NoSQL

Data Science

Relational Databases

Rational Databases have a well-established, strong and on-going position in modern business, supported by:

- an extensive volume of existing applications;
- a large suite of tools;
- a huge pool of expert labor.

However, with massive changes in the nature and volume of data that companies have to deal with alternative more flexible systems are required.

NoSQL: The Name

- SQL = Traditional Relational DBMS
- Recognition over the last decade or so that not every data management/analysis problem is best solved using a traditional relational DBMS
- NoSQL = “Not only SQL”, not using traditional relational DBMS
- Some NoSQL systems use SQL-like language

Pro's of a NoSQL System

As an alterative to traditional relational DBMS, NoSQL has a number of advantages:

1. Flexible Schema
2. Quicker/cheaper to set up
3. Massive scalability
4. Relaxed consistency giving higher performance and availability

Con's of a NoSQL System

However, there is a trade-off.

In achieving such gains in flexibility, scalability and performance, NoSQL databases have forgone, among other thing:

1. Expressive querying language;
2. Secondary Indexing;
3. Consistency;
4. Strong support & expertise.

Types of Data

Its not only the volume of data that companies need has deal with that has changed, but also the nature of the data:

- Structured Data
- Semi-structured data
- Unstructured data
- Polymorphic data

Thus the need for greater flexibility.

Example 1: Web Log Analysis

- Each record has userid, url, timestamp, additional info
- The task is to load into a database system
- Task: find all records for:
  - Given userid
  - Given url
  - Given timestamp
  - Certain construct appearing in additional-info

Example 1: Web Log Analysis

- Task: Find the average age of a user accessing the URL
- Age is not a field in the database but it is contained in the additional info block

Example 2: Social-network graph

- Each record: userid
- Separate records: userid, name, age, gender
- Task: Find all friends of friends of … of friends of a given user

Example 3: Wikipedia pages

- Large collection of documents with a combination of structured and unstructured data
- Task: Retrieve introductory paragraph of all pages about US presidents before 1990

NoSQL

- Class of non-relational data storage systems.
- Usually do not require a fixed table schema nor do they use the concept of joins
- All NoSQL offerings relax one or more of the ACID properties – Atomicity, consistency, isolation, durability
- BASE – basically available, soft state, eventual consistency

The NoSQL Movement

- Three major papers were the seeds of the NoSQL movement:
- BigTable (Google)
- Dynamo (Amazon)
  - Gossip protocol (discovery and error detection)
  - Distributed key-value data store
  - Eventual consistency
- CAP Theorem

CAP Theorem

It is impossible for a distributed computer system to simultaneously provide all three of the following guarantees:

- **Consistency:** all nodes see the same data at the same time
- **Availability:** a guarantee that every request receives a response on whether it was successful or failed
- **Partition tolerance:** the system continues to operate despite arbitrary message loss or failure of part of the system

According to the theorem, a distributed system can satisfy any two of these guarantees at the same time but not all three

NoSQL - Data Model

The main types of Data Models are:

1. Document Model

2. Graph Model

3. Key-Value Model/ Wide Column Model

Document Model

- Data is stored as documents/ collections:

  - each record is a document &

  - documents are gathered together in collections;

  - documents have a JSON like structure.

- Guardian Newspaper – each article is stored as a document & articles are brought together in collections.

Document Model

- Method of storage is most closely aligned to Object Oriented Programming… Documents viewed as Objects.
- Document: JSON, XML, other semistructured formats
- Basic operations: Insert(key,document), fetch(key), update(key), delete(key)
- Also fetch based on document contents
- Examples: CouchDB, MongoDB, SimpleDB

Graph Model

- Data is stored using nodes, edges and properties;

- Data is modelled as a network built on the relationships between specific elements.
- Used to model data concerning relationships

  … social networking data/ Fraud detection

- Neo-4J (graph-based)

Key-Value Model

- Data is stored using Key-Value pairs.
- Key/Value or ‘the big hash table’
- Values are accessed through keys only.
- Amazon S3 (Dynamo),
- Voldemort,
- Scalaris,
- Memcached