

Document Databases

Definición

Document Database

Es una base no-relacional que almacena los datos como documentos estructurados.

El concepto principal es el **documento**

- Las BD almacena y recupera documentos.
- Los documentos pueden ser XML, **JSON**, BSON, etc

Definición

Document Database

Es una base no-relacional que almacena los datos como documentos estructurados.

El concepto principal es el **documento**

- Las BD almacena y recupera documentos.
- Los documentos pueden ser XML, **JSON**, BSON, etc

Documento

Es una colección de pares: nombre de campo y valor. Los valores pueden ser un valor simple o una estructura compleja como listas, otro documento o listas de documentos hijos

Definición

Document Database

Es una base no-relacional que almacena los datos como documentos estructurados.

El concepto principal es el **documento**

- Las BD almacena y recupera documentos.
- Los documentos pueden ser XML, **JSON**, BSON, etc

Documento

Es una colección de pares: nombre de campo y valor. Los valores pueden ser un valor simple o una estructura compleja como listas, otro documento o listas de documentos hijos

Ejemplos

MongoDB, RavenDB, eXist, CouchDB, CouchBase, ArangoDB

XML vs JSON

```
<order id="1234">
<customer id="52">Adam Fowler</customer>
<items>
<item qty="2" id="456" unit_price="2.00" price="4.00">Hammer</item>
<item qty="1" id="111" unit_price="0.79" price="0.79">Hammer Time</item>
</items>
<delivery_address lon="-43.24" lat="54.12">
<street>Some Place</street>
<town>My City</town>
...
</delivery_address>
</order>
```

```
{
  "orderid": 1234,
  "Customer": { "id": 52, "Nombre": "Jhon Doe" },
  "items": [ { "qty": 2, "id": 456, "unit_price": 2, "price": 4 },
             { "qty": 1, "id": 111, "unit_price": 0.79, "price": 0.79 } ],
  "delivery_address": { "lon": -43.24, "lat": 54.12, "street": "Some Place", "ciudad": "My City" }
}
```

Metodología

A Big Data Modeling Methodology for NoSQL Document Databases

Gerardo ROSSEL, Andrea MANNA

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

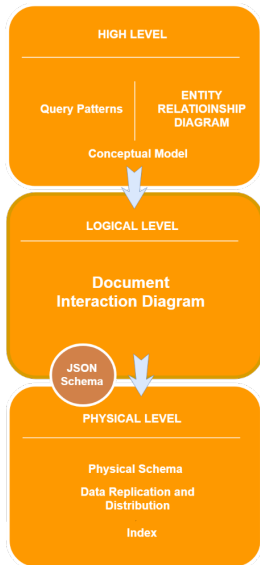
Departamento de Computación. Buenos Aires, Argentina

grossel@dc.uba.ar, amanna@dc.uba.ar

In recent years, there has been an increasing interest in the field of non-relational databases. However, far too little attention has been paid to design methodology. Key-value datastores are an important component of a class of non-relational technologies that are grouped under the name of NoSQL databases. The aim of this paper is to propose a design methodology for this type of database that allows overcoming the limitations of the traditional techniques. The proposed methodology leads to a clean design that also allows for better data management and consistency

Keywords: NoSQL, Document Databases, Conceptual Modeling, Data Modeling, NoSQL Database developing.

Modelización



Consideraciones de Diseño

Desnormalización

```
{  
  order_item_ID : 834838,  
  order_ID: 8827,  
  quantity: 3,  
  cost_per_unit: 8.50,  
  product_ID: 3648  
}
```

```
{  
  product_ID: 3648,  
  product_description: "1 package laser printer paper.  
    100% recycled.",  
  product_name : "Eco-friendly Printer Paper",  
  product_category : "office supplies",  
  list_price : 9.00  
}
```

Desnormalización

```
{  
  order_item_ID : 834838,  
  order_ID: 8827,  
  quantity: 3,  
  cost_per_unit: 8.50,  
  product_ID: 3648  
}
```

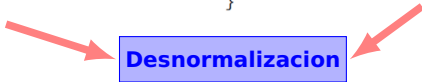
```
{  
  product_ID: 3648,  
  product_description: "1 package laser printer paper.  
    100% recycled.",  
  product_name : "Eco-friendly Printer Paper",  
  product_category : "office supplies",  
  list_price : 9.00  
}
```

Desnormalizacion

Desnormalización

```
{  
  order_item_ID : 834838,  
  order_ID: 8827,  
  quantity: 3,  
  cost_per_unit: 8.50,  
  product_ID: 3648  
}
```

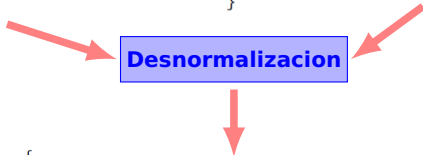
```
{  
  product_ID: 3648,  
  product_description: "1 package laser printer paper.  
    100% recycled.",  
  product_name : "Eco-friendly Printer Paper",  
  product_category : "office supplies",  
  list_price : 9.00  
}
```



Desnormalización

```
{
order_item_ID : 834838,
order_ID: 8827,
quantity: 3,
cost_per_unit: 8.50,
product_ID: 3648
}
```

```
{
product_ID: 3648,
product_description: "1 package laser printer paper.
100% recycled.",
product_name : "Eco-friendly Printer Paper",
product_category : "office supplies",
list_price : 9.00
}
```



Desnormalizacion

```
{
order_item_ID : 834838,
order_ID: 8827,
quantity: 3,
cost_per_unit: 8.50,
product :
{
product_description: "1 package laser printer
paper. 100% recycled.",
product_name : "Eco-friendly Printer Paper",
product_category : "office supplies",
list_price : 9.00
}
}
```

Desnormalización

¿Cuanta desnormalización es demasiada?

- Generar facturas y remitos para los clientes (95 %)
- Generar reportes para la gerencia (5 %)

```
{
  order_item_ID : 834838,
  order_ID: 8827,
  quantity: 3,
  cost_per_unit: 8.50,
  product :
    {
      product_description: "1 package laser printer
        paper. 100% recycled.",
      product_name : "Eco-friendly Printer Paper",
      product_category : "office supplies",
      list_price : 9.00
    }
}
```

```
{
  product_description: "1 package laser printer paper.
    100% recycled.",
  product_name : "Eco-friendly Printer Paper",
  product_category : 'office supplies',
  list_price : 9.00
}
```


Desnormalización

¿Cuanta desnormalización es demasiada?

- Generar facturas y remitos para los clientes (95 %)
- Generar reportes para la gerencia (5 %)

```
{  
  order_item_ID : 834838,  
  order_ID: 8827,  
  quantity: 3,  
  cost_per_unit: 8.50,  
  product :  
    {  
      product_description: "1 package laser printer  
        paper. 100% recycled.",  
      product_name : "Eco-friendly Printer Paper",  
      product_category : "office supplies",  
      list_price : 9.00  
    }  
}
```

```
{  
  product_description: "1 package laser printer paper.  
    100% recycled.",  
  product_name : "Eco-friendly Printer Paper",  
  product_category : 'office supplies',  
  list_price : 9.00  
}
```



```
{  
  order_item_ID : 834838,  
  order_ID: 8827,  
  quantity: 3,  
  cost_per_unit: 8.50,  
  product_name : "Eco-friendly Printer Paper"  
}
```

Diseño Físico


Documentos mutables

```
{  
  truck_id: 'T87V12',  
  time: '08:10:00',  
  date : '27-May-2015',  
  driver_name: 'Jane Washington',  
  fuel_consumption_rate: '14.8 mpg',  
  ...  
}
```


Diseño Físico

Documentos mutables

```
{  
  truck_id: 'T87V12',  
  time: '08:10:00',  
  date : '27-May-2015',  
  driver_name: 'Jane Washington',  
  fuel_consumption_rate: '14.8 mpg',  
  ...  
}
```



```
{  
  truck_id: 'T87V12',  
  date : '27-May-2015',  
  driver_name: 'Jane Washington',  
  operational_data:  
    [  
      {time : '00:01',  
        fuel_consumption_rate: '14.8 mpg',  
        ...},  
      {time : '00:04',  
        fuel_consumption_rate: '12.2 mpg',  
        ...},  
      {time : '00:07',  
        fuel_consumption_rate: '15.1 mpg',  
        ...},  
      ...]  
    ]  
}
```

Diseño Físico

Documentos mutables

```
{  
  truck_id: 'T87V12',  
  time: '08:10:00',  
  date : '27-May-2015',  
  driver_name: 'Jane Washington',  
  fuel_consumption_rate: '14.8 mpg',  
  ...  
}
```

```
{  
  truck_id: 'T87V12',  
  date : '27-May-2015',  
  driver_name: 'Jane Washington',  
  operational_data:  
    [  
      {time : '00:01',  
        fuel_consumption_rate: '14.8 mpg',  
        ...},  
      {time : '00:04',  
        fuel_consumption_rate: '12.2 mpg',  
        ...},  
      {time : '00:07',  
        fuel_consumption_rate: '15.1 mpg',  
        ...},  
      ...]  
    ]  
}
```

200 Embedded Documents with Default Values

```
{truck_id: 'T8V12'  
  date: '27-May-2015'  
  operational_data:  
    [[{time: '00 : 00',  
        fuel_consumption_rate: 0.0}  
      {time: '00 : 00',  
        fuel_consumption_rate: 0.0}  
      .  
      .  
      .  
      {time: '00 : 00',  
        fuel_consumption_rate: 0.0}  
    ]  
}
```

Considerar el ciclo de vida

Modelo Conceptual -> DID -> Documentos

- DER - Modelo conceptual de alto nivel.
- DID (Modelo/Diagrama de Interrelación de Documentos).
- JSON Schema: especificación de la estructura de los documentos.

¿Cómo resolvemos la interrelación entre documentos?

- DER - Modelo conceptual de alto nivel.
- DID (Modelo/Diagrama de Interrelación de Documentos).
- JSON Schema: especificación de la estructura de los documentos.

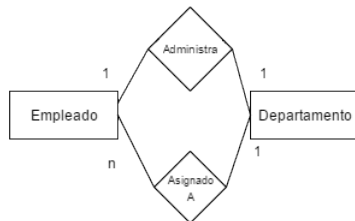
¿Cómo resolvemos la interrelación entre documentos?

Incrustar o Referenciar

La decisión más importantes es si incrustar o referenciar, lo que determinará el grado de desnormalización de los documentos

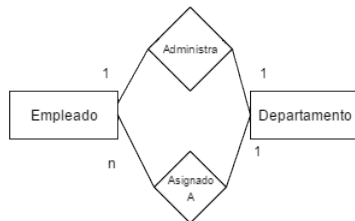
Cardinalidad 1 a N / 1 a 1

Se omiten los atributos por razones didácticas



Cardinalidad 1 a N / 1 a 1

Se omiten los atributos por razones didácticas



- Incrustar el departamento en el empleado
- Incrustar los empleados en el departamento
- Referenciar los empleados e incrustar el departamento en empleado.
- Referenciar de ambos lados
- Incrustar de ambos lados
- etc, etc...

Cardinalidad 1 a N / 1 a 1

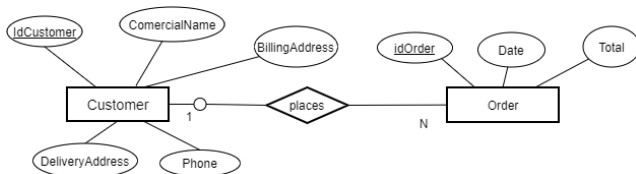
¿Que es Referenciar?

En un documento se hace referencia a un ID o una lista de ID de otro documento

Cardinalidad 1 a N / 1 a 1

¿Que es Referenciar?

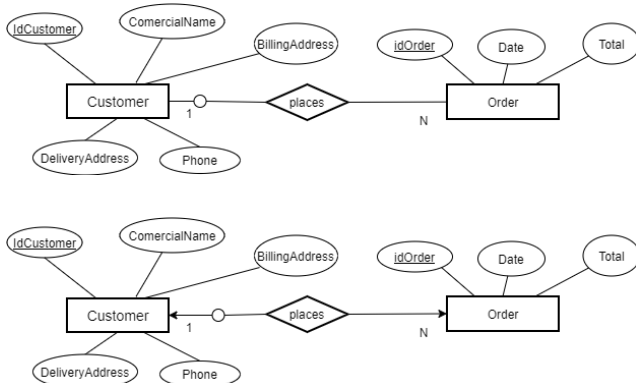
En un documento se hace referencia a un ID o una lista de ID de otro documento



Cardinalidad 1 a N / 1 a 1

¿Que es Referenciar?

En un documento se hace referencia a un ID o una lista de ID de otro documento



Cardinalidad 1 a N / 1 a 1

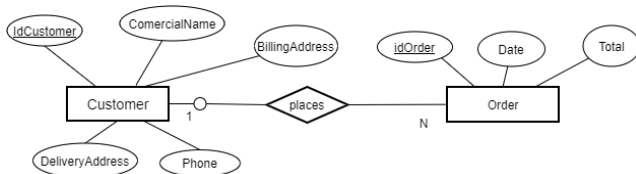
¿Que es Incrustar?

En un documento se incluyen todos los datos (en principio) de otro documento

Cardinalidad 1 a N / 1 a 1

¿Que es Incrustar?

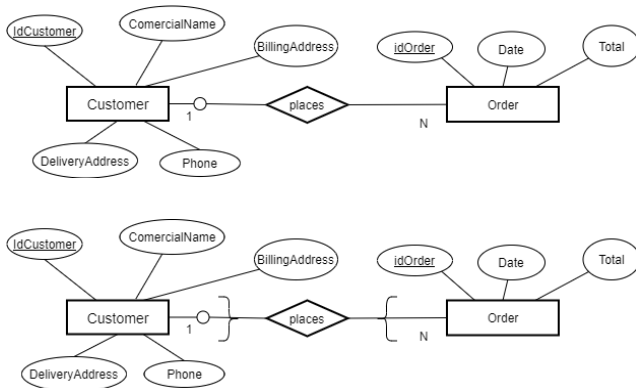
En un documento se incluyen todos los datos (en principio) de otro documento



Cardinalidad 1 a N / 1 a 1

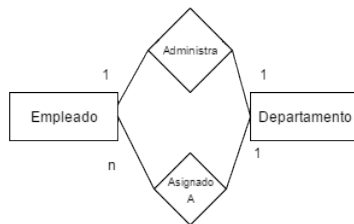
¿Que es Incrustar?

En un documento se incluyen todos los datos (en principio) de otro documento



Cardinalidad 1 a N / 1 a 1

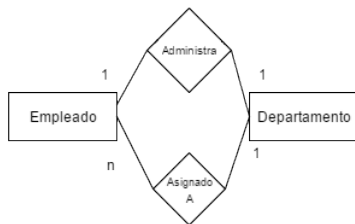
Se omiten los atributos por razones didácticas



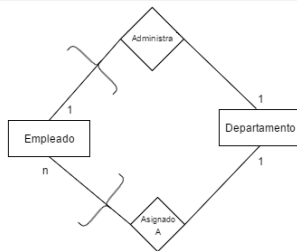
Modelo conceptual: DER

Cardinalidad 1 a N / 1 a 1

Se omiten los atributos por razones didácticas



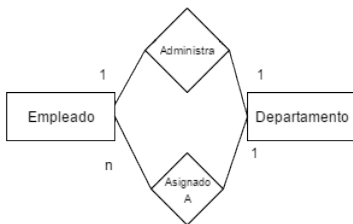
Modelo conceptual: DER



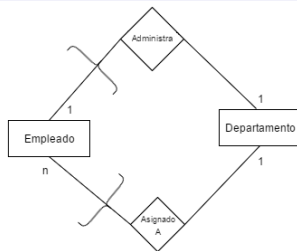
DID: Alternativa 1 Todo Incrustado en Depto

Cardinalidad 1 a N / 1 a 1

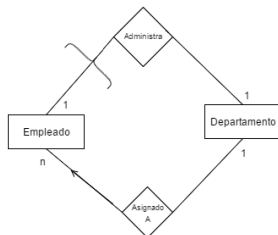
Se omiten los atributos por razones didácticas



Modelo conceptual: DER



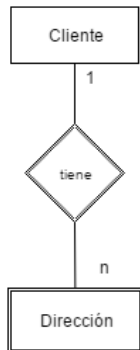
DID: Alternativa 1 Todo Incrustado en Depto



DID: Alternativa 2 - Incrustar sólo Gerente.

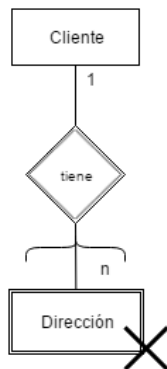
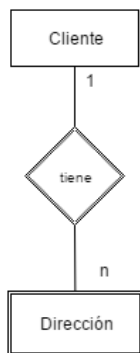
Entidades débiles

Se omiten los atributos por razones didacticas



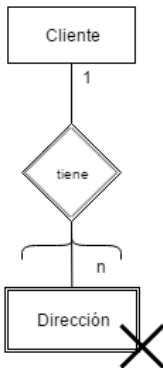
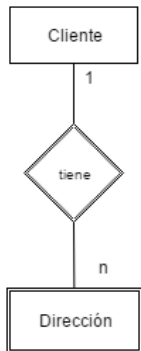
Entidades débiles

Se omiten los atributos por razones didacticas



Entidades débiles

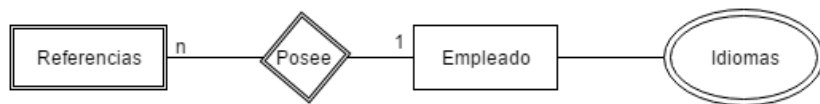
Se omiten los atributos por razones didacticas



```
{
  {
    "cliente_id": 76123,
    "nombre": "Acme Data Modeling
    Services",
    "tipo_de_cliente": "business",
    "direcciones" :
      [
        {calle: "San Martin 2222",
        ciudad: "Caseros",
        provincia: "Buenos Aires",
        codigo_postal: 99076} ,
        {calle: "9 de Julio 2223",
        ciudad: "CABA",
        codigo_postal: 01097}
      ]
  }
}
```

Entidades Débiles - Atributos Multivaluados

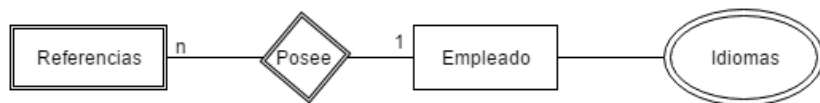
Se omiten los atributos por razones didácticas



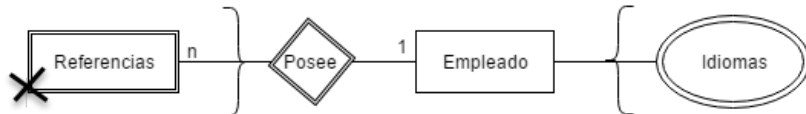
DER. Empleados Idiomas y Referencias

Entidades Débiles - Atributos Multivaluados

Se omiten los atributos por razones didácticas



DER. Empleados Idiomas y Referencias



DID. Empleados Idiomas y Referencias

Cardinalidad M a N

Se omiten los atributos por razones didácticas



DER. M a N

Cardinalidad M a N

Se omiten los atributos por razones didácticas



DER. M a N



DID. M a N con referencias

Cardinalidad M a N

Se omiten los atributos por razones didácticas



DER. M a N



DID. M a N con referencias

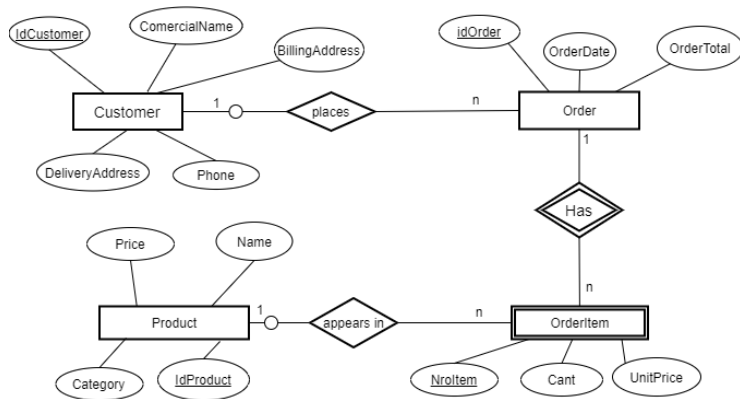
```

{
  {
    courseID: 'C1667',
    title: 'Introduction to Anthropology',
    instructor: 'Dr. Margret Austin',
    credits: 3,
    enrolledStudents: ['S1837', 'S3737', 'S9825' ...
                      'S1847'] },
  {
    courseID: 'C2873',
    title: 'Algorithms and Data Structures',
    instructor: 'Dr. Susan Johnson',
    credits: 3,
    enrolledStudents: ['S1837', 'S3737', 'S4321', 'S9825'
                      ... 'S1847'] },
  {
    courseID: 'C3876',
    title: 'Macroeconomics',
    instructor: 'Dr. James Schulen',
    credits: 3,
    enrolledStudents: ['S1837', 'S4321', 'S1470', 'S9825'
                      ... 'S1847'] },
  ...
}
    
```

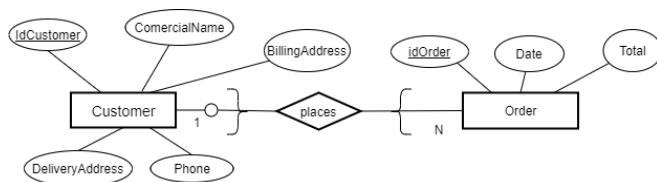
```

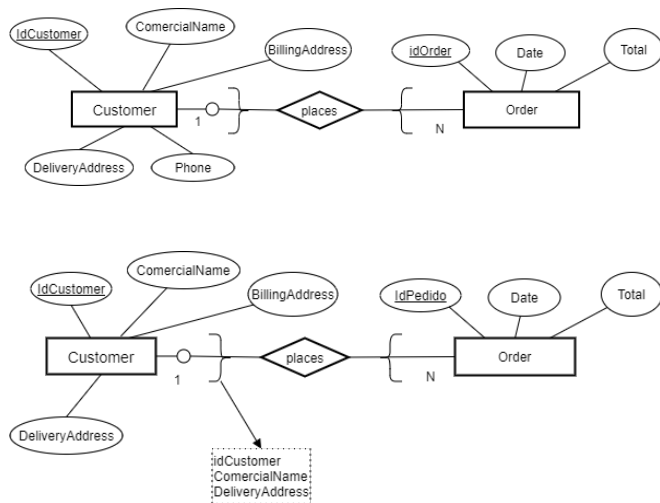
{
  {
    studentID: 'S1837',
    name: 'Brian Nelson',
    gradYear: 2018,
    courses: ['C1667', 'C2873', 'C3876'] },
  {
    studentID: 'S3737',
    name: 'Yolanda Deltor',
    gradYear: 2017,
    courses: ['C1667', 'C2873'] },
  ...
}
    
```


Desnormalización parcial



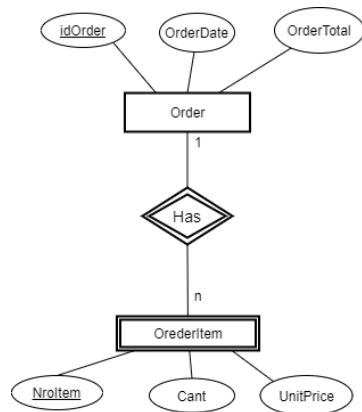
DER





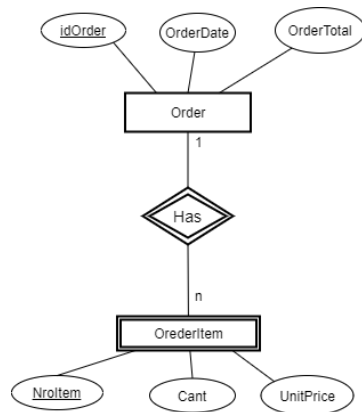
DID con desnormalización parcial

Desnormalización parcial

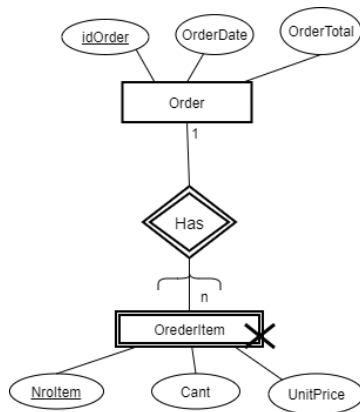


DID con desnormalización parcial

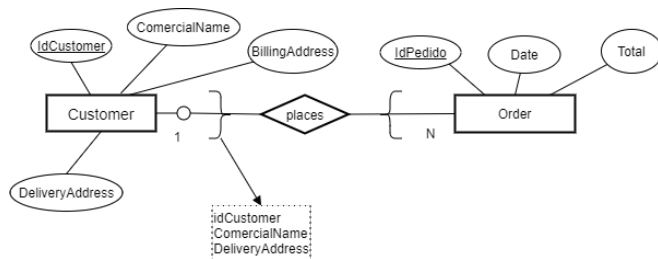
Desnormalización parcial

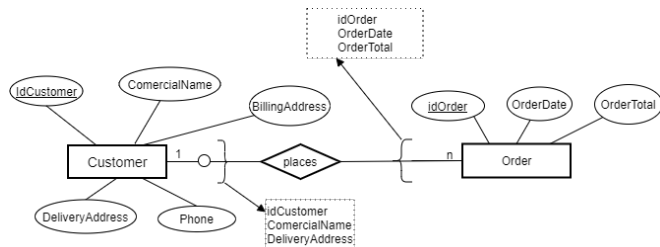
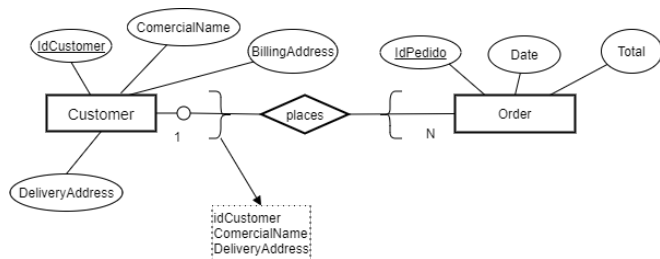


DID con desnormalización parcial

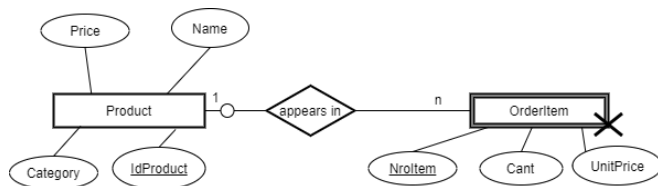


DID con desnormalización parcial

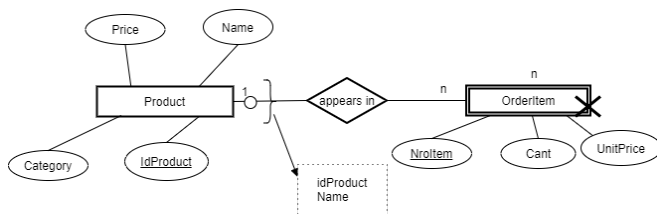
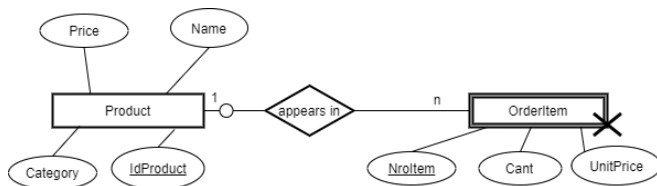




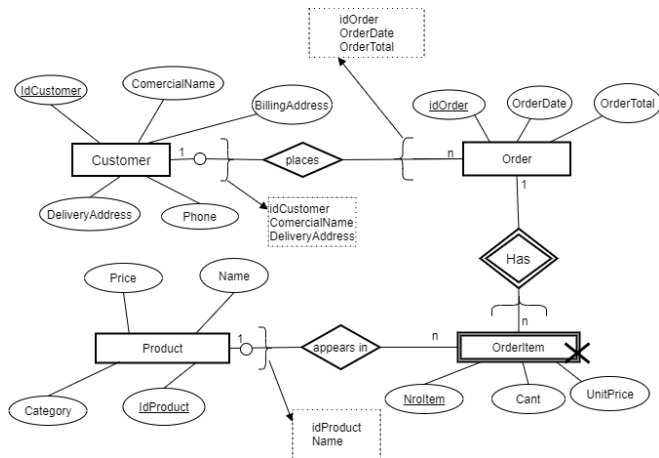
Producto-OrderItem



Producto-OrderItem



DID COMPLETO



DID con desnormalización parcial

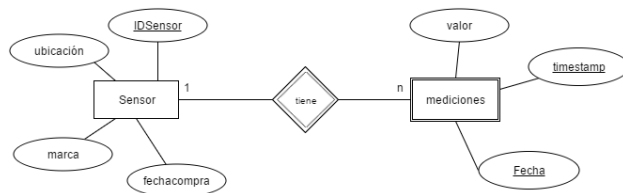
JSON Schema para Documento Orden

```

"Order": { "type": "object",
  "properties": {
    "idOrder": { "type": "integer" },
    "Date": { "type": "string", "format": "date-time" },
    "Total": { "type": "integer" },
    "Customer": {
      "type": "object",
      "properties": {
        "idCustomer": { "type": "integer" },
        "ComercialName": { "type": "string" },
        "DeliveryAddress": { "type": "string" }
      }
    },
    "OrderItem": {
      "type": "Array",
      "items": [
        {
          "type": "object",
          "properties": {
            "Cant": { "type": "integer" },
            "IdProduct": { "type": "integer" },
            "Name": { "type": "string" },
            "UnitPrice": { "type": "string" }
          }
        }
      ]
    }
  }
}

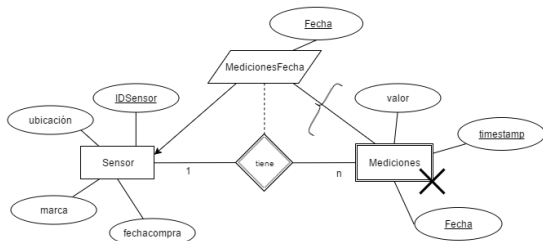
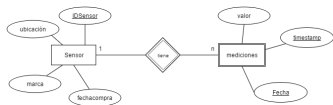
```

Uso de Documentos Auxiliares



¿Que pasa cuando la cantidad de mediciones es muy grande y además se actualiza permanentemente?

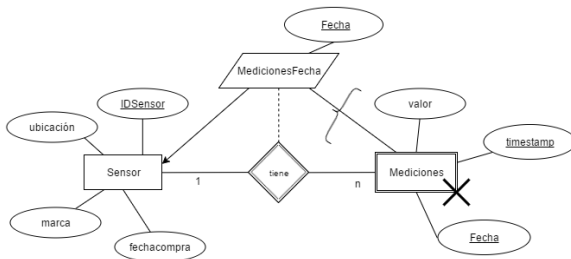
Uso de Documentos Auxiliares



Se necesita crear un tipo de documento auxiliar que permita particionar las mediciones

Uso de Documentos Auxiliares

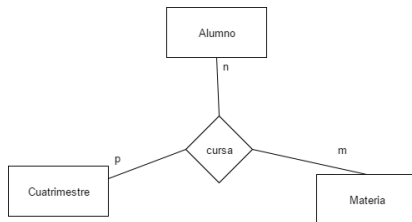
JSON Schema



```
"MedicionesFecha": {"type": "object",
"properties": {
  "IDSensor": {"type": "integer" },
  "Fecha": {"type": "format": "date-time"},
  "Mediciones": {"type": "array",
    "items": {"type": "object",
      "properties": {"timestamp": {"type": "string", "format": "date-time"}, "valor": {"type": "decimal"}}}
  }
}
```

Interrelaciones Ternarias

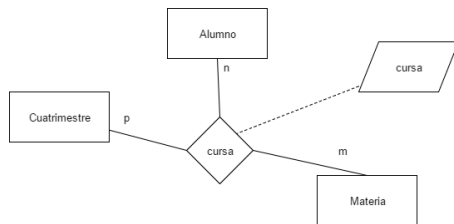
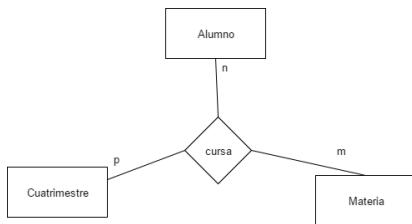
Se omiten los atributos por razones didácticas



¿Como resolvemos la interrelación *curso*?

Interrelaciones Ternarias

Se omiten los atributos por razones didácticas

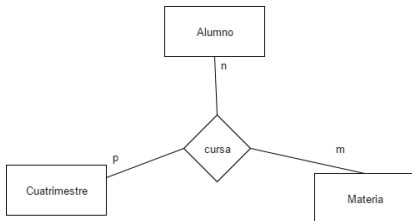


Opción básica:

Se genera un documento con las claves de cada uno

Interrelaciones Ternarias

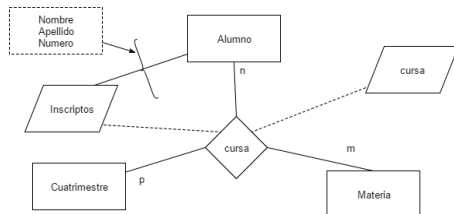
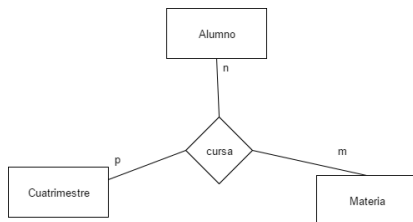
Se omiten los atributos por razones didácticas



Supongamos

Una consulta muy común es saber cuales son los alumnos anotados en una materia en un cuatrimestre

Se omiten los atributos por razones didácticas

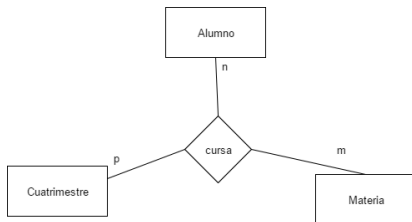


Supongamos

Una consulta muy común es saber cuales son los alumnos anotados en una materia en un cuatrimestre

Interrelaciones Ternarias

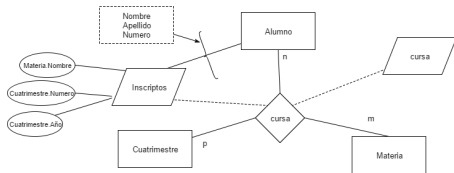
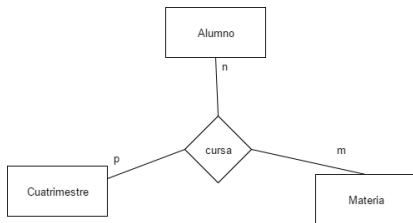
Se omiten los atributos por razones didácticas



¿Y si queremos además el nombre de la materia y el número y año del cuatrimestre?

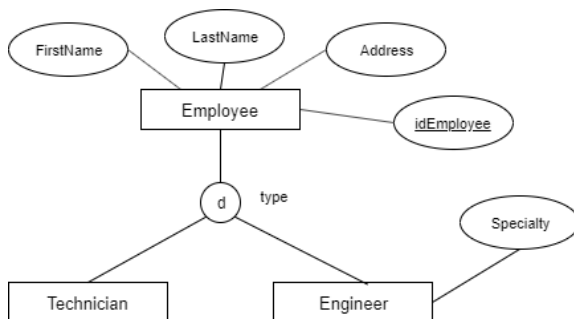
Interrelaciones Ternarias

Se omiten los atributos por razones didácticas

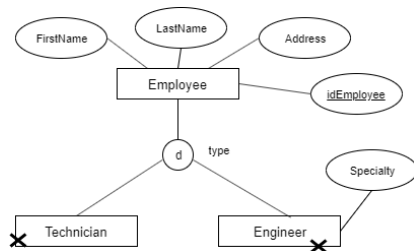
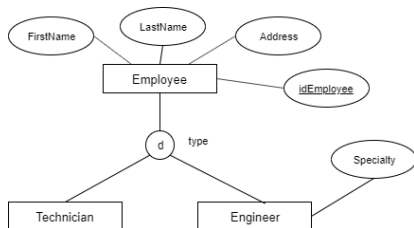


La propia semántica de la cardinalidad de la ternaria nos facilita este modelo

Jerarquías



Jerarquías



La facilidad de tener esquema flexible nos facilita el diseño. Podemos usar sólo un tipo de documentos para toda la jerarquía.

Jerarquías

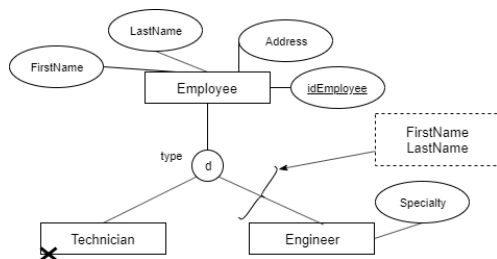
Supongamos que:

Queremos conservar la entidad *Engineer* como tipo de documento independiente porque una consulta importante es listar todos los ingenieros con sus datos.

Jerarquías

Supongamos que:

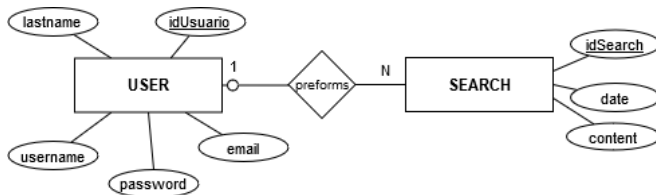
Queremos conservar la entidad *Engineer* como tipo de documento independiente porque una consulta importante es listar todos los ingenieros con sus datos.



Caso especial

Supongamos que:

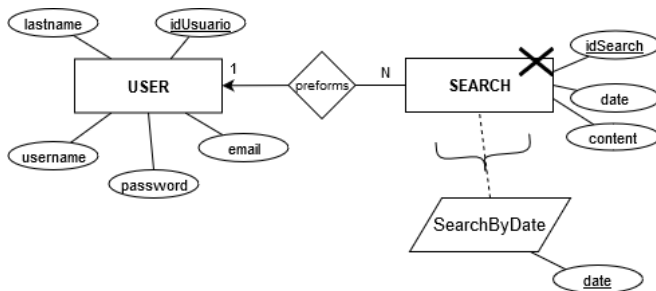
Un caso a considerar es cuando se hace necesario agrupar múltiples instancias de una entidad, por uno o más atributos, en un solo documento.



Caso especial

Supongamos que:

Un caso a considerar es cuando se hace necesario agrupar múltiples instancias de una entidad, por uno o más atributos, en un solo documento.



Bibliografía

- *NoSQL for Mere Mortals* - Dan Sullivan
- *NoSQL Distilled. A Brief Guide to the Emerging World of Polyglot Persistence* - Pramod J. Sadalage y Martin Fowler
- *A Big Data Modeling Methodology for NoSQL Document Databases* . Database Systems Journal, Vol XI, 2020. ISSN 2069 - 3230. Gerardo Rossel, Andrea Manna
- *Diseño de Bases de Datos Basadas en Documento: Modelo de Interrelación de Documentos* - Gerardo Rossel y Andrea Manna
- *MongoDB Applied Design Patterns* - Rick Copeland
- *CouchDB- The Definitive Guide* - J. Chris Anderson, Jan Lehnardt, Noah Slater
- *RavenDB in Action* - Itamar Syn-Hershko