Heartland Escapes: Business Data Analysis

Big Data Analytics – CS686

Aidan Polivka

June 16, 2024

Table of Contents

Organizational Background

Throughout my time at Colorado Technical University, I've developed a fictional bookstore called Heartland Escapes. Previous projects include a migration plan from an on-premises system to an Azure Cloud environment, a system networking plan for cloud computing, a system security management and maintenance plan, and a database schema development plan. I think it's only natural to continue working with this fictional store here to analyze the business uses of their data and identify new utilization opportunities of both structured and unstructured data relevant to Heartland Escapes.

As stated previously, Heartland Escapes is a bookstore in Lincoln Nebraska with two locations. They have seen recent growth and consumer interest due to their expanding presence in social media. Their business model is heavily centered around hosting events at their stores including author meet and greets, summer reading programs, holiday-oriented events (like scary story readings on Halloween or Santa reads Christmas stories), and many others. Because of their marketing of these events, they've received a lot of local publicity and foot traffic in store. With the growth they've seen over the past year, Heartland Escapes would like to expand to two neighboring cities in Nebraska with a new location in Omaha Nebraska, and a new location in Beatrice Nebraska. (Polivka, 2024)

Organization Inception

Heartland Escapes started out as a form of community necessity. The public libraries were well stocked with books but lacked in providing young people with the spark to enjoy reading. The owners of the company are avid book readers themselves and have always enjoyed sharing their literary journeys with other like-minded individuals. Heartland Escapes mission is to be a shared “Escape” for others to enjoy captivating stories and their love for books with each other. The first store opened in 2010, and it was a slow start to get people in the door with competing companies like Barnes and Noble around. As people began seeing the benefits of community-oriented storytelling, traction accrued to the point that Heartland Escapes had the financial stability to open a second store in 2017. The first year of the pandemic was difficult for the company, but after the guidelines loosened and the public started feeling more comfortable leaving their homes, business caught its stride once again. Since then, Heartland Escapes has seen nothing but growth and enthusiastic customers. (Polivka, 2024)

Data Structures and Management

When Heartland Escapes started their journey, they developed their own point-of-sale system and inventory management API. Married to these systems are the Accounting Database and the Inventory Database. Additionally, they have a public facing website where users can search store inventory with and see store hours and event schedules. (Polivka, 2024)

The data stores that Heartland Escapes currently has available are Microsoft SQL Databases, excel spreadsheets, generated reports from their Inventory API and Point-of-Sale system, social media platforms, google reviews, and potentially their third-party payment processor.

Identification of Big Data

Data Utilization

The structured data within the Heartland Escapes system is primarily Microsoft SQL Server managed databases with very limited breadth. These databases are mostly used for inventory and purchase/sale management, although a large amount of Heartland Escapes accounting is done manually using excel spreadsheets. They also run reports through their point-of-sale system and the inventory API to generate needed information about monthly revenue, inventory restocking needs and other business operation required reports. Heartland Escapes also uses a third-party payment processor that may have API endpoints to retrieve data about purchases, but that hasn't been fully explored by stakeholders.

This structured data could be utilized to create more in-depth reports about book genre sales over time. Heartland Escapes could also use some of the structured data from their social media profiles for what book genres and authors are sold after a viral post based on demographic engagement analysis.

Unstructured Data

Heartland Escapes has a lot of unutilized unstructured data. Their social media standing mainly comes from spurts of local virality from their talented staff rather than true

advertising. There is a lot of untapped potential from the available demographic data of followers, high account traffic times, analysis of viral posts, and comment analysis.

They could use the demographic data for targeted advertising and time of day interaction rates for optimal posting times. Additionally, the viral posts can be analyzed for what they have in common that resulted in a viral response, leading to better engagement from followers and therefore more foot traffic through the doors of Heartland Escapes. Being a company with social media influence, they also have a large amount of data from google reviews that could be utilized to determine what they're doing well and what needs improvement from the perspective of customers.

Data Types and Business Analysis

Both the structured data and unstructured data are crucial for the business analyst's work. This data is a tool in the analyst's arsenal for investigating what aspects of the business can be improved to lead to more profit. The underutilized data from Heartland Escapes' social media accounts presents a fantastic opportunity to gain new insights and find ways to attract more customers to their stores.

Structured data is important for the analyst in that it offers concrete, easy to understand insight into business processes and metrics. This also provides quantitative information to bring back to the stakeholders to explain issues, solutions and improvements to their company. Unstructured data is equally important in that it provides more qualitative information about the company. Whether this is coming from google reviews, social media comments or some other source, this data is crucial to understand customer sentiment and make improvements based on consumer feedback. Business analysts use both structured and unstructured data to form a holistic view of the business.

Big Data Analytical Tools

Heartland Escapes needs a data warehouse capable of storing multiple types of data (structured, unstructured, and semi-structured). They also need an ETL tool to inject data from their social media sites into their warehouse, and they also need a business intelligence tool to analyze the data stored in the warehouse. Considering their organization's size, they need a tool that is easily maintained and updated and offers simple querying capabilities for analysis. Additionally, they'll need a suite of resources that are budget friendly, as their organization is not large enough to be spending exorbitant amounts of money on business analysis.

Industry Standard Big Data Tooling

Extract, Transform and Load (ETL) Tools

Most organizations that care about business intelligence require the use of data from many sources. In this case for Heartland Escapes, they care about data from their own databases and data from their Instagram and TikTok accounts. In order to access this data and store it in a centralized location, an Extract, Transform and Load Tool (or ETL tool for short) is required. Most of the time it's ideal to extract data from other sources and store it yourself to improve the performance of your business intelligence (BI) tools. In some cases, it might be necessary for your BI tool to perform some querying itself.

There are many ETL tools available today in the big data industry. Your choice of tooling should be influenced by utility requirements, employee experience, platform harmonization (Microsoft integrates well with other Microsoft products, for example), and cost. For example, it wouldn't make much sense to choose an AWS ETL tool if your data sources and warehouses are in Google Cloud SQL and Big Query. Also, if your desired tool only offers custom scripting in Go and your team is a .NET shop, that would be an abrupt change for most engineers. When considering platform harmony, it's important to not choose platform harmony at the cost of platform dependence.

There are a couple popular ETL options that would make sense for Heartland Escapes, considering these criteria. SQL Server Integration Studio (SSIS) would be a solid option considering they are a Microsoft shop. All their existing applications are written in C# .NET MVC and .NET Core Rest APIs. Also, their database are Microsoft SQL Server. SSIS offers custom scripting in C#, which would allow Heartland Escapes to perform requests to Instagram and TikTok APIs for data. Finally, SSIS can integrate with most data solutions, including Google Big Query and Cloud SQL. Since the rest of their system is hosted in Google Cloud Platform (GCP), integration with GCP is a strict requirement. SSIS's licensing is rolled into the enterprise license for SQL Server as well, however the infrastructure maintenance and cost may pose problems for Heartland Escapes. SSIS is not managed by Google Cloud platform, who hosts their other applications. This means that SSIS will need to run on a separate server, or within a Compute Engine Virtual Machine which is an expensive and high maintenance cloud component on GCP (Google, 2024b).

Another popular option would be ETLBox. This is a new ETL tool that's built on .NET Core, which would fit in incredibly well with Heartland Escape's existing software stack. ETLBox is built into an .NET Rest API, using network calls to trigger ETL actions. This includes Pub/Sub integration out of the box, and integration with the rest of Google Cloud Platform. Since it's built on a .NET Rest API, the Rest API can be containerized and deployed to Cloud Run, making it a fully managed tool with no infrastructure management needed. The biggest downside to ETLBox is that there is no UI pipeline development. It can only be maintained and updated by individuals with C#/Visual Basic skills (Lennartz, 2024).

The final ETL option for Heartland Escapes would be Google Cloud Dataflow. Dataflow meets many of Heartland Escapes requirements: It is a fully managed Google Cloud Platform tool. Therefore, it doesn't require any infrastructure maintenance, and it provides platform harmony by remaining within the GCP toolset. As a result of this harmony, Dataflow is easily integrated into other GCP components and tools. Dataflow also offers templates for commonly scripted flow types, making creating ETL flows quicker and simpler. One of the major bonuses to Dataflow is the integrated monitoring and observability that comes with using this tool on Google Cloud Platform. The downside is that Dataflow does not offer C# scripting, the only scripting available is in Python or Java. An additional downside is platform dependency (Google, 2024c).

Data Warehousing

After the ETL Tool "extracts" and "transforms" the data, it needs to be able to "load" it into some resource or collection of resources. This is the host for the organization's big

data and will be the data source for the organization's business intelligence tool. In choosing a Data Warehousing solution, it's important to have some of the same considerations as for the ETL Tool. It needs to be able to fulfill your company's requirements, platform harmony is a bonus, it needs to be highly performant and cost effective. For Heartland Escapes, there are really two natural options that match all these specifications.

The first solution would be a combination of Google Cloud Storage and Google Big Query. Big Query offers storage of both unstructured and semi-structured data types. Big Query's uses of columnar storage making it optimized for fast query times and efficient data processing. Google Cloud Storage is a great tool for unstructured data storage, and it offers different storage costs per document usage frequency. It also allows you to create automated retention policies and storage type resolutions. For example, if Heartland Escapes does not typically use documents three years or older, those can be moved to archive storage therefore reducing the storage cost. Both Cloud Storage and Big Query operate on a pay-as-you-go pricing model.

The second solution would be a combination of Snowflake and Google Cloud Storage. Snowflake has somewhat recently partnered with Google and can be fully integrated into GCP (Ichhpurani, 2019). Snowflake offers very similar benefits to Big Query, although it's slightly more performant and slightly more costly (also a pay-as-you-go pricing model) than Big Query. Snowflake is easily integrated into Google Cloud Storage as well. The major benefit that Snowflake offers over Big Query is that it's able to be used across

various platforms. This removes some platform dependency while also providing platform harmony.

Business Intelligence (BI)

The business intelligence tool is the presentation side of a big data technology stack. This is the tool that is used to query the compiled data and display it to business analysts. Heartland Escapes has a couple natural options for a business intelligence tools, considering the options that were selected for data warehousing.

The first natural choice would be Google's Looker Studio. Looker Studio free to use and integrates natively with Google Cloud Storage and Google Big Query. It also offers multi-cloud support, meaning that if Heartland Escapes chooses to migrate cloud platforms to AWS, they'd be able to do so without destroying the existing BI analytics dashboards. Looker offers third-party connectors, which means that it will be able to integrate with Snowflake as well. Supposedly, the user interface is intuitive, which makes development of dashboards easy for technical and non-technical individuals (Google, 2024a). As a result, these dashboards should be easily maintained for the Heartland Escapes stakeholders. There is no guarantee that third party connectors are free to use, so that may be a pitfall for Looker if Snowflake is chosen as the data warehousing tool (Snowflake, Inc, 2024).

The second choice for Heartland Escapes is Power BI. Power BI is a very common business intelligence tool that offers many similar benefits to Looker Studio. It also offers direct integration with Big Query, Cloud Storage, and Snowflake. Power BI offers a free

edition, however there are more collaboration and processing capabilities using the Power BI paid versions (Kenneth, 2022).

Selection for Heartland Escapes

Because the rest of the Heartland Escapes system is containerized and/or portable, it's important that the chosen big data technology stack doesn't pigeonhole them into a cloud platform dependency. Any configuration of the options outlined within this section are completely viable solutions for Heartland Escapes. The best organization of these tools that would offer Heartland Escapes a cost effective, easily maintainable solution with a familiar tech stack would be:

- ETL Tool: ETLBox
    - Platform independence and containerization will make this tool highly portable for Heartland Escapes
    - Familiar development language will make this tool easily maintained by the Heartland Escapes team
    - Containerized deployment with Cloud Run removes any need for infrastructure management, making the system easily maintainable
- Warehousing Tool: Snowflake & Google Cloud Storage
    - Snowflake is fully managed, so there is no need to manage infrastructure
    - Snowflake being multi-cloud offers more simple portability

    - Snowflake in combination with cloud storage provides coverage of all potential data formats
    - Although Snowflake is slightly more costly, removing the platform dependency is worth it.
- BI Tool: Google Looker Studio
    - Looker is free to use without any paywall behind collaboration
    - It's easy to use and the UI is intuitive enough for non-technical individuals to participate in dashboard development
    - Offers native support for Cloud Storage, and there are third-party connectors to easily integrate Snowflake

Big Data Analytics System Architectures

Data Architecture Selection Recap

The selected big data architecture consists of three primary components: An ETL tool, a warehousing platform, and a BI tool. The selected Extract, Transform, and Load tool, ETLBox, was chosen because it is highly compatible with the existing architecture, it's highly portable, platform independent, and suitable for the current development skills of Heartland Escapes engineers. ETLBox can also be containerized, this paired with Google Cloud Platform's Cloud Run component removes any need for infrastructure management.

The chosen warehousing tool is a combination of Snowflake and Google Cloud Storage. Between the two, there is storage coverage of all data types. Snowflake is also multi-cloud, reducing platform dependencies to GCP. Snowflake is a commonly used data warehousing solution, making it easier to develop upon and well documented. The final piece to Heartland Escapes data analytics architecture is the chosen business intelligence tool. Google Looker Studio was chosen because it's free, multi cloud, and has an intuitive dashboard building environment for non-engineers.

All of these sub-components of the larger data analytics system were chosen because they integrate well with the existing system, and they integrate well with one another. These sub-components are highly compatible with the current cloud platform, and together they form a cost effective, maintainable data analytics solution for the Heartland Escapes stakeholders.

Data Analytics Alignment with Existing Architecture

The data analytics architecture fits well with the existing architecture. Parts of the e-commerce system and the point-of-sale system are hosted using GCP's Cloud Run component. This is the same component that will host the ETL tool, which doesn't require any infrastructure management. Both the point-of-sale system and e-commerce system are written in C#, which is the same language used for ETLBox. ETLBox supports event-based messaging support, allowing for other platform tools to interact with the ETL tool like Google Cloud Scheduler and Google Pub/Sub.

Snowflake has native support in Google Cloud Platform, which ensures a basic level of alignment with the existing architecture. Snowflake also offers third-party connectors, extending portability and connectivity with software systems. ETLBox should have no problems connecting to Snowflake, and other parts of the system like Google Cloud Logging and Cloud Storage should be able to stream data to Snowflake without too much additional configuration. Google Cloud Storage is a part of Google Cloud Platform and is easily connected to in .NET. Google offers .NET libraries to write files to and read files from Google Cloud Storage.

Looker Studio also has native support in Google Cloud Platform. Although it doesn't natively support Snowflake, it also has third-party connectors that are easily configured.

Data Analytics Training Requirements

New System Analytics Support

Big Data Analytics Techniques

Big Data Analytics Value Creation

References

Polivka, A. D. (2024). Heartland Escapes: Database System Overview. Accessed September 8, 2024.

AWS. (n.d.). What is Structured Data? - Structured Data Explained - AWS. Amazon Web Services, Inc. https://aws.amazon.com/what-is/structured-data/. Accessed September 8, 2024.

AltexSoft Inc. (2023, November 1). Unstructured Data: Examples, Tools, Techniques, and Best Practices. Medium. https://altexsoft.medium.com/unstructured-data-examples-tools-techniques-and-best-practices-c0fefa57f741#:~:text=Unstructured%20data%20storage%201%20Scalability.%20Unstructured%20data%20has. Accessed September 8, 2024.

Terra, J. (2020, May 5). What is a Business Analysis and What does Business Analyst Do. Simplilearn.com. https://www.simplilearn.com/what-is-a-business-analyst-article. Accessed September 8, 2024.

Google. (2024a). Welcome to Looker Studio! - Looker Studio Help. Support.google.com. https://support.google.com/looker-studio/answer/6283323?hl=en. Accessed September 15, 2024.

Google. (2024b, September 10). SQL Server Integration Services (SSIS). Google Cloud. https://cloud.google.com/sql/docs/sqlserver/ssis. Accessed September 15, 2024.

Google. (2024c, September 11). Dataflow overview. Google Cloud. https://cloud.google.com/dataflow/docs/overview. Accessed September 15, 2024.

Herzberg, B. (2021, December 19). What Is Snowflake Data Warehouse? A Tutorial | Built In. Built In. https://builtin.com/articles/snowflake-data-warehouse. Accessed September 15, 2024.

Ichhpurani, K. (2019, June 4). Announcing Snowflake on Google Cloud Platform. Google Cloud Blog; Google Cloud. https://cloud.google.com/blog/products/data-analytics/announcing-snowflake-on-google-cloud-platform. Accessed September 15, 2024.

Kenneth, G. (2022, June 1). Top 6 Power BI Benefits For Your Business. Panoply.io; Panoply. https://blog.panoply.io/benefits-of-power-bi. Accessed September 15, 2024.

Lennartz, A. (2024). Quickstart. ETLBox. https://www.etlbox.net/docs/getting-started/quick-start/. Accessed September 15, 2024.

Rogojan, B. (2021, March 1). Snowflake vs BigQuery Pricing and Performance Comparison. Panoply.io; Panoply. https://blog.panoply.io/snowflake-vs-bigquery-comparing-pricing-performance-and-usability. Accessed September 15, 2024.

Snowflake, Inc. (2024). Use the Snowflake Connector for Google Looker Studio | Snowflake Documentation. Snowflake.com. https://other-docs.snowflake.com/en/connectors/google-looker-studio-connector. Accessed September 15, 2024.