INDEX

Subject Code: BCDS601 Big data Analytics

ST1: Syllabus

UNIT1:

Introduction to Big Data: Types of digital data, history of Big Data innovation, introduction to Big Data platform, drivers for Big Data, Big Data architecture and characteristics, 5 Vs of Big Data, Big Data technology components, Big Data importance and applications. Big Data features – security, compliance, auditing and protection, Big Data privacy and ethics, Big Data Analytics, Challenges of conventional systems, intelligent data analysis, nature of data, analytic processes and tools, analysis vs reporting, modern data analytic tools.

UNIT2:

Hadoop: History of Hadoop, Apache Hadoop, the Hadoop Distributed File System, components of Hadoop, data format, analyzing data with Hadoop, scaling out, Hadoop streaming, Hadoop pipes, Hadoop Echo System.

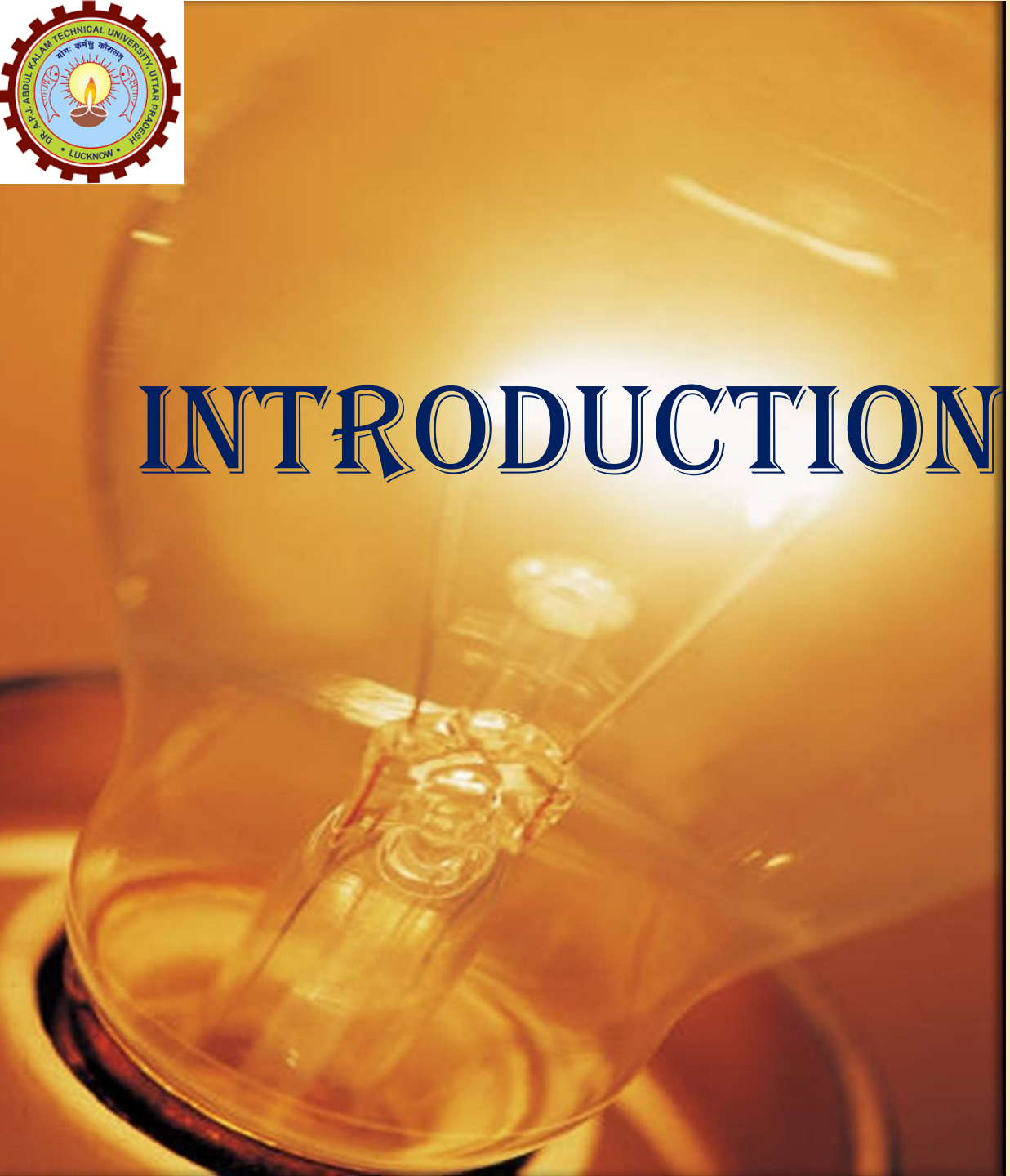
BCDS 601	BIG DATA AND ANALYTICS	
Course Outcome (CO)		**Bloom's Knowledge Level (KL)**
At the end of course , the student will be able to		
CO 1	Demonstrate knowledge of Big Data Analytics concepts and its applications in business.	K_1,K_2
CO 2	Demonstrate functions and components of Map Reduce Framework and HDFS.	K_1,K_2
CO 3	Discuss Data Management concepts in NoSQL environment.	K_6
CO 4	Explain process of developing Map Reduce based distributed processing applications.	K_2,K_5
CO 5	Explain process of developing applications using HBASE, Hive, Pig etc.	K_2,K_5

DETAILED SYLLABUS		**3-0-0**
Unit	**Topic**	**Proposed Lectures**
I	**Introduction to Big Data**: Types of digital data, history of Big Data innovation, introduction to Big Data platform, drivers for Big Data, Big Data architecture and characteristics, 5 Vs of Big Data, Big Data technology components, Big Data importance and applications. Big Data features – security, compliance, auditing and protection, Big Data privacy and ethics, Big Data Analytics, Challenges of conventional systems, intelligent data analysis, nature of data, analytic processes and tools, analysis vs reporting, modern data analytic tools.	**06**
II	**Hadoop:** History of Hadoop, Apache Hadoop, the Hadoop Distributed File System, components of Hadoop, data format, analyzing data with Hadoop, scaling out, Hadoop streaming, Hadoop pipes, Hadoop Echo System. **Map Reduce:** Map Reduce framework and basics, how Map Reduce works, developing a Map Reduce application, unit tests with MR unit, test data and local tests, anatomy of a Map Reduce job run, failures, job scheduling, shuffle and sort, task execution, Map Reduce types, input formats, output formats, Map Reduce features, Real-world Map Reduce	**08**
III	**HDFS (Hadoop Distributed File System):** Design of HDFS, HDFS concepts, benefits and challenges, file sizes, block sizes and block abstraction in HDFS, data replication, how does HDFS store, read, and write files, Java interfaces to HDFS, command line interface. Hadoop file system interfaces, data flow, data ingest with Flume and Scoop, Hadoop archives, Hadoop I/O: compression, serialization, Avro and file-based data structures. **Hadoop Environment:** Setting up a Hadoop cluster, cluster specification, cluster setup and installation, Hadoop configuration, security in Hadoop, administering Hadoop, HDFS monitoring & maintenance, Hadoop benchmarks, Hadoop in the cloud	**08**
IV	**Hadoop Eco System and YARN**: Hadoop ecosystem components, schedulers, fair and capacity, Hadoop 2.0 New Features - NameNode high availability, HDFS federation, MRv2, YARN, Running MRv1 in YARN. **NoSQL Databases:** Introduction to NoSQL **MongoDB:** Introduction, data types, creating, updating and deleing documents, querying, introduction to indexing, capped collections **Spark:** Installing spark, spark applications, jobs, stages and tasks, Resilient Distributed Databases, anatomy of a Spark job run, Spark on YARN **SCALA:** Introduction, classes and objects, basic types and operators, built-in control structures, functions and closures, inheritance.	**09**
V	**Hadoop Eco System Frameworks**: Applications on Big Data using Pig, Hive and HBase **Pig** - Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases, Grunt, Pig Latin, User Defined Functions, Data Processing operators,	**09**

	Hive - Apache Hive architecture and installation, Hive shell, Hive services, Hive metastore, comparison with traditional databases, HiveQL, tables, querying data and user-defined functions, sorting and aggregating, Map Reduce scripts, joins & subqueries. **HBase** – Hbase concepts, clients, example, Hbase vs RDBMS, advanced usage, schema design, advance indexing, Zookeeper – how it helps in monitoring a cluster, how to build applications with Zookeeper. IBM Big Data strategy, introduction to Infosphere, BigInsights and Big Sheets, introduction to Big SQL.	3

Text books and References:

1. Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley
2. Big-Data Black Book, DT Editorial Services, Wiley
3. Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch, “Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill.
4. Thomas Erl, Wajid Khattak, Paul Buhler, “Big Data Fundamentals: Concepts, Drivers and Techniques”, Prentice Hall.
5. Raj Kamal, Preeti Saxena, “Big Data Analytics”, McGraw Hill Education
6. Bart Baesens “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications (WILEY Big Data Series)”, John Wiley & Sons
7. ArshdeepBahga, Vijay Madisetti, “Big Data Science & Analytics: A HandsOn Approach “, VPT
8. Anil Maheshwari, “Big Data”, Second Edition, McGraw Hill
9. Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CUP
10. Tom White, "Hadoop: The Definitive Guide", O'Reilly.
11. Eric Sammer, "Hadoop Operations", O'Reilly.
12. Chuck Lam, “Hadoop in Action”, MANNING Publishers
13. Deepak Vohra, “Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools”, Apress
14. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilly
15. Lars George, "HBase: The Definitive Guide", O'Reilly.
16. Alan Gates, "Programming Pig", O'Reilly.
17. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer
18. Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley & sons
19. Glenn J. Myatt, “Making Sense of Data”, John Wiley & Sons
20. Pete Warden, “Big Data Glossary”, O’Reilly



INTRODUCTION TO BIG DATA

Basic: Data

- The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

- Data can be defined as figures or facts that can be stored in or can be used by a computer.

Conti…

Data Measurement	Size
Bit	Single Binary Digit (1 or 0)
Byte	8 bits
Kilobyte (KB)	1,024 Bytes
Megabyte (MB)	1,024 Kilobytes
Gigabyte (GB)	1,024 Megabytes
Terabyte (TB)	1,024 Gigabytes
Petabyte (PB)	1,024 Terabytes
Exabyte (EB)	1,024 Petabytes

Basic: Big Data

- Data which are very large in size is called Big Data.
- Normally we work on data of size MB (Word Doc, Excel) or maximum GB (Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data.
- It is stated that almost 90% of today's data has been generated in the past 3 years.

Conti…

- Big Data is a **collection of data** that is **huge in volume**, yet growing exponentially with time.
- It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.
- "Big data" is ***high-volume***, ***velocity***, and ***variety*** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Example of Big Data

- **New York Stock Exchange** is an example of Big Data that generates about ***one terabyte*** of new trade data per day.
- The statistic shows that ***500+terabytes*** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.
- A single **Jet engine** can generate ***10+terabytes*** of data in ***30 minutes*** of flight time. With many thousand flights per day, generation of data reaches up to many ***Petabytes.***

Sources of Big Data

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

Advantages of Big Data

- Big Data analytics tools can predict outcomes accurately, thereby, allowing businesses and organizations to ***make better decisions***, while simultaneously optimizing their operational efficiencies and reducing risks.

- Big Data provides insights into the customer pain points and allows companies to improve upon their products and services.

Conti…

- Big Data analytics could help companies generate more sales leads which would naturally mean a ***boost in revenue***.

- Big Data insights allow you to learn customer behavior to understand the customer trends and provide a highly 'personalized' experience to them.

Applications of Big Data

Healthcare

- With the help of predictive analytics, medical professionals are able to provide personalized healthcare services to individual patients. Apart from that, fitness wearables, telemedicine, remote monitoring – all powered by Big Data and AI – are helping change lives for the better.

Academia

- Education is no more limited to the physical bounds of the classroom – there are numerous online educational courses to learn from. Academic institutions are investing in digital courses powered by Big Data technologies to aid the all-round development of budding learners.

Conti…

Banking

- The banking sector relies on Big Data for fraud detection. Big Data tools can efficiently detect fraudulent acts in real-time such as misuse of credit/debit cards, archival of inspection tracks, faulty alteration in customer stats, etc.

Manufacturing

- According to TCS Global Trend Study, the most significant benefit of Big Data in manufacturing is improving the supply strategies and product quality. In the manufacturing sector, Big data helps create a transparent infrastructure, thereby, predicting uncertainties and in competencies that can affect the business adversely.

Conti…

IT

- One of the largest users of Big Data, IT companies around the world are using Big Data to optimize their functioning, enhance employee productivity, and minimize risks in business operations. By combining Big Data technologies with ML and AI, the IT sector is continually powering innovation to find solutions even for the most complex of problems.

Transportation

- Big Data Analytics holds immense value for the transportation industry. In countries across the world, both private and government-run transportation companies use Big Data technologies to optimize route planning, control traffic, manage road congestion, and improve services.

Conti…

Retail

- Big Data has changed the way of working in traditional brick and mortar retail stores. Over the years, retailers have collected vast amounts of data from local demographic surveys, POS scanners, RFID, customer loyalty cards, store inventory, and so on. Now, they've started to leverage this data to create personalized customer experiences, boost sales, increase revenue, and deliver outstanding customer service.
- Retailers are even using smart sensors and Wi-Fi to track the movement of customers, the most frequented aisles, for how long customers linger in the aisles, among other things.

Big Data Analysis Tools and Software

- The tools that are used to store and analyze a large number of data sets and processing these complex data are known as big data tools.

- Xplenty
- Atlas.ti
- Analytics
- Microsoft HDInsight
- Talend
- R-Programming

Xplenty

- A cloud-based ETL solution providing simple visualized data pipelines for automated data flows across a wide range of sources and destinations. Xplenty's powerful on-platform transformation tools allow you to clean, normalize, and transform data while also adhering to compliance best practices.

Features:

- Powerful, code-free, on-platform data transformation offering
- Rest API connector – pull in data from any source that has a Rest API
- Destination flexibility – send data to databases, data warehouses, and Salesforce
- Security focused – field-level data encryption and masking to meet compliance requirements
- Rest API – achieve anything possible on the Xplenty UI via the Xplenty API
- Customer-centric company that leads with first-class support

Atlas.ti

- All-in-one research software. This big data analytic tool gives you all-in-one access to the entire range of platforms. You can use it for qualitative data analysis and mixed methods research in academic, market, and user experience research.

Features:

- You can export information on each source of data.
- It offers an integrated way of working with your data.
- Allows you to rename a Code in the Margin Area
- Helps you to handle projects that contain thousands of documents and coded data segments.
- Supported platforms: Mac, Windows, Web, Mobile App

Analytics

- Tool that provides visual analysis and dash boarding. It allows you to connect multiple data sources, including business applications, databases, cloud drives, and more.

Features:

- Offers visual analysis and dash boarding.
- It helps you to analyze data in depth.
- Provides collaborative review and analysis.
- You can embed reports to websites, applications, blogs, and more.

Azure HDInsight

- Spark and Hadoop service in the cloud. It provides big data cloud offerings in two categories, Standard and Premium. It provides an enterprise-scale cluster for the organization to run their big data workloads.

Features:

- Reliable analytics with an industry-leading SLA
- It offers enterprise-grade security and monitoring
- Protect data assets and extend on-premises security and governance controls to the cloud
- High-productivity platform for developers and scientists
- Integration with leading productivity applications
- Deploy Hadoop in the cloud without purchasing new hardware or paying other up-front costs

Talend

- Big data analytics software that simplifies and automates big data integration. Its graphical wizard generates native code. It also allows big data integration, master data management and checks data quality.

Features:

- Accelerate time to value for big data projects
- Simplify ETL & ELT for big data
- Talend Big Data Platform simplifies using MapReduce and Spark by generating native code
- Smarter data quality with machine learning and natural language processing
- Agile DevOps to speed up big data projects
- Streamline all the DevOps processes

R-Programming

- Language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

- Effective data handling and storage facility,
- It provides a suite of operators for calculations on arrays, in particular, matrices,
- It provides coherent, integrated collection of big data tools for data analysis
- It provides graphical facilities for data analysis which display either on-screen or on hardcopy

Others

- **Apache Hadoop:** A framework that allows you to store big data in a distributed environment for parallel processing.
- **Apache Pig:** A Platform that is used for analyzing large datasets by representing them as data flows. Pig is designed to provide an abstraction over MapReduce which reduces the complexities of writing a MapReduce program.
- **Apache Hbase:** A multidimensional, distributed, open-source, and NoSQL database written in Java. It runs on top of HDFS providing Bigtable-like capabilities for Hadoop.
- **Apache Spark:** Open-source general-purpose cluster-computing framework. It provides an interface for programming all clusters with implicit data parallelism and fault tolerance.

Big Data Case studies

- Walmart leverages Big Data and *Data Mining* to create personalized product recommendations for its customers. With the help of these two emerging technologies, Walmart can uncover valuable patterns showing the most frequently bought products, most popular products, and even the most popular product bundles (products that complement each other and are usually purchased together).

- Based on these insights, Walmart creates attractive and customized recommendations for individual users. By effectively implementing Data Mining techniques, the retail giant has successfully increased the conversion rates and improved its customer service substantially. Furthermore, Walmart uses *Hadoop* and NoSQL technologies to allow customers to access real-time data accumulated from disparate

Conti…

- Uber is one of the major cab service providers in the world. It leverages customer data to track and identify the most popular and most used services by the users. Once this data is collected, Uber uses data analytics to analyze the usage patterns of customers and determine which services should be given more emphasis and importance.

- Apart from this, Uber uses Big Data in another unique way. Uber closely studies the demand and supply of its services and changes the cab fares accordingly. It is the surge pricing mechanism that works something like this – suppose when you are in a hurry, and you have to book a cab from a crowded location, Uber will charge you double the normal amount!

Conti…

- Netflix is one of the most popular on-demand online video content streaming platform used by people around the world. Netflix is a major proponent of the recommendation engine. It collects customer data to understand the specific needs, preferences, and taste patterns of users. Then it uses this data to predict what individual users will like and create personalized content recommendation lists for them.

- Today, Netflix has become so vast that it is even creating unique content for users. Data is the secret ingredient that fuels both its recommendation engines and new content decisions. The most pivotal data points used by Netflix include titles that users watch, user ratings, genres preferred, and how often users stop the playback, to name a few. Hadoop, Hive, and Pig are the three core components of the data structure used by Netflix.

THANK YOU

TYPES OF BIG DATA

Basic: Big Data

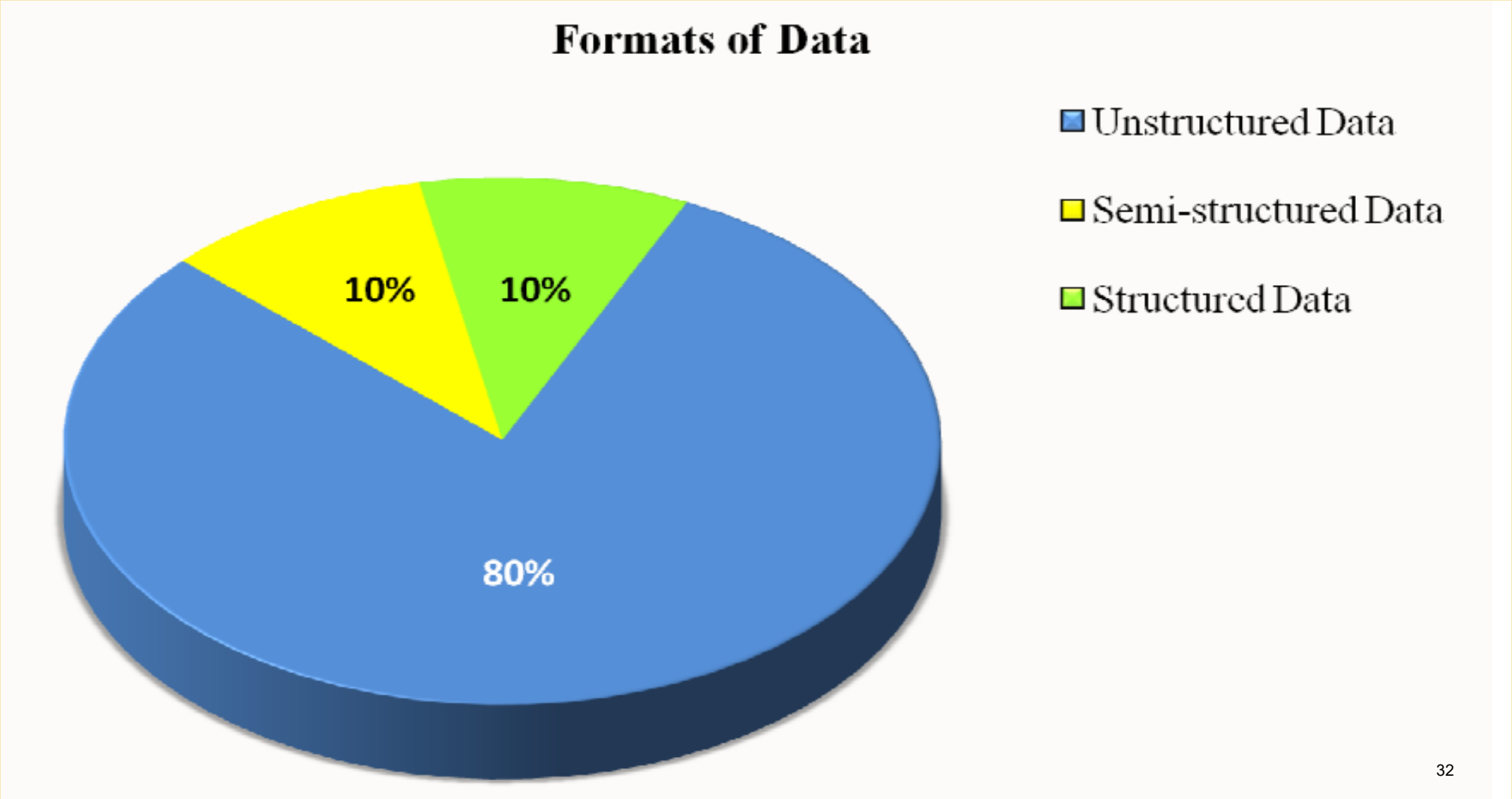
- Generally, Data which are very large in size is called Big Data.
- Normally we work on data of size MB (Word Doc, Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data.
- It is stated that almost 90% of today's data has been generated in the past 3 years.

Types of Big Data

- Following are the types of Big Data:
  - Structured
  - Unstructured
  - Semi-structured

- According to Merrill Lynch, 80–90% of business data is either unstructured or semi-structured.
- Gartner also estimates that unstructured data constitutes 80% of the whole enterprise data.

Conti…


Formats of Data
Unstructured Data
Semi-structured Data
Structured Data
10%
10%
80%

Types of Big Data: Structured

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

10^{21} bytes = 1 zettabyte

or ***one billion terabytes*** forms ***a zettabyte***.

- Data stored in a relational database management system is one example of a **'structured'** data.

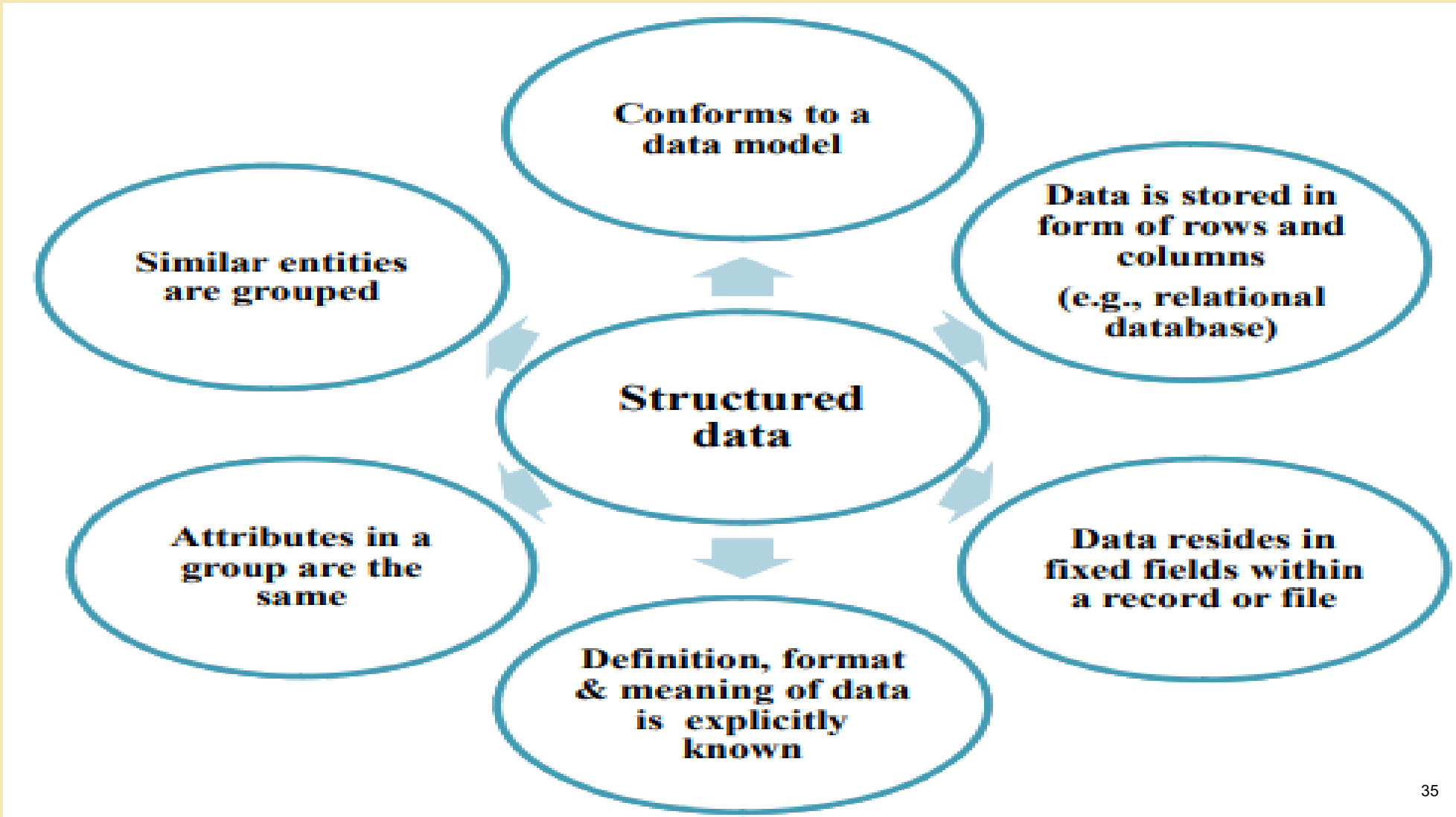
Conti…

- Structured is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.

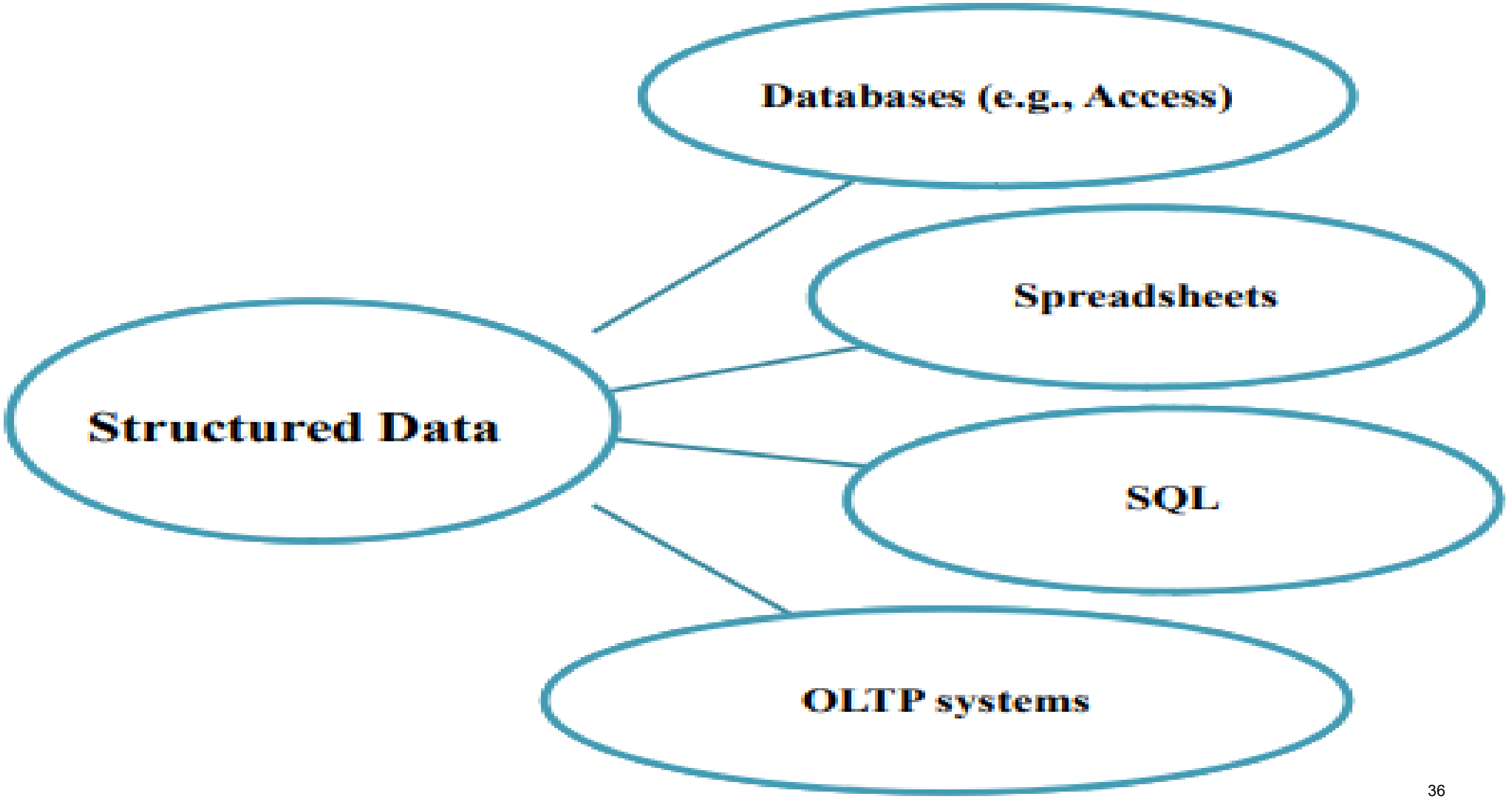
- Relationships exist between entities of data, such as classes and their objects.

- Data stored in databases is an example of structured data.

Conti…


Conforms to a data model
Similar entities are grouped
Data is stored in form of rows and columns
(e.g., relational database)
Structured data
Attributes in a group are the same
Data resides in fixed fields within a record or file
Definition, format & meaning of data is explicitly known

Structured Data Come from…



Structured V/s Semi-structured Data

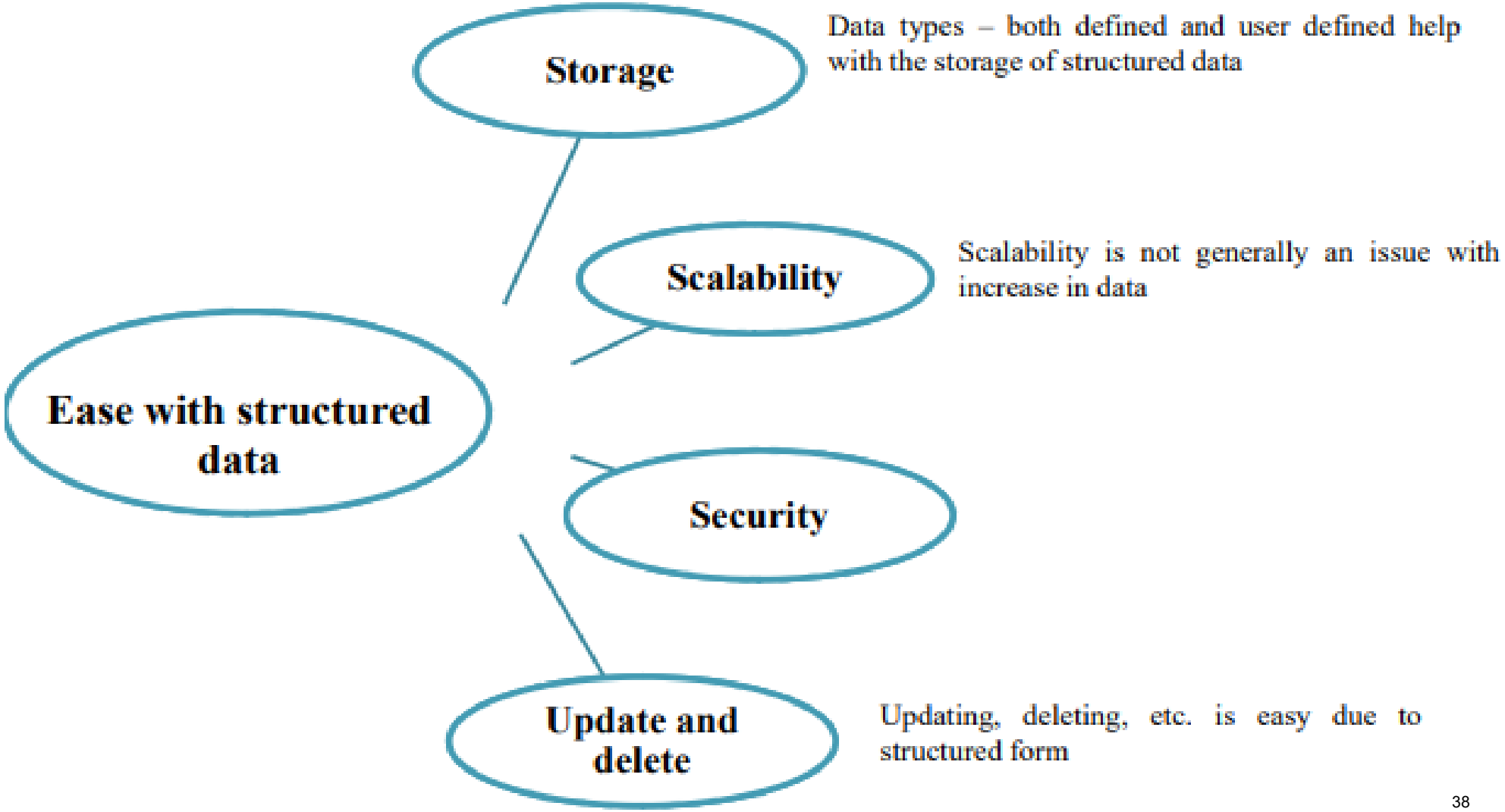
Semi-structured

Name	E-mail
Patrick Wood	ptw@dcs.abc.ac.uk, p.wood@ymail.uk
First name: Mark Last name: Taylor	MarkT@dcs.ymail.ac.uk
Alex Bourdoo	AlexBourdoo@dcs.ymail.ac.uk

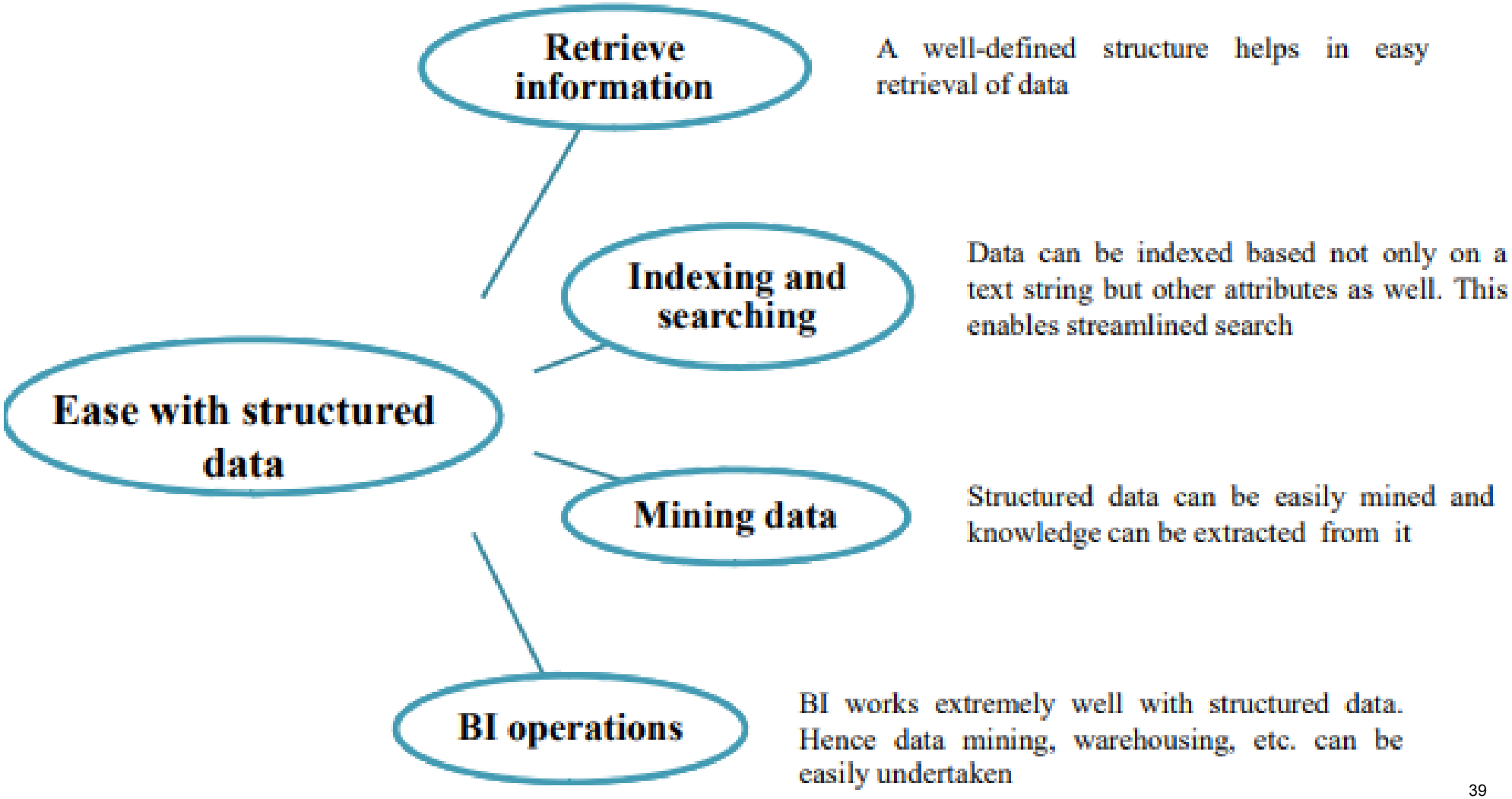
Structured

First Name	Last Name	E-mail Id	Alternate E-mail Id
Patrick	Wood	ptw@dcs.abc.ac.uk	p.wood@ymail.uk
Mark	Taylor	MarkT@dcs.ymail.ac.uk	
Alex	Bourdoo	AlexBourdoo@dcs.ymail.ac.uk	

Structured Data



Structured Data Retrieval



Structured Data Example

- An ‘Employee’ table in a database is an example of Structured Data.

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Types of Big Data: Unstructured

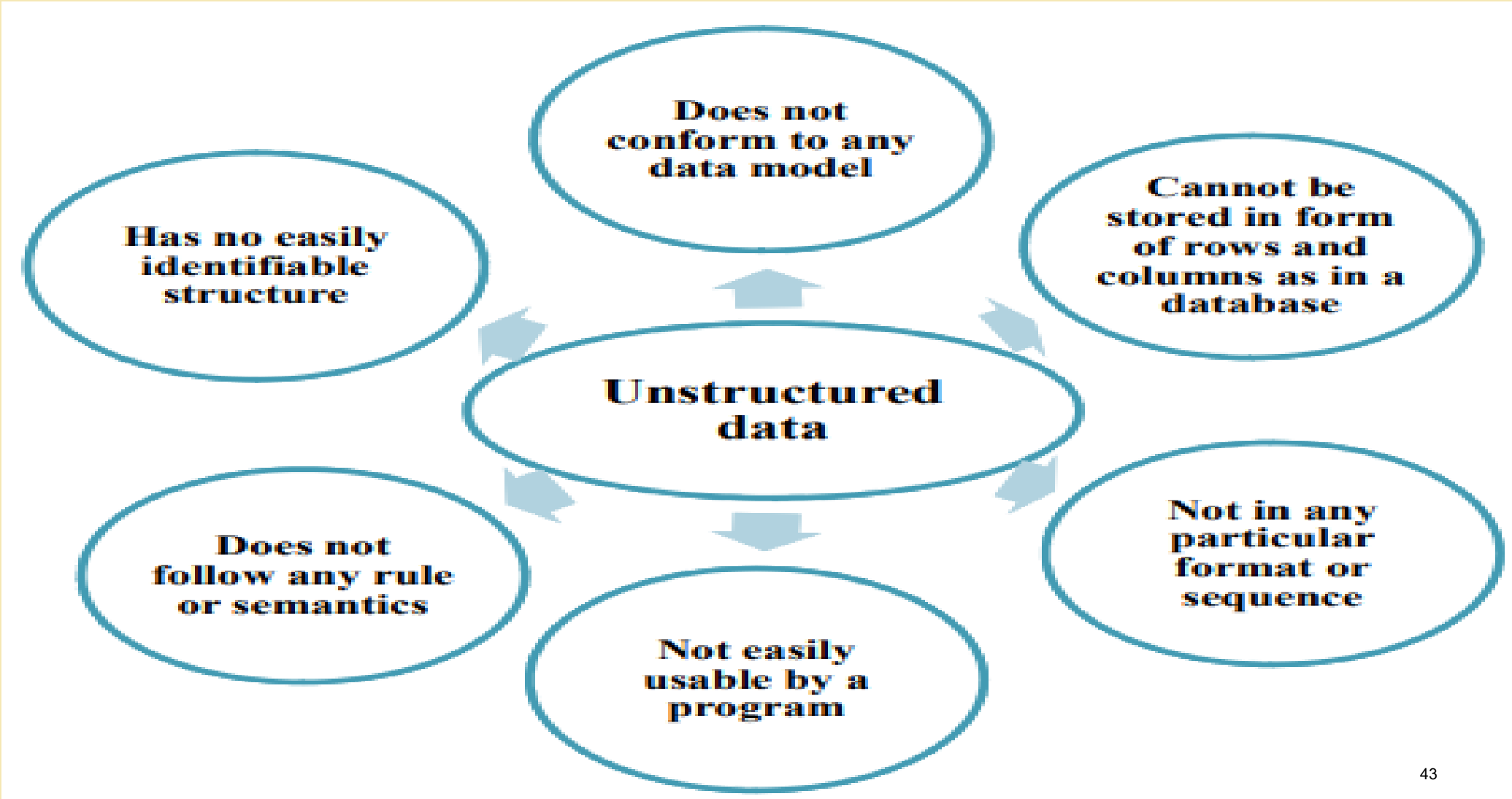
- Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.

Conti…

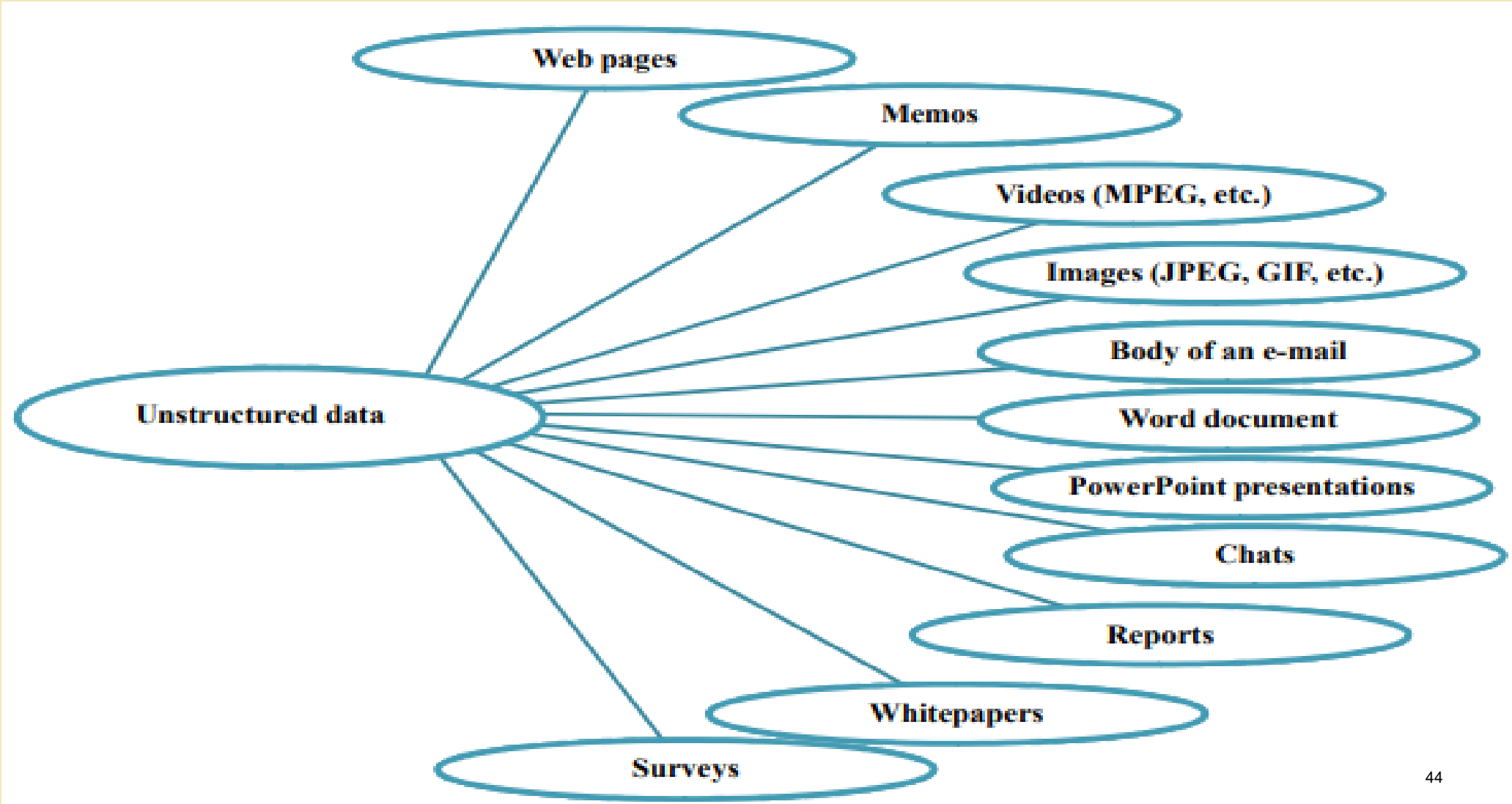
- Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80—90% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

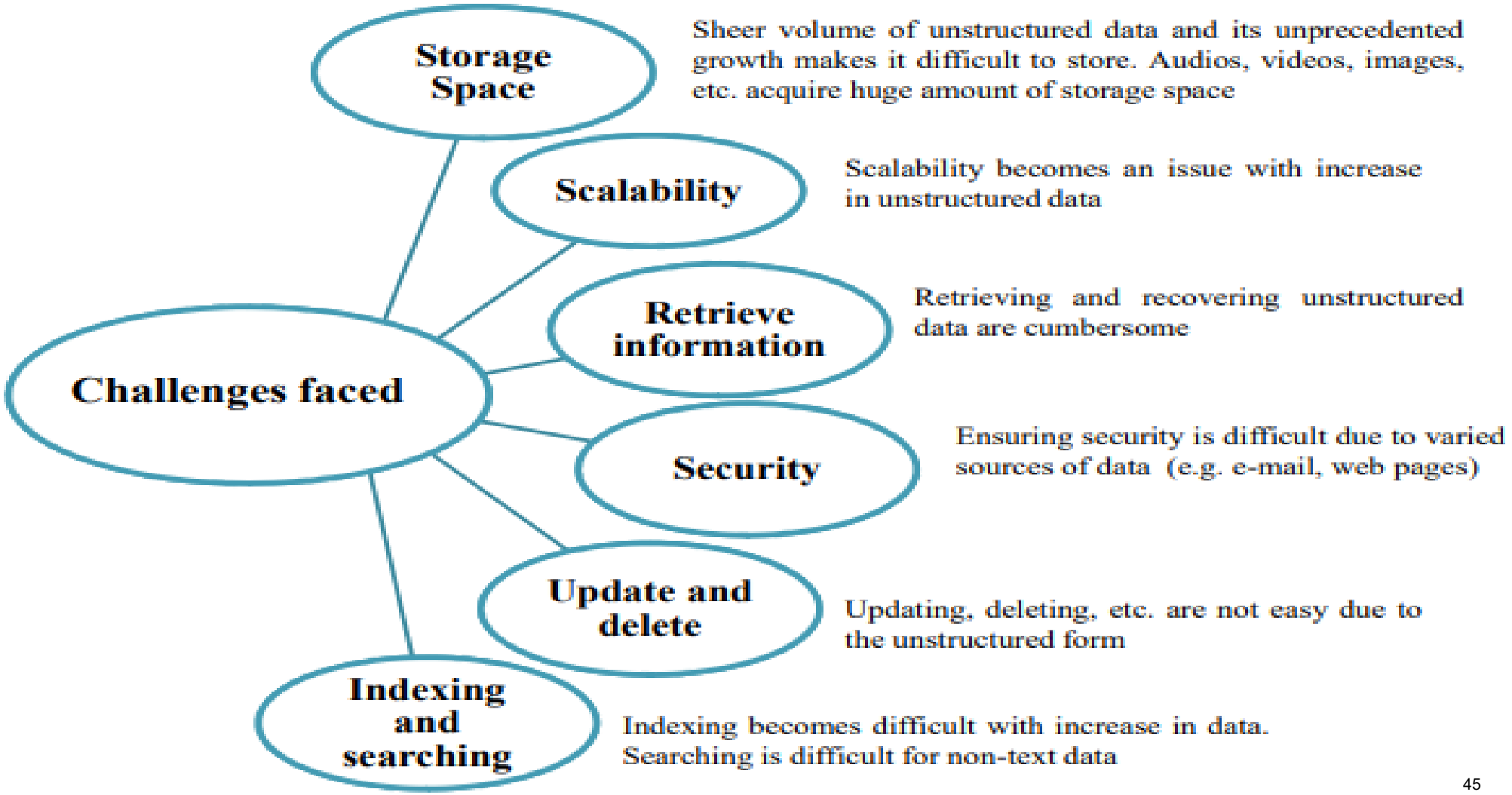
Unstructured data



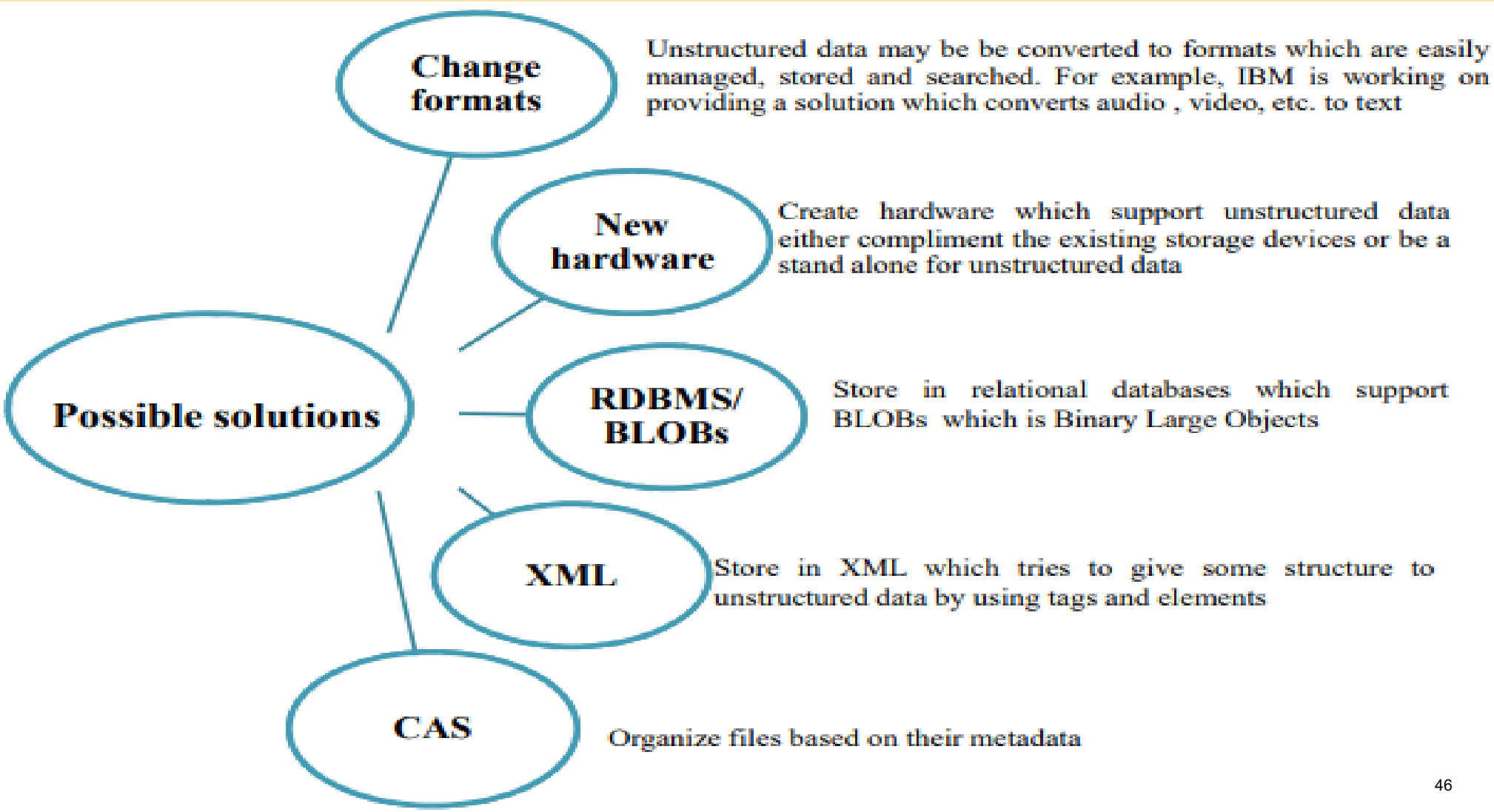
Unstructured data Come from…



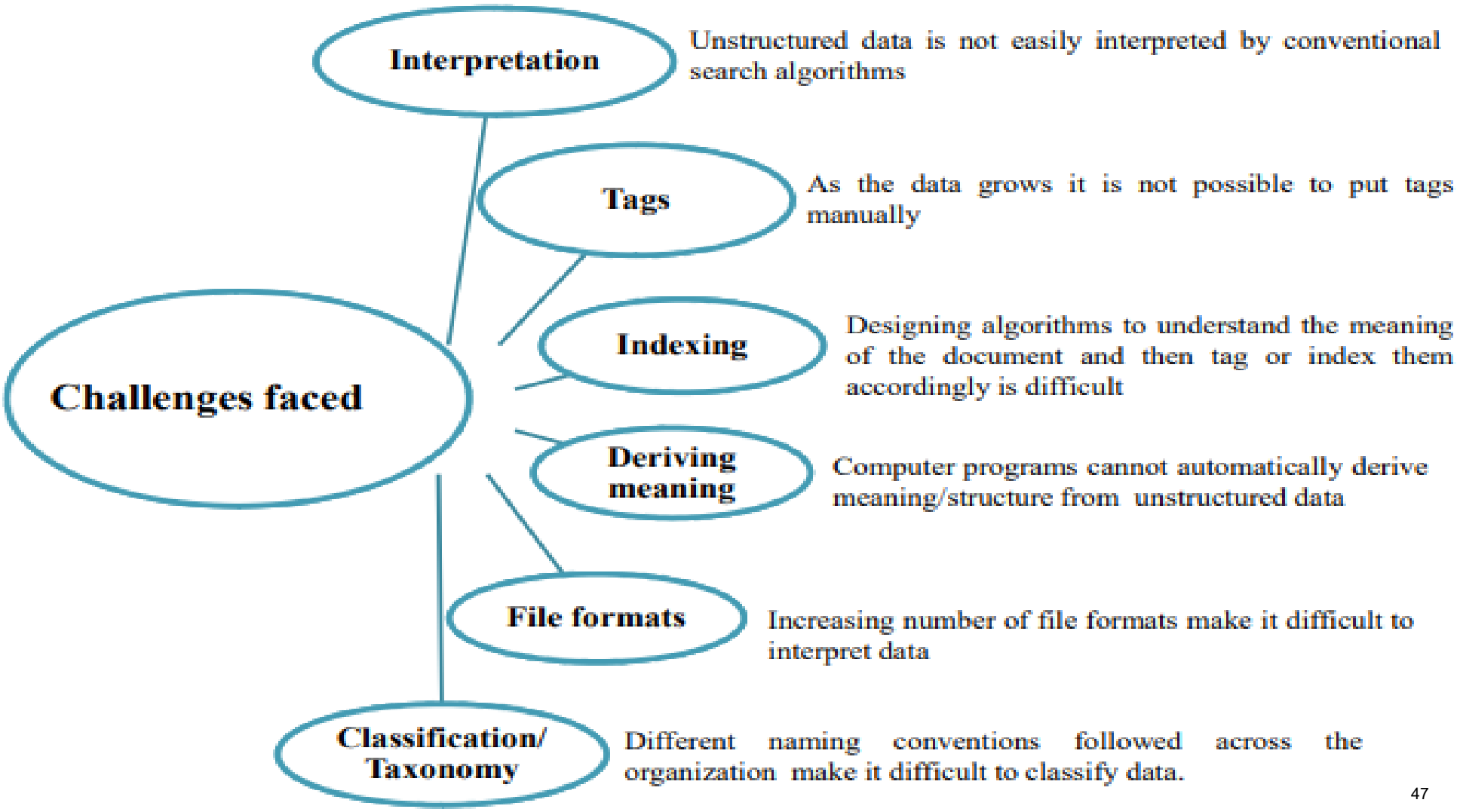
Store Unstructured data



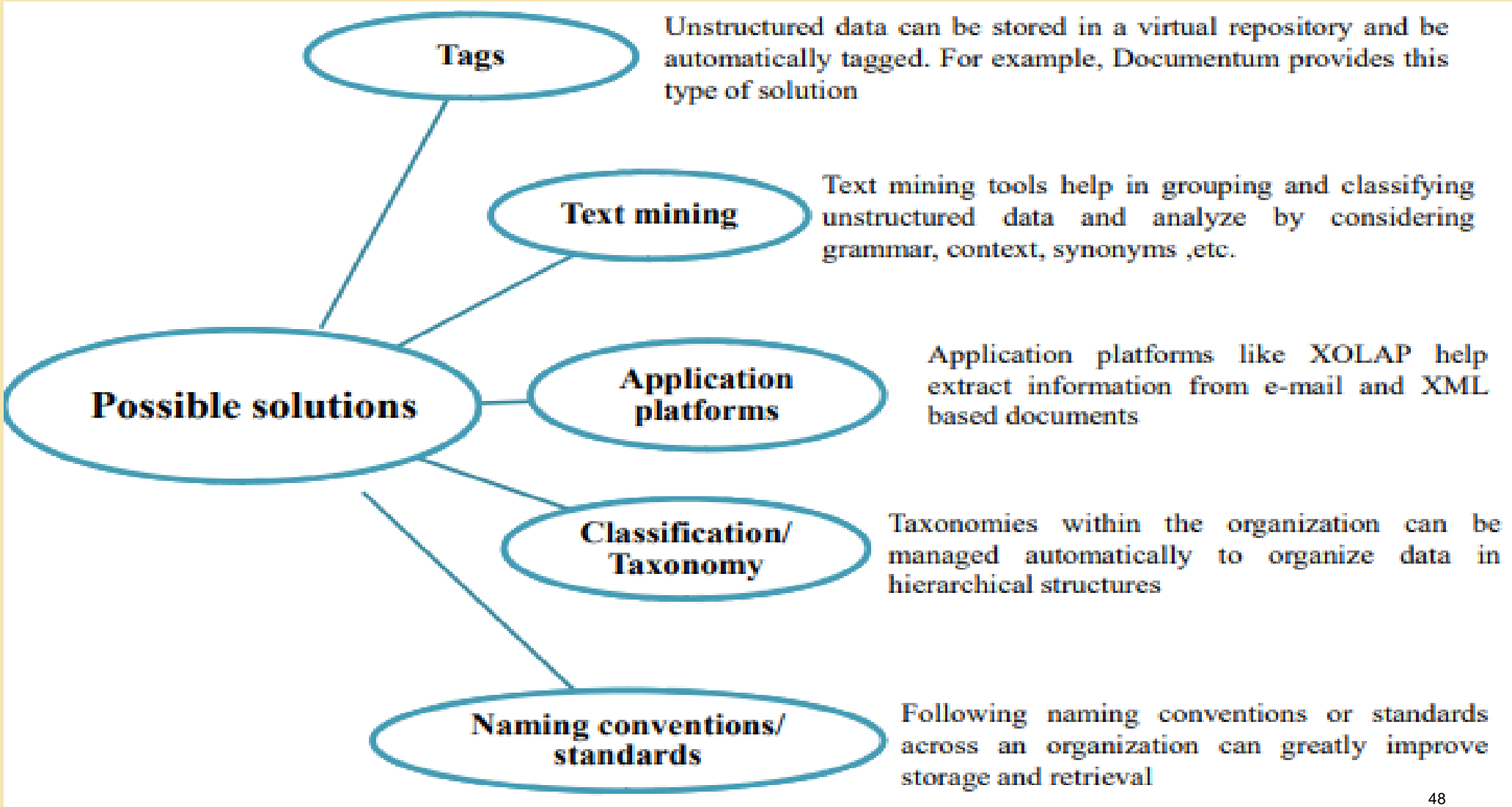
Conti…


Possible solutions
Change formats
Unstructured data may be be converted to formats which are easily managed, stored and searched. For example, IBM is working on providing a solution which converts audio , video, etc. to text
New hardware
Create hardware which support unstructured data either compliment the existing storage devices or be a stand alone for unstructured data
RDBMS/ BLOBs
Store in relational databases which support BLOBs which is Binary Large Objects
XML
Store in XML which tries to give some structure to unstructured data by using tags and elements
CAS
Organize files based on their metadata

Extract information from Unstructured data

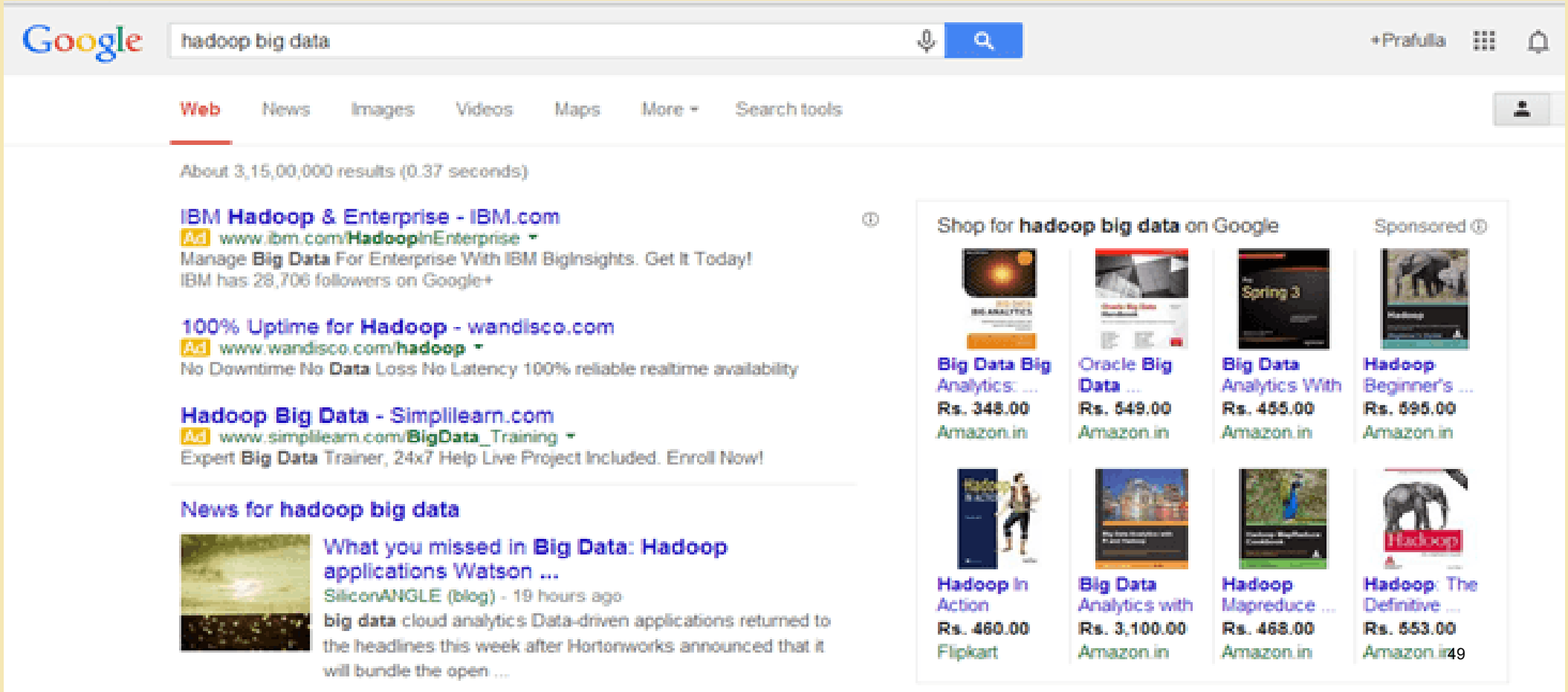


Conti…


Possible solutions
Tags
Unstructured data can be stored in a virtual repository and be automatically tagged. For example, Documentum provides this type of solution
Text mining
Text mining tools help in grouping and classifying unstructured data and analyze by considering grammar, context, synonyms ,etc.
Application platforms
Application platforms like XOLAP help extract information from e-mail and XML based documents
Classification/ Taxonomy
Taxonomies within the organization can be managed automatically to organize data in hierarchical structures
Naming conventions/ standards
Following naming conventions or standards across an organization can greatly improve storage and retrieval

Unstructured Data: Example

- The output returned by ‘Google Search’



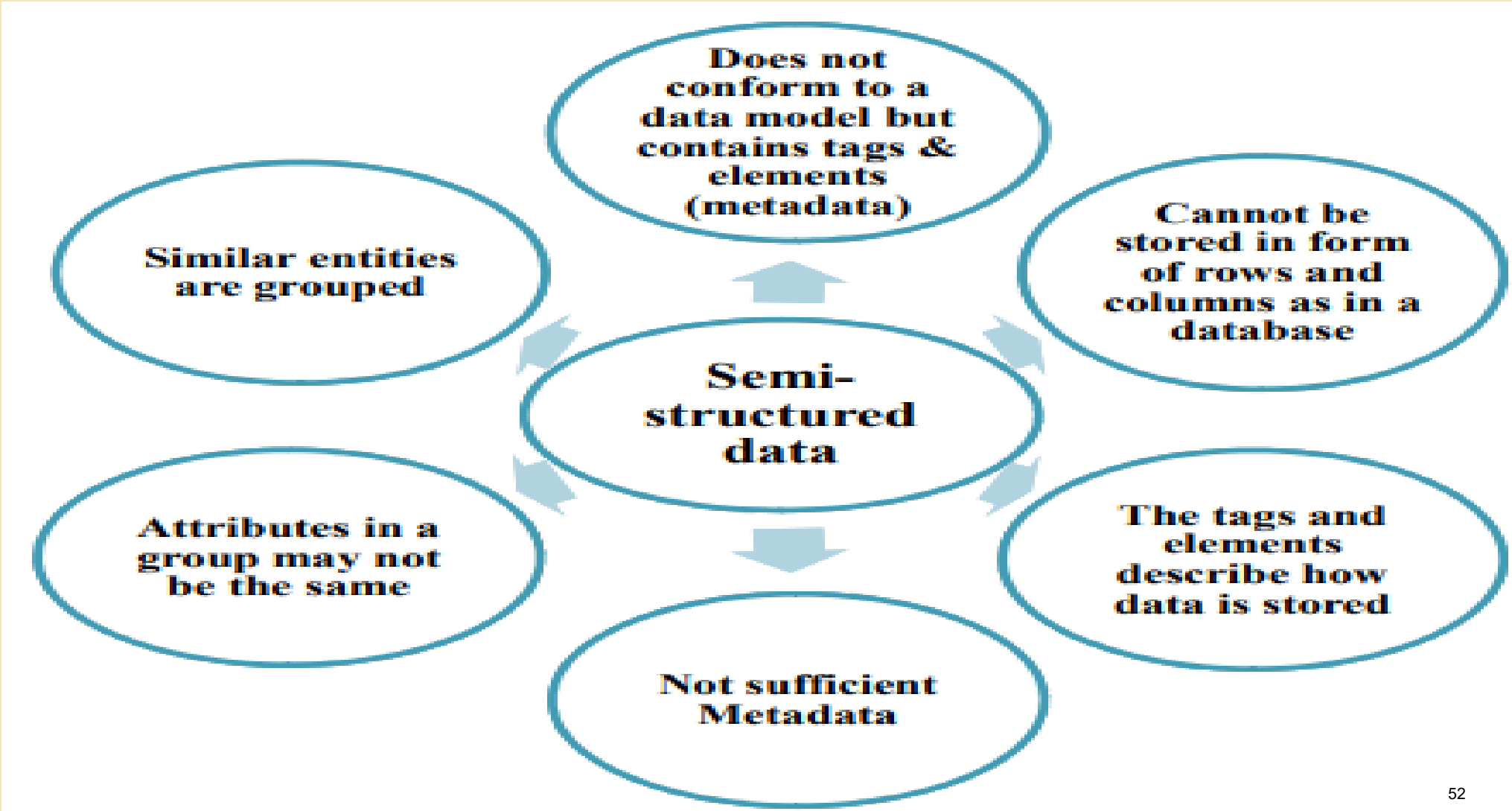
Types of Big Data: Semi-structured

- Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

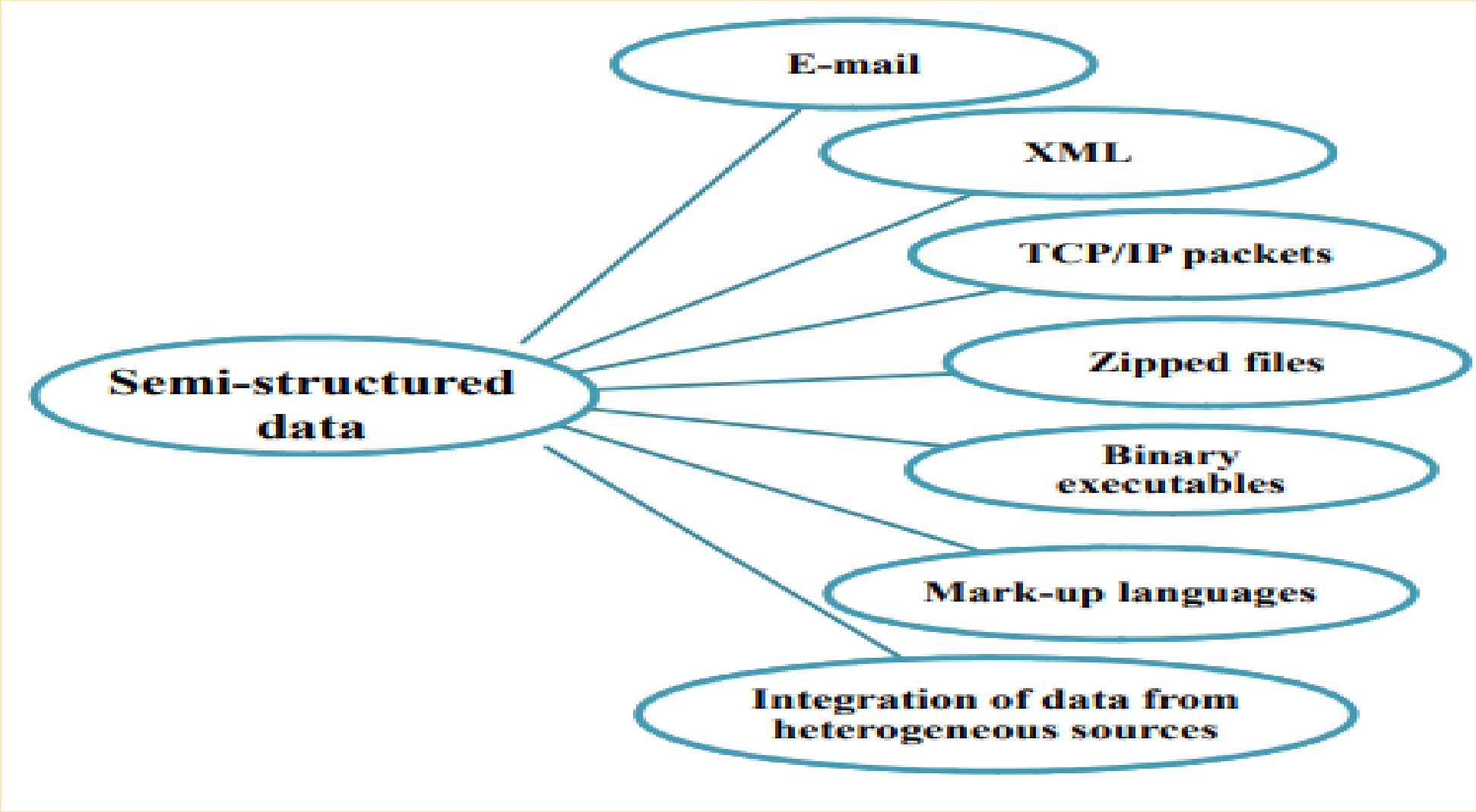
Conti…

- Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data.
- To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

Semi-structured data



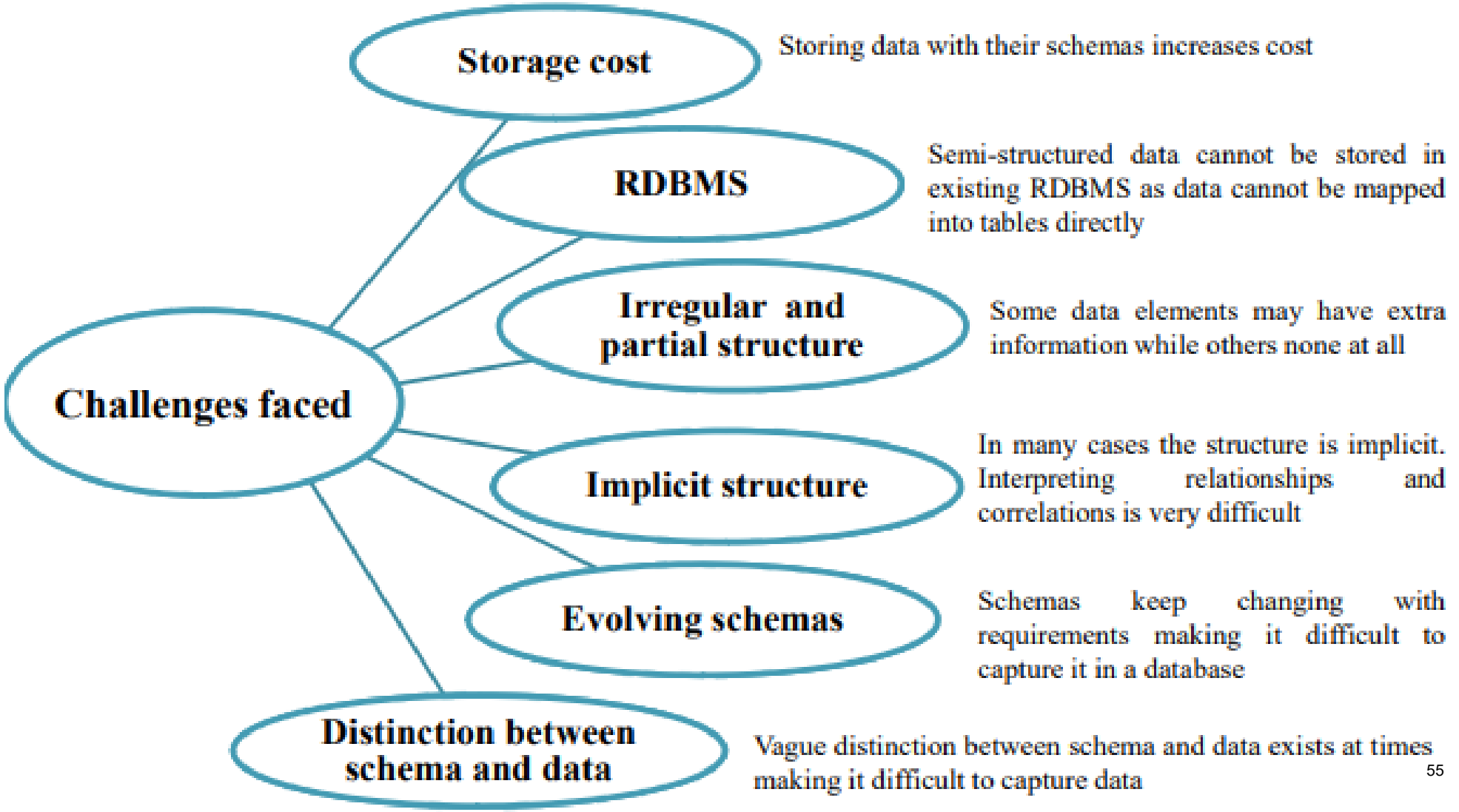
Semi-structured data come from…



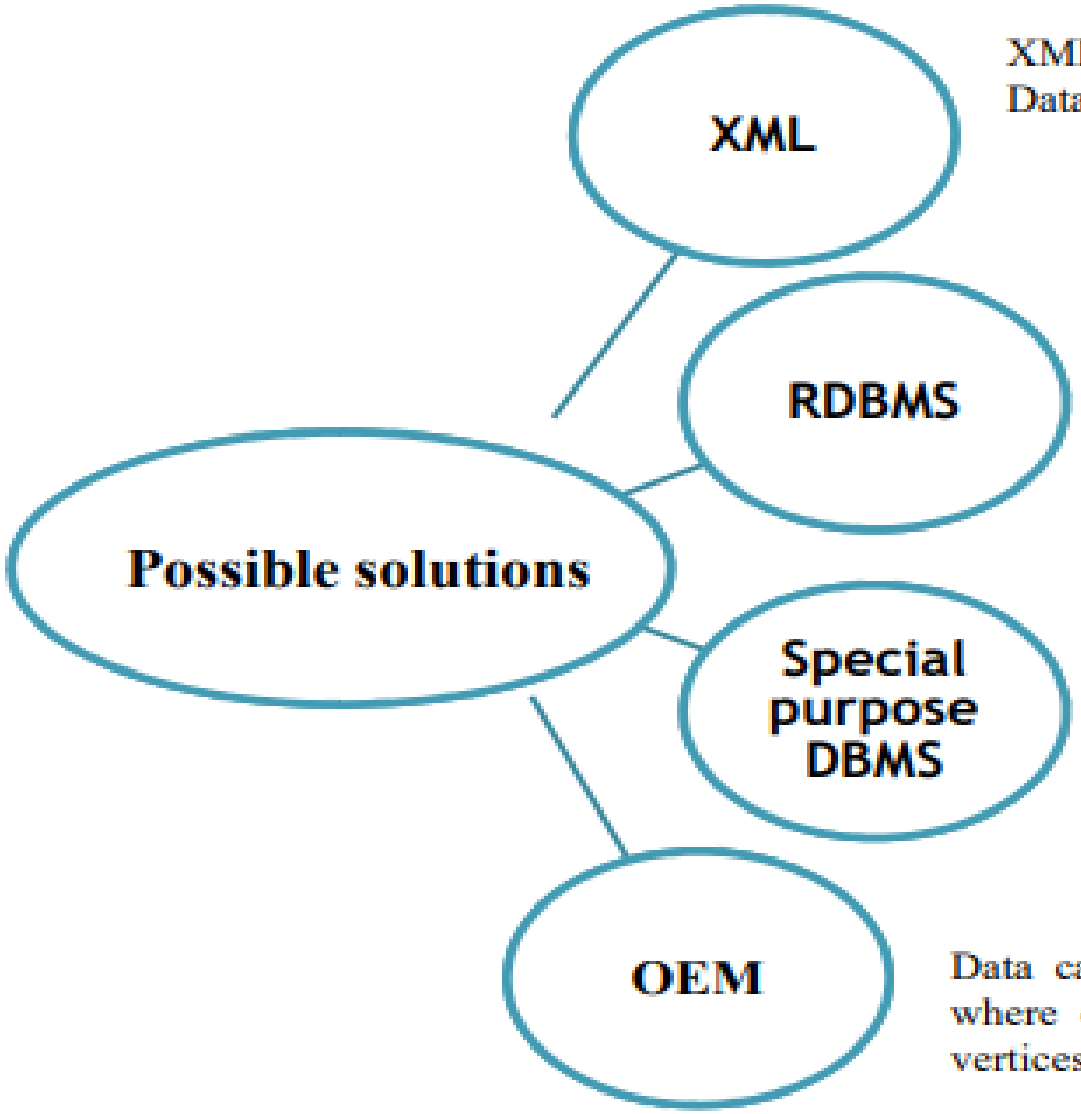
Manage Semi-structured Data

Schemas	Graph-based data models	XML
• Describe the structure and content of data to some extent • Assign meaning to data hence allowing automatic search and indexing	• Contain data on the leaves of the graph. Also known as 'schema less' • Used for data exchange among heterogeneous sources	• Models the data using tags and elements • Schemas are not tightly coupled to data

Store Semi-structured Data



Conti…



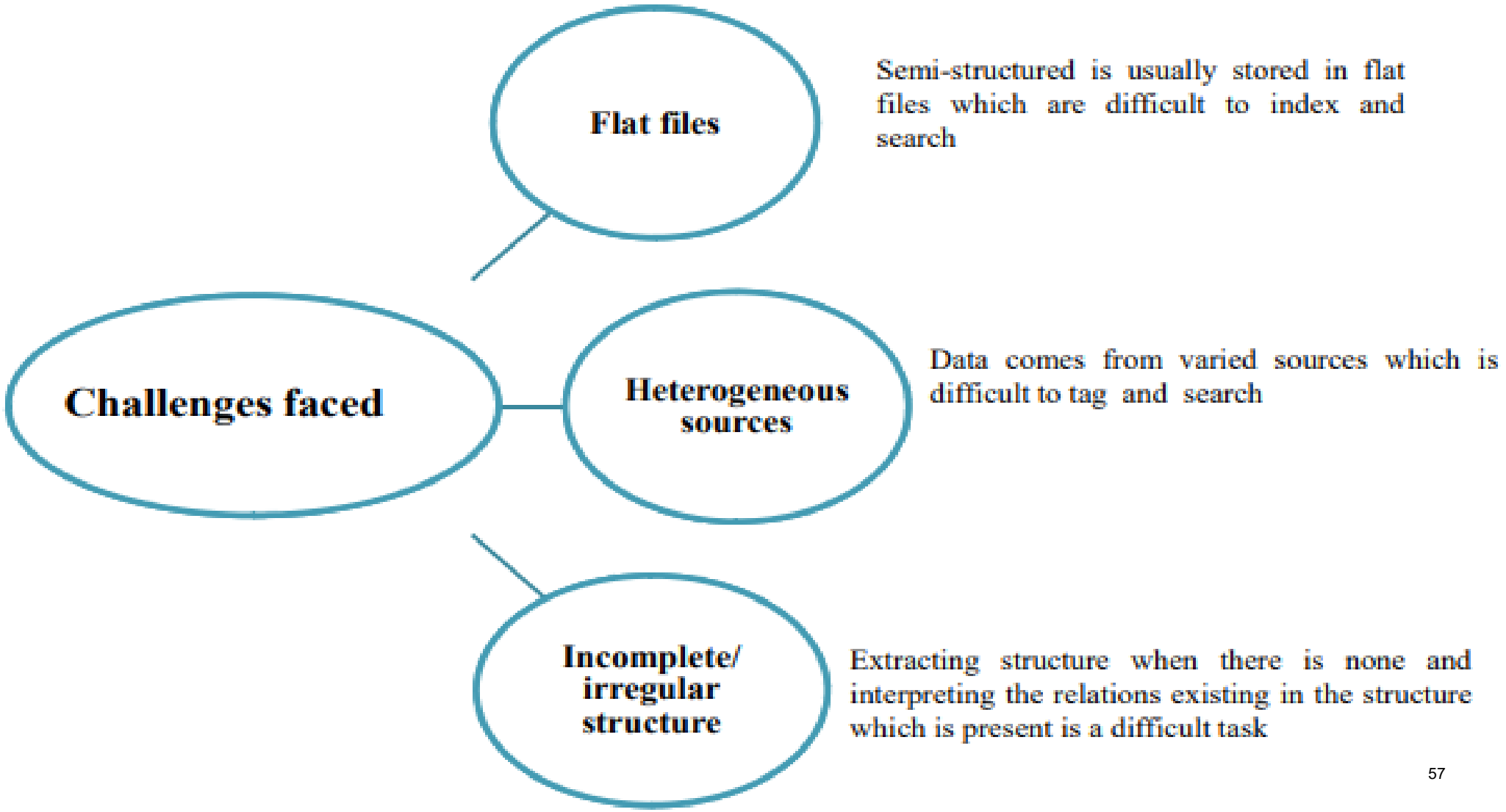
XML allows to define tags and attributes to store data. Data can be stored in a hierarchical/nested structure

Semi-structured data can be stored in a relational database by mapping the data to a relational schema which is then mapped to a table

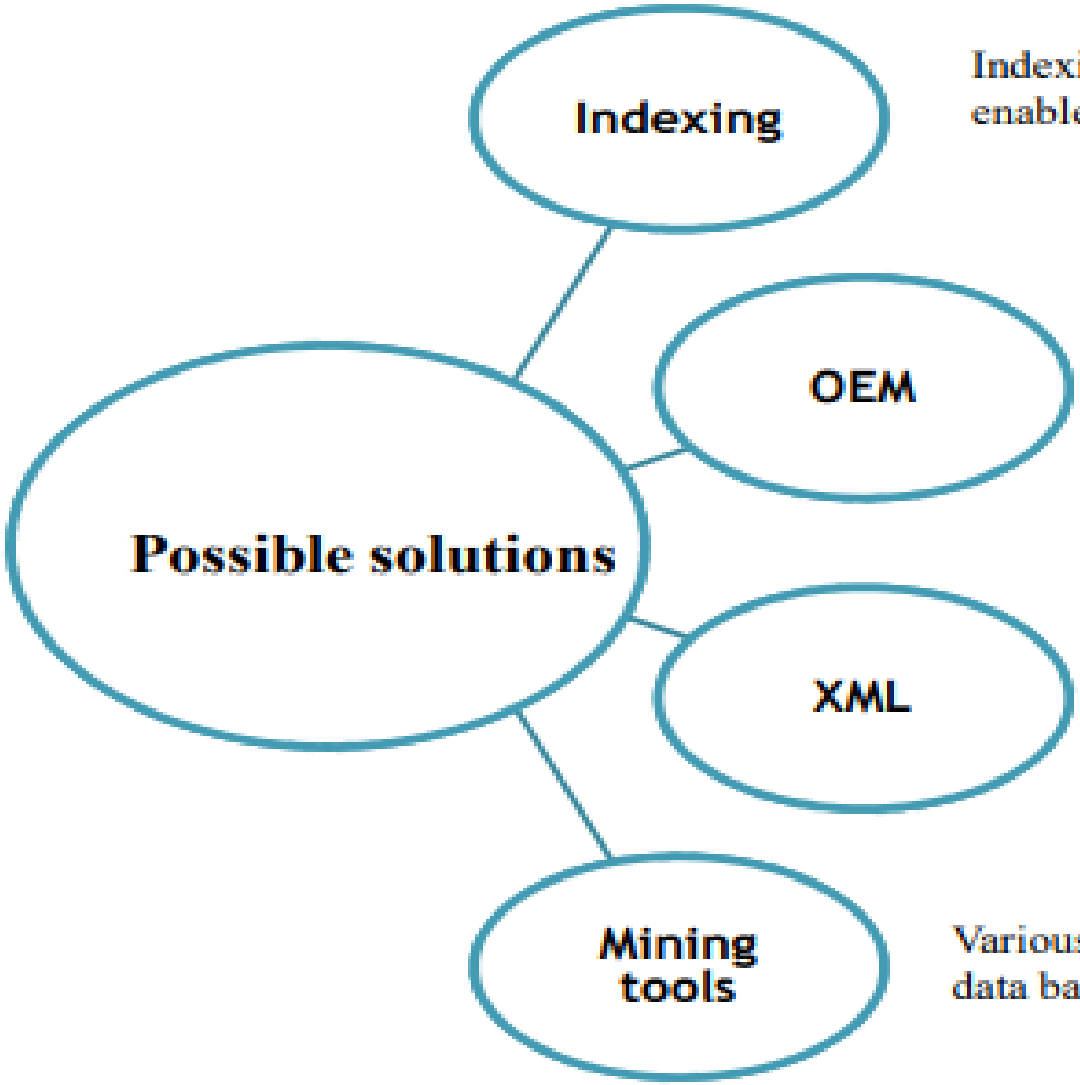
Databases which are specifically designed to store semi-structured data

Data can be stored and exchanged in the form of graph where entities are represented as objects which are the vertices in a graph

Extract Information from Semi-structured data



Conti…



Indexing data in a graph-based model enables quick search

Allows data to be stored in a graph-based data model which is easier to index and search

Allows data to be arranged in a hierarchical or tree-like structure which enables indexing and searching

Various mining tools are available which search data based on graphs, schemas, structure, etc.

Semi-structured Data example

- Personal data stored in an XML file

<rec><name>Anil Dubey</name><sex>Male</sex><age>33</age></rec>
<rec><name>Rohit Rastogi</name><sex>Male</sex><age>43</age></rec>
<rec><name>Sikha</name><sex>Female</sex><age>31</age></rec>

THANK YOU

HISTORY OF BIG DATA INNOVATION

Basic

- Big Data has been described by some Data Management pundits (with a bit of a snicker) as “huge, overwhelming, and uncontrollable amounts of information.” In 1663, **John Graunt** dealt with “overwhelming amounts of information” as well, while he studied the bubonic plague, which was currently ravaging Europe.

- Graunt used statistics and is credited with being the first person to use statistical data analysis. In the early 1800s, the field of statistics expanded to include collecting and analyzing dat

History

- Data became a problem for the U.S. Census Bureau in 1880. They estimated it would take eight years to handle and process the data collected during the 1880 census, and predicted the data from the 1890 census would take more than 10 years to process.
- Fortunately, in 1881, a young man working for the bureau, named Herman Hollerith, created the Hollerith Tabulating Machine. His invention was based on the punch cards designed for controlling the patterns woven by mechanical looms. His tabulating machine reduced ten years of labor into three months of labor.

History

- **1881:** One of the first instances of data overload was experienced during the 1880 census. The ***Hollerith Tabulating Machine is invented*** and the work of ***processing census data*** is cut from ten years of labor to under a year.

- **1928:** German-Austrian engineer Fritz Pfleumer develops magnetic data storage on tape, which led the way for how digital data would be stored in the coming century.

History

- **1948:** Shannon's Information Theory is developed, laying the foundation for the information infrastructure widely used today.

- **1970:** Edgar F. Codd, a mathematician at IBM, presents a "relational database" displaying how information in large databases can be accessed without knowing its structure or location. This was formerly reserved for specialists or those with extensive computer knowledge.

History

- **1976:** Commercial use of Material Requirements Planning (MRP) systems are developed to organize and schedule information, becoming more common for catalyzing business operations.

- **1998:** John Mashey (Chief Scientist at SGI) presented a paper titled "Big Data… and the Next Wave of Infrastress." at a USENIX meeting.

- **1989:** World Wide Web was created by Tim Berners-Lee.

History

- **2001:** Doug Laney presented a paper describing the "3 Vs of Data," which becomes the fundamental characteristics of big data. That same year the term “software-as-a-service” was shared for the first time.
- **2005:** Hadoop, the open-source software framework for large dataset storage is created.
- **2007:** The term “big data” is introduced to the masses in the Wired article "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete."

History

- **2008:** A team of computer science researchers published the paper "Big Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society," describing how big data is fundamentally changing the way companies and organizations do business.
- **2010:** Google CEO Eric Schmidt reveals that every two days people are creating as much information as people created from the beginning of civilization until 2003.
- **2014:** More and more companies begin moving their Enterprise Resource Planning Systems (ERP) to the cloud.
- The Internet of Things (IoT) became widely used with an estimated 3.7 billion connected devices or things in use, transmitting large amounts of data every day.

History

- **2016:** Obama administration releases the "Federal Big Data Research and Strategic Development Plan," designed to drive research and development of big data applications that will directly benefit society and the economy.
- **2017:** IBM study says 2.5 quintillion bytes of data are created daily and that 90% of the world's data has been created in the last two years.

THANK YOU

INTRODUCTION TO BIG DATA PLATFORM

Basic

- Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.
- An enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure /environment.
- Refers to IT solutions that combine several Big Data Tools and utilities into one packaged answer, and this is then used further for managing as well as analyzing Big Data.

Conti..

- Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities.
- It also supports custom development, querying and integration with other systems.
- The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/ solutions into a one cohesive solution.
- Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services.

Platforms

- Hadoop Delta Lake Migration Platform
- Data Catalog Platform
- Data Ingestion Platform
- IoT Analytics Platform
- Data Integration and Management Platform
- ETL Data Transformation Platform

Conti…

Hadoop - Delta Lake Migration Platform

- It is an open-source software platform managed by Apache Software Foundation.
- It is used to manage and store large data sets at a low cost and with great efficiency.

Conti…

Data Catalog Platform

- It provides a single self-service environment to the users, helping them find, understand, and trust the data source.
- It also helps the users to discover the new data sources if there are any. Discovering and understanding data sources are the initial steps for registering the sources.
- Users search for the Data Catalog Tools based on the needs and filter the appropriate results.
- In Enterprises, Data Lake is needed for Business Intelligence, Data Scientists, ETL Developers where the right data needed. The users use catalog discovery to find the data which fits their needs.

Conti…

Data Ingestion Platform

- This layer is the first step for the data coming from variable sources to start its journey. This means the data here is prioritized and categorized, making data flow smoothly in further layers in this process flow.

IoT Analytics Platform

- It provides a wide range of tool to work upon big data; this functionality of it comes handy while using it over the IoT case.

Platforms

Big Data Integration and Management Platform

- Our ElixirData provides a highly customizable solution for Enterprises. ElixirData provides Flexibility, Security, and Stability for an Enterprise application and Big Data Infrastructure to deploy on-premises and Public Cloud with cognitive insights using Machine Learning and Artificial Intelligence.

ETL Data Transformation Platform

- This Platform can be used to build pipelines and even schedule the running of the same for data transformation.

Essential Components of Big Data Platform

There are many essential components which are given as follows:

- Data Ingestion, Management, ETL, and Warehouse
- Stream Computing
- Analytics/ Machine Learning
- Integration
- Data Governance
- Provides Accurate Data
- Scalability
- Price Optimization
- Reduced Latency

Conti…

- **Data Ingestion, Management, ETL, and Warehouse** – It provides these resources for effective data management and effective data warehousing, and this manages data as a valuable resource.

- **Stream Computing** – Helps compute the streaming data that is used for real-time analytics.

- **Analytics/ Machine Learning** – Features for advanced analytics and machine learning.

Conti…

- **Integration** – It provides its user with features like integrating big data from any source with ease.
- **Data Governance** – It also provides comprehensive security, data governance, and solutions to protect the data.
- **Provides Accurate Data** – It delivers with analytic tools which in turn helps to omit any inaccurate data that has not been analyzed. This also helps the business to make the right decision by utilizing accurate information.

Conti…

- **Scalability** – It also helps scale the application to analyze all time climbing data; it sizes to provide efficient analysis. It offers scalable storage capacity.

- **Price Optimization** – Data analytics with the help of a big data platform provides insight for B2C and B2B enterprises which helps the business to optimize the prices they charge accordingly.

- **Reduced Latency** – With the set of the warehouse, analytics tools, and efficient Data transformation, it helps to reduce the data latency and provide high throughput.

Big Data Platform Use Cases

- **Insurance Fraud Detection**: Companies handling a large number of financial transactions use tools provided by this platform to look for any fraud that's happening.

- **In Real Life:** It can be used for various use cases of real-time stream processing like in the field of Media and Entertainment, Weather patterns, the Transportation industry, Banking sector, and so on.

Drivers for Big Data

Six main business drivers

1) Digitization of Society
2) Plummeting of Technology costs
3) Connectivity through Cloud Computing
4) Increased Knowledge about Data Science
5) Social Media Applications
6) Upcoming Internet-of-Things (IoT)

1. Digitization of Society

- Big Data is largely consumer driven and consumer oriented. Most of the data in the world is generated by consumers, who are nowadays ‘always-on’.
- Most people now spend 4-6 hours per day consuming and generating data through a variety of devices and (social) applications.
- With every click, swipe or message, new data is created in a database somewhere around the world. Because everyone now has a smartphone in their pocket, the data creation sums to incomprehensible amounts.

2. Plummeting of Technology Costs

- The costs of data storage and processors keep declining, making it possible for small businesses and individuals to become involved with Big Data.
- For storage capacity, the often-cited Moore's Law still holds that the storage density (and therefore capacity) still doubles every two years. The plummeting of technology costs has been depicted in the figure below.
- Besides the plummeting of the storage costs, a second key contributing factor to the affordability of Big Data has been the development of open source Big Data software frameworks.
- Most popular software framework (nowadays considered the standard for Big Data) is Apache Hadoop for distributed storage and processing.

3. Connectivity through Cloud Computing

- Cloud computing environments (where data is remotely stored in distributed storage systems) have made it possible to quickly scale up or scale down IT infrastructure and facilitate a pay-as-you-go model.
- This means that organizations that want to process massive quantities of data (and thus have large storage and processing requirements) do not have to invest in large quantities of IT infrastructure.
- Instead, they can license the storage and processing capacity they need and only pay for the amounts they actually used.

4. Increased Knowledge about Data Science

- The knowledge and education about data science has greatly professionalized and more information becomes available every day.
- While statistics and data analysis mostly remained an academic field previously, it is quickly becoming a popular subject among students and the working population.

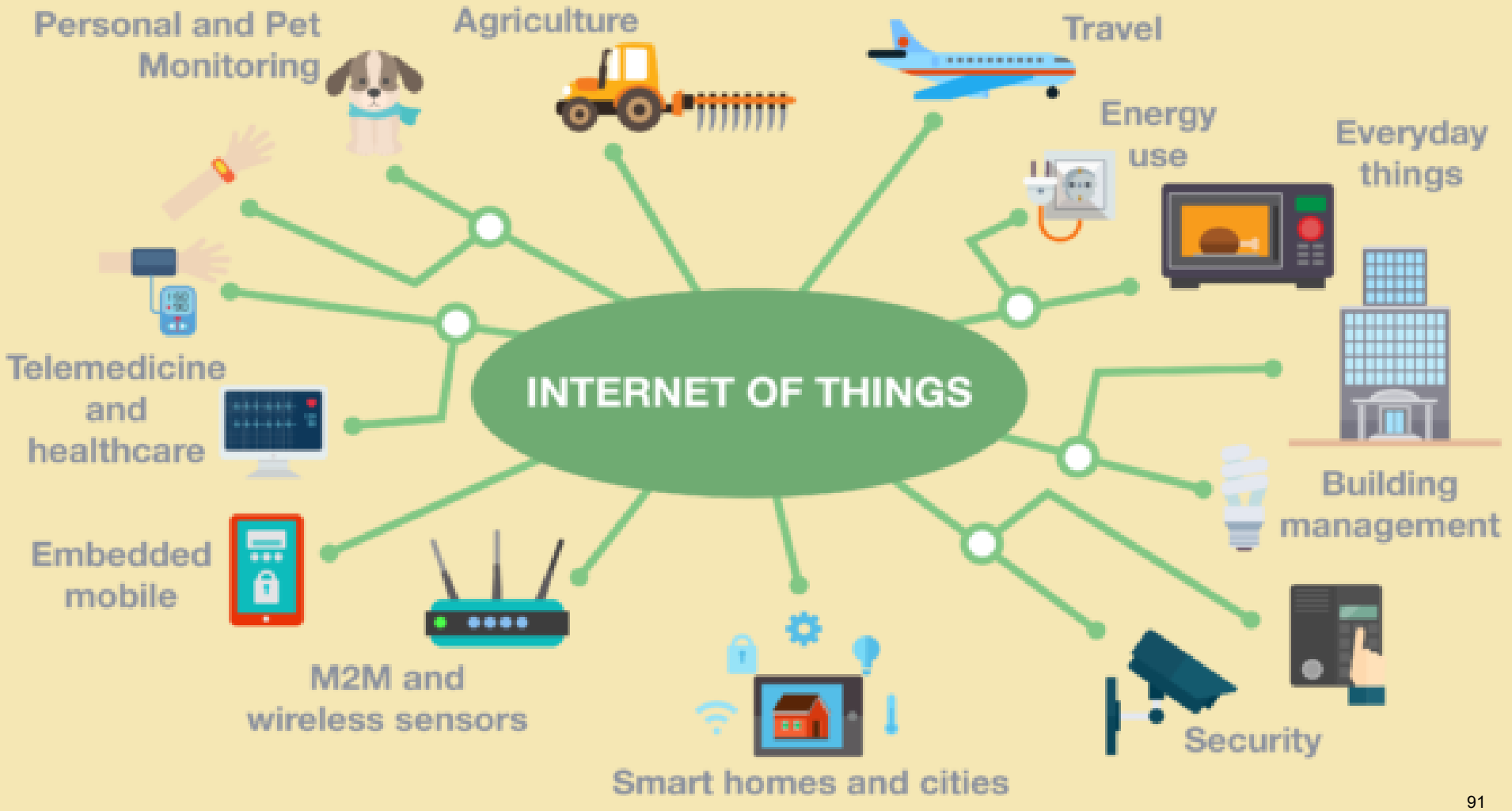
5. Social Media Applications

- Social media data provides insights into the behaviors, preferences and opinions of ‘the public’ on a scale that has never been known before.
- Due to this, it is immensely valuable to anyone who is able to derive meaning from these large quantities of data.
- Social media data can be used to identify customer preferences for product development, target new customers for future purchases, or even target potential voters in elections.
- Social media data might even be considered one of the most important business drivers of Big Data.

6. Upcoming Internet of Things (IoT)

- IoT is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and network connectivity which enables these objects to connect and exchange data.
- It is increasingly gaining popularity as consumer goods providers start including ‘smart’ sensors in household appliances.
- Examples of these devices include thermostats, smoke detectors, televisions, audio systems and even smart refrigerators.

Conti…


Personal and Pet Monitoring
Agriculture
Travel
Energy use
Everyday things
Telemedicine and healthcare
INTERNET OF THINGS
Embedded mobile
Building management
M2M and wireless sensors
Smart homes and cities
Security

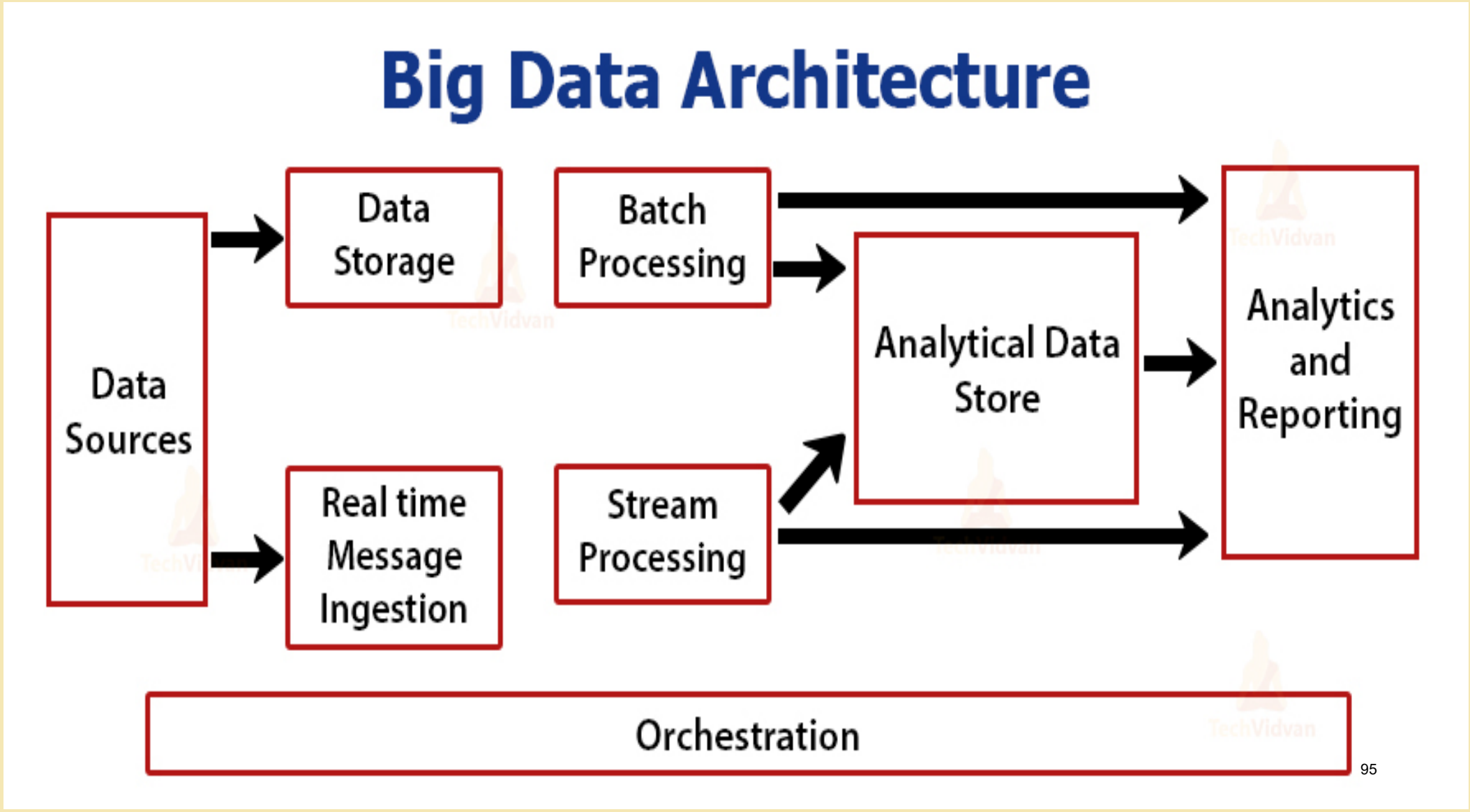
THANK YOU

BIG DATA ARCHITECTURE AND CHARACTERISTICS

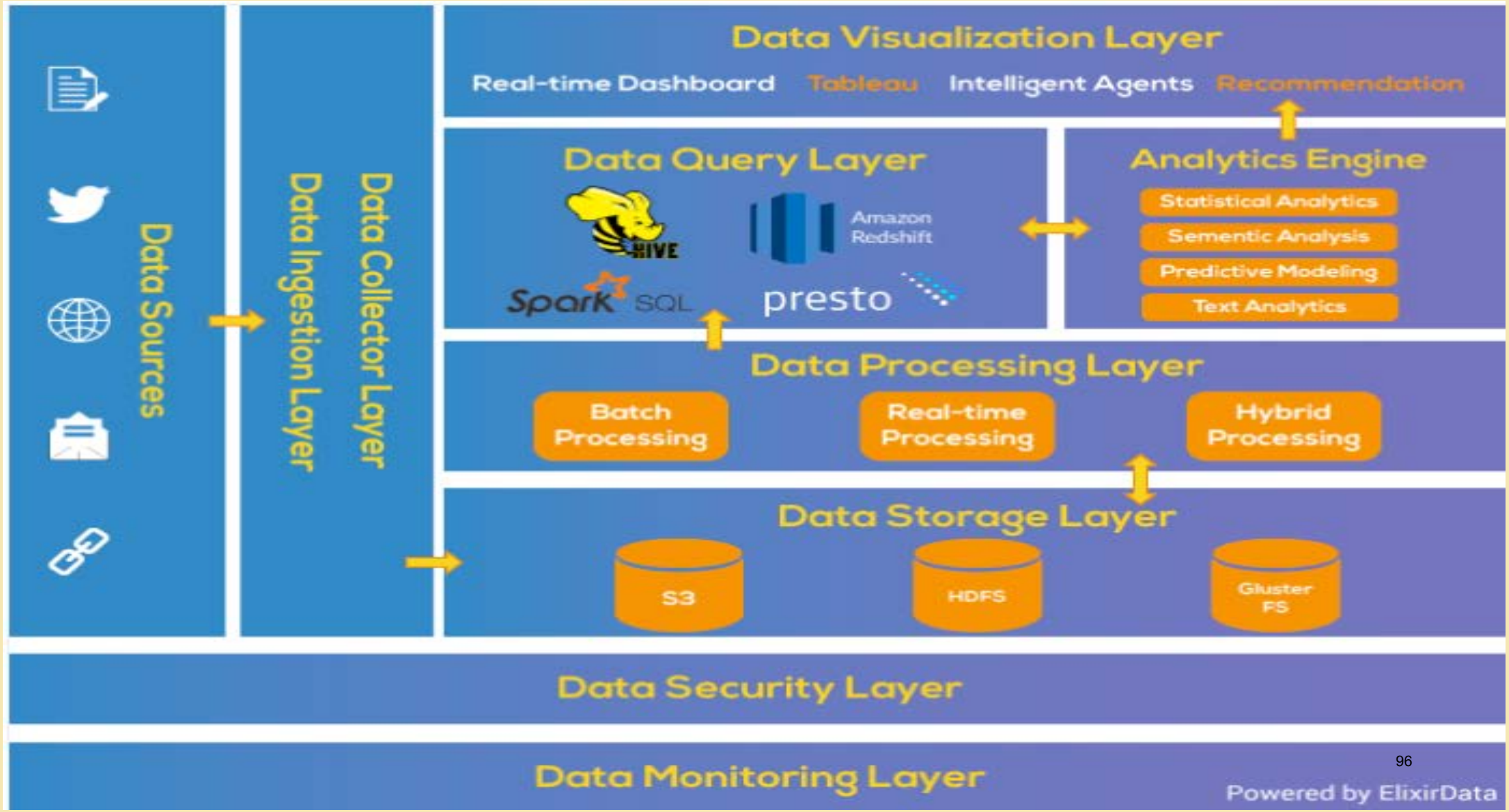
Big Data Architecture

- Big data architecture refers to the logical and physical structure that dictates how high volumes of data are ingested, processed, stored, managed, and accessed.

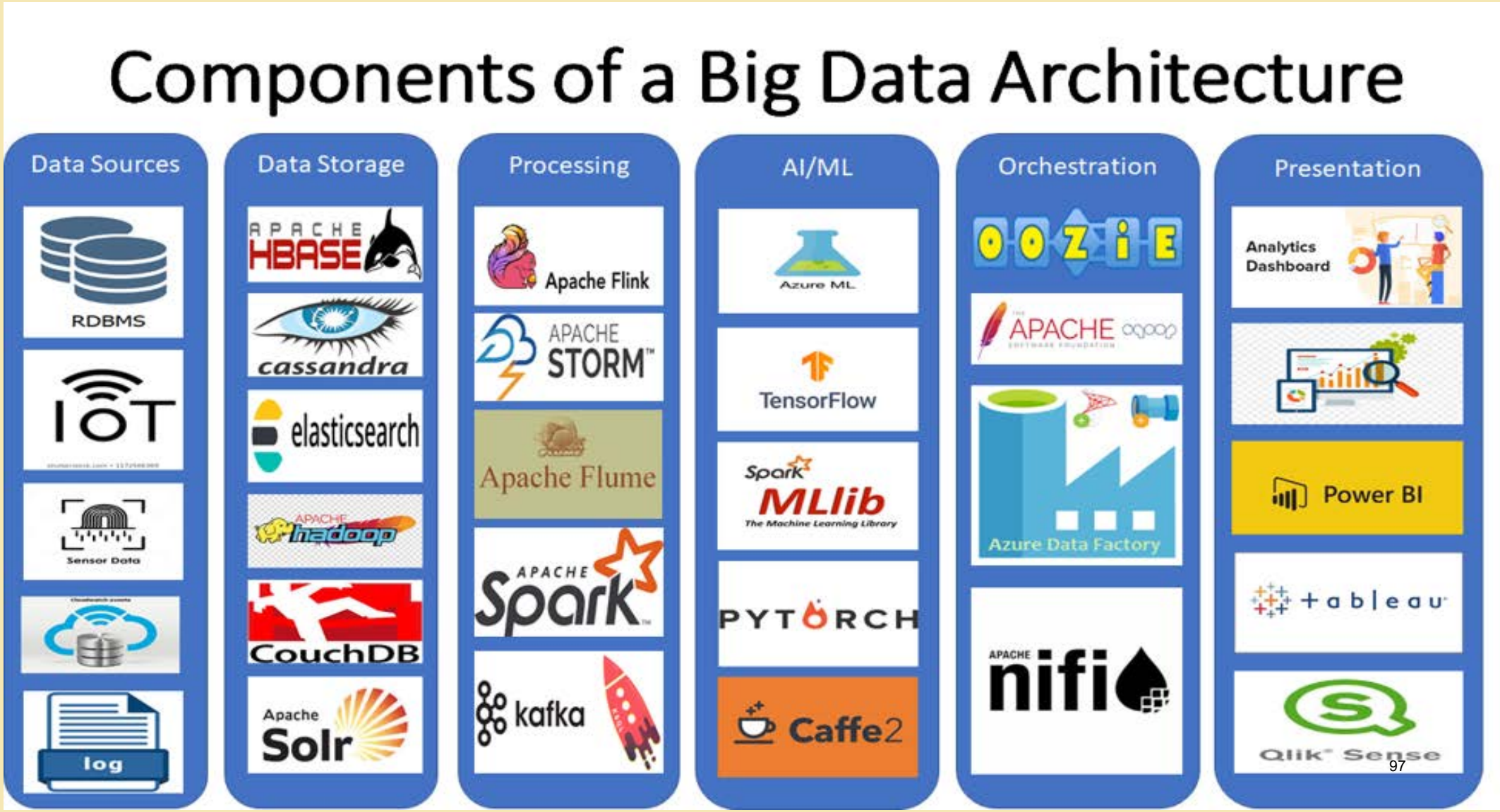
Conti…


Big Data Architecture
Data Sources
Data Storage
Real time Message Ingestion
Batch Processing
Stream Processing
Analytical Data Store
Analytics and Reporting
Orchestration

Conti…


Data Sources
Data Collector Layer
Data Ingestion Layer
Data Visualization Layer
Real-time Dashboard
Tableau
Intelligent Agents
Recommendation
Data Query Layer
HIVE
Amazon Redshift
Spark SQL
presto
Analytics Engine
Statistical Analytics
Sementic Analysis
Predictive Modeling
Text Analytics
Data Processing Layer
Batch Processing
Real-time Processing
Hybrid Processing
Data Storage Layer
S3
HDFS
Gluster FS
Data Security Layer
Data Monitoring Layer
Powered by ElixirData

Conti…


Components of a Big Data Architecture
Data Sources
RDBMS
IOT
Sensor Data
log
Data Storage
APACHE HBASE
cassandra
elasticsearch
APACHE hadoop
CouchDB
Apache Solr
Processing
Apache Flink
APACHE STORM
Apache Flume
APACHE Spark
kafka
AI/ML
Azure ML
TensorFlow
Spark MLlib
The Machine Learning Library
PYTORCH
Caffe2
Orchestration
OOZIE
APACHE
Azure Data Factory
APACHE nifi
Presentation
Analytics Dashboard
Power BI
+ableau
Qlik Sense

Conti…

- **Big Data Sources Layer**

Big data environment can manage both batch processing and real-time processing of big data sources, such as data warehouses, relational database management systems, SaaS applications, and IoT devices.

- **Management & Storage Layer**

Receives data from the source, converts the data into a format comprehensible for the data analytics tool, and stores the data according to its format.

Conti…

- **Analysis Layer**

Analytics tools extract business intelligence from the big data storage layer.

- **Consumption Layer**

Receives results from the big data analysis layer and presents them to the pertinent output layer - also known as the business intelligence layer.

Conti…

- **Connecting to Data Sources**

Connectors and adapters are capable of efficiently connecting any format of data and can connect to a variety of different storage systems, protocols, and networks.

- **Data Governance**

Includes provisions for privacy and security, operating from the moment of ingestion through processing, analysis, storage, and deletion.

Conti…

- **Systems Management**

Highly scalable, large-scale distributed clusters are typically the foundation for modern big data architectures, which must be monitored continually via central management consoles.

- **Protecting Quality of Service**

Quality of Service framework supports the defining of data quality, compliance policies, and ingestion frequency and sizes.

Big Data Architecture Best Practices

- **Preliminary Step**

It is also important to have a thorough *understanding of the elements* of the current business technology landscape, such as business strategies and organizational models, business principles and goals, current frameworks in use, governance and legal frameworks, IT strategy, and any pre-existing architecture frameworks and repositories.

Conti…

- **Data Sources**

Data sources can be *categorized* as either structured data, which is typically formatted using predefined database techniques, or unstructured data, which does not follow a consistent format, such as emails, images, and Internet data.

Conti…

- **Big Data ETL** (Extract, Transform, Load)

Data should be *consolidated into a single Master Data Management system* for querying on demand, either via batch processing (through Hadoop) or stream processing.

For querying, the Master Data Management system can be stored in a data repository such as NoSQL-based or relational DBMS

Conti…

- **Data Services API**

Consider whether or not there is a standard query language, how to connect to the database, the ability of the database to scale as data grows, and which security mechanisms are in place.

Conti…

- **User Interface Service**

Big data application architecture should have an intuitive design that is customizable, available through current dashboards in use, and accessible in the cloud.

Standards like *Web Services for Remote Portlets* (WSRP) facilitate the serving of User Interfaces through Web Service calls.

How to Build a Big Data Architecture

Designing a big data reference architecture, while complex, follows the same general procedure:

- **Analyze the Problem**

First determine if the business does in fact have a big data problem, taking into consideration criteria such as data variety, velocity, and challenges with the current system.

Common use cases include data archival, process offload, data lake implementation, unstructured data processing, and data warehouse modernization.

Conti…

- **Select a Vendor**

Hadoop is one of the most widely recognized big data architecture tools for managing big data end to end architecture.

Popular vendors for Hadoop distribution include Amazon Web Services, BigInsights, Cloudera, Hortonworks, Mapr, and Microsoft.

Conti…

- **Deployment Strategy**

Deployment can be either on-premises, which tends to be more secure; cloud-based, which is cost effective and provides flexibility regarding scalability; or a mix deployment strategy.

- **Capacity Planning**

When planning hardware and infrastructure sizing, consider daily data ingestion volume, data volume for one-time historical load, the data retention period, multi-data center deployment, and the time period for which the cluster is sized.

Conti…

- **Infrastructure Sizing**

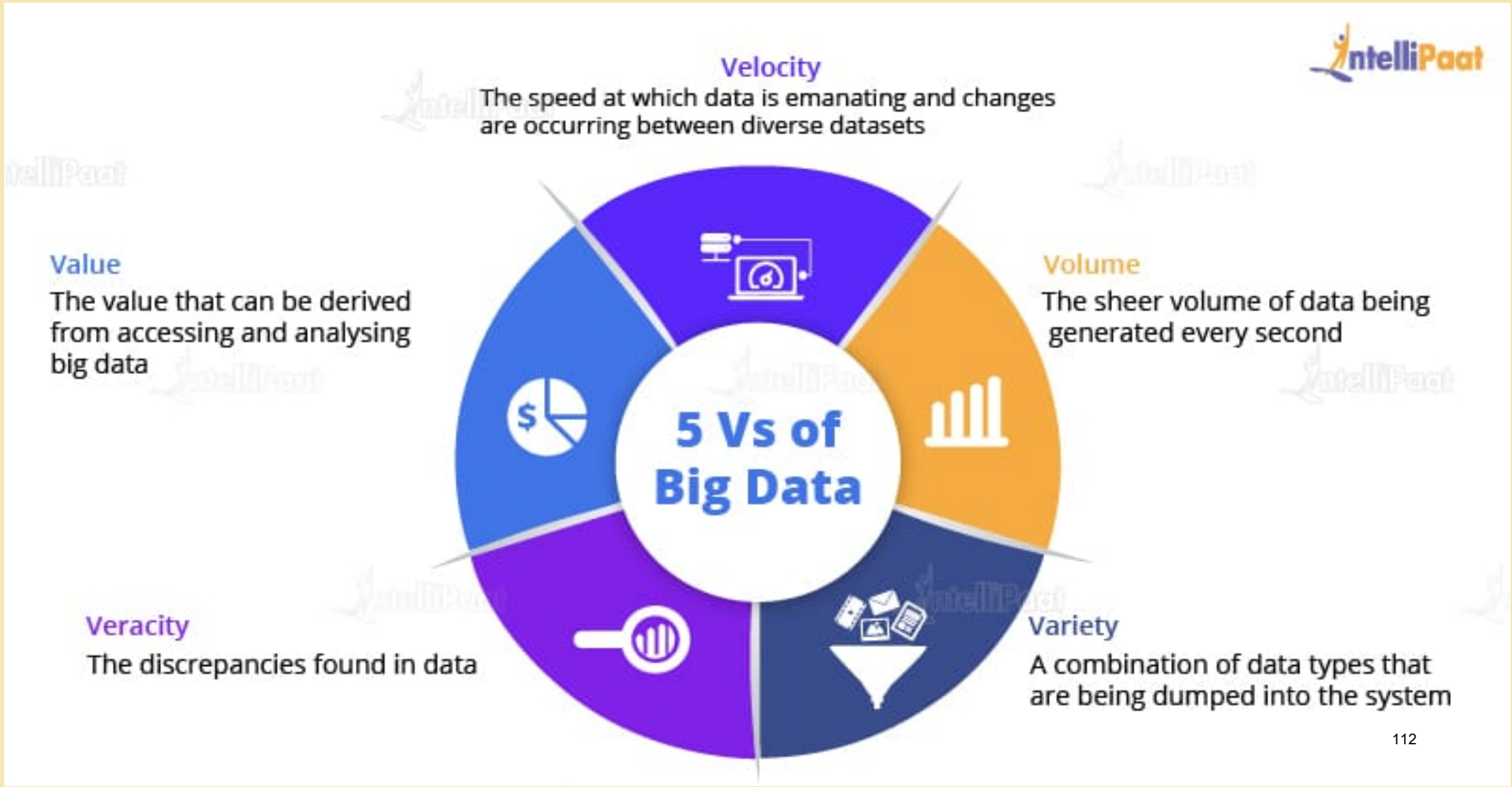
Based on capacity planning and determines the number of clusters/environment required and the type of hardware required. Consider the type of disk and number of disks per machine, the types of processing memory and memory size, number of CPUs and cores, and the data retained and stored in each environment.

Conti…

- **Plan a Disaster Recovery**

In developing a backup and disaster recovery plan, consider the criticality of data stored, the Recovery Point Objective and Recovery Time Objective requirements, backup interval, multi datacenter deployment, and whether Active-Active or Active-Passive disaster recovery is most appropriate.

Characteristics of Big Data



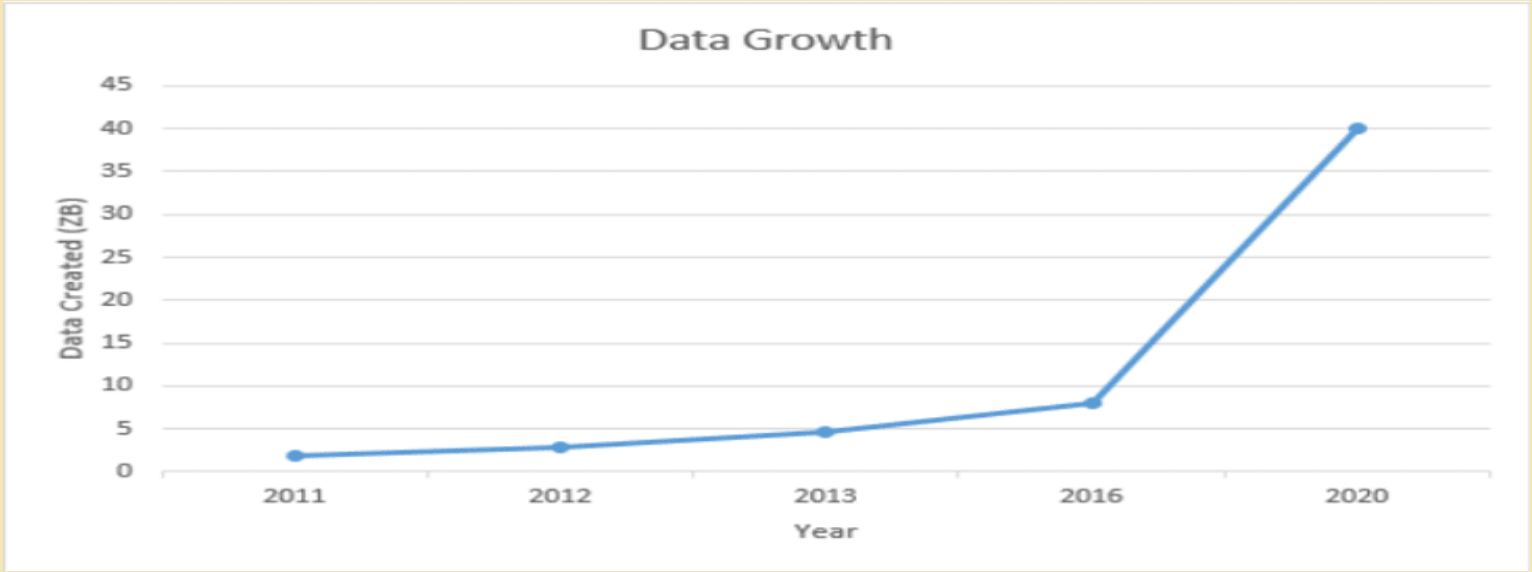
Conti…

Big data can be described by the following characteristics

1) Volume
2) Variety
3) Velocity
4) Variability
5) Value

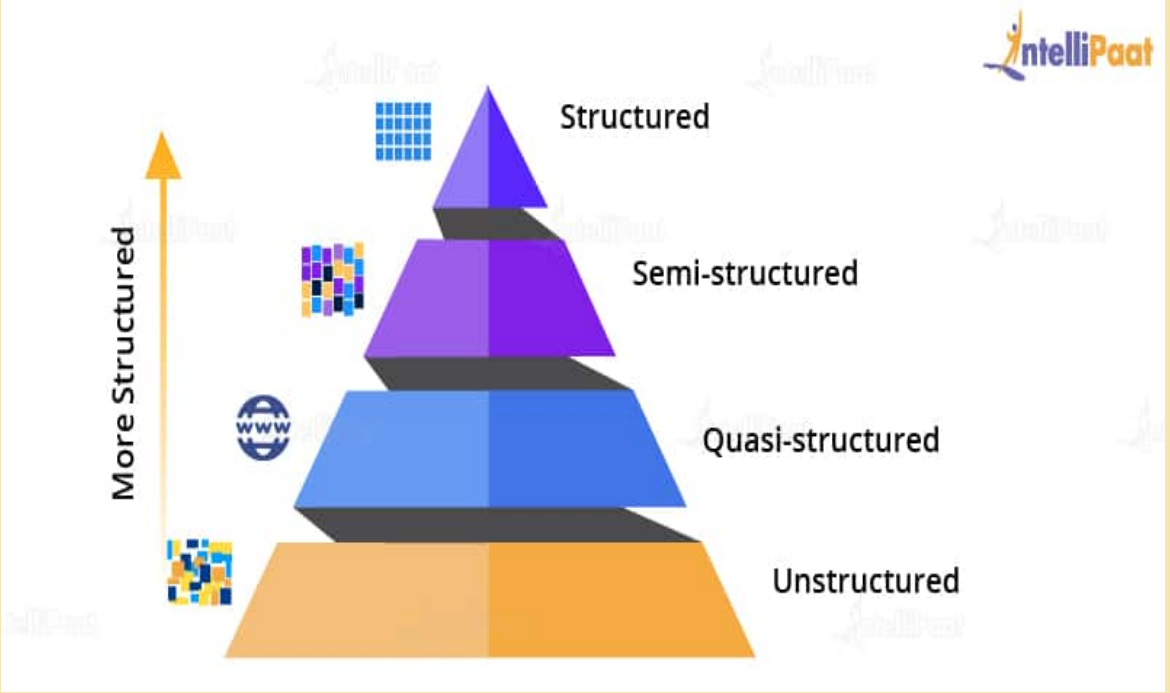
1. Volume

- Volume is one of the characteristics of big data. We already know that Big Data indicates huge '**volumes**' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. As you can see from the image, the volume of data is rising exponentially. In 2016, the data created was only 8 ZB; it is expected that, by 2020, the data would rise to 40 ZB, which is extremely large.



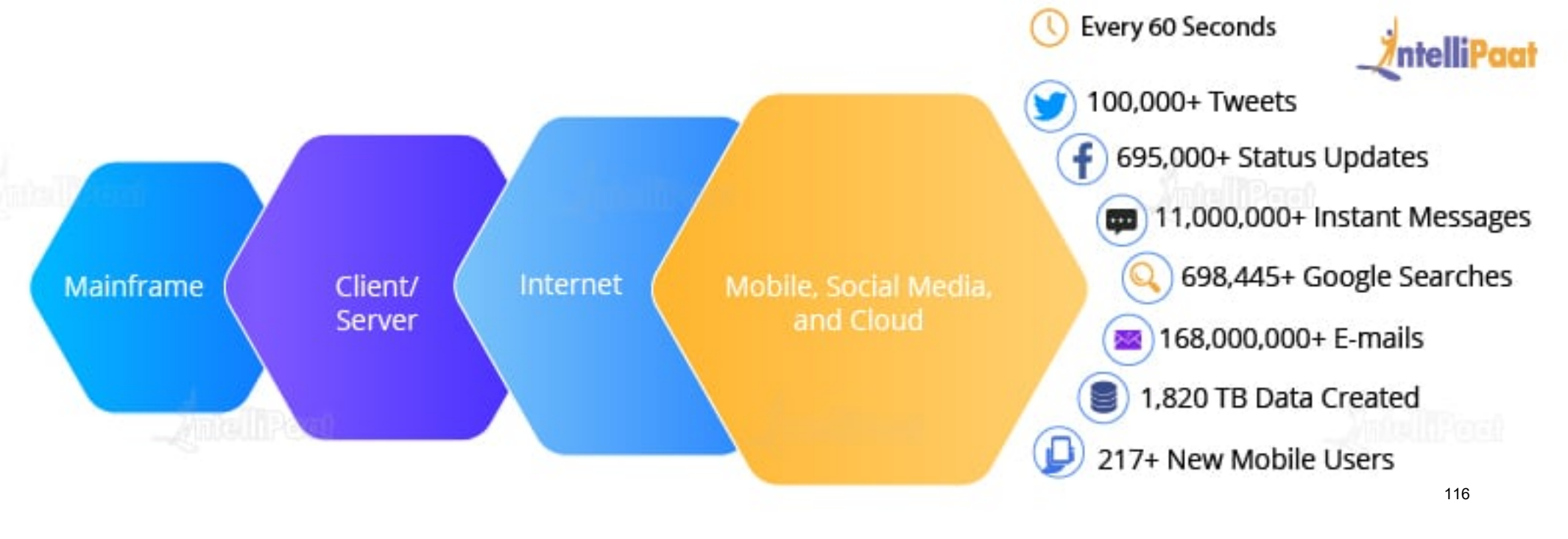
2. Variety

- Variety of Big Data refers to structured, unstructured, and semistructured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more.



3. Velocity

- The term **‘velocity’** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data networks, and social media sites, sensors, Mobile devices, etc. As can be seen in the image below, mainframes were initially used when fewer people were using computers.

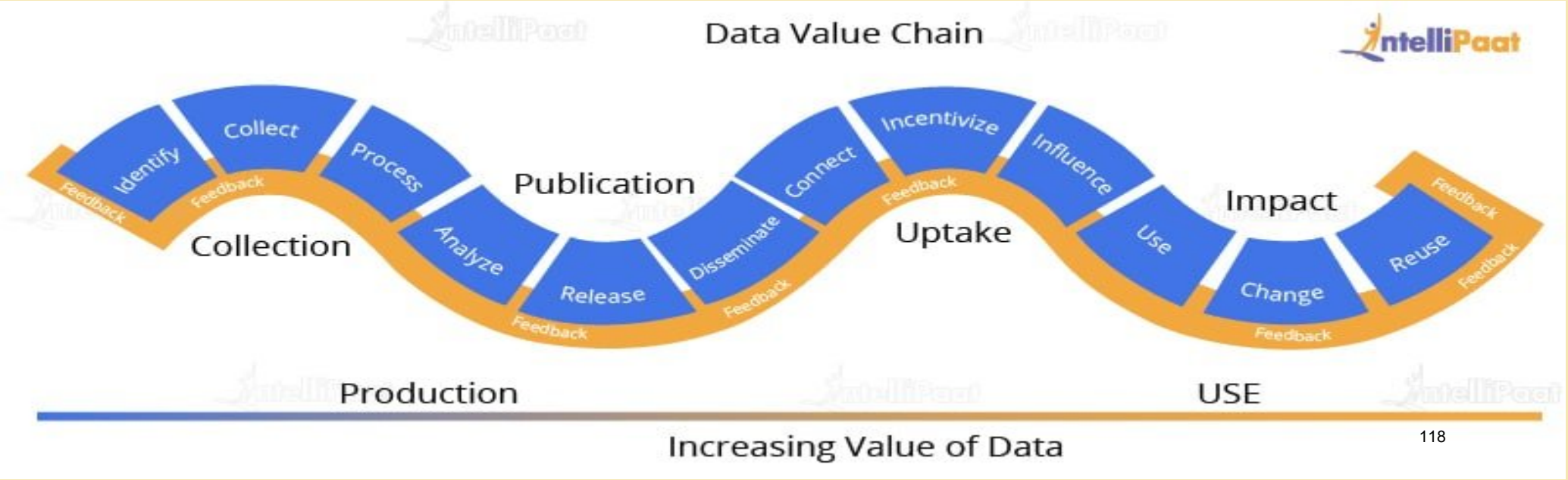


4. Variability

- This refers to the uncertainties and inconsistencies in the data that can be overcome by veracity. Veracity means the trustworthiness and quality of data.
- The veracity of data must be maintained.
- For example, think about Facebook posts, hashtags, abbreviations, images, videos, etc., which make the posts unreliable and hamper the quality of their content. Collecting loads and loads of data is of no use if the quality and trustworthiness of the data are not up to the mark.

5. Value

- It deals with a mechanism to bring out the correct meaning of data. First of all, you need to mine data, i.e., the process to turn raw data into useful data. Then, an analysis is done on the data that you have cleaned or retrieved from the raw data. Then, you need to make sure whatever analysis you have done benefits your business, such as in finding out insights, results, etc., in a way that was not possible earlier.



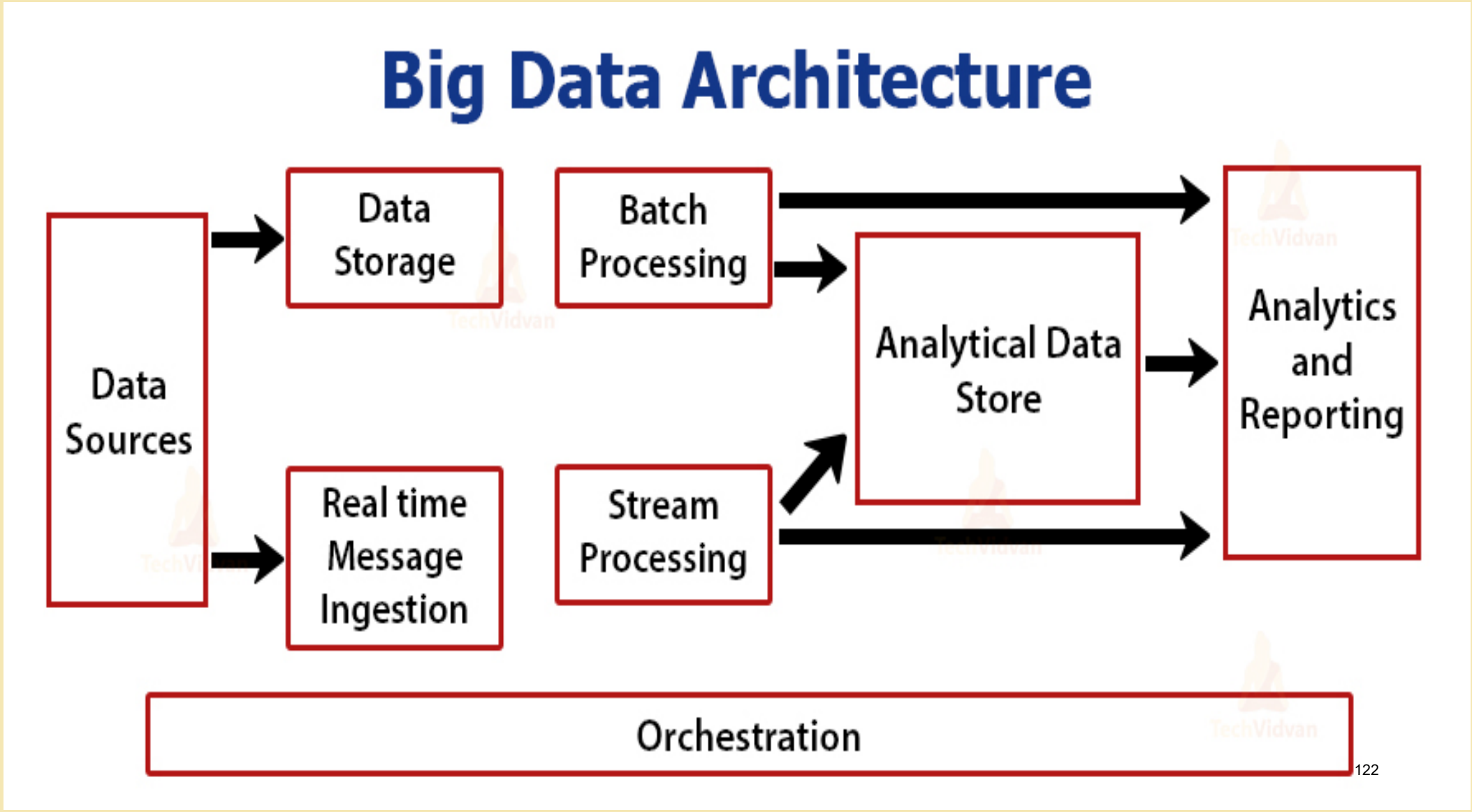
THANK YOU

BIG DATA ARCHITECTURE AND CHARACTERISTICS

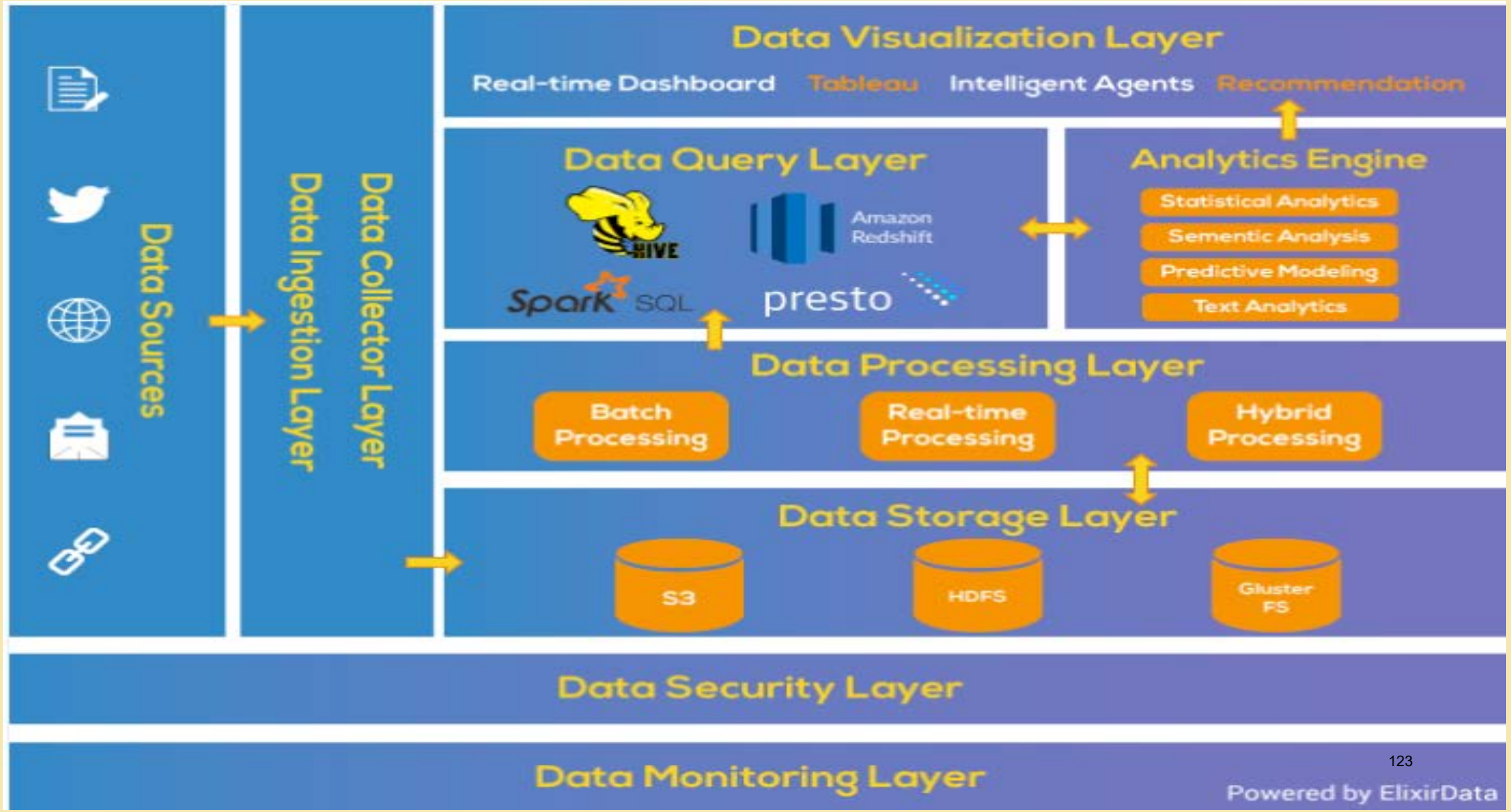
Big Data Architecture

- Big data architecture refers to the logical and physical structure that dictates how high volumes of data are ingested, processed, stored, managed, and accessed.

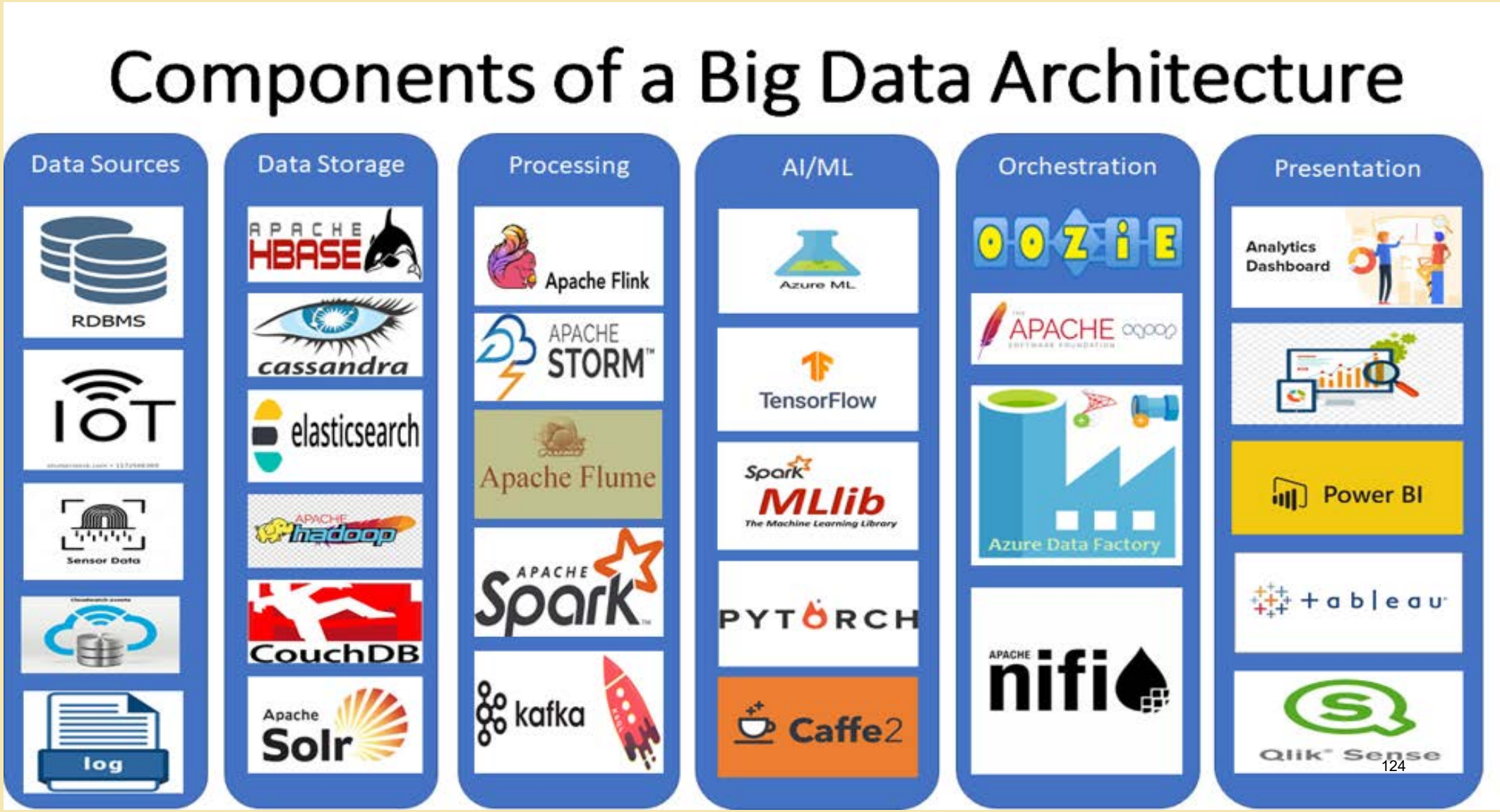
Conti…


Big Data Architecture
Data Sources
Data Storage
Batch Processing
Real time Message Ingestion
Stream Processing
Analytical Data Store
Analytics and Reporting
Orchestration

Conti…


Data Sources
Data Collector Layer
Data Ingestion Layer
Data Visualization Layer
Real-time Dashboard
Tableau
Intelligent Agents
Recommendation
Data Query Layer
HIVE
Amazon Redshift
Spark SQL
presto
Analytics Engine
Statistical Analytics
Sementic Analysis
Predictive Modeling
Text Analytics
Data Processing Layer
Batch Processing
Real-time Processing
Hybrid Processing
Data Storage Layer
S3
HDFS
Gluster FS
Data Security Layer
Data Monitoring Layer
Powered by ElixirData

Conti…


Components of a Big Data Architecture
Data Sources
RDBMS
IOT
Sensor Data
log
Data Storage
APACHE HBASE
cassandra
elasticsearch
APACHE hadoop
CouchDB
Apache Solr
Processing
Apache Flink
APACHE STORM
Apache Flume
APACHE Spark
kafka
AI/ML
Azure ML
TensorFlow
Spark MLlib
The Machine Learning Library
PYTORCH
Caffe2
Orchestration
OOZIE
APACHE
Azure Data Factory
APACHE nifi
Presentation
Analytics Dashboard
Power BI
+ableau
Qlik Sense

Conti…

- **Big Data Sources Layer**

Big data environment can manage both batch processing and real-time processing of big data sources, such as data warehouses, relational database management systems, SaaS applications, and IoT devices.

- **Management & Storage Layer**

Receives data from the source, converts the data into a format comprehensible for the data analytics tool, and stores the data according to its format.

Conti…

- **Analysis Layer**

Analytics tools extract business intelligence from the big data storage layer.

- **Consumption Layer**

Receives results from the big data analysis layer and presents them to the pertinent output layer - also known as the business intelligence layer.

Conti…

- **Connecting to Data Sources**

Connectors and adapters are capable of efficiently connecting any format of data and can connect to a variety of different storage systems, protocols, and networks.

- **Data Governance**

Includes provisions for privacy and security, operating from the moment of ingestion through processing, analysis, storage, and deletion.

Conti…

- **Systems Management**

Highly scalable, large-scale distributed clusters are typically the foundation for modern big data architectures, which must be monitored continually via central management consoles.

- **Protecting Quality of Service**

Quality of Service framework supports the defining of data quality, compliance policies, and ingestion frequency and sizes.

Big Data Architecture Best Practices

- **Preliminary Step**

It is also important to have a thorough *understanding of the elements* of the current business technology landscape, such as business strategies and organizational models, business principles and goals, current frameworks in use, governance and legal frameworks, IT strategy, and any pre-existing architecture frameworks and repositories.

Conti…

- **Data Sources**

Data sources can be *categorized* as either structured data, which is typically formatted using predefined database techniques, or unstructured data, which does not follow a consistent format, such as emails, images, and Internet data.

Conti…

- **Big Data ETL** (Extract, Transform, Load)

Data should be *consolidated into a single Master Data Management system* for querying on demand, either via batch processing (through Hadoop) or stream processing.

For querying, the Master Data Management system can be stored in a data repository such as NoSQL-based or relational DBMS

Conti…

- **Data Services API**

Consider whether or not there is a standard query language, how to connect to the database, the ability of the database to scale as data grows, and which security mechanisms are in place.

Conti…

- **User Interface Service**

Big data application architecture should have an intuitive design that is customizable, available through current dashboards in use, and accessible in the cloud.

Standards like *Web Services for Remote Portlets* (WSRP) facilitate the serving of User Interfaces through Web Service calls.

How to Build a Big Data Architecture

Designing a big data reference architecture, while complex, follows the same general procedure:

- **Analyze the Problem**

First determine if the business does in fact have a big data problem, taking into consideration criteria such as data variety, velocity, and challenges with the current system.

Common use cases include data archival, process offload, data lake implementation, unstructured data processing, and data warehouse modernization.

Conti…

- **Select a Vendor**

Hadoop is one of the most widely recognized big data architecture tools for managing big data end to end architecture.

Popular vendors for Hadoop distribution include Amazon Web Services, BigInsights, Cloudera, Hortonworks, Mapr, and Microsoft.

Conti…

- **Deployment Strategy**

Deployment can be either on-premises, which tends to be more secure; cloud-based, which is cost effective and provides flexibility regarding scalability; or a mix deployment strategy.

- **Capacity Planning**

When planning hardware and infrastructure sizing, consider daily data ingestion volume, data volume for one-time historical load, the data retention period, multi-data center deployment, and the time period for which the cluster is sized.

Conti…

- **Infrastructure Sizing**

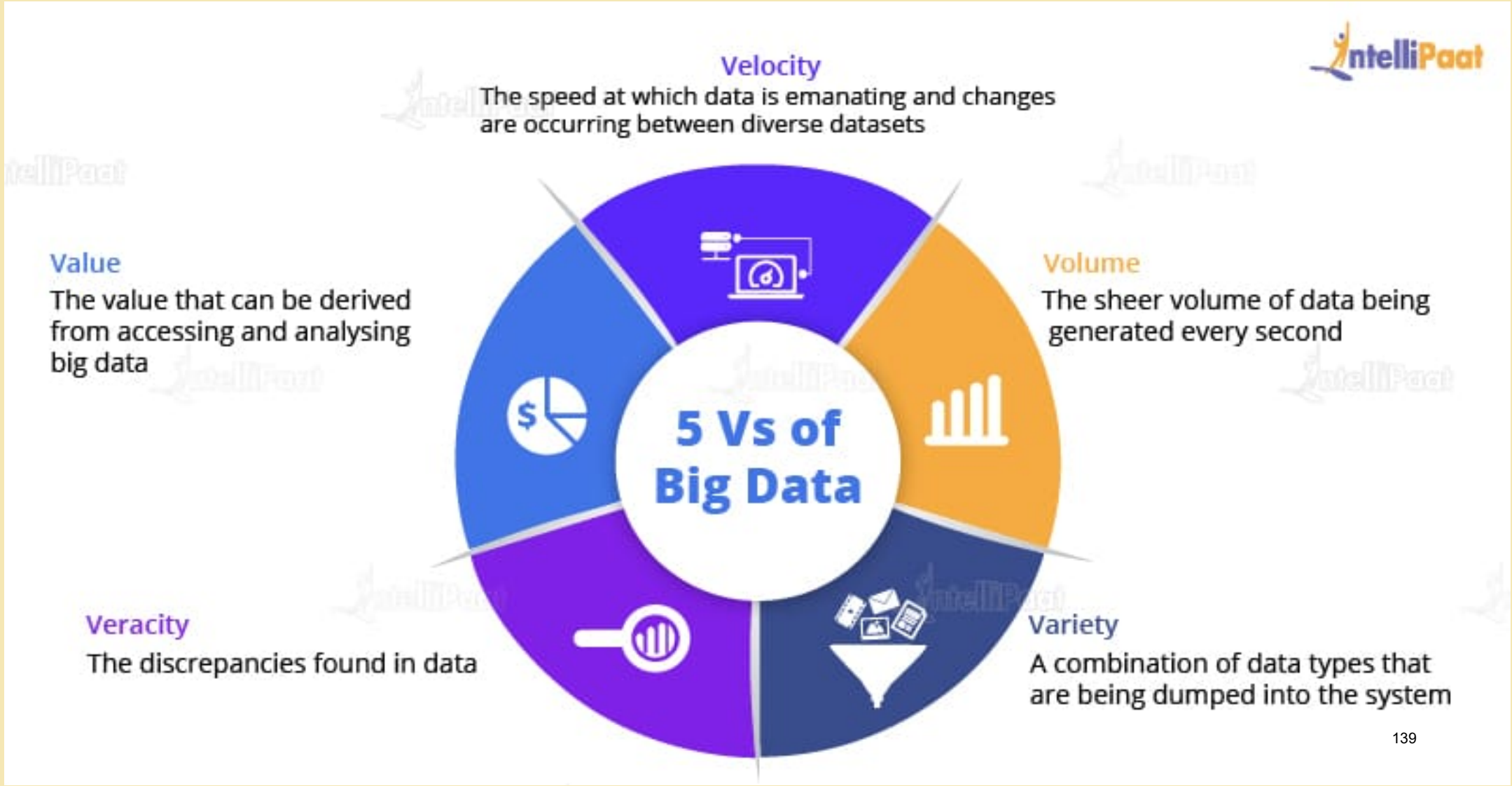
Based on capacity planning and determines the number of clusters/environment required and the type of hardware required. Consider the type of disk and number of disks per machine, the types of processing memory and memory size, number of CPUs and cores, and the data retained and stored in each environment.

Conti…

- **Plan a Disaster Recovery**

In developing a backup and disaster recovery plan, consider the criticality of data stored, the Recovery Point Objective and Recovery Time Objective requirements, backup interval, multi datacenter deployment, and whether Active-Active or Active-Passive disaster recovery is most appropriate.

Characteristics of Big Data



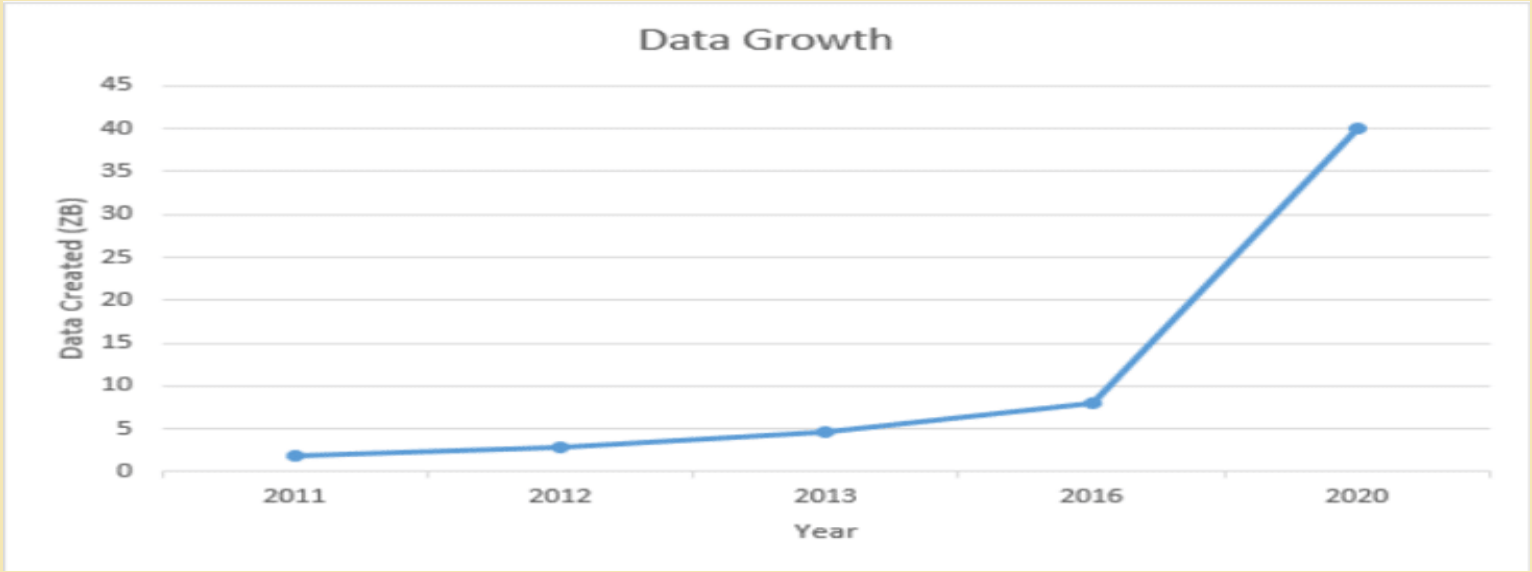
Conti…

Big data can be described by the following characteristics

1) Volume
2) Variety
3) Velocity
4) Variability
5) Value

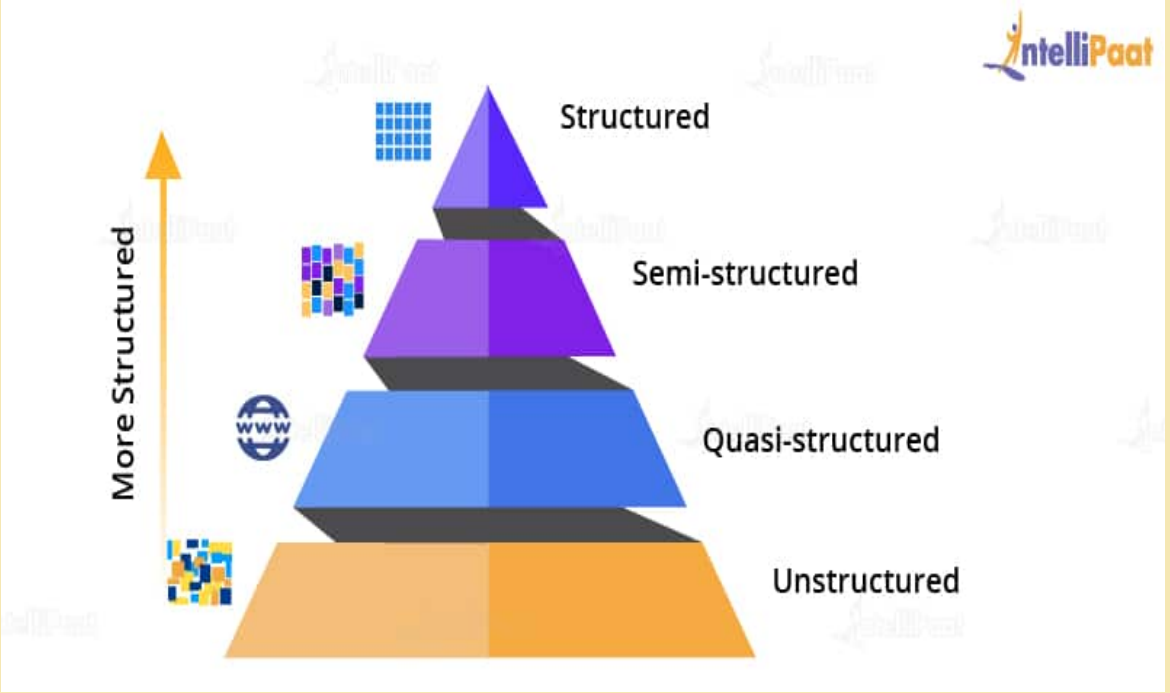
1. Volume

- Volume is one of the characteristics of big data. We already know that Big Data indicates huge '**volumes**' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. As you can see from the image, the volume of data is rising exponentially. In 2016, the data created was only 8 ZB; it is expected that, by 2020, the data would rise to 40 ZB, which is extremely large.



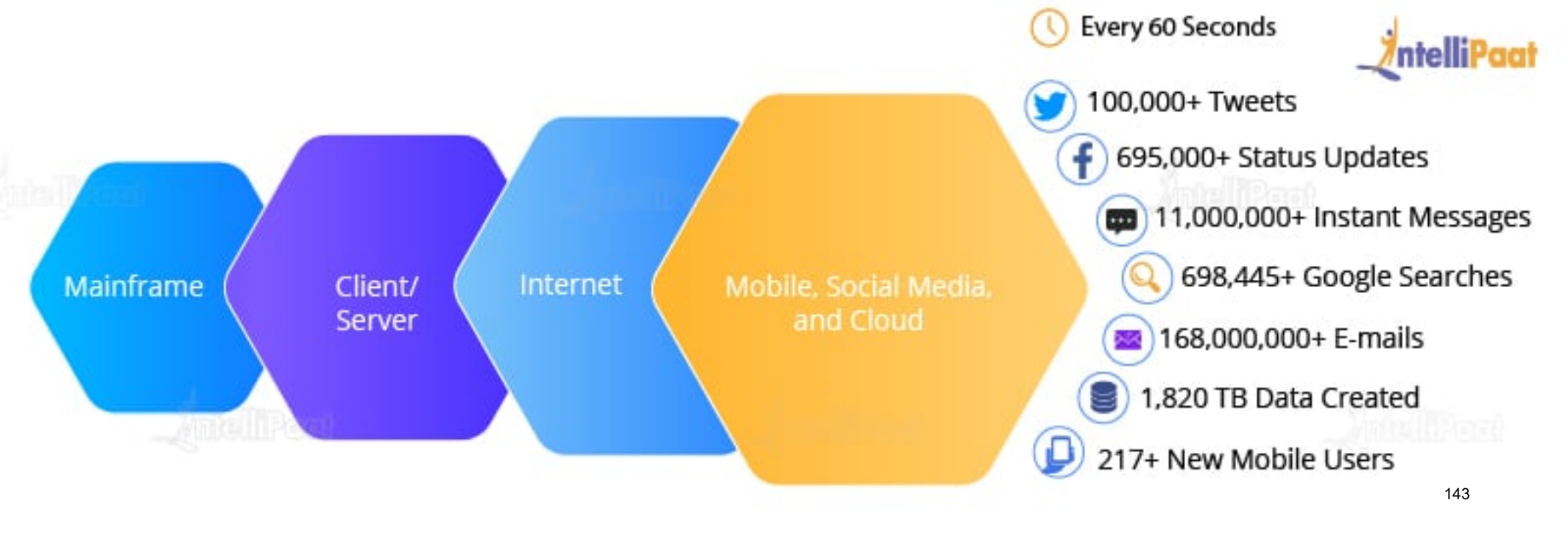
2. Variety

- Variety of Big Data refers to structured, unstructured, and semistructured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more.



3. Velocity

- The term **‘velocity’** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data networks, and social media sites, sensors, Mobile devices, etc. As can be seen in the image below, mainframes were initially used when fewer people were using computers.

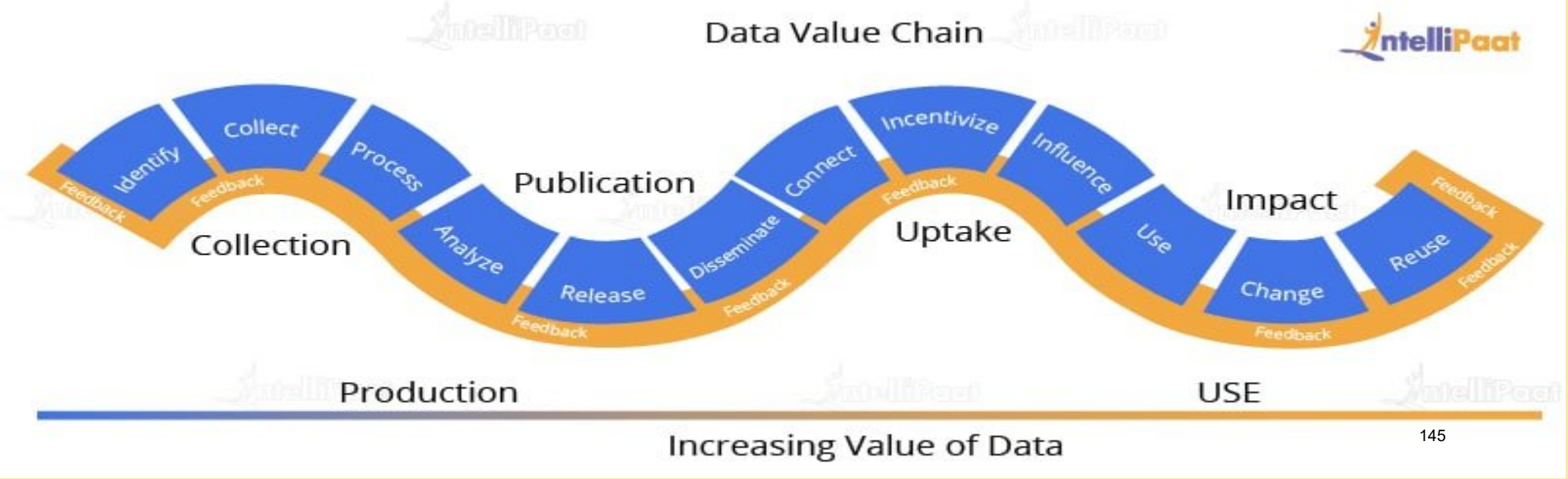


4. Variability

- This refers to the uncertainties and inconsistencies in the data that can be overcome by veracity. Veracity means the trustworthiness and quality of data.
- The veracity of data must be maintained.
- For example, think about Facebook posts, hashtags, abbreviations, images, videos, etc., which make the posts unreliable and hamper the quality of their content. Collecting loads and loads of data is of no use if the quality and trustworthiness of the data are not up to the mark.

5. Value

- It deals with a mechanism to bring out the correct meaning of data. First of all, you need to mine data, i.e., the process to turn raw data into useful data. Then, an analysis is done on the data that you have cleaned or retrieved from the raw data. Then, you need to make sure whatever analysis you have done benefits your business, such as in finding out insights, results, etc., in a way that was not possible earlier.



THANK YOU

BIG DATA FEATURES – SECURITY

Data Security

- Big data security can be termed as the tool and measures which are used to guard both data and analytics processes.

- **Main purpose of Big data security is to provide protection against attacks, thefts, and other malicious activities that could harm valuable data.**

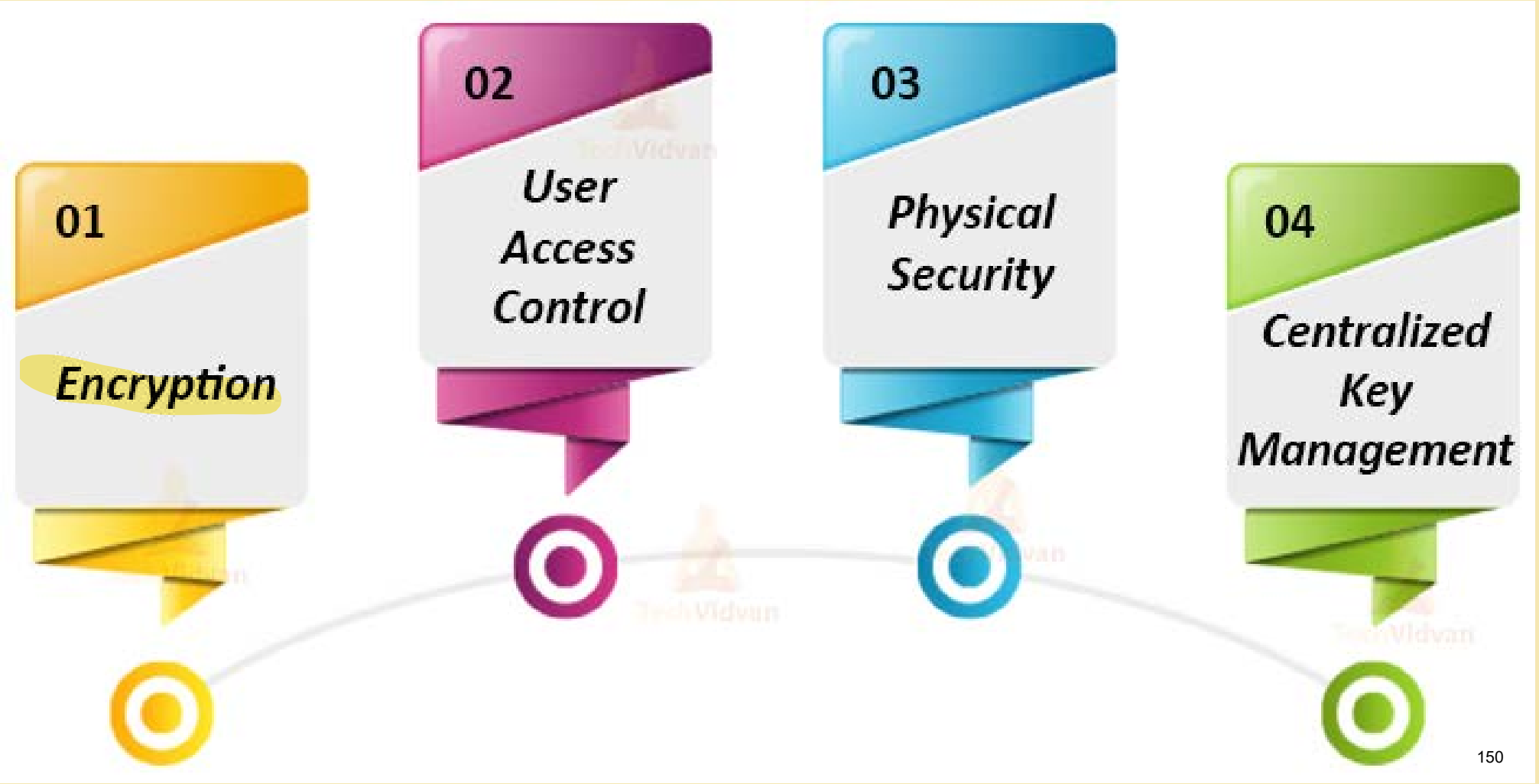
Conti…

- Big data security challenges are multi-faced for the companies that operate on the cloud.

- This challenging threat includes the theft of information stored online, ransomware, or DDoS attacks that could crash a server.

- These threats can cause serious financial repercussions such as losses, litigation costs, and fines or sanctions of an organization.

Big Data Security Technologies



1. Encryption

- Encryption of data is generally done to secure a massive volume of data, different types of data.
- It can be user-generated or machine-generated code.
- Encryption tools along with different analytics toolsets format or code the data.
- They also get applied to data from different sources like relational database management system (RDBMS), specialized file systems like Hadoop Distributed File System (HDFS), etc.

2. User Access Control

- It is the most basic network security tool. But few companies practice this because it involves high management overhead, this can be dangerous at the network level and not good for the Big data platforms.
- Automated strong user access control is a must for organizations.
- Automation control manages complex user control levels that protects the Big data platform against the inside attack.

3. Physical Security

- It is generally built in when you deploy the Big data platform in your own center.
- It can also be built around your cloud provider's data center security.
- They are important as they can deny data center access to strangers or suspicious visitors.
- Video surveillance and security logs are also used for the same purpose.

4. Centralized Key Management

- It is applied in Big data environments, especially on those having wide geographical distribution.

- Best practices under centralized key management include policy-driven automation, on-demand key delivery, logging, and abstracting key management from key usage.

Conti…

Some of the companies practicing Big Data Securities are:

- **Cloudwick**

CDAP (Cloudwick Data Analytics Platform) is a managed security hub that integrates security features from multiple analytics toolsets and machine learning projects.

- **IBM**

IBM security Guardium is used to monitor Big data and NoSQL environments. It includes the discovery and classification of sensitive data.

Conti…

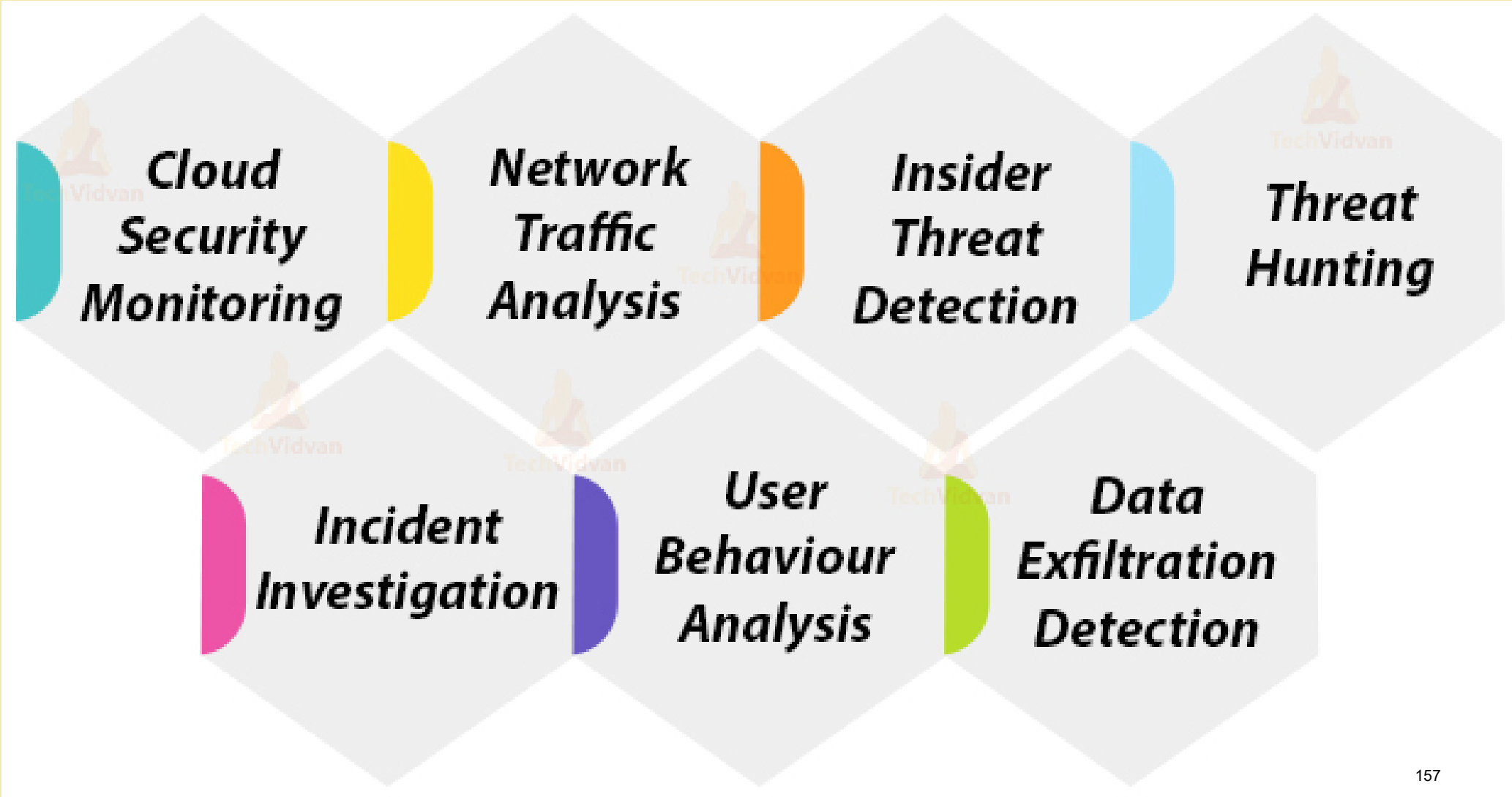
- **Logtrust**

Logtrust is partnered with Panda Security in order to provide the ART (Advanced Reporting Tool) and Panda Adaptive Defense.

- **Gemalto**

Gemalto SafeNet protects Big data platforms. Usually, it protects the Big data platforms in the cloud, data center, and virtual environments.

Big Data Security Use Cases



A. Cloud Security Monitoring

- Cloud computing generally offers more efficient communication and increased profitability for all businesses. This communication needs to be secure.

- Big data security offers cloud application monitoring.

- This provides host sensitive data and also monitors cloud-hosted infrastructure. Solutions also offer support across several relevant cloud platforms.

B. Network Traffic Analysis

- Due to the high volume of data over the network, it is difficult to maintain transactional visibility over the network traffic.
- Security analytics allow your enterprise to watch over this network traffic. It is used to establish baselines and detect anomalies.
- It is used to analyze traffic in and out of cloud infrastructure. It also illuminates dark spaces that are hidden in infrastructures and analyze encrypted sensitive data. Thus, ensuring the proper working of channels.

C. Insider Threat Detection

- Insider threats are as much as a danger to your enterprise as external threats.

- An active malicious user can do as much damage as any malware attack.

- But it is only in some rare cases that an insider threat can destroy a network.

D. Threat Hunting

- Security analytics helps to automate this threat of hunting.
- It acts as an extra set of eyes for your threat hunting efforts.
- Threats hunting automation can help in detecting malware beaconing activity and thus alerts for its stoppage as soon as possible.

E. Incident Investigation

- Generally, the sheer number of security alerts from SIEM solutions would overwhelm your IT security team. These continuous alerts can cause more fostering burnout and frustration.
- Thus to minimize this issue, security analytics automates the incident investigation by providing contextualization to alerts.
- Thus your team has more time to prioritize incidents and can deal with potential breach incidents first.

F. User Behaviour Analysis

- Organization's users generally interact with your IT infrastructure all the time. Mainly it is the user's behavior that decides the success or failure of your cybersecurity.
- The security analytics monitor the unusual behavior of employees.
- An example of one such renowned security analytics use case is UEBA (**User and entity behavior analytics)**. It helps to provide visibility into the IT environment.
- Thus compiling user activities from multiple datasets into complete profiles.

G. Data Exfiltration Detection

- Data exfiltration is termed as any unauthorized movement of data moving in and out of any network.
- Unauthorized data movements can cause theft and leakage of data.
- Thus there is a need to protect data from such unauthorized access.
- Security analytics helps to detect the data exfiltration over a network. It is generally used to detect data leakage in encrypted communications.

Big Data Security Issues

Conti…

1. Access Controls

- It is critically important for an organization to have a system which is fully secure.
- Permission to exchange the data should be permitted to authenticated users only.
- Access control needs to be such that it would not get hacked by attackers, hackers, or by any malicious activities.

2. Non-relational data stores

- Non-relational databases like NoSQL usually lack security by themselves.

Conti…

3. Storage

- In Big data architecture, we store data on multiple tiers.
- Its storage depends on business needs in terms of performance and cost.
- For example, high-priority data is generally stored on flash media. So locking down storage means creating a tier-conscious strategy.

Conti…

4. Endpoints

- Security solutions that usually draw logs from endpoints will need to validate the authenticity of those endpoints or the analysis is not going to do much.

5. Real-time security/ compliance tools

- Real-time tools generally generate a large amount of information.
- The key is to find a way to ignore false or rough information. So that human talent can be focused on true breaches or valuable information.

Conti…

6. Data mining solutions

- Data mining solutions generally find a pattern that suggests business strategies.
- For this reason, there is a need for ensuring that it is secured from both internal and external threats.

Big Data Privacy and Ethics

- Private customer data and identity should remain private
- Shared private information should be treated confidentially
- Customers should have a transparent view
- Big Data should not interfere with human will
- Big data should not institutionalize unfair biases

Conti…

- **Private customer data and identity should remain private**

Privacy does not mean secrecy, as private data might need to be audited based on legal requirements, but that private data obtained from a person with their consent should not be exposed for use by other businesses or individuals with any traces to their identity.

Conti…

- **Shared private information should be treated confidentially**

Third party companies share sensitive data — medical, financial or locational — and need to have restrictions on whether and how that information can be shared further.

- **Customers should have a transparent view**

How our data is being used or sold, and the ability to manage the flow of their private information across massive, third-party analytical systems.

Conti…

- **Big Data should not interfere with human will**

Big data analytics can moderate and even determine who we are before we make up our own minds. Companies need to begin to think about the kind of predictions and inferences that should be allowed and the ones that should not.

- **Big data should not institutionalize unfair biases**

Machine learning algorithms can absorb unconscious biases in a population and amplify them via training samples.

Challenges of Conventional Systems

Three Challenges that big data face

- Data or Volume
- Process
- Management

Conti…

- The volume of data, especially machine-generated data, is exploding and How fast that data is growing every year, with new sources of data that are emerging.

- For example, in the year 2000, 800,000 petabytes (PB) of data were stored in the world, and it is expected to reach 35 zettabytes (ZB) by2020 (according to IBM).

Reporting v/s Analysis

- **Reporting:** The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

- Reporting translates raw data into information.

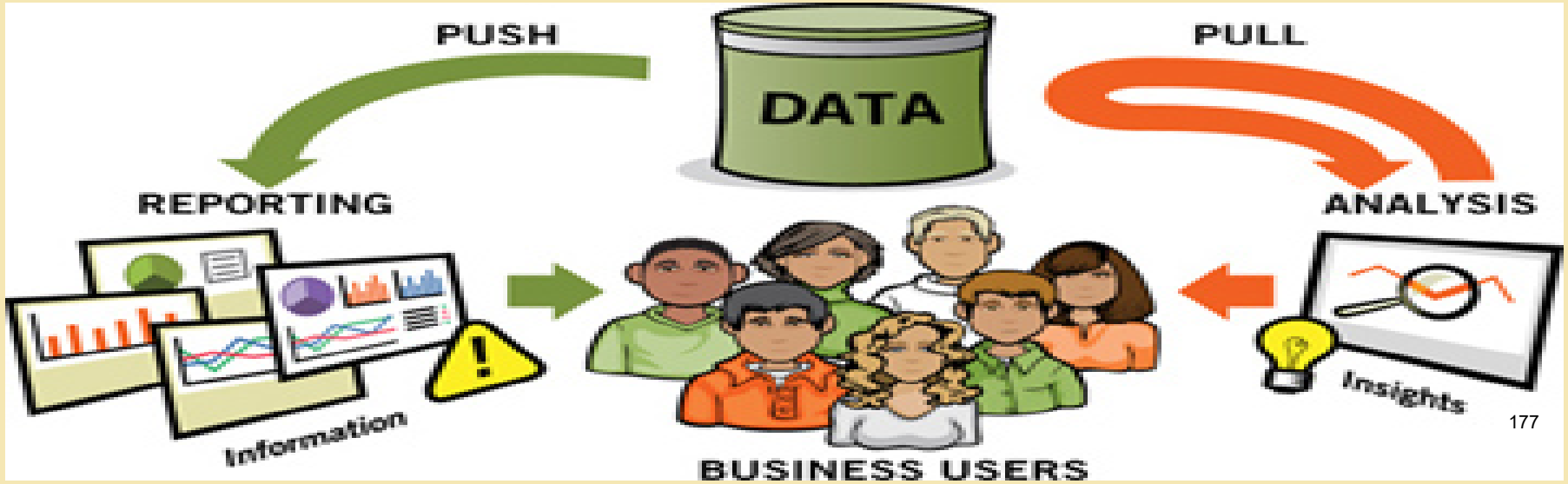
- **Analysis:** The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

- Analysis transforms data and information into insights.

Conti…

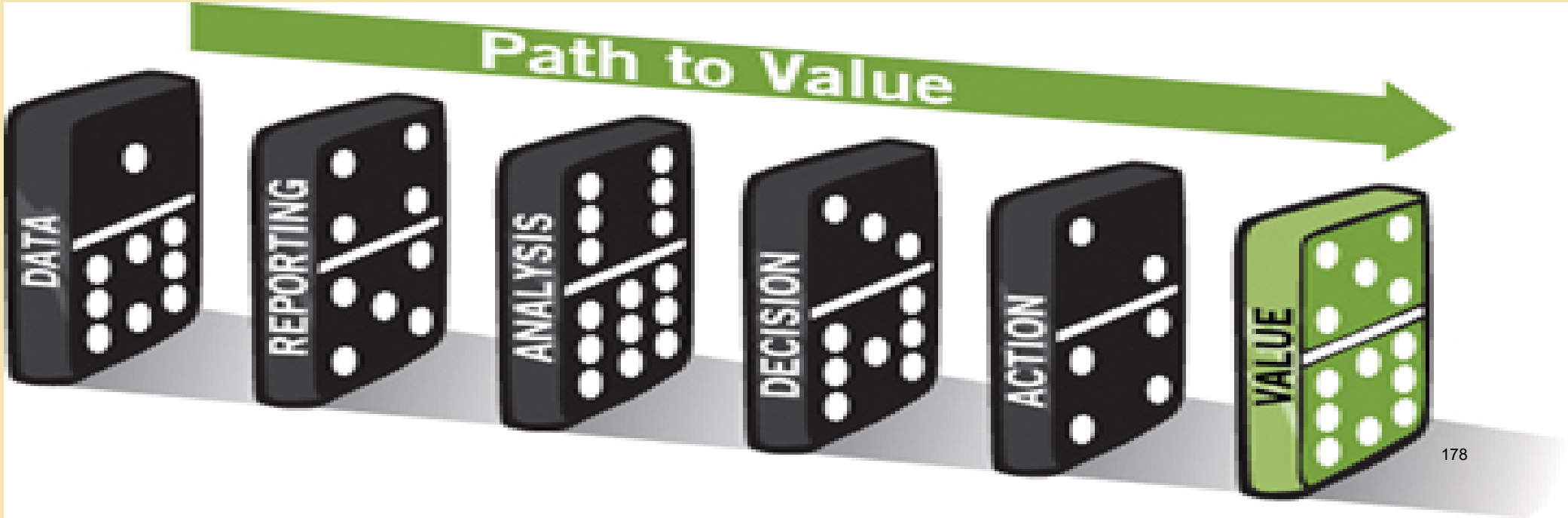
- **Data reporting:** Gathering data into one place and presenting it in visual representations.

- **Data analysis:** Interpreting your data and giving it context.



Conti…

- Analysis transforms data and information into insights.
- Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges.



Conti…

	Reporting	Analysis
Purpose	Monitor and alert	Interpret and recommend actions
Tasks	Build, Configure, Consolidate, Organize, Format, Summarize	Question, Examine, Interpret, Compare, Confirm
Outputs	Canned reports, Dashboards, Alert	Ad hoc responses, Analysis, presentations (findings + recommendations)
Delivery	Accessed via tool, Scheduled for delivery	prepared and shared by analyst
Value	Distills data into, information for further analysis, Alerts company to exceptions in data	provides deeper insights into business, Offers recommendations to drive action

THANK YOU

ABES ENGINEERING COLLEGE, GHAZIABAD

Unit-1

QUESTION BANK

SUBJECT NAME: BIG DATA
UNIT: 1

Q.No	Description	CO	year	Long-short
Part 1: Types of Digital Data, History of Big Data Innovatum.				
1	Write short note on digital data	CO1		Long
2	Explain different types of digital data.	CO1		Long
3	Differentiate between structured, unstructured and semi – structured data.	CO1		Long
4	Describe the history of Big data innovation.	CO1		Long
Part 2: Introduction to Big Data Platform, Drivers for Big Data				
5	Write a short note on: Big data platform.	CO1		Long
6	Describe the drivers of Big data.	CO1		Long
Part 3: Big Data Architecture and Characteristics, 5Vs of Big Data				
7	Describe the architecture of Big Data.	CO1		Long
8	What are the characteristics of Big data? OR Explain 5Vs of Big data. OR Discuss about the three dimensions of Big data.	CO1		Long
Part 4: Big Data Technology Component				
9	What are the Big data technology components?	CO1		Long
Part 5: Big Data Importance and Application				
10	Why Big data is important?	CO1		Long
11	What are the applications of Big Data?	CO1		Long
Part 6: Big Data Features, Security, Compliance, Auditing and Protections				

ABES ENGINEERING COLLEGE, GHAZIABAD

12	What is Big Data Security? What are the steps for securing Big Data?	CO1		Long
13	Write a short note on: Big data and Compliance	CO1		Long
14	Write a short note on: Protecting big data analytics.	CO1		Long
	Part 7: Big Data Privacy and Ethics			
15	What is big data privacy? Mention big data privacy concerns.	CO1		Long
16	Explain principles of Big data ethics.	CO1		Long
	Part 8: Big Data Analytics			
17	Describe briefly Big Data analytics.	CO1		Long
18	What are the advantages and disadvantages of Big data analytics?	CO1		Long
	Part 9: Challenges of Conventional System, Intelligent Data Analysis, Nature of Data			
19	What is conventional system? List some of the challenges of conventional systems.	CO1		Long
20	Explain intelligent data analysis.	CO1		Long
21	What is data? List the properties of data. Describe the types of data.	CO1		Long
	Part 10: Analytics Process and Tools			
22	Explain the steps involved in analytic process.	CO1		Long
23	What are the tools used for analytic processes?	CO1		Long
	Part 11: Analytics Vs Reporting, Modern Data Analytics Tools			
24	What is analysis? What is reporting? Differentiate between analysis and reporting.	CO1		Long
25	Describe modern data analysis tools.	CO1		Long
	Short Questions			
26	What do you mean by Big Data? OR How you can define the term big data?	CO1		Short
27	What are the advantages of Big data?	CO1		Short
28	What are the disadvantages of Big data?	CO1		Short

29	Explain major challenges of Big data.	CO1		Short
30	Define digital data.	CO1		Short
31	What is the strength of digital data?	CO1		Short
32	What are the different types of digital data?	CO1		Short
33	Define structured data.	CO1		Short
34	Define semi – structured data.	CO1		Short
35	What is unstructured data?	CO1		Short
36	What is Big data platform?	CO1		Short
37	What are the 5Vs of Big data? OR List the characteristics of Big data.	CO1		Short
38	Define veracity.	CO1		Short
39	What are different Big data technology?	CO1		Short
40	What are the components of Big data technology?	CO1		Short
41	What are the application of Big data? OR What comes under Big data application?	CO1		Short
42	What are Big data risks?	CO1		Short
43	Define intelligent data analysis.	CO1		Short
44	What are the tools used for data analytics?	CO1		Short
45	Explain the difference between operational and analytical system.	CO1		Short

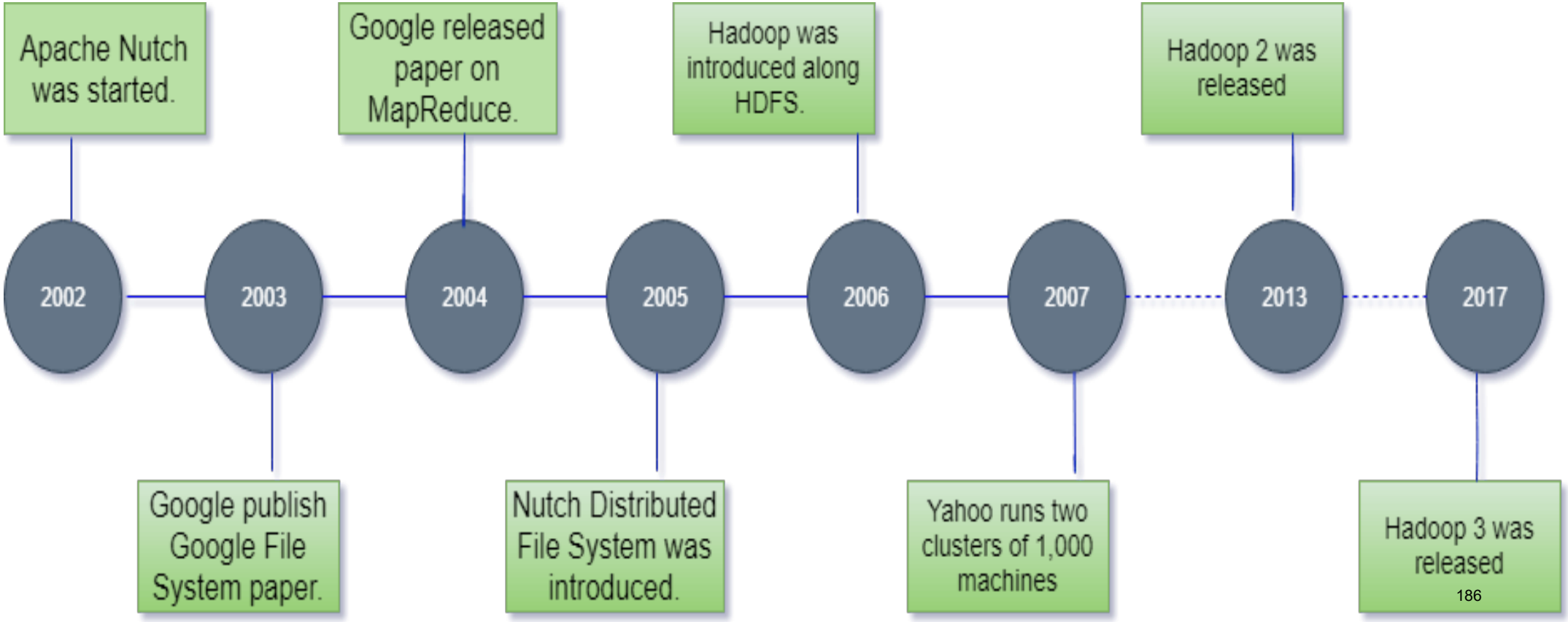
HISTORY OF HADOOP

Basic

- Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume.
- Hadoop is written in Java and is not OLAP (online analytical processing).
- It is used for batch/offline processing.
- It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

History of Hadoop

- The Hadoop was started by Doug Cutting and Mike Cafarella in 2002.



Conti...

- **2002**: Doug Cutting and Mike Cafarella started to work on a project, **Apache Nutch.** It is an open source web crawler software project. While working on Apache Nutch, they were dealing with big data. To store that data they have to spend a lot of costs which becomes the consequence of that project. This problem becomes one of the important reason for the emergence of Hadoop.
- **2003**: Google introduced a file system known as GFS (Google file system). It is a proprietary distributed file system developed to provide efficient access to data.

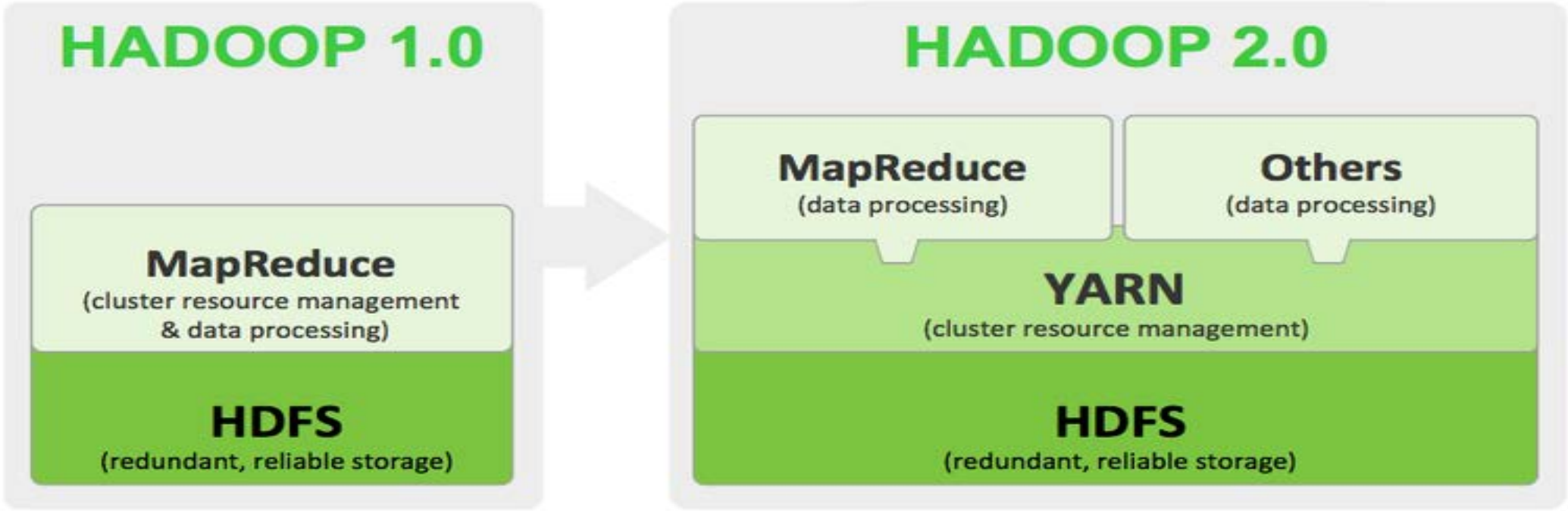
Conti...

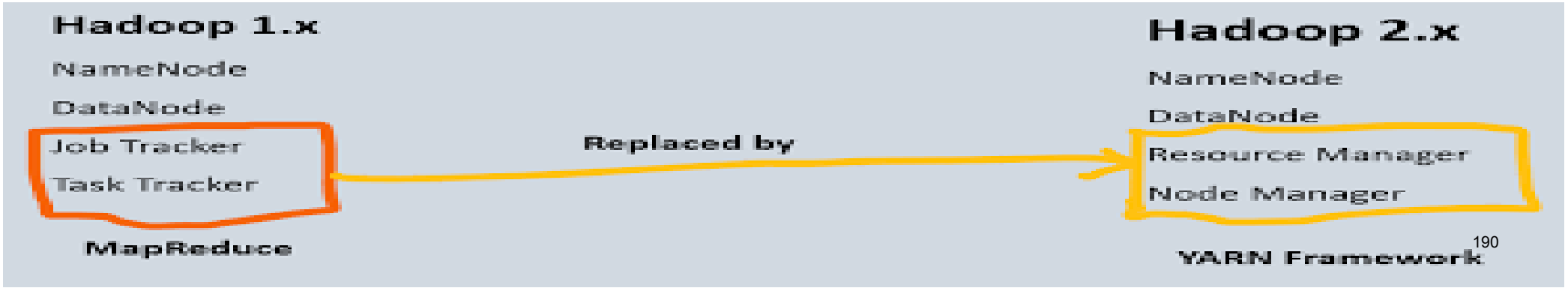
- **2004**: Google released a white paper on Map Reduce. This technique simplifies the data processing on large clusters.
- **2005**: Doug Cutting and Mike Cafarella introduced a new file system known as NDFS (Nutch Distributed File System). This file system also includes Map reduce.
- **2006**: Doug Cutting quit Google and joined Yahoo. On the basis of the Nutch project, Dough Cutting introduces a new project Hadoop with a file system known as HDFS (Hadoop Distributed File System). Hadoop first version 0.1.0 released in this year. Doug Cutting gave named his project Hadoop after his son's toy elephant.

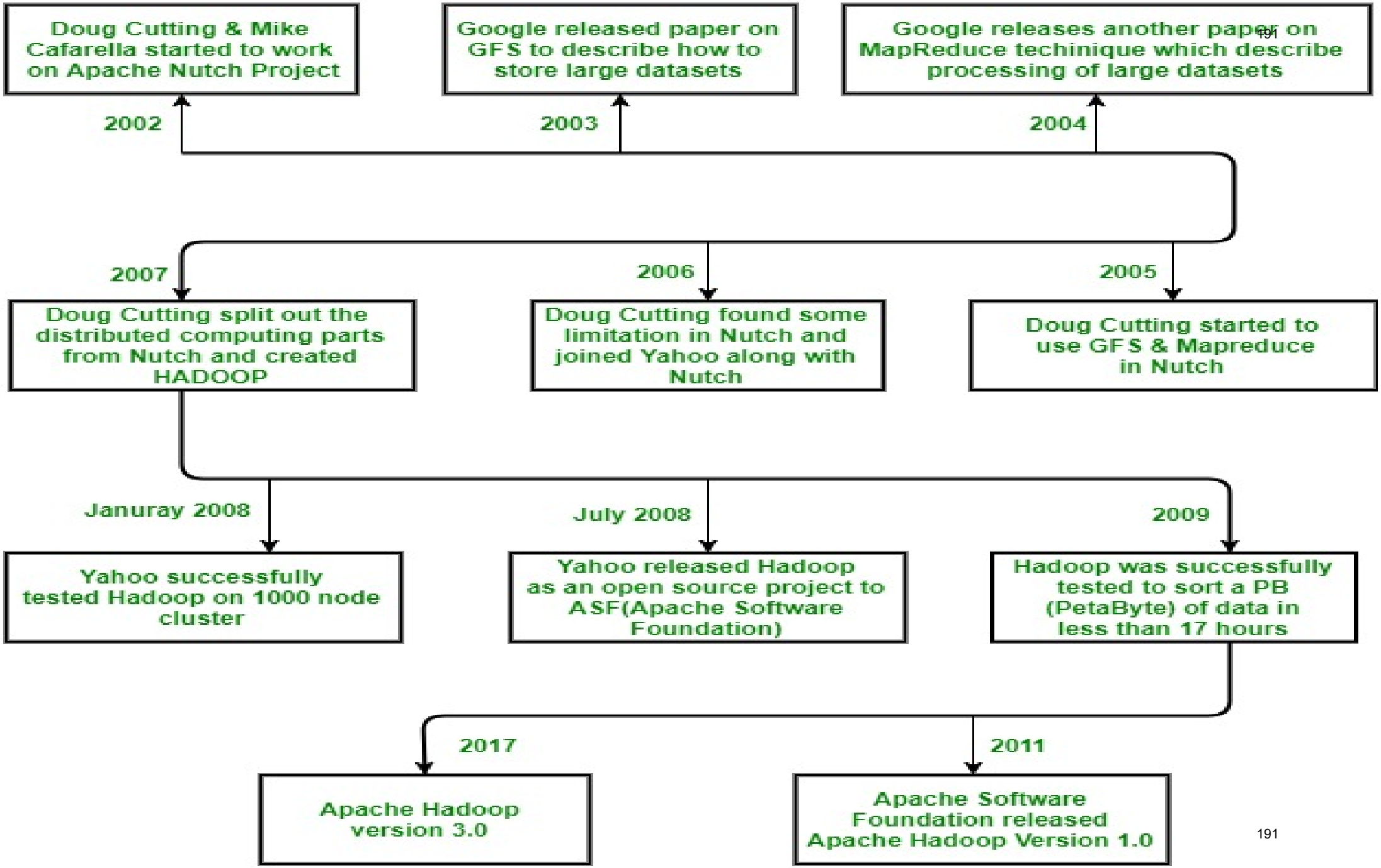
Conti...

- **2007**: Yahoo runs two clusters of 1000 machines.
- **2008**: Hadoop became the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds.
- **2013**: Hadoop 2.2 was released.
- **2017**: Hadoop 3.0 was released.

Conti…


HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)


Hadoop 1.x
NameNode
DataNode
Job Tracker
Task Tracker
MapReduce
Replaced by
Hadoop 2.x
NameNode
DataNode
Resource Manager
Node Manager
YARN Framework


Doug Cutting & Mike Cafarella started to work on Apache Nutch Project
Google released paper on GFS to describe how to store large datasets
Google releases another paper on MapReduce techinique which describe processing of large datasets
2002
2003
2004
2007
2006
2005
Doug Cutting split out the distributed computing parts from Nutch and created HADOOP
Doug Cutting found some limitation in Nutch and joined Yahoo along with Nutch
Doug Cutting started to use GFS & Mapreduce in Nutch
Januray 2008
July 2008
2009
Yahoo successfully tested Hadoop on 1000 node cluster
Yahoo released Hadoop as an open source project to ASF(Apache Software Foundation)
Hadoop was successfully tested to sort a PB (PetaByte) of data in less than 17 hours
2017
2011
Apache Hadoop version 3.0
Apache Software Foundation released Apache Hadoop Version 1.0

Conti...

Year	Event
2003	Google released the paper, Google File System (GFS).
2004	Google released a white paper on Map Reduce.
2006	•Hadoop introduced. •Hadoop 0.1.0 released. •Yahoo deploys 300 machines and within this year reaches 600 machines.
2007	•Yahoo runs 2 clusters of 1000 machines. •Hadoop includes HBase.

Conti…

Year	Event
2008	•YARN JIRA opened •Hadoop becomes the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds. •Yahoo clusters loaded with 10 terabytes per day. •Cloudera was founded as a Hadoop distributor.
2009	•Yahoo runs 17 clusters of 24,000 machines. •Hadoop becomes capable enough to sort a petabyte. •MapReduce and HDFS become separate subproject.

Conti...

Year	Event
2010	•Hadoop added the support for Kerberos. •Hadoop operates 4,000 nodes with 40 petabytes. •Apache Hive and Pig released.
2011	•Apache Zookeeper released. •Yahoo has 42,000 Hadoop nodes and hundreds of petabytes of storage.

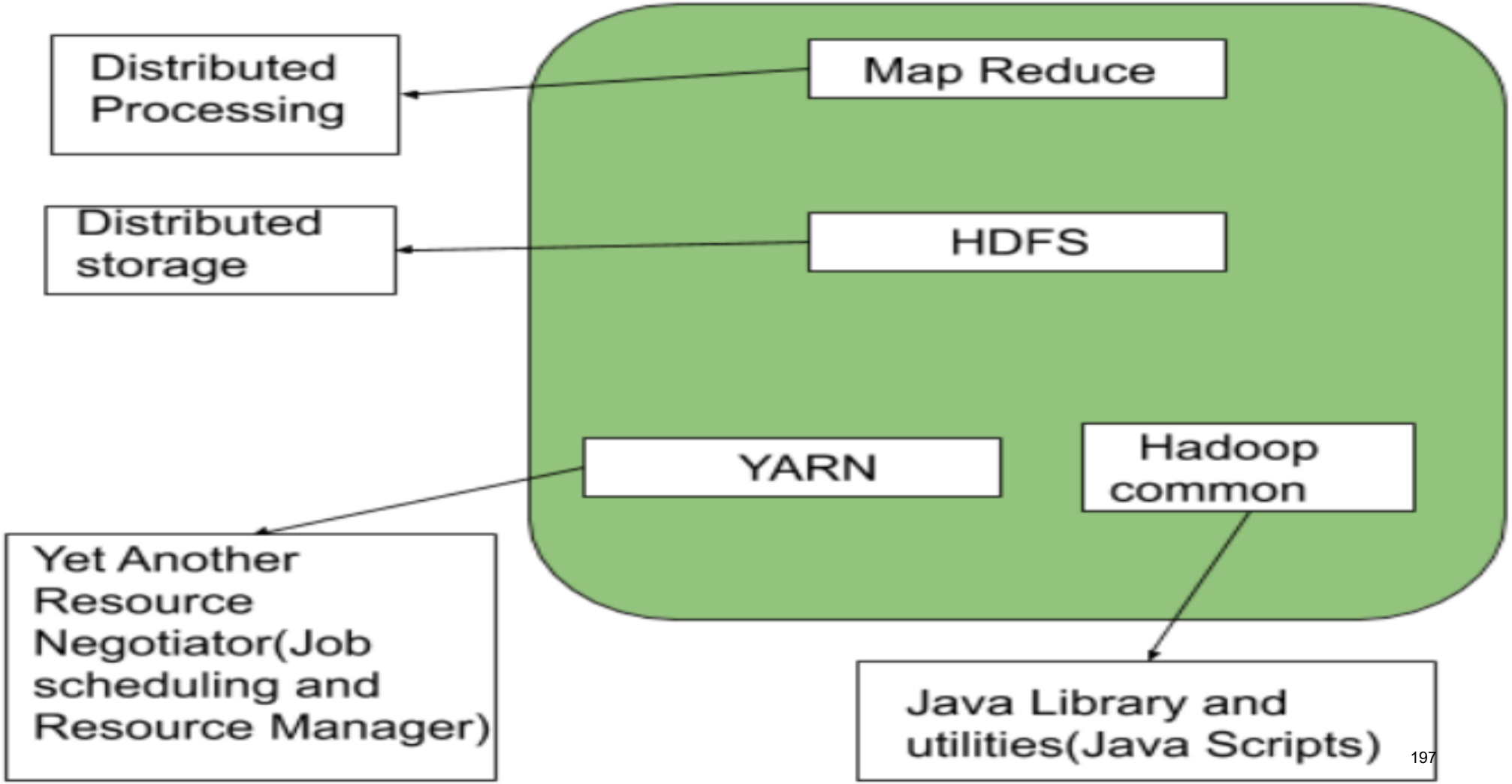
Conti...

Year	Event
2012	Apache Hadoop 1.0 version released.
2013	Apache Hadoop 2.2 version released.
2014	Apache Hadoop 2.6 version released.
2015	Apache Hadoop 2.7 version released.
2017	Apache Hadoop 3.0 version released.
2018	Apache Hadoop 3.1 version released.

Modules of Hadoop

- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet another Resource Negotiator
- **Map Reduce:** Map task convert input data into a Key value pair based dataset. The output of Map task consumed by reduce task to provide desired result.
- **Hadoop Common:** Java libraries used to start Hadoop

Conti...


Distributed Processing
Map Reduce
Distributed storage
HDFS
YARN
Hadoop common
Yet Another Resource Negotiator(Job scheduling and Resource Manager)
Java Library and utilities(Java Scripts)

Conti…

- **HDFS:** The files will be broken into blocks and stored in nodes over the distributed architecture.
- **YARN:** is used for job scheduling and manage the cluster.
- **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
- **Hadoop Common:** These Java libraries are used to start Hadoop and are used by other Hadoop modules.

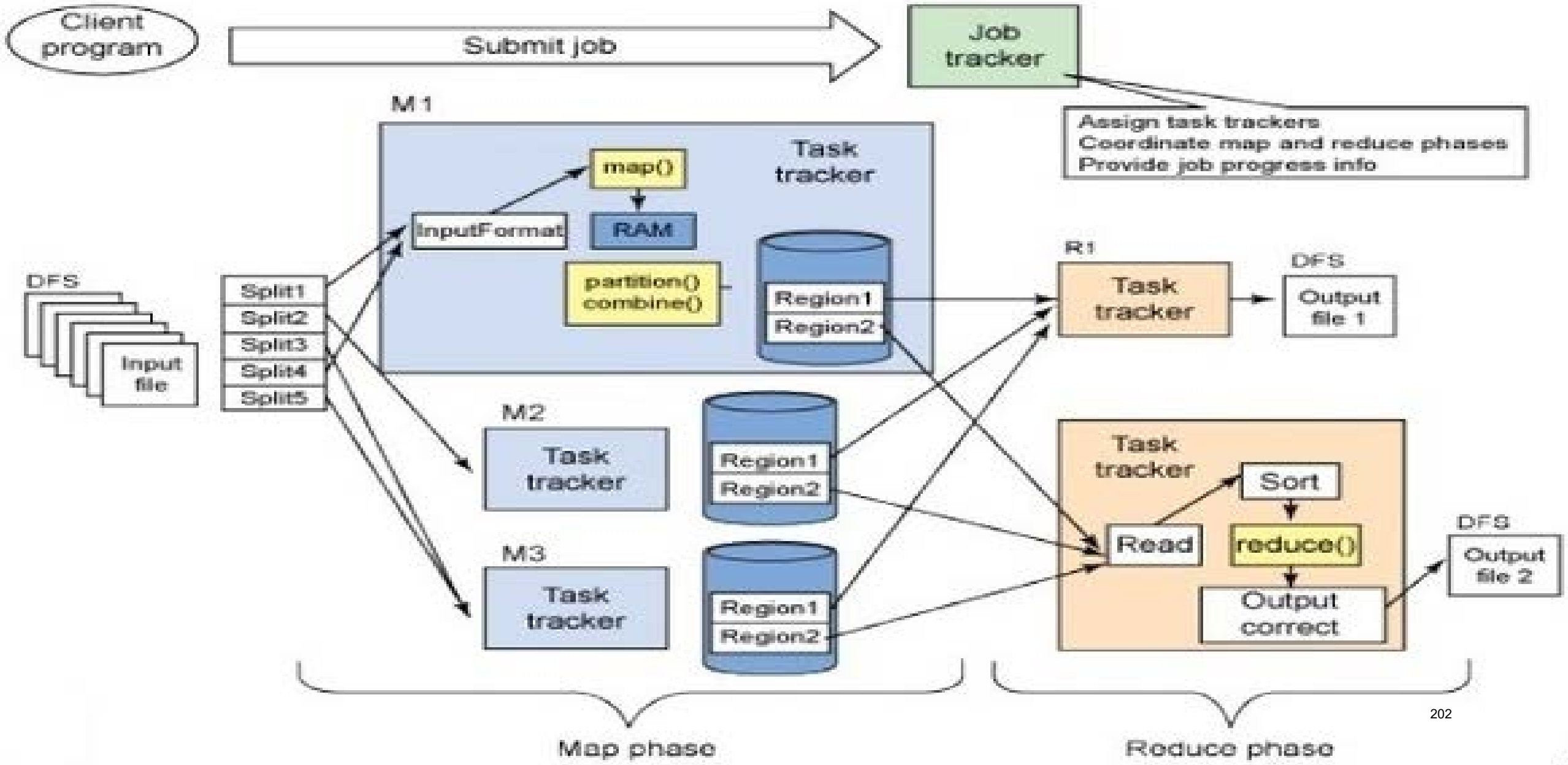
THANK YOU

HADOOP ARCHITECTURE

Hadoop Architecture

- Hadoop architecture is a package of the file system, MapReduce engine and the HDFS.
- A Hadoop cluster consists of a single master and multiple slave nodes.
- Master node includes: Job Tracker, Task Tracker, NameNode, and DataNode
- Slave node includes: DataNode and TaskTracker.

Hadoop Architecture



Advantages of Hadoop

Fast

- In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval.
- Even the tools to process the data are often on the same servers, thus reducing the processing time.
- It is able to process terabytes of data in minutes and Peta bytes in hours.

Scalable

- Hadoop cluster can be extended by just adding nodes in the cluster.

Conti...

Cost Effective

- Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.

Resilient to failure

- HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it.
- Normally, data are replicated thrice but the replication factor is configurable.

Apache Hadoop

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

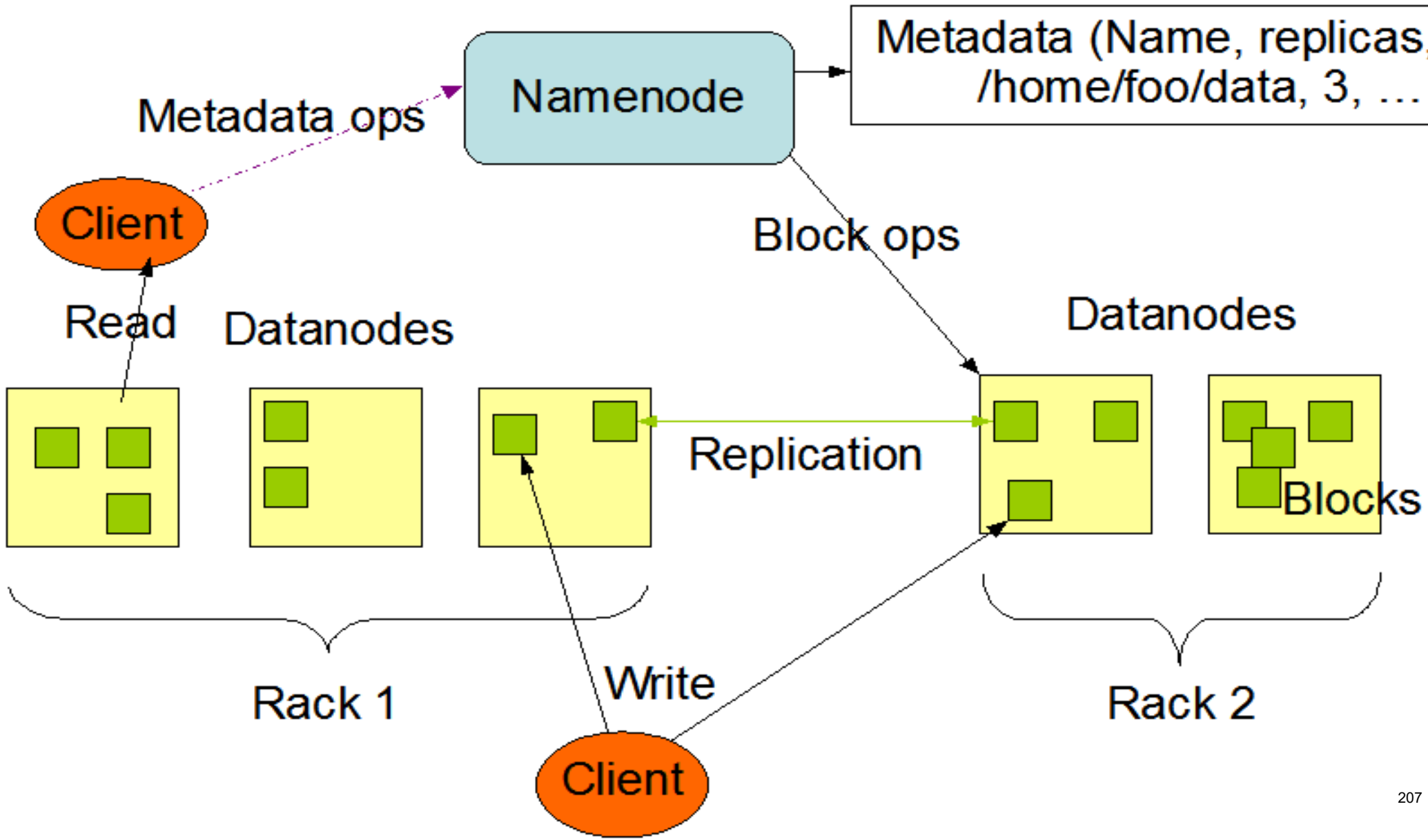
- It is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Hadoop Distributed File System (HDFS)

- HDFS is a distributed file system for Hadoop. It contains a master/slave architecture.
- This architecture consist of a single NameNode performs the role of master, and multiple DataNodes performs the role of a slave.
- Both NameNode and DataNode are capable enough to run on commodity machines.
- The Java language is used to develop HDFS. So any machine that supports Java language can easily run the NameNode and DataNode software.

HDFS Architecture



HDFS Terminology

- Namenode
- Datanode
- DFS Client
- Files/Directories
- Replication
- Blocks
- Rack-awareness

Conti...

NameNode

- It is a single master server exist in the HDFS cluster.
- As it is a single node, it may become the reason of single point failure.
- It manages the file system namespace by executing an operation like the opening, renaming and closing the files.
- It simplifies the architecture of the system.

Conti...

DataNode

- The HDFS cluster contains multiple DataNodes.
- Each DataNode contains multiple data blocks.
- These data blocks are used to store data.
- It is the responsibility of DataNode to read and write requests from the file system's clients.
- It performs block creation, deletion, and replication upon instruction from the NameNode.

Conti...

Job Tracker

- The role of Job Tracker is to accept the MapReduce jobs from client and process the data by using NameNode.
- In response, NameNode provides metadata to Job Tracker.

Task Tracker

- It works as a slave node for Job Tracker.
- It receives task and code from Job Tracker and applies that code on the file. This process can also be called as a Mapper.

Conti...

MapReduce Layer

- It comes into existence when the client application submits the MapReduce job to Job Tracker.
- In response, the Job Tracker sends the request to the appropriate Task Trackers.
- Sometimes, the TaskTracker fails or time out.
- In such a case, that part of the job is rescheduled.

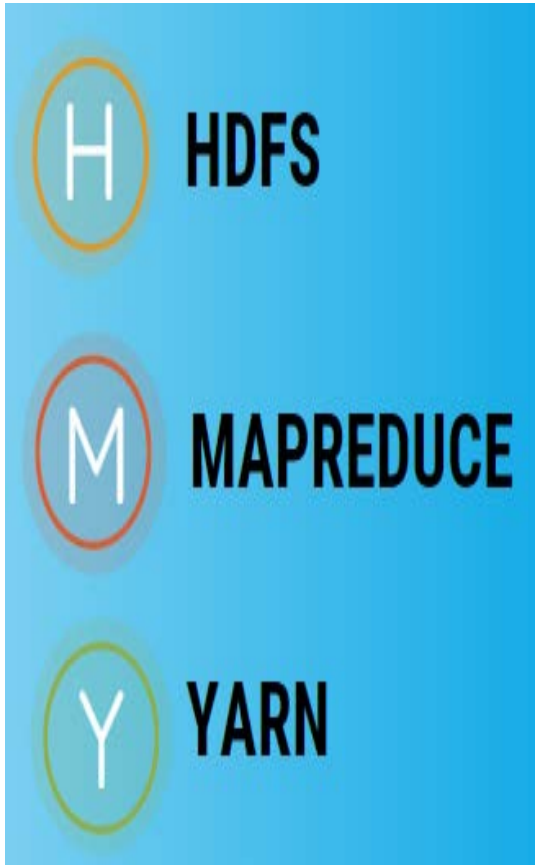
Components of Hadoop

Three major Components of Hadoop

- **HDFS:** Hadoop Distributed File System

- **Map Reduce:** Map task convert input data into a Key value pair based dataset. The output of Map task consumed by reduce task to provide desired result.

- **YARN:** Yet another Resource Negotiator



Data format in Hadoop

Input File Formats in Hadoop

1. Text/CSV Files
2. JSON Records
3. Avro Files
4. Sequence Files
5. RC Files
6. ORC Files
7. Parquet Files

Conti...

Most common formats of the Hadoop ecosystem:

Text/CSV

- A plain text file or CSV is the most common format both outside and within the Hadoop ecosystem.
- The disadvantage in the use of this format is that it does not support block compression, so the compression of a CSV file in Hadoop can have a high cost in reading.
- Text and CSV files are quite common and frequently Hadoop developers and data scientists received text and CSV files to work upon.

Conti...

SequenceFile

- The SequenceFile format stores the data in binary format.
- This format accepts compression; however, it does not store metadata and the only option in the evolution of its scheme is to add new fields at the end.
- This is usually used to store intermediate data in the input and output of MapReduce processes.

Conti...

Avro

- Avro is a row-based storage format.
- Avro files include markers that can be used to split large data sets into subsets suitable for Apache MapReduce processing.
- Avro also allows block compression in addition to its divisibility, making it a good choice for most cases when using Hadoop.
- Avro is quickly becoming the top choice for the developers due to its multiple benefits.
- You can rename, add, delete and change the data types of fields by defining a new independent schema.
- Avro files are splittable, support block compression and enjoy broad, relatively mature, tool support within the Hadoop ecosystem.

Conti...

Parquet

- Parquet is a column-based (column-based) binary storage format that can store nested data structures.
- This format is very efficient in terms of disk input / output operations when the necessary columns to be used are specified.
- This format is very optimized for use with Cloudera Impala.
- Parquet file is another columnar file given by Hadoop founder Doug Cutting during his Trevni project. Like another Columnar file RC & ORC, Parquet also enjoys the features like compression and query performance benefits but is generally slower to write than non-columnar file formats.

Conti...

RCFile (Record Columnar File)

- RCFile is a columnar format that divides data into groups of rows, and inside it, data is stored in columns.
- This format does not support the evaluation of the scheme and if you want to add a new column it is necessary to rewrite the file, slowing down the process.
- RC file was the first columnar file in Hadoop and has significant compression and query performance benefits.
- But it doesn't support schema evaluation and if you want to add anything to RC file you will have to rewrite the file. Also, it is a slower process.

Conti...

ORC (Optimized Row Columnar)

- ORC is considered an evolution of the RCFile format and has all its benefits alongside with some improvements such as better compression, allowing faster queries.
- This format also does not support the evolution of the scheme.
- ORC is the compressed version of RC file and supports all the benefits of RC file with some enhancements like ORC files compress better than RC files, enabling faster queries.
- But it doesn’t support schema evolution. Some benchmarks indicate that ORC files compress to be the smallest of all file formats in Hadoop.

Conti...

JSON Records

- JSON file is **a file that stores simple data structures and objects in JavaScript Object Notation (JSON) format**, which is a standard data interchange format. It is primarily used for transmitting data between a web application and a server.
- In the case of JSON files, metadata is stored and the file is also splittable but again it also doesn't support block compression.
- The only issue is there is not much support in Hadoop for JSON file but thanks to the third party tools which helps a lot.

Conti...

Sequence File

- Sequence file stores data in binary format and has a similar structure to CSV file with some differences.
- It also doesn’t store metadata and so only schema evolution option is appending new fields but it supports block compression.
- Due to complexity, sequence files are mainly used in flight data as an intermediate storage.

THANK YOU

HADOOP ECHO SYSTEM

Scaling Out

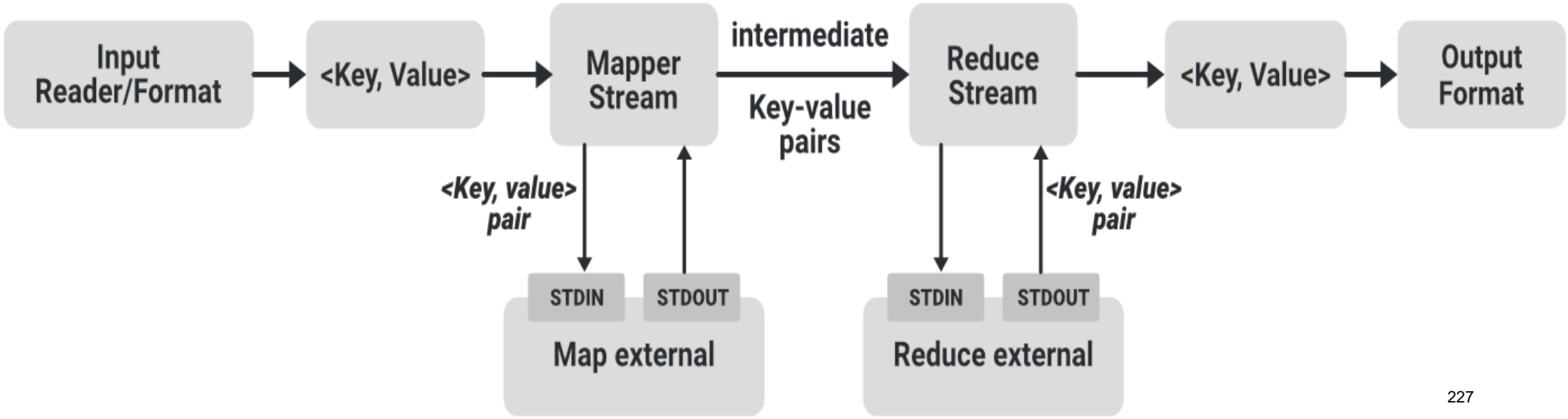
- Scale out is the process of selling off portions of the total held shares while the price increases.
- To scale out means to get out of a position (e.g., to sell) in increments as the price climbs.
- Scaling out of a stock lets an investor reduce exposure to a position when momentum seems to be slowing. This strategy allows the investor to take profits while the price is increasing, rather than trying to time the peak price. If the actual value continues to increase, however, the investor could be selling a winner too early.

Conti...

- To scale out of a trade is to incrementally sell a portion of one's long position as the price rises.
- This profit-taking strategy can help reduce the risk of mistiming the market's high; however, it could also risk selling shares too early in a rising market and limit potential upside.
- Scaling out is seen as a risk-averse strategy that can reward investors if the price of a stock subsequently reverses trend and falls.

Hadoop Streaming

- Hadoop streaming is a utility that comes with the Hadoop distribution. This utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.



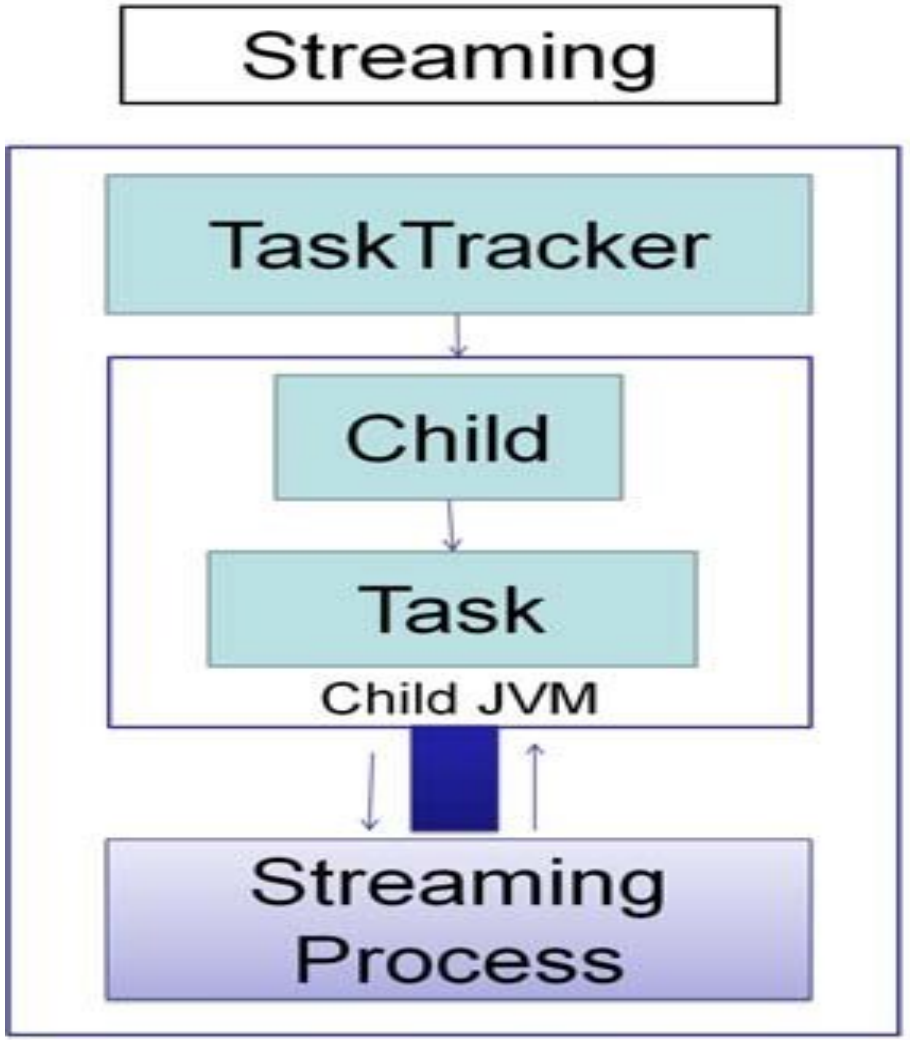
Conti…

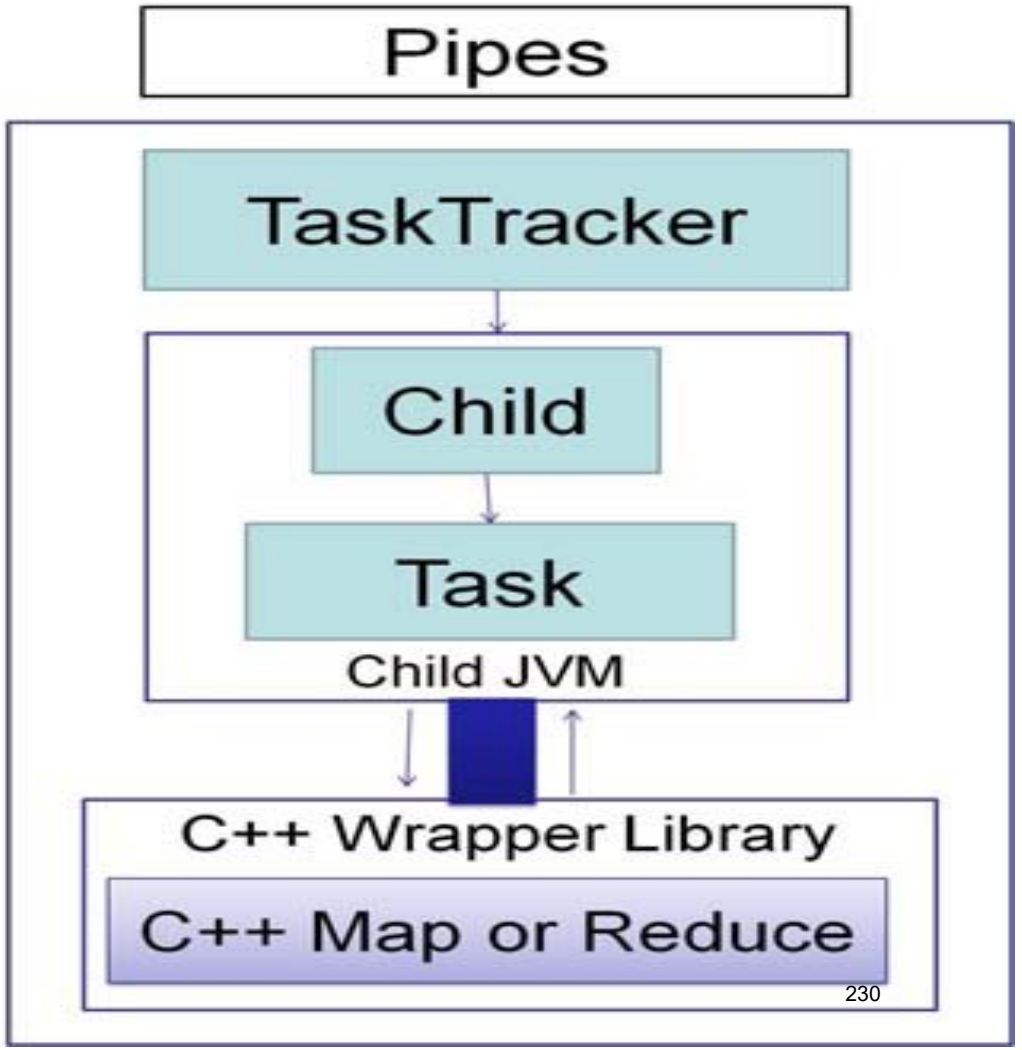
- For Hadoop streaming, we are considering the word-count problem.
- Any job in Hadoop must have two phases: mapper and reducer.
- We have written codes for the mapper and the reducer in python script to run it under Hadoop.

Hadoop Pipes

- Hadoop Pipes is the name of the C++ interface to Hadoop MapReduce.
- Unlike Streaming, which uses standard input and output to communicate with the map and reduce code, Pipes uses sockets as the channel over which the tasktracker communicates with the process running the C++ map or reduce function.

Conti…


Streaming
TaskTracker
Child
Task
Child JVM
Streaming Process


Pipes
TaskTracker
Child
Task
Child JVM
C++ Wrapper Library
C++ Map or Reduce

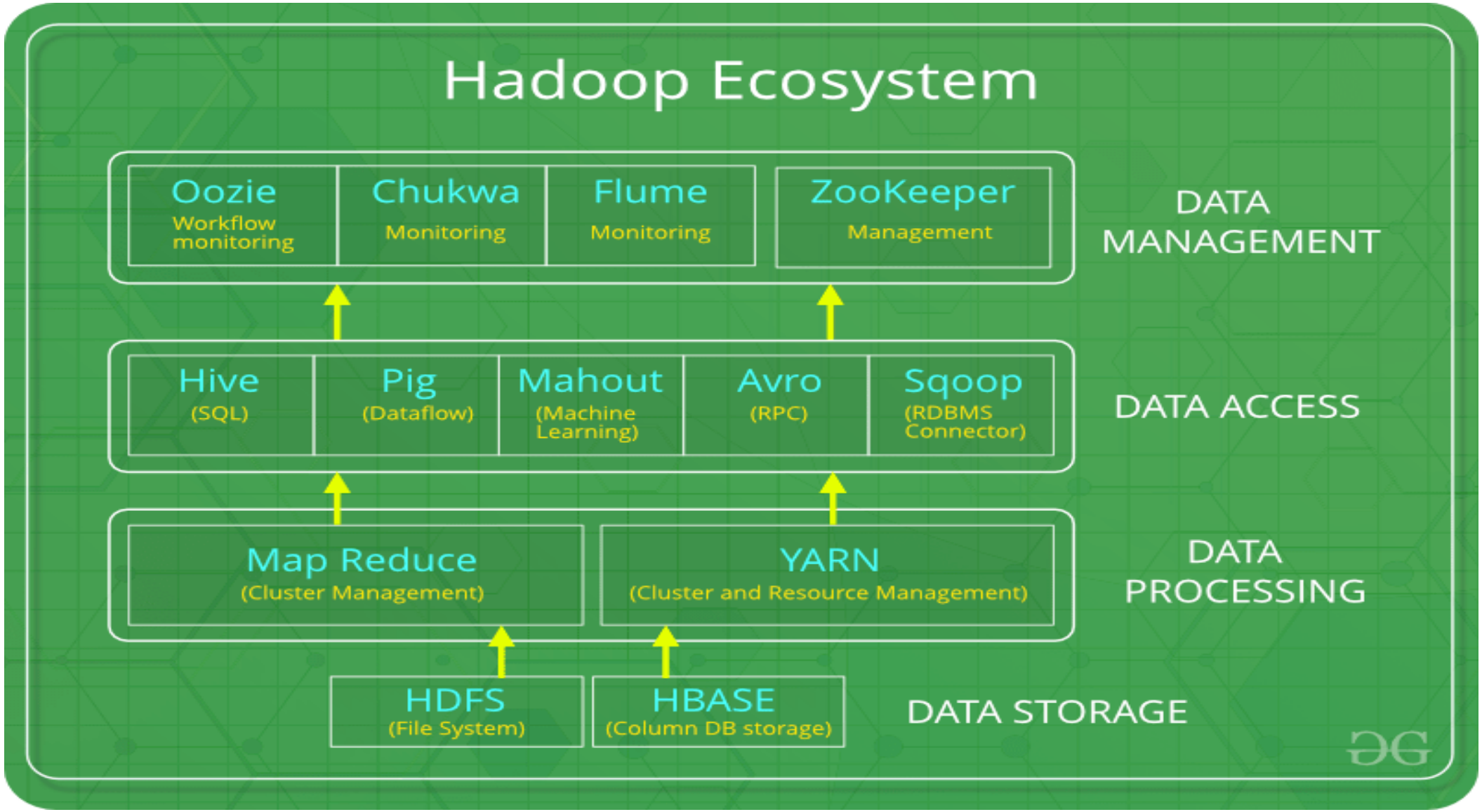
Hadoop Echo System

- Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems.
- It includes Apache projects and various commercial tools and solutions.
- There are *four major elements of Hadoop* i.e. HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these major elements.
- All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

Conti...

Components of Hadoop ecosystem:

- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet Another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing
- **PIG, HIVE:** Query based processing of data services
- **HBase:** NoSQL Database
- **Mahout, Spark MLLib:** Machine Learning algorithm libraries
- **Solar, Lucene:** Searching and Indexing
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling


Hadoop Ecosystem
Oozie
Workflow monitoring
Chukwa
Monitoring
Flume
Monitoring
ZooKeeper
Management
DATA MANAGEMENT
Hive
(SQL)
Pig
(Dataflow)
Mahout
(Machine Learning)
Avro
(RPC)
Sqoop
(RDBMS Connector)
DATA ACCESS
Map Reduce
(Cluster Management)
YARN
(Cluster and Resource Management)
DATA PROCESSING
HDFS
(File System)
HBASE
(Column DB storage)
DATA STORAGE

THANK YOU

ABES ENGINEERING COLLEGE, GHAZIABAD

Unit-2

QUESTION BANK

SUBJECT NAME: BIG DATA
UNIT: 2

Q.No	Description	CO	year	Long-short
	Part 1: History of Hadoop, Apache Hadoop			
1	Describe the history of Hadoop.	CO2		Long
2	Write short note on Apache Hadoop.	CO2		Long
	Part 2: The Hadoop Distributed File System, Components of Hadoop, Data Format			
3	What is Hadoop Distributed File System (HDFS)? How does HDFS work? Also explain the features of HDFS.	CO2		Long
4	Describe the goals of HDFS.	CO2		Long
5	What are the benefits of using HDFS?	CO2		Long
6	What are various components of the Hadoop?	CO2		Long
7	What are the various data formats used in Hadoop?	CO2		Long
	Part 2: Analyzing Data with Hadoop			
8	Give reasons why Hadoop can be considered a helpful tool to analyze the big data?	CO2		Long
9	Which tools are used to analyze data using Hadoop?	CO2		Long
	Part 4: Scaling Out, Hadoop Streaming, Hadoop Pipes, Hadoop Echo System			
10	Describe the term scaling out. OR Differentiate “Scale up and scale out”. Explain with an example how Hadoop uses scale out feature to improve the performances.	CO2		Long
11	What is Hadoop Streaming?	CO2		Long
12	Write short note on Hadoop Pipes.	CO2		Long
13	Describe briefly Hadoop Ecosystem.	CO2		Long
	Part 5: Map Reduce, Map Reduce Framework and Basics, How Map Reduce Works			
14	Write short note on MapReduce.	CO2		Long

ABES ENGINEERING COLLEGE, GHAZIABAD

15	Explain different phases of MapReduce. OR What is MapReduce? Explain the stages of MapReduce program execution.	CO2		Long
16	Describe how MapReduce works. OR Explain in detail about MapReduce workflows.	CO2		Long
Part 3: Developing a Map Reduce Application, Unit Test with MR Unit, Test Data and Local Tests, Anatomy of a Map Reduce Job Run.				
17	Give the phases of developing a MapReduce application.	CO2		Long
18	How to write a program for MapReduce application?	CO2		Long
19	Write a short note on: Unit tests with MRUnit.	CO2		Long
20	Write a short note on: Test data and local tests in MapReduce.	CO2		Long
21	How does Hadoop executes a MapReduce program?	CO2		Long
22	Explain anatomy of job run in classic MapReduce (MapReduce 1).	CO2		Long
23	How the scalability shortcomings of classic MapReduce is overcome by YARN?	CO2		Long
24	Explain anatomy of job run in YARN (MapReduce2).	CO2		Long
Part 4: Failures, Job Scheduling, Shuffle and Sort, Task Execution, MapReduce Types, Input Formats, Output Formats, Map Reduce Features, Real – World Map Reduce				
25	What are various failures in classic MapReduce (MapReduce1)?	CO2		Long
26	What are various failures in YARN (MapReduce2)?	CO2		Long
27	What are the types of schedulers in MapReduce?	CO2		Long
28	What are the advantages and disadvantages of different types of scheduler?	CO2		Long
29	What is shuffle and sort in Hadoop MapReduce?	CO2		Long
30	Write a short note on: i) Speculative Execution ii) Task JVM Reuse iii) Skipping Bad Records	CO2		Long
31	What are the different types of input formats in Hadoop?	CO2		Long
32	What are the different types of output formats in Hadoop?	CO2		Long
33	What are the various advanced features of MapReduce?	CO2		Long

ABES ENGINEERING COLLEGE, GHAZIABAD

Short Questions				
34	What do you mean by Hadoop?	CO2		Short
35	What is Hadoop architecture?	CO2		Short
36	What are the layers of Hadoop?	CO2		Short
37	What are the advantages of Hadoop?	CO2		Short
38	What are the disadvantages of Hadoop?	CO2		Short
39	Define Apache Hadoop.	CO2		Short
40	Define Hadoop Distributed File System.	CO2		Short
41	What are the features of HDFS?	CO2		Short
42	What are the goals of HDFS?	CO2		Short
43	What are the components of Hadoop data format?	CO2		Short
44	What is Hadoop streaming?	CO2		Short
45	Define Hadoop pipes? OR State the usage of Hadoop pipes. OR State the purpose of Hadoop pipes.	CO2		Short
46	Define Hadoop Ecosystem.	CO2		Short
47	What are the elements of Hadoop?	CO2		Short
48	What is MapReduce?	CO2		Short
49	What are the different phases of MapReduce?	CO2		Short
50	What are the types of task used in Hadoop?	CO2		Short
51	What are the types of schedulers in Hadoop?	CO2		Short
52	What are the features of MapReduce?	CO2		Short
53	List down the tools related with Hadoop.	CO2		Short