

Data Management Open Source Tools

This reading provides a summary of key open-source Data Management Tools

They are broadly classified as -

Data Management Tools - Facilitates the storage, organization, and retrieval of data. Includes Relational Databases, NoSQL Databases, and Big Data platforms.

Data Management Tools:

1. MySQL

- Popular open source relational database management system (RDBMS)
- Uses structured query language (SQL) to manage and store data.
- Common uses:
 - Web applications
 - Data warehousing
 - E-commerce

2. PostgreSQL

- Powerful and open source relational database management system (RDBMS)
- Emphasizes extensibility and SQL compliance.
- Offers advanced features such as:
 - Support for JSON
 - Full-text search
 - Spatial data

3. Apache CouchDB

- Document-oriented NoSQL database
- Uses JSON to store data
- Highly scalable
- Fault-tolerant
- Easy to use

4. MongoDB

- Document-oriented NoSQL database
- Stores data in a flexible JSON
- Provides:

- Scalability
- High availability
- Data distribution
- Suitable for modern web applications that handle large volumes of unstructured data

5. Apache Cassandra

- Highly scalable, distributed Document-oriented NoSQL database
- Can handle large amounts of structured and unstructured data across many commodity servers.
- Offers:
 - High availability
 - Fault tolerance
 - Tunable consistency levels
 - Suitable for mission-critical applications

6. Hadoop Distributed File System (HDFS)

- Designed to work with large datasets like Apache Hadoop in a distributed computing environment
- High-throughput data processing by splitting files into blocks (default 128MB), and these blocks are distributed across multiple DataNodes
- Data is replicated across different DataNodes ensuring high availability and fault tolerance
- Scalable and efficient

7. Ceph

- Free, open source software-defined storage platform suitable for hybrid cloud environments
- Designed for modern data centers
- Provides highly scalable, unified storage system that can be used for object storage (like AWS S3), block storage (like virtual disks for VMs), and file storage (like NFS) under one unified system
- High performance, availability and reliability

8. Elasticsearch

- Primarily a distributed RESTful search engine and analytics tool
- Based on the Lucene library.
- Full-text search, real-time data analytics
- Highly scalable
- Easy to use
- Powerful querying capabilities
- Real-time data indexing for fast document retrieval.

