Case Study: Instigation of a Data Mart at the Great Academies Education Trust

Introduction

The Great Academies Education Trust (GAET) is an organisation which oversees the governance of four schools in East Manchester, UK. (*Home | Great Academies Education Trust*, n.d.) The trust acts as the primary budget holder and decision maker for each of the schools under its governance, allowing the institutions to be funded by their own pooled resources as opposed to being individually funded by the government. This structure allows school trusts to pool certain resources for cost savings and allows larger scale projects to be undertaken as the collective funding pool is larger than that for individually run institutions.

The key goals of GAET are outlined in their company values (*Vision and Values | Great Academies Education Trust*, n.d.), to support their academies to become outstanding schools where pupils are valued as individuals and achieve exceptional academic progress to become college ready. To do this the trust aims to support its teachers and support staff with internal and external systems throughout their own academy and other academies in the trust, improving teaching progression and growth as professionals.

One of the key features of the trust is to provide cover support for staff during absences, ensuring that staff can take absences when required and the students consequently do not miss out on those resources. Estimates show that on average the education sector loses 8.9 working days per employee per year (CIPD, 2016). This is reflected in the 2019 annual report for the trust which shows that staff cover payments accounted for 3.7% of their annual renumeration costs (*Our Finance | Great Academies Education Trust*, n.d.). As a Multiple Academy Trust (MAT) however GAET could pool all resource other than just money, allowing staff to also be mobile and able to serve multiple locations and reduce the need for costly cover staffing. To do so however the trust will need to address one of the common issues of MATs which is the diversity of data sources from all its overseen academies, a problem which will only get worse if the trust grows and adds more institutions ('The Challenges MAT's are Facing,' 2020).

This project looks to assess the potential for a data mart approach to be taken for GAET to provide a system to help them address staffing absences across their multiple sites. Additionally such a system will allow the organisation to centralise support staff for all of their schools, allowing for facilities, catering, IT and cleaning staff to be efficiently deployed across sites as required. The project will focus on three main aspects which are briefly introduced below and explored in more detail later in the document, these are:

1. Big Data
    - Big Data is defined by the research data alliance as "the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis" (*Big Data - Definition, Importance, Examples & Tools*, 2019). Typically big data is characterised by the three v's of: **volume** indicating the large amounts of data utilised,

velocity indicating the high speed at which the data is collected and processed, and **variety** indicated the diverse types of data collected from different sources.

2. Data Warehouse
    - A Data warehouse is a database that is designed to enable business intelligence activities (*Definition of Data Warehouse - Gartner Information Technology Glossary*, n.d.). Data warehouses collate data from multiple different sources and separate analytical and transactional analysis, allowing users to maintain historical records and analyse the data to better understand and improve the business.
3. OLAP
    - OnLine Analytical Processing (or OLAP) is software for high-speed analysis of large data sets stored within the data warehouse (*What is OLAP?*, 2021). A key feature of OLAP is the ability to retrieve data from multiple tables stored within a data warehouse to return a multidimensional table, allowing easy analysis/visualisation to enable business decisions.

Data Sources of the Organisation

Data is of the upmost importance to any organisation allowing it to make informed strategic decisions and efficiently monitor the results. The education sector has been identified as being large benefactors of robust data systems, with the UK Department of Education highlighting this as a key feature to reduce staff administrative workloads (*Reducing teacher workload: Data Management Review Group report*, n.d.). Data is especially important in an MAT such as GAET due to the siloed nature of the academies under its management. Accurate, open and responsive data between the sites allows for greater co-ordination and efficiencies to be found that would be very difficult to achieve if all academies operated independently. This section looks at the data sources currently available to GAET, how each is stored, their importance to the organisation and their potential usefulness in a data mart designed at facilitating absent staff cover.

Internal Data Sources

HR Data

For monitoring staff details and history the trust will use standard HR software available to the education market. Interestingly the trust will potentially act as 5 different organisations with HR records for each of the 4 academy sites and the centralised head office, an issue addressed in the ETL system later in this document. Typically HR software is controlled by a centralised HR office with open read access to each individual of their records on file, and contain data such as contact details, attendance records, document records, disciplinary records, and compensation/pension details. Most of this data will be stored as text in a relational database structure but other data sources such as JPEGs (for staff ID photos) and PDFs (copies of signed contracts and documents) will also be present. Individuals will have a responsibility to ensure their records are up to date and the ability to submit change requests to their HR office if any edits are required (e.g. change of contact details or error in the data). HR software is also an ideal storage location for job detail roles within the organisation as all staff will require a role designation along with a pay grade and duties description, this information can be stored as a separate data table to staff personal information and simply assigned to individuals on their records. HR data is an example of slowly changing data as contact details and job specifics are often infrequently changed with potentially years between any edits being made.

HR data is vital to the organisation for many reasons. Accurate record keeping is essential to ensure contact details are up to date for emergency purposes, correct lines of communication are used to individuals and salaries are paid to employees accurately. HR data is also a valuable resource for higher management level decision making such as planning staffing for certain activities, assessment of factors influencing staff retention rates, and audits of staff for legal reporting purposes. For the specific purpose of staff absence covering HR data will allow staff to be screened for suitability to cover other areas throughout the trust.

Facilities Data

Facilities data will be stored via facilities management software and controlled by a site custodian or facilities manager, as with HR data this may consist of 5 different management systems across the different schools contained within the trust and require centralisation. The facilities management software will store site locations, location asset register, occupancy limits, maintenance schedules and records, current works underway and emergency evacuation procedures. Most of this data will be stored as text data in a relational database but other sources such as PDFs (scanned copies of contracts or work permits) will also be present.

Facilities data is essential to the organisation for correct maintenance and job tracking across multiple locations. It also provides vital information for high level management activities such as age of assets for calculating depreciation in annual finance audits or emergency procedure planning across different sites. For the purposes of staff cover facilities data by itself will not be of much use, combined with HR data however it can provide a list of staff which are designated to a given site or able to travel between sites contained by the education trust.

Time and Attendance Data

Time and attendance data is collected via an RFID contact card reader system across the locations run by the trust. This system operates by all personnel (staff and students) having individual ID cards which are used to tap into RFID readers fitted in rooms around the locations within the trust. Each contact is logged on a tracking system to assess how many people are in each site location. Each entry is reset to a new location when an ID card is tapped into a different location or signed out of a location if viewed as inactive for a period of 3 hours to ensure no open logs are maintained, with the data stored as text data within a relational database. Time and attendance data is an example of a big data stream for GAET due to the high velocity of data input into the system.

Time and attendance data is vital for the trust to assess the attendance of pupils in classes as per the legal requirement guidelines set out by the department of education (Department for Education, 2020), as well as providing a complete onsite attendance register if required for emergency evacuations such as fire alarms. Additionally the information helps with high level facilities planning such as room usage tracking to allow for accurate scheduling and planning of future activities. For staff cover purposes time and attendance data will be vital in showing which staff are on which site to potentially provide cover.

Timetable Data

Timetabling is a relatively complex task and as such is often done in specialised software to eliminate the risk of clashes for students and teachers within an institution. Planning and storage of data will be performed by a timetabling officer with final approval at upper management level of each finished schedule. For GAET this will be data that is stored separately across all the individual academies within the trust.

Timetabling is vital data for any education institution as it dictates the flow of students and teachers across its site. Good timetables will allow accurate planning of teaching curriculums, ensure teachers have time for lesson planning and enable students to work in a structured environment (*The Importance of a School Timetable*, n.d.). For GAET it will also allow assessment for vulnerable posts where staff are potentially overworked or departments are prone to requiring high levels of cover if staff members are absent.

External Data Sources

Whilst the business issue we are looking at does not directly require outside data sources to operate that does not mean that they will not be useful in a wider context for the business. For example active monitoring and recording of job adverts for teachers in the local area of the trust on websites such as eteach.com (*Teaching Jobs and Recruitment - eTeach*, n.d.), could allow decision makers to set pay grades based on active open positions in the area at the time to help retain talent and to assess for demand in neighbouring areas to see if they are vulnerable to staff loss to other academies. Monitors such as this may require a separate data mart to the one explored in this project but clearly has direct links to many of the dimensions being recorded.

Data Warehouse

A data warehouse or mart is a single store for information within a business that provides a simple searchable system to allow both high level long-term and short-term business decisions. Many people have released publications promoting the benefits of such a system (Furlow, 2001; Watson et al., 2002; Jayanthi Ranjan, 2012), this section looks to explore the direct benefits such a system could have for The Great Academies Education Trust.

Data Consistency

With all data stored in a warehouse consistency across multiple systems will be a given as the data will be formatted this way before uploading. For an organisation such as GAET this is of additional importance as it is a trust which is made up of several individual schools all of which have the potential to contain several data storage methods for various departments that will provide logistical problems when attempting to search across sites. As the trust expands newly added schools will most likely also have different data storage methods adding to the complexity of data searches. Such complex multi methodology structures can also lead to data loss as some schools may store different information to others or simply not store a particular metric that may prove vital to another school. A data warehouse system will drastically improve on this issue and allow for dynamic high-level decision making within the trust as data from all associated schools will be stored in the same location in a consistent manner.

Speed of Data Retrieval

With all data stored in one location data retrieval becomes much quicker than when spread across several locations. If data were stored at the academies separately for GAET a simple query would require input from at least four different data managers, with potentially more if the query spanned across several databases adding further communication time delays. With all data stored in a central warehouse however a single operator can access data from across all sites seamlessly. For our exploration into covering staff absences this increased speed is essential, allowing for internal cross site cover options to be assessed almost instantly once staff absences have been identified.

Informed Decision Making

One of the largest benefits to any organisation implementing a data warehouse is the increased ability for informed decision making. Often in business scenarios decision making can take place in a vacuum with assumptions made about the wider organisation as the required data to make an informed decision is too cumbersome or costly to collect.

A prime example for GAET would be on budget release for facilities maintenance across its academies. With data dispersed across the academies such a decision would require a report from facilities managers across the academies followed by a session to compare and contrast the results with assumptions made on usage and utility to departments. However using a data warehouse approach all site usage figures across all the academies can be found in the same database for direct comparison for budgetary decisions.

Disadvantages

A project as complex as a data warehouse will also have some negatives that should be carefully considered by any organisation before implementation, namely these will be:

- **Set up cost and time:** Data warehouses and marts are very complex labour intensive endeavours where costs of set up can quickly spiral. The systems can take a long time to implement, so long in fact that large portions of the data required at the start of the project may in fact be considered out of date by the time of completion. Any institution must carefully weight the balance of these risks compared to the potential benefits before implementation.
- **Maintenance cost:** In addition to installation, upkeep of data warehouses may also be expensive and require dedicated staff. Refreshing of data into the warehouse will be costly and careful considerations on frequency will need to be taken. For GAET another important factor will be the potential additional costs this may incur if new academies are integrated into the trust.
- **Data security:** Having all data stored in one central location will come with some inherent security risks, as any data corruption or theft will affect all stored data from the entire business. For GAET this will be of particular concern as one data breach will then affect all academies under the trust rather than individually if their data is stored separately.

Data Warehouse Structure

Having explored the potential benefits of initiating a data warehouse at The Great Academies Education Trust we can put together a schema to represent how the data would be structured to meet our primary goal of providing staff monitoring data to enable speedy internal cover in case of absence. Figure 1 overleaf shows a snowflake schema for GAET utilising data from the HR, facilities, timetable and time/attendance data stores to generate an active fact table monitoring up to date staff activities. This is followed by a detailed breakdown of the elements of the schema, including the granularity required for the data.

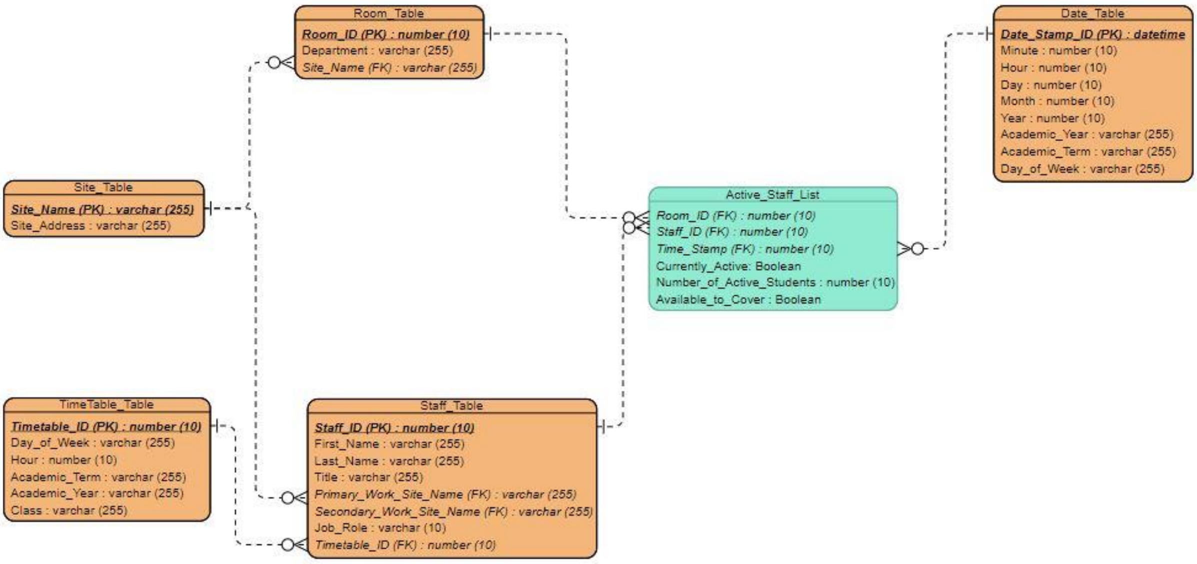
Room_Table
Room_ID (PK) : number (10)
Department : varchar (255)
Site_Name (FK) : varchar (255)
Date_Table
Date_Stamp_ID (PK) : datetime
Minute : number (10)
Hour : number (10)
Day : number (10)
Month : number (10)
Year : number (10)
Academic_Year : varchar (255)
Academic_Term : varchar (255)
Day_of_Week : varchar (255)
Site_Table
Site_Name (PK) : varchar (255)
Site_Address : varchar (255)
Active_Staff_List
Room_ID (FK) : number (10)
Staff_ID (FK) : number (10)
Time_Stamp (FK) : number (10)
Currently_Active: Boolean
Number_of_Active_Students : number (10)
Available_to_Cover : Boolean
TimeTable_Table
Timetable_ID (PK) : number (10)
Day_of_Week : varchar (255)
Hour : number (10)
Academic_Term : varchar (255)
Academic_Year : varchar (255)
Class : varchar (255)
Staff_Table
Staff_ID (PK) : number (10)
First_Name : varchar (255)
Last_Name : varchar (255)
Title : varchar (255)
Primary_Work_Site_Name (FK) : varchar (255)
Secondary_Work_Site_Name (FK) : varchar (255)
Job_Role : varchar (10)
Timetable_ID (FK) : number (10)

Figure 1: Proposed snowflake schema for a staff attendance data mart at the Great Academies Education Trust

Fact Table

A fact table is a section of the data warehouse which contains the up-to-date measurements of the business. Fact tables tend to be small with only the measurements and dimension keys to allow access to the wider stored data (Ponniah, 2004). The fact table for GAET of the **Active_Staff_List** contains the following components:

- **Composite key:** rather than a distinct primary key the fact table contains a composite key of the Room_ID, Staff_ID, and Time_Stamp foreign keys, giving each entry a unique key figuration and providing searchable access to all the dimension tables
- **Facts:** our star schema stores three key facts for the check in of each member of staff as follows:
  - **Currently_Active** is a boolean fact (1 = True, 0 = False) and calculated on the timing of the last active tap in by staff. This will be replaced by false and closed if the entry for each new tap in or if no tap ins have been recorded by the staff member for the last 3 hours
  - **Number_of_Active_Students** is a running count of the number of active pupils checked into the room at the same time as the staff member
  - **Available_to_Cover** is a boolean fact (1 = True, 0 = False) calculated on the amount of free time indicated in the member of staffs teaching timetable for the day (this can potentially be based on a set block of time allowing a staff member time to travel to another site)

Dimension Tables

Dimension tables provide structured labelling data describing the fact measurement data. In the GAET schema there are five dimension tables supporting the central fact table. Details contained within the tables as follows:

- Staff_Table
    - Primary key of a ***Staff_ID*** which is also used as a foreign key for the main fact table
    - Staff name details (*Title*, *First_Name*, *Last_Name*)
    - Available work sites for the member of staff (*Primary_Work_Site_Name*, *Alternative_Work_Site_Name*), this will be defined as *Site_Name* foreign key and indicate where a member of staff is contracted to work on a regular basis and where they will also be able to provide cover if needed
    - *Job_Role* will be selected from a pre-set list of job roles within the academy and will identify the role of the staff member and what they will be able to cover. Some staff may potentially have more than one job role allowing them to cover multiple areas. It is noted that staff may change role as they progress at the academies, the data mart shall handle this by refreshing roles at the beginning of each academic term when job roles typically begin in schools
    - A *Timetable_ID* foreign key will be used to link to the TimeTable_Table for the staff member to allow for easy searching of staff availability. Each *Timetable_ID* will correspond to one teaching period meaning a teaching staff member will have as many IDs as periods that they teach per week. These IDs will then be reset at the start of each academic term as a new teaching cycle begins
    - Note that while the staff table is primarily centred towards teaching staff, non-teaching support staff (for example IT specialists) can also be included with non-relevant fields left as blank allowing for a dynamic support staff system using a centralised source to also be implemented
- Date_Table
    - Primary key of a ***Date_Stamp_ID*** which is also used as a foreign key for the main fact table
    - Time (*Hour*, *Minute*) collected on the 24h clock and used for accurate tap in times
    - Date (*Day*, *Month*, *Year*) to establish the day of each tap in
    - Academic calendar (*Academic_Term*, *Academic_Year*) used to allow granularity down to standard academic term cycles
    - *Day_of_Week* (Mon, Tue etc.) to allow search capability within the TimeTable_Table which is set out as the standard teaching five day week
- TimeTable_Table
    - Primary key of a ***Timetable_ID*** which is also used as a foreign key for the Staff dimension table and represents one teaching period
    - Time of teaching period (*Day_of_Week*, *Hour*) denoting the scheduling of the teaching period
    - Academic period (*Academic_Term*, *Academic_Year*) denoting the academic calendar cycle for the teaching period
    - *Class* identifying which class is due to be taught for the teaching period

- Room_Table
    - Primary key of a ***Room_ID*** which is also used as a foreign key for the main fact table
    - *Department* can be selected from a pre-set list of subject departments defined for the academies and is included to allow searches into room uses for academic subjects
    - *Site_Name* is a foreign key from the Site_Table defining which academy the room is in within the trust
- Site_Table
    - Primary key of a ***Site_Name*** which will be the name of an academy within the trust also used as foreign keys for the Room and the Staff dimension tables
    - *Site_Address* is included as one cell entry as further granularity is not required due to the very small number of sites that are in the trust
    -

OLAP

As previously defined OnLine Analytical Processing (OLAP) is the system used to query the data warehouse and produce relevant data tables for specific business needs (*What is OLAP?*, 2021). In this section we shall explore this by presenting three business questions and generating OLAP queries to produce the required data. All queries shall be demonstrated in the Oracle SQL developer, a commonly used SQL language based programme used to generate such queries (*SQL Developer | Oracle United Kingdom*, n.d.).

Query 1: Finding Immediate Teaching Cover

For the first query we shall explore the main query which the data mart was designed for, looking for available staff to cover an unexpected absence at one of the trust academy sites. In this scenario a list of the available staff to cover the specific subject available at the designated site would be required along with the site they are currently on (to be able to contact them). This is done in the SQL query outlined below using the following steps:

- Join the Active_Staff_List with the Staff_Table on the Staff_ID key
- Join the Active_Staff_List with the Room_Table on the Room_ID key
- Filter the list on the following conditions:
    - Staff are Currently_Active (1 = True)
    - Staff are Available_to_Cover (1 = True)
    - Job_Role is equal to the role needed for cover ('Geography Teacher' has been selected in the example)
    - Primary_Work_Site or Secondary_Work_Site are equal to the academy where cover is needed ('Copley Academy' has been chosen in the example)
- Reduce the data to the First_Name, Last_Name and current tapped in Site_Name of the staff members

SQL Query Code

```
SELECT Staff_Table.First_Name, Staff_Table.Last_Name, Room_Table.Site_Name
FROM Active_Staff_List, Staff_Table, Room_Table
WHERE Active_Staff_List.Staff_ID = Staff_Table.Staff_ID
AND Active_Staff_List.Room_ID = Room_Table.Room_ID
AND Job_Role = ‘Geography_Teacher‘
AND Primary_Work_Site = ‘Copley Academy’ OR Secondary_Work_Site = ‘Copley Academy’
AND Currently_Active = 1
AND Available_to_Cover = 1;
```

Query 2: Assessing Vulnerable Areas for Requiring Staff Cover

For this query we shall explore areas vulnerable within the trust for not having sufficient cover on site during any unplanned absence, to allow sufficient management decisions to mitigate such a risk. To look at this query we shall generate a table which groups staff available to cover by subject, site, and day of the week over the current term and sum the number of available staff for cover. Low numbers will indicate areas of risk as cover has been light up to the present for that term. To generate such a list we shall invoke a SQL query using the following steps:

- Join the Active_Staff_List with the Date_Table on the Date_Stamp_ID key
- Join the Active_Staff_List with the Staff_Table on the Staff_ID key
- Filter the academic year and term of interest (‘2020/2021’ and ‘Summer’ term have been selected in the example)
- Reduce the data to the Job_role, Primary_Work_Site_Name, Day_of_Week and the sum of the Available_to_Cover entries in the list (note that entries where no cover is available are classed as 0 so will not contribute to the sum)
- Group the data on the Job_role, Primary_Work_Site_Name, Day_of_WeekOrder the data by the sum of the Available_to_Cover entries in descending order (this way the most vulnerable areas will appear at the top of the list)

SQL Query Code

```
SELECT Staff_Table.Job_Role, Staff_Table.Primary_Work_Site_Name, Date_Table.Day_of_Week,
SUM(Active_Staff_List.Available_to_Cover),

FROM Active_Staff_List, Staff_Table, Date_Table, Site_Table

WHERE Date_Table.Date_Stamp_ID = Active_Staff_List.Date_Stamp_ID

AND Staff_Table.Staff_ID = Active_Staff_List.Staff_ID

AND Academic_Year = ‘2020/2021’ AND Academic_Term = ‘Summer’

GROUP BY (Job_Role, Primary_Work_Site_ID, Day_of_Week)

ORDER BY SUM(Active_Staff_List.Available_to_Cover) ASC;
```

Query 3: Assessing Room Usage to Revise Maintenance Schedules and Budget

For this query we shall look at the frequency and average room usage figures for each teaching department which would be used to review maintenance schedules for a facilities manager and to help set maintenance budgets for a finance manager within the trust. Whilst the query was not initially planned for in the data mart it demonstrates the power of the system to allow for a range of business decisions to be investigated once a data structure is in place. This query utilises a GROUP BY CUBE function which will allow for multiple breakdowns of the data by individual site, individual department and site department for rapid assessment of the data at various granularity levels. The SQL query required to generate the data for the review is outlined below using the following steps:

- Join the Active_Staff_List with the Date_Table on the Date_Stamp_ID key
- Join the Active_Staff_List with the Room_Table on the Room_ID key
- Filter the academic year of interest (‘2019/2020’ has been selected in the example)
- Reduce the data to the Department, Site_Name, the count of the number of entries featured in the filtered list and the average number of students checked in on the filtered list (note that the average figure is rounded to 2 decimal places to avoid inconvenient very high levels of precision which are presented as a default in SQL)
- Group by cube the data on the Department and Site_Name

SQL Query Code

```
SELECT Room_Table.Department, Room_Table.Site_Name, COUNT(Active_Staff_List.*),
ROUND(AVG(Active_Staff_List.Number_of_Active_Students), 2)
FROM Active_Staff_List, Room_Table, Date_Table
WHERE Date_Table.Date_Stamp_ID = Active_Staff_List.Date_Stamp_ID
AND Room_Table.Room_ID = Active_Staff_List.Room_ID
AND Academic_Year = ‘2019/2020’
GROUP BY CUBE (Site_Name, Department);
```

Extract, Transform, Load (ETL)

The Extract, Transform and Load (ETL) process is the next step after architecture design when implementing a data warehouse, and is vital for facilitating the initial data upload into the system along with subsequent data updates. The ETL process ensures data integrity for the data warehouse by formatting and cleaning the data along with ensuring data from all required sources is compatible and stored in the same format (Zhao, 2017). In this section we shall explore each step of the ETL process in detail and how it can be used for the creation and maintenance of the Great Academies Education Trust data mart.

The extraction step is the first component of ETL where the data is sourced from multiple locations for use in the data warehouse. For GAET this will involve extracting data from all the sources identified earlier in the **Data Sources of the Organisation** section. Typically extraction is done in either a full or incremental fashion (Kakish and Kraft, 2012), described as follows:

- **Full Extraction***:* This is where the entirety of the data is collected from a source for use in the data warehouse. This method requires tracking of changes within the data as all previous data from the source in the warehouse will simply be replaced by the new extraction.
- **Incremental Extraction***:* For this method only a portion of the target data source is extracted for use in the mart, any updates to the source data will need to be tracked so only new or required data is loaded into the data warehouse. This method can either be done on a fixed time schedule (e.g. weekly or monthly) or using an update notification from the source which indicates when further data extraction is required.

GAET will likely utilise both methods for the maintenance and update of the designed data mart. For the initial creation stage of the data mart a full extraction will be utilised to extract the required data from all the sources from all academy sites. If new academies are added to the trust they will also require an initial full extraction of data to be integrated into the functioning data mart. We have also identified slow moving variable dimensions that will require incremental extraction as they are updated, namely the Staff_Table and the TimeTable_Table. As previously stated these will be scheduled for extraction at the beginning of every school term to account for the new teaching timetable and any personnel changes within the academies. Finally the Active_Staff_List fact table will require frequent extraction into the database to allow for any dynamic teaching cover decisions to be made. One tactic for this could be for the system to deliver a notification once a given number of tap ins has been made to load into the system (say 1,000), care should be taken with this data to ensure it is correctly formatted at source so it can be automatically extracted and loaded into the data mart to avoid any data refresh delays.

Extract
Transform
Load

The transform step is arguably the most important in the ETL process as this is where the data is prepared in a useful format for analytic analysis (*ETL Transform — ETL Database*, n.d.). This step has a heightened importance for GAET due to the multi-site nature of academy trusts where historic data collection may take very different forms. Occasionally the data may be loaded directly into the data warehouse before any transforms take place (called the ELT process (Kakish and Kraft, 2012)) but this method comes with risks that corrupted data may be imported directly into the warehouse. A far more secure method is to host the extracted data into a staging area where the data transforms can take place and then subsequently loading into the warehouse. Common data transforms include but are not exclusive to:

- **Cleaning:** adjusting the data for consistency in format, e.g. same date format from all sources, or represented gender as ‘M’ instead of ‘Male’, dealing with blank data entries
- **Deduplication:** Removal of all duplicated data
- **Format Revision:** Adjusting the granularity of the data, e.g. adjusting time measurements to also record seconds not just hours and minutes, modifying the measurement units of metrics
- **Key Restructuring:** Establishing primary and foreign key relationships across data tables
- **Derivation:** creating new metrics calculated from the data e.g. calculating a profit field from raw revenue and cost data
- **Splitting:** Splitting multiple data points in single columns over multiple columns
- **Merging:** Merging separate data points in multiple columns to one column (commonly done on address data if high granularity is not needed)
- **Data Validation:** Assessing for non-sensical data in the set such as typos or extreme outliers

As previously mentioned the data transformation stage for GAET will be of vital importance due to the variety of different data sources and increased risk for data inconsistency. Key features for GAET to look out for at the transformation stage would be format revision between sites, deduplication, cleaning of missing data and key restructuring. Due to the potential high levels of data transformation required the vast majority would need to be done in a staging area before loading into the data mart. The only exception to this would be the previously mentioned time and attendance data, where great care should be taken for the data to be collected in a consistent manner to allow direct loading into the database for quick turnover for analysis when required.

Extract
Transform
Load

The load step is the final stage where the data is imported into the data warehouse. This is typically done using three different strategies (*ETL (Extract, Transform, and Load) Process in Data Warehouse*, n.d.), described as follows:

- **Initial Load:** This is done during the first creation of the data warehouse when all data is loaded into the system
- **Incremental Load:** Changes in data are periodically applied to the data warehouse, either on a fixed interval schedule or with update notifications from the data source
- **Full Refresh:** Existing data from the warehouse is erased and updated data uploaded in its place, typically this would be done on something like a complete stock audit where physically counted figures replace ongoing calculated ones

For the GAET data mart all three of these strategies will be utilised for various data types. As with all data marts an initial load step will be utilised upon creation but for GAET this can also take place when new academies are added to the trust which require a full integration into the ongoing data mart. For incremental loads we have already identified that the Staff_Table and TimeTable_Table dimensions will require an incremental load at the beginning of every academic term to update the data, as well as the Active Staff_List which will require frequent incremental loads as new tap in data is received. Full refresh data loads will be far less common for GAET but there will be instances when required. A particular example will be if one of the academies has major structural changes where departments are shifted to different locations around site buildings, such a scenario would require erasing (or potential archiving if required for future use) of the existing data for the site and a full refresh of the new Room_Table allocations for each department.

Big Data

As described earlier big data is the term used for the large amounts of data generated by modern businesses on a daily basis defined by volume (amount of data), velocity (speed at which the data is received) and variety (different types of data received). For the Great Academies Education Trust one obvious source of big data is that which is collected by the time and attendance log in system utilised in the earlier outlined data mart. This system is responsible for logging room entries of all staff of pupils for attendance purposes and can generate in excess of 5,000 data entries a day (average secondary school size of 1,000 pupils with 5 teaching periods a day (Gov.uk, 2021)). This will not be the only means of big data stored within the academies however, pupil grade scores can also be classed as this due to their frequent updating for each pupil with a variety of potential feedback metrics. Both of these data types are vital to an educational facility in no small part due to their legal obligations to monitor the attendance and progress of their pupils (Department for Education, 2020). Storage of such data will help GAET monitor the key metrics for their business of helping students achieve exceptional academic progress, something which can only be done if academic scores are monitored.

Storage for data such as attendance and grades will need to be done in a transparent fashion allowing pupils to also see their own records, providing the students with the ability to map their own progress. Such systems are quite common for use in the education sector and specialist software such as that provided by Gradelink (*Gradelink Student Information System*, n.d.) can be used to store attendance data generated from the on-site RFID readers and pupils grades generated by the teaching staff.

One of the key roles of big data is to allow management within an organisation to make data led strategic decisions, which for GAET can take several forms:

- Highly accurate attendance data can be used to generate the flow of movement around the academy sites throughout the duration of a day as pupils move between lessons. This information can allow management to assess for high traffic areas within buildings and plan alternative walking routes or lesson plans to ease congestion, an issue which has taken particular prominence in the last year with the global corona-virus pandemic and requirements for social distancing.
- Up to date and historic grading data can identify problem areas in a school's academic record that require improvement and guide strategic decisions to rectify. In particular if this data is coupled with attendance data trends such as class size vs grading can be closely monitored and teaching staff deployment adjusted accordingly to provide the best learning environment for the pupils.

Legal, Ethical, Social and Cultural Issues

Data handling responsibility by organisations has become a hot topic in recent years as developments in the field often occur too quickly for laws and regulations to be put in place at the time of implementation. For the UK the Data Protection Act of 2018 (*Data protection*, n.d.) outlines the legal obligations of companies when it comes to data storage as indicated by the General Data Protection Regulation and sets out 7 key principles:

- **Fairness and transparency:** any decisions made using personal information must be transparent and explainable
- **Purpose limitation:** data collected shall only be used for its designated purpose which people are to be made aware of before submission
- **Data minimisation:** only required data is to be collected, no 'just in case' data collection
- **Accuracy:** there is a responsibility from the data collector to ensure it is accurate and up to date
- **Storage limitation**: personal data should not be stored indefinitely and people have a right to be forgotten and erased from a system unless there is a legal requirement for continued storage
- **Security:** data collectors have a responsibility to ensure all data is stored securely and unavailable for theft, this applies to both digital and physical forms of data
- **Accountability:** an institution must have an accountability system in place for data handling

As an education institute the Great Academies Education Trust will also have added responsibilities as they are dealing with information from minors, in particular having to report all collected personal information annually to the Information Commissioner's Office (Liz Burton, 2018).

On top of legal requirements GAET will have to carefully consider social and cultural issues regarding data collection and usage. For example in the UK children over the age of 13 are allowed to be legally responsible for their own data, but naturally parents will still want to be heavily involved in the safeguarding of their children's information at this age. Bullying and cyber-bullying are also large problems within schools with the BBC reporting up to a fifth of all students have been the victim of bullying at some point (BBC News, 2019). Sensitive personal information can often provide additional opportunities for bullying, meaning extra care and attention should be paid to the security of student information.

Data Management Process

GAET has already undertaken one of the key steps in an efficient data management process in the hiring of a data officer to oversee and provide accountability for the trust (*Our Support | Great Academies Education Trust*, n.d.). From this role key decisions will need to be made for the trust, including:

- Revision cycle for data, to audit accuracy and deletion requirements for stored data
- Data security strategy such as filtering systems to limit data access as required for staff and pupils, monitoring and regulation of data usage within the trust and its academies and education of all staff and pupils of their responsibilities with data usage
- Tracking and logging of personal data usage permissions and ensuring there are no breaches outside of agreed upon fair use

Conclusion

For this report we have demonstrated how a data warehouse strategy can be used by the Great Academies Education Trust for the purposes of reducing expensive staff cover costs across its academy sites. We have also demonstrated how such a system could be an important tool for the organisation in making data driven strategic decisions both directly relating to the main role of the data mart and wider areas of the organisation by utilising the ease of reporting access to the captured information. Whilst this data mart only captures a small part of the overall operations of the trust it could easily be expanded into a comprehensive data warehouse, encompassing all areas of the company.

References

BBC News (2019) 'Bullying: Fifth of young people in UK have been victims in past year - report.' UK. [Online] 11th November. [Accessed on 4th May 2021] https://www.bbc.com/news/uk-50370667.

Big Data - Definition, Importance, Examples & Tools (2019) RDA. [Online] [Accessed on 11th March 2021] https://www.rd-alliance.org/group/big-data-ig-data-development-ig/wiki/big-data-definition-importance-examples-tools.

CIPD (2016) *Annual Survey Report - Absence Management 2016*. CIPD.

Data protection (n.d.) GOV.UK. [Online] [Accessed on 4th May 2021] https://www.gov.uk/data-protection.

Definition of Data Warehouse - Gartner Information Technology Glossary (n.d.) Gartner. [Online] [Accessed on 10th March 2021] https://www.gartner.com/en/information-technology/glossary/data-warehouse.

Department for Education (2020) *School attendance guidance*. Department for Education, p. 22.

ETL (Extract, Transform, and Load) Process in Data Warehouse (n.d.) Guru99. [Online] [Accessed on 4th May 2021] https://www.guru99.com/etl-extract-load-process.html.

ETL Transform — ETL Database (n.d.) Stitch. [Online] [Accessed on 4th May 2021] https://www.stitchdata.com/etldatabase/etl-transform/.

Furlow, G. (2001) 'The case for building a data warehouse.' *IT Professional*, 3(4) pp. 31–34.

Gov.uk (2021) *Schools, pupils and their characteristics, Academic Year 2019/20*. [Online] [Accessed on 3rd May 2021] https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics.

Gradelink Student Information System (n.d.). [Online] [Accessed on 3rd May 2021] https://www.gradelink.com/.

Home | Great Academies Education Trust (n.d.). [Online] [Accessed on 22nd March 2021] https://www.greatacademieseducationtrust.org.uk/.

Jayanthi Ranjan (2012) 'Business Intelligence: Concepts, Components, Techniques and Benefits.' *SSRN Electronic Journal*, 9(1) pp. 60–70.

Kakish, K. and Kraft, T. (2012) 'ETL Evolution for Real-Time Data Warehousing.' *In*. New Orleans, USA: Education Special Interest Group of the AITP.

Liz Burton (2018) 'UK Data Protection Act Compliance: A Free Guide for Schools.' The Hub | High Speed Training. 4th September. [Online] [Accessed on 4th May 2021] https://www.highspeedtraining.co.uk/hub/data-protection-act-compliance-schools/.

Our Finance | Great Academies Education Trust (n.d.). [Online] [Accessed on 24th March 2021] https://www.greatacademieseducationtrust.org.uk/our-finance/.

Our Support | Great Academies Education Trust (n.d.). [Online] [Accessed on 4th May 2021] https://www.greatacademieseducationtrust.org.uk/our-support/.

Ponniah, P. (2004) 'Principles of Dimensional Modeling.' *In Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. John Wiley & Sons, pp. 203–224.

Reducing teacher workload: Data Management Review Group report (n.d.) GOV.UK. [Online] [Accessed on 25th March 2021] https://www.gov.uk/government/publications/reducing-teacher-workload-data-management-review-group-report.

SQL Developer | Oracle United Kingdom (n.d.). [Online] [Accessed on 2nd May 2021] https://www.oracle.com/uk/database/technologies/appdev/sqldeveloper-landing.html.

Teaching Jobs and Recruitment - eTeach (n.d.). [Online] [Accessed on 25th March 2021] https://www.eteach.com/.

'The Challenges MAT's are Facing' (2020) Assembly. 27th October. [Online] [Accessed on 23rd March 2021] https://assembly.education/the-challenges-mats-are-facing/.

The Importance of a School Timetable (n.d.). [Online] [Accessed on 6th May 2021] https://education.gov.gy/web/index.php/teachers/tips-for-teaching/item/2028-the-importance-of-a-school-timetable.

Vision and Values | Great Academies Education Trust (n.d.). [Online] [Accessed on 25th March 2021] https://www.greatacademieseducationtrust.org.uk/vision-and-values/.

Watson, H. J., Goodhue, D. L. and Wixom, B. H. (2002) 'The benefits of data warehousing: why some organizations realize exceptional payoffs.' *Information & Management*, 39(6) pp. 491–502.

What is OLAP? (2021). [Online] [Accessed on 11th March 2021] https://www.ibm.com/cloud/learn/olap.

Zhao, S. (2017) *What is ETL? (Extract, Transform, Load) | Experian*. Experian Data Quality. [Online] [Accessed on 3rd May 2021] https://www.edq.com/blog/what-is-etl-extract-transform-load/.