**Step-by-step Approach of Building Data Pipelines as a Data Scientist or a Machine Learning Engineer**

Learn how to build interesting pipelines in the field of machine learning to develop a system that can perform AI capabilities



Photo by [夜 咔罗](https://unsplash.com/@karosu?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

Oftentimes we are asked in either **interviews** or in our job roles as data scientists to build an application that is capable of performing machine learning predictions for **continuous** streaming data. There is often expectation from our boss that we are going to be delivering the results on time and generate these high-quality predictions with the use of machine learning and data science.

When looking at a large number of job descriptions state usually that to become a data scientist, one should be having **3+ years** of experience along with a few other things such as knowledge of SQL and Python. In addition, there is one more important component that is often being highlighted which is being able to build**data pipelines** and ensure that there are timely predictions. It is in these situations that a candidate is expected to have a strong understanding of building the data pipelines that are often the case to become a data scientist or a machine learning engineer.

**What is Data Pipeline?**



Photo by [Towfiqu barbhuiya](https://unsplash.com/@towfiqu999999?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

It is initially important to understand the definitions of a data pipeline before understanding ways in which we can successfully build and create a data pipeline. It is basically performing the **automation** of data and how it is actually processed and given to machine learning models during test time without human effort.

In other words, we would be creating a pipeline where the data that has been streaming would reach the pre-processing state after which, it would go with the machine learning model predictions with minimal effort.

In this article, we would be looking at the data pipelines and how to build them and ensure that we do a good job in building machine learning predictions. We will go in step-by-step about how to build the data pipelines that are important.

**Steps for Successful Data Science Pipeline**



Photo by [Volodymyr Hryshchenko](https://unsplash.com/@lunarts?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

**Understand the Business Constraints**

Before we build a data pipeline, we start off by asking some fundamental questions about the data and **its size**. Furthermore, we also look for **business constraints** such as whether there is a requirement for low-latency systems. If our business requires us to build low latency systems such as in internet applications, it is advisable to go ahead with simple ML models rather than relying on more complex models though they might be accurate in our cases. On the other hand, there can be other constraints such as model accuracy and it is important for the models to have higher accuracy. This is particularly true when we are using machine learning in the field of **healthcare** where the cost of misdiagnosing a patient suffering from a disease can have a significant impact. Therefore, the first step would be understanding the business constraints before trying to build interesting machine learning solutions.

**Data Collection**

Now that you have understood the requirements of the business and then have determined that machine learning and artificial intelligence can be the most useful according to the requirements, it is now time to collect the data that is important for ML predictions. Therefore, different departments can have access to different amounts and **variations** of the data that could be merged to create a unique dataset for our models to make predictions. Therefore, it would be a great practice to talk with different departments such as Sales teams and Data Science teams to get access to relevant data that could power your ML models. Therefore, we have explored the step of collecting the data that would be the most useful for our models for predictions.

**Data Pre-processing**

Now that the data is ready for use for ML applications, it is time to **preprocess** the data and make it easier for the computer (ML models) to understand them. Usually, data consists of a large number of missing values that do not make a significant impact on predictions. If you are considering the task of **natural language processing (NLP)**, there are often words that do not actually add a lot of value to the meaning of the text. These words are also called stop words such as ‘and’, and ‘or’ that do not have a lot of meaning to the text. Therefore, we perform the task of data preprocessing such as filling in the missing values or removing words such as stop words before giving it to our models.

If you are looking for ways to pre-process the data before feeding it to the machine learning models, feel free to explore my **earlier article** where I mention various ways in which we could perform feature engineering with the data. Below is the link.

[What Are the Most Important Preprocessing Steps in Machine Learning and Data Science? | by Suhas Maddali | Towards Data Science (medium.com)](https://medium.com/towards-data-science/what-are-the-most-important-preprocessing-steps-in-machine-learning-and-data-science-a7606d18f32a)

**Machine Learning Training**

After the data is processed and converted to a form that is more computer-friendly (machine learning-friendly, to be precise), the next step would be to**feed**this data to our ML models to make predictions. It is also important to note that we would have to divide the data into 2 parts: training and testing parts. This is because we would not want to evaluate the model for its performance on the training data itself. This is because since the model is trained on this data, it can be expected for it to perform extremely well on it. The only concern that we have is how will the ML model perform once it is put into**production**. Therefore, the real question that we must be asking before we deploy the model is how well it performs on the data that it has not seen before.

That is the case where the test set can be quite handy as this data could represent the data that the model might be facing in the**future**as well. However, there can be plenty of scenarios where this does not always hold. Assuming that the situation holds, we can then go ahead and train the model with the training data and test it with the data that we have split for testing. After trying out a large number of models along with hyperparameter tuning, we determine the best model that must be put into production. Note that though the performance of the model on the test set was excellent, we might not be putting that model sometimes to production due to business constraints such as low latency or other requirements depending on the nature of the business. But it is always a good idea to test the models and see how they are performing on the data that they have not seen and to get a general idea about their performance in real-time respectively.

**Model Deployment**

Once the data that we have is given to the models for predictions and we determine the best model to deploy in real-time, the next step that we could be taking is to **deploy**the product in real-time so that the end users get access to the impressive predictions that the model gives based on the history at which it was earlier trained. The parameters that the model learns during training are used to determine the output for real-time data that is easily accessible to the model respectively. While the machine learning predictions can seem impressive, failing to put the model into production can mean that we have been wasting quite our valuable time looking at a piece of technology that is impressive but was not able to provide any business value. Therefore, we would have to spend good amount of time ensuring that the best model that we have obtained is deployed in real-time to make a business impact.

**Data Monitoring**

Efforts have been put to deploy the best model in real-time based on the experiments that we have run on the test data. Now is the time that we constantly monitor the performance and regularly retrain the model if necessary after finding that it is not doing well after a few days with the change in events that cause the relationship between the input and the output to no longer exist as was earlier **determined** by our ML model.

Hence this is a stage where we try to maintain our product without **lowering** its quality in the long run. We can constantly monitor our data and see if there are any changes in relationships between the features or the distribution of the output respectively. If we find that there is a significant difference between the data that was used for model training, we again retrain the model to give better predictions with the present real-time data. If you are more interested to know how the data could change and the distribution might not always be the same, you can click on the link below that explains this phenomenon in great detail.

[Why is it Important to Constantly Monitor Machine Learning and Deep Learning Models after Production? | by Suhas Maddali | Towards Data Science (medium.com)](https://medium.com/towards-data-science/why-is-it-important-to-constantly-monitor-machine-learning-and-deep-learning-models-after-4136f6067547)

**Conclusion**

While going through this article, you might have gotten a good idea about the**importance**of building data pipelines and the list of steps that could be followed to effectively build robust pipelines and ensure that we are giving good value to the business from our machine learning models. Steps such as data monitoring must be taken to ensure that the quality of predictions is not reduced. This can be done with the help of retraining the model in cycles and ensuring that it is getting access to the most recent data. Thank you for taking the time to read this article.

If you like to get more updates about my latest articles and also have unlimited access to the medium articles for just **5 dollars per month**, feel free to use the **link** below to add your support for my work. Thanks.