INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
Garden of Knowledge and Virtue
LEADING THE WAY
KHALĪFAH • AMĀNAH • IQRA' • RAHMATAN LIL-ĀLAMĪN
LEADING THE WORLD
INTERNATIONAL MULTI-AWARD WINNING INSTITUTION FOR SUSTAINABILITY
OCT 2024

LILI MARZIANA ABDULLAH
DEPT. OF INFORMATION SYSTEMS
lmarziana@iium.edu.my

Data Profiling: Helping to turn Raw Data into Business Intelligence (BI)

INFO 4311 DATA WAREHOUSING

Structure Discovery

Content Discovery

relationship DISCOVERY

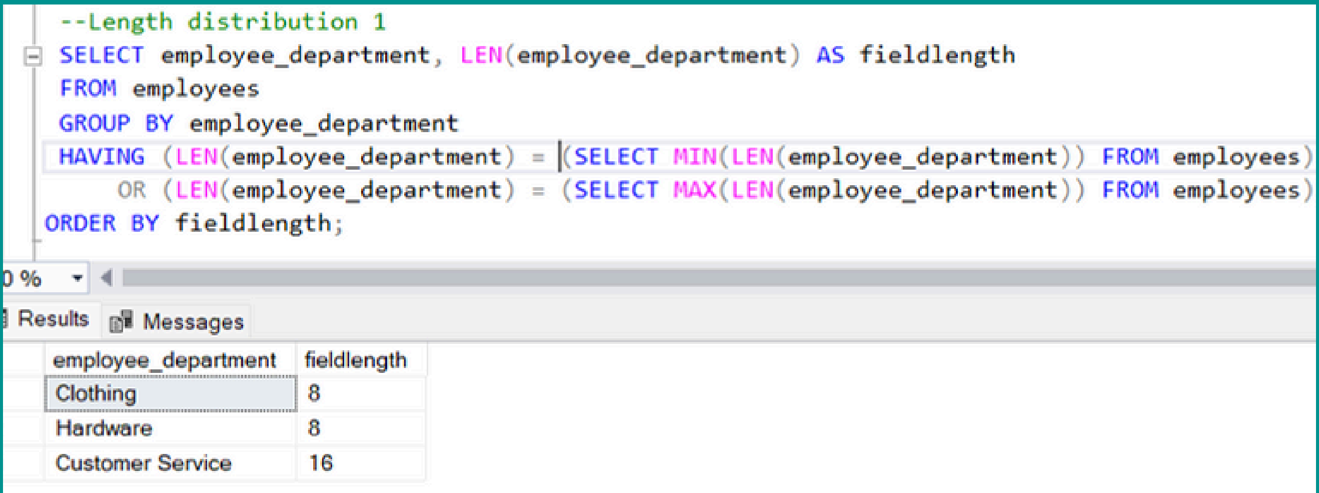
Organizations can become so focused on collecting data and managing day-to-day operations that the effectiveness and quality of the data becomes compromised. Poor-quality data can lead to incorrect analysis, flawed reporting, and misguided decisions. Organisations should be able to know that their data is healthy when they can demonstrate that it is valid, complete and of sufficient quality to produce analytics that they can confidently rely on for decisions. The health of the organisation's data depends on how well the organisation profiles it. Data profiling enables organisations to organize and analyse their data so that it can yield its maximum value and give a clear, competitive edge in the market.

Data Profiling and its Relationship with Business Intelligence (BI)

Data profiling is a process that involves analysing and assessing the quality, structure, and content of data from multiple sources, while creating meaningful summaries of the data. Its primary focus is on understanding, cleansing, and validating data before it can be used for further analytical purposes. Meanwhile, BI refers to the strategies, technologies, and tools used to collect, analyse, and present data to support decision-making. Data profiling serves as a foundational activity for BI by ensuring the quality, accuracy, and consistency of data before it is used in BI systems.

Exposure to Data Profiling

Due to data profiling immense importance to any BI project, students in INFO 4311 Data Warehousing are introduced to data profiling through hands-on group project and/or assignments. Before building a data warehouse in a BI system, students are required to perform data profiling to thoroughly understand the structure, quality, and content of the data. This step is essential for identifying any inconsistencies, errors, or patterns that need to be addressed, ensuring the data is fit for effective use in the data warehouse and subsequent analytics. Students are provided with raw datasets but are also encouraged to source publicly available open data from the internet. However, publicly available data is sometimes already in summarised and clean form, thus, it can be a challenge for the students to find data that they can profile extensively.



```
 --Length distribution 1
 SELECT employee_department, LEN(employee_department) AS fieldlength
 FROM employees
 GROUP BY employee_department
 HAVING (LEN(employee_department) = (SELECT MIN(LEN(employee_department)) FROM employees)
     OR (LEN(employee_department) = (SELECT MAX(LEN(employee_department)) FROM employees)
ORDER BY fieldlength;
```

employee_department	fieldlength
Clothing	8
Hardware	8
Customer Service	16

Figure 1: Maximum and minimum lengths of string values using SQL queries

Getting to know the data is an important step in building the data warehouse because students must understand the data that they must work with before they can leverage it as part of a solution.

To perform data profiling, a wide range of tools can be used depending on the scale and complexity of the data, from specialized data quality tools like Talend or Informatica to general-purpose tools like SQL queries. Students in the course use basic data profiling by writing SQL queries to analyse and summarize data, and the free tool used for the class.

An example of a data profiling technique that lists the minimum and maximum lengths of string values from a string column using SQL queries is shown in Figure 1. Similar question can be asked using a data profiling tool as shown in Figure 2 where the length distribution for every string value can be seen in the result.

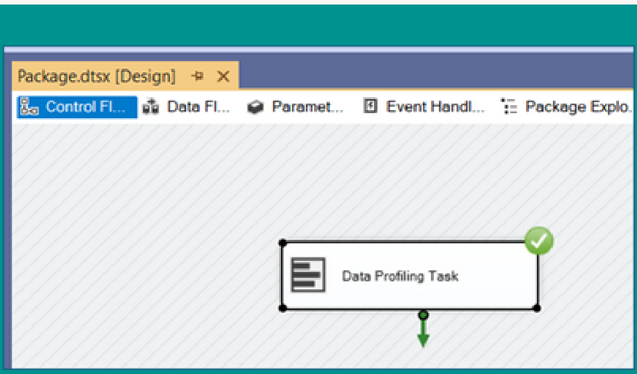
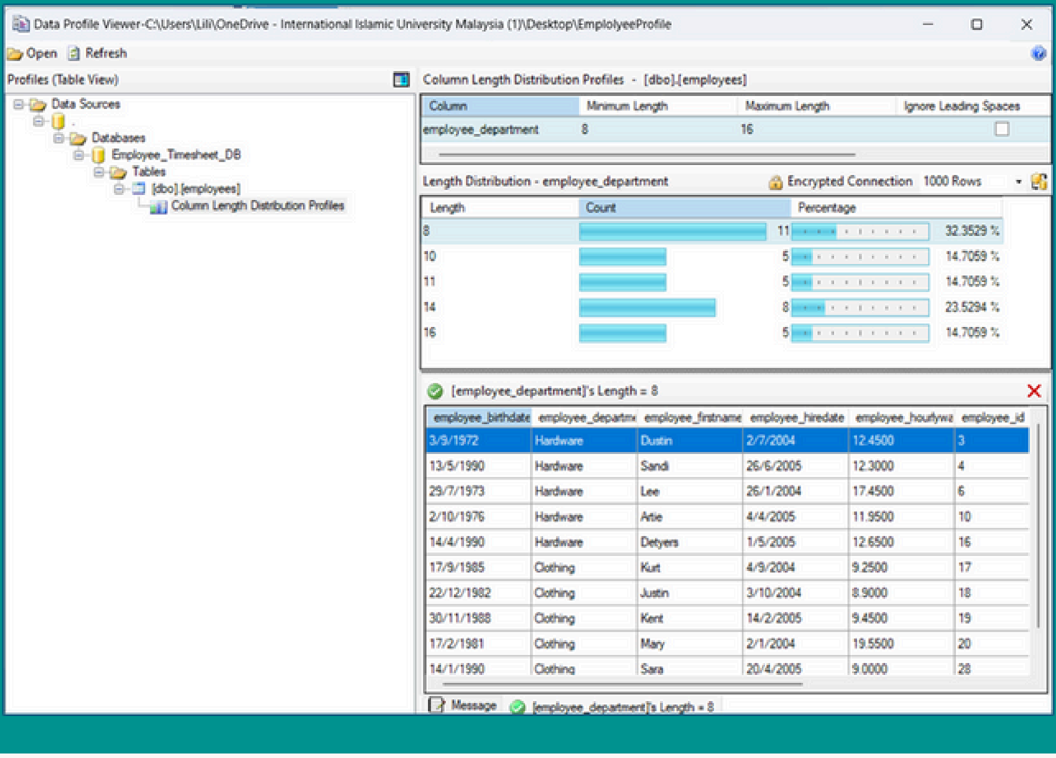


Figure 2: Length distribution of string values using a data integration tool



Data Profile Viewer-C:\Users\Lili\OneDrive - International Islamic University Malaysia (1)\Desktop\EmplolyeeProfile

Column Length Distribution Profiles - [dbo].[employees]

Column	Minimum Length	Maximum Length	Ignore Leading Spaces
employee_department	8	16	□

Length Distribution - employee_department

Length	Count	Percentage
8	11	32.3529 %
10	5	14.7059 %
11	5	14.7059 %
14	8	23.5294 %
16	5	14.7059 %

[employee_department]'s Length = 8

employee_birthdate	employee_departm	employee_firstname	employee_hiredate	employee_hourlywa	employee_id
3/9/1972	Hardware	Dustin	2/7/2004	12.4500	3
13/5/1990	Hardware	Sandi	26/6/2005	12.3000	4
29/7/1973	Hardware	Lee	26/1/2004	17.4500	6
2/10/1976	Hardware	Artie	4/4/2005	11.9500	10
14/4/1990	Hardware	Detyers	1/5/2005	12.6500	16
17/9/1985	Clothing	Kurt	4/9/2004	9.2500	17
22/12/1982	Clothing	Justin	3/10/2004	8.9000	18
30/11/1988	Clothing	Kent	14/2/2005	9.4500	19
17/2/1981	Clothing	Mary	2/1/2004	19.5500	20
14/1/1990	Clothing	Sara	20/4/2005	9.0000	28

[employee_department]'s Length = 8