# **Document AI**

## **Build and Run Two Document AI models**

In this section you will learn how to build and train a Document AI Model which will enable you to extract important structured content out of unstructured documents.

* Within the AI & ML area, click on **Document AI**:  
  alt text
* Click on the + Build icon to build a new model  
  alt text
* Under Build name, type in **ANALYST\_REPORTS**
* Choose the **DATAOPS\_EVENT\_PROD** for the database and **DOCUMENT\_AI** for the Schema

create build

Press **Create**

### **Downloading Documents**

We now need upload some documents into Document AI in order to build a train model. The first Model we will build are analyst reports. The reports we we will be using are completely fictional.

* **Download** The Analyst Reports [**here**](https://app.dataops.live/artifacts/df/a5/dfa50b41ffb5d6a1d36676f46f3e1fe3c586da6cab295ffbcda9a941466f45dd/2025_07_23/60142298/123186362/Homepage/dataops/event/homepage/site/downloads/fake_docs.zip)

# Uploading the Documents to the model

* Unzip the Analyst reports to a file location which you can access
* Open the model which you have just created and press upload documents

upload_docs

You will see that you can upload a variety of formats. Today we will be uploading PDF documents. You can either select the files by searching in your file system **or** drag and drop.

upload_docs

Press **Upload**

You will see that 4 documents are now uploaded but need attention. To use the model, you will first need to decide which fields you would like to extract. A warning will also let you know that for better accuracy you should really train the model with more documents. For this excercise we will just use 4 but in reality you will need to train the model with at least 10 documents for good accuracy.

### **Review Documents and Extract Fields**

* With in the Documents in Dataset section, press **See all documents**

You willl see that all 4 documents are in need of review

* Click on the first document in the list. This will take you straight to the Analyst Report. The right hand side of the screen you will need to add the fields that you wish to extract. We will be extracting specific fields.
* Press **+ Value** to add a new field
* Enter **DATE\_REPORT** in the first field and in the second field, as the question **When was the Report Created?**

Document AI will search through the multi page document and then will give you an answer with a probability score

upload_docs

* If you are happy with the answer, mark it with the **tick**. When the press the tick, the probability score will change to 1.00 - In otherwords, you have told document AI that you are 100% sure that this answer is correct.
* Create 4 additional fields with the corresponding questions - If document AI get's the answer right, mark with a tick. press **+ Value** to create each field.

| **Field** | **Question** |
| --- | --- |
| NAME\_OF\_REPORT\_PROVIDER | name of the report provider - not snowflake |
| RATING | What is the rating? is it BUY, SELL OR HOLD? |
| MARKET\_PRICE | Close Price Value |
| GROWTH | What is the revenue Growth - YoY |

If the answer is wrong, overtype the answer with the correct answer. This will automatically mark it as correct with the new answer.

Sometimes document AI comes up with no answer at all. If this is correct, still mark with a tick.

When you have finished entering the questions and getting the answers you should see something that looks like this:

upload_docs

Now it's time to evaluate the same questions and answers with another document.

* Click on the arrow to review on the bottom right of this pane the next document. Currently you are reviewing 1 of 4 documents.

The next document may populate more of the answers. You will need to simply mark them as to whether they are correct or not. If any are incorrect, type over with the correct answer.

* When you have finished reviewing 2 of 4 documents, continue to review the final 2 documents.

Once you have completed your document review press the back button on the top left to go back to the previous screen

upload_docs

All 4 documents should now be marked as **Accepted**.

* Go back to Build Details

You will see that there is a model accuracy. This accuracy figure will be due to how many answers document ai initially predicted correctly or not. If the accuracy score is low, you may wish to **Train model** this will train the model with the **corrected answers**.

* Press **Publish Version** which will publish your first version of the model. Today, we will not **train model** as this can take some time - especially if you are training large documents of multiple pages. In addition, the model is calculating a high model accuracy score of 0.97. Being concise with your questions can help the accuracy with little need for training.

Once complete, you will see a model version number and example queries on how to utilise the model in **SQL**

**create build**

You will have examples of these queries preloaded into a notebook. This model is iterative, if you find the results are not what you expect in terms of accuracy, simply add more documents and retrain. The more documents you train, the better results you will get. Every time you re-train or republish a new version, you will get a new version number. You will need this version number in the query. The screenshot above, shows that you are using the **model version 1**.

* Press the **back button** to return to the Document AI home screen.

### **Create a second model**

The Second model we will build are consistant infographic style reports. These are quartely publically available snowflake reports.

* **Download** the Infographics [**here**](https://app.dataops.live/artifacts/df/a5/dfa50b41ffb5d6a1d36676f46f3e1fe3c586da6cab295ffbcda9a941466f45dd/2025_07_23/60142298/123186362/Homepage/dataops/event/homepage/site/downloads/infographics.zip)

You will now create a new model - this time, we will call it **INFOGRAPHICS**

* Go back to the document AI home page and click on the **+ build** button to build a new model and populate the fields with the following:

Build name: **INFOGRAPHICS** Location:

* Database **DATAOPS\_EVENT\_PROD**
* Schema **DOCUMENT\_AI**

Press **Create**

Upload the previously downloaded **Infographic** files using the Upload documents button.

* Train the model with the following fields

| **Field** | **Question** |
| --- | --- |
| PROD\_REVENUE | What is the product revenue (choose Q if it specifies) |
| NET\_REVENUE\_RETENTION | What is the net revenue retention rate |
| TOTAL\_CUSTOMERS | TOTAL CUSTOMERS |
| TOTAL\_M\_CUSTOMERS | Number of $1 + Customers |
| GLOBAL\_2000\_CUSTOMERS | How many forbes global 2000 customers are there? |
| MARKETPLACE\_LISTINGS | how many marketplace listings are there? |
| DATE\_OF\_REPORT | What is the date of the report |
| QUARTER | What is the quarter of the report Q1, Q2, Q3 or Q4 |

* Add all the fields the same way as previously
* Remember to evaluate each field by pressing the tick when correct, or correcting manually if incorrect
* Remember to review ALL documents

You should finish with answers similar to this. Remember to correct answers which are incorrect.

create build

* Press **Publish Version** to publish a new version. Do not press train as this may take too much time for the lab. The accuracy score is also very high as the model has given the correct answers without training.

You will now have a second model which will allow you to extract the infographic information into a structured table.

You have now created two models to process documents.

### **Process Documents at Scale**

Let's now process some documents. We have more documents both for Analyst Reports and infographics in a stage. We can use our new models to extract structured information out of them. In addition, we will extract all the text out of the Analyst reports for search purposes.

For the next Document AI steps, you will use **Snowflake Notebooks** to process the documents as well as **visualise** the outputs.

#### **Model 1 - Analyst Reports**

* Go back to the home page and click on **Projects > Notebooks**
* Click on the **Document AI Analyst Reports** notebook
* Run through the notebook to process the **Analyst Reports**.

This notebook will perform the following:

* Extract the values defined in the **ANALYST\_REPORTS** model you have created from a directory of files residing in a **Snowflake Stage**
* Create a structured table of information
* Extract all remaining text using **Cortex Parse Document**

#### **Model 2 - Infographics**

You will run a second notebook which will also extract the structured text out of the **Infographic** image files stored in a **Snowflake Stage**.

* Go back to the Notebooks and open the **INFOGRAPHICS** notebook
* Run through the notebook to process the **Infographics** which will extract all values and store the results in a new structured table.