Hybrid Recommendation System

Chien-Te Lee

2020/8/20

Outline

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

Outline

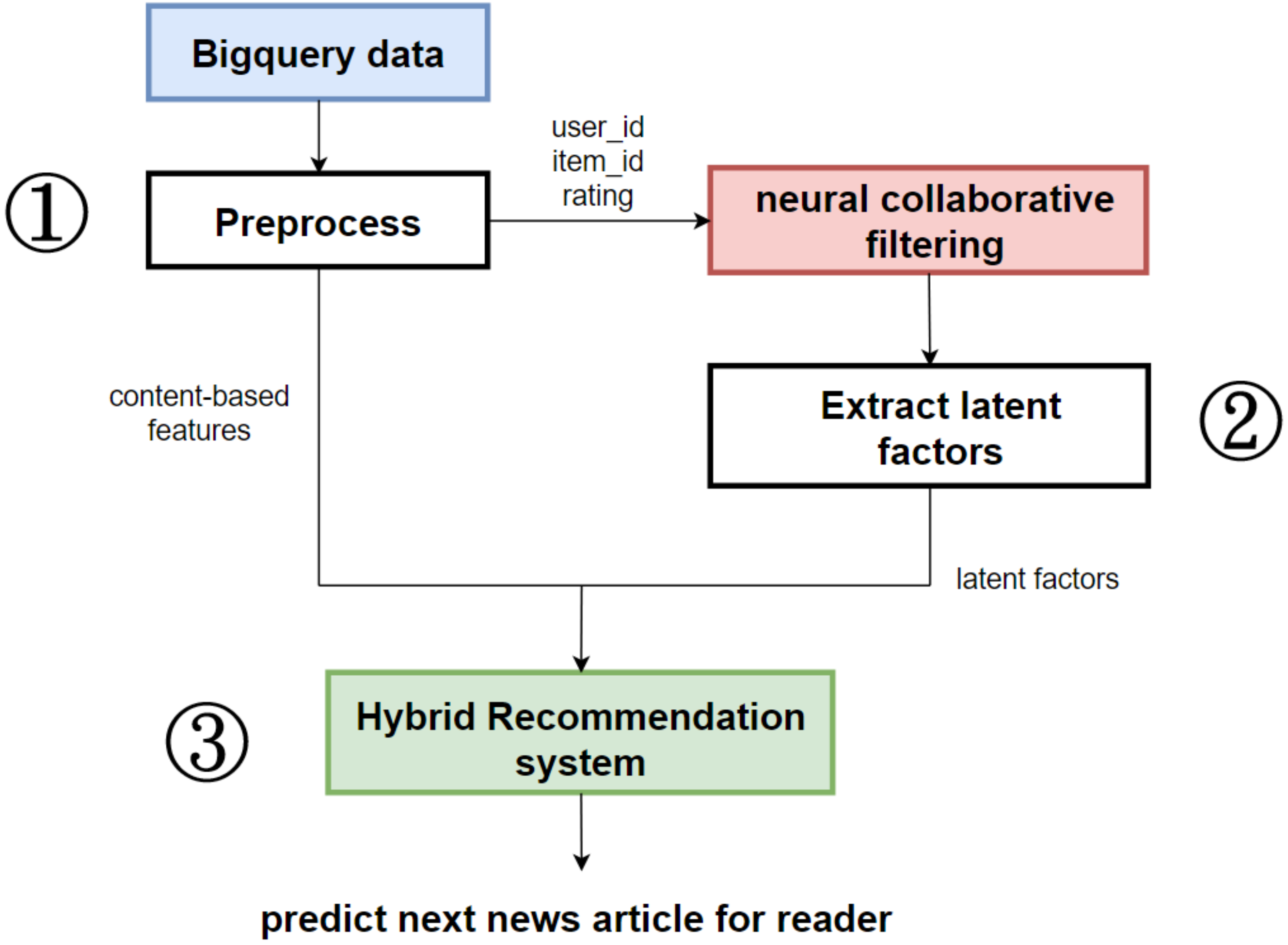
- Problem definition and workflow
- Preprocess BigQuery dataset
- Extract latent factors
- Hybrid recommendation system
- Reference

Problem definition

- Given a reader is reading an article on the news website, how to figure out what the reader would like to read for the next article?

Workflow

- Preprocess dataset → Extract latent factors → Train hybrid model

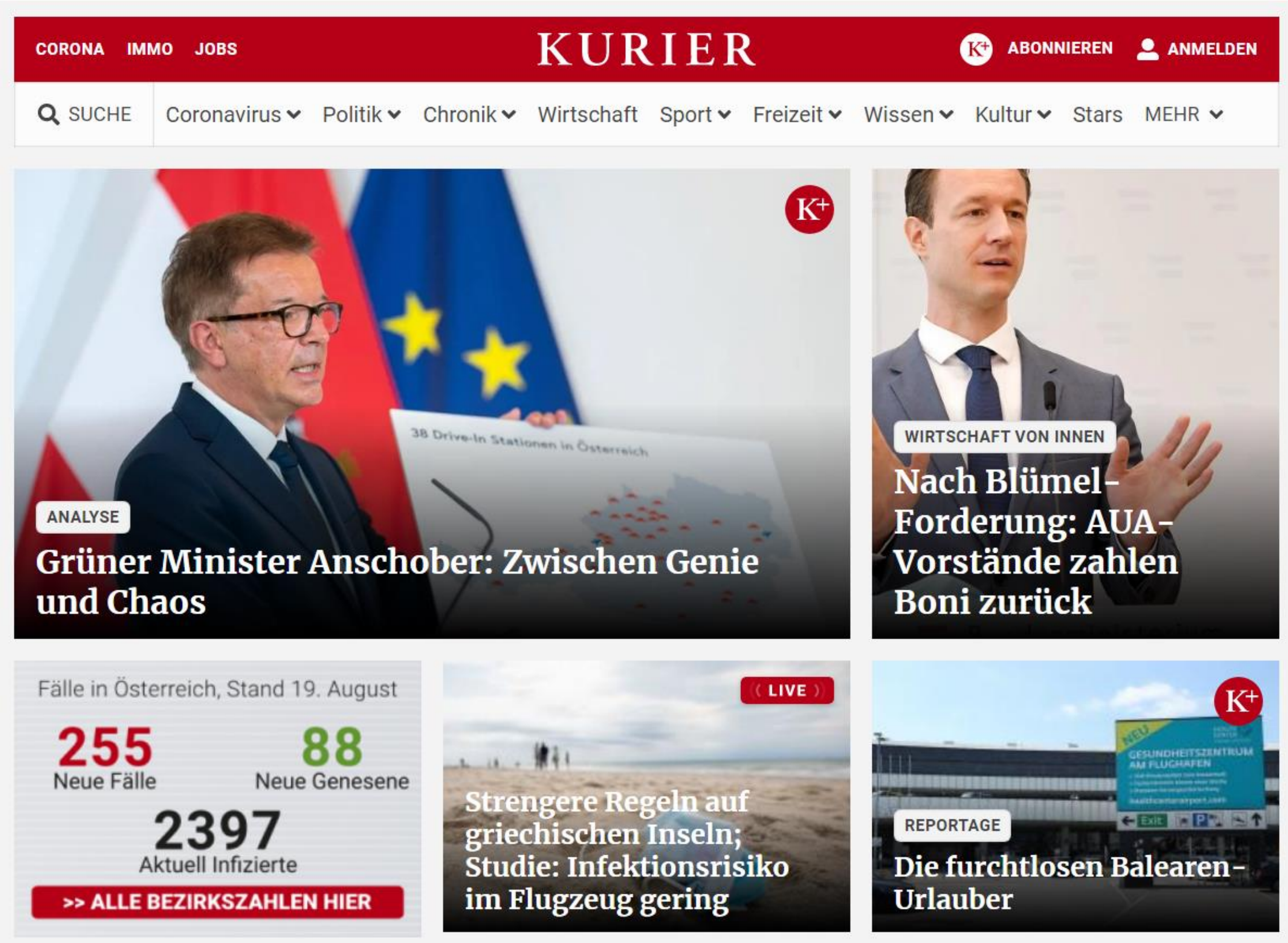


Outline

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

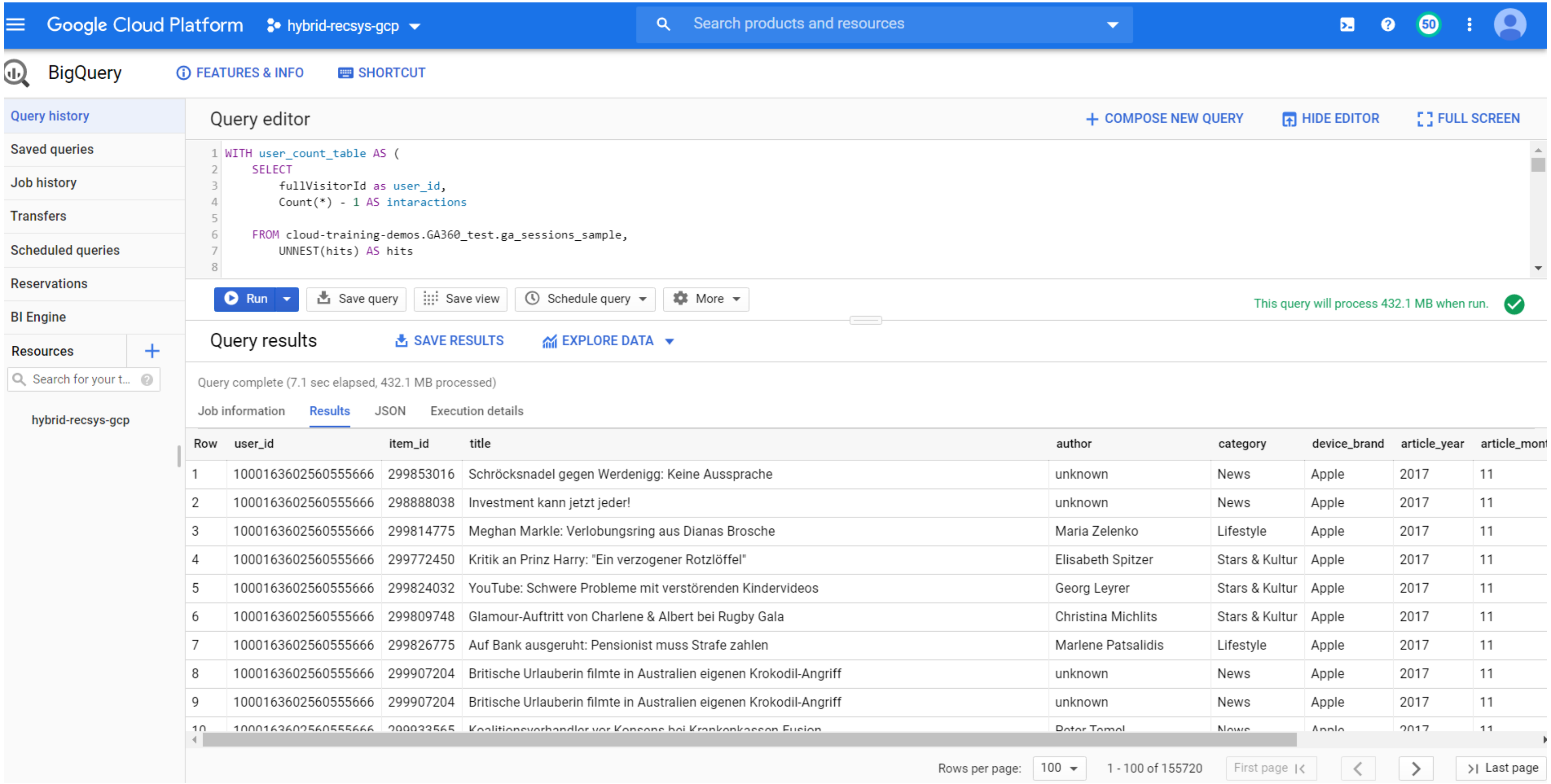
Preprocess BigQuery dataset

- “cloud-training-demos.GA360_test.ga_sessions_sample” is the Google Analytic data from Austrian news website Kurier.at.



Preprocess BigQuery dataset

- Use standard SQL to query the public BigQuey dataset, and select customDimensions as content-based features.



Preprocess BigQuery dataset

- Selected features:

	user_id	item_id	title	author	category	device_brand	article_year	article_month	rating	next_item_id	fold
0	1000196974485173657	299910994	Direktorensprecherin Isabella Zins: So könnte ...	Ute Brühl	News	unknown	2017	11	1.000000	299899819	0
1	1000196974485173657	299930679	Wintereinbruch naht: Erster Schnee im Osten mö...	Daniela Wahl	News	unknown	2017	11	1.000000	299972194	0
2	1004209053768679755	18976804	Heimskandal - Brigitte Wanker: Die Landesverrä...	Georg Hönigsberger	News	Huawei	2013	7	1.000000	299695400	0
3	1004555043399129313	299837992	Das erste TV-Interview von Prinz Harry & Megha...	Christina Michlits	Stars & Kultur	unknown	2017	11	0.979912	299824032	0
4	1004555043399129313	299836841	ÖVP will Studiengebühren FPÖ in Verhandlungen...	Raffaela Lindorfer	News	unknown	2017	11	1.000000	299899819	0

- Use 0.5*time/median_time as rating
- Use ABS(MOD(FarmFigerprint(visitor_id + visit_time), 10)) as hash_id

Outline

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

Extract Latent Factors

- Collaborative filtering use matrix factorization to split rating matrix into user matrix and item matrix.
- User and item latent factors are variables which represent similarities in high dimensional space.

Item

User	W	X	Y	Z
A		4.5	2.0	
B	4.0		3.5	
C		5.0		2.0
D		3.5	4.0	1.0

Rating Matrix

=

A	1.2	0.8
B	1.4	0.9
C	1.5	1.0
D	1.2	0.8

User Matrix

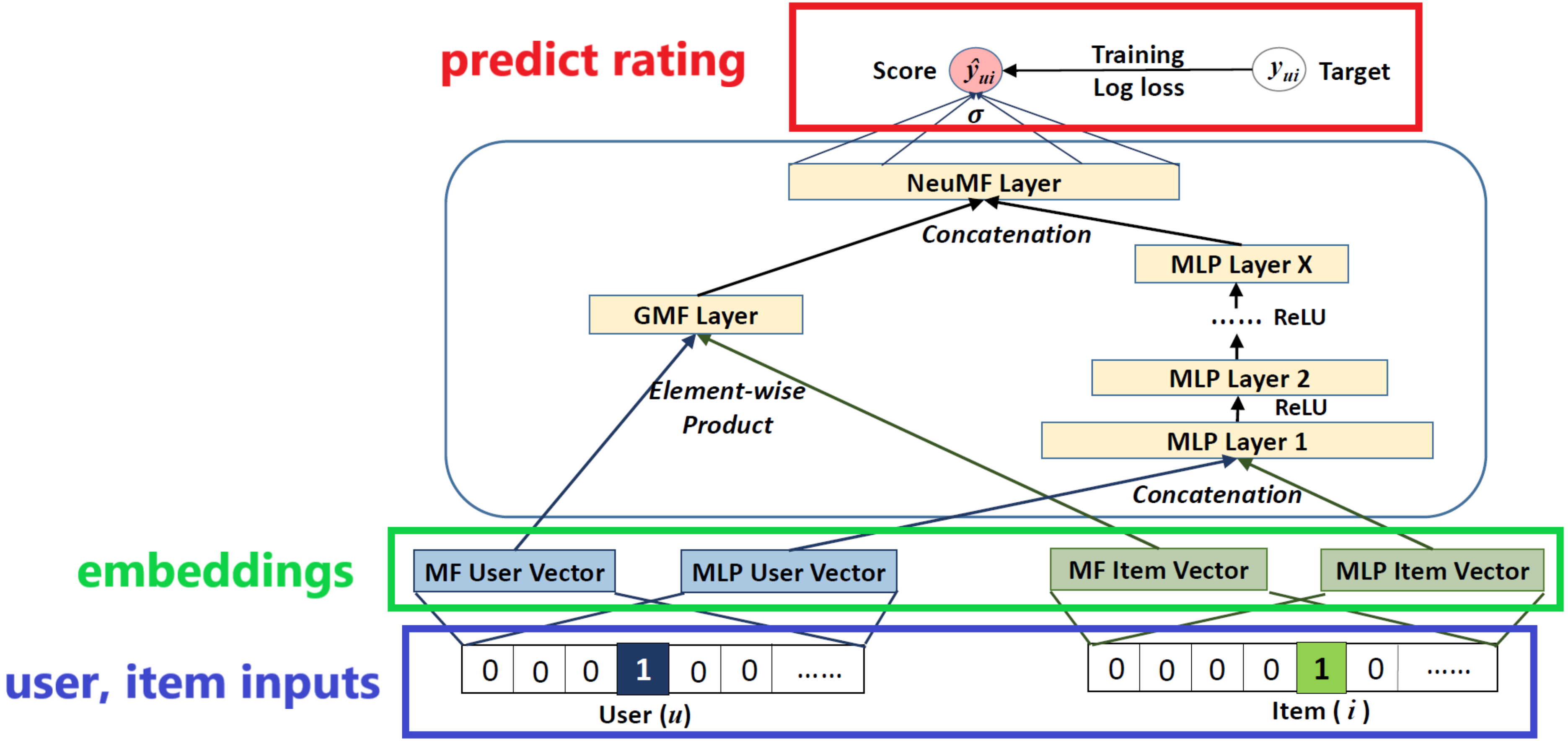
X

W	X	Y	Z
1.5	1.2	1.0	0.8
1.7	0.6	1.1	0.4

Item Matrix

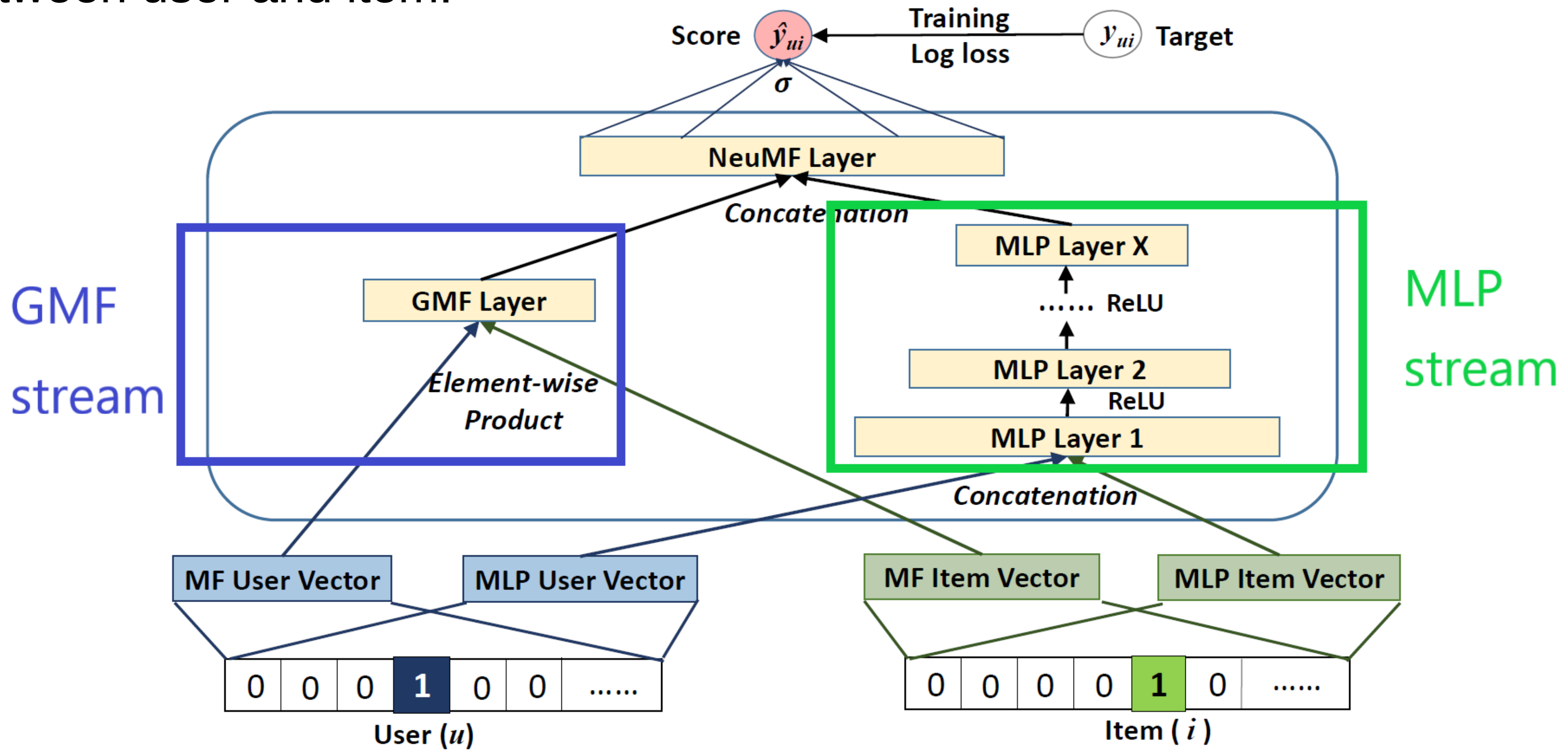
Extract Latent Factors

- Use neural collaborative filtering to predict rating
- Extract embedding as latent factors



Extract Latent Factors

- The Generalized Matrix Factorization (GMF) stream represents the matrix factorization.
- The Multi-Layer Perceptron (MLP) stream captures the non-linear relation between user and item.



Extract Latent Factors

- User latent factors

	user_id	u_latent_0	u_latent_1	u_latent_2	u_latent_3	u_latent_4	u_latent_5	u_latent_6	u_latent_7	u_latent_8	...	u_latent_10
0	1000163602560555666	-0.223890	0.013522	-0.197976	0.217497	-0.053681	-0.006720	-0.117004	0.113265	0.187496	...	-0.043796
1	1000196974485173657	-0.033306	0.020547	0.104502	-0.003414	0.063732	0.086023	-0.062370	0.030699	-0.115149	...	-0.034464
2	1002090131595000997	-0.179897	-0.139295	0.073862	-0.047588	0.047952	-0.000489	0.117391	0.058213	-0.077938	...	-0.012818
3	1002109532017576768	-0.079408	-0.174885	0.014121	-0.081578	0.140167	-0.137453	0.088288	0.162533	-0.106551	...	0.016511
4	1004209053768679755	-0.000192	-0.134218	0.076557	-0.169822	-0.072396	0.000815	-0.026878	-0.070867	0.092746	...	0.002242

- Item latent factors

	item_id	i_latent_0	i_latent_1	i_latent_2	i_latent_3	i_latent_4	i_latent_5	i_latent_6	i_latent_7	i_latent_8	...	i_latent_10	i_latent_11	i_latent_12
0	100170790	-0.044401	-0.054478	-0.024215	-0.095297	0.030977	-0.051534	-0.087727	0.066595	-0.116718	...	0.027045	-0.022293	-0.038569
1	100292889	0.044174	0.018957	-0.020329	0.005043	-0.066686	-0.046977	-0.011907	0.023122	-0.024344	...	0.048059	-0.057492	-0.052943
2	100735153	0.004435	-0.092585	-0.101787	-0.067878	0.077632	0.000198	-0.068222	0.012467	-0.053971	...	-0.001989	0.011519	0.056933
3	100915139	0.009406	0.015461	-0.031027	-0.006515	0.015776	-0.004458	0.006125	-0.020394	-0.046054	...	0.044172	0.006846	-0.025532
4	101092112	-0.063698	0.059048	0.049322	-0.023419	0.039215	0.036990	0.013302	-0.031852	-0.001982	...	0.043176	-0.017732	-0.049217

Extract Latent Factors

- Concatenate latent factors with preprocessed content-based features
- Each specific user_id matches with specific user_latent (INNER JOIN)
- Each specific item_id matches with specific item_latent (INNER JOIN)

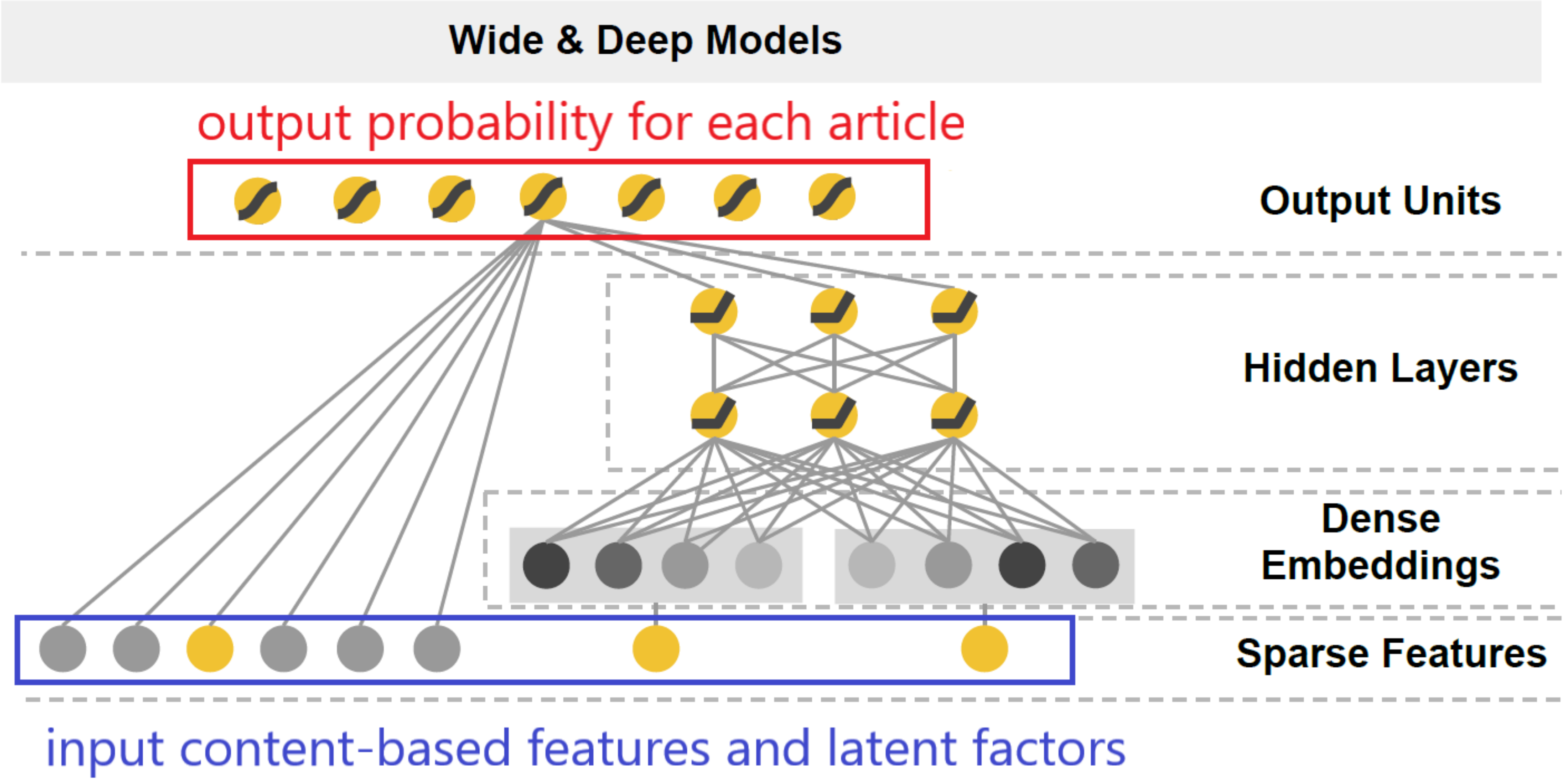
	user_id	item_id	content-based features	user_latent	item_latent
0					
1					
2					

Outline

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

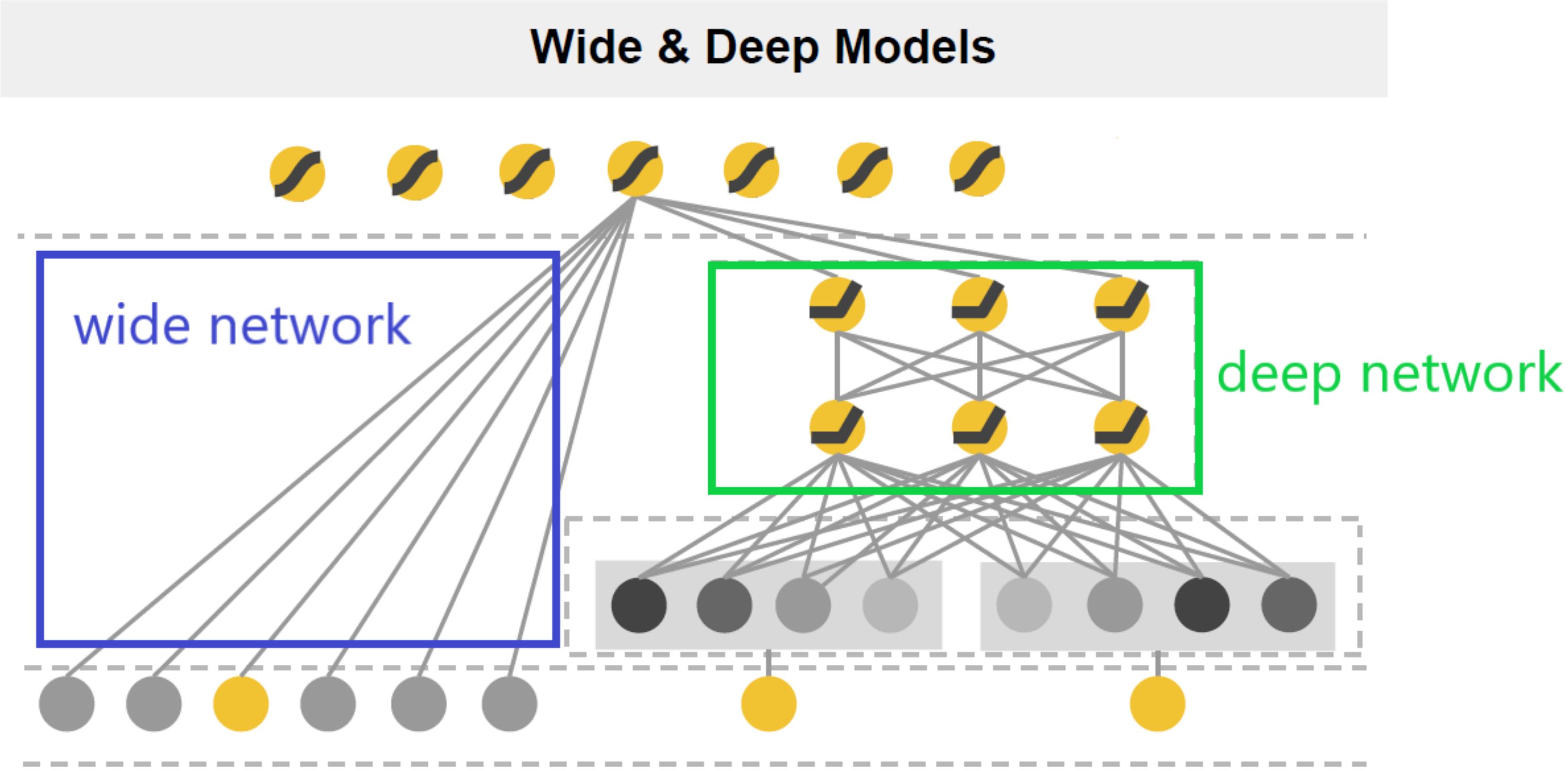
Hybrid Recommendation System

- Apply wide & deep network for hybrid model.
- Use content-based features and user, item latent factors as input.
- Predict the probability for each article as next item.



Hybrid Recommendation System

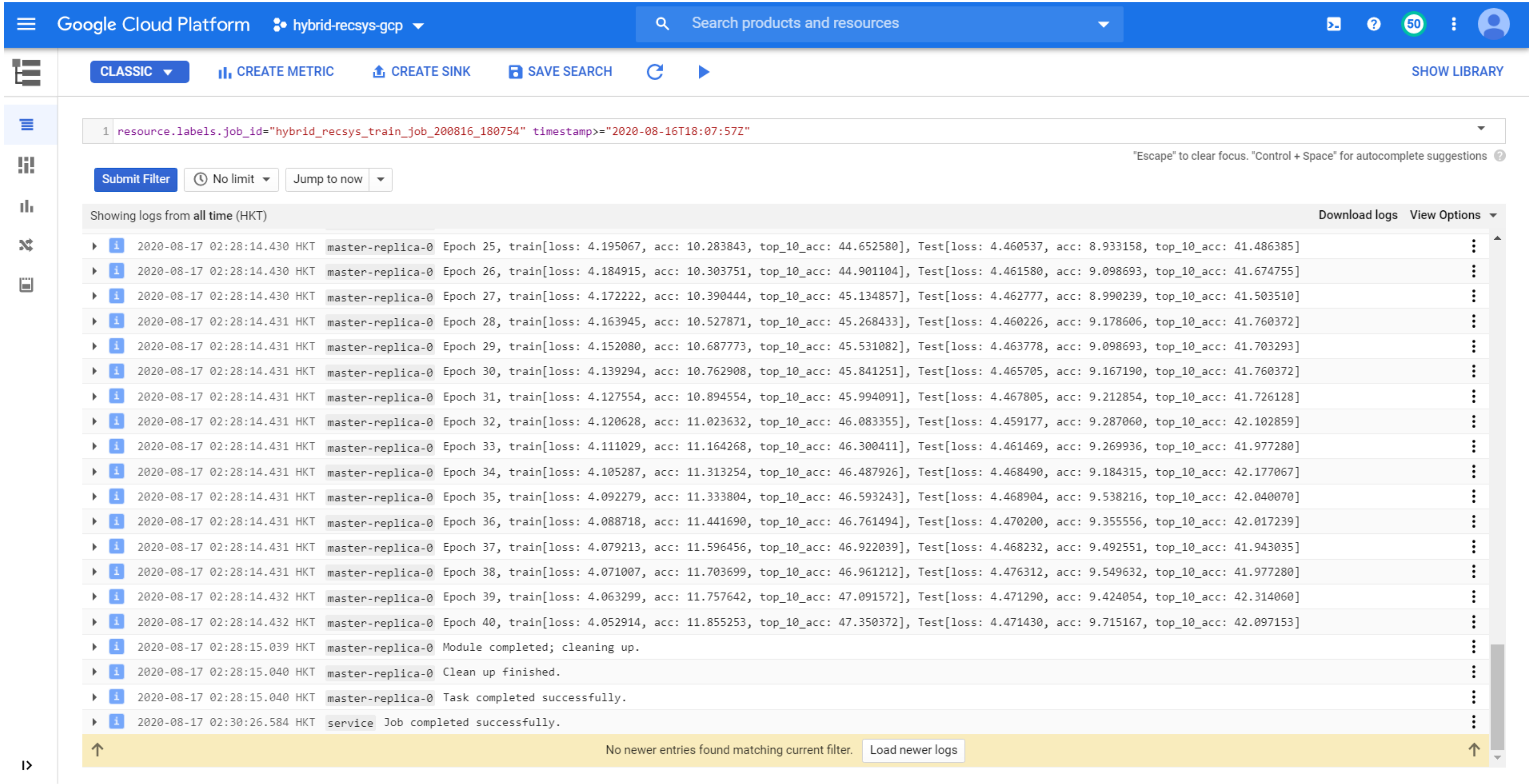
- The deep network takes dense embedding features, and the wide network takes sparse features.
- Dense embedding feature: user_id, item_id, author, device_brand, title (NNLM)
- Sparse feature: author, cross_date, category, device_brand



Hybrid Recommendation System

- Result: Train [bc_loss: 4.05, acc: 11.85, top_10_acc: 47.35]

  Test [bc_loss: 4.47, acc: 9.71, top_10_acc: 42.09]



Hybrid Recommendation System

- The model has 42.09% chance to correctly predict the next news article the reader would like to view if our model provide 10 recommended items.
- If randomly picking 10 items from total 2421 news articles, the top 10 accuracy would be only 0.413%. Our hybrid model has around 100 times better top 10 accuracy than random picking.

Outline

- **Problem definition and workflow**
- **Preprocess BigQuery dataset**
- **Extract latent factors**
- **Hybrid recommendation system**
- **Reference**

Reference

- Neural Collaborative Filtering
- Wide & Deep Learning for Recommender Systems
- Recommendation Systems with TensorFlow on GCP
- End-to-end Machine Learning with TensorFlow on GCP
- Collaborative Filtering using Deep Neural Networks (in Tensorflow)
- Get started with TensorBoard
- Deploying models
- Method: projects.predict
- Simple Matrix Factorization example on the Movielens dataset using Pyspark

Thank you for your attention!!