**Information Density and Subset Ratio and its relation with different Key’s (Perfect Key, Primary Key and Key)**

**Definitions of Information Density and Subset Ratio (Source – Reference Guide):**

**Information Densit**y is the number of records that have values (i.e. not NULL) in this field as compared to the total number of records in the table.

**Subset ratio** is the number of distinct values of this field found in this table as compared to the total number of distinct values of this field (that is other tables as well).

No of unique/distinct values of a field in the table  
Subset Ratio = -----------------------------------------------------------------------------------  
 No of unique values of the field in the whole data model

Investigate wherever Information Density is less than 100% and inform the Architect about the

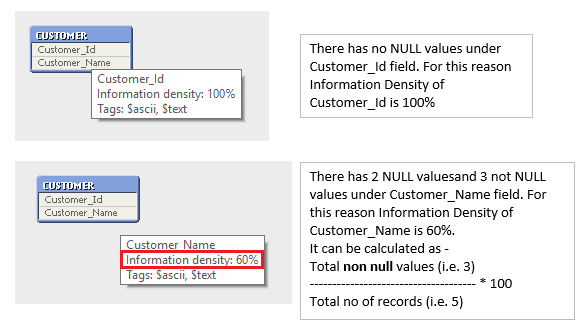
potential issue(s) with the NULL values. I would always check for Subset Ratio whenever I perform a QlikView Join. This way you know how many key field distinct values are associated to another table.

**Simple example to understand Information Density and Subset Ratio:**

Input Data:

|  |  |
| --- | --- |
| **Customer\_Id** | **Customer\_Name** |
| A | Abc |
| B |  |
| C | Ghi |
| D |  |
| E | Mno |

Output:



**Primary Key:**

Each row is unique (i.e. no repeating value present under the field).

No NULL value present.

Subset Ratio **<** 100%.

**Perfect Key**:

Each row is unique (i.e. no repeating value present under the field).

No NULL value present.

Subset Ratio **=** 100%.

In the example 2: you will notice that it's downgraded to primary key, because it does not have the

value **Inv003**, therefore now the subset ratio is 67% (which means 2 values out of the 3 existing).

**Key**:

Rows are not unique (i.e. repeated values present under the field).

NULL values may/may not present.

Subset Ratio <= 100%.

In the example 3, InvoiceID is a Key with Subset Ratio = 100%, where as in the

example 4, InvoiceID is a Key with Subset Ratio = 67%.

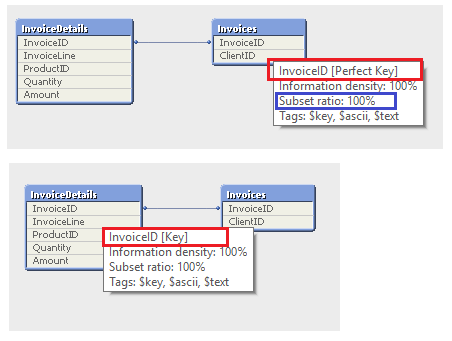
So, it’s clear that Subset Ratio has no impact to define one field as **Key** butrepeated values have.

**Example 1:**

InvoiceDetails:  
**LOAD** \* Inline   
[  
 InvoiceID, InvoiceLine, ProductID, Quantity, Amount  
 Inv001, 1, PR01, 10, 50  
 Inv001, 2, PR02, 10, 40  
 Inv002, 1, PR01, 30, 150  
 Inv002, 2, PR03, 10, 800

];  
  
Invoices:  
**LOAD** \* Inline   
[  
 InvoiceID, ClientID  
 Inv001, CL0001   
 Inv002, CL0002  
];

**Output:**

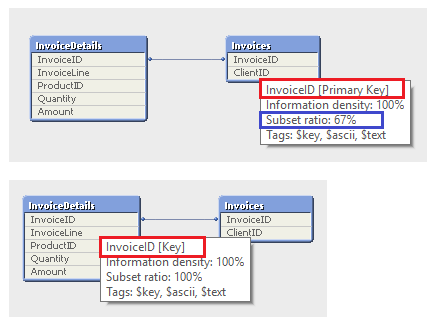


Here in the **Invoices** table, InvoiceID has 2 values which are unique and there has no NULL values and Subset Ratio is 100%. So, it’s a **Perfect Key**.

**Example 2:**

InvoiceDetails:  
**LOAD** \* Inline   
[  
 InvoiceID, InvoiceLine, ProductID, Quantity, Amount  
 Inv001, 1, PR01, 10, 50  
 Inv001, 2, PR02, 10, 40  
 Inv002, 1, PR01, 30, 150  
 Inv002, 2, PR03, 10, 800  
 Inv003, 1, PR01, 10, 50  
];  
  
Invoices:  
**LOAD** \* Inline   
[  
 InvoiceID, ClientID  
 Inv001, CL0001  
 Inv002, CL0002  
];

**Output:**



Here in the **Invoices** table, InvoiceID has 2 values which are unique and there has no NULL values but Subset Ratio is 67% as a new value ‘Inv003’ is available under the same field name in the InvoiceDetails table. So, it’s a **Primary Key**.

Subset Ratio can be calculated in the following way:

No of unique values under InvoiceID field in the Invoices table = 2 (Inv001, Inv002)  
No of unique values under InvoiceID field in the whole data model = 3 (Inv001, Inv002, Inv003)

No of unique values under InvoiceID field in the Invoices table (i.e. 2)

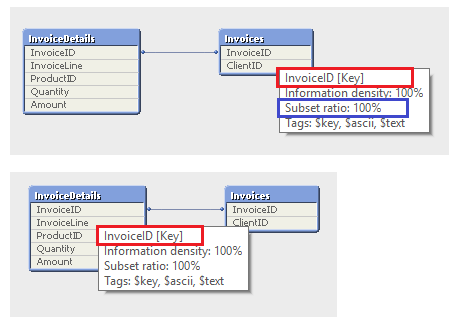
So, Subset Ratio = ---------------------------------------------------------------------------------------------------- \* 100  
 No of unique values under InvoiceID field in the whole data model (i.e. 3)

**Example 3:**

InvoiceDetails:  
**LOAD** \* Inline   
[  
 InvoiceID, InvoiceLine, ProductID, Quantity, Amount  
 Inv001, 1, PR01, 10, 50  
 Inv001, 2, PR02, 10, 40  
 Inv002, 1, PR01, 30, 150  
 Inv002, 2, PR03, 10, 800

];  
  
Invoices:  
**LOAD** \* Inline   
[  
 InvoiceID, ClientID  
 Inv001, CL0001  
 Inv001, CL0003  
 Inv002, CL0002  
];

**Output:**



Here InvoiceID of the Invoices table has repeated values (Inv001 has 2 entries). For this reason, it is defined as **Key**.

**Example 4:**

InvoiceDetails:  
**LOAD** \* Inline   
[  
 InvoiceID, InvoiceLine, ProductID, Quantity, Amount  
 Inv001, 1, PR01, 10, 50  
 Inv001, 2, PR02, 10, 40  
 Inv002, 1, PR01, 30, 150  
 Inv002, 2, PR03, 10, 800  
 Inv003, 1, PR01, 10, 50  
];  
  
Invoices:  
**LOAD** \* Inline   
[  
 InvoiceID, ClientID  
 Inv001, CL0001  
 Inv001, CL0003  
 Inv002, CL0002  
];

**Output:**

