International Journal of Business Analytics and Intelligence
13 (1) 2025, 33-43
www.publishingindia.com/ijbai/
https://doi.org/10.21863/ijbai/2025.13.1.004

Submitted: 05 March, 2025
Accepted: 20 March, 2025

Importance of Master Data Management (MDM) in Artificial Intelligence (AI)

Avinash Jha*

Abstract

A typical organisation today has a process that spans across multiple entities like Order Procurement, Financials, Customer Relations, Inventory, Supply Chain & Distribution. This indicates critical business data flowing between the entities interwoven together to conclude the business transaction at the end of the day. With the industry shifting its gears towards welcoming Artificial Intelligence it is important to ensure readiness of the data flowing in the system across these entities. In 2023, McKinsey surveyed more than 80 large global organisations to understand their top objectives in maturing their Master Data Management and the report indicated the top four objectives as: Improved customer experience and satisfaction, Revenue growth by presenting better cross-selling and up-selling opportunities, Increased sales productivity or operational efficiency with seamless access to master data and Streamlined reporting (Kayvaun Rowshankish et al., 2024). The importance of clean and relevant data in Artificial Intelligence cannot be stressed enough. The Supervised Learning in Neural Networks & Deep Learning relies on the Labeled Data [Features mapped with Target] for training the model and aiming for high accuracy in prediction. The words predicted in Generative AI (Gen AI) over Generative Pre-trained Transformer (GPT) can rely on the clean business data flowing in the system.

Keywords: Master Data Management (MDM), Artificial Intelligence (AI), Single Source of Truth (SST), Hub and Spoke System, Mastered Data

Executive Summary

Research in the field of "Artificial Intelligence" has shown that meaningful data help avoid both under-fitting and over-fitting. Clean data would give the right amount of information to train the model, thereby increasing its efficacy and striking a balance between "Accuracy" and "Performance".

"Master Data Management (MDM)" is the solution for providing such clean data that adheres to Data Integrity, Data Consistency, Data Validity, Data Quality and Data Governance, thereby providing an integrated, consistent, relevant and non-redundant data source that can be used in Regression and Classification models.

Importance of Master Data Management (MDM)

The prevailing challenges that an organisation faces in terms of the data quality are incompleteness, inaccuracy, non-uniqueness, non-compliance, obsoleteness and non-relevance of data flowing across multiple business units. While running for several years, business units tend to operate in silos, maintaining their own data sources. This eventually result in data inconsistency between them that later becomes a pain to assimilate in terms of time, labour and cost.

These challenges pose hindrances during the year-end consolidation and reporting. What is even more annoying is one or more days per week spent in identifying and correcting the faulty data and bringing it into the acceptable state. Such an additional effort is unacceptable when it comes to ROI, as the effort could have been channelised instead to win Sales and increase revenue.

Master Data Management (MDM) can be implemented as a solution to the problem with a collaborative effort between Organisations' Stake Holders, Business Units Analyst, Functional Members and Engineering Team

* Sr. Manager & SME – AIA MDM, Cognizant Technologies, United States of America. Email: jha_avin@rediffmail.com

in identifying the data sources and understanding the business processes along with dependencies that integrate them. The idea is to shape a Single Source of Truth (SST) composed of critical, unique, complete and relevant data that is referred by various business units in their day to day business use case. The MDM repository gets updated with the latest data pulled from various source systems on a frequent basis (eg. daily, weekly, bi-weekly). The data brought in the SST is then cleansed by process of "Mastering" and then published back to the source systems for assimilating. This two-way flow of data helps to keep source systems in synchronisation enabling them to transact on the same page without any ambiguity.

With this kind of central repository of clean data shared across all source systems in a "Bi-Spoke" model or "Hub-and-Spoke" model, business units can be confident about data consistency and can rely on the unique data for cross-integration. This synchronisation between source systems is important as it enables the system to work in harmony.

Implementing Master Data Management – A Collaborative Effort Between Stakeholders, Business, Functional and Engineering

Challenges and Approach: It all starts with convincing the organisations' stakeholders about the importance of implementing MDM. This sometimes is challenging since the benefits reaped are not immediate and it gradually becomes more apparent as the system matures and tunes up to the technology and process. So cost benefit analysis can be sensed only using existing case studies of successful implementations.

Further, it is important for a business/functional analyst to work with architects and infrastructure team to understand the overall system and process that binds various sources together for data flow. Understanding of the overall architecture, platform and infrastructure is significant for implementation of MDM. It helps to reveal the footprint of the existing hardware and network topology on which the existing source systems are setup and operate. The topology discloses some of the significant aspects like capacity (CPUs, Memory and vNIC) of servers, virtualisation exposing virtual machines and nodes, routing table and security list for controlling outgoing and incoming traffic between the subnets, local peering gateways for dedicated routing within private subnets, firewalls and ports opened for traffic, operating system on which the application server is configured and the database.

Then comes the time when engineering needs to get involved to configure and develop the MDM solution as per the topology and ensure that the "Source of Truth" is accessible to various source systems as part of integration. The clean data is provided to all source systems by way of publishing in the preferred format (eg. CSV, JSON).

In this entire journey of implementation, the business needs to get their engineering team to make necessary changes in their existing technology to consume the clean data published by MDM. This is critical for having their source system database updated with this clean data on a frequent basis. It is also equally important for engineering team to provide a file in the preferred format to the MDM containing the unclean data that needs to get mastered or cleansed in MDM on a regular basis and sent back for consumption.

Overall Value: By this MDM process, all the source systems come into synchronisation with each other over the unified data published from the MDM "Hub and Spoke System" through mastering, thereby letting business units operate in harmony, performing their day-to-day business activities and taking smart decisions with accurate information. This information architecture provides a single definition of customers, orders, product, suppliers etc. that are important for a business. One good example of such an industry specific implementation of MDM is the Customer Data Hub (CDH) offering from Oracle e-Business Suite that helps companies to efficiently manage customers, orders, manufacturing, inventory, billing and payments in finance and shipping due to the single enterprise view of the customer base (Oracle e-Business Suite, n.d.).

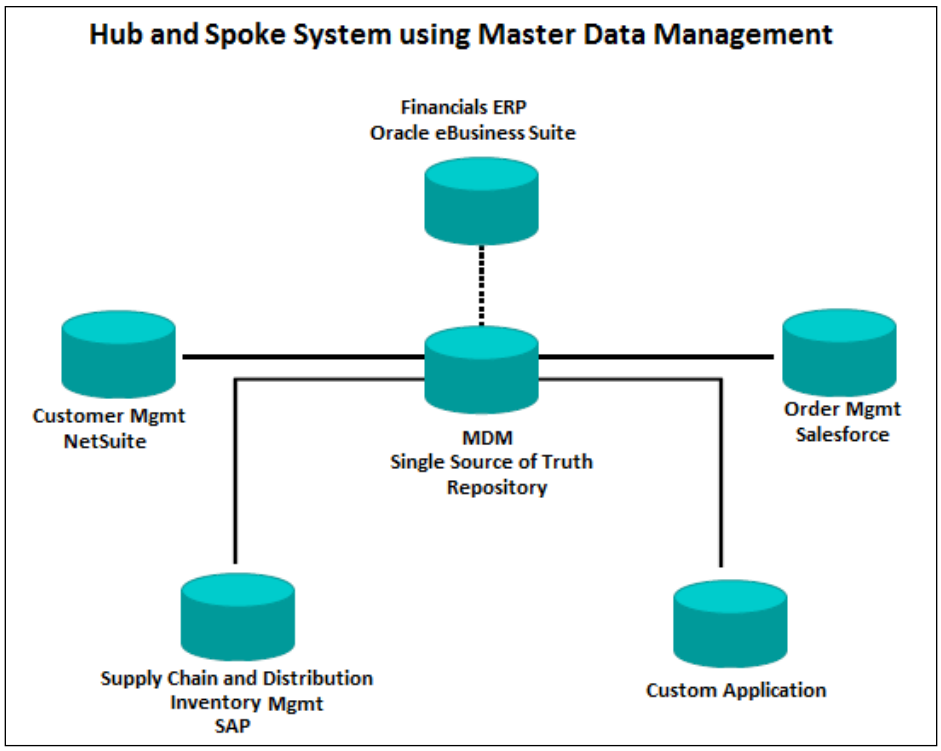

Fig. 1

Features and Tools of Master Data Management

Master Data Management (MDM) comes with key features: "Match and Merge" for not only recognising duplicates and preventing their creation in the repository but also resolving existing duplicates already sitting in the repository; "Data Quality Check" for ensuring completeness of data; and "Third Party Tool Integration" for enriching and validating data. These features guarantee valid, accurate and consistent data across the system being referred by various sources to communicate with each other.

Duplicate Identification and Prevention Using "Match & Merge"

Identifying and Preventing Duplicates Using "Match"

MDM ensures no duplicate record gets created in the MDM repository. This feature identifies duplicates and then leaves onto the user discretion to decide whether to allow such duplicates to get created within the system.

Example: Let us consider a scenario where the incoming record from a source system comprises attributes: organisation name, phone number, address line 1 and city.

Incoming Record from Source System			
Organization Name	*Phone Number*	*Address Line 1*	*City*
ABC Corporation	123-456-7891	1234 Cross Blvd.	Foster

Using the Duplicate Identification process, the duplicates are identified using the "Match Rules" setup, where the attributes are assigned respective "Matching Scores" based on their weightage or importance.

Match Rule Setup: Attributes with Matching Scores Defined		
Attribute	*Usage*	*Matching Score*
Organization Name	Scoring	20
URL	Scoring	30
Phone Number	Scoring	25
Address Line 1	Scoring	15
City	Scoring	10

The attributes of the incoming record are matched with attributes of existing records within MDM repository based on the EXACT or FUZZY match algorithm logic and Match % is calculated. Match % is the consolidated

score as a percentage of the total possible score. The higher the Match %, the closer it is considered as a "Match".

Match % = (Matching Score / Total Possible Score) x 100

Total Possible Score is the sum of all the Matching Scores for incoming attributes; in this example, it is 70.

Organisation Name (20) + Phone Number (25) + Address Line 1 (15) + City (10) = 70

Now that the Matching Score and the Total Possible Score are known, the Match % can be calculated based on the attributes undergoing a match.

The Match % for Phone Number (25) + Address Line 1 (15) + City (10) that matched with Record #1:

(50 / 70) * 100 = 71.4%

The Match % for Phone Number (25) that matched with Record #2:

(25 / 70) * 100 = 35.7%

The Match % for Organisation Name (20) + Phone Number (25) + Address Line 1 (15) matched with Record #3:

(60 / 70) * 100 = 85.7%

The Match % for Organisation Name (20) + City (10) that matched with Record #4:

(30 / 70) * 100 = 42.9%

Those records whose Match % meet or exceed the Threshold % are returned as potential duplicates. So if the Threshold % is set as 70% then Record #1 and Record #3 in the repository would be returned as potential duplicates.

	Potential Duplicates Identifed				
	Organization Name	*Phone Number*	*Address Line 1*	*City*	*Match%*
#3	ABC Corporation	123-456-7891	1234 Cross Boulevard	Foster City	85.7
#1	ABC Corp	123-456-7891	1234 Cross Boulevard	Foster	71.4

Resolving Duplicates Using "Merge"?

Practically, before the implementation of MDM since there is no duplicate identification check in place there is a high possibility of duplicates getting created in the system. These duplicates need to be resolved by the process of "Merge".

In the process of Merge, the golden record from the preferred source system, also known as the "Winner" or "Survivor" record, is retained as the only active record while others are inactivated. The process ensures that the "Looser" or "Non-Survivor" source systems are made to refer to the golden record so that at the end of the day, all the sources refer to one record entity. While performing this switch over to the new reference, the old reference is maintained as history that might be needed during un-merge.

Example: Since MDM has identified two duplicate records matching the incoming record, it needs to make sure that only one of them is the active golden record and the other is inactivated.

Record #3 is owned by Source System: "Salesforce" and Record #1 is owned by Source System: "SAP". Salesforce points to Organisation ID: 101 and SAP points to Organisation ID: 102.

	Potential Duplicates Identifed					
	Organization ID	*Organization Name*	*Phone Number*	*Source System*	*Source System Ref #*	*Status*
#3	101	ABC Corporation	123-456-7891	Salesforce	SLF1234	Active
#1	102	ABC Corp	123-456-7891	SAP	SAP5678	Active

Further assume that "Salesforce" has been declared as "Winner" or "Survivor" by the business. In that case the "Merge" process ensures that "SAP" is made to switch and refer to Organisation ID: 101 instead. This is done by inactivating the old reference and creating a new active reference mapped to Organisation ID: 101. As a result, "SAP" would now see "ABC Corporation" instead of "ABC Corp".

	Potential Duplicates Identifed					
	Organization ID	*Organization Name*	*Phone Number*	*Source System*	*Source System Ref #*	*Status*
#3	101	ABC Corporation	123-456-7891	Salesforce	SLF1234	Active
#1	102	ABC Corp	123-456-7891	SAP	SAP5678	Inactive
#1	101	ABC Corporation	123-456-7891	SAP	SAP5678	Active

This way all the source systems refer to the single golden record from the source of truth.

Data Transformation Rules

"Transformation Rules" help to check the incoming data and perform necessary transformation for standardisation before it gets created or updated in the MDM repository. These rules are driven by the business needs. Transformation Rules can also be applied to the outgoing data published back to various sources after it gets mastered.

Examples of such transformation rules are: "Limit postal code to only 5 digits", "Change Organization Name to upper case", "Concatenate Address Line 1, City, State and Postal Code into single Address attribute", etc.

Veeva Network is an MDM offering for the pharmaceutical industry that provides features for such transformation rules. The Network Expression (NEX) Rules can be applied on any attribute to enforce a transformation before it is ingested into the system as part of Source Subscription or before it is published as part of Target Subscription (Veeva Network, n.d.).

Data Quality Check

Another powerful feature of MDM is the "Data Quality Check" performed on the incoming and outgoing data. These are again driven by business needs and ensure the basic completeness of the data before it is absorbed or published.

Examples of such Data Quality Check are: "Address Line 1 cannot be null", "Country Code needs to be replaced with USA if null", "Phone Number has to be 10 digits", etc.

These checks are necessary for data to be qualified for further processing. Any failures are reported back to the respective source systems to correct the data and sent back.

Symarchy is an MDM offering that provides features of such data quality checks. The SemQL Enrichers can be applied on any attribute to enforce a quality check before it is ingested or published (Semarchy, n.d.).

Integration with Third Party Tools

MDM is flexible to integrate with other third party offerings to acquire and maintain data for its accuracy.

One such example is integration with "Dun & Bradstreet", which is a premium provider of business information related to organisation's financial and credit report, legal report and hierarchies. This can be used to verify and enrich existing data like tax identification number, D-U-N-S number, which are unique nine digit numbers assigned to a business.

Another example is integration with "Vertex" or "Trillium" for address validation. This can be used to validate and complete the user entered address, which is a combination of Address Line 1, City, State and Postal Code (Precisely, n.d.; Vertex, n.d.).

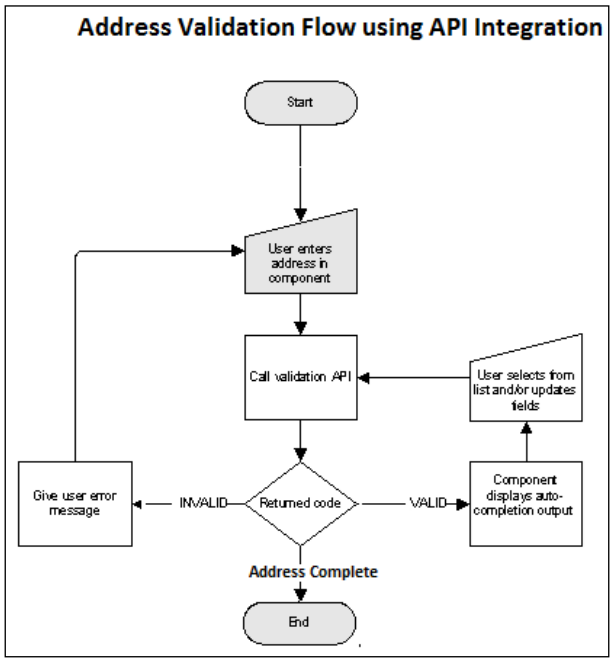

Fig. 2

Example: Let the user enter a partial address without a zip code: "400 CROSS BLVD., BRIDGEWATER, NJ" and get it validated by third party tool: "Vertex" or "Trillium".

The API interacts with such tools and returns back the correct and validated address: "400 Crossing Boulevard, Bridgewater, NJ 08807".

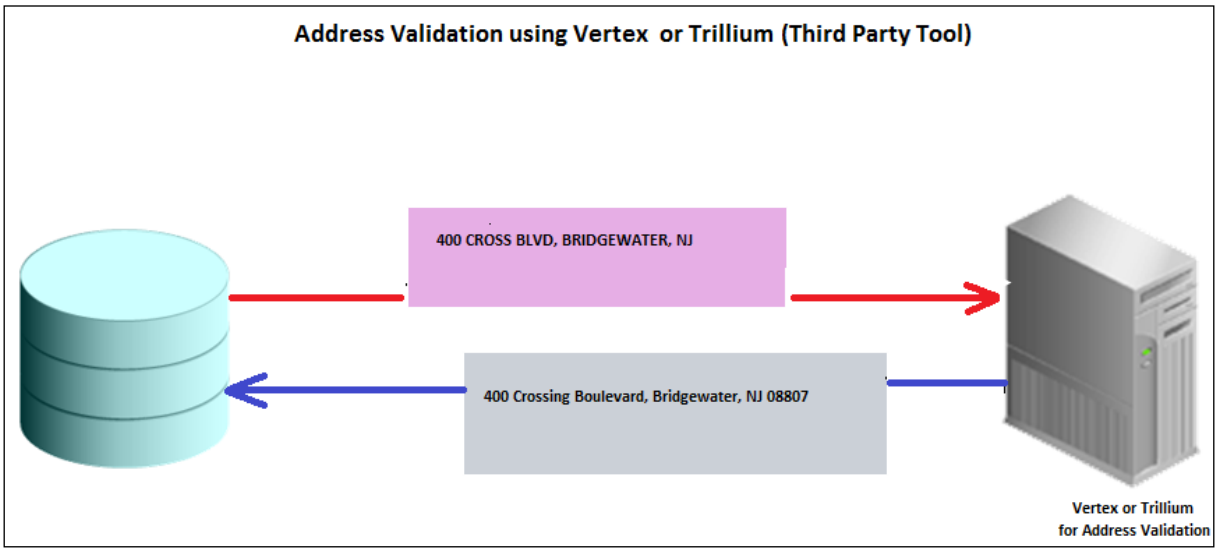

Fig. 3

Master Data Management Configured within Demilitarised Zone (DMZ)

Configuring MDM behind a "Demilitarized Zone (DMZ)" is significant for having a network perimeter isolating HUB from the intranet and providing controlled access to all source systems sending and consuming data. This includes important components, namely: firewall, virtual cloud network (VCN), security list, routing table, gateway, subnet, reverse proxy, secure communication protocol (SSL/TLS) to define such a perimeter, allowing authorised access to HUB from various source systems.

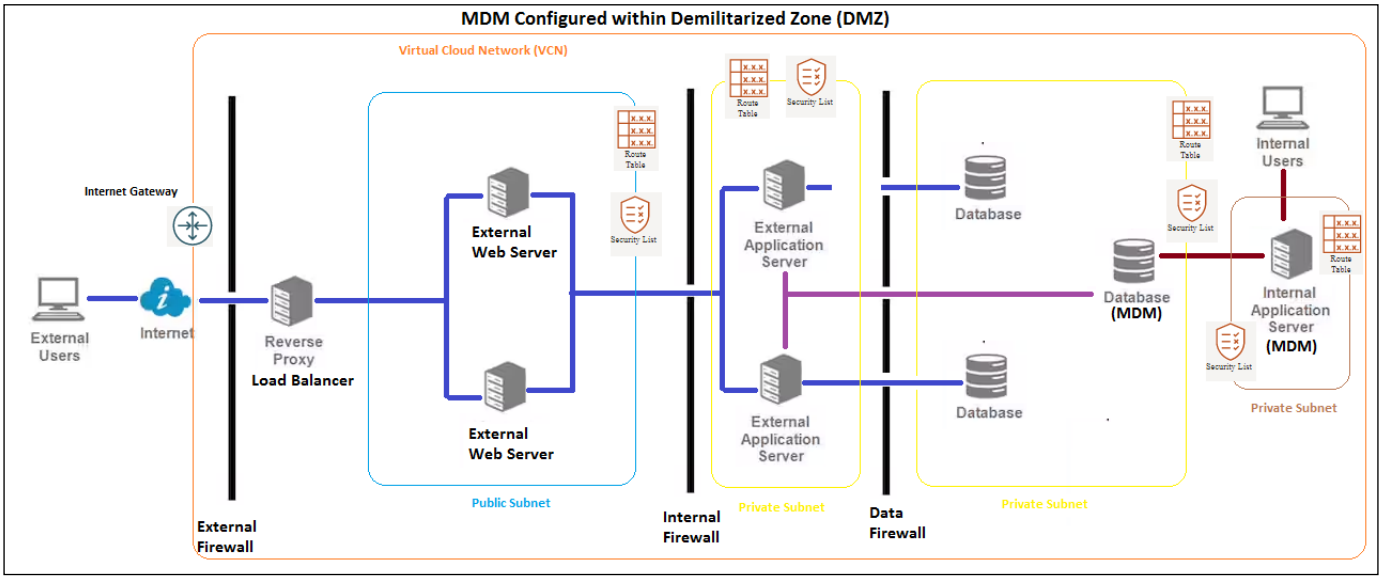

Fig. 4

Artificial Intelligence (AI) Fed with Cleansed Data from Master Data Management

For a good performance of an "AI Model" during training and for the most accurate prediction during testing, it is significant to feed clean and relevant data. That is why MDM provided mastered data is important, whether we talk in terms of recurrent neural network, convolution neural network or transformers.

Generative Pre-Trained Transformers (GPT)

The cleansed data from the MDM fed to the Generative Pre-trained Transformers (GPT) model would result in better predictability of words in Natural Language Processing (NLP) tasks in terms of possibilities relevant to the context.

In Generative Pre-trained Transformers, each word is maintained as "KEY Vector" in a multi-dimensional vector space. The contextual words that are closely related to each other in meaning are semantically maintained close to each other as "VALUE Vector". Some of the answers to questions associated with these words are maintained in "QUERY Vector" for example: Is the word preceded with an adjective?

The DOT PRODUCT between these three vectors together shows how valid the words are and how closely they are related to each other contextually. Mathematically, the closer the DOT PRODUCT of two vectors to "1", the closer the match is for prediction.

-1 <= DOT PRODUCT of VECTORS <= 1

DOT PRODUCT between these three vectors is explained in the Google Published Paper (Ashish Vaswani et al., 2023) under "Attention".

$$\text{Attention}(Q, K, V) = \textbf{Sigmoid}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Note the function: "Sigmoid" is used in the expression to tune the value between 0 and 1 indicating probability. Softmax and Rectified Linear Unit (ReLU) are some of the other similar alternatives.

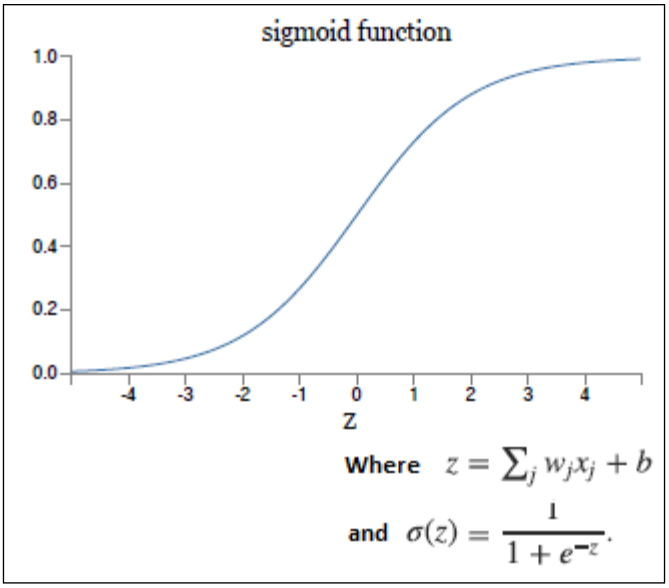

Fig. 5

The input coming in from the user is stored in the form of Embedding Vector in a multi-dimensional vector space and undergoes the VECTOR DOT product with KEY, QUERY and VALUE vector to determine different possibilities of prediction.

Speaking about recurrent neural network or convolution neural network with encoder-decoder architecture or speaking about the recent transformers, the multi headed "Attention" mechanism has gained focus.

Back Propagation in Neural Networks and Deep Learning (Nielson, 2015)

The cleansed data from MDM fed to neural networks model with multi-layer perceptron (MLP) would result in better predictability with the highest accuracy and lowest cost in image recognition tasks.

Back Propagation, as the name indicates, is about thinking backward how the output is related to a set of inputs with associated weights and bias.

Consider the figure below containing three "Input Neurons": x1, x2 and x3 as user input. These "Input Neurons" have associated "Weights": w1, w2 and w3 all connected to one single "Output Neuron" with "Bias": b.

The "Weighted Sum": z = [(w1*x1 + w2*x2 + w3*x3) + b] is passed to the Sigmoid Function σ(z) to get the value of "Output Neuron" reflecting the probability of a match.

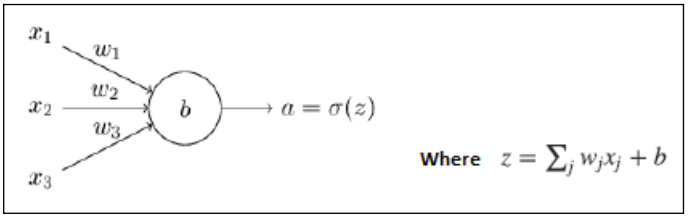

Fig. 6

While training the model, these "Weights" and "Bias" are adjusted according to the expected output. However, while validating or testing the epoch-batch the data is fed one at a time to the model and the probability for various possible outputs are determined by assigning "Weights" and "Bias" and by tuning "Regularization Parameters" as per "Stochastic Gradient Descent". Finally, the difference between the "Predicted Value" and "Expected Value" is determined for each of the possible outputs, also known as "Cost": C(w,b).

$$C(w, b) = |y - a|$$

Where y is expected value
and a is predicted value

The average Cost from all the outputs: x is calculated for each test data and that would be the Cost for the specific test run. Various Machine Learning algorithms can be used for determining the average cost namely "Mean Absolute Error", "Mean Absolute Percentage Error", "Mean Square Error", "Root Mean Square Error". The lower the "Cost", the better is the "Accuracy".

$$C(w, b) = \frac{1}{n} \sum_x |y(x) - a|^2$$

Where y is expected value
a is predicted value
x is the number of possible outputs

The Cost of each validation data or test data is determined in the same way and finally the average is taken to find the overall "Cost" of the model.

Thinking mathematically from the point of view of calculus, if we plot a graph between "Cost" and "Test Data Features", then we can find the "Point of Minima" where the rate of change of "Cost" w.r.t "Test Data Features" is almost equal to 0. The goal is to find such a point indicating the best combination of weights and bias, under specific regularisation parameters, at which the cost is minimum.

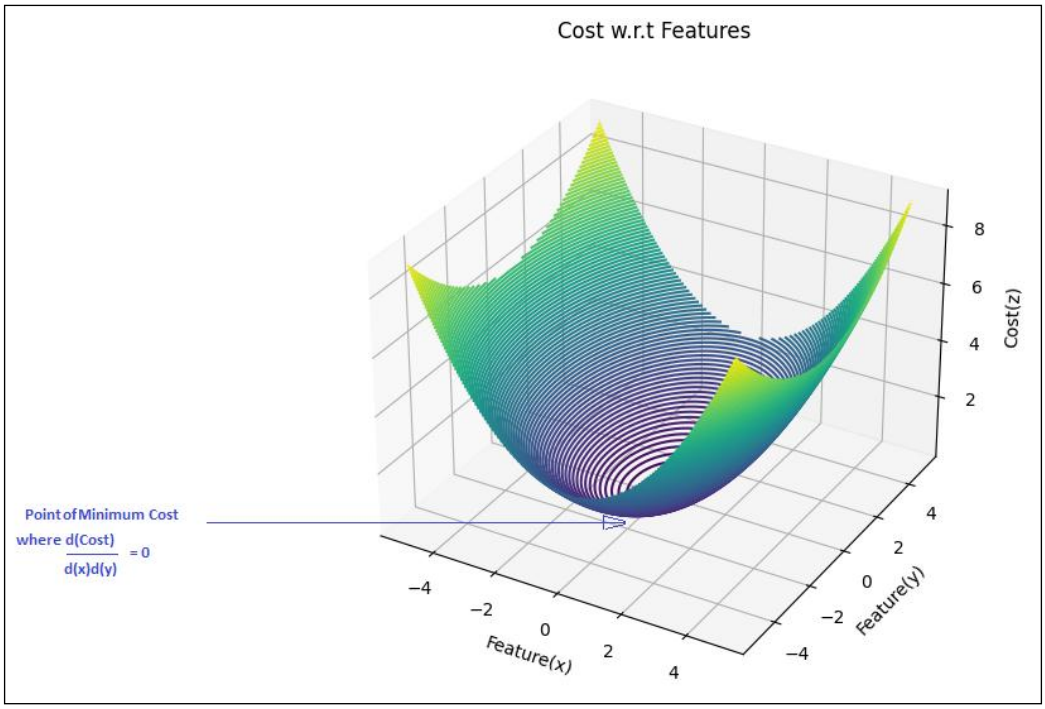

Fig. 7

$\nabla_a C$ (Stochastic Gradient Descent) is the amount of nudge needed on Weights and Bias to lower down the Cost. From "Differential Calculus" perspective, it is $\partial C/\partial w^l_{jk}$ (the rate of change of Cost w.r.t Weight) and $\partial C/\partial b^l_j$ (the rate of change of Cost w.r.t Bias), almost tending to 0, that decides the extent of such nudging. $\partial C/\partial w^l_{jk}$ and $\partial C/\partial b^l_j$ can be expressed in terms of the error δ^l_j (the rate of change of Cost w.r.t Output Neuron).

Error in the output layer, δ^L:

$$\delta^L_j = \frac{\partial C}{\partial a^L_j}\sigma'(z^L_j).$$

where $\partial C/\partial a^L_j$ measures how fast the cost is changing as a function of the j^{th} output activation

$\sigma'(z^L_j)$, measures how fast the activation function σ is changing at z^L_j

$$\delta^L = \nabla_a C \odot \sigma'(z^L)$$

Rate of change of the cost with respect to any weight in the network:

$$\frac{\partial C}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j$$

Rate of change of the cost with respect to any bias in the network:

$$\frac{\partial C}{\partial b^l_j} = \delta^l_j$$

So by knowing the Stochastic Gradient Descent, we can make necessary nudging on the weight and bias to have the predicted output very close to the expected output thereby reducing the overall cost to a minimum. The idea of such a nudge is to increase the accuracy of prediction and minimise the cost.

AI Model Overfitting

Think about incomplete, irrelevant and inconsistent data being fed to the AI model during training. This would result in "Overfitting" which means the line of best fit needs to be drawn to cater to most of the scattered data points. This is good for prediction but not for efficient training, as it takes into account the unwanted data points. Cleansed data from MDM would help avoid overfitting, as only relevant data is being accounted for.

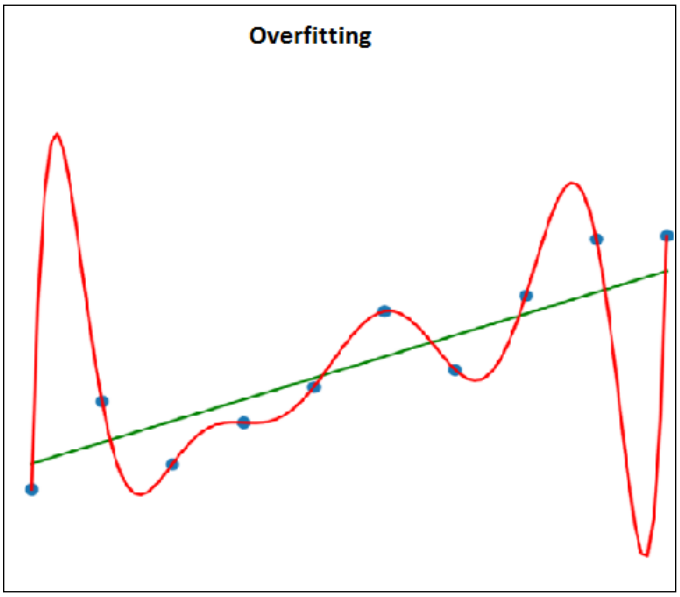

Fig. 8

AI Model Accuracy

It is noted that the AI model takes certain amount of time initially to achieve the state of maximum accuracy and minimum cost. As in the figure below the accuracy is reached at maximum and remains unchanged after 280 epochs (Nielson, 2015). Usually it takes several epoch-batch-test data runs before the accuracy reaches maximum. Validation data is specifically meant to find such a threshold. By feeding the model with cleansed data from MDM, such threshold can be met more quickly.

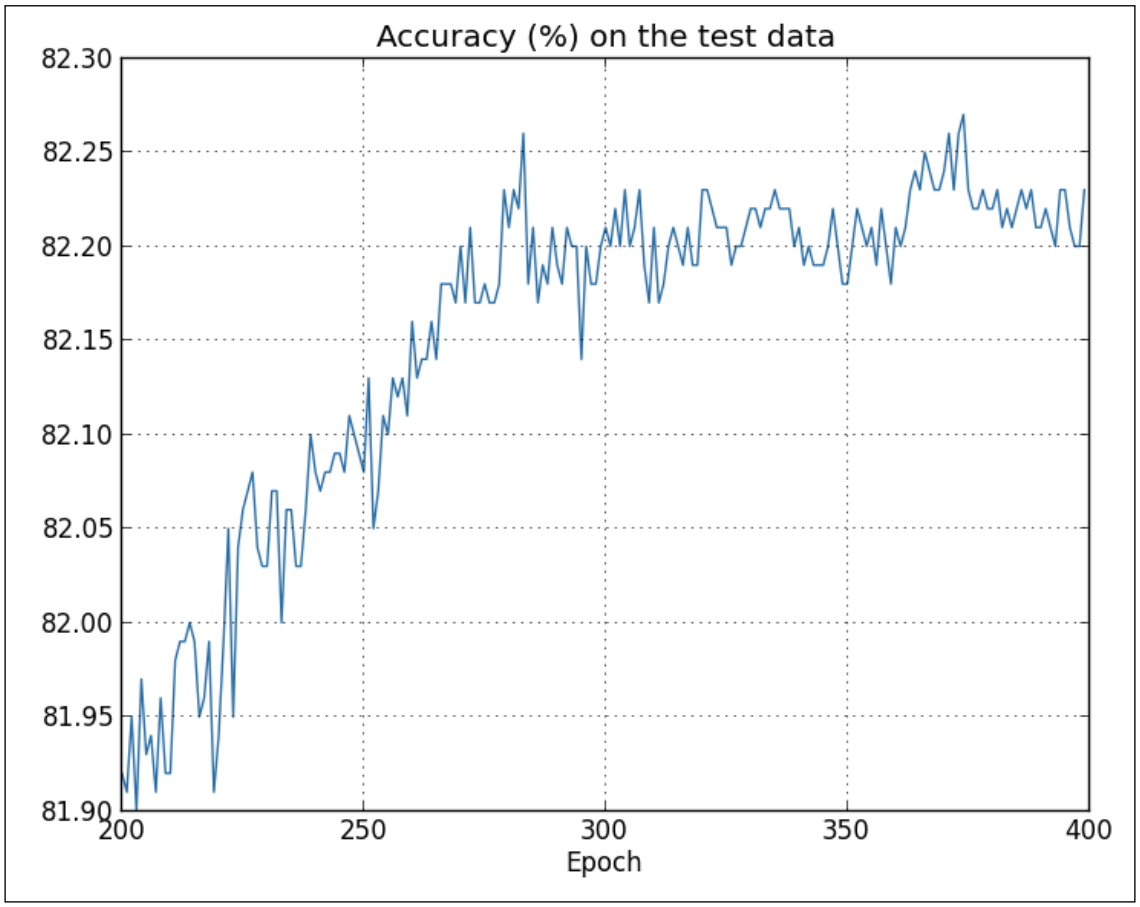

Fig. 9

Conclusion

With the demand for Artificial Intelligence (AI) increasing in the industry, it is time to start thinking about MDM feeding the cleansed data into the model while training. The more accurate, consistent and complete the data, the better would be the accuracy of prediction by the model. This would significantly improve the ROI from a business standpoint, as the effort would be channelised to win sales and increase revenue instead of spending days in identifying and correcting the faulty data and bringing it into the acceptable state.

References

Shaikh, A., Harreis, H., Machado, J., & Rowshankish, K. (2024). *Master data management: The key to getting more from your data* (pp. 1-9). McKinsey Digital. Retrieved from https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/master-data-management-the-key-to-getting-more-from-your-data

Oracle e-Business Suite. (n.d.). *Oracle customer data hub* (pp. 1-8). Oracle Corporation. [Online] Retrieved from https://www.oracle.com/a/ocom/docs/applications/ebusiness/oracle-customer-data-hub-data-sheet.pdf

Veeva Network. (n.d.). *Transforming data in Network > Transformation rules* (pp. 1-10). Veeva Systems Inc. [Online]. Retrieved from https://docs-vdm.veevanetwork.com/doc/vndocad/Content/Network_topics/Data_export/Transformation_rules.htm

https://semarchy.com/blog/how-to-measure-data-quality/

https://www.precisely.com/resource-center/productsheets/trillium-geolocation

https://www.vertexinc.com/solutions/products/vertex-o-series-address-cleansing

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L.,…Polosukhin, I. (2023). Attention is all you need. *arXiv preprint arXiv: 1706.03762v7,* pp. 1-15. Retrieved from https://arxiv.org/pdf/1706.03762

Nielson, M. (2015). *Neural networks and deep learning* (pp. 1-293.) Determination Press. Retrieved from http://neuralnetworksanddeeplearning.com/

Glossary

1. Master Data Management (MDM): Technology and Architecture for providing cleansed data that adheres to Data Integrity, Data Consistency, Data Validity, Data Quality and Data Governance.
2. Artificial Intelligence (AI): Computer Systems with sufficient capacity capable of performing tasks with intelligence including learning, problem solving, prediction and comprehension.
3. Single Source of Truth (SST): Repository of integrated, consistent, relevant and non-redundant data source.
4. Hub and Spoke System: Source systems working in synchronization with each other over the unified and accurate data published from MDM.
5. Mastered Data: Cleansed data from MDM after performing Duplicate Identification, Match and Merge, Validation, Data Quality Check.
6. Match and Merge: Recognize duplicates and prevent its creation in the repository. Resolves existing duplicates already sitting in the repository.