



# Machine Learning avec Elastic Stack

David THIBAU – 2025

david.thibau@gmail.com

# Agenda

- **Introduction**

- Rappels Elastic Stack
- Machine Learning pour l'IT
- Assets ELK-ML
- Vues Kibana

- **Détection d'anomalies**

- Single et multi-metric jobs
- Autres jobs
- API et jobs avancés
- Optimisations
- Analyse de cause
- Alertes
- Prévisions

- **Analyse de trame de données**

- Introduction
- Détection de valeurs aberrantes
- Régression
- Classification
- Inférence

- **Annexes**

- URLs personnalisés
- Les différentes visualisations pour le ML

# Introduction

## **Rappels Elastic Stack**

Machine Learning pour l'IT

Assets ELK-ML

Vues Kibana

# Introduction

- **Elastic Stack** : Suite d'outils facilitant l'analyse et la visualisation temps-réel de données volumineuses
- Offre OpenSource et Commerciale de la société *Elastic*

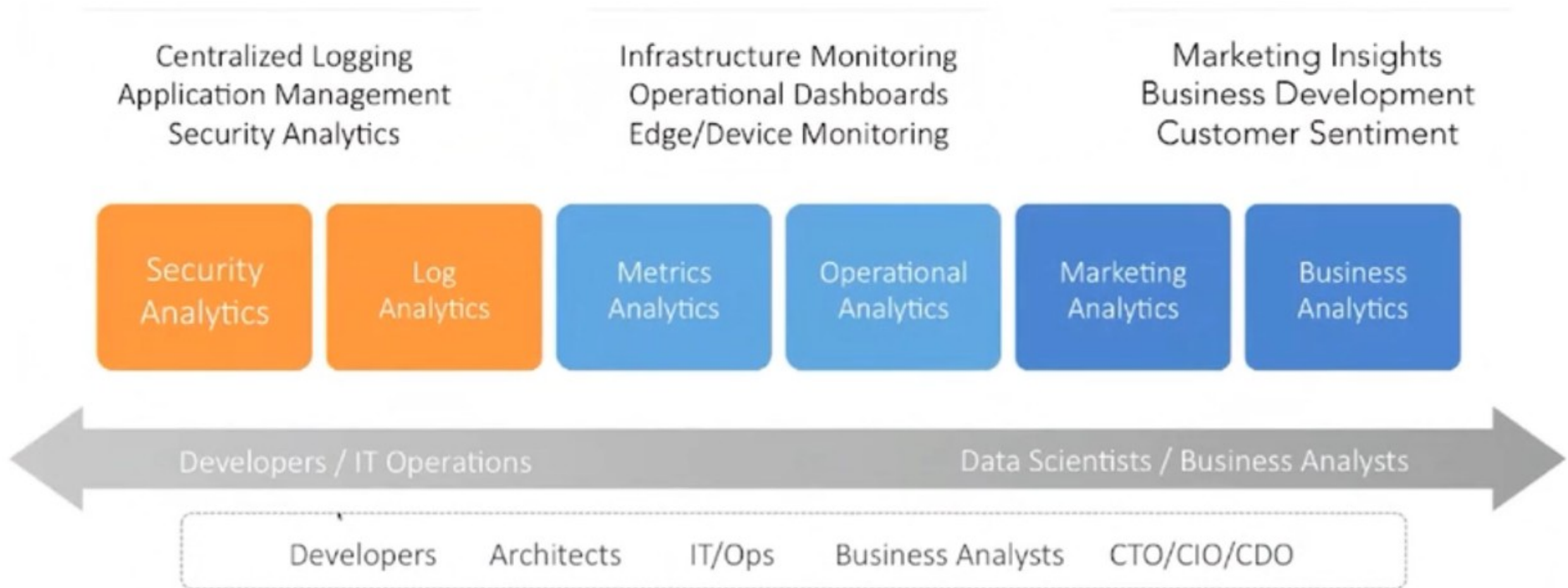
# La pile

- La suite est composé des outils suivants :
  - ***Integration / Elastic Agent / Beats*** : Différentes techniques pour ingérer des données en continu
  - ***ES Pipelines / Logstash*** : Moyens pour transformer, enrichir, filtrer les données sources
  - ***Elasticsearch*** : Base documentaire NoSQL intégrant différentes techniques de recherche (SQL~like, Full-text, Langage naturel, Vecteur) et d'agrégations
  - ***Kibana*** : UI End-user personnalisable pour le métier, les administrateurs et les développeurs

# Offre connexe

- ***X-Pack*** : Inclus dans la distribution, certaines fonctionnalités doivent être activées avec une licence commerciale ou d'évaluation
  - Sécurité (Libre à partir de 6.3+)
  - Alerting, Reporting, Machine Learning
- ***Elastic Cloud / Enterprise*** : ELK As A Service,
- Solutions « prêtes » à l'emploi :
  - ***Elastic Enterprise Search*** : Moteur de recherche, Site Web,
  - ***Elastic Observability*** : Stack pour surveiller un SI : Logs, APM, ...
  - ***Elastic Security*** : Protéger un SI contre les menaces

# Use cases de la suite





# Cluster ElasticSearch

- Un **cluster** est un ensemble de serveurs (nœuds) qui contient l'intégralité des données et offre des capacités de recherche sur les différents nœuds
    - Il est identifié par son nom unique sur le réseau local (par défaut : "*elasticsearch*").
- => Un cluster peut n'être constitué que d'un seul nœud
- => Un nœud ne peut pas appartenir à 2 clusters distincts

# Spécialisation des nœuds

- Tous les nœuds d'un cluster se connaissent mutuellement et peuvent rediriger des requêtes HTTP vers le nœud approprié.  
Il est possible de spécialiser les nœuds afin de contrôler les ressources allouées à chaque fonction d'ES
- Les différents types de nœuds sont :
  - Nœuds pouvant être **maître**
  - Nœuds de **données** : Détient les données et effectue les tâches d'indexation et de recherche
  - Nœuds **d'ingestion** : Exécute les pipelines d'ingestion
  - Nœuds de **coordination** : Nœuds acceptant les requêtes et redirigeant vers les nœuds appropriés
  - Nœuds **cross-cluster** : Nœuds pouvant effectuer des recherches vers plusieurs cluster.
  - Nœuds **ML** : Dédiés à l'exécution des jobs de Machine Learning

# Index

- Un **index** est une collection de documents qui ont des caractéristiques similaires
  - Par exemple un index pour les données client, un autre pour le catalogue produits et encore un autre pour les commandes
- Un index est identifié par un nom (en minuscule)
  - Le nom est utilisé pour les opérations de mise à jour ou de recherche
- Dans un cluster, on peut définir autant d'index que l'on veut

# Sharding et réplication

- Afin de pouvoir scaler en volume et augmenter les performances, ELS permet de découper un index en shards
  - Chaque *shard* est un index indépendant hébergé sur un des nœuds contenant un sous-ensemble des données
- Afin de tolérer des défaillances et des instabilités, les shards peuvent être répliqués

# Document

- Un **document** est l'unité basique d'information stocké par les shards.
- Il est constitué d'un ensemble de champs typés
- Il peut être représenté avec JSON

# Data Stream

- Un flux de données ou ***data stream*** stocke des évènements sur plusieurs index tout en offrant une seule ressource nommée pour les requêtes d'indexation ou de recherche
  - Les flux de données sont bien adaptés aux données générées en continu.
- Le data stream achemine automatiquement les requêtes d'indexation vers le bon index.
- Des stratégies d'ILM (*Index Lifecycle Management*) permettent de préciser le cycle de vie des données.

# Introduction

L'offre Elastic Stack  
**Machine Learning pour l'IT**  
Assets ELK-ML  
Vues Kibana

# Introduction

- L'offre de ML d'Elastic même si elle peut être appliquée à de nombreux domaines se concentre sur la surveillance IT :
  - Surveillance : Détection de panne, de dysfonctionnement
  - Sécurité : Détection d'intrusion, d'exfiltration
- Dans la surveillance IT, le constat est :  
*On génère beaucoup de données mais on a pas le temps de les regarder et analyser*  
=> Automatisation de ces analyses par des systèmes pouvant apprendre seul



# Anomalies

- La majorité des exigences de surveillance sont des variations sur le thème :  
***Trouver quelque chose qui est différent de l'ordinaire***
- Les jobs d'analyse tentent alors de détecter des anomalies/changements par rapport à l'observation de l'historique  
=> Cela nécessite que l'historique soit assez volumineux
- Exemples :
  - Des messages d'erreur qui apparaissent soudainement dans un fichier journal
  - Une baisse soudaine du nombre de commandes traitées par un système en ligne
  - Un nombre excessif soudain de tentatives d'accès à quelque chose (authentification par brute force)

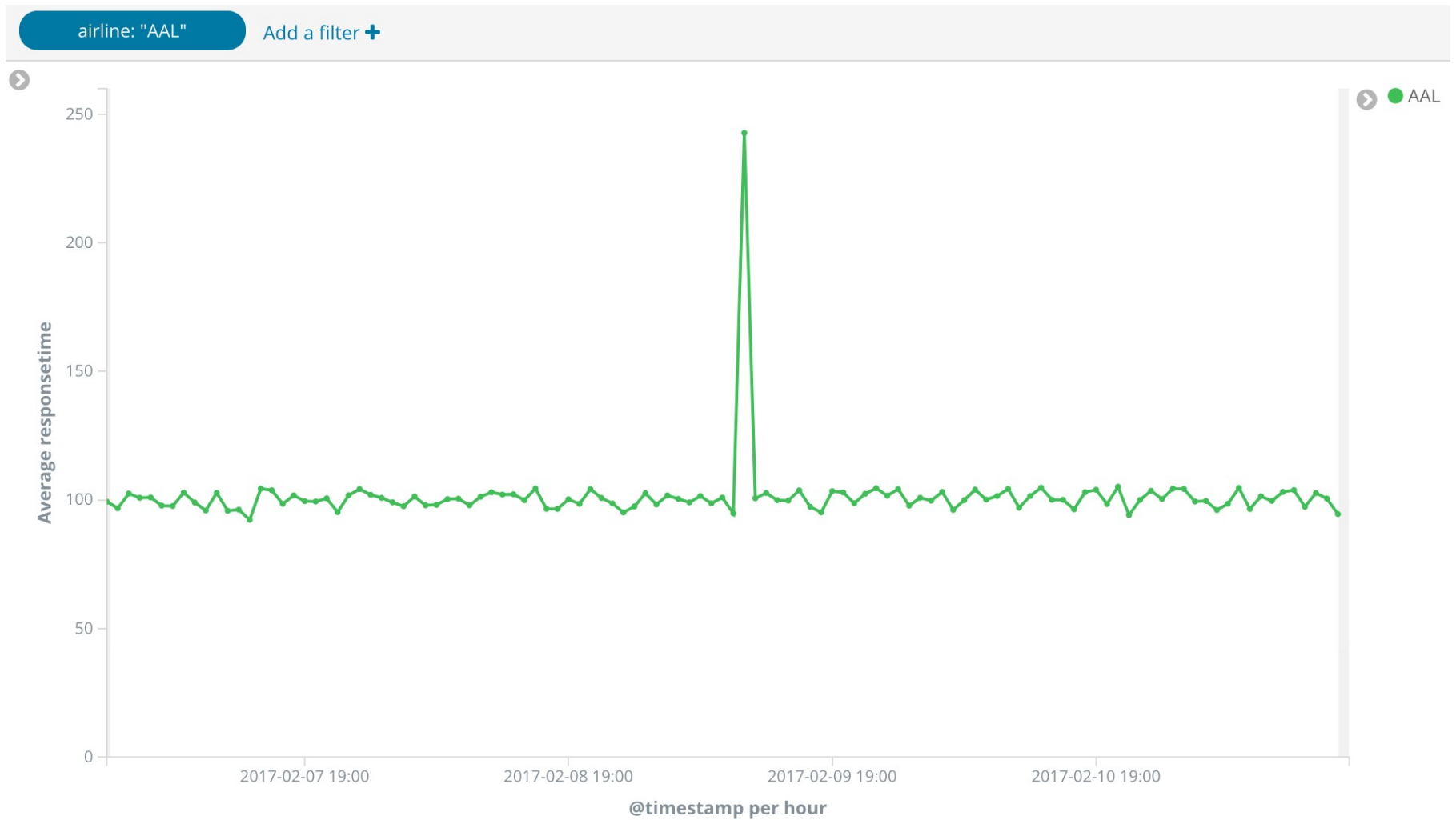
# Objectifs des solutions ML

- **Rapidité**: Notification d'une panne, d'une brèche ou de toute autre anomalie significative pro-activement et le plus rapidement possible afin de l'atténuer.
- **Scalability**: Les algorithmes doivent pouvoir scaler linéairement avec les données.
- **Performance** : Utiliser du matériel modeste plutôt que des super-ordinateurs
- **Applicabilité**: Prise en compte de la diversité des données dans les environnements informatiques.
- **Adaptabilité**: Les environnements informatiques en constante évolution peuvent rapidement rendre fragile un algorithme statique.
- **Précision**: Ne pas générer de fausses alarmes

# Qu'est-ce qu'une anomalie ?

- 2 définitions d'une anomalie :
  - Quelque chose est inhabituel si son comportement a dévié de manière significative d'un modèle établi basé sur son histoire passée
  - Quelque chose est inhabituel si certaines de ses caractéristique sont significativement différentes des mêmes caractéristiques des autres membres d'un ensemble ou d'une population

# Ex : Histoire passée



# Ex : Population



# Apprentissage non supervisé

- En ML, l'apprentissage non supervisé ne nécessite aucune intervention humaine pour la mise en place et l'affinement du modèle.
  - => Le modèle se construit, s'affine, se modifie en fonction des données qui lui sont présentées
  - => Plus il y a de données, plus le modèle se précise
- Elastic Stack propose 2 types d'analyse **non supervisé**
  - La détection d'anomalie : Basée sur des données temporelles
  - Détection de valeurs aberrantes : Basée sur des densité de valeurs

# Apprentissage supervisé

- Les apprentissages supervisés nécessitent des données d'entraînement.
- ELK-ML propose de types d'analyse supervisés :
  - **Classification** : Classification d'évènements  
Ex : Requêtes normales ou malicieuses
  - **Régression** : Prédiction de valeurs numériques  
Ex : Temps de réponse pour une requête Web.

# Identification du modèle

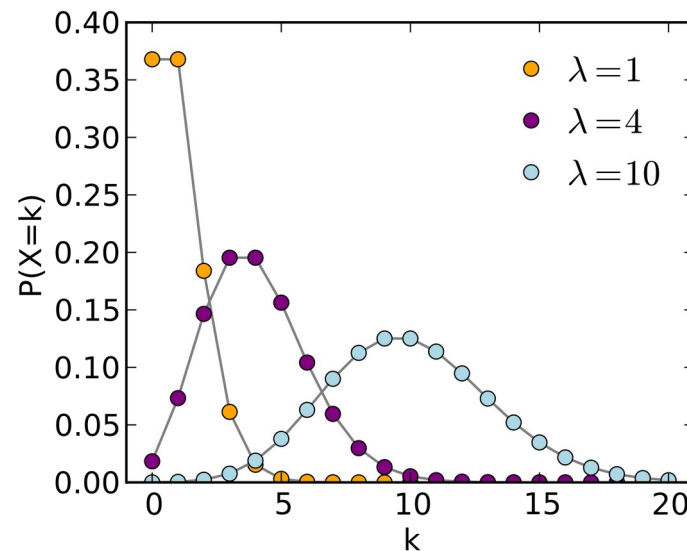
- Les processus de ML sélectionne à partir de leurs observations le modèle statistique approprié (poisson, gaussien, log-normal, etc.) et ses coefficients
- Des techniques bayésiennes sont utilisées pour évaluer les probabilités des valeurs du modèle compte tenu de l'ensemble de données observées.
- La modélisation effectuée est continue, de sorte que de nouvelles informations sont considérées avec les anciennes, avec une pondération exponentielle de l'information plus fraîche
- La modélisation peut être arrêtée puis redémarrée plus tard, d'où la nécessité de persister le modèle.



# Ex : Distribution de poisson

- Par exemple, la distribution de poisson permet de modéliser des événements tels que
  - Le nombre de météorites de plus d'un mètre de diamètre qui frappent la Terre chaque année
  - Le nombre de patients arrivant dans une salle d'urgence entre 10h00 et 23h00
  - Le nombre de photons frappant un détecteur dans un intervalle de temps particulier

$$p(k) = \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

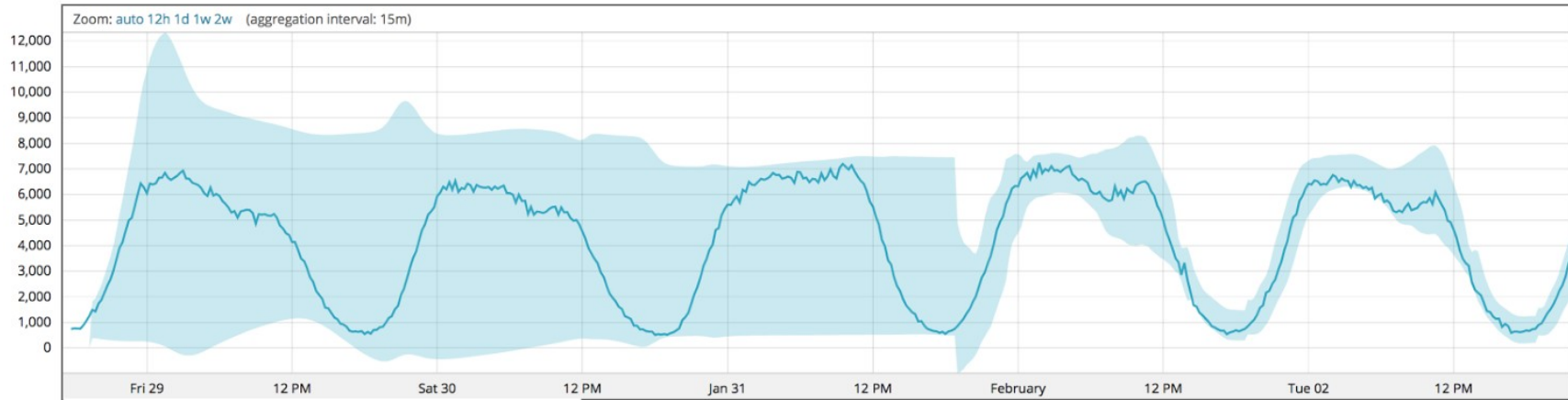


Si un certain type d'événements se produit en moyenne 4 fois par minute, pour étudier le nombre d'événements se produisant dans un laps de temps de 10 minutes, on choisit comme modèle une loi de Poisson de paramètre  $\lambda = 10 \times 4 = 40$

# Identification des Cycles

- Un autre aspect important de la modélisation de données réelles est de prendre en compte les tendances harmoniques importantes qui se produisent naturellement.
- ELK-ML recherche automatiquement les tendances marquantes dans les données (croissance linéaire, harmoniques cycliques, etc.) et les factorise

# Exemple



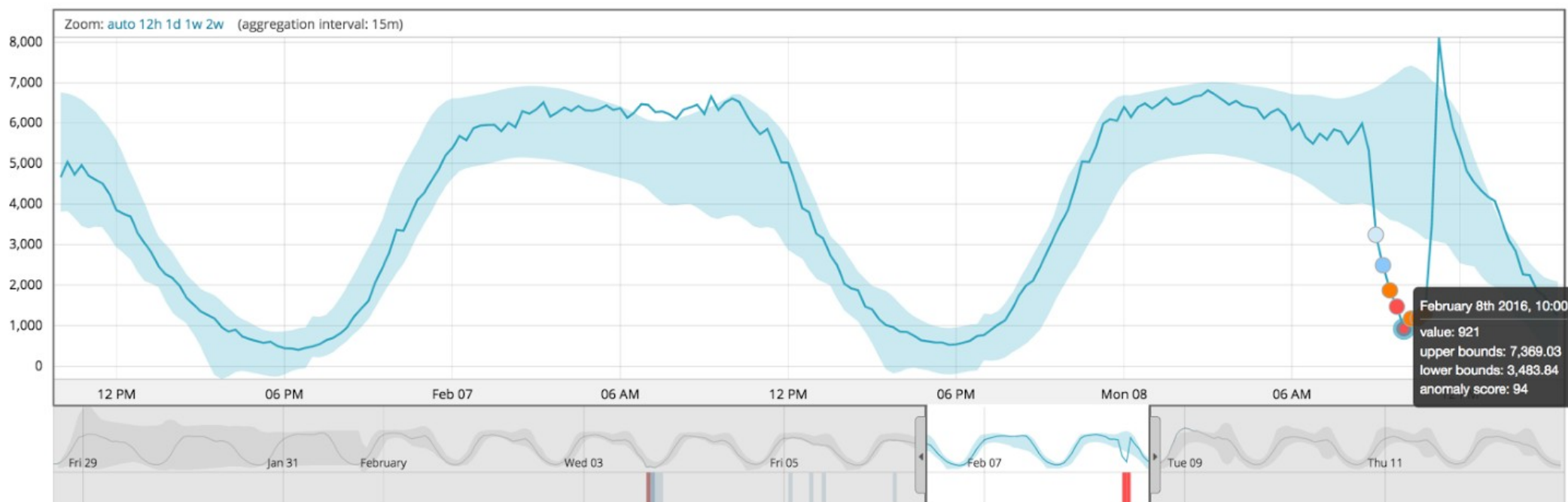
- Le cycle quotidien périodique est appris, puis factorisé.  
La prédiction du modèle s'ajuste après la détection automatique de trois itérations successives de ce cycle.

# Score d'anomalie

- Les modèles statistiques permettent de calculer le taux de probabilité qu'une mesure prenne une certaine valeur.
- Les taux de probabilité oscillant entre 0 et 1 sont normalisés en **scores d'anomalie** qui oscille entre 0 et 100
- Des niveaux de sévérité sont associés à des seuils des scores d'anomalie

# Exemple

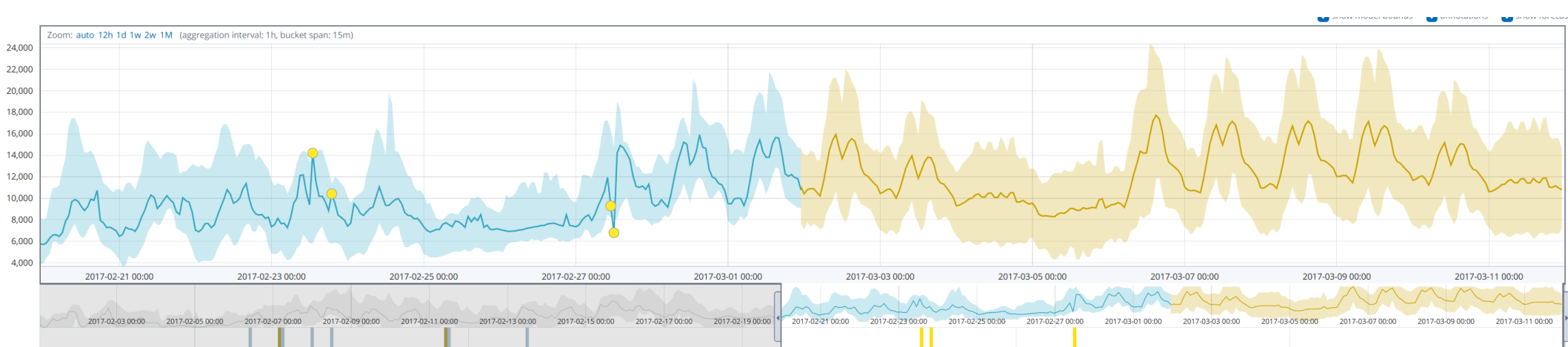
- La probabilité de la valeur 921 était de  $6.3634e-7$
- Cela donne un score d'anomalie de 94
- Ce qui donne une anomalie critique ( $> 75$ )



# Prévisions / Forecast

- Une fois que ML a créé son modèle comportemental, il l'utilise pour extrapoler le comportement futur :
  - Estimer une valeur à une date future
  - Probabilité qu'une valeur atteigne un seuil
- Chaque prévision a un *id* unique, une durée, un délai d'expiration
- Les prévisions ne peuvent pas s'effectuer sur tous les types de valeurs
- Elles sont limitées à 3 pour le même job
- Elles nécessitent pas mal de mémoire

# Example



# Introduction

L'offre Elastic Stack  
Machine Learning pour l'IT  
**Assets ELK-ML**  
Vues Kibana



# Job ML

- Un **job** est l'unité de travail.
- Chaque job a un identifiant unique
- Il existe des jobs pour :
  - La détection d'anomalie sur des données temporelles
  - L'analyse de trame de données
- Un job s'exécute sur un nœud ML
- Il peut s'exécuter en continu ou à la demande

# Nœuds ML

- La ML peut être activée sur tout ou partie des nœuds, mais il est recommandé dans un système de production d'avoir dédié des nœuds au ML  
***xpack.ml.enabled*** et ***node.roles = [ml]***
- Contrairement aux nœuds de données effectuant bcp d'I/O, les nœuds ML nécessitent davantage de CPU et de mémoire
- Les algorithmes ML ne s'exécutent pas dans la JVM.  
Ce sont des exécutables C++ qui utilisent la RAM laissée par la JVM

# Données d'entrée

- Les données d'entrée des jobs proviennent des index ES
  - Les jobs de détection d'anomalie utilisent un **datafeed**
  - Les jobs d'analyse de trame utilise des pipelines d'ingestion
- Le datafeed définit comment les données sont extraites pour alimenter un job. Il est généralement constitué
  - D'une **requête DSL** utilisant des filtres, des agrégations, des transformations
  - D'un **bucket-span**, i.e intervalle temporel, permettant une agrégation des données.

# Exemple : DSL d'un datafeed

```
{
  "query": {
    "bool": {
      "filter": [
        { "term": { "service": "webapp" } },
        { "range": { "@timestamp": { "gte": "now-1h" } } }
      ]
    }
  },
  "aggs": {
    "avg_response_time": {
      "avg": { "field": "response_time" }
    }
  }
}
```

# Index ML

- Elastic ML utilise plusieurs index internes pour le ML :
  - **.ml-config** : Stocke les configurations des jobs.
  - **.ml-state** : Informations internes sur le modèle statistiques appris
  - **.ml-notifications** : Stocke les messages d'audit qui apparaissent dans l'UI : *Job* → *Messages*
  - **.ml-anomalies-\*** : contient les résultats détaillés des jobs d'anomalies. L'index *.ml-anomalies-shared* contient les informations de plusieurs jobs.
  - **.ml-inference-\*** : Stocke les données liées aux modèles d'inférence
  - **.ml-stats** : Contient des métriques liées aux jobs (performance, consommation de ressource)
  - **.ml-telemetry** : Informations télémétriques liées à l'utilisation des fonctionnalités ML pour améliorer les fonctionnalités ML.

# Kibana vs ML APIs

- L'UI de Kibana offre une simplification de l'utilisation de l'API Rest mais celle-ci reste la plus puissante et la plus flexible.
- Avec l'API, il est possible de :
  - Gérer les jobs : créer/éditer/exécuter, déclencher des prévisions
  - Gérer les datafeeds : Créer/démarrer/arrêter/surveiller
  - Accéder aux résultats des analyses
  - Évaluer la performance de modèles supervisés
  - ....

# Exemple : Création de Job

**PUT \_ml/anomaly\_detectors/total-requests**

```
{
  "description" : "Total sum of requests",
  "analysis_config" : {
    "bucket_span": "10m",
    "detectors": [
      {
        "detector_description": "Sum of total",
        "function": "sum",
        "field_name": "total"
      }
    ]
  },
  "data_description" : {
    "time_field": "timestamp",
    "time_format": "epoch_ms"
  }
}
```

# Introduction

L'offre Elastic Stack  
Machine Learning pour l'IT  
Assets ELK-ML  
**Vues Kibana**



# Introduction

- L'interface Kibana propose 5 menus principaux :
  - **Détection d'anomalies** : ML non supervisé, permet de définir les jobs de détection d'anomalie sur un historique et leurs visualisations
  - **Analyse de trame de données** : Gestion des jobs et visualisations résultat
  - **Gestion de modèles** : Modèles d'inférences pour les modèles supervisés
  - **Data Visualizer** : Outils pour visualiser les données d'entrées
  - **AIOps Labs**: Outils ML

# Data Visualizer

- Data Visualizer propose 4 entrées :
  - **File** : Permet d'uploader des fichiers de données sans les ingérer dans ES
  - **ES|QL** : Permet de faire des requêtes d'agrégation et de générer des visualisations
  - **Data Viewer** :
    - Identifie les champs non vides,
    - propose un graphe de distribution des valeurs de chaque champ  
=> Aide à choisir les champs à analyser avec Elastic ML
  - **Data Drift** : Permet de comparer 2 périodes différentes d'une série temporelle

# Dataviewer

- Les champs dans les index sont répertoriés dans 2 sections :
  - Les champs **numériques** ("métriques").  
Pour chaque champ, le Visualiseur indique le nombre de documents contenant le champ dans la période sélectionnée, les valeurs minimales, médianes et maximales, le nombre de valeurs distinctes et leur distribution.
  - **keyword** :  
Nombre de valeurs distinctes, Top values, % de documents contenant le champ
  - Les **autres** champs (text, date, boolean)  
*date*:  
Plus ancienne, plus récente, % de documents contenant le champ

# Tableau de bord des jobs

- Les jobs sont séparés dans les menus « Détection d'anomalie » et « Analyse de données »
- Les tableaux présentent :
  - L'id du job, sa description
  - Son statut mémoire, i.e. la mémoire utilisée par le modèle :
    - ok,
    - soft\_limit : les vieux modèles vont être nettoyés
    - hard\_limit : toutes les données n'ont pas pu être traitées
  - Le statut du job
- Une vue détaillée permet de voir les métriques précises d'exécution

# Job Management

Machine Learning / Job Management

30 seconds

[Job Management](#) [Anomaly Explorer](#) [Single Metric Viewer](#) [Data Visualizer](#) [Settings](#)

Active ML Nodes: 0 Total jobs: 1 Open jobs: 0 Closed jobs: 1 Active datafeeds: 0

Refresh

+ Create new job

Search...

Opened Closed Failed Started Stopped Group ▾

<input type="checkbox"/>	ID ↑	Description	Processed records	Memory status	Job state	Datafeed state	Latest timestamp	Actions
<input type="checkbox"/>	> total-requests	Total sum of requests	14,040	ok	closed	stopped	2017-04-01 23:59:00	

Rows per page: 10 ▾

# Détection d'anomalie

- 2 menus permettent de visualiser les résultats des jobs d'anomalie :
  - **Anomaly Explorer** : Couloirs indiquant le score d'anomalie maximal au fil du temps.
  - **Single Metric Viewer** : Graphique représentant les valeurs réelles et attendues dans le temps.  
Uniquement disponible pour les jobs de type Single Metric.
- **Settings** : permet de configurer des calendriers pour exclure certaines périodes particulières de l'analyse et des filtres
- **Supplied configurations** : Des configurations prêtes à l'emploi pour des événements générés par des systèmes classiques (Apache, Nginx, Docker, Beats, ..)

# Analyse de trame de données

- Outre la visualisation des jobs, le menu « Data frame analytics » offre 2 visualisation permettant d'exploiter les résultats :
  - **Result Explorer** : Permet de parcourir les résultats produits par des jobs comme la classification, la régression ou la détection de valeurs aberrantes.  
Les résultats incluent des prédictions générées par le modèle ou des scores d'anomalies.
  - **Analytics Map** : Visualiser la structure et les relations entre les données d'un ou plusieurs jobs.  
Explorer les dépendances et les transformations appliquées pendant l'analyse.

# Gestion des modèles

- Permet de déployer des modèles d'inférence :
  - Issus de jobs supervisés
  - Fournis par d'autres sources (Ex : Hugging Face)
- 3 modèles pour le NLP sont fournis par elastic :
  - ***.elser\_model\_2\_linux-x86\_64, .multilingual-e5-small\_linux-x86\_64***: Recherche sémantique sur du texte
  - ***lang\_ident\_model\_1*** : Modèle permettant d'identifier la langue d'un texte source



# Outils ML

- AIOps Labs propose 3 outils
  - ***Log Rate Analysis*** : Identifie les pics ou baisses dans les taux d'événements (log rate spikes) ainsi que leur cause.
  - ***Log Pattern Analysis*** : Regroupe des messages texte similaires
  - ***Change Point Detection*** : Détecte automatiquement les points de changement significatifs dans une série temporelle.

# Détection d'anomalies

## **Single et multi-metric Jobs**

Autres jobs

API et jobs avancés

Optimisations

Analyse de cause et jobs multiples

Alertes

Prévisions

# Types de job

- « **Single Metric** » permet d'analyser une unique métrique dans des données temporelles
- « **Multi-metric** » permet de corréler différentes analyses portant sur des métriques différents.  
Ce type de job est adapté à l'analyse de cause
- « **Population** » permet de comparer des entités avec un modèle incluant tous les autres membres observés au fil du temps
- « **Categorization** » groupe des événements texte (typiquement des messages de logs) dans des catégories et y détecte des anomalies
- « **Rare** » : Détecte des valeurs rares dans des données temporelles
- Le job « **Advanced** » permet d'avoir plus de contrôle sur la configuration de l'analyse.  
Les autres types de job ne sont que des simplifications des job « *Advanced* »

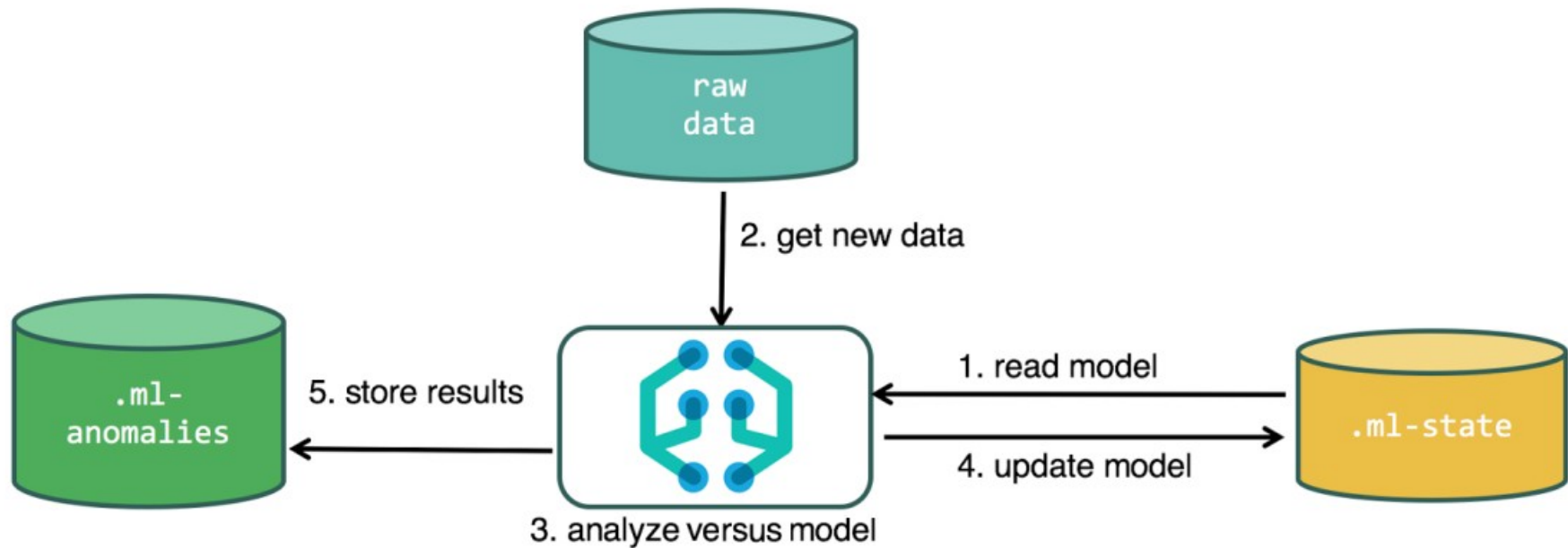
# Bucketization

- Le ***bucket-span*** correspond à la fenêtre de temps pour laquelle les données sont agrégées à des fins d'analyse
  - Plus la durée de *bucket\_span* est courte, plus l'analyse est fine, mais aussi plus l'influence du bruit dans les données est élevée
- Pour la configurer prendre en compte :
  - La granularité de l'analyse
  - La fréquence des données d'entrée
  - La durée typique d'une anomalie
  - La rapidité d'alerte voulue
- Typiquement entre 5mn et 1h

# Influence du bucket



# Workflow pour un bucket span



# Datafeed

- Le ***datafeed*** est le mécanisme par lequel les données sont régulièrement récupérées et présentées aux algorithmes ML
- Un datafeed peut être démarré (il s'exécute tous les x temps) ou arrêté (désactivation)
- Il est défini par :
  - ***Une requête DSL*** : La requête DSL est exécutée avec un intervalle de date correspondant au bucket recherché
  - ***query\_delay*** : La latence entre l'indexation et l'exécution de la requête ; afin d'être sûr que toutes les données du bucket\_span ont bien été ingérées
  - ***frequency*** : Fréquence d'exécution de la requête. Égal au *bucket* sauf si celui-ci est supérieur à 20 mn, dans ce cas la fréquence est + courte
  - ***scroll\_size*** : Taille de la pagination lors de la requête elastic search.

# Cycle de vie d'un job

- Ouverture du job : Allocation des ressources nécessaires sur les nœuds ML, préparation du datafeed
- Démarrage du datafeed : Extraction des données source par lots et transmission au job pour analyse.
- Traitement des données : Détection des anomalies, résultats stockés dans *.ml-anomalies-\**
- Pause du job ou Arrêt du datafeed : Collecte des données mise en pause, mais le job reste ouvert.
- Fermeture du job : Libère les ressources allouées au job, ferme automatiquement le datafeed associé, ses résultats et ses configurations restent accessibles.
- Suppression du job : Supprime les configurations associées et les résultats



# Fermeture du datafeed

- Le datafeed se ferme dans les cas suivants :
  - Manuellement : Fermeture du job ou arrêt du datafeed :
  - Fin des données :  
Si le datafeed atteint la fin des données disponibles et qu'il est configuré pour ne pas fonctionner en temps réel (par exemple, un job historique), il se termine automatiquement après avoir traité toutes les données.
  - Job inactif pendant trop longtemps :  
Elasticsearch peut fermer un datafeed inactif ou associé à un job qui ne reçoit plus de nouvelles données, pour économiser les ressources.

# Détecteurs

- Lors de la création de job, on spécifie un ou plusieurs **détecteurs**, qui définissent les champs de données du job et le type d'analyse à effectuer :
- Un détecteur est défini par les propriétés suivantes :
  - ***field\_name*** : Le champ utilisé
  - ***function*** : La fonction d'analyse utilisée. Comptage, moyenne, somme, etc ..
  - ...

# Fonctions

- Les fonctions permettent :
  - une agrégation du métrique pour le bucket span
  - La détection d'une valeur anormales
- Il existe :
  - Des fonctions de comptage
  - Des fonctions de sommes
  - Des métriques (moyenne, min, max, variance)
  - Des fonctions pour la rareté
  - Des fonctions temporelles (anomalie par rapport à l'heure ou au jour de la semaine)
  - Des fonctions géographiques
  - Des fonctions sur une volume de contenu

# Fonctions

- La plupart des fonctions détectent des anomalies dans les valeurs basses et hautes. Dans la terminologie statistique, ils appliquent un test *bilatéral*. Certaines fonctions ne font un test que d'un côté (*high\_count*, *low\_count*)
- Si les données sont éparées, les buckets vides peuvent être ignorés.

# Fonctions de comptage

- EL-ML propose 3 familles de comptage de documents :
  - ***count***, ***high\_count*** et ***low\_count*** : compte le nombre de documents dans le bucket-span. Détecte des anomalies ou des anomalies hautes ou basses
  - ***non\_zero\_count***, ***non\_zero\_high\_count*** et ***non\_zero\_low\_count*** : Idem mais ne prend pas en compte les buckets qui n'ont pas d'évènements
  - ***distinct\_count***, ***distinct\_high\_count*** et ***distinct\_low\_count*** : Détecte les anomalies sur le nombre de valeurs distinctes
- Il est quelquefois plus avantageux d'utiliser 2 détecteurs *low*, *high* plutôt que le *count*.  
Surtout si le nombre d'anomalies high et low sont très différents.

# Fonctions de somme

- Dans la même idée, EL-ML propose 2 familles de somme de valeur :
  - *sum, high\_sum* et *low\_sum* : détecte les anomalies concernant la somme d'un champ
  - *non\_null\_sum, high\_non\_null\_sum* et *low\_non\_null\_sum* : Idem mais ne prend pas en compte les buckets qui n'ont pas d'évènements

# *Single Metric Viewer*

- Les valeurs réelles sont indiquées par une ligne bleue. Une zone bleue ombrée représente les limites des valeurs attendues.
- Si une valeur est en dehors de la zone ombragée, elle est anormale.
  - Un score d'anomalie est calculé pour le bucket et donne une couleur au point
  - Les détails des anomalies est visible dans la partie inférieure
- Il est possible de faire glisser le sélecteur de temps

# Single Metric Viewer

Machine Learning / Single Metric Viewer

Auto-refresh ◀ ⌚ March 23rd 2017, 06:00:00.000 to April 22nd 2017, 04:59:00.000 ▶

Job Management Anomaly Explorer Single Metric Viewer Data Visualizer Settings

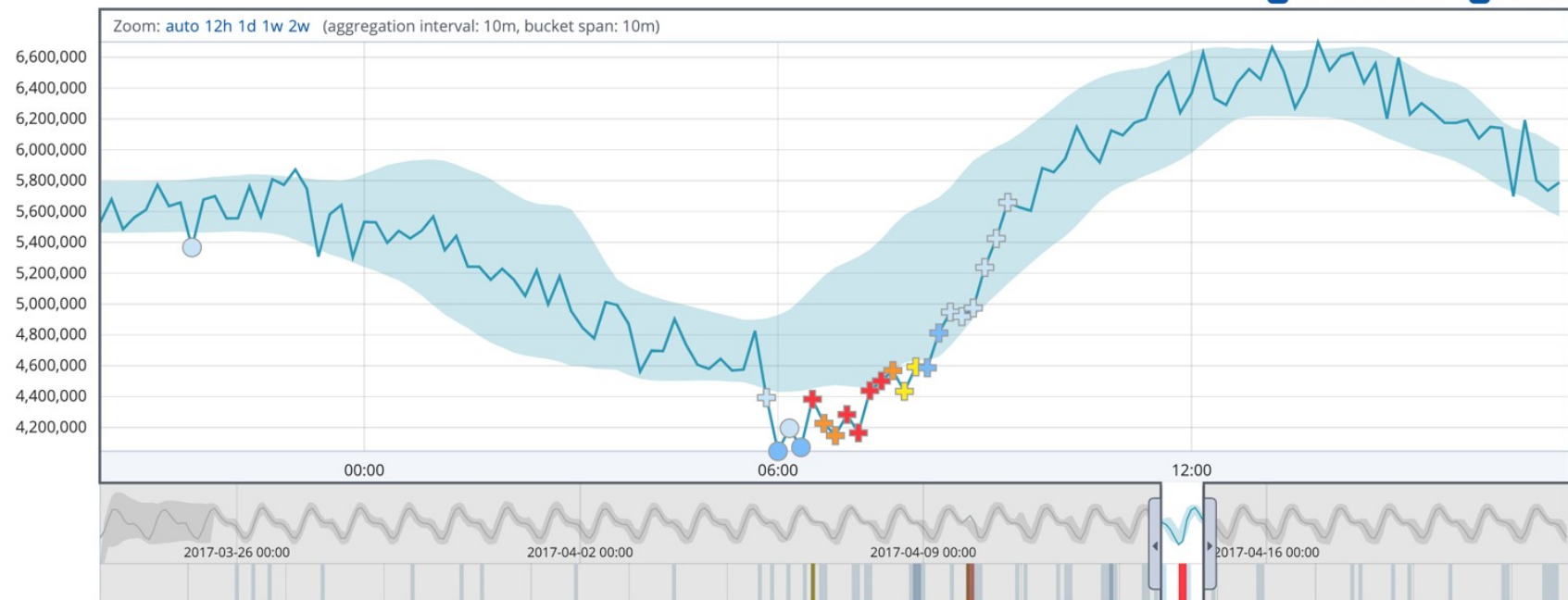
Job total-requests

Detector: sum(total) ▶

Forecast

Single time series analysis of sum total

☒ show model bounds ☒ annotations



## Anomalies

Severity threshold

☒ warning

Interval

Auto

time	max severity ↓	detector	actual	typical	description	job ID	actions
> April 14th 2017, 07:00	● 96	sum(total)	4,283,309	4,869,925.771	↓ 1.1x lower	total-requests	⚙
> April 14th 2017, 06:00	● 88	sum(total)	4,382,911	4,779,628.322	↓ 1.1x lower	total-requests	⚙



# Multi-bucket impact

- Il se peut que des événements anormaux n'existent pas au sein d'une seule plage mais se produisent sur une gamme de plages contigus.
  - Pour prendre en compte cette situation, EL-ML effectue une analyse *multi-bucket* en prenant en compte les buckets contigus
- Lorsque l'on visualise les résultats, une propriété ***multi\_bucket\_impact*** indique la force avec laquelle le score final d'une anomalie est influencé par l'analyse multi-bucket
- Dans Kibana, les anomalies avec des impacts `multi_bucket` élevés sont indiquées avec des croix plutôt que des points

# Prévisions

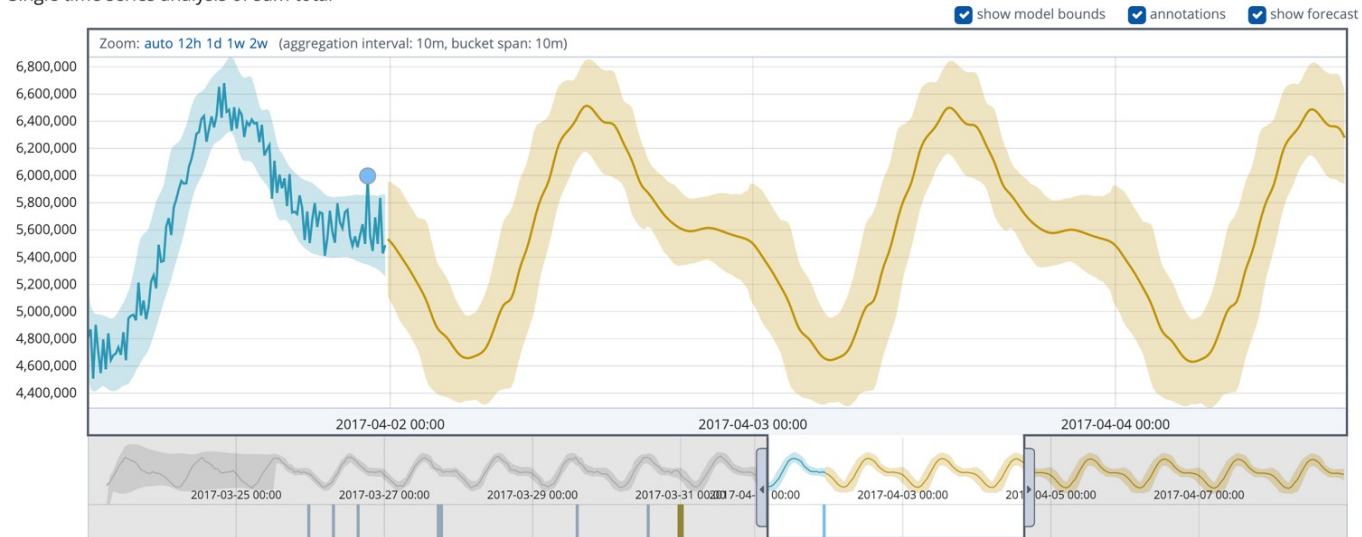
- Les prévisions sont visibles dans le *Single Metric Viewer*
- La ligne jaune dans le graphique représente les valeurs de données prédites.
- La zone jaune ombrée représente les limites des valeurs prédites, ce qui donne également une indication de la confiance des prédictions.  
Les limites augmentent généralement avec le temps (i.e. les niveaux de confiance diminuent).
- Si les niveaux de confiance sont trop bas, la prévision s'arrête

# Prévisions

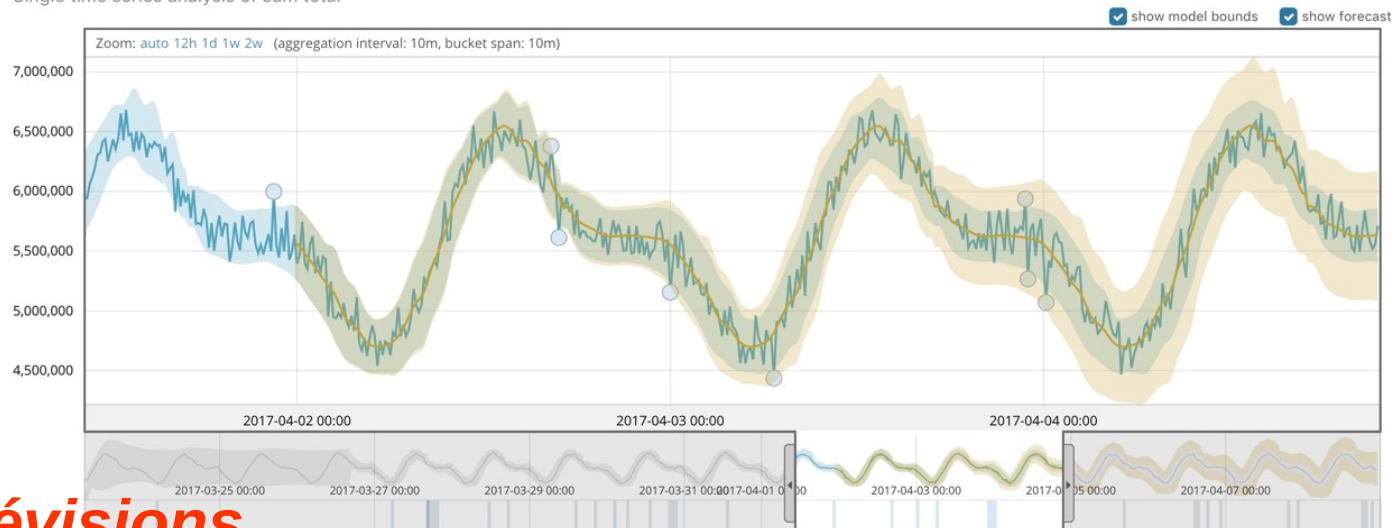
Detector:  

Forecast

Single time series analysis of sum total



Single time series analysis of sum total



# Jobs multi-metric

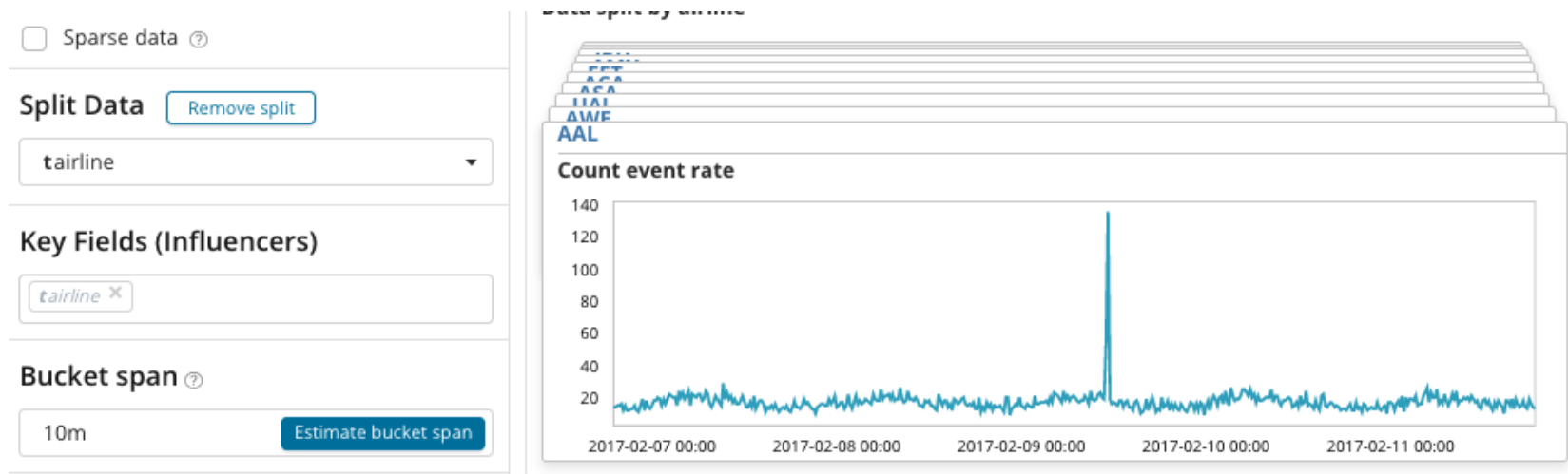
- Les jobs multi-metrics peuvent avoir :
  - plusieurs détecteurs,
  - des influenceurs permettant d'identifier la cause de l'anomalie
  - Un champ de split permettant de diviser les données d'entrées
- Ils peuvent être vus comme de multiple jobs single metric indépendants qui permettent cependant :
  - D'obtenir un score d'anomalie global
  - De définir des influenceurs s'appliquant à tous les détecteurs
  - D'effectuer plusieurs analyses indépendantes en fonction de la valeur du champ de split
- Le Single Metric Viewer n'est pas disponible pour ce type de job

# Influenceurs

- Les influenceurs sont des champs que l'on soupçonne influencer ou contribuer aux anomalies
- Ce sont des métadonnées qui accompagnent les anomalies détectées, mais ils ne participent pas directement aux calculs d'anomalies
- Leurs usages sont :
  - Identifier les causes de l'anomalie
  - Faciliter la recherche et la navigation des anomalies

# Split Data

- Les fonctions peuvent être associées à un champ de split et ainsi les détecteurs sont appliqués pour des sous-ensemble des données d'entrée



# Anomaly Explorer

- Adaptée à tout type de job, cette vue permet d'explorer la chronologie globale des anomalies.
  - Pour chaque section de la période spécifiée, le score d'anomalie maximal est indiqué.
    - Les sections ne correspondent pas nécessairement au bucket span. Elles s'adaptent à la période sélectionnée. La taille la plus petite cependant est le bucket span.
    - La section est clickable et permet d'accéder aux détails des anomalies (graphique et tableau)
- Elle met en lumière les influenceurs
  - À gauche, la liste des principaux influenceurs avec leur score maximal d'anomalies.
  - Un couloir par influenceur affichant la chronologie des anomalies
- Sur une anomalie un menu action permet :
  - D'afficher l'anomalie dans un Single Metric Viewer
  - D'afficher les données dans Discover
  - De créer une règle excluant l'anomalie des résultats ou de la mise à jour du modèle

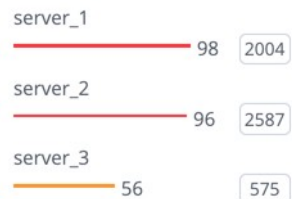
# Example

## Top Influencers

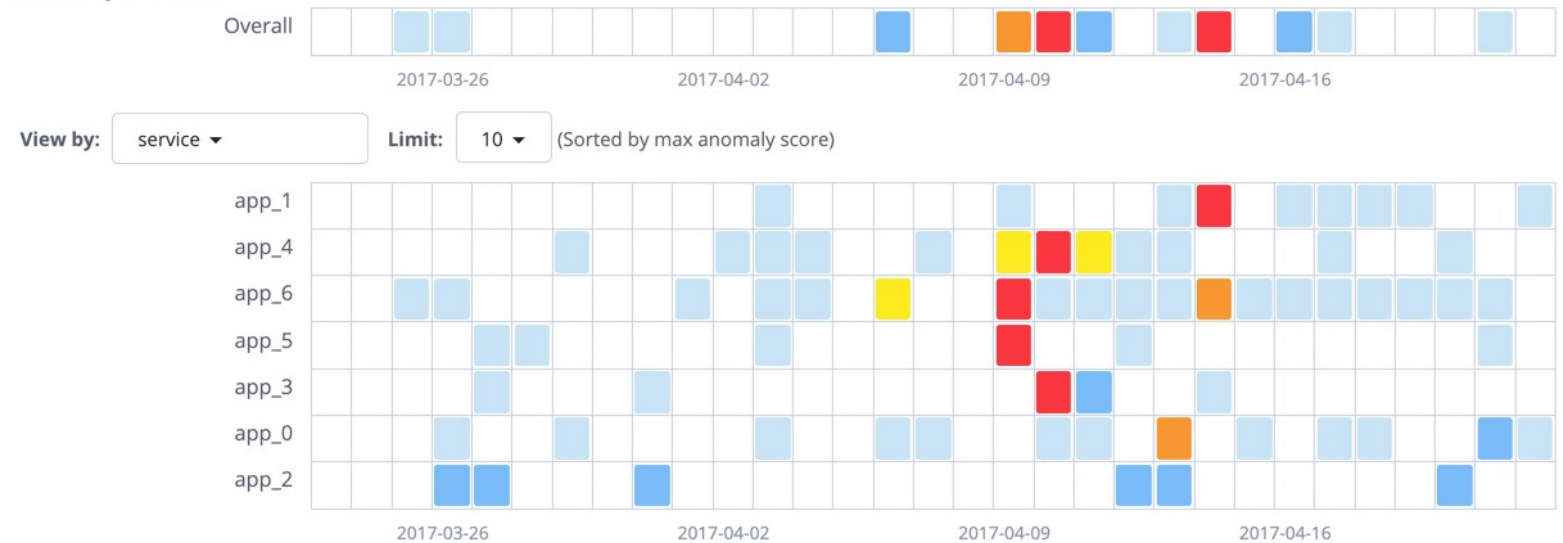
### service



### host



## Anomaly timeline



## Anomalies

Severity threshold

**warning**

Interval

**Auto**

	time	max seve...	detector	found for	influenced by	actual	typical	description	job ID	actions
>	April 14th 2017	● 99	sum(total)	app_1	service: app_1	707,990	1,301,631.575	↓ 2x lower	response_requ ests_by_app	⚙
>	April 9th 2017	● 99	high_mean(res ponse)	app_6	host: server_2 service: app 6	2.762	2.45	↑ 1.1x higher	response_requ ests by app	⚙



# Détection d'anomalies

Single et multi-metric Jobs

**Autres jobs**

API et jobs avancés

Optimisations

Analyse de cause et jobs multiples

Alertes

Prévisions

# Analyse de population

- Des entités ou des événements peuvent être considérés comme anormaux lorsque :
  - Leur comportement change au fil du temps, par rapport à leur propre comportement antérieur, ou
  - Leur comportement est différent de celui des autres entités d'une population spécifiée.C'est l'analyse de population
- Kibana propose un assistant dédié à l'analyse de population

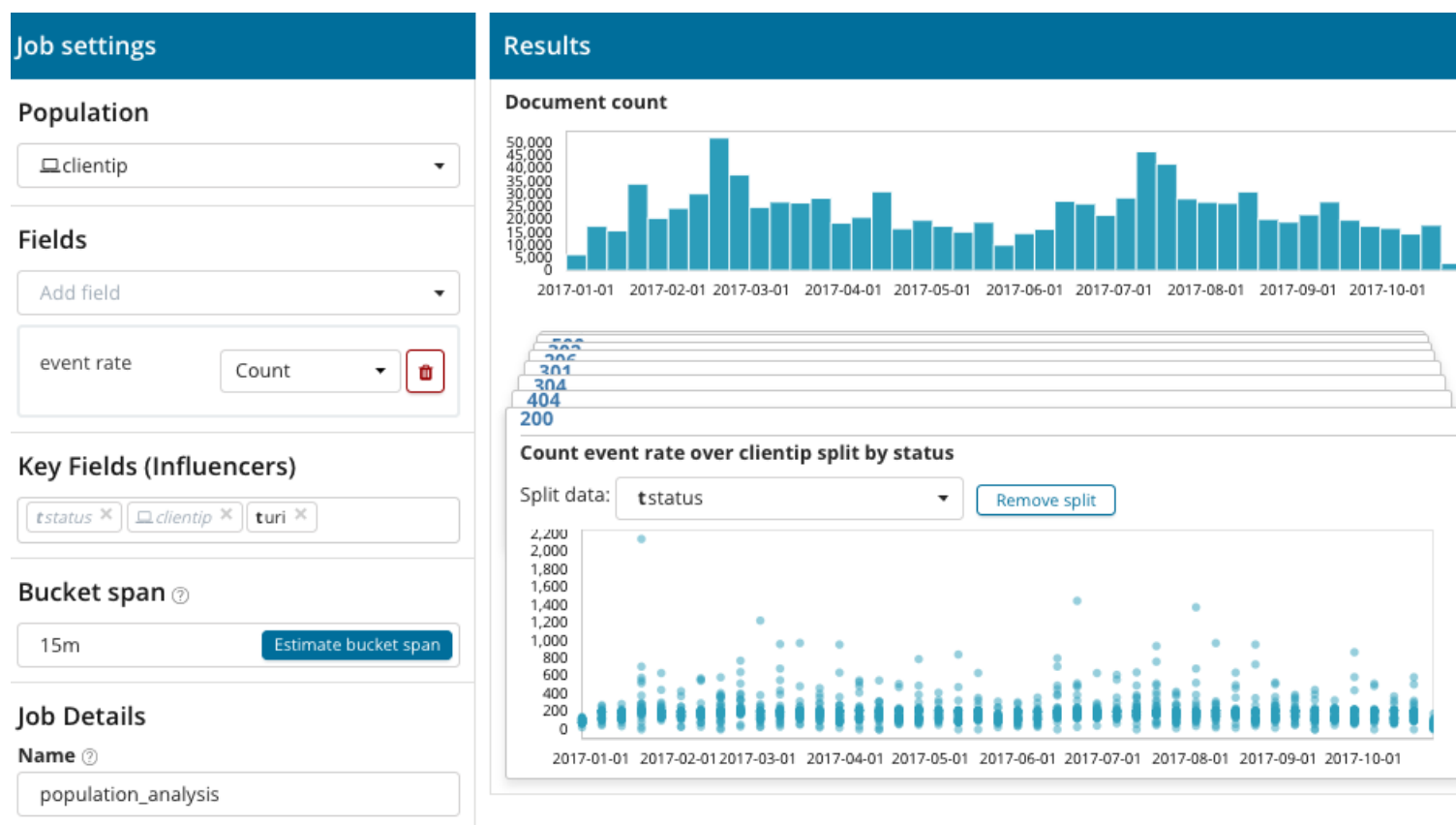
# Comptage et analyse de population

- Comptabiliser les événements au travers d'une population a de nombreux cas d'utilisation :
  - Trouver des serveurs qui génèrent beaucoup plus de traces que leurs pairs
  - Un utilisateur qui effectue beaucoup plus de requêtes que les autres
  - ....

# Job population

- Un job population divise les données par les valeurs distinctes d'un champ.
- Ce champ définit ce que l'on appelle une population.
- Les divisions sont alors analysées dans le contexte de toutes les divisions pour trouver des valeurs inhabituelles dans la population.
- La configuration consiste donc à spécifier :
  - Le champ de population
  - La fonction à appliquer sur un champ

# Exemple Configuration



# Catégorisation de message

- Les jobs de catégorisation
  - regroupent les valeurs de texte similaires,
  - les classent en catégories
  - Puis détectent les anomalies au sein des catégories.
- Ces jobs fonctionnent mieux sur le texte écrit par des machine comme les messages de logs d'un service applicatif.

# Algorithme de clustering

- EL-ML permet de retrouver des messages de logs similaires et ainsi de les catégoriser.
  - Les messages de log doivent être générés par une machine (Pas de texte libre saisi par un humain)
- EL-ML utilise un algorithme de similarité de chaîne :
  - Se concentre sur les mots du dictionnaire (anglais) et non pas les chaînes variables (serveur, adresse, etc..)
  - Utilise un algorithme proche de celui de Levenshtein pour calculer une distance entre 2 messages
    - Si la distance est petite, positionne les 2 messages dans la même catégorie
    - Sinon, crée une nouvelle catégorie

# Exemple

Error writing file "foo" on host "acme6"

Error writing file "bar" on host "acme5"

Opening database on host "acme7"

- Dans ce cas EL-ML trouve 2 catégories, il compte alors le nombre d'occurrence pour chaque catégorie qu'il nomme *mlcategory* *N* :

mlcategory 1: 2

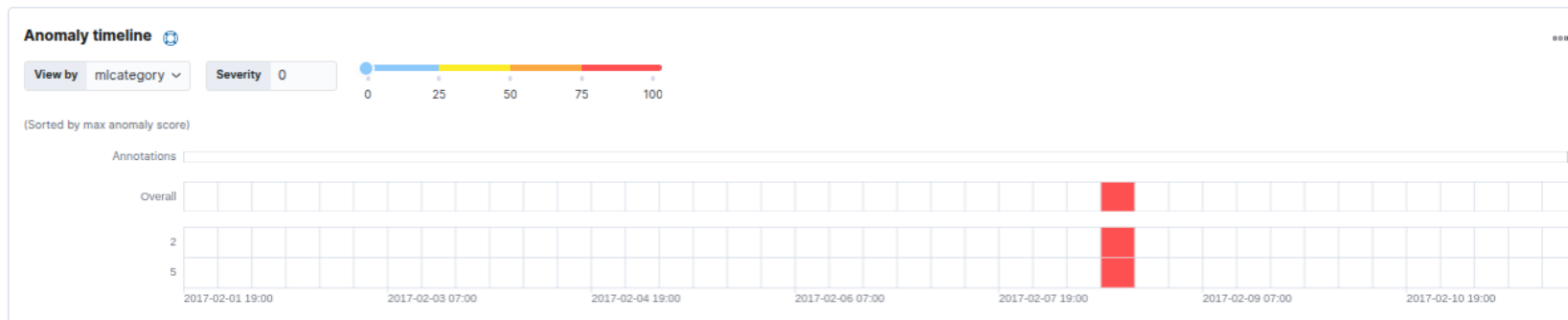
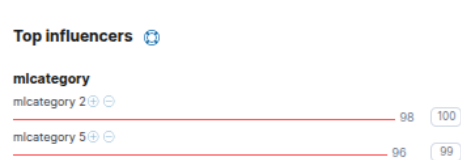
mlcategory 2: 1



# Fonctions proposées par l'assistant

- Lors d'un job de catégorisation, les fonctions disponibles sont :
  - **Count** : Détection d'anomalie sur le taux d'évènements d'une catégorie, i.e un message anormalement produit sur une période
  - **high\_count** : Détection d'une haute production d'un type de message
  - **Rare** : Recherche des message qui apparaissent rarement

# Anomaly explorer



> Annotations Total: 1

## Anomalies

Severity warning Interval Auto

Time	Severity	Detector	Found for	Influenced by	Actual	Typical	Description	Category examples	Actions
> February 8th 2017, 16:00	+ 98	count by micategory	micategory 2	micategory: 2	49	0.0539	More than 100x higher	REC Not INSERTED [DB TRAN] Ta...	
> February 8th 2017, 16:00	+ 96	count by micategory	micategory 5	micategory: 5	50	0.271	More than 100x higher	Opening Database = DRIVER=(SQ... Opening Database = DRIVER=(SQ... Opening Database = DRIVER=(SQ...	
> February 8th 2017, 17:00	< 1	count by micategory	micategory 5	micategory: 5	0	0.351	Unexpected zero value	Opening Database = DRIVER=(SQ... Opening Database = DRIVER=(SQ... Opening Database = DRIVER=(SQ...	
> February 8th 2017, 17:00	< 1	count by micategory	micategory 2	micategory: 2	0	0.138	Unexpected zero value	REC Not INSERTED [DB TRAN] Ta...	

# Fonction rare

- La notion de rareté tient compte du contexte :
  - S'il y a beaucoup de choses uniques, alors rien n'est rare.
  - S'il y a beaucoup de choses identiques et peu de choses uniques, alors elles sont rares.
- La fonction **rare** s'applique sur un champ et permet de détecter les valeurs rares d'un champ. Par exemple :
  - Un message de log rare
  - Un process s'exécutant rarement
  - Des destinations de connexion rares

# Fonctions rare

- Les fonctions *rare* détectent des valeurs qui arrivent rarement dans le temps
- Elles peuvent être segmentées en population
- 2 fonctions existent :
  - ***rare*** : détecte les anomalies selon le nombre de valeurs rares distinctes
  - ***freq\_rare*** : détecte les anomalies en fonction du nombre de fois (fréquence) qu'apparaissent de valeurs rares

# Kibana

- *Kibana* propose désormais un assistant pour créer ce type de job

## Create job: Rare

Using index pattern logstash-apache\*



Time range

2

Pick fields

3

Job details

### Pick fields

#### Rare detector

##### Rare

Find rare values over time.

✓ Selected

##### Rare in population

Find members of a population that have rare values over time.

Select

##### Frequently rare in population

Find members of a population that frequently have rare values.

Select

#### Rare field

Select a field in which to detect rare values.

Rare field

< Previous

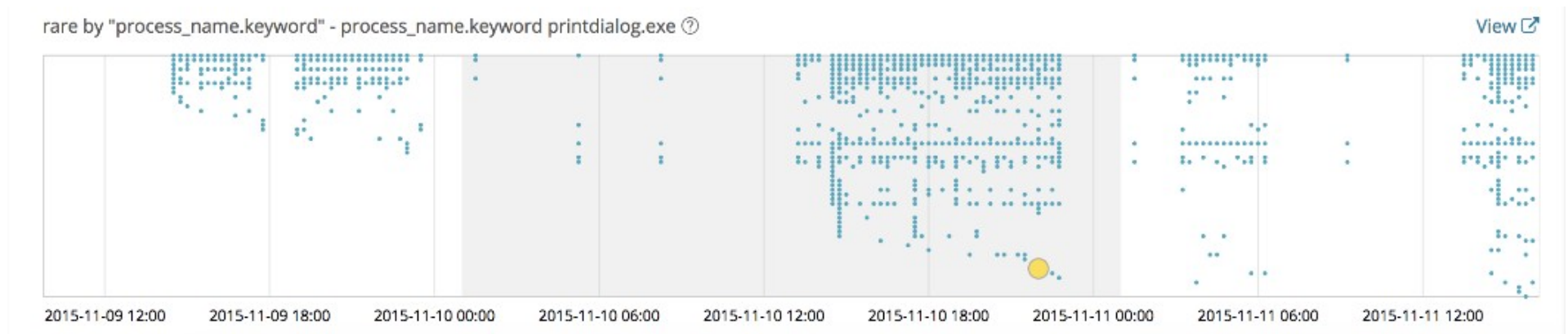
Next >

# Explication et cas d'usage

- Les 3 choix proposés par l'assistant Kibana
  - **Rare** : Détection d'événements rares dans l'ensemble des données.  
Ex : Une adresse IP qui apparaît rarement par rapport à la moyenne globale
  - **Rare in population** : Identifier des événements rares dans des sous-groupes.  
Ex : Détecter les adresses IP rares par région.
  - **Frequently rare in population** : Identifier des patterns rares qui se produisent fréquemment dans un groupe spécifique.  
EX : Des transactions rares sont souvent effectués par un utilisateur particulier

# Vue « Rareté »

- Les points bleus indiquent les taux d'occurrence des valeurs de champ dans le temps.
- Ceux qui se retrouvent près du bas sont les plus rares et l'anomalie sélectionnée est affichée sous forme d'un point agrandi



# Geo jobs

- Les jobs de détection d'anomalies géographiques détectent les valeurs inhabituelles dans les coordonnées géographiques.
  - L'index doit contenir des champs de type geo\_loc
  - La fonction *lat\_long* dans le détecteur est utilisée.
- On peut, par exemple, identifier les transactions initiées à partir d'emplacements inhabituels par rapport aux emplacements du reste des transactions.



# Détection d'anomalies

Single et multi-metric Jobs

Autres jobs

**API et jobs avancés**

Optimisations

Analyse de cause et jobs multiples

Alertes

Prévisions

# Jobs avancés

- Les jobs avancées offrent toute la flexibilité possible de l'API de création de jobs.
- L'assistant dans Kibana permet de voir le JSON envoyé à l'API lors de la création du job.
- Tous les autres types jobs présentées précédemment peuvent être créés en tant que jobs avancés.

# API

- ELK-ML propose une API avec la base ***\_ml***
  - ***/anomaly\_detectors/***: Création et gestion des jobs
  - ***/calendars/***: Création et gestion des calendriers
  - ***/datafeeds/***: Sélection des données à analyser
  - ***/filters/***: Création et gestion des filtres utilisés par les règles personnalisées
  - ***/results/***: Accès aux résultats des jobs
  - ***/model\_snapshots/***: Gestion des instantanés de mémoire

# Création de job

- Les principaux blocs JSON pour la création de jobs sont :
  - ***job\_id, description, groups*** : Identification du job
  - ***analysis\_config*** : La configuration d'analyse. Après avoir créé un job, on ne peut pas modifier la configuration d'analyse.
    - ***bucket\_span***
    - ***categorization\_analyzer, categorization\_field\_name, categorization\_filters*** : Champs pour la catégorisation des données d'entrée texte
    - ***detectors*** : Tableau de détecteur
    - ***influencers*** : Liste des champs influenceurs
    - ***summary\_count\_field\_name*** : Utile lorsqu'on le pré-agrège les données du datafeed. Indique le nombre de documents qui ont été agrégés
  - ***analysis\_limits*** : Des limites pour conserver les modèles mathématiques en mémoire.
  - ***data\_description*** : Spécification du champ @timestamp
  - ***datafeed\_config*** : Configuration du data feed
  - ***model\_plot\_config*** : stocke les informations du modèle avec les résultats pour une vue plus détaillée de la détection des anomalies.
  - ***results\_index\_name*** : L'index stockant les résultats. Par défaut *shared*

# Configuration du datafeed

- Les principaux attributs du bloc datafeed :
  - ***query, query\_delay, frequency, scroll\_size***
  - ***datafeed\_id*** : Un identifiant
  - ***indices*** : Tableau d'index sur lesquels porte la recherche. Le caractère \* supporté
  - ***runtime\_mappings*** : Permet d'ajouter des champs calculés via le script Painless
  - ***aggregations*** : Permet d'indiquer une requête d'agrégations de type date\_histogram. Moyen d'améliorer performance mais limitations
  - ***delayed\_data\_check\_config*** : Indique si le datafeed vérifie si des données ont été ajoutées à posteriori après une requête initiale
  - ***max\_empty\_searches*** : Le maximum de recherches vides avant la fermeture du datafeed

# Objets détecteur

- Chaque détecteur configuré peut avoir les propriétés suivantes :
  - ***function*** : La fonction analytique
  - ***field\_name*** : Le champ utilisé par la fonction. Pas applicable pour les calculs d'occurrence (count ou rare)
  - ***by\_field\_name*** : Le champ utilisé pour fractionner les données, permet de trouver des valeurs inhabituelles dans le contexte du fractionnement. Par exemple, une catégorisation de message
  - ***over\_field\_name*** : Pour une analyse de population
  - ***partition\_field\_name*** : Champ pour segmenter l'analyse. Les analyses sont alors indépendantes
  - ***use\_null*** : Définit si une série est créée pour les valeurs nulles de *by\_field\_name* ou *partition\_field\_name*
  - ***exclude\_frequent*** : *all*, *none*, *by*, ou *over*. Si défini, les entités fréquentes (en temps ou par population) sont exclues pour le calcul d'anomalie. On peut distinguer les champs *by* ou *over*
  - ***custom\_rules*** : Règles personnalisant le fonctionnement du détecteur. Permet par exemple d'indiquer de ne pas prendre en compte certains résultats

# *by\_field\_name* versus *partition\_field\_name*

- La configuration de ses 2 champs divise les données et sépare les affichages
- Ils peuvent être utilisés séparément ou ensemble dans un détecteur
  - Si on veut séparer complètement les analyses, utiliser “*partition\_field\_name*”  
=> Le calcul du score des anomalies est **indépendant**
  - Pour une séparation « douce », utiliser “*by\_field\_name*”  
=> Le calcul du score prend en compte les autres valeurs du champ

# Example API

```
PUT _ml/anomaly_detectors/test-job1?pretty
{
  "analysis_config": {
    "bucket_span": "15m",
    "detectors": [
      {
        "detector_description": "Sum of bytes",
        "function": "sum",
        "field_name": "bytes"
      }
    ]
  },
  "data_description": {
    "time_field": "timestamp",
    "time_format": "epoch_ms"
  },
  "analysis_limits": {
    "model_memory_limit": "11MB"
  },
  "model_plot_config": {
    "enabled": true,
    "annotations_enabled": true
  },
  "results_index_name": "test-job1",
```



# Example API (2)

```
"datafeed_config":
{
  "indices": [
    "kibana_sample_data_logs"
  ],
  "query": {
    "bool": {
      "must": [
        {
          "match_all": {}
        }
      ]
    }
  },
  "runtime_mappings": {
    "hour_of_day": {
      "type": "long",
      "script": {
        "source": "emit(doc['timestamp'].value.getHour());"
      }
    }
  },
  "datafeed_id": "datafeed-test-job1"
}
```

# APIs sur les jobs

- L'APIs permet beaucoup d'autres opérations que l'interface Kibana ne permet pas.
  - En particulier, il est plus simple de mettre à jour un job via l'API
- Les autres opérations sont :
  - Suppression ou reset d'un job
  - Ajout, suppression de calendriers associés
  - Get info ou statistics
  - Poster directement des données à un jobs
  - Créer ou supprimer des prévisions

# Exemple : Mise à jour

```
POST _ml/anomaly_detectors/spring-cat-high-count/_update
{
  "model_plot_config": {
    "enabled": true
  }
}
```

# Démarrage du job et accès aux résultats

- Ouverture du job

POST `_ml/anomaly_detectors/test-job1/_open`

- Démarrage du datafeed associé

POST `_ml/datafeeds/datafeed-test-job1/_start`

- Vérification du statut

GET `_ml/anomaly_detectors/test-job1`

- Accès aux résultats

GET `_ml/anomaly_detectors/test-job1/results/buckets`

GET `_ml/anomaly_detectors/test-job1/results/records`

GET `_ml/anomaly_detectors/test-job1/results/influencers`

- Ou

GET `.ml-anomalies-*/_search`

# Détection d'anomalies

Single et multi-metric Jobs

Autres jobs

API et jobs avancés

**Optimisations**

Analyse de cause et jobs multiples

Alertes

Prévisions

# Comptage résumé

- Si les données dans notre index sont déjà agrégées et qu'un champ indique le nombre d'évènements agrégés, il est avantageux d'indiquer la propriété ***summary\_count\_field\_name*** dans la définition du job
- Dans ce cas EL-ML ne compte pas les documents mais utilise ce champ pour trouver le nombre d'occurrence et appliquer les fonctions

# Agrégation de données

- Utiliser une agrégation dans la query du datafeed peut améliorer les performances d'un job
- L'agrégation doit
  - Inclure une agrégation de haut-niveau de type **date\_histogram**
  - Les nom des agrégations calculant des métriques doivent correspondre au nom du champ sur lequel elle s'applique
  - Pour les agrégation de type term, le paramètre size doit correspondre à la cardinalité du champ
  - Les champs influenceurs et partion doivent être inclus
- Du côté du job et du datafeed :
  - Le bucket span doit être divisible par l'agrégation date\_histogram
  - La fréquence d'interrogation doit également être divisible  
=> Pas de date histogram sur les mois !

# Example

```
PUT _ml/anomaly_detectors/kibana-sample-data-flights
{
  "analysis_config": {
    "bucket_span": "60m",
    "detectors": [{ "function": "mean", "field_name": "responsetime", "by_field_name": "airline" }],
    "summary_count_field_name": "doc_count"
  },
  "data_description": { "time_field": "time" },
  "datafeed_config": {
    "indices": ["kibana-sample-data-flights"],
    "aggregations": {
      "buckets": {
        "date_histogram": {
          "field": "time",
          "fixed_interval": "360s",
          "time_zone": "UTC"
        },
        "aggregations": {
          "time": {
            "max": { "field": "time" }
          },
          "airline": {
            "terms": {
              "field": "airline",
              "size": 100
            },
            "aggregations": {
              "responsetime": {
                "avg": {
                  "field": "responsetime"
                }
              }
            }
          }
        }
      }
    }
  }
}
```



# Adapter les données du datafeed

- En utilisant les *runtime fields* et le scripting *Painless*, on peut adapter les données d'entrées aux besoins de l'analyse.
  - Concaténer des strings, les passer en minuscule
  - Faire du remplacement de chaîne
  - Ajouter des champs numériques
  - Transformer en geo\_point
  - ...

# Example

```
PUT _ml/anomaly_detectors/test1
```

```
{
  "analysis_config":{
    "bucket_span": "10m",
    "detectors":[ { "function":"mean", "field_name": "total_error_count" }
    ]
  },
  "data_description": { "time_field":"@timestamp" },
  "datafeed_config":{
    "datafeed_id": "datafeed-test1",
    "indices": ["my-index-000001"],
    "runtime_mappings": {
      "total_error_count": {
        "type": "long",
        "script": {
          "source": "emit(doc['error_count'].value + doc['aborted_count'].value)"
        }
      }
    }
  }
}
```

# Affiner l'analyse

- ELK ML fournit 2 techniques pour exclure de l'analyse certaines données
  - Au niveau détecteur, ***exclude\_frequent*** permet d'exclure les entités trop fréquentes dans les champs over et by pour le calcul du score d'anomalie
  - Au niveau détecteur, les règles personnalisées permettent que les données ne mettent pas à jour les résultats ou le modèle si certaines conditions surviennent

# Règles personnalisées

- Les règles personnalisées décrivent quand un détecteur doit effectuer une certaine action plutôt que de suivre son comportement par défaut.
- Elles s'appliquent en fonction de ***scopes*** (portée) ou de ***conditions***
- Elles peuvent s'appuyer sur des objets ***filter*** stockés au préalable dans EL, qui listent des valeurs à inclure ou exclure

# Structure d'une règle personnalisée

- **scope** (optionnel) : Par défaut, la portée inclut toutes les séries de donnée. Il est possible de spécifier des champs *by\_field\_name*, *over\_field\_name* ou *partition\_field\_name*. Le scope définit également le filtre à appliquer :
  - **filter\_id** : L'id du filtre à utiliser
  - **filter\_type** : include ou exclude
- **actions** [ ] : L'ensemble des actions à déclencher lorsque la règle s'applique :
  - **skip\_result** : Le résultat ne sera pas créé. Ceci est la valeur par défaut. Le modèle sera mis à jour.
  - **skip\_model\_update** : La valeur de cette série ne sera pas utilisée pour mettre à jour le modèle. Les résultats seront créés comme d'habitude.
- **conditions** ([ ] optionnel) : Plusieurs conditions sont combinées avec un ET logique.
  - **applies\_to** : Spécifie la propriété de résultat à laquelle la condition s'applique. Les options disponibles sont *actual*, *typical*, *diff\_from\_typical*, *time*.
  - **operator** : Spécifie l'opérateur de condition. Les options disponibles sont *gt*, *gte*, *lt* et *lte*
  - **value** : La valeur comparée au champ *apply\_to* à l'aide de l'opérateur

# Exemple

- Création d'un filtre

```
PUT _ml/filters/safe_domains
```

```
{  
  "description": "Our list of safe domains",  
  "items": ["safe.com", "trusted.com"]  
}
```

# Exemple

- Création d'un détecteur avec une règle personnalisée utilisant un filtre

PUT \_ml/anomaly\_detectors/dns\_exfiltration\_with\_rule

```
{
  "analysis_config" : {
    "bucket_span": "5m",
    "detectors" : [{
      "function": "high_info_content",
      "field_name": "subdomain",
      "over_field_name": "highest_registered_domain",
      "custom_rules": [{
        "actions": ["skip_result"],
        "scope": {
          "highest_registered_domain": {
            "filter_id": "safe_domains",
            "filter_type": "include"
          }
        }
      }]
    }]
  },
  "data_description" : {
    "time_field": "timestamp"
  }
}
```

# Exemple avec conditions

```
PUT _ml/anomaly_detectors/rule_with_range
{
  "analysis_config" : {
    "bucket_span": "5m",
    "detectors" : [{
      "function": "count",
      "custom_rules": [{
        "actions": ["skip_result"],
        "conditions": [
          {
            "applies_to": "actual",
            "operator": "gt",
            "value": 30
          },
          {
            "applies_to": "actual",
            "operator": "lt",
            "value": 50
          }
        ]
      }]
    }]
  },
  "data_description" : {
    "time_field": "timestamp"
  }
}
```



# Détection d'anomalies

Single et multi-metric Jobs

Autres jobs

API et jobs avancés

Optimisations

**Analyse de cause**

Alertes

Prévisions

# Anomalie et KPI

- Il est naturel que la détection d'anomalie se base sur les KPI d'un service.
- Ces KPI peuvent être aussi diverses que :
  - Axé utilisateur : Temps de réponse, Nombre d'erreurs
  - Opérations : Disponibilité, délai moyen de réparation
  - Métier : Transactions à la minute, CA/Bénéfices , Nombre d'utilisateurs actifs
- Ces KPIs sont souvent présentés dans des tableaux de bord (Kibana)

# Limitations des dashboards Kibana

- L'inspection manuelle de ces tableaux de bord a quand même des limitations :
  - Interprétation: difficulté à comprendre la différence entre fonctionnement normal et anormal, à moins que cette différence ne soit déjà comprise intrinsèquement.
  - Défis d'échelle: Le nombre de KPIs peut être nombreux résultant dans des tableaux de bord surchargés
  - Manque de proactivité: Les tableaux de bord n'ont pas leurs métriques liées aux alertes, ce qui nécessite une surveillance constante.

# Contexte global

- Toutefois, pour avoir une vision plus globale de ce qui peut contribuer à un problème opérationnel, l'analyse doit être élargie aux systèmes et aux technologies sous-jacents supportant l'application.  
=> Beats (Metric, Packet), APM
- Les données remontées par les beats sont souvent segmentées (ex host, server) mais cette segmentation ne correspond pas toujours à une segmentation applicative.  
=> Pour préparer un bon bon contexte d'analyse, il faut souvent retravailler ses données

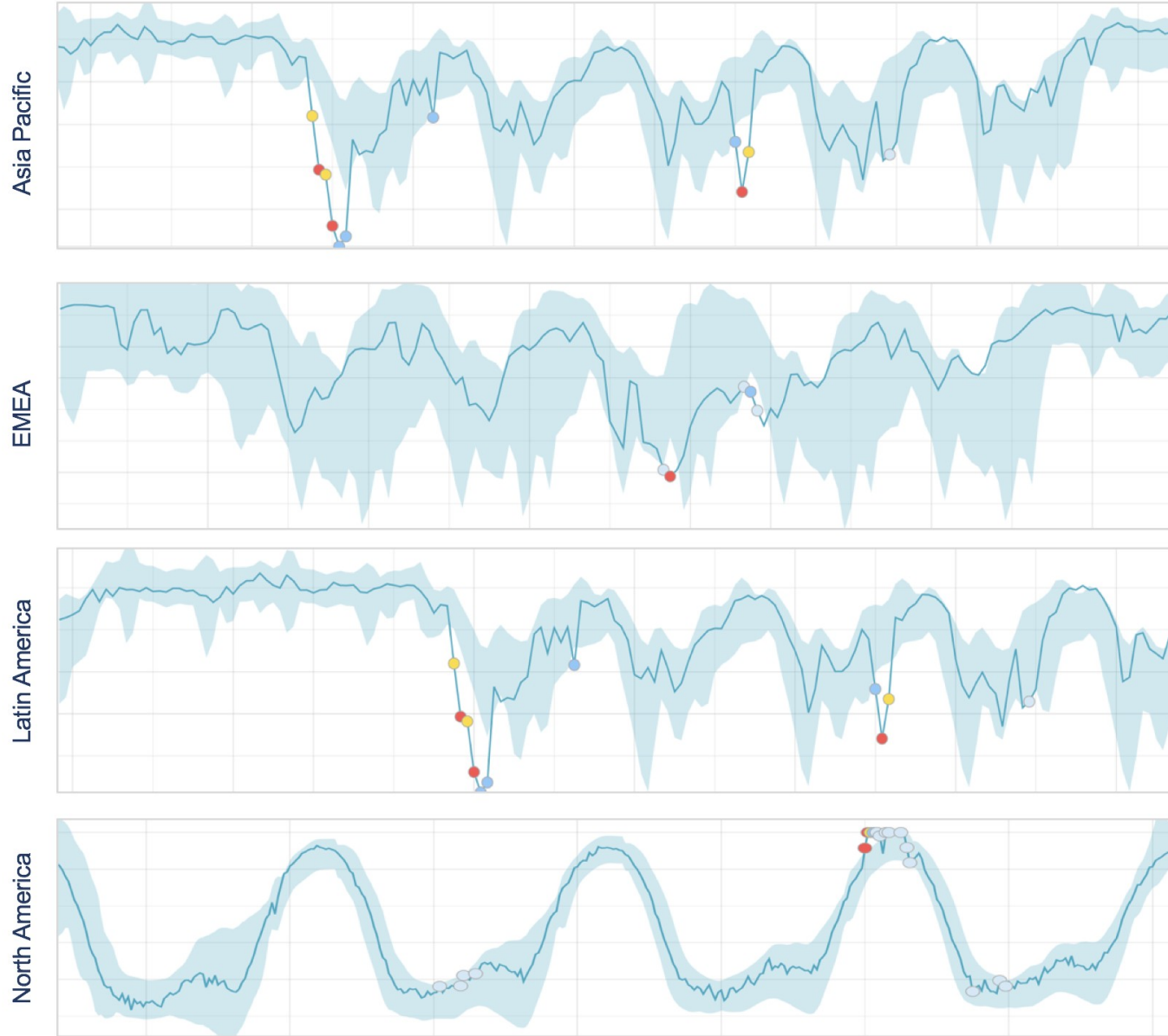
# Segmentation et enrigissement des données

- Les données des index bruts peuvent être segmentées via :
  - La requête DSL du datafeed
  - Via des requêtes enregistrées Kibana
- *logstash* et ses filtres peut ajouter des données qui pourront être utilisées :
  - Comme filtre
  - Comme catégorisation
  - Comme influenceur
- Des champs scriptés peuvent être rajoutés au niveau du datafeed

# Tirer parti des informations contextuelles

- Une fois les données efficacement organisées, 2 méthodes principales sont disponibles pour en tirer parti :
  - Diviser l'analyse (split) en séparant les données pour identifier des modèles comportementaux distincts.  
Ex : Une région, un type de serveur, etc ...
  - Et les influenceurs.

# Example Split



# Influenceur

- Un **influenceur** est la valeur d'un champ que ML identifie comme responsable de l'existence de l'anomalie, ou du moins qui a eu une contribution significative.
- La recherche d'influenceurs se produit après que ML ait découvert l'anomalie :
  - ML examine systématiquement toutes les valeurs de chaque influenceur.
    - Pour chaque valeur distincte, il supprime les données correspondantes du bucket span.
    - Et si les données restantes ne sont plus anormales, alors la contribution de cette valeur est marquée comme influente.



# Score de l'influenceur

- Un **score d'influence** est attribué à chaque valeur du champ influenceur  
Plus le score est élevé, plus cette entité aura contribué ou sera responsable des anomalies.
- Les scores sont calculés pour chaque détecteurs du job et sont rassemblés dans la même vue

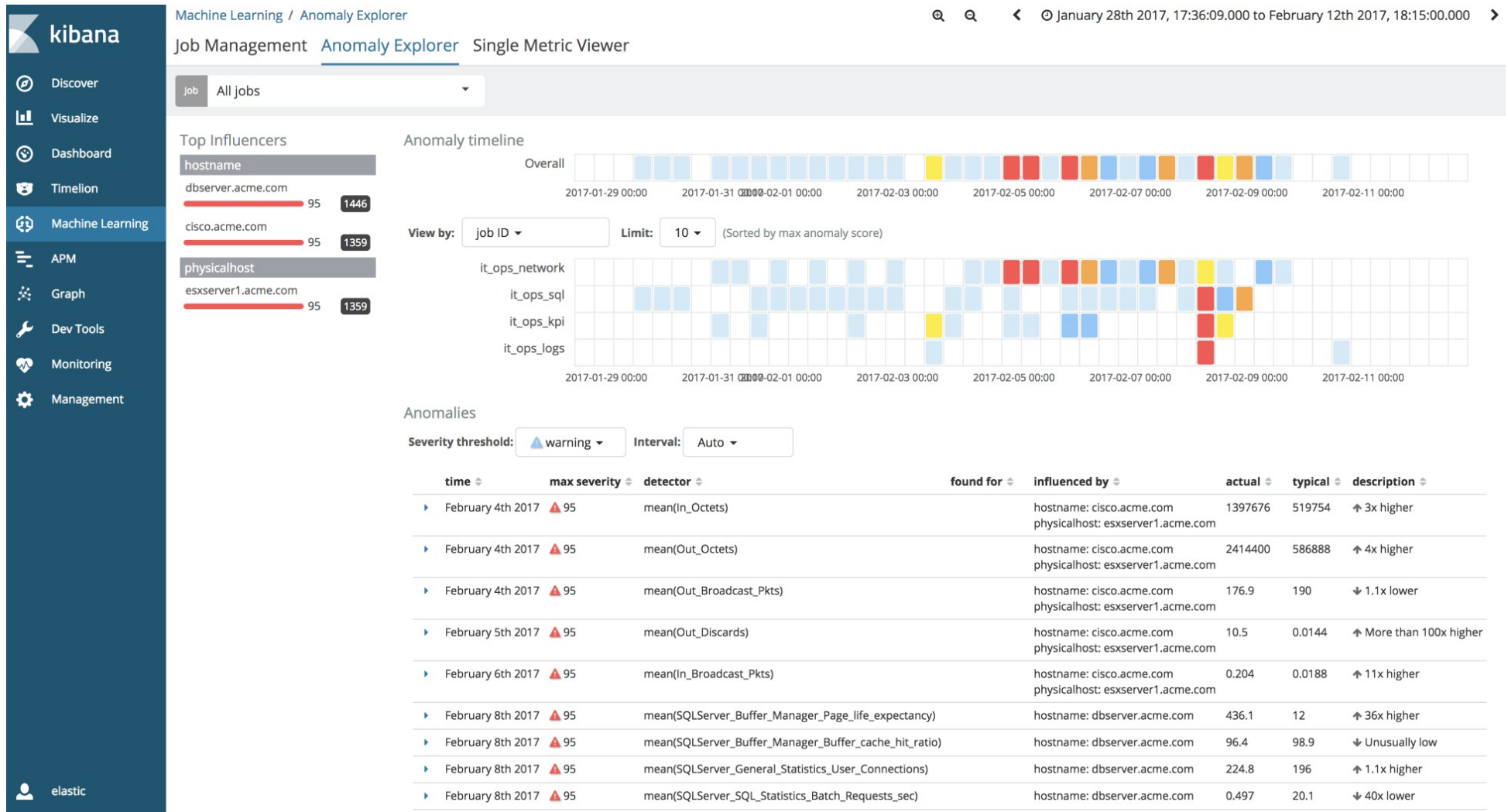
# Corrélation de données

- Pour analyser la cause d'une anomalie dans un KPI, il est souvent nécessaire de configurer d'autres jobs surveillant d'autres métriques
- Par exemple, pour un magasin on-line :
  - Comptage (1 minute) du volume de transactions (le KPI)
  - Logs applicatifs du moteur de transaction (filebeat/logstash/Categorisation de message)
  - Mesures de performances SQL du Serveur de base de données associé au moteur de transactions (metricbeats avec module bas de données)
  - Mesures de performance d'utilisation du réseau (packetbeat)

# Corrélation de données

- L'explorateur d'anomalies permet de superposer différents jobs sur la même période de temps
  - Avec les couloirs temporels, il est facile d'identifier des corrélations entre les anomalies des jobs
- Les tops influenceurs sont également présentés, les chiffres correspondent à :
  - Le score max de l'influenceur dans un bucket
  - Le score total sur tous les buckets de la sélection temporelle
- Les influenceurs qui sont communs à tous les jobs ont leur score additionné  
=> ce qui les font monter dans la top-list

# Corrélation



# Détection d'anomalies

Single et multi-metric Jobs

Autres jobs

API et jobs avancés

Optimisations

Analyse de cause

**Alertes**

Prévisions

# Introduction

- Les alertes Kibana prennent en charge le machine learning.
- La règle déclenchant la règle peut s'exprimer en fonction :
  - D'une détection d'anomalie
  - Du statut d'un job ML
- Par exemple, une règle qui vérifie toutes les 15 minutes si des anomalies critiques ont été détectées et si oui envoie un mail
- Les alertes se créent :
  - Soit dans **Stack Management** → **Rules**
  - Ou dans l'application Machine Learning, à partir de l'assistant de création de job

# Règle

- Une règle est basée sur un score d'anomalie s'appliquant sur les différents types d'un d'analyse :
  - **Bucket** : Score d'anomalie atteint pour le bucket
  - **L'enregistrement** : Anomalie individuelle
  - **Influenceur** : Score d'influence supérieur au seuil

# Anomaly detection

Alert when anomaly detection jobs results match the condition. [Learn more](#)

Select job

high\_sum\_total\_sales\_1708561338743

Result type

Bucket

How unusual was the job within the bucket of time?

✓ Selected

Record

What individual anomalies are present in a time range?

Select

Influencer

What are the most unusual entities in a time range?

Select

Severity75

Include interim results

Advanced settings

Lookback interval

Time interval to query the anomalies data during each rule condition check. By default, is derived from the bucket span of the job and the query delay of the datafeed.

Lookback interval

123m

Number of latest buckets

The number of latest buckets to check to obtain the highest anomaly.

Number of latest buckets

1

Check the rule condition with an interval

1y

Test

Check every1

minute



# Détection d'anomalies

Single et multi-metric Jobs

Autres jobs

API et jobs avancés

Optimisations

Analyse de cause

Alertes

**Prévisions**

# Introduction

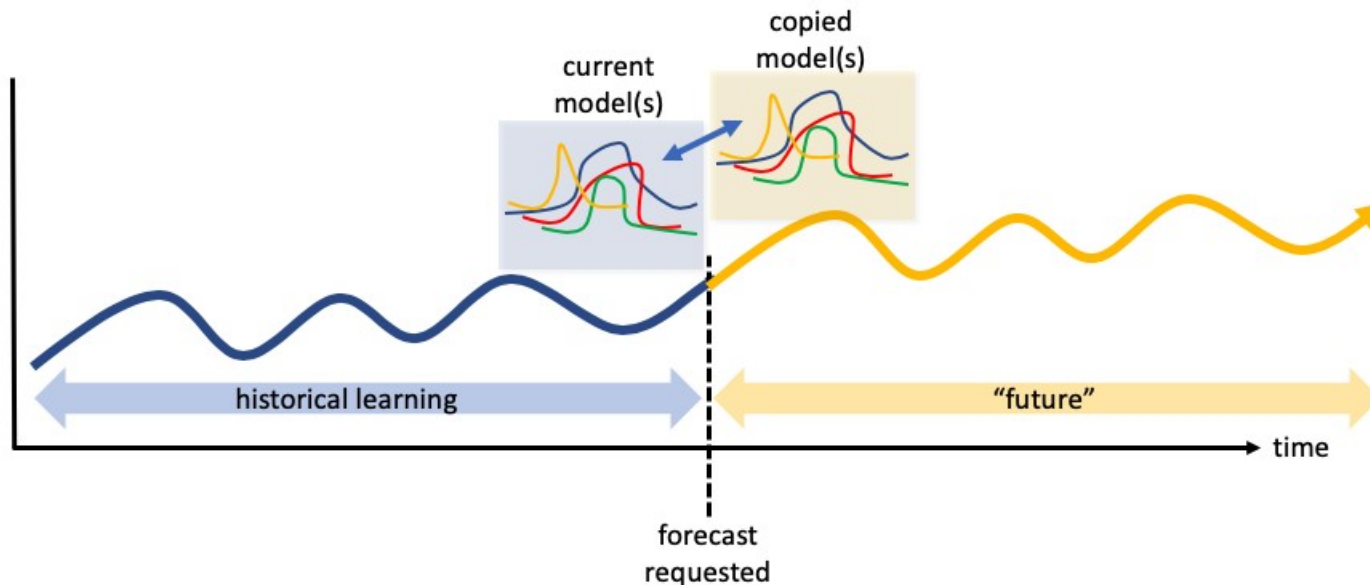
- La prévision est une extension naturelle de la modélisation comportementale d'ELK- ML. Il est donc possible de projeter ces informations dans le temps et de prédire le comportement futur.
- Mais attention, il n'est pas toujours possible de prédire une tendance si un facteur externe inconnu est en jeu (En IT, configuration manuelle incorrecte, matériel défaillant, etc.).
- On peut donc utiliser une analyse probabiliste pour donner la meilleure estimation de l'avenir, mis à part les facteurs externes possibles.

# Cas d'utilisation

- Avec EL-ML, il y a 2 cas d'utilisation des prévisions :
  - **Axé sur la valeur** : Extrapoler une série chronologique dans le futur pour comprendre une valeur future probable.  
Ex : "combien de widgets vais-je vendre par jour dans deux mois?"
  - **Axé sur le temps** : Obtenir une probabilité qu'une valeur atteigne un seuil à un temps donné  
Ex : "Est-ce que je compte atteindre une utilisation de 80% dans la semaine prochaine?"

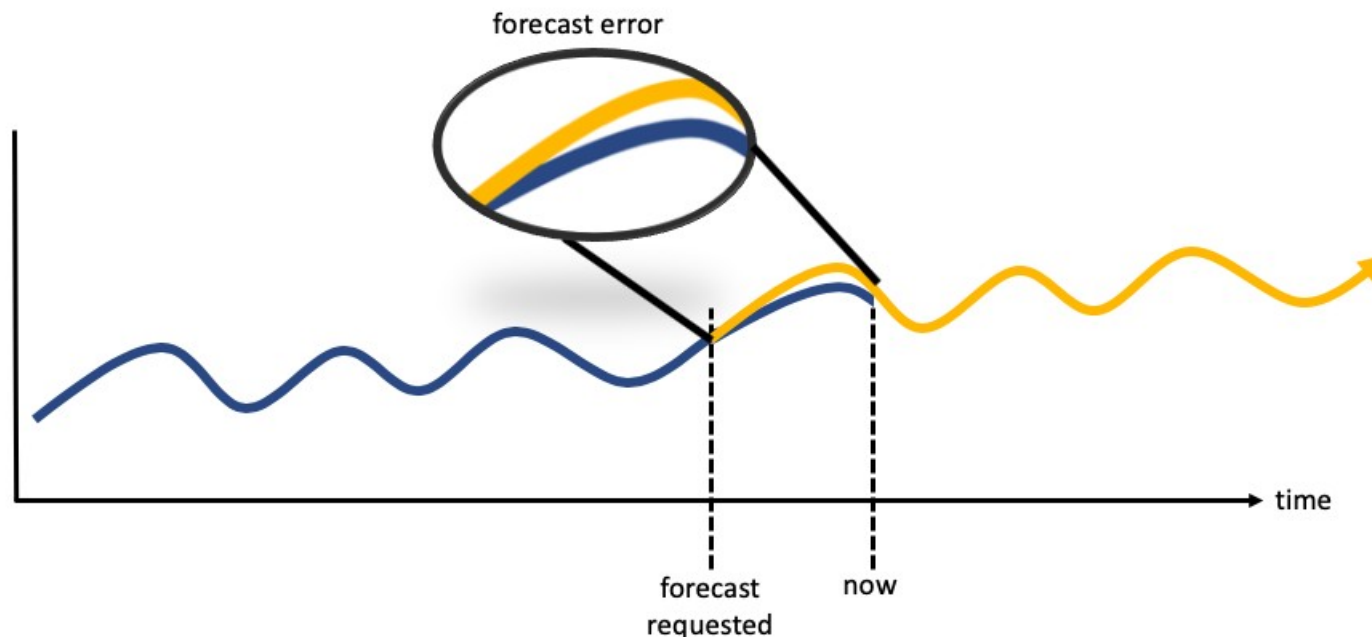
# Processus séparés

- Lorsqu'une prévision est demandée, une copie des modèles du job est créée et un processus séparé est utilisé pour extraire ces modèles et les extrapoler dans le "futur".
- Ce processus est exécuté en parallèle pour ne pas interférer avec les modèles originaux et leur évolution.



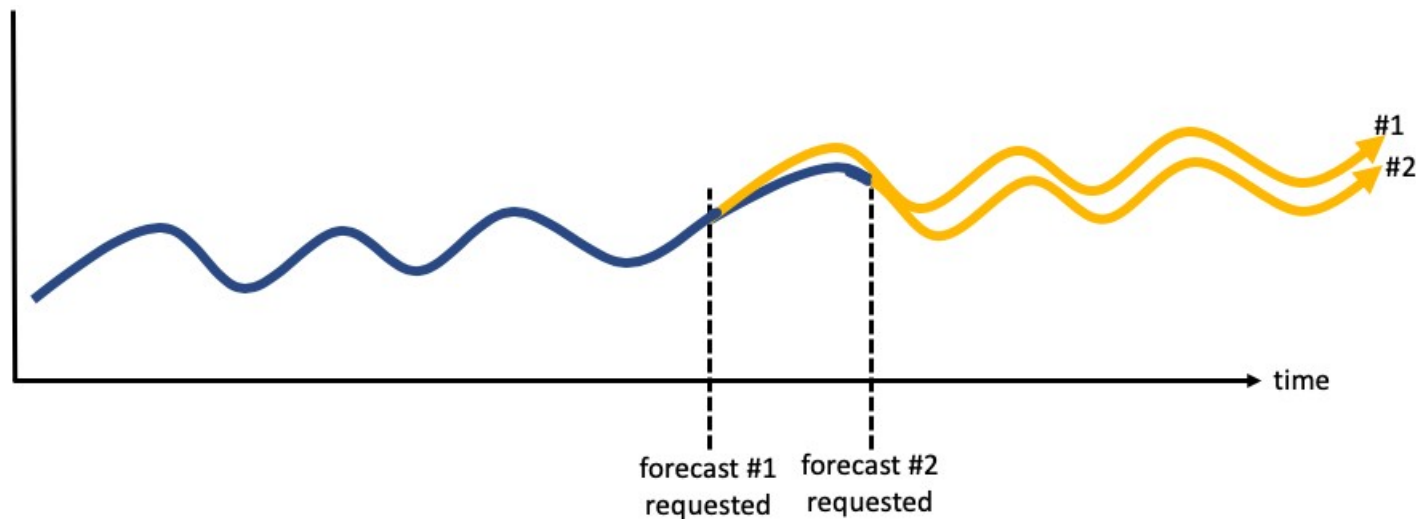
# Comparaison avec le réel

- Le job ML analysant les données réelles se poursuit (s'il fonctionne en temps réel) et qu'après un certain laps de temps, une différence peut apparaître entre la valeur prédite et la valeur réelle



# Plusieurs prévisions

- Plusieurs prévisions peuvent être demandées par l'utilisateur. Chaque prévision est stockée séparément.



# Exécution

- Indiquer une durée :
  - limité à 8 semaines pour le moment
  - Ne pas indiquer une durée plus grande que l'historique
  - L'historique doit avoir un minimum de 3 cycles de pattern périodiques
- Si le job est configuré pour afficher plusieurs séries temporelles, la prévision s'exécute pour tous les détecteurs et les partitions de données
- Les résultats ont une durée de vie par défaut de 14 jours. Après cela, les résultats sont définitivement supprimés.  
Il est possible d'indiquer un autre délai d'expiration via la ressource REST *\_forecast*,

# Single Metric

Machine Learning / Single Metric Viewer

Untitled Workpad - Kibana

Auto-refresh

January 31st 2017, 19:00:00.000 to March 1st 2017, 00:00:00.000

Job Management Anomaly Explorer **Single Metric Viewer** Data Visualizer Settings

Job a\_forecast\_example

Detector: sum(amount)

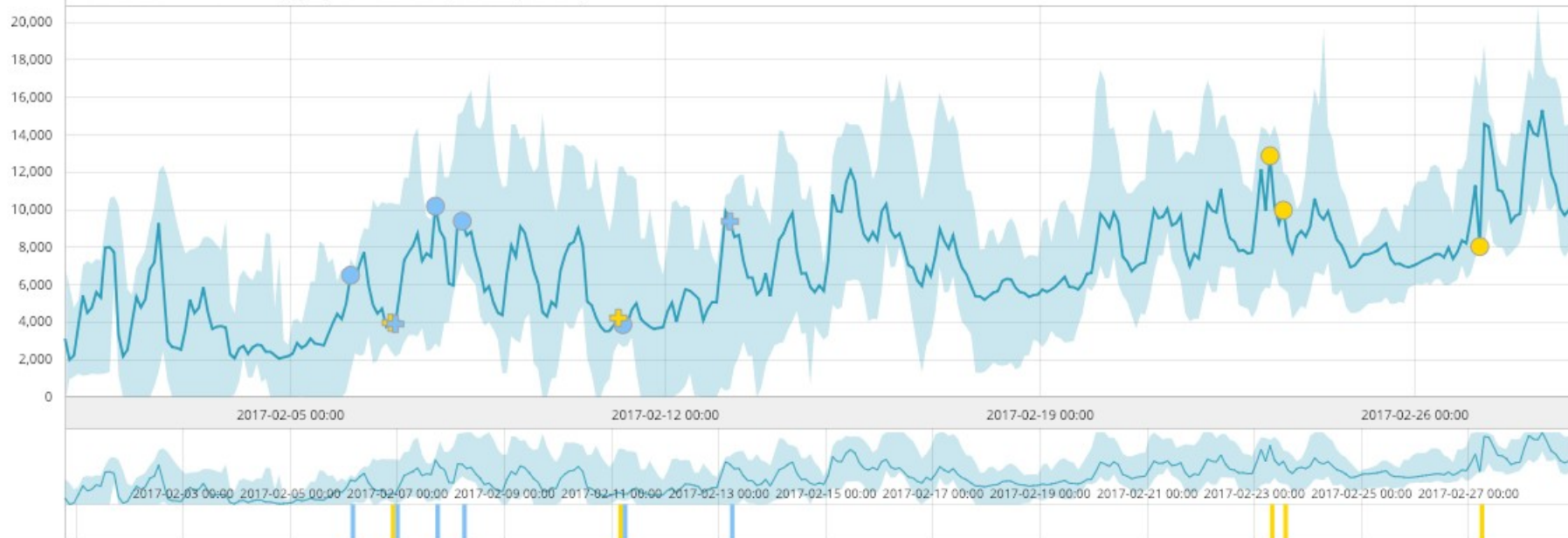


Forecast

Single time series analysis of sum amount

☒ show model bounds

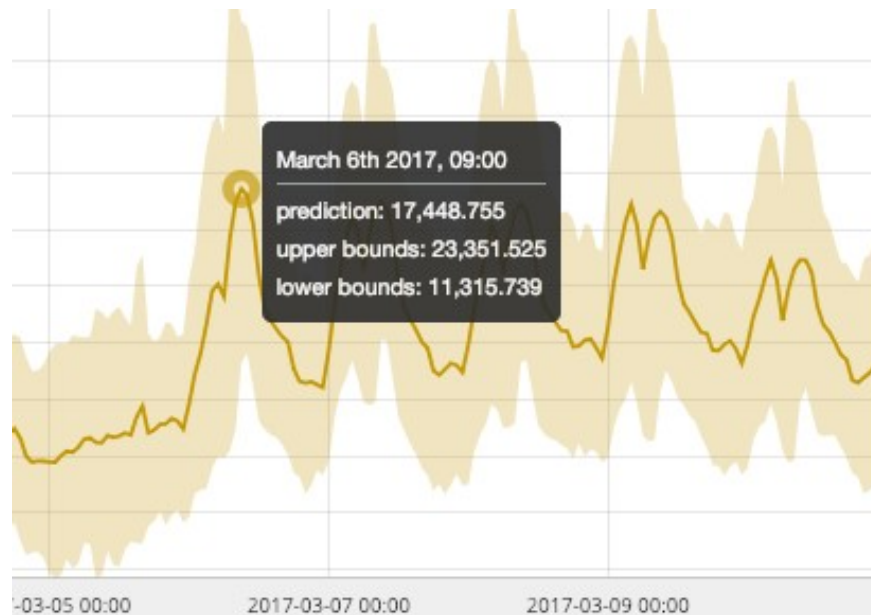
Zoom: auto 12h 1d 1w 2w (aggregation interval: 2h, bucket span: 15m)





# Résultats de prévisions

- Une fenêtre contextuelle, disponible pour chaque point de données des résultats, affiche la valeur de prédiction, la limite supérieure et la valeur de limite inférieure. (intervalle de confiance du 95 e centile.)



# API

- Les résultats sont également accessible via l'API search

GET .ml-anomalies-\*/\_search

```
{
  "query": {
    "bool": {
      "filter": [
        {"term": {"timestamp": "1488808800000"}},
        {"term": {"result_type": "model_forecast"}},
        {"term": {"job_id": "a_forecast_example"}}
      ]
    }
  }
}
```

# Réponse

```
{
  ... "hits" : [
    {
      "_index" : ".ml-anomalies-shared",
      "_type" : "doc",
      "_id" :
      "a_forecast_example_model_forecast_i2DxbGgBITRq2rXM21p4_1488808800000_900_0_961_0",
      "_score" : 0.0,
      "_source" : {
        "job_id" : "a_forecast_example",
        "forecast_id" : "i2DxbGgBITRq2rXM21p4",
        "result_type" : "model_forecast",
        "bucket_span" : 900,
        "detector_index" : 0,
        "timestamp" : 1488808800000,
        "model_feature" : "'bucket sum by person'",
        "forecast_lower" : 11315.739312779506,
        "forecast_upper" : 23080.83486433322,
        "forecast_prediction" : 17198.287088556364
      }
    }
  ]
}
```

# Prévision axée sur le temps

- Pour une prévision axée sur le temps, il faut exécuter une query
- Par exemple : Query avec elastic-sql

```
POST /_xpack/sql?format=txt
```

```
{  
  "query": "SELECT timestamp FROM \".ml-anomalies-*\" WHERE  
job_id='a_forecast_example' AND result_type='model_forecast' AND  
forecast_prediction>'17700' ORDER BY timestamp DESC"  
}
```

- Réponse

```
timestamp
```

```
-----
```

```
2017-03-06T14:45:00.000Z
```

# Plusieurs séries temporelles

- Une prévision peut également être démarrée via l'API
- Par exemple :

POST

```
_xpack/ml/anomaly_detectors/web_traffic_per_country/_forecast  
{  
  "duration": "7d"  
}
```

Réponse :

```
{  
  "acknowledged" : true,  
  "forecast_id" : "DGT6bWgBITRq2rXMb1Rr"  
}
```

# Analyse de trame de données

## **Introduction**

Détection de valeurs anormales

Régression

Classification

Inférence

# Introduction

- Pour ce type d'analyse les données sources doivent être structurées sous la forme d'un tableau à 2 dimensions
- Un module de transformation est fourni pour restructurer les données provenant d'index
- Permet différentes analyses et l'annotation des données avec les résultats

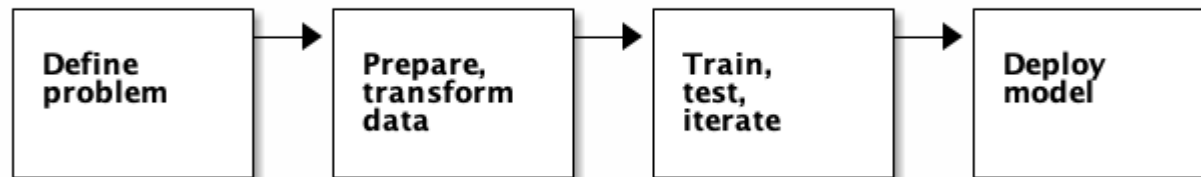
# Analyses

- Les différentes analyses disponibles :
  - Détection de valeurs anormales (non supervisés)
  - Prévisions sur des données
  - Classification des données
  - Inférence permet de confronter les modèles ML entraînés vis à vis des données réelles



# Apprentissage supervisé

- Permet de créer un modèle ML en fournissant des exemples d'entraînement
- Le modèle peut être alors utilisé pour des prédictions
- Le workflow est alors :



# Définition du problème

- Quels types de modèles souhaitez-vous découvrir dans vos données ?
- Quel type de valeur voulez-vous prédire : une catégorie ou une valeur numérique ?
- ELK-ML propose 2 types d'analyse :
  - **regression**: prédit des valeurs **numériques continues**  
Ex : le temps de réponse d'une requête Web.
  - **classification** : prédit des valeurs **discrètes**  
Ex : Requête DNS malveillant ou normale.

# Préparation des données

- Pour les analyses supervisées, il faut fournir un jeu de données **étiqueté, important**
  - Par exemple, pour entraîner un modèle de classification qui décide si un e-mail est un spam ou non, on a besoin d'un ensemble de données étiqueté contenant d'exemple de bons emails et de spams.
- Les données doivent être structurées sous un format tabulaire

# Entraînement du modèle

- L'entraînement est un processus itératif : chaque itération est suivie d'une évaluation pour évaluer l'efficacité du modèle.
  - La première étape consiste à définir les champs pertinents dans l'ensemble de données
  - Ensuite, séparer les données en un ensemble pour entraîner et un ensemble pour évaluer le modèle
- Les données d'apprentissage sont transmises à l'algorithme d'apprentissage.

Le modèle prédit la valeur et la compare à la vérité terrain  
Puis le modèle est affiné pour rendre les prédictions plus précises.

# Déploiement du modèle

- Après avoir obtenu une bonne performance, le modèle est déployé et utilise les nouvelles données
- L'inférence permet de faire des prédictions sur les nouvelles données en utilisant :
  - un processeur dans une pipeline d'ingestion
  - Dans un processus de transformation continue (Roll-up d'index)
  - Une agrégation lors d'une recherche

# Jobs d'analyse

- Les jobs d'analyse sont constitués de 4 ou 5 phases :
  - **Réindexation** (API Reindex) : Copie des documents de l'index source vers un autre index (possibilité de changer le settings et le mapping)
  - **Chargement des données** : Chargement des données et conversions dans la structure demandée par l'analyse
  - **Analyse** : Une seule phase pour la détection de valeurs anormales, plusieurs phases pour la régression et la classification :
    - Identification des champs significatifs
    - Affinement de paramètres (les hyperparamètres)
    - Entraînement du modèle
  - **Écriture des résultats** : Les résultats issus de l'analyse sont réécrits dans l'index
  - **Inférence** : Pour la régression et la classification, validation du modèle sur les données d'évaluation

# Recommandations

- Démarrer petit et itérer rapidement
- Fournir un petit pourcentage de données pour l'apprentissage
- Désactiver le calcul de l'importance des champs (Feature importance) pour réduire le temps d'exécution de l'analyse
- Optimiser le nombre de champs inclus dans l'analyse
- Augmenter le nombre de threads (par défaut 1)
- Optimiser la taille de l'index source (par des filtres par exemple)
- Configurer manuelle les *hyperparamètres*

# Évaluation des résultats

- ELK fournit une API pour évaluer les résultats d'analyse  
`POST _ml/data_frame/_evaluate`
- Cela permet de comprendre les distributions d'erreurs et identifier les points où le modèle d'analyse de trame de données fonctionne bien ou de manière moins fiable
- Les métriques fournis sont :
  - La matrice de confusion
  - La précision
  - Le recall
  - La courbe caractéristique de fonctionnement du récepteur (ROC).



# Matrice de confusion

- Une matrice de confusion fournit quatre mesures :
  - **Vrais positifs** (TP) : membres de la classe que l'analyse a identifiés comme étant des membres de la classe.
  - **Vrais négatifs** (VN) : Les membres du groupe que l'analyse a correctement identifiés comme n'étant pas membres du groupe.
  - **Faux positifs** (FP) : Les membres que l'analyse a identifiés à tort comme des membres du groupe.
  - **Faux négatifs** (FN) : Les membres du groupe que l'analyse a identifiés à tort comme n'étant pas membres du groupe.
- Les résultats de l'analyse ne sont pas des valeurs exactes mais des probabilités. Il faut alors donner des seuils où le considère la valeur comme exacte
- L'API fournit par défaut des matrices de confusion prenant comme seuils 0,25 ; 0,5 et 0,75

# Précision et recall

- Les valeurs de **précision** et de **recall** résument les performances de l'algorithme sous la forme d'un nombre unique qui facilite la comparaison des résultats de l'évaluation
  - La précision indique combien de points que l'algorithme a identifiés comme membres de la classe étaient en fait des membres de la classe. C'est le nombre de vrais positifs divisé par la somme des vrais positifs et des faux positifs ( $TP/(TP+FP)$ ).
  - Recall répond à une question légèrement différente. Cette valeur indique combien de points de données qui sont des membres réels de la classe ont été correctement identifiés en tant que membres de la classe. C'est le nombre de vrais positifs divisé par la somme des vrais positifs et des faux négatifs ( $TP/(TP+FN)$ ).
- Les seuils s'appliquent également pour ces 2 mesures

# ROC

- **ROC** est un tracé qui représente les performances du processus de classification binaire à différents seuils.
- Il compare le taux de vrais positifs au taux de faux positifs aux différents niveaux de seuil pour créer la courbe.
- À partir du tracé, on peut calculer la valeur de l'aire sous la courbe (AUC), qui est un nombre compris entre 0 et 1.  
Plus la valeur est proche de 1, meilleures sont les performances de l'algorithme.

# Analyse de trame de données

Introduction

**Détection de valeurs anormales**

Régression

Classification

Inférence

# Introduction

- La détection des valeurs aberrantes est une analyse permettant d'identifier les points de données (valeurs aberrantes) dont les valeurs de caractéristiques sont différentes de celles des points de données normaux.
- Les valeurs aberrantes peuvent indiquer des erreurs ou un comportement inhabituel.

# Algorithme

- L'algorithme calcule 4 valeurs pour déterminer si une valeur est aberrante :
  - La distance au Kème voisin
  - La distance des K plus proches voisins
  - Lof (local outlier factor) : Calcule sur la densité, (les voisins proches ont ils également des voisins)
  - Ldof (local distance-based outlier factor) : Ratio, également sur la densité
- Ces 4 valeurs détectent des valeurs aberrantes. ELK-ML, agrège et normalise ces données et fournit alors une probabilité (entre 0 et 1) qu'une valeur soit aberrante.

# Résultats

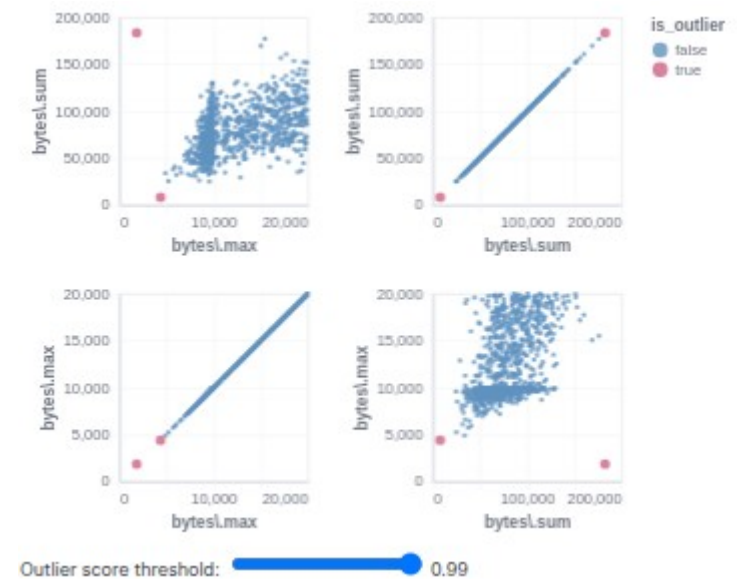
- Les résultats sont présentés sous 2 formes :
  - Forme tabulaire ou les lignes sont triées par ml.outlier\_score
  - Forme de nuage de points

Results Total docs 1001 Feature influence score 0 0.2 0.4 0.6 0.8 1

Columns 7/8 Sort fields 1 Histogram charts

0 documents contain field. ml.feature\_influence

ml.outlier_score	@timestamp.value_count	bytes.max	bytes.sum	top 20 of 1001 categories clientip	request.value_count	Actions
1	100	1,831	183,100	30,156.16.164	100	
0.992	5	4,372	7,998	81.84.213.90	5	
0.989	29	17,722	160,518	164.85.94.243	29	
0.956	24	15,471	176,720	111.237.144.54	24	
0.947	5	12,858	29,086	184.31.17.237	5	
0.875	26	8,671	115,577	50.184.59.162	26	
0.852	21	15,035	169,258	177.120.218.48	21	
0.824	25	18,551	153,285	16.241.165.21	25	
0.79	5	19,326	36,358	148.48.158.9	5	
0.785	15	9,821	104,303	25.181.253.214	15	
0.707	5	11,243	34,419	30.104.245.98	5	
0.686	26	18,851	138,839	236.212.255.77	26	
0.682	7	8,395	36,545	114.88.248.26	7	
0.675	14	6,420	39,069	58.139.164.111	14	
0.674	12	9,735	79,510	214.57.83.208	12	
0.66	10	11,185	47,774	228.130.217.137	10	
0.613	6	12,695	43,930	140.156.212.70	6	
0.612	7	8,836	35,563	50.101.245.255	7	
0.603	23	18,605	105,255	15.225.65.207	23	
0.601	7	9,659	42,927	74.184.0.64	7	
0.564	22	17,148	156,375	210.180.52.151	22	
0.546	15	8,679	82,436	239.23.215.100	15	
0.531	20	12,045	104,111	21.141.237.239	20	



# Feature Influence

- Une autre valeur est calculée durant l'importance : le facteur d'influence d'une feature
- Pour les champs de l'index, ELK-ML évalue l'influence du champ sur le résultat



# Analyse de trame de données

Introduction

Détection de valeurs anormales

**Régression**

Classification

Inférence

# Introduction

- But : Estimer les relations entre les différents champs de vos données, puis faire d'autres prédictions basées sur ces relations
  - Par exemple, prédire la valeur d'achat d'un appartement à partir de sa superficie, sa localisation, son étage, etc.

# Feature variables

- La première étape consiste donc à identifier les champs de notre index qui serviront au modèle pour prédire la valeur d'un autre champ
- Les types supportés sont :
  - Numérique
  - Catégorie : Ensemble fixe de valeurs
  - Booléen

# Apprentissage et évaluation du modèle

- Il faut fournir des ensemble de données étiquetés qui contiennent les feature variables et la variable dépendante
- La régression consiste à identifier une fonction mathématique qui permet de calculer la variable dépendante en fonction des variables de feature  
Algorithme : *eXtreme Gradient boost*
- ELK-ML propose différents indicateurs pour évaluer la performance du modèle.  
Principalement :
  - **mse : Mean Squarred Error**, moyenne des carrés des erreurs entre les prédictions et les valeurs réelle. Plus c'est petit mieux c'est
  - **r\_squared: Coefficient of determination**, métrique relative qui montre à quel point votre modèle capture la relation entre les variables. Plus c'est proche de 1 mieux c'est
- Le modèle créé est stocké dans des index internes d'Elasticsearch
- Des résultats d'impacts des champs (Feature importance) sont également évalués

# Mise en place

- La mise en place du job d'analyse est un processus itératif.
- Après avoir défini le problème, vous devez produire un ensemble de données de haute qualité et créer le job d'analyse appropriée.
- Différentes configurations, paramètres et méthodes de transformation doivent être expérimentés avant d'arriver à un résultat qui réponde au cas d'utilisation.

# Paramètres du job

- La création du job consiste donc à fournir :
  - La variable dépendant
  - Les champs inclus ou exclus de l'analyse
  - La requête sélectionnant les données de l'index source
  - Le pourcentage de document à utiliser dans les données de test

# Indicateurs de performance

- A la suite du job, les indicateurs de performances sont fournies pour les données d'entraînement et les données de test.  
(Generalization)
- Si les indicateurs sont bons sur les données d'entraînement et proche sur les données de généralisation, le modèle est performant

# Mesure de la performance du modèle (2)

- 2 cas où le modèle n'est pas très efficace
  - Le **sous-ajustement** se produit lorsque le modèle ne peut pas capturer la complexité de l'ensemble de données. Les indicateurs sur les données de test ne sont pas bons.
  - Le **surapprentissage** se produit lorsque le modèle est trop spécifique à l'ensemble de données d'apprentissage et capture des détails qui ne se généralisent pas aux nouvelles données.  
Typiquement, une valeur MSE faible sur l'ensemble de données d'entraînement et une valeur MSE élevée sur l'ensemble de données de test



# Analyse de trame de données

Introduction

Détection de valeurs anormales

Régression

**Classification**

Inférence

# Introduction

- Application typique : Accord de prêt bancaire  
En fonction du montant du prêt et du profil de l'emprunteur, accorder ou pas le prêt
- Lors de la création d'un job de classification, il faut spécifier :
  - La variable dépendante : Max de 30 valeurs discrètes
  - Les feature variables : Variables qui influent

# Apprentissage

- Comme pour la régression, il faut fournir un ensemble de données étiquetées divisées en :
  - Un ensemble pour l'apprentissage
  - Un ensemble pour l'évaluation
- La division s'effectue en fournissant un pourcentage
- L'algorithme utilisé est appelé *boosted tree regression model*.

Il utilise des arbres de décision pour prédire la probabilité d'une valeur discrète
- Le modèle est stocké dans un index interne à Elasticsearch

# Résultats

- Le résultat est fourni via 2 indicateurs :
  - ***class\_probability*** : valeur comprise entre 0 et 1, qui indique la probabilité qu'un point donné appartienne à une certaine classe
  - ***class\_score*** : Fonction de *class\_probability* qui a une valeur  $\geq 0$ .  
Il prend en considération l'objectif (défini via la configuration de job `class_assignment_objective`) :
    - Précision : Score pondéré avec les bonnes classifications
    - Recall : Prend en compte également les mauvaises classifications

# Mesure de la performance du modèle

- On peut mesurer l'efficacité du modèle via l'API `POST _ml/data_frame/_evaluate` {  
    `"evaluation" : "classification"`  
}
- Le retour est une matrice de confusion

# Analyse de trame de données

Introduction  
Détection de valeurs anormales  
Régression  
Classification  
**Inférence**

# Introduction

- L'inférence permet d'utiliser les analyses régression ou classification de manière continue.
- Cela veut dire appliquer les modèles prédictifs sur les nouvelles données
- Le modèle doit être déployé, ensuite plusieurs alternatives
  - Utiliser une pipeline d'ingestion
  - Utiliser des agrégations

# Déploiement du modèle

*Machine Learning > Model Management > Trained models*

## Trained Models




Auto refresh 30 seconds


Refresh

Total trained models: 2

Search...

Type ▾

<input type="checkbox"/>	ID ↑	Description	Type	State	Created at	Actions
<input checked="" type="checkbox"/>	.elser_model_1	Elastic Learned Sparse Encoder v1	<span>TECHNICAL PREVIEW</span>	<span>elastic</span>	-	
<input type="checkbox"/>	> flight-delay-min-regression-1692262551907		<span>tree_ensemble</span> <span>regression</span>		Aug 17, 2023 @ 10:55:51.907	
<input checked="" type="checkbox"/>	> lang_ident_model_1	Model used for identifying language from arbitrary input text.	<span>lang_ident</span> <span>classification</span> <span>built-in</span>		Dec 5, 2019 @ 13:28:34.594	

 View training data

 Analytics map

 Deploy model

 Delete model

Rows per page: 10 ▾

< 1 >



# Inference processor

- La configuration d'un processeur contient les champs :
  - **"*model\_id*"** : Le modèle
  - **"*target\_field*"** : Le champ ajouté au document contenant la prédiction
  - **"*inference\_config*"** : Le type d'analyse *regression* ou *classification* avec un bloc dépendant du type

# Examples

```
"inference":{
  "model_id":"my_model_id"
  "inference_config": {
    "regression": {
      "results_field": "my_regression"
    }
  }
}
```

```
"inference":{
  "model_id":"my_model_id"
  "inference_config": {
    "classification": {
      "num_top_classes": 2,
      "results_field": "prediction",
      "top_classes_results_field": "probabilities"
    }
  }
}
```

# Agrégation

- L'inférence peut également être utilisée via une pipeline d'agrégation. Le modèle est référencé dans l'agrégation pour inférer sur le champ de l'agrégation parente

```
{
  "inference": {
    "model_id": "a_model_for_inference",
    "inference_config": {
      "regression_config": {
        "num_top_feature_importance_values": 2
      }
    },
    "buckets_path": {
      "avg_cost": "avg_agg", # avg_cost référence une agrégation parente
      "max_cost": "max_agg"  # avg_agg un champ d'entrée dans le modèle
    }
  }
}
```

MERCI !!

Pour votre attention

# Annexe

## **Calendriers**

Visualisations Kibana  
Alertes

# Calendriers

- EL – ML permet d'indiquer des plages de temps ou le comportement du système est inhabituel (« black friday » par exemple).  
=> Les jobs peuvent être configurés pour exclure ses plages lors de l'analyse
- Les DST Calendars sont une variante particulière des Calendriers. Ils prennent en compte les changements d'heure liés à l'heure d'été

# Annexe

Calendriers  
**Visualisations Kibana**  
Alertes

# URLs personnalisées

- Via des *Advanced jobs*, il est possible d'ajouter des hyperliens dans les vue *Anomaly Explorer* ou *Single Metric Viewer*. Ces liens peuvent être dirigés vers :
  - Un tableau de bord Kibana
  - La page Discovery de Kibana
  - Une URL externe



# Configuration

- Pour chaque URL, on indique :
  - Un label
  - Optionnellement un intervalle de temps.

Edit response\_requests\_by\_app

Job Details

Detectors

Datafeed

Custom URLs

## Create new custom URL

Label ⓘ

My link 1

Link to ⓘ

☐ Kibana dashboard

☒ Discover

☐ Other URL

Index pattern ⓘ

server-metrics\*

Query entities ⓘ

service ✕

Time range ⓘ Interval

interval ▼

2h

Add

Update

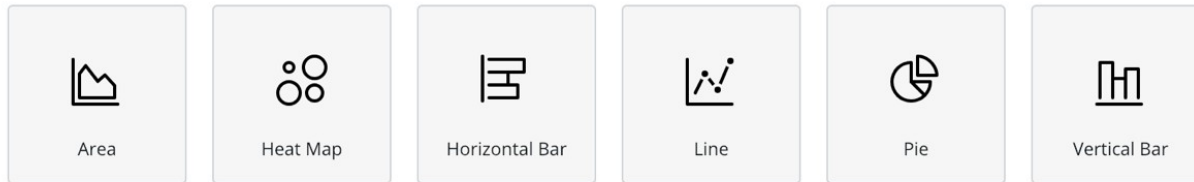
Cancel

# Expression dans l'URL

- L'URL peut contenir des expressions entourées par des **\$** qui sont substituées à l'exécution :
  - Soit par les champs disponibles dans le document anomalie
  - Soit par des valeurs prédéfinies :
    - **\$earliest\$ \$latest\$** : le début et la fin de la période de l'anomalie sélectionnée
    - **\$mlcategoryregex\$ \$mlcategoryterms\$** : Utile lors de la catégorisation de messages
      - **\$mlcategoryregex\$** expression régulière de l'anomalie sélectionnée (champ mlcategory)
      - **\$mlcategoryterms\$** valeur des termes de la catégorie de l'anomalie sélectionnée

# Visualisations de Kibana

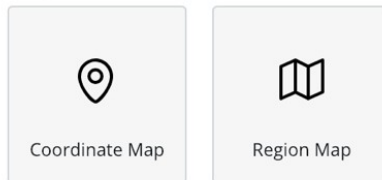
## Basic Charts



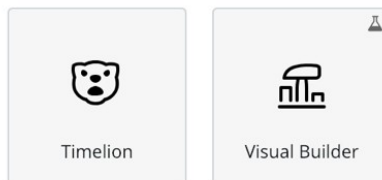
## Data



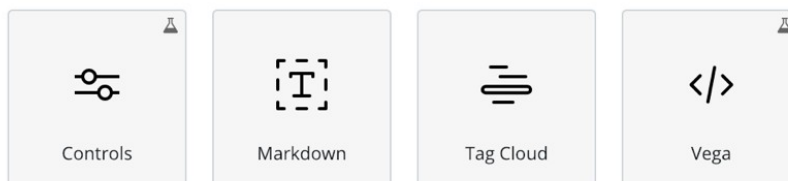
## Maps



## Time Series



## Other



# Visualisations utiles pour ML

- Data Table : Liste de données
- Heat map :
  - Sévérité d'une anomalie
  - Corrélation de plusieurs
- Timelion :
  - Basée sur une expression
  - Permet de combiner plusieurs sources de données
- Time series visual builder
  - Permet les agrégations et les agrégations en pipeline
  - Propose différents types de graphiques
  - Permet les annotations

# Exemple TSVB

- La configuration d'un TSVB contient 3 onglets :
  - **Data** : Configuration de l'agrégation et de la fonction à appliquer
  - **Panel** : Source de données
  - **Annotations** : Possibilité d'utiliser une autre source de données, par exemple le job ML

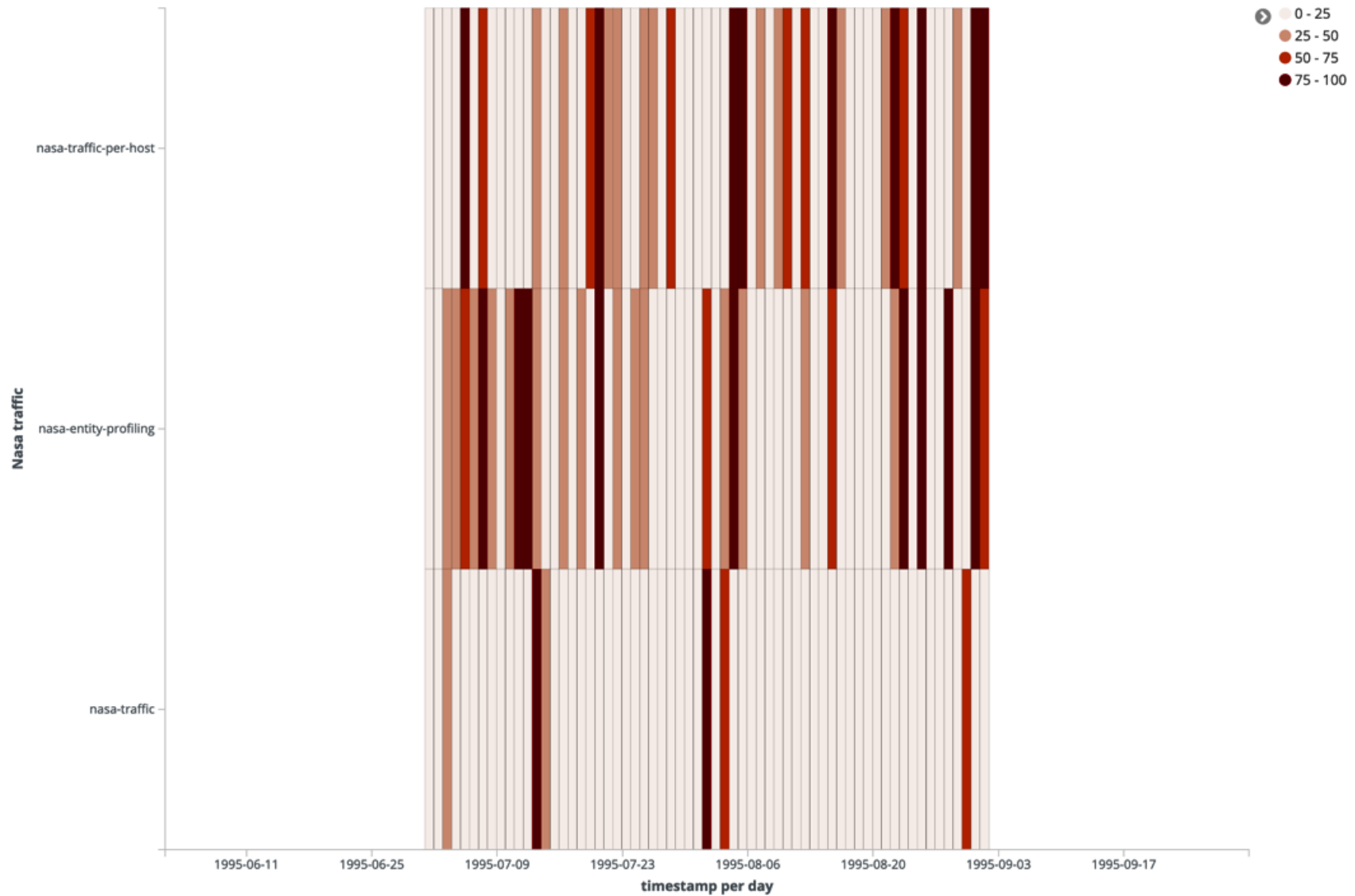
# Affichage des anomalies via les annotations



# Exemple heat map

- Un heat map peut être utilisé pour visualiser toutes les anomalies de tous les jobs ML :
  - Choisir l'index pattern ***.ml-anomalies\****
  - 1 bucket de type Date Histogram
  - 1 sub-bucket de type term sur le ***job id***
  - Metrics : Max ***anomaly\_score***
  - Color schema : Rouge

# Exemple heat map



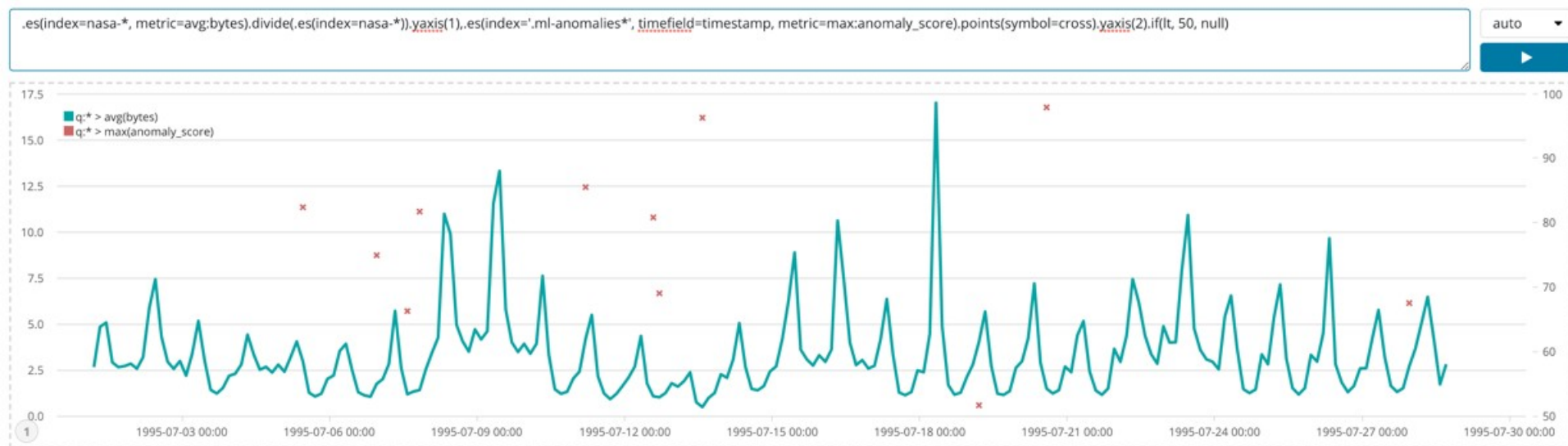


# Timelion

- *Timelion* peut également être utilisé pour combiner dans la même vue les données et les anomalies détectées
- Par exemple afficher la taille moyenne des requêtes d'un site web en même temps que les anomalies de trafic détectées
- L'expression pourrait être :

```
.es(index=nasa-*, metric=avg:bytes)
  .divide(.es(index=nasa-*)).yaxis(1),
.es(index='.ml-anomalies*', timefield=timestamp,
metric=max:anomaly_score)
  .points(symbol=cross).yaxis(2).if(lt, 50, null)
```

# Exemple timelion



# Canvas

- Kibana Canvas, permet de créer des rapports complètement personnalisés
- Dans *Canvas*, les projets sont appelés "présentations", elles sont analogues aux présentations habituelles Powerpoint et peuvent comporter plusieurs pages ...  
Et en plus les données sont dynamiques !
- Canvas peut s'appliquer sur les index d'anomalies

# Workspace

- Canvas propose un workspace où il est possible de placer des éléments directement connectés aux données EL
- Les éléments peuvent être personnalisés en travaillant directement sur le CSS

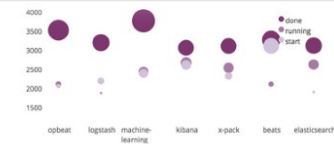
# Éléments

🔍 Filter elements



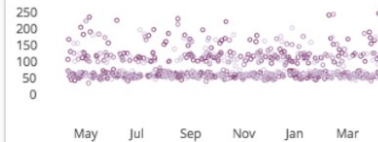
## Area chart

A line chart with a filled body



## Bubble chart

A customizable bubble chart



## Coordinate plot

Mixed line, bar or dot charts

cost #	username #	state #	cost #	price #
22.99	acollinsd9	done	22.99	51
23.43	kphillipsmv	done	23.43	54
21.84	jfloresn0	running	21.84	71
23.12	jcarpenteric	done	23.12	53
21.93	sbutlerb1	running	21.93	67
23.13	agardnerd0	done	23.13	60

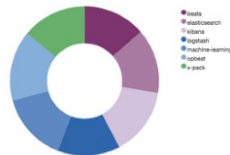
## Data table

A scrollable grid for displaying data in a tabular format

```
"time": 1460444400000,
"username": "swhitejp",
},
{
  "age": 74,
  "cost": 22.69,
  "country": "CN",
  "price": 79,
```

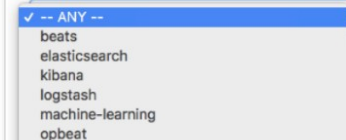
## Debug

Just dumps the configuration of the element



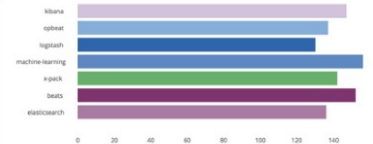
## Donut chart

A customizable donut chart



## Dropdown filter

A dropdown from which you can select values for an "exactly" filter



## Horizontal bar chart

A customizable horizontal bar chart

Dismiss

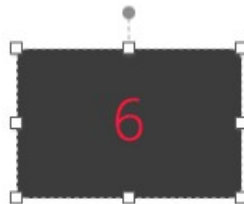
# Canvas expression

- Derrière chaque élément configuré dans l'interface, il y a une **expression Canvas** éditable via l'*Expression editor* qui définit comment ce composant est construit
- Une expression est composée de plusieurs fonctions chaînées par |.
- Les fonctions disponibles permettent
  - De définir des jeux de données : démo, query, expression timeline
  - De filtrer, transformer les données
  - d'exécuter des fonctions mathématiques complexes
  - De la logique
  - Définir les axes de visualisation, le style CSS, formater les dates, ...

# Example

- Exemple Markdown

```
filters
| essql
query="SELECT timestamp, anomaly_score FROM \".ml-anomalies-*\" WHERE
result_type = 'bucket' AND anomaly_score > 10 AND job_id = 'nginx-traffic'"
| markdown "
#
#
# {{rows.length}}"
| render css="h1 {
text-align: center;
color: #ff1744;
}
"
containerStyle={containerStyle backgroundColor="#444444" border="5px none
#FFFFFF" borderRadius="7px" padding="px"}
```



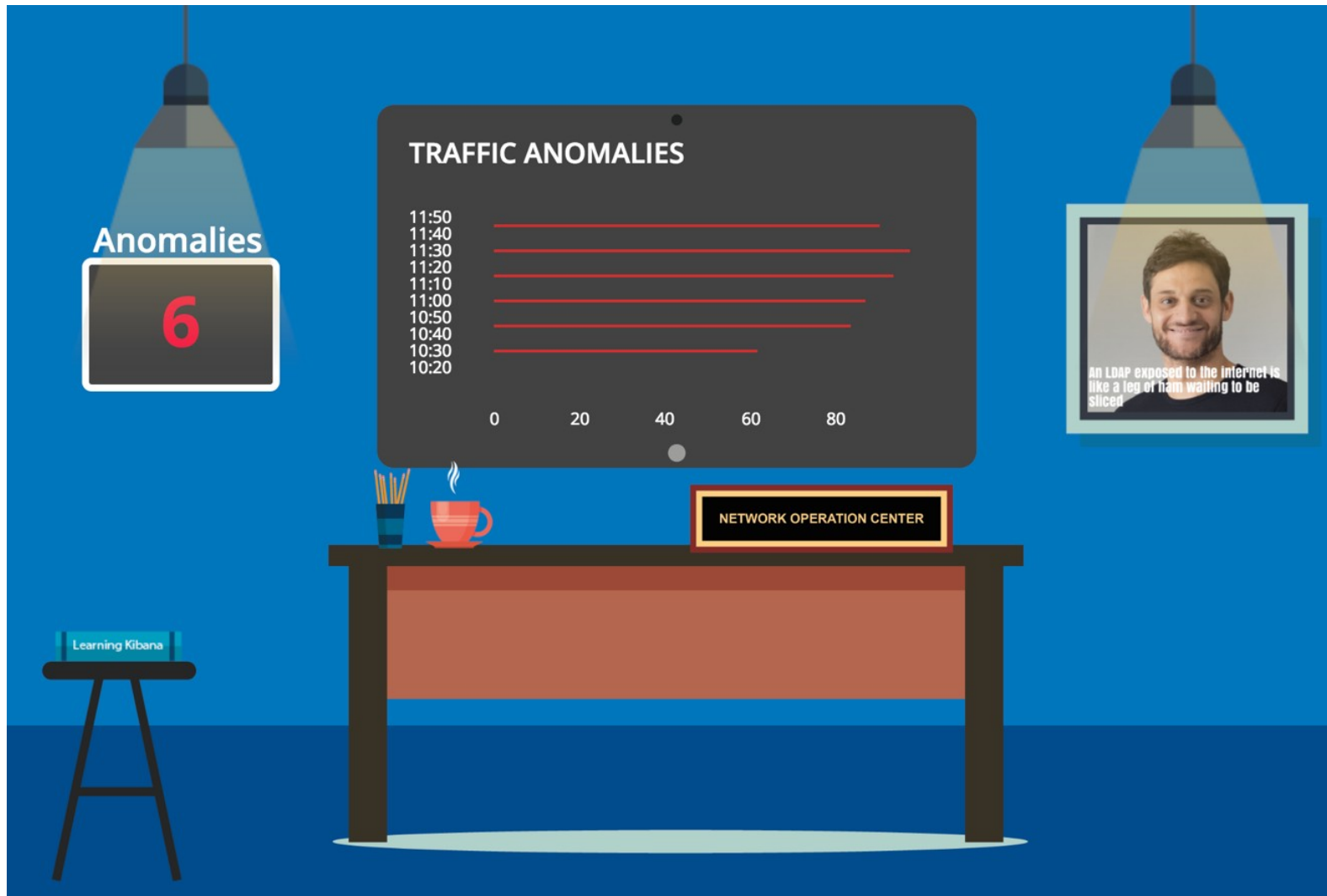
# Example

- Exemple personnalisation d'un graphique (bar)

```
filters
| essql
query="SELECT timestamp, anomaly_score FROM \".ml-anomalies-shared\"
WHERE result_type = 'bucket' AND anomaly_score > 10 AND job_id =
'nginx-
traffic'"
| pointseries x="anomaly_score" y="timestamp"
| plot
defaultStyle={seriesStyle lines=0 bars="2" points=0 horizontalBars=true
color="#d32f2f"} legend=false xaxis=true yaxis=true
font={font family="'Open Sans', Helvetica, Arial, sans-serif" size=12
align="left" color="#FFFFFF" weight="normal" underline=false
italic=false}
| render containerStyle={containerStyle backgroundColor="#444444"}
```



# Exemple Canvas



# Référence

- « *Machine Learning with Elastic Stack* » :  
Rich Collier et Bahaaldine Azarmi
- Web Site :

## Overview

<https://www.elastic.co/guide/en/elastic-stack-overview/current/xpack-ml.html>

## Data structure

<https://www.elastic.co/guide/en/elasticsearch/reference/7.0/api-definitions.html>

# Résultats des jobs

- Les résultats des jobs ML sont présentés à trois différents niveaux d'abstraction
  - **Bucket** : Représente comment ce bucket est inhabituel vis à vis des détecteurs du job
  - **L'enregistrement** : Ce sont les informations les plus détaillées sur chaque occurrence anormale ou entité anormale dans un intervalle de temps
  - **Influenceur** :
- Pour accéder aux résultats :
  - Utiliser l'API ML */results*
  - Rechercher dans les index créés par ML. Méthode la plus flexible

# L'index résultat

- ML analyse les données et stocke les résultats dans un index nommé ***.ml-anomalies-shared*** par défaut ou ***.ml-anomalies-custom-myname*** si on l'a indiqué dans la configuration du job le champ *results\_index\_name*
- De plus, un alias d'index est également créé sous la forme ***.ml-anomalies-jobname*** :

```
GET ml-anomalies-*/_alias
".ml-anomalies-farequote": {
  "filter": {
    "term": {
      "job_id": {
        "value": "farequote",
        "boost": 1
      }
    }
  }
}
```

# Types de document

- Dans l'index des résultats, il existe une variété de documents différents, chacun ayant sa propre utilité en ce qui concerne Alerting :
  - ***result\_type:bucket***: pour donner des résultats au niveau du bucket.  
1 document par bucket, timestamp égal au démarrage du bucket
  - ***result\_type:record***: Pour donner des résultats au niveau de l'enregistrement.  
1 document pour chaque anomalie trouvé dans le bucket span.  
timestamp égal au démarrage du bucket
  - ***result\_type:influencer***: Pour donner des résultats au niveau des influenceurs.  
1 document pour chaque influenceur d'une anomalie, timestamp égal au démarrage du bucket

# Documents bucket

- Les champs d'un document bucket :
  - ***timestamp***
  - ***anomaly\_score*** : Le score normalisé. La valeur peut fluctuer à mesure que de nouvelles données arrivent
  - ***initial\_anomaly\_score*** : Le score normalisé lors de la première analyse
  - ***event\_count*** : Le nombre de documents analysés pendant le bucket span
  - ***is\_interim*** : Un flag indiquant si ML attend encore des données pour ce bucket
  - ***bucket\_influencers*** : Un tableau des influenceurs identifiés pour ce bucket. (Il y a toujours l'influenceur par défaut `influencer_field_name:bucket_time`)

# Détails des influenceurs

```
"bucket_influencers": [  
  {  
    "job_id": "farequote",  
    "result_type": "bucket_influencer",  
    "influencer_field_name": "airline",  
    "initial_anomaly_score": 85.06429298617539,  
    "anomaly_score": 99.7634,  
    "raw_anomaly_score": 15.040566947916583,  
    "probability": 6.5926436244031685e-18,  
    "timestamp": 1486656000000,  
    "bucket_span": 900,  
    "is_interim": false  
  },  
  {  
    "job_id": "farequote",  
    "result_type": "bucket_influencer",  
    "influencer_field_name": "bucket_time",  
    "initial_anomaly_score": 85.06429298617539,  
    "anomaly_score": 99.76353,  
    "raw_anomaly_score": 15.040566947916583,  
    "probability": 6.5926436244031685e-18,  
    "timestamp": 1486656000000,  
    "bucket_span": 900,  
    "is_interim": false  
  }  
],
```

# Document enregistrement

- Les champs d'un document enregistrement :
  - ***timestamp***
  - ***record\_score*** : Le score normalisé pouvant fluctuer
  - ***initial\_record\_score*** : Le score normalisé de la première analyse
  - ***detector\_index*** : Indique le détecteur ayant provoqué l'anomalie
  - ***function*** : La fonction utilisé par le détecteur
  - ***is\_interim***
  - ***actual*** : La valeur réelle observée pour ce bucket.
  - ***typical*** : Une représentation de la valeur attendue par le modèle
- Si le job a été catégorisé (par *by\_field\_name*, *partition\_field\_name* ou des influenceurs), des données additionnelles sur les valeurs de la catégorie ayant provoqué l'anomalie sont disponibles



# Analyse de population

- Si le job effectue une analyse de population via le champ *over\_field\_name*, les résultats sont présentés différemment
  - Un premier bloc identifie l'individu anormal
  - Un tableau causes liste toutes les anomalies provoquée par cet individu

# Example

```
{ ...
  "_source": {
    "job_id": "gallery",
    "result_type": "record",
    "..."
  },
  "over_field_name": "clientip",
  "over_field_value": "173.203.78.60",
  "causes": [
    {
      "probability": 4.593248987780688e-31,
      "by_field_name": "status",
      "by_field_value": "404",
      "function": "count",
      "typical": [ 1.1177332137173952 ],
      "actual": [ 1215 ],
      "over_field_name": "clientip",
      "over_field_value": "173.203.78.60" } ],
    "influencers": [
      { "influencer_field_name": "uri",
        "influencer_field_values": [ "/wp-login.php" ] },
      { "influencer_field_name": "status",
        "influencer_field_values": [ "404" ] },
      { "influencer_field_name": "clientip",
        "influencer_field_values": [ "173.203.78.60" ] }
    ],
    "clientip": [ "173.203.78.60" ],
    "uri": [ "/wp-login.php" ],
    "status": [ "404" ]
  }
```

# Document influenceur

- Les documents influenceurs ont les champs suivants :
  - ***timestamp***
  - ***influencer\_score*** : Le score normalisé pouvant fluctuer
  - ***initial\_influencer\_score*** : Le score normalisé de la première analyse
  - ***influencer\_field\_name*** : Le nom de l'influenceur
  - ***influencer\_field\_value*** : La valeur de l'influenceur
  - ***is\_interim***

# Définition d'une alerte via l'UI

- La première méthode pour définir une alerte liée à un job ML consiste à utiliser l'assistant de l'UI, lorsque l'option *Continue job in real time*, est cochée

☒ Continue job in real-time

☒ Create watch for real-time job

**Time range**  
Now - 30m

**Severity threshold**  
critical ▼

☒ Send email

admin@elastic.co

Apply

# Paramètres de l'alerte

- Les paramètres demandés par l'UI sont :
  - **Time range** : Par défaut : now-2xbucket span. La valeur minimum étant : now -(bucket span + query delay)
  - **Seuil de sévérité** : Minimum score du bucket. Par exemple critical = 75
  - Adresse mail à alerter. L'alerte écrit toujours un message de log
- Après la création, l'alerte est visible, éditable et peut être testée via l'UI Watcher

# Implications

- La condition principale pour l'alerte est un score d'anomalie du bucket. L'alerte n'est pas déclenchée pour des anomalies qui ne provoqueraient pas le dépassement du seuil au niveau du bucket.
- Par défaut, seul un maximum des trois scores les plus élevés dans le bucket est indiqué dans la sortie, et uniquement si l'action d'envoi de mail est choisie.
- L'alerte existe toujours même si le job est supprimé.
- Les seules actions possibles de l'alerte sont la journalisation et l'envoi de mail.

# Alertes manuelles

- Des exemples beaucoup plu complexes peuvent être effectué manuellement.
- Voir :  
[https://github.com/PacktPublishing/Machine-Learning-with-the-Elastic-Stack/blob/master/Chapter06/custom\\_ML\\_watch.json](https://github.com/PacktPublishing/Machine-Learning-with-the-Elastic-Stack/blob/master/Chapter06/custom_ML_watch.json)
- Cet exemple
  - produit une alerte si 2 jobs différents détecte une anomalie pour le même bucket
  - Consolide les informations des 2 jobs dans le résultat