Data Ingestion Pipeline and Data Marts to Empower UK Researchers, Academics, and Business and Economic Decision Makers

Dr. Qicheng Yu[1], Stephanie Healy[1], Indrajitrakuraj Ravi[1] and Dr. Preeti Patel[1]

[1]London Metropolitan University, London N7 8DB, United Kingdom

Abstract: The data integration problem from the voluminous data generated from different sources in disparate formats coupled with a large number of diverse requirements related to the data have made the need for a reconciliation of them into a unique model, the identification of relationships, and the enabling of data analytics processes extremely vital. In light of the unabated growth of data volume and the need for data sharing across various stakeholders there is a requirement for the design and implementation of a data ingestion pipeline with a set of data marts. In this paper, we present a data ingestion pipeline which empowers hitherto impeded data users to easily access shared big data sources. We aim to improve the effectiveness and efficiency of open 993source data sharing capability so that researchers, academics, policy makers, businesses and government departments can all benefit from the use of these sophisticated data management techniques. In this work, we propose a novel data ingestion pipeline and data marts approach to utilise data generated from big data systems and effectively integrate them to a unified form, ready for use. Currently, the data ingestion pipeline focuses on UK data, as our primary aim is to support the City of London and the various communities within it. An additional benefit is the potential for developing collaboration across disciplines to tackle the economic and social challenges faced by cities in innovative ways.

Keywords: Big data, Data Pipeline, Data Marts, Data Ingestion, Extracting Loading and Transforming (ETL), Star Schema

1 Introduction

The amount of data being created each day in the world has been increasing at an extremely rapid pace. A study conducted by Forbes indicates that from 2010 to 2020, this amount has increased from 1.2 trillion gigabytes to 59 trillion gigabytes, showcasing a growth rate close to 5000%. Statista forecasts that this growth will be even faster in the years to come, projecting data volume in 2025 to be up to 180 zettabytes. In fact, current data growth estimates suggest the world's data is doubling approximately every 1.5 years. [1] As the amount of data increases, the opportunities to conduct exploratory analysis, to enhance reporting capabilities, and to address business requirements to support informed decisions is growing as well. For these reasons, it is imperative to leverage innovative data management techniques to support big data.

Businesses and world economies are starting to rely on data analytics to make relevant decisions. This is true across all industries and sectors, big or small companies and economies. Large and small businesses, as well as economies are aiming to apply business intelligence analytics for decision making. Modern business intelligence (BI) focuses on the transformation of raw data to knowledge, the

visualization of data and the data tools and infrastructure in order to promote better data-driven decision-making [2] Not only this, but researchers and academics also rely on data to conduct studies; some of them with the aim of creating an impact in the world. Therefore, we see that in many contexts, there is a need to maximize the data sharing capabilities so that all interested parties can access and utilize the power of the same data.

Data integration is an important factor in the provision of consistent and reliable data, as it consolidates previously disparate sources into a single dataset. The main challenges for optimal data integration are volume and variety, particularly when utilizing data created by other companies or institutions. The early phase of data integration is the data ingestion process, where the transportation of data from assorted sources to a storage medium (cloud), where it can be accessed and analysed is conducted. Sources of data ingestion include spreadsheets, databases, custom apps, SaaS data or web scraping. Typical destinations are data warehouses, data marts, databases or data repositories.

There are various ways of ingesting data, for example, streaming ingestion is necessary for analytics that require constantly refreshed data. Meehan et al argue that older ETL environments cause a latency, which they report can be overcome by considering ETL as a streaming problem [3]. However, there continues to be a requirement for the integration of data in a batch processing format, particularly where the issue of latency is not of direct concern.

Data warehouses and data marts are well established as destinations for ingested data and there are continuing opportunities to enhance the usage of them. Open source datasets have the potential to provide readily available data to users to gain valuable insights. However, not all users are able to perform the data ingestion process. The open datasets used in this research are UK-based and the focus is on empowering the City of London to tackle the challenges which affect all cities, such as health, crime, poverty, transportation, housing, sustainability and social wealth. Bellini et al consider the architectural requirements for data ingestion for smart city management and cite the difficulties of processing datasets that are rarely semantically interoperable [4]. Open datasets are typically generated by different people at different times for different purposes. Some open datasets (or parts of them) may or may not be private and some may have restrictions on usage.

Within this project, data has been proved to be a powerful tool, which is why we have decided to use big data as the energy behind our aim to support the City of London is faced by several challenges, ranging from areas such as ecosystems, health, crime, poverty, and social wealth. The various communities, companies and institutions generate innumerable amounts of public datasets that can provide valuable insights to help tackle some of these issues. We aim to take full advantage of available data and use it to help solve some of these issues and empower all concerned.

Use is made of the Companies House REST API, which holds company information delivered under the UK Companies Act and related legislation. Whilst this data is freely and widely available, it can be complicated to interpret. Given its complexity, it is unlikely that anyone who is not a data scientist will use it to support decision-making. Furthermore, to extract companies' data from the API, the company number must be specified; making it difficult to conduct analysis involving multiple

companies or to conduct exploratory analysis to discover new companies. This paper proposes the use of a data ingestion pipeline to explore public data and studies how using data marts provides advantages over using the API directly.

This research has the capacity to contribute to the research community, academic researchers and students, small and medium-sized enterprises (SMEs), and local councils and charity organisations. We propose a novel idea of data marts which are delivered from shared outcomes of various big data systems through effective data ingestion pipelines compared with traditional data marts that are extended from institutional data warehouses. This means that new data marts could have a wide range of data sources to meet various needs as no single company or institution itself could have so much detailed data belonging to others, even if vast data storage was available.

2 Key Techniques and Related Work

In this section, we describe and discuss key techniques involved in the project as well as referring to some related work. Firstly, different data storage options for the management of shared open source big data is presented and data marts as a chosen option is justified and explained in more detail. Secondly, what a data pipeline is and how a data pipeline works, its main objectives, and its architecture process are presented. Finally, the data ingestion stage in the data pipeline architecture is described and its methods are presented.

2.1 Data Marts

There are different data storage options that could be considered for a destination system, which can be a data warehouse, a data mart, a data lake or other data management system. A data lake is a data storage repository that can store large quantities of both structured and unstructured data. Data in a data lake is raw data without any prior processing. It simplifies the data pipeline, hence, leading to the reduction of data transport costs. It is particularly beneficial for processing large unstructured stream data such as videos or music tracks. However, since data in a lake is raw, it is difficult for inexperienced user to manipulate. Although it is becoming increasingly popular in big data management, it is not a suitable option for this project because our aim is to improve the effectiveness and efficiency of open source data sharing capability.

Data warehouse is another well-known data storage approach which offers excellent support for decision making. A data warehouse is a central platform for data storage that helps businesses collect and integrate data from various operational sources. Data warehouses serve as the backbone for mission-critical aspects of operations and have been used by many larger leading corporations. However, to build a data warehouse is time consuming and costly and it also requires completed and detailed historical data to ensure high levels of quality and accuracy. Since this project is seeking to utilise the open shared data where there may be some unpredictability concerning the level of detail or quality. Since the open datasets do not belong to any specific corporation and since we anticipate the semantic interoperability to be low, a data warehouse approach is not applicable.

According to BM Cloud Education, a data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.[5] A data mart inherits many advantages from a data warehouse, whilst being much simpler to create and with lower cost since it only focuses on a specific subject.[6]

Based on the evaluation of the different destination options, data marts are the most suitable destination system for this project. This type of destination offers subject-oriented data, which makes it more affordable and enables faster analytical performance. Generic data mart architecture is based on a dimensional model, with a Star schema design. The model uses a dimensional database management system. The schema consists of fact tables to store the metrics of the specified event and referencing dimension tables to store data regarding the context of facts, in the same way as a warehouse schema, in a relational database. [7]

2.2 Data Pipeline

A data pipeline is the mechanism used for this transportation process. A data pipeline can be defined 'as the tool that enables the handling of data flow from the source to the destination'.[8] The data pipeline architecture involves the design and structure of code to clean and transform data as needed, and route source data to the destination system. The design can be divided into four layers: the first layer consists of the data sources; the second layer concerns the ingestion of data or 'pushing' it into the pipeline; the third layer is the transformation of the data structure; and the last layer is to transport the data to its destination.[9] The basic steps involved in a data pipeline are extract, transform, and load (ETL) data.

Given the scope of this project, a data pipeline is a key element to transport data from a chosen data source to tables of a data mart. Also, it is imperative to optimize data throughput, that is the amount of data a pipeline can process within a set amount of time to ensure the pipeline works smoothly.

2.3 Data Ingestion

After selecting the most suitable destination system, we considered data ingestion methods. This was one of the most important steps for the success of this project, given we needed to find an efficient strategy to ingest big data in a short amount of time. Data ingestion is a crucial step in the architecture of a successful data pipeline, because here the components of the pipeline extract and read the data from the data sources and push it into the destination system.

A common extraction method is to use application programming interfaces (API) to read data from the source. Data can be transferred with the technique of batch streaming, which extracts groups of records in a sequential, timely manner. [9] The data analyst can then set up "jobs" to periodically start up and shut down the pipeline. Several tools have been built to simplify this process. Azure Data Studio, for example, has an extension designed to support the creation of jobs. At first we planned on using this extension to schedule a job to extract data from the API,

however we found an alternate and more efficient solution which is explained in Section 4.

2.4 Related Work

Previous research has acknowledged the challenges of big data sharing and management. Positive shifts in corporate culture is key to overcoming the situation of data silos and thereby realising the full of potential of big data sharing. [10] However, it is not just individual companies that can harness this potential, entire industry sectors, given the advancements in network technology, can enhance the new perspective of data sharing at enterprise level. [11] Further, the connectivity afforded by computer networks can break the barriers of knowledge flow inside and outside enterprises. [12]

Data lakes can store large quantities of raw structured and unstructured data, thereby simplifying the data pipeline. Indeed data lakes can create new challenges such as dataset discovery whilst addressing newer requirements for classic problems such as ETL. [13] Data lakes can be problematic in that organizational datasets can remain unused and uninterpreted. [14] However, existing approaches such as data marts are a tried and tested approach that is particularly suitable for certain types of sources. Researchers continue to make good use of advanced OLAP-based tools when working with big social data [15]. Where heterogeneous data sources are concerned, traditional ETL systems have proved to be an efficient solution. [16]

In previous research, data pipelines have transported data to destination systems other than data marts. Data warehouses, for example, have been used in cases where data has to be retrieved from disparate sources and where a broader range of data is needed rather than just a subset of it. Moreover, other researchers have used data lakes as the pipeline's destination systems. In cases where data lakes have been used, data was unprocessed rather than processed, which often needed data scientists to understand it. [17]

Data marts in utilities sector have played an important part. Water and Power utilities have integrated their stations with microprocessor based devices and supervisory control and data acquisition (SCADA) technology to monitor and store operational and non-operational data. However, these automated systems do not have the ability to transfer the data to a central data warehouse. To solve this the data mart servers are linked to the SCADA and other systems to retrieve the data and store them in a central storage system. The datasets are integrated into unique information that are made accessible to specific user groups within the utility. [18]

The COVID-19 pandemic has created many unprecedented challenges in various fields such as medicine, biology, public health , social science etc. Researchers have developed a Data warehouse called as the COVID-WAREHOUSE. This warehouse models, integrates and stores the Covid-19 data that are published on a daily basis by the Italian Protezione Civile Department and the pollution and climate data published by the Italian Regions. The data relating to Covid-19, pollution measures and climate are subjected to an automatic Extraction , Transformation and

Loading (ETL) process and then integrated and organized using the Dimensional Fact Model using time and geographical location as main dimensions. The Warehouse is designed to support On-Line Analytical Processing (OLAP) analysis and provides visualisation in the form of a heatmap and extract data for further analysis. The initial findings of the research indicate that the spread of the Covid-19 virus is significant in regions of high concentration of particulate in the air and in the absence of rain and wind. [19]

3 Proposed Data Ingestion Pipeline and Data Marts and Design

We based our methods on the studies conducted from previous research as well as the gaps identified. In this research, we propose a novel data ingestion pipeline and data marts approach. Instead of creating data marts from a data warehouse, we create data marts directly from the open shared data from big data systems to make the task more flexible and effective. It can avoid reliance on a data warehouse and will also be more scalable since it is much easier to add a data mart when a new shared data source is made available. The proposed data marts can better utilise data generated from big data systems and effectively integrate them to a unified form, ready for use. In the proposal, data marts are supported with independent data pipelines to ingest data gathered from multiple shared UK big data sources. In the following sections, we present details of our proposed data ingestion pipeline and data marts.

3.1 Project Setup

To make the destination system more accessible and easy to connect to open data sources provided by various big data systems, a cloud based approach will be a natural choice. We have made use of the MS Azure cloud platform as it is competitively priced and relatively easy to set up. We then added cloud-based Azure SQL database to host data marts and used cross-platform Azure Data Studio as the main database access tools. To conduct the ETL process, we used Python programming language and SQL via Azure Data Studio GUI. Python programming language offers rich data science and web access packages to achieve our objectives. Both Python and SQL are natively supported by the Azure SQL database and we can easily switch the kernel from SQL to Python programming in Azure Data Studio to support the project development.

3.2 Data Marts Design

Before we began the set up process of the data marts, the scope of the project had to be selected. With the selected scope, the fact and dimension tables for the relational database were chosen.

a) Selected Subject

Hundreds of public UK datasets from various big data systems were divided into eight different categories were analysed and evaluated. Topics cover key challenging areas which local communities face ranging from health,

economy, demography, social activities, education, crime, environment, and housing. Datasets from various data sources are carefully explored and analysed in depth to determine the accessibility of data sources as well as detailed contents. Based on a careful and thorough inspection of available datasets, we decided to start with economy and demography categories as test cases for the project, since both categories contain larger volumes of data from UK official Companies House and Office for National Statistics. We determine that if the two selected cases work properly, others can simply follow suit.

b) Dimension Tables

Business Dimension: works as a business directory for all UK companies registered at UK Companies House which includes attributes containing key descriptive information about a company such as company number, company name, company postcode (as backup), the date of incorporation, and more.

Date Dimension: covers date range 2010 to 2030 which includes attributes such as day, week, month, year, season, and holiday.

Location Dimension: includes the entire hierarchy of location information for the UK at postcode level. It includes postcode, output area, ward, LSOA, MSOA, and region.

Demographic Dimension: includes attributes regarding the demographics of the general population of an area.

c) Fact Table

Business Fact: includes measures such as company size and company turnover, which will be loaded periodically.

Demographic fact: includes measures such as population of age, race, ethnicity, gender, marital status, income, education, and employment.

Schema Design

A star schema for the data mart is used for design. In the relational schema, each dimension table consists of a list of attributes required to understand the background of the measures in the fact tables. In the project, multiple stars will be created gradually, with all stars will sharing the common dimension of data and location. At this beginning stage, only the business data mart will be created.

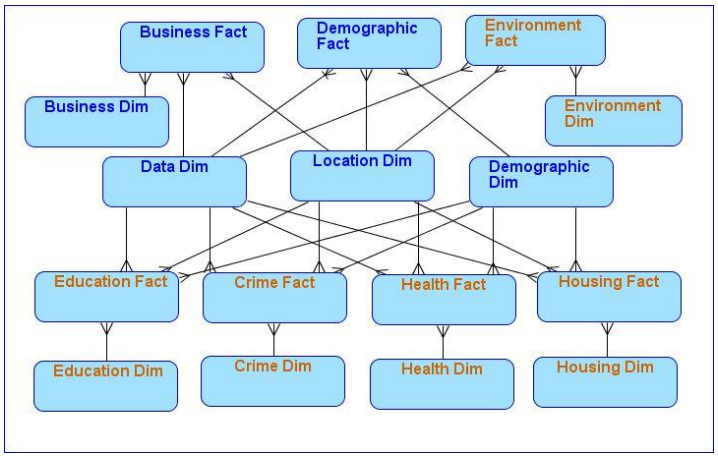

Figure 1. Star cluster schema design of the project's data marts

Figure 1 depicts the design of a star cluster schema consisting of nine dimensions tables and seven facts tables. In Figure 1, stars labelled in blue colour are chosen to be implemented initially and other stars in Orange colour are for future work, with the benefit that they can share the same dimensions already created in the project. The Date and Location dimension tables are most important dimensions as they will be used for all stars. In the project, these two dimensions are used to support both the Business and the Demographic facts. Date and location will provide the descriptive information when querying the database to perform analysis. This star cluster schema displays how the data in the data marts is related and can give the users an insight into the type of queries that can be made to the data marts.

3.3 Data Pipeline Architecture

The data pipeline architectural design shown in Figure 2 and described in the previous section was implemented by using external data sources, ingesting data, extracting it from the sources, pushing it to the source systems, transforming it and finally using it for querying.

While implementing the architecture based on our research and study on previous projects, we discovered that it was not the best approach. Batch streaming, the ingestion method proposed in the literature was not found to obtain the optimal throughput rate. For this reason, we arrived at a faster and more efficient method to extract data stored in the API surface. In this research, we propose a way to optimize the ingestion layer of the data pipeline architecture by using an alternative method for data ingestion.

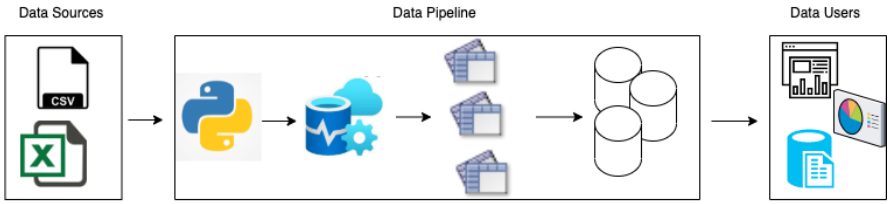

Figure 2. Data Pipeline Architecture

4 Implementation and Challenges

Based on the logical design of the data marts, the physical database dimension tables and fact tables can be created with SQL language using Azure Data Studio.

Once the physical design of the data mart was created, we populated the dimension tables. To do this, we used public datasets downloaded from shared open data sources from official websites. These datasets were cleaned in Python and converted into CSV files. After these components were created, the source destination was ready for the data pipeline to be implemented. The data ingestion is not a trivial job as the data contains more than serval millions of rows, and the whole process becomes hugely memory-hungry very quickly. To address the issues, we have to read and process data in trunks and garage them if it needed. Since the process was done and recorded in program, it could be simply reused or amended with very little effort in the future.

The first layer of the data pipeline architecture regards the data sources. For this project, an API was used to read data from the source, website Endole.co.uk. This website is part of UK Companies House which stores big data regarding UK businesses. The API enables users to access UK business data for free, but it can only return one record per call. Therefore, many users would find it impossible to obtain a broader perspective of their business environment in terms of market segments and competition.

The second layer of the architecture regards the ingestion of data. For this step, we created an application account and used Python code to extract the data from the API necessary to populate the fact table. Based on our studies, previous research recommends the use of scheduled jobs to conduct this step. The source website, Endole, has a limit established which allows only 300 rows of data to be downloaded every 5 minutes. At first, we created a scheduled job using an extension in Azure Data Studio to ingest 300 rows of data from Endole every 5 minutes. This method used a single account and was able to extract only 20,000 records a day.

Given this slow pace, some optimisation was needed. Therefore, instead of creating a single scheduled job, we created twenty different application accounts to extract the data simultaneously. This method of ingestion allowed 360,000 rows of data to be ingested in a single day rather than the original 20,000 records completed using the method proposed in previous research studies. By using multiple application accounts and making each transaction more than 5 minutes to enable the process to complete without waiting, it was no longer necessary to use scheduled jobs to extract data from Endole, making the ingestion faster and arguably less complicated.

The last layer of the architecture is to transport the extracted data to its source destination. For this step, we formulated functions in Azure Data Cloud to transport the financial data extracted from Endole to its destination, the Business fact table. This concluding step finalized the population of the data mart, making it ready to solve analytical queries and reports. The data ingestion pipeline could be scheduled to run periodically to gather dynamic data about companies. In the project, we intend to run the process every quarter to capture the changes such as turnover and size of a company.

5 Results and Evaluation

We have focused our efforts in making sure the data mart provides outputs that will support the researchers, academics, and business and economic decision makers within the City of London. Cities constantly face issues that affect multiple areas: instability in the economy, high crime rates, over-stretched health system, and pollution are amongst a few of them. For this research, we have decided to pursue a data mart that possesses reporting and analytical capabilities that can help tackle some of these. In this paper, we highlight and present a scenario in the area of pollution, which represents what could also be possible for the other areas.

Pollution and harm to the ecosystem is one of the most predominant problems being faced by the city of London and the entire UK. Poor air quality is a growing concern in the main capital, having tens of thousands of citizens breathing polluted air and having 99% of Londoners living in areas where pollution exceeds World Health Organization recommended guidelines for particulate matter, an extremely harmful pollutant. Exposure to pollution affects everyone who lives and works in London and most of the UK's population; from the most vulnerable groups, like children and seniors, to the healthiest and wealthiest communities. It is well known that poor air quality stunts the growth of children’s lungs and worsens chronic illness, such as asthma, lung and heart disease.

Becoming green is increasingly important for the wellbeing of citizens and companies are acutely aware of this. Businesses are opting to migrate to greener and more sustainable solutions and operations. Furthermore, there are companies such as Renewable Energy Waste Solutions (REWS) who are continually fighting to create a greener United Kingdom. This company has re-invented the concept of waste, eliminating it where possible and processing waste to create fuels that reduce the end-users reliance on polluting fossil fuels. Their efforts help reduce pollution to the environment as well as people and businesses carbon footprints. REWS is seeking to expand to other parts in London and finding a suitable location is essential. A location where there exists recycling centres around to provide them with raw materials, such as plastic waste, would be the most appropriate. Small businesses such as REWS would find it challenging to get a list of the areas in London with recycling companies and the names of these businesses, as the dedicated resources required may not be available. REWS could easily access our data mart and issue simple queries to obtain the necessary results.

As explained in the implementation section of this paper, the data mart contains a location dimension storing data of London addresses and a business

dimension storing data regarding businesses in London. We wrote the following SQL query to obtain a list of all the recycling companies in London and their locations.

```
SELECT TOP (100) BusinessDim.EB_Code
      ,BusinessDim.CompanyName
      ,BusinessDim.CompanyNumber
      ,pcds
  FROM BusinessDim, BusinessFact, LocationDim
  where BusinessDim.EB_Code=BusinessFact.EB_Code
  and LocationDim.ID=BusinessFact.PostcodeKey
  and BusinessDim.CompanyName like '%RECYCLE%' AND CompanyStatus like 'Active%'
```

Figure 3: SQL Query to return list of recycling companies in London

rcycleCompany

EB_Code	CompanyName	CompanyNumber	pcds
51893	ADVANCED SUSTAINABLE DEVELOPMENTS RECYCLE EXCHANGE LIMITED	11663196	W1S 2GF
86219	ANDREWS RECYCLED MATERIALS LIMITED	6027175	RM13 8UG
152022	BEST 4 SKIPS & RECYCLE LTD	6601092	SE16 7DL
178139	BOYD RECYCLE LTD	8701227	SE3 9SD
240308	CINCHONA RECYCLE LTD	11408133	N1 7GU
248205	CLEAR REUSE RECYCLE LTD	5628431	IG11 0AQ
344675	EBAUX RECYCLE SERVICES LTD	7317034	SE6 3DJ
368292	ENVIRECYCLE LTD	4185753	SE19 1TS
368269	ENVIART RECYCLE LTD	11856364	WC2H 9JQ
386974	EZY RECYCLE LIMITED	7959072	UB8 3PS
408482	FLASH RECYCLE SERVICES LTD	12321493	UB4 0SE
432010	GADGETS RECYCLE LTD	10343784	IG1 4TA
450942	GLOBAL INTER RECYCLE LTD	11007991	E13 9PJ
468090	GREEN LIFE RECYCLE AND RECOVERY LTD	12087860	W3 7JZ
520691	HUMAN RECYCLE COMPANY LIMITED	10846183	SW1Y 5EA
520690	HUMAN RECYCLE COMPANY INVEST LIMITED	10846075	SW1Y 5EA
533082	IMAXRECYCLE LIMITED	9200511	SE15 2NB
524296	I CARE RECYCLE LIMITED	7543873	SE21 8BS
524704	I RECYCLE DATA LIMITED	11414460	SE10 8EY
555960	IT EQUIPMENT RECYCLE LTD	12521086	HA9 7RE
556465	ITALAT RECYCLE UK LIMITED	10404996	W1W 7LT
591115	K.O.C RECYCLE LTD	11408199	N9 9BU

Figure 4: SQL Query output of recycling companies in London and their postcodes

Figure 4 displays a screenshot of the output provided by performing the query in Figure 3. The list in Figure 4 includes all of the recycling companies in London, with the company name, company number and postcode. This list, for example, could be used by REWS to obtain valuable insights to support their decision of where to expand next. This could save them valuable time and resources by providing fast and actionable information.

Moreover, imagine that now that REWS knows the names of the recycling companies and their postcodes, it would know which areas have the most recycling businesses. We performed additional queries such as this one to provide an output that shows a list of all of the areas in London that have recycling companies in it as well as the number of recycling companies there.

```
SELECT ladcd, ladnm, count(msoa11cd) as Companycount
  FROM BusinessDim, BusinessFact, LocationDim
  where BusinessDim.EB_Code=BusinessFact.EB_Code
  and LocationDim.ID=BusinessFact.PostcodeKey
  and BusinessDim.CompanyName like '%RECYCLE%' AND CompanyStatus like 'Active%'
  group by ladcd, ladnm
  order by Companycount desc
```

Figure 5: SQL Query to show areas in London and the number of recycling companies there

recyclecompanyinarea

ladcd	ladnm	Companycount
E09000007	Camden	7
E09000012	Hackney	6
E09000019	Islington	6
E09000033	Westminster	6
E09000026	Redbridge	4
E09000016	Havering	4
E09000017	Hillingdon	4
E09000009	Ealing	4
E09000003	Barnet	4
E09000002	Barking and Dagenham	3
E09000024	Merton	3
E09000025	Newham	2
E09000027	Richmond upon Thames	2
E09000028	Southwark	2
E09000010	Enfield	2
E09000011	Greenwich	2
E09000001	City of London	2
E09000008	Croydon	2
E09000022	Lambeth	2
E09000023	Lewisham	2
E09000021	Kingston upon Thames	1
E09000018	Hounslow	1

Figure 6: Output of SQL query performed in Figure 5

By looking at the output displayed in Figure 6, REWS can make an informed decision when it formulates its expansion strategy. By using the data marts we have designed, REWS can not only look at the recycling companies in London and their addresses, but also analyse which area in London would be the most appropriate for their new location, based on the number of recycling companies in the given areas.

Going further, given that REWS not only operates within the City of London, but the entire UK, we decided that this analysis could expand and include businesses from around the United Kingdom. To make it even more specific, we decided to include another filter to the analysis. Suppose REWS not only wishes to find a location where there are many recycling centres around that can provide them with raw materials, but they also desire it to be far from the city and residential areas and closer to rural areas. Taking this new requirement, we configured a query that would only output UK recycling companies that are located in rural areas.

UKrecyclecompanyinruralarea

CompanyName	CompanyNumber	pcds	wzc11nm	ladnm
BMS RECYCLE LIMITED	9883292	BL1 5AL	Primarily residential suburbs	Bolton
NEWRY CYCLE RECYCLE C.I.C.	NI674437	BT35 6PH	Primarily residential suburbs	Newry, Mourne and Down
SALVADOR RECYCLERS LIMITED	4987533	DA11 0JF	Self-employed tradespeople in multicultural metro suburbs	Gravesham
PACKBUILD (RECYCLE) LIMITED	SC517640	G77 6TN	Primarily residential suburbs	East Renfrewshire
ASHWELL RECYCLED TIMBER PRODUCTS LIMITED	4358838	IP3 8GD	Rural with mining or quarrying	East Suffolk
FUTURE RECYCLE LTD	11527195	LE3 6AH	Primarily residential suburbs	Leicester
100% RECYCLED PANEL COMPANY LIMITED	7370670	M25 1PY	Professional home-workers in outer suburbs	Bury
HV OIL RECYCLE LTD	12635868	N15 5NS	Self-employed tradespeople in multicultural metro suburbs	Haringey
ASHCROFT ASSET RECYCLERS LTD	12082059	RM19 1LA	Self-employed tradespeople in multicultural metro suburbs	Thurrock
SAVEWORLDBYRECYCLE LTD	12832658	B17 8EN	Teachers and carers in metro suburbs	Birmingham
ECO RECYCLERS LTD	11416114	B21 9NN	Self-employed tradespeople in multicultural metro suburbs	Birmingham
SLW RECYCLE LIMITED	9865725	B80 7EE	Primarily residential suburbs	Stratford-on-Avon
PRIMERECYCLE UK LIMITED	10227831	BL3 3QE	Rural with mining or quarrying	Bolton
SUPERGOOD RECYCLES LTD	13403532	BR5 2DZ	Primarily residential suburbs	Bromley

Figure 7: Output of query

The output displayed in Figure 7 shows a list of UK recycling companies with their company names, company number, post code, type of area (rural, residential, etc), and the city where it is located. By looking at this scenario, one can start to grasp the extent to which this data mart can be taken. Various analytical reports can be conducted by simply adding or modifying basic queries to filter desired data. The queries themselves are basic because the data mart has been designed appropriately.

6 Recommendation for Future Work

An important facet of data pipelines is scalability and an ability to incorporate multiple additional sources. We have constructed the pipeline to accommodate other dimensions to the data mart, in order to provide deep analytics. The challenges that most cities face tend to be overlapped and arriving at innovative solutions can be greatly facilitated by the construction of a single ingestion pipeline that encompasses a number of data marts. We have identified areas for potential future work such as health, housing, crime, unemployment and ecosystem. We believe the new dimensions can be added to the existing data mart to provide relevant analytical reports to help address issues faced in London such as crime, health, homelessness, and pollution.

For example, we found multiple public datasets concerning the UK Health System in the National Health Service (NHS) website. The NHS publishes open datasets daily, with infinite amounts of data. By using our proposed technique, a data pipeline could be implemented to stream data from the NHS website and transport it to a fact table in the data mart. With this additional fact table, users could perform exploratory analysis that integrates health data with the existing data in the data mart. We have considered various analyses that could be performed such as exploring if a person's profile has an impact on the types of illnesses he/she has. This would be done using variables in the demographic dimension along with variables in the health dimension and link both through the postcode or living area of the different populations.

Furthermore, another pertinent source from open UK data are the multiple datasets concerning housing information. These datasets include relevant information such as renting and selling prices, area population density, and area council taxes. Adding a housing dimension with all of this information can be a valuable extension to the data mart, given it can take analysis even further. For example, if a business is looking for the best area to open its new location, it can now look not only at the business dimension information to understand competition in the area, but also look at rent prices and council tax details to evaluate which area would be the most profitable overall. The same concept can apply for people who wish to look for the best location to move to; they can evaluate not only the demographic profile of people living in the area by analysing variables in the demographic dimension, but also integrate the information in the housing dimension to evaluate which is the most appropriate living area for them.

Additionally, our data pipeline and data mart opens up the potential to include dimensions such as electricity consumption, air pollution, traffic light systems and under/over ground public transport systems. In the longer term, more complex and subtle challenges such as ageing, mobility and social segregation could be

considered for evaluation and analysis. Also, there are opportunities for expansion to a national level, at which point the optimising of the data pipeline and data marts could be considered.

7 Conclusion

In this paper, we present an effective and efficient way to use APIs by developing a data pipeline that ingests big data stored in the interface and directs it to subject-oriented data marts. We have proposed a creative idea of data marts which are delivered from shared outcomes of various big data systems through effective data ingestion pipelines compared with traditional data marts that are extended from institutional data warehouses. This means that our new data marts could have a wide range of data sources to meet an infinitely broader set of requirements, as no single company or institution itself could have so much detailed data belonging to others, even if vast data storage was available.

The work has proved important to combine modern technologies with established technologies to leverage the power of big data management. Data understanding and preparation typically consumes about 50% to 70% of researchers and students time, given generally available data are not properly organised, cleaned and integrated. Our proposed data marts could effectively address those issues, hence, enhancing productivity of academic researchers and students. Furthermore, the results and insights drawn from the data marts could allow local councils and charity organisations a wider view of the society to help them make better, more informed decisions. Lastly, this research offers a cost effective solution for small and medium-sized enterprises to comprehend their business as well as the environment in which they operate to understand and know the trends of the market, its development, and help avoid unnecessary competition.

The scalable foundations have been laid for potential future growth at an accelerated pace, since this research has the capacity to contribute to the research community, academic researchers and students, small and medium-sized enterprises (SMEs), and local councils and charity organisations.

References

1. Siddiqui S. Big data process analytics : a survey. Int J Emerg Res Manag Technol. 2014;3(7):117–23.
2. O'Donovan, P. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. Journal (2016)
3. Meehan, J., Aslantas, C., Zdonik, S., Data Ingestion for the Connected World CIDR 2017. CIDR'17 January 8–11, 2017, Chaminade, CA, USA

4. Bellini, P., Nesi, P., Paolucci, M., Zaza, I., Smart City architecture for data ingestion and analytics: processes and solutions, 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications
5. IBM, http://www.springer.com/lncs, last accessed 2021/07/14.
6. Jaleel Talib, A., Abbas, M.J.: Design and Implementation of Efficient Decision Support System Using Data Mart Architecture,2020 IEEE International Conference on Electrical, Communication, and Computer Engineering (ICECCE).
7. Jenny, T., Set Up the Data Mart, Oracle® Business Intelligence Standard Edition One Tutorial. Oracle(2007).
8. TechTarget, https://searchdatamanagement.techtarget.com/feature/How-to-build-an-all-purpose-big-data-pipeline-architecture, last accessed 2021/07/13
9. Stitchdata, https://www.stitchdata.com/resources/data-pipeline-architecture/, last accessed 2021/07/06
10. Youyung Hyun, Ryuichi Hosoya, Taro Kamioka The Implications of Big Data Analytics Orientation upon Deployment, ICIT 2018: Proceedings of the 6th International Conference on Information Technology: IoT and Smart City, December 2018, pp 42–48https://0-doi-org.emu.londonmet.ac.uk/10.1145/3301551.3301566
11. Li Kun-fa, Chen Jing-chun, Wang Yan-xi Big Data Informatization Applied to Optimization of Human Resource Performance Management, IMMS 2019: Proceedings of the 2019 2nd International Conference on Information Management and Management Sciences, August 2019, pp 12–17https://0-doi-org.emu.londonmet.ac.uk/10.1145/3357292.3357302
12. Baohua Qi Research on Knowledge Management System Construction of High-tech Enterprises Based on Big Data, AICS 2019: Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, July 2019, pp 803–807https://0-doi-org.emu.londonmet.ac.uk/10.1145/3349341.3349516
13. Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, Patricia C. Arocena, Data lake management: challenges and opportunities, Proceedings of the VLDB Endowment, Volume 12, Issue 12August 2019, pp 1986–1989https://0-doi-org.emu.londonmet.ac.uk/10.14778/3352063.3352116
14. Yihan Gao, Silu Huang, Aditya Parameswaran, Navigating the Data Lake with DATAMARAN: Automatically Extracting Structure from Log Datasets, SIGMOD '18: Proceedings of the 2018 International Conference on Management of DataMay 2018, pp 943–958https://0-doi-org.emu.londonmet.ac.uk/10.1145/3183713.3183746
15. Alfredo Cuzzocrea, OLAPing Big Social Data: Multidimensional Big Data Analytics over Big Social Data Repositories, ICCBDC '20: Proceedings of the 2020 4th International Conference on Cloud and Big Data Computing, August 2020, pp 15–19https://0-doi-org.emu.londonmet.ac.uk/10.1145/3416921.3416944
16. Alexey Samoylov, Nikolay Sergeev, Margarita Kucherova, Boris Denisov, Methodology of Big Data Integration from A Priori Unknown Heterogeneous Data Sources, CSAI '18: Proceedings of the 2018 2nd International Conference on Computer Science and Artificial IntelligenceDecember 2018, pp 131–135https://0-doi-org.emu.londonmet.ac.uk/10.1145/3297156.3297249
17. Zuar, https://www.zuar.com/blog/data-mart-vs-data-warehouse-vs-database-vs-data-lake/, last accessed 2021/07/08

18. J. D. McDonald, J.D., Rajagopalan, S., Waizenegger, J.R., Pardo, .P, "Realizing the Power of Data Marts," in *IEEE Power and Energy Magazine*, vol. 5, no. 3, pp. 57-66(2007)
19. Agapito, G., Zucco, C., Cannataro, M.: COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data *International Journal of Environmental Research and Public Health* (2020)