



Maykin Media

Technisch ontwerp Newsroom



Inhoudsopgave

Inleiding.....	3
Overzicht Newsroom applicatie.....	3
Bronnen in Newsroom.....	4
Infrastructuur overzicht.....	5
Architectuur Newsroom applicatie.....	8



Inleiding

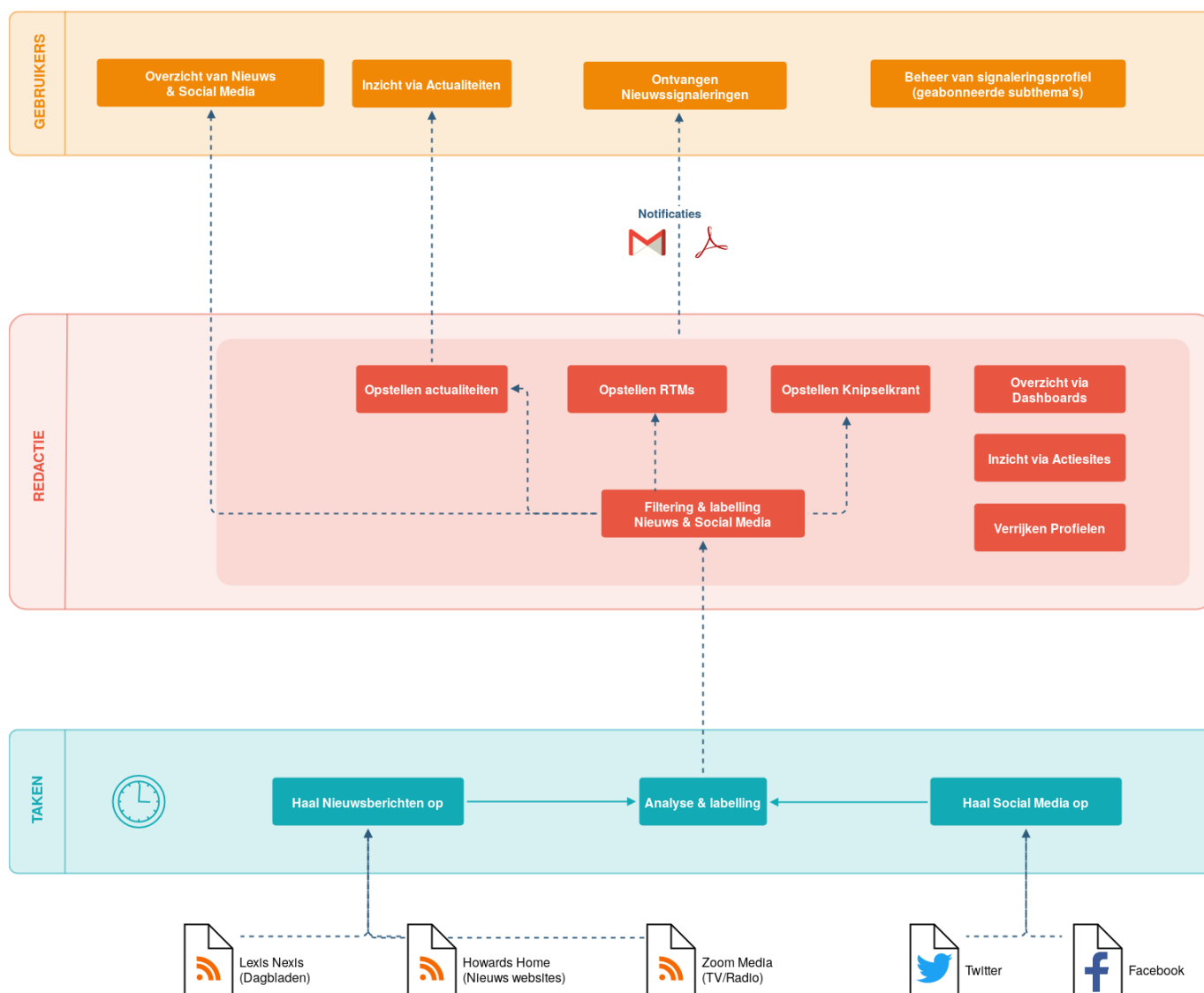
In 2019 heeft Maykin Media in opdracht van de Ministerie Infrastructuur en Waterstaat Directie Communicatie (MinlenW DCO) gerealiseerd. In dit document beschrijven wij de opgeleverde oplossing in technische zin. De applicatie draait in productie op newsroom.minienw.nl.

Overzicht Newsroom applicatie

De Newsroom applicatie is ontwikkeld om DCO te ondersteunen in:

1. het filteren en analyseren van nieuwsberichten en sociale media,
2. het labelen van deze berichten onder (sub)thema's of actualiteiten,
3. het van redactionele context voorzien van de actualiteiten
4. het op de hoogte brengen van medewerkers van lenW over deze actualiteiten

Een schematische weergave van de functionaliteiten in Newsroom (exclusief beheer):





Bronnen in Newsroom

Momenteel zijn de volgende 5 bronnen beschikbaar via Newsroom:

- Lexis Nexis – Dagbladen / Krantenartikelen
- Howards Home – Artikelen van nieuwswebsites
- Zoom Media – Een selectie van TV/Radio
- Twitter – posts, profielgegevens & statistieken
- Facebook – posts & statistieken

De eerste 3 bronnen worden via RSS ingeladen via bron-specifieke 'collectors'. Voor Twitter en Facebook wordt gebruik gemaakt van de APIs van beide platformen. De redactie- en beheeromgevingen van Newsroom biedt een uitgebreide reeks aan mogelijkheden om in te stellen hoe deze bronnen worden bevestigd.

RSS en Twitter integraties

In Newsroom zijn voor de 3 RSS bronnen specifieke RSS collectors opgebouwd, dit aangezien Lexis Nexis, Howards Home en Zoom Media net op een andere manier additionele metadata (identifiers om de artikelen te ontdebelen) in de RSS hebben opgenomen waar Newsroom gebruik van maakt. Behalve dat zijn dit reguliere RSS parsers en de feeds zijn conform RSS versie 2.0.

Voor de informatie vanuit Twitter wordt momenteel (nog) gebruik gemaakt van de Twitter API v1.1. Een reeks Newsroom taken zijn beschikbaar om de verschillende Twitter API endpoints aan te roepen. Wat er wordt ingesteld en welke gegevens worden opgevraagd vanuit deze endpoints wordt gestuurd vanuit de redactie/beheeromgeving van Newsroom (self-service).

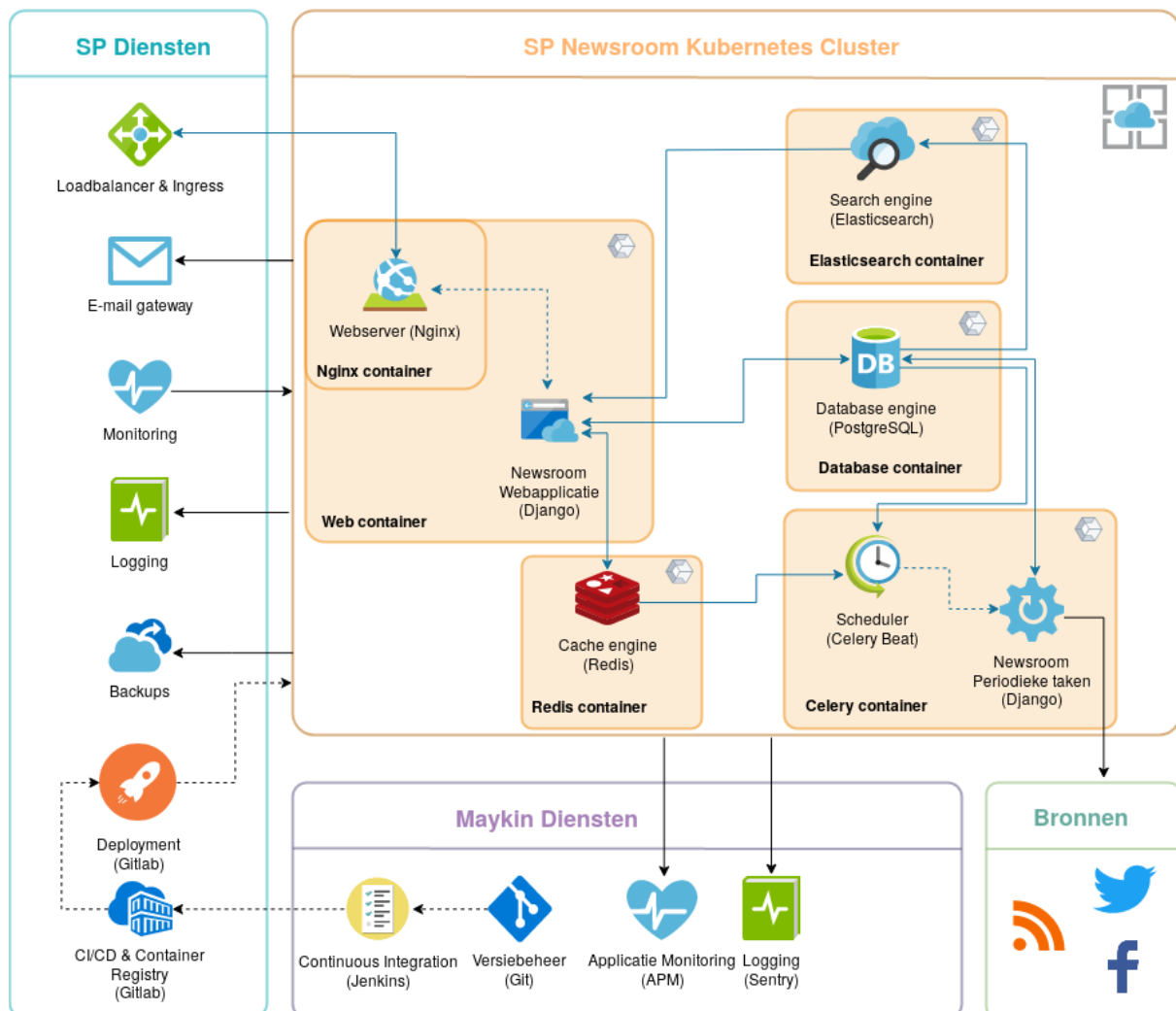
- Twitter profielen worden opgehaald met de users/show endpoint
- Twitter profielwijzigingen worden gedetecteerd met de statuses/show endpoint
- Twitter timelines worden opgehaald met de statuses/user_timeline endpoint
- Twitter favorieten worden opgehaald met de favorites/list endpoint
- Followers van profielen worden opgehaald via de followers/ids endpoint
- Namen van followers worden opgehaald via de users/lookup endpoint
- Tweets worden opgehaald door middel van specifieke zoekopdrachten via de search/tweets endpoint
- Rate limits worden gemonitored door middel van de application/rate_limit_status endpoint

Deze taken worden asynchroon door de applicatie uitgevoerd op door de beheerder ingestelde momenten met door de beheerder ingestelde collectors, waarin bijvoorbeeld zoektermen of profielnamen opgegeven kunnen worden. Tevens zijn er dataminimalisatietaken om bijvoorbeeld tweets ook automatisch te verwijderen.



Infrastructuur overzicht

Newsroom maakt gebruik van een Kubernetes cluster draaiend op het Standaard Platform (SP) van MinlenW. De installatie is als volgt schematisch weergegeven:



De Newsroom installatie op de Kubernetes cluster van het SP bestaat uit de volgende 6 containers per test/acceptatie/(pre)productie namespace. Wanneer we het hebben over een Newsroom 'cluster' bedoelen we de 6 containers binnen een namespace.

- **Web:** Hier in draait de Newsroom webapplicatie, welke in Python/Django gebouwd is conform de voorgestelde architectuur. Handelt de web-verzoeken (request/response) van browsers af.
- **Nginx:** De Nginx installatie binnen deze container is geconfigureerd als reverse-proxy, deze webserver is voornamelijk ingericht om media en statische bestanden rechtstreeks te serveren. Vanwege restricties op het SP was het noodzakelijk om deze container te 'side-loaden' in de web pod.
- **Database:** Een standaard PostgreSQL container is ingericht voor het gestructureerd opslaan van alle Newsroom data.



- Elasticsearch: Een standaard Elasticsearch container is ingericht voor het snel volledig doorzoekbaar maken van alle artikelen in Newsroom. De index in de zoekmachine wordt dagelijks ververs, en wanneer er wijzigingen zijn aan de artikelen in Newsroom.
- Redis: Een standard Redis container is ingericht voor het snel ophalen van tijdelijke data (caching) en voor het opslaan van de Celery task queue data.
- Celery: Een kopie van Newsroom (web container) ingericht om de asynchrone taken uit te voeren naar aanleiding van een gebruikeractie of op gezette tijden. Een voorbeeld is het herhaaldelijk ophalen van de gegevens uit de bronnen van Newsroom.

Vanwege de SP / IenW-specifieke eisen is er geen gebruik gemaakt van de reguliere installatie playbooks van Maykin. De verschillende componenten zijn voor een webapplicatie zoals Newsroom standaard, echter de configuratie, deployment en beheer van Newsroom vereisen een hoge mate van afwijking t.o.v. onze standaard 'operating procedures' dat ervoor gekozen is om dit volledig los te splitsen.

SP diensten

Naast de Newsroom cluster zijn er een reeks diensten van het Standaard Platform een afhankelijkheid is voor het goed kunnen functioneren van Newsroom en waar Maykin gebruik van maakt voor technisch applicatiebeheer. Uiteraard heeft het SP meer diensten in gebruik, slechts een deel zijn voor het technisch ontwerp van Newsroom onderkent:

- Loadbalancer / Ingress: Verkeer van en naar Newsroom verloopt via dit onderdeel van het SP. Handelt SSL-termination af en stuurt de verzoeken (afhankelijk van de domeinnaam) door naar een Newsroom cluster. Zie voor meer details de documentatie vanuit het SP.
- E-mail gateway: Verkeer verzonden vanuit Newsroom wordt via de centrale SP E-mailgateway ienwprd1-mailrelay001.external-cloud.nl verstuurd. Er is geen inkomende e-mailverkeer naar Newsroom, dit wordt met behulp van een centrale redactie-mailbox afgehandeld. Zie voor meer details de documentatie vanuit het SP.
- Monitoring: Via de OpenShift beheeromgeving van het SP is inzicht mogelijk in het correct draaien van de containers. Zie voor meer details de documentatie vanuit het SP.
- Logging: Via de Kibana omgeving van het SP is inzicht in de Kubernetes logs van de verschillende containers. Zie voor meer details de documentatie vanuit het SP.
- Backups: Wordt door het SP uitgevoerd, hiervoor is wel vanuit de Kubernetes deployment het met 'velero' taggen van de persistent volumes vereist. Zie voor meer details de documentatie vanuit het SP.

Maykin diensten

Vanwege de beschikbaar gestelde SP diensten waar Newsroom op aangesloten dient te worden zijn maar een deel van de reguliere diensten van Maykin in gebruik voor Newsroom.

- Logging: Via een centrale Sentry-omgeving wordt loginformatie verzameld en wordt (bij warnings/errors/exceptions in Newsroom) het technische beheerteam van Maykin er van op



de hoogte gebracht. Hierdoor valt bij problemen van een applicatie snel een analyse uit te voeren en kan Maykin inspelen op problemen zonder dat ze door een gebruiker of beheerder gerapporteerd hoeven te worden.

- APM: Applicatie-monitoring wordt ingezet voor het controleren van de beschikbaarheid en de performance van Newsroom. Net zoals bij het SP wordt hiervoor de ELK-stack gebruikt.

Deployment van nieuwe releases

Een nieuwe Newsroom release volgt de volgende stappen:

1. Bij Maykin wordt in git een pull-request van develop naar master uitgevoerd. Na het slagen van alle tests en een code-review worden de wijzigingen samengevoegd in de master-branch ter release.
2. Via Ansible wordt door Maykin op de eigen OTAP-infrastructuur de Newsroom test- en staging-omgevingen automatisch bijgewerkt (newsroom-test.maykin.nl, newsroom-staging.maykin.nl)
3. De Newsroom repository wordt door Maykin gepushed naar de Gitlab omgeving van het SP, inclusief de Kubernetes configuratiebestanden. De images worden via de Gitlab Build Pipeline gebouwd en in de Harbor registry geupload. De master-branch stuurt de acceptatie-uitrol aan, de prod-branch stuurt de productie-uitrol aan.
4. Op basis van de images in de Harbor registry wordt de Kubernetes acceptatie/productiecluster opgezet conform de in Newsroom ingerichte Kubernetes configuratie, wat leidt tot het bijwerken van de acceptatieomgeving (newsroom-acc-sp.maykin.nl) en de productieomgeving (newsroom.minienw.nl).



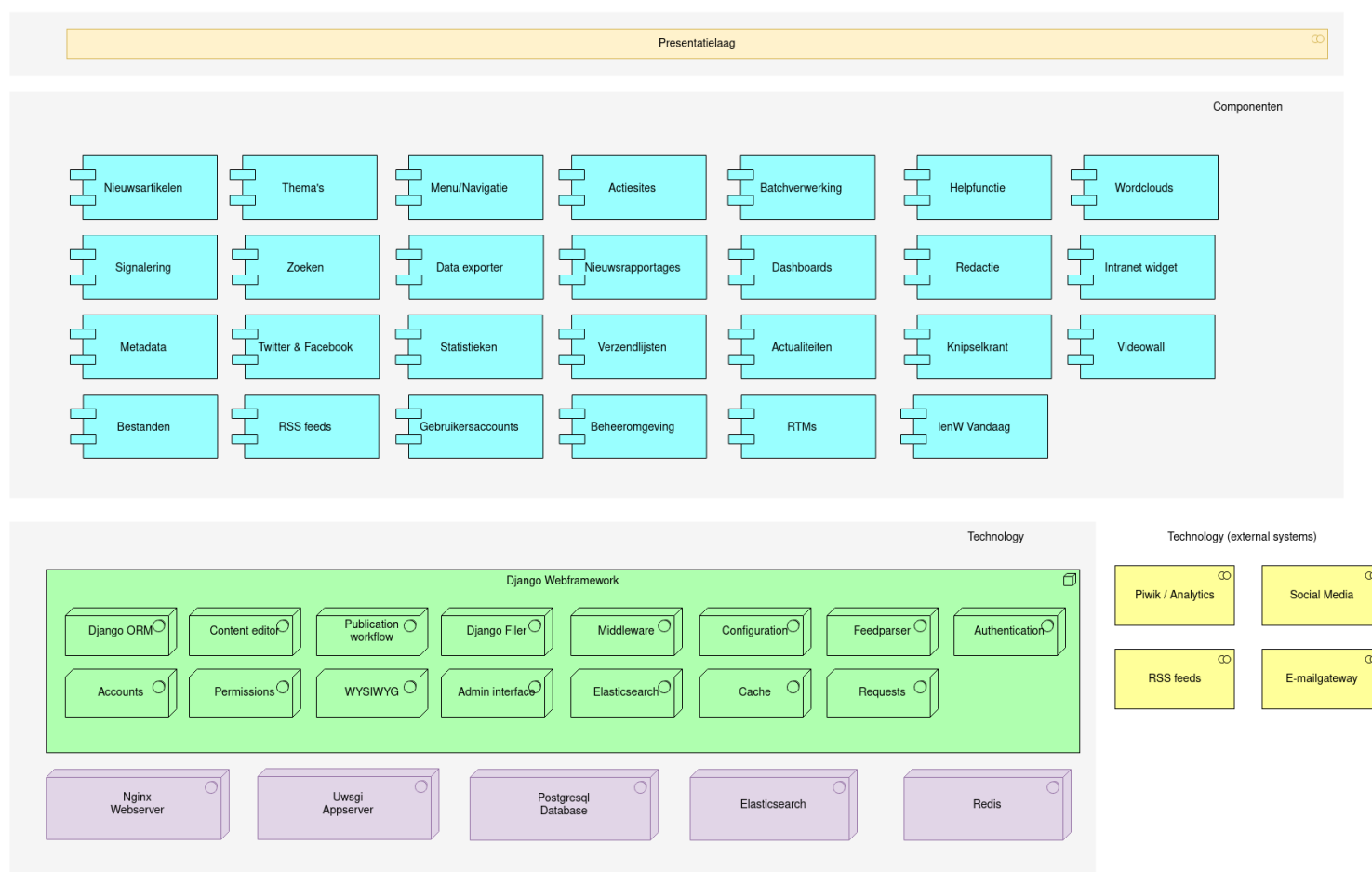
Architectuur Newsroom applicatie

Wij hebben voor Newsroom omgeving deels gebruik maken van standaard (open source) componenten, gecombineerd met modules die wij voor vergelijkbare projecten hebben gerealiseerd. Daar bovenop is maatwerk gerealiseerd om de door DCO gewenste workflow mogelijk te maken.

Voor de onderliggende componenten zijn gebruik gemaakt voor de (voor Maykin) standaard onderdelen: PostgreSQL als database, Nginx als webserver, Elasticsearch als search engine en Redis als caching server. Deze en de overige gebruikte componenten van haar oplossingen vallen onder een open source licentie, door onze “open source tenzij” werkwijze sluiten wij hier in goed aan op de werkwijze van het SP en wordt “vendor lock-in” voorkomen.

In onderstaand architectuurschets worden de verschillende modules en componenten weergegeven. Deze modulaire aanpak zorgt ervoor dat bij doorontwikkeling het toevoegen van nieuwe modules relatief eenvoudig is.

Architectuuroverzicht Newsroom



Het webframework Django wat wij gebruiken voor de realisatie van vrijwel al ons projecten is modulaair van aard. Door nieuwe modules toe te voegen kunnen bestaande website functionaliteiten worden uitgebreid of nieuwe onderdelen worden toegevoegd. Mochten modules niet meer nodig zijn dan zijn deze ook eenvoudig weer uit te schakelen.