***Apply your understanding of the Data Governance pillars to your industry profession or one that you are seeking. Describe the important pillars of an organizational issue that you are aware of and discuss the reasons why. Support your answers addressing the benefits and focus of the pillars.***

I work for a startup in Fintech Space, so essentially I am touching both the financial industry and technology. So in this context, I wanted to highlight the data breach issue that occurred with Capital One. Capital one is in the list of top 10 banks in the United States. In this data breach scenario, an outside person gained access to personal information of the customers who had applied for credit cards from 2005 to 2019, this included names, addresses, zip, mobile, email, dates of birth, income, credit scores, credit limits, balances, payment history, contact information, etc. This issue was primarily due to misconfiguration of the firewall with additional permissions to read and list files from S3 buckets in AWS, the individual able to retrieve identity and access management credentials and assume a role on the server which had broader permissions for S3 buckets where the sensitive information was residing. So considering this data breach, if the data governance frameworks were in place along with the implementation of four pillars, then this data breach could have been easily avoided. Below are the 4 pillars

1. Policies, Standards, and Strategy
2. Information Quality
3. Privacy, Compliance, and Security
4. Architecture and Integration.

In this case, specifically, these 2 pillars would have been of higher significance

* Privacy, Compliance, and Security
* Policies, Standards, and Strategy

Pillar 3 highlights

1. Access management

Fine-grained, role-based, access controls based on the principle of least privilege (POLP) could have limited data loss. With POLP, users are granted permission to read, write or execute only the files or resources they need to do their jobs. Security needs to be configured across multiple layers in the IT or cloud infrastructure from firewalls to the data stored.

2. Securing and protecting sensitive data.

For protecting sensitive data encryption need to be in place as well data can be anonymized. This could have prevented exposure to sensitive (PII - Personally Identifiable Information) data.

Apart from Pillar 3, Pillar 1 (Policies, Standards, and Strategy) could also have been of significance over here. It talks about decision-makers deciding the access rights, which was a failure over here. Decision-makers did not factor in the cloud-related roles and permissions and did not enforce any controls or establish IAM roles related policy. Even if the IAM related policy was established it was not properly audited or was not implemented correctly.

***Identify one pillar in the Data Governance framework and define the activities to develop a governance strategy for a particular industry challenge. Describe the organizational issue/challenge and discuss why this type of governance program would be beneficial.***

Another pillar of the data governance framework which is of importance is "Information Quality".

Information/data quality is essential for accurate outputs/results from data analysis. Here are some of the dimensions against which the data quality can be measured

* Completeness – Identify missing data.
* Consistency – Data does not provide conflicting information. Values from the data are interpreted accurately.
* Conformity – Data stored in with consistent data type, same data columns do not have data stored in inconsistent data types.
* Accuracy – Data represent the exact inventory that exists at any point in time. E.g. there is inconsistency in the data inventory of a particular product in the system and the physical inventory in the data warehouse.
* Redundant data - Avoid duplicates.
* Data Integrity.

Following actions can be taken as part of data governance strategy for improving Information Quality

1. Typically data governance framework consists of the DG Steering committee, practice and working group. So as part of the working group, a lead should be identified who is responsible for data quality.
2. Identify and implement tools to raise data quality issues which can then be tracked till closure. For each data quality issue raise identify and perform root cause analysis. Data quality issues will shy away if and only if the solution addresses the root cause.
3. Educating developers, system users about information quality and raising awareness in general for reporting any data quality issues.
4. The big organization has large business-specific entities. So as part of the data governance activity for each of those business entities, data quality metrics need to be defined. E.g. Customer entity should not have any duplicates so the metrics should be Redundancy = 0%. These metrics have to be reported, tracked and reviewed periodically. In this case, as part of the initial phase of DG, goals should be defined and all reporting should be tracked against those goals.
5. Profiling of the data collected. This could be done using a good data profiling tool. Identify and implement tools for data cleansing, scrubbing
6. In the case of a data warehouse of big data, an appropriate data ingestion pipeline could be build which ensures the data quality.
7. Publish centralized metadata across all business entities.

Organizational Issues/Challenges related to "Data Quality"

* In an organization, data is generated by its applications or it comes from third-party data sources e.g. SAAS application. In these cases data model could be different, the quality of the data cannot be guaranteed.
* Organizations have N number of applications and application teams. These application teams are working in silos and in the process are storing the redundant data for the same business entity.
* Large Organizations with many applications are looking at data-lake for getting the unified view of their data asset, but in most case, the data lake becomes a data dumping site without any quality and any other type of gates in place.
* Data consolidation – In the case where systems are merged due to acquisitions or older systems getting replaced, this leads to overlapping of data entities.

All the above issues of "Data Quality" can be addressed through Data Governance and there are benefits to organization due to following reasons

* Unified data view (360-degree view) across all business entities across the entire organization.
* Improvement in the quality of data results in making better business decisions that are derived from the data.
* The success of the adoption of newer technologies like Artificial intelligence and Machine learning depends on data quality.
* Saves cost and time by automating the data quality process. Less cost and less time lead to better productivity.
* Improved data quality results in higher customer satisfaction, customers are ensured that their PII data is handled with care and they are handled with better service resulting due to increased productivity and reduced operational costs.

Responses

I could not agree more with the diagram which articulates better the data quality better the business outcome. With quality gates and infrastructure in place it becomes easy for data driven organization to scale up with higher confidence with the assumption that their data is trustworthy.

When the data quality is of high standard then organization can perform targeted marketing in an effective way. This is possible because of rich demographics data available for customers. Earlier without this data, it would be to broader audience, leading to waste of money on advertising with people who are unlikely to but the product or service.

For large organizations simply fixing the data errors should not be the only solution, they should embrace the path to changing the culture which results in better information quality. This does not happen with a single employee or team. It should come from top, right from the leadership team which will eventually trickle downstream to individual teams.

Good data quality leads to single source of truth to which everyone agrees and relies on. No more confusion or finger pointing about which system has most accurate set of data.

Response 2

Just wanted to share some more details on Data Architecture and Integration. It provides an information on where really data exists and how the data flows through various systems, it also highlights the data transformation performed when data moves from one system to another.

In general data architecture has to be aligned to business drivers, should only focus on core data, and should be actionable as well as evolutionary.

With SOA based models it become easy maintain the data in single source instead of moving it through various systems. Integration between applications is through web services, nowadays light weight micro service based architecture has become very common theme for integrating data across various systems.