BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF BUSINESS AND MANAGEMENT
FAKULTA PODNIKATELSKÁ

INSTITUTE OF INFORMATICS
ÚSTAV INFORMATIKY

DATA EXCHANGE AUTOMATION
AUTOMATIZACE DATOVÉ VÝMĚNY

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR — Adrián Spaček
AUTOR PRÁCE

SUPERVISOR — Ing. Jiří Kříž, Ph.D.
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF BUSINESS AND MANAGEMENT

Assignment Bachelor's Thesis

Department: Institute of Informatics

Student: **Adrián Spaček**

Supervisor: **Ing. Jiří Kříž, Ph.D.**

Academic year: 2022/23

Study programme: Managerial Informatics

Pursuant to Act no. 111/1998 Coll. concerning universities as amended and to the BUT Study Rules, the degree programme supervisor has assigned to you a Bachelor's Thesis entitled:

Data Exchange Automation

Characteristics of thesis dilemmas:

Introduction
Aim of the Thesis
Theoretical Background
Problem Analysis and Current Situation
Proposals and Contribution of Suggested Solutions
Conclusions
References
Appendices

Objectives which should be achieved:

The aim of this thesis is to design a data exchange automation system for a specific subject. The outputs of the work are focused on the issue of ETL processes on the SQL Server platform.

Basic sources of information:

BASL, J. a R. BLAŽÍČEK. Podnikové informační systémy - podnik v informační společnosti. 3. vyd.Praha: Grada Publishing a.s., 2012. 328 s. ISBN 978-80-247-4307-3.

BEGG, C., R. HOLOWCZAK a T. CONOLLY. Mistrovství - Databáze: Profesionální průvodce tvorbou efektivních databází. Praha: Computer Press, 2009. 584 s. ISBN 978-80-251-2328-7.

LABERGE, R. Datové sklady - Agilní metody a business intelligence. Praha: Computer Press, 2012. 352 s. ISBN 978-80-251-3729-1.

LACKO, L. Mistrovství v Microsoft SQL Server 2012. 1 vyd. Praha: Computer Press, 2013. 640 s. ISBN 978-80-251-3773-4.

MITCHELL, R. Web Scraping with Python: Collecting More Data from the Modern Web. 2. vyd. Massachusetts: O'Reilly Media, 2018. 300 s. ISBN 978-1491985571.

Deadline for submission Bachelor's Thesis is given by the Schedule of the Academic year 2022/23

In Brno dated 5.2.2023

L. S.

Ing. Jiří Kříž, Ph.D.	doc. Ing. Vojtěch Bartoš, Ph.D.
Branch supervisor	Dean

Abstract

This thesis has the goal of describing and analyzing the problematics of data exchange and automation of processes in the environment of a live database and external sources of information. In the digital domain, there exist numerous ways of acquiring, transforming and transporting data. The work aims to describe these, supplying arguments for and against their use in specific circumstances. Subsequently, a solution of an automated data exchange scenario is presented.

Key words

Data, system, process, exchange, import, advancement, technology, efficiency, API, script

Abstrakt

Cílem této práce je popsání a analýza problematik datové výměny a automatizace procesů v prostředí živé databáze a externích zdrojů informací. V digitálním prostředí existuje mnoho způsobů získávání, transformace a transportu dat. Tato práce má za cíl jich opsat, dodávajíc argumenty pro a proti jejich použití v specifických situacích. Následně je prezentováno řešení situace vyžadující automatizaci datové výměny.

Rozšířený abstrakt

Tato bakalářská práce se zaměřuje na automatizaci datové výměny pomocí jazyka Python pro stažení dat a T-SQL pro import a transformaci dat. Cílem práce je navrhnout a implementovat řešení, které umožní automatizovanou výměnu dat mezi společností a jejím klientem, za účelem snížení nákladů a zvýšení efektivity procesů datové výměny.

V první části práce bude popsáno teoretické pozadí i související technologie, které jsou důležité pro porozumění tématu a řešení. Budou diskutovány různé přístupy k automatizaci datové výměny a technologie, které se používají pro zajištění bezpečnosti a kvality dat.

Začneme vysvětlením základních pojmů jako jsou data, informace a znalosti. Podíváme se také na databázové a informační systémy a jejich význam ve společnosti. Dále se budeme zabývat konkrétními technologiemi a postupy pro automatizaci datové výměny, jako jsou například ETL (Extraction, Transformation, and Loading) nástroje, které umožňují extrahovat data z různých zdrojů, transformovat je do požadovaného formátu a načíst je do cílového systému za předem specifikovaných podmínek. Budou popsány faktory ovlivňující kvalitu dat a vlastnosti, které tyto data mají.

V této části se také zaměříme na technologie, které obecně umožňují online komunikaci, jako jsou protokoly použity při samotné výměně dat vrámci vrstev referenčního modelu ISO. Popíšeme také koncepci a využití API jako způsobu výměny dat na různých místech online sféry.

Zmíníme se o bezpečnostních aspektech automatizace datové výměny, konkrétně na příkladu protokolů určených k transportu souborů jako jsou FTP a SFTP včetně šifrování, a o dalších technologiích pro zajištění ochrany dat při jejich užití.

Část končí popisem technologii souvisejících s tvorbou HTTP dotazování a internetem jako platformou pro výměnu dat.

První část práce tedy poskytne teoretické pozadí a vysvětlí důležité technologie a koncepty, které jsou součástí práce.

Druhá část této bakalářské práce se zaměřuje na analýzy, které jsou důležité pro plánování implementace automatizace datové výměny ve společnosti. Konkrétně jsou použity modely 7S, PESTLE a SWOT, aby byla provedena hloubková analýza společnosti a umožněno tak úspěšné nasazení nového systému.

Model 7S analyzuje sedm klíčových faktorů, které jsou důležité pro úspěšnou transformaci organizace: strategii, strukturu, systémy, styl vedení, spolupráci, zaměstnance a sdílené hodnoty. Zkoumání těchto faktorů umožňuje identifikovat jakékoliv nedostatky, které by mohly omezovat úspěch implementace nového řešení pro automatizaci datové výměny a obecně operaci společnosti ve které je řešení realizováno. Analýza 7S byla provedena abychom se ujistili, že všech sedm faktorů je plně zohledněno a optimalizováno pro úspěšnou implementaci.

PESTLE analýza pak zkoumá politické, ekonomické, sociální, technologické, právní a environmentální faktory, které mohou ovlivnit způsob, jakým společnost funguje. Tato analýza se zaměřuje na to, jak by změny v těchto činitelích mohly ovlivnit plánování a provoz společnosti. Některé faktory, jako například změny zákonných předpisů týkajících se ochrany osobních údajů, mohou mít významný dopad na to, jak budou data ve společnosti zpracovávána a zabezpečována. PESTLE analýza umožní provést podrobnou kontrolu a přizpůsobit se případným změnám v prostředí ve kterém firma operuje.

SWOT analýza je použita pro identifikaci silných a slabých stránek společnosti, stejně jako pro nalezení příležitostí a hrozeb v souvislosti s jejím provozem. Interní a externí analýzy byly provedeny, za účelem identifikování faktorů ovlivňujících konkurenční schopnosti společnosti, jako například vývoj technologií, vývoj finanční situace a zaměstnanecké zdroje. SWOT analýza umožní lépe porozumět současnému stavu společnosti a identifikovat místa, kde je potřeba zlepšení.

Poslední sekce práce se zaměřuje na analýzu současného procesu, cíle nového automatizovaného procesu a prostředí, ve kterém se proces odehrává. Následně

popisuje technické aspekty nového řešení, přičemž konkretizuje bezpečnostní uvážení při nasazení. Díl končí určením možností dalšího rozšíření projektu a ekonomickým zhodnocením navrhovaného řešení.

Tato část popisuje každý krok v průběhu tvoření nástroje na extrakci, transformaci a následné načtení dávek dat z klientova úložiště do databáze pro další zpracování.

Současný proces datové výměny v nejmenované společnosti je manuální a zahrnuje mnoho ručních kroků, což vede k chybám a zdržením. Cílem nového automatizovaného procesu je odstranit co nejvíce ručních kroků a snížit čas potřebný pro výměnu dat. Tento proces je klíčový pro společnost, protože se jedná o základní krok pro řadu důležitých funkcí souvisejících s daty jednoho z nejdůležitějších klientů.

Technickým řešením pro automatizaci datové výměny je použití Pythonu a T-SQL skriptů, které dokáží automatizovat stažení, import a transformaci dat. Bezpečnostní aspekt nového řešení je kritickým bodem, a proto jsou v rámci bezpečnosti implementovány opatření jako šifrování dat a logování přístupů uživatelů.

Dále jsou popsány možnosti rozšíření projektu, jako například implementace systému validace osobních údajů, který pomůže společnosti splnit požadavky na ochranu osobních údajů a zároveň odstraní další ruční krok z procesu výměny. Další popsanou možností rozšíření projektu je integrace ChatGPT pro automatické získávání dat pomocí rozeznávání struktur. Jako příklad je uvedeno získání PSČ z různě naformátovaných textových řetězců adres.

V závěru práce je provedeno ekonomické zhodnocení nového řešení. Je zohledněna cena vývoje a úspory nákladů na pracovní sílu s předpokládaným relativním nárůstem zisků z tohoto procesu. Výsledkem nákladové analýzy je, že společnost může očekávat rychlejší, spolehlivější a levnější výměnu dat, na základě čeho je jí doporučena implementace navrženého řešení.

Hlavním přínosem této práce je vytvoření fungujícího řešení, které zprostí zaměstnance zpravující databázi, kteří by jinak byli nuceni všechny úlohy vykonávat ručně, obírajíc je o čas a tím pádem společnost o peníze. Čtenáři práce mohou následně získat vědomosti o způsobech tvoření řešení k obdobným zadáním.

Klíčová slova

Data, systém, proces, výměna, importování, pokrok, technologie, účinnost, API, skript

Bibliographic citation

SPAČEK, Adrián. *Data Exchange Automation* [online]. Brno, 2023 [cit. 2023-04-30]. Available at: https://www.vutbr.cz/studenti/zav-prace/detail/142310. Bachelor's Thesis. Brno University of Technology, Fakulta podnikatelská, Ústav informatiky. Supervisor Ing. Jiří Kříž, Ph.D.

Affidavit

I declare that the present bachelor project is an original work that I have written myself. I declare that the citations of the sources used are complete, that I have not infringed upon any copyright (pursuant to Act. no 121/2000 Coll.).

Brno dated 30th Apr 2023

Adrián Spaček

author's signature

Table of contents

Introduction

With the rise of online communication in big part due to the introduction of numerous social networks, the need for file sharing systems has been increasing rapidly. Whether it is sharing of photos and video files from the significant moments in a person's life, seamlessly streaming music and films when commuting or sharing documents in official communication. Although this kind and scale of information exchange may seem recent, in the corporate world, the shift towards systematic and instant exchange of information has been happening for far longer. This is mainly caused by the fact that the implementation of such mechanisms provides an extensive competitive advantage when it comes to processing information. Companies soon recognized the value of being able to exchange data with their business partners, subsidiaries and of course clients and were in a timed race to implement such systems into their business model. Those successful saw increased margins due to lower costs. Costs dropped as a result of there not being a need to manually transfer and process data, freeing up labor for other tasks. Nowadays, automated data exchange is almost a necessity for companies that wish to be competitive in one form or another. It is therefore beneficial for them to know what options they can use to solve which tasks. As soon as there is a need for a database, there is subsequently the need for a way of filling it with information and perhaps for its further distribution. This is what automated data exchange options can achieve.

Aim of the thesis

The aim of this thesis is to address the inefficiencies present within the process of data exchange in the specified company by way of automation which is meant to decrease labor costs and increase reliability.

The first part is used for explaining the terms and technologies used when dealing with data exchange solutions. It delves into the meaning and use cases for these and should serve as a basis for understanding the problematics.

The second part describes the current situation and structures within the company and several types of analysis outline the positives and negatives that apply to it from the internal and external factors, to its strengths and weaknesses, to the threats it faces and the opportunities it should be on the lookout for.

The third and main part of the thesis is aimed specifically at the technologies of automation and the grasping of those opportunities by describing a proposal for automation of one of the company's processes and by extention for lowering the labor costs that specific process entails. This part also describes the exact steps this improved process takes, as well as the environments in which it takes place and the technologies it uses and attempts to approximate the financial gain from its implementation and its overall viability of application.

1 Theoretical Background

This part of the thesis is reserved for the description of basic terms used throughout the next parts and introduction of the environment the solution is to be implemented in. These are needed to fully grasp the extent of the solution.

1.1 Data, information and knowledge

Data by itself is a raw resource, and we store it in the hope of it having a certain value for us. Data has several inherent properties:

- Data is factual – meaning there is no way to interpret it.
- Data is collected – the existence of data itself is due to someone bothering to collect it for its potential value.
- Data is stored – data always has a storage medium, this can be as simple as a log of sales in a notebook.
- Data is accessible – data is collected to be used in the future and must therefore be accessible in the future.
- Data is unchanged – data remains in the state it has been collected. Because of this it has integrity.

Data becomes information once we seek patterns within it and give these patterns meaning. Information is a meaningful interpretation of data. While data is factual, we define important data characteristics and information is the meaning we choose to give them. For example if we have a set of animal pictures and we ask which pictures are of cats, the pictures of cats from this dataset become relevant information because we chose to give these specific data points a meaning [33].

Knowledge can be a skill or a theoretical understanding of a subject. It is how we analyze and structure information in relation to other information. To know means to have an idea of how information relates and being able to perform operations within this structure, to create representations for information patterns and to be able to make decisions based on these. For example, if data is a set of animal pictures and information is which are the pictures of cats, then the knowledge can be what a cat looks like and a use of this knowledge can be the ability to recognize images of other cats among animals [33].

1.2 Databases and information systems

The following part describes the prerequisites for implementing a complex solution for the task of information exchange commonly known as Electronic Data Interchange.

„Electronic Data Interchange (EDI) is the computer-to-computer exchange of business documents in a standard electronic format between business partners.“[26]

1.2.1 Databases

The database is a single, possibly large repository of data, which can be used simultaneously by many users. All data that is required by these users is integrated with minimal duplication. And importantly, the database is normally not owned by a single department or user but is a shared corporate resource.

As well as holding the organization's operational data, the database also holds a description of this data. This is called meta-data and is what creates the database's self describing nature. This also provides what is called data independence. This means if structures are created or altered that are not directly involved in the functioning of an application, this application will not be affected by these changes.[2]

The interconnection of systems through a common database is one of the core drives in the innovation of business information systems employed from the nineteen nineties onward. Data accessible whenever and from wherever necessary has represented the technological shift away from file oriented data processing. Relational databases have helped to make business data available online, lowering costs of material, shortening spans to realize orders and making faster and more accurate decision making possible. The first effects were felt in the operative transactional applications of ERP systems and later, more advanced applications of information systems were affected, making business intelligence, datamining and web mining tools available for all sorts of data analysis.[1]

1.2.2 Information systems

Information systems (IS) involve a variety of information technologies (IT) such as computers, software, databases, communication systems, the Internet, mobile devices and much more, to perform specific tasks, interact with and keep informed all the

various specified actors in different organizational or social contexts. However, the IS field is not primarily concerned with the technical and computational aspects of IT. What matters to the IS instead is how technology is utilized in order to allow the IS to fulfill the information needs and requirements of those various actors – individuals or groups - in regards to specific goals and practices [24].

The definition of IS according to Symons: *"The system utilizes computer hardware and software; manual procedures; models for analysis, planning, control and decision making; and a database. The emphasis is on information technology (IT) embedded in organizations"* [25, p. 181].

1.2.3 Exchanged data problematics

The information gathered in exchanges often has encoded data, primary keys with multiple parts. Example being an identifier that contains the year and location of origin. There are several issues that regularly occur when transforming data:

- Uncertainty of data
- Missing values
- Duplicates
- Non uniform currencies, names, objects, concepts, numeric and text formats
- Reference integrity
- Missing date

Uncertain data is one of the most numerous issues that show up in datasets. Mostly when a case parameter is being specified - gender being stated as Male/Female or 1/0 or M/F.

Missing values are an issue that together with duplicates lower the data quality, however duplicates are much less of an issue since additional data can be easily added. Resolving missing values in columns can sometimes be done by hand by supplying it from other sources, or ignored if the missing data is not critical.

Non uniform data includes differing numbers of decimal places in numeric values or for example negative values in a column that is only supposed to contain positive integers. Monetary data is often mashed together and the values can require converting them to a single currency for uniformity's sake. Sometimes, data that is used to specify an address

for instance 305/10 can be recognized as a mathematical expression. Same goes for bank accounts.

Reference integrity issues can arise when the organizational structure of a company is changed – if a department is dissolved and employees remain without it specified, their records can skew data and negatively affect its quality.

Date is often the most important metric for a single record in a database - when an order was made or when a transaction was carried out. These have to be present in the data batch before it‘s loading. [4]

1.2.4 Choosing the correct solution for the given task

The reasons for choosing a data exchange method are rarely clear, and they frequently necessitate balancing the technique's benefits and drawbacks, as well as local and corporate requirements. When it comes to data exchange, there is no such thing as a "one-size-fits-all" solution. The following factors should be taken into account [6]:

1.2.4.1 Data set characteristics:

Data Complexity

Direct database access may be the most effective approach when the data entity to be transferred has numerous linked elements or the specific components are unknown in advance, i.e. the required data items have to vary if necessary [6].

A REST API for example is entity-based, which is one of its major design principles. While this has the benefit of providing a consistent location for each entity (e.g., Plan 123 is always found at /plans/123), it has the downside of making it more difficult to connect multiple related things. To reassemble the relationships among the various data parts, an API technique may necessitate many calls and coding. It's worth noting that using an integration platform or enterprise service bus can help with data complexity.

It's vital to remember that flat files are 'flat' in terms of data formatting and can't easily handle hierarchical data. Although JSON and XML can express more complicated data models, the REST architecture was created to avoid dealing with complex query and result data. [6]

Frequency of data update

The cost of replacing a whole dataset via file transfer or direct database access can be significant. These problems are amplified if the data set is updated often (and the number of updates is huge). APIs and messaging system methods are likely superior solutions in this circumstance because they handle transactional updates more readily and avoid continuous bulk resynchronization. [6]

Data set size

For performance reasons, transferring very big data sets often necessitates the use of a file transfer or a direct database connection. Although there are approaches for improving performance when transferring massive data sets or high numbers of large messages using REST or comparable APIs, other methods are often preferred. [6]

1.2.4.2 Data environment characteristics:

Data flows and breadth of solution

What is the process for moving data from one application to another? The best data sharing techniques will be determined by analyzing the numerous planned and possible data flows.

Middleware that mediates one-to-one, one-to-many, and many-to-one interactions is commonly used in the message broker approach. The Message Broker design can be used to create extremely massive and high-performance push-based fan-in / fan-out systems. Asynchronous communication, unreliable networks, and big data applications all benefit from this.

A streaming method is likely the ideal approach in circumstances where a large number of data sources continuously feed data to a single receiving system, such as log or other instrumentation data. [6]

Frequency of data usage

Is there any advantage to using real-time data? If you need access to the most up-to-date version of the data, you'll need to use a synchronous remote procedure call or API. [6]

Data versions

When a data provider needs to deliver different versions or schemas to support different consuming applications an API may not be the best choice. An API is, ideally, a single

consistent representation of a set of resources. Maintaining multiple schemas or versions in a single API is complex and will often accrue technical debt within a codebase. [6]

Data security

The tools and techniques required to safeguard data while it is in transit and at rest are typically external to the data sharing method. For example, an API can be created to require a key, a direct database connection can be restricted by security restrictions in the database management system, and web servers can be configured to secure data files, among other things. Although not directly related to the data sharing mechanism, it is critical to consider the required restrictions and alternatives when deciding between two options. [6]

Data transformation complexity

A direct database connection utilizing ETL tools is suggested if the data interchange requires extensive processing and data transformation, especially if the transformations are based on sophisticated business requirements. This is especially true if the transfer's goal is to shift data from one location to another. Some API management solutions can also perform less substantial modifications, therefore API methods may be useful depending on the unique requirements [6].

Certain operations are required when ultimately loading data to the new location. An important step when loading data can be the blocking of immediate indexing, which could depending on the size of the dataset significantly slow the loading process. Indexing which allows a more efficient search should in these cases be created only after the data is already stored in the new location. Among the most important steps in the final phase of ETL is usually the creation of replacement primary keys. Data concerning a single person can be stored in multiple forms with multiple primary keys. These unintentional duplicate records should therefore be consolidated under a single new primary key in the new storage location [3].

Connection persistence

Long-lived protocols are those that are designed to keep the connection open indefinitely. SSH is an example of such a protocol. Short-lived protocols are more

transactional in nature; they perform a single operation, or a sequence of activities, and then terminate the connection. HTTP POST is a good example.

Long-lasting connections are especially beneficial for streaming data to end-user clients like browsers or mobile devices. They can also be beneficial within the network when specific receivers, such as job processors, are not addressable. WebSockets is a protocol for creating bi-directional, long-lived connections [6].

Scope Constraints

Every project has limitations, and choosing a data transfer technology is no exception. The fundamental scope triangle of time, cost, and quality cannot be overlooked at the highest level. Time refers to the amount of time available to complete the project, cost refers to the amount of money or resources available, and quality refers to the project's fit-for-purpose. One or more of these elements is usually constant, while the others vary. Reduced completion time, for example, will have an impact on quality and/or prices.

Factors such as available technical skills, business strategies and organizational culture may also represent constraints. In addition, it is unlikely that all, or even many, of the data exchange methods discussed above will be supported in a particular case. This is particularly true in the case of Software as a Service (SaaS) applications where the customer has no control over the data exchange methods available in the product. However, after taking these larger considerations into account, more than one option may remain. This discussion is focused on those cases [6].

Organizational Considerations

It is important to view individual project decisions within this enterprise data management framework and to balance project and application specific requirements with broader organizational requirements. Uncoordinated approaches by various segments of the organization can result in data conflicts and quality inconsistencies that reduce efficiency and stifle innovation.

Three of the basic data exchange mechanisms listed above, file transfer, direct database connection and remote procedure calls have traditionally been used to allow different applications and systems to communicate and exchange data. Unfortunately, because each of these approaches requires detailed knowledge of the operational database or

application involved, they are tightly coupled and difficult to change. More importantly, as the number of individual point-to-point exchanges grow, the overall environment becomes increasingly complex and difficult to manage over time. Database links in particular are normally created and maintained by external groups. Wide use of this approach can lead to a substantial access management burden. While there are circumstances in which point-to-point custom integrations are appropriate, they should be carefully considered as they are difficult to evolve based on changing requirements.

Brokered Messaging and Web Services more easily support wider enterprise data integration designs such as the Publish/Subscribe and Gateway patterns. These, and other similar patterns can be used to isolate applications and databases from one another by using a middle service layer to decouple systems. This provide a number of advantages including increased flexibility, better visibility, reduced administration costs, reduced dependencies and the ability to support real time updates [6].

1.2.4.3 Consumer characteristics:

Human beings and front-facing applications

Text files, such as the 'comma delimited file' format, are easily readable by humans and may be used with popular tools. If this type of direct use is the most prevalent scenario, file transmission is the best option. Similarly, while APIs are most commonly used by programmers, they typically deliver text or hypermedia. REST APIs are a good choice if the receiving system is front-facing, such as a web browser or equivalent agent. Optimized RPC protocols, rather than REST APIs, are more likely to benefit systems sharing confidential data and offering 'back-end' services [6].

Receiving system processes

Assumptions built into a receiving system related to the business processes it supports may make one or another exchange method a better choice. For example, the designers of a system oriented to batch processing of transactions may have assumed that that data transfers are always file based. While selecting an alternative data exchange method may be possible, the cost/benefit ratio may not be favorable [6].

Usage by the receiving system

Is the information being utilized to support a feature or to build a platform? If the data is being utilized to support a 'feature' that addresses a specific need, such as a person query to return a set of attributes, an API is most likely the best option. A file or database solution, on the other hand, would be more appropriate if a huge dataset is being transferred and utilized as the foundation of a 'platform' or reporting system [6].

1.3 Technologies used in information Exchange in the OSI model

This part describes a couple of technologies which can be used to set up an information exchange process within a company.

1.3.1 Transport layer protocols

TCP

The Transmission Control Protocol fits into a layered protocol architecture just above a basic Internet Protocol which provides a way for the TCP to send and receive variable-length segments of information enclosed in internet datagram "envelopes". The internet datagram provides a means for addressing source and destination TCPs in different networks. The internet protocol also deals with any fragmentation or reassembly of the TCP segments required to achieve transport and delivery through multiple networks and interconnecting gateways. The internet protocol also carries information on the precedence, security classification and compartmentation of the TCP segments, so this information can be communicated end-to-end across multiple networks[7].

When two processes wish to communicate, their TCP's must first establish a connection (initialize the status information on each side). When their communication is complete, the connection is terminated or closed to free the resources for other uses. Since connections must be established between unreliable hosts and over the unreliable internet communication system, a handshake mechanism with clock-based sequence numbers is used to avoid irregular initialization of connections [7].

Although TCP establishes a connection, this by itself does not mean this connection is secured. For example FTP (File Transfer Protocol) works over TCP/IP but sends unencrypted data. Security is usually handled by separate protocols such as SSH-TLP (Secure Shell Transport Layer Protocol) or TLS (Transport Layer Security). FTPS (File

Transfer Protocol Secured) or SFTP (SSH File Transfer Protocol) are then the secured versions [7].

UDP

UDP (User Datagram Protocol) provides a procedure for application programs to send messages to other programs with the minimum of protocol mechanisms – providing minimal latency. The protocol is transaction oriented, and delivery and duplicate protection are not guaranteed. Applications requiring ordered reliable delivery of streams of data should use the Transmission Control Protocol [8].

UDP is usually applied in situations where the importance lies in the on-time delivery of data rather then its accuracy. Typical applications include audio and video streaming and online multiplayer games.

1.3.2 Session layer

API

An API (Application Programming Interface) defines rules of communication between computers or applications. APIs sit between an application and the server creating an intermediary layer for data transfer processing.

First, the client's application initiates an API request that is processed by its URI (Uniform Resource Identifier) containing request verbs, headers and the body if data is being posted.

Verbs used include GET, POST, HEAD, PUT, PATCH and DELETE. The last 3 are sometimes used for replacing, modifying or deletion of entire collections [9].

The GET method requests transfer of a current selected representation for the target resource. GET is the primary mechanism of information retrieval and the focus of almost all performance optimizations. Hence, when people speak of retrieving some identifiable information via HTTP, they are generally referring to making a GET request.

The POST method requests that the target resource process the representation enclosed in the request according to the resource's own specific semantics [9].

The HEAD method is almost identical to GET however the message body must not be sent by the server. GET and POST together function as main parts of the data exchange between subjects [10].

After a valid request is received by the API, it makes a call to the external program or server. Subsequently the server sends a response with the requested information and then the API transfers the data to the initial requesting application.

APIs offer security by design because their position between systems facilitates the abstraction of functionality between those systems. The API endpoint decouples the consuming application from the infrastructure providing the service. API requests usually include authorization credentials to reduce the risk of attacks on the server, and an API gateway can limit access to minimize security threats. Also, during the exchange, HTTP headers, cookies, or query string parameters provide additional security layers to the data [10].

There are several services that use an API for distinctly different purposes. An API example is the function that enables people to log in to websites by using their Facebook, Twitter, or Google profile login details. This feature allows any website to leverage an API from one of the widely used services to quickly authenticate the user, saving them from setting up a new profile for every website service or new membership. Another use is the PayPal payment option that can be found on many ecommerce sites. This allows for safe transfer of funds as the API does not expose sensitive data nor does it grant access to unauthorized users. The Google Maps service uses multiple APIs concurrently for displaying static and dynamic maps, directions or points of interest. Through geolocation, you can communicate with the Maps API when plotting travel routes or tracking items on the move, such as a delivery vehicle.

With the popularization of APIs, several protocols have been introduced that facilitate the de facto standardization of data exchange:

SOAP (Simple Object Access Protocol) is an API protocol built with XML, enabling users to send and receive data through SMTP and HTTP. With SOAP APIs, it is easier to share information between apps or software components that are running in different environments or written in different languages.

XML-RPC is a protocol that relies on a specific format of XML to transfer data, whereas SOAP uses a proprietary XML format. XML-RPC is older than SOAP, but much simpler, and relatively lightweight in that it uses minimum bandwidth.

JSON-RPC is a protocol similar to XML-RPC, as they are both remote procedure calls (RPCs), but this one uses JSON instead of XML format to transfer data. Both protocols are simple. While calls may contain multiple parameters, they only expect one result.

REST (Representational State Transfer) is a set of web API architecture principles, which means there are no official standards. To be a REST API, the interface must adhere to certain architectural constraints. It's possible to build RESTful APIs with SOAP protocols, but the two standards are usually viewed as competing specifications [10].

1.3.3 Presentation layer

FTP

FTP is a client-server protocol and relies on two communication channels between the client and the server: a command channel used to control the conversation and a data channel used to transfer file content. The client initiates a dialogue with the server by requesting to download the file. Using FTP, the client can upload, download, delete, rename, move, and copy files on the server. Users usually need to log in to the FTP server, although some servers can make some or all of the content available without logging in. This is called anonymous FTP [11].

FTP sessions work in passive or active mode. In active mode, after the client requests to start a session through the command channel, the server will start a data connection with the client and start transmitting data. In passive mode, the server instead uses the command channel to send the information needed to open the data channel to the client. Because passive mode enables clients to initiate all connections, it works well between firewalls and network address translation (NAT) gateways [11].

Users can use FTP through a simple command-line interface (for example, from the console or terminal window of Microsoft Windows, Mac-OS, or Linux) or a dedicated graphical user interface (GUI). Web browsers can also be used as FTP clients [11].

SFTP

One of the widely used methods of secured data transfer is using SFTP (SSH file transfer protocol) through one of its available clients or servers:

- Tectia SSH client
- WinSCP client
- FileZilla client
- PuTTY client
- Cyberduck client
- Tectia SSH server for Windows
- Open SSH server

SSH itself is a protocol that works on a client – server model (Same as SFTP) meaning the connection is established by the SSH client to the SSH server. The SSH client uses public key cryptography to verify the identity of the SSH server. After the connection has been set up, SSH uses symmetric encryption and hashing algorithms (Such as 128 bit AES and SHA) to ensure the security and integrity of the data being exchanged between the client and the server. One of its advantages is the fact that SFTP operates only on the SSH dedicated port number 22 as established by IANA (Internet Assigned Numbers Authority) without the need of specifying other ports as is the case with FTPS [12,17].

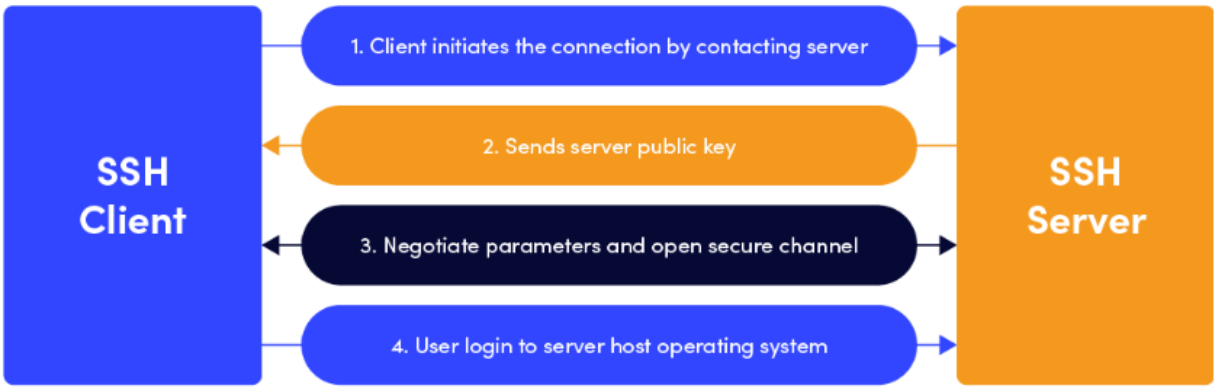

Image 1: Interaction between the client and server sides of SSH

(Source: 12)

In terms of ciphers used, SSH uses the required 3des – applying the DES (data encryption standard) 3 times to all blocks and typically uses 1 optional cipher, aes128 being the recommended one [13].

SHA or Secure Hash Algorithm is used to hash the message authentication code or MAC – creating an HMAC. Any attempt to alter this message results in failed authentication due to a different digestion of the message. These algorithms are considered secure due to the computational infeasibility of trying to find a message with the corresponding digest or trying to find 2 messages with the same digest [14].

The protocol supports multiple concurrent operations. Each operation is identified by a unique number assigned by the client, and servers response contains the same identifying number. Server may process requests asynchronously and may return responses out-of-order. For performance reasons, file transfer clients often send multiple requests before stopping to wait for responses [13].

1.4 Additional terminology

Next, some terms used throughout the world of information exchange will be described.

1.4.1 The Internet – ensuring communication

The basis for the functioning of transfer systems is the ability to connect individual parts to a common network – the internet. The internet is based on a series o protocols, namely TCP/IP, which enable communication and data exchange. One of the crucial services is WWW or World Wide Web. Its characteristic is the interconnection of sites with links, adhering to the hypertext principals [15].

WWW as a service is based on 3 main pillars – the HTML language, the HTTP protocol and addressing through URL. Now we will take a look at these [15].

1.4.2 HTML

The Hypertext Markup Language (HTML) is a simple markup language for creating platform-independent hypertext texts. HTML documents are SGML documents that have generic semantics and can be used to represent data from a wide range of areas. HTML markup can be used to create hypertext news, mail, documentation, and hypermedia; menus of options; database query results; simple structured documents with in-lined graphics; and hypertext views of existing data sets [16].

1.4.3 HTTP

HTTP (The Hypertext Transfer Protocol) is a stateless application-level protocol for hypertext information systems that are dispersed and collaborative. The semantics of HTTP/1.1 communications are defined in this document, as stated through request methods, request header fields, response status codes, and response header fields, as well as the payload of messages (metadata and body content) and content negotiation mechanisms [9].

The request method token is the primary source of request semantics. It indicates the purpose for which the client has made this request and what is expected by the client as a successful result. By convention, standardized methods are defined in all-uppercase US-ASCII letters [9].

Method	Description
GET	Transfer a current representation of the target resource.
HEAD	Same as GET, but only transfer the status line and header section.
POST	Perform resource-specific processing on the request payload.
PUT	Replace all current representations of the target resource with the request payload.
DELETE	Remove all current representations of the target resource.
CONNECT	Establish a tunnel to the server identified by the target resource.
OPTIONS	Describe the communication options for the target resource.
TRACE	Perform a message loop-back test along the path to the target resource.

Image 2: Commonly used HTTP request methods

(Source: 9)

1.4.4 URI/URL

A URI (Uniform Resource Identifier) is a short string of letters that identifies a digital or physical resource.

„The term "Uniform Resource Locator" (URL) refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource by describing its primary access mechanism (e.g., its network "location").“ [18, p. 7]

2 Problem Analysis and Current Situation

In the following part, the company in question will be introduced, its structure together with its information systems. After that follows the analysis of the current state of the company.

2.1 Introduction of the company

In accordance with the anonymity of the thesis and taking into account the requests of the company, names and other specifications in this thesis will be omitted. Since the aim of the thesis is directly linked to its internal systems and processes, its competitive ability could become compromised. The area of business of the company is debt collection. Its one of the most important debt collection agencies in the Czech Republic. Its main activities include in court and out of court settlement of debts, call center operation devoted to these, various court proceedings and the acquisition of portfolios with their subsequent management. After acquiring lists of items from a client, initially an out of court approach is taken for the collection of owed funds. Call center operators work with personal information that was given to the company by the client or was obtained from one of the freely accessible public registries. They contact the debtor and attempt to strike a deal that specifies payment intervals. If the deal cannot be made, the item is taken over by the department specializing in in court dealings lead by the main attorney of that specific debt portfolio. Depending on the type of case, the item is handled by specific consecutive processes. The company also manages a significant number of purchased investment portfolios [34,35].

2.2 Organizational structure of the company

Now, the structure and systems used throughout the company‘s processes will be described to show what kind of an environment the project in this thesis would have to be implemented in.

2.2.1 Departmentalization

The organizational structure can be divided into four blocks. At the top of the company stand the owners along with the CEO (Chief Executive Officer). The 4

functional blocks are comprised of the department for out of court dealings, the in court dealings department, the IT department and the financial and human relations department [34].

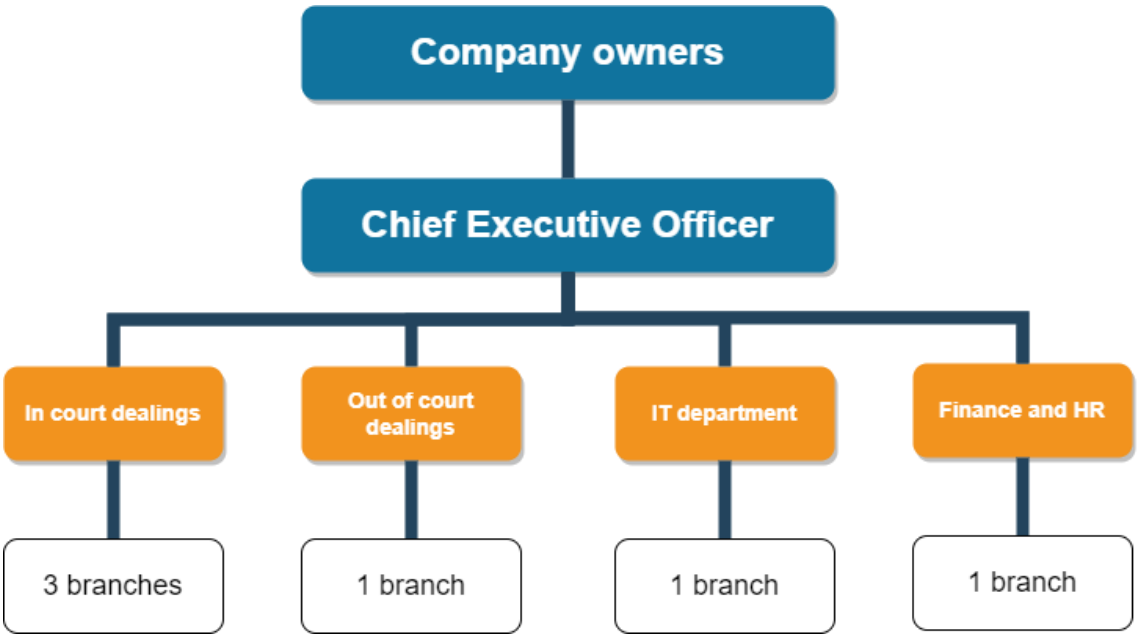

Chart 1 : Organizational structure of the company

(Source : Own solution based on [34])

The out of court department is made of its director, which oversees its entirety, individual operators and employees in administrative functions. The operators are part of the call center, the administrative work force takes care of all email communication and other correspondence [34,35].

The in court department is divided between 3 branches. Each one specializes in a certain type od debt portfolios. All of them have a specific set of employee positions. Each of the branches is by itself helmed by a leading attorney. Them and other leading attorneys together with additional attorneys, apprentices and paralegals make up the main part of the collection force. These are then backed up by case managers and similar administrative employees. Law students working as apprentices often continue in the company full-time after completing their studies. Even many of the attorneys now managing entire portfolios have begun working in the company during their university studies [34,35].

Support for the IS, along with ensuring data exchange with individual clients is handled by the IT department. It also handles the regular update of data itself. The

department is made of database specialists. These work closely with the supplier of the IS. This system is only managed by the department in a limited way. Front-end changes are mainly made by the supplier of the IS, while the data manipulation and management processes are overseen by the IT department [35].

The financial and HR department facilitates all financial services in the company, audits, or invoicing of customers. HR involves itself in personal needs, acquisition of new employees and filling of vacancies [34].

2.2.2 Utilized systems

Although some departments use systems specifically for their purposes that do not meaningfully affect the goal of this thesis. There are however systems used across most departments and most positions. Among the most important ones are [35]:

- Microsoft Teams – its use has been widely adopted approximately 3 years ago. It's used by employees for common daily communication, allowing online meetings to take place, which became especially vital during the pandemic that began during 2019, after which many employees were forced to work from home.
- Microsoft Outlook – this e-mail client is used mainly when communicating with representatives from client's ranks or when sending information to large groups within the company. It also supports the planning of meetings, shared event schedules and much more. Its interconnectivity with other applications in the Microsoft suite makes it the optimal choice since the company deploys more and more of its products.
- CRM – a direct communications channel for clients that is enveloped by the main IS of the company.
- Formic – the main ERP (Enterprise resource planning) IS of the company that is more closely described in the chapter 2.3.4. on Systems of the following 7S analysis.
- Other Internally developed apps include systems for work attendance management, work shift management and for the days-off records of individual employees.
- Other systems mainly include products of the Office 365 suite [35].

The company uses the infrastructure of a law firm it has close ties to. This firm then manages every piece of hardware this company uses in terms of procurement and lifetime management. They also manage the company network and servers and help with implementation of new systems when needed [34,35].

2.3 7S analysis

The 7s model, also known as the McKinsey 7s model, is a framework for analyzing an organization's structure and strategy. It is used to assess how well an organization's various components (the "7 S's") are aligned and working together to achieve its goals. The 7 S's are:

1. Strategy: The overall plan for achieving an organization's goals.
2. Structure: The way in which the organization is organized, including its formal reporting relationships and decision-making processes.
3. Systems: The formal and informal procedures and processes that the organization uses to manage its operations.
4. Shared values: The core beliefs, principles, and philosophy that guide the behavior of the organization and its employees.
5. Skills: The abilities and knowledge of the organization's employees.
6. Style: The leadership and decision-making style of the organization's management.
7. Staff: The people who work for the organization.

To conduct a 7s model analysis, information about the organization's strategy, structure, systems, shared values, skills, style, and staff will be needed. This information can be gathered through interviews with key stakeholders, surveys, and document reviews. Additionally, it's important to keep in mind that the analysis should be done by keeping in view the context of the organization and the industry it operates in.

Once this information is gathered, you can use it to identify any misalignment or gaps between the 7 S's and develop recommendations for addressing them. It's important to keep in mind that the 7s model is a framework, not a "one size fits all" solution, and that the analysis needs to be tailored to the specific organization and context.

In the following part of the thesis, 7S will shed light on the inner workings of the company.

2.3.1 Strategy

The company's vision is to become a dominant force in its field of business, offering clients the best way of recuperating the funds they are owed, and outperforming the competing companies by having a higher percentage of successfully collected debts [34].

According to this, its aim is to strengthen its business and working relations with its clients, allowing it to maximize its efficiency and developing solutions to struggles it encounters while reaching the client's demands [34].

The company is set on developing an ongoing mutually beneficial relationship with every one of its customers and by doing so is striving to make its vision a reality [34].

2.3.2 Structure

The company has over the years of operation developed its own structure model meant to streamline the management process. It has a matrix structure for the allocation of employees. This means that individual employees can be assigned to work on processes for a certain client. During this time, they remain on their respective positions in the context of the organizational structure. They complete assignments predominantly for their specified client, but can be temporarily assigned tasks that serve the company in a more general sense. In their individual assignments, they report to a leading lawyer – a so called officer assigned for that specific clients needs. The company also possesses a call center. Operators of which report to the leader of the out of court department [34,35].

These leaders – officers - of their respective departments all fall under the CEO (Chief Executive Officer). On the highest level, work is discussed between the CEO and the owners of the company. Financial analysis is done by a standalone financial analyst. Executive power is decentralized and the company always realizes the cost of expanding its scope of operation. It also realizes the need for automating processes in order to drive the costs attached to employees down. The aim of this very bachelor's

thesis is to advance this front and lower costs by automating a process that has been done by hand beforehand [34,35].

2.3.3 Staff

The company tries to limit the effect of differences in work experience and authority among individuals to ensure the smoothest and least combative work environment possible. The same approach is kept when hiring. The leadership always attempts to remove barriers across the diverse set of educational or performance requirements for all the different positions. An example of this would be the team building events that take place each year. During these, people who do not normally work in the same groups get to know each other, regardless of their position in the company. The atmosphere is always friendly and everyone in the company is invited. From the youngest of law students to the most experienced lawyers together with everyone in executive functions. [34,35].

2.3.4 Systems

Multiple systems are used to ensure the running of the company. Reason being the fact that in the wide range of services the company offers to its clients, lots of very useful licensed but also freely available products exist, as well as many public registries and portals [34,35].

- Formic – The company uses a main IS for a large number of internal processes. It was developed by an external company which developed this tailor made system in exact accordance to the company's needs and requests. An advantage of this for the company is the unlimited additional development that was guaranteed by the system developer. On the other hand, the developer has gained an industry proven product that has given them notoriety and allowed them to expand their services into several European countries [35].

The company primarily uses this IS, however the company's work in certain areas of business also utilizes other systems. These mainly include systems that maintain registries of insolvency records, court registries or registries of foreclosures. Additional systems also maintain automated filling in of court documents and their delivery [35].

2.3.5 Shared values

The company promotes a professional and thoroughly thought-out communication with its clients among its employees, which has been evaluated positively by clients on multiple occasions. The company also strives for constant progress and innovation through the monitoring of international trends and the most recent know-how and technologies. As well as this, the company actively engages in socially beneficial projects and promotions [34].

2.3.6 Style

Projects are worked on in small, specialized teams, while the usual office has 2-4 employees in it. This means high level of training within these teams in their respective areas. Members of these teams are usually rewarded when gaining new professional or soft skills. Every year, almost a quarter of employees sees some form of career advancement and many highly positioned employees rose through the ranks from the position of an ordinary student or apprentice [34].

Within each group there are regular meetings that discuss the goals and their fulfillment, current situation within the company and also opportunities and various threats to operation [35].

Each employee has their authority taken into account whenever an issue or an opportunity arises that requires their input and overall, after speaking to several employees, it can be ascertained that the company's professional approach meets the expectation of even the most highly qualified individuals [34,35].

2.3.7 Skills

Since the company caters to many different clients, each with their own specific needs and a different area of business, the work can be very diverse and can be hard to grasp. Despite this, the company's leadership is of the opinion that it is in the company's interest to meet even the difficult demands of each client and come to a satisfactory solution [34].

Most of the higher positioned employees fall into the category of longtime experts in the law profession and pride themselves in their highly competent approach. At the top

of each organizational structure, these employees possess some form of executive authority. With the company’s long-term approach to the promotion of its employees with higher-than-average performance results, it eliminates the risk of assigning an inexperienced employee to a role they have insufficient skills for [34].

2.4 PESTLE analysis

PESTLE analysis is a framework used to analyze the external factors that can affect an organization. PESTLE stands for Political, Economic, Sociocultural, Technological, Legal, and Environmental factors. The objective of PESTLE analysis is to identify and understand the external factors that can impact an organization's performance and strategy, so that the organization can develop effective plans to manage these factors.

1. Political: The impact of government policies and regulations on the organization, such as tax laws, trade policies, and labor laws.
2. Economic: The impact of economic conditions, such as GDP growth, inflation, and interest rates, on the organization's sales, costs, and profitability.
3. Sociocultural: The impact of demographic and cultural factors, such as population growth, age distribution, and cultural attitudes, on the organization's target market and customers.
4. Technological: The impact of technological advancements, such as automation, artificial intelligence, and the internet, on the organization's products, services, and operations.
5. Legal: The impact of laws and regulations, such as health and safety laws, data protection laws, and environmental laws, on the organization's operations and compliance.
6. Environmental: The impact of environmental factors, such as climate change, natural disasters, and resource depletion, on the organization's operations and sustainability.

The PESTLE analysis is usually a preliminary step to other strategic analysis and planning, such as SWOT or BCG matrix and it can be used to support decision making and identify opportunities or threats [23].

A PESTLE analysis for the company in question might look something like this:

Political

The company needs to consider the impact of government policies and regulations on its business. The company needs to be aware of any political changes that could impact its clients' businesses, such as trade policies and taxes which in turn means the company should consider both the laws that apply to debtors, as well as the laws that apply to the businesses they are in debt to [34].

Economic

The company needs to consider the impact of economic conditions on its business, such as changes in GDP growth, inflation, and interest rates, as well as changes in the demand for its legal services. Especially because of the current geopolitical circumstances, the company needs to closely pay attention to changes in the cost of doing business, such as changes in the cost of labor, rent, and materials. With the ongoing recession and uncertain projected energy prices, these factors have become an even more important part of the company ‘s operating costs then in the years before [35].

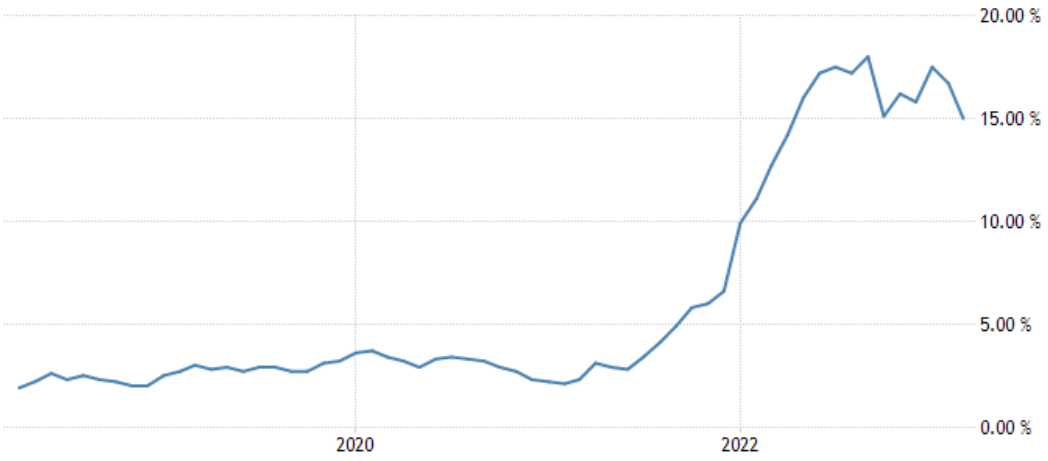

Chart 2 : 5 year chart of the rate of inflation for the Czech republic

(Source: 36)

The Czech Republic itself has seen a rise of the inflation rate, especially from the beginning of 2022 onwards reaching a record high and remains on a similar level to this day. Even though this is expected to change by 2024, these expectations are predicated on the geopolitical situation not deteriorating even further [36].

Sociocultural

The company needs to actively evaluate the impact of demographic and cultural factors on its business, such as changes in the population and age distribution, and shifts in cultural attitudes towards legal and debt collection services as a whole. Effects of these can easily make their way into the political sphere where pressure from citizens demanding or needing change in a certain part of for example insolvency legislation can directly affect the company's ability to operate [23,34].

Technological

With the constant advance in the development of data management technologies, to keep the company ahead of its competition at all times, the company must pay attention to these. API exchange technologies are becoming the norm in the business as a way to exchange large datasets with clients, and with artificial intelligence showing itself as a useful tool when manipulating data that cannot be easily transformed into a useable form, as well as the impact of the internet on the way debt collection services are delivered, the company needs to constantly push for the widescale adoption of these throughout its processes to get any competitive advantage it can [35].

```
Create Procedure [dbo].[ChatGPT] (@prompt varchar(8000),@temperature float)
AS
DECLARE @URL NVARCHAR(MAX) = 'https://api.openai.com/v1/chat/completions';
DECLARE @Object AS INT;
DECLARE @ResponseText AS VARCHAR(8000);
DECLARE @Body AS VARCHAR(8000) =
'{
"model": "gpt-3.5-turbo",
"messages": [{"role": "user",
"content": "'+@prompt+'"}],
"temperature": '+cast(cast(@temperature as decimal(1,1)) as varchar(3))+'
}'
EXEC sp_OACreate 'MSXML2.XMLHTTP', @Object OUT;
EXEC sp_OAMethod @Object, 'open', NULL, 'post',@URL,'false'
EXEC sp_OAMethod @Object, 'setRequestHeader', null, 'Content-Type', 'application/json'
EXEC sp_OAMethod @Object, 'setRequestHeader', null, 'Authorization', 'Bearer AUTH_TOKEN'
EXEC sp_OAMethod @Object, 'send', null, @body
EXEC sp_OAMethod @Object, 'responseText', @ResponseText OUTPUT
select @ResponseText
```

Image 3: Procedure that creates calls to the API of OpenAI's ChatGPT

(Source: Own solution)

Legal

The company has to consider the impact of laws and regulations on its business, such as changes in laws related to legal services, data protection, and intellectual property, regulation of the debt collection business by way of regulations concerning insolvency as well as different forms of salary adjusting legislation. Additionally, the firm would need to be aware of changes in laws and regulations related to its clients' businesses, such as changes in environmental laws and labor laws. regulation of the debt collection business by way of regulations concerning insolvency as well as different forms of salary adjusting legislation that affect the debtor's ability to pay [34,35].

Environmental

The company and its client's business usually does not get significantly affected by environmental factors, however there may be situations when this is not the case. For example if some environmental regulations affect the income status for some debtors, this can directly hinder the company's ability to collect their dues for the client. If on large enough scale, this could have dire consequences and so the company has to closely pay attention to such changes [34].

2.5 SWOT

„SWOT stands for Strengths, Weaknesses, Opportunities, and Threats, and so a SWOT analysis is a technique for assessing these four aspects of your business. SWOT Analysis is a tool that can help you to analyze what your company does best now, and to devise a successful strategy for the future. SWOT can also uncover areas of the business that are holding you back, or that your competitors could exploit if you don't protect yourself."[27]

Based on previous analysis and knowledge gained from working within the company for a considerable amount of time, below are some points identified within the structure of a SWOT analysis, precisely the company's strengths, weaknesses, opportunities and last but not least threats. These are visualized in the following image.

	Useful For reaching goals	Harmful For reaching goals
Internal environmental attributes	Strengths: •The foresight to implement operational change to adhere to legislation •The ability to integrate predictions for economic developments into its operational strategy •Implementation of strategies while paying attention to demographics they affect	Weaknesses: •The direct link between the company's profitability and the solvency of its client's debtors •Difficulty in retaining a larger percentage of students, apprentices and part-time employees beyond their studies •Difficulty of smoothly performing sweeping changes throughout its body of work
External environmental attributes	Opportunities: •The change of public conciousness concerning debt collection •Implementation of advanced technologies to help with decision making	Threats: •Unexpected personnel changes – replacement of certain key employees •Unexpected large scale changes in legislation concerning the company's area of business

Image 4: SWOT analysis of the company
(Source: Own solution based on 27)

Strengths

Among the strong suits of the company is the ability to partially implement or test in advance solutions concerning new legislation. In the recent years because of the geopolitical situation and a pandemic outbreak, the legislation affecting the company's business such as insolvency laws, has been altered several times. The company was forced to alter its operations to accommodate these changes, and has gained the knowledge necessary on how to do the same the next time such changes are required. The same goes for economic developments and their effect on the operations of the company and its clients, which required the company to rapidly implement several measures to lessen their impact [34].

One of the company's strengths is also the adjustment of its strategies when it comes to client's that serve different kinds of demographics. This for example takes the form of implementing modern solutions when it comes to the younger generation [35].

Weaknesses

One of the company's weaknesses or rather the inherent weakness of its business model is its exposure to volatility within the economy of the country it operates in. Effect of

this is generally weakened by the fact that the company has amassed a large number of diversified portfolios over the years, which reduces the effect of this volatility by distributing its business across multiple portfolios that benefit or lose from downward or upward momentum within the economy [34].

Another weakness can be the difficulty with which the company retains employees that do not work in it full-time. To not insult its long term employees and limit the salaries within a reasonable range, the company utilizes a salary progression system that uses more or less set percentage ranges of regular increase that reward the employees more the longer they are at the company. This may be a fair way to evaluate longer term employees but it does not particularly motivate short-term employees to remain at the company [35].

The company also struggles to implement large scale changes to its processes, due to how its organizational structure is set up. Since employees are divided into specialized teams, changes that affect the work of multiple teams at once are difficult for them to agree upon [34].

Opportunities

Although the company conducts only B2B operations, it may find it beneficial to expand the public's consciousness concerning debt collection. The topic is widely viewed as taboo and therefore does not garner a lot of sympathy. This however is not really justified. From being involved in its processes, it can be surmised that the operation of the company and its client's is much more ethical then its usually given credit for being. If this were put across well, the company could potentially benefit from public attention and expand its sphere of business [35].

Another opportunity lies in further advancement of the company used technologies. Implementation of new technologies into the company's decision making processes can greatly optimize its internal processes and increase its profit margins [34].

Threats

One of the threats to the company is its lack of ability to supplant an employee in case of an unexpected departure from the company. Several processes rely heavily on a

couple of people and are therefore susceptible to personnel changes. This is especially true within the IT department [35].

An additional threat is the populist nature of the country's politics, which can see the introduction of legislation that has unforeseen effects on the company's business. Since political candidates are incentivized to use bold legislation proposals as a way to gain an audience, uninformed changes can directly affect the company's work [34].

3 Proposals and Contribution of Suggested Solutions

The following chapter evaluates the information gained from the previous one, requirements of change within the data exchange process to accommodate data exchange with a new client and the choice of a solution for this change.

Knowledge gained from analysis

Summarizing the knowledge we have gained from previous analysis, it can be ascertained that the company should continue with automation of its internal processes. As mentioned in the opportunities from SWOT analysis and technological factors from PESTLE analysis, the company operates in a business sphere that is highly competitive when it comes to utilizing new technologies. The company should further investigate the option of introducing technologies such as AI to always stay ahead of the curve when it comes to progress. The company should also invest effort to make sure that integral parts of its IS are as efficient as possible, since the potentially threatened stability of this systems as mentioned in threats of the SWOT analysis, by extension threaten the company's ability to operate since the IS is such an integral part of its infrastructure. This project was chosen to tackle the automation and efficiency points of the analysis, in addition to how it reduces the effect of the weakness identified in the SWOT analysis as an inability to supplant absent employees by reducing the number of steps they are required to do during the process of exchanging data.

3.1 Initial findings

The purpose of this initial study is to introduce the problematics and to put forward solutions for specific obstacles in the process of creating a reliable data exchange solution. The introduction will shed light on the broader requirements for this project which will be described in more detail in the following parts. Next come the specifications, which describe why the project was created and what it is supposed to achieve. Further along will be a description of the data this solution will work with, and of the structures created for its manipulation throughout the iterative building process. All parts of the ETL (Extract, transform, load) cycle will be included, described in detail and shown on the specific solution. At last, several variants for

different use cases will be shown, that use the knowledge obtained from this development for further automation in the environment of this company

3.1.1 Introduction of the task

After a new client was obtained by the company, the initial expectations for the data exchange were specified, including both handing over of materials needed for enforcement and also the analytical data that would be sent back to the client for evaluation and invoicing. We were given detailed documentation regarding the varied types of data we would receive, including every data format, size and the overall structure of the files. We were assigned to create new structures that would contain the contents of several specified types of files imported into our database and server storage.

These files were specified as two types. First one being compressed archive files containing debt item documentation. These contain individual folders of documents in all sorts of formats – mostly .pdf and .docx, rarely also .xls or .csv. The second type is a delimited .txt file with a varied number of columns, depending whether it contains information about payments, new or updated items. When asking about later automation of this exchange, we were given the information that because of the way the data storage was set up on the client's end, file transfer using SFTP, FTPS or SCP that were already being used for similar applications were not an option.

3.1.2 Analysis of the currently employed process

With the aim of this thesis, we found it important to analyze specifically the information management abilities of the company and its system overall. For this purpose, a separate SWOT analysis was done to determine in which directions the company could improve its standing in this area.

	Useful For reaching goals	Harmful For reaching goals
Internal environmental attributes	Strenghts: •Data available from public registries •Stable, long term development •Reporting services •Easily accessible UI of systems	Weaknesses: •Time consuming loading and updating of data •Inadequate network speeds •Working in a production database •Inadequate ability to supplant an employee
External environmental attributes	Opportunities: •Business intelligence •Cloud services •Efficiency of processes •Automated payment loading	Threats: •Threat to security and data integrity caused by development •Difficulty of the development for the IS •Low efficiency of the IS

Image 5: SWOT analysis of the company's IT solutions
(Source: Own solution based on 27)

Strengths

Among the strong suits of the company are processes that gain useful data from publicly accessible registries. One of such registries is ISIR – Information system of the insolvency registry. It contains data concerning insolvencies of debtors from 2008 onwards that haven't been removed in accordance to the §425 of the Insolvency law in the Czech Rebublic. An automated process parses XML data from the registry and updates relevant information for each of the debtors in the company's system on a daily basis.[28,35]

Another strength is the long-term development of systems not only guaranteed by the IS developer but also conducted by internal employees. These offer flexibility that allows the company to adapt even to urgent changes in the client's needs. The creation and regular use of detailed reports for the bank, insurance company and even the telecommunications sector can also be considered a strength [35].

Finally, a relatively simple UI of the IS system makes for easy use by the employees, reducing the likelihood of faults showing up in most usual operations within it [35].

Weaknesses

Processes that regularly update data by loading it from client's external systems, could be considered a weakness due to how often they can cause issues. Most often they are parsed from the XML format. With the ever growing amount of data that is in need of

updating, it is necessary for these processes to be evolved into a more efficient and overall more practical solution. Another possible weak point is the low network speed currently available in 2 of the 3 company's branches due to older infrastructure. This can cause issues if rapid transfer of data is needed for an urgent task. However, this should be alleviated by the moving of these 2 branches to new locations that is going forward this year. Another weakness can be attributed to the lack of a testing environment to ensure development outside of the production database. Also, there is a lack of ability to supplant certain employees when they are unexpectedly absent [35].

Opportunities

The company should consider a serious implementation of Business Intelligence as a support mechanism for decision making. There is a lot of data available in the company from all the different portfolios it has collected debts from. With the detailed records of actions taken when in contact with each specific debtor, along with payment records and other financial operations, there is an option to utilize this data in future decisions and use it for an even more efficient debt collection process [35].

Another possible opportunity lies in the use of a cloud based data storage system. This approach is becoming more and more popular and widely used. With the growing interest of companies, there is a lot of competition offering safe, efficient and fast cloud services for a reasonable price [35].

Further automation and improvement of existing processes is another way the company can advance. With always growing numbers of various clients, it is critical to remove the need for human interference in as many processes as possible in order to reduce the operating costs and make business viable. This is why this thesis is based on data exchange automation [34].

Threats

Due to constant development of software and systems within the company, there is a heightened risk of endangering security and integrity of data. Several measures are set

up to mitigate this. Even then the risk is present when trying to adhere to client's safety and performance requirement, which can be difficult to reach within the IS [34,35].

The difficulty of the IS development is directly linked to this. With the mentioned considerations, the IS can get to a point where new features, modules or reports are being added and the IS then becomes too large and at the same time inefficient and hard to navigate [35].

Description of the current process

This project will ultimately serve as a replacement for the process currently used to import this client's data into the company IS's database. This process that lacks automation is described by this EPC diagram:

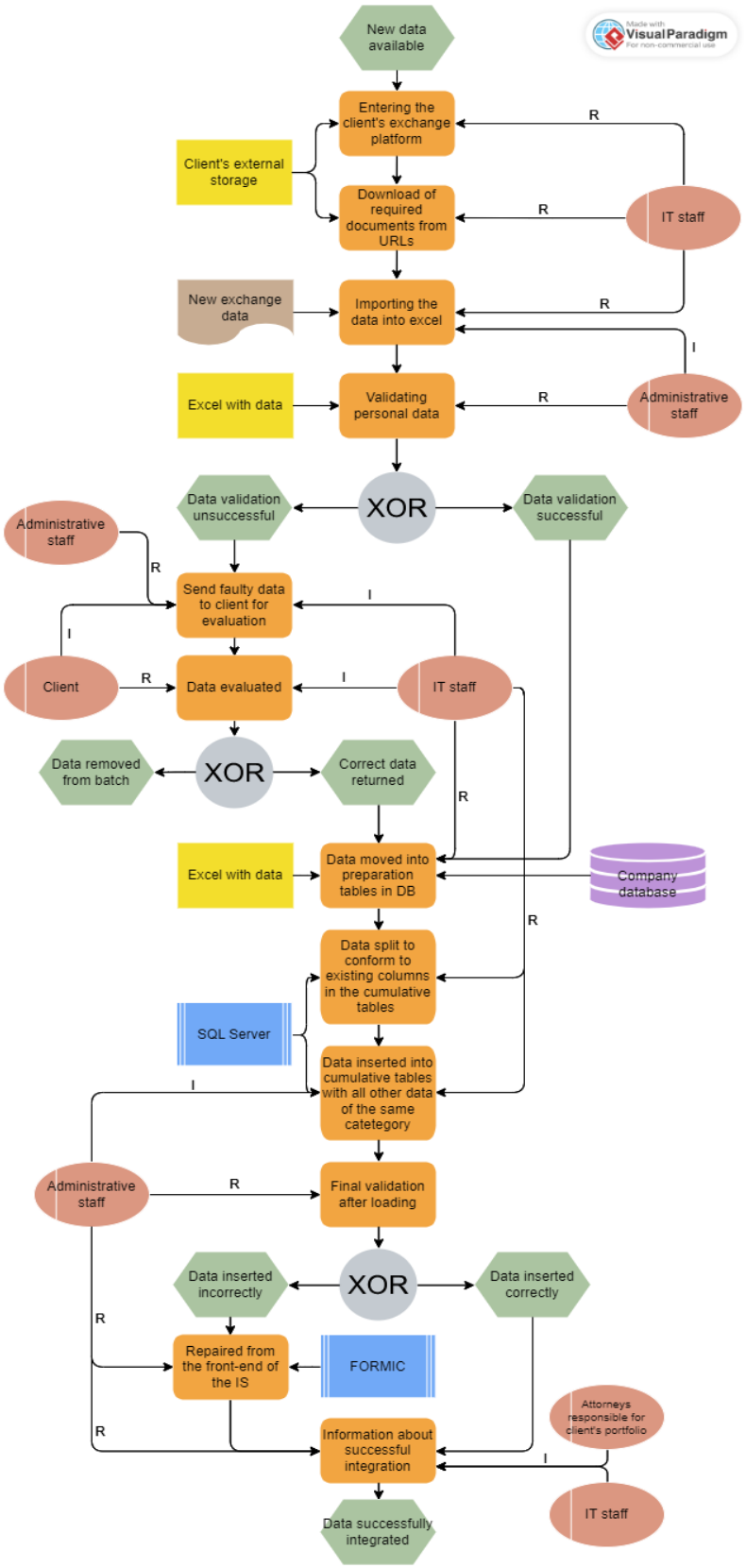

Image 6 : EPC diagram of the current process

(Source: Own solution)

As can be seen in the diagram above, the current process makes employees responsible for 9, in some cases up to 11 steps within the process of loading client's data. This creates many possible points where human error can cause issues.

This is only exaggerated by the fact that the data sent by the client is growing in size. Since every time we expand the number of items in a client's portfolio, the amount of additional data like payments or updates of status grows as well, the current process has a large possibility of being unmanageable in the future. The chart below shows the time required for managing the exchange with the current process up till now with a forecast of the increase :

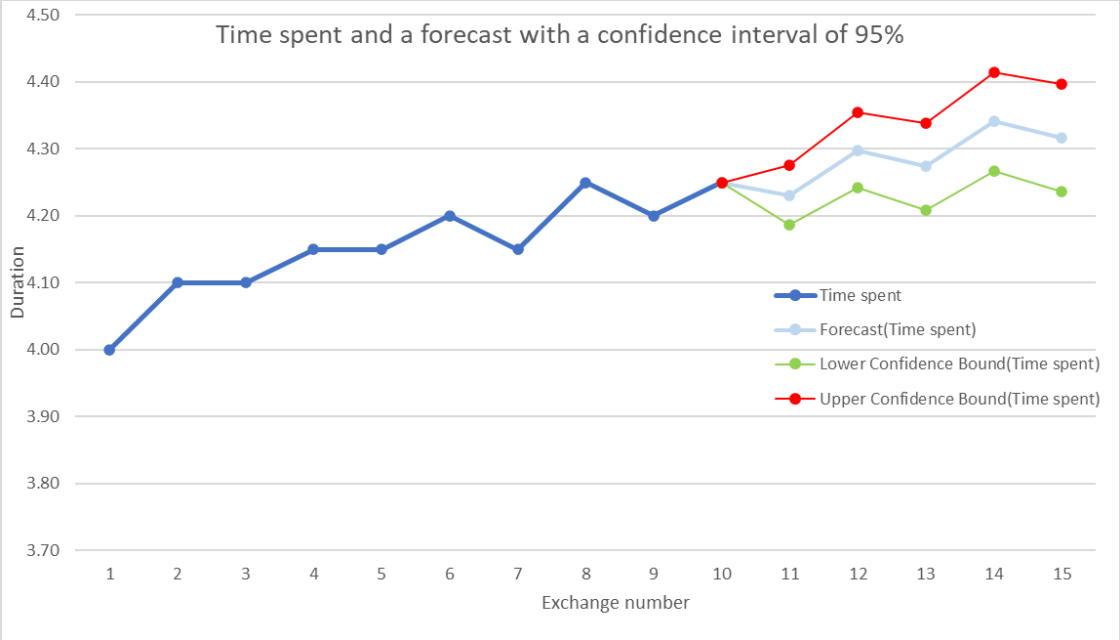

Chart 3 : Time required for an exchange and its projected increase over time

(Source: Own solution based on gathered data)

With data from the current process, we can create a forecast for the next 5 exchanges. This forecast shows that even an optimistic estimation shows the time required rising over time. It is therefore important to offload as many of these steps as possible to automated sub-processes that require minimal human input and cut the time spent to less than 15% of the currently used process. This will be examined further in the chapter dealing with expenses.

3.1.3 Goals of the proposed solution

Goals of this project were to create a simple, safe and dependable way of downloading, transforming and ultimately loading data into our database. This data would later be used in both court proceedings and out of court settlements and so the correctness of this data is essential to the general operation of the company. This process had to be fully automated and with built in redundancies to minimize the impact of human interference on our end. It also had to be easily expandable for other file types that would be added into the exchange further along.

The basic principal that this project follows is the acceleration of repetitive tasks that would normally require a person to manipulate the data and then painstakingly load it themselves, and to remove the human factor from as many sensitive areas of data formatting as possible, as human interference is the most prevalent source of faulty data in data sciences [2].

This work focuses on the process of building this type of a system using advanced data manipulation functionality of T-SQL and describes some options for automated web-scraping in Python and alternatively VBScript.

3.1.4 Environment analysis

There are several structures that are important to mention. First is the location of the data storage. The client uses an online folder in the form of a table that contains information about the files that are being uploaded there. This folder is rid of older files regularly and only holds around a month of latest data at the most. Every file in the folder has its own date of upload and size. While the compressed archives download, .txt files open in the browser on a new window.

File_name	Size(bytes)	Date_altered
payments_45258.txt	7 854	10.1.2022
payments_45256.txt	242	7.1.2022
payments_45240.txt	168	6.1.2022
payments_45210.txt	6832	5.1.2022
Items_A04012022.txt	85604	5.1.2022
DOC_01_2022.zip	148 654 425	5.1.2022

Image 7 : Visual representation of the client‘s data storage

(Source: Own solution)

Second is the database itself. The database, because of its use case, is required to be a live one. The risk that this entails is however mitigated by the multitude of regular backup processes and any mishaps that had happened in the past were resolved quickly and without excessive damages. Because most of the clients are Czech in origin, the database is set up with CZECH_CI_AS collation and is several hundreds of gigabytes in size. There are also secondary databases that are used mostly for analytical processes.

Data from the database is then directly connected to the company's purpose built ERP system. Here, each item sent by clients has its own set of connected information. This includes debt items, payments, debtors, other connected individuals, their addresses, contact information, contact status, item status – whether it's in enforcement, insolvency, restructuring, or whether it's archived, in court proceedings or being paid off. All of these areas is where the information from this data exchange will get to.

3.1.5 Functionality

- Download of the latest data onto server storage
- Bulk insert into database temporary tables
- Transformation of data in a cursor
- Loading of useful columns through loading procedures into tables used on the front end
- Loading into log tables for reporting and further analysis

3.1.6 Automated execution

The automation of these steps is divided into two segments. first is the download done by running a python script from the server. This happens twice a day for redundancy. Always 10 minutes before the second part, which is transformation and import of the data into the database. Since we receive data every day, an automatic email alerts are set for each of these, in case of their failure. Information about the loaded data is then relayed to lawyers that are assigned to this specific client via an SSRS report.

3.2 Proposed solution including technical aspects

This section describes the specifics of the tools used to create the sub-processes necessary to achieve the goals mentioned in chapter 3.1.3.

3.2.1 Used programming languages and libraries

For initial download using python, utilization of Requests, Beautifulsoup - bs4 and Datetime libraries was required. For dealing with archive folders, patool library and os module were used.

- Requests: Requests allows you to send HTTP/1.1 requests extremely easily. There's no need to manually add query strings to your URLs, or to form-encode your PUT & POST data.[19]
- BeautifulSoup: Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.[20]
- Datetime: The datetime module supplies classes for manipulating dates and times.[21]
- Patool: Various archive formats can be created, extracted, tested, listed, searched, compared and repacked by patool. The advantage of patool is its simplicity in handling archive files without having to remember a myriad of programs and options.[22]
- OS: this module provides a portable way of using operating system dependent functionality.

Commands from these are used to pass login parameters and subsequently filter out URLs of files that have been uploaded to the warehouse in the last day.
A flow chart of this process could then look something like this:

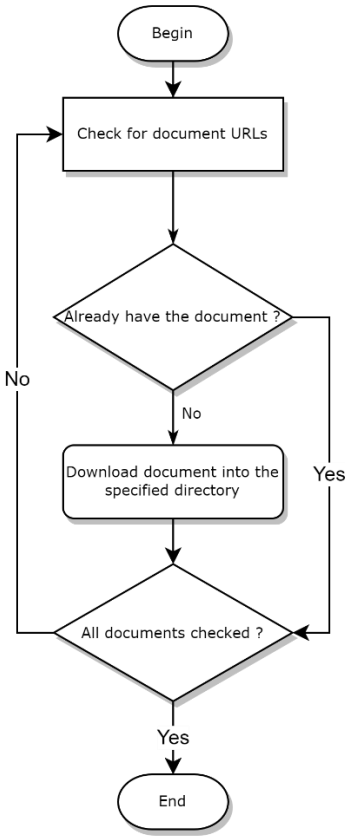

Image 8 : Flow diagram for a crawler that collects document URLs

(Source: Own solution based on 5)

„The nice thing about breaking up tasks into simple functions such as „find all external links on this page“ is that the code can be later be easily refactored to perform a different crawling task.“ [5, p. 46]

Archive folders are also extracted into folders and made ready for file distribution into folders that each of the cases have made for them.

Following the download, the bulk insert utility loads the downloaded .txt files into a database table as delimited text, splitting it into columns. For documents uploaded in archive folders, the xcopy command is run through the xp_cmdshell extended stored procedure in sql – with a slight delay as to not interfere with loading the .txt files.

3.2.2 Main steps taken by the used scripts

- Parsing of datetime information in the HTML structure
- Selection of files that were added within the set interval, in this case the last day
- Establishing a connection to the URL and passing of login parameters
- Parsing of URL file links from the HTML table used to store them
- Making a get request for the download of selected files
- Returning HTTP metadata with the request response, file encoding and content types
- Extraction of the archive folders
- Loading .txt files into a database table
- Attaching transformed data to cases using procedures that split it into tables
- Selecting and moving files from the original folder into their target folders in the NAS.

A flow chart of the new process proposed within the project would look something like this:

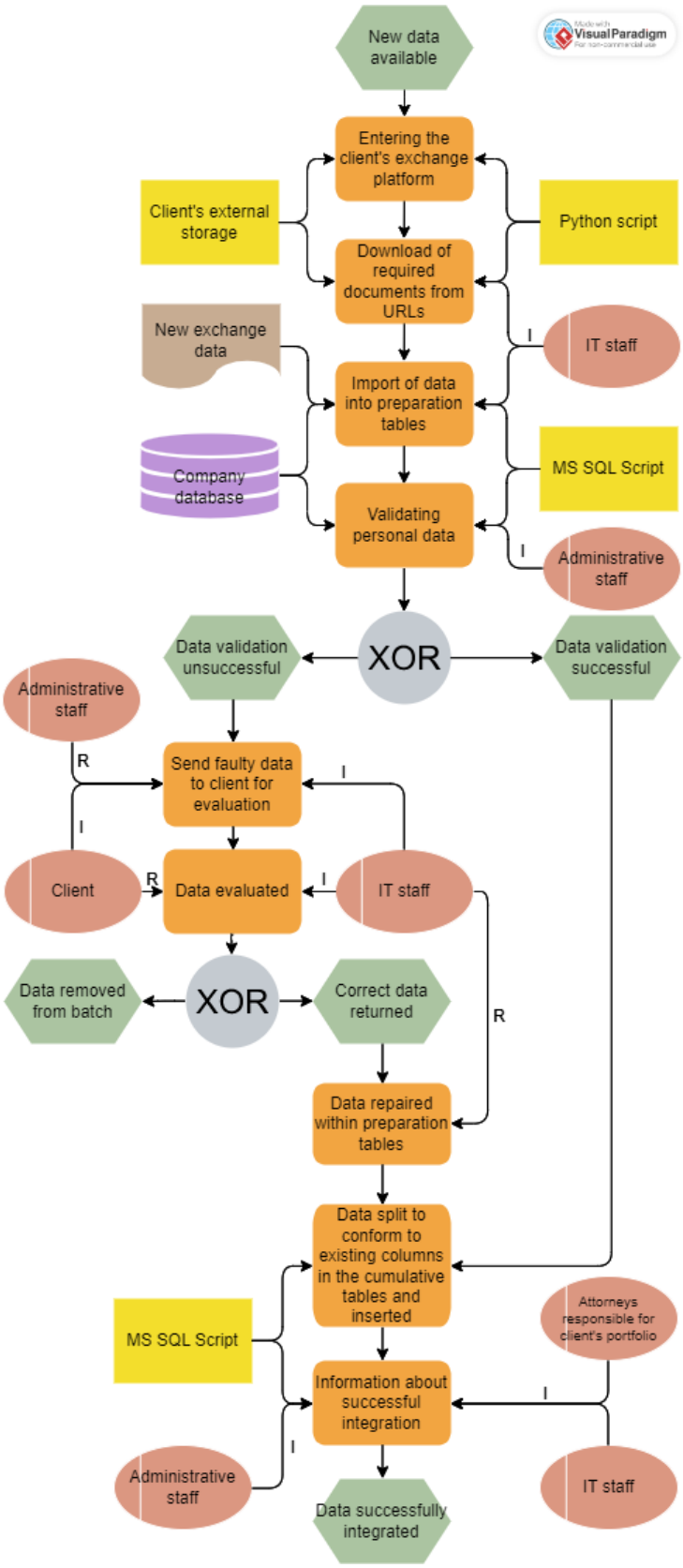

Image 9 : EPC diagram of the proposed new process

(Source: Own solution)

3.2.3 Overview of used queries

Although the list of steps taken to download and insert the data is not extensive, some analysis is necessary to explain the process. Now the used queries will be picked apart step by step to explain their meaning, including the first iteration of the file downloading script made using VBA in excel which although functional, was deemed not ideal for this use.

Possible way of acquiring the data with the use of http requests in MS Excel VBA

The responsible module within the workbook can be set up to open and close by itself. Execution can therefore be fully automated by using the task scheduler. Either on a user machine connected to the server or on the server itself. (This company's server runs Windows Server)

Microsoft WinHTTP Services must be selected from the available references for this script to work.

A request saves the contents of the homepage into the „FileData" variable.

An Open statement writes the saved contents in „FileData" into the „Placeholder.xls" file as seen below :

```
filePath = "\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_PAY\Placeholder.xls"

With auth
.Open "GET", "https://data.XYZ.cz/exp/XXX001122/", False
.SetCredentials "XXX001122", "bpzLaedD:q7:GpJ", _
HTTPREQUEST_SETCREDENTIALS_FOR_SERVER
.SetClientCertificate ("CURRENT_USER\Personal")
sResponse = StrConv(.responseBody, vbUnicode)
End With
FileData = auth.responseBody

Set auth = Nothing
FileNum = FreeFile
Open filePath For Binary Access Write As #FileNum
  Put #FileNum, 1, FileData
Close #FileNum
```

Image 10 : WinHttp request that writes the contents of home page URL into a .xls file

(Source: Own solution)

A loop is set up for an increment of 1 for the variable I. I is first selected as 10 because the structure of the homepage in the client's storage space places the first record on the 10th row of the file.

Within this loop, the script selects the names of .txt files from a column and runs until the „X“ variable becomes empty, or in other words, the loop reaches the last record of the column. This terminates the loop as no more files can be found.

Later a second request writes a new .txt file in the directory since the original file name was chosen together with the format specification. The temporary file is closed automatically.

Overall, this solution works as intended and reliably, however its use would require allowing MS Excel files to run macros on the server, which was quite understandably deemed an unnecessary security risk and because of existing alternatives that would not require additional permissions to be set up. Although it still served as an interesting proof of concept.

The chosen solution for acquiring the data by way of scraping the client's storage using a simple „crawler“

The script initially parses datetime from the selected object supplied as string according to the specified format.
It then returns the string format of current datetime according to the specified format.
After that it checks for the difference between those selected dates, to be exact it checks for the difference being lower than one day and if there are incompatible formats being passed in the used variables.
It then specifies the homepage URL and creates a requests session for creating the connection**.** Then it creates a request for the content on the main URL with login parameters passed in the „auth“ variable.
Then it runs a conversion of the document and HTML entities into Unicode and parses the specified data, in this case using python's onboard parser. [20]
After which it creates loops and subloops for going through all the URLs in the following steps and in between of these finds all rows of the table and passes them into the „link“ variable. Then finds cells in these rows and passes those into „link2“. Then passes all attributes into „link3“ and searches them for filled href attributes and fills the „file“ variable with their contents. The image below shows this process:

```
for link in soup.find_all('tr'):
    for link2 in link.find_all('td'):
        for link3 in link2.find_all('a'):
            if (link3.get('href')!= none):
                file = link3.get('href')
        if (validate(link2.text)):
            file_link_url = "https://data.xyz.cz/EXP/XXX0001122/" + file
            print(file_link_url)
            r = s.get(file_link_url, auth=HttpNtlmAuth('domain\\XXX0001122','bpzLaedD:q7:GpJ') ,allow_redirects = True)
```

Image 11 : Python loops for collecting URLs from the href HTML tags

(Source: Own solution)

As also shown above, it then fills „file_link_url“ variable with the URL of the homepage joined with file name found in the href attribute in the previous step.

In the following step it creates requests for data in all of the URLs.

And if the requests are successful, writes the data into a file in the specified directory. („Sample directory/“ + file)

```
with open(path ,'wb') as f:
    f.write(r.content)
```

Image 12 : Python snippet that writes the content gained from a request into a specified directory

(Source: Own solution)

The two following scripts work with these newly written files in the transformation and loading steps of the ETL cycle.

Handling payment data via a bulk insert statement within T-SQL

Because of the lenght of the procedure that loads the main set of data sent by the client, the script that inserts payment data was chosen instead to show the following steps of this project. Since these are functionally similar and the only difference is the amount of inserted data, it encompasses everything about the parts of the exchange that still need explanation within this thesis [Appendix 3].

The script first clears the supporting tables into which the data is initially loaded. The payment script is different in the fact that the data does not require preliminary validation after insertion. This was agreed upon by the company and the client after the client stated that their data concerning payments is internally validated and does not require further interference. Since the company’s aim is to build trust with its clients, it was determined the data would not indeed be revalidated on the company‘s side [Appendix 3].

Following this the script selects the directory of the folder into which all of the downloaded files were written. It then loads all of the filenames within this directory into a temporary table, using the xp_dirtree extended stored procedure. It then specifies which files are meant for insert and declares a cursor for directly inserting the delimited .txt file contents into a table created with specific columns to accommodate these [Appendix 3].

```
Open C
Fetch next from C into @file
While @@FETCH_STATUS=0
begin
    declare @concat NVARCHAR(100) = @folder+'\'+@file
    declare @script varchar(8000)
set @script='bulk insert YYYYY.dbo.XX_PAY_data_AUTO from '''+@concat+'''
with
(FIRSTROW=1,
FIELDTERMINATOR = ''|'',
ROWTERMINATOR= ''\n'',
CODEPAGE = ''ACP'',
TABLOCK)'
    Exec (@script)

Fetch next from C into @file
end
Close C
Deallocate C
```

Image 13 : Bulk insert statement for loading payment files

(Source: Own solution)

To understand the bulk insert statement settings, here is a sample of a .txt payment file:

```
XXX|00000|28.04.2023|0000000|000000000|788.00|CZK|27.04.2023|XXX|175.00|0
XXX|00000|28.04.2023|0000000|000000000|2600.00|CZK|28.04.2023|XXX|-2216.0
XXX|00000|28.04.2023|0000000|000000000|991.52|CZK|28.04.2023|XXX|166.52|0
XXX|00000|28.04.2023|0000000|000000000|448.00|CZK|28.04.2023|XXX|55.00|00
XXX|00000|28.04.2023|0000000|000000000|890.78|CZK|28.04.2023|XXX|87.78|00
```

Image 14 : Sample payment data

(Source: Own solution)

The next sub step takes care of the data transformation, changing the formats of columns in accordance to previously agreed upon parameters. Some of the data that was being handed over, initially required correction and these adjustments were discussed when dealing with the client's representative.

The cursor that follows then selects information for loading only if the corresponding item exists in our database and the payment is valid and does not already exist based on

the „Client payment identifier“ which the client specified as a unique payment identifier [Appendix 3].

Since the data contains payments not intended for regular loading, a distinction is made based on what account number came attached to the specific payment and is either loaded with a payment insertion procedure – „ChangePayment“ or a procedure for loading secondary information – „ChangeAdditionalInfo“ and therefore does not interfere in the allocation of resources within the debt item that the regular payments are connected to [Appendix 3].

Distribution of documentation after the removal of folder compression

An additional step that does not concern data meant for database storage, but rather documentation sent in compressed .zip files is the distribution of files throughout item documentation directories. Queries in this part are used to download files into separate folders based on an identifier that that their names contain.

First, and intermediate step in python to „unzip“ the received documentation [Appendix 4]:

```
import os, patoolib

dir_name = r"\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_DATA"

os.chdir(dir_name)

for item in os.listdir(dir_name):
        file_name = os.path.abspath(item)
    patoolib.extract_archive(file_name, outdir=dir_name)
        os.remove(file_name)
```

Image 15 : Python script that removes folder compression, or „Unzips“ the folders

(Source: Own solution)

File names are put into a temporary table the same way as in the payment_autoload.sql query. In a table containing the log of this process, the item IDs are kept along with the number of files received that correspond to them that are added after the following cursor [Appendix 5].

A cursor is filled with an identifier, the path of downloaded files in a specific depth and the path of the destination folder from a scalar-valued function dbo.GetDocumentsPathForItem which takes the creation date and identifier of a case to

reconstruct its documentation directory. The following part then creates a list of all the cases that have had documentation sent and counts the files inside their sent folders. A shell command filled with variables then has its output inserted into a temporary table and then these contents are filled into the Copy_Docs table that serves as a log and a source for reporting data [Appendix 5]:

```
set @order = 'xcopy "' + @path_from + '\*" "' + @path_to + '\" /S /Y /D'

Create table #Output (id int Identity (1,1), output nvarchar(255))
insert #Output (output) exec xp_cmdshell @order

Update XXXXX_XXXXX.dbo.Copy_docs
set
[status] = isnull(select Output from #output where output like '%copied%'),'error')
,[to] = @path_to
,[done] = 1
Where item_id = @item_id and [status] is null

drop table #output
```

Image 16 : Shell command for copying documentation

(Source: Own solution)

In case it does not exist yet, an action is attached to the selected cases that signifies that documentation was received from the client and is shown on the front-end of the IS.

3.2.4 Security considerations for the proposed solution

The security of the exchanged data and the proposed solution itself has also been considered. The security of the exchanged data is handled by an IP whitelist set up within the client's storage system. An ntlm authorization has also been set up, as can be seen within the http requests in the scripts mentioned above or in the Appendices. This should limit the potential threats to internal ones, and since the company deals with sensitive data, access to these means of exchange are monitored. The client monitors access to their storage, while the company logs all activity within it's database. Since employees have limited access to these logs and are therefore unable to cover their tracks, this solution is considered sufficiently secure [35].

3.2.5 Possible future expansion of the project

Immediately during the creation of the proposal, several possible expansion of this project were mentioned.

First of these was the removal of the validation step when handling client‘s item personal data. This step is by percentage a considerable part of the remaining requirement for human interference during the exchange process. Elimination of this step would further shorten the execution time of the already shortened time of the proposed solution and cause an even larger reduction in labor cost. This would however require the introduction of an automated validation process. A use of an identity validation portal (identitaobcana.cz) along with a registry of addresses for Czech residents (RÚIAN) was proposed as a solution to this but has not yet undergone the approval process. If it were adopted however, it could become the tool for further optimization of this process.

The other proposed expansion was the use of artificial intelligence to validate and parse through all sorts of data. A ChatGPT stored procedure was created to directly integrate the AI language model into the database and attempt to determine whether it could be used for the company’s purposes. This can be seen in the Technology part of the PESTLE analysis. The idea of the proposal was to use ChatGPT to divide data that was being sent by the client in a form that was difficult to parse. Addresses are of particular interest as they are usually sent as a single line of text in all sorts of configurations. Since the data needs to be divided into individual parts for further processing, these different configurations pose an issue. What ChatGPT allows, is to divide these addresses using its excellent pattern recognition. An example of this would be 2 addresses, one of which has its zip code at the beginning and the second at the end. While a traditional query would need to specify where the zip code can be found, ChatGPT easily recognizes the patterns within these addresses and simply knows which part is the zip code based on it knowing what a zip code looks like.

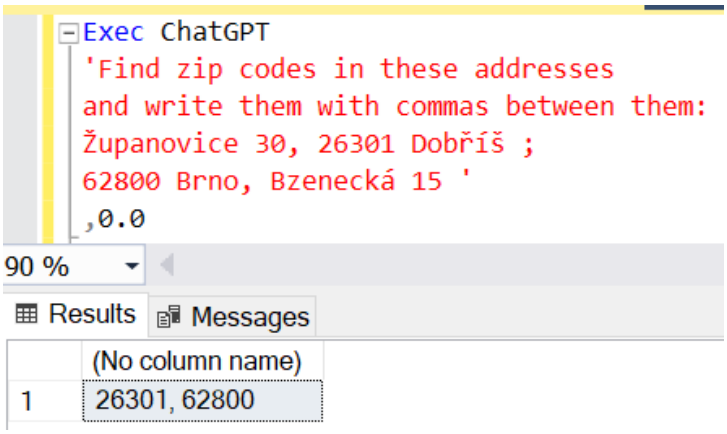

Image 17 : ChatGPT pattern recognition abilities

(Source: Own solution)

Specification of the output format also enables the user to integrate it into exchanges dealing with JSON, XML and other formats. This proposal however is only in the preliminary testing phase and is not yet expected to be integrated into existing data processing applications within the company.

3.3 Economic impact

We will go into more detail on the economic analysis of the proposed improvements to the IS in the section below. In the first part, we examine the expenses associated with the change, and in the second, we list the advantages of the suggested improvements. First it is important to establish a method for analysis.

Since the CBA analysis (Cost Benefits Analysis) has had a considerable influence on how the economic evaluation of past proposed improvements in the IS of the company was structured, it was chosen again to evaluate this proposal. It is quite frequently utilized in the appraisal of potential investment initiatives at the company in question. We can relatively reliably determine the linked economic indicators by utilizing this methodology. When considering whether to invest in a new project, changes in the IS, etc., these might be used as supporting arguments when finding the benefits and drawbacks of various project variants in the process of creating a CBA analysis. In addition, CBA provides options and the optimal strategy for reaching the objective while protecting the parties' investments [29].

It is advised to follow a few steps in the process leading to the preparation of the CBA analysis. The majority of them, including the project goals, alternatives, and interested parties, have already been outlined in the work. The phrase „discount rate“ which shows the amount of interest as a percentage at the end of a given period, will be added in the chapters that follow. A separate concept is the net present value, which measures profit as a percentage of the present value of each share of monetary revenue throughout the time period under consideration. It indicates how much money the specified project will make us in the given time frame, or how much money the project will lose. A recommendation regarding whether or not to begin investing in the project is the outcome of the CBA analysis [29].

3.3.1 Expenses

When developing a new sizeable project for a company IS, it is important to first gather the requirements by all parties and then come up with a proposal for the project. After that, it is required to go through an approval process by the management responsible – this usually means first going before the IT manager that determines whether or not the proposal is executed well enough to go in front of general management. Here the leading attorney of a department together with employees whose portfolios the project affects are present. Since this project concerns a specific portfolio, apart from the department lead only a portfolio officer and supporting attorneys are needed for approval. When the project gets approved, the IT staff develop a solution which usually has a few alternatives. The preferred solution then gets selected and begins testing – first in a mock exchange with prototype data, then in a real world scenario – to check for inconsistencies, errors and bugs that need to be addressed. After the solution is sufficiently reliable, it is introduced to the employees that will come in contact with it. Documentation is created by the IT staff and the solution is introduced into the regular process.

Expenses are then made of labor costs from all personnel involved in the project, Summary of these capital expenditures along with approximate time requirements are in the table below.

Table 1 : Capital expenditures for the proposed project summary

(Source: Own solution)

Phase	Action	N.o.people	N.o.hours spent	hourly wage	Overall expenses
1.	Putting together requirements by the client	4	1	500	2000
1.	Putting together requirements by the company	3	1.5	400	1800
2.	Proposal by IT department staff	2	4	400	3200
2.	Evaluation of the proposal by the IT manager	1	1	600	600
3.	The go-ahead by company representatives	4	0.5	500	1000
4.	Development of the solution, alternatives	2	12	400	9600
4.	Selection, testing of the solution in a mock exchange	2	2	400	1600
5.	Testing of the solution in real conditions, debug	1	5	400	2000
6.	Introduction to employees	6	2	400	4800
7.	Documentation	2	2	400	1600
8.	Introduction into the existing process	1	2	400	800

The capital expenditures reach an estimated 29.000,- CZK. The abbreviated CAPEX or capital expenditures is used to describe a set on one time capital or investment expenses. These expenses take the form of an investment. They make up the budget of a project or a proposal and therefore cause change within the company. [30].

It is currently not necessary for the company to spend additional resources on the project beyond the scope of employee wages. Other supporting systems are sufficient for the introduction of the proposed solution. Expenses for similarly sized projects could in the current state therefore be calculated in a similar fashion.

For a comprehensive summary of the project expenditures, it is important to consider OPEX – operational expenditures. These apply to the continual operation of systems and services that are connected to a regularly running process within the scope of a timeframe of months or years[31].

An OPEX summary of operational expenditures of both the current and proposed states is in the table below.

Table 2 : Operational expenditures for the project summary

(Source: Own solution)

Activity	Required time in hours/year (Current process)	Required time in hours/year (Proposed process)	Expenses (Current process)	Expenses (Proposed process)
Work on the exchange process by IT staff (400CZK/H wage)	120	20	48000	8000
Work on the exchange process by administrative staff (400CZK/H wage)	80	0	36000	0
Lifetime management of the process by IT (400CZK/H wage)	8	8	3200	3200
Summary:	**208 hours**	**28 hours**	**87.200,-**	**11.200,-**

From the table 2 values, it can be surmised that the proposed solution will decrease the time spent by IT and administrative staff in comparison to the current process yearly by more than 85%. Even though the process will still require some management and upkeep throughout it's operational period, it is still expected to cut a significant 76.000,- CZK from yearly operational expenditures.

3.3.2 Benefits of implementation

The following table shows the proposal's effect on expenditures in the first 5 years following implementation.

Table 3 : Yearly evolution in expenditures in case of adoption

(Source: Own solution)

Years	1.	2.	3.	4.	5.
CAPEX	29000	0	0	0	0
OPEX(Current process)	87200	87200	87200	87200	87200
OPEX(Proposed process)	11200	11200	11200	11200	11200
Decrease in expenditures	47200	76000	76000	76000	76000

According to the process of CBA, we can use these calculated values along with the discount rate to determine the Net Present Value (NPV) of the project. NPV can be calculated using this formula:

$$NPV = -C_0 + \frac{C_1}{1+r} + \frac{C_2}{(1+r)^2} + \ldots + \frac{C_T}{(1+r)^T}$$

$-C_0 = Initial\ Investment$
$C = Cash\ Flow$
$r = Discount\ Rate$
$T = Time$

Image 18 : Net present value formula

(Source: 32)

Since the company did not agree to share financial data, the discount rate has been, for our purposes, set to 10%. By filling in the variables in the formula, we get this:

$$NPV = -29000 + \frac{47200}{1+0.1} + \frac{76000}{(1+0.1)^2} + \frac{76000}{(1+0.1)^3} + \frac{76000}{(1+0.1)^4} + \frac{76000}{(1+0.1)^5}$$

$$NPV = 259099.8$$

The Net Present Value for our project in the horizon of 5 years comes out to 259 099,- CZK and this result indicates the investment has all the prerequisites to be profitable within the horizon of the next 5 years.

One of the project's important parts is also the ease with which it can be expanded and utilized in many other similar applications within the company's use. This would further increase the relative profits that this project brings and it is therefore reasonable to advise the company to go ahead with implementation.

Conclusions

The overall importance of the systems explained in this thesis cannot be understated. Their use has changed how information is perceived throughout the world of data. More and more companies are realizing just how vital they can be in reaching the top. This thesis attempts to find a solution to do this for this specific company. This thesis can serve as inspiration to those companies that should also implement such solutions to their own processes to automate where possible, and advance with the rest.

In the introduction, several terms and technologies that are needed for the overall understanding of the subject were outlined. These along with the database systems that were described create a backbone of the processes in the company that the project in this thesis was made for. They all serve as an introduction to the analysis in the parts that follow.

In the analytical part, the company itself was introduced, with its internal structures and systems of management. It was analyzed from both the internal and external points of view and a clear picture was made of the sort of company it is. Then using the SWOT analysis, its positive and negative aspects were brought forth and these serve as a sort of backing for the introduction of the proposed project.

The form of the project was then outlined and processes that assure its function were introduced together with the description of their functions. These make up the practical part of the thesis and are meant to optimize the company's processes with the aim of reducing operating costs. The accomplishment of this is seen in the projected cost reduction analysis and the overall recommendation for the company to go ahead with this project.

References

[1] BASL, J. a R. BLAŽÍČEK. *Podnikové informační systémy - podnik v informační společnosti.* 3. vyd. Praha: Grada Publishing a.s., 2012. 328 s. ISBN 978-80-247-4307-3.

[2] BEGG, C., R. HOLOWCZAK a T. CONOLLY. *Mistrovství - Databáze: Profesionální průvodce tvorbou efektivních databází.* Praha: Computer Press, 2009. 584 s. ISBN 978-80-251-2328-7.

[3] LABERGE, R. *Datové sklady - Agilní metody a business intelligence.* Praha: Computer Press, 2012. 352 s. ISBN 978-80-251-3729-1.

[4] LACKO, L. *Mistrovství v Microsoft SQL Server 2012.* 1 vyd. Praha: Computer Press, 2013. 640 s. ISBN 978-80-251-3773-4.

[5] MITCHELL, R. *Web Scraping with Python: Collecting More Data from the Modern Web.* 2. vyd. Massachusetts: O'Reilly Media, 2018. 300 s. ISBN 978-1491985571.

[6] Charest, G. and Rogers, M. *Data Exchange Methods and Considerations*. For Enterprise Architecture Advisory. 2020. [Online] Available from: https://enterprisearchitecture.harvard.edu/data-exchange-mechanisms#:~:text=Three%20of%20the%20basic%20data,to%20communicate%20and%20exchange%20data.

[7] USC/Information sciences institute, „*TRANSMISSION CONTROL PROTOCOL - DARPA INTERNET PROGRAM - PROTOCOL SPECIFICATION,"* RFC 793, September 1981. [Online] Available from: https://www.ietf.org/rfc/rfc793.txt

[8] Postel, J., "*User Datagram Protocol,"* RFC 768, USC/Information sciences institute, August 1980. [Online] Available from: https://www.ietf.org/rfc/rfc768.txt

[9] R. Fielding (Adobe Systems Incorporated), J. Reschke (greenbytes GmbH), "*Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*," RFC 7231, June 2014. [Online] Available from: https://datatracker.ietf.org/doc/html/rfc7231#section-4.3.3

[10] "*What is an API (application programming interface)?,"* 2023 [Online] Available from: https://www.ibm.com/cloud/learn/api

[11] *What is File Transfer Protocol (FTP)?* [Online] Available from: https://www.raysync.io/news/what-is-file-transfer-protocol

[12] Ylonen, T. *SSH - Secure Login Connections over the Internet*. Proceedings of the 6th USENIX Security Symposium, pp. 37-42, USENIX, 1996. [Online] Available from: https://www.ssh.com/academy/ssh/protocol

[13] Ylonen, T. (SSH Communications Security Corp), Lonvick, C. (Cisco Systems, Inc.) *„The Secure Shell (SSH) Transport Layer Protocol,“* RFC 4253, January 2006. [Online] Available from: https://datatracker.ietf.org/doc/html/rfc4253

[14] Eastlake, D. 3rd (Huawei), Hansen, T. (AT&T Labs) *„US Secure Hash Algorithms - (SHA and SHA-based HMAC and HKDF),“* RFC 6234, May 2011. [Online] Available from: https://datatracker.ietf.org/doc/html/rfc6234

[15] Berners-Lee, T. (CERN) *„Universal Resource Identifiers in WWW“* RFC 1630, June 1994. [Online] Available from: https://datatracker.ietf.org/doc/html/rfc1630

[16] Berners-Lee, T. (MIT/W3C), Conolly, D. *„Hypertext Markup Language - 2.0“* RFC 1866, November 1995. [Online] Available from: https://datatracker.ietf.org/doc/html/rfc1866

[17] „*SSH File Transfer Protocol (SFTP): Get SFTP client & server*“ 2023. [Online] Available from: https://www.ssh.com/academy/ssh/sftp-ssh-file-transfer-protocol

[18] Berners-Lee, T. (MIT/W3C), Fielding, R. (Day Software), Masinter L. (Adobe []Systems) *„Uniform Resource Identifier (URI): Generic Syntax“* RFC 3986, January 2005. [Online] Available from: https://datatracker.ietf.org/doc/html/rfc3986

[19] Reitz, K. „Requests: HTTP for Humans“ February 2011. [Online] Available from: https://requests.readthedocs.io/en/latest/

[20] Richardson, L. *„Beautiful Soup Documentation“* January 2014. [Online] Available from: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[21] Zope foundation and Contributors *„DateTime“* January 2008. [Online] Available from: https://pypi.org/project/DateTime/

[22] Kleineidam, B. *„patool“* March 2010. [Online] Available from: http://wummel.github.io/patool/

[23] Wood, A. *„What is a PESTLE Analysis? A Complete PESTLE Analysis Guide“* December 2022. [Online] Available from: https://onstrategyhq.com/resources/pestle-analysis/

[24] BOELL, Sebastian K. a Dubravka CECEZ-KECMANOVIC. *What is an Information System?* 2015 48th Hawaii International Conference on System Sciences [Online]. IEEE, 2015, 2015, , 4959-4968 [cit. 2021-02-21]. ISBN 978-1-4799-7367-5. Available from : doi:10.1109/HICSS.2015.587

[25] SYMONS, V.J. *Impacts of information systems: four perspectives. Information and Software Technology* [Online]. 1991, 33(3), 181-190 [cit. 2021-02-21]. ISSN 09505849. Available from : doi:10.1016/0950-5849(91)90132-U

[26] „*What is EDI (Electronic Data Interchange)?*“ [Online] Available from : https://www.edibasics.com/what-is-edi/

[27] Mind Tools Content Team *„SWOT analysis”* [Online] [cited 2023-04-24] Available from : https://www.mindtools.com/amtbj63/swot-analysis

[28] „Insolvenční rejstřík“ [Online] [cited 2023-04-24] Available from : https://isir.justice.cz/isir/common/index.do

[29] LANDAU, Peter. *„Cost Benefit Analysis for Projects – A Step-by-Step Guide.“* 09.06.2021 [Online] Available from : https://www.projectmanager.com/blog/cost-benefit-analysis-for-projects-a-step-by-step-guide

[30] CAPEX (Capital Expenditures). ManagementMania [online][cited 2023-04-27]. Available from : https://managementmania.com/sk/capex-capital-expenditures

[31] OPEX (Operational Expenditures). ManagementMania [online][cited 2023-04-27]. Available from : https://managementmania.com/sk/opex-operational-expenditures

[32] „*Discount rate formula: Calculating discount rate [WACC/APV]“* [online][cited 2023-04-27]. Available from : https://www.paddle.com/resources/discount-rate-formula

[33] Shlomo, E. “*Chapter 4 : Data, Information & Knowledge”* [Online][cited 2023-04-28] Available from : https://dataloop.ai/book/data-information-knowledge/

[34] *Interview with the company's CEO* conducted 17.01.2023

[35] *Interviews with leading attorneys and the lead of the IT department* conducted between 19.-20.01.2023

[36] *"Czech Republic Inflation Rate"* [Online] Available from : https://tradingeconomics.com/czech-republic/inflation-cpi#:~:text=Inflation%20rate%20in%20the%20Czech,to%20market%20forecasts%20of%2016.6%25

Used images

Image 18 : Net present value formula (Source: [32])

Used tables

Used charts

Appendices

Appendix 1 : Initially thought up solution using Excel VBA

```
Dim auth As New WinHttp.WinHttpRequest
Dim FileData() As Byte
Dim sResponse As String

'Kill "\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_PAY\*.txt"

filePath = "\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_PAY\Placeholder.xls"

With auth
.Open "GET", "https://data.XYZ.cz/exp/XXX001122/", False
.SetCredentials "XXX001122", "bpzLaedD:q7:GpJ", _
HTTPREQUEST_SETCREDENTIALS_FOR_SERVER
.SetClientCertificate ("CURRENT_USER\Personal")
sResponse = StrConv(.responseBody, vbUnicode)
End With
FileData = auth.responseBody

Set auth = Nothing
FileNum = FreeFile
Open filePath For Binary Access Write As #FileNum
  Put #FileNum, 1, FileData
Close #FileNum

Workbooks.Open "\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_PAY\Placeholder.xls"
Dim X As String
Dim I As Integer
Do
For I = 10 To 10000

X = Workbooks("Placeholder.xls").Sheets("Sheet1").Range("B" & I).Value
  If X = "" Then
  Exit For
  End If
If InStr(X, "XX_pay") = 0 Then
GoTo Konec
End If


filePath = "\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_PAY\" & X
With auth
.Open "GET", "https://data.XYZ.cz/exp/XXX001122/" & X, False
.SetCredentials "XXX001122", "bpzLaedD:q7:GpJ", _
HTTPREQUEST_SETCREDENTIALS_FOR_SERVER
.SetClientCertificate ("CURRENT_USER\Personal")
.send
sResponse = StrConv(.responseBody, vbUnicode)
End With
FileData = auth.responseBody
Set auth = Nothing
FileNum = FreeFile
Open filePath For Binary Access Write As #FileNum
  Put #FileNum, 1, FileData
Close #FileNum

Konec:
Next I
Loop While X <> ""
Workbooks("Placeholder.xls").Close SaveChanges:=False
```

Appendix 2 : Python for scraping the client's storage for document URLs

```python
import requests
from requests_ntlm import HttpNtlmAuth
from bs4 import BeautifulSoup
from datetime import datetime

def validate(date_text):
    try:
        time = datetime.strptime(date_text, '%d. %m. %Y %H:%M:%S')
        now = datetime.now().strftime('%d. %m. %Y %H:%M:%S')
        now = datetime.strptime(now, '%d. %m. %Y %H:%M:%S')

        if ( (now - time).days < 1):
            return True
        else:
            return False
    except ValueError:
        return False

    url = "https://data.xyz.cz/EXP/XXX0001122/"

    with requests.Session() as s:
        page = s.get(url, auth=HttpNtlmAuth('domain\\XXX0001122','bpzLaedD:q7:GpJ'))
        soup = BeautifulSoup(page.content,"html.parser")
    file = ""
    for link in soup.find_all('tr'):
        for link2 in link.find_all('td'):
            for link3 in link2.find_all('a'):
                if (link3.get('href')!= none):
                    file = link3.get('href')
            if (validate(link2.text)):
                file_link_url = "https://data.xyz.cz/EXP/XXX0001122/" + file
                print(file_link_url)
                r = s.get(file_link_url, auth=HttpNtlmAuth('domain\\XXX0001122','bpzLaedD:q7:GpJ') ,allow_redirects = True)
            if(r.status_code == 200):
                path = "files/" + file
                with open(path ,'wb') as f:
                    f.write(r.content)
                    print(r.status_code)
                    print(r.headers['content-type'])
                    print(r.encoding)
```

Appendix 3 : T-SQL script for importing payments

```
delete from YYYYY.dbo.XX_PAY_data
delete from YYYYY.dbo.XX_PAY_data_AUTO

SET @Folder = '\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_PAY'

 Create table #XX_PAY(subdirectory nvarchar(max),depth bit,isfile bit);
 Insert into #XX_PAY
 EXEC master.sys.xp_dirtree @Folder, 0, 1;


Declare C cursor for

select subdirectory from #XX_PAY
where subdirectory like 'XX_pay%' and subdirectory not like 'XX_pay_xy%'

Open C
Fetch next from C into @file
While @@FETCH_STATUS=0
begin
    declare @concat NVARCHAR(100) = @folder+'\'+@file
    declare @script varchar(8000)
set @script='bulk insert YYYYY.dbo.XX_PAY_data_AUTO from '''+@concat+'''
with
(FIRSTROW=1,
FIELDTERMINATOR = ''|'',
ROWTERMINATOR= ''\n'',
CODEPAGE = ''ACP'',
TABLOCK)'
    Exec (@script)

Fetch next from C into @file
end
Close C
Deallocate C
```

```
Insert into YYYYY.dbo.XX_PAY_data
([Company ID], [Load ID], [Load creation date],[Specific Item identifier]
,[Variable symbol],[Exchange sum],[Exchange currency],[Exchange date]
,[Exchange code],[Current remainder],[Payment identifier],[Payment source code]
,[Client account code],[Payment sum],[Payment currency],[Payment allocation date]
,[Payment variable symbol],[Payment specific symbol],[Client account number]
,[Client bank code],[Client account name],[Due date])

select
X.[Company ID],
    [Load ID],
    cast(RIght([Load creation date],4)+
        Right(Left([Load creation date],5),2)+
        LEFT([Load creation date],2) as date),
    [Specific Item identifier],
    [Variable symbol],
    cast(REPLACE([Exchange sum],',','.') as money),
    [Exchange currency],
    cast(RIght([Exchange date],4)+
        Right(Left([Exchange date],5),2)+
        LEFT([Exchange date],2) as date),
    [Exchange code],
    cast(REPLACE([Current remainder],',','.') as money),
    [Payment identifier],
    [Payment source code],
    [Client account code],
    cast(REPLACE([Payment sum],',','.') as money),
    [Payment currency],
    cast(RIght([Payment allocation date],4)+
        Right(Left([Payment allocation date],5),2)+
        LEFT([Payment allocation date],2) as date),
    [Payment variable symbol],
    [Payment specific symbol],
    [Client account number],
    [Client bank code],
    [Client account name],
    cast(RIght([Due date],4)+
        Right(Left([Due date],5),2)+
        LEFT([Due date],2) as date)

    from YYYYY.dbo.XX_PAY_data_AUTO X
```

```
Declare J cursor for

select
it.Item_id,
ez.[Exchange sum],
ez.[Payment allocation date],
Concat('Client payment identifier: '
,ez.[Payment identifier]) as addinfo
, ez.[Client account number]
, ez.[Client account name]
, ez.[Client bank code]

from YYYYY.dbo.XX_PAY_data ez
left join
(select i.VS,i.Item_id from dbo.item i
            join Package p on p.Package_Id = i.Package_id1 and p.PackageGroup_id1=XXX
            )it on it.VS = CONVERT(varchar(100), ez.[Variable symbol])
left join
(select I.VS,p.Comment from payment p join Item I on I.Item_Id=p.Item_id1
            join package pg on pg.Package_Id=I.Package_id1 and pg.PackageGroup_id1=XXX)a1
                on a1.Comment=Concat('Client payment_identifier: ',ez.[Payment identifier])
left join
(select * from ItemAdditionalInfo) IAI
        on IAI.Item_Id1=it.Item_Id
        and IAI.DisplayText='Returned court fees:'+convert(varchar(255),ez.[Exchange sum])
        and IAI.Date=ez.[Payment allocation date]

where it.VS is not null and ez.[Exchange sum]>0 and a1.Comment is null and IAI.Item_Id1 is null
```

```
Open J
fetch next from J into @Item,@value,@added,@addi1,@fromaccount,@fromaccountname,@frombank
while @@FETCH_STATUS = 0 begin
set @addinforeturn= ''
set @datereturn = ''

if @fromaccount != '000000000XXXXXXX' or @fromaccount is null
begin
       execute XXXXX.dbo.ChangePayment
       1,
       'I',
       NULL,
       NULL,
       NULL,
       @Item,
       2,
       @value,
       @added,
       @added,
       null,
       null,
       0,
       1,
       1,
       @addi1
end

if @fromaccount = '000000000XXXXXXX'  and @frombank = 'XXXX'
begin
set @addinforeturn = 'Returned court fees: ' + convert(varchar(255),@value)
set @datereturn = convert(varchar(255), @added,112)

execute XXXXX.dbo.ChangeItemAdditionaInfo

        1,
        'i',
        NULL,
        Null,
        NULL,
        1,
        @Item,
        @datereturn,
        @addinforeturn
end

fetch next from J into @Item,@value,@added,@addi1,@fromaccount,@fromaccountname,@frombank

end
close J
deallocate J
Drop Table #XX_PAY
```

Appendix 4 : Python script to „Unzip“ received documentation

```
import os, patoolib

dir_name = r"\\SERDB00\XXXXX_Storage\XXXXX_TEMP_DOCS\XX_DATA"

os.chdir(dir_name)

for item in os.listdir(dir_name):
        file_name = os.path.abspath(item)
    patoolib.extract_archive(file_name, outdir=dir_name)
        os.remove(file_name)
```

Appendix 5 : T-SQL script used for distributing documentation into item documentation directories

```
declare @order varchar(500)
declare @file varchar(max)
declare @path_to varchar(max)
declare @path_from varchar(max)
declare @path_fromX varchar(max) = '\\SERDB00\XXXXX_storage\XXXXX_TEMP_DOCS\XX_DATA'
declare @item_id bigint
declare @Today datetime = Getdate()
declare @archived bit
declare @input_table table (subdirectory varchar(max), depth int, isfile bit)
declare @count_input table (subdirectory varchar(max), depth int, isfile bit)
declare @number_of_files int
declare @order varchar(8000)

Insert into @input_table
exec xp_dirtree @path_fromX,2,1

Declare C cursor for

select
I.item_id,
@path_fromX + '\' + subdirectory
,XXXXX.dbo.GetDocumentsPathforItem(I.item_id) + '\client_data'

from @input_table
    left join XXXXX.dbo.Item I on subdirectory like '%'+I.ExternalReference+'%'
where isfile = 0 and depth = 2

Open C
Fetch next from C into @item_id,@path_from,@path_to
while @@Fetch_Status = 0 begin
```

```
Insert into XXXXX_XXXXX.dbo.Copy_docs(Item_id)
values (@item_id)

Insert into @count_input
exec xp_dirtree @path_from,6,1

set @number_of_files = (select count(subdirectory) from @count_input where isfile = 1)

Update XXXXX_XXXXX.dbo.Copy_docs
set file_count = @number_of_files
where Item_id = @item_id

set @order = 'xcopy "' + @path_from + '\*" "' + @path_to + '\" /S /Y /D'

Create table #Output (id int Identity (1,1), output nvarchar(255))
insert #Output (output) exec xp_cmdshell @order
```

```
Update XXXXX_XXXXX.dbo.Copy_docs
set
[status] = isnull(select Output from #output where output like '%copied%'),'error')
,[to] = @path_to
,[done] = 1
Where item_id = @item_id and [status] is null

drop table #output

If XXXXX.dbo.GetActionDate(@item_id,257) is null
    begin
    exec XXXXX.dbo.ChangeAction 1,'I',257,null,@item_id,null,null,@today,null,0,null,1,null,0,0
    end

Fetch next from C into @item_id,@path_from,@path_to
end
Close C
Deallocate C
```