# Présentation des technologies

## Elastic Stack

### Présentation de la suite

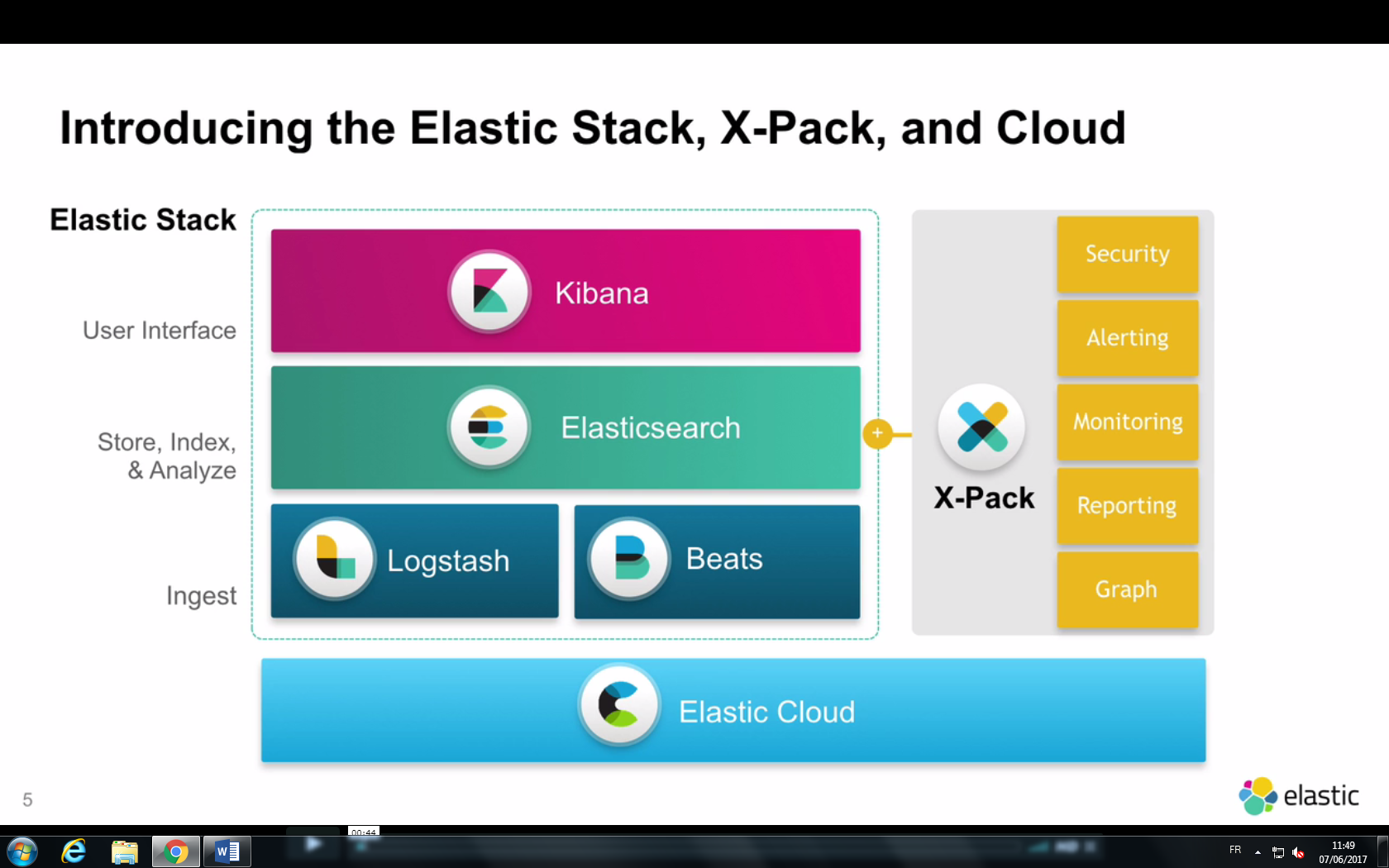
Site Web : <https://www.elastic.co> 07/06/2017

Elasticsearch : <https://www.elastic.co/fr/products/elasticsearch>

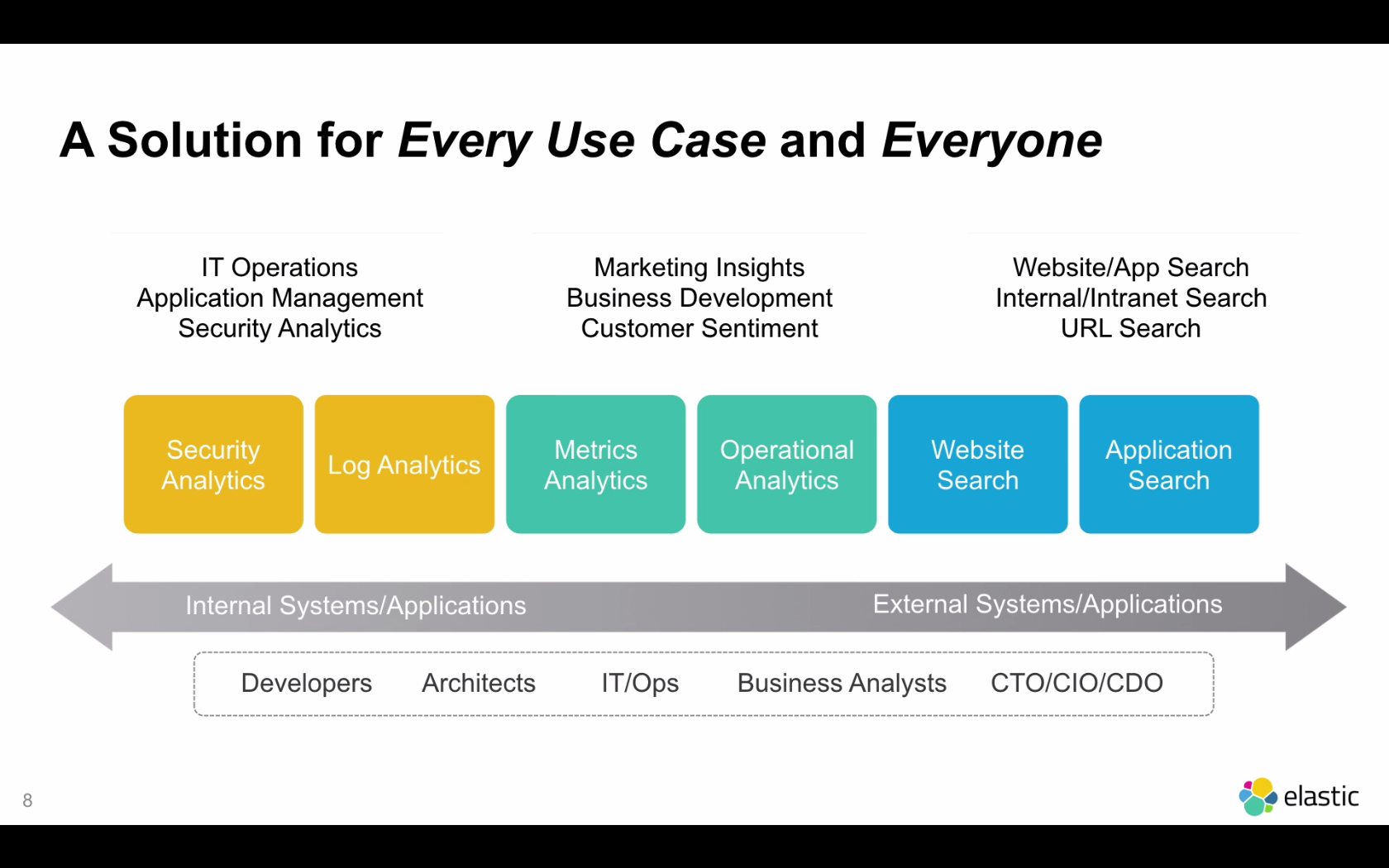
Logstash : <https://www.elastic.co/fr/products/logstash>

Kibana : <https://www.elastic.co/fr/products/kibana>

Elastic Stack est une architecture qui permet de récupérer des données (Logstash, Beats), de les stocker, indexer et analyser (Elasticsearch), et de les visualiser (Kibana). Elastic Cloud permet d’héberger le serveur sur le cloud et d’y accéder à distance (par opposition à la version locale) et X-Pack est une bibliothèque de plugins permettant par exemple de créer des graphes ou de sécuriser l’accès aux différents outils.



La suite Elastic (capture vidéo)



Champ d’utilisation de la suite Elastic

#### Elasticsearch

Au cœur de la suite Elastic, Elasticsearch est un outil de base de données sans schéma et au modèle de données flexible. Il est programmé en java. C’est un serveur d’API sur lequel on fait les requêtes via le protocole HTTP.

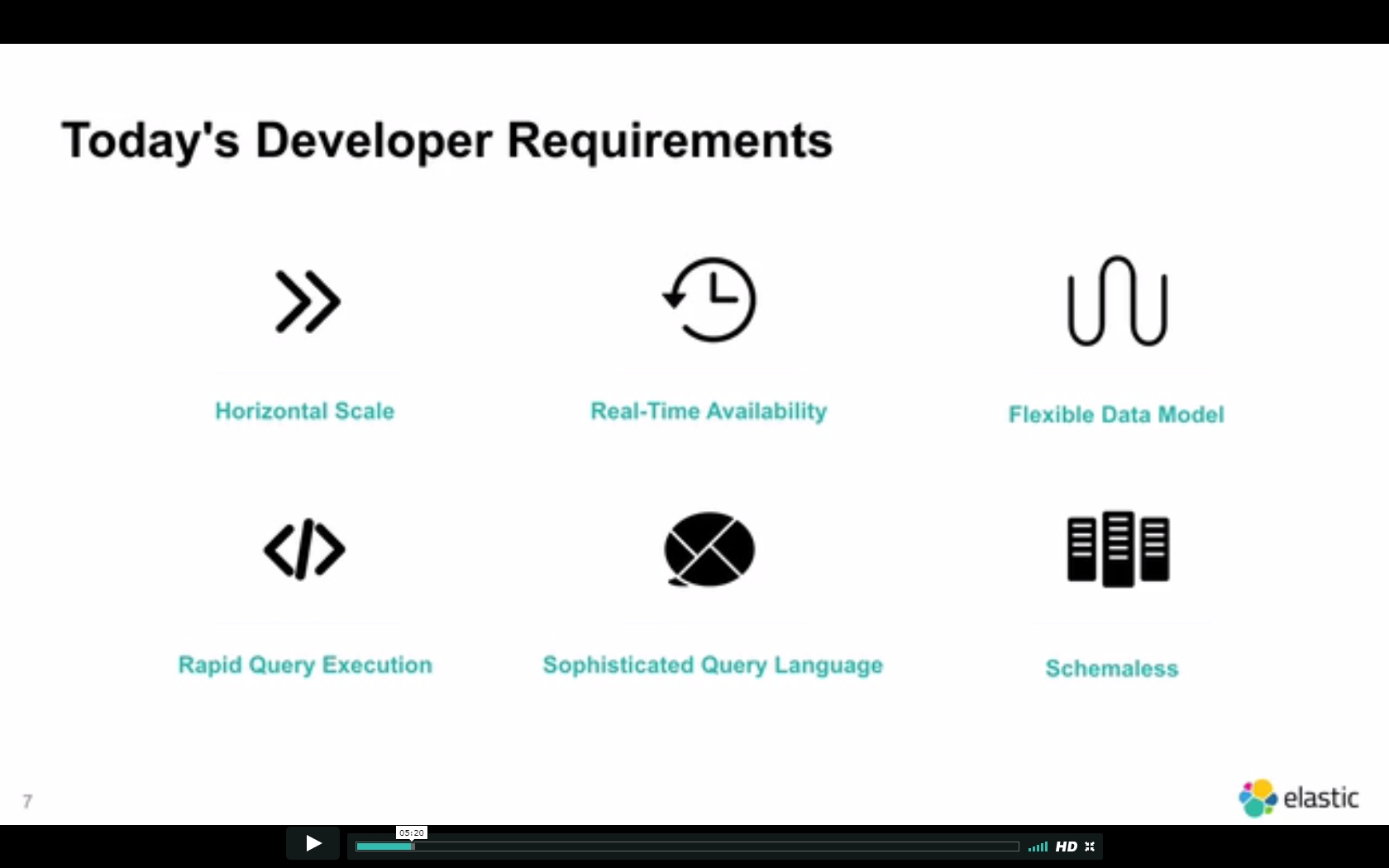
Clients :

* PSA Peugeot Citroën : efficience et amélioration des processus
* Orange : création d’un moteur de recherche intégré
* AXA : enrichissement et visualisation des données utilisateur
* Github : accélérer le développement de logiciels

Les avantages mis en avant sont les suivants :

* Scalabilité horizontale : si on veut traiter deux fois plus de données aussi rapidement, il faudra deux fois plus de machines
* Disponibilité en temps réel : les données envoyées au serveur sont disponibles pour les requêtes après une seconde environ
* Modèle de données flexible : ajout de champs sans réindexation
* Exécution rapide des requêtes
* Langage de requête sophistiqué : efficace avec les données structurées (mots-clés) ou numériques par exemple
* Pas de schéma : cela permet de se décharger de la définition d’un schéma qui est fastidieuse. Les champs sont détectés automatiquement ainsi que leurs caractéristiques, et on peut ensuite apporter les corrections nécessaires.

Une fonctionnalité clé est la géolocalisation que l’on peut affecter à des documents.



Points forts de Elastic search (Capture vidéo)

#### Logstash

Logstash est l’outil qui collecte les données en continu depuis différentes sources (logs Windows, twitter, Amazon), les filtre et les envoie vers une base de données (ici Elasticsearch).

#### Kibana

Kibana est un outil de visualisation des données avec des histogrammes, des graphes, des plans sur des tableaux de bord. Il permet aussi de faire les requêtes http et visualiser le résultat au format JSON.

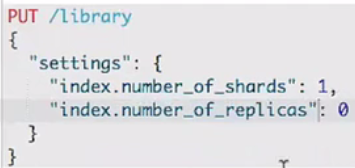


Tableau de bord avec Kibana

### Fonctionnement de l’API

#### Création d’un index

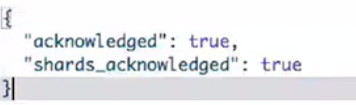
On peut comparer l’index à une table en BD relationnelle, ou à une collection sur les BD NoSQL. A un index peut être associé un mapping (schéma) et de settings (paramètres). Pour créer un index il suffit d’utiliser le verbe http PUT comme suit :



Avec :

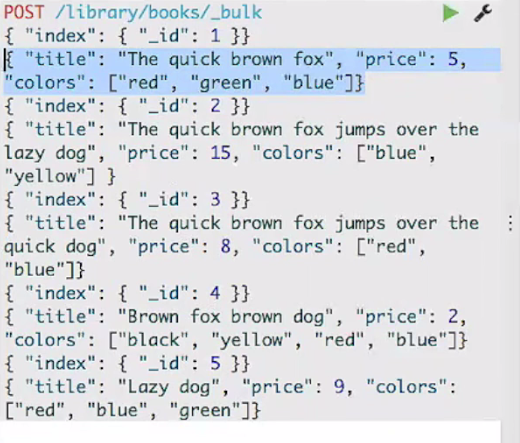
* Number of shards : Les shards (ou primary shards) sont des conteneurs qui permettent de faciliter la recherche des documents. Ces derniers sont répartis sur les shards en fonction de leur ID ou bien sur le même shard que le document parent s’il existe.
* Number of replicas : C’est le nombre de shards répliques (ou replica shards). Un document créé est ensuite créé dans les répliques.

Résultat :



#### \_bulk : Ajout de documents

L’Api \_bulk permet de créer plusieurs documents à la fois. On peut donner un type aux documents (ici « books »). Les documents, qui sont au format JSON, peuvent être précédés d’une ligne qui indique leur ID. On peut laisser le serveur attribuer l’ID en laissant le champ « index » vide.



#### \_search : recherche de documents

L’API \_search permet de faire une recherche sur les documents. On peut chercher par type ou bien sur la totalité d’un index.



Le résultat est le suivant :



En plus des infos de base des documents on trouve pour chaque document un champ « \_score ». Ce score est en fait un indice de pertinence des résultats. Ici ils sont tous égaux à 1 car on n’a pas de critère de recherche. ***Nous reviendrons sur la méthode de calcul du score.***

Comprendre le calcul de score

<https://www.elastic.co/guide/en/elasticsearch/guide/current/relevance-intro.html>

## Apache Spark et Spark streaming

### Spark <https://databricks.com/spark/about>

Spark est un outil de Big data performant. Il est beaucoup plus rapide que Hadoop car il utilise pleinement la mémoire alors que Hadoop sollicite plus le disque (mémoire physique). En plus de la rapidité il est facile d’utilisation et propose un grand nombre d’opérateurs (plus de 100) pour transformer les données.

### Spark streaming <https://spark.apache.org/streaming/>

<http://blog.ippon.fr/2014/12/10/introduction-a-spark-streaming/>

Spark streaming fait partie de l’écosystème de Apache Spark. Il permet de traiter des données qui sont fournies régulièrement. Il propose entre autres de récupérer des messages depuis Kafka ou des tweets.

## Docker

<https://www.docker.com/community-edition>

<http://putaindecode.io/fr/articles/docker/>

<http://putaindecode.io/fr/articles/docker/dockerfile/>