Final Report: Data Source Report

Dawn Stratchko

Capella University

ANLT 5020: Data Sources for Analytics

Professor Kyle Camac

June 9, 2024

Data Collection Methods

Organizations like Vila Health navigate complex challenges and opportunities to maintain a continuous flow of accurate and comprehensive data. As a healthcare system encompassing numerous hospitals, clinics, and healthcare providers, Vila Health illustrates the complexity of modern healthcare. However, Vila Health needs help with data maintenance issues, particularly acquiring the Uptown Wellness Center. Integrating different data sources, accuracy of demographic records, and regulatory compliance pose significant barriers to the organization's mission of delivering high-quality care.

Over three meetings with Rick Susskind, a data analytics mentor, the complexities of data quality and ETL processes in healthcare, mainly focusing on challenges faced by Vila Health, were discussed. He emphasized the importance of understanding data sources and quality before creating solutions, citing issues with redundant data and the need for a clear understanding of business problems (*Analytics Internship: ETL and Data Quality*, n.d.). Rick also discussed the importance of understanding data creation processes by scheduling interviews at the Uptown Clinic. He emphasized the importance of consulting with IT personnel regarding interoperability issues.

There were a total of eight staff members available to interview. Rick highlighted the need to prioritize interviews due to time constraints that allowed time for only four out of six possible interviews after the second meeting and then went on to offer two more staff members after the third meeting. He suggested the importance of understanding how data is created and entered into various systems and the significance of understanding system constraints and strategic vision for data warehousing. Through these interviews, the following information was obtained:

- Amber McBride, Clinic Manager
    - Amber had two interviews. In them, she revealed that the clinic collects patient demographic data through multiple points, such as patient registration, provider interactions, and billing departments, and faces challenges integrating external data into the electronic health record (EHR). She emphasized the difficulties with merging records from newly acquired practices, highlighting the risks to patients and the institution's credibility due to data inaccuracies, and suggested that top-level leadership commit to standardizing data collection and management practices.

- George Fink, Director of Clinical Operations
    - Over two interviews, George highlighted the critical role of data quality in strategic planning and the necessity of reliable data for informed decision-making, stressing the need to identify and understand conflicting data sources. He also emphasized the interdependencies between different systems and how changes in one area affect others, noting that the absence of a top-down data management strategy impeded previous efforts to address master patient index (MPI) problems.
- Janie Poole, Nurse Practitioner
    - In her interview, Janie explained the limited role of nurses in data creation and maintenance. Still, she described instances where they may correct or update demographic data during patient interactions or through the nurse hotline.
- Elizabeth Douglas, Office Manager
    - Elizabeth's interviews highlighted the patient demographic data collection and updating processes at the clinic, interactions with billing and insurance departments affecting this data, the absence of standardized workflows for data entry, and the informal training process for new hires, noting that mistakes in the MPI are typically identified reactively rather than proactively (*Vila Health Analytics Internship: Identifying Data Sources*, n.d.).
- Blake Carter, Director of Information Technology
    - Blake's interview highlighted EHR reporting and ETL issues, especially with acquisition data migration. He discussed the importance of adhering to state and federal regulations and stressed the need for close communication with EHR vendors to address non-standard data. Lastly, he suggested that the main challenge is communicating the importance of data quality to all stakeholders.
- Jason Armstrong, Data Analyst
    - In Jason's interview, he noted that the MPI is not integrated with the EHR, leading to inconsistencies. Discrepancies between the billing system and the EHR regarding address and race/ethnicity data were discussed. He stressed the need for a comprehensive strategy

for data management to avoid patchwork solutions (*Analytics Internship: Limitations on ETL Modeling Transcript*, n.d.).

Data Collection Instruments

As Vila Health has expanded its reach through mergers and acquisitions, challenges have emerged, particularly regarding the integration and management of data. In this context, the Uptown Wellness Center faces broader data complexities, where the electronic, paper-based, and image scan data collection methods intersect with accuracy, completeness, and integration issues. An in-depth examination of the data collection instruments and methods utilized within the Uptown Wellness Center revealed critical insights into the complexities and opportunities inherent in healthcare data management. Based on the knowledge gained from the interviews, the data collection instruments and methods used included:

- Electronic Systems: The primary method for data collection seems to be through electronic systems like the EHR, which stores patient demographic data, clinical notes, and lab results. Some patient data may be electronically updated or modified based on information received from insurance companies or other external sources.
- Paper Records: Paper forms are relied on for pre-registration, referral forms from other systems, and specialized services like counseling and case coordination, especially for populations such as LGBTQ individuals and transgender patients. However, these paper forms are not always designed with integration into the EHR in mind, leading to potential inconsistencies.
- Image Scans: Information not directly entered into the EHR is scanned and linked as images. This system is not ideal as it limits data accessibility, but it is currently used for any documentation created outside Vila Health.
- Manual Entry: Registration clerks manually enter or update patient data from paper forms into the electronic systems. Fixes to data issues are often done manually when discrepancies are identified.
- Verbal Interactions: Clinical staff might verbally verify or update patient demographic data during interactions with patients, such as when patients arrive for appointments (*Vila Health Analytics Internship: Identifying Data Sources Transcript*, n.d.).

Several issues and challenges have arisen from the current circumstances, including data quality, integration problems, and process inefficiencies. Concerns about data quality involve inaccuracies, omissions, and duplicates in demographic data, undermining the reliability of reports and analyses. Integration challenges are evident in the difficulty of integrating the MPI of the Uptown Wellness Center with Vila Health's EHR, leading to discrepancies and data retrieval issues. Additionally, inefficiencies in the data cleansing process, marked by unpredictable intervals for checking duplicate records, allow errors to persist and impact data reliability.

Assessing data maintenance approaches reveals challenges based on the database and system type; relational databases are typically used for EHR and billing systems to manage structured data, but the MPI, crucial for patient data maintenance across systems, is often separate from the EHR, causing integration issues (Lee et al., 2020). EHR systems face data migration and integration hurdles, while billing systems operate independently and handle key data attributes differently. Custom interfaces connecting these disparate systems often exacerbate integration challenges, and a historical patchwork approach to data issues has resulted in a lack of a cohesive data management strategy.

Data Maintenance Approaches

Practical data maintenance approaches ensure operational efficiency, strategic decision-making, and quality patient care delivery. For Vila Health, an in-depth exploration of the data maintenance approaches employed is necessary to create the best plan. By analyzing the data deficits, integration challenges, and regulatory considerations, we can determine the importance of solid data maintenance strategies, which can help drive organizational success and improve patient outcomes. Based on the knowledge gained from the interviews, several data maintenance approaches are evident:

- Patient demographic data is collected at multiple touch points, including pre-registration, provider interactions, discharge, billing, and payments. This data is gathered electronically and through paper forms, with varying levels of consistency and integration with the EHR system.
- Integrating acquired clinics' MPI with Vila Health's EHR poses significant challenges. Discrepancies in data accuracy and completeness hinder the seamless integration of patient records, leading to errors, redundancies, and difficulties in data reconciliation.

- While periodic audits are conducted to address duplicate records, the frequency and effectiveness of these audits are inadequate. The interval between audits is too long, and proactive measures are lacking to detect and address inaccuracies within non-duplicate records. Data cleansing rules are not sufficiently stringent or consistently applied (*Vila Health Analytics Internship: Identifying Data Sources*, n.d.).
- Data is sourced from various systems and processes, including patient registration, provider interactions, billing, and referrals. However, inconsistencies in data formats and collection methods exist, especially concerning documentation originating outside Vila Health, which is often scanned as images rather than entered directly into the EHR.
- Poor data quality and integrity affect strategic planning, operational efficiency, and organizational decision-making processes. Inaccurate or incomplete demographic data hinder trend analysis, population health management, and financial management, leading to increased costs and inefficiencies.
- The healthcare industry's complex regulatory environment, including requirements such as HITECH, ICD-10, and HIPAA, presents additional challenges for data management. Interoperability with electronic health records, billing systems, and health information exchanges is essential but difficult to achieve.

More stringent data cleansing standards, improved integration processes, and proactive measures to detect and correct data inaccuracies are needed to address these challenges. Investing in technology upgrades, staff training, and standardized data collection procedures can enhance data quality and streamline data maintenance across the organization.

Data Quality Approach

"A data quality framework (DQF) includes the procedures, methods, standards, and tools businesses use to analyze, manage, and improve the quality and ethical standards of their data" (Zendata, 2024). To create a DQF for the Vila Health Clinic, we will include the standard elements:

- Accuracy
- Completeness
- Consistency

- Timeliness
- Transparency
- Bias identification and mitigation

To establish the DQF, we will need to complete the following for Vila Health Clinic:

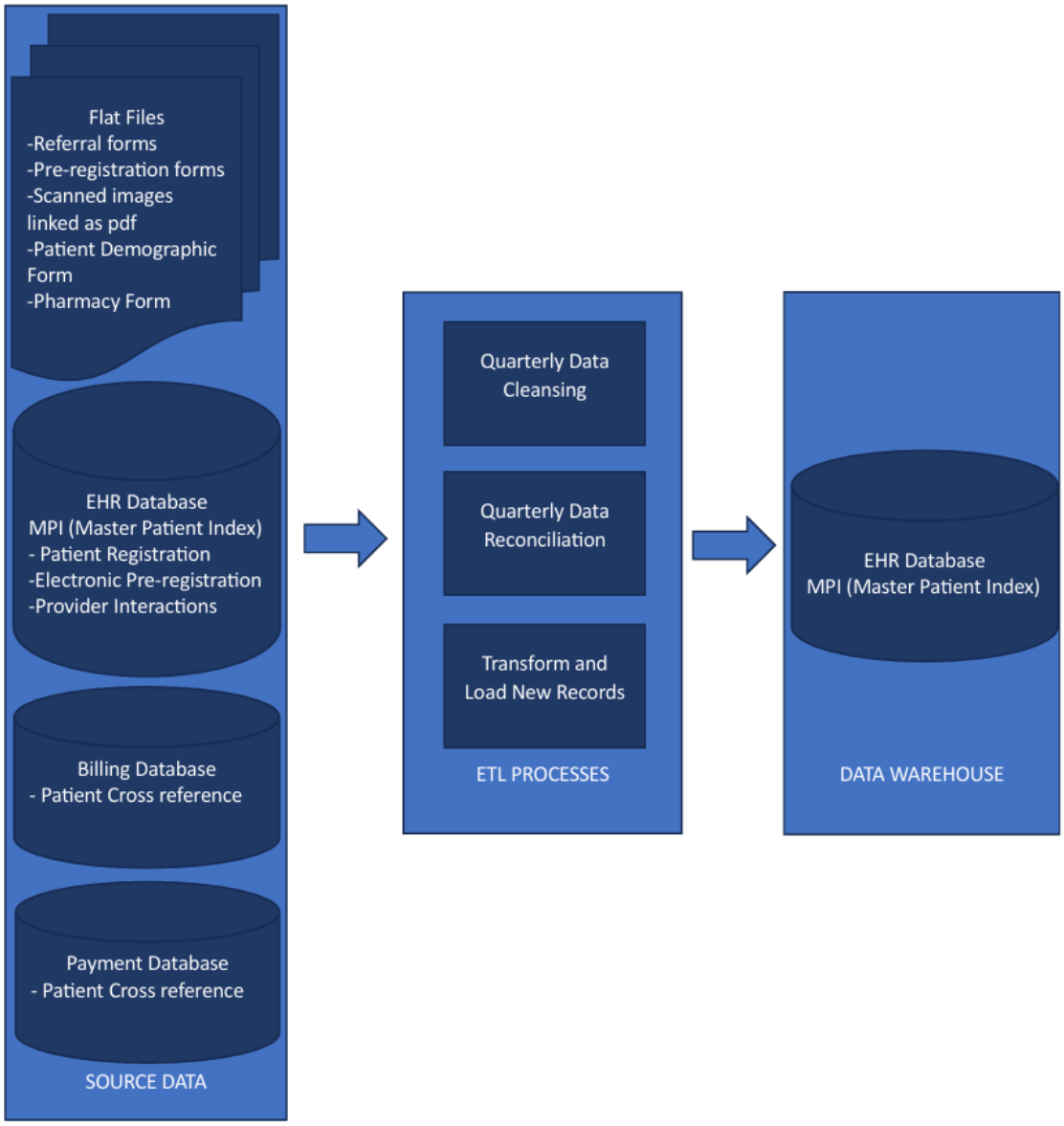
- Assess the data quality and determine where issues are.
- Identify and set objectives and standards for data quality.
- Design and implement data quality management policies.
- Solidify the tools and infrastructure needed to manage and monitor data quality.
- Implement a monitoring and continuous improvement mechanism.

Current ETL Process Used in Vila Health

The ETL process for Vila Health has several critical issues in data collection, extraction, and storage (see Figure 1 for the current ETL process). Integration of external data sources, such as paper forms and image scans, has caused redundant data entries, requiring methods to eliminate duplication. Multiple data sources lead to conflicting details due to inconsistent data labeling, necessitating a synthesis of standard field names. Urgent updates to patient demographic information are crucial, with changes needing to be immediate and trackable by timestamp. Current data cleansing techniques fail to accurately match patient information with insurers' records, causing billing issues. Additionally, there is a substantial backlog of scanned documents from external sources. To improve accuracy, EHR reconciliation, currently done every six months, should be increased to quarterly or monthly. The ETL process will enhance data accuracy and efficiency by addressing these challenges comprehensively.

Figure 1

Vila Health Current Data Flow and ETL Process



Improved Data Cleansing

A strong strategy must be developed to enhance data cleansing and pipeline efficiency at Vila Health. Data pipeline monitoring involves “continuous tracking, observing, and evaluating data through different stages” to ensure compliance with the General Data Protection Regulation (GDPR) and other regulations (Atlan, 2023). Key recommendations include:

- Set data monitoring objectives.
- Capture relevant timestamps, data sources, and transformation details to maintain detailed logs.

- Ensure accuracy, completeness, consistency, reliability, and timeliness. Implement tools and scripts to evaluate these metrics and set necessary benchmarks.
- Implement immediate compliance adherence.
- Consistently review system metrics and maintain data lineage documentation.
- Conduct bi-monthly audits and enable feedback loops to address issues promptly.
- Ensure the system can handle growing data volumes and complexities.
- Encourage continuous monitoring practices across the organization (Atlan, 2023).

Implementing an OCR program for document management can streamline converting scanned text into usable data within a Java GUI interface and structured query language in the database. Verification steps must be adhered to for compliance with new policies. Microsoft Fabric, specifically its OneLake and Data Wrangler components, is recommended for creating a unified and manageable data environment. These tools integrate seamlessly for ETL processes, with Data Wrangler aiding in data cleansing through a combination of GUI and Python code using Pandas. The cleansed data can be saved in various formats like CSV or JSON. Leveraging PowerBI alongside Data Wrangler for ETL tasks provides a robust data science experience (*How to Use OCR Software for PDFs in 4 Easy Steps | Adobe Acrobat*, n.d.).

Maintaining a familiar environment is crucial, and combining SAS with Microsoft Fabric is strategic for Vila Health. SAS is known for its compliance in the healthcare industry and offers efficient data cleansing techniques:

- Use programming to extract and test samples to effectively identify and solve data issues (see Figure 2).

Figure 2

SAS Data Import of Selected 8000 Cells

```
PROC CONTENTS DATA=WORK.IMPORT;

data import;
set work.import (obs=8000);
run;
```

- Use 'data import' and 'set' statements for sampling and 'IF-THEN' statements for data corrections. Data from Fabric can be exported to SAS for further processing (see Figure 3).

Figure 3

SAS IF, THEN Statements for Name Corrections

```
data work.import;
set work.import;
 if LastName = "Simth" then LastName= "Smith";
 if LastName = "Jone s" then LastName= "Jones";
 if LastName = "Hernadez" or LastName= "Hernandnez " or LastName="Hernadnez" then LastName= "Hernandez";
 run;
```

Communication

Transitioning to Microsoft PowerBI for report development will streamline communication and sharing. Reports can be generated and exported in multiple formats (Word, Excel, CSV, PowerPoint, PDF) for ease of use (Pykes, 2023). Key performance indicators (KPIs) must be established before implementing the ETL process:

- Regularly scheduled meetings to review and develop policies for naming conventions, data structure, and compliance.
- Develop and launch training programs for data compliance.
- Document the outline of the data pipeline.
- Conduct bi-monthly audits with thorough tracking and follow-through.

Vila Health can improve its data cleansing processes by following these structured steps, ensuring efficient, compliant, and high-quality data management.

Data Collection Design

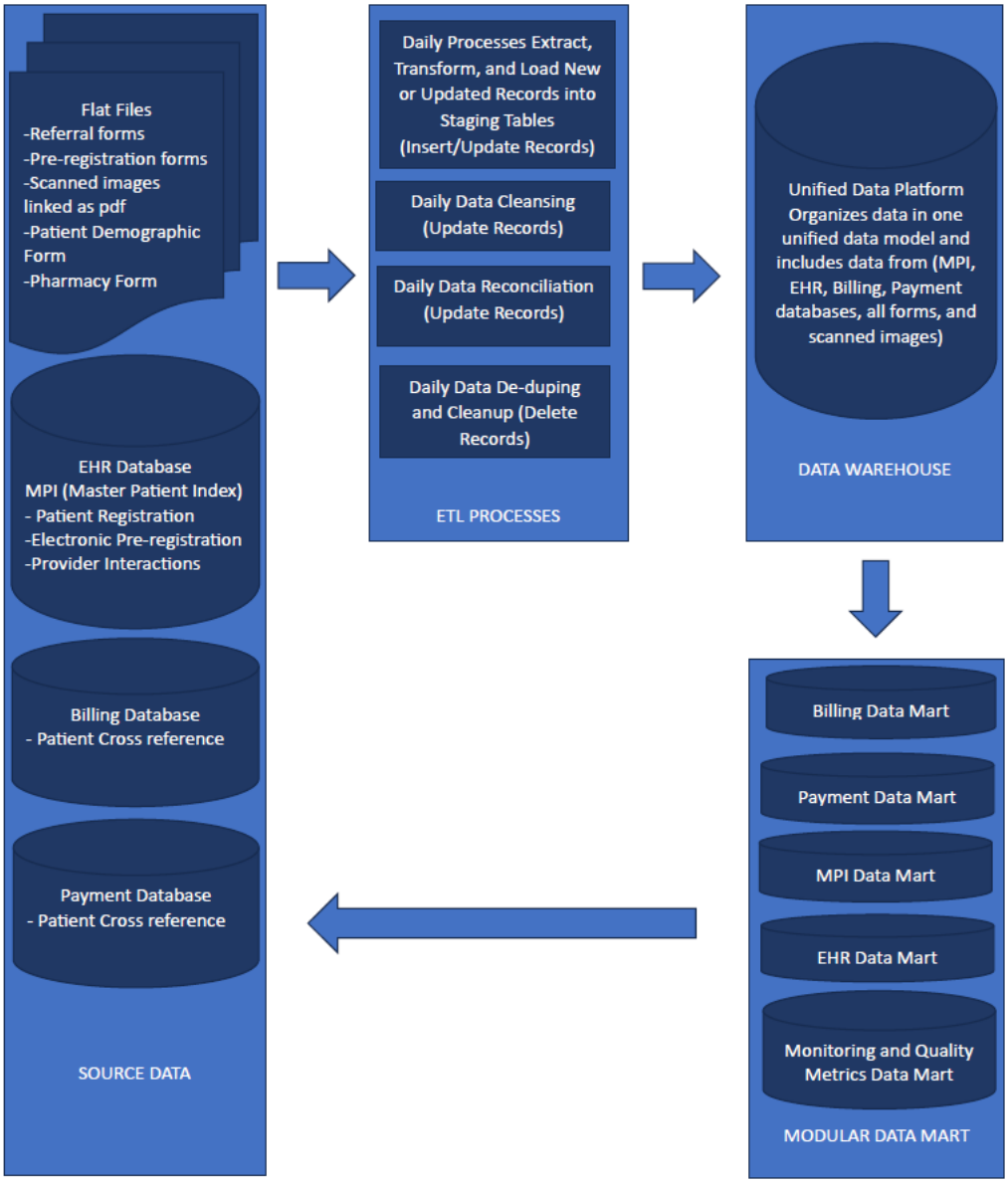
One of the significant focus areas will be integrating the MPI and EHR systems to work together seamlessly to eliminate discrepancies and improve data consistency. To do this, we propose investing in cloud-based unified data platforms like Databricks that can integrate data from many sources to ensure a smooth migration and compatibility with newly acquired practices (Databricks, 2023). See Figure 4 for an updated ETL process. The benefits of a unified data platform include:

- Simplifying big data
- Unifying data analytics and data engineering
- Enabling faster innovation
- Better understanding of the current and future architecture

- Documenting processes and understanding critical paths
- Balancing data governance and data management for high-quality deliverables
- Promoting a business glossary for better understanding of data terminology
- Promoting a culture of responsibility for data governance
- More easily understanding the impact of changes

Figure 4

Proposed Improved ETL Process



Due to its current data management practices, Vila Health faces significant privacy, security, and ethical concerns. Privacy issues arise from handling sensitive patient data through various means, including electronic systems, paper records, and verbal interactions, with compliance with privacy laws like HIPAA being crucial. Integrating data from new acquisitions like Uptown Wellness Center and using image scans for external documents add to these challenges, increasing the risk of unauthorized access and data breaches. Security concerns center on data accuracy and integrity, as inaccuracies can lead to incorrect patient information being used in decisions, posing risks to patient care and security. Interoperability and system integration issues can create security gaps, and reliance on paper forms and manual data entry increases the risk of data loss and breaches. Ethically, poor data quality can result in harmful clinical decisions and disparities in patient care, highlighting the need for standardized workflows and proactive error prevention to ensure data accuracy and equity in healthcare services. By addressing these concerns through targeted improvements, Vila Health can enhance its data management practices, ensuring patient information privacy, security, and ethical handling.

A structured operational procedure must be implemented as part of the new BI infrastructure to manage external metadata from outside sources and other divisions acquired through mergers and acquisitions. Using an operational source system as a staging area for all incoming data ensures clean, structured, and manageable data. The data is then extracted and combined into a single unit before being transformed into the data warehouse.

Improving the data collection system with scannable software that electronically tags necessary information is advantageous for form, paper input, and voice data. Tagged documentation will be saved using predefined criteria relevant to Vila Healthcare terms. This metadata can include details such as author, file size, and creation date, making it easy to search for relevant documents within your organization. This system can handle various documents, from supply chain billing and shipment tracking for medical devices and medication orders to patient insurance proof and mail-in surveys.

Implementing a metadata procedure and a formal data collection system in unison will enhance adherence to security and government regulations while streamlining communication with vendors regarding non-standard data (Ramakolote et al., 2023). As Blake Carter, Director of IT, and Amber

McBride highlighted, improving the quality of these matrices will provide consistency, flexibility, traceability, reliability, and recoverability.

Strategies and Recommendations

Our group outlined practical data maintenance approaches for Vila Health to implement to ensure operational efficiency, strategic decision-making, and quality patient care delivery. One limitation Vila Health, or organizations in general, could face when implementing changes is change management follow-through, training and retraining, and continuous improvement to adapt to any unforeseen impacts.

We indicated Vila Health implements a data quality framework when managing its data for accuracy, completeness, consistency, timeliness, transparency, and bias identification and mitigation. Additionally, we developed a strategy to enhance data cleansing and pipeline efficiency. This included implementing an OCR program for document management and utilizing tools that would integrate seamlessly into the improved ETL process. Lastly, we outlined a thorough data collection design and communication methodology to enable the MPI and EHR systems to work together seamlessly, eliminating discrepancies and improving data consistency. This all ties in to address the privacy, security, and ethical concerns.

References

Analytics Internship: ETL and Data Quality. (n.d.). https://media.capella.edu/CourseMedia/ANLT5020element23263/transcript.asp

Analytics Internship: Limitations on ETL Modeling Transcript. (n.d.). https://media.capella.edu/CourseMedia/VilaHealth/ANLT5020/LimitationsOnETLModeling/transcript.asp

Atlan, T. (2023, November 24). Data Pipeline Monitoring: Steps, Metrics, Tools & More! *Atlan*. https://Atlan.com/data-pipeline-monitoring/

Databricks. (2023, May 10). *What is the Databricks Unified Data Analytics Platform?* https://www.databricks.com/glossary/unified-data-analytics-platform

How to use OCR software for PDFs in 4 easy steps | Adobe Acrobat. (n.d.). https://www.adobe.com/acrobat/how-to/ocr-software-convert-pdf-to-text.html

Lee, S., Xu, Y., D'Souza, A. G., Martin, E. A., Doktorchik, C., Zhang, Z., & Quan, H. (2020). Unlocking the potential of electronic health records for health research. *International Journal of Population Data Science*, *5*(1). https://doi.org/10.23889/ijpds.v5i1.1123

Pykes, K. (2023, December 21). *What is Microsoft Fabric?* https://www.datacamp.com/blog/what-is-microsoft-fabric

Ramakolote, J. M., John Andrew Van, d. P., & Dongmo, C. (2023). Towards a Conceptual Framework for Data Management in Business Intelligence. *Information, 14*(10), 547. https://doi.org/10.3390/info14100547

Vila Health Analytics Internship: Identifying Data Sources. (n.d.). https://media.capella.edu/CourseMedia/VilaHealth/ANLT5020/IdentifyingDataSources/wrapper.asp?sso=true#scene6

Zendata. (2024, May 2). *Establishing a Data Quality Framework: A Comprehensive guide*. https://www.zendata.dev/post/data-quality-framework-a-comprehensive-guide