DSBL Capstone

# Step 0 - Introduction. 100-day Data Science Plan: Build a Data Science Strategy

Upon assuming a new leadership role within a company (whether from an internal move or joining the company anew), it is common for an executive to be asked to prepare a plan for their first 100 days in the job.

As part of this project, you will build/create the following:

1. Identification of six data science opportunities for the organization
   1. Opportunities must be spread across three different functional areas
   2. Detail the risks, challenges, and key factors for success for each of these opportunities
2. Prepare a roadmap for executing these six data science opportunities.
   1. Rack and stack evaluation of these opportunities
3. Prepare a Human Capital plan for your data science organization
4. Prepare a Technical plan for your data science organization
   1. Data and Data Architecture Strategy
   2. Machine Learning Architecture

The work product for this Capstone project will be a detailed presentation to the CEO, detailing your plan and the rationale behind your decisions.

This project asks you to prepare that 100-day data science plan for a company of your choosing; this could be your current company or some other existing company.

**Name of Company Chosen:**  AA Consumer Goods Corporation (AACG)

**Brief Company Description:** The company in scope is a local operating company belonging to a large corporation. They sell consumer and household goods ranging from personal hygiene, beauty, cleaning and OTC supplements .

They are a sales organisation which means that the manufacturing plants are elsewhere and products are supplied in the region.

The company operates in one of the European countries and is subject to GDPR and other privacy laws when it comes to data operations.

When it comes to the data maturity level: the company’s HQ have been investing a lot in building the data lake and standardising basic reporting for finance and business intelligence across all markets in the world. The local leadership wants to use data science to go beyond traditional reporting and descriptive analytics and optimize its operations.

# Step 1 - Identify Data Science Opportunities in the Business

Throughout the course, you have been exposed to multiple examples of data science projects implemented in a business setting. Now, based on your knowledge of your specific business context, you will generate six potential projects to be considered by the executive leadership team. These projects must span three unique functional areas of the business, with any one functional area representing no more than 3 projects:

Acceptable Project Mixes

\* 2 marketing + 2 supply chain + 2 finance

\* 2 marketing + 1 human resources + 1 procurement + 1 product + 1 manufacturing

\* 3 finance + 1 legal + 2 marketing

Unacceptable Project Mixes:

\* 3 marketing + 3 finance

\* 4 marketing + 1 product + 1 manufacturing

### Please identify your six projects here:

**Project 1:** Marketing Mix Modelling for a core brand with large investments

**Project 2:** Predicting customer churn

**Project 3:** Short term forecasting automation

**Project 4:** Predicting launch curves for future products

**Project 5:** Supplier Quality Optimization

**Project 6:** Localisation of global ML model for vendor selection

**Note: You may choose to represent this information on slide 5 of the CEO Presentation Template**

## Project 1 Name: Measuring the impact of the marketing mix used in one of the core brands

**Business Functional Area:** Marketing

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

* Business Problem Addressed:

The company is spending large amounts of money yearly in advertising and marketing campaigns across both digital and offline channels (TV, radio, out of home banners, etc). Currently there is little understanding about which of these channels are more efficient and the teams are struggling to explain the underlying sales performance drivers and therefore justify their budgets.

Having a better understanding of which marketing channels are driving sales in order

* Role of data science in addressing the business problem: we would be able to combine data from multiple sources and using the appropriate marketing mix modelling approach provide a very detailed answer to the existing questions around marketing investments
* Targeted business objective(s): Marketing Investments optimization

**2. Data Science Classification**

* Approach:  **Prescriptive**
* Type of Model: Supervised Learning (Bayesian Regression) + Optimization algorithms

**3. Data needed for project and sources for that data**

For this project a mix of data sources will be needed, both internal ( sales, pricing, campaign information, internal events , etc) as well as external which are to be provided by the various marketing agencies we are currently working with.

**4. Magnitude of opportunity (with justification)**

Having a lack of clear understanding around what drives the sales for the product with the biggest revenue or the point of diminishing returns generally leads to bad marketing investments, poor budget allocations and with it missed opportunities.

**5. Cost and complexity of development and implementation**

There will be a low cost as the data is available and always paid for via the agency partnerships. The project will have a medium to high implementation complexity due to the difficulties of getting the data in the right format and with the right granularity.

We need data for modelling (such as impressions from digital channels ) and net spending data to be used for budget optimisation and ROI & profit calculations. This is usually scarce or aggregated.

Moreover, in order to achieve the prescriptive aspect of this implementation, we need to use off the shelf optimization algorithms which may or may not work as expected for our problem.

**6. Likelihood of value capture (Low/Medium/High) with justification**

If the data has the quality and granularity we need and if the optimization algorithms perform as expected, there is a high likelihood of value capture.

Due to the unknowns I would mark this as “medium”. I would expect that even a better understanding of the sales drivers would be of a huge benefit for the marketing teams.

**7. Key Business Stakeholders**

Core business stakeholders : CMO ( chief marketing officer) and the marketing performance team. Extended stakeholders : the finance team.

## **Project 2 Name:** Predicting customer churn for the beauty products online shop

**Business Functional Area:** Marketing

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

* Business Problem Addressed: while in the past, the online beauty shop has been less popular, following the pandemic, a lot of the business has been moved online . The most common items sold are monthly beauty box subscriptions, while purchasing of individual items became more common . The latter has been a great source of revenue in the past year.

While AA CG Corporation sells also through 3rd parties shops such as Sephora, they have changed their strategy in 2020 to focus more on their direct customers (regular consumers) .

* Role of data science in addressing the business problem: data science can help with identifying which customers are most likely to churn . Once identified, the marketing teams can run appropriate campaigns to stimulate retention and loyalty .
* Targeted business objective(s): The overall focus is to retain the existing customers and in parallel run a separate initiative to find and attract new customers.

**2. Data Science Classification**

* Approach:  **Predictive approach**
* Type of Model: supervised learning

**3. Data needed for project and sources for that data**

Customer attributes from the CRM system ( anonymised, following GDPR rules) and data from past purchases.

**4. Magnitude of opportunity (with justification)**

A great opportunity to grow an area of the business which will likely be a core sales driver in the next few years as loyal customers will prefer to shop from the original online shop & get various discounts , rather than going through 3rd parties .

**5. Cost and complexity of development and implementation**

Low cost and low complexity as all data is readily available and reliable as it has been managed internally by the CRM team for the past few years.

This would be a great case study for the local opco to use existing data infrastructure developed by the HQ and autoML tools (SaaS) as the sales and CRM data are centralised across all markets.

**6. Likelihood of value capture (Low/Medium/High) with justification**

High - churn prediction models are quite common in cases like this .They can be relatively easy to develop and scale and can provide useful and actionable insights fast.

**7. Key Business Stakeholders**

CMO , Marketing and CRM teams . The Marketing and CRM teams will have to work closely together to release relevant digital content via email campaigns to the selected customers.

## **Project 3 Name:** Short term financial forecasting automation

**Business Functional Area:** Finance

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

* Business Problem Addressed:  the internal finance teams are currently doing manual forecasts in Excel spreadsheets for all products sold by the OpCo across the various business lines . This is done on an SKU level which means that the local small finance team needs to do all of this across hundreds of SKUs. This process is time consuming and error prone .
* Role of data science in addressing the business problem: using existing data and forecasting algorithms, this whole process can be automated almost end to end. Moreover, the forecasting accuracy will also increase and be less biased .
* Targeted business objective(s): Resource optimisation ; contributing to overall corporate financial planning accuracy

**2. Data Science Classification**

* Approach:  **Predictive**
* Type of Model: Supervised learning

**3. Data needed for project and sources for that data**

Revenue data on SKU level, several years back in time. This data is readily available in the financial systems.

**4. Magnitude of opportunity (with justification)**

Even though it is part of a large corporation, the OpCo has minimal resources to deliver on the various internal corporate activities. Thus any level of automation which will free people’s time from “busy work” and direct it towards strategic thinking will be of great benefit for the company.

Moreover, the HQ offices have introduced revenue forecasting accuracy as a core KPI for the operating companies. Thus this approach will help the OpCos yearly evaluation.

**5. Cost and complexity of development and implementation**

Low cost and medium complexity. The different divisions are using different ways to log their financials so this data will have to be extracted and transformed for this exercise. Regarding the algorithmic approach: while various complex algorithms could be selected (including deep learning approaches) , for short term forecasting standard approaches will do . We do not expect to spend significant time in the first phases of the project to fine tune these algorithms.

**6. Likelihood of value capture (Low/Medium/High) with justification**

High value capture as most of the products have a quite stable sales pattern and seasonality which means that these short term forecasts can perform quite well.

So we will likely be able to achieve the automation and high accuracy goals.

**7. Key Business Stakeholders**

The Finance team (finance analysts).

## **Project 4 Name: Predicting launch curves**

**Business Functional Area:** Finance, Business insights & Supply Chain (cross functional)

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

* Business Problem Addressed: historically, the OpCo has been struggling to provide long term volume and value forecasts for non launched products as part of the internal business planning efforts. Over or under forecasting the first couple of years of a product's lifecycle can have a significant impact to a product’s success on the market as all of the efforts allocated for launching (both human and financial) and supporting a new product are tied to the corporate financials.
* Role of data science in addressing the business problem:using data on past launches across years and external factors which might have influenced these we might be able to provide an estimate of what the likely launch curve in terms of sales volumes we would have under certain circumstances.
* Targeted business objective(s): business optimization (financial forecasting accuracy, demand optimization, resource planning)

**2. Data Science Classification**

* Approach: **Predictive**
* Type of Model: unsupervised models to isolate types of “ product launch curves” . Supervised models (such as random forest) to predict the shapes of these curves.

**3. Data needed for project and sources for that data**

Data on the first 5 years past a product launch across hundreds of similar products both for the corporate brands and and competitors in similar markets.

Additional data sources which can provide historical prices , discounts and other relevant information for these products.

All data needed on non-AACG's products will need to be purchased from external data vendors.

**4. Magnitude of opportunity (with justification)**

High opportunity for optimising operations, planning for the right amount of FTEs and budgets to support new brands without affecting core operations.

**5. Cost and complexity of development and implementation**

High complexity and high cost - the data needed is quite pricey and available across different vendors. There is a quite high analytical complexity and I would probably consider sourcing external partners which have solved this problem in the past to support us with this initiative. This will significantly increase the cost .

**6. Likelihood of value capture (Low/Medium/High) with justification**

Low if executed internally; Medium to high via an external partners which have worked on these problems and are very familiar with the data and common analytical approaches here

**7. Key Business Stakeholders:**

Finance and Supply Chain teams. Business insights teams would also use this information in their long range strategic forecasting.

### **Project 5 Name:** Supplier quality optimization

**Business Functional Area:** Supply Chain

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

* Business Problem Addressed: The various business areas (beauty, household goods, personal hygiene) have hundreds of products under their umbrella. This means that the company is working with different suppliers to ship the product from the manufacturing plant to the partner shops or other customers.

Despite this complexity, the quality and speed of the shipments needs to be at the same standard across suppliers.

* Role of data science in addressing the business problem: using data on suppliers, past quality assurance programs and shipments we can predict future non adherence to quality standards.
* Targeted business objective(s): business process optimization

**2. Data Science Classification**

* Approach:Semi supervised
* Type of Model: Predictive

**3. Data needed for project and sources for that data**

Data from internal logistics and supply chain systems.

**4. Magnitude of opportunity (with justification)**

The brand perception and the sales linked to it are as high as the end product quality is perceived. With an increasing more complex supplier network, there is a higher risk of product damage which means that intensifying these suppliers or individual events which could have bad consequences to the company would allow us to mitigate risks.

**5. Cost and complexity of development and implementation**

Low cost and medium complexity as it is unclear whether the data includes the necessary information that can help us identify events or suppliers who have a high risk of non compliance.

**6. Likelihood of value capture (Low/Medium/High) with justification**

Medium. If after the discovery phase we do not have the data to specifically answer the question, we can at least identify some commonalities between “poor suppliers” and add additional tracking in the systems that can allow us to easily predict in real time which are the suppliers most likely to perform poorly.

**7. Key Business Stakeholders**

Supply Chain

### Project 6 Name: Localisation of global ML model for vendor selection

**Business Functional Area:** Procurement

**1. Description of the project (including business problem to be addressed, how data science will address that business problem, and the targeted business objective (revenue? customer acquisition? cost reduction?):**

* Business Problem Addressed: The central data science team has built a model for vendor selection designed to be used by the procurement department. This model has been designed and tested on data and parameters relevant for a selected number of countries which are significantly larger in size and with a different vendor ecosystem than the one where our operating company is located. This means that the local procurement team cannot benefit from it, although there is a huge need for such a tool. Moreover, the central team does not have future development plans for this solution on their roadmap for the next year.
* Role of data science in addressing the business problem: use existing historical data from purchases, invoices, budgets, project types and internal stakeholder rating to recommend the best supplier for a project.

In this case , the business value has been shown and the goal here is to take an existing model and adapt it to the specifics of the local market .

* Targeted business objective(s): optimise supplier selection to ensure successful project completion and good value for the investment.

**2. Data Science Classification**

* Approach:  **Predictive b**
* Type of Model: Supervised Machine Learning

**3. Data needed for project and sources for that data**

Use existing historical data from purchases, invoices, budgets, project types and internal stakeholder rating .

**4. Magnitude of opportunity (with justification)**

Medium opportunity in an often overlooked area. With a plethora of vendors out there, it can be difficult to always make the right choice of supplier. Often expensive projects do not turn out as expected which leads to losses at many levels ( resources, monetary, time, missing out on deadlines for lanches etc)

**5. Cost and complexity of development and implementation**

Low cost and low complexity. The model has actually been implemented so the core thinking and approach have been validated.

**6. Likelihood of value capture (Low/Medium/High) with justification**

Medium value : the project has been proven a success by the central team and it is currently used in production by various procurement teams. There is a risk however in the efforts to localise it as the data set might be significantly smaller and we will thereby have a high risk of overfitting.

**7. Key Business Stakeholders**

Procurement, with impact across all teams in the OpCo which are engaging with external vendors.

# Step 2 - Developing a Roadmap: Prioritizing Data Science Opportunities in the Business

A strategic approach to data science requires the business to consider the relative opportunities, costs, and risks of potential projects to identify the best order to carry out the projects. What should be tackled first? What is best pushed off until later? Completing the Data Science Roadmap requires stepping through key considerations to determine which project(s) should be considered ‘top priority’ and at what pace these and subsequent projects should be initiated.

**1. Complete this “Rack and Stack Exercise” worksheet to determine the relative strategic alignment, cost, complexity of implementation, certainty of value capture, and magnitude of benefit for each of the six projects**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Direct Alignment with Strategic Goals?** | **Cost** | **Complexity of Implementation** | **Certainty of Value Capture** | **Magnitude of Benefit** |
|  | 1=Low; 5=High | 1=High; 5=Low | 1=High; 5=Low | 1=Low; 5=High | 1=Small; 5=Large |
| **Project 1:**  **[MMM for a core brand with large investments]** | 3 | 2 | 1 | 3 | 5 |
| **Project 2:**  **Predicting customer churn** | 5 | 4 | 4 | 5 | 4 |
| **Project 3:**  **Short term forecasting automation** | 4 | 5 | 4 | 5 | 5 |
| **Project 4:**  **Predicting launch curves for existing products** | 5 | 1 | 2 | 3 | 4 |
| **Project 5:**  **Supplier quality optimization** | 4 | 3 | 2 | 3 | 4 |
| **Project 6:**  **Localisation of global ML model for vendor selection** | 1 | 5 | 5 | 3 | 3 |

**Project 1:** Marketing Mix Modelling for a core brand with large investments

**Project 2:** Predicting customer churn

**Project 3:** Short term forecasting automation

**Project 4:** Predicting launch curves for future products

**Project 5:** Supplier Quality Optimization

**Project 6:** Localisation of global ML model for vendor selection

**Note: You may choose to represent this information on slide 8 of the CEO Presentation Template**

**Please complete Step 2, Part 2, the Data Science Opportunity Matrix, using slide 1 of the CEO Presentation Template (You may or may not decide to include this slide as part of your CEO presentation)**

**Step 2, Part 3: Complete the table below by referencing the first four data science projects chosen for implementation. Include your justification for each project's order of implementation (e.g., how will the third project benefit from being implemented after the completion of the first two projects?)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Project Order** | **Project Title** | **Order Justification** | | | | | |
| 1 | **Project 3:** Short term forecasting automation | There is an urgent need as the financial teams workload is very high and the company has experienced a decrease in the continuous forecast accuracy. One of the company’s corporate KPIs is to increase revenue forecasting accuracy across the planning cycles. This project also has a low cost and low implementation effort with a high likelihood for success which will help in getting traction for the other upcoming data science projects. | | | | | |
| 2 | **Project 2:** Predicting customer churn | The strategic goal during the pandemic is to grow the online business. In order to do so, we need to better understand our direct customers and ensure that we have a retention strategy. | | | | | |
| 3 | **Project 4:** Predicting launch curves for future products | This project has a high complexity, but it is linked to overall corporate strategic goals and can have a large long impact on business growth. | | | | | |
| 4 | **Project 1:** Marketing Mix Modelling for a core brand with large investments | This project has the highest complexity in terms of the data needed (availability and quality). A successful completion would have a large impact as it would ensure better budget allocation, increase profit and growth. | | | | | |

**Note: You may choose to represent this information on slides 6 and 7 of the CEO Presentation Template**

# Step 3 - Establishing a Data Science Human Capital Strategy for your Data-driven Business

Now that we have established a roadmap for carrying out data science projects, our attention must turn to building and configuring the organization we will leverage to carry out this roadmap. The Data Science Human Capital Plan completed in this step will cover the organizational structure and talent configuration best suited to carry out the business’s roadmap, as well as the activities that the organization in particular -- and business more broadly -- must complete in order to promote a data-driven culture throughout the business.

**1. Identify the organizational model best suited for the data science organization that your business will need to deliver on the roadmap completed in Step 2. Provide justification for your selection based on the needs, scope, and timing of projects to be implemented in the Data Science Roadmap. If your organization should start with one model and evolve toward a different model, you may provide that detail and justification in your response.**

**Organizational Model:** Federated

**Justification:**

The current organisation is an operating company in a medium sized market in Europe, part of a large corporation. There is a central data science team running various high profile - high cost projects . The OpCo is planning to start their own initiative which can better serve the local needs.

While the data science projects will be generally executed locally, they will need to be in alignment with the company’s strategic goals ( corporate and local ) . Moreover, since there is already existing infrastructure built for data science, the OpCo will leverage it as much as possible from that.

The local Head of Data & Analytics will oversee all the initiatives and manage the data science resources , but there will be a functional business owner for each of the projects . These individuals will not only contribute with domain expertise, but they will also make sure that the outcome of the project is incorporated back in the business within their respective areas.

The local Head of Data & Analytics will be connected to the central data science team via the corporate forums and will closely follow the developments within the area.

**2. Complete the “Human Capital Plan” Worksheet for your data science organization.**

**- Identify the first ten professional roles for which you would recruit. How would you organize these roles into teams within the organization?**

For example, if you had 4 data scientists split evenly into two teams, your response would look like this:

|  |  |  |
| --- | --- | --- |
|  | **Position** | **Team** |
| 1 | Data Scientist | 1 |
| 2 | Data Scientist | 1 |
| 3 | Data Scientist | 2 |
| 4 | Data Scientist | 2 |

Identify your roles and teams below:

|  |  |  |
| --- | --- | --- |
|  | **Position** | **Team** |
| 1 | [Business] Data Scientist | Team 1 |
| 2 | Machine Learning Engineer | Team 2 |
| 3 | Data Analyst | Team 1 |
| 4 | Data Engineer | Team 2 |
| 5 | Dashboard/BI Developer | Team 2 |
| 6 | Project Manager | Team 1 |
| 7 | Project Manager | Team 2 |
| 8 | Change Manager | Team 2 |
| 9 | Data Science Engagement Manager | Across teams |
| 10 | Data Manager | Across teams |

**Note: You may choose to represent this information on slide 9 of the CEO Presentation Template**

**Assume that leadership will allocate four new FTEs for your data science organization during the current fiscal year. How would you prioritize your organizational buildout?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Order of Hire** | **Position** | **Justification** | | | | | |
| 1 | Business Data Scientist | We can use a data scientist with business knowledge for a wide range of projects in the initial phase. These individuals are data science generalists and can probably develop a POC project on their own (data collection, cleaning, translating business problems to data science solution and implementation). | | | | | |
| 2 | Data Engineer | Following a successful POC , we will need to establish a data pipeline that can help us scale our project. Thus we need to hire data engineers that can help do that and help further with data collection and preparation for the other data science initiatives. | | | | | |
| 3 | Machine Learning Engineer | For ML algorithm development and fine tuning as well as building algorithms and solutions which can be scalable. | | | | | |
| 4 | Change Manager | We need an individual who can facilitate change and introduce the new ways of working which will be changed by introducing data science initiatives in this workflow. This role would be crucial to ensure that the business will embrace these new initiatives. | | | | | |

**Craft a “Data-Driven Transformation Strategy” by identifying six specific initiatives that you would recommend the data science organization and/or the business undertake in order to promote a data-driven culture across the business.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Strategy** | | | | | | |
| 1 | Executive Data Science Education - training leadership role in what data science is the benefits of incorporating data driven approaches in business practice. | | | | | | |
| 2 | Engagement programs through sharing sessions where Data Science POC results are shared across the organisation | | | | | | |
| 3 | Data literacy training across all roles | | | | | | |
| 4 | Use external consultants for unique skills/ first time projects only and ensure that they work on the infrastructure together with the internal teams rather then delivering the results in ppt decks. | | | | | | |
| 5 | Create concrete learning programs for traditional analysts who would like to transition into a data science career. They will be great assets for these initiatives in the long term. | | | | | | |
| 6 | Encourage a try & fail fast culture for data science initiatives in the initial phases to stimulate the interest across teams for trying out new things and incorporating learnings into the next project. | | | | | | |

**Note: You may choose to represent this information on slide 10 of the CEO Presentation Template**

# Step 4 - Establishing the Technical Infrastructure to Support the Data Science Organization

With a completed Data Science Roadmap and a Human Capital Plan for executing the data science strategy, we turn our attention to the technological capabilities that must be built to support the new Data Science organization.

Complete the table on the next page by entering strategic aspects your business might consider to meet its Data and Data Architecture needs.

**Data and Data Architecture Strategy for the business**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Component** | | **Strategy** | | | | | | |
| Data Requirements | What data should be included in the Data Strategy? | The financial and CRM data is readily available via the centralised data lake. The next steps for the OpCo would be to integrate their marketing data from various sources and external partners as well as develop processes for external data acquisition that can help answering the specific questions | | | | | | |
| Data Governance | How will we promote data availability? (provide at least two ideas) | If the data is highly sensitive (such as financials) , it will be made available for the project team with the appropriate clearance only.  If the data is not sensitive, it will be made available for access to the appropriate data dictionaries for all team members .  Regardless of access level, the set up needs to be very easy to understand so I would make sure that the data is clearly labelled and documented. | | | | | | |
| How will we promote usability? (provide at least two ideas) | Providing use cases from the central teams where they have worked on similar data  Providing external use cases either published or from startup partnerships where they have managed to get insights from similar data. | | | | | | |
| How will we guarantee integrity? (provide at least two ideas) | * All data coming from the internal systems will be subjected to a thorough process of cleaning, documentation and validation which will follow an automated process and periodic quality and security reviews. * Security certification will be required to operate with the internal data * The external data required for a POC will not be uploaded to our environment before it has been approved by IT, legal ( GDPR and privacy compliance) and security. | | | | | | |
| How will we guarantee security? (provide at least two ideas) | Data can only be accessed in the company’s secure environment (AWS via virtual machines or via the data stack already enable on the cloud infrastructure)  Each user will receive the right level of access depending on their role and scope of project . | | | | | | |
| Technology | Identify the components of your Data Architecture | Centralised Data Lake for all internal sources (Amazon Redshift)  Amazon S3 buckets for one-off /POC projects where users can upload their data directly  AutoML tool ( DataRobot)  Amazon AWS EC2 for advanced ML development & Python  Rstudio server for R users  Data pipeline (Apache Airflow : open source)  Tableau for data visualisation | | | | | | |
| Skills and Capacity | How will we promote development of data literacy skills and capacity throughout the organization (provide at least three ideas) | * While the individual data science specialists are skilled within their own areas, they would require additional training on the bigger picture and how the business operates. This will help that the specialists do not work in silos and solutions which are far fetched from how the business operates. * SImilarly, other teams outside data sciences will require some high level training in data science across the ranks. This will make sure that everyone speaks the same language. * Developing internal learning paths for data upskilling across roles. These would be internal “data champions” that will support in different capacities our data initiatives. | | | | | | |
| Support for Machine Learning | Give a brief description of the machine learning architecture and how it will interface with the data architecture | DataRobot (autoML) will be connected directly to the Datalake and it will help with data access , preparation and basic ML for the easy use cases. The users can also upload their data in S3 and access it directly via this infrastructure. The output from the ML models can be pushed into S3 and visualised in Tableau.  If we need to implement something that cannot be easily done using the above, we can do so using Python , Apache Airflow and Docker from the same data source.  This would be a more rare use case but suitable in the initial phase ( up to 2 years) | | | | | | |

**Note: You may choose to represent this information on slide 11 of the CEO Presentation Template**

# Step 5 (OPTIONAL) - Record a short video of you presenting your final slide deck to your CEO or Executive Committee (5 minutes)

You may wish to submit a short video of you presenting your final presentation to your CEO; while this is not a formal requirement for the Capstone project, it does provide an outstanding way to gain practice with communicating about data science in business contexts.