Tracking recent events through recent Wikipedia changes using Storm

by Gustaf Helgesson

Aim

- Correlate # of article changes *within* a language to recent events.
  - For English, German, Spanish and Japanese.

- Correlate article changes *between* languages to recent events.
  - By using Wikipedia's "in another language: English" feature.

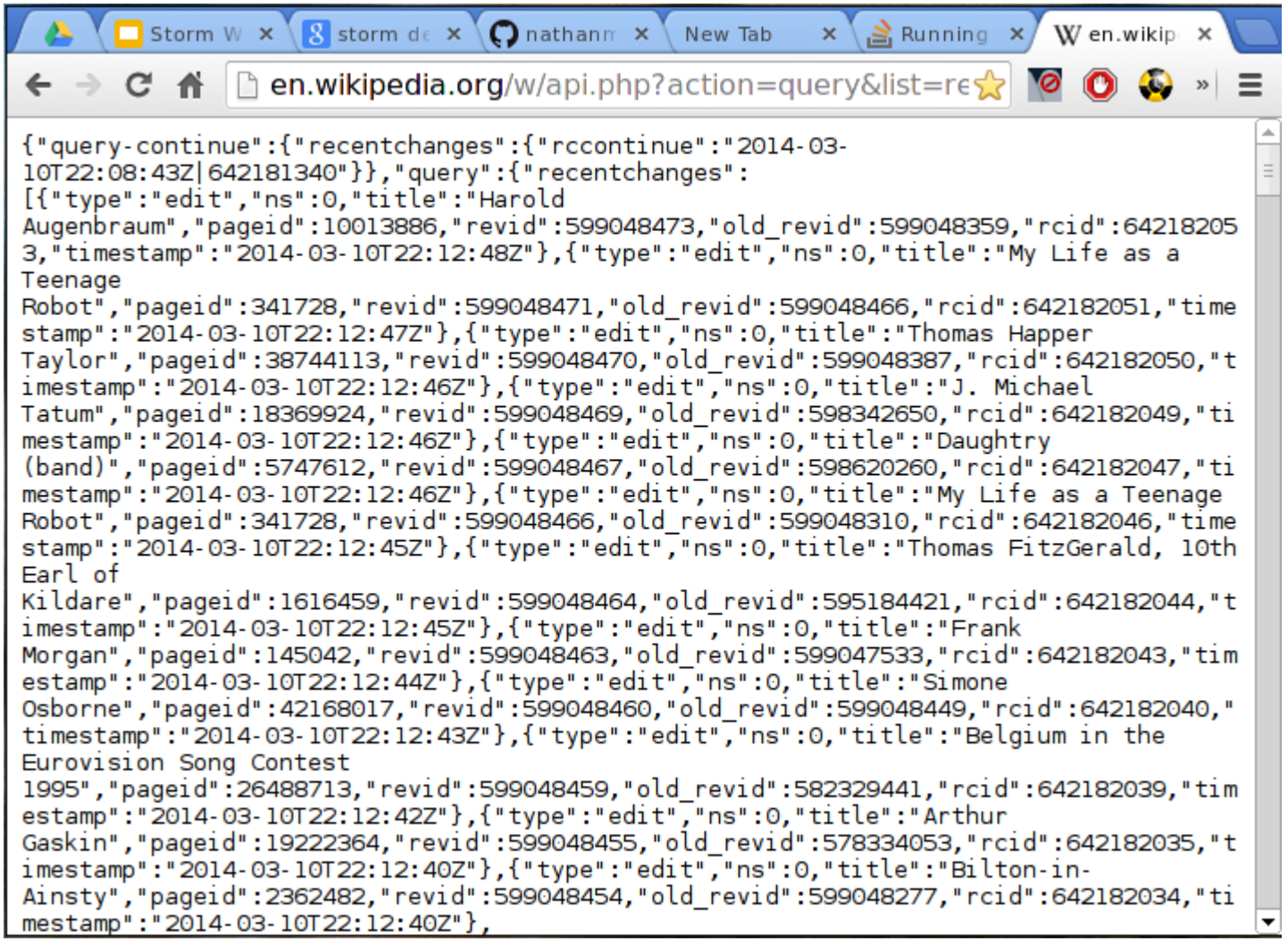
Data collection

- #Recent changes per article per language
  - For: English, Spanish, German and Japanese

- Use streaming windows of 2-6 hours and see how event changes for the top 100 events

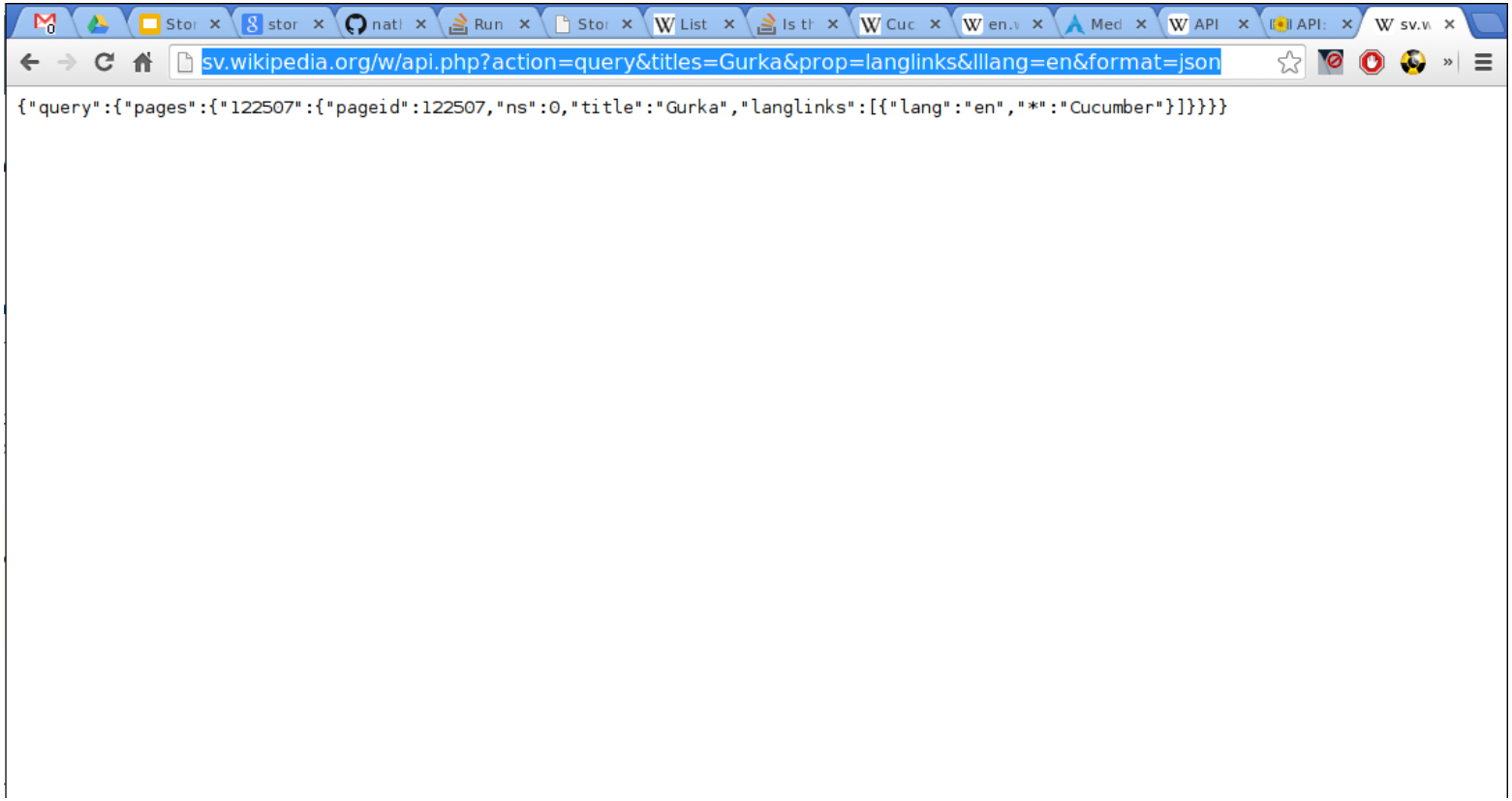
- Depending on necessity I may make use of approximate counting in the counting phases.

Input stream - JSON data!




```
{"query-continue":{"recentchanges":{"rccontinue":"2014-03-
10T22:08:43Z|642181340"}},"query":{"recentchanges":
[{"type":"edit","ns":0,"title":"Harold
Augenbraum","pageid":10013886,"revid":599048473,"old_revid":599048359,"rcid":64218205
3,"timestamp":"2014-03-10T22:12:48Z"},{"type":"edit","ns":0,"title":"My Life as a
Teenage
Robot","pageid":341728,"revid":599048471,"old_revid":599048466,"rcid":642182051,"time
stamp":"2014-03-10T22:12:47Z"},{"type":"edit","ns":0,"title":"Thomas Happer
Taylor","pageid":38744113,"revid":599048470,"old_revid":599048387,"rcid":642182050,"t
imestamp":"2014-03-10T22:12:46Z"},{"type":"edit","ns":0,"title":"J. Michael
Tatum","pageid":18369924,"revid":599048469,"old_revid":598342650,"rcid":642182049,"ti
mestamp":"2014-03-10T22:12:46Z"},{"type":"edit","ns":0,"title":"Daughtry
(band)","pageid":5747612,"revid":599048467,"old_revid":598620260,"rcid":642182047,"ti
mestamp":"2014-03-10T22:12:46Z"},{"type":"edit","ns":0,"title":"My Life as a Teenage
Robot","pageid":341728,"revid":599048466,"old_revid":599048310,"rcid":642182046,"time
stamp":"2014-03-10T22:12:45Z"},{"type":"edit","ns":0,"title":"Thomas FitzGerald, 10th
Earl of
Kildare","pageid":1616459,"revid":599048464,"old_revid":595184421,"rcid":642182044,"t
imestamp":"2014-03-10T22:12:45Z"},{"type":"edit","ns":0,"title":"Frank
Morgan","pageid":145042,"revid":599048463,"old_revid":599047533,"rcid":642182043,"tim
estamp":"2014-03-10T22:12:44Z"},{"type":"edit","ns":0,"title":"Simone
Osborne","pageid":42168017,"revid":599048460,"old_revid":599048449,"rcid":642182040,"
timestamp":"2014-03-10T22:12:43Z"},{"type":"edit","ns":0,"title":"Belgium in the
Eurovision Song Contest
1995","pageid":26488713,"revid":599048459,"old_revid":582329441,"rcid":642182039,"tim
estamp":"2014-03-10T22:12:42Z"},{"type":"edit","ns":0,"title":"Arthur
Gaskin","pageid":19222364,"revid":599048455,"old_revid":578334053,"rcid":642182035,"t
imestamp":"2014-03-10T22:12:40Z"},{"type":"edit","ns":0,"title":"Bilton-in-
Ainsty","pageid":2362482,"revid":599048454,"old_revid":599048277,"rcid":642182034,"ti
mestamp":"2014-03-10T22:12:40Z"},
```

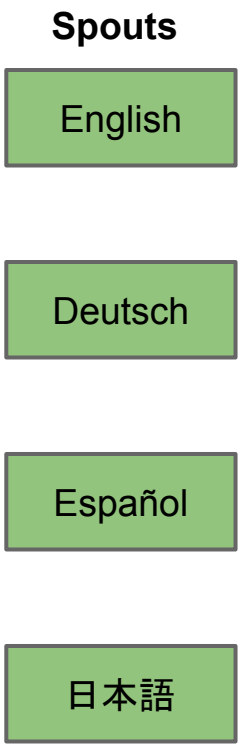
Article conversion to English Wikipedia



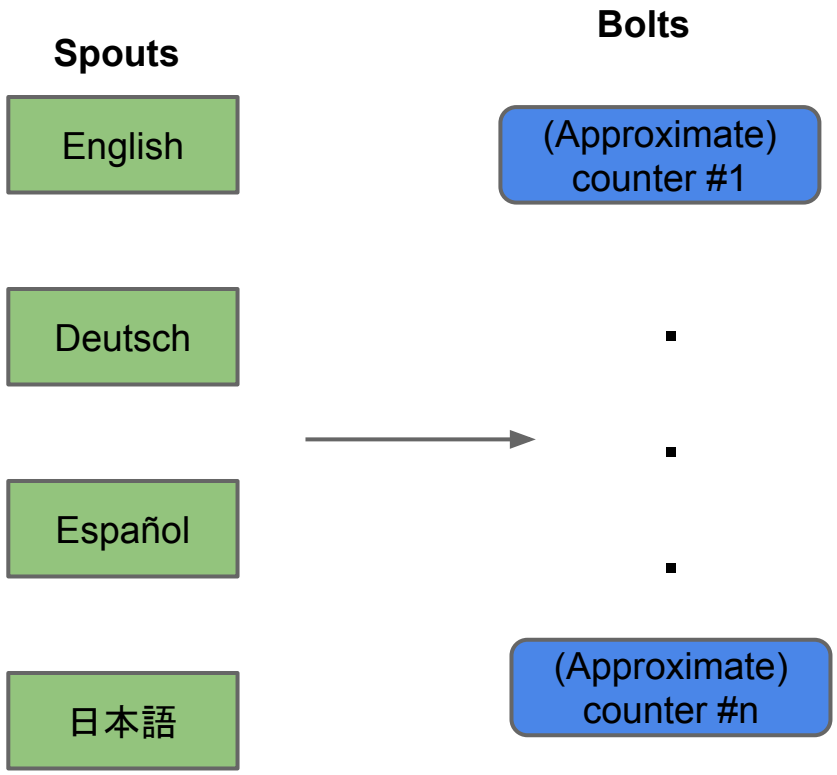
Storm Intro/Recap

- Stream Processing Engine
- Programmers create explicit DAGs (topologies) of custom or built in functions
- External inputs (spouts), external outputs (sinks), processing elements (bolts)

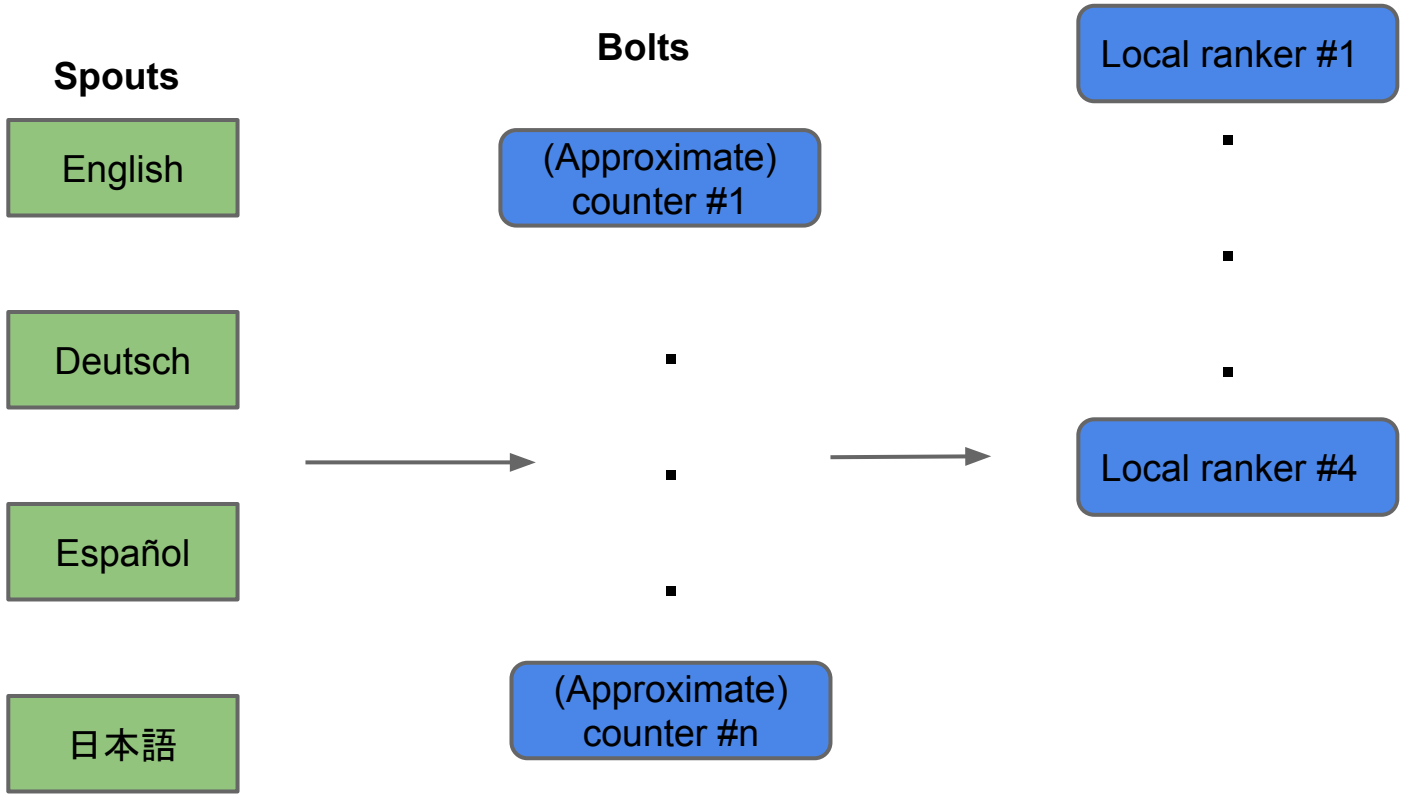
Storm topology



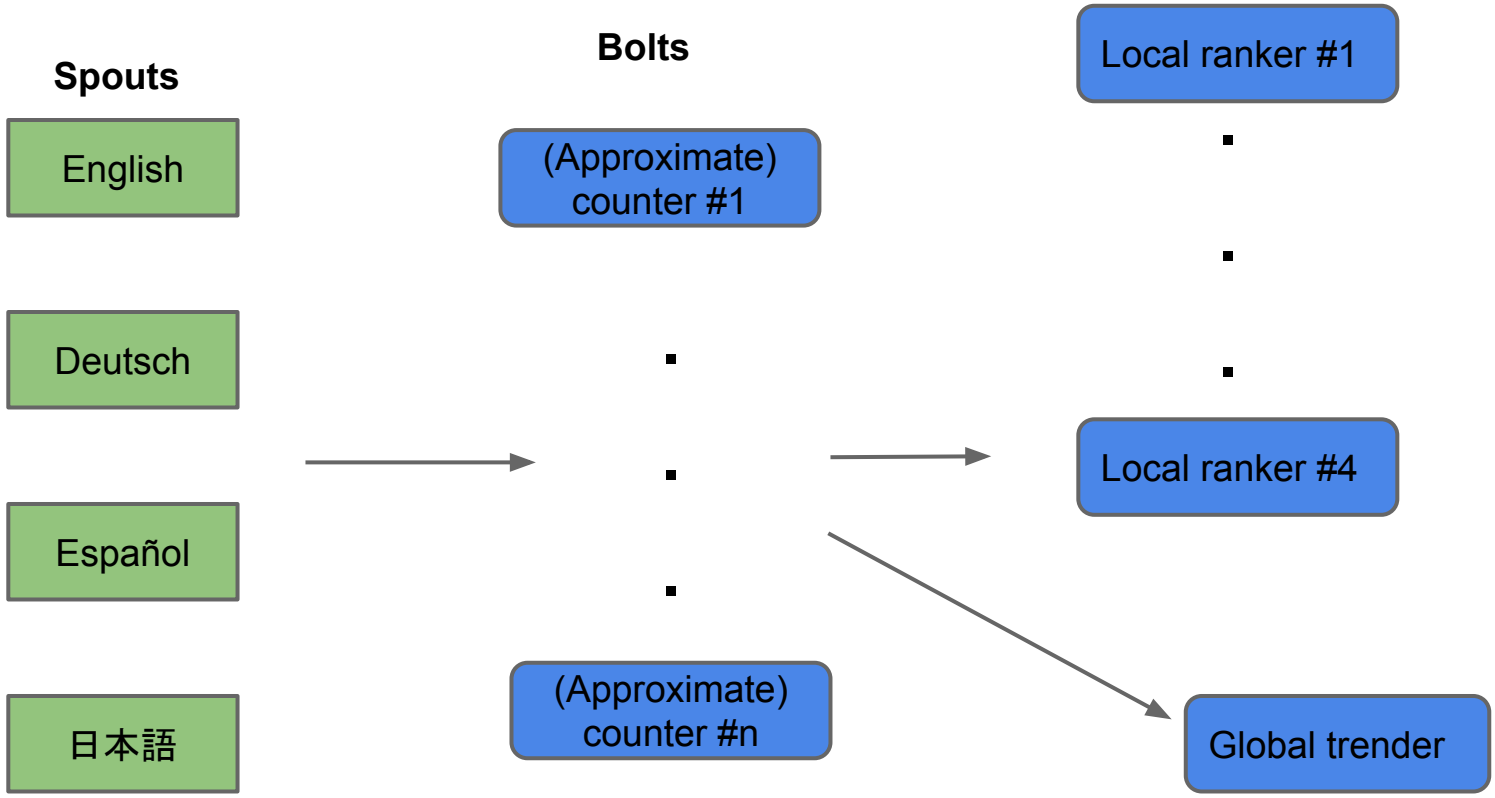
Storm topology



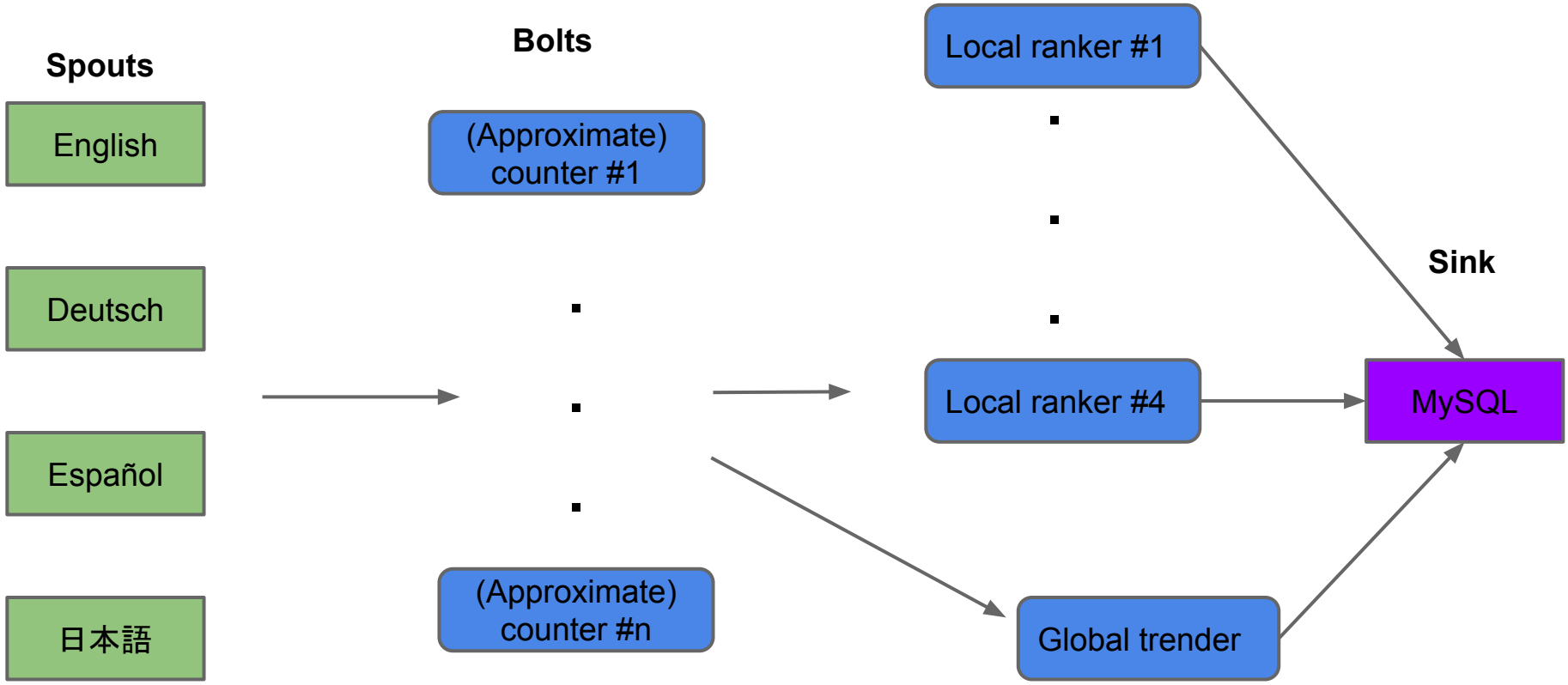
Storm topology



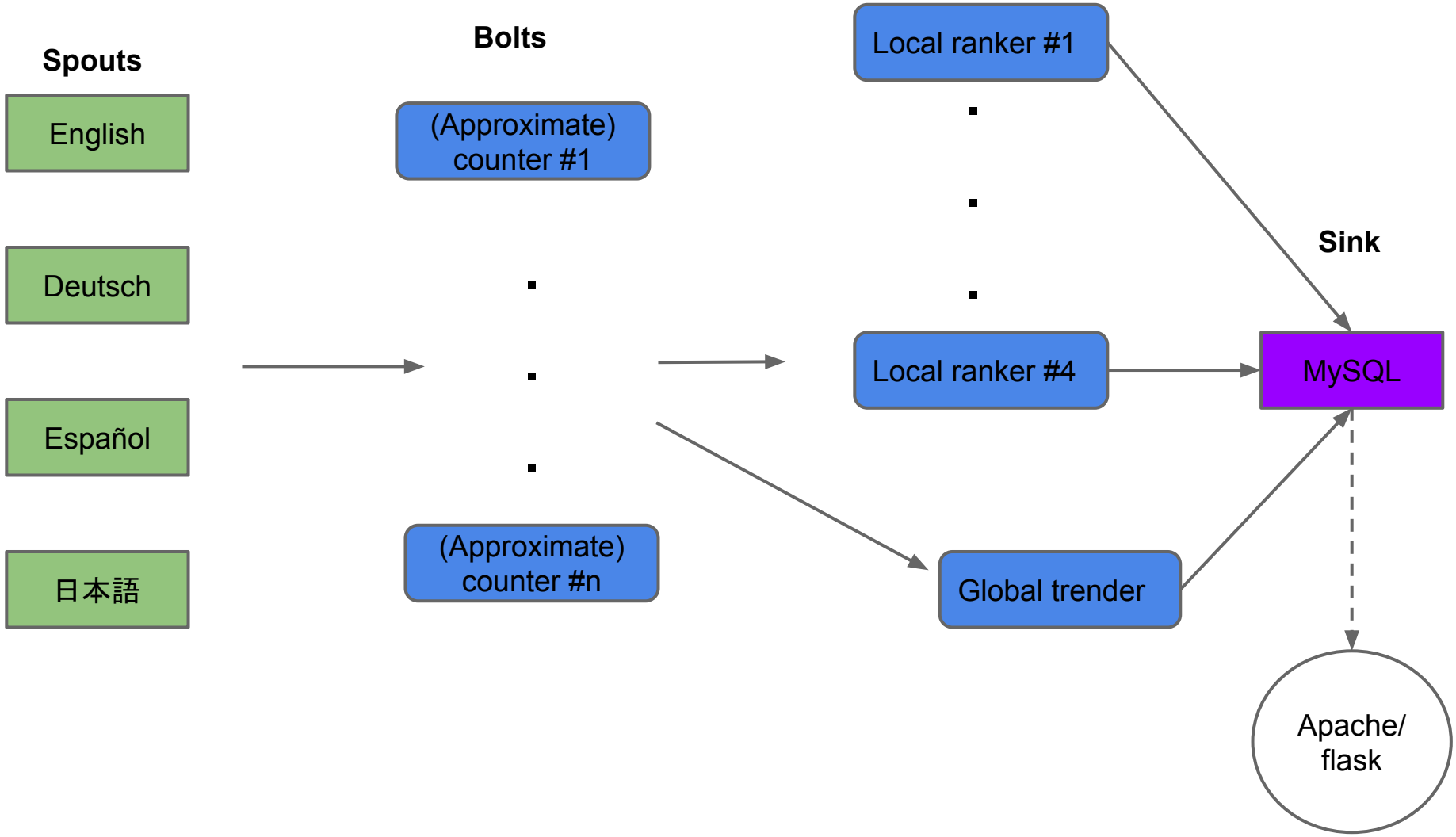
Storm topology



Storm topology



Storm topology



Expected Results

- Recent news locally and globally between the languages visible in trending topics and related people
  - E.g. Sotji medal count, Canada hockey team, Sidney Crosby.

- To a smaller degree article propagation
  - Minor changes in an English article being picked up and added to other languages.

Potential pitfalls

- Missed events
  - One person making a single, large change to a topic
  - May be solvable by comparing against similar pages which should hopefully be edited too!

- Potential noise
  - Spammers may trigger many changes and community undos will add to the number of changes!

Deployment

- Rent 4-5 Amazon EC2 instances for a two day period

- m3.large instances
  - Dual core Intel Xeon E5-2680 @2.6GHz, 32GB SSD 7.5GB RAM

- Use the Storm-deploy tool to deploy the Storm program over a

Current Progress

- Design plan
- Got the sample Storm program and a development environment locally
- Set up an EC2 account
- Able to scrape recent changes from Wikipedia in JSON format

Plan

- Create a Storm program with the proposed topology
- Setup a simple web interface to easily observe recent trends between languages
- Deploy the program on EC2
- Try to see how different topologies can make the program more efficient
- Look into page view counts as opposed to edits and see if these correspond better with recent events

Questions / Suggestions?