



Machine Learning in the Elastic Stack

Elvis Saravia  (@omarsar0)

Education Engineer

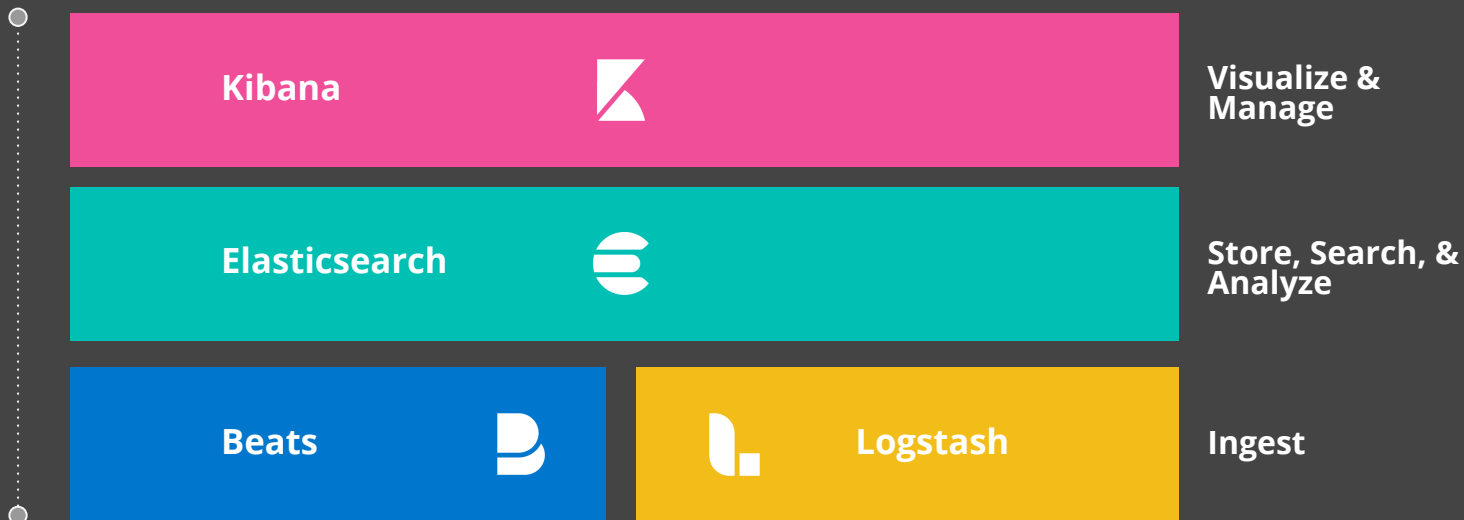
Virtual Meetup - April 2020

Slides: <https://ela.st/ml-meetup>



Elastic Stack

Search, Observe, and Protect



Machine Learning with the Elastic Stack

Machine Learning in the Elastic Stack

An Elasticsearch cluster can include the following ML capabilities:

- Data visualizer
- Data transformation
- Modeling
- Evaluation
- Visualization

Classification

**Regression
Analysis**

Supervised ML

**Outlier
Detection**

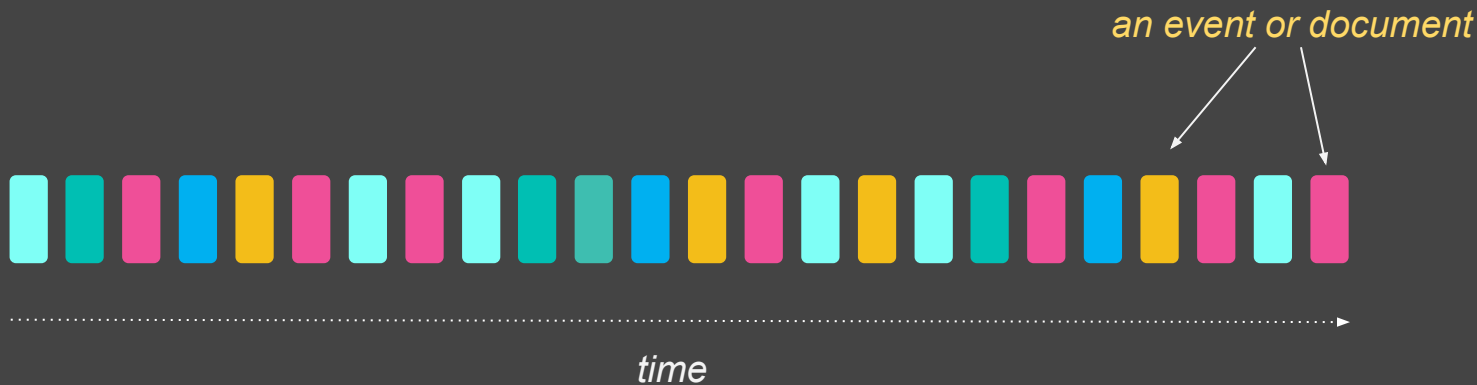
Unsupervised ML

Data Transformation

Event-Centric vs Entity-Centric Data

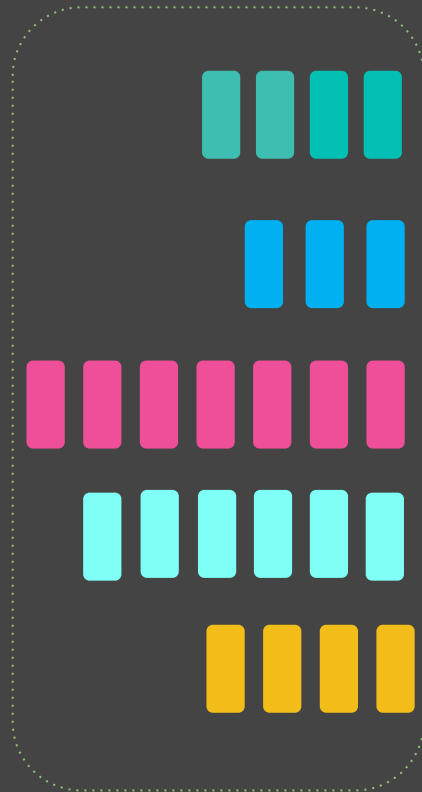
We typically store data as *event-centric*:

- tweets
- web apache logs
- network activity
- customer transactions



Aggregations

Using aggregations (metrics or buckets) we can summarize our data in different ways and obtain insights or create powerful features



Typical Aggregations



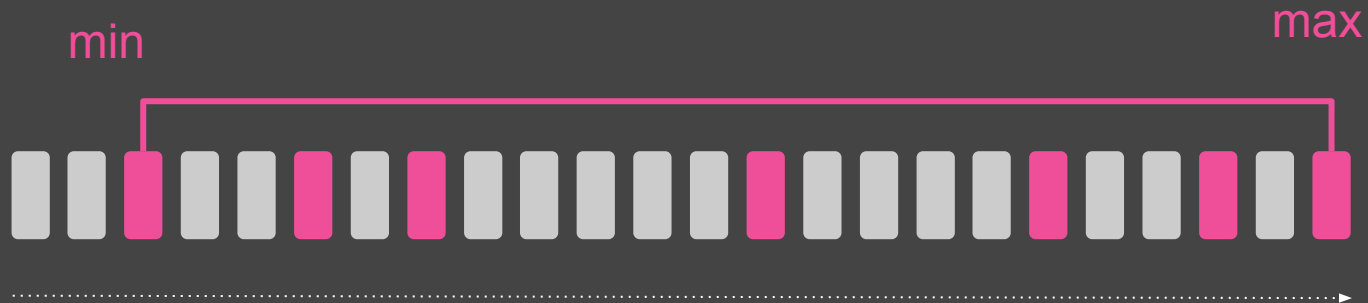
most frequent URLs?
blogs?
terms?
user?
agent?

terms aggregation

page views per minute?
published blogs per month?
published comments per day?

date histogram
aggregation

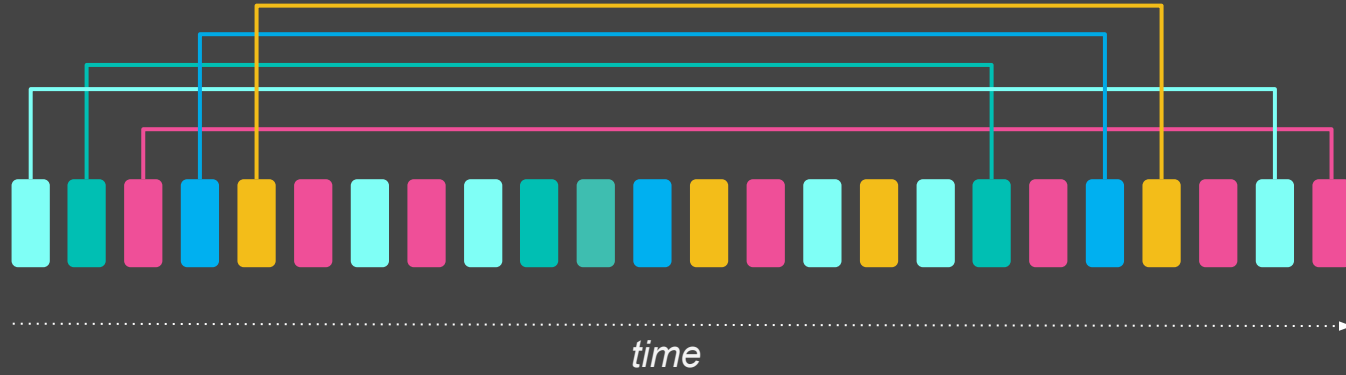
Clickstream data



how long was session x?

scripted aggregations
(**max** - **min**)

Clickstream data

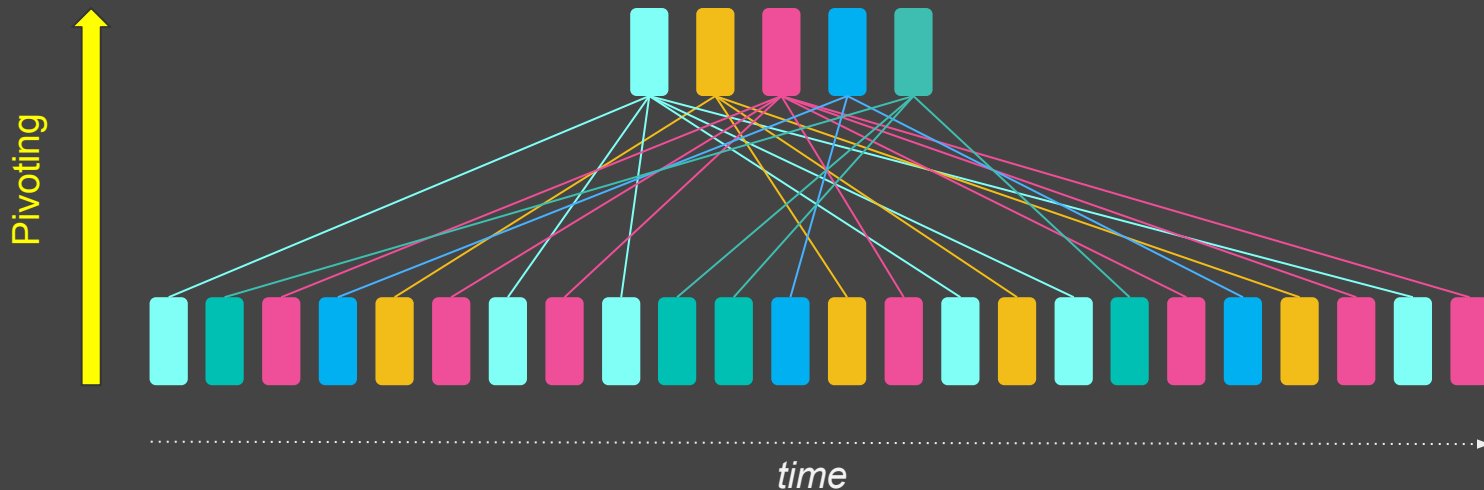


Average session duration?

We are doing behavioral
analytics!

Transform

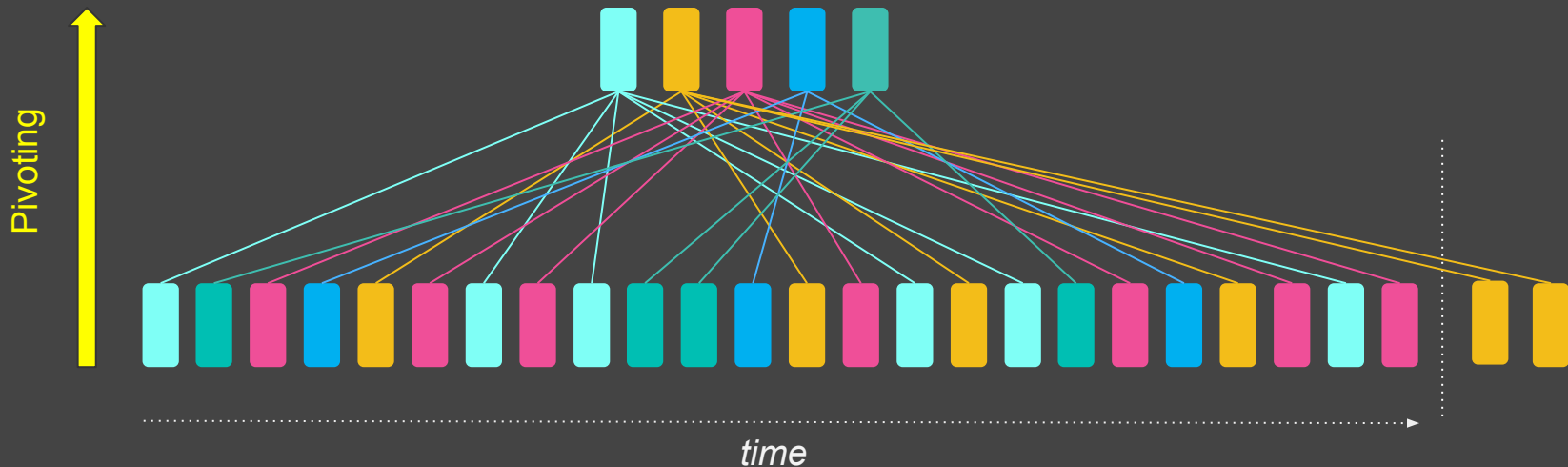
```
{  
  "session_id": "yellow",  
  "duration": 15.00  
  "count_hits": 4  
}
```



```
{  
  "session_id": "yellow",  
  "timestamp": "2019-30-09",  
  "url": "http://elastic.com/home"  
}
```

Continuous Pivot Transform

```
{  
  "session_id": "yellow",  
  "duration": 16.50  
  "count_hits": 6  
}
```



```
{  
  "session_id": "yellow",  
  "timestamp": "2019-30-09",  
  "url": "http://elastic.com/home"  
}
```

ML Modeling

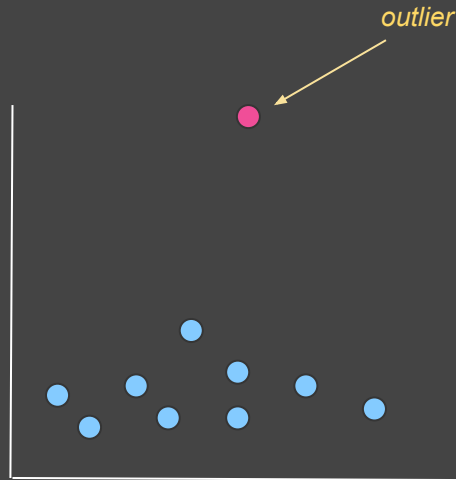
Data Frame Analytics

- Allows different analyses of the data to gather insights
- Allows to train a model and evaluate it:
 - Outlier Detection (*unsupervised*)
 - Regression Analysis (*supervised*)
 - Classification (*supervised*)



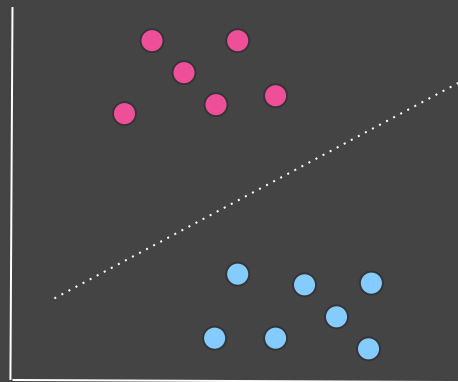
Outlier Detection

- The goal is to identify data points that do not follow the model of the data
- Uses an ensemble of distance and density based outlier detection methods
- Outputs an **outlier score** and the **feature influence score**
- Applications:
 - Bank fraud
 - Threat detection
 - Medical problems



Classification

- The goal is to predict the category/class of a given data point in a dataset
- Users a boosted tree regression model
- Requires **feature variables** and a **dependent variable**
- Applications:
 - Detect cancer
 - Classifying music or text
 - Predict loan risk



Demo

References

Demos

[Data Transforms Overview](#)

[Machine Learning Data Frame Analytics](#)

[Transforms API](#)

[Elastic Stack 7.3 Release](#)

[Introducing Transforms in Elastic Machine Learning](#)