University of Economics, Prague

Faculty of Informatics and Statistics

Standardisation and integration of data using ETL for integrated customer view

DIPLOMA THESIS

study programme Applied Informatics

field of study Information Systems and Technologies

Author: Bc. Volha Krukovich

Diploma thesis supervisor: doc. Ing. Jan Pour, CSc.

Prague, April 2018

Prohlášení

Prohlašuji, že jsem diplomovou práci Standardization and integration of data using ETL for integrated customer view vypracovala samostatně za použití v práci uvedených pramenů a literatury.

V Praze dne 25. dubna 2018 . . . . . .

Bc. Volha Krukovich

Acknowledgments

I would like to acknowledge and express gratitude to the supervisor of this thesis doc. Ing. Jan Pour, CSc. not only for the valuable advice but also for the guidance and feedback provided.

Abstract

The main objective of the thesis is the analysis, design and implementation of the ETL tool and processes used for a logistics organisation that went through recent mergers and acquisitions and procured multiple other organizations, to provide a consolidated 360 view of their client base by the standard various data elements available across the legacy systems. The first sections of the thesis cover the theory and strategy for the implementation CRM solution.

The remaining section of this thesis covers design and implementation of the Extract-Transform-Load processes used to create a single customer view from interfacing the existing legacy systems within the organisation. This diploma thesis will emphasise the need of the importance of a robust Customer Relationship Management system within an organization.

Keywords

Siebel, Oracle, Consolidated customer view, ETL, Data Stage, Quality Stage, Data Quality, Data Integration Tools, ERP, CRM

Abstrakt

Hlavním cílem teto práce je analýza, návrh, implementace ETL a procesů používaných pro logistickou organizaci, která prošla nedávnými fúzemi a akvizicemi a obstarala několik dalších organizací. Poskytnout konsolidovaný 360-stupňový pohled na svou klientskou základnu standardními datovými prvky které jsou dostupne ve starších systémech. První část práce se zabýva teorií a strategií implementace CRM řešení.

Zbývající část této práci se zabývá návrhem a implementací procesů Extract-Transform-Load používaných k vytvoření jednotného pohledu na zákazníka z propojení stávajících starších systémů v rámci organizace. Tato diplomová práce bude zdůrazňovat potřebu důležitosti robustního systému řízení vztahů se zákazníky v rámci organizace.

Klíčová slova

Siebel, Oracle, Konsolidovaný pohled na zákazníka, ETL, Data Stage, Quality Stage, Data Quality, Nástroje pro integraci dat, ERP, CRM

Table of Contents

List of Figures

List of Tables

Introduction

Topic Definition

Every day organisations are looking for opportunities to expand business through mergers and acquisitions (M&A). Reasons for organisations to merge or acquire other organisations are:

- **Synergies:** in an attempt to gain complementary strengths and weaknesses businesses would seek to acquire another company with the desired market position.
- **Diversification/Business focus:** these two goals are used to describe a merger with the aim of acquisition of a company from another industry. The reason for this kind of acquisition would be the reduction of negative financial impact from the acquired business.
- **Growth:** the company simply acquires a competitor to gain an extra market share, which helps to eliminate the significant effort by acquiring it step by step.
- **Increase Supply Chain Pricing Power:** by acquiring one of the suppliers or distributors a significant saving can be achieved; for example in the case of buying out one of the suppliers there are usually direct savings from cutting out the margins that were added to the cost by the supplier.
- **Eliminate Competition:** purchasing of a competitor and gaining larger market share is an obvious benefit of M&A. There is a negative side of this transaction: significant premium is usually a must to convince the target company to accept an offer. There is also a risk for the shareholders of the buying company to sell their shares in order to lower the price to balance out the high cost of acquisition (Investopedia, 2017).

Along with the benefits and disadvantages of the M&A transactions the buying company has to deal with the established IT infrastructure and customer base (among other factors) of the acquired company. For the buying company it means the alignment and integration of multiple IT systems into a single Customer Relationship Management solution to enable the management of multiple accounts and associated information in a single customer view, whilst retaining the underlying legacy systems and underlying data content and structure.

Reasons for choosing the topic

The reasoning for choosing this topic, is that the M&A has been in place since the late 1800's and has been evolving since this point, which means there is a continuous need to understands and manage information availability and visibility for the benefit and growth of the organisation. The integration of data is now a daily occurrence within most organisations, and by using the correct tools it can be applied in a systemised way, enabling full visibility for the organisations without incurring a major financial outlay (Need for Infrastructure change, Business Processes and New Systems). Another reason is that the use of data integration tools can be used by organisations of any size, as well as the author's hands on experience in this area (Source: Economywatch, 2010).

Scope and goals of diploma thesis

This thesis is intended to show how the correct use of integration and transformation tools and processes can be used within organisations of any size which have completed an M&A to create and manage a complete 360 degree view of the clients to help not only achieve the targets / ROI stated as part of the M&A, but to also ensure that the clients can be managed in a more effective and economical method.

Structure of diploma thesis

This thesis is divided into 2 main sections – theoretical and practical.

The first section covers the reasoning, benefits and design for the implementation of a 360 degree customer view within a CRM solution and the tools and processes to be implemented The second section covers the solution design, and implementation of the Extract-Transform-Load processes required to interface existing legacy systems within an organization´s IT landscape to provide the 360 customer view.

Approach to the solution

To deliver the solution theoretical information and methods are used as well as company documents. Interviews were conducted with the company resources/business owners to understand the selected CRM and the business/data requirements. At the start of the practical section the author sets goals and specifies how to achieve them.

Assumptions and limitations

The implementation would require sub-projects that would need to be completed in parallel and are not in scope of this thesis:

- Evaluation, selection, sizing and installation of a new CRM solution

This thesis will concentrate on the tools and process used to in the standardised and integrated data to create a 360 view of the new client base.

Assumptions:

- The CRM has been implemented and signed off
- All Requirements in terms of data requirements are finalised and documented (data dictionary)

Master thesis outcome

The outcome of this work is the implementation of a set of ETL processes to create a 360 degree customer view from different account data sources/systems. Expected benefit for the organisation is to provide a holistic view of their customers and a decrease in cost.

1 Literature review

Sources used to compose this thesis will be listed and described in this chapter. Literature review covers three sections: in the first one the publications are covered, the second one deals with online sources used and the third one describes used academic theses. All the sources used to create this paper are listed in the bibliography part.

1.1 Literary review of publications

Business Intelligence v podnikové praxi by Jan Pour, Milos Maryška and Ota Novotný, 2012 – this book is dedicated to the main principles of BI including the dimensional modelling and implementation. The book illustrate the cases described using Microsoft SQL server technologies. This publication was useful to gain an overall view of BI and as an introduction to this problem.

Make or Break: The Critical Role of IT in Post-Merger Integration by ITKerney – this study deals with the research of the business and financial impact of mergers as well as gives a guidance on smooth transition for the IT sector of the merged companies. This paper was used for writing the Theoretical part of this thesis.

Oracle Business Intelligence The Condensed Guide to Analysis and Reporting: The Condensed Guide to Analysis and Reporting by Juri Vasiliev, 2010 – this book is a guide to get started with Oracle BI solution. It covers the whole process of BI starting from getting the information from data to reporting. In terms of this thesis was useful as a reference for the practical part.

1.2 Literary review of online sources

Garner IT glossary – is an online glossary of IT terms which is run by the Gartner research and development company. This resource was mainly used to compose the glossary of terms for this paper.

IBM Knowledge Center – this web portal contains all the up-to-date and reliable information about the products offered including InfoSphere software, documentation and practical examples.

MBI – is a portal for Management Business Informatics which includes generalized IT management solutions and development of business informatics. Its purpose is to share the practical and theoretical knowledge in the given area. This source was very helpful while writing the theoretical part about CRM as it provides well-structured information on multiple areas of business informatics topics.

1.3 Literary review of master and bachelor theses

Business Intelligence implementation in the MVNO Company by Alena Kamenchsikova, 2017 – this master thesis covers both theoretical base as well as practical guidance on how to create a BI solution using Microsoft tools.

Analýza a návrh prototypového ře-šení Business Intelligence pro KMPS a.s. by Radim Stralka, 2015 – this bachelor thesis offers a prototypical BI solution with a focus on a financial area, alongside provides general analysis of the company and analyses of source data.

2 Theoretical part

2.1 Current situation

A logistics company has procured multiple smaller organisations to further expand their product offering and their market share within specific regions. With the acquisition there are multiple constraints that would prove to be negative for the new parent company. These are a mixture of quantitative and qualitative items.

One major concern for the parent organisation is that with the acquisition there is a disjointed IT landscape with segregated systems across different technologies, all at various stages of their product lifecycle support by multiple business processes and support organisations.

This therefore makes it difficult for the organisation as a whole to engage with it customers, validate the net worth of each customer; verify the actual opportunity potential across the various product offerings etc.

There is the additional concern of application management and harmonisation of information governance and management as well as extended costs, related to not only IT, but also resource allocation and duplication of efforts which can lead to extended efforts and costs, in addition to bringing a negative view of the company from the perspective of the customer as well as their employees.

2.2 IT's role in post-merger integration

There is a very small margin for a mistake nowadays as the results are expected to be delivered strongly, reliably and quickly. One of the vital components of a smooth transition is IT, which can change the game after the merger has happened. IT brings not only post-merger results but also cannot be ignored in a long term perspective – it can define the difference between a successful and failing merger. IT not only creates a foundation of a new company but also helps to maintain the vital parts of every company in both business operations and customer relationships. Below is the figure pointing out the short and long term roles of IT.

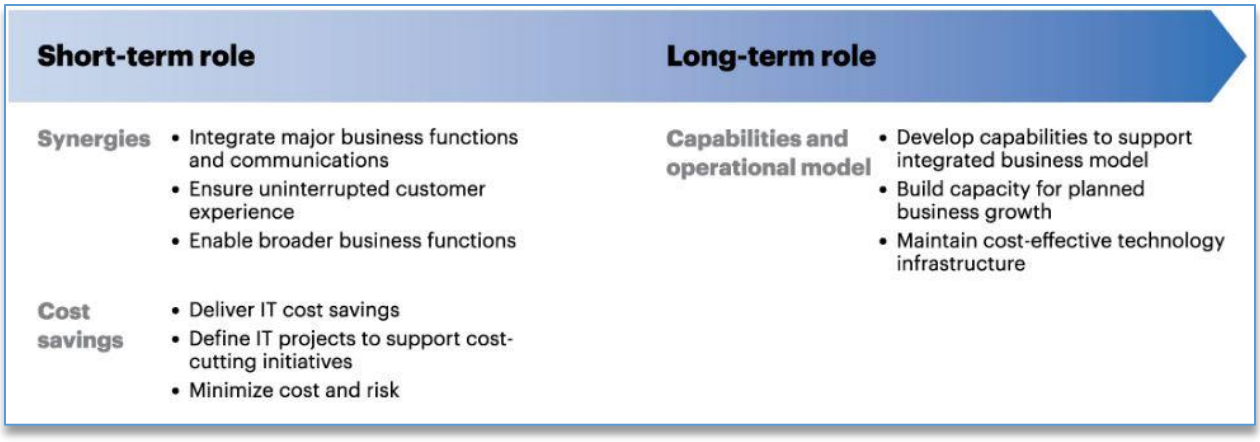

Figure 1: IT's role in post-merger integration (Source: Hagen, 2010)

As we can see from the Figure 1 both short and long-term goals are focused on the cost saving goals. Companies of any size can face hard times especially in after merger, when different systems, applications and resources have to be combined together, there is an increased risk of going into critical situation. Problems connected to poorly managed IT in after merger period can lead to disappointed customers and damaged brand name (Hagen, 2010).

2.3 Perceived Benefits

The main benefit of IT is that it brings a sizeable amount of a post-merger value. It plays a significant role in supporting a business case, which is most likely depends on a significant cost-savings. The savings are achieved by combining the two companies IT cost structures, which helps to plan future cost savings as well as reducing operational risks. Another frequent occasion is that companies find many similarities between IT functions, even if the merged companies do not overlap in terms of industry or product produced. The figure below shows possible potential savings in different categories in IT integration (numbers are valid if based on incremental spending in the merger).

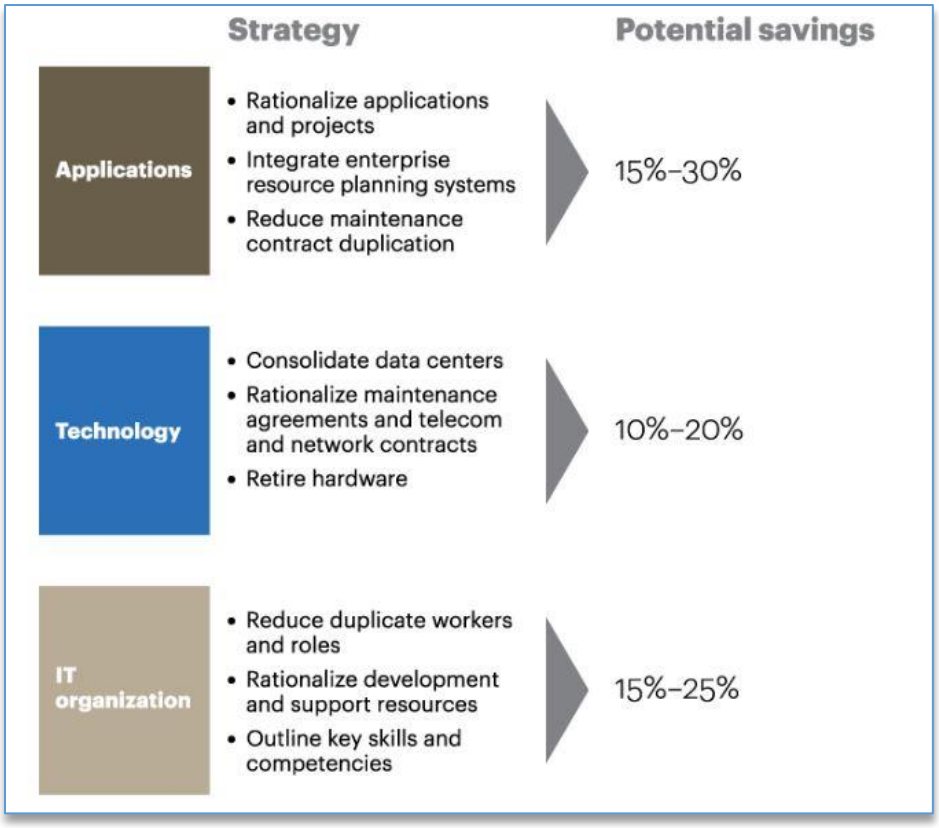

Figure 2: Value categories in IT integration (Source: Hagen, 2010)

From the Figure 2 we can see that companies can save up to 30% by shrinking the portfolios, defining better skills necessary and skills needed for the IT model. In this case IT binds the business together by integrating major business functions, intensifying processes as well as making sure customers receive the continuous service. The last point is crucial as the customers will be the judges to decide if the new company succeeded or not (Hagen, 2010).

There are other detected benefits derived from the M&A. Some of the benefits can be measured in terms of cost – tangible benefits. Several other benefits could be considered difficult to quantify or intangible and these can also be applied to the customer as well as the organisation. Below are listed both tangible and intangible advantages for the company.

Tangible benefits:

- **Revenue Enhancement:** Identification of reasonable revenue enhancement factors due to sales force automation (SFA) such as cross selling, up selling and improved customer retention.

- **Incremental Gross Margin on time saving:** Identification of time saved due to SFA. For example, SFA will save sales representatives time when pulling together information about a particular customer.

- **Savings on Headcount:** Identification of potential headcount cost savings. For example the correct classification of the customer base and ease of access to consolidated customer and financial information.

- **Savings on baseline costs:** Cost savings that result from not having to support and maintain the various existing applications used to support the sales force due to the implementation of a SFA and CRM.

Intangible Benefits:

1. **Customer**

    - Improved customer relationship through the provision of better and more relevant information.

    - Ability to deal with customers more efficiently and effectively.

    - Faster communication and response time, resulting in a better service and increased customer satisfaction.

2. **Organization**

    - Improved communication between departments, as system will contain all customer information, e.g. customer services and finance to sales.

    - Enforces a consistent use of terms in order to compare financial performance of one country to the next, e.g. 'first time buyer' means different things in different countries.

    - Enables a standard way of working - everyone can access the information in a standard way.

    - Improved flexibility - another sales person can manage the account while the usual sales person is away or leaves.

    - Improved employee satisfaction and motivation – more effective use of sales time.

    - Improved speed and efficiency of staff, hence increased revenue, customer satisfaction etc.

    - Improved data accuracy.

    - Enables leads/customers to be 'handed over' more effectively.

    - Better forecasting (linked to accuracy of data/consistent use of terms).

    - Managers will be able to see the 'current situation' on a day to day basis, e.g. what leads have been created or not created.

- Reporting will be more accurate.

Some advantages and potential savings have been listed in this sub-chapter to illustrate a great necessity of understanding not only business perspective and technological capabilities, but also to underline a role of the customer.

2.4 Issues with current set-up

As the author has illustrated possible strengths we cannot underestimate the value of the correct detection of the possible weaknesses, so that they could be taken into the account in the further proposed solution. The problems detected are:

- Extended IT landscape, meaning the large number of existing applications in use.

- IT hardware and software at various stages of the lifecycle and also with various versions for example Win 2003, Win 2008 R2 etc.

- Multiple vendors supporting different applications, infrastructure, software which lead to duplicated efforts in terms of costs (maintenance, licenses etc.) which reduces the opportunity for the organisation to gain synergies and therefore lower costs and effort

- Multiple business processes being used by the various organisations

- No holistic view of the Customer base (360 View) as the data content and visibility is spread across multiple systems

- Data format and content problem: as the systems were owned by multiple organisations, all with their own applications/data stores and business requirements, the ways that data is captured and stored is different and requires manual efforts to collate and standardise information to enable even a simplistic view, with no single source of truth within the organisation, can result in different output from different departments causing confusion and issues at all levels from Top Down to bottom up

- Cost of sales is increased due to the fact that the organisation is segmented. Sales representatives do not have a complete view of the client and sales are completed by the sales teams from all companies within the organization, hence causing duplicated costs (IT HW, Cars, Fuel, and Salaries etc.)

We have to bear in mind that for the successful IT integration there would also be a need for the detailed evaluations of the actual state of the merging IT functions. Applications, organisations, infrastructure and merging organisations preparedness for change have to be assessed. Also the crucial point in understanding would be to develop an understanding of the main business capabilities of the newly combined IT setup.

2.5 Chapter conclusion

The author has dedicated this chapter to the general overview of the default business situation of the company which went through M&A process. The potential benefits and issues were detected for the company's situation, as well as the importance of short and long-term focus on the customer.

3 BI overview

In the previous chapter benefits flowing from the correct IT setup in post-merger period were described as well as the weaknesses of the existing systems being detected. As in the further practical part of this thesis, it will also describe the ETL process enabling the holistic view of the customer. This chapter will provide a general overview of Business Intelligence stages and show the linkage to the Customer Relationship Management.

3.1 BI framework and architecture

One of the multiple definitions of BI could be: BI is the use of computing technologies for the identification, discovery and analysis of business data - like sales revenue, products, costs and incomes (Techopedia, 2018).

BI is divided into different categories according to its level of utilisation:

- **Strategic** – supports long-term company goals, applications include aggregations, statistical analysis data mining. BI of this type is supposed to give senior management a holistic view of the company.
- **Tactical** – supports short-term business decisions, created for business analysts who access and analyse data on a daily basis. Examples of this type of BI applications could be CRM, which is used to analyse customer behaviour and market segmentation.
- **Operational** – supports daily business operations, created to respond to specific operational events. The targeted audience of operational BI is the customer-facing staff (Quaddus, 2015)

If we would look at the BI from the architectural perspective we could break it down to the following main stages and components illustrated in Figure 3.

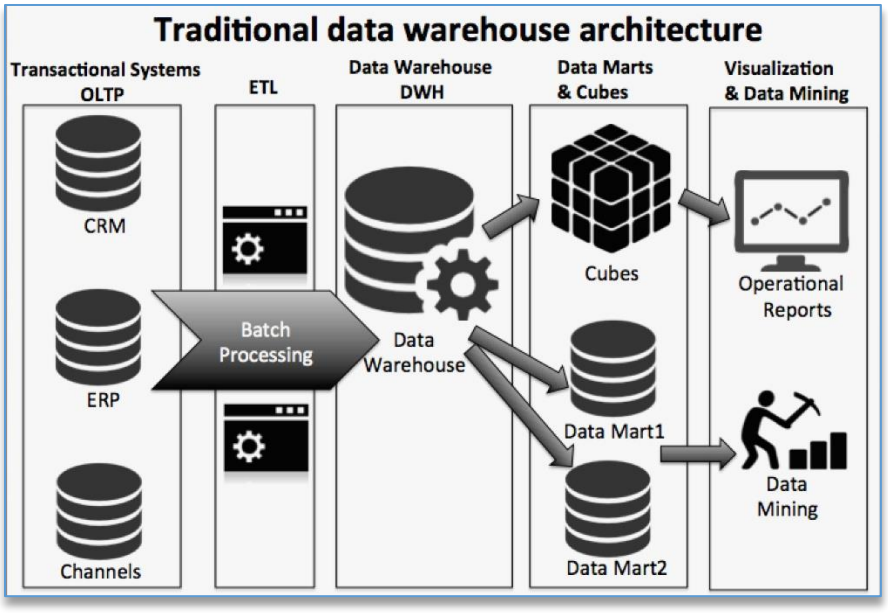

Figure 3: Typical Architecture for a Complex Data Warehouse (Source: Oracle, 1999)

Data sources – could be represented by transactional informational systems (ERP, CRM, HRM), by databases both on-premises and in cloud from various vendors (Oracle, DB2, Informix and so on), by flat files, Excel or Access files as well as it is possible to get data from web services.

ETL – is the pipeline of a data which collects data from different sources, performs a data transformation according to business rules and loads transformed data into a data store. All the phases can run in parallel. During the data processing the staging area (DSA) is used to store the intermediate data. One of the key functions of this area is to consolidate data from different source systems, alignment of the reference data as well data cleansing to detect and update the invalid data.

Data warehouse – this system is used for data analysis and reporting, in other words it could be called a multidimensional source. In that source multidimensional data models are used to perform a complex analysis of historical data. To perform the effective analysis data is organised along dimensions and then used for building cubes. Dimensions included in a cube define its edges or dimensionality (Time, Product and so on). Dimension itself is defined by a set of levels which represents a level of data aggregation. As an example we can have a look at the Time dimension which could aggregate data at yearly, quarterly, monthly, weekly and daily levels (Vasiliev, 2010).

Data mart – in contrast with data warehouse, data mart deals not with multiple subject areas across the company but focused on a single functional area, such as Sales, Marketing, and Inventory. Data is normally drawn from only a few sources. The example of these sources could be a centralised data warehouse, external data from other sources or operational systems. One more key objective of a data mart is it provides users with the most relevant data available in short response time (Oracle, 1999)

Online analytical processing – OLAP systems were designed for extraction of information from business intelligence data in an efficient way as they are optimised for heavy read and low write workloads. To read and process data and perform fast calculations and aggregations the OLAP cubes are used. *"An OLAP cube is a data structure that overcomes limitations of relational databases by providing rapid analysis of data. OLAP cubes can display and sum large amounts of data while also providing users with searchable access to any data points so that the data can be rolled up, sliced, and diced as needed to handle the widest variety of questions that are relevant to a user's area of interest." (*Technet, 2016*).* In order to filter, group and label the data dimensions are used where the data is categorised into the hierarches and categories to allow a more in-depth analysis. Dimensions may also have natural hierarches to allow users to "drill down" to more detailed levels of detail (Technet, 2016).

Figure 4: OLAP Cube (Source: Frattelone, 2012)

Reporting – is a multi-perspective view on a processed dataset the goal of which is to enable the end-user to understand the data through visualised aids like summarised and structured reports and dashboards of a different kind. Usually focused on one specific set of data, it presents a trend and gives a solid base for identifying a potential problems.

Having a stable BI architecture helps companies to have a better overview and control not only the implementation but also the operation of the entire BI environment. The five layers described in the above text are important to make sure company has a good data quality and smooth information flow. (Ong, 2011)

3.2 CRM types

In terms of software solution design, CRM applications are a purposeful combination of transactional, analytical, and infrastructure applications. This combination is manifested by the existence of three basic

functional parts of CRM applications, which are the operational, cooperative and analytical part. All three parts are closely connected.

Operational CRM deals with the everyday customer related activities which are customer search and retention. This area is focused on:

- **Sales force automation (SFA)**, where the agenda of sales representatives is managed. The example could be contact, opportunity and potential customers and order management.
- **Marketing Automation System (MAS)**, supports activities oriented towards support of one-to-one marketing in internet, management of internet campaigns and their evaluation.
- **Customer Service and Support (CSS)**, this area is focused on providing detailed product information, complaints management, warranty services and post warranty services. The trend is to deploy this kind of applications as self-service applications, so that customers can get all the product information.

Cooperative CRM applications are focused on managing the communication channels both with potential and loyal customers. All the communications are managed through call centres, where all possible customers' related information is gathered. Call centres provide such services as support, automatic voice responses, and marketing campaigns.

Analytical CRM collects customer information from every customer activity and transaction: operational, transactional, interaction, customer profile, behavioural or geographic. Analytical CRM uses the data collected in operational and cooperative parts or another source - ERP systems. All this data has to be analysed and processed to achieve business process quality. To provide business with analysis BI applications are used. This kind of combination is called Customer intelligence. CI represents functionality focused on customer knowledge, value and likelihood of leaving.

To add another dimension to the above mentioned types of CRM there is a so called **Social CRM (sCRM)** technology. It was developed due to the increased social network usage, so companies had to look for a way how to adjust a new way of communication with customers through social networks. sCRM is oriented on including the customer into conversation within the community and context, which is driven by the customer (Gala, 2015)

There are multiple solutions for any type and size (from Enterprise level to small business) of a company is offered on the market. Among the companies which develop CRM solutions are Microsoft, Oracle and Salesforce. The solution chosen for the purposes of practical part of the thesis is introduced later in Proposed solution part.

3.3 Correlation of BI and CRM

Since the general concepts of BI architecture were introduced in previous text, in this chapter the author will attempt to give the overview of the relationship between ETL and CRM technologies.

As we can see from the previous chapter, data warehouse plays a role of a repository for customer related data. Customer data are used to identify and improve such areas as selling opportunities, inefficiencies and improve retention of existing customers to build a 360-degreee view around the customer. Data warehouses which fulfil CRM's basic requirements of granular customer transaction data are known as customer data repositories (CDR). Effective CRM collects data at every customer interaction and then analyse it for future improvements. The data warehouse becomes the repository for all customer information from all sources. The key focus of a data warehouse is to support an enterprise decision support system and is not restricted to a specific Line of Business.

Data warehousing and business intelligence solutions are the key to customer identification. Companies plan to enhance the ability to better understand their customers. Better customer identification can aid in profiling best customers and the rate at which they are buying products. Trend information gathered can eventually lead to making better business rules, marketing strategies and trained sales forces. Another benefit data warehouses try to accomplish is to understand customer profitability. When a customer receives benefits, the company profits automatically. In short, availability of ERP-driven information provides enhanced customer relationships, identification of new products and services and improved market segmentation (Khan, 2011).

3.4 Chapter Conclusion

In this chapter the author has provided the reader with the basic architectural concepts of BI, introduces the importance of data quality as well as to underline the BI-CRM correlation to achieve the holistic view of the customer.

4 Proposed solution

There were several solutions proposed to resolve the current situation ranging from mandating all acquired companies to migrate to the parent organizations applications, this was rejected due to the multitude of complexities, timelines and cost associated to make the change.

The decision was to purchase an off the shelve CRM solution which would be possible to use only required modules and, as needed, expand the solution to incorporate other needs such as Marketing. CRM normally consists of sales force automation, marketing automation and customer support. As part of ERP, CRM is one of the five ERP pillars.

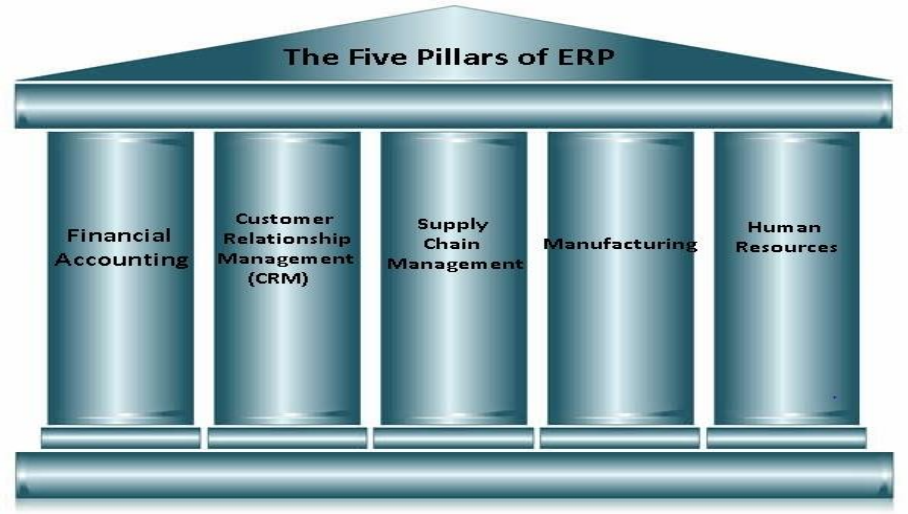

Figure 5: Five pillars of ERP (Source: DCS, 2015)

To enable the implementation of a new CRM solution, there is an underlying need to review and agree the following:

- Would the application be business process driven or process independent
- What would be the actual scope of the initial implementation
- What are the Organisations data content
- What should be the target data model

To complete the steps listed above, several work streams are required under the umbrella of a programme to enable the correct analysis, then document all findings and recommendations in regards to which is the correct approach and solution that should be implemented, citing the costing savings that can be achieved.

There were multiple workshops to agree on what should be the design for the new CRM solution. Based on the requirement from the organisation that will enable interaction for the clients at a consolidated view, there was a need to ensure that certain goals and targets can be realised, the interaction with the client is streamlined and that there is a reduction in the duplication of the contact by the account management teams. The modules to be implemented are Sales CRM which consists of the following subcomponents:

- Accounts
- Address
- Contact
- Revenue
- Opportunities

There is a single component that is missing from the above that will bring all the data together to enable the 360 view is Customer. The CRM selected was Siebel Sales which is part of the Oracle offering as it is amongst the market leaders within this segment (see Figure 5 below).

Figure 6: Magic Quadrant CRM 2017 (Source: PEGA, 2017)

Siebel Sales is designed to improve pipeline visibility, sales effectiveness, and bottom-line results; Siebel Sales enables your organization to share information across teams. Oracle's Siebel Sales is fully integrated

with the entire Siebel product family, including CRM On-Demand enabling flexible, phased deployments for constantly changing and growing companies (Oracle, 2016)

Table 1: Features and benefits of Siebel CRM (Source: Oracle, 2017)

Features	**Benefits**
Account Management	Provides a comprehensive, 360 degree view of your customer, including service history, order management, interactions, and account profile
Opportunity Management	Including management of leads, territories, opportunities, contacts, and all account activities
Sales Methodologies	Standardize on common best practices to ensure consistent sales performance and sales coaching throughout the sales cycle
Sales Forecasting	Including real-time insight into sales and employee performance
Order Management	Allows you to create quotes, proposals, and product configuration
Territory Management	To pipeline leads and more
Integration to Microsoft Applications	Siebel Server Sync for Microsoft Exchange Server enables employees to centralize customer information across Microsoft Outlook and Siebel applications

Siebel Sales can also be integrated into other Oracle offerings such as Oracle Business Intelligence Suite (OBI EE) or other business intelligence solutions which will allow a single source of reporting across the organization and provide valuable overview and status the customer based to not only the sales resources but to other divisions with the organisation. Even without implementing an analytical solution with the Sales CRM if there is a consolidated view of the customer then this will still enhance the organizations possibilities to:

a. Reduce Cost of sales with the sales forces (no longer multiple sales selling different products).
b. Enhance/solidify the customer relationship by creating a single point of contact for all needs.
c. Ability to understand the customer in terms of current revenue and needs across all products.

d. Ability to manage all opportunities and provide strategic pricing based on all usage and not a silo approach due lack of data consolidation.

As the IT landscape and the CRM solution has to be preordained there several questions that need to be addressed to enable a successful implementation:

a. What is the available schema from Siebel Sales (Out of the Box) for the required components?
b. What are the data elements that are required additional by the organization?
c. What is the data schema from the source systems?
d. How will the data mapping be completed?
e. What is the validity of each data element?
f. How can a 360 customer view be completed?

4.1 Design applied

With the CRM and IT Infrastructure agreed, there is a need to validate what process/tool should be used to enable the collection, standardisation, integration of data from the individual legacy systems to the new Standardized CRM data model to provide not only a standardised data set, but a single 360 degree view of the Customers and their associated data.

Data integration involves a framework of applications, techniques, technologies, and products for providing a unified and consistent view of enterprise business data.

Applications are custom-built and vendor-developed solutions that utilise one or more data integration products.

a. Products are off-the-shelf commercial solutions that support one or more data integration technologies.
b. Technologies implement one or more data integration techniques.
c. Techniques are technology-independent approaches for doing data integration

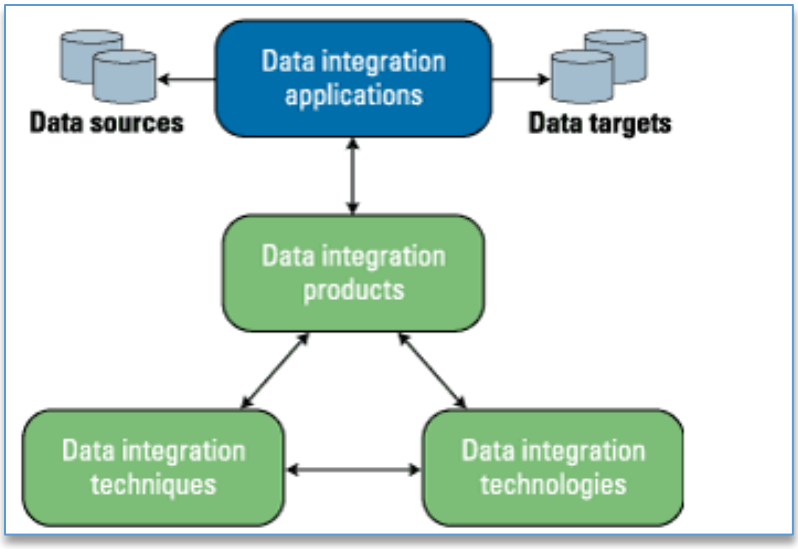

Figure 7*:* Data Integration (Source: White, 2003)

Following is the short description of the Figure 8 components used in data integration projects. **Data integration techniques** can be divided into three main groups:

- **Data consolidation** captures data from multiple systems and integrates is into a single data store. With data consolidation there is often latency which is connected to a time delay in between data transition from the source to the target system. Depending on business needs it can be a few seconds or many days.

- **Data Federation** provides a unified view of source data files, it always *pulls* data from source systems on an on-demand basis. As data is retrieved from source any data transformation is done. Enterprise information integration (EII) is an example of a technology that supports a federated approach to data integration.

- **Data Propagation** applications copy data from one location to another. These applications usually operate online and *push* data to the target location; in other words they are event-driven. Updated are represented by synchronous and asynchronous propagation. Regardless of the type of synchronization used, propagation guarantees the delivery of the data to the target. Enterprise application integration (EAI) and enterprise data replication (EDR) are examples of technologies that support data propagation.
- **Hybrid Approach** is used quite often due to the business and technology requirements

Data Integration technologies are represented by four main pillars:

- **Extract, Transform, Load (ETL)** – this technology was already mentioned in BI overview chapter. To add in context of Data integration scope is that ETL supports a consolidation approach to data integration. Data transformation may involve data record restructuring and reconciliation, data content cleansing or data content aggregation. Data loading may cause a complete refresh of a target data store or may be done by updating the target destination. Interfaces include de facto standards like ODBC, JBDC, JMS or native database and application interfaces.

- **Enterprise Information Integration (EII)** according to (White, 2003) "*provides a virtual business view of dispersed data. This view can be used for demand-driven query access to operational business transaction data, a data warehouse, and/or unstructured information. EII supports a data federation approach to data integration. The objective of EII is to enable applications to see dispersed data as though it resided in a single database. EII shields applications from the complexities of retrieving data from multiple locations, where the data may differ in semantics and formats, and may employ different data interfaces*".

- The next technology is represented by **Enterprise Application Integration (EAI)** which is supposed to integrate application systems by allowing them communicate, exchange different kinds of messages and information. Data transformation and metadata features in an EAI are designed to process simple transactions and message structures, and they cannot support the complex data structures handled by ETL products. In this regard, EAI does not compete with ETL (White, 2003).

4.2 Data Requirements

For each entity that is required within the CRM there needs to be a full understanding of the source as well as what are the requirements within the new system, these are for example:

a. Source schema
b. Target schema
c. Data content
d. Data transformation
    I. New agreed values
    II. List of Values (LOV's)
e. Standardisation rules / requirements
f. Data retention (how long should data types be retained / should it be migrated)
g. Data validity (What is valuable)

There is a need to complete a detailed analysis of each source system and where possible standardise as much as possible, and in other cases collaboration is required between the IT focused teams and the country / regional business teams to identify the correct transformation.

4.3 Data mapping

As there is an available data source and a target application and a requirement by the author to complete a one-off data mapping exercise for each source system from source to target, this will be completed as part of an initial migration. For this chapter only first seven rows of each table are captured. To see the complete table fields please see the Attachment.

Table 2: Customer Data Fields (Source: Author)

Customer Data Field	Data Type	Description
Customer ID	Char	Unique Identifier for a Customer Record. Country Code and C to be prefixed for a sequence number generated.
Customer Name	Char	A customer is a trading entity that actively generates revenue with organization and is located at a specific site, and will have one or more separate agreements with the organization
Telephone Number	Char	Should be a 10 or more digit number. "+" and Country code to be prefixed to the telephone number. For ex., for Norway, +4798435675 would be an example of a telephone number
Industry Code	Char	The Industry code assigned to this customer. This is a 2 digit industry code.
Sales Territory Code	Char	The Sales Territory the customer belongs to. This is the 'Managed' territory information.
Call Frequency	Number	The call frequency schedule this customer has been assigned
Customer Sales Stage	Char	The Sales stage the customer is in at the moment

Table 3: Account data fields (Source: Author)

Account Data Field	Data Type	Description
Customer ID	Char	Uniquely identifies a Customer Record
Customer Name	Char	Uniquely identifies a Customer Record. Should be the same Customer Name as available in the Customer File for this Customer ID.
Account ID	Char	Required for maintaining the Customer – Account Relationship within COMET.
Account Number	Char	The unique identifier of this account. This is the account number generated in the legacy billing systems
Account Name	Char	Account Name
Parent Account Number	Char	Master Account Number of an account.
Credit Stop Flag	Boolean	Indicates that this account is on credit stop

Table 4: Address data fields (Source: Author)

Address Data Field	Data Type	Description

Record Type	Char	Indictor if the Address is for a Customer (C) or Account (A)
Customer Id / Account Id	Char	Uniquely identifies a Customer (or) Account Record
Customer/Account Name	Char	Uniquely identifies a Customer (or) Account Record. Should be the same Customer/Account Name as available in the Customer/Account File for this Customer ID.
Address Unique Id	Char	Uniquely identifies the address. It should have the following format -
Address Type	Char	The role for the address type. Number of address will differ per country
Address Line 1	Char	The first line of the address specification
Address Line 2	Char	The second line of the address specification

Table 5: Contact data fields (Source: Author)

Contact Data Field	Data Type	Description
Record Type	Char	Indictor if the Contact is for a Customer (C) or Account (A)
Customer Id / Account Id	Char	Uniquely identifies a Customer (or) Account Record
Customer / Account Name	Char	Uniquely identifies a Customer (or) Account Record.
Contact Unique Id	Char	Unique Identifier to identify the contact.
Contact Last Name	Char	Contact last name
Contact Type	Char	Type of the Contact; Billing, Shipping, main, Sales, Pick up
Contact First Name	Char	Contact first name

Table 6: Activity data fields (Source: Author)

Activity Data Field	Data Type	Description
Customer ID	Char	Uniquely identifies a Customer Record
Customer Name	Char	Uniquely identifies a Customer Record.
Priority	Char	Priority of the Activity
Activity Type	Char	Activity Type
Activity Objective	Char	Activity Objective.
Activity Purpose	Char	General Purpose of the Activity; Maintenance, acquisition, penetration, retention

Planned Date	Date	If Activity Status is "Scheduled", then Planned Start date to be defaulted to current date (System Date).

Table 7: Opportunity fields (Source: Author)

Data Field	Data Type	Description
Customer ID	Char	Uniquely identifies a Customer Record
Customer Name	Char	Uniquely identifies a Customer Record.
Opportunity Name	Char	The name of the opportunity. This should reflect the opportunity and is unique under the customer site.
Opportunity Type	Char	Indicates the nature of the opportunity. E.g. penetration, acquisition
Potential Revenue	Number	The potential revenue of this opportunity.
Committed Revenue	Number	The committed revenue of this opportunity.
Pipeline Stage	Char	The stage this opportunity has reached.

4.4 Entity Relationships

The below shows the basic data relationships that are to be implemented in the CRM, this does not cover standard tables/extensions found in the out of the box CRM solution. The entity relationship does not reflect the overall database schema.

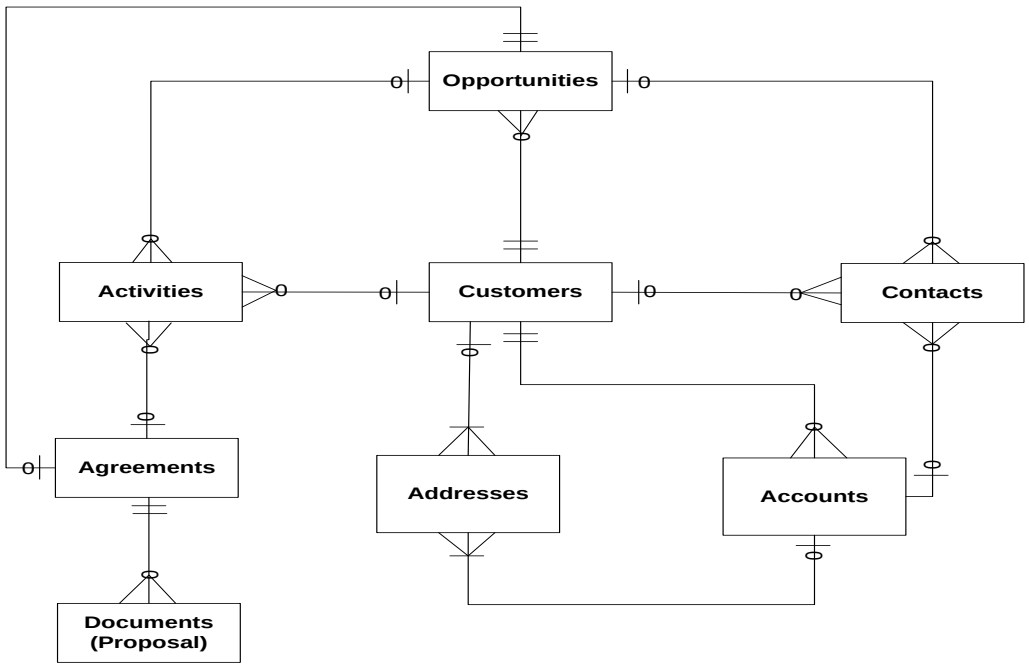

Figure 8: 360 Customer Entity Relationships (Source: Author)

The entity relationship shows the different relationships between each entity for example:

- Contacts/Addresses can be child records to both Customers and Accounts.
- Opportunities/Development Leads can only be the child record of a Customer.
- Activities can be linked to Customers, Accounts, Contact or Opportunities.
- Customer to Account is 1 to many relationship.

The tool selected for the data integration is part of the IBM WebSphere suite and the 2 main products are DataStage and QualityStage:

- It is already used by the parent organisation hence there are available skill sets available within the organisation without having to procure external resources
- IBM are one of the market leaders in regard to Data Integration Tools (see Figure 8 below)

Figure 9: Magic Quadrant Data Integration Tools 2017 (Source: Informatica, 2017)

4.5 Chapter Conclusion

The author shows in this chapter the importance of a 360 degree customer view within an organization was described in this chapter. Organizations that expand and increase their market share by M&A need to ensure that all the various data components are integrated using a standardised process with pre-defined data models and processes. This will enable an organisation to manage not only their Client base but also their data to ensure data quality and reliability to help further increase their business by the understanding of the Customer and their needs, as well as decrease in organisational costs to manager their customer base.

With the correct management and understanding of the customer data it enables the organization to provide quality data analysis and reporting.

5 Practical part

5.1 General description

In this chapter the author identifies, designs and develops the ETL solution that is required to enable the logistic company to have a 360 degree view of all their customers and enable a sustainable and scalable solution for the integration of the new acquisitions as well as changes within the existing organisational environment. Author will concentrate on the standardization and matching of data to convert data from multiple account systems from different organization into a single customer.

To enable the correct set-up there is a need to set-up 2 different processes:

- Initial Migration
- On-Going interfaces

To enable the correct collation between account data and final customer even for an initial migration there is a need for defining several different aspects.

These are not only relevant for the initial migration but also for the on-going interfaces. These are outlined in the next sections:

- The initial migration is to be run to create initial group of customers. It is a one-off process and once this has been completed the on-going interface process can be implemented.
- The on-going interface process enables the daily operations to be managed as per the organisational requirements in relation to not only managing a single customer view. In additional it enables the inclusion of new account data as well as managing new opportunities.
- As part of the initial migration an initial exercise will be complete to validate mappings. These mappings, which are named List of Values (LOV's), are required from difference source systems to the new CRM solution to enable a consolidated and atomised data set where needed.

The practical part will concentrate on the business as usual process as this addresses the criticality of not only the one-off single customer view but having the ability to maintain that view in an automated and systemized way.

After the initial migration there will be a set of internal reference files in place. These are hash files and are used in the processes to ensure the correct linkage of data is maintained and used as the correct reference points. The overall high level process in regards to the extraction, standardisation and matching processes is illustrated below. There is a feedback loop with the individual company that is in scope. It is needed to note that the below process is only a sub-set of a larger ETL process.

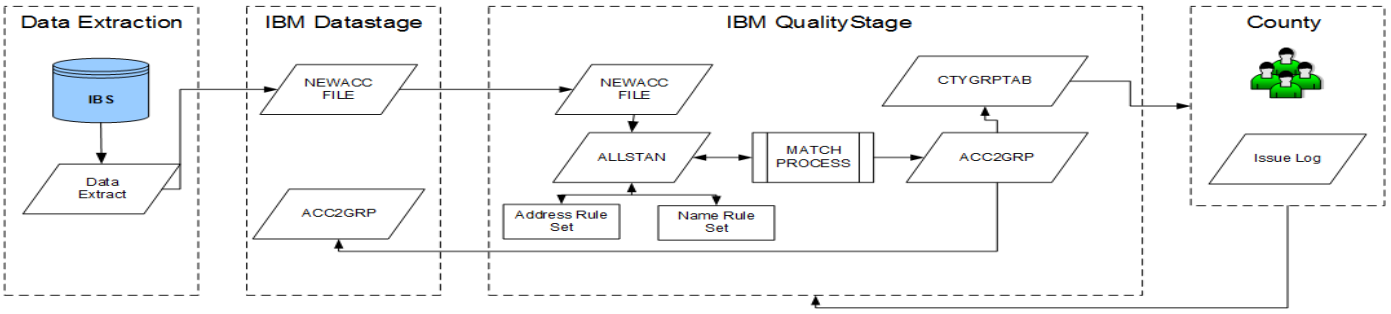

Figure 10: ETL - High level process (Source: Author)

The process can be separated into 6 phases:

- **Extract –** The process for pulling the information from the legacy Customer Account Handling (CAH) systems
- **Analyze –** The process for understanding the data structure, format etc.
- **Data patterns –** Understanding the various data patterns to enable the correct standardisation of data
- **Standardise –** The process for taking the original data and creating a standardised interim file
- **Matching Algorithms** – The rules / logic used to match the different data elements
- **Output / Results** – The original data with the grouping results

The below shows a more detailed view of the overall process used which will encompass all the above phases.

Grouping Process

Extract

Data Extraction

NEWACC - Fixed Width file
- Source
- Acno
- Custid
- Acname
- Adtyco
- Coad1
- Coad2
- Coad3
- Coad4
- Coad5
- Ccode
- Pcode
- City
- Bldname
- Distric
- Provinc
- Stname
- Pobox
- Phone
- Vatno

- Standardised input file for grouping
- Different field avalibel for company data to encompass most countries data structures

Standardisation

Account Name Fields
Account Name
(All processed as organisations)

NAME RULE SET
- Classifications (Abbreviations / common words/ etc...)
- Pattern action (Hard code)

Address Fields
Coad1
Coad2

ADDRESS RULE SET
- Classifications (Abbreviations / common words/ etc...)
- Pattern action (Hard code)

Area Fields
City
Pcode

AREA RULE SET
- Classifications (Abbreviations / common words/ etc...)
- Pattern action (Hard code)

Telephone Fields
Phone

TELEPHONE RULE SET
- Classifications (Abbreviations / common words/ etc...)
- Pattern action (Hard code)

ALLSTAN FILE
- Standardize file
- Output is fix fielded addressable data
- Phonetics (NYSISS, Soundex)

Name (Example output fields)
Account name word 1
Account name word 2
Packedkey account name (1st 20 chars)
NYSIIS account word 1
NYSIIS account word 2

Address (Example output Fields)
Street Name
NYSIIS Street name
Building Number
Sub Building Number
Sub Building description
PO Box Number
PO Box descripter

Area (Example output fields)
City Name
NYSIIS City Name
Postal Code

Telephone (Example output fields)
Telephone number

- Countries define fields to be used in the grouping process
- Data patterns Will differ between countries
- Grouping can and in most cases will differ between countries
- Country / Group specific classification files and pattern action files
- Pattern action files look at the patterns of the data, the system will not resolve data quality / entry issues For example

Account Name	TURBORCO	INDUSTRIAL	LIMITED
Pattern	+	W	O
Descripton	Unknown	Non-Individual Word	Organisation Name Suffix

- Not all patterns will be handled
- Countries should ensure data entry standards algined across different systems

Grouping (Example)

Grouping Pass 1 – Blocking Fields
- NYSIIS Account word 1
- NYSIIS Account word 2
- NYSIIS Street name
- Building Number

Conditins Applied
- None

Grouping Pass 2 – Blocking Fields
- NYSIIS Account word 1
- NYSIIS Street name
- Building Number

Conditins Applied
- NYSISS Account word 2
- Sub Building number

Grouping Pass 2 – Blocking Fields
- NYSIIS Account word 1
- NYSIIS Street name

Conditins Applied
- NYSISS Account word 2
- Sub Building number
- Building number

- Upto 7 passes of the data can be completed
- Groups can be modified based on feedback from validation

Output

ACC2GRP:
NEWACC with group key
- Source
- Acno
- Custid
- Acname
- Adtyco
- Coad1
- Coad2
- Coad3
- Coad4
- Coad5
- Ccode
- Pcode
- City
- Bldname
- Distric
- Provinc
- Stname
- Pobox
- Phone
- Vatno
- DF (Group Key)

- Original input records provided back with a group key
- Tab delimited version delivered to the country for validation purposes

Figure 11: End to End grouping process overview (Source: Author)

5.2 Extract

The extract phase of the process designed by the author refers to the pulling of the account related data from the customer account handling (Customer Account Handling) for example the billing system or ERP systems. The naming convention for that system "CAH" is taken from the main organisation and applied to the acquired business as all organisations can use different naming conventions for their Account and Customer systems.

This was applied to create a unified input file to enable subsequent phases of the whole process. There are multiple key fields that are required in the extract phase

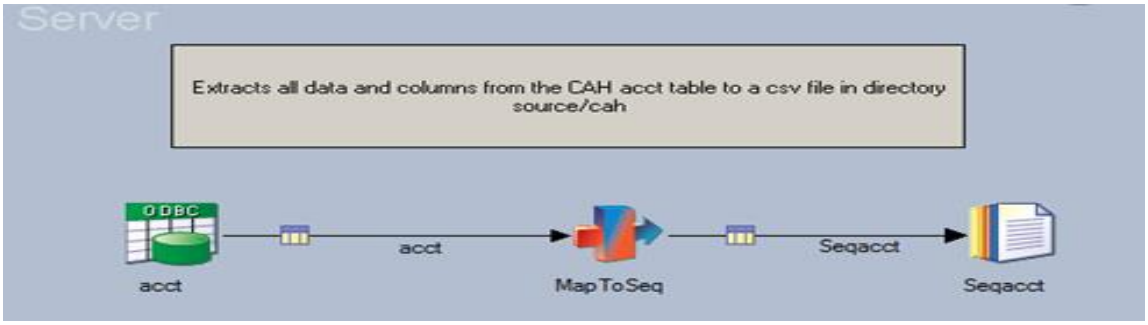

Figure 12*:* DataStage Extraction job (Source: Author)

In order to process the initial account data there is a need to use a connector to the source system. In this case the author used the ODBC connector. With the help of this connector we can pull the data from the data ODBC source which contains all the technical level data for the accounts.

Then the data from the source system are (cleaned, trimmed, matched) ... with the help of the transformation stage (named Map To Seq on Figure 6)

After the data is transformed it is written to the file (see Seqacct on Figure 6). This file represents the extract of the individual account information from the various account sources. It will also be used by the author in the next process. Depending on the amount of sources the author can use multiple extraction jobs to get all the necessary data for further processing. The output of all the extraction jobs are consolidated in the next stage to create a single account file on the picture below it is called the InterimAccount file.

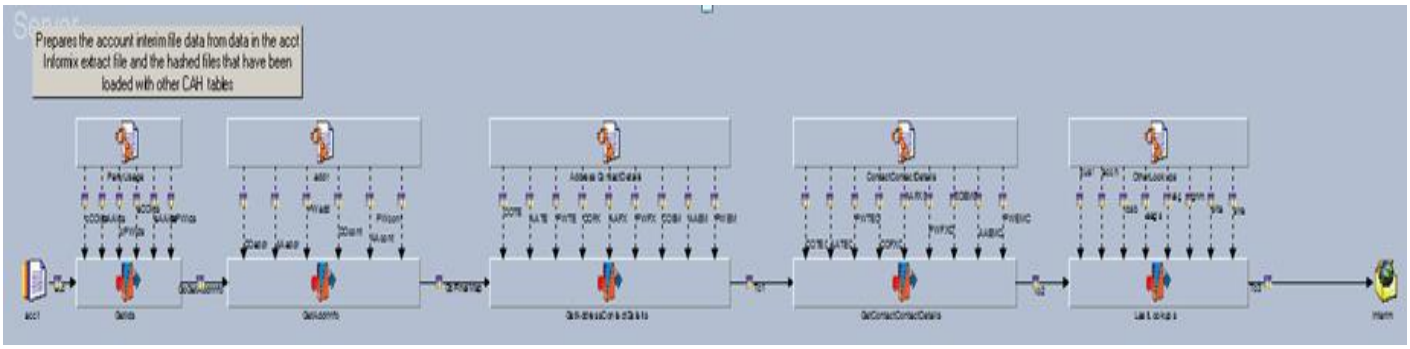

Figure 13: Consolidation of data sources (Source: Author)

The InterimAccount file is then carried into the next stage where specific filters and look-ups are applied based on the Hash files that were created during the initial migration into the CRM. A Hash file is a specific set of files that contain either key reference data for example:

Hash_valid_stc – This is a reference file that maintains a look-up list of valid Sales Territory Codes (STC) that enable data that is not required to be filtered (for example accounts that are not required to be managed by the sales team or are not valid).

Hash_cah_delta – This is a reference set of accounts that are already available within the CRM application. These are accounts that have been processed previously within the tool (Initial Migration, previous on-going interface runs)

This Hash file enables only the delta between the sources extract to be taken into account for the next stage. This also stops duplicate from coming in the system as there is a fixed reference point within the process for anything previously seen.

The result of this job is the consolidated package of all accounts after extraction and filtering. Further process will involve this package to complete multiple further transformations in filtering job.

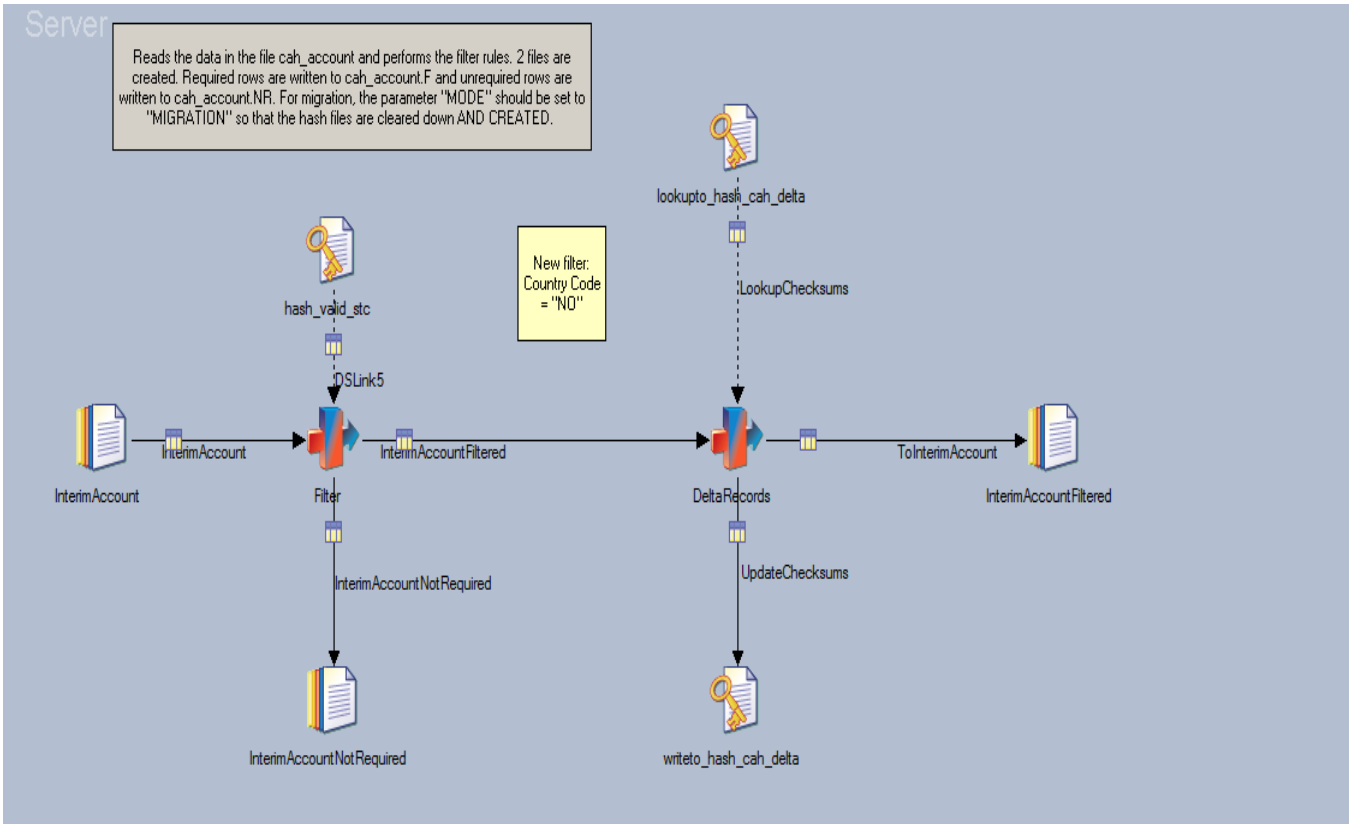

Figure 14: InterimAccount filtering (Source: Author)

After the necessary transformations, the output of this job is the file (see InterimAccountFiletered on figure 8), which contains the data fields (according to the Table 6 below) and is used by the author for the Analysis, Standardization and Matching phases. This file will be called NEWACC, this file is a static format in regards your structure and content.

The author then completed a one-off exercise to map the data from the original CAH systems to the NEWACC format. As the NEWACC file is a standard format, it simplifies the process for duplicating the process for implementation within any other country.

Table 8: NEWACC (Source: Author)

Field Name	Description
SOURCE:	This is the identifier to enable the organization to trace the account back to the original legacy source system
ACNO:	This is the account number that is populated from the original legacy source system and is used by other legacy systems still and is also used as a unique identifier across other data sources and business intelligence solution to enable the complete 360 degree customer view without impacting legacy reporting and billing etc.
CUSTID:	Customer identification number is the unique number that is assigned to each record and is populated after the matching process has been completed
ACNAME:	Account name is the name from the legacy system and is maintained

ADTYCO:	Type of address and is a standardized list and the available values are Billing, Shipping, Main
COAD1:	First line of the address from the legacy system
COAD2:	Second line of the address from the legacy system
COAD3:	Third line of the address from the legacy system
COAD4:	Fourth line of the address from the legacy system
COAD5:	Fifth line of the address from the legacy system
CCODE:	The Country code related to the account
PCODE:	Postal Code / Zip code related to the account
CITY:	City related to the account
BLDNAME:	Building name where the account is location
DISTRIC:	District of where the account is located
PROVINC:	Province of where the account is located
STNAME:	Street name related to where the account is located
POBOX:	Postal Box related to the account
PHONE:	Telephone number of the account
VATNO:	VAT Number
BYTE3CC:	2 byte country code

5.3 Analyse

In order to ensure that the account level information is correctly matched together to collate the customer information there is an analysis that was undertaken by the author to understand the different content of the individual fields as well as the actual structure of the data.

As the main matching criteria from Account to Customer is related to the name and address of any specific organizations the analysis relates to the main fields that are identified as key attributes within the matching process predominantly:

- Name
- Address Line 1
- Address line 2
- City
- State (If applicable)
- Province (If applicable)
- Postal / Zip Code
- VAT Number

For example the address format for the United Kingdom is different to that of France and there are even bigger differences between continents. The first stage of the analysis after the NEWACC file has been created is to verify is there is a default country rule set. If there is a standard rule set available then this will be used as the starting point by the author.

The approach taken by the author for matching is sampling where she works with a small, manageable amount of data in order to build and run analytical models more quickly, while

still producing accurate findings. Sampling can be particularly useful with data sets that are too large to efficiently analyse in full - for example, in big data analytics applications.

An important consideration, though, is the size of the required data sample. In some cases, a very small sample can tell all of the most important information about a data set. In others, using a larger sample can increase the likelihood of accurately representing the data as a whole, even though the increased size of the sample may impede ease of manipulation and interpretation. Either way, samples are best drawn from data sets that are as large and close to complete as possible.

The analyse phase is also iterative and will also continue during the subsequent processes such as 6.4 Data patterns and 6.5 Standardise and 6.6 Matching Algorithms.

When looking at the data it is also imperative to understand what data can be considered as specific to the data set in question within the various rule sets being used, specifically the Name and Address rule sets. This can encompass various types of data:

- Common Words
- Abbreviations
- Company types

Once the analysis is completed the output can be used by the author in the Standardize phase.

Table 9: Sample Data (Source: Author)

ACNO	DF	ACNAME	COAD1	COAD2	PHONE	PCODE	CITY
950343952	621	3S SILICON TECH INC TW	NO 169-2 SEC 1	KANG LO RD HSIN FENG HSIANG	03-5577668	304	HSINCHU HSIEN TAIWAN
620023425	621	3S SILICON TECH. INC.	NO 169-2 SEC 1	KANG LO RD HSIN FENG HSIANG	03-5577668	304	HSINCHU HSIEN TAIWAN BH00140
620022576	579	A & J ENT CO LTD	22F-2 NO 447 SEC 3	WEN HSIN RD	04-22971998	406	TAICHUNG TAIWN AX01250
963743035	579	A & J ENT CO LTD TW	22F-2 NO 447 SEC 3	WEN HSIN RD PEI TUN DIST	04-22971998	406	TAICHUNG TAIWAN
620012340	296	A C P (TAIWAN) INC	4F NO 267 SEC 3	CHEN TEH RD	02-25979153	103	TAIPEI TAIWAN AP01120
951233768	296	A C P (TAIWAN) INC TW	4F NO 267 SEC 3	CHEN TEH RD	02-25979153	103	TAIPEI TAIWAN
620016090	416	A ONE UNION CO LTD	5F NO 99 SEC 3	NAN KANG RD	02-27882828	115	TAIPEI TAIWAN AP14100
950671077	416	A ONE UNION CO LTD	5F NO 99 SEC 3	NAN KANG RD	02-27882828	115	TAIPEI TAIWAN
620022307	569	A.T.G. SOURCING LTD (TAIWAN LIAISON OFFICE)	9F NO 152 SEC 1.	CHUNG KANG RD	04-23211911	403	TAICHUNG TAIWAN AX00590
962264728	569	A.T.G. SOURCING LTD (TAIWAN LIAISON OFFICE)	9F NO 152 SEC 1.	CHUNG KANG RD	04-23211911	403	TAICHUNG TAIWAN AX00590
620013947	353	ABA UFO INTERNATIONAL CORP.	3F NO 649-5	CHUNG CHENG RD HSIN CHUANG CITY	02-29064736	242	TAIPEI HSIEN TAIWAN AP06290
951232538	353	ABA UFO INTERNATIONAL CORP. TW	3F NO 649-5	CHUNG CHENG RD HSIN CHUANG CITY	02-29064736	242	TAIPEI HSIEN AIWAN
620013848	351	ABACUS DISPLAY INFINITY CORP	5F NO 131 SEC 3	NAN KING E. RD	02-27180895	104	TAIPEI TAIWAN
620601221	351	ABACUS DISPLAY INFINITY CORP	5F NO 131 SEC 3	NAN KING E. RD	02-27180895	104	TAIPEI TAIWAN
962268902	351	ABACUS DISPLAY INFINITY CORP TW	5F NO 131 SEC 3	NAN KING E. RD	02-27180895	104	TAIPEI TAIWAN
620014445	366	ABB LTD	NO 11	WU QUAN 5 RD WU KU DIST	02-22993299	248	NEW TAIPEI CITY TAIWAN
620014458	366	ABB LTD	NO 11	WU CHUAN 5 RD	02-22993299	248	TAIPEI HSIEN TAIWAN
620013484	338	ABBOTT LABORATORIES SERVICES CORP	6F NO 51 SEC 3	MIN SHENG E. RD	02-25050828	104	TAIPEI TAIWAN AP04090
950345820	338	ABBOTT LABORATORIES SERVICES CORP	6F NO 51 SEC 3	MIN SHENG E. RD	02-25050828	104	TAIPEI TAIWAN
620711856	201	ABC TAIWAN ELECTRONICS CORP	NO 422 SEC 1	YANG HU RD YANG MEI CHENG	03-4788188	326	TAOYUAN HSIEN TAIWAN
620004884	201	ABC TAIWAN ELECTRONICS CORP.	NO 422 SEC 1	YANG HU RD YANG MEI CHENG	03-4788188	326	TAOYUAN HSIEN TAIWAN AA00010
950667456	201	ABC TAIWAN ELECTRONICS CORP.	NO 422 SEC 1	YANG HU RD YANG MEI CHENG	03-4788188	326	TAOYUAN HSIEN TAIWAN
620012311	295	ABILITY ENT CO LTD	3F NO 33 LN 76	REI KUANG RD	02-66028668	114	TAIPEI TAIWAN AP0106A
620017684	464	ABILITY ENT CO LTD	NO 147	FU HSIN N. RD	02-27168266	105	TAIPEI TAIWAN
963975018	464	ABILITY ENT CO LTD	NO 147	FU HSIN N. RD	02-27168266	105	TAIPEI TAIWAN
964930490	295	ABILITY ENT CO LTD TW	3F NO 33 LN 76	REI KUANG RD	02-66028668	114	TAIPEI TAIWAN AP0106A
960359242	464	ABILITY ENT CO LTD TW	NO 147	FU HSIN N. RD	02-27168266	105	TAIPEI TAIWAN AP18810
620021557	542	ABLE TOUCH ENTERPRICE CO LTD	8F-2 NO 206 SEC 2	NAN KING E. RD	02-25083015	104	TAIPEI TAIWAN AP23820
965341228	542	ABLE TOUCH ENTERPRISE CO LTD TW	8F-2 NO 206 SEC 2	NAN KING E. RD	02-25083015	104	TAIPEI TAIWAN AP23820
620015662	403	ABLEREX ELECTRONICS CO LTD	4F NO 1 LN 7	PAO KAO RD HSIN TIEN CITY	02-29176857	231	TAIPEI HSIEN TAIWAN AP13080
960929920	403	ABLEREX ELECTRONICS CO LTD	4F NO 1 LN 7	PAO KAO RD HSIN TIEN CITY	02-29176857	231	TAIPEI HSIEN TAIWAN
620016298	420	ACCTEL LIMITED TAIWAN BRANCH (HK)	6F NO 181	CHOU Z ST	02-27993539	114	TAIPEI TAIWAN

5.4 Data Patterns

Pattern Action language is used to manipulate data during the standardisation phase. Using an investigation phase, the data can be analysed and it is possible to identify patterns in data. Once the

pattern has been identified it is possible to perform actions against the data using the identified patters. Data Simple pattern classes are represented by single characters which are introduced in the table below.

Table 10: Pattern Classes (Source: IBM, 2018)

Class	Description
A - Z	User-supplied class from the classifications The classes A - Z correspond to classes that you code in the classifications. For example, if APARTMENT is given the class of U in the classifications, then APARTMENT matches a simple pattern of U.
^	Numeric The class ^ (caret) represents a single number, for example, the number 123. However, the number 1,230 uses three tokens: the number 1, a comma, and the number 230.
?	One or more consecutive words that are not in classifications. The class ? (question mark) represents one or more consecutive alphabetic words. For example, MAIN, CHERRY HILL, and SATSUMA PLUM TREE HILL each match to a single ? class provided none of these words are in the classifications for the rule set. Class ? is useful for street names when multi-word and single-word street names must be treated identically.

+	A single alphabetic word that is not in classifications
	The class + (plus sign) is useful for separating the parts of an unknown string. For example, in a name like OWAIN LIAM JONES, copy the individual words to columns with given name, middle name, and family name as follows:
	`+ | + | +`
	`COPY [1] {GivenName}`
	`COPY [2] {MiddleName}`
	`COPY [3] {FamilyName}`

&	A single token of any type The class & (ampersand) represents a single token of any class. For example, a pattern to match to a single word following an apartment type is: `U | &` SUITE 11 is recognized by this pattern. However, in a case such as APT 1ST FlOOR, only APT 1ST is recognized by this pattern.
\&	Type the backslash (\) escape character before the ampersand to use the ampersand as a literal. `< | \& | ? | T` 1ST & MAIN ST is recognized by this pattern.
>	Leading numeric The class > (greater than symbol) represents a token with numbers that is followed by letters. For example, a house number like 123A MAPLE AVE can be matched as follows: `> | ? | T` 123A is recognized by this pattern. The token contains numbers and alphabetic characters but the numbers are leading. In this example, T represents street type.
<	Leading alphabetic character

Class	Description
	The class < (less than symbol) matches itself to leading alphabetic letters. It is useful with the following examples: A123 ALPHA77 The token contains alphabetic characters and numbers but the alphabetic characters are leading.
@	Complex mix The class @ (at sign) represent tokens that have a complex mixture of alphabetic characters and numerics, for example: A123B, 345BCD789. For example, area information like Hamilton ON L8N 2P1 can be matched as follows: `+ | P | @ | @` In this example, P represents Province. The first @ represents L8N and the second @ represents 2P1.
~	Special punctuation The class ~ (tilde) represents special characters that are not in the **SEPLIST**. For example, if a **SEPLIST** does not contain the dollar sign and percent sign, then you might use the following pattern: `~ | +` In this example, $ HELLO and % OFF match the pattern.
k	One or more Chinese numeric characters
/	Literal The class / (slash) is useful for fractional addresses like 123 ½ MAPLE AVE, which matches to the following pattern: `> | ^ | / | ^ | ? | T`
\/	Backslash, forward slash You can use the backslash (\) escape character with the slash in the same manner that you use the / (slash) class.
-	Literal The class - (hyphen) is often used for address ranges, for example, an address range like 123-127 matches the following pattern: `^ | - | ^`
\-	You can use the backslash (\) escape character with the hyphen in the same manner you use the - (hyphen) class.
\#	Literal. You must use with the backslash (\) escape character, for example: \#. The class # (pound sign) is often used as a unit prefix, for example, an address like suite #12 or unit #9A matches the following pattern: `U | \# | &`
()	Literal The classes (and) (parentheses) are used to enclose operands or user variables in a pattern syntax. An example of a pattern syntax that includes a leading numeric operators and a trailing character operator is as follows: `> | ? | T`

Class	Description
	COPY [1](n) {HouseNumber} COPY [1](-c) {HouseNumberSuffix} COPY [2] {StreetName} COPY_A [3] {StreetSuffixType} EXIT The pattern syntax example, can recognize the address 123A MAPLE AVE. The numbers 123 are recognized as the house number and the letter A is recognized as a house number suffix. Use the backslash (\) escape character with the opening parenthesis or closing parenthesis to filter out parenthetical remarks. To remove a parenthetical remark such as (see Joe, Room 202), you specify this pattern: \(| ** | \) RETYPE [1] 0 RETYPE [2] 0 RETYPE [3] 0 The code example removes the parentheses and the contents of the parenthetical remark. In addition, when you retype these fields to **NULL** you essentially remove the parenthetical statement from consideration by any patterns that are further down in the pattern-action file. The **NULL** class (0) is not included in this list of single character classes. The **NULL** class is used in the classifications or in the **RETYPE** action to make a token **NULL**. Because a **NULL** class never matches to anything, it is never used in a pattern.
\(and \)	Use the backslash (\) escape character with the opening parenthesis or closing parenthesis to filter out parenthetical remarks. To remove a parenthetical remark such as (see Joe, Room 202), you specify this pattern: \(| ** | \) RETYPE [1] 0 RETYPE [2] 0 RETYPE [3] 0 The code example removes the parentheses and the contents of the parenthetical remark. In addition, when you retype these fields to **NULL** you essentially remove the parenthetical statement from consideration by any patterns that are further down in the pattern-action file.

Data patterns are key element as they are driver for enabling good quality standardization. There is a specific phase implemented by the author for analysing the different data patterns that are available within the full data set.

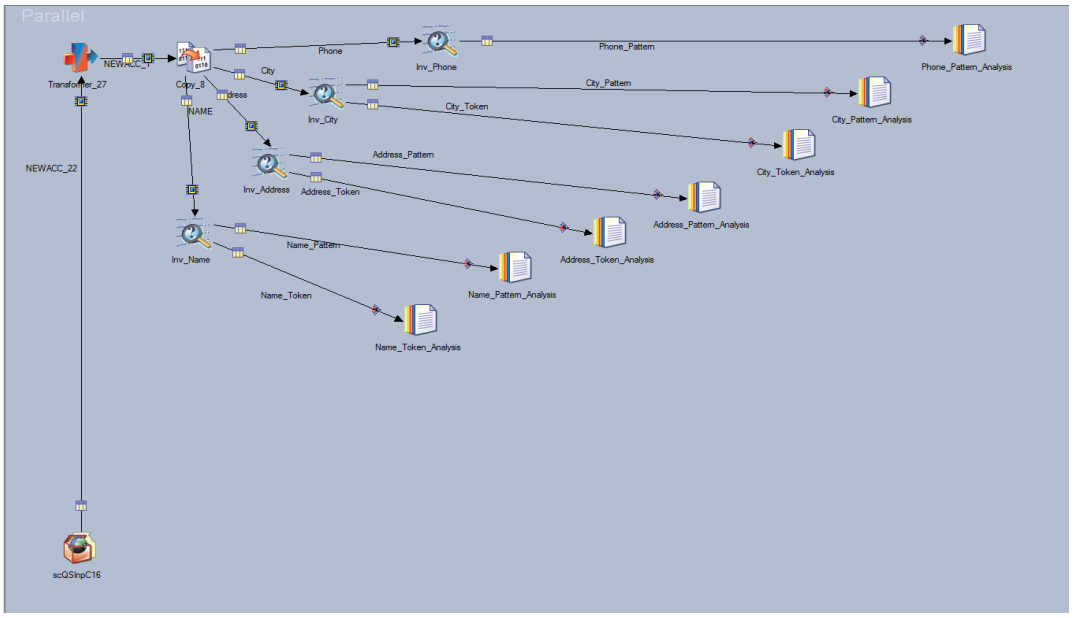

Figure 15: Identification of Data Patterns (Source: Author)

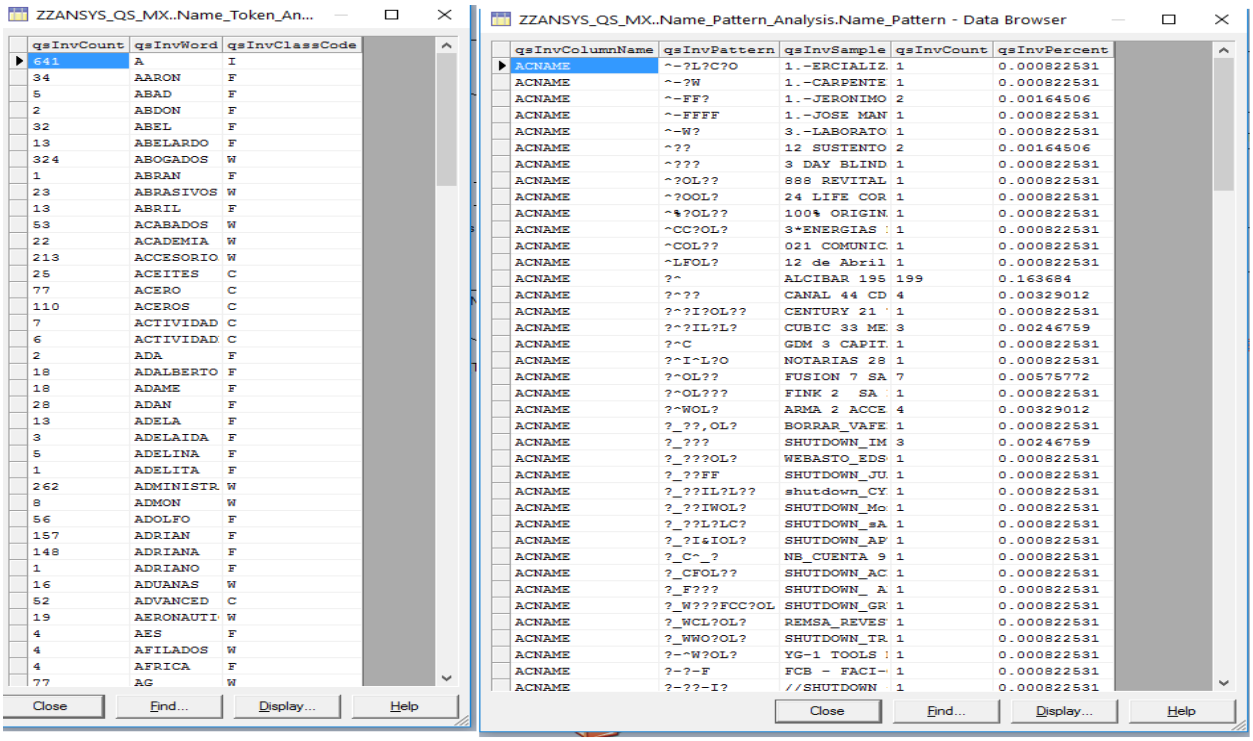

Figure 16: Pattern Analysis results (Source: Author)

The output of the analysis will be used by the author to make any required updates / additions in the Pattern Action file that is described in 6.3 Standardise to ensure that the standardisation is a high quality meaning that a better quality match can be created.

5.5 Standardise

Standardisation works based on special instructions called rule sets. Some rule sets are:

- Domain pre-processor, such as CTYPREP
- Domain-specific, such as CTYNAME, CTRYADDR

Most of the pre-packaged rule sets are country-specific. For example, there are different name standardisation rule sets for the United States and Japan. During the implementation there is the possibility for the user to either modify existing rule sets or create new sets based on the output from the analysis phase. Each rule set contains 3 main components and using a NAME rule set as an example:

Classification Table (.cls) which contains, keywords, standard values and user defined categories as well as how they should be read i.e.

- A - Abbreviations (Misspellings)
  - Tiphony = Tiffany
  - Moreen = Maureen
- C - Connector
  - THE = THE
  - TO = TO
- F - First Name
  - TAMIE = TAMIE
  - TAMIKA = TAMIKA
- G - Generational
  - SENIOR = SR
  - SNR = SR
- I - Initial
  - E = E
  - B = B
- L - Last Name Prefix
  - SAINT = ST
  - ST = ST
- O - Organization Name Suffix
  - PTNS = PARTNERS
  - PTRS = PARTNERS
- P - Prefix
  - PROF = PROF
  - PROFESSOR = PROF
- Q - Qualifier
  - ATTENTION = ATTN
  - ATTN = ATTN
- S – Suffix
  - ESQ = ESQ
  - ESQUIRE = ESQ
- W - Non-Individual Word (Common)
  - COMPUTER = COMPUTER

- COMPUTERS = COMPUTERS

The output from the Analyse phase will also be incorporated at this point: Seeing the analysis from Norway the following company words were added in addition to a Name Rule Set as common words:

- AS
- OG
- ASA
- AB
- LTD

Dictionary File (.dic) which contains the following:

- Layout of the output columns and can be broken into different categories
  - Business Intelligence Fields, these are filed that could be passed through to a business intelligence solutions i.e.
    - NameType
    - NamePrefix
    - FirstName
    - MiddleName
  - Matching fields, these are mainly used after the standardisation to enable the various levels of matching for i.e.
    - NYSIIS
    - Rsoundex
    - Match
    - Hashkey
    - Packedkey
  - Reporting fields, these are used to help report on the different information that can be in reporting i.e.
    - Unhandled data
    - Unhandled patterns

Pattern-Action File (.pat) which contains the logic to populate standardized output data

- STRIPLIST – The various chars that should be stripped out of any standardization. i.e.
  - !@#$%^&*()_-+={}[]
- SEPLIST – The various chars that should be read as separators. i.e.
  - []|\\:;\"<>
- Uses the classification from the .cls to identify various patterns and apply the correct algorithms i.e.
  - If a word is unknown it will be given the classification "+"
- If required additional standardization rules can be created based on the various data patterns

Once all the required rules sets defined the author assigns the correct rule set to the associated field that requires standardization

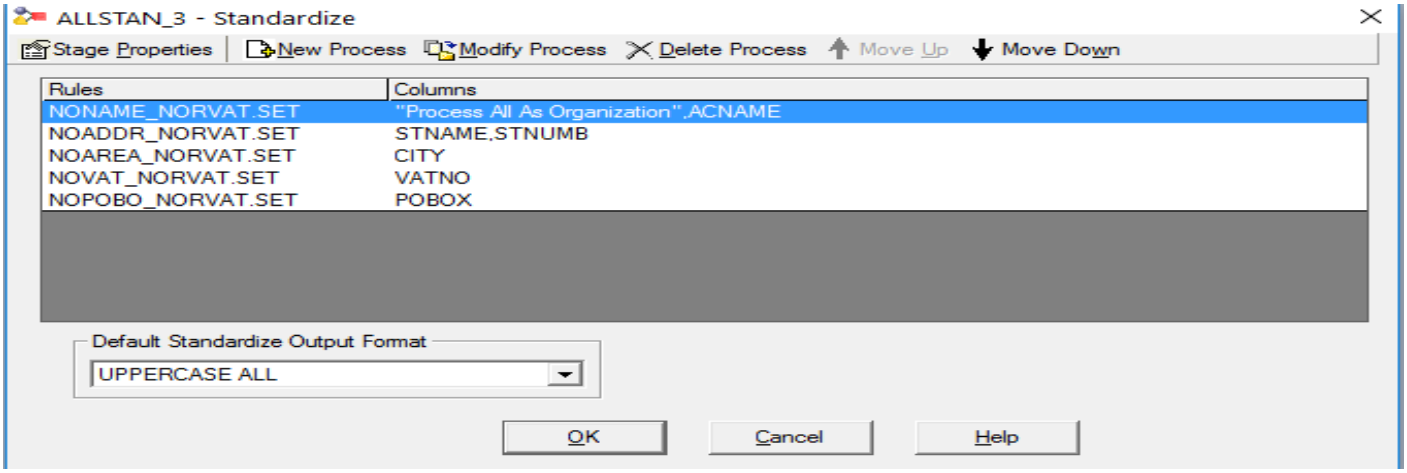

Figure 17: Assignation of rules (Source: Author)

The reason for standardisation of data is to create a clean data set with additional output fields that will enable different methods for creating a more effective matching. Some the additional fields that are created during the standardisation are:

New York State Identification and Intelligence System (NYSIIS): "*Is a phonetic algorithm for creating indices for words based on their pronunciation. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling"* (Rosettacode, 2018)

SOUNDEX: "*Returns a character string containing the phonetic representation of char. This function lets you compare words that are spelled differently, but sound alike in English*" (Oracle, 2018)

RSOUNDEX: Reverse SOUNDEX action is the same as the SOUNDEX action except that the phonetic code is generated from the last non-blank character of the field and proceeds to the first (IBM, 2010).

Match Words: These are the various words that are separated during the standardisation process

Input Pattern: The sequence of class labels assigned to the values in a data record which can be used to identify a subset of records that might be standardised the same way (IBM knowledge center, 2012)

The standardisation can be checked using the rules management functionality without having to run the full end to end process allowing.

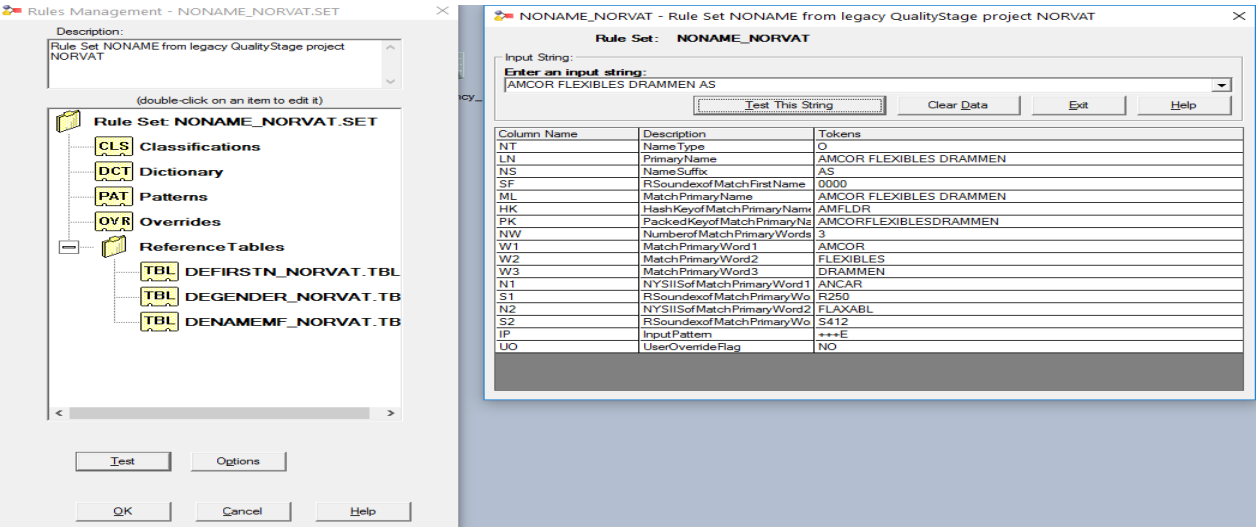

Figure 18: Testing of Standardisation (Source: Author)

The output from the standardisation stage is a file called ALLSTAN which an extension of the NEWACC file but containing all the fields that are used for enabling a better quality match without making any modifications to the original source data.

As the data being standardized relates to companies there are additional settings that can be applied in relation to the name:

Process All as Individual: All columns are standardised as individual names.

Process All as Organization: All columns are standardised as organization names.

Process Undefined as Individual: All undefined (unhandled) columns are standardised as individual names.

Process Undefined as Organization: All undefined (unhandled) columns are standardised as organisation names.

NAMES handling options enhance performance by eliminating the processing steps of determining the type of name. This option is useful when you know the type of name information that your input file contains. For example, if you know that your file contains only organisation names, specify Process All as Organisation. Every name entry in the file will be treated as an organisation name. Even if you have an entry that is obvious to you as the name of an individual, it will be parsed as an organisation (Source: IBM knowledge center, 2012).

As it is already known that the field ACNAME contains company names the additional setting Process All as Organisation is used. The following standardisation rules were applied for the following fields:

NONAME Domain-Specific Rule Set for NO Names

Field Name	Start Position	Length	Description or Literal
STNAME	606	50	ACCOUNT NAME

NOADDR Domain-Specific Rule Set for NO Addresses

Field Name	Start Position	Length	Description or Literal

COADR1	656	50	ADDRESS LINE 1

NOAREA Domain-Specific Rule Set for NO Localities

Field Name	Start Position	Length	Description or Literal
CITY	456	50	CITY

NOPOBO Domain-Specific Rule Set for NO Post Box Numbers

Field Name	Start Position	Length	Description or Literal
POBOX	706	50	PO Box Numbers

NOVAT Domain-Specific Rule Set for NO VAT numbers

Field Name	Start Position	Length	Description or Literal
VATNO	781	30	VAT Numbers

By using these rule sets it is possible to split the data into manageable and readable sets of information. This helps us to ensure the matching is a higher quality than is possible with the original data set. The ALLSTAN file is only used for the matching process and is not used for data cleansing and updating the original legacy systems.

LNNONAM	NGNONAM	NSNONAM	ANNONAM	MFNONAM	NFNONAM	SFNONAM	MLNONAM	HKNONAM	PKNONAM	NWNONAM	W1NONAM	W2NONAM	W3NONAM	W4NONAM	W5NONAM	N1NONAM	S1NONAM
07 MEDIA		AS				0	07 MEDIA	07ME	07MEDIA	2	7	MEDIA				7	7000
07 OSLO		AS				0	07 OSLO	07OS	07OSLO	2	7	OSLO				7	7000
07 SØR		AS				0	07 SØR	07SØ	07SØR	2	7	SØR				7	7000
123 COMMUNICATION		AS				0	123 COMMUNICATION	12CO	123COMMUNICATION	2	123	COMMUNICATION				123	3210
123CONCEPT		AS				0	123CONCEPT	12	123CONCEPT	1	123CONCEPT					123CANCA	T125
123FRILUFT DA						0	123FRILUFT DA	12DA	123FRILUFTDA	2	123FRILUFT	DA				123FRALA	T146
123						0	123	12	123	1	123					123	3210
150 YARDS AHEAD		AS				0	150 YARDS AHEAD	15YAAH	150YARDSAHEAD	3	150	YARDS	AHEAD			150	510
1755 RETAIL NORGE		AS				0	1755 RETAIL NORGE	17RENO	1755RETAILNORGE	3	1755	RETAIL	NORGE			1755	5571
19 PILOGBUE V / SKEIE						0	19 PILOGBUE V / SKEIE	19PIV / SK	19PILOGBUEV/SKEIE	5	19	PILOGBUE	V	/	SKEIE	19	9100

Figure 19: Example of name standardisation from ALLSTAN (Source: Author)

5.6 Matching Algorithms

The matching is carried out using specific fields from the ALLSTAN file. it is possible are also able to complete more than one pass of the data, using different fields and applying different conditions, and this can be completed up to seven times.

Within the matching process a concept called “blocking” is used. This means that for each individual match pass, fields are selected at the start of the each pass and all these fields must match 100% to be even considered as a possible match.

In the first pass the highest amount of "blocking" fields as possible are used so that the starting point give the highest possible match based on the data content and quality, then for each subsequent pass the "blocking" fields are reduced / modified.

As well as using the concept of blocking fields different conditions can be applied to each pass to ensure that matches can still be made if there is information missing from different legacy systems.

For the grouping process in Norway, the author completed seven different passes of the data to ensure that she captured as many matches as possible. Below is the breakdown of each of the passes and the fields and conditions used:

Pass One

Blocking fields:

- Primary word of the Account Name
- House Number
- NYSIIS (phonetic) Street Name Root
- NYSIIS (phonetic) City

Conditions Applied:

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO BOX number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 2 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Pass Two

Blocking Fields:

- Primary word of the Account Name
- NYSIIS (phonetic) City
- NYSIIS (phonetic) Street Name Root

Conditions Applied:

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 2 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO Box number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Pass Three

Blocking Fields:

Hash Key of the primary Account Name

NYSIIS (phonetic) City

NYSIIS (phonetic) Street Name Root

Conditions Applied:

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO Box number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Pass Four

Blocking Fields:

- NYSIIS (phonetic) Street Name Root word 1
- Primary word of the Account Name
- House Number
- Postal Code

Conditions Applied:

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Postal Code - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 2 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO Box number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Pass Five

Blocking Fields

- NYSIIS (phonetic) Street Name Root word 2
- Primary word of the Account Name
- House Number
- Postal Code

Conditions Applied

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO BOX number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 2 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

<u>Pass Six</u>

Blocking Fields:

- NYSIIS (phonetic) Street Name Root word 2
- Primary word of the Account Name
- Postal Code

Conditions Applied:

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO BOX number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 2 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Post Box Value - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Building Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Postal Code - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Pass Seven

Blocking Fields:

- VAT Number
- 3 Letters of City

Conditions Applied:

Total Original Input Record - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 1 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account words 2-5 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Street Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

House Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Packed key of Street name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

City Name - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

PO BOX number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Number - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

Account Primary word 2 - If the data is populated and does not match then do not group the account, but if the data is missing and all the blocking fields match then group the accounts.

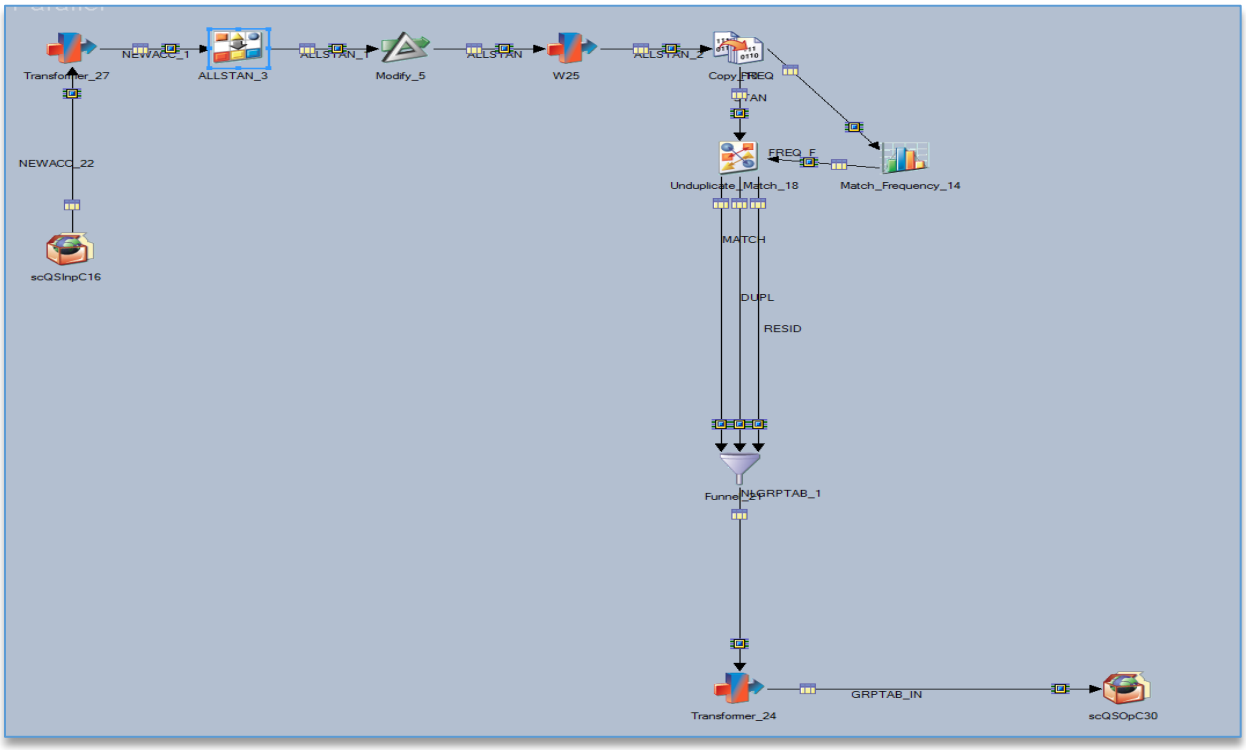

Figure 20: Matching Process (Source: Author)

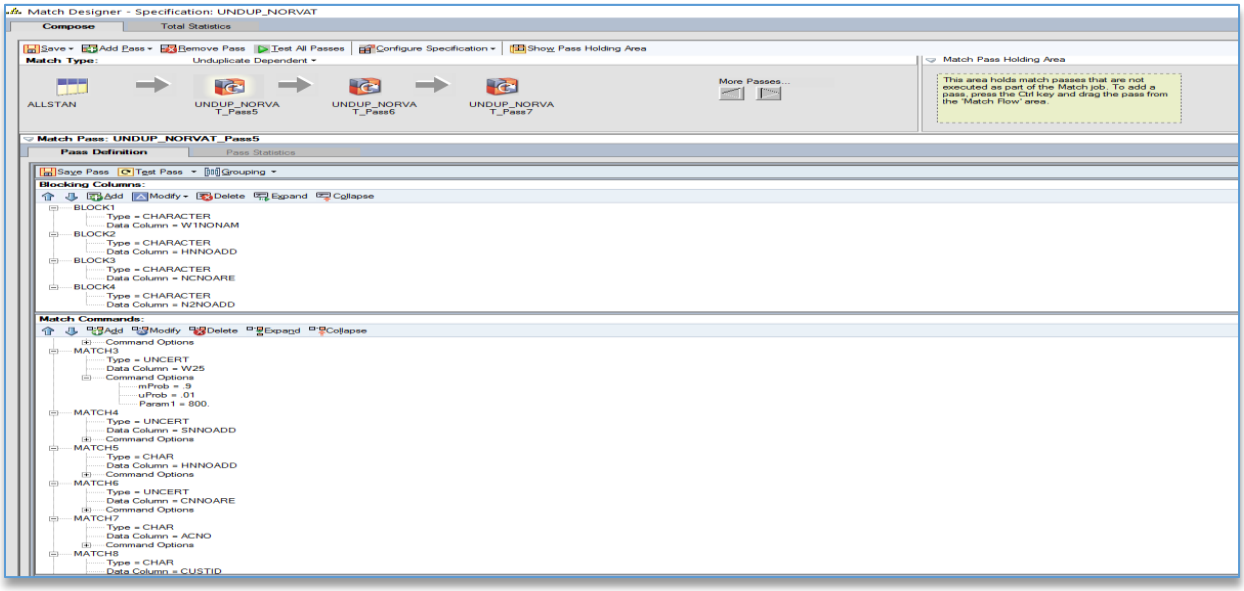

Figure 21: Matching Specification (Source: Author)

Once the matching process has completed, there is a validation phase carried out by the author that again uses sampling. The output from the Matching phase is a new file called ACC2GRP which is the

NEWACC file with 1 additional field Group ID and this is actual matched key which is also the Customer ID.

5.7 Output

Once the ACC2GRP file is created, there is a secondary output file created which is a Tab Delimited version of ACC2GRP and is called GRPTAB.

SOURCE	ACCOUNT_NUMBER	GSFA_CUSTOMER_ID	SUSPECT_NAME	ADTY	ADDRESS_LINE1	AD2	AD3	AD4	AD5	CC	POSTCODE	CITY	BNAME	DISTRICT	PROVINCE	STREET NAME	STREET NUMBER	PO_BOX	PHONE	VATNO	CC	GROUP_ID
	NOS28163SS	REFERENCE DATA	07 MEDIA AS		Peter Møllers vei 8172				NOS28163SS	NO	585	oslo							4745425934		NO	10724
	NOC2428	REFERENCE DATA	07 OSLO As		Peter Møllers vei 8				NOC2428	NO	585	OSLO				PETER MØLLERS VEI			4722799500		NO	6186
	NOC345052968	REFERENCE DATA	07 SØR AS		INDUSTRIGATA 13	4632 KRISTIANSAND			NOC345052968	NO	4632	KRISTIANSAND				INDUSTRIGATA	13				NO	8196
	NOC116619	REFERENCE DATA	123 COMMUNICATION AS		Postboks 123	4302 SANDNES			NOC116619	NO	4302	STAVANGER						123	4799390123		NO	5280
	NOS26899SS	REFERENCE DATA	123CONCEPT AS		Johan Stangs plass 2				NOS26899SS	NO	1767	HALDEN							4790564379		NO	18383
	NOC265440568	REFERENCE DATA	123FRILUFT DA		Lindholmveien 14	1788 BERG I ØSTFOLD			NOC265440568	NO	1788	BERG I ØSTFOLD				ISEBAKKEVEIEN	43		4.79056E+17		NO	16940
	COMET_1-1CZIRD5	REFERENCE DATA	123		Olav V's gate 5				COMET_1-1CZIRD5	NO	161	Oslo							4723114950		NO	115
	COMET_1-4UWJD7L	REFERENCE DATA	150 YARDS AHEAD AS		Magasinvegen 14	Magasinvegen 14			COMET_1-4UWJD7L	NO	5705	Voss							4748036083		NO	4192
	NOC283782843	REFERENCE DATA	1755 RETAIL NORGE AS		GRØNNEGATA 1	0350 OSLO			NOC283782843	NO	350	OSLO				GRØNNEGATA	1				NO	16506
	NOS3520SS	REFERENCE DATA	19 PILOGBUE V/SKEIE		STOEPERIVEIEN 4 B				NOS3520SS	NO	4517	MANDAL							4.748E+11		NO	10660

Figure 22: GRPTAB sample (Source: Author)

The GRPTAB is then used by various teams as well as the Author to complete multiple validations where the output enables further enhancements to the following input files.

- Classification file
- Pattern Action file

After each change to either of the above files the standardisation and match phases are re-run by the author. This is iterative until there has been an agreement that the matching is of the required quality

The number of accounts that entered the matching process was 58,000 and the number of customers outputted form the process was 32,000 which means an average of 1.81 accounts to a Customer.

Once the matching is completed a coalition can be made by matching the accounts to the legacy revenue systems and then rolling up based on the Customer ID

6 Feedback from Organization

The author has requested feedback in relation to the effect the standardization and matching of the data has had on the company. The company has supplied the pros and cons in relation to the final results.

After the matching and the creation of 32,000 Customers and distributed them across different sales channel which is revenue based only, there have been improvements in the following areas:

- Reduction in sales force number as working at Customer and not Account level
- Reduction in cost of sale (Reduction in Salesforce hardware, mobility needs)
- Improvement in quality of the customer base
- Reduction in number of sales executives
- Improvement in overall visibility of all services/products being used
- Assigned to correct sales channel based on combined revenue

The following drawback was highlighted

- Potential that some accounts may not be matched correctly

Conclusion

This master thesis was dedicated to the usage of integration tools within organizations to create a 360 degree view of the customer.

It shows that the use of integration tools can not only create a better visibility in relation to a 360 degree customer view, but also save costs and drive expansion of existing customer revenue.

It is also apparent that after mergers / acquisitions there is not a need to invest heavily either in new infrastructure/software or by using integration tools they can maintain the existing solutions in the background.

Using integration tools the legacy systems does not affect any embedded processes which will result in additional costs to complete new process reviews, modelling and implementation which can cause an employee retention issues.

Glossary of terms

Term	Definition
Customer Relationship Management (CRM)	A strategy for managing all your company's relationships and interactions with your customers and potential customers. It helps you improve your profitability (Salesforce, 2018).
Mergers and acquisitions (M&A)	A general term that refers to the consolidation of companies or assets. M&A can include a number of different transactions, such as mergers, acquisitions, consolidations, tender offers, purchase of assets and management acquisitions. In all cases, two companies are involved. The term M&A also refers to the department at financial institutions that deals with mergers and acquisitions (Investopedia, 2016).
Enterprise resource planning (ERP)	A process by which a company (often a manufacturer) manages and integrates the important parts of its business. An ERP management information system integrates areas such as planning, purchasing, inventory, sales, marketing, finance and human resources (Investopedia, 2018).
Extract Transform Load (ETL)	Its task is to get data from source systems and select (Extraction),transform due requested form (Transformation) and load to specified data structures, data scheme, data warehouse (Loading) (Novotny, 2005).
Sales force automation (SFA)	Direct sales software builds on the attributes of technology, functionality and value of order management systems and also includes the functionality for sales execution and sales operations. The direct B2B sales organization is the traditional sales channel, composed of internal sales resources focused on the selling of products or services directly into the client, customer and prospect base as employees of the provider company. Direct sales resources may be field-based, calling on customers face to face at their locations, or inside sales, selling from a desk environment over the phone (Gartner, 2018).

Big Data	Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation (Gartner, 2018).
Marketing automation system (MAS)	A marketing automation system is a system that helps marketers execute multichannel marketing campaigns by providing a scripting environment for authoring business rules and interfaces to a variety of third-party applications (Gartner, 2018).
Customer service and support (CSS)	Once known as the "complaint department," customer service and support or CSS is responsible for retaining and extending customer relationships once a product or service is sold. Due to the increasing complexity of customer interactions, customer service organizations need a complex technological infrastructure that is flexible, extensible and scalable and that integrates front-office applications with back-end processes and data (Gartner, 2018).
Customer Intelligence (CI)	Customer intelligence (CI) is information derived from customer data that an organization collects from both internal and external sources. The purpose of CI is to understand customer motivations better in order to drive future growth (Searchbusinessanalytics, 2010)
Customer Data Repository (CDR)	Enhanced Customer Data Repository (ECuRep) is a secure and fully supported data repository with problem determination tools and functions. It updates problem management records (PMR) and maintains full data life cycle management (Searchbusinessanalytics, 2010).
Line of Business (LOB)	A line of business is a corporate subdivision focused on a single product or family of products (Gartner, 2018).
Data Quality Management (DQM)	Data quality management is an administration type that incorporates the role establishment, role deployment, policies, responsibilities and processes with regard to the acquisition, maintenance, disposition and distribution of data. In order for a data quality management initiative to succeed, a strong partnership between technology groups and the business is required (Techopedia, 2018).

Enterprise Information Management (EIM)	EIM is an integrative discipline for structuring, describing and governing information assets across organizational and technological boundaries to improve efficiency, promote transparency and enable business insight (Gartner, 2018).

Bibliography

POUR, Jan, Miloš MARYŠKA a Ota NOVOTNÝ. *Business intelligence v podnikové praxi*. Praha: Professional Publishing, 2012. ISBN 978-80-7431-065-2.

NOVOTNÝ, Ota, POUR, Jan a SLÁNSKÝ, David. *Business Intelligence: Jak využít bohatství ve vašich datech.* Praha : Grada Publishing a.s., 2005. 80-247-1094-3.

Gartner. *Gartenr IT Glossary* [online]. [cit. 2018-04-17]. Available: https://www.gartner.com/it-glossary

HAGEN, Christian, Sumit CHANDRA a Jason MILLER. Mergers & Acquisitions: Make or Break: The Critical Role of IT in Post-Merger Integration. *ATKerney* [online]. 2010 [cit. 2018-04-11]. Available: https://www.atkearney.com/mergers-acquisitions/article?/a/make-or-break-the-critical-role-of-it-in-post-merger-integration

Salesforce. *How do we define CRM* [online]. [cit. 2018-03-11]. Available: https://www.salesforce.com

Investopedia: Why do companies merge with or acquire other companies?. *Investopedia*[online]. 2017 [cit. 2018-04-11]. Available: https://www.investopedia.com/ask/answers/why-do-companies-merge-or-acquire-other-companies/

Investopedia: What is 'Enterprise Resource Planning - ERP'. *Investopedia* [online]. 2017 [cit. 2018-04-11]. Available: https://www.investopedia.com/terms/e/erp.asp

DCS. *DCS Solutions: Our new series: ERP frequently asked questions* [online]. 2015 [cit. 2018-04-02]. Available on: https://www.dcs-solutions.co.uk/news/what_is_erp_software/

Pega. *Gartner Recognizes Pega as a Leader in MQ for CRM Customer Engagement Center* [online]. [cit. 2018-04-02]. Available on: https://www.pega.com/gartner-crm-cec-2017

Oracle. *Siebel Sales* [online]. 2017 [cit. 2018-04-02]. Available on: http://www.oracle.com/us/products/applications/siebel/sales/siebel-sales/features/index.html

Oracle. *Siebel Sales Applications* [online]. 2016 [cit. 2018-04-03]. Available on: http://www.oracle.com/us/products/applications/siebel/051148.pdf

Informatica. *Informatica Named a Leader for 12th Consecutive Year* [online]. 2017 [cit. 2018-04-03]. Available on: https://www.informatica.com/data-integration-magic-quadrant.html#fbid=mMBgJsPlKOF

Economywatch. *History of Mergers and Acquisitions* [online]. 2010 [cit. 2018-04-03]. Available on: http://www.economywatch.com/mergers-acquisitions/history.html

Techopedia: Business Intelligence (BI). *Techopedia* [online]. 2018 [cit. 2018-04-14]. Available: https://www.techopedia.com/definition/345/business-intelligence-bi

QUADDUS, Mohammed and Arch WOODSDE. *Sustaining Competitive Advantage via Business Intelligence, Knowledge Management, and System Dynamics* [online]. Emerald Publishing Limited, 2105 [cit. 2018-04-15]. ISBN 9781784417635. Available: https://ebookcentral.proquest.com/lib/vsep/detail.action?docID=4339861&query=Sustaining%20Competitive%20Advantage%20via%20Business%20Intelligence%2C%20Knowledge%20Management%2C%20and%20System%20Dynamics

VASILIEV, Yuli. *Oracle Business Intelligence The Condensed Guide to Analysis and Reporting : The Condensed Guide to Analysis and Reporting* [online]. Packt Publishing, 2010 [cit. 2018-04-15]. ISBN 9781849681193. Available: https://ebookcentral.proquest.com/lib/vsep/reader.action?docID=944000&ppg=5

Oracle. *Oracle8i Data Warehousing Guide Release 2 (8.1.6): Data Marts* [online]. 1999 [cit. 2018-04-16]. Available: https://docs.oracle.com/cd/A81042_01/DOC/server.816/a76994/marts.htm

Oracle. *Oracle8i Data Warehousing Guide Release 2 (8.1.6): Data warehousing concepts* [online]. 1999 [cit. 2018-04-16]. Available: https://docs.oracle.com/cd/A84870_01/doc/server.816/a76994/concept.htm

Microsoft. *Online analytical processing (OLAP)* [online]. 2017 [cit. 2018-04-16]. Available on: https://docs.microsoft.com/en-us/azure/architecture/data-guide/relational-data/online-analytical-processing

Microsoft. *About OLAP Cubes* [online]. 2016 [cit. 2018-04-16]. Available on: https://technet.microsoft.com/en-us/library/hh916536(v=sc.12).aspx

ONG, Lih, Pei SIEW a Siew WONG. *A Five-Layered Business Intelligence Architecture* [online]. 2011 [cit. 2018-04-17]. DOI: 10.5171/2011.695619. Available on: https://pdfs.semanticscholar.org/e700/16f3b82bebd130bf58746d2a4d681b7ec18d.pdf

WHITE, Colin. *Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise (Report Excerpt)* [online]. 2003 [cit. 2018-04-17]. Available on: http://www.bi-bestpractices.com/view-articles/4737

Rosettacode. NYSIIS [online]. 2018 [cit. 2018-04-18]. Available on: https://rosettacode.org/wiki/NYSIIS

IBM Knowledge Center. *Configuring the Standardize stage* [online]. 2015 [cit. 2018-04-18]. Available on:https://www.ibm.com/support/knowledgecenter/en/SSZJPZ_11.3.0/com.ibm.swg.im.iis.qs.ug.doc/topics/t_Configuring_the_Standardize_stage.html

Oracle. *Soundex: Database SQL Reference* [online]. 2018 [cit. 2018-04-18]. Available on: https://docs.oracle.com/cd/B19306_01/server.102/b14200/functions148.htm

IBM. *RSoundex* [online]. 2010 [cit. 2018-04-18]. Available on: https://www.ibm.com/support/knowledgecenter/en/SSZJPZ_8.5.0/com.ibm.swg.im.iis.qs.patguide.doc/topics/r_SOUNDEX_Phonetic_Coding.html

IBM knowledge center. *InfoSphere QualityStage glossary* [online]. 2012 [cit. 2018-04-18]. Available on: https://www.ibm.com/support/knowledgecenter/en/SSZJPZ_9.1.0/com.ibm.swg.im.iis.qs.glossary.doc/topics/glossary_qs.html

IBM knowledge center. *Identifying simple pattern classes* [online]. 2018 [cit. 2018-04-24]. Dostupné z: https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.qs.patguide.doc/topics/c_Identifying_Simple_Pattern_Classes.html

Searchbusinessanalytics. *Customer intelligence* [online]. 2010 [cit. 2018-04-18]. Available on: https://searchbusinessanalytics.techtarget.com/definition/customer-intelligence-CI

GALA, L. Řízení vztahů se zákazníky (CRM- Customer Relationship Management). *MBI* [online]. 2015 [cit. 2018-04-18]. Available: http://mbi.vse.cz/mbi/index.html#obj/FACTOR-157

MBI. *MBI* [online]. [cit. 2018-04-18]. Available on: http://mbi.vse.cz/

IBM. *Enhanced Customer Data Repository* [online]. 2015 [cit. 2018-04-18]. Available on: https://www-05.ibm.com/de/support/ecurep/index.html

KHAN, Abeer a Nadeem EHSAN. Integration between Customer Relationship Management (CRM) and Data Warehousing. *Sciencedirect* [online]. 2011, **2011**, 11 [cit. 2018-04-18]. Available on: https://www.researchgate.net/publication/257743696_Integration_between_Customer_Relationship_Management_CRM_and_Data_Warehousing

ROUSE, Margaret. Data Qulity. *Searchdatamanagement* [online]. 2005 [cit. 2018-04-18]. Available on: https://searchdatamanagement.techtarget.com/definition/data-quality

FRATTELONE, Paul. Business Intelligence and Data Quality. *Stickyminds* [online]. 2012 [cit. 2018-04-18]. Available on: https://www.stickyminds.com/article/business-intelligence-and-data-quality

Attachments

Attachment A: complete tables from the Data mapping chapter

Table 11: Customer Data Fields complete table (Source: Author)

Customer Data Field	Data Type	Description
Customer ID	Char	Unique Identifier for a Customer Record. Country Code and C to be prefixed for a sequence number generated.
Customer Name	Char	A customer is a trading entity that actively generates revenue with organization and is located at a specific site, and will have one or more separate agreements with the organization
Telephone Number	Char	Should be a 10 or more digit number. "+" and Country code to be prefixed to the telephone number. For ex., for Norway, +4798435675 would be a telephone number
Industry Code	Char	The Industry code assigned to this customer. This is a 2 digit industry code.
Sales Territory Code	Char	The Sales Territory the customer belongs to. This is the 'Managed' territory information.
Revenue Band	Char	The Revenue band of the Customer
Sales Channel	Char	
URL	Char	Website for the Customer
Call Frequency	Number	The call frequency schedule this customer has been assigned
Customer Sales Stage	Char	The Sales stage the customer is in at the moment
Regular Pick-up	Boolean	A flag to indicate if this account has regular pick-up
Division	Char	A code to show which division this customer has been associated with.
Source Type	Char	A code to show the source of the lead
Country Specific Fields 1	Char	Can be used for inserting country specific details related to the Customer
Country Specific Fields 2	Char	Can be used for inserting country specific details related to the Customer
Country Specific Fields 3	Char	Can be used for inserting country specific details related to the Customer
Country Specific Fields 4	Char	Can be used for inserting country specific details related to the Customer
Global Agreement Flag	Boolean	If there is a global agreement related to the Customer
Regional Agreement Flag	Boolean	If there is a reginal agreement related to the Customer
Country Agreement Flag	Boolean	If there is a country agreement related to the Customer
Country Industry Class	Char	A code that identifies the industry in which a site customer operates, as defined locally.

Overall Opportunity Potential	Number	Summation of Customers Opportunity Potential Revenue
Qualification Potential Revenue	Number	Potential Revenue for the Customer – will drive Sales Channel/ Revenue Band
Loyalty Code	Char	The loyalty code associated with this customer. Mastered in the Data Mart
Cash Customer	Boolean	A flag to indicate if the customer is a cash customer.
Solvency	Boolean	A flag to indicate that the customer is solvent overall and does not have an accounts that are bankrupt
Share of Wallet	Number	Potential over actual revenue of this customer site.
Committed revenue	Number	The sum of all committed revenue for each opportunity for a customer.
Organization	Char	Default to the Organization of the Country to which the Customer belongs (Old company etc)
Business at Risk Reason	Char	
Business at Risk	Boolean	A flag to indicate the business is at risk at the Customer.
Enterprise ID	Char	National/Mini-National Indicator
Currency Code	Char	Currency code of the organization the Customer belongs to
Site Agreement Flag	Boolean	
Created	Date	Created Date. Format of the date should be "YYYY-MM-DD HH24:MI:SS"
Created By	Char	Created By.
Customer Name Local	Char	Customer Name in Local Language
Last Call Date	Date	Last Call Date is the date when customer was last called or met. The format of this field will be YYYY-MM-DD HH24:MI:SS
Lead Qualification Date	Date	Date on which the particular lead was qualified in.The format of this field will be YYYY-MM-DD HH24:MI:SS
Last Shipment/Last International Package	Char	When the Last Shipment took place. It takes values 1 to 3 months/More than 3 months.
All Products Total Potential Revenue	Number	Total Potential revenue
Overall Qualification Revenue	Number	Overall Qualification Revenue
Qualification Potential Revenue	Number	Potential Revenue
Billing Entity/Time Definite Paid By	Char	The consignee/The Sender/Third Party
Air Freight	Boolean	Value can be 'Y' or 'N' or Null.
Sea Freight	Boolean	Value can be 'Y' or 'N' or Null.
Mail	Boolean	Value can be 'Y' or 'N' or Null.
Sales Lead Originator	Char	The employee who sent the lead.
Sales Lead Originator E-Mail	Char	The email of the employee who sent the lead.

Table 12: Account data fields complete table (Source: Author)

Account Data Field	Data Type	Description
Customer ID	Char	Uniquely identifies a Customer Record
Customer Name	Char	Uniquely identifies a Customer Record. Should be the same Customer Name as available in the Customer File for this Customer ID.
Account ID	Char	Required for maintaining the Customer – Account Relationship within COMET.
Account Number	Char	The unique identifier of this account. This is the account number generated in the legacy billing systems
Account Name	Char	Account Name
Parent Account Number		Master Account Number of an account.
Credit Stop Flag	Boolean	Indicates that this account is on credit stop
Tariff Reference Code	Char	A contract code applicable to this account
First Shipment Date	Date	The date the last shipment was sent. Date format is "YYYY-MM-DD HH24:MI:SS". Data type and format need not match. Database has its own representation.
Last Shipment Date	Date	The date the last shipment was sent. Date format is "YYYY-MM-DD HH24:MI:SS". Data type and format need not match. Database has its own representation.
Service Area Code	Char	The Service Center the account belongs to.
Country		Value should not be provided.
Country Specific Fields 1	Char	Can be used for inserting country specific details related to the Account
Country Specific Fields 2	Char	Can be used for inserting country specific details related to the Account
Country Specific Fields 3	Char	Can be used for inserting country specific details related to the Account
Country Specific Fields 4	Char	Can be used for inserting country specific details related to the Account
Account Status	Char	Status of the Account
Major Account Code	Char	A code to identify the Major Account code.
Archive Flag	Boolean	Flag to indicate that the Account is archived by the source system
Account Create Date	Date	The date the last shipment was sent. Date format is "YYYY-MM-DD HH24:MI:SS". Data type and format need not match. Database has its own representation.
Account Closed Date	Date	The date the last shipment was sent. Date format is "YYYY-MM-DD HH24:MI:SS". Data type and format need not match. Database has its own representation.
Credit Limit	Number	The credit limit to which the Account is allowed to ship
VAT number	Char	Customers VAT number
Contract End Date	Date	The day the contract ends for account
Global Agreement Code	Char	

Regional Agreement Code	Char	
Country Agreement Code	Char	
Legacy Industry Code	Char	The legacy industry code to which the Account belongs to.
Organization	Char	Default to the Organization of the Country to which the Customer belongs t
Division	Char	Organization of the legacy Account
Site Agreement Code	Char	
Currency Code	Char	Currency code of the organization the Customer belongs to
Created	Date	Created date. Format of the date should be "YYYY-MM-DD HH24:MI:SS"
Created By	Char	Created By.
Account Name Local	Char	Account Name in Local Language

Table 13: Address data fields complete table (Source: Author)

Address Data Field	Data Type	Description
Record Type	Char	Indictor if the Address is for a Customer (C) or Account (A)
Customer Id / Account Id	Char	Uniquely identifies a Customer (or) Account Record
Customer / Account Name	Char	Uniquely identifies a Customer (or) Account Record. Should be the same Customer/Account Name as available in the Customer/Account File for this Customer ID.
Address Unique Id	Char	Uniquely identifies the address. It should have the following format -
Address Type	Char	The role for the address type. Number of address will differ per country
Address Line 1	Char	The first line of the address specification
Address Line 2	Char	The second line of the address specification
Address Line 3	Char	The Third line of the address specification
Building Name	Char	The name of the building
Po Box	Number	A number issued by the country's postal authority
Street Name	Char	The name of the street
Street Number	Char	The number within the street
District	Char	The area of the city
City	Char	Name of the city
Province	Char	Area of the country
State	Char	Mandatory based on condition. If State is a mandatory field for the Country, then value to be available in the data file. For ex., State is mandatory for US.
Postcode	Char	A number issued by the country's postal authority indicating the geographical location of the address
Country	Char	The international country code

Primary Address	Boolean	One of the addresses must be primary
Organization	Char	Default to the Organization of the Country to which the Customer belongs (Old company etc.)

Table 14: Contact data fields complete table (Source: Author)

Contact Data Field	Data Type	Description
Record Type	Char	Indictor if the Contact is for a Customer (C) or Account (A)
Customer Id / Account Id	Char	Uniquely identifies a Customer (or) Account Record
Customer / Account Name	Char	Uniquely identifies a Customer (or) Account Record.
Contact Unique Id	Char	Unique Identifier to identify the contact.
Contact Last Name	Char	Contact last name
Contact Type	Char	Type of the Contact; Billing, Shipping, main, Sales, Pick up
Contact First Name	Char	Contact first name
Title	Char	The title of the Contact Mr./Ms./Dr. etc.
Work Phone Number	Char	Should be a 10 or more digit number. "+" and Country code to be prefixed to the telephone number
Home Phone Number	Char	Should be a 10 or more digit number. "+" and Country code to be prefixed to the telephone number.
Email	Char	Email id
Contact Method	Char	The preferred contact method of the contact, e.g. e-mail, phone, fax
Buying Role	Char	The role the contact has in the procurement; process, approver, decision maker, influencer, user
Mail Stop	Boolean	An indicator that this contact should not be included in any mailing
Department	Char	The department the contact is in
Comments	Char	Any comments related to this contact
Language Code	Char	The language code of the contact
Never e-mail	Boolean	Flag indicating the contact has not authorized email communication
Primary Contact	Boolean	Indicator if the contact is the Primary contact for the Customer or Account.
Job Description	Char	The job title of the contact. This is a pick list with predetermined DHL Job description
Organization	Char	Default to the Organization of the Country to which the Customer belongs (Old company etc)
Job Title	Char	Job Title of the contact.
Contact Last Name Local	Char	Contact Last Name in Local Language
Contact First Name Local	Char	Contact First Name in Local Language
Contact Department Local	Char	Contact Department in Local Language

Table 15: Activity data fields complete table (Source: Author)

Activity Data Field	Data Type	Description
Customer ID	Char	Uniquely identifies a Customer Record
Customer Name	Char	Uniquely identifies a Customer Record.
Priority	Char	Priority of the Activity
Activity Type	Char	Activity Type
Activity Objective	Char	Activity Objective.
Activity Purpose	Char	General Purpose of the Activity; Maintenance, acquisition, penetration, retention
Planned Date	Date	If Activity Status is "Scheduled", then Planned Start date to be defaulted to current date (System Date).
Planned Completion	Date	Date that the activity has been planned for completion.
Description	Char	A further description of the Activity
Status	Char	The status of the activity. e.g. completed, scheduled, in progress etc.
Call Outcome	Char	The result of the Call linked to this Activity
Comments	Char	Comments related to the activity
Actual completion	Date	Date the activity was actually completed Date format is "YYYY-MM-DD HH24:MI:SS".
Start time	Time	The time the activity is planned to start. Time format is "HH24:MI:SS". This field should contain only the time format.
End time	Time	The time the activity is planned to end. Time format is "HH24:MI:SS". This field should contain only the time format.
Loaded Via Interface	Boolean	Flag to indicate the record was loaded via interface and not created through UI
Organization	Char	Default to the Organization of the Country to which the Customer belongs to
Activity UID	Char	Uniquely identifies an Activity record in CRM
Created	Date	Created date. Format of the date should be "YYYY-MM-DD HH24:MI:SS"
Created By	Char	Created By.
UTC Planned Start Date	Date	Planned Start Date in UTC format. Format will be "YYYY-MM-DD HH24:MI:SS"
UTC Planned End Date	Date	Planned End Date in UTC format. Format will be "YYYY-MM-DD HH24:MI:SS"

Table 16: Opportunity fields complete table (Source: Author)

Data Field	Data Type	Description
Customer ID	Char	Uniquely identifies a Customer Record

Customer Name	Char	Uniquely identifies a Customer Record.
Opportunity Name	Char	The name of the opportunity. This should reflect the opportunity and is unique under the customer site.
Opportunity Type	Char	Indicates the nature of the opportunity. E.g. penetration, acquisition
Potential Revenue	Number	The potential revenue of this opportunity.
Committed Revenue	Number	The committed revenue of this opportunity.
Pipeline Stage	Char	The stage this opportunity has reached.
Expected Close Date	Date	Date format is "YYYY-MM-DD HH24:MI:SS".
Actual Close Date	Date	Date format is "YYYY-MM-DD HH24:MI:SS"
Reason Lost	Char	The reason the opportunity was won or lost.
Lead Priority Level	Char	Have Values of High - Medium - Normal.
Source	Char	The name of the campaign that originated this opportunity.
Source Type	Char	The source type refers to the channel through which the lead was received.
Lead Originator	Char	The employee who sent the lead.
Lead Originator email	Char	The email of the employee who sent the lead.
Department Type	Char	The customer's department we are dealing with for this opportunity.
Competitor	Char	The competitor linked to this opportunity.
Last Update Date	Date	Date format is "YYYY-MM-DD HH24:MI:SS".
Reason for Lead	Char	Field to store the reason the lead was generated
Organization	Char	Default to the Organization of the Country to which the Customer belongs (Old company etc)
Currency Code	Char	Currency code of the organization the Customer belongs to
Created	Date	Format of the date should be "YYYY-MM-DD HH24:MI:SS"
Created By	Char	Created By.