News Hunter infrastructure and architecture

Slides by Marc Gallofré Ocaña

The News Hunter infrastructure

Service nodes
Web scraping, API, user interfaces, semantic lifting processes
- Light-to-medium processing
- Python, REST API, ...

Computation-intensive nodes
Complex AI services and training processes.
- CPU, RAM, GPU intensive
- *Python, spaCy, ...*

Harvesting news-related information from social media and other sources; analysing, organising, enriching and presenting news-related information to journalists. Implemented using state-of-the-art big data and distributed technologies.

Management nodes
Service orchestration and monitoring
- Lighter processing
- Docker Swarm

Message queue nodes
Message exchange, queueing (TBD)
- Lighter processing
- Kafka

Raw data nodes
Distributed storage for raw data files (textual, multimedia)
- Disk intensive
- *Cassandra, ...*

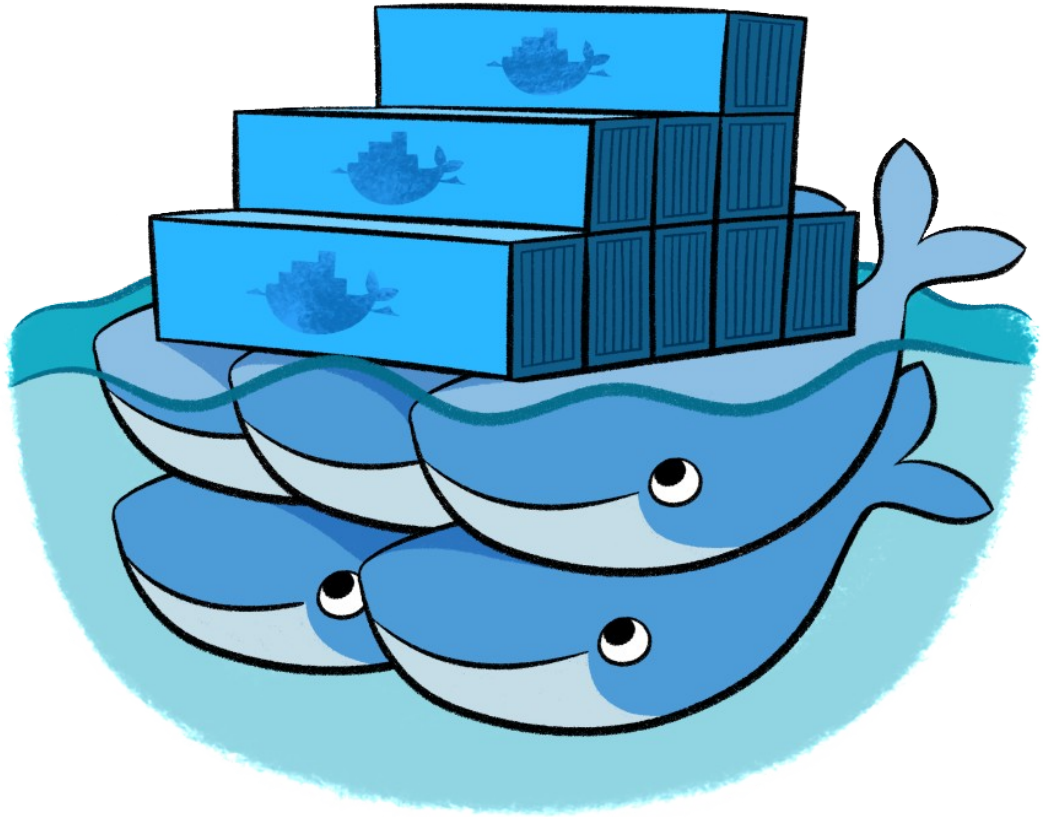
Configuration nodes
- Lighter processing
- *MongoDB, files*

Knowledge graph nodes
News semantic representation storage.
- Disk, CPU and RAM intensive
- *Blazegraph*

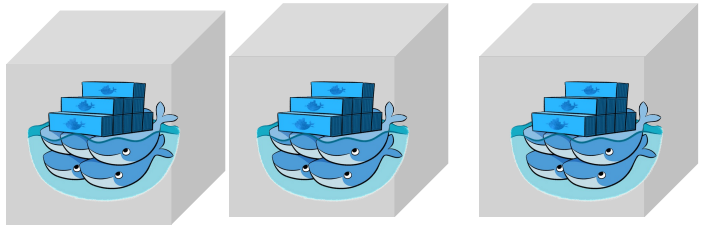
M. Gallofré Ocaña & A.L. Opdahl (2021)

Cloud infrastructure deployment tools

Service orchestration (Docker Swarm)

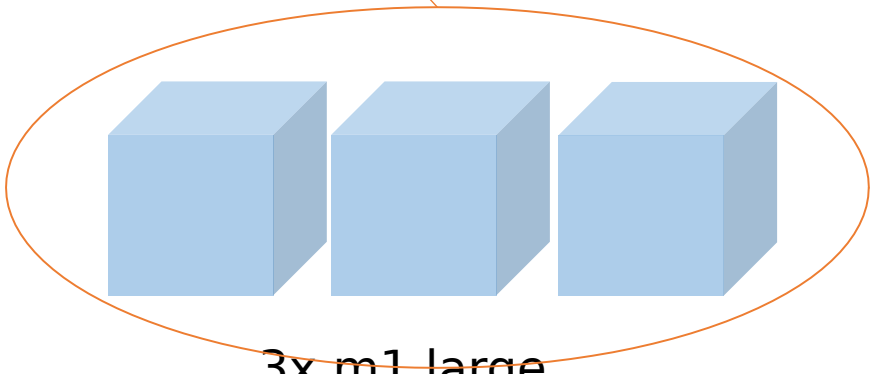


Service instances.
For running services
for AI, Web
scraping, API ...

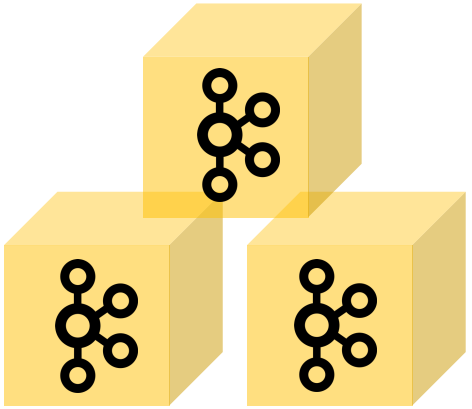


3x m1.medium
- 1 (3) vCPUs
- 4 (12) GB RAM

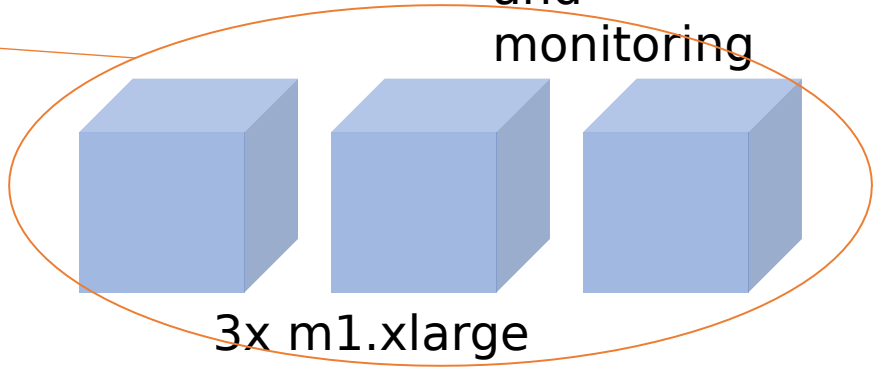
Manager nodes. For Swarm orchestration and monitoring



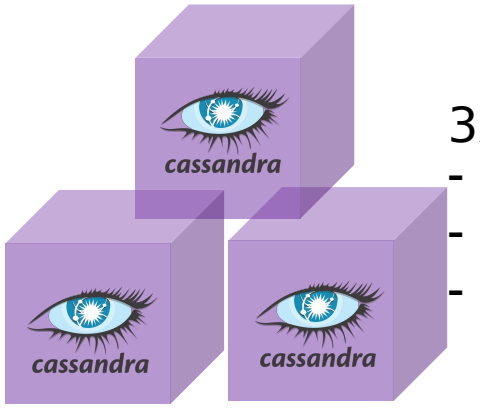
3x m1.large
- 2 (6) vCPUs
- 8 (24) GB RAM



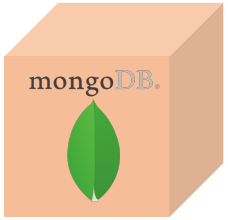
3x m1.large
- 2 (6) vCPUs
- 8 (24) GB RAM



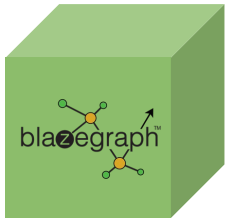
3x m1.xlarge
- 4 (12) vCPUs
- 16 (48) GB RAM



3x m1.large
- 2 (6) vCPUs
- 8 (24) GB RAM
- 3 (9) TB Disk



1x m1.medium
- 1 vCPUs
- 4 GB RAM
- 20 GB Disk



1x m1.xlarge
- 4 vCPUs
- 16 GB RAM
- 11 TB Disk

1 vCPU = 0.5CPU

News Hunter Platform:

- **38 vCPUs**
- **152GB RAM**
- **20TB Disk**
- **17 Instances**

+

1 Launcher instance for deploying the cloud infrastructure:

- **1 vCPU**
- **4 GB RAM**

1 vCPU = 0.5CPU

Technologies

- Docker Swarm
- Kafka (as pub/sub message queue to communicate between all services in the platform)
- Zookeeper
- Cassandra (storing raw data in a distributed cluster)
- Blazegraph (Knowledge graph with news and events representations)
- MongoDB (configuration and metadata)

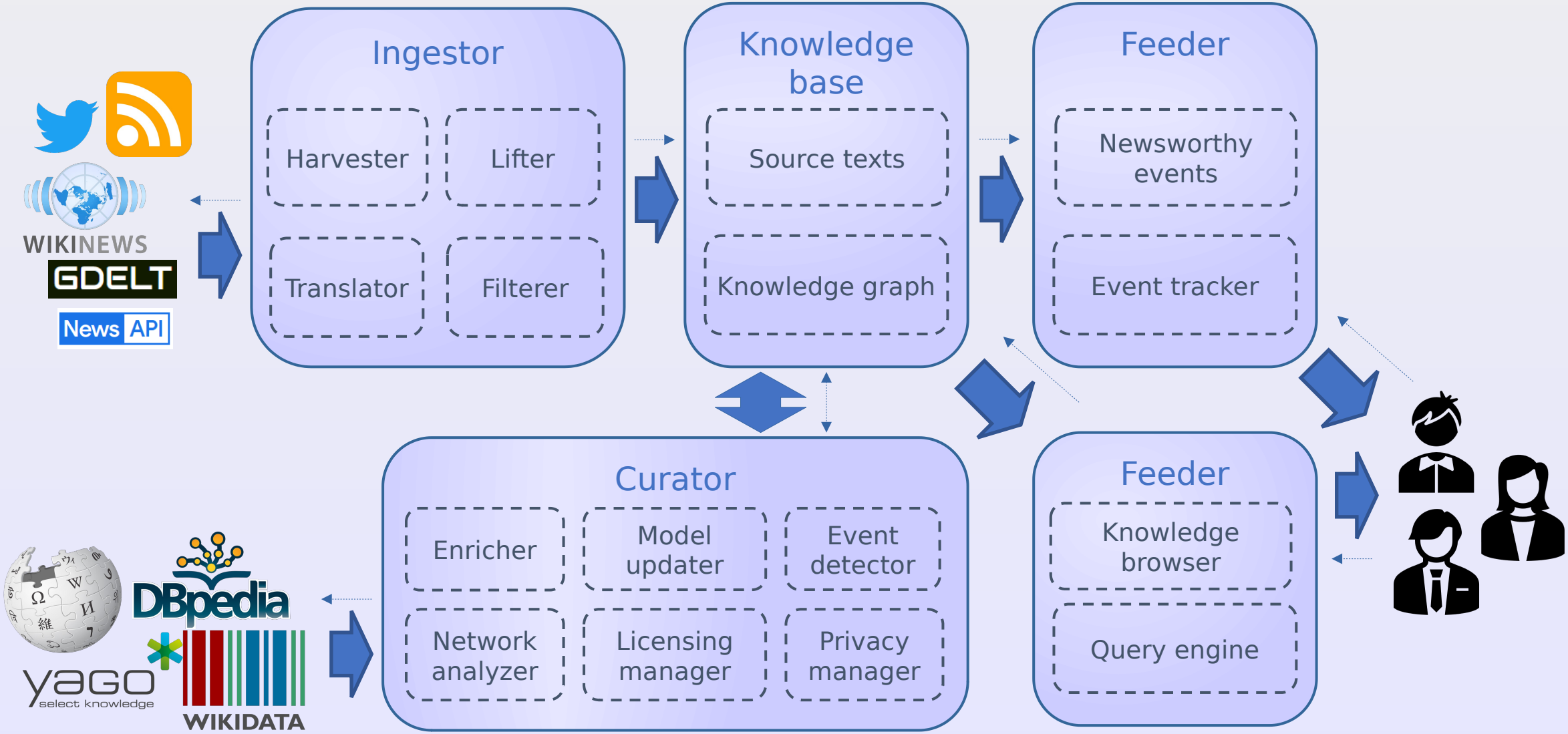
* All of them have been deployed using Docker containers

Services

- Written in Python 3.8-3.9
- All services are deployed in docker containers
- FastAPI as the main python library for writing APIs

The News Hunter architecture

Harvesting news-related information from social media and other sources; analysing, organising, enriching and presenting news-related information to journalists. Implemented state-of-the-art big data and distributed technologies.



M. Gallofré Ocaña & A.L. Opdahl (2021)

Services - harvesters

- Twitter harvester: connects to the Twitter API to read streams of tweets from news organizations accounts
- RSS harvester: downloads RSS feeds from news organisations
- GDELT harvester: gets the events and GKG datasets from GDELT projects
- NewsAPI harvester: use NewsAPI.org API to get real-time feeds of news from thousands of news outlets

Services - lifters

Lifters for news and GDELT that use NER to represent the information into knowledge graphs

- DbpediaSpotlight NEL: using DBpediaSpotlight for named entity linking
- SpaCy NEL: using SpaCy for named entity linking
- Kolitsas NEL: using Kolitsas algorithm for named entity linking