ROOTSTRAP

Airflow Summit

Building a Scalable & Isolated Architecture for Preprocessing Medical Records

Mikaela Pisani - Anthony Figueroa

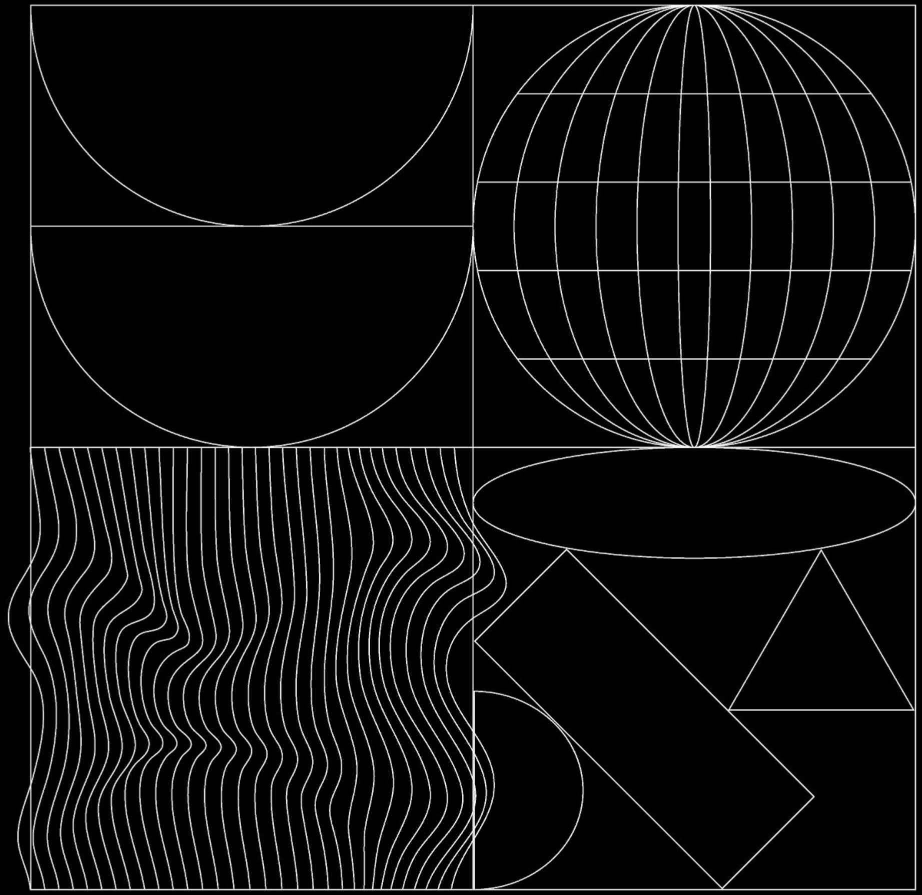


Table Of Contents

Motivation

Process Medical Records at Scale

Why?

Physicians don’t have a lot of time to review historical data during an appointment. They usually have to ask.

Historical data often contains errors

It’s hard to **Summarize**.

Validation of the problem

Interviewing Physicians

“Having a view that **summarizes** main points and trends of current hospital stay, as well as points important to **my specialty.** “

“Too many clicks; Lack of good data visualization; clunky and **hard to read** with excessive note bloat.”

“Use AI to provide a current updated **summary** of patient clinical encounters within a requested date range”

“Having results in **one single place**. Not scattered everywhere”

“I spend more time dealing with my EMR than attending my patient”

Challenges

NLP

A lot of data is non-structured, written using plain english but using medical lexicon and abbreviations.

Each patient has decades of history, potentially gigas of data if we include imagenology .

Big question:

What is relevant?

```
<TEXT><![CDATA[

Record date: 2088-03-08

Patient Name: YOUNT, PATRICIA;  MRN: 4711083
Dictated at: 03/08/2088 by BRANDON VICENTE, M.D.

RENAL CONSULT NOTE

It is a pleasure seeing seeing Ms. Yount in consultation.  She was referred to us by Dr.
also has a long history of hypertension.  She retired from IBM in 2085 after working the
well.  She denies headaches, nausea, vomiting, abdominal pain.  She also denies shortnes
fever, chills.  She is taking her medications regularly.

ALLERGIES
No known drug allergies.

PAST MEDICAL HISTORY
1.  Diabetes mellitus for about seven years.
2.  Hypertension.
3.  PVD
4.  Hypercholesterolemia


SOCIAL HISTORY
She denies smoking or drinking.  She has one son living with her.


FAMILY HISTORY
Positive for diabetes.

MEDICATIONS
Norvasc 10 mg q. day,

Glucophage 500 mg twice a day,
```

Data Quality

- Typing errors
- OCR Errors (data was imported from older EMRs, or from handwritten notes)
- Contradictory information
- Different medical conventions for namings and abbreviations
- Non-structured (we can’t distinguish easily what’s important and what isn’t important)
- A LOT of data.

Semantics

- We had to convert plain English to something manageable and semantic.
- It’s not a challenge of language recognition, it’s fairly more complicated: many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT. Hierarchies, definitions, and other relationships and attributes.
- We focused our efforts on transforming plain text to a semantic network.
- Unified Medical Language System® (UMLS®):
- UMLS Metathesaurus: 215+ Vocabularies

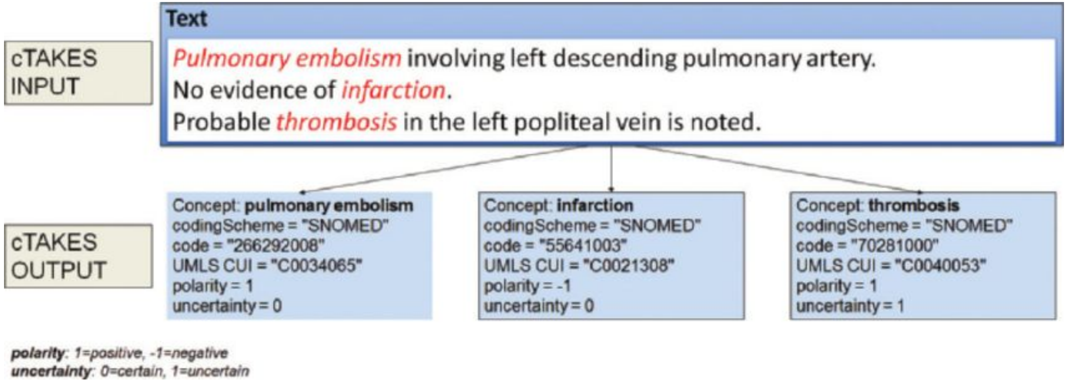
Apache CTakes

Apache cTAKES™ is a natural language processing system for extraction of information from electronic medical record clinical free-text.

Input ⇒ NLP Files (potentially XML).

Output ⇒ XMI files with a representation of the semantic network

Output is a graph with thousands of nodes and relationships.



Why Airflow?

Data Preparation and Processing

- Before using CTakes, data cleaning has to take place.
- Diversity of problems (depending of the source of the Medical Record).
- Traceability and reproducibility is key.
- The better data preparation, the better outputs from CTakes
- Examples of Tasks:
  - Fix typos and OCR problems (dates 01-01-2088 instead of 01-01-1988)
  - Separate each appointment
  - Run CTakes scripts
  - Measure quality of results
  - Process output: Summarize and Personalize

Reproducibility

- Try different DAGs. Different versions of each task.
- Isolate inputs and outputs of each task, to recognize opportunities to improve.
- Each step should store its outputs (we use Redshift).
- Data Preparation is challenging due to its diversity.
- Post-processing (Summarization) is challenging because it’s ambiguous.

Scalability

- Being able to process thousands/millions of medical records.
- Parallelize everything that can be parallelized
- Machine Learning algorithms for summarization require a lot of data

Architecture

1 Data Sources

Assemble Data (from EMRs and public available data)

Analyze

We manually analyzed medical records and detected different types of problems.

Cloud Storage

Data will initially be stored in a cloud storage platform: Amazon S3

DAG and infrastructure

After an initial import, the input/output of each task is done in Redshift.

2

Data Preparation

Prepare Data making it ready for CTakes

Tasks

Create Tasks for each one of the problems detected. Use Redshift table partitions to store intermediate results

Measure & Trazable data

We are able to measure inputs and outputs of each task, so we can improve each one of them on each increment.

Spark

Better parallelism, and better CPU utilization for computing intensive work.

Isolated

Plus, if there is an error we have to change only one task.

3 CTakes

Convert clean Medical Records to a semantic network, following UMLS standard.

CTakes

Just another Airflow task that executes CTakes command line processor in a Kubernetes Pod

Reproducibility

Each input/output has to be stored, so we can analyze opportunities to improve.

Customize

CTakes comes with a pre-trained NLP processor that can be customized.

4

Summarize

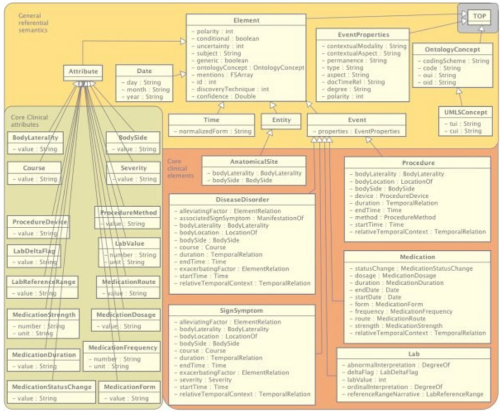
Process a Semantic network recognize what’s important

Machine Learning

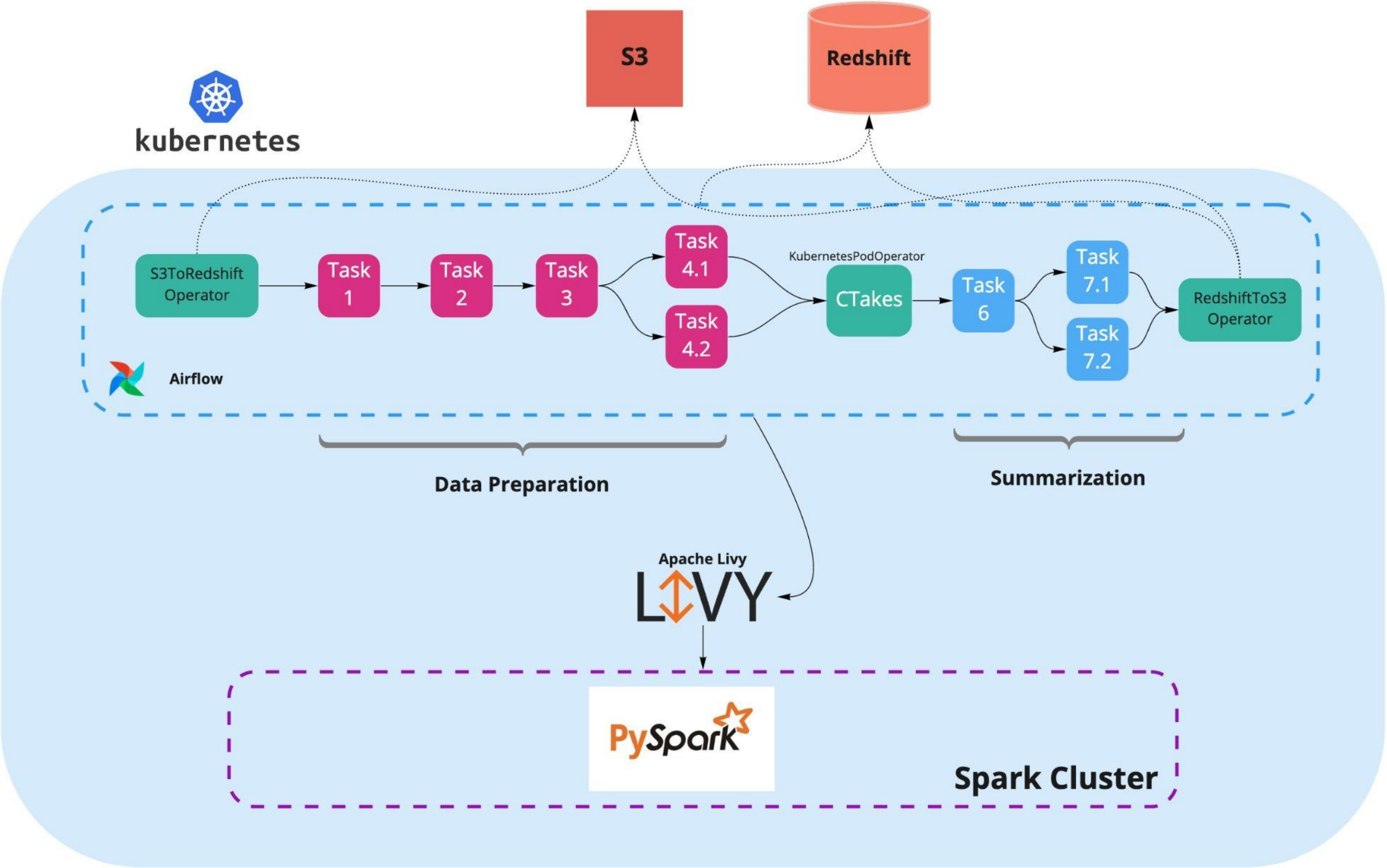
Train ML Models to be able to summarize semantic networks. Compare different results.

Spark MLlib

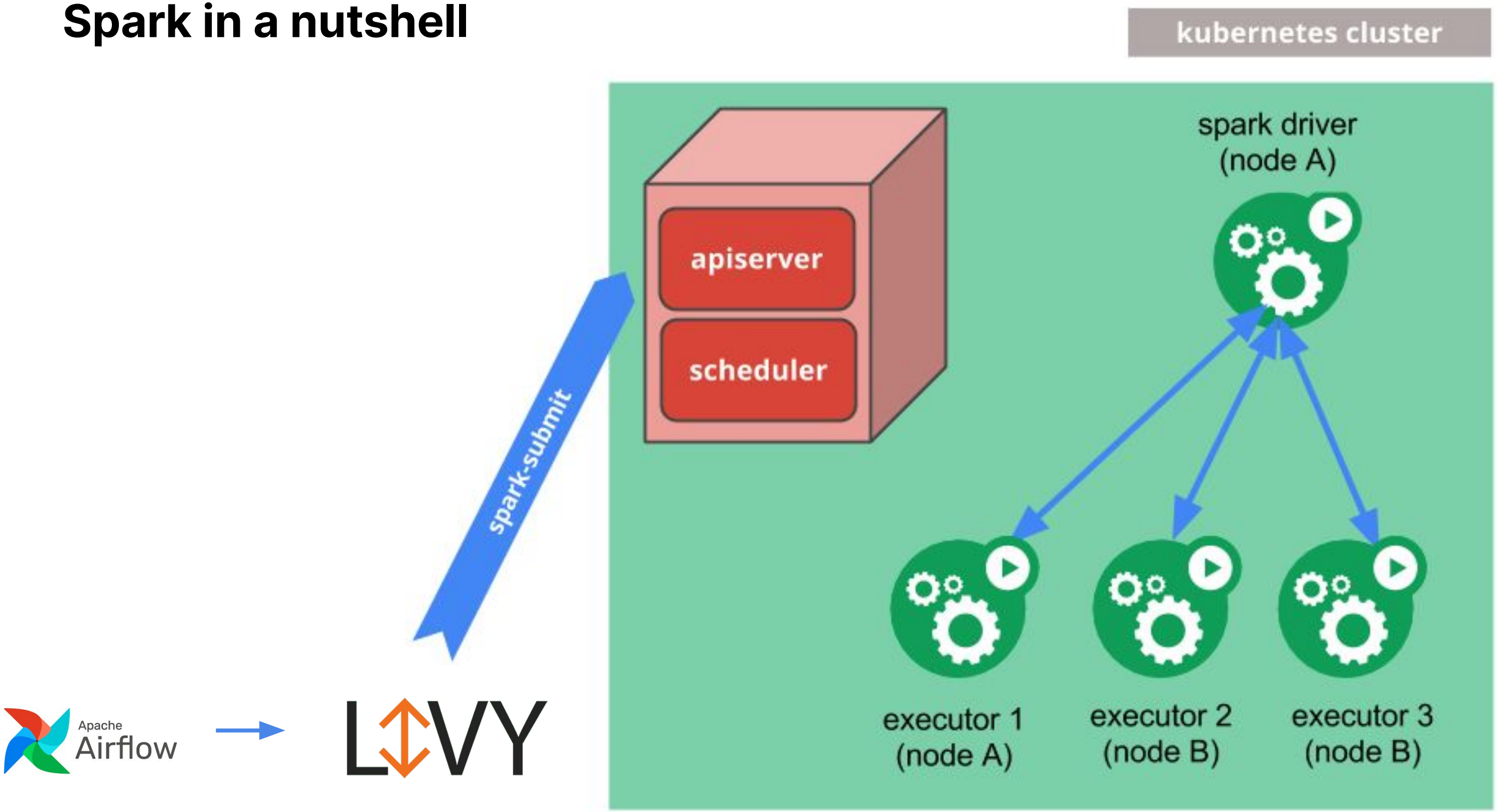
We are able to load ML models in Spark.



Architecture



Spark in a nutshell



https://spark.apache.org/docs/latest/running-on-kubernetes.html

Conclusions

Airflow

- Reproducible & trackable tasks in the data pipeline
- Integration with Kubernetes (EKS), Git, Apache Livy, S3 and Redshift
- Being able to see DAG execution in a visual way
- DAG versions are extremely important

Kubernetes and Spark

- Isolation tasks, scalable nodes
- Parallel processing for large datasets
- Different versions of Airflow had different challenges
- DevOps and environments configuration take a lot of effort

Work in Progress

- Spark Streaming
- Break-down architecture into more independent units and open source.
- Summarization: Strongly related to dimension reduction.
- Personalization: “Important” means different things to different people/specialties.

Thanks!

please contact our Rootstrap team
for a consultation:

hello@rootstrap.com

Los Angeles

8913 1/2 Sunset Blvd.
West Hollywood, CA 90069
(310) 907-9210

New York City

455 West 23rd St. Suite 1B
New York, NY 10011

Montevideo

Sarandi 690D EP,
Montevideo, UY 11100
(+598) 2909-0655

Buenos Aires

Av. Corrientes 800, Caba,
Buenos Aires, AR C1008