DALHOUSIE UNIVERSITY

Inspiring Minds

CONVENTIONAL DATABASES AND BIG DATA

Lecture # 1

Course: CSCI 5408 Data Management, Warehousing, and Analytics

Prepared by: Suhaib Qaiser (suhaibqaiser@dal.ca)

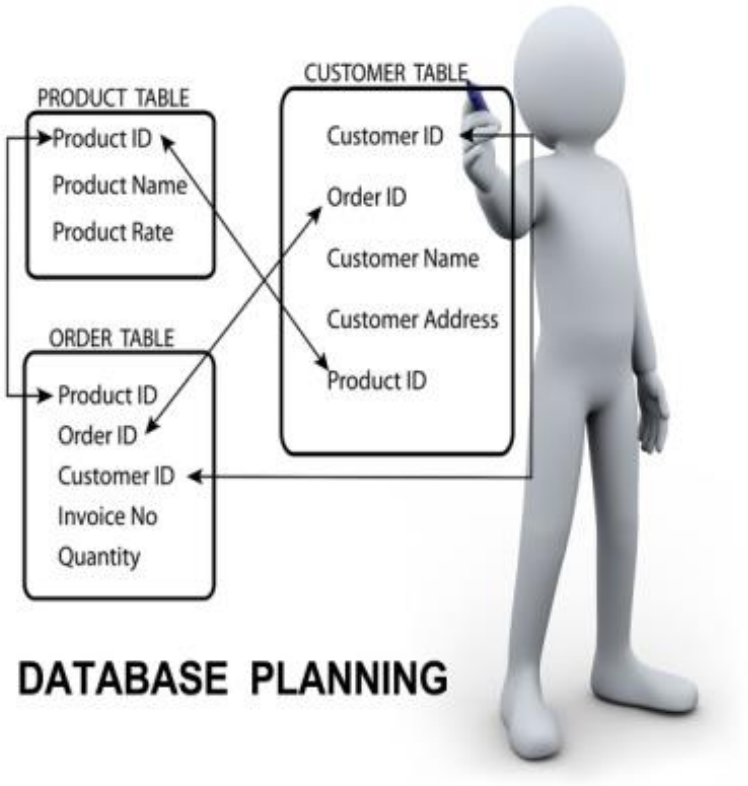
CONTENTS

CONTENTS

Relational Database Concepts

What is a RDBMS
A Relational Database Management System is a kind of database Management system where entities/data are connected to each other through relation
RDBMS is more structured and entities are closely linked together
All tables have a structure that contains at least one column/attribute
In general, a relational database is composed of tables and records

Fields and attributes
Table has fields which can be either a relational key or data field
Each field can be defined with name, data type, constraint type (PK, FK, Not NULL, etc)
Fields can also be indexed that supports fast searching



http://www.relationaldbdesign.com/relational-database-design/module2/rdbms-functions.php

Relational Database Concepts

RECORDS

Records in a database are also known as tuples

SQL is normally used to retrieve and store data inside a database

Data needs to be normalized before storing in RDBMS

Students

IDSt	LastName	IDProf	Prof	Grade
1	Mueller	3	Schmid	5
2	Meier	2	Borner	4
3	Tobler	1	Bernasconi	6

Startsituation

Result after normalisation

Students

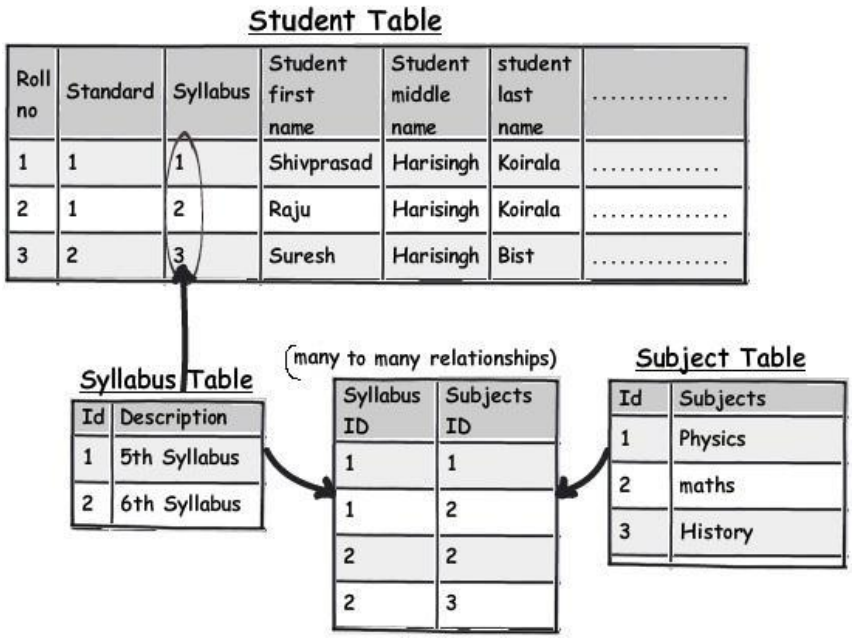
ID	LastName
1	Mueller
2	Meier
3	Tobler

Professors

IDProf	Professor
1	Bernasconi
2	Borner
3	Schmid

Grades

IDStIDProf	Grade	
1	3	5
2	2	4
3	1	6



Sources:

- http://www.gitta.info/LogicModelin/en/image/2NF.gif
- https://www.codeproject.com/KB/database/359654/aa9.jpg

Indexes in RDBMS

Indexes
Each table in RDBMS can contain multiple records and multiple attributes
Each attribute in RDBMS can be indexed to make searching more efficient and faster
As a standard practice we use integer data type for indexing because it is more efficient that way. However, there are situations where other data types can be used
Primary Keys are by default indexed in RDBMS table
Primary keys are also by default constraint to be Not Null and Unique
An index can be primary, secondary or clustered

SQL Queries

SQL Queries
SELECT * FROM EMPLOYEE
SELECT * FROM EMPLOYEE WHERE DEPT = ‘A’
SELECT EMPID, MAX(SALARY) FROM EMPLOYEE

QUESTIONS
A. Select maximum salary of an employee in department ‘A’
B. Select maximum earning salary employee from every department
C. Select maximum earning salary employee who reports to manager ‘John Chen’

Triggers in RDBMS

Triggers
Triggers are special events that are executed before or after a DML statement

Trigger can be executed at:
BEFORE DELETE, AFTER DELETE
BEFORE INSERT, AFTER INSERT
BEFORE UPDATE, AFTER UPDATE

Q. Suppose you want to audit track an accounts table and for every debit of amount you want to create an audit entry in another table.
Which trigger you will use?

Conventional Databases

Conventional databases like MySQL, Oracle, MS SQL Server are good to hold data with limited functionality

IT industry needs have grown and they want high performing database with low cost

Conventional database are not able to provide very high performance, searching from Tera bytes of data in milliseconds

NoSQL Databases are replacing conventional databases where big data transaction are required

Conventional databases like MySQL, Oracle provides more structure but less performance. They require high performing servers to boost speed

Limitation of Servers

Centralization:
Data is more centralized and in a controlled environment. But this increases maintenance cost on the company

Elasticity and Scalability:
Resources are not freely available to upgrade instantly. When user demand increases it takes sometimes weeks to upgrade infrastructure

Accessibility:
Private servers are not always accessible outside intranet. They require VPN connections which are very costly and not possible for every company to utilize

Limitation of Servers

Security:
Every company implementing their own security has advantages and disadvantages. It is more personalized but not upgraded regularly. A high skilled team is required to manage security infrastructure of central databases
Cost:
Private central servers increase cost on organizations
Performance limitation:
There is a limit on infrastructure on memory and speed utilization and availability
Fault tolerant:
Private central servers are not always fault tolerant

Big Data

"extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions."

Source: Google Search

The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."

Source: Source: Wikipedia

"Big data is changing the way people within organizations work together. It is creating a culture in which business and IT leaders must join forces to realize value from all data. Insights from big data can enable all employees to make better

Source: " Source: IBM

NoSQL

NoSQL

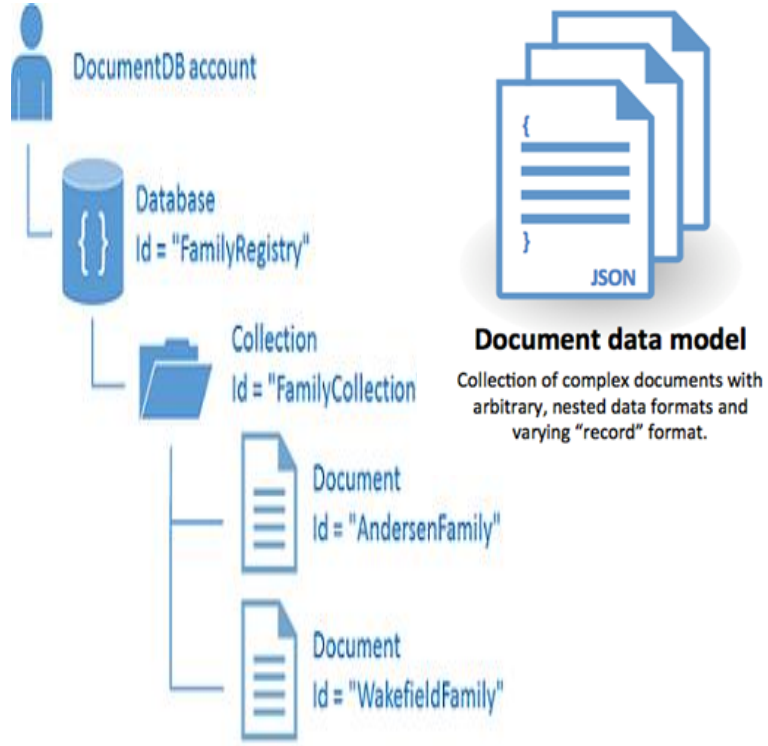
NoSQL is a way to retrieve data or information from Big Data databases. It is much faster and provides more comprehensive analytical information than normal SQL

NoSQL databases can be of various types:
Document Databases
Graph Databases
Key-Value stores
Wide column stores

What is a Document DB?

- Document databases store documents in the value part of the key-value store where:
  - Documents are indexed using a BTree
  - and queried using a JavaScript query engine

```
{
  name: "sue",          ← field: value
  age: 26,              ← field: value
  status: "A",          ← field: value
  groups: [ "news", "sports" ]   ← field: value
}
```



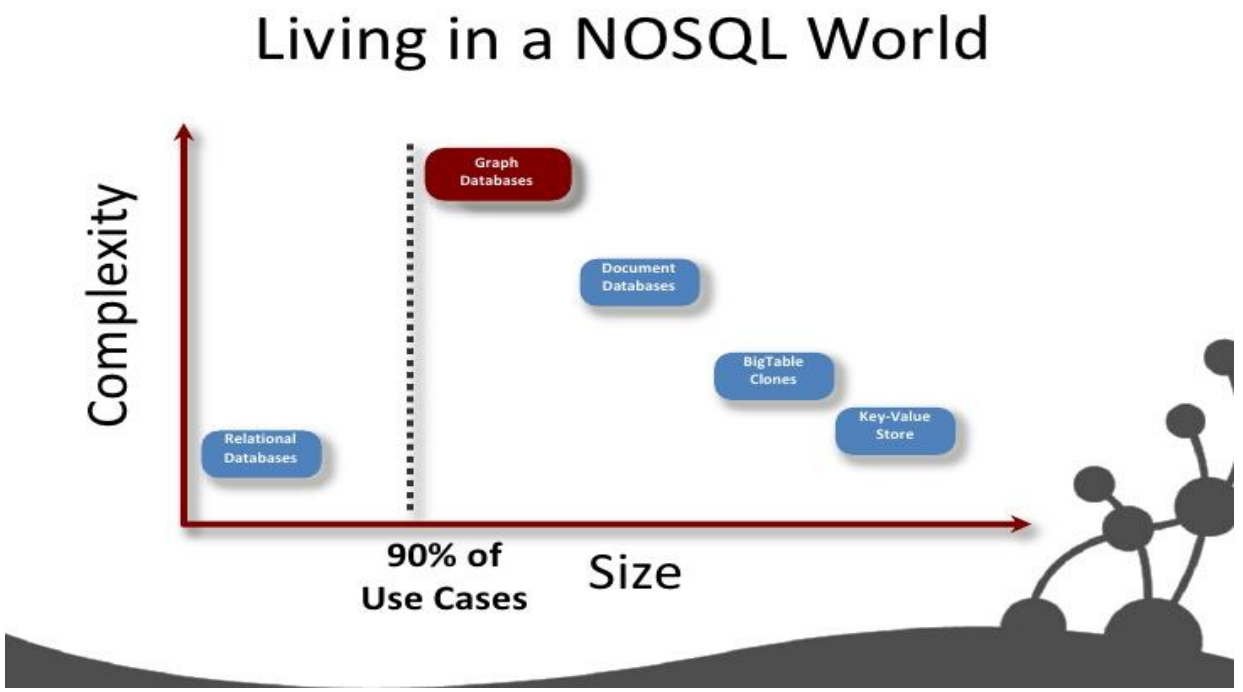
NoSQL

NoSQL

NoSQL is a way to retrieve data or information from Big Data databases.
It is much faster and provides more comprehensive analytical information than normal SQL.

NoSQL databases can be of various types:

Graph Databases



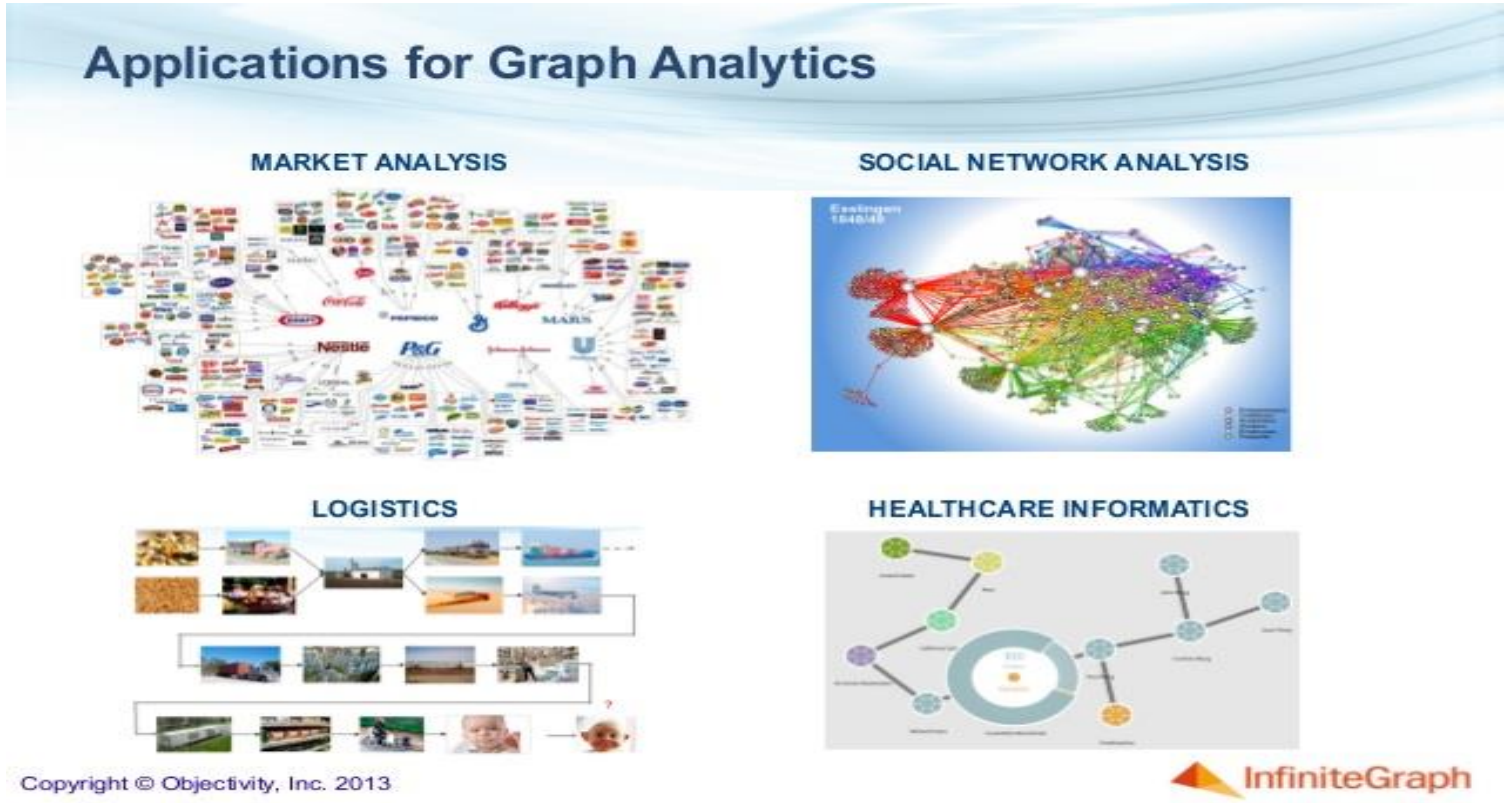
NoSQL

NoSQL

NoSQL is a way to retrieve data or information from Big Data databases
It is much faster and provides more comprehensive analytical information than normal SQL

NoSQL
Databases can be of various types:

Graph Databases



NoSQL

NoSQL

NoSQL is a way to retrieve data or information from Big Data databases
It is much faster and provides more comprehensive analytical information than normal SQL

NoSQL

Databases can be of various types:

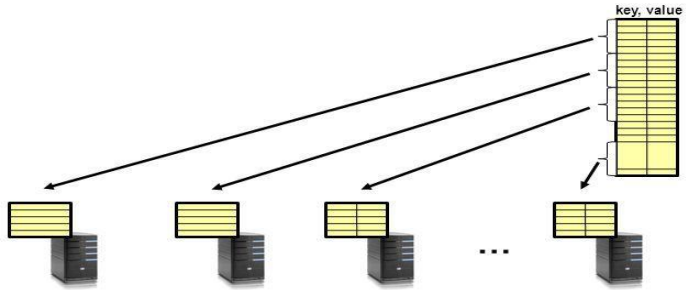
Key Value Store

A **key-value store**, or **key-value** database, is a data storage paradigm designed for storing, retrieving, and managing associative arrays, a data structure more commonly known today as a dictionary or hash.

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Key Value Store

- Also called Distributed Hash Tables (DHT)
- Main idea: partition set of key-values across many machines



NoSQL

NoSQL databases can be of various types:

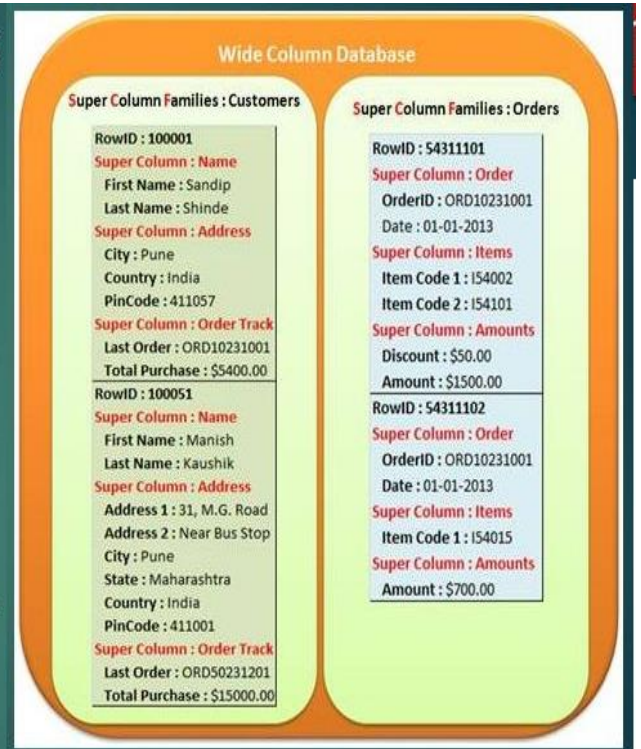
Wide column stores

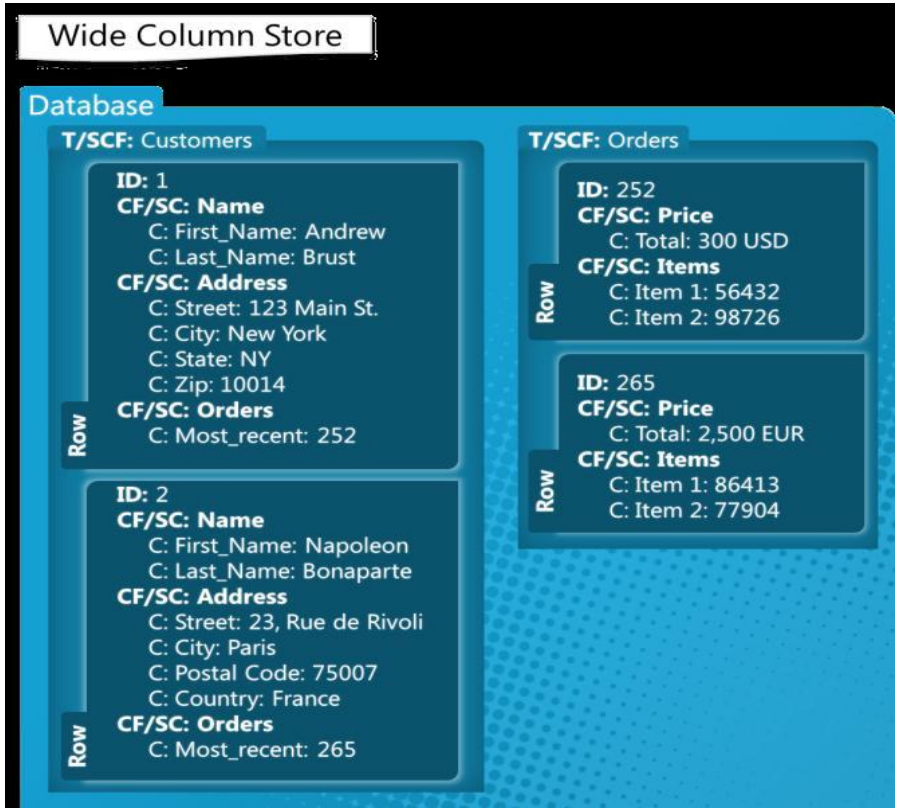
A **wide column store** is a type of key-value database. It uses tables, rows, and **columns**, but unlike a relational database, the names and format of the **columns** can vary from row to row in the same table. Examples: Bigtable.

No SQL Database Types

Wide Column Stores

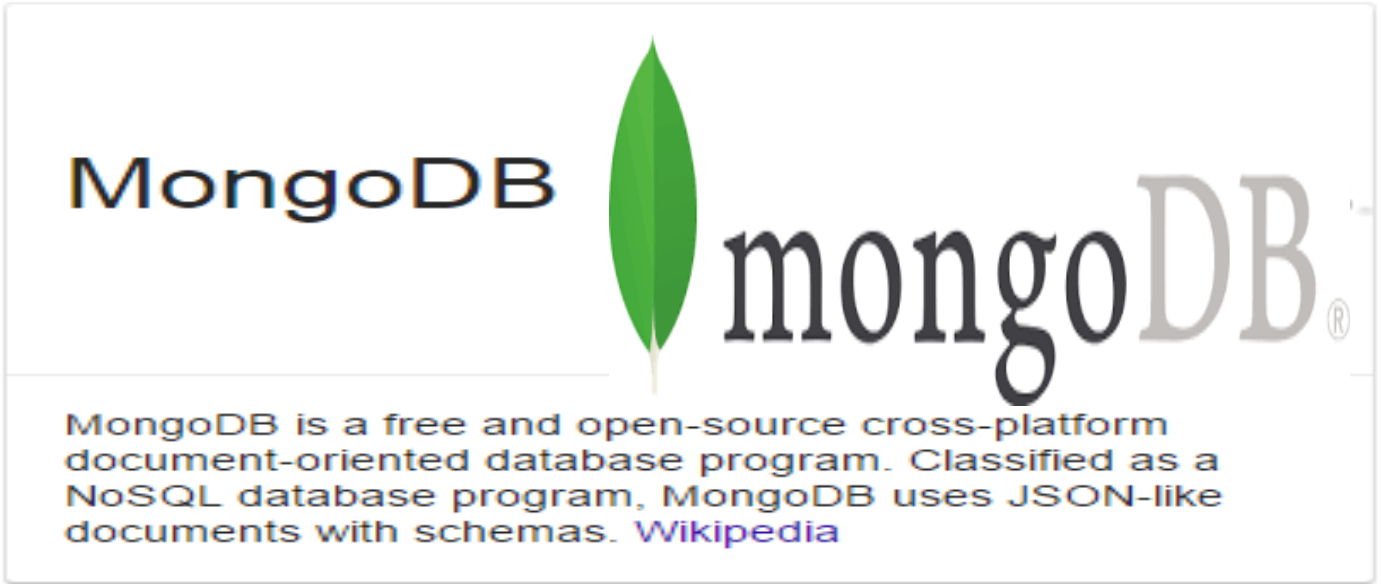
- Wide column stores have tables which contains columns
- Stores data tables as sections of columns of data
- A column family we can have different columns in each row (Super Column Families)
- E.g.: Hybertable, SimpleDB





NoSQL

Some popular NoSQL databases are:
MongoDB
Apache Casandra
Redis Document Databases



https://en.wikipedia.org/wiki/MongoDB

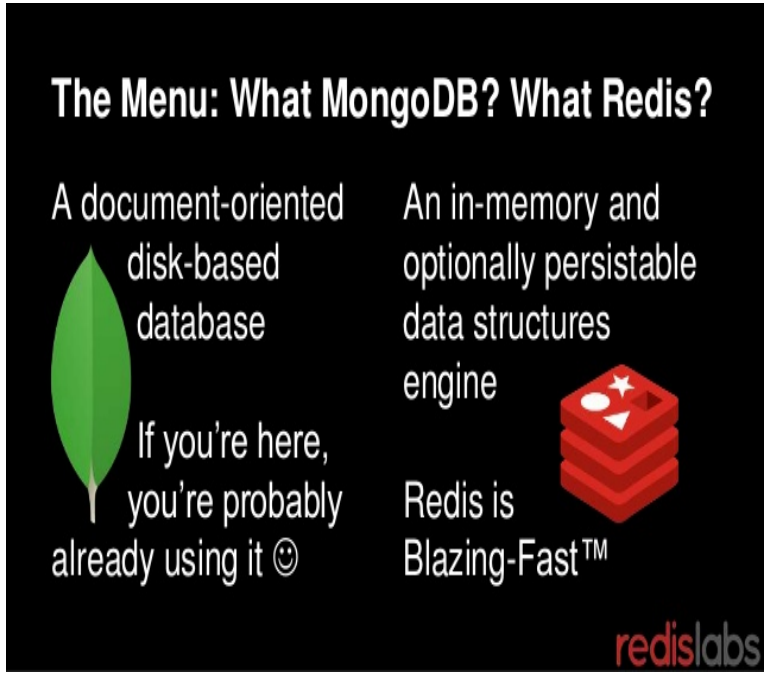
Data Partitioning - **Apache Cassandra** is a distributed database system using a shared nothing **architecture**. A single logical database is spread across a cluster of nodes and thus the need to spread data evenly amongst all participating nodes. ... Each node is responsible for part of the data.

https://dzone.com/articles/introduction-apache-cassandras

NoSQL

Some popular NoSQL databases are:

Redis Document Databases



http://www.slideshare.net/itamarhaber/why-your-mongodb-needs-redis

Some popular NoSQL databases are:

NoSQL Database are less structured and more flexible to store Terabytes of information and provides efficient ways of comparison between them

Great features

Redis is an advanced **key-value store**, where keys can contain data structures such as strings, hashes, lists, sets, and sorted sets. Redis supports a set of **atomic operations** on these data types.

Redis also supports trivial-to-setup master-subordinate replication, with very fast non-blocking first synchronization, auto-reconnection on net split and so forth.

Other features include transactions, publish/subscribe, Lua scripting, keys with a limited time to live, and configuration settings to make Redis behave like a cache.

You can use Redis from most of today's programming languages.

Azure Redis Cache uses Redis authentication and also supports SSL connections to Redis.

https://azure.microsoft.com/en-us/services/cache/

Comparison between conventional database and big data

Conventional Databases:
Supports more structure
Provides more control on data and less analytical operations
Has limitation on quantity of data stored
More costly and requires high performance speed to operate

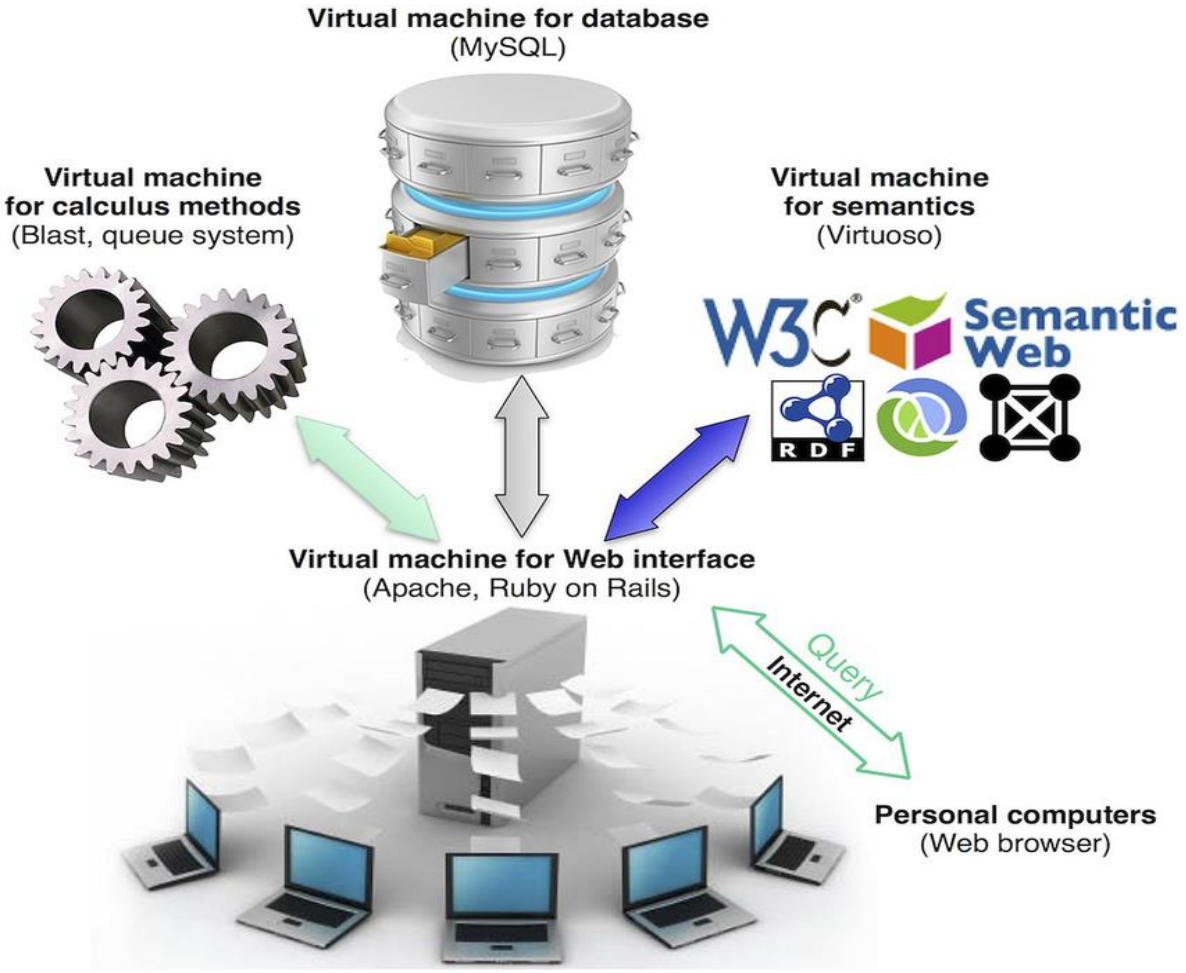
NoSQL Databases:
Supports more flexibility less structure
Provides more analytical operations and less control on data
Support huge sets of data sometimes ranging to Terabytes
Less costly and provides less resource to operate in high performance mode

Query search

Searching is the most useful and required feature from databases. It is one of the primary goal why databases are used

SQL is famous because of its aggregation, filtering, sorting and grouping functionality from a given set of data

SQL uses features like GROUP BY, SELECT, WHERE, JOIN, SUBQUERY to perform various operations on database



https://www.researchgate.net/figure/281426956_fig4_Figure-1-Database-organization-in-virtual-machines-Using-Web-browsers-users-will

Query search

Query search can be as simple as selecting an attribute value from a table

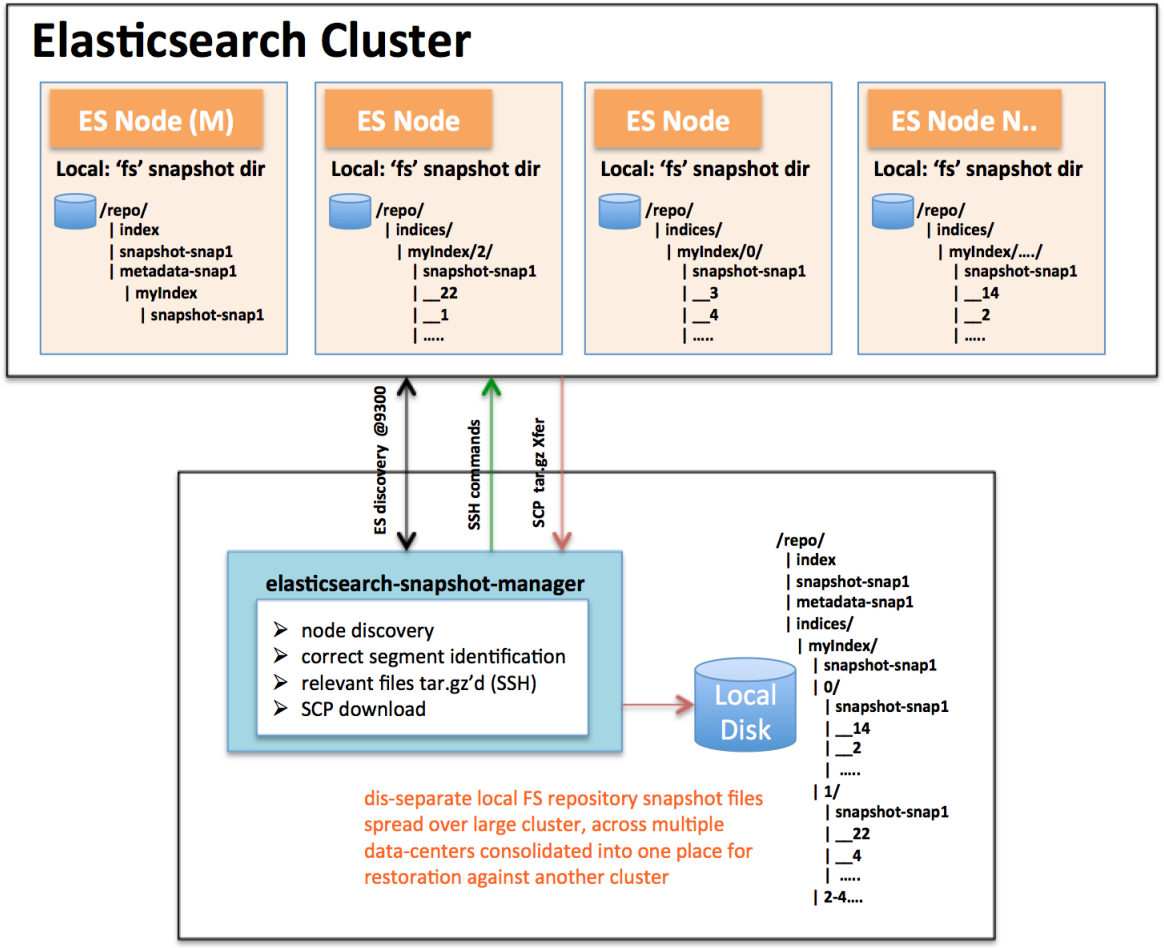
Query search can involve complex operations such as pattern recognition, sorting of millions of record and grouping or various patterns of data

With advancement of IT systems we need more advanced tools for query searching

Elastic Search

Elastic search is an open-source, broadly-distributable, readily-scalable, enterprise-grade search engine. Accessible through an extensive and elaborate API, Elastic search can power extremely fast searches that support your data discovery applications

Elastic search is able to achieve fast search responses because, instead of searching the text directly, it searches an index instead



https://bitsofinfo.wordpress.com/2015/05/29/aggregate-backup-elasticsearch-fs-snapshots-across-a-widely-distributed-cluster/

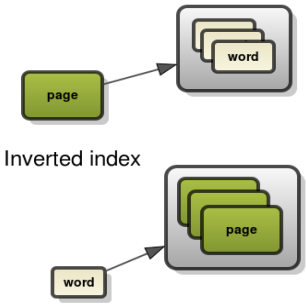
Elastic Search

This is like retrieving pages in a book related to a keyword by scanning the index at the back of a book, as opposed to searching every word of every page of the book

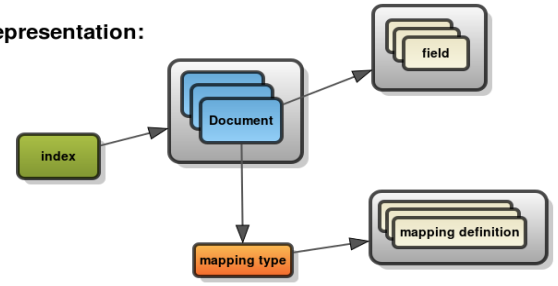
This type of index is called an **inverted index**, because it inverts a page-centric data structure (page->words) to a keyword-centric data structure (word->pages)

Elasticsearch uses Apache Lucene to create and manage this inverted index.
Source: ElasticSearchTutorial.com

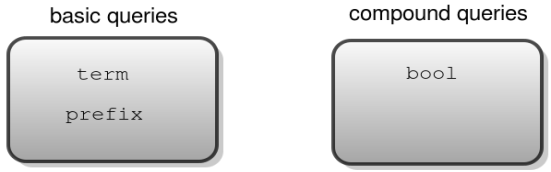
Indexing:



Data representation:



Query DSL:



main query structure

```
curl -X POST "http://localhost:9200/blog/_search?pretty=true" -d '{
                                    "from": 0,
                                    "size": 10,
                                  "query" : QUERY_JSON,
                                      FILTER_JSON,
                                      FACET_JSON,
                                       SORT_JSON
                                          }'
```

Case Study: Amazon

How does Amazon evolved their business model by using prediction and Big Data?

Source: http://www.investopedia.com/articles/insights/090716/7-ways-amazon-uses-big-data-stalk-you-amzn.asp

1. Personalized Recommendation System
2. Book Recommendations From Kindle Highlighting
3. One-Click Ordering
4. Anticipatory Shipping Model
5. Supply Chain Optimization
6. Price Optimization
7. Amazon Web Services

Reading Material

Resource # 1

Jignesh M. Patel, "Operational NoSQL Systems: What's New and What's Next?", *Computer*, vol. 49, no. , pp. 23-30, Apr. 2016, doi:10.1109/MC.2016.118
URL: https://www.computer.org/csdl/mags/co/2016/04/mco2016040023.pdf
Copy available in Dal Bright Space - https://dal.brightspace.com

Resource # 2

Online Tutorial: http://www.studytonight.com/dbms/overview-of-dbms

Resource # 3

A. Parssian, W. Yeoh and M. S. Ee, "Quality-Based SQL: Specifying Information Quality in Relational Database Queries," in *Computer*, vol. 48, no. 9, pp. 69-74, Sept. 2015.
doi: 10.1109/MC.2015.264
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7274413&isnumber=7274307
Copy available in Dal Bright Space - https://dal.brightspace.com


ANY
QUESTIONS
?