Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Engineering

Diploma Thesis

The Effect of Affordable Data Integration Services on Start-up Businesses

Ahmed Ramzi Abdulkarem Muthana

Supervised by

Doc. Ing. Vojtěch Merunka, Ph.D.

© 2023 CZU Prague

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Ahmed Ramzi Abdulkarem Muthana

Systems Engineering and Informatics
Informatics

Thesis title

The Effect of Affordable Data Integration Services on Startup Businesses

---

Objectives of thesis

The aim of this thesis is to investigate how Data Integration services can help small to medium-sized companies (or startups) in identifying, analyzing and studying the performance of their work. The thesis will look into the differences between different types of data integration methods. For example, ETL vs ELT integration methods. The thesis will also explore how pricing for such services could affect how customers and data integration scientists find benefits in data transformation and data loading. The end expected result will be to have a clear understanding of why data is the new trend for marketing teams and how it helps boost small businesses using data pipelines to flow customers' data to their dashboard reports in order to analyze them.

Methodology

This work will be a case study. It will include the Theoretical exploration of Data Architecture, Data Transformation, Data Manipulation and other Data techniques and knowledge that revolves around the use of data. The case study will also look into some of the very few Data Integration companies that are available in the Czech Republic and abroad.

In addition, an interview will be inducted with some customers who use integration solutions, and to what extent it benefits them to have such a tool in their arsenal. Finally, the case study will also dive into how to construct a data pipeline using the API offered by many services, and how to make a business out of such development.

The proposed extent of the thesis

80 – 120 pages

Keywords

ETL, ELT, Data Transformation, Data Pipeline, Data Manipulation

---

Recommended information sources

Fensel, D., Ying Ding, B. Omelayenko, E. Schulten, G. Botquin, M. Brown, and A. Flett. “Product Data Integration in B2B e-Commerce.” IEEE Intelligent Systems 16, no. 4 (2001): 54–59. https://doi.org/10.1109/5254.941358.

Hansen, Mark, Stuart Madnick, and Michael Siegel. “Data Integration Using Web Services.” Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web, 2003, 165–82. https://doi.org/10.1007/3-540-36556-7_15.

Chung, Ping-Tsai, and Sarah H. Chung. “On Data Integration and Data Mining for Developing Business Intelligence.” 2013 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2013. https://doi.org/10.1109/lisat.2013.6578235.

Expected date of thesis defence

2021/22 SS – FEM

The Diploma Thesis Supervisor

doc. Ing. Vojtěch Merunka, Ph.D.

Supervising department

Department of Information Engineering

Electronic approval: 4. 11. 2022

Ing. Martin Pelikán, Ph.D.

Head of department

Electronic approval: 28. 11. 2022

doc. Ing. Tomáš Šubrt, Ph.D.

Dean

Prague on 28. 11. 2022

The Effect of Affordable Data Integration Services on Start-up Businesses

Abstract

Historically, the term Data Integration has become more familiar and famous in the last decade. Data integration is essential in business when it comes to collecting data for market research. The process of combining data from several sources and providing individuals with a single view of it is known as data integration. This technique is crucial in a number of circumstances, including those in the commercial (such as when two businesses with comparable products or services must consolidate their databases).

Businesses attempt to make the raw data they have collected from customers cohesive data while deciding what actions to do next. Data mining is becoming more popular among businesses as a way to gather data and trends from databases, which aids in the creation of fresh company plans that improve operations and speed up economic analysis. To increase their chances of success, business intelligence has modified a type of data integration by compiling the massive amount of data they gather into their system.

With the eruption of large data and the necessity to share existing data, data integration arises more frequently. It has been the subject of significant theoretical research, yet there are still many unresolved issues. The integration of data promotes internal and external user collaboration. In order to offer synchronous data across a network of files for clients, the data being integrated must be obtained from a heterogeneous database system and turned into a single coherent data store for better analysis.

Keywords: ETL, Reverse ETL, Data Pipeline, API, Aggregation, Virtualisation, Transformation, Migration, Integration, Silo, Data Governance, Replication

Vliv Cenově Dostupných Služeb Integrace Dat na Začínající Podniky

Abstrakt

Historicky se termín datová integrace stal známějším a slavnějším v posledním desetiletí. Pokud jde o shromažďování dat pro průzkum trhu, je integrace dat v podnikání klíčová. Integrace dat je proces slučování dat z několika zdrojů a poskytování jednotného obrazu o nich. Tato technika je klíčová za řady okolností, včetně těch komerčních (například když dva podniky se srovnatelnými produkty nebo službami musí konsolidovat své databáze) .

Podniky se pokoušejí ze surových údajů, které shromáždily od zákazníků, udělat ucelené údaje a zároveň se rozhodnout, jaké kroky podniknout dále. Data mining je mezi podniky stále populárnější jako způsob shromažďování dat a trendů z databází, což napomáhá vytváření čerstvých podnikových plánů, které zlepšují činnost a urychlují ekonomické analýzy. Aby zvýšili své šance na úspěch, upravili business intelligence typ integrace dat tím, že shromažďují obrovské množství dat, která shromažďují, do svého systému.

S erupcí velkého množství dat a nutností sdílet existující data vzniká datová integrace stále častěji. Je předmětem významného teoretického výzkumu, přesto stále existuje mnoho nevyřešených otázek. Integrace dat podporuje interní i externí spolupráci uživatelů. Aby bylo možné klientům nabídnout synchronní data napříč sítí souborů, je třeba integrovaná data získat z heterogenního databázového systému a přeměnit je v jedno ucelené datové úložiště pro lepší analýzu.

Klíčová slova: ETL, reverzní ETL, Datový kanál, API, Agregace, Virtualizace, Transformace, Migrace, Integrace, Sila, Správa dat, Replikace

Table of contents

1 Introduction **15**

2 Objectives and Methodology **16**

2.1. Objectives of the Study 16

2.2. Methodology of the Study 16

3 Literature Review **17**

3.1. Business Intelligence 17

3.2. Data Migration 17

3.3. Data Integration with Business Process 18

3.3.1. Defining Data Integration 18

3.3.2. Types of Data Integration Process 19

3.3.3. ELT vs ETL 19

3.3.3.1. ETL 19

3.3.3.2. ELT 21

3.3.3.3. Key Differences 22

3.3.3.4. Current Available Solutions Worldwide 23

3.3.4. Data Integration Challenges, Solutions and Examples 25

3.4. Key Applications of Data Integration 27

3.4.1. Data Warehouse 27

3.4.2. Data Virtualisation 27

3.4.3. Data Replication 28

3.4.3.1. Benefits of Data Replication and its Effects on Business Strategy 28

3.4.3.2. Schemes of Data Replication for Databases 29

3.4.4. Data Streaming 30

3.5. What does Data Transformation mean? 30

3.5.1. Introduction 30

3.5.2. How important is Data Transformation? 31

3.5.3. Data Transformation Benefits 31

3.5.4. Challenges that arise from Data Transformation 32

3.5.5. Methods for Data Transformation 33

3.5.6. Data Transformation Solutions in the Market 34

3.6. How vital Data Integration is to business nowadays? 35

3.7. Data Integration Flows for BI 35

3.8. Data Integration vs Application Integration 37

3.9. Data Silos & Big Data Integration in Enterprises 38

3.9.1. What are Data Silos? 38
3.9.2. What causes Data Silos? 38
3.9.3. Challenges with Data Silos? 38
3.9.4. How bad can Siloed Data be to any enterprise? 39
3.9.5. Big Data Integration? 39
3.9.6. Big Data Integration for Bridging Siloed Data 40
3.10. Data Governance 41
3.10.1. Introduction 41
3.10.1. Challenges of Data Governance 41
3.10.2. Benefits of Data Governance 42
3.10.3. Who should enforce Data Governance? 43
3.10.4. How to develop a framework for Data Governance 43
3.10.5. Steps to model your Data Governance 44
3.10.6. Tools for Data Governance 45
3.11. Data Integration Deployment Models for this Project 46
3.11.1. Data Transformation Tools 46
3.11.2. Personal Backend API 47
3.11.3. Google Cloud Tools 48
3.11.4. Reporting Data over API 48

4 Practical Part 50
4.1. Use Cases 50
4.1.1. No-Code Integration 50
4.1.1.1. Data Virtualization: Data Source to Dashboarding Application 50
4.1.1.2. Data Storage: Data Source to Data Warehouse 57
4.1.1.3. Data Blending: Multiple Sources, One Destination 61
4.1.2. Custom Data Integration (With Coding) 65
4.1.2.1. Data Migration: From Spreadsheet to a better Data Warehouse 65
4.1.2.2. Public Data: API to Data Virtualisation 76
4.1.2.3. Data from Private API to DWH 82

5 Results and Discussion 92

6 Future Work Recommendation 93

Conclusion 93

7 References 94

8 Appendix 98

List of Figures

Figure 1: Fundamentals of Data Integrations 18
Figure 2: Simple diagram of ETL process 20
Figure 3: Detailed ETL Process 21
Figure 4: Detailed ELT Process 22
Figure 5: Full Data Replication 29
Figure 6: Partial Data Replication 30
Figure 7: Fundamentals of Data Transformation 32
Figure 8: Traditional architecture for BI 36
Figure 9: Modern architecture for BI 37
Figure 10: Data Governance Framework 41
Figure 11: Data Governance for all reasons 43
Figure 12: Steps to Data Governance 45
Figure 13: Data Source to Dashboarding Application 51
Figure 14: Use Case 1: Activity Diagram 51
Figure 15: Selecting the YouTube Analytics connector 52
Figure 16: Dataddo's YouTube Analytics Dataset 53
Figure 17: YouTube Analytics Metrics and Attributes 53
Figure 18: Snapshot Automation 54
Figure 19: Data Preview 54
Figure 20: Data Flow Configuration - Google Data Studio 55
Figure 21: Dataddo Config with Google Data Studio 55
Figure 22: Data Studio Report Configuring 56
Figure 23: Google Data Studio report 57
Figure 24: Dataddo Data Warehouse Table Creation 58
Figure 25: YouTube Data Schema in Google BigQuery 58
Figure 26: YouTube SQL query in Google BigQuery 59
Figure 27: Use case 2: Activity Diagram 60
Figure 28: Use case 3: Activity Diagram 61
Figure 29: Blending Two Data Sources 62
Figure 30: Data Blending JOIN key 63
Figure 31: Data Blending JOIN types 63

Figure 32: Select Field for Data Blending 63
Figure 33: Blended Data Source 64
Figure 34: Blended Data Sources to Google BigQuery 64
Figure 35: Blended Data Sources Migrated to BigQuery 65
Figure 36: Use case 4: Activity Diagram 66
Figure 37: Mock Google Sheet 67
Figure 38: JSON Universal Connector 67
Figure 39: Google Sheet RAW Data 70
Figure 40: Deconstructing Google Sheet Data 71
Figure 41: Deconstructing Index from Values 72
Figure 42: Constructing Proper Array of Objects 73
Figure 43: Google Sheet Connector Creation 74
Figure 44: Custom Google Sheets Data Flow 75
Figure 45: Google Sheet Data Migrated to Google BigQuery 76
Figure 46: Use case 5: Activity Diagram 77
Figure 47: Public API Raw Response 78
Figure 48: Public API Data Transformation 79
Figure 49: Public API Data Flow Configuration 79
Figure 50: Public Data Virtualised 80
Figure 51: Public API Data Source Settings 81
Figure 52: Public API Data Source Snapshotting Settings 81
Figure 53: Use case 6: Activity Diagram 82
Figure 54: API Code Base Repository 83
Figure 55: Vercel: API Hosting Server 84
Figure 56: API HTTP Request Home Route 85
Figure 57: API HTTP Response Home Route 85
Figure 58: Personal Firestore Database with Empty Records 86
Figure 59: Creating Single Record in Collection 86
Figure 60: API Error 422 87
Figure 61: Bulk data creation 87
Figure 62: Database populated 88
Figure 63: Personal API request on Dataddo's platform 89

Figure 64: Data Source Creation for Personal API 90
Figure 65: Data flow configuration Personal API to Google BigQuery 91
Figure 66: Personal API Migrated to Google BigQuery 91

List of Tables

Table 1: ETL vs ELT 23
Table 2: Data Integration Platform Type Comparisons 24
Table 3: Data Integration Platform Type Comparisons 24
Table 4: Data Integration Platform Type Comparisons 24
Table 5: Data Integration Platform Type Comparisons 25
Table 6: List of Parameters for Google Sheet API 69

List of abbreviations

DI	Data Integration
DS	Data Source
DF	Data Flow
DM	Data Management
DB	Database
DWH	Data Warehouse
DV	Data Virtualisation
HTTP	HyperText Transport Protocol
URL	Uniform Resource Locator
API	Application Programming Interface
UI	User Interface
SQL	Structured Query Language
JS	JavaScript
OODB	Object Oriented Database
RDMSB	Relational Database Management Systems
JSON	JavaScript Object Notation

1 Introduction

The general purpose of information system integration is to bring together specific systems into a unified, fresh whole, giving users the appearance that they are interacting with only one system. The need for integration stems from two factors: In order to make it simpler to access and reuse information from a single information access point, an integrated view can first be constructed from a collection of already-existing information systems. Second, data from numerous complementary information systems are combined in response to a piece of specific information required to produce a more comprehensive basis for doing so.

Nowadays, it is typical for firms to run several information systems simultaneously. Businesses that employ these techniques miss out on lucrative business opportunities in markets with intense competition. The integration of current information systems is becoming more essential in this circumstance since long-term investments in present IT infrastructure are being utilized while dynamically satisfying business and customer objectives.

I started my IT career two years ago with a company called Dataddo, that is specialised in meeting clients' needs to provide data integration solutions and manage their data and the flow of migrations pipelines. It's also a company that is based in Prague. I have learned in these two years the needs of clients and common and unique use cases on why they want to integrate their data. I have also had first-hand experiences in developing data pipelines and also data transformation. This project illustrates the knowledge I have gained in dealing with clients and handling their projects and developing new technologies to automate these processes.

2 Objectives and Methodology

2.1. Objectives of the Study

- This thesis aims to investigate how Data Integration services can help small to medium-sized companies (or start-ups) in identifying, analyzing and studying the performance of their work.

- The thesis will investigate the differences between different types of data integration methods. For example, ETL vs ELT integration methods.

- The thesis will also explore how pricing for such services could affect how customers and data integration scientists find benefits in data transformation and data loading.

- The end expected result will be to have a clear understanding of why data is the new trend for marketing teams and how it helps boost small businesses using data pipelines to flow customers' data to their dashboard reports to analyse them.

2.2. Methodology of the Study

This work will be a case study. It will include the theoretical exploration of Data Architecture, Data Transformation, Data Manipulation and other Data techniques and knowledge that revolves around the use of data. The case study will also look into some of the very few Data Integration companies that are available in the Czech Republic and abroad.

In addition, an interview will be inducted with some customers who use integration solutions, and to what extent it benefits them to have such a tool in their arsenal. Finally, the case study will also dive into how to construct a data pipeline using the API offered by many services, and how to make a business out of such development.

5 Results and Discussion

In this project, we look at the multiple use cases and trends on how to imply Data Integration into your business process. As we have become a world filled with data and information, it is simple to ignore how pieces of data help your business. The lack of utilising those data can cause an enterprise performance damage and not a clear understanding of its audience. As more services like YouTube Analytics, Facebook Ads, LinkedIn Ads, HubSpot CRM, ADP and many more services emerges to help enterprise clients do a specific task, its good to remember that most of these services support API services and your data can be integrated into your preferred destination to do your business analytics.

You do not need to be confided to you use the charts and graphs that the service your use gives you. For example, let us say you make advertisement campaigns on Facebook Ads, they also give you an analytics virtualisation section in their platform to observe the performance of your campaigns, but they could be very limited in terms of defining scopes for your dataset and reporting. Therefore, you do not have to be limited by their tools, You can extract your data and send it to much better Dashboarding platforms that are specialised in defining scopes and generating reports, Google Data Studio, Power BI …etc.

Not only that, but most services also give you the power to extract historical data from their API. These historical data can be crucial to most enterprises to draw a big picture that can help them in defining the trajectory of their goals and performances, past and also plan for the future.

In this study, we only look into one form of Data Integration, and that is ETL methodology, and from this study, we see in section 4.1.2.1, that ETL comes in very handy when migrating from one Datawarehouse to another. Especially when we want to migrate from SQL to a NoSQL type of database could be tricky and not so easy for those people who do not have much knowledge of the migration process or are afraid to drop some tables in the process of migration.

6 Future Work Recommendation

As Data Integration usage increases daily and a new type of use case emerges. Data Integration specialists should study more the about the term Reverse ETL. Reverse ETL is now being adopted in a few platforms, like Dataddo and Integrate.io, where the data are sent backwards from DWH to Data Sources. The reason for this usage may vary and a detailed study of enterprises' usage will be required as it slowly becomes popular with the crowd.

Conclusion

Working in a firm that builds Data Integration tools for clients and handles their projects made this project topic an interesting study that I have been very keen on learning more about it for the last two years. In my thesis, I elaborate on how DI solutions used to be very costly, 2 or more decades ago, and with the explosion of big data, social media marketing and so on, small businesses started to utilise these free data-oriented platforms, whether for sales, marketing, CRM purposes …etc. With time, affordable integration platforms have emerged and made the process much easier for data scientists to analyse their data.

This thesis was divided into two parts, theoretical and practical parts. The theoretical part investigated multiple terms and concepts of Data Integrations, including key methods that were further elaborated on and demonstrated in the practical part. It also looked at how it is vital to business and how can some DI platforms can be affordable to enterprises, especially start-up and middle-sized companies, depending on their use case

The methodology then focuses on taking those key points and I lay multiple use cases for different client usages. It starts with easy day-to-day integrations, to more complicated ones. However, there are many more use cases in that many clients come to me for consultation, and every client becomes a learning experience when I take on their project.

Different scenarios were taken in the studies to simulate the overall picture for the use case analyses.

7 References

Abraham, Rene, Johannes Schneider, and Jan vom Brocke. (2019). "Data Governance: A Conceptual Framework, Structured Review, and Research Agenda." *International Journal of Information Management* 49:424–38. https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

Chung, Ping-Tsai, and Sarah H. Chung. (2013). "On Data Integration and Data Mining for Developing Business Intelligence." IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2013. https://doi.org/10.1109/lisat.2013.6578235.

Dayal, Umeshwar, Malu Castellanos, Alkis Simitsis, and Kevin Wilkinson. (2009). "Data Integration Flows for Business Intelligence." *Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT '09*, 2009. https://doi.org/10.1145/1516360.1516362.

Ehtisham Zaidi, W. Roy Schulte, Eric Thoo (2019). Adopt Stream Data Integration to Meet Your Real-Time Data Integration and Analytics Requirements. Accessed October 1, 2022 .https://www.gartner.com/en/documents/3904668

Farmer, Donald. (2022). "6 Key Steps to Develop a Data Governance Strategy." SearchDataManagement. TechTarget, May 17, 2022. https://www.techtarget.com/searchdatamanagement/tip/6-key-steps-to-develop-a-data-governance-strategy.

Fensel, D.; Ying Ding,; Omelayenko, B.; Schulten, E.; Botquin, G.; Brown, M.; Flett, A. (2001). Product data integration in B2B e-commerce. IEEE Intelligent Systems, 16(4), 54–59. https://doi.org/10.1109/5254.941358.

Hansen, MarkStuart Madnick; and Michael Siegel. “Data Integration Using Web Services.” Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web, 2003, 165–82. https://doi.org/10.1007/3-540-36556-7_15.

J Sreemathy; K Naveen Durai; E Lakshmi Priya; R Deebika; K Suganthi; PT Aisshwarya; (2021). Data Integration and ETL: A Theoretical Perspective . 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), (), – . doi:10.1109/ICACCS51430.2021.9441997

Lenzerini, Maurizio (2002). [ACM Press the twenty-first ACM SIGMOD-SIGACT-SIGART symposium - Madison, Wisconsin (2002.06.03-2002.06.05)] Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02 - Data integration. , (), 233–.https://doi:10.1145/543613.543644

ManageEngine. “Data Replication in Distribution System.” Data Replication Types - Benefits - Schemes in DBMS - ManageEngine Device Control Plus. Accessed October 16, 2022. https://www.manageengine.com/device-control/data-replication.html.

Mary K. Pratt. (2022). “What Is Data Preparation? an in-Depth Guide to Data Prep.” SearchBusinessAnalytics. TechTarget, February 2, 2022. https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation.

Patel, Jayesh. (2019).“Bridging Data Silos Using Big Data Integration.” *International Journal of Database Management Systems* 11, no. 3:01–06. https://doi.org/10.5121/ijdms.2019.11301.

Reeve, April. (2013). “Managing Data in Motion.” *Data Integration Best Practice Techniques and Technologies*. https://doi.org/10.1016/c2011-0-07758-x.

S, Manikandan, (2010). “Data Transformation.” *Journal of Pharmacology and Pharmacotherapeutics* 1, no. 2 (2010): 126–27. https://doi.org/10.4103/0976-500x.72373.

Salmi, Christina. (2022). “8 Steps to Start a Data Governance Program.” Analytics8, October 10, 2022. https://www.analytics8.com/blog/8-steps-to-start-your-data-governance-program/.

SAP. (n.d.). “What Is Data Governance?: Definition, Importance, & Types: SAP Insights.” SAP. Accessed December 13, 2022. https://www.sap.com/insights/what-is-data-governance.html.

Song, IY. (2009). Data Warehouse. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_882

Stedman, Craig (2022). “What Is Data Governance and Why Does It Matter?” SearchDataManagement. TechTarget, May 31, 2022. https://www.techtarget.com/searchdatamanagement/definition/data-governance.

Stedman, Craig (2019). “What Is Data Management and Why Is It Important?” SearchDataManagement. TechTarget, October 30, 2019. https://www.techtarget.com/searchdatamanagement/definition/data-management.

Stitch. (n.d.). “Understanding Data Replication and Its Impact on Business Strategy.” Stitch. Accessed November 17, 2022. https://www.stitchdata.com/resources/data-replication/.

Stitch. (n.d.). “What Is Data Transformation: Definition, Benefits, and Uses.” Stitch. Accessed November 29, 2022. https://www.stitchdata.com/resources/data-transformation/.

Tatbul, Nesime (2010). Streaming data integration: Challenges and opportunities. [IEEE 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010) - Long Beach, CA, USA (2010.03.1-2010.03.6)] 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010).https://doi:10.1109/icdew.2010.5452751

Ziegler, Patrick, Dittrich, Klaus R. (2007). Data Integration — Problems, Approaches, and Perspectives. In: Krogstie, J., Opdahl, A.L., Brinkkemper, S. (eds) Conceptual Modelling in Information Systems Engineering. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72677-7_3

8 Appendix

Personal API Codes:

https://github.com/ahmed-ramzi/api-master-thesis

Google Data Studio Reports:

~Use case 1:

https://datastudio.google.com/reporting/79fc5c59-5c2d-4bdc-a448-cc7345682da1/page/rtO7C

~Use case 4:

https://datastudio.google.com/reporting/4b27e059-0bbd-4ec7-bd24-780ab80f77a5/page/B3d7C

Documentations :

YouTube API:

https://developers.google.com/youtube/analytics/data_model

Google Sheet API:

https://developers.google.com/sheets/api/reference/rest/v4/spreadsheets/get

Firebase Cloud Store

https://firebase.google.com/docs/firestore