Data Warehouse Modelling

**ie university Master in business analytics & big data**

Bi & DW | Australian federal contracts

GROUP f Addison P., Alberto I., Daniel R., Fidelis M., Vania C.

2020

Table of Contents

[**Data Set Analysis** 2](#_Toc33888354)

[Data Set information 2](#_Toc33888355)

[Definition of user requirements. Business questions 3](#_Toc33888356)

[Detection of potential data quality issues 5](#_Toc33888357)

[**Data Warehouse Approach Selection** 9](#_Toc33888358)

[**Data Warehouse Design** 10](#_Toc33888359)

# **Data Set Analysis**

### Data Set information

**DESCRIPTION**

Australian Government contract spending since the 2007 fiscal year.

**LINKS TO SOURCE:**

* <https://www.tenders.gov.au/>
* <https://data.gov.au/dataset/ds-dga-5c7fa69b-b0e9-4553-b8df-2a022dd2e982/details>

The fields in the data set and their corresponding details are the following:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Column Description** | **Data Type** | **Number of Unique Values** |
| agencyname | Name of government agencies | varchar | 155 |
| value | Contract value in Australian Dollars | float | Not Applicable |
| suppliername | The name of awarded suppliers | varchar | 54 076 |
| description | Description of the contract | varchar | 160 851 |
| publishdate | Date when the contract was made public | date | 1 856 |
| contractstart | Contract period - start date | date | 3 075 |
| contractend | Contract period - end date | date | 3 386 |
| procurementmethod | Procurement method - open, select or direct. Competitive public procurement methods used for acquiring goods, services and infrastructure works. Executed in accordance with established procedures set out in the procurement guidelines and detailed in the standard bidding documents. | varchar | 3 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Column Description** | **Data Type** | **Number of Unique Values** |
| category | Category of service where contract belongs to | varchar | 2 354 |
| agencyabn | ID corresponding to Agency Australian Business Name (ABN) | char | 125 |
| supplierabn | ID corresponding to Supplier Australian Business Name (ABN) | char | 21 072 |
| categoryunspsc | ID corresponding to category name.  UNSPSC - Category: United Nations Standard Products and Services Code (UNSPSC) | char | 2 315 |
| cnid | ID unique to each contract | char | 241 164 |
| supplierid | ID of suppliers | varchar | 35 147 |
| sourceurl | URL of each record | varchar | 241 164 |

### Definition of user requirements. Business questions

This dataset has been extracted from public information regarding the public contracts awarded by the Australian Government. As this is public information, it could be relevant for several business users, such as:

* **Australian Government**

This Government as any other is composed by several smaller bodies or agencies that have authority to award contracts using different methods: open, select or direct.

* + Open: Any supplier can participate in the process as far as it is compliance with the solvency (economic and technical) requirements.
  + Select: There is an initial phase to select some suppliers based on certain specific requirements, once shortlisted they can participate in the tender
  + Direct: There is no competition, only one supplier is invited to present its best proposal.

Australian Government could use the final DW/DataMart to get the following insights:

* + How many contracts start/end in any specific year?
  + What is the average duration of a contract?
  + Which contracts have the highest value?
  + Which agencies are spending more/less on contracts in any given budget cycle?
* **Internal auditors / internal control**

Every governmental body has an internal control unit. This unit takes cares of transparency and is responsible for quality assurance in the public contracts processes. An internal auditor could use the final DW/DataMart to get the following insights:

* + Which is the most used procurement method?
  + Is there any agency which uses Select or Direct methods more than Open?
  + Which supplier wins most of the tenders?
* **Suppliers**

As a supplier of the public sector, it could be interesting to know who is providing what and at what price to the Australian Government:

* + Which supplier is selling to which agency?
  + Which categories of products and/or services are the most requested by the Australian Government?
  + What is the market share of the suppliers?
  + Which services/products are provided by different suppliers to different agencies?

### Detection of potential data quality issues

Column: **[agencyname]**

**Missing Values:** No Missing Values.

**Error in Values:** No errors.

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Action Taken:** Retain the column

Column: **[value]**

**Missing Values:** No Missing Values.

**Error in Values:** There are 9 records with value equal to 1. Need to validate if this is an actual AUD amount or a default value for missing contract price. Can be an outlier or a data quality issue.

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column and assume that this is a valid default contract price value.

Column: **[suppliername]**

**Missing Values:** No Missing Values.

**Error in Values:** No errors

**Incomplete Values:** Some of the supplier names in the records have incomplete character but pertaining to the same supplier. Example, "3 Dimensional Consulting Pty Ltd", some of the records show "3 Dimensional Pty" only. Another records show "3 Dimensional" only. Some contained numeric values only such as "6326" or "849. Some contain strings like "A" only.

**Duplicated column data:** data is not duplicated

**Actions Taken:** Perform data clean-up of supplier names. Standardize the naming convention to ensure records fall under one and the same supplier.

Column: **[description]**

**Missing Values:** Has Missing Values for 52 records.

**Error in Values:** There are trailing blanks before the start of the first character of the string such as " Spare parts Military Vehicles ". The trailing blanks are different for a number of records. Some have 4. Some have 7.

**Incomplete Values:** No incomplete values.

**Actions Taken:** Remove trailing blanks. Removing them will not only standardize values in this column but also save memory space.

Column: **[publishdate]**

**Missing Values:** No Missing Values.

**Error in Values:** Publish dates are ahead of Start dates

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Don’t include this column, its values are irrelevant

Column: **[contractstart]**

**Missing Values:** No Missing Values.

**Error in Values:** No errors

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[contractend]**

**Missing Values:** No Missing Values.

**Error in Values:** 1936, 1970 could be wrong years. Also, some dates are exactly same as “start date” which is very suspicious.

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[procurementmethod]**

**Missing Values:** Has Missing Values for 62,961 records.

**Error in Values:** No errors

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Missing values in the dataset but present in the URL (example, row 2). Might consider checking the extraction of data from the source.

Column: **[category]**

**Missing Values:** No Missing Values.

**Error in Values:** No errors

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[agencyabn]**

**Missing Values:** Has Missing Values for 1,219 records.

**Error in Values:** No errors

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[supplierabn]**

**Missing Values:** Has Missing Values for 20,021 records.

**Error in Values:** Aside from the 20,021 missing values, there are 6,693 records with value equal to "0". They do not conform with the 11-digit values for supplier “abn”.

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[categoryunspsc]**

**Missing Values:** Has Missing Values for 92 records.

**Error in Values:** There are 14 records that do not conform to the 8-digit values of categoryunspsc. The value is "461823".

**Duplicate Columns (Containing same information):** Same information as Category Name. ID for Category Name.

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[cnid]**

**Missing Values:** No Missing Values.

**Error in Values:** No errors

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Retain the column as it is

Column: **[supplierid]**

**Missing Values:** No Missing Values.

**Error in Values:** Contains values of either [supplierabn] or [suppliername] columns.

**Duplicated column data:** data is not duplicated

**Actions Taken:** Consider to be removed from the dataset as values contain either a supplier “abn” or a supplier name.

Column: **[sourceurl]**

**Missing Values:** No Missing Values.

**Error in Values:** No errors

**Incomplete Values:** No incomplete data

**Duplicated column data:** data is not duplicated

**Actions Taken:** Consider to be removed from the dataset as it is just a link.

# **Data Warehouse Approach Selection**

Due to simplicity of the given dataset, it is practical to choose a star schema dimensional modelling technique as it is a valid approach to derive a model which is easy to understand, which answers the pre-set business questions stated above in a straightforward manner, and which is optimized regarding performance.

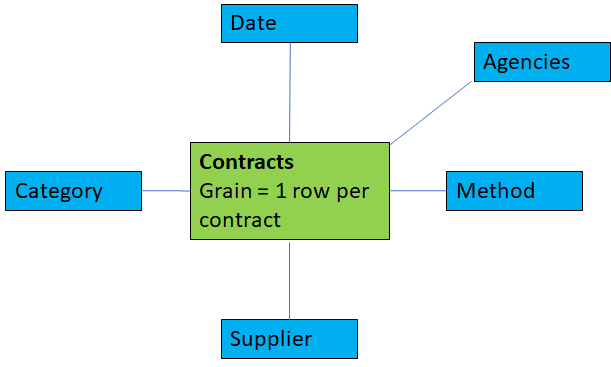
*A star schema is a dimensional model implemented in relational database management systems where a facts table is associated with one or more dimension tables. This model is illustrated as a “star” because the points of the star represent dimensions and the center of the star represents the facts table.[[1]](#footnote-2)*

On the other hand, Data Vault methodology (DV 2.0) is an agile data modelling solution for a system of record. It solves primary business requirements and addresses flexibility and scalability. It is a hybrid data modelling methodology that provides historical data representation from multiple sources that is designed to be resilient to environmental changes. This methodology is mostly suitable for multi-source environments that need a fast adaptation to changes. Looking at our data set we see data is coming from either one Government ERP system or at most three different sources hence we see no need to complicate the design by using Data Vault 2.0 modelling.

# **Data Warehouse Design**

In this specific case for Australian Government Spending, the project team followed a thoughtful process to set up the Dimensional Design:

* First step has been the identification of the facts table. In this case, **contracts**. It is located at the centre of the Star Schema.
* Secondly, the team identified the dimensions. These dimensions help categorize the measurement, give context and allow the final user to answer business questions. Dimensions are connected to the fact table and located at the edges of the Star Schema. Dimensions are what users want to sort, group and filter when analysing data. In our case, **dates**, **suppliers**, **agencies**, **contract categories** and **procurement method**.



**High-level bubble chart**

* Thirdly, the team allocated each of the attributes to the respective dimension tables. Attributes are understood as descriptive information related to the correspondent Dimension Table.
* Fourthly, data types for each of the attributes were defined.

As a result of steps 3 and 4 the following tables were obtained:

|  |  |  |
| --- | --- | --- |
| **CONTRACTS (fact)** |  |  |
| ***Tag*** |  | ***data type*** |
| Contract Value (AUD): Value of the contract | value | DECIMAL (2) |
| Description: Description of the contract | description | VARCHAR (200) |
| Contract Notice ID: Contract Notice ID | cnid | VARCHAR (10) |
| Contract Start Date: Contract Start Date | contractstart | DATE |
| Contract End Date: Contract End Date | contractend | DATE |

|  |  |  |
| --- | --- | --- |
| **CATEGORY (dimension)** | | |
| ***Tag*** |  | ***data type*** |
| Category: Category | category | VARCHAR (200) |
| Category: UNSPSC - Category: United Nations Standard Products and Services Code (UNSPSC) | categoryunspsc | CHAR (8) |

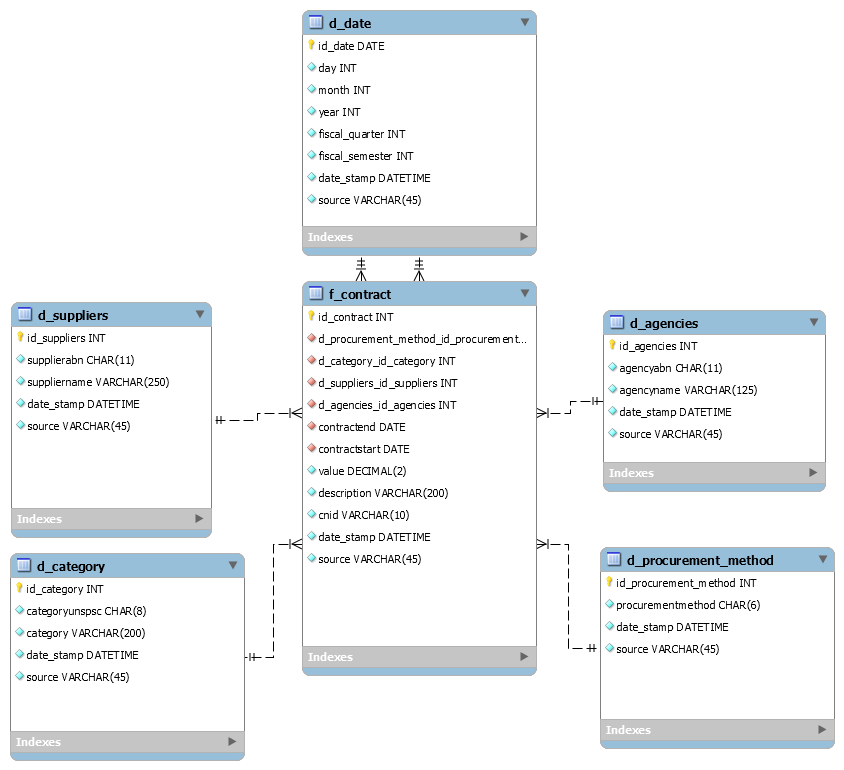
|  |  |  |
| --- | --- | --- |
| **PROCUREMENT METHOD (dimension)** | | |
| ***Tag*** |  | ***data type*** |
| Procurement Method: Competitive public procurement methods used for acquiring goods | procurementmethod | CHAR (6) |

|  |  |  |
| --- | --- | --- |
| **AGENCIES (dimension)** | | |
| ***Tag*** |  | ***data type*** |
| Agency: Agency | agencyname | VARCHAR (125) |
| Agency ABN: Agency Australian Business Name (ABN) | agencyabn | CHAR(11) |

|  |  |  |
| --- | --- | --- |
| **SUPPLIERS (dimension)** | | |
| ***Tag*** |  | ***data type*** |
| Supplier Name: Supplier Name | suppliername | VARCHAR (250) |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **DATE (dimension)** | | |
| ***Tag*** |  | ***data type*** |
| Day | DAY | INT |
| Year | YEAR | INT |
| Month | MONTH | INT |
| Fiscal\_Quarter | Fiscal\_Quarter | INT |
| Fiscal\_Semester | Fiscal\_Semester | INT |

Once all this information became clear, the team used MySQL Workbench to implement the final design.



Note that:

* “d\_” and “f\_” were added to the dimensions and facts tables to make the model clearer.
* Each of the dimensions tables contains a primary key (**id\_date**, **id\_suppliers**, **id\_agencies**, **id\_category** and **id\_procurement\_method**). The same occurs for the facts table, that includes its own primary key (**id\_contract**).
* Metadata is added to each of the tables in the model. See the **date\_stamp** (to account for the specific time when the data was created) and **source** (to account for its System of Record origin).
* The primary keys from the dimensions are added to the facts table as foreign keys (relationships) and resemble a star.
* Facts table **f\_contract** has several dates contract start date and contract end date), each of which is represented by a foreign key to the date dimension (**d\_date**). Each reference is linking to a logically distinct role for the dimension. These separate dimension views (with unique attribute column names) are called roles.
* Initially, it was considered the use of a natural key or business keys as a primary key for tables. For example, we could have used supplier abn, agency abn and category unspsc as natural keys with a relevant meaning to the business. However, it is not recommended because of performance and future problems if any key is reused after some time. That is why it was decided to use Surrogate Keys (integers) that are assigned sequentially in the dimension table.
* Data usage for our model:
  + URL was removed as an attribute, as it does not contain relevant information.
  + supplier ID was not used, since it is a duplicate of supplier name and supplier abn.
  + Publish date not used for being considered irrelevant.

1. Kimball Group (2013). Kimball Dimensional Modeling Techniques [online]. Available at: <http://www.kimballgroup.com/wp-content/uploads/2013/08/2013.09-Kimball-Dimensional-Modeling-Techniques11.pdf>. [Accessed 2020-02-10]. [↑](#footnote-ref-2)