Week 1: Data Storage

Categories of Data Management Tools

https://www.futurelearn.com/courses/data-management-and-visualisation-data-storage-and-privacy/3/steps/1600262

Databases and data storage options allow you to Find, Select, Appky analysis functions, Calculate reports as data is updated and ensure consistency and protection of data.

You can consider the following characteristics to be found in most data storage methods

1. **Structure** - tables, documents, “chunks”
2. **Minimise redundancy** - efficient storage, normalisation
3. **Maintain consistency** - updates, transactions, deletions
4. **Multiple user and concurrent access**
5. **Query options** - eg. language like SQL, SPARQL, Cypher
6. **Security**

five general categories of data storage:

1. Relational databases (traditional & modern)

- not suitable for really big data, especially if query results are greater than available working memory.
- Modern relational DBs and individual implementations often have considerable extra functionality that enables them to support big data or unstructured, non-relational data.

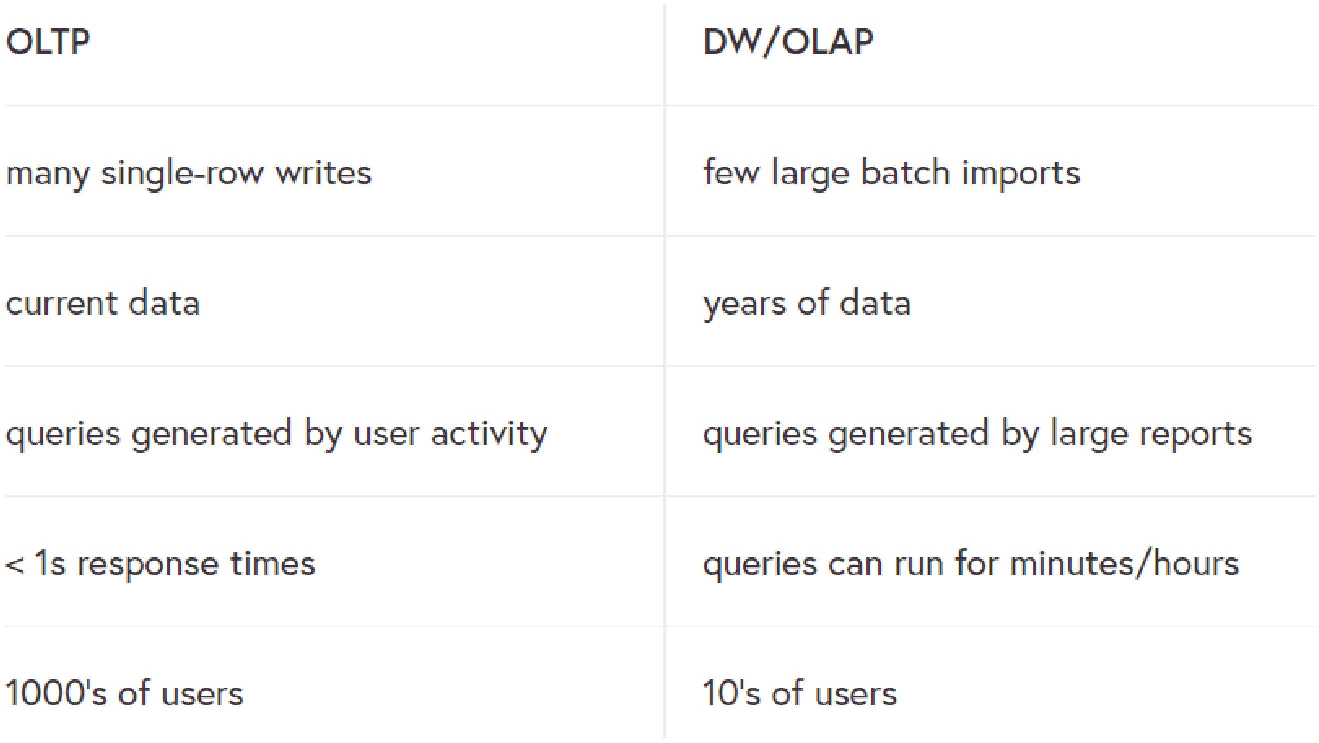
2. Column databases

- rather than indexing rows like a traditional relational database, index by column
    - key benefit of using col:
        - **Compression**. Column stores are very efficient at data compression and/or partitioning.
        - **Aggregation queries**. Due to their structure, columnar databases perform particularly well with aggregation queries (such as SUM, COUNT, AVG, etc).
        - **Scalability**. Columnar databases are very scalable. They are well suited to massively parallel processing (MPP), which involves having data spread across a large cluster of machines – often thousands of machines.
        - **Fast to load and query**. Columnar stores can be loaded extremely fast. A billion-row table could be loaded within a few seconds. You can start querying and analysing almost immediately.
    - key drawbacks:
        - generally considered as the best option for data mining (see the next article) and can be slow to add, update or purge data though modern implementations are often optimised for streaming data.

3. MPP (Massively Parallel Processing Databases) or Data Warehouses

- Data Warehouses are often used for business intelligence analytics particularly where there are large volumes of historical data.

- Two terms you'll often see used with DW are **OLAP** and **OLTP**.



OLTP	DW/OLAP
many single-row writes	few large batch imports
current data	years of data
queries generated by user activity	queries generated by large reports
< 1s response times	queries can run for minutes/hours
1000's of users	10's of users

- Generally, OLTP can support big data for many concurrent requests for small amounts of data each time while DW/OLAP can support big data for low concurrent requests for very large amounts of data each time.

4. No SQL data storage methods

- loose term for data storage methods that don't fit into a traditional relational database model
- open source, non-relational
- Schema-Less - so adaptable to changes in data structure
- Horizontally Scalable – to increase capacity you can add new separate servers rather than increasing the power of the existing server
- Lack of Adherence to ACID Principles – **ACID principles** (atomicity, consistency, isolation, and durability) are the accepted properties of any transaction run on a DBMS

- Can be sub-classified into four types though many implementations offer a blend or mix of services:

    1. Key-Value
        - A key-value store is the simplest data model
        - like a Python Dictionary
    2. Document store
        - storing semi-structured document object data and metadata (JSON)
        - can query for documents by their content or metadata but you don't need to adhere to a strict scheme
        - eg: MongoDB
    3. Column store (see Step 1.6)
    4. Graph (see course 1, Step 3.10)

5. Big data methods (e.g., MapReduce, Hadoop/HDFS)

Acronyms: .... SPARQL: SPARQL Protocol and RDF Query Language (we looked at this in course 1, topic 3)

wk1 quiz: https://www.futurelearn.com/courses/data-management-and-visualisation-data-storage-and-privacy/3/steps/1600264/quiz/introduction

---

Week 2: Big Data Storage

- 3 Vs of big data – Volume, Velocity and Variety
- grouped data storage methods into five classes:

1. Relational
2. Column
3. Data Warehouses and MPP
4. NoSQL
5. Big Data

> Modern relational databases, data warehouses, columnar databases and NoSQL (Key-Value, Document) can manage **very large volumes of data**.
>
> The schema-less options of NoSQL can **adapt well to the variety** of big data
>
> the efficient implementations of key-value and columnar databases are adept at ingesting **high-velocity, streaming data**.

However, there are other paradigms for processing, indexing, analysing and storing big data......

Map/Reduce