Data Science Strategy Consulting Project

Assessment type: Presentation

Assessment 2: Solution design and strategy proposal

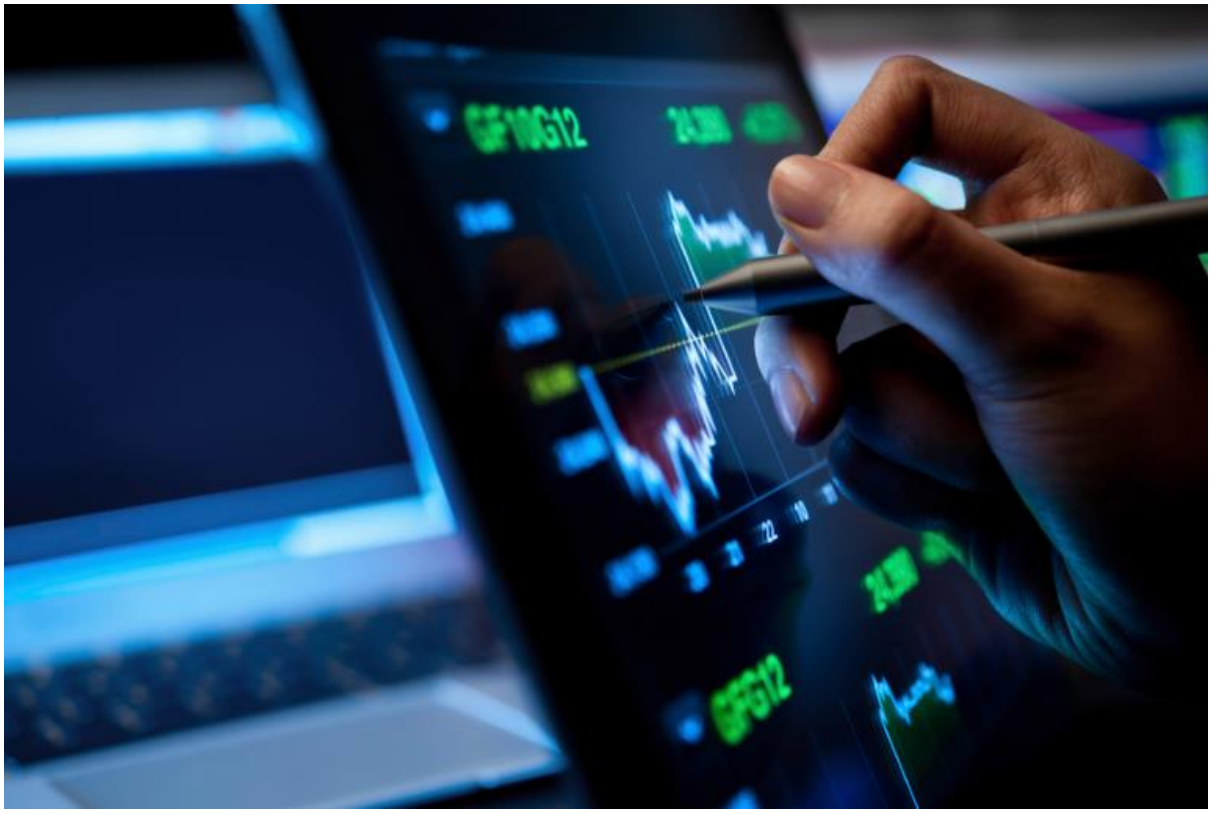
Slide 1: Hi, I am Rokshana Pervin. Today, I am going to narrate my presentation in regard to solution design and strategy proposal.

Slide 2: Overview of initial solution approach and why it will work

During research about Power Up, there are many challenges have been identified that can be overcome using some solutions where data science solutions such as data strategy, data engineering, data analysis, data models and data visualisations are remarkable (Shree 2022). Using the components of data science solutions, any organisations such as Power Up can mitigate their present and future problems. Apart from this, different types of data science solutions such as descriptive analysis and predictive analysis can help business to provide an informed and sustainable decisions that bring large amount of profit the industry (Paul et al. 2022).

Benefits: In terms of benefits, to make the strong backbone of data science projects, it is considered to apply the combination of data science solutions where data strategy or machine learning approach such as supervised machine learning approach is significant to predict the viability of the business that ensures the financial performance of the project (Kumar and Zymbler 2019). Here, viability indicates a business success and profitability 'which means it has more revenue coming in' (Murray 2022). Apart from this, as logistic regression model is one of the supervised machine learning approaches that helps to predict highly accurate data and classified the customers accurately, it is anticipated that it will be beneficial for Power Up to find the accurate customers behaviour and their attributes who are using their product in a highly manner (Manďák and Hančlová 2019).

Costs: Here, it needs to be considered about the cost of the data science solution though the benefits of the above solutions have been designed to this proposal (Inan et al. 2022). As every project has some cost and effect, therefore, it is crucial to find the accurate estimate of the cost of the desired project. Or else, the project will be vulnerable or might be failed because of the unplanned project estimation (Incze 2019). Moreover, while adding data science solutions such as data strategy, data engineering, predictive analytics into the project, we need to be aware about the budget of these solutions that can be hinders on the way of business success (Incze 2019).



Therefore, the business requirement documents (BRD) or business objectives and functional requirement document (FRD) or the purpose of the project need to set in accordance with the cost of the data science solution to increase the scalability of the project (Kaushal 2022). Eventually, it is hypothesised that the data science strategy such as supervised machine learning approach can be used, as it can predict the future plan and factors accurately using various algorithm that indicates to reduce the cost of the project (Inan et al. 2022).

Slide 3: Overview of the computational architecture and data science approaches.

The term of Architecture indicates the structure of the construction (Ackerman et al. 2023). As this presentation is highlighting the data science architecture, therefore, we need to build the framework where data can be 'collected, stored and used in an information system' (Simplilearn 2023). These are known as collect layer, organise layer, analyse layer and infuse layer (Ahluwalia 2022).

In order to collect, transform and loading the data, organisations are adopting different approaches where ETL (Extract, Transform and Load) is one of the processes 'for moving data from source system' (Wainstein 2024).

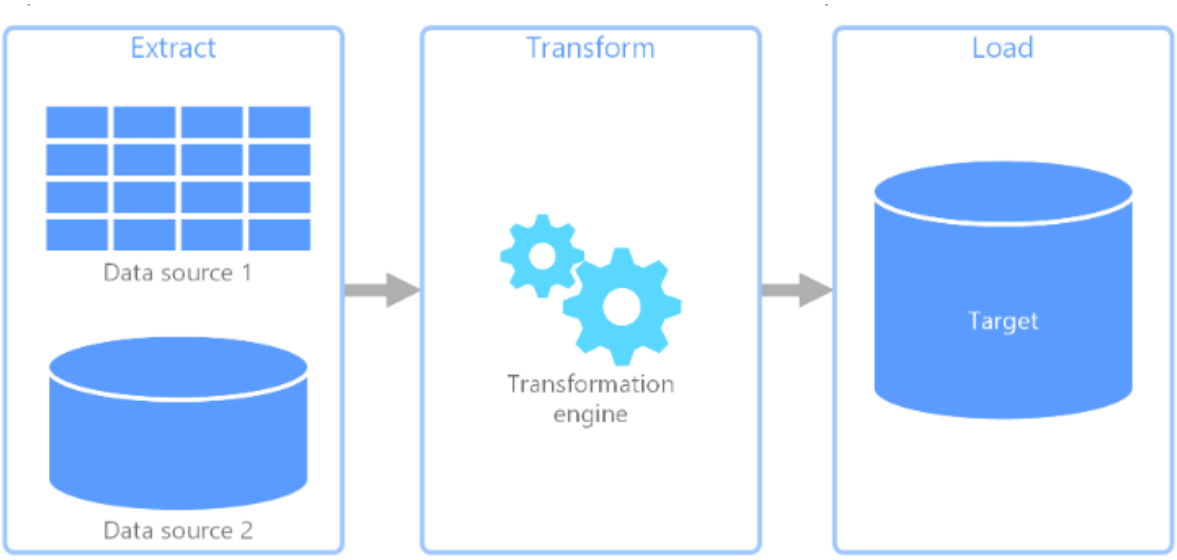


Figure 1: ETL (Extract, Transform, Load)

In this layer, the end goal is to enable businesses to data-driven business decisions using the targeted system which is known as data repositories or data warehouses (Wainstein 2024).

Apart from this, the modern data integration process, ELT (Extract, Load, Transfer) is becoming more common and practical to data warehouse as it is cloud-based data warehouse.

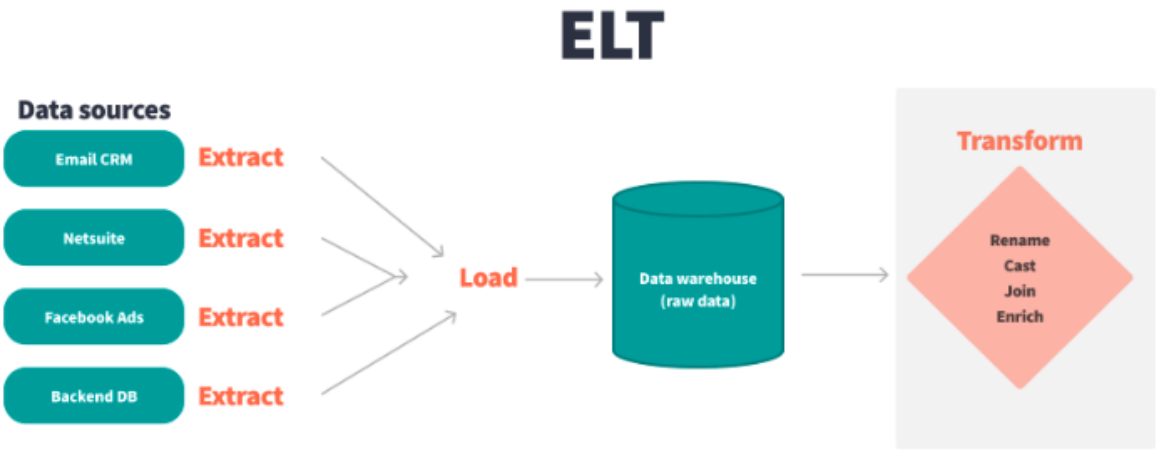


Figure 2: ELT (Extract, Load, Transform)

Eventually, it is observed that Power Up is looking for different data gathering innovations where scalability and security need to work together.

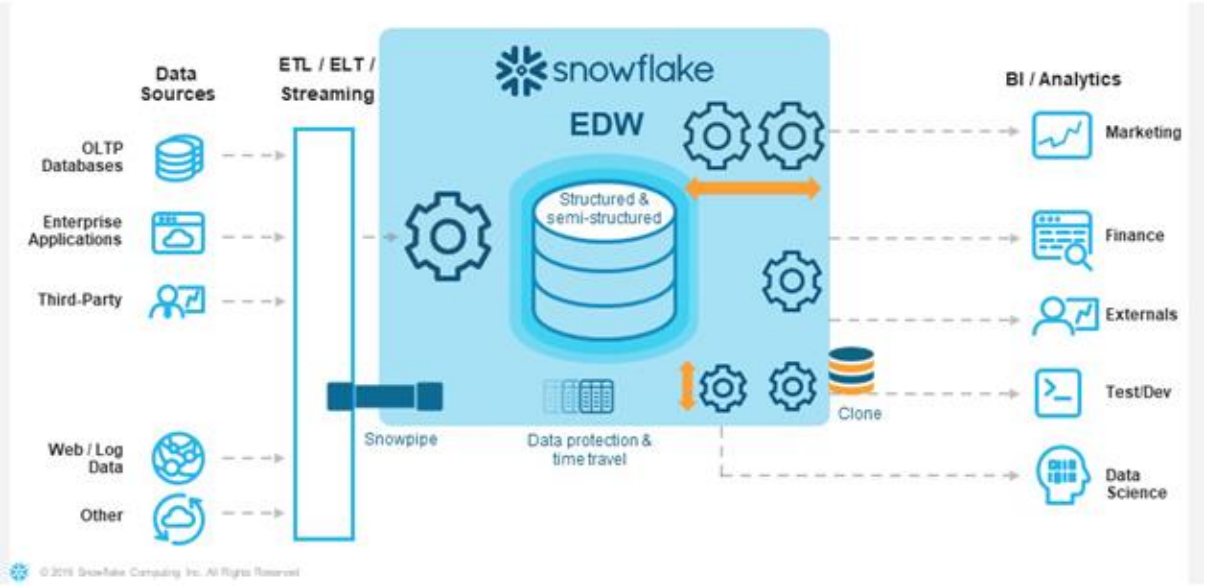


Figure 3: Snowflake

The cloud-based data platform or data warehouse known as Snowflake can support all the above three terms together including extract, loading, and transform while integrating data that indicates bringing more benefit for the Power Up energy industry (Allam et al. 2023).

Slide 4 and 5: Overview of how this solution approach will be implemented.

Snowflake is implemented in energy industry:

At the beginning of the implementation of the solution approach, it is vital to focus on all stakeholders' perspectives that ensure about the effective solution and create actual user's need (Myob 2024). It is observed from the case study that Power Up can gather those aspects of stakeholders using different kind of tools such as persona map, marketing campaigns and so on. Power Up can combine these perspectives and send it to the centre of excellence team where data, analytics and cloud roadmap can be created for the industry (SNOWFLAKE n.d.).

Now, in order to implement the solution approach or Snowflake data repositories into the data science project, it is necessary to add different types of data sources such as open sources, persona map for various stakeholders, as well as data lakes, as Snowflake can integrate all the data formats such as structured, unstructured, and semi-structured (Snowflake 2022). After integrating all data into the cloud-data warehouse, it needs to be analysed using some tools such as machine learning algorithms to find the real insights and provide informed data -driven decisions for improving the customer experiences (Malony 2023).

As mentioned in our previous work that we will use machine learning algorithms to find the accurate results of the customer experiences and customer satisfactions, here we will use logistic regression which is one of the supervised machine learning algorithms (Kumar and

Zymbler 2019). Here, logistic regression is a specialised form of a very common statistical method which helps to predict the real-life insights by tackling various problems that arise from real-life data (Anand and Bansal 2016). As Power Up is struggling to track the accurate customer experiences, it is anticipated that logistics regression model can help to classify the customers with satisfied and dissatisfied classes as it is a binary classification method. According to the logistic regression model, there will be two variables such as target variable or dependent variable and independent variable or predictor where target variable is known as customer satisfaction and independent variables will be customer's age, gender, monthly income, monthly interest, service quality, preferable industry to deal with and so on (Anand and Bansal 2016). To run the model, training data with input features and target outcome (satisfied or dissatisfied) will be created to develop the relationship between input features and the outcome.

$$\log \frac{Y}{1+Y} = b_0 + b_1 X1 + b_2 X2 + C$$

Figure 4: Screenshot of the Sigmoid Function

The above equation is highlighting Y as a probability of outcome, b0, b1, b2 are the co-efficients, X1, X2 are the independent variables and C is the constant.

After training the data, p-value (0.05) will indicate the accuracy and confusion matrix of the model and will remove the collinear features (most features are same with each other) and the list of variables by showing the percentage of specificity and Sensitivity (Anand and Bansal 2016).

Sensitivity: probability that the predicted result will be true when the actual case is also true (true positive rate, expressed as a percentage)

$$Sensitivity = \frac{true\ positive}{true\ positive\ +\ false\ negative}$$

Specificity: probability that the predicted value will be false when the actual case is also false (true negative rate, expressed as a percentage)

$$Specificity = \frac{true\ negative}{true\ negative + flase\ positive}$$

Figure 5: Sensitivity and Specificity

which are most relevant with independent variables. Apart from this, the AUC or area under curve of the final reduced model would be closely 1 that will provide the great insight into which customers are more likely to satisfy with the desired product (Anand and Bansal 2016).

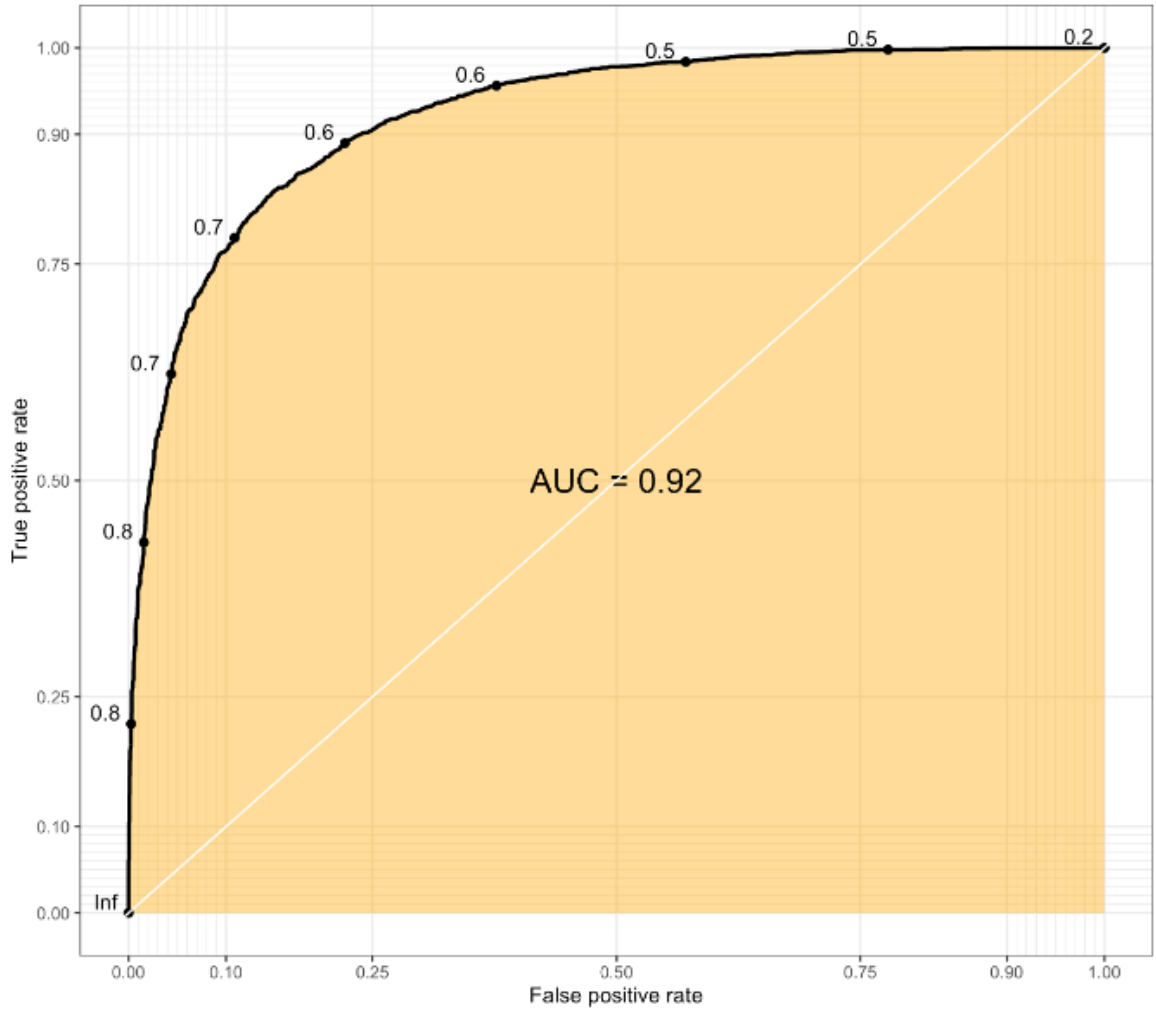


Figure 6: Screenshot of the AUC or area under curve

Additionally, co-efficient of the lr or logistic regression model will provide the customer insights who are likely to be refused the usage of the product.

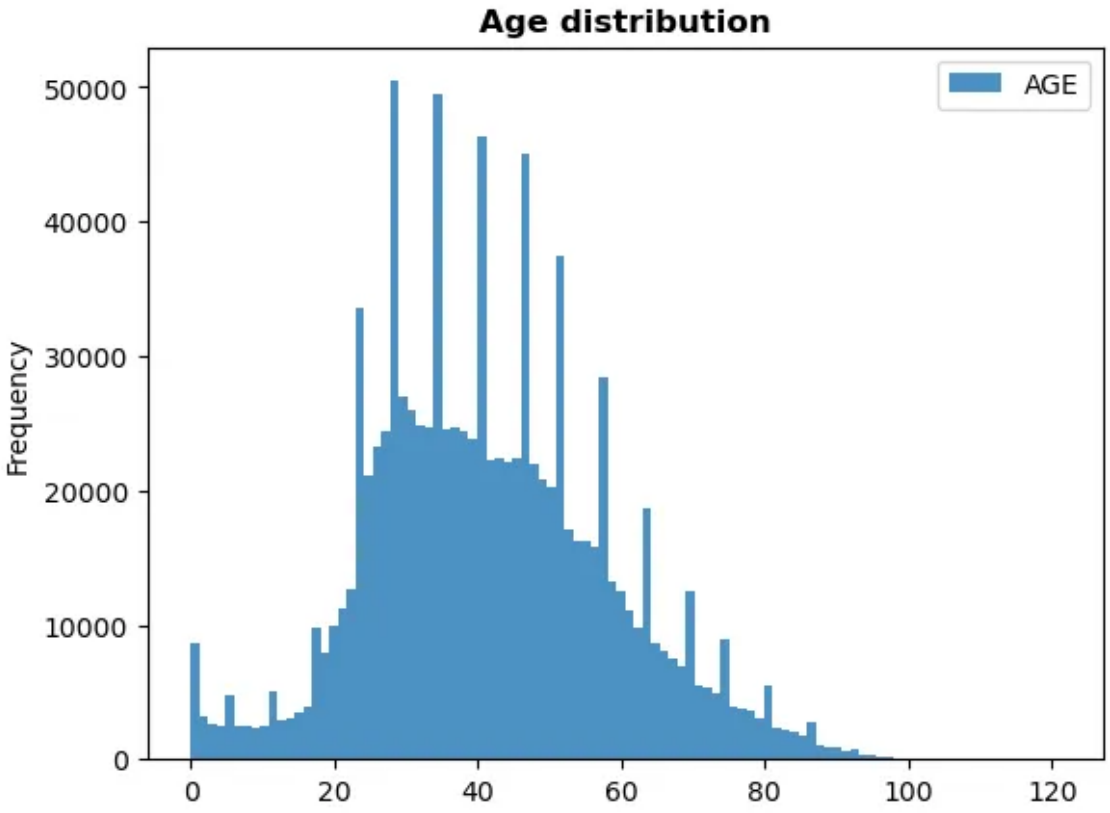


Figure 4: Age distribution for high usage

To display the visualisation, IDE or Integrated development environment such as Jupyter notebook with Python programming language can be used. Here, it is indicated from the above histogram plot that the vast majority of products have been used by 25-year-old to 55-year-old customers as they are active users and engage with their work and other surrounding environments actively.

From the above analysis, it is observed that the logistic regression model can predict the precise results of the customers who are actually satisfied with the product of the industry. To get the accurate number of satisfied customers, Power Up data science team can train the model repeatedly by 'centralising ModelOPs or Model Operations' strategic approach and find the predictors that are intensively relate to the target variable (ModelOPs 2020).In this way, the Power Up team can manipulate the real number of customers who are using the network or product less or more extensively.

Slide 6: Overview of how the chosen solution achieves business value.

From this presentation, it is anticipated that if the presented solution can be applied in the Power Up industry, it would be hybrid infrastructure where centre of excellence team will work collaboratively using both cloud-based data warehouse 'Snowflake' and on-premises

infrastructure machine learning model such as logistic regression model (PwC 2015). In this case, it is observed from the Power Up industry that centre of excellence team has already been involved with adding business value for the organisation. Apart from this, as Snowflake is the cloud-based repositories, it is researched that Snowflake can enhance scalability, speed, highest security as well as 'provide cost-efficiency to the organisations with its pay-as-you-go pricing model' (Allam et al. 2023). Additionally, as machine learning is the subset of AI and Power Up is proud with Centralised AI, it is represented that centralised AI will help Power Up to enhance the interoperability between different operating systems which will generate the sustainability of the project (Lio 2023). Eventually, it is anticipated that the chosen solution will bring the more value for the entire organisation in near future, if they transfer their whole architecture to the cloud including machine learning processes which will bloom as 'deepest set of machine learning services' (Aws 2024).

Why this is the preferred solution.

There are some attributes that make the solution preferable such as,

1) It increases the scalability and efficiency of computing and storage resources (Bennett 2023).
2) It provides robust security including data encryption, and network isolation (Bennett 2023).
3) It reduces hardware and software arrangement difficulty as it has cloud-based functionality (Carrero 2021).
4) It has real-time insights with analysing accurate data from the real-life scenario.
5) It increases the business value that indicates viability or financial performance across the industry (Murray 2022).
6) This solution is sustainable as it follows the ModelOPs strategic approach that allows the continuous scaling up onto the data science project (ModelOPs 2020).

Slide 7: A brief discussion of non-functional requirements such as privacy, security, and performance.

To generate the successful project for the business, non-functional requirements are playing a vital role as non-functional requirements implies 'how the product must perform'. Here, non-functional requirements such as privacy, security and performance can be described as below:

NFR	Description	Metric	It's important
Privacy	This requirement refers 'understanding in terms of the social relationships that underlie the application domain' (Yu and Cysneiros n.d.).	Social relationship	If privacy is not maintained in any organisation, they will fail to get future success as the success is completely relate to the

			customer's privacy.
Security	This requirement refers Authenticate Access and Identification (Yu and Cysneiros n.d.).	Passwords and digital certificate.	Security is crucial for every business as it provide compliance, rules and regulations for any enterprise.
Performance	This requirement refers the speed of the system or product (Altexsoft n.d.).	Greater speed	If the network is low, the system will not work accurately which results customer dissatisfaction (Altexsoft n.d.).

Slide 8: The persona we're trying to solve an issue for, and their key wants and needs.

Persona Map	
Internal Stakeholder	External stakeholder
Name: Gene Collins **Position:** CEO of the Power Up company.	**Name:** John Cutler **Position:** working in the IT company (Customer)
Demographics: **Age:** 55, **Residence:** Victoria	**Demographics:** **Age:** 35, **Residence:** Victoria
Attributes: He is a hardworking and innovative as well as trying to resolve all problem collaboratively.	**Attributes**: He is very co-operative and friendly.
Desired Goal: Providing uninterrupted network to the customer even if nationwide.	**Desired Goal:** Not having any network outages.
Needs: Using advanced technology to mitigate network outages.	**Needs:** Getting uninterrupted network.

Governance Solution Presentation:

Slide 9: Goals and planned outcomes of the engagement

Goals of the engagement:

- Maximise the business value using KPIs or business drivers such as customer satisfaction and customer acquisition.
- Applying advanced technology both in cloud and on-premises infrastructure to get the vast network for the customers.
- Ensuring transparent governance for the industry by minimising ethical issues and mitigating the risk that occurs from bias.

Planned outcomes of the engagement:

From the above engagement, it is analysed using high-level solution that customers monthly interest can be predicted which will assist Power Up to compute the number of the customers who are using their network extensively.

Tracking of the outcomes at the end of the engagement, and in future

Although, the above solution added ModelOPs to make the potential sustainability and greater speed, Power Up can make a link with the sustainability and viability to bring the maximum value for the future business.

Slide 11: Governance plans for the project.

In this era, Governance is taking a large part for the organisation that encompasses the establishment of systems and processes (Australian COMMISSION ON SAFETY AND QUALITY IN HEALTHCARE 2024). As the above solution attached with cloud-based repository and ModelOPs strategic approach, it is analysed that cloud-based data repository places a strong emphasis on security and data governance (Allam et al. 2023). Apart from this, ModelOPs approach is also represented greater speed with scalability, risk control with higher security that indicates the best governance practice for the project (ModelOPs 2020). Moreover, as the project has business requirements (BRDs) and functional requirements (FRDs) where the objective of the project is indicating scalability, the purpose of the project is representing sustainable and consistent solution, therefore it is notion that the project met the governance plan adequately (Kaushal 2022).

Slide 12: Risk, dependencies, and mitigation – including analysis of project viability and feasibility.

Here, viability implies that the ability to survive which the financial performance for the business (Murray 2022). On the contrary, feasibility exposes different aspects of a proposed project such as technical, financial, environmental considerations and so on to deliver the project in an easy way (Simplilearn 2023).

Therefore, if the aspects of the project will not meet with the project requirements, it will not add value to the business rather it will form a risk for the project and call for dependencies to mitigate the emerged risk (RISKID 2018).

Eventually, organisation needs to monitor continuously to mitigate the risks and to get the higher-level of success of the project.

Highlight key potential risks around ethics and bias.

The potential risks associated with ethics:

Ethics in business or project ensures high levels of standards, principles, policies and values that comes with high level of trust between consumers and industry (Twin 2024). Here, it is analysed from the case study that customers are core part for the organization. Therefore, if we do not keep our customers happy and satisfied, it will create a high risk for Power Up industry.

The potential risk around with bias:

Bias is an important term for any business specifically in working with data as data is the core thing for business (Kangralkar 2021). Bias can occur anytime that reflect on the solution and brings unpredictable loss for the organisation. In data science, bias happened in different ways such as sample bias. Sample bias created when the algorithms are trained on data that doesn't represent the real-world scenarios (Kangralkar 2021).

Hence, we should be conscious about all kind of biases to secure the future success of the business and organisation.

Slide 14: Recommendation for the Proposed Solution:

From the above observations, it is analysed that the above solution will help Power Up company to mitigate their arisen problem using the highlighted cloud-based data repositories with machine learning approach which is remarkable. Though this solution has some advantage and disadvantages, it will be worth for the Power Up company as they were looking for innovative and sustainable approach to mitigate their arisen problem. Eventually, Power Up can change their entire infrastructure and transfer it to the cloud which will be cost-effective and maximise value for their future business.

Slide 15: Thank you.

Slide 16: References

Allam K , Ankam M and Nalmala M (2023) 'CLOUD DATA WAREHOUSING: HOW SNOWFLAKE IS TRANSFORMING BIG DATA MANAGEMENT', *International Journal of Computer Engineering and Technology (IJCET),* 14(3), 156-162, https://iaeme.com/Home/issue/IJCET?Volume=14&Issue=3

Ackerman SJ, Gowans A and Scruton R (2023) *architecture,* britannica website, accessed 13 January 2024. Architecture | Definition, Techniques, Types, Schools, Theory, & Facts | Britannica

Ahluwalia H (2022) *Layers of the Data Platform Architecture*, analyticsvidhya website, accessed 11 January 2024. Layers of the Data Platform Architecture - Analytics Vidhya

Anand A and Bansal G (2016) 'Predicting Customer's Satisfaction (Dissatisfaction) Using Logistic Regression', *International Journal of Mathematical, Engineering and Management Sciences*, 1(2), 77-88, ISSN: 2455-7749.

Australian COMMISSION ON SAFETY AND QUALITY IN HEALTHCARE (2024), *Goernance, leadership and culture*, Safetyandquality website, accessed 23 January 2024. Governance, leadership and culture | Australian Commission on Safety and Quality in Health Care

Altexsoft (n.d.) *Nonfunctional Requirements in Software Engineering: Examples, Types, Best Practices,* altessooft website, accessed 12 January 2024. Nonfunctional Requirements (NFR):Examples, Types, Approaches (altexsoft.com)

Aws (2024) *Free Machine Learning Services on AWS,* aws website, accessed 20 January 2024. Free Machine Learning Services - AWS (amazon.com)

Bennett T (2023) *What is Snowflake | A Comprehensive Overview with Pros and Cons*, dreamfactory blog, accessed 23 January 2024. What is Snowflake | A Comprehensive Overview with Pros and Cons - DreamFactory Software- Blog

Shree M (2022) *The 7 key components of data science – A Quick Overview,* medium website, accessed 01 February 2024. The 7 key components of data science — A Quick Overview | by Madhu Shree | Medium

Paul S, Sah D, Das R, Goswami S, Samanta S, Roy B and Pal D (2022) 'COMPONENTS OF DATA SCIENCE AND ITS APPLICATION', *International Journal of Engineering Applied Sciences and Technology,* 7(5), 237-240, DOI: 10.33564/IJEAST.2022.v07i05.039

Kumar S and Zymbler M (2019) 'A machine learning approach to analyze customer satisfaction from airline tweets', *Journal of Big Data,* 6(62), 1-16, https://doi.org/10.1186/s40537-019-0224-1

Murray J (2022) *What is Business Viability?* , the balance website, accessed 25 January 2024. What Is Business Viability? (thebalancemoney.com)

Manďák J and Hančlová J (2019) 'Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications', *STATISTIKA,* 99(2), 129 – 141, Use of Logistic Regression_understanding and predicting customer churn in telecommunications.pdf.

Inan T, Narbaev T and Hazir O (2022) 'A Machine Learning Study to Enhance Project Cost Forecasting', *IFAC PapersOnLine*, 55(10), 3286–3291, 10.1016/j.ifacol.2022.10.127.

Incze R (2019) *The Cost of Machine Learning Projects,* medium website, accessed 15 January 2024. The Cost of Machine Learning Projects | by Raul Incze | Cognifeed | Medium

Kaushal D (2022) *Why Business and Functional Requirements are Vital for a Project's Success,* netsolutions website, accessed 14 January 2024.

Kangralkar S (2021) *Types of Biases in Data*, towardsdatascience website, accessed 23 January 2024. Types of Biases in Data. Biases in data that we should all be… | by Swapnil Kangralkar | Towards Data Science

Simplilearn (2023) *What Is Data Architecture? Overview and Best Practices*, simplilearn website, accessed 11 January 2024.

Myob (2024) *How to work with business stakeholders*, myob website, accessed 16 January 2024. How To Work With Business Stakeholders (myob.com)

SNOWFLAKE (n.d.) CENTRAL DATA CLOUD IN ENERGY TRADING, *INFORMATION*, accessed 17 January 2024. SnowcaseEnergyINFOMOTIONCentralDataCloudinEnergyTradingSolutionSheet.pdf

Snowflake (2022) *How Retailers Increase Customer Satisfaction and Retention with the Snowflake Retail Data Cloud,* snowflake website, accessed 15 January 2024. How Retailers Increase Customer Satisfaction and Retention (snowflake.com)

Malony M (2023) *Uplift Modeling for e-commerce with Snowplow and Snowflake*, mikeymalony website, accessed 20 January 2024. Uplift Modeling for e-commerce with Snowplow and Snowflake | by Mikey Maloney | Medium

ModelOPs (2020) 'ModelOPs Essentials: Best Practices for Success with Enterprise AI', *modelops pdf,* accessed 21 January 2024. ModelOps_Essential_Guide.pdf

PwC (2015) 'Powerful predictive data analytics strategies help clinical trials get healthier: A global pharmaceutical company left traditional methods behind as it deployed advanced predictive data analysis and visualization techniques to improve the health of its clinical trial sites around the world' Pharmaceuticals & Life Sciences (Consulting) [PDF], accessed 22 January 2024. Pharmaceuticals & Life Sciences Consulting - Powerful predictive data analytics strategies help clinical trials get healthier: March 2015 (pwc.com)

Lio D M (2023) *The Pros and Cons of Centralizing the AI Industry: A Detailed Examination,* medium website, accessed 22 January 2024. The Pros and Cons of Centralizing the AI Industry: A Detailed Examination | by Michael De Lio | Medium

Carrero C (2021) *Building Better Machine Learning Models using Snowflake Data Cloud*, medium website, accessed 20 January 2024. Building Better Machine Learning Models using Snowflake Data Cloud | by Carlos Carrero | Snowflake | Medium

Yu E and Cysneiros M L (n.d.) 'Designing for Privacy and Other Competing Requirements', Faculty of Information Studies, accessed 23 January 2024. Designing for privacy requirements.pdf

RISKID (2018) *What is Risk*, video, YouTube, 8 August, accessed 25 January 2024.

Simplilearn (2023) *Feasibility Study and Its Importance in Project Management,* simplilearn website, accessed 23 January 2024. Feasibility Study and Its Importance in Project Management (simplilearn.com)

Twin A (2024) *What Is Business Ethics? Definition, Principles, and Importance*, Investopedia website, accessed 22 January 2024. https://www.investopedia.com/terms/b/business-ethics.asp

Wainstein L (2024) *The Community Blog for Business Analysts*, modernanalysts website, accessed 10 January 2024. Is ETL Still Relevant? > Business Analyst Community & Resources | Modern Analyst