Seminario de Estadística I
Introducción a la Ciencia de Datos y Machine Learning
NoSQL

Jimmy Hernández

Universidad Nacional Autónoma de México
Facultad de Ciencias

Contenido

Sobre Hadoop

- Un sistema manejador o de gestión de base de datos es un software para describir, almacenar y consultar datos.
- Todos los sistemas de gestión de base de datos contienen una componente de almacenamiento y una de gestión.
- El componente de gestión contiene un lenguaje de consulta y manipulación de datos para evaluar y editar los datos y la información (SQL).

Modelo Relacional

Una de las formas más simples e intuitivas de recopilar y presentar datos es en una tabla.

Definición de Tabla

Una tabla o relación es un conjunto de tuplas presentadas en forma tabular y que cumplen los siguientes requisitos:

- **Nombre de la tabla**, una tabla tiene un nombre de tabla único.
- **Nombre de atributo**, todos los nombres de atributo son únicos dentro de una tabla y etiquetan una columna específica con la propiedad requerida.
- **Sin orden de columnas**, no se establece el número de atributos y no importa el orden de las columnas dentro de la tabla.
- **Sin orden de filas**, no se establece el número de tuplas y no importa el orden de las filas dentro de la tabla.
- **Clave de identificación** un atributo o una combinación de atributos identifica de forma única las tuplas dentro de la tabla y se declara la clave de identificación.

Modos

El modelo relacional representa tanto datos como relaciones entre datos como tablas.

Modelo Relacional

Cualquier relación R es simplemente un subconjunto de un producto cartesiano de dominios: $R \subseteq D1 \times D_2 \times ... \times D_n$ con D_i como el dominio del i-ésimo atributo/propiedad. Cualquier tupla r es, por lo tanto, un conjunto de valores de datos específicos o manifestaciones, $r = (d_1, d_2, ..., d_n)$. Tenga en cuenta que esta definición significa que cualquier tupla solo puede existir una vez dentro de cualquier tabla, es decir, $R = \{r_1, r_2, ..., r_m\}$.

RDBMS

Un sistema RDBMS, Relational database management system por sus siglas en ingles tiene las siguientes propiedades

Propiedades

- **Modelo**, el modelo de base de datos sigue el modelo relacional, todos los datos y las relaciones de datos se representan en tablas.
- **Esquema**, las definiciones de tablas y atributos se almacenan en el esquema de base de datos relacional. El esquema contiene además la definición de las claves de identificación y las reglas para el aseguramiento de la integridad.
- **Lenguaje**, el sistema de base de datos incluye SQL para la definición, selección y manipulación de datos.
- **Arquitectura**, independencia de datos, es decir, los datos y las aplicaciones están en su mayoría segregados.
- **Operaciones multiusuario**, varios usuarios pueden consultar o manipular la misma base de datos al mismo tiempo.
- **Garantía de coherencia**, el sistema de gestión de bases de datos proporciona herramientas para garantizar la integridad de los datos.

Limitaciones de una RDBMS

Limitaciones

- Uno de los principales problemas del modelo relacional es lo que algunos autores conocen como $\ll$ *impedancia* $\gg$ o desajuste por impedancia, que es la imposibilidad de adaptar las estructuras de datos en memoria al modelo relacional, por ejemplo convertir objetos en registros y viceversa.
- Otra limitante de las bases de datos relacionales es su integración en un clúster, ya que no están diseñadas para escalarse de manera horizontal.

Caracteristica de las NoSql

Debido a estas limitantes y con advenimiento de Big Data y la variedad de la información es que surgen la llamadas NoSql (Not Only SQL) o no relacionales, las cuales son muy faciles de escalar de manera horizontal y permiten trabajar con un conjunto mas rico de datos.

Características

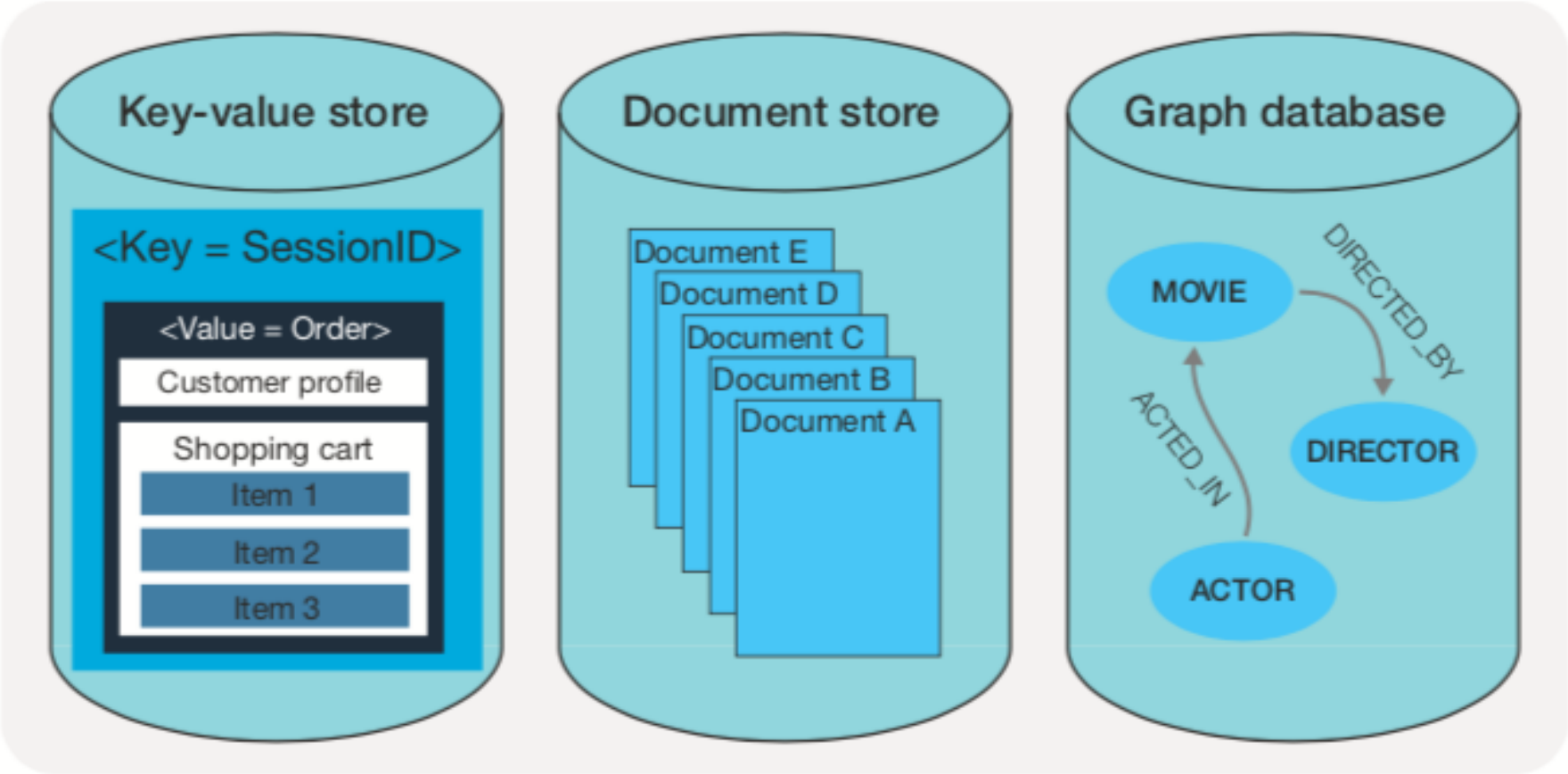
- No usan SQL como lenguaje de consultas, sin embargo, algunas de ellas utilizan lenguajes de consultas similares.
- Fácilmente escalables, trabajan muy bien en modo clúster, aunque no todas están diseñadas para ello.
- Las bases de datos relacionales usan transacciones ACID para manejar la consistencia de la base de datos, sin embargo, esto choca con un entorno de clu?ster, de manera que ofrecen diversas opciones para implementar la consistencia y la distribución.
- Las bases de datos NoSQL no tienen un esquema fijo.

Familias

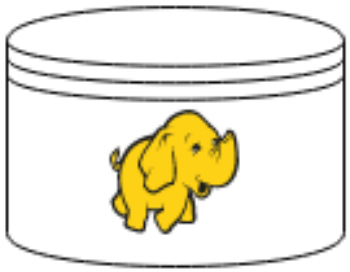
Clasificación

Por sus modelos de datos podemos clasificar a las no sql en las siguientes familias

- Bases de datos clave-valor: Riak, Redis, Dynamo, Voldemort, **ArangoDB**.
- Bases de datos orientadas a documento: MongoDB, CouchDB, **ArangoDB**.
- Bases de datos basadas en columnas: Cassandra, Hypertable, HBase, SimpleDB
- Bases de datos de grafos: Neo4J, Infinite Graph, **ArangoDB**.

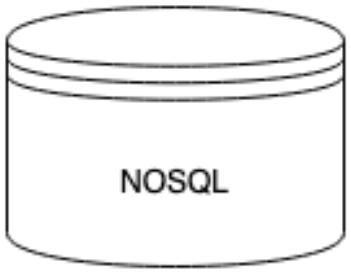

Key-value store
<Key = SessionID>
<Value = Order>
Customer profile
Shopping cart
Item 1
Item 2
Item 3
Document store
Document E
Document D
Document C
Document B
Document A
Graph database
MOVIE
DIRECTED_BY
DIRECTOR
ACTED_IN
ACTOR

- Analitica como deteccion de fraudes
- Manejo y procesamiento de gran cantidad de información
- Data Archiving



- Batch Processing
- Computo masivo en
- Analítica Predictiva

- Manejo óptimo de datos semiestructados, no estructurados.
- Logs, datos de sensores para consulta rápida
- Perfiles de usuarios



- Manejo de streams no transacciones donde la baja latencia es critica.
- Rapida lectura y escritura