2020-12-16 - Special Topic - VIVO Kafka Ingest

Date

16 Dec 2020

- Time: 10:00 am, Eastern Time (New York, GMT-04:00)
- See in your timezone

Call-in Information

To join the online meeting:

- Go to: https://lyrasis.zoom.us/my/vivo1
- One tap mobile:
  - US: +16699006833,,9358074182# or +19292056099,,9358074182#
- Or Telephone:
  - US: +1 669 900 6833 or +1 929 205 6099 or 877 853 5257
  - Meeting ID: 935 807 4182
- International numbers available: https://zoom.us/u/aeANHanzED

Slack

- https://vivo-project.slack.com
  - Self-register at: http://bit.ly/vivo-slack

Attendees

⭐ Indicating note-taker

1. Michel Héon
2. Ralph O'Flinn
3. Andrew Woods ⭐
4. Mike Conlon
5. Sandra Mierz
6. Benjamin Gross
7. Benjamin Kampe
8. Brian Lowe
9. Bruce Herbert
10. Christian Hauschke
11. Huda Khan
12. Maxime Belanger
13. Nicolas Dickner
14. Paul Albert
15. Sarbajit Dutta
16. Tatiana Walther
17. William Welling
18. Rachid Belkouch

Context

1. Team UQAM and team TIB had a nice first investigative sprint with the objective to learn about Apache Kafka. We would like to start the work of the data ingest task force as soon as possible. We would like to present the outcome of our first investigation as long as it is fresh, and to see if we have alignment with others, and to get valuable feedback.

Agenda

1. Introduction: Apache Kafka as a central component for data ingest in VIVO? (10 minutes by Michel Héon)
   Entry point of the presentation 2020-12-16 VIVO-DataConnect ORCID Demo and https://github.com/vivo-community/vivo-data-connect/tree/POC-extract-orcid for code
2. Work at UQAM (5-10 min)
3. Work at TIB (5-10 min)
4. General discussion

Recording

- 2020-12-16-vivo-kafka.mp4

Notes

Draft notes in Google-Doc

VIVO - DataConnect - ORCID - UQAM Demo

1. Walking through context and use case
    - "A professor wishes to add the reference to a scientific article, irrespective of whether he chooses ORCID or VIVO, the information he will enter in either of these platforms will be mutually updated"
2. Goal of using Kafka with VIVO:
    - VIVO is a component in the enterprise, instead of the center
3. Main idea of Kafka
    - An event-driven messaging system
    - Allow for multi-to-multi producers and consumers
4. Recent sprint
    - Ingest ORCID data into VIVO
5. Walkthrough of flow:
    - Extract all ORCID_IDs associated with UQAM members
        - './orcid_get_all_records.sh'
        - Converting ORCID JSON into RDF
    - Transform RDF into VIVO representation
    - Send to Kafka
        - Then pass to VIVO
6. Demo
    - 25,171 statements pushed through Kafka
    - 763 users, with name, org, and competencies
7. Summary
    - The ORCID ontology needs to be refined and clarified.
    - The mapping between ORCID and VIVO also needs to be worked on
    - The structure of the Kafka message has to be designed to respect the add/delete/modify record actions
    - Several minor bugs need to be fixed in the scripts.
8. Future plans
    - Building a POC VIVO Kafka ORCID
    - Proving the architecture to operate in event-driven and real-time mode
    - Getting POCs to Java
    - Redesigning the mapping process, ORCID ontology structure and message structure

TIB

1. Using Kafka as a consumer of VIVO messages
2. Tasks
    - Listener in VIVO to capture internal changes
    - Producer to send to Kafka
3. VIVO Kafka-Module
    - ModelChangedListener and ChangeListener
    - Kafka start-up listener
    - Http connection
4. VIVO producer
    - Spring-boot service
5. Code will be in GitHub soon

Discussion

1. Interest in the architecture presented
    - Allows for integration with any number of source systems
2. This initiative allows for outputs from VIVO
3. Can past initiatives be used in this context?
    - ..such as ORCID-to-VIVO
    - ..such as Dimensions-to-VIVO
4. Could this support large-scale ingest?
    - +100M triples?
    - Are there Kafka buffer limits, throttling
    - Kafka is designed for "big data"
5. Next steps
    - TIB: VIVO to other systems by Feb/May
    - TIB: Other systems to VIVO... timeline is further out
    - UQAM: Ingest timeframe in Q1 of 2021
    - Next meeting in January? - Ralph to organize

Actions

-