M.DS
Aide à la Décision

MARRAKECH
جامعة القاضي عياض
UNIVERSITÉ CADI AYYAD

كليـــــــة العلوم
السملالية - مراكش
FACULTÉ DES SCIENCES
SEMLALIA - MARRAKECH

Warehouse & Reporting
Disability in India

BI Project

Es sanhaji El houcine
Elasri Abdelati

Supervised by
Pr.J.Zahir

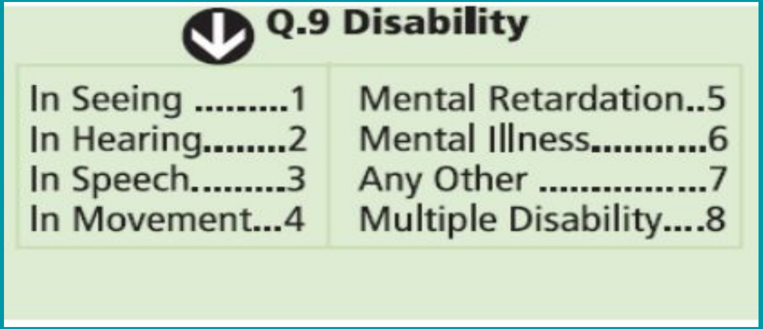
Plan

- Data Integration
  - Data Preparation
  - Data warehouse
  - ETL
- OLAP
  - ROLAP
  - Mondrian Schema
  - MDX Queries
  - MDX Demo
- Reporting Demo

Data Integration

How the data was collected ?

- Collected by Census India on 2001 & 2011.
- On 32 Indian States.
- Used Questionnaire to collect, City, Area Type, Age Group, and Total number of males/females disabled for every disability type in the Questionnaire.



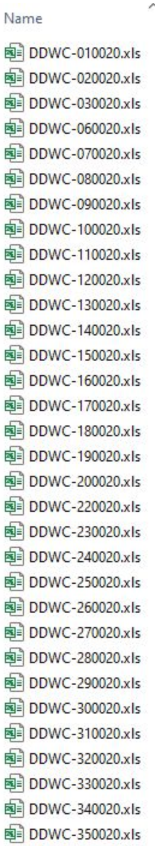
Q.9 Disability

In Seeing 1
In Hearing.........2
In Speech..........3
In Movement...4
Mental Retardation..5
Mental Illness...........6
Any Other 7
Multiple Disability....8

data.gov.in
Open Government Data (OGD) Platform India

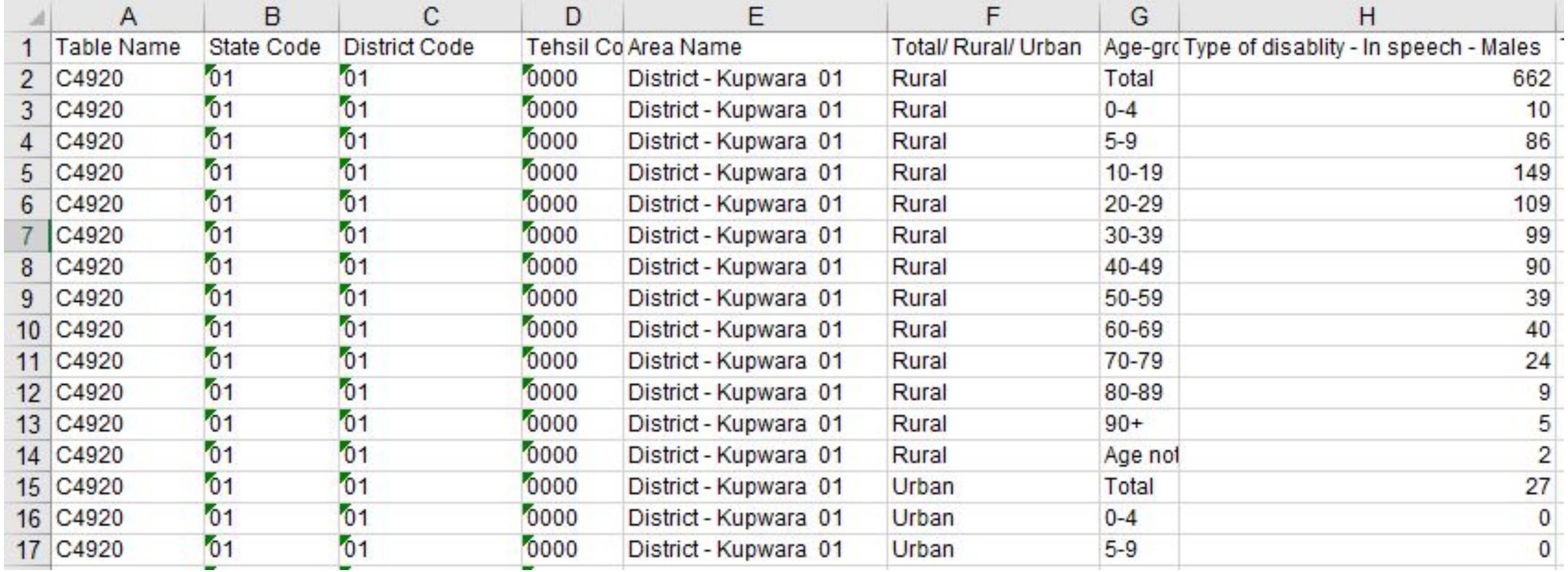
Data Format

32 States x 2



Name

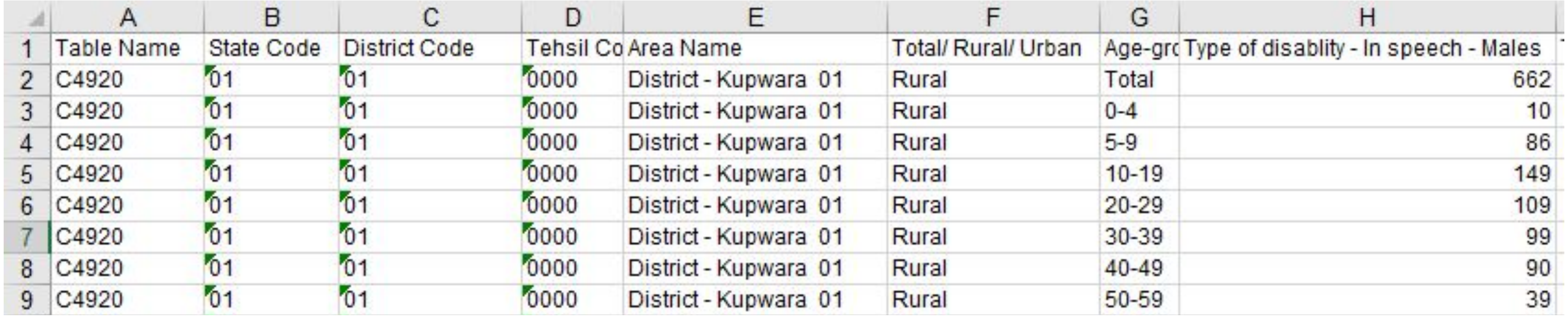
DDWC-010020.xls
DDWC-020020.xls
DDWC-030020.xls
DDWC-060020.xls
DDWC-070020.xls
DDWC-080020.xls
DDWC-090020.xls
DDWC-100020.xls
DDWC-110020.xls
DDWC-120020.xls
DDWC-130020.xls
DDWC-140020.xls
DDWC-150020.xls
DDWC-160020.xls
DDWC-170020.xls
DDWC-180020.xls
DDWC-190020.xls
DDWC-200020.xls
DDWC-220020.xls
DDWC-230020.xls
DDWC-240020.xls
DDWC-250020.xls
DDWC-260020.xls
DDWC-270020.xls
DDWC-280020.xls
DDWC-290020.xls
DDWC-300020.xls
DDWC-310020.xls
DDWC-320020.xls
DDWC-330020.xls
DDWC-340020.xls
DDWC-350020.xls



	A	B	C	D	E	F	G	H
1	Table Name	State Code	District Code	Tehsil Co	Area Name	Total/ Rural/ Urban	Age-grc	Type of disablity - In speech - Males
2	C4920	01	01	0000	District - Kupwara 01	Rural	Total	662
3	C4920	01	01	0000	District - Kupwara 01	Rural	0-4	10
4	C4920	01	01	0000	District - Kupwara 01	Rural	5-9	86
5	C4920	01	01	0000	District - Kupwara 01	Rural	10-19	149
6	C4920	01	01	0000	District - Kupwara 01	Rural	20-29	109
7	C4920	01	01	0000	District - Kupwara 01	Rural	30-39	99
8	C4920	01	01	0000	District - Kupwara 01	Rural	40-49	90
9	C4920	01	01	0000	District - Kupwara 01	Rural	50-59	39
10	C4920	01	01	0000	District - Kupwara 01	Rural	60-69	40
11	C4920	01	01	0000	District - Kupwara 01	Rural	70-79	24
12	C4920	01	01	0000	District - Kupwara 01	Rural	80-89	9
13	C4920	01	01	0000	District - Kupwara 01	Rural	90+	5
14	C4920	01	01	0000	District - Kupwara 01	Rural	Age not	2
15	C4920	01	01	0000	District - Kupwara 01	Urban	Total	27
16	C4920	01	01	0000	District - Kupwara 01	Urban	0-4	0
17	C4920	01	01	0000	District - Kupwara 01	Urban	5-9	0

Problem with this format ?

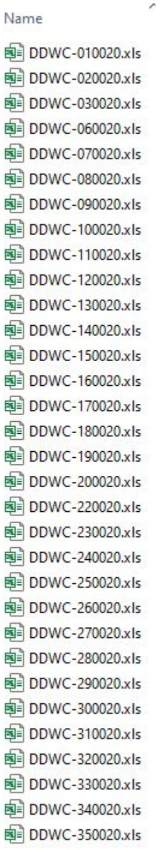
The Information of Gender, Disability Type, Disability Rate are all **condensed** in one column. Not Normalized and Would be complicated to handle with and ETL tool.



	A	B	C	D	E	F	G	H
1	Table Name	State Code	District Code	Tehsil Co	Area Name	Total/ Rural/ Urban	Age-grc	Type of disablity - In speech - Males
2	C4920	01	01	0000	District - Kupwara 01	Rural	Total	662
3	C4920	01	01	0000	District - Kupwara 01	Rural	0-4	10
4	C4920	01	01	0000	District - Kupwara 01	Rural	5-9	86
5	C4920	01	01	0000	District - Kupwara 01	Rural	10-19	149
6	C4920	01	01	0000	District - Kupwara 01	Rural	20-29	109
7	C4920	01	01	0000	District - Kupwara 01	Rural	30-39	99
8	C4920	01	01	0000	District - Kupwara 01	Rural	40-49	90
9	C4920	01	01	0000	District - Kupwara 01	Rural	50-59	39

Data Preparation

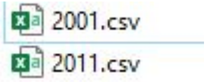
64 Files



Name

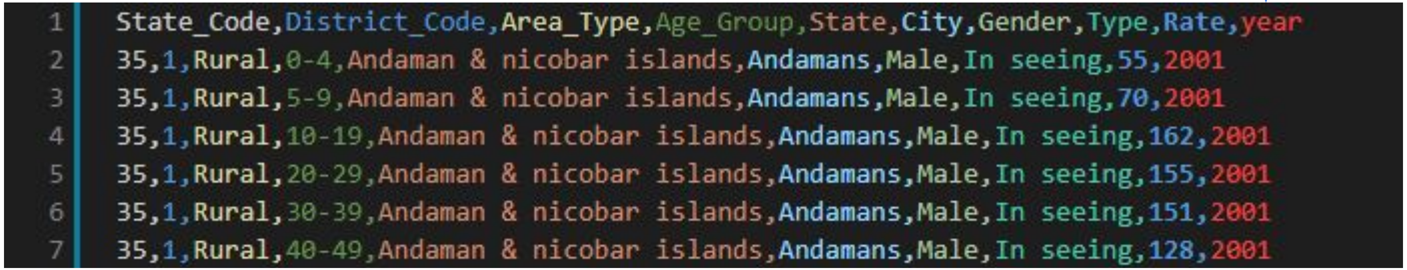
DDWC-010020.xls
DDWC-020020.xls
DDWC-030020.xls
DDWC-060020.xls
DDWC-070020.xls
DDWC-080020.xls
DDWC-090020.xls
DDWC-100020.xls
DDWC-110020.xls
DDWC-120020.xls
DDWC-130020.xls
DDWC-140020.xls
DDWC-150020.xls
DDWC-160020.xls
DDWC-170020.xls
DDWC-180020.xls
DDWC-190020.xls
DDWC-200020.xls
DDWC-220020.xls
DDWC-230020.xls
DDWC-240020.xls
DDWC-250020.xls
DDWC-260020.xls
DDWC-270020.xls
DDWC-280020.xls
DDWC-290020.xls
DDWC-300020.xls
DDWC-310020.xls
DDWC-320020.xls
DDWC-330020.xls
DDWC-340020.xls
DDWC-350020.xls

Pandas



2001.csv
2011.csv

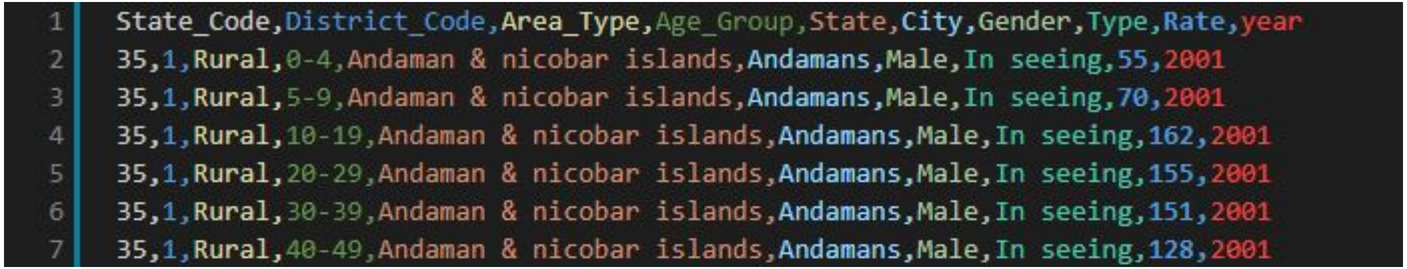
Prepared Data



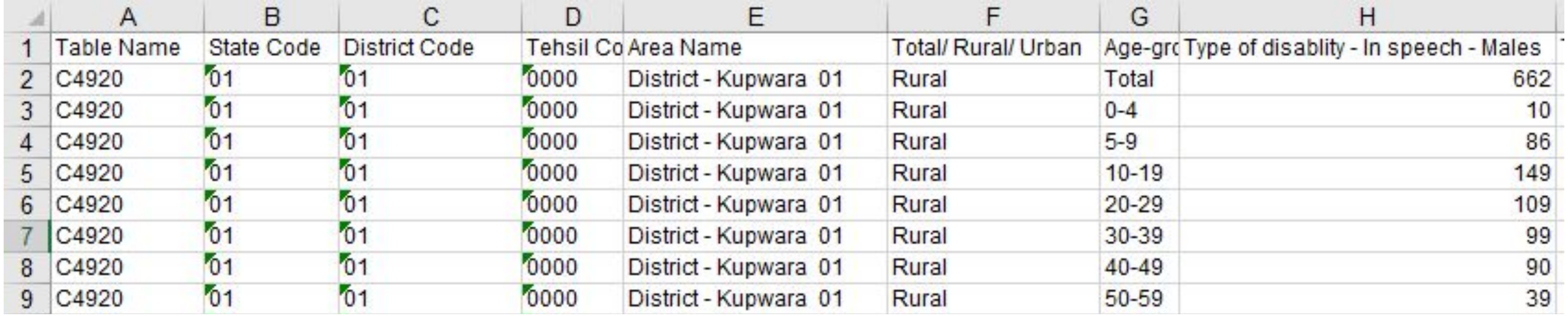
```
State_Code,District_Code,Area_Type,Age_Group,State,City,Gender,Type,Rate,year
35,1,Rural,0-4,Andaman & nicobar islands,Andamans,Male,In seeing,55,2001
35,1,Rural,5-9,Andaman & nicobar islands,Andamans,Male,In seeing,70,2001
35,1,Rural,10-19,Andaman & nicobar islands,Andamans,Male,In seeing,162,2001
35,1,Rural,20-29,Andaman & nicobar islands,Andamans,Male,In seeing,155,2001
35,1,Rural,30-39,Andaman & nicobar islands,Andamans,Male,In seeing,151,2001
35,1,Rural,40-49,Andaman & nicobar islands,Andamans,Male,In seeing,128,2001
```

Normalization

We Used Python, Pandas Package and some String Operations.



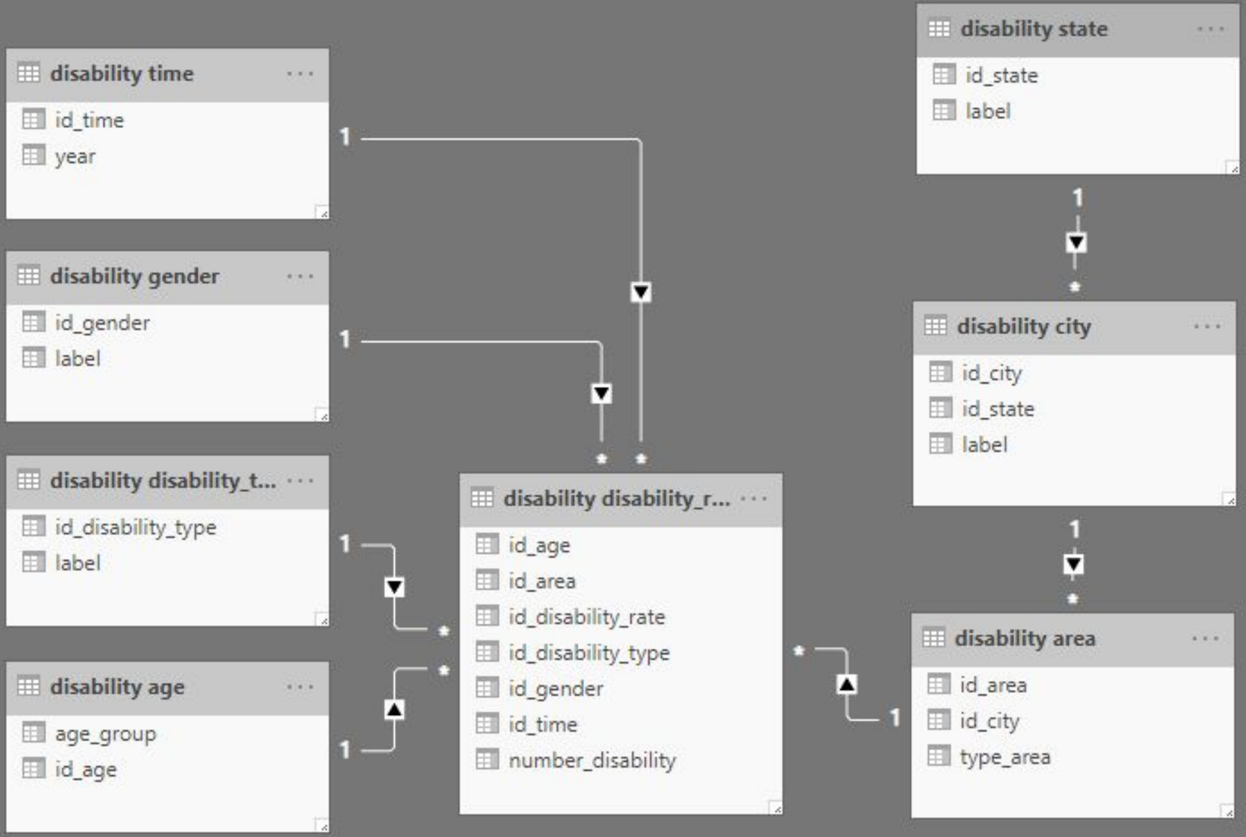
```
State_Code,District_Code,Area_Type,Age_Group,State,City,Gender,Type,Rate,year
35,1,Rural,0-4,Andaman & nicobar islands,Andamans,Male,In seeing,55,2001
35,1,Rural,5-9,Andaman & nicobar islands,Andamans,Male,In seeing,70,2001
35,1,Rural,10-19,Andaman & nicobar islands,Andamans,Male,In seeing,162,2001
35,1,Rural,20-29,Andaman & nicobar islands,Andamans,Male,In seeing,155,2001
35,1,Rural,30-39,Andaman & nicobar islands,Andamans,Male,In seeing,151,2001
35,1,Rural,40-49,Andaman & nicobar islands,Andamans,Male,In seeing,128,2001
```



	A	B	C	D	E	F	G	H
1	Table Name	State Code	District Code	Tehsil Co	Area Name	Total/ Rural/ Urban	Age-grc	Type of disablity - In speech - Males
2	C4920	01	01	0000	District - Kupwara 01	Rural	Total	662
3	C4920	01	01	0000	District - Kupwara 01	Rural	0-4	10
4	C4920	01	01	0000	District - Kupwara 01	Rural	5-9	86
5	C4920	01	01	0000	District - Kupwara 01	Rural	10-19	149
6	C4920	01	01	0000	District - Kupwara 01	Rural	20-29	109
7	C4920	01	01	0000	District - Kupwara 01	Rural	30-39	99
8	C4920	01	01	0000	District - Kupwara 01	Rural	40-49	90
9	C4920	01	01	0000	District - Kupwara 01	Rural	50-59	39

Data Warehouse

Disability Warehouse SnowFlake Schema



disability time

- id_time
- year

disability gender

- id_gender
- label

disability disability_t...

- id_disability_type
- label

disability age

- age_group
- id_age

disability disability_r...

- id_age
- id_area
- id_disability_rate
- id_disability_type
- id_gender
- id_time
- number_disability

disability state

- id_state
- label

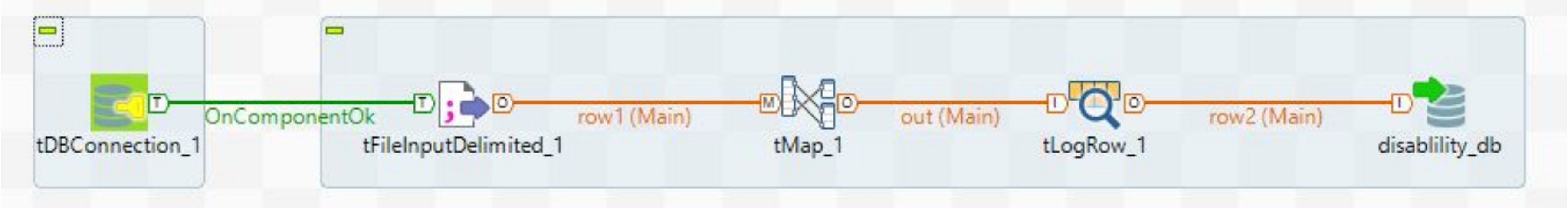
disability city

- id_city
- id_state
- label

disability area

- id_area
- id_city
- type_area

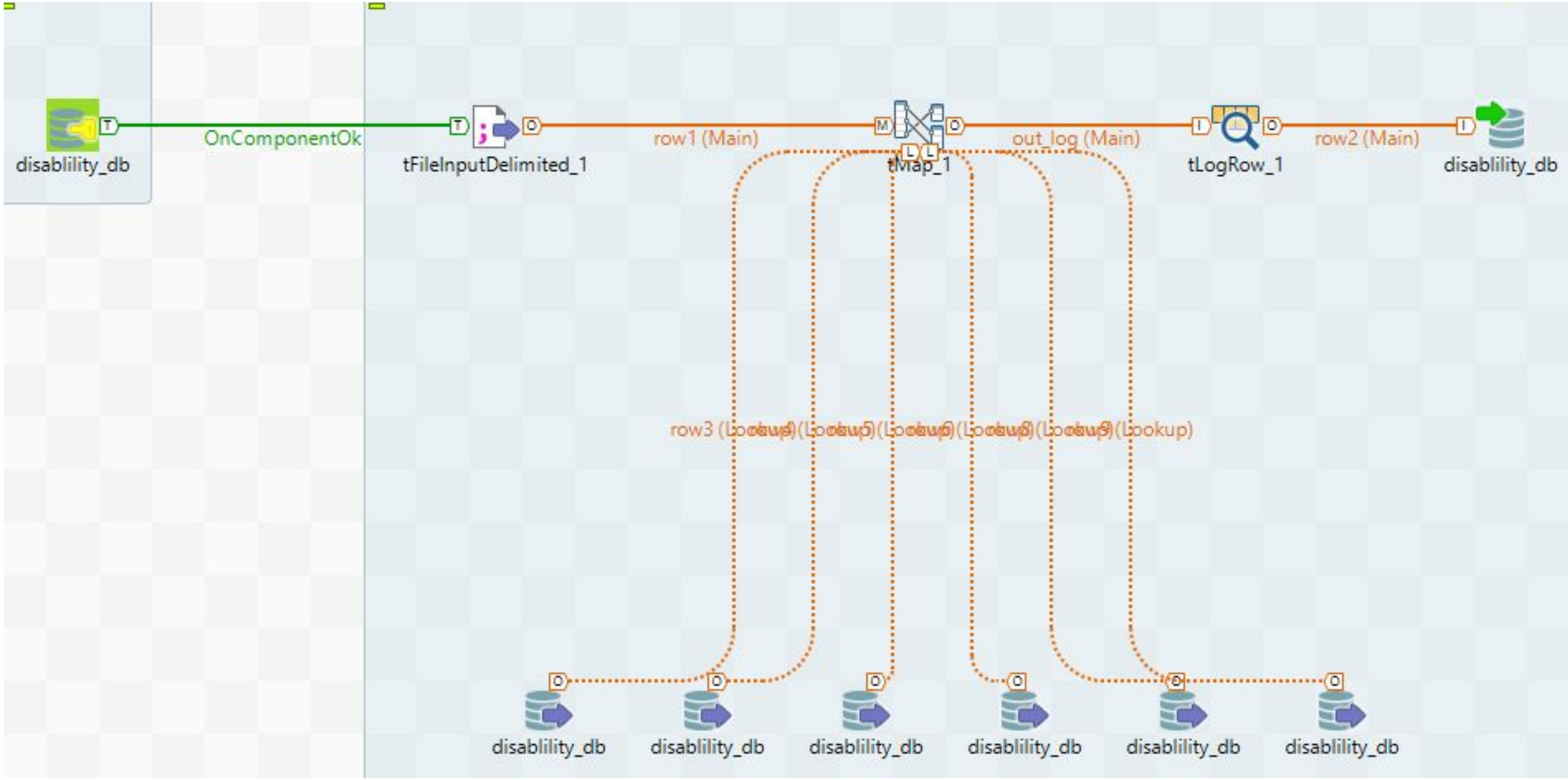
Talend - Extract Transform Load - Dimensions



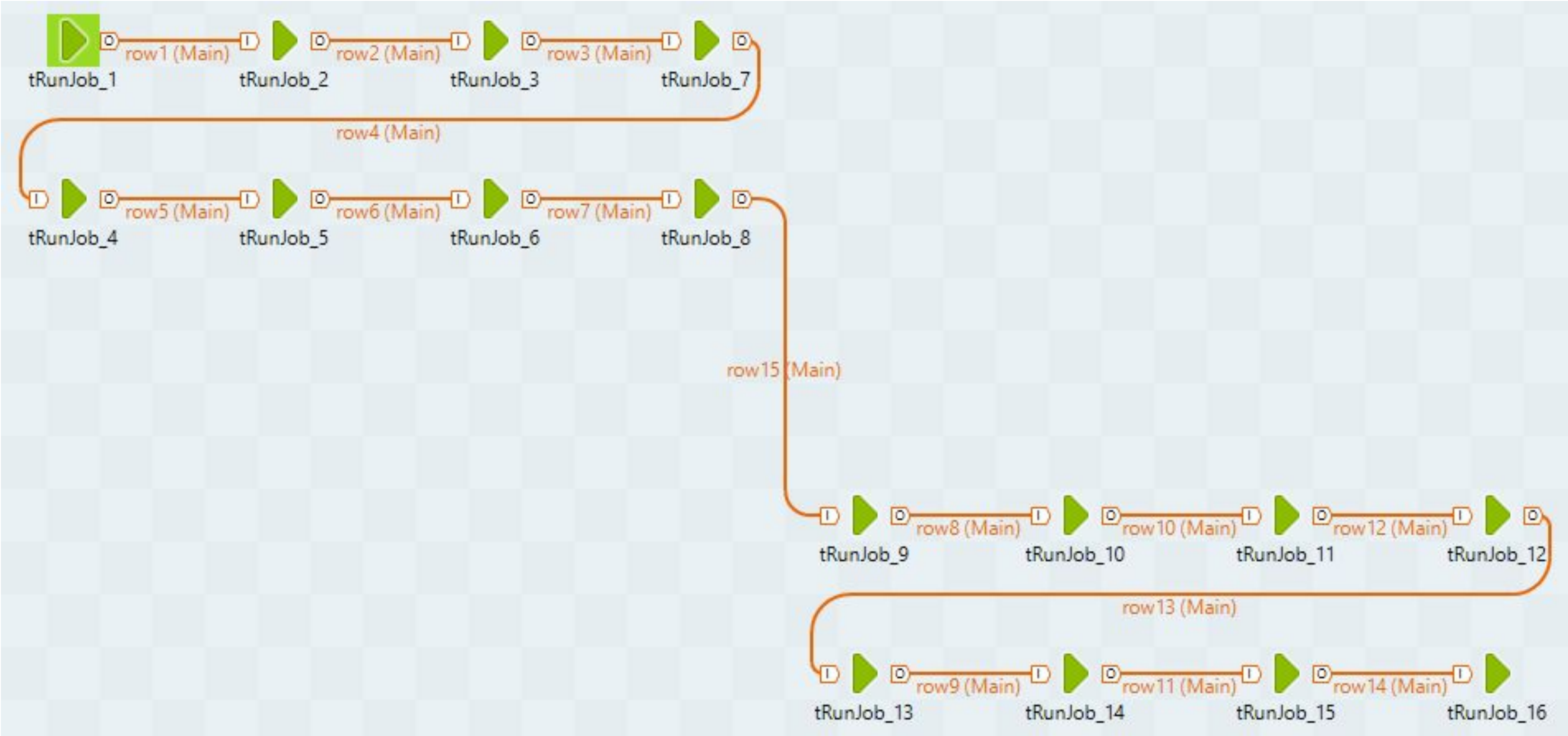
We have 1 job for each one of the 7 dimensions, similar the one above:

1. Connecting to Data warehouse.
2. Getting the csv data file.
3. Mapping the columns from csv file to the corresponding database table column.
4. Tlog visualizing output rows.
5. Insert into data warehouse dimension table

Talend - Extract Transform Load - Fact

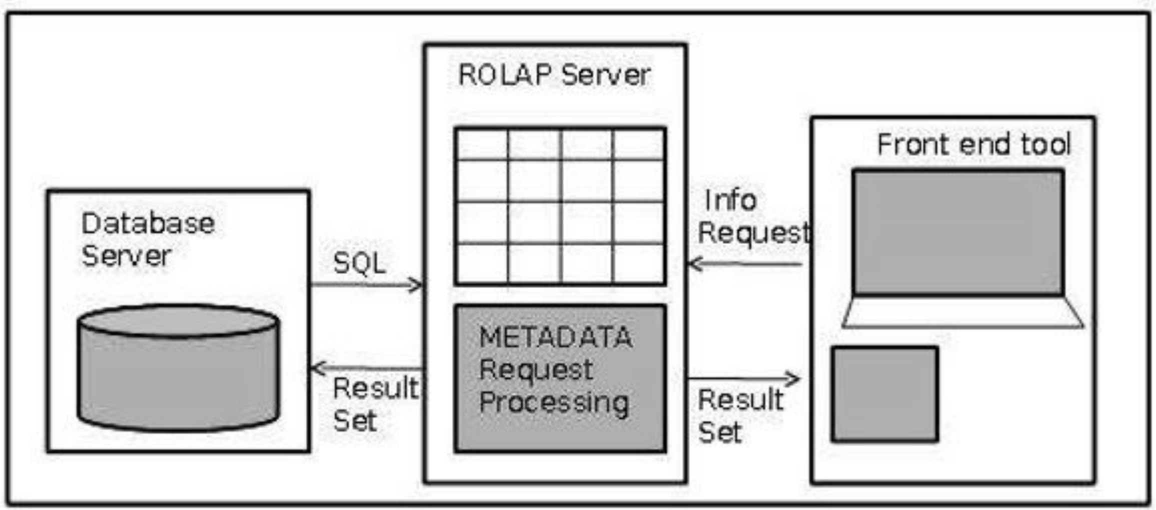


Talend - Extract Transform Load - Global Job



OnLine Analysis Server
OLAP

ROLAP

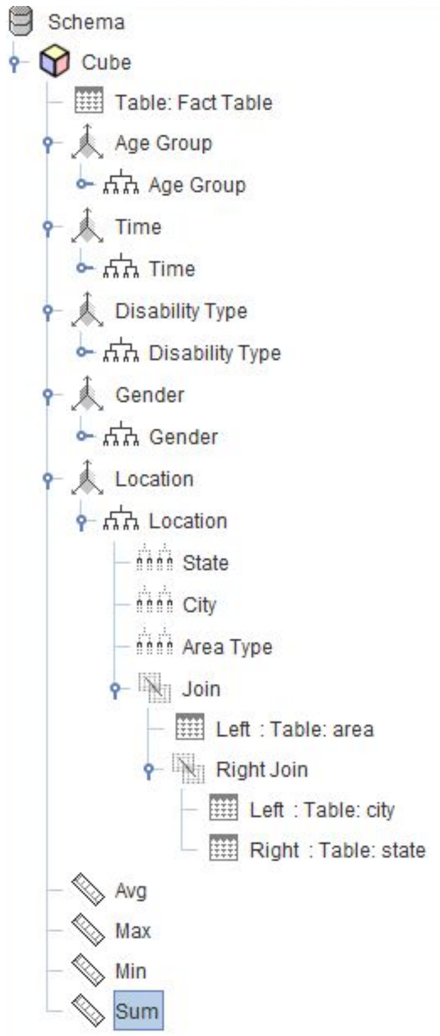


Technique to query relational database using cube-like queries.

Using XML Mondrian Schemas

ROLAP - Mondrian Schema

Encountered Problem: Snowflake Schema when dimension has 3 levels it requires 2 join tables.

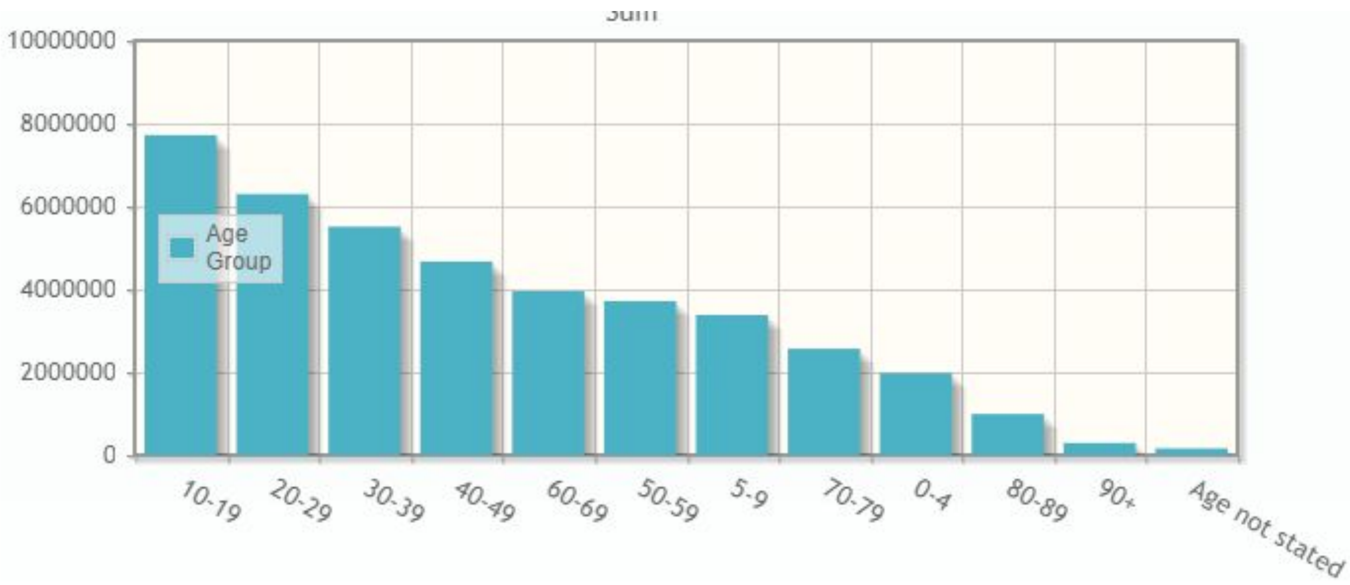


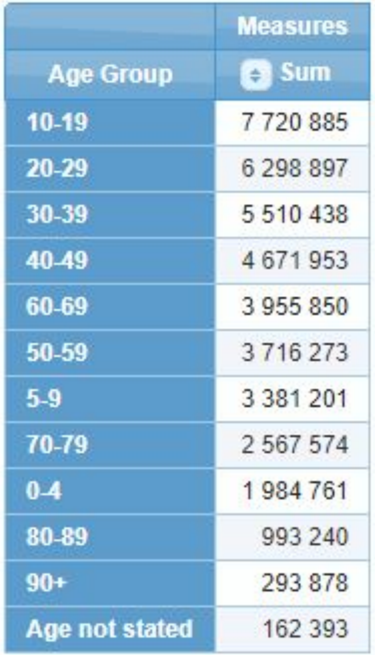
- Schema
  - Cube
    - Table: Fact Table
    - Age Group
      - Age Group
    - Time
      - Time
    - Disability Type
      - Disability Type
    - Gender
      - Gender
    - Location
      - Location
        - State
        - City
        - Area Type
        - Join
          - Left : Table: area
          - Right Join
            - Left : Table: city
            - Right : Table: state
    - Avg
    - Max
    - Min
    - Sum

MDX Queries
Pentaho

What age group has the most and the least disabled people?

```
SELECT {[Measures].[Sum]} ON COLUMNS,
Order({[Age Group].children}, [Measures].[Sum], DESC) ON ROWS
FROM [Cube]
```

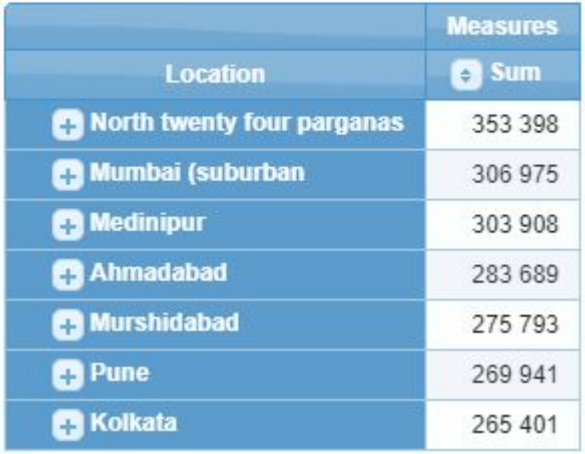




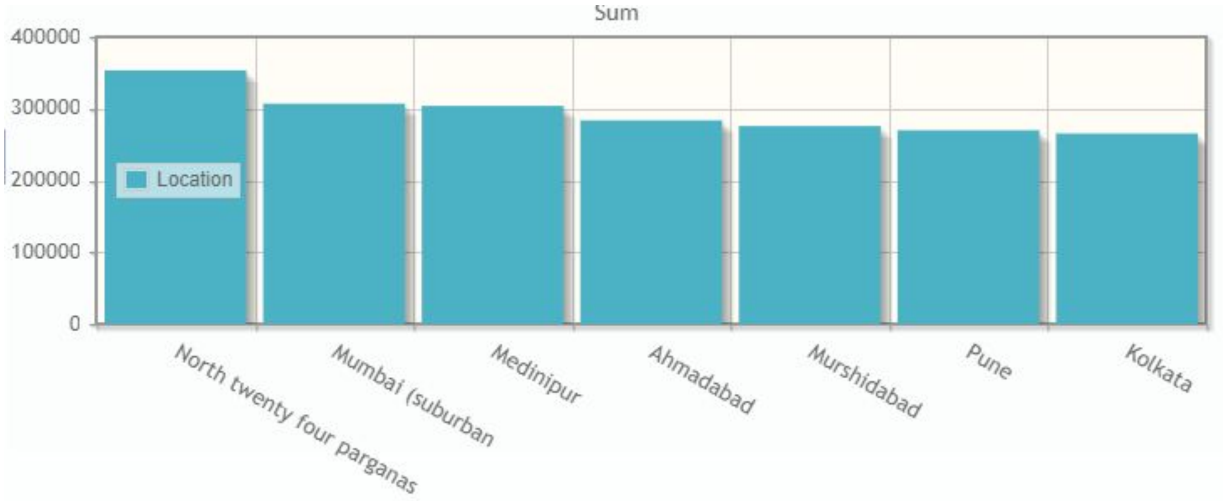
	Measures
Age Group	Sum
10-19	7 720 885
20-29	6 298 897
30-39	5 510 438
40-49	4 671 953
60-69	3 955 850
50-59	3 716 273
5-9	3 381 201
70-79	2 567 574
0-4	1 984 761
80-89	993 240
90+	293 878
Age not stated	162 393

what are the top 7 Indian cities with disabled people?

```
SELECT {[Measures].[Sum]} ON COLUMNS,
TOPCOUNT(Descendants([Location], 2.0), 7.0, [Measures].[Sum]) ON ROWS
FROM [Cube]
```

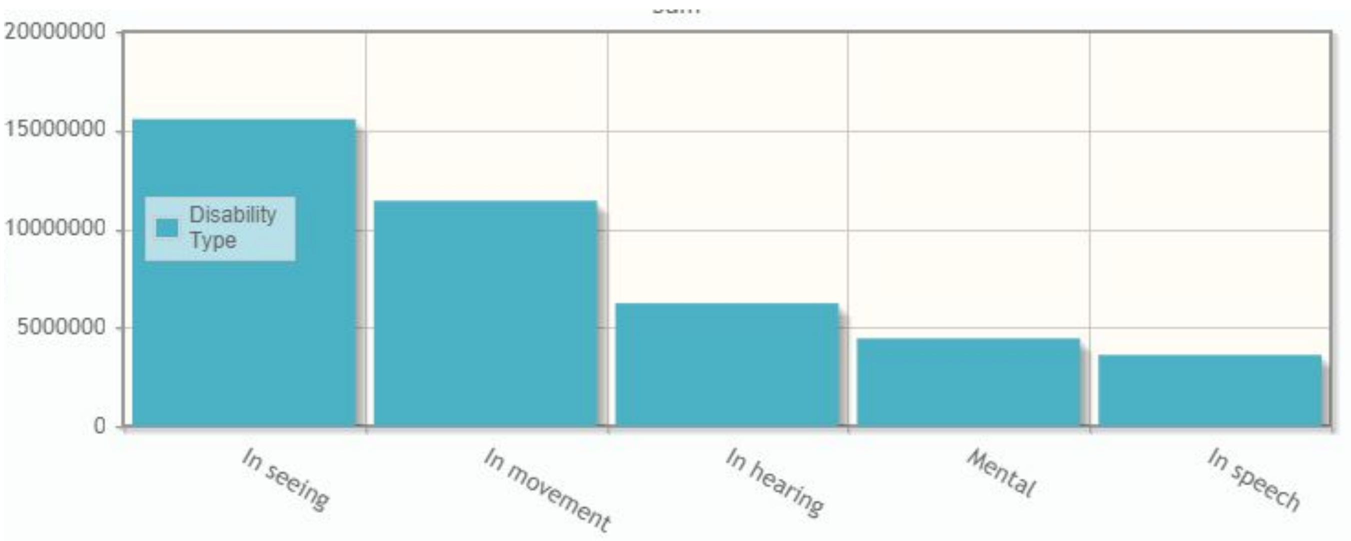


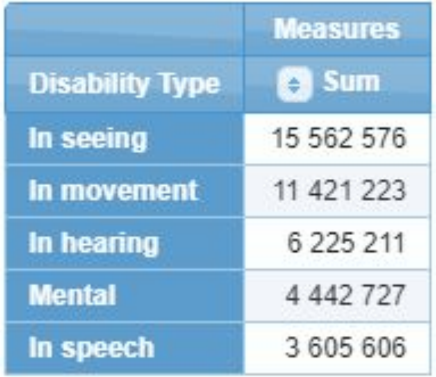
	Measures
Location	Sum
⊞ North twenty four parganas	353 398
⊞ Mumbai (suburban	306 975
⊞ Medinipur	303 908
⊞ Ahmadabad	283 689
⊞ Murshidabad	275 793
⊞ Pune	269 941
⊞ Kolkata	265 401



What is the most occurring disability type in India?

```
SELECT {[Measures].[Sum]} ON COLUMNS,
Order({[Disability Type].children}, [Measures].[Sum], DESC) ON ROWS
FROM [Cube]
```

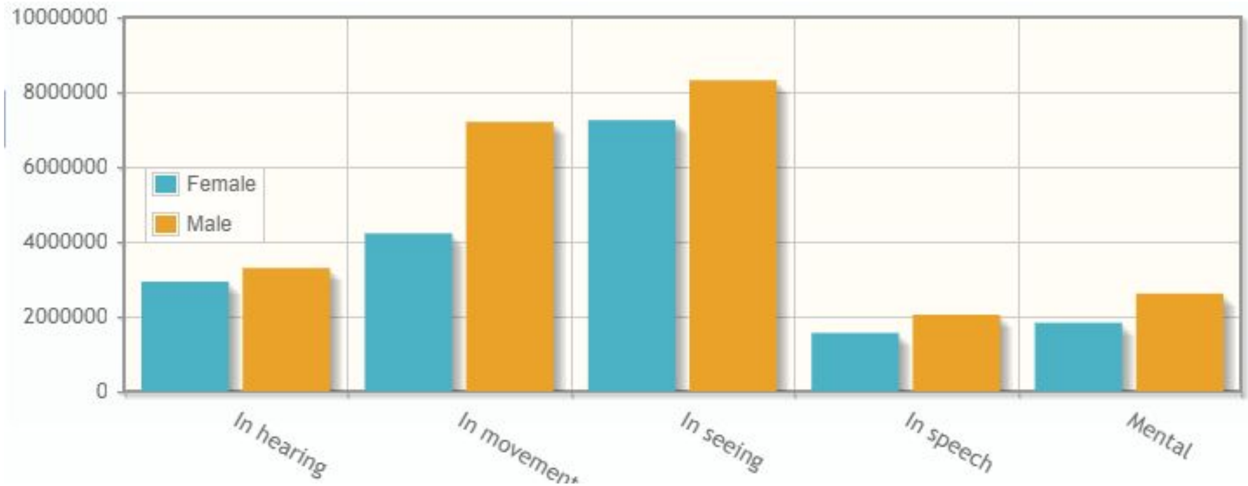


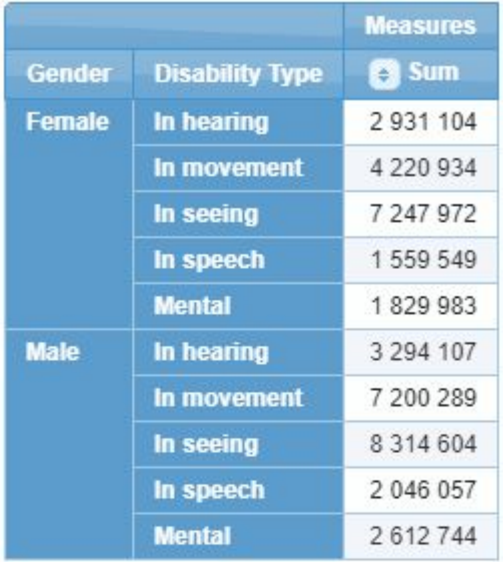


	Measures
Disability Type	Sum
In seeing	15 562 576
In movement	11 421 223
In hearing	6 225 211
Mental	4 442 727
In speech	3 605 606

Are Females More Disabled than Males In every Disability Type?

```
SELECT {[Measures].[Sum]} ON COLUMNS,
CrossJoin({[Gender].children}, {[Disability Type].children}) ON ROWS
FROM [Cube]
```





		Measures
Gender	Disability Type	Sum
Female	In hearing	2 931 104
	In movement	4 220 934
	In seeing	7 247 972
	In speech	1 559 549
	Mental	1 829 983
Male	In hearing	3 294 107
	In movement	7 200 289
	In seeing	8 314 604
	In speech	2 046 057
	Mental	2 612 744

Reporting Demo With Power BI