Module 4

Relationship between Big Data and Cognitive Computing

Introduction

- A cognitive computing environment requires sufficient amount of data to discover patterns or anomalies within that data.
- a cognitive system it is important to have enough data that the results of analytics are trustworthy and consistent.
- A cognitive system requires the ingestion and mapping of data so that the system can begin to discover where there are connections between data sources to begin discovering insights.
- To accomplish the goal of finding insights in data, a cognitive system includes both structured and unstructured data. Structured data, such as data in a relational database, is created for processing by a computer.
- Unstructured data in the form of written material, video, and images, is designed for human consumption and interpretation.

Dealing with Human-Generated Data

1. Intentional data
2. photos,
3. videos,
4. record audio
5. put text on a social network,
6. click “like” if you’re on Facebook.
7. When you do web searches,
8. a record of the web pages that you have viewed are bookmarked.
9. Your emails and your text messages.
10. Your cell phone calls.
11. If you read an eBook, the highlights, the notes in the bookmarks and online purchases.

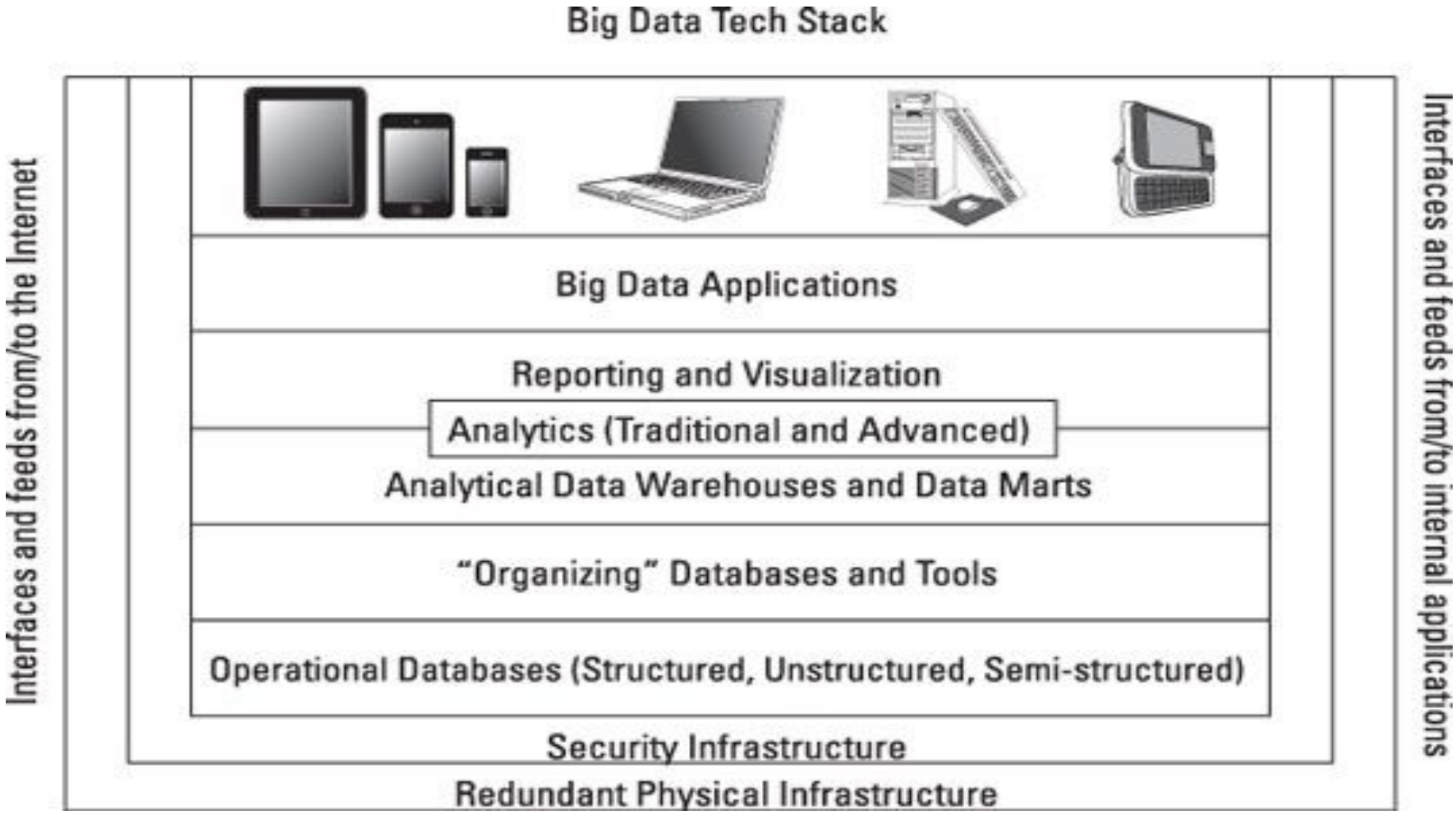
All of these are kinds of data that do not exist until the person deliberately makes them happen, so these are records of human actions.

This type of data requires a level of processing beyond the capabilities of relational database systems because the goal is to interpret the meaning and create human readable data.
Hence the concept of Metadata came in to picture that is data about data

Defining Big Data

5 V's of Big Data

- Volume : Scale of data.
- Velocity : Speed of data.
- Variety : Diversity of data.
- Veracity : accuracy gained from data.
- Value : Insights gained from data.


Big Data Tech Stack
Interfaces and feeds from/to the Internet
Big Data Applications
Reporting and Visualization
Analytics (Traditional and Advanced)
Analytical Data Warehouses and Data Marts
"Organizing" Databases and Tools
Operational Databases (Structured, Unstructured, Semi-structured)
Security Infrastructure
Redundant Physical Infrastructure
Interfaces and feeds from/to internal applications

Interfaces and feeds for big data

To understand how big data works in the real world, it is important to start by understanding the necessity of interfaces and feeds. In fact, what makes big data big is the fact that it relies on picking up lots of data from lots of sources.

Therefore, open application programming interfaces (APIs) will be core to any big data architecture. In addition, keep in mind that interfaces exist at every level and between every layer of the stack. Without integration services, big data can’t happen.

Redundant big data physical infrastructure

The supporting physical infrastructure is fundamental to the operation and scalability of a big data architecture. In fact, without the availability of robust physical infrastructures, big data would probably not have emerged as such an important trend. To support an unanticipated or unpredictable volume of data, a physical infrastructure for big data has to be different than that for traditional data.

The physical infrastructure is based on a distributed computing model. This means that data may be physically stored in many different locations and can be linked together through networks, the use of a distributed file system, and various big data analytic tools and applications.

Redundancy is important because you are dealing with so much data from so many different sources. Redundancy comes in many forms. If your company has created a private cloud, you will want to have redundancy built within the private environment so that it can scale out to support changing workloads.

If your company wants to contain internal IT growth, it may use external cloud services to augment its internal resources. In some cases, this redundancy may come in the form of a Software as a Service (SaaS) offering that allows companies to do sophisticated data analysis as a service. The SaaS approach offers lower costs, quicker startup, and seamless evolution of the underlying technology.

Big Data security infrastructure

The more important big data analysis becomes to companies, the more important it will be to secure that data. For example, if you are a healthcare company, you will probably want to use big data applications to determine changes in demographics or shifts in patient needs. This data about your constituents needs to be protected both to meet compliance requirements and to protect the patients' privacy.

You will need to take into account who is allowed to see the data and under what circumstances they are allowed to do so. You will need to be able to verify the identity of users as well as protect the identity of patients.

Operational big data sources

- Incorporate all the data sources that will give a complete picture of the business and see how the data impacts the way it operate the business.
- As the world changes, it is important to understand that operational data now has to encompass a broader set of data sources, including unstructured sources such as social media data in all its forms.
- approaches to data management in the big data world, including document, graph, columnar, and geospatial database architectures.
- These are referred to as NoSQL, or not only SQL, databases. In essence, you need to map the data architectures to the types of transactions.

All these operational data sources have several characteristics in common:

- Represent systems of record that keep track of the critical data required for real-time, day-to-day operation of the business.
- Continually updated based on transactions happening within business units and from the web.
- Sources to provide an accurate representation of the business, they must blend structured and unstructured data.
- Must be able to scale to support thousands of users on a consistent basis. These might include transactional e-commerce systems, customer relationship management systems, or call center applications.

Role of Structured and Unstructured Data

Structured data refers to data that has a defined length and format and whose semantics are explicitly defined in metadata, schemas, and glossaries.

Much of structured data is stored in traditional relational databases and data warehouses. In addition, even more structured data is machine-generated from devices such as sensors, smart meters, medical devices, and Global Positioning Systems (GPS). These data sources are instrumental in creating cognitive systems.

Unlike structured data, unstructured or semi-structured data does not follow a specified format, and the semantics of these data types are not explicitly defined. Rather, the semantics must be discovered and extracted through techniques such as natural language processing, text analytics, and machine learning.

The need to find ways to collect, store, manage, and analyze unstructured data has become increasingly urgent. As much as 80 percent of all data is unstructured, with the amount of unstructured data growing at a rapid pace.

These unstructured data sources include data from documents, journal articles and books, clinical trials, customer support systems, satellite images, scientific data (seismic imagery, atmospheric data, and high-energy physics), radar or sonar data, mobile data, website content, and social media sites.

All these types of sources are important elements of a cognitive system because they may provide context for understanding a specific issue.

Unlike most relational databases, unstructured or semi-structured data sources are typically not transactional in nature. Unstructured data follows a variety of structures and may be large. These unstructured data are typically used with non relational databases such as NoSQL databases and include the following structures:

- *Key-Value Pair (KVP):* It is often used with semi-structured data from XML documents and EDI systems.KVP are used in lookup tables, hash tables, and configuration files.
- *Document databases:* provide a technique for managing repositories of unstructured and semi-structured data such as text documents, web pages, complete books
- *Columnar databases* : are an efficient database structure that stores data in columns rather than rows
- *Graph databases* : make use of graph structures with nodes and edges to manage and represent data
- *Spatial databases:* are those that are optimized to store and query geometric objects that can include points, lines, and polygons.

Data Services and Tools

- A distributed file system that is needed to manage the decomposition of structured and unstructured data streams. A distributed file system is often a requirement for doing complex data analytics when data comes from a variety of sources.
- Serialized services are required to support persistent data storage as well as supporting remote procedure calls.
- Coordination services are essential for building an application that leverages highly distributed data.
- Extract, transform, and load (ETL) services are required to both load and convert structured and unstructured data to support Hadoop (a key technique for organizing big data).
- Workflow services are the technique for synchronizing processing elements across a big data environment.

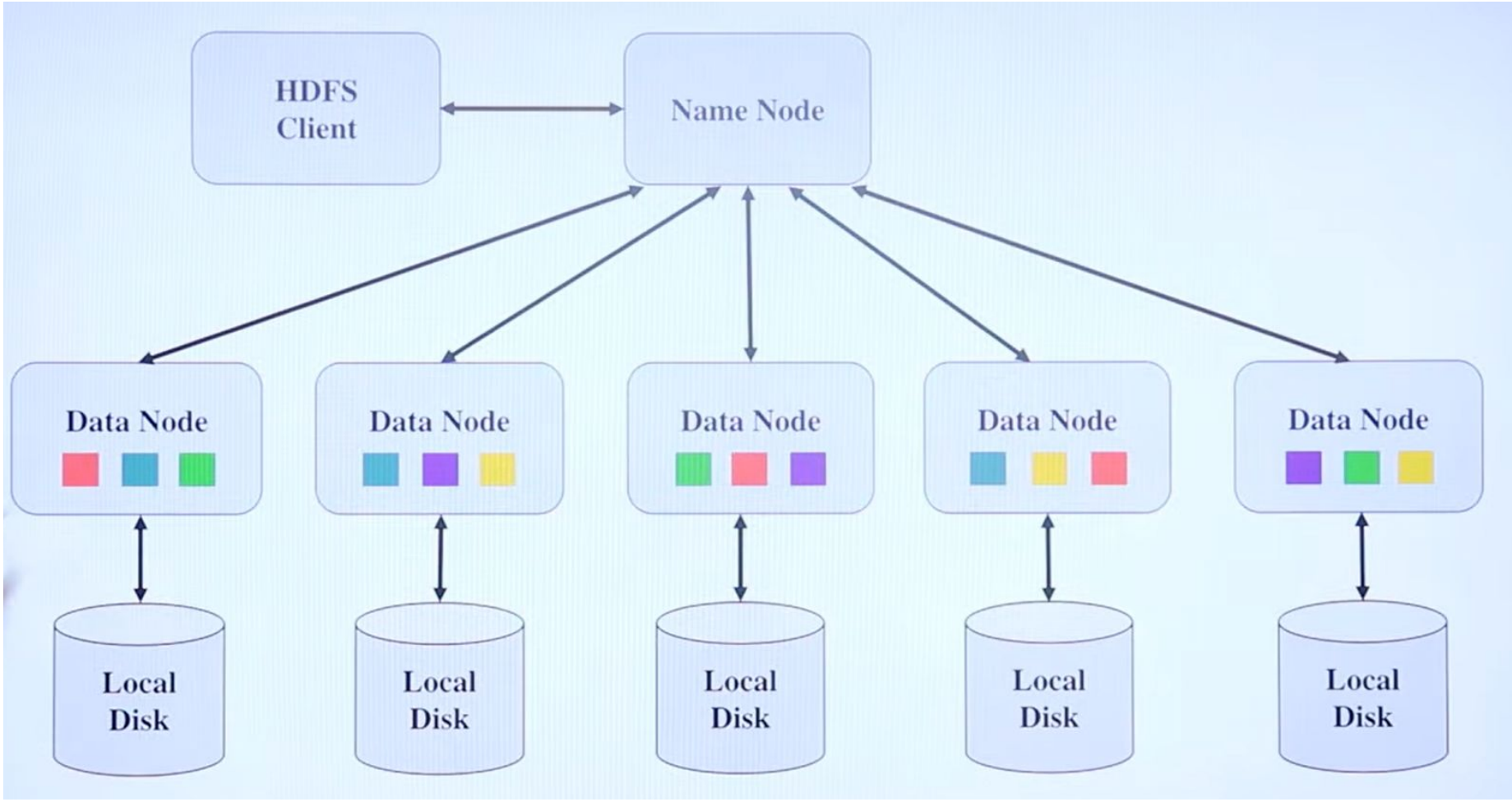
Analytical Data Warehouses

There are a number of available and emerging algorithms that support big data analytics including:

- **Sketching and streaming**: These algorithms are used when analyzing streaming data from sensors. Data elements are small but must be moved at a fast speed and require frequent updating.
- **Dimensionality reduction**: These algorithms help to convert data that is highly dimensional into much simpler data. This type of reduction is necessary so it will be easier to solve machine learning problems for classification and regression tasks.
- **Numerical linear algebra**: These algorithms are used when data includes large matrices. For example, retailers use numerical linear algebra to identify customer preferences for a large variety of products and services.
- **Compressed sensing**: These algorithms are useful when the data is sparse or signal data from a streaming sensor are limited to a few linear or time-based measurements. These algorithms enable the system to identify the key elements present in this limited data.

Hadoop

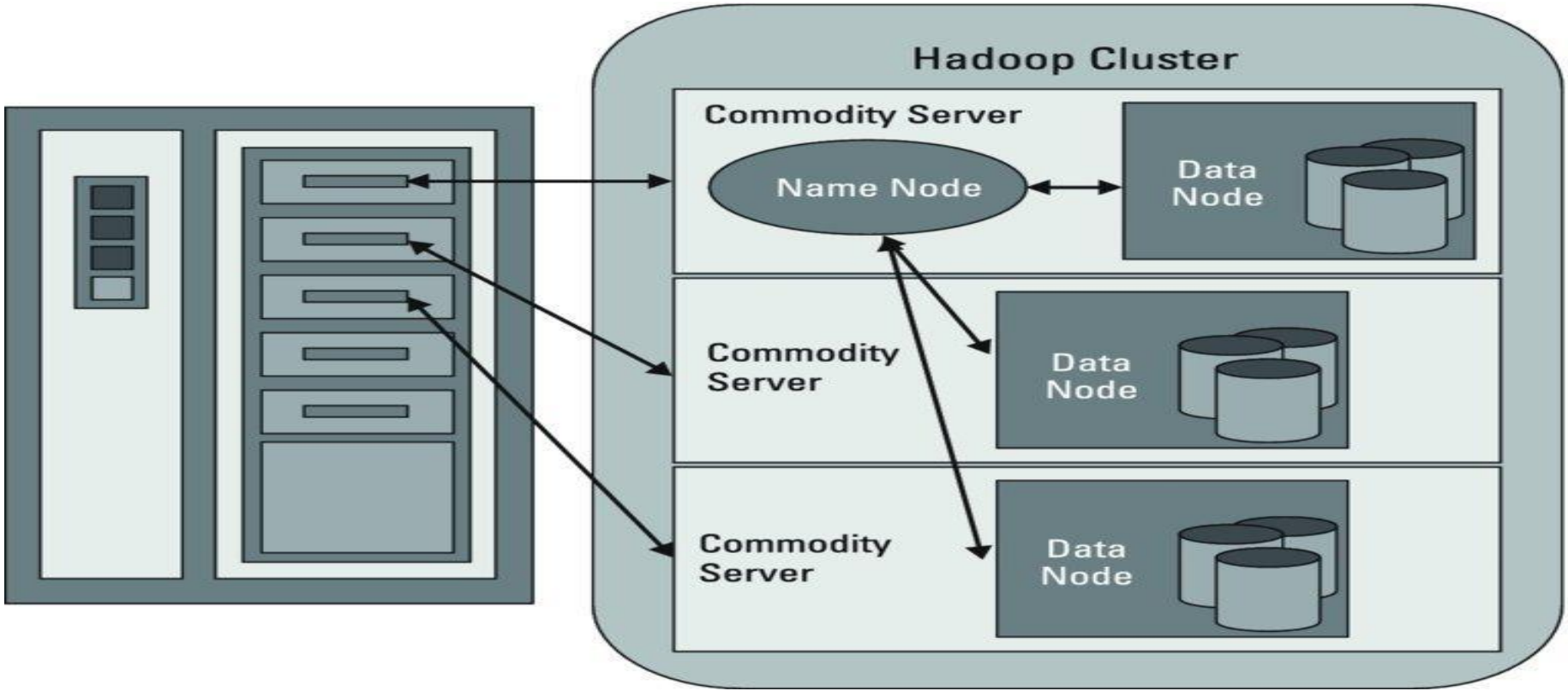
- Hadoop has emerged as one of the most important technologies for managing large amounts of unstructured data in an efficient manner because it uses distributed computing techniques.
- Hadoop enables you to use parallelization techniques to improve efficiency.
- It is an open-source community, codebase, and market for a big data environment that is designed, among other things, to parallel-execute code written to MapReduce.
- Text documents, ontologies, social media data, sensor data, and other forms of nontraditional data types can be efficiently managed in Hadoop.


HDFS
Client
Name Node
Data Node
Data Node
Data Node
Data Node
Data Node
Local
Disk
Local
Disk
Local
Disk
Local
Disk
Local
Disk

Two key components to Hadoop are described here:

- **Hadoop Distributed File System (HDFS):** A data storage cluster that is both highly reliable and low cost used to make it easy to manage related files across different machines.
- **MapReduce engine:** Provides a way to distribute the processing of the analytics algorithms across a large number of systems. After the distributed computation is complete, all the elements are aggregated back together to provide a result.

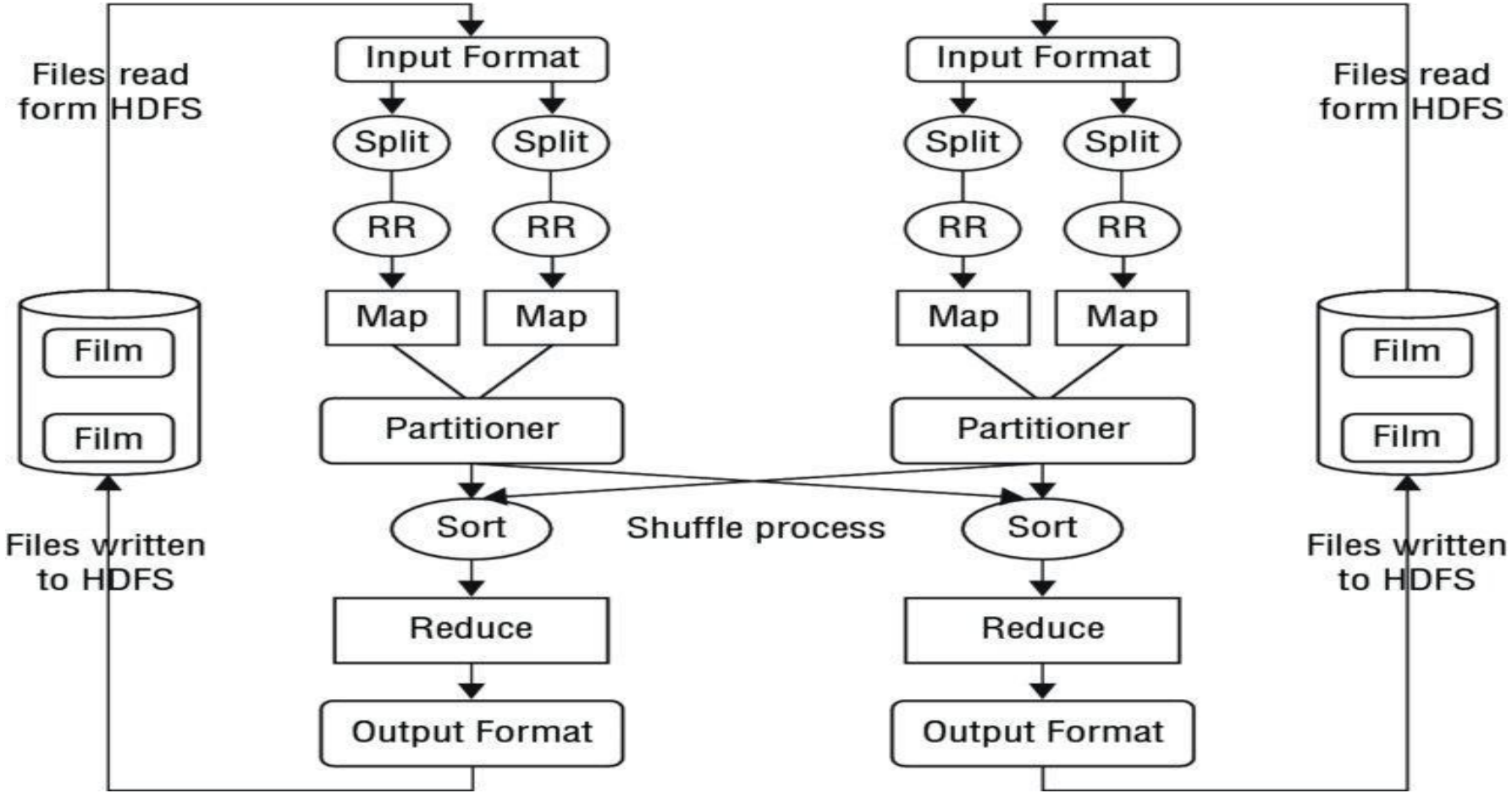
Example of a Hadoop cluster



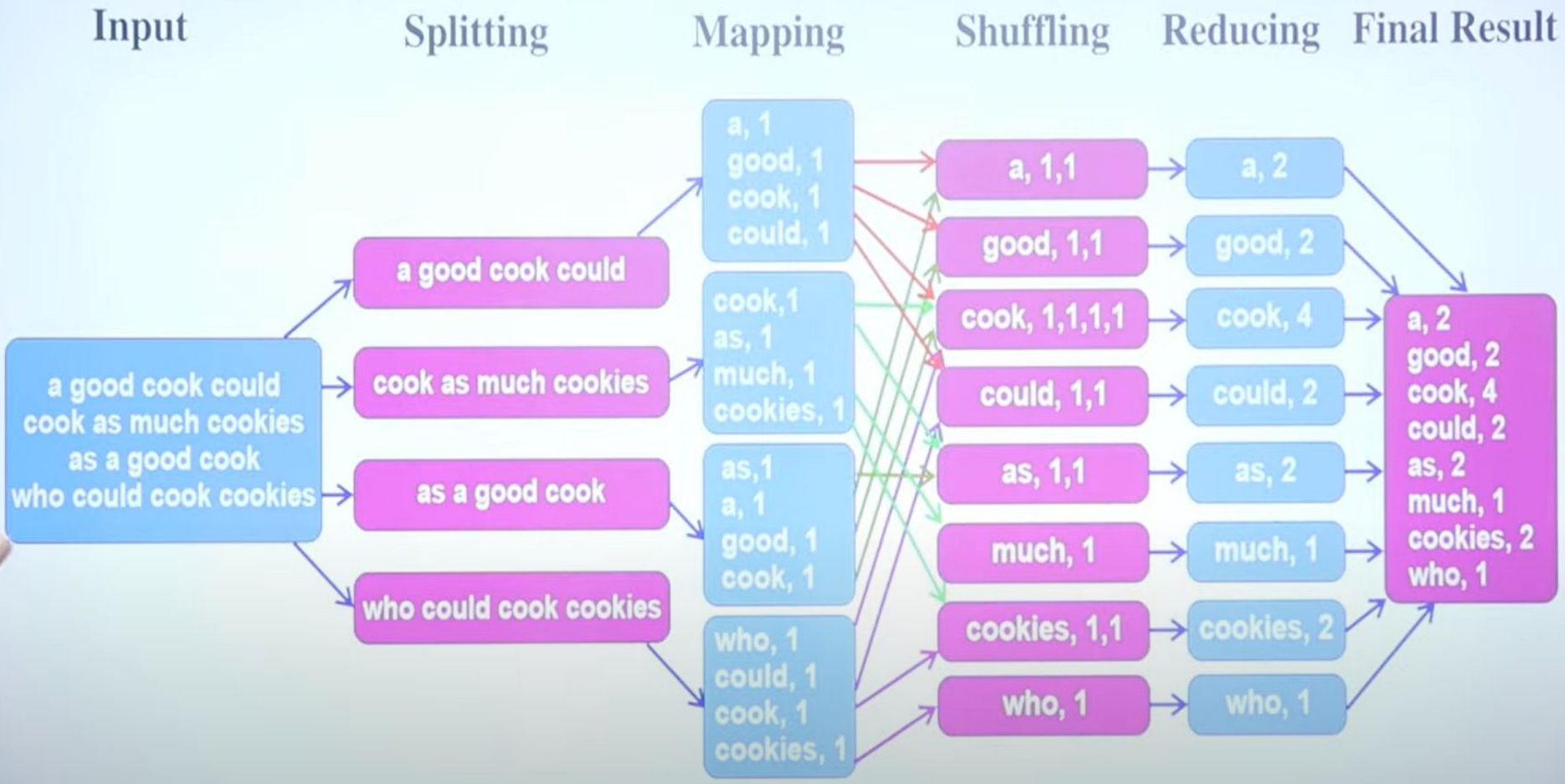
These architectural elements are described here:

- **NameNodes:** The role of the NameNode is to keep track of where data is physically stored in the cluster. To maintain this knowledge, the NameNode needs to understand which blocks on which data nodes make up the complete file. The NameNode manages all access to the files, including reads, writes, creates, deletes, and replication of data blocks on the data nodes. In addition, NameNode has the important responsibility of telling the Data Nodes if there is anything for them to do. Because NameNode is critical to keep the HDFS working, it should be replicated to protect against a single point of failure.

- **Data Nodes:** Data Nodes act as servers that contain the blocks for a set of files. Of the two components, NameNodes have some intelligence, whereas Data Nodes are more simplistic. However, they are also resilient and have multiple roles to play. They store and retrieve the data blocks in the local filesystem of the server. They also store the metadata of a block in the filesystem. In addition, Data Nodes send reports to the NameNode about what blocks are available for file operations. Blocks are stored on Data Nodes.



Workflow and data movement in a small Hadoop cluster


Input
Splitting
Mapping
Shuffling
Reducing
Final Result
a good cook could
cook as much cookies
as a good cook
who could cook cookies
a good cook could
cook as much cookies
as a good cook
who could cook cookies
a, 1
good, 1
cook, 1
could, 1
cook,1
as, 1
much, 1
cookies, 1
as,1
a, 1
good, 1
cook, 1
who, 1
could, 1
cook, 1
cookies, 1
a, 1,1
good, 1,1
cook, 1,1,1,1
could, 1,1
as, 1,1
much, 1
cookies, 1,1
who, 1
a, 2
good, 2
cook, 4
could, 2
as, 2
much, 1
cookies, 2
who, 1
a, 2
good, 2
cook, 4
could, 2
as, 2
much, 1
cookies, 2
who, 1

Data in Motion and Streaming Data

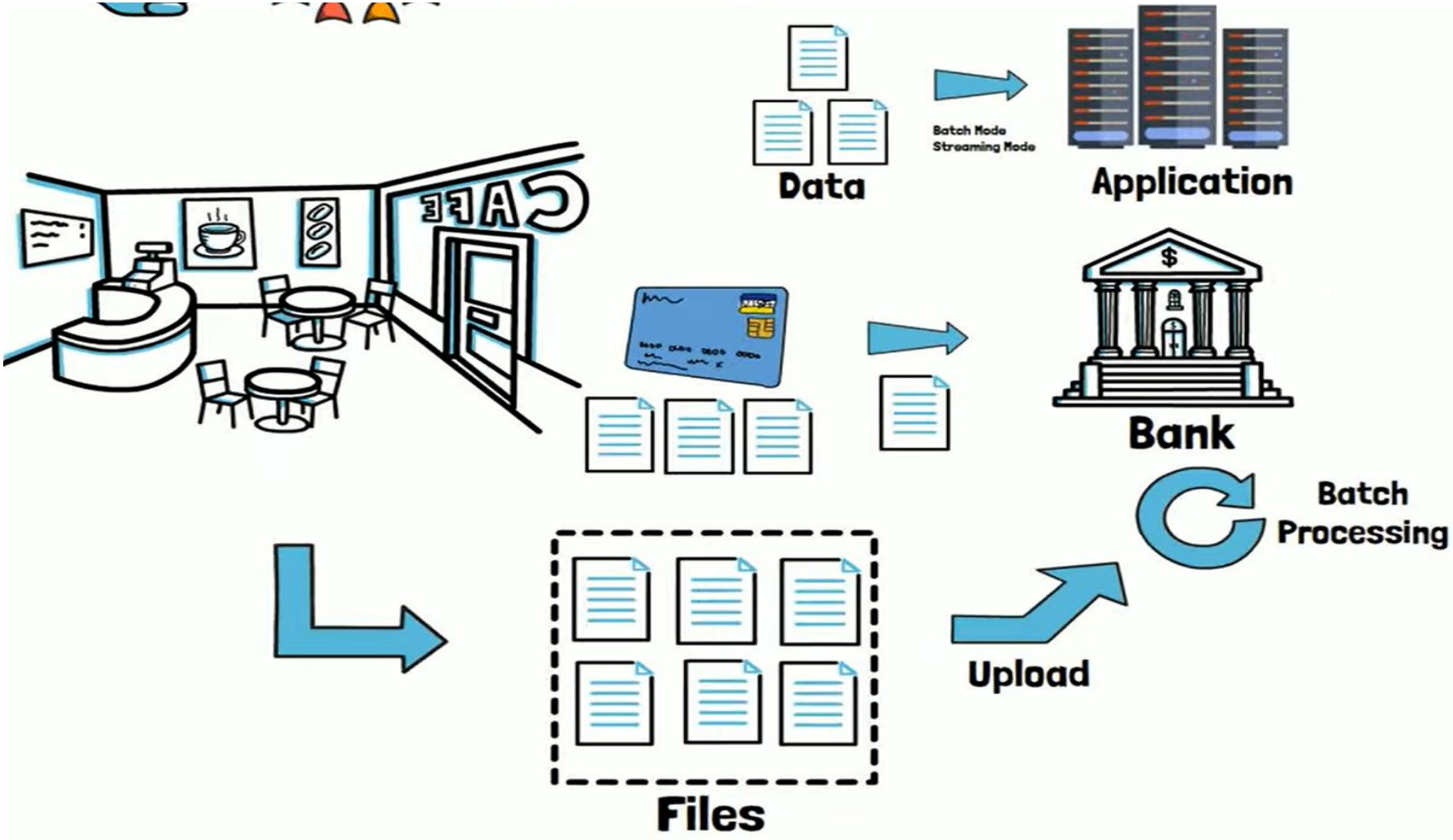
Streaming data is a continuous sequence of data that is moving at fast speeds. There are many examples of streaming data ranging from data coming from equipment sensors to medical devices to temperature sensors to stock market financial data and video streams. Streaming data platforms are designed to process this data at high speeds. Speed is of the highest priority when processing streaming data, and it can't be compromised or the results will not be useful. Streaming data is useful when analytics need to be done in real time while the data is in motion
The uses for streaming data include the following:

- In power plant management, there is the need for a highly secure environment so that unauthorized individuals do not interfere with the delivery of power to customers. Companies often place sensors around the perimeter of a site to detect movement. But not all forms of movement represent a threat. For example, the system needs to be able to detect if an unauthorized person is accessing a secure area

DATA
AT REST

DATA
IN USE

DATA
IN MOTION


Data
Batch Mode
Streaming Mode
Application
CAFE
Bank
Batch
Processing
Upload
Files

- versus an animal walking around. Clearly, the innocent rabbit does not pose a security risk. Therefore, the vast amount of data coming from these sensors needs to be analyzed in real time so that an alarm is sounded only when an actual threat exists.
- In manufacturing, it will be important to use the data coming from sensors to monitor the purity of chemicals being mixed in the production process. This is a concrete reason to leverage the streaming data. However, in other situations, it may be possible to capture a lot of data, but no overriding business requirement exists. In other words, just because you can stream data doesn’t mean that you always should.

- In medical applications, sensors are connected to highly sensitive medical equipment to monitor performance and alert technicians of any deviations from expected performance. The recorded data is continuously in motion to ensure that technicians receive information about potential faults with enough lead time to make a correction to the equipment and avoid potential harm to patients.
- In the telecommunications industry it is critical to monitor large volumes of communications data to ensure that service levels meet customer expectations.
- In the retail industry, point-of-sale data is analyzed as it is created to try to influence customer decision making. Data is processed and analyzed at the point of engagement and maybe used in combination with location data or social media data.
- Understanding the context of data collected is critical in at-risk physical locations. The system has to be able to detect the context of the incident and determine if there is a problem.
- Medical organizations can analyze complex data from medical devices. The resulting analysis of this streaming data can determine different aspects of a patient's condition and then match results against known conditions or other abnormal indicators.

Integration of Big Data with Traditional Data

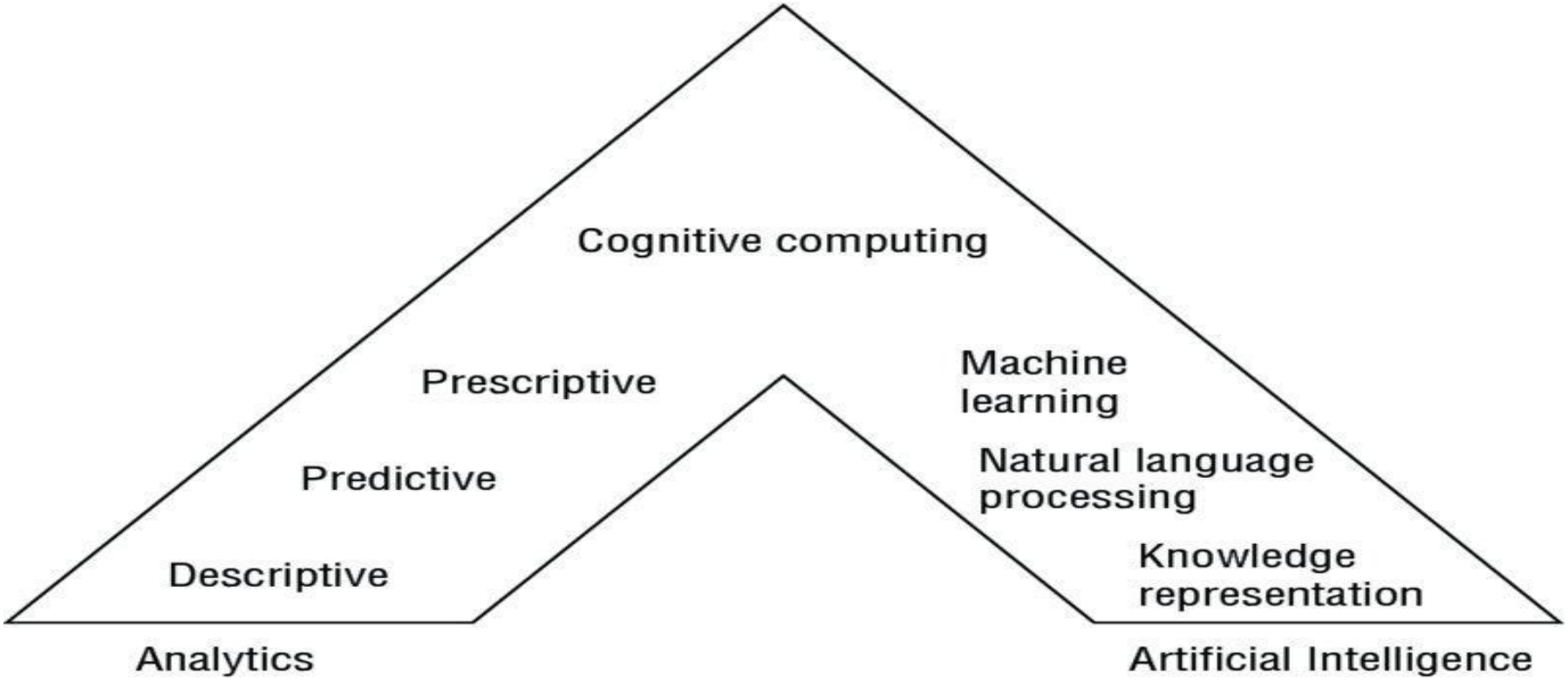
Although much of the attention in big data has been focused on accessing and analyzing complex unstructured data, it is important to understand that the results of analysis of this data has to be integrated with traditional relational databases, data warehouses, and line of business applications. To create a cognitive system requires that an organization have a holistic view of the required data so that the context is correct.

Therefore, building a cognitive system requires that the massive amounts of data be managed and analyzed. It also requires that there are the right data integration tools and techniques in place to effectively create the corpus. This is not a static process. To be effective, all types of big data must be moved, integrated, and managed based on the problem being addressed.

Advanced analytics is on a path to cognitive computing,

ANALYTICS TYPE	DESCRIPTION	EXAMPLES OF QUESTIONS ANSWERED
Descriptive Analytics	Understand what happens when using analytic techniques on historical and current data.	Which product styles are selling better this quarter as compared to last quarter? Which regions are exhibiting the highest/lowest growth? What factors are impacting growth in different regions?
Predictive Analytics	Understand what might happen when using statistical predictive modeling capabilities, including data mining and machine learning. Predictive models use historical and current/real-time data to predict future outcomes. Models look for trends, clusters of behavior, and events. Models identify outliers.	What are the predictions for next quarter's sales by product and region? How does this impact raw material purchases, inventory management, and human resource management?
Prescriptive Analytics	Use to create a framework for making a decision about what to do or not do in the future. The "predictive" element should be addressed in prescriptive analytics to help identify the relative consequences of your actions. Use an iterative process so that your model can learn from the relationship between actions and outcomes.	What is the best mix of products for each region? How will customers in each region react to advertising promotions and offers? What type of offer should be made to each customer to build loyalty and increase sales?
Machine Learning and Cognitive Computing	Collaboration between humans and machines to solve complex problems. Assimilate and analyze multiple sources of information to predict outcomes. Need depends on the problems you are trying to solve. Improve effectiveness of problem solving and reduce errors in predicting outcomes.	How secure is the city environment? Are there any alerts from the vast amount of information streaming from monitoring devices (video, audio, and sensing devices for smoke or poisonous gases)? Which combination of drugs will provide the best outcome for this cancer patient based on the specific characteristics of the tumor and genetic sequencing?

Converging technologies: analytics and artificial intelligence



Key Capabilities in Advanced Analytics

Machine learning is applied to improve the accuracy of the models and make better predictions.

It is an essential technology for advanced analytics, particularly because of the need to analyze big data sources that are primarily unstructured in nature. In addition to the following

- Machine learning
- Advanced analytics capabilities including
  - Predictive analytics
  - Text analytics
  - Image analytics
  - Speech analytics

The Relationship Between Statistics, Data Mining, and Machine Learning

Statistics, data mining, and machine learning are all included in advanced analytics

The following highlights how these capabilities relate to each other.

- **Statistics** is the science of learning from data. Classical or conventional statistics is inferential in nature, meaning it is used to reach conclusions about the data (various parameters).
- **Data mining** which is based on the principles of statistics, is the process of exploring and analyzing large amounts of data to discover patterns in that data.
- **Machine learning** uses some of the same algorithms that are used in data mining. One of the key differences in machine learning as compared to other mathematical approaches is the focus on using iterative methods to reduce the errors.

Using Machine Learning in the Analytics Process

Machine learning is essential to improving the accuracy of predictive models in a cognitive environment.

Supervised and unsupervised machine learning algorithms are used in a variety of analytics applications.
Labeled data refers to the identification or tag that provides some information about the data.

Supervised Learning

Supervised learning typically begins with an established set of data and a certain understanding of how that data is classified.
The following tools and techniques are often used to implement supervised learning algorithms.

- **Regression**—Regression models were developed in the statistical community.
- **Decision tree**—A decision tree is a representation or data structure that captures the relationships among a set of categories.
- **Neural networks—**Neural network algorithms are designed to emulate human/animal brains.
- **Support Vector Machine (SVM)**—SVM is a machine learning algorithm that works with labeled training data and output results to an optimal hyperplane.
- **k-Nearest Neighbor (k-NN)**—k-NN is a supervised classification technique that identifies groups of similar records.

Unsupervised Learning
Unsupervised learning algorithms can solve problems that require large volumes of unlabeled data.

The following tools and techniques are typically used in unsupervised learning.

- **Clustering** techniques are used to find clusters that exist in the data sample.
  - The **K-means algorithm** can estimate the unknown means based on the data. This is probably the most widely used unsupervised learning algorithm. It is a simple local optimization algorithm.
  - **EM-Algorithm for clustering** can maximize the mixture density given the data.
- **Kernel density estimation (KDE)** estimates the probability distribution or the density of a data set.

- **Nonnegative matrix factorization (NMF)** is useful in pattern recognition and to solve challenging machine learning problems in fields such as gene expression analysis and social network analysis.
- **Principal Components Analysis (PCA)** is used for visualization and feature selection.
- **Singular Value Decomposition (SVD)** can help to eliminate redundant data to improve the speed and overall performance of the algorithm.

Predictive Analytics

Predictive analytics is a statistical or data mining solution consisting of algorithms and techniques that can predict future outcomes.

Business Value of Predictive Analytics
Companies use predictive analytics to solve many business challenges, including reducing customer churn, improving the overall understanding of customer priorities, and reducing fraud.

Text Analytics
Given the business value of text-based unstructured sources, text analytics is a critical element of cognitive systems.

Business Value of Text Analytics
The business value of text analytics increases with an organization’s capability to understand how to act or make decisions based on the content.

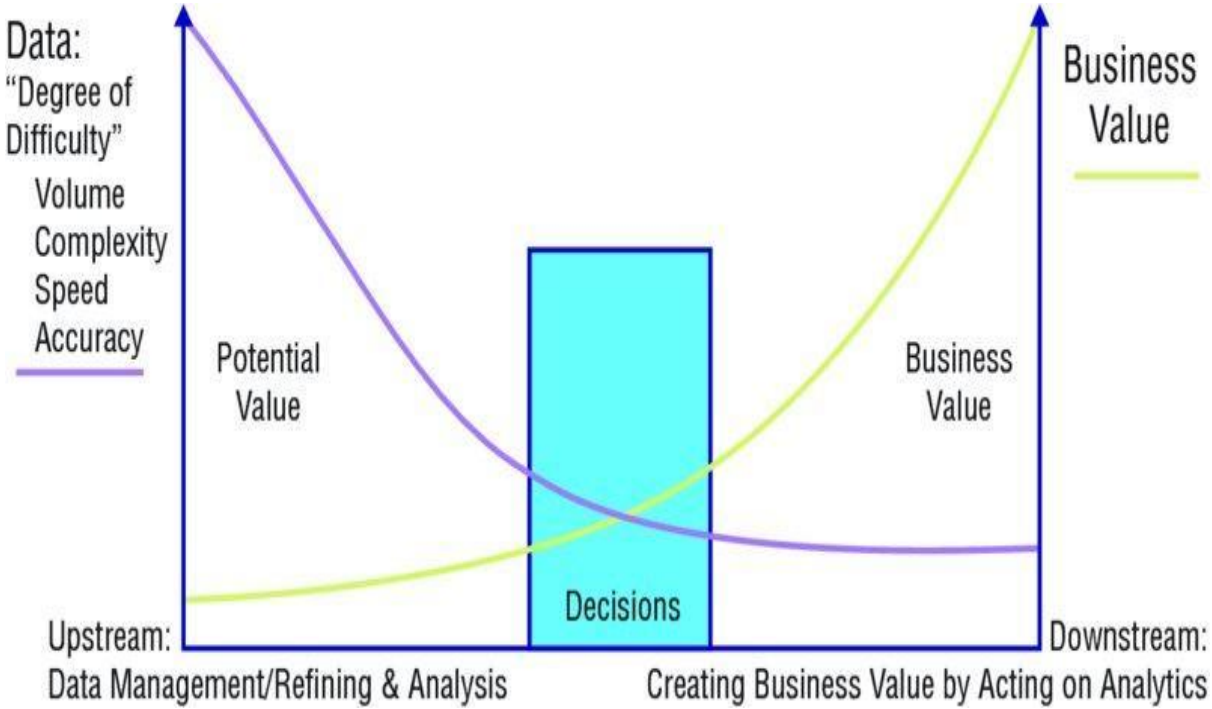
Image Analytics

The sources used to develop knowledge corpora for a cognitive system are likely to include videos, photos, or medical images.

Speech Analytics

Text, image, and speech analytics can be used in a cognitive system to provide the right context to answer a question correctly or make an accurate prediction.

Using advanced analytics to create value



The ability to make faster and more execution-oriented decisions depends on reducing the "degree of difficulty" of interpreting the right data from all input streams. You need to manage the volume and complexity of the data. At the same time, you need to sample high-velocity data in a meaningful way. Raw data as captured by systems and sensors has potential value, but needs to be processed and analyzed to build business value. Business value increases as volume, complexity, and speed are managed during the analytics process.

Impact of open source tools on advanced analytics

- Open source analytics tools are having a major impact on the growth of predictive analytics at many organizations.
- The open source software environment and programming language, R, is fast becoming one of the primary tools for data scientists, statisticians, and other enterprise users.
- R, which is designed for computational statistics and data visualization, is the language of choice for graduate students doing research in advanced analytics and cognitive computing.
- Strong interest in R has led to a very active open source community. Members of the community share information on models, algorithms, and coding best practices.
- Users like the flexibility that a special-purpose programming language and environment offers for building custom applications.
- Some of the benefits of R include its flexibility and adaptability.
- R is actually an implementation of the statistical programming language S, developed at Bell Laboratories, as a higher-level alternative to using FORTRAN statistical subroutines.