# Lab 2 - Log Categorization - Detect Unusual Log Entries

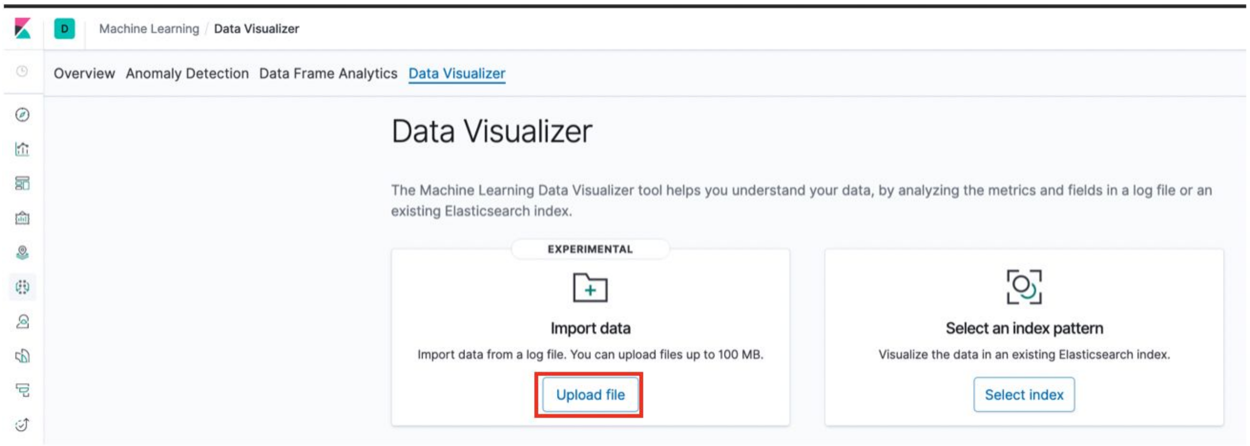
In this lab, we will be performing the following:

a. Import a sample log file into Kibana
b. Use Log Categorization to find unusual log entries

## A - Import Sample Log Data into Kibana

1. Download the sample log file from:
   https://drive.google.com/open?id=1EdXwMc0gtQFQFf47eDJODQGMAzu6iJzm

2. Go to Kibana > Machine Learning > Data Visualiser. Click on the “Upload file” link



Machine Learning / Data Visualizer

Overview Anomaly Detection Data Frame Analytics Data Visualizer

Data Visualizer

The Machine Learning Data Visualizer tool helps you understand your data, by analyzing the metrics and fields in a log file or an existing Elasticsearch index.

EXPERIMENTAL

Import data

Import data from a log file. You can upload files up to 100 MB.

Upload file

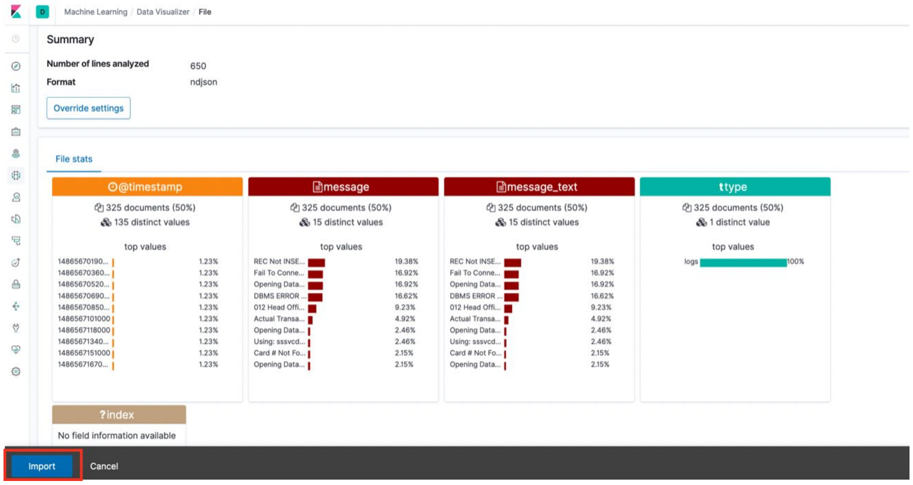
Select an index pattern

Visualize the data in an existing Elasticsearch index.

Select index

3. Upload the “it_ops_app_logs.json” file from Step 1 above.

4. Accept the default mapping and click on the “Import” button at the bottom



Machine Learning / Data Visualizer / File

Summary

| | |
|---|---|
| Number of lines analyzed | 650 |
| Format | ndjson |

Override settings

File stats

@timestamp
325 documents (50%)
135 distinct values
top values

| | |
|---|---|
| 14865670190... | 1.23% |
| 14865670360... | 1.23% |
| 14865670520... | 1.23% |
| 14865670690... | 1.23% |
| 14865670850... | 1.23% |
| 1486567101000 | 1.23% |
| 1486567118000 | 1.23% |
| 14865671340... | 1.23% |
| 1486567151000 | 1.23% |
| 14865671670... | 1.23% |

message
325 documents (50%)
15 distinct values
top values

| | |
|---|---|
| REC Not INSE... | 19.38% |
| Fail To Conne... | 16.92% |
| Opening Data... | 16.92% |
| DBMS ERROR ... | 16.62% |
| 012 Head Offi... | 9.23% |
| Actual Transa... | 4.92% |
| Opening Data... | 2.46% |
| Using: sssvcd... | 2.46% |
| Card # Not Fo... | 2.15% |
| Opening Data... | 2.15% |

message_text
325 documents (50%)
15 distinct values
top values

| | |
|---|---|
| REC Not INSE... | 19.38% |
| Fail To Conne... | 16.92% |
| Opening Data... | 16.92% |
| DBMS ERROR ... | 16.62% |
| 012 Head Offi... | 9.23% |
| Actual Transa... | 4.92% |
| Opening Data... | 2.46% |
| Using: sssvcd... | 2.46% |
| Card # Not Fo... | 2.15% |
| Opening Data... | 2.15% |

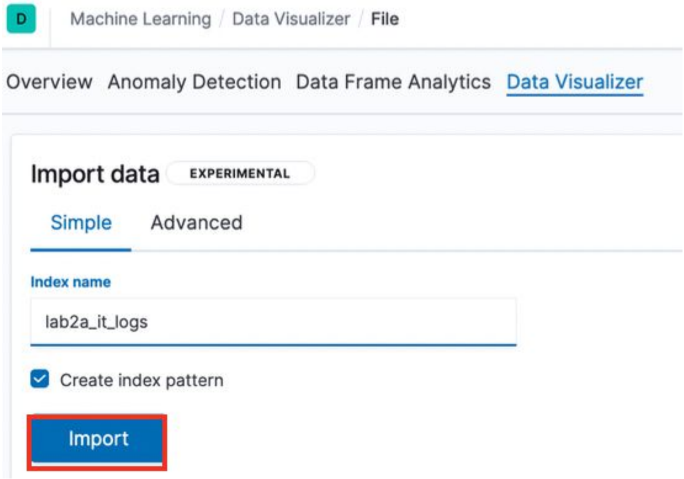
type
325 documents (50%)
1 distinct value
top values

| | |
|---|---|
| logs | 100% |

?index
No field information available

Import Cancel

5. Name the index “lab2a_it_logs” and click on the “Import” button

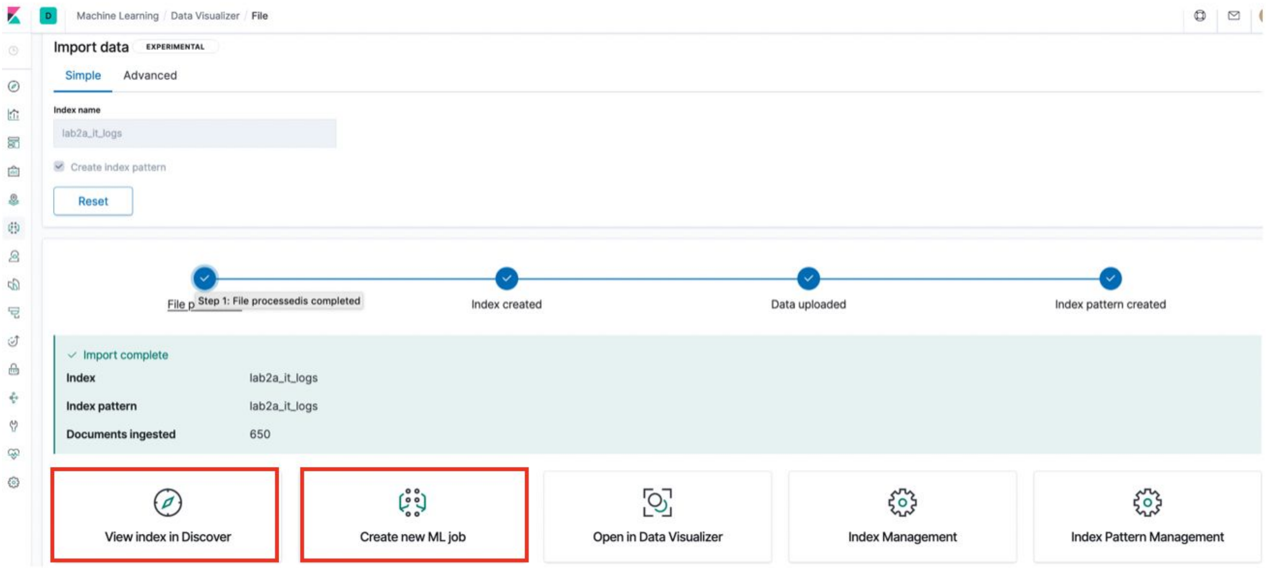


Index name
lab2a_it_logs

☑ Create index pattern

6. Once the import is done, click on the link to “View index in discover” (In the next steps we will also be creating a new ML job for this index)

   To speed things up, you might want to right-click on the links and open up different tabs for “Discover” and “ML”



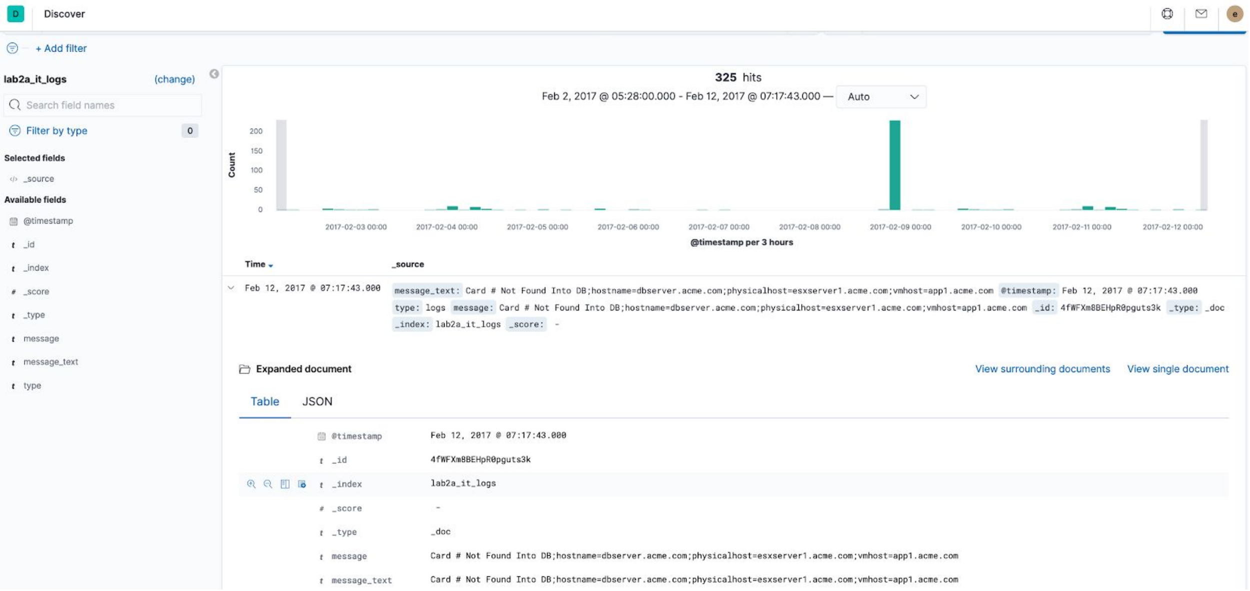
Index name
lab2a_it_logs

☑ Create index pattern

✓ Import complete

| | |
|---|---|
| Index | lab2a_it_logs |
| Index pattern | lab2a_it_logs |
| Documents ingested | 650 |

7. In Discover, note how the raw data looks like. Note that the "message" and "message_text" fields contain the log message entry



325 hits

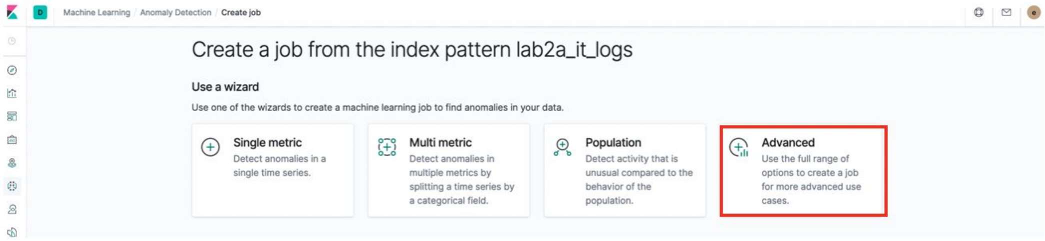
Feb 2, 2017 @ 05:28:00.000 - Feb 12, 2017 @ 07:17:43.000 —

| Time | _source |
|---|---|
| Feb 12, 2017 @ 07:17:43.000 | message_text: Card # Not Found Into DB;hostname=dbserver.acme.com;physicalhost=esxserver1.acme.com;vmhost=app1.acme.com @timestamp: Feb 12, 2017 @ 07:17:43.000 type: logs message: Card # Not Found Into DB;hostname=dbserver.acme.com;physicalhost=esxserver1.acme.com;vmhost=app1.acme.com _id: 4fWFXm8BEHpR0pguts3k _type: _doc _index: lab2a_it_logs _score: - |

Expanded document

| | |
|---|---|
| @timestamp | Feb 12, 2017 @ 07:17:43.000 |
| _id | 4fWFXm8BEHpR0pguts3k |
| _index | lab2a_it_logs |
| _score | - |
| _type | _doc |
| message | Card # Not Found Into DB;hostname=dbserver.acme.com;physicalhost=esxserver1.acme.com;vmhost=app1.acme.com |
| message_text | Card # Not Found Into DB;hostname=dbserver.acme.com;physicalhost=esxserver1.acme.com;vmhost=app1.acme.com |

8. Now, let's create a ML job using the index, to detect unusual log entries. Select the "Advanced" job link



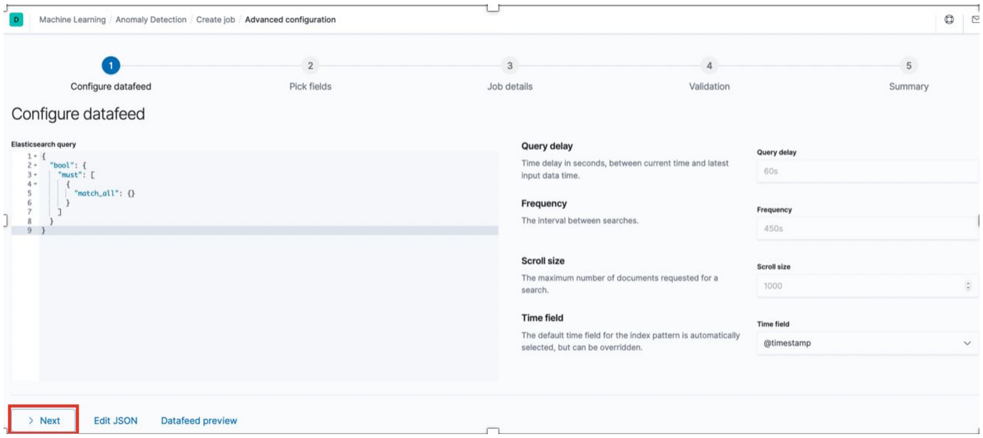
Create a job from the index pattern lab2a_it_logs

Use a wizard

Use one of the wizards to create a machine learning job to find anomalies in your data.

- Single metric: Detect anomalies in a single time series.
- Multi metric: Detect anomalies in multiple metrics by splitting a time series by a categorical field.
- Population: Detect activity that is unusual compared to the behavior of the population.
- Advanced: Use the full range of options to create a job for more advanced use cases.

9. Keep the default settings for the dataset (to use all available data in the index without filtering) and click on the next.



Configure datafeed

Elasticsearch query

```
{
  "bool": {
    "must": [
      {
        "match_all": {}
      }
    ]
  }
}
```

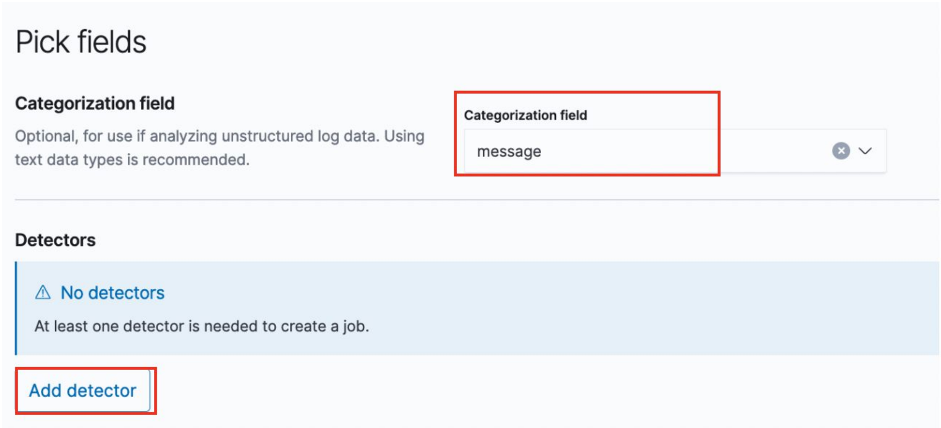
Query delay: Time delay in seconds, between current time and latest input data time. Query delay: 60s

Frequency: The interval between searches. Frequency: 450s

Scroll size: The maximum number of documents requested for a search. Scroll size: 1000

Time field: The default time field for the index pattern is automatically selected, but can be overridden. Time field: @timestamp

10. Select “message” under Categorization field and click on “Add Detector”



Pick fields

**Categorization field**

Optional, for use if analyzing unstructured log data. Using text data types is recommended.

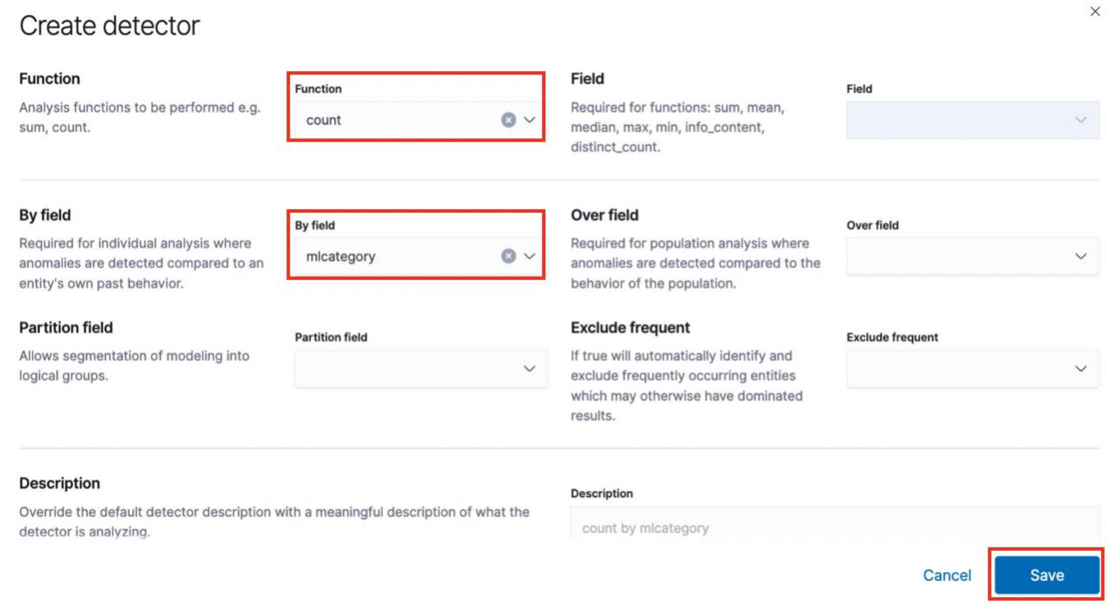
Categorization field: message

**Detectors**

⚠ No detectors

At least one detector is needed to create a job.

11. Set the detector to a count by field mlcategory, as below, and click on the “Save” button Click on “Next”



Create detector

**Function**

Analysis functions to be performed e.g. sum, count.

Function: count

**Field**

Required for functions: sum, mean, median, max, min, info_content, distinct_count.

Field:

**By field**

Required for individual analysis where anomalies are detected compared to an entity's own past behavior.

By field: mlcategory

**Over field**

Required for population analysis where anomalies are detected compared to the behavior of the population.

Over field:

**Partition field**

Allows segmentation of modeling into logical groups.

Partition field:

**Exclude frequent**

If true will automatically identify and exclude frequently occurring entities which may otherwise have dominated results.

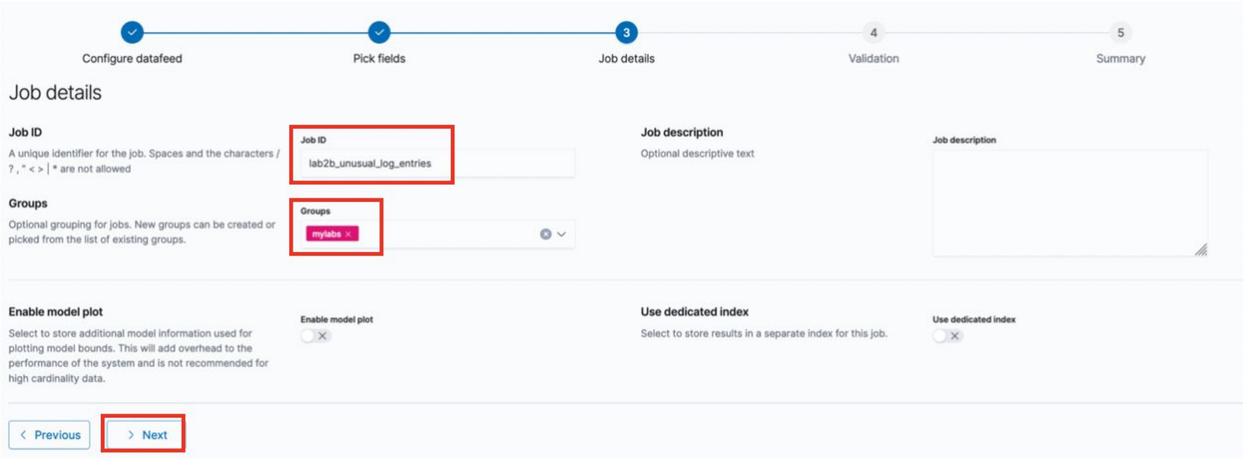
Exclude frequent:

**Description**

Override the default detector description with a meaningful description of what the detector is analyzing.

Description: count by mlcategory

12. Name the job “lab2b_unusual_log_entries”, place it in “mylabs” group and click “Next”



Job details

**Job ID**
A unique identifier for the job. Spaces and the characters / ? , " < > | * are not allowed

Job ID: lab2b_unusual_log_entries

**Groups**
Optional grouping for jobs. New groups can be created or picked from the list of existing groups.

Groups: mylabs

**Job description**
Optional descriptive text

Job description

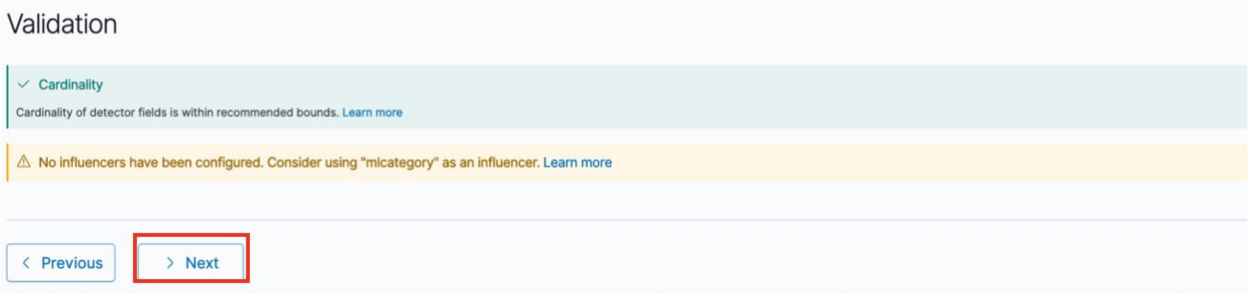
**Enable model plot**
Select to store additional model information used for plotting model bounds. This will add overhead to the performance of the system and is not recommended for high cardinality data.

Enable model plot

**Use dedicated index**
Select to store results in a separate index for this job.

Use dedicated index

13. Click “Next” to progress after Job Validation

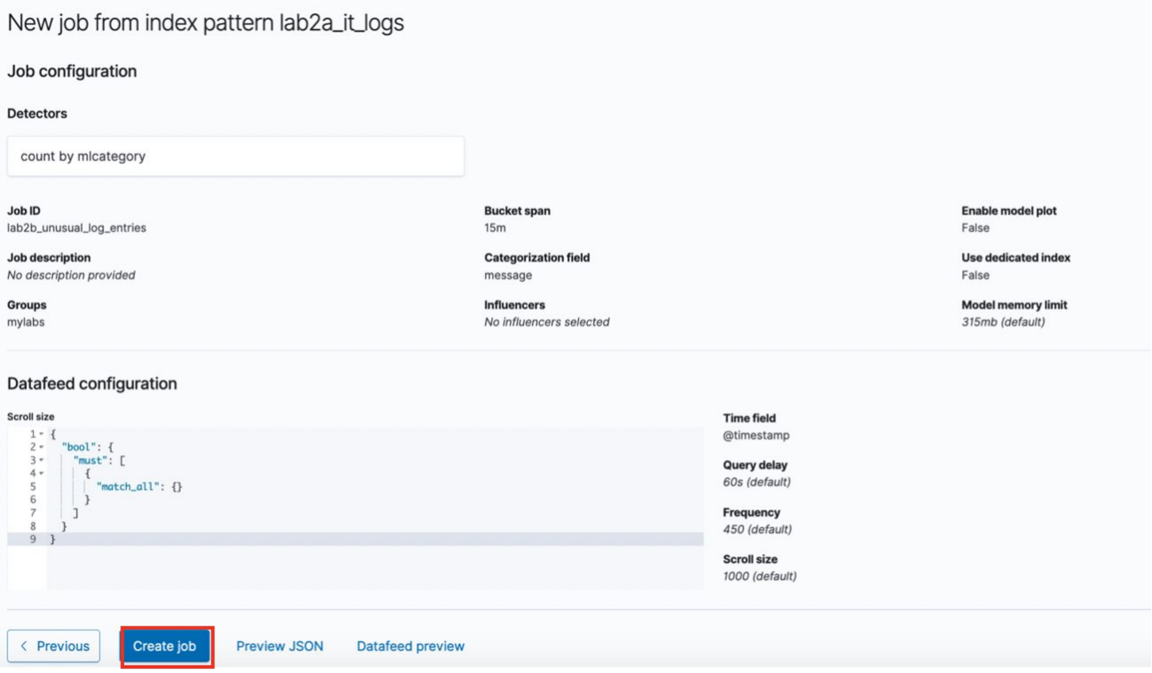


Validation

✓ Cardinality
Cardinality of detector fields is within recommended bounds. Learn more

⚠ No influencers have been configured. Consider using "mlcategory" as an influencer. Learn more

14. Review the final job configuration and click on “Create Job” to start the job



New job from index pattern lab2a_it_logs

**Job configuration**

**Detectors**

count by mlcategory

**Job ID**
lab2b_unusual_log_entries

**Bucket span**
15m

**Enable model plot**
False

**Job description**
*No description provided*

**Categorization field**
message

**Use dedicated index**
False

**Groups**
mylabs

**Influencers**
*No influencers selected*

**Model memory limit**
*315mb (default)*

**Datafeed configuration**

Scroll size

```
{
  "bool": {
    "must": [
      {
        "match_all": {}
      }
    ]
  }
}
```

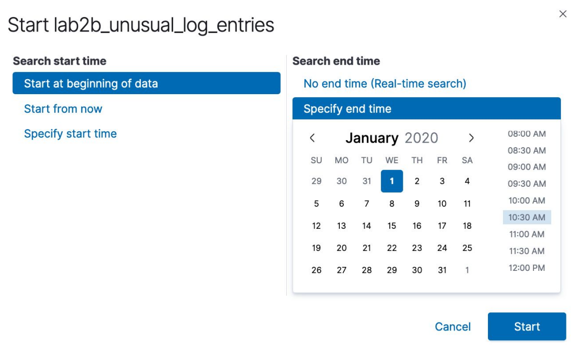
**Time field**
@timestamp

**Query delay**
*60s (default)*

**Frequency**
*450 (default)*

**Scroll size**
*1000 (default)*

15. Accept the default settings and click on "Start" to begin the ML job

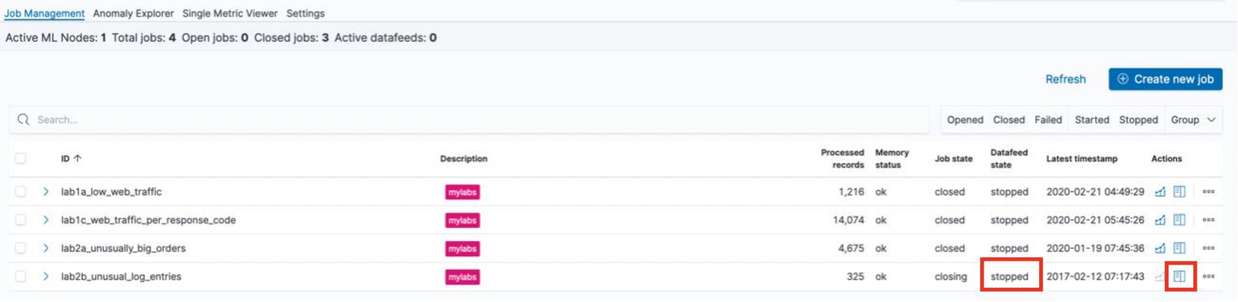


Start lab2b_unusual_log_entries

Search start time: Start at beginning of data

Search end time: Specify end time

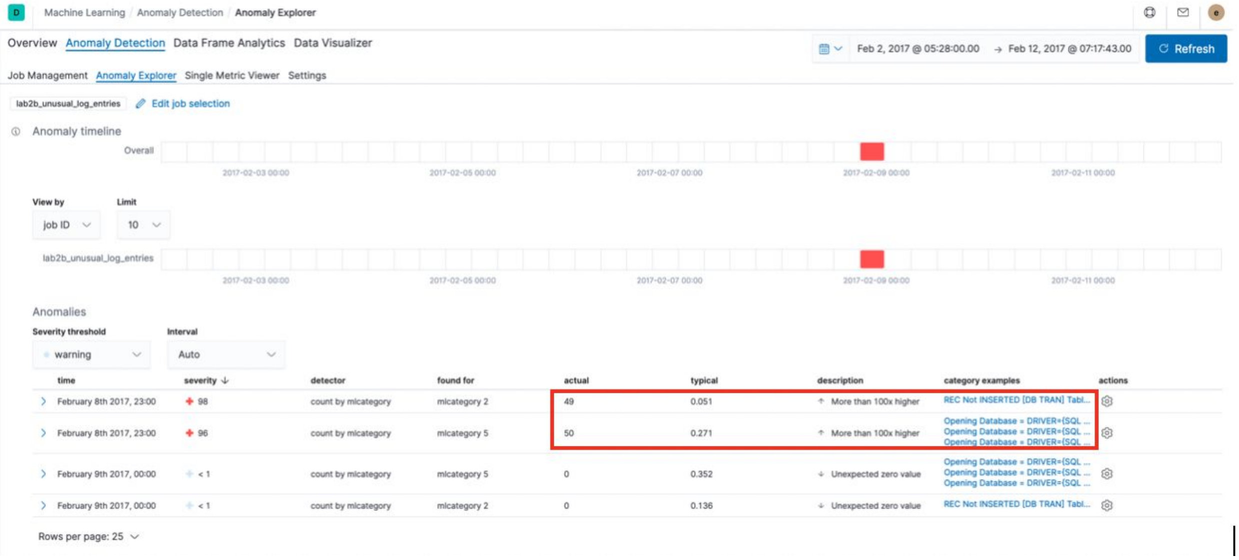
16. When the job has completed, click on the "Anomaly Viewer" to view the results



Active ML Nodes: **1** Total jobs: **4** Open jobs: **0** Closed jobs: **3** Active datafeeds: **0**

| ID ↑ | Description | Processed records | Memory status | Job state | Datafeed state | Latest timestamp |
|---|---|---|---|---|---|---|
| lab1a_low_web_traffic | mylabs | 1,216 | ok | closed | stopped | 2020-02-21 04:49:29 |
| lab1c_web_traffic_per_response_code | mylabs | 14,074 | ok | closed | stopped | 2020-02-21 05:45:26 |
| lab2a_unusually_big_orders | mylabs | 4,675 | ok | closed | stopped | 2020-01-19 07:45:36 |
| lab2b_unusual_log_entries | mylabs | 325 | ok | closing | stopped | 2017-02-12 07:17:43 |

17. We can see that the ML job flagged out abnormal log entries



Machine Learning / Anomaly Detection / Anomaly Explorer

Feb 2, 2017 @ 05:28:00.00 → Feb 12, 2017 @ 07:17:43.00

lab2b_unusual_log_entries

View by: job ID Limit: 10

Anomalies

Severity threshold: warning Interval: Auto

| time | severity ↓ | detector | found for | actual | typical | description | category examples |
|---|---|---|---|---|---|---|---|
| February 8th 2017, 23:00 | 98 | count by mlcategory | mlcategory 2 | 49 | 0.051 | ↑ More than 100x higher | REC Not INSERTED [DB TRAN] Tabl... |
| February 8th 2017, 23:00 | 96 | count by mlcategory | mlcategory 5 | 50 | 0.271 | ↑ More than 100x higher | Opening Database = DRIVER={SQL ...<br>Opening Database = DRIVER={SQL ...<br>Opening Database = DRIVER={SQL ... |
| February 9th 2017, 00:00 | < 1 | count by mlcategory | mlcategory 5 | 0 | 0.352 | ↓ Unexpected zero value | Opening Database = DRIVER={SQL ...<br>Opening Database = DRIVER={SQL ...<br>Opening Database = DRIVER={SQL ... |
| February 9th 2017, 00:00 | < 1 | count by mlcategory | mlcategory 2 | 0 | 0.136 | ↓ Unexpected zero value | REC Not INSERTED [DB TRAN] Tabl... |

Rows per page: 25

Typically these categories of log messages only appear once in the time bucket (default 15 mins) but the count went up to 49 & 50 during that time period. Hence ML has flagged that out as an anomaly.