Big Data Systems Design Report:
An Analysis for O-Ren's Vipers

Gary Leung, Maaz Saad, and Syed Qadri
Faculty of Business & IT, Ontario Tech University
MBAI 5110G – Big Data Systems Design
Professor Carolyn McGregor
March 20, 2023

Introduction

O-Ren's Vipers is an e-commerce organization based in Hamilton, Ontario. The company started operations in early 2020, just before the COVID-19 pandemic began. Due to the pandemic, the company had to make significant changes to its operational workflow within a few weeks of starting. These changes helped the company stay in business, and now, three years later, the company has established itself in a strong position. Currently, O-Ren's Vipers employs twenty-five people who work in a hybrid model.

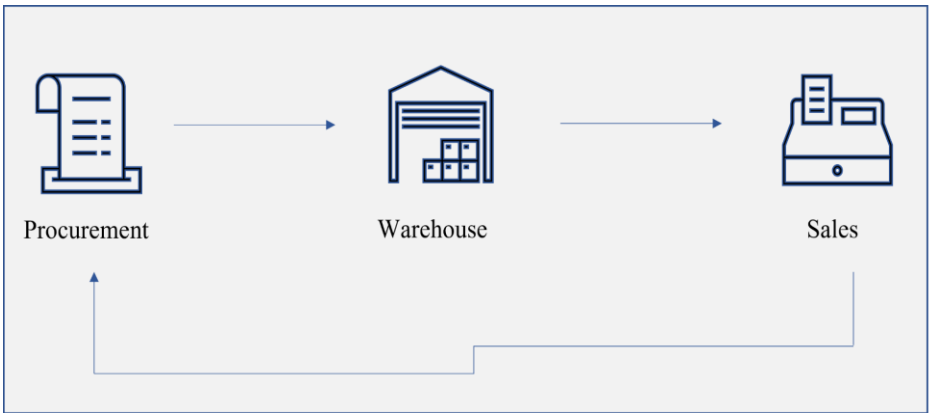
O-Ren's Vipers has noticed its competitors taking advantage of the new technological tools that have gained traction in recent years. Management has decided that, given their current stable financial situation, the time is ripe for them to revamp their operations and move away from their current data collection, storage, and analytics landscape. However, it is crucial for management to have a genuine understanding of the tools they wish to utilize, rather than simply selecting them because they have been endorsed by public figures, or because they are impressed by the interface, or, worst of all, because they believe new software can magically transform their business. Therefore, the organization must conduct a thorough analysis of the transformation it wishes to undergo, in order to better understand the steps involved in the process and the technologies required.

Our organization has been hired to prepare a big data systems design report for O-Ren's Vipers. The report will outline the organization's problems that need to be solved, the type of data it generates, and how the data can be cleaned. Once we have a better understanding of the cleaned data, we will present our recommendations on the structure of a data pipeline and how the data should be stored. After constructing the data pipeline and accumulating data on the servers, we will suggest some data analytics approaches to O-Ren’s Vipers to help them take full advantage of the accumulated data. Finally, we will examine whether it is feasible for O-Ren's Vipers to migrate from an on-premises storage module to the cloud and recommend a suitable cloud provider.

Case Study Problem Statement

O-Ren's Vipers have selling rights on Amazon Canada. However, due to the current organizational landscape of the company, they are unable to expand to other selling platforms such as Amazon US, Walmart, and Shopify. The organization documents its data in Excel files that need manual updating at the end of each day. These files are then shared among employees. Consequently, there is always a significant delay between the version of data an employee is working with and the most up-to-date version, leading to mistakes and workflow delays.

The organization consists of three main teams that must work together in complete sync for optimal output and overall efficiency. Unfortunately, this is not the case, and we have identified the reason for this as the manual nature of the operational workflow. The following diagram breaks down the teams, their responsibilities, and how they interact with each other.



The procurement team is responsible for placing orders with their vendors. These orders are then received and processed at O-Ren's Vipers' warehouse, which is located in the basement of the company's office building, by the warehouse team. The products are repackaged, listed on Amazon, and shipped, enabling the sales team to set prices and generate revenue. The sales team is also responsible for informing the procurement team when the quantity of a product is running low, and more units need to be procured. Likewise, the warehouse team checks with the sales team to determine when additional units of a product need to be sent to Amazon's fulfillment centers.

In summary, the need for synchronization between the three teams is apparent, and the lack of up-to-date data causes significant bottlenecks, resulting in lower revenue generation.

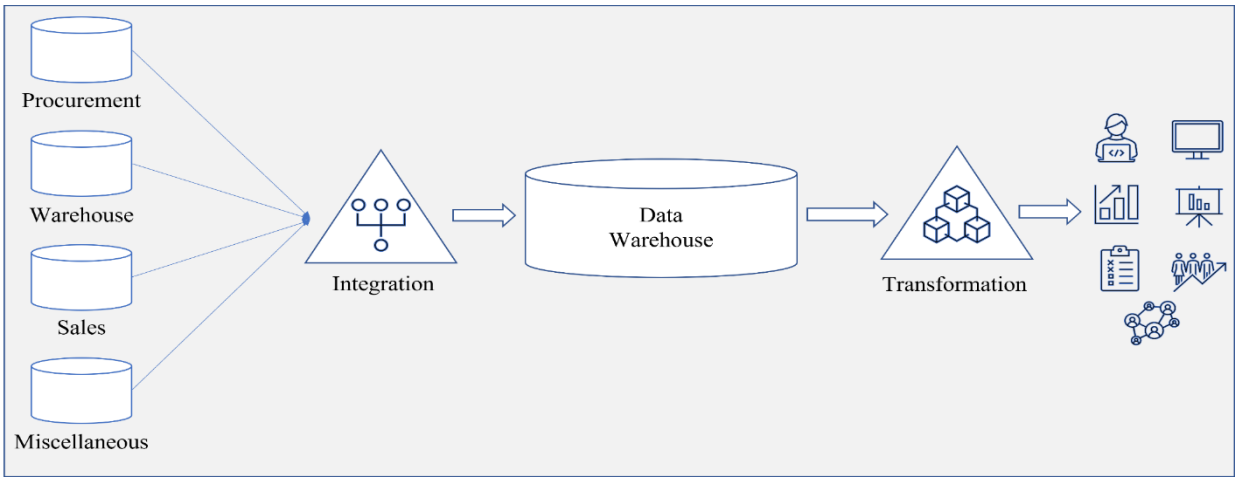
The organization has set aside funds and would like to embark on a project to migrate from its manual workflow to an automated one. They believe that it is the right time to invest in a database management system, which will help improve the organization's privacy and security, team integration, and overall efficiency and performance (Pipeline).

Database management systems enable users to share data quickly and allow employees to access data in real-time, leading to more accurate results (Pipeline). This helps the sales team increase sales and generate revenue more quickly (Pipeline). By implementing a database management system, the connections between different teams within the organization become more prominent and easier to illustrate (Pipeline). Additionally, data inconsistency is a significant issue for an organization, as different versions of the same data may exist in different places within the company. By investing in a database management system and data quality tools, O-Ren's Vipers can ensure an accurate and up-to-date view of the data for all team members at all times (Pipeline).

Furthermore, a database management system provides a framework for the successful implementation of privacy policies, helps maintain confidentiality, and keeps information safe from hackers (Pipeline). These measures lead to increased productivity, as a well-structured database management system allows employees to spend their time working on revenue-generating activities and initiatives rather than mundane and troubleshooting tasks (Pipeline).

Before moving on to the next section, we would like to present a simple diagram that showcases what the organization hopes to achieve from this project. This diagram represents the

final operational architecture of O-Ren's Vipers, and the way we achieve it will be presented in the next few sections of this report.



Data Cleaning

Data cleaning is the process of identifying errors in data, which can manifest as inconsistencies, spelling errors, missing information, duplicate entries, or other types of invalid data (Rahm et al., 3). In the case of O-Ren's Vipers, data will be collected from multiple sources, including procurement, warehousing, and sales, which greatly increases the need for data cleaning (Rahm et al., 3). Combining data from various sources increases the likelihood of duplicate data points, as the same information may be recorded twice: once in the procurement module and then again in sales (Rahm et al., 3). Therefore, for accurate data analysis to occur, the data must be processed in a way that ensures its accuracy and consistency (Rahm et al., 3).

Data cleaning is a crucial process that transforms raw data into a usable format, and it occurs in a separate staging area before the data is loaded into the data warehouse for analysis (Rahm et al., 3). Although there are several tools available to simplify and streamline this process, the fact remains that data cleaning is often a manual process due to the variability in the type of cleaning that a dataset requires (Rahm et al., 3). Therefore, O-Ren's Vipers would need to hire entry-level programmers to prepare the data before experienced analysts can analyze it to generate insights.

When cleaning data, it is critical to perform the cleaning process in conjunction with schema-related data transformations based on comprehensive metadata (Rahm et al., 4). According to IBM, a database schema can be explained as the "blueprint" of a database that showcases the links in data across different tables and models. The mapping functions used for data cleaning and other data transformation tasks need to be documented in a way that breaks down the purpose of the function and not just the process (Rahm et al., 4). The reason mapping functions are so vital is that they can be reused for other data sources and query processing, which boosts the consistency and accuracy of data processing and improves the overall efficiency and effectiveness of the organization (Rahm et al., 4).

Now that we have examined the reasons for the paramount importance of data cleaning, let's review a five-step process for data cleaning, as outlined by Ridzuan et al. The five steps are:

data analysis, defining the transformation workflow and mapping rules, data verification, data transformation, and the backflow of cleaned data into the data pipeline.

- **Data Analysis**: In this step, errors and inconsistencies in the data are identified, and all outliers and anomalies are pinpointed. There are two approaches to data analysis: data profiling and data mining. Data profiling involves analyzing individual variables, while data mining involves analyzing overall patterns in the data.
- **Transformation Workflow**: The anomalies in the data are removed through a series of operations that are pre-defined during the mapping process. The challenging part is creating the mapping rules and workflow design because the amount of cleaning required varies from one data set to another.
- **Verification**: In this stage, the transformed data is tested and evaluated for correctness and accuracy. It is possible that the data goes through several iterations of cleaning and analysis before a final version can be passed onto the next stage.
- **Transformation**: This process involves updating the structure of the data to make it compatible with the requirements of the data warehouse. This step requires a considerable amount of metadata to ensure that the transformation process is performed successfully, and the resulting data is of high quality.
- **Backflow of Cleaned Data**: Once the finalized version of the data is prepared, the raw or dirty data is replaced by the transformed data, which is ready for further data analysis.

In the end, it is up to O-Ren's Vipers to decide on the data cleaning approach and tools that work best for them, as every organization must adapt these tools to their specific situation. Furthermore, here is an example of what the raw data generated by O-Ren's Vipers looks like and what it would look like after undergoing a few iterations of data cleaning.

- **Raw Data**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Notificatic	Payload'	UniqueId	Publish	SellerId	MarketplaceId	KittMarket	NP_AOCN_OC	NP_AOCN_(	NP_AOCN_	NP_AOCN_OCT_Of	NP_AOCN_S_Lis	NP_AOCN_S_List	NP_AOCN_S_Sugg	NP_AOCN_S
2	AnyOfferC	1	2871078c-a8a3-459	30:23.9	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B08SMMP43	30:14.2	FeaturedOffer	49.99	USD	toy_display_on_w	43.17
3	AnyOfferC	1	74e1893e-dfaa-47c	43:14.5	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B07WTQ3DI	42:57.5	Internal	NULL	NULL	166225011	43.17
4	AnyOfferC	1	daa4179e-6462-4ab	33:03.8	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B081FWTX7	32:56.6	Internal	39.99	USD	toy_display_on_w	17.75
5	AnyOfferC	1	0c194fb6-699a-467	35:05.0	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B003LZVTFG	35:03.9	FeaturedOffer	NULL	NULL	2514571011	16.31
6	AnyOfferC	1	37f675e8-5684-433	31:19.3	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B01IA9F6B0	31:15.0	Internal	NULL	NULL	toy_display_on_w	Merchant
7	AnyOfferC	1	57ccf9fd-9a1f-4a9e	40:38.4	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B0984Z3TCC	40:23.9	Internal	18.99	USD	2514571011	Amazon
8	AnyOfferC	1	26215253-4870-4c8	33:16.6	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B078GTYC7K	33:04.0	Internal	NULL	NULL	art_and_craft_sup	Merchant
9	AnyOfferC	1	d010b4f4-7f87-4d5	36:05.0	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B08XPRTDQ	35:53.3	FeaturedOffer	NULL	NULL	262625011	Amazon
10	AnyOfferC	1	c4330b26-d825-4e9	45:10.5	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B0002V9INN	44:43.7	Internal	74.9	USD	toy_display_on_w	Merchant
11	AnyOfferC	1	fe9a7b33-1368-43c	35:36.1	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B002JGP8PU	35:27.6	FeaturedOffer	25.9	USD	art_and_craft_sup	Merchant
12	AnyOfferC	1	4687ec61-2490-467	33:34.2	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B00AZX0GO	33:25.5	FeaturedOffer	37	USD	2514571011	3.6
13	AnyOfferC	1	39a8dbaf-7689-46a	39:19.1	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B07BKMKR6	39:15.6	FeaturedOffer	59.95	USD	toy_display_on_w	-7
14	AnyOfferC	1	7eeecc98-a5f8-405	35:01.4	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B07117ZDTF	34:55.4	FeaturedOffer	49.9	USD	beauty_display_o	-17.6
15	AnyOfferC	1	d1791183-e291-4dc	34:02.0	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B00DU3XSG	33:53.4	FeaturedOffer	37.5	USD	11057251	-28.2
16	AnyOfferC	1	639c1e66-8371-4de	35:56.1	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B01IADZWZ	35:44.6	FeaturedOffer	NULL	NULL	166363011	Merchant
17	AnyOfferC	1	31405030-655b-4e5	46:52.5	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B07RLL8LLN	46:24.6	Internal	24.95	USD	kitchen_display_c	Amazon
18	AnyOfferC	1	d7841152-27f9-497	36:41.3	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	1947494503	36:36.6	FeaturedOffer	50	USD	2474049011	Merchant
19	AnyOfferC	1	e6f5d7fb-3bd1-40d	41:58.9	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B00JYEBTRC	41:26.8	FeaturedOffer	34.99	USD	761520	Amazon
20	AnyOfferC	1	2d4c27db-b910-4b	41:54.4	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B07Z7J4V7F	41:26.6	FeaturedOffer	NULL	NULL	toy_display_on_w	Merchant
21	AnyOfferC	1	683bbcc9-039e-47c	41:58.7	A2TZPY9I	ATVPDKIKX0DER	1	ATVPDKIKX0D	B07H8R2MT	41:37.7	FeaturedOffer	40	USD	2491829011	Merchant

- **Cleaned Data**

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ASIN	MSP	Selling Price	Referral Fee	Margin	Margin %	Sale Quantity	Discounted Price - New	New Margin	Discount*	New Margin %	New Margin - Old Margin	Total Loss
2	B000F8R46A	$ 17.19	$ 16.17	15%	$ (0.87)	-5.04%	1.00	$ 16.17	$ (0.87)	FALSE	-5.04%	$ -	$ -
3	B000FCEQ08	$ 13.16	$ 23.71	15%	$ 8.97	68.14%	10.00	$ 22.91	$ 8.29	TRUE	62.96%	$ (0.68)	$ (6.82)
4	B000GCNCXO	$ 19.25	$ 21.63	15%	$ 2.02	10.51%	10.00	$ 21.10	$ 1.57	TRUE	8.18%	$ (0.45)	$ (4.48)
5	B000HM6EFG	$ 41.49	$ 36.00	15%	$ (4.67)	-11.25%	4.00	$ 36.00	$ (4.67)	FALSE	-11.25%	$ -	$ -
6	B000IGOXLS	$ 78.72	$ 78.72	15%	$ -	0.00%	6.00	$ 78.72	$ -	FALSE	0.00%	$ -	$ -
7	B000IZSHQ6	$ 22.04	$ 23.65	15%	$ 1.37	6.21%	4.00	$ 23.53	$ 1.27	TRUE	5.76%	$ (0.10)	$ (0.40)
8	B000J1HDWI	$ 8.31	$ 8.31	15%	$ -	0.00%	8.00	$ 8.31	$ -	FALSE	0.00%	$ -	$ -
9	B000J5AK9M	$ 27.22	$ 31.90	15%	$ 3.98	14.61%	6.00	$ 31.43	$ 3.58	TRUE	13.14%	$ (0.40)	$ (2.40)
10	B00006IFJD	$ 5.71	$ 5.00	15%	$ (0.60)	-10.57%	1.00	$ 5.00	$ (0.60)	FALSE	-10.57%	$ -	$ -
11	B0000AZ6DA	$ 6.24	$ 10.93	15%	$ 3.99	63.89%	6.00	$ 10.66	$ 3.76	TRUE	60.25%	$ (0.23)	$ (1.36)
12	B00018A3US	$ 29.76	$ 34.00	15%	$ 3.60	12.11%	5.00	$ 33.83	$ 3.46	TRUE	11.63%	$ (0.14)	$ (0.72)
13	B000296LQ6	$ 26.05	$ 23.99	15%	$ (1.75)	-6.72%	4.00	$ 23.99	$ (1.75)	FALSE	-6.72%	$ -	$ -
14	B0002APIKK	$ 16.08	$ 24.12	15%	$ 6.83	42.50%	8.00	$ 23.65	$ 6.43	TRUE	40.00%	$ (0.40)	$ (3.22)
15	B0002AQS72	$ 18.01	$ 27.02	15%	$ 7.66	42.52%	4.00	$ 26.75	$ 7.43	TRUE	41.26%	$ (0.23)	$ (0.91)
16	B0002BSP4K	$ 23.82	$ 23.82	15%	$ -	0.00%	12.00	$ 23.82	$ -	FALSE	0.00%	$ -	$ -
17	B0002I5QK4	$ 13.27	$ 14.05	15%	$ 0.66	5.00%	11.00	$ 14.05	$ 0.66	FALSE	5.00%	$ -	$ -

Data Pipeline Integration

The integration of a data pipeline architecture is essential to realize the potential benefits it provides in storing data. O-Ren's Vipers can automate their workflow, integrate their teams and data, generate revenue more frequently, and at a more rapid pace by proposing this architecture. The lack of synchronization between the three teams and lower revenue generation is mainly due to the manual nature of their operational workflow, wherein data is documented in Excel files that need manual updating at the end of each day. Therefore, big data integration in data pipelines is crucial to this project.

Implementing a database management system makes connections between different teams more prominent and easier to illustrate, in the case of O-Ren’s Vipers, where there are teams who are not able to keep in constant communication, this aspect of pipelining is extremely attractive. This also allows employees to share data quickly and access data in real-time, resulting in more accurate results. The sales team can increase sales and generate revenue more quickly with a well-structured database management system. Moreover, it enables employees to spend their time working on revenue-generating activities and initiatives, leading to increased productivity. Data pipelines have become an essential component of modern business operations, as they provide a structured way to move, transform, and analyze data, enabling organizations to optimize their workflow and automate time-consuming and error-prone tasks.

Here are the steps O-Ren’s Vipers can take to integrate data pipelines into an organization and the steps they can follow to ensure complete and seamless transition from singular online platforms to multiple distribution channels:

- **Identify your data sources and define your data requirements**: The first step in integrating data pipelines into your organization is to identify your data sources and define your data requirements. This involves identifying the types of data that you need to collect and the sources from which you will collect them. Once you have identified your data sources, you should define your data requirements, including the types of data you need, the format of the data, and any other requirements that are specific to your business.

- **Choose a data pipeline tool**: The next step is to choose a data pipeline tool that will allow you to extract, transform, and load your data into a target database or data warehouse. There are many data pipeline tools available, including open-source tools like Apache Kafka and commercial tools like Google Cloud Dataflow and AWS Glue. You should choose a tool that is compatible with your data sources and data requirements and that is easy to use and maintain.
- **Build your data pipeline**: Once you have chosen your data pipeline tool, you can begin building your data pipeline. This involves configuring your data pipeline tool to extract data from your sources, transform the data as necessary, and load it into your target database or data warehouse. You should test your data pipeline thoroughly to ensure that it is working as expected and that it is meeting your data requirements.
- **Implement data governance policies**: Data governance policies are essential for ensuring that your data is accurate, up-to-date, and secure. You should implement data governance policies that define who have access to your data, how your data is managed and maintained, and how your data is used. You should also define data quality metrics and monitoring procedures to ensure that your data is accurate and up to date.
- **Monitor and maintain your data pipeline**: Once your data pipeline is up and running, you should monitor it regularly to ensure that it is working correctly and that it is meeting your data requirements. You should also perform regular maintenance tasks, such as updating your data pipeline tool, to ensure that it continues to function properly.

There are however positives and negatives in implementing data pipelines, and being aware of these aspects can help make an informed decision as to whether to implement this data management solution. The pros and cons of the pipeline are as follows:

Pros:

- **Improved data quality**: Data pipelines can help ensure that data is accurate, complete, and up to date by automating data extraction, transformation, and loading processes.
- **Increased efficiency**: Data pipelines can help automate data processing tasks, reducing the amount of time and effort required to manage and maintain data.
- **Faster access to insights**: Data pipelines can help make data available more quickly, allowing businesses to make more informed decisions in real-time.
- **Greater flexibility**: Data pipelines can help businesses integrate data from a variety of sources, allowing them to gain a more comprehensive understanding of their operations.
- **Scalability**: Data pipelines can be designed to handle large volumes of data, making them a good solution for businesses that need to process high volumes of data.

Cons:

- **Initial setup can be complex**: Setting up a data pipeline can be complex, requiring expertise in data engineering and data management.
- **Cost**: Implementing and maintaining a data pipeline can be expensive, especially if it involves purchasing or licensing data pipeline tools.
- **Data governance challenges**: Ensuring data quality and compliance with data governance policies can be challenging when integrating data pipelines, especially if data is being integrated from multiple sources.

- **Security risks**: Integrating data pipelines can increase the risk of data breaches or data loss, especially if data is being moved between different systems or stored in different locations.
- **Dependency on technology**: Implementing a data pipeline can create dependencies on specific technologies or tools, which can make it difficult to switch to alternative solutions in the future.

Integrating data pipelines into an organization requires careful planning and execution. You need to identify your data sources and requirements, choose a data pipeline tool, build your data pipeline, implement data governance policies, and monitor and maintain your data pipeline. By following these steps, you can ensure that your data is accessible, accurate, and up to date, which can help to drive insights and decision-making in your business.

Data Storage

The connection between creating and monitoring a data pipeline, revolves around the secure methodologies used by this company with regards to data storage. Data storage and the integration of data pipelines go together and are both critical aspects of the other. Data storage refers to the process of storing digital information or data in a centralized location, which can be accessed, managed, and retrieved as needed. In today's digital age, data storage is a crucial aspect of creating an efficient workplace, as businesses and organizations generate vast amounts of data daily. Effective data storage can help businesses manage their data in a more organized, secure, and cost-effective way, leading to improved productivity and profitability.

Data storage covers several prominent reasons as to why a company may be wanting a change, which includes efficient data management, scalability, data integrity and compliance. Although these may seem like the ones mentioned before, these aspects are now with respect to how this data is stored and handled. Data storage provides a central location where all data can be stored and accessed by different users and applications within an organization. This ensures that data is managed efficiently, and everyone has access to the same version of the data, which reduces the risk of errors and improves decision-making. When data is stored in a centralized location, it becomes easier for decision-makers to access and analyze the data in a timely and efficient manner. This allows them to make better-informed decisions based on accurate and up-to-date data.

Moreover, by leveraging data storage systems with advanced analytics and reporting capabilities, decision-makers can gain insights into trends, patterns, and correlations in the data. These insights can help them identify opportunities, risks, and areas for improvement, which can inform their decision-making process.

Data storage systems provide secure storage for sensitive data, and this is crucial for protecting company data from unauthorized access, data breaches, and cyber-attacks. By implementing robust data storage systems with proper security measures, companies can mitigate the risks of data loss and ensure that their data is always safe and protected. Data integrity is critical for ensuring that the data being processed is accurate, complete, and consistent. An example is the use of access controls, such as user authentication and authorization, to restrict access to data based on an individual's role, level of clearance, or need-to-know basis. This helps

to prevent unauthorized users from accessing sensitive data and ensures that only authorized personnel can view or modify it.

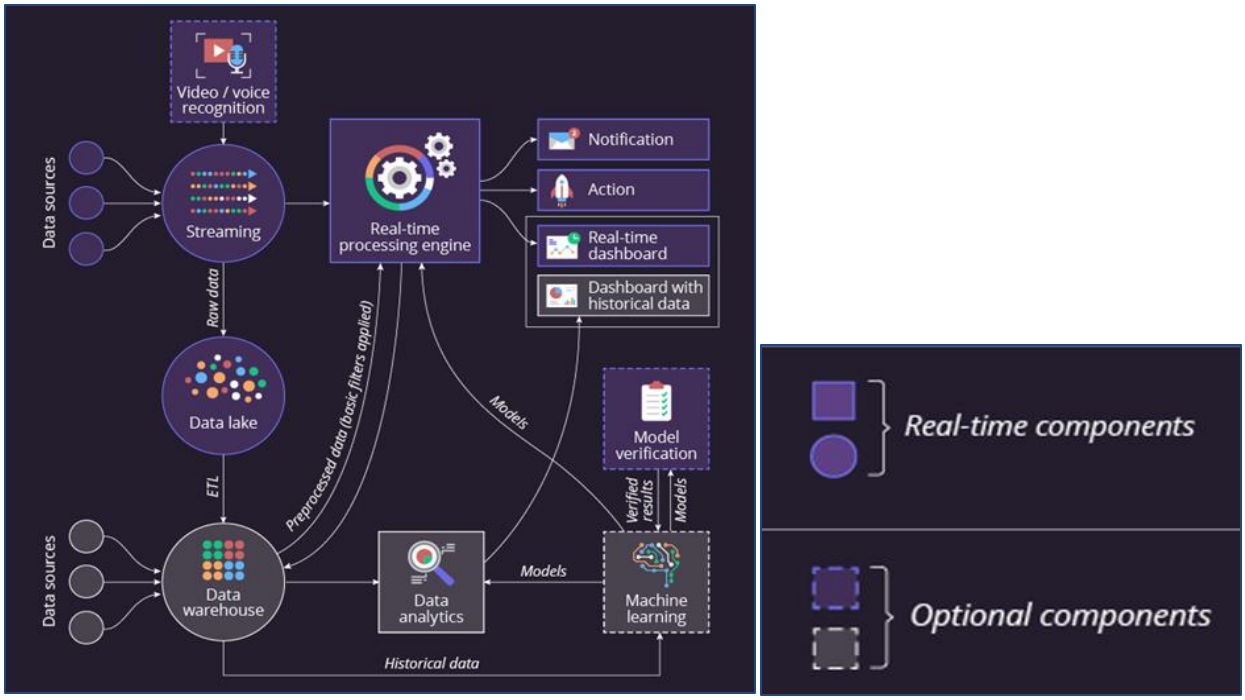
Additionally, data backup and disaster recovery solutions are also important security measures to protect against data loss. These solutions involve creating and storing copies of data in multiple locations and implementing processes to restore data in the event of a disaster or system failure.

Data storage systems provide tools for data quality management, which enables companies to maintain data integrity by detecting and correcting data errors, inconsistencies, and duplicates. Data storage systems play a crucial role in ensuring compliance with data privacy laws and regulations, such as GDPR and CCPA. Tools like data enrichment, monitoring, validation and cleansing, companies can ensure that their data is of high quality and less prone to error. By implementing a compliant data storage system, companies can ensure that they are adhering to these regulations and avoid costly fines and penalties.

Real-time Analytics and Retrospective Data Analysis

After performing data cleansing, completing the data integration across different systems, and identifying the data storage solution, the next part is the data analytic component. For O-Ren's Vipers, it is crucial to use the real-time data for improving customer experience during their visit. For example, the information of real-time available stock inventory. If you are a customer and would like to place an order in a shop, you will know the expected delivery date on the Amazon website. After placing the purchase order, what is the customer's feeling if you receive an email saying your order was delayed due to a lack of inventory? They must be disappointed and unsatisfied this shopping experience happened in the shop, so you may receive bad comments or even complaints from them. Besides, they may request a refund or return the product. This also creates extra cost to the business.

After acknowledging the importance of real-time data analytics, we would like to introduce what big data analytics we can use, especially for real-time analytics. There are four types of data analytics in the big data pipeline which are descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. Below is the big data analytics architecture combining both real-time data and historical data.



https://www.scnsoft.com/blog/real-time-big-data-analytics-comprehensive-guide[1]

Descriptive Analytics

Descriptive analysis is the way to discover customer behavior. For example, how many customers have visited our shop, which product is the most popular and what is the average order value per customer etc. This is the simplest form among those analytics but the easiest to illustrate what's happening to the business. One of the popular solutions is using Google analytics which can trace some basic web analytics in real-time such as number of visits, number of unique customers, the source of traffic or web browsing journey. This provides a free data source to monitor the real-time traffic in high level summary. We can also build another form of dashboard with historical data using data visualization tools for customized dashboard such as Tableau, Power BI. This dashboard can consist of different charts or tables with interactive features such as filter, grouping to facilitate different levels of user. Management can monitor high level KPI while data analysts can drill down the details for actionable insights by using the same dashboard.

Diagnostic Analytics

Same as descriptive analytics which is using historical data, diagnostic analysis aims to understand and explain the customer behavior. Normally, we will combine with descriptive analytics to use. For example, you may see the trend for total sales volume has dropped significantly in these few weeks (descriptive analysis) and would like to know why this happens. Then you will perform an investigation and drill down to the root cause such as specific location sales or specific customer segment. This is the diagnostic analytics to find the hidden pattern. Nowadays, many data visualization tools already support diagnostic analytics to drill down the

details for root-cause investigation. Some powerful tools like Tableau, have already offered a story telling module which allows users to create a story from descriptive analysis trend discovery to diagnostic analysis finding illustration. This allows O-Ren's Vipers management to better understand the entire analytical flow for why the sales are dropped and give the direction for the sales team to action.

Predictive Analytics

O-Ren's Vipers as an e-commerce company, one of the most important analytics should be understanding what their customers need. This is what predictive analytics can provide. Based on the historical data and understanding from customer behavior to predict the customer's needs. For example, what is the sales pattern for different customer groups and does any product have a seasonality effect. This can determine the upcoming sales strategy, provide better time management for the procurement process and allow better utilization of the warehouse on inventory management. The insights do not only support business revenue generation, but also perform cost management to avoid extra cost for delivery and less warehouse storage cost. Besides, predictive analytics allow us to predict the preference for each individual customer, so that we can customize the offer and send notification message to stimulate the sales.

Prescriptive Analytics

This is the advanced analytics area that is not yet widespread, but useful for recommendation to optimize the decision making. You may see many ecommerce websites with a section "Product you may interest", this is based on previous user purchase history and recent browsing record, using machine learning algorithms to provide the best product recommendation option. For the recommendation, they may provide a special pricing, bundle offer or even free shipment in order to optimize the profit and conversion rate for customers. Besides, O-Ren's Vipers can do some A/B testing strategy on new sales tactics which provide the data for the recommendation engine validation and improvement.

Last but not the least, it is important that all the analytics applied should align with the purpose stated in the customer consent statement. Otherwise, it may have potential ethical and privacy risk issues. This unethical behavior may lead to public backlash, negative media attention and the damage of company reputation is unrecoverable. Also, it will violate the privacy law and be sued by governance parties.

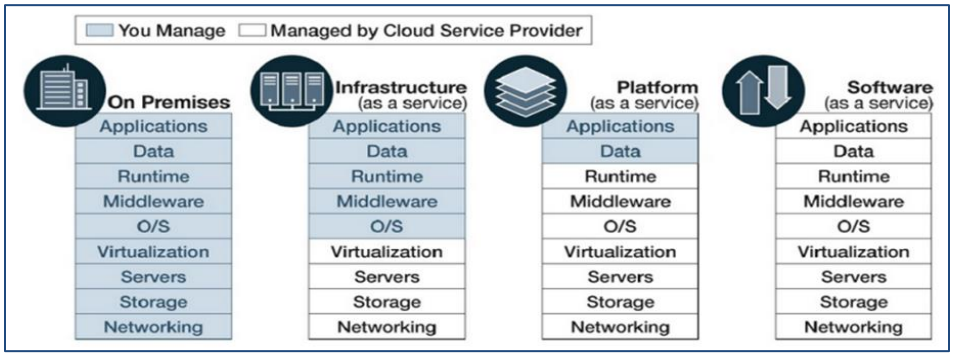
Cloud Provisioning or Local

The last session is on the storage infrastructure configuration. O-Ren's Vipers is operating an e-commerce business and currently using on-premises storage modules for their infrastructure. We would recommend the company migrate their infrastructure to cloud computing in order to catch up the big data requirement in our plan.

On-premises storage modules required O-Ren's Vipers to manage all the storage infrastructures including physical server setup, system configuration, networking connection and

runtime monitoring. It involved different aspects of skilled IT experts and resources to configure and maintain. From a cost perspective, the setup and operating cost to support on-premises storage modules is much larger than cloud computing. Besides, the traffic to e-commerce websites is quite seasonal. For example, there will be huge traffic to visit O-Ren's Vipers website during black Friday promotion. It may require many redundant server resources in order to capture those sudden growth traffic if we are using on-premises storage modules.

Compared with on-premises storage modules, cloud computing is the most popular solution for big data storage infrastructure. They offer various service models such as Infrastructure as a service, Platform as a service and Software as a service. Below is the function provided for each cloud computing service model compared with on-premises storage modules.



Siebel, T. M., & Rice, C. (2019). *Digital Transformation: Survive and thrive in an era of mass extinction*[2]

The more functions that the service provides, the less inhouse IT resources are required. SaaS offers all the features including software application which is the completed service suite and requires the least IT resource to maintain. Besides, we can use hybrid cloud or multi cloud approach to formulate the best cloud computing service model for O-Ren's Vipers. For example, we may store customer information in on-premises storage modules for security control, using SaaS for data analytics to leverage their well-build machine learning model. This approach can adopt the benefit for each cloud computing service model. Besides, there are many benefits for using public cloud compared with on-premises such as low latency, up-to-date cyber security technology and reliable disaster recovery resource. These required heavy IT resources to invest if we use on-premises modules.

Based on the above benefit, we would recommend O-Ren's Vipers to use hybrid cloud and multi cloud approach for their infrastructure configuration. They retain on-premises storage modules on confidential data storage such as customer personal information or payment details for security control. For cloud storage service providers, Amazon website service, Google cloud or Microsoft Azure are the most popular cloud platforms, we can select the one which is most suitable for O-Ren's Vipers based on their subscription pricing model and technical performance. We may also consider IBM Watson or Databricks Lakehouse platform which are designed for big data analytics to strengthen O-Ren's Vipers analytical abilities.

References

Benefits of database management systems (Dbms) | zoominfo. (n.d.). Pipeline Blog. Retrieved March 19, 2023, from https://pipeline.zoominfo.com/operations/6-benefits-of-using-database-management-systems-dbms

Ridzuan, F., & Wan Zainon, W. M. N. (2019). A review on data cleansing methods for big data. Procedia Computer Science, 161, 731–738. https://doi.org/10.1016/j.procs.2019.11.177

What is a database schema? | IBM. (n.d.). Retrieved March 19, 2023, from https://www.ibm.com/topics/database-schema

Rahm, Erhard & Do, Hong. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull.. 23. 3-13.

Bekker, A. (2023, February 10). A comprehensive guide to real-time Big Data Analytics. ScienceSoft footer icon. Retrieved March 19, 2023, from https://www.scnsoft.com/blog/real-time-big-data-analytics-comprehensive-guide

Siebel, T. M., & Rice, C. (2019). Digital Transformation: Survive and thrive in an era of mass extinction. RosettaBooks.

Agarwal, A., & Mittal, P. (2017). Big Data Pipeline Architecture. In Big Data Analytics (pp. 59-88). Springer, Singapore. https://doi.org/10.1007/978-981-10-4438-0_4

IBM. (n.d.). The benefits of good data storage. IBM. https://www.ibm.com/topics/data-storage

Jaiswal, A., Agarwal, A., & Mittal, P. (2018). Data Pipeline Design and Management for Big Data Analytics. In Big Data Analytics (pp. 187-223). Springer, Singapore. https://doi.org/10.1007/978-981-10-7849-8_7

Kshetri, N., & Voas, J. (2019). Blockchain-enabled data storage and management: A review. International Journal of Information Management, 46, 36-44. https://doi.org/10.1016/j.ijinfomgt.2018.11.005

Oracle. (n.d.). Data quality management. Oracle. https://www.oracle.com/master-data-management/data-quality/

PWC. (2019). GDPR compliance and data protection risks. PWC. https://www.pwc.com/gx/en/services/cyber-security/data-protection/gdpr-compliance-data-protection-risks.html