NoSQL

Need for maintaining a separation between Data Management and Data Storage in these databases-Mainly focuses on high-performance scalable data storage and provides low level access to data management layer

Features

- Non Relational-They are either key-valuepairs or document oriented or column oriented or graph based databases
- Distributed- Data is distributed across several nodes in a cluster constituted of low cost commodity hardware
- No Support for ACID Properties – On contrast they have adherence to Brewer’s CAP (Consistency , Availability and partition tolerance theorem)
- No Fixed Table Schema

Types of NoSQL Databases

- Key value – It maintains a bug hash table of key and values.e.g. Dynamo

Key	Value
First Name	Ramesh
Last name	Kumar

- Document – It maintains data in collections constituted of documents. E.g. MongoDB, Apache CouchDB , CouchBase etc

contd…

- Column – Each storage block has data from only one column e.g. Cassandra , Hbase
- Graph – also called network database – A Graph stores data in nodes e.g. Neo4j , HyperGraphDB etc

Why NoSQL

- It has scale out architecture instead of the monolithic architecture of relational databases
- It can house large volumes of structured , semi structured and unstructured data
- Dynamic Schema – NoSQL database allows insertion of data without a predefined schema .It facilitates application changes in real time, which thus supports faster development, easy code integration and requires less data adminsitration

Contd…

- Auto sharding – It automatically spreads data across an arbitrary number of servers .It balances the load of data and query on the available servers ; and if and when a server goes down , it is quickly replaced without any major activity disruptions
- Replication – It offers good support for replication which in turn guarantees hugh availability , fault tolerance and disaster recovery

Advantages of NoSQL

1. Can Easily scale up and down-NoSQL Database supports scaling rapidly and elastically and even allows to scale to the cloud

(a) Cluster Scale- It allows distribution of database across 100+ nodes often in multiple data centres

(b) Performance Scale – It sustains over 100000+ database reads and writes per second

(c) Data Scale – It supports housing of 1 billion+ documents in the database

Contd…

2. Doesn’t require a predefined schema –It is pretty flexible
3. Cheap, easy to implement- deploying NoSQL properly allows for all of the benefits of scale, high availability , fault tolerance , etc while also lowering operational costs
4. Relaxes the data consistency requirements – adherence to CAP theorem .Most of the NoSQL Databases compromise on consistency in favor of availability and partition tolerance , however they go for eventual consistency

Contd...

5. Data can be replicated to multiple nodes and can be partitioned

(a) sharding – Sharding is when different pieces of data are distributed across multiple servers .NoSQL databases support auto-sharding meaning they can natively and automatically spread data across an arbitrary number of servers , without requiring the application to even be aware of the composition of the server pool.Servers can be added or removed from the data layer without application downtime.This would mean that data and query load are automatically balanced across servers, and when a server goes down, it can be quickly and

Contd…

(b) Replication – when multiple copies of data are stored across the cluster and even across data centres.This promises high availability and fault tolerance

What we miss with NoSQL

- JOINS
- Group by
- ACID Properties
- SQL
- Easy integration with other applications that support SQL

Use of NoSQL in Industry

- Key value pairs
  Shopping carts
  web user data analysis (Amazon,Linkedin)
- Graph Based
  Network modeling
  Recommendation
  Walmart – upsell and cross sell

Contd…

- Column Oriented
  Analyze huge web
    User actions
   sensor feeds

(Facebook , Twitter , eBay, Netflix)

- Document Based
  Real Time
  Analytics, logging ,
  document achive management

NoSQL Vendors

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

Scale-up

- Issues with scaling up when the dataset is just too big
- RDBMS were not designed to be distributed
- Began to look at multi-node database solutions
- Known as ‘scaling out’ or ‘horizontal scaling’
- Different approaches include:
  - Master-slave
  - Sharding

Benefits of No-SQL Databases

- Scale (Horizontal)
- Simple Data Model (Fewer Joins)
- Streaming/Volume
- Reliability
- Schema-Less (No Modeling/Prototyping)
- Rapid Development
- Flexible-can handle all types of data
- Cheaper Than Relational Database
- Wide Data Type-Variety
- Uses Large Binary Objects for storing Large Data
- Bulk upload
- Graphs
- Distributed Storage-Lower Administration-Real Time Analysis

Challenges against NoSQL Databases

- ACID Transactions
- Can not use SQL
- Ecosystem/ tools/ass-ons
- Can not perform searches
- Data Loss
- No Referential Integrity
- Lack of availability of Expertise

Aggregate Data Models

Data Model	Performance	Scalability	Flexibility	Functionality
Key-Value	High	High	High	Variable
Column-Oriented	High	High	Moderate	Minimal
Document Oriented	High	Variable(High)	High	Variable (Low)
Graph	Variable	Variable	High	Graph Theory
Relational	Variable	Variable	Low	Relational Algebra

Master-Slave

- All writes are written to the master. All reads performed against the replicated slave databases
- Critical reads may be incorrect as writes may not have been propagated down
- Large data sets can pose problems as master needs to duplicate data to slaves

Sharding

- Partition or sharding
  - Scales well for both reads and writes
  - Not transparent, application needs to be partition-aware
  - Can no longer have relationships/joins across partitions
  - Loss of referential integrity across shards

Overview of Hadoop Eco-System

1. HDFS- It simply stores data files as close to the original form as possible
2. Hbase- It is Hadoop’s database and compares well with an RDBMS.It supports structured data storage for large tables.
3. Hive – It enables analysis of large datasets using a language very similar to standard ANSI SQL. This implies that anyone familiar with SQL should be able to access data stored on a adoop cluster

Contd…

4. Pig – It is an easy to understand data flow language .It helps with the analysis of large datasets which is quiet the order with Hadoop. Even if one does not have the proficiency in MapReduce programming, the analysts and the persons entrusted with the task of comprehending data will still be able to analyze the data in a Hadoop cluster as the Pig scripts are automatically converted into MapReduce jobs by the Pig interpreter.

5. ZooKeeper – It is a coordination service for distributed applications.
6. Oozie – It is a workflow scheduler system to manage apache Hadoop jobs
7. Mahout – It is a scalable machine learning and data mining library
8. Chukwa – It is a data collection system for managing large distributed systems
9. Sqoop – It is used to transfer bulk data between Hadoop and structured data stores such as relational databases
10. Ambari – It is a web based tool for provisioning , managing and monitoring Apache Hadoop clusters

Hadoop Distributions

- Hadoop is an open source Apache project .The core aspect of hadoop include following :

1. Hadoop Common
2. HDFS
3. Hadoop YARN (Yet another resource negotiator)
4. Hadoop Map Reduce

Contd...