Virtual Event 15-18 June 2020

2020 Asia-Pacific Statistics Week

Leaving no one and nowhere behind

Big Data for Official Statistics: Administrative Area Identification from Plain Text Address

Action Area D. Modernizing statistical business processes (SD1)

Six modernization approaches for your statistical business

Presenter:

Wa Ode Zuhayeni Madjida

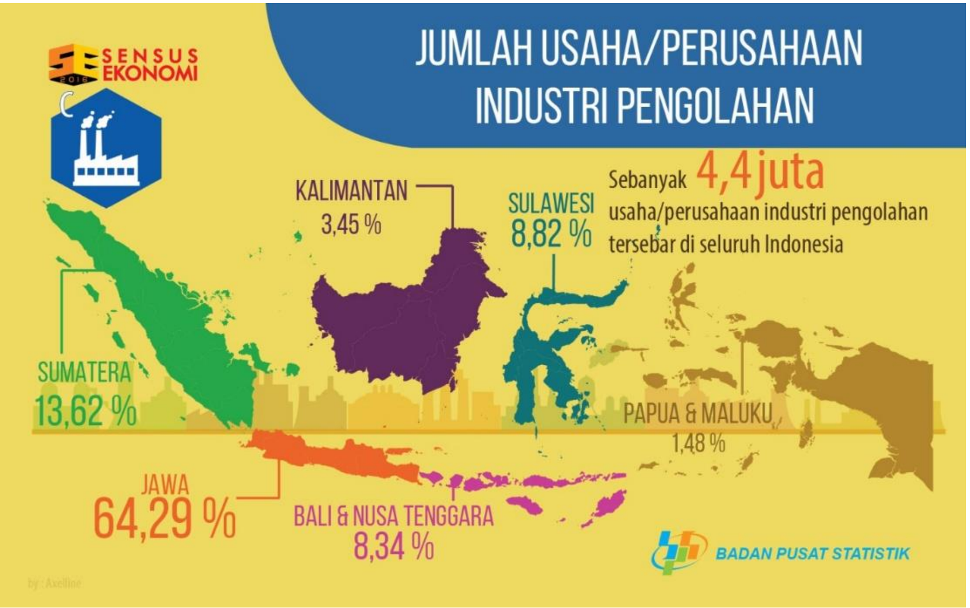
Statistics Indonesia

DECADE OF ACTION

UNITED NATIONS ESCAP Economic and Social Commission for Asia and the Pacific

UKaid from the British people

Administrative Area in Official Statistics

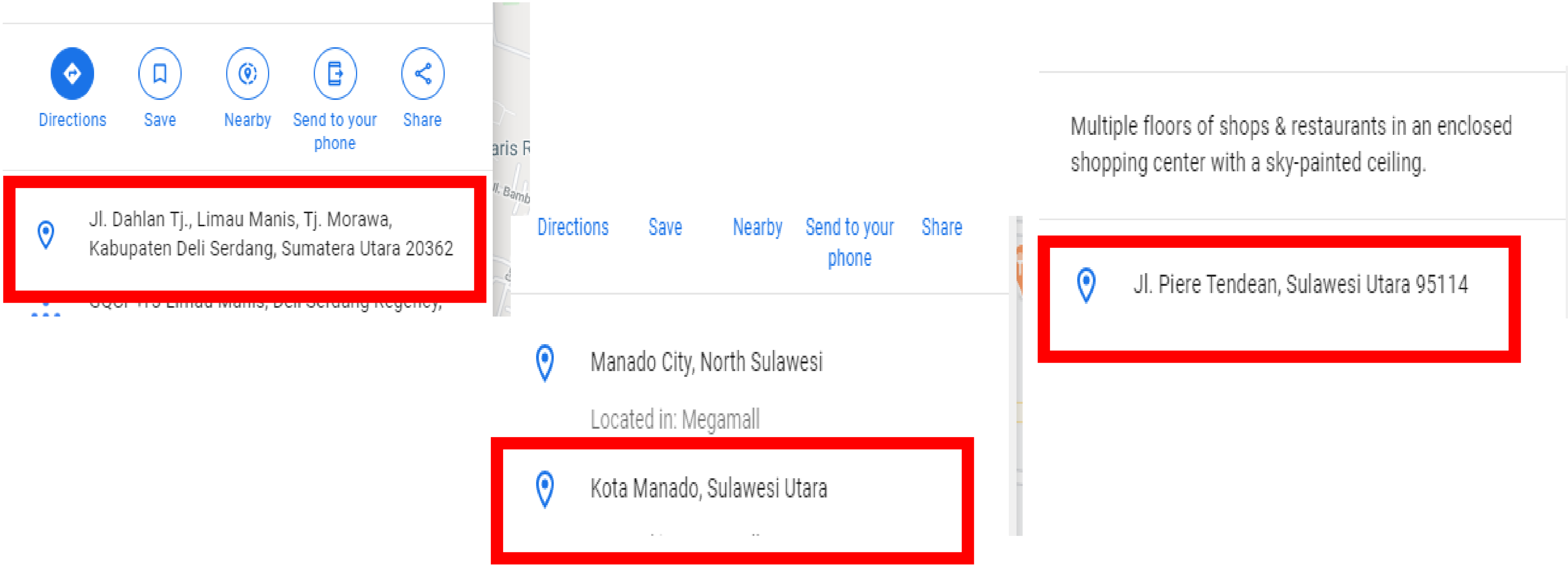

SENSUS EKONOMI 2016
JUMLAH USAHA/PERUSAHAAN INDUSTRI PENGOLAHAN
Sebanyak 4,4 juta usaha/perusahaan industri pengolahan tersebar di seluruh Indonesia
KALIMANTAN 3,45 %
SULAWESI 8,82 %
SUMATERA 13,62 %
PAPUA & MALUKU 1,48 %
JAWA 64,29 %
BALI & NUSA TENGGARA 8,34 %
BADAN PUSAT STATISTIK
by : Axelline

- ❑ Region/Administrative Area information is often used as a dimension in data collection and dissemination
- ❑ Statistics Indonesia has an administrative area standard that used in each statistical process, namely **Master File Desa (MFD)**
- ❑ MFD consist of area code and region name for every level of area

Administrative Area in Official Statistics (Master File Desa in Indonesia)

Province Code	Province Name	Regency code	Regency Name	Sub-district Code	Sub-district Name	Village Code	Village Name
11	Aceh	16	Aceh Jaya	030	Krueng Sabee	010	Kabong
72	Sulawesi Tengah	09	Tojo Una-Una	070	Togean	002	Awo
94	Papua	01	Merauke	051	Eligobel	011	Tof-Tof

Area/Region Information in Big Data



Jl. Dahlan Tj., Limau Manis, Tj. Morawa, Kabupaten Deli Serdang, Sumatera Utara 20362

[illegible] Limau Manis, Deli Serdang Regency,

Manado City, North Sulawesi

Located in: Megamall

Kota Manado, Sulawesi Utara

Multiple floors of shops & restaurants in an enclosed shopping center with a sky-painted ceiling.

Jl. Piere Tendean, Sulawesi Utara 95114

Issue

- ❑ How to integrate big data and census/survey?
- ❑ How to compare big data and census/survey result?
- ❑ Making regional information on big data follow the official statistics standard
- ❑ **Are you sure want to do it manually?**

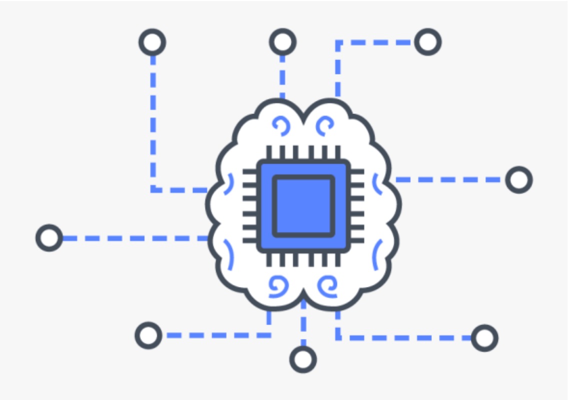
Text Processing as a Solution

MFD as corpus

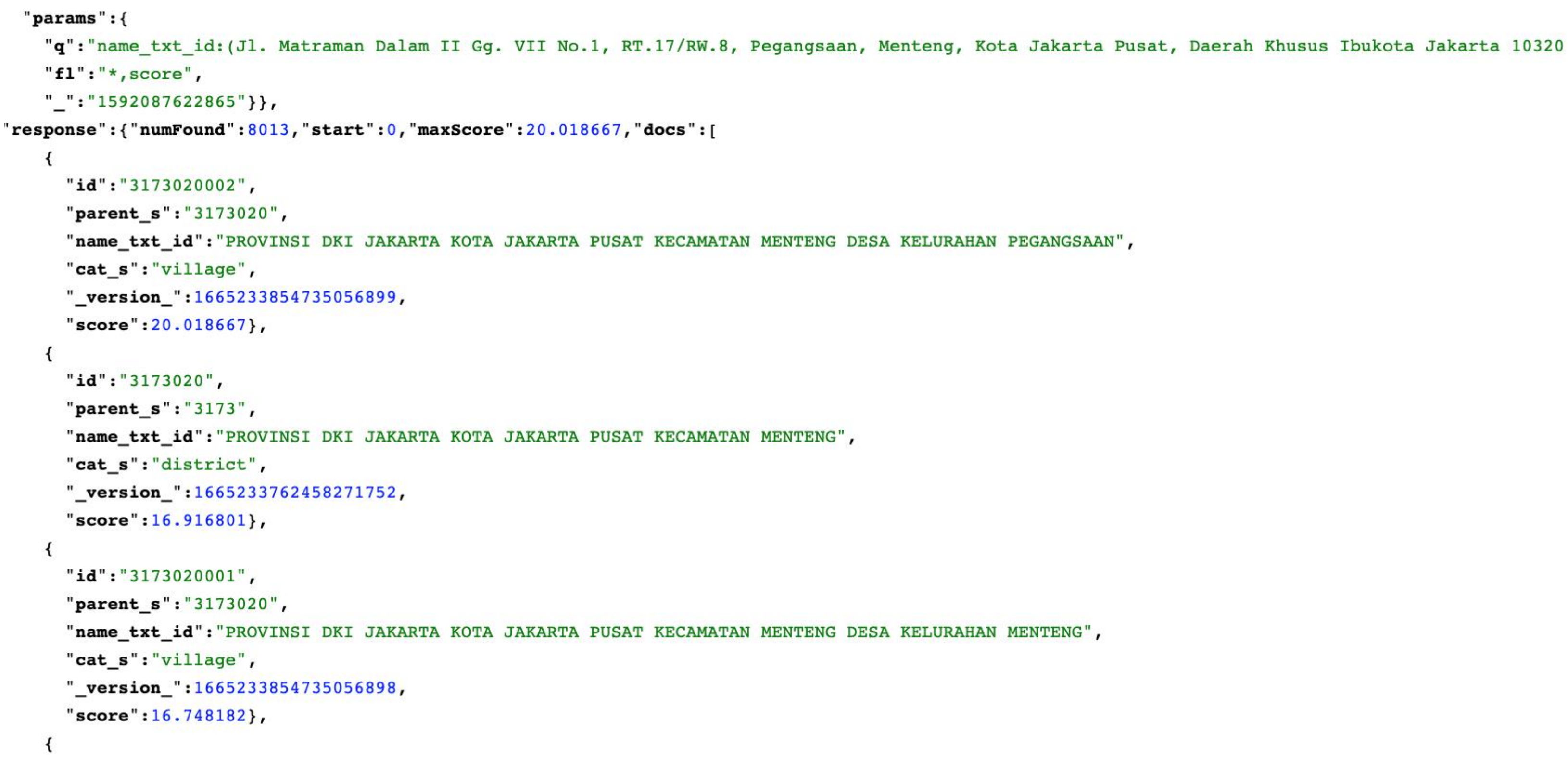
Indexing

Searching

Scoring

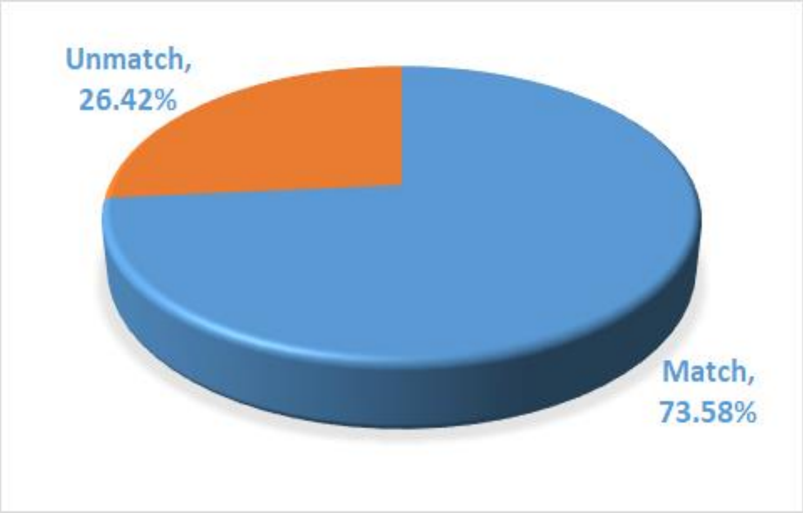


Text Processing as a Solution



```
  "params":{
    "q":"name_txt_id:(Jl. Matraman Dalam II Gg. VII No.1, RT.17/RW.8, Pegangsaan, Menteng, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10320
    "fl":"*,score",
    "_":"1592087622865"}},
"response":{"numFound":8013,"start":0,"maxScore":20.018667,"docs":[
    {
      "id":"3173020002",
      "parent_s":"3173020",
      "name_txt_id":"PROVINSI DKI JAKARTA KOTA JAKARTA PUSAT KECAMATAN MENTENG DESA KELURAHAN PEGANGSAAN",
      "cat_s":"village",
      "_version_":1665233854735056899,
      "score":20.018667},
    {
      "id":"3173020",
      "parent_s":"3173",
      "name_txt_id":"PROVINSI DKI JAKARTA KOTA JAKARTA PUSAT KECAMATAN MENTENG",
      "cat_s":"district",
      "_version_":1665233762458271752,
      "score":16.916801},
    {
      "id":"3173020001",
      "parent_s":"3173020",
      "name_txt_id":"PROVINSI DKI JAKARTA KOTA JAKARTA PUSAT KECAMATAN MENTENG DESA KELURAHAN MENTENG",
      "cat_s":"village",
      "_version_":1665233854735056898,
      "score":16.748182},
    {
```

Result



Future Challenge

- ❑ Expansion of region
- ❑ The use of abbreviation in region information of big data source
- ❑ Unstructured grammar

DECADE OF ACTION
UNITED NATIONS ESCAP Economic and Social Commission for Asia and the Pacific
UKaid from the British people