
CollectionHouse
Group
Collecting Insights
our journey towards analytical maturity
Lucas Carpenter
Senior Business Analyst

Welcome; today's journey

- Introduction – about the Collection House Group
- Drowning in data and debt; our analytical challenges
- Our path to analytical maturity through data discovery
- Putting data discovery into practice
- To maturity and beyond

Who are we and what do we do?

- Banking and Finance, Insurance, Government, Telecom and Utility debt

- 250,000+ active purchased accounts with an aggregate value of $1.5 Billion
- Call-centre based; agents contact customers to negotiate an outcome

CollectionHouse Group

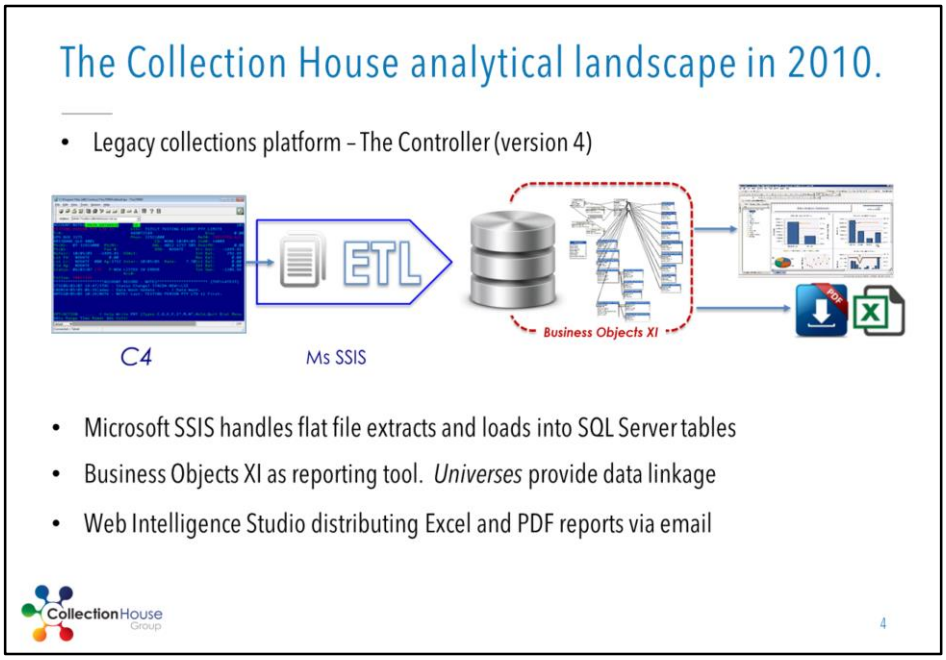
3

Collection House is one of Australia's leading cross-functional receivables manager, operating in debt purchase, receivables management and collections outsourcing.

Debt comes from a wide range of industries including banking and finance – defaulted credit cards and personal loans - insurance, government - national, state and local, a range of fines enforcement and unpaid bills and taxes -, telecom providers and utilities.

To understand the scale of the business in which we operate, in the debt-purchase business alone we currently have just over 250,000 active accounts that have an aggregate face value in excess of $1.5 Billion.

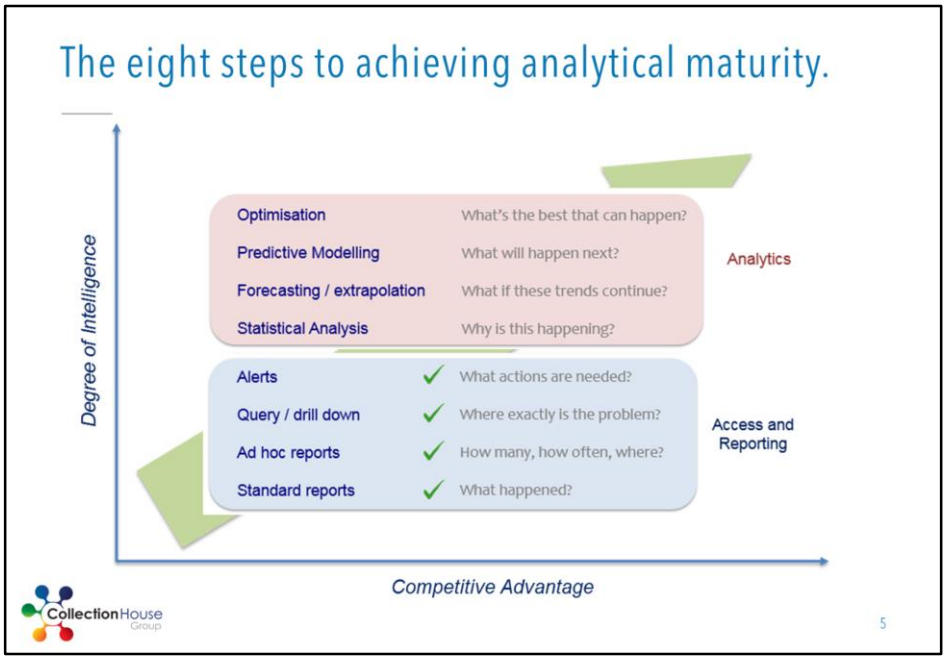
Unlike the traditional "door knocking" image of debt collection, the majority of our contact with customers is through the mail and telephone and now email and SMS. Our Account Representatives (ARs) operate within a typical call-centre environment.



- C4 (Controller 4) was the main collection system since inception within CLH, where all reporting was executed within the prod system, via bespoke scripting

In 2007, CLH moved across to an ETL process, loading flat files from C4, via Microsoft SSIS, into SQL server. With this data now available in an external space outside of C4, Business Objects (SAP) was selected for usage in the management of reporting and analytics within the company.

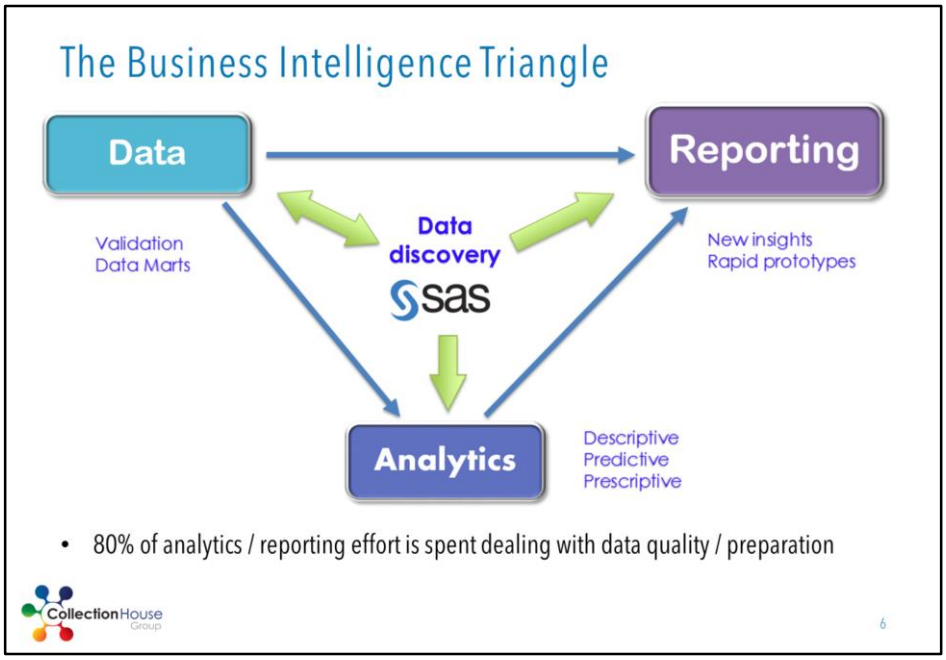
These reports were then scheduled for distribution to end consumers as either PDF files or Excel workbooks, sent out via email.



In his book "Competing on Analytics", Thomas Davenport introduced a concept which broke the path to analytical maturity down into eight stages. Achieving each of these stages would demonstrate a higher degree of business intelligence, leading to greater competitive advantage.

In using BO, CLH was able to achieve the first four steps categorised as Access and reporting -
Standard Reporting
Adhoc Reporting
Data querying/drill down
Alerts

What was missing however the next steps – categorised as Advanced or Genuine Analytics.
Statistical analysis,
where forecasting based on that analysis,
modelling to identify patterns in historical activities that would predict future outcomes
fine tuning of activities to deliver the optimal outcome according to that analysis.



I believe that there are three key elements that need to be handled successfully to satisfy the demands of any business:

- Reporting is first and foremost.

- Underlying data is key to reporting but the focus on ensuring it is robust and reliable can often be overlooked.

- The third area of focus for an analytics team is genuine or advanced analytics

How do we ensure that all three sides of the triangle are given sufficient attention?

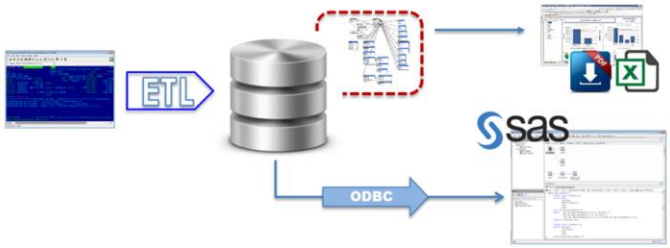
Data Discovery. This is where SAS came in to play in CLH.

A combination of drag and drop tool (using the Enterprise Guide interface) coupled with a powerful, bespoke programming language allows skilled users to extract data from source systems, explore it offline in the SAS environment and uncover new insights about the business.
ETL processes

Statistic procedures such as cluster analysis, regression and data sampling

Introducing SAS – a data discovery proof of concept.

- SAS Enterprise Guide running on virtual machine



- SAS access to existing SQL Server tables through ODBC
- Visual (GUI) and hand coded "data discovery" using Enterprise Guide
- New insights into existing business issues

CollectionHouse Group

7

In order to complete the analytics triangle, we were able to organise through SAS a trial of the Enterprise Guide initially.

This was conducted on a virtual machine, independent from but connected to Collection House systems and accessing the underlying SQL Server data tables through ODBC.

Once established, we were able to demonstrate over a period of weeks the power of SAS in digging into Collection House's data at a very granular level and providing new insight into the business.

Data manipulations as a proof of concept.

Account	Action Date	Action Time	Action Code	Number Dialled	Direction	Result	Talk Time (s)
46002182547	21-Nov-14	14:43:35	RCUS	0212345678	Outbound	AMD	6
46002182547	1-Dec-14	16:47:18	RCUS	0212345678	Outbound	WPC	0
46002182547	5-Dec-14	14:43:38	RCUS	0298765432	Outbound	AMD	4
46002182547	8-Dec-14	7:12:48	CUSR	0212345678	Inbound	WPC	0
46002182547	18-Dec-14	14:42:33	RGTP	0212345678	Outbound	RPC	224
46002182547	24-Dec-14	10:56:28	RGTP	0212345678	Outbound	RPC	34

36M rows of transactional data

Transform / summarise to 1 row per account.

Account	Calls Last 3 Mths	Calls Last 6 Mths	Calls Last 12 Mths	All Calls	First RPC	Last RPC
46006801831	1	8	8		5-Nov-12	4-Jun-15
46007801858	0	0	1	33	17-Oct-12	17-Oct-12
46002182547	26	33	55	164	19-Oct-12	7-Aug-15
46009701826	0	0	5	39	15-Oct-12	20-Jan-15
46009301885	18	30	69	124	.	.

CollectionHouse Group

8

An example of providing new insight was the generation of “Segmentation Reports” or “Cherry sheets” for ARs. These summarized transactional data to an account level to allow our operators to quickly assess the level of activity that had been performed on individual accounts across different time periods.

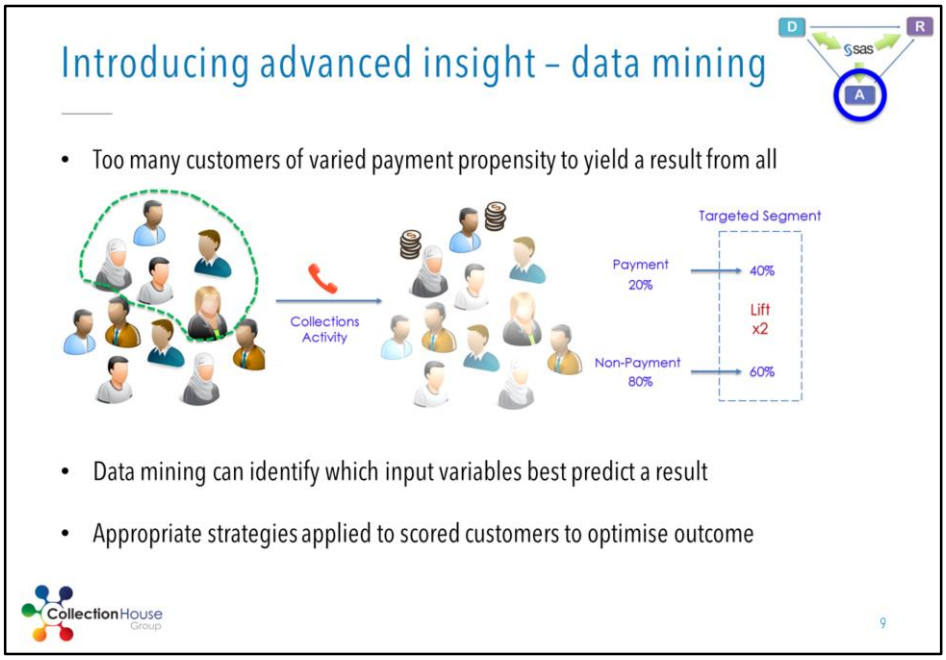
Using SAS to extract data, we were able to take large transactional data sets and condense them into a readable format. For example, the data set that lists individual telephone calls made to each account (with information such as the number dialed, who was contacted, the length of the call and the outcome category) amounts to 36 million records.

SAS is able to read through this data in a reasonably quick time (minutes) and with the application of some simple code, to transpose this data into a row-per-account summary, showing such measures as the number of call attempts over the last 3, 6, 12 months etc, the last time a successful contact was made with the customer, the total time spent talking to the customers or third parties etc.

Using this information, our agents are able to determine if an account has recently been under or over worked, and utilize filters to identify groups of accounts to work –

e.g. accounts where no commitment to pay had been made but we had spoken with a customer within the last three months.

Note that this operation would “time-out” in Business Objects owing to the large number of records being handled (the SAS data sets produced were 2-300,000 records by more than 100 variables).



Just as predictive modelling and data mining have a role elsewhere in the financial services industry – credit scoring in banking, marketing insights in insurance – there is a need to predict the behaviour of our customers in the collections industry. At the basic level this means the likelihood of a customer to pay, but we also model other factors such as the propensity for customers on an existing arrangement to default or the likely success of legal action.

In the example shown, we have a large customer base – in our case over a quarter of a million accounts – and a limited number of staff to attempt to locate and negotiate with each customer. If we were to handle all customers the same, we will make some recoveries but may miss other potential opportunities that could yield results with a little more effort.

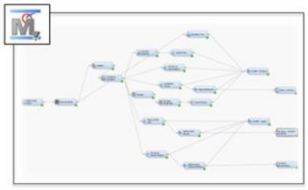
The aim with data mining is to use independent input variables, sourced from customers' transactional history and other metrics that are indicative of behaviour such as credit history, to predict a target outcome – in our primary case the likelihood of a customer paying all or part of their debt within a defined time period.

Once the relationship between these input variables and the target has been

established, it then becomes possible to focus efforts on the customers that are most likely to engage with us to make payments and minimise efforts expended on those customers that are almost certainly never going to pay. Conversely, these models can also be used to identify those individuals that are so likely to pay that they require minimal up front effort (e.g. just a single letter), allowing our agents to focus on those customers that require a little more work but who will eventually pay.

Data mining enables better account segmentation.

- SAS Enterprise Miner is used to analyse data and produce predictive models



- Transformation
- Imputation
- Cluster Analysis

- Decision Tree
- Regression
- Neural Network

- "CH Score" developed – used to drive campaigns and decide account treatments
- Over 180 independent variables assessed; the most predictive are selected by the model

CollectionHouse Group

10

In order to mine our data, we used SAS Enterprise Miner to prepare and model the data using the SEMMA approach (Sample, Explore, Modify, Model and Assess).

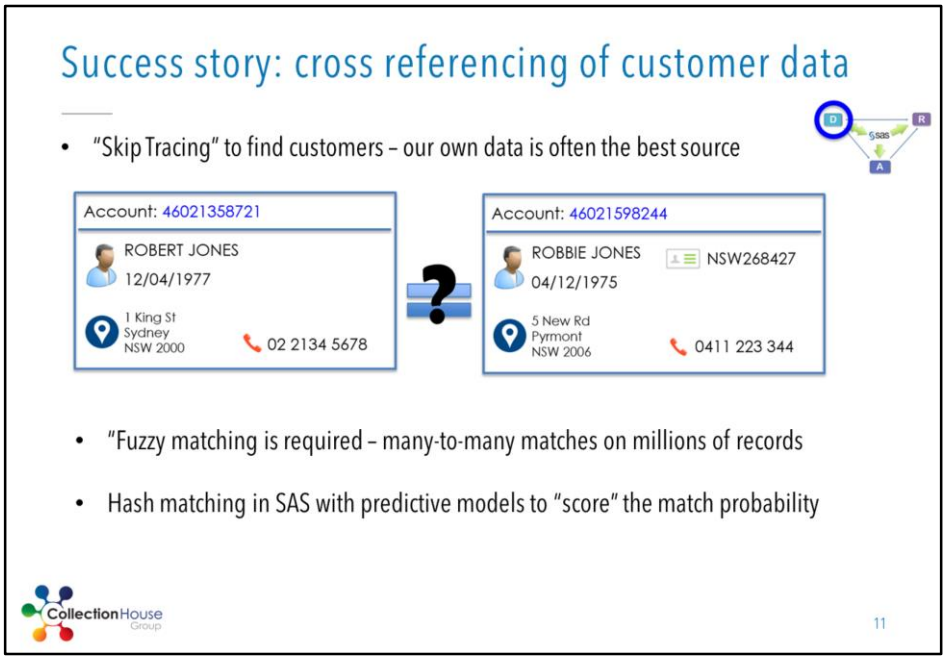
Enterprise Miner allows us to sample and explore the data – to identify variables that are most likely to be predictive, or those that may interact and create over fitted models – to modify the variables – e.g. to transform skewed data or to impute missing values in preparation for performing regression analysis – to generate models – using a range of decision trees, regressions, neural networks either individually or in combination – and to assess the worth of the different models to choose the best performing one.

There are some pretty complex statistics being performed here, but the beauty of Enterprise Miner is that most of it is performed by the modelling nodes, requiring little or no coding knowledge. An understanding of the underlying statistics however is required!

At CH we have built several models that perform different roles – the main one being the CH Score which is used to assess the propensity to pay for individual customers and enable us to tailor our treatments accordingly (e.g. a high scoring account will

receive more contact effort than a low scoring one, as that customer is more likely to pay).

In building these models, over 180 independent variables – mostly derived from customer credit bureau files – were assessed to build an optimal model for ranking the customers by the CH Score.



Skip tracing is a vital part of the debt collection process – finding contact information for customers whose details are no longer current and are therefore considered to have "skipped". Whilst for some customers this is intentional, for many it is not – e.g. if they have moved house and are no longer receiving mail from the service provider.

We search both external data sources – for example credit bureau data, property ownership data, electronic white pages data and of course Google searches – but we also are sitting on a large repository of customer intelligence already; our own data base of existing and past customers.

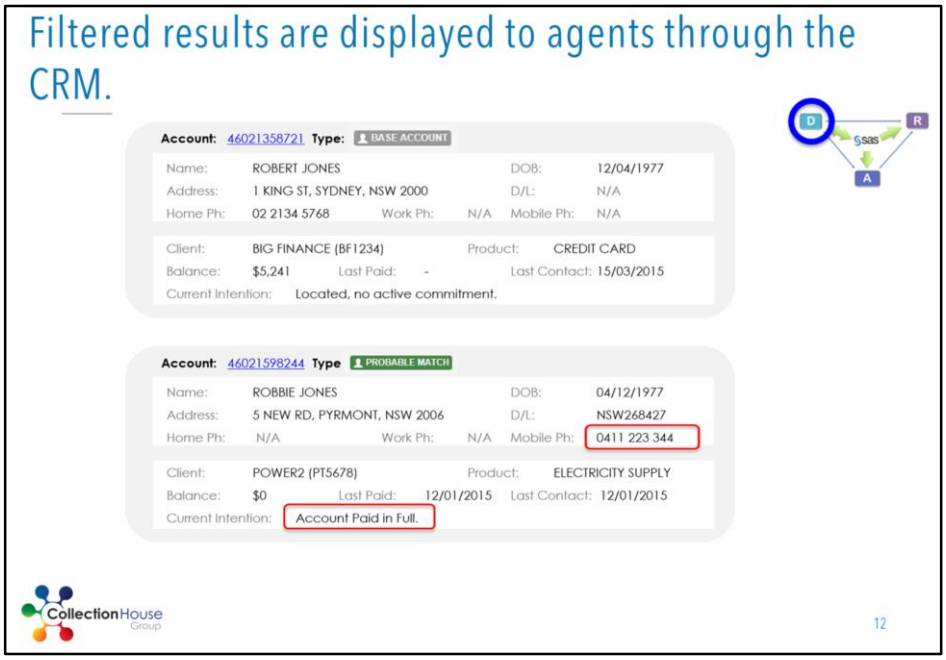
Searching through existing records would previously have been conducted manually, with ARs running a search by name, address etc. on an account by account basis. Added up over the course of months and years, it equated to a substantial amount of time when ARs were performing support activities rather than being on the phone talking to customers. We identified it as a task that would benefit from automation.

Making matters more difficult, matching is not a black and white process. The two examples above could be the same person – the same surname, the same initial, living in close proximity and at first glance having a different date of birth, but this

may be due to the month and day being inverted. A traditional matching process would class this as a non-match, but we would want to consider it as a “possible” match and investigate further. Thus a “fuzzy” matching approach that matches on combinations of similar values is preferred.

The volumes of data are huge. When considering a search of all active purchased debt accounts (300K) against all purchased accounts both current and historic, a full table merge would result in 300 billion possible permutations – not something that would be practical to perform using traditional data base merges. Within SAS we were able to take advantage of the Hash Match function – an in memory process for performing quick merges on large data sets. By implementing this process, we were able to significantly speed up the matching process and extend the range of variables upon which the matching logic was conducted.

In addition, we utilized the predictive modelling capabilities of SAS Enterprise miner to assess the quality of the matches produced and generate a likelihood score of two customers being the same individual. This used credit bureau data as a source of validation to confirm whether the matches that we were able to generate with our own data matched the linking that the bureau considered to be valid.



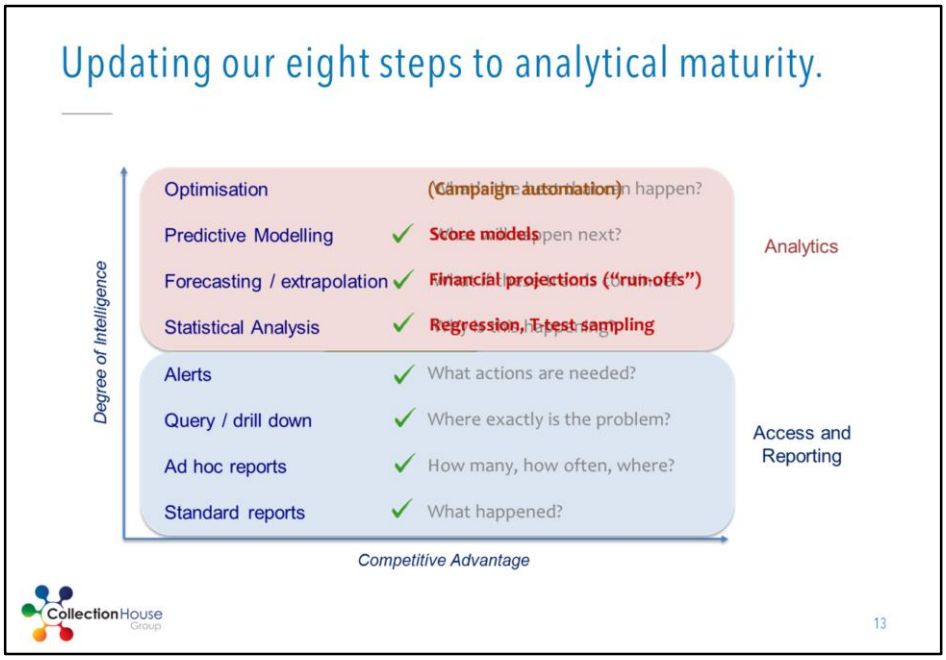
The output data (several hundred thousand rows worth) is loaded into the CRM and an alert flag placed on each account that has one or more matches. Pertinent information is displayed that allows ARs to assess the match quality, identify new leads and assess a customer's potential willingness to pay.

In the example here, the model has assessed this as a probable match (DOB match, same surname and moniker, plus similar locality but different address).

The cross reference reveals a new lead – a mobile phone number for the customer. It also shows that this customer has previously paid an account held by us. This is a good indicator of the customer's likelihood to pay the base account, and combined with the phone number we have discovered, will improve the chances of achieving a good outcome for this account in a timely, efficient manner.

This automation, possible using SAS software has saved us more than 20 FTE by reducing manual cross referencing. It is conducted on a weekly basis – to take account of new, or changed accounts, so that fresh data is available for assessment as accounts are being worked.

My previous manager, Stuart Edwards, has in the past delivered a more detailed presentation on this topic, the link to which will be available in my speaker notes.
http://www.sas.com/content/dam/SAS/en_au/doc/User%20Groups/QUEST/2015-presentations/QUEST-Q2-Cross-Referencing-at-CHL-Stuart-Edwards.pdf



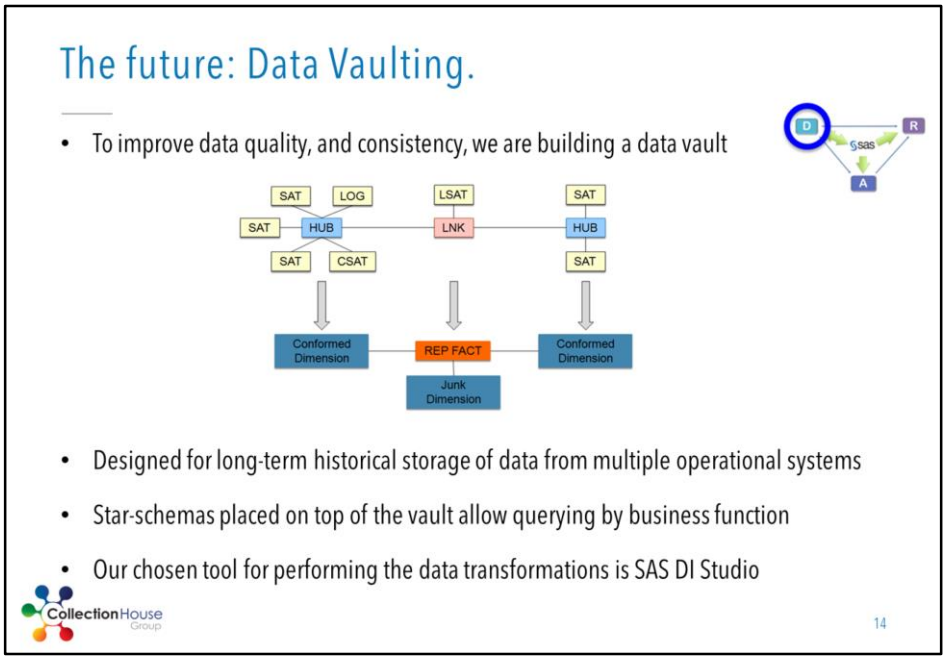
Revisiting our eight steps to analytical maturity, we can see that the adoption of a more robust toolset has helped us to make significant inroads into the analytical space.

We have been able to introduce a range of statistical techniques to the business, making use of the broad range of statistical procedures that ship with SAS.

The ability to apply those statistical methods to large sets of data that we are able to transform and manipulate with ease has enabled us to make projections of revenue streams many years into the future.

The introduction of data mining has enabled us to deploy multiple predictive models that have helped us to segment our portfolio into tranches that can have different treatments applied.

And in future, we aim to apply predictive modelling techniques to help us to optimize and automate in near real-time the way in which our campaigns are conducted.



The next step for us is to re-focus on our underlying data and take the lessons that we have learned from our data discovery exercises to help re-shape the way our information is captured and retained.

We have recently begun a project to build an enterprise data warehouse using the data vault methodology. This is driven by events generated directly from our collections system, rather than by traditional pull-extraction of the data.

The data vault methodology has been selected due to its robustness and flexibility in dealing with long term historical storage of data derived from multiple operating systems.

Once data is captured and validated, star schema views can be placed on top of the data to expose the information to the business for reporting and analysis. By building the vault to well-planned architecture, we will reduce the pain of data preparation that often accompanies any new analysis or report building.

Our tool for performing the event stream driven ETL and building the Data Vault and accompanying star schemas is SAS Data Integration Studio.

In conclusion.

- Identified a need for analysis and data discovery beyond simple reporting
- Arranged with SAS a demonstration of their toolset's capabilities for data discovery
  - Powerful GUI for quick development or code based for flexibility and optimisation
  - Extensive range of built in functions and statistical procedures
- Adopted as long term analytical and prototyping tool to enhance legacy reporting
- Specialist tools – Enterprise Miner and DI Studio – increase analytical maturity

CollectionHouse Group

15

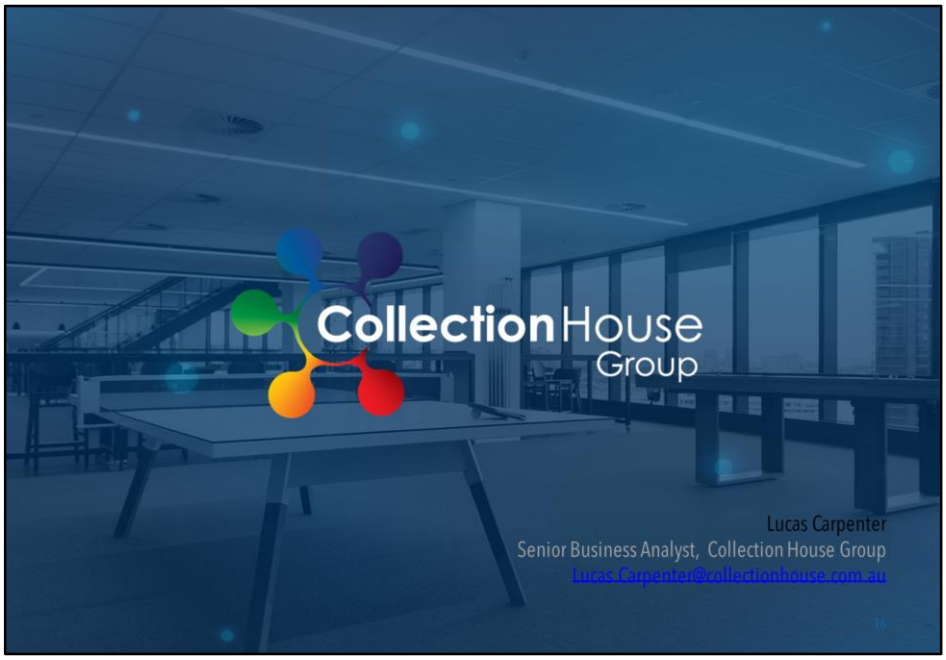
The next step for us is to re-focus on our underlying data and take the lessons that we have learned from our data discovery exercises to help re-shape the way our information is captured and retained.

We have recently begun a project to build an enterprise data warehouse using the data vault methodology. This is driven by events generated directly from our collections system, rather than by traditional pull-extraction of the data.

The data vault methodology has been selected due to its robustness and flexibility in dealing with long term historical storage of data derived from multiple operating systems.

Once data is captured and validated, star schema views can be placed on top of the data to expose the information to the business for reporting and analysis. By building the vault to well-planned architecture, we will reduce the pain of data preparation that often accompanies any new analysis or report building.

Our tool for performing the event stream driven ETL and building the Data Vault and accompanying star schemas is SAS Data Integration Studio.


Collection House
Group
Lucas Carpenter
Senior Business Analyst, Collection House Group
Lucas.Carpenter@collectionhouse.com.au