Welcome!

July 27th 2016

wattpad

Wattpad?

- Wattpad is a free app that lets people discover and share stories about the things they love.
- The Wattpad community is made up of 40 million people around the world. The majority of users are under 30.
- People read and create serialized stories on Wattpad. They connect around the content they love.
- We work with leading brands like Sony Pictures, Kraft, and Unilever on telling their brands stories on Wattpad.

Why do I care?

- Over 45M people use Wattpad each month
- 100k+ Signups per day
- Over 250,000,000 stories on the platform
- Spend 15.5 Billion (with a B) minutes
- Huge repository of data
- Let us tell you about how we use it all...

Spark Stories

Wattpad

Agenda

- First Spark project
- Running Spark on EMR
- Issues with Spark & Luigi
- Recommendations
- Doc2Vec & S3 Import
- Search

First Project - Event Processing

1. *https://github.com/pinterest/secor*

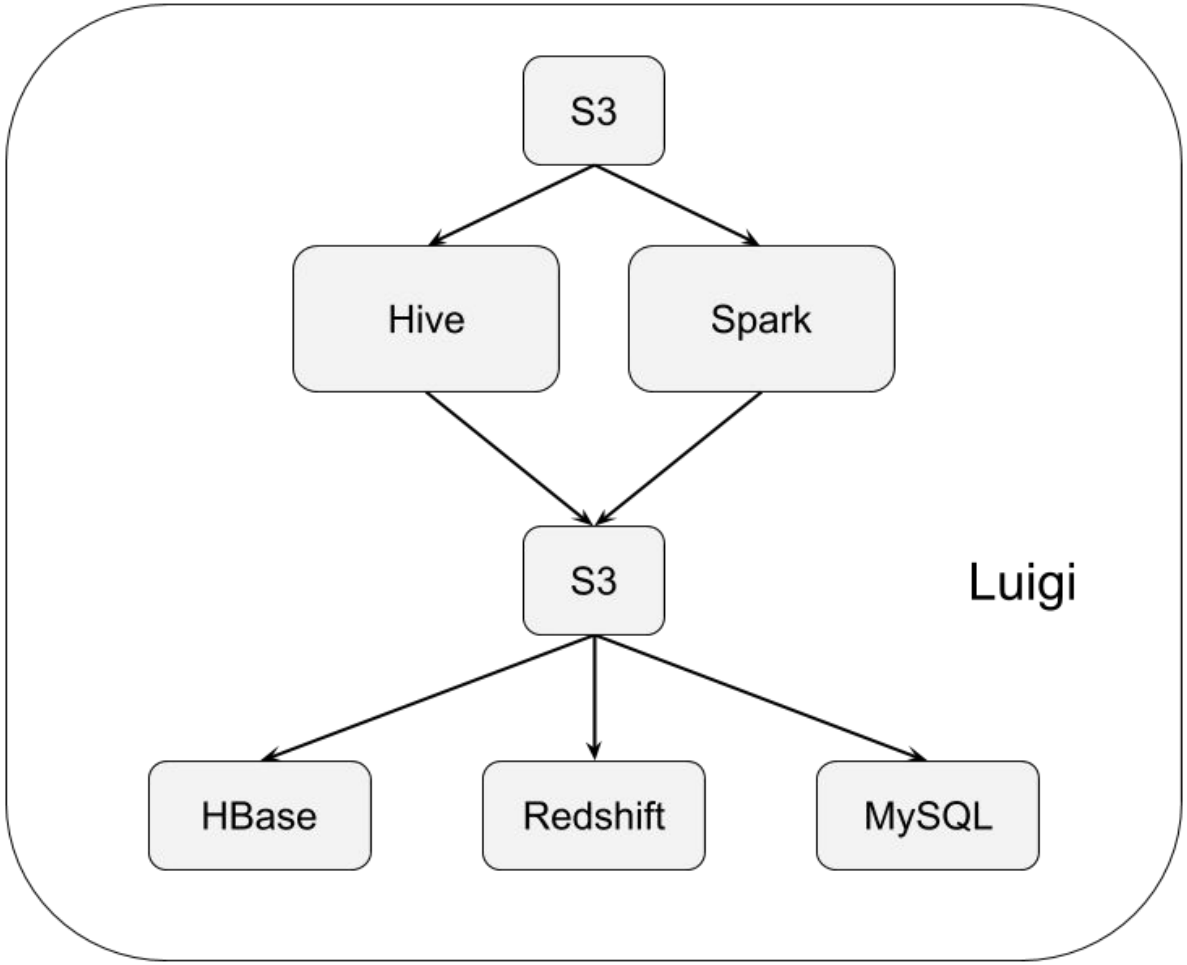
First Project - Event Processing

- Originally using Shark & Spark 0.9
  - Did not perform well as number of events scaled
  - As JSON complexity increased, became difficult to parse in Hive
- Moved to Scala & Spark 1.1
- Currently on Spark 1.3
  - Accessing JSON schema had changed

First Project - Event Processing

- **1.2B** events per day
- Run on 3 r3.8xlarge instances < 3 hours

- Settings
  - SPARK_NUM_EXECUTORS ~ 90% of available CPUs
  - SPARK_EXECUTOR_MEMORY ~85% of available memory
    - NUM_EXECUTORS * EXECUTOR_MEMORY


S3
Hive
Spark
S3
Luigi
HBase
Redshift
MySQL

EMR

- Run new clusters every night
- No permanent cluster or metastore
  - Requires loading of schemas and data
- Spot instances
- Tied to Amazon’s release cycle
  - Spark availability is tied to AMI version

Cluster settings

- NUM_EXECUTORS ~ 20% of available cores
- EXECUTOR_MEMORY ~ 13% of available memory
  - NUM_EXECUTORS * EXECUTOR_MEMORY
- DEFAULT_PARALLELISM = 3x NUM_EXECUTORS

Luigi

- Workflow Management Tool
  - Dependency resolution
  - Job scheduling
  - Visualization
- Written in python
- Community plugins

1. *https://github.com/spotify/luigi*

Luigi Task Status

Task List | Dependency Graph | Workers | Resources

TASK FAMILIES

- 8 BayesianWeighting
- 8 BayesianWeightingInput
- 8 BayesianWeightingPost
- 1 Categories
- 37 ClientShares
- 12 Comments
- 8 DailyDiscoverExports
- 8 DailyETLExports
- 8 DailyExports
- 7 DailyMySQLExports
- 7 DailyMySQLImports
- 7 DailyMySQLJobs
- 7 DailyStoryReadingProgress
- 7 DailyStoryReadingProgressCha
- 8 DailyWriterAnalyticsExports
- 8 DemographicActivitiesCountM
- 8 DemographicCompletedReads
- 8 DemographicReadsCount
- 8 DemographicStoryChapterCou
- 8 DemographicStoryRankingCor
- 8 DemographicStoryRankingMer

PENDING TASKS	48
RUNNING TASKS	2
DONE TASKS	2398
FAILED TASKS	0
UPSTREAM FAILURE	0
DISABLED TASKS	0
UPSTREAM DISABLED	0

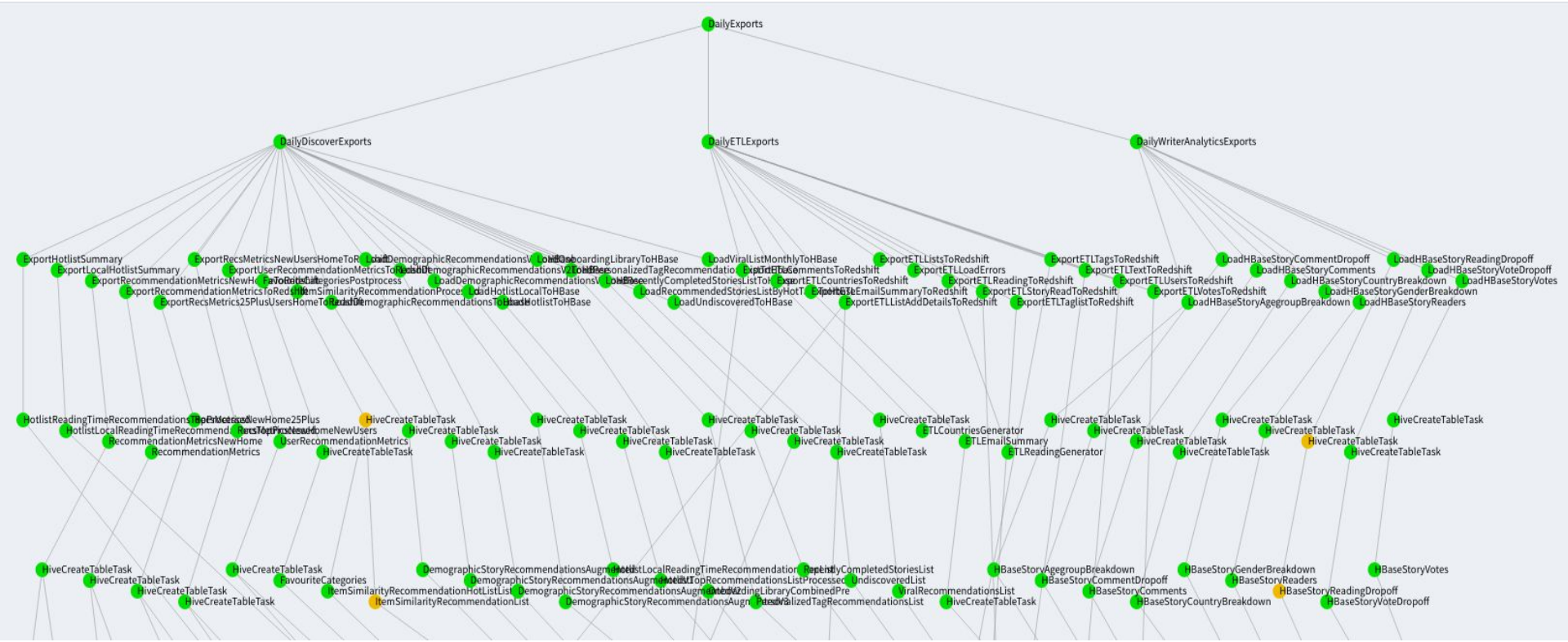
Show 10 entries

Filter table: Filter on Server

	Name	Details	Priority	Time	Actions
DONE	ExportETLListAddDetailsToRedshift	date=2016-07-21	1000	7/22/2016, 1:59:49 AM	
DONE	LoadPersonalizedTagRecommendationsListToHbase	date=2016-07-22, pool=None	0	7/23/2016, 10:41:52 AM	
DONE	HiveCreateTableTask	table=load_recommended_stories_list_by_hottags_hbase, dependencies=(RecommendedStoriesListByHotTags(date=2016-06-16),), is_hbase_task=True, pool=None, create_query= CREATE EXTERNAL TABLE load_recommended_stories_list_by_hottags_hbase (key STRING, value STRING) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with SERDEPROPERTIES ("hbase.columns.mapping" = ":key,cf:a") tblproperties ("hbase.table.name" = "hot_tags");	0	7/26/2016, 10:06:27 AM	
DONE	HBaseStoryReadingDropoff	date=2016-07-21, pool=None	0	7/22/2016, 1:57:14 PM	
DONE	ExportLocalHotlistSummary	date=2016-07-21	0	7/22/2016, 4:22:44 AM	
DONE	RecsMetricsNewHome25Plus	date=2016-07-22, pool=None	1000	7/23/2016, 2:23:11 AM	
DONE	HBaseStoryGenderBreakdown	date=2016-07-23, pool=None	0	7/24/2016, 10:49:50 AM	
DONE	LoadPersonalizedTagRecommendationsListToHbase	date=2016-07-19, pool=None	0	7/20/2016, 11:30:57 AM	
DONE	ETLListAddDetails	date=2016-07-18, pool=None	1000	7/26/2016, 1:49:03 AM	
DONE	RecentlyCompletedStories	date=2016-07-19, pool=None	0	7/20/2016, 5:27:53 AM	

Showing 1 to 10 of 2,448 entries

Previous 1 2 3 4 5 ... 245 Next


DailyExports
DailyDiscoverExports
DailyETLExports
DailyWriterAnalyticsExports
ExportHotlistSummary
ExportLocalHotlistSummary
ExportRecommendationMetricsNewHo
ExportRecommendationMetricsToRedshift
ExportRecsMetrics25PlusUsersHomeToRedshift
ExportRecsMetricsNewUsersHomeToR
ExportUserRecommendationMetricsToRedshift
FavouriteCategoriesPostprocess
ItemSimilarityRecommendationProces
LoadDemographicRecommendationsV
LoadDemographicRecommendationsV2ToHBase
LoadDemographicRecommendationsV
LoadHotlistLocalToHBase
LoadDemographicRecommendationsToHBase
LoadHotlistToHBase
LoadOnboardingLibraryToHBase
LoadPersonalizedTagRecommendatio
LoadRecentlyCompletedStoriesListToH
LoadRecommendedStoriesListByHotT
LoadViralListMonthlyToHBase
LoadUndiscoveredToHBase
ExportETLCommentsToRedshift
ExportETLCountriesToRedshift
ExportETLEmailSummaryToRedshift
ExportETLListAddDetailsToRedshift
ExportETLListsToRedshift
ExportETLLoadErrors
ExportETLReadingToRedshift
ExportETLStoryReadToRedshift
ExportETLTaglistToRedshift
ExportETLTagsToRedshift
ExportETLTextToRedshift
ExportETLUsersToRedshift
ExportETLVotesToRedshift
LoadHBaseStoryCommentDropoff
LoadHBaseStoryComments
LoadHBaseStoryCountryBreakdown
LoadHBaseStoryGenderBreakdown
LoadHBaseStoryAgegroupBreakdown
LoadHBaseStoryReadingDropoff
LoadHBaseStoryVoteDropoff
LoadHBaseStoryVotes
LoadHBaseStoryReaders
HotlistReadingTimeRecommendations
HotlistLocalReadingTimeRecommenda
RecsMetricsNewHome25Plus
RecsMetricsNewHomeNewUsers
RecommendationMetricsNewHome
RecommendationMetrics
UserRecommendationMetrics
HiveCreateTableTask
ETLCountriesGenerator
ETLEmailSummary
ETLReadingGenerator
FavouriteCategories
ItemSimilarityRecommendationHotListList
ItemSimilarityRecommendationList
DemographicStoryRecommendationsAugmented
DemographicStoryRecommendationsAugmentedV2
DemographicStoryRecommendationsAugmentedV3
HotlistLocalReadingTimeRecommendation
HotlistTopRecommendationsListProcessed
OnboardingLibraryCombinedPre
PersonalizedTagRecommendationsList
RecentlyCompletedStoriesList
UndiscoveredList
ViralRecommendationsList
HBaseStoryAgegroupBreakdown
HBaseStoryCommentDropoff
HBaseStoryComments
HBaseStoryCountryBreakdown
HBaseStoryGenderBreakdown
HBaseStoryReaders
HBaseStoryReadingDropoff
HBaseStoryVoteDropoff
HBaseStoryVotes

Luigi Tasks

```python
import luigi

class MyTask(luigi.Task):
    param = luigi.Parameter(default=42)

    def requires(self):
        return SomeOtherTask(self.param)

    def run(self):
        f = self.output().open('w')
        print >>f, "hello, world"
        f.close()

    def output(self):
        return luigi.LocalTarget('/tmp/foo/bar-%s.txt' % self.param)

if __name__ == '__main__':
    luigi.run()
```

The business logic of the task

Where it writes output

What other tasks it depends on

Parameters for this task

Luigi - Spark Shared Context

- Included spark support starts & tears down the spark context for every task
  - Adds overhead
  - Duplicate work
  - Interim storage can be required

How can we do better?

- Create our own Spark Task
- Keep a reference count of the number of tasks using the context
- Overwrite the callbacks

Manipulate the context counter

```
def get_context(self):
    if SharedSparkContext.applications == 0:
        self.initialize_context()

    SharedSparkContext.applications += 1
    return self

def release_context(self):
    SharedSparkContext.applications -= 1

    if SharedSparkContext.applications == 0:
        self.sc.stop()
```

Create a singleton

```
class SharedSparkContext(object):
    __metaclass__ = core.Singleton

    # Records the number of Spark applications currently using this Spark context
    applications = 0

    # Spark configuration and context information that will be shared across multiple
Spark applications
    conf = None
    sc = None
    sqlc = None
```

Create our Luigi Spark Task

```
class BaseSparkApplication(luigi.Task):
    task_namespace = ‘spark’
    ctx = None

    def complete(self):
        is_complete = super(BaseSparkApplication, self).complete()
        if not is_complete and not self.ctx:
            self.ctx = SharedSparkContext().get_context()
        return is_complete

    def on_success(self):
        self.ctx.release_context()

    def on_failure(self, exception):
        self.ctx.release_context()
        return super(BaseSparkApplication, self).on_failure(exception)
```

Luigi - Spark UDFs

- Luigi code is only loaded onto the Master node
- Needed a way to make UDFs available to all nodes
- Extend our custom Spark Class

On initialization

- Walk the UDF directory to get the files we need
- Zip the files
- Add the files to the Spark Context & register the UDFs

```
self.sc.addPyFile('file://' + zipfilename.filename)

# Create Hive context and register any UDFs available
self.sqlc = HiveContext(self.sc)
self._register_udfs(udfs)
```

```
def _register_udfs(self, udfs):
    # udfs are a list of named tuples ('SparkSQLUDF', 'name, func, return_type')

    for udf in udfs:
        self.sqlc.registerFunction(udf.name, udf.func, udf.return_type)
```

Scala vs. Python

- Developers & Data Scientists tend to know Python better
- Integrates better with Luigi
- Better build process in Python
- Libraries not available in Scala
  - Nltk, scipy
- But scala is better sometimes
  - Event processing
  - Recommendations

What’s Next

- Open source
- Investigate SparkSession in Spark 2.0
- Investigate Tachyon

Recommendations

Recommendations

Recommendations
A fresh set of stories, just for you

The Good Girl's Bad Boys: The
by Rubix
HUMOR
41.1M ★ 1.3M

Evolution (Book 1 of POE
by Nam
SCIENCE FICTION
1.1M ★ 29.3K

News & Updates
by Wattpad
RANDOM
6.5M ★ 73.5K

THE DENTIST
by Kelly Anne Bl...
HORROR
21.7K ★ 1.6K

Manor Dark
by Stephen Clark...
FANTASY
7.9K ★ 421

WIN: The Atlantis Grail (Book Thre...
by VeraNazarian
SCIENCE FICTION
312K ★ 8.8K

Reading lists by Written In Action >
Collections from a top profile

Follow

WIA Undiscovered 2
42 Stories
#futuristic #fantasy #adventure #epic #action

WIA Undiscovered
172 Stories
#futuristic #action #adventure #fantasy #romance

Stories by Pantopicon >
Stories written by a profile you follow

Following

The Butcher's Daughter
by Pantopicon
GENERAL FICTION
197 ★ 21
The Butcher's Daughter follows 24 hours in the life of Cookie; club kid, mad genius, and controlled sociopath.

Air Thin and Eager
by Pantopicon
GENERAL FICTION
1.5K ★ 92
My name is Loretta but I prefer Lottie I'm closing in on my fifteenth year And if you think you have seen a pair of eyes more green Then you sure didn't see them around...

Stories by 女王 >
A profile we think you'll love

Follow

Completed Stories
Recently completed stories just for you

Enhancement #Wattys2015
by kelly
SCIENCE FICTION
702K ★ 47.7K

Evolution (Book 1 of POE
by Nam
SCIENCE FICTION
1.1M ★ 29.3K

The Purge
by tobi
SCIENCE FICTION
115K ★ 4.8K

Breeder Nation (#Wattys2016)
by Kara Michelle
SCIENCE FICTION
815K ★ 44K

Unbrokenworld: The Crystal Caves
by Carla
SCIENCE FICTION
185K ★ 13.3K

Experiment (Completed)
by Anna
SCIENCE FICTION
117K ★ 7.5K

#spock stories >
Because you searched for startrek

Star Trek Reader Inserts
by Whovian3135
FANFICTION
314K ★ 10.8K

Star Trek Reader Inserts 2
by Whovian3135
FANFICTION
165K ★ 7.3K

Boarding Cassandra (A Star
by
FANFICTION
80.8K ★ 2.1K

Shooting Star : Star Trek One-
by owenharper
FANFICTION
52.5K ★ 2K

Book 1: Love In The Stars (Star
by CrazyDreamGirl
FANFICTION
27.8K ★ 862

Emotionless (Spock Love
by UnaNova
FANFICTION
21.2K ★ 707

#mystery stories >
Because you searched for wasted life

The Girl He Never Noticed
by
ROMANCE
104M ★ 2.6M

The Shy Girl Has a Gun
by makeandoffer
ACTION
28.4M ★

I'm a Model that's Undercover as T...
by KittyKattt
TEEN FICTION
17.6M ★

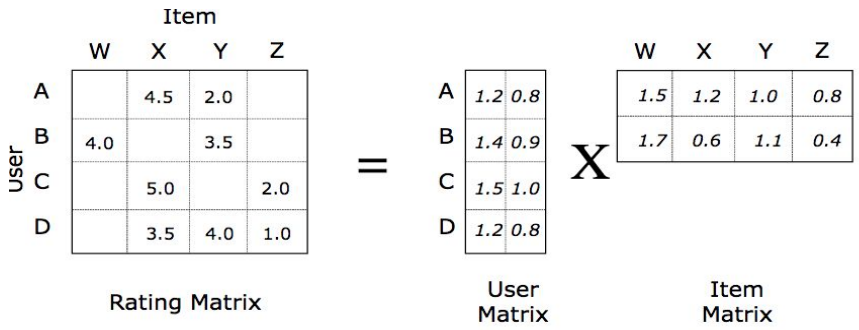
Struck (A Vampire Novel) ✔
by CaitSarai
VAMPIRE
12.4M ★

Fatal Alliances
by CeilneMahadeo
ROMANCE
10M ★ 361K

Marked by the Alpha
by Zabellerain
WEREWOLF
12M ★ 319K

Recommendations: MLlib

- ALS Matrix Factorization
  - Implicit signals
- Scala:
  - Performance
  - Library compatibility
- Personalized Recs:
  - **800M** ratings
  - 20M users x 10M stories
- Similar Items:
  - **1.3B** ratings
  - 30M users x 25M stories



Recommendations: Challenges

- 200 Trillion operations
  - Breeze (Runtime: 20+ hours)
  - JBLAS (Runtime: no change)
  - Native BLAS (Runtime 8 hours 60%↓)
- Approximate Nearest Neighbors (Run time: 3 hours 62.5%↓)

1. *Fast and Accurate Maximum Inner Product. Recommendations on Map-Reduce. Rob Hall and Josh Attenberg . Etsy Inc.*
2. *Improved Asymmetric Locality Sensitive Hashing (ALSH) for Maximum Inner Product Search (MIPS), Anshumali Shrivastava and Ping Li*

Recommendations: ANN With pySpark

- What is ANN?
  - searches subset of items
  - keeping accuracy as high as possible
- Spotify's Annoy[1] Library
  - speed
  - accuracy
  - python
- Run time: 40 mins (78% ↓)
- Also available in Scala now[2]

```
# Add annoy index to spark
sc.addPyFile(os.path.join(base_path, annoy_fname))

# KNN method
def get_KNN(user_features):
        t = AnnoyIndex(rank, metric='euclidean')
        t.load(SparkFiles.get(annoy_fname))
        return [t.get_nns_by_vector(vector=features,
n=k, search_k=-1, include_distances=True)) for
user_id, features in user_features]

# Call to KNN method
recommendations = user_features.mapPartitions
(get_KNN)
```

https://github.com/spotify/annoy
https://github.com/karlhigley/spark-neighbors

Recommendations: Diversification

- Balances accuracy and diversity of items in a list.
- Algorithms:
  - Pairwise IntraList diversification[1]
  - Clustering based genre diversification[2]
- Runtime:
  - 20 mins

1. *The use of MMR, diversity-based reranking for reordering documents and producing summaries, Jaime Carbonell & Jade Goldstein*
2. *Improving recommendation lists through topic diversification, Cai-Nicolas Ziegler et al*

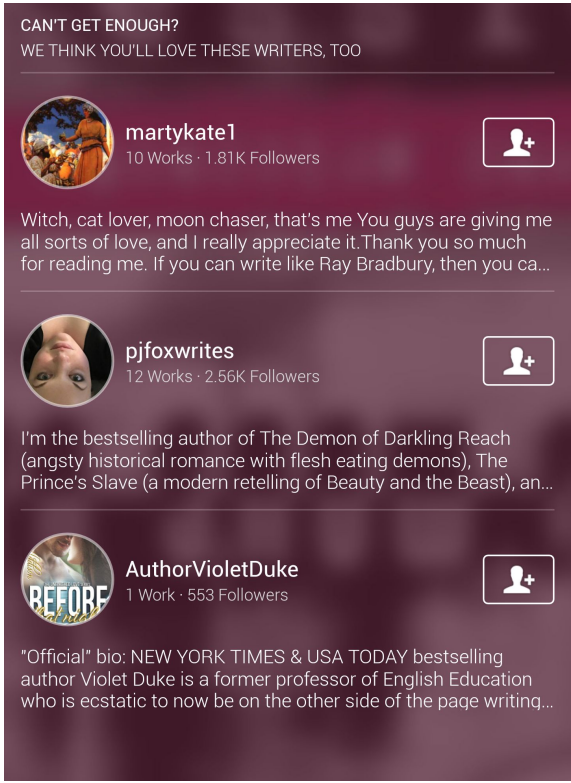
Doc2Vec & S3 Import

Doc2Vec

- A neural net technique for vectorizing segments of text. Extension of Word2Vec.
- Useful for performing content-based tasks: recommendations, clustering. Especially useful for cases where little user activity is available.

Text → **Doc2Vec Model** → $v=(v_1,\ldots,v_m)$

Use case: Content based user recommendation.



Use case: Content based user recommendation.

Trained model (single machine): process >1M English stories (use implementation in **Gensim**), corresponding to ~450k users ⇒ vectorize stories & users.

For a given story (distributed): Vectorize its text. Run KNN

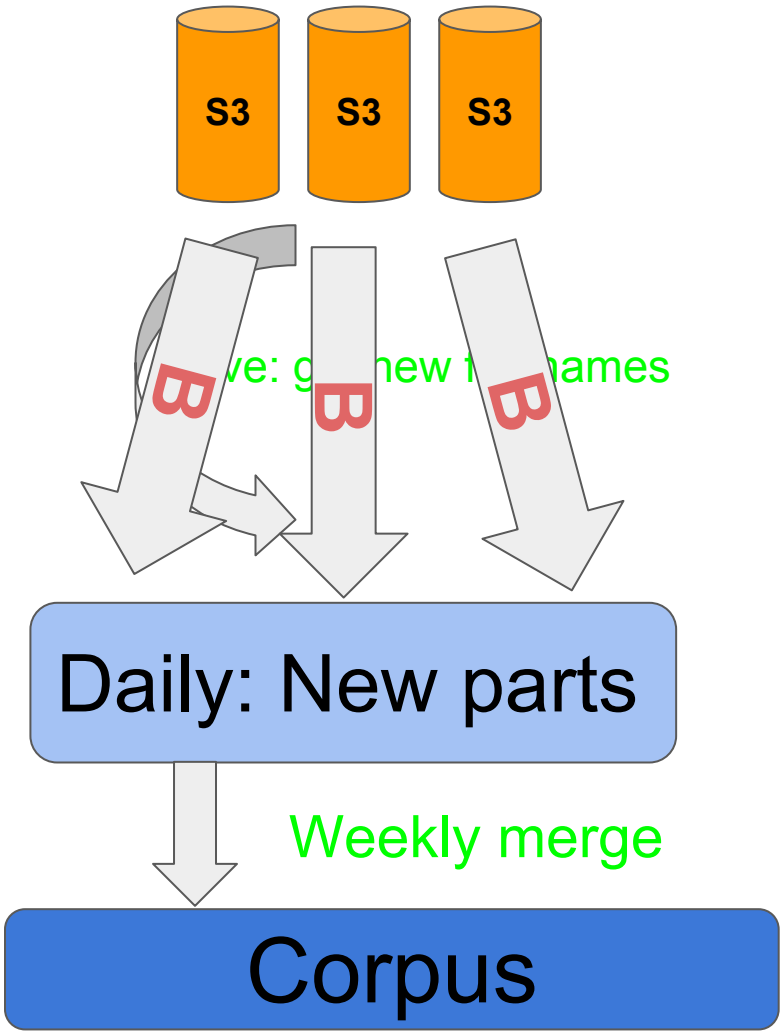
```
Doc2vec_model = sc.broadcast(Doc2Vec.load(model_loc, 'd2v_writers_model'))

def get_story_vectors(story_id, text):
    words = extract_story_data(story_id, text)
    return doc2vec_model.value.infer_vector(words)

def get_nearest_writers(model, story_data, k):
    story_id, story_vec = story_data
    return nsmallest(k,
                     model.value.docvecs.doctag_syn0,
                     key=lambda writer_vec: cosine(vec, story_vec))
```

Text import from S3

- Corpus is stored in folders in an S3 bucket.
- Process text (English, Spanish, Tagalog).
- Daily throughput: **~450-500k** story chapters (**~3GB**).
- Outline of the process:
  - Retrieve list of new parts added yesterday.
  - Boto+Spark: Retrieve raw text; process & clean-up text.



Text import from S3 (continued)

- Previously implemented in Pig + Bash.
- Running-time improvement:
  - Previously: ~3 hours (for English only)
  - Currently: ~1 hour (for three languages; roughly **x2** the total amount of text).

Next steps:

- Add support for all languages.
- Optimize the process by switching to Spark dataframe operations.

Search

Search

- Use Elasticsearch to power our search system
- API queries Elasticsearch
- Calculate tf-idf for parts of the documents
  - Titles
  - Descriptions
  - Interaction data: Votes, Reads, etc.
  - Many other fields
- Index is updated as documents change

Search: Metrics

- Track search and click events on results
- Allows us to calculate performance metrics:
  - DCG, P@N, etc.
- Use this to identify poor search performance for certain queries
- Use this to run A/B tests on search
- Side note: Use Spark to quickly iterate on these metrics

ROGERS 3:21 PM 92%

wattpad.com

Wattpad ★★★★★ 1,000,000+ ratings

Get the App

canada

Refine by: selected tag unselected

Showing results for **Canada** instead of **canadaa**

Other suggestions Canadian Canuck Kaneda

Can't find the story you're looking for?
Try adjusting your **settings**.

2.1K STORIES

Love Can Change You - 2ª Temporada
Cat_Somerhalder • 28 Parts • Fanfiction
71.4K 4.2K
A rapariga que perdeu os seus pais e que mudou de vida agora descobriu algo que a fez ganhar curiosidade. Os pais que ela tinha afinal não eram os seus verdad...

wattpad.com

Discover Create mystery

mystery

Filters

Related tags: thriller murder adventure horror suspense

334K STORIES Filters

Confessions About Colton
colourlessness • 48 Parts • Mystery / Thriller
2.9M 148K
"Seven letters seven confessions seven clues." When smart and charismatic Colton Crest disappears his small country town in rural New South Wales jumps int...
Completed

The Secret - Book 1 (Watty Award Winner)
Monrosey • 54 Parts • Mystery / Thriller
3.8M 128K
THANK YOU, Wattpad, for the HQ Love Watty Award! Concept: A mysterious young woman flees from her troubled past only to discover it is impossible to escape TH...
Completed

Dear Diary (Wattys2016)
Mr__Writer • 52 Parts • Mystery / Thriller
720K 36.2K

Search: Reformulations

- Suggest new query or automatically replace query for common typos
- Sessionize queries
  - Amount of time between actions
  - Dwell time on document
  - String similarity between queries
  - How likely is this query going to result in a click
- Result is a mapping of commonly misspelled queries to correct queries

Search: Query Suggestions

- Help people browse search by suggesting related queries
- Again using sessionization similar to before
- Vectorize queries and items clicked on using word2vec
- Other searchers where a similar group of stories were clicked on is a good signal

Search: Reorder results

- Develop a mapping of documents => queries that people use to get to this document
- Sessionization comes in handy here again because we can get more context (other queries in this session)
- This addition signal can be plugged into Elasticsearch as another field and used at query time
- All these jobs are done in batch daily

Questions?

We're Hiring
Come see us!