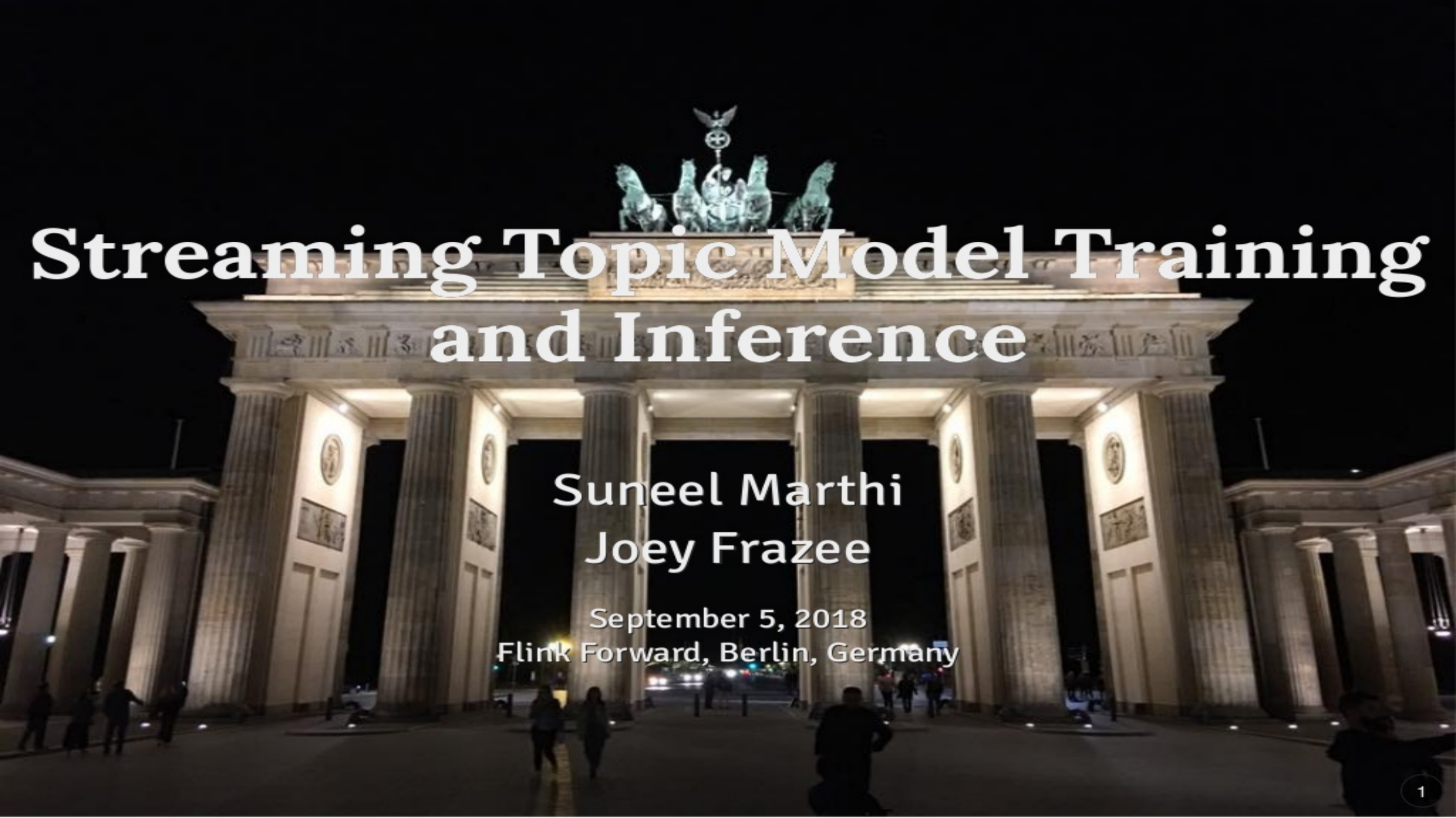


Streaming Topic Model Training and Inference

Suneel Marthi
Joey Frazee

September 5, 2018
Flink Forward, Berlin, Germany

$WhoAreWe

Joey Frazee

@jfrazee

- Member of Apache Software Foundation
- Committer on Apache NiFi, and PMC on Apache Streams

Suneel Marthi

@suneelmarthi

- Member of Apache Software Foundation
- Committer and PMC on Apache Mahout, Apache OpenNLP, Apache Streams

Agenda

- Motivation for Topic Modeling
- Existing Approaches for Topic Modeling
- Topic Modeling on Streams

Motivation for Topic Modeling

Topic Models

- Automatically discovering main topics in collections of documents
  - Overall
    - What are the main themes discussed on a mailing list or a discussion forum ?
    - What are the most recurring topics discussed in research papers ?

Topic Models (Contd)

- Per Document
  - What are the main topics discussed in a certain (single) newspaper article ?
  - What is the most distinctive topic that makes a certain research article more interesting than the others ?

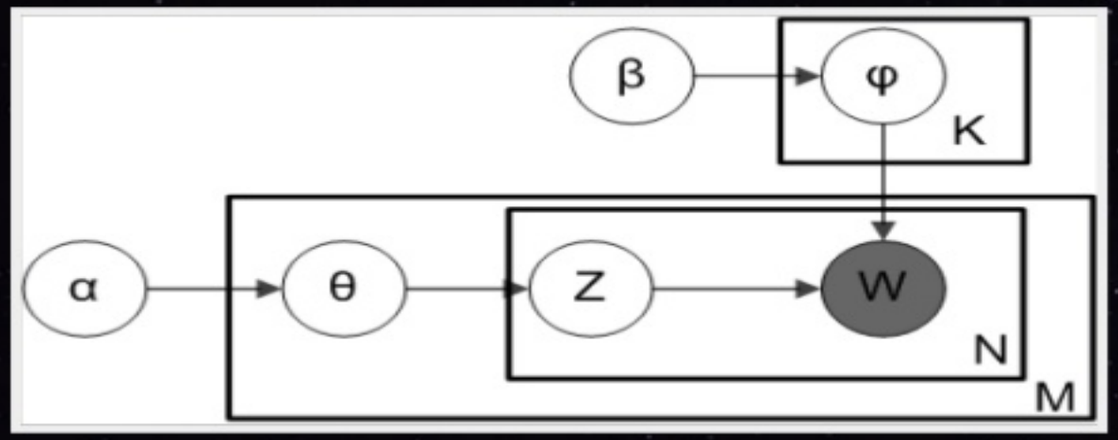
Uses of Topic Models

- Search Engines to organize text
- Annotating Documents according to these topics
- Uncovering hidden topical patterns across a document collection

Common Approaches to Topic Modeling

- Latent Dirichlet Allocation (LDA)
  - Topics are composed by probability distributions over words
  - Documents are composed by probability distributions over Topics
  - Batch Oriented approach

LDA Model



- *N*: vocab size
- *M/D*: # of documents
- *alpha*: prior on per-document topic dist.
- *beta*: prior on topic-word dist.
- *theta*: per-document topic dist.
- *phi*: per-topic word dist.
- *z*: topic for word w

Intuition: Documents are a mixture of topics, topics are a mixture of words. So documents can be generated by drawing a sample of topics, and then drawing a sample of words.

- Latent Semantic Analysis (LSA)
  - SVDed TF-IDF Document-Term Matrix
  - Batch Oriented approach

- Drawbacks of Traditional Topic Modeling Approach
  - Problem : labelling topics is often a manual task

```
E.g. LDA topic:30% basket,15% ball,10% drunk,…
                    can be tagged as basketball
```

~~Common Approaches~~ Deep Learning (of course)

- Learn to perform LDA: LDA supervised DNN training (inference speed up)
  https://arxiv.org/abs/1508.01011
- Deep Belief Networks for Topic Modelling
  - https://arxiv.org/abs/1501.04325

What are Embeddings ?

Represent a document as a point in space, semantically similar docs are close together when plotted

Such a representation is learned via a (shallow) neural network algorithm

Embeddings for Topic Modeling

- LDA2Vec: Mixing LDA and word2vec word embeddings https://arxiv.org/abs/1605.02019
- Navigating embeddings: geometric navigation in word and doc embeddings space looking for topics

Static vs Dynamic Corpora

- Learning Topic Models over a fixed set of Documents
  - Fetch the Corpus
  - Train and Fit a topic model
  - Extract topics
  - ...other downstream tasks

Static vs Dynamic Corpora (Contd)

- What to do with Corpus updates?
  - Document collections are not static in real world
  - There may be no document collections to begin with
  - Memory limits with large corpora
  - Reprocess entire corpora when new documents are added

Static vs Dynamic Corpora (Contd)

- Streams, Streams, Streams
  - Twitter stream
  - Reddit posts
  - Papers on ResearchGate and Arxiv
  - ....

Questions for Audience

- Who's doing inference or scoring on streaming data?
- Who's doing online learning and optimization on streaming data?

(2) is hard, sometimes because the algorithms aren't there. But, also because people don't know the algorithms are there.

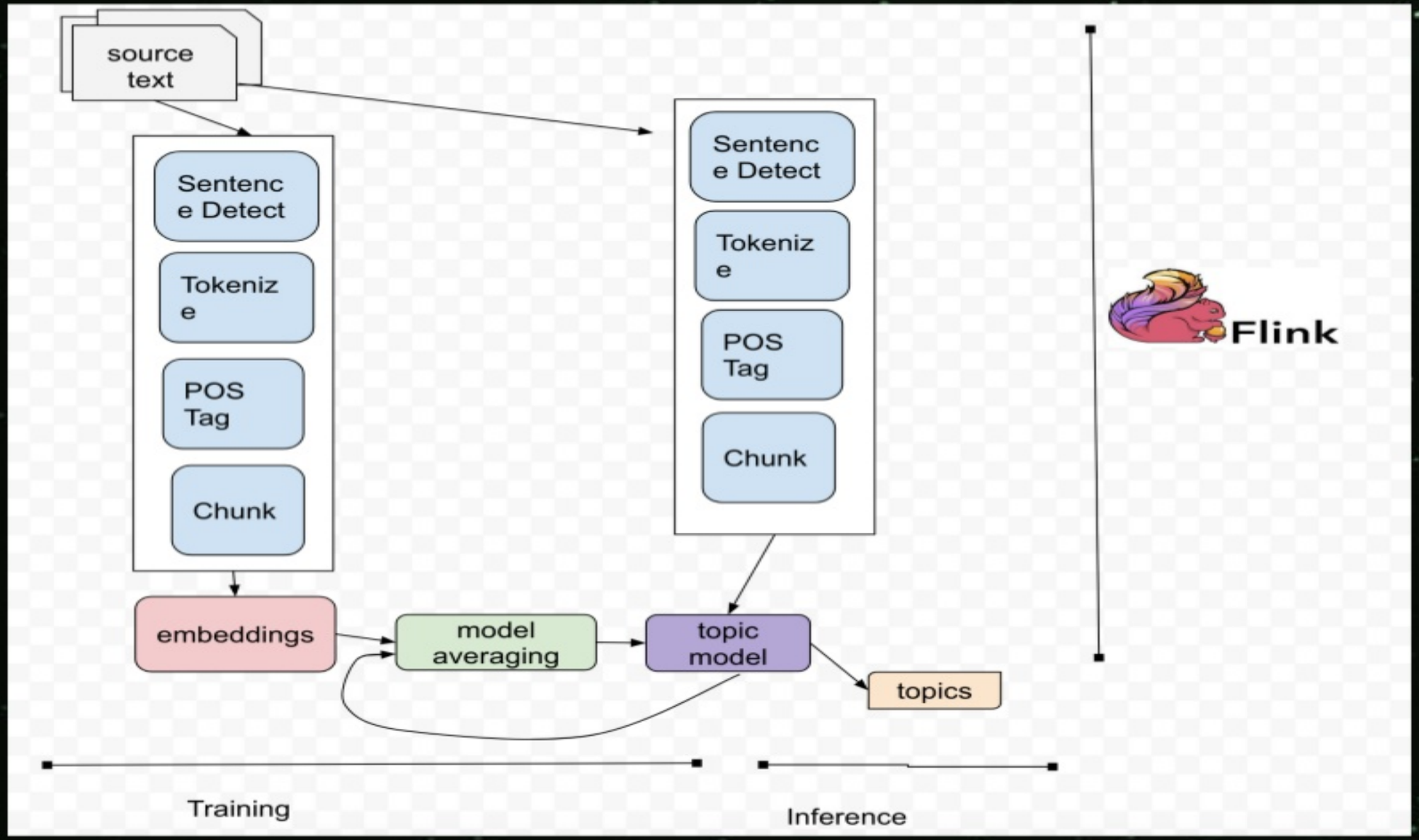
Topic Modeling on Streams

2 Streaming Approaches

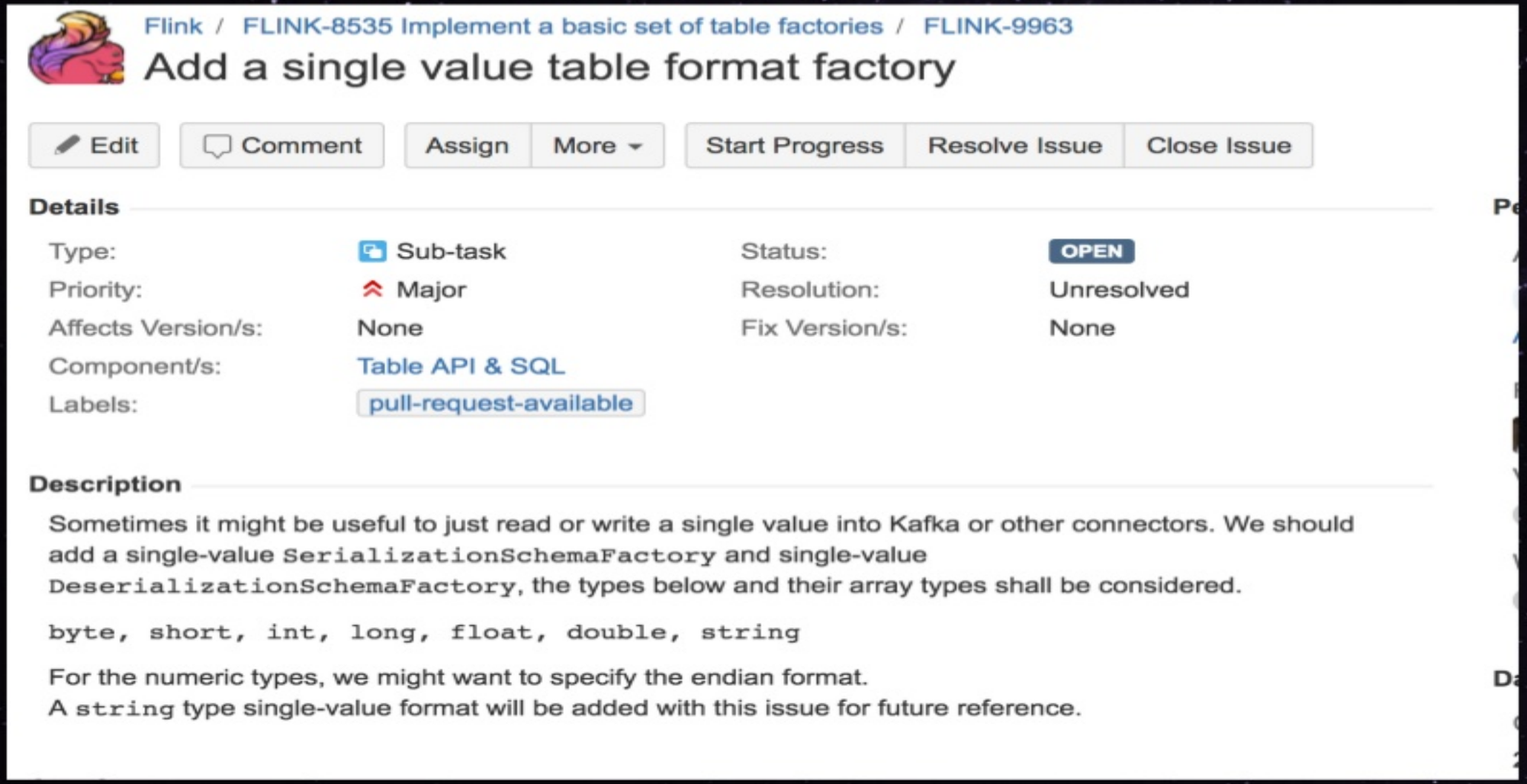
- Learning Topics from Jira Issues
- Online LDA

Learning Topics on Jira Issues

Workflow


source text
Sentenc e Detect
Tokeniz e
POS Tag
Chunk
Sentenc e Detect
Tokeniz e
POS Tag
Chunk
Flink
embeddings
model averaging
topic model
topics
Training
Inference

FLINK-9963: Add a single value table format factory



Flink / FLINK-8535 Implement a basic set of table factories / FLINK-9963

Add a single value table format factory

Edit | Comment | Assign | More | Start Progress | Resolve Issue | Close Issue

Details

Type:	Sub-task	Status:	OPEN
Priority:	Major	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Component/s:	Table API & SQL		
Labels:	pull-request-available		

Description

Sometimes it might be useful to just read or write a single value into Kafka or other connectors. We should add a single-value `SerializationSchemaFactory` and single-value `DeserializationSchemaFactory`, the types below and their array types shall be considered.

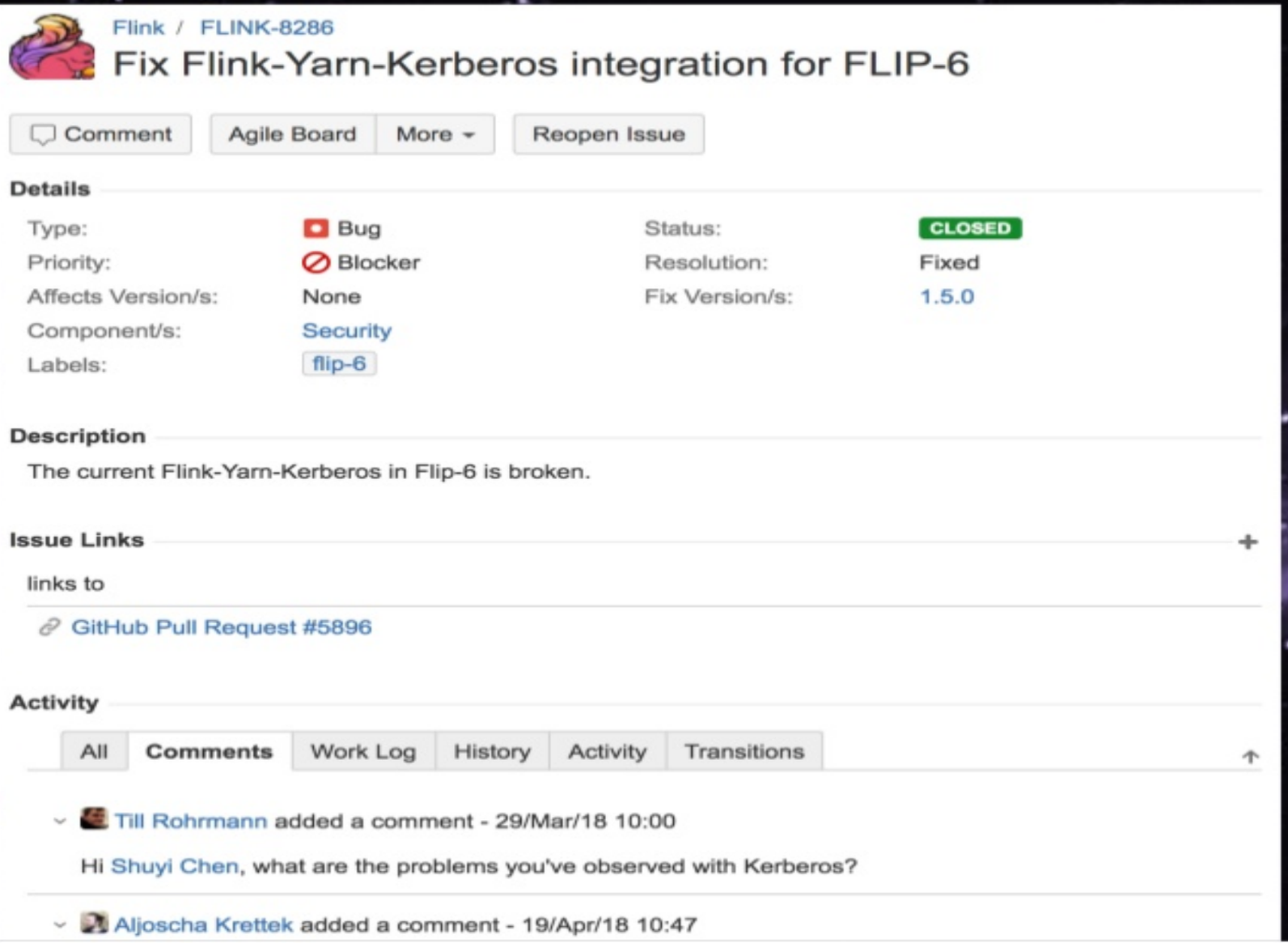
```
byte, short, int, long, float, double, string
```

For the numeric types, we might want to specify the endian format.
A `string` type single-value format will be added with this issue for future reference.

Topics extracted for Flink-9963

- format factory
- table schema
- table source
- sink factory
- table sink

FLINK-8286: Fix Flink-Yarn-Kerberos integration for FLIP-6



Flink / FLINK-8286

Fix Flink-Yarn-Kerberos integration for FLIP-6

Comment | Agile Board | More | Reopen Issue

Details

Type:	Bug	Status:	CLOSED
Priority:	Blocker	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	1.5.0
Component/s:	Security		
Labels:	flip-6		

Description

The current Flink-Yarn-Kerberos in Flip-6 is broken.

Issue Links

links to

GitHub Pull Request #5896

Activity

All | Comments | Work Log | History | Activity | Transitions

Till Rohrmann added a comment - 29/Mar/18 10:00

Hi Shuyi Chen, what are the problems you've observed with Kerberos?

Aljoscha Krettek added a comment - 19/Apr/18 10:47

Shuyi Chen Is there any update on this?

Shuyi Chen added a comment - 20/Apr/18 05:55

Hi Till Rohrmann and Aljoscha Krettek, the context is that there is a regression in flink kerberos yarn integration in 1.4, which is addressed in FLINK-8275. This task is created at that time to make sure that there is no regression on flip6 as well. I'll take a look the next few days

Topics extracted for Flink-8286

- yarn kerberos integration
- kerberos integrations
- yarn kerberos

Bayesian Inference

Goal: Calculate the posterior distribution of a probabilistic model given data

- Gibbs sampling/MCMC: Repeatedly sample from conditional distribution(s) to approximate the posterior of the joint distribution; typically a batch process
- Variational Bayes: Optimize the parameters of a function that approximates joint distribution

Variational Bayes

- Expectation Maximization-like optimization
- Faster and more accurate
- Can often be computed online (on windows over streams)

Online LDA (1/2)

- Variational Bayes for LDA
- Batch or online
- Iterate between optimizing per-word topics and per-document topics (“E” step) and topic-word distributions (“M” step)
- Default topic model optimization in Gensim, scikit-learn, Apache Spark MLlib -- not because it’s amenable to streaming per se, but more because it’s memory efficient and can be applied to large corpora.

Hoffman, Blei & Bach. Online Learning for Latent Dirichlet Allocation. Proceedings of Neural Information Processing Systems. 2010.

Online LDA (2/2)

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $t = 0$ to ∞ **do**
  E step:
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{twk} \propto \exp\{\mathbb{E}_q[\log \theta_{tk}] + \mathbb{E}_q[\log \beta_{kw}]\}$
    Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$
  until $\frac{1}{K} \sum_k |\text{change in} \gamma_{tk}| < 0.00001$
  M step:
  Compute $\tilde{\lambda}_{kw} = \eta + D n_{tw} \phi_{twk}$
  Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}}$.
end for

job-constant
- *K*: number of topics
- *alpha*: topic-document dist. prior.
- *eta*: topic-word dist, prior
- *tau0*: learning rate
- *kappa*: decay rate

(mini)batch-local
- *gamma*: theta prior
- *theta*: param. for per-document topic dist
- *phi*: param. for per-topic word dist.
- *rho*: update weight

job-global
- *lambda*: parameter for topic-word dist.
- *t*: documents/batches seen (index)

Goal: Optimize the topic-word dist. parameter lambda; it tells us what words belong to the topics.

[Image source: Hoffman, Blei & Bach, 2010.]

Job-global parameters can be tracked in a state store. Batch-local parameters can be discarded.

Online LDA on Flink

job-constant
- *K*: number of topics
- *alpha*: topic-document dist. prior
- *eta*: topic-word dist. prior
- *tau0*: learning rate
- *kappa*: decay rate

Set job parameters globally with setGlobalJobParameters() on execution environment, or pass as arguments to operators

---

(mini)batch-local
- *gamma*: theta prior
- *theta*: param. for per-document topic dist
- *phi*: param. for per-topic word dist.
- *rho*: update weight

Operator/task-internal and don't need to be accessible by other tasks

---

job-global
- *lambda*: parameter for topic-word dist.
- *t*: documents/batches seen (index)

Several options:
- External key-value store
- Accumulators: *lambda* update as increment, accumulator + broadcast variables
- Broadcast state

And, (mini)batches selected by Window size

So what?

- For many tasks, the default is batch training/optimization
- Often, even if good online training and optimization algorithms exist (why?)
- So if your algorithms only require 1-step of parameter state, why not just do streaming?
- You can get instant, always up-to-date results by putting the effort into using a different class of optimization algorithms an using a stream processing engine, so just do that

Links

- http://papers.nips.cc/paper/3902-online-learning-for-latentdirichlet-allocation
- Learning from LDA using Deep Neural Networks - https://arxiv.org/pdf/1508.01011.pdf
- Deep Belief Nets for Topic Modeling - https://arxiv.org/pdf/1501.04325.pdf
- Topic Modeling on Jira issues: https://github.com/tteofili/jtm

Credits

- Tommaso Teofili, Simone Tripodi (Adobe - Rome)
- Joern Kottmann, Bruno Kinoshita (Apache OpenNLP)
- Fabian Hueske (Apache Flink, Data Artisans)

Questions ???