Implementation of topic modeling in industry

Madhurima Nath

Agenda

Intro to topic modeling

Available algorithms

Little bit of math

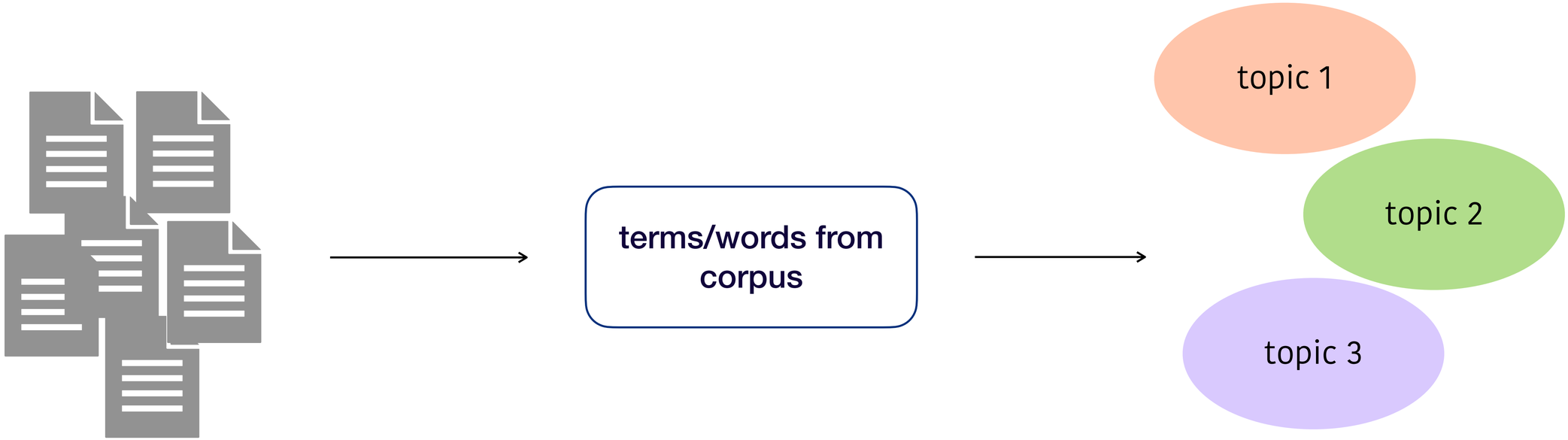
Kaggle dataset

Large scale end-to-end implementation

Q & A

What is topic modeling?

Topic modeling is a part of natural language processing (NLP) which enables end-users to identify themes/topics within a collection of documents. It has applications in multiple industries for text mining and gaining relevant insights from textual data.

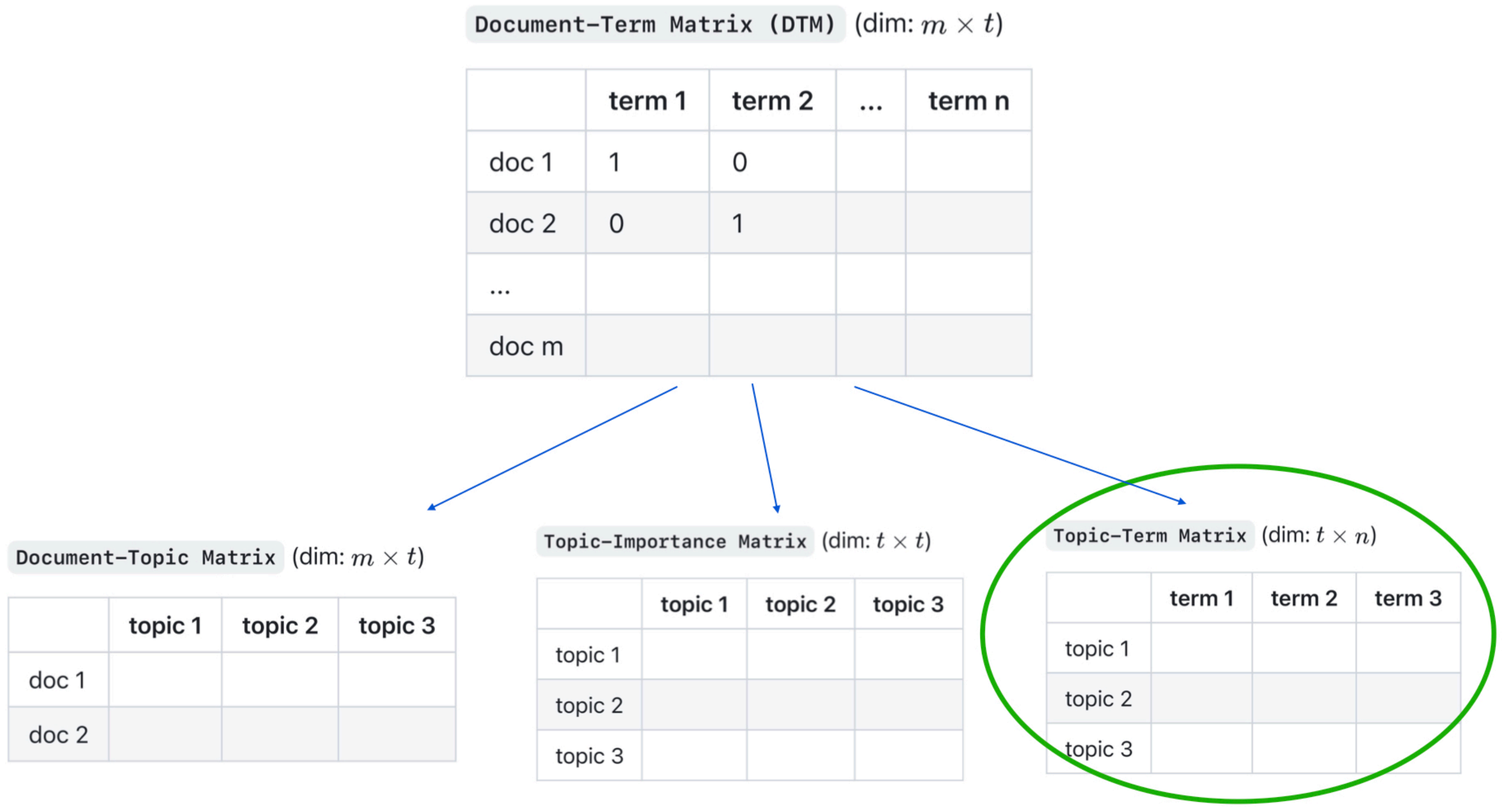


What is topic modeling?

- Unsupervised machine learning problem
- Unlike text classification or clustering, it is not aiming to find similar documents
- Makes clusters of 3 types of words - co-occurring words, distribution of words, and histogram of words topic-wise
- Some well-known algorithms are
  - Latent Semantic Analysis (LSA)
  - Probabilistic Latent Semantic Analysis (pLSA)
  - Latent Dirichlet Allocation (LDA)
  - Hierarchical Dirichlet Process (HDP)
  - Non-Matrix Factorization (NMF)
  - BERTopic

Topic modeling algorithms

Most algorithms try to decompose the document-term matrix into two or more matrices to obtain the matrix containing terms and topics.



Schematic to understand topic modeling algorithms

Topic modeling algorithms - Step 1

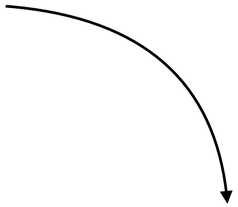
Start with the conversion of a textual corpus into a Document -Term Matrix (DTM), a table where each row is a document, and each column is a distinct word.

Corpus	
doc1	I like books
doc2	I recently read two bestseller books
doc3	Some movies are based on bestseller books

Topic modeling algorithms - Step 1

Create Document-Term Matrix by term frequency (TF), TF-IDF, or converting into embedding representation etc.

Corpus	
doc1	I like books
doc2	I recently read two bestseller books
doc3	Some movies are based on bestseller books



	I	like	books	recently	read	two	bestseller	some	movies	are	based	on
doc1	1	1	1	0	0	0	0	0	0	0	0	0
doc2	1	0	1	1	1	1	1	0	0	0	0	0
doc3	0	0	1	0	0	0	1	1	1	1	1	1

Topic modeling algorithms - Step 2

`Document-Term Matrix (DTM)` (dim: $m \times t$)

	term 1	term 2	...	term n
doc 1	1	0		
doc 2	0	1		
...				
doc m				

`Document-Topic Matrix` (dim: $m \times t$)

	topic 1	topic 2	topic 3
doc 1			
doc 2			

`Topic-Importance Matrix` (dim: $t \times t$)

	topic 1	topic 2	topic 3
topic 1			
topic 2			
topic 3			

`Topic-Term Matrix` (dim: $t \times n$)

	term 1	term 2	term 3
topic 1			
topic 2			
topic 3			

$$DTM\ (m \times n) = Document - Topic\ (m \times t)\ Topic - Importance\ (t \times t)\ Topic - Term\ (t \times n)$$

Topic modeling algorithms - Step 2

Decompose the **Document-Term Matrix DTM** and extract topics.

- LSA uses matrix factorization - Singular Value Decomposition (SVD)
- pLSA uses probabilistic model, calculates the joint probability of seeing a word and a document together as a mixture of conditionally independent multinomial distributions
- LDA uses Dirichlet priors to estimate the document-topic and term-topic distributions in a Bayesian approach
- HDP mixture model is a non-parameteric generalization of LDA - number of topics can be unbounded and learnt from the data
- NMF decomposes the matrix into 2 lower rank matrices
- BERTopic uses custom class-based TF-IDF to extract topics from clusters formed from embedding representation

$$DTM\ (m \times n) = Document - Topic\ (m \times t)\ Topic - Importance\ (t \times t)\ Topic - Term\ (t \times n)$$

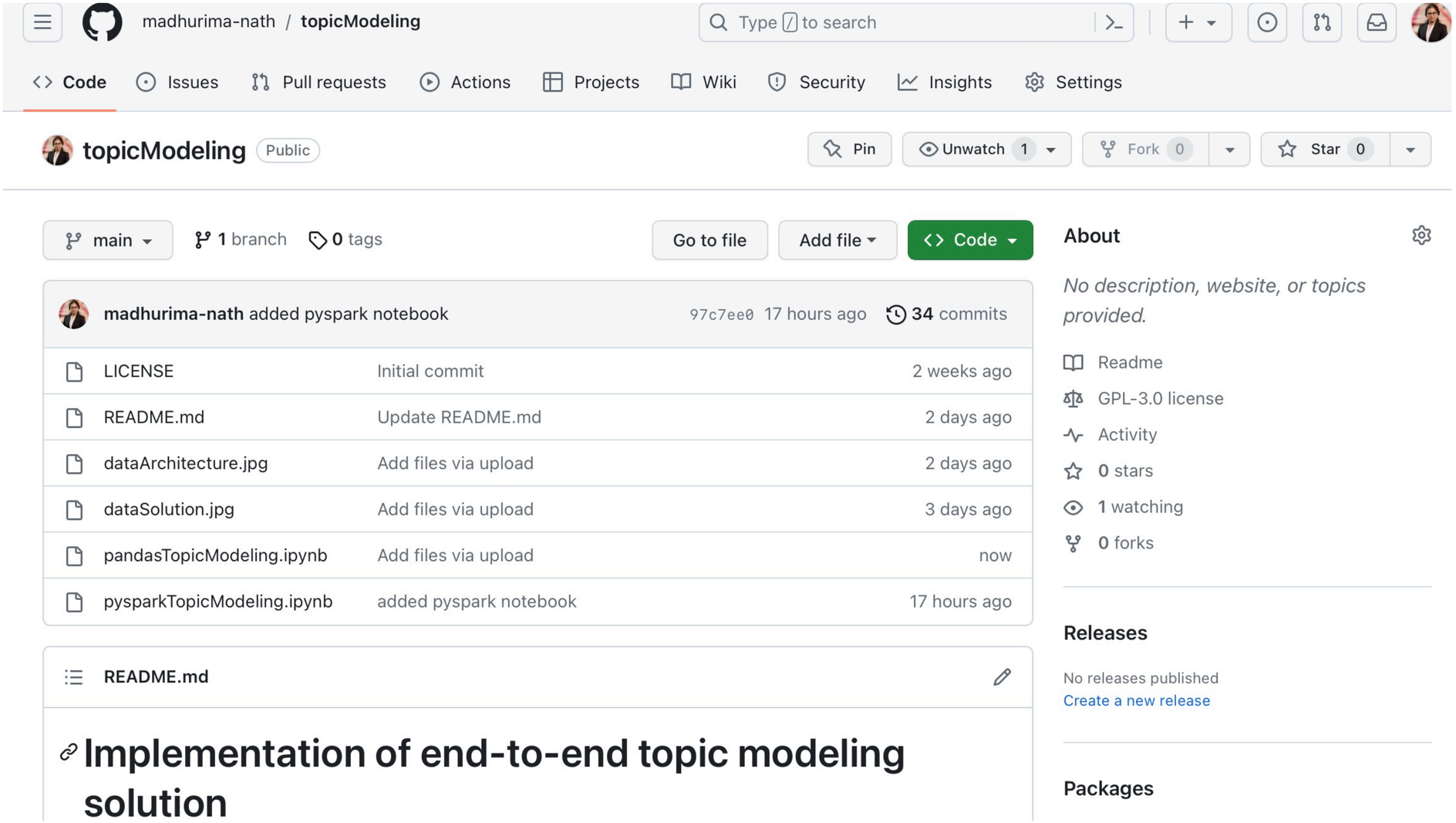
Kaggle dataset

- Researchers have access to large online archives of scientific articles. Tagging or topic modeling provides a way to give token of identification to research articles which facilitates recommendation and search process.

- The dataset has abstract and title for a set of research articles and each article is assigned to one or more of the following topics:
    - Computer Science
    - Physics
    - Mathematics
    - Statistics
    - Quantitative Biology
    - Quantitative Finance

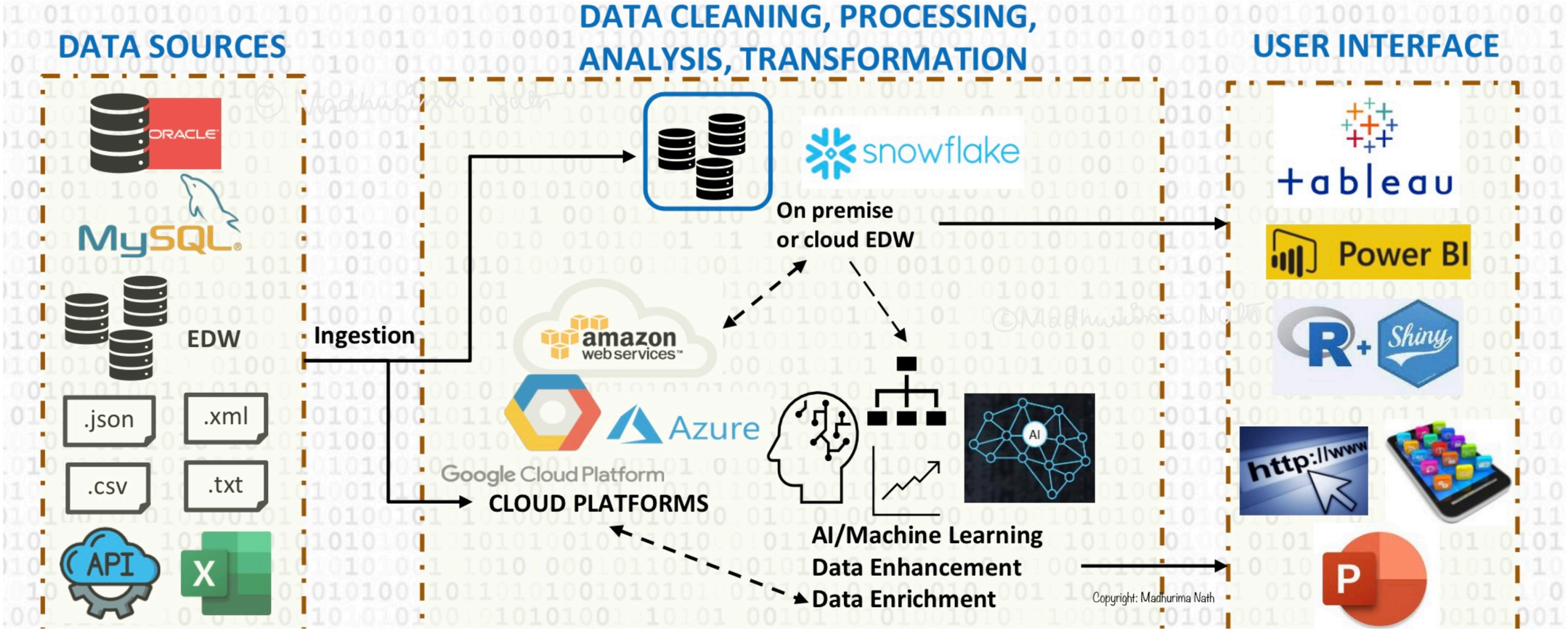
- Download the data here

GitHub link


madhurima-nath / topicModeling
Type / to search
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
topicModeling
Public
Pin
Unwatch 1
Fork 0
Star 0
main
1 branch
0 tags
Go to file
Add file
Code
madhurima-nath added pyspark notebook
97c7ee0 17 hours ago
34 commits
LICENSE Initial commit 2 weeks ago
README.md Update README.md 2 days ago
dataArchitecture.jpg Add files via upload 2 days ago
dataSolution.jpg Add files via upload 3 days ago
pandasTopicModeling.ipynb Add files via upload now
pysparkTopicModeling.ipynb added pyspark notebook 17 hours ago
README.md
Implementation of end-to-end topic modeling solution
About
No description, website, or topics provided.
Readme
GPL-3.0 license
Activity
0 stars
1 watching
0 forks
Releases
No releases published
Create a new release
Packages

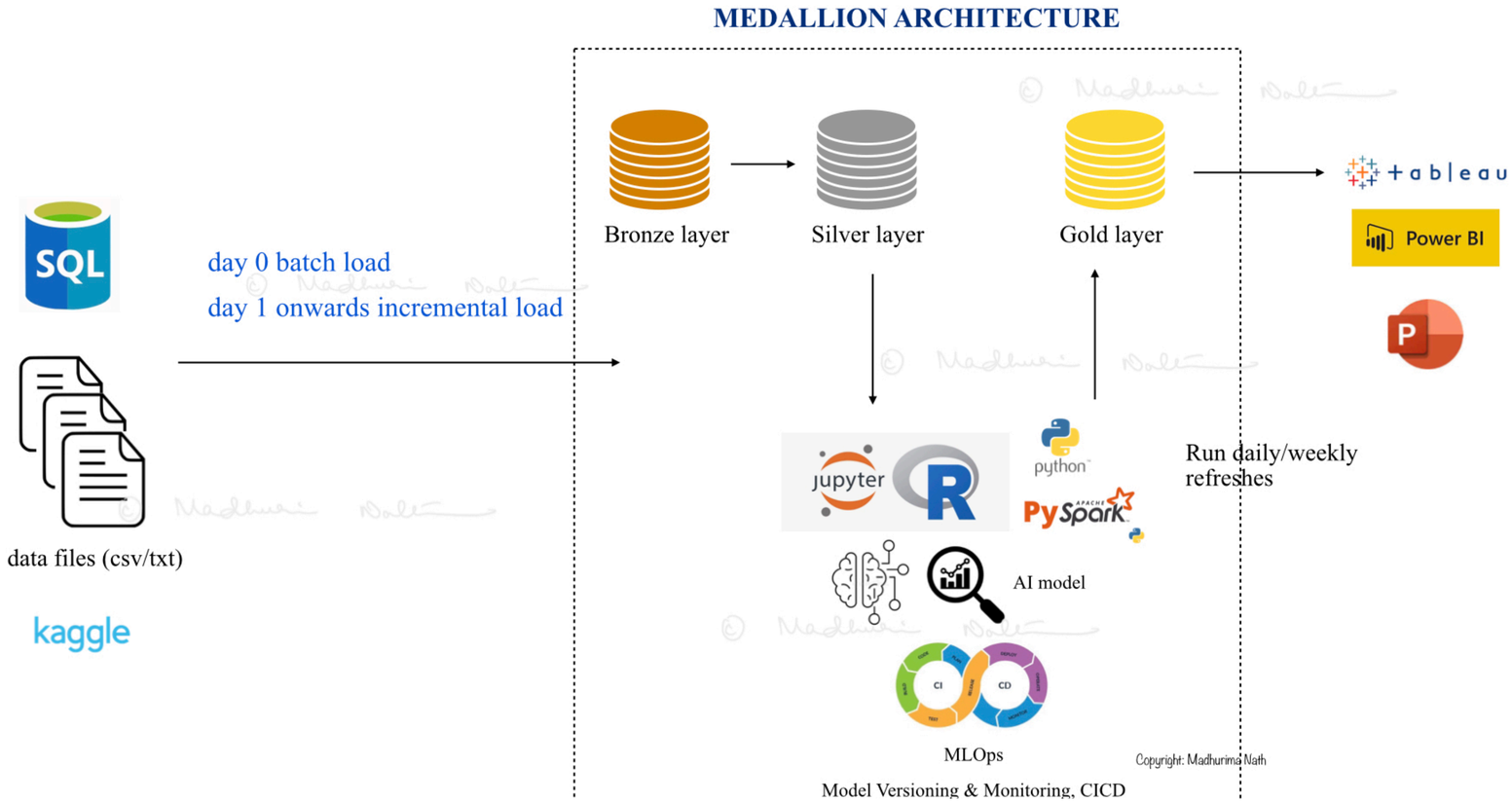
Large scale end-to-end solution

In a large scale solution, machine learning part is one of the pieces of the entire data architecture.



Large scale end-to-end solution

Here's an example of end-to-end architecture which incorporates medallion architecture and MLOps.



What are the changes?

- Every component needs to be integrated seamlessly
- Data cleaning/processing could take longer - use Spark to handle data
- Additional data might be required to enrich the data
- Parameterization is important
- CI/CD is important
- Incremental changes/updates required
- Re-training or incorporating feedback might be required
- Scheduled execution and updates with new data or latest changes

Q & A