ZOMBIE DATA

WHAT IS ZOMBIE DATA HOW USEFUL IN DATA ANALYTICS

BY AJIT DASH

DEAD DATA
Data not being used from years

GARBAGE DATA
Data thrown away

DATA INSIDE CLOSET
Data sitting in one place from years

REFERENCE DATA
Missing validation of reference data

PAST VALIDATION DATA
Validated data used in the past

Abstract:

This paper discusses the utilization of data and Zombie data. In short, if the data is not used, it becomes Zombie data. To restore data from a Zombie state to an active state, this book provides useful techniques for data analytics. Furthermore, the subsequent legacy data is useful for the accuracy of the predictive analytics and helpful in the decision-making process. These data could be helpful in time to market, time to customers, and time to business. This book explains the use of data wisely and effectively throughout the decision-making process, to avoid having the data become a Zombie data.

Introduction:

Zombie & Zombie Data

Before we get into the details of Zombie Data, let us first understand what a zombie is: A Zombie is described as a mute and will-less body (noun)[1]. Something that was declared as concluded, finished, or dead, but surprisingly continues to linger, or comes back in a different version (adjective)[2].

In our context, Zombie Data describes data that is not being used in the business operation or decision-making process for a very long time but can be useful if explored. Data in an organization is directly or indirectly related to some business necessity. This information from the data has never been explored to find the intrinsic values that lie within

Sometimes it is described as **"Dead Data", "Garbage Data", "Data Inside a Closet", "Unused Reference Data" and "Past Validated Data"**

[1] "Zombie" *Noun, n.d.,* https://www.dictionary.com/browse/zombie July 20th, 2021.

[2] "Zombie" *Adjective, n.d.,* https://www.dictionary.com/browse/zombie July 20th, 2021.

Dead Data:

These are data that was once actively used in the past, but now has become obsolete. A good example of this practice is when a company changes it's product line by creating another novel product, and subsequently, they sunset the old product. Now after sunset, they would have stopped using the information related to the old product. Hence, data such as ~~like~~ who had purchased the (old) product, the geographical areas with the most and least orders, the customer with frequent orders, top ten customers, and financial information such as profit and total revenue generated, for example, would be now considered as irrelevant.

Garbage Data:

Garbage data also known as Junk data are information that is kept separately for future use or as needed. One common example of this involves a dimension table for the reference keys or zip codes. This information is used as ~~a~~ reference as required. These are also known as Garbage or Junk data because it's not the primary data and does not connect to the primary table. Dropping or removing this information will have ~~a~~ minimal impact on the business operation.

Data Inside a Closet:

Data Inside a Closet is referred to as data that was once used for the business operation is not used now for any kind of operation. This data is sitting ideal without any value to the current operation. No one tried recently to determine the values it can bring to the current business operations.

Unused Reference Data:

Reference data could be a third-party data or data from any of the operations of the business unit such as sales leads. Reference data are like parasites, and they only bring value when they are being used with the main data. Reference data are quite often stored in a table and is a standalone table.

Past Validated Data:

Past Validated data could be of any kind; either an internal business operation data or outside vendor data. This data already being collected or used in past operations is also being validated by the end consumers. This data at present not in use.

How Data Convert to Zombie Data:

Usually in an organization, business directions change with time for various reasons, eventually impacting the products and services of the organization. The changes in the products and services, also affect the markets, customers, and manpower requirements. With time, all

the old information about the products, services, markets, customers & manpower get obsolete due to minimal or no use. Consequently, this once active information (or data) becomes Zombie Data (or Inactive).

Examples:

Fashion Industry:

In fashion industries, product line changes frequently along with the customers, and demand for the product changes. However, the old customer's data could be helpful to determine the past experiences of the customer and could be used as a reference to the new customer experience. This may create a relationship between the old and the new product lines; thereby, determining the product's future, change in customers' behavior, market geographies, and amounts spent by the customers.

If the old product line data has never been used after the product line being sunset then the data related to the sunset product line such as customer, revenue generated, the geography of the customers, etc. could become Zombie Data.

Technology Product Manufacturer:

Companies frequently change their products to add various features that meet customer requirements. For example, to enhance efficiency and user-friendliness of the system. One good example is a cell phone manufacturer. Cell manufacturing companies have two components, hardware

and software. Hardware changes are easily captured in drawings, which illustrates the internal changes of the electronic components. The software changes come as an add-on version, wherein the software updates are pushed over-the-air to the device by the manufacturer.

With hardware changes comes the release of a new product as per the product release cycle. Once the new products are launched the old ones get obsolete. Eventually, these old versions are not supported by the manufacturer. With each new release, new software versions are pushed as per the release cycle.

Each new release rollout by the manufacturer is actively monitored by the manufacturer to understand how the market is reacting to the new release. Eventually, the manufacturer takes those feedback and comes with new changes for a future rollout.

Mostly, the data being captured for each release is hardly compared with only a few old releases or the old release data never been used.

These releases could impact market segments, customer usages, hardware issues, and software issues. If this information is not well compared with old release information, this may create a gap in release for the product manufacturer to understand ~~the~~ customer sentiments and ~~the~~ direction of the business.

Eventually, those feedback data not being used get obsolete and become zombie data. ~~But~~ these data actually have equivalent insight as compared to the new data being captured for the new release. If the data for the old and new releases are being used together it will give

historical data inside and this may help understand the customer sentiments, time to business, time to customer, and time to market for a better comparison.

Also, the use of old data prevents it from being named a zombie data.

Service Industry:

Service companies help clients with businesses or technology consultation by providing consultants. The success of the service industry is directly proportional to the solution expertise and problem-solving skills it has. Service companies provide their expertise as required by the client to meet the on-demand requirements. Each consultant brings their own knowledge and expertise, to solve problems and provide unique solutions to the client. In turn, they gain experience and learn from the client's project and problems faced. It is a continuous process of delivery of expertise and learning.

If the service company does not have a way to keep track of the use cases and pain points resolved by the consultants, then the knowledge becomes obsolete. If the knowledge captured by the consultants has never been exposed or recorded, due to a change in the job role or iteration, then the learning eventually becomes useless .The knowledge, information, and data collected becomes a Zombie Data.

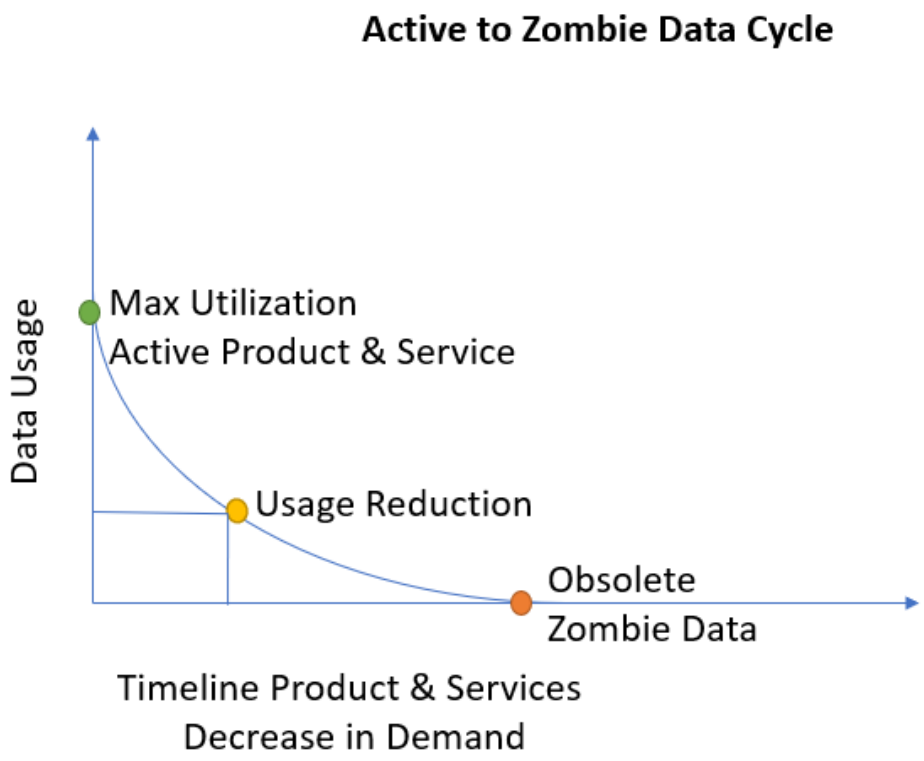
Assembly line: Automobile Manufacturer

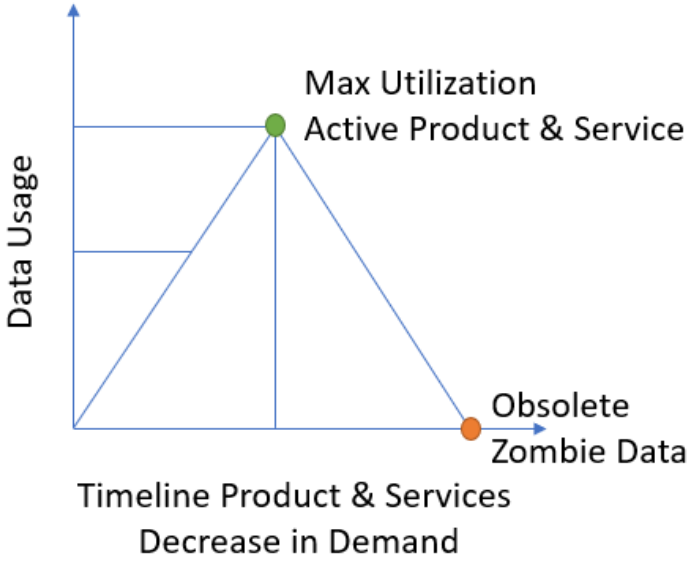
Let's discuss about an automobile manufacturer whose sunsets for a particular automobile model. The data for the automobile model collected before sunset if not used for any of the future data analysis, then those data may become obsolete. Eventually the data becomes a Zombie data

Similarly, for other domains such as Insurance, Finance, Banking, Travel, Media, and Hi-tech, etc.. The data becomes obsolete and usually into Zombie data if data collected before the sunset is not used after sunset of the product

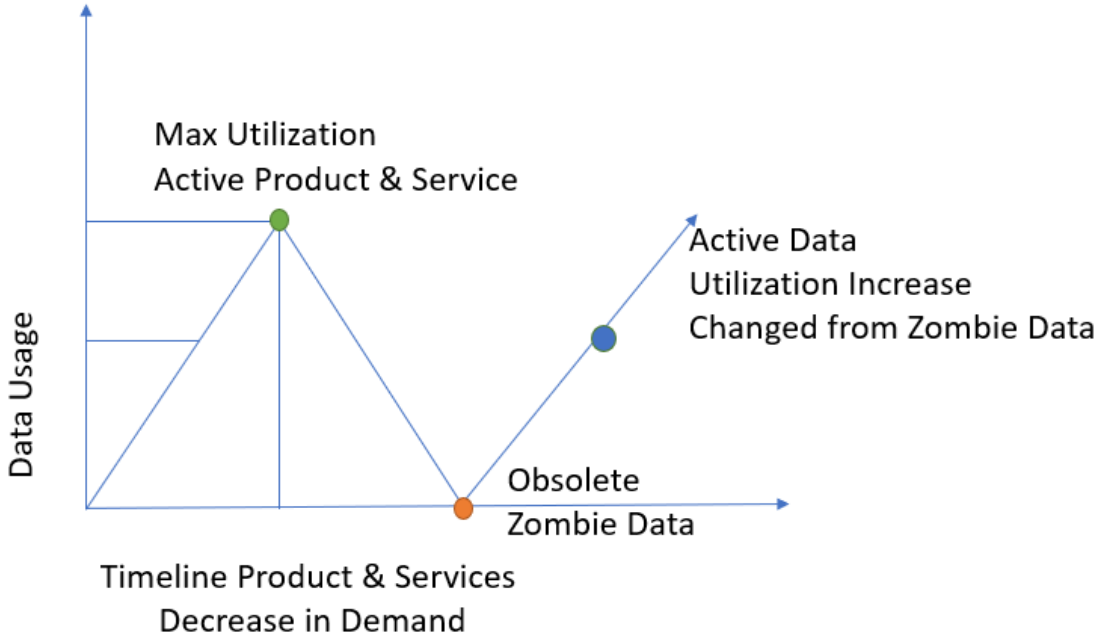
Life Cycle of Zombie Data:

The zombie data utilization cycle is given in the below chart. The maximum utilization of data is shown when products or services are active. Similarly, when their usefulness is reduced then the usage of the data reduces accordingly. If the products or services are not in use, then it is said to be in its inactive state, in which case becomes obsolete and thereby, Zombie





Zombie Data Transfer to Active Data



Zombie Data Framework:

Zombie data framework is all about the utilization of data to find out valuable insight that can used for any decision-making process. Also, it is crucial to know the data maturity and incorporate the data in the corporate data culture.

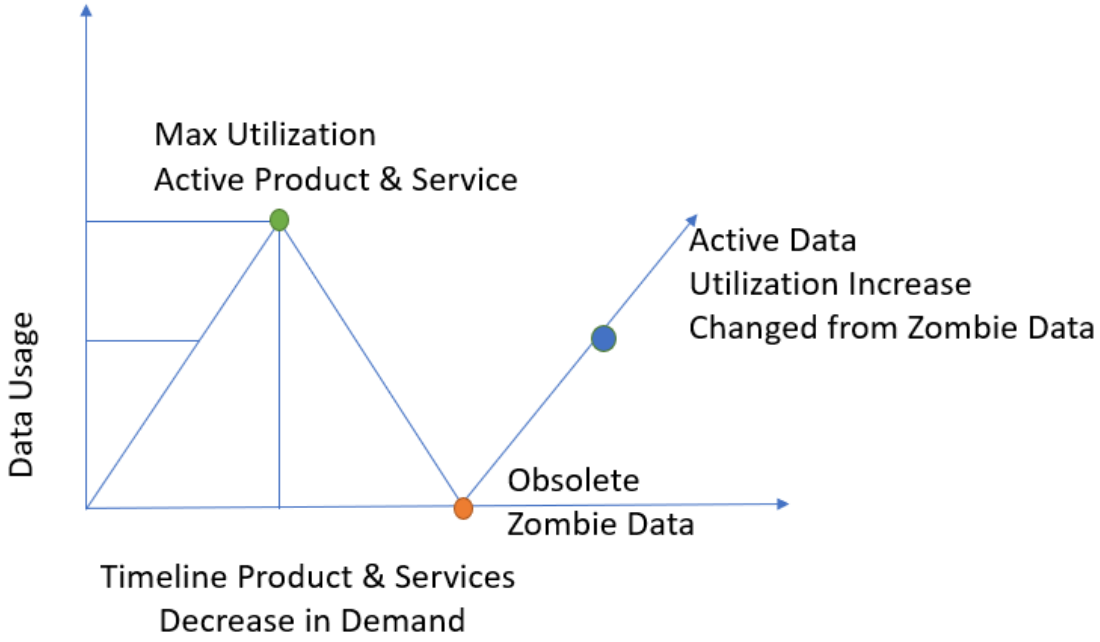
Keeping Track of Active Products and Services Information/Data	**Using the Data to Find out Valuable Insight Making Sure the Data is Part of the Decision-Making Process**
Utilization of the Data /Information Over the Time Period for Active Products and Services	Understand the Data Maturity and the Information /Data Part of the Corporate Data Culture

Time To live

As long as the data for products and services are active and in a live state, the utilization of this data is always in progress, known as "Time to Live". When these data are neglected or ~~less~~ seldom, used perhaps due to the reduction of demand, it goes out of Time to Live state. Then, to Zombie data state.

Activation and Utilization of Zombie Data

Data from Zombie State to active state converted as per the utilization. If the data remain unused then the data move from the active data to Zombie data. Only Zombie Data converted to Active data after utilization. This is the life cycle of the Active -Zombie to Active data.



Keeping track of Zombie Data & Advantages:

When Zombie data is used for business analytics or business insight analysis, it is necessary to keep track of the utilization and benefits. Most companies use Zombie data along with the current data captured from products or services. It is nothing but a mixture of recent data with the history (Zombie) data. Usually, this is helpful in predictive analytics in training ML models

These are major advantages of using Zombie data with current data:

Determination of Value Proposition

- Problem
- Competition
- Market Fit

Strategy:

It helps in decision making, finding strategy for the business Founders, Advisors and Partners

Marketing Strategy:

Zombie data along with the current data will help in marketing strategy such as price, promotions, and placement

Financial Strategy:

A better financial strategy could be possible with the combination of Zombie data ~~to~~ with the current data to create efficient and accurate financial models for Sale, cost, revenue & funding

Data Culture and Zombie Data

Data culture provides information about the organization's data maturity and data-driven capabilities.

“Data culture” refers to a company’s ability to use data to make decisions. Companies with a strong data culture have robust analytic capabilities. All the strategies are driven by analytical information derived from a strong data culture. Zombie data gives information

about past anomalies, challenges for products and services the company offers. It's always recommended to have the Zombie data part of the data culture.

These are the advantages/disadvantages of Data Culture when combined with Zombie Data:

Advantages of Data Culture

- Improve the process and understanding of data usage
- Accelerate customer experience
- Act as a catalyst for business strategy and decision making
- Determine the data maturity of an organization
- Enhance business ROI and creates better business opportunities

Disadvantages of Data Culture

- Without proper Data Culture creates trust issues for data
- Takes longer for decision making and organizational strategy
- Creates data silos and gaps between the different business units
- Repetition of duplicate data and data clarity

How to integrate zombie data with existing data:

Integrating Zombie data with the existing active data is a similar process to the active data. However, the Zombie data requires less data quality process as compared to the active data. As well, it may need more data filtration than the active data based on ~~the~~ data usage requirements.

The data integration process depends on the ecosystem (i.e., Data Warehouse, Data Lake, Lake House, or a Delta Lake and downstream consumptions).

Data integration happens in cleanses/integration/query layer before it goes for the downstream consumptions. Data could be integrated using a common existing key or through an artificial key generated by the system.

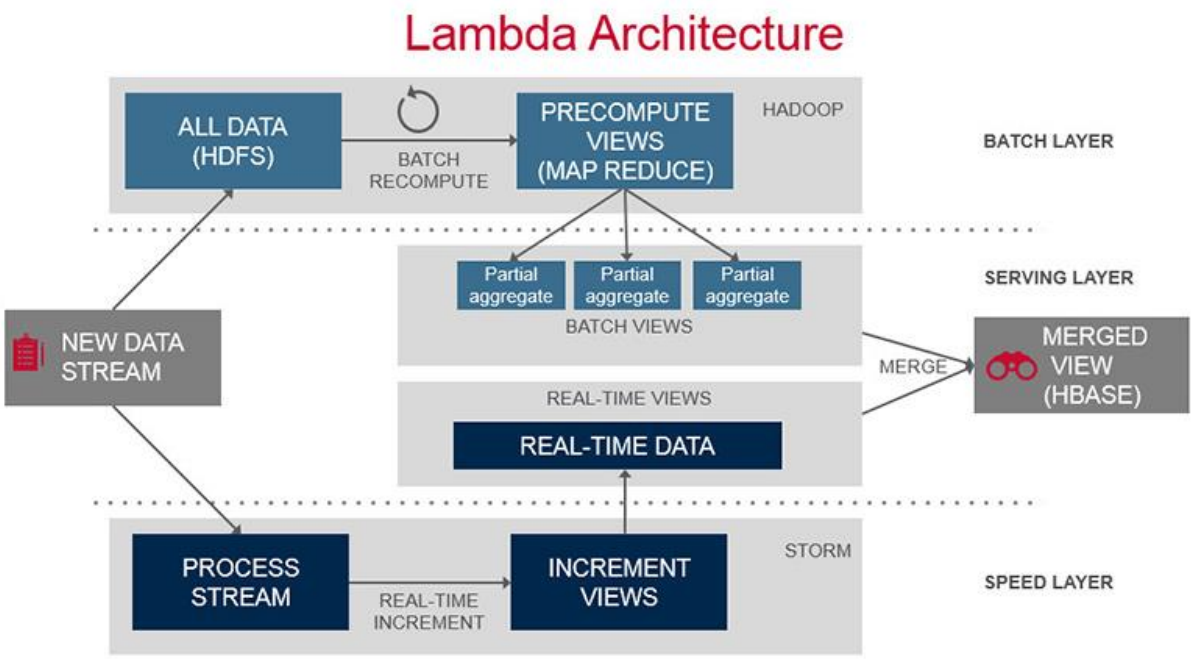
Cleansing and Partitioning of zombie data

Cleansing and partitioning of Zombie data requires the use of the same ETL and ETL process as the active data. This process mainly depends on business requirements. It is not necessary to consume all Zombie data for processing, only consume as per the business needs. Also, it is recommended to use data partition and filter.

Processing of Zombie data

Zombie data could be processed either in real-time, batch, or mix mode (batch + real). To process it a special kind of architecture is recommended such as Lambda Architecture. The main advantage of the combination of real-time data with the history data increases the percentage of accuracy for an analytic model.

Also, the training of your ML model could perform well with training the data model with historic data.



Lambda Architecture[3]

Data types, Profiling & Governance of Zombie data

Zombie data does not have any data restriction as was used in the past. Zombie data may require to profile the data as required for the new operation this could be achieved through a data profiling tool.

[3] Image Reference from MapR

Conclusion:

As per our above discussion, usage of the data is much more important to keep the “data” in an active state. Unused data may move to a Zombie state. If the data went to a Zombie state, it could move to an active state by using it again.

With a proper data culture, data could save from moving to a Zombie state from an Active state. Also, data prediction accuracy increases by using the history and current data together.

About the Author

Ajit Dash has spent more than 24+ years in data and analytics in various capacities, led various projects as a Sr. Director Data / Cloud Advisor/ Solution Architect / Cloudy data strategy manager/Advance Analytics/Data Scientist Lead, Pre Sales Lead providing Enterprise and Cross platform integration solutions to various corporations.

His expertise includes strong hands-on analysis and design of Enterprise Solution Architecture, Cloud Advisor, Data Lake, Bigdata, Data Sc, Data warehouse and Data mining, Database Management/Integration, BI Reporting, and Cross-Platform

Domain Expertise: Telecommunication, Biotech, Finance, Banking, Media, Aerospace, Insurance and Technology etc.: (Clients: Fox, Oshkosh, Otis, Travelers, Apple, Qualcomm, IBM, LPL Fin. etc.)

Education:

Ajit Dash holds a master's degree in General Management from Harvard University

Master's in Computer Information Systems from University of Phoenix

Bachelor's Degree in Electrical Engineering from India

Blog: http: //www.thedataworld.org