Search Engine using Apache Spark

Muad Abu-Ata

1. Summary:

The system implemented using Apache Spark for scalability and query speed.

The system first builds an inverted index of the document collection.

The inverted index contains the word, document and TFIDF score for each word in each document.

The system scores the relevant documents for a given query using tfidf weights.

The system returns the top N relevant documents using a relevance scoring function.

2. Implementation:

The system implemented in Spark using two modules. One for building the inverted index (buildIndex.py in Figure 1) and the second for querying the system (query.py in Figure 2). The architecture of the system is shown in the diagram below for building the index file and querying.

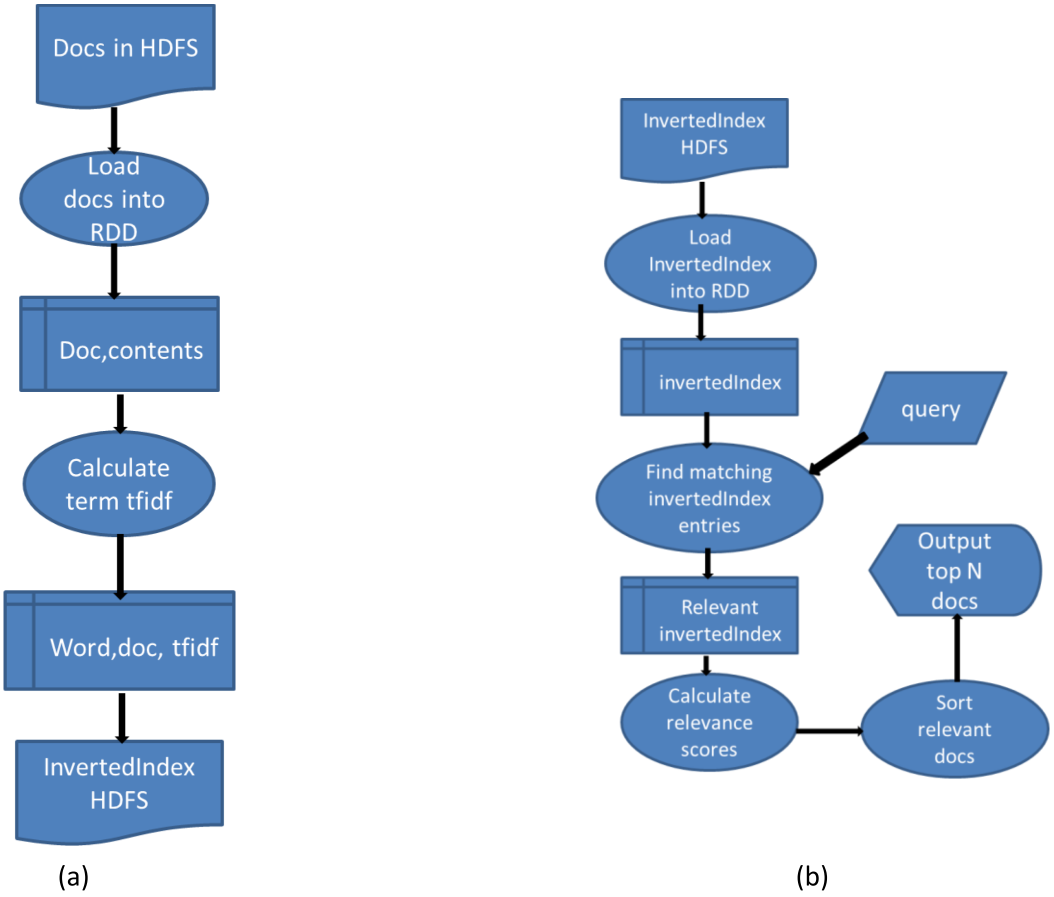


Fig. 1 Architecture of the system two modules. Flowchart (a) for building the inverted index module. Flowchart (b) for querying module.

2.1 Building the inverted Index:

Please refer to the source code scripts for better line by line documentation of the scripts.

First, the collection of documents is loaded into RDD using whole wholeTextFiles. Each RDD element is a pair of the document name and its contents (doc, contents). (Line 4 in Figure 2)

Preprocess the documents RDD by changing the case to lowercase and remove stop words. (Lines 6-9 in Figure 2)

Calculating the tf weight of each word. This is done by counting the number of times each word appearing in each document. (Lines 10-12 in Figure 2)

Calculating document frequency (df) weight for each word in the collection. (Line 13 in Figure 2)

Combining tf and df into tfidf for each word per document. (Line 14-16 in Figure 2)

The inverted index is stored on the cluster HDFS. (Lines 17-18 in Figure 2)

```
from pyspark import SparkConf, SparkContext
import math
def main(sc):
# reading input from HDFS. Reading the whole directory containing the documents. Assuimh each replace is small enough.
    textFile = sc.wholeTextFiles("/user/root/final/bbcsport/athletics")
    N=textFile.count() # total number of documents
    rawWordList1 = textFile.flatMap(lambda (doc,contents): [(word.strip(".;,\'\"?!()").lower(),doc) for word in contents.split()])
    rawWordList2 = rawWordList1.filter(lambda (word,doc):word!='')
    stopWords=sc.textFile("file:///root/lab/stopWords.txt").collect() # reading stop words from local file
    wordList=rawWordList2.filter(lambda (word,doc):word not in stopWords ) # removing stop words
 # mapping/counting (1) to each word occurrence in a file
    wordCount = wordList.map(lambda (word,doc):((word,doc),1))
#summming same word counts within a file i.e. calculating TF. TF is actually the inverted index ((word,file),tf)
    TF = wordCount.reduceByKey(lambda v1, v2: v1+v2)
    TF.persist()
# counting the document frequency of each word. Result RDD is ((word,file),df)
    DF=TF.map(lambda ((word,doc),v): (word,1)).reduceByKey(lambda v1, v2: v1+v2)
    TFWord=TF.map(lambda((word,doc),v):(word,(doc,v)))
    TF.unpersist()
    invertedIndex=TFWord.join(DF).map(lambda(word,((doc,tf),df)):(word,(doc,tf*math.log10(N/float(df)))))
    invertedIndexOutFormat= invertedIndex.map(lambda (word,(doc,tfidf)):word+" "+doc+" "+str(tfidf))
    invertedIndexOutFormat.saveAsTextFile("/user/root/final/inverted.txt")
if __name__  == "__main__":
    conf = SparkConf().setAppName("buildIndex")
    sc = SparkContext(conf = conf)
    main(sc)
    sc.stop()
#spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m buildIndex.py
```

Fig. 2: buildIndex.py. For full documentation of the code, Please refer to the source code file.

2.2 Querying the system:

First the inverted index is loaded from the HDFS into RDD. (Lines 9-11 in Figure 3)

Find matching elements of the inverted index. Each matching element is a triple (word,doc,tfidf) with word equal any query word (term.) This would result is those documents containing at least one query term. (Line 15 in Figure 3)

For each document in the matching set, calculate the relevance score of each document (Lines 16-18 in Figure 3) according to the following formula:

$$Score(Q, Doc) = \frac{|Q \cap Doc|}{|Q|} \sum_{q \in Q} TFIDF(q, Doc)$$

Retrieve the top N relevant documents from those matching the query. (Line 19 in Figure 3)

```
from pyspark import SparkConf, SparkContext
import math
import sys
reload(sys)
sys.setdefaultencoding('UTF8')
def main(sc,argv):
  query=argv[0:len(argv)-2]
  n=int(argv[len(argv)-1])
# reading the inverted index from HDFS into RDD as string
  rawInvertedIndex = sc.textFile("/user/root/final/inverted.txt")
  inverted1=rawInvertedIndex.map(lambda line:line.split())
  invertedIndex=inverted1.map(lambda (word,doc,tfidf): (word,doc,float(tfidf)))
  invertedIndex.persist()
  for i, qWord in enumerate(query):
     query[i]=qWord.lower().strip(".:;,\'\"?!()")
  matchDocs=invertedIndex.filter(lambda (word,doc,tfidf):word in query)
  matchDocsScore1=matchDocs.map(lambda(word,doc,tfidf):(doc,tfidf)).reduceByKey(lambda
tfidf1,tfidf2:tfidf1+tfidf2)
  matchDocsScore2=matchDocs.map(lambda(word,doc,tfidf):(doc,1)).reduceByKey(lambda v1,v2:v1+v2)
  matchDocScore=matchDocsScore1.join(matchDocsScore2).map(lambda
(doc,(score1,score2)):(doc,score1*score2/float(len(query))))
  rerteivedDocs=matchDocScore.takeOrdered(n,key= lambda (doc,score):-score)
  print(rerteivedDocs)
  for (doc,score) in rerteivedDocs:
     docText=sc.textFile(doc).collect()
     print "%s %s"% (doc,score)
     print('\n'.join(docText))
     print"------------------------\n"
  invertedIndex.unpersist()
if __name__  == "__main__":
  conf = SparkConf().setAppName("query")
  sc = SparkContext(conf = conf)
  main(sc,sys.argv[1:])
  sc.stop()
#spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-
memory 512m query.py query n
```

Fig. 3: query.py. . For full documentation/comments of the code, Please refer to the source code

3. Results:

3.1 Query 1: Olympic Medal

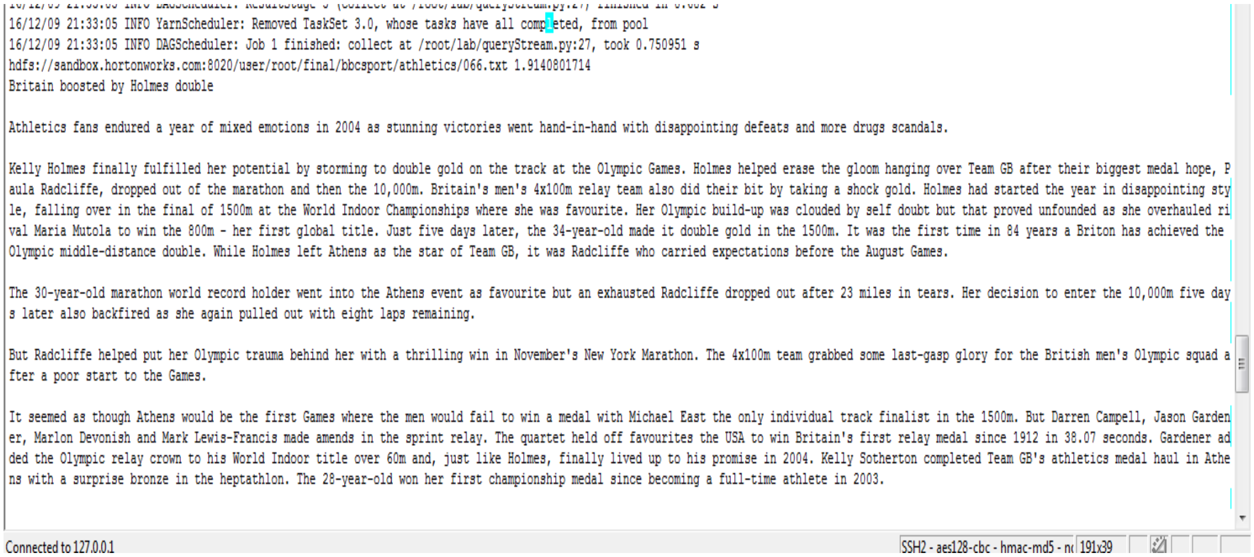
3.1.1 Query 1: Olympic Medal N=1

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m query.py Olympic Medal 1

Results:

hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/066.txt 1.9140801714

File: …/athletics/066.txt is too big to put here in the document.



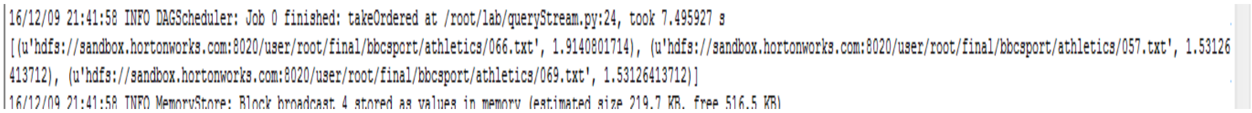
```
16/12/09 21:33:05 INFO DAGScheduler: ResultStage 3 (collect at /root/lab/queryStream.py:27) finished in 0.602 s
16/12/09 21:33:05 INFO YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/12/09 21:33:05 INFO DAGScheduler: Job 1 finished: collect at /root/lab/queryStream.py:27, took 0.750951 s
hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/066.txt 1.9140801714
Britain boosted by Holmes double

Athletics fans endured a year of mixed emotions in 2004 as stunning victories went hand-in-hand with disappointing defeats and more drugs scandals.

Kelly Holmes finally fulfilled her potential by storming to double gold on the track at the Olympic Games. Holmes helped erase the gloom hanging over Team GB after their biggest medal hope, P
aula Radcliffe, dropped out of the marathon and then the 10,000m. Britain's men's 4x100m relay team also did their bit by taking a shock gold. Holmes had started the year in disappointing sty
le, falling over in the final of 1500m at the World Indoor Championships where she was favourite. Her Olympic build-up was clouded by self doubt but that proved unfounded as she overhauled ri
val Maria Mutola to win the 800m - her first global title. Just five days later, the 34-year-old made it double gold in the 1500m. It was the first time in 84 years a Briton has achieved the
Olympic middle-distance double. While Holmes left Athens as the star of Team GB, it was Radcliffe who carried expectations before the August Games.

The 30-year-old marathon world record holder went into the Athens event as favourite but an exhausted Radcliffe dropped out after 23 miles in tears. Her decision to enter the 10,000m five day
s later also backfired as she again pulled out with eight laps remaining.

But Radcliffe helped put her Olympic trauma behind her with a thrilling win in November's New York Marathon. The 4x100m team grabbed some last-gasp glory for the British men's Olympic squad a
fter a poor start to the Games.

It seemed as though Athens would be the first Games where the men would fail to win a medal with Michael East the only individual track finalist in the 1500m. But Darren Campell, Jason Garden
er, Marlon Devonish and Mark Lewis-Francis made amends in the sprint relay. The quartet held off favourites the USA to win Britain's first relay medal since 1912 in 38.07 seconds. Gardener ad
ded the Olympic relay crown to his World Indoor title over 60m and, just like Holmes, finally lived up to his promise in 2004. Kelly Sotherton completed Team GB's athletics medal haul in Athe
ns with a surprise bronze in the heptathlon. The 28-year-old won her first championship medal since becoming a full-time athlete in 2003.
```

3.1.2. Query 1: Olympic Medal N=3

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m query.py Olympic Medal 3

Result:

[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/066.txt', 1.9140801714), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/057.txt', 1.53126413712), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/069.txt', 1.53126413712)]



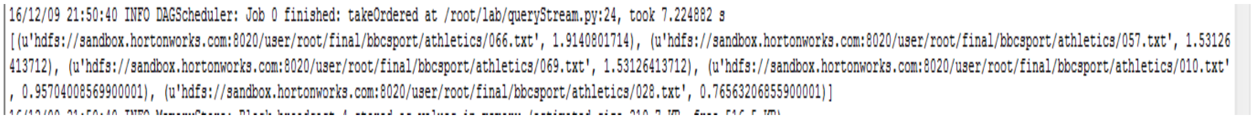
```
16/12/09 21:41:58 INFO DAGScheduler: Job 0 finished: takeOrdered at /root/lab/queryStream.py:24, took 7.495927 s
[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/066.txt', 1.9140801714), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/057.txt', 1.53126
413712), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/069.txt', 1.53126413712)]
16/12/09 21:41:58 INFO MemoryStore: Block broadcast 4 stored as values in memory (estimated size 219.7 KB, free 516.5 KB)
```

2.1.3. Query 1: Olympic Medal N=5

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m query.py Olympic Medal 5

Result:

[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/066.txt', 1.9140801714), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/057.txt', 1.53126413712), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/069.txt', 1.53126413712), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/010.txt', 0.95704008569900001), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/028.txt', 0.76563206855900001)]



```
16/12/09 21:50:40 INFO DAGScheduler: Job 0 finished: takeOrdered at /root/lab/queryStream.py:24, took 7.224882 s
[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/066.txt', 1.9140801714), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/057.txt', 1.53126
413712), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/069.txt', 1.53126413712), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/010.txt'
, 0.95704008569900001), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/028.txt', 0.76563206855900001)]
```

3.2 Query 2: doping investigation marion jones

3.2.1 Query 2: doping investigation marion jones N=1

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m query.py doping investigation marion jones 1

Result:

hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/088.txt
9.35248234875

3.2.2 Query 2: doping investigation marion jones N=3

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m query.py doping investigation marion jones 3

Result:

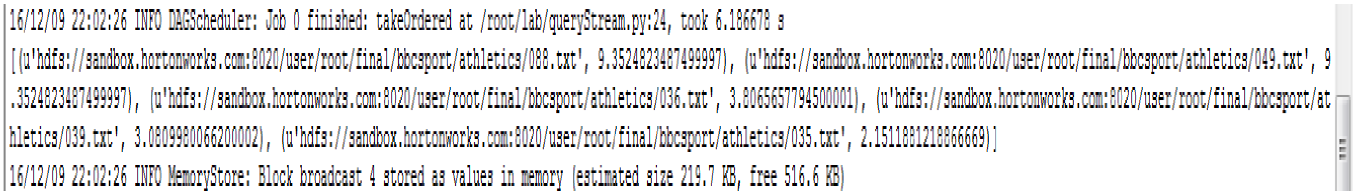
[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/088.txt', 9.3524823487499997), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/049.txt', 9.3524823487499997), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/036.txt', 3.8065657794500001)]

3.2.3 Query 2: doping investigation marion jones N=5

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m query.py doping investigation marion jones 5

Result:

[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/088.txt', 9.3524823487499997), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/049.txt', 9.3524823487499997), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/036.txt', 3.8065657794500001), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/039.txt', 3.0809980066200002), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/035.txt', 2.1511881218866669)]



```
16/12/09 22:02:26 INFO DAGScheduler: Job 0 finished: takeOrdered at /root/lab/queryStream.py:24, took 6.186678 s
[(u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/088.txt', 9.3524823487499997), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/049.txt', 9
.3524823487499997), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/036.txt', 3.8065657794500001), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/at
hletics/039.txt', 3.0809980066200002), (u'hdfs://sandbox.hortonworks.com:8020/user/root/final/bbcsport/athletics/035.txt', 2.1511881218866669)]
16/12/09 22:02:26 INFO MemoryStore: Block broadcast 4 stored as values in memory (estimated size 219.7 KB, free 516.6 KB)
```

File athletics/088.txt: (score 9.3524823487499997)

Jones files lawsuit against Conte

Marion Jones has filed a lawsuit for defamation against Balco boss Victor Conte following his allegations that he gave her performance-enhancing drugs.

The Sydney Olympic gold medallist says Conte damaged her reputation and she is seeking $25m (Â£13m) in the suit. Conte, whose company is at the centre of a doping investigation, made the claims in a US television programme. He and three others were indicted in February by a federal grand jury for a variety of alleged offences. In an email to the Associated Press on Wednesday, Conte said: "I stand by everything I said". Jones won three gold medals and two bronzes in Sydney in 2000. Her lawsuit, filed in the US District Court in San Francisco, said the sprinter had passed a lie detector test and that she "has never taken banned performance-enhancing drugs". Conte's statements, the suit added, were "false and malicious". After the ABC television program earlier this month, Jones' lawyer Richard Nicholls said: "Marion has steadfastly maintained her position throughout: she has never, ever used performance-enhancing drugs. "Victor Conte is a man facing a 42-count federal indictment, while Marion Jones is one of America's most decorated female athletes. Mr Conte's statements have been wildly contradictory. "Mr Conte chose to make unsubstantiated allegations on television, while Marion Jones demanded to take and then passed a lie detector examination.

"Mr Conte is simply not credible. We challenge him to submit to the same lie detector procedure that Marion Jones passed." The sport's ruling body, the IAAF, is taking a cautious approach to Conte's allegations but contacted the US Anti-Doping Agency. Communications director Nick Davies said the IAAF would seek to contact Conte "for further information". But Davies stressed it would be up to the American authorities to decide whether they will take action against Jones in light of Conte's television interview and the world governing body would monitor the situation closely. "If it is felt there is case to answer, it would be for its national governing body (USA Track and

Field) to take the appropriate disciplinary action," he added. "The US Anti-Doping Agency has proved itself to be very diligent in its anti-doping war. "And I am sure, like ourselves, they will be watching the television programme with great interest." Jones, who is under investigation for steroid use by the US Anti-Doping Agency, has continually denied ever taking illegal substances since being investigated in the Balco scandal, although she praised a zinc supplement Conte marketed. Jones, who did not win any medals in Athens in August, has never failed a drugs test. Meanwhile, Conte, who has been charged along with three other men of distributing illegal steroids and money laundering, is due to face trial in March.

File athletics/049.txt: (score 9.3524823487499997)

Jones files Conte lawsuit

Marion Jones has filed a lawsuit for defamation against Balco boss Victor Conte following his allegations that he gave her performance-enhancing drugs.

The Sydney Olympic gold medallist says Conte damaged her reputation and she is seeking $25m (Â£13m) in the suit. Conte, whose company is at the centre of a doping investigation, made the claims in a US television programme. He and three others were indicted in February by a federal grand jury for a variety of alleged offences. In an email to the Associated Press on Wednesday, Conte said: "I stand by everything I said". Jones won three gold medals and two bronzes in Sydney in 2000. Her lawsuit, filed in the US District Court in San Francisco, said the sprinter had passed a lie detector test and that she "has never taken banned performance-enhancing drugs". Conte's statements, the suit added, were "false and malicious". After the ABC television program earlier this month, Jones' lawyer Richard Nicholls said: "Marion has steadfastly maintained her position throughout: she has never, ever used performance-enhancing drugs. "Victor Conte is a man facing a 42-count federal indictment, while Marion Jones is one of America's most decorated female athletes. Mr Conte's statements have been wildly contradictory. "Mr Conte chose to make unsubstantiated allegations on television, while Marion Jones demanded to take and then passed a lie detector examination.

"Mr Conte is simply not credible. We challenge him to submit to the same lie detector procedure that Marion Jones passed." The sport's ruling body, the IAAF, is taking a cautious approach to Conte's allegations but contacted the US Anti-Doping Agency. Communications director Nick Davies said the IAAF would seek to contact Conte "for further information". But Davies stressed it would be up to the American authorities to decide whether they will take action against Jones in light of Conte's television interview and the world governing body would monitor the situation closely. "If it is felt there is case to answer, it would be for its national governing body (USA Track and Field) to take the appropriate disciplinary action," he added. "The US Anti-Doping Agency has proved itself to be very diligent in its anti-doping war. "And I am sure, like ourselves, they will be watching the television programme with great interest." Jones, who is under investigation for steroid use by the US Anti-Doping Agency, has continually denied ever taking illegal substances since being investigated in the Balco scandal, although she praised a zinc supplement Conte marketed. Jones, who did not win any medals in Athens in August, has never failed a drugs test. Meanwhile, Conte, who has been charged along with three other men of distributing illegal steroids and money laundering, is due to face trial in March.

File athletics/036.txt: (score 3.80656577945)

Jones doping probe begins

An investigation into doping claims against Marion Jones has been opened by the International Olympic Committee.

IOC president Jacques Rogge has set up a disciplinary body to look into claims by Victor Conte, of Balco Laboratories. Jones, who says she is innocent, could lose all her Olympic medals after Conte said he gave her performance-enhancing drugs before the Sydney Olympics. But Rogge said it was too early to speculate about that, hoping only that "the truth will emerge".

Any decision on the medals would be taken by the IOC's executive board and could hinge on interpretation of a rule stating that Olympic decisions can only be challenged within three years of the Games closing. The Sydney Olympics ended more than four years ago, but World Anti-Doping Agency chief Dick Pound said the rule may not apply because the allegations are only coming out now. "We will find a way to deal with that," Pound said. In a statement released through her attorney Rich Nichols, Jones repeated her innocence and vowed she would be cleared. "Victor Conte's allegations are not true and the truth will be revealed for the world to see as the legal process moves forward," she said. "Conte is someone who is under federal indictment and has a record of issuing contradictory, inconsistent statements."

File athletics/039.txt: (score 3.0809980066200002)

Jones medals 'must go if guilty'

World Anti-Doping Agency (WADA) chief Dick Pound says Marion Jones should be stripped of all her medals if found guilty of taking banned substances.

Victor Conte, of Balco Laboratories, claims the American sprinter regularly used drugs to enhance her performance. "If she is found guilty she should be stripped of all her medals and banned for two years," said Pound. Asked if there was a timescale as to what medals could be taken, Pound said: "That is not an issue at all." However, under International Olympic Committee (IOC) rules, athletes can only be stripped of their medals if caught within three years of the event. Jones, who won five medals at the 2000 Olympics, denies using drugs and says she will take legal action over Conte's allegations. Balco Laboratories is the firm at the centre of a wide-reaching investigation into doping in the US. Pound continued: "If she has indeed taken drugs it is going to be a big disappointment for a lot of people."

File athletics/035.txt: (score 2.1511881218866669)

Collins banned in landmark case

Sprinter Michelle Collins has received an eight-year ban for doping offences after a hearing at the North American Court of Arbitration for Sport (CAS).

America's former world indoor 200m champion is the first athlete to be suspended without a positive drugs test or an admission of drugs use. Collins' ban is a result of her connection to the federal inquiry into the Balco doping scandal. The 33-year-old was found guilty of using performance-enhancing drugs. The US Anti-Doping Agency (USADA) decided to press charges against Collins in the summer. The sprinter has consistently protested her innocence but the CAS has upheld USADA's findings. "The USADA has proved, beyond a reasonable doubt, that Collins took EPO, the testosterone/epitestosterone cream and THG," said a CAS statement. "Collins used these substances to enhance her performance and elude the drug testing that was available at the time." So far a total of 13 athletes have been sanctioned for violations involving drugs associated with the Balco doping scandal. World record holder Tim Montgomery is also facing a lifetime ban after being charged by the USADA. His hearing before the CSA has been rescheduled for June next year.

Drug enforcement chiefs in the US have vowed to crack down on cheats. USADA chief executive officer Terry Madden said the action taken against Collins was further proof of that. "The CAS panel's decision confirms that those who violate the rules will be sanctioned as part of USADA's ongoing efforts to protect the rights of the overwhelming majority of US athletes that compete drug-free," said Madden. The USADA has built its cases on verbal evidence given to the federal investigation into Balco rather than test results. The San Francisco-based Balco laboratory faces steroid distribution and money laundering charges. The trial is expected to open next March.

4. Performance:

It took around 1 minute to execute the query scripts. This time includes preparing/ loading the program and spark environment preparations and loading the index file. Query retrieval results are actually faster (real time) especially if the query module is implemented using spark stream such that environment preparation and loading the inverted index is done once.

5. Why Spark?

I used spark to implement the search engine as it provides scalability and cluster computing framework to work on the data in memory in parallel. It beats other platforms as Map-Reduce, Hive,Pig,..etc. as it implement in-memory data transformation streaming /flow instead of storing intermediate results on disks used by Map-Reduce.