Big Data Analytics

Dr. Sonali Agarwal
Associate Professor, Department of IT
Indian Institute of Information Technology Allahabad, India

Big Data Trends for 2024

01
Data
Gravity
VOLUME

02
Democratizing
Real-Time Data
VELOCITY

03
Augmented
Data Management
VARIETY

04
Identity
Authenticity
VERACITY

05
Data
Monetization
VALUE

06
Adaptive
Data Governance
VIRTUE

07
AI-Driven Storytelling &
Augmented Analytics
VISUALIZATION

08
Data
Marketplace
VIRALITY

09
DevOps –
DataOps – XOps
VISCOSITY

Trend 1: Big Data Gravity

Volume

Data attracts more data and an ecosystem of applications and services

SharePoint, OneDrive, Google Drive, and Dropbox offer APIs and integration opportunities for developers to enhance their products.

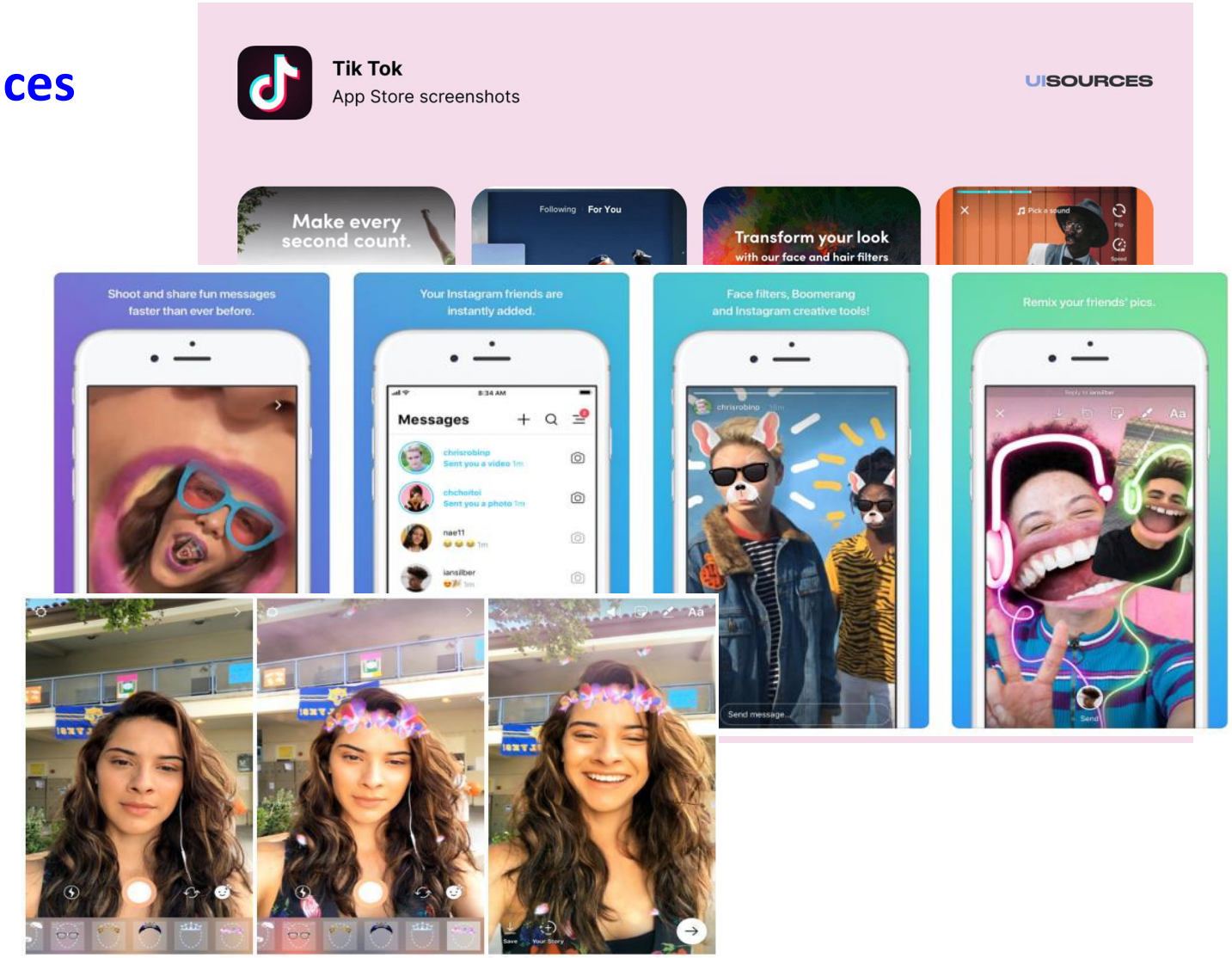
Social media platforms thought about this early by allowing for an ecosystem of filters, apps, games, and effects that engage their users with little to no additional effort from internal resources.

SharePoint

Google Drive

Dropbox

OneDrive



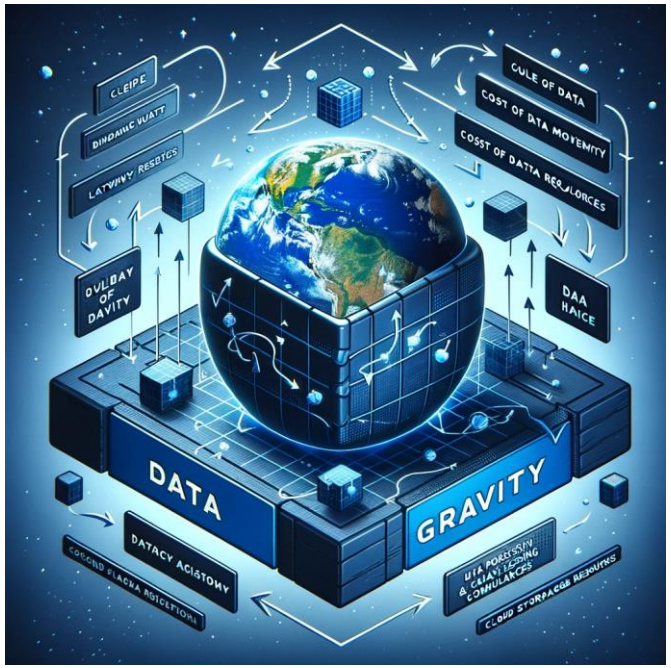
Trend 1: Big Data Gravity

Volume

Demand for Storage and Bandwidth CONTINUES to GROW

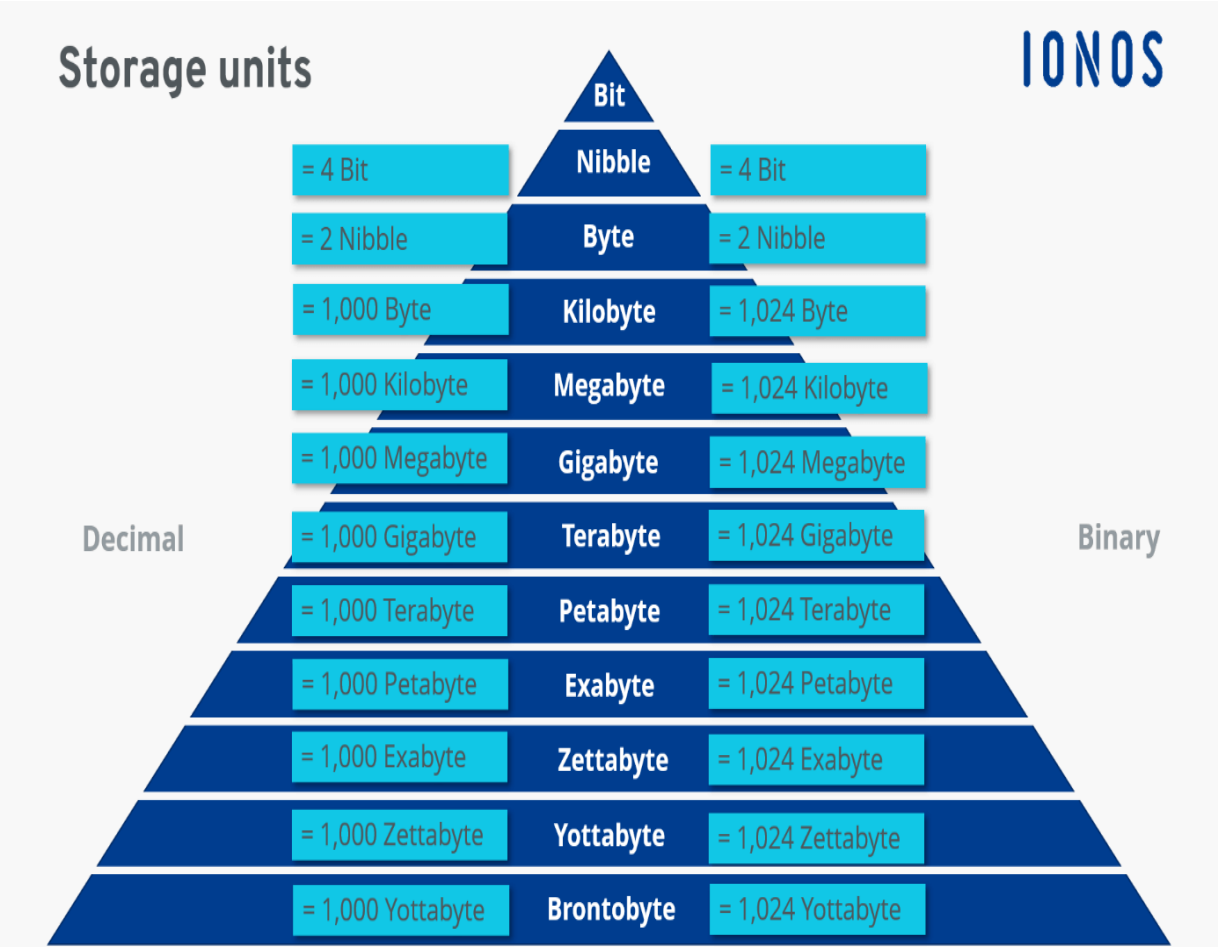
- Data Volume as a conceptual Indicator
  - Bandwidth and Latency Considerations
  - Costs Associated with Data Movement
  - Data Processing and Analytics Concentration
  - Expansion of Cloud and Computing Resources

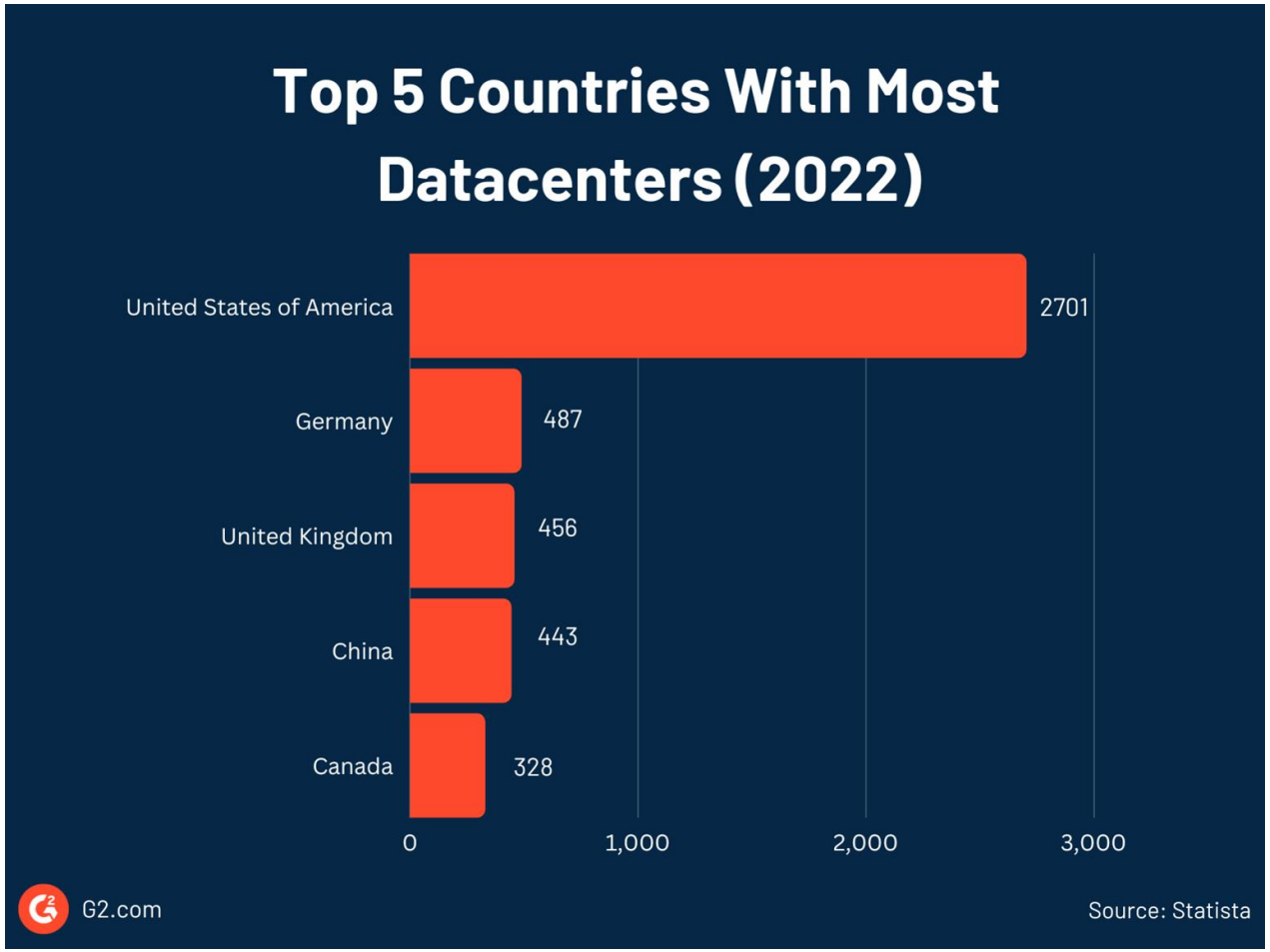
What worked for terabytes is ineffective for petabytes



- The largest data center in the world is a citadel in Reno, Nevada, that stretches over 7.2 million square feet! ***Source: Cloudwards, 2022***
- IoT devices will generate 79.4 zettabytes of data by 2025. ***Source: IDC, 2019***
- There were about 97 zettabytes of data generated worldwide in 2022. *Source:* ***"Volume of Data," Statista, 2022***

Trend 1: Data Gravity





Demo: Microsoft Data Center under SEA

Big Data Trends for 2024

01
Data Gravity
VOLUME

03
Augmented Data Management
VARIETY

05
Data Monetization
VALUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

09
DevOps – DataOps – XOps
VISCOSITY

VELOCITY
Democratizing Real-Time Data
02

VERACITY
Identity Authenticity
04

VIRTUE
Adaptive Data Governance
06

VIRALITY
Data Marketplace
08

Trend 2: Democratizing Real-Time Data

Trend 2: Democratizing Real-Time Data

Velocity

Real-time ANALYTICS presents an important differentiator

- The velocity element of data can be assessed from two standpoints: the speed at which data is being generated and how fast the organization needs to respond to the incoming information through capture, analysis, and use.
- Traditionally data was processed in a batch format **(all at once or in incremental nightly data loads).** There is a growing demand to process data continuously using streaming data-processing techniques.
- Emerging technologies and capabilities: **Edge Computing**

Trend 2: Democratizing Real-Time Data

Velocity

Make data accessible to everyone in real time

- 90% of an organization's data is replicated or redundant.
- Build API and web services that allow for live access to data.
- Most social media platforms, like Twitter and Facebook, have APIs that offer access to incredible amounts of data and insights.

- Data democratization means data is widely accessible to all stakeholders without bottlenecks or barriers. Success in data democratization comes with ubiquitous real-time analytics.

6G will push Wi-Fi connectivity to 1 terabyte per second!
This is expected to become commercially available by 2030.

Nearly 70% of all new vehicles globally will be connected to the internet by 2026.

Source: "Connected light-duty vehicles," Statista, 2022

Big Data Trends for 2024

01
Data Gravity
VOLUME

02
Democratizing Real-Time Data
VELOCITY

03
Augmented Data Management
VARIETY

04
Identity Authenticity
VERACITY

05
Data Monetization
VALUE

06
Adaptive Data Governance
VIRTUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

08
Data Marketplace
VIRALITY

09
DevOps – DataOps – XOps
VISCOSITY

Trend 3: Augmented Data Management

Variety

Need to manage unstructured data

- The variety of data types is increasingly diverse. Structured data often comes from relational databases, while unstructured data comes from several sources such as photos, video, text documents, cell phones, etc. The variety of data is where technology can drive business value. However, unstructured data also poses a risk, especially for external data.

The number of IoT devices could rise to 30.9 billion by 2025.

Source: "IoT and Non-IoT Connections Worldwide," Statista, 2022

The global edge computing market is expected to reach $250.6 billion by 2024.

Source: "Edge Computing," Statista, 2022

Genomics research is expected to generate between 2 and 40 exabytes of data within the next decade.

Source: NIH, 2022

Big Data Trends for 2024

01
Data Gravity
VOLUME

02
Democratizing Real-Time Data
VELOCITY

03
Augmented Data Management
VARIETY

04
Identity Authenticity
VERACITY

05
Data Monetization
VALUE

06
Adaptive Data Governance
VIRTUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

08
Data Marketplace
VIRALITY

09
DevOps – DataOps – XOps
VISCOSITY

Trend 4: Identity Authenticity

Varacity

Veracity of data is a true test of your Data Capabilities

Data veracity is defined as the accuracy or truthfulness of a data set.

More and more data is created in semi-structured and unstructured formats and originates from largely uncontrolled sources (e.g. social media platforms, external sources).

Data quality affects overall labor productivity by as much as 20%, and 30% of operating expenses are due to insufficient data.

Source: Pragmatic Works, 2017

Bad data costs up to 15% to 25% of revenue.

Source: MIT Sloan Management Review, 2017

Big Data Trends for 2024

01
Data Gravity
VOLUME

03
Augmented Data Management
VARIETY

05
Data Monetization
VALUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

09
DevOps – DataOps – XOps
VISCOSITY

VELOCITY
Democratizing Real-Time Data
02

VERACITY
Identity Authenticity
04

VIRTUE
Adaptive Data Governance
06

VIRALITY
Data Marketplace
08

Trend 5: Data Monetization

Value

- Data monetization is the transformation of data into financial value.
- However, this does not imply selling data alone.
- Monetary value is produced by using data to improve and upgrade existing and new products and services.
- Data monetization demands an organization-wide strategy for value development.

Netflix uses big data to save $1 billion per year on customer retention.

Source: Logidots, 2021

Big Data Trends for 2024

01
Data Gravity
VOLUME

02
Democratizing Real-Time Data
VELOCITY

03
Augmented Data Management
VARIETY

04
Identity Authenticity
VERACITY

05
Data Monetization
VALUE

06
Adaptive Data Governance
VIRTUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

08
Data Marketplace
VIRALITY

09
DevOps – DataOps – XOps
VISCOSITY

Trend 6: Adaptive Data Governance

Virtue

- **Five CORE VIRTUES: Resilience, Humility, Grit, Liberal Education, Empathy (FORBES, 2020)**

- Resilience
  - Adapt to situations and recover quickly.
  - Engineer conditions that explore the full solution space and feasible scenarios available.
  - Account for the effects of local constraints and overcome such conditions to mitigate premature stopping.

Trend 6: Adaptive Data Governance

Virtue

- **Five CORE VIRTUES: Resilience, Humility, Grit, Liberal Education, Empathy (FORBES, 2020)**

- Humility
    - Take responsibility for results.
    - Continually learn and adapt with reinforcement learning.
    - Recognize and engineer outcomes that appreciate how little can be known or controlled.

Trend 6: Adaptive Data Governance

Virtue

- **Five CORE VIRTUES: Resilience, Humility, Grit, Liberal Education, Empathy (FORBES, 2020)**

Grit

- Be obsessed with being productive and getting stuff done in new and innovative ways.
- Provide auditable and interpretable results.

Trend 6: Adaptive Data Governance

Virtue

- **Five CORE VIRTUES: Resilience, Humility, Grit, Liberal Education, Empathy (FORBES, 2020)**

Liberal Education

- Welcome and work with complexity, diversity and change.
- Review opportunities or business problems critically, analyzing data fully and considering feasible methods to formulate solutions.
- Communicate clearly, cleanly and reasonably with documentation that develops healthy and accountable solutions for society.

Trend 6: Adaptive Data Governance

Virtue

- **Five CORE VIRTUES: Resilience, Humility, Grit, Liberal Education, Empathy (FORBES, 2020)**

- Empathy
  - Recognize and account for social impact and the feelings of others.
  - Develop objective functions or constraints based on understanding and compassion.
  - Identify interdependence and direct connections with a higher purpose or consciousness.

Big Data Trends for 2024

01
Data Gravity
VOLUME

02
Democratizing Real-Time Data
VELOCITY

03
Augmented Data Management
VARIETY

04
Identity Authenticity
VERACITY

05
Data Monetization
VALUE

06
Adaptive Data Governance
VIRTUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

08
Data Marketplace
VIRALITY

09
DevOps – DataOps – XOps
VISCOSITY

Trend 6: AI-Driven Storytelling & Augmented Analytics

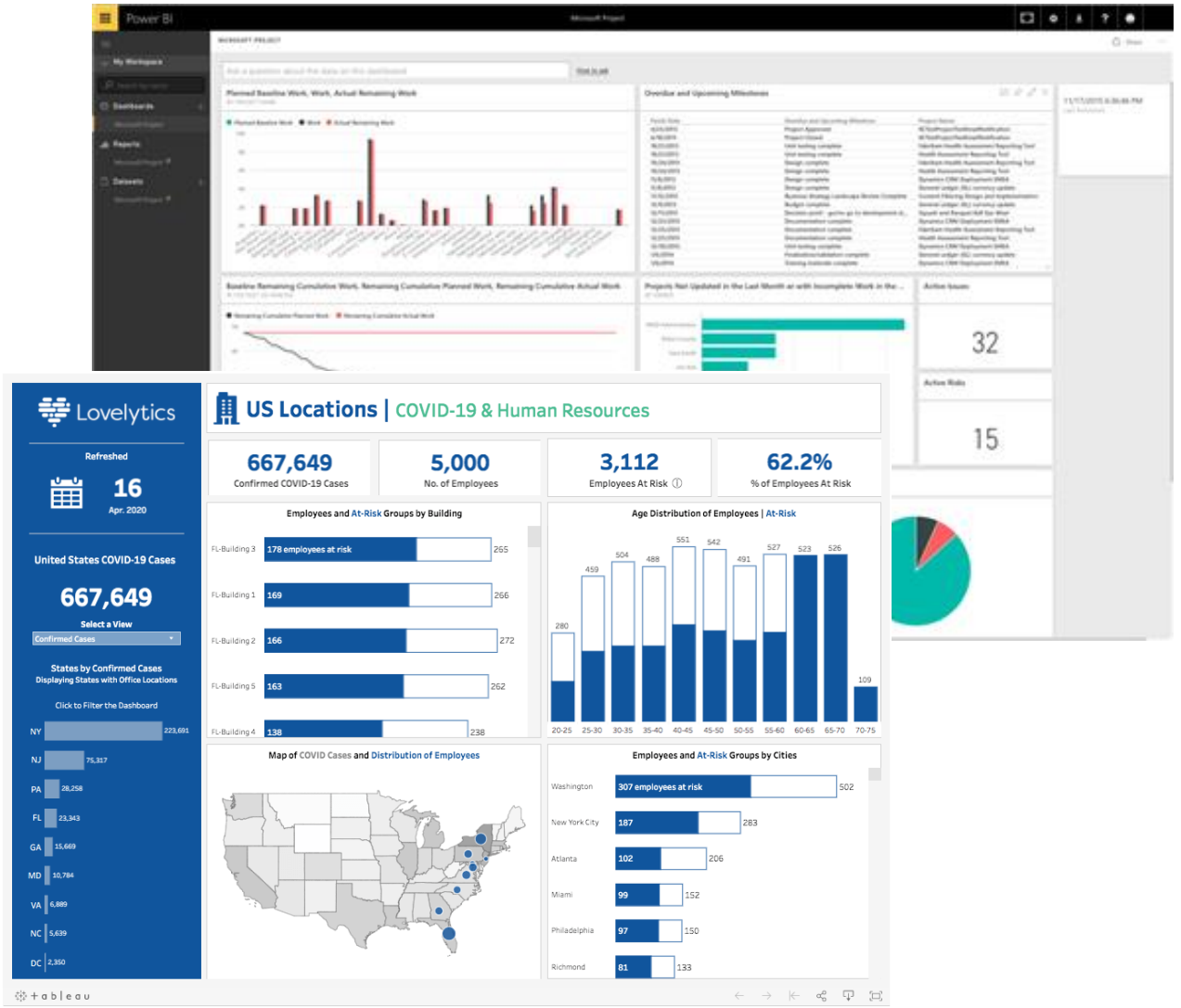
Visualization

Use AI to enhance data storytelling

- Tableau, Power BI, and many other applications already use AI-driven analytics.
- Power BI and SharePoint can use AI to generate visuals for any SharePoint list in a matter of seconds.

Emerging technologies and capabilities:

AI-Powered Data Catalog and Metadata Management,
Automated Data Policy Enforcement



Big Data Trends for 2024

01
Data Gravity
VOLUME

03
Augmented Data Management
VARIETY

05
Data Monetization
VALUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

09
DevOps – DataOps – XOps
VISCOSITY

VELOCITY
Democratizing Real-Time Data
02

VERACITY
Identity Authenticity
04

VIRTUE
Adaptive Data Governance
06

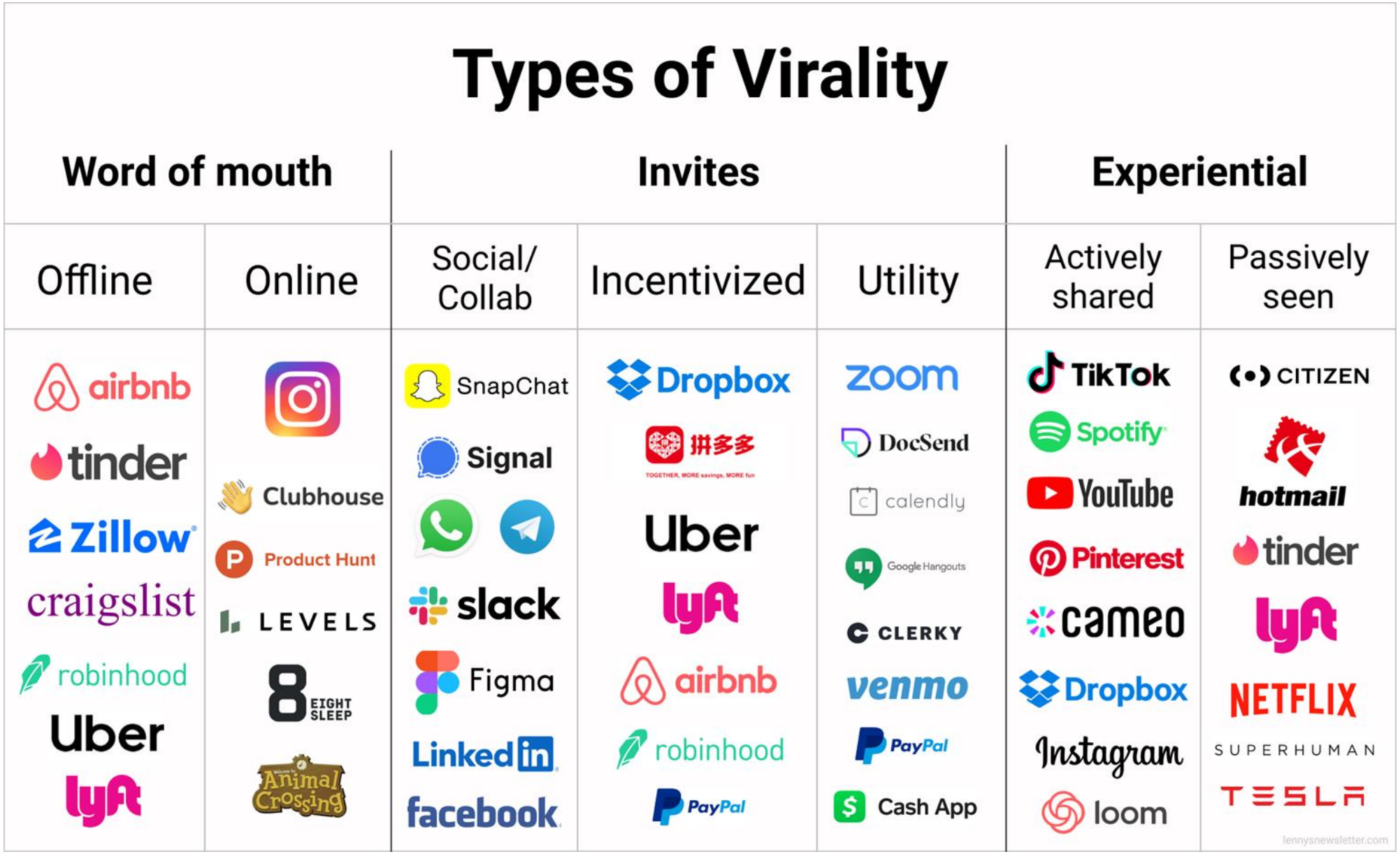
VIRALITY
Data Marketplace
08

Trend 8: Data Marketplace

Make data easily accessible

- Making data accessible to a broader audience is the key to successful virality.
- Data marketplaces provide a location for you to make your data public.
- Why do this? Contributing to public data marketplaces builds credibility, just like contributing to public GitHub projects.
- Big players like Microsoft, Amazon, and Snowflake already do this!
- Snowflake introduced zero-copy cloning, which allows users to interact with source data without compromising the integrity of the original source.
- Emerging technologies and capabilities: AI-Powered Data Catalog and Metadata Management, Automated Data Policy Enforcement

Big Data Trends for 2024



Data Trends for 2024

01
Data Gravity
VOLUME

02
Democratizing Real-Time Data
VELOCITY

03
Augmented Data Management
VARIETY

04
Identity Authenticity
VERACITY

05
Data Monetization
VALUE

06
Adaptive Data Governance
VIRTUE

07
AI-Driven Storytelling & Augmented Analytics
VISUALIZATION

08
Data Marketplace
VIRALITY

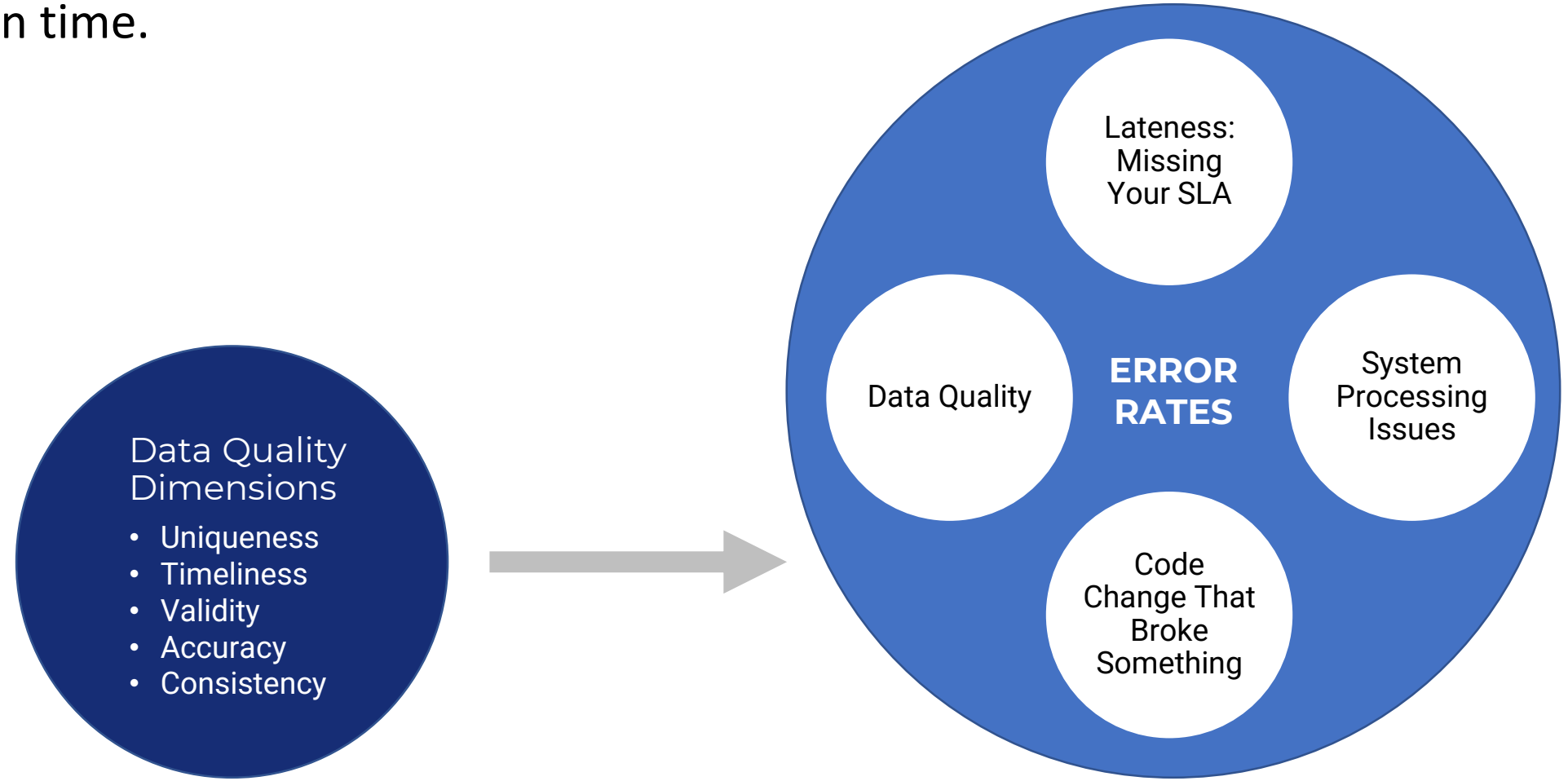
09
DevOps – DataOps – XOps
VISCOSITY

DevOps – DataOps – XOps

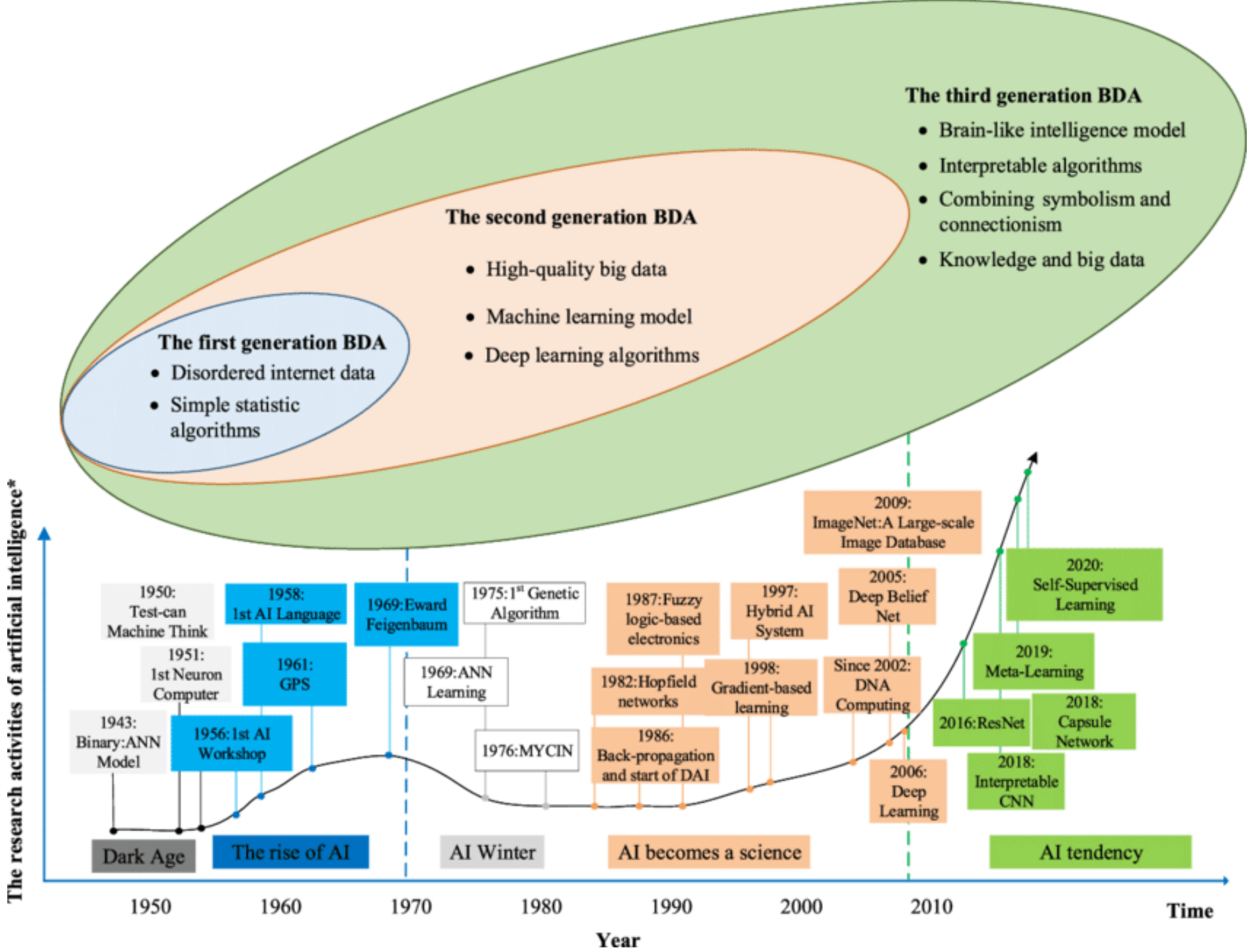
- Compared to water, a fluid with a high viscosity flows more slowly, like honey.
- Data viscosity measures the resistance to flow in a volume of data.
- The data resistance may come from other Vs (variety, velocity, etc.).
- The merger of development (Dev) and IT Operations (Ops) started in software development with the concept of DevOps. Since then, new Ops terms have formed rapidly (AIOps, MLOps, ModelOps, PlatformOps, SalesOps, SecOps, etc.).

DataOps → Data Observability

- Data observability, a subcomponent of DataOps, is a set of technical practices, cultural norms, and architecture that enables low error rates.
- Data observability focuses on error rates instead of only measuring data quality at a single point in time.



Generations of BDA


The third generation BDA
Brain-like intelligence model
Interpretable algorithms
Combining symbolism and connectionism
Knowledge and big data
The second generation BDA
High-quality big data
Machine learning model
Deep learning algorithms
The first generation BDA
Disordered internet data
Simple statistic algorithms
The research activities of artificial intelligence*
1950: Test-can Machine Think
1951: 1st Neuron Computer
1943: Binary:ANN Model
1956:1st AI Workshop
1958: 1st AI Language
1961: GPS
1969:Eward Feigenbaum
1969:ANN Learning
1975:1st Genetic Algorithm
1976:MYCIN
1982:Hopfield networks
1986: Back-propagation and start of DAI
1987:Fuzzy logic-based electronics
1997: Hybrid AI System
1998: Gradient-based learning
2005: Deep Belief Net
2006: Deep Learning
2009: ImageNet:A Large-scale Image Database
Since 2002: DNA Computing
2016:ResNet
2018: Interpretable CNN
2018: Capsule Network
2019: Meta-Learning
2020: Self-Supervised Learning
Dark Age
The rise of AI
AI Winter
AI becomes a science
AI tendency
1950
1960
1970
1980
1990
2000
2010
Time
Year

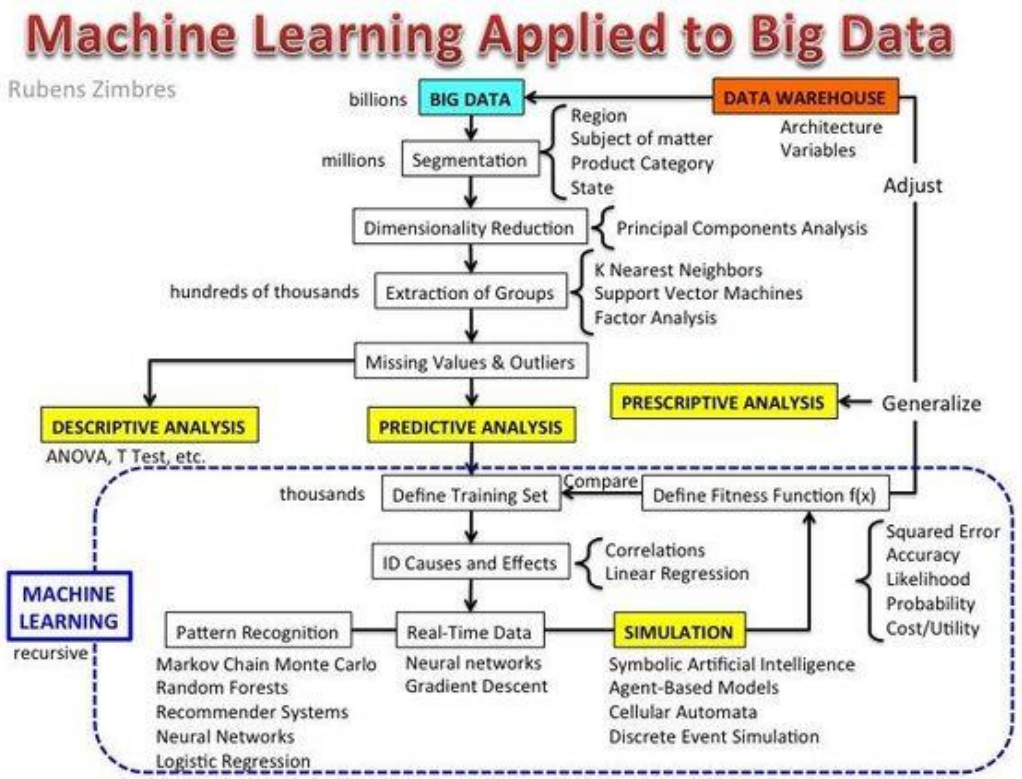
Why Big Data need different types of Machine Learning algorithms

Why Big Data need different types of Machine Learning algorithms

- Learning for different types of data
- Learning for high speed of streaming data
- Learning for uncertain and incomplete data
- Learning for data with low value density and meaning diversity

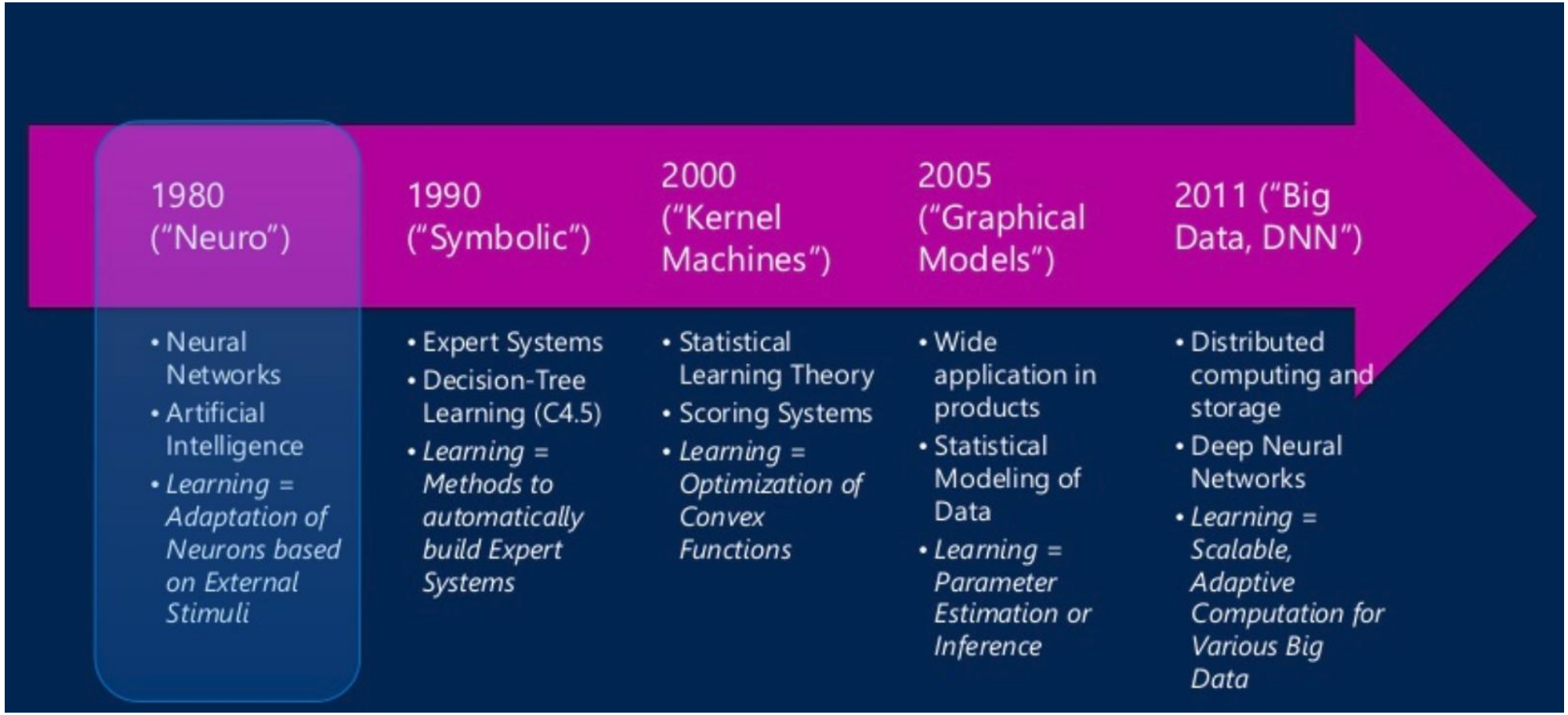
Machine Learning for Big Data

- Rubens Zimbres captures the synergistic interaction between machine learning and big data and is shown as a mindmap as follows:

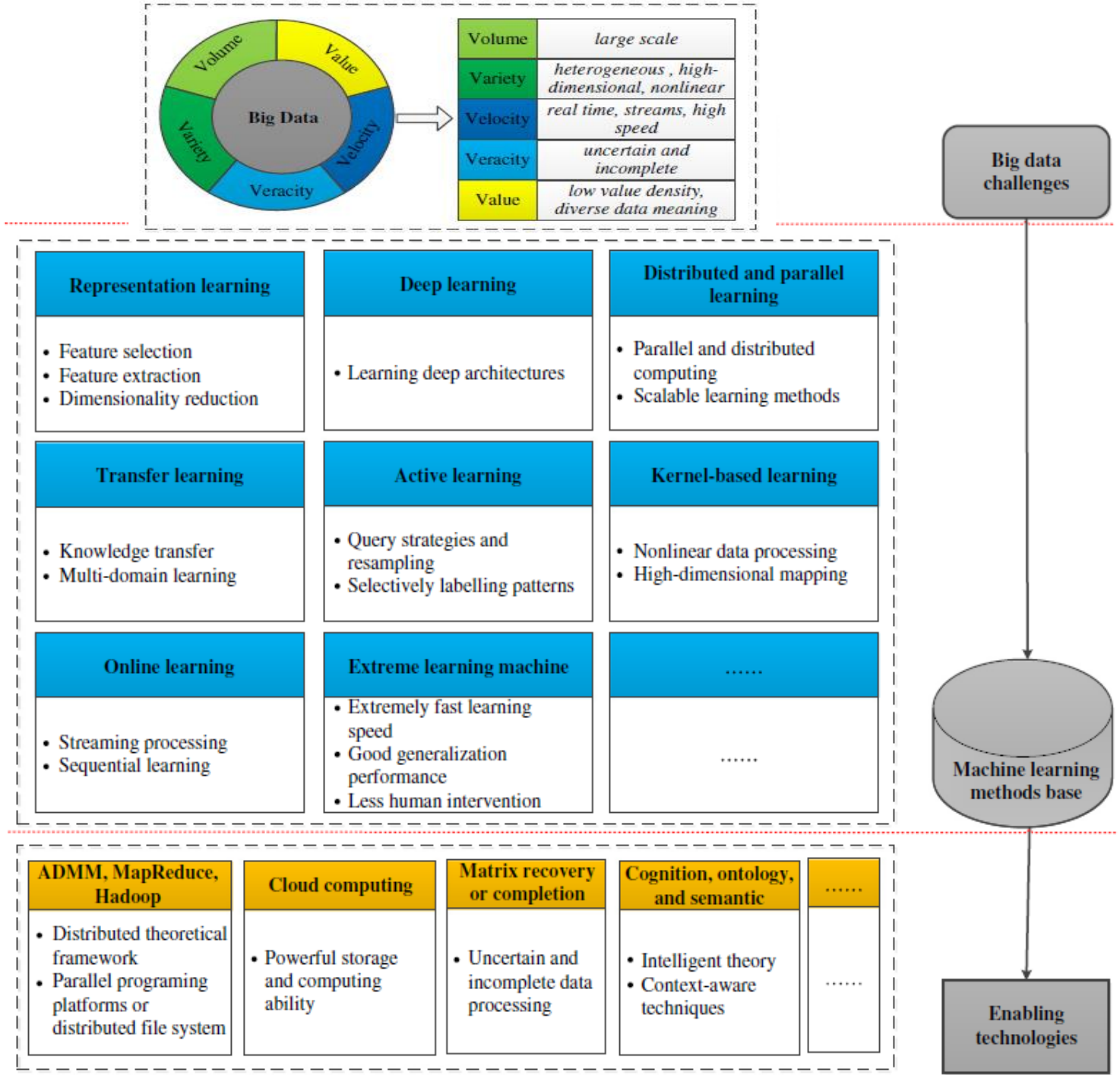


https://www.datasciencecentral.com/profiles/blogs/key-tools-of-big-data-for-transformation-review-case-study

Machine Learning Advancement and Big Data



Hierarchical framework of efficient machine learning for big data processing



Big Data Challenges and Machine Learning Advantages

Big Data Challenges: Volume

- According to a recent study by (MIT), 1 the average processing speed in humans is around 45 bits per second, with skilled individuals going up to 60 bits per second.
- That means that if a person were to spend 24 hours a day just reading (without ever stopping, not even to eat or drink or sleep), that person would process slightly above 5 MB of data.
- It would take such a person around 536 days to go through a 1TB data set – and at the end of the process, the person would very likely not remember the entire data set.

Machine Learning Advantages

- ML has no issue with volume; in fact, the more data you have, the higher-quality results you expect.
- In terms of capability to process volume, the machine is constrained only by the size of the machine (a modern laptop can likely read 1TB data in a few seconds)

Big Data Challenges and Machine Learning Advantages

Big Data Challenges: Velocity

- Data changes too quickly for a human to understand the consequences and react in a positive manner.
- For example, the New York Stock Exchange typically produces 1 TB of trade information during each trading session.
- In simplistic terms, 1 TB is equal to a couple million decent-sized books. Humans simply cannot absorb the data fast enough and, even more important, are unable to sift through the data to find thevaluable actionable insight

Machine Learning Advantages

- ML operates at machine speeds: a successfully trained model can react to input data at speeds impossible for a human.
- Recent advances in neural nets allow for immediate reaction to inputs “close to edge”– where the event happens.
- Reaction times of an efficient model can be measured in tens of milliseconds.

Big Data Challenges and Machine Learning Advantages

Big Data Challenges :Variety

- Most enterprises now have data from multiple sources (e.g., databases, real-time social media streams, IoT reporting) and in various formats (e.g., structured, unstructured).
- The average human intellect will almost certainly be unable to construct and hold in memory a data model the considers every single possible correlation and relationship among data from multiple sources and in multiple formats.
- The variety of data needs to be simplified for human consumption either via description, aggregation, or summarization, but simplification loses the potential of big data and makes finding hidden insights unlikely

Machine Learning Advantages

- Variety is not a problem for machines, as extremely complex models can be built and held in machine memory.
- ML operates at machine scale with modern machines capable of holding and making use of terabytes of information in milliseconds

Big Data Challenges and Machine Learning Advantages

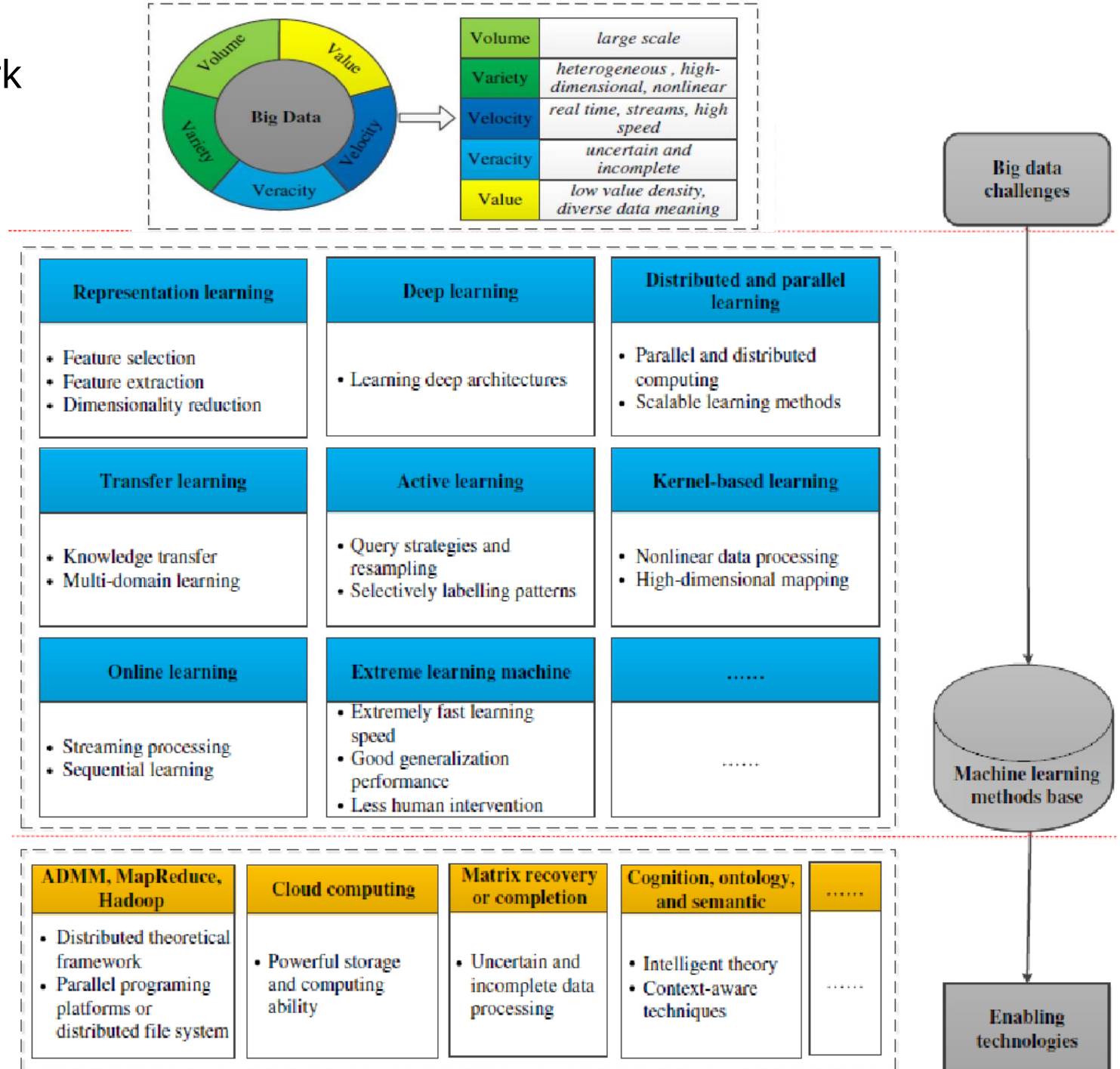
Big Data Challenges :Veracity

- We cannot be confident that the data we have is clean, usable, and of high quality – and even worse, humans are prone to errors of judgement such as confirmation bias and will tend to attribute more worth to data that confirms biases – even if it means losing out financially

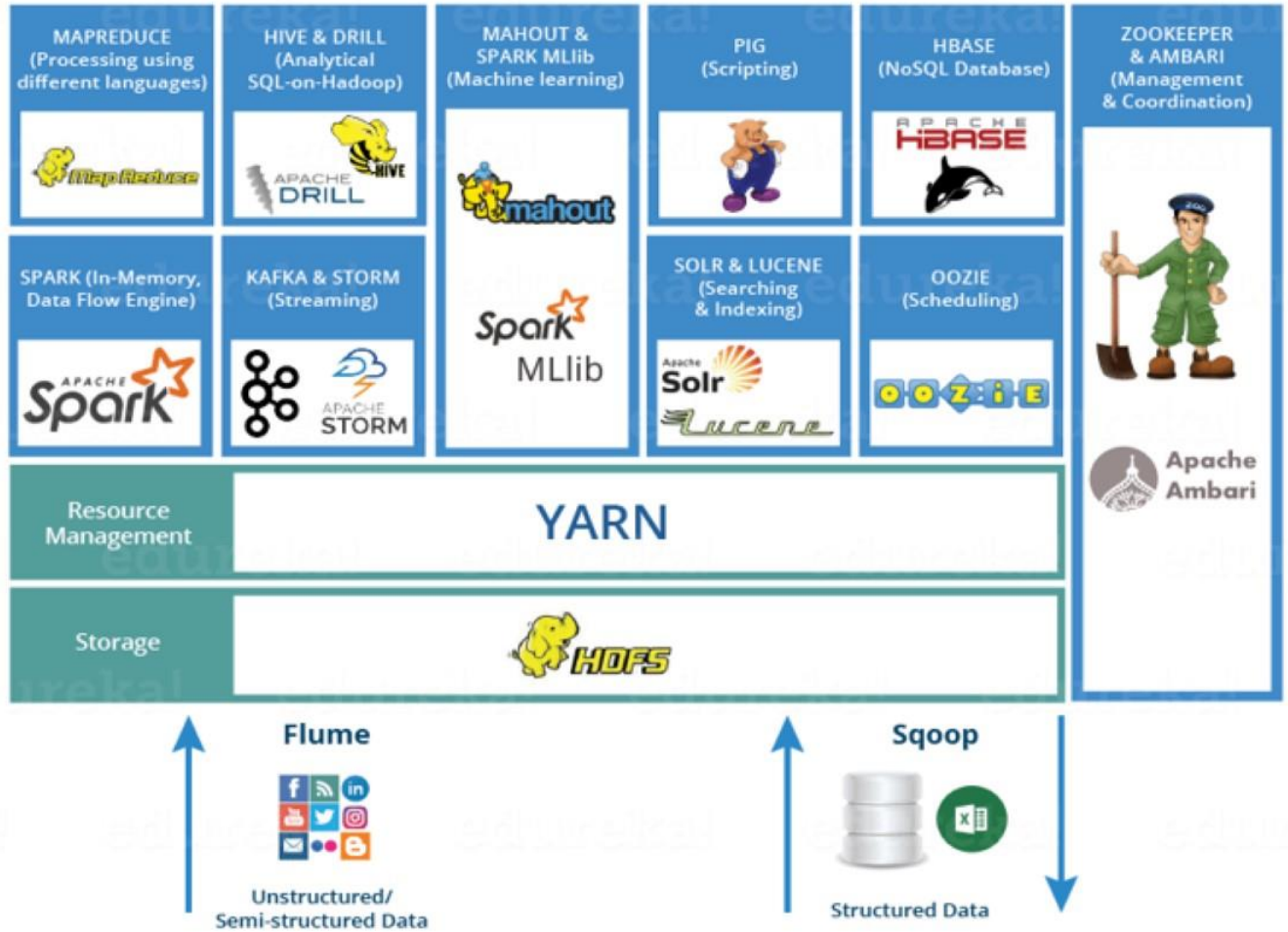
Machine Learning Advantages

- ML has no confirmation (or other) bias when analysing data.
- Proper algorithms and models will also trend to recognize “bad” data.
- Significant advances in automated algorithms exist that allow the machine to effectively ignore noise and outliers.

Hierarchical framework of efficient machine learning for big data processing



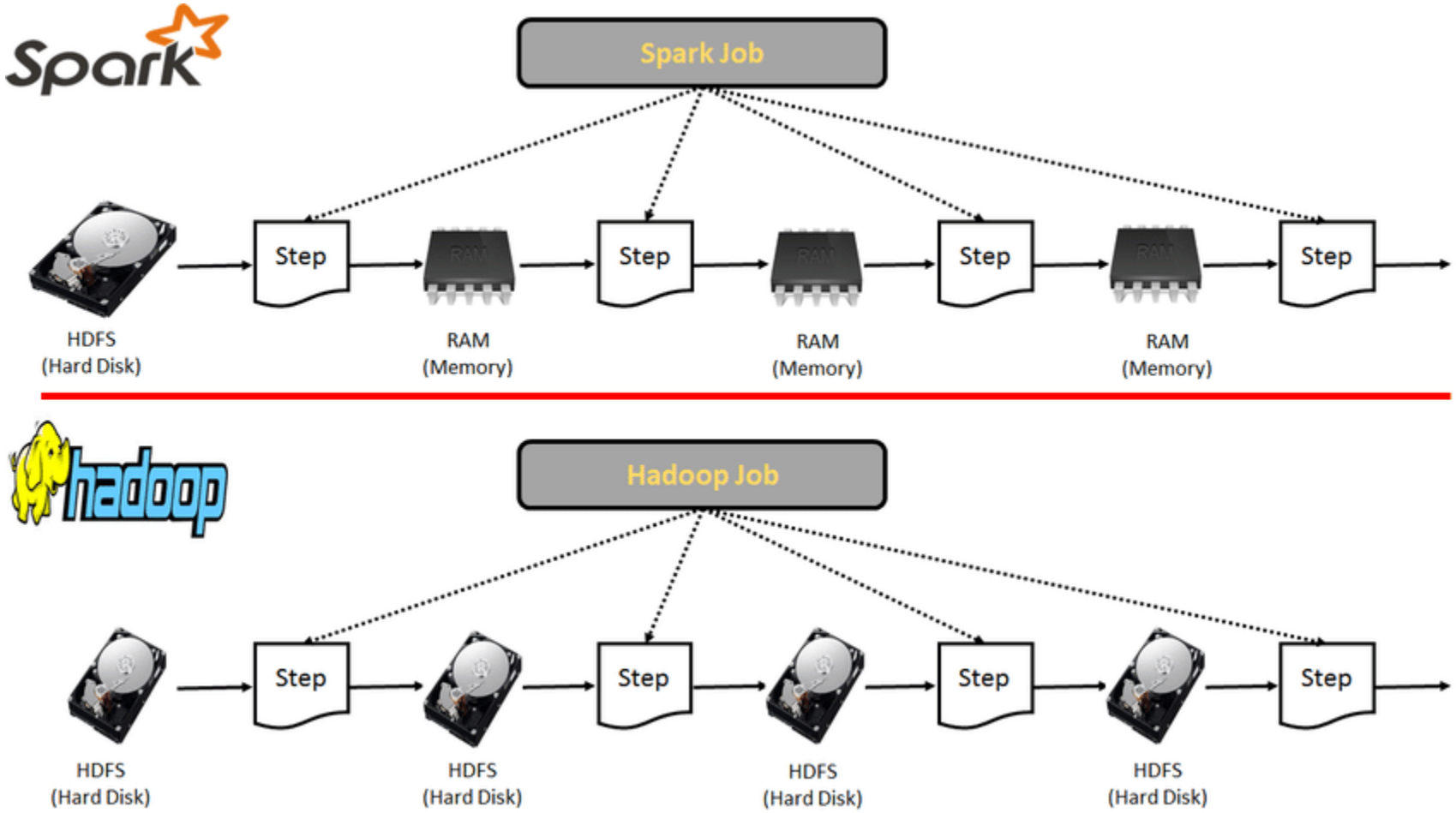
Hadoop Ecosystem



HDFS -> Hadoop Distributed File System
YARN -> Yet Another Resource Negotiator
MapReduce -> Data processing using programming
Spark -> In-memory Data Processing
PIG, HIVE-> Data Processing Services using Query (SQL-like)
HBase -> NoSQL Database

Mahout, Spark MLlib -> Machine Learning
Apache Drill -> SQL on Hadoop
Zookeeper -> Managing Cluster
Oozie -> Job Scheduling
Flume, Sqoop -> Data Ingesting Services
Solr & Lucene -> Searching & Indexing
Ambari -> Provision, Monitor and Maintain cluster

In Memory Versus Disk based Computation



Big Data Formats

Consideration 1: ROW VS. COLUMN

- It is the most important consideration.
- A row based format or a column-based format.
- At the highest level, column-based storage is most useful when performing analytics queries that require only a subset of columns examined over very large data sets.
- If your queries require access to all or most of the columns of each row of data, row-based storage will be better suited to your needs.

Big Data Formats

- To help illustrate the differences between row and column-based data, consider this table of basic transaction data.
- For each transaction, we have the customer name, the product ID, sale amount, and the date.

EXAMPLE: SAMPLE TRANSACTION DATA

Customer Name	Product ID	Sale Amount	Transaction Date
Emma	Prod 1	100.00	2018-04-02
Liam	Prod 2	79.99	2018-04-02
Noah	Prod 3	19.99	2018-04-01
Olivia	Prod 2	79.99	2018-04-03

Big Data Formats

- **Row-based storage** is the simplest form of data table and is used in many applications, from web log files to highly-structured database systems like MySql and Oracle.

EXAMPLE: SAMPLE TRANSACTION DATA

Customer Name	Product ID	Sale Amount	Transaction Date
Emma	Prod 1	100.00	2018-04-02
Liam	Prod 2	79.99	2018-04-02
Noah	Prod 3	19.99	2018-04-01
Olivia	Prod 2	79.99	2018-04-03

```
Emma,Prod1,100.00,2018-04-02;Liam,Prod2,79.99,2018-
04-02;Noah,Prod3,19.99,2018-04-01;Oliv-
ia,Prod2,79.99,2018-04-03
```

Big Data Formats

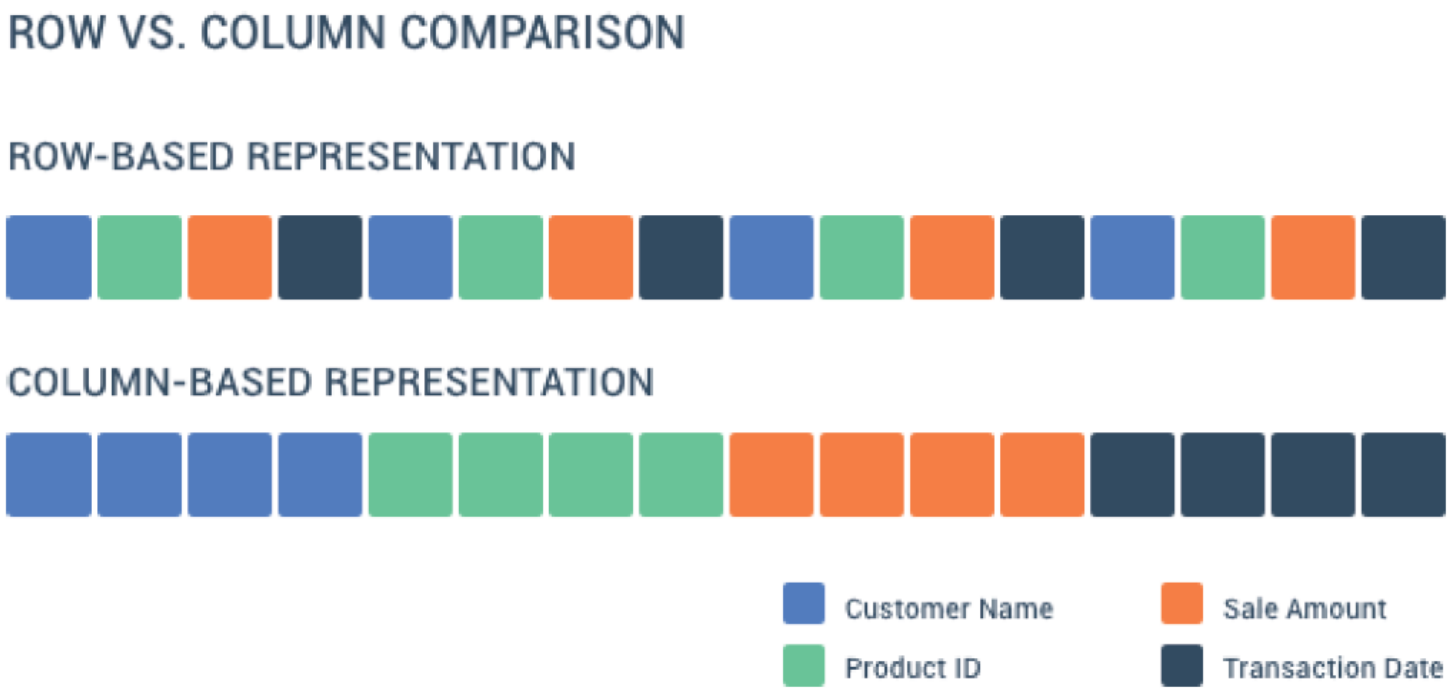
- In columnar formats, data is stored sequentially by column, from top to bottom—not by row, left to right. Having data grouped by column makes it more efficient to easily focus computation on specific columns of data. Column-based storage is also ideal for sparse data sets where you may have empty values.

EXAMPLE: SAMPLE TRANSACTION DATA

Customer Name	Product ID	Sale Amount	Transaction Date
Emma	Prod 1	100.00	2018-04-02
Liam	Prod 2	79.99	2018-04-02
Noah	Prod 3	19.99	2018-04-01
Olivia	Prod 2	79.99	2018-04-03

```
Emma,Liam,Noah,Olivia;Prod1,Prod2,Prod3;Pr
od2;100.00,79.99,19.99,79.99;2018-04-02,2018-04-02,
2018-04-01, 2018-04-03
```

Big Data Formats



To analyze “Sale Amount” (orange) and “Transaction Date” (navy blue) Column Based format is better than Row Based format

Big Data Formats

Consideration 2: **SCHEMA EVOLUTION**

- "Schema" in a database context, is its organization—the tables, columns, views, primary keys, relationships, etc.
- When evaluating schema, there are a few key questions to ask of any data format:
  - How easy is it to update a schema (such as adding a field, removing or renaming a field)?
  - How will different versions of the schema "talk" to each other?
  - Is it human-readable? Does it need to be?
  - How fast can the schema be processed?
  - How does it impact the size of data?

Big Data Formats

Consideration 3: **SPLITABILITY**

- Processing such datasets efficiently usually requires breaking the job up into parts that can be farmed out to separate processors.
- In fact, large-scale parallelization of processing is key to performance.
- Your choice of file format can critically affect the ease with which this parallelization can be implemented.
- For example, if each file in your dataset contains one massive XML structure or JSON record, the files will not be “splittable”, i.e. decomposable into smaller records that can be handled independently.

Big Data Formats

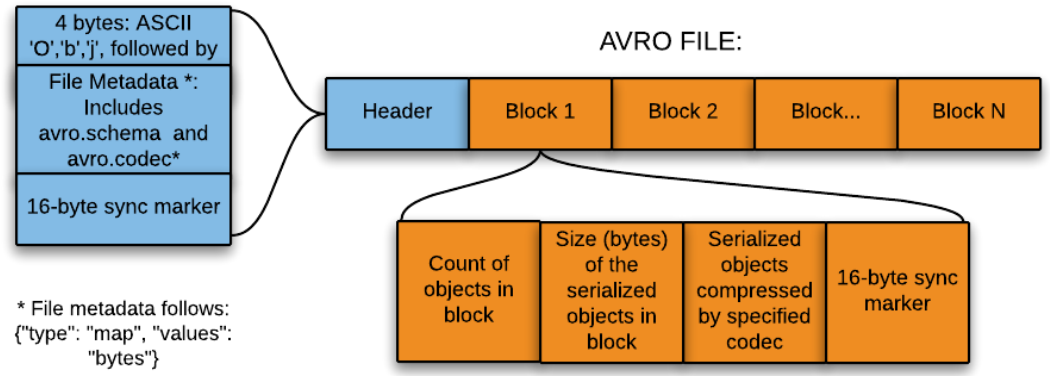
Consideration 4: **COMPRESSION**

- Data compression reduces the amount of information, the resources required to store and transmit data, typically saving time and money.
- Columnar data can achieve better compression rates than row-based data.
- Storing values by column, with the same type next to each other, allows you to do more efficient compression on them than if you're storing rows of data.
- For example, storing all dates together in memory allows for more efficient compression than storing data of various types next to each other—such as string, number, date, string, date.

Big Data Formats

APACHE AVRO: **A ROW BASED FORMAT**

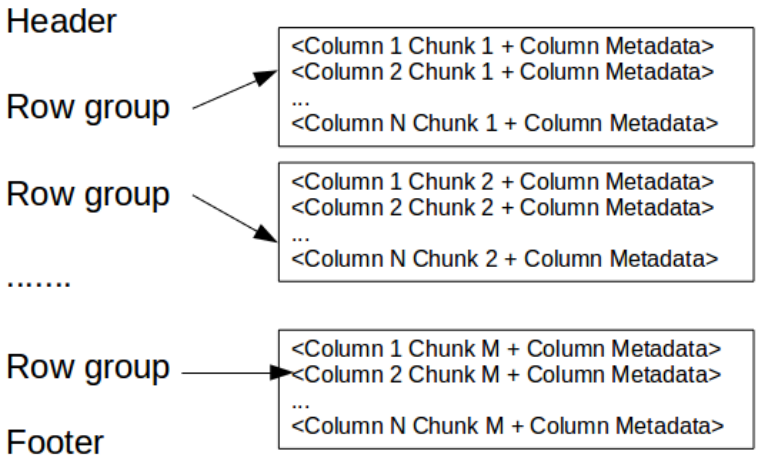
- Released by the Hadoop working group in 2009, is highly splittable.
- Schema travels with data.
- The data definition is stored in JSON format while the data is stored in binary format, minimizing file size and maximizing efficiency.
- Easy to change.



Big Data Formats

APACHE PARQUET: **A COLUMN BASED FORMAT**

- Launched in 2013, Parquet was developed by Cloudera and Twitter to serve as an optimized columnar data store on Hadoop.
- Because data is stored by columns, it can be highly compressed and splittable (for the reasons noted above).
- The column metadata for a Parquet file is stored at the end of the file, which allows for fast, one-pass writing. Metadata can include information such as, data types, compression/encoding scheme used (if any), statistics, element names, and more.



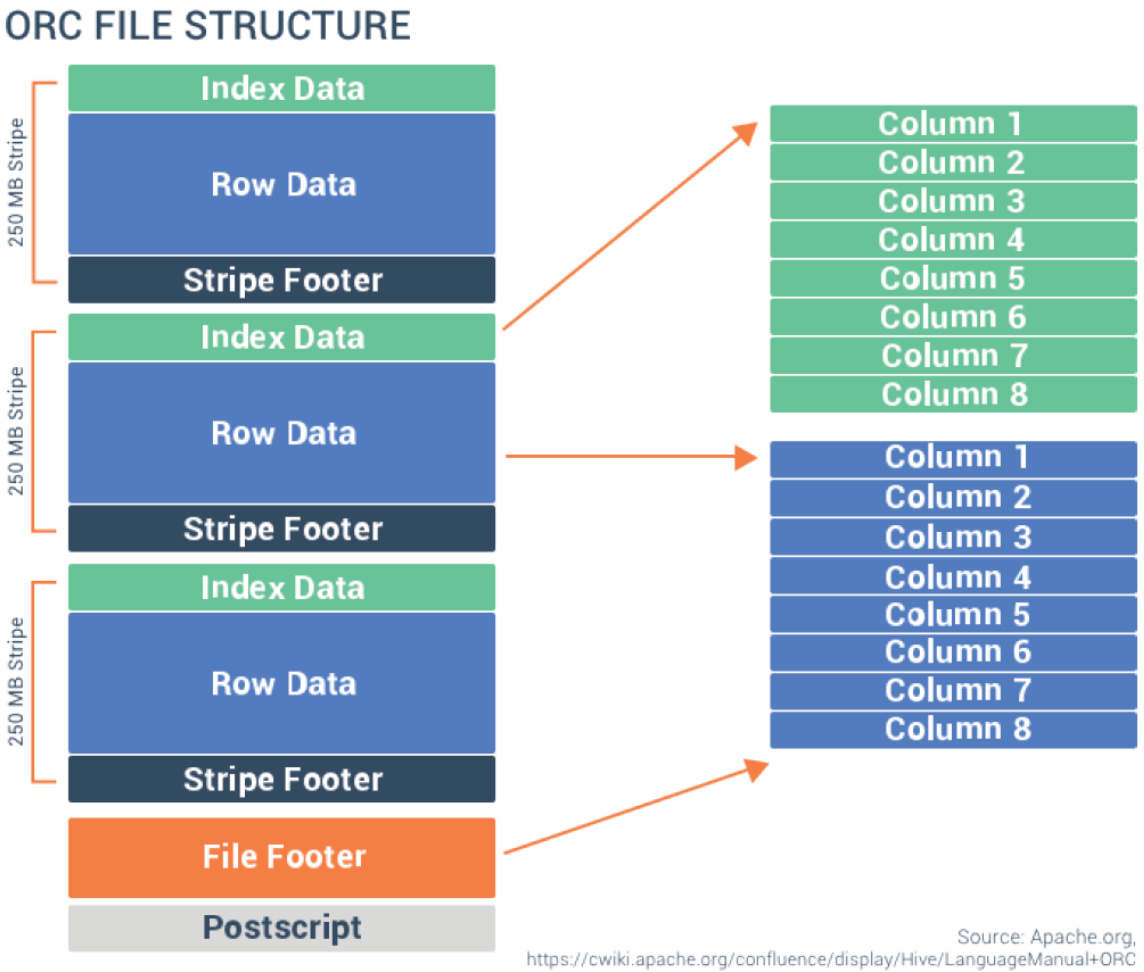
Big Data Formats

APACHE ORC: **A ROW-COLUMNAR BASED FORMAT**

- Optimized Row Columnar (ORC) format was first developed at Hortonworks to optimize storage and performance in Hive, a data warehouse for summarization, query and analysis that lives on top of Hadoop.
- Hive is designed for queries and analysis, and uses the query language HiveQL (similar to SQL). ORC files are designed for high performance when Hive is reading, writing, and processing data.
- ORC stores row data in columnar format. This row-columnar format is highly efficient for compression and storage.
- It allows for parallel processing across a cluster, and the columnar format allows for skipping of unneeded columns for faster processing and decompression.

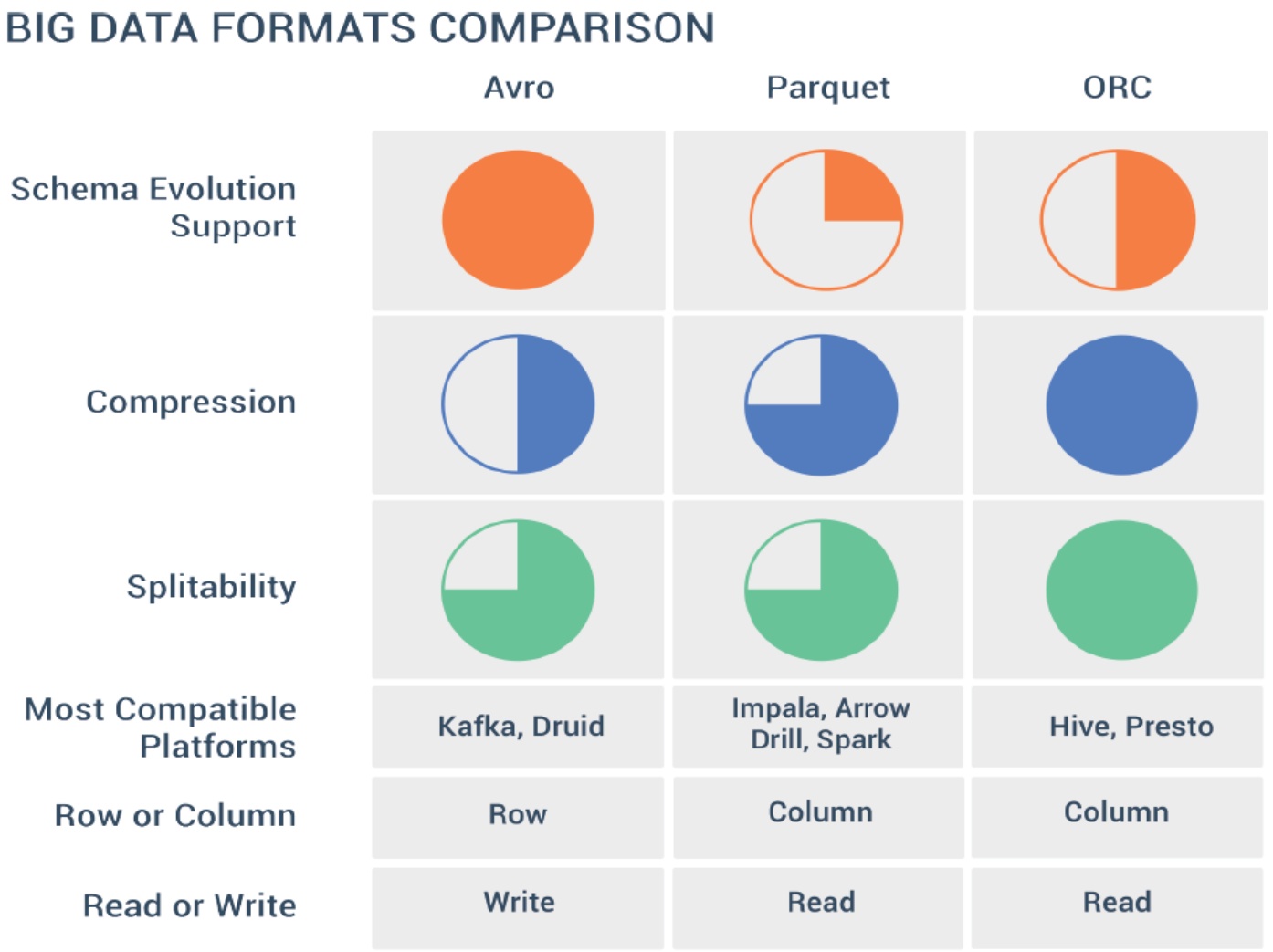
Big Data Formats

- APACHE ORC: **A ROW-COLUMNAR BASED FORMAT**



Source: Apache.org, https://cwiki.apache.org/confluence/display/Hive/LanguageManual+ORC

Big Data Format Comparison



References

- Official Websites of Apache Software

References

- **Advanced Analytics with Spark** by S. Ryza, U. Laserson, S. Owen and J. Wills
- **Apache Spark documentation**
  - http://spark.apache.org/
  - http://spark.apache.org/docs/latest/programming-guide.html
- **Pyspark**
  - http://spark.apache.org/docs/latest/api/python/pyspark.html
- **Resilient Distributed Dataset: A Fault-tolerant Abstraction for in-Memory Cluster Computing.** M. Zaharia et al.
  - https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf

Bibliography

Bean, Randy. “Why Becoming a Data-Driven Organization Is So Hard.” *Harvard Business Review*, 24 Feb. 2022. Accessed Oct. 2022.

Brown, Annie. “Utilizing AI And Big Data To Reduce Costs And Increase Profits In Departments Across An Organization.” *Forbes*, 13 April 2021. Accessed Oct. 2022.

Burciaga, Aaron. “Five Core Virtues For Data Science And Artificial Intelligence.” *Forbes*, 27 Feb. 2020. Accessed Aug. 2022.

Cadwalladr, Carole, and Emma Graham-Harrison. “Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach.”
The Guardian, 17 March 2018. Accessed Aug. 2022.

Carlier, Mathilde. “Connected light-duty vehicles as a share of total vehicles in 2023.” *Statista*, 31 Mar. 2021. Accessed Oct. 2022.

Carter, Rebekah. “The Ultimate List of Big Data Statistics for 2022.” Findstack, 22 May 2021. Accessed Oct. 2022.

- Castelvecchi, Davide. “Underdog technologies gain ground in quantum-computing race.” *Nature*, 6 Nov. 2023. Accessed Feb. 2023.

Clark-Jones, Anthony, et al. “Digital Identity:” *UBS*, 2016. Accessed Aug 2022.

“The Cost of Bad Data – Infographic.” *Pragmatic Works*, 25 May 2017. Accessed Oct. 2022.

Demchenko, Yuri, et al. “Data as Economic Goods: Definitions, Properties, Challenges, Enabling Technologies for Future Data Markets.“ *ITU Journal: ICT Discoveries*, Special Issue, no. 2, vol. 23, Nov. 2018. Accessed Aug 2022.

Feldman, Sarah. ”20 Years of Quantum Computing Growth.” *Statista*, 6 May 2019. Accessed Oct. 2022.

“Genomic Data Science.” *NIH, National Human Genome Research Institute*, 5 April 2022. Accessed Oct. 2022.

Bibliography

Hasbe, Sudhir, and Ryan Lippert. “The democratization of data and insights: making real-time analytics ubiquitous.” *Google Cloud*, 15 Jan. 2021.
Accessed Aug. 2022.

Helmenstine, Anne. “Viscosity Definition and Examples.” *Science Notes*, 3 Aug. 2021. Accessed Aug. 2022.

“How data storytelling and augmented analytics are shaping the future of BI together.” *Yellowfin*, 19 Aug. 2021. Accessed Aug. 2022.

“How Netflix Saves $1B Annually using AI?” *Logidots*, 24 Sept. 2021. Accessed Oct. 2022

Hui, Kenneth. “The AWS Love/Hate Relationship with Data Gravity.” *Cloud Architect Musings*, 30 Jan. 2017. Accessed Aug 2022.

ICD. “The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast.” *Business Wire*, 18 June 2019. Accessed Oct 2022.

Internet of Things (IoT) and non-IoT active device connections worldwide from 2010 to 2025” *Statista*, 27 Nov. 2022. Accessed Nov. 2022.

Koch, Gunter. “The critical role of data management for autonomous driving development.” *DXC Technology*, 2021. Accessed Aug. 2022.

Morris, John. “The Pull of Data Gravity.” *CIO*, 23 Feb. 2022. Accessed Aug. 2022.

Nield, David. “Google's Quantum Computer Is 100 Million Times Faster Than Your Laptop.” *ScienceAlert*, 9 Dec. 2015. Accessed Oct. 2022.

Redman, Thomas C. “Seizing Opportunity in Data Quality.” *MIT Sloan Management Review*, 27 Nov. 2017. Accessed Oct. 2022.

Segovia Domingo, Ana I., and Álvaro Martín Enríquez. “Digital Identity: the current state of affairs.” *BBVA Research*, 2018. Accessed Aug. 2022.

Bibliography

"State of IoT 2022: Number of connected IoT devices growing 18% to 14.4 billion globally." *IOT Analytics*, 18 May 2022. Accessed. 14 Nov. 2022.

Strod, Eran. "Data Observability and Monitoring with DataOps." *DataKitchen*, 10 May 2021. Accessed Aug. 2022.

Sujay Vailshery, Lionel. "Edge computing market value worldwide 2019-2025." *Statista*, 25 Feb. 2022. Accessed Oct 2022.

Sujay Vailshery, Lionel. "IoT and non-IoT connections worldwide 2010-2025." *Statista*, 6 Sept. 2022. Accessed Oct. 2022.

Sumina, Vladimir. "26 Cloud Computing Statistics, Facts & Trends for 2022." *Cloudwards*, 7 June 2022. Accessed Oct. 2022.

Taulli, Tom. "What You Need To Know About Dark Data." *Forbes*, 27 Oct. 2019. Accessed Oct. 2022.

Taylor, Linnet. "What is data justice? The case for connecting digital rights and freedoms globally." *Big Data & Society*, July-Dec 2017. Accessed Aug 2022.

"Twitter: Data Collection With API Research Paper." *IvyPanda*, 28 April 2022. Accessed Aug. 2022.

"Using governance automation to reduce data risk." *Nintex*, 15 Nov. 2021. Accessed Oct. 2022

"Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025." *Statista*, 8 Sept. 2022. Accessed Oct 2022.

Wang, R. "Monday's Musings: Beyond The Three V's of Big Data – Viscosity and Virality." *Forbes*, 27 Feb. 2012. Accessed Aug 2022.

"What is a data fabric?" *IBM*, n.d. Accessed Aug 2022.

Yego, Kip. "Augmented data management: Data fabric versus data mesh." *IBM*, 27 April 2022. Accessed Aug 2022.

Thank you