BDA601—Big Data and Analytics

Design Data Pipeline Assessment 1

Assesement 1 Report

A00097084   Farshid Keivanian

**Academic Integrity**

I declare that except where I have referenced, the work I am submitting for this assessment task is my work. I have read and am aware of Torrens University Australia Academic Integrity Policy and Procedure viewable online at https://www.torrens.edu.au/policies-and-forms I am aware that I need to keep a copy of all submitted material and their drafts, and I will do so accordingly.

Contents

[Introduction 3](#_Toc163169126)

[Potential Data Sources 3](#_Toc163169127)

[Internal data sources: 3](#_Toc163169128)

[Transaction data record: 3](#_Toc163169129)

[Email Marketing metrics**:** 4](#_Toc163169130)

[Online activity (placing items in an online shopping cart): 4](#_Toc163169131)

[External data source: 4](#_Toc163169132)

[Social media data (Facebook, YouTube or LinkedIn): 4](#_Toc163169133)

[Online Market Research data: 5](#_Toc163169134)

[Data integration challenges and strategies: 5](#_Toc163169135)

[Challenges: 5](#_Toc163169136)

[Data Quality issues: 5](#_Toc163169137)

[Scalability Problems: 5](#_Toc163169138)

[Disparate formats: 6](#_Toc163169139)

[Strategies to mitigate the data integration challenges: 6](#_Toc163169140)

[ETL: 6](#_Toc163169141)

[Figure1 6](#_Toc163169142)

[Extract data: 6](#_Toc163169143)

[Transform data: 7](#_Toc163169144)

[Load data: 7](#_Toc163169145)

[Data lake: 7](#_Toc163169146)

[Figure 2 8](#_Toc163169147)

[Data Pipeline Architecture 8](#_Toc163169148)

[Figure 3 9](#_Toc163169149)

[Conclusion 9](#_Toc163169150)

[References 10](#_Toc163169151)

# Introduction

This report is about the design of the data pipeline which helps businesses to critically analyse and get business goals.in this report, we are discussing the various sources of data to establish a data pipeline. Also, we will discuss the challenges come from the different sources of data and how to resolve it. Moreover, we will demonstrate some of the strategies and retrieval systems for the data that can be used for commercial purposes and or open-source big data.

According to the research (IBM, n.d.), we can determine that a data pipeline is a process of ingesting data from different sources which helps to ported data storage (data warehouse, or data lake) later that data will be used for analysis. There are two main types of data pipelines which are:

* Batch processing.
* Streaming data

# Potential Data Sources

The potential data sources that align with the objective for this organisation’s data driven strategy **are divided into two sources. that are:**

## Internal data sources:

Internal data is mainly generated inside the business or organization. This kind of data is produced by the company’s operations, some of the internal data sources from organisation are:

### Transaction data record:

The transactional data record is a detailed record of Purchases, sales transactions, and revenues.

Characteristics of transaction data record: it contains a time aspect (TIBCO, 2024). Transactional data is generally included in the structured data. These transaction data records are store in RDBMS, critically important for financial reporting and inventory management. Or it can be stored in data warehouse.

### Email Marketing metrics**:**

According to the research, we can determine that the email marketing metrics are indicators that assist to the success in the email campaign. the important to email marketing metrics are:

* Email deliverability rate.
* email open rate.
* click-through rate.
* Spam complaints.
* Unsubscribe rate.
* Revenue per email
* List growth rate and so on.

### Online activity (placing items in an online shopping cart):

Online activity also helps to analyse the data and gives insights results to the business for business goals. By tracking the information from an online shopping cart.

Characteristics:

this online activity is structured data because there is user information, cart information product.

These data are mainly stored in a structure database table or XML or CSV.

## External data source:

According to the research (Future Learn, n.d.) we can determine that the “External data can include almost everything from historical demographical data to, market prices or forecasting weather to social media trends”. Especially organization believes as a important assets on this external data source because they will analyse the model economic, political social or environmental factored that included their business goals.

According to the business case study we have some of the external data sources, that are:

### Social media data (Facebook, YouTube or LinkedIn):

According to the research (Newberry, 2022) we can determine that the social media data is kind of any type of data that can collected through social media and get meaning information for business purpose.

Characteristic of social media data are:

* clicks, comments, share, etc.
* Reach
* Impressions and videos
* Profile visit.
* Social share of voice
* Demographic data: age, gender, location, language, behaviours, etc.

Social media data is semi-structured data which can be access via API calls and also through Json format, providing real time feedback and trends.

### Online Market Research data:

Market research data is collection of data from consumer behaviour and economic trends which makes business improve their productivity and profit.

Characteristics:

It provides the market trends and customer behaviour which helps and analysis the data for business profit and can be used for analysis the competitor’s market. The reasonable assumption about the field of market research can be customer preference, marketing trends, product differentiate, marketing segmentation (Twin, 2023). It is a structure data which is store into the datasets, provided by the research firms.

# Data integration challenges and strategies:

## Challenges:

### Data Quality issues:

According to the research (Smallcombe, 2023) we can determine that the data can be seen vast number of inconsistencies when data are streaming in from different sources. we cannot filter the types of data coming from different sources (Smallcombe, 2023) A data lake may be store with the duplicate records and insufficient data which is no usable for management team or It management (Smallcombe, 2023).

### Scalability Problems:

Scalability problem is also a challenge for data integration in a data lake because data can come in large volumes from different sources which may strain system resources and can cause performance issues. during data integration and processing methods (Smallcombe, 2023).

### Disparate formats:

Disparate formats is also one of the challenges of integrating data in a data lake because all the data form different sources come in different types and formats which takes a lot of time effort, specialized tools and expertise to convert all the data into a unified and usable format (Smallcombe, 2023).

## Strategies to mitigate the data integration challenges:

### ETL:

According to the research (AWS, n.d.-c) we can determine that the ETL is a process of Extract, transform , and load where all the external and internal data are move form sources system to the destination system at periodic intervals.

### Figure1

Simple understanding of ETL process (AWS, n.d.-c)

A diagram of a cloud

Description automatically generated

These are the ETL process:

### Extract data:

Extract data involves puling data from different sources like relational and non-relational database, and API, website application. Social media mentions, online transaction, and so on.

### Transform data:

Transform data is the place where data are sorted and organised. And also, involves cleaning and validating the data, deduplication records and applying data transformations (Smallcombe, 2023).

Load data: Load data is a process of transfer of transform data into data storage repository (Smallcombe, 2023)

# Data lake:

According to the research (Google Cloud, n.d.) we can determine that a Data Lake is a centralized repository design to store, process and secure large amounts of structured, semi-structured and unstructured data. Most important a data lake can store data in any native format which helps to process data by ignoring size of the data format.

According to the research (AWS, n.d.) we can determine that the Data lakes on AWS (Amazon S3) helps us break down data silos to increase end-to-end data insights. The benefits of data leks with AWS are as follows:

* It stores all the data.
* Increase innovation.
* Seamlessly integrate and move data.
* It easily enables purpose-built analytics.
* It quickly deploys machine learning.

According to the research (AWS, n.d.) we can determine that the all the external data sources and internal data sources are integrated into the amazon S3 .And all the data sources are move to amazon S3 from where all external data and internal data are store in amazon S3 .

AmazonS3 is a object storage built to store all the data from different sources which access controls and management options. And allow uploading any amount of data (AWS, n.d.-a)

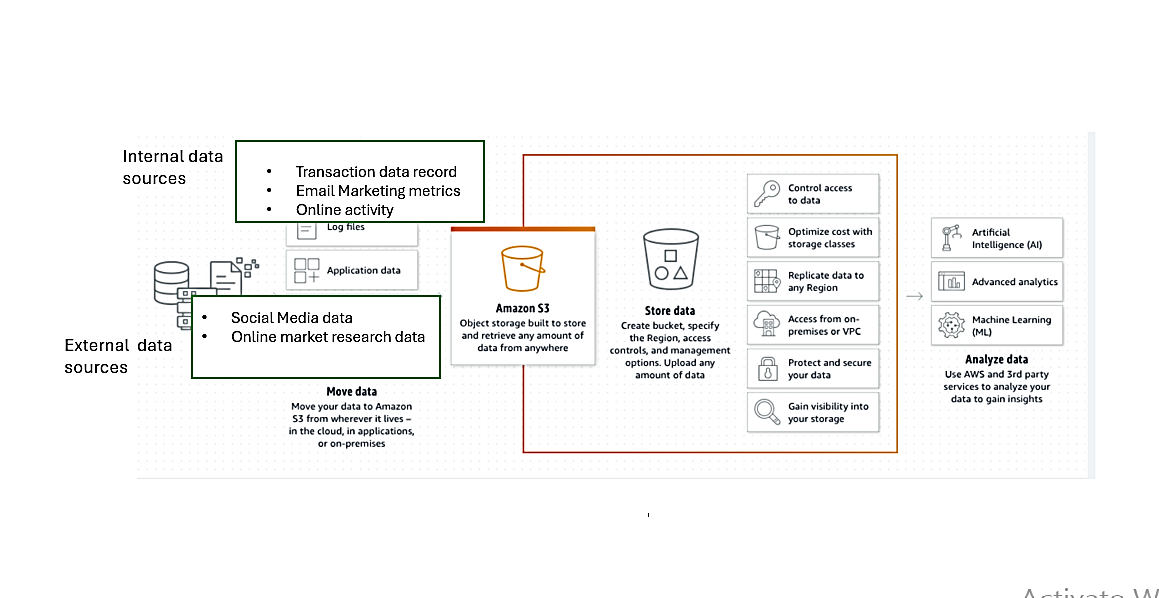
The feature of this Amazon S3 are:

* It gives access to control the data.
* It optimizes cost with storage classes.
* It replicates data any region.
* It gives access from on premises or VPC.
* Protect and secure our data.
* It gives visibility into storage to track the data.

Which is very useful for **management team**, It department, Advanced analytics and for machine learning where data can be analyse for business purpose.

### Figure 2

Amazon S3 (data Lake)





Note: This diagram is adopted from given resources.

# Data Pipeline Architecture

### Figure 3

A diagram of data storage

Description automatically generated

# Conclusion

In conclusion from this report, we learned how to identify various sources of a data for building data pipeline.in short we know about the challenges and how to mitigate the challenge for integrating data from different sources. Most important we demonstrate some important Etl process. we discussed about the amazon s3 storage service which helps to store data from different sources and analyse the data for business propose. At last, we developed schematic of the overall data piepeline and clearly show some sources data integration steps, and the component of the data lake and the interactions among all entities.

# References

AWS. (n.d.-a). *Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service*. Amazon Web Services, Inc. <https://aws.amazon.com/s3/>

AWS. (n.d.-b). *Data Lakes on AWS*. Amazon Web Services, Inc. Retrieved March 15, 2024, from <https://aws.amazon.com/big-data/datalakes-and-analytics/datalakes/>

AWS. (n.d.-c). *What is ETL? ETL Explained - AWS*. Amazon Web Services, Inc. <https://aws.amazon.com/what-is/etl/>

BasuMallick , C. (2022, August 22). *What is a Data Lake? Definition, Architecture, Tools, and Applications*. Spiceworks. <https://www.spiceworks.com/tech/cloud/articles/what-is-data-lakes/>

DataDrip Solutions. (2023, October 4). *Understanding the Magic of Data Warehouses and the ETL Process*. Www.linkedin.com. <https://www.linkedin.com/pulse/understanding-magic-data-warehouses-etl-process-datadrip/>

Future Learn. (n.d.). *Types and sources of data*. FutureLearn. Retrieved March 14, 2024, from <https://www.futurelearn.com/info/courses/data-analytics-python-statistics-and-analytics-fundamentals/0/steps/186574>

Google Cloud. (n.d.). *What is a Data Lake?* Google Cloud. Retrieved March 15, 2024, from <https://cloud.google.com/learn/what-is-a-data-lake>

IBM. (n.d.). *What is a data pipeline | IBM*. Www.ibm.com. <https://www.ibm.com/topics/data-pipeline>

Newberry, C. (2022, January 24). *Social Media Data Collection: Why and How You Should Do It*. Social Media Marketing & Management Dashboard. <https://blog.hootsuite.com/social-media-data-collection/#What_is_social_media_data>

Smallcombe, M. (2023, July 19). *Data Lakes: Integrating Data From Multiple Sources with ETL Tools*. Integrate.io. <https://www.integrate.io/blog/etl-tools-in-data-lake/>

TIBCO. (2024). *What is Transactional Data? | TIBCO*. TIBCO. <https://www.tibco.com/glossary/what-is-transactional-data#:~:text=The%20defining%20characteristic%20of%20transactional>

Twin, A. (2023, April 14). *How to Do Market Research, Types, and Example*. Investopedia. <https://www.investopedia.com/terms/m/market-research.asp>