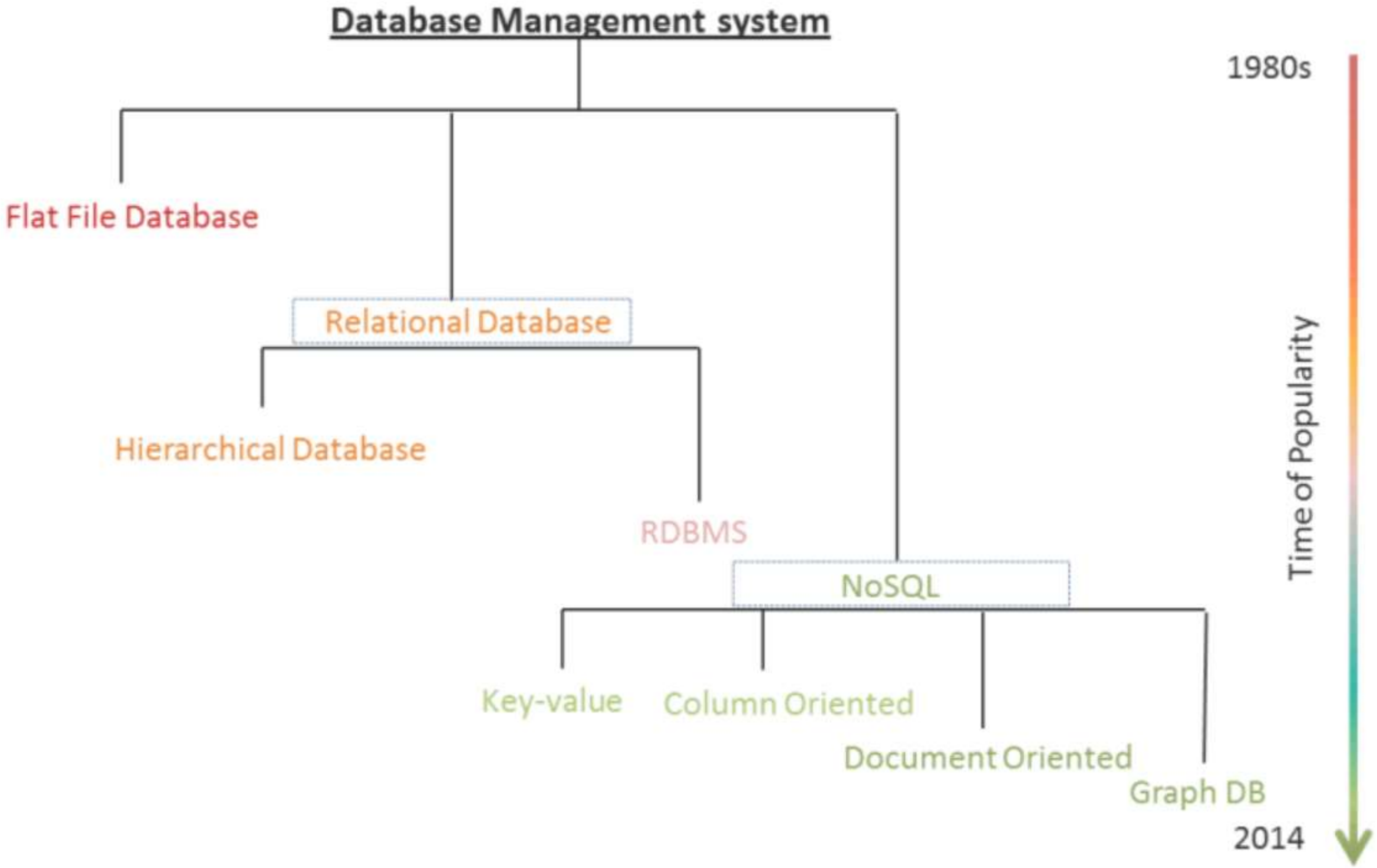
Index

- Introduction to Database
- Why Do we need database in Big Data
- Type of Database
- What is Columnar database
- Type of Columnar database
- Introduction to Document DB
- Introduction to MPP database
- Conclusion



Database Management system

Flat File Database

Relational Database

Hierarchical Database

RDBMS

NoSQL

Key-value

Column Oriented

Document Oriented

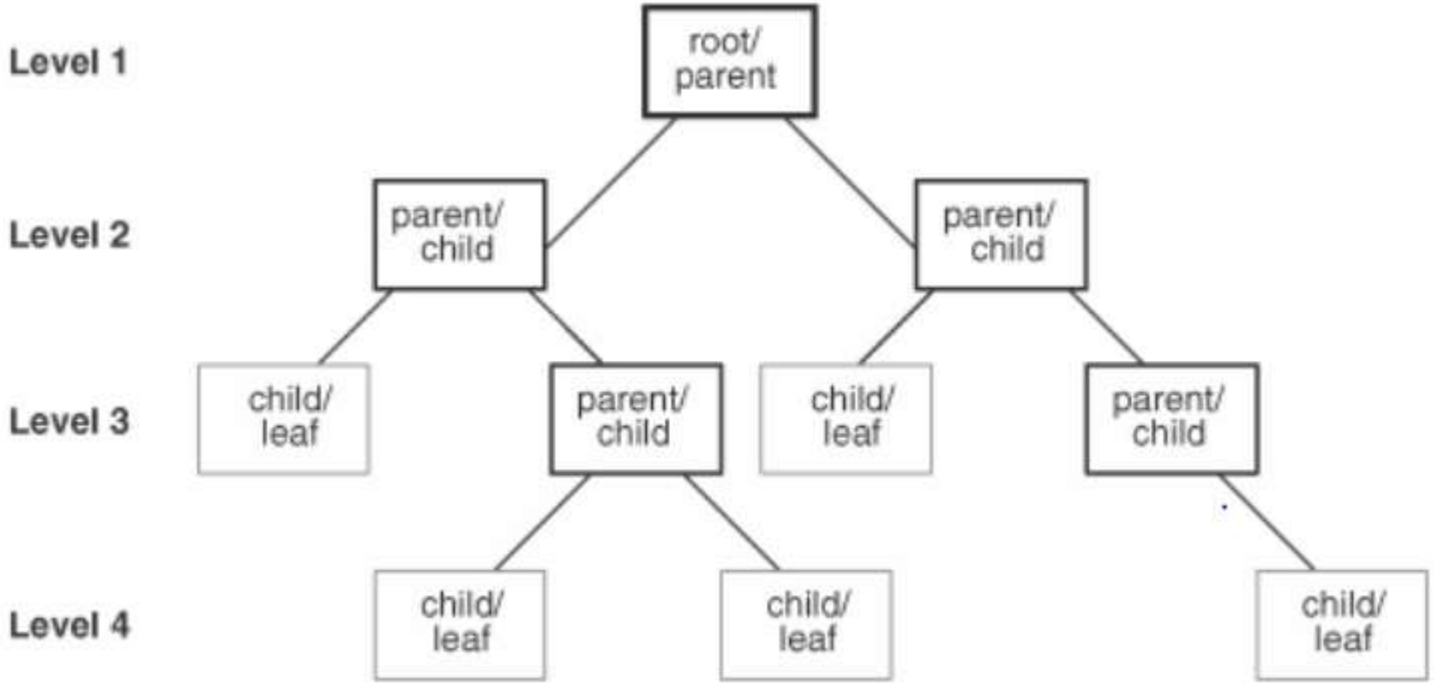
Graph DB

1980s

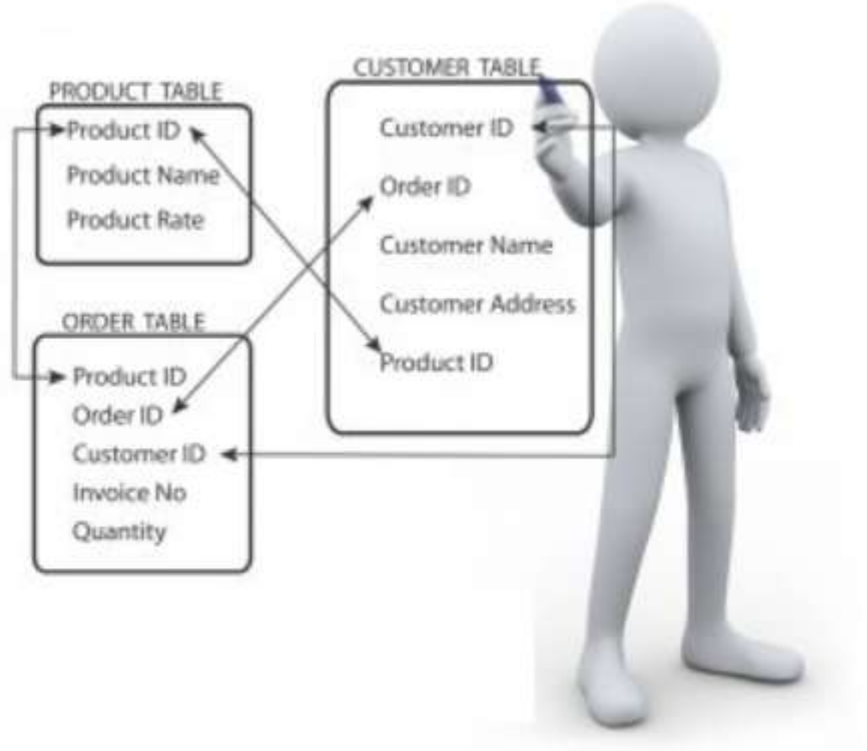
Time of Popularity

2014

Hierarchal Database



RDBMS



PRODUCT TABLE

Product ID
Product Name
Product Rate

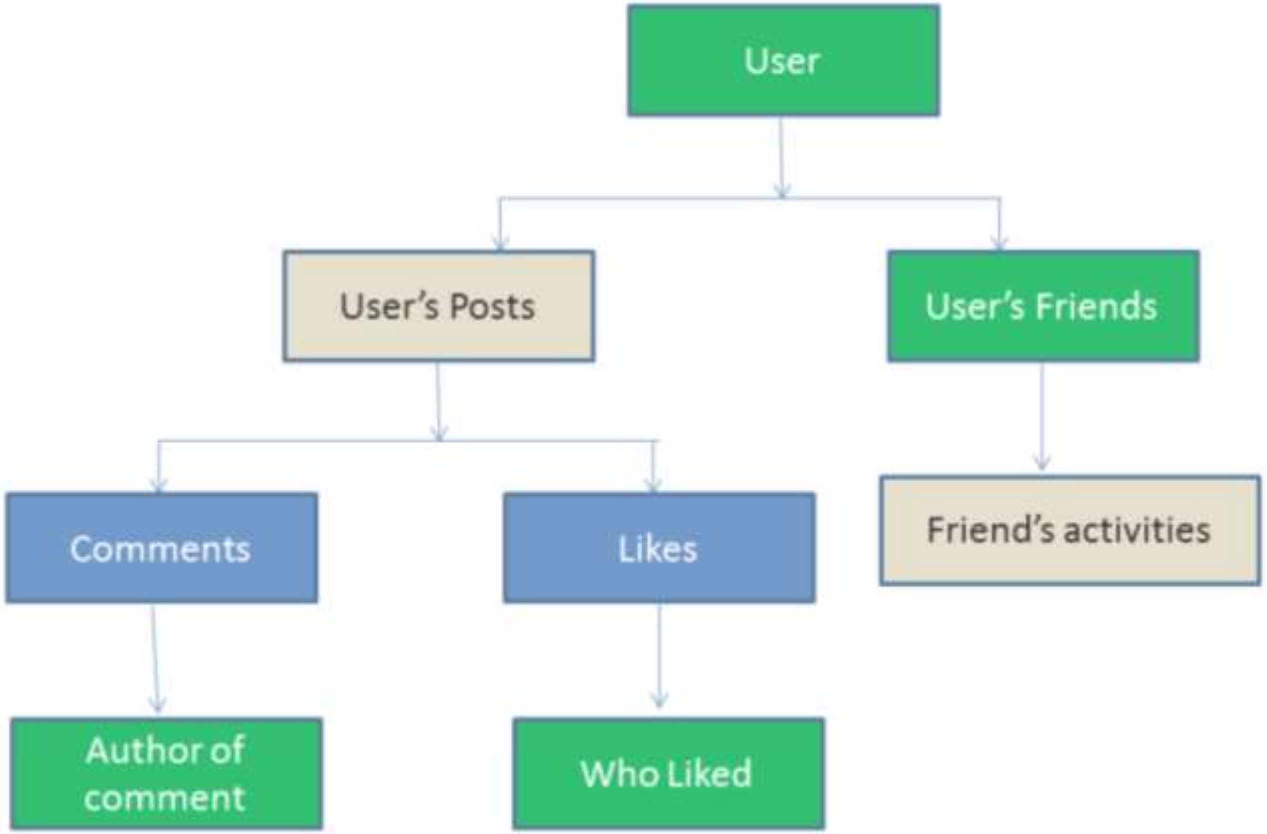
ORDER TABLE

Product ID
Order ID
Customer ID
Invoice No
Quantity

CUSTOMER TABLE

Customer ID
Order ID
Customer Name
Customer Address
Product ID

NO SQL DATABASE



User

User's Posts

User's Friends

Comments

Likes

Friend's activities

Author of comment

Who Liked

Why do we need database in Big Data

- There are three type of data we receive
  - Structured Data
    - **current data warehouse contains structured data and only structured data. It's structured because when you placed it in your relational database system a structure was enforced on it.**
  - Semi-Structure
    - **Text: XML, email or electronic data interchange messages (EDI).**
    - **Web Server Logs and Search Patterns**
    - **Sensor Data**
    - **Marketing and Sales Campaigns**
    - **Ecommerce**
    - **Brick and Mortar Retail**
    - **Supply Chain**
  - Unstructured
    - **documents produced in company,**
    - **images**
    - **videos, audio files,**
    - **social media**

WHAT IS COLUMNAR DATABASE

A columnar database is a database management system (DBMS) that stores data in columns rather than in rows as relational DBMSs do. The main differences between a columnar database and a traditional row-oriented database are centered around performance, storage necessities and schema modifying techniques. The goal of this type of database is to effectively read and write data to and from the secondary storage in order to be able to speed up the processing time in returning a query.

Note : A columnar database also be known as a column-oriented database

Different Vendors for Columnar Database

1) **SAP Sybase IQ**
2) **InfoBright**
3) **Vertica (HP)**
4) **MonetDB**
5) **ParAccel**

SAP Sybase IQ :

A highly optimized analytics server designed specifically to deliver superior performance for mission-critical business intelligence, analytics and data warehousing solutions on any standard hardware and operating system. Its column oriented grid-based architecture, patented data compression, and advanced query optimizer delivers high performance, flexibility, and economy in challenging reporting and analytics environments.

InfoBright :

Offering both a commercial (IEE) and a free community (ICE) edition, the combination of a column oriented database with their Knowledge Grid architecture delivers a self-managed, scalable, high performance analytics query platform. Allowing 50Tb using a single server, their industry-leading data compression (10:1 up to 40:1) significantly reduces storage requirements and expensive hardware infrastructures. Delivered as a MySQL engine, Infobright runs on multiple operating systems and processors needing only a minimum of 4Gb of RAM (however 16Gb is a recommended starting point).

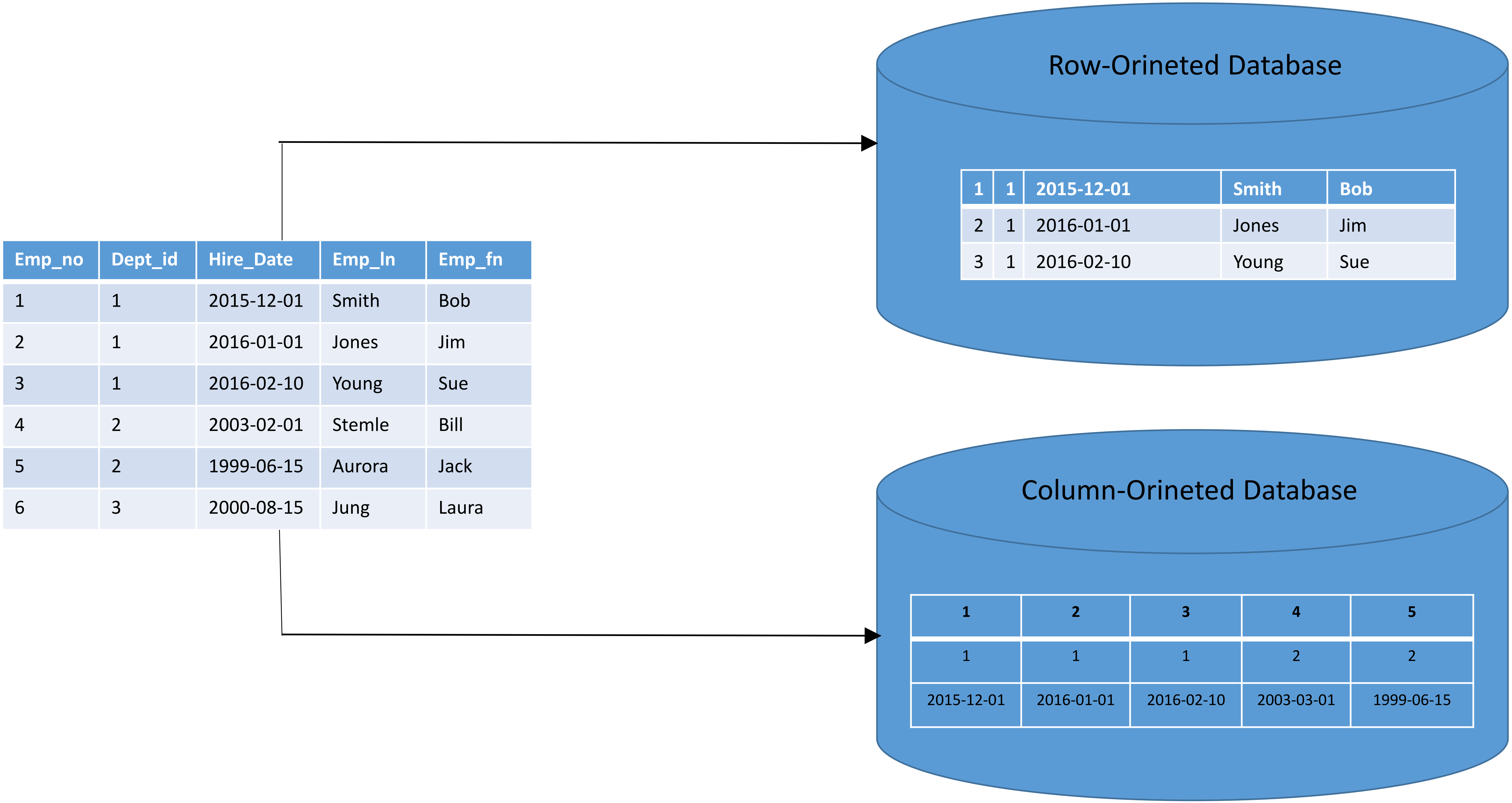
HP Vertica :

This platform was purpose built from the ground up to enable data values having high performance real-time analytics needs. With extensive data loading, queries, columnar storage, MPP architecture, and data compression features, diverse communities can develop and scale with a seamless integration ecosystem.

ParAccel :

Analytic-driven companies need a platform, not just a database where speed, agility, and complexity drive the data ecosystem. The ParAccel Analytic Platform streamlines the delivery of complex business decisions through its high performance analytic database. Designed for speed, its extensible framework supports on-demand integration and embedded functions

Column-oriented databases provide significant advantages over traditional row oriented system applied correctly; In particular for data warehouse and business intelligence environments where aggregations prevail. It would not be fair however to ignore the disadvantages.



Emp_no	Dept_id	Hire_Date	Emp_ln	Emp_fn
1	1	2015-12-01	Smith	Bob
2	1	2016-01-01	Jones	Jim
3	1	2016-02-10	Young	Sue
4	2	2003-02-01	Stemle	Bill
5	2	1999-06-15	Aurora	Jack
6	3	2000-08-15	Jung	Laura

Row-Orineted Database

1	**1**	**2015-12-01**	**Smith**	**Bob**
2	1	2016-01-01	Jones	Jim
3	1	2016-02-10	Young	Sue

Column-Orineted Database

1	**2**	**3**	**4**	**5**
1	1	1	2	2
2015-12-01	2016-01-01	2016-02-10	2003-03-01	1999-06-15

Column-Oriented Advantages

- Efficient storage and data compression
- Fast data loads
- Fast aggregation queries
- Simplified administration & configuration

‘Column-Oriented Disadvantages

- Transactions are to be avoided or just not supported
- Queries with table joins can reduce high performance
- Record updates and deletes reduce storage efficiency
- Effective partitioning/indexing schemes can be difficult to design

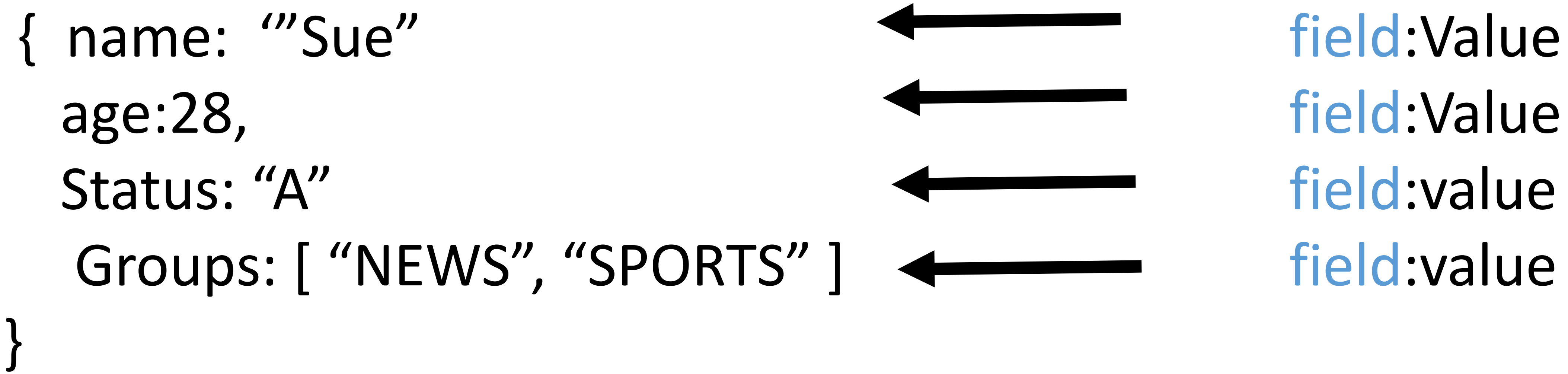
Document Database Introduction

- Documents are the main concept.
- A Document-Oriented database store and retrieves documents (XMP,JSON, BSON and so on).
- Documents are :

  - Self-describing

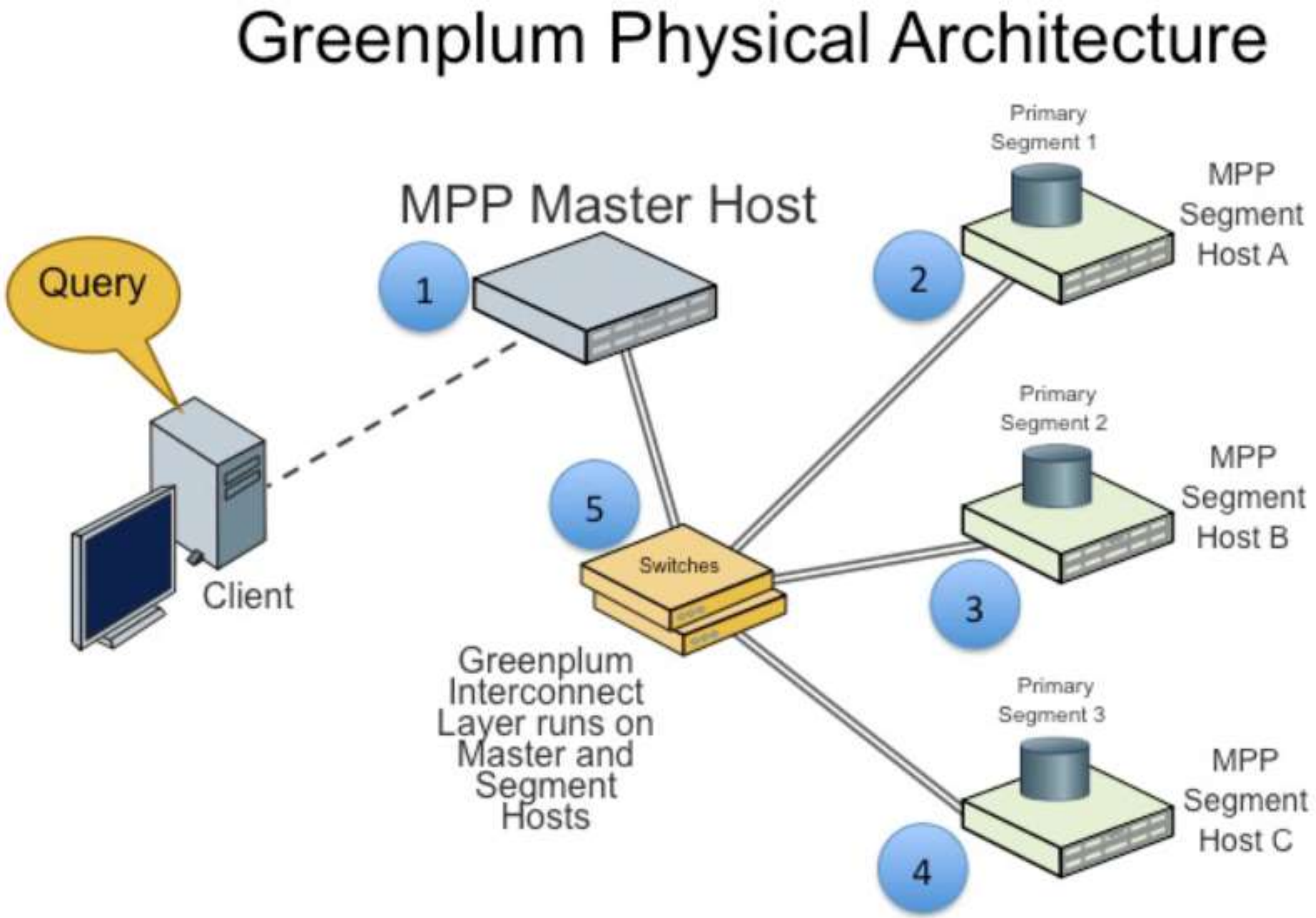
  - Hierarchical tree data structure (Map, collection and scalar values)

- Document databases stores document in the value part of the key-value store where :
  - Documents are indexed using a Btree
  - and quarried using a JavaScript Query engine



```
{  name:  “”Sue”
   age:28,
   Status: “A”
    Groups: [ “NEWS”, “SPORTS” ]
}
```

MPP (Massively Parallel Processing) Database



https://dwarehouse.wordpress.com/2012/12/28/introduction-to-massively-parallel-processing-mpp-database/

Conclusion

- ❑ Column architecture doesn’t read unnecessary columns
- ❑ Avoids decompression costs and perform operations faster.
- ❑ Use compression schemes allow us to lower our disk space requirements.