Linkedin DataHub 💡

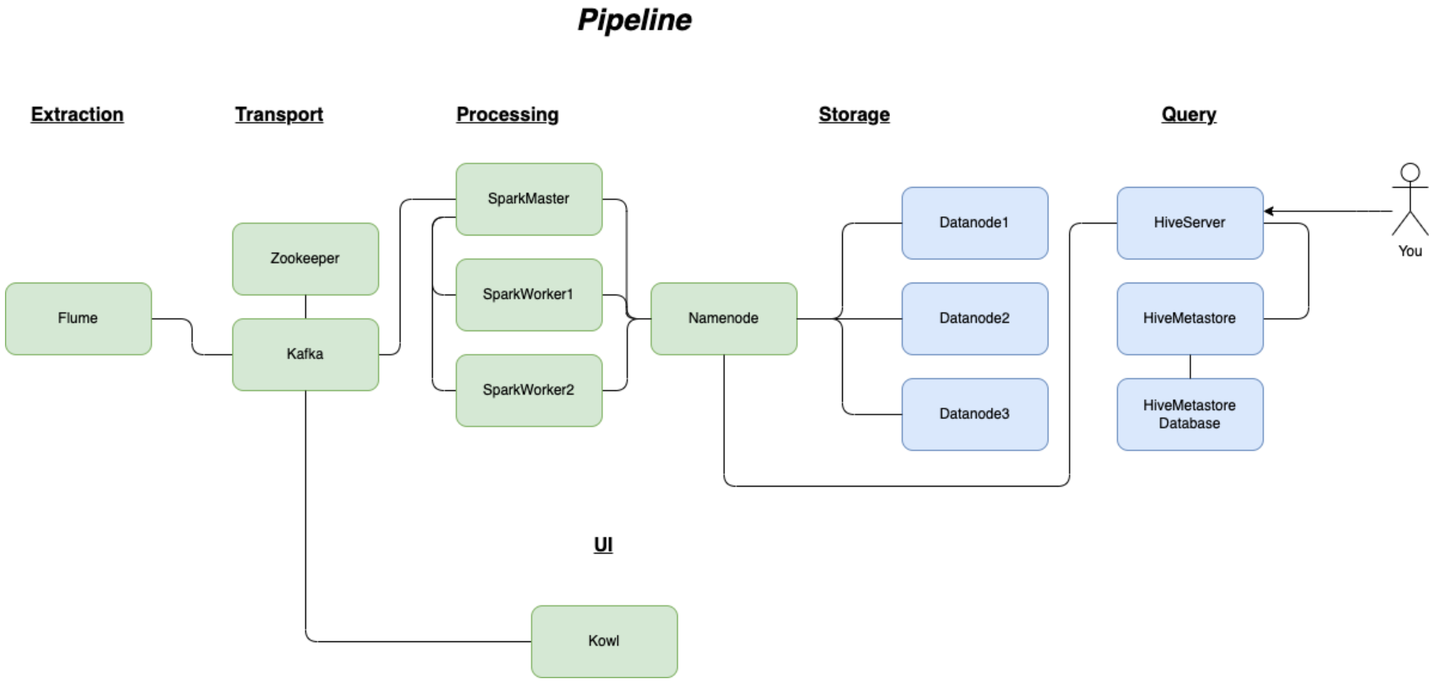
Big Data E22

Nikolai Emil Damm - nidam16@student.sdu.dk
Bende Siewertsen - besie18@student.sdu.dk

Context

What have we been doing?

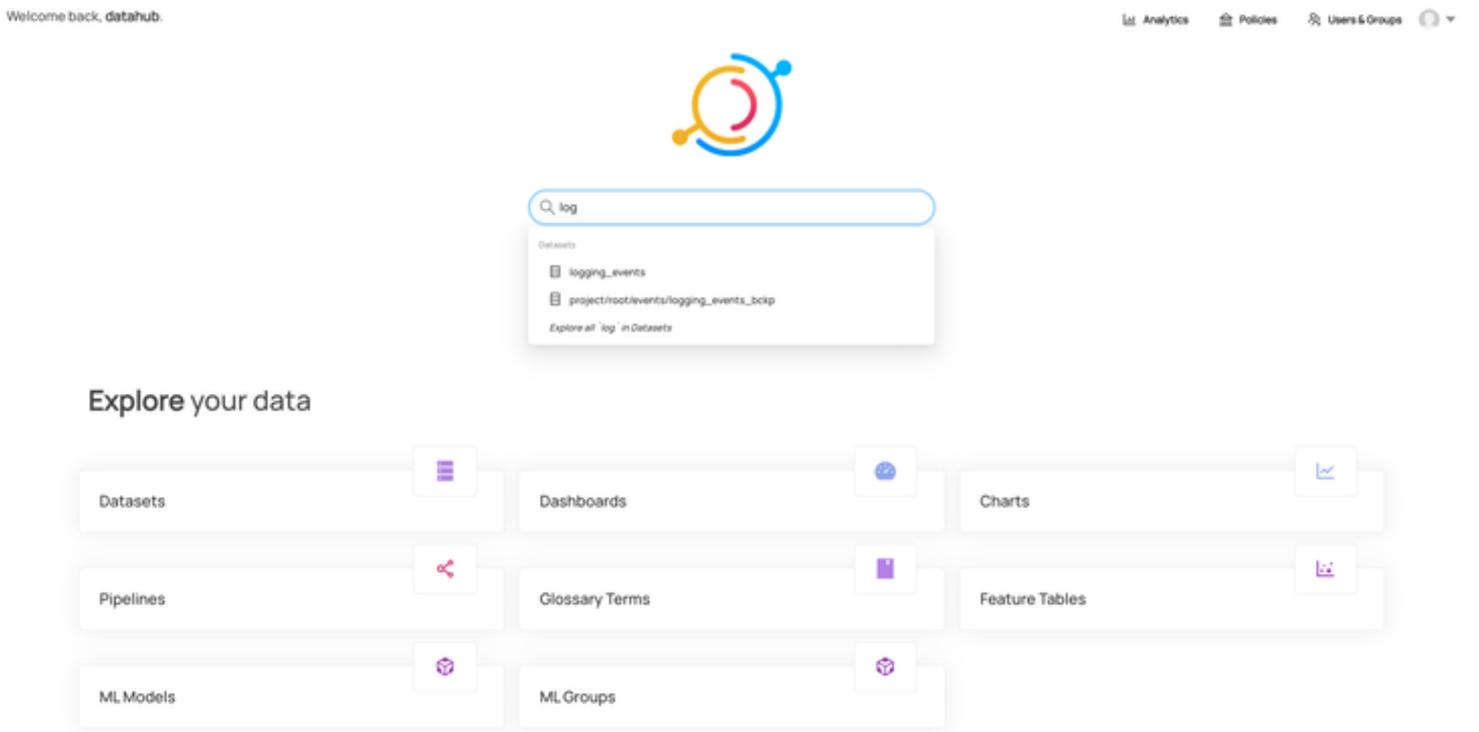
- A data-pipeline to ingest, process, store and query data.
- Last time we looked at Hive a Data Warehouse built on top of Hadoop.



Context

What are we doing today?

- Experimenting with Linkedin DataHub.
- Linkedin DataHub has two main approaches to ingest metadata, a pull-based approach, and a push-based approach.
  - The pull-based approach allow us to ingest data from a source on a schedule.
  - The push-based approach allow us to ingest data send to the platform.
  - The push-based approach is generally accepted as the best approach, as it ensures your metadata is consistent with your source, but it requires granular control of the source.
- As these exercises aim to give you some familiarity with Linkedin DataHub we will use the pull-based approach, as this allows you to do most tasks from the platform.



Exercise 1

Setting up Linkedin DataHub with Docker

- To work with Linkedin Datahub we need to setup a set of services it depends on.
- To do this we will use its official CLI
    1. Follow the first five steps on: https://datahubproject.io/docs/quickstart/

Exercise 2
Organizing metadata

- As more and more metadata is added to the platform the need for organising the metadata becomes important. To do this we have three options. Domains, Glossaries, and Tags.
    - Domains are used to organise metadata according to what domain they belong to.
    - Glossaries are used to organise metadata to business terms, such that metadata can be found from generally accepted terms e.g. metadata related to AccountSavings.
    - Tags are also useful for organising metadata, but there is not a strict policy on how to use them. They can be used to e.g. add versioning to an entity, or whether an entity is legacy.

1. Play around in the UI, and see if you can add the following:
    1. A new domain
    2. A new Glossary Group
    3. 1-2 new glossaries in your new Glossary Group.
    4. 1 new glossary that inherits from another glossary.
2. Now try and use your glossaries on some of the existing dummy data.
3. What metadata does a glossary term that is inherited by another term show? Can you explain this?

Exercise 3

Checkout analytics!

- DataHub comes with a nice analytics overview. Here we can gain an overview of how the platform is used.

1. Play around in the Analytics UI.
2. Can you find analytics for your newly added Domain?

Exercise 4

Add a Kafka Ingestion Source

- Obs! The Kafka Ingestion Source is not able to extract metadata on a topics key and values without setting up Kafka schemas. As this is not something we have touched upon, we will only set up a Kafka Ingestion source to extract topic names.

1. Compose the stack in ./exercise07/docker-compose.yml to set up a simple Kafka cluster + Kowl.

2. Use the ingestion UI in Linkedin DataHub to create a Kafka Ingestion source that uses the Kafka broker in the stack you just set up.

3. Check that the ingestion worked, and you can see newly added topics in Linkedin DataHub.

- You can remove SSL authentication by going to the YAML view and deleting the lines related to SSL authentication.

- If you want to keep the stateful ingestion turned on (under advanced) you need to platform_instance = “somename” to your config. This can also be done in the YAML view.

Exercise 5

Add Linkedin DataHub's internal MySQL database as an ingestion source

- Linkedin DataHub creates a new MySQL database as part of its stack. We will try to add this database as an ingestion source for some data inception.
- The MySQL database has the following config:
  - host_port: mysql:3306
  - database: datahub
  - username: datahub
  - password: datahub
  - profiling: enabled: true # this allows the ingestion source to collect metadata on e.g. sample data and other niceties. It is disabled by default as it can mean a performance hit.

1. Set up a new MySQL Ingestion Source using the above config.
2. Check out your new metadata entities. Can you spot some of the niceties that come with enabling profiling?

Exercise 6

Alice in Metaverse Part 1 👸🏼

- Your task is to create a new table in the database "datahub" and ingest all the words from Alice in Wonderland!

1. Navigate to ./lecture07-exercises/upload-alice
    1. Change the python file to:
        1. Create a new database called "alice"
        2. Read alice-in-wonderland.txt
        3. Upload the individual words to the table
2. Look at datahub, and see what changed and what you can see about alice in datahub!

Exercise 7

Alice in Metaverse Part 2👸🏽

- Now that we have Alice in Linkedin DataHub, we should add some more metadata to describe the tables contents.

1. Add the following to the table alice:
    1. a new domain called books
    2. a new glossary called word
    3. a new tag with the author name
    4. a description of what the table contains
    5. yourself as the owner (you are called datahub, you can change this under: your profile -> edit profile)