AIDR Persister File Parsing

|  |  |
| --- | --- |
| REQ No.  1 | Project  AIDR Persister Files to Database |

Abstract

This document covers requirements for aidr persister files to database.

| **Rev** | **Date** | **Author** | **Status** |
| --- | --- | --- | --- |
| 1 |  | Ji Lucas | First draft |
| 2 |  |  |  |
|  |  |  |  |

Distribution List

|  |  |
| --- | --- |
| **Reviewers** | **FYI** |
| Ji Lucas |  |

Table of Contents

1. Introduction 4

2. Related Documents/Links/People 4

3. Glossary 4

4. Scope 5

4.1. Goals and Objectives 5

4.2. Assumptions 5

4.3. Constraints 5

4.4. Dependencies 5

4.5. Risks 5

5. Analysis 5

5.1. Overview 5

5.2. Database Requirements 6

5.2.1. General Requirements 6

5.2.2. File Details 6

5.3. Databases 7

5.3.1. PostgreSQL 7

5.3.2. Table : data\_collection 7

5.3.3. Table : data\_geo 8

5.4. Non-functional Related 9

5.4.1. Security 9

5.4.2. QA/Testing 9

Issues/Questions 9

6. Revision History 10

AIDR Persister File Parsing

# Introduction

# Related Documents/Links/People

References in the text throughout this document appear in square brackets (e.g., [1], [JS]).

| **Reference** | **Document/Link/Person/Application** |
| --- | --- |
|  | MicroMappers http://clickers.micromappers.org/ |
|  | AIDR http://aidr-dev.qcri.org/AIDRFetchManager/ |
|  | Digital Humanitarian volunteers coordinator |
|  | Image Clicker http://clickers.micromappers.org/app/MM\_ImageClicker/ |
|  | Text Clicker http://clickers.micromappers.org/app/MM\_TextClicker/ |
|  | MicroFilters |
|  | Video Clicker http://clickers.micromappers.org/app/MM\_VideoClicker/ |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Glossary

| **Term** | **Definition** |
| --- | --- |
| AIDR | Artificial Intelligence for Disaster Response : <http://aidr-dev.qcri.org/AIDRFetchManager/>  To use the application, user has to have twitter account |
| MicroMappers | A cloned customized of Pybossa platform for tagging by Digital Humanitarians http://clickers.micromappers.org/ |
| AIDR Collection Data | Twitter data that is collected by AIDR based on configuration. AIDR Collection output |
| Text Clicker | App that displays tweet text only. Then, user selects one of options |
| Image Clicker | App that displays image only. Then, user selects one of options |
| Video Clicker | App that displays Youtube video only. Then, user selects all scenes that is related damage assessment. |
| Aerial Clicker |  |
|  |  |
|  |  |
|  |  |

# Scope

## Goals and Objectives

## Assumptions

* Predictable data access pattern
* Well-structured dataset(JSON)

## Constraints

* Social Media api
* Application driven

## Dependencies

## Risks

# Analysis

## Overview



## Database Requirements

### General Requirements

| **ID** | **Requirement** | **Cat** |
| --- | --- | --- |
| 1 | Able to handle big data | M |
| 2 | Able to handle JSON | M |
| 3 | Community supports | M |
| 4 | Geospatial support | M |
| 5 | Proven user cases | M |
| 6 | Developer Momentum / Developer consideration | M |
| 7 | Driver fit |  |
| 8 | Does it simplify your world? |  |
| 9 | Does it fail nicely? |  |
| 10 | Low latency, real-time access required |  |
| 11 | Be accessed by users & applications |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### File Details

|  |  |
| --- | --- |
| **Filename:** | {**short\_name**}\_{yyyymmdd}\_vol-{sequence}.json |
| **Zipped?** | No |
| **Zip filename:** | n/a |
| **Field Delimiter:** | n/a |
| **Record Delimiter:** | n/a |
| **Character set:** | UTF-8 |
| **Generation Frequency:** | Record size or daily |
| **Header row?** | N/A |
| **Footer row?** | No |
| **Retrieval location:** | User’s local or URL |
| **Import database:** | NO |

Keys:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **keys** | **Type** | **Len** | **M/O** | **Comments/Data Mapping** |
| 1 | filter\_level |  |  |  |  |
| 2 | retweeted |  |  |  |  |
| 3 | in\_reply\_to\_screen\_name |  |  |  |  |
| 4 | possibly\_sensitive |  |  |  |  |
| 5 | truncated |  |  |  |  |
| 6 | lang |  |  |  |  |
| 7 | in\_reply\_to\_status\_id\_str |  |  |  |  |
| 8 | id |  |  |  |  |
| 9 | extended\_entities |  |  |  |  |
| 10 | in\_reply\_to\_user\_id\_str |  |  |  |  |
| 11 | timestamp\_ms |  |  |  |  |
| 12 | in\_reply\_to\_status\_id |  |  |  |  |
| 13 | created\_at |  |  |  |  |
| 14 | favorite\_count |  |  |  |  |
| 15 | place |  |  |  |  |
| 16 | coordinates |  |  |  |  |
| 17 | retweeted\_status |  |  |  |  |
| 18 | contributors |  |  |  |  |
| 19 | text |  |  |  |  |
| 20 | geo |  |  |  |  |
| 21 | entities |  |  |  |  |
| 22 | **Aidr** |  |  |  | * Features * crisis\_code * nominal\_labels * doctype * crisis\_name |
| 23 | source |  |  |  |  |
| 24 | favorited |  |  |  |  |
| 25 | in\_reply\_to\_user\_id |  |  |  |  |
| 26 | retweet\_count |  |  |  |  |
| 27 | user |  |  |  |  |

## Database

### PostgreSQL

After reviewing various choices, we have decided to go Postgre9.4 that offers jsob for json object.

* Any json object should be stored as jsonb data type.
* In addition, proper index for jsonb handling is required. Review GIN indexs in Postgresql documentation(http://www.postgresql.org/docs/9.4/static/datatype-json.html)

### Table : data\_collection

| **ID** | **Field** | **Note** |
| --- | --- | --- |
| 1 | id | Unique ID |
| 2 | collection\_id | collection\_id : aidr collection\_id |
| 3 | data | Tweet json. Data type should be **jsonb** |
| 4 | source | tweet, gdelt, facebook and etc. it should be string |
| 5 | created \_at | timestamp |
| 6 |  |  |

### Table : data\_geo

The below is the geo info from twitter schema. If data\_collection.data doesn’t have geometric info, data\_geo table shouldn’t be populated.

|  |  |  |  |
| --- | --- | --- | --- |
| **coordinates** | [Coordinates](https://dev.twitter.com/overview/api/tweets#obj-coordinates) | Nullable. Represents the geographic location of this Tweet as reported by the user or client application. The inner coordinates array is formatted as [geoJSON](http://www.geojson.org/)(longitude first, then latitude).  Example:   |  | | --- | | "coordinates":  {      "coordinates":      [          -75.14310264,          40.05701649      ],      "type":"Point"  } | |

You can find twitter json schema from https://dev.twitter.com/overview/api/tweets

| **ID** | **Field** | **Note** |
| --- | --- | --- |
| 1 | id | Unique ID |
| 2 | data\_collection\_id |  |
| 2 | collection\_id | collection\_id : aidr collection\_id |
| 3 | data | geometric format should be used as data type  http://www.postgresql.org/docs/9.4/static/datatype-geometric.html |
| 5 | created \_at | timestamp |
| 6 |  |  |

### Table : data\_aidr

This table will store aidr attribute json. Any aidr generated json should be saved into the table.

| **ID** | **Field** | **Note** |
| --- | --- | --- |
| 1 | id | Unique ID |
| 2 | collection\_id | collection\_id : aidr collection\_id |
| 3 | data | Tweet json. Data type should be **jsonb** |
| 5 | created \_at | timestamp |
| 6 |  |  |

## Non-functional Related

### Security

| **ID** | **Requirement** | **Source** | **Cat** |
| --- | --- | --- | --- |
| .1 |  |  |  |

### QA/Testing

| **ID** | **Use Case** | **Date/Status** |
| --- | --- | --- |
| .1 |  |  |

# Issues/Questions

| **Issue #** | **Issue/Resolution Description** | **Date/Status** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Revision History

Changes to the text of this document are indicated by bars in the outside margin adjacent to the affected text.

| **Date** | **Change Description** |
| --- | --- |
|  | Initial draft. |
|  |  |
|  |  |
|  |  |