Utilizing Docker and Kafka for Highly Scalable Bulk Processing of Electronic Theses and Dissertations (ETDs)

ECE 5904: Project and Report

By Dhanush Dinesh

Chair: Dr. Edward Fox

Committee Members:
Dr. Creed Jones
Dr. Nektaria Tryfona
Dr. Prashant Chandrasekar

May 9, 2023 **Virginia Tech MEng Defense**
Blacksburg, VA 24061

VIRGINIA TECH

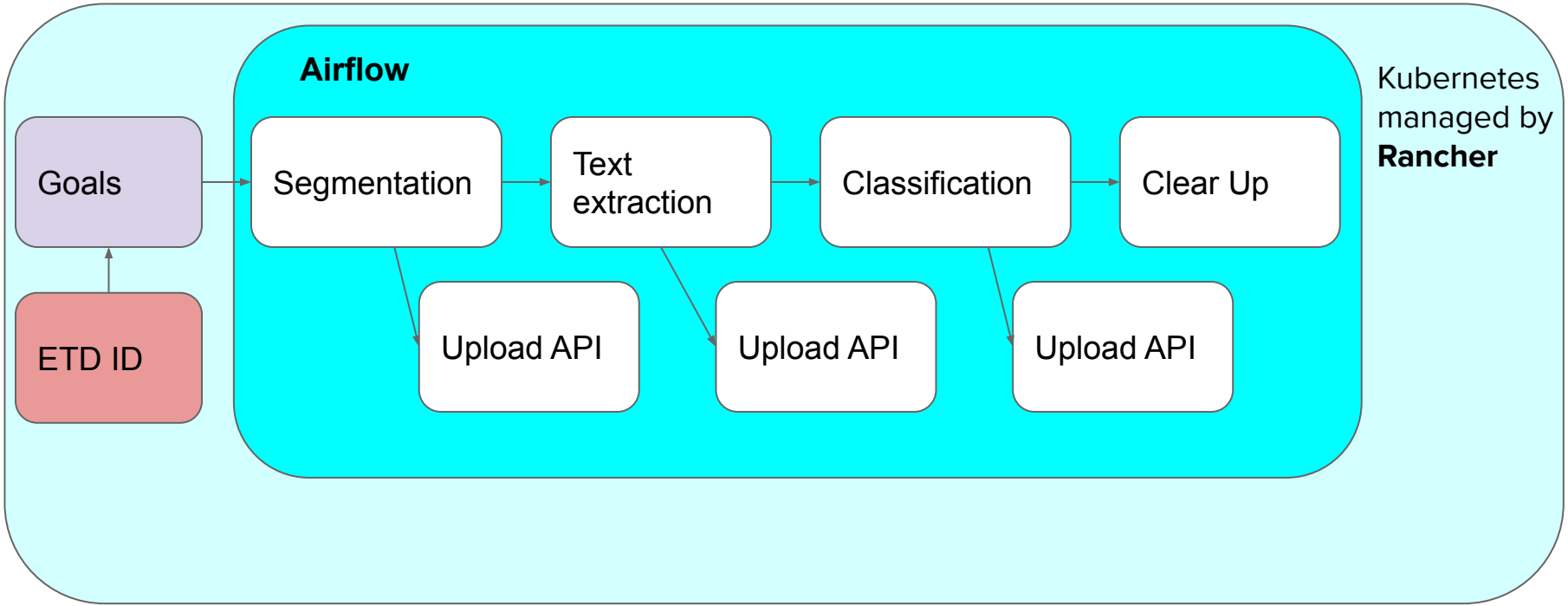
Introduction

- Introduction to Electronic Thesis and Dissertation (ETD) data
- Cloud computing and Container as a Service (CaaS)
- Advantages of containerization and platform-agnostic software
- Challenges in managing large infrastructure with numerous interconnected Docker containers [1]
- Overview of previous system architecture[2, 3]
- Challenges in processing bulk ETD data

Introduction

- Introduction to Kafka
- Advantages of Kafka
  - High performance
  - Low latency
  - High availability
- Role of Zookeeper
  - Tracking the status of Kafka nodes
  - Elects a leader of a specific partition and topic

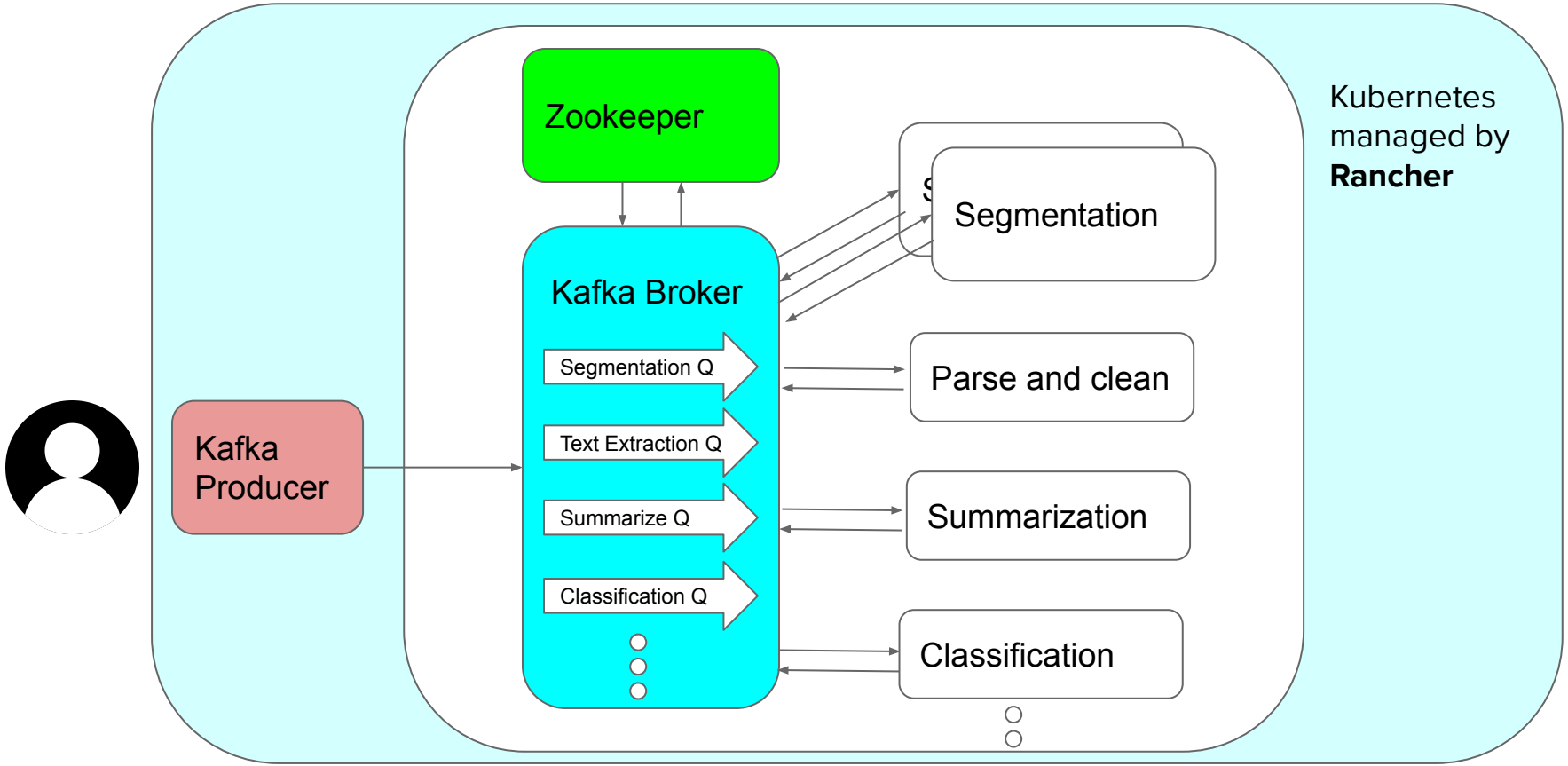
Previous pipeline architecture



Developing a Pipeline for ETD Data Processing

- Overview of previous pipeline and limitations
  - Large set up time (6- 8 min)
  - Serial processing of ETD
  - Processing one ETD at a time
- Need for a pipeline to process bulk ETD data
- Utilizing Kafka as an intermediary between services to enable parallel processing

Current pipeline architecture



Live Demo - Propagation of ETD Data through the Pipeline

cloud.cs.vt.edu

Advantages of current system

- Fast deployment and scalability
- Faster migrations
- Achieving parallel processing
- Significantly reducing processing times (one set up time)
- Bulk processing
- Decoupling of services

Conclusion

- Utilized Docker and Kafka for a high-performance and scalable bulk processing of ETD data pipeline
- Future scope for further improvements and optimization of the pipeline
    - Pipeline
    - CI/CD
    - User Interfaces

Questions ?


Thank You

References

1. Mitesh Soni. End to end automation on cloud with build pipeline: The case for DevOps in insurance industry, continuous integration, continuous testing, and continuous delivery. In 2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), pages 85–89, 2015. doi:10.1109/CCEM.2015.29.
2. Harish Babu, Manogaran Pallavi Sisodiya, Yuze Li, Aaron Travasso, Anmol Shukla. 2022. Team 5 final submission CS 5604: Information storage and retrieval. URL: http://hdl.handle.net/10919/114078
3. Suraj Gupta, Xingyu Long, Yash Mahajan, Mohit Thazhath, Hsinhan Hsieh, Alex Hicks, Cherie Poland. 2020. INT team final submission CS5604: Information storage and retrieval. URL: http://hdl.handle.net/10919/101544

2022 teams system architecture

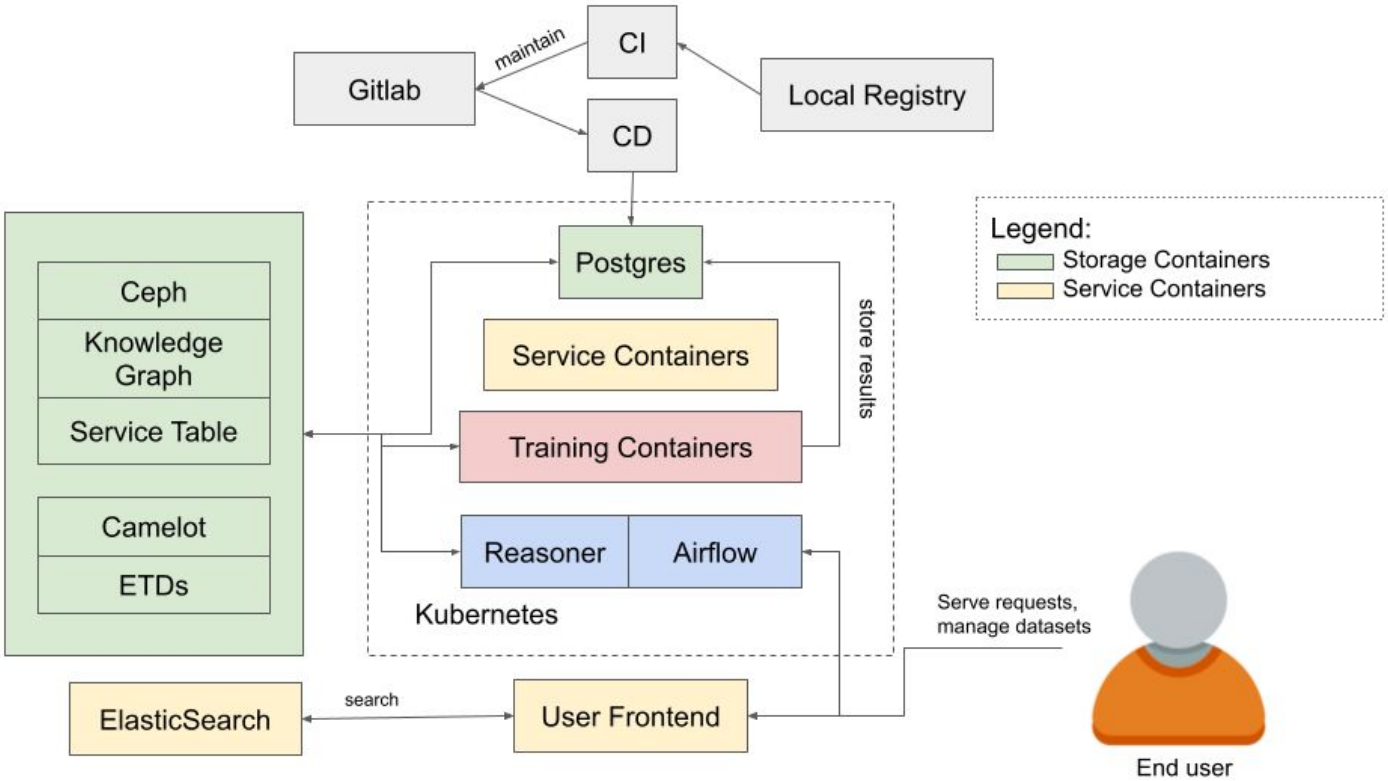


Fig: System Architecture 2022 [2]

2020 teams system architecture

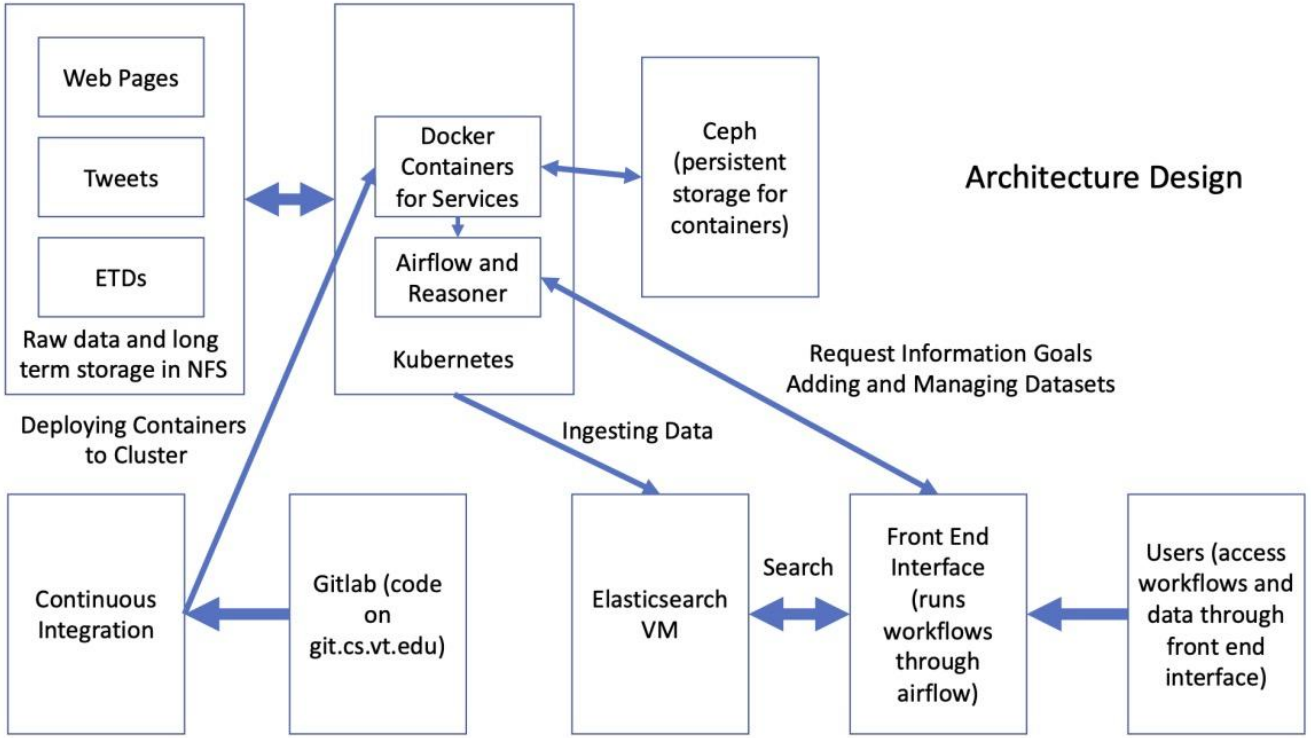


Fig: System diagram for information retrieval and analysis system 2020 [3]

Discovery cluster - Rancher

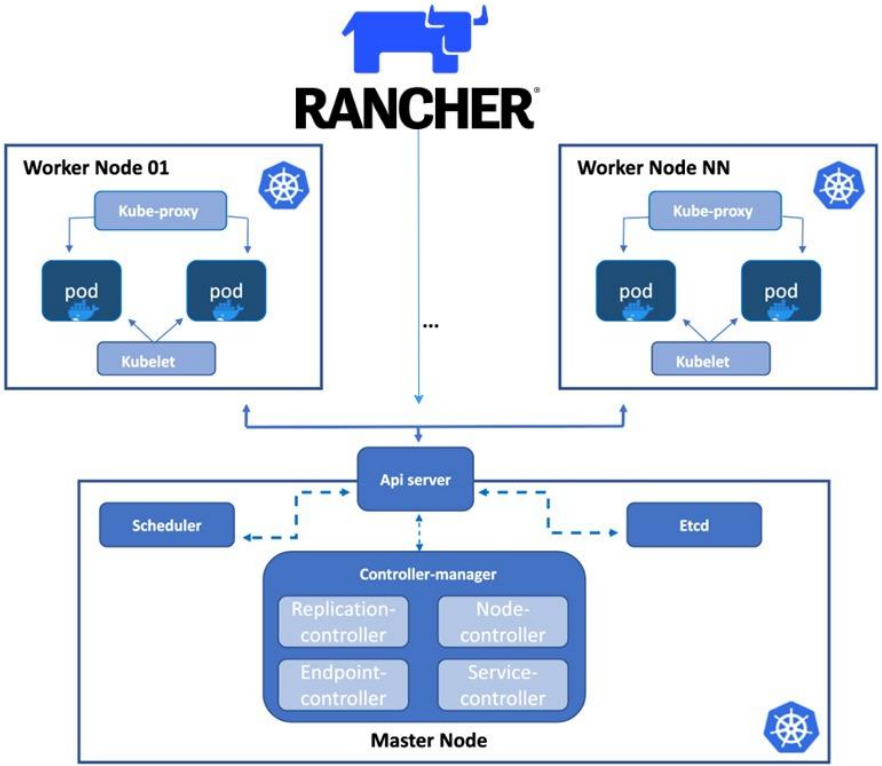


Fig: Rancher architecture as used in discovery cluster