Universidad Católica San Pablo | Department of Computer Science

Data Storage

Erick Gomez Nieto, PhD

emgomez@ucsp.edu.pe

Data Storage

Big data storage technologies are referred to as storage technologies that in some way specifically address the volume, velocity, or variety challenge and do not fall in the category of relational database systems. This does not mean that relational database systems do not address these challenges, but alternative storage technologies such as columnar stores and clever combinations of different storage systems, e.g. using the Hadoop Distributed File System (HDFS), are often more efficient and less expensive (Marz and Warren 2014).

Marz, N., & Warren, J. (2014). A new paradigm for Big Data. In N. Marz & J. Warren (Eds.), Big Data: Principles and best practices of scalable real-time data systems. Shelter Island, NY: Manning Publications.

Main challenges for Data Storage

1. Infrastructure: If you plan on storing vast amounts of data, you'll need the infrastructure necessary to store it, which often means investing in high-tech servers that will occupy significant space in your office or building. One of the easiest workarounds is to use cloud hosting and cloud storage.
2. Cost: Again, the best solution here is to outsource the work; you'll probably have to pay a monthly fee, but it will save you money in the long run.
3. Security: There are many layers of security that can help you prevent this unauthorized access, including encryption and reliance on third-party providers, but there's a limit to how well these can protect you

Extracted from: https://www.smartdatacollective.com/7-biggest-problems-data-storage-overcome/

Main challenges for Data Storage

4. Corruption: Practically every form of data storage has the potential to be corrupted. Stray particles can interfere with most forms of data storage, and anything relying on magnetic strips or electric storage can be corrupted by electromagnetic interference. Your best bet for protection here is utilizing multiple backups.

5. Scale: Your data storage solution needs some capacity to scale. Here, it pays to give yourself as many options as possible, since you won’t be sure exactly how your needs will change in the future.

Extracted from: https://www.smartdatacollective.com/7-biggest-problems-data-storage-overcome/

Main challenges for Data Storage

6. UI and Accessibility: you’ll need some kind of system with an intuitive, accessible user interface (UI), and clean accessibility for whatever functionality you want.

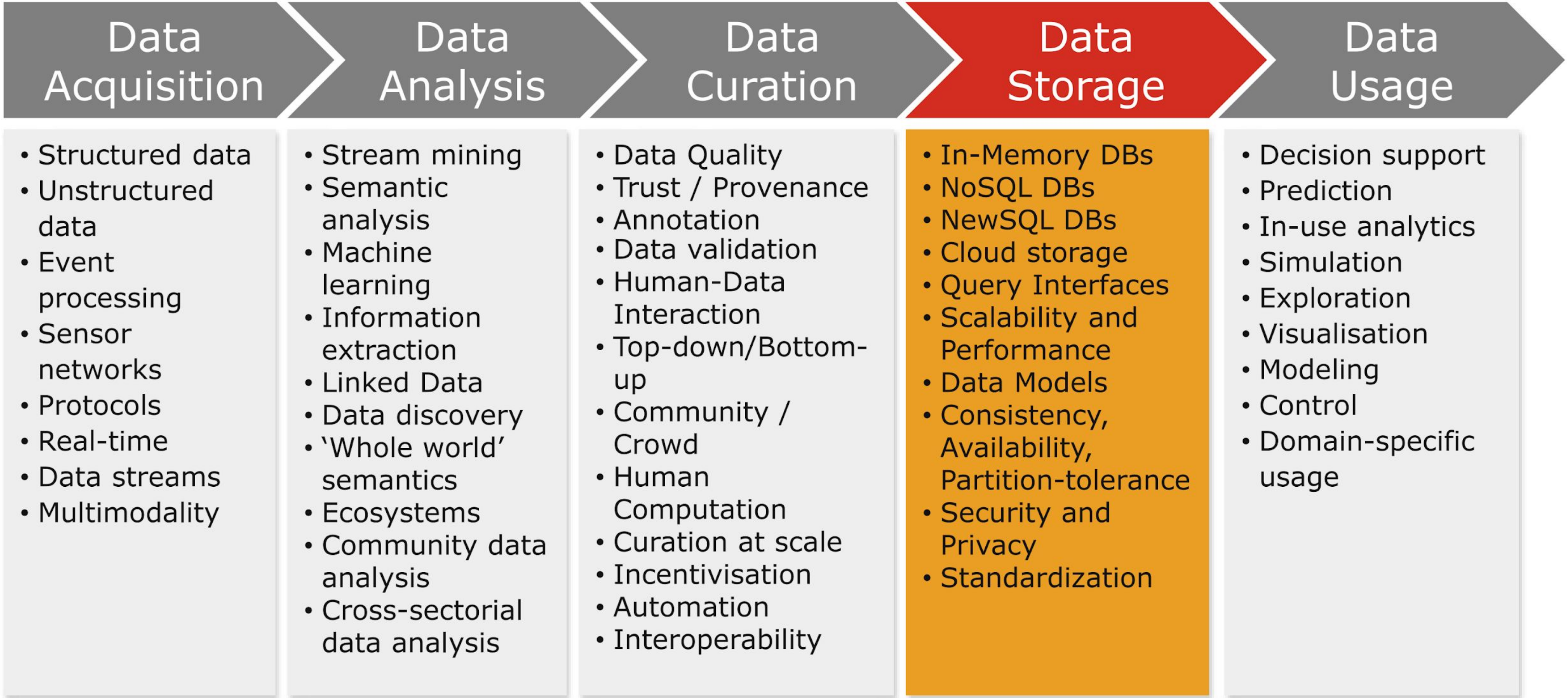
7. Compatibility. If you plan on using multiple systems or applications with your data, you’ll need to ensure they’re compatible. For that, you’ll need to find a data storage partner with an open API and a clean system of transition.

8. Volume

9. Predictability

Extracted from: https://www.smartdatacollective.com/7-biggest-problems-data-storage-overcome/

Big Data Value Chain



Data Acquisition

- Structured data
- Unstructured data
- Event processing
- Sensor networks
- Protocols
- Real-time
- Data streams
- Multimodality

Data Analysis

- Stream mining
- Semantic analysis
- Machine learning
- Information extraction
- Linked Data
- Data discovery
- 'Whole world' semantics
- Ecosystems
- Community data analysis
- Cross-sectorial data analysis

Data Curation

- Data Quality
- Trust / Provenance
- Annotation
- Data validation
- Human-Data Interaction
- Top-down/Bottom-up
- Community / Crowd
- Human Computation
- Curation at scale
- Incentivisation
- Automation
- Interoperability

Data Storage

- In-Memory DBs
- NoSQL DBs
- NewSQL DBs
- Cloud storage
- Query Interfaces
- Scalability and Performance
- Data Models
- Consistency, Availability, Partition-tolerance
- Security and Privacy
- Standardization

Data Usage

- Decision support
- Prediction
- In-use analytics
- Simulation
- Exploration
- Visualisation
- Modeling
- Control
- Domain-specific usage

Data Storage Technologies

- Traditional DataBase Management Systems (DBMS)
- Distributed File Systems
- NoSQL databases
- NewSQL Databases
- Graph-based databases
- Big Data Query Platforms
- Cloud Storage

DB vs. DBMS

- DB
  - Is a collection of information organized to provide efficient retrieval.
  - Physical or Digital
  - E.g. phone book, address book

- DBMS
  - Computer software
  - "is system software for creating and managing databases. The **DBMS** provides users and programmers with a systematic way to create, retrieve, update and manage data" (searchsqlserver.techtarget.com)

mongoDB MySQL PostgreSQL IBM DB2 ORACLE DATABASE MariaDB

Relational Databases

A relational database is a type of database that stores and provides access to data points that are related to one another. Relational databases are based on the relational model, an intuitive, straightforward way of representing data in tables. In a relational database, each row in the table is a record with a unique ID called the key. The columns of the table hold attributes of the data, and each record usually has a value for each attribute, making it easy to establish the relationships among data points.

ORACLE Extracted from https://www.oracle.com/database/what-is-a-relational-database/

Relational Databases | Benefits

The simple yet powerful relational model is used by organizations of all types and sizes for a broad variety of information needs. Relational databases are used to track inventories, process ecommerce transactions, manage huge amounts of mission-critical customer information, and much more. A relational database can be considered for any information need in which data points relate to each other and must be managed in a secure, rules-based, consistent way.

Relational databases have been around since the 1970s. Today, the advantages of the relational model continue to make it the most widely accepted model for databases.

ORACLE Extracted from https://www.oracle.com/database/what-is-a-relational-database/

Relational Databases | Limitations

- Few relational databases have limits on field lengths which can't be exceeded.
- Relational databases can sometimes become complex as the amount of data grows, and the relations between pieces of data become more complicated.
- Complex relational database systems may lead to isolated databases where the information cannot be shared from one system to another.

ORACLE Extracted from https://www.oracle.com/database/what-is-a-relational-database/

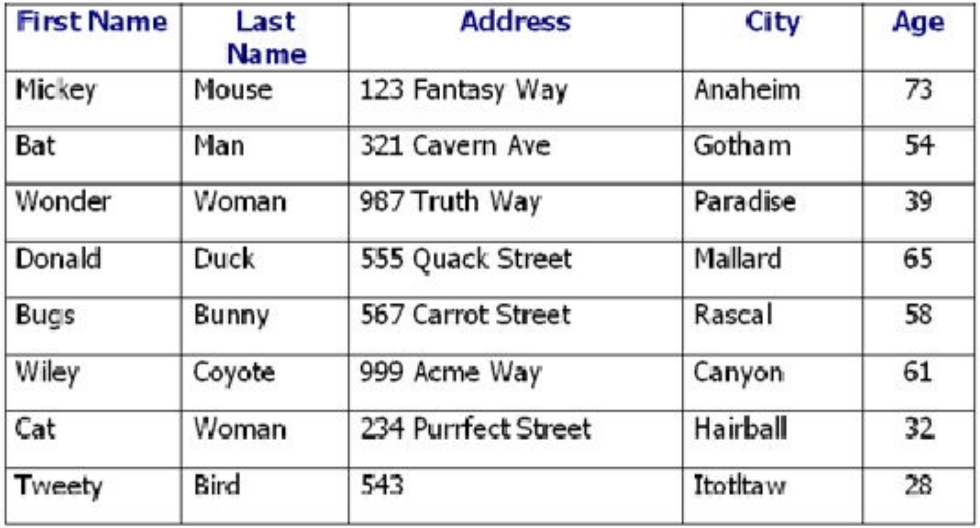
Relational Model

Document Model

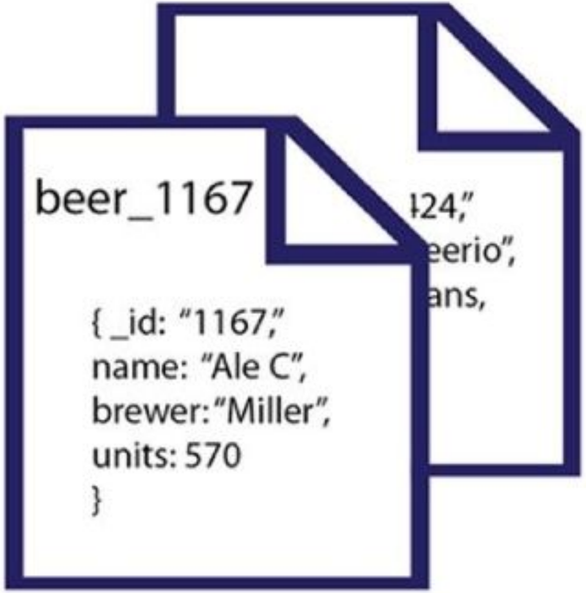
```
{
  "_id" : "rp-prod132546",
  "name" : "Marvel T2 Athena”,
  "brand" : "Pinarello",
  "category" : "bike",
  "type" : "Road Bike”,
  "price" : 2949.99,



  "size" : "55cm",
  "wheel_size" : "700c",
  "frameset" : {
      "frame" : "Carbon Toryaca",
      "fork" : "Onda 2V C"
  },
  "groupset" : {
      "chainset" : "Camp. Athena 50/34",
      "brake" : "Camp."
  },
  "wheelset" : {
      "wheels" : "Camp. Zonda",
      "tyres" : "Vittoria Pro"
  }
}
```

select * from users;

db.users.find()



First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28



beer_1167

```
{ _id: "1167,"
name: "Ale C",
brewer:"Miller",
units: 570
}
```

Distributed File Systems

File systems such as the Hadoop File System (HDFS) (Shvachko et al. 2010) offer the capability to store large amounts of unstructured data in a reliable way on commodity hardware. Although there are file systems with better performance, HDFS is an integral part of the Hadoop framework (White 2012) and has already reached the level of a de-facto standard. It has been designed for large data files and is well suited for quickly ingesting data and bulk processing.

* Shvachko, K. H. K., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In IEEE 26th Symposium on Mass Storage Systems and Technologies (pp. 1–10).
* White, T. (2012). Hadoop: The Definitive Guide. O'Reilly.

NoSQL Databases

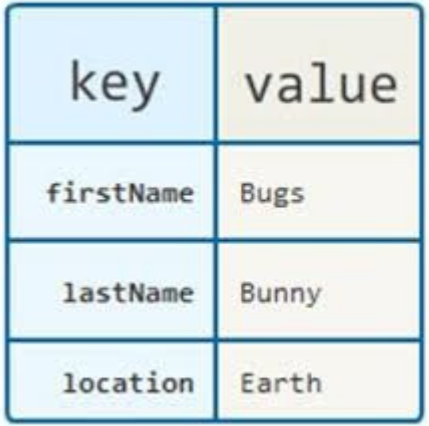
NoSQL databases use data models from outside the relational world that do not necessarily adhere to the transactional properties of atomicity, consistency, isolation, and durability (ACID).

NoSQL databases are designed for scalability, often by sacrificing consistency. Compared to relational databases, they often use low-level, non-standardized query interfaces, which make them more difficult to integrate in existing applications that expect an SQL interface.

Erick Gomez-Nieto, PhD. What is Data?. Data Science Program.

NoSQL Databases | Models I

Key-Value Stores: Key-value stores allow storage of data in a schema-less way. Data objects can be completely unstructured or structured and are accessed by a single key. As no schema is used, it is not even necessary that data objects share the same structure.



key	value
firstName	Bugs
lastName	Bunny
location	Earth

amazon DynamoDB

ORACLE NOSQL DATABASE

redis

AEROSPIKE

NoSQL Databases | Models II

Columnar Stores: Such databases are typically sparse, distributed, and persistent multi-dimensional sorted maps in which data is indexed by a triple of a row key, column key, and a timestamp. The value is represented as an uninterrupted string data type. Data is accessed by column families, i.e. a set of related column keys that effectively compress the sparse data in the columns. Column families are created before data can be stored and their number is expected to be small. In contrast, the number of columns is unlimited. In principle columnar stores are less suitable when all columns need to be accessed. However in practice this is rarely the case, leading to superior performance of columnar stores.

cassandra

APACHE HBASE

NoSQL Databases | Models III

Document Databases: In contrast to the values in a key-value store, documents are structured. However, there is no requirement for a common schema that all documents must adhere to as in the case for records in relational databases. Thus document databases are referred to as storing semi-structured data. The capability of the query interface is typically dependent on the encoding format used by the databases. Common encodings include XML or JSON.

NoSQL Databases | Models IV

Graph databases, such as Neo4J (2015), store data in graph structures making them suitable for storing highly associative data such as social network graphs. A particular flavour of graph databases are triple stores such as AllegroGraph (Franz 2015) and Virtuoso (Erling 2009) that are specifically

Neo4j. (2015). Neo4j Company Website. http://neo4j.com/. Accessed Feb 6, 2015.

NewSQL Databases

A modern form of relational databases that aim for comparable scalability as NoSQL databases while maintaining the transactional guarantees made by traditional database systems.

They solve some of the major problems associated with traditional online transaction processing (OLTP).

They maintain the (Atomicity, Consistency, Isolation and Durability) ACID guarantees of traditional DBMS.

Cockroach DB

Vitess

TiDB

FAUNA

NewSQL Databases | Features

Partitioning/Sharding: almost all NewSQL database management systems scale out by dividing the database into separate subsets known as partitions or shards. The tables are horizontally split into several fragment with boundaries based on column values. Related fragments from different tables are joined to create a partition.

Replication: This feature allows database users to create and maintain copies of a database or a part of a database. Copies of the database are stored in a remote site next to the main site or in a distant site.

Secondary Indexes: Secondary indexes allow database users to efficiently access database records by using a different value other than the primary key.

https://www.predictiveanalyticstoday.com/newsql-databases/

Concurrency Control: This feature addresses the problem that might occur in a multi-user system when many user access or modify data simultaneously. NewSQL systems use this feature to ensure simultaneous transaction while maintaining data integrity.

Crash Recovery: NewSQL databases have a mechanism that enables them to recover data and move to a consistent state when the system crashes.

https://www.predictiveanalyticstoday.com/newsql-databases/

Some of the benefits include:

- Database partitioning reduces the system’s communication overhead making it possible to access data with ease.
- ACID transactions ensure data integrity even if there is a system failure or error.
- NewSQL databases can handle complex data.
- NewSQL systems are highly scalable.

https://www.predictiveanalyticstoday.com/newsql-databases/

Cloud Storage

Cloud storage is a cloud computing model that stores data on the Internet through a cloud computing provider who manages and operates data storage as a service. It’s delivered on demand with just-in-time capacity and costs, and eliminates buying and managing your own data storage infrastructure. This gives you agility, global scale and durability, with “anytime, anywhere” data access.

https://aws.amazon.com/what-is-cloud-storage/?nc1=h_ls

Cloud storage providers

Amazon S3
Azure
DAV
WebDAV
Google Cloud Platform Live
Dropbox
Google Drive
Yandex Disk
amazondrive
openstack.
OneDrive for Business
hubiC
BACKBLAZE

Cloud Storage | **Types I**

There are three types of cloud data storage: object storage, file storage, and block storage. Each offers their own advantages and have their own use cases:

Object Storage - Applications developed in the cloud often take advantage of object storage's vast scalablity and metadata characteristics. Object storage solutions like Amazon Simple Storage Service (S3) are ideal for building modern applications from scratch that require scale and flexibility, and can also be used to import existing data stores for analytics, backup, or archive.

File Storage - Some applications need to access shared files and require a file system. This type of storage is often supported with a Network Attached Storage (NAS) server. File storage solutions like Amazon Elastic File System (EFS) are ideal for use cases like large content repositories, development environments, media stores, or user home directories.

Block Storage - Other enterprise applications like databases or ERP systems often require dedicated, low latency storage for each host. This is analagous to direct-attached storage (DAS) or a Storage Area Network (SAN). Block-based cloud storage solutions like Amazon Elastic Block Store (EBS) are provisioned with each virtual server and offer the ultra low latency required for high performance workloads.

Cloud Storage | Benefits

Total Cost of Ownership

With cloud storage, there is no hardware to purchase, storage to provision, or capital being used for "someday" scenarios. You can add or remove capacity on demand, quickly change performance and retention characteristics, and only pay for storage that you actually use. Less frequently accessed data can even be automatically moved to lower cost tiers in accordance with auditable rules, driving economies of scale.

Time to Deployment. When development teams are ready to execute, infrastructure should never slow them down. Cloud storage allows IT to quickly deliver the exact amount of storage needed, right when it's needed. This allows IT to focus on solving complex application problems instead of having to manage storage systems.

Information Management. Centralizing storage in the cloud creates a tremendous leverage point for new use cases. By using cloud storage lifecycle management policies, you can perform powerful information management tasks including automated tiering or locking down data in support of compliance requirements.

Erick Gomez-Nieto, PhD. What is Data?. Data Science Program.

Cloud Storage | Limitations

Internet Connection

Cloud based storage is dependent on having an internet connection. If you are on a slow network you may have issues accessing your storage. In the event you find yourself somewhere without internet, you won't be able to access your files.

Costs

There are additional costs for uploading and downloading files from the cloud. These can quickly add up if you are trying to access lots of files often.

Hard Drives

Cloud storage is supposed to eliminate our dependency on hard drives right? Well some business cloud storage providers require physical hard drives as well.

Support

Support for cloud storage isn't the best, especially if you are using a free version of a cloud provider. Many providers refer you to a knowledge base or FAQs.

Privacy

When you use a cloud provider, your data is no longer on your physical storage. So who is responsible for making sure that data is secure? That's a gray area that is still being figured out.

References

Strohbach, Martin & Daubert, Jörg & Ravkin, Herman & Lischka, Mario. (2016). Big Data Storage. 10.1007/978-3-319-21569-3_7.