

Big Data Applications and Analytics - Overview

Topic Outline

- **The lesson covers:**
 - Concepts and Terminology
 - Big Data Applications
 - History of Big Data
 - Evolution from 3Vs, 4Vs, 5Vs, to 6Vs of Big Data
 - Analytics Stack
 - Transition from Big Data Vs to Ms

Learning Outcomes

At the end of this topic, you should be able to:

- Describe the concepts related to Big Data, including big data technologies, analytics, NoSQL, and applications.

Big Data

Big Data

Big data refers to the vast and complex sets of data that exceed the capabilities of traditional data processing tools.

Traditional data processing tools have several capabilities: -

Structured Data Handling

They are well-suited for structured data like relational databases.

Batch Processing

These tools are efficient in batch processing tasks.

Limited Scalability

Traditional tools may not scale well for large datasets.

Standard Query Languages

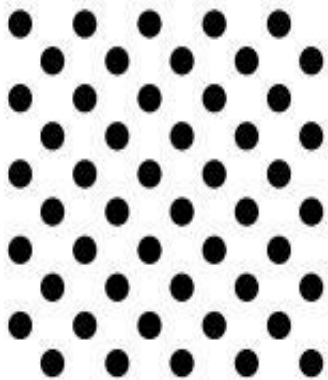
They often use standard query languages like SQL.

Big Data

- Big data is used to extract valuable insights and patterns through advanced analytics and technology.
- It informs decision-making, improves operations, and gains a competitive advantage in various fields, such as business, science, and technology.

Big Data V's

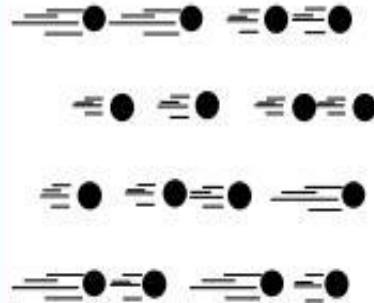
Volume



Data at Rest

Terabytes to exabytes of existing data to process

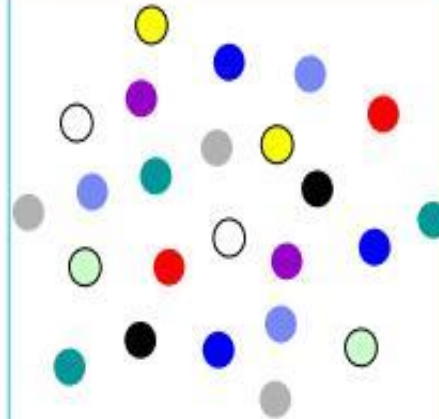
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

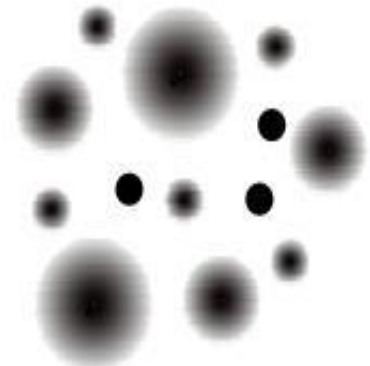
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

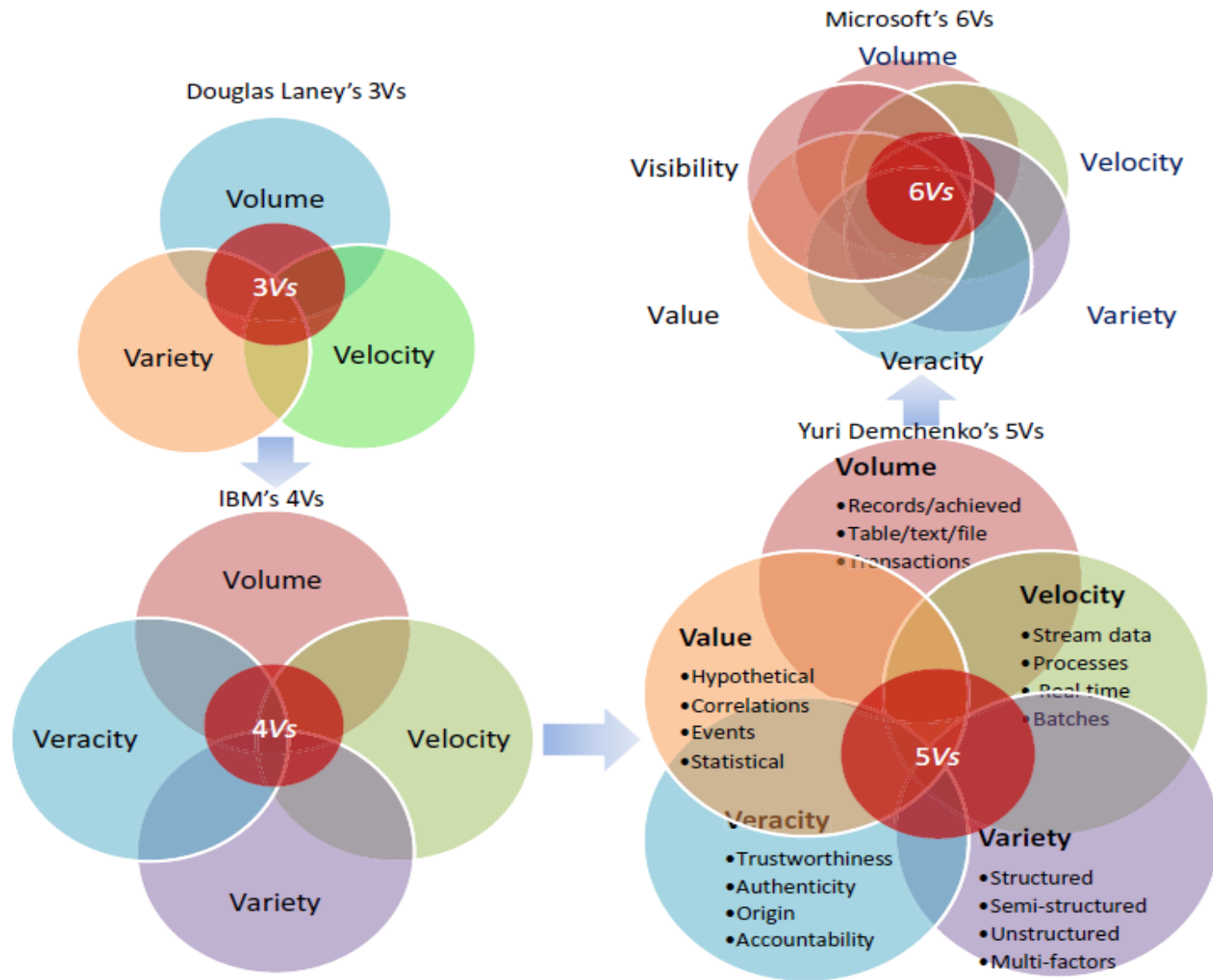
Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

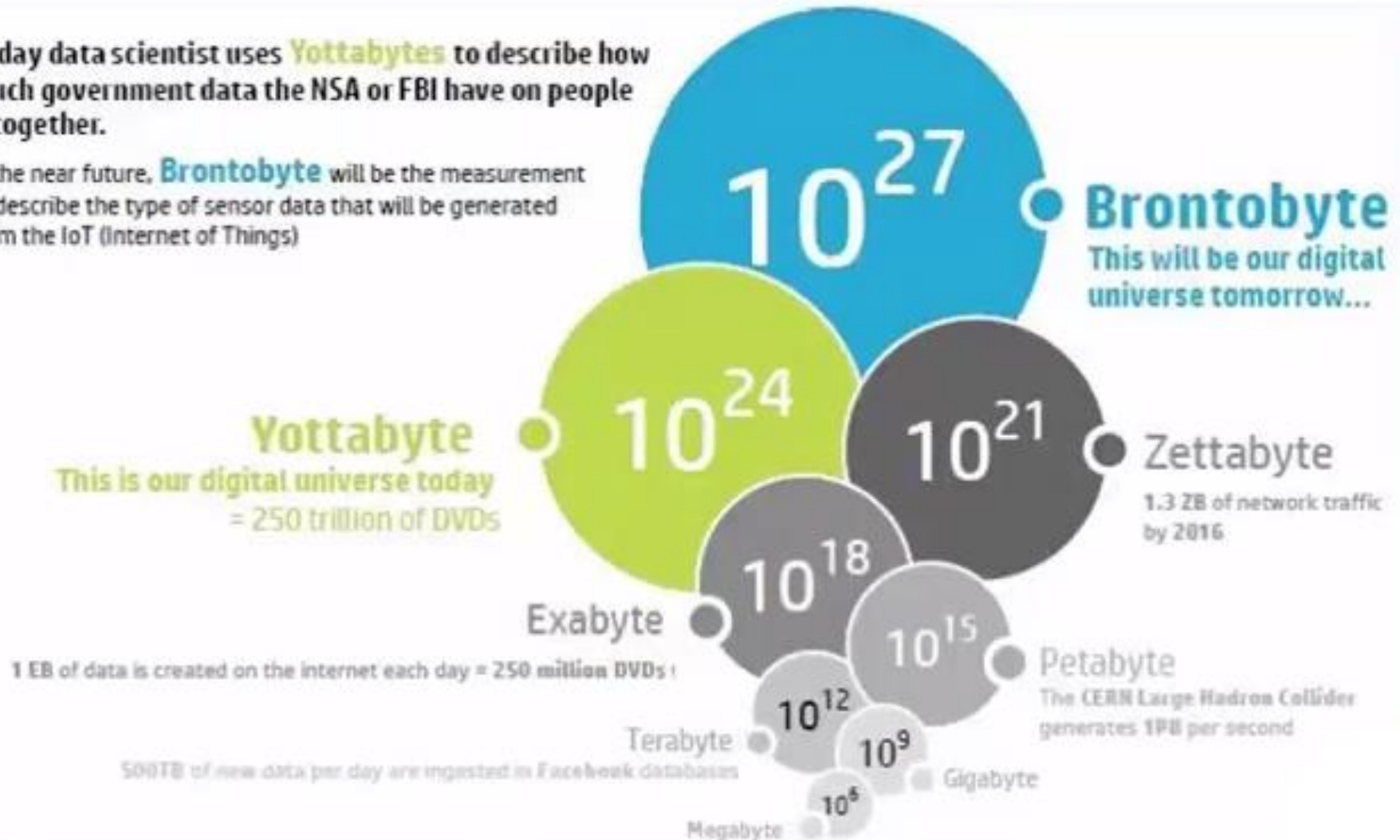
From 3Vs, 4Vs, 5Vs, and 6Vs big data



History of Big Data

Today data scientist uses **Yottabytes** to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)



Big Data Analytics

Big data analytics is the process of examining **large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences** and other useful business information.

Big Data Application

- Gaming
- Image Recognition
- Speech Recognition
- Recommender Systems
- Internet Search
- Digital Advertisements (Targeted Advertising and Re-targeting)
- Price Comparison Websites
- Airline Route Planning
- Delivery Logistics
- Fraud and Risk Detection

Application

How Big Data is Generated

Gaming

- Description of data sources and types for gaming.

Image Recognition

- How image data is collected and analyzed for recognition.

Speech Recognition

- Sources of speech data and its use in recognition systems.

Recommender Systems

- Data sources and algorithms for recommendation engines.

Internet Search

- How user queries and web data contribute to search results.

Digital Advertisements

- Use of user behavior and ad impressions for targeting.

Price Comparison Websites

- Data sources for product prices and comparisons.

Airline Route Planning

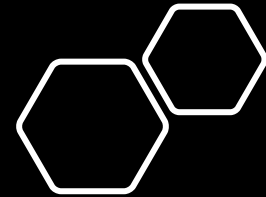
- Utilization of flight and passenger data for planning.

Delivery Logistics

- Data sources and optimization methods for delivery.

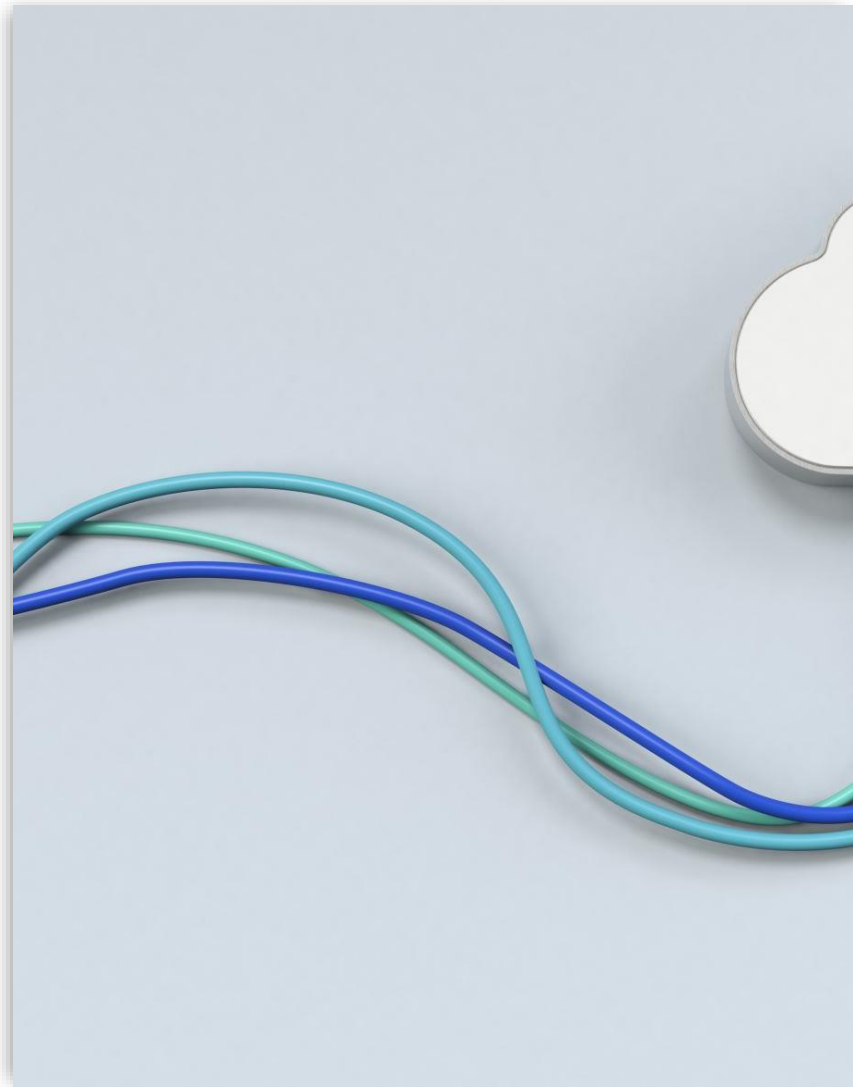
Fraud and Risk Detection

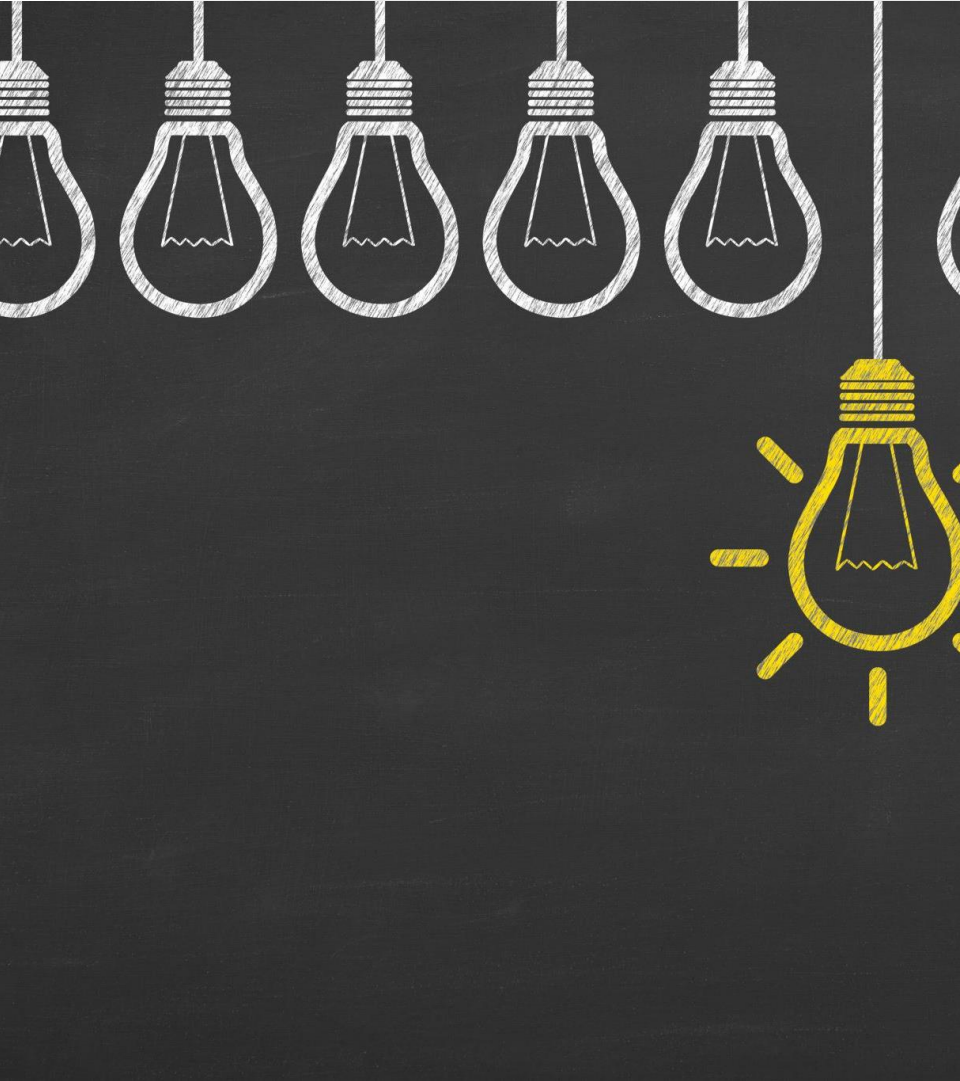
- How various data sources are used to detect fraud and risk.



Cloudera

- Cloudera offers a comprehensive platform for data management and analytics.
- It provides a distribution of the Apache Hadoop ecosystem.
- Cloudera supports various related big data tools and technologies.
- The company specializes in large-scale data processing and storage solutions.
- Cloudera's products contribute to efficient data processing and analytics.





Cloudera QuickStart VM

- Cloudera QuickStart VM is a pre-configured virtual machine for big data.
- It simplifies setting up a development environment for learning and testing.
- Comes with sample data and tutorials for hands-on experience.
- Ideal for exploring big data tools and analytics without complex setup.
- Useful for developers and data analysts to experiment with Cloudera's platform.

The Apache Software Foundation™

cloudera

Hadoop is so much more
than just Hadoop.



Big Data




Big Data



Big Data



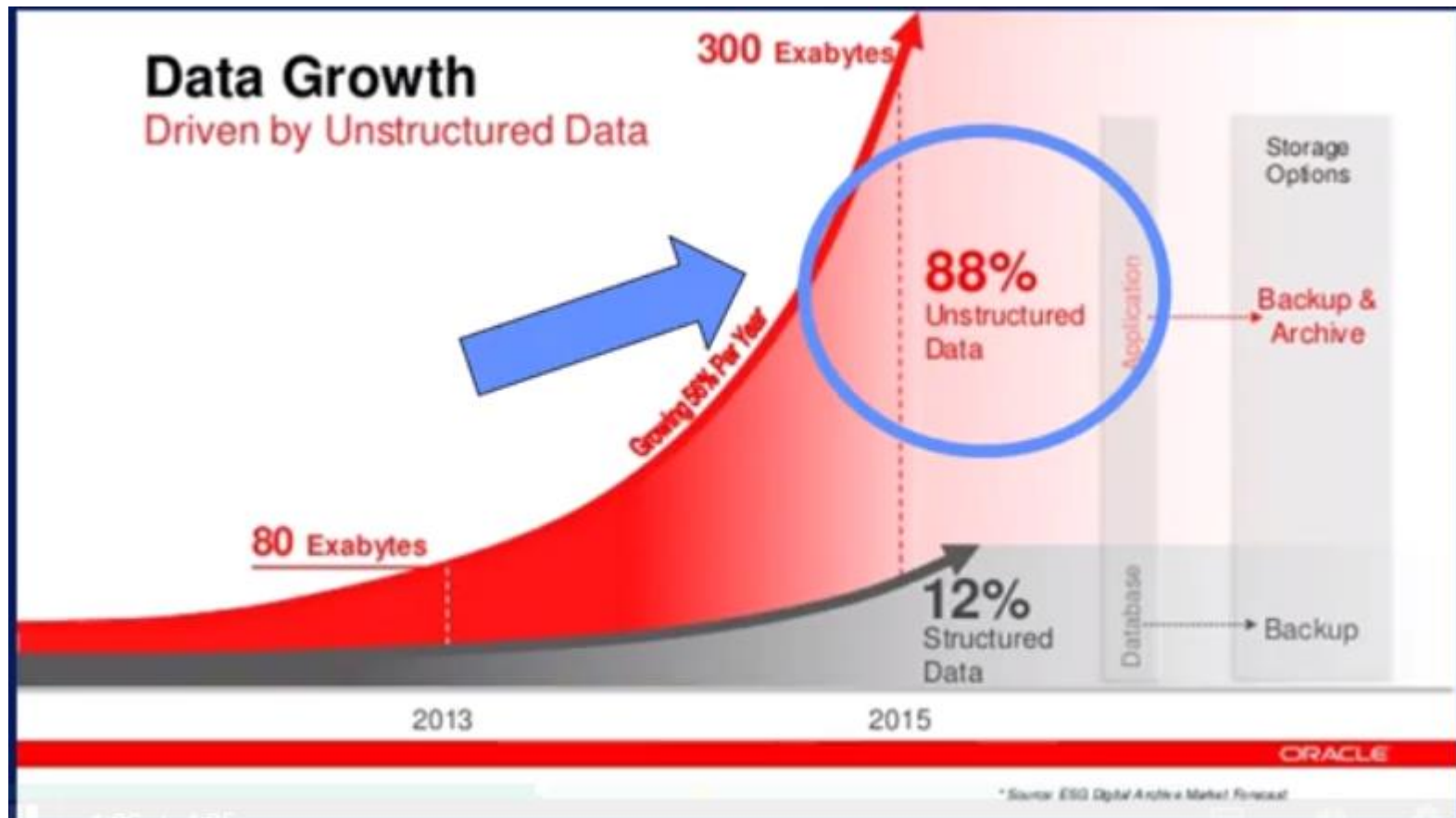
Big Data



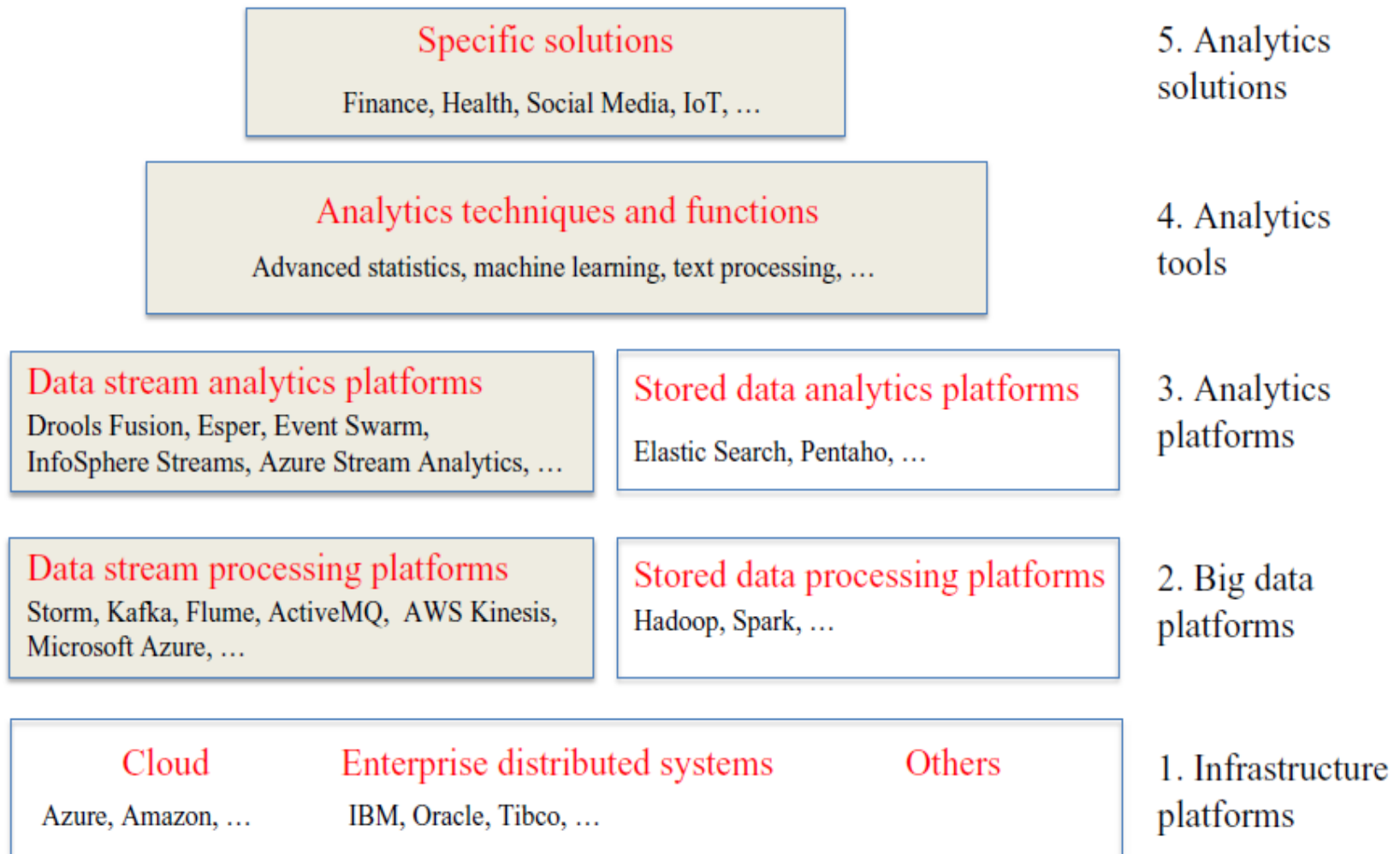
Data is moving in from a variety of sources – how can we keep up?

Big Data

What is changing in the land of Big Data ?



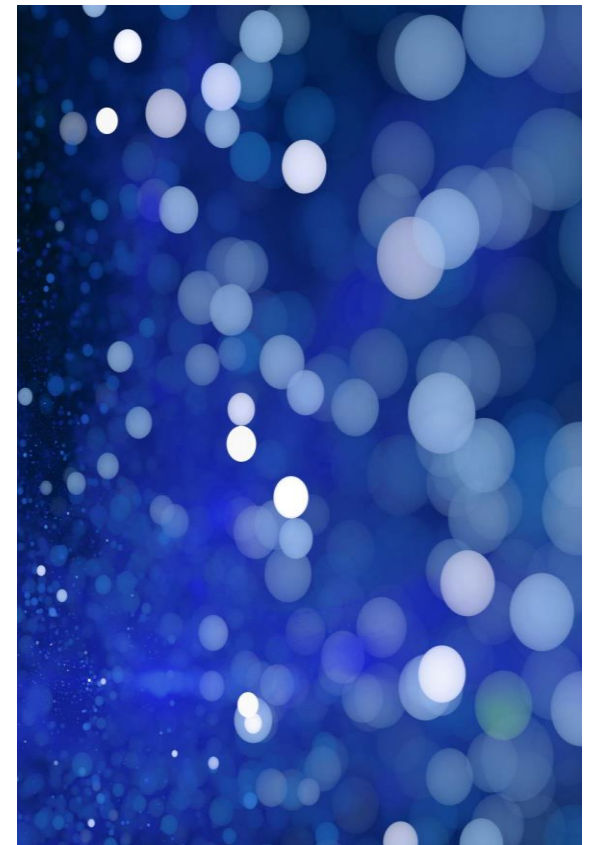
Analytics stack



Concepts and Terminology

To begin, we must define and understand several fundamental concepts and terms.

Big Data	Hadoop	MapReduce	NoSQL
Machine Learning	Data Mining	Predictive Analytics	Data Warehouse
ETL (Extract, Transform, Load)	IoT (Internet of Things)	Streaming Analytics	Data Visualization
Cloud Computing	Data Governance	Data Privacy	Scalability
	Real-Time Analytics	Business Intelligence (BI)	



Concepts and Terminology

- **Big Data:** Refers to the vast volume, variety, and velocity of data that traditional methods struggle to manage and analyze.
- **Hadoop:** An open-source framework for distributed storage and processing of big data.
- **MapReduce:** A programming model for processing and generating large datasets that Hadoop uses.
- **NoSQL:** A class of database systems that don't rely on traditional SQL-based relational database management systems.

Concepts and Terminology

- **Machine Learning:** The field of study that gives computers the ability to learn without being explicitly programmed.
- **Data Mining:** The process of discovering patterns, trends, and insights in large datasets.
- **Predictive Analytics:** The use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes.
- **Data Warehouse:** A central repository for storing and managing data that's used for reporting and analysis.

Concepts and Terminology

- **ETL (Extract, Transform, Load):** The process of collecting data from various sources, transforming it for analysis, and loading it into a data warehouse.
- **IoT (Internet of Things):** The network of physical devices, vehicles, and other objects embedded with sensors, software, and network connectivity, enabling them to collect and exchange data.
- **Streaming Analytics:** The analysis of data in motion, allowing real-time insights and decision-making.
- **Data Visualization:** The representation of data in graphical or pictorial format to help users understand and interpret information.

Concepts and Terminology

- **Cloud Computing:** The delivery of computing services (such as servers, storage, databases, networking, software, and analytics) over the internet.
- **Data Governance:** The process of managing data availability, usability, consistency, and data quality.
- **Data Privacy:** Protecting sensitive and personal information from unauthorized access and usage.
- **Scalability:** The ability of a system to handle growing amounts of data or users without compromising performance.

Concepts and Terminology

- **Real-Time Analytics:** The process of analyzing data as it's generated or collected, allowing immediate action.
- **Business Intelligence (BI):** The technologies, processes, and tools for analyzing data and delivering actionable information to help organizations make informed decisions.

Concepts and Terminology- cont'd

- Datasets

- Collections or groups of related data are generally referred to as datasets.
- Each group or dataset member shares the same set of attributes or properties as others in the same dataset.
- Examples:
 - Tweets stored in a flat file
 - A collection of image files in a directory
 - An extract of rows from a database table stored in a CSV formatted file
 - Historical weather observations that are stored as XML files



NOSQL

- Non-relational
- Don't require schema
- Data are replicated to multiple nodes (so, identical & fault-tolerant) and can be partitioned:
 - down nodes easily replaced
 - no single point of failure
- Horizontal scalable
- Cheap, easy to implement (open-source)
- Fast key-value access

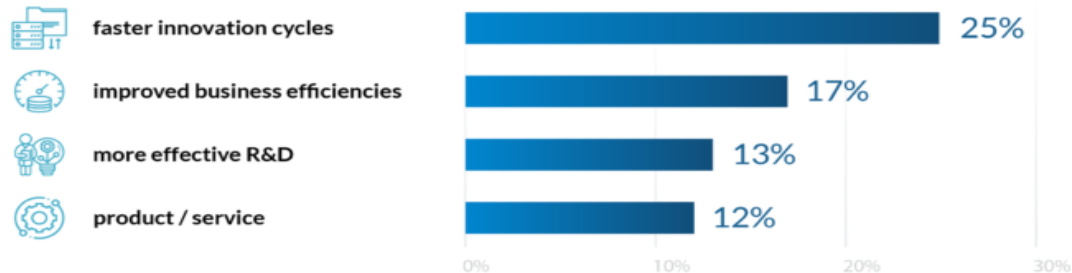
Big Data V's to M's

- Seeing competitors jumping onto the Big Data bandwagon, many organisations follow suit.
- Some realising the struggle of keeping up with its maintenance.
- Hidden costs and processes emerge which either slows organisations, or splits it up into silos.
- Many forgetting or violating the four **M**s of big data: **M**ake **M**e **M**ore **M**oney.

3 Key Big Data Trends You Should Know

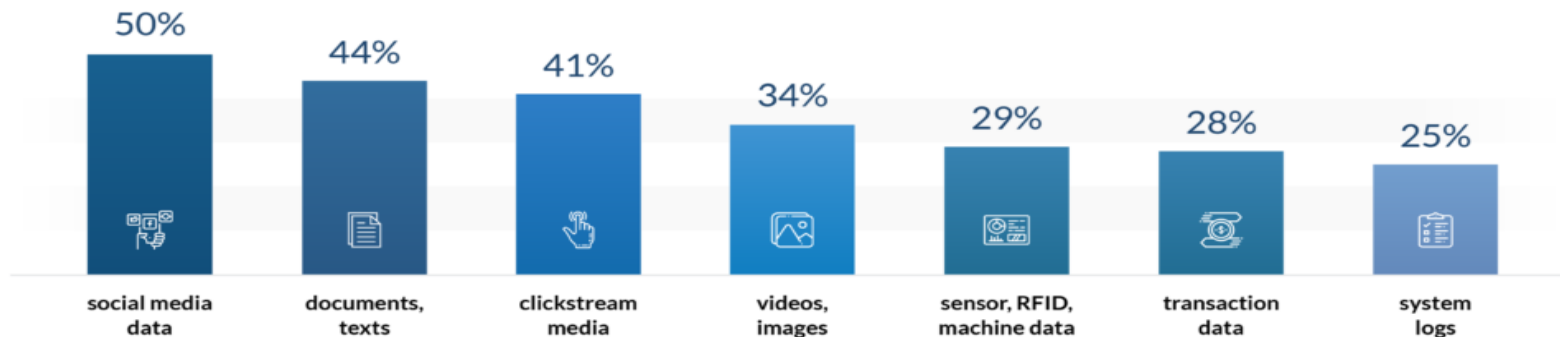
1 Top benefits that drive the use of data analytics

Source: Chicago Analytics Group 2019



2 Areas where companies plan to increase their big data analysis investment

Source: BI Survey 2019



3 How do companies plan to monetize their data?

Source: BARC survey, April 2019



Useful Books and Resources

- Real-time big data processing for anomaly detection: A Survey
- <https://www.sciencedirect.com/science/article/abs/pii/S0268401218301658>
- <https://www.researchgate.net/publication/327538806> Real-time big data processing for anomaly detection A Survey
- Deep learning and big data technologies for IoT security
- <https://www.sciencedirect.com/science/article/abs/pii/S0140366419315361>
- <https://www.researchgate.net/publication/338523843> Deep learning and big data technologies for IoT security

Useful Books and Resources

- Big Data For Dummies by [Judith Hurwitz](#), [Alan Nugent](#), [Fern Halper](#), and [Marcia Kaufman](#).
- Hadoop, The Definitive Guide [[pdf](#)]
- <http://bigdata.andreamostosi.name/>
- Big Data Fundamentals Concepts, Drivers & Techniques.
- <http://www.sciencedirect.com/science/article/pii/S0306437914001288>

Quick Review Question

- What exactly is Big Data?
- What are the biggest challenges of big data?
Creating / collecting the right data? Identifying / blending multiple external data sources?
- How big data analysis helps businesses increase their revenue?