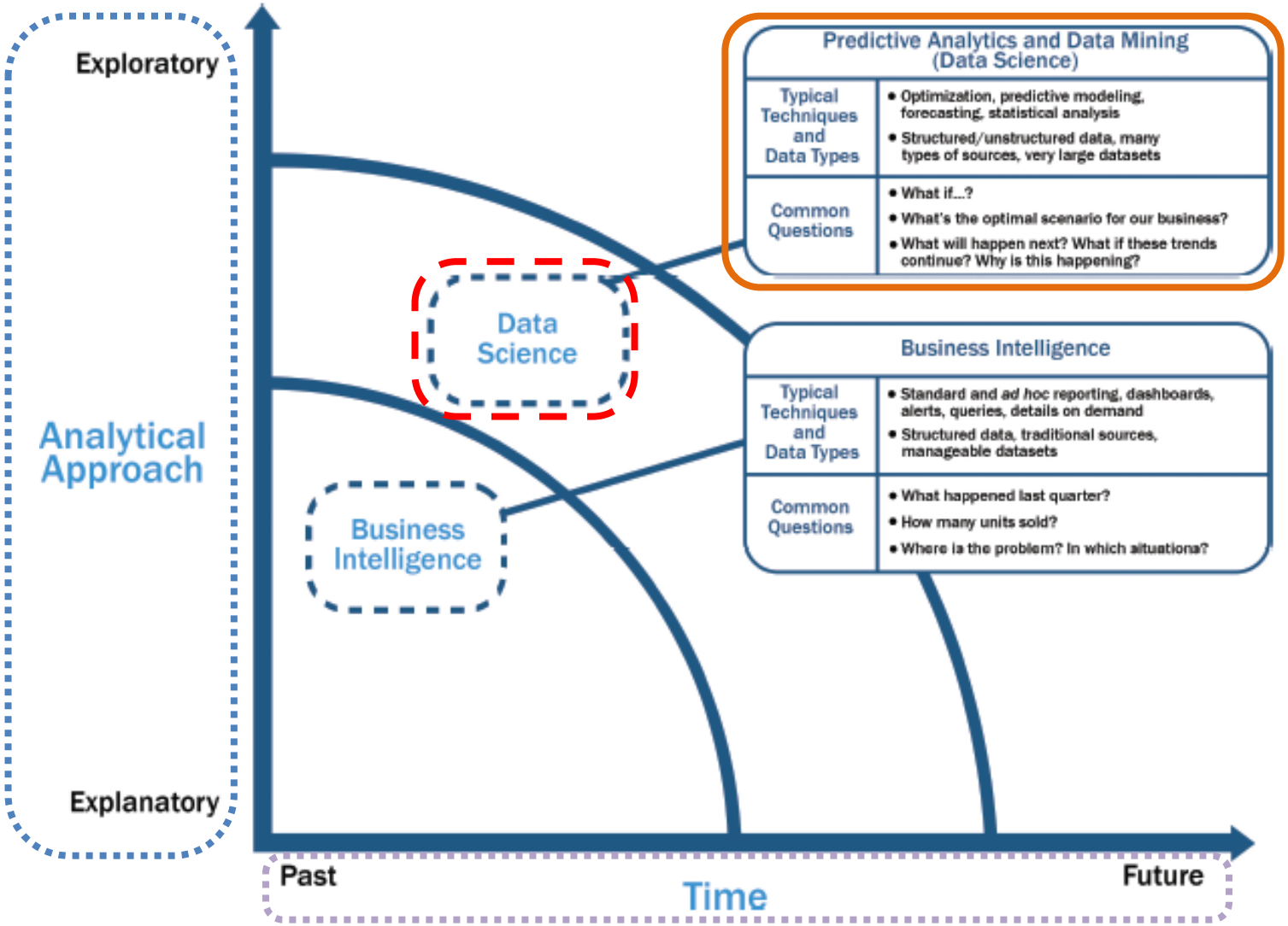


Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

The Evolution of BI Capabilities

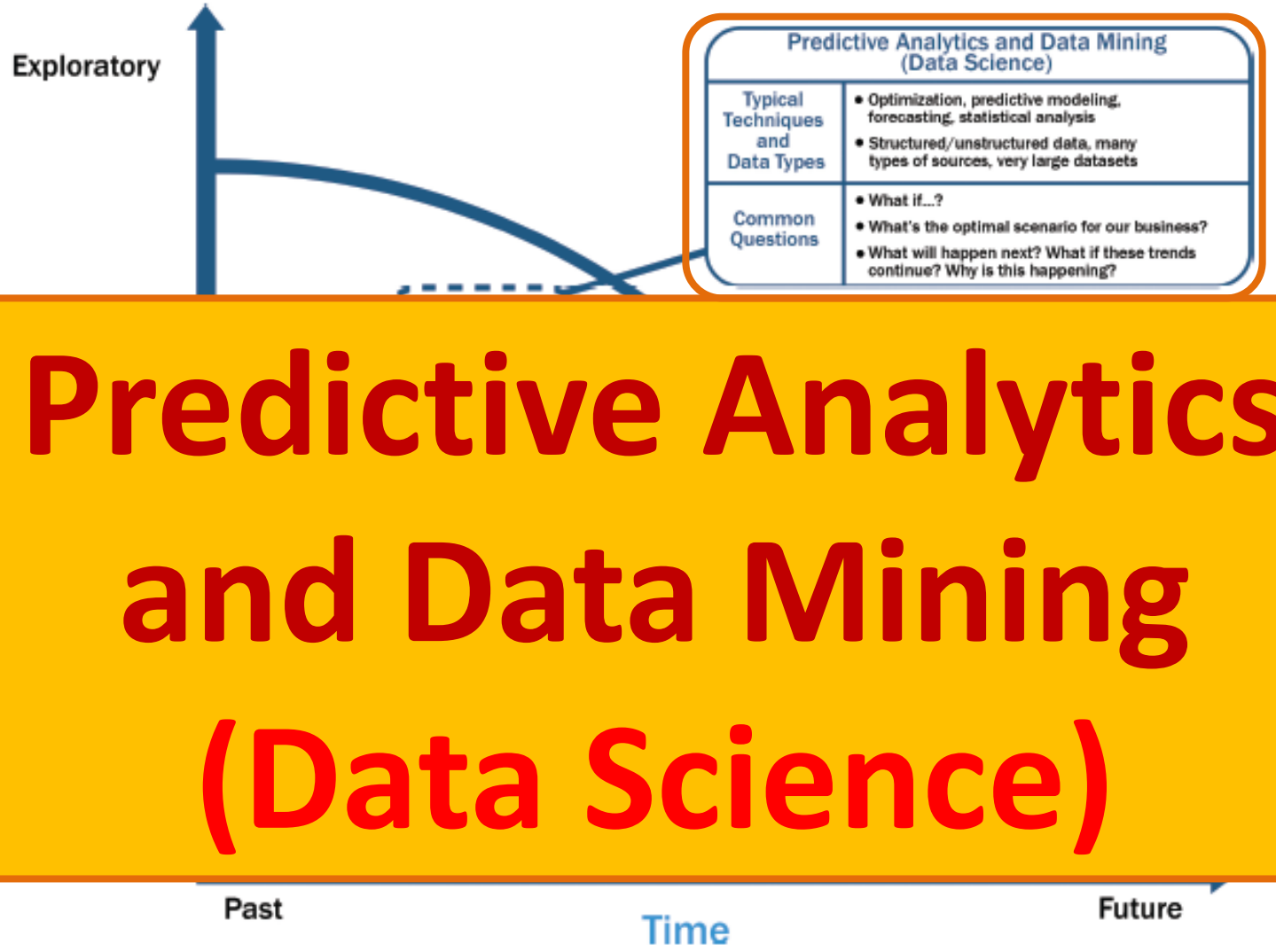
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Science and Business Intelligence



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

Data Science and Business Intelligence



Predictive Analytics and Data Mining (Data Science)

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

Predictive Analytics and Data Mining (Data Science)

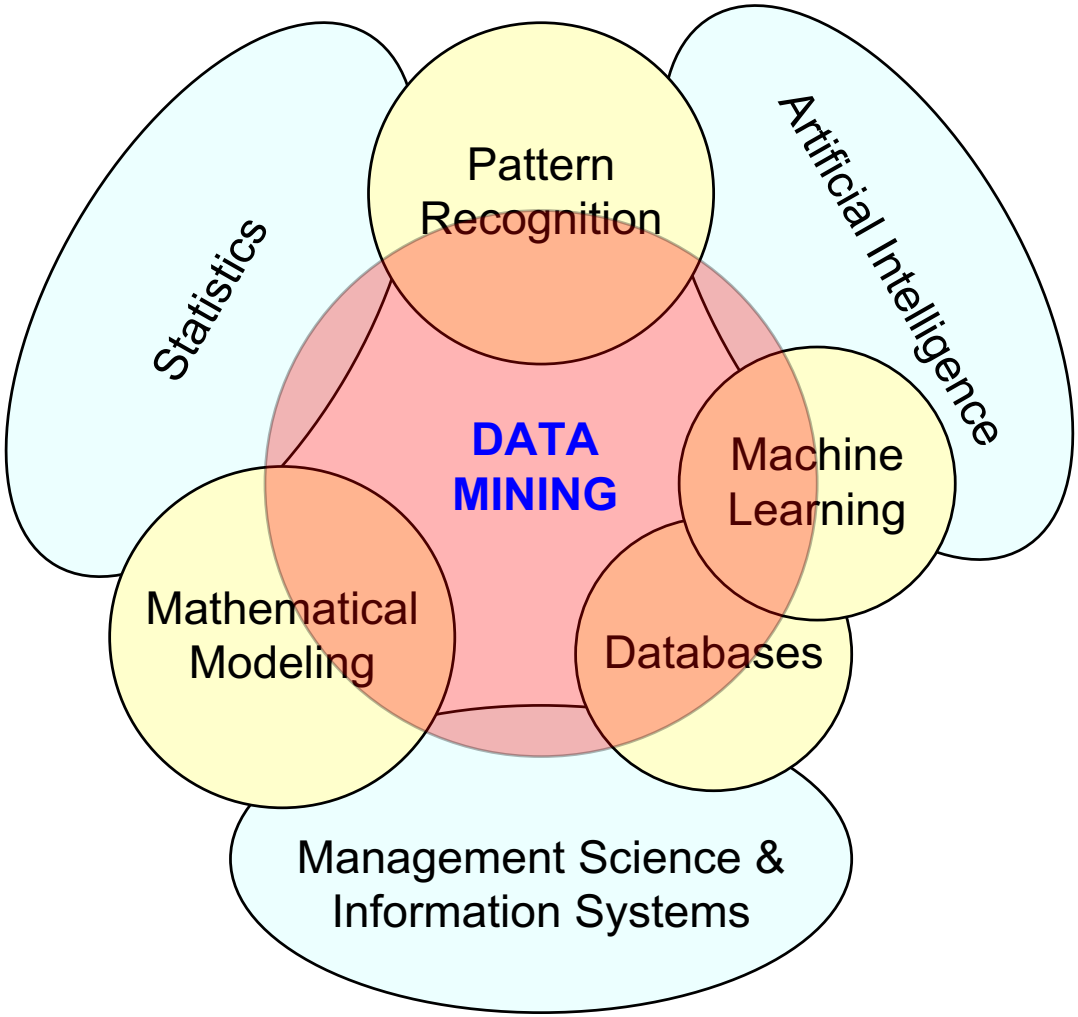
Structured/unstructured data, many types of sources, very large datasets

Optimization, predictive modeling, forecasting statistical analysis

What if…?
What's the optimal scenario for our business?
What will happen next?
What if these trends countinue?
Why is this happening?

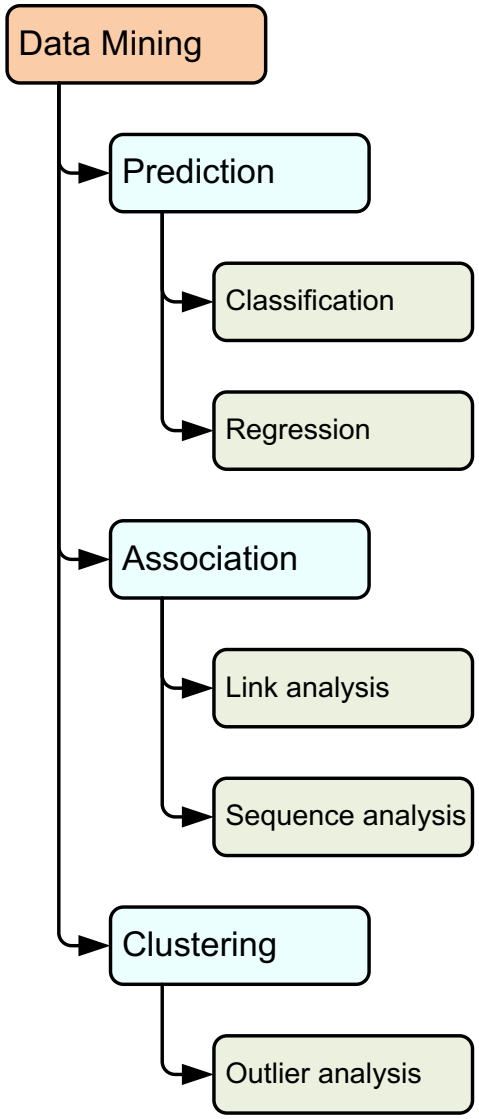
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Mining at the Intersection of Many Disciplines



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

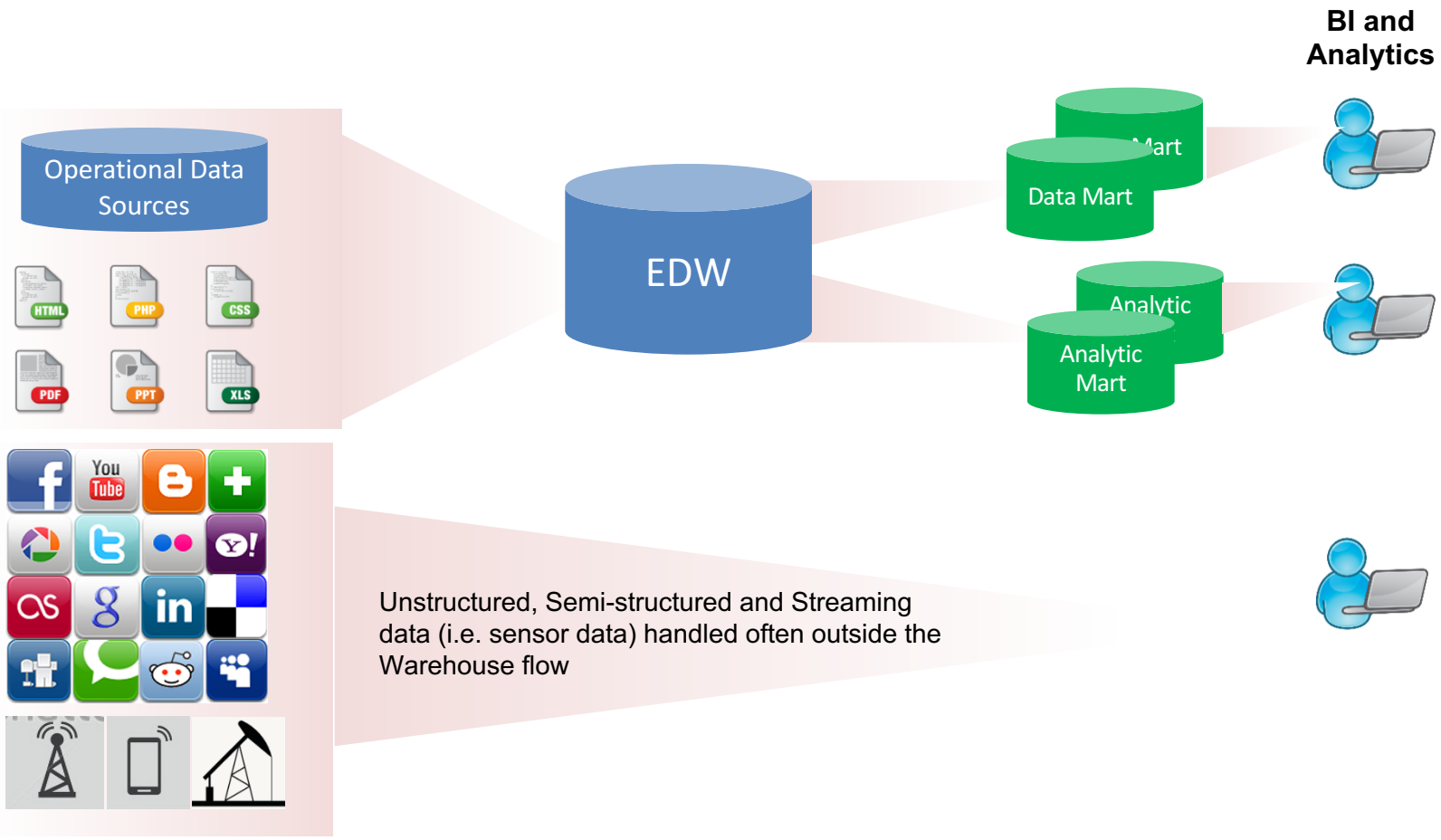
A Taxonomy for Data Mining Tasks



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)

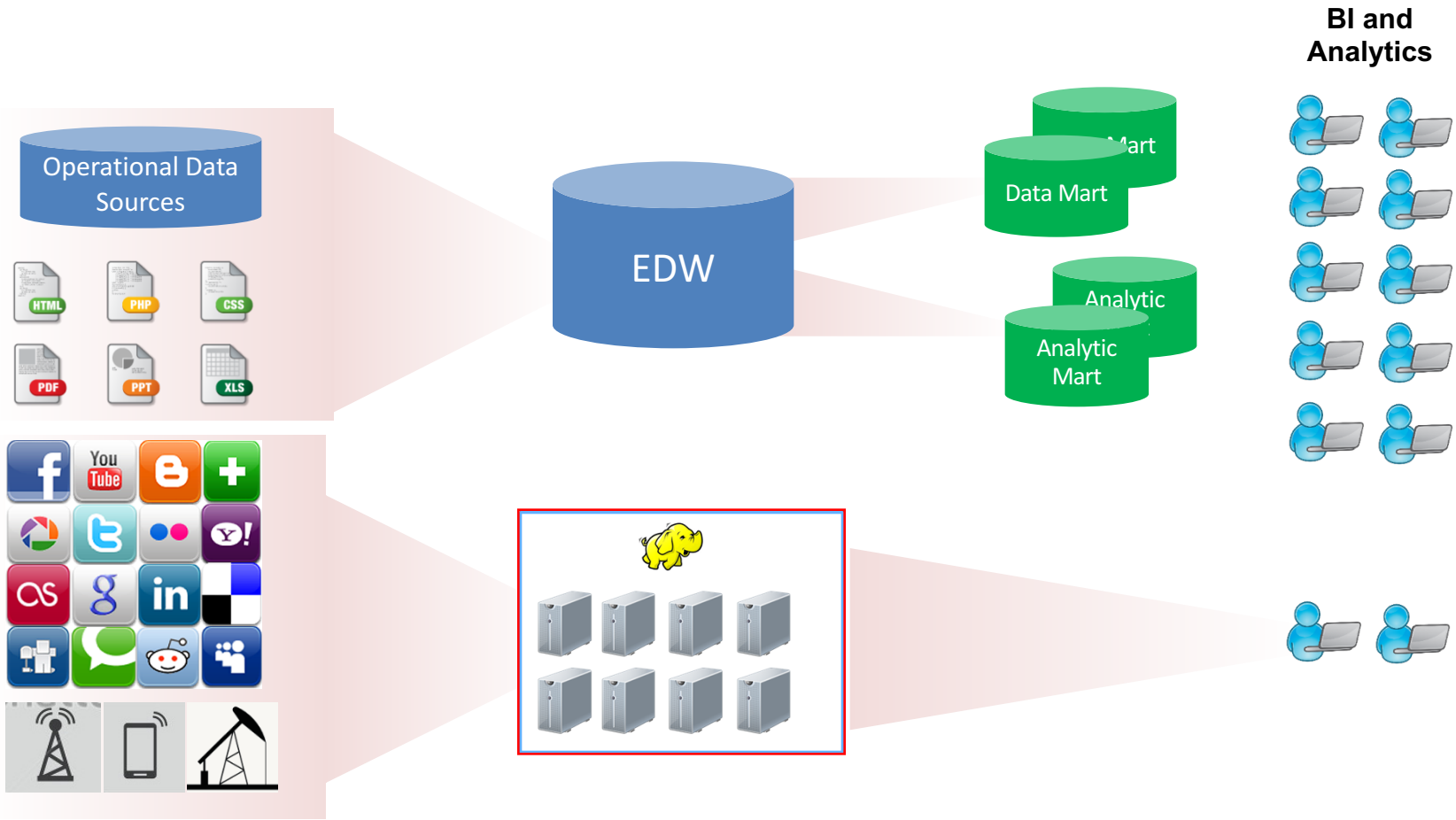
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Traditional Analytics



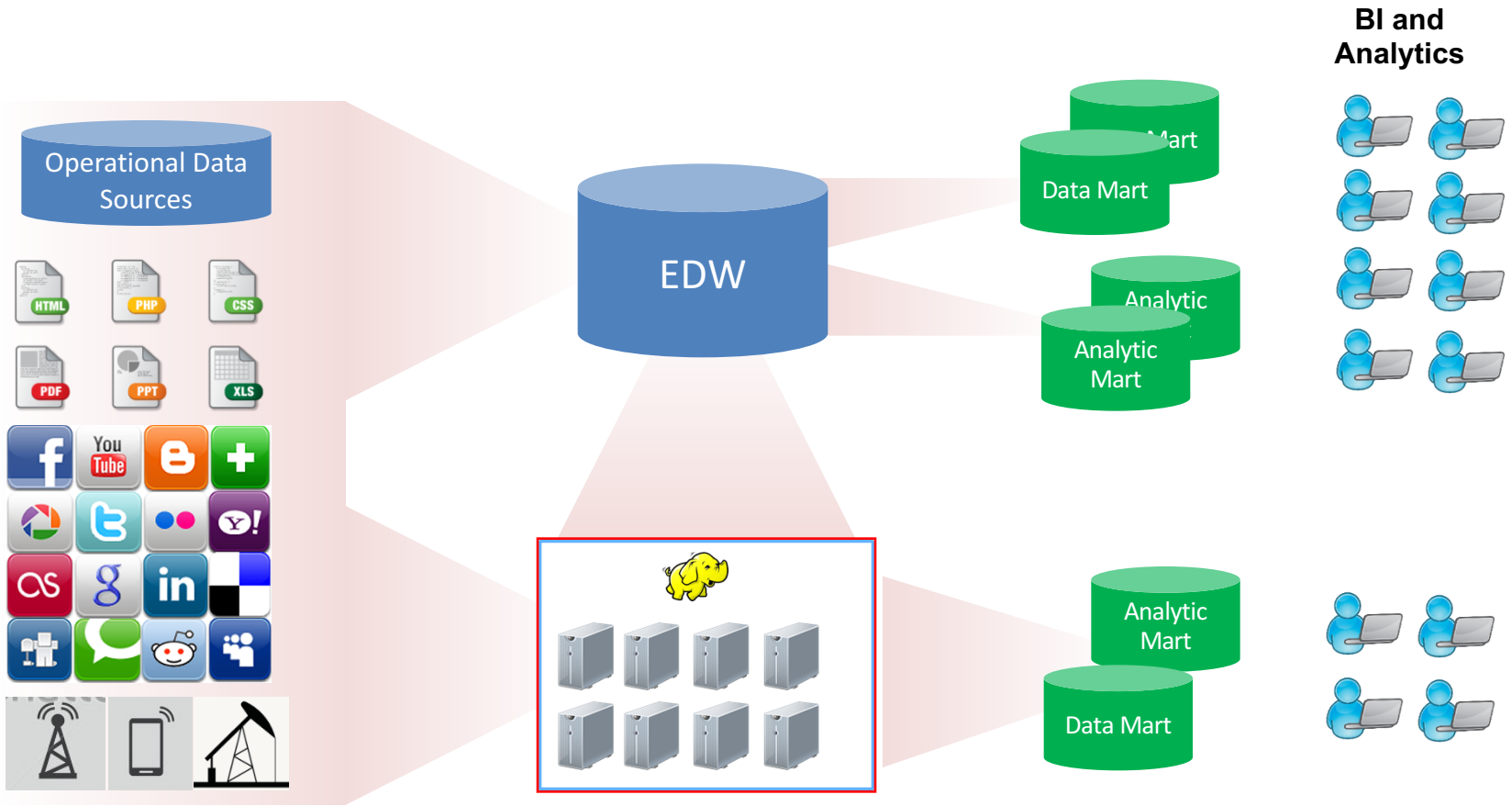
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

Hadoop as a “new data” Store



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

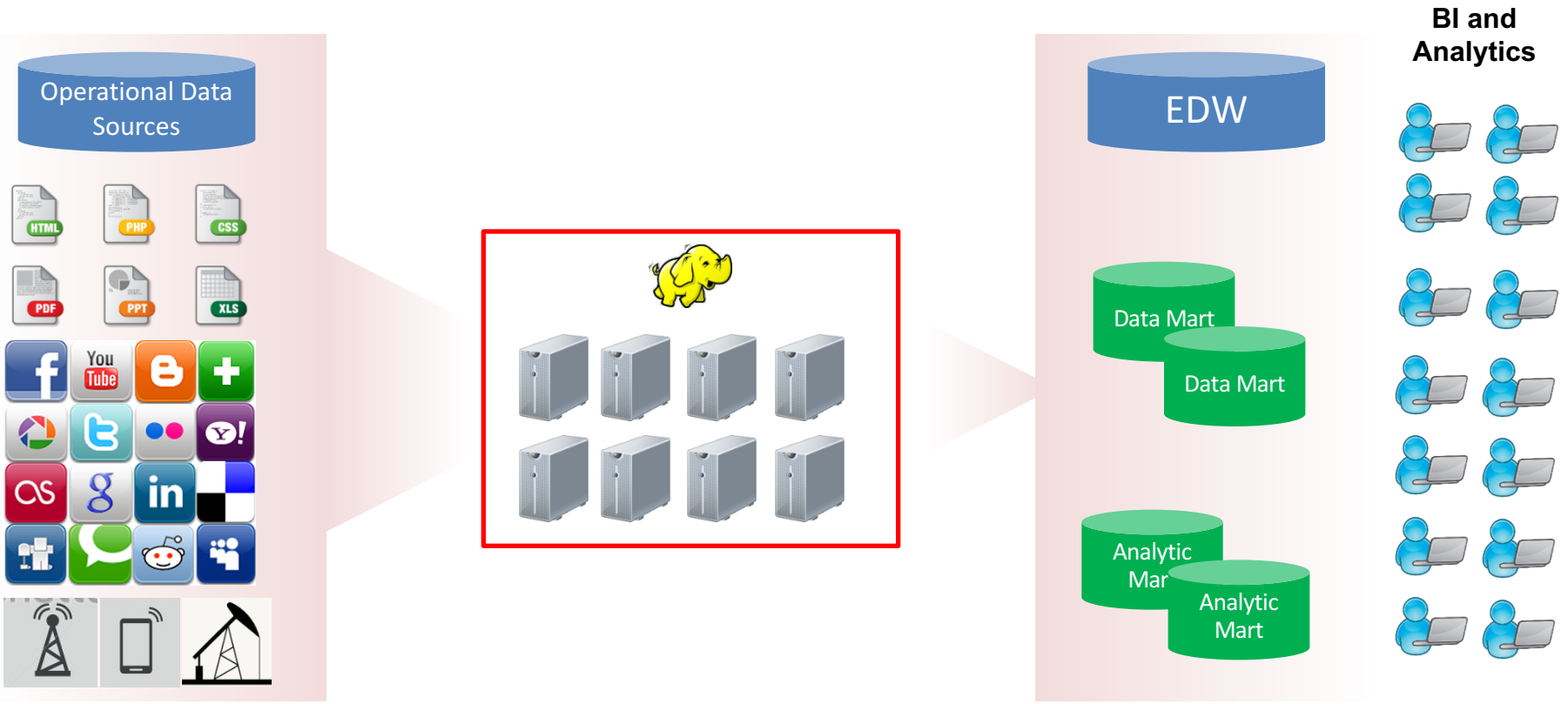
Hadoop as an additional input to the EDW



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

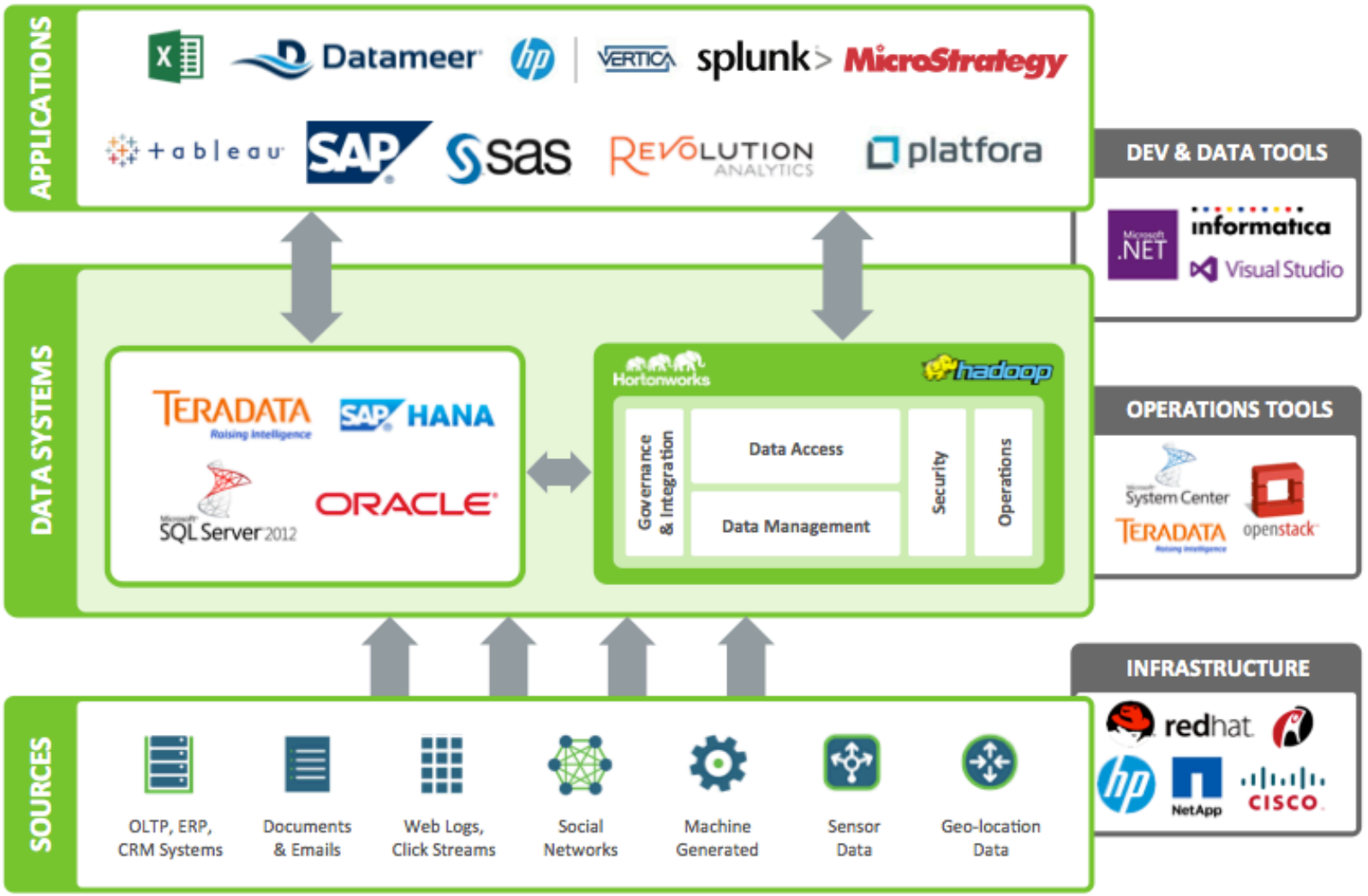
Hadoop Data Platform As a "staging Layer" as part of a "data Lake"

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



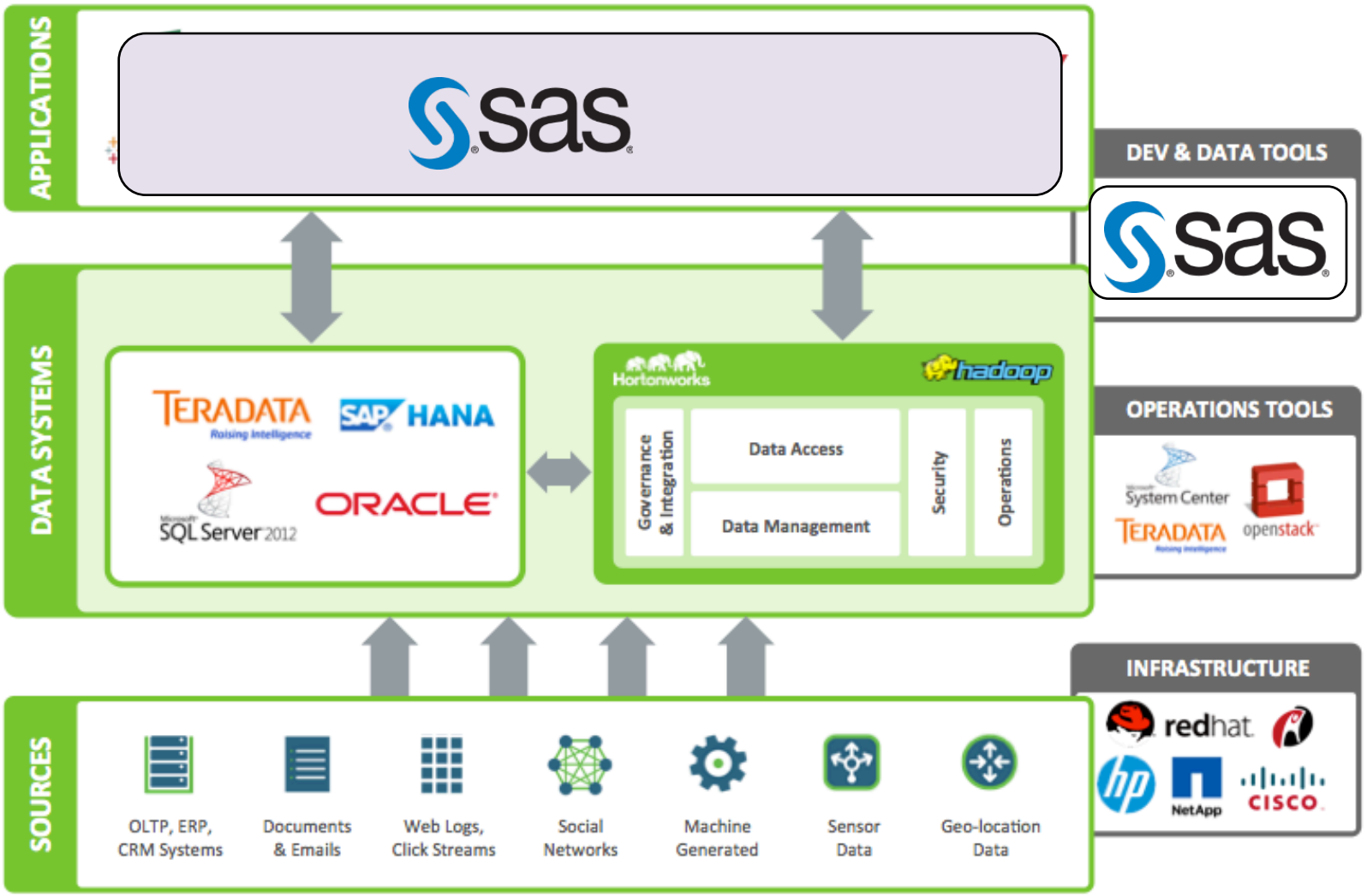
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



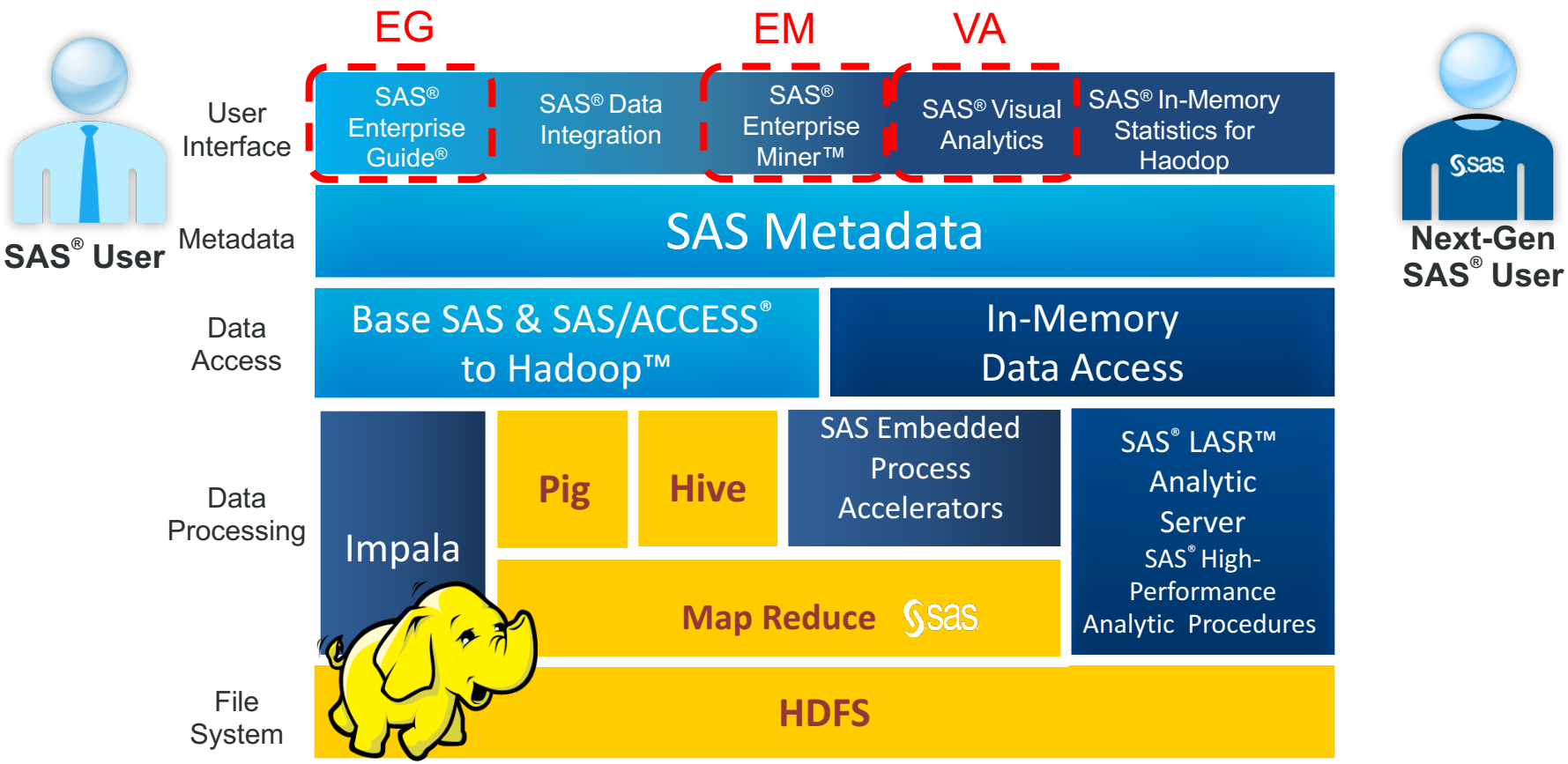
Source: Deepak Ramanathan (2014), SAS Modernization architectures - Big Data Analytics

SAS Big data Strategy – SAS areas



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

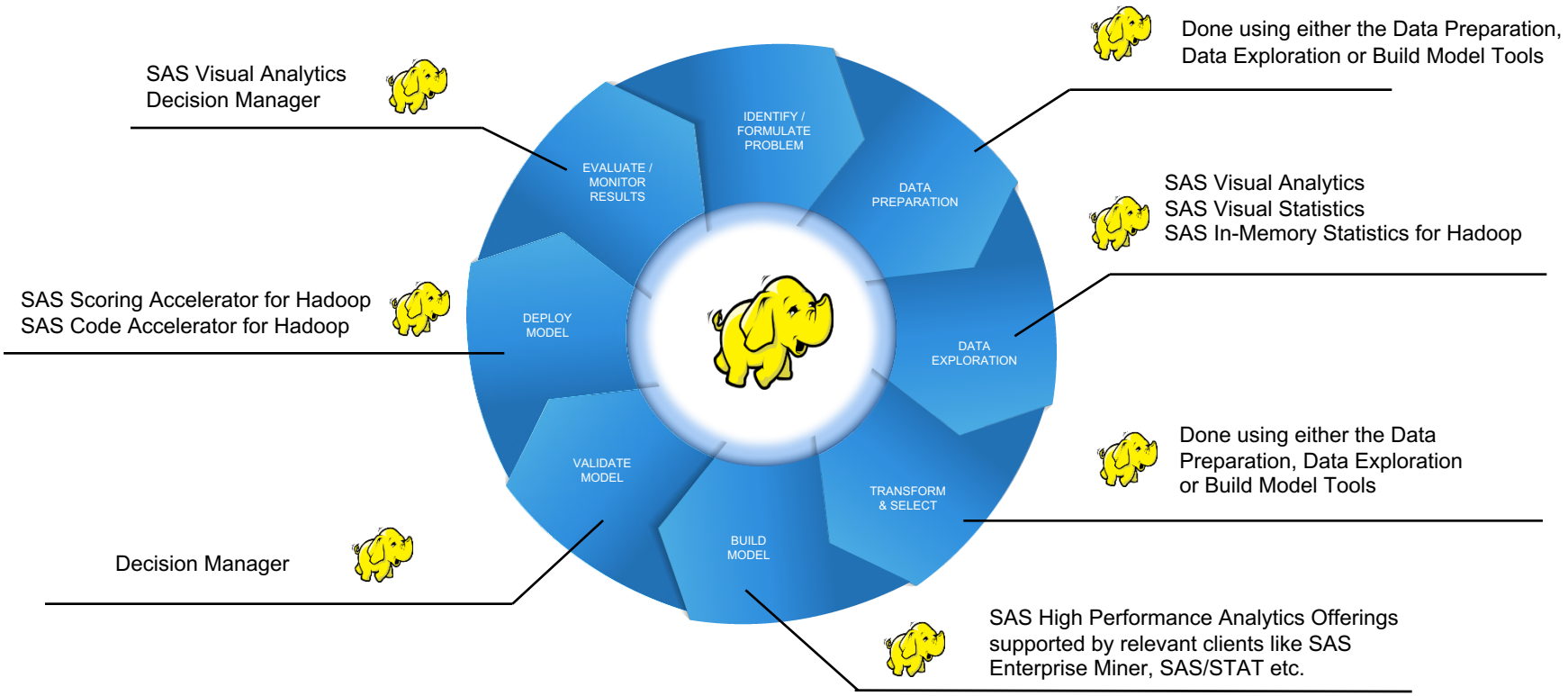
SAS® Within the HADOOP ECOSYSTEM



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

Data Mining Process

Data Mining Process

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard processes:
  - CRISP-DM
    (Cross-Industry Standard Process for Data Mining)
  - SEMMA
    (Sample, Explore, Modify, Model, and Assess)
  - KDD
    (Knowledge Discovery in Databases)

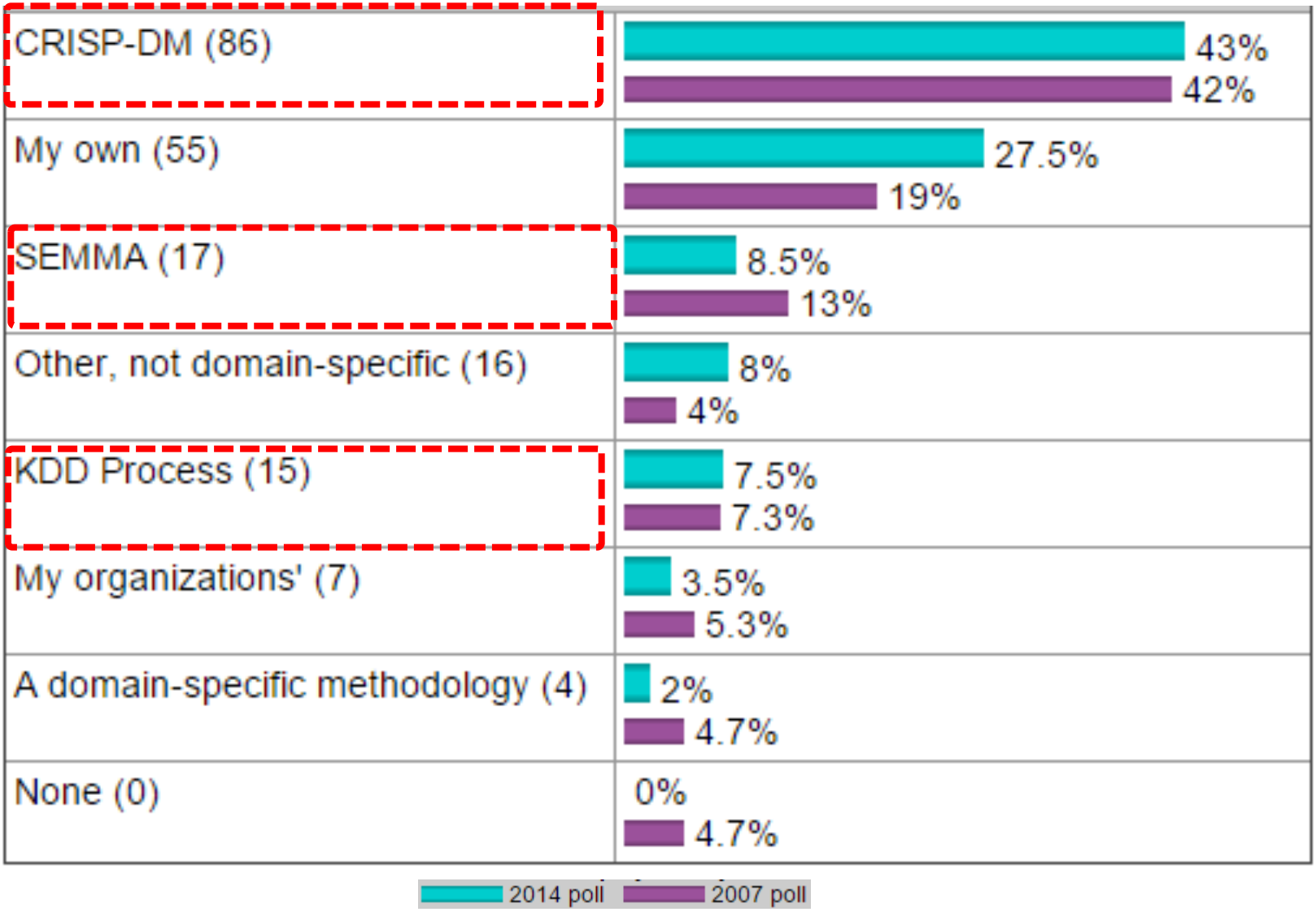
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Process (SOP of DM)

What main methodology are you using for your **analytics**, **data mining**, or **data science** projects ?

Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Data Mining Process



Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

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.

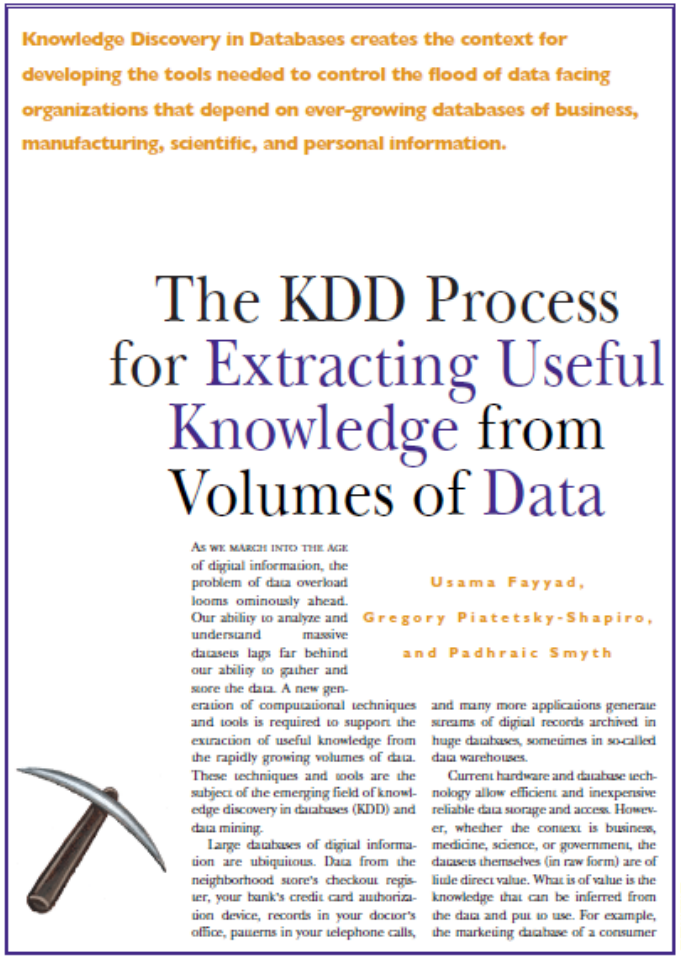
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

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.

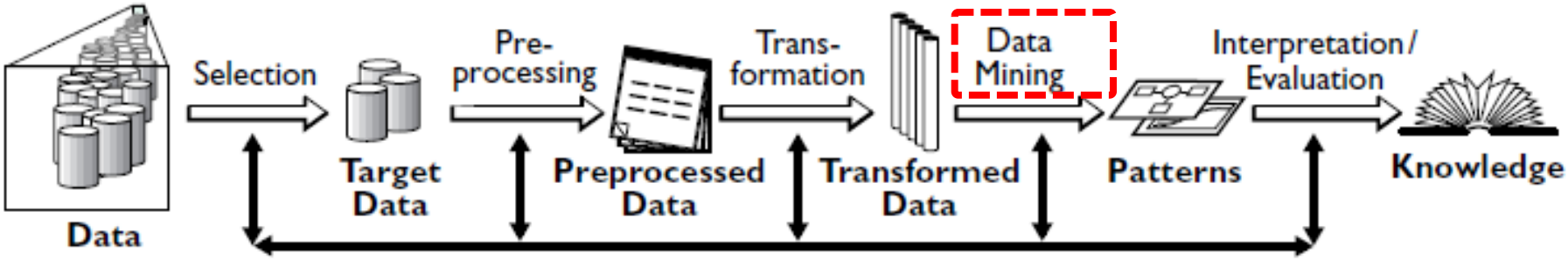
Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred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

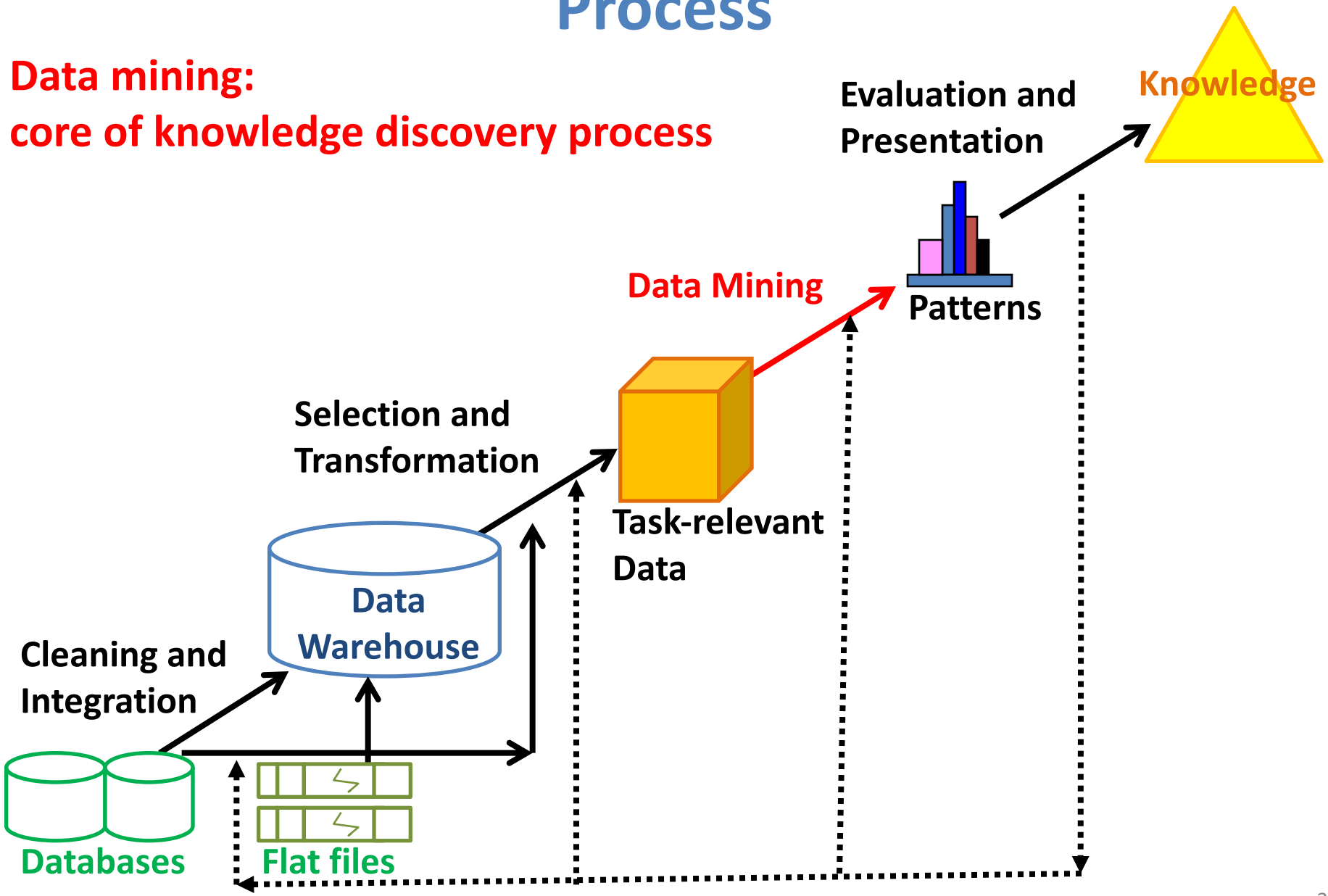
(Fayyad et al., 1996)



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.

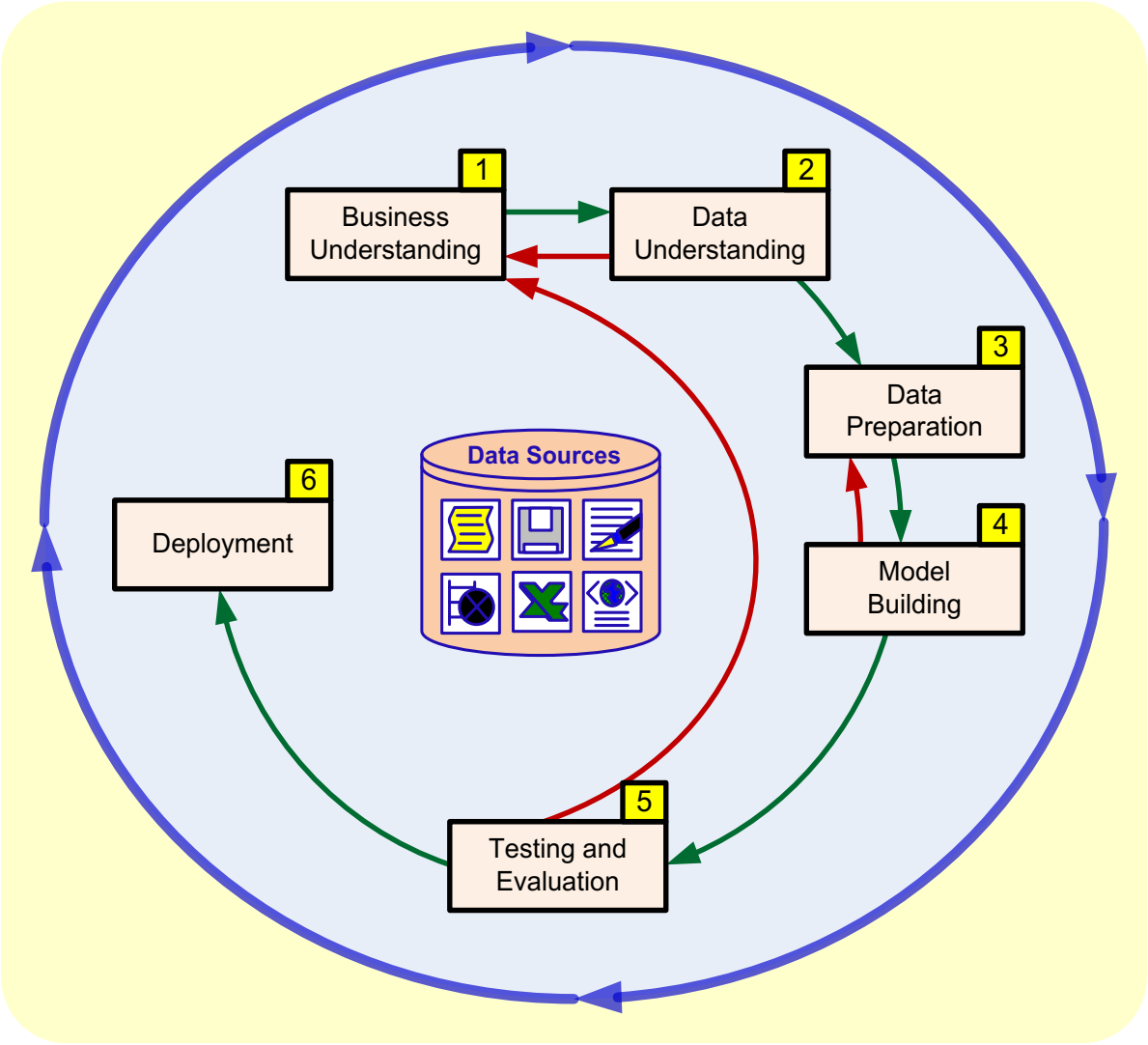
Knowledge Discovery in Databases (KDD) Process

Data mining:
core of knowledge discovery process



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

Data Mining Process:
CRISP-DM



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

Data Mining Process:
CRISP-DM

Step 1: Business Understanding

Step 2: Data Understanding

Step 3: Data Preparation (!)

Step 4: Model Building

Step 5: Testing and Evaluation

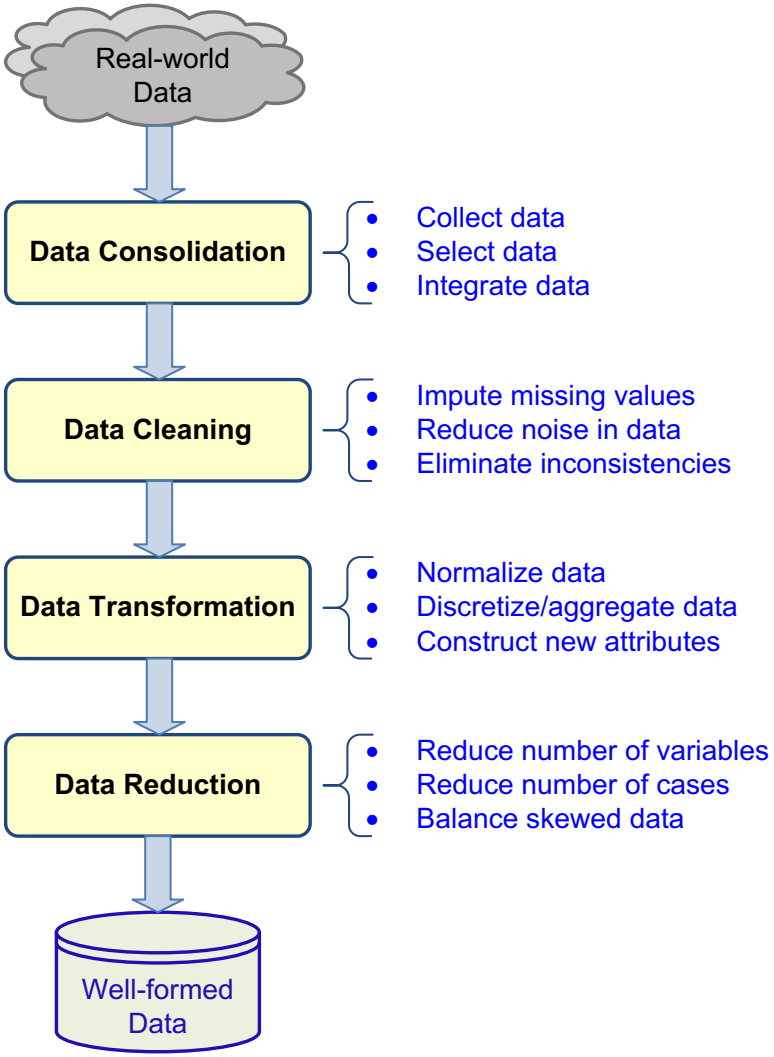
Step 6: Deployment

Accounts for ~85% of total project time

- The process is highly repetitive and experimental (DM: art versus science?)

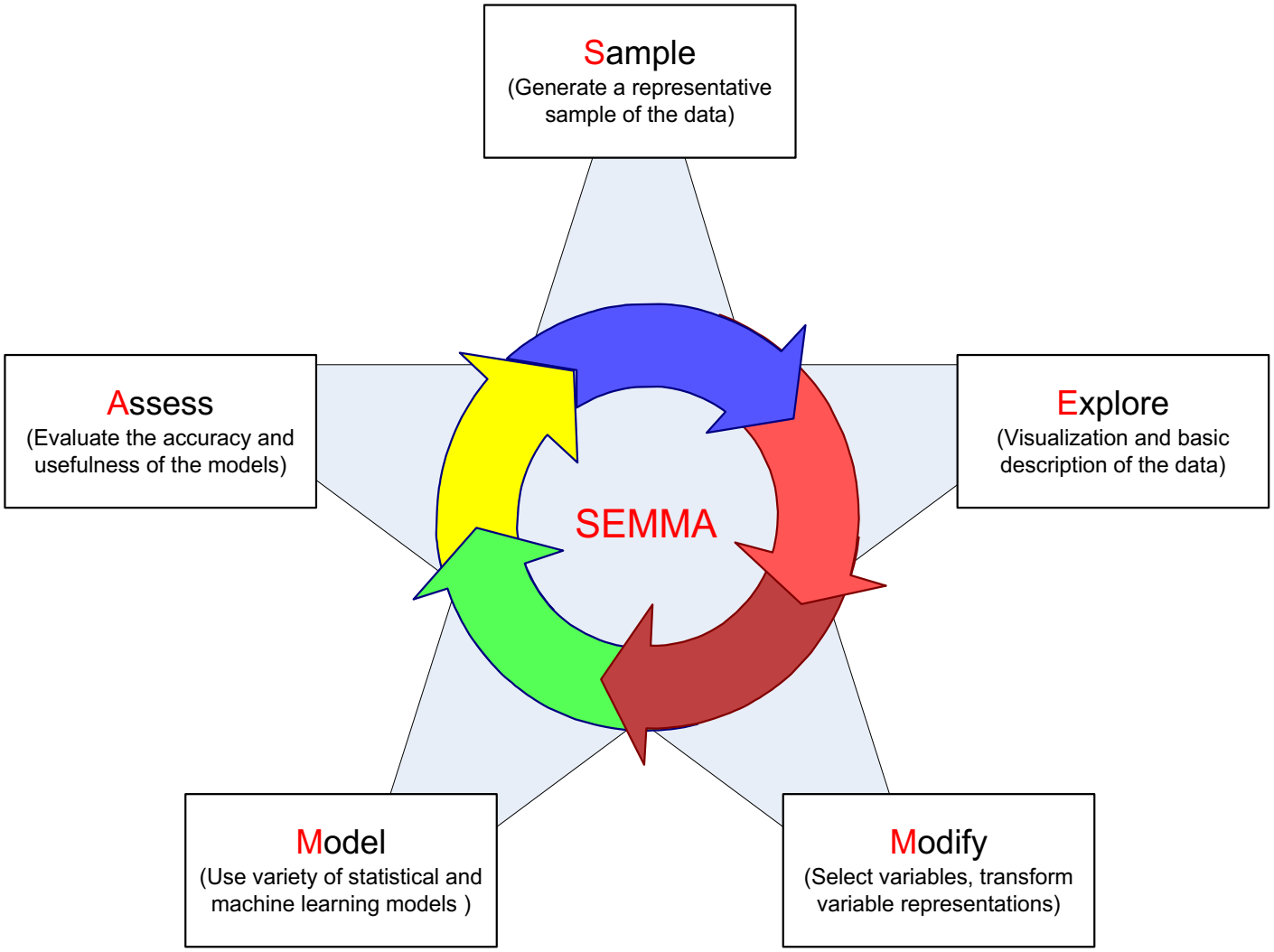
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Preparation – A Critical DM Task



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

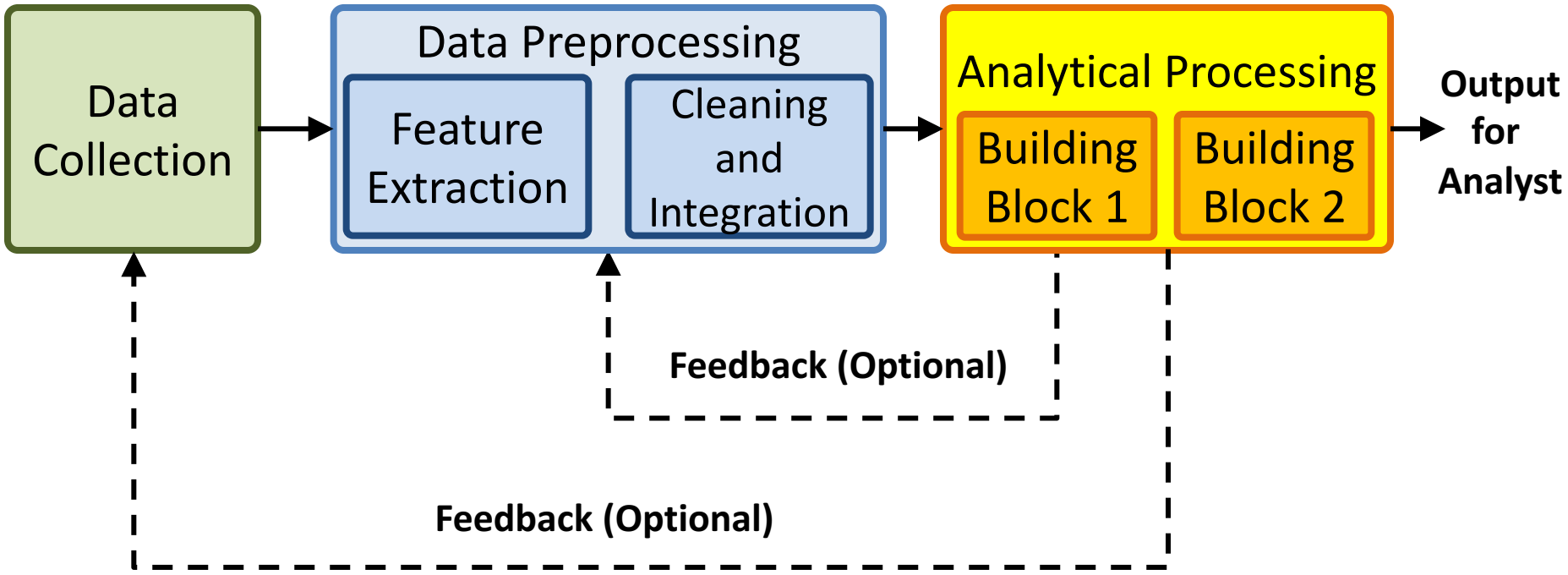
Data Mining Process: SEMMA



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

Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



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

Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem

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

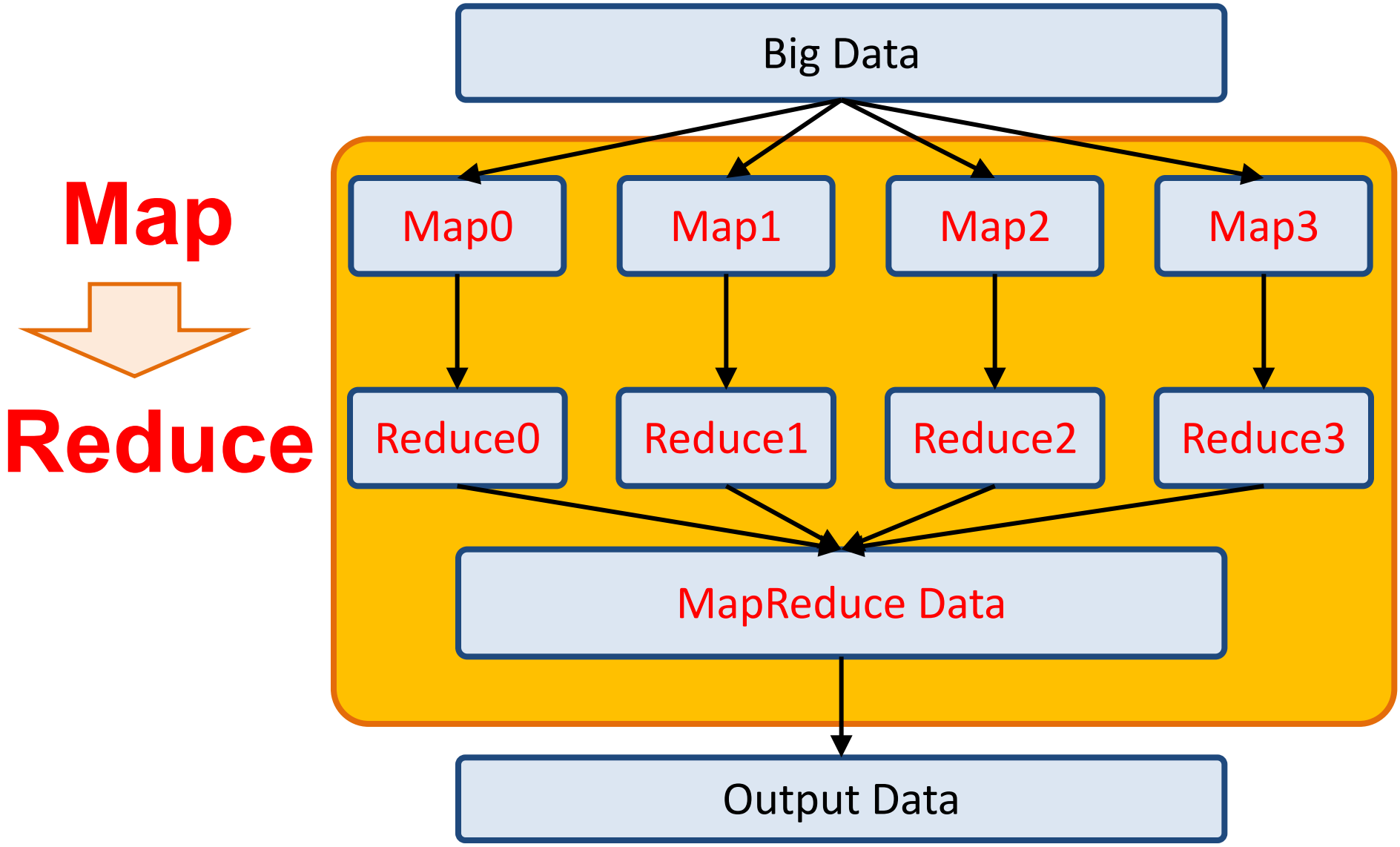
Complex by nature

Complex by structure

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

MapReduce Paradigm

MapReduce Paradigm



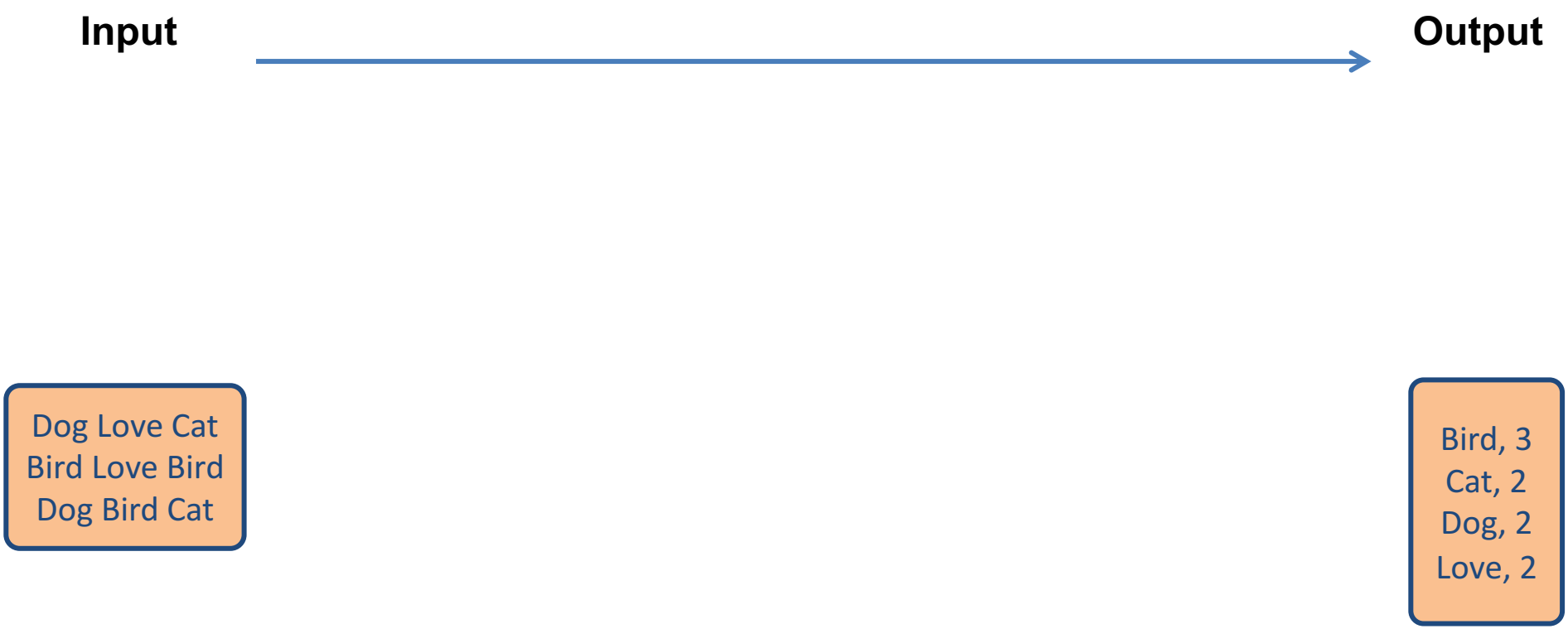
MapReduce Word Count

Input

Dog Love Cat
Bird Love Bird
Dog Bird Cat

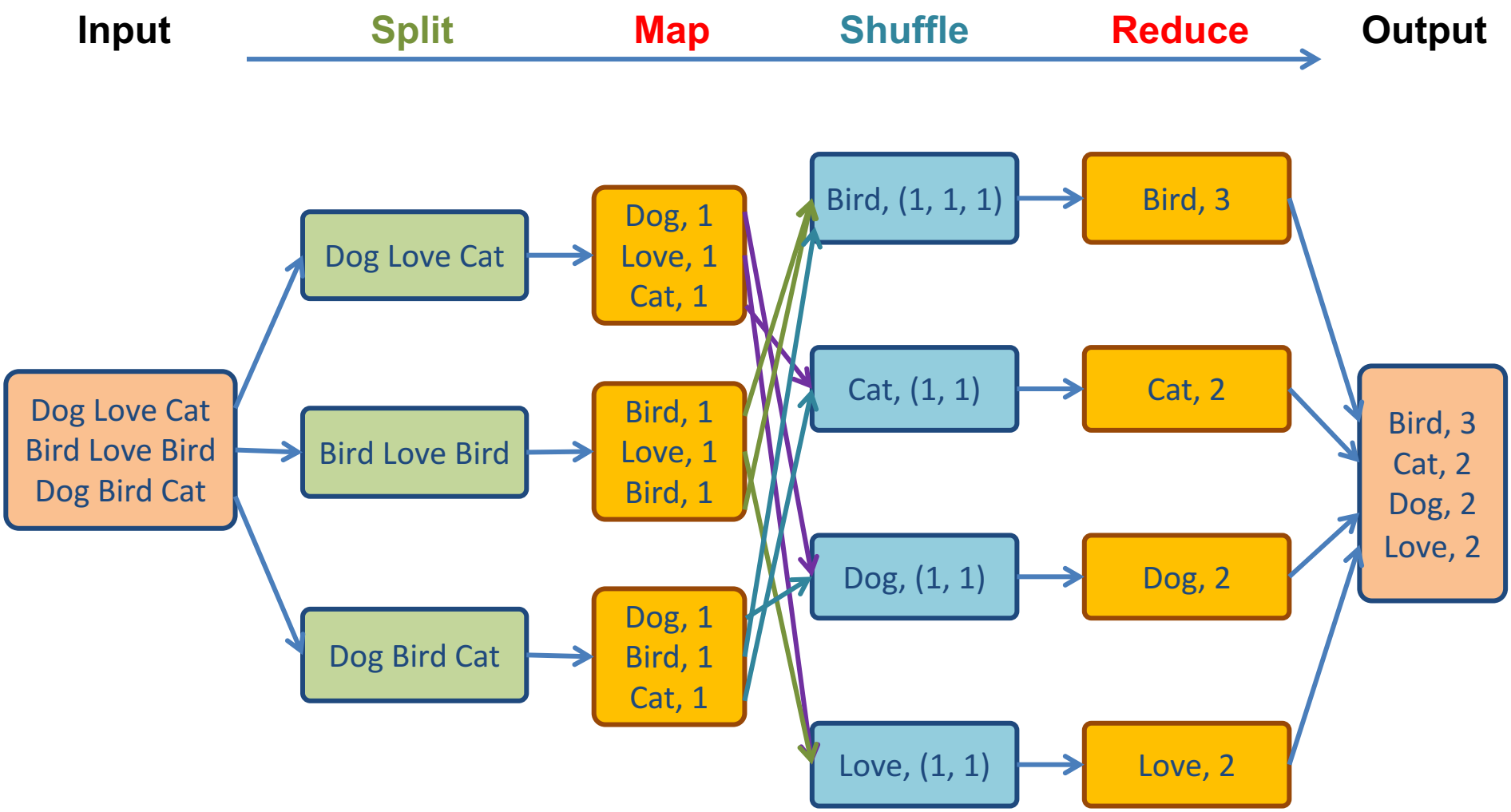
Source: https://www.edureka.co/blog/mapreduce-tutorial/

MapReduce Word Count



Source: https://www.edureka.co/blog/mapreduce-tutorial/

MapReduce Word Count



Source: https://www.edureka.co/blog/mapreduce-tutorial/

Hadoop Ecosystem

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

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

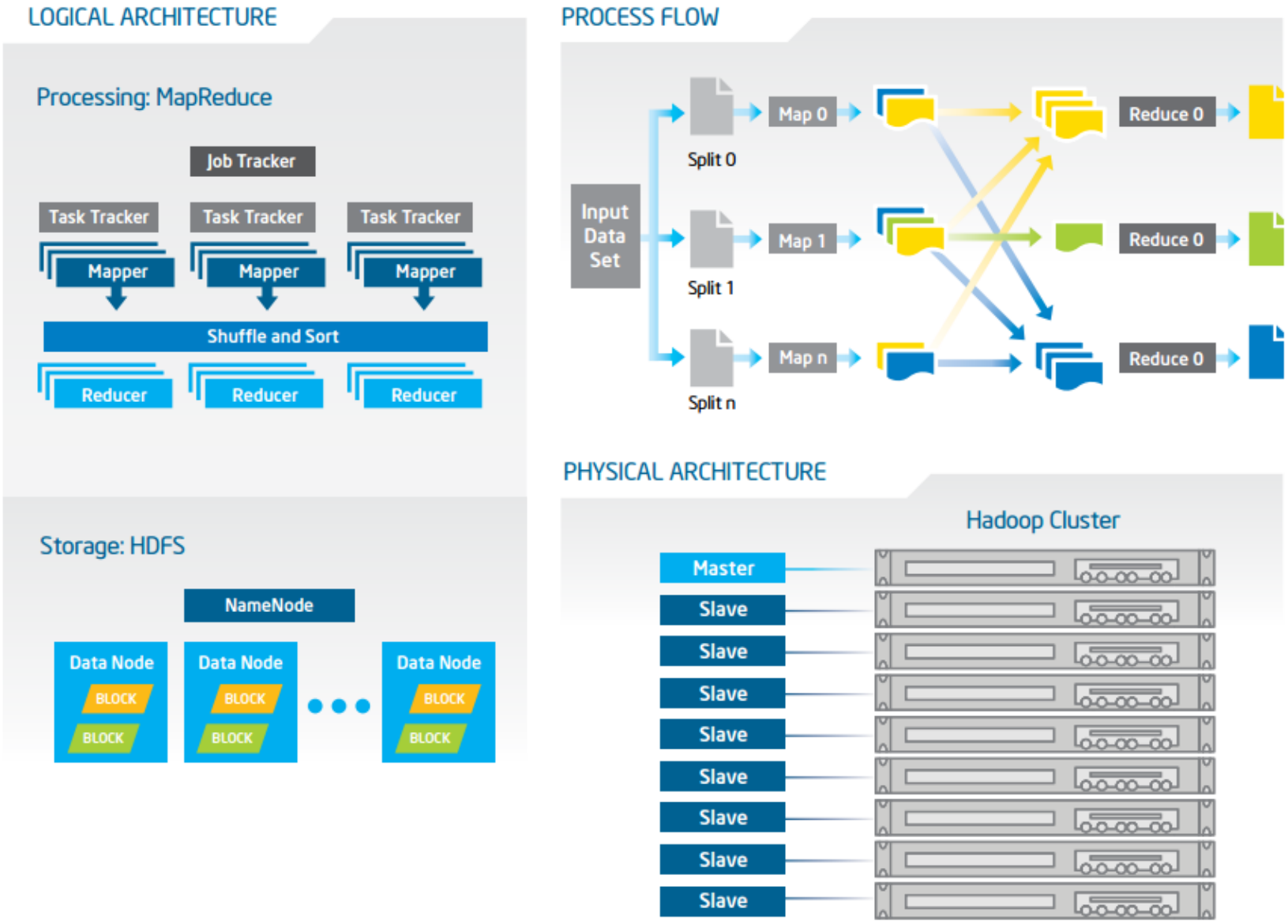
MapReduce Processing

\+

HDFS Storage

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

Big Data with Hadoop Architecture

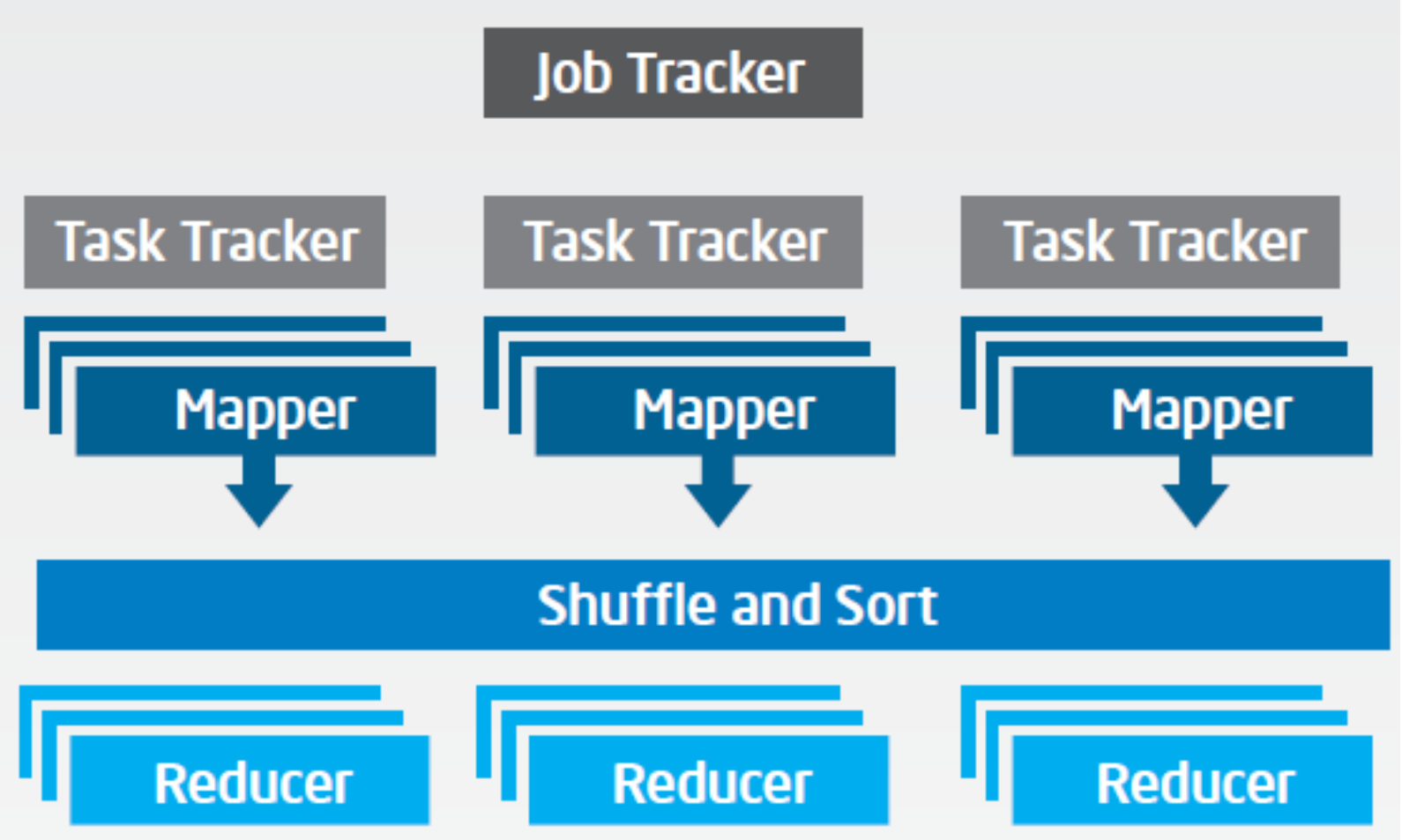


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

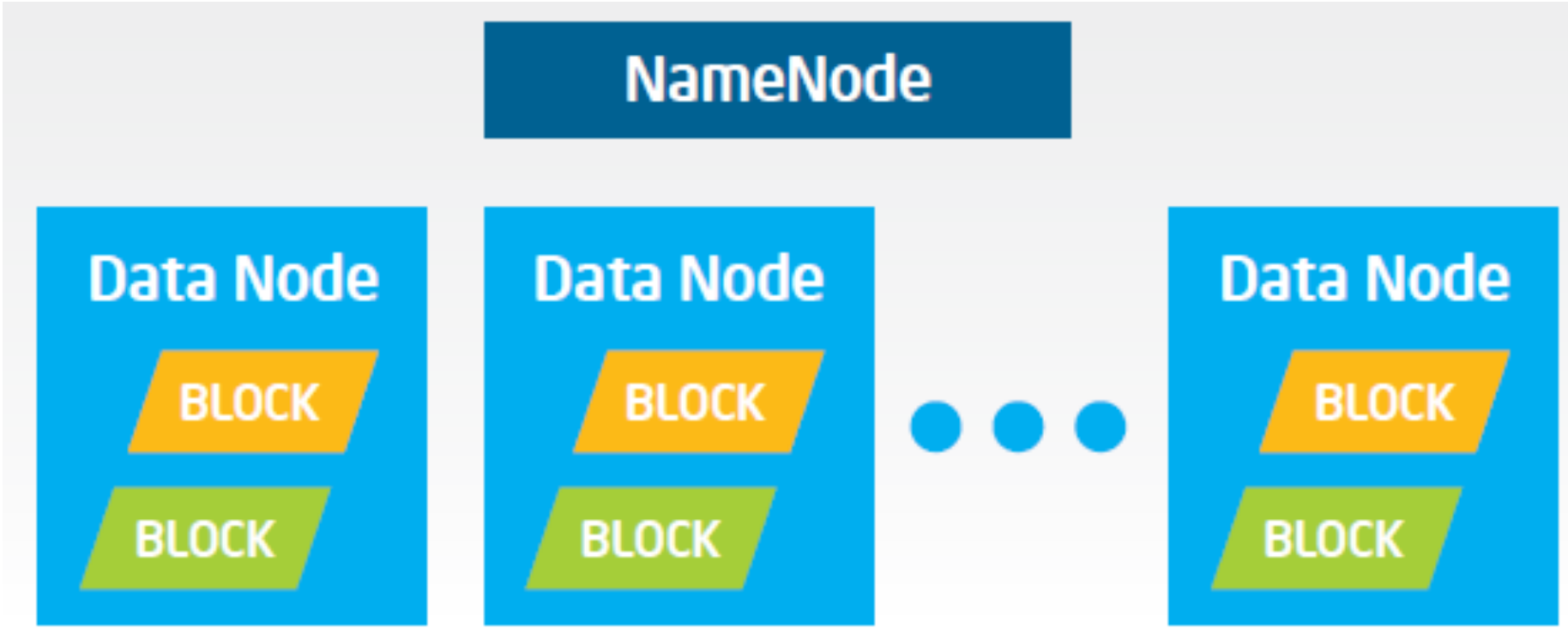


Source: https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf

Big Data with Hadoop Architecture

Logical Architecture

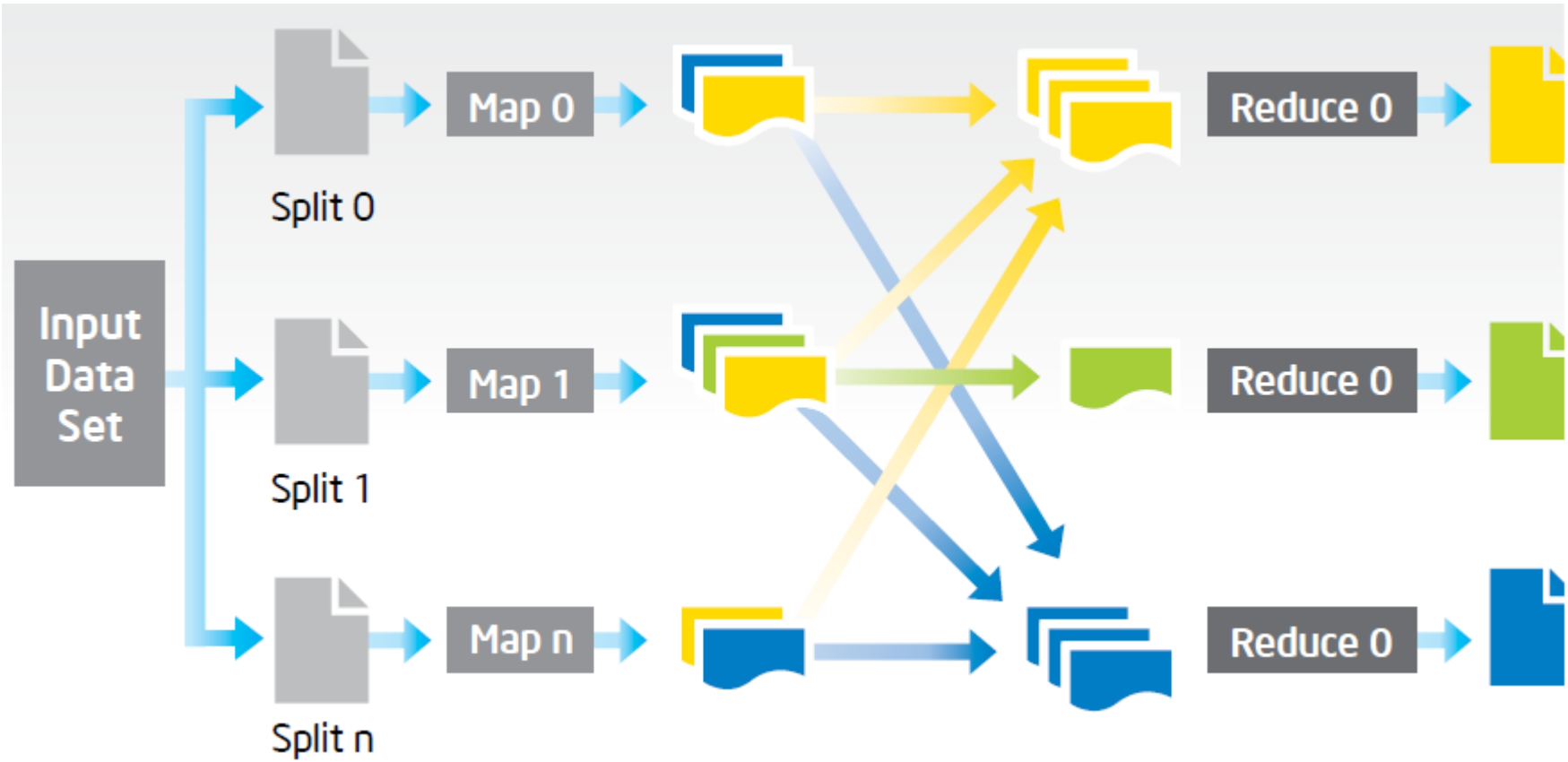
Storage: HDFS



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

Big Data with Hadoop Architecture

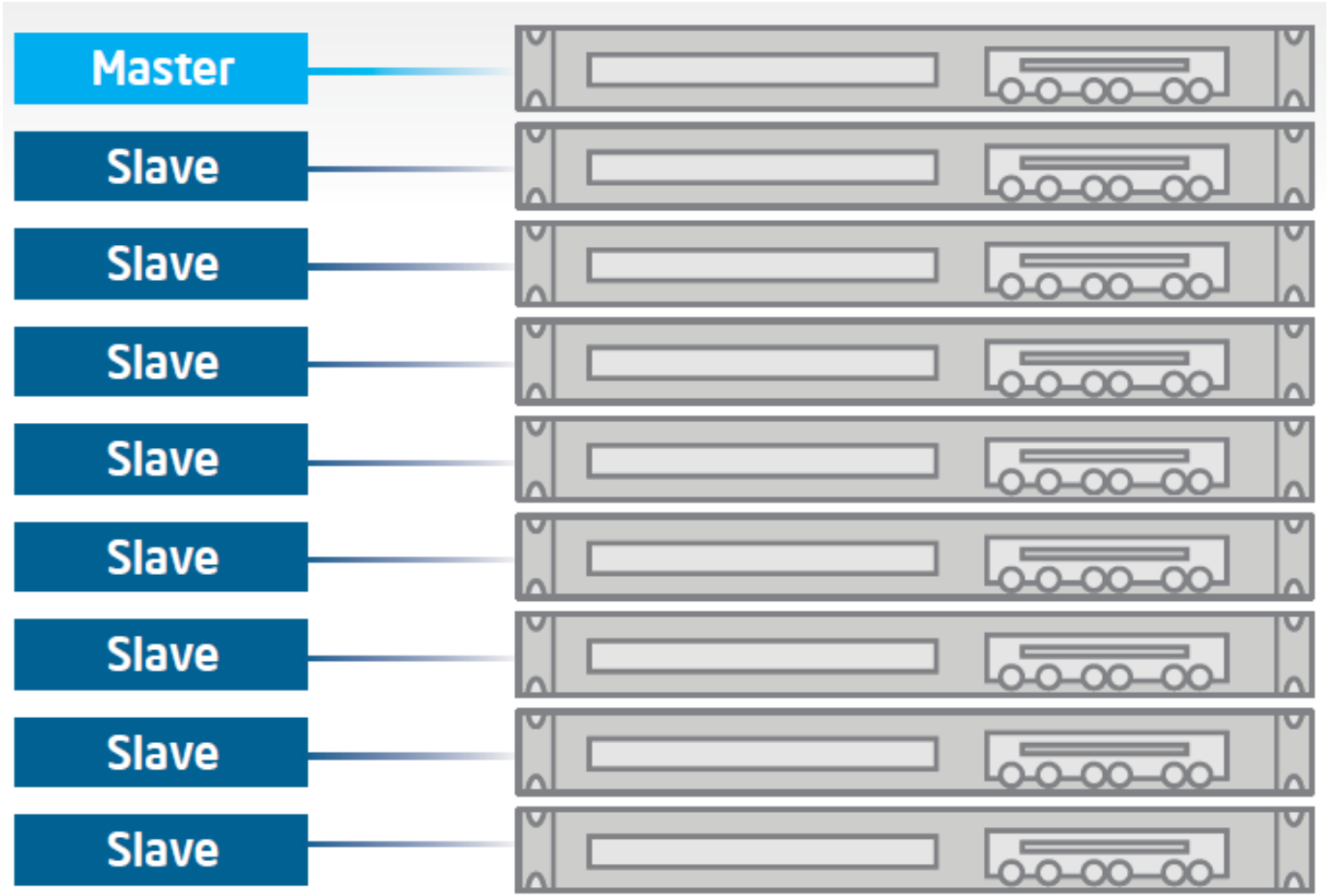
Process Flow



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

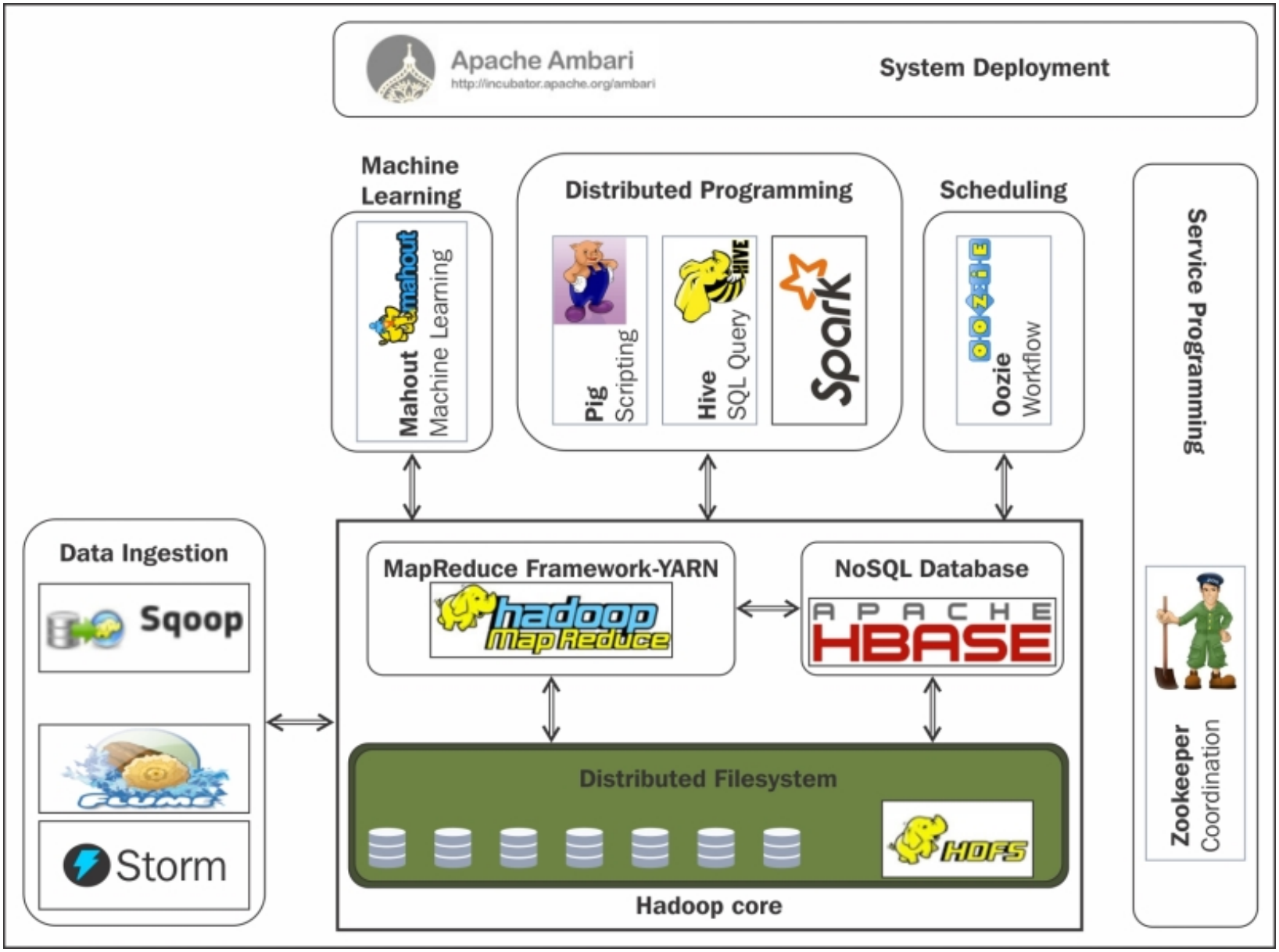
Big Data with Hadoop Architecture

Hadoop Cluster



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

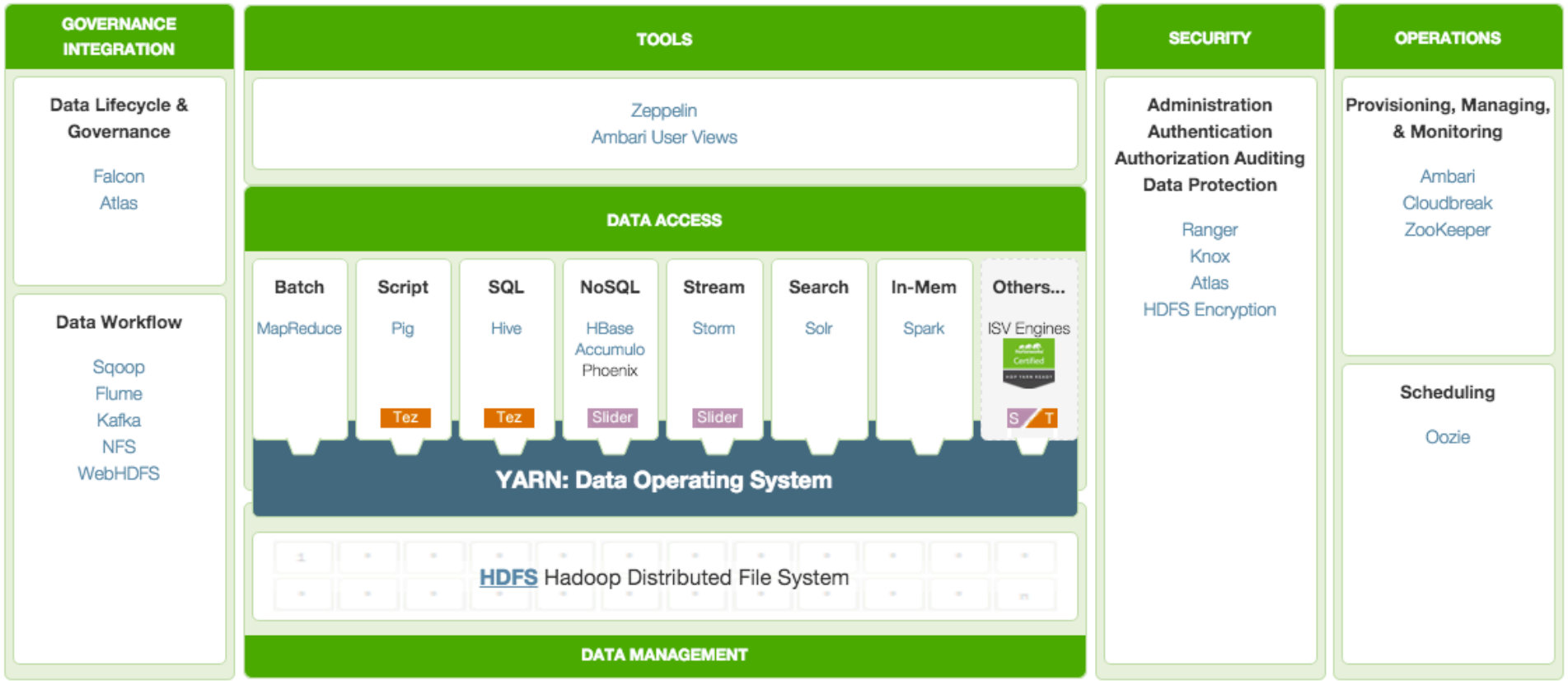
Hadoop Ecosystem



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

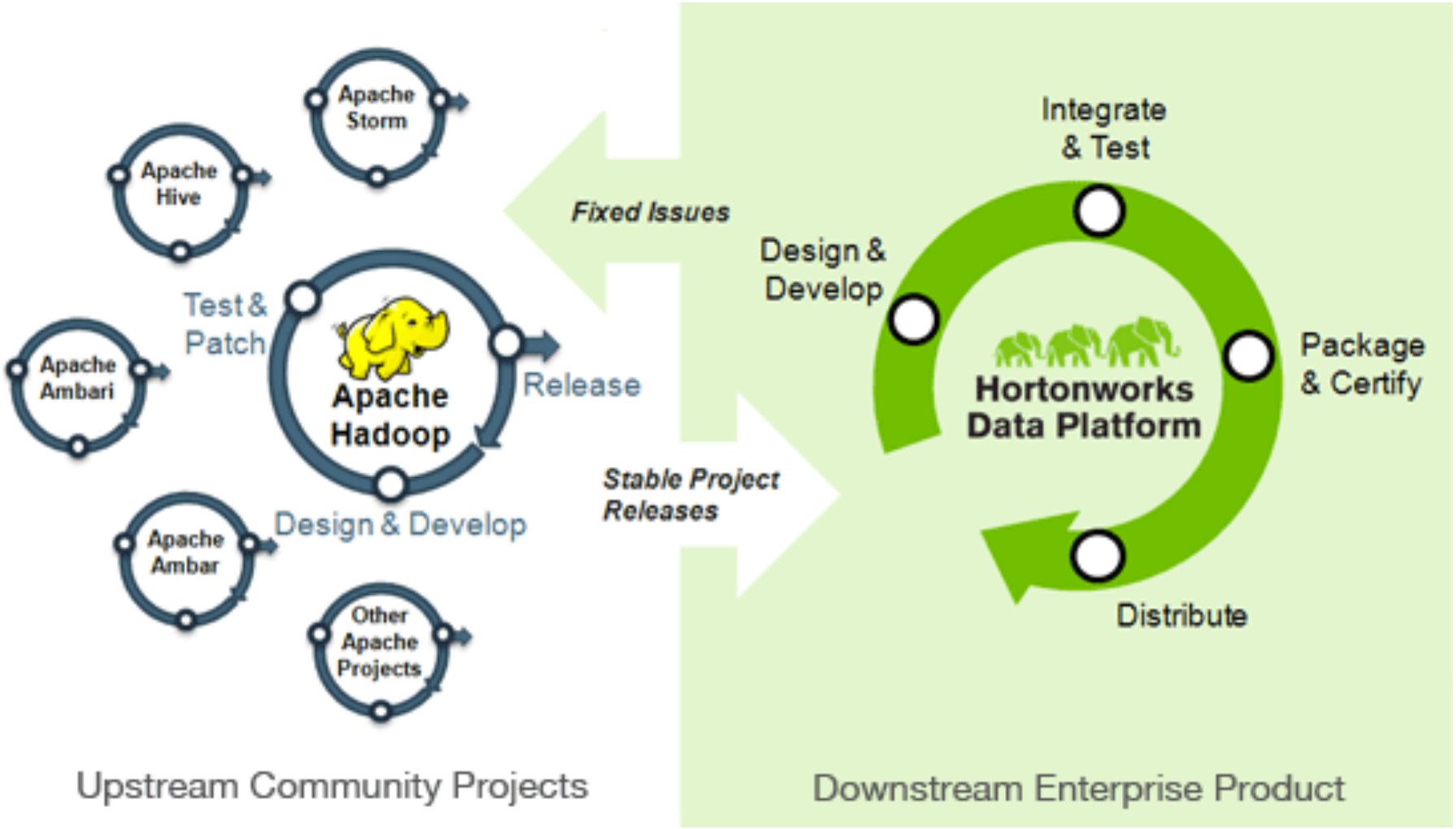
HDP (Hortonworks Data Platform)

A Complete Enterprise Hadoop Data Platform



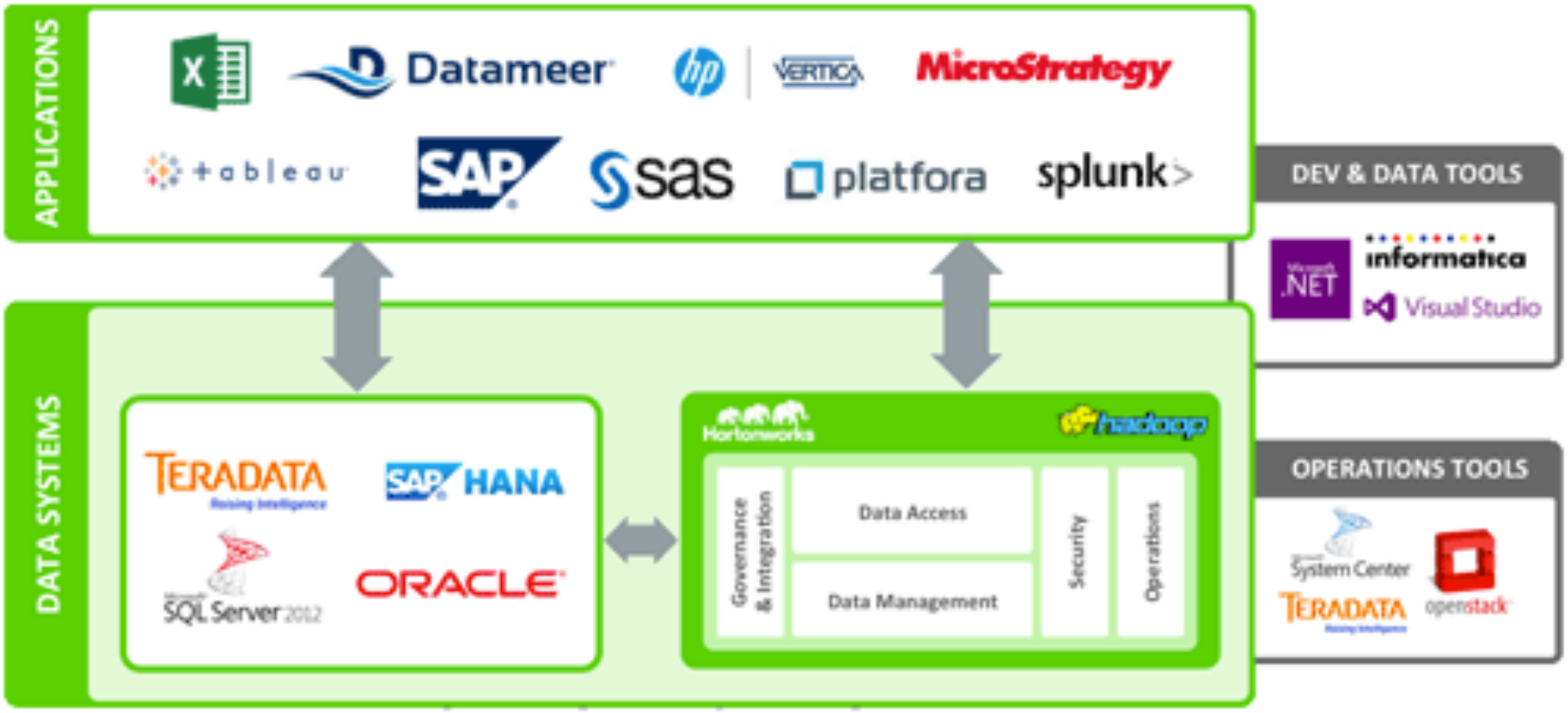
Source: http://hortonworks.com/hdp/

Apache Hadoop
Hortonworks Data Platform



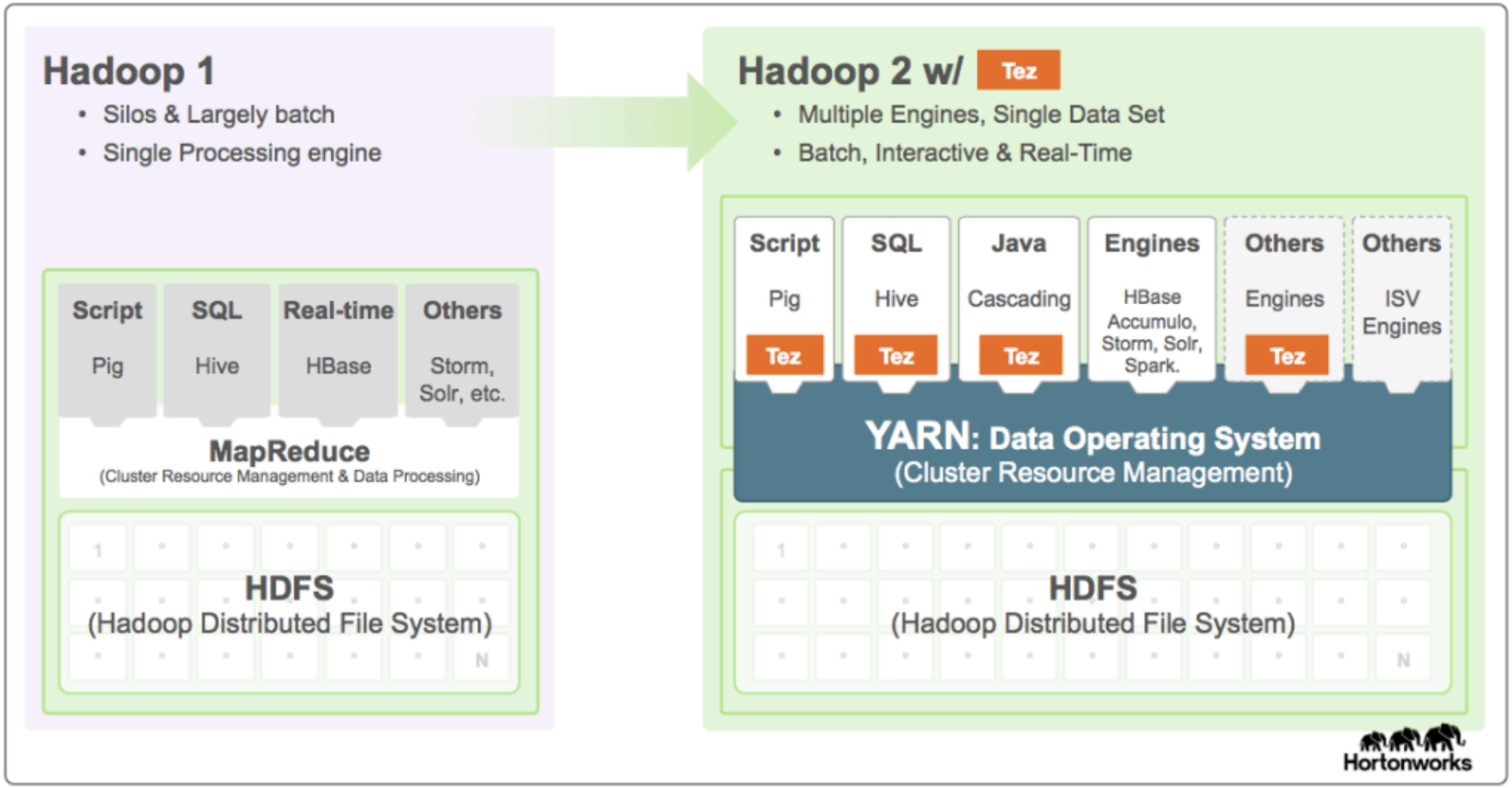
Source: http://hortonworks.com/hdp/

Hadoop and Data Analytics Tools



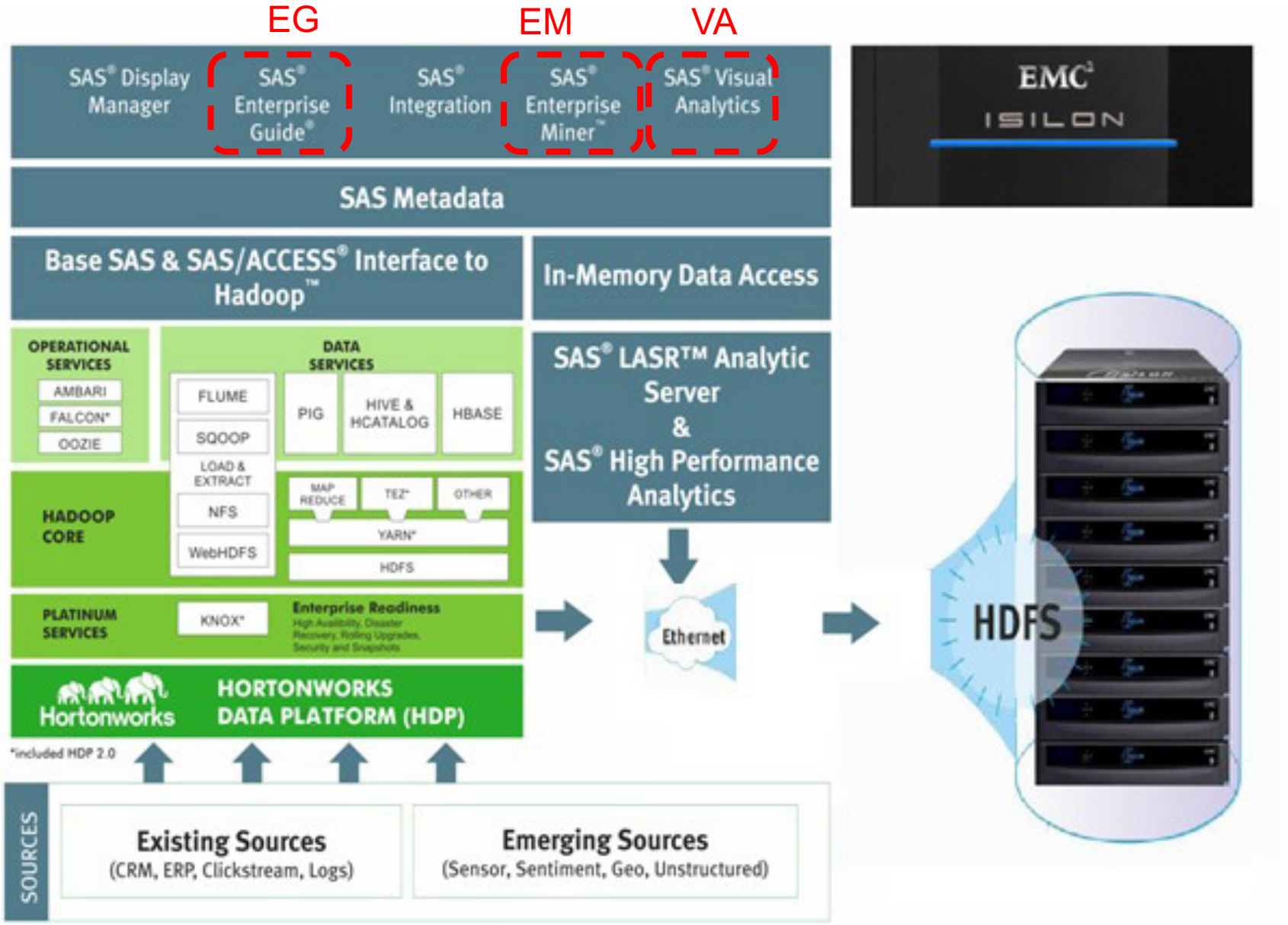
Source: http://hortonworks.com/hdp/

Hadoop 1 → Hadoop 2



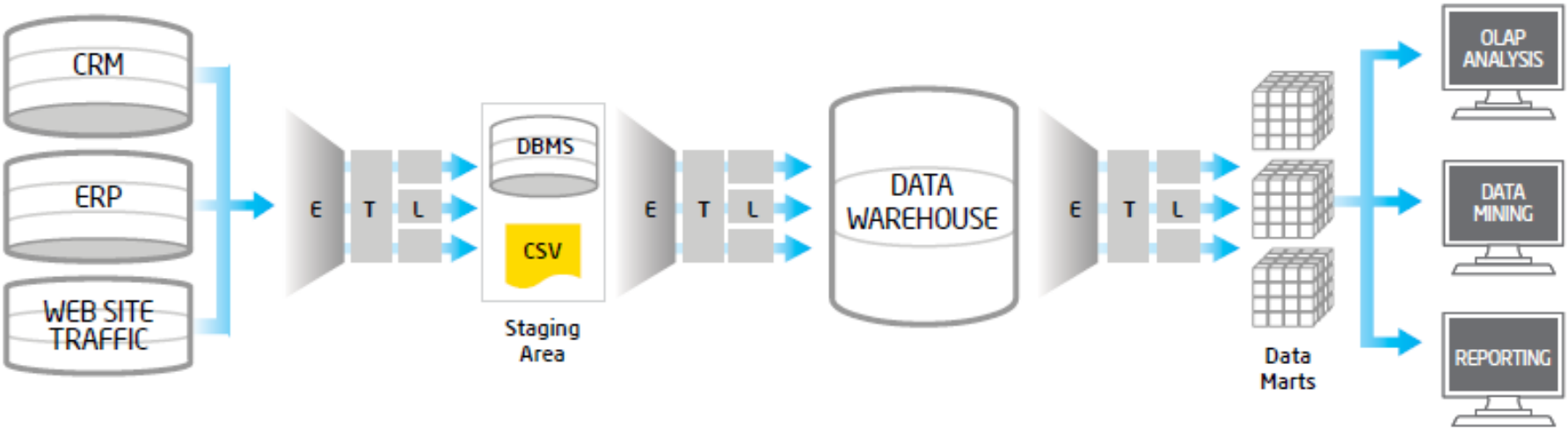
Source: http://hortonworks.com/hadoop/tez/

Big Data Solution



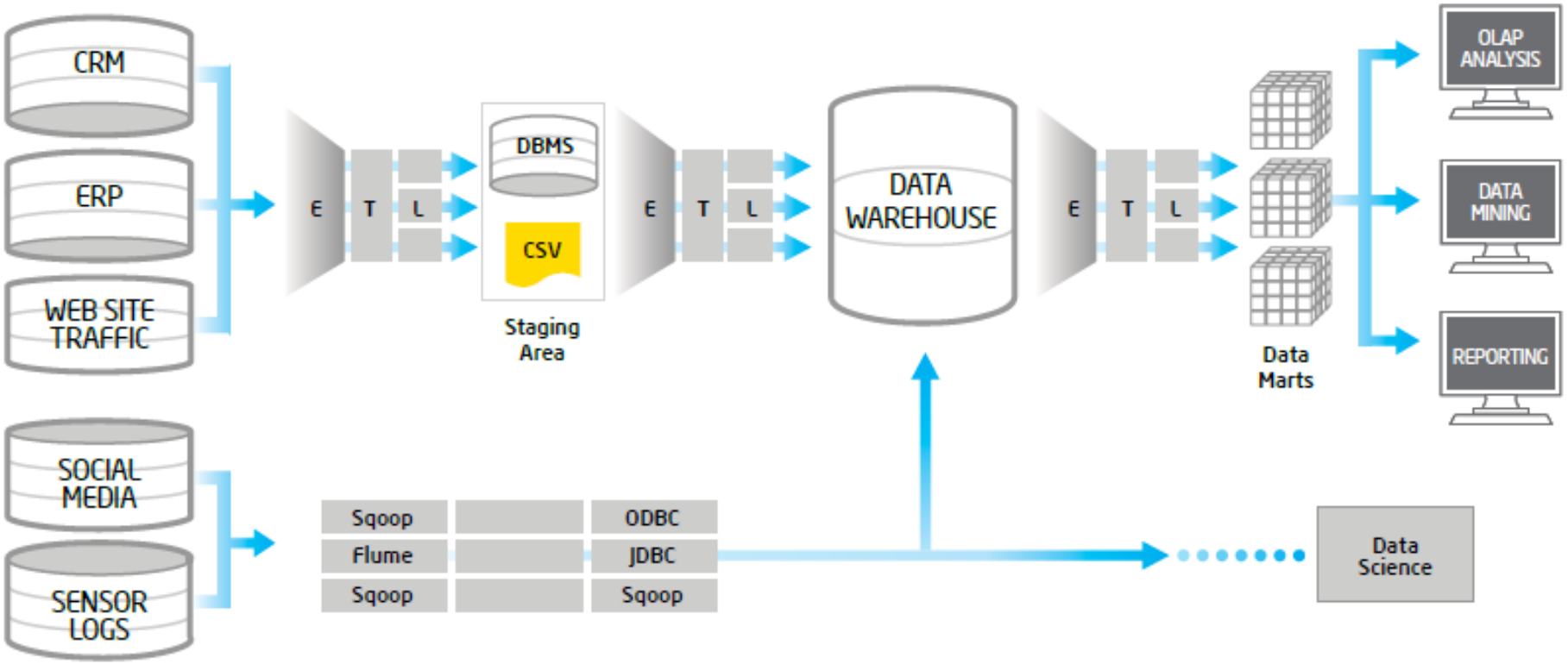
Source: http://www.newera-technologies.com/big-data-solution.html

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

Spark Ecosystem

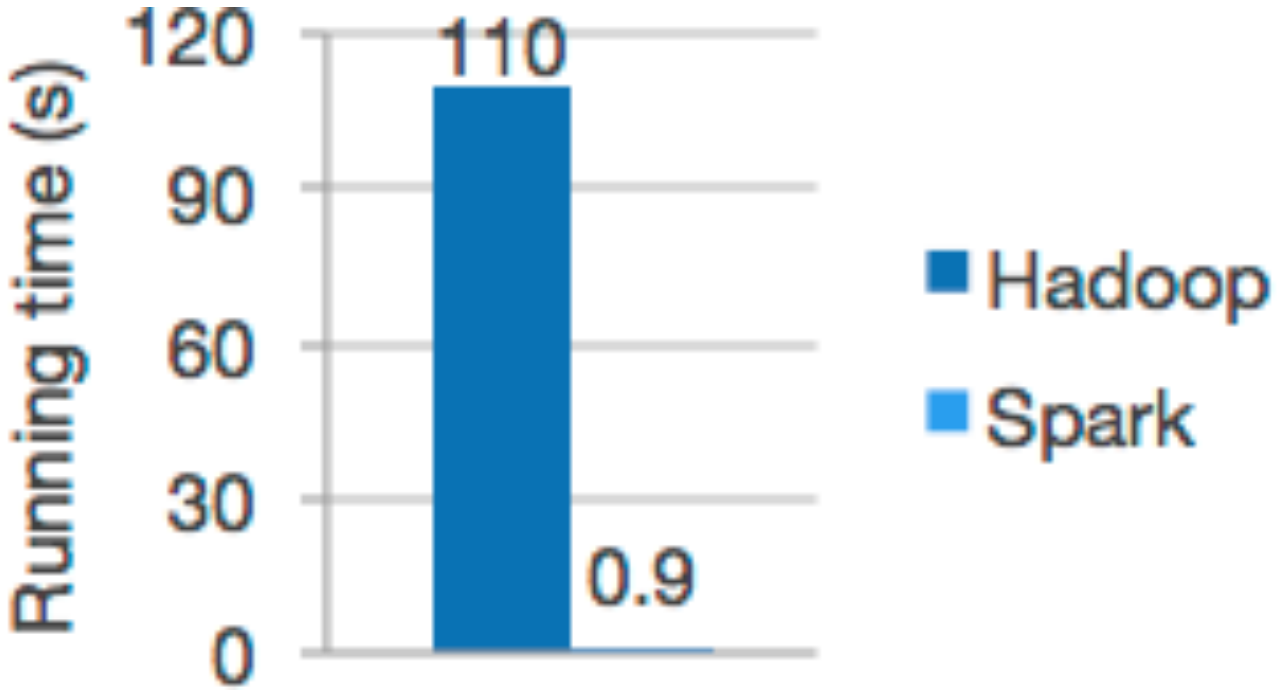
Lightning-fast cluster computing

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

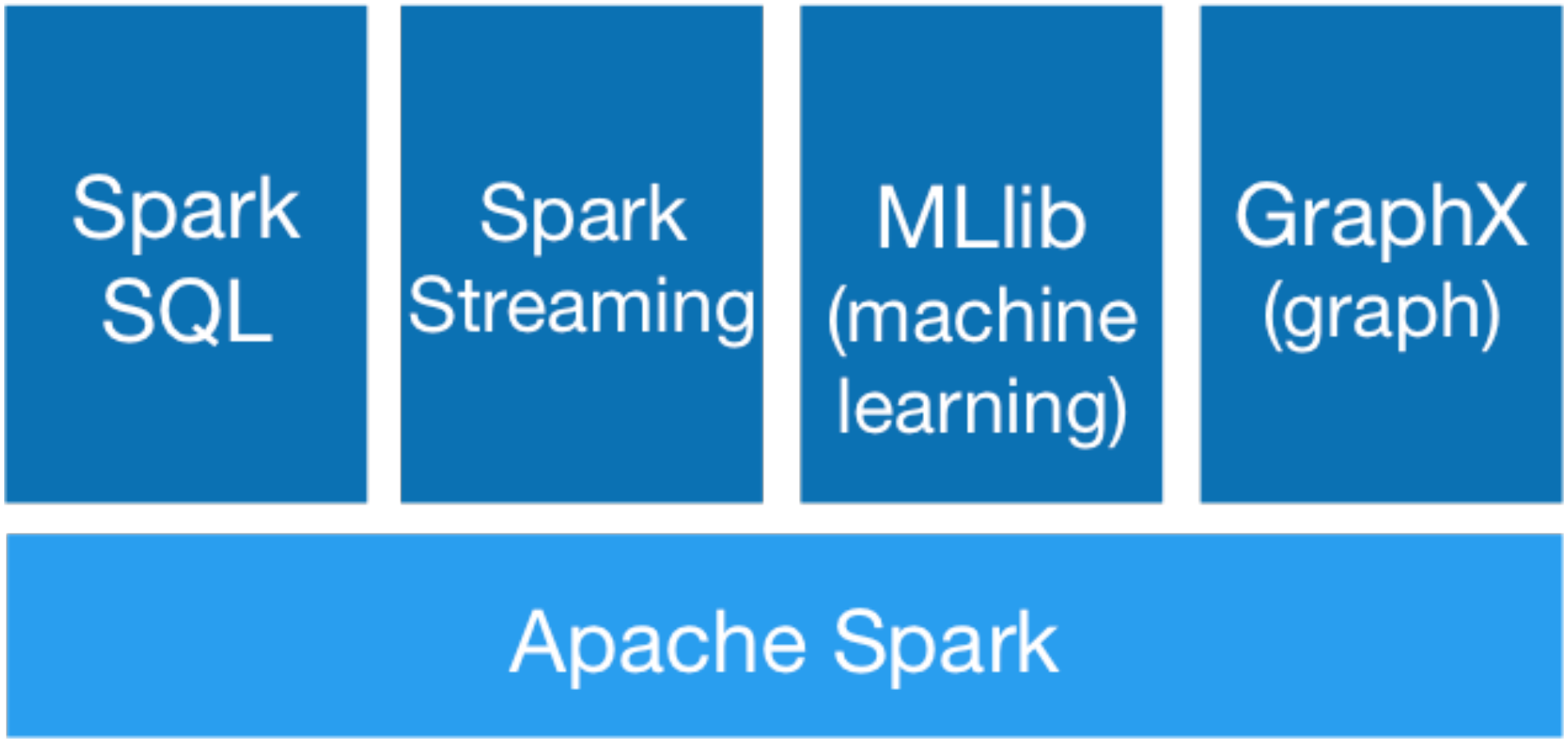
```
text_file = spark.textFile("hdfs://...")

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

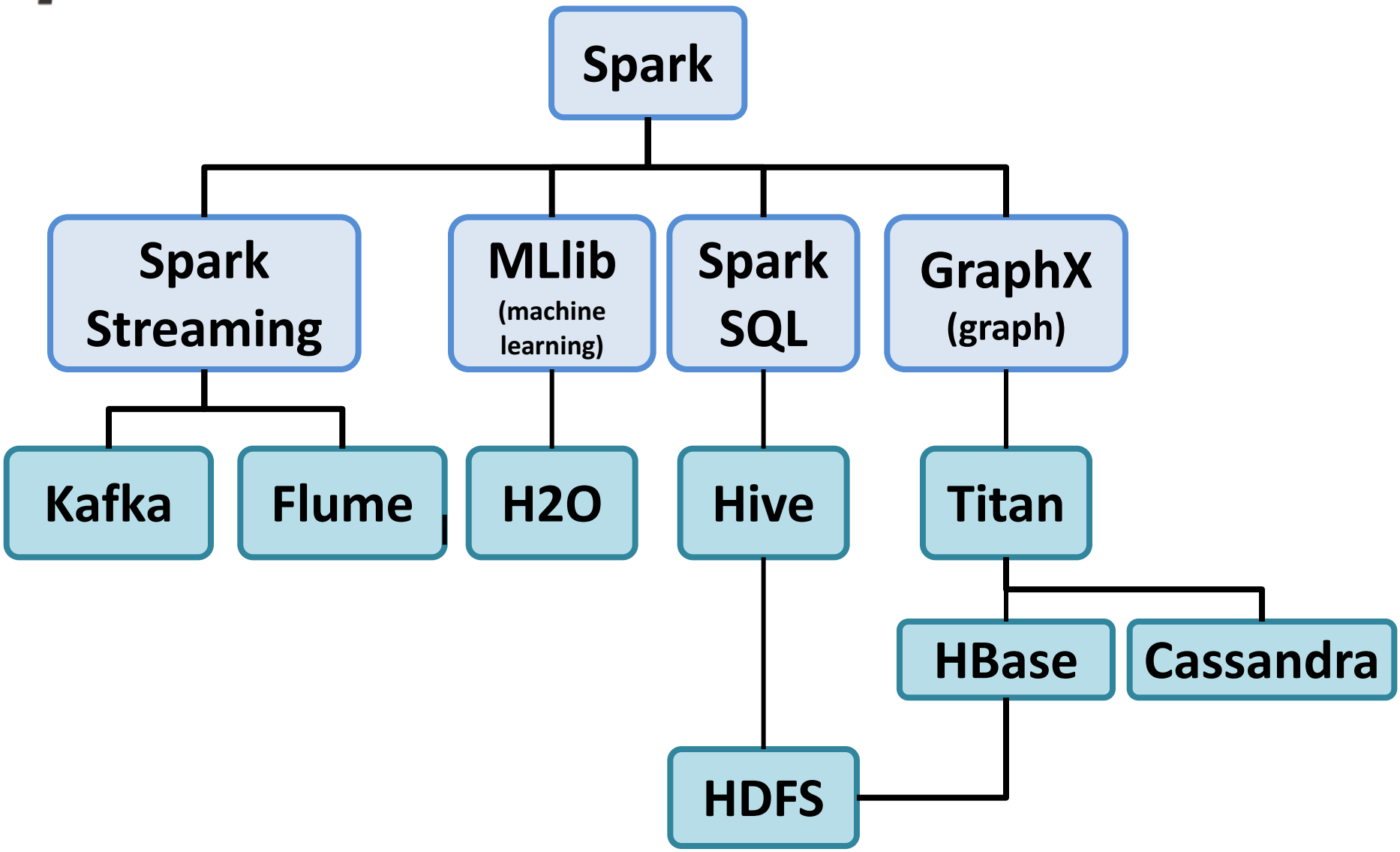
Source: http://spark.apache.org/

Spark Ecosystem



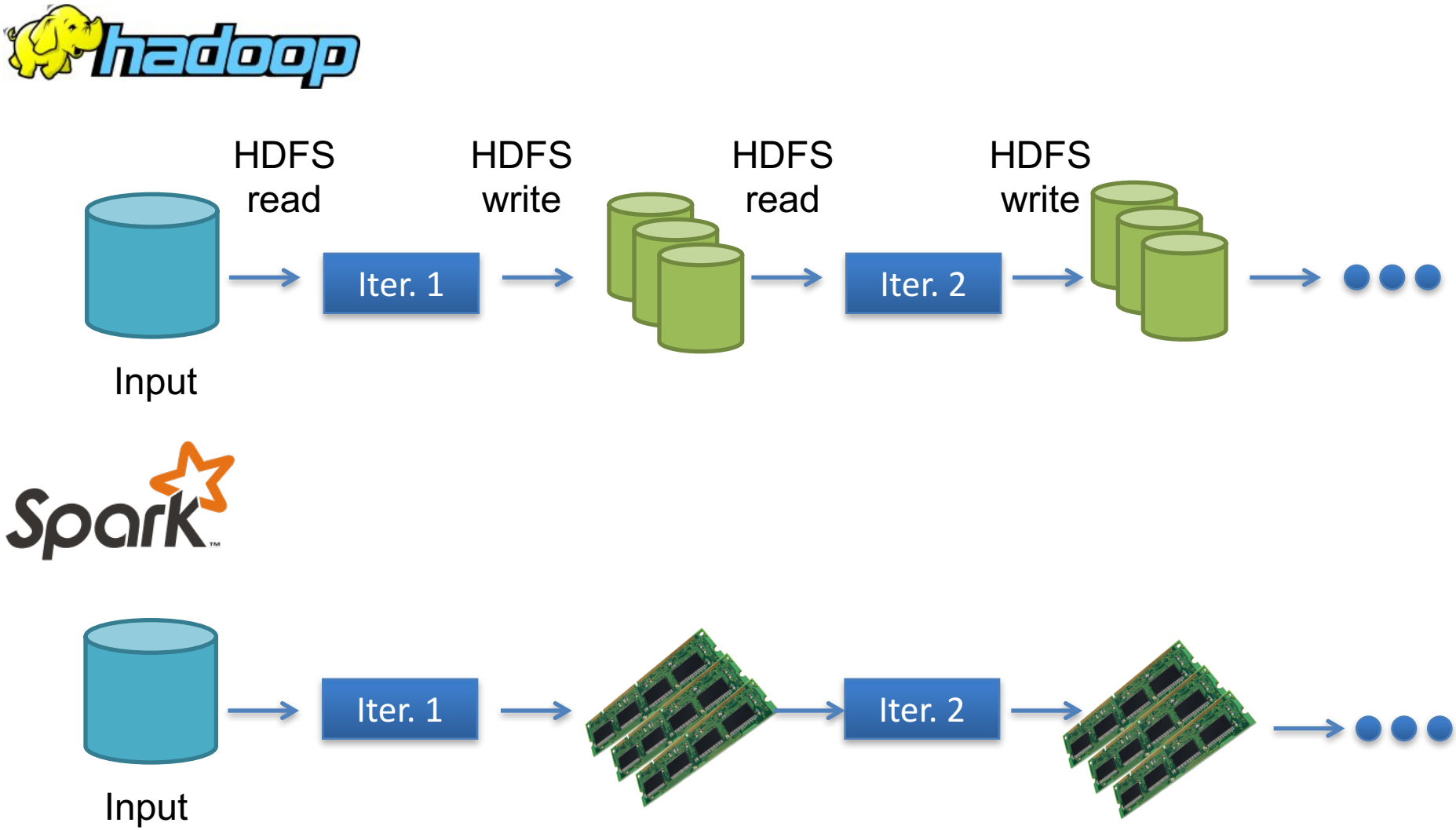
Source: http://spark.apache.org/

Spark Ecosystem



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

Hadoop vs. Spark



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

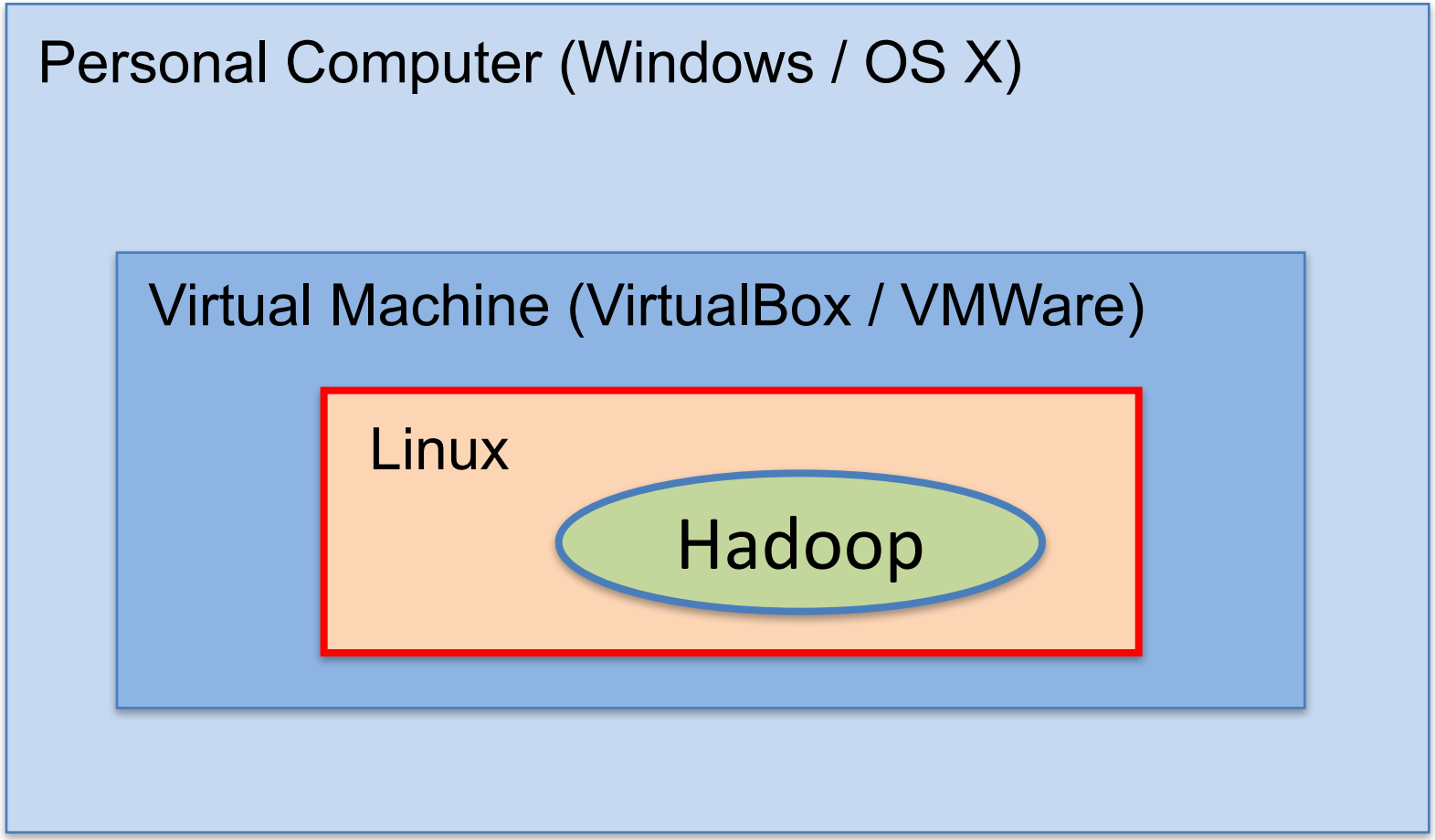
Steps to Install Hadoop on a Personal Computer (Windows/OS X)

Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

Hodoop: Linux Based Software

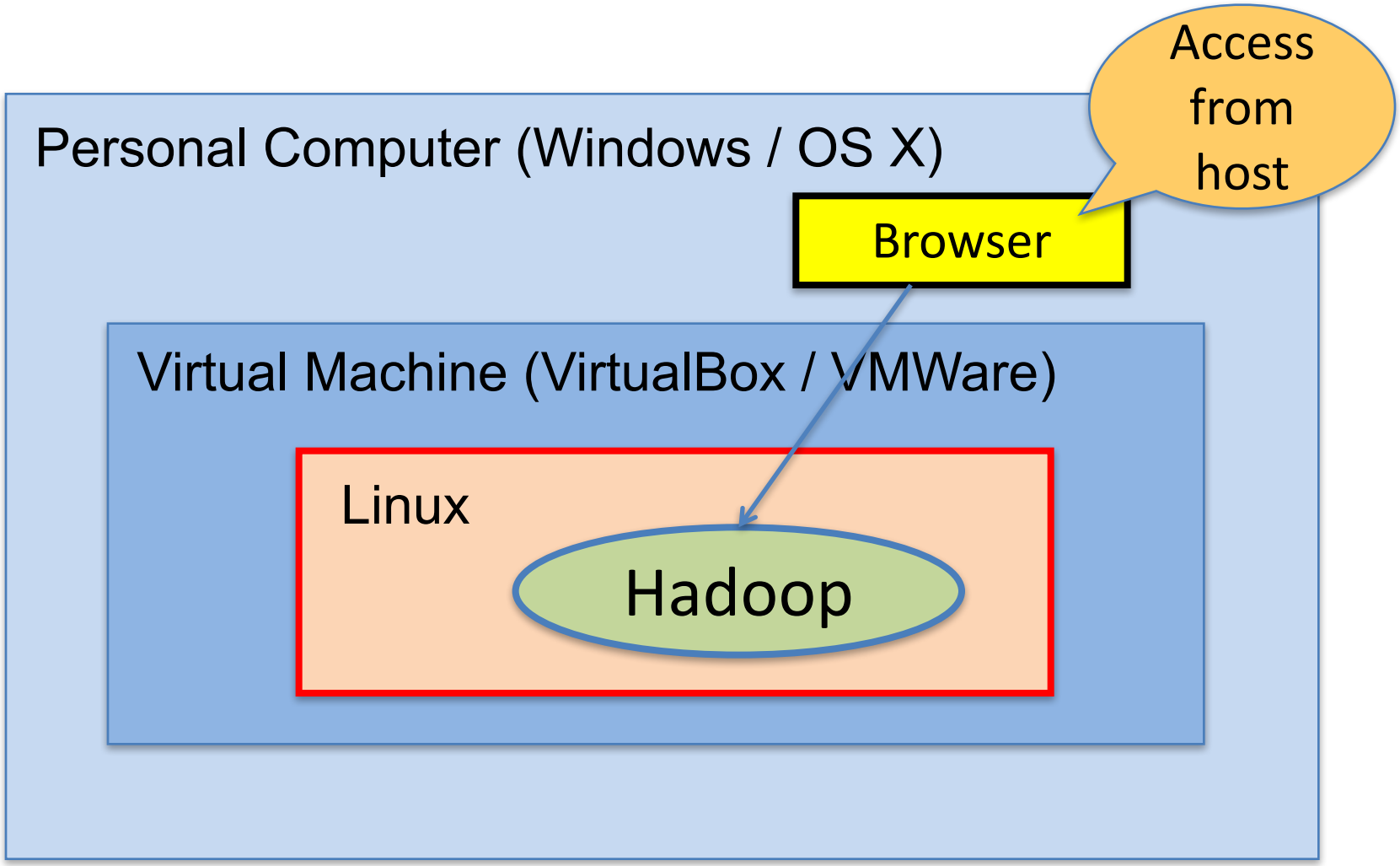
Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

Appliance



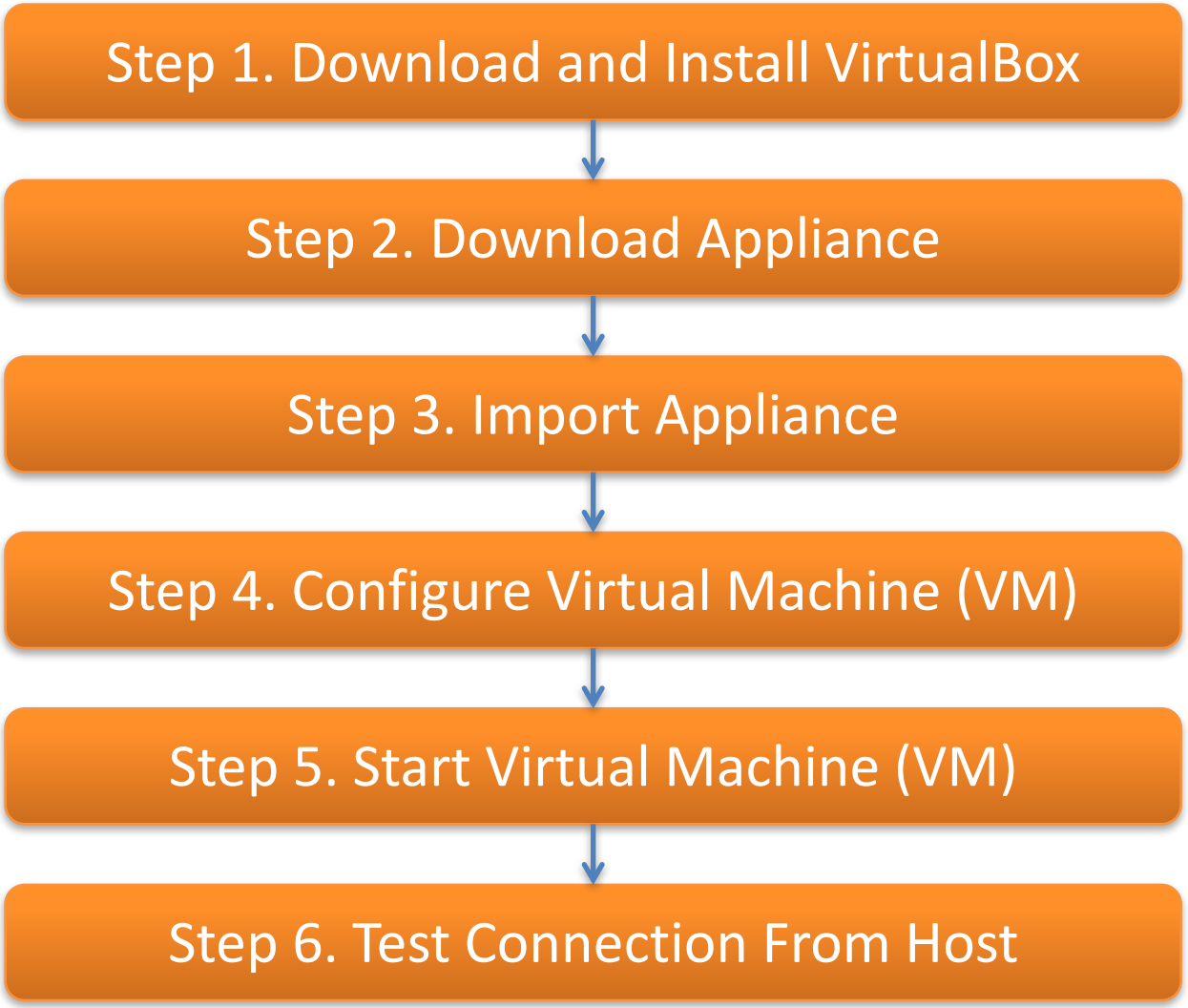
Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

Connection to Hadoop



Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

Steps to Install Hadoop on a Personal Computer (Windows/OS X)

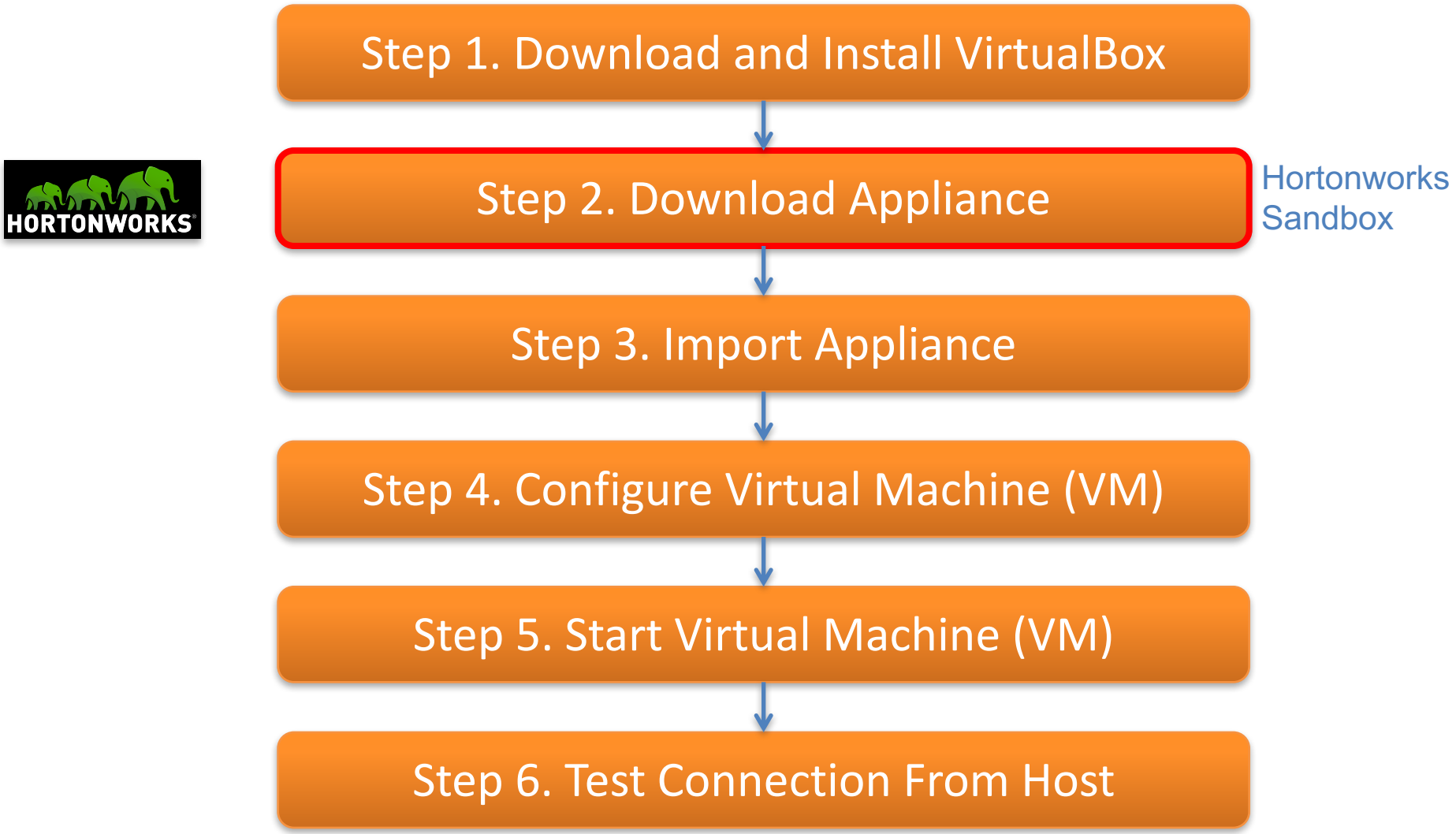


Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

Virtual Box

https://www.virtualbox.org/

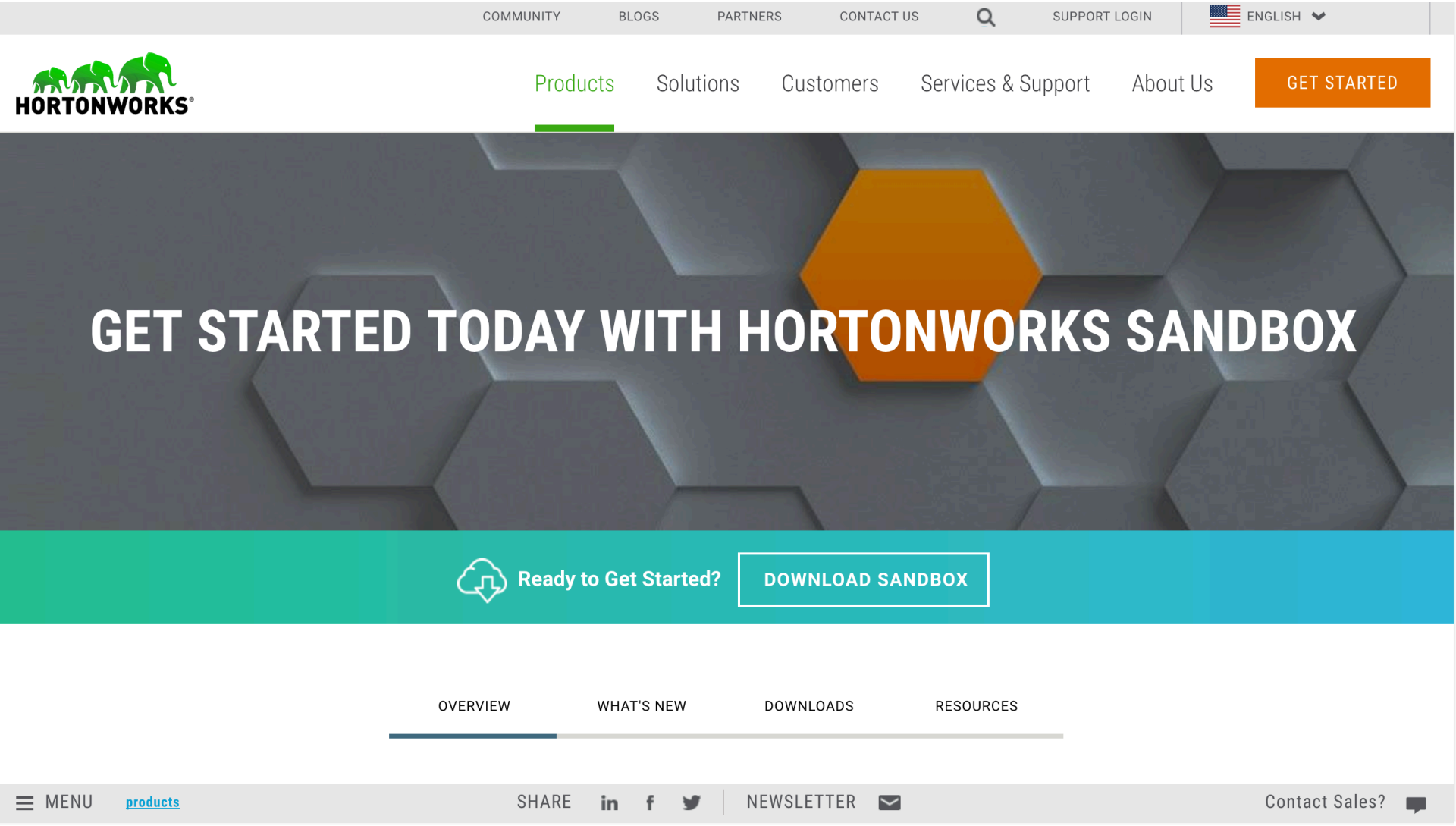
Steps to Install Hadoop on a Personal Computer (Windows/OS X)



Source: https://www.youtube.com/watch?v=rO-V1mxhzcM&list=PLyZEf-TOnZen8E5m5TIpIsdok2fyKDNRa&index=5

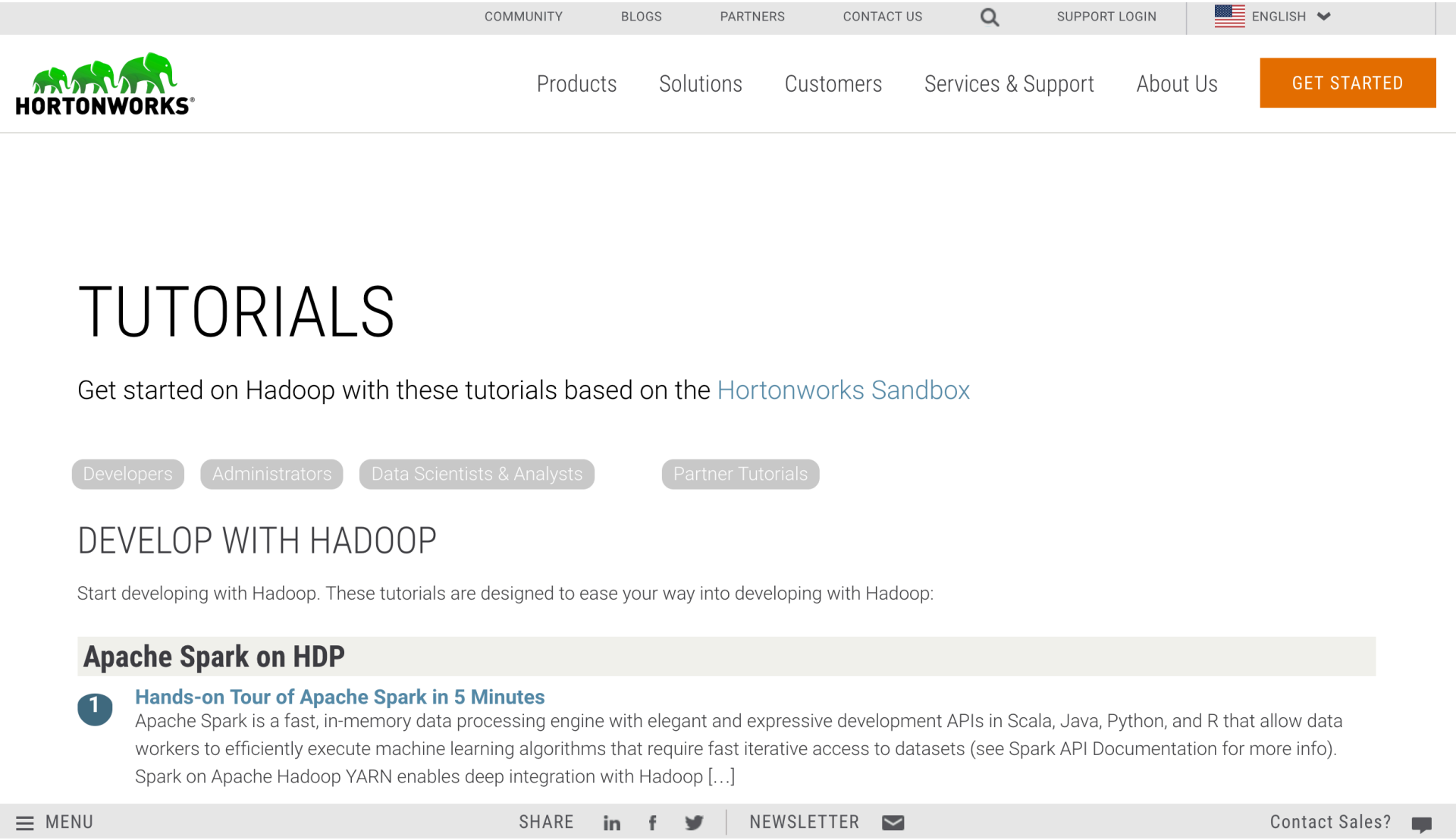
Hortonworks Sandbox

The easiest way to get started with Enterprise Hadoop



http://hortonworks.com/products/hortonworks-sandbox/#install

Get started on Hadoop with these tutorials based on the Hortonworks Sandbox



http://hortonworks.com/tutorials/

Apache Hadoop

Apache > Hadoop >

Top | Wiki

Search with Apache Solr | Search

Last Published: 01/27/2017 02:33:46

- About
  - Welcome
    - What Is Apache Hado...
    - Getting Started ...
    - Download Hadoop
    - Who Uses Hadoop?...
    - News
  - Releases
  - Release Versioning
  - Mailing Lists
  - Issue Tracking
  - Who We Are?
  - Who Uses Hadoop?
  - Buy Stuff
  - Sponsorship
  - Thanks
  - Privacy Policy
  - Bylaws
  - Committer criteria
  - License
- Documentation
- Related Projects

built with Apache Forrest

Welcome to Apache™ Hadoop®!

PDF

What Is Apache Hadoop?

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

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

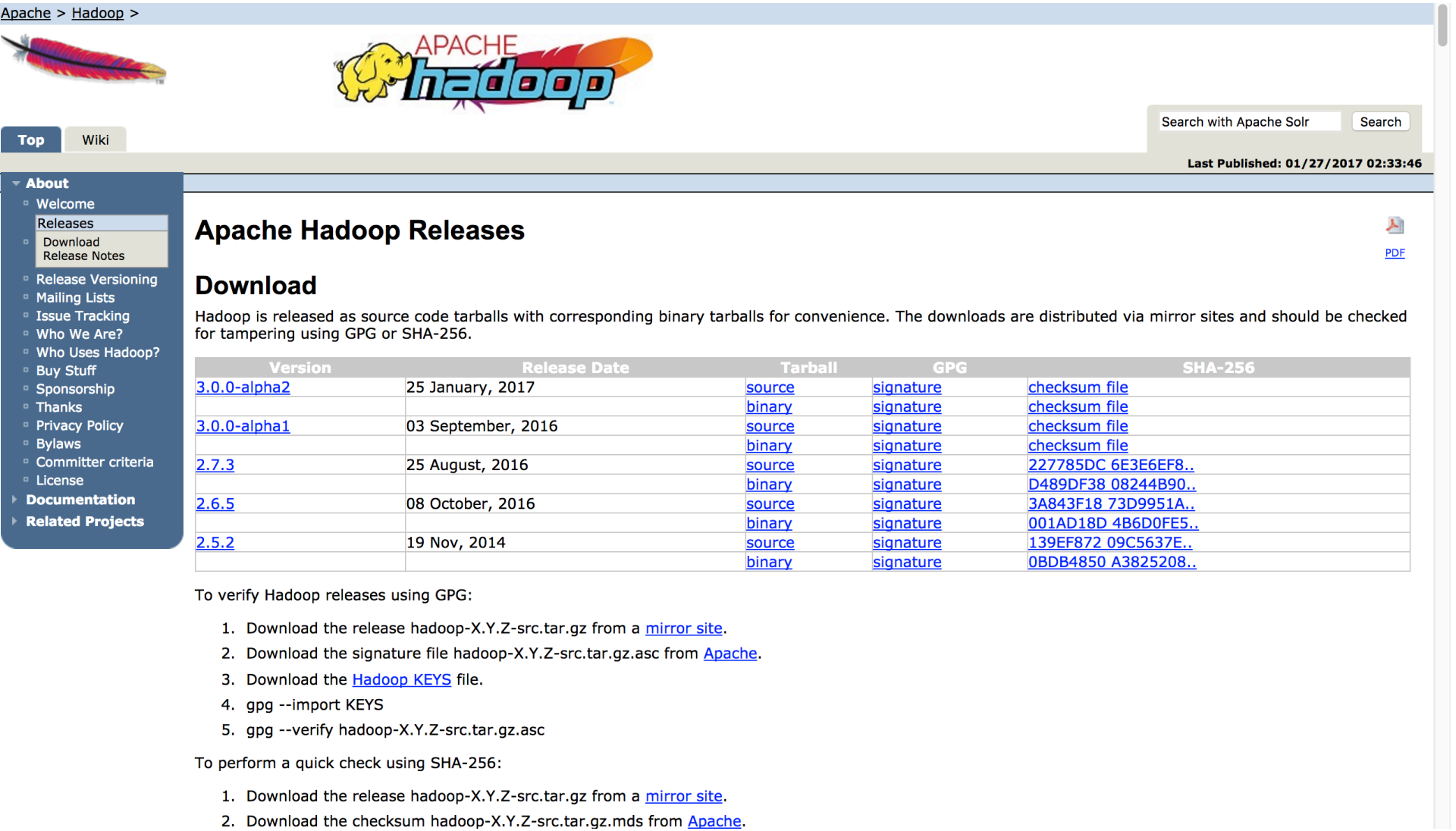
Other Hadoop-related projects at Apache include:

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.

http://hadoop.apache.org/

Apache Hadoop

http://hadoop.apache.org/releases.html#Download



Apache > Hadoop >

Top | Wiki

Search with Apache Solr | Search

Last Published: 01/27/2017 02:33:46

- About
  - Welcome
  - Releases
    - Download
    - Release Notes
  - Release Versioning
  - Mailing Lists
  - Issue Tracking
  - Who We Are?
  - Who Uses Hadoop?
  - Buy Stuff
  - Sponsorship
  - Thanks
  - Privacy Policy
  - Bylaws
  - Committer criteria
  - License
- Documentation
- Related Projects

Apache Hadoop Releases

PDF

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release Date	Tarball	GPG	SHA-256
3.0.0-alpha2	25 January, 2017	source	signature	checksum file
		binary	signature	checksum file
3.0.0-alpha1	03 September, 2016	source	signature	checksum file
		binary	signature	checksum file
2.7.3	25 August, 2016	source	signature	227785DC 6E3E6EF8..
		binary	signature	D489DF38 08244B90..
2.6.5	08 October, 2016	source	signature	3A843F18 73D9951A..
		binary	signature	001AD18D 4B6D0FE5..
2.5.2	19 Nov, 2014	source	signature	139EF872 09C5637E..
		binary	signature	0BDB4850 A3825208..

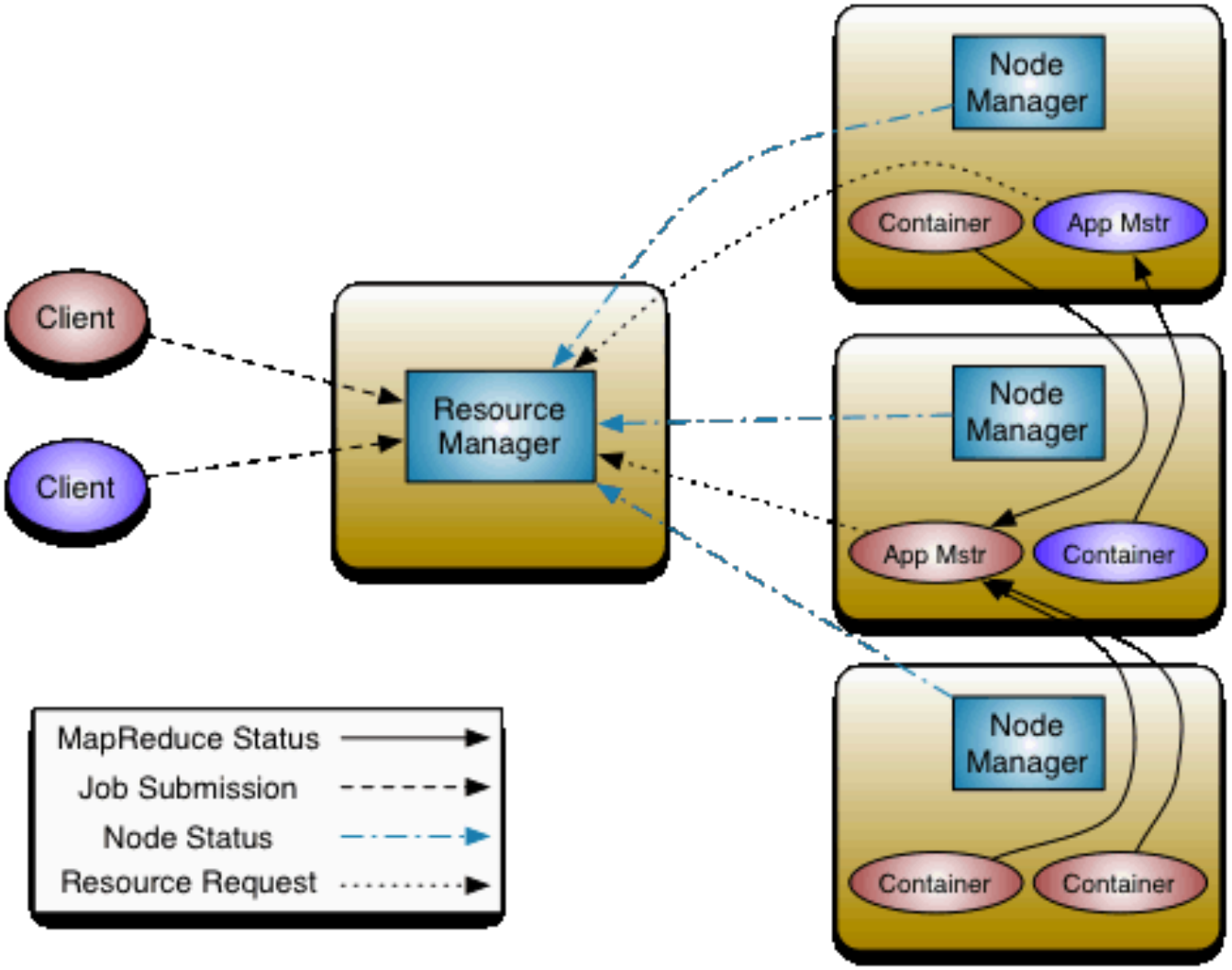
To verify Hadoop releases using GPG:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a mirror site.
2. Download the signature file hadoop-X.Y.Z-src.tar.gz.asc from Apache.
3. Download the Hadoop KEYS file.
4. gpg --import KEYS
5. gpg --verify hadoop-X.Y.Z-src.tar.gz.asc

To perform a quick check using SHA-256:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a mirror site.
2. Download the checksum hadoop-X.Y.Z-src.tar.gz.mds from Apache.

Apache Hadoop YARN



Source: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

Apache Spark

Lightning-fast cluster computing

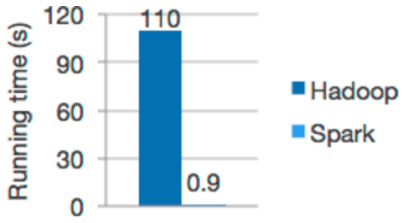
Download | Libraries | Documentation | Examples | Community | Developers | Apache Software Foundation

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

Speed

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

Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

Ease of Use

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

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python and R shells.

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

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

Word count in Spark's Python API

Latest News

Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (Jan 04, 2017)

Spark 2.1.0 released (Dec 28, 2016)

Spark wins CloudSort Benchmark as the most efficient engine (Nov 15, 2016)

Spark 2.0.2 released (Nov 14, 2016)

Archive

Download Spark

Built-in Libraries:

SQL and DataFrames
Spark Streaming
MLlib (machine learning)
GraphX (graph)

Third-Party Projects

http://spark.apache.org/

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