Cloud Computing for Big Data

Content

- Big data
- Cloud computing for big data
- Hadoop
- MapReduce
- Conclusion
- References

Introduction

What is Big Data?

- A collection of very huge data sets with a great diversity of types.
- It becomes difficult to process by using traditional data processing.

More Definitions

Gartner's Definition (2012 to 2020):

- Big Data are high-volume, high-velocity, and/or high-variety information assets.
- They require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

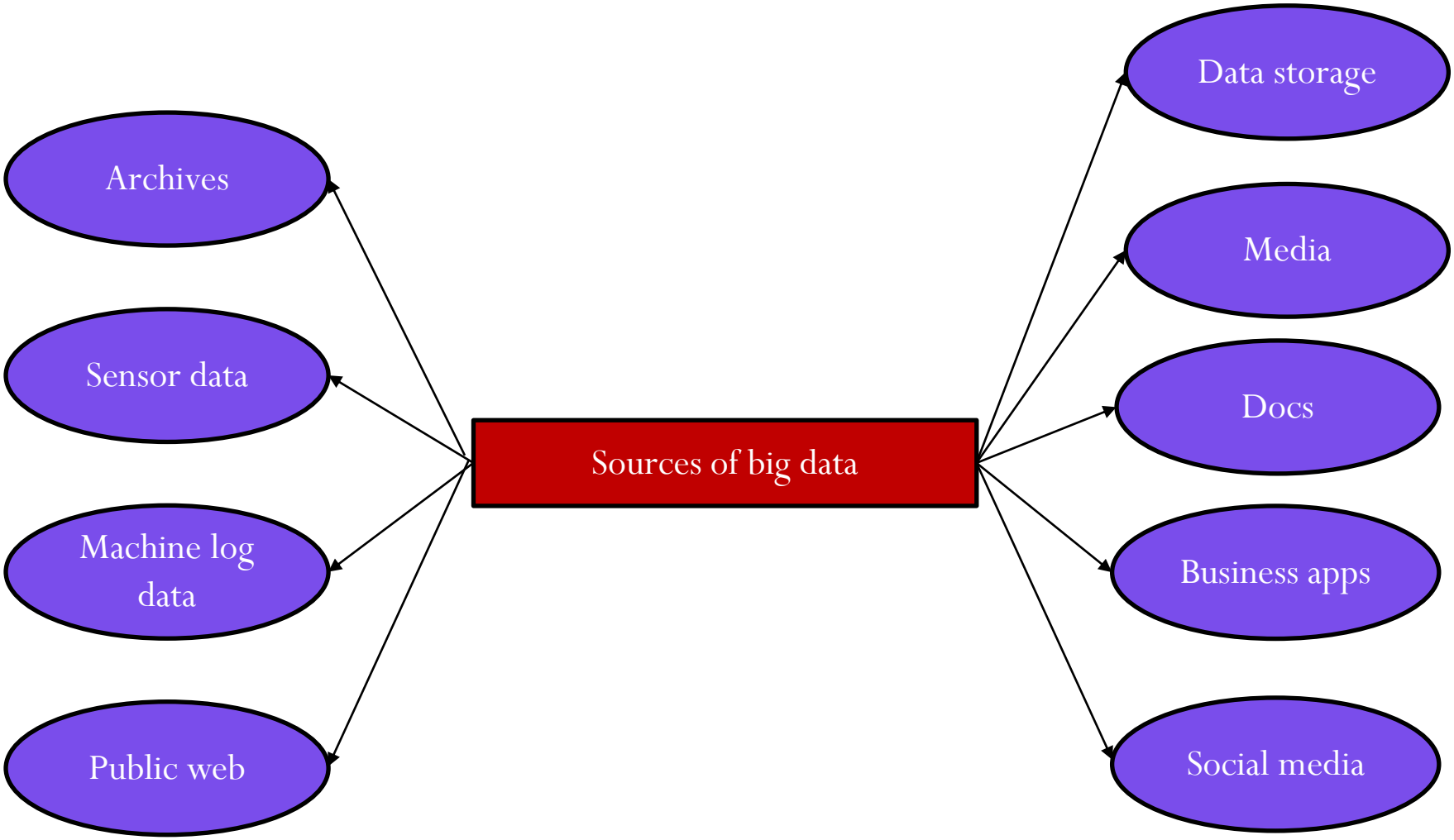
More generally:

- A data set can be called Big Data if it is difficult to perform capture, curation, analysis and visualization on it at the current technologies.

Some Facts

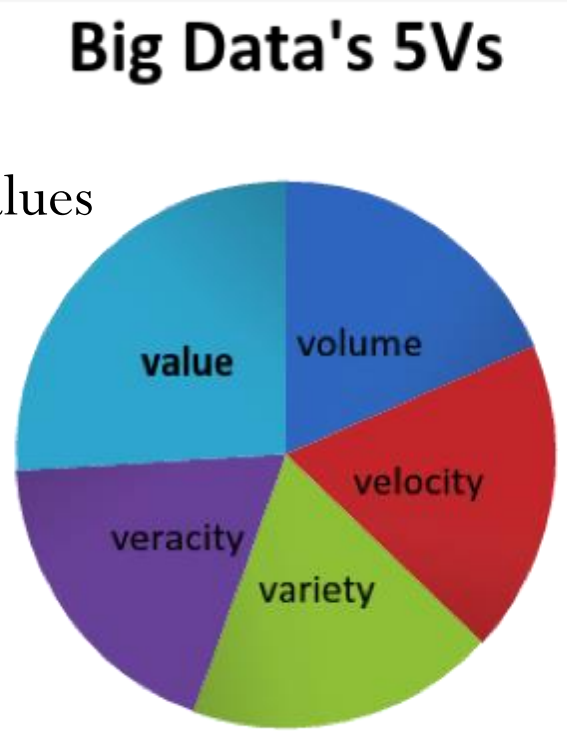
- Data is growing at 40% annual rate that will reach 88 ZB by 2022.
- Over 2.5 quintillion bytes of data are created every single day, and it's only going to grow from there. By 2020, it's estimated that 1.7MB of data will be created every second for every person on earth. ***Reported on Jun 15, 2018***
- Business data doubles every 1.2 years.
- 1 million customer transactions are processed by Wal-Mart per hour.
- 500 million “tweets” are posted every day.
- 3.2 billion “Likes” and comments are posted on Facebook.

Sources of Big Data

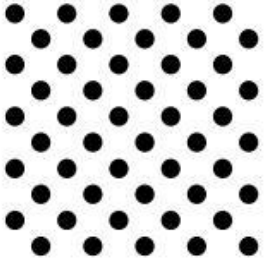
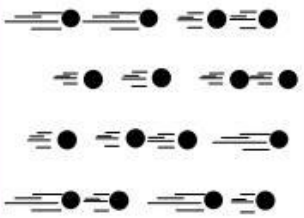
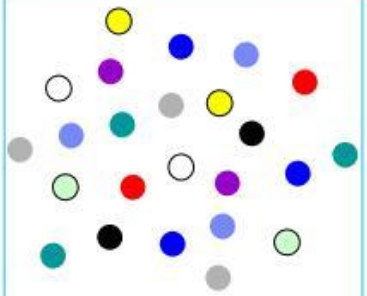
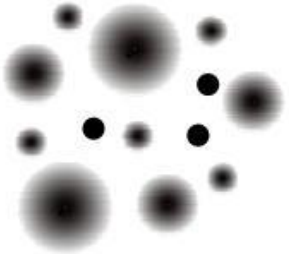
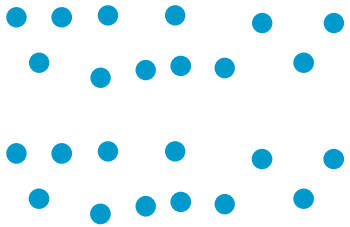


Big Data Characteristics

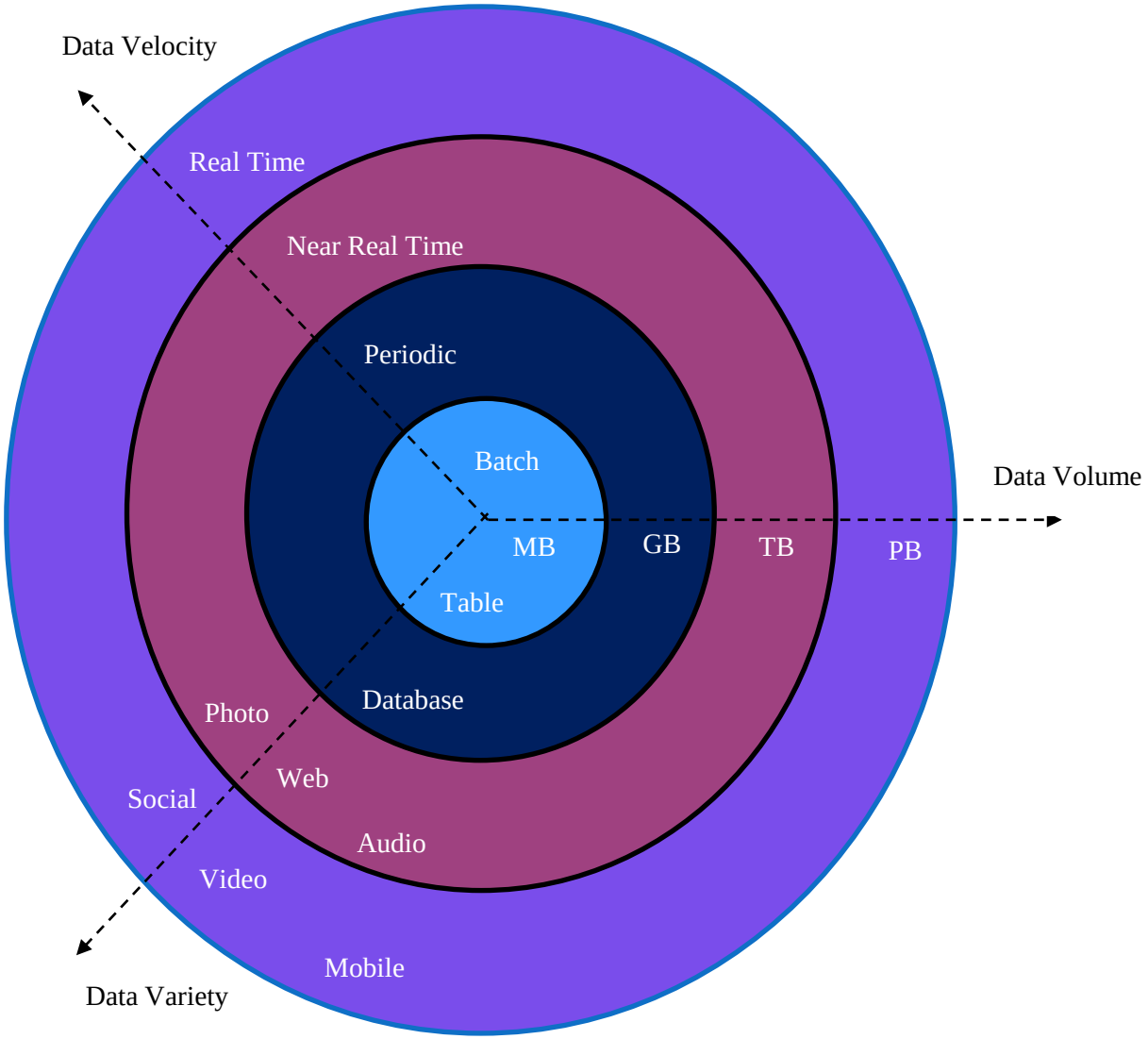
- Volume: The size of the data set
- Velocity: The speed of data in and out
- Variety: The range of data types and sources.
- Value : The process of discovering huge hidden values
- Veracity: Accuracy or correctness



5V Picture

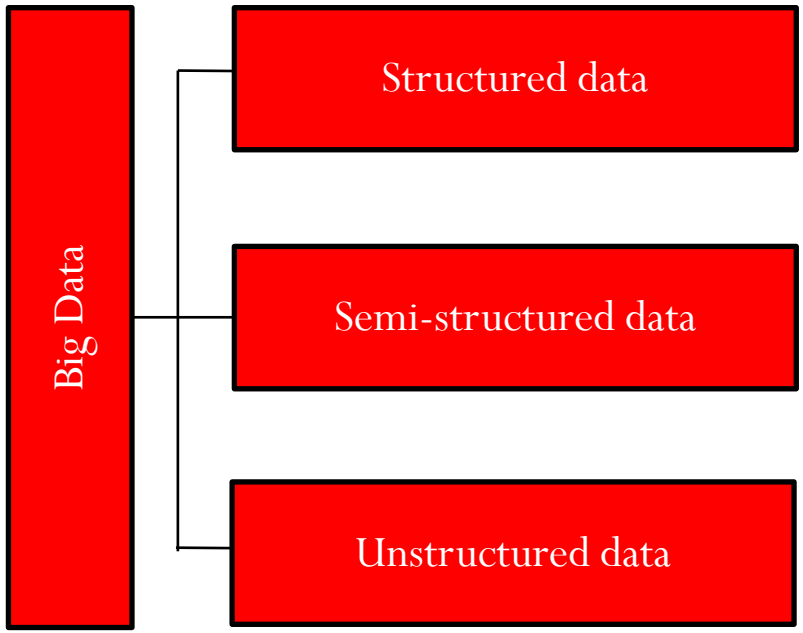
Volume	Velocity	Variety	Veracity*	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Hidden Values**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Highly useful values hidden in the huge volume of data

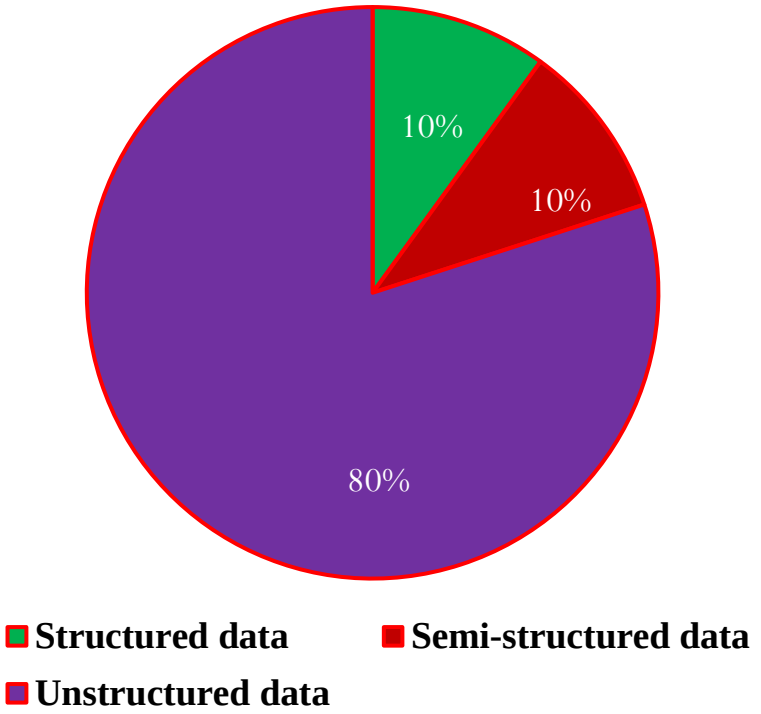
Data: Big in volume, variety & velocity



Growth of data	
MB	2^{20} bytes
GB	2^{30} bytes
TB	2^{40} bytes
PB	2^{50} bytes
EB	2^{60} bytes
ZB	2^{70} bytes
YB	2^{80} bytes

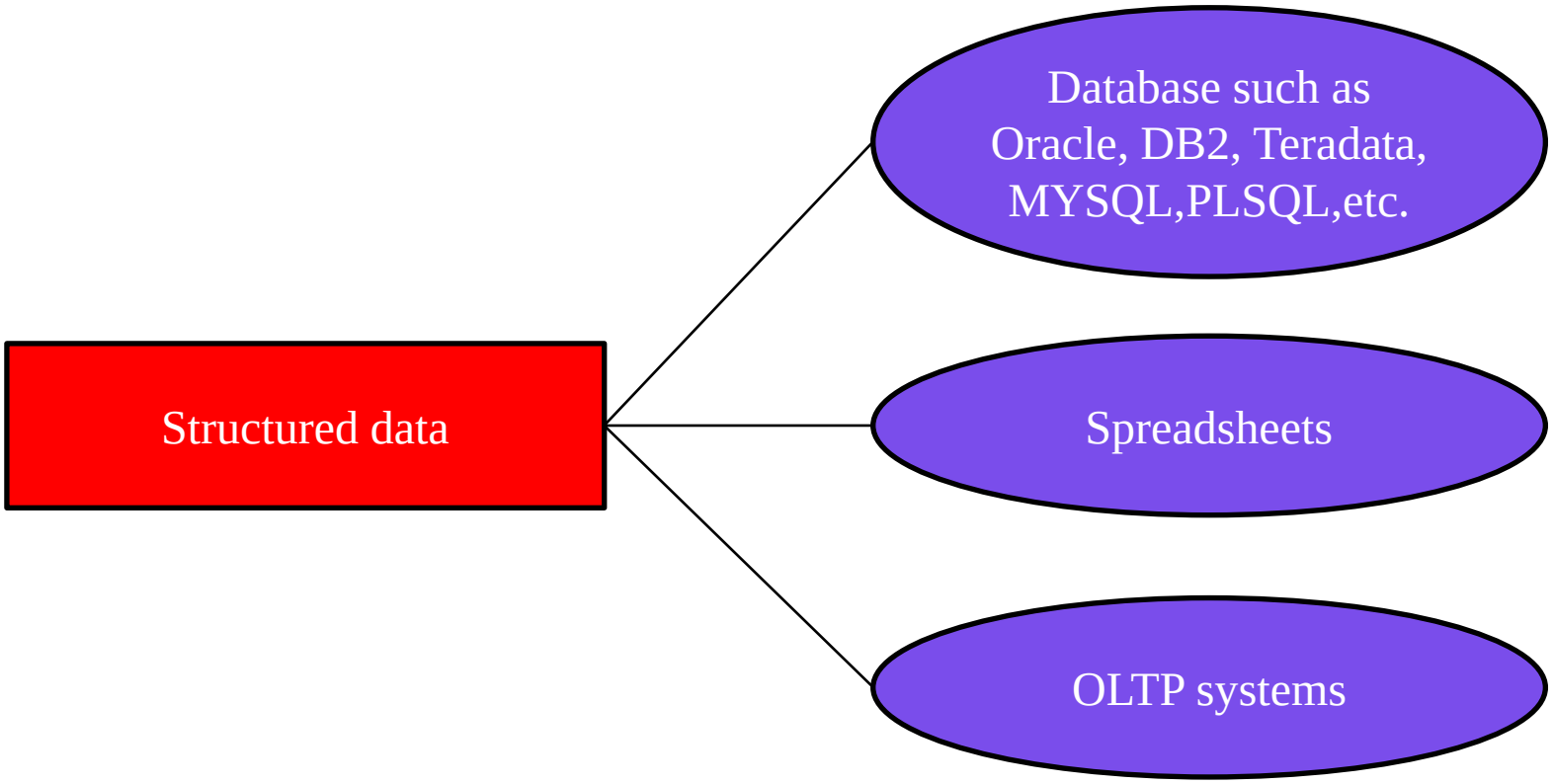
Classification of Big Data



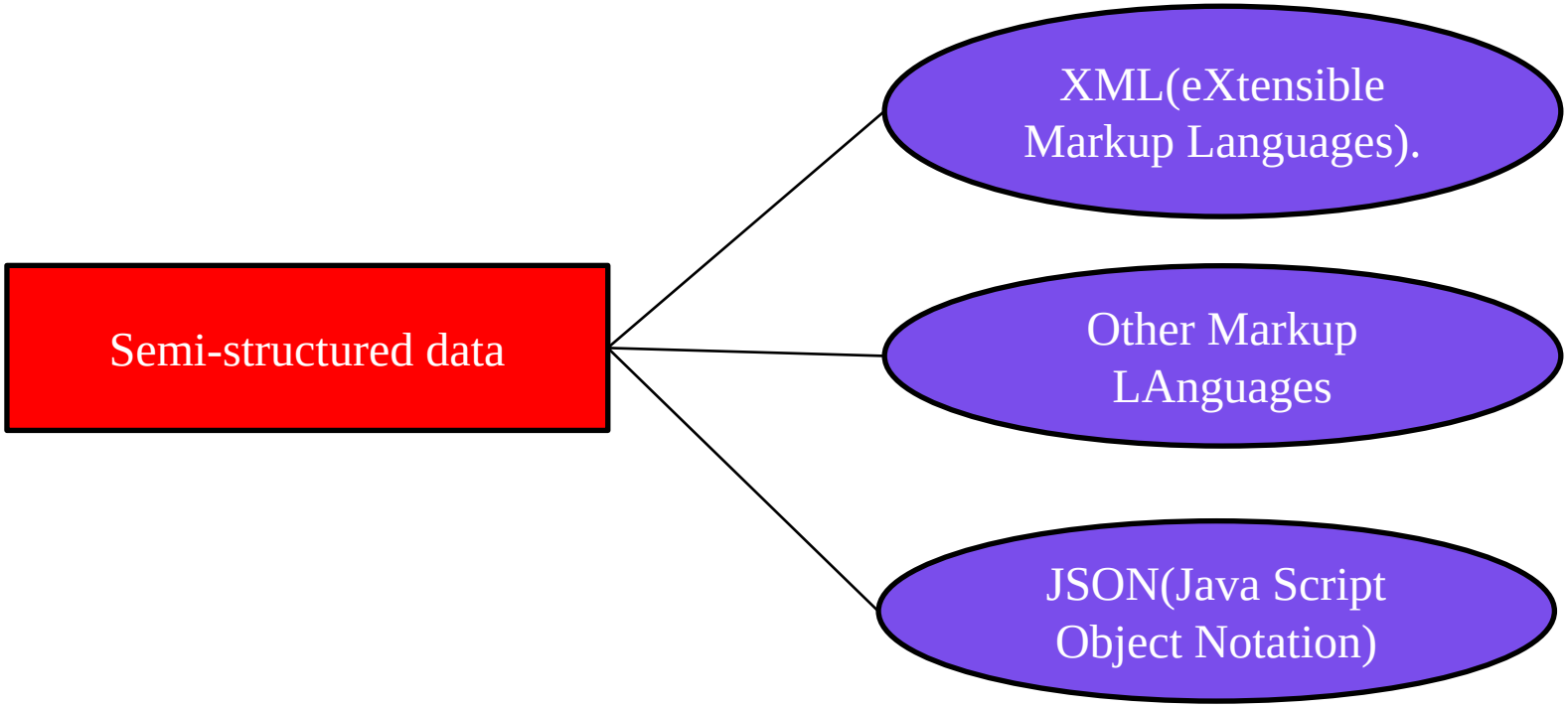


Approximate percentage distribution of Big data

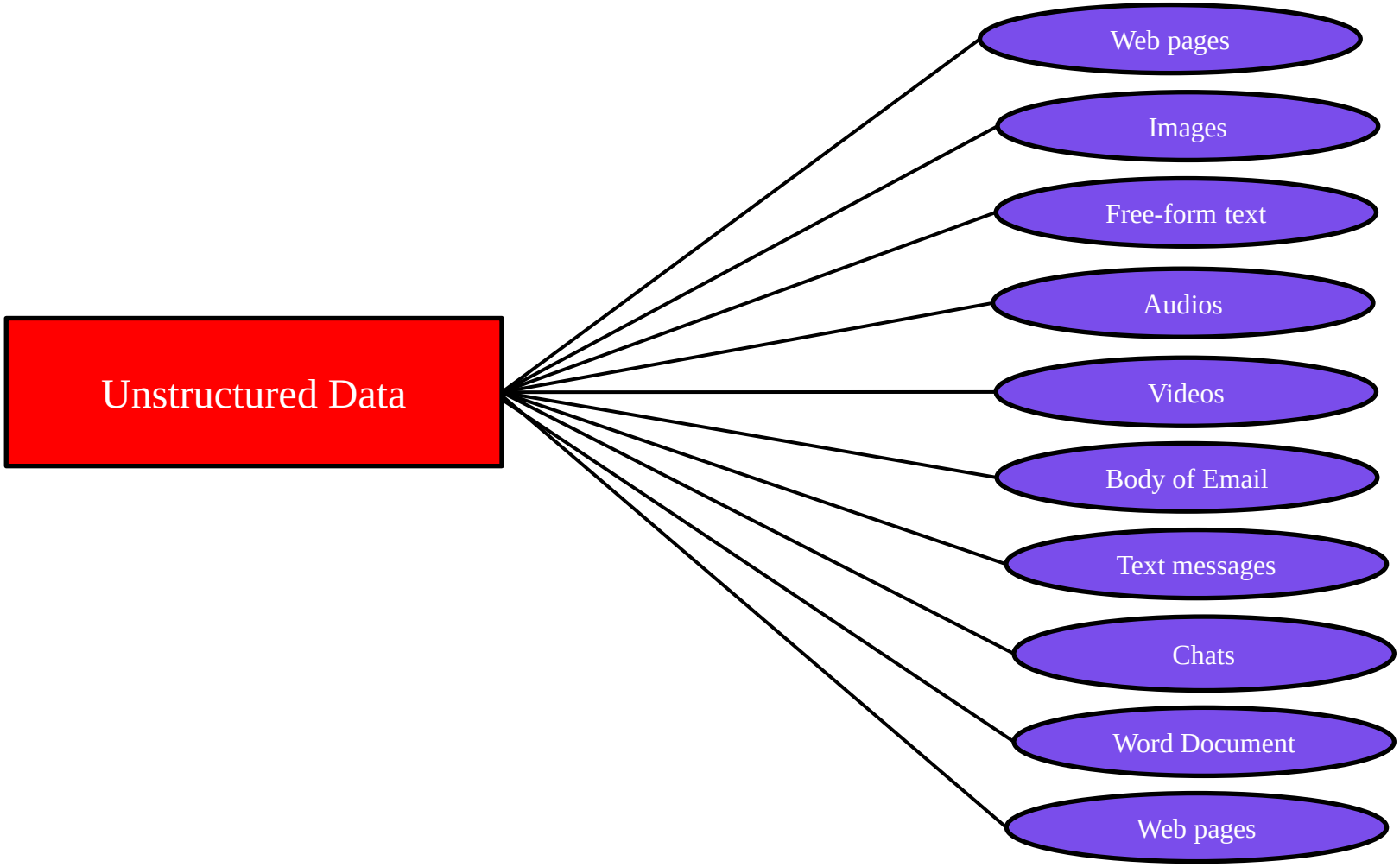
Sources of structured data



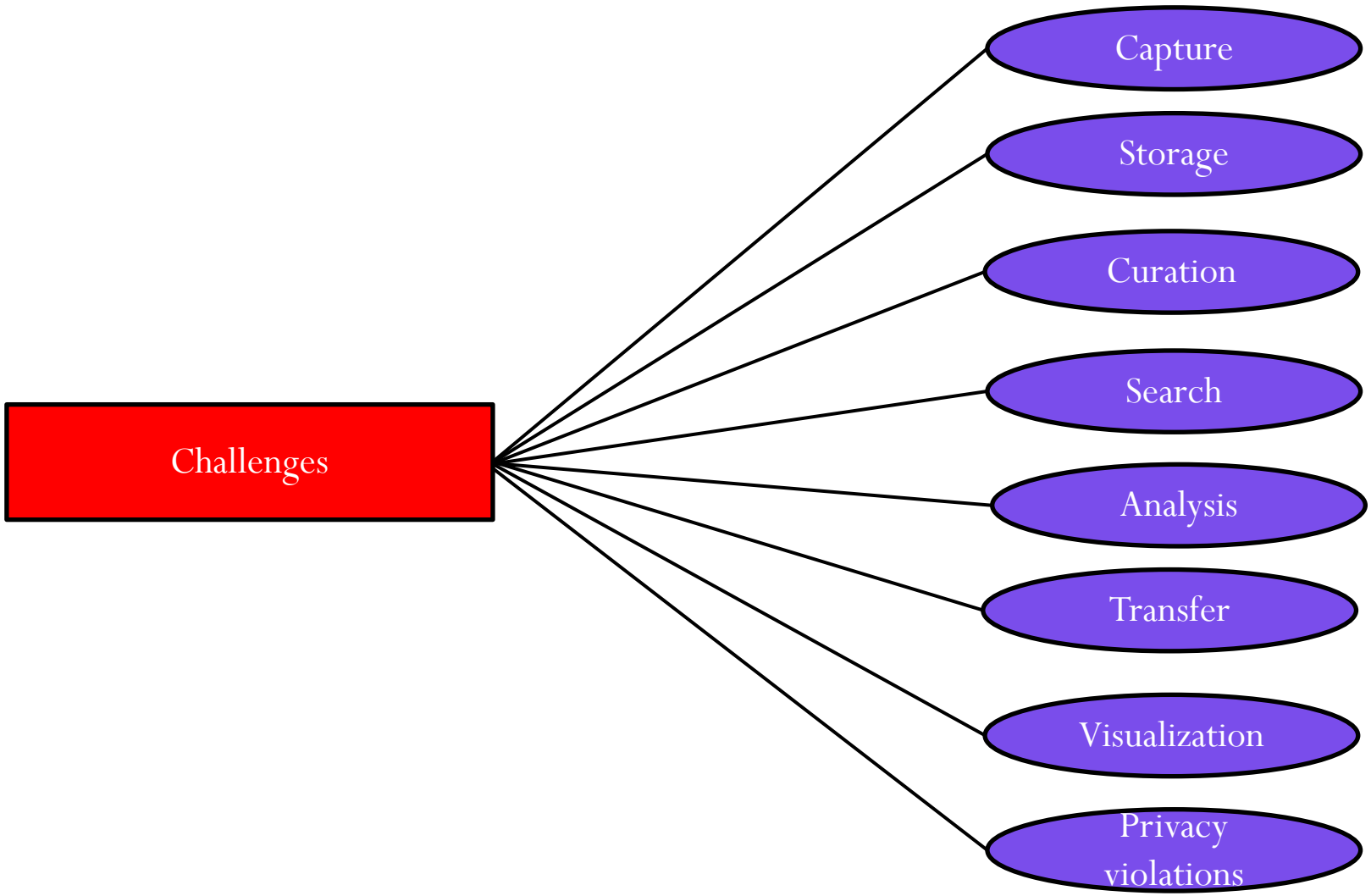
Sources of semi-structured data



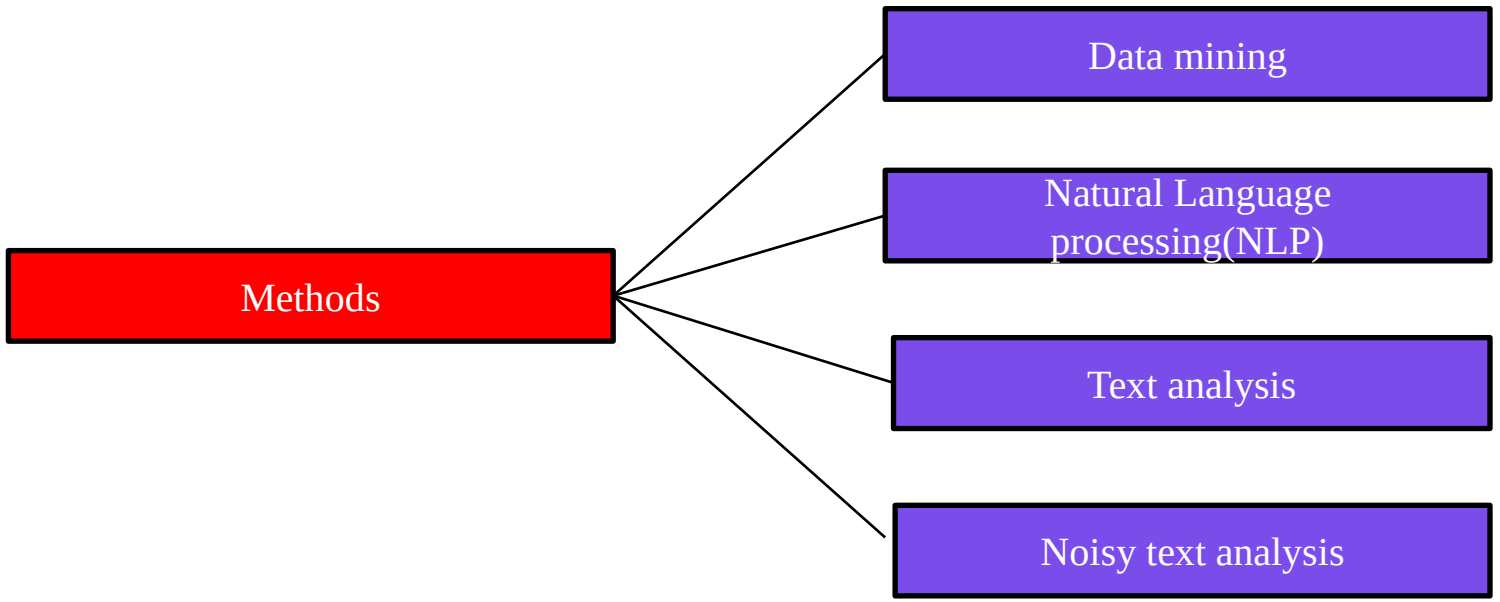
Sources of unstructured data



Challenges in Big Data



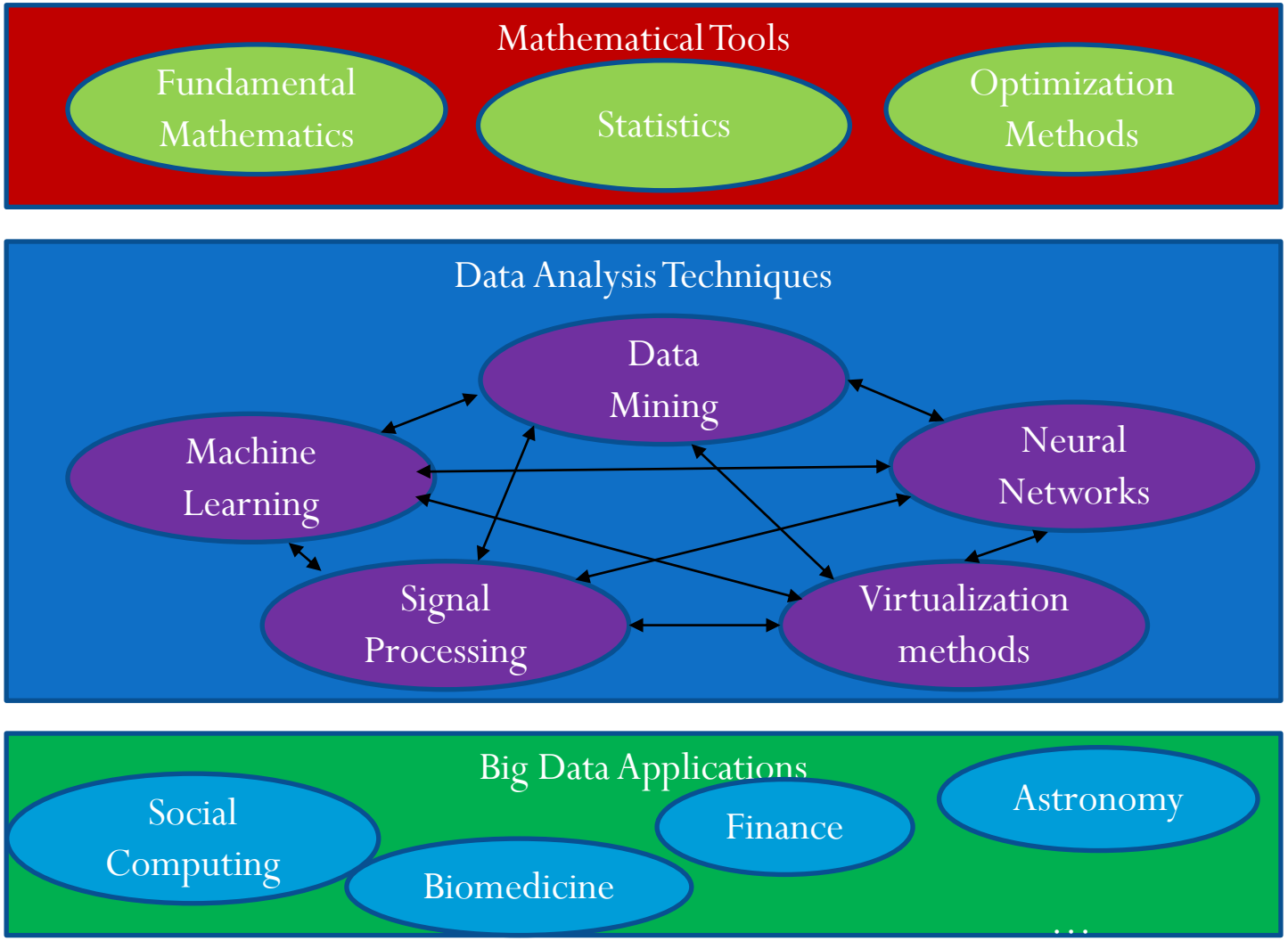
Dealing with unstructured data



Business Intelligence vs. Big Data

- Data housed in central server in BI
- Data resides in distributed file system in big data.
- BI analyses the data in offline mode
- Big data can do in real time / offline mode.
- BI can handle only structured data
- Big data can handle any data type

Big Data Techniques



Cloud Computing for Big Data

- Best model for allowing on-demand network access to computing resources

- Cloud computing provides underlying engine for big data to process distributed queries across multiple data sets through the use of Hadoop

Cloud Computing Features to Handle Big Data

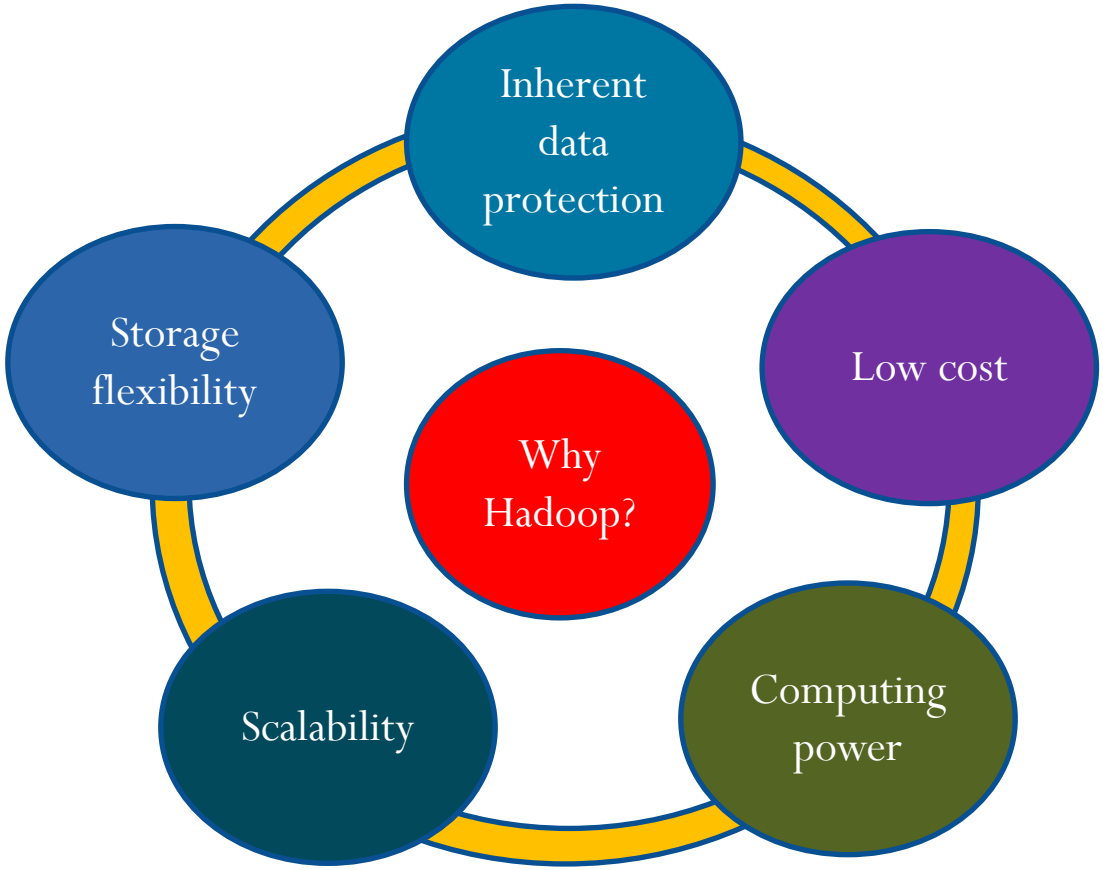
- Scalability : New resources can be added easily.
- Elasticity : Hiring resources only when required.
- Resource pooling : Sharing of the resources by organizations
- Self Service : Direct access to the cloud services.
- Low cost : Pay-as-you-use, lower initial investment.
- Fault tolerance : Uninterrupted service in case of component failure.

Hadoop

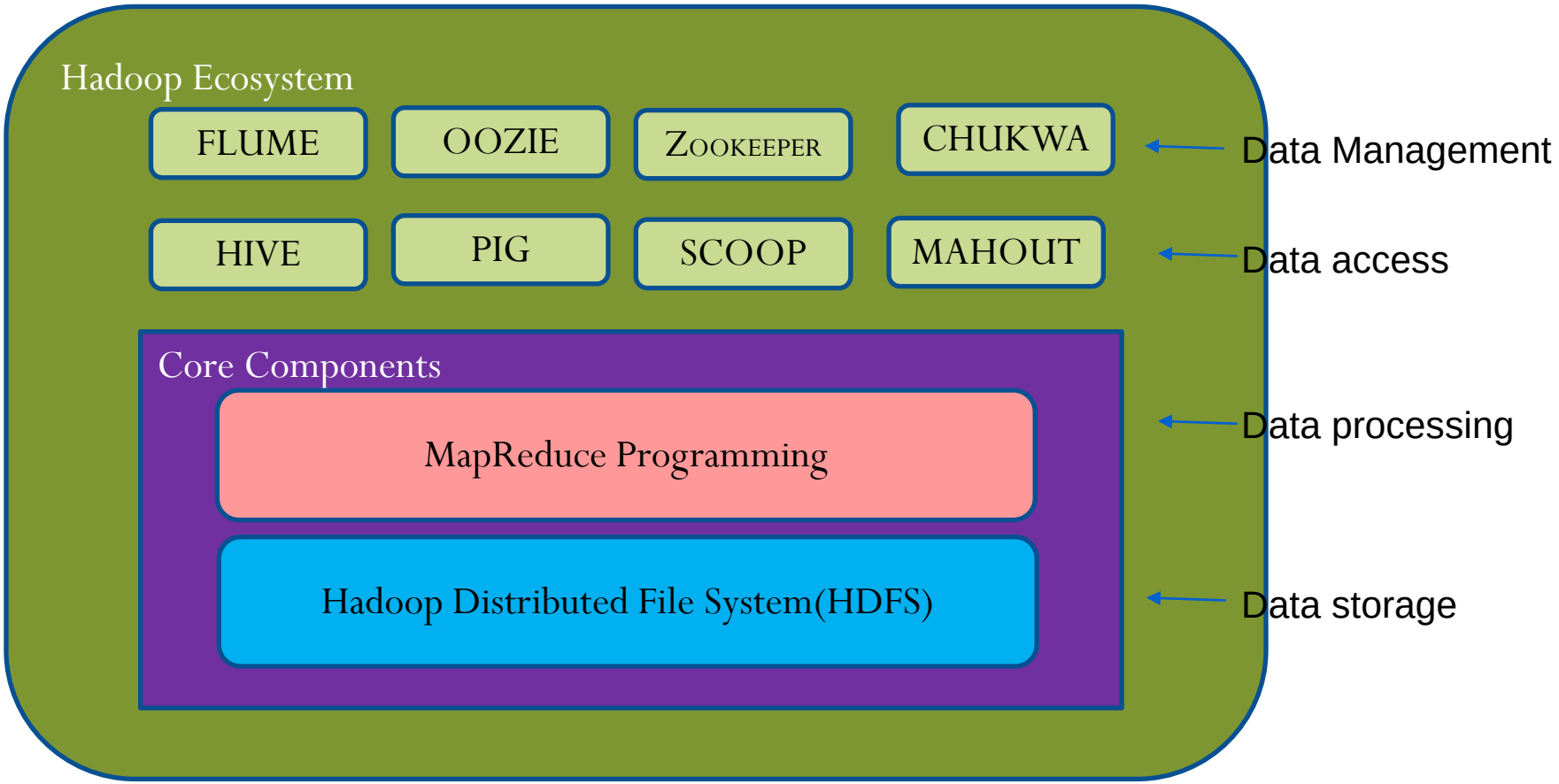
- A framework for distributed processing of large data sets across clusters of computers using simple programming model

  According to Apache

Key Considerations of Hadoop



Components of Hadoop



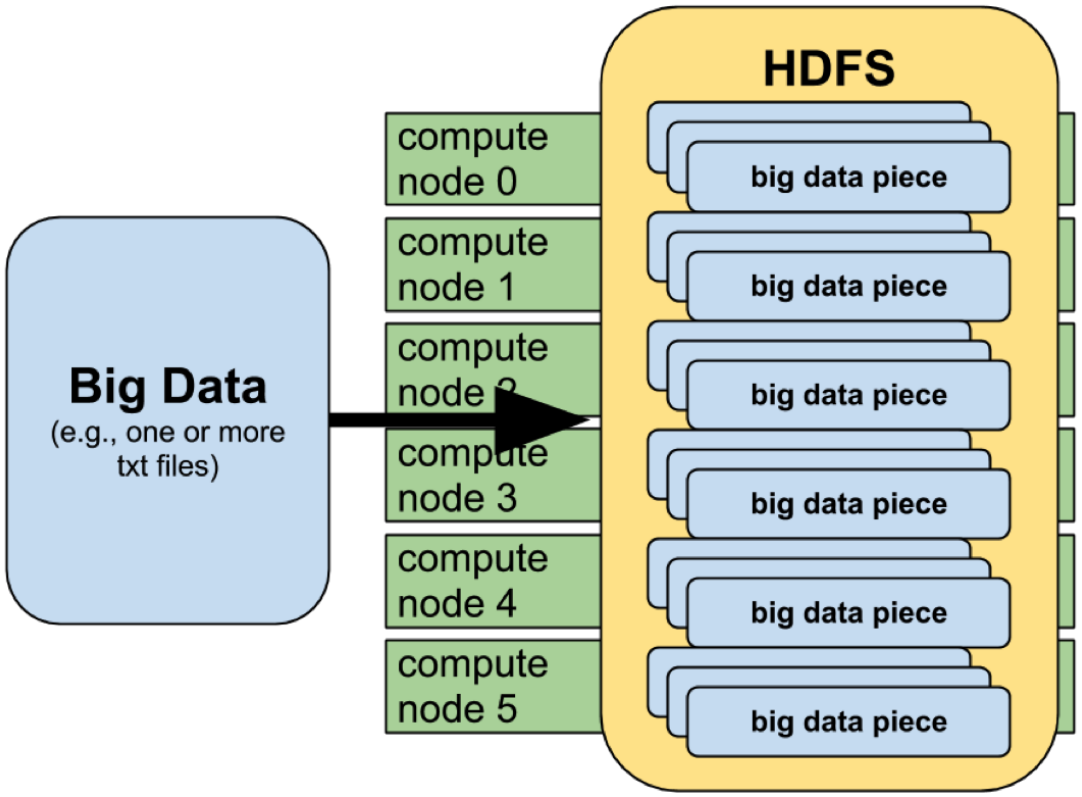
Hadoop Ecosystem

- Hive: Analysis of large data using simple query language similar to SQL.
- Pig: Data flow language. Analyses large data sets.
- Sqoop: Transfers bulk data between Hadoop and structured data stores.
- Mahout: Scalable machine learning and data mining library.
- Oozie: Workflow scheduler system to manage jobs.
- Zookeeper: Coordination service for distributed applications.

HDFS

- Distributed file system
- Inherently fault tolerant.
- Highly scalable.

HDFS Architecture
Ref: Google images



MapReduce

- Simplified programming model for processing large number of data sets in parallel.
- Efficiently utilizes the cloud resources, accelerates processing of large amount data on cloud.

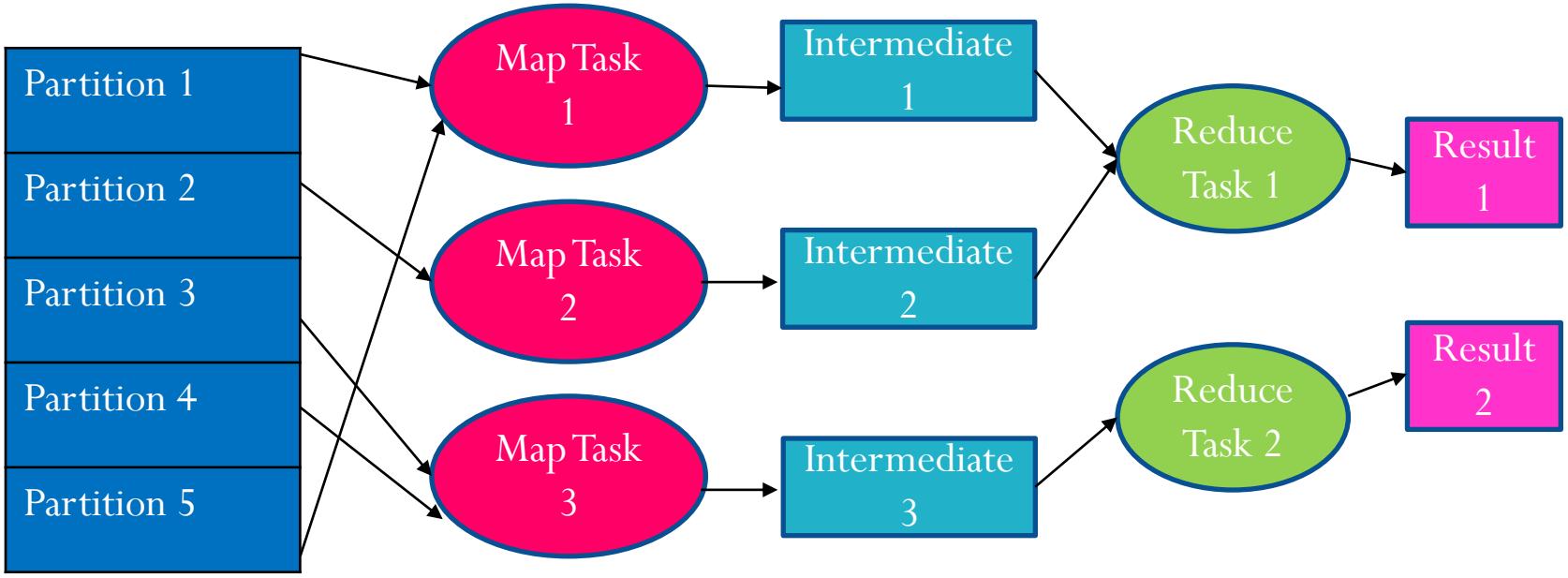
Drawbacks:

- MapReduce is not a database ,lacks features like indexing, security, query optimization etc.
- Assumes every job is an entity, no knowledge of parallely running jobs.
- New Concept, Very few people know about the configuration, coding and usage of MapReduce.
- Continuously evolving.

Logical Flow of Data in MapReduce

- Input is provided in form of key-value pair.
- Input is divided into small pieces, master and slaves are created
- Master node usually work where data is present , slave nodes work remotely on data.
- Map operation is performed in parallel on all data pieces

Working of MapReduce



Working of MapReduce

- MapReduce worker receives data from the master, processes it, and sends back the generated result to the master.
- MapReduce workers work on the same code on received data, no communication between the co-workers.
- Master receives result from each worker, integrates the result, processes integrated result, and generates final output.

Technology Expertise

- Good database knowledge such as RDBMS.
- Good NoSQL database knowledge such as MangoDB, Cassandra, HBase, etc.
- Programming language such as Java, Python, C++,etc.
- Open source tool such as Hadoop.
- Data warehousing.
- Data mining.
- Visualization such as Tableau, Flare, Google visualization APIs, etc.

Subject Expertise

- Mathematics.
- Statistics.
- Artificial Intelligence(AI).
- Algorithms.
- Machine learning.
- Pattern recognition.
- Natural Language Processing.

Big Data requires a broad set of Skills:

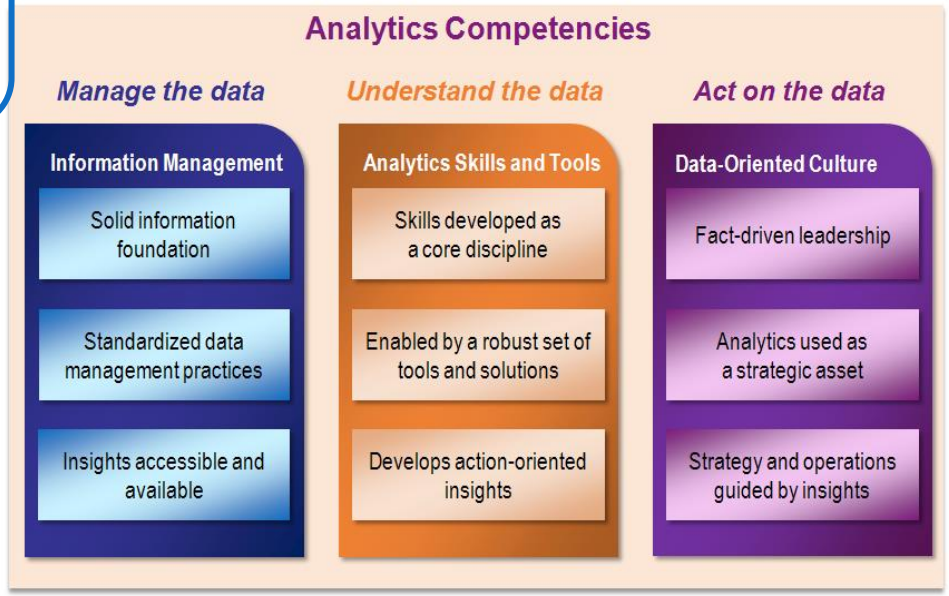
"By 2015, big data demand will reach 4.4 million jobs globally, but only one-third of those jobs will be filled.“ *Source: Gartner "Gartner's Top Predictions*

Math and Operations Research Expertise

Develop analytic algorithms

Data Experts

Data architecture, management, governance, policy



Decision Making Executive and Management

Apply information to solve business issues

Tool Developers

Mask complexity and analytics to lower skills boundaries

Industry Vertical Domain Expertise

Develop hypothesis, identify relevant business issues,

ask the right questions

Visualization Expertise

Interpret data sets, determine correlations and present in meaningful ways


Thank
You


Any
Questions

References

- Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information Sciences* 275 (2014): 314-347.
- Hashem, Ibrahim Abaker Targio, et al. "The rise of "big data" on cloud computing: review and open research issues." *Information Systems* 47 (2015): 98-115.
- DT Editorial Services, "Big Data Black Book",Dreamtech press,2015.
- Seema Acharya,Subhasini Chellappan,"Big Data and Analytics",Wiley,2015.
- Wu, Xindong, et al. "Data mining with big data." *Knowledge and Data Engineering, IEEE Transactions on* 26.1 (2014): 97-107.
- "Apache Hadoop," Jun. 2011. [Online]. Available: http://hadoop.apache.org
- Kambatla, Karthik, et al. "Trends in big data analytics." *Journal of Parallel and Distributed Computing* 74.7 (2014): 2561-2573.