

Spark and Elasticsearch for real-time data analysis

Costin Leau, @costinl



What is Elasticsearch?

Scalable, real-time search and analytics engine

Open-source (on Github, Apache 2 License)



WIKIMEDIA
COMMONS

WIKIPEDIA

The Free Encyclopedia



WIKIDATA



WIKIMEDIA
FOUNDATION



Wikiquote



WIKIVERSITY

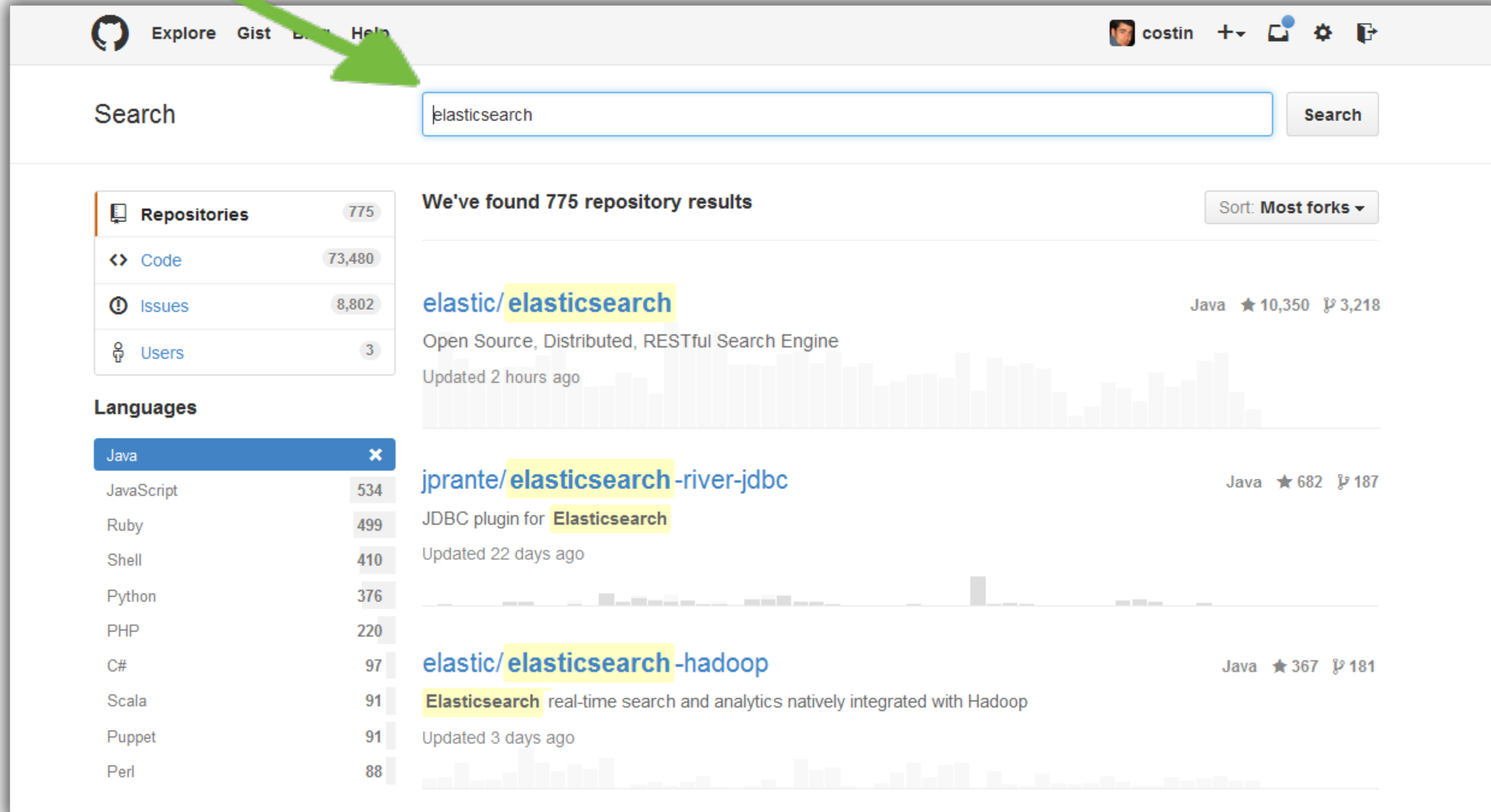


WIKINEWS



WIKIBOOKS
Open books for an open world

Unstructured search



The screenshot shows the GitHub search interface. A green arrow points to the search bar containing the text 'elasticsearch'. The search results are sorted by 'Most forks' and show 775 repository results. The top results are:

- elastic/elasticsearch**: Open Source, Distributed, RESTful Search Engine. Updated 2 hours ago. Java, 10,350 stars, 3,218 forks.
- jprante/elasticsearch-river-jdbc**: JDBC plugin for Elasticsearch. Updated 22 days ago. Java, 682 stars, 187 forks.
- elastic/elasticsearch-hadoop**: Elasticsearch real-time search and analytics natively integrated with Hadoop. Updated 3 days ago. Java, 367 stars, 181 forks.

On the left sidebar, the 'Languages' section is visible, with 'Java' selected. Other languages listed include JavaScript (534), Ruby (499), Shell (410), Python (376), PHP (220), C# (97), Scala (91), Puppet (91), and Perl (88).

Sorting

The screenshot shows the GitHub search interface. At the top, the search bar contains 'elasticsearch' and the 'Search' button is visible. A green arrow points from the search bar to the 'Sort: Most forks' dropdown menu. The search results are sorted by 'Most forks' and show three repository results:

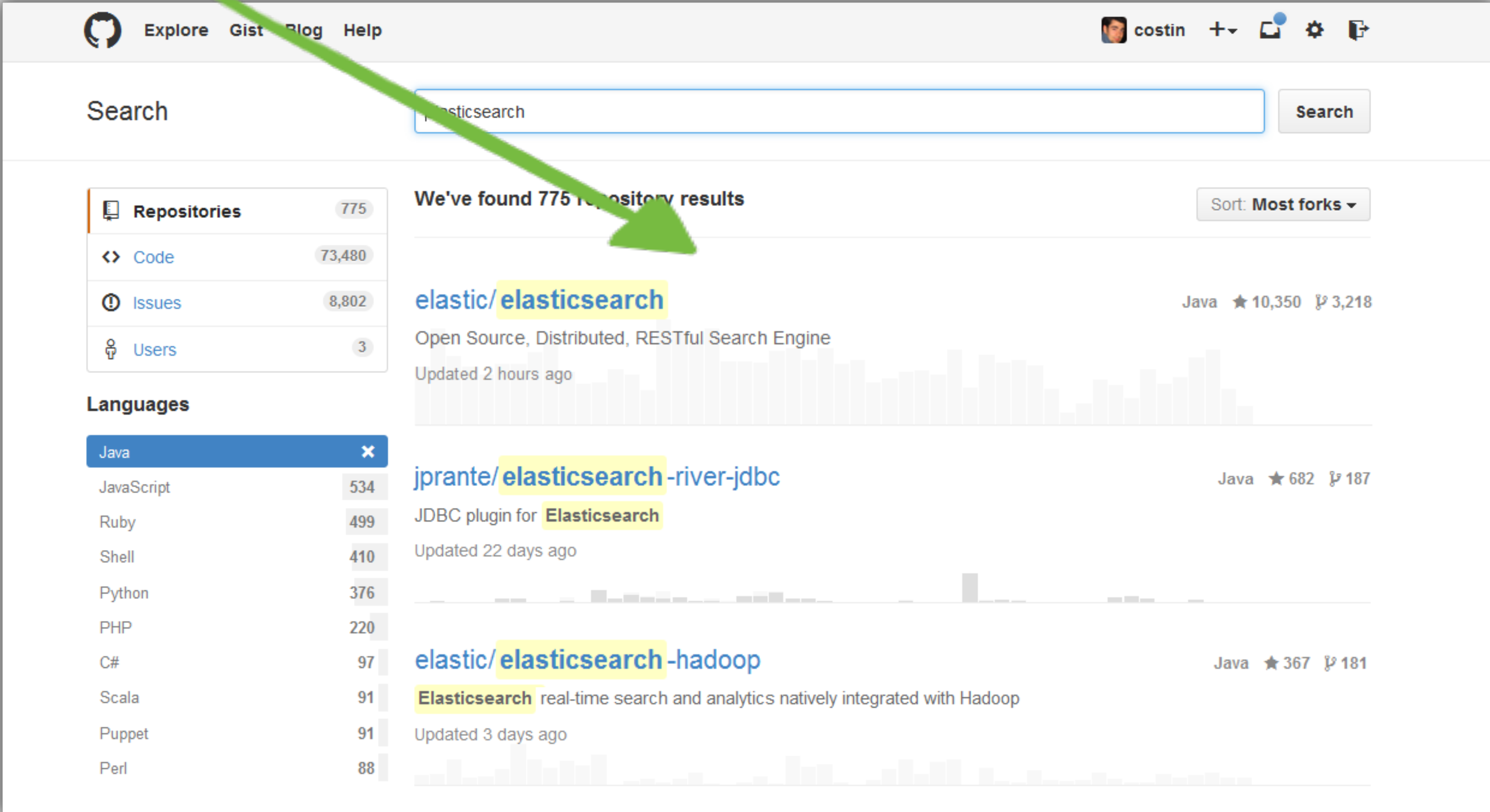
- elastic/elasticsearch**: Java, 10,350 stars, 3,218 forks. Description: Open Source, Distributed, RESTful Search Engine. Updated 2 hours ago.
- jprante/elasticsearch-river-jdbc**: Java, 682 stars, 187 forks. Description: JDBC plugin for Elasticsearch. Updated 22 days ago.
- elastic/elasticsearch-hadoop**: Java, 367 stars, 181 forks. Description: Elasticsearch real-time search and analytics natively integrated with Hadoop. Updated 3 days ago.

On the left side, there are filters for 'Repositories' (775), 'Code' (73,480), 'Issues' (8,802), and 'Users' (3). Below these are language filters, with 'Java' selected.

Pagination

The screenshot shows a GitHub search interface for the query 'elasticsearch'. The search bar at the top contains 'elasticsearch' and a 'Search' button. Below the search bar, a sidebar on the left lists categories: Repositories (775), Code (73,480), Issues (8,802), and Users (3). Under 'Languages', 'Java' is selected, with a list of other languages and their counts: JavaScript (54), Ruby (49), Shell (410), Python (376), PHP (220), C# (97), Scala (91), Puppet (91), and Perl (88). The main content area displays search results. The first result is 'elastic/elasticsearch', described as 'Open Source, Distributed, RESTful Search Engine', updated 2 hours ago, with 10,350 stars and 3,218 forks. The second result is 'jprante/elasticsearch-river-jdbc', described as 'JDBC plugin for Elasticsearch', updated 22 days ago, with 682 stars and 187 forks. The third result is 'elastic/elasticsearch-hadoop', described as 'Elasticsearch real-time search and analytics natively integrated with Hadoop', updated 3 days ago, with 367 stars and 181 forks. At the bottom of the page, a pagination bar shows 'Previous', '1' (selected), '2', '3', '4', '5', '...', '77', '78', and 'Next'. A green arrow points from the top left towards the pagination bar.

Enrichment



The screenshot shows the GitHub search interface. At the top, there are navigation links: Explore, Gist, Blog, and Help. On the right, the user 'costin' is logged in. The search bar contains 'elasticsearch' and a 'Search' button. Below the search bar, a sidebar on the left shows filters for Repositories (775), Code (73,480), Issues (8,802), and Users (3). Under 'Languages', 'Java' is selected. The main content area displays search results. The top result is 'elastic/elasticsearch', a Java repository with 10,350 stars and 3,218 forks. It is described as an 'Open Source, Distributed, RESTful Search Engine' and was updated 2 hours ago. Below this are two other results: 'jprante/elasticsearch-river-jdbc' (Java, 682 stars, 187 forks) and 'elastic/elasticsearch-hadoop' (Java, 367 stars, 181 forks). A green arrow points from the top left towards the first result.

Repository	Stars	Forks
elastic/elasticsearch	10,350	3,218
jprante/elasticsearch-river-jdbc	682	187
elastic/elasticsearch-hadoop	367	181

Suggestions



GitHub This repository ▾ debian [Sign up](#) [Sign in](#)

PUBLIC [elasticsearch](#) [★ Star](#) 4,683 [Fork](#) 1,097

[Browse Issues](#) [New Issue](#)

Everyone's Issues

Labels

- Lucene 4.5 Upgrade
- breaking
- bug
- enhancement
- feature
- non-issue

elasticsearch/elasticsearch#1726 **debian** package violates naming convention

elasticsearch/elasticsearch#3571 **debian** package init-script: start-stop-daemon ne

elasticsearch/elasticsearch#1681 **Debian** pkg

elasticsearch/elasticsearch#3286 There is no official **debian/ubuntu** repository

elasticsearch/elasticsearch#3500 Elasticsearch should include **debian's** standard j

elasticsearch/elasticsearch#1526 Moving **debian** package to maven

Search elasticsearch/elasticsearch for 'debian'

Search GitHub for 'debian'

1	Opened by s1monw 14 hours ago	
11	NoShardAvailableActionException in ES 0.90.3 on startup	#3700
10	Opened by richardwilly98 a day ago	
9	Feature Request: Don't reindex the document when updating non-indexed fields	#3696
1	Opened by ddorian 2 days ago 4 comments	

Forms #3702

roducible #3701

Info

Code

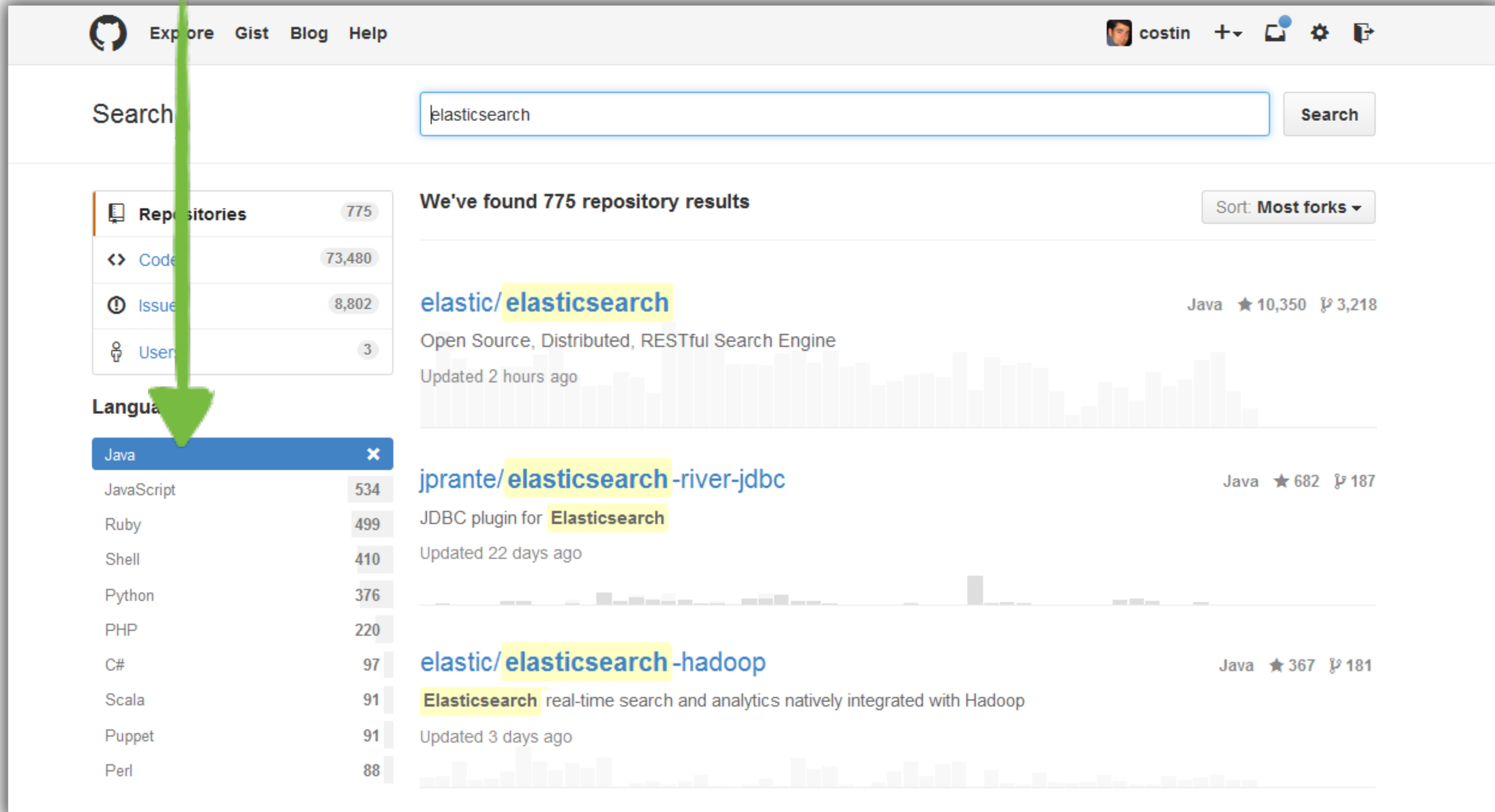
Alert

Activity

Search

Profile

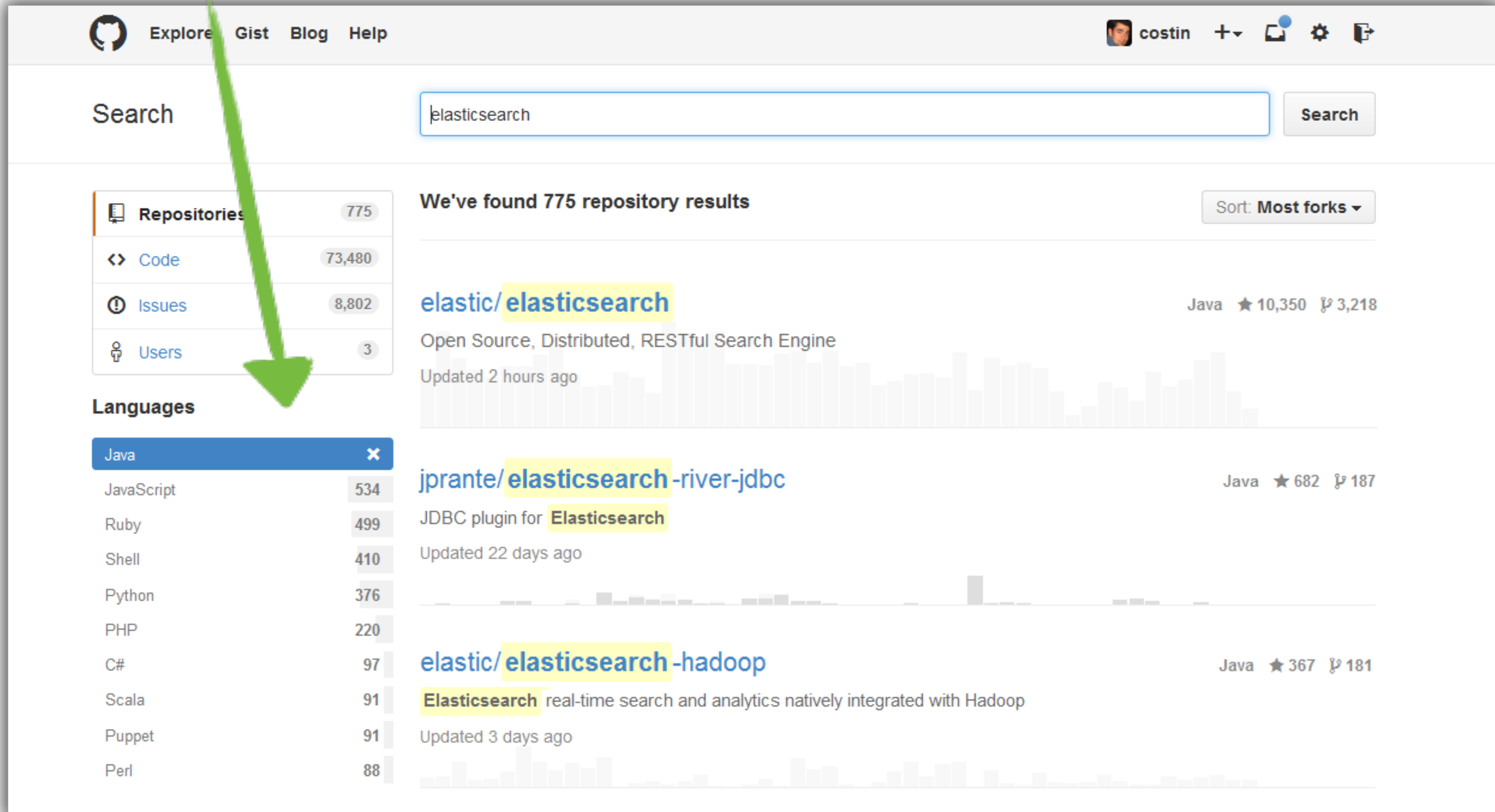
Structured search



The screenshot shows the GitHub search interface. At the top, there are navigation links: Explore, Gist, Blog, and Help. The user's profile 'costin' is visible in the top right. The search bar contains 'elasticsearch' and a 'Search' button. Below the search bar, a summary indicates 'We've found 775 repository results' with a 'Sort: Most forks' dropdown. A sidebar on the left shows search filters: Repositories (775), Code (73,480), Issues (8,802), and Users (3). Under the 'Language' filter, 'Java' is selected and highlighted with a green arrow. The main content area displays three repository results, each with a commit history bar:

- elastic/elasticsearch** (Java, 10,350 stars, 3,218 forks): Open Source, Distributed, RESTful Search Engine. Updated 2 hours ago.
- jprante/elasticsearch-river-jdbc** (Java, 682 stars, 187 forks): JDBC plugin for Elasticsearch. Updated 22 days ago.
- elastic/elasticsearch-hadoop** (Java, 367 stars, 181 forks): Elasticsearch real-time search and analytics natively integrated with Hadoop. Updated 3 days ago.

Aggregations



The screenshot shows the GitHub search interface. At the top, there are navigation links for 'Explore', 'Gist', 'Blog', and 'Help'. The user 'costin' is logged in. A search bar contains the text 'elasticsearch' and a 'Search' button. Below the search bar, a summary states 'We've found 775 repository results' with a 'Sort: Most forks' dropdown. A green arrow points from the top left towards the 'Repositories' filter in the left sidebar. The sidebar includes filters for 'Repositories' (775), 'Code' (73,480), 'Issues' (8,802), and 'Users' (3). Under the 'Languages' section, 'Java' is selected. The main content area displays three repository results, each with a star count and a commit count, and a commit history bar chart.

Repository	Language	Stars	Forks
elastic/elasticsearch	Java	10,350	3,218
jprante/elasticsearch-river-jdbc	Java	682	187
elastic/elasticsearch-hadoop	Java	367	181

elasticsearch-hadoop

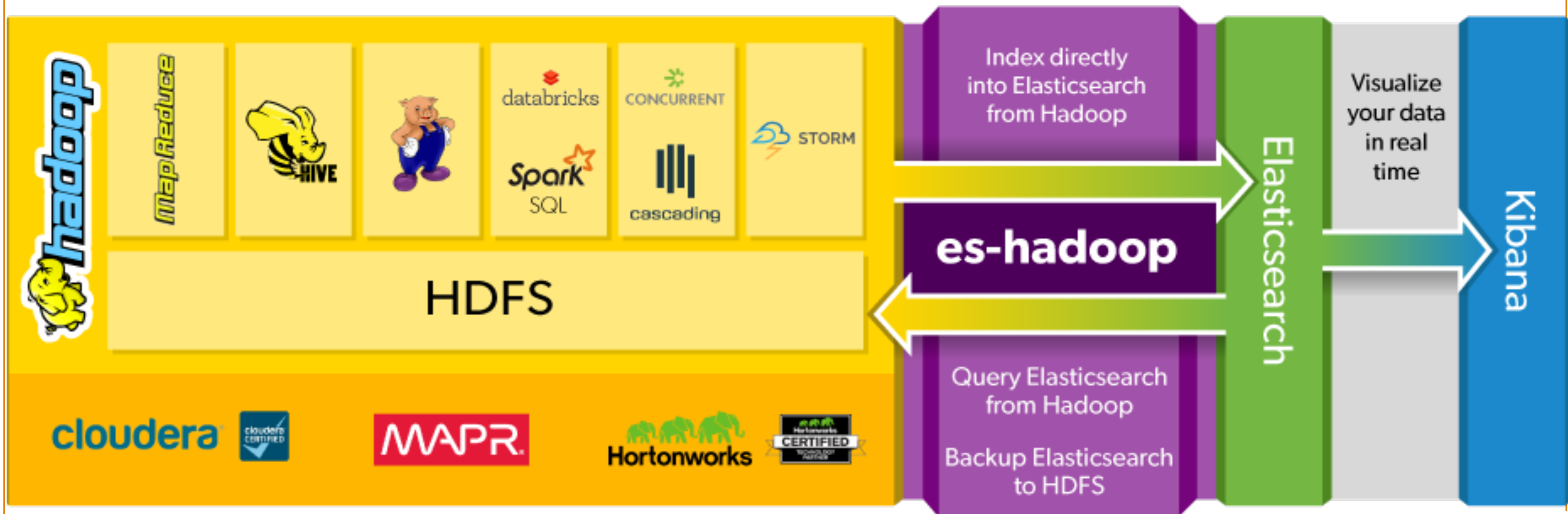
Java ★ 362 📄 178

Elasticsearch real-time search and analytics natively integrated with Hadoop

Updated a day ago



Elasticsearch for Apache Hadoop



Map/Reduce integration

```
import org.elasticsearch.hadoop.mr._

val conf = new Configuration()
conf.set("es.resource", "radio/artists")
conf.set("es.query", "?q=me*")

val mrNewApiRDD = sc.newAPIHadoopRDD(conf,
    classOf[EsInputFormat[Text, MapWritable]],
    classOf[Text], classOf[MapWritable])

val mrOldApiRDD = sc.hadoopRDD(conf,
    classOf[EsInputFormat[Text, MapWritable]],
    classOf[Text], classOf[MapWritable])
```

Scala API

```
import org.elasticsearch.spark._  
  
val sc = new SparkContext(new SparkConf())  
val rdd = sc.esRDD("radio/artists", "?me*")
```

```
import org.elasticsearch.spark._  
  
case class Artist(name: String, albums: Int)  
  
val u2 = Artist("U2", 12)  
val bh = Map("name" -> "Buckethead", "albums" -> 95, "age" -> 45)  
  
sc.makeRDD(Seq(u2, bh)).saveToEs("radio/artists")
```

Java API

```
import org.elasticsearch.spark.java.api.JavaEsSpark;  
  
JavaSparkContext jsc = new JavaSparkContext(sc);  
JavaPair jrdd = JavaEsSpark.esRDD("radio/artists", "?me*");
```

```
import org.elasticsearch.spark.java.api.JavaEsSpark;  
  
JavaSparkContext jsc = ...  
  
Map doc1 = ImmutableMap.of("name", "u2", "albums", 12);  
Object doc2 = ArtistJavaBean("buckethead", 95, 45);  
JavaRDD javaRDD = jsc.parallelized(ImmutableList.of(doc1, doc2));  
  
JavaEsSpark.saveToEs(javaRDD, "radio/artists");
```

Spark SQL support

```
import org.elasticsearch.spark.sql._

val sql = new SQLContext...
val artists = sql.esRDD("radio/artists", "?me*")
val a80s = sql.sql("SELECT * FROM artists
                   WHERE formed <= 1990 AND formed >=1980")
```

```
import org.elasticsearch.spark.sql._

val sql = new SQLContext...
val people = sql.parquetFile("artists.dat")

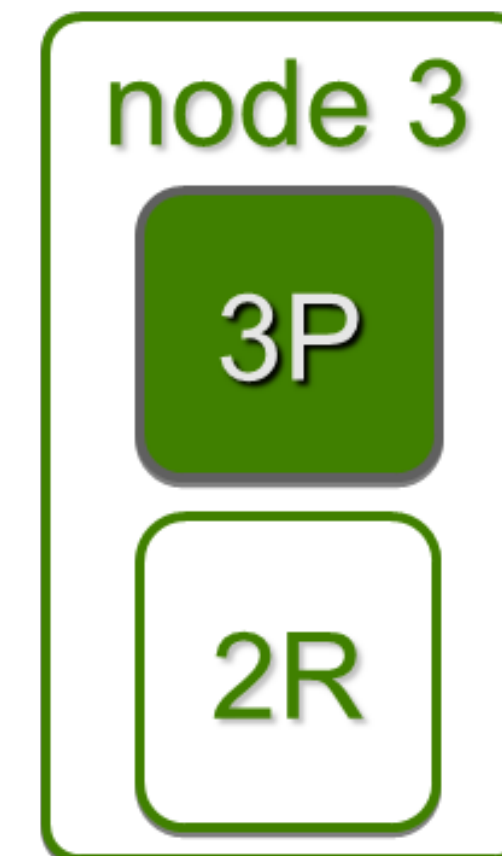
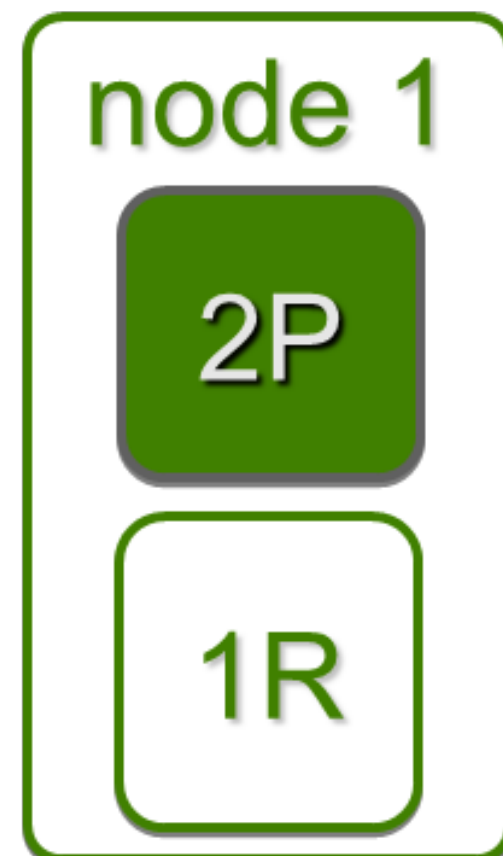
sql.saveToEs("radio/artists")
```


Spark SQL Data Sources

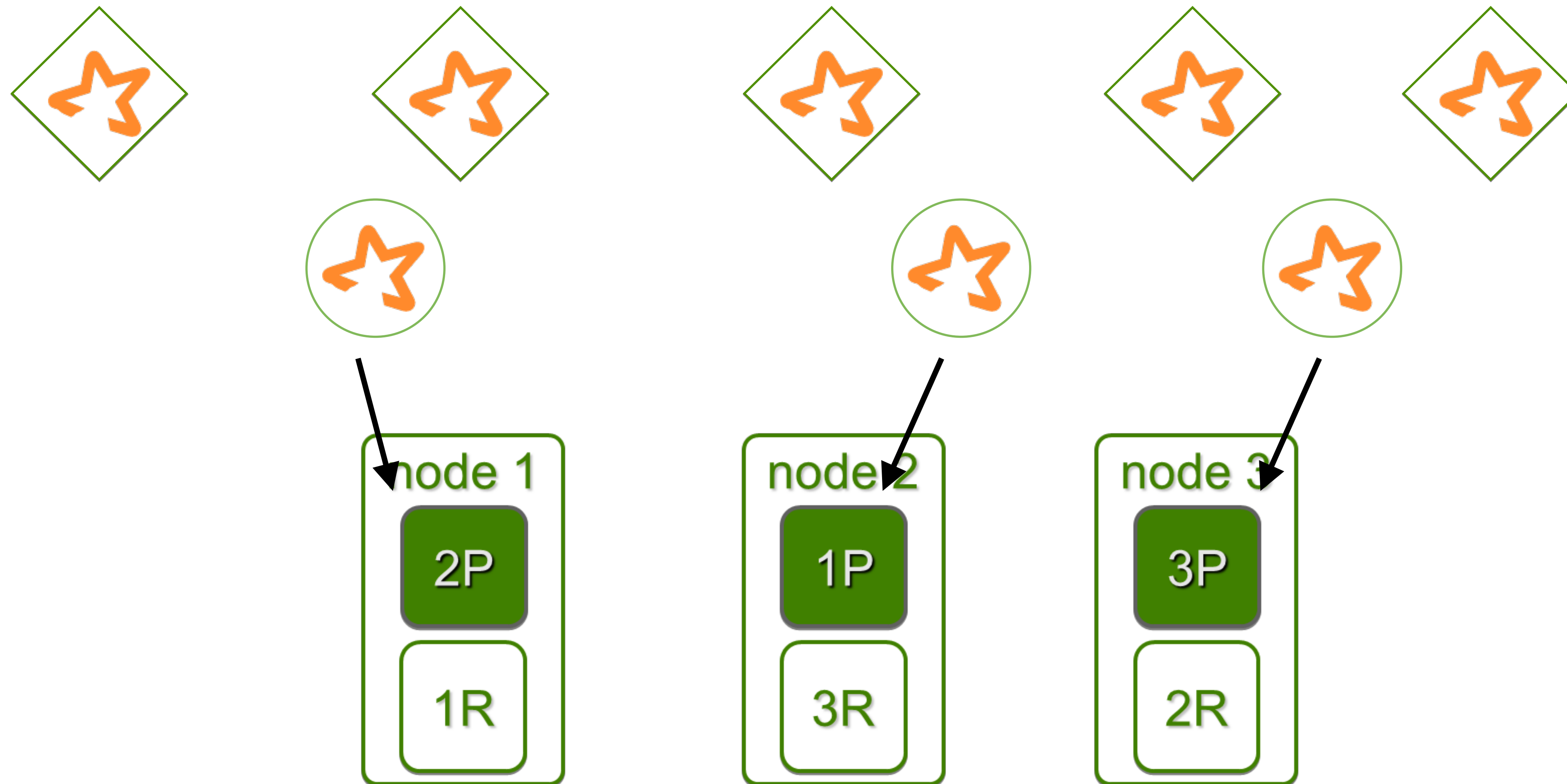
```
val sql = new SQLContext...  
val df = sql.load("radio/artists", "org.elasticsearch.spark.sql")  
df.filter(df("age") > 40)
```

```
val sql = new SQLContext...  
val table = sql.sql("CREATE TEMPORARY TABLE artists " +  
                    "USING org.elasticsearch.spark.sql " +  
                    "OPTIONS(resource=`radio/artists`) ")  
  
val names = sql.sql("SELECT name FROM artists")
```

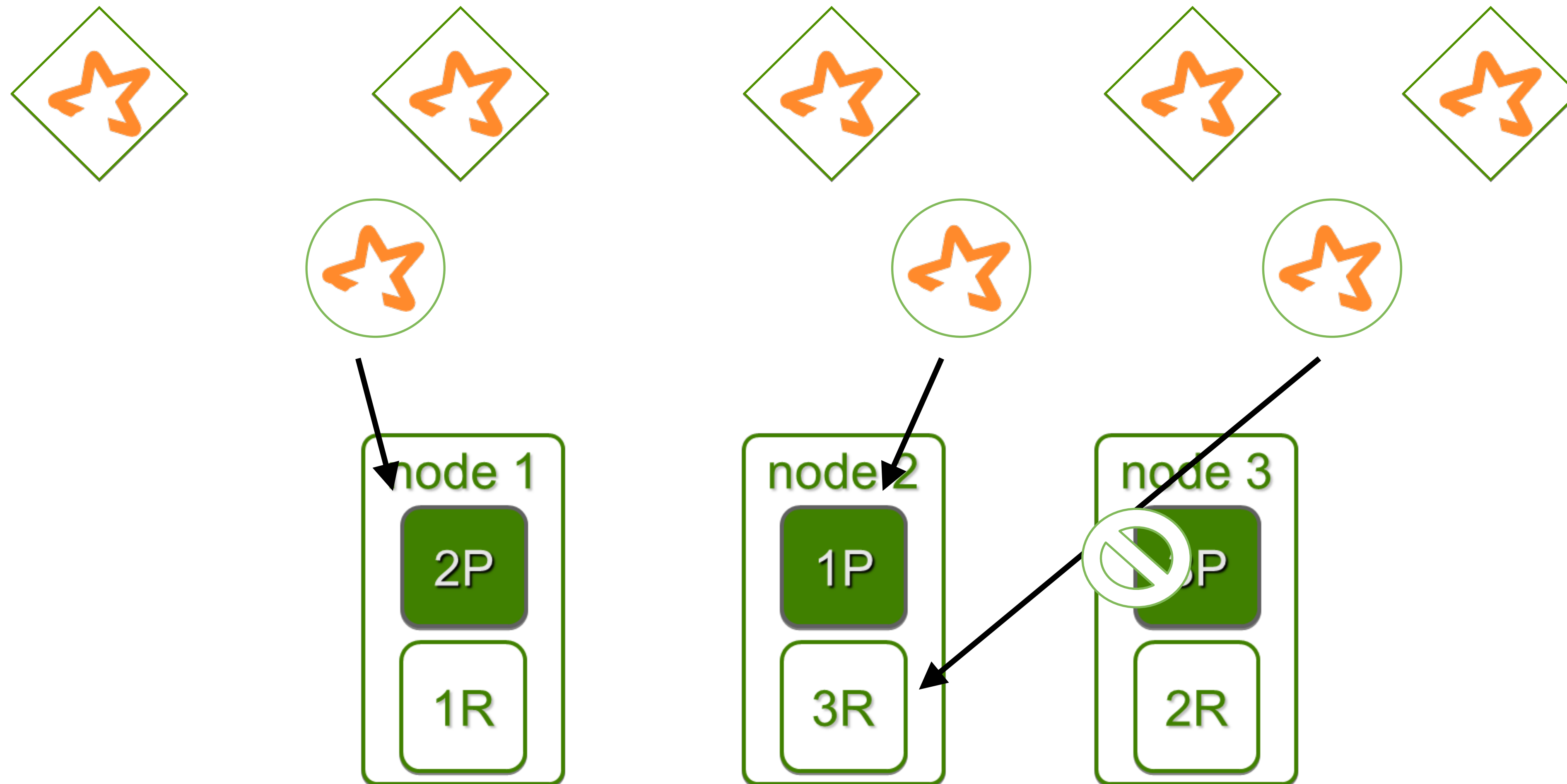
Partition-to-Partition Architecture



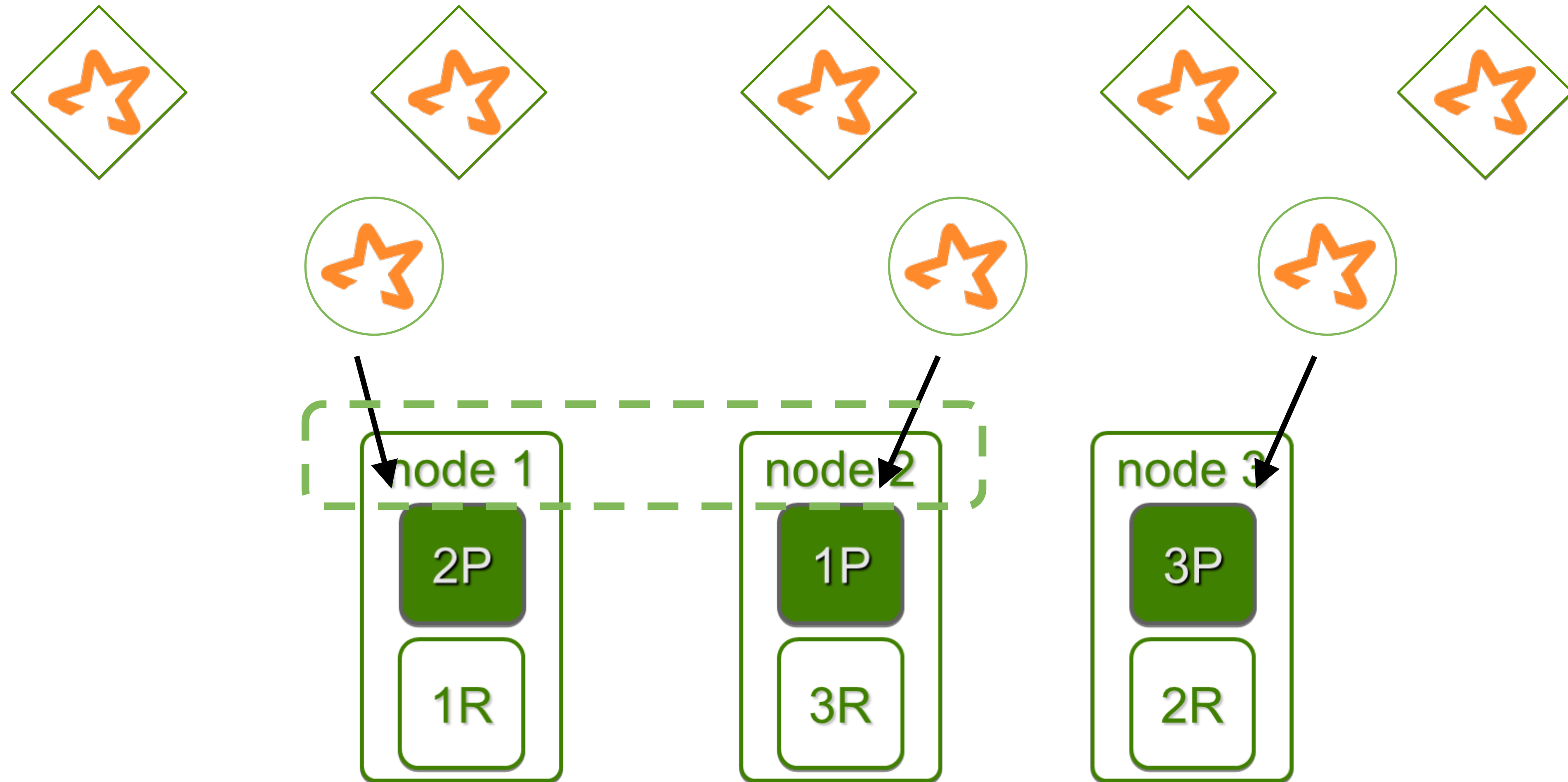
Dynamic Runtime Matching



Failure Handling



Co-location



Reacting to streaming data



Live loops

Data keeps on changing

Adapt set of rules

Improves reaction time

Build a model for fast decision making

Keeps the prevention rate high

Categorize data on the fly

Finding interesting data basic approach

```
import org.apache.spark.SparkContext._
import org.elasticsearch.spark._

val sc = new SparkContext(conf)

// Spark PairRDD
val esRDD = sc.esRDD("bank/operations",
  "{ \"fields\" : [\"location\", \"owner\"],
  \"query\" : { \"term\" : { \"transaction\" : \"cash\" } } }")
```


Finding interesting data analytics

```
// get outstanding cash transactions

val esRDD = sc.esRDD("branch/transactions",
  "{ \"query\" : { \"terms\" : { \"transaction\" : [ \"cash\" ] } } },
  \"aggregations\" : {
    \"unusual\" : {
      \"significant_terms\" : { \"field\" : \"amount\" }
    }
  } }")
```

Finding interesting data through a ML model

```
import org.apache.spark.mllib.linalg.Vector
import org.apache.spark.mllib.feature.HashingTF
import org.apache.spark.mllib.clustering.KMeans

val vectors = esRDD.map(_("location")).map(featurize)

val model = KMeans.train(vectors, numClusters, numIterations)

// predict / flag
val pendingTx = sc.esRDD("branch/pending", "amount gt 1000")
val flaggedTx = pendingTx.filter( t =>
    model.predict(featurize(t)) == clusterGroup)
```

MLlib integration - wip

Hashing and featurize functions

Expose the Elasticsearch engine data structures

- term vectors

- term frequency

- document frequency

- (vectorize API in the works)

Thank you!

@costinl

github.com/elastic

elastic.co