# OpenstreetMap Project Data Wrangling with MongoDB

# 1 PROCESS FOLLOWED IN PROJECT

This is the process followed

1. We clean the data using Audit.py, the Code has been commented in Audit.py
2. We convert data to json using data.py
3. Once converted to json we use the mongo import clause to import json to mongo db this step is documented in mongoimport.txt
4. We have also included module in Audit.py to clean the phone numbers utilizing pcre re Expression Operations module
5. We have commented additional problems and challenges in implementing the improvement

# 2 PROBLEMS ENCOUNTERED IN THE MAP

After initial Downloading the file for Dubai- Abu Dhabi Area , and running against provisional data.py File. I have noticed following issue

## 2.1 NAMES ARE IN ARABIC

The data contains “name” attribute value in Arabic, the correct “name” is in attribute of K tag with the name “Name:en”

<tag k="name" v="کیش"/>

<tag k="name:en" v="Kish"/>

### 2.1.1 Solution for Names in Arabic

While creating the Json file , we first check for existence of “name:en" and the same equivalent values are now passed to json attribute “name”

## 2.2 DATA HAS IN:COUNTRY FROM IRAN

The Data contains tag element with attribute “is_in:country” with text value of “Iran” which is different country than United Arab Emirates, which we are interested in, these are errors in the data.

The Kish island is not part of United Arab Emirates , it is part of iran, but somehow it is part of this Openstreet Map data, which is to be corrected.

<tag k="is_in:country" v="Iran"/>

### 2.2.1 Solution for data contains IRAN

While creating the Json file , we first check for existence of "is_in:country" and the same equivalent values of V attribute if it is Iran, we simply skip the row

## 2.3 INCORRECT DATA FOR AMENITY

On querying dubai collection, we found the distinct value of Amenity using following clause

we figured out that there is an incorrect value of Amenity as “Jebel Ali Station”, the location is correct and it is Jebel Ali Station , so we corrected the Amenity as bureau_de_change",

### 2.3.1 Updating Incorrect data for amenity

using update_amenity.py

> db.dubai.find({"amenity":"Jebel Ali Station"}).count()

1

> db.dubai.find({"amenity":"Jebel Ali Station"})

{ "_id" : ObjectId("572a225072edaabb5f02fb92"), "amenity" : "Jebel Ali Station", "network" : "RTA", "created" :

{ "changeset" : "37140259", "user" : "ConsEbt", "version" : "8", "uid" : "313448", "timestamp" : "2016-02-11T07:

03:44Z" }, "pos" : [ 55.0910256, 24.9774785 ], "station" : "subway", "address" : { "city" : "Dubai", "country" :

"AE" }, "railway" : "station", "type" : "node", "id" : "1379661746", "name" : "UAE Exchange" }

> db.dubai.find({"amenity":"restaurant"}).count()

176

After executing update_amenity.py

$ sudo python3 update_amenity.py

$ mongo

> use local

switched to db local

> db.dubai.find({"amenity":"Jebel Ali Station"})

> db.dubai.find({"amenity":"restaurant"}).count()

177

> db.dubai.find({"amenity":"Jebel Ali Station"}).count()

0

> db.dubai.find({"name" : "UAE Exchange"})

{ "_id" : ObjectId("572a225072edaabb5f02fb92"), "amenity" : "bureau_de_change", "type" : "node", "railway" : "st

ation", "name" : "UAE Exchange", "id" : "1379661746", "address" : { "city" : "Dubai", "country" : "AE" }, "pos"

: [ 55.0910256, 24.9774785 ], "created" : { "uid" : "313448", "changeset" : "37140259", "timestamp" : "2016-02-1

1T07:03:44Z", "version" : "8", "user" : "ConsEbt" }, "network" : "RTA", "station" : "subway" }

{ "_id" : ObjectId("572a225072edaabb5f032d59"), "amenity" : "bureau_de_change", "name" : "UAE Exchange", "create

d" : { "changeset" : "9881953", "user" : "MathewJ", "version" : "1", "uid" : "451206", "timestamp" : "2011-11-20

T10:25:22Z" }, "pos" : [ 55.3384777, 25.2772196 ], "type" : "node", "id" : "1509578887" }

{ "_id" : ObjectId("572a225172edaabb5f0337f7"), "amenity" : "bureau_de_change", "name" : "UAE Exchange", "create

d" : { "changeset" : "10410357", "user" : "Walter Schlögl", "version" : "2", "uid" : "78656", "timestamp" : "201

2-01-16T17:51:39Z" }, "pos" : [ 55.3476672, 25.2785833 ], "type" : "node", "id" : "1536329967" }

This confirms the update clause works fine.

# 3 DATA OVERVIEW

We used  following script to get unique amenity in the output.json file

mongo --quiet local --eval 'printjson(db.dubai.distinct("amenity"))' > output.json

The section contains basic statistics about the dataset and the MongoDB Queries used to gather them

## 3.1 FILE SIZES

dubai_abu-dhabi.osm 353.5 MB

dubai_abu-dhabi.osm.json 157.9 MB

## 3.2 Number of Documents

```
> use local

switched to db local

> db.dubai.find().count()

655515
```

## 3.3 Number of Documents of type node

```
>db.dubai.find({"type":"node"}).count()

655509
```

## 3.4 Number of Documents of type way

```
>db.dubai.find({"type":"way"}).count()

0
```

## 3.5 Number of distinct users

```
> db.dubai.distinct("created.user").length

612
```

## 3.6 Top contributing user

```
> db.dubai.aggregate([{$group:{_id:"$created.user","count":{"$sum":1}}},{"$sort":{"count":-
1}},{"$limit":1}])

{ "_id" : "eXmajor", "count" : 158029 }
```

## 3.7 Number of contributing user with one post

```
> db.dubai.aggregate([{$group:{_id:"$created.user","count":{$sum:1}}},{$match:{"count":
1}},{$group:{_id:null,"count1":{$sum:1}}}, {"$limit":1}])

{ "_id" : null, "count1" : 119 }
```

# 4 ADDITIONAL IDEAS

## 4.1 DATA CLEANSING AND ANALYSIS PROCESS

I agree on the basic premise that data cleansing and importing has to be iterated , but since the amount of data to be queried and corrected is Huge

I believe a better way is to Import all data into staging collection with all key/Value of data form the tags of K and V

And then using MongoDB we can do lot more faster analysis of distinct keys and values and start creating additional collections with corrected keys and value pair which we care about

## 4.2 ADDITIONAL ELEMENT TO CAPTURE RENT

In Dubai, there is lot of analysis going on to find the accommodation with cheaper rent and better Amenity, I believe we can add another element called Rent and include keys of 1BHK, 2BHK, 3BHK, Studio, Villa Etc and populate these values and ask people to update these values, which can be eventually used for analysis to find areas within relevant budget and room sizes

### 4.2.1 Anticipated Issues / Challenges in implementing enhancement

Since there is no incentive to update, the data generated could be not trusted as it is not updated frequently leading to incorrect value as of date and information partly available

One solution could be to advertise the benefits of this initiative to Newspaper which will frequently utilize such resources and publish rent statics, stating the sources, which will incentive general public to actively participate in this initiative.

## 4.3 ADDITIONAL DATA EXPLORATION USING MONGODB QUERIES

# Top 10  amenities

>
db.dubai.aggregate([{"$match":{"amenity":{"$exists":1}}},{"$group":{"_id":"$amenity","count":{"$sum":
1}}},{"$s

ort":{"count":-1} }, {"$limit":10}])

{ "_id" : "restaurant", "count" : 177 }

{ "_id" : "fuel", "count" : 172 }

{ "_id" : "parking", "count" : 154 }

{ "_id" : "fast_food", "count" : 94 }

{ "_id" : "place_of_worship", "count" : 77 }

{ "_id" : "bank", "count" : 51 }

{ "_id" : "cafe", "count" : 40 }

{ "_id" : "hospital", "count" : 40 }

{ "_id" : "pharmacy", "count" : 39 }

{ "_id" : "atm", "count" : 39 }

# Biggest Contributor of Valid Address Name

>
db.dubai.aggregate([{"$match":{"name":{"$exists":1}}},{"$group":{"_id":"$created.user","count":{"$sum
":1}}},{"

$sort":{"count":-1} }])

{ "_id" : "Metehyi", "count" : 1242 }

{ "_id" : "csdf", "count" : 258 }

{ "_id" : "eXmajor", "count" : 172 }

{ "_id" : "Tiramon", "count" : 92 }

{ "_id" : "MathewJ", "count" : 73 }

{ "_id" : "chachafish", "count" : 70 }

{ "_id" : "Tommy", "count" : 67 }

{ "_id" : "aighes", "count" : 67 }

{ "_id" : "Sal73x", "count" : 43 }

{ "_id" : "Seandebasti", "count" : 34 }

{ "_id" : "yahya", "count" : 32 }

{ "_id" : "jphilipz", "count" : 31 }

{ "_id" : "Vector82", "count" : 29 }

{ "_id" : "PavelPS", "count" : 24 }

{ "_id" : "jiggybee", "count" : 23 }

{ "_id" : "amlapierre", "count" : 22 }

{ "_id" : "xybot", "count" : 22 }

{ "_id" : "hno2", "count" : 21 }

{ "_id" : "wheelmap_visitor", "count" : 20 }

{ "_id" : "kesler", "count" : 19 }

# Biggest Contributor without Valid Address Name

> db.dubai.aggregate([{"$group":{"_id":"$created.user","count":{"$sum":1}}},{"$sort":{"count":-1} }])

{ "_id" : "eXmajor", "count" : 158029 }

{ "_id" : "csdf", "count" : 85638 }

{ "_id" : "AndrewBuck", "count" : 53447 }

{ "_id" : "Skywave", "count" : 33760 }

{ "_id" : "crackers250", "count" : 33404 }

{ "_id" : "Tommy", "count" : 24643 }

{ "_id" : "richard worl", "count" : 16978 }

{ "_id" : "DOM91", "count" : 15913 }

{ "_id" : "Seandebasti", "count" : 13792 }

{ "_id" : "hno2", "count" : 13246 }

{ "_id" : "chachafish", "count" : 11278 }

{ "_id" : "sebbehr", "count" : 10381 }

{ "_id" : "Oberaffe", "count" : 9386 }

{ "_id" : "Supercarwaar", "count" : 9107 }

{ "_id" : "Scrup", "count" : 7482 }

{ "_id" : "Ben", "count" : 7319 }

{ "_id" : "Asterix200", "count" : 6007 }

{ "_id" : "Muokkaaja", "count" : 5702 }

{ "_id" : "robgeb", "count" : 5215 }

{ "_id" : "kesler", "count" : 4567 }

# Most Number of addresses with valid name by city

>
db.dubai.aggregate([{"$match":{"name":{"$exists":1}}},{"$group":{"_id":"$address.city","count":{"$sum":1}}},{"

$sort":{"count":-1} }])

{ "_id" : null, "count" : 2902 }

{ "_id" : "Abu Dhabi", "count" : 61 }

{ "_id" : "Dubai", "count" : 25 }

{ "_id" : "Ras Al Khaimah", "count" : 3 }

{ "_id" : "Dubai Media City, Dubai", "count" : 3 }

{ "_id" : "AE", "count" : 2 }

{ "_id" : "Dubai Media City", "count" : 2 }

{ "_id" : "Khasab", "count" : 2 }

{ "_id" : "Yas Island, Abu Dhabi", "count" : 2 }

{ "_id" : "New Shahama", "count" : 2 }

{ "_id" : "Yas Island", "count" : 1 }

{ "_id" : "Ras al Khaimah", "count" : 1 }

{ "_id" : "Khalifa City A", "count" : 1 }

{ "_id" : "Dubai Marina, Dubai", "count" : 1 }

{ "_id" : "الورقـــاء", "count" : 1 }

{ "_id" : "Sharjah", "count" : 1 }

{ "_id" : "Al Jabeeb", "count" : 1 }

{ "_id" : "Khorfakkan", "count" : 1 }

{ "_id" : "New Shahama, Abu Dhabi", "count" : 1 }

{ "_id" : "Al Aqqa, Fujairah", "count" : 1 }

# Most Number of restaurants by city

> db.dubai.aggregate([{"$match":{"amenity":{"$exists":1},"amenity":"restaurant"}},
{"$group":{"_id":"$address.

city","count":{"$sum":1}}},{"$sort":{"count":-1}}])

{ "_id" : null, "count" : 139 }

{ "_id" : "Abu Dhabi", "count" : 28 }

{ "_id" : "Dubai", "count" : 4 }

{ "_id" : "Dubai Media City, Dubai", "count" : 2 }

{ "_id" : "Dubai Media City", "count" : 1 }

{ "_id" : "Ras al Khaimah", "count" : 1 }

{ "_id" : "Al Jabeeb", "count" : 1 }

{ "_id" : "Yas Island, Abu Dhabi", "count" : 1 }

# Biggest religion

> db.dubai.aggregate([{"$match":{"amenity":{"$exists":1},
"amenity":"place_of_worship"}},{"$group":{"_id":"$reli

gion", "count":{"$sum":1}}},{"$sort":{"count":-1}}])

{ "_id" : "muslim", "count" : 64 }

{ "_id" : null, "count" : 11 }

{ "_id" : "christian", "count" : 2 }

# Most popular cuisines

>
db.dubai.aggregate([{"$match":{"amenity":{"$exists":1},"amenity":"restaurant"}},{"$group":{"_id":"$cuisine","c

ount":{"$sum":1}}},{"$sort":{"count":-1}}])

{ "_id" : null, "count" : 80 }

{ "_id" : "italian", "count" : 13 }

{ "_id" : "indian", "count" : 11 }

{ "_id" : "pizza", "count" : 10 }

{ "_id" : "international", "count" : 7 }

{ "_id" : "french", "count" : 6 }

{ "_id" : "lebanese", "count" : 6 }

{ "_id" : "regional", "count" : 5 }

{ "_id" : "kebab", "count" : 4 }

{ "_id" : "asian", "count" : 3 }

{ "_id" : "steak", "count" : 3 }

{ "_id" : "steak_house", "count" : 3 }

{ "_id" : "chicken", "count" : 3 }

{ "_id" : "seafood", "count" : 2 }

{ "_id" : "sushi", "count" : 2 }

{ "_id" : "japanese", "count" : 2 }

{ "_id" : "pakistani", "count" : 1 }

{ "_id" : "vegetarian", "count" : 1 }

{ "_id" : "sandwich", "count" : 1 }

{ "_id" : "american", "count" : 1 }