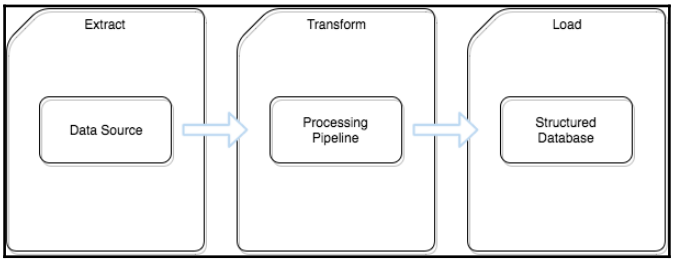
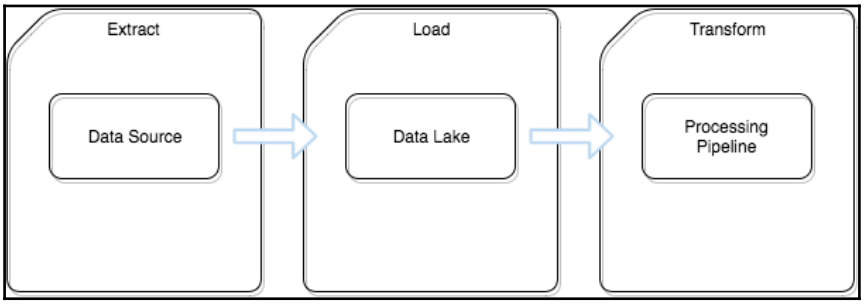
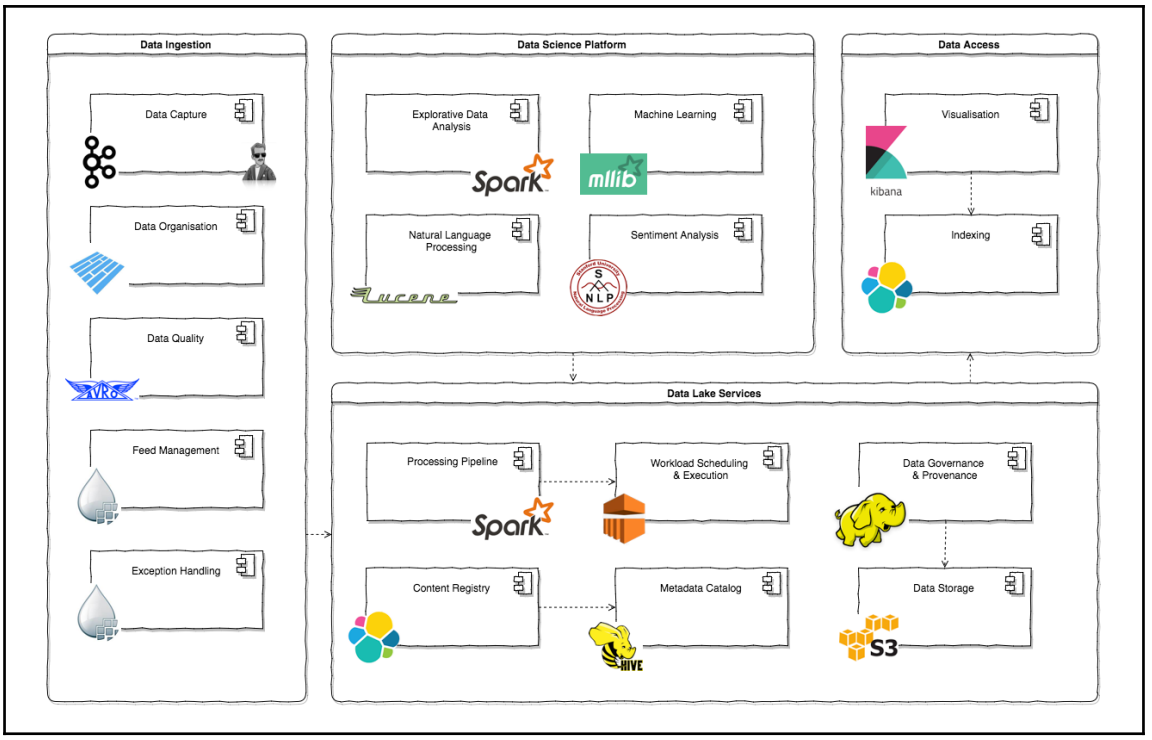
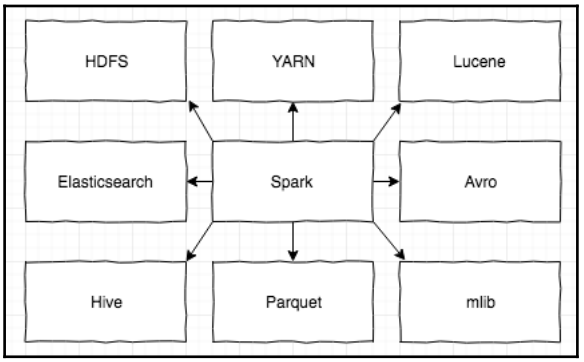
Chapter 01: The Big Data Science Ecosystem

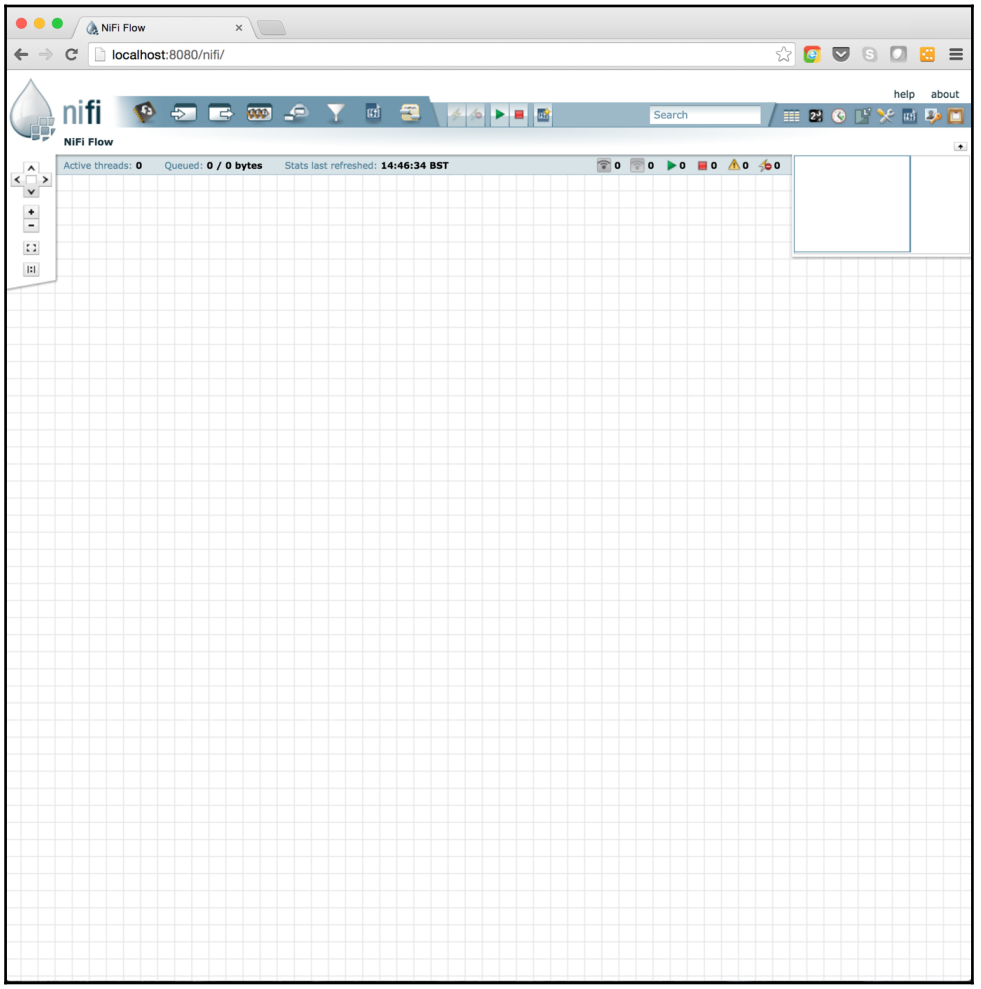


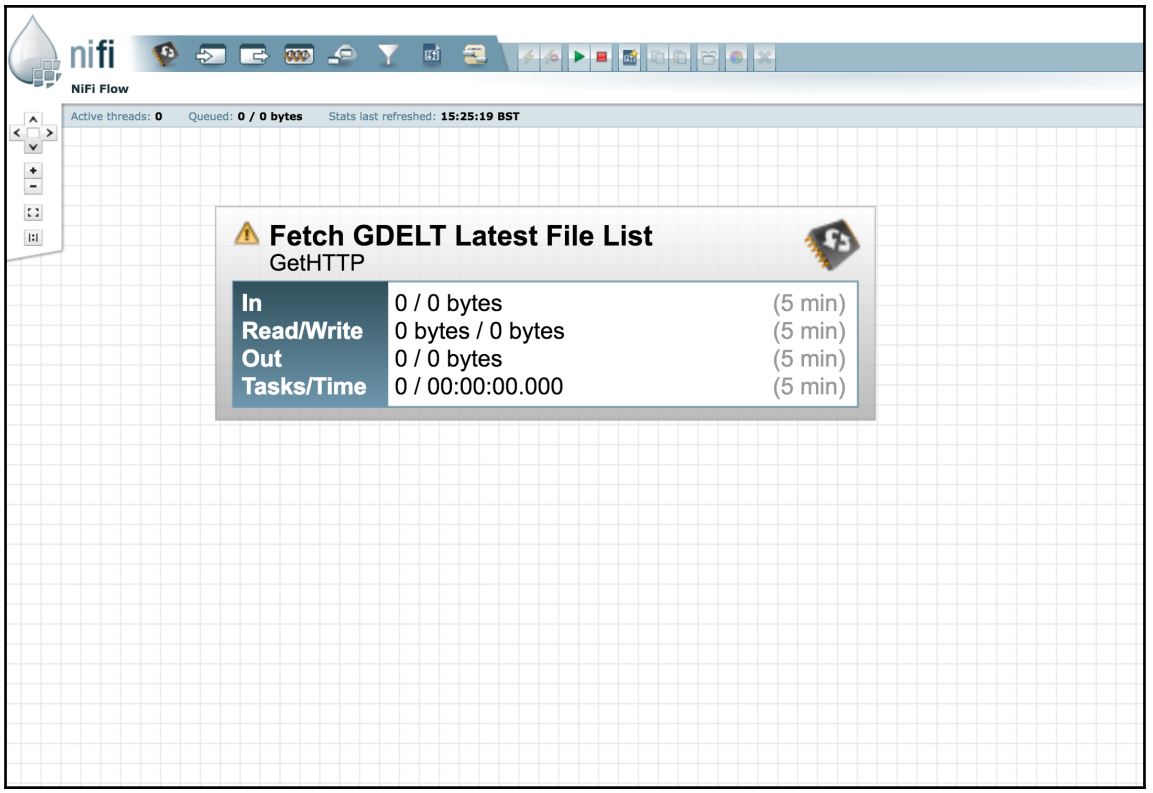


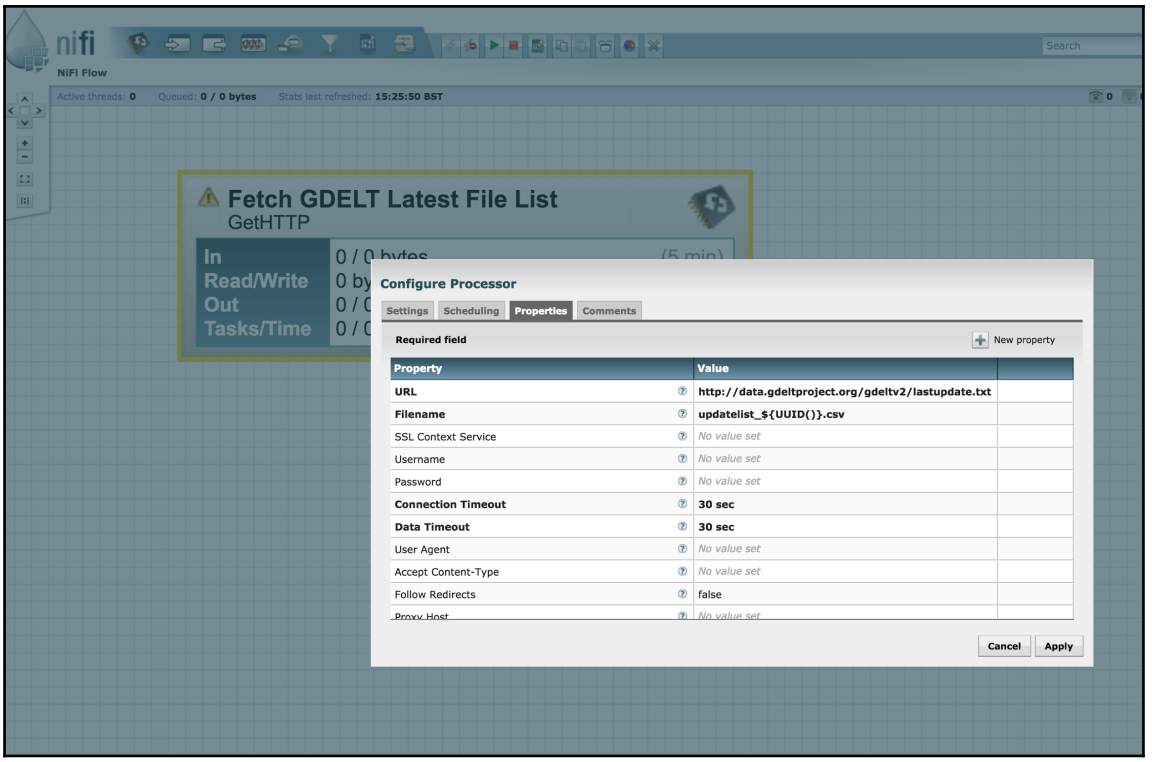

Data Ingestion
Data Capture
Data Organisation
Data Quality
Feed Management
Exception Handling
Data Science Platform
Explorative Data Analysis
Spark
Machine Learning
mllib
Natural Language Processing
lucene
Sentiment Analysis
S NLP
Data Access
Visualisation
kibana
Indexing
Data Lake Services
Processing Pipeline
Spark
Workload Scheduling & Execution
Data Governance & Provenance
Content Registry
Metadata Catalog
HIVE
Data Storage
S3

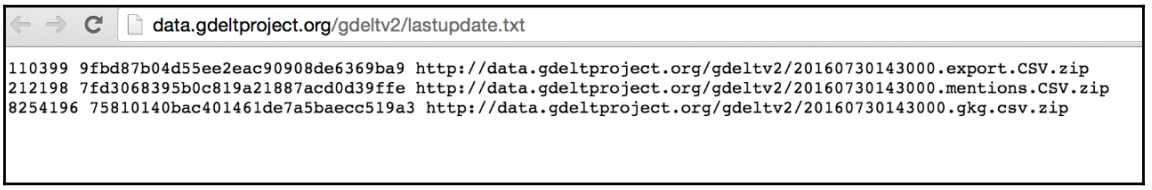

HDFS
YARN
Lucene
Elasticsearch
Spark
Avro
Hive
Parquet
mlib

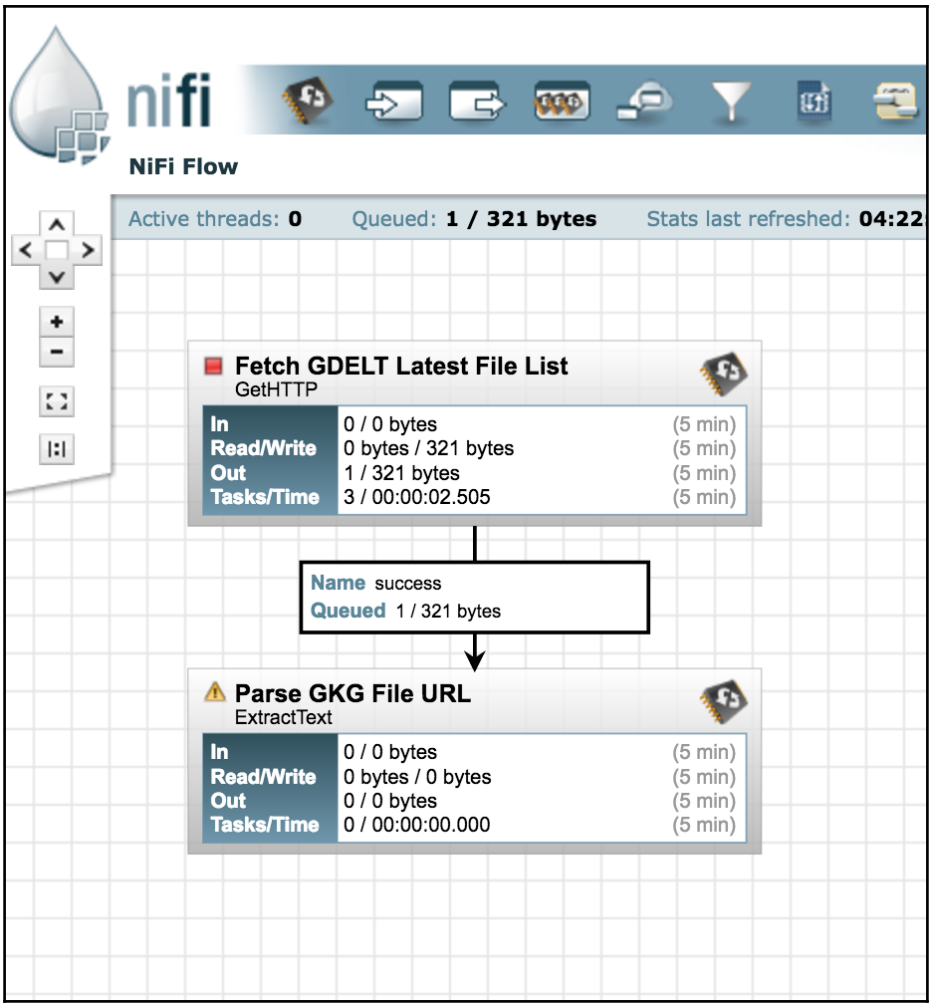
Chapter 02: Data Acquisition

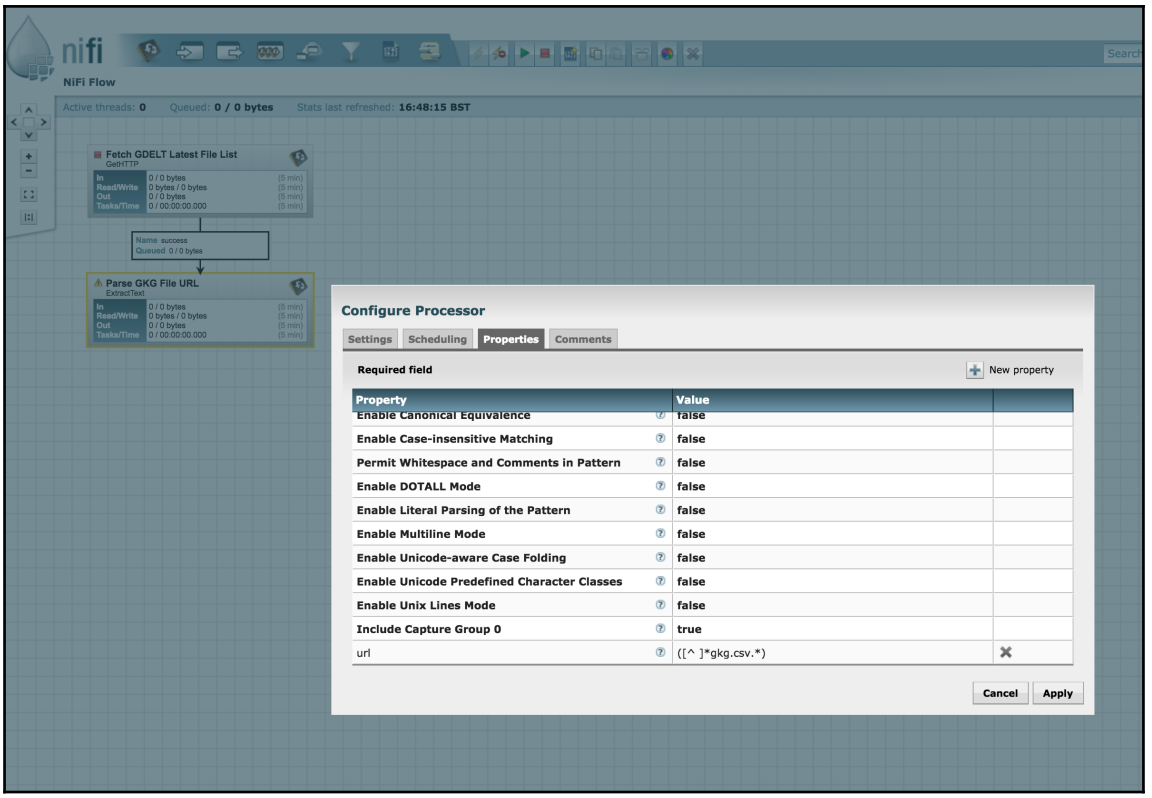


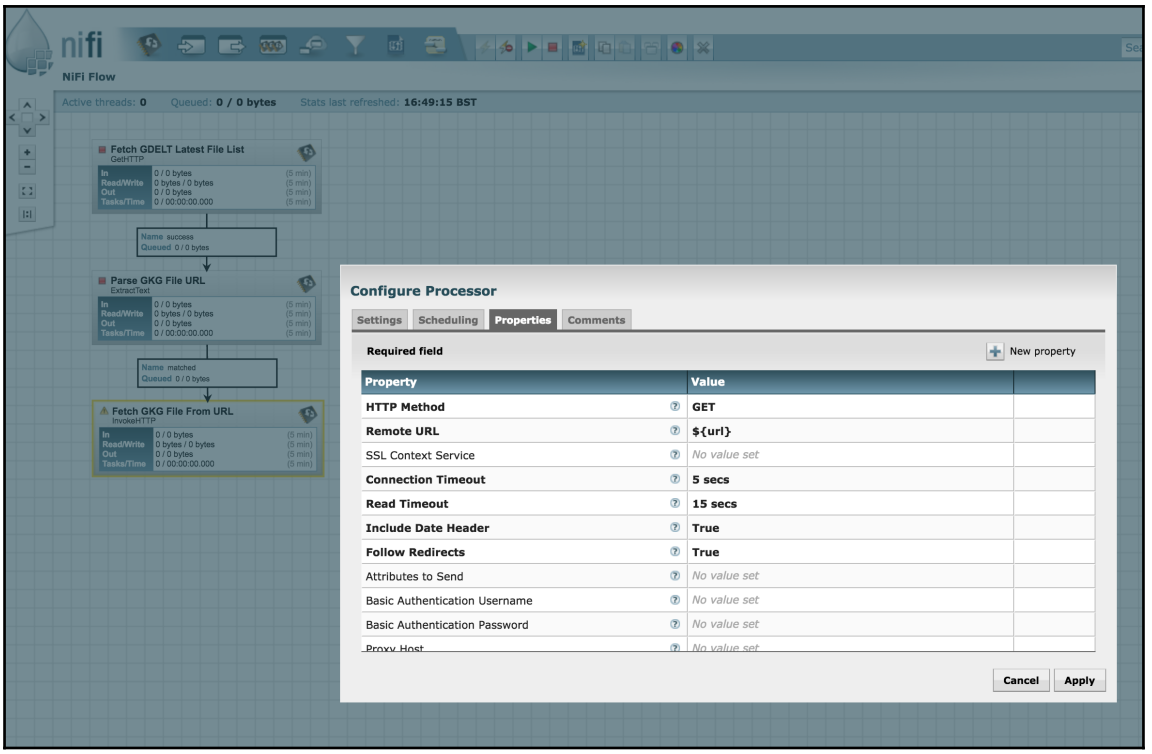

nifi
NiFi Flow
Active threads: 0
Queued: 0 / 0 bytes
Stats last refreshed: 15:25:19 BST
Fetch GDELT Latest File List
GetHTTP
In
0 / 0 bytes
(5 min)
Read/Write
0 bytes / 0 bytes
(5 min)
Out
0 / 0 bytes
(5 min)
Tasks/Time
0 / 00:00:00.000
(5 min)

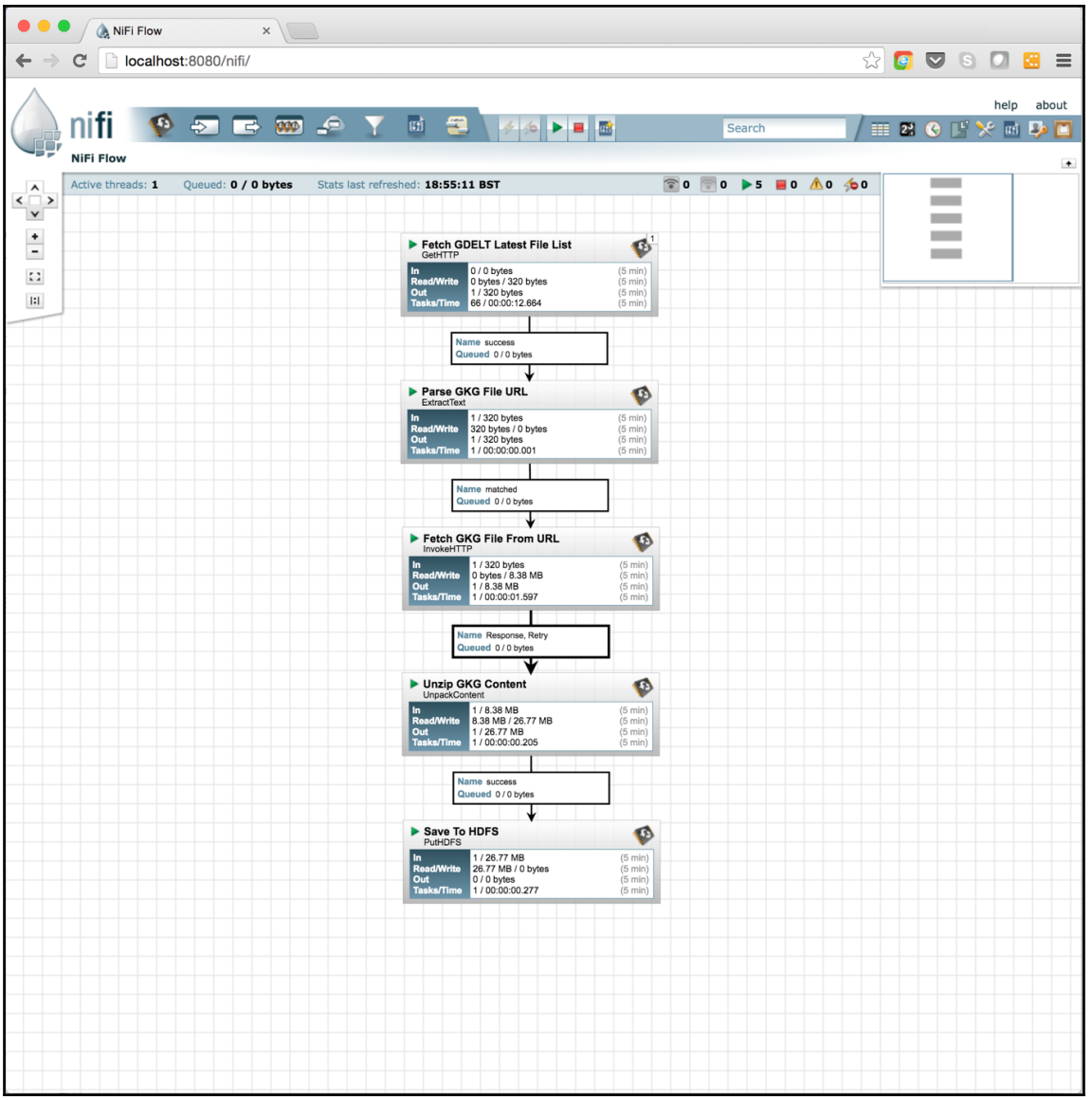

nifi
NiFi Flow
Search
Active threads: 0
Queued: 0 / 0 bytes
Stats last refreshed: 15:25:50 BST
Fetch GDELT Latest File List
GetHTTP
In
Read/Write
Out
Tasks/Time
Configure Processor
Settings
Scheduling
Properties
Comments
Required field
New property
Property
Value
URL
http://data.gdeltproject.org/gdeltv2/lastupdate.txt
Filename
updatelist_${UUID()}.csv
SSL Context Service
No value set
Username
No value set
Password
No value set
Connection Timeout
30 sec
Data Timeout
30 sec
User Agent
No value set
Accept Content-Type
No value set
Follow Redirects
false
Proxy Host
No value set
Cancel
Apply

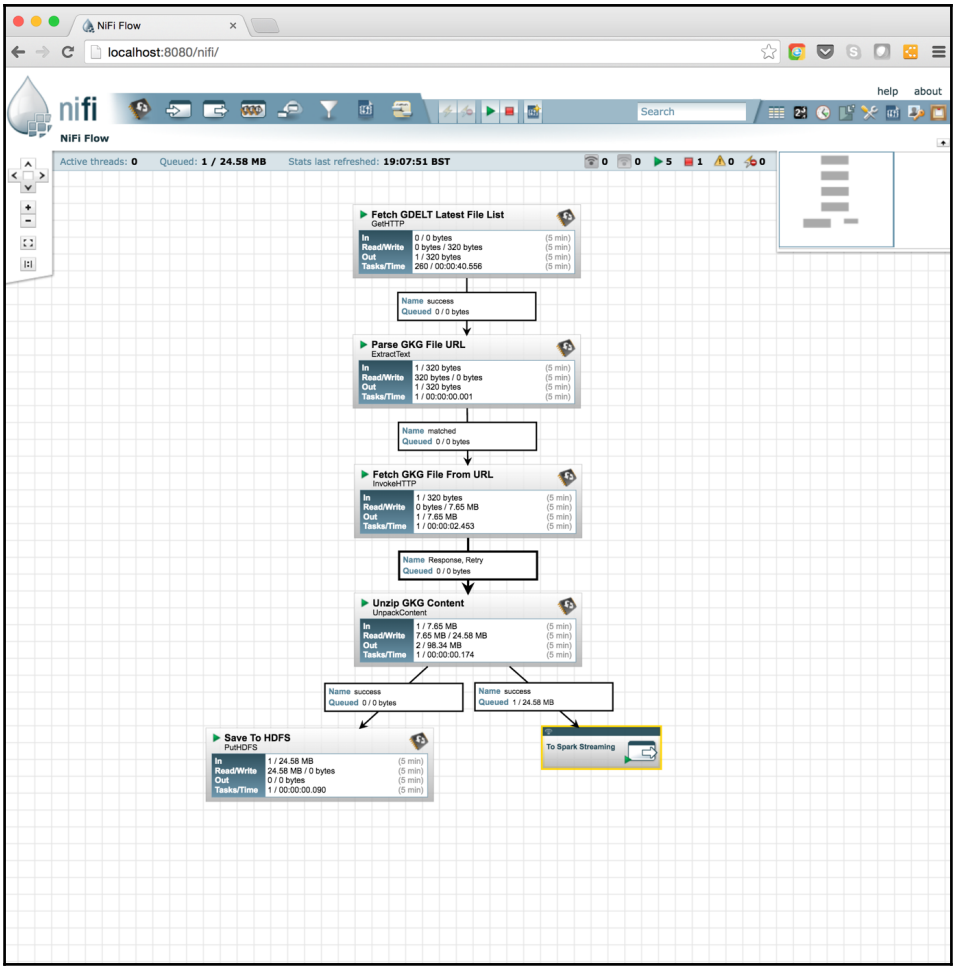

data.gdeltproject.org/gdeltv2/lastupdate.txt
110399 9fbd87b04d55ee2eac90908de6369ba9 http://data.gdeltproject.org/gdeltv2/20160730143000.export.CSV.zip
212198 7fd3068395b0c819a21887acd0d39ffe http://data.gdeltproject.org/gdeltv2/20160730143000.mentions.CSV.zip
8254196 75810140bac401461de7a5baecc519a3 http://data.gdeltproject.org/gdeltv2/20160730143000.gkg.csv.zip

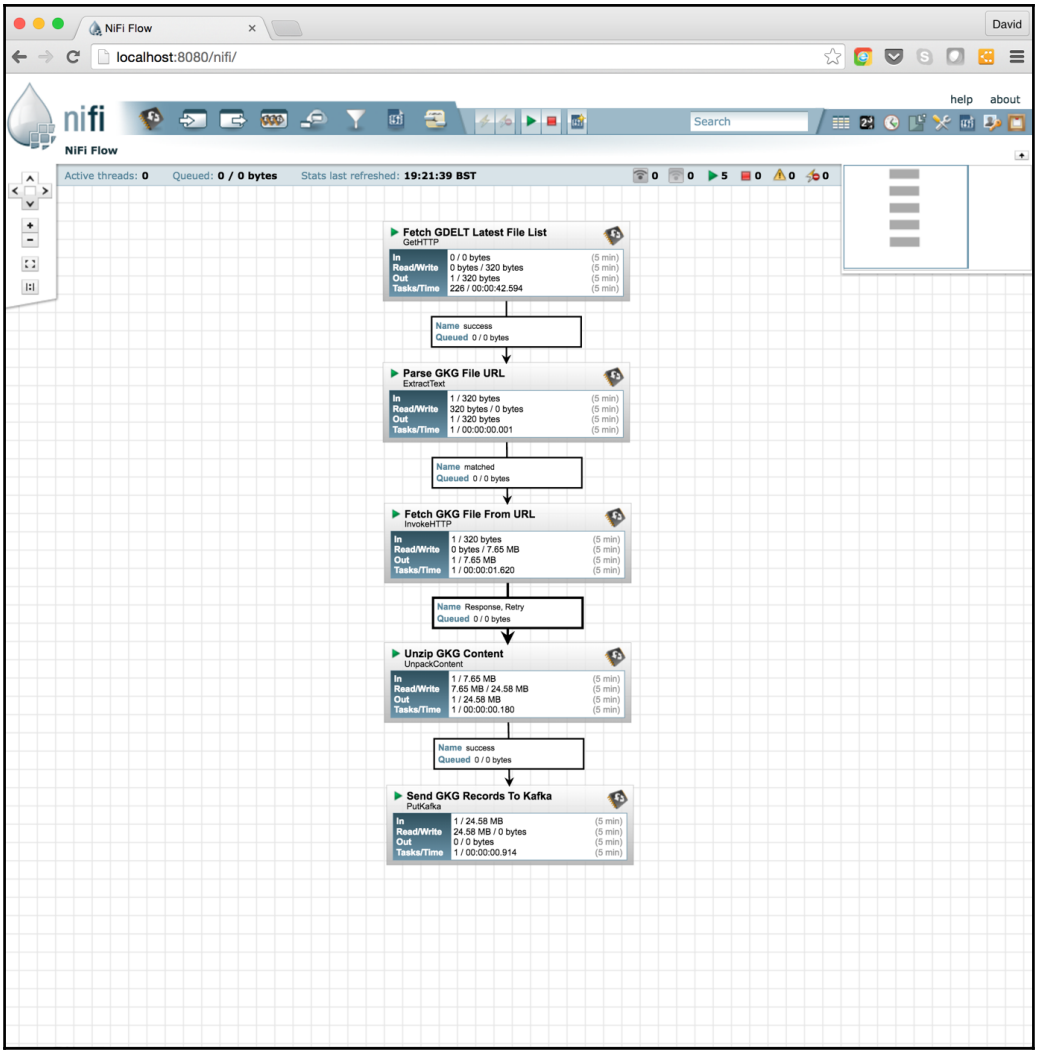

nifi
NiFi Flow
Active threads: 0
Queued: 1 / 321 bytes
Stats last refreshed: 04:22
Fetch GDELT Latest File List
GetHTTP
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 321 bytes (5 min)
Out 1 / 321 bytes (5 min)
Tasks/Time 3 / 00:00:02.505 (5 min)
Name success
Queued 1 / 321 bytes
Parse GKG File URL
ExtractText
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)

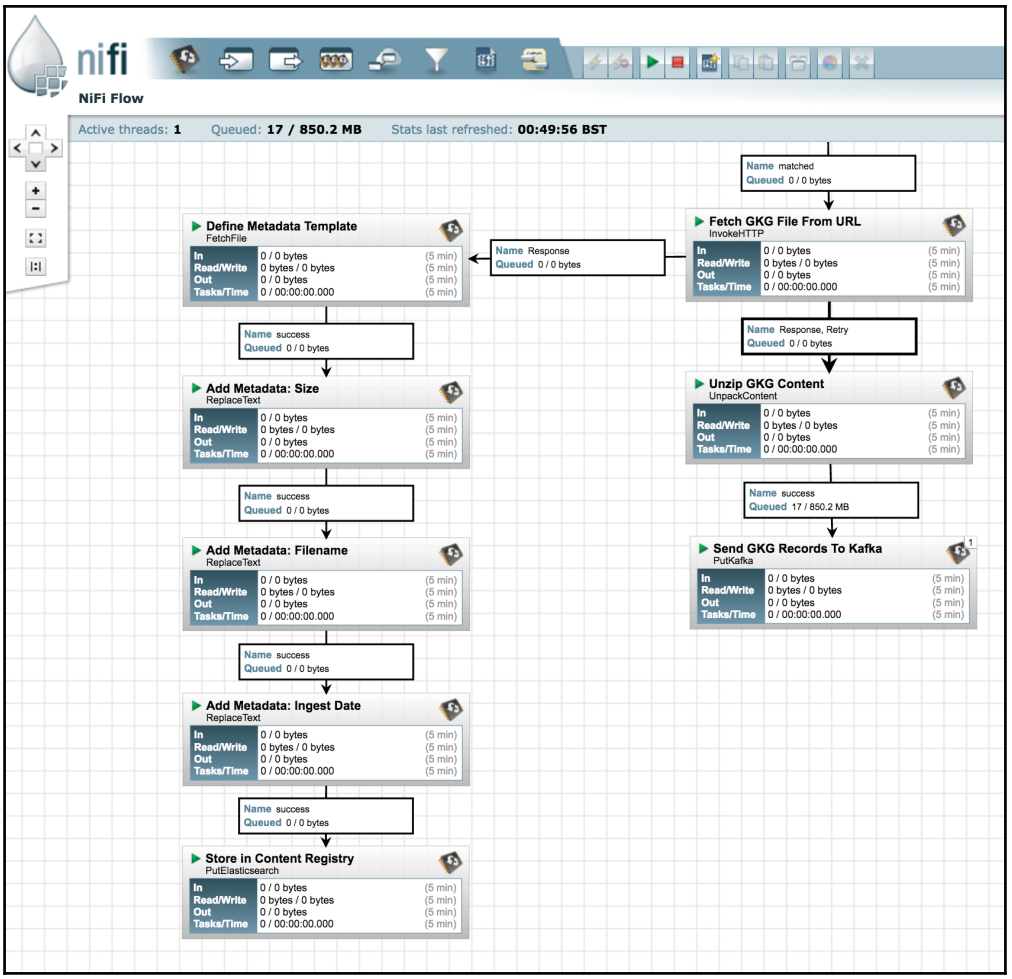

nifi
NiFi Flow
Active threads: 0
Queued: 0 / 0 bytes
Stats last refreshed: 16:48:15 BST
Fetch GDELT Latest File List
GetHTTP
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name success
Queued 0 / 0 bytes
Parse GKG File URL
ExtractText
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Configure Processor
Settings
Scheduling
Properties
Comments
Required field
New property
Property
Value
Enable Canonical Equivalence
false
Enable Case-insensitive Matching
false
Permit Whitespace and Comments in Pattern
false
Enable DOTALL Mode
false
Enable Literal Parsing of the Pattern
false
Enable Multiline Mode
false
Enable Unicode-aware Case Folding
false
Enable Unicode Predefined Character Classes
false
Enable Unix Lines Mode
false
Include Capture Group 0
true
url
([^]*gkg.csv.*)
Cancel
Apply

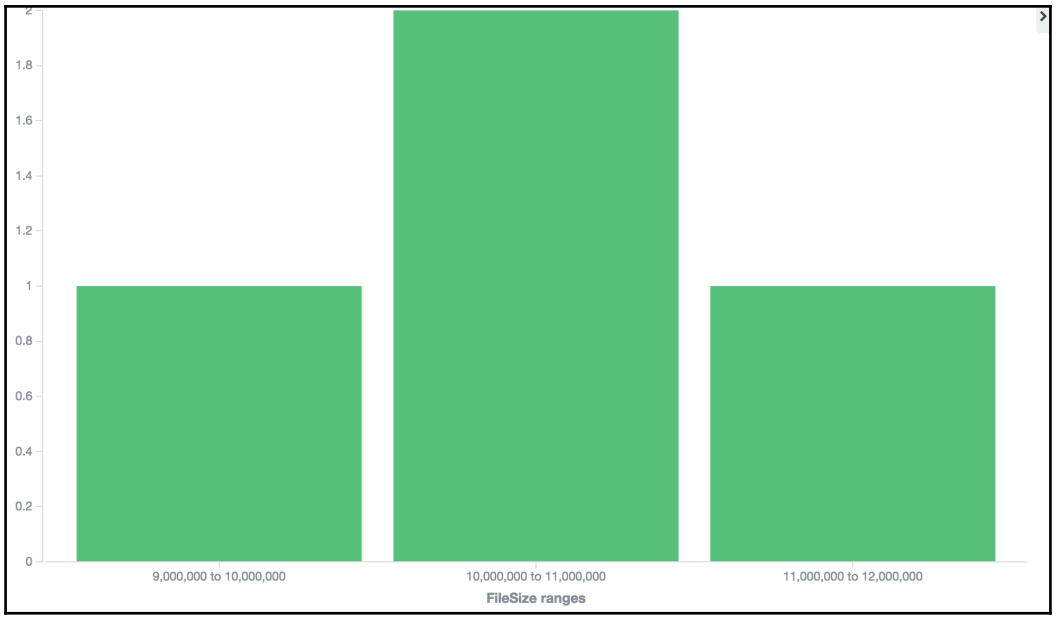

nifi
NiFi Flow
Active threads: 0
Queued: 0 / 0 bytes
Stats last refreshed: 16:49:15 BST
Fetch GDELT Latest File List
GetHTTP
Name success
Queued 0 / 0 bytes
Parse GKG File URL
ExtractText
Name matched
Queued 0 / 0 bytes
Fetch GKG File From URL
InvokeHTTP
Configure Processor
Settings
Scheduling
Properties
Comments
Required field
New property
Property
Value
HTTP Method
GET
Remote URL
${url}
SSL Context Service
No value set
Connection Timeout
5 secs
Read Timeout
15 secs
Include Date Header
True
Follow Redirects
True
Attributes to Send
No value set
Basic Authentication Username
No value set
Basic Authentication Password
No value set
Proxy Host
No value set
Cancel
Apply


NiFi Flow
localhost:8080/nifi/
nifi
NiFi Flow
help
about
Search
Active threads: 1
Queued: 0 / 0 bytes
Stats last refreshed: 18:55:11 BST
Fetch GDELT Latest File List
GetHTTP
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 320 bytes (5 min)
Out 1 / 320 bytes (5 min)
Tasks/Time 66 / 00:00:12.664 (5 min)
Name success
Queued 0 / 0 bytes
Parse GKG File URL
ExtractText
In 1 / 320 bytes (5 min)
Read/Write 320 bytes / 0 bytes (5 min)
Out 1 / 320 bytes (5 min)
Tasks/Time 1 / 00:00:00.001 (5 min)
Name matched
Queued 0 / 0 bytes
Fetch GKG File From URL
InvokeHTTP
In 1 / 320 bytes (5 min)
Read/Write 0 bytes / 8.38 MB (5 min)
Out 1 / 8.38 MB (5 min)
Tasks/Time 1 / 00:00:01.597 (5 min)
Name Response, Retry
Queued 0 / 0 bytes
Unzip GKG Content
UnpackContent
In 1 / 8.38 MB (5 min)
Read/Write 8.38 MB / 26.77 MB (5 min)
Out 1 / 26.77 MB (5 min)
Tasks/Time 1 / 00:00:00.205 (5 min)
Name success
Queued 0 / 0 bytes
Save To HDFS
PutHDFS
In 1 / 26.77 MB (5 min)
Read/Write 26.77 MB / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 1 / 00:00:00.277 (5 min)

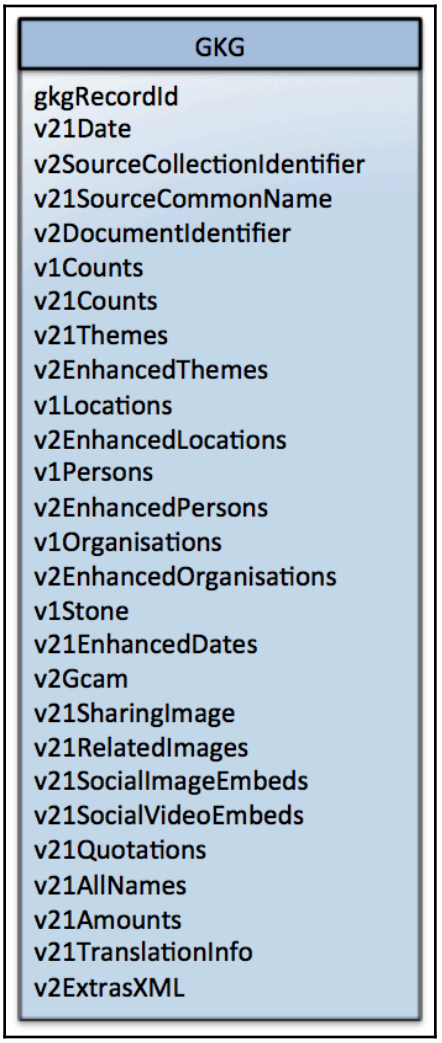

NiFi Flow
localhost:8080/nifi/
help
about
nifi
NiFi Flow
Search
Active threads: 0
Queued: 1 / 24.58 MB
Stats last refreshed: 19:07:51 BST
Fetch GDELT Latest File List
GetHTTP
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 320 bytes (5 min)
Out 1 / 320 bytes (5 min)
Tasks/Time 260 / 00:00:40.556 (5 min)
Name success
Queued 0 / 0 bytes
Parse GKG File URL
ExtractText
In 1 / 320 bytes (5 min)
Read/Write 320 bytes / 0 bytes (5 min)
Out 1 / 320 bytes (5 min)
Tasks/Time 1 / 00:00:00.001 (5 min)
Name matched
Queued 0 / 0 bytes
Fetch GKG File From URL
InvokeHTTP
In 1 / 320 bytes (5 min)
Read/Write 0 bytes / 7.65 MB (5 min)
Out 1 / 7.65 MB (5 min)
Tasks/Time 1 / 00:00:02.453 (5 min)
Name Response, Retry
Queued 0 / 0 bytes
Unzip GKG Content
UnpackContent
In 1 / 7.65 MB (5 min)
Read/Write 7.65 MB / 24.58 MB (5 min)
Out 2 / 98.34 MB (5 min)
Tasks/Time 1 / 00:00:00.174 (5 min)
Name success
Queued 0 / 0 bytes
Name success
Queued 1 / 24.58 MB
Save To HDFS
PutHDFS
In 1 / 24.58 MB (5 min)
Read/Write 24.58 MB / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 1 / 00:00:00.090 (5 min)
To Spark Streaming


NiFi Flow
David
localhost:8080/nifi/
nifi
help
about
Search
NiFi Flow
Active threads: 0
Queued: 0 / 0 bytes
Stats last refreshed: 19:21:39 BST
Fetch GDELT Latest File List
GetHTTP
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 320 bytes (5 min)
Out 1 / 320 bytes (5 min)
Tasks/Time 226 / 00:00:42.594 (5 min)
Name success
Queued 0 / 0 bytes
Parse GKG File URL
ExtractText
In 1 / 320 bytes (5 min)
Read/Write 320 bytes / 0 bytes (5 min)
Out 1 / 320 bytes (5 min)
Tasks/Time 1 / 00:00:00.001 (5 min)
Name matched
Queued 0 / 0 bytes
Fetch GKG File From URL
InvokeHTTP
In 1 / 320 bytes (5 min)
Read/Write 0 bytes / 7.65 MB (5 min)
Out 1 / 7.65 MB (5 min)
Tasks/Time 1 / 00:00:01.620 (5 min)
Name Response, Retry
Queued 0 / 0 bytes
Unzip GKG Content
UnpackContent
In 1 / 7.65 MB (5 min)
Read/Write 7.65 MB / 24.58 MB (5 min)
Out 1 / 24.58 MB (5 min)
Tasks/Time 1 / 00:00:00.180 (5 min)
Name success
Queued 0 / 0 bytes
Send GKG Records To Kafka
PutKafka
In 1 / 24.58 MB (5 min)
Read/Write 24.58 MB / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 1 / 00:00:00.914 (5 min)

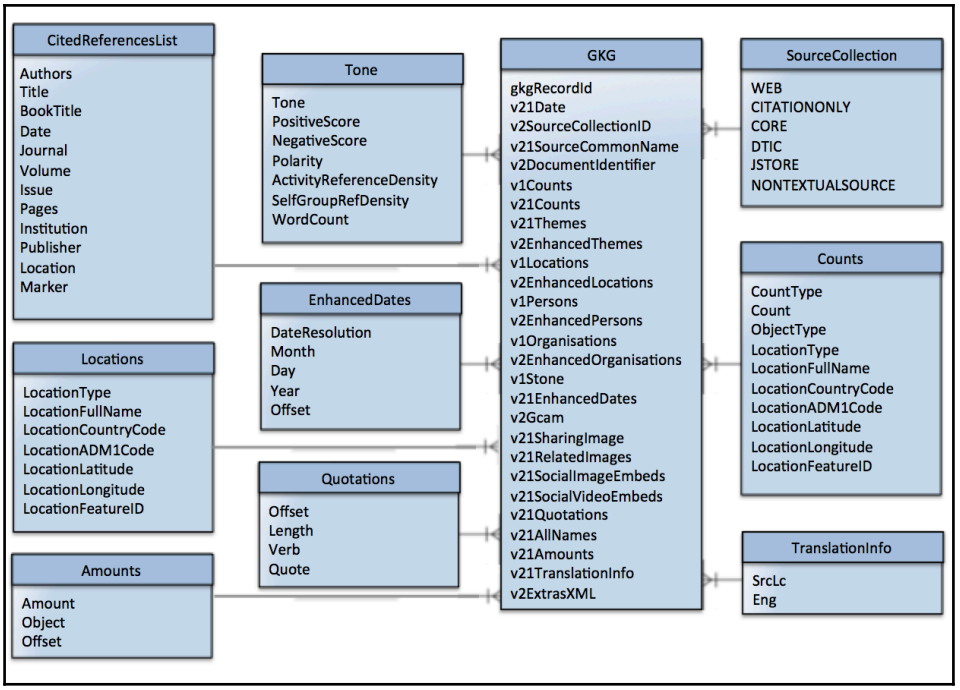

nifi
NiFi Flow
Active threads: 1
Queued: 17 / 850.2 MB
Stats last refreshed: 00:49:56 BST
Name matched
Queued 0 / 0 bytes
Fetch GKG File From URL
InvokeHTTP
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name Response
Queued 0 / 0 bytes
Define Metadata Template
FetchFile
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name Response, Retry
Queued 0 / 0 bytes
Unzip GKG Content
UnpackContent
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name success
Queued 0 / 0 bytes
Add Metadata: Size
ReplaceText
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name success
Queued 17 / 850.2 MB
Send GKG Records To Kafka
PutKafka
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name success
Queued 0 / 0 bytes
Add Metadata: Filename
ReplaceText
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name success
Queued 0 / 0 bytes
Add Metadata: Ingest Date
ReplaceText
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)
Name success
Queued 0 / 0 bytes
Store in Content Registry
PutElasticsearch
In 0 / 0 bytes (5 min)
Read/Write 0 bytes / 0 bytes (5 min)
Out 0 / 0 bytes (5 min)
Tasks/Time 0 / 00:00:00.000 (5 min)

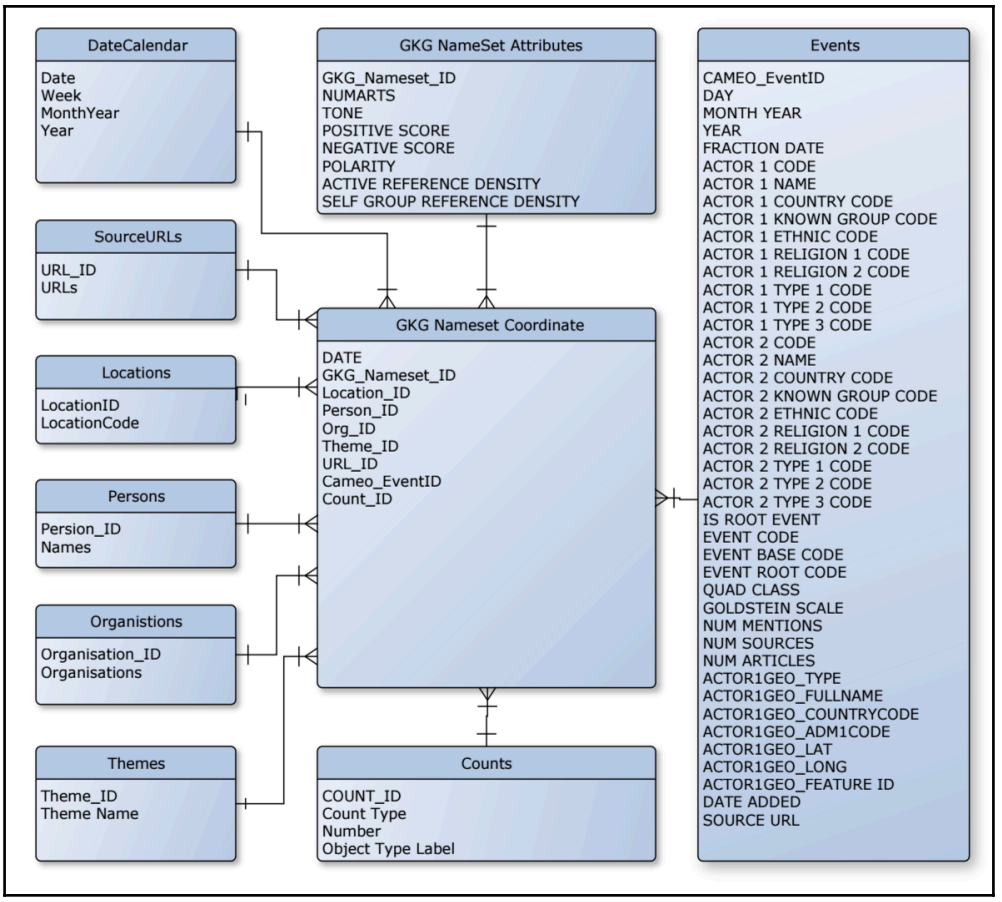

2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
9,000,000 to 10,000,000
10,000,000 to 11,000,000
11,000,000 to 12,000,000
FileSize ranges

Chapter 03: Input Formats and Schema

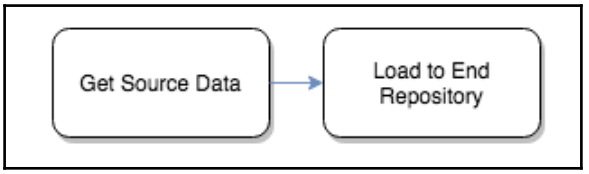


Type	Count
WordCount	125
General Inquirer Bodypt	4
SentiWordNet	40
SentiWordNet average	3.21111111

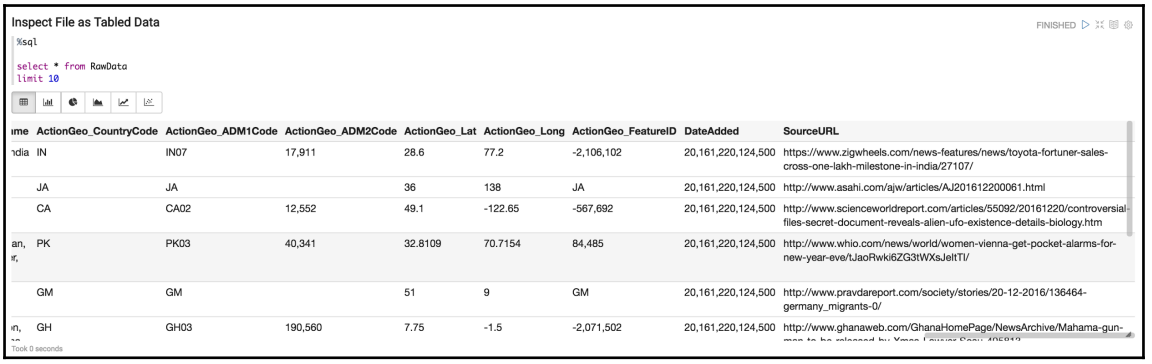


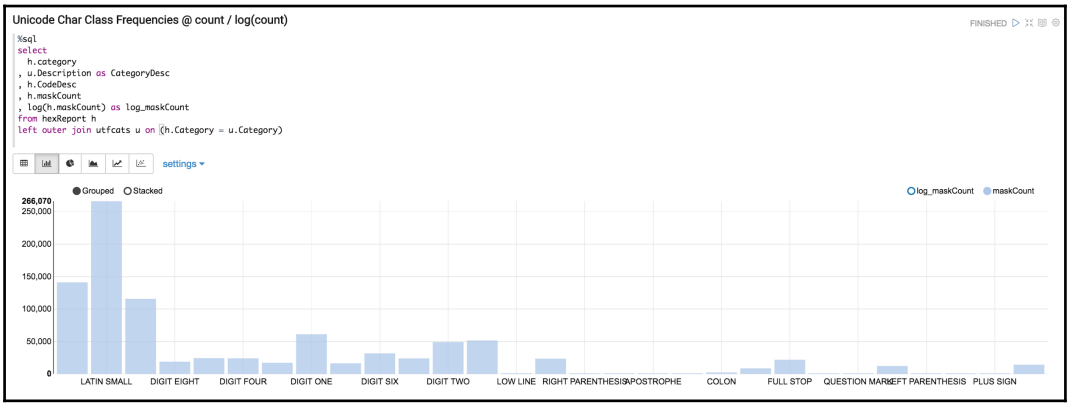

DateCalendar
Date
Week
MonthYear
Year
GKG NameSet Attributes
GKG_Nameset_ID
NUMARTS
TONE
POSITIVE SCORE
NEGATIVE SCORE
POLARITY
ACTIVE REFERENCE DENSITY
SELF GROUP REFERENCE DENSITY
Events
CAMEO_EventID
DAY
MONTH YEAR
YEAR
FRACTION DATE
ACTOR 1 CODE
ACTOR 1 NAME
ACTOR 1 COUNTRY CODE
ACTOR 1 KNOWN GROUP CODE
ACTOR 1 ETHNIC CODE
ACTOR 1 RELIGION 1 CODE
ACTOR 1 RELIGION 2 CODE
ACTOR 1 TYPE 1 CODE
ACTOR 1 TYPE 2 CODE
ACTOR 1 TYPE 3 CODE
ACTOR 2 CODE
ACTOR 2 NAME
ACTOR 2 COUNTRY CODE
ACTOR 2 KNOWN GROUP CODE
ACTOR 2 ETHNIC CODE
ACTOR 2 RELIGION 1 CODE
ACTOR 2 RELIGION 2 CODE
ACTOR 2 TYPE 1 CODE
ACTOR 2 TYPE 2 CODE
ACTOR 2 TYPE 3 CODE
IS ROOT EVENT
EVENT CODE
EVENT BASE CODE
EVENT ROOT CODE
QUAD CLASS
GOLDSTEIN SCALE
NUM MENTIONS
NUM SOURCES
NUM ARTICLES
ACTOR1GEO_TYPE
ACTOR1GEO_FULLNAME
ACTOR1GEO_COUNTRYCODE
ACTOR1GEO_ADM1CODE
ACTOR1GEO_LAT
ACTOR1GEO_LONG
ACTOR1GEO_FEATURE ID
DATE ADDED
SOURCE URL
SourceURLs
URL_ID
URLs
GKG Nameset Coordinate
DATE
GKG_Nameset_ID
Location_ID
Person_ID
Org_ID
Theme_ID
URL_ID
Cameo_EventID
Count_ID
Locations
LocationID
LocationCode
Persons
Persion_ID
Names
Organistions
Organisation_ID
Organisations
Themes
Theme_ID
Theme Name
Counts
COUNT_ID
Count Type
Number
Object Type Label


Get Source Data
Extract the Useful Bits
Transform Those Bits
Load to End Repository


Get Source Data
Load to End Repository

Chapter 04: Exploratory Data Analysis

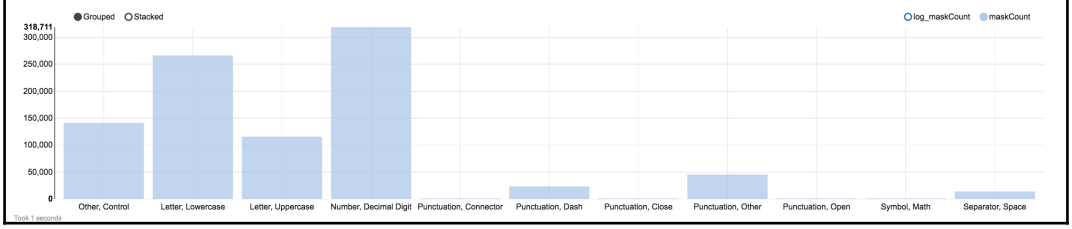




Unicode Category Frequencies @ count / log(count)

FINISHED

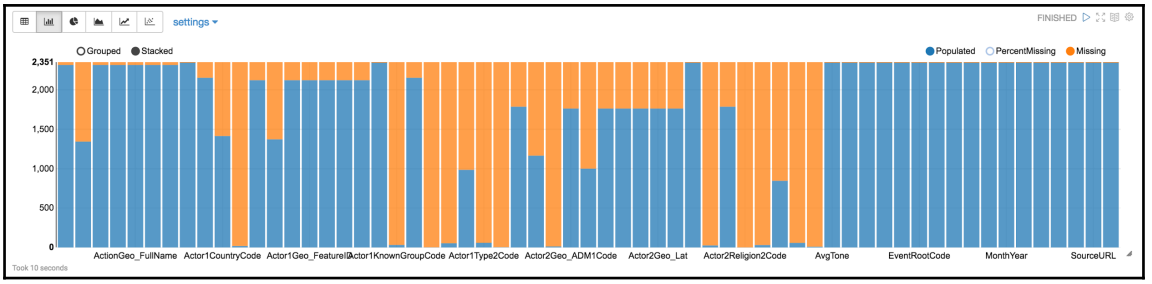
```
%sql
select
  h.category
, u.Description as CategoryDesc
, h.CodeDesc
, h.maskCount
, log(h.maskCount) as log_maskCount
from hexReport h
left outer join utfcats u on (h.Category = u.Category)
```



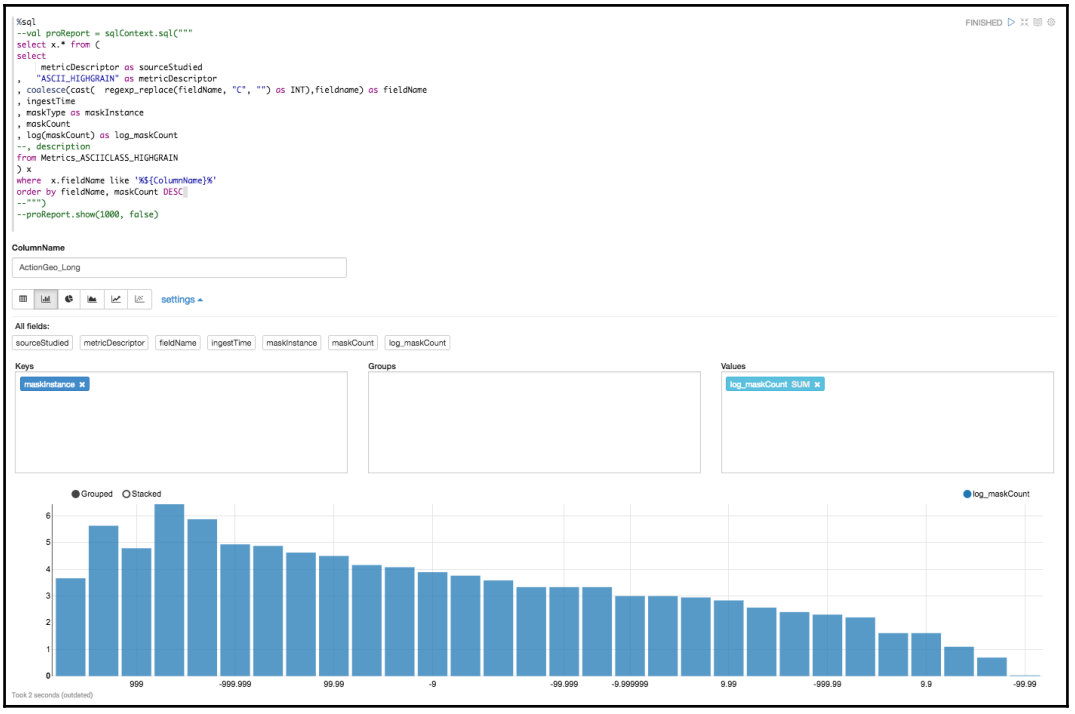

```
Metrics_POPCHECKS.toDF.show(1000, false)
```

metricDescriptor	sourceStudied	ingestTime	maskType	fieldName	occurrenceCount	keyCount	maskCount	description
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	Actor1Type2Code	100525	61	59	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	Actor1Geo_Lat	100525	61	2122	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	Actor2Type2Code	100525	61	58	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	NumSources	100525	61	2351	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	ActionGeo_Long	100525	61	2315	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	Actor1Geo_FullName	100525	61	2122	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	AvgTone	100525	61	2351	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	Actor1Geo_Type	100525	61	2351	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	ActionGeo_ADM2Code	100525	61	1343	<function1>
/user/feeds/gdelt/events/*.export.CSV		2016-12-23	0	DateAdded	100525	61	2351	<function1>

fieldname	Missing	Populated	PercentMissing	fileName
ActionGeo_ADM1Code	36	2315	1.53	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_ADM2Code	1008	1343	42.88	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_CountryCode	36	2315	1.53	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_FeatureID	36	2315	1.53	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_FullName	36	2315	1.53	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_Lat	36	2315	1.53	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_Long	36	2315	1.53	/user/feeds/gdelt/events/*.export.CSV
ActionGeo_Type	0	2351	0.0	/user/feeds/gdelt/events/*.export.CSV
Actor1Code	198	2153	8.42	/user/feeds/gdelt/events/*.export.CSV
Actor1CountryCode	937	1414	39.86	/user/feeds/gdelt/events/*.export.CSV
Actor1EthnicCode	2332	19	99.19	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_ADM1Code	229	2122	9.74	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_ADM2Code	979	1372	41.64	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_CountryCode	229	2122	9.74	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_FeatureID	229	2122	9.74	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_FullName	229	2122	9.74	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_Lat	229	2122	9.74	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_Long	229	2122	9.74	/user/feeds/gdelt/events/*.export.CSV
Actor1Geo_Type	0	2351	0.0	/user/feeds/gdelt/events/*.export.CSV




```
proReport: org.apache.spark.sql.DataFrame = [sourceStudied: string, metricDescriptor: string, fieldName: string, ingestTime: date, maskInstance: string, maskCount: bigint, description: st
+--------------------------+----------+--------------------+----------+------------+---------+
|sourceStudied             |metricDescriptor|fieldName           |ingestTime|maskInstance|maskCount|
+--------------------------+----------+--------------------+----------+------------+---------+
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_ADM1Code  |2016-12-23|A9          |1207     |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_ADM1Code  |2016-12-23|A           |1108     |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_ADM1Code  |2016-12-23|            |36       |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_ADM2Code  |2016-12-23|9           |1204     |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_ADM2Code  |2016-12-23|            |1008     |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_ADM2Code  |2016-12-23|A9          |139      |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_CountryCode|2016-12-23|A           |2315     |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_CountryCode|2016-12-23|            |36       |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_FeatureID |2016-12-23|-9          |1108     |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_FeatureID |2016-12-23|A           |810      |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_FeatureID |2016-12-23|9           |397      |
|/user/feeds/gdelt/events/*.export.CSV|ASCII_LOWGRAIN  |ActionGeo_FeatureID |2016-12-23|            |36       |
```



```
%sql
Select
case maskType
    when '0' then 'Populated'
    when '1' then 'Missing'
end as Status
, *
from PopChecks
--where fieldName like '%Themes%' or fieldName like '%Persons%' or fieldNAme like '%Loc%'
```

FINISHED

settings

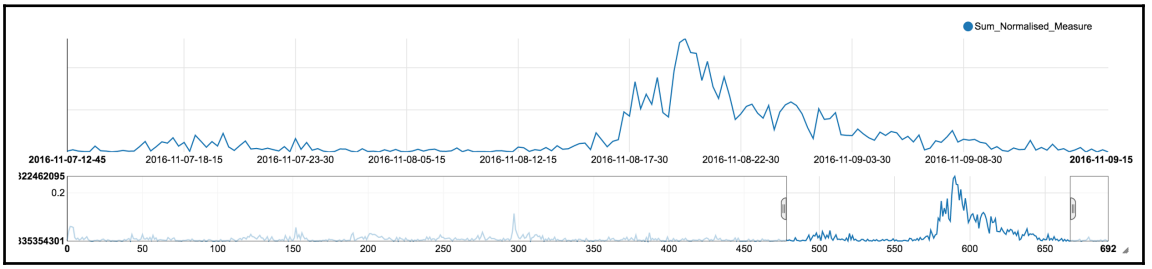
Stacked Stream Expanded

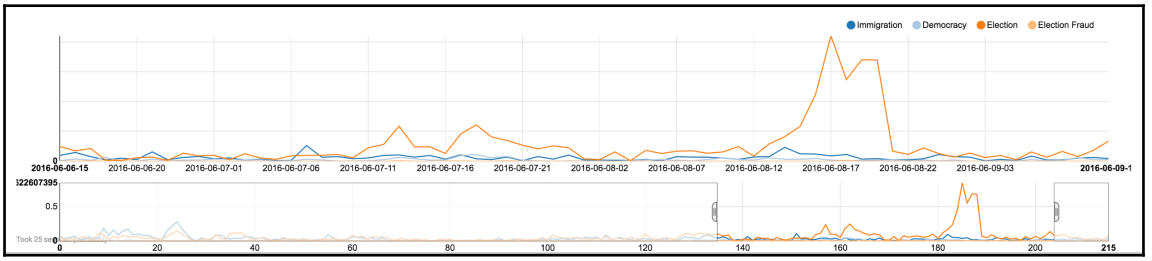
Gkg English Files.Populated.maskCount | Gkg English Files.Missing.maskCount | Gkg Tranlsated Files.Populated.maskCount | Gkg Tranlsated Files.Missing.maskCount

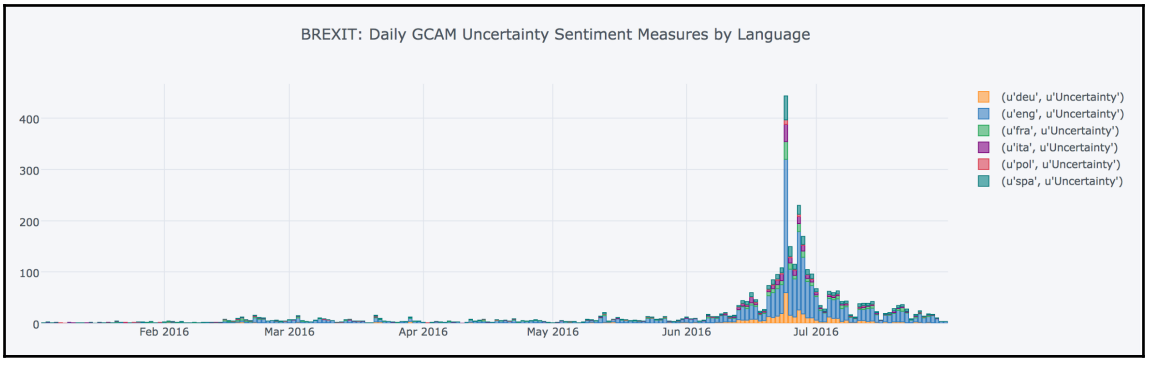
100% 80% 60% 40% 20% 0%

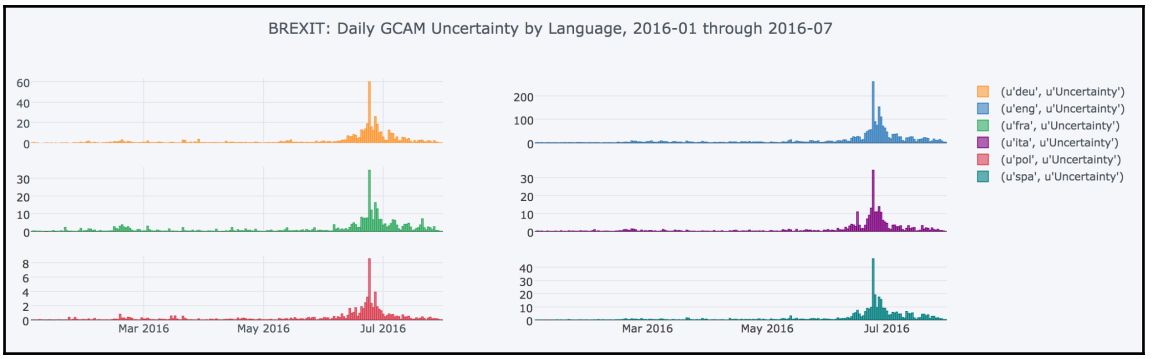
V1Counts V2EnhancedThemes V1Persons V21AllNames V2EnhancedPersons V1Locations GkgRecordId V1OrganisazationsV21SocialVideoEmbeds V21SocialImageEmbeds V2EnhancedLocations V21Amounts V21Quotations V2ExtrasXML

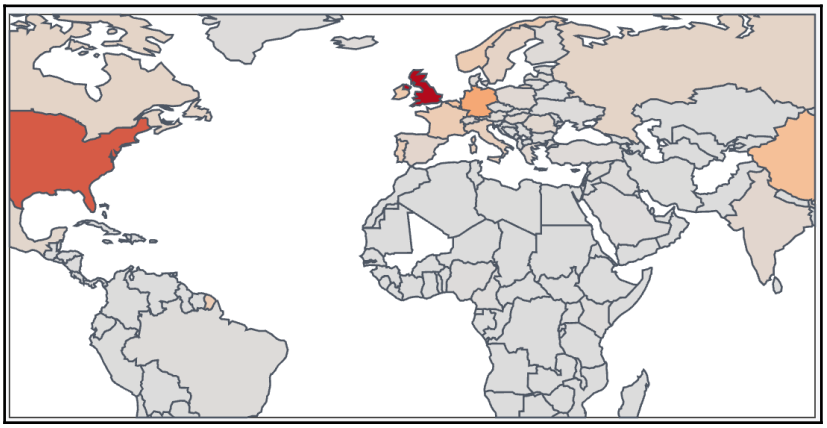
Took 0 seconds

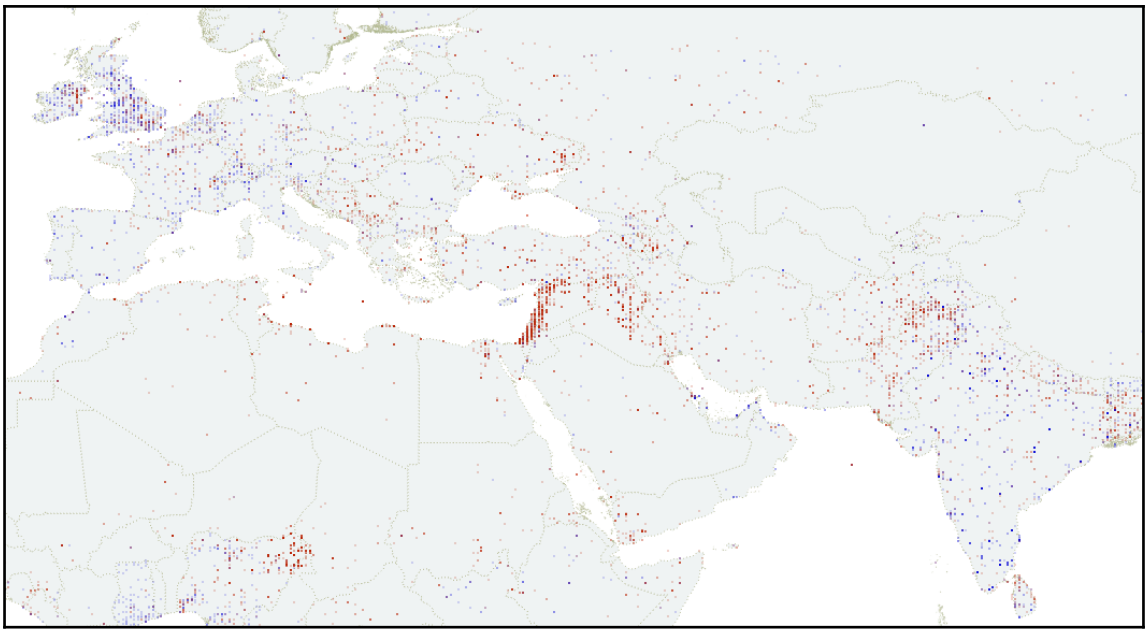






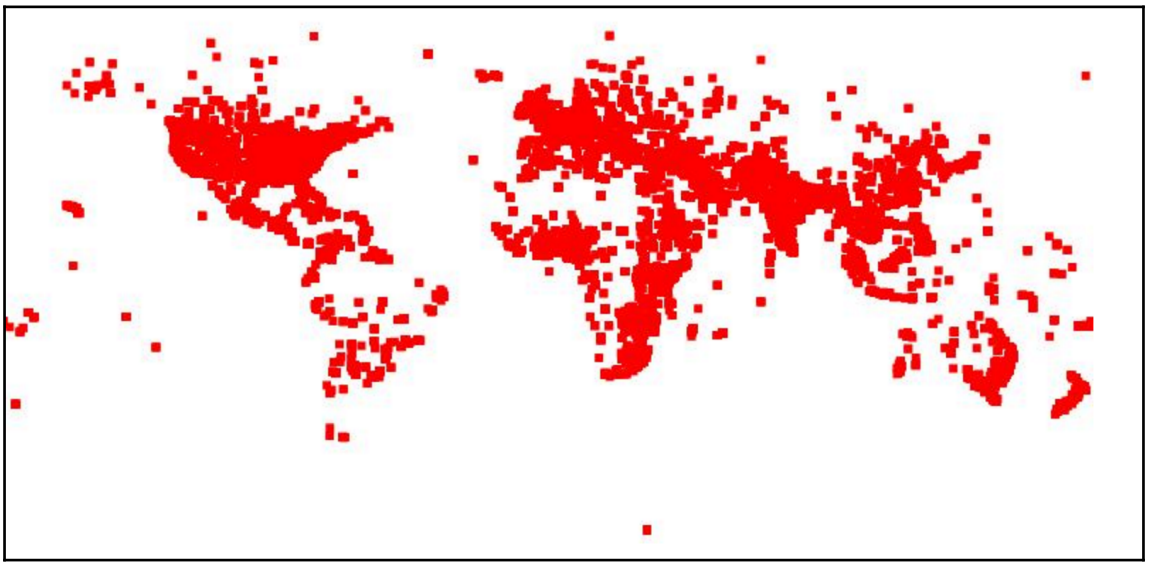

BREXIT: Daily GCAM Uncertainty by Language, 2016-01 through 2016-07
60
40
20
0
30
20
10
0
8
6
4
2
0
Mar 2016
May 2016
Jul 2016
200
100
0
30
20
10
0
40
30
20
10
0
Mar 2016
May 2016
Jul 2016
(u'deu', u'Uncertainty')
(u'eng', u'Uncertainty')
(u'fra', u'Uncertainty')
(u'ita', u'Uncertainty')
(u'pol', u'Uncertainty')
(u'spa', u'Uncertainty')

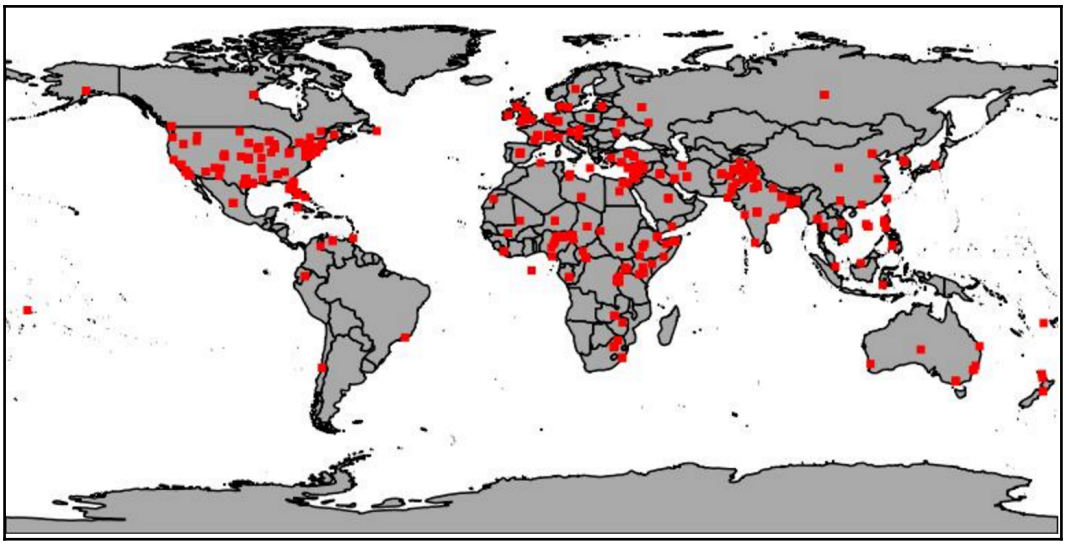




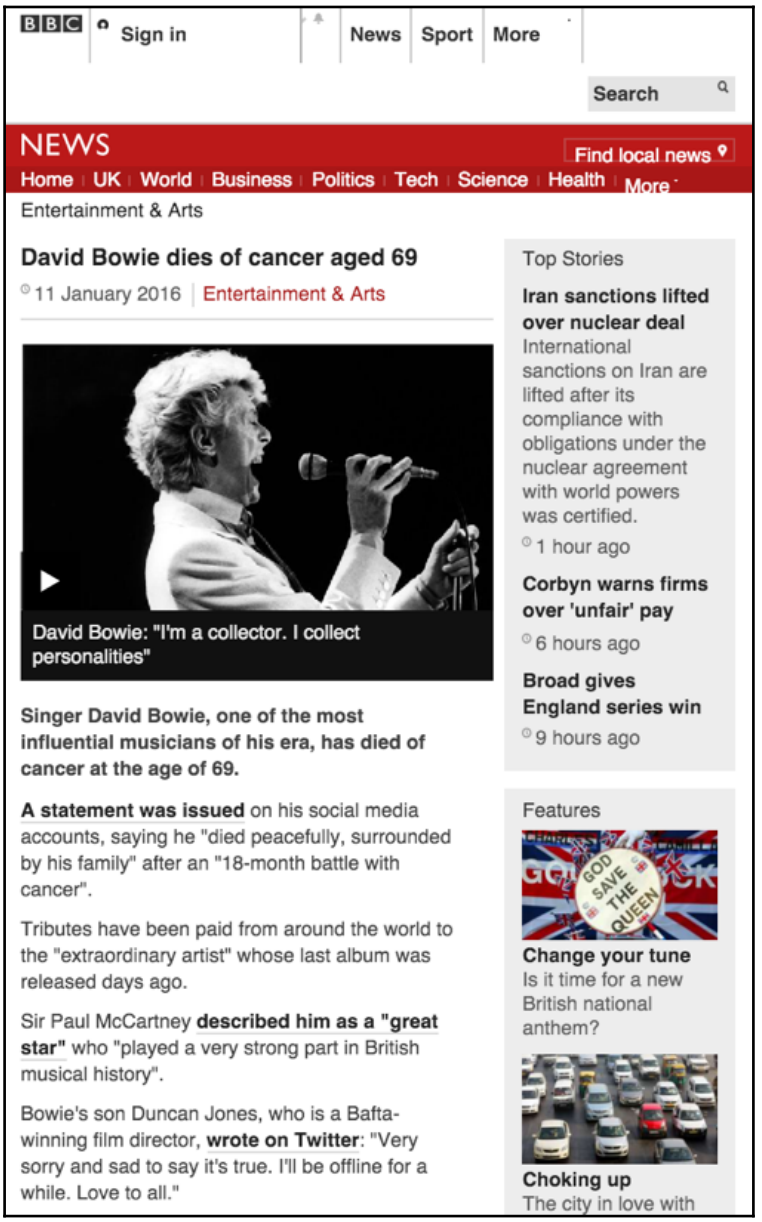
Chapter 05: Spark for Geographic Analysis

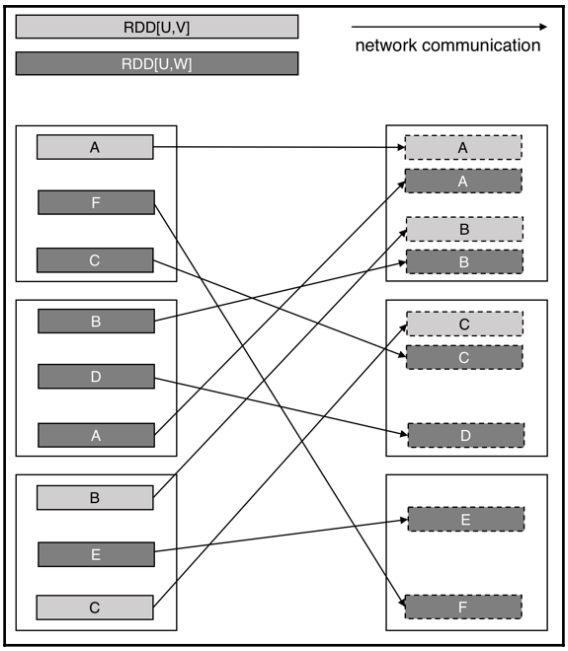
$$2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

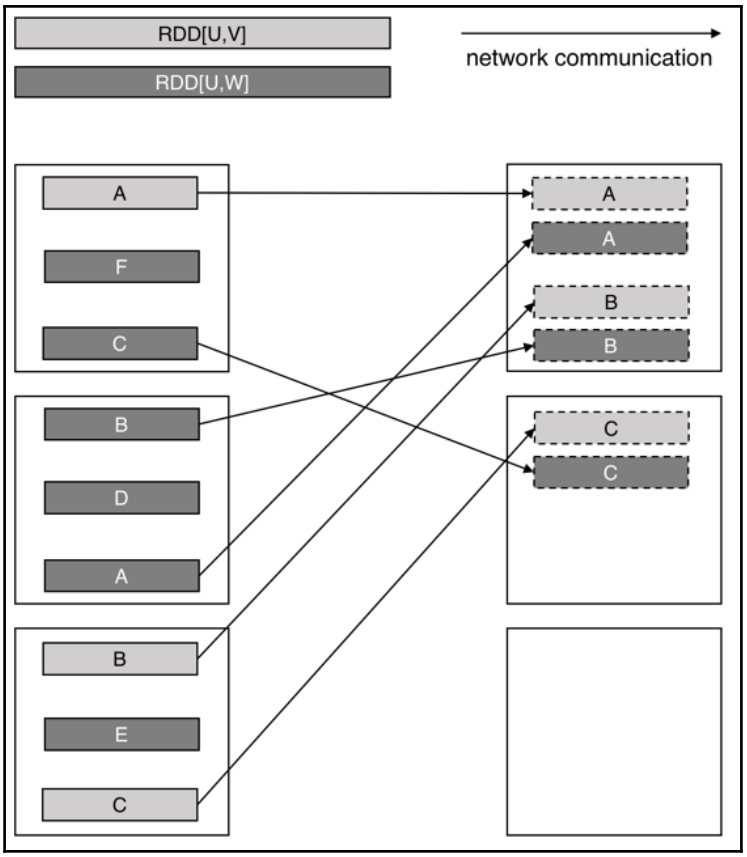


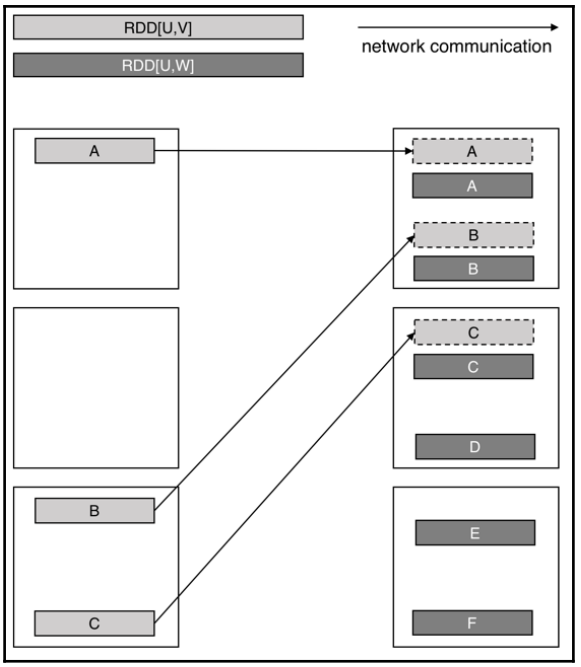


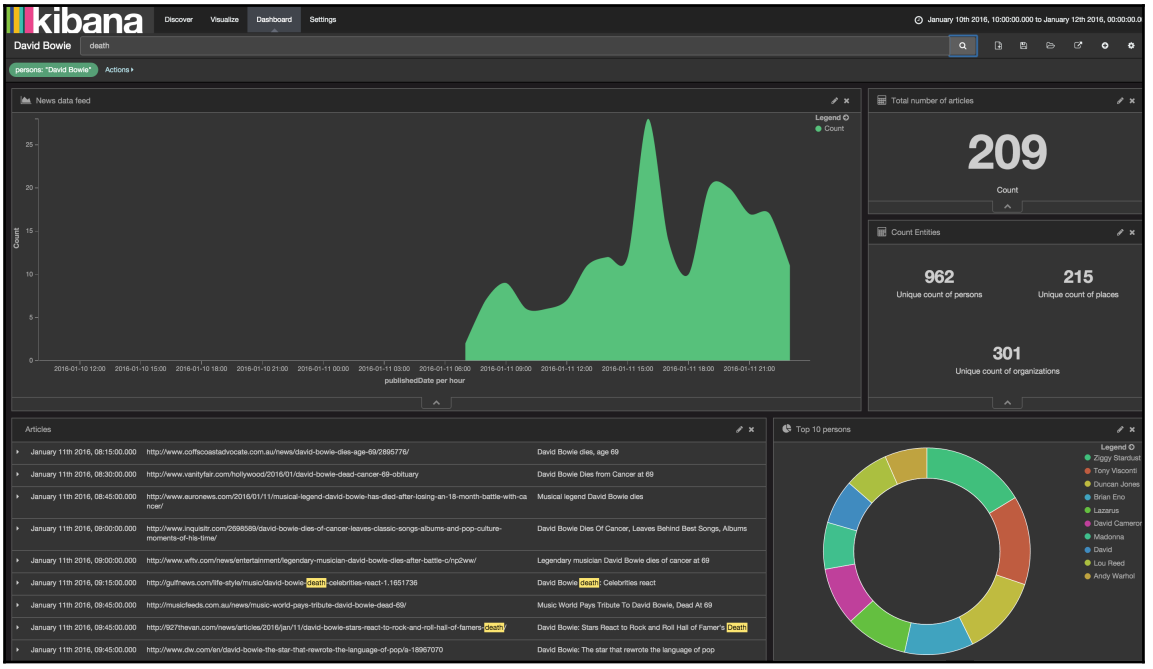
Chapter 06:Scraping Link-Based External Data




RDD[U,V]
RDD[U,W]
network communication
A
F
C
B
D
A
B
E
C
A
A
B
B
C
C
D
E
F

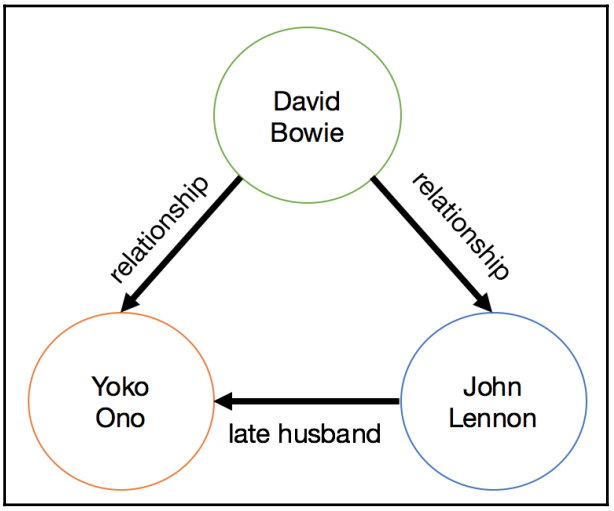

RDD[U,V]
RDD[U,W]
network communication
A
F
C
B
D
A
B
E
C
A
A
B
B
C
C


RDD[U,V]
RDD[U,W]
network communication
A
A
A
B
B
B
C
C
C
D
E
F

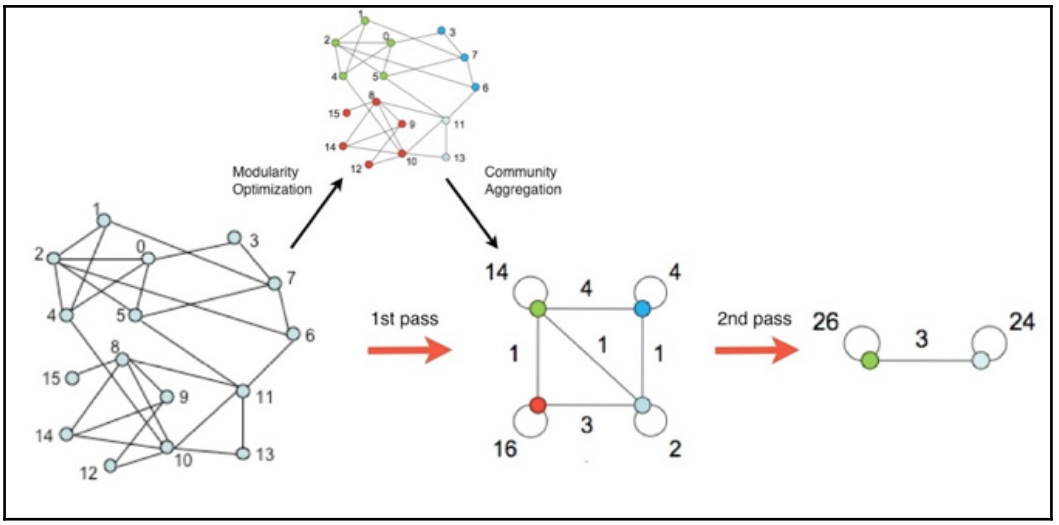

kibana
Discover
Visualize
Dashboard
Settings
January 10th 2016, 10:00:00.000 to January 12th 2016, 00:00:00.0
David Bowie
death
persons: "David Bowie"
Actions
News data feed
Legend
Count
publishedDate per hour
Total number of articles
209
Count
Count Entities
962
Unique count of persons
215
Unique count of places
301
Unique count of organizations
Articles
January 11th 2016, 08:15:00.000 http://www.coffscoastadvocate.com.au/news/david-bowie-dies-age-69/2895776/ David Bowie dies, age 69
January 11th 2016, 08:30:00.000 http://www.vanityfair.com/hollywood/2016/01/david-bowie-dead-cancer-69-obituary David Bowie Dies from Cancer at 69
January 11th 2016, 08:45:00.000 http://www.euronews.com/2016/01/11/musical-legend-david-bowie-has-died-after-losing-an-18-month-battle-with-cancer/ Musical legend David Bowie dies
January 11th 2016, 09:00:00.000 http://www.inquisitr.com/2698589/david-bowie-dies-of-cancer-leaves-classic-songs-albums-and-pop-culture-moments-of-his-time/ David Bowie Dies Of Cancer, Leaves Behind Best Songs, Albums
January 11th 2016, 09:00:00.000 http://www.wftv.com/news/entertainment/legendary-musician-david-bowie-dies-after-battle-c/np2ww/ Legendary musician David Bowie dies of cancer at 69
January 11th 2016, 09:15:00.000 http://gulfnews.com/life-style/music/david-bowie-death-celebrities-react-1.1651736 David Bowie death Celebrities react
January 11th 2016, 09:45:00.000 http://musicfeeds.com.au/news/music-world-pays-tribute-david-bowie-dead-69/ Music World Pays Tribute To David Bowie, Dead At 69
January 11th 2016, 09:45:00.000 http://927thevan.com/news/articles/2016/jan/11/david-bowie-stars-react-to-rock-and-roll-hall-of-famers-death/ David Bowie: Stars React to Rock and Roll Hall of Famer's Death
January 11th 2016, 09:45:00.000 http://www.dw.com/en/david-bowie-the-star-that-rewrote-the-language-of-pop/a-18967070 David Bowie: The star that rewrote the language of pop
Top 10 persons
Legend
Ziggy Stardust
Tony Visconti
Duncan Jones
Brian Eno
Lazarus
David Cameron
Madonna
David
Lou Reed
Andy Warhol


Gianfranco Ravasi
@CardRavasi
Follow
Ground Control to Major Tom
Commencing countdown,
engines on
Check ignition
and may God's love be with you (David Bowie)
8:35 AM - 11 Jan 2016
4,226
4,146

Chapter 07: Building Communities



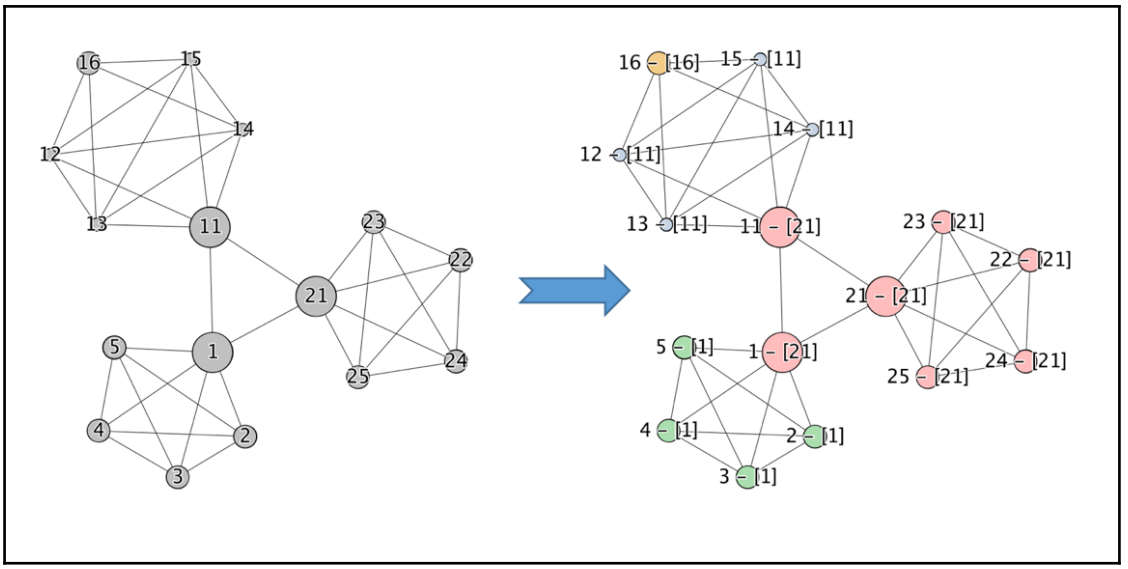
row key	column families					
	relationA			relationB		
	qualifier	value	visibility	qualifier	value	visibility
personA	personB	1	INTERNAL	personD	1	SECRET
	personC	1	CONFIDENTIAL			
personB	personC	1	INTERNAL	personD	1	CONFIDENTIAL

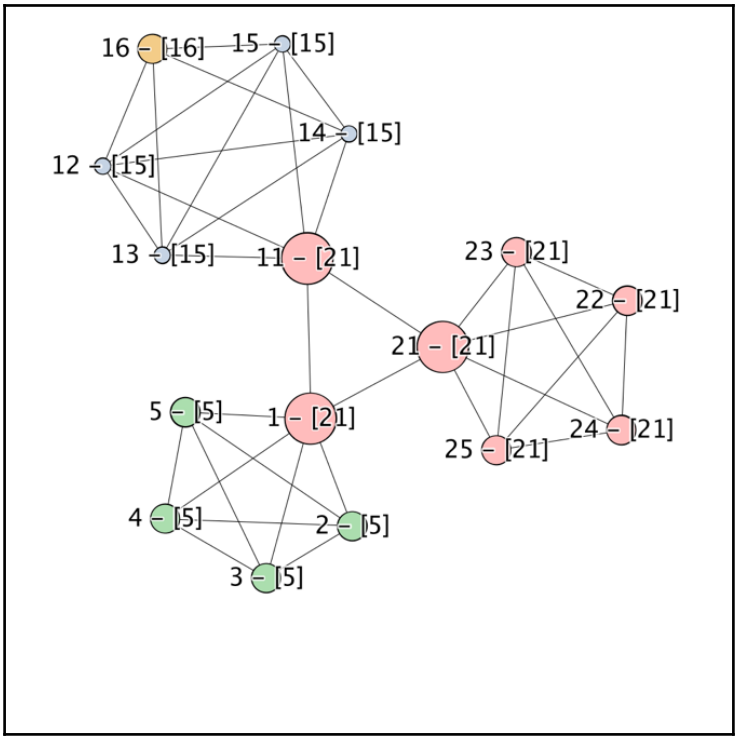


$$WCC(x,C) = \begin{cases} \frac{t(x,C)}{t(x,V)} \cdot \frac{vt(x,V)}{|C\setminus\{x\}|+vt(x,V\setminus C)} & \text{if } t(x,V) \neq 0; \\ 0 & \text{if } t(x,V) = 0. \end{cases}$$

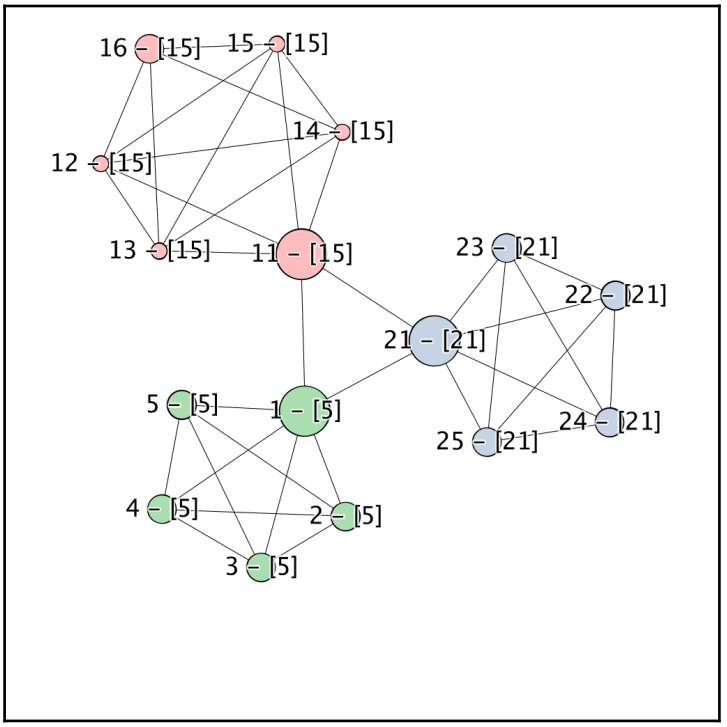
$$WCC(S) = \frac{1}{|C|} \sum_{x \in S} WCC(x,C).$$

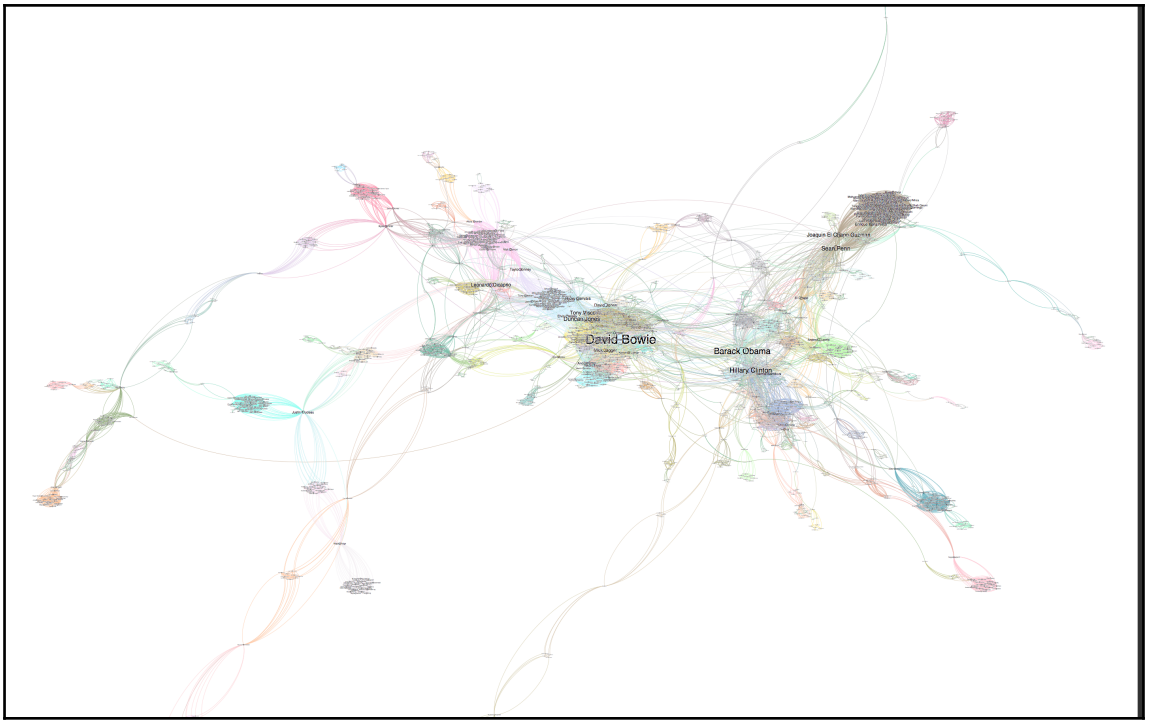
$$WCC(\mathcal{P}) = \frac{1}{|V|} \sum_{i=1}^{n} \left(|C_i| \cdot WCC(C_i) \right).$$

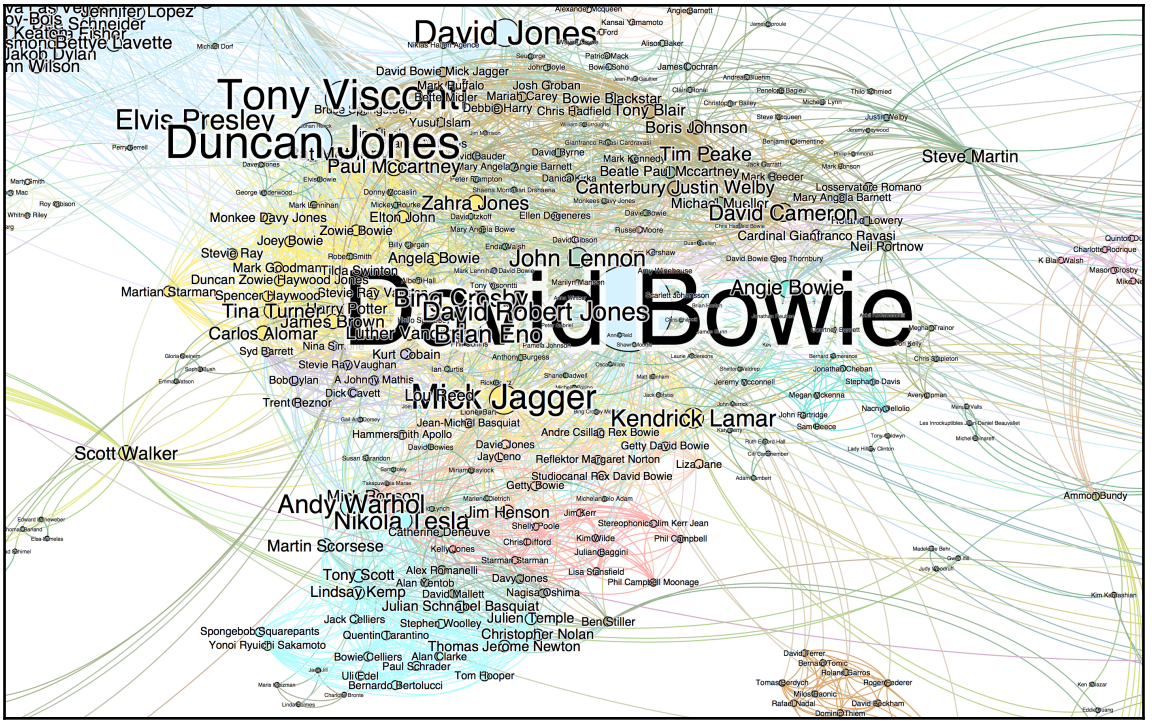

16
15
14
12
13
11
23
22
21
5
1
24
25
4
2
3
16 – [16]
15 – [11]
14 – [11]
12 – [11]
13 – [11]
11 – [21]
23 – [21]
22 – [21]
21 – [21]
5 – [1]
1 – [21]
25 – [21]
24 – [21]
4 – [1]
2 – [1]
3 – [1]

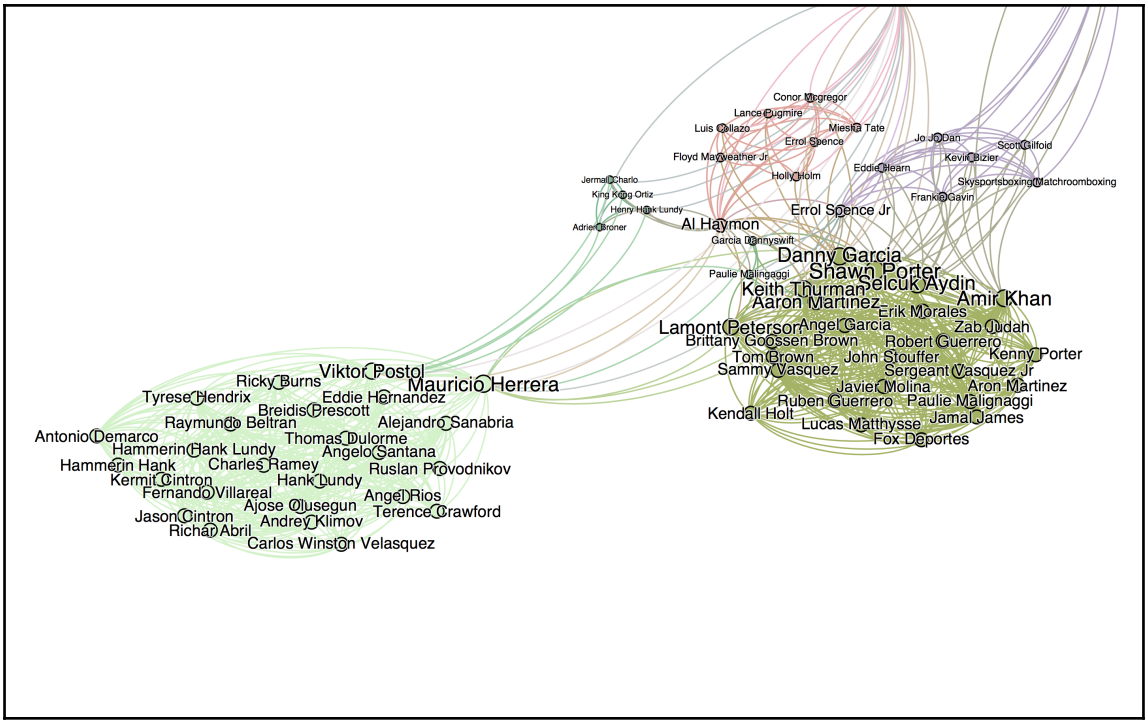

16 – [16]
15 – [15]
14 – [15]
12 – [15]
13 – [15]
11 – [21]
23 – [21]
22 – [21]
21 – [21]
5 – [5]
1 – [21]
24 – [21]
25 – [21]
4 – [5]
2 – [5]
3 – [5]

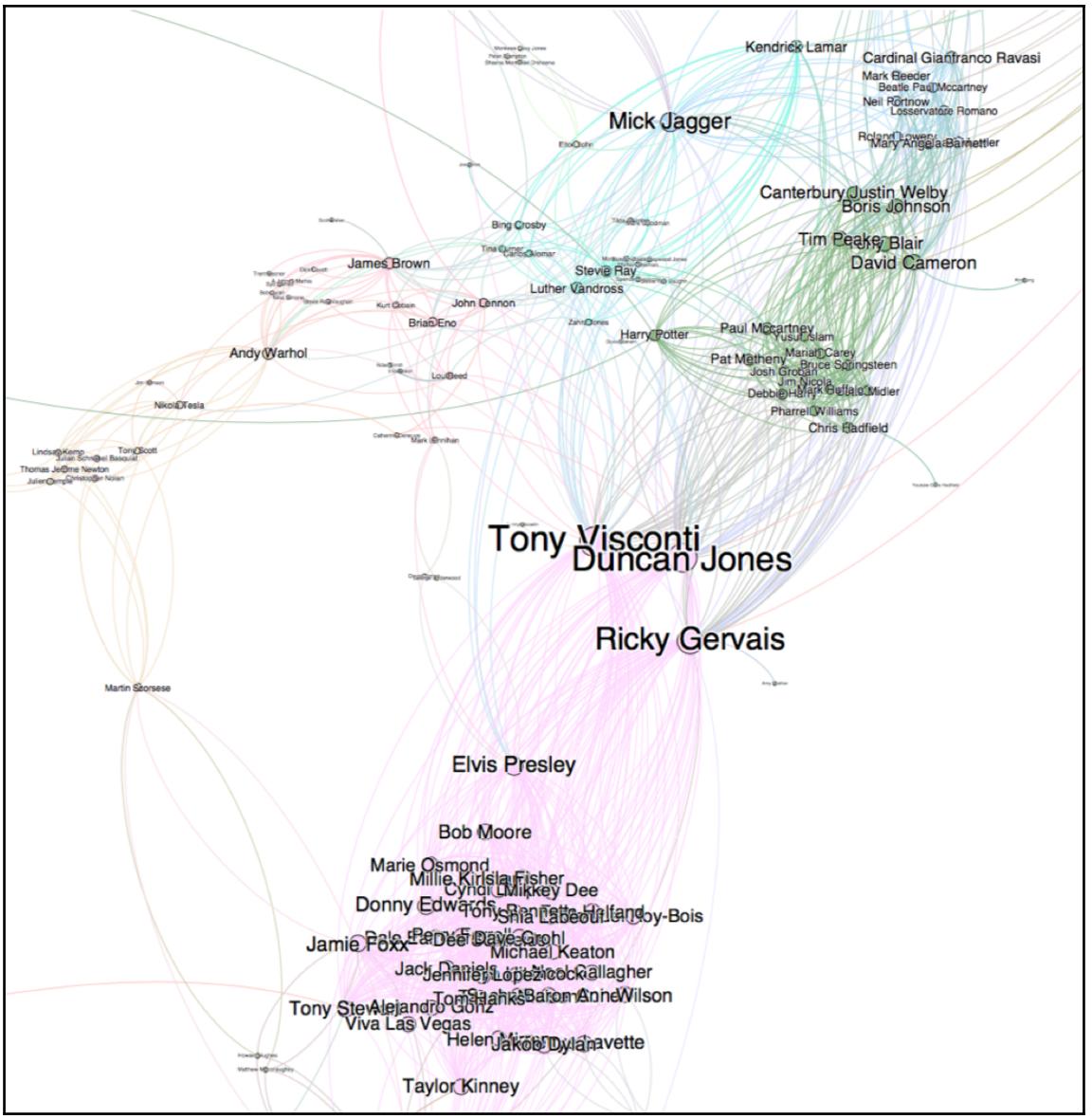
$$WCC(P') - WCC(P) = WCC'_I(v, C)$$
$$= \frac{1}{V} \cdot (d_{in} \cdot \Theta_1 + (r - d_{in}) \cdot \Theta_2 + \Theta_3)$$



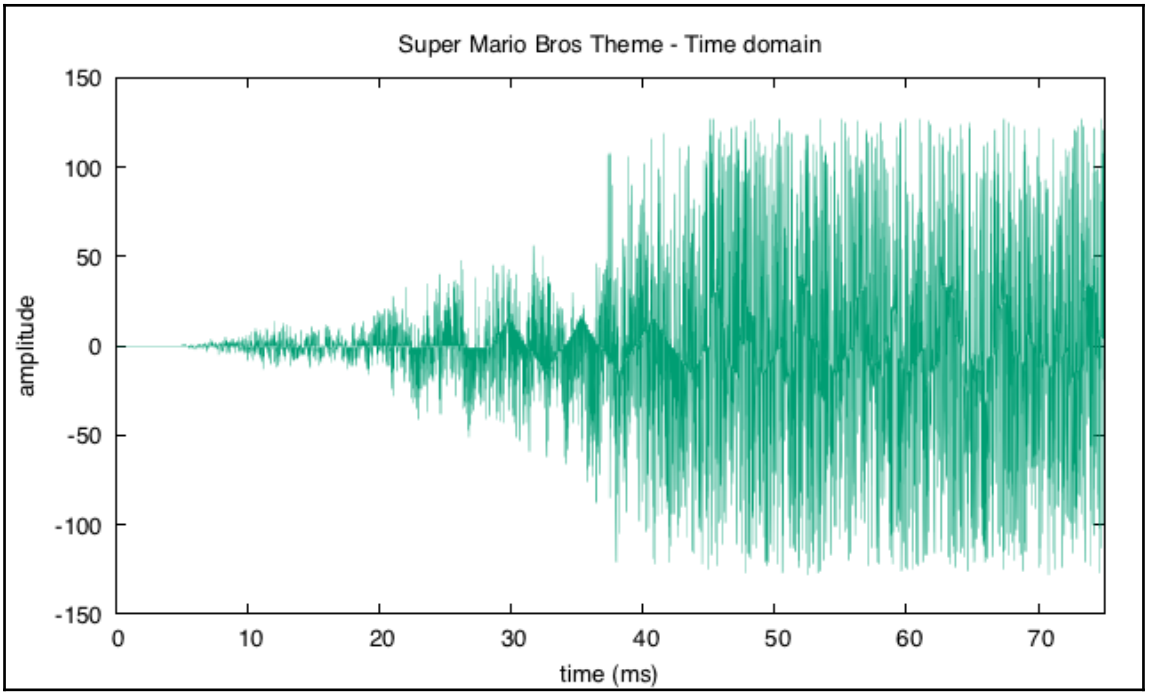

David Bowie
Barack Obama
Hillary Clinton
Joaquin S. Quinn-Guzman
SeanPenn
Leonard DiCaprio
Tom Hanks
Duncan Jones
John Rhodes


David Jones
Tony Visconti
Duncan Jones
Elvis Presley
David Bowie
Mick Jagger
Paul Mccartney
Zahra Jones
Angela Bowie
John Lennon
Tina Turner
David Robert Jones
Brian Eno
Carlos Alomar
Mick Jagger
Kendrick Lamar
Scott Walker
Andy Warhol
Nikola Tesla
Martin Scorsese
Tony Scott
Lindsay Kemp
Julian Schnabel Basquiat
Thomas Jerome Newton
Tim Peake
Steve Martin
Boris Johnson
Tony Blair
Bowie Blackstar
Canterbury Justin Welby
David Cameron
Cardinal Giantranco Ravasi
Neil Portnow
Angie Bowie
Ammon Bundy
Steve Martin

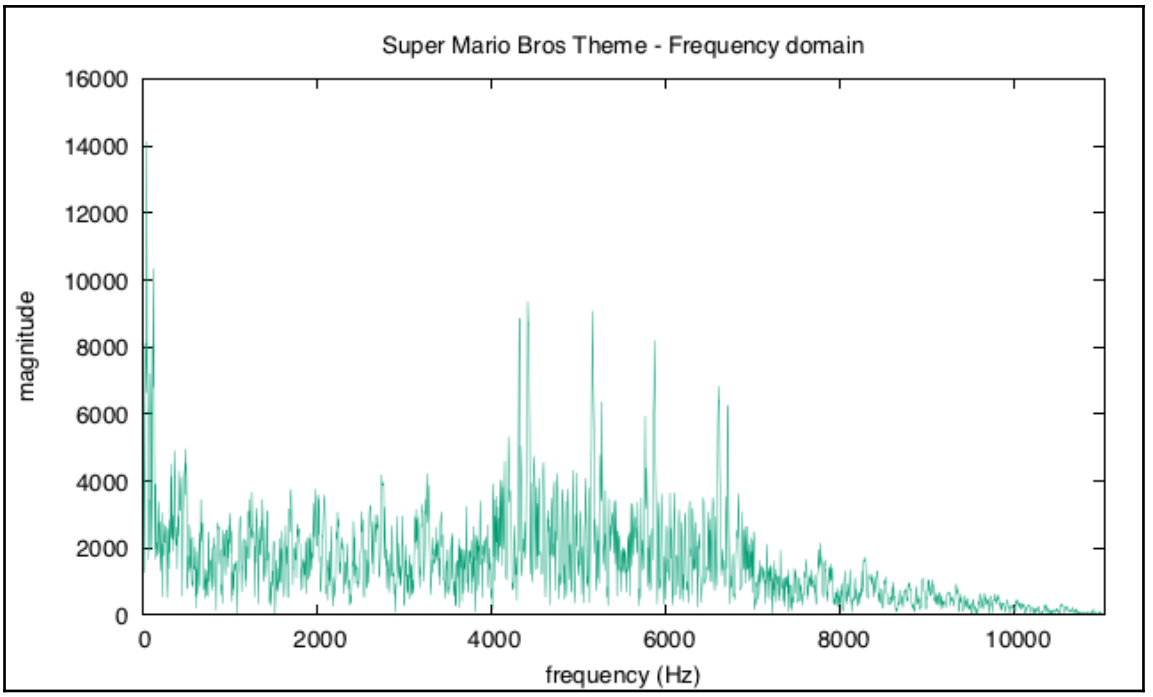

Conor Mcgregor
Lance Cugmire
Luis Collazo
Miesha Tate
Errol Spence
Jo Jo Dan
Scott Gilfoid
Floyd Mayweather Jr
Kevin Bizier
Holly Holm
Eddie Hearn
Jermall Charlo
Skysportsboxing
Matchroomboxing
King Kong Ortiz
Frankie Gavin
Henry Hook Lundy
Errol Spence Jr
Adrien Broner
Al Haymon
Garcia Donnyswift
Danny Garcia
Paulie Malignaggi
Shawn Porter
Keith Thurman
Selcuk Aydin
Aaron Martinez
Amir Khan
Erik Morales
Lamont Peterson
Angel Garcia
Zab Judah
Brittany Goossen Brown
Robert Guerrero
Tom Brown
John Stouffer
Kenny Porter
Sammy Vasquez
Sergeant Vasquez Jr
Javier Molina
Aron Martinez
Ruben Guerrero
Paulie Malignaggi
Kendall Holt
Jamal James
Lucas Matthysse
Fox Deportes
Viktor Postol
Ricky Burns
Mauricio Herrera
Tyrese Hendrix
Eddie Hernandez
Breidis Prescott
Raymundo Beltran
Alejandro Sanabria
Antonio Demarco
Thomas Dulorme
Hammerin Hank Lundy
Angelo Santana
Hammerin Hank
Charles Ramey
Ruslan Provodnikov
Kermit Cintron
Hank Lundy
Fernando Villareal
Angel Rios
Ajose Olusegun
Jason Cintron
Terence Crawford
Andrey Klimov
Richar Abril
Carlos Winston Velasquez


Kendrick Lamar
Cardinal Gianfranco Ravasi
Mark Beeder
Beatle Paul Mccartney
Neil Portnow
Losservatore Romano
Mick Jagger
Mary Angela Barnettler
Canterbury Justin Welby
Boris Johnson
Bing Crosby
Tim Peake
Tony Blair
David Cameron
James Brown
Stevie Ray
Luther Vandross
John Cannon
Brian Eno
Harry Potter
Paul Mccartney
Andy Warhol
Mariah Carey
Pat Metheny
Bryce Springsteen
Josh Groban
Jim Nicola
Debbie Harry
Mark Gellard Midler
Pharrel Williams
Chris Hadfield
Nikola Tesla
Lou Reed
Tony Scott
Thomas Jerome Newton
Tony Visconti
Duncan Jones
Ricky Gervais
Martin Scorsese
Elvis Presley
Bob Moore
Marie Osmond
Millie Kirsla Fisher
Mikkey Dee
Donny Edwards
Shia Labeouf
Boy-Bois
Jamie Foxx
Dave Grohl
Michael Keaton
Jack Daniels
Jennifer Lopez
Cock
Gallagher
Tom Hanks
Anne Wilson
Tony Stew
Alejandro Gonz
Viva Las Vegas
Jakob Dylan
Taylor Kinney

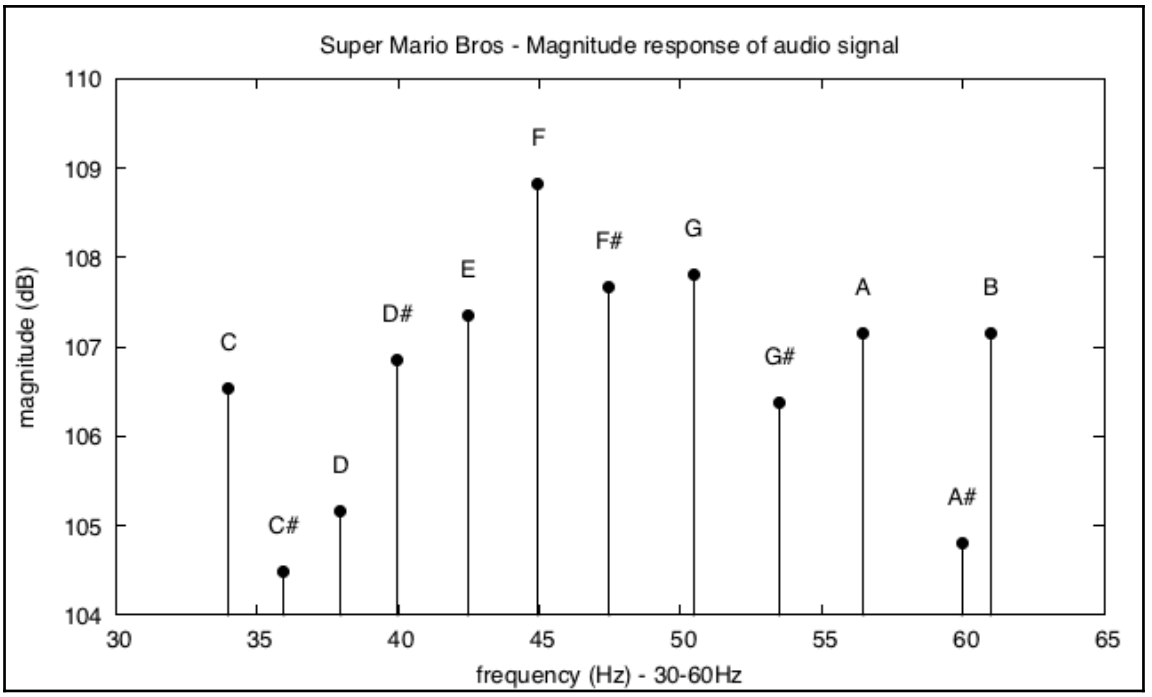
Chapter 08: Building a Recommendation System

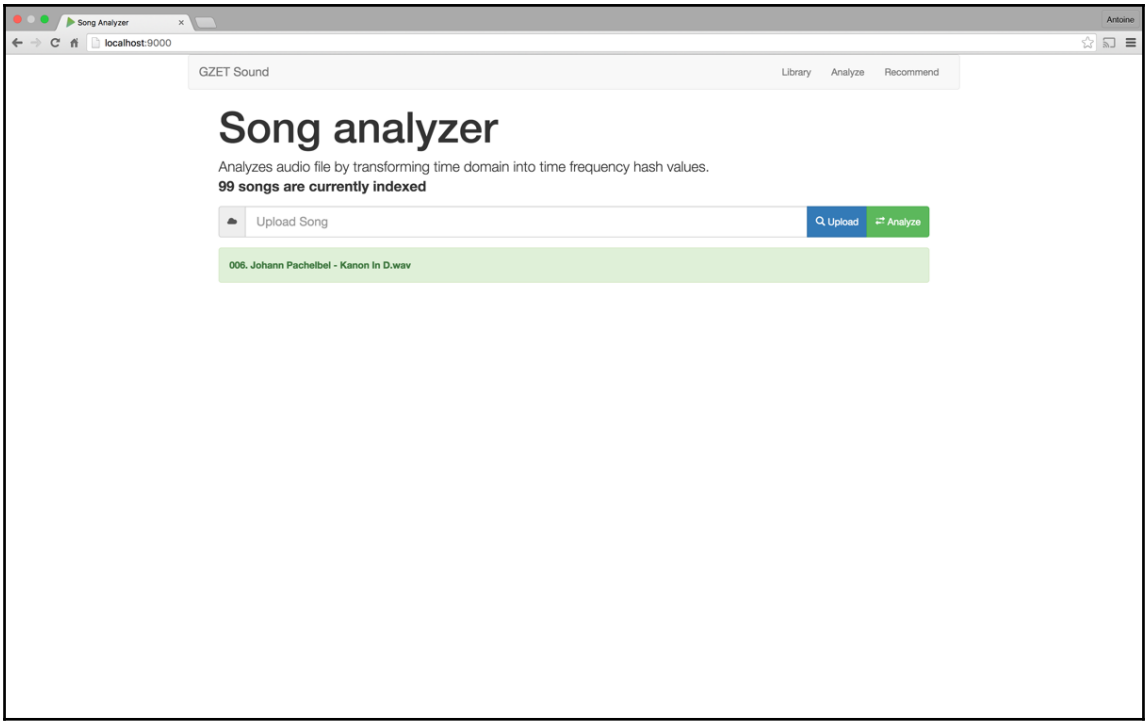


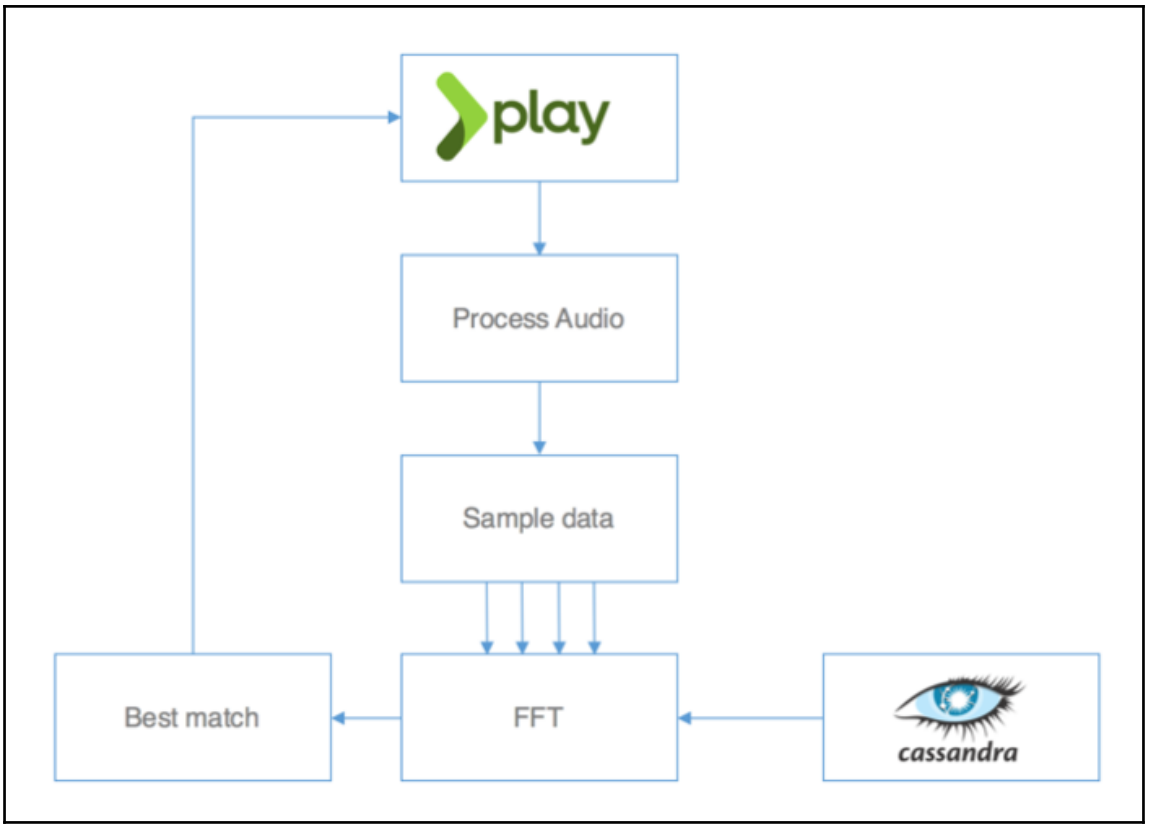
$$F(n) = \sum_{k=0}^{N-1} x(k) e^{\frac{-j2\pi kn}{N}}$$

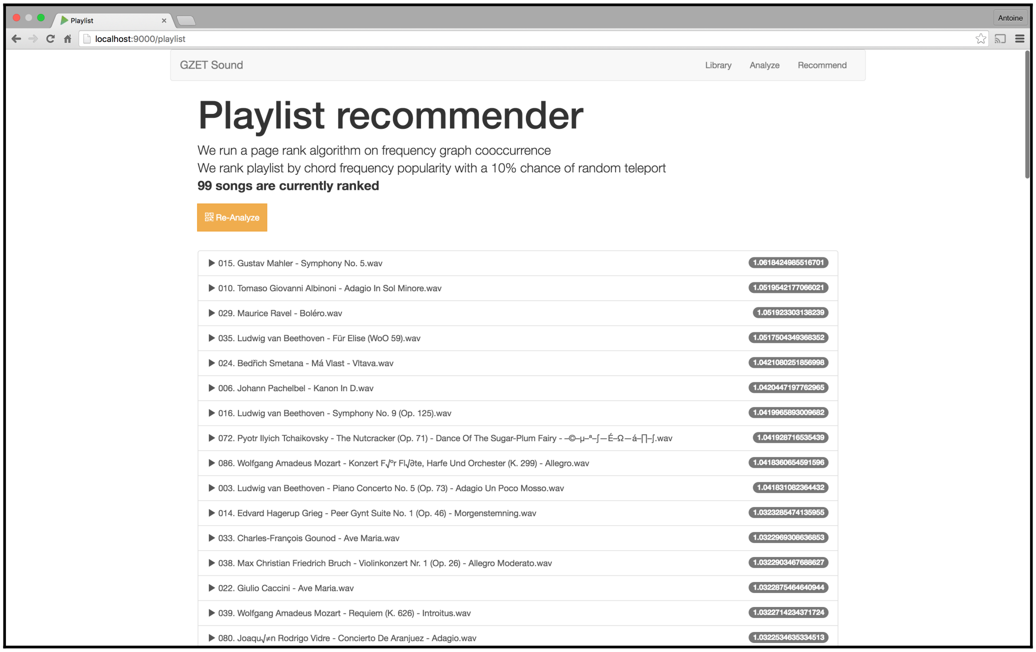


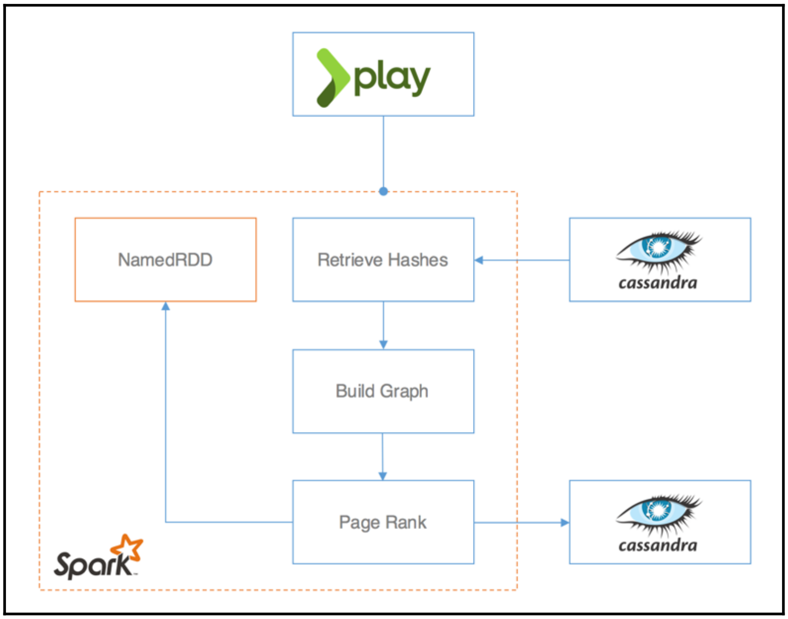
	C	C#	D	Eb	E	F	F#	G	G#	A	Bb	B
0	16.35	17.32	18.35	19.45	20.60	21.83	23.12	24.50	25.96	27.50	29.14	30.87
1	32.70	34.65	36.71	38.89	41.20	43.65	46.25	49.00	51.91	55.00	58.27	61.74
2	65.41	69.30	73.42	77.78	82.41	87.31	92.50	98.00	103.8	110.0	116.5	123.5
3	130.8	138.6	146.8	155.6	164.8	174.6	185.0	196.0	207.7	220.0	233.1	246.9
4	261.6	277.2	293.7	311.1	329.6	349.2	370.0	392.0	415.3	440.0	466.2	493.9
5	523.3	554.4	587.3	622.3	659.3	698.5	740.0	784.0	830.6	880.0	932.3	987.8
6	1047	1109	1175	1245	1319	1397	1480	1568	1661	1760	1865	1976
7	2093	2217	2349	2489	2637	2794	2960	3136	3322	3520	3729	3951
8	4186	4435	4699	4978	5274	5588	5920	6272	6645	7040	7459	7902

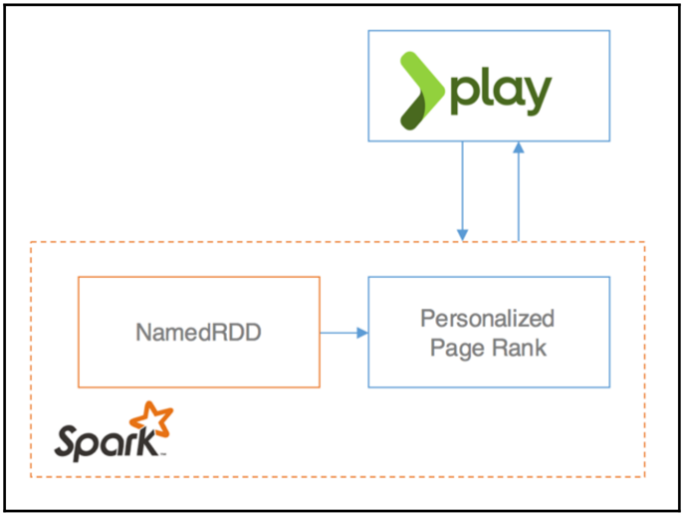

Super Mario Bros - Magnitude response of audio signal
magnitude (dB)
frequency (Hz) - 30-60Hz
110
109
108
107
106
105
104
30
35
40
45
50
55
60
65
C
C#
D
D#
E
F
F#
G
G#
A
A#
B

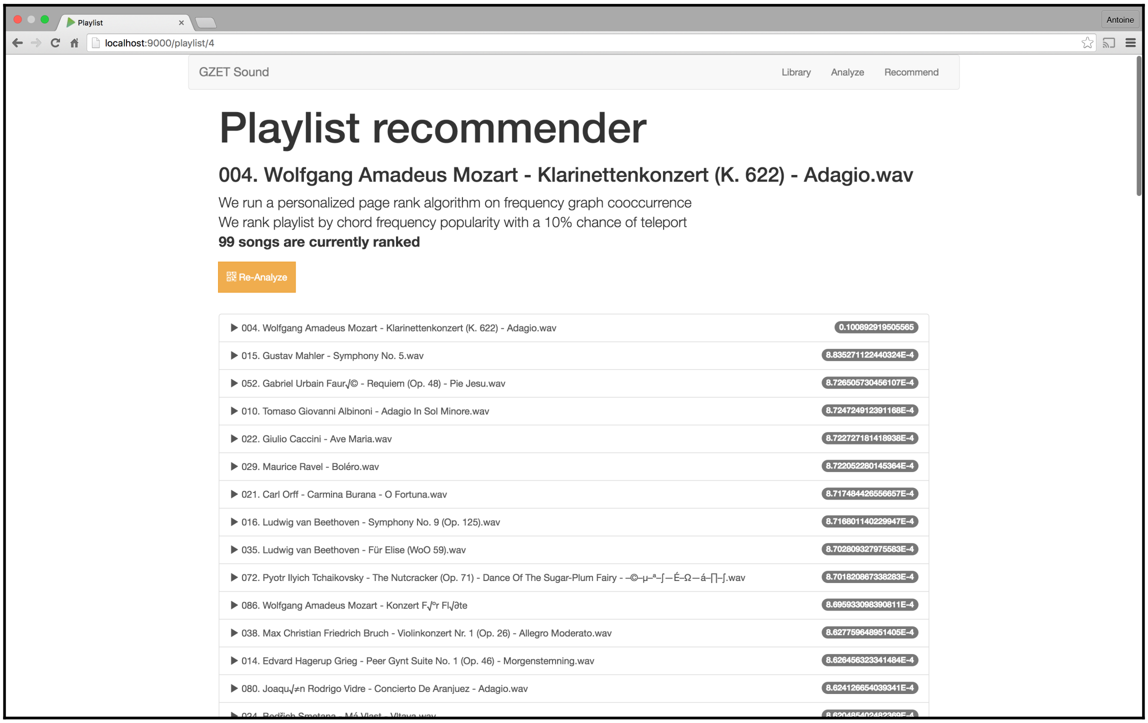

Song Analyzer
Antoine
localhost:9000
GZET Sound
Library
Analyze
Recommend
Song analyzer
Analyzes audio file by transforming time domain into time frequency hash values.
99 songs are currently indexed
Upload Song
Upload
Analyze
006. Johann Pachelbel - Kanon In D.wav


play
Process Audio
Sample data
Best match
FFT
cassandra

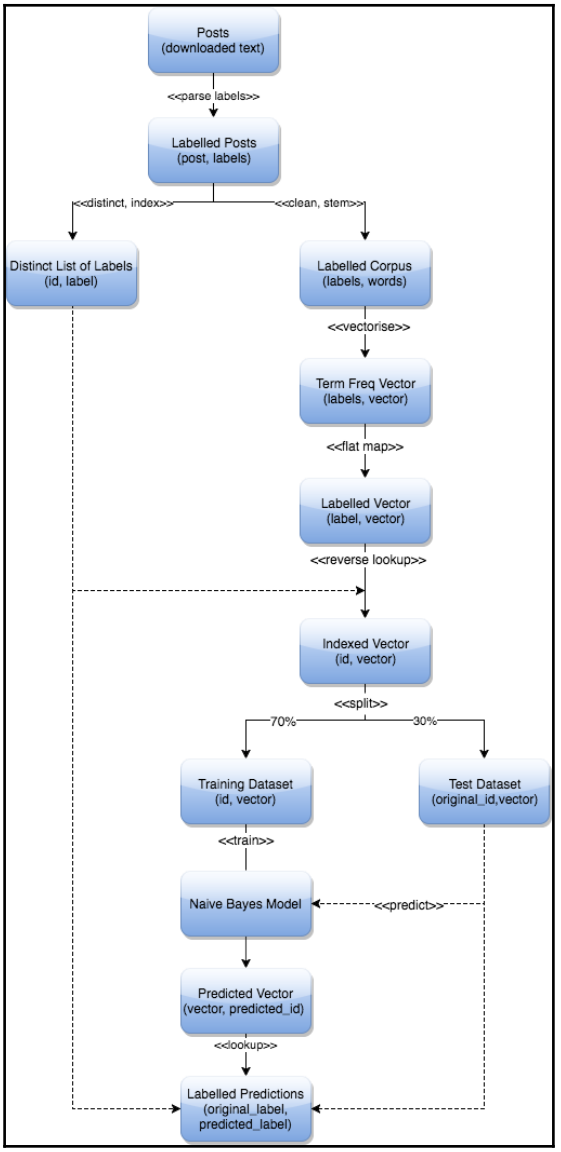

Playlist
Antoine
localhost:9000/playlist
GZET Sound
Library
Analyze
Recommend
Playlist recommender
We run a page rank algorithm on frequency graph cooccurrence
We rank playlist by chord frequency popularity with a 10% chance of random teleport
99 songs are currently ranked
Re-Analyze
015. Gustav Mahler - Symphony No. 5.wav 1.0618424985516701
010. Tomaso Giovanni Albinoni - Adagio In Sol Minore.wav 1.0519542177066021
029. Maurice Ravel - Boléro.wav 1.051923303138239
035. Ludwig van Beethoven - Für Elise (WoO 59).wav 1.0517504349368352
024. Bedřich Smetana - Má Vlast - Vltava.wav 1.0421080251856998
006. Johann Pachelbel - Kanon In D.wav 1.0420447197762965
016. Ludwig van Beethoven - Symphony No. 9 (Op. 125).wav 1.0419965893009682
072. Pyotr Ilyich Tchaikovsky - The Nutcracker (Op. 71) - Dance Of The Sugar-Plum Fairy - –©–µ–ª–∫—É–Ω—á–∏–∫.wav 1.041928716535439
086. Wolfgang Amadeus Mozart - Konzert F√ºr Fl√∂te, Harfe Und Orchester (K. 299) - Allegro.wav 1.0418360654591596
003. Ludwig van Beethoven - Piano Concerto No. 5 (Op. 73) - Adagio Un Poco Mosso.wav 1.041831082364432
014. Edvard Hagerup Grieg - Peer Gynt Suite No. 1 (Op. 46) - Morgenstemning.wav 1.0323285474135955
033. Charles-François Gounod - Ave Maria.wav 1.0322969308636853
038. Max Christian Friedrich Bruch - Violinkonzert Nr. 1 (Op. 26) - Allegro Moderato.wav 1.0322903467688627
022. Giulio Caccini - Ave Maria.wav 1.0322875464640944
039. Wolfgang Amadeus Mozart - Requiem (K. 626) - Introitus.wav 1.0322714234371724
080. Joaqu√≠n Rodrigo Vidre - Concierto De Aranjuez - Adagio.wav 1.0322534635334513


play
NamedRDD
Retrieve Hashes
cassandra
Build Graph
Page Rank
cassandra
Spark


play
NamedRDD
Personalized
Page Rank
Spark


Playlist
Antoine
localhost:9000/playlist/4
GZET Sound
Library
Analyze
Recommend
Playlist recommender
004. Wolfgang Amadeus Mozart - Klarinettenkonzert (K. 622) - Adagio.wav
We run a personalized page rank algorithm on frequency graph cooccurrence
We rank playlist by chord frequency popularity with a 10% chance of teleport
99 songs are currently ranked
Re-Analyze
004. Wolfgang Amadeus Mozart - Klarinettenkonzert (K. 622) - Adagio.wav
0.100892919505565
015. Gustav Mahler - Symphony No. 5.wav
8.835271122440324E-4
052. Gabriel Urbain Faur√© - Requiem (Op. 48) - Pie Jesu.wav
8.726505730456107E-4
010. Tomaso Giovanni Albinoni - Adagio In Sol Minore.wav
8.724724912391168E-4
022. Giulio Caccini - Ave Maria.wav
8.722727181418938E-4
029. Maurice Ravel - Boléro.wav
8.722052280145364E-4
021. Carl Orff - Carmina Burana - O Fortuna.wav
8.717484426556657E-4
016. Ludwig van Beethoven - Symphony No. 9 (Op. 125).wav
8.716801140229947E-4
035. Ludwig van Beethoven - Für Elise (WoO 59).wav
8.702809327975583E-4
072. Pyotr Ilyich Tchaikovsky - The Nutcracker (Op. 71) - Dance Of The Sugar-Plum Fairy - –©–µ–ª–∫—É–Ω—á–∏–∫.wav
8.701820867338283E-4
086. Wolfgang Amadeus Mozart - Konzert F√ºr Fl√∂te
8.695933098390811E-4
038. Max Christian Friedrich Bruch - Violinkonzert Nr. 1 (Op. 26) - Allegro Moderato.wav
8.627759648951405E-4
014. Edvard Hagerup Grieg - Peer Gynt Suite No. 1 (Op. 46) - Morgenstemning.wav
8.626456323341484E-4
080. Joaqu√≠n Rodrigo Vidre - Concierto De Aranjuez - Adagio.wav
8.624126654039341E-4

Chapter 09: News Dictionary and Real-Time Tagging System

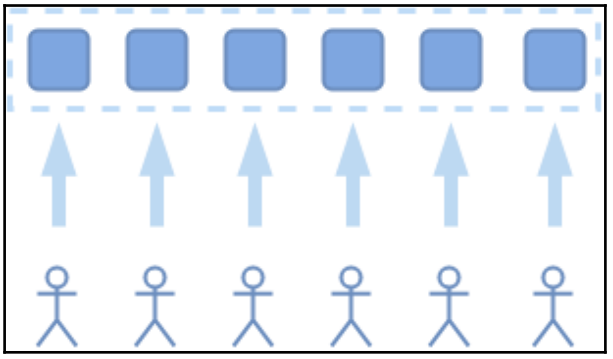


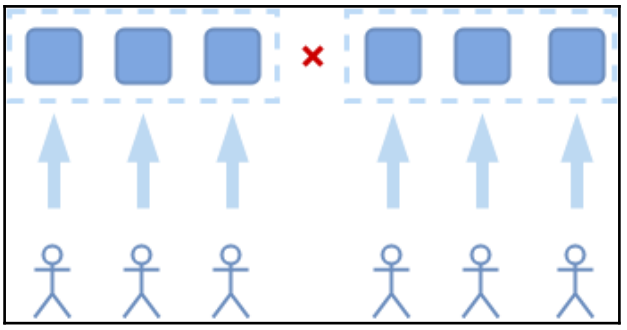
WFMY News 2 @WFMY · Apr 24

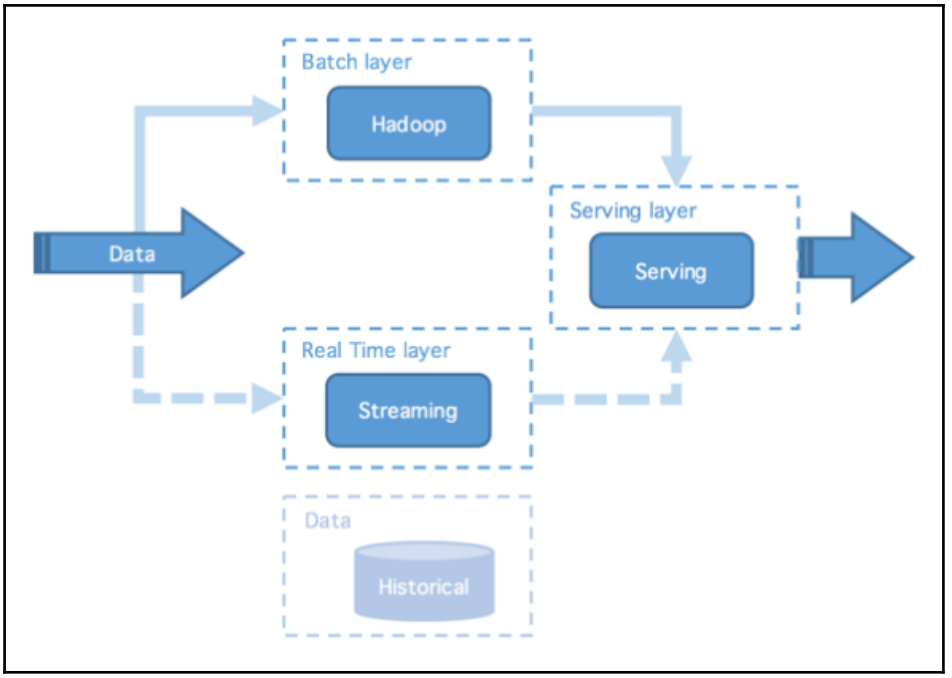
Adorable #Prince **George** misses bedtime, meets President #Obama -- ow.ly/4n2snm

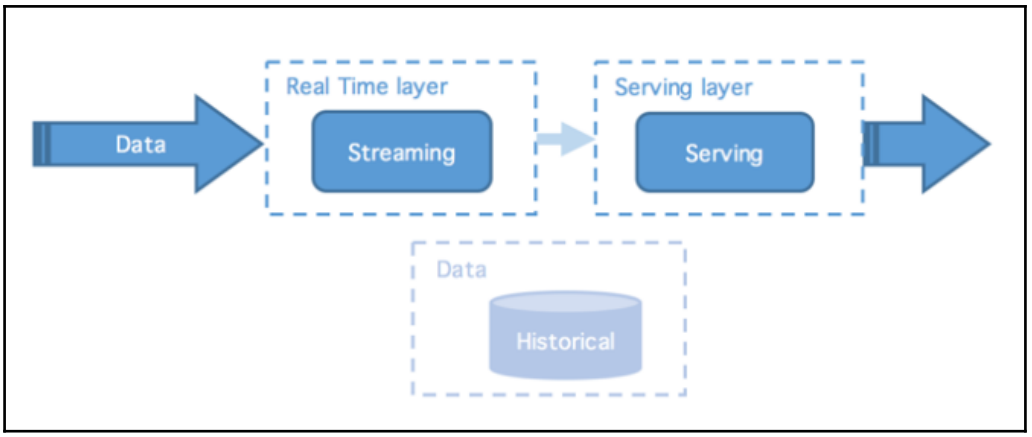
1 4

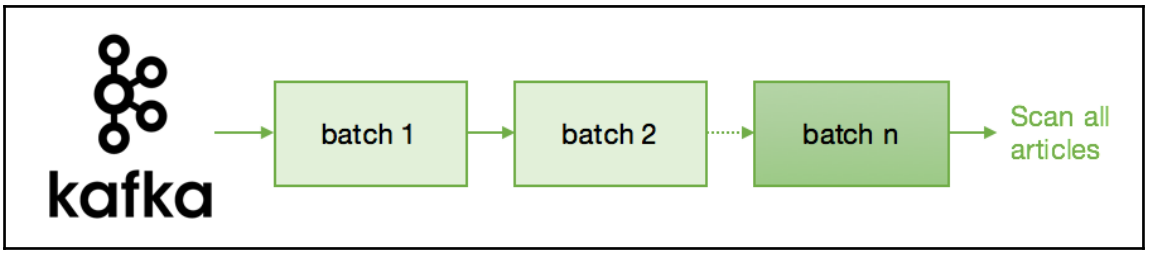

Kids of the Eighties @kidsofeighties · 30 Dec 2016
Death stars: music's great losses of 2016 via @guardian
theguardian.com/music/2016/dec… #DavidBowie #PRINCE #georgemichael
#LeonardCohen
1
3

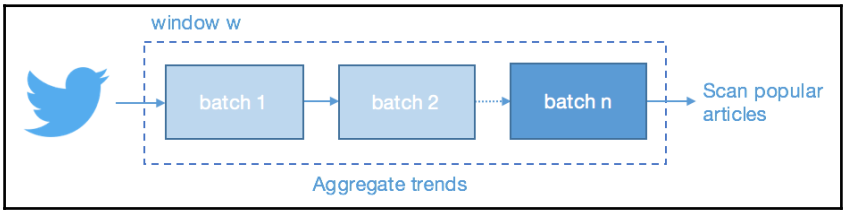


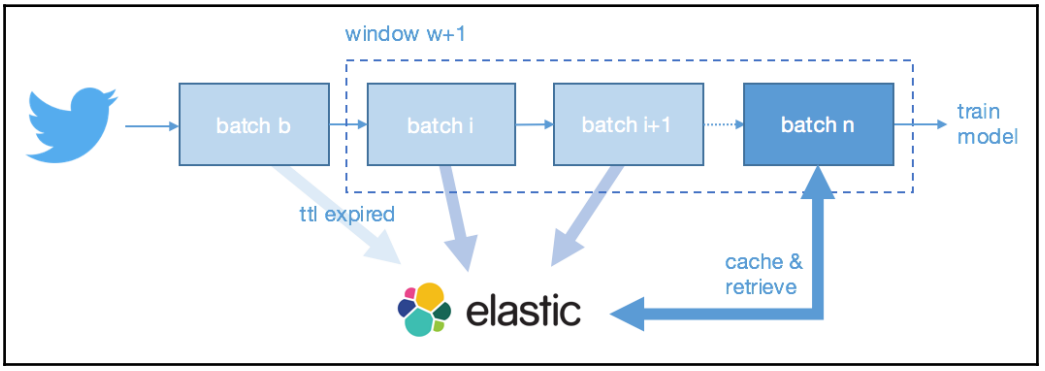


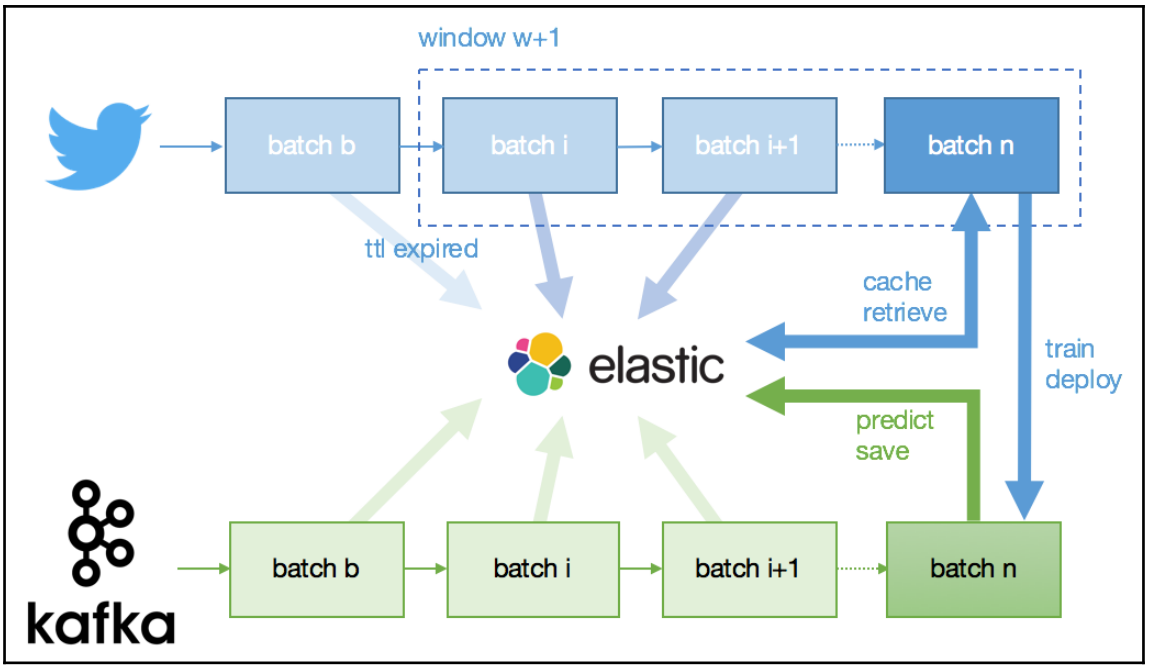

Batch layer
Hadoop
Data
Serving layer
Serving
Real Time layer
Streaming
Data
Historical


Real Time layer
Serving layer
Data
Streaming
Serving
Data
Historical

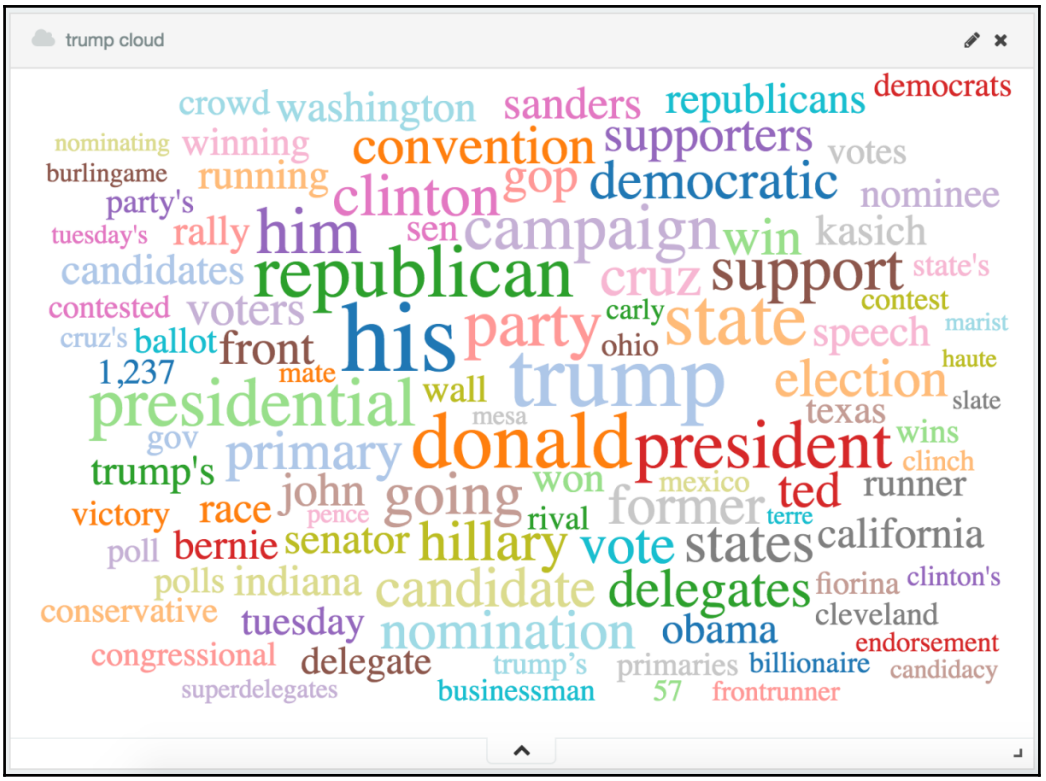

kafka
batch 1
batch 2
batch n
Scan all articles


window w
batch 1
batch 2
batch n
Scan popular articles
Aggregate trends


window w+1
batch b
batch i
batch i+1
batch n
train model
ttl expired
cache & retrieve
elastic


window w+1
batch b
batch i
batch i+1
batch n
ttl expired
cache
retrieve
train
deploy
elastic
predict
save
kafka
batch b
batch i
batch i+1
batch n


predicted articles
predicted articles
unique hashtags
1,000
800
600
400
200
0
19:00
22:00
01:00
04:00
07:00
10:00
13:00
time per 30 minutes
hashtag cloud
#makeamericagreatagain
#AyrtonSenna
#gameofthrones
#leadersdebate
#hiring
#handmade
#inprimary
#toryparty
#ublraid
#swamyrocks
#giveaway
#Zodiac
#health
#nevertrump
#retail
#business
#Baaghi
#follow
#travel
#edchat
#cruzcrew
#iartg
#trump
#nowplaying
#startup
#asmsg
#deals
#pjnet
#books
#mayday
#auspol
#yahoo
#usury
#Management
#fashion
#softrock
#writing
#wedding
#GLvKXIP
#feelthebern
#crafthour
#sports
#indianaprimary
#freebiafra
#indiana
#teamfollowback
#CyberLoversBook
#fastcompany
#hospitality
#mothersday


trump cloud
democrats
crowd
washington
sanders
republicans
nominating
winning
convention
supporters
votes
burlingame
running
clinton
gop
democratic
nominee
party's
tuesday's
rally
him
sen
campaign
win
kasich
candidates
republican
cruz
support
state's
contested
voters
carly
contest
his
party
state
speech
marist
cruz's
ballot
front
ohio
haute
1,237
mate
trump
election
presidential
wall
slate
mesa
texas
gov
primary
donald
president
wins
trump's
won
mexico
clinch
john
going
ted
runner
victory
race
pence
former
terre
rival
poll
bernie
senator
hillary
vote
states
california
polls
indiana
candidate
delegates
fiorina
clinton's
conservative
tuesday
nomination
obama
cleveland
endorsement
congressional
delegate
trump's
primaries
billionaire
candidacy
superdelegates
businessman
57
frontrunner

Chapter 10: Story De-duplication and Mutation

Antoine Amend
Just now · London ·

http://www.forbes.com/…/samsung-south-korean-government-i…/…

Galaxy Note 7 Fiasco: Samsung, South Korean Government Launches Investigation

Industry analysts say investigating why the Note 7 devices caught fire is crucial for the world's largest smartphone maker.

FORBES.COM | BY JOHN KANG

Like Comment Share

Write a comment...

RELATED ARTICLES

Why Samsung Abandoned Its Galaxy Note 7 Flagship Phone

The unprecedented move by the South Korean electronics giant is an embarrassing reversal for a…

THE NEW YORK TIMES · 11,181 SHARES Share Save

Singapore Airlines bans Samsung Galaxy Note 7 on its flights

SINGAPORE Singapore Airlines said on Saturday it has banned Samsungs Galaxy Note 7 mobile phon…

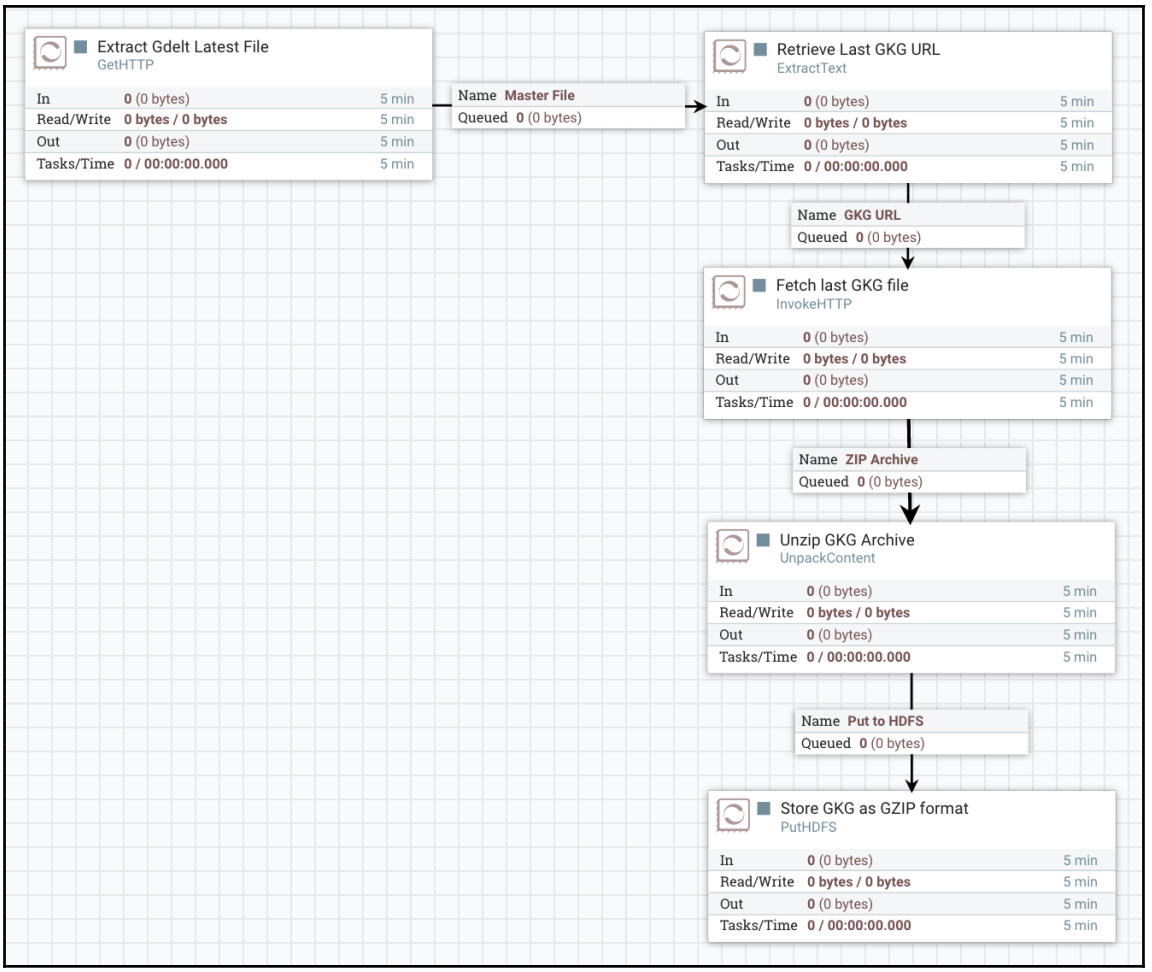
GMA NEWS · 4,386 SHARES Share Save

	hashcodes											
he	1	1	0	0	1	1	1	1	1	1	0	1
el	1	1	0	0	1	0	1	0	0	1	1	1
ll	1	1	0	1	1	0	0	0	0	0	0	0
lo	1	1	0	1	1	0	0	0	0	0	1	1
si	1	1	1	0	0	1	0	1	0	1	1	0
im	1	1	0	1	0	0	1	0	0	1	0	0
mh	1	1	0	1	1	0	0	1	1	0	1	1
ha	1	1	0	0	1	1	1	1	1	0	0	1
as	1	1	0	0	0	0	1	1	0	0	1	0
sh	1	1	1	0	0	1	0	1	0	1	0	1

	10	10	-6	-2	2	-2	0	2	-4	0	0	2
hello simhash	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**

hello simhash	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**
hello minhash	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**1**
hello world	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**0**

hello simhash	1	1	0	0	1	0	0	1	0	0	0	1
hello minhash	1	1	0	0	1	0	1	1	0	0	0	1
XOR	0	0	0	0	0	0	**1**	0	0	0	0	0


Extract Gdelt Latest File
GetHTTP
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name Master File
Queued 0 (0 bytes)
Retrieve Last GKG URL
ExtractText
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name GKG URL
Queued 0 (0 bytes)
Fetch last GKG file
InvokeHTTP
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name ZIP Archive
Queued 0 (0 bytes)
Unzip GKG Archive
UnpackContent
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name Put to HDFS
Queued 0 (0 bytes)
Store GKG as GZIP format
PutHDFS
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Note 7 fiasco leaves Samsung's smartphone brand in question

By ASSOCIATED PRESS
PUBLISHED: 20:02, 12 October 2016 | **UPDATED:** 20:02, 12 October 2016

Share

SEOUL, South Korea (AP) — The fiasco of Samsung's fire-prone Galaxy Note 7 smartphones — and its stumbling response to the problem — has left consumers from Shanghai to New York reconsidering how they feel about the South Korean tech giant and its products.

Samsung Electronics said this week that it would stop making the Note 7 for good, after first recalling some devices and then recalling their replacements , too. Now, like the makers of Tylenol, Ford Pintos and other products that faced crises in the past, it must try to restore its relationship with customers as it repairs damage to its brand.

Samsung shares plunged as much as 8 percent in Seoul, their biggest one-day drop since the 2008 financial crisis, after the company apologized for halting sales of the Note 7 .

Note 7 Fiasco Leaves Samsung's Smartphone Brand in Question

A visitor passes by an advertisement of the Samsung Electronics Galaxy Note 7 smartphone at its shop in Seoul, South Korea, Oct. 11, 2016.

SEOUL, SOUTH KOREA — The fiasco of Samsung's fire-prone Galaxy Note 7 smartphones - and its stumbling response to the problem - has left consumers from Shanghai to New York reconsidering how they feel about the South Korean tech giant and its products.

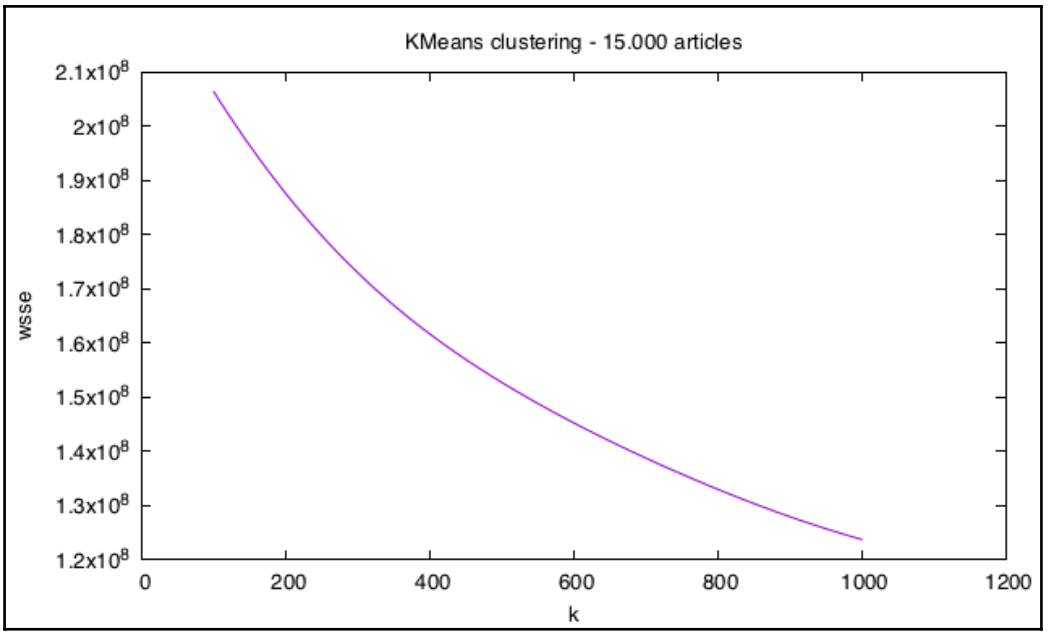
Related

Samsung Drastically Cuts Profit Estimates As Malfunctions Continue

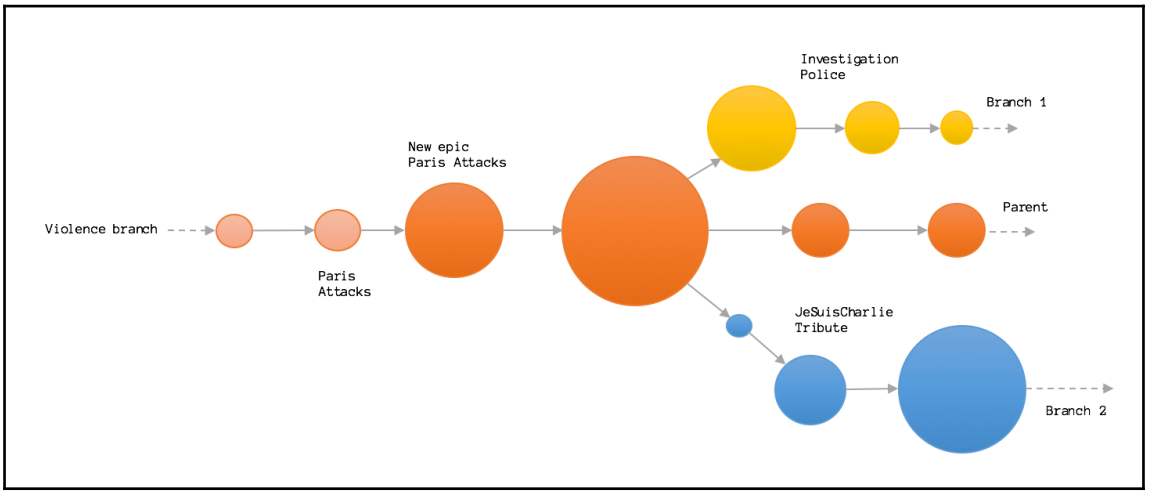
$$idf_i = log\left(\frac{n+1}{df_i+1}\right)$$

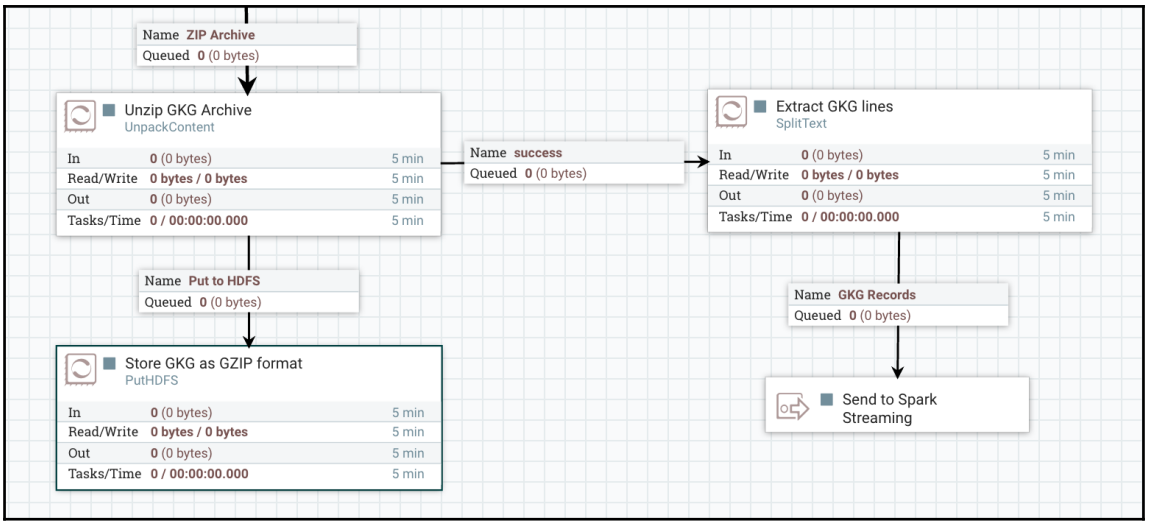
	Random Contexts								
mastering	1	0	0	0	0	1	1	0	-1
spark	0	1	0	0	1	1	1	-1	0
for	1	0	0	0	1	1	-1	0	1
data	0	0	0	0	0	0	0	1	0
science	-1	1	1	0	1	1	0	0	-1

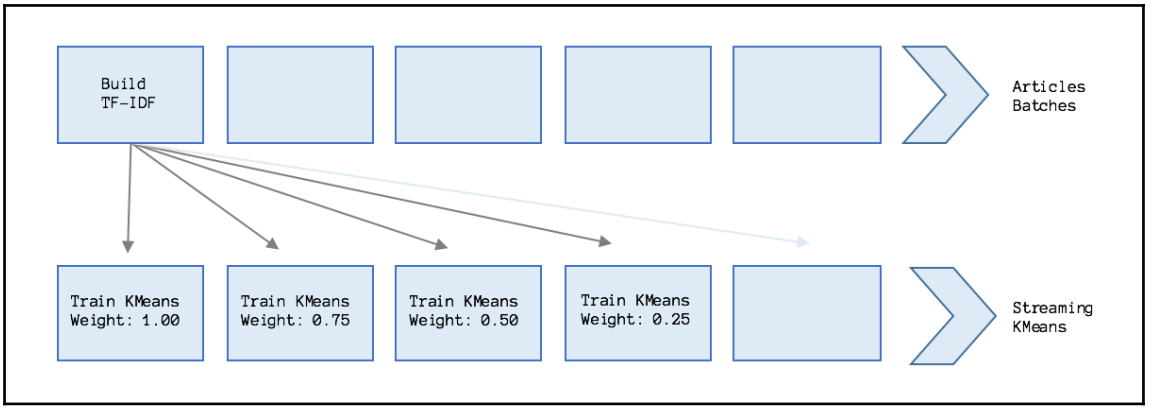
	Aggregated Context								
mastering spark for data science	1	2	1	0	3	4	1	0	-1

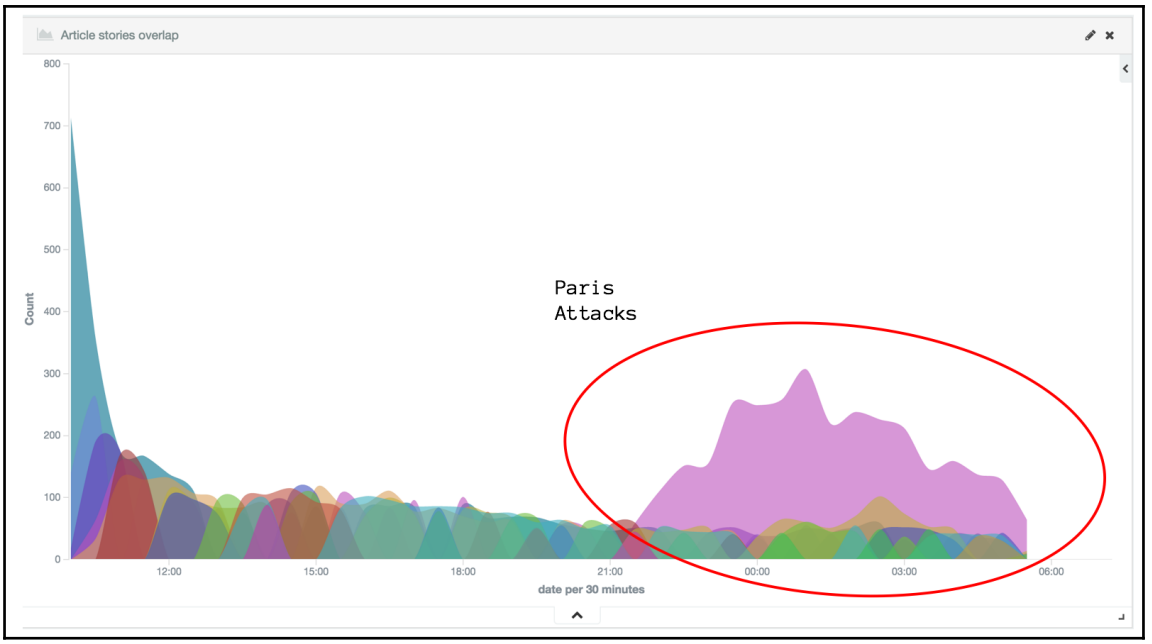


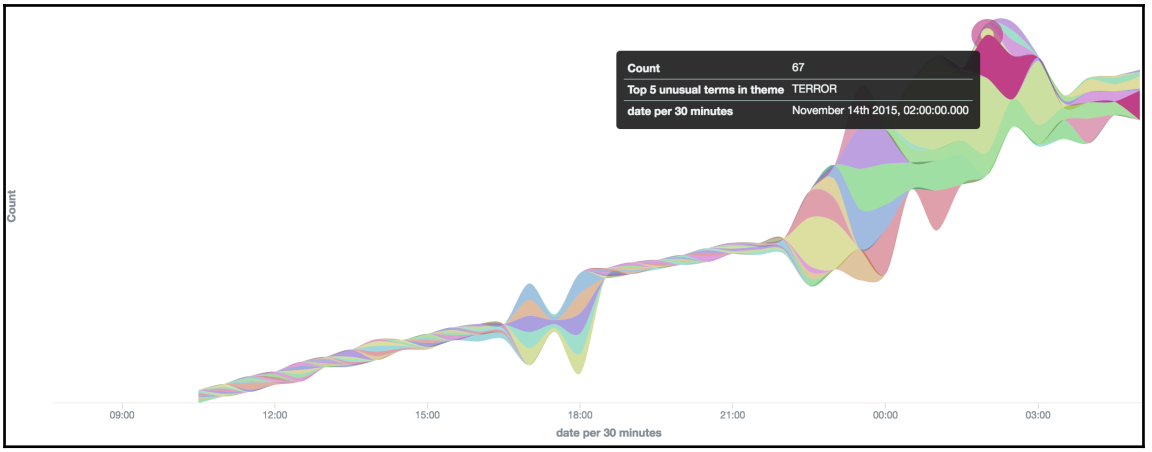
$$k \approx \sqrt{\frac{n}{2}}$$

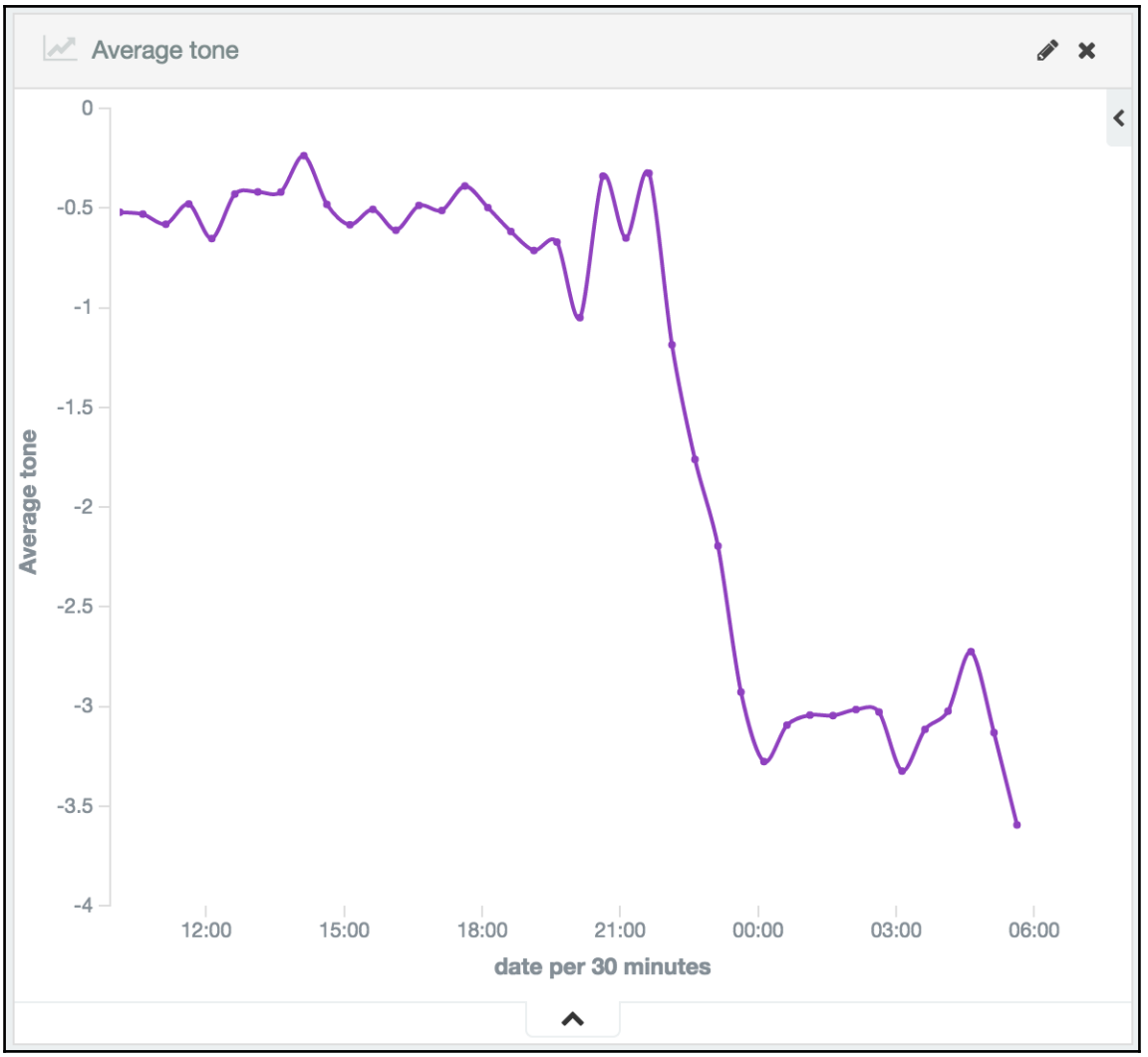


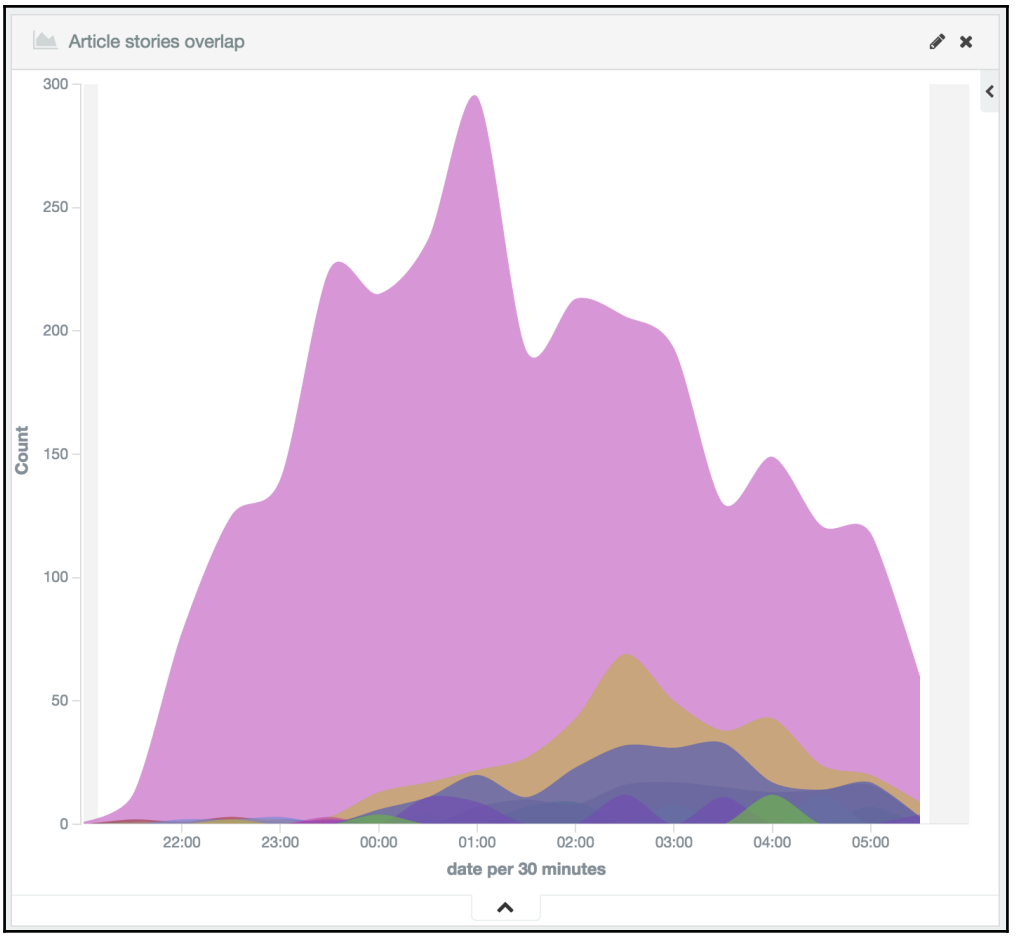

Name ZIP Archive
Queued 0 (0 bytes)
Unzip GKG Archive
UnpackContent
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
Extract GKG lines
SplitText
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name Put to HDFS
Queued 0 (0 bytes)
Name GKG Records
Queued 0 (0 bytes)
Store GKG as GZIP format
PutHDFS
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Send to Spark Streaming

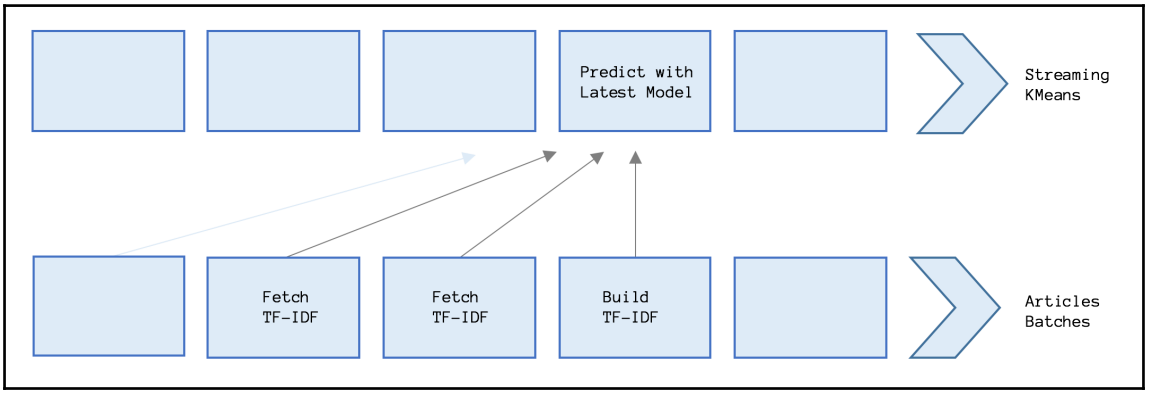

Build TF-IDF
Articles Batches
Train KMeans Weight: 1.00
Train KMeans Weight: 0.75
Train KMeans Weight: 0.50
Train KMeans Weight: 0.25
Streaming KMeans

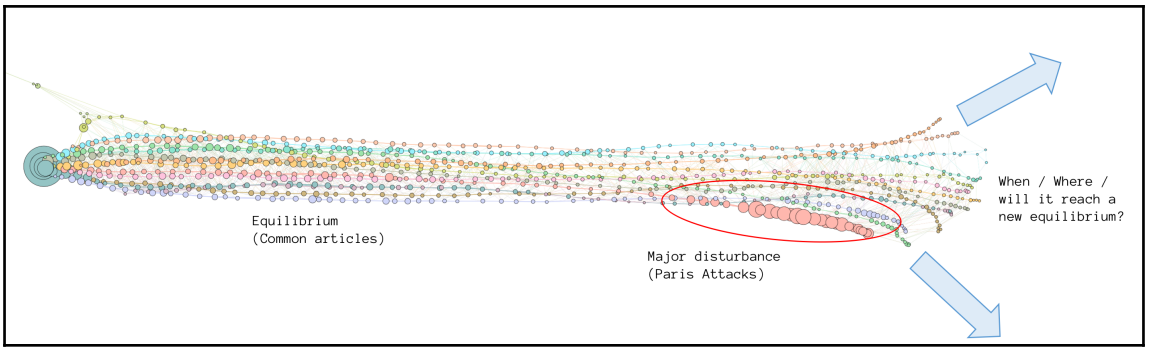

Article stories overlap
800
700
600
500
400
300
200
100
0
Count
Paris
Attacks
12:00
15:00
18:00
21:00
00:00
03:00
06:00
date per 30 minutes

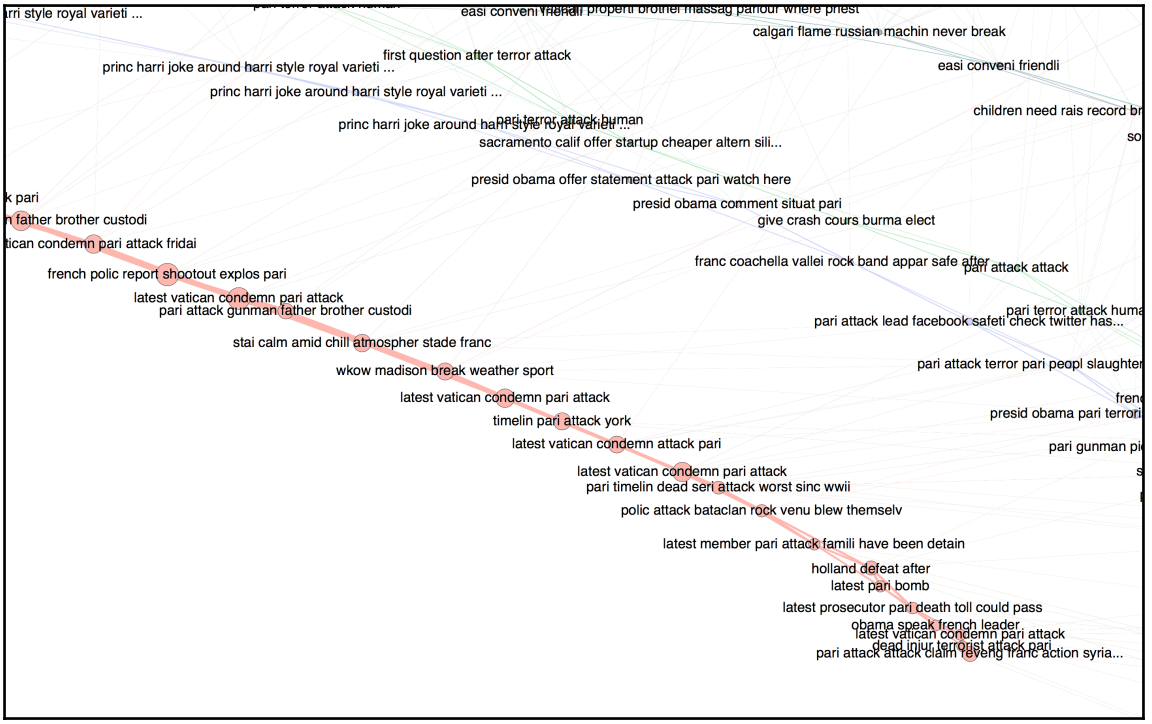

Count
67
Top 5 unusual terms in theme
TERROR
date per 30 minutes
November 14th 2015, 02:00:00.000
Count
09:00
12:00
15:00
18:00
21:00
00:00
03:00
date per 30 minutes


Average tone
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Average tone
12:00
15:00
18:00
21:00
00:00
03:00
06:00
date per 30 minutes


Article stories overlap
300
250
200
150
100
50
0
Count
22:00
23:00
00:00
01:00
02:00
03:00
04:00
05:00
date per 30 minutes


Predict with
Latest Model
Streaming
KMeans
Fetch
TF-IDF
Fetch
TF-IDF
Build
TF-IDF
Articles
Batches

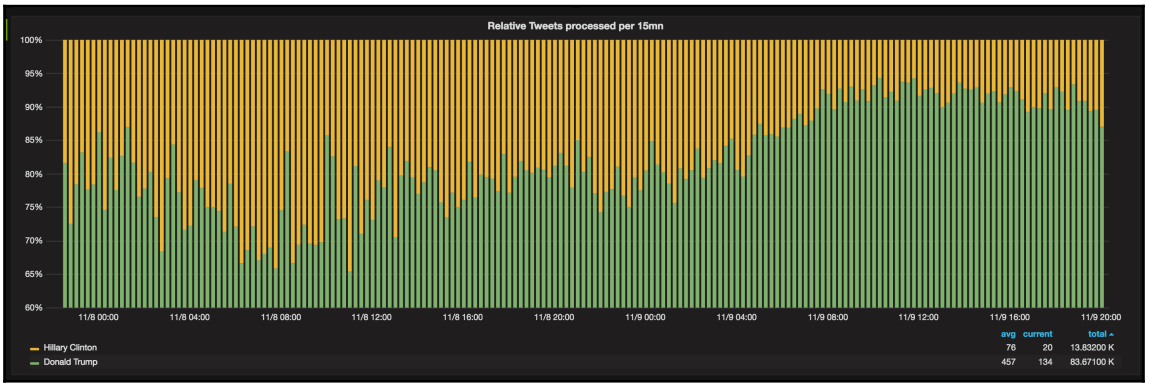

Equilibrium
(Common articles)
Major disturbance
(Paris Attacks)
When / Where /
will it reach a
new equilibrium?


rri style royal varieti ...
princ harri joke around harri style royal varieti ...
first question after terror attack
princ harri joke around harri style royal varieti ...
princ harri joke around harri style royal varieti ...
pari terror attack human
sacramento calif offer startup cheaper altern sili...
calgari flame russian machin never break
easi conveni friendli
children need rais record br
presid obama offer statement attack pari watch here
presid obama comment situat pari
give crash cours burma elect
k pari
father brother custodi
tican condemn pari attack fridai
franc coachella vallei rock band appar safe after
pari attack attack
french polic report shootout explos pari
latest vatican condemn pari attack
pari attack gunman father brother custodi
pari terror attack huma
pari attack lead facebook safeti check twitter has...
stai calm amid chill atmospher stade franc
wkow madison break weather sport
pari attack terror pari peopl slaughter
latest vatican condemn pari attack
timelin pari attack york
presid obama pari terrori
latest vatican condemn attack pari
pari gunman pi
latest vatican condemn pari attack
pari timelin dead seri attack worst sinc wwii
polic attack bataclan rock venu blew themselv
latest member pari attack famili have been detain
holland defeat after
latest pari bomb
latest prosecutor pari death toll could pass
obama speak french leader
latest vatican condemn pari attack
dead injur terrorist attack pari
pari attack attack claim revenq franc action syria...

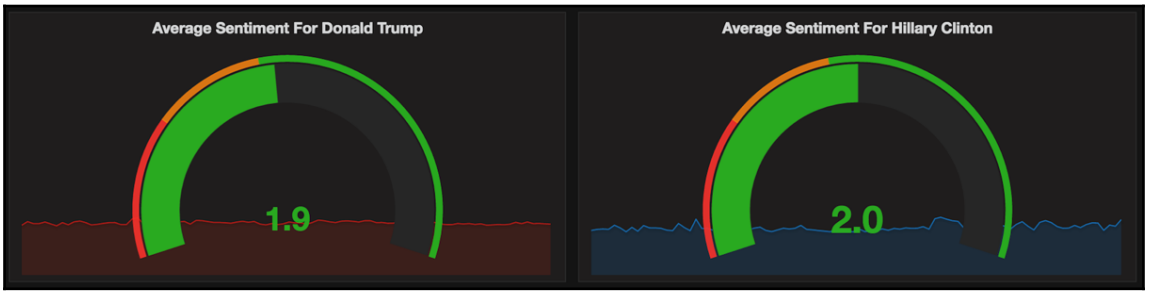
Chapter 11: Anomaly Detection on Sentiment Analysis

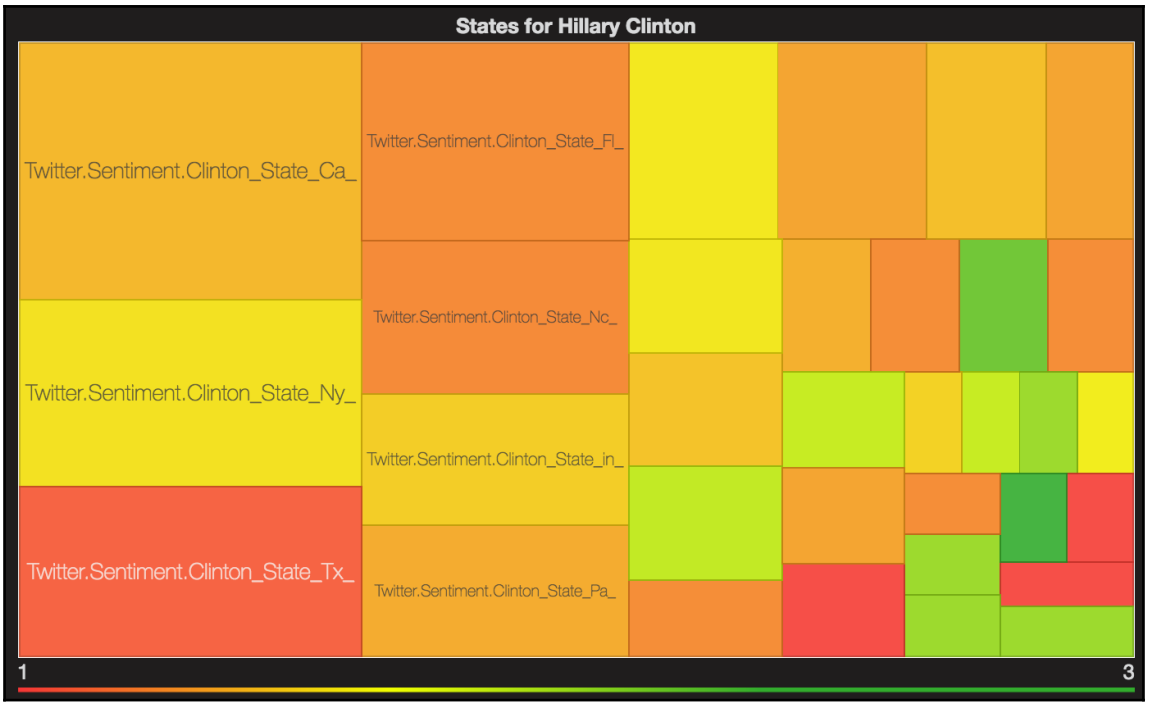
```
scala> textWithEmojis
res1: String = A 🐱 🐶and a 🐭became friends💕 For 🐶s birthday party, they all had 🍔, 🍟, 🍪 and 🍰

scala> EmojiUtils.shortCodify(textWithEmojis)
res2: String = A :cat:, :dog: and a :mouse: became friends:heart:. For :dog:'s birthday party, they all had :hamburger:s, :fries:s, :cookie:s and :cake:.

scala> EmojiUtils.removeAllEmojis(textWithEmojis)
res3: String = A ,  and a  became friends. For 's birthday party, they all had s, s, s and .
```

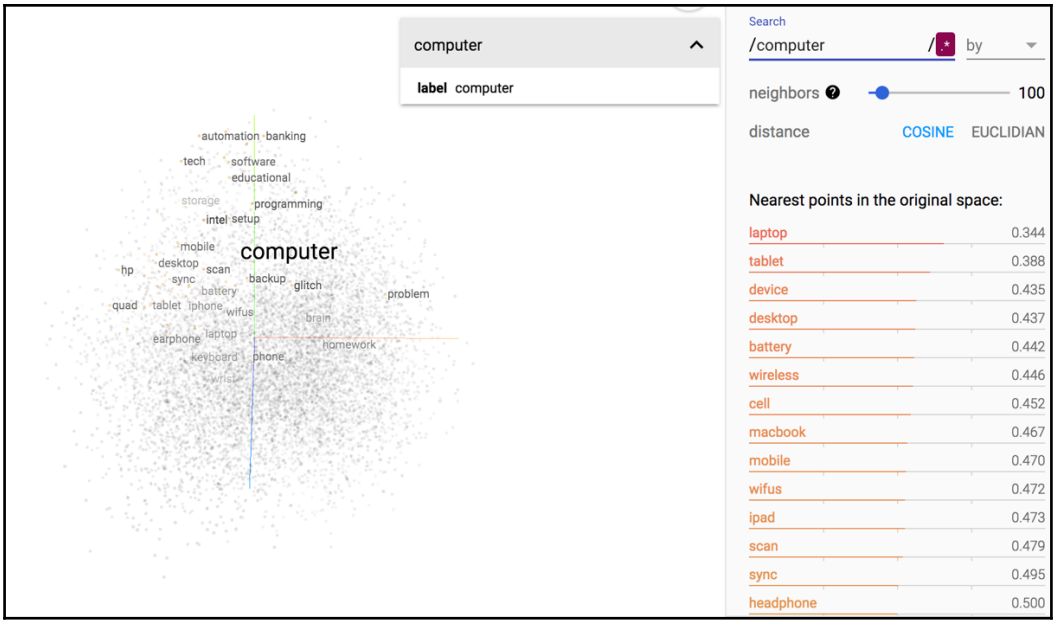
```
scala> import io.gzet.timeseries.twitter.Twitter._
import io.gzet.timeseries.twitter.Twitter._

scala> println(text)
RT @johnanjos: Michelle Obama for president in 2020 🇺🇸
Michelle Obama for president in 2020 🇺🇸
Michelle Obama for president in 2020 🇺🇸
#Not…

scala> println(text.clean)
michelle obama for president in 2020 michelle obama for president in 2020 michelle obama for president in 2020 #not

scala> println(text.emojis.mkString(" "))
us us us
```

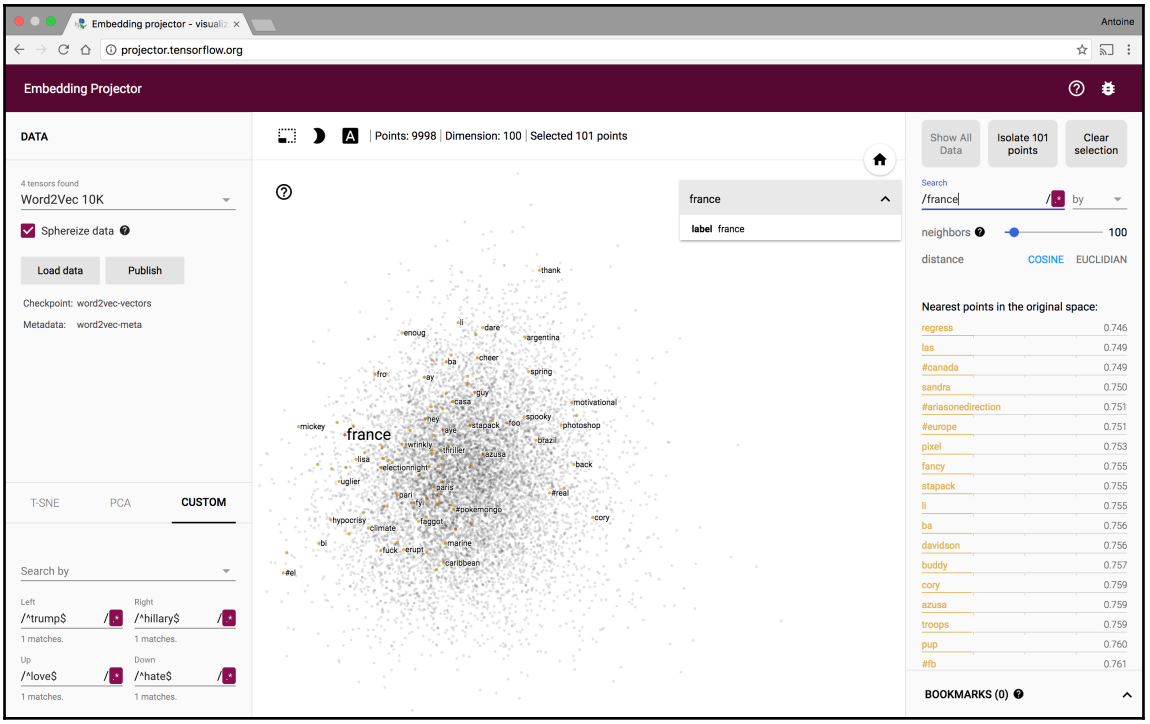


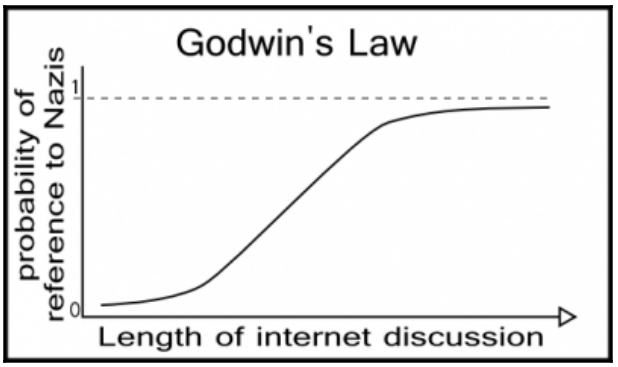

https://localhost:3000/dashboard/db/twitter
Twitter
Zoom Out
Nov 7, 2016 21:42:10 to Nov 9, 2016 20:20:39
Average Sentiment For Donald Trump
1.4
Average Sentiment For Hillary Clinton
1.3
Total Tweets received per 6h
Time
Donald Trump
Hillary Clinton
2016-11-07 21:42:10
2.19 K
577.00
2016-11-08 03:42:10
1.55 K
579.00
2016-11-08 09:42:10
2.31 K
673.00
2016-11-08 15:42:10
5.57 K
1.39 K
2016-11-08 21:42:10
15.07 K
3.64 K
Sentiment over time - Donald Trump
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
11/8 00:00
11/8 04:00
11/8 08:00
11/8 12:00
11/8 16:00
11/8 20:00
11/9 00:00
11/9 04:00
11/9 08:00
11/9 12:00
11/9 16:00
11/9 20:00
min
max
avg
current
Donald Trump
0.750
2.333
1.358
1.235

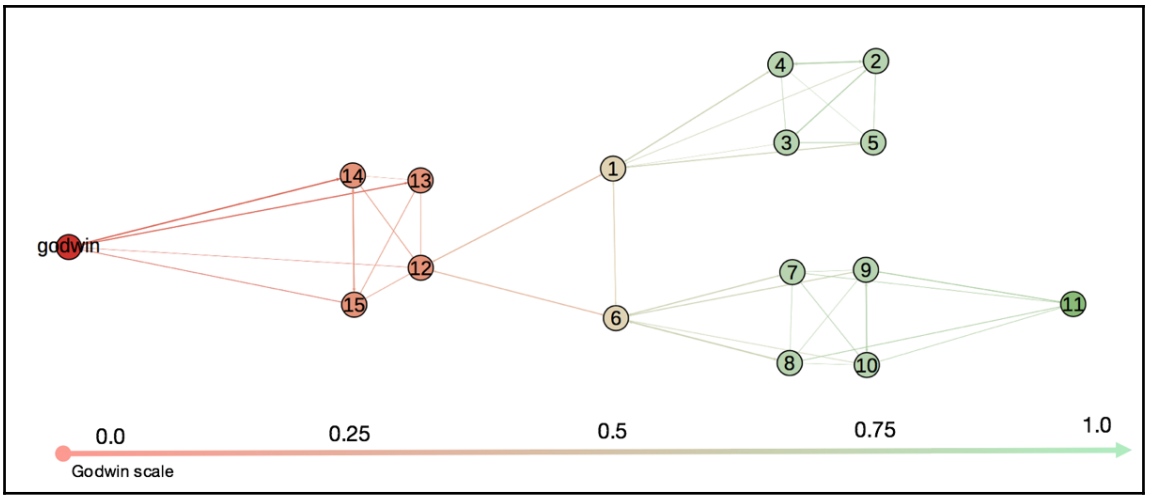

Average Sentiment For Donald Trump
1.9
Average Sentiment For Hillary Clinton
2.0


States for Hillary Clinton
Twitter.Sentiment.Clinton_State_Ca_
Twitter.Sentiment.Clinton_State_Fl_
Twitter.Sentiment.Clinton_State_Nc_
Twitter.Sentiment.Clinton_State_Ny_
Twitter.Sentiment.Clinton_State_in_
Twitter.Sentiment.Clinton_State_Tx_
Twitter.Sentiment.Clinton_State_Pa_
1
3


computer
label computer
automation
banking
tech
software
educational
storage
programming
intel
setup
mobile
computer
desktop
scan
hp
sync
backup
glitch
problem
battery
quad
tablet
iphone
wifus
brain
earphone
laptop
homework
keyboard
phone
wrist
Search
/computer
by
neighbors
100
distance
COSINE
EUCLIDIAN
Nearest points in the original space:
laptop 0.344
tablet 0.388
device 0.435
desktop 0.437
battery 0.442
wireless 0.446
cell 0.452
macbook 0.467
mobile 0.470
wifus 0.472
ipad 0.473
scan 0.479
sync 0.495
headphone 0.500


Embedding projector - visualiz
Antoine
projector.tensorflow.org
Embedding Projector
DATA
4 tensors found
Word2Vec 10K
Sphereize data
Load data
Publish
Checkpoint: word2vec-vectors
Metadata: word2vec-meta
T-SNE
PCA
CUSTOM
Search by
Left
/^trump$
1 matches.
Right
/^hillary$
1 matches.
Up
/^love$
1 matches.
Down
/^hate$
1 matches.
Points: 9998 | Dimension: 100 | Selected 101 points
france
label france
thank
dare
argentina
enoug
cheer
spring
motivational
mickey
france
spooky
photoshop
brazil
back
wrinkly
thriller
azusa
electionnight
ugher
paris
fyi
spokemongo
#real
cory
hypocrisy
climate
faggot
marine
fuck
erupt
caribbean
#el
Show All Data
Isolate 101 points
Clear selection
Search
/france
by
neighbors
100
distance
COSINE
EUCLIDIAN
Nearest points in the original space:
regress 0.746
las 0.749
#canada 0.749
sandra 0.750
#ariasonedirection 0.751
#europe 0.751
pixel 0.753
fancy 0.755
stapack 0.755
li 0.755
ba 0.756
davidson 0.756
buddy 0.757
cory 0.759
azusa 0.759
troops 0.759
pup 0.760
#fb 0.761
BOOKMARKS (0)


Godwin's Law
probability of reference to Nazis
1
0
Length of internet discussion



$$M = USV^T$$

$$M^{2k} = US^{2k}U^T$$

$$M^{2k+1} = US^{2k+1}V^T$$

Retrieve anomalies for each cluster

FINISHED

```
%sql
SELECT body, trump, clinton, love, hate, sentiment, emojis FROM tweet
WHERE cluster = ${cluster}
ORDER BY distance DESC
```

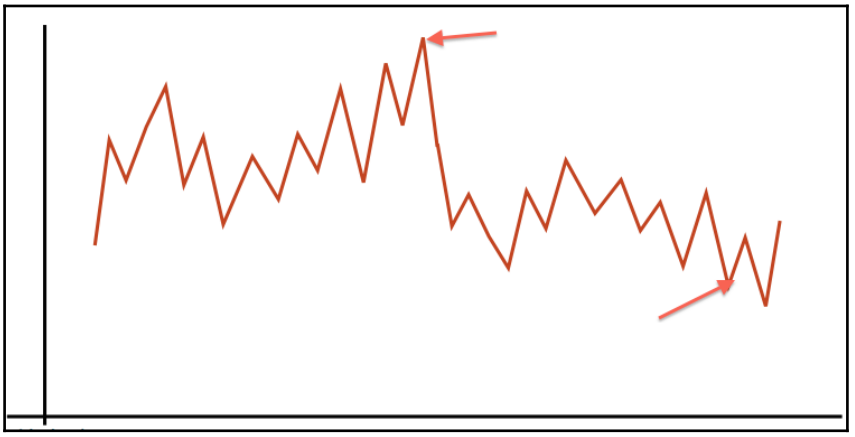
cluster

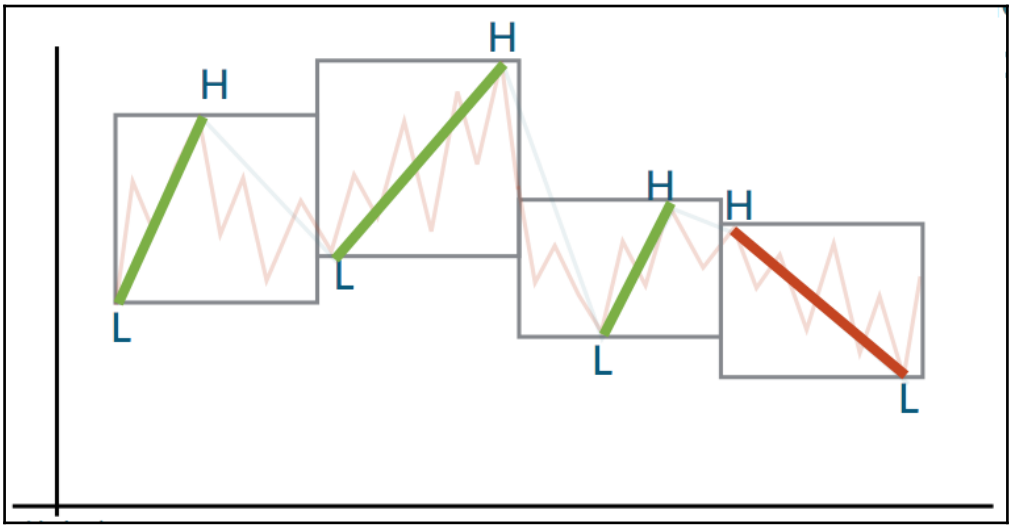
0

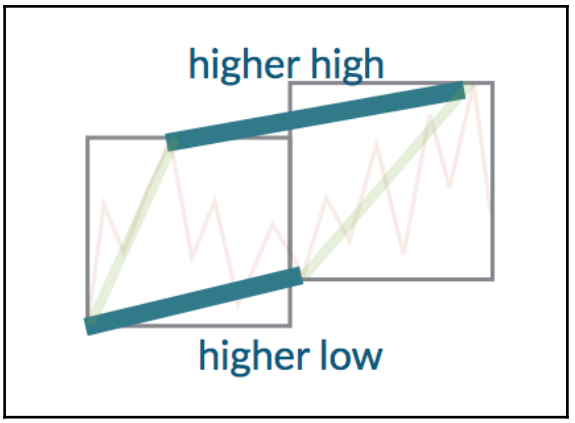
body	trump	clinton	love	hate	sentiment	emojis
please negan please take trump and give we back glenn #trump #thewalkingdead	0.51166	0.45357	0.00572	0.06561	3	WrappedArray(cry, pray)
even we movie title hint the celebrated #trump age ago #achhedin	0.44397	0.37753	-0.01636	-0.00007	3	WrappedArray(smiley)
soydeandaya #soyandayer good morning friend #happyday #electionday	0.4067	0.38863	0.11563	0.06771	3	WrappedArray(grinning)
i love the fact that #toblerone be trend above #uselection where i live #ilovegreatbritain	0.47246	0.39321	0.13134	0.09768	3	WrappedArray(joy)
get we good champagne glass out towel and the brut #maga x enjoy the show #schadenfreude all around the #msm	0.51371	0.39599	0.08136	0.07178	3	WrappedArray(sunglasses, sweat_smile, joy)
well have a permanente patime job for the next year #electionfinalthoughts #electionnight #heartbroken	0.58486	0.52501	0.0372	0.07604	3	WrappedArray(broken_heart)
guy so what be the back up plan if #drumpf win #electionnight #imwithher	0.49184	0.54833	0.1178	0.10663	3	WrappedArray(sob)
to have a super power nation with a leader that do not believe in #climatechange i fear for the future of #biodiversity #uselection	0.26681	0.25984	0.03077	0.07656	1.5	WrappedArray(broken_heart)
nevada latino ur beautiful #bluewave grow into a huge tidal wave #votebluеballot #votehrc #strongertogether	0.52024	0.52676	0.02254	0.02089	3	WrappedArray(clap, clap, ocean, ocean, ocean)

Took 2 sec. Last updated by anonymous at December 18 2016, 1:08:42 AM. (outdated)

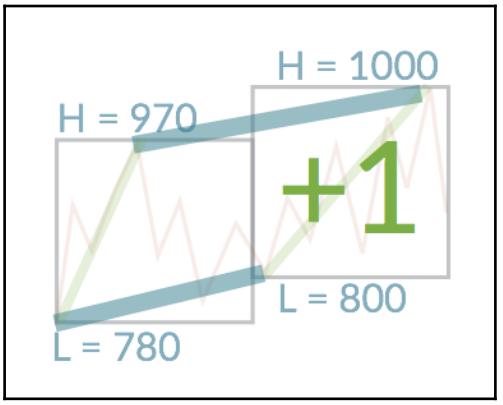
Chapter 12: TrendCalculus

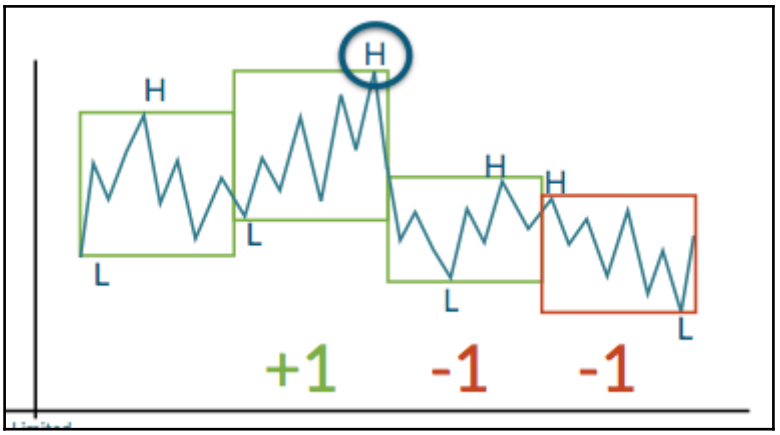


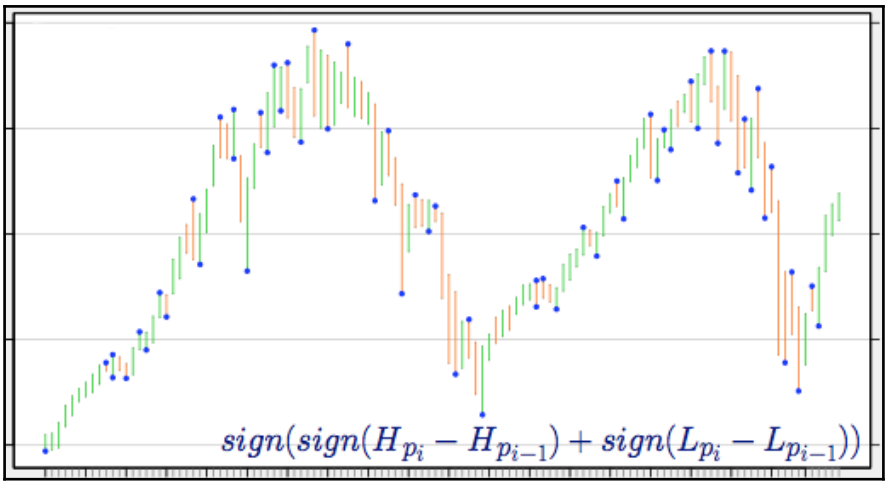


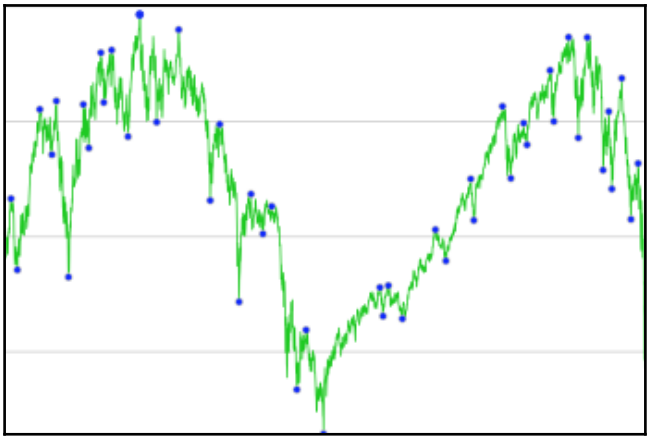


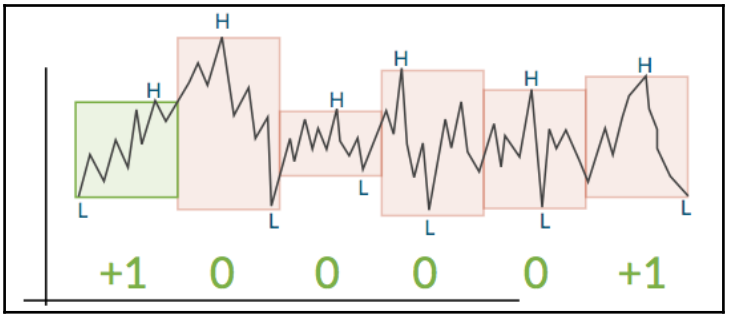
$$sign(sign(H_{p_i} - H_{p_{i-1}}) + sign(L_{p_i} - L_{p_{i-1}}))$$

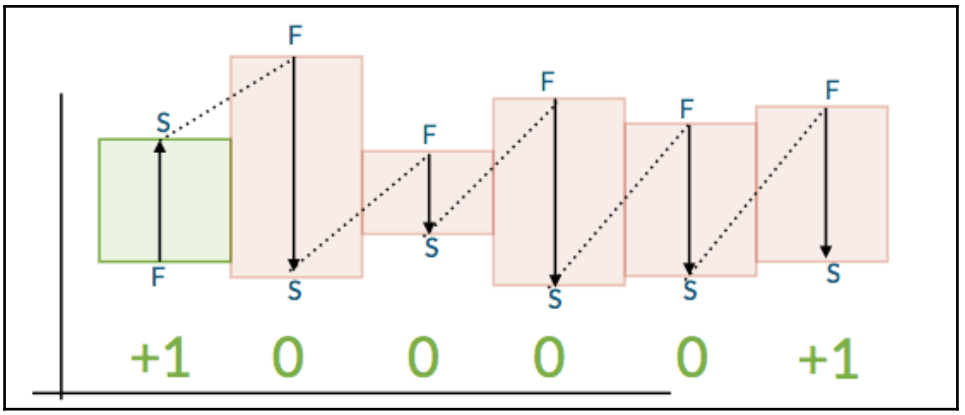


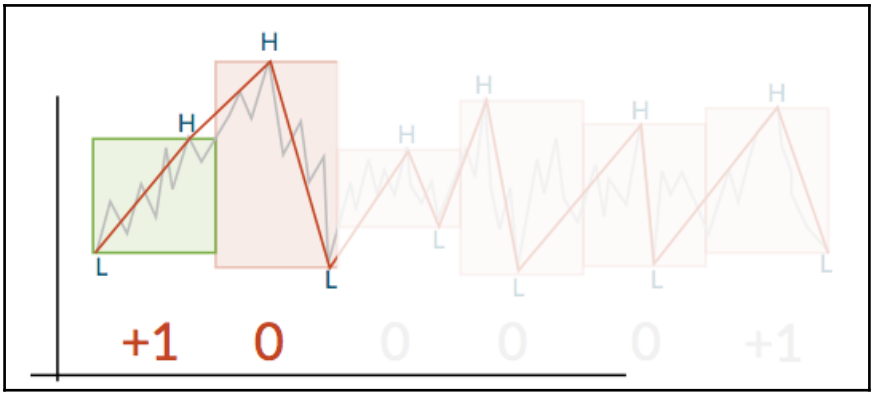

H
H
H
H
L
L
L
L
+1
-1
-1

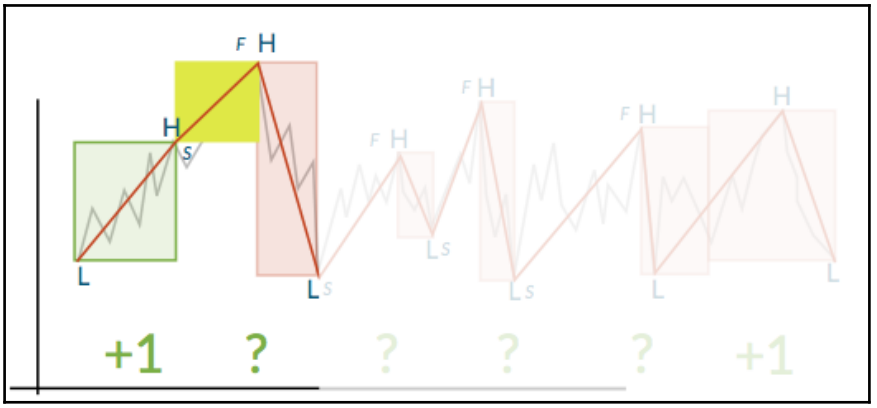

$sign(sign(H_{p_i} - H_{p_{i-1}}) + sign(L_{p_i} - L_{p_{i-1}}))$

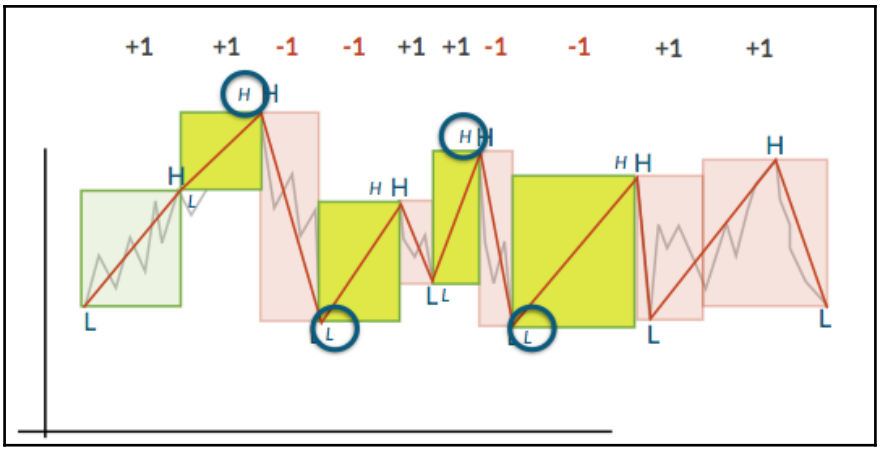


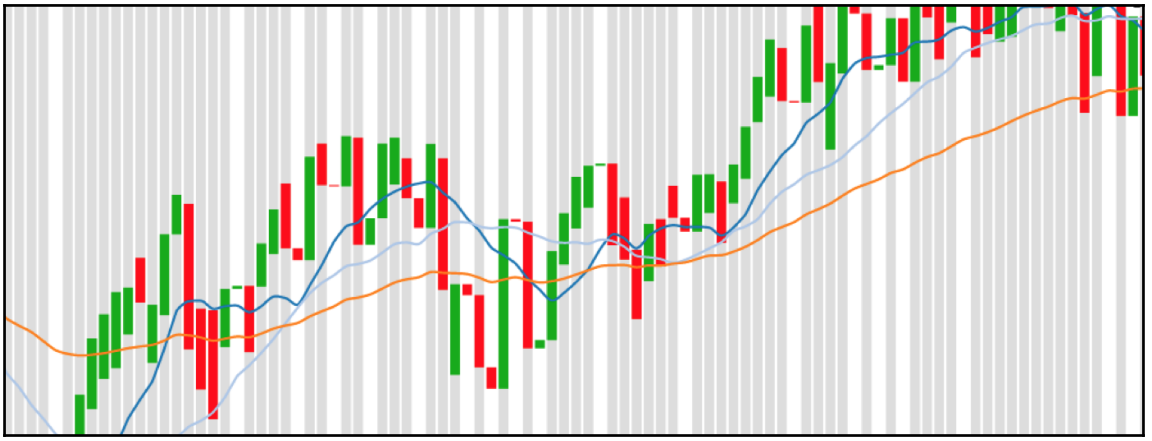

H
H
H
H
H
H
L
L
L
L
L
L
+1
0
0
0
0
+1


F
S
F
S
F
S
F
S
F
S
F
S
+1
0
0
0
0
+1

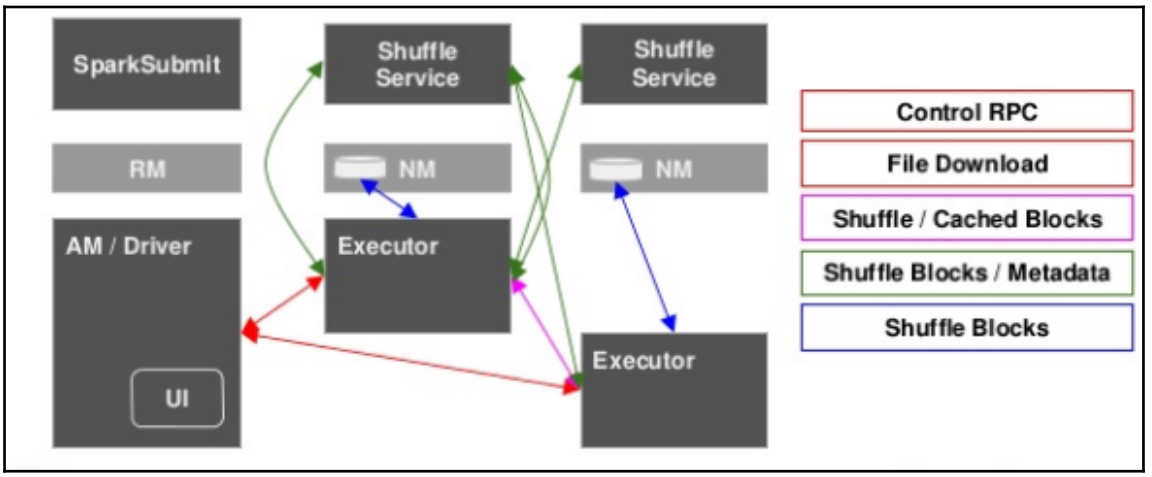

H
H
L
L
H
L
H
L
H
L
+1
0
0
0
0
+1

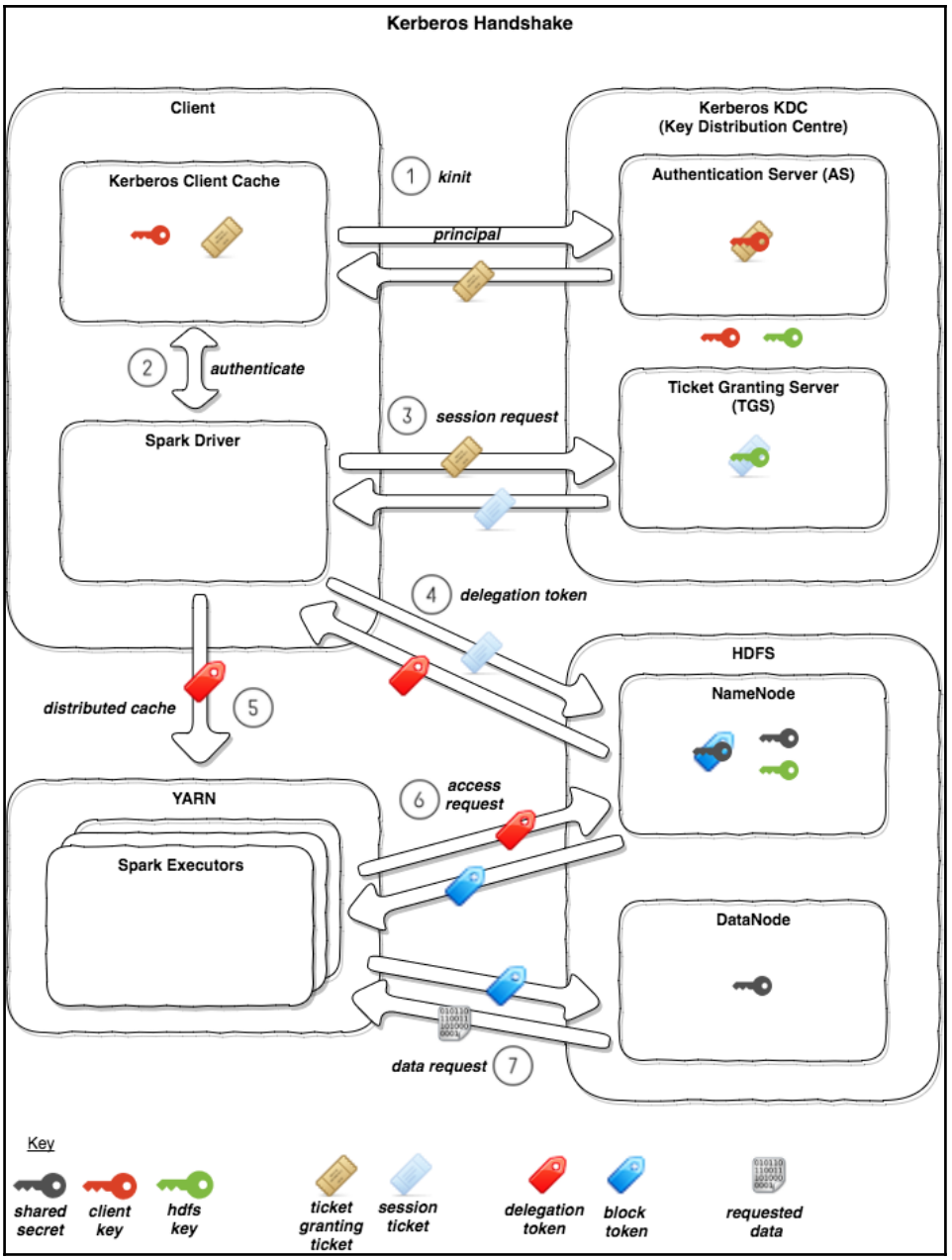

F H
H
S
L
L S
F H
L S
F H
L S
F H
L
H
L
+1
?
?
?
?
+1

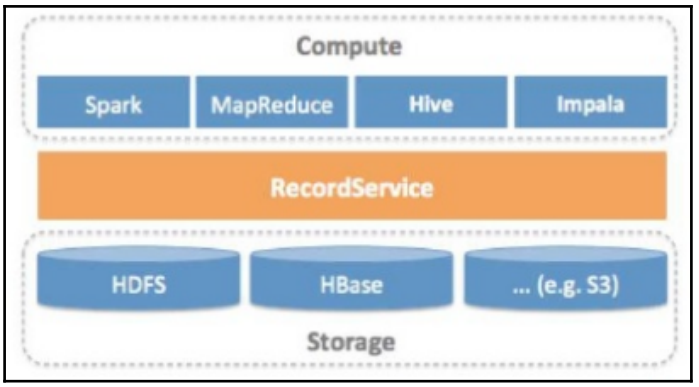

+1 +1 -1 -1 +1 +1 -1 -1 +1 +1
H H H H H H H H H H
L L L L L L L L L L



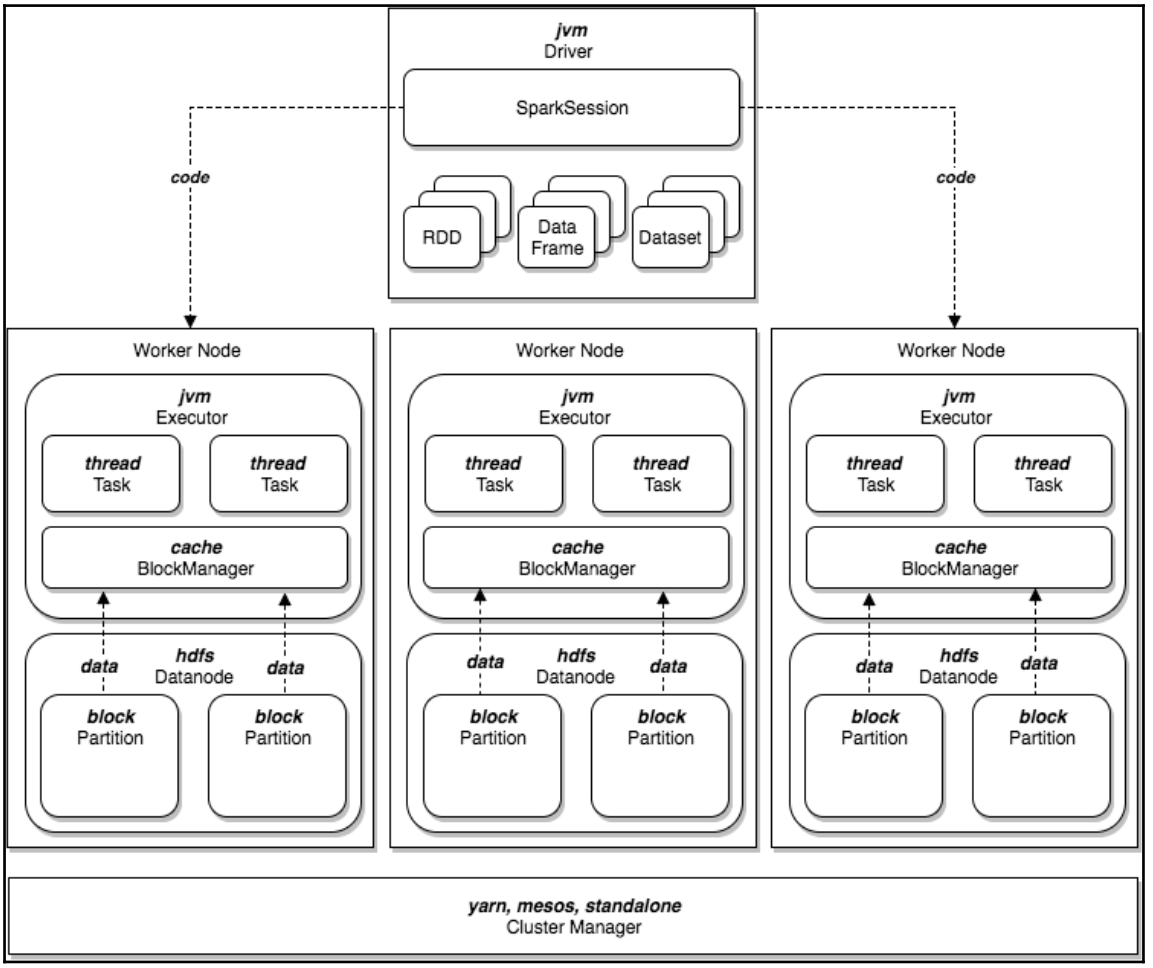
Chapter 13: Secure Data

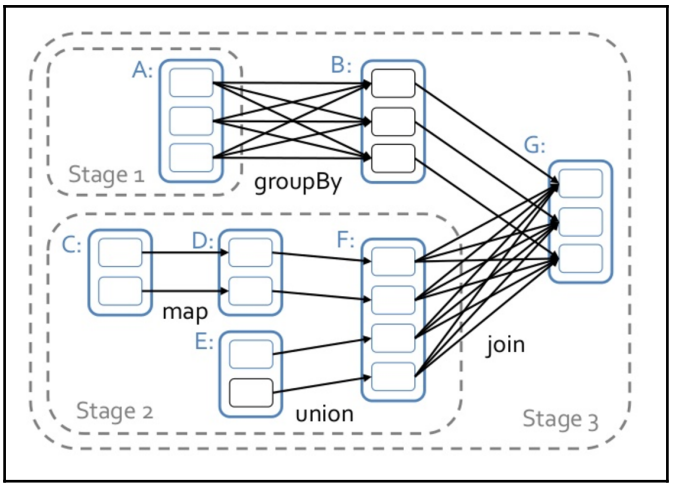


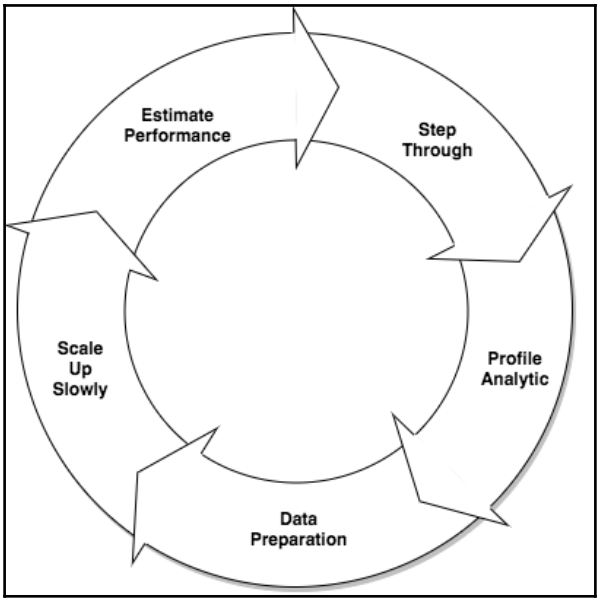

Kerberos Handshake
Client
Kerberos Client Cache
Spark Driver
Kerberos KDC
(Key Distribution Centre)
Authentication Server (AS)
Ticket Granting Server
(TGS)
1 kinit
principal
2 authenticate
3 session request
4 delegation token
distributed cache
5
HDFS
NameNode
DataNode
YARN
Spark Executors
6 access request
data request
7
Key
shared secret
client key
hdfs key
ticket granting ticket
session ticket
delegation token
block token
requested data


Compute
Spark
MapReduce
Hive
Impala
RecordService
HDFS
HBase
... (e.g. S3)
Storage

Chapter 14: Scalable Algorithms




A:
B:
G:
Stage 1
groupBy
C:
D:
F:
map
E:
join
Stage 2
union
Stage 3


Estimate Performance
Step Through
Profile Analytic
Data Preparation
Scale Up Slowly

Summary Metrics for 168 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	71 ms	7 s	12 s	13 s	17 s
Scheduler Delay	1 ms	2 ms	3 ms	3 ms	5 ms
Task Deserialization Time	0 ms	1 ms	1 ms	2 ms	24 ms
GC Time	3 ms	0.2 s	0.4 s	0.4 s	0.6 s
Result Serialization Time	0 ms	0 ms	0 ms	0 ms	1 ms
Getting Result Time	0 ms	0 ms	0 ms	0 ms	0 ms
Peak Execution Memory	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B
Input Size / Records	119.3 KB / 11	16.5 MB / 1397	32.1 MB / 2576	32.1 MB / 2726	35.2 MB / 2869
Shuffle Write Size / Records	1591.0 B / 32	61.9 KB / 6445	98.4 KB / 11595	104.9 KB / 12861	120.2 KB / 17005

Aggregated Metrics by Executor

Executor ID ▲	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Write Size / Records
0	CANNOT FIND ADDRESS	14 min	88	0	88	2035.1 MB / 170712	6.7 MB / 809037
1	CANNOT FIND ADDRESS	14 min	80	0	80	2.0 GB / 174489	6.9 MB / 832305

Tasks (168)

Page: 1 2 >

2 Pages. Jump to 1 . Show 100 items in a page. Go

Index ▲	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	Scheduler Delay	Task Deserialization Time	GC Time	Result Serialization Time	Getting Result Time	Peak Execution Memory	Input Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	336	0	SUCCESS	PROCESS_LOCAL	1 / 192.168.1.67	2016/12/22 06:07:19	12 s	4 ms	24 ms	0.4 s	0 ms	0 ms	0.0 B	32.1 MB / 2644	48 ms	104.9 KB / 15620	
1	337	0	SUCCESS	PROCESS_LOCAL	0 / 192.168.1.67	2016/12/22 06:07:19	7 s	4 ms	24 ms	0.3 s	0 ms	0 ms	0.0 B	15.7 MB / 1286	44 ms	56.4 KB / 6727	
2	338	0	SUCCESS	PROCESS_LOCAL	1 / 192.168.1.67	2016/12/22 06:07:19	12 s	4 ms	23 ms	0.4 s	0 ms	0 ms	0.0 B	32.1 MB / 2627	45 ms	101.6 KB / 13084	
3	339	0	SUCCESS	PROCESS_LOCAL	0 / 192.168.1.67	2016/12/22 06:07:19	7 s	3 ms	24 ms	0.3 s	0 ms	0 ms	0.0 B	17.0 MB / 1397	37 ms	62.1 KB / 7047	
4	340	0	SUCCESS	PROCESS_LOCAL	0 / 192.168.1.67	2016/12/22 06:07:26	12 s	3 ms	2 ms	0.4 s	0 ms	0 ms	0.0 B	32.1 MB / 2720	35 ms	105.4 KB / 13534	
5	341	0	SUCCESS	PROCESS_LOCAL	0 / 192.168.1.67	2016/12/22 06:07:27	6 s	3 ms	1 ms	0.2 s	0 ms	0 ms	0.0 B	15.5 MB / 1306	39 ms	57.5 KB / 6523	
6	342	0	SUCCESS	PROCESS_LOCAL	1 / 192.168.1.67	2016/12/22 06:07:32	13 s	3 ms	1 ms	0.4 s	0 ms	0 ms	0.0 B	32.1 MB / 2717	35 ms	110.0 KB / 13248	

Spark 2.0.1 Jobs Stages Storage Environment Executors SQL **Spark shell** application UI

Thread dump for executor 1

Updated at 2016/12/22 06:00:41

Expand All

Search:

Thread ID	Thread Name	Thread State
58	Executor task launch worker-0	RUNNABLE
59	Executor task launch worker-1	RUNNABLE

```
org.json.simple.JSONValue.escape(Unknown Source)
org.json.simple.JSONValue.escape(Unknown Source)
org.json.simple.JSONValue.toJSONString(Unknown Source)
org.json.simple.JSONObject.toJSONString(Unknown Source)
org.json.simple.JSONObject.toJSONString(Unknown Source)
org.json.simple.JSONObject.toJSONString(Unknown Source)
org.json.simple.JSONValue.toJSONString(Unknown Source)
org.json.simple.JSONArray.toJSONString(Unknown Source)
org.json.simple.JSONArray.toJSONString(Unknown Source)
org.json.simple.JSONValue.toJSONString(Unknown Source)
org.json.simple.JSONObject.toJSONString(Unknown Source)
org.json.simple.JSONObject.toJSONString(Unknown Source)
org.json.simple.JSONObject.toJSONString(Unknown Source)
io.gzet.utils.gdelt.GkgJSON.parseV2ToJson(GkgJSON.java:248)
io.gzet.utils.spark.gdelt.GKGParser$.toJsonGKGV2(GKGParser.scala:23)
$line20.$read$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$anonfun$1.apply(<console>:31)
$line20.$read$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$iw$$anonfun$1.apply(<console>:31)
scala.collection.Iterator$$anon$11.next(Iterator.scala:409)
scala.collection.Iterator$$anon$11.next(Iterator.scala:409)
scala.collection.Iterator$$anon$12.nextCur(Iterator.scala:434)
scala.collection.Iterator$$anon$12.hasNext(Iterator.scala:440)
scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:408)
org.apache.spark.shuffle.sort.BypassMergeSortShuffleWriter.write(BypassMergeSortShuffleWriter.java:147)
org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:79)
org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:47)
org.apache.spark.scheduler.Task.run(Task.scala:86)
org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:274)
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
java.lang.Thread.run(Thread.java:745)
```

31	dispatcher-event-loop-0	WAITING

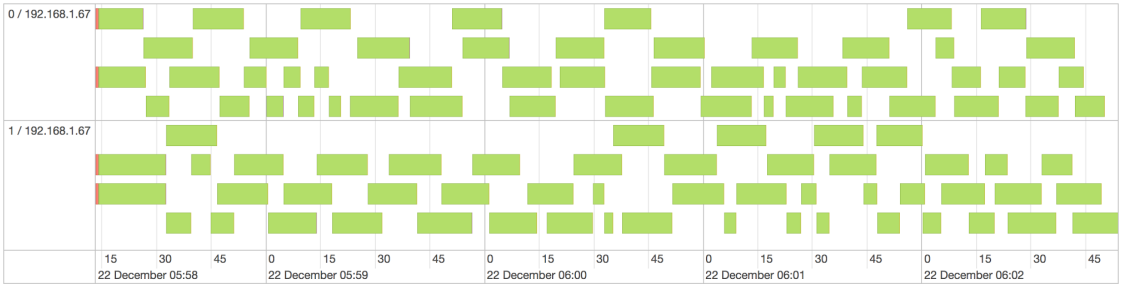
Spark 2.0.1 Jobs Stages Storage Environment Executors SQL **Spark shell** application UI

Details for Stage 0 (Attempt 0)

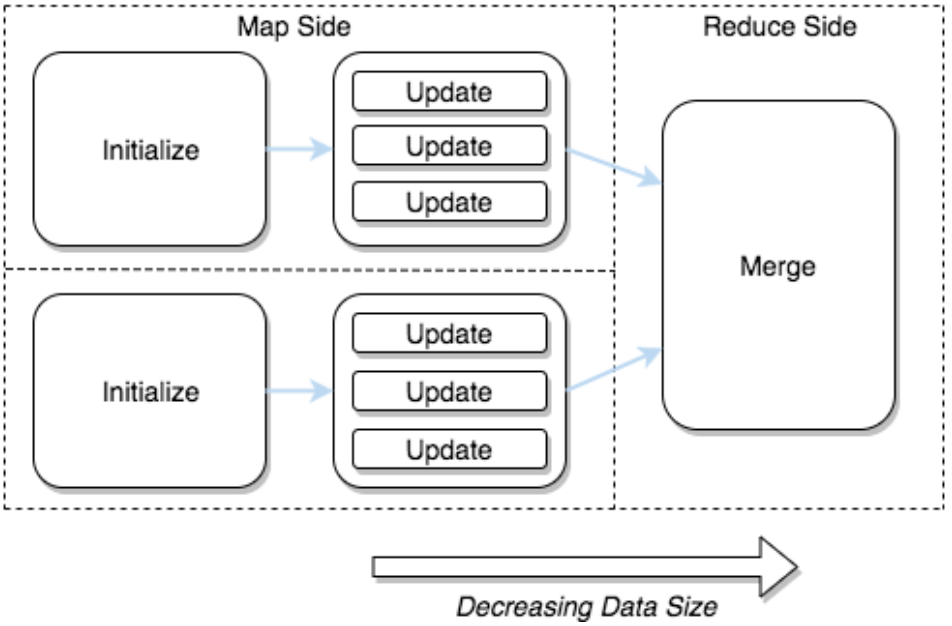
Total Time Across All Tasks: 18 min
Locality Level Summary: Process local: 114
Input Size / Records: 2.6 GB / 216023
Shuffle Write: 8.7 MB / 1070998

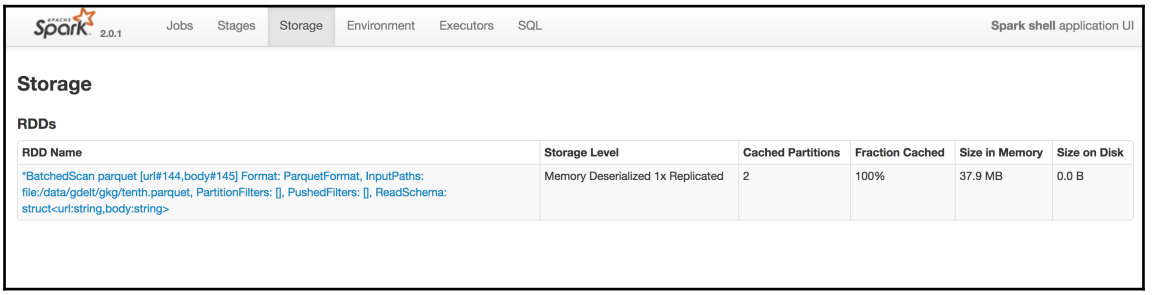
▸ DAG Visualization
▸ Show Additional Metrics
▾ Event Timeline
☐ Enable zooming

Scheduler Delay
Task Deserialization Time
Shuffle Read Time
Executor Computing Time
Shuffle Write Time
Result Serialization Time
Getting Result Time



Summary Metrics for 110 Completed Tasks


Map Side
Reduce Side
Initialize
Update
Update
Update
Merge
Initialize
Update
Update
Update
Decreasing Data Size


Spark 2.0.1
Jobs
Stages
Storage
Environment
Executors
SQL
Spark shell application UI
Storage
RDDs
RDD Name
Storage Level
Cached Partitions
Fraction Cached
Size in Memory
Size on Disk
*BatchedScan parquet [url#144,body#145] Format: ParquetFormat, InputPaths: file:/data/gdelt/gkg/tenth.parquet, PartitionFilters: [], PushedFilters: [], ReadSchema: struct<url:string,body:string>
Memory Deserialized 1x Replicated
2
100%
37.9 MB
0.0 B