JTA

THE DATA SCIENTISTS

The Adoption of Telemetry

A Market Researcher's Tale

Contents

INTRODUCTION

This is a story that describes our work with one of our customers. It is a typical project that ends up being very different from what was intended at the outset. The initial steps became a journey. Every time we got to the destination someone asked us what was round the next corner, and the journey continued. We often weren't sure ourselves what the next destination would be but by venturing into the unknown we always managed to get to a better place.

We were working at our client when the discussion turned to telemetry data. The client had recently acquired data which was interesting to our client as it contained insights into competitor products as well as their own products. It covered millions of devices installed in hundreds of thousands of organizations in most countries in the world, but the files were large and hard to handle.

During the first conversations our client explained that the data was delivered in files which were sent to a company to be and returned with weighting applied. Their business issue was that the supplier took 5 weeks to calculate and apply these weights and return the files.

It was against this backdrop that the journey begins.

CHALLENGE ONE: LOAD DATA IN LESS THAN 5 WEEKS

Simple Database Solution

We discussed a simple solution where would create a script to read the files into a database. After import we could run further scripts that would summarise the sample, compare it with known population targets (also stored in the database) and calculate the weights.

Our client liked the approach and they set us a 48-hour challenge. We gladly accepted. After a few weeks of understanding and documenting the data we had a working prototype.

Data was delivered, and we received an email saying that the clock was ticking. Our system processed the data overnight and we were able to easily beat the 48-hour challenge.

The results

When the original supplier sent the files, we compared results and analysed any differences. This analysis confirmed three important facts:

The database method was always consistent whereas the previous manual approach varied each month.

The database allowed us to keep a repository of metadata associated with each organization in the sample. This allowed us to lookup missing data and generally helped with accuracy and ease of use.

Although our original remit had been to just weight and deliver individual files each month, we also maintained a consolidated fact table with all the deliverables combined. This would allow us to bring new and exciting innovations.

Time to Move Away from Excel

Overall, we were pleased with the improvements in accuracy and processing times, but the client still received monthly files which they opened in Excel. Some files had 80 million rows and 100 columns and so were really outside the scope of Excel. The good news was that we had successfully ingested telemetry data for the first time.

We knew that this was a block to unlocking insights, but we also knew that changing this would require a change of culture.

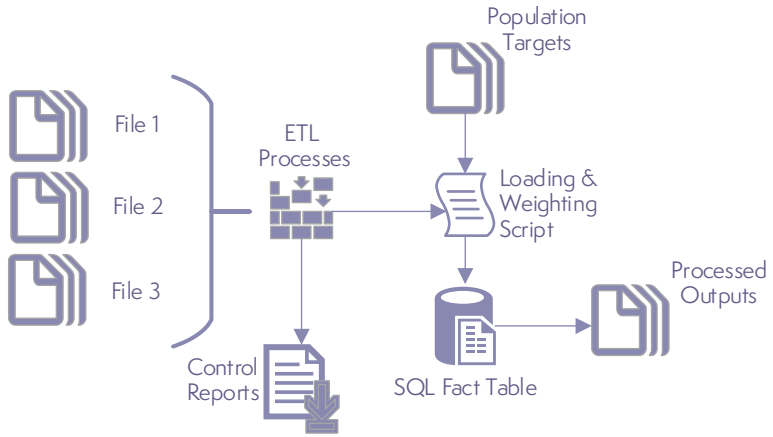


Figure 1 Architecture post challenge one

CHALLENGE TWO: LONGITUDINAL SETS

Consistency over time

Our client wanted to work with sets where each month of data had the same cohort of organizations. The telemetry had new organizations each month. New organizations did not show the same behaviour as older organizations and, as a result, trends were being influenced by changes in the sample.

We also noticed that organizations sometimes disappeared for a month or two and then returned to the sample.

From our client's perspective this was a hard problem to solve. They had an Excel spreadsheet for each month of data. Calculating the longitudinal set would require all the months to be open and Excel could barely open one month.

In our database the problem was trivial: A simple SQL query could count the number of periods associated with each organization and filter any organizations with incomplete history.

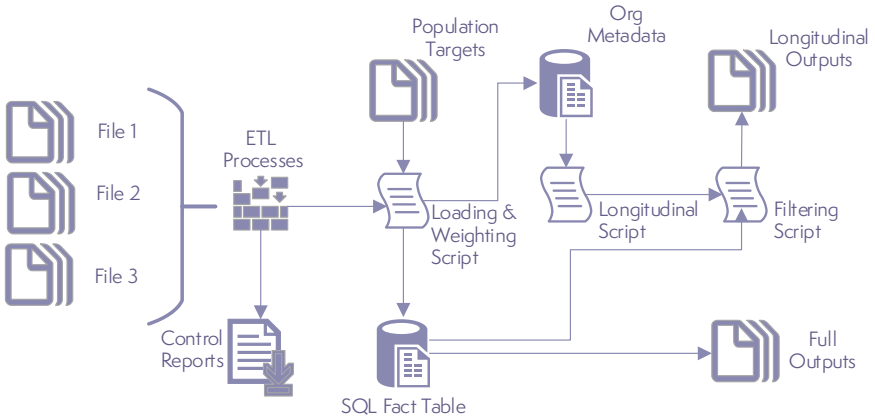


Figure 2 Architecture post challenge two

CHALLENGE THREE: INTEGRATED MODELLING

Business Models need to Change

The clients market models used telemetry data. The data was cleaned every month with SPSS and then merged with other information. The data was weighted manually using Rim Weighting and Bayes Theorem. These models were very important to our client and the next challenge was to see if we could automate their production.

Many scientific and statistical studies use the R language, and this was ideal for our needs. R is open source and it is becoming embedded in many mainstream technologies. R can analyze big data. We judged that it had all the flexibility needed to build the models and it continuously grows in power and sophistication and so is not likely to become obsolete.

Early Experiments with R

Our first experiment was to write a function to perform Rim Weighting. R was able to handle the matrix algebra we needed and could also control the iterative processes used in the technique. We had proved conclusively that R interfaced with a database could do complex modelling. Consequently, our client invited us to automate the models.

By using R, we could drastically reduce the time needed to produce the models to a few hours. This opened opportunities to rerun the model with different assumptions allowing us to finesse the results before publishing. Previously this had not been practical to do.

There were other advantages from having integrated modelling. Users could quickly drill into the results because R sent the results directly to an OLAP cube. We also shared dashboards with the business by linking the cube to Power BI. Finally, telemetry data was revealing insight.

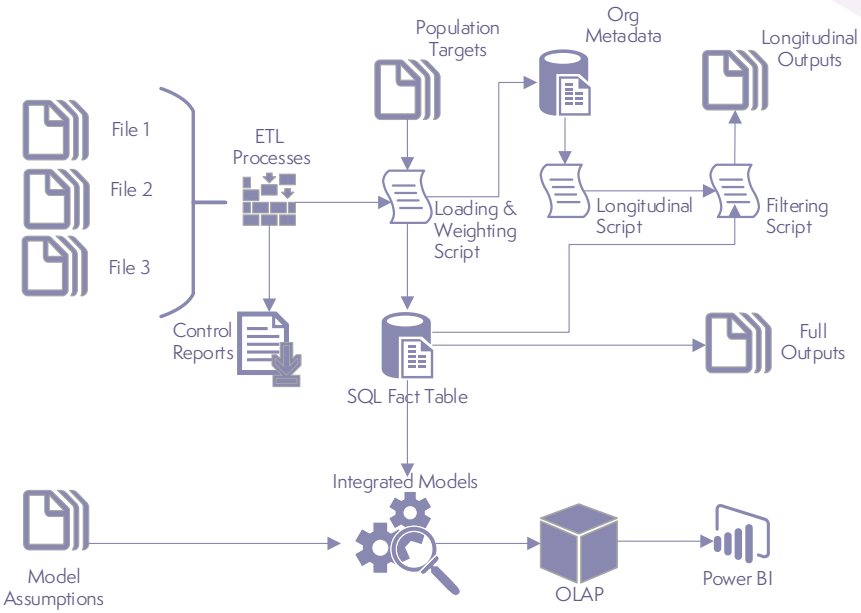


Figure 3 Architecture post challenge three

CHALLENGE FOUR: GIVING UP THE DATABASE

Modern Data Storage Solutions

After some time with the database we started to have conversations with our client about how we could make ad-hoc analysis simpler for the market researchers.

We started by recommending replacing the database with a more flexible Big Data storage solution. Databases are of vital importance in commercial applications but are not ideal for the analysis of large volumes of data. This is because a database has many sophisticated mechanisms that help in transaction processing but can slow down research. A database expects to process transactions whilst guaranteeing not to fail and this requires constant monitoring and backup. The problem is when we want to analyze millions of rows of data at once the overhead of these control mechanisms slows everything down.

Data Lakes

Our suggestion was to move the data into a Data Lake stored in the cloud. Data Lakes allow us to hold both structured and unstructured data and are massively scalable. They also open the door to big data style processing whereby a task may be performed by clusters of machines, all working together.

By adopting a cloud solution, we would be able to scale the computation power by connecting as many machines as we want to the lake. More importantly, we could disconnect machines as well which helps us to make the analysis highly cost effective.

The challenge of coping with the volume of telemetry data was suddenly not so bad.

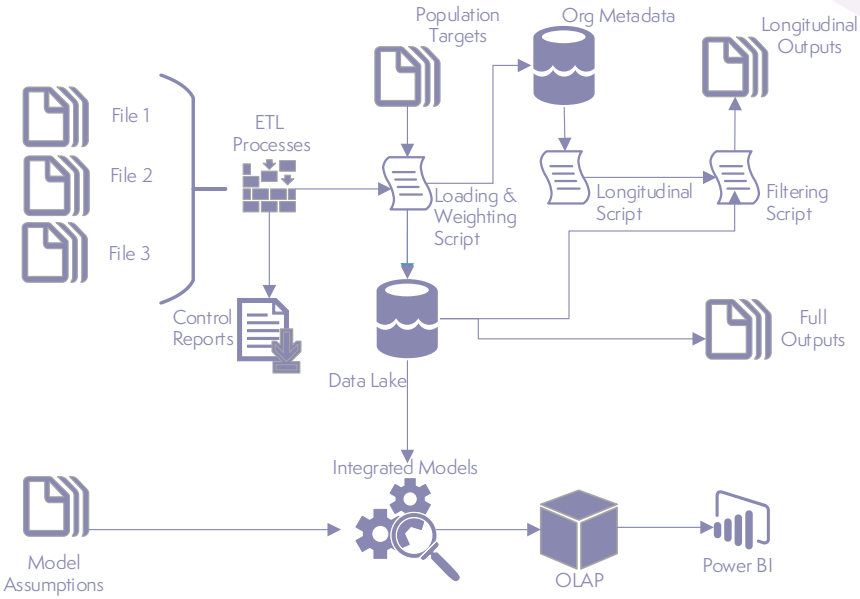


Figure 4 Architecture post challenge four

CHALLENGE FIVE: UNDERSTAND DATA FORMANTS

Conditioned to Use Spreadsheets

Many users prefer a specific format for their files and with good reason. They like to have certain things in rows and other things in columns. Files like this implicitly have two dimensions. This often helps users who like to put totals or averages at the foot of each column using Excel.

When we put the telemetry data into a grid of rows and columns only around 10% of the cells had numbers present. Unfortunately, the remaining cells still had values of zero. Consequently, we were using large amounts of storage to hold hundreds of millions of zero values that were just padding the data.

Quite frequently the number of columns changed from one month to the next. There was a risk that when we joined two files with different columns data would get mixed up.

The vertical Format

The solution was to store everything in one single column. Unfortunately, the client team found our approach hard to accept as it was used to using excel. Fortunately, in challenge eight we had a solution to this by making this single column of data easier to manipulate.

Conditioned to Use Text

We had often used the popular *csv* format for storing files. Whilst there is nothing wrong with this it is not a very efficient mechanism.

Consider, as an example, three digits which can range from 000 to 999. We can have one thousand different values in our three digits. We need only 10 bits of data to store One thousand combinations. However, when we store our values as text each character will take at least eight bits. The three original digits occupy at least 24 bits and so we will have used more than twice the space required by opting for a text file.

It is very rare for market researchers to be concerned with efficient storage and for good reason. Storage has become so cheap and plentiful that one normally doesn't need to think if we are using it efficiently or not.

Data Structures Matter

When we move to a world of telemetry data, however, the volumes of data become so large that we must consider storage efficiencies.

Our solution was to suggest a binary file format that we had measured as saving over 60% of the storage required by *csv*.

CHALLENGE SIX: MANAGE MASTER DATA

Coping with Meta Data

Companies like to view their markets a certain way. They arrange the countries of the world into sales regions, they categorize organizations by industry and by size and they group their products into sets.

Data coming from internal systems is easy to understand and fits the company's view of the world. In contrast data sources externally will not follow the rules. Changing external data to match a company's view of the market is not trivial. Some data may reference the United Kingdom; others may reference England, Scotland, Northern Ireland and Wales. We need business rules to ensure that internal data matches with external data. The rules change incoming data to create one single, consistent view.

The MDR

A Master Data Repository ("MDR") system stores and curates these rules. Typically, the system will be able to let users browse and understand the rules and to make changes to the rules under version control. A log is generated when rules are changed, and approval is often required before the change takes effect.

The MDR and internal systems must be integrated. Firstly, it must supply the business rules to any data import and cleaning processes. It is also extremely useful for the MDR to link to reporting engines. This is because reports also benefit from showing a consistent view of the market. It is very common for business data to be arranged into hierarchies. A sales region may break down into sales offices, countries, states and towns. Defining which countries are handled by which sales office is just another example of a business rule that can be managed by the MDR. When viewing hierarchies, it is great to be able to drill into the detail and to return to higher levels.

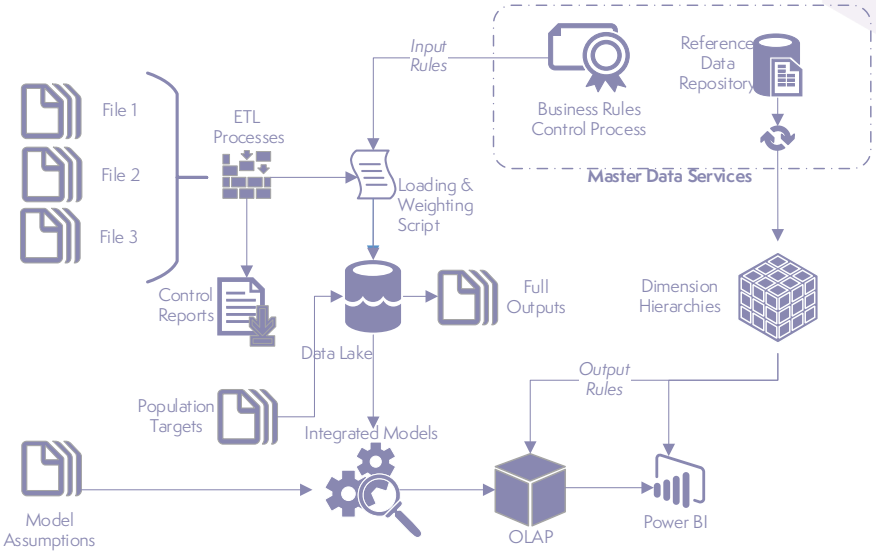


Figure 5 Architecture post challenge six

CHALLENGE SEVEN: SIMPLIFY ANALYSIS OF TELEMETRY

The need for New Tools

By this point in the journey our client had great storage, great processes, great models and it all worked together in harmony. There was still more required, however. The telemetry data had a lot of hidden insight. The client's Market Researchers had some wonderful ideas for lines of enquiry, but they were not altogether happy working with Data Lakes and 'R' to query the data.

Our solution was to build an 'R' repository. It needs only one 'R' command to load this library of functions. Once loaded a user can call any of the functions from a script. We constantly add new functions to the repository along with documentation, help pages and examples of how to use them.

Example Tools

- Fetching and storing files on the data lake
- Publishing reports
- Cleaning and weighting data
- Joining files, filtering files and aggregating files
- Adding new columns and calculations to files
- Computing longitudinal sets
- Starting parallel processes to compute results
- Fitting polynomial regressions
- Fitting probability distributions
- Machine Learning

Essentially what we do is to encase our data science knowhow into a simple tool for others to leverage. We now have many more research managers at our client running ad-hoc investigations to positive effect.

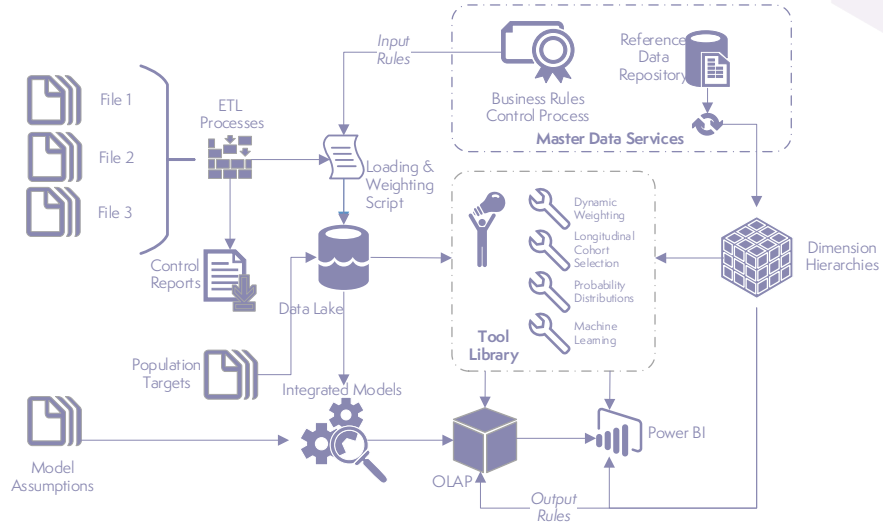


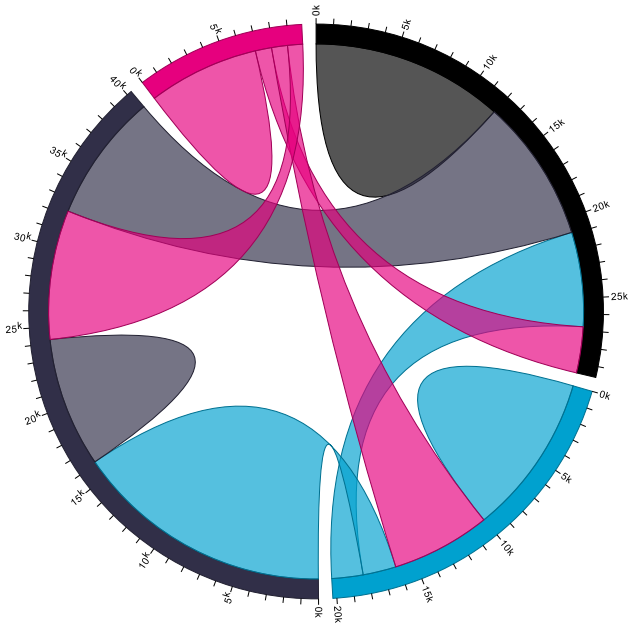
Figure 6 Architecture post challenge seven

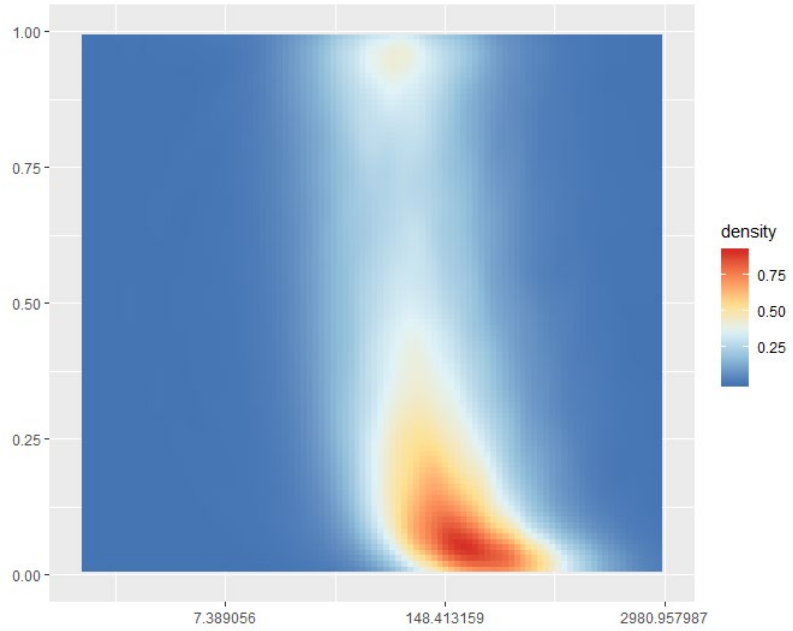
CHALLENGE EIGHT: VISUALIZE TELEMETRY

The Need for Custom Visualizations

A traditional Market Research report is often a series of numbers on a page with extra columns to show growths or declines and variances. That is fine when we are reviewing a calculation of market share but in the world of telemetry data things can get a lot more complicated. What happens when we look at Bayesian decision trees or an ecosystem of various products?

In these cases, we create custom visualizations. These are typically add-ins to Power BI that allow us to create custom-made controls and charts to suit the data and to suit the customer's requirements.





CONCLUSION

At JTA we guide, inspire and support our clients, whatever their industry, to unlock significant value from both their own and sourced data. Data Science projects will often evolve over time. The recipes are similar but the data and the skill in the client's team will decide the best paths to take. During our journey with our clients, value is added at each stage and that value increases over time.

Innovation... that's why JTA exists...

and it's how the company forged itself in the first place. Providing innovative solutions to its clients complex and intricate data dilemmas.

Whilst many service providers within the industry protest to 'think outside the box', JTA think outside the industry. This begins with the specialist team that have been handpicked from a multitude of outer-industry specialisms to join the JTA family and assist them in giving clients the knowledge, answers and power to progress their technologies.

JTA

THE DATA SCIENTISTS

Rua Alexandre Herculano 351, 5º Andar,

4000-055 Porto, Portugal

T: **+351 225 371 539**

E: info@thedatascientists.com