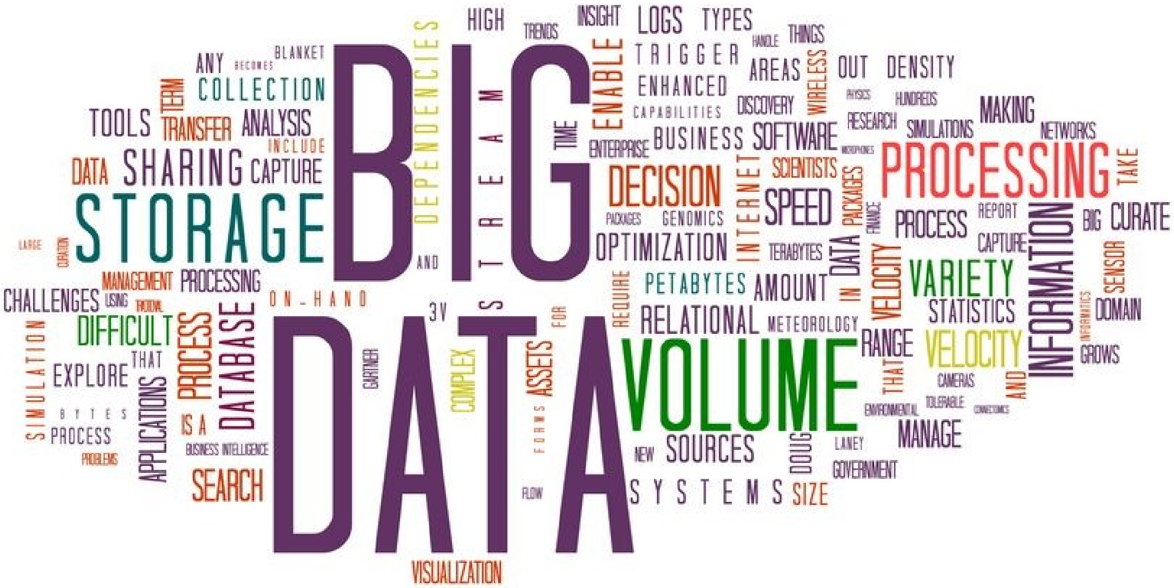
INFO319 - Big Data, Autumn 2022



Andreas L Opdahl <Andreas.Opdahl@uib.no>

The big data challenge

- How to deal with data that are too big for one machine?
  - the data explosion: availability of open, social data; IoT
  - standards for data exchange and identifiers
  - cheaper mass storage and communication
  - powerful *multi-core* processing
- Last two decades:
  - new, distributed technologies for large-scale data management
  - organisational and societal impacts

Course background

- Big Data for Emergency Management (BDEM)
  - an INTPART project (2018-2021)
  - original INFO319: big data for *emergency management*
    - critique: more emergency management than big data
  - revised INFO319: more focussed on big data
    - *news production* as focussed domain

GEORGE MASON UNIVERSITY

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

SAN DIEGO STATE UNIVERSITY

VESTLANDSFORSKING

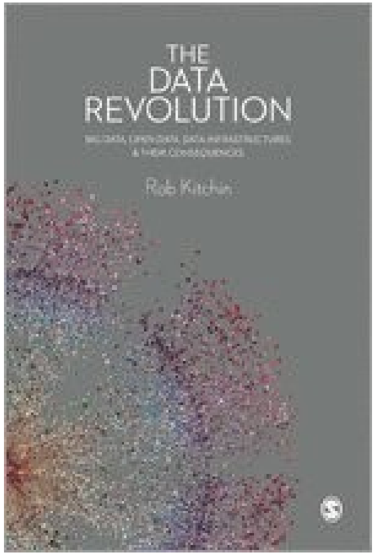
THE UNIVERSITY OF AIZU To Advance Knowledge for Humanity 1993

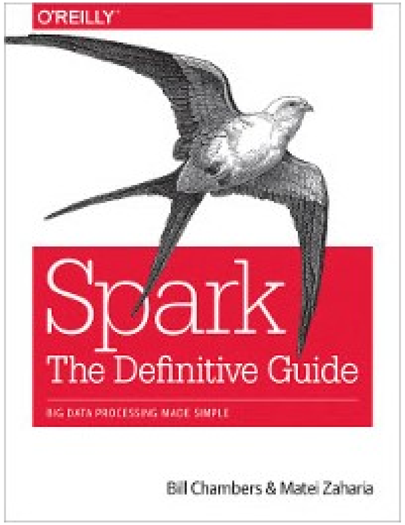
UNIVERSITAS BERGENSIS

Themes

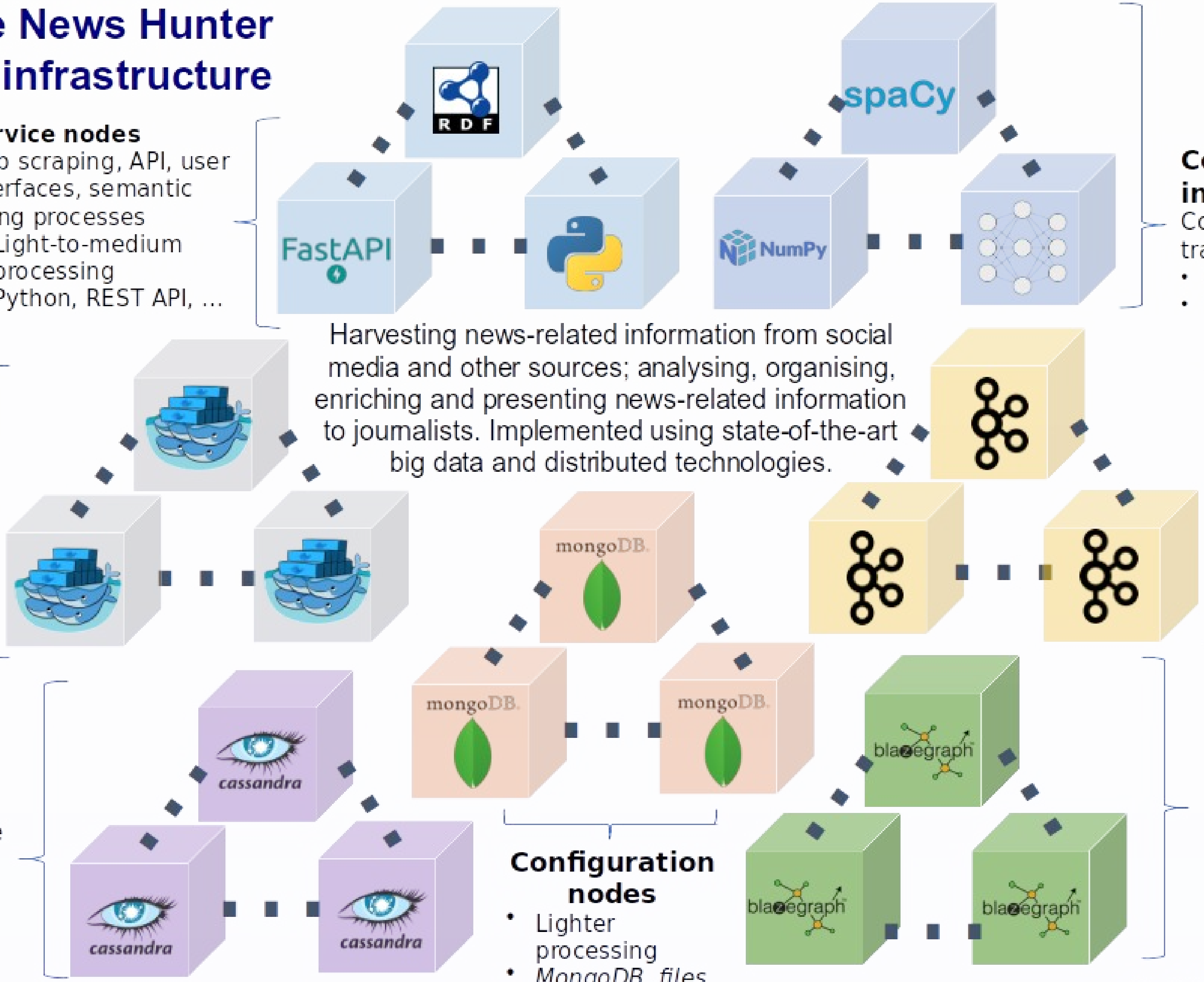
- Big data technologies:
  - Spark, Streaming Spark
  - Cloud:
    - Openstack
    - Terraform/Ansible
  - Kafka, Docker
- Big data processing:
  - big data architectures
  - the News Hunter platform
- Focus on text from the start
  - open for graphs, images...

- Societal concerns:
  - privacy, GDPR
- Textbooks:


THE DATA REVOLUTION
BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES
Rob Kitchin


O'REILLY
Spark
The Definitive Guide
BIG DATA PROCESSING MADE SIMPLE
Bill Chambers & Matei Zaharia

The News Hunter infrastructure



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

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, ...*

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*

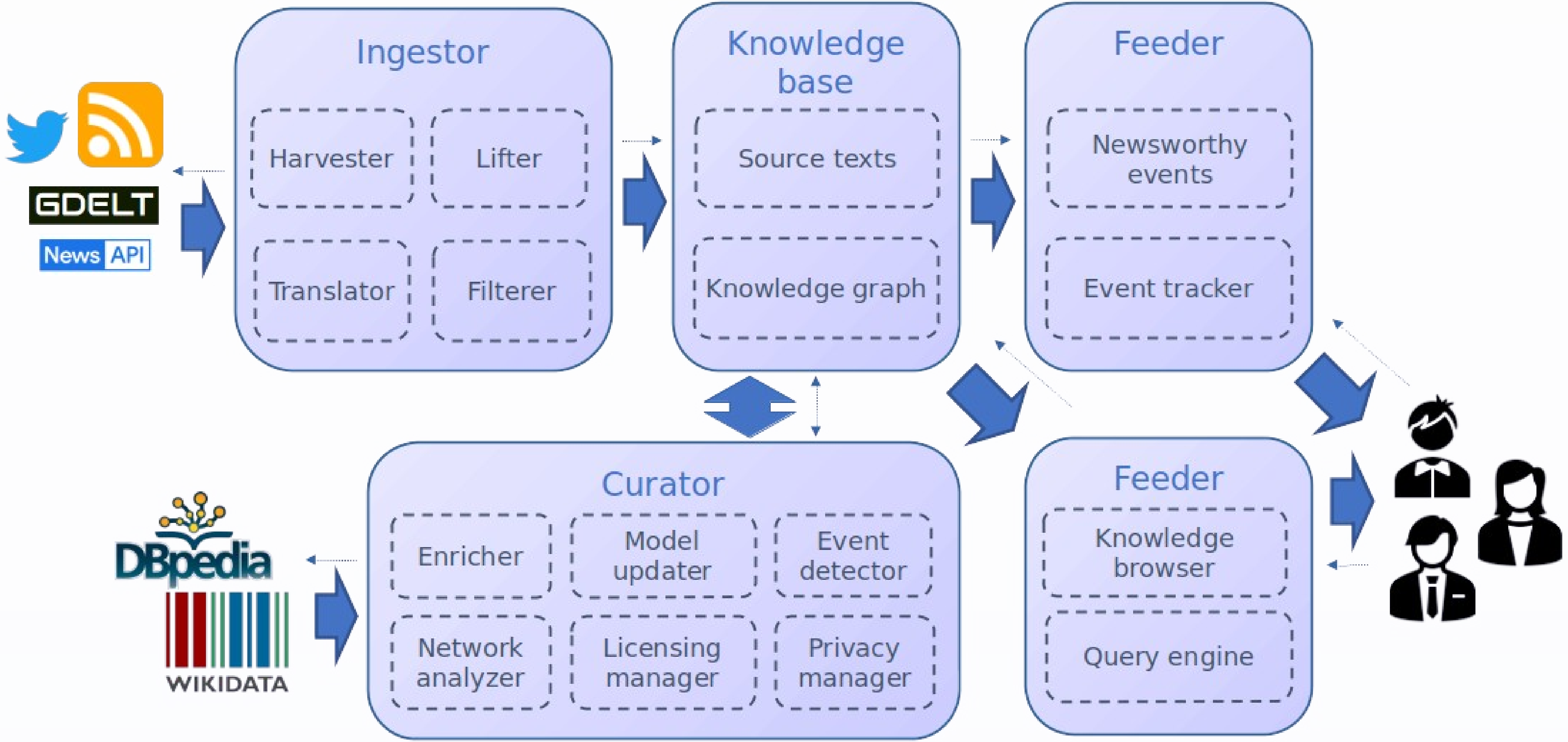
Knowledge graph nodes
News semantic representation storage.

- Disk, CPU and RAM intensive
- *Blazegraph*

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

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)

Organisation

- Teaching
  - 8 bi-weekly 6-hour sessions
    - starting Thursday September 1st 1015
    - lectures and discussions, work with assignments, presentations
  - exercises between the sessions
  - individual essay
  - group programming project (1-3 persons, 3 recommended)
- Participation
  - physical presence, no streaming/recording
  - participation at 80% of course seminars is mandatory
  - compulsory requirements are only valid the same semester

Example projects

- Harvest and analyse tweet streams.
  - analyse them for sentiment, topics and names
  - detect changes and trends
  - analyse associated images
  - can the data be geo-located?
- Similar project, but using GDELT (the Global Database of Events, Language, and Tone)
  - aggregate similar events into textual summaries or graph representations.
- Find unexpected connections between selected locations, people, and/or organisarions

Technology

- Linux (Ubuntu) and Python centric
- Virtual instances in NREC (Norwegian Research and Education Cloud)
  - http://nrec.no
- Starts with a quota of 20 instances in various sizes
  - expandable on demand
  - GPU instances should be available

Assessment

- Portfolio (55%)
  - practical assignment in groups (group programming project)
    - can be an extension of the exercises
    - proposals by October 12th 1500
  - individual, theoretical essay
    - can be about the group programming project
    - proposals by October 12th 1500
- Oral presentations (15%)
    - essay presentations: November 24th
    - project presentations: December 8th
- Written exam (30%)
  - 3 hours, December 19th 0900

Welcome to Session 1 on Thursday September 1st 1015-1600!

Follow the wiki: http://wiki.uib.no/info319