Chapter 6

Integration - pt 2

[Come back and write this.]

6.1 When to start

[Come back to write this.]

Data Volume and Complexity

As organizations grow, they accumulate data from various sources. This data growth becomes difficult to manage and analyze efficiently. Data could be in different formats, both structured and unstructured, and reside in disparate systems. This ever-growing load cannot be encompassed by ad hoc tools like Excel, once commonly associated with the task. Organizations must look to new data tools and a more integrated data framework. Centering around data helps alleviate the current stress on the system and ensures future data needs are met.

Data Management Tools

Data stored across disparate locations creates a hindrance and challenge to optimal efficiency. Data is stored in Excel, PDFs, databases, and much more all across the organization. As demand for this data increases, so does the need for robust data

management tools to support integrating data. This alleviates the strain of maintaining current data systems and ensures valuable data-driven insights.

Scalability

We've discussed in great detail that Excel spreadsheets have long been a primary method of handling data. As data becomes more intricate and industry dynamics change, these traditional solutions sacrifice scalability. Spreadsheets, flat files, and even disparate databases require maintenance. They require personnel to read through data and extract important information. Organizations cannot scale and add additional data sources because they are too busy trying to understand what they currently have. Organizations are throttled on their ability to use data creatively. These sources also lack compliance standards for sensitive information. With these traditional data solutions, organizations throttle their data scalability.

Growth and Innovation

Choosing a simple run-and-maintain organization will no longer suffice within a competitive landscape. This compounds when competing enterprises embrace new data techniques and technologies. Organizations must integrate advanced data analytics methodologies and adopt robust tools for data transformation. Doing this will ensure they enhance their decision-making and create nuanced strategic insights. As machine learning and artificial intelligence advance, so does an organization's ability to extract deeper insights into its data. Organizations can no longer afford to maintain the status quo.

Collaboration and Coordination

Another criterium affecting administrative rigor is adequate team coordination. A cohesive data infrastructure, with implementations of standardized APIs, establishes smooth communication between departments working on company data sets. As

businesses' data use intensifies, it's worth scrutinizing the consistency of information flow and introducing appropriate improvements.

6.2 Metrics for a data-centric organization

Organizations must first create a strong data culture. Data culture is how an organization interacts with data and data outcomes. Do teams trust their data, or are they skeptical? Are they able to make decisions based on data? What will it take to build trust in that data? Confronting these is key to building trust with leadership and empowering the data team. It encourages the team to experiment with data in new ways, enabling new insights and opportunities. A strong data culture ultimately builds trust and unlocks data as a strategic asset.

Next, organizations must establish a baseline for comparison. Assessing the current state of the organization and its struggles will inform goal-setting for data-driven initiatives. It's essential to consider what pain points are currently holding the organization back. This could be costs, data quality, speed, or something else entirely. However, identifying these blockers will help inform your data's best- and worst-case end states. A thorough understanding of the organization's current position will allow future improvements to be observed more intelligibly. If this is too challenging to assess, look to the competition. Understand how the competition utilizes data and makes data-driven decisions.

This baseline sets the stage for your data metrics. This point will be measured as the data journey progresses. Below, we walk through a couple of metrics organizations can utilize. These may not all apply to every data journey. However, they can help get the data organization to start capturing the data's value.

6.2.1 Data quality

We talked in section 3.5.3 about the importance of data quality. Three things should be captured for data quality: accuracy, completeness, and consistency. Accuracy refers to how well the data aligns with reality. Completeness gauges whether all

required data elements have been gathered. Both involve manual data checks for spreadsheets. For small data sets, this works fine. For larger data sets, accuracy and completeness should be automated. This can be done by setting up data quality rules in your data ingestion process. These automated data quality rules flag errors and inconsistencies in the data. These rules can be set up to check for things like missing values and incorrect data. The data quality rules can be run regularly to identify issues and improve accuracy and completeness.

Consistency measures how well data elements correspond to each other. One way to determine consistency is to compare individual data elements, searching for differences in data types and formats. For instance, a non-numerical value in a field intended for numerical data should be identified as inconsistent. Like above, automated tools can be utilized to discover inconsistencies by scrutinizing data patterns.

Data duplication, missing data, and validation are additional things to consider. These can identify potential data quality problems or risks in decision-making.

6.2.2 Data freshness

Data freshness refers to how up-to-date the data is in your system. Fresh data ensures that the information being used for decision-making, analysis, or reporting is current and relevant to the present situation. Some metrics to consider for data freshness include:

- Data Latency: The time delta between a data update in the source system and when it's available in the user-facing system, e.g., data warehouse.

- Data Refresh Frequency: The frequency with which the ETL process is run to refresh the target data repository. This metric can be captured in as little as minutes and up to monthly.

- Source Data Staleness: Monitoring frequency with which the data source is updated to understand the data source's freshness.

- User Data freshness feedback: Feedback loop from users about the freshness of the data they consume explicitly. Poor feedback will potentially reveal issues with the freshness that may be addressed.

These metrics can be captured automatically and displayed in reports. An example is an ETL Completion Report. This summarizes your ETL workflow's missing data, duplicates, unhandled exceptions, and data rejections. It's used for troubleshooting issues and improving the overall data freshness.

6.2.3 Speed

We discussed speed for Excel and the cloud in sections 3.5 and 4.4.1. As organizations grow along their data journey, speed starts to increase dramatically. Manual data manipulations that took days or weeks in Excel are happening in seconds. However, there is always room for improvement. Organizations should now expand their metric collection to encompass the entire data lifecycle. This starts from data ingestion to processing and transformation to end-user data retrieval. Organizations can also consider metrics like how long it takes to onboard a new data user. Some metrics are captured below:

- Data Extraction Time: Measure the time taken to extract data from source systems. This can be recorded in terms of records per second or the total time taken to complete the extraction process. It can also encompass full vs incremental data load times.

- Data Transformation Time: Measure the time taken to transform the extracted data using reference data and business rules.

- Query Benchmarking: Measure how quickly the data warehouse or data store can respond to a query. This can double as a good set of test cases to ensure your data warehouse is up and running.

- User onboarding time: Measure the time it takes for a new user to start using transformed data.

Capturing these metrics and analyzing their fluctuations can identify potential causes for slowdowns and bottlenecks in your data workflows. Regularly monitoring and improving these metrics will ensure a stable, efficient, and fast data processing environment.

6.2.4 Cost

As an organization moves along the data journey, cost becomes increasingly difficult to capture. This is largely due to the high investment cost required to succeed in the journey. This often results in leadership wanting to see the return on their investment as soon as possible. It is important to articulate the value in the data journey, whether that is improved revenue, cost savings or avoidance, and/or better or faster decision-making to exceed the costs identified. Over time, calculate the return on investment, ROI, to understand the overall value. Some things organizations can capture are:

- Decision-making: Centralized and standardized data storage can lead to quicker and more reliable data decisions. Reducing time to decision can generate financial incentives for organizations. This leads to better business strategies and cost savings in the long run.

- IT infrastructure costs: Organizations can avoid investing in expensive on-premises hardware and software solutions by moving to a cloud-based data warehouse. They also save by removing ongoing maintenance and upgrade costs.

- Risk and data security: Data warehouses often include built-in data encryption features, role-based access controls, and activity monitoring. These provide better data protection as compared to maintaining data in disparate locations. This risk reduction can be a sizable cost savings for companies with highly sensitive data.

- Reduced overhead: Organizations can retire legacy data management tools for the cloud. This reduces the software licensing costs for the organization. Addi-

tionally, this consolidation requires much less time and energy spent maintaining disparate systems.

When creating metrics, ensure they are specific. Ie. decision time reduced from one month to one week. This leads to faster decisions on prospects reducing uncertainty in our investments. In this example, it can be difficult to calculate the value of reducing time from one month to a week. However, it can be much easier to understand the benefits of deciding on prospects faster than the competition.

6.3 Exploring trade-offs and getting buy-in

Moving toward a data-centric organization can be a complex process. It will require your data teams to deeply understand the consequences of this transition, both good and bad. This must be communicated to leadership to ensure their support throughout the process. Below we walk through some ways for organizations to obtain buy-in and trade-offs to consider with each.

Develop a clear and compelling strategy:

- Buy-in: Leaders can understand and weigh the benefits by presenting a clear strategy. Highlighting the impact on key organizational goals and objectives can help leaders recognize the value of investing in data-driven initiatives. Focus on the value added to the organization and how the strategy achieves that value.
- Trade-off: Creating a comprehensive strategy requires time and effort to gather and analyze relevant data. It requires the data organization to articulate the value proposition of a data-centric organization. It may involve conducting workshops, in-depth data analysis, and modeling return on investment.

Demonstrate quick wins:

- Buy-in: Quick wins demonstrate the tangible benefits of data initiatives and provide evidence of their potential. By showcasing early successes, leaders can gain confidence in the value of data-driven decision-making and be more inclined to support further investments.

- Trade-off: While quick wins can generate buy-in, they may not capture the full potential of data-driven decision-making. It's essential to balance delivering immediate results and pursuing longer-term, transformative initiatives that require more time and resources.

Engage with stakeholders:

- Buy-in: Leaders can create a sense of ownership and alignment across the organization by involving stakeholders in the data journey. Engaged stakeholders are more likely to support and advocate for data initiatives, facilitating buy-in at various organizational levels.

- Trade-off: Engaging with stakeholders requires additional overhead to communicate the benefits of data initiatives and address their concerns and feedback. It may also involve aligning different departments and teams, which can be challenging.

Measure and communicate success:

- Buy-in: By measuring and communicating the success of data initiatives, leaders can showcase the positive outcomes and the impact on the organization's objectives. Quantitative and qualitative evidence of success reinforces the value of data-driven decision-making and strengthens buy-in among leaders.

- Trade-off: Establishing robust measurement and evaluation mechanisms requires ongoing effort and resources to collect, analyze, and report relevant data. It may involve implementing data tracking systems, defining KPIs, and conducting regular performance assessments. Being strategic about capturing only value add metrics helps with unnecessary overhead.

6.4 Real-world use cases

In this section, we dive into some real world use cases. We will talk about the challenges of data and how organizations are looking to alleviate those challenges.

6.4.1 Ford

Ford launched Smart Mobility in 2014 in an attempt at digital transformation. The goal was to build, "digitally enabled cars with enhanced mobility." [22] However, this project was designed as a separate entity to Ford's central business. This siloed digital effort was a huge failure and led to significant financial repercussions.

Ford realized becoming data centric cannot happen in a silo. It takes everyone to buy-in to build this organization with data. They realized data had to exist at the very forefront of what the cars they build. To do this, they partnered with Google to modernize IT systems and leverage data from dealerships, repairs, warranty services, customer purchase patterns, etc. Over 4,600 data sources, both internal and external, were fed into a central data lake. They are now able to leverage this data to launch usage-based auto insurance programs, offer personalized technology and services, accelerate product development. They are also able to help their customers by providing real-time notifications for maintenance and other activities. [22]

6.4.2 Biology

The biology community is also undergoing a data transformation. Wratten et al. discuss the use of workflow managers to enable easy and reproducible pipeline development. However, there is a gap in expertise when it comes to data. Ph.D. biologists often don't have data engineering skills. On top of that, there are hundreds of these workflow managers. This leads to fragmentation in their tooling. Biologists are left confused and deterred from starting their data journey.[30]

Many organizations suffer from this fragmentation. One way to mitigate this is to create clear standards. This allows organizations and communities to work together to create better, more accessible tools. Fragmentation can also be mitigated by level-setting on a common language. This common language can enable teams to work together to create data tools together and not in silos. Both of these enable the creation of reusable tools. Creating reusable and portable tools allows for non-technical users to plug and play in their data workflows with limited data skills.

6.4.3 Banking industry

The banking industry is facing challenges due to cybersecurity and data integration. With increased cybercrime, data breaches, and hacking, the banking industry faces significant challenges in ensuring data security and protecting customer information. Integrating various data sources, both structured and unstructured, poses challenges for the banking industry. Banks need efficient and effective data integration mechanisms and strategies to identify better business insights from their integrated data.

Companies use AI to detect cybercrime by leveraging Cyber Data Lakes as a central platform housing multiple data sources. These also act as analytics platforms for detection engines. Advanced Artificial Intelligence (AI) techniques, such as Deep Learning and Graph analysis, can then be run on top of these data lakes. AI-based, early warning, multi-stage systems are being used to detect malicious Trojan activities. These detections work across the whole data lifecycle and encompass internal and external sources. They can even detect malicious activity ahead of the actual spear phishing campaign.[15]

Chase Bank is working to get fully integrate their data pipelines. They are working to establish a data marketplace, build data products, curate data assets, and deliver data through modern pipelines. This allows them to integrate various types of structured and unstructured data to achieve a full picture of the customer. The data marketplace also allows them to break down data silos between different departments. All users must access their data through this central repository. Data marketplaces also enforce data governance helping to secure confidential business data.[2]

6.5 Conclusion