Data Governance and Compliance Essentials

Stanton Man, Kagen Quiballo, Krishanu Sengupta

MSDS 403: Data Science in Practice

February 14, 2021

1. Overview | Understanding and mitigating potential risks in a digital transformation are crucial when launching a new product or service, creating a global marketing brand, or closing down sites to minimize profit loss. In order to understand these risks, companies must have a data governance plan, fundamental rules and regulations, and universal compliance with laws and ethics that govern their industry. Data governance can include managing highly sensitive data which poses newer levels of security risks and ethical understandings of data use. Along with complying to laws and regulations, data governance also allows for effective data management which helps with accuracy and reliability of data. These four points (data security, data accuracy, data reliability, and data ethics) create a framework for data governance essentials that companies need to adopt in order to have a successful digital transformation.

2. Roadmap | To aid in a company's data governance guidelines, The Consultancy has created a Data Governance and Data Compliance Matrix to rate use cases based on the four criteria: data security, data accuracy, data reliability, and data ethics. Based on a one to five scale (with one being the most complicated, and five being the most manageable), the scores in the matrix are averaged into a final "ease of management" score to determine which use cases have an attainable data governance implementation for a digital transformation. Specific data governance examples from the highest scoring use case, Amazon's Comprehend customer sentiment analysis, will be used as a reference throughout the suggested plan.

3. Guidelines for Implementation

Data Security: When looking into security governance, companies should focus on cybersecurity methods that encourage safe data usage and provide privacy protection for customers' data. Evaluation of this criteria depends on how complex of a cybersecurity architecture is needed to protect highly sensitive data. By taking into account data compliance from existing regulations and relevant cyber security threats, each use case may require different security measures for different data. From the Amazon Comprehend use case, data security is addressed using AWS Cloud Security. With customizable options for who can access the data and strong encryption to protect data in transit and at rest, one can ensure the protection of customer reviews (emails, calls, reviews, social media) and analyses to authorized employees ("Data Privacy" 2021). This use case scored highly because it did not require a complex security set up, whereas a use case that had no architecture in place yet may have scored lower.

Data Accuracy: Garbage in, garbage out. Models and insights that are generated from inaccurate data cause inaccurate predictions and prescriptions. This creates skepticism in the recommendations for decision makers, creating potentially harmful decisions. By focusing on data ingestion points, the data governance plan ensures clean data environments and quick data quality management. Practically, this can mean standardized formats, naming conventions, and validation techniques upon ingestion. From Amazon Comprehend's use case, the company reports using confidence intervals when identifying customer sentiment, so the analysts know exactly how accurate the data is. A study for The International Conference on Machine Learning and Applications reports Amazon Comprehend's service to "process large amounts of unstructured text with high accuracy." (Bhatia et al 2019). This use case scored highly because data input is deemed accurate, whereas a use case with unclean or untrustworthy sources would score lower. Once the accuracy of data is addressed and executed, a company then can focus on the continual improvement of their data pipeline and create a system for ensuring data reliability.

Data Reliability: The third essential point of data governance addresses the reliability of data from an ongoing maintenance perspective. This point is crucial for any sort of automated work a company would be interested in pursuing. Building off of accurate data, a company can automate processes and reduce time needed for deliverables, but this poses risks to automated systems malfunctioning and creating a need for regular maintenance. A company should establish periodic quality checks to ensure systems are operating exactly as intended. Connecting back to Amazon's Use case, accurate and trusted data sources (company's database and review sites) were established allowing automation to run smoothly. Automating a monthly or quarterly report on new customer sentiment analysis can provide useful insight and is a simple process that would not require high maintenance once built. This use case scored highly because maintenance is simple and infrequent whereas a use case that requires constant bug fixes and complex maintenance would score lower. Once a company's system is operating effectively and efficiently, data scientists can start to build incredible predictive models, work on artificial intelligence for solutioning, or building machine learning capabilities to address a need. However, data scientists need to ensure their work isn't compromised by risk or biases.

Data Ethics: The last essential point revolves around ethics of data, specifically for what data scientists are building. While building models, analytics, and insights create excitement in a company, it is important to recognize the fundamental importance of risks and biases in each use case. When looking into building models, prescriptive or predictive, there are inherent biases that a data scientist could unknowingly adopt into the model. Providing training and guidelines will allow a data scientist to ensure their models are ethically sound. False positives and negatives naturally occur in predictive spaces, and it is important to build in support systems to address these issues. Amazon's use case is definitely prone to false positives and false negatives, but one can find comfort in the vast training this model has undergone. With over 20+ client companies with large databases, this model has gone through extensive testing to minimize biases and produce insights for any demographic. This use case scored highly because it's model is extensively tested to deal with multiple types of data and eliminate biases, whereas a data structure with no rules, regulations, experience, or training would have scored lower

4. Conclusion | By applying the four main points from the Data Governance and Data Compliance Matrix as we did to Amazon Comprehend customer sentiment analysis use case, this process can be followed to determine the ease of implementing any data governance plan. With a strong data governance leadership, detailed rules and documentation, and careful data management planning, any company can be assured that their data is private, protected and precise.

Data Governance and Data Compliance Matrix

Use Case	Reliability / Maintenance	Accuracy / Data Quality	Security / Privacy	Ethics	Overall Ease of Data Governance and Data Compliance
Customer Sentiment Analysis: Amazon Reviews	5	4	5	4	4.5
Customer Churn: Valpak	5	3	4	3	3.75
Customer Segmentation: MetLife	4	3	3	3	3.25
Targeted Marketing: 7FAM	3	3	4	2	3
Customer loyalty program: Starbucks	4	2	3	3	3
Augmented Reality: Home Depot	2	3	3	2	2.5

References

“Amazon Comprehend.” *Amazon Web Services, Inc.*, 2021. https://aws.amazon.com/comprehend/.

Bhatia, Parminder, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. “Comprehend Medical: a Named Entity Recognition and Relationship Extraction Web Service.” *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019.

“Data Privacy FAQ.” *Amazon Web Services*, Inc., 2021. https://aws.amazon.com/compliance/data-privacy-faq/.