Designing and Implementing a Data Warehouse

https://drive.google.com/drive/folders/1Bazs7gc2mQLL11LoFrDtNTXJJvFgYtKn?usp=sharing

Background

Galleria Holdings is a fast-food chain which originated in Italy but has now acquired a number of businesses in the UK. The company believes that the UK menus are too large and inconsistent across the outlets they have acquired. They also believe that the menus should be consolidated to suit the preferences of the customers and therefore reduce unnecessary purchases of food which is not chosen by customers.

In order to achieve this, the parent company wishes to create a data warehouse for Business Intelligence purposes.

They wish to analyse sales to determine the most popular menu items and also those which produce the most revenue. They also wish to find out the most popular product groups and those which provide the most revenue. Finally, they wish to construct a “league table” of the total weekly and monthly sales of the various outlets so that their sales performance can be monitored easily.

**The project brief is as follows:**

* Review the case study and identify the business requirements from a reporting perspective. Your overall objective is to design a data warehouse capable of reporting on the questions the business wish to answer.
* Perform an assessment of the initial dataset to:
  + Understand the data items and how they relate together.
  + Identify any data quality issues which could affect your implementation.
* Design the schema to satisfy the requirements.
* Implement the ETL to create and populate the tables in PowerBI ensuring all counts reconcile.
* Create a set of visualisations to meet the reporting requirements contained in the brief.

**Detailed requirements**

In its simplest terms, your task can be broken down into six parts which should be evidenced as a report that includes screen shots and commentary.

Part 1: Problem Domain Understanding

**Task 1**

Given the nature of the business you have been asked to conduct a comparative analysis of on-premises, cloud-based, and hybrid data engineering solutions e.g., ETL and ELT.

The evaluation criteria should consider the following:

* Security
* Compliance
* Scalability
* Efficiency
* Reliability
* Fidelity
* Flexibility
* Portability

|  |  |  |  |
| --- | --- | --- | --- |
| **On-Premise vs. Cloud** | | | |
| **Top Drivers** | **On-Premise** | **Cloud** | **Hybrid** |
| Security | On-premises security might as well be a synonym for “hands-on security.” With this approach, security is you responsibility and yours alone. This means constant monitoring and maintenance.  You get to retain all of your data and remain in control of what happens to it.  Complete on-premises security includes both physical and network security measures, especially if you need to stay within compliance.  On-premises allows you to configure your system the way you like it, but this means you need a high level of expertise.  Separate and sometimes costly security tools are needed to protect each layer of an enterprise.  Security measures and resources are limited by location. | Security is the responsibility of both the enterprise and the vendor providing the software, meaning there’s less of a burden on IT teams  Cloud security is highly automated thanks to APIs, which can mean less work for IT staff members  Since you can access your files anywhere, your security goes everywhere with you, too  Since everything in the cloud is located in a centralized platform security, boundaries are no obstacle  Cloud may offer specialized options otherwise out of reach because of cost.  Public cloud-based services involves trusting a third-party with your most precious data—but on the other hand, they’re experts.  You should ensure your cloud provider allows for compliance with necessary regulations.  Some cloud-based security services are pre-configured, so if you don’t like the way the system is set up, you may not have options to information as and when needed. | **Manage your security risk:** Having public and private storage options lets you lock down your most sensitive and/or highly regulated data in infrastructure under your control while saving on overhead by storing less sensitive data with a third party. |
| Compliance | Many companies these days operate under some form of regulatory control, regardless of the industry. Perhaps the most common one is the Health Insurance Portability and Accountability Act (HIPAA) for private health information, but there are many others, including the Family Educational Rights and Privacy Act (FERPA), which contains detailed student records, and other government and industry regulations. For companies that are subject to such regulations, it is imperative that they remain compliant and know where their data is at all times.  Whether you’re dealing with local, international or industry-specific regulations—staying compliant means you’ll need specialized knowledge and expertise. If you choose to build or continue with your on-prem infrastructure, it can be costly. You’ll have to make sure you properly configure and maintain your systems to stay in compliance, as well as have the right employees and resources at hand to get it done.  Meeting compliance rules means hiring staff familiar with your particular regulations, continuously monitoring systems/logins, creating incident procedures and using data encryption. If something goes wrong, the responsibility falls 100% on your business, as it owns the servers and storage. If you’re audited or fined for being out of compliance, that can add to the cost. | Using a compliant cloud provider can take much of the costly compliance burden off your plate. Unlike an on-premise solution, a top cloud provider has teams of experts who have compliance certifications in dozens of key industries, including health, government, finance, education, manufacturing and media. Not to mention groups that engage globally with governments, regulators, standards bodies and non-governmental organizations.  Enterprises that do choose a cloud computing model must do their due diligence and ensure that their third-party provider is up to code and in fact compliant with all of the different regulatory mandates within their industry. Sensitive data must be secured, and customers, partners, and employees must have their privacy ensured. | **Navigate international data governance:** By taking advantage of a hybrid cloud, especially a multi-cloud environment hosted on different cloud service providers, your organization can more readily comply with privacy and sovereignty regulations like CCPA and GDPR. |
| Scalability | When your on-premises infrastructure can no longer handle its changing workload, you’ll need to scale by adding resources like new hardware and software and increased memory and computing power. Expanding your computing resource requires money, labour, expertise, procurement, hardware, software, monitoring systems and precious time. And if your spike in demand is short-lived, your spend will be highly inefficient. | With cloud storage, there’s no need to rack and stack servers. In the cloud, businesses can scale workloads based on certain metrics as needed—in literally a few clicks—using built-in features. This capabilityallows you to easily scale up, down, out, in—and even auto-scale. This flexibility dramatically cuts your overhead costs associated with monitoring and scaling resources manually. |  |
| Efficiency | Most on- premise setups, resources cannot be rapidly provisioned to scale up or down based on real time need. This can affect performance, overall efficiency and general output in case of high demand or peak season | Increased efficiency is a major difference between on-premise and cloud. Integrated automatic data back-ups and recovery come as standard. Cloud providers maintain software updates, meaning subscribers will always have the latest technology at their fingertips. Flexibility enable companies to instantly react to changing customers, markets, and technology requirements. |  |
| Reliability | Some businesses like the idea of on-premises because it does not require an internet connection for onsite employees to access storage, making it fully available without the need for a good connection. However, recognize that having all of your data on the server limits access to only those working in the office—excluding today’s many remote workers. In addition, on-premises will require power and backup power (such as a generator), and a storage backup system, which will add to your costs. | A good, fast, reliable internet connection is critical when it comes to cloud storage. No connection means no access to your files—and a slow connection can be equally daunting. A break in connectivity can delay your operations and send productivity into a slump. Be sure you’re happy with your internet connection before switching to cloud storage. Many opt for a backup internet connection if all (or most) of their workload is in the cloud. |  |
| Fidelity | Trust is a given, and run from within the organization  Reproducibility protocol will be enacted and managed by the organisation. | Trust has to be extended to the service provider  Reproducibility will be managed by the cloud service provider |  |
| Flexibility | Features can be incorporated as per the needs of the business. Applications or resources can be prioritized, based on hierarch. Customization can be carried out whenever desired | High-level functionality, but some features might not align with business requirements.In the event of any problem, redressal will be carried out irrespective of organizational hierarchyPlenty of choices in terms of customization, although not necessarily business-specific |  |
| Portability |  |  |  |
| Functionality |  |  |  |
| Capability | Resources with specialized knowledge are required to operate and maintain on premise applications | Service providers employ staff with niche skills, ensuring efficient operations and seamless delivery. |  |
| Cost | Investment in assets is a must, which will then be subject to wear and tear, and will gradually have to be replaced.  During peak periods, substantial expenses will be incurred to meet demand | Subscription based model is followed and can be categorized as operational expenses; further, there will be tax benefits |  |
| Maintenance | Slightly difficult to monitor the systems around the clock  In the event of a breakdown, remedial measures have to be implemented | Systems will be under the scanner 24 x7.Financially backed SLAs are provided, resulting in credit for downtime and eliminating the headaches associated with restoring services |  |

**Task 2**

Review the case study and identify the business requirements from a reporting perspective and identify additional relevant business questions for reporting purposes.

Part 2: Data Understanding

**Task 1**

Perform an assessment of the initial dataset to:

* Understand the data items and how they relate together.
* Identify any data quality issues which could affect your implementation.

Part 3: OLAP Schema Design

**Task 1**

Identify and justify your OLAP schema design.

**Star schemas** are optimized for querying large data sets, data warehouses, and data marts. They support OLAP cubes, analytic applications, ad hoc queries, and business intelligence (BI). They also allow you to quickly aggregate a large number of fact records using count, sum, average, and other methods. These aggregations can be filtered and grouped (sliced and diced) by users.

**Task 2**

Design and implement the schema in PowerBI to satisfy the requirements.

Part 4: Data deliverables

**Task 1**

Produce a document (Source to Target Mapping) that contains the mapping of source system fields to the fields of the target system.

## Steps Involved in Source to Target Mapping

You can map your data from a source of your choice to your desired destination by implementing the following steps:

* [Step 1: Defining the Attributes](https://hevodata.com/learn/source-to-target-mapping/#step1)
* [Step 2: Mapping the Attributes](https://hevodata.com/learn/source-to-target-mapping/#step2)
* [Step 3: Transforming the Data](https://hevodata.com/learn/source-to-target-mapping/#step3)
* [Step 4: Testing the Mapping Process](https://hevodata.com/learn/source-to-target-mapping/#step4)
* [Step 5: Deploying the Mapping Process](https://hevodata.com/learn/source-to-target-mapping/#step5)
* [Step 6: Maintaining the Mapping Process](https://hevodata.com/learn/source-to-target-mapping/#step6)

### Step 1: Defining the Attributes

Before data transfer between the source and the destination begins, the data to be transferred has to be defined. This means defining which tables and which attributes in those tables are to be transferred. If data integration is being performed, the frequency of integration is also defined in this step.

### Step 2: Mapping the Attributes

Once the data to be transferred has been defined, it has to be mapped according to the destination system’s attributes. If the data is being integrated into a data warehouse, some amount of denormalization would be required, and hence, the mapping would be complex and error-prone.

### Step 3: Transforming the Data

This step involves converting the data into a form suitable to be stored in the destination system and homogenized to maintain uniformity.

### Step 4: Testing the Mapping Process

Once the first three steps have been completed, it has to be tested on some sample data sources to ensure that the right data attributes in the proper form are mapped correctly with the destination system.

### Step 5: Deploying the Mapping Process

Upon completion of testing and successful data transfer, migration or integration can be scheduled on the live data as per the user’s requirements.

### Step 6: Maintaining the Mapping Process

This step is only required for data integration since migration is a one-time process. Data integration will take place regularly after certain intervals of time. Hence, the Source to Target Mapping process must be maintained and updated periodically to handle large datasets and any new data sources if required.

**Task 2**

Create the tables in PowerBI.

**Task 3**

Extract Transform and Load the data using PowerBI.

Part 5: Test approaches

**Task 1**

Perform some tests to check the four different levels of data testing.

**Entity Level:**

Review the reporting requirements that were listed as part of the initial data warehouse project brief. Can your design support the requirements?

**Record Level:**

Run some queries to get the counts for each of your populated tables.

Look back at the original source data and analyse what each entity count should be.

Do your counts align? If not, consider what needs to be fixed.

**Column Level:**

Check your table structures against the original source data. Have you accounted for all the data items?

**Column Value Level:**

Run some queries to get frequency counts of your columns. As a minimum, pick one column per table and do a frequency count of that column.

Do the same in your source data. Do the values match?

Part 6: Measures and Visualisations

**Task 1**

Create visualisations to meet the original reporting requirements. Include any additional visualisations to support the extra business questions you identified in Part 1.

**Task 2**

Identify and create measures in Power BI for non-additive and semi-additive facts, so they can be used in visualisations and reports.

Useful Links

[Power BI Desktop Playground](https://cloudacademy.com/lab/powerbi-desktop-playground/)

[Visualisation tips](https://www.elsevier.com/connect/a-5-step-guide-to-data-visualization)

Marking Scheme

The skills evaluated within this project are described within the SFIA 8 framework; please see https://sfia-online.org/en/framework for further information.

The skills which this project will evaluate are the following:

* Data Engineering
  + Evaluating, designing, and implementing on-premises, cloud-based, and hybrid data engineering solutions.
  + Structuring and storing data for uses including — but not limited to — analytics, machine learning, data mining, and sharing with applications and organisations.
  + Integrating, consolidating, and cleansing data.
  + Building in security, compliance, scalability, efficiency, reliability, fidelity, flexibility, and portability.
* Data Visualisation
  + Presenting findings and data insights in creative ways to facilitate the understanding of data across a range of technical and non-technical audiences.

Submissions

Completed project report should be emailed to: [AcademyQAC@qa.com](mailto:AcademyQAC@qa.com).

The email should be titled “DFEDATA2 Final Project”.