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.

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

Declaration

I declare that I have worked on my master's thesis titled " **The Effect of Affordable Data Integration Services on Start-up Businesses**" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the master's thesis, I declare that the thesis does not break any copyrights.

In Prague on: 29/11/2022

Ahmed Ramzi Abdulkarem Muthana

Acknowledgement

I would like to express my gratitude and acknowledgement to my supervisor, Doc. Ing. Vojtěch Merunka, PhD, who made this work possible. I was able to complete all my project's writing stages because of his or her direction and assistance, for making this thesis a fun learning experience and for their insightful comments and ideas.

Additionally, I want to especially express my gratitude to my wife, Belqis, for always pushing me to give out the best for this thesis, for helping me with the proofreading and for motivating me toward accomplishing my goal.
Secondly, I want to extend my gratitude to my mother and. father, and my entire family for their unwavering support and tolerance as I conducted my research and wrote my project.

Finally, I want to express my gratitude to God for guiding me through all the challenges. Every day, I have felt your guidance. I was able to complete my degree thanks to you.

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

List of Figures

List of Tables

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.

3 Literature Review

3.1. Business Intelligence

Business intelligence (BI), in the words of (**Dayal et al. 2009)** is a technology stack, tools, and techniques for gathering, integrating, purifying, and mining corporate data for decision-making. These include data mining, data warehousing, analytics, reporting, and visualisation. The BI architecture of today was created for strategic decision-making, where a small group of knowledgeable users evaluate historical data to create reports or models, and where decision-making cycles might take weeks or months.

The primary objective of business intelligence is to assist organizations in analysing customer behaviour. It aids in identifying current market trends and identifying concerns or challenges. Using past data, business intelligence aids in making decisions during critical situations. This historical data is viewable but not editable. It provides extensive information about the processes of a business nature, hence increasing the value of those processes and allowing for more accurate conclusions. **(Sreemathy et al. 2021)**

3.2. Data Migration

Data migration is the process of moving data from one system to another. Enterprises migrate their data for a variety of reasons. They may be required to remodel a whole system, upgrade databases, create a new data warehouse, or combine fresh data from an acquisition or other source.

(Sreemathy et al., 2021) say that despite the apparent simplicity of the definition above, this requires a change in storage, database, or application. In addition, data migration is required when establishing a new system that coexists with old applications. Data must be prepared, extracted, and, if necessary, transformed. Supporting the migration of processed data from one storage location to another are numerous other technologies, such as ETL tools. There may be any unmatched data type that gets transferred, and the migrated data may be in number, date, sub-records, and multiple characters set that can be encoded. This is one of the most essential concerns in migration.

3.3. Data Integration with Business Process

B2B e-commerce presents significant constraints in today's age. Therefore, we require intelligent data structure, standardisation, alignment, and personalization automation solutions. To function well, a market must accommodate users on many different devices and operating systems by establishing a standard protocol for data transfer, specifically on various hardware and software technologies.

3.3.1. Defining Data Integration

Hence comes data integration. **(Lenzerini 2002)** describes data integration as the process of merging pieces of information from several sources (usually referred to as "data sources") and presenting a uniform picture of these data to the user. The topic of creating data integration systems is significant in contemporary real-world applications and is characterized by a variety of theoretically intriguing difficulties.

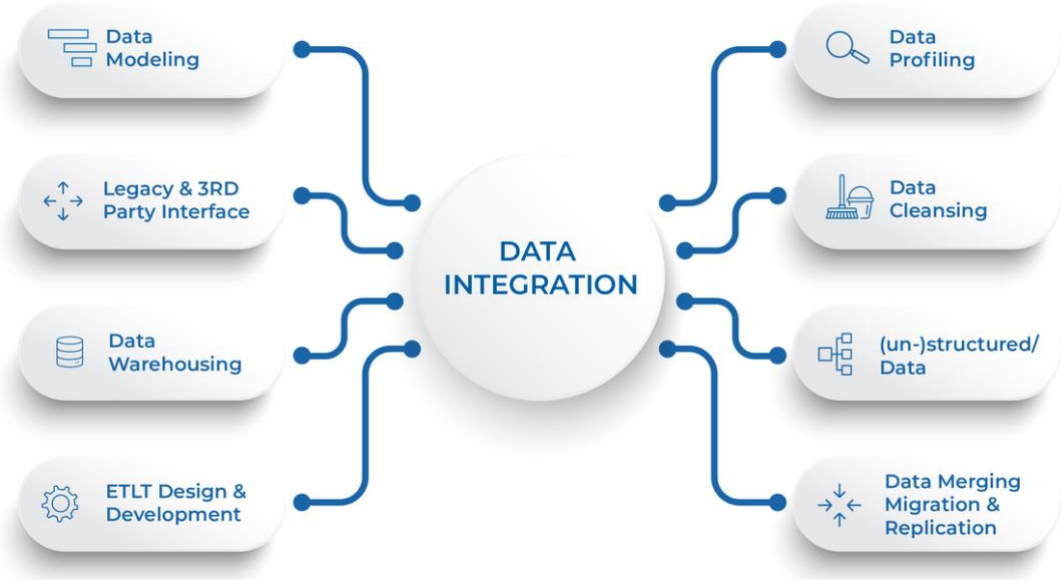


Figure 1: Fundamentals of Data Integrations

According to (**Talend's, n.d.)**, -a company that specialised in Data Integration solution that helps enterprises transform, load, and govern data- B2B e-commerce tends to combine data (for example, marketing and advertising performance data, employees data …etc.) from various data sources into single or multiple destinations, to serve the information needs of all applications and business processes, and to provide users with consistent access to and

delivery of data across a variety of themes and structure types, for example, making monthly virtualised reports of said company's marketing and sales performance. One of the key steps in the total data management process, Data Integration is used more frequently as big data integration and the demand for current data sharing increases.

3.3.2. Types of Data Integration Process

With time, various data integration tools emerged, and many companies started making platforms that are created by data integration architects to make it easier for companies to integrate and route data from source systems to target systems through an automated data integration process. (**Stedman 2019)** highlights multiple methods of data integration that are commonly known in the world that may be used to accomplish this, including:

- **ETL:** is short for **E**xtract, **T**ransform and **L**oad. Simply, It gathers, unifies, and loads datasets from several sources into a data warehouse or data virtualisation tools (for reporting purposes).

- **ELT: this** is the same thing, but the order of the operations is different, and it stands for **E**xtract, **L**oad, then **T**ransform. Similarly, in ETL, the transformation of the extracted data occurs after the data is loaded into databases for analytic and reporting uses.

3.3.3. ELT vs ETL

3.3.3.1. ETL

Data warehouses have traditionally been designed to read and query massive databases quickly for accurate business analytics A data warehouse, however, was a multi-million-dollar project that required purchasing hardware, obtaining software licenses, and designing and maintaining the system. Developers would only load cleansed, converted, and aggregated data into their warehouses in order to reduce costs, and they would delete any data that wasn't required for the analysis to increase efficiency.

ETL is indispensable for data integration. With the use of ETL, businesses are able to collect data from disparate locations and transform it into a cohesive view for future use.

Organizations have to gather data from several databases, harmonise it, and weed out extraneous details before putting it into the warehouse in order to prepare it in this way. ETL (extract, transform, load) tools were created as a result, and they prepare and process data in the following order:

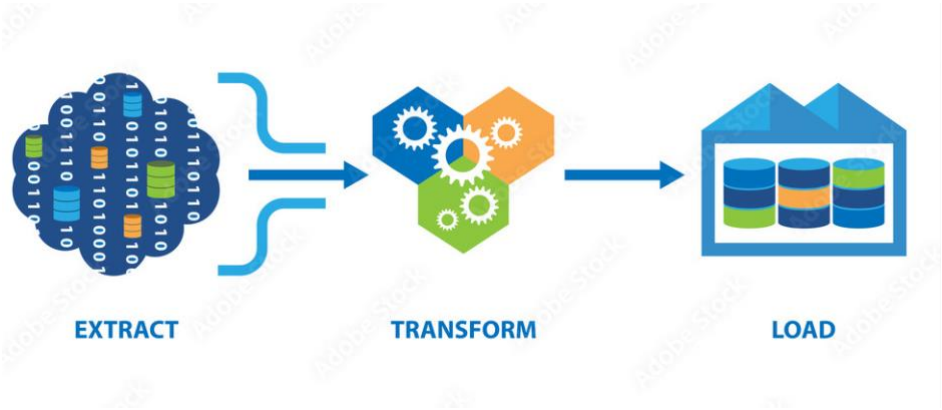


Figure 2: Simple diagram of ETL process

Extraction is the process of obtaining raw data that is unprocessed and unprepared from various data sources (like marketing, sales, employee, customer relationship management (CRM) applications, etc.)

Transformation is normally defined as the "data pipeline "where data is prepared by changing its structure to the preferred way by users, utilizing consolidating, aggregating, mixing it with other data from different data sources, normalising it etc., and rendering it useful for business intelligence.

Loading is the final process stage, and that is simply taking the transformed and prepared data from the data pipeline to the specified destination database or warehouse.

In brief, the data can be extracted from many sources, subjected to transformations such as calculations, and then loaded into a repository known as a data warehouse. Its primary purpose is to combine data from many sources, and it is frequently employed in data warehouses. **(Sreemathy et al, 2021)**

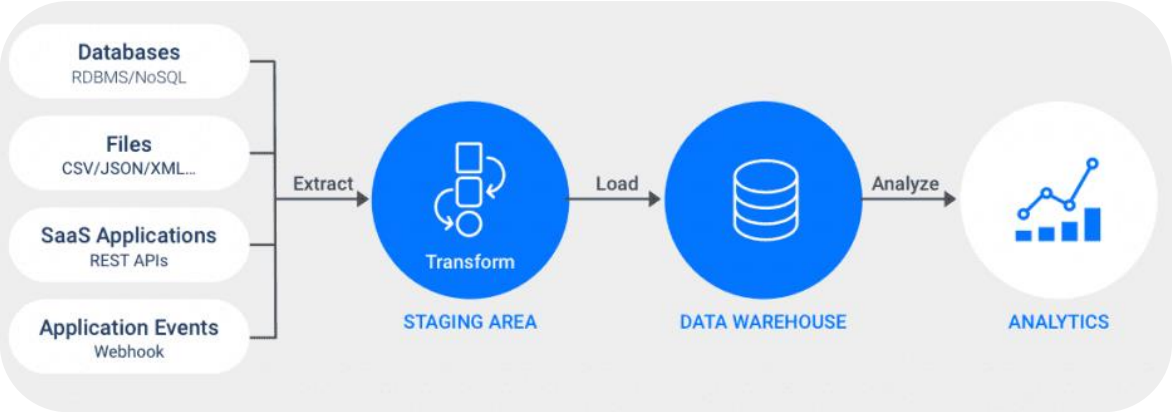


Figure 3: Detailed ETL Process

3.3.3.2. ELT

ELT has the same terms as ETL, but their data integration processes differ from one another. With the help of tools like Apache Hadoop as modern cloud-based data warehouses that offer the computing power to manage write operations on huge data sets efficiently, it became an interesting option for users **(Talend, n.d.)**.

This has ultimately resulted in the development of a new data integration technique, ELT, which bypasses the ETL staging area for faster data input and more agility. ELT transmits unprocessed, unprepared data directly to the data warehouse and relies on the data warehouse to perform transformations after loading. **(Talend, n.d. & Dearmer 2020)** In fact, some cloud data warehouses are so efficient at data processing that they have rendered ETL obsolete in many use cases.

In contrast to ETL, the process of **Transformation** and **Loading** data are reversed. The following explains the integration process.

Extraction remains the initial stage for extracting data from various data sources.

Loading becomes the second step of the integration process, making ELT unique as it diverges from its ETL counterpart. ELT migrates a massive amount of raw data to the users' destination, where it will ultimately reside, as opposed to loading it onto a temporary

processing server for modification. Although the time between data extraction and transmission is shortened, much more work must still be done before the data is usable.

Transformation becomes the final stage here as the data gets sorted and normalised in the database or data warehouse as the final process for data integration, and some or all of it is kept on hand and available for specialised reporting. Although maintaining this much data has a higher cost, there are more options to custom-mine it for timely business analytics.

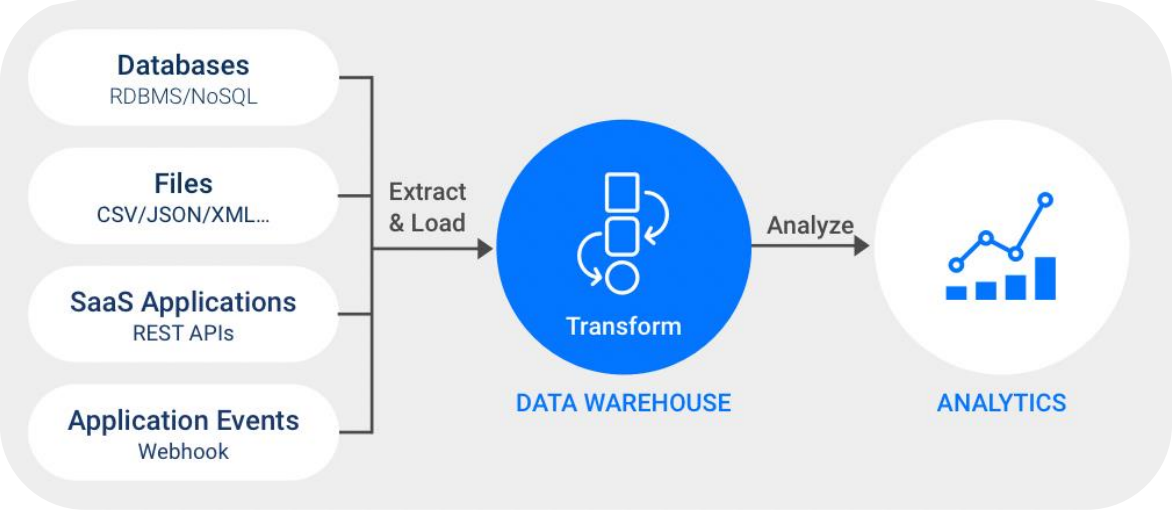


Figure 4: Detailed ELT Process

3.3.3.3. Key Differences

To summarise everything that was mentioned above

Section	ETL	ELT
History	Has been around for more than 20 years; With many use cases examples available online	A much newer concept for Data Integration and its complexity of implementation varies based on the use case
Transformation Process	Raw data that is extracted from the Data Source is transformed and processed on a pipeline server	Raw data that is extracted from the Data Source is transformed after the data is loaded inside of the destination. (i.e., Data Warehouse Cloud Management)
Data	Structured data is transformed and then loaded to the destination	Raw data is loaded to the destination then it is transformed

Cost	May cost more for small and medium-sized firms depending on the data servers or pipelines	Costs are less due to fewer data pipelines, especially when using cloud platforms. It is usually cheaper at first when you have limited data, but the price can scale up with more data.
Maintenance	The additional processing server increases the maintenance workload.	The amount of maintenance work is decreased with fewer systems.
Volume	Handy when complex transformation needs must be met with limited data sets.	Perfect for massive datasets that call for both effectiveness and speed.
Privacy	When data are transformed before loading, it can help prevent personally identifiable information	Requires a layer of safeguards when direct loading data
Speed	Users must wait for the data to be transformed before being loaded into a destination system. Transformation time increases as data size does.	Speed in the ELT process is never based on data size. Data is directly imported into a destination system and transformed simultaneously, making it faster in comparison.
Complexity of development	Easier to develop when it is at an earlier stage	Developers and analysts must have in-depth knowledge and high skill in their data and integrations tools

Table 1: ETL vs ELT

3.3.3.4. Current Available Solutions Worldwide

I list here 4 of the top-rated companies that are affordable to start-ups and middle-sized companies. The following information was obtained directly from their website

- **Integration Type**

Feature/Platform	Stitch	Dataddo	Fivetran	Supermetrics
DI Type	ELT	ETL + Reverse ETL	ELT	ETL
Data Encryption	Yes	Yes	Yes	Yes

Data Residence	Can be selected	Can be selected. Data are stored as cache and then deleted by the next data sync	Can be selected. Data Warehouse is required	Cannot be selected. However, data are stored as cache and then deleted by the next data sync

Table 2: Data Integration Platform Type Comparisons

- **Pre-Build Connectors**

Feature/Platform	Stitch	Dataddo	Fivetran	Supermetrics
No. of Available Connectors	100+	200+	100+	100+
Duration to Build New Integration	Users should develop it by themselves using their platform	10 days	Users should develop it by themselves using their platform	Not Transparent

Table 3: Data Integration Platform Type Comparisons

- **Destinations Supported**

Feature/Platform	Stitch	Dataddo	FiveTran	Supermetrics
Dashboarding Applications	No	Yes	No	Yes
Data Warehouses	Yes	Yes	Yes	Yes

Table 4: Data Integration Platform Type Comparisons

- **Pricing**

Feature/Platform	Stitch	Dataddo	Fivetran	Supermetrics
Cheapest plan	$100/month	$99/month billed annually	Not Transparent	$39/month billed annually
Based On	5 Data Pipelines, and only one destination	10 Data Pipelines	Based on Connectors & rows in Data Source	1 Data Source
Includes Free Trial	14 days of full access to the platform	14 days of full access to the platform	Yes	14 days of full access to the platform

Has Free Tier	No	Yes, but limited to 3 Data Pipelines	No	No
Support Multiple Pricing Tiers	Yes, depends on the number of destinations and records per Data Source	Yes, depends on the type and number of destinations you desire	Yes, depends on your use case	Yes, depends on the type and number of destinations you desire

Table 5: Data Integration Platform Type Comparisons

3.3.4. Data Integration Challenges, Solutions and Examples

Over my journey with Dataddo, I have come to understand that BI and analytics systems, it is the same: Data analysts, corporate executives, and business managers use data integration solutions to help them provide a comprehensive picture of key performance indicators (KPIs), clients, operations in the manufacturing and supply chain, efforts to comply with regulations, financial risks, and other facets of business processes. Consequently, they can track business performance, manage operations, and plan advertising and marketing campaigns with greater analytical information at their disposal. **(Stedman 2019)**

Let us take financial traders as an example. **Stedman (2019)** explains how they must monitor the inflow of market data originating from both internal and external platforms. Another example can be loan officers when granting mortgages. Before doing so, you are required to review account information, credit histories, property prices, and other data. Data from various sensors are used by pipeline operators and plant managers to monitor equipment. Data integration saves users from having to manually integrate the essential data in these and other apps by automatically gathering it.

However, unique technological problems arise when several data sources must be combined into a single structure. Businesses are charged with developing pre-built procedures for reliably getting data that needs to be sent to the preferred destination as more businesses develop data integration solutions. While doing so saves money and time in the near run, there are many potential barriers to implementation. **Ziegler Dittrich**

(2007) and **Talend's research** highlighted a few challenges that are quite common when businesses build their integration arsenal.

- The goal of data integration for businesses is often the resolution of a particular problem. They frequently fail to consider the route that will be necessary to get there. The various types of data that must be gathered and analysed, the sources from which those data must be obtained, the systems that will use the data, the types of analyses that will be carried out, and the frequency at which data and reports must be updated are all things you must be aware of when implementing data integration solutions.

- The work isn't finished until an integration system is operational. The data team now must keep data integration initiatives compliant with best practices and the most recent demands from the business and regulatory bodies.

- Today's new systems generate many types of data from a range of sources, including movies, IoT devices, sensors, and the cloud, such as unstructured or real-time data. In order for your organisation to succeed, you must immediately adjust your solution infrastructure to match the needs of integrating all of these data. This is particularly challenging due to the additional issues that the volume, pace, and new format of the data provided.

- Data from old systems may need to be included in integration attempts. However, more contemporary systems frequently incorporate markers for activities, including times and dates, in their data.

- It may be challenging to conduct the same rigorous analysis of data obtained from external sources since they may not give the same degree of information as internal sources. Additionally, it could be challenging for the company to exchange data due to contracts with other providers.

3.4. Key Applications of Data Integration

3.4.1. Data Warehouse

According to **(Song 2009),** DWH is an integrative database of information that has been organized in a way that makes it simple for business decision-makers to understand, interpret, and analyse the information. DWHs have gained popularity by meeting the demand for a consolidated storehouse of business data used in decision-making, Online transaction processing (OLTP) systems, also referred to as operational database systems,

Data warehouse aids in routine company operations. The business intelligence processing typically supported by a DWH, however, is tactical or strategic. A DW system is optimized for complicated decision-support queries, whereas an OLTP system is optimized for quick transactions. As a result, operational database systems and data warehouse systems are often maintained independently. DWH systems and OLTP systems differ greatly from one another because of this disparity. Examples of DWH solutions are, PostgreSQL, Google Cloud (Like BigQuery), MariaDB …etc

3.4.2. Data Virtualisation

(Reeve 2013) describes Data Virtualisation (DV) in her book as a solution that enables an organization to deliver a real-time integrated picture of data gathered from diverse sources and technologies and formatted into the desired form to their consumer. It's only that earlier technology solutions tended to be too slow to enable real-time transformation and consolidation, even though this is not a new business desire.

It is largely used to deliver integrated business insight, a function previously exclusive to the data warehouse. The concept of a data warehouse can be expanded to include data that is not directly under the physical data warehouse's control thanks to DV.

Because it wasn't possible to do so in real-time with a response time suitable to business analysts, data warehouses were developed largely as an instantiation of an integrated view of data. The information is merged and transformed through DV from both structured Data

Sources of business intelligence and unstructured Data Sources, which is the most fascinating part. Examples of DV tools are Google Data Studio, Power BI, Zoho Analytics …etc

3.4.3. Data Replication

Data replication is similar to data mirroring, which means that data is copied from one database to another for backup purposes and to maintain data synchronization with operational needs. Both can be used on both servers and individual computers. The same system, on-site and off-site servers, and cloud-based hosts can all store duplicate data.

Asynchronous data replication, in which replication begins only when the database receives the commit statement, is an alternative to synchronous data replication, which replicates any changes made to the original data.

Modern database solutions frequently leverage third-party tools or built-in features to replicate data. Although Microsoft SQL and Oracle Database actively provide data replication, some traditional technologies might not come with this feature by default.

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

- When accessing data from multiple places around the world, consumers may encounter some latency in enterprises with numerous branch offices dispersed throughout the globe. Users benefit from quicker data access and query execution times when replicas are placed on local servers. By storing data at many nodes around the network, data replication improves the resilience and dependability of systems. Hence, increasing data accessibility

- Distributing the load among the distributed system's nodes, also lessens the burden on the main server, enhancing network efficiency. IT managers can reserve the primary server for write tasks that require more processing power by sending all read operations to a replica database.

- By maintaining reliable backups at places that are closely watched, data replication makes it easier to retrieve lost or damaged data in the event of a data breach or hardware malfunction that results in data load. This improves data protection.

3.4.3.2. Schemes of Data Replication for Databases

Replication of DWH comes in two schemes: **(ManageEngine, n.d.)**

- **Full Replication -** when a database's entirety is replicated for use across several hosts. Slower update procedures and trouble maintaining consistency between each site are drawbacks of this scheme, especially if the data is continually changing.

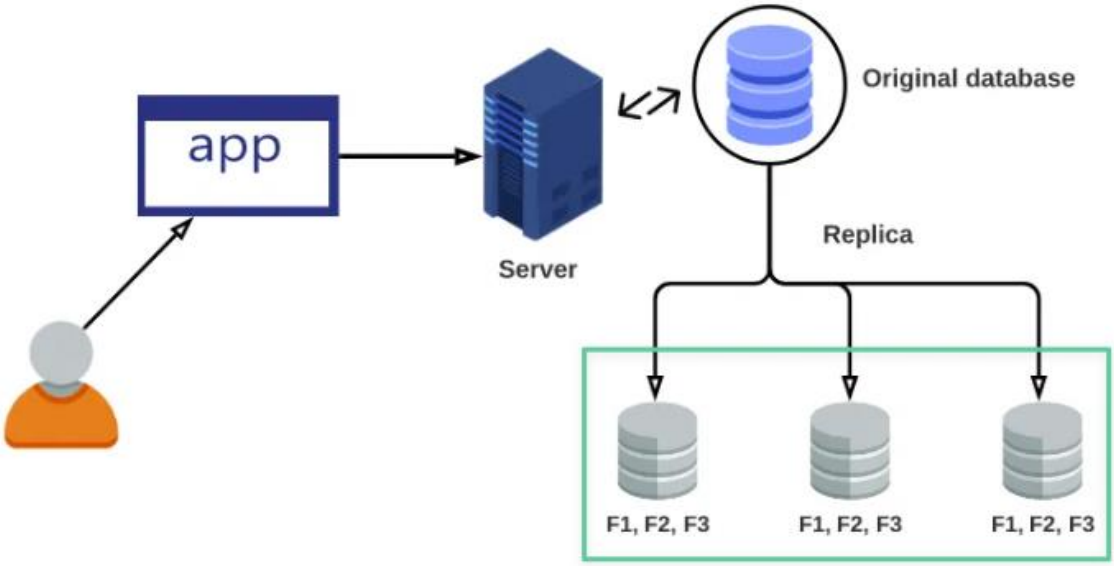


Figure 5: Full Data Replication

- **Partial Replication -** The database is divided into pieces, and each section is stored in a distinct location according to the importance of that area. While the headquarters maintains a complete collection of data for high-level analysis, it may be more effective for analysts to store specific categories of data where they should be stored. For example, employees' data of a specific enterprise with multiple branches would be stored in the respective branch, and so on. This keeps the data close to the analysts, supervisors, etc.

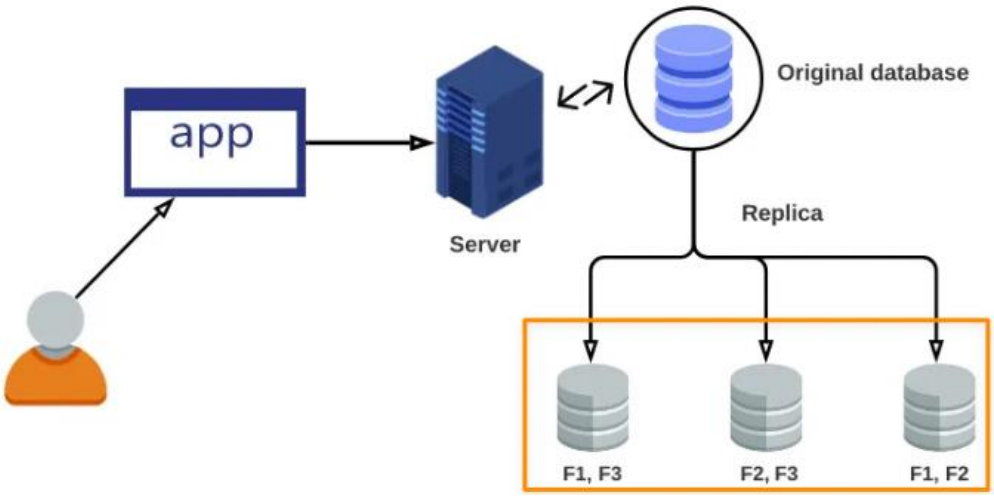


Figure 6: Partial Data Replication

3.4.4. Data Streaming

Steaming refers to data integration in real-time by a technique that continually integrates and feeds various data streams into analytics systems and data repositories. To support data-driven decisions to enhance customer experience, reduce fraud, and optimise operations and resource utilisation, stream data integration increase the efficacy of event data from across the company and makes it accessible in real-time.

Organisations' use of streaming data is becoming more complicated, opening up possibilities for more in-depth analysis and improved operational intelligence. Data management plans for real-time integration and analytics must include stream data integration. This is a requirement for data and analytics leaders. **(Gartner 2019 & Tatbul 2010)**

3.5. What does Data Transformation mean?

3.5.1. Introduction

Manikandan S. (2010) describes data transformation as data preparation and checking, generating derived data from the original values, statistically controlling for outliers, and data transformation—all steps in the preparation of the data that enable statistical analysis.

For instance, it is the process of converting data from one format to another, such as from an Excel spreadsheet, an XML document, or a database file.

Transformations often require transforming raw data sources into formats that have been cleaned, verified, and are suitable for use. Data integration, data migration, data warehousing, and data preparation are all steps in the data management process that require data transformation. ETL is known to be a perfect example of data transformation. During the extraction stage, data must be located, extracted from the numerous source systems that provide it, and then moved to a single repository. The raw data is then cleaned, if necessary. It is then converted into a target format that may be used by business intelligence and analytics applications or fed into operational systems, a data warehouse, a data lake, or another repository. Data types may be changed, redundant information may be eliminated, and the source data may be enhanced as part of the transformation**. (Pratt 2022)**

3.5.2. How important is Data Transformation?

Data transformation is necessary for a variety of processes, including data management, data transmission, data warehousing, and data wrangling. It is also a vital component for any company intending to use its data to deliver pertinent business insights. As the volume of data has expanded, organizations require a trustworthy way for utilizing it in order to use it effectively for business goals. Data transformation, when done correctly, ensures that the data is trustworthy, consistent, safe, and trusted by the consumers. Data transformation is one element of utilising this data effectively.

Data transformation can be constructive (adding, copying, and replicating data, combining columns in a database, and/or renaming them), destructive (eliminating fields and records), or aesthetic (standardizing salutations or street names, for example). **(Stitch, n.d.)**

3.5.3. Data Transformation Benefits

Data analysis is required by businesses of all sizes for a variety of business processes, from supply chain management to customer support. Additionally, they require data to support

the growing number of automated and intelligent systems in their organisation. **(Pratt 2022)**

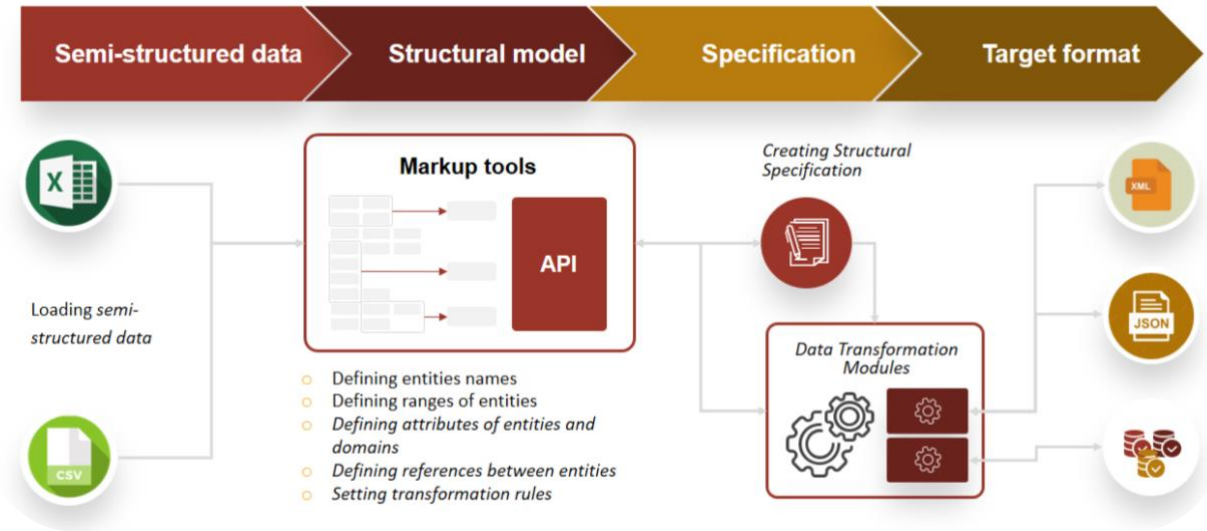


Figure 7: Fundamentals of Data Transformation

Organisations require high-quality data in appropriate formats for the systems receiving the data in order to gain insight into and enhance their operations. The interoperability or adaptability of applications, systems, and data formats is made possible via data transformation. It could be necessary to modify data differently if it is utilised for different reasons or how a system inputs and accepts data.

Data is altered (or transformed) to improve its structure. Data that has been transformed could be simpler to utilise for both people and machines. Improved data quality and protection from possible landmines like null values, unexpected duplication, wrong indexing, and incompatible formats are two benefits of correctly structured and verified data. **(Stitch, n.d.)**

3.5.4. Challenges that arise from Data Transformation

According to **(Pratt 2022 & Stitch, n.d.)**:

- Costly data transformation is a possibility. The cost is determined by the particular infrastructure, software, and processing-related instruments employed. The cost of recruiting the required employees and purchasing system resources are only a few examples of expenses.

- Transformations that don't meet an organisation's needs can be carried out. For one application, a company could convert information to a certain format, only to change it back for another application to the original format.

- Lack of knowledge can cause issues during transformation. Lacking the necessary subject-matter expertise, data analysts are less likely to catch typos or inaccurate data since they are less knowledgeable about the gamut of accurate and allowed numbers. For instance, a person working with patient records who is not familiar with the terminologies used in the field may overlook misspellings or fail to indicate illness names that need to be mapped to a specific value.

- Processes for transforming data can be resource intensive. Transforming data in an on-premises data warehouse after importing it or before feeding it into apps might add computational overhead and impede other processes. If you use a cloud-based database or warehouse, you can do the transformations after loading because the platform can scale up to meet demand.

3.5.5. Methods for Data Transformation

Data transformation comes in various methods: **(Pratt 2022)**

- **Discretisation -** It includes breaking down continuous data values into groups of intervals with predetermined values to organise the data and make it easier to analyse.

- **Generalisation -** to have a more complete understanding of the data, low-level data characteristics are transformed into high-level data attributes. Like transforming data from several sets divided by age (like old and young),

- **Aggregation -** where information is gathered from several sources and kept in a single manner.

- **Constructing Attribute -** the process by which new characteristics are produced from existing attributes or added.

- **Manipulation:** when information is modified or updated to make it easier to understand and organise.

- **Smoothing:** It uses algorithms to lessen "noise" (also known as "white noise" in econometric terms) in datasets, making it easier and more efficient to spot patterns within the data.

- **Normalization is** a method used to reduce the amount of duplicate data by converting the original data into a different format.

3.5.6. Data Transformation Solutions in the Market

There are a variety of solutions that are available out there on the market for data professionals to aid in the ETL process. There are tools for both commercial and open-source data transformation, some of which are intended for on-premises transformation procedures and others for cloud-based transformation operations.

For expensive solutions, there are the likes of SAP and IBM Info Sphere, for big companies that can afford it and had a large number of data records. And there are a few emerging affordable solutions for small and medium-sized companies, like Stitch, Fivetran, or the company I work for, Dataddo, which is based here in the Czech Republic. However, solution companies that offer a wide variety of capabilities for handling enterprise data are what other ETL solutions on the market are made up of.

These technologies replace much, if not all, of the manual scripting and hand coding that had previously been a significant component of the data transformation process by automating many of the procedures involved in data transformation. Additionally, some solutions for data transformation are concentrated on the actual data transformation procedure, managing the series of steps needed to change data.

3.6. How vital Data Integration is to business nowadays?

The majority of businesses have a variety of data sources, frequently including external ones. Business applications and operational staff frequently require access to data from several sources to carry out transactions and other operations. For instance, in order to fulfil orders, a contact centre employee needs access to the same combination of databases for customers, products, and logistics that an online order entry system requires. **(Stedman 2019)**

Big data, with all its benefits and challenges, is being embraced by businesses that want to remain innovative and competitive. An integrated view of key performance indicators (KPIs), financial risks, clients, activities in the manufacturing and supply chain, attempts to comply with regulations, and other aspects of business processes are provided to firm managers and data analysts by integrating customer data.

3.7. Data Integration Flows for BI

Historically, ETL design and execution used to be generally disregarded by researchers since it was seen as a supporting function for the data warehouse. However, it is still a costly, time-consuming, and primarily manual operation. In reality, ETL may account for a significant portion of the work in a data warehouse project.

Correct functionality and acceptable performance have been the emphasis of ETL, which means that the functional mappings from data sources to the warehouse must be accurate and the ETL mappings must finish within a specific time frame. Even though they are more difficult to quantify, additional business objectives that are crucial to success are missed by an ETL project that just concentrates on functionality and performance.

Several additional needs for the architecture—in particular for the back-end data integration processes—are imposed by operational BI. These include managing a much wider variety of data types, including unstructured and semi-structured streaming data; low latency requirements to support online decision-making; quick refresh cycles; more

advanced analytic and reporting tools; a greater number of data connections; round-the-clock availability; and so forth.

ETL processes are becoming more general data flows in the evolving operational BI architecture instead of being a one-way, batch pipeline. For example, events from the sources are streamed through transformation operations toward the data warehouse, and cleansed data can be sent back to its source, this is a term that we used quite often in my company, Dataddo, **as Reverse ETL**

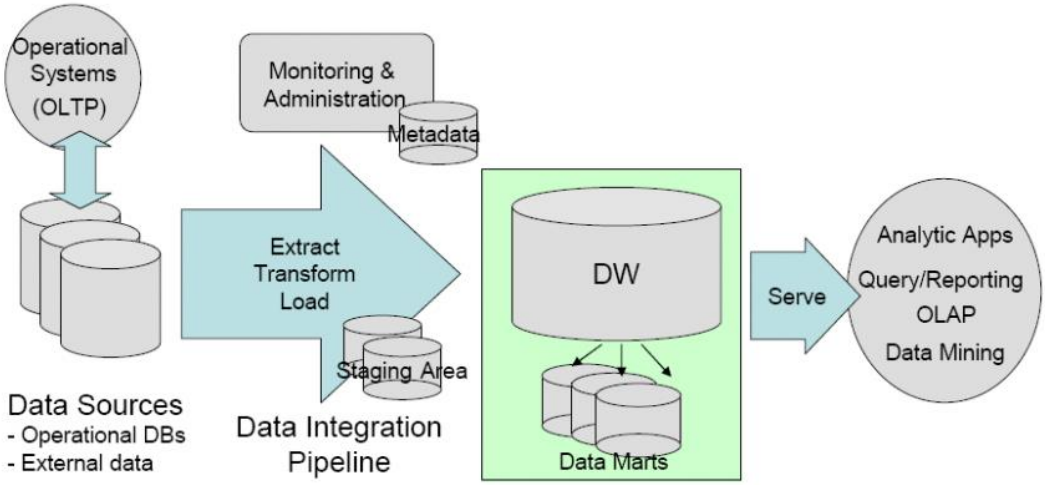


Figure 8: Traditional architecture for BI

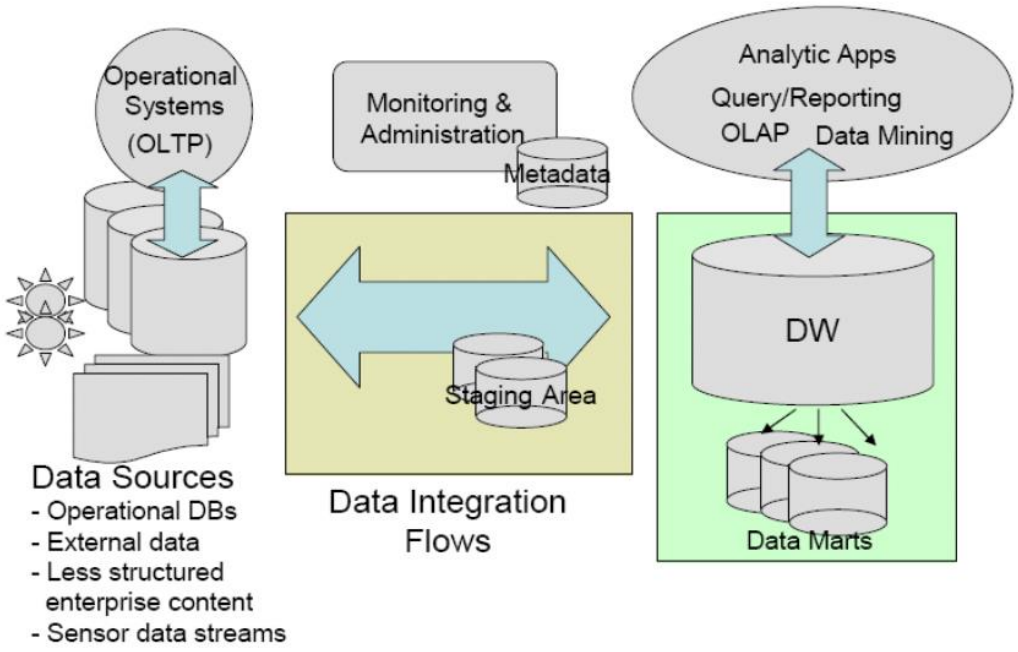


Figure 9: Modern architecture for BI

3.8. Data Integration vs Application Integration

As a result of the widespread use of relational databases and the increasing demand for efficient information transfer across them, usually including data at rest, data integration technologies were developed. Application integration, on the other hand, controls the real-time integration of operational, live data between two or more applications. **(Ziegler, Dittrich 2007)**

The main objective of application integration is to allow independently developed applications to work together. This method makes use of integration programs, which connect to numerous data sources and provide the user with consolidated results. For a limited set of component systems, this technique is workable. In order to achieve this, it is necessary to coordinate the coordinated flow of a variety of operations carried out by various applications, guarantee data consistency among various copies of the same data, and provide a single user interface or service from which to access information and functionality from independently developed applications.

On the other hand, a frequent approach to completing application integration is through cloud data integration, which connects several applications for the real-time exchange of data and processes and allows access by many devices through a network or the internet.

3.9. Data Silos & Big Data Integration in Enterprises

3.9.1. What are Data Silos?

(Patel 2019) describes "data silos" as separate collections of data kept in several corporate applications and refers to them as "data silos." Data silos prevent teams from collaborating and exchanging information. According to research by PricewaterhouseCoopers (PwC), it causes bad decisions and has a detrimental effect on profitability.

3.9.2. What causes Data Silos?

Data silos develop for a number of reasons, including structural, vendor contractual, and political factors. Data from silos must be manually gathered and integrated, which can take days, months, or even years. Even if such were the case, doing so without a sound plan would take a long time. Data silos are created by the hundreds, if not thousands, of servers or business applications that are becoming more prevalent in the corporate sector. When mergers and acquisitions occur, it gets worse, as it becomes difficult for developers to integrate the data into the new merger organisation's system. On the other hand, granting access to all applications to every employee would not be practicable, because, for example, it can cause clients, employees, or any sensitive data to be exposed to everyone, which is a major security risk to our current standards.

3.9.3. Challenges with Data Silos?

Based on **(Patel 2019)** research, 97% of executives believe that silos have a negative impact. Data silos limit visibility across industries and might have various representations of the same issue or circumstance. This results in inefficiencies and extra efforts to ascertain whether the source is reliable or not for organisations

3.9.4. How bad can Siloed Data be to any enterprise?

Data Silos can be detrimental to any enterprise as a whole. The team's ability to see the data is constrained by an information silo. Multiple organisational silos can cause serious issues with how well individuals and teams work together to achieve a common objective.

It is simple for data to go out of sync and produce errors when it is segregated in other data sets or a separate data warehouse. Without active data management, users can be using outdated spreadsheets or be unable to access the most recent information.

Using outdated data might have a detrimental effect on the company. Executives and workers might run into major issues when they can't trust the data they have. Business intelligence can suffer from a lack of openness that causes individuals to doubt the data, draw competing conclusions, and make bad business decisions. This may restrict how much money a firm may make, as well as how much money salespeople can make.

3.9.5. Big Data Integration?

"Big data integration" refers to the sophisticated data integration techniques developed to manage the enormous amount, diversity, and velocity of big data. It unifies data from sources such as online data, data from the Internet of Things (IoT), data from social media, and machine-generated data. Different volumes of data are produced by large companies in real-time, very near to real-time, or not at all. Around 80% of the data from businesses is either unstructured or semi-structured. Big data technologies make it simpler to merge numerous data sources at scale. The integration of data silos used to be possible with Hadoop and Map Reduce.

The requirement for scalable and quick big data analytics solutions emphasizes the necessity for a standardized data integration platform. This platform should make it possible to profile and quality-check data, and it should also produce insights by giving users the most complete and up-to-date view of their business.

Big data integration services use real-time integration techniques in addition to traditional ETL techniques to provide continuously flowing data with dynamic context. Users should develop real-time systems and applications, parallel and coordinated ingestion engines, resilience in each stage of the pipeline to prepare for component failure, and standardization of data sources with APIs for improved insights. The moving, soiled, and temporal characteristics of real-time data are addressed by these best practices.

3.9.6. Big Data Integration for Bridging Siloed Data

According to **(Patel 2019),** the following are summarised.

- Since many applications and processes create data silos, data are stored in a variety of locations, including databases, flat files, cloud storage, on-premises servers, and application servers.
- Finding value in data and defining which data silos should produce the most value if they are interconnected is crucial here.
- Planning sources of data silos to improve collaborations and communications among various departments and teams is one strategy to overcome data silos.
- Data silos will be destroyed by merging various processes and apps after first isolating them.
- However, it necessitates significant work and a shift in the organization's entire culture.
- Utilizing tools and techniques for integration, this issue may also be resolved by integrating various data silos.
- There are many frameworks and technologies available for data integration but owing to the long-term advantages of big data integration, we will concentrate on it.
- Use a straightforward illustration to comprehend the necessity of merging data silos in the business sector.
- A clearer and more accurate understanding of the whole picture will be very helpful when executives and investors evaluate a firm as a whole.

3.10. Data Governance

3.10.1. Introduction

The exercise of power and control over the administration of data is known as "data governance". Any firm that uses big data must have a solid data governance plan. It defines the procedures and roles necessary to guarantee the accuracy and safety of the data utilised within a company or organisation. Data governance makes ensuring that responsibilities and accountability are established within the organisation and that roles connected to data are clearly defined. Therefore, with data governance, you seek to maximise the value of data while reducing the cost and risk associated with its use. **(Abraham, Schneider Broke, 2019)**

Figure 10: Data Governance Framework

3.10.1. Challenges of Data Governance

The main difficulty may be people (as in personnel issues) and organisational problems. Every company transformation requires a champion to drive the change, along with accountable roles and duties. Additionally, a cultural transformation is necessary from one of low relevance and boredom for data management to one of high importance. Employees

who generate, modify, utilise, or otherwise transfer data—especially sensitive data—need to be aware of their responsibility for its appropriate maintenance and accept responsibility for it.

The exponential growth of data, which is only growing more common over time, is another significant difficulty. A large portion of this data is unstructured. This necessitates new procedures and increased governance needs.

3.10.2. Benefits of Data Governance

Eliminating data silos inside a company is one of data governance's main objectives. Data governance is a collaborative process in which stakeholders from diverse business units work to synchronise the data in those systems. Additionally, data governance can assist in finding a balance between required privacy protections and data-gathering methods.

Data governance benefits include better data quality, cheaper data management expenses, and expanded access to necessary data for data scientists, other analysts, and business users, in addition to more accurate analytics and higher regulatory compliance. In the end, data governance may assist in enhancing company decision-making by providing executives with better data. **(Stedman 2022, Farmer 2022)**

Figure 11: Data Governance for all reasons

3.10.3. Who should enforce Data Governance?

The data governance process involves a variety of personnel in the majority of firms. Business leaders, data management specialists, and IT workers are all included. These are the main players and what their main governance duties are. Frequently, the top executive in charge of a data governance programme is the chief data officer. In addition, Data managers are in charge of keeping data sets organised by keeping watch on them. **(Stedman 2022)**

3.10.4. How to develop a framework for Data Governance

The model that serves as the cornerstone for data strategy and compliance is referred to as a data governance framework. It might be appealing to invest in a governance plan that guarantees a mostly pre-packaged set of guidelines and resources. Another problem is when new frameworks are introduced to existing ones. This may cause existing governance procedures to fail. Therefore, the best frameworks are the ones that are

developed in-house, that can also be built with the help of specialised vendors **(Farmer, 2022).**

The governance model first overlays the rules, actions, responsibilities, protocols, and operations that define how those data flows are governed and regulated, starting with the data model that represents the data flows, as in storage settings (input and output)

Consider the model to be akin to a template for the way data governance is handled inside a given business. Keeping in mind that each organization's governance structure will be different, reflecting the particulars of its data systems, activities, and responsibilities, as well as legal and industry standards. The framework should include the following:

- **Structure of the organisation:** tasks, responsibilities and duties of the executive sponsor, head of data, CFO, business team, and accountable owner.
- **Scope of Data:** Big Data, Master Data Manager, Transactional, Operational, Analytical, etc.
- **Metrics and supervision:** metrics for gauging the effectiveness of a plan.
- **Data guidelines and standards:** guidelines that describe what you are overseeing and managing, as well as the intended results.

3.10.5. Steps to model your Data Governance

Every firm is different when it comes to data governance, but all can learn from using the same straightforward steps to create a programme. Each programme for data governance is unique and ultimately depends on the requirements and assets of a company. Consider these eight actions to set yourself up for successful data governance. **(Salmi 2022)**

Figure 12: Steps to Data Governance

3.10.6. Tools for Data Governance

Search for open-source, scalable technologies that can be easily and affordably integrated with the organization's current environment to identify the best data governance strategy for your business. **(Talend, n.d.)** A cloud-based platform will also enable you to easily connect to powerful capabilities that are affordable and simple to utilise. Additionally, cloud-based solutions do not require the overhead associated with on-site servers. Focus on choosing solutions that will enable you to achieve the business advantages outlined in your data governance strategy.

For these tools to help you improve your data's relevance, searchability, accessibility, likability, and compliance, you should document the data. Self-service technologies should enable those with the most in-depth knowledge of the data to participate in data stewardship duties. Discover your data with tools and capabilities for profiling, benchmarking, and discovery. Improve your data quality by performing data cleansing, validation, and enrichment, while monitoring and reviewing your data actively and regularly. It will also help you in managing your data with the help of metadata driven ETL and ELT, as well as data integration technologies, that provide end-to-end data tracing for data pipelines. **(Talend n.d., Farmer 2022)**

3.11. Data Integration Deployment Models for this Project

The following tools are used to develop the practical part of this study.

3.11.1. Data Transformation Tools

Throughout all my practical parts, I will be using the following tools for all the use cases of this project. Starting with,

- **Dataddo Platform**

A completely managed, no-code data integration platform called Dataddo links dashboarding tools, data warehouses, and data lakes with cloud-based applications. Any degree of data maturity in a business can be supported by it.

Dataddo, which was established in Prague in 2015 and now has its headquarters there, provides services to more than 3000 businesses, individuals, and organizations worldwide, including some of the most well-known names in the business. **(Dataddo n.d.)**

It supports more than 200 prebuild connectors and support manual data transformation for the more experienced user or those who are looking to make dynamic data source and send them to their destinations for analytics purposes. (See section 4.1)

- **Mongo Aggregation Pipeline**

A unique flow of actions that processes, alters, and provides results is referred to as the aggregation pipeline. As Dataddo supports manual data transformations, they use the MongoDB aggregation pipeline tool for transforming RAW and unstructured data into more coherent data. (See section 4.1.2)

3.11.2. Personal Backend API

For this section, I am building an API server that is hosted publicly and only accessed with a secret token (See section 4.2.2.3). The API and the hosting server are composed of

- **Node JS + Express (TypeScript)**

Node JS refers to the framework that is built on Chrome's JavaScript runtime, while Express JS is a framework of Node JS that is added as a layer on top of it, to provide hybrid web applications and manage routes and servers. The project is also you using Typescript to make sure that errors with data are caught during the development process

- **Firestore**

It is a sub-solution of Firebase (a Google product aiding web applications). A NoSQL document database called Firestore was created for quick application development, high speed, and automatic scaling. Although the Firestore interface shares many features with conventional databases, it differs from them as a NoSQL database by how it constructs relationships between data objects. **(Google Cloud, n.d.)**

- **Vercel**

It is a cloud platform that gives programmers the ability to host frontend apps and web services that can be deployed immediately, scale automatically, and don't need any management.

Note: All the codes and repositories are highlighted in the Appendix

3.11.3. Google Cloud Tools

- **Google Data Studio**

A web-based application for data visualization called Google Data Studio enables users to create personalized dashboards and clear reports. It aids in visualizing patterns, comparing performances over time, and measuring important KPIs for clients. (See section 4.1.1 & 4.1.2)

- **Google BigQuery**

It is a managed data warehouse service for enterprises with built-in technologies like machine learning, geographic analysis, and business intelligence that assist you in managing and analysing your data. (See section 4.1.1 & 4.1.2)

- **Google Spreadsheet**

Users can create, update, and modify spreadsheets using this web-based tool, and they can share the data instantly online. The Google product has capabilities that are common to spreadsheets, like the ability to add, delete, and sort rows and columns. (See section 4.1.2.1)

3.11.4. Reporting Data over API

- **YouTube Analytics**

With the aid of the show more option, YouTube Analytics enables producers to get essential & fine-grained information on the subscribers to your channel and their source.

You can learn more about your audience's demographics, including its age, gender, and location, by selecting the "display more" option. Additionally, it aids developers in gaining technical knowledge about things like operating systems, the language used, etc. (See section 4.1.1)

- **Covid Data**

This is an open-sourced API using data from the Robert Koch Institute. The JSON Rest API may be used to search through all pertinent corona data for Germany, like the number of cases, deaths, recovered patients, number of hospitalisations…etc. It is developed by Marlon Lükert (See section 4.1.2.2)

4 Practical Part

The main objective of this part is to shed light on some use cases for clients that use Data Integration, in collaboration with Dataddo and with the help of our platform. During my time working with Dataddo, I found that the following 6 use cases are common among our clients, and thus, this project will shed light on how different Data Sources can be transformed and migrated to the Data Destination of our choice

4.1. Use Cases

For the following use cases, the only thing that will be in common is using Dataddo's data pipeline for transformation and migrating the data (as mentioned in section 3.11.1) . As I am an employee at the firm, I have unlimited access to their platform, and I can create as many data flows as I wish to conduct my study cases.

4.1.1. No-Code Integration

4.1.1.1. Data Virtualization: Data Source to Dashboarding Application

Throughout my time with Dataddo and the first experiences I had with dealing with our clients, I understood that the most common use case for Data Integration for the client is to virtualise their data. The need to virtualise the data for this use case is to make reports for analytic purposes. For example, if a YouTube content creator wants to track the performance of their videos and creates daily, monthly, or weekly reports using their favourite dashboarding tool (for this example, we will use Google Data Studio). They can create unique reports with Google's YouTube Analytics data using the YouTube Analytics API. Reports for channels and content owners are supported via the API. (as mentioned in section 3.11.4)

The data can be called using Postman, or for our use case, we will use Dataddo's platform. In most cases, Data Pipeline platforms usually have YouTube Analytics pre-coded Data Transformation templates, usually called Connectors. Most companies prebuild famous services, like YouTube Analytics, Facebook Ads, LinkedIn Ads …etc, to make it easier for

the user to connect their data by just simply authorizing it to the Data Pipeline platform, then you can configure it to send the data to your preferred destination.

Figure 13: Data Source to Dashboarding Application

Here is an overall activity concept diagram to simulate the goal of this use case

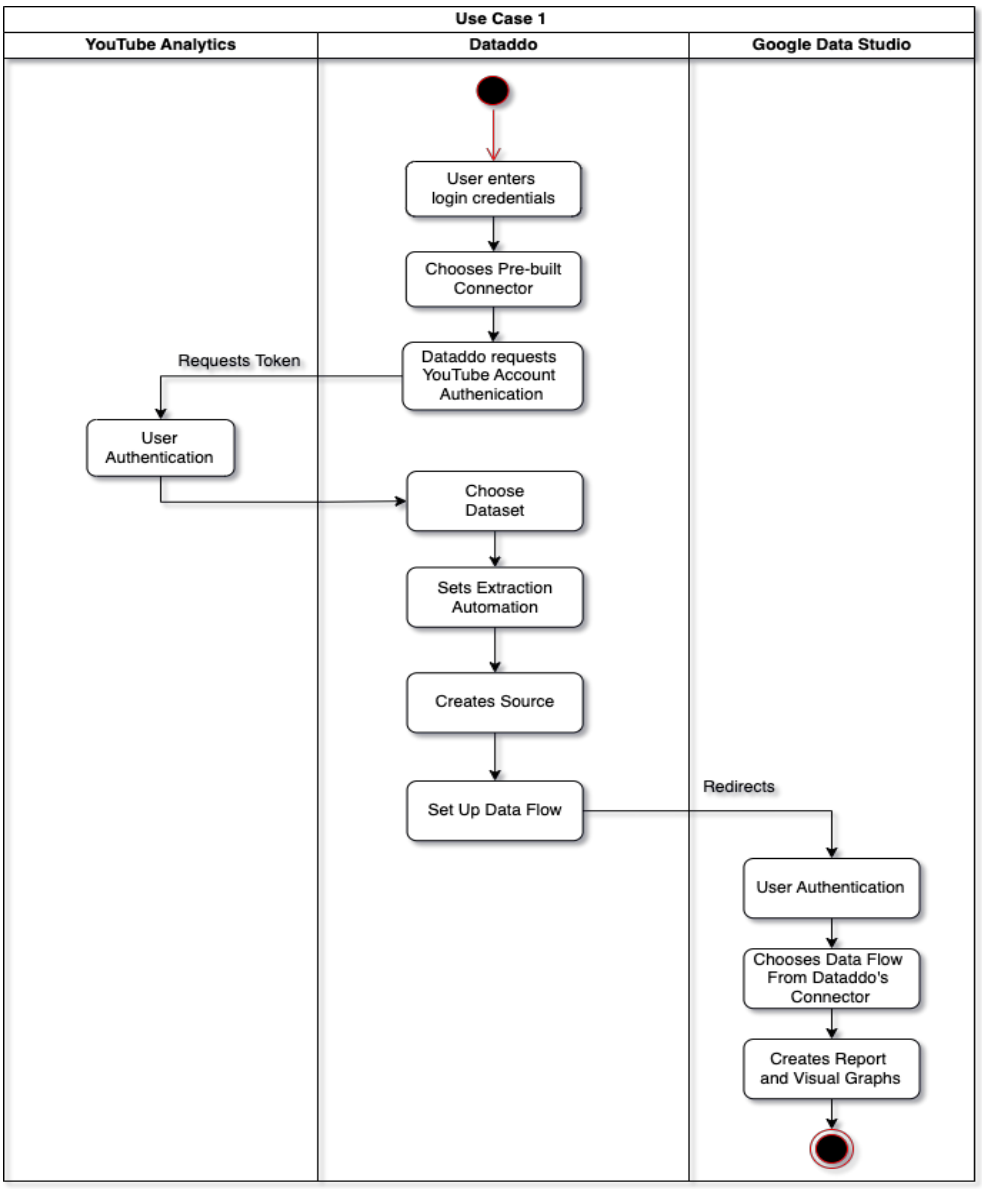


Figure 14: Use Case 1: Activity Diagram

For this use case, I will be using my YouTube account to track my video performance and send it to Google Data Studio to generate reports for analytics. Therefore. I log in to the Dataddo Platform and search for the YouTube Analytics connector.

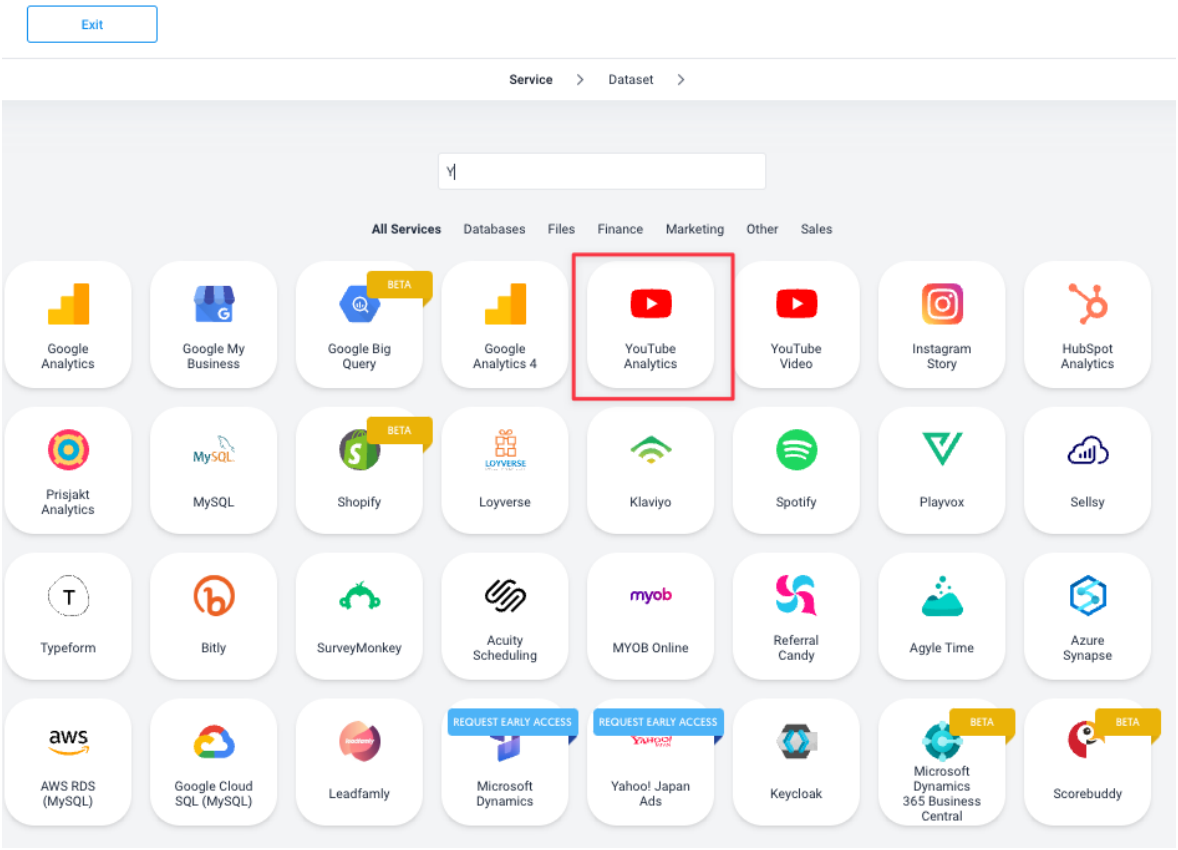


Figure 15: Selecting the YouTube Analytics connector

Once selected, you get to see these *Dataset* sections, which are different prebuild reports based on the API endpoints that YouTube Analytics provide. Some of these endpoints I have personally created or helped in creating (including Datasets from other connectors). These datasets make it easier for the client to choose the type of reporting they want to analyse without writing any code themselves.

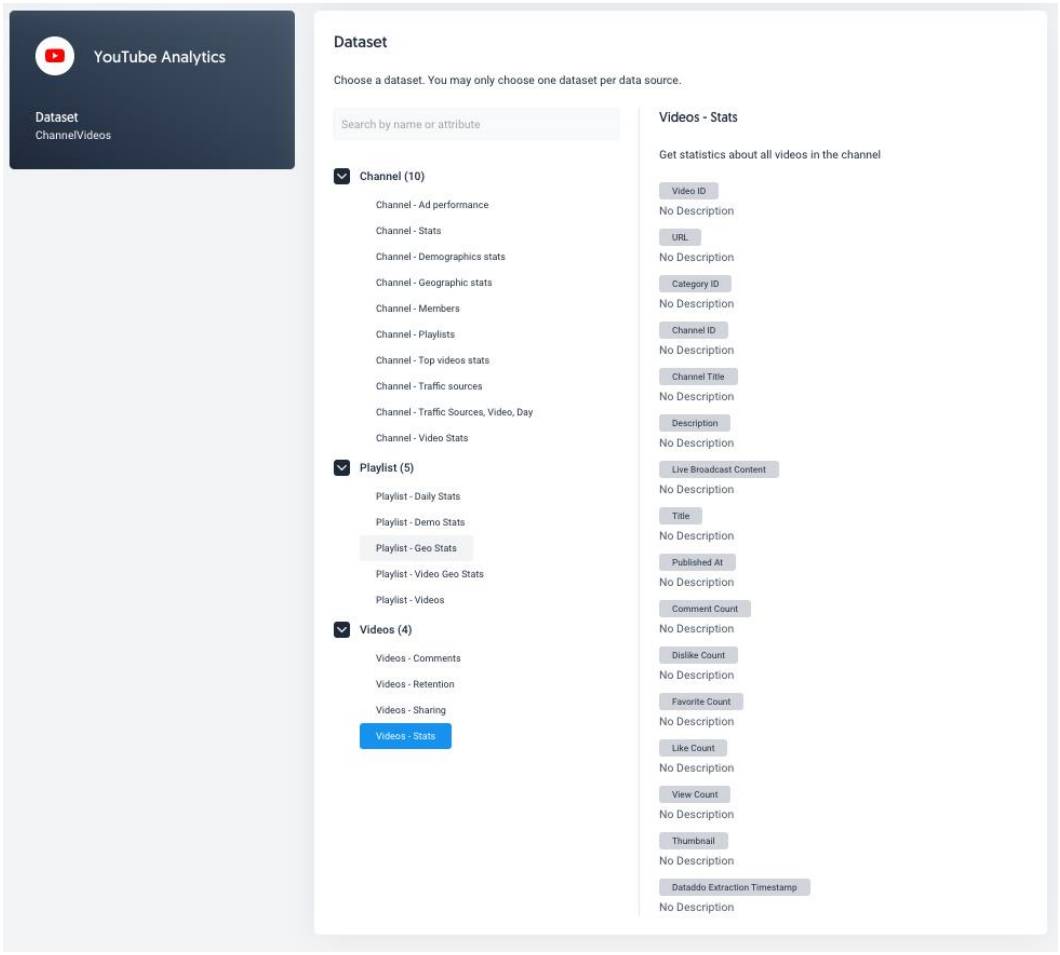


Figure 16: Dataddo's YouTube Analytics Dataset.

In this step, I will be selecting the Video Stats dataset, to be able to track and extract the video views, dislikes, and comment count.

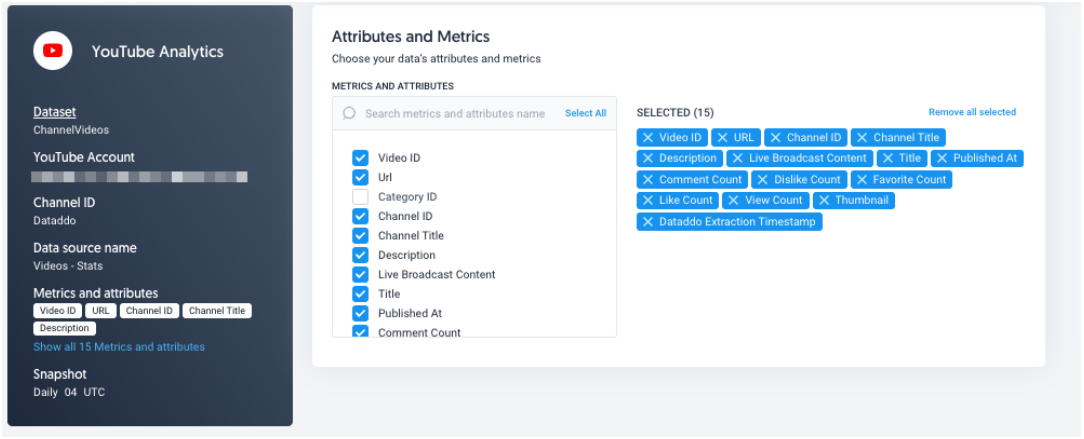


Figure 17: YouTube Analytics Metrics and Attributes

The Metric and Attributes section gives the ability to choose what metrics to include and exclude from your desired report.

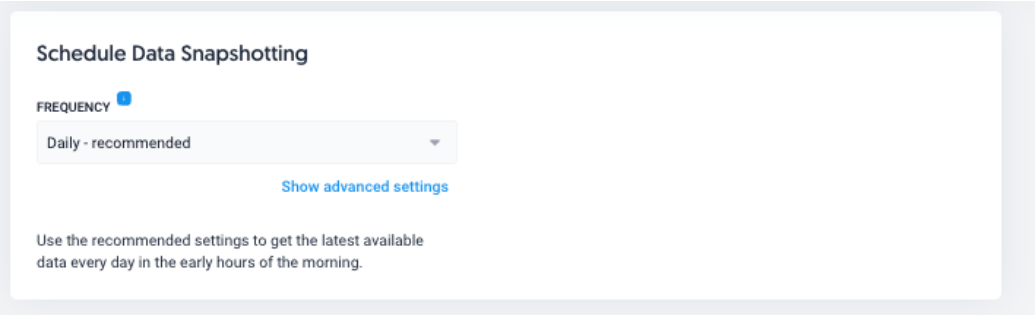


Figure 18: Snapshot Automation

One of the most reasons why clients resort to Data Integration solution platforms to migrate their data is because they lack the skill to automate the extraction on a specific time and date, repeatedly. With such platforms, you can extract new data, whether daily, monthly, weekly, or hourly, and help in constructing better reports with both historical data and new data available to clients on demand.

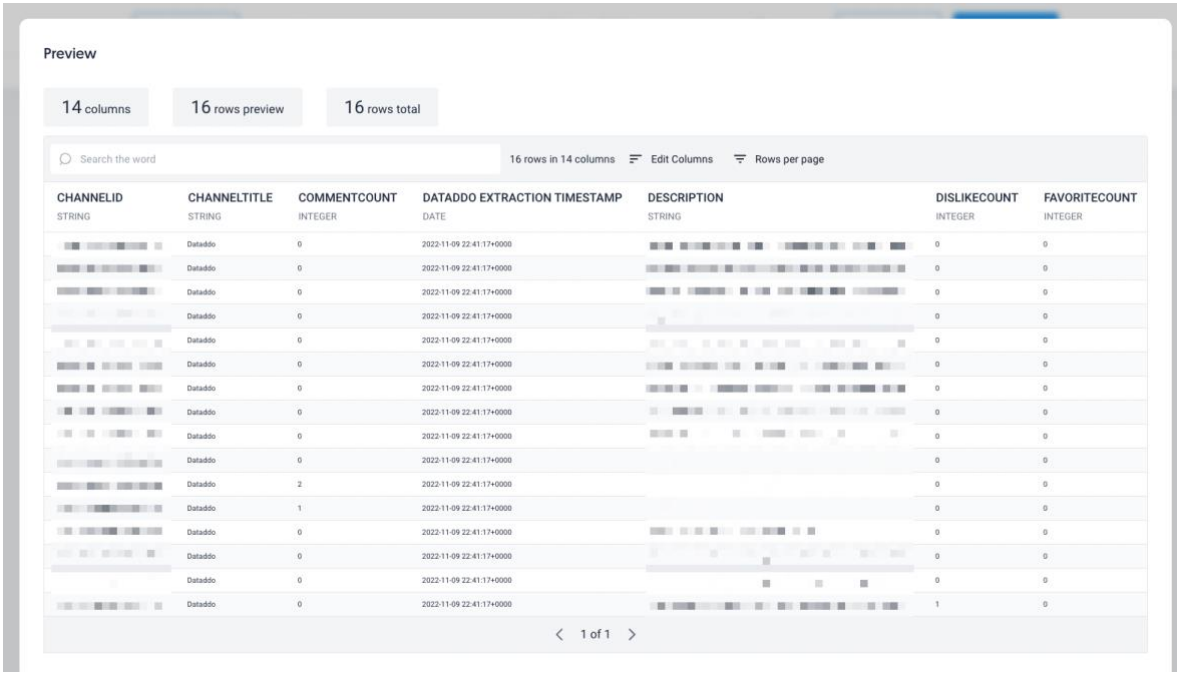


Figure 19: Data Preview

The Data Source is created, and the data can be previewed and ready to be sent to the Dashboard. The next step is to create the Data Flow and configure the Dashboarding report in Google Data Studio

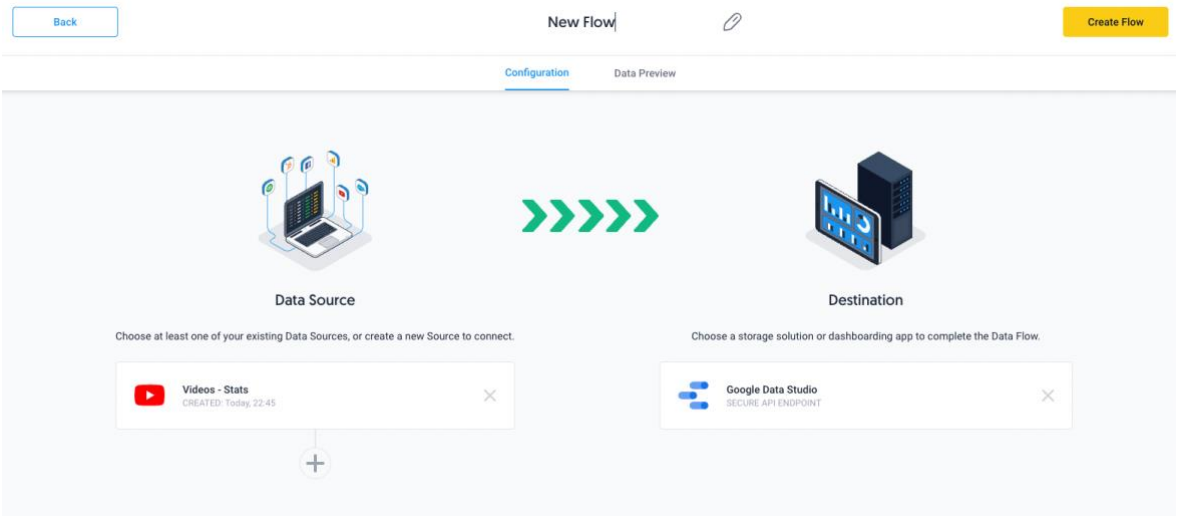


Figure 20: Data Flow Configuration - Google Data Studio

Once the data flow is created, an API key will be generated for the Data Flow to connect with Google Data Studio connectors

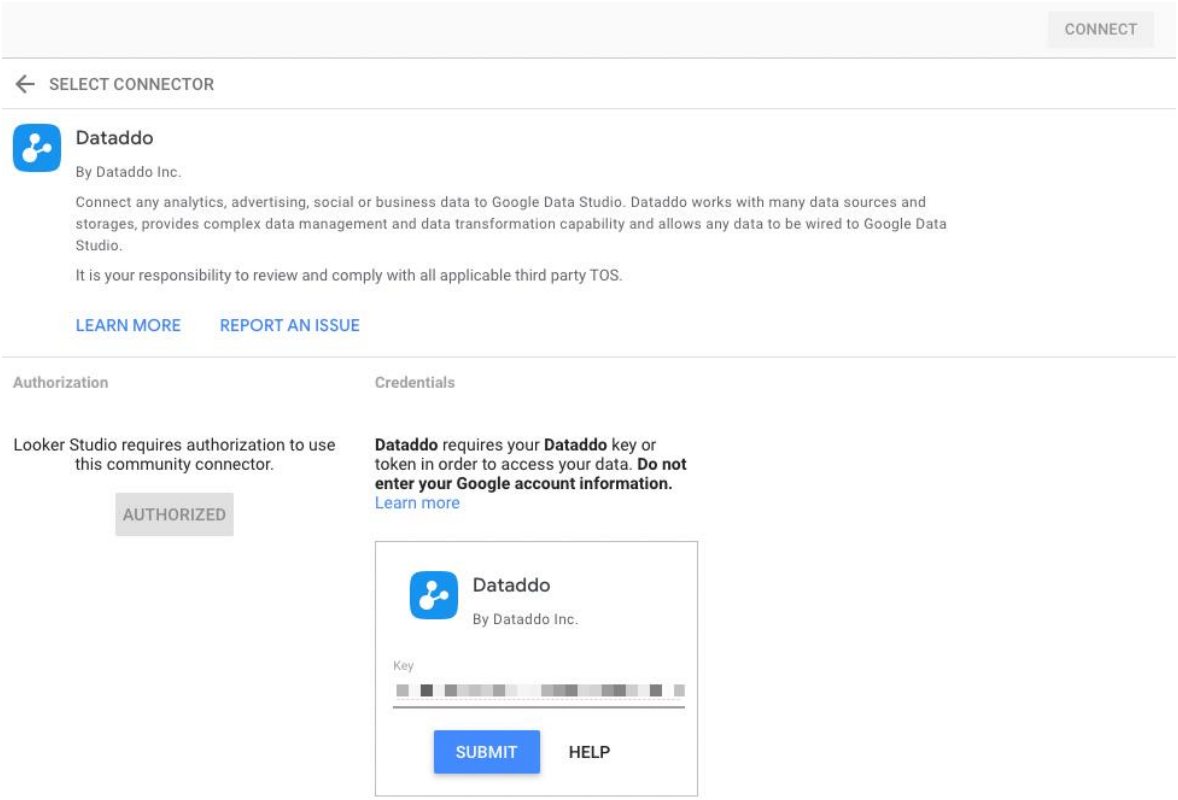


Figure 21: Dataddo Config with Google Data Studio

After configuring the Dataddo API with Google Data Studio, we get to see the last confirmation menu of all the fields from the Data Source before creating the report.

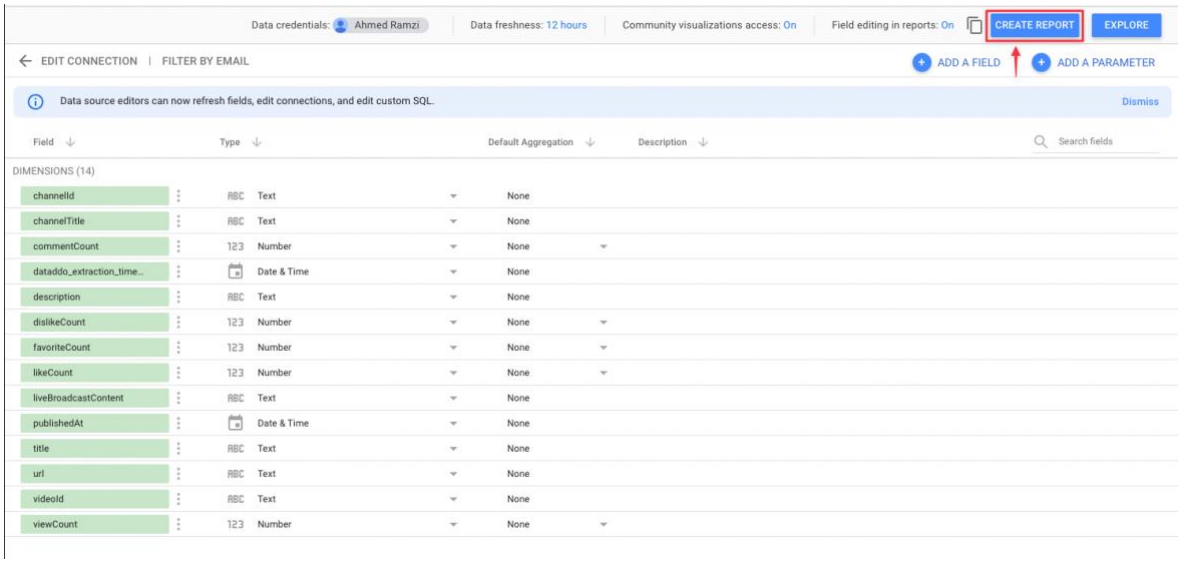


Figure 22: Data Studio Report Configuring

The last step of this stage is to create the report charts and table of your choice that better fits your report requirement. You can customise the Dimension and Metrics of the charts and table to view only the metrics you need and how to virtualise them.

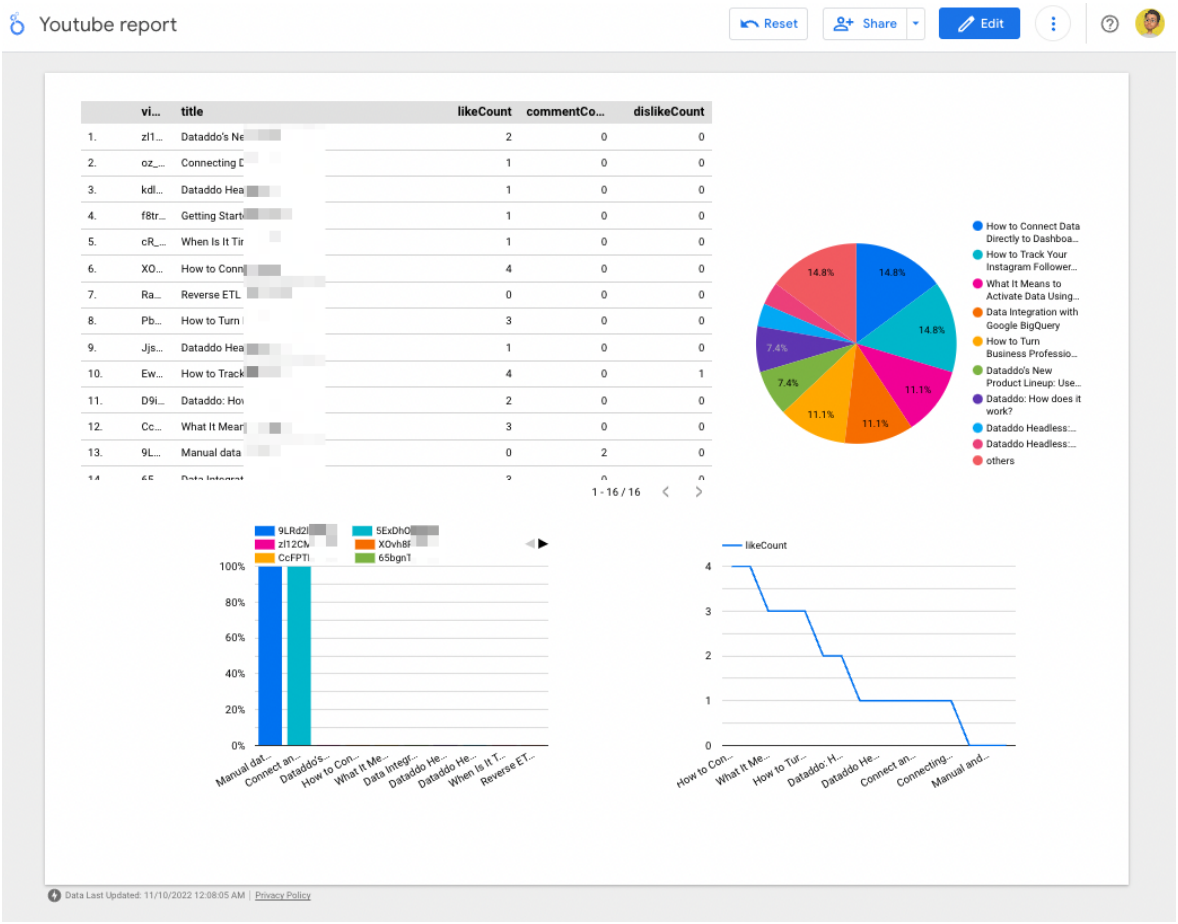


Figure 23: Google Data Studio report

Note: The report link is available in the Appendix

4.1.1.2. Data Storage: Data Source to Data Warehouse

The second use case resembles the first by having the same steps when creating the Data Source. However, when creating the Data Flow, I use a Data Warehouse solution. For this use case, I created a Google BigQuery account, and connect it with Dataddo. Once on the flow creation, I can specify the Table name that will be created in my Database.

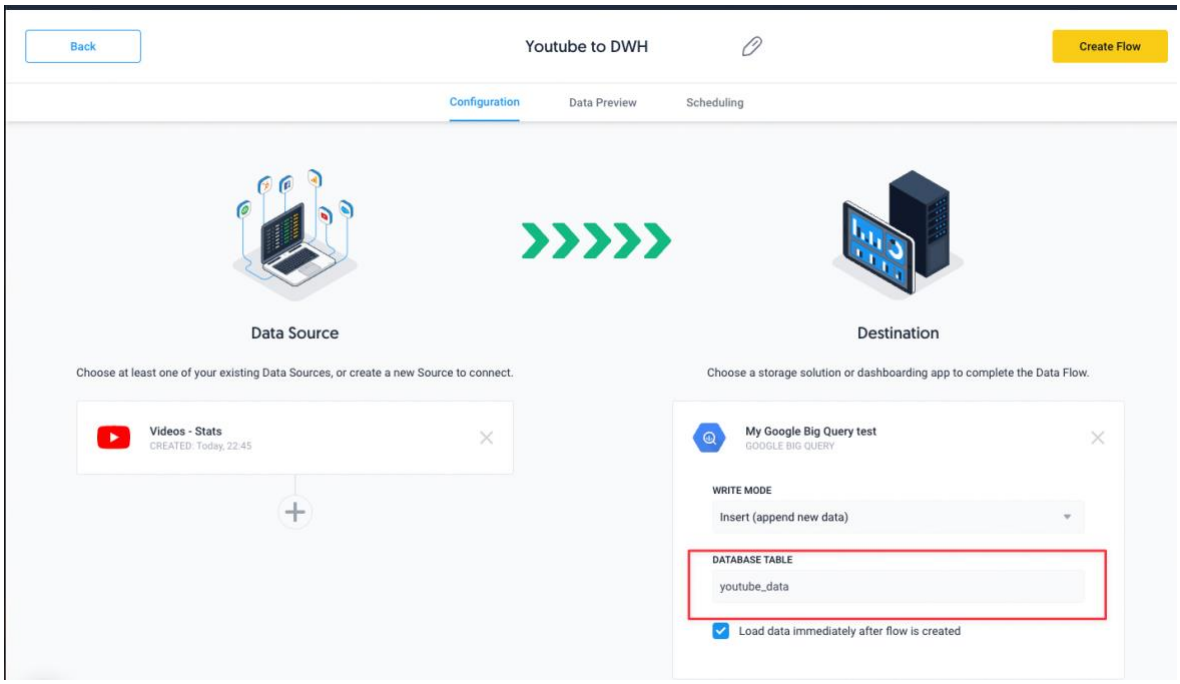
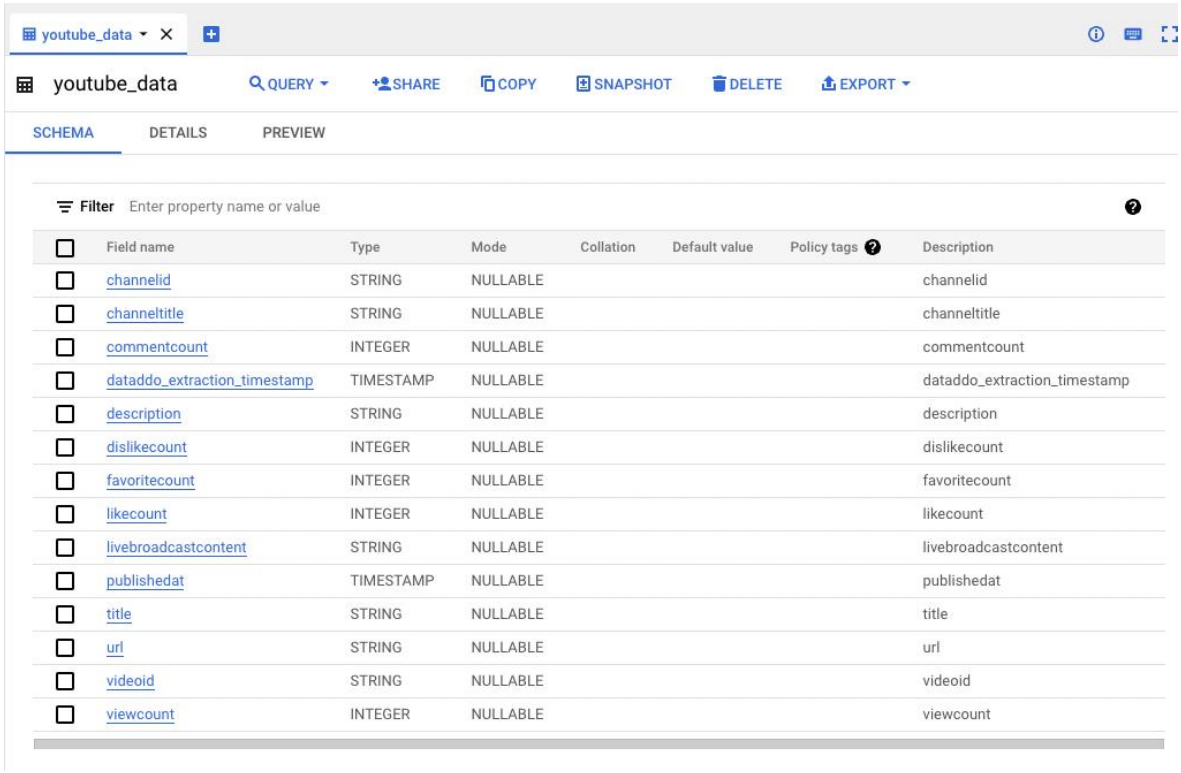


Figure 24: Dataddo Data Warehouse Table Creation

After creating the flow, the data should be synced automatically, and the data should appear in my Database. Once the table is created, they can see the schema of the table



Field name	Type	Mode	Collation	Default value	Policy tags	Description
channelid	STRING	NULLABLE				channelid
channeltitle	STRING	NULLABLE				channeltitle
commentcount	INTEGER	NULLABLE				commentcount
dataddo_extraction_timestamp	TIMESTAMP	NULLABLE				dataddo_extraction_timestamp
description	STRING	NULLABLE				description
dislikecount	INTEGER	NULLABLE				dislikecount
favoritecount	INTEGER	NULLABLE				favoritecount
likecount	INTEGER	NULLABLE				likecount
livebroadcastcontent	STRING	NULLABLE				livebroadcastcontent
publishedat	TIMESTAMP	NULLABLE				publishedat
title	STRING	NULLABLE				title
url	STRING	NULLABLE				url
videoid	STRING	NULLABLE				videoid
viewcount	INTEGER	NULLABLE				viewcount

Figure 25: YouTube Data Schema in Google BigQuery

I can now run SQL queries to construct the table view as I want. With new data being imported from the automated Data Pipeline

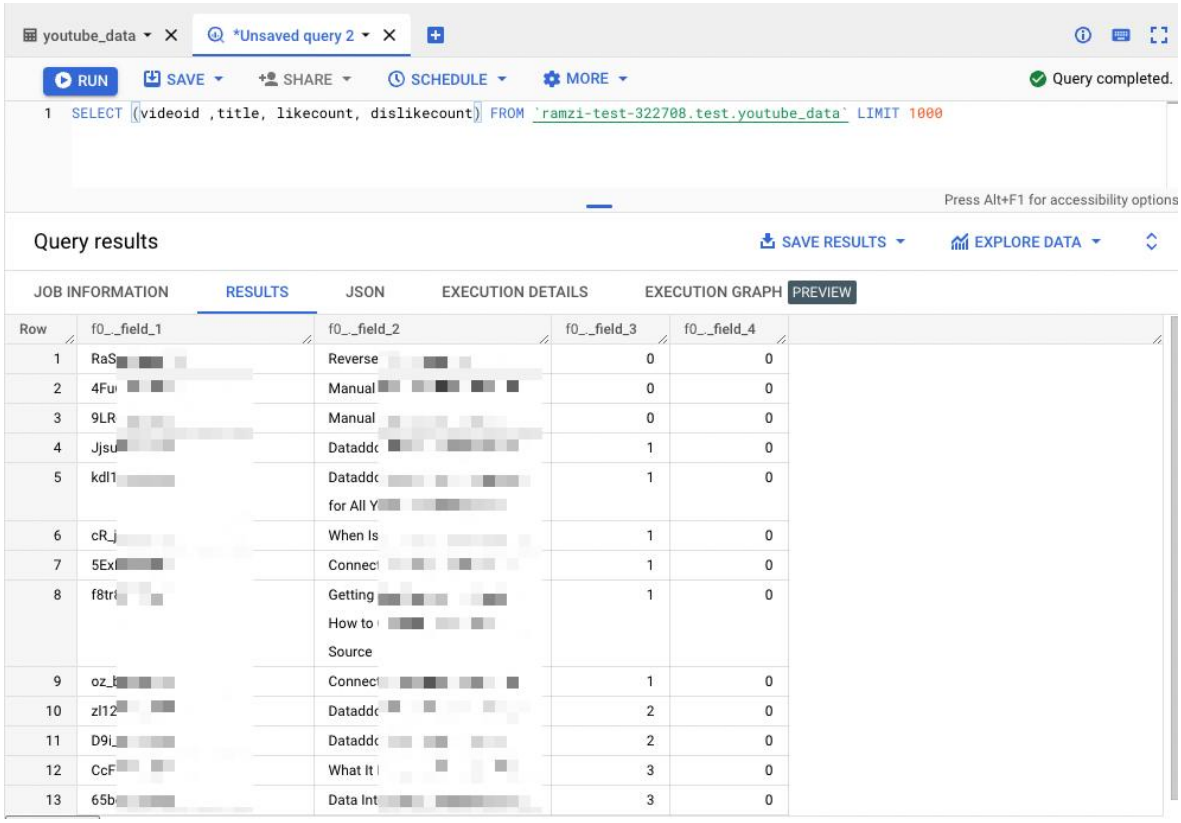


Figure 26: YouTube SQL query in Google BigQuery

The following diagram illustrates the full process in detail

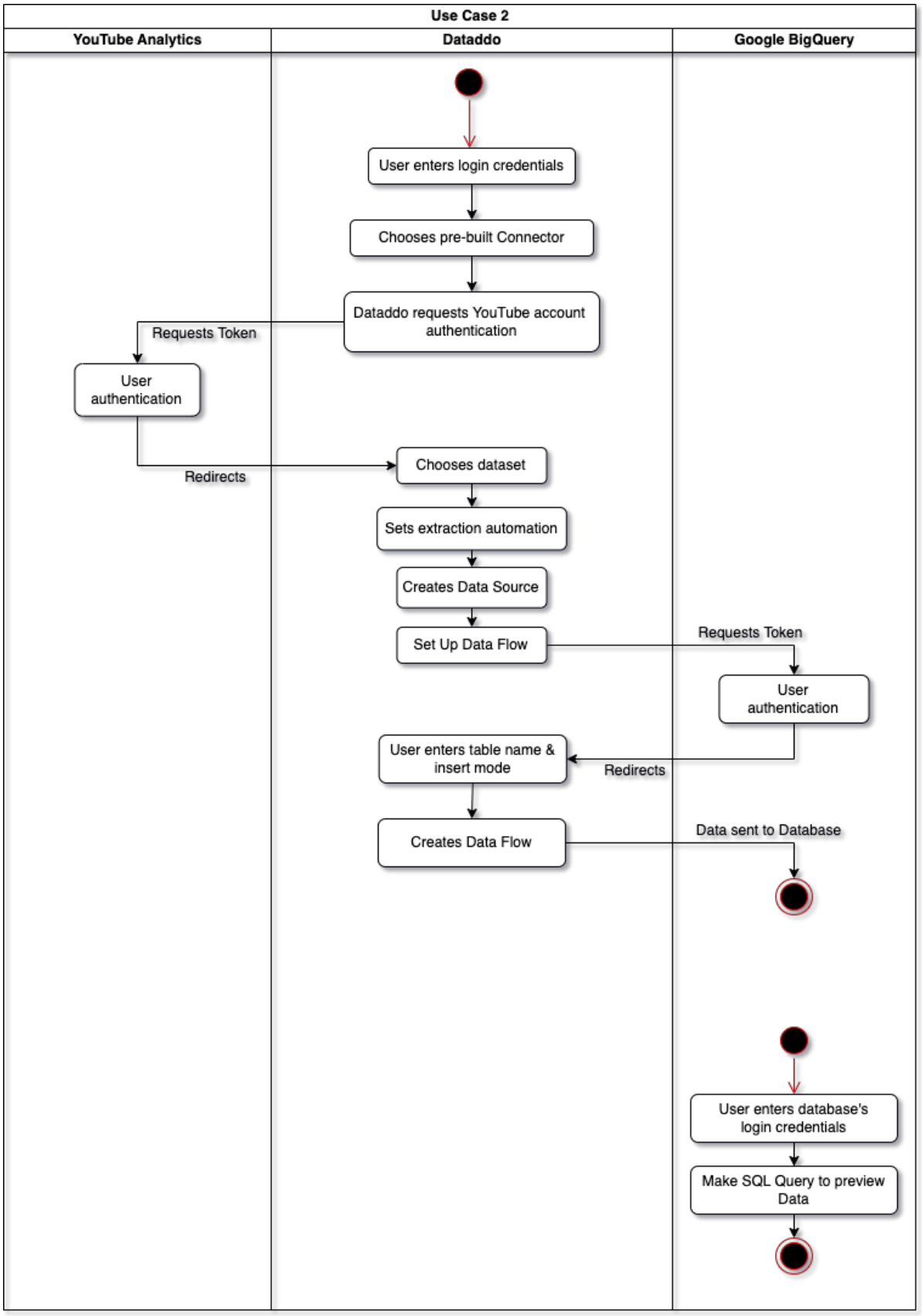


Figure 27: Use case 2: Activity Diagram

4.1.1.3. Data Blending: Multiple Sources, One Destination

For this use case. I make two different YouTube Analytics Data Sources. Both Data Sources are extracted from two different API endpoints, with different datasets. The following figure illustrates the process of this use case.

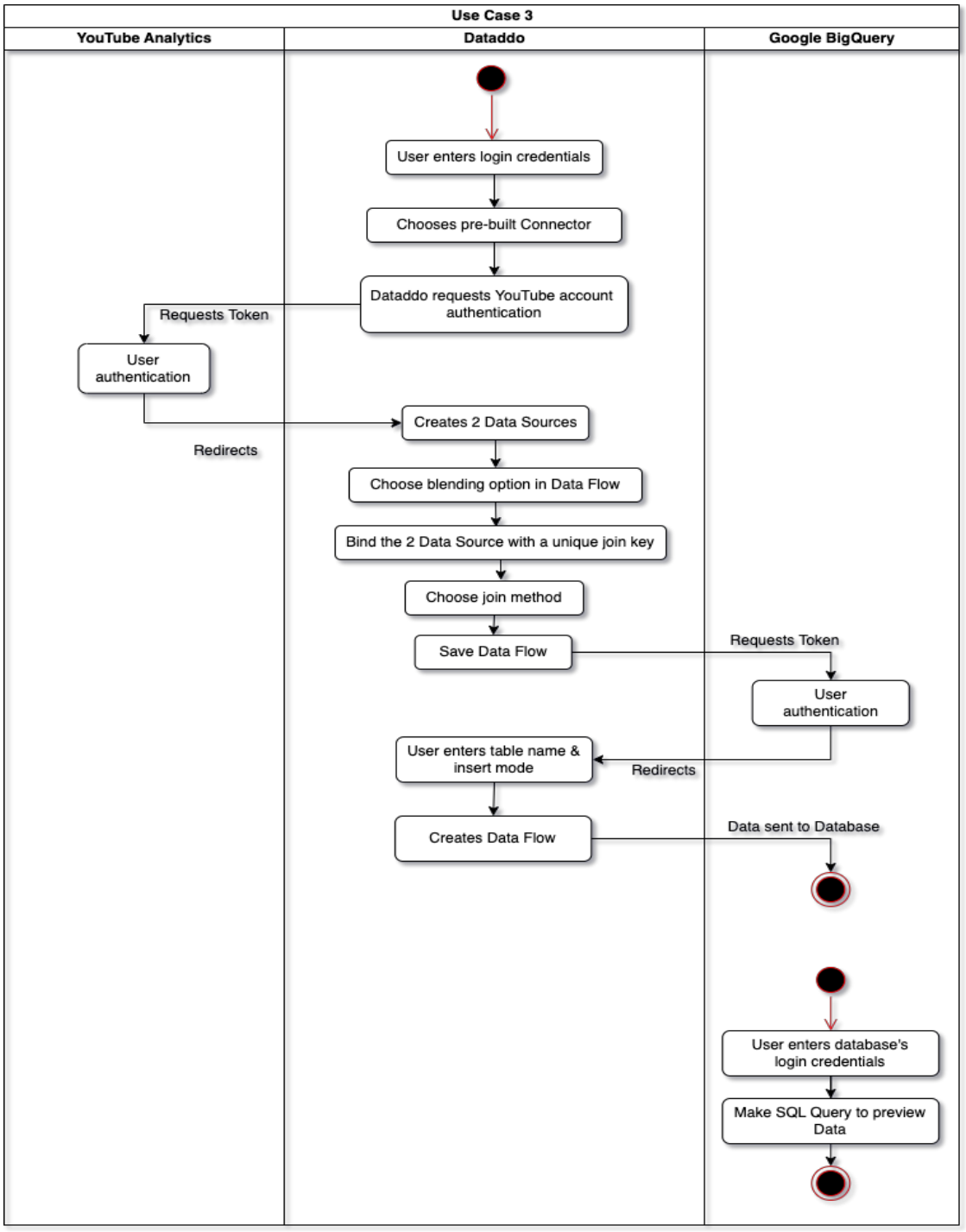


Figure 28: Use case 3: Activity Diagram

One source contains my channel's video performance statistics (e.g., total number like, dislikes, comments, and video details), and another source has the sharing statistics of the videos, but no video details (just the Video ID), thus making it hard to identify the video details when doing my analysis at the destination. Therefore, this is a perfect example to use Data Blending.

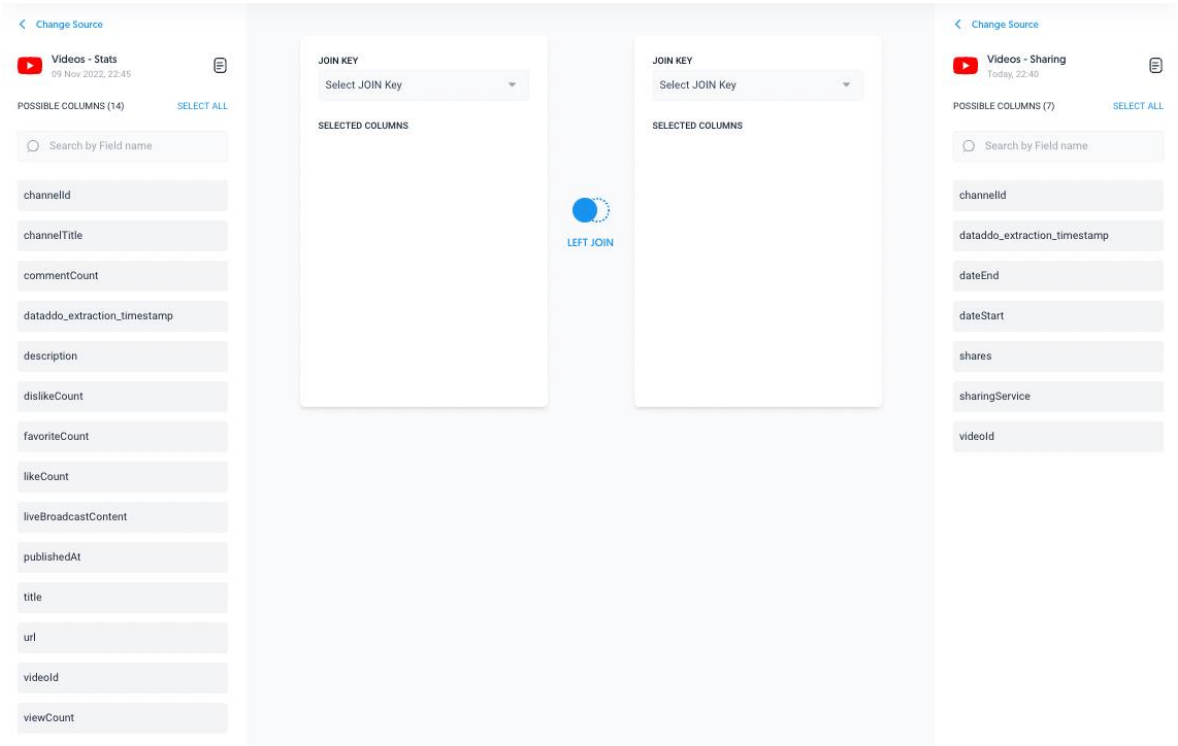


Figure 29: Blending Two Data Sources

With the help of Data Blending, helps me in creating a personalised Data Source that has information about my video's details, in addition to the likes, dislikes and comments numbers, and the total numbers of sharing and the sharing method/

Data Blending helps in joining the table permanently, rather than having to query JOIN SQL commands in any database, or even when you do not have the option to do JOIN commands in the destination, like most Data Virtualisation applications. Therefore, to join the two tables together, I need a unique field that is common between the two data sources, and that is the **videoId**

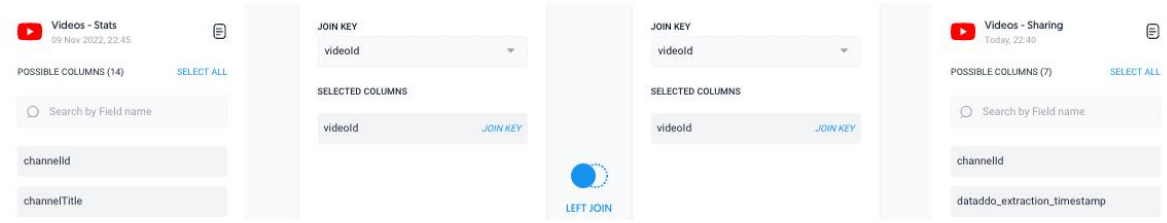


Figure 30: Data Blending JOIN key

Data Blending sources can be blended with Dataddo in two ways, **Left Join** or **Inner Join.**

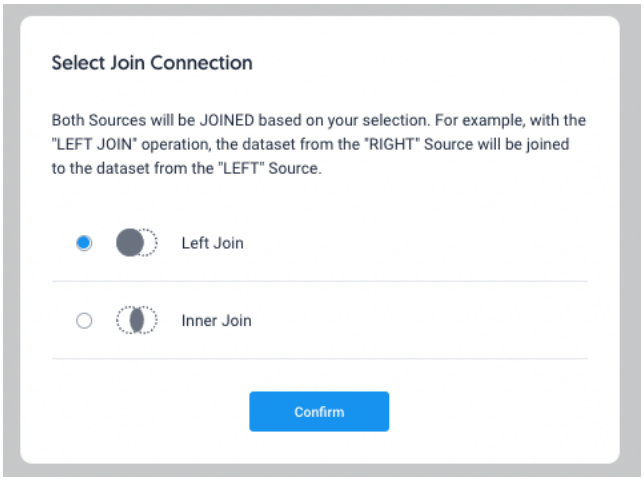


Figure 31: Data Blending JOIN types

I am using Left Join for this use case. After that, I can select the fields I want to have from both sources to be constructed as a unique and personalised Data Source.

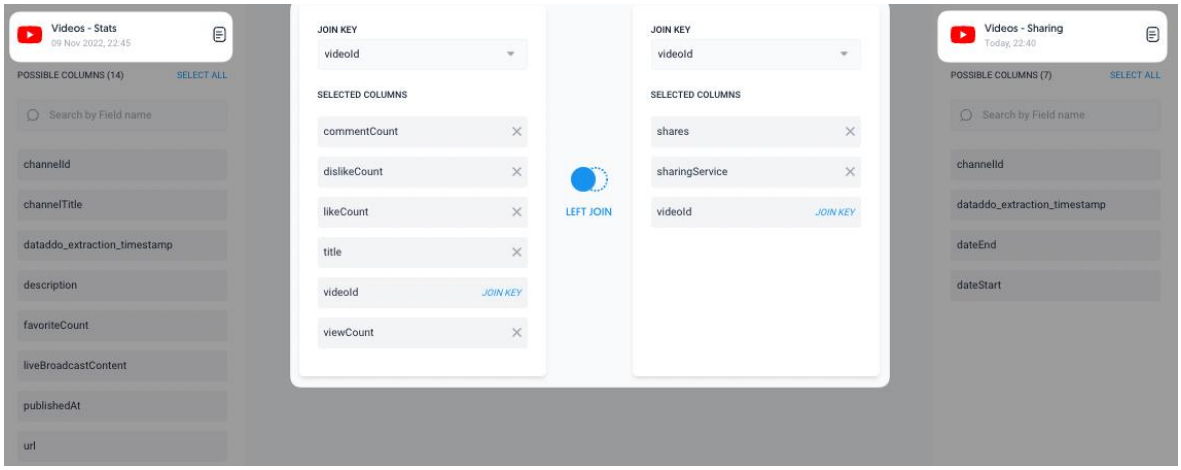


Figure 32: Select Field for Data Blending

Once the fields are selected, the new Data Blended Source should be created.

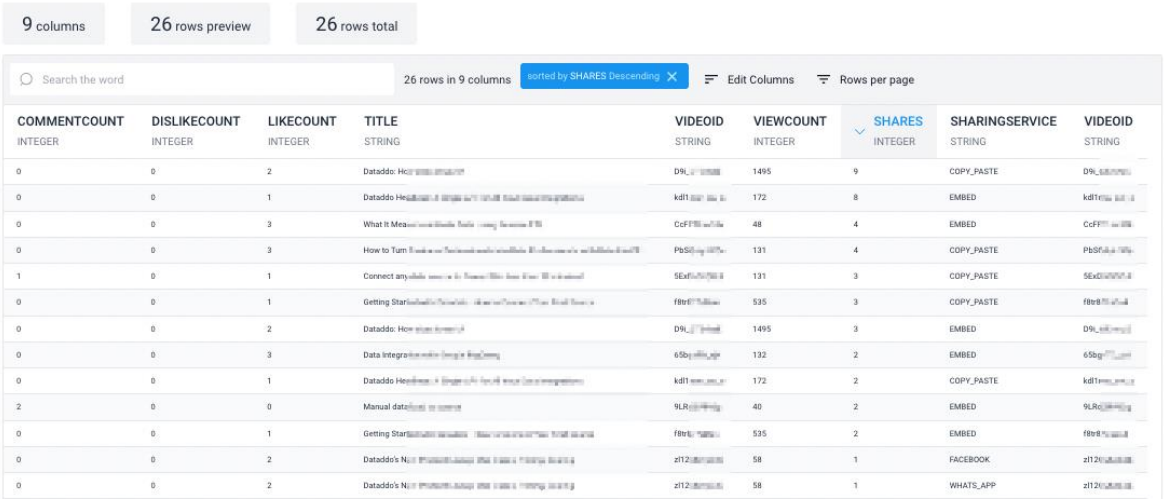


Figure 33: Blended Data Source

Now I can configure the Data Flow to send the data to Google BigQuery

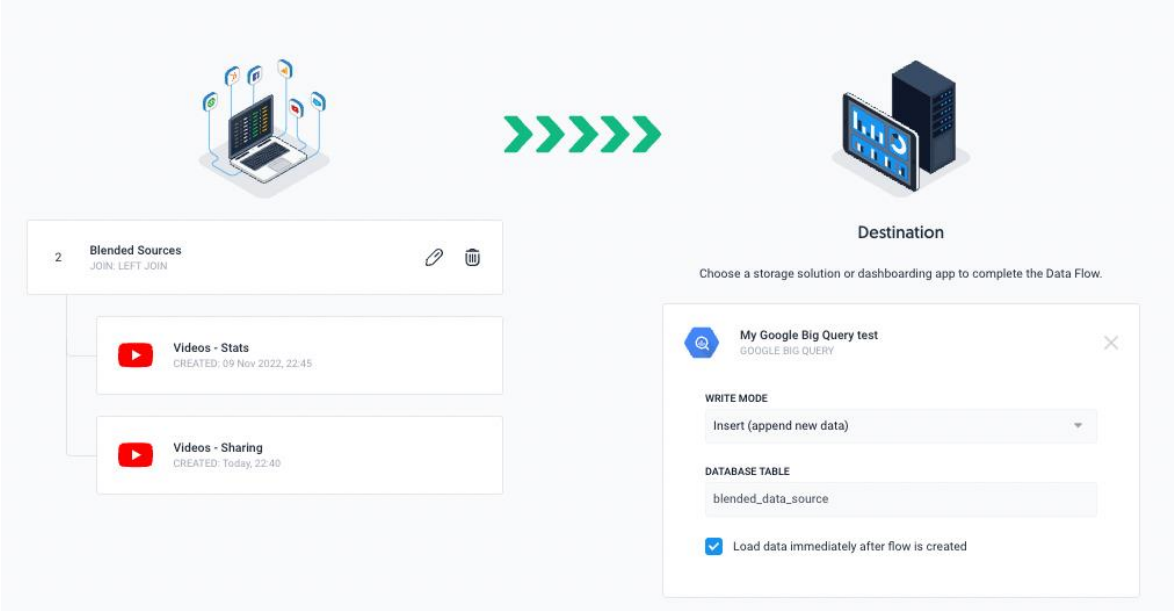
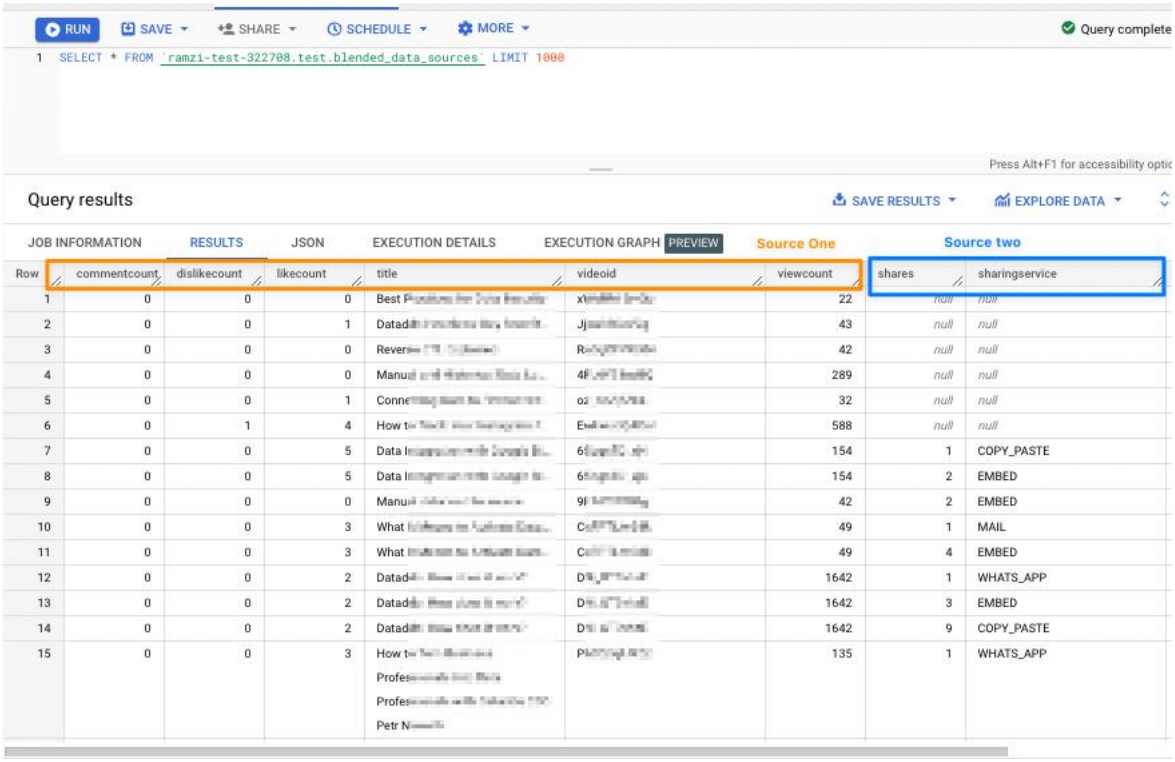


Figure 34: Blended Data Sources to Google BigQuery

Once the Data Flow is created, data should be migrated to the database, the data should be just displayed with a simple “SELECT *”, rather than querying a JOIN SQL command each time we want the join report from two different sources or table, and that is the beauty of automating such tasks.



RUN SAVE SHARE SCHEDULE MORE Query complete

```
1  SELECT * FROM `ramzi-test-322708.test.blended_data_sources` LIMIT 1000
```

Press Alt+F1 for accessibility optic

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW | Source One | Source two

Row	commentcount	dislikecount	likecount	title	videoid	viewcount	shares	sharingservice
1	0	0	0	Best	x	22	null	null
2	0	0	1	Datad	J	43	null	null
3	0	0	0	Reverse	R	42	null	null
4	0	0	0	Manual	4F	289	null	null
5	0	0	1	Conne	o	32	null	null
6	0	1	4	How to	E	588	null	null
7	0	0	5	Data	6	154	1	COPY_PASTE
8	0	0	5	Data	6	154	2	EMBED
9	0	0	0	Manual	9I	42	2	EMBED
10	0	0	3	What	C	49	1	MAIL
11	0	0	3	What	C	49	4	EMBED
12	0	0	2	Datad	D	1642	1	WHATS_APP
13	0	0	2	Datad	D	1642	3	EMBED
14	0	0	2	Datad	D	1642	9	COPY_PASTE
15	0	0	3	How to Profes Profes Petr N	P	135	1	WHATS_APP

Figure 35: Blended Data Sources Migrated to BigQuery

New data should be automatically extracted at the time of my choosing for both data sources, and blended before the next data transfer to the database daily.

4.1.2. Custom Data Integration (With Coding)

4.1.2.1. Data Migration: From Spreadsheet to a better Data Warehouse

In this use case, I am taking an example of an unprofessional inventory manager who uses Google sheets to track the products that they sell and track the stock of each product and wants to start using a data management tool with enhanced storage and scale it's as your data grows, like Google BigQuery. The following figure describes the process of this use case implementation.

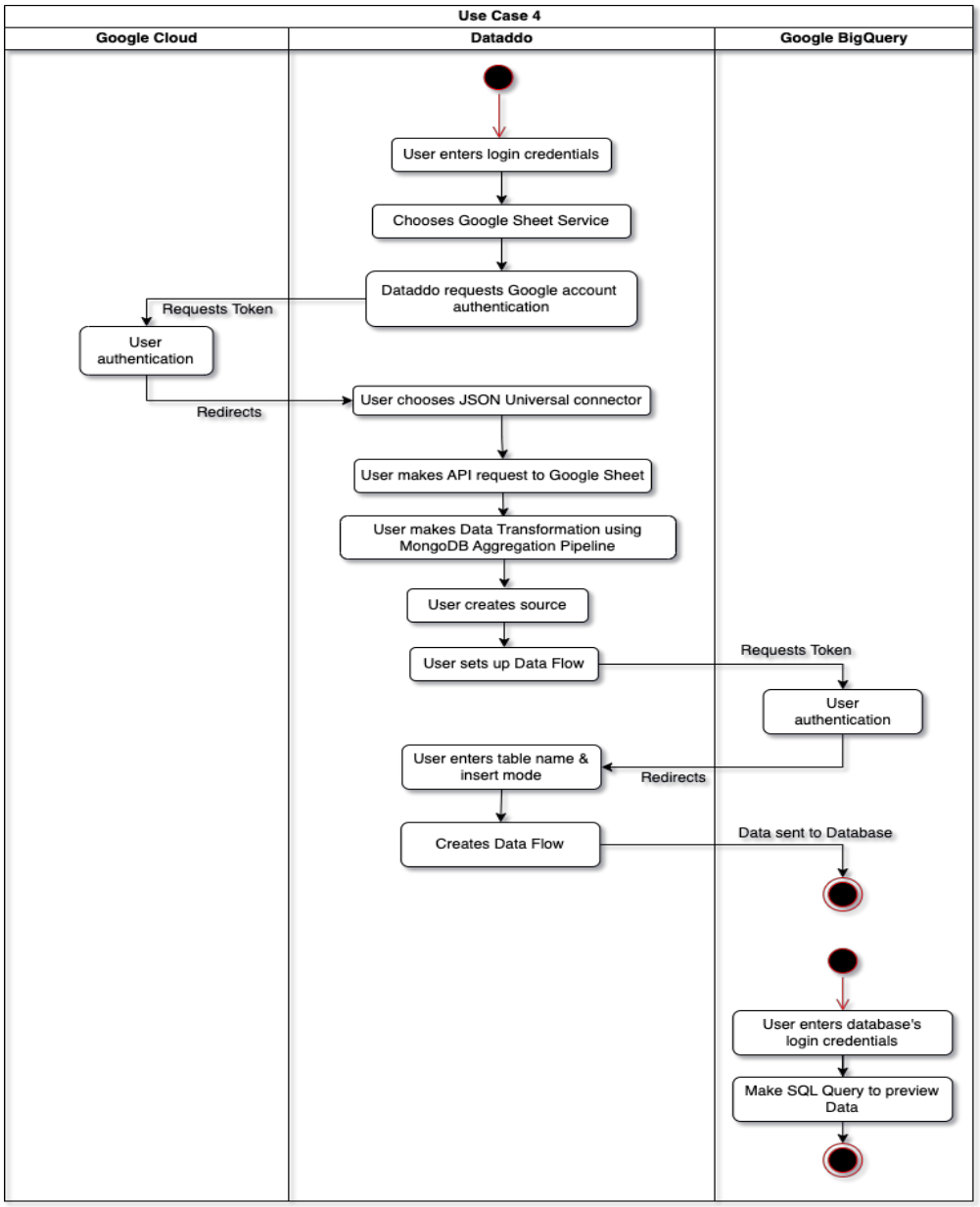


Figure 36: Use case 4: Activity Diagram

Therefore, for this example, I create a sample shop inventory sheet that has information about my products, price, and stock.

description	extraction_date	filename	height	id	price	rating	title	type	width	stock
Baking cake in ri	2022-11-09 23:0	5.jpg	450	5Y9h51elBwPoi	11.14	4	Baking cake	dairy	675	3
Glass of red stra	2022-11-09 23:0	30.jpg	600	9WHhpTNE0CE	25.26	5	Smoothie with cl	fruit	900	98
Homemade tradi	2022-11-09 23:0	28.jpg	450	9sC3eZJrzc7kD	18.5	4	Cuban sandwich	bakery	300	33
Rosemary, lemo	2022-11-09 23:0	9.jpg	450	BlOTGeLJT6R7	15.79	5	Lemon and salt	fruit	299	22
Concept of healt	2022-11-09 23:0	22.jpg	450	Cc4da7aWaDljX	22.7	2	Breakfast with m	dairy	299	54
Vegan sandwich	2022-11-09 23:0	32.jpg	600	FeZGulyl8t8MaF	22.48	5	Sandwich with s	vegetable	398	55
Sweet fresh pea	2022-11-09 23:0	18.jpg	600	FqhzCVO2pL2S	15.12	5	Fresh pears	fruit	398	222
Raw legums on t	2022-11-09 23:0	4.jpg	450	GAMU11wjkw1N	17.11	2	Raw legums	vegetable	299	45
Raw fresh aspar	2022-11-09 23:0	34.jpg	600	HBj5nlsyQt5ELx	22.97	4	Raw asparagus	vegetable	400	867
Cherry with suga	2022-11-09 23:0	33.jpg	600	IdrJPAWz7WrfIC	14.35	5	Cherry	fruit	400	93
Asparagus with l	2022-11-09 23:0	2.jpg	450	Iqh3zMdO8qbe6	18.95	3	Asparagus	vegetable	299	56
Sweet fresh stav	2022-11-09 23:0	8.jpg	600	Izjc2sciceJWGK	28.59	4	Fresh stawberry	fruit	399	209
Raw organic bro	2022-11-09 23:0	0.jpg	600	LUkCudWOEgP	28.1	4	Brown eggs	dairy	400	3
Italian ciabatta b	2022-11-09 23:0	39.jpg	450	LXBV7Y4qDJlpk	15.18	1	Italian ciabatta	bakery	565	5
Raw organic gre	2022-11-09 23:0	14.jpg	450	MDSkh5eCh5np	28.79	1	Green beans	vegetable	300	63
Homemade yogu	2022-11-09 23:0	31.jpg	450	MbuEOSmqb4J	27.61	4	Yogurt	dairy	299	2
Fresh pears juic	2022-11-09 23:0	17.jpg	600	N5dEWfnXHMlj	19.49	4	Pears juice	fruit	398	5
Healthy breakfas	2022-11-09 23:0	37.jpg	450	PrKlitUjXQWHzf	21.01	4	Fresh blueberrie	fruit	321	6
Concept of healt	2022-11-09 23:0	36.jpg	600	QOLT7e3ymcpU	28.96	5	Vegan	vegan	398	7
Sweet fresh stav	2022-11-09 23:0	1.jpg	450	RM1hBdSvckPX	29.45	4	Sweet fresh stav	fruit	299	63
Healthy breakfas	2022-11-09 23:0	13.jpg	450	TYGUEEGlb73e	13.02	2	Healthy breakfas	fruit	350	564
Homemade brea	2022-11-09 23:0	10.jpg	450	UqGJS9klB20aS	17.48	3	Homemade brea	bakery	301	63
Vegan sandwich	2022-11-09 23:0	38.jpg	450	V7FezTqeMHAv	25.88	0	Smashed avoca	fruit	450	4
Rustic healthy bı	2022-11-09 23:0	40.jpg	450	XNweAeyHPX6f	21.32	0	Rustic breakfast	dairy	307	5
Fresh tomato jui	2022-11-09 23:0	12.jpg	600	YaGxe8oYnX6y	16.3	2	Fresh tomato	vegetable	903	3
Concept of vega	2022-11-09 23:0	21.jpg	450	Zlnst3xQXfbGKl	29.66	4	Vegan food	vegetable	299	4
Homemade bak	2022-11-09 23:0	15.jpg	600	c1wrkDTDnKrjC	20.31	1	Baked stuffed pc	bakery	400	63

Figure 37: Mock Google Sheet

Google sheets provide an API service that converts the sheet to JSON format to run queries and assign variables to objects in a list of documents via the API. I use the JSON universal connector of Dataddo to make API requests and code the transformation.

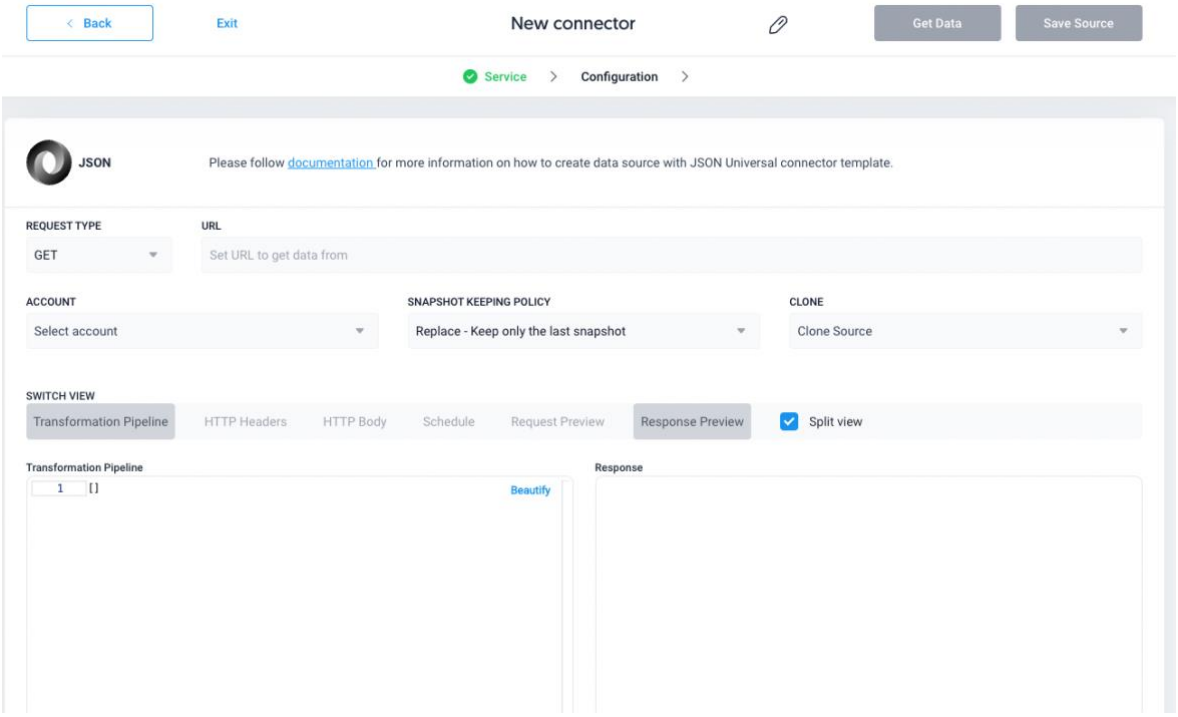
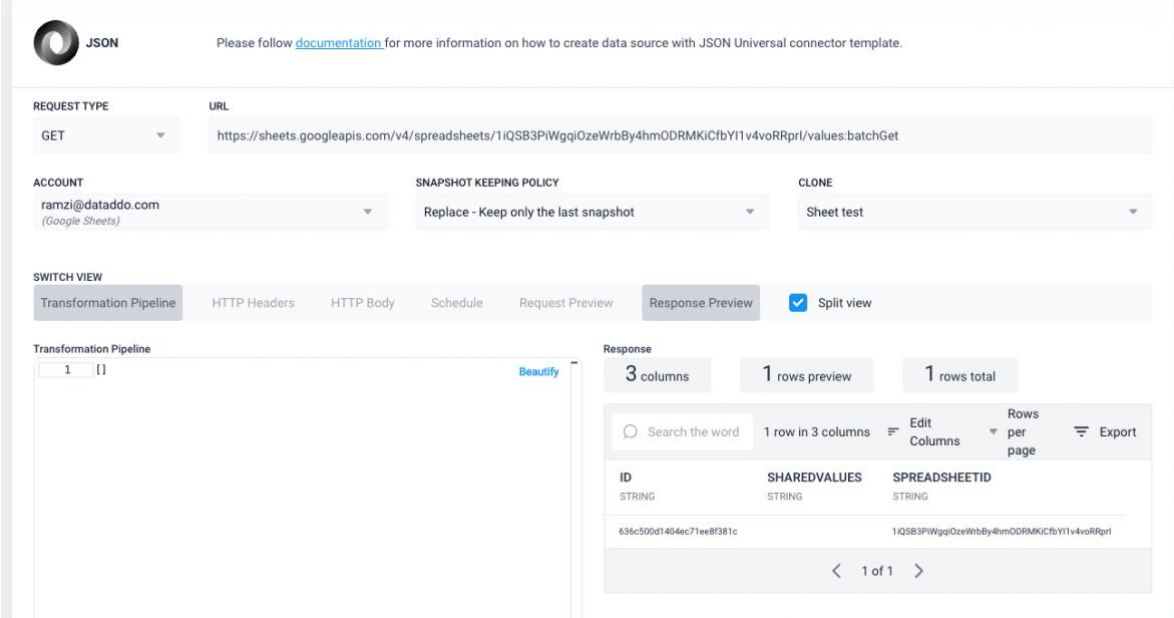


Figure 38: JSON Universal Connector

Example request call:

GET https://sheets.googleapis.com/v4/spreadsheets/{spreadsheetId}/values:batchGet

Using my spreadsheet ID which can be found from the URL of the Google Sheet, I make a GET method request.



Once I have that, I will need to query the flowing parameters to access the correct tab and the range of the table

Parameters	
ranges[]	string The field specifies the sheet name and the range of cells that houses data. Mainly, I need to know how many columns the sheet has, and the row range can be random and not less than the amount of data you have. For my case, I have my table's sheet name as Sheet1 and columns starting from A to K. Thus, my range field will be specified as **ranges=Sheet1!A1:K1000**
majorDimension	**enum (**Dimension**)** The main dimension should be applied to the results. For instance, requesting range=A1:B2,**majorDimension=ROWS** produces [[1,2],[3,4]], whereas requesting range=A1:B2,**majorDimension=COLUMNS** returns [[1,3],[2,4]]. This is because the spreadsheet data is A1=1, B1=2, A2=3, B2=4, for instance. Since my table's data are entered vertically, my major dimension should be specified as Columns

valueRenderOption	**enum (**ValueRenderOption**)** How values should be represented in the output. The default renders option is **UNFORMATTED_VALUE.** I want the data as raw as possible, therefore, the **UNFORMATTED_VALUE** best suits my needs
dateTimeRenderOption	**enum (**DateTimeRenderOption**)** How the output should display dates, times, and durations. If **valueRenderOption** is **FORMATTED_VALUE**, this is disregarded. The default DateTime render option is **[DateTimeRenderOption.SERIAL_NUMBER],** I went with the **FORMATTED_STRING** option

Table 6: List of Parameters for Google Sheet API

My final URL query should be like this

https://sheets.googleapis.com/v4/spreadsheets/**1iQSB3PiWgqiOzeWrbBy4hmODRMKiCfbYI1v4voRRprI**/values:batchGet?valueRenderOption=**UNFORMATTED_VALUE**&ranges=**Sheet1!A1:K1000**&majorDimension=**COLUMNS**&dateTimeRenderOption=**FORMATTED_STRING**

The following is the result.

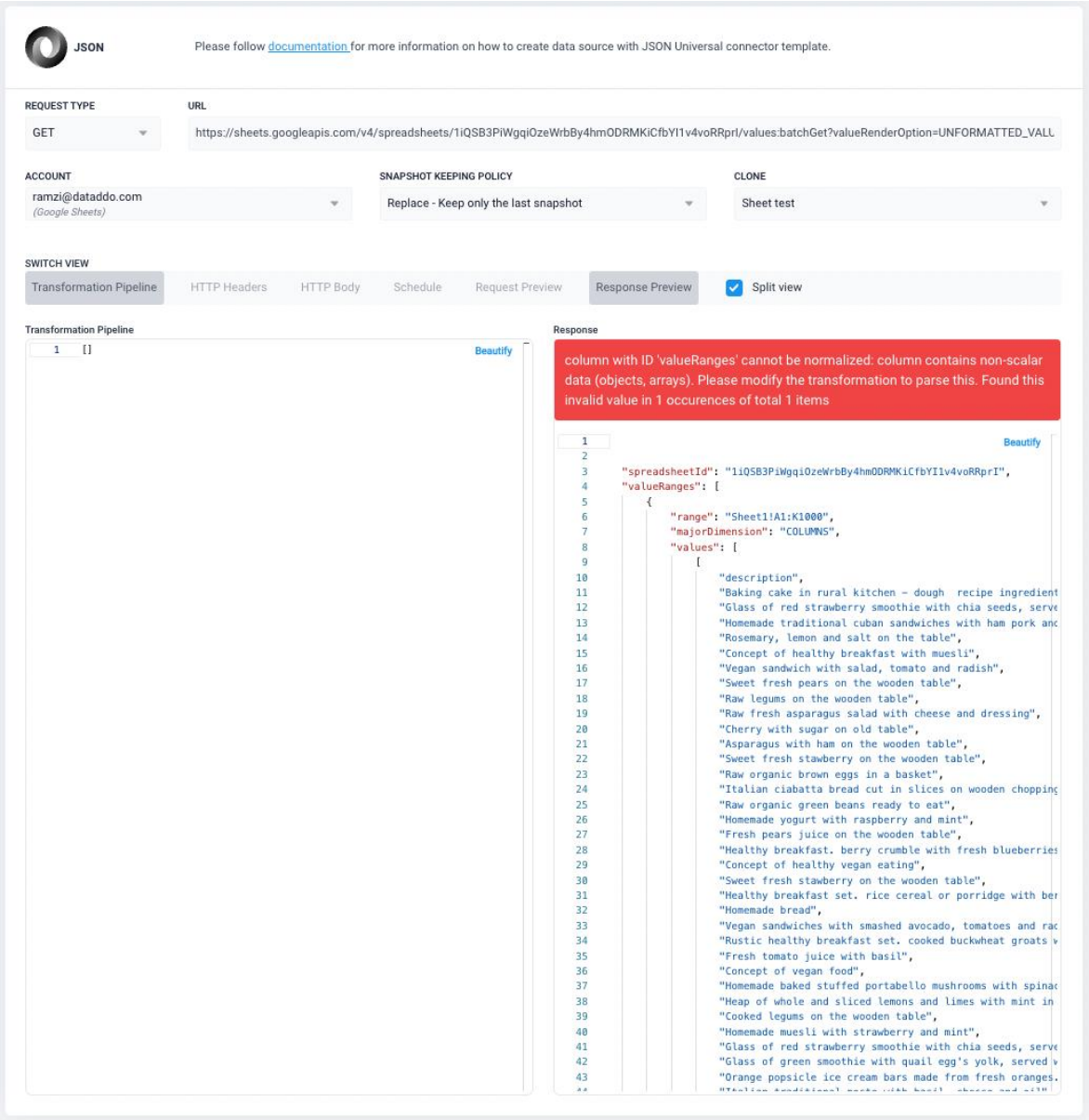


Figure 39: Google Sheet RAW Data

As we see for the API response, that response data object is not structured, as all the data are stored in **values** variable that has the column name then its values respectively. Therefore, I use Mongo Aggregation Pipelines (see section 3.11.1) to help in transforming and manipulating the response data. I need to deconstruct the values array field to have a document output for each element of the value of the array.

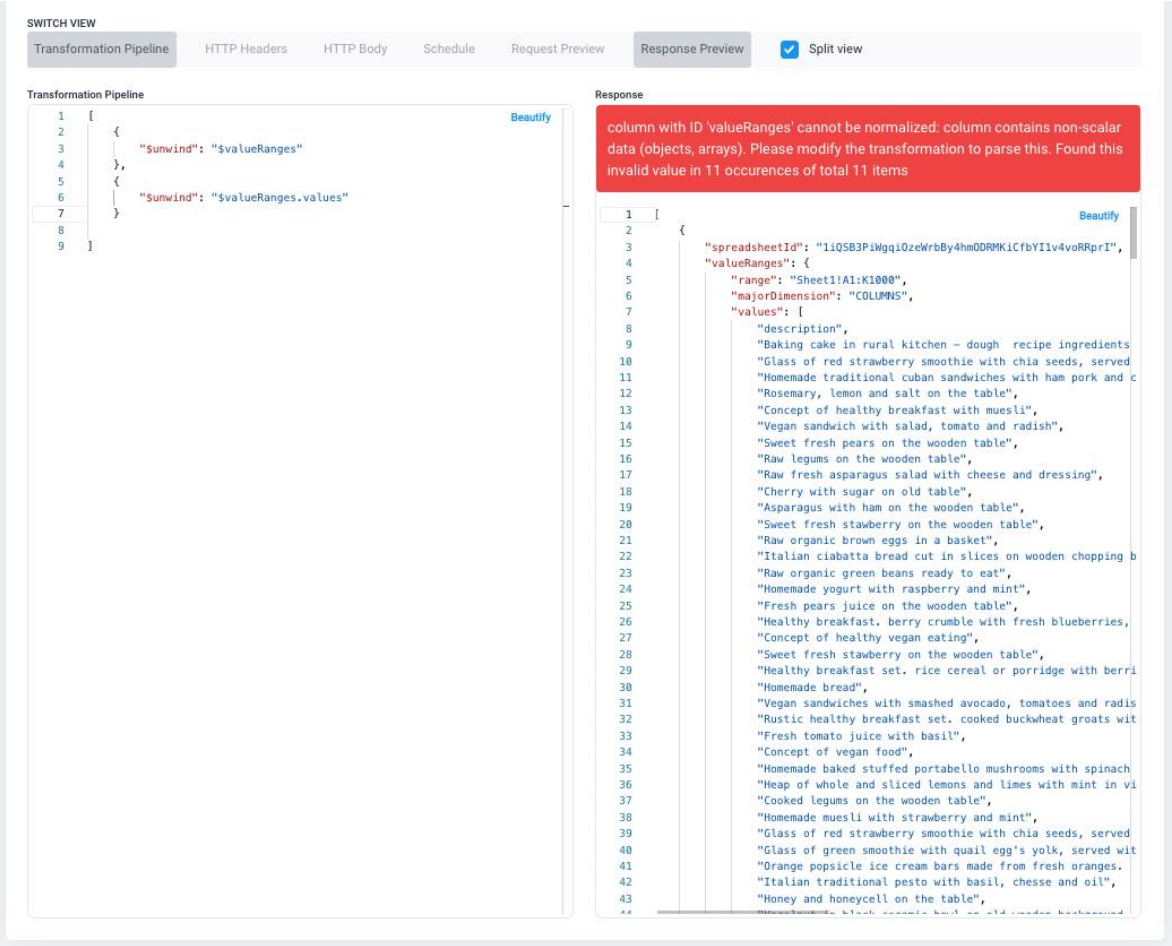


Figure 40: Deconstructing Google Sheet Data

Now, in this step, I need to take the first value from each document for the values field that is shown in the response because they are the columns' index in my google sheet.

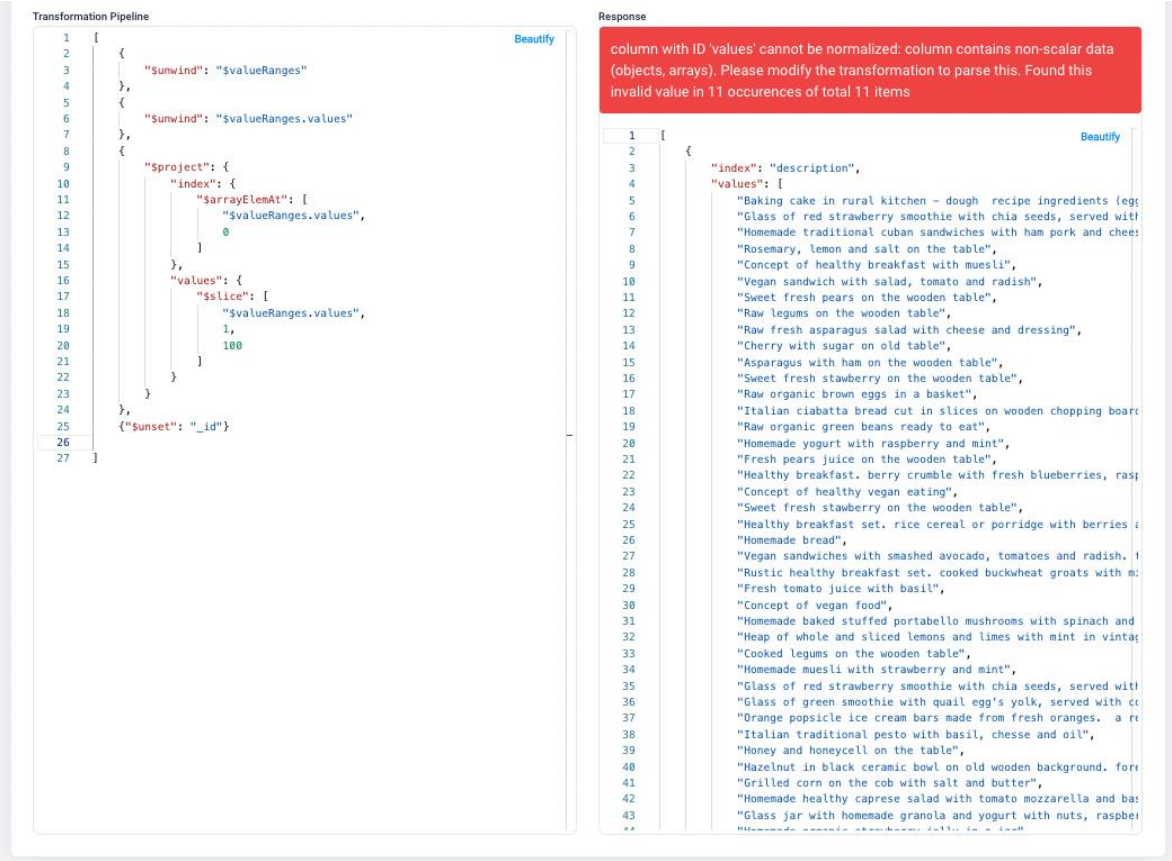


Figure 41: Deconstructing Index from Values

The next step is to join the index with its respective value. I do this by taking the position (or index of the values) in the array. This helps in constructing a proper document object that has proper keys and their respective values. I group the index with each value based on the index or each length of the array. I give the document a parent key, and I name its **products**.

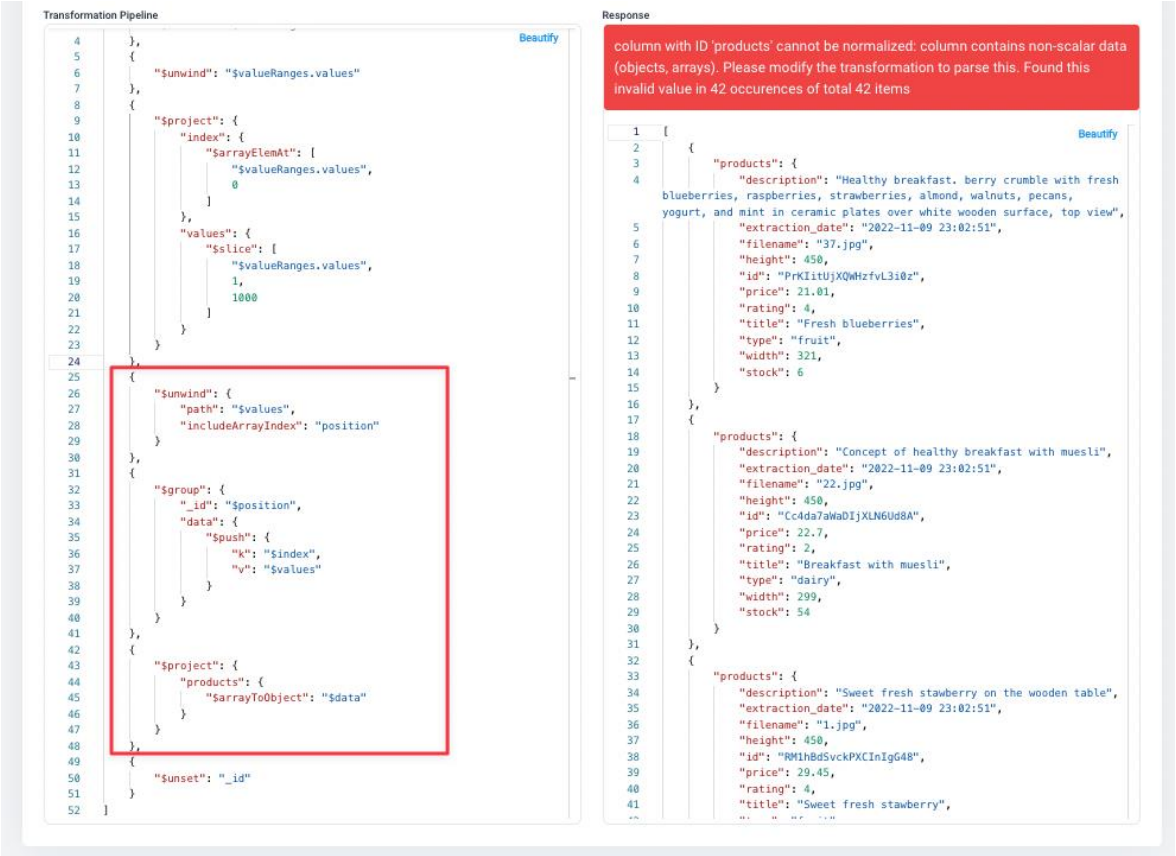


Figure 42: Constructing Proper Array of Objects

The last step in the transformation is to create the table. Using the **$project** operator, I can create the index column of the table and assign its value from the now newly **transformed** API response. The transformation should be saved for future automated extractions. I also added an extraction date field, that will be set every time the automated extraction is executed. This will help to differentiate between new and old data.

I can also set up null values, should any field have a null or undefined value, to have consistency in the flow of the data

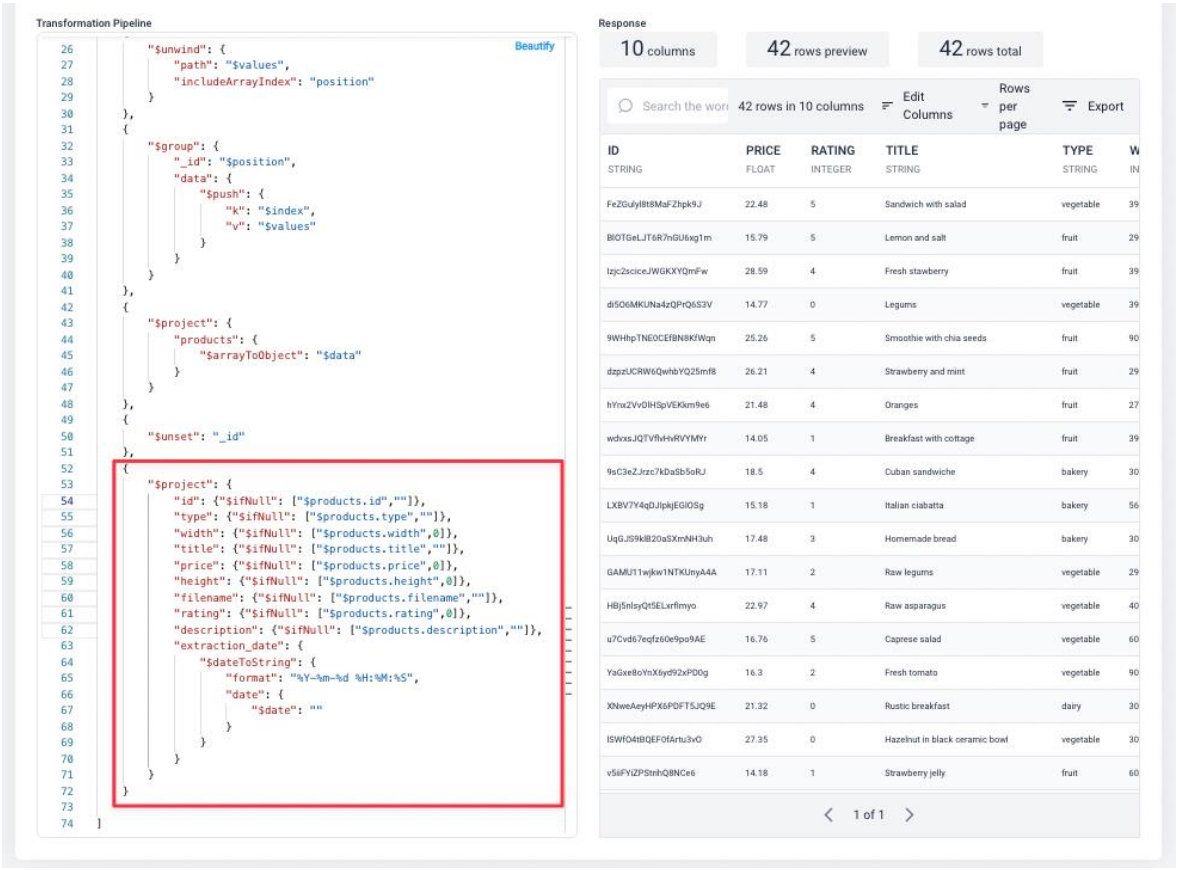


Figure 43: Google Sheet Connector Creation

As the connector is created, now I configure the Data Flow, as I want to send the data to migrate the data to a proper database. I am using Google Data Studio again to store the data on the cloud. I set up the method of writing the data in the database to Truncate Insert. This will clean the old data

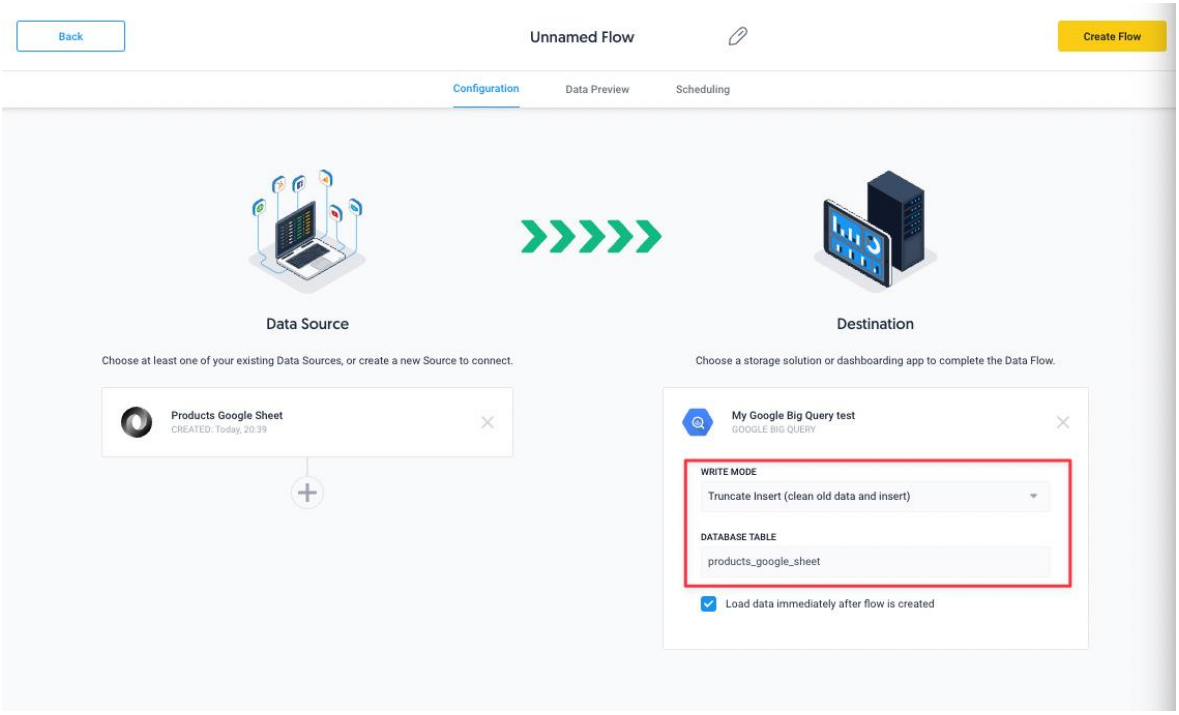
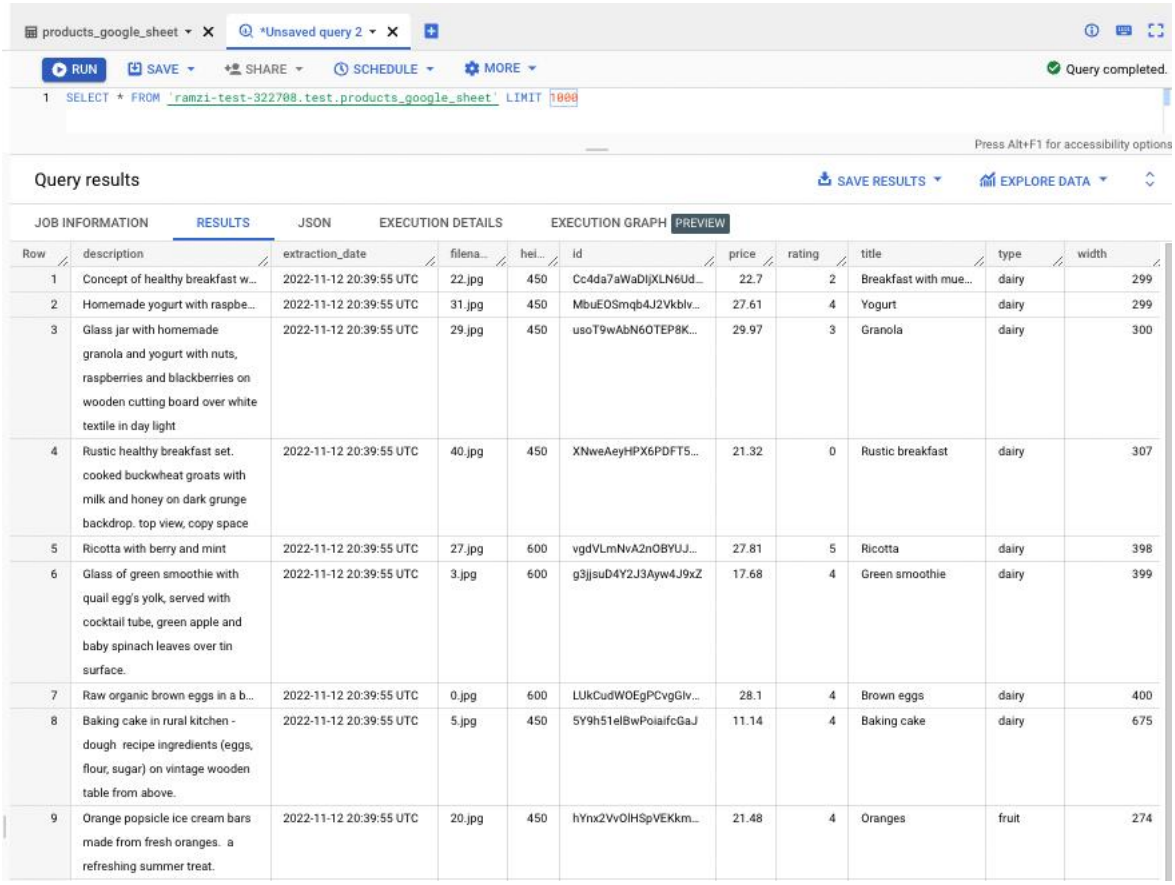


Figure 44: Custom Google Sheets Data Flow

Now the data should be sent to the database and automated for future insertion. The following table should be shown on the SQL query



products_google_sheet | *Unsaved query 2

RUN | SAVE | SHARE | SCHEDULE | MORE | Query completed.

```
1  SELECT * FROM `ramzi-test-322708.test.products_google_sheet` LIMIT 1000
```

Press Alt+F1 for accessibility options

Query results | SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH | PREVIEW

Row	description	extraction_date	filena...	hei...	id	price	rating	title	type	width
1	Concept of healthy breakfast w...	2022-11-12 20:39:55 UTC	22.jpg	450	Cc4da7aWaDljXLN6Ud...	22.7	2	Breakfast with mue...	dairy	299
2	Homemade yogurt with raspbe...	2022-11-12 20:39:55 UTC	31.jpg	450	MbuEOSmqb4J2Vkblv...	27.61	4	Yogurt	dairy	299
3	Glass jar with homemade granola and yogurt with nuts, raspberries and blackberries on wooden cutting board over white textile in day light	2022-11-12 20:39:55 UTC	29.jpg	450	usoT9wAbN6OTEP8K...	29.97	3	Granola	dairy	300
4	Rustic healthy breakfast set. cooked buckwheat groats with milk and honey on dark grunge backdrop. top view, copy space	2022-11-12 20:39:55 UTC	40.jpg	450	XNweAeyHPX6PDFT5...	21.32	0	Rustic breakfast	dairy	307
5	Ricotta with berry and mint	2022-11-12 20:39:55 UTC	27.jpg	600	vgdVLmNvA2nOBYUJ...	27.81	5	Ricotta	dairy	398
6	Glass of green smoothie with quail egg's yolk, served with cocktail tube, green apple and baby spinach leaves over tin surface.	2022-11-12 20:39:55 UTC	3.jpg	600	g3jjsuD4Y2J3Ayw4J9xZ	17.68	4	Green smoothie	dairy	399
7	Raw organic brown eggs in a b...	2022-11-12 20:39:55 UTC	0.jpg	600	LUkCudWOEgPCvgGlv...	28.1	4	Brown eggs	dairy	400
8	Baking cake in rural kitchen - dough recipe ingredients (eggs, flour, sugar) on vintage wooden table from above.	2022-11-12 20:39:55 UTC	5.jpg	450	5Y9h51elBwPoiaifcGaJ	11.14	4	Baking cake	dairy	675
9	Orange popsicle ice cream bars made from fresh oranges. a refreshing summer treat.	2022-11-12 20:39:55 UTC	20.jpg	450	hYnx2VvOlHSpVEKkm...	21.48	4	Oranges	fruit	274

Figure 45: Google Sheet Data Migrated to Google BigQuery

This use case is also applicable for migrating data from OODBs to RDBMS, and vice versa. A good example would be migrating collections from MongoDB to MySQL tables.

4.1.2.2. Public Data: API to Data Virtualisation

For this example, let us simulate the role of a medical practitioner who is tasked to make a report of the Covid Hospitalisation cases for the last 30 days in Germany (see section 3.11.4), and to have new data each day added to the report. The following is the activity flow for this use case

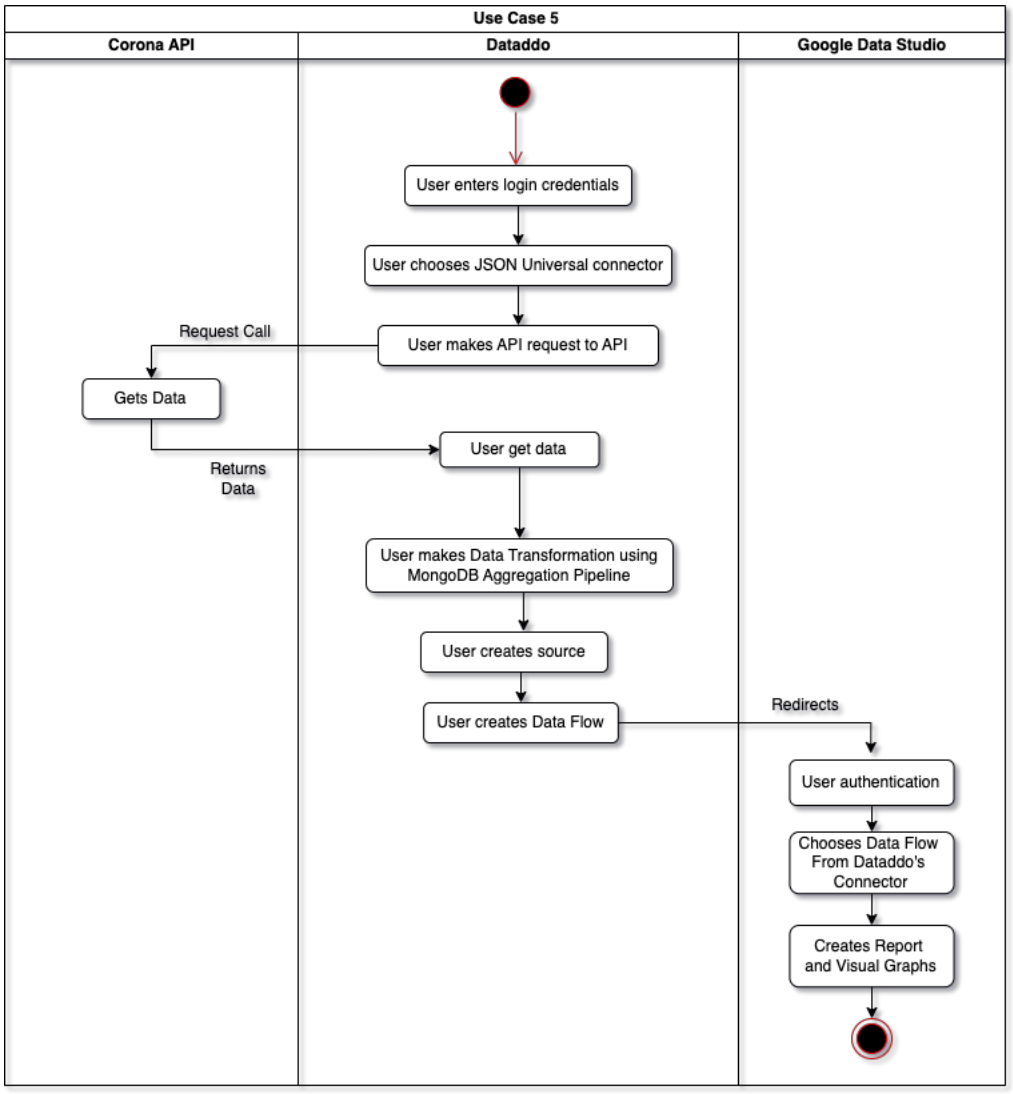


Figure 46: Use case 5: Activity Diagram

Therefore, for this use case, I found a source that has a public API with data on the Covid hospitalisation rates, with historical data. I start by creating a custom JSON Universal Data Source, just as I did with the Google Sheet use case. I use the following API service, and I query the historical data to be **30 days in the** URL parameters (as mentioned in their documentation)

https://api.corona-zahlen.org/germany/**history**/hospitalization/**30**

The following is the result of the RAW data.

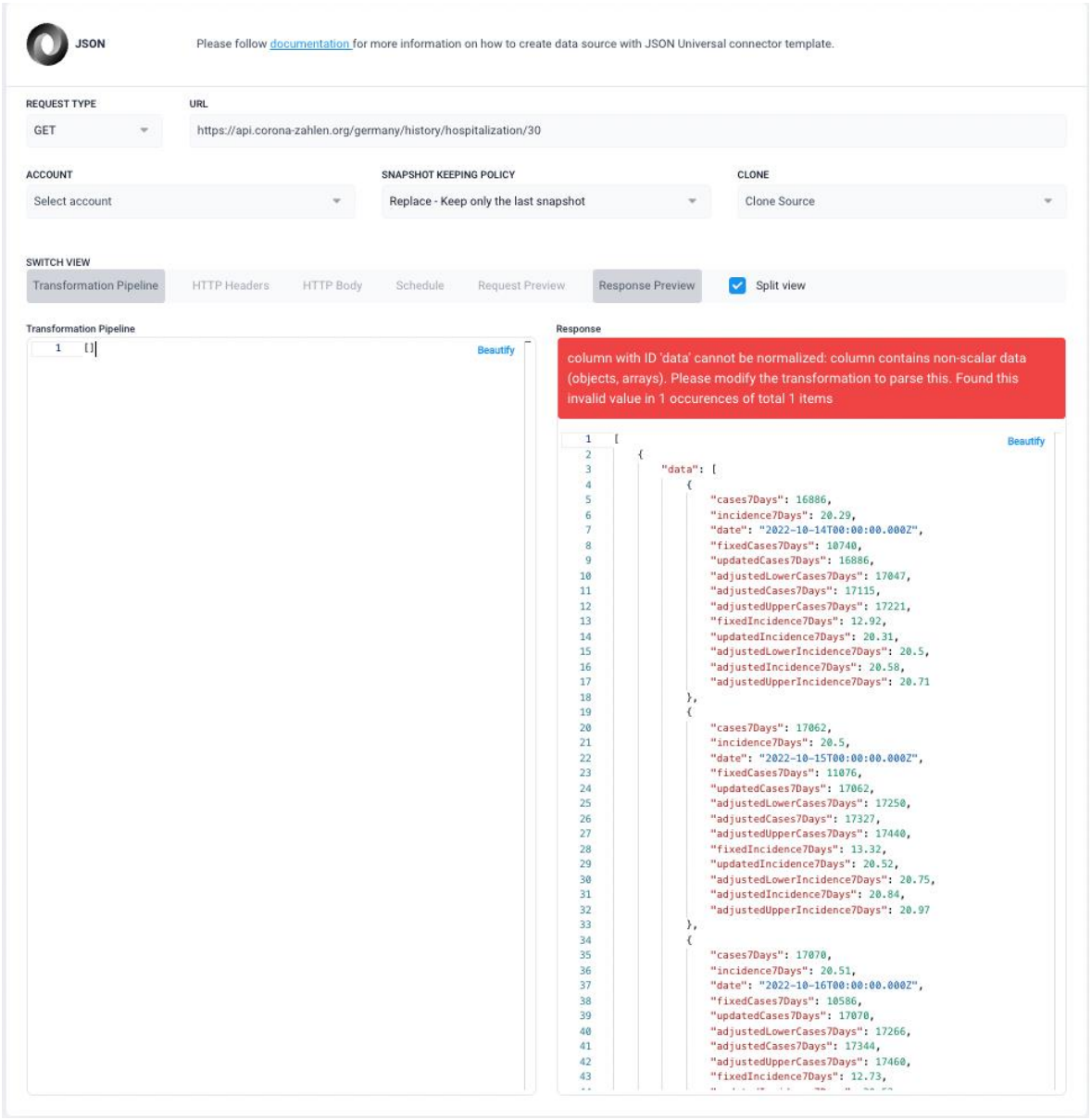


Figure 47: Public API Raw Response

Here, as we see the data is almost well structured, I need to do a small transformation in the pipeline to create the Data Source. First, I am breaking the root array of the response to make the array values to document structure using the **$unwind** operator, then construct the table again by using the **$project** operator

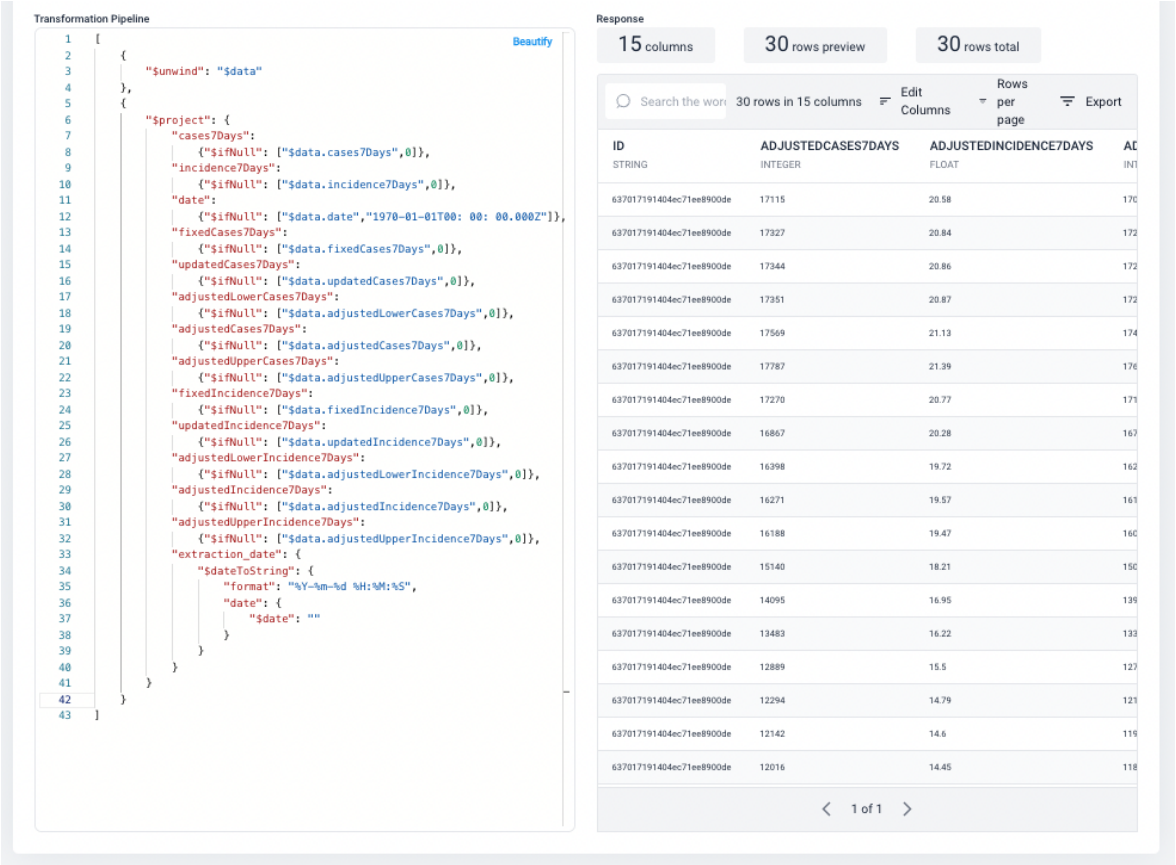


Figure 48: Public API Data Transformation

Now the Data Source can be created and the Data Flow can be set up to Google Data Studio as the final destination.

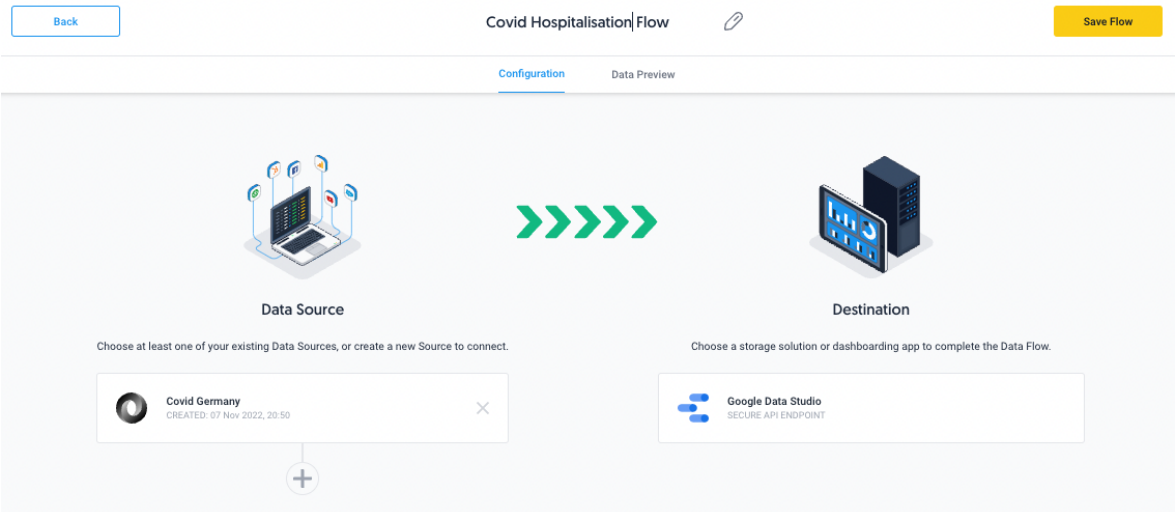
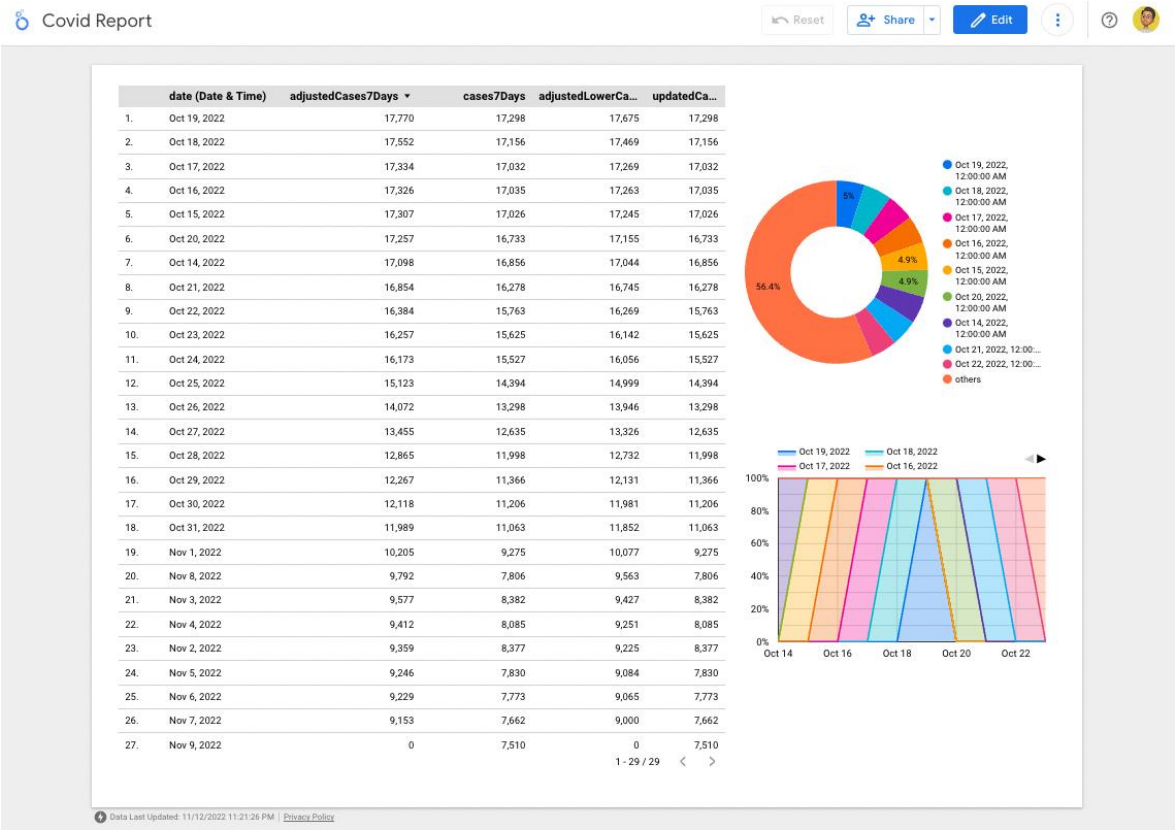


Figure 49: Public API Data Flow Configuration

Now reports charts and tables can be created to virtualise the covid cases based on the metrics the data management needs.



	date (Date & Time)	adjustedCases7Days	cases7Days	adjustedLowerCa...	updatedCa...
1.	Oct 19, 2022	17,770	17,298	17,675	17,298
2.	Oct 18, 2022	17,552	17,156	17,469	17,156
3.	Oct 17, 2022	17,334	17,032	17,269	17,032
4.	Oct 16, 2022	17,326	17,035	17,263	17,035
5.	Oct 15, 2022	17,307	17,026	17,245	17,026
6.	Oct 20, 2022	17,257	16,733	17,155	16,733
7.	Oct 14, 2022	17,098	16,856	17,044	16,856
8.	Oct 21, 2022	16,854	16,278	16,745	16,278
9.	Oct 22, 2022	16,384	15,763	16,269	15,763
10.	Oct 23, 2022	16,257	15,625	16,142	15,625
11.	Oct 24, 2022	16,173	15,527	16,056	15,527
12.	Oct 25, 2022	15,123	14,394	14,999	14,394
13.	Oct 26, 2022	14,072	13,298	13,946	13,298
14.	Oct 27, 2022	13,455	12,635	13,326	12,635
15.	Oct 28, 2022	12,865	11,998	12,732	11,998
16.	Oct 29, 2022	12,267	11,366	12,131	11,366
17.	Oct 30, 2022	12,118	11,206	11,981	11,206
18.	Oct 31, 2022	11,989	11,063	11,852	11,063
19.	Nov 1, 2022	10,205	9,275	10,077	9,275
20.	Nov 8, 2022	9,792	7,806	9,563	7,806
21.	Nov 3, 2022	9,577	8,382	9,427	8,382
22.	Nov 4, 2022	9,412	8,085	9,251	8,085
23.	Nov 2, 2022	9,359	8,377	9,225	8,377
24.	Nov 5, 2022	9,246	7,830	9,084	7,830
25.	Nov 6, 2022	9,229	7,773	9,065	7,773
26.	Nov 7, 2022	9,153	7,662	9,000	7,662
27.	Nov 9, 2022	0	7,510	0	7,510

1 - 29 / 29

Figure 50: Public Data Virtualised

Once the last thing that needs to be done is to set the automation of the Data Source to extract data from the previous day's record daily. Therefore, I modify the URL parameters from the settings of the Data Source from 30 to 1 (as in yesterday's data)

Note: The report link is available in the Appendix

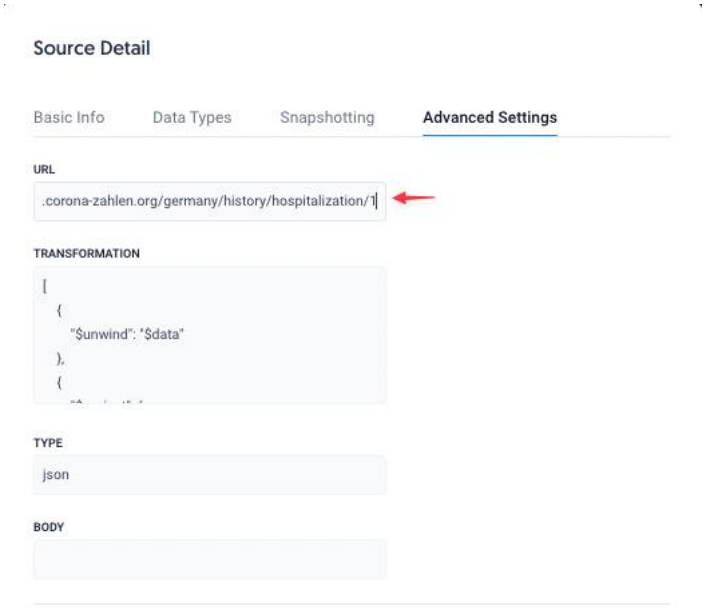


Figure 51: Public API Data Source Settings

Also, I make sure the snapshotting settings are daily at a specific time, to get the data daily at the time of my needs

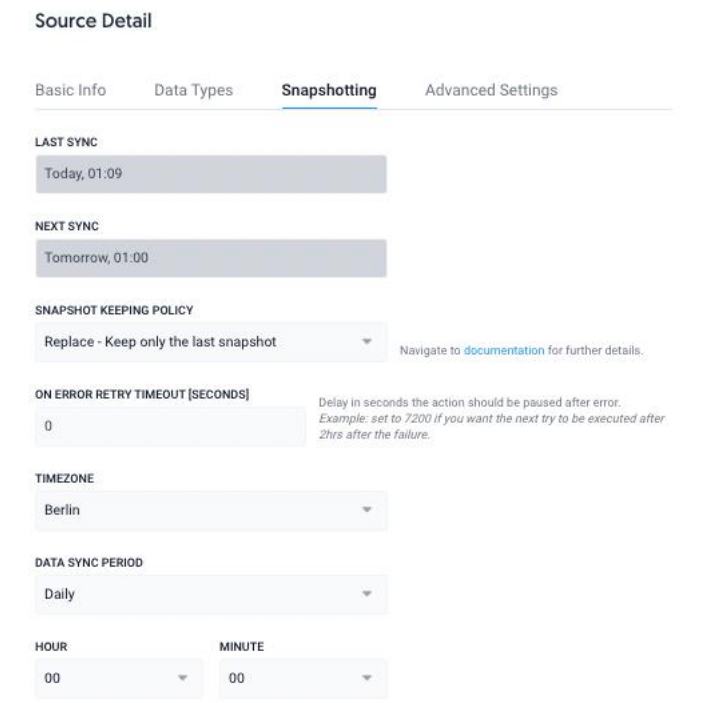


Figure 52: Public API Data Source Snapshotting Settings

4.1.2.3. Data from Private API to DWH

Private data migration can be a hassle. Especially when you use a personal database and you do not want to give the access credential to anyone, and only host the via API server. This sometimes can cause to data be siloed and could affect teams that work from the same cooperation. The following illustrates the flow for the process of this use case

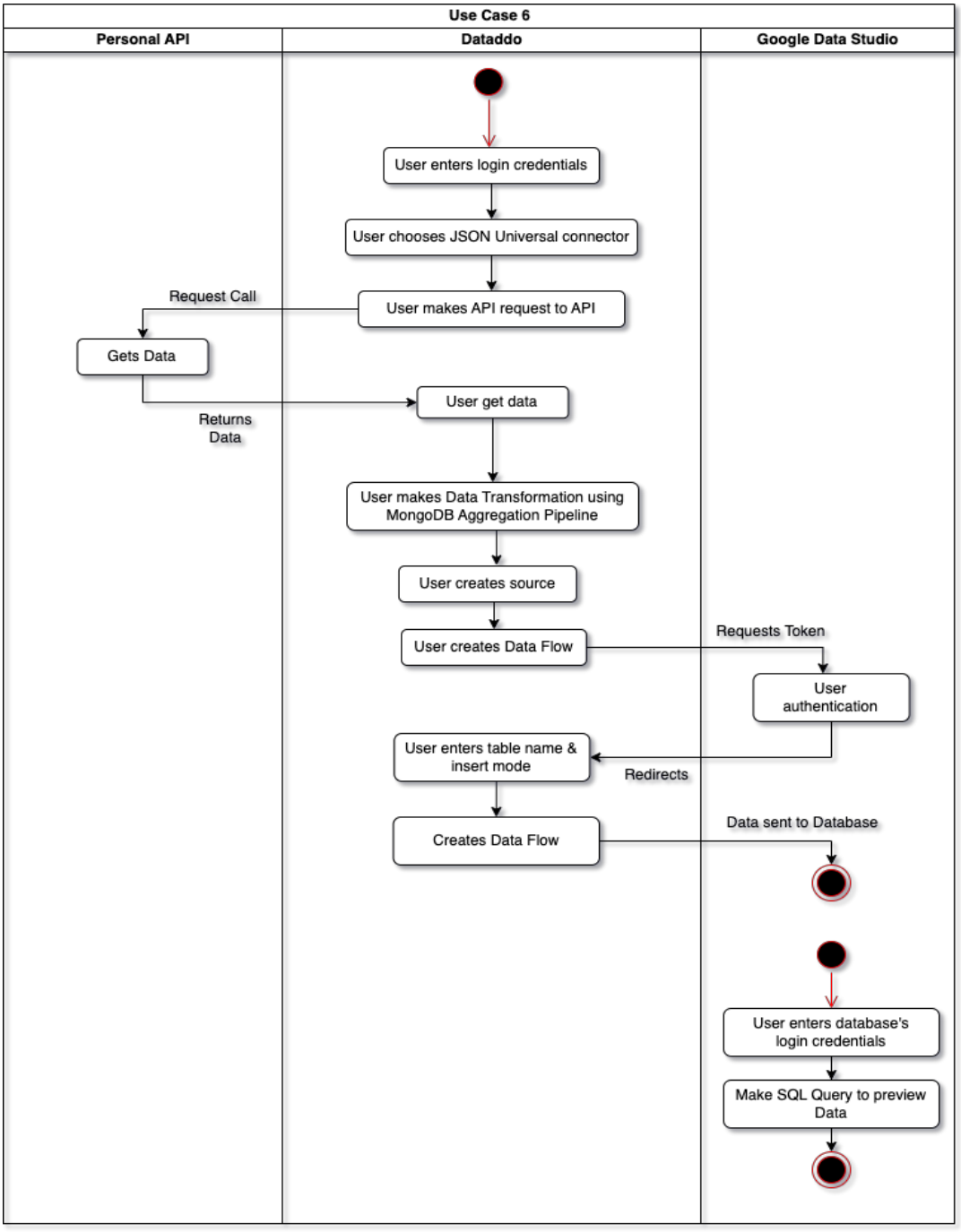


Figure 53: Use case 6: Activity Diagram

I simulate this use case by developing an API using Node JS. (See section 3.11.2)

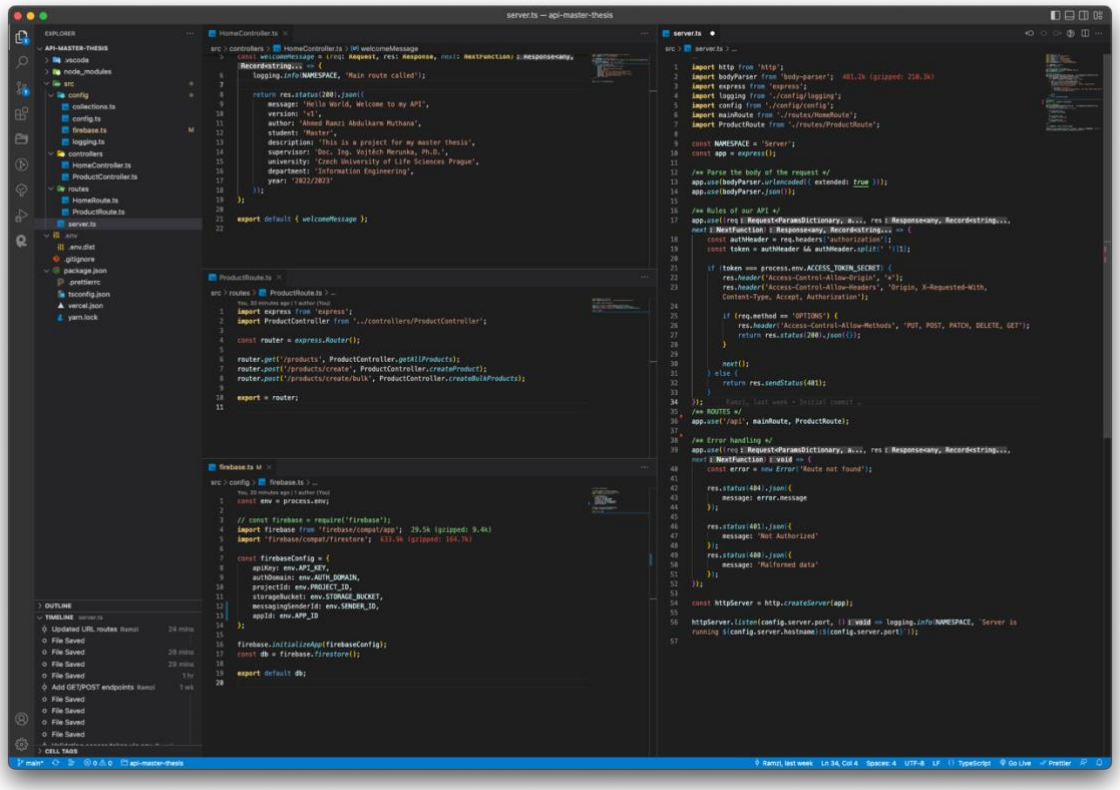


Figure 54: API Code Base Repository

The API is being hosted online on Vercel's cloud.

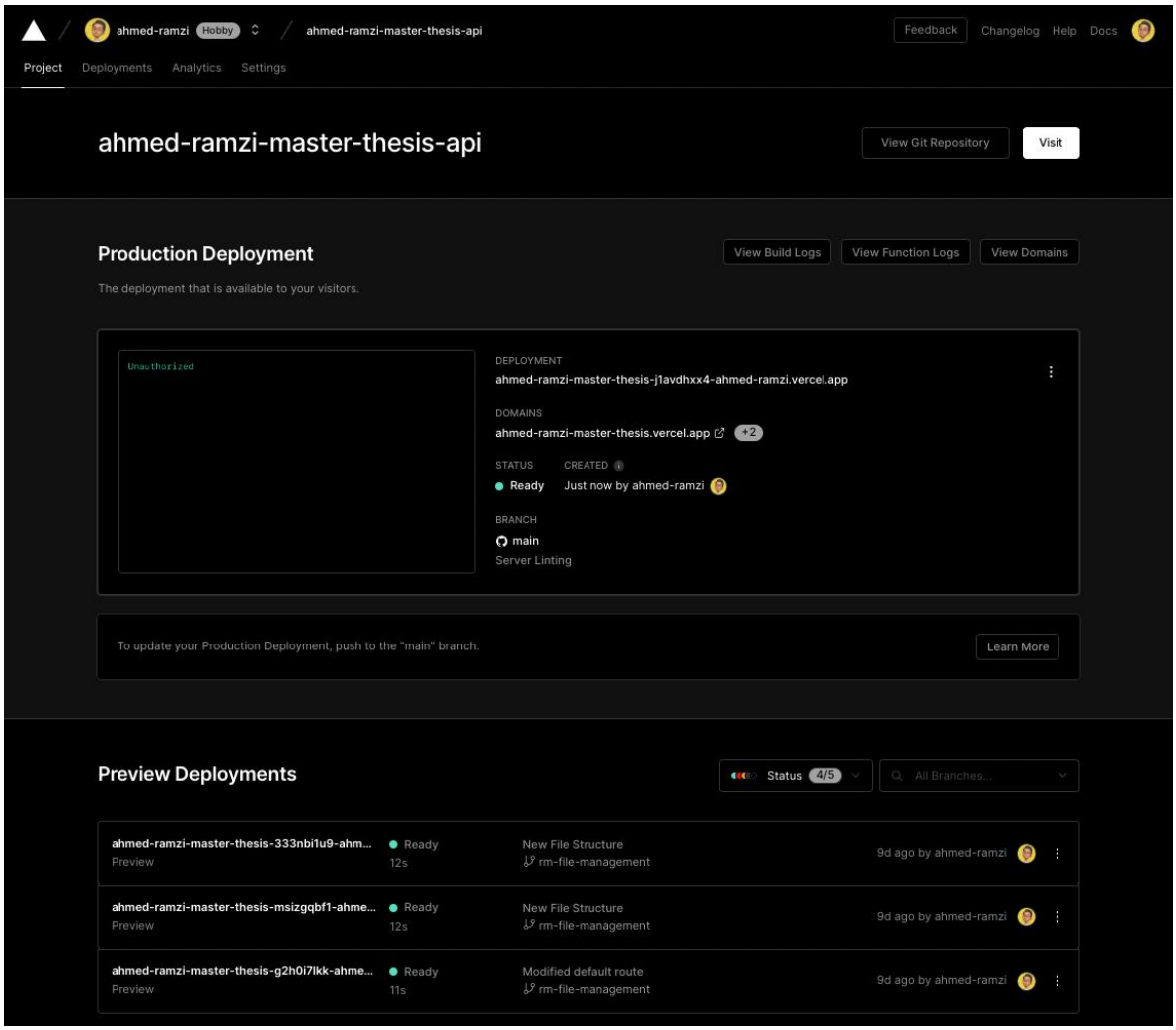


Figure 55: Vercel: API Hosting Server

The API is accessible online publicly. However, because the code base repository is saved online public, therefore, for this security reasons, I made a requirement for an access token to be able to read and write data to the database. The access token is not stored in the Code Base but stored as an environment variable. Therefore, making the token not visible in the code files.

HTTP URL:

```
https://ahmed-ramzi-master-thesis.vercel.app/api
```

The access token needs to be entered in the header of the HTTP request.

```
1231j23h12nssj1k2h12jnswj12nskjn12jwhk1nsnkahdskfjalcn
```

I use the Insomnia REST tool as my API Client to make HTTP requests

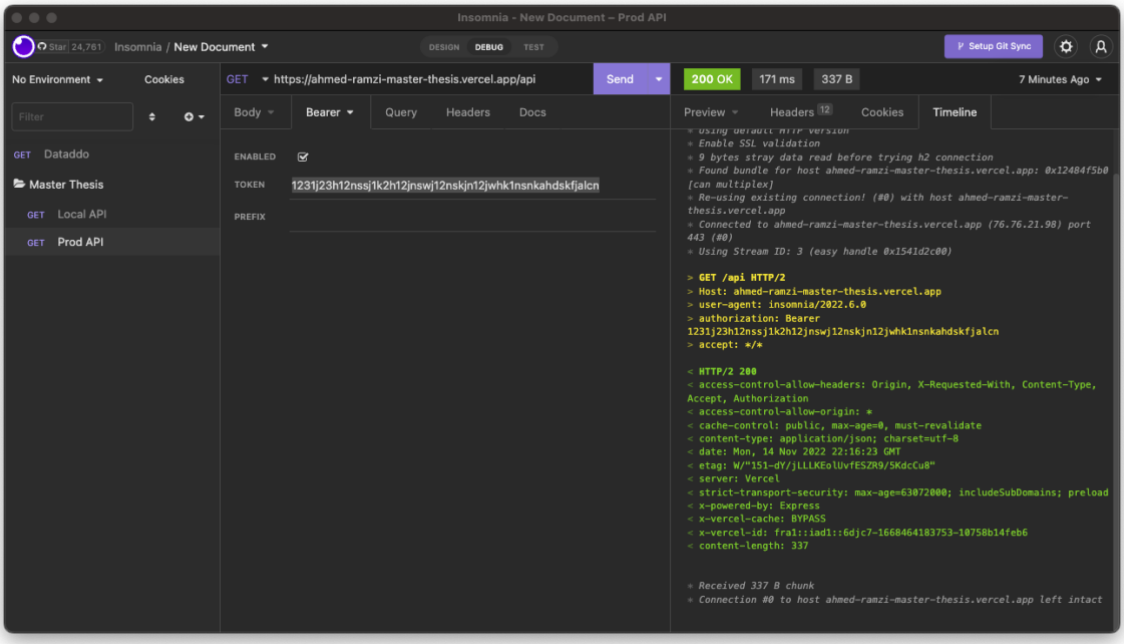


Figure 56: API HTTP Request Home Route

The following is the response to the HTTP request above

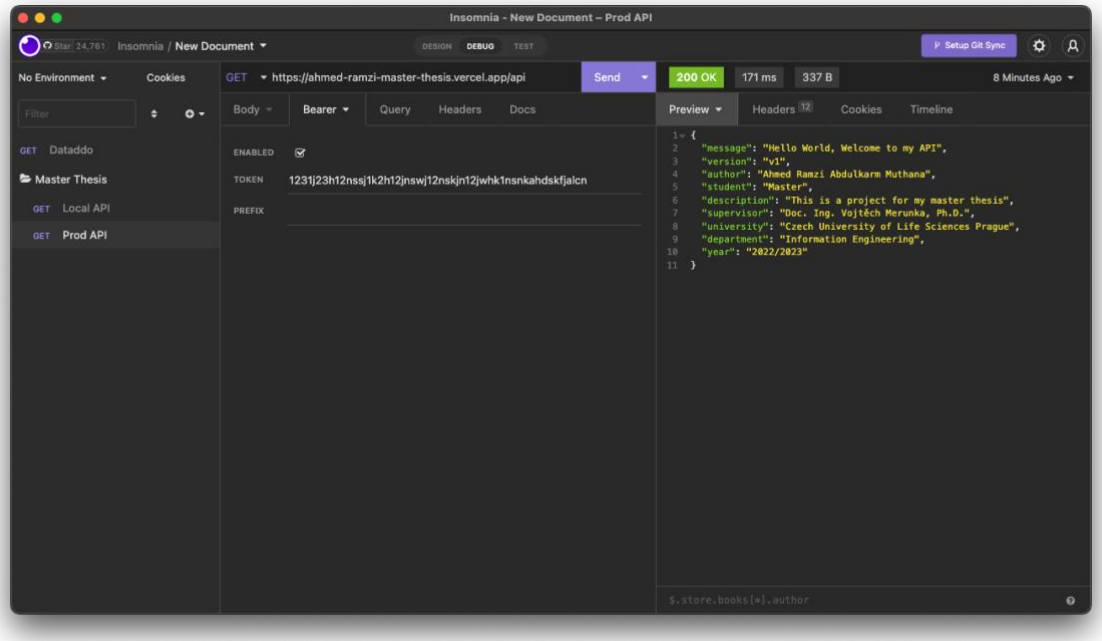


Figure 57: API HTTP Response Home Route

I used Firestore (by Firebase, Google) as the database that I will use to migrate data from it to another relational database (Google BigQuery). Firestore is a NoSQL object-oriented database, like MongoDB.

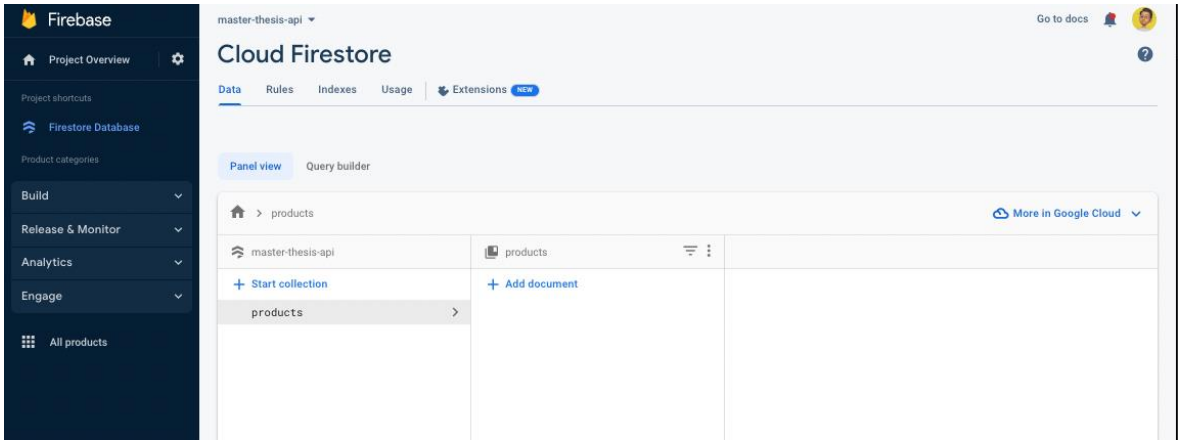


Figure 58: Personal Firestore Database with Empty Records

First, I populate my database by making a mock collection of random product data. I made the process of population easier by creating an API POST endpoint, to create one document of bulk documents.

Single product creation endpoint

https://ahmed-ramzi-master-thesis.vercel.app/api/products/create

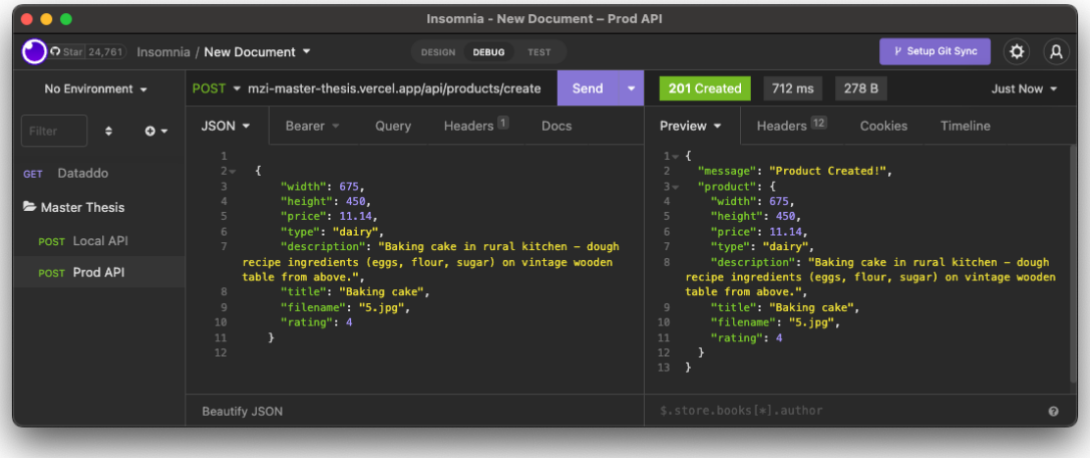


Figure 59: Creating Single Record in Collection

Data must be entered in singular object form otherwise the API will return an error with status code 422.

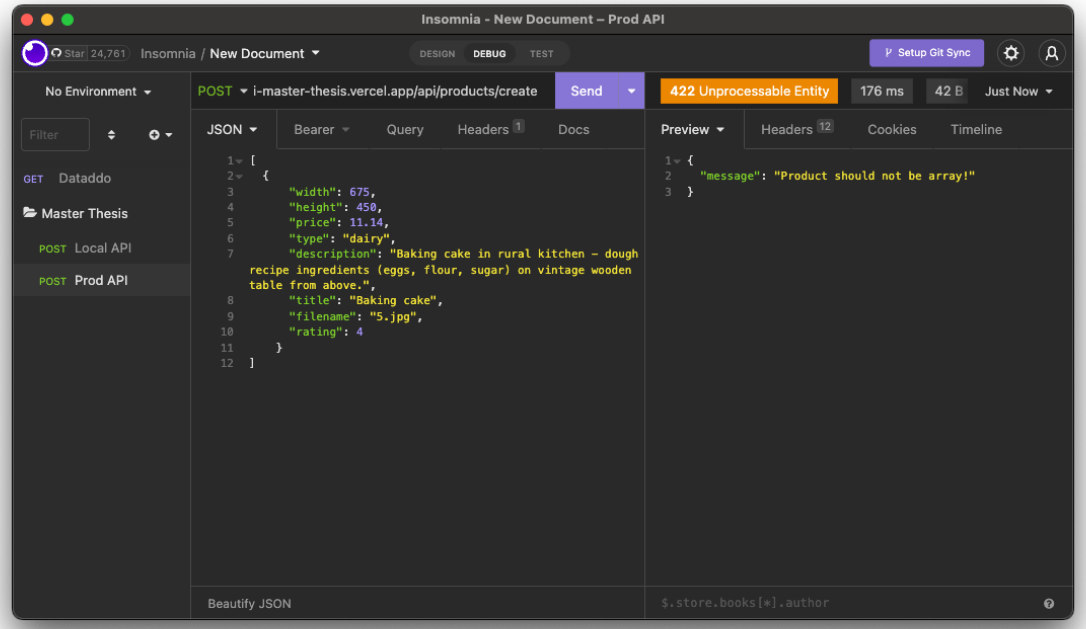


Figure 60: API Error 422

The following data should be used for bulk data population

https://ahmed-ramzi-master-thesis.vercel.app/api/products/create/bulk

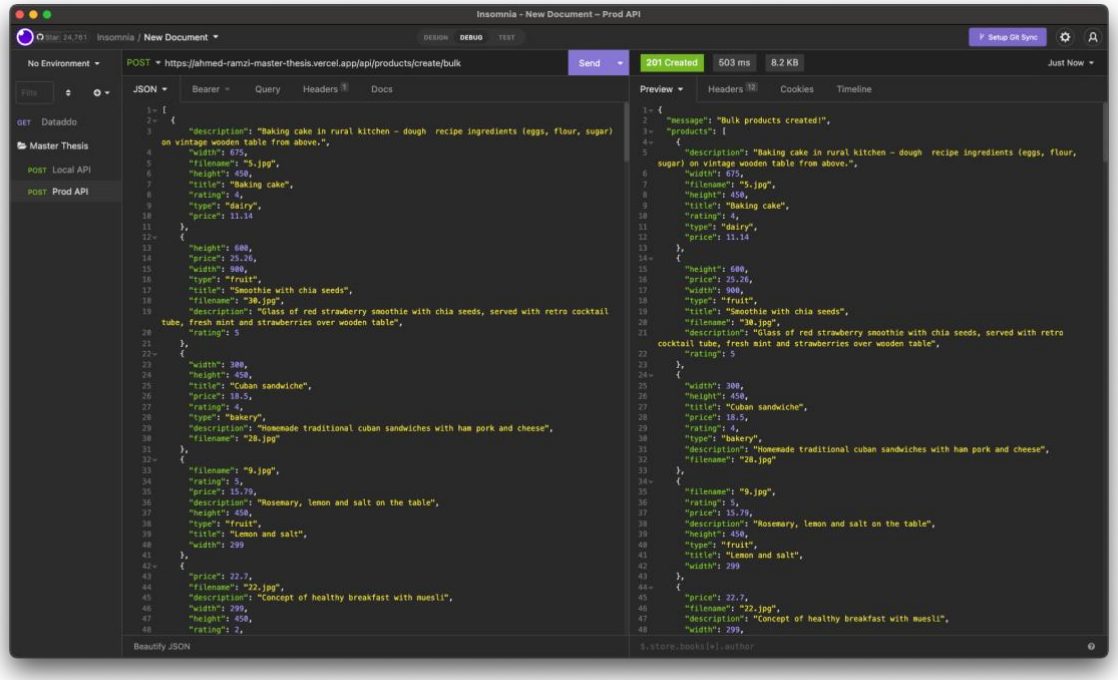


Figure 61: Bulk data creation

In seconds, the database should be filled with the mock data.

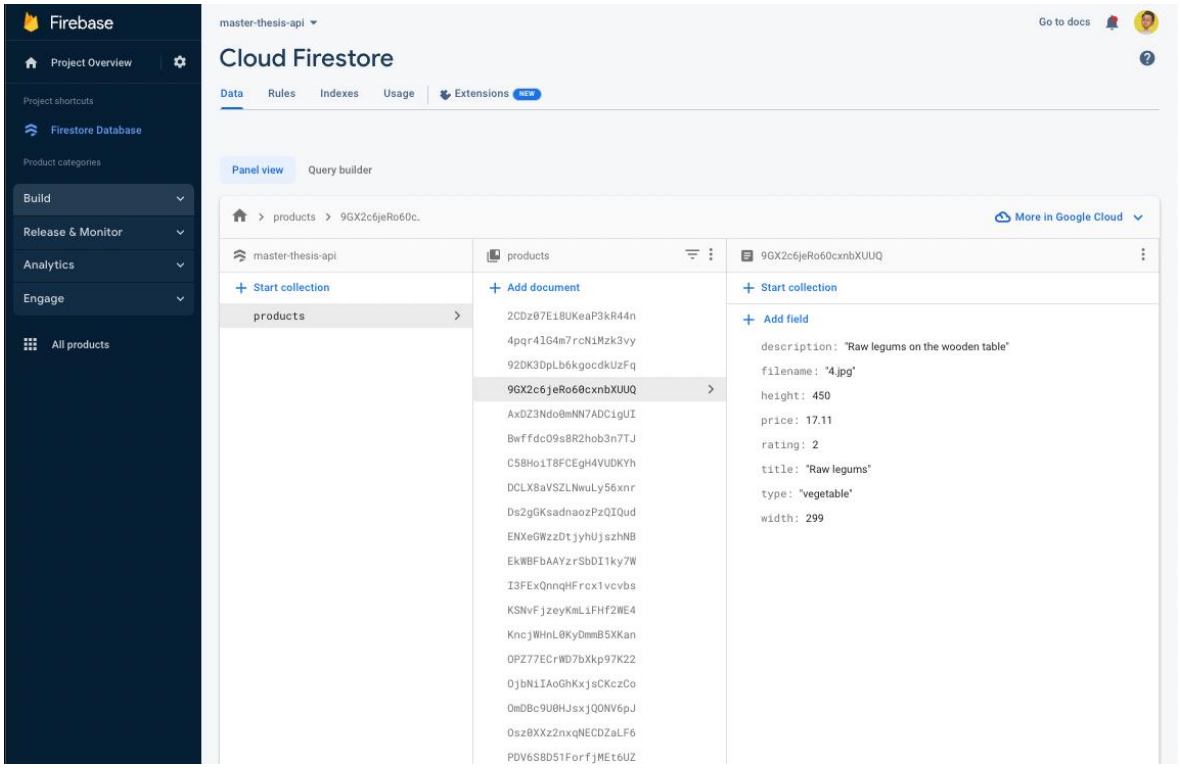


Figure 62: Database populated

Now I can use the GET endpoint to read all my products that are in the database using Dataddo's JSON Universal connector

https://ahmed-ramzi-master-thesis.vercel.app/api/products

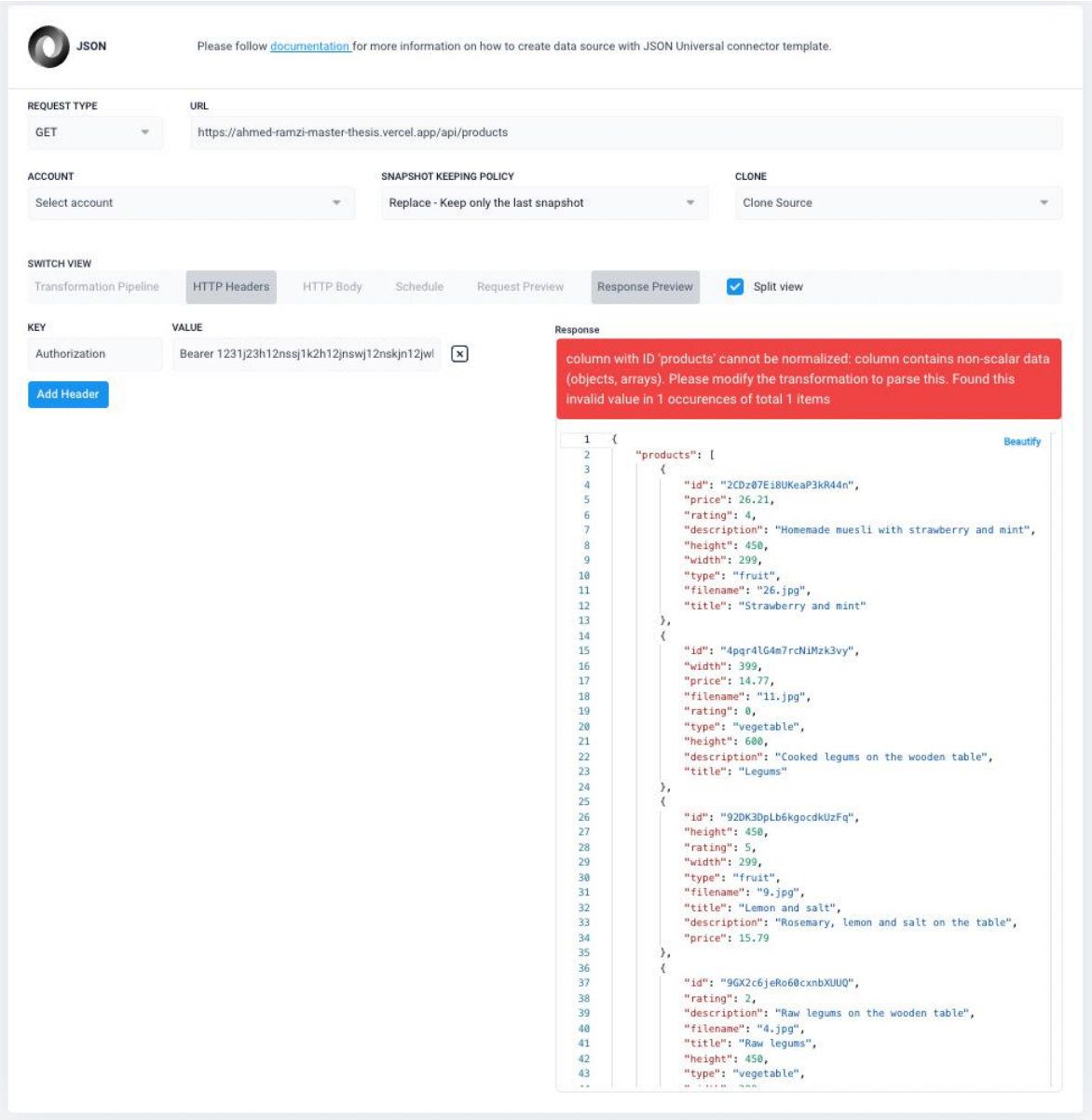


Figure 63: Personal API request on Dataddo's platform

I use the following transformation to create the table for the Data Source

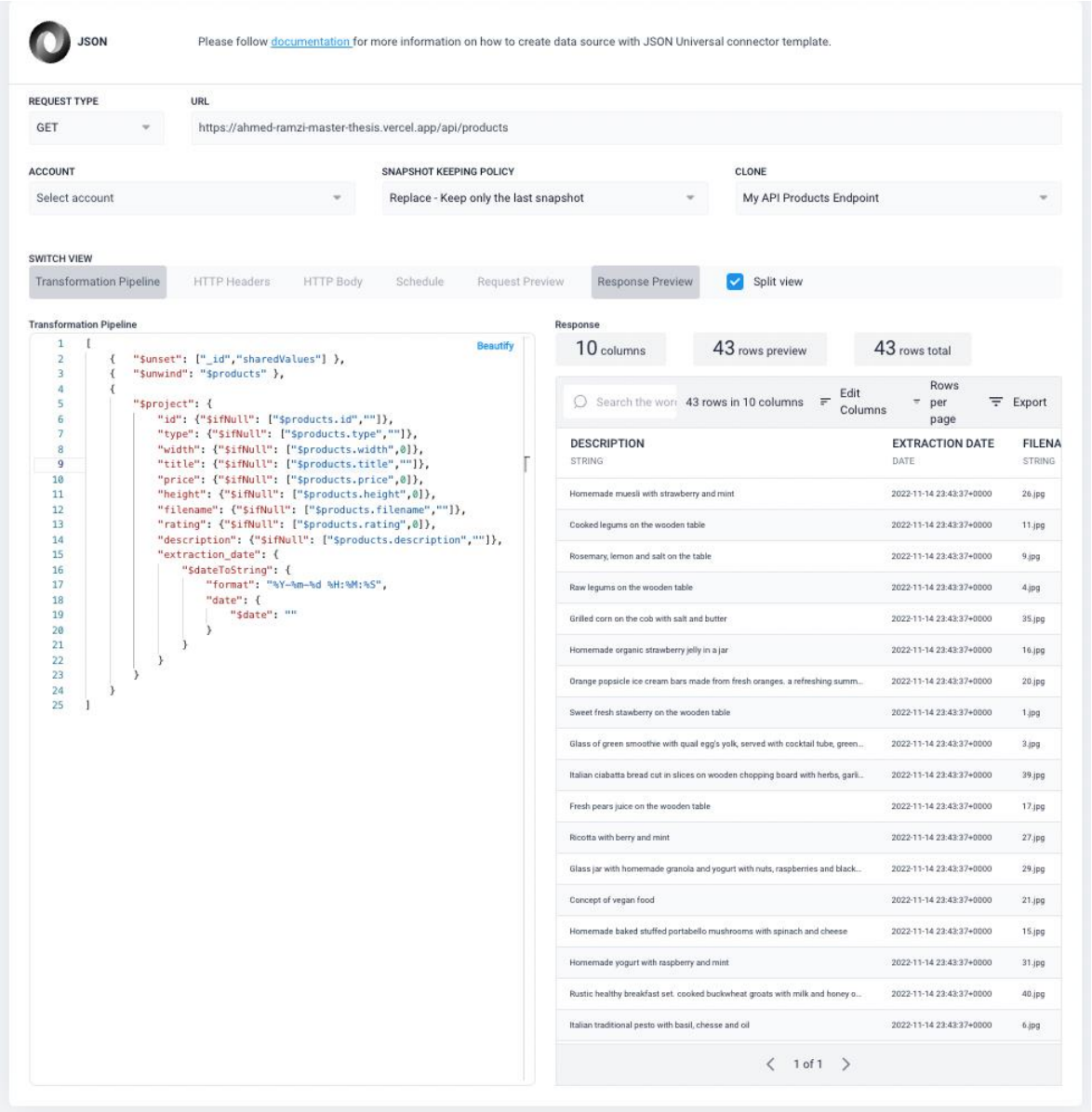


Figure 64: Data Source Creation for Personal API

The following step will be to configure the Data Flow to Google BigQuery

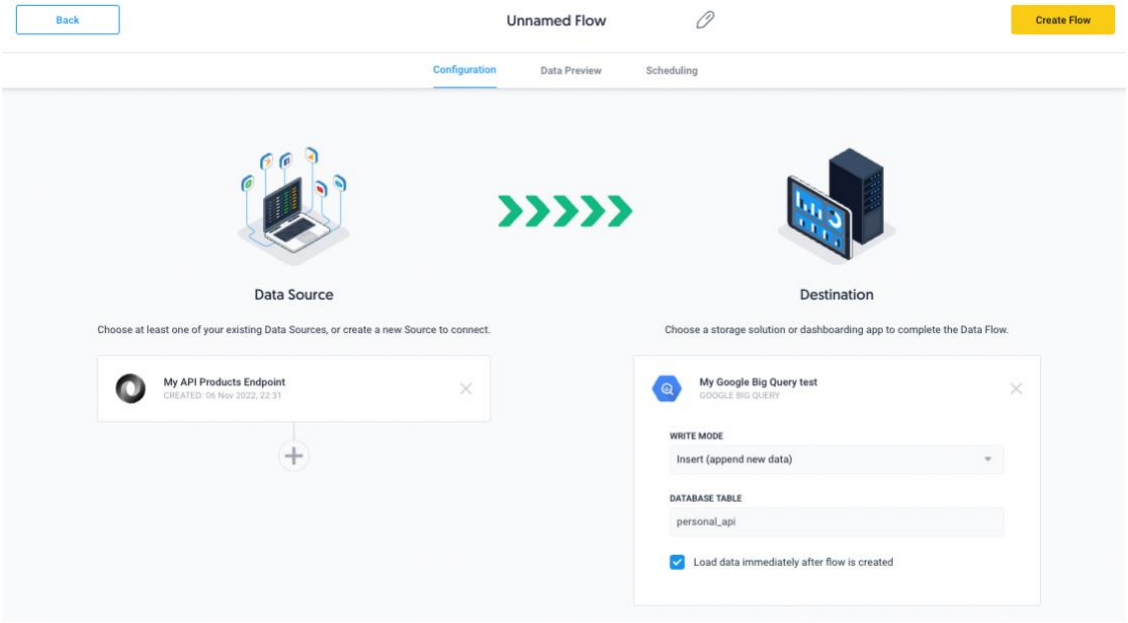
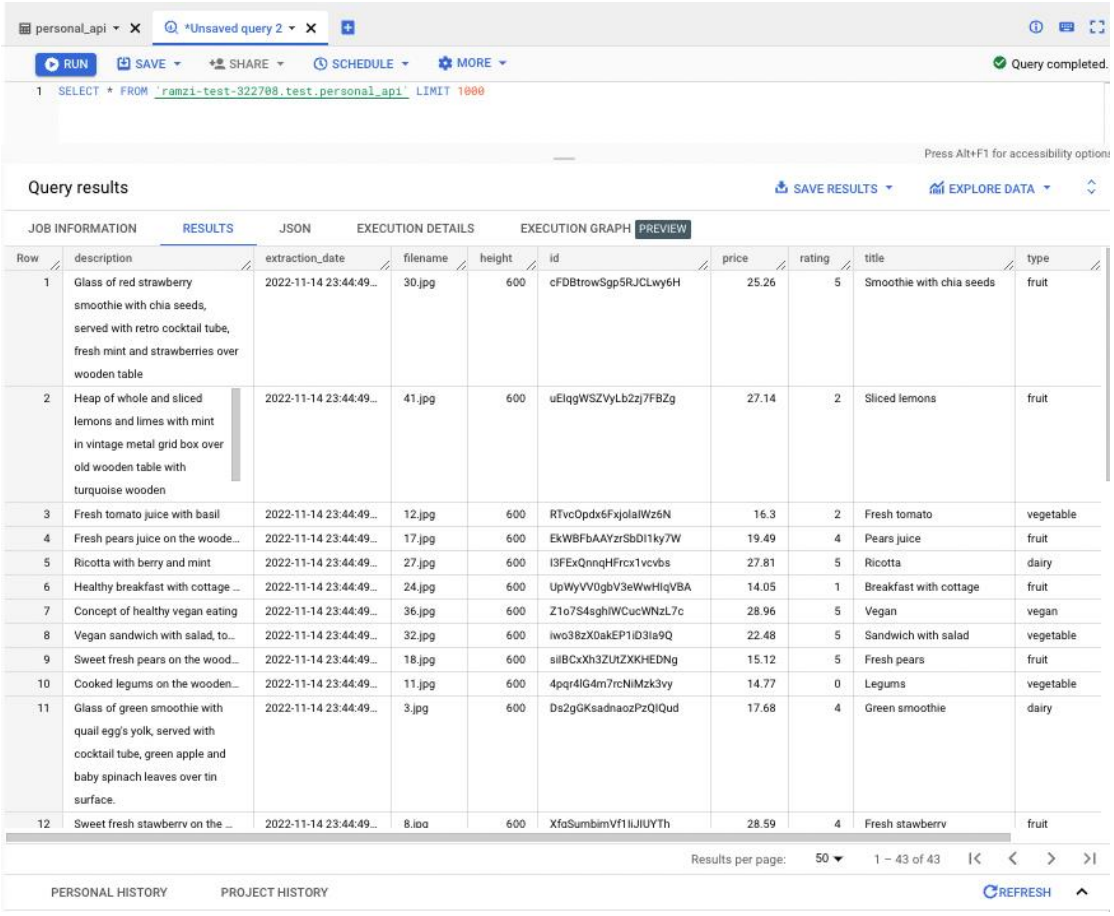


Figure 65: Data flow configuration Personal API to Google BigQuery

The data should be written to the Data Warehouse.



Row	description	extraction_date	filename	height	id	price	rating	title	type
1	Glass of red strawberry smoothie with chia seeds, served with retro cocktail tube, fresh mint and strawberries over wooden table	2022-11-14 23:44:49...	30.jpg	600	cFDBtrowSgp5RJCLwy6H	25.26	5	Smoothie with chia seeds	fruit
2	Heap of whole and sliced lemons and limes with mint in vintage metal grid box over old wooden table with turquoise wooden	2022-11-14 23:44:49...	41.jpg	600	uElqgWSZVyLb2zj7FBZg	27.14	2	Sliced lemons	fruit
3	Fresh tomato juice with basil	2022-11-14 23:44:49...	12.jpg	600	RTvcOpdx6FxjolaIWz6N	16.3	2	Fresh tomato	vegetable
4	Fresh pears juice on the woode...	2022-11-14 23:44:49...	17.jpg	600	EkWBFbAAYzrSbDI1ky7W	19.49	4	Pears juice	fruit
5	Ricotta with berry and mint	2022-11-14 23:44:49...	27.jpg	600	I3FExQnnqHFrcx1vcvbs	27.81	5	Ricotta	dairy
6	Healthy breakfast with cottage ...	2022-11-14 23:44:49...	24.jpg	600	UpWyVV0gbV3eWwHlqVBA	14.05	1	Breakfast with cottage	fruit
7	Concept of healthy vegan eating	2022-11-14 23:44:49...	36.jpg	600	Z1o7S4sghIWCucWNzL7c	28.96	5	Vegan	vegan
8	Vegan sandwich with salad, to...	2022-11-14 23:44:49...	32.jpg	600	iwo38zX0akEP1iD3Ia9Q	22.48	5	Sandwich with salad	vegetable
9	Sweet fresh pears on the wood...	2022-11-14 23:44:49...	18.jpg	600	silBCxXh3ZUtZXKHEDNg	15.12	5	Fresh pears	fruit
10	Cooked legums on the wooden...	2022-11-14 23:44:49...	11.jpg	600	4pqr4lG4m7rcNiMzk3vy	14.77	0	Legums	vegetable
11	Glass of green smoothie with quail egg's yolk, served with cocktail tube, green apple and baby spinach leaves over tin surface.	2022-11-14 23:44:49...	3.jpg	600	Ds2gGKsadnaozPzQIQud	17.68	4	Green smoothie	dairy
12	Sweet fresh stawberry on the ...	2022-11-14 23:44:49...	8.jpg	600	XfqSumbimVf1liJIUYTh	28.59	4	Fresh stawberry	fruit

Figure 66: Personal API Migrated to Google BigQuery

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