Understand Business Use of Data Objective

**Purpose**

To understand how data is used to decide the most appropriate technology for a task.

**Problems highlighted by the data**

1. Size
   1. So big processing time is immense – often timing out. Some reports are around 16TB in size.
2. Inaccuracies
   1. Due to human input inaccuracies are throughout the dataset, such as dates not put in in the same format (e.g. mm-dd-yyyy and dd-mm-yyyy). It is time consuming to check for these.
   2. 80% of Savings analysts time is spent extracting and cleaning data.
3. Not meaningful to anyone not fully to grips with everything
   1. Lack of definition of codes and terms, meaning not everyone can easily understand the data.
   2. Most of the organisation need to gain access to this, a lot of who haven’t been trained on the codes or shorthand used in this dataset.
4. Hard to interpret
   1. Not visualised – all data kept in a massive spread sheet that takes time and patience to go through.
5. Spread across multiple systems
   1. Not all contained in one system – spread across oracle and different excel files.

**Things to consider when looking for a solution**

* What does it mean to be the best solution?
  + Most suitable solution for the company looking at – in this case NHS SCCL. Might be different for each company since imagine it would be different per company.
* What criteria should be used to decide this?
  + Does anyone in the team have any previous experience with tools? If not how easy is it to implement?
  + Cost
  + Adoptability – where is the data stored?
  + Scalability – how easy is it to integrate in more data?
  + ETL tools capabilities – can you accomplish these types of tasks quickly?
  + Visualisation techniques
  + Processing time
  + Manpower required to run it – no point implementing if a new tool will take the same amount of time as an old one.
  + Adaptability

**Best possible solutions**

Identified suitable tools for this within the NHS’s Scope are erwin Data Intelligence Suite, Oracle APEX and Oracle Data Integrator.

|  |  |  |  |
| --- | --- | --- | --- |
| Criteria | erwin DI Suite | Oracle APEX | Oracle Data Integrator |
| Previous Experience | A couple of team members with experience. | A lot of experience with Oracle and APEX is like other Oracle tools, so could be learnt quickly. | A lot of experience with Oracle and Data Integrator is like other Oracle tools, so could be learnt quickly. |
| Cost | Expensive, but in budget. | Free with Oracle which is already owned by SCCL. | Requires a subscription in addition to Oracle, still not that expensive. |
| Adaptability | Quite difficult to start fresh and quickly remove previous work. | Easy to start over and adapt to new situations. | Easy to start over and adapt to new situations. |
| Adoptability | Easy to pick up and has a glossary for ease of use. | Easy to pick up. | Easy to pick up. |
| Scalability | Is made to deal with vast quantities of data – easy to integrate with multiple systems. | Easy to add in Oracle datasets, however hard to integrate anything else. | Easy to add in Oracle datasets, however hard to integrate anything else. |
| ETL Activities | Very good at picking out errors in data and performing ETL activities. | Not ideal for ETL activities, does have some functionality but it is designed for App and Web development mainly. | Designed for ETL activities for Oracle Datasets. |
| Visualisation Techniques | Report creation easy – charts and summaries provided. | Access to a wide variety of visualisation techniques. | Limited range of visualisation techniques. |
| Processing Time | Used to running large datasets so very quick to accomplish this. | Not quick at processing large datasets – tends to time out. | It is okay at handling large datasets. Not the fastest but can cope with in. |
| Manpower | Requires 3 main roles to run it, however multiple users can access at any one time. | Requires 1 person to constantly be overseeing it. Hard to have multiple users on it. | Requires 1 person to constantly be overseeing it. Hard to have multiple users on it. |

Therefore, I would recommend erwin Data Intelligence Suite.

**Colleagues examples**

* Discuss with colleagues how they make these decisions; can they give you any specific examples? Including if you can find any examples of the wrong choices: what led them to believe it was the best choice? Why was it wrong? Could they reasonably have foreseen the problems?

Paul Comis:

* Analysis to correlate business increased activity on the mainframe – done in excel

David Brooke:

* System Upgrades or moving to a different application – Rolls Royce, back in the 90s investigating on bespoke mainframe applications – too large.
* People have retired who wrote them – hard to understand
* Large database put into another system – buying off the shelf packing solutions
* Hard to make it fit whilst preserving. Mandatory fields prove problematic for missing data
* ETL was used to address it – multiple extracts merged in the transform step.
* User errors high – not specific enough mandatory fields
* 2 or 3 times to upgrade