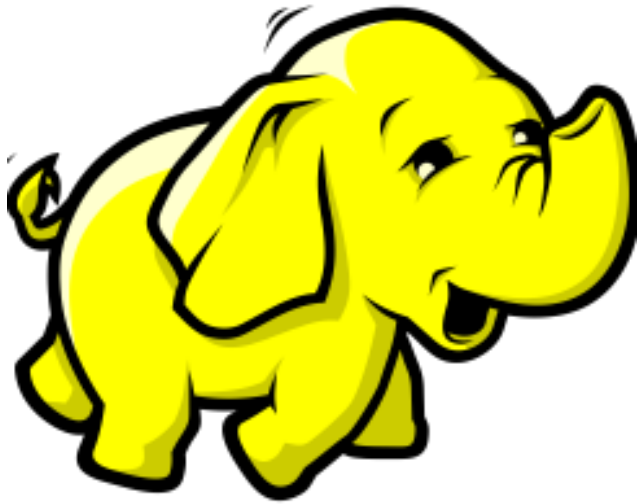
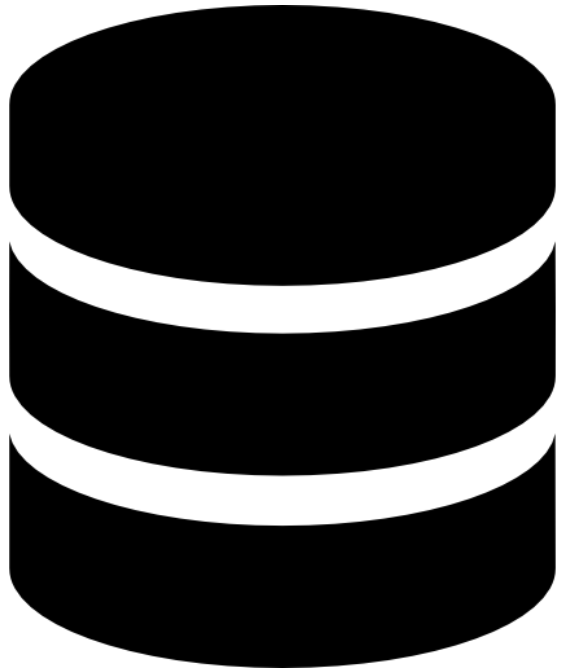




HADOOP VS. TRADITIONAL DATA STORAGE AND PROCESSING

WHAT IS HADOOP?

HADOOP IS FUNDAMENTALLY AN OPEN-SOURCE INFRASTRUCTURE SOFTWARE FRAMEWORK THAT ALLOWS DISTRIBUTED STORAGE AND PROCESSING A HUGE AMOUNT OF DATA I.E. BIG DATA.

IT'S A CLUSTER SYSTEM WHICH WORKS AS A MASTER-SLAVE ARCHITECTURE.

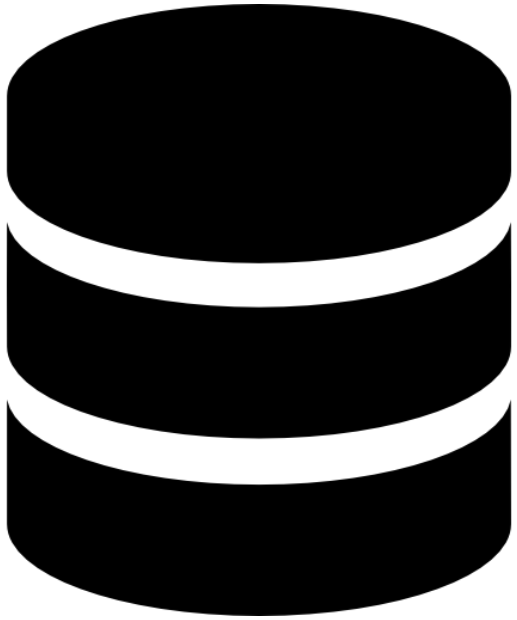
HENCE, WITH SUCH ARCHITECTURE, LARGE DATA CAN BE STORED AND PROCESSED IN PARALLEL.

DIFFERENT TYPES OF DATA CAN BE ANALYZED, STRUCTURED(TABLES), UNSTRUCTURED (LOGS, EMAIL BODY, BLOG TEXT) AND SEMI-STRUCTURED (MEDIA FILE METADATA, XML, HTML).

WHAT IS RDBMS?

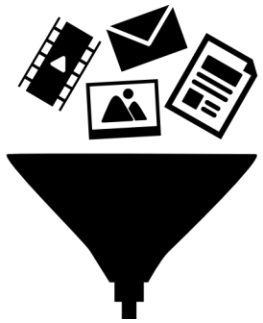
RDBMS stands for the relational database management system. It is a database system based on the relational model specified by Edgar F. Codd in 1970. The database management software like Oracle server, My SQL, and IBM DB2 are based on the relational database management system.

The data represented in the RDBMS is in the form of the rows or the tuples. This table is basically a collection of related data objects and it consists of columns and rows. Normalization plays a crucial role in RDBMS. It contains the group of the tables, each table contains the primary key.



#1 DATA VARIETY

HDOOP



Used for Structured, Semi-Structured and Unstructured data

RDBMS

Mainly for Structured data

#2 DATA STORAGE

HDOOP

RDBMS

Use for large data set
(Tbs and Pbs)

Average size data
(GBS)

#3 QUERYING

HDOOP

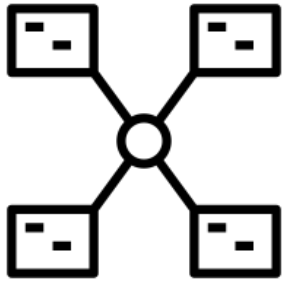
RDBMS

HQL (Hive Query Language)

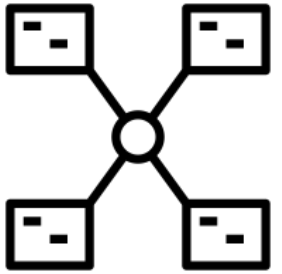
SQL Language

#4 SCHEMA

HDOOP



RDBMS



Required on reading
(dynamic schema)

Required on write
(static schema)

#5 COST

HDOOP

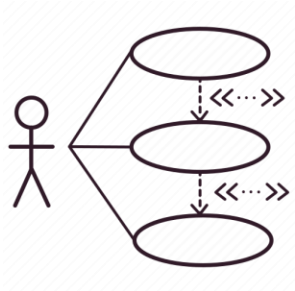
RDBMS

Free

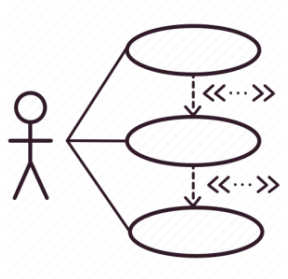
License

#6 USE CASE

HDOOP



RDBMS

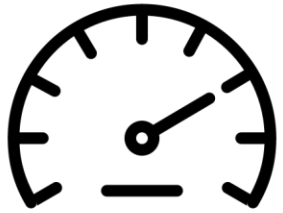


Analytics (Audio, video, logs etc), Data Discovery

OLTP (Online transaction processing

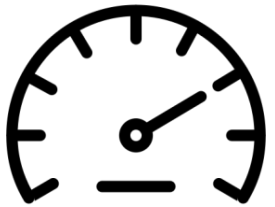
#7 SPEED

HDOOP



Both reads and writes are fast

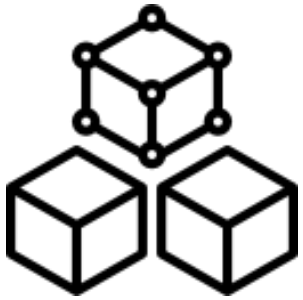
RDBMS



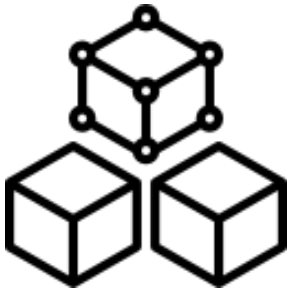
Reads are fast

#8 DATA OBJECTS

HDOOP



RDBMS



Works on Key/Value Pair

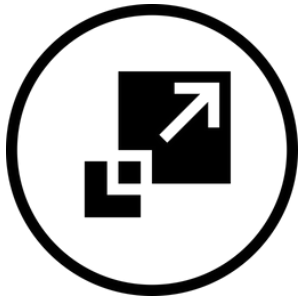
Works on Relational Tables

#9 THROUGHPUT

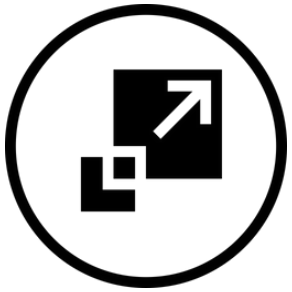
HDOOP	RDBMS
	
High	Low

#10 SCALABILITY

HDOOP



RDBMS



Horizontal

Vertical

#11 HARDWARE PROFILE

HDOOP

RDBMS

Commodity/Utility Hardware

High-End Servers

#12 INTEGRITY

HDOOP	RDBMS
	
Low	High (ACID)

SO…

By the above comparison, we have come to know that HADOOP is the best technique for handling Big Data compared to that of RDBMS.

As day by day, the data used increases and therefore a better way of handling such a huge amount of data is becoming a hectic task.

Analysis and storage of Big Data are convenient only with the help of the Hadoop eco-system than the traditional RDBMS.

Hadoop is a large-scale, open-source software framework dedicated to scalable, distributed, data-intensive computing.

This framework breakdowns large data into smaller parallelizable data sets and handles scheduling, maps each part to an intermediate value, Fault-tolerant, reliable, and supports thousands of nodes and petabytes of data, currently used in the development, production and testing environment and implementation options.