# Chapter 7

## Big Data Concepts and Tools

# Learning Objectives (1 of 2)

7.1 Learn what Big Data is and how it is changing the world of analytics

7.2 Understand the motivation for and business drivers of Big Data analytics

7.3 Become familiar with the wide range of enabling technologies for Big Data analytics

7.4 Learn about Hadoop, MapReduce, and NoSQL as they relate to Big Data analytics

7.5 Compare and contrast the complementary uses of data warehousing and Big Data technologies

# Learning Objectives (2 of 2)

**7.6** Become familiar with select Big Data platforms and services

**7.7** Understand the need for and appreciate the capabilities of stream analytics

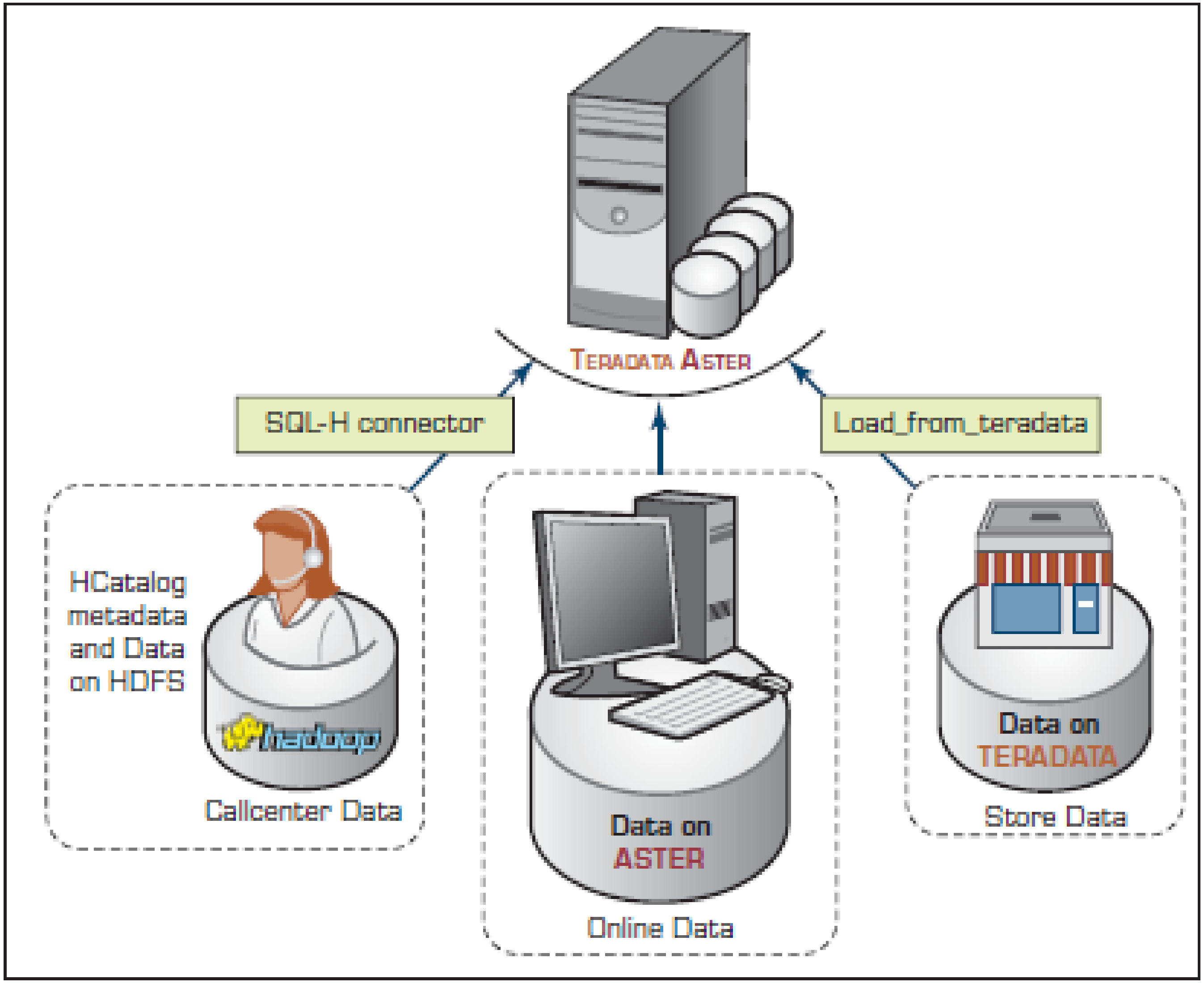
**7.8** Learn about the applications of stream analytics

# Opening Vignette (1 of 4)

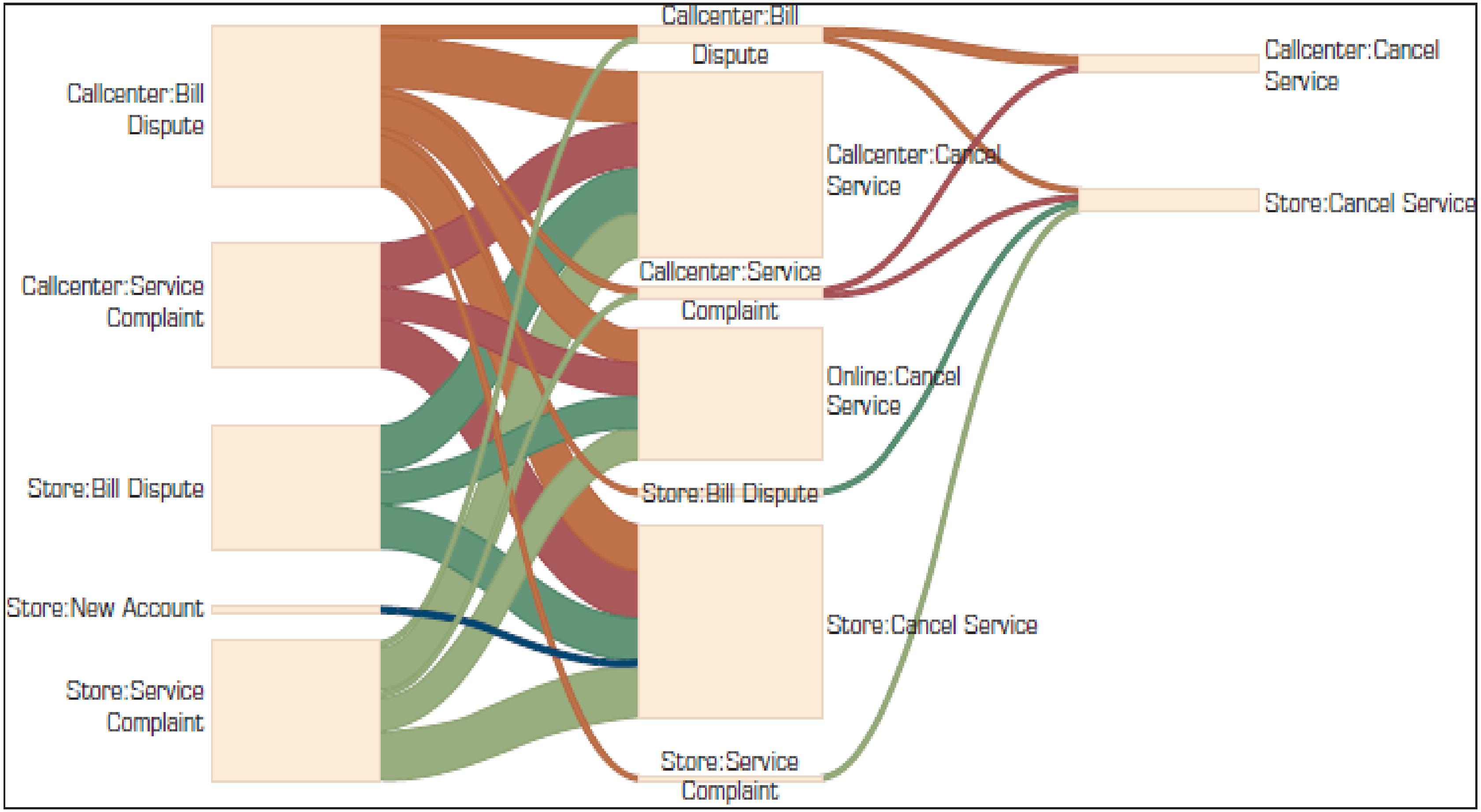
**Analyzing Customer Churn in a Telecom Company Using Big Data Methods**

- Telecom – a highly competitive market segment
- Customer churn rate is higher than most other markets
- A good example of Big Data analytics
- Challenges
  - Data from multiple sources
  - Data volume is higher than usual
- Solution
- Results

# Opening Vignette (2 of 4)



# Opening Vignette (3 of 4)



# Opening Vignette (4 of 4)

## Discussion Questions

1. What problem did customer service cancellation pose to AT’s business survival?
2. Identify and explain the technical hurdles presented by the nature and characteristics of AT’s data.
3. What is sessionizing? Why was it necessary for AT to sessionize its data?
4. Research other studies where customer churn models have been employed. What types of variables were used in those studies? How is this vignette different?

# Big Data - Definition and Concepts (1 of 2)

- Big Data means different things to people with different backgrounds and interests
- Traditionally, “Big Data” = massive volumes of data
  - Example, volume of data at CERN, NASA, Google, …
- Where does the Big Data come from?
  - Everywhere! Web logs, RFID, GPS systems, sensor networks, social networks, Internet-based text documents, Internet search indexes, detail call records, astronomy, atmospheric science, biology, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, multimedia archives, …

# Technology Insights 7.1 (1 of 2)

**The Data Size Is Getting Big, Bigger, and Bigger**

- Hadron Collider - 1 PB/sec
- Boeing jet - 20 TB/hr
- Facebook - 500 TB/day
- YouTube – 1 TB/4 min
- The proposed Square Kilometer Array telescope (the world’s proposed biggest telescope) – 1EB/day

# Technology Insights 7.1 (2 of 2)

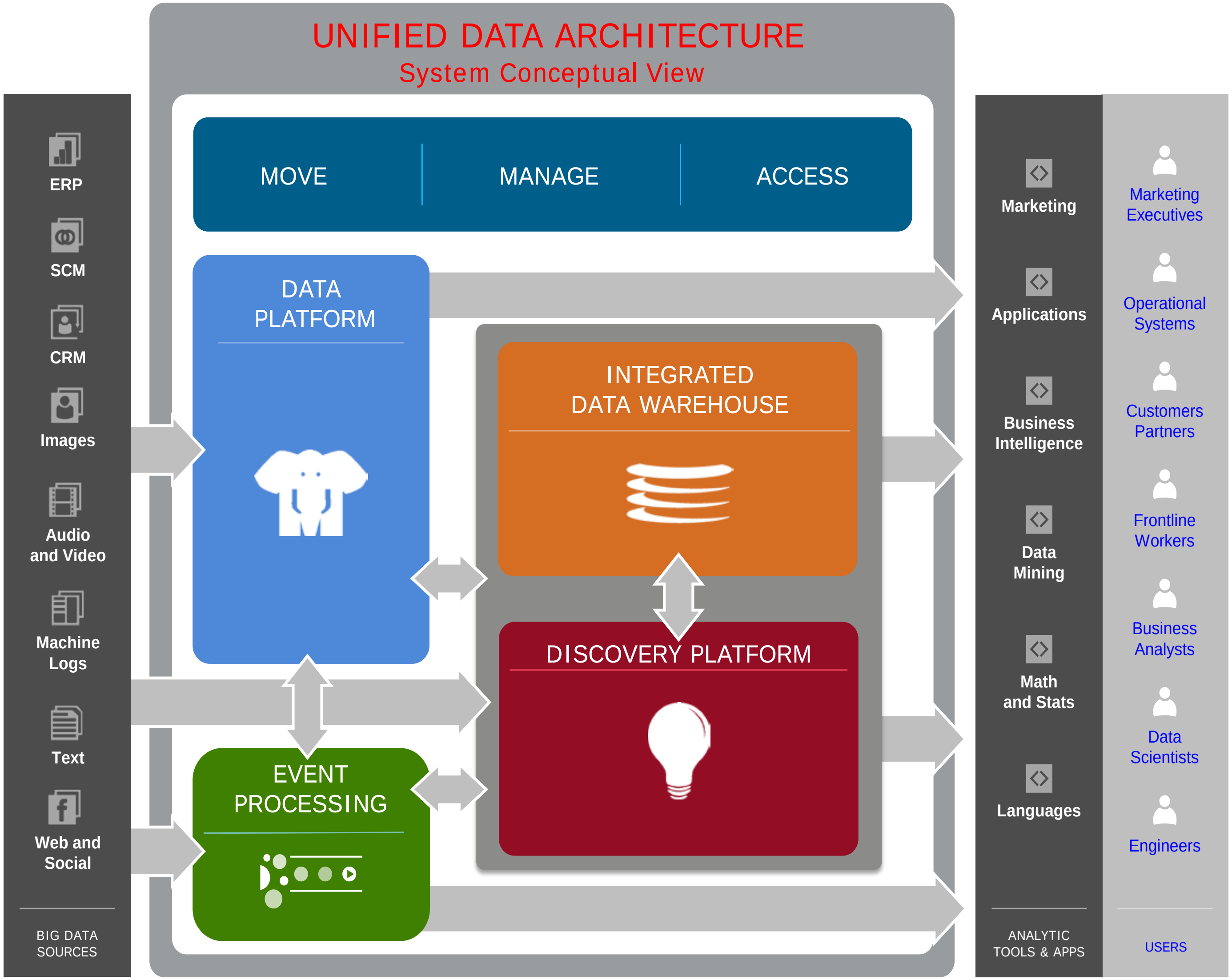
| Name | Symbol | Value |
|---|---|---|
| Kilobyte | kB | $10^3$ |
| Megabyte | MB | $10^6$ |
| Gigabyte | GB | $10^9$ |
| Terabyte | TB | $10^{12}$ |
| Petabyte | PB | $10^{15}$ |
| Exabyte | EB | $10^{18}$ |
| Zettabyte | ZB | $10^{21}$ |
| Yottabyte | YB | $10^{24}$ |
| Brontobyte* | BB | $10^{27}$ |
| Gegobyte* | GeB | $10^{30}$ |

*Not an official SI (International System of Units) name/symbol, yet.

# Big Data - Definition and Concepts (2 of 2)

- Big Data is a misnomer!
- Big Data is more than just “big”
- The Vs that define Big Data
  - **Volume**
  - **Variety**
  - **Velocity**
  - **Veracity**
  - **Variability**
  - **Value**
  - …

# A High-Level Conceptual Architecture for Big Data Solutions (by AsterData / Teradata)



# Application Case 7.1

**Alternative Data for Market Analysis or Forecasts**

**Questions for Discussion**

1. What is a common thread in the examples discussed in this application case?
2. Can you think of other data streams that might help give an early indication of sales at a retailer?
3. Can you think of other applications along the lines presented in this application case?

# Fundamentals of Big Data Analytics

- Big Data by itself, regardless of the size, type, or speed, is worthless
- Big Data + “big” analytics = value
- With the value proposition, Big Data also brought about big challenges
  - Effectively and efficiently capturing, storing, and analyzing Big Data
  - New breed of technologies needed (developed or purchased or hired or outsourced …)

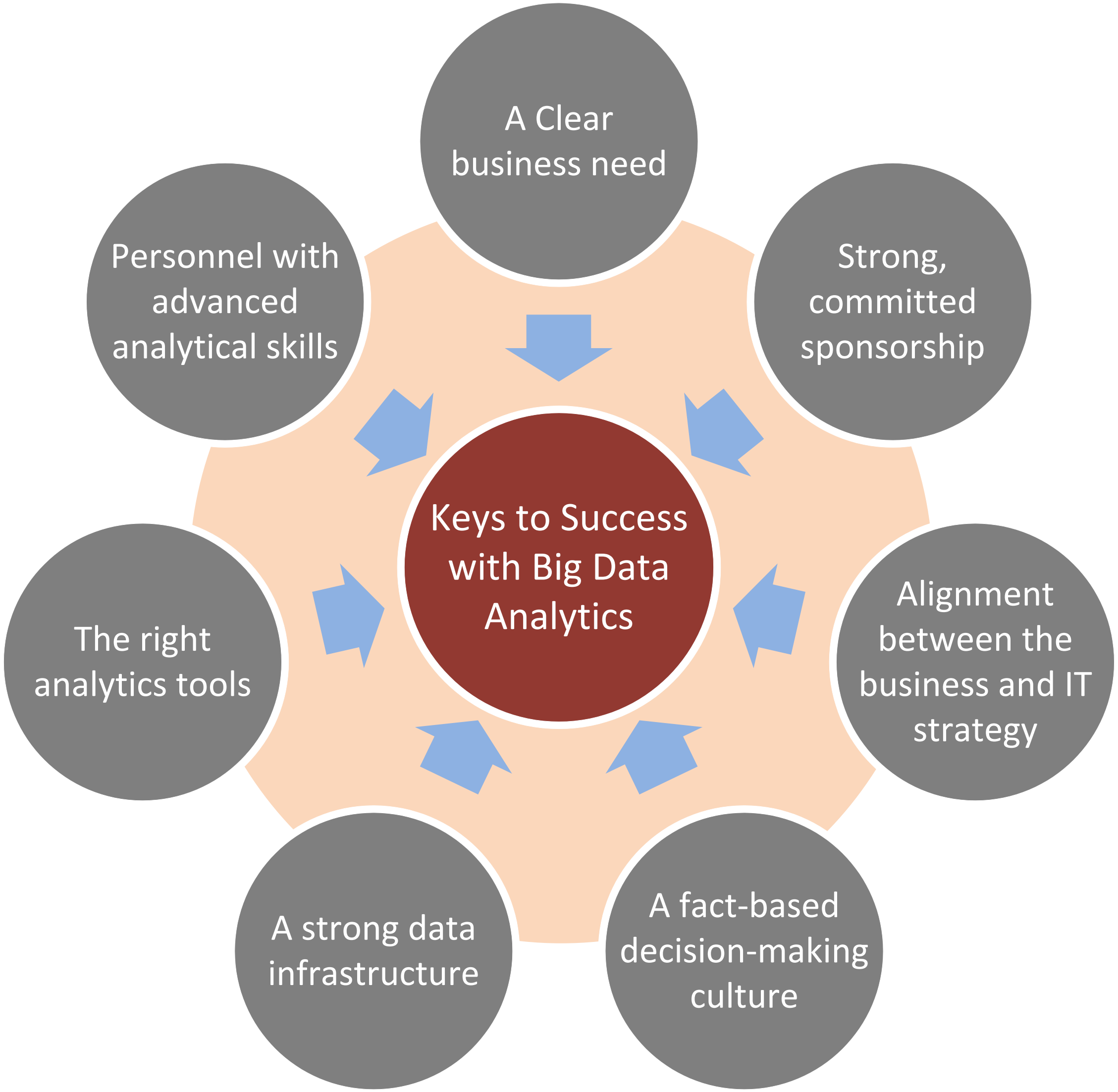
# Big Data Considerations

- You can’t process the amount of data that you want to because of the limitations of your current platform.
- You can’t include new/contemporary data sources (example, social media, RFID, Sensory, Web, GPS, textual data) because it does not comply with the data storage schema
- You need to (or want to) integrate data as quickly as possible to be current on your analysis.
- You want to work with a schema-on-demand data storage paradigm because the variety of data types involved.
- The data is arriving so fast at your organization’s doorstep that your traditional analytics platform cannot handle it.
- …

# Critical Success Factors for Big Data Analytics (1 of 2)

- A clear business need (alignment with the vision and the strategy)
- Strong, committed sponsorship (executive champion)
- Alignment between the business and IT strategy
- A fact-based decision-making culture
- A strong data infrastructure
- The right analytics tools
- Right people with right skills

# Critical Success Factors for Big Data Analytics (2 of 2)



Keys to Success with Big Data Analytics

A Clear business need

Strong, committed sponsorship

Alignment between the business and IT strategy

A fact-based decision-making culture

A strong data infrastructure

The right analytics tools

Personnel with advanced analytical skills

# Enablers of Big Data Analytics

- In-memory analytics
  - Storing and processing the complete data set in RAM
- In-database analytics
  - Placing analytic procedures close to where data is stored
- Grid computing & MPP
  - Use of many machines and processors in parallel (MPP - massively parallel processing)
- Appliances
  - Combining hardware, software, and storage in a single unit for performance and scalability

# Challenges of Big Data Analytics

- Data volume
  - The ability to capture, store, and process the huge volume of data in a timely manner
- Data integration
  - The ability to combine data quickly and at reasonable cost
- Processing capabilities
  - The ability to process the data quickly, as it is captured (i.e., stream analytics)
- Data governance (… security, privacy, access)
- Skill availability (… data scientist)
- Solution cost (ROI)

# Business Problems Addressed by Big Data Analytics (1 of 2)

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling/up-selling
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities

# Business Problems Addressed by Big Data Analytics (2 of 2)

- Risk management
- Regulatory compliance
- Enhanced security capabilities
- …

# Application Case 7.2 (1 of 2)

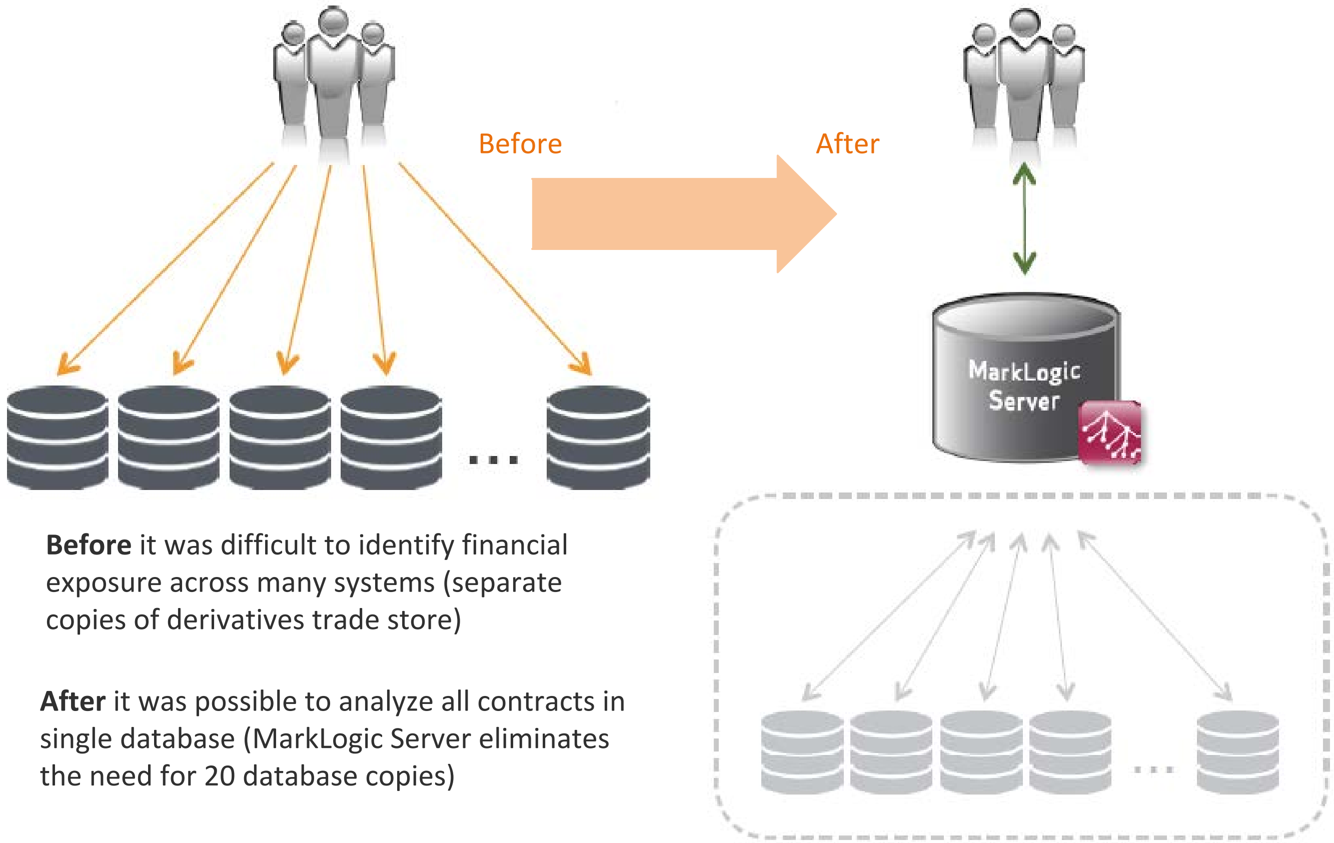
**Top Five Investment Bank Achieves Single Source of the Truth**

**Questions for Discussion**

1. How can Big Data benefit large-scale trading banks?
2. How did MarkLogic infrastructure help ease the leveraging of Big Data?
3. What were the challenges, the proposed solution, and the obtained results?

# Application Case 7.2 (2 of 2)

- Moving from many old systems to a unified new system

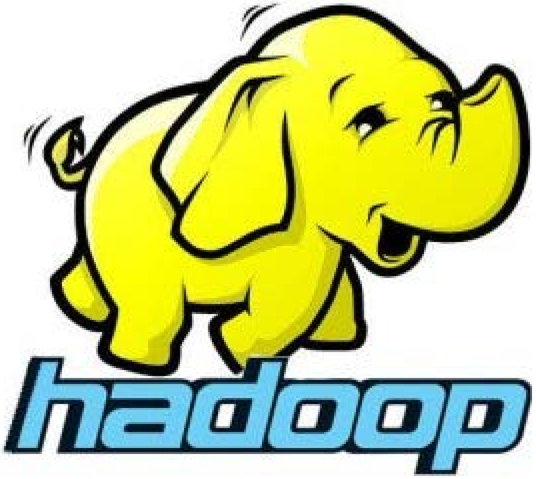


**Before** it was difficult to identify financial exposure across many systems (separate copies of derivatives trade store)

**After** it was possible to analyze all contracts in single database (MarkLogic Server eliminates the need for 20 database copies)

# Big Data Technologies (1 of 2)

- MapReduce …
- Hadoop …
- Hive
- Pig
- Hbase
- Flume
- Oozie
- Ambari



# Big Data Technologies (2 of 2)

- Avro
- Mahout
- Sqoop, Hcatalog, ….

# VDO

- MapReduce Explained
  - https://www.youtube.com/watch?v=lgWy7BwIKKQ

- What is Hadoop?
  - https://www.youtube.com/watch?v=9s-vSeWej1U

- Hadoop Tutorial for Beginners.
  - https://www.youtube.com/watch?v=oT7kczq5A-0

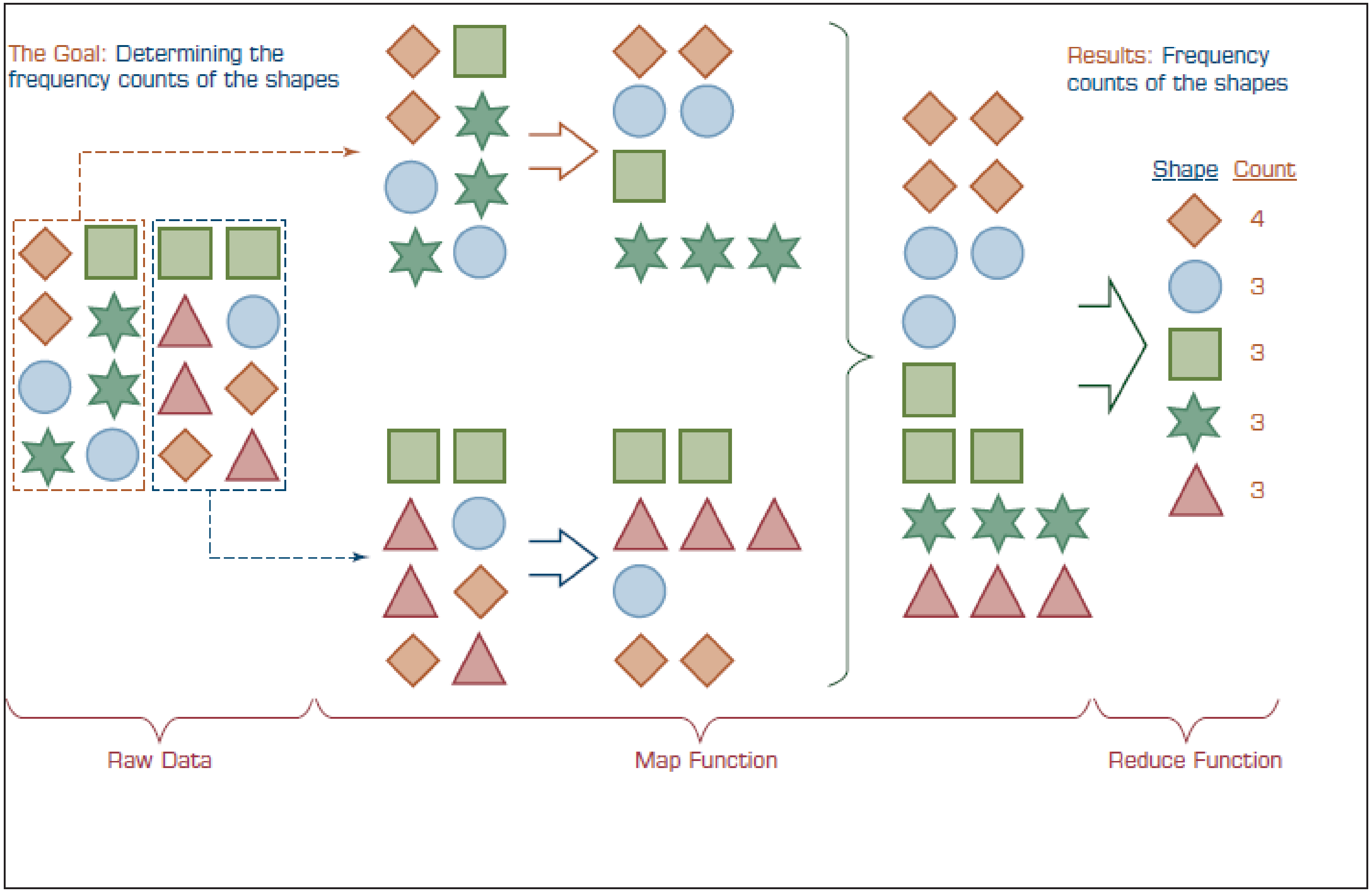
# What is MapReduce?

# Big Data Technologies--MapReduce (1 of 2)

- MapReduce distributes the processing of very large multi-structured data files across a large cluster of ordinary machines/processors
- Goal - achieving high performance with “simple” computers
- Developed and popularized by Google
- Good at processing and analyzing large volumes of multi-structured data in a timely manner
- Example tasks: indexing the Web for search, graph analysis, text analysis, machine learning, …

# Big Data Technologies--MapReduce (2 of 2)

- How does MapReduce work?


The Goal: Determining the frequency counts of the shapes
Results: Frequency counts of the shapes
Shape Count
4
3
3
3
3
Raw Data
Map Function
Reduce Function

# Big Data Technologies--Hadoop (1 of 3)

- Hadoop is an open source framework for storing and analyzing massive amounts of distributed, unstructured data
  - Originally created by Doug Cutting at Yahoo!
- Hadoop clusters run on inexpensive commodity hardware so projects can scale-out inexpensively
  - Hadoop is now part of Apache Software Foundation
  - Open source - hundreds of contributors continuously improve the core technology

# What is Hadoop?

# Big Data Technologies--Hadoop (2 of 3)

- **How Does Hadoop Work?**
  - Access unstructured and semi-structured data (example, log files, social media feeds, other data sources)
  - Break the data up into “parts,” which are then loaded into a file system made up of multiple nodes running on commodity hardware using HDFS
  - Each “part” is replicated multiple times and loaded into the file system for replication and failsafe processing
  - A node acts as the **Facilitator** and another as **Job Tracker**
  - Jobs are distributed to the clients, and once completed the results are collected and aggregated using *MapReduce*

# Big Data Technologies--Hadoop (3 of 3)

- **Hadoop Technical Components**
  - Hadoop Distributed File System (HDFS)
  - Name Node (primary facilitator)
  - Secondary Node (backup to Name Node)
  - Job Tracker
  - Slave Nodes (the grunts of any Hadoop cluster)
  - Additionally, Hadoop ecosystem is made up of a number of complementary sub-projects: NoSQL (Cassandra, Hbase), DW (Hive), …
    - NoSQL = not only SQL

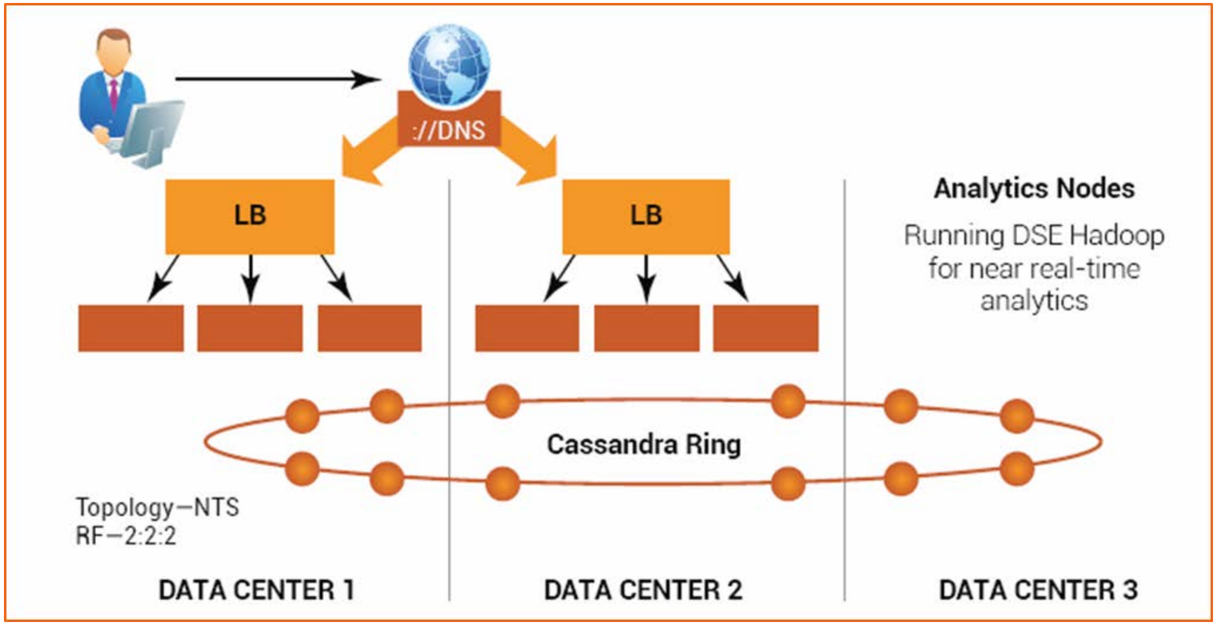
# Technology Insights 7.2

## A Few Demystifying Facts about Hadoop

- Hadoop consists of multiple products
- Hadoop is open source but available from vendors, too
- Hadoop is an ecosystem, not a single product
- HDFS is a file system, not a DBMS
- Hive resembles SQL but is not standard SQL
- Hadoop and MapReduce are related but not the same
- MapReduce provides control for analytics, not analytics
- Hadoop is about data diversity, not just data volume

**Page 31**

# Application Case 7.3 - eBay’s Big Data Solution



## Questions for Discussion

1. Why did eBay need a Big Data solution?
2. What were the challenges, the proposed solution, and the obtained results?

# Application Case 7.4

**Understanding Quality and Reliability of Healthcare Support Information on Twitter**

**Questions for Discussion**

1. What was the data scientists' main concern regarding health information that is disseminated on the Twitter platform?
2. How did the data scientists ensure that nonexpert information disseminated on social media could indeed contain valuable health information?
3. Does it make sense that influential users would share more objective information whereas less influential users could focus more on subjective information? Why?

***Page 33***

# Big Data and Data Warehousing

- What is the impact of Big Data on DW?
  - Big Data and RDBMS do not go nicely together
  - Will Hadoop replace data warehousing/RDBMS?
- Use Cases for Hadoop
  - Hadoop as the repository and refinery
  - Hadoop as the active archive
- Use Cases for Data Warehousing
  - Data warehouse performance
  - Integrating data that provides business value
  - Interactive BI tools

# Hadoop Versus Data Warehouse When to Use Which Platform (1 of 2)

## Table 7.1 When to Use Which Platform—Hadoop versus DW

| Requirement | Data Warehouse | Hadoop |
|---|---|---|
| Low latency, interactive reports, and OLAP | ☑ | |
| ANSI 2003 SQL compliance is required | ☑ | ☑ |
| Preprocessing or exploration of raw unstructured data | | ☑ |
| Online archives alternative to tape | | ☑ |
| High-quality cleansed and consistent data | ☑ | ☑ |
| 100s to 1,000s of concurrent users | ☑ | ☑ |
| Discover unknown relationships in the data | | ☑ |

# Hadoop Versus Data Warehouse When to Use Which Platform (2 of 2)

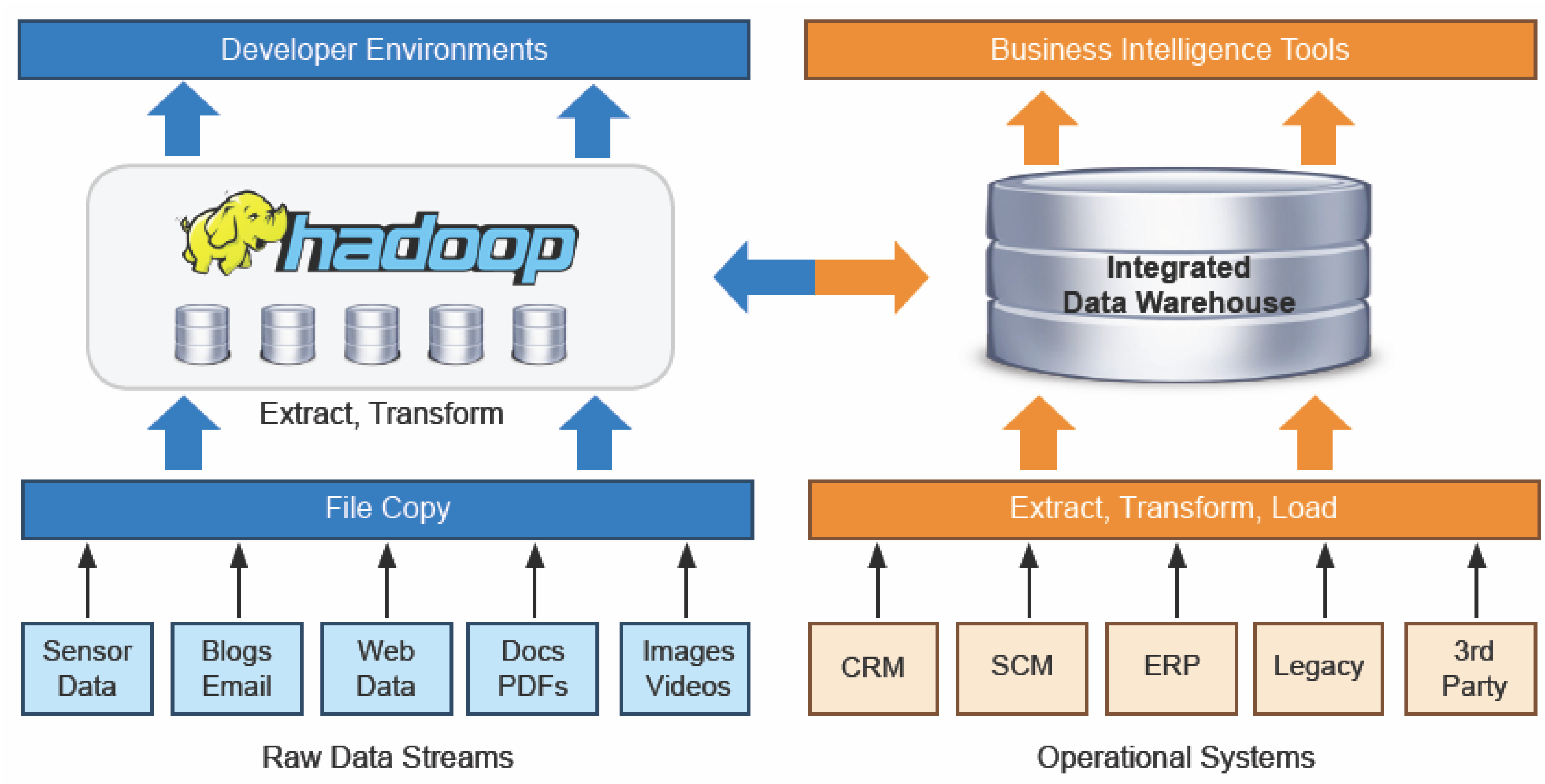
## Table 7.1 [continued]

| Requirement | Data Warehouse | Hadoop |
|---|---|---|
| Parallel complex process logic | ☑ | ☑ |
| CPU intense analysis | ☑ | |
| System, users, and data governance | | ☑ |
| Many flexible programming languages running in parallel | | ☑ |
| Unrestricted, ungoverned sandbox explorations | | ☑ |
| Analysis of provisional data | ☑ | |
| Extensive security and regulatory compliance | ☑ | ☑ |

# Coexistence of Hadoop and DW (1 of 2)

1. Use Hadoop for storing and archiving multi-structured data
2. Use Hadoop for filtering, transforming, and/or consolidating multi-structured data
3. Use Hadoop to analyze large volumes of multi-structured data and publish the analytical results
4. Use a relational DBMS that provides MapReduce capabilities as an investigative computing platform
5. Use a front-end query tool to access and analyze data

# Coexistence of Hadoop and DW (2 of 2)

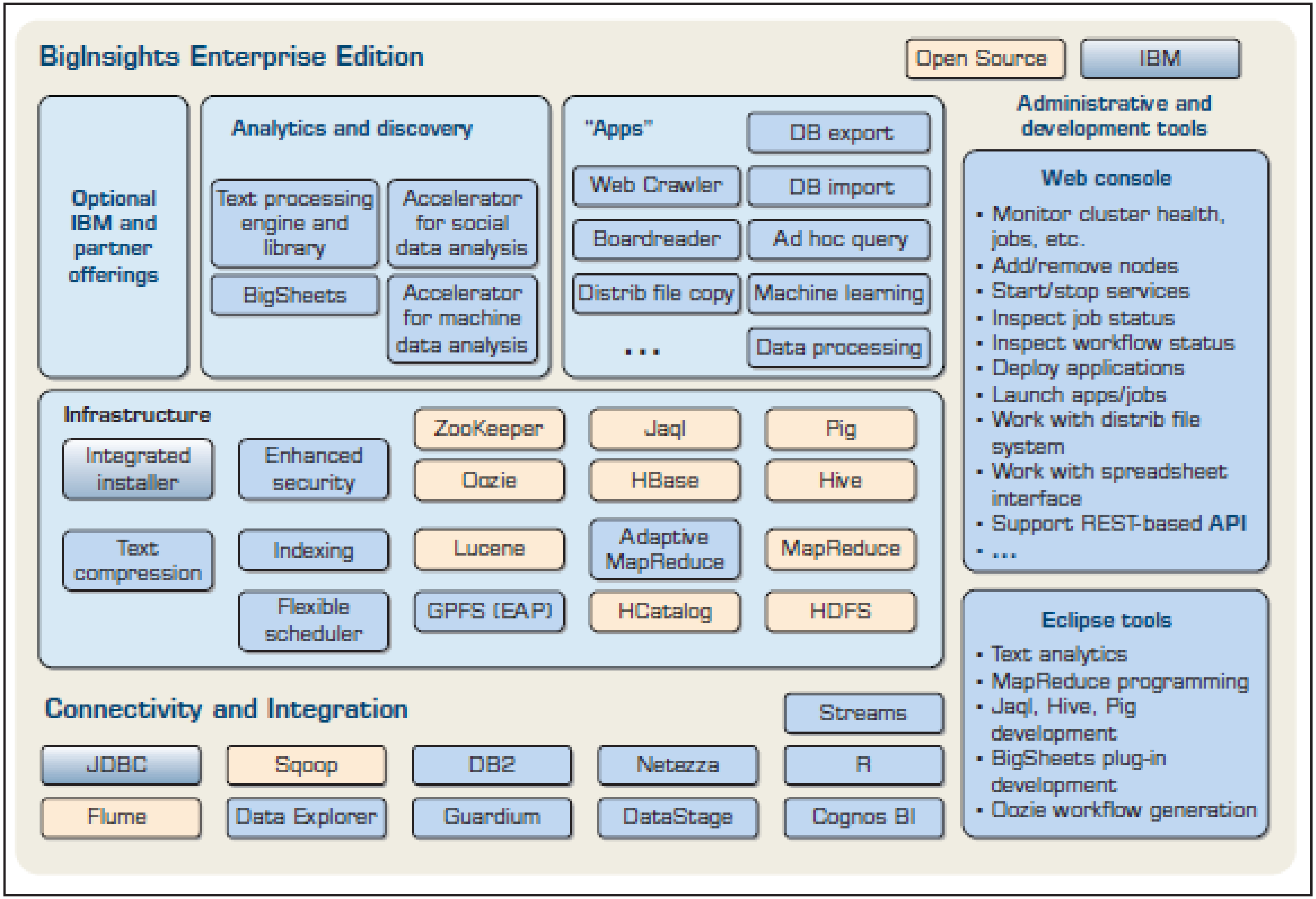


# Big Data Vendors

**Software, Hardware, Service, …**

- Big Data vendor landscape is developing very rapidly
- A representative list would include
  - Cloudera - cloudera.com
  - MapR – mapr.com
  - Hortonworks - hortonworks.com
  - Also, IBM (Netezza, InfoSphere), Oracle (Exadata, Exalogic), Microsoft, Amazon, Google, …

# IBM InfoSphere BigInsights



BigInsights Enterprise Edition

Open Source

IBM

Optional IBM and partner offerings

Analytics and discovery

- Text processing engine and library
- Accelerator for social data analysis
- BigSheets
- Accelerator for machine data analysis

"Apps"

- DB export
- Web Crawler
- DB import
- Boardreader
- Ad hoc query
- Distrib file copy
- Machine learning
- . . .
- Data processing

Administrative and development tools

**Web console**

- Monitor cluster health, jobs, etc.
- Add/remove nodes
- Start/stop services
- Inspect job status
- Inspect workflow status
- Deploy applications
- Launch apps/jobs
- Work with distrib file system
- Work with spreadsheet interface
- Support REST-based API
- . . .

**Eclipse tools**

- Text analytics
- MapReduce programming
- Jaql, Hive, Pig development
- BigSheets plug-in development
- Oozie workflow generation

Infrastructure

- Integrated installer
- Enhanced security
- ZooKeeper
- Jaql
- Pig
- Oozie
- HBase
- Hive
- Text compression
- Indexing
- Lucene
- Adaptive MapReduce
- MapReduce
- Flexible scheduler
- GPFS (EAP)
- HCatalog
- HDFS

Connectivity and Integration

- Streams
- JDBC
- Sqoop
- DB2
- Netezza
- R
- Flume
- Data Explorer
- Guardium
- DataStage
- Cognos BI

# Application Case 7.5

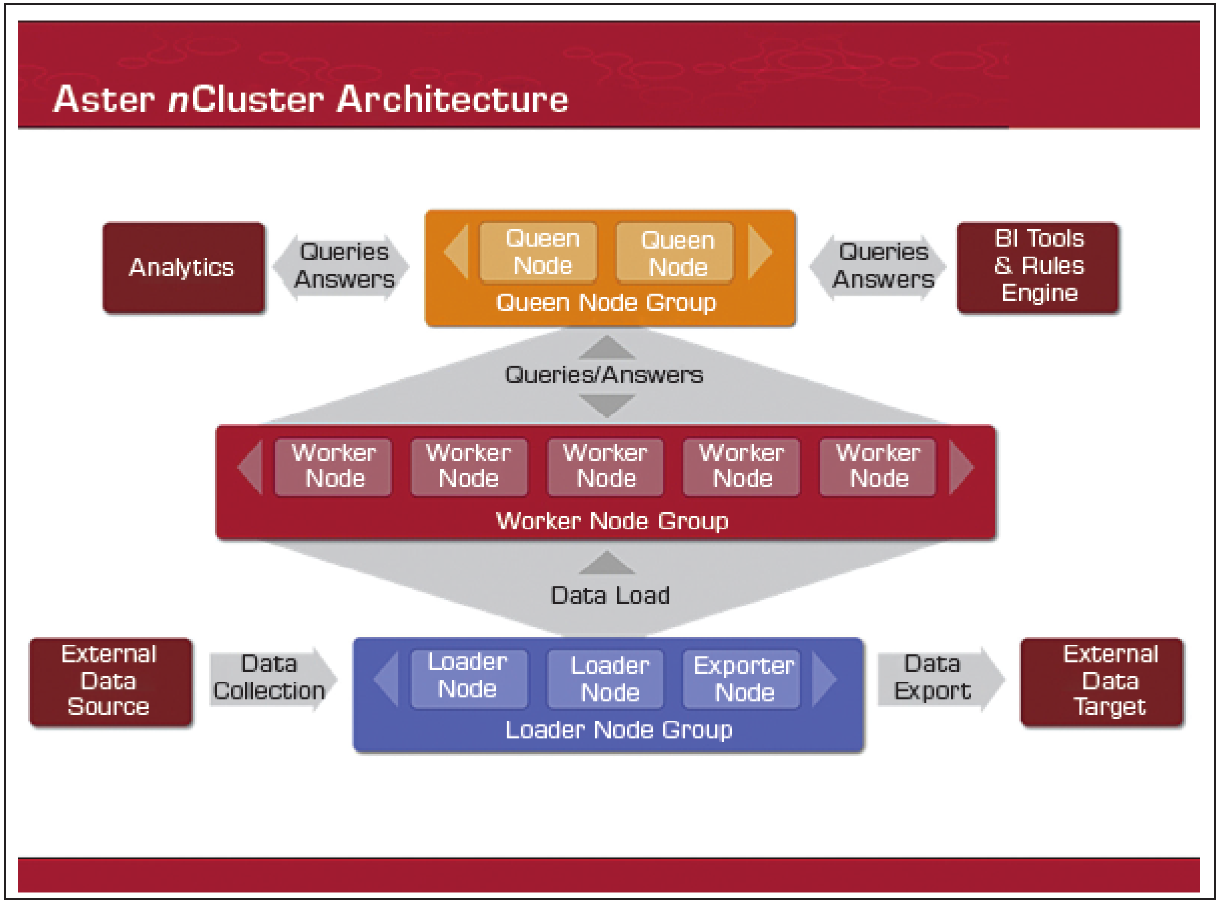
**Using Social Media for Nowcasting the Flu Activity**

**Questions for Discussion**

1. Why would social media be able to serve as an early predictor of flu outbreaks?
2. What other variables might help in predicting such outbreaks?
3. Why would this problem be a good problem to solve using Big Data technologies mentioned in this chapter?

# Big Data Platforms

## Teradata Aster



# Application Case 7.6

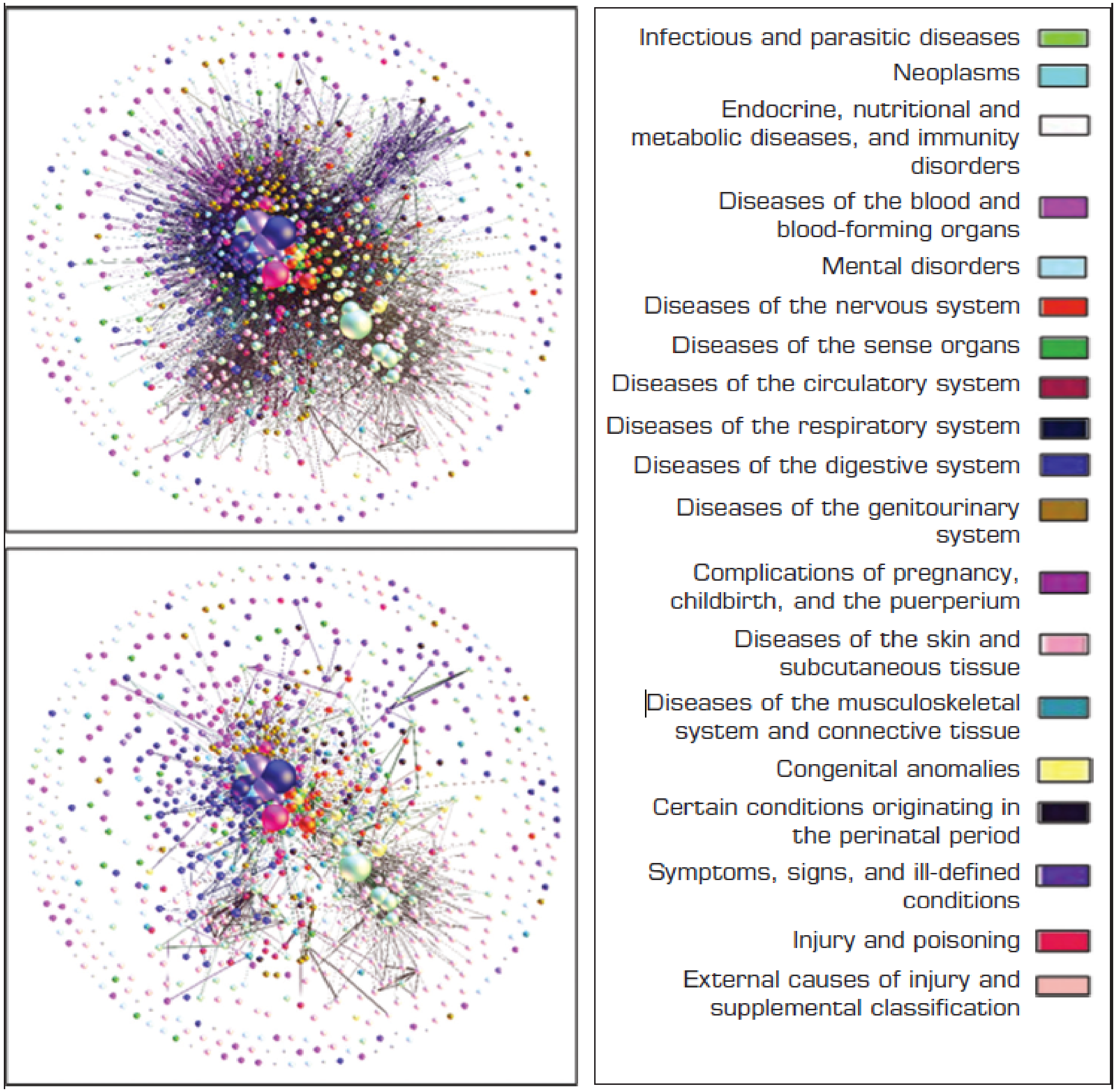
**Analyzing Disease Patterns from an Electronic Medical Records Data Warehouse**

**Questions for Discussion**

1. Why could comorbidity of diseases be different between rural and urban hospitals?
2. What is the issue about the huge difference between rural and urban patient encounters?
3. What are the main components of a network?
4. Where else can you apply the network approach?

***Page 43***

# Figure 7.11 Urban and Rural Comorbidity Networks


Infectious and parasitic diseases
Neoplasms
Endocrine, nutritional and metabolic diseases, and immunity disorders
Diseases of the blood and blood-forming organs
Mental disorders
Diseases of the nervous system
Diseases of the sense organs
Diseases of the circulatory system
Diseases of the respiratory system
Diseases of the digestive system
Diseases of the genitourinary system
Complications of pregnancy, childbirth, and the puerperium
Diseases of the skin and subcutaneous tissue
Diseases of the musculoskeletal system and connective tissue
Congenital anomalies
Certain conditions originating in the perinatal period
Symptoms, signs, and ill-defined conditions
Injury and poisoning
External causes of injury and supplemental classification

# Technology Insights 7.3

## How to Succeed with Big Data

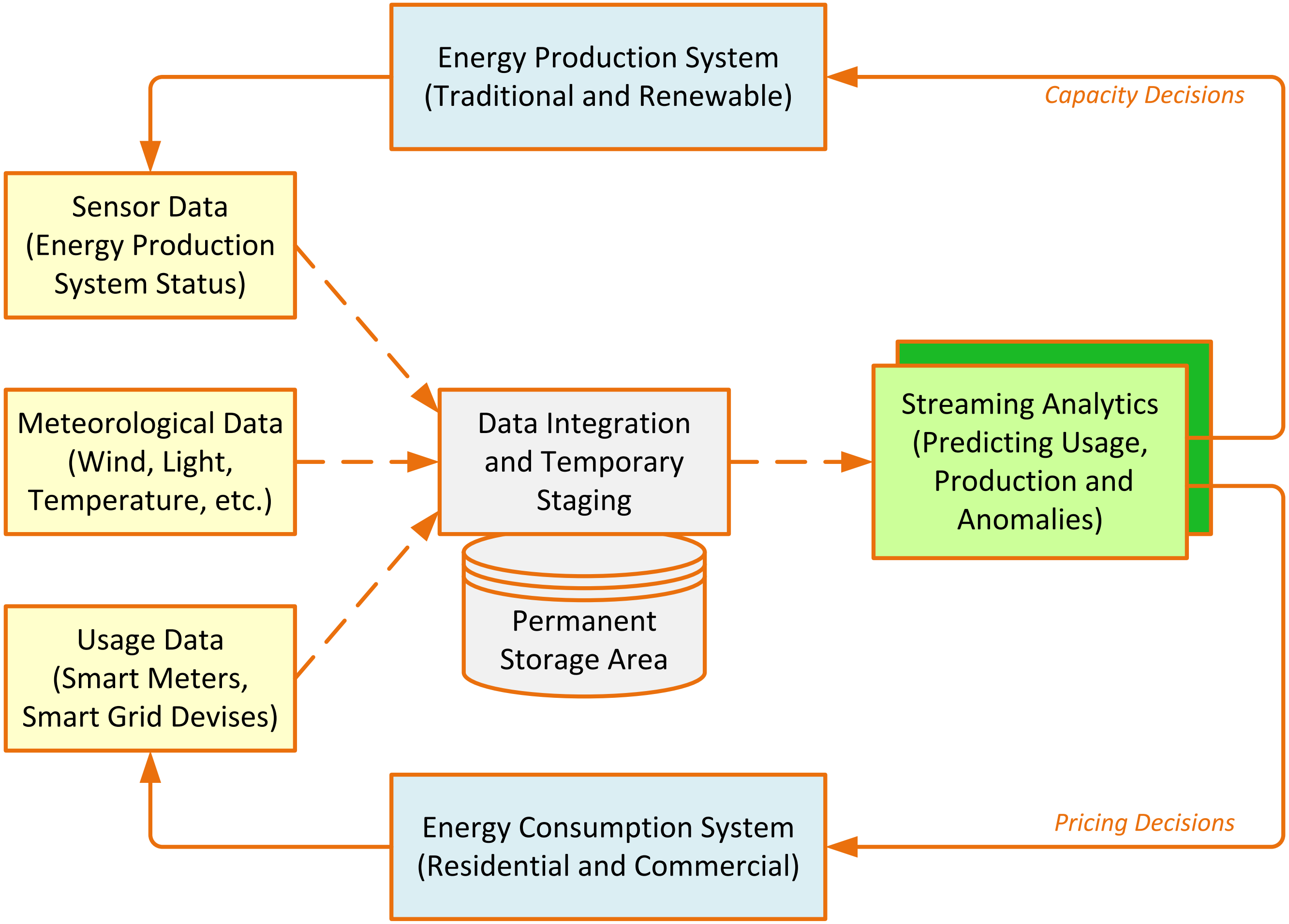
1. Simplify
2. Coexist
3. Visualize
4. Empower
5. Integrate
6. Govern
7. Evangelize

# Big Data and Stream Analytics

- Data-in-motion analytics and real-time data analytics
  - One of the Vs in Big Data = Velocity
- Analytic process of extracting actionable information from continuously flowing data
- Why Stream Analytics?
  - It may not be feasible to store the data, or lose its value
- Stream Analytics Versus Perpetual Analytics
- Critical Event Processing?

# Stream Analytics

## A Use Case in Energy Industry


Energy Production System (Traditional and Renewable)
Capacity Decisions
Sensor Data (Energy Production System Status)
Meteorological Data (Wind, Light, Temperature, etc.)
Usage Data (Smart Meters, Smart Grid Devises)
Data Integration and Temporary Staging
Permanent Storage Area
Streaming Analytics (Predicting Usage, Production and Anomalies)
Pricing Decisions
Energy Consumption System (Residential and Commercial)

# Stream Analytics Applications

- e-Commerce
- Telecommunication
- Law Enforcement and Cyber Security
- Power Industry
- Financial Services
- Health Services
- Government

# Application Case 7.7

**Salesforce Is Using Streaming Data to Enhance Customer Value**

**Questions for Discussion**

1. Are there areas in any industry where streaming data is irrelevant?
2. Besides customer retention, what are other benefits of using predictive analytics?