



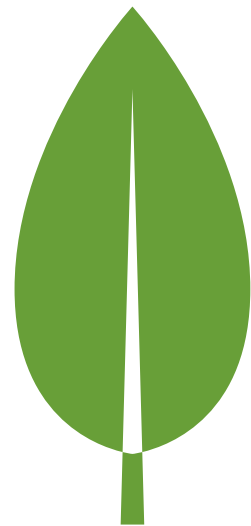
Data Science – 2024

REST for Big-Data

ElasticSearch for Big-Data

Master Data Science

Last update: Déc. 2024



Introduction à Elasticsearch

- **Définition :** Elasticsearch est un moteur de recherche et d'analyse distribué basé sur Apache Lucene.
- **Utilisation principale :**
 - Recherche rapide dans de grands volumes de données.
 - Analyse en temps réel.
- **Caractéristiques clés :**
 - Basé sur un modèle NoSQL.
 - Supporte la recherche full-text et les agrégations.
 - Extensible et scalable.
- **Cas d'utilisation :**
 - Journaux système (log management).
 - Monitoring (observabilité avec Elastic Stack).
 - Recherche textuelle sur des sites web (e.g., moteur de recherche e-commerce).

Pourquoi Elasticsearch ?

- **Avantages par rapport à d'autres solutions :**
 - Performances : Recherche en millisecondes grâce aux index.
 - Scalabilité : Fonctionnement distribué par nature.
 - Flexibilité : Supporte des documents JSON semi-structurés.
- **Comparaison avec des bases de données traditionnelles :**
 - Optimisé pour la recherche rapide et les agrégations.
 - Indexation des données pour accélérer les requêtes.
- **Importance dans le Big Data :**
 - Gestion et analyse de données massives.
 - Intégration avec Hadoop, Logstash, Kibana, etc.

Concepts fondamentaux – Cluster et Nœuds

- **Cluster : Un ensemble de nœuds Elasticsearch travaillant ensemble.**
 - Nom unique : Chaque cluster a un identifiant unique.
 - Scalabilité horizontale : Ajout ou suppression de nœuds facilement.
- **Nœuds : Instances Elasticsearch dans un cluster.**
 - Types :
 - Master Node : Coordonne le cluster.
 - Data Node : Stocke et traite les données.
 - Ingest Node : Traite les documents avant l'indexation.
 - Communication via API RESTful.

Concepts fondamentaux – Index

- Index: Équivalent à une "base de données" dans Elasticsearch.
- **Structure :**
 - Shards : Sous-divisions d'un index, permettant la distribution.
 - Replicas : Copies redondantes pour la tolérance aux pannes.
- **Création d'un index :**
 - PUT /mon_index
 - {
 - "settings": {
 - "number_of_shards": 3,
 - "number_of_replicas": 1
 - }
 - }
- **Requêtes associées :**
 - Création, mise à jour, suppression.
 - Recherche et agrégations.

Document et mappage

- Document : Unité de données dans Elasticsearch (JSON).
 - Exemple :
 - `{"titre": "Elasticsearch", "auteur": "Elastic", "date": "2024-01-01", "tags": ["recherche", "analyse"]}`
- Mapping : Décrit la structure et le type des données d'un index.
 - Exemple :
 - `PUT /mon_index`
 - `{`
 - `"mappings": {`
 - `"properties": {`
 - `"titre": { "type": "text" },`
 - `"date": { "type": "date" }`
 - `}`
 - `}`
 - `}`
- Elasticsearch gère dynamiquement les types si aucun mappage n'est défini.

Recherche dans Elasticsearch

- Requêtes basiques :

- Recherche par correspondance :
- GET /mon_index/_search
- {
- "query": {
- "match": {
- "titre": "Elasticsearch"
- }
- }
- }

Recherche dans Elasticsearch

- Requêtes booléennes :
 - Combinaison de conditions (must, should, must_not).
- Requêtes full-text :
 - Prise en charge de la recherche floue, correspondance partielle.
 - Exemple :
 - GET /mon_index/_search
 - {
 - "query": {
 - "match": {
 - "tags": "analyse"
 - }
 - }
 - }

Agrégations

- Définition : Permet de résumer, analyser et agréger des données.
- **Types d'agrégations :**
 - Statistiques : Moyenne, minimum, maximum.
 - Bucket : Groupement de données (e.g., par date, catégorie).
- **Exemple : Compter les documents par "tags".**
- GET /mon_index/_search
 - {
 - "aggs": {
 - "par_tags": {
 - "terms": {
 - "field": "tags.keyword"
 - }
 - }
 - }
 - }

Indexation

- Indexation :
 - Ajout de documents dans un index.
- POST /mon_index/_doc
- {
- "titre": "Exemple d'indexation",
- "contenu": "Introduction à Elasticsearch"
- }

Pipeline d'ingestion

- Transformation des données avant indexation.

- Exemple : Ajout d'un champ calculé.

```
PUT _ingest/pipeline/mon_pipeline
{
  "processors": [
    {
      "set": {
        "field": "nouveau_champ",
        "value": "valeur par défaut"
      }
    }
  ]
}
```

Monitoring et Sécurité

- **Monitoring :**

- Outils natifs : Kibana (Tableaux de bord).
- Analyse des performances via `/_cat` API.
 - `GET /_cat/indices?v`

- **Sécurité :**

- Authentification avec Elastic Security.
- Communication chiffrée via SSL/TLS.
- Gestion des rôles et utilisateurs.
 - `POST /_security/user/admin`
{
 "password": "password",
 "roles": ["superuser"]
}

Intégrations et Cas d'utilisation

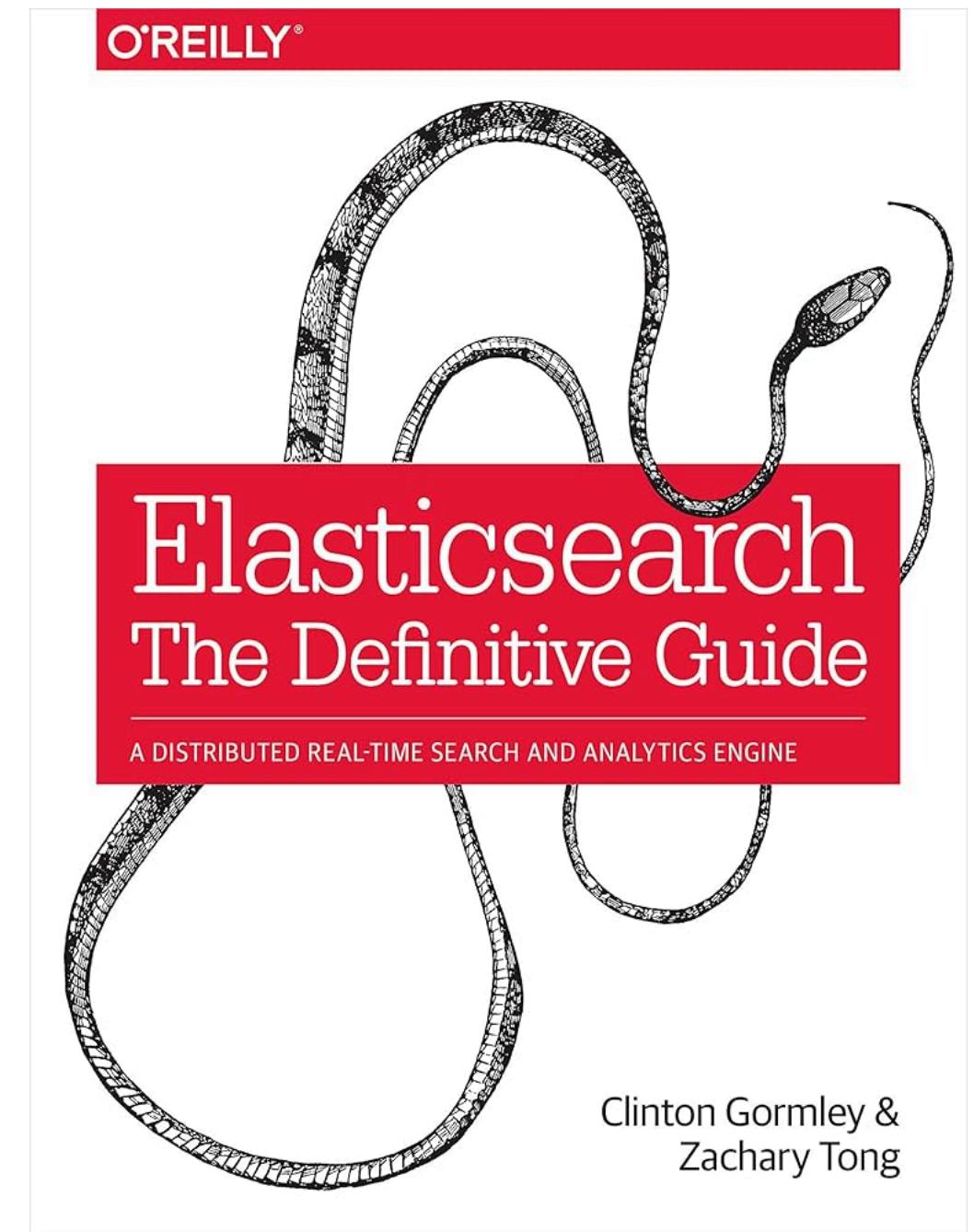
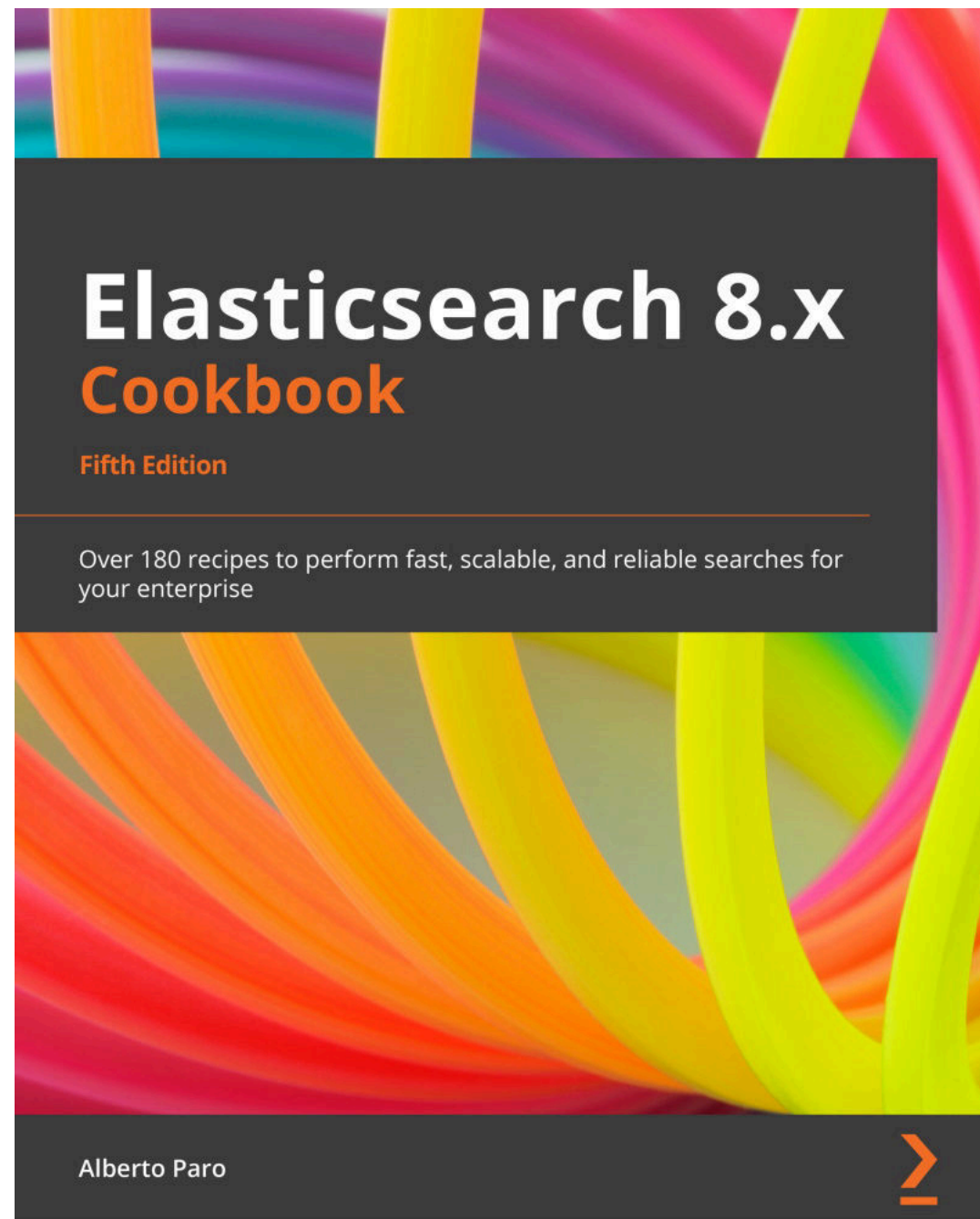
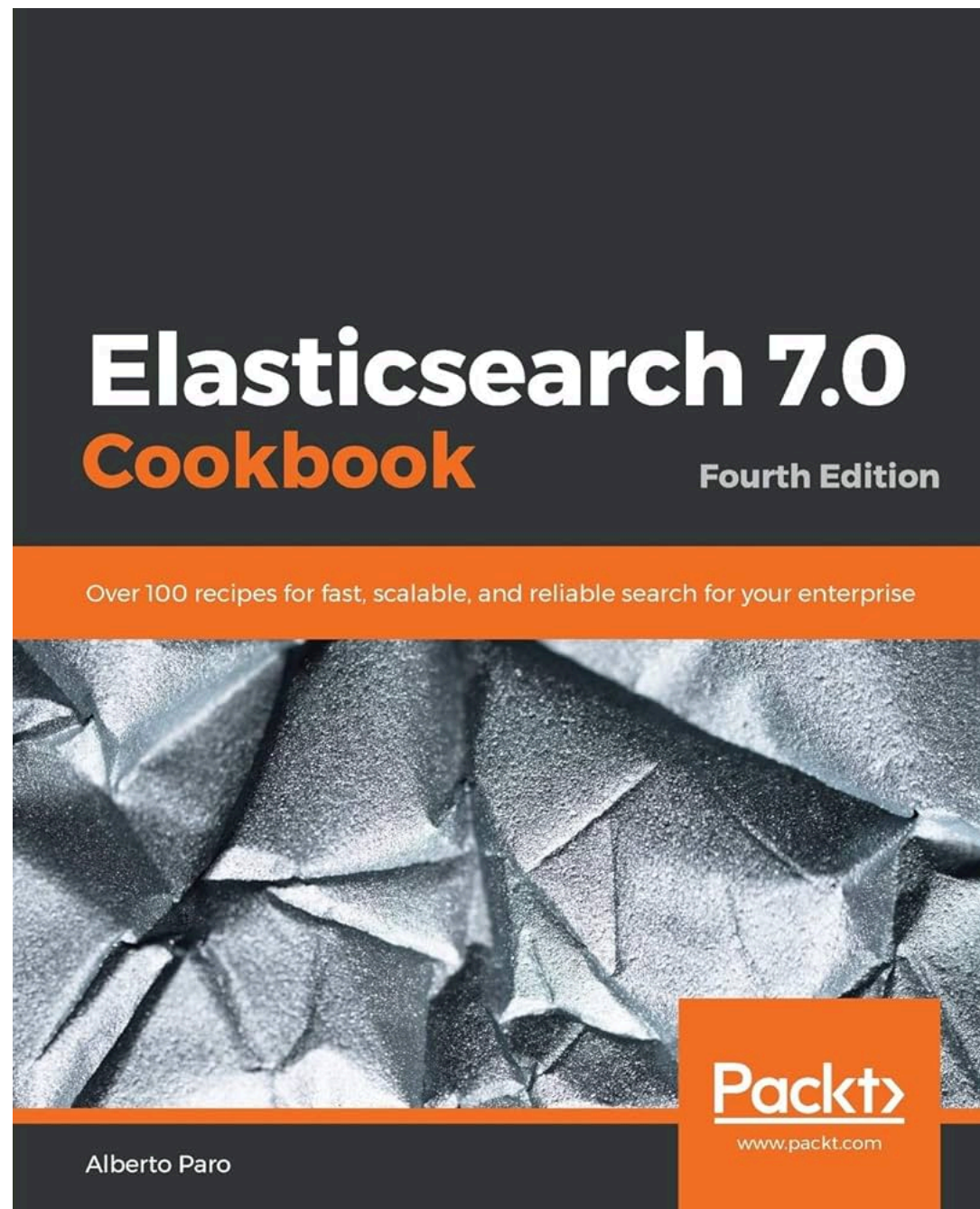
- **Intégrations :**

- Logstash : Import et transformation de données.
- Kibana : Visualisation et tableaux de bord.
- Beats : Collecte de journaux.

- **Cas pratiques :**

- E-commerce : Recherche de produits.
- Analytics : Suivi des utilisateurs en temps réel.
- Observabilité : Gestion des logs (stack ELK).

Documentation et lecture



+ la documentation sur site Officiel : www.elastic.co