# Before Lecture…

- Go over Syllabus
- Take a look at an example portfolio (link)
  - 9/21 will distribute out:
    - Portfolio Template (will try to set up in class)
    - Assignment 1 Description
- Student Research Proposal and Final Project
  - Proposal Due: 9/28
  - Final Project Due: 12/7

# Student Final Project Format (50%)

Research Proposal (Due: 9/28)

- **Overview, and Motivation**: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.
- **Related Work and Audience**: Anything that inspired you, such as a paper, a web site, or something we discussed in class. Who is your audience. Who will you show your results?
- **Initial Questions**: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?
- **Data**: Source, scraping method, cleanup, storage, etc.

Final Paper (also includes Abstract, Overview, Motivation, Related Work, Initial Questions and Data from the Research Proposal):

- **Exploratory Data Analysis**: What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?
- **Final Analysis**: What did you learn about the data? How did you answer the questions? How can you justify your answers?
- **Presentation**: Present your final results in a compelling and engaging way using text, visualizations, images, and videos on your project web site.

# Lecture 2: Understanding Big Data and Motivation/Drivers for Big Data Adoption

Benjamin Simeon Harvey (Ben)
Adjunct Faculty, EMSE Department
The George Washington University
E-mail: bsharve@nsa.gov
Web: TBD

# Agenda

Course Timeline and Course Goals and "How does this lecture fit?"

Chapter 1: Understanding Big Data

Fundamental Concepts and Terminology

- Data Analysis
- Data Analytics
- Types of Data Analytics
  - Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics
- Business Intelligence (BI)
- Key Performance Indicators (KPI)

Big Data Characteristics

- Volume, Velocity, Variety, Veracity, Value

Different Types of Data

- Structured Data
- Unstructured Data
- Semi-structured Data
- Metadata

# Agenda

Chapter 2: Business Motivations and Drivers for Big Data Adoption

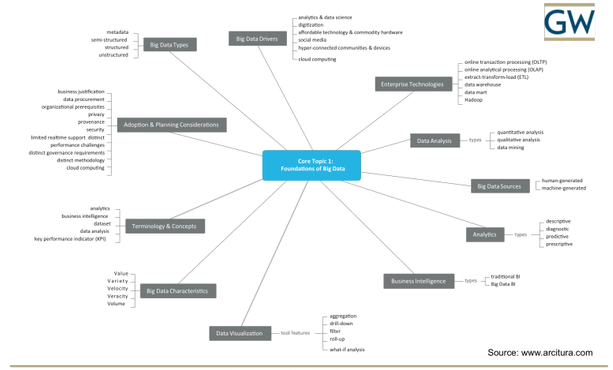
- Marketplace Dynamics
- Business Architecture
- Business Process Management
- Information and Communication Technology
  - Data Analytics and Data Science
  - Digitization
  - Affordable Technology and Commodity Hardware
  - Social Media
  - Hyper-Connected Communities and Devices
  - Cloud Computing

Internet of Everything

# Course Timeline and How does this fit?

## Foundations of Big Data

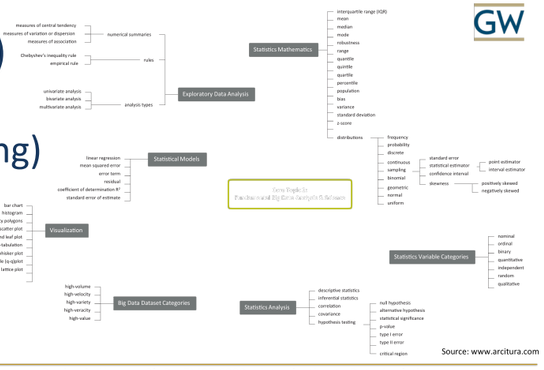
- **Understanding Big Data, Ch. 1&2 – Erl (9/7)**
  - **Fundamental Terminology & Concepts**
  - **Big Data Business Intelligence & Technology Drivers**
- Big Data Adoption, Planning and Enterprise Technologies, Ch. 3&4 – Erl (9/14)
  - Big Data Adoption & Planning Considerations
  - Traditional Enterprise Technologies Related to Big Data
- Big Data Storage, Ch. 5 – Erl (9/21)
  - Characteristics of Data in Big Data Environments
  - Dataset Types in Big Data Environments
- Big Data Processing, Ch. 6 – Erl (10/5)
- Big Data Storage Technology, Ch. 7 – Erl (10/12)
  - Data Engineering (Ch14 - Schutt)
- Big Data Fundamental Analysis and Analytics, Ch. 8 – Erl (10/19)



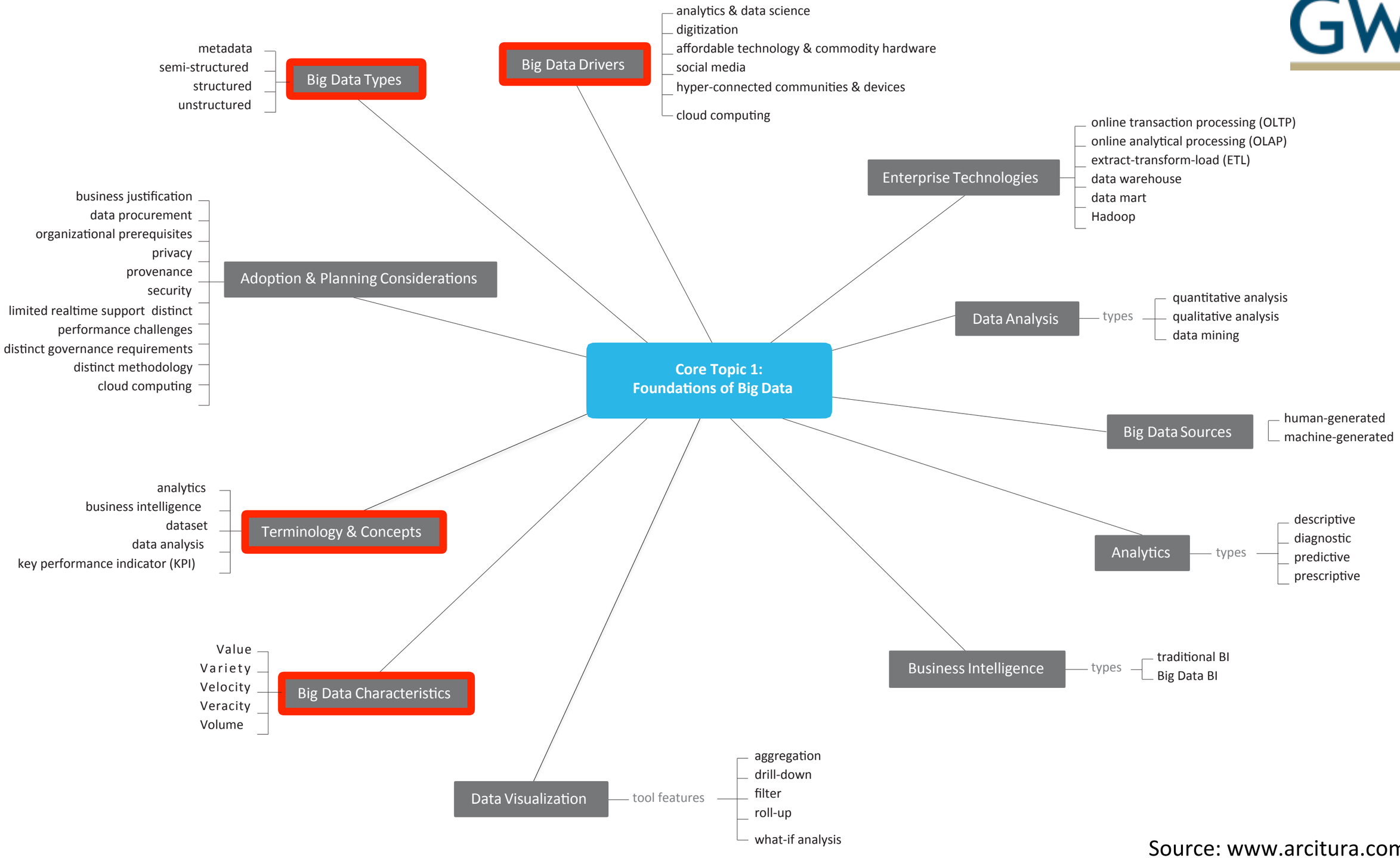



## Fundamental Big Data Analysis and Science

- Statistical Inference, EDA, and extracting meaning from data, Ch. 2 & 7 - Schutt (10/26)
  - Essential Statistics
  - Statistics Analysis (including descriptive, inferential, correlation, covariance & hypothesis testing)
- Algorithms and Regression, Ch. 3&5 - Schutt (11/2)
- Machine Learning I, Ch. 4 & 6 - Schutt (11/16)
- Data Visualization, Ch. 9 - Schutt (11/30)
- Machine Learning II (12/7)

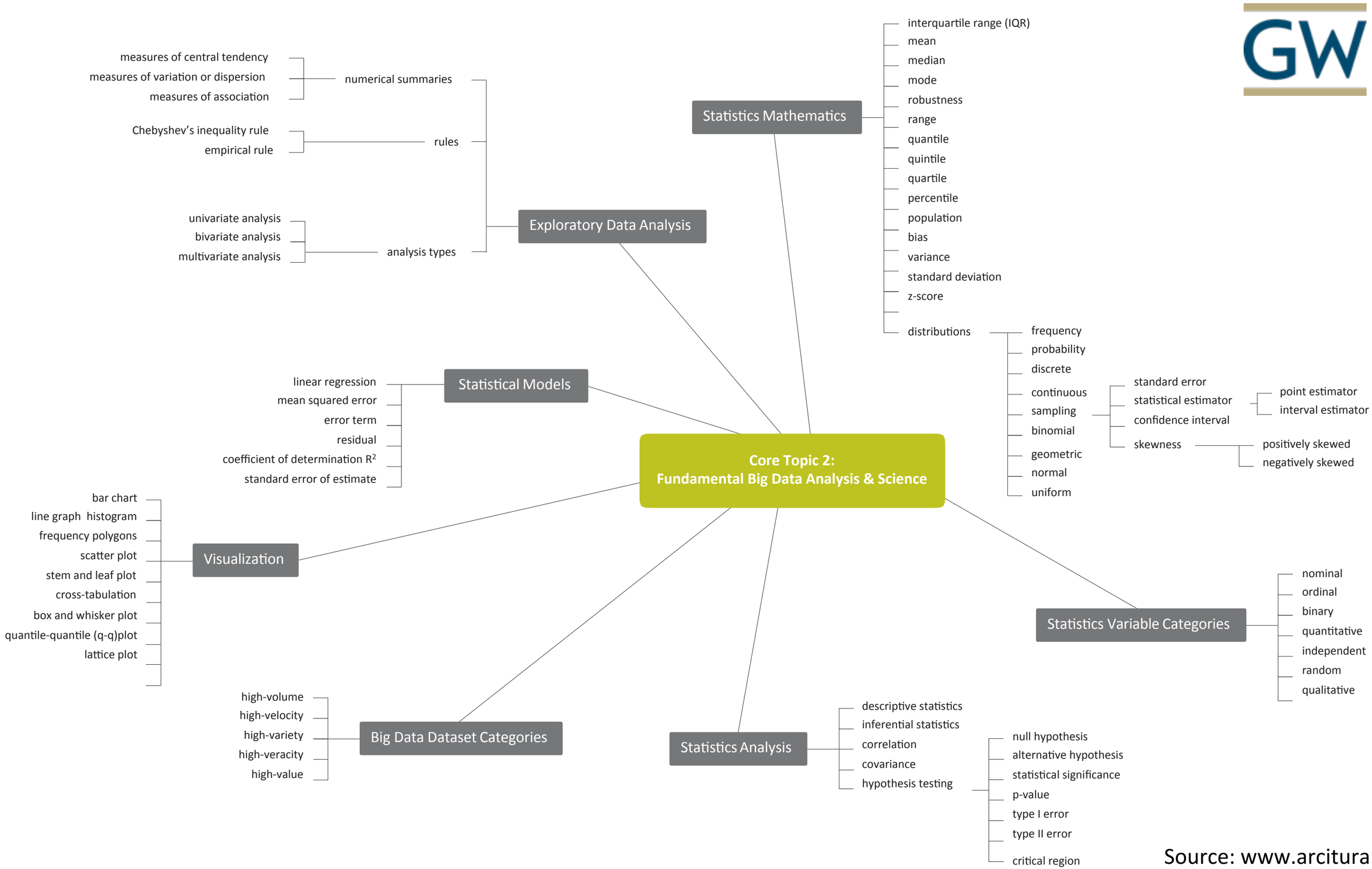


# Core Topic 1. Foundations of Big Data




Source: www.arcitura.com

# Core Topic 2. Fundamental Big Data Analysis and Science




Source: www.arcitura.com

# Concepts and Terminology

Data Analysis

- Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends. The overall goal of data analysis is to support better decision making.

Data Analytics

- Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and **governing data**.

Big Data Analytics

- The lifecycle generally involves identifying, procuring, preparing and analyzing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches.

Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone.

- descriptive analytics
- diagnostic analytics
- predictive analytics
- prescriptive analytics

# Concepts and Terminology

Descriptive Analytics

- Descriptive analytics are carried out to answer questions about events that have already occurred.

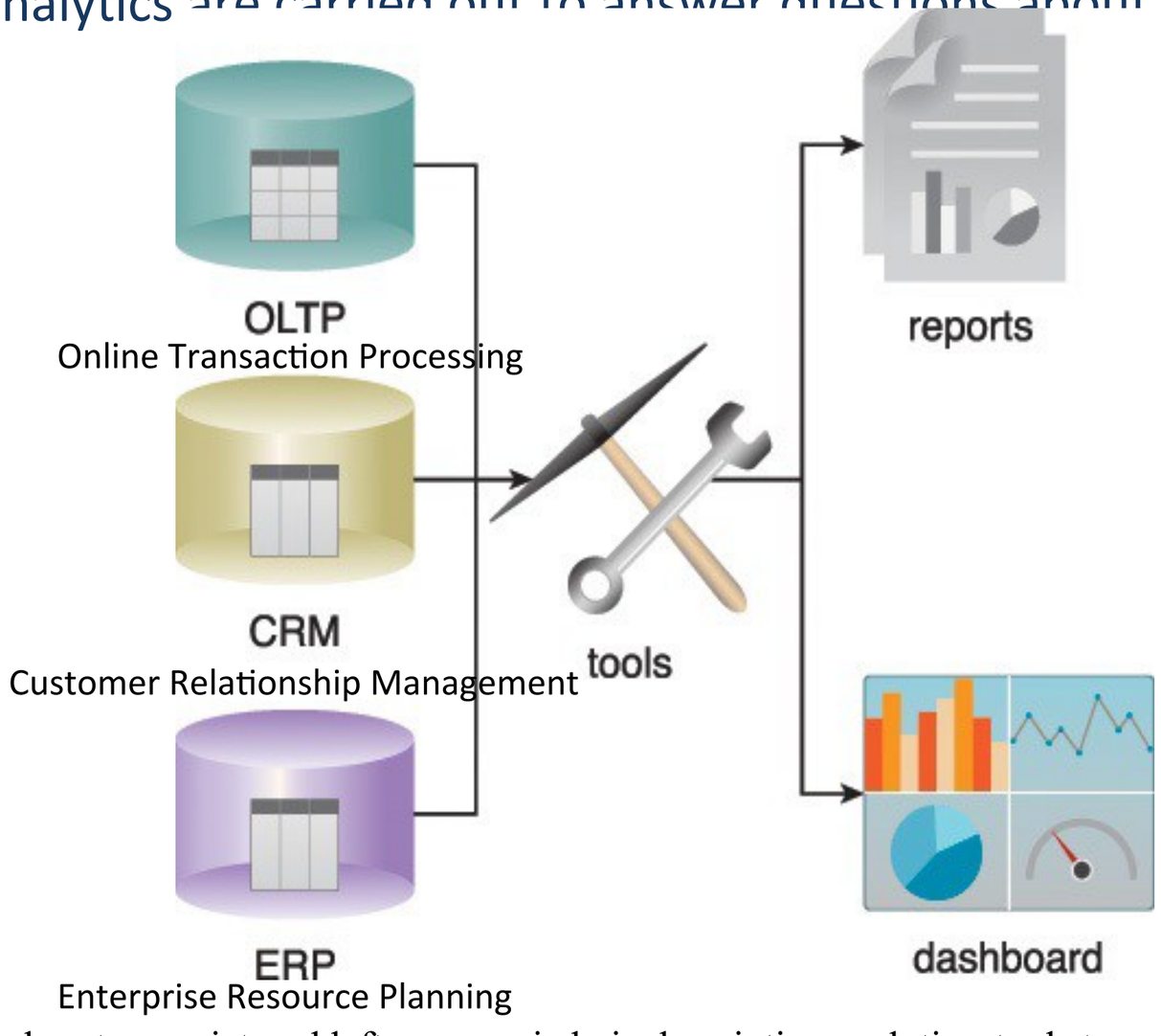



**Figure 1.5** The operational systems, pictured left, are queried via descriptive analytics tools to generate reports or dashboards, pictured right.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Concepts and Terminology

Diagnostic Analytics

- Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
- The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something h

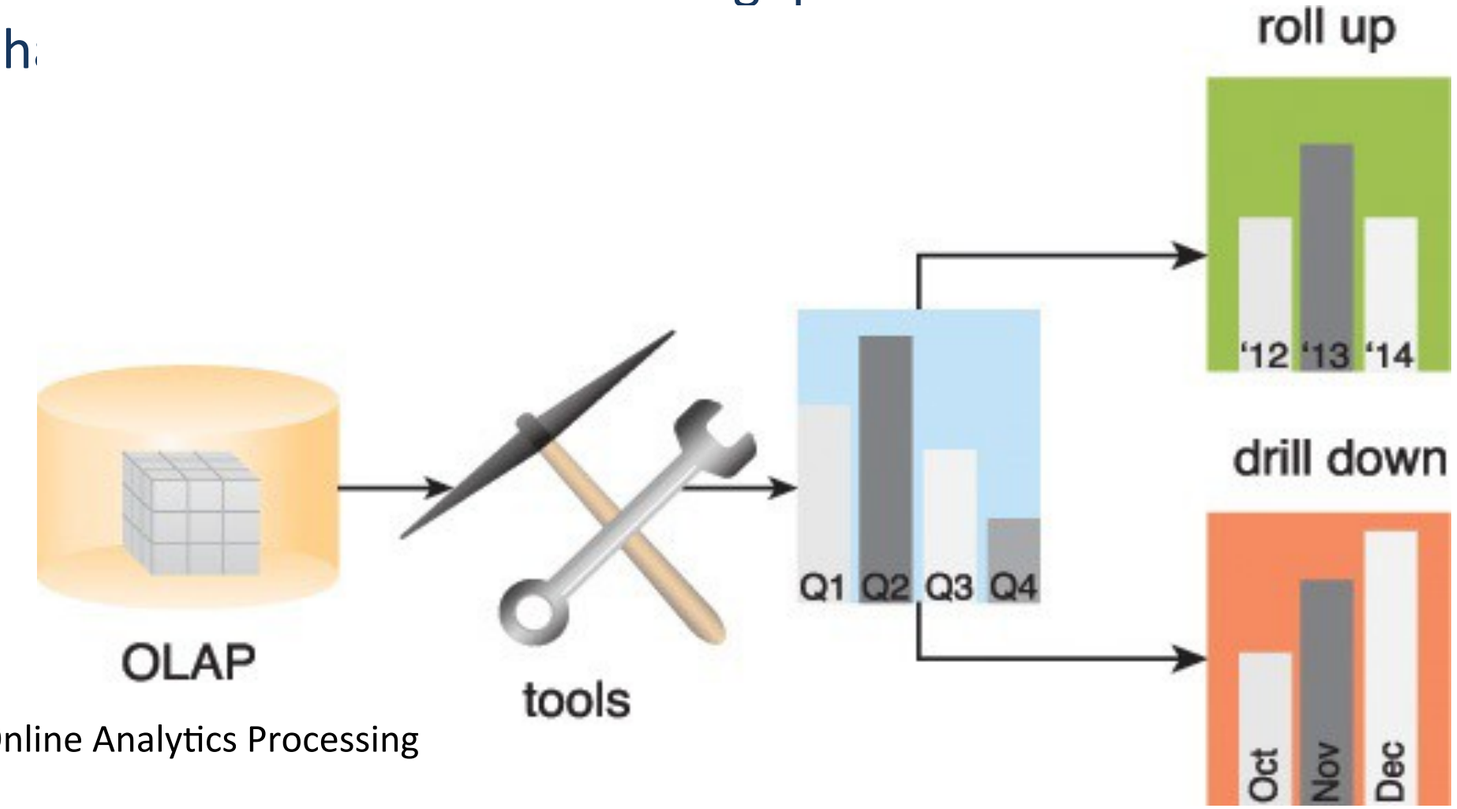



**Figure 1.6** Diagnostic analytics can result in data that is suitable for performing drill- down and roll-up analysis.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Concepts and Terminology

Predictive Analytics

- Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. Information is enhanced with meaning to generate knowledge that conveys how that information is related.

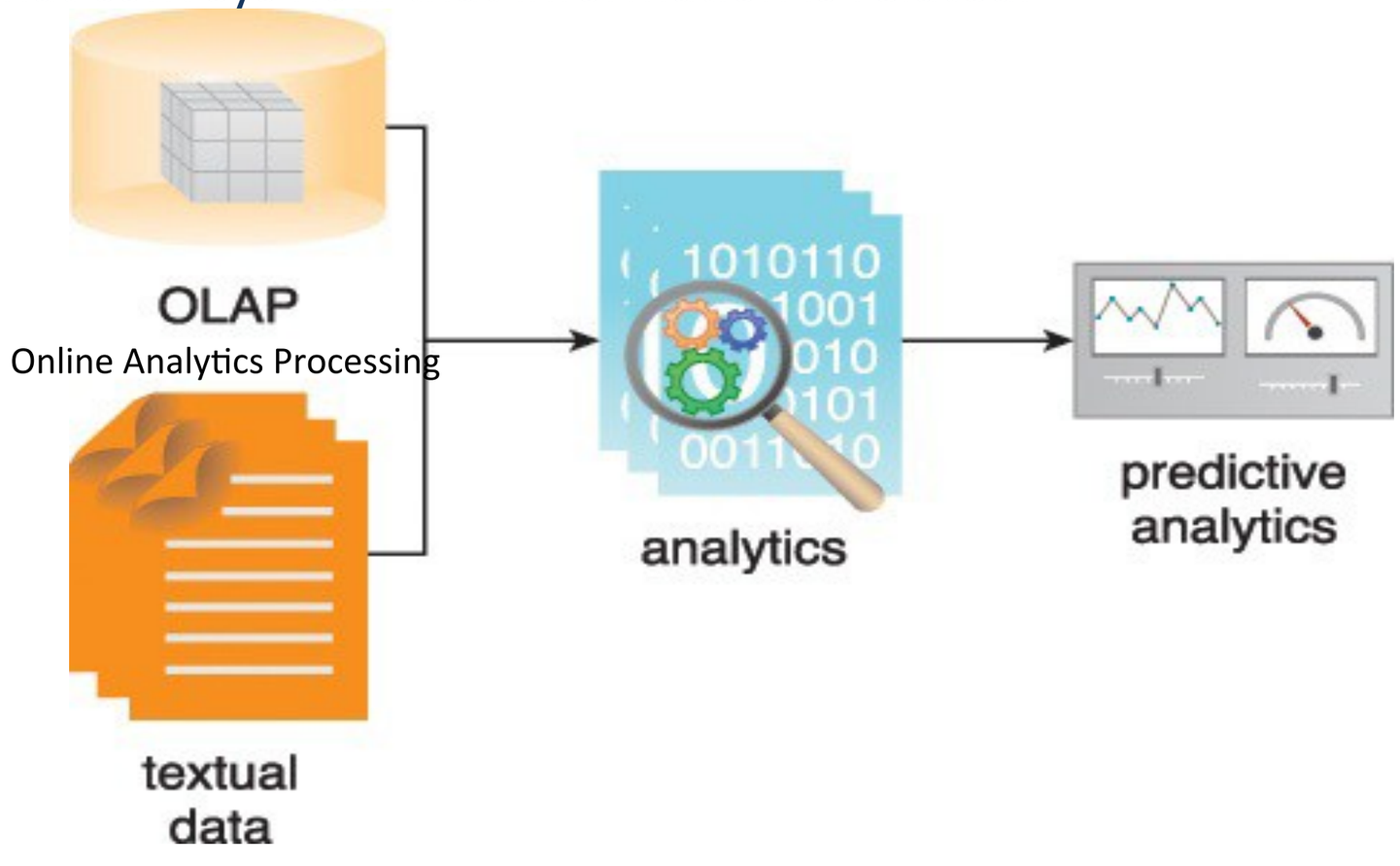



**Figure 1.7** Predictive analytics tools can provide user-friendly front-end interfaces.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Concepts and Terminology

Prescriptive Analytics

- Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken (best option and why).

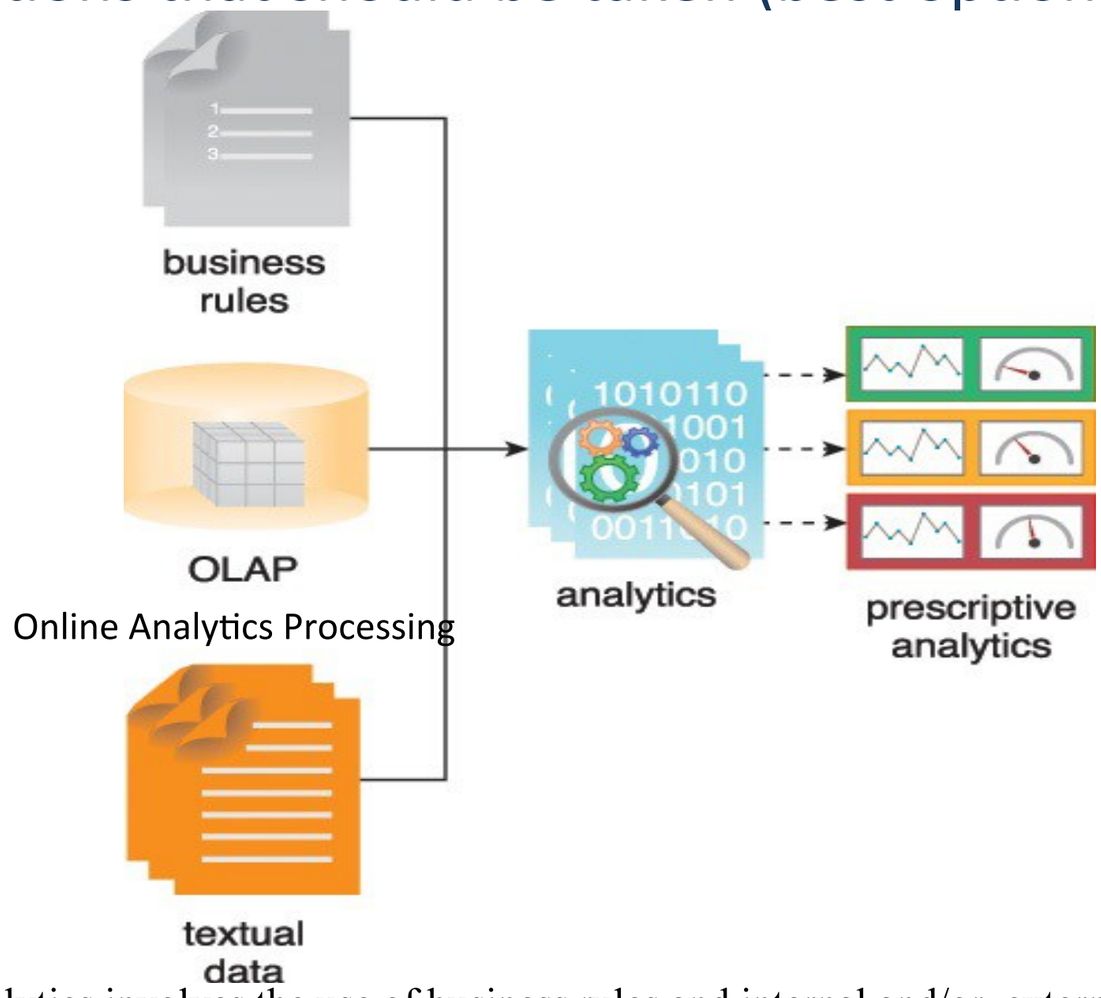



**Figure 1.8** Prescriptive analytics involves the use of business rules and internal and/or external data to perform an in-depth analysis.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Concepts and Terminology

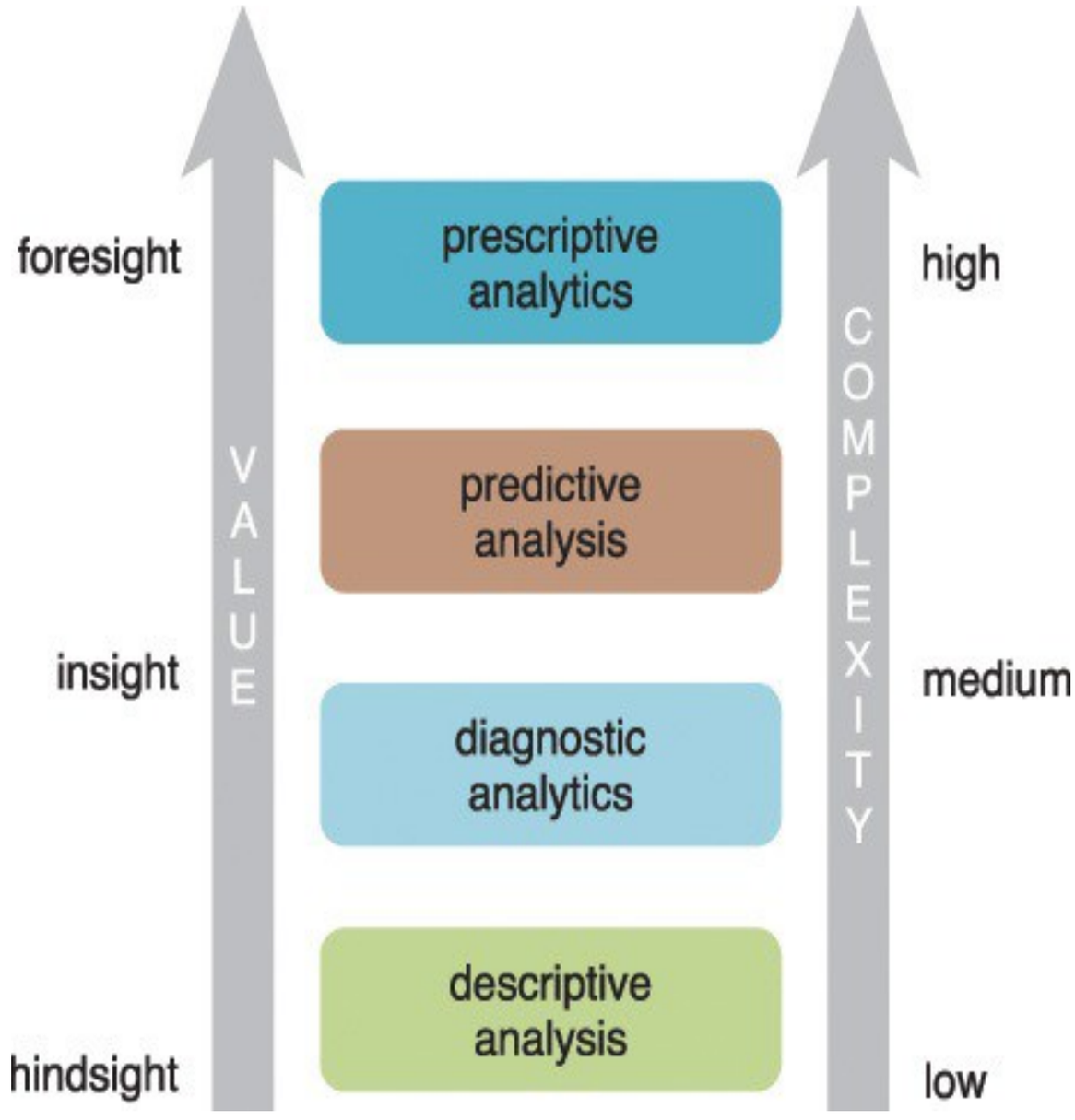



**Figure 1.4** Value and complexity increase from descriptive to prescriptive analytics.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Concepts and Terminology

Key Performance Indicators (KPI)

- A KPI is a metric that can be used to gauge success within a particular business context.
- KPIs are linked with an enterprise's overall strategic goals and objectives.
- They are often used to identify business performance problems and demonstrate regulatory compliance.

Business Intelligence (BI)

- BI enables an organization to gain insight into the performance of an enterprise by analyzing data generated by its business processes and information systems.
- The results of the analysis can be used by management to steer the business in an effort to [illegible]formance

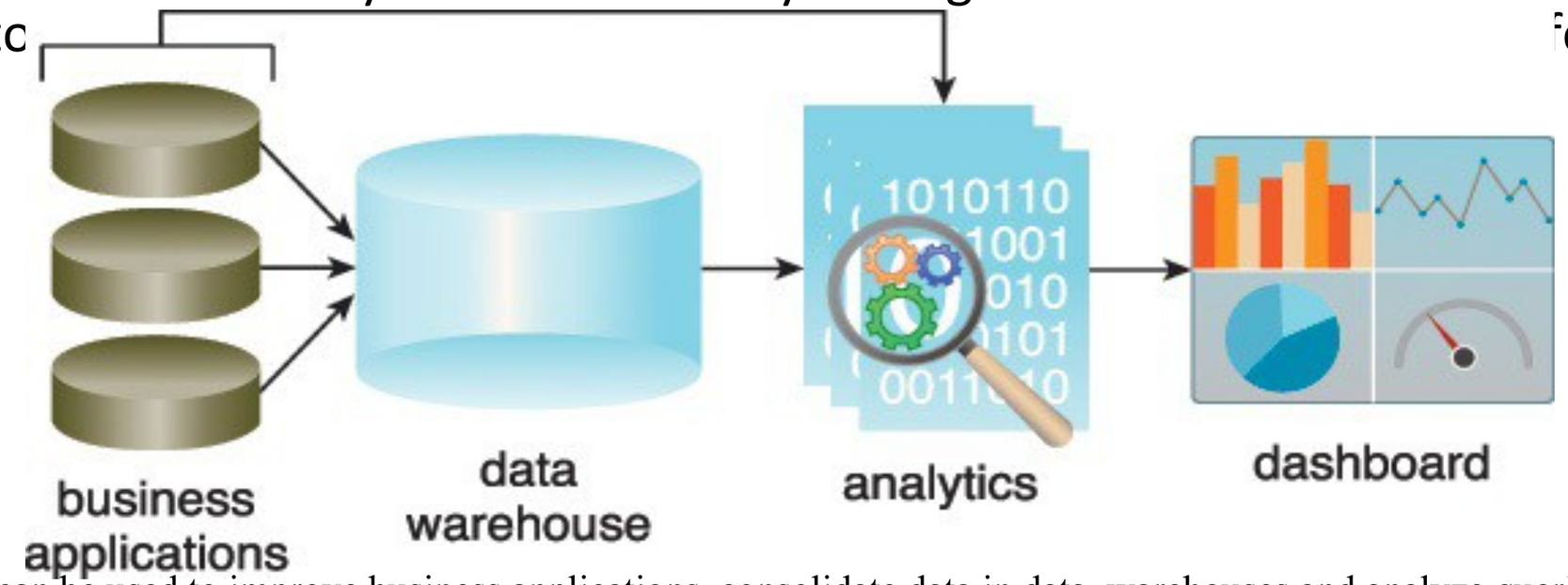



**Figure 1.9** BI can be used to improve business applications, consolidate data in data warehouses and analyze queries via a dashboard.
Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Big Data Characteristics

Volume

- **The anticipated amount/volume of data that is processed** by Big Data solutions is substantial and ever-growing. High data volumes impose distinct data storage and processing demands, as well as additional data preparation, curation and management processes.

Velocity

- In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time. From an enterprise's point of view, the **velocity of data translates into the amount of time it takes for the data to be processed once it enters the enterprise's perimeter.**

Variety

- Data variety refers to **the multiple formats and types of data that need to be supported** by Big Data solutions. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.

Veracity

- Veracity refers to the **quality or fidelity of data**. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise.

Value

- Value is defined as the **usefulness of data for an enterprise**. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.

# Big Data Characteristics

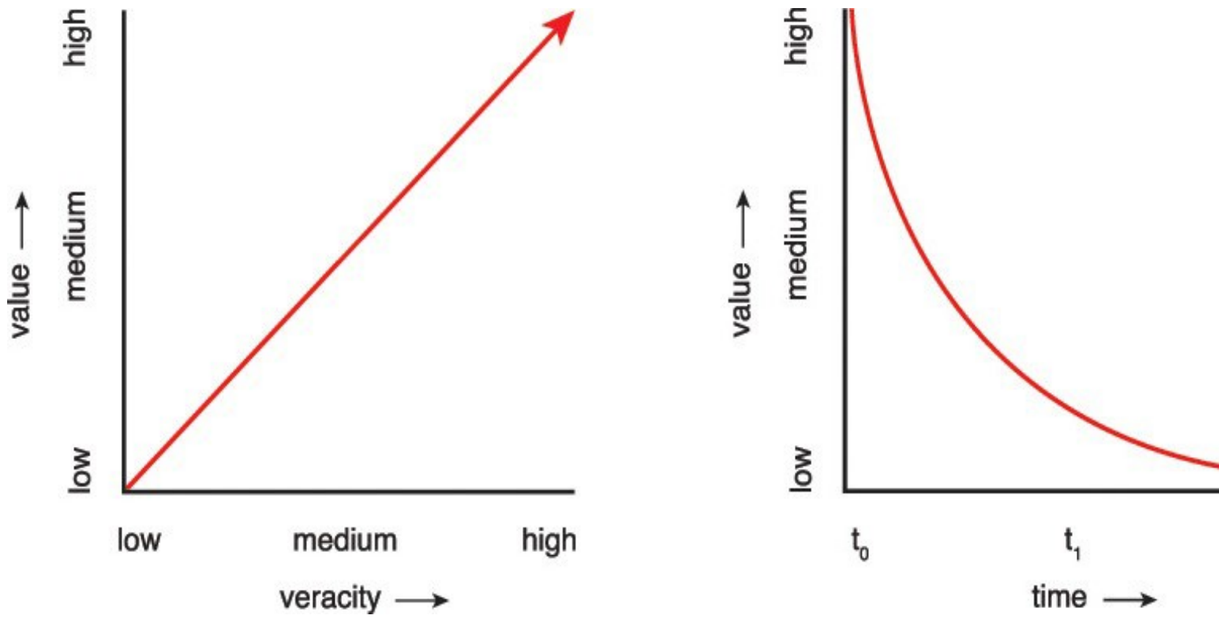
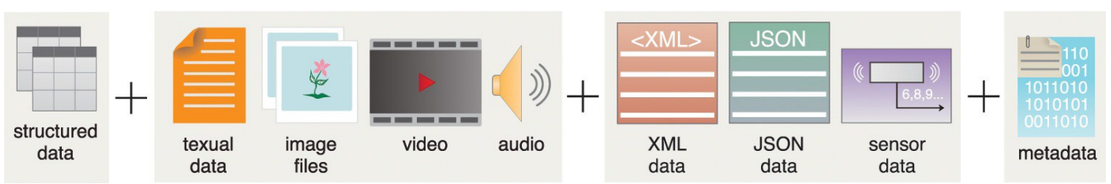
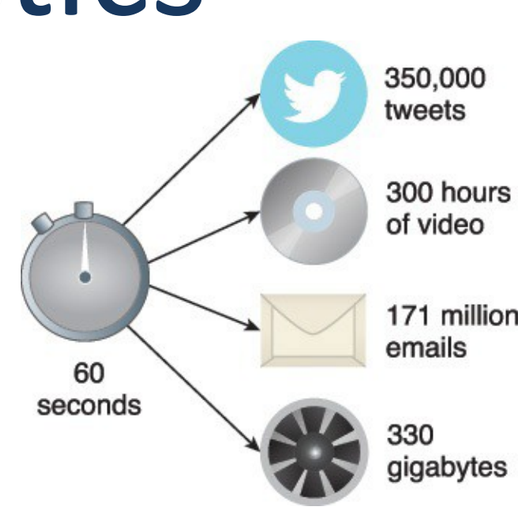
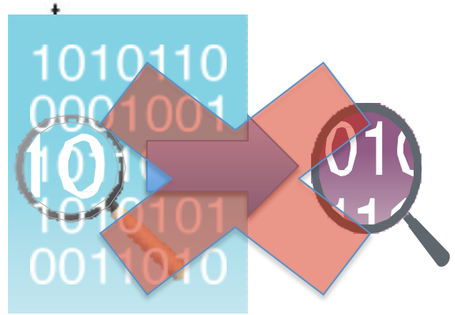



**Figure 1.15** Data that has high veracity and can be analyzed quickly has more value to a business





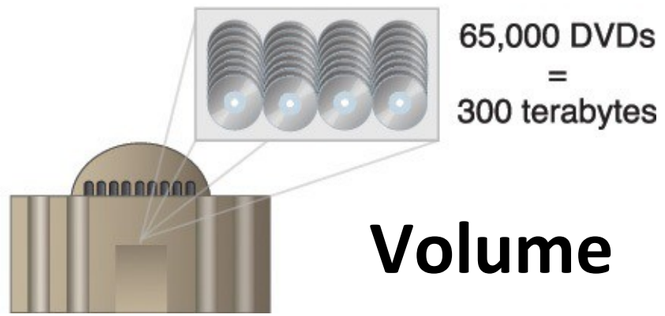



**Figure 1.12-15** The Five Vs of Big Data.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Concepts and Terminology

Structured Data

- **Structured data conforms to a data model or schema** and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database.

Semi-structured Data

- Semi-structured data has a **defined level of structure and consistency, but is not relational in nature**. Instead, semi-structured data is hierarchical or graph-based.

Unstructured Data

- Data that does not conform to a data model or data schema is known as unstructured data.

Metadata

- provides information about a dataset's characteristics and structure.

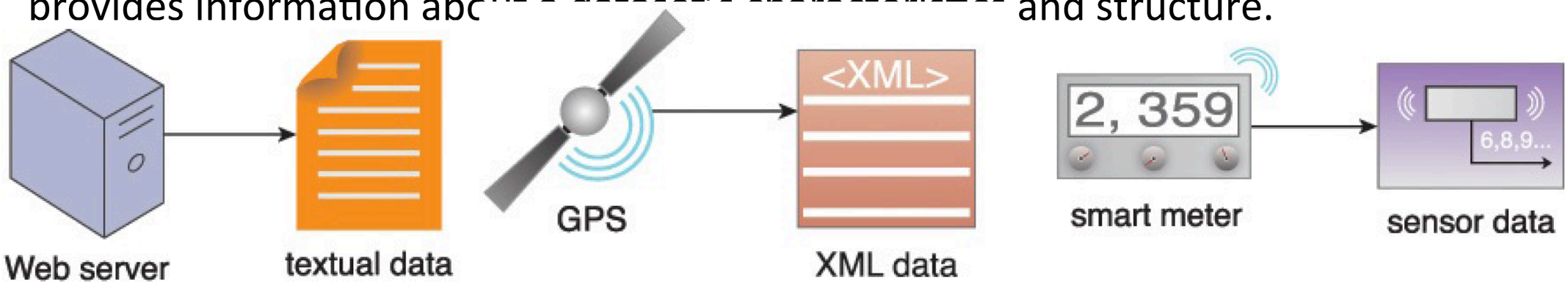
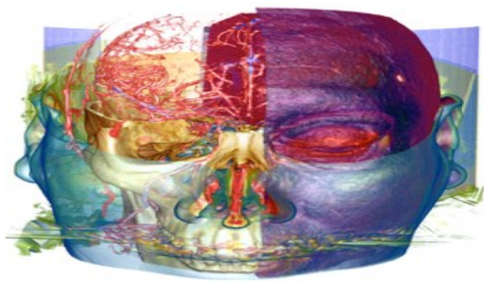



**Figure 1.17** Examples of machine-generated data include web logs, sensor data, telemetry data, smart meter data and appliance usage data.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Summary: Understanding Big Data

Data Science: To gain insights through computation, statistics, and visualization




- In this course, you will learn four core topics that have been associated with Data Science
  - Big Data Foundations, Big Data and Technology Concepts, Big Data Analysis and Science, Advanced Big Data Analysis and Science
- The definition of Big Data is subjective and really depends of the following:
  - Volume, Velocity, Veracity, Variety of data
  - Current CPU capacity

# Case Study: Ensure to Insure

Introduction to Case Studies

Identifying Types of Data

- The IT team members go through a categorization exercise of the various datasets that could be identified:
  - Categorize them as either: Structured, Unstructured and Semi-Structured

Identifying Data Characteristics

- The IT team members want to gauge different datasets that are generated inside ETI's boundary as well as any other data generated outside ETI's boundary that may be of interest to the company in the context of volume, velocity, variety, veracity and value characteristics.
- The team members take each characteristic in turn and discuss how different datasets manifest that characteristic.

Do you think the new Big Data strategy will work? Justify your answer.  What' s missing (data types and characteristics)?

# Motivation and Drivers for Big Data Adoption

Here we will explore the business motivations and drivers behind the adoption of Big Data solutions and technologies:

- Marketplace Dynamics,
- An appreciation and formalism of Business Architecture (BA),
- The realization that a business' ability to deliver value is directly tied to Business Process Management (BPM),
- Innovation and Information and Communications Technology (ICT), and
- The Internet of Everything

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Motivation and Drivers for Big Data Adoption

Motivation and drivers include:

- Marketplace Dynamics,
- An appreciation and formalism of Business Architecture (BA),
- The realization that a business' ability to deliver value is directly tied to Business Process Management (BPM),
- Innovation and Information and Communications Technology (ICT), and
- The Internet of Everything

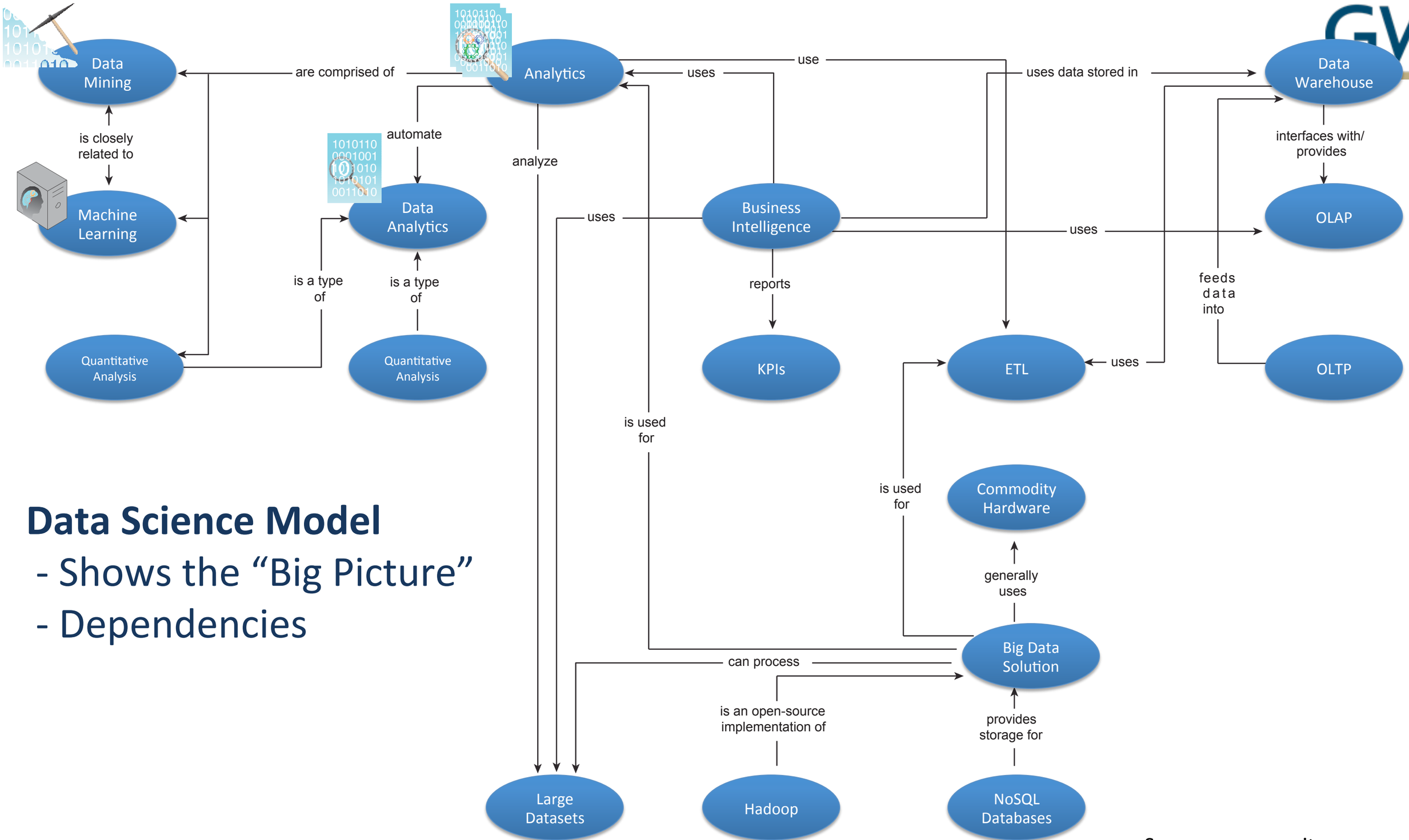
# Motivation and Drivers for Big Data Adoption

Marketplace Dynamics:

- Businesses working to improve efficiency and effectiveness to stabilize their profitability by reducing costs.
- Companies conduct transformation projects to improve their corporate processes to achieve savings.
- Data – a discrete, objective facts about events.
- Information – data that makes a difference
- Knowledge – a fluid mix of framed experience, values, contextual information and expert insight that provides a framework for evaluating and incorporating new experiences and information.

# Data Science Model

- Shows the “Big Picture”
- Dependencies




Source: www.arcitura.com

# Motivation and Drivers for Big Data Adoption

- Latitude − Longitude
- Jacksonville FL
- Benjamin Harvey − 10668 Pine Rd Jax FL
- Ben will be at home at 6pm today

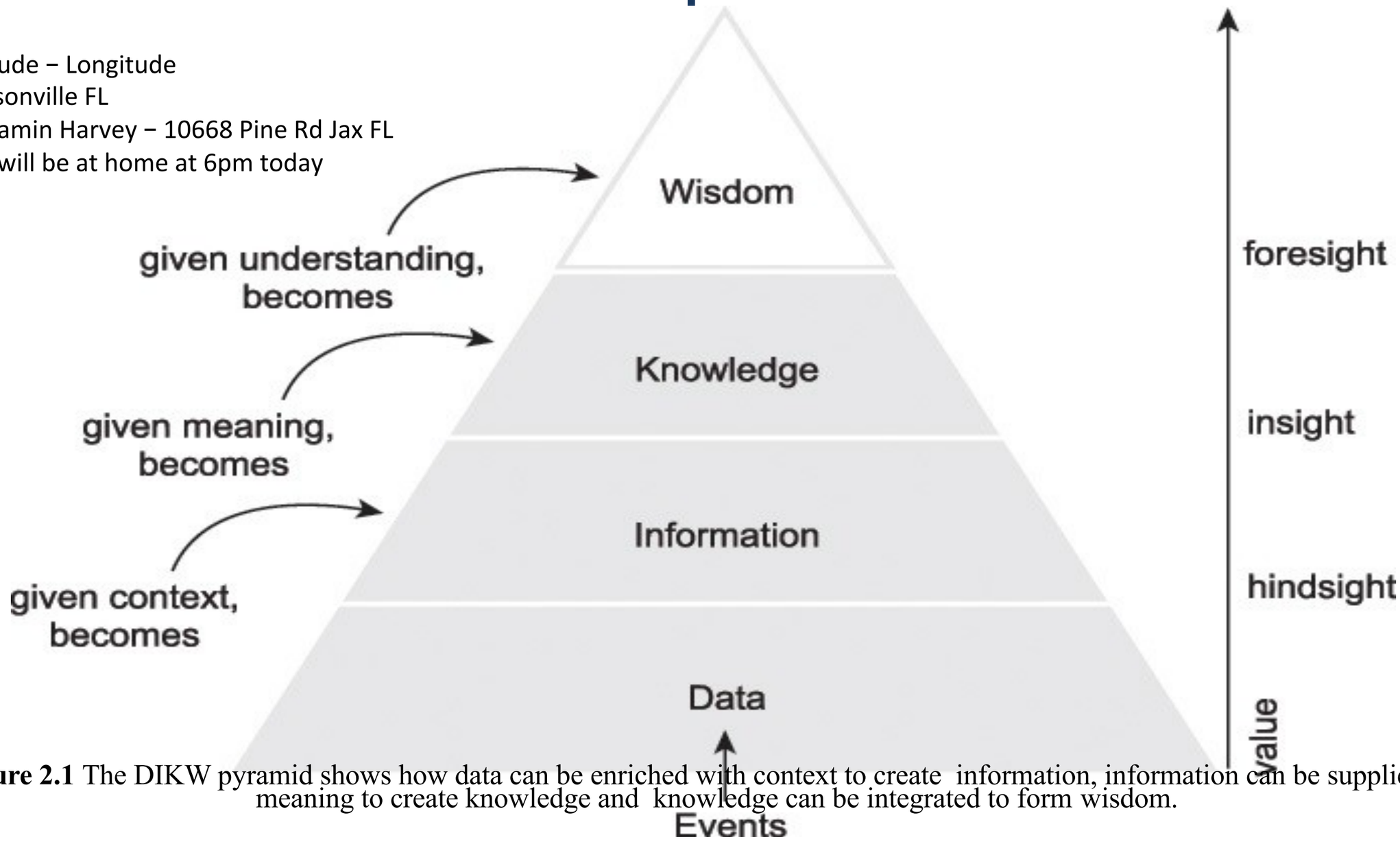



**Figure 2.1** The DIKW pyramid shows how data can be enriched with context to create information, information can be supplied with meaning to create knowledge and knowledge can be integrated to form wisdom.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Motivation and Drivers for Big Data Adoption

Business Architecture:

- Provides a means of blueprinting or concretely expressing the design and the processes of the business
- Helps and organization align its strategic vision, goals, objectives with its underlying execution.
- Layered system:
- Strategic Layer– C-Level Executives and Advisory Groups
- Managerial Layer – seeks to steer the organization in alignment with the strategy
- Operations Layer – where a business executes its core processes and delivers value to its customers

# Motivation and Drivers for Big Data Adoption

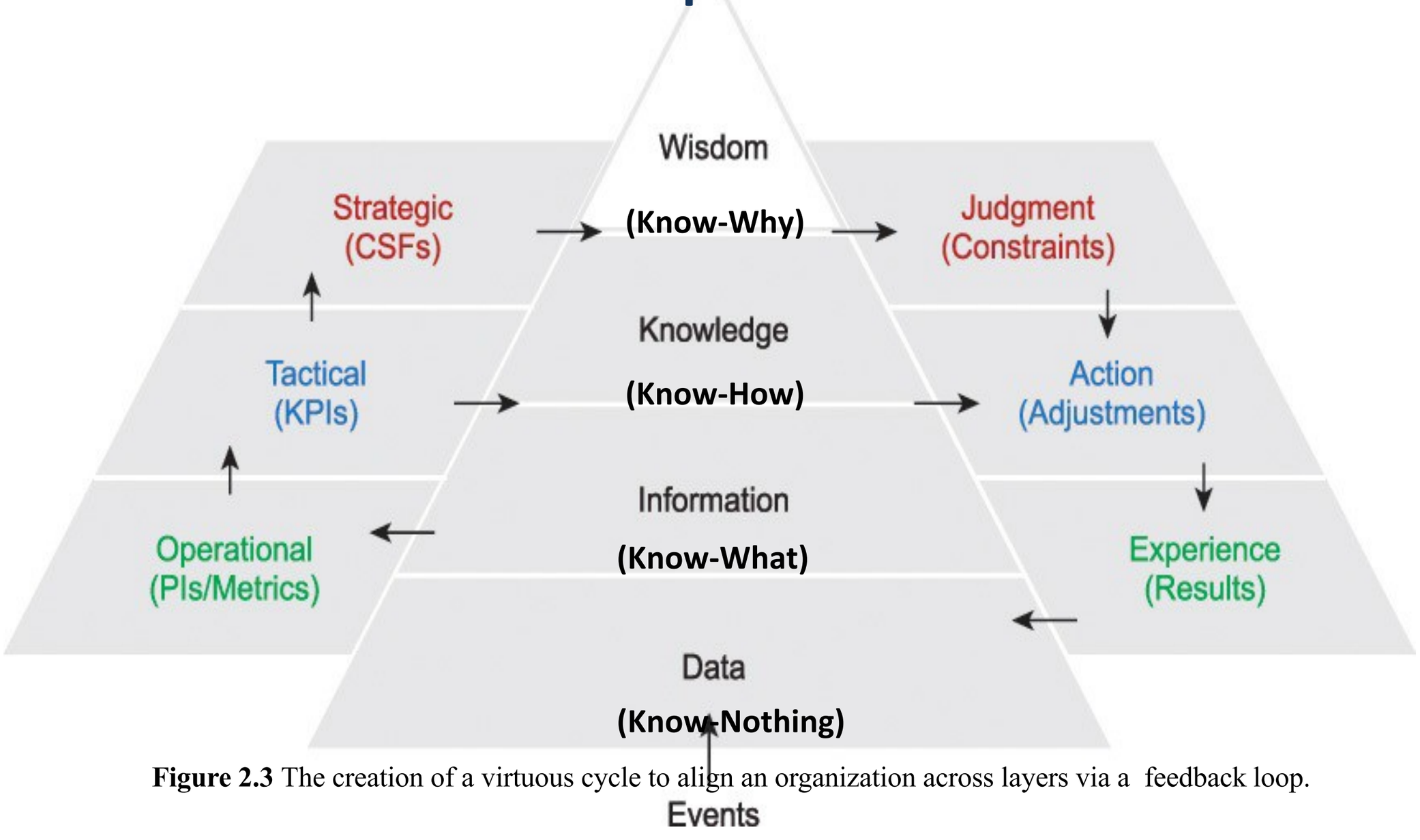



**Figure 2.3** The creation of a virtuous cycle to align an organization across layers via a feedback loop.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Motivation and Drivers for Big Data Adoption

Business Process Management:

- A business process is a description of how work is performed in an organization. It describes all work-related activities and their relationships, aligned with the organizational actors and resources responsible for conducting them.
- When the combination of Big Data analytic results and goal-driven behavior are used together, process execution can become adaptive to the marketplace and responsive to environmental conditions.

# Motivation and Drivers for Big Data Adoption

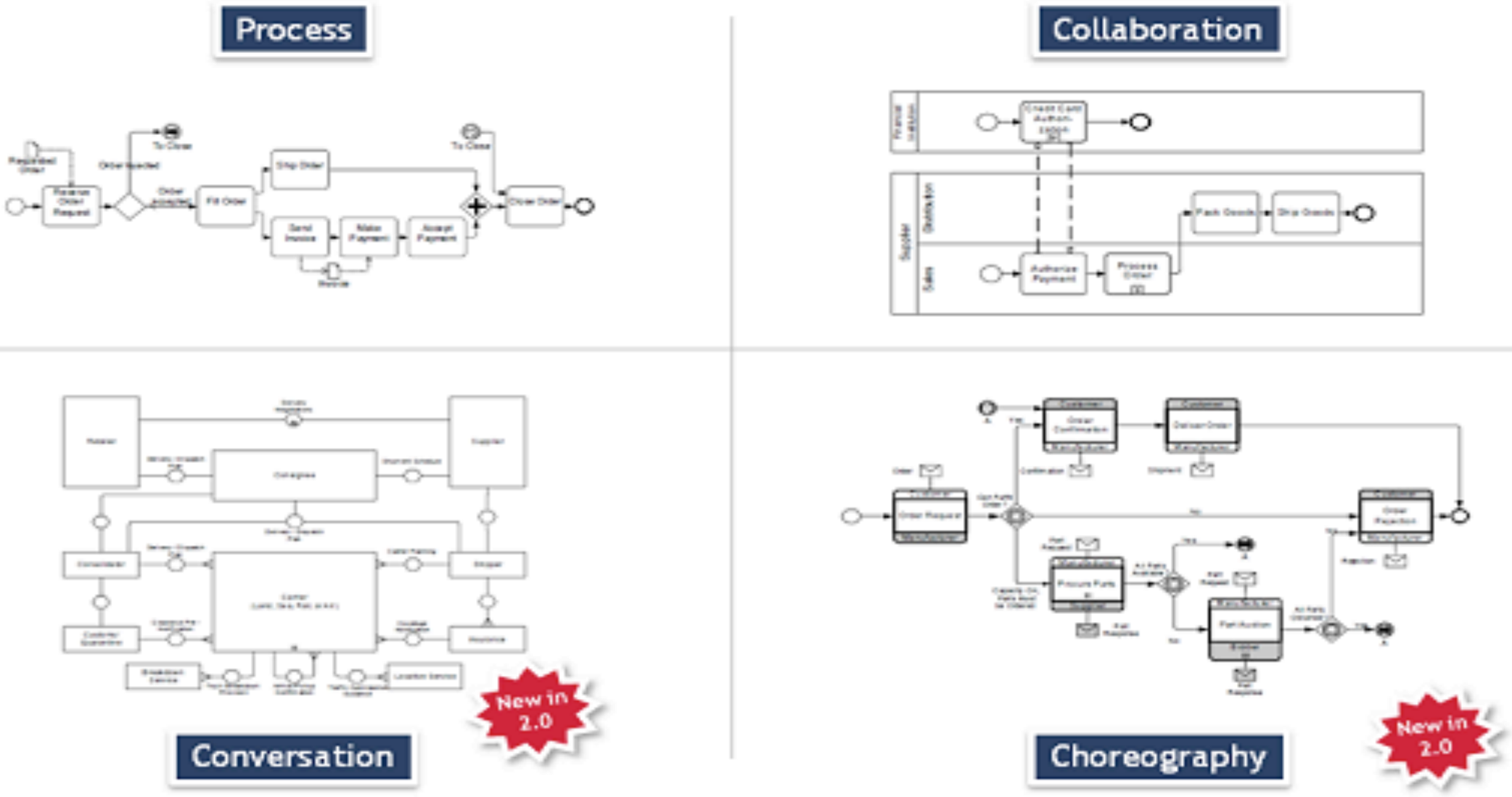



**Figure**: BPMN four model types
Source: http://www.ariscommunity.com/users/roland-woldt/2011-01-28-learning-bpmn-2-which-models-are-available-bpmn

# Motivation and Drivers for Big Data Adoption

Information and Communications Technology

- ICT developments that have accelerated the pace of Big Data adoption include:
  - Data analytics and data science
  - Digitization
  - Affordable technology and commodity hardware
  - Social media
  - Hyper-connected communities and devices
  - Cloud computing

# Information and Communications Technology

- Data analytics and data science
  - Enterprises are collecting, procuring, storing, curating, and processing increasing quantities of data to find new insights that can drive more efficient and effective operations.
  - Companies are looking for new ways to gain a competitive edge by using Data Science tools and technologies to extract meaningful information and insights.
- Digitization
  - Businesses have replaced physical mediums as the de facto communications and delivery mechanisms.
  - This allows for the opportunity to collect secondary data for analysis
- Affordable technology and commodity hardware
  - Technology capable of storing and processing large quantities of diverse data has become increasingly affordable.
  - Big data solutions also often leverage open-source software that executes on commodity hardware, further reducing costs.

# Information and Communications Technology

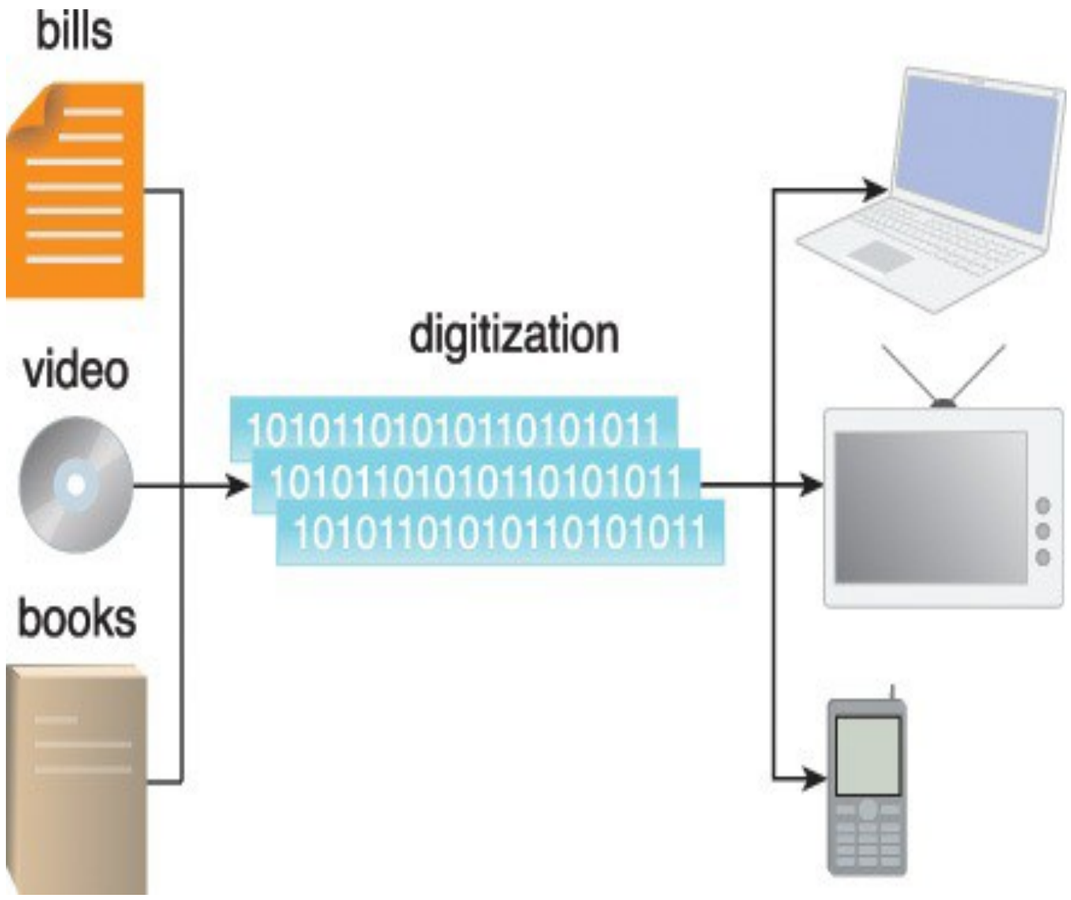



**Figure 2.4** Examples of digitization including online banking, on-demand television and streaming video.

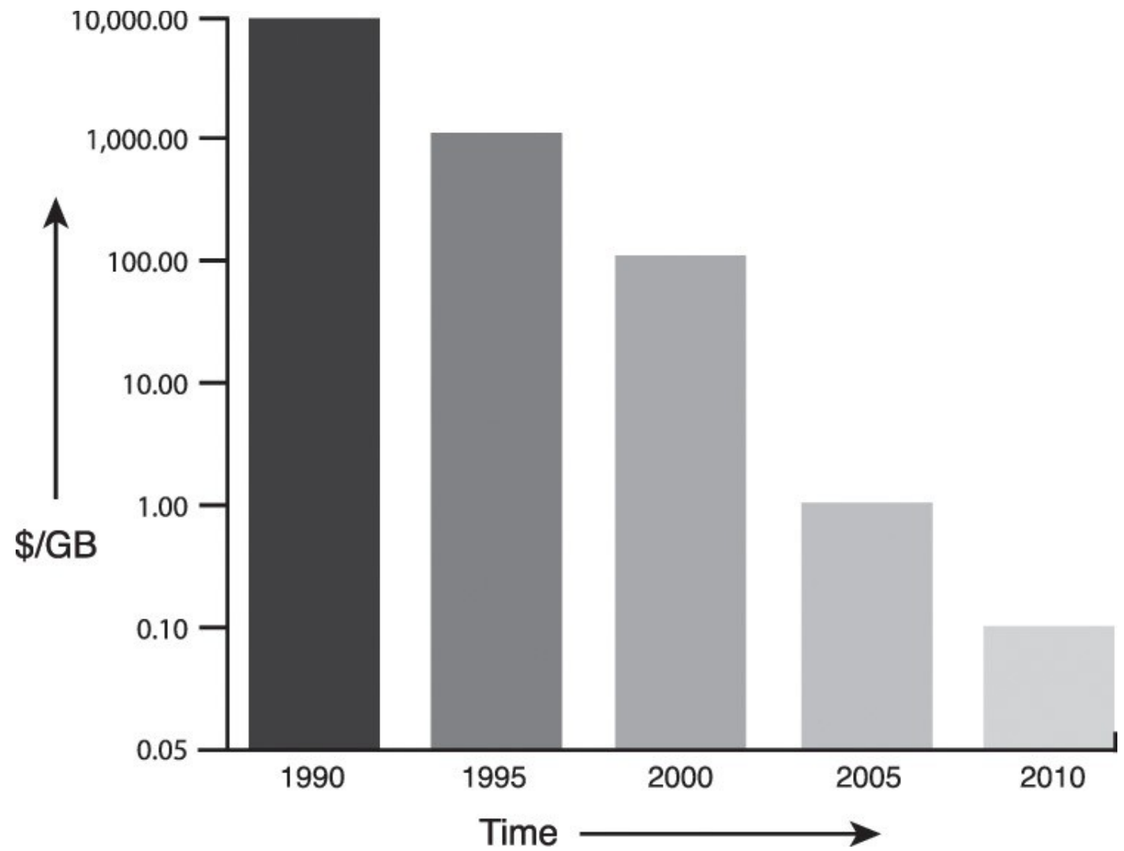



**Figure 2.5** Data storage prices have dropped dramatically from more than $10,000 to less than $0.10 per GB over the decades.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

# Information and Communications Technology

- Social media
  - The emergence of social media has empowered costumers to provide feedback in near real-time via open and public mediums.
  - The shift has forced businesses utilize CRM's to analyze customer feedback on their service and product offerings
- Hyper-connected communities and devices
  - The broadening coverage of the Internet and the proliferation of cellular, Wi-Fi and sensor networks has enabled more people and their devices to be continuously active in virtual communities.
  - IoT – a vast collection of smart Internet-connected devices
  - IoE – combines the services provided by smart connected devices of the IoT into meaningful business processes that posses the ability to provide unique differentiating value propositions.
- Cloud computing
  - Cloud computing advancements have led to the creating of environments that are capable of providing highly scalable, parallel, distributed on-demand IT resources.

# Information and Communications Technology

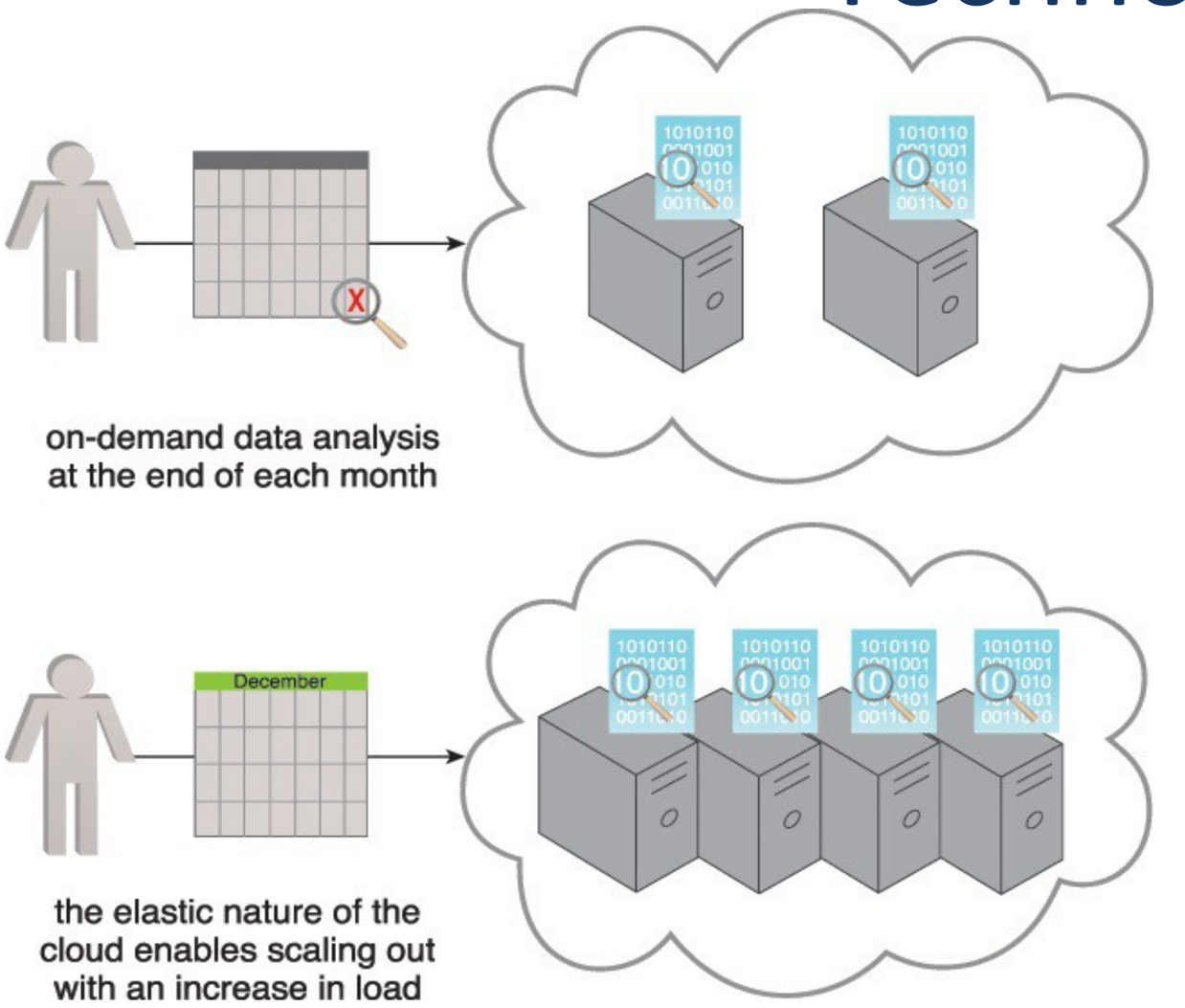



**Figure 2.7** The cloud can be used to complete on-demand data analysis at the end of each month or enable the scaling out of systems with an increase in load.

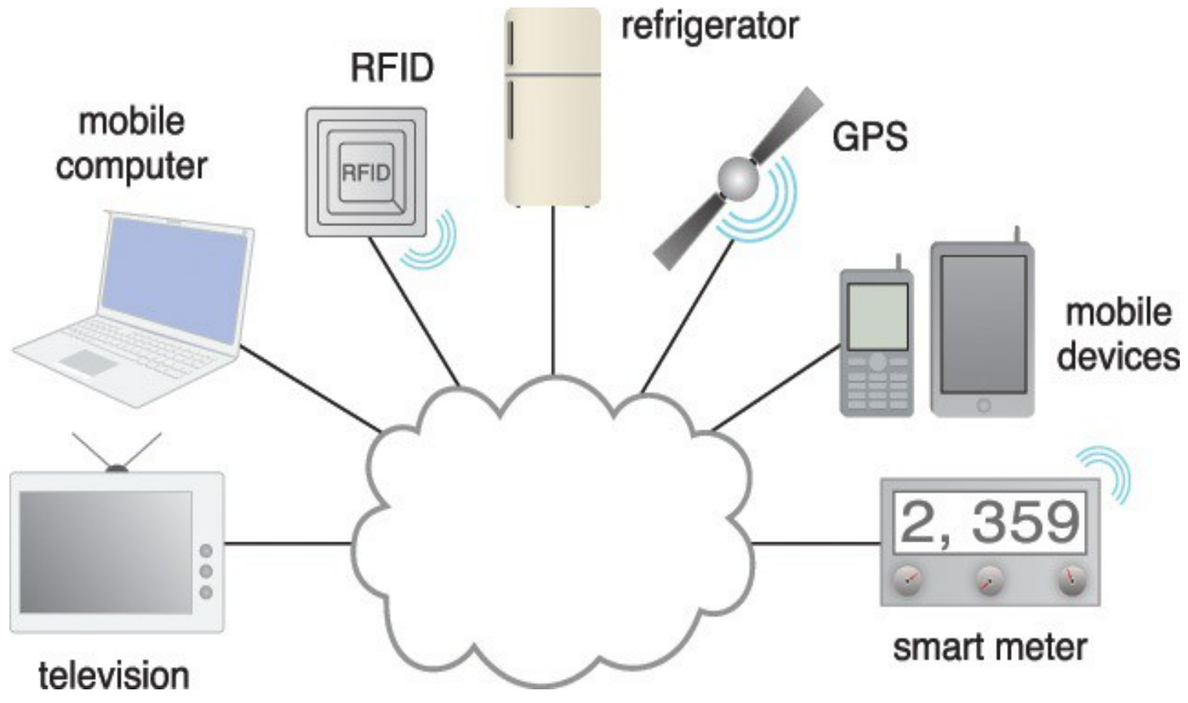



**Figure 2.6** Hyper-connected communities and devices include television, mobile computing, RFIDs, refrigerators, GPS devices, mobile devices and smart meters.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

DATA CLOUD

amazon web services™

Amazon Elastic Elastic MapReduce (EMR)


STORAGE CLOUD

amazon web services™

GW

Amazon Simple Storage Service (S3)

APACHE HBASE

UTILITY CLOUD

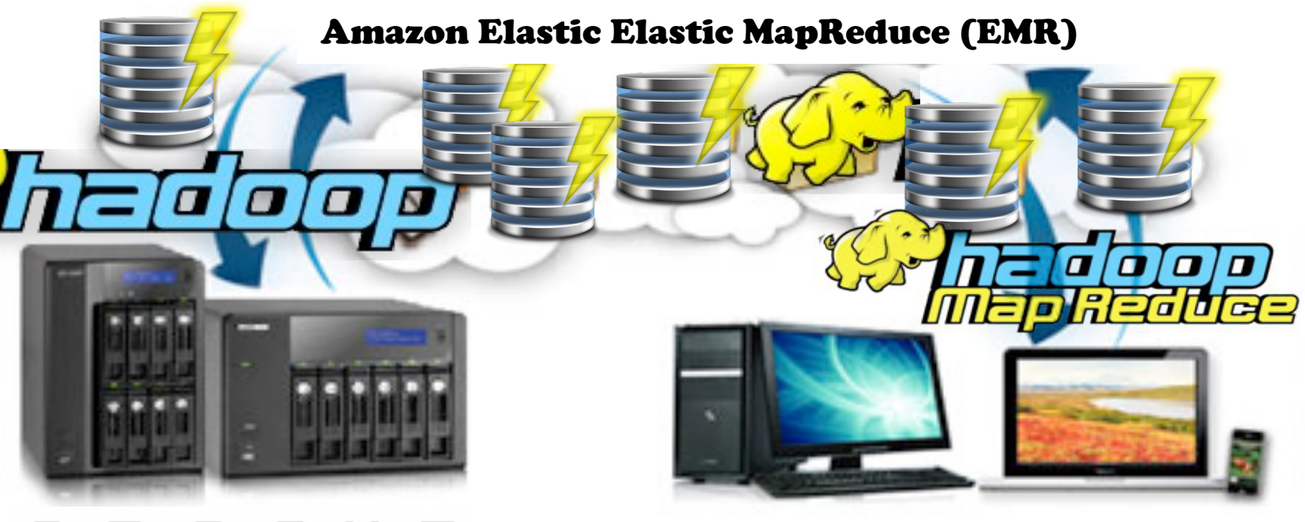
amazon web services™

Amazon Elastic Compute Cloud (EC2)

# DATA CLOUD

amazon web services™

**Amazon Elastic Elastic MapReduce (EMR)**




APACHE HBASE

# STORAGE CLOUD

amazon web services™

GW

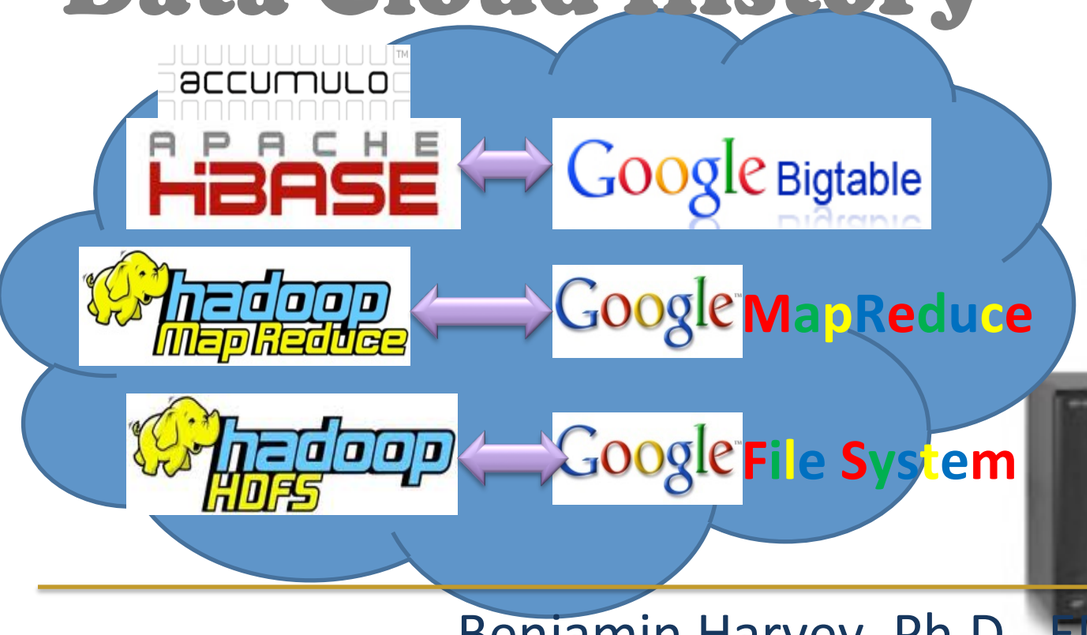
**Amazon Simple Storage Service (S3)**

# Utility CLOUD

amazon web services™

**Amazon Elastic Compute Cloud (EC2)**

# Data Cloud History



# Cloud Infrastructures

## Data Cloud (Comp.)

ACCUMULO

APACHE HBASE ⟷ Google Bigtable

hadoop MapReduce ⟷ Google MapReduce

hadoop HDFS ⟷ Google File System

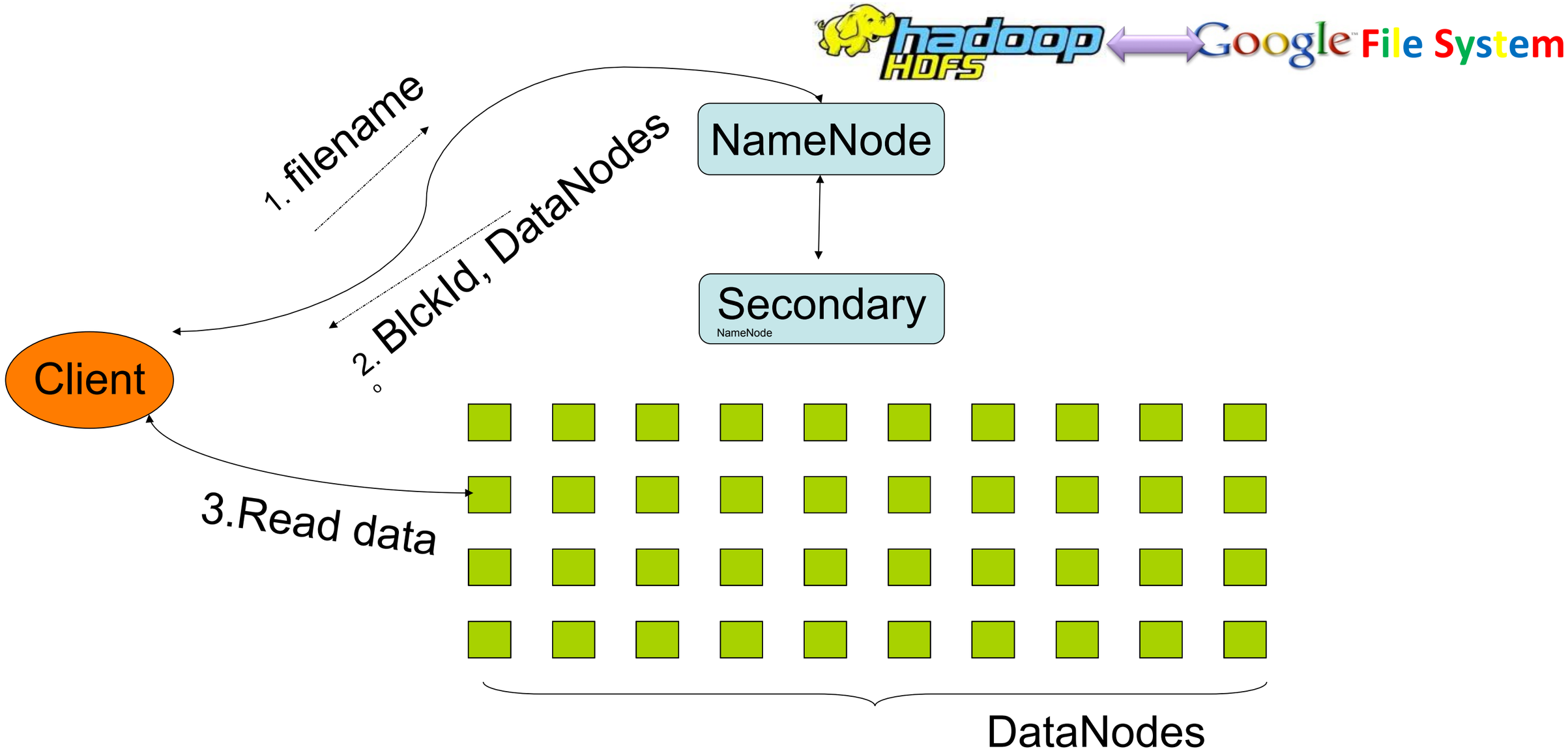
## Storage Cloud

Introducing amazon cloud drive

SkyDrive

## Utility Cloud

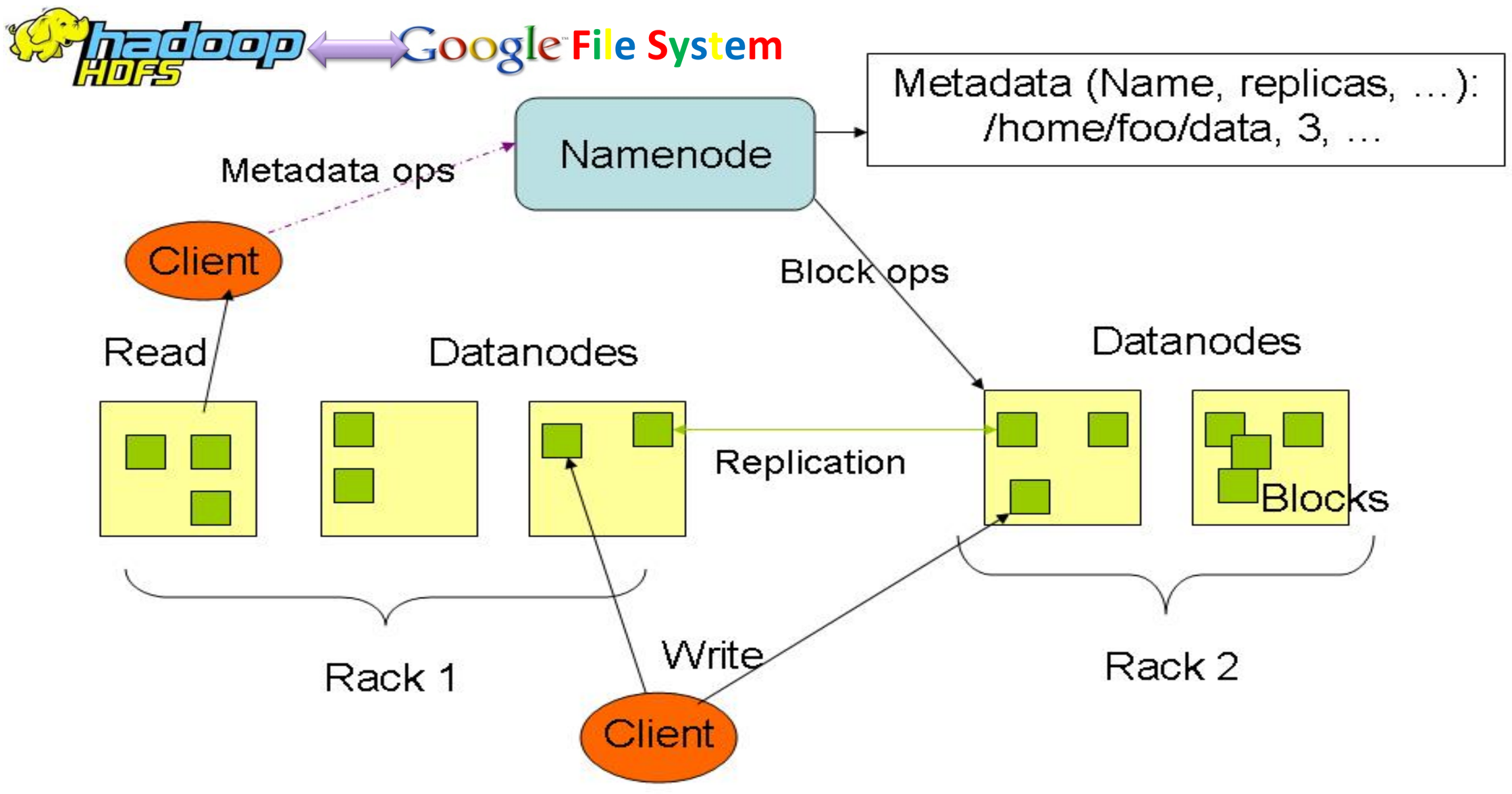

# HDFS: The System Structure




NameNode : Maps a file to a file-id and list of MapNodes
DataNode : Maps a block-id to a physical location on disk

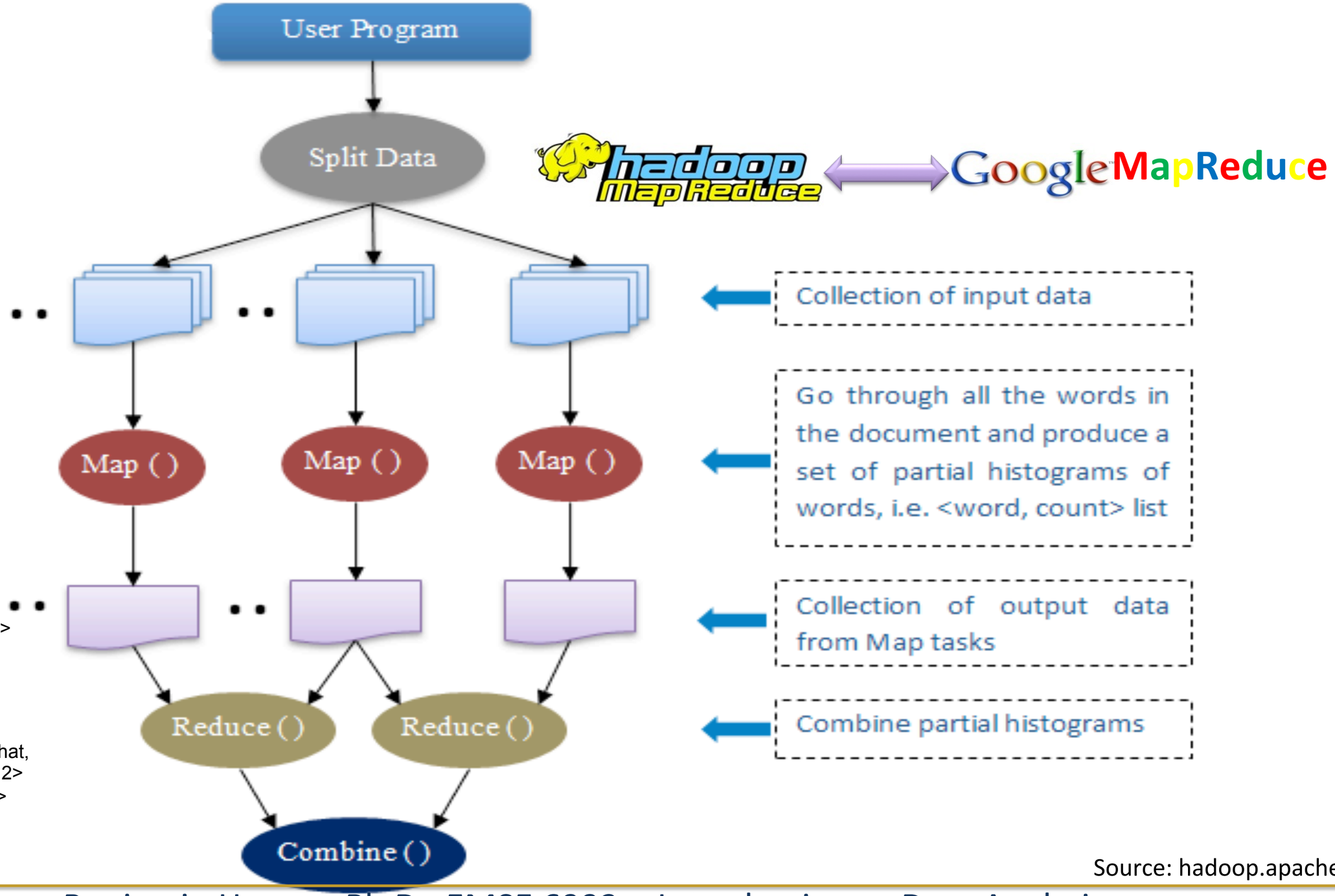
Source: hadoop.apache.org

# HDFS Architecture



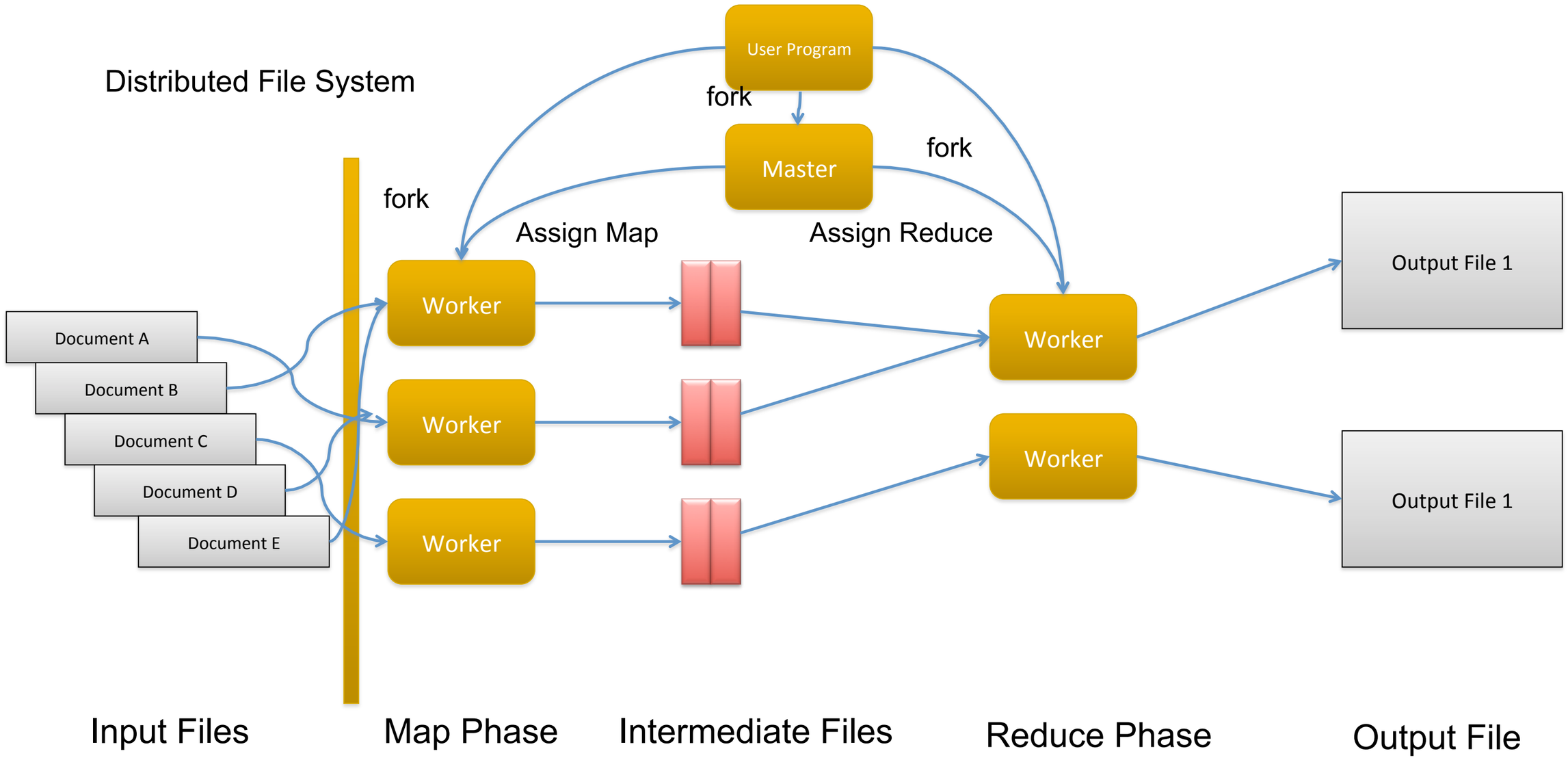

Source: hadoop.apache.org

# MapReduce WordCount




Source: hadoop.apache.org

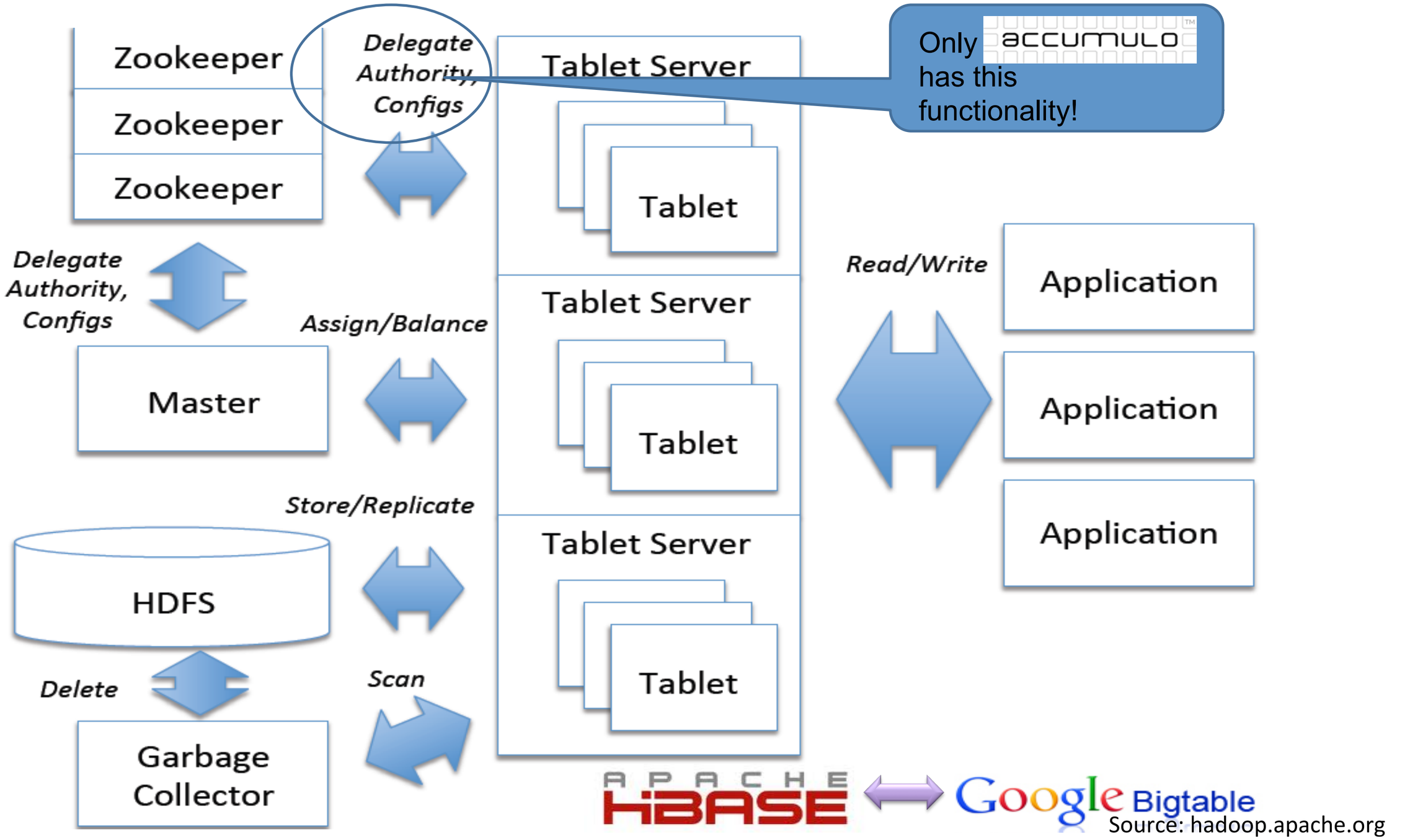
# MapReduce Execution Overview




hadoop MapReduce ⟺ Google MapReduce

Source: hadoop.apache.org

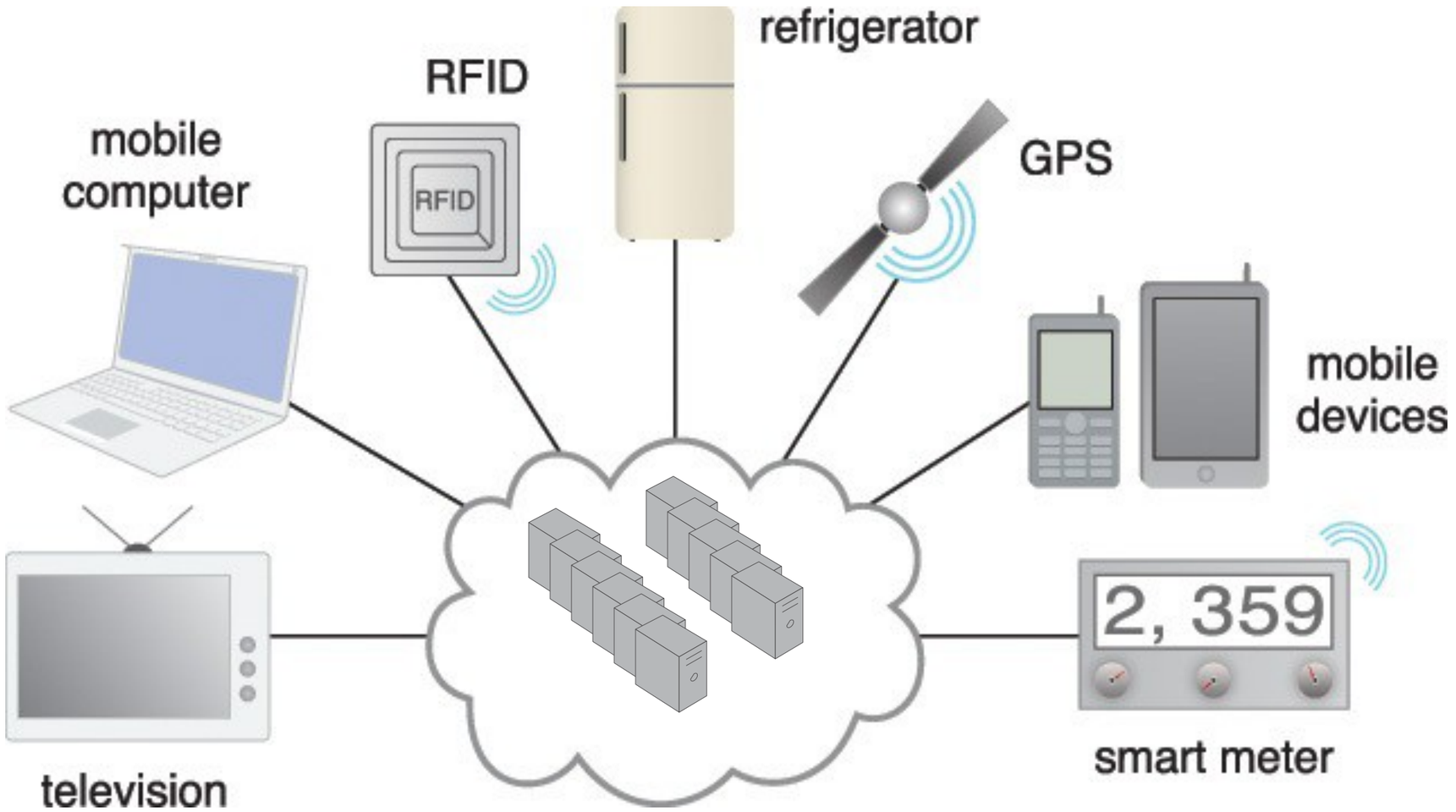
# Hbase/Acculumo Architecture




Source: hadoop.apache.org

# Internet of Everything

- Hyper-connected communities and devices running on affordable technology and commodity hardware stream digitized data that is **subject to analytic processes hosted in elastic cloud computing environments**.



# Internet of Everything

- The results of the analysis can provide insights as to how much value is generated by the current process and whether or not the process should proactively seek opportunities to further optimize itself.
- (1) Business processes in combination with analysis of (2) streaming data and customer context, being able to (3) adapt the execution of these process to align with the customer's goals will become a key corporate differentiator in the future

# Summary: Drivers and Motivation for Big Data Adoption

# Case Study: Ensure to Insure

Introduction to the Case Study

ETI new transformation and innovation corporate priorities:

- Considering transformation, business process management disciplines will be adopted to document, analyze, and improve the processing of claims
- Risk assessment and fraud detection will be enhanced with the application of innovative Big Data technologies that will produce analytic results that can drive data-drive decision-making.
- Redefining CSFs and KPIs has helped ETI link and align the strategic, tactical, and operating levels of the business
- Organization role responsible for innovation management.

Do you think the new transformation and innovations priorities will help? Justify your answer.

What's missing?