Hierarchical Of Grid Partition (HGP) For Measuring The Similarity Of Data In Optimizing Data Accuracy

Verdi Yasin¹, Muhammad Zarlis², Opim Salim Sitompul³· Poltak Sihombing⁴

¹Graduate Doctoral Program in Computer Science, Universitas Sumatera Utara, Medan, Sumatera Utara 20155, Indonesia.

²Department of Computer Science, Universitas Sumatera Utara, Medan, Sumatera Utara 20155, Indonesia.

³Department of Information Technology, Universitas Sumatera Utara, Medan, Sumatera Utara 20155, Indonesia.

⁴Department of Computer Science, Universitas Sumatera Utara, Medan, Sumatera Utara 20155, Indonesia.

Abstract
Data redundancy of candidate votes during general elections held once every 5 years in Indonesia is one of the reasons for election malpractice and rigging. Therefore, this study aims to determine the similarity of data object attributes using the Hierarchical of Grid Partition (HGP) to minimize data redundancy (Chenhun et al., 2018; Ding, 2021; Hamza & Recep, 2018; Nur & Karima, 2020; Yijie, 2021). One of such attributes includes having a double ID to prevent data manipulation. Other tools employed to measure or test for similarity are Flowchart measurement, Relation Model, and Class Diagram (Bin et al., 2018; Kaitlyn & Aspen, 2017; Wisal et al., 2019). Data were collected from a sample of 6,215 candidate votes stored in a SQL database to determine their Main Variable (Vu), Secondary Variable (Vs), and Alternate Variable (Va). The result showed 5,211 normal data (K=0) without similarity (single ID) at 83.85%. Furthermore, there are 922 (K=1) Invalid data with 3 types of object attribute similarities (potentially fraud occurrence) at a medium similarity risk of 14.84%. There are 78 invalid data (K=1) with 2 object similarities and 4 valid data (K=1) with a similarity ratio of 0.06%. Results find similar,
effective, optimal and accurate data.

Keywords
Data Redundancy; Hierarchical Grid Partition (HGP); Data Variabel (Vu, Vs, Va); Result

Introduction
General elections are held once every 5 years by the General Election Commission of the Republic of Indonesia to elect State officials, such as President and Vice President, Province Governor and Vice Governor, and city Mayor and Vice Mayor, etc (Kateryn & Oleksandr, 2020; Nur & Karima, 2020; Ranjani, 2020; Republik Indonesia, 2017, 2020).

However, these elections are subject to data redundancy due to similarities, making it challenging to retrieve civilian data stored in the Population Data department of the Home Affairs Ministry of the Republic of Indonesia (Ding, 2021; Nur & Karima, 2020; Republik Indonesia, 2013). Therefore, it is imperative to develop a method or model capable of maximizing and using the right method to extract data during the importation process to avoid redundancy, thereby doubling the population data of permanent voters (Republik Indonesia, 2006, 2013).

The Hierarchical of Grid Partition (HGP) method is one of the solutions that can be used to overcome the problem associated with doubled data entity attributes (Bin et al., 2018; Zhenxiu et al., 2021; Zhengxi et al., 2020; Ruizhe et al., 2020). The mechanism of this method is carried out by importing data from within the population and civil database into the application system of the General Election Commission of the Republic of Indonesia (KPU) (Nur & Karima, 2020). After the importation process, the data is filtered to determine similar attributes associated with the primary, secondary, and type using the hierarchical method. The primary data measured is the population ID number, a unique identification number issued to the population for the general election. Secondary data mainly comes from birth records to amplify the process of determining similarity using resident addresses (Verdi et al., 2019, 2020; Zhenxiu et al., 2021; Ruizhe et al., 2020; Zhengxi et al., 2020; Bin et al., 2018; Nur & Karima, 2020).

The main reason for data classification is to eliminate redundancy due to similarity (Ding, 2021; Zhengxi et al., 2020). The secondary data, which contains birth records and other attributes used to determine one's real identity, are weil to data fraud. Therefore, it is important to use alternative data, such as resident addresses, to ensure someone's data are accurate (Kateryn & Oleksandr, 2020; Zhengxi et al., 2020). Although people can have similar attributes, it is impossible for them to same resident address under the same ID card. However, it is possible for someone to have more than one ID card by modifying their names, thereby creating data fraud (Verdi et al., 2020, 2019).
Hierarchy is a tree-like structure used to simplify complex data by separating them into groups before storing in files, databases, bits, bytes (characters), records, and fields (Buchholz & Werner, 1962; Laudon & Jane, 2011; McLeod et al., 2008; Relita, Muhammad, et al., 2021; Relita, Syahril, et al., 2021; Wisal et al., 2019). Generally, data is divided into 6 hierarchies in the database from the least to the most, namely Bit, Byte, Field, Record, File, and Database. The smallest part of a database is a bit, a synonym of a binary containing only 1 or 0. A combination of 8 bits is called a byte, which forms a field consisting of words, characters, symbols, and pictures. Furthermore, related fields form a record, defined as a single row in a database. Similar types of records from a file or a table and a group of related tables are called a database (Ding, 2021; Kaitlyn & Aspen, 2017; Wisal et al., 2019). This data hierarchy does not exist at the database level because there are also data mart and data warehouse, which of course, are a collection of databases (Hamza & Recep, 2018; Rainer & Cegielski, 2019; Relita, Syahril, et al., 2021; Zulfian & Verdi, 2021).

The "data hierarchy" is a systematic organization of data in characters, fields, records to help determine the connection between the smallest and largest part inside a database. It is also used to share more understanding related to data components and their relationship (Denis et al., 2012; Rainer & Cegielski, 2019). The referential integrity is the third normal form or perfect key of data used to determine redundancy precisely and discover its relational structure (Ding, 2021; Rainer & Cegielski, 2019; Zulfian & Verdi, 2021).

The data grid is configured to enable people to access, modify and transfer extensive data in a virtual media environment. It supports data management and is also used to control its usages. The CPU scavenging grid is a project system transferable from one computer to another according to users’ needs. Furthermore, it is capable of moving and easier to use to achieve desired goals. In simple words, the data grid is a combination of data between computers or database servers inside a distributed network system that are integrated under one management to reduce or avoid data redundancy (Arni & Uli, 2021; Ding, 2021). Data can also be partitioned to enable users to choose the records or files to be used during the general election.

A partition is an object divided into various parts to achieve a specific goal. This is literally an English term, which means divider, and it is commonly used to define a hard drive with 2 or more storage areas. Partition is also a method used to divide data inside a hard drive in a computer system for effective data management (Arni & Uli, 2021; Buchholz & Werner, 1962; Kateryn & Oleksandr, 2020; Laudon & Jane, 2011; McLeod et al., 2008; Rainer & Cegielski, 2019; Verdi, 2012; Zhenxiu et al., 2021).
This study used an object-oriented system and a Unified Modeling Language to model data and processed its design. Furthermore, the HGP method simplifies and measures data accuracy, using Case, Activity, and Class Diagrams (Denis et al., 2012; Kaitlyn & Aspen, 2017; Laudon & Jane, 2011; McLeod et al., 2008; Verdi, 2012; Yijie, 2021). A processes algorithm was also created to display the workflow of the Hierarchical of Grid Partition (HGP) Method using flowcharts (Verdi et al., 2019, 2020; Zhenxiu et al., 2021; Ruizhe et al., 2020; Zhengxi et al., 2020).

This study aims to develop a data optimization model for permanent voters to eliminate redundancy during general elections by using the Hierarchical of Grid Partition (HGP) method (Kateryn & Oleksandr, 2020; Ruizhe et al., 2020; Verdi et al., 2020, 2019; Zhengxi et al., 2020; Zhenxiu et al., 2021).

Research Method

This is a scientific research used to analyze data redundancy and other related problems to produce optimal results.

The Research Variable Instruments

The research variables used are as follows:

- Main Variable (Vu): Population Data
- Secunder Variables (Vs): Birth Data
- Alternative Variables (Va): Other Data

Table 1 Research Data Instrument Variables

<table>
<thead>
<tr>
<th>Main Variable (Vu): Population Data</th>
<th>Secondary Variables (Vs): Birth Data</th>
<th>Alternative Variables (Va): Other Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Name</td>
<td>• Gender</td>
<td>• Village</td>
</tr>
<tr>
<td>• Place of Birth</td>
<td>• Blood Type</td>
<td>• City</td>
</tr>
<tr>
<td>• Date of Birth</td>
<td>• Religion</td>
<td>• Province</td>
</tr>
</tbody>
</table>

Research Process Model (Hierarchical of Grid Partitions/HGP)

This study explains the steps used to solve the problems related to population data analysis of permanent voters (DPT) during the general election in the Republic of Indonesia. The steps used to carry out the HGP methods are shown in Figure 1:
Figure 1 Research Methods (Hierarchical of Grid Partitions/HGP)

The steps in the HGP method are described as follows:
1. Research Object Observation is a process taken from the general election commission (KPU) office in March 2018.
2. Research Instrument Needs Analysis is a process used to set data instruments to be processed or analyzed, further measured with HGP.
3. Required Data Import is a process used to gather data sets from the database server of the population (PCR/DUKCAPIL) to the General Election Commission (KPU).
4. Data Processing Engineering creates a measurement model to determine data similarity using the HGP method.
5. Measuring Data Similarity is a process used to measure data with correlating attributes.
6. Create a Similarity Partitions (Classify Data Similarities) is a process used to categorize measurement results.
7. Determine the Percentage Similarity Data is a process to calculate the percentage ratio of data attributes with accuracy.
8. Results of Accurate Similarity Data (Valid Similarity Data) is used to set data accuracy results with the most similar attributes and less percentage ratio

System Configuration Import Data Sets and Distribution of Fixed Voter Data (FVD/DPT)
The process of importing the data model from the population database server (PCR/DUKCAPIL) into the general election commission database server and the distribution of permanent voters is shown in Figure 2:

**Figure 2 System Configuration Import Data Sets and Distribution of Fixed Voter Data (FVD/DPT)**

The System Configuration Import Data Sets and Distribution of Fixed Voter Data (FVD/DPT), shown in figure 2, are as follows:
1. Integrate population and birth data from the Client-Server of Villages and Districts into the population and civil registration (PCR/DUKCAPIL).
2. Import population data sets from Population and civil registration (PCR/DUKCAPIL) servers as well as the Electoral Commission servers of the Republic of Indonesia.
4. Measure similarities in the data stored for prospective voters.
5. Non-Similar Voter Data (only has a Single Identity) is directly input into the Fixed Voter database Partition (FVD/DPT).
6. Prospective Voter data with similar attributes are input directly into the database partition with similar valid objects (Valid data Accuracy Similarity)
7. Voter data with more than 1 identity are not registered(PCR/DUKCAPIL), rather they are directly input into the database partition Invalid Data Voter (Poor Data / Fake Data / Fraud Data)
8. Distribution of Fixed Voter data to Client-Servers of The Fixed Voter database (FVD/DPT) at the Provincial and Regency/City/Districts/Villages level.

Result and Discussion

This study described, analyzed, and measured data validity to determine similar data validity and accuracy attributes (Bin et al., 2018; Ding, 2021; Hamza & Recep, 2018; Wisal et al., 2019; Yijie, 2021). The model used to categorize population data sets are data variable instruments such as Main Variable (Vu): Population Data, Secunder Variables (Vs): Birth Data, Alternative Variables (Va), etc.

Sample Data Set as Research Instrument

A data sample of 6,215 was used to carry out this study, as shown in Table 2.

Table 2 Sample Data Set as Research Instrument

<table>
<thead>
<tr>
<th>Full name</th>
<th>Place of birt</th>
<th>Date of birt</th>
<th>Gender</th>
<th>Blood type</th>
<th>Religion</th>
<th>Village</th>
<th>City</th>
<th>Provinsi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abby Puda Bimantara</td>
<td>Pontianak</td>
<td>1997-11-14</td>
<td>L</td>
<td></td>
<td>Islam</td>
<td>Tanjung Hulu</td>
<td>Pontianak</td>
<td>Pontianak</td>
</tr>
<tr>
<td>Ahmad Maulana</td>
<td>Lebak</td>
<td>1994-08-06</td>
<td>L</td>
<td></td>
<td>Islam</td>
<td>Muara Ciujung Timur</td>
<td>Lebak</td>
<td>Lebak</td>
</tr>
<tr>
<td>Name</td>
<td>City</td>
<td>Born</td>
<td>L</td>
<td>A</td>
<td>Isla m</td>
<td>Cibubur</td>
<td>Jakarta Timur</td>
<td>Jakarta Timur</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>----</td>
<td>---------</td>
<td>--------</td>
<td>----------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Airlangga Ekatama</td>
<td>Jakarta</td>
<td>1989-07-31</td>
<td>L</td>
<td>A</td>
<td>Isla m</td>
<td>Cibubur</td>
<td>Jakarta Timur</td>
<td>Jakarta Timur</td>
</tr>
<tr>
<td>Amelia Rosmeli</td>
<td>Jakarta</td>
<td>1989-11-13</td>
<td>P</td>
<td>-</td>
<td>Isla m</td>
<td>Kedoya</td>
<td>Jakarta Barat</td>
<td>Jakarta Barat</td>
</tr>
<tr>
<td>Andi Maulana</td>
<td>Lebak</td>
<td>1993-03-25</td>
<td>L</td>
<td>A</td>
<td>Isla m</td>
<td>Cibubah</td>
<td>Lebak</td>
<td>Lebak</td>
</tr>
<tr>
<td>Anis Yoon Deswari</td>
<td>Pekanbaru</td>
<td>2000-12-18</td>
<td>P</td>
<td>O+</td>
<td>Isla m</td>
<td>Rumbai</td>
<td>Pekanbaru</td>
<td>Pekanbaru</td>
</tr>
<tr>
<td>Anna</td>
<td>Jakarta</td>
<td>1990-04-23</td>
<td>P</td>
<td>AB</td>
<td>Isla m</td>
<td>Sukmajaya</td>
<td>Mekarjaya</td>
<td>Mekarjaya</td>
</tr>
<tr>
<td>Annisa Rahmadadina</td>
<td>Pekanbaru</td>
<td>2002-08-16</td>
<td>P</td>
<td>-</td>
<td>Isla m</td>
<td>Rumbai</td>
<td>Pekanbaru</td>
<td>Pekanbaru</td>
</tr>
<tr>
<td>Apriliani</td>
<td>Tangerang Selatan</td>
<td>1999-04-01</td>
<td>P</td>
<td>O</td>
<td>Isla m</td>
<td>Pondo</td>
<td>Tangerang Selatan</td>
<td>Tangerang Selatan</td>
</tr>
<tr>
<td>Aqidatul izzah</td>
<td>Minas</td>
<td>1999-11-19</td>
<td>P</td>
<td>A</td>
<td>Isla m</td>
<td>Minas Jaya</td>
<td>Pekanbaru</td>
<td>Pekanbaru</td>
</tr>
<tr>
<td>Ariska Dwi Prasetyani</td>
<td>Jakarta</td>
<td>1994-07-03</td>
<td>P</td>
<td>O</td>
<td>Isla m</td>
<td>Periuk</td>
<td>Tangerang</td>
<td>Tangerang</td>
</tr>
<tr>
<td>Asshifa Sabrina</td>
<td>Jakarta</td>
<td>1998-12-15</td>
<td>P</td>
<td>O</td>
<td>Isla m</td>
<td>Gandaria</td>
<td>Jakarta Selatan</td>
<td>Jakarta Selatan</td>
</tr>
<tr>
<td>Aulia Nursuciani Mohamaf</td>
<td>Talaga</td>
<td>1999-11-14</td>
<td>P</td>
<td>-</td>
<td>Isla m</td>
<td>Tinelo</td>
<td>Gorontalo</td>
<td>Gorontalo</td>
</tr>
<tr>
<td>Awaludin Fahmi</td>
<td>Tangerang</td>
<td>1989-06-12</td>
<td>L</td>
<td>B</td>
<td>Isla m</td>
<td>Tanah</td>
<td>Tangerang</td>
<td>Banten</td>
</tr>
<tr>
<td>Bahari Dikba Regianto</td>
<td>Jakarta</td>
<td>1998-08-13</td>
<td>L</td>
<td>O</td>
<td>Isla m</td>
<td>Semper</td>
<td>Jakarta Utara</td>
<td>Jakarta Utara</td>
</tr>
<tr>
<td>Banyu Puruhito</td>
<td>Tangerang</td>
<td>1990-10-23</td>
<td>L</td>
<td>B</td>
<td>Isla m</td>
<td>Karang</td>
<td>Tangerang</td>
<td>Karang Tengah</td>
</tr>
<tr>
<td>Bastian Ramadoni</td>
<td>Pekanbaru</td>
<td>2000-01-03</td>
<td>L</td>
<td>A</td>
<td>Isla m</td>
<td>Marpoyan</td>
<td>Pekanbaru</td>
<td>Pekanbaru</td>
</tr>
<tr>
<td>Bayu Putra Pratama</td>
<td>Lubuklinggau</td>
<td>1992-09-21</td>
<td>L</td>
<td>B</td>
<td>Isla m</td>
<td>Taba Koji</td>
<td>Lubuklinggau</td>
<td>Lubuklinggau</td>
</tr>
<tr>
<td>Bima SN</td>
<td>Kebumen</td>
<td>1990-03-02</td>
<td>L</td>
<td>B</td>
<td>Isla m</td>
<td>Wonokriyo</td>
<td>Kebumen</td>
<td>Kebumen</td>
</tr>
<tr>
<td>Chandra Hermawan</td>
<td>Serang</td>
<td>1993-04-06</td>
<td>L</td>
<td>B</td>
<td>Isla m</td>
<td>Cijoropasir</td>
<td>Lebak</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>City</td>
<td>Date</td>
<td>Code</td>
<td>Province</td>
<td>City</td>
<td>Code</td>
<td>Province</td>
<td>Code</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------</td>
<td>---------------</td>
<td>------</td>
<td>----------</td>
<td>---------------</td>
<td>-------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>Chitra Deviana Irianty</td>
<td>Biak</td>
<td>1993-12-09</td>
<td>P</td>
<td>Isla</td>
<td>Muara Ciujung Timur</td>
<td>Lebak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dani</td>
<td>Jakarta</td>
<td>1993-03-23</td>
<td>L</td>
<td>Isla</td>
<td>Penjaringan</td>
<td>Jakarta Selatan</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Dedi Gunawan</td>
<td>Tangerang</td>
<td>1991-12-06</td>
<td>L</td>
<td>A</td>
<td>Ciledug</td>
<td>Tangerang</td>
<td>Banten</td>
<td></td>
</tr>
<tr>
<td>Della Montia</td>
<td>Painan</td>
<td>1990-01-23</td>
<td>P</td>
<td>B</td>
<td>Muara Fajar</td>
<td>Pekanbaru</td>
<td>Riau</td>
<td></td>
</tr>
<tr>
<td>Desi Guinda</td>
<td>Pekanbaru</td>
<td>1999-12-06</td>
<td>P</td>
<td>O</td>
<td>Muara Fajar</td>
<td>Pekanbaru</td>
<td>Riau</td>
<td></td>
</tr>
<tr>
<td>Devi Sry Atmaja</td>
<td>Jakarta</td>
<td>2001-03-22</td>
<td>P</td>
<td>O</td>
<td>Pegangsaan</td>
<td>Jakarta Pusat</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Dinda</td>
<td>Jakarta</td>
<td>1998-09-04</td>
<td>P</td>
<td>A</td>
<td>Kebon Baru</td>
<td>Jakarta Selatan</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Dinda Azzura</td>
<td>Pekanbaru</td>
<td>2002-06-11</td>
<td>P</td>
<td>O</td>
<td>Simpang Tiga</td>
<td>Pekanbaru</td>
<td>Riau</td>
<td></td>
</tr>
<tr>
<td>Dini Hidayah</td>
<td>Pontianak</td>
<td>1997-11-09</td>
<td>P</td>
<td>AB</td>
<td>Batulayang</td>
<td>Pontianak</td>
<td>Kalimantan Barat</td>
<td></td>
</tr>
<tr>
<td>Dwima Velin Heryulika</td>
<td>Dharmasraya</td>
<td>2000-03-14</td>
<td>P</td>
<td>A+</td>
<td>cari sendiri</td>
<td>Dharmasraya</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egi</td>
<td>Lebak</td>
<td>1993-12-26</td>
<td>L</td>
<td>B</td>
<td>Narimbang Mulia</td>
<td>Lebak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eka</td>
<td>Jakarta</td>
<td>1990-07-12</td>
<td>P</td>
<td>A</td>
<td>Kebon Baru</td>
<td>Tebet</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Febriyan Mailano</td>
<td>Jakarta</td>
<td>1992-02-28</td>
<td>L</td>
<td>B</td>
<td>Bungur</td>
<td>Jakarta Selatan</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Fika Kaesarani</td>
<td>Lebak</td>
<td>1993-07-09</td>
<td>P</td>
<td>A</td>
<td>Kadaungung</td>
<td>Lebak</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filqi Albilal</td>
<td>Jakarta</td>
<td>1999-07-14</td>
<td>L</td>
<td>O</td>
<td>Kembangan</td>
<td>Jakarta Barat</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Firda</td>
<td>Jakarta</td>
<td>1990-08-20</td>
<td>P</td>
<td>O</td>
<td>Kebon Melati</td>
<td>Jakarta Pusat</td>
<td>DKI</td>
<td>Jakarta</td>
</tr>
<tr>
<td>Ganjar</td>
<td>Cianjur</td>
<td>1991-08-28</td>
<td>L</td>
<td>AB</td>
<td>Susukan</td>
<td>Cianjur</td>
<td>Jawa Barat</td>
<td></td>
</tr>
<tr>
<td>Gigih</td>
<td>Tangerang Selatan</td>
<td>1989-05-24</td>
<td>P</td>
<td>O</td>
<td>Parigi Lama</td>
<td>Tangerang Selatan</td>
<td>Banten</td>
<td></td>
</tr>
</tbody>
</table>
**Flowchart for Measuring Data Object Attribute Similarity**

This section illustrates the process flow of the data object attributes and similarities using the HGP method, as shown in Figure 3:

![Flowchart Image]

*and so on a total dataset of 6,215*
**Figure 3 Flowchart for Measuring Data Similarity**

The following are a detailed description of the flowchart:

1. The start is the initial process that imports the population dataset into the measurement model of the database server.
2. Read dataset is the initial process used to read the population dataset from the database system.
3. Preprocessing is the initial process used to filter and count temporary voters’ data.
4. The next step is data object attributes similarity measurement, based on the instrument variables of this study, namely Main Variable (Vu), Secondary Variable (Vs), and Alternative Variable (Va).
   - K = Measurement result similarity category
   - Vu = Main variable to measure data similarity
   - Vs = Secondary Variable to measure data similarity
   - Va = Alternative Variable to measure data similarity

The following formula was used to determine similar data.

Vu = Id Number, Name, Place of Birth, Date of Birth, Blood Type, Gender, Religion, Address, Village, District, City, Province.
Vs = Birth Certificate Number, Name, Place of Birth, Date of Birth, Father’s Name, Mother’s name
Va = Family ID Number, Face phrases, Eye phrases, Fingerprint, Previous Address, Current Address.

The HGP model was used to measure similarity as follows:

K = Vu, Vs, Va.

If “K = 0%”, the resulting data is Valid Data Similarity or Data Accurate Similarity, high-Risk data Similarity.
If “K < 1%", then the resulting data is invalid data or Low-Risk Double Data
If “K = > 1%”, then the resulting data is Normal data or Single Identity Data Rule:

**Similarities**

**Rule I: Grid Data**
- IF “K= 0%” Then “Similarities”
- IF “K<1%” Then “Similarities”
- IF “K= >1%” Then “Not Similar”

http://www.webology.org
Rule 2: Partition: Result of Data Grid
   • IF “Similarities =0%” Then “Accurate data object attribute similarity (indication of data fraud)”
   • IF “Similarities <1%” Then "Low-Risk Double Data"
   • IF “Not Similar = >1%” Then “Single Identity Data”

Rule 3: Data Accuracy
   • IF “Similarities = 0%” or “Least Percentage Ratio” Then “Accurate Data Similarity Measurement Results.”

Rule 4: Result:
   • IF “Not Similar = >1%” Then “Accurate Data Certainty is not similar.

Class Diagram of Entity-Relationship Model

This study creates a data entity to describe instrument variables, as shown in Figure 4.
Figure 4 Class Diagram of Entity Relationship Model

Component Diagram of Flow Model System Voter Data Synchronization
Figure 5 is the component diagram of the population dataset flow distributed to client-server databases by the general election commission representative offices across Indonesia.

![Component Diagram of Flow Model System Voter Data Synchronization](image)

**Figure 5 Component Diagram of Flow Model System Voter Data Synchronization**

The Process of Measuring Object Attribute Similarity using a System Database

This study employed a MySQL database server to test and measure the similarity of data object attributes, as shown in Figure 6.

![Process of Measuring Object Attribute Similarity using a System Database](image)

**Figure 6 The Process of Measuring Object Attribute Similarity using a System Database**
An example of a Command used to determine the similarities performed using a grid dataset into a database system is shown in Figure 7.

Figure 7 Display Results Test Similarity of Object Attributes Inter-Variable Data Set

The test/measurement result of data object attributes similarity shown in the following figure of Excel database table.

Figure 8 Measurement Result of Data Attributes Similarity
The data attribute objects similarity measurement result is summarized in Table 2.

If “K = 0”, there are 4,111 data categorized as normal (Single Identity Number)
If “K = -1”, there are 932 data categorized as invalid with medium risk due to the similarities in the 3 attribute types (Single Identity Number)
If "K = -1", 138 data, categorized as invalid, low-risk data because of two attribute types has similarities (Single Identity Number)
K = 1, there are 4 data categorized as valid with high similarity in data object attributes, which can potentially create data fraud.

The following table shows the tabulation of data object attributes measurement results.

**Table 3 Results Measuring the Similarity of Data Objects**

<table>
<thead>
<tr>
<th>Measurement/Test Result Category of Data Attribute similarity</th>
<th>Measurement/Test Result of Data Attribute similarity</th>
<th>Percentage Ratio of Data Attribute similarity</th>
<th>Summary of Measurement/Test Result of Data Attribute similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>If “K = 0”; Normal Data</td>
<td>5,211 normal data that are unique</td>
<td>5,211: 6,215 = 83.85%</td>
<td>Single Identity Data (Non-risky data)</td>
</tr>
<tr>
<td>If “K= -1”; Invalid Data</td>
<td>922 invalid data with 3 attribute types that are similar</td>
<td>922: 6,215 = 14.84%</td>
<td>Data fraud possibility occurred (Data with medium risk)</td>
</tr>
<tr>
<td>If “K= -1”; Invalid Data</td>
<td>78 invalid data with 2 attribute types that are similar</td>
<td>78: 6,215 = 1.26%</td>
<td>No data fraud (Low-risk Data)</td>
</tr>
<tr>
<td>If “K= 1”; Valid Data</td>
<td>4 valid data with major object attributes and similar instrument variables tested/measured.</td>
<td>4: 6,215 = 0.06%</td>
<td>Bad or fraudulent data due to inaccuracy. (High-risk Data)</td>
</tr>
<tr>
<td>Sample tested/measured</td>
<td>6,215 data</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9 shows the similarities between the attributes of this data object.
Conclusion

In conclusion, this study found some similarities in categorizing data attributes from a sample size of 6,215, such as Main Variable (Vu), Secondary Variable (Vs), and Alternate Variable (Va), using hierarchical grid partition method (HGP), as follows:

1. Normal Data category (K= 0), 5,211 data with a percentage ratio of 83.85%, and summarized as Single Data ID (no similarity occurrence).
2. Invalid Data category (K= -1), 78 data with a percentage ratio of 1.26%, and summarized as No Fraud likely to occur (low-risk data) because the similarity is minor.
3. Invalid Data category (K= -1), 922 data with a percentage ratio 14.84%, and summarized as Data Fraud is likely to occur (medium risk data) because the similarity is minor.
4. Valid Data category (K= 1), 4 data with a percentage ratio 0.06%, and summarized as Data Fraud Occurred (High-risk Data) because the similarity is accurate with major data similarity.

This study's measurement or testing model is likely to be developed for more object attributes or other entities.

Acknowledgement

The authors are grateful to the Election Commission of the Republic of Indonesia for the opportunity to carry out this study as a Dissertation Doctoral Program in Computer Science.
aimed at reducing the occurrence of Data Redundancies. In addition, the authors are also grateful to the Promoter and Co-promoter of The Strata Three Dissertation for their guidance in drafting this research at the University of North Sumatra and all colleagues and parties for their financial and material assistance.

References


Conference on Computer and Informatics Engineering (IC2IE, 978-1-7281-8247-6 20 31 00).


