Unit 2

Big Data Analytics and Big Data Analytics Techniques

Big Data Analytics and Big Data Analytics Techniques

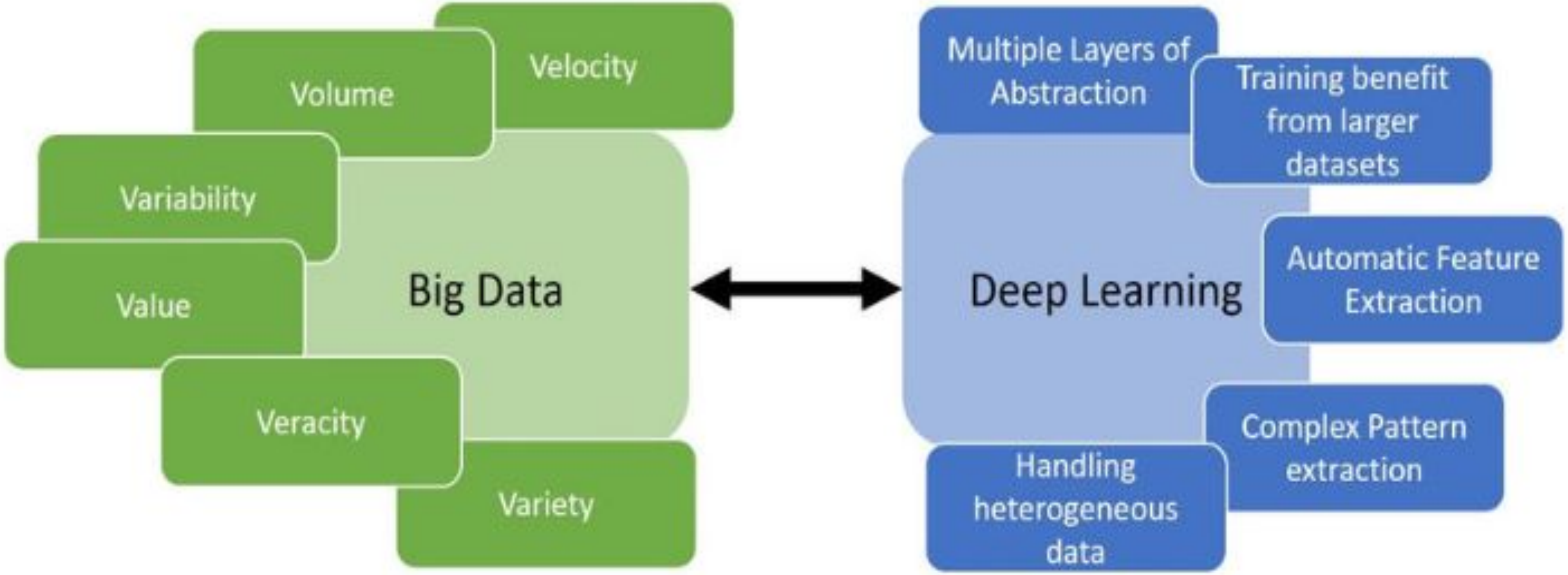
- Big Data and its Importance, Drivers for Big data, Optimization techniques, Dimensionality Reduction techniques, Time series Forecasting,
- Social Media Mining and Social Network Analysis and its Application,
- Big Data analysis using Hadoop, Pig, Hive, Mongodb, Spark and Mahout, Data analysis techniques like Discriminant Analysis and Cluster Analysis,
- Introduction to NOSQL (Neo4j) and MongoDB, Hive Architecture, HBase concepts, PIG, Zookeeper - how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper, No SQL databases: Cassandra and HBase (columnar), MongoDB and Elastic Search (document-based), Neo4j (graph based) 12 hrs

BDA Final Evaluation-30 Marks

1. Quiz-CO1, CO2-Unit 1 ,Unit 2-**12** Marks- Individual submission-31 Aug 24.
2. Poster-CO3-Unit 3-**6** Marks- Group submission-22 Sept 24.
3. Case study report-CO4,CO5-Unit 4, Unit 5-5 Marks-**12** Marks- Individual submission-20 Oct 24.

https://www.bigdataframework.org/business-drivers-for-big-data/

- Six main business drivers can be identified:
- The digitization of society;
- The plummeting of technology costs;
- Connectivity through cloud computing;
- Increased knowledge about data science;
- Social media applications;
- The upcoming Internet-of-Things (IoT).


Velocity
Volume
Variability
Value
Veracity
Variety
Big Data
Deep Learning
Multiple Layers of Abstraction
Training benefit from larger datasets
Automatic Feature Extraction
Complex Pattern extraction
Handling heterogeneous data

Optimization techniques in big data

- Optimization is the process of making something, such as a system or process, perform as efficiently and effectively as possible. It involves finding the best solution among a set of possible solutions, by adjusting various parameters and variables. Optimization is important because it helps to improve the performance of a system, process or model, and can also help to reduce costs, increase efficiency and improve the quality of the results.
- **In the context of big data**, optimization can refer to techniques used to improve the performance of data processing and analysis tasks, such as reducing the amount of time it takes to complete a task, or reducing the amount of resources (e.g., memory, storage) required to perform a task.

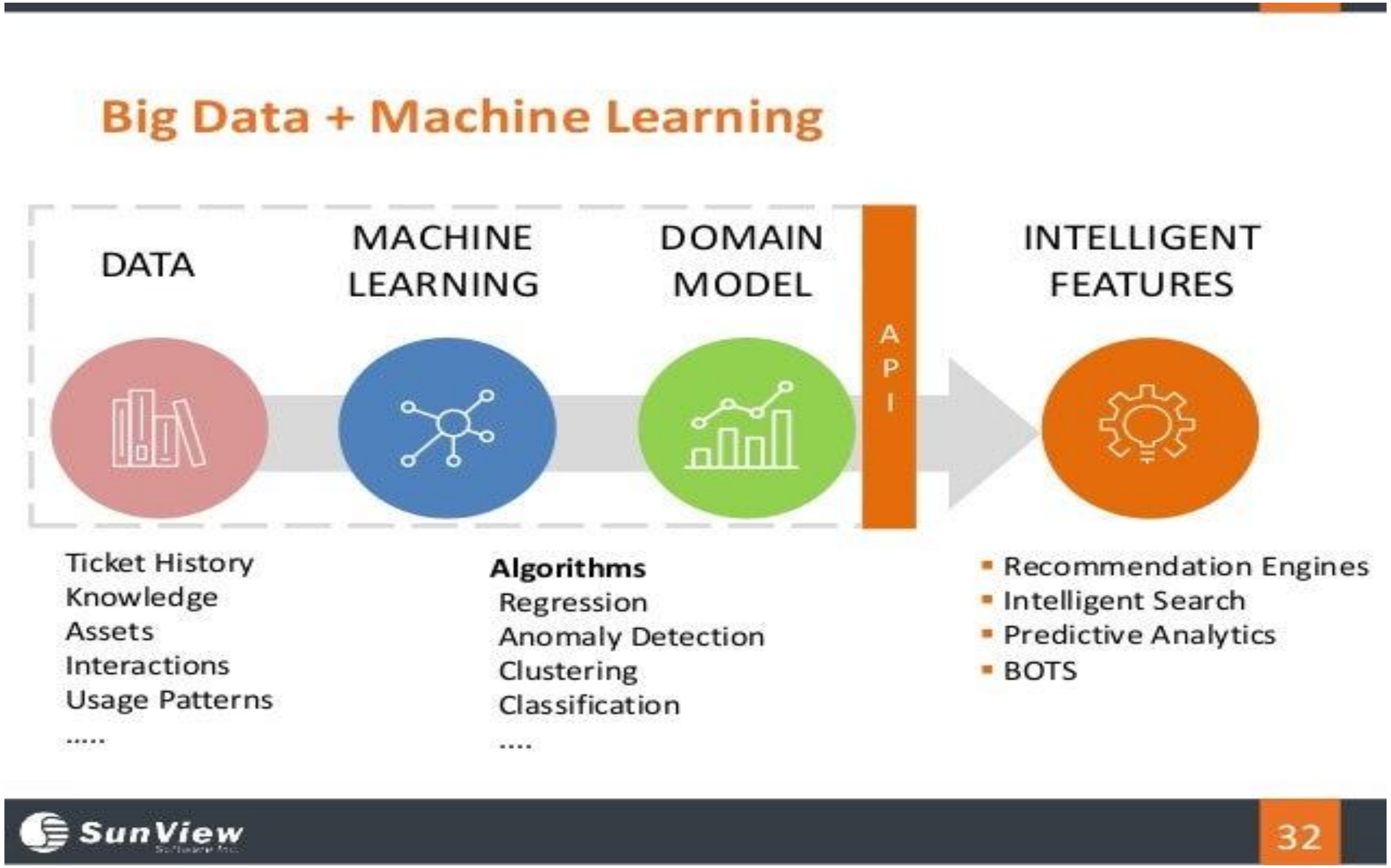
Why is optimization important ?

- **Improved Performance:** Optimization can improve the performance of a system, process or model by reducing the time and resources required to execute it, or by improving the accuracy and scalability of the results.
- **Cost Reduction:** Optimization can help to reduce costs by finding more efficient ways of performing a task, or by reducing the amount of resources required.
- **Increased Efficiency:** Optimization can increase the efficiency of a system, process or model by finding ways to perform the same task with fewer steps, or by reducing the amount of resources required.
- **Better Quality of Results:** Optimization can improve the quality of the results by increasing the accuracy of a model or by finding the best solution among a set of possible solutions.
- **Scalability:** Optimization can help to improve the scalability of a system, process or model, allowing it to handle increasing amounts of data or workloads.
- **Security and Privacy :** Optimization can also help to improve the security and privacy of a system by reducing the attack surface and minimizing the risk of data breaches.
- **Innovation:** Optimization can also play an important role in innovation, as finding new and more efficient ways of performing a task can lead to new products and services.

There are several optimization techniques that can be used to improve the performance of big data systems, some of which include:

- **Data Compression:** Compressing data before it is stored or transmitted can reduce the amount of storage and network resources required, and can also improve the performance of read and write operations.
- **Data Partitioning:** Splitting large data sets into smaller chunks, which are processed in parallel across multiple nodes in a cluster, can improve the performance of operations that require data shuffling, such as join and groupBy operations
- **Data Caching:** Caching frequently accessed data in memory can reduce the number of times the data needs to be read from disk, which can improve performance.
- **Query Optimization:** Rewriting or optimizing queries can improve the performance of data retrieval operations.
- **Data Sampling:** Analyzing a representative subset of a large data set can improve the performance of data analysis tasks, without sacrificing the accuracy of the results.
- **Algorithm Optimization:** Optimizing the algorithms used to process and analyze big data can improve performance.
- **Distributed Processing:** Using a distributed computing system can improve the scalability of big data systems by allowing data to be processed in parallel across multiple machines.

- **Automation:** Using automation techniques like machine learning and artificial intelligence to optimize the big data process can improve performance.
- **Cloud computing:** Using cloud-based big data services can provide more resources for processing and storing large data sets and can help to optimize costs.
- **Data indexing:** Creating indexes on data can improve the performance of search and query operations, by reducing the amount of data that needs to be scanned.
- **Data summarization:** Summarizing data can reduce the amount of data that needs to be processed, and can also improve the performance of query and analysis operations.
- **Data replication:** Replicating data across multiple nodes in a cluster can improve the performance of read operations and provide high availability of data.
- **Predicate pushdown:** Pushing down filtering conditions to the data source can reduce the amount of data that needs to be read and processed.
- **Column pruning:** Eliminating unnecessary columns from the dataset can reduce the amount of data that needs to be processed.
- **Code generation:** Generating optimized code for operations such as filtering and aggregations can improve the performance of these operations.
- **Data Skew handling :** Handling skewed data distribution by using techniques such as bucketing, sampling or replication can improve the performance of query and analysis operations.



Big Data + Machine Learning

DATA

MACHINE LEARNING

DOMAIN MODEL

API

INTELLIGENT FEATURES

Ticket History
Knowledge
Assets
Interactions
Usage Patterns
.....

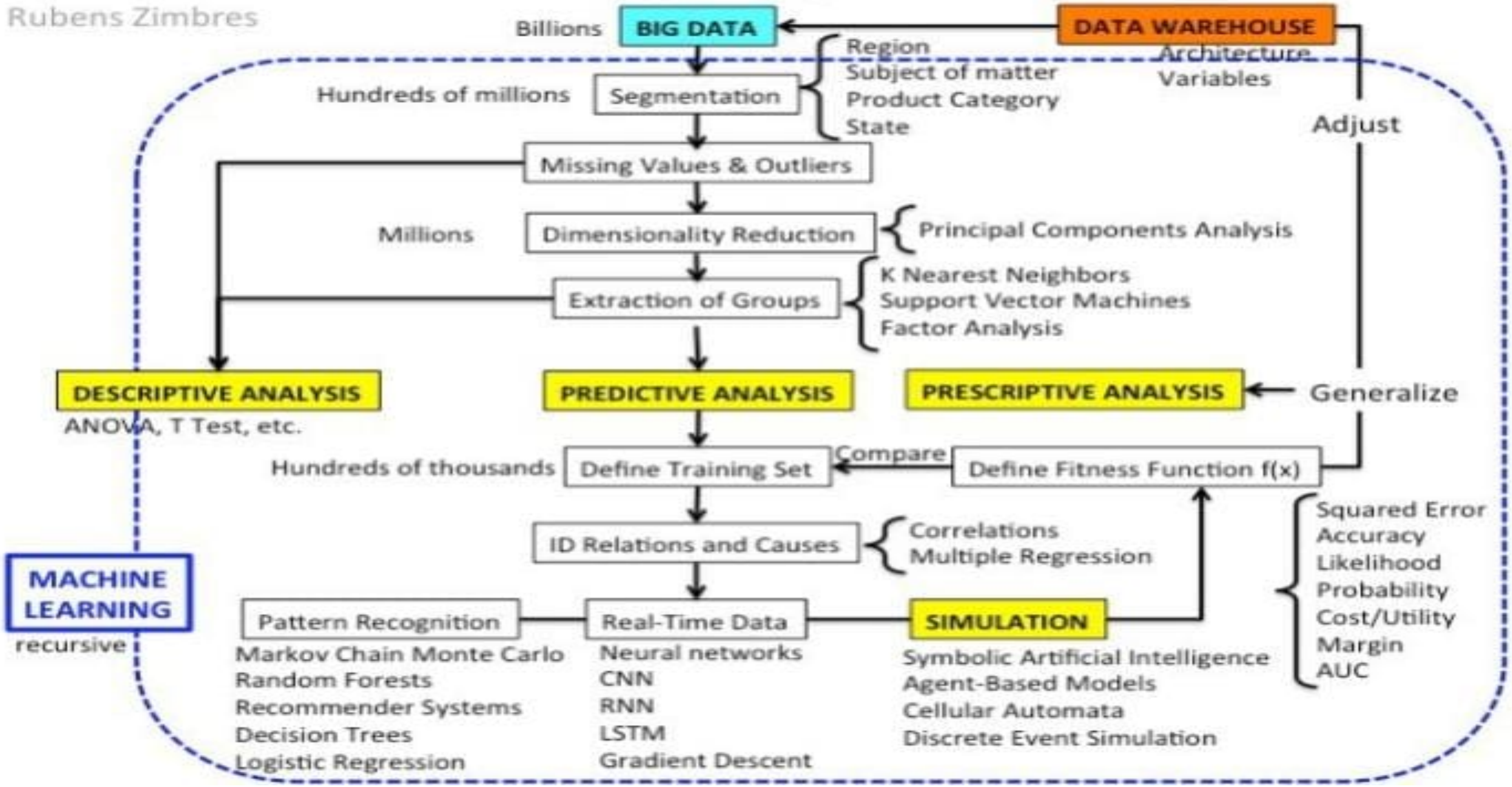
Algorithms
Regression
Anomaly Detection
Clustering
Classification
....

- Recommendation Engines
- Intelligent Search
- Predictive Analytics
- BOTS

SunView Software Inc.

Machine Learning Applied to Big Data

Rubens Zimbres



https://www.kdnuggets.com/2017/07/machine-learning-big-data-explained.html

Dimensionality Reduction

- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9036908
- Hence, to counter the curse of dimensionality, dimensionality reduction techniques come to the rescue. The answers to the question of why dimensionality reduction is useful are:
- The model performs more accurately since redundant data will be removed, which will lead to less room for assumption.
- Less usage of computational resources, which will save time and financial budget
- A few machine learning/Deep Learning techniques do not work on high-dimensional data, a problem that will be solved once the dimension reduces.
- Clean and non-sparse data will give rise to more statistically significant results because clustering of such data is easier and more accurate.

Advantages

- Storage space and the processing time are less
- Multi-collinearity of the dependent variables is removed
- Reduced chances of overfitting the model
- Data Visualization becomes easier

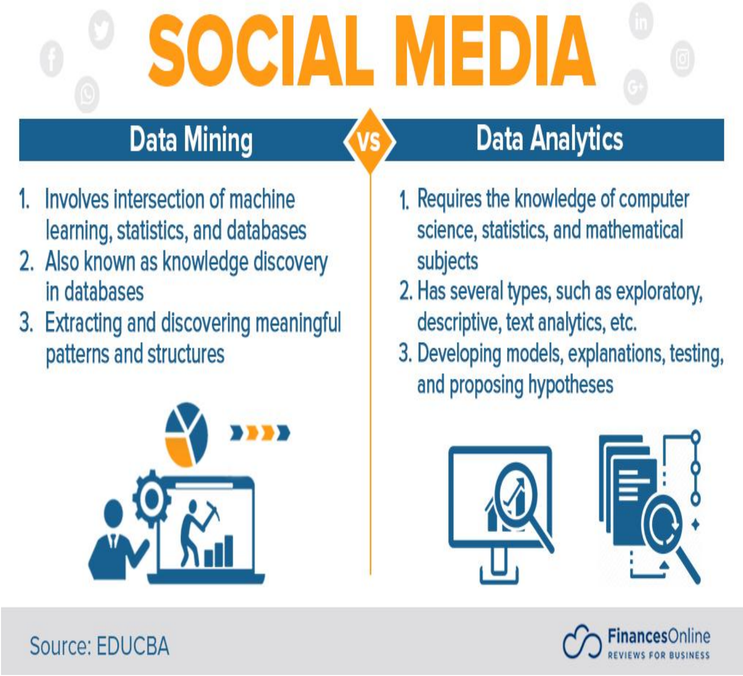
Disadvantages

- Some amount of data is lost.
- PCA cannot be applied where data cannot be defined through mean and covariance.
- Not every variable needs to be linearly correlated, which PCA tends to find.
- Labeled data is required for LDA to function, which is not available in a few cases.

Time series forecasting

Time series forecasting is a technique for the prediction of events through a sequence of time. It predicts future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends. It is used across many fields of study in various applications including:

- Astronomy
- Business planning
- Control engineering
- Earthquake prediction
- Econometrics
- Mathematical finance
- Pattern recognition
- Resources allocation
- Signal processing
- Statistics
- Weather forecasting
- https://www.geeksforgeeks.org/time-series-analysis-and-forecasting/



SOCIAL MEDIA

Data Mining VS **Data Analytics**

Data Mining

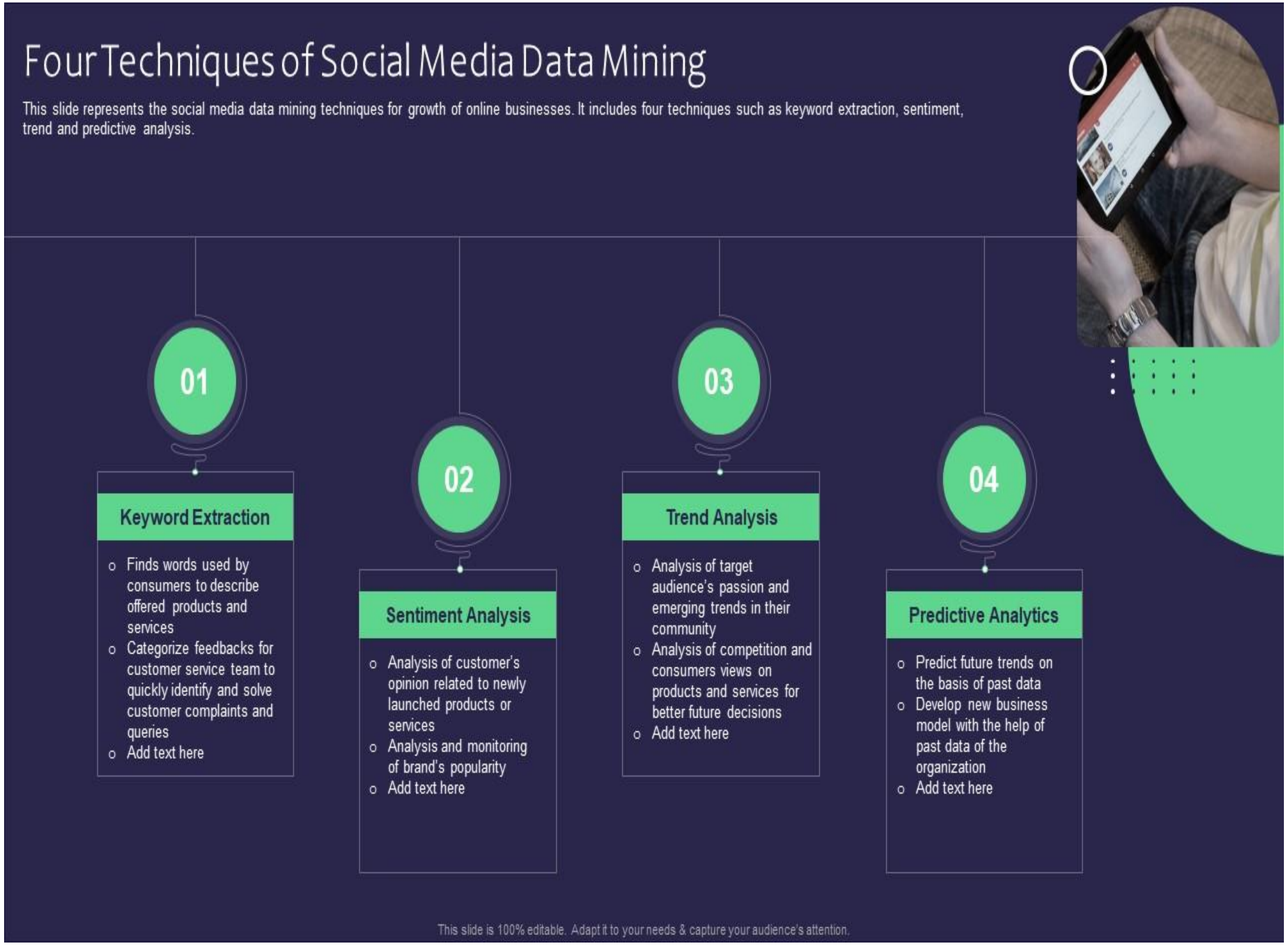
1. Involves intersection of machine learning, statistics, and databases
2. Also known as knowledge discovery in databases
3. Extracting and discovering meaningful patterns and structures

Data Analytics

1. Requires the knowledge of computer science, statistics, and mathematical subjects
2. Has several types, such as exploratory, descriptive, text analytics, etc.
3. Developing models, explanations, testing, and proposing hypotheses

Source: EDUCBA

FinancesOnline REVIEWS FOR BUSINESS



Four Techniques of Social Media Data Mining

This slide represents the social media data mining techniques for growth of online businesses. It includes four techniques such as keyword extraction, sentiment, trend and predictive analysis.

01 **Keyword Extraction**

- Finds words used by consumers to describe offered products and services
- Categorize feedbacks for customer service team to quickly identify and solve customer complaints and queries
- Add text here

02 **Sentiment Analysis**

- Analysis of customer's opinion related to newly launched products or services
- Analysis and monitoring of brand's popularity
- Add text here

03 **Trend Analysis**

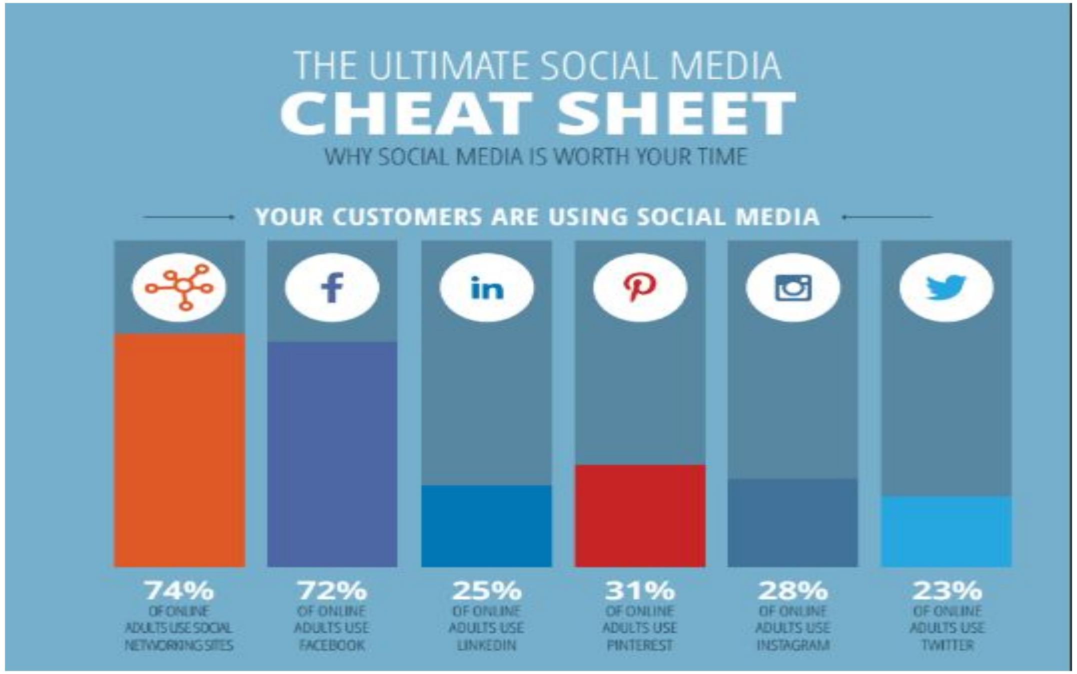
- Analysis of target audience's passion and emerging trends in their community
- Analysis of competition and consumers views on products and services for better future decisions
- Add text here

04 **Predictive Analytics**

- Predict future trends on the basis of past data
- Develop new business model with the help of past data of the organization
- Add text here

This slide is 100% editable. Adapt it to your needs & capture your audience's attention.

https://www.suntecindia.com/blog/social-media-data-mining/

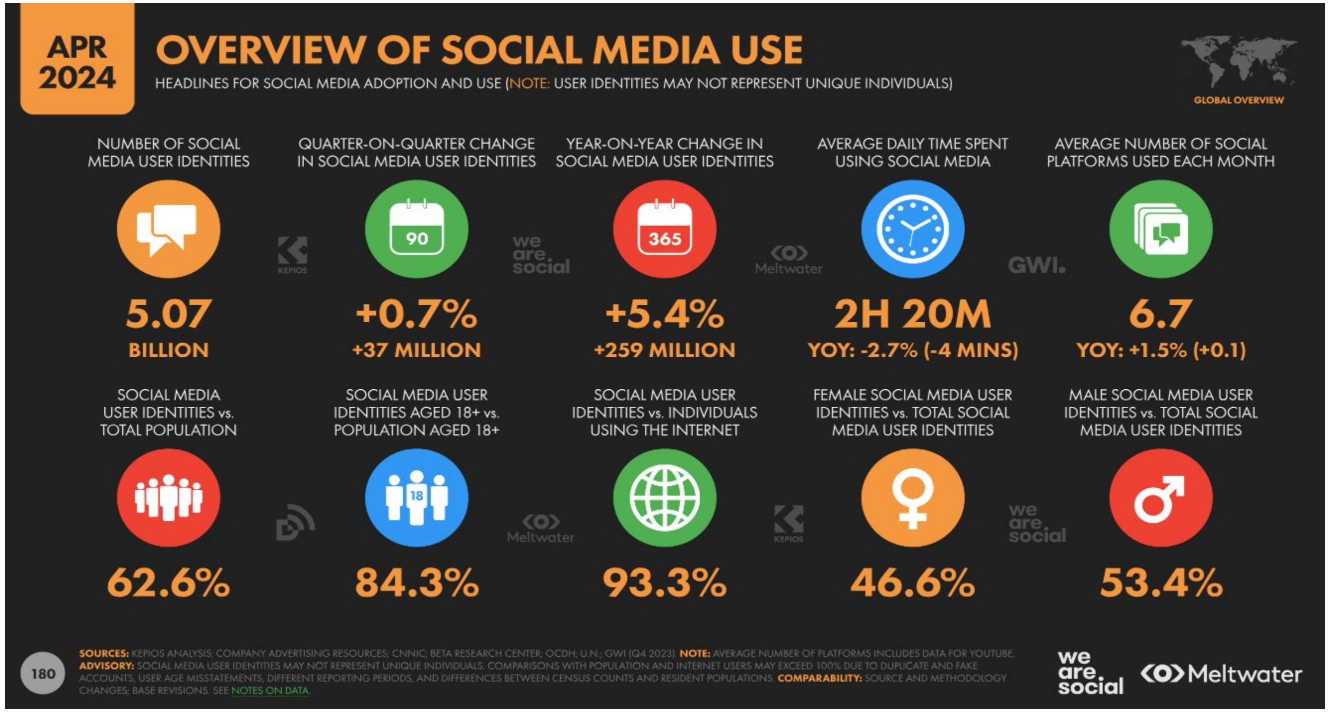


THE ULTIMATE SOCIAL MEDIA

CHEAT SHEET

WHY SOCIAL MEDIA IS WORTH YOUR TIME

YOUR CUSTOMERS ARE USING SOCIAL MEDIA



APR 2024

OVERVIEW OF SOCIAL MEDIA USE

HEADLINES FOR SOCIAL MEDIA ADOPTION AND USE (NOTE: USER IDENTITIES MAY NOT REPRESENT UNIQUE INDIVIDUALS)

GLOBAL OVERVIEW

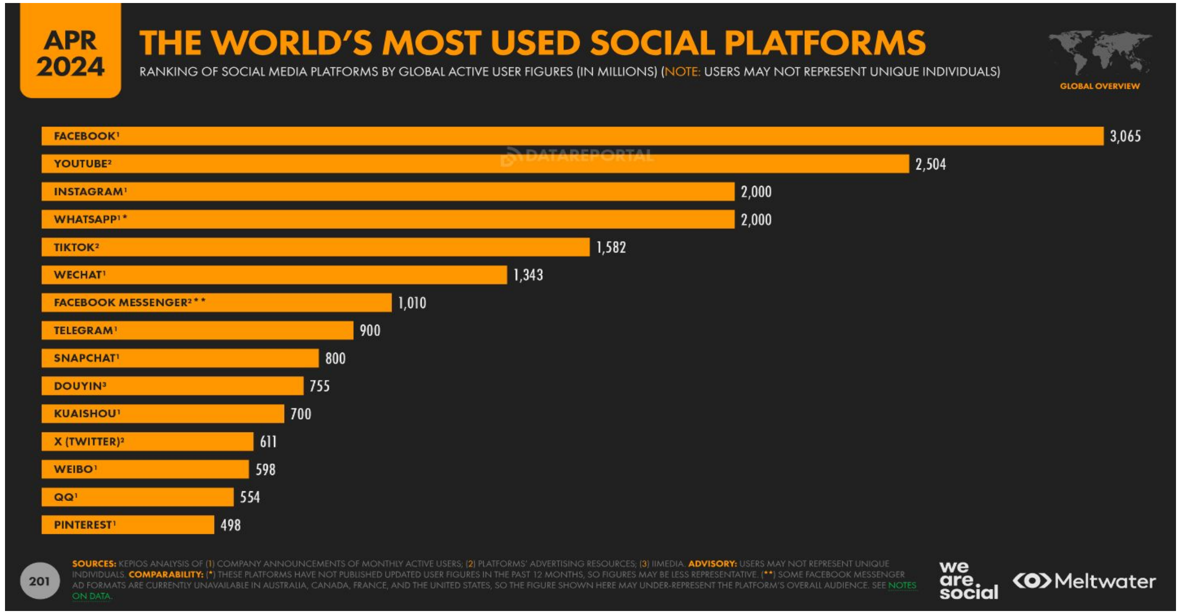
NUMBER OF SOCIAL MEDIA USER IDENTITIES	QUARTER-ON-QUARTER CHANGE IN SOCIAL MEDIA USER IDENTITIES	YEAR-ON-YEAR CHANGE IN SOCIAL MEDIA USER IDENTITIES	AVERAGE DAILY TIME SPENT USING SOCIAL MEDIA	AVERAGE NUMBER OF SOCIAL PLATFORMS USED EACH MONTH
5.07 BILLION	+0.7% +37 MILLION	+5.4% +259 MILLION	2H 20M YOY: -2.7% (-4 MINS)	6.7 YOY: +1.5% (+0.1)

SOCIAL MEDIA USER IDENTITIES vs. TOTAL POPULATION	SOCIAL MEDIA USER IDENTITIES AGED 18+ vs. POPULATION AGED 18+	SOCIAL MEDIA USER IDENTITIES vs. INDIVIDUALS USING THE INTERNET	FEMALE SOCIAL MEDIA USER IDENTITIES vs. TOTAL SOCIAL MEDIA USER IDENTITIES	MALE SOCIAL MEDIA USER IDENTITIES vs. TOTAL SOCIAL MEDIA USER IDENTITIES
62.6%	84.3%	93.3%	46.6%	53.4%

180

SOURCES: KEPIOS ANALYSIS; COMPANY ADVERTISING RESOURCES; CNNIC; BETA RESEARCH CENTER; OCDH; U.N.; GWI (Q4 2023). NOTE: AVERAGE NUMBER OF PLATFORMS INCLUDES DATA FOR YOUTUBE. ADVISORY: SOCIAL MEDIA USER IDENTITIES MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARISONS WITH POPULATION AND INTERNET USERS MAY EXCEED 100% DUE TO DUPLICATE AND FAKE ACCOUNTS, USER AGE MISSTATEMENTS, DIFFERENT REPORTING PERIODS, AND DIFFERENCES BETWEEN CENSUS COUNTS AND RESIDENT POPULATIONS. COMPARABILITY: SOURCE AND METHODOLOGY CHANGES; BASE REVISIONS. SEE NOTES ON DATA.

we are social Meltwater



APR 2024

THE WORLD'S MOST USED SOCIAL PLATFORMS

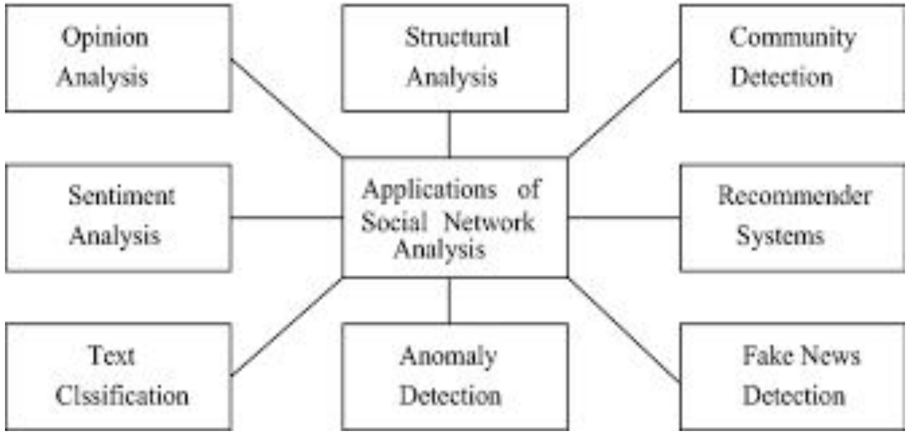
RANKING OF SOCIAL MEDIA PLATFORMS BY GLOBAL ACTIVE USER FIGURES (IN MILLIONS) (NOTE: USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS)

GLOBAL OVERVIEW

201

SOURCES: KEPIOS ANALYSIS OF (1) COMPANY ANNOUNCEMENTS OF MONTHLY ACTIVE USERS; (2) PLATFORMS' ADVERTISING RESOURCES; (3) IIMEDIA. ADVISORY: USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARABILITY: (*) THESE PLATFORMS HAVE NOT PUBLISHED UPDATED USER FIGURES IN THE PAST 12 MONTHS, SO FIGURES MAY BE LESS REPRESENTATIVE. (**) SOME FACEBOOK MESSENGER AD FORMATS ARE CURRENTLY UNAVAILABLE IN AUSTRALIA, CANADA, FRANCE, AND THE UNITED STATES, SO THE FIGURE SHOWN HERE MAY UNDER-REPRESENT THE PLATFORM'S OVERALL AUDIENCE. SEE NOTES ON DATA.

we are social Meltwater



Social Media Data Mining Methods

- As its name implies, social media data mining refers to the process of mining social data. Unlike regular data mining, social media data mining explores beyond the internal databases and systems of a given company or research firm.
- It typically involves the collection, processing, and analysis of raw data obtained from social media platforms such as Facebook, Instagram, Twitter, TikTok, LinkedIn, YouTube, and others, to uncover meaningful patterns and trends, draw conclusions, and provide insightful and actionable information.

What data is collected for social media data mining?

- Social media data mining harvests various types of social data that are either publicly available (e.g., age, gender, job profession, geographic location, etc.) or are generated on a daily basis on social media platforms (e.g., comments, likes, clicks, etc.).
- Typically, the data represents people’s attitudes, connections, behavior, and feelings towards a certain topic, product, or service. Depending on the social media platform in question, this data may include the number of followers, comments, likes, or shares, if the targeted social media data comes from Facebook, Twitter’s retweets or the number of impressions, or Instagram’s engagement rates and hashtag usage.
- When trying to optimize your social content, promote your online business, discover influential customers, or improve marketing and engagement strategies, you should always focus on gathering the above-mentioned types of data.

How does social data mining work?

- Combination of **statistical techniques, mathematics, and machine learning**.
- The first step is to gather and process social data from different social media sources.
- Apart from social media platforms such as Facebook, Twitter, or YouTube, data miners also extract data from various blogs, news sites, forums, or any other public pages
- Once data is collected and processed, what follows is the application of various data mining techniques which allow for easier identification of common patterns and the correlation of various data points in large datasets.
- Classification, association, tracking patterns, predictive analytics, keyword extraction, sentiment analysis, and market/trend analysis.
- Ex.Microsoft SharePoint, Sisense, IBM Cognos, RapidMiner, and Dundas BI.
- This is usually done by using social media analytics or a variety of data visualization tools, such as Infogram, ChartBlocks, Tableau, and Datawrapper.

How is social media used and who’s using it?

Some of its major uses in businesses include **targeted marketing campaigns, market research, sales enablement, predictive analytics, influencer marketing, and monitoring of brand reputation.**

- Trend analysis - Businesses use social media data mining to gain valuable insights into currently trending keywords, mentions, and topics on social media platforms.
- Event detection (social heat mapping) - This metric is of great importance for agencies and researchers who use social media monitoring.
- Social spam detection - Social media data mining allows for easier detection of spammers and bots on social media platforms like Instagram and Twitter.
- Ecommerce - Social media data mining is used to analyze how people talk about products.
- Digital media –The content that is to be shown on a particular digital billboard may be decided upon through conducting a social media data mining process in order to cater to the audience’s preferences or needs.
- Bloggers and social media influencers - Social media data mining is often used by bloggers and social media influencers to help them analyze the attitudes and feelings of their followers, what they are talking about, and how they feel about certain topics of discussion.
- Brands - Social media data mining helps brands with important decision-making, for example, when deciding about potential future markets.
- Research purposes - Social media data mining can be applied in different research domains, including social science, research, health research, and technology research.
- Government agencies - Social media data mining is also increasingly being used by government agencies for the purpose of welfare-focused interventions.

Key capabilities of effective social media analytics

The first step for effective social media analytics is developing a **goal**. Goals can range from **increasing revenue to pinpointing service issues**.

Sources also need to be specified — responses to YouTube videos, Facebook conversations, Twitter arguments, Amazon product reviews, comments from news sites. It is important to select sources pertinent to a given product, service or brand.

These steps are typical of a general social media analytics approach that can be made more effective by capabilities found in social media analytics platforms.

Techniques

- **Natural language processing and machine learning technologies** identify entities and relationships in unstructured data — information not pre-formatted to work with data analytics. Virtually all social media content is unstructured.
- **Segmentation** is a fundamental need in social media analytics. It categorizes social media participants by geography, age, gender, marital status, parental status and other demographics. It can help identify influencers in those categories.
- **Behavior analysis** is used to understand the concerns of social media participants by assigning behavioral types such as user, recommender, prospective user and detractor.
- **Sentiment analysis** measures the tone and intent of social media comments. It typically involves natural language processing technologies to help understand entities and relationships to reveal positive, negative, neutral or ambivalent attributes.
- **Share of voice** analyzes prevalence and intensity in conversations regarding brand, products, services, reputation and more. It helps determine key issues and important topics. It also helps classify discussions as positive, negative, neutral or ambivalent.
- **Clustering analysis** can uncover hidden conversations and unexpected insights. It makes associations between keywords or phrases that appear together frequently and derives new topics, issues and opportunities.
- **Dashboards and visualization** charts, graphs, tables and other presentation tools summarize and share social media analytics findings — a critical capability for communicating and acting on what has been learned.

The best social media analytics tools for 2024

- 1. Hootsuite
- 2. Sprout Social
- 3. Buffer
- 4. Hubspot
- 5. Later
- 6. Rival IQ
- 7. Talkwalker
- 8. Brandwatch
- 9. Keyhole
- 10. Channelview Insights
- 11. Mentionlytics
- 12. Panoramiq Insights
- 13. Quintly
- 14. Iconosquare
- 15. Google Analytics
- 16. Meta Business Suite Insights
- 17. Instagram Insights
- 18. TikTok Analytics
- 19. X Analytics
- 20. Pinterest Analytics
- 21. LinkedIn Page Analytics

Some links to explore

- https://sproutsocial.com/insights/social-media-analytics/
- https://buffer.com/library/social-media-analytics-tools/
- https://www.ibm.com/topics/social-media-analytics

What is the importance of social media analytics?

1. Trendspotting

- Trendspotting is the act of pinpointing upcoming trends before they're mainstream. Keeping a close eye on your social media analytics can help you do just that. Some of the trends that your social media analytics can help you determine include:
- Which platforms are gaining or losing traction and popularity
- Topics of interest that your audience is talking about (and brand mentions in conversations)
- Types of ads that interest your audience
- Rising influencers and products in your niche or industry
- Types of content that your audience engages with most

2. Brand sentiment

- Brand sentiment illustrates how people are feeling about your brand. It includes all positive, neutral and negative feelings that are discussed online. By looking through your social media analytics, you can review and measure your brand sentiment through a **sentiment analysis software**.
- This helps ensure your audience is happy with your business and enables you to detect opportunities to make amends with unsatisfied customers. And you can uncover opportunities to improve your business.
- For instance, through sentiment analysis you could discover your customers are asking the same questions about a particular product feature, enabling you update your FAQ page or help center. Sentiment analysis can be used with competitor analysis because you can pinpoint new competitors and related topics your customers are buzzing about that you may have not considered before.

3. Value perception

- Value perception (or perceived value) refers to the overall customer opinion of your brand's product or service and whether or not it can meet their needs. Perceived value is key to determining demand and the price point of a product or service. For example, if your product has a low perceived value, customers won't be willing to pay much for it.
- You can measure value perception by using **social listening tools** and monitoring data from other digital marketing dashboards, such as **Google Analytics**. This can help guide the content you create to improve value perception and make sure you're showcasing how your product or service can hit key pain points.
- https://blog.hubspot.com/service/social-listening-tools

4. Setting social media goals

- Social media analytics can also help you see which channels and content are performing well, so you can create actionable, realistic **social media goals and objectives**.
- The key word here is realistic. If you take a look at your social media analytics reports and realize your Instagram account is growing by 10 followers per week, trying to jump from 5,000 followers to 10,000 followers in a single quarter is not a realistic goal, even if you revamp your posting strategy. You might instead try to make a goal where your account starts growing by 20 followers per week instead and steadily increase that goal from there.

5. Proving ROI

- Finally, your social media analytics can help prove the ROI of your social media marketing efforts.
- Each time you run a new campaign, monitor your social analytics to see how the content is performing, if people are clicking over to your website and if you're generating new sales.
- Doing this demonstrates **social media ROI** so teams can earn more buy-in and resources. **UTM tracking and URL shortening** are two ways that make proving ROI via analytics even easier.
- This way, you can attribute specific pipeline and purchases to your social media efforts.

What are the types of social media analytics?

- There are several different types of social media analytics you should monitor in your **social media dashboard** that will guide your strategy and discover valuable insights. We'll walk you through the six main types of analytics below.
- **Performance analysis**
- First and foremost, you need to measure the overall performance of your social media efforts. This includes **social media metrics** including:
- Impressions
- Reach
- Likes
- Comments
- Shares
- Views
- Clicks
- Sales

Audience analytics

- Next, you'll want to take a look at your audience analytics. This will help you discover which demographics your content is reaching—and ensure they match up to your target audience. If not, you may need to adjust your content strategy to better attract your ideal customer profile.
- Audience analytics will include data like:
- Age
- Gender
- Location
- Device

Competitor analysis

- Another key area to look into is how your competitors perform on social media. How many followers do they have? What is their **engagement rate**? How many people seem to engage with each of their posts?
- You can then compare this data to your own to see how you stack up—as well as set more realistic growth goals. Using a tool like Sprout, you can gather all of this data in one place and measure it network by network.
- Pay attention to how your benchmarks stand up to your competitors and consider adjusting your social media strategy to take advantage of opportunity gaps.

Paid social analytics

- When you're putting money behind specific social media posts, you want to make sure they're driving results. This is why you absolutely need to pay close attention to your **paid social** analytics.
- Some of the most important ad analytics to measure include:
- Total number of active ads
- Clicks
- Click-through rate
- Cost-per-click
- Cost-per-engagement
- Cost-per-action
- Conversion rate
- Total ad spend
- Each **social media platform** that you run ads through will have its own dashboard to provide you with all of this information, but you may want to create your own spreadsheet as well to track total ads and ad spend.

Influencer analysis

- If you're running **influencer marketing campaigns**, tracking the success of these partnerships is essential to proving ROI. We recommend using the **five W's + H of influencer marketing** to inform your strategy and measure ROI at each stage of the buyer journey.
- Some of the data you'll want to keep track of includes:
- Number of posts created per influencer
- Total number of interactions per post
- Audience size of each influencer
- Hashtag usage and engagement
- This can help you gauge overall engagement from your influencer campaigns. If you have an **affiliate marketing** program, you can designate promo codes for each individual influencer to use so your team can track how many sales each partner drives as well.

Sentiment analysis

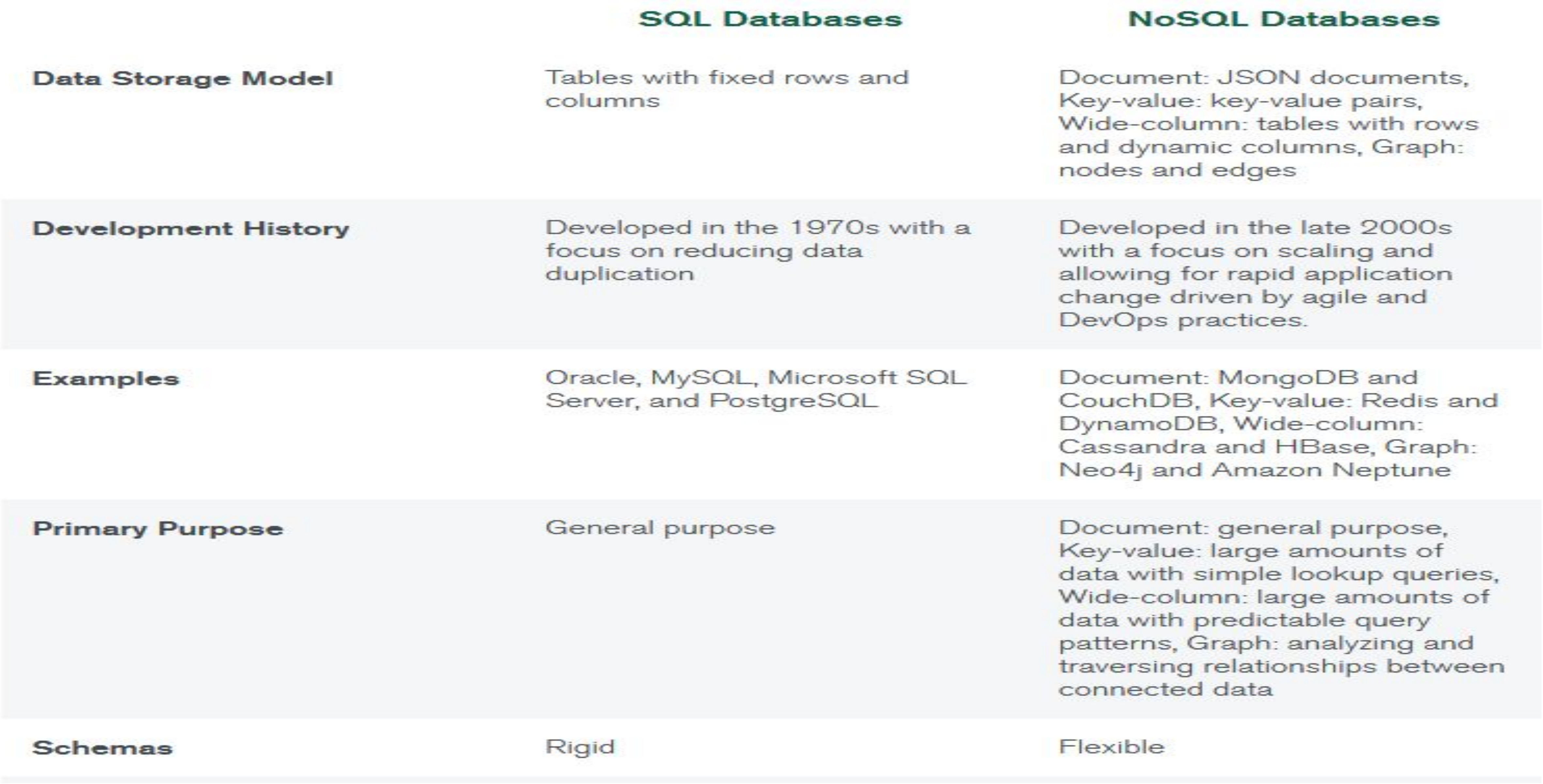
- The last major segment of social media analysis you'll want to track is brand sentiment. Earlier, we talked about how social media analytics tools can help you determine and measure sentiment analysis. But if you want to dig even deeper, use **social listening** to gauge specific connotations around your brand.
- Sprout's Social Listening dashboard helps measure your brand sentiment, showcasing how users feel about your brand or relevant keywords and topics. You can also use **sentiment analysis** in Sprout’s Inbox and Reviews Feed.

SQL (Structured Query Language)

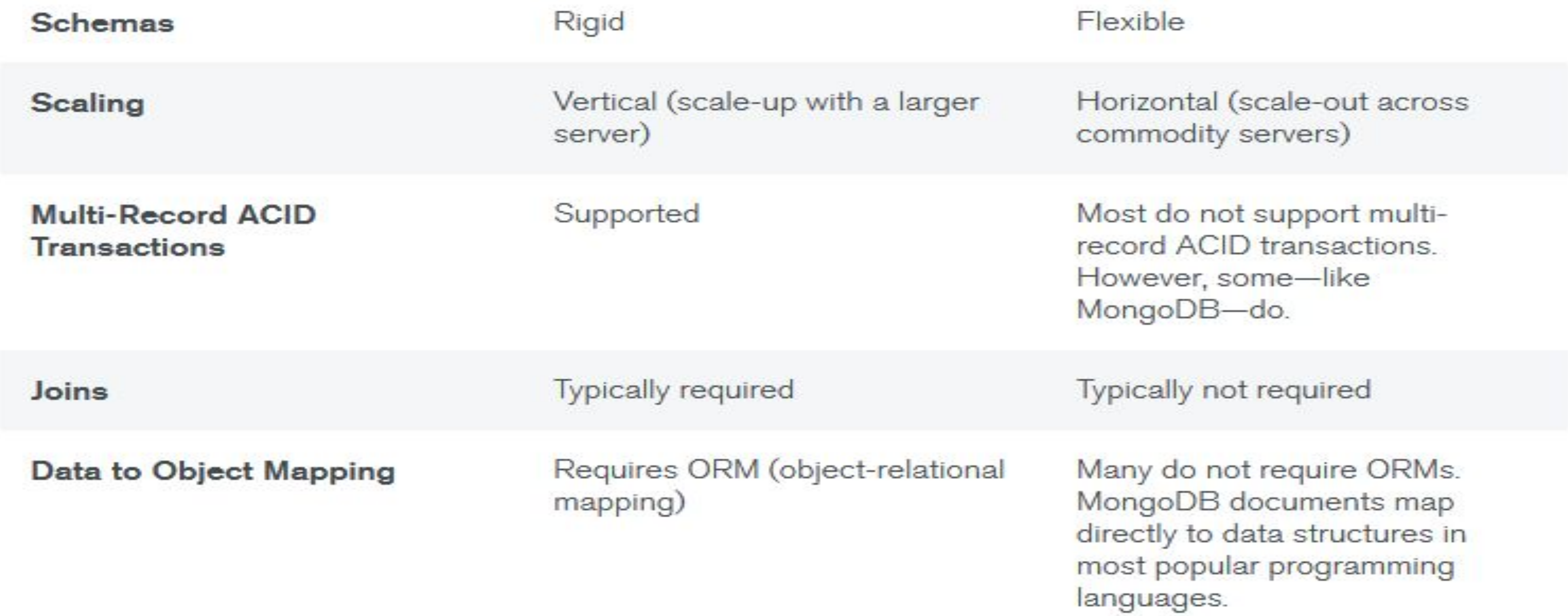
- Most popular databases
- Relational databases accessed by SQL
- Fixed columns and rows
- Early 1970s-storage was extremely expensive
- Efforts to reduce data duplication
- Follow waterfall software development model
- Complex entity-relationship (E-R) diagrams
- Struggled to adapt if requirements changed during the development cycle
- Limitations-Over budget, exceeded deadlines and failed to deliver against user needs.

NoSQL

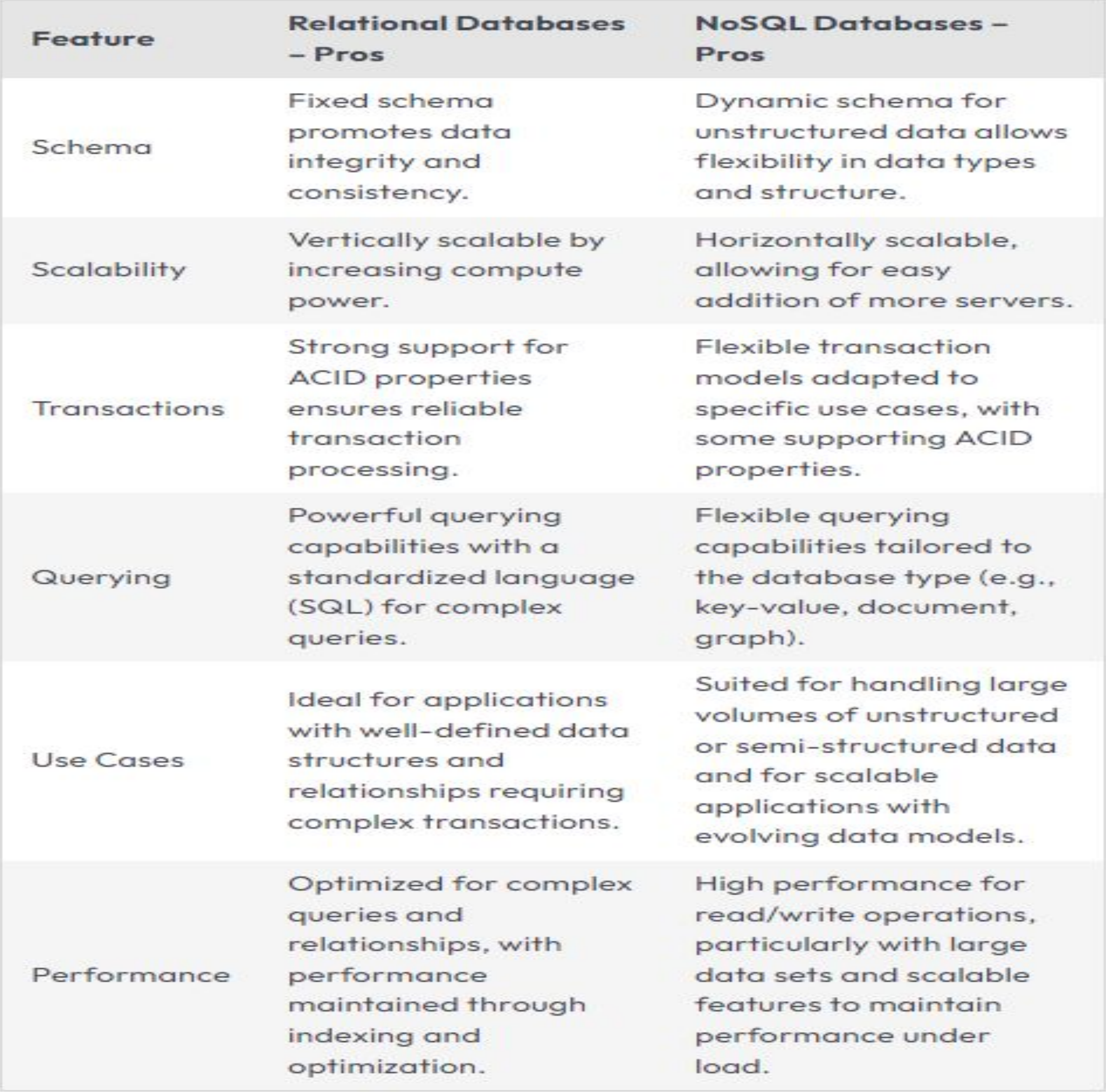
- “NoSQL” stands for “non SQL” “not only SQL.”
- Non-relational database/non tabular database
- Misconception -NoSQL databases or non-relational databases don’t store relationship data well.
- Store **differently** than relational databases do.
- *Easier* than in SQL databases, because related data doesn’t have to be split between tables.
- Nested within a single data structure.
- Late 2000s as the cost of storage dramatically decreased
- Changed the purposes of reducing data duplication
- Developers (rather than storage) were becoming the primary cost of software development
- NoSQL databases optimized for developer productivity
- NoSQL databases come in a variety of types based on their data model.



	SQL Databases	NoSQL Databases
Data Storage Model	Tables with fixed rows and columns	Document: JSON documents, Key-value: key-value pairs, Wide-column: tables with rows and dynamic columns, Graph: nodes and edges
Development History	Developed in the 1970s with a focus on reducing data duplication	Developed in the late 2000s with a focus on scaling and allowing for rapid application change driven by agile and DevOps practices.
Examples	Oracle, MySQL, Microsoft SQL Server, and PostgreSQL	Document: MongoDB and CouchDB, Key-value: Redis and DynamoDB, Wide-column: Cassandra and HBase, Graph: Neo4j and Amazon Neptune
Primary Purpose	General purpose	Document: general purpose, Key-value: large amounts of data with simple lookup queries, Wide-column: large amounts of data with predictable query patterns, Graph: analyzing and traversing relationships between connected data
Schemas	Rigid	Flexible



Schemas	Rigid	Flexible
Scaling	Vertical (scale-up with a larger server)	Horizontal (scale-out across commodity servers)
Multi-Record ACID Transactions	Supported	Most do not support multi-record ACID transactions. However, some—like MongoDB—do.
Joins	Typically required	Typically not required
Data to Object Mapping	Requires ORM (object-relational mapping)	Many do not require ORMs. MongoDB documents map directly to data structures in most popular programming languages.



Feature	Relational Databases – Pros	NoSQL Databases – Pros
Schema	Fixed schema promotes data integrity and consistency.	Dynamic schema for unstructured data allows flexibility in data types and structure.
Scalability	Vertically scalable by increasing compute power.	Horizontally scalable, allowing for easy addition of more servers.
Transactions	Strong support for ACID properties ensures reliable transaction processing.	Flexible transaction models adapted to specific use cases, with some supporting ACID properties.
Querying	Powerful querying capabilities with a standardized language (SQL) for complex queries.	Flexible querying capabilities tailored to the database type (e.g., key-value, document, graph).
Use Cases	Ideal for applications with well-defined data structures and relationships requiring complex transactions.	Suited for handling large volumes of unstructured or semi-structured data and for scalable applications with evolving data models.
Performance	Optimized for complex queries and relationships, with performance maintained through indexing and optimization.	High performance for read/write operations, particularly with large data sets and scalable features to maintain performance under load.

Benefits of NoSQL Databases

- Handle Large Volumes of Data at High Speed with a Scale-Out Architecture
- Store Unstructured, Semi-Structured, or Structured Data
- Enable Easy Updates to Schema and Fields-Flexible data models
- Developer-Friendly
- Take Full Advantage of the Cloud to Deliver Zero Downtime-Fast queries
- Horizontal scaling

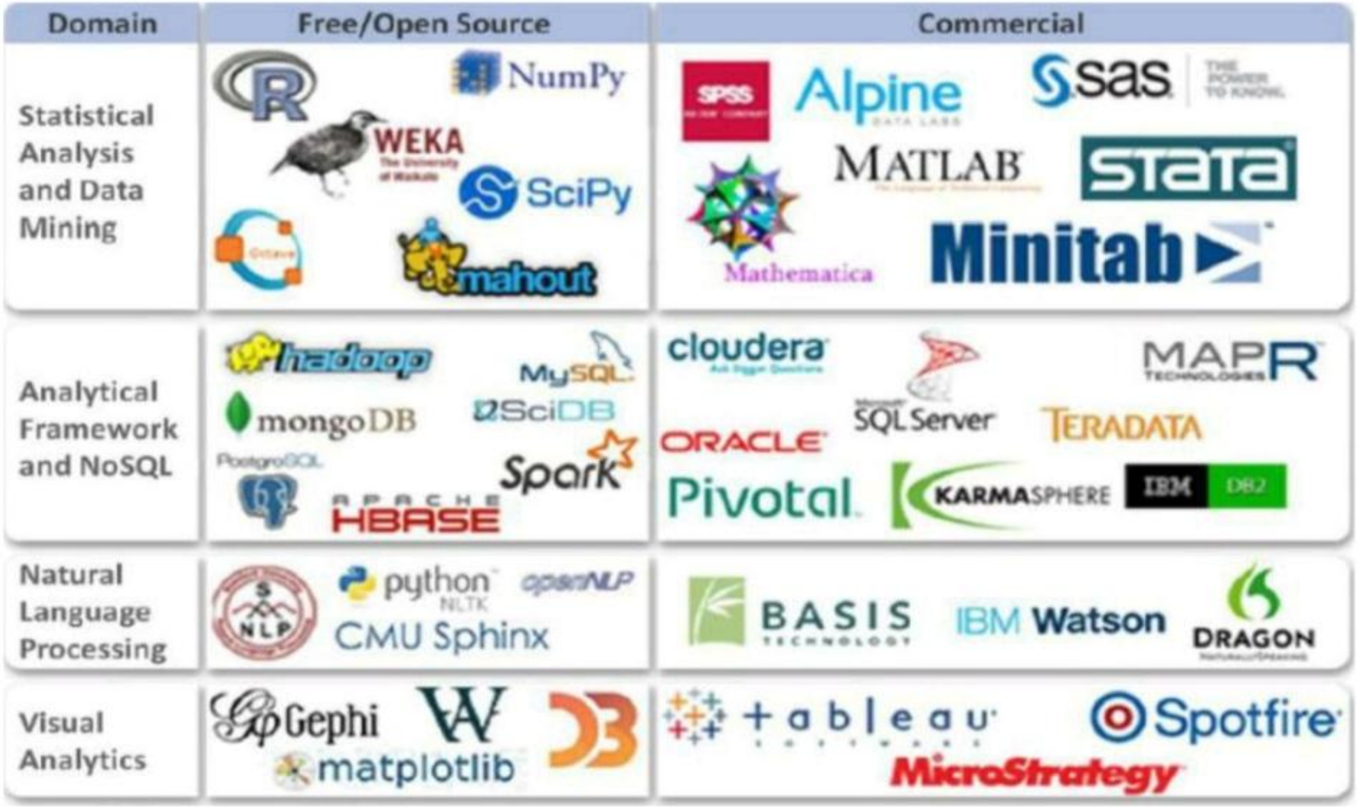
Drawbacks of NoSQL Databases

- Don’t support ACID (atomicity, consistency, isolation, durability)
- Since data models in NoSQL databases are typically optimized for queries and not for reducing data duplication, NoSQL databases can be larger than SQL databases.
- Depending on the NoSQL database type you select, you may not be able to achieve all of your use cases in a single database.
- For example, graph databases are excellent for **analyzing relationships** in your data but may not provide what you need for everyday retrieval of the data such as range queries.
- When selecting a NoSQL database, consider what your use cases will be and if a general purpose database like MongoDB would be a better option.

Motivation to use NoSQL

- SQL databases tend to have rigid, complex, tabular schemas and typically require expensive vertical scaling.
- Provide flexible schemas and scale easily with large amounts of data and high user loads.
- RDBMS focuses on reducing data duplication as storage was much more costly than developer time.
- NoSQL Focuses on scaling, fast queries, allowing for frequent application changes, and making programming simpler for developers.

NoSQL Today (a partial, unrefined list)



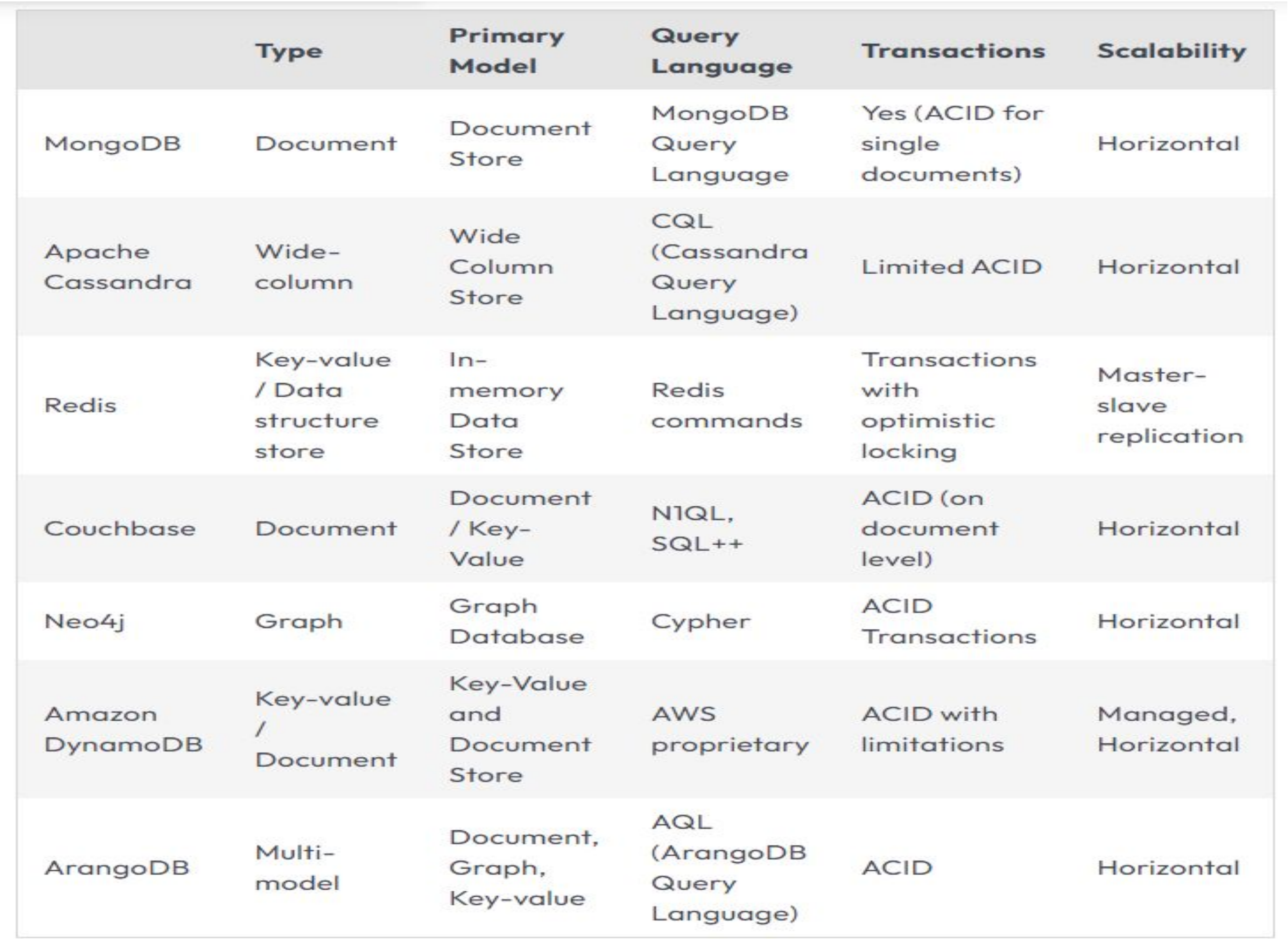
Domain	Free/Open Source	Commercial
Statistical Analysis and Data Mining	R NumPy WEKA SciPy Octave mahout	SPSS Alpine sas MATLAB stata Mathematica Minitab
Analytical Framework and NoSQL	hadoop MySQL mongoDB SciDB PostgreSQL Spark APACHE HBASE	cloudera SQL Server MAPR ORACLE TERADATA Pivotal KARMASPHERE IBM DB2
Natural Language Processing	NLP python NLTK openNLP CMU Sphinx	BASIS IBM Watson DRAGON
Visual Analytics	Gephi W D3 matplotlib	tableau Spotfire MicroStrategy

Figure 1 - Tools and technologies for big data analytics[2]

Resource- https://encyclopedia.pub/entry/10083

https://www.trustradius.com/nosql-databases

https://www.bairesdev.com/blog/nosql-databases/

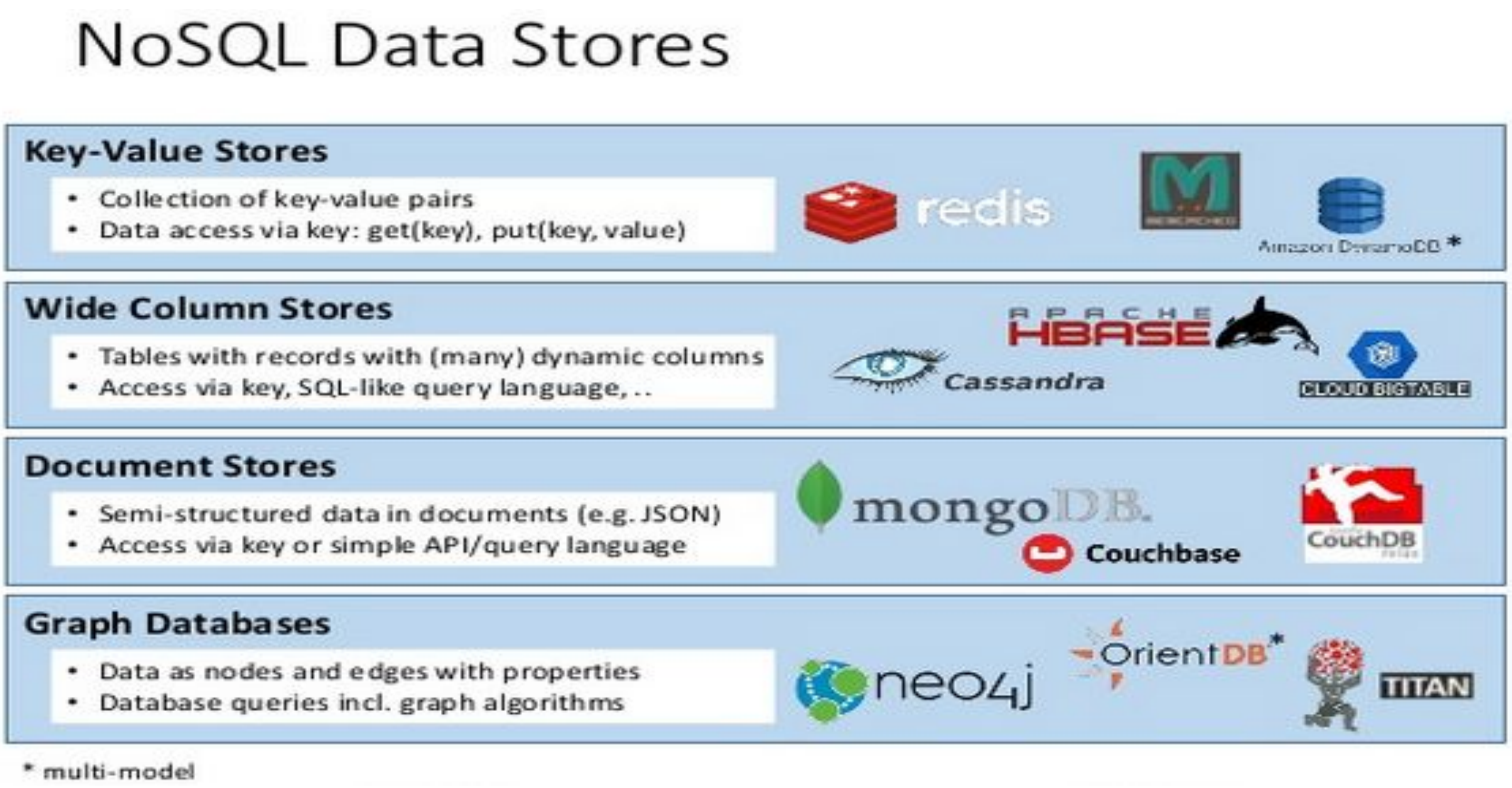


	Type	Primary Model	Query Language	Transactions	Scalability
MongoDB	Document	Document Store	MongoDB Query Language	Yes (ACID for single documents)	Horizontal
Apache Cassandra	Wide-column	Wide Column Store	CQL (Cassandra Query Language)	Limited ACID	Horizontal
Redis	Key-value / Data structure store	In-memory Data Store	Redis commands	Transactions with optimistic locking	Master-slave replication
Couchbase	Document	Document / Key-Value	N1QL, SQL++	ACID (on document level)	Horizontal
Neo4j	Graph	Graph Database	Cypher	ACID Transactions	Horizontal
Amazon DynamoDB	Key-value / Document	Key-Value and Document Store	AWS proprietary	ACID with limitations	Managed, Horizontal
ArangoDB	Multi-model	Document, Graph, Key-value	AQL (ArangoDB Query Language)	ACID	Horizontal

Types of NoSQL Databases

Four major types of NoSQL databases emerged:

- Document databases
- Key-value databases
- Wide-column stores
- Graph databases.



NoSQL Data Stores

Key-Value Stores

- Collection of key-value pairs
- Data access via key: get(key), put(key, value)

Wide Column Stores

- Tables with records with (many) dynamic columns
- Access via key, SQL-like query language, ..

Document Stores

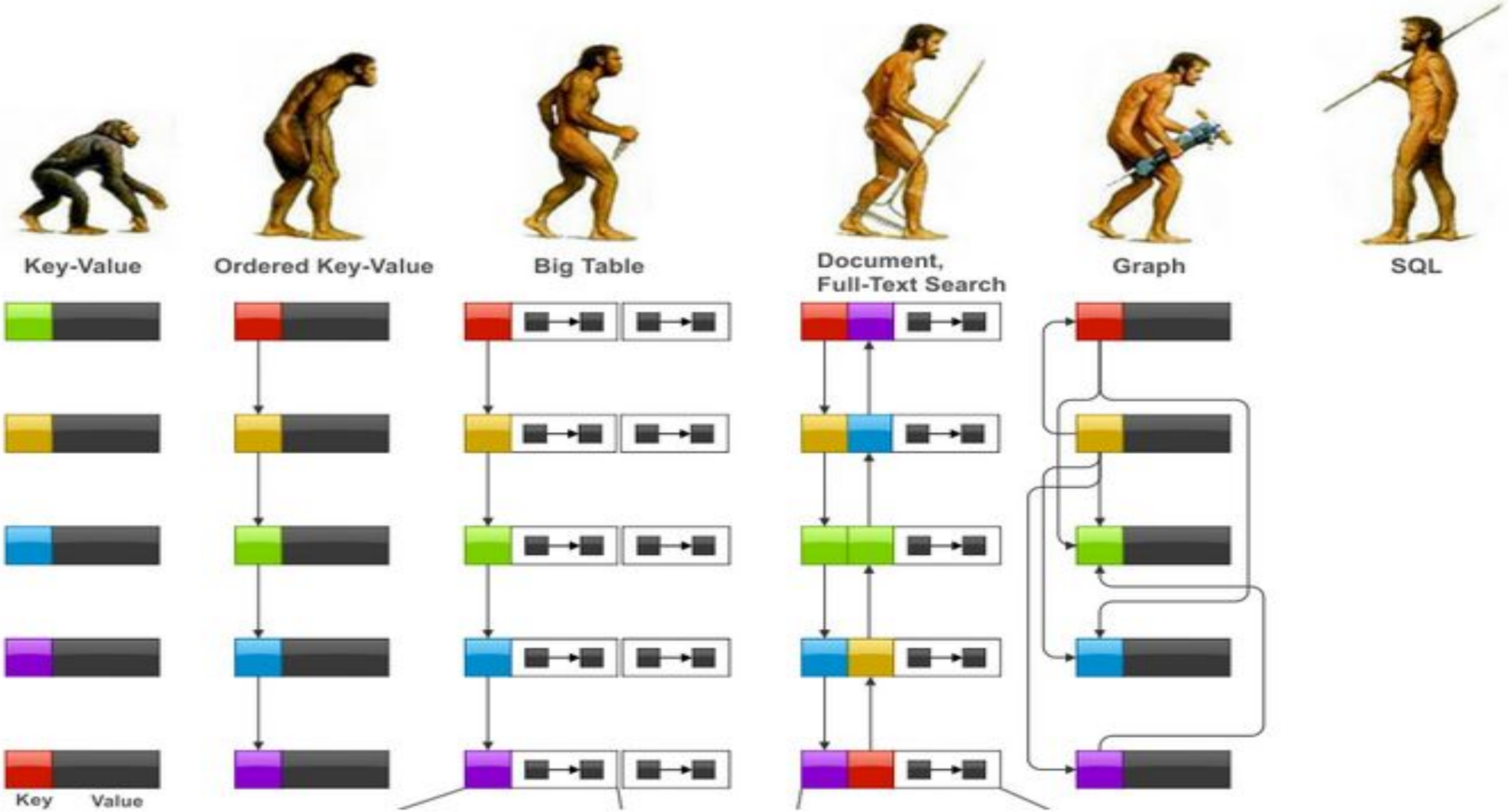
- Semi-structured data in documents (e.g. JSON)
- Access via key or simple API/query language

Graph Databases

- Data as nodes and edges with properties
- Database queries incl. graph algorithms

* multi-model

A multimodel database is a data processing platform that supports multiple data models, which define the parameters for how the information in a database is organized and arranged.


Key-Value
Ordered Key-Value
Big Table
Document,
Full-Text Search
Graph
SQL
Key
Value

Overview of four types

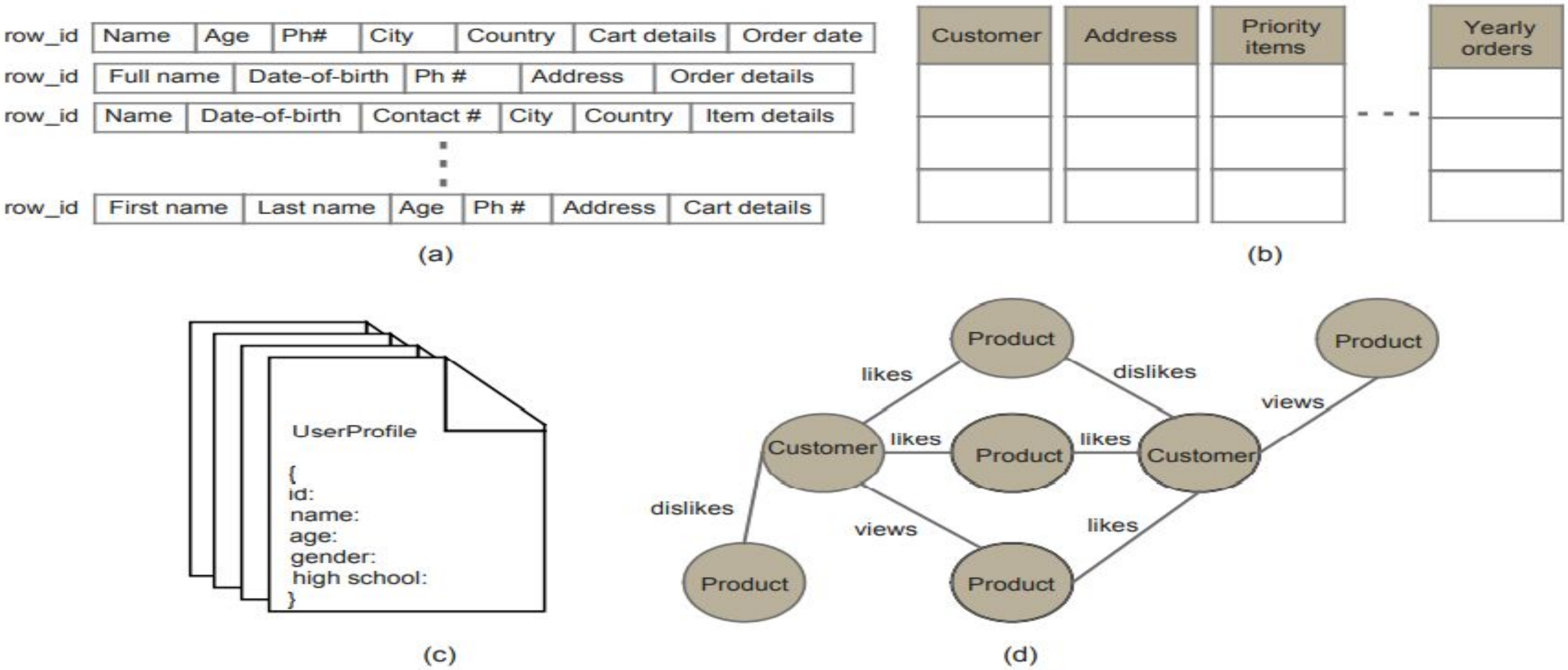
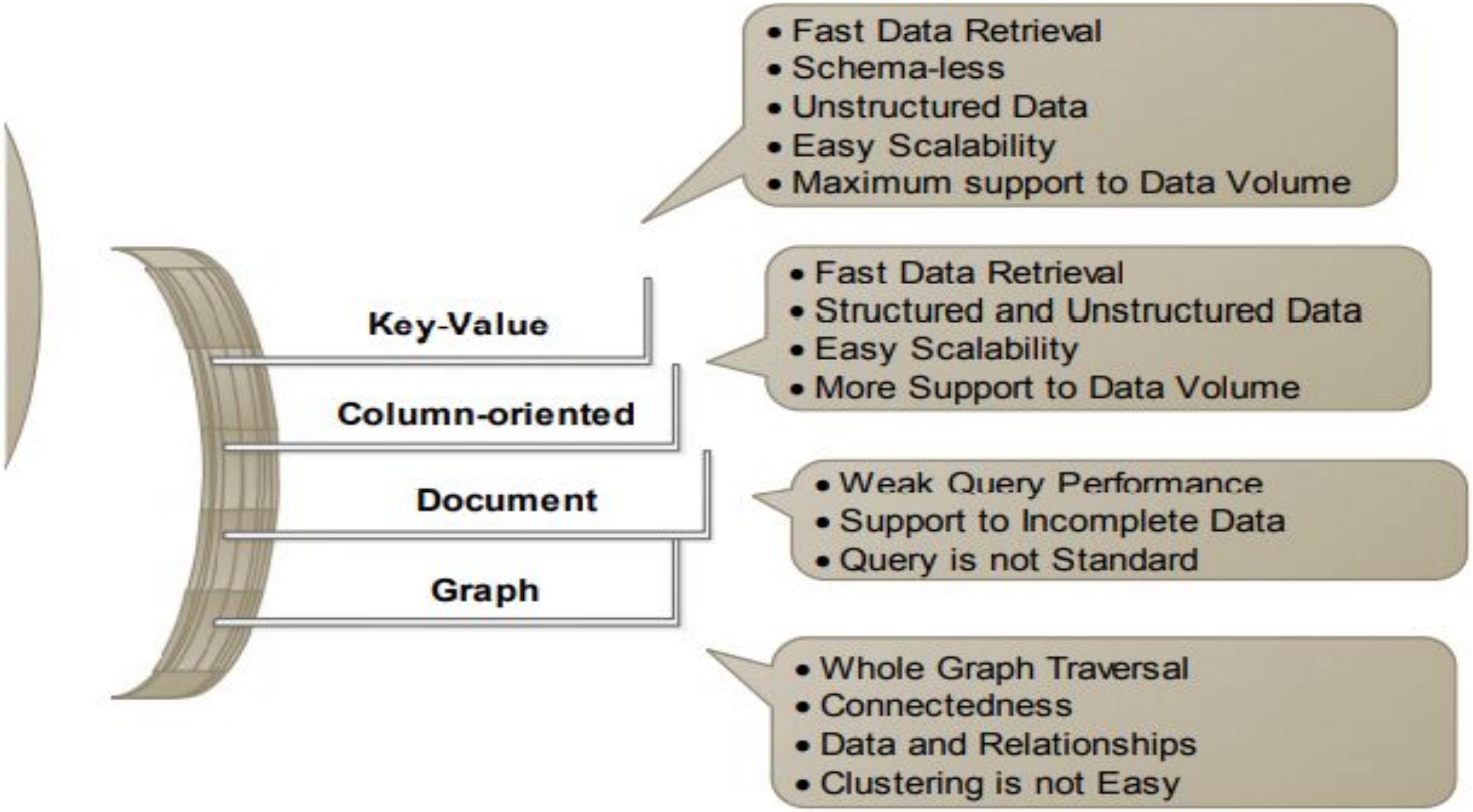


Fig. 4 Examples of data models: (a) key-value; (b) column-oriented; (c) document-oriented; (d) graph



Key-Value

Column-oriented

Document

Graph

- Fast Data Retrieval
- Schema-less
- Unstructured Data
- Easy Scalability
- Maximum support to Data Volume

- Fast Data Retrieval
- Structured and Unstructured Data
- Easy Scalability
- More Support to Data Volume

- Weak Query Performance
- Support to Incomplete Data
- Query is not Standard

- Whole Graph Traversal
- Connectedness
- Data and Relationships
- Clustering is not Easy

Key-Value Databases

- Simpler type of database where each item contains keys and values
- A value can typically only be retrieved by referencing its value
- Learning how to query for a specific key-value pair is typically simple
- Key-value databases are great for use cases where you need to store large amounts of data but you don’t need to perform complex queries to retrieve it.
- Common use cases include storing user preferences or caching.
- Eg. Redis and DynamoDB.

Document Databases

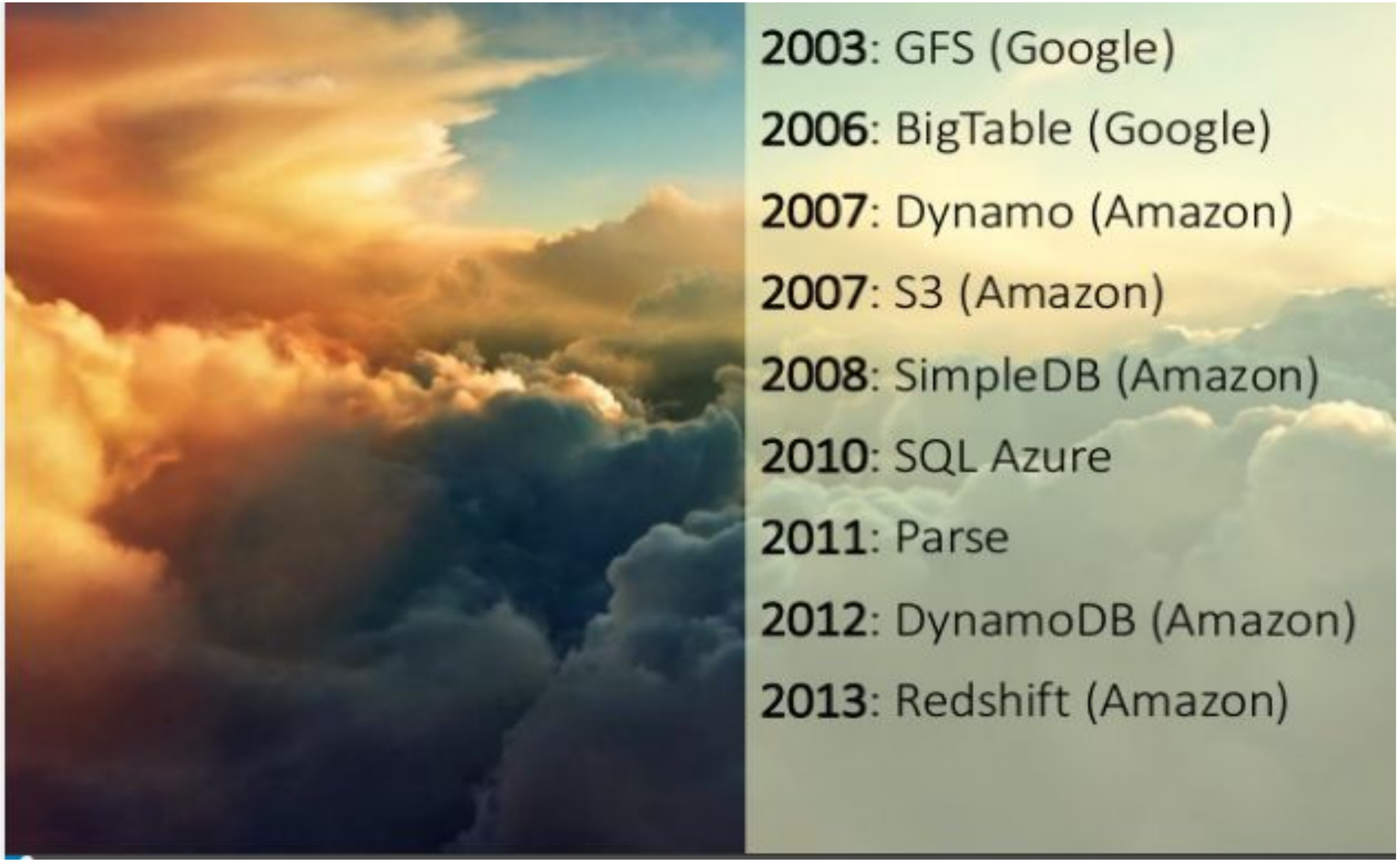
- Store data in documents similar to JSON (JavaScript Object Notation) objects
- Each document contains pairs of fields and values.
- The values can typically be a variety of types including things like strings, numbers, Booleans, arrays, or objects
- Variety of field value types and powerful query languages
- Higher developer productivity, and faster evolution with application needs
- Both natural and flexible for developers to work with
- Great for a wide variety of use cases
- Used as a general purpose database
- Horizontally scale-out to accommodate large data volumes.
- Ex. MongoDB

Wide-Column Stores

- Store data in tables, rows, and dynamic columns
- a lot of flexibility because each row is not required to have the same columns.
- Two-dimensional key-value databases Common
- Great for when you need to store large amounts of data and you can predict what your query patterns will be.
- Commonly used for storing Internet of Things data and user profile data.
- Ex. Cassandra and HBase

Graph Databases

- Store data in nodes and edges.
- Nodes typically store information about people, places, and things
- Edges store information about the relationships between the nodes
- Graph databases excel in use cases where you need to traverse relationships to look for patterns such as social networks, fraud detection, and recommendation engines.
- Ex. Neo4j and JanusGraph



2003: GFS (Google)

2006: BigTable (Google)

2007: Dynamo (Amazon)

2007: S3 (Amazon)

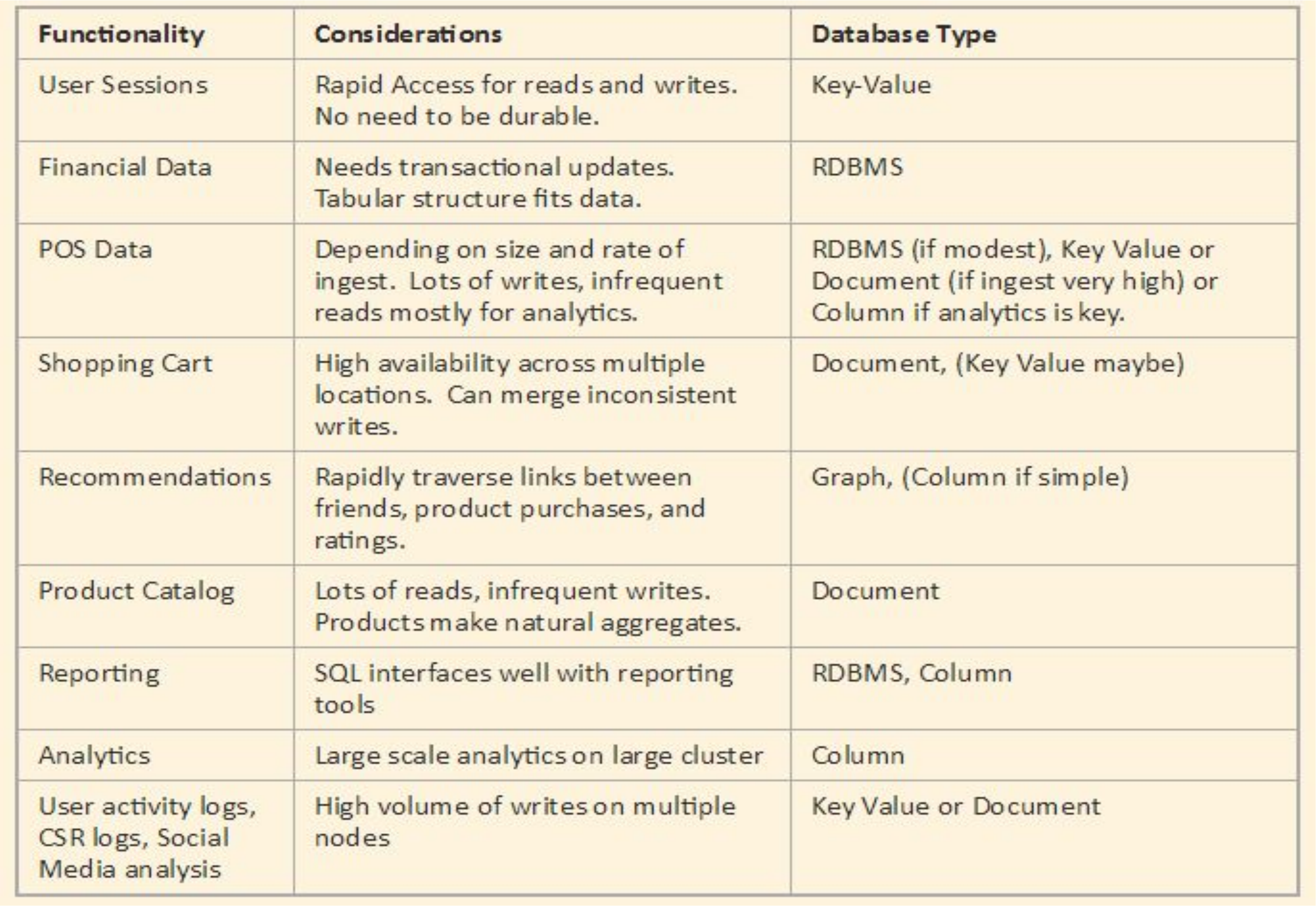
2008: SimpleDB (Amazon)

2010: SQL Azure

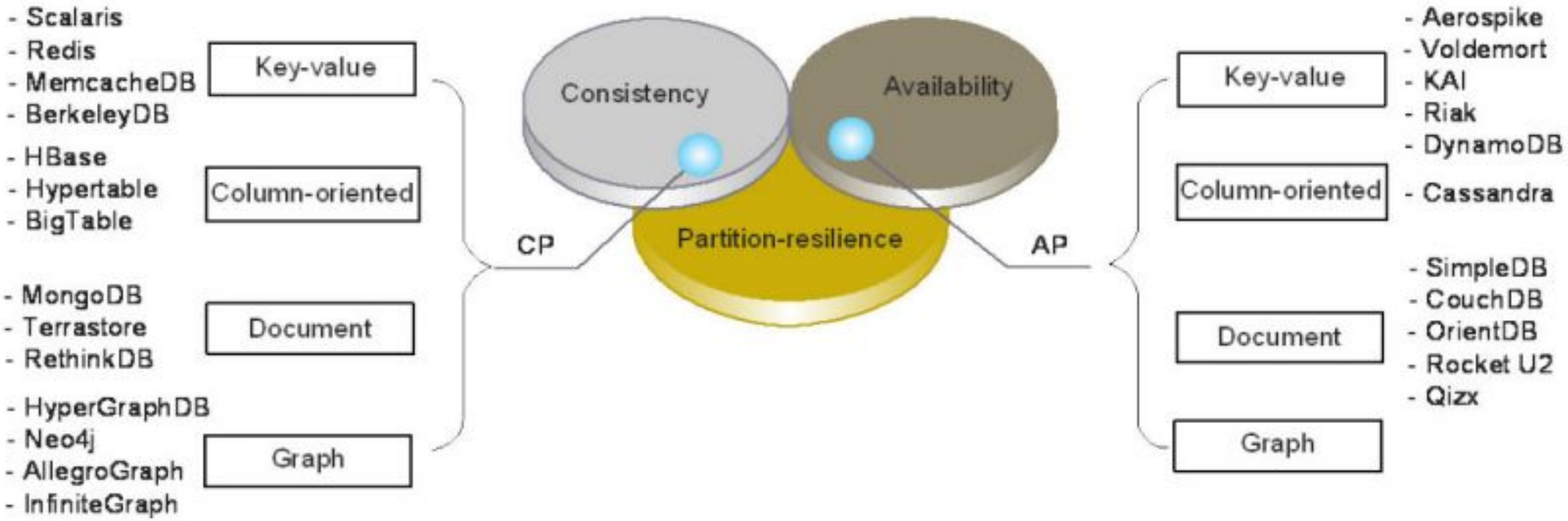
2011: Parse

2012: DynamoDB (Amazon)

2013: Redshift (Amazon)



Functionality	Considerations	Database Type
User Sessions	Rapid Access for reads and writes. No need to be durable.	Key-Value
Financial Data	Needs transactional updates. Tabular structure fits data.	RDBMS
POS Data	Depending on size and rate of ingest. Lots of writes, infrequent reads mostly for analytics.	RDBMS (if modest), Key Value or Document (if ingest very high) or Column if analytics is key.
Shopping Cart	High availability across multiple locations. Can merge inconsistent writes.	Document, (Key Value maybe)
Recommendations	Rapidly traverse links between friends, product purchases, and ratings.	Graph, (Column if simple)
Product Catalog	Lots of reads, infrequent writes. Products make natural aggregates.	Document
Reporting	SQL interfaces well with reporting tools	RDBMS, Column
Analytics	Large scale analytics on large cluster	Column
User activity logs, CSR logs, Social Media analysis	High volume of writes on multiple nodes	Key Value or Document



Consistency

Availability

Partition-resilience

CP

Key-value

- Scalaris
- Redis
- MemcacheDB
- BerkeleyDB

Column-oriented

- HBase
- Hypertable
- BigTable

Document

- MongoDB
- Terrastore
- RethinkDB

Graph

- HyperGraphDB
- Neo4j
- AllegroGraph
- InfiniteGraph

AP

Key-value

- Aerospike
- Voldemort
- KAI
- Riak
- DynamoDB

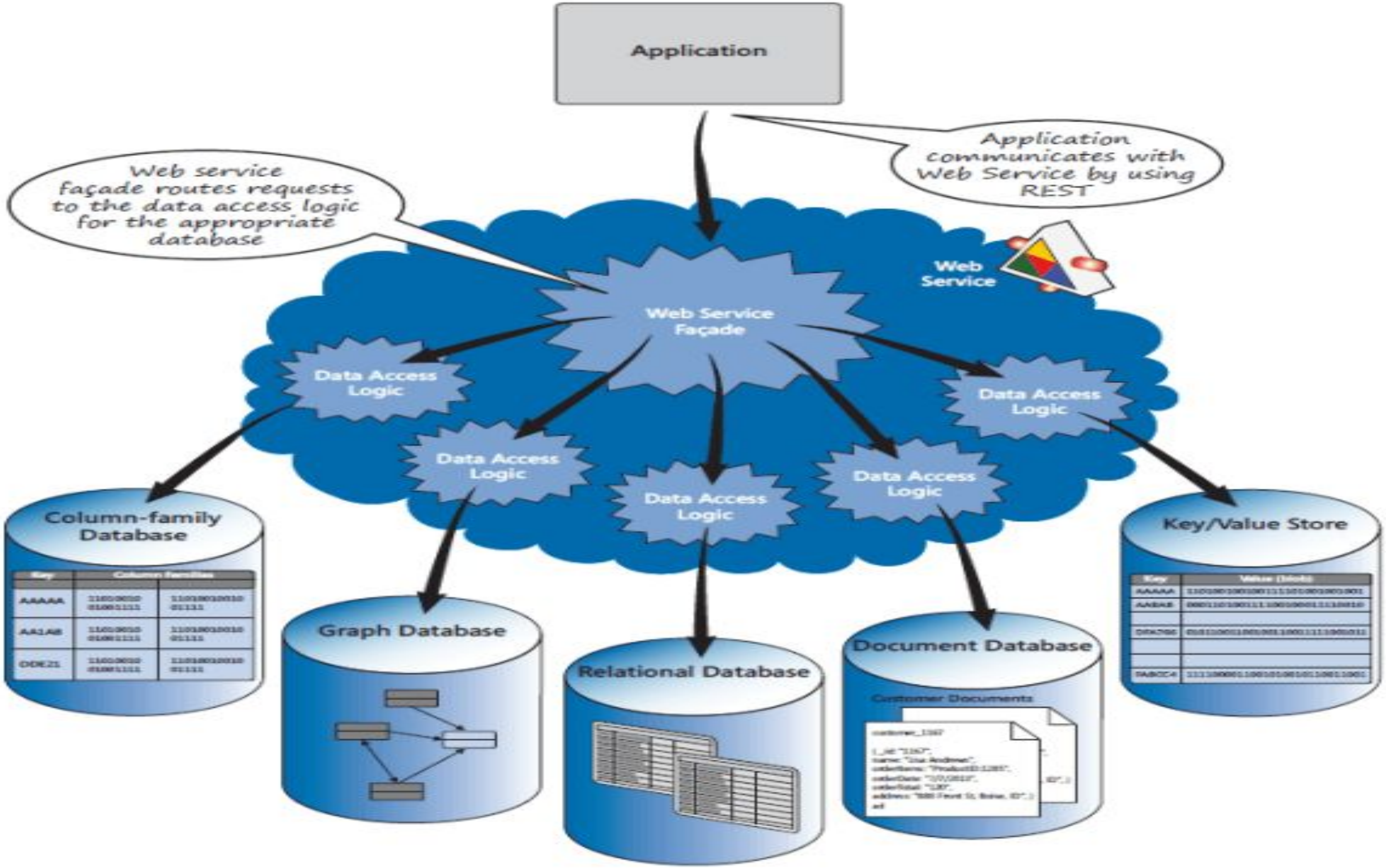
Column-oriented

- Cassandra

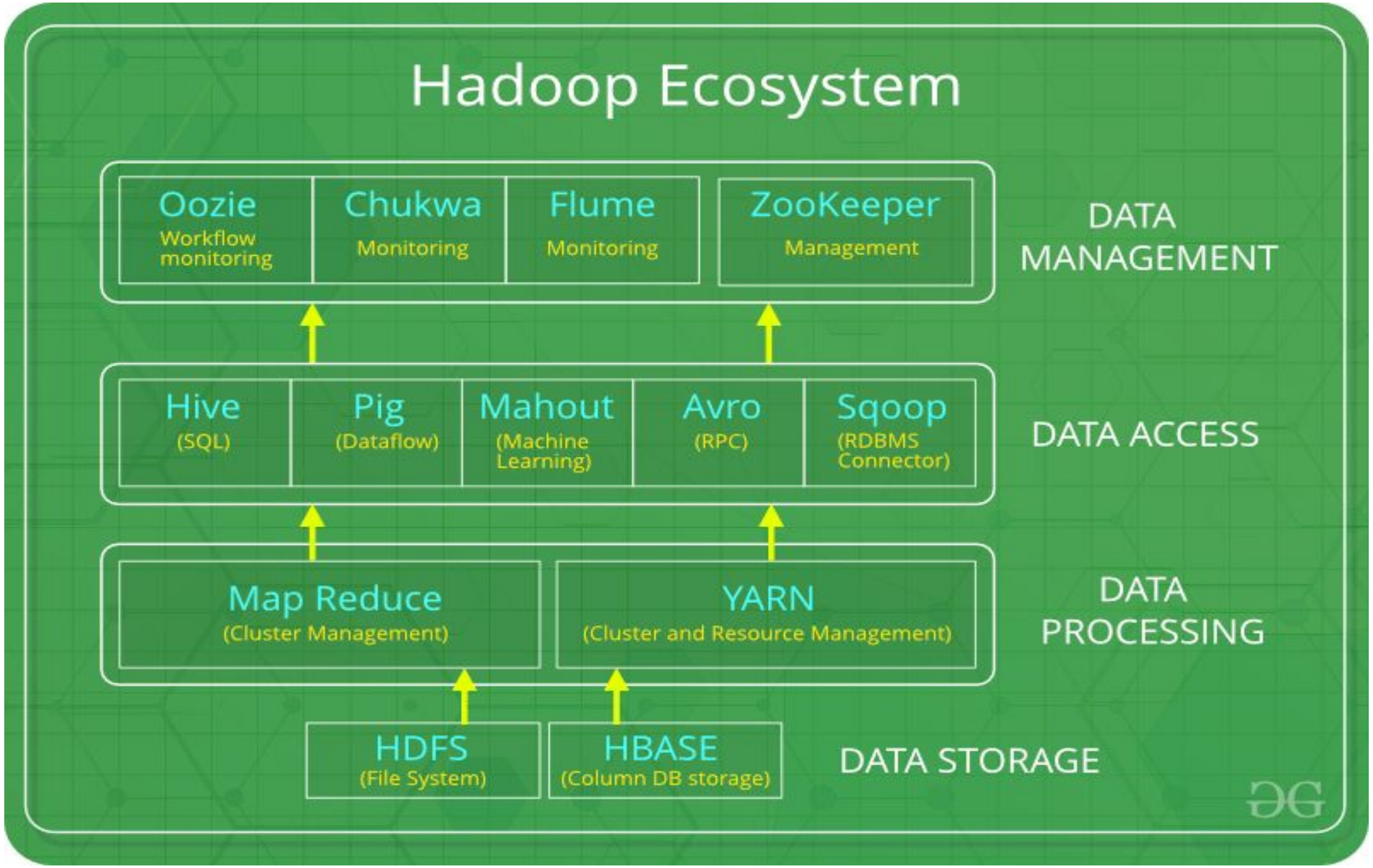
Document

- SimpleDB
- CouchDB
- OrientDB
- Rocket U2
- Qizx

Graph



Application
Application communicates with Web Service by using REST
Web service façade routes requests to the data access logic for the appropriate database
Web Service
Web Service Façade
Data Access Logic
Data Access Logic
Data Access Logic
Data Access Logic
Data Access Logic
Column-family Database
Key
Column Families
AAAAA
AA1AB
DDE21
Graph Database
Relational Database
Document Database
Customer Documents
Key/Value Store
Key
Value (blob)
AAAAA
AABAB
DFA766
FABCC4



Hadoop Ecosystem

DATA MANAGEMENT

Oozie
Workflow monitoring

Chukwa
Monitoring

Flume
Monitoring

ZooKeeper
Management

DATA ACCESS

Hive
(SQL)

Pig
(Dataflow)

Mahout
(Machine Learning)

Avro
(RPC)

Sqoop
(RDBMS Connector)

DATA PROCESSING

Map Reduce
(Cluster Management)

YARN
(Cluster and Resource Management)

DATA STORAGE

HDFS
(File System)

HBASE
(Column DB storage)

Here is a list of the key components in Hadoop:

- **HDFS**: Hadoop Distributed File System
- **HIVE**: Data warehouse that helps in reading, writing, and managing large datasets
- **PIG**: helps create applications that run on Hadoop, allowing to execute jobs in MapReduce
- **MapReduce**: System used for processing large data sets
- **YARN**: Yet Another Resource Negotiator
- **Spark**: Popular analytics engine that works in-memory
- **Oozie**: Open-source workflow scheduling program
- **Zookeeper**: Centralized service for maintaining config info, naming, providing distributed synchronization, and more
- **Mahout**: Helps create ML applications

- https://www.bizety.com/2020/06/20/hadoop-ecosystem-mapreduce-yarn-hive-pig-spark-oozie-zookeeper-mahout-and-kube2hadoop/
- https://www.geeksforgeeks.org/hadoop-ecosystem/
- https://datascienceguide.github.io/opensource-bigdata-tools

- https://www.mongodb.com/resources/products/compatibilities/hadoop-and-mongodb https://www.mongodb.com/resources/basics/databases/nosql-explained