# SNS COLLEGE OF TECHNOLOGY

**Coimbatore-35**
**An Autonomous Institution**

## Department of Information Technology


### 19ITE305 – BIG DATA ANALYTICS

III B.Tech. IT/ VI SEMESTER

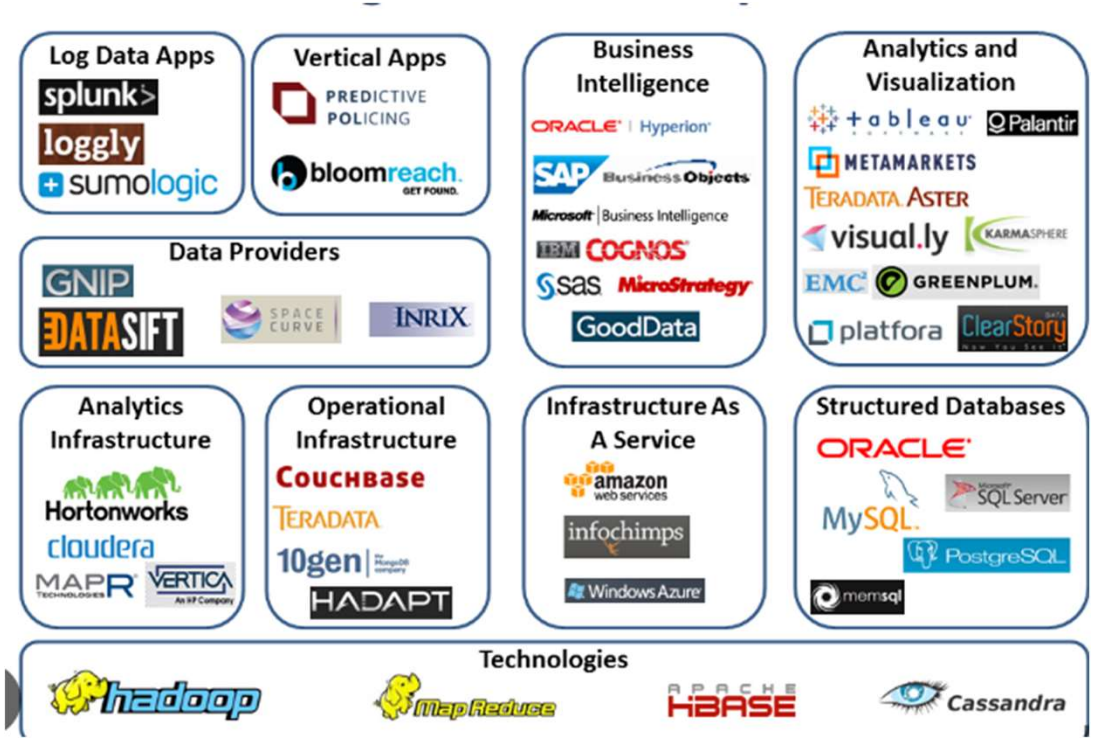
# UNIT II : INTRODUCTION TO TECHNOLOGY LANDSCAPE

**Topic 1 : NoSQL, Comparison of SQL and NoSQL**

NoSQL, Comparison of SQL and NoSQL, Hadoop - RDBMS Versus Hadoop - Distributed Computing Challenges – Hadoop Overview - Hadoop Distributed File System - Processing Data with Hadoop - Managing Resources and Applications with Hadoop YARN - Interacting with Hadoop Ecosystem

# Big Data Technology Landscape

- NoSQL
- Hadoop



# NoSQL

- NOT ONLY SQL
- First coined by Carlo Strozzi in 1998 – Lightweight, Open Source, Relational database
- Johan Oskarsson - 2009 reintroduced the term NoSQL – Open Source Distributed Network

**Features**

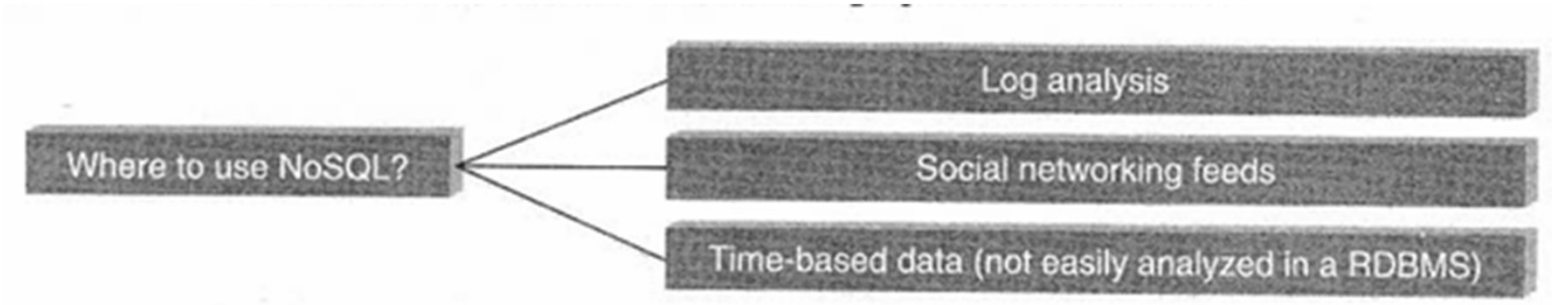
- Open Source
- Mom relational
- Distributed Scheme Less
- Cluster Friendly

It is triggered by the needs of Web 2.0 companies such as Facebook, Google, and Amazon.com.

Most NoSQL databases offer a concept of "eventual consistency" in which database changes are propagated to all nodes "eventually" (typically within milliseconds).

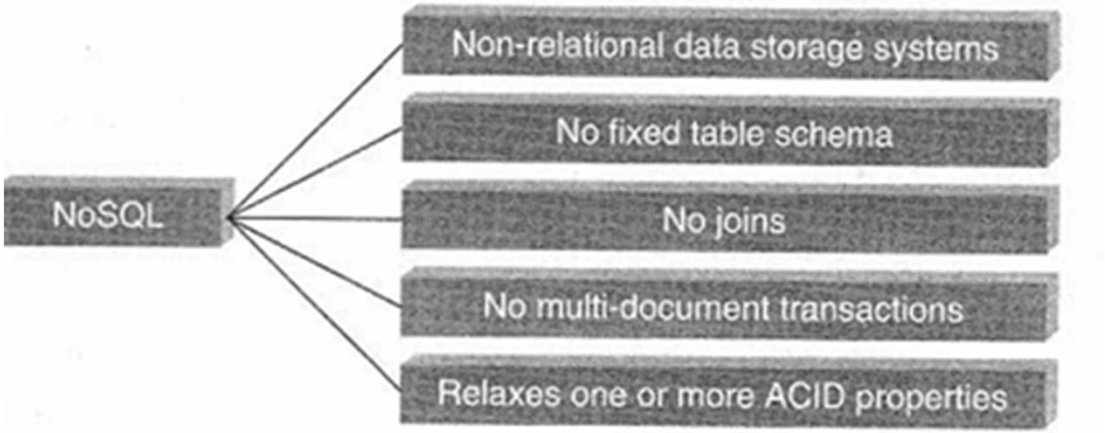
# Where is it used?

- Used in Big data and Real Time Application
- Stock log data which can then be pulled for Analysis
- **Data which cannot be stored and analyzed comfortably in RDBMS**



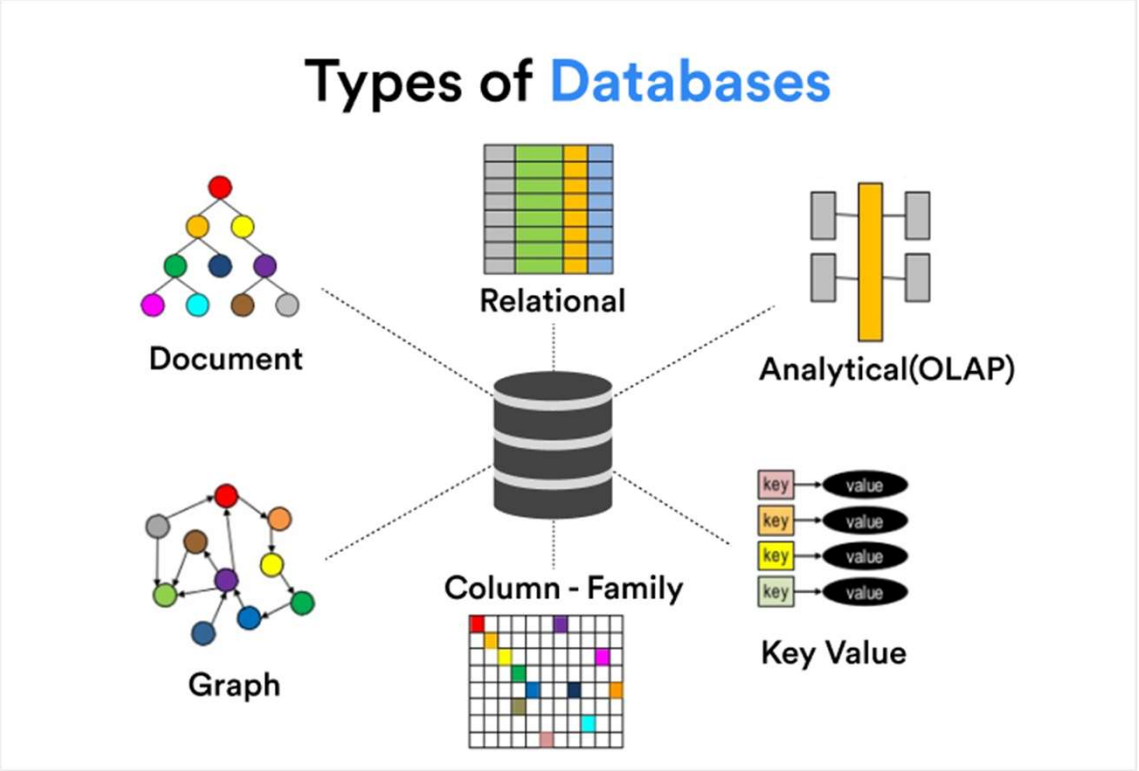
# What is it?

- Stands for NOT ONLY SQL
- Non relational , Opensource, and Distributed Databases
- Dealing with variety of data – Structured, Semi- Structured and Unstructured



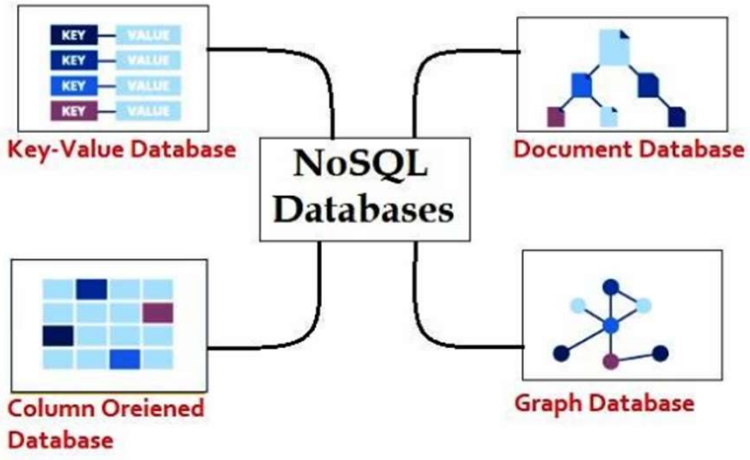
# What is it?

- Non relational
- Data is distributed across several nodes
- No Support of ACID
- No Fixed Table Schema



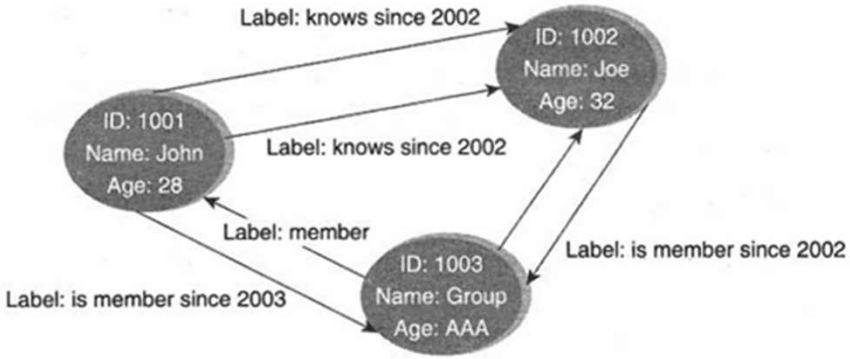
# Types of NoSQL Databases

- Broadly Classified into 2 types
  - Key – value or Big Hash table
  - Schema - Less



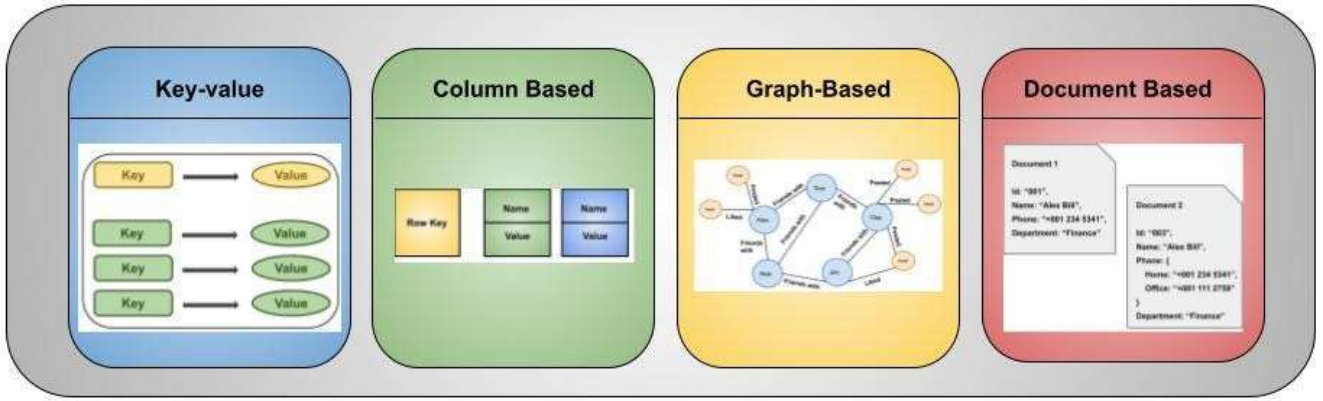

| Key Value | Column Based | Document Database | Graph Database |
|---|---|---|---|
| • In a key-value NoSQL Database, all of the data within consists of an indexed key and a value<br>• *Examples include :*<br>• *DynamoDB*<br>• *Cassandra* | • In Column Based NoSQL Database, DB is designed for storing data tables as sections of columns of data, rather than as rows of data<br>• *Examples include :*<br>• *HBase*<br>• *SAP HANA* | • This NoSQL Database expands the key-value stores where "documents" contain more complex in that they contain data and each document is assigned a unique key, which is used to retrieve the document<br>• *Examples include :*<br>• *MongoDB*<br>• *CouchDB* | • This No SQL database IS designed for data whose relations are well represented as a graph and has elements which are interconnected, with an undetermined number of relations between them<br>• *Examples include :*<br>• *Polyglot*<br>• *Neo4J* |

# Types of NoSQL Databases




*Sample Document in Document Database*

```
{
"Book Name":    "Fundamentals of Business Analytics",
"Publisher":    "Wiley India",
"Year of Publication":    "2011"
}
```



# Popular Schema Less Databases

| Key–Value Data Store | Column-Oriented Data Store | Document Data Store | Graph Data Store |
|---|---|---|---|
| • Riak | • Cassandra | • MongoDB | • InfiniteGraph |
| • Redis | • HBase | • CouchDB | • Neo4j |
| • Membase | • HyperTable | • RavenDB | • AllegroGraph |

# Why NoSQL?

- Scale out Architecture
- Large Volume of Structured, Semi- Structured and Unstructured
- Dynamic Scheme – insert the data without pre defined Schema
- Auto Sharding - Automatically Spread data Across an arbitrary number of nodes
- Replication – high Availability, Fault tolerance, and Disaster Recovery

# Advantage of NoSQL

1. Support elastic scaling:
   a) Cluster scale: It allows distribution of database across 100+ nodes often in multiple data centers.
   b) Performance scale: It sustains over 100,000+ database reads and writes per second.
   c) Data scale: It supports housing of 1 billion+ documents in the database.
2. Doesn't require a pre-defined schema: Does not require any adherence to pre-defined schema and supports flexible schema

# Advantage of NoSQL

3. Cheap and easy to implement and supports benefits of scale, high availability, fault tolerance in low cost.
4. Relaxes the data consistency requirements: Adopts CAP theorem
5. Data can be replicated on multiple nodes and can be distributed:
   a) Sharding: Automatically spread data across an arbitrary number of servers
   b) Replication: Multiple copies of the data across the cluster.

# What we miss with NoSQL?

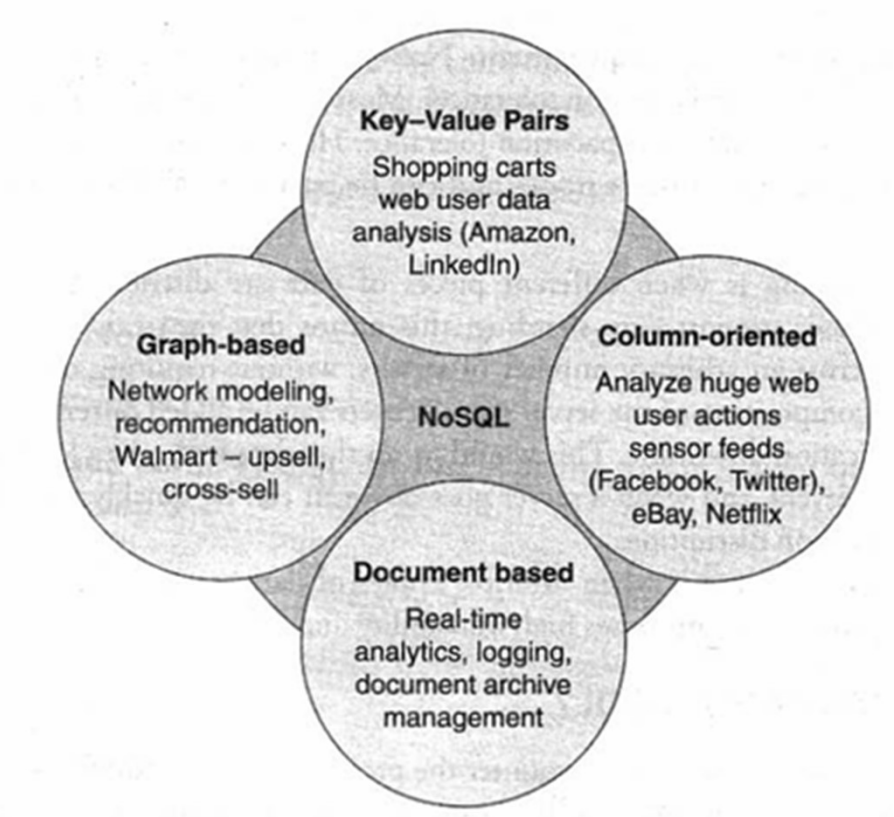
- Does not support Joins.
- Group by
- ACID properties
- Does not have standard SQL interface. But supports MQL and CQL) [M- MongoDB, C- Cassandra]
- Does not support easy integration with other applications that support SQL.

# Use of NoSQL in Industry

NoSQL is being used in varied industries. They are used to support analysis for applications such as:

- Web user data analysis
- Log analysis
- Sensor feed analysis
- Making recommendations for upsell and cross-sell, etc.

# Use of NoSQL in Industry



# NoSQL Vendors

| Company | Product | Most widely used by |
|---|---|---|
| Amazon | DynamoDB | LinkedIn, Mozilla |
| Facebook | Cassandra | Netflix, Twitter, eBay |
| Google | BigTable | Adobe Photoshop |

## TEXT BOOKS

Seema Acharya, Subhashini Chellappan, “Big Data and Analytics”, Wiley Publications, First Edition,2015

## REFERENCES


1. Judith Huruwitz, Alan Nugent, Fern Halper, Marcia Kaufman, “Big data for dummies”, John Wiley & Sons, Inc. (2013)
2. Tom White, “Hadoop The Definitive Guide”, O’Reilly Publications, Fourth Edition, 2015
3. Dirk Deroos, Paul C.Zikopoulos, Roman B.Melnky, Bruce Brown, Rafael Coss, “Hadoop For Dummies”, Wiley Publications, 2014
4. Robert D.Schneider, “Hadoop For Dummies”, John Wiley & Sons, Inc. (2012)
5. Paul Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill, 2012