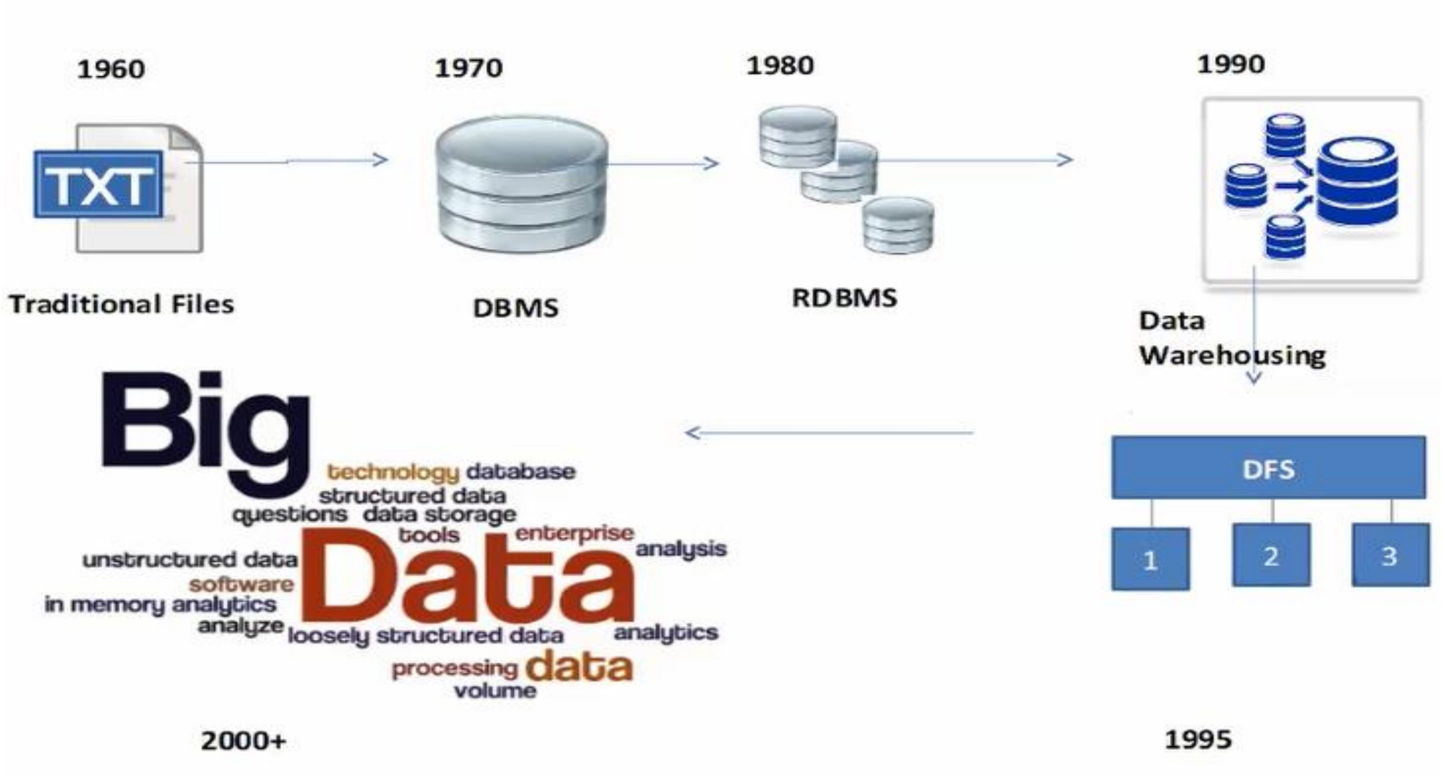
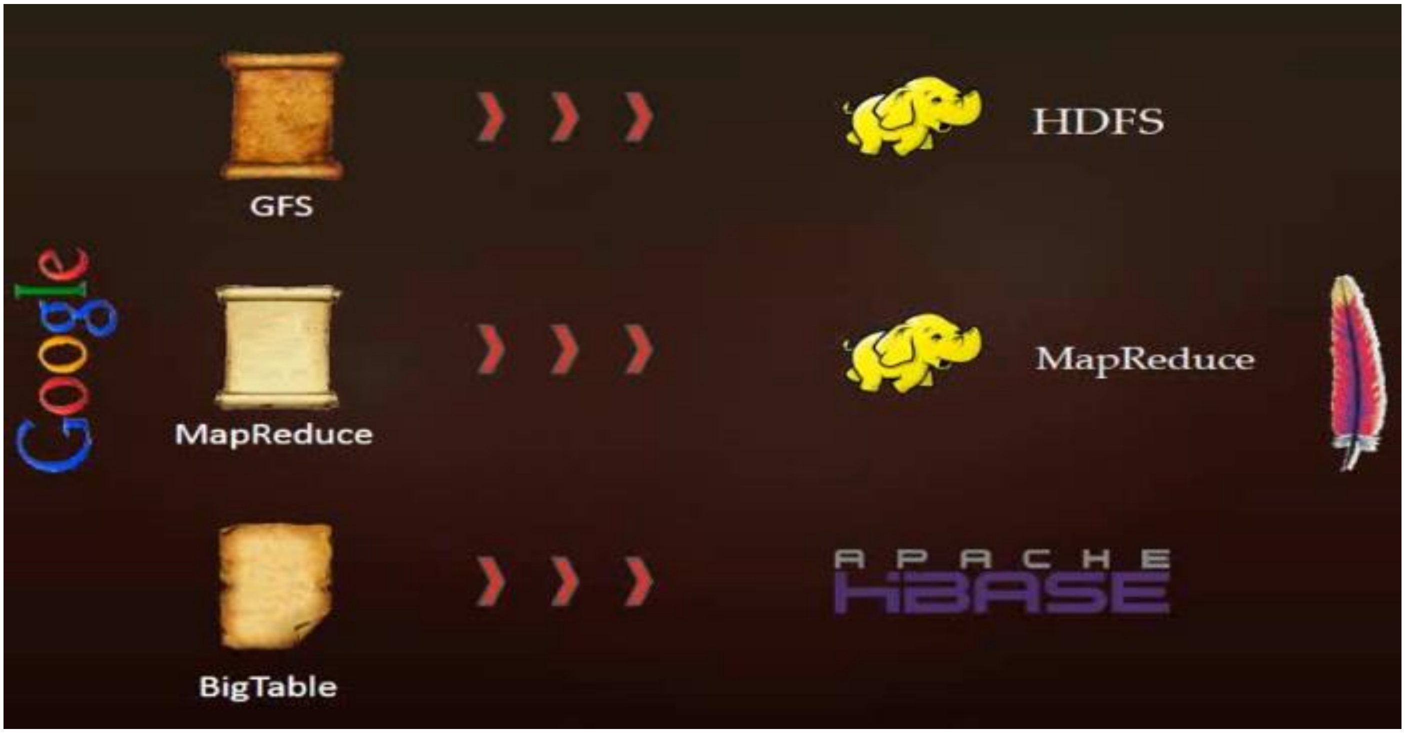
Evolution of Big Data:



Challenges:

- Store & Process this large volume of data.

Google Papers:



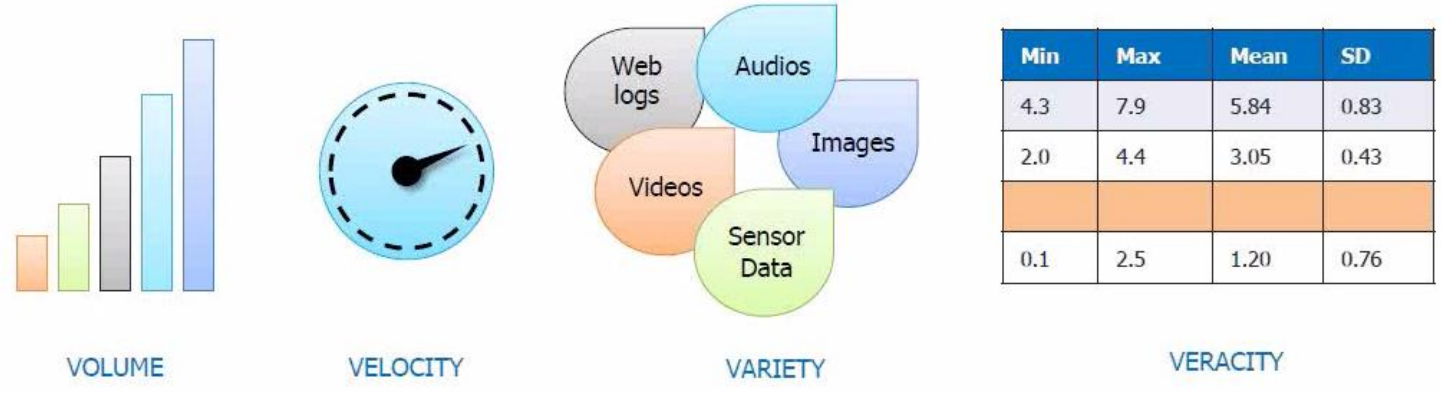
What is Big Data?

- Lots of Data (Terabytes and Petabytes)
- Big Data is the term for collection of data i.e. so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

→ Systems / Enterprises generate huge amount of data from Terabytes to and even Petabytes of information

Stock market generates about one terabyte of new trade data per day to perform stock trading analytics to determine trends for optimal trades

Types of Big Data & their Significance:



Min	Max	Mean	SD
4.3	7.9	5.84	0.83
2.0	4.4	3.05	0.43
0.1	2.5	1.20	0.76

Understanding:

Map the following to corresponding data type:
» XML files, e-mail body
» Audio, Video, Images, Archived documents
» Data from Enterprise systems (ERP, CRM etc.)

Structured: Relational Data, Data from Enterprise Systems (ERP, CRM etc.)
Semi-Structured: XML files, E-mail body
Un-Structured: Audio, Video, Images, Archived documents, Logs etc.

What is Hadoop?

- Hadoop is a framework that allows distributed processing of large data sets across clusters of commodity computers using a simple programming model.
- It is an Open-Source Data Management with scale out storage and distributed processing.

Hadoop Distributions available in Market:

- **Apache:** Vanilla flavour, as the actual code is residing in Apache repositories.
- **Hortonworks:** Popular distribution in the industry.
- **Cloudera:** It is the most popular in the industry.
- **MapR:** It has rewritten HDFS and its HDFS is faster as compared to others.
- **IBM:** Proprietary distribution is known as Big Insights.

All flavours are almost same and if you know one, you can easily work on other flavours as well.

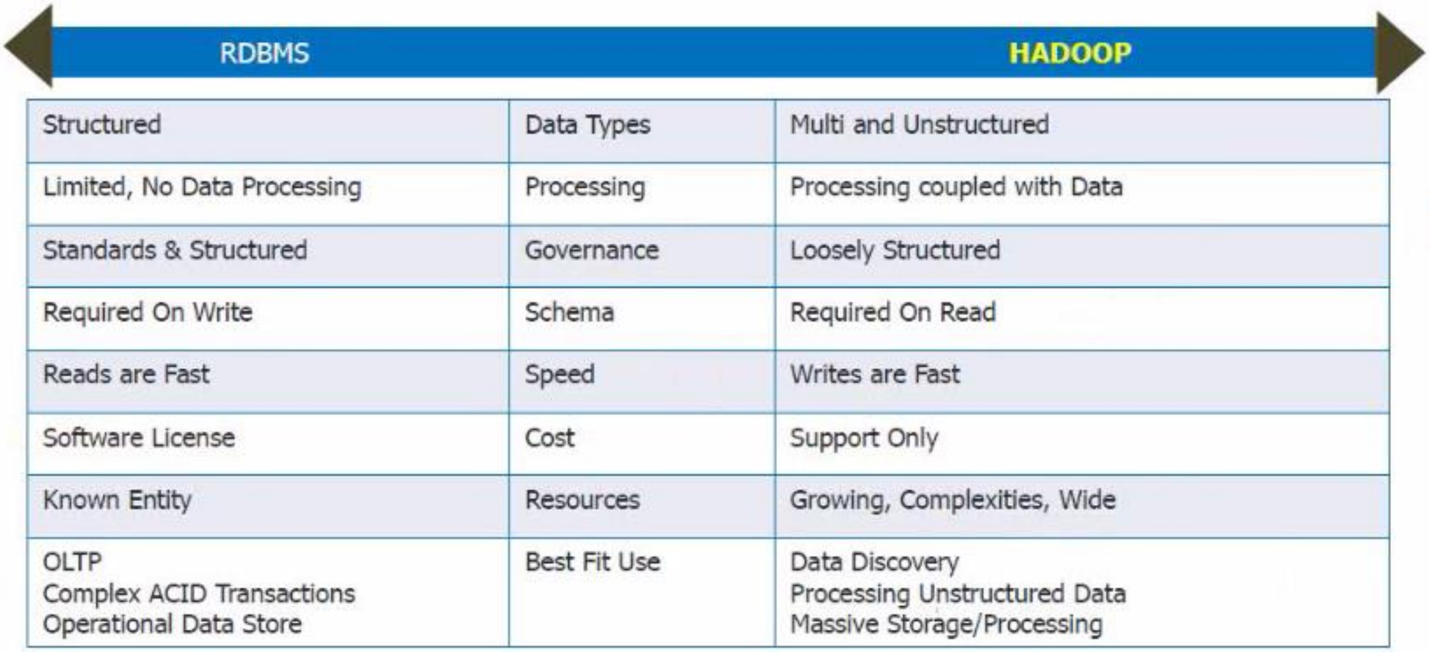
Features of Hadoop:

a) Open Source
b) Distributed Processing
c) Fault Tolerance
d) Reliability
e) High Availability
f) Scalability
g) Data Locality

Limitations of Hadoop:

a) Issue with Small Files
b) Support for Batch Processing only
c) No Real-time Processing
d) Lengthy Code

RDBMS Vs HADOOP



RDBMS		HADOOP
Structured	Data Types	Multi and Unstructured
Limited, No Data Processing	Processing	Processing coupled with Data
Standards & Structured	Governance	Loosely Structured
Required On Write	Schema	Required On Read
Reads are Fast	Speed	Writes are Fast
Software License	Cost	Support Only
Known Entity	Resources	Growing, Complexities, Wide
OLTP Complex ACID Transactions Operational Data Store	Best Fit Use	Data Discovery Processing Unstructured Data Massive Storage/Processing