Elastic Stack

Data Set – Denver Crime Reports

2013228419.0,20132284192202002,2202,0,**burglary-residence-by-force,burglary,2013-05-21 22:15:00,**2013-05-22 07:30:00,2013-05-22 08:29:00,**2276 S OSCEOLA ST,**3130211.0,1671262.0,**-105.0374051,39.6755022,**4,421,**harvey-park,**1,0

How do you make sense of 370719 Rows?

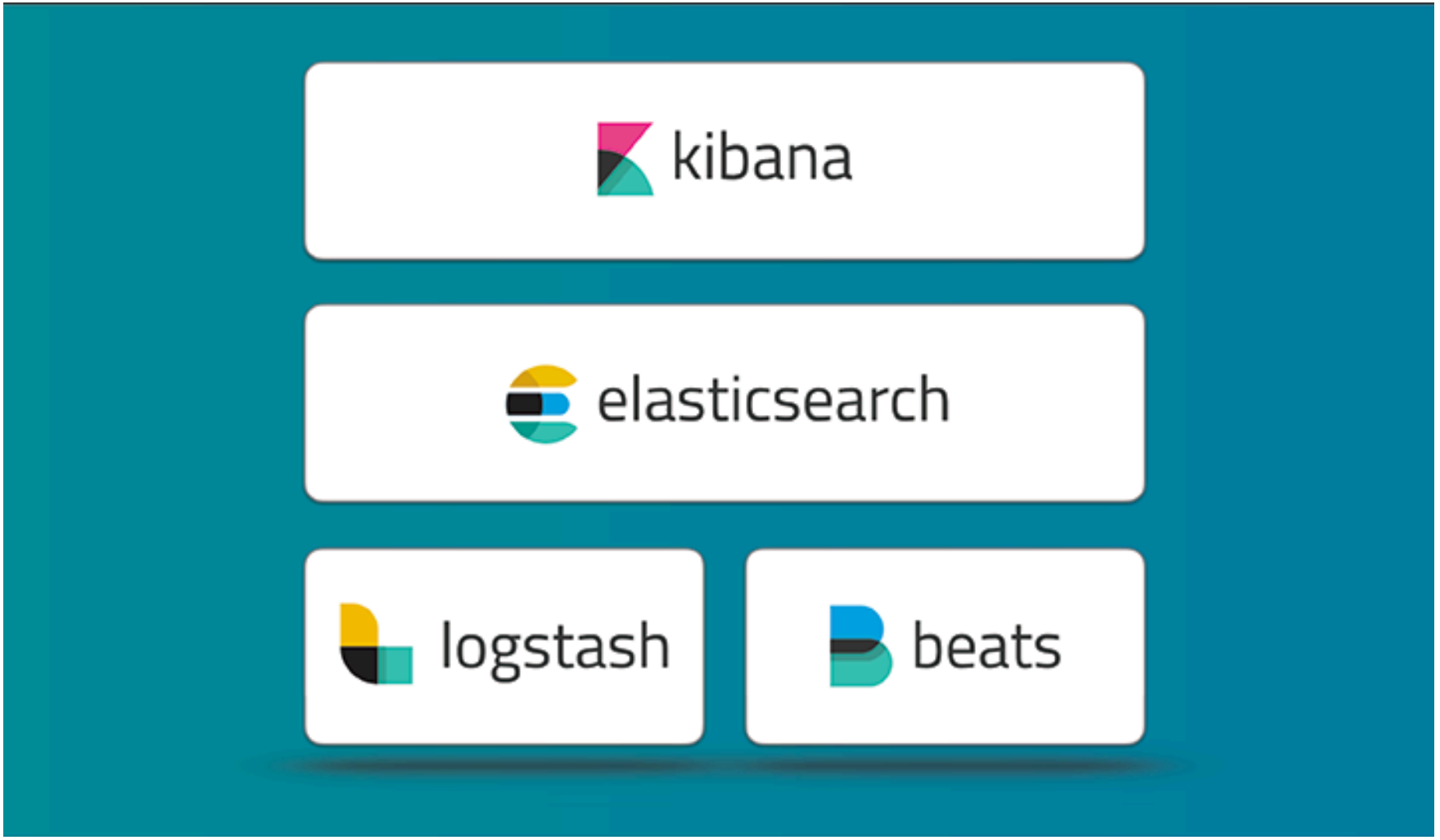
```
2013228419.0,2013228419220200,2202,0,burglary-residence-by-force,burglary,2013-05-21 22:15:00,2013-05-22 07:30:00,2013-05-22 08:29:00,2276 S
OSCEOLA ST,3130211.0,1671262.0,-105.0374051,39.6755022,4,421,harvey-park,1,0
201333126.0,201333126540100,5401,0,traffic-accident-hit-and-run,traffic-accident,2013-01-20 19:30:00,,2013-01-21 08:48:59,1111 N ASH ST,
3157814.0,1692426.0,-104.938879,39.733175,2,222,hale,0,1
2013216603.0,2013216603501600,5016,0,violation-of-restraining-order,all-other-crimes,2013-05-08 06:59:59,2013-05-14 23:00:00,2013-05-15
14:55:59,3460 N HUMBOLDT ST,3149133.0,1704190.0,-104.9695001,39.7656123,2,211,cole,1,0
2013216639.0,2013216639230500,2305,0,theft-items-from-vehicle,theft-from-motor-vehicle,2013-05-14 21:00:00,2013-05-15 09:59:59,2013-05-15
13:15:00,1100 BLK S TENNYSON ST,3128298.0,1678874.0,-105.0440655,39.6964255,4,421,mar-lee,1,0
201321664.0,2013216642604000,2604,0,fraud-identity-theft,white-collar-crime,2011-08-31 12:00:00,,2013-01-14 12:26:00,1650 S QUIETO CT,
3138580.0,1675380.0,-105.0075942,39.6866863,4,422,ruby-hill,1,0
2013216649.0,2013216649356200,3562,0,drug-marijuana-possess,drug-alcohol,2013-05-15 12:30:00,,2013-05-15 13:11:00,1735 N PONTIAC ST,
3166995.0,1696633.0,-104.9061383,39.7445637,2,223,south-park-hill,1,0
2013216662.0,2013216662299901,2999,1,criminal-mischief-mtr-veh,public-disorder,2013-05-14 20:00:00,2013-05-15 12:49:00,2013-05-15 12:49:00,5000
BLOCK N CLAY ST,3134854.0,1712116.0,-105.0201438,39.7875883,1,111,chaffee-park,1,0
201326982.0,201326982131300,1313,0,assault-simple,other-crimes-against-persons,2013-01-17 16:08:00,,2013-01-17 16:11:00,5559 N XANADU ST,
3188230.0,1715815.0,-104.8301227,39.7968152,5,512,montbello,1,0
201326982.0,201326982299900,2999,0,criminal-mischief-other,public-disorder,2013-01-17 16:08:00,,2013-01-17 16:11:00,5559 N XANADU ST,
3188230.0,1715815.0,-104.8301227,39.7968152,5,512,montbello,1,0
nreese-mbp-2015:denver_data nreese$ head -20 crime.csv
INCIDENT_ID,OFFENSE_ID,OFFENSE_CODE,OFFENSE_CODE_EXTENSION,OFFENSE_TYPE_ID,OFFENSE_CATEGORY_ID,FIRST_OCCURRENCE_DATE,LAST_OCCURRENCE_DATE,REPORTE
D_DATE,INCIDENT_ADDRESS,GEO_X,GEO_Y,GEO_LON,GEO_LAT,DISTRICT_ID,PRECINCT_ID,NEIGHBORHOOD_ID,IS_CRIME,IS_TRAFFIC
2013228419.0,2013228419220200,2202,0,burglary-residence-by-force,burglary,2013-05-21 22:15:00,2013-05-22 07:30:00,2013-05-22 08:29:00,2276 S
OSCEOLA ST,3130211.0,1671262.0,-105.0374051,39.6755022,4,421,harvey-park,1,0
201333126.0,201333126540100,5401,0,traffic-accident-hit-and-run,traffic-accident,2013-01-20 19:30:00,,2013-01-21 08:48:59,1111 N ASH ST,
3157814.0,1692426.0,-104.938879,39.733175,2,222,hale,0,1
2013216603.0,2013216603501600,5016,0,violation-of-restraining-order,all-other-crimes,2013-05-08 06:59:59,2013-05-14 23:00:00,2013-05-15
14:55:59,3460 N HUMBOLDT ST,3149133.0,1704190.0,-104.9695001,39.7656123,2,211,cole,1,0
2013216639.0,2013216639230500,2305,0,theft-items-from-vehicle,theft-from-motor-vehicle,2013-05-14 21:00:00,2013-05-15 09:59:59,2013-05-15
13:15:00,1100 BLK S TENNYSON ST,3128298.0,1678874.0,-105.0440655,39.6964255,4,421,mar-lee,1,0
201321664.0,2013216642604000,2604,0,fraud-identity-theft,white-collar-crime,2011-08-31 12:00:00,,2013-01-14 12:26:00,1650 S QUIETO CT,
3138580.0,1675380.0,-105.0075942,39.6866863,4,422,ruby-hill,1,0
2013216649.0,2013216649356200,3562,0,drug-marijuana-possess,drug-alcohol,2013-05-15 12:30:00,,2013-05-15 13:11:00,1735 N PONTIAC ST,
3166995.0,1696633.0,-104.9061383,39.7445637,2,223,south-park-hill,1,0
2013216662.0,2013216662299901,2999,1,criminal-mischief-mtr-veh,public-disorder,2013-05-14 20:00:00,2013-05-15 12:49:00,2013-05-15 12:49:00,5000
BLOCK N CLAY ST,3134854.0,1712116.0,-105.0201438,39.7875883,1,111,chaffee-park,1,0
201326982.0,201326982131300,1313,0,assault-simple,other-crimes-against-persons,2013-01-17 16:08:00,,2013-01-17 16:11:00,5559 N XANADU ST,
3188230.0,1715815.0,-104.8301227,39.7968152,5,512,montbello,1,0
201326982.0,201326982299900,2999,0,criminal-mischief-other,public-disorder,2013-01-17 16:08:00,,2013-01-17 16:11:00,5559 N XANADU ST,
3188230.0,1715815.0,-104.8301227,39.7968152,5,512,montbello,1,0
2013269894.0,2013269894131302,1313,2,assault-dv,other-crimes-against-persons,2013-06-14 20:00:00,2013-06-14 21:10:00,2013-06-14 21:17:59,9825 E
GIRARD AVE,3176185.0,1663877.0,-104.8742605,39.6544725,3,323,hampden,1,0
2013269901.0,2013269901131302,1313,2,assault-dv,other-crimes-against-persons,2013-06-14 22:30:00,2013-06-14 22:44:59,2013-06-15 01:08:00,S CLAY
ST / W WARREN AVE,3135000.0,1671755.0,-105.0203825,39.6767874,4,422,college-view-south-platte,1,0
2013269950.0,2013269950542000,5420,0,traffic-accident-dui-duid,traffic-accident,2013-06-15 00:24:59,,2013-06-15 02:06:59,7800 E HAMPDEN AVE,
3168969.0,1663071.0,-104.899907,39.6523935,3,323,hampden-south,0,1
2013373705.0,2013373705549900,5499,0,traf-other,all-other-crimes,2013-08-09 13:15:00,,2013-08-09 13:24:00,E 26TH AVE / WELTON ST,
3146752.0,1700233.0,-104.9780505,39.7547875,2,211,five-points,1,0
2013373750.0,2013373750521201,5212,1,weapon-by-prev-offender-powpo,all-other-crimes,2013-08-09 14:25:00,2013-08-09 14:29:59,2013-08-09
15:14:59,3300 BLOCK N DAHLIA ST,3159794.0,1703641.0,-104.9315896,39.7639282,2,221,northeast-park-hill,1,0
2013373764.0,2013373764544100,5441,0,traffic-accident,traffic-accident,2013-08-09 13:34:59,,2013-08-09 13:35:59,S FEDERAL BLVD / W MEXICO AVE,
3133667.0,1675062.0,-105.0250568,39.6858852,4,421,mar-lee,0,1
2013373768.0,2013373768540100,5401,0,traffic-accident-hit-and-run,traffic-accident,2013-08-09 12:30:00,,2013-08-09 14:39:59,1320 S FEDERAL BLVD,
3133874.0,1677374.0,-105.0242783,39.6922292,4,422,ruby-hill,0,1
2013373783.0,2013373783544100,5441,0,traffic-accident,traffic-accident,2013-08-09 13:51:59,,2013-08-09 13:51:59,LAWRENCE ST / 22ND ST,
```

How do you

- Ingest data
- Expose data web services
- Expose analytic web services
- Explore data
- Provide rich GUIs
- Scale to billions of records

All without writing a single line of code?

Welcome to the Elastic Stack



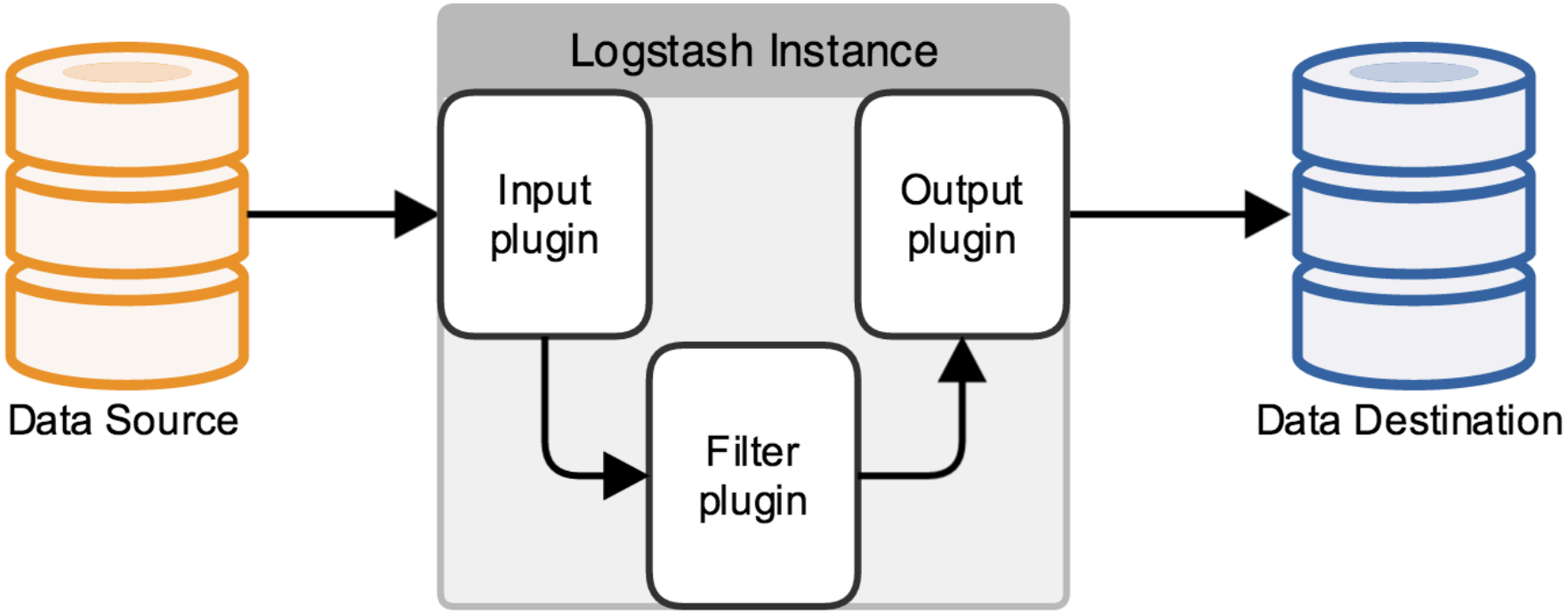
One more thing…

- ElasticStack is Open Source
- No
  - Licensing fees
  - Proprietary black holes

Logstash

Parse, Enrich & Transport Data

Logstash Processing Pipeline



Input: Read CSV file

```
1 input {
2   file {
3     path => "/Users/nreese/projects/denver_data/crime.csv"
4     start_position => "beginning"
5     sincedb_path => "/dev/null"
6     ignore_older => 0
7     type => "report"
8   }
9 }
```

Filter: CSV row -> JSON document

```
11 filter {
12   csv {
13     columns => [
14       "INCIDENT_ID", "OFFENSE_ID", "OFFENSE_CODE",
15       "OFFENSE_CODE_EXTENSION", "OFFENSE_TYPE_ID",
16       "OFFENSE_CATEGORY_ID", "FIRST_OCCURRENCE_DATE",
17       "LAST_OCCURRENCE_DATE", "REPORTED_DATE",
18       "INCIDENT_ADDRESS", "GEO_X", "GEO_Y",
19       "GEO_LON", "GEO_LAT", "DISTRICT_ID", "PRECINCT_ID",
20       "NEIGHBORHOOD_ID", "IS_CRIME", "IS_TRAFFIC"
21     ]
22     separator => ","
23   }
```

Filter: CSV row -> JSON document

```
24
25   mutate {
26     convert => { "GEO_LON" => "float" }
27     convert => { "GEO_LAT" => "float" }
28     convert => { "IS_CRIME" => "boolean" }
29     convert => { "IS_TRAFFIC" => "boolean" }
30   }
31
32   mutate {
33     add_field => {
34       "[location][lat]" => "%{GEO_LAT}"
35       "[location][lon]" => "%{GEO_LON}"
36     }
37     remove_field => ["GEO_X", "GEO_Y", "path", "host", "message"]
38   }
```

Filter: CSV row -> JSON document

```
39
40    ruby {
41      code =>
42        event['FIRST_OCCURRENCE_DATE'][10] = 'T'
43
44        hours = Integer(event['FIRST_OCCURRENCE_DATE'][11..12], 10)
45        mins = Integer(event['FIRST_OCCURRENCE_DATE'][14..15], 10)
46        secs = Integer(event['FIRST_OCCURRENCE_DATE'][17..18], 10)
47
48        event['secsOfDay'] = (hours * 3600) + (mins * 60) + secs
49    }
50  }
```

Output: insert into ElasticSearch

```
51
52 output {
53   elasticsearch {
54     index => "denver_crime"
55   }
56 }
```

ElasticSearch

Search and Analytics

Everything is a web service

- Health and Status
- Create/Read/Update/Delete
- Query
- Full text search
- **Aggregation**

Aggregation

- Metrics - compute metrics over document set
  - avg, cardinality, stats, geo_bounds, max, min, percentiles, sum, and more
- Bucket - segment document set
  - date_histogram, filters, geo_distance, geohash_grid, histogram, range, terms, and more
- Nested
  - Sub-aggregations are computed for each bucket from the parent aggregation.
  - No depth limit.

Which neighborhood has the most crime?

```
POST /denver_crime/report/_search HTTP/1.1
User-Agent: curl/7.37.1
Host: localhost:9200
Accept: */*
Content-Length: 180
{
  "size": 0,
  "aggs": {
    "mostCrime": {
      "terms": {
        "field": "NEIGHBORHOOD_ID”
      }
    }
  }
}
```

```
HTTP/1.1 200 OK
Content-Type: application/json
Content-Length: 1031
{
  "aggregations" : {
    ”mostCrime" : {
      "buckets" : [ {
        "key" : "five-points",
        "doc_count" : 17557
      }, {
        "key" : "stapleton",
        "doc_count" : 14168
      }, {
        "key" : "cbd",
        "doc_count" : 13896
      } ]
    } } }
```

When are the most accidents caused by DUIs?

```
{
  "query": {
    "match": {
      "OFFENSE_TYPE_ID": {
        "query": "traffic-
accident-dui-duid"
      }
    }
  },
  "size": 0,
  "aggs": {
    "mostDUIs": {
      "histogram": {
        "field": "secsOfDay",
        "interval": 3600
      }
    } } }
```

```
{
  "aggregations" : {
    "mostDUIs" : {
      "buckets" : [ {
        "key" : 0,
        "doc_count" : 380
      }, {
        "key" : 3600,
        "doc_count" : 409
      }, {
        "key" : 7200,
        "doc_count" : 525
      }, {
        "key" : 10800,
        "doc_count" : 175
      }, {
        "key" : 14400,
        "doc_count" : 68
      },
```

When are the most accidents caused by DUIs?

```
{
  "query": {
    "match": {
      "OFFENSE_TYPE_ID": {
        "query": "traffic-
accident-dui-duid"
      }
    }
  },
  "size": 0,
  "aggs": {
    "mostDUIs": {
      "histogram": {
        "script":
"doc['secsOfDay'].value / 3600",
        "interval”:1
       }
    }
  } }
```

```
{
  "aggregations" : {
    "mostDUIs" : {
      "buckets" : [{
        "key" : 0,
        "doc_count" : 380
      }, {
        "key" : 1,
        "doc_count" : 409
      }, {
        "key" : 2,
        "doc_count" : 525
      }, {
        "key" : 3,
        "doc_count" : 175
      }, {
        "key" : 4,
        "doc_count" : 68
      },
```

How do crime rates trend over time?

```
"aggs": {
  "topNeighborhoods": {
    "terms": {
      "field": "NEIGHBORHOOD_ID",
      "size": 5,
      "order": { "_count": "desc” }
    },
    "aggs": {
      "yearlyTrend": {
        "date_histogram": {
          "field": "FIRST_OCCURRENCE_DATE",
          "interval": "1y",
            "time_zone": "America/Denver”,
          }
        }
      }
    } }
```

```
"topNeighborhoods" : {
  "buckets" : [ {
    "key" : "five-points",
    "doc_count" : 17557,
    "yearlyTrend" : {
      "buckets" : [ {
        "key_as_string" : "2014-01-01T00:00:00.000",
        "doc_count" : 3911
      }, {
        "key_as_string" : "2015-01-01T00:00:00.000",
        "doc_count" : 4562
      } ]
    }
  }, {
    "key" : "stapleton",
    "doc_count" : 14168,
    "yearlyTrend" : {
      "buckets" : [ {
        "key_as_string" : "2014-01-01T00:00:00.000",
        "doc_count" : 3375
      }, {
        "key_as_string" : "2015-01-01T00:00:00.000",
        "doc_count" : 3580
      } ]
    }
```

Wouldn't it be cool if there was a **UI** that did all of this for us?

- Build JSON requests
- Graphically display results

Kibana

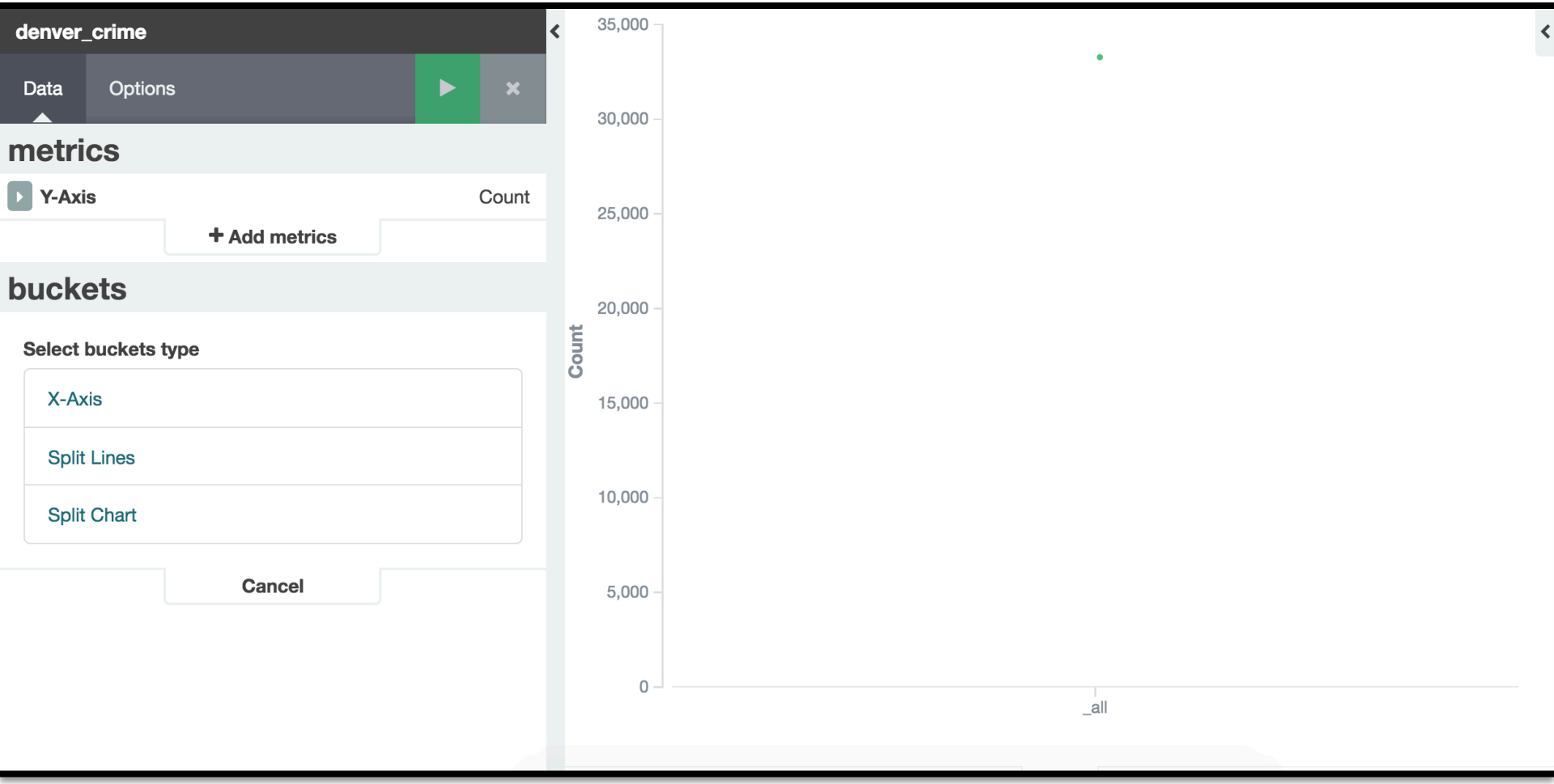
Visualization platform

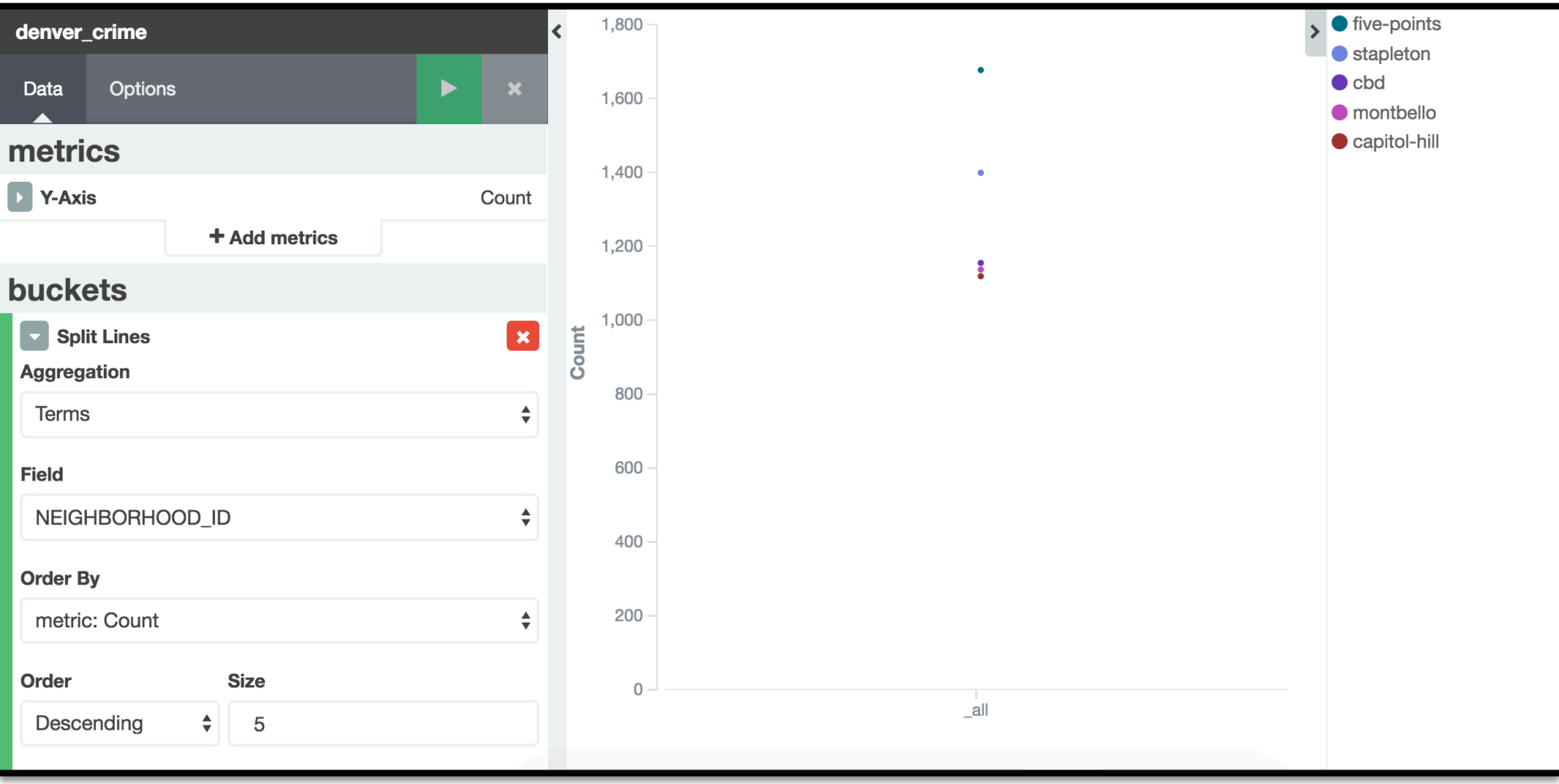
How would you like the results displayed?

Create a new visualization

Step 1

Area chart	Great for stacked timelines in which the total of all series is more important than comparing any two or more series. Less useful for assessing the relative change of unrelated data points as changes in a series lower down the stack will have a difficult to gauge effect on the series above it.
Data table	The data table provides a detailed breakdown, in tabular format, of the results of a composed aggregation. Tip, a data table is available from many other charts by clicking grey bar at the bottom of the chart.
Line chart	Often the best chart for high density time series. Great for comparing one series to another. Be careful with sparse sets as the connection between points can be misleading.
Markdown widget	Useful for displaying explanations or instructions for dashboards.
Metric	One big number for all of your one big number needs. Perfect for showing a count of hits, or the exact average a numeric field.
Pie chart	Pie charts are ideal for displaying the parts of some whole. For example, sales percentages by department.Pro Tip: Pie charts are best used sparingly, and with no more than 7 slices per pie.
Tile map	Your source for geographic maps. Requires an elasticsearch geo_point field. More specifically, a field that is mapped as type:geo_point with latitude and longitude coordinates.
Vertical bar chart	The goto chart for oh-so-many needs. Great for time and non-time data. Stacked or grouped, exact numbers or percentages. If you are not sure which chart your need, you could do worse than to start here.


denver_crime
Data
Options
metrics
Y-Axis
Count
+ Add metrics
buckets
Select buckets type
X-Axis
Split Lines
Split Chart
Cancel
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Count
_all

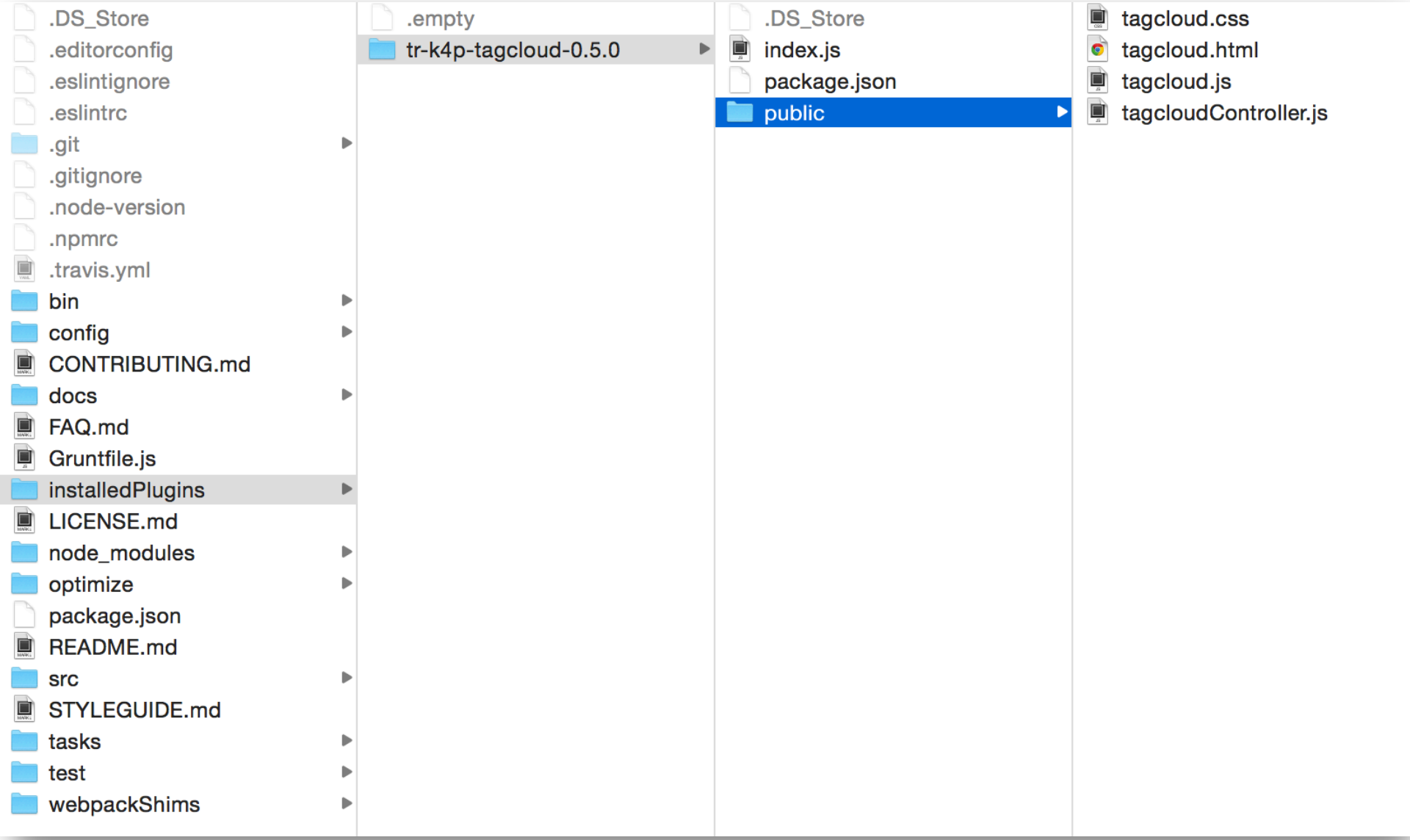

denver_crime
Data
Options
metrics
Y-Axis
Count
+ Add metrics
buckets
Split Lines
Aggregation
Terms
Field
NEIGHBORHOOD_ID
Order By
metric: Count
Order
Descending
Size
5
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
Count
_all
five-points
stapleton
cbd
montbello
capitol-hill


denver_crime
Data
Options
metrics
Y-Axis
Count
+ Add metrics
buckets
Split Lines NEIGHBORHOOD_ID: Descending
X-Axis
Sub Aggregation
Date Histogram
Field
FIRST_OCCURRENCE_DATE
Interval
Weekly
Advanced
Add sub-buckets
100
80
60
40
20
0
Count
2015-11-30
2015-12-31
2016-01-31
2016-02-29
2016-03-31
FIRST_OCCURRENCE_DATE per week
five-points
stapleton
cbd
montbello
capitol-hill

Write your own visualizations

```
module.exports = function(kibana) {
  return new kibana.Plugin({
    uiExports: {
      visTypes: ['plugins/tr-k4p-tagcloud/tagcloud']
    }
  });
};
```

```
<div ng-controller="TagcloudController" class="tagcloud">
  <span class="tag" ng-click="filter(tag)" ng-repeat="tag in tags" ng-
  style="{'font-size': tag.fontSize + 'px'}">{{tag.label}}</span>
</div>
```


.DS_Store
.editorconfig
.eslintignore
.eslintrc
.git
.gitignore
.node-version
.npmrc
.travis.yml
bin
config
CONTRIBUTING.md
docs
FAQ.md
Gruntfile.js
installedPlugins
LICENSE.md
node_modules
optimize
package.json
README.md
src
STYLEGUIDE.md
tasks
test
webpackShims
.empty
tr-k4p-tagcloud-0.5.0
.DS_Store
index.js
package.json
public
tagcloud.css
tagcloud.html
tagcloud.js
tagcloudController.js

Create a new visualization

Step 1

Area chart	Great for stacked timelines in which the total of all series is more important than comparing any two or more series. Less useful for assessing the relative change of unrelated data points as changes in a series lower down the stack will have a difficult to gauge effect on the series above it.
Data table	The data table provides a detailed breakdown, in tabular format, of the results of a composed aggregation. Tip, a data table is available from many other charts by clicking grey bar at the bottom of the chart.
Line chart	Often the best chart for high density time series. Great for comparing one series to another. Be careful with sparse sets as the connection between points can be misleading.
Markdown widget	Useful for displaying explanations or instructions for dashboards.
Metric	One big number for all of your one big number needs. Perfect for showing a count of hits, or the exact average a numeric field.
Pie chart	Pie charts are ideal for displaying the parts of some whole. For example, sales percentages by department.Pro Tip: Pie charts are best used sparingly, and with no more than 7 slices per pie.
Tagcloud	Tagcloud visualization
Tile map	Your source for geographic maps. Requires an elasticsearch geo_point field. More specifically, a field that is mapped as type:geo_point with latitude and longitude coordinates.
Vertical bar chart	The goto chart for oh-so-many needs. Great for time and non-time data. Stacked or grouped, exact numbers or percentages. If you are not sure which chart your need, you could do worse than to start here.

Summary

Try Elastic Stack

- Explore your data
- Rapidly prototype
  - Does data set even contain the answer?
- Save
  - Time
  - Money
  - Future maintenance

Expose all functionality through services

Jeff Bezos, Amazon CEO, ~2002

- All teams will henceforth expose their data and functionality through service interfaces.
- Teams must communicate with each other through these interfaces.
- There will be no other form of inter-process communication allowed: no direct linking, no direct reads of another team's data store, no shared-memory model, no back-doors whatsoever. The only communication allowed is via service interface calls over the network.
- It doesn't matter what technology they use.
- All service interfaces, without exception, must be designed from the ground up to be externalizable. That is to say, the team must plan and design to be able to expose the interface to developers in the outside world. No exceptions.

Anyone who doesn't do this will be fired. Thank you; have a nice day!

Resources

- https://github.com/FJbob/denver_data
- https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations.html
- https://www.timroes.de/2015/12/06/writing-kibana-4-plugins-visualizations-using-data/
- https://qbox.io/blog/elasticsearch-aggregations
- http://apievangelist.com/2012/01/12/the-secret-to-amazons-success-internal-apis/