

Elasticsearch & Bigdata

Medcl

About me

- Medcl，曾勇（Zeng Yong）
- Developer @ Elastic
  - Follow Elasticsearch since v0.5, 2010
  - Joined Elastic since September, 2015
- @medcl
- medcl@elastic.co
- http://github.com/medcl
- Based in Changsha, Hunan, China

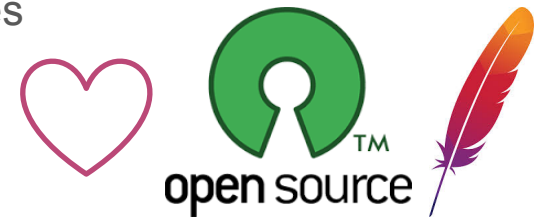


elastic

About Elastic



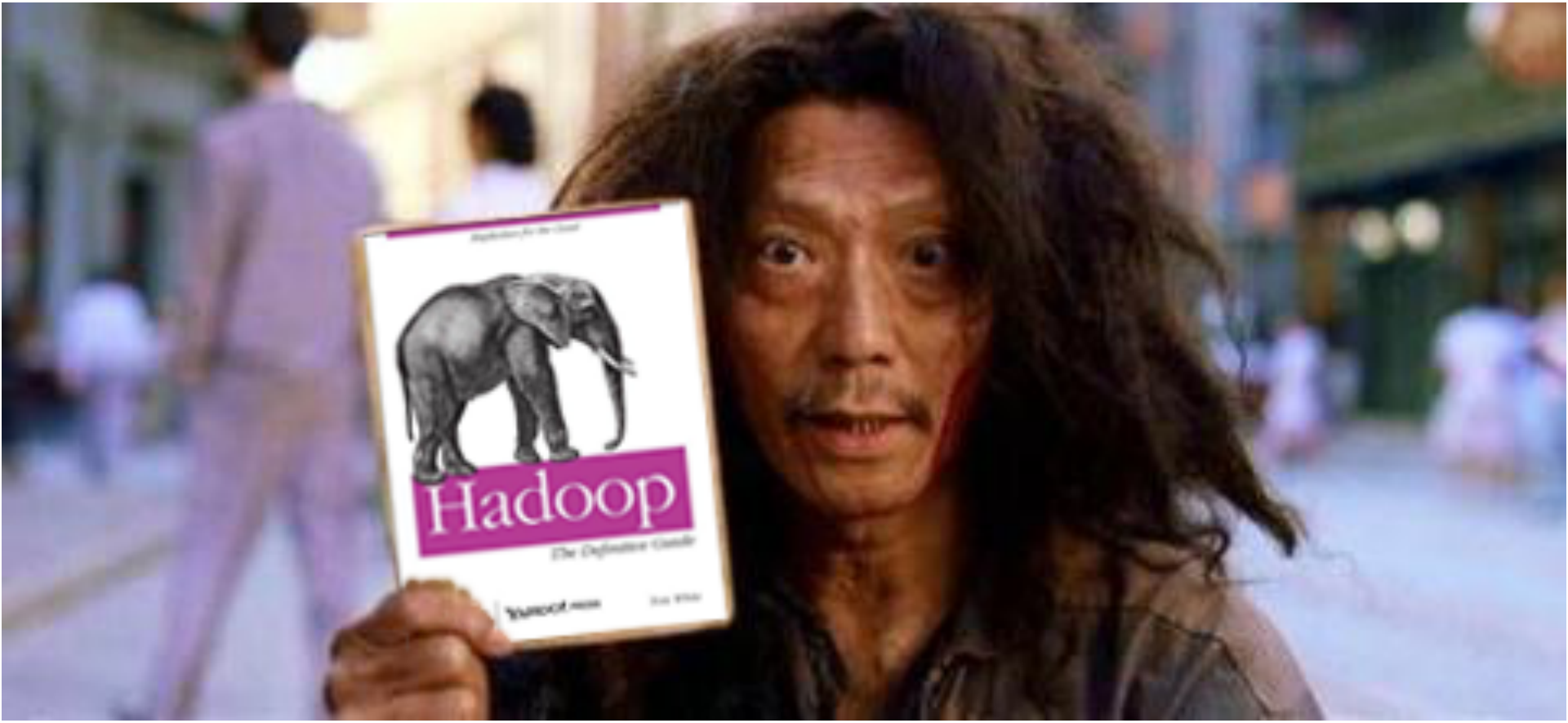
- A distributed startup company，since 2012
  - HQ: Mountain View, CA AND Amsterdam, Netherlands
  - With employees in 27 countries (and counting),
  - spread across 18 time zones, speaking over 30 languages
- We are working on Open Source projects!
  - (Luckily some of them are popular, eg:elasticsearch)



- Offering Support Subscription，X-pack，Cloud and Trainings
- Find us on: https://github.com/elastic and https://www.elastic.co



elastic



elastic

Hadoop Ecosystem



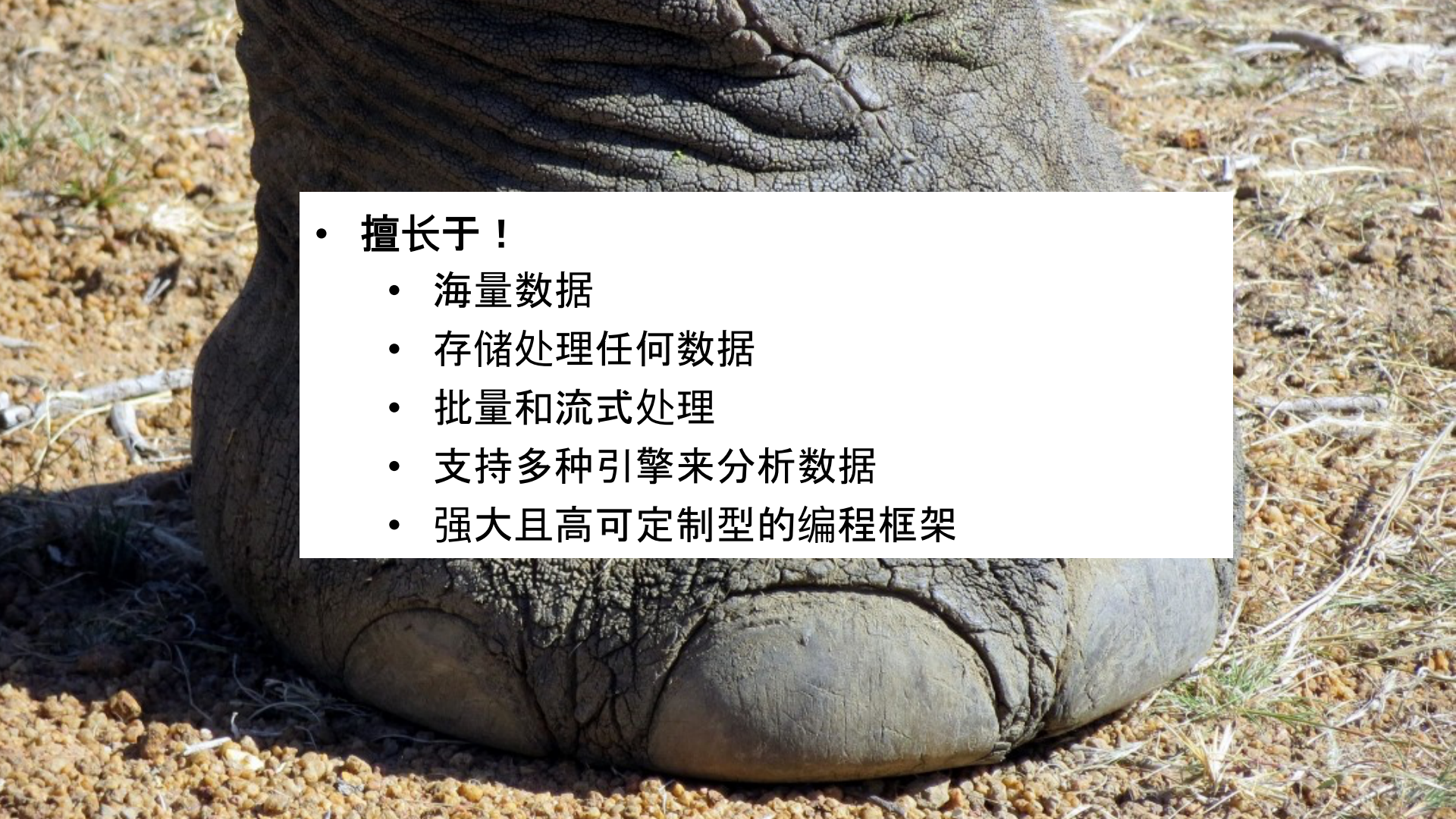
Map / Reduce

Other

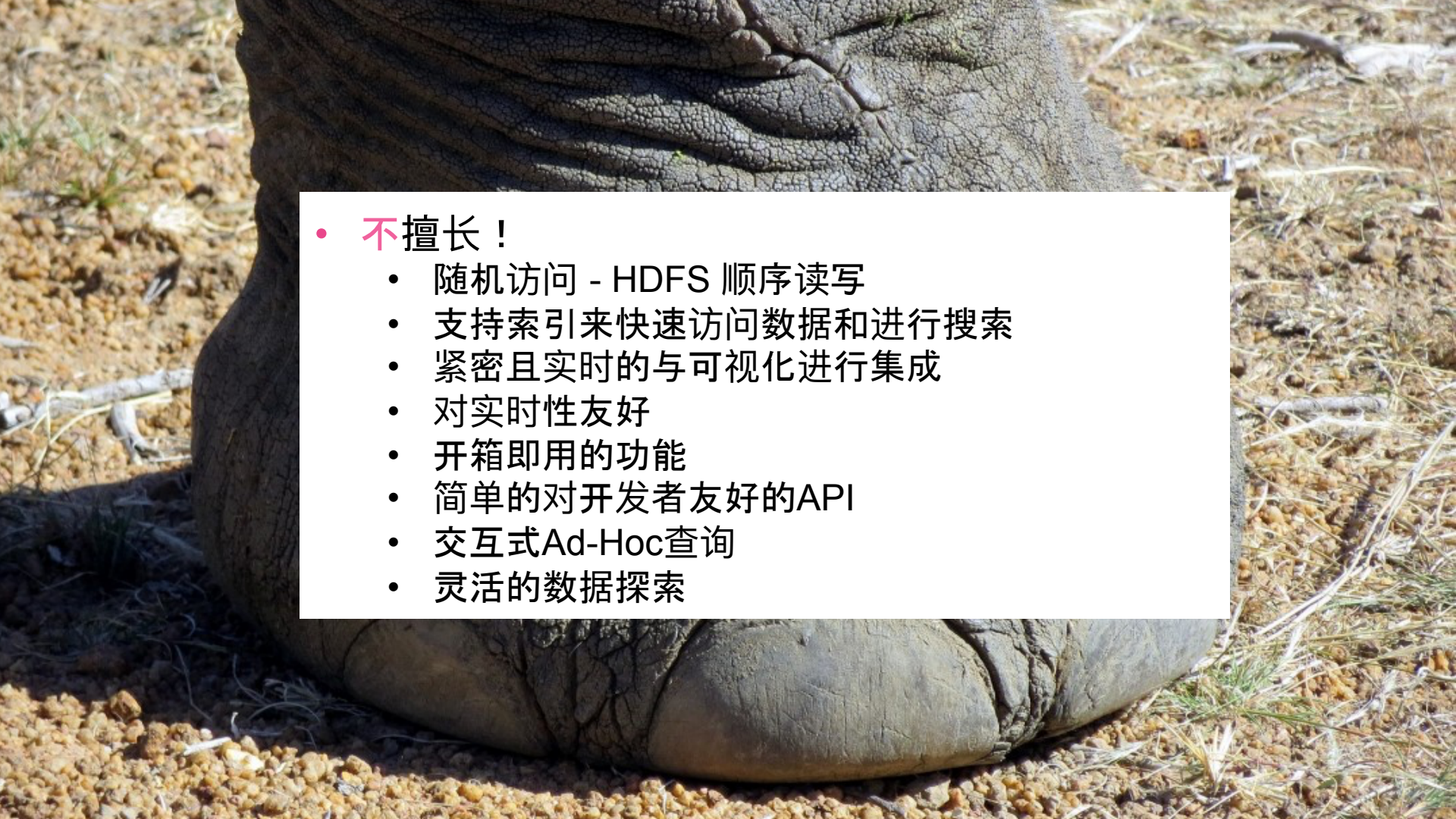
YARN

Hadoop Distributed File System (HDFS)

elastic



- **擅长于！**
  - 海量数据
  - 存储处理任何数据
  - 批量和流式处理
  - 支持多种引擎来分析数据
  - 强大且高可定制型的编程框架



- 不擅长！
  - 随机访问 - HDFS 顺序读写
  - 支持索引来快速访问数据和进行搜索
  - 紧密且实时的与可视化进行集成
  - 对实时性友好
  - 开箱即用的功能
  - 简单的对开发者友好的API
  - 交互式Ad-Hoc查询
  - 灵活的数据探索



elastic



elasticsearch

an open source, distributed, scalable, highly available, document-oriented, RESTful, full text search engine with real-time search and analytics capabilities

http://github.com/elastic/elasticsearch

elastic



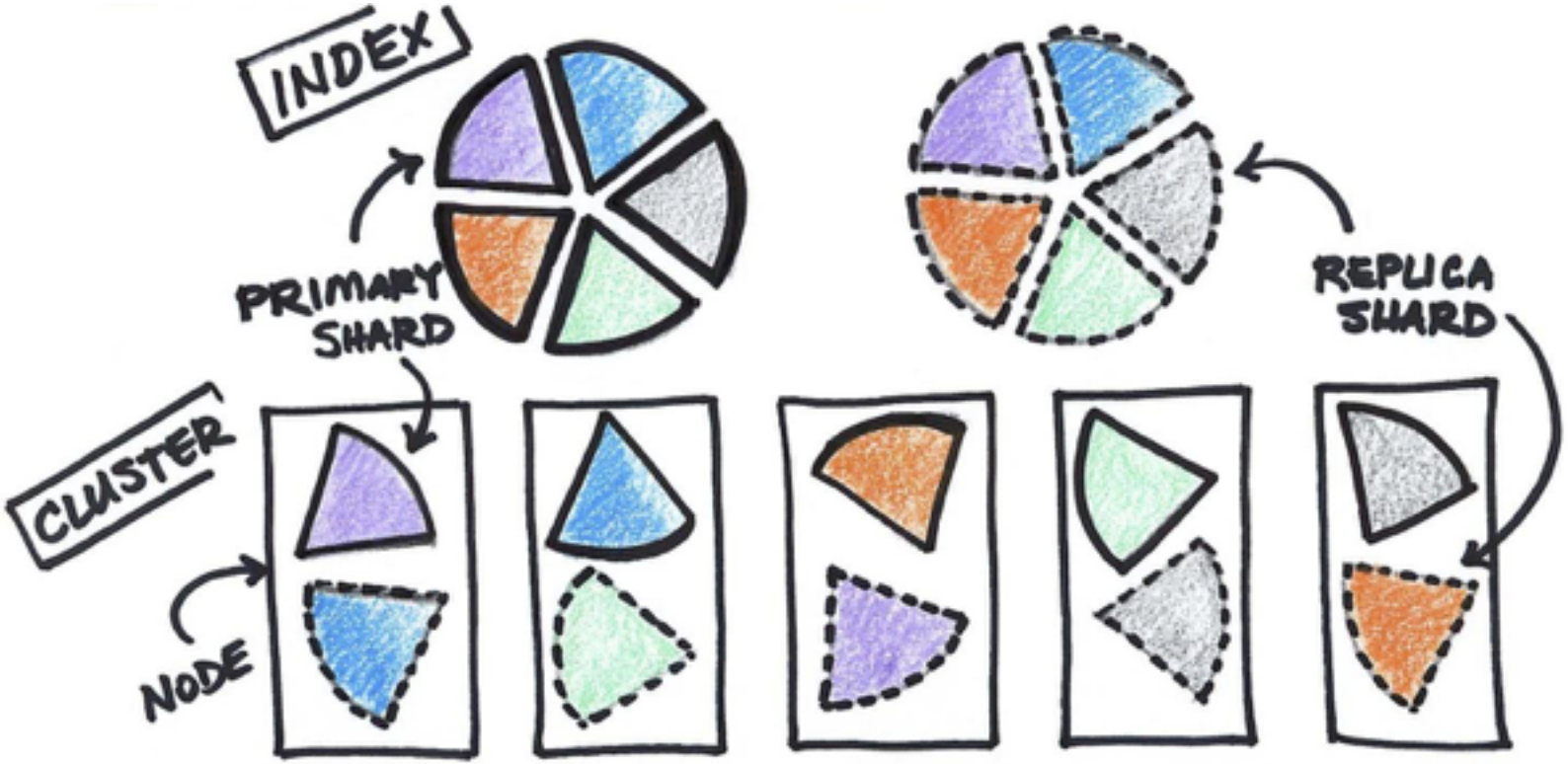
elasticsearch

*an open source, distributed, scalable, highly available, document-oriented, RESTful, full text search engine with **real-time** search and analytics capabilities*

http://github.com/elastic/elasticsearch

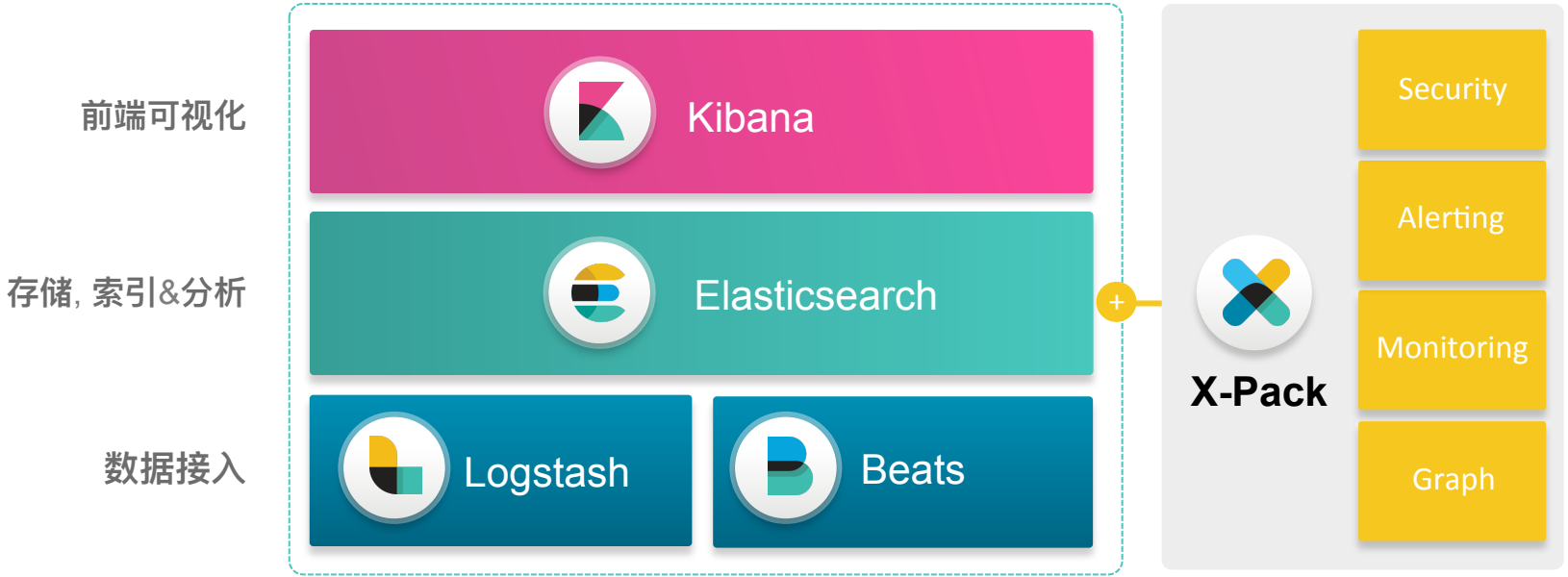
elastic

Glossary



elastic

Elastic Stack



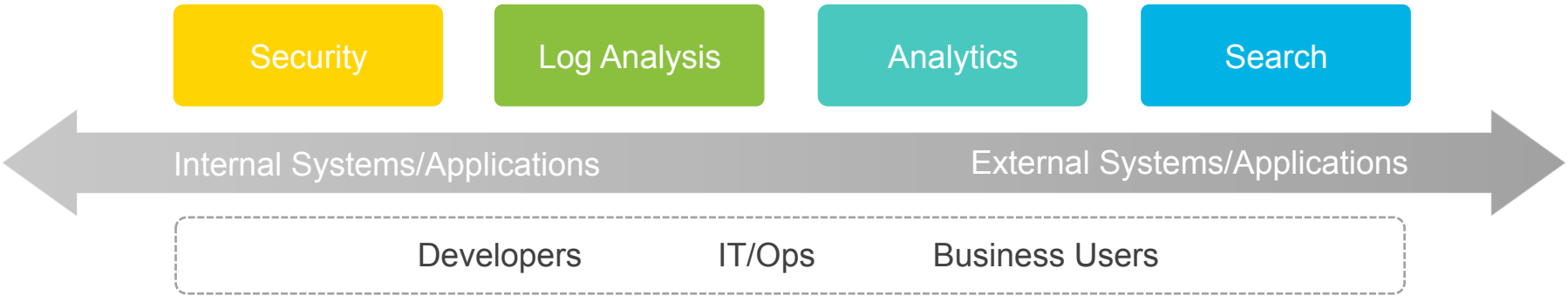
elastic

Solving Many Use Cases Within Any Organization

IT Operations
Application Management
Security Analytics

Marketing Insights
Business Development
Customer Sentiment

Website/App Search
Internal/Intranet Search
URL Search





elastic



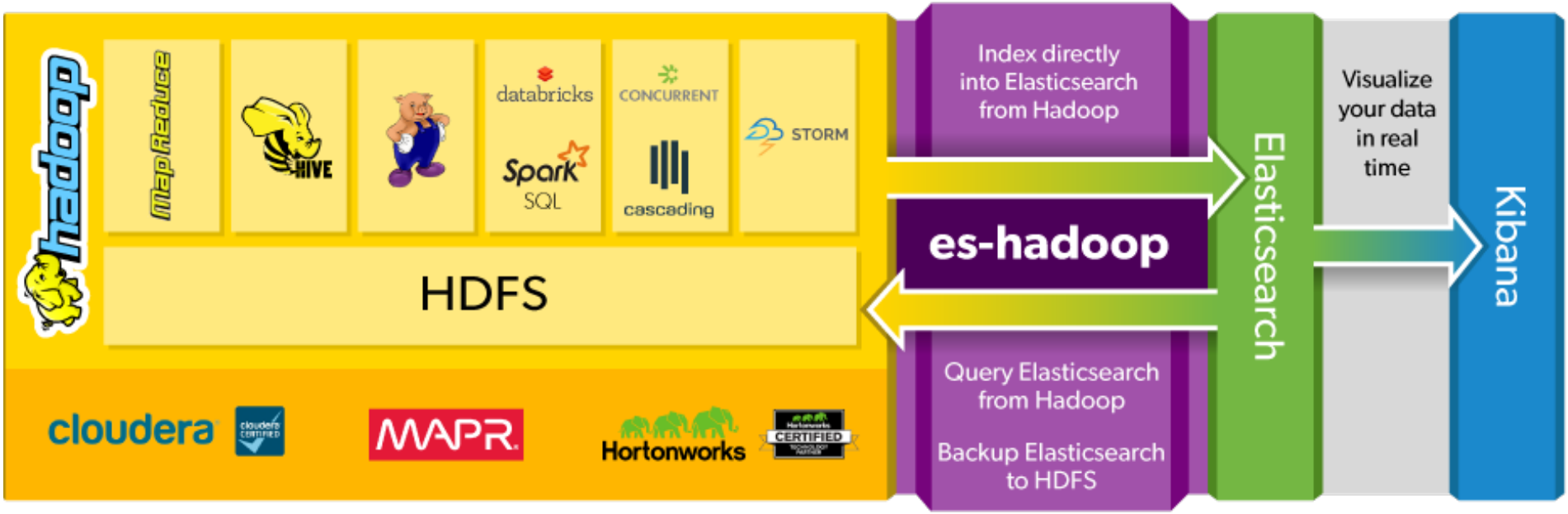
elastic



Elastic Stack + Hadoop Stack

Elasticsearch for Apache Hadoop ™

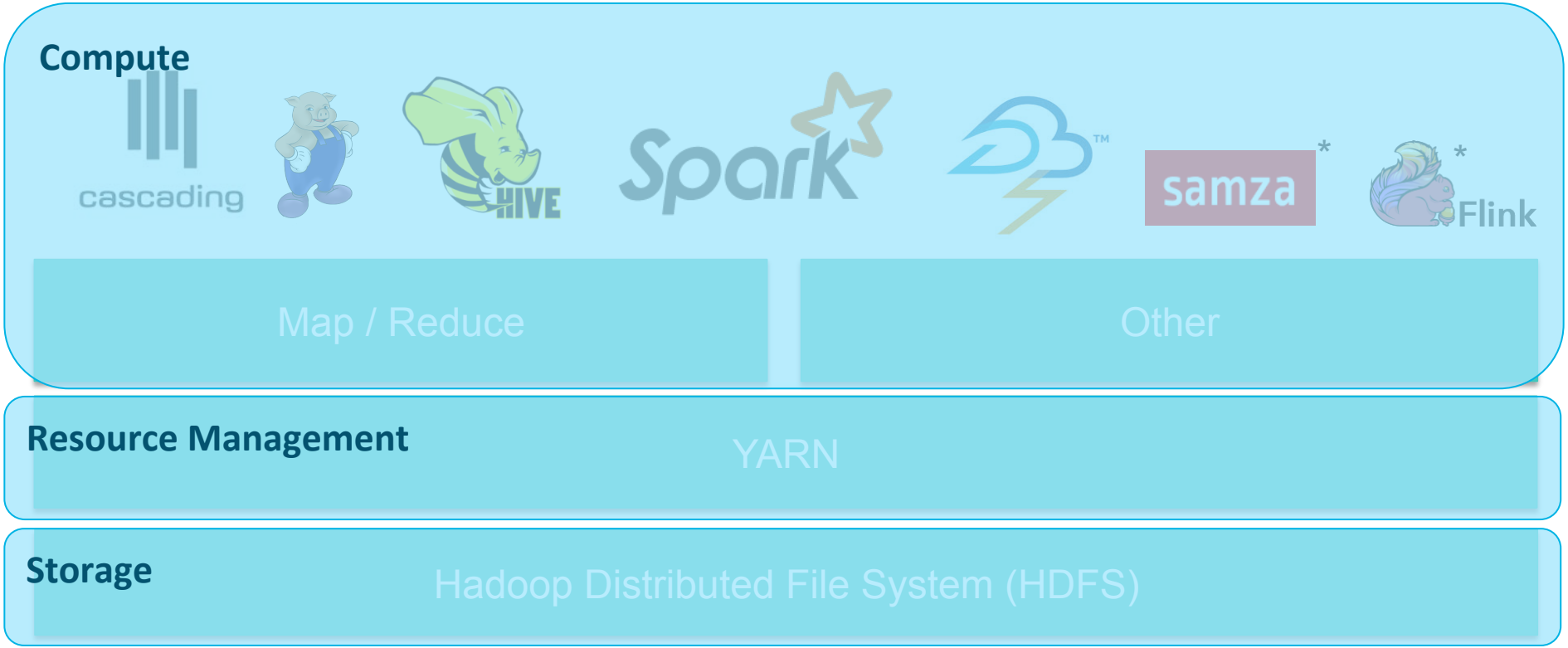
aka ES-Hadoop



https://github.com/elastic/elasticsearch-hadoop

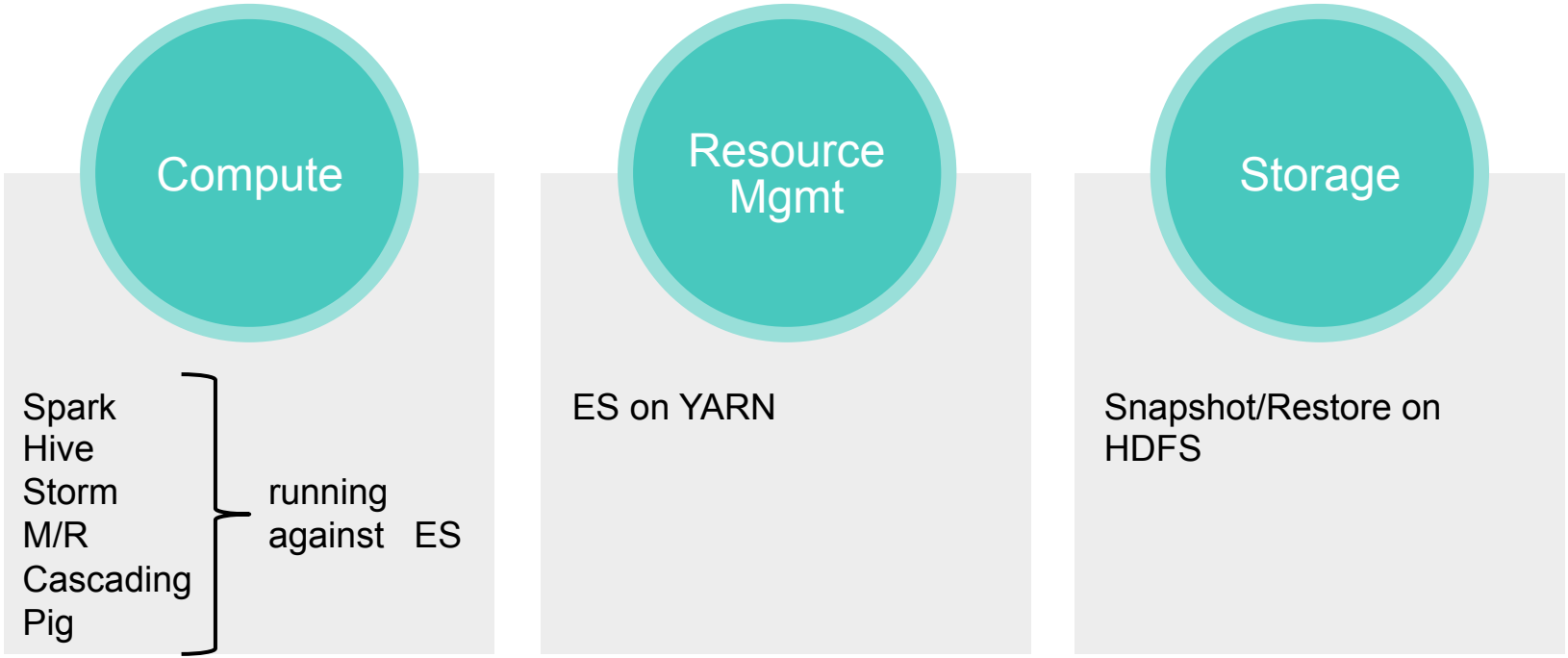
elastic

Hadoop Ecosystem



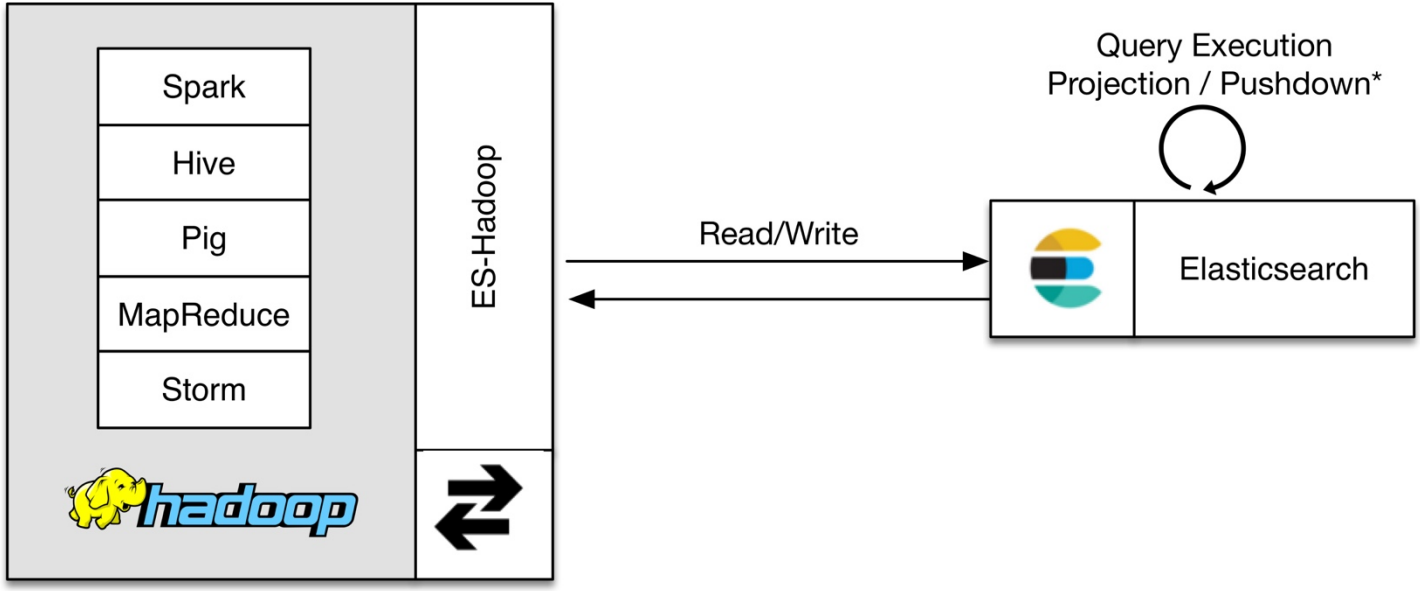
elastic

ES-Hadoop components



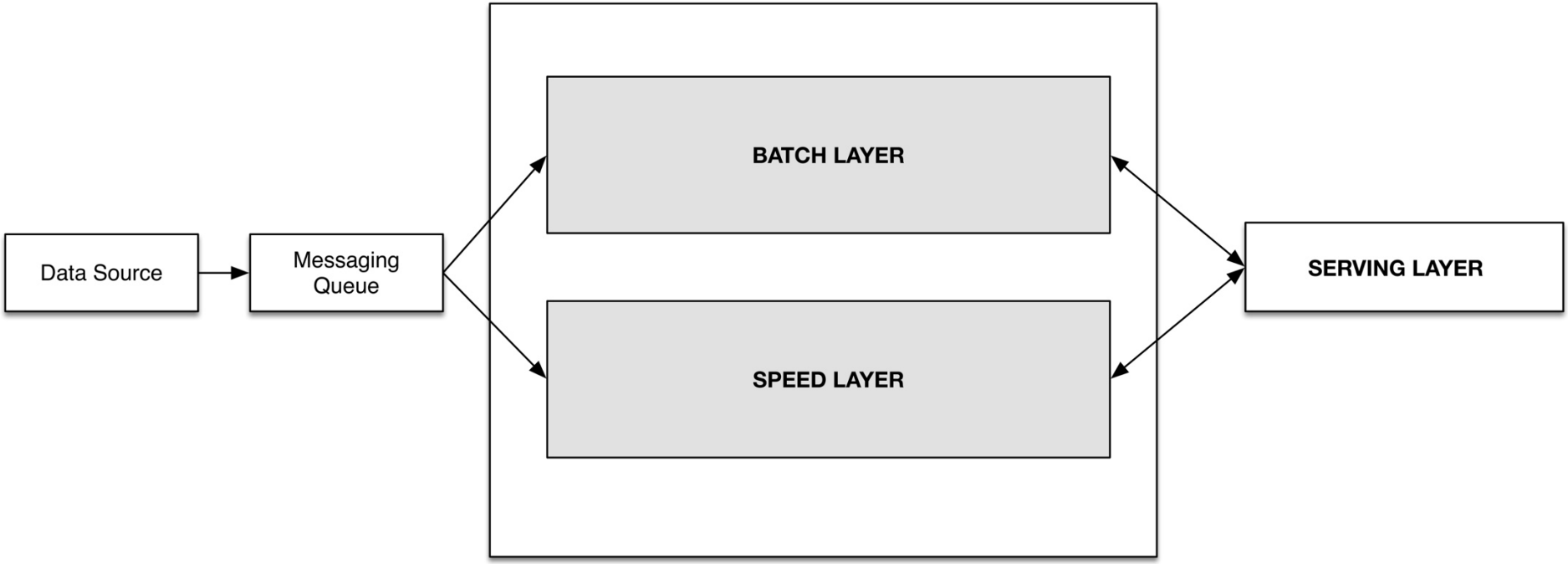
elastic

ES-Hadoop



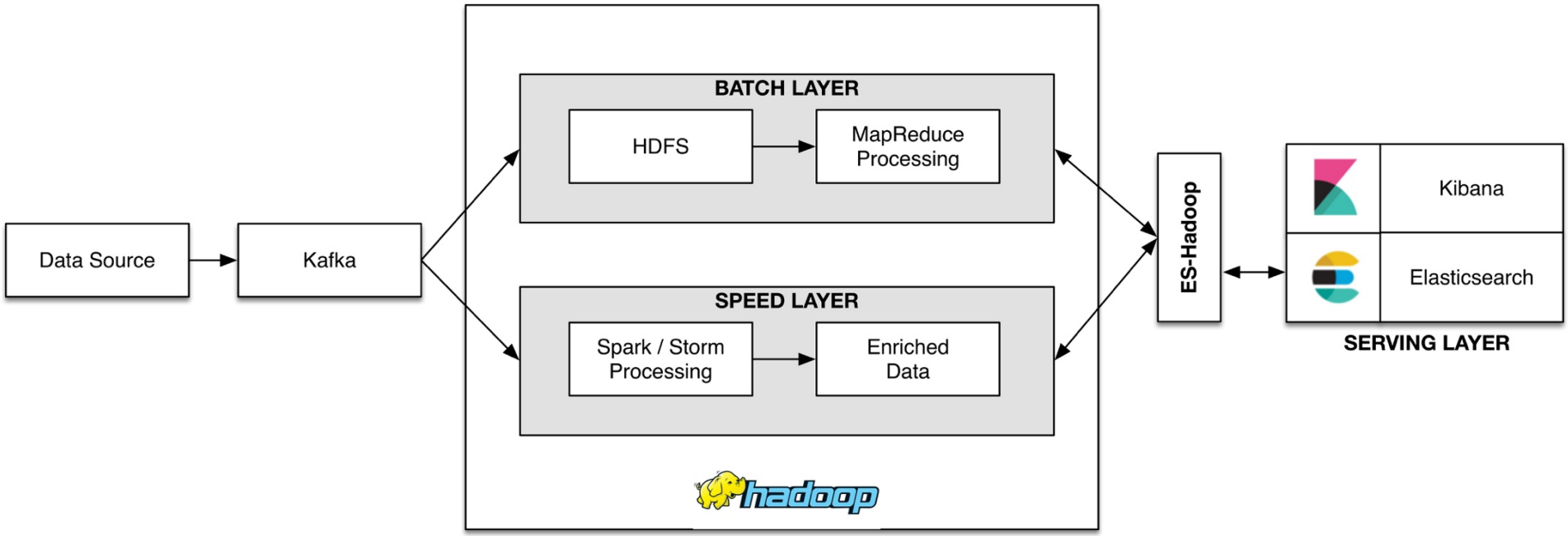
elastic

Lambda Architecture



elastic

Lambda Architecture



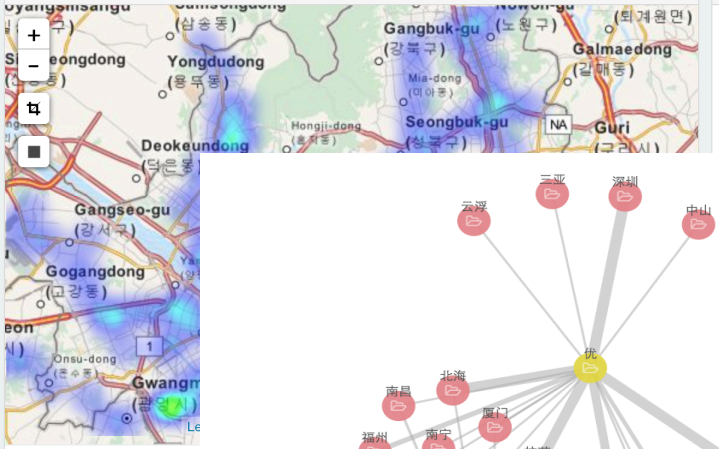
elastic

Why ?

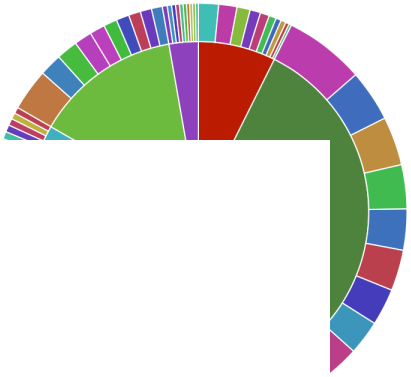
- Index Index Index !
- Native, Real-time data access – Kibana platform
- Easy powerful APIs
- Real-time even at scale
- Simple data model - JSON
- Built-in Analyzer
- Out of the box functionality – Relevance, Full-Text Search, Geo, Aggregation
- Embedding Elasticserach in custom applications
- One stone, two mangoes

elastic

Seoul Metro - Geomap people In



Seoul Metro - people in per line and station



Seoul Metro - top 50 stations people in

Station	People in
Gangnam	82,359
Seolleung	52,841
Samseong (World Trade Center Seoul)	49,845
Seoul Station	46,908
Jongno 3(sam)ga (Tapgol Park)	46,711
Yeoksam	45,153
Seoul Nat'l Univ. of Education (Court & Public Prosecutor's Office)	44,475
Jamsil (Songpa-gu Office)	44,151

Export: Raw Formatted

Seoul Metro - Geomap

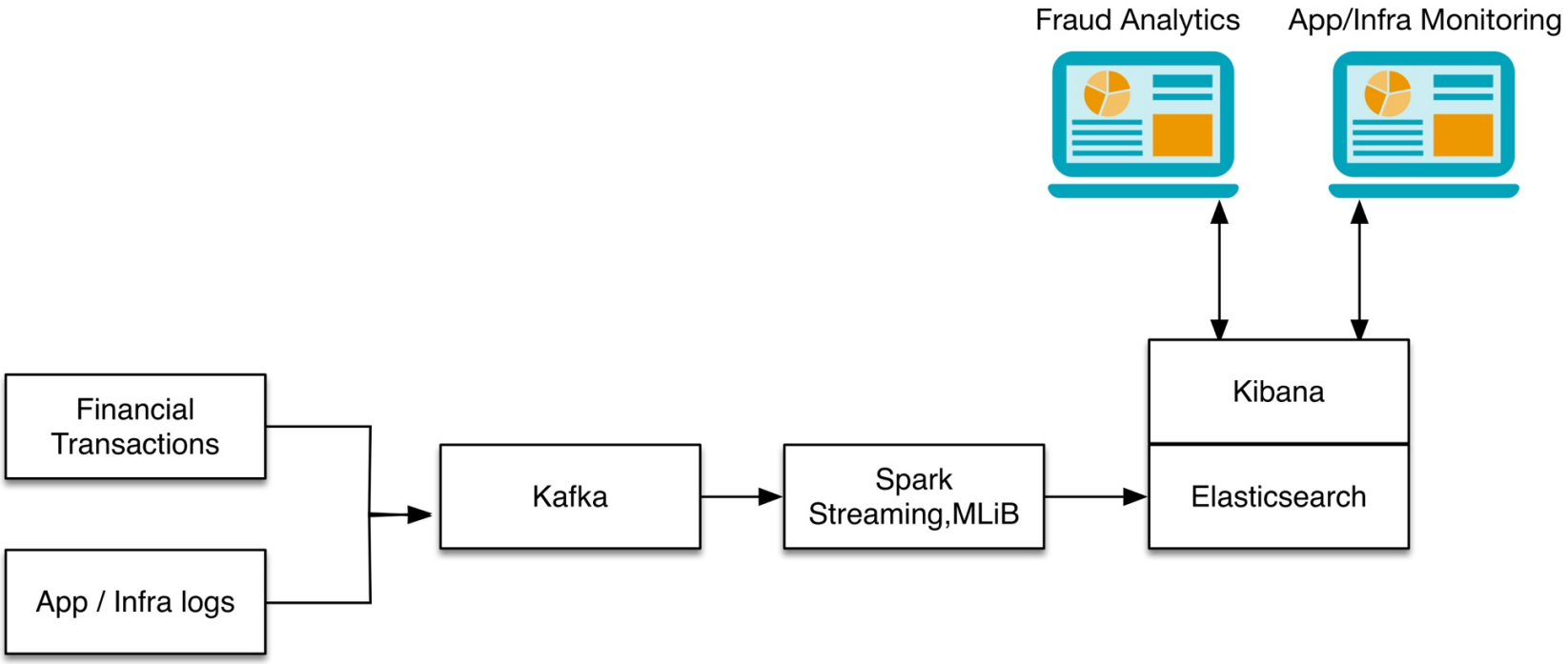
Seoul Metro - top 50 stations people out

Station	People out
Sillim	53,417
Sadang	47,835
Jamsil (Songpa-gu Office)	43,873
Gangnam	43,024
Konkuk Univ.	39,863
Sindorim	39,780
Hongik Univ.	37,308
Yeonsinnae	36,846

Export: Raw Formatted

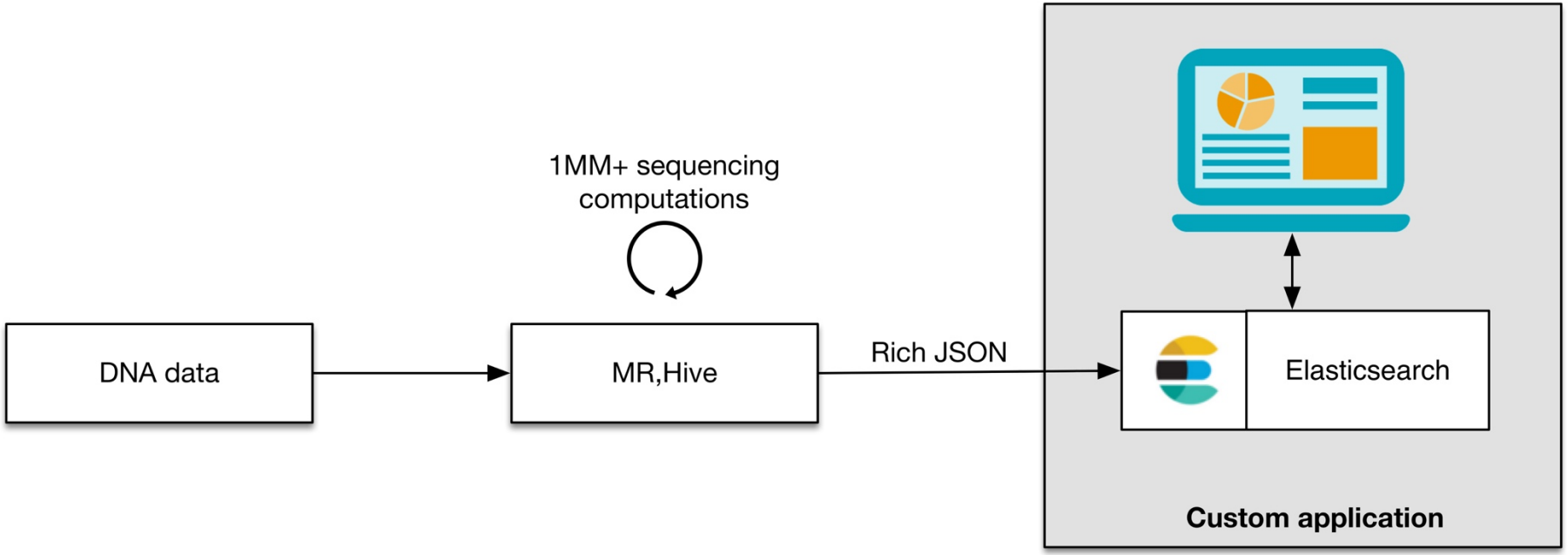
1 2 3 4 5 ...7 »

Case: Financial Services Company



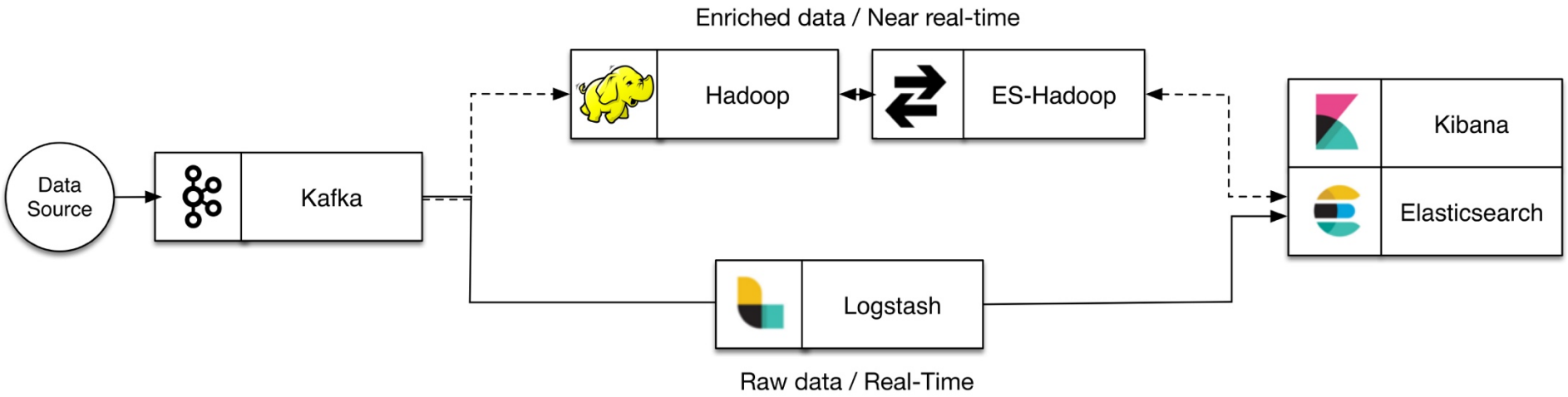
elastic

Case: Bioinformatics Company



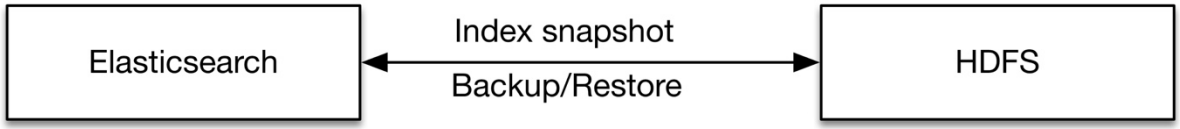
elastic

Case: Raw Real-time Architecture



elastic

Case: HDFS as a backup store



elastic

ES-Hadoop compute integrations

Library / API		ES-Hadoop exposed as
Map/Reduce		`Input / OutputFormat`
Cascading		`Tap / Sink`
Apache Pig		`Loader / Storage`
Apache Hive		`(EXTERNAL) TABLE`
Apache Storm		`Spout / Bolt`
Apache Spark		RDD, DataFrame, DataSource

elastic

Apache Hive




```
CREATE EXTERNAL TABLE playlist (name STRING, year:long)
STORED BY 'org.elasticsearch.hadoop.hive.EsStorageHandler'
TBLPROPERTIES('es.resource'='buckethead/albums','es.query'='?q=pikes');

SELECT name FROM playlist WHERE year > 2010;
```

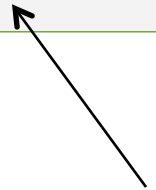
```
CREATE EXTERNAL TABLE playlist(name STRING, year:long)
STORED BY 'org.elasticsearch.hadoop.hive.EsStorageHandler'
TBLPROPERTIES('es.resource'='buckethead/albums');

INSERT INTO TABLE playlist VALUES ('Buildor', 2016), ('Florrmat', 2016);
```


elastic

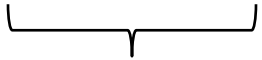
Query Analysis

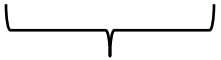
```
SELECT name FROM playlist WHERE year > 2010;
```





```
{ name : “shapeless”, year : 2015 }
```





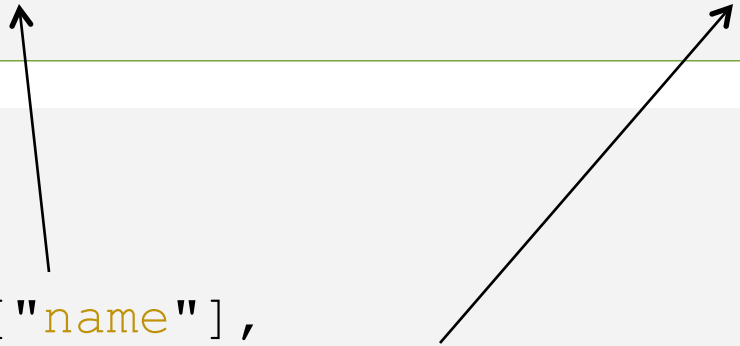
Target fields

Filtering/Operating fields

elastic

Query DSL Conversion

```
SELECT name FROM playlist WHERE year > 2010;
```

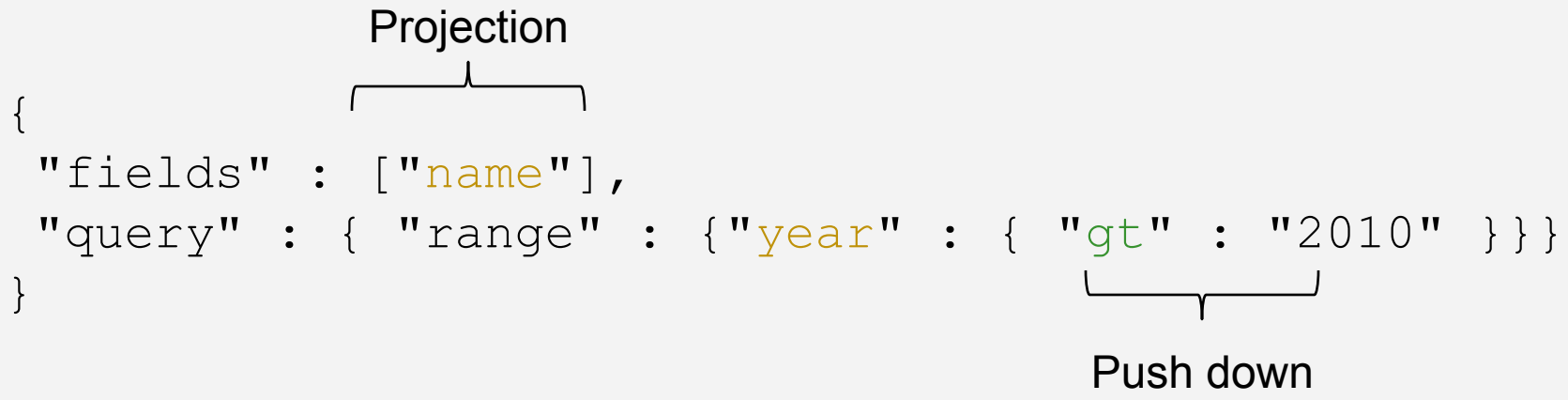


```
{
 "fields" : ["name"],
 "query" : { "range" : {"year" : { "gt" : "2010" }}}
}
```


elastic

Query DSL Conversion

```
SELECT name FROM playlist WHERE year > 2010;
```



elastic

Query optimization

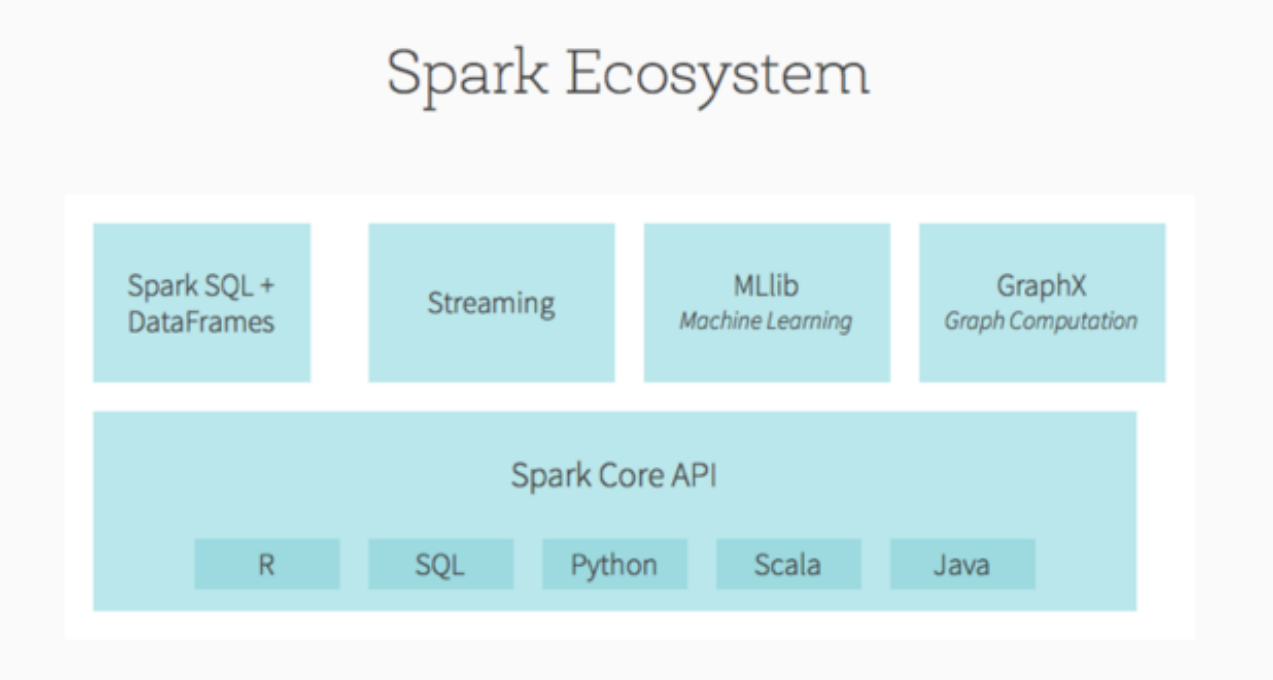
Where possible, convert the API query to ES query

Return only the results without any intermediate data

Library / API	Operation Awareness
Map/Reduce	Does not apply
Cascading	Projection
Apache Pig	Projection
Apache Hive	Projection
Apache Storm	Projection
Apache Spark	Projection & Push-Down

elastic

Apache Spark



elastic

ES-Hadoop native Spark integration

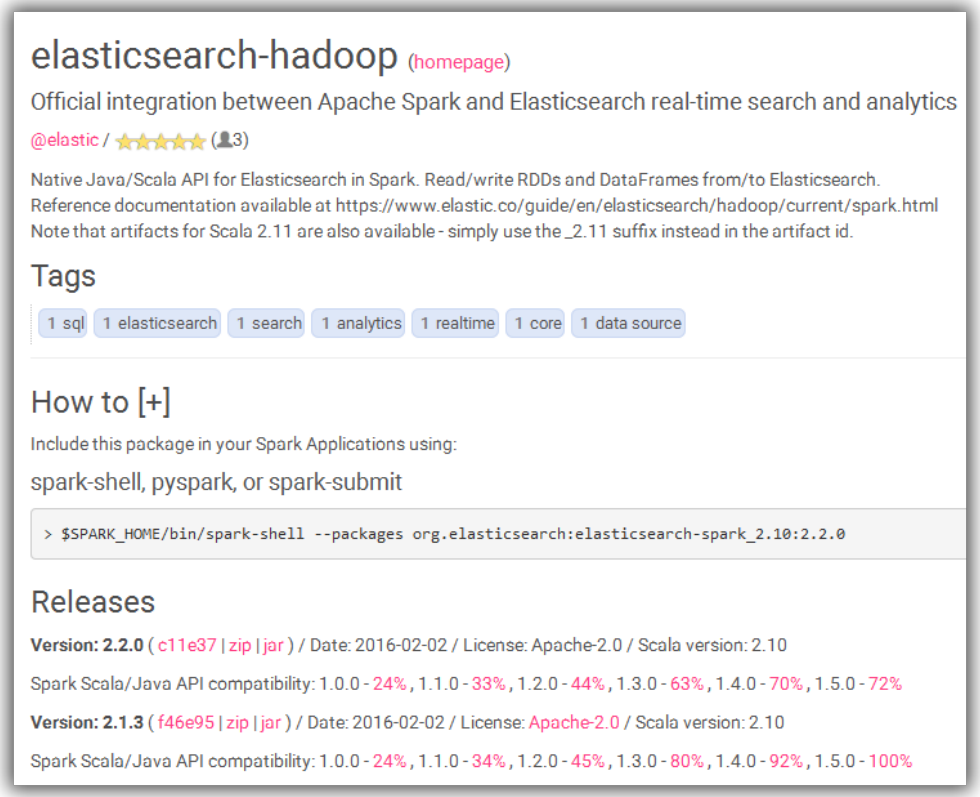
Offers both Scala & Java API

Available as Spark package

Supports Spark Core & SQL

all 1.x version (1.0-1.6)

Available for Scala 2.10 and 2.11



elasticsearch-hadoop (homepage)

Official integration between Apache Spark and Elasticsearch real-time search and analytics

@elastic / (3)

Native Java/Scala API for Elasticsearch in Spark. Read/write RDDs and DataFrames from/to Elasticsearch.
Reference documentation available at https://www.elastic.co/guide/en/elasticsearch/hadoop/current/spark.html
Note that artifacts for Scala 2.11 are also available - simply use the _2.11 suffix instead in the artifact id.

Tags

1 sql | 1 elasticsearch | 1 search | 1 analytics | 1 realtime | 1 core | 1 data source

How to [+]

Include this package in your Spark Applications using:

spark-shell, pyspark, or spark-submit

```
> $SPARK_HOME/bin/spark-shell --packages org.elasticsearch:elasticsearch-spark_2.10:2.2.0
```

Releases

Version: 2.2.0 (c11e37 | zip | jar) / Date: 2016-02-02 / License: Apache-2.0 / Scala version: 2.10

Spark Scala/Java API compatibility: 1.0.0 - 24%, 1.1.0 - 33%, 1.2.0 - 44%, 1.3.0 - 63%, 1.4.0 - 70%, 1.5.0 - 72%

Version: 2.1.3 (f46e95 | zip | jar) / Date: 2016-02-02 / License: Apache-2.0 / Scala version: 2.10

Spark Scala/Java API compatibility: 1.0.0 - 24%, 1.1.0 - 34%, 1.2.0 - 45%, 1.3.0 - 80%, 1.4.0 - 92%, 1.5.0 - 100%

spark-packages.org/package/elastic/elasticsearch-hadoop

Apache Spark – Resilient Distributed Dataset (RDD)

```
import org.elasticsearch.spark._

val sc = new SparkContext()

sc.esRDD("buckethead/albums", ?q=pikes")
```

```
import org.elasticsearch.spark._

case class Album(name: String, year: Long)
val lent = Map("name" -> "Celery", "year" -> 2014)
val onMyDesk = Album("Electric Tears", 2002)

sc.makeRDD(Seq(lent, onMyDesk)).saveToEs("buckethead/albums")
```

Apache Spark – JSON RDDs

```
jsonRDD : RDD[(String, String)]

jsonRDD = sc.esJsonRDD("buckethead/albums", "?q=pikes")
```

```
val p95 = """{"name" : "Hold Me Forever", "year" : 2014 }"""
val p180 = """{"name" : "Heaven Is Your Home", "year" : 2015}"""

sc.makeRDD(Seq(p95, p180)).saveJsonToEs("buckethead/albums")
```

Apache Spark SQL – DataFrames



"Spark SQL is Spark's module for working with structured data"

RDD + schema = DataFrame (inspired by Python Pandas)

Allows usage of SQL

Integrates with Hive*

** trivia – the project was initially based on Hive (Shark)*

elastic

Apache Spark SQL Support

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

val df = sqlCtx.read.format("es").load("buckethead/albums")

df.filter(df("category").equalTo("pikes").and(df("year").gt(2015)))
```

```
CREATE TEMPORARY TABLE dfAsTable USING org.elasticsearch.spark.sql
                OPTIONS ('path' = 'buckethead/albums');

SELECT name FROM dfAsTable WHERE year > 2015 and category = "pikes";
```

```
val df = sqlContext.read.json("buckethead/2015/albums.json")
df.saveToES("buckethead/albums")
```

Spark SQL to Query DSL

```
df.filter(df("category").equalTo("pikes").and(df("year").geq(2015)))
```

```
{ "query" :
   { "bool" : { "must" : [
      "match" : { "category" : "pikes" }
    ],
    "filter" : [
      { "range" : { "year" : {"gte" : "2015" }}}
    ]
  }}
}
```

Advanced Spark SQL features in ES-Hadop

Spark SQL 1.3 - 1.6 (`DataFrame`)

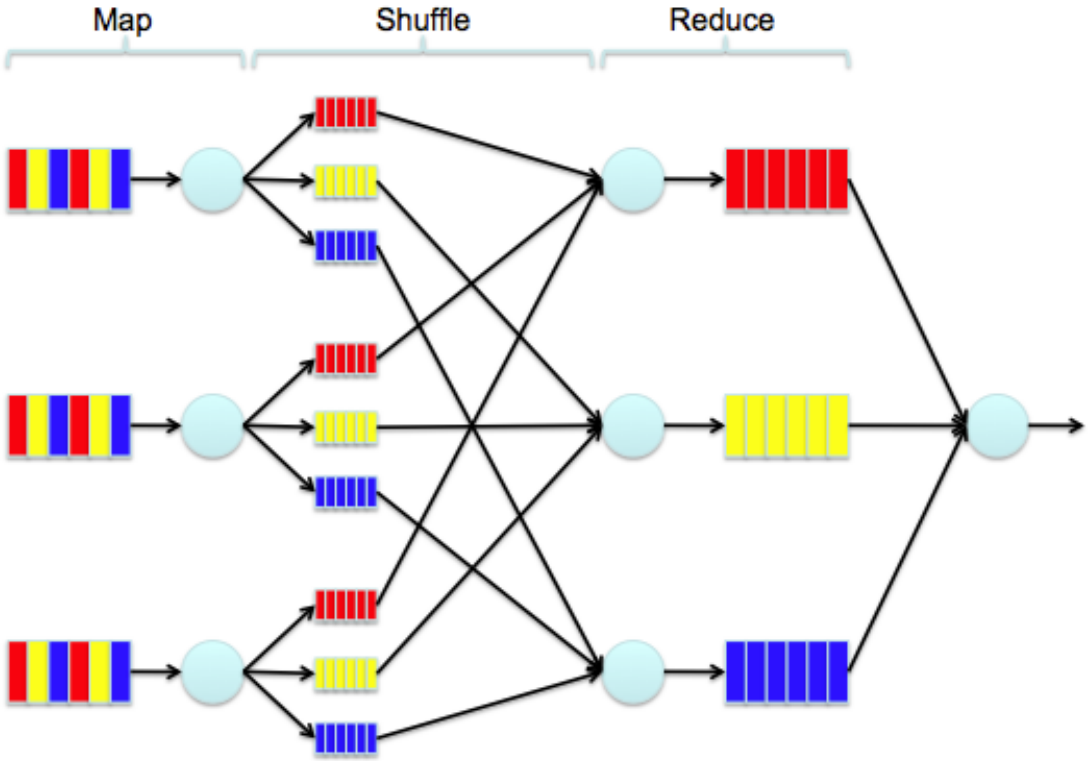
Spark SQL 1.1 - 1.2 (`SchemaRDD`)

Supports all filters in Spark SQL

- EqualTo/EqualNullSafe
- GreaterThan/GreaterThanOrEqual/LessThan/LessThanOrEqual
- In/ IsNull/ IsNotNull
- And/Or/Not
- StringStartsWith/StringEndsWith/StringContains

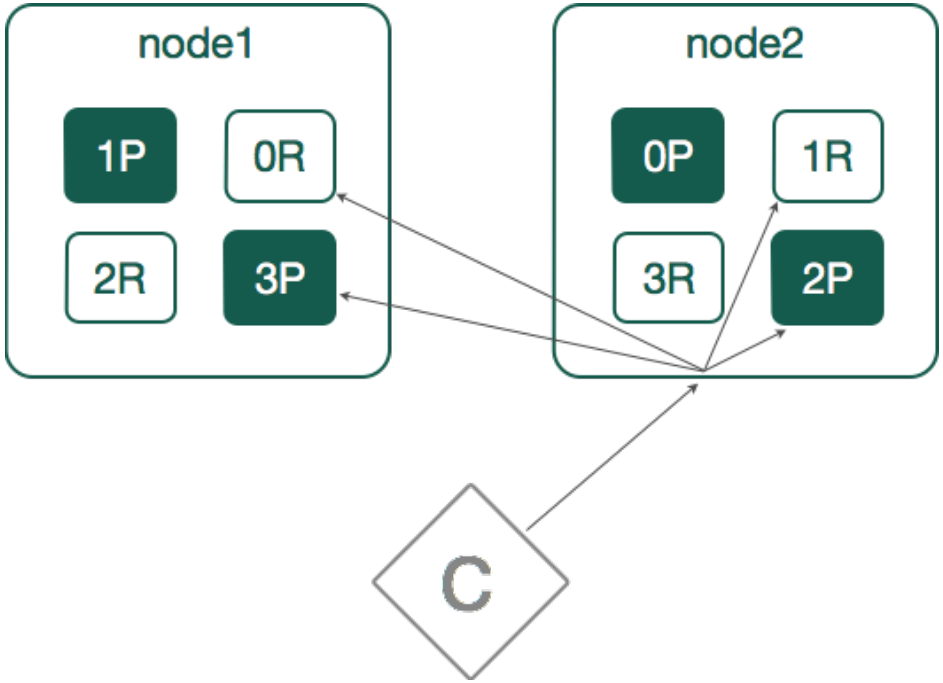
How it works?

Map / Reduce overview



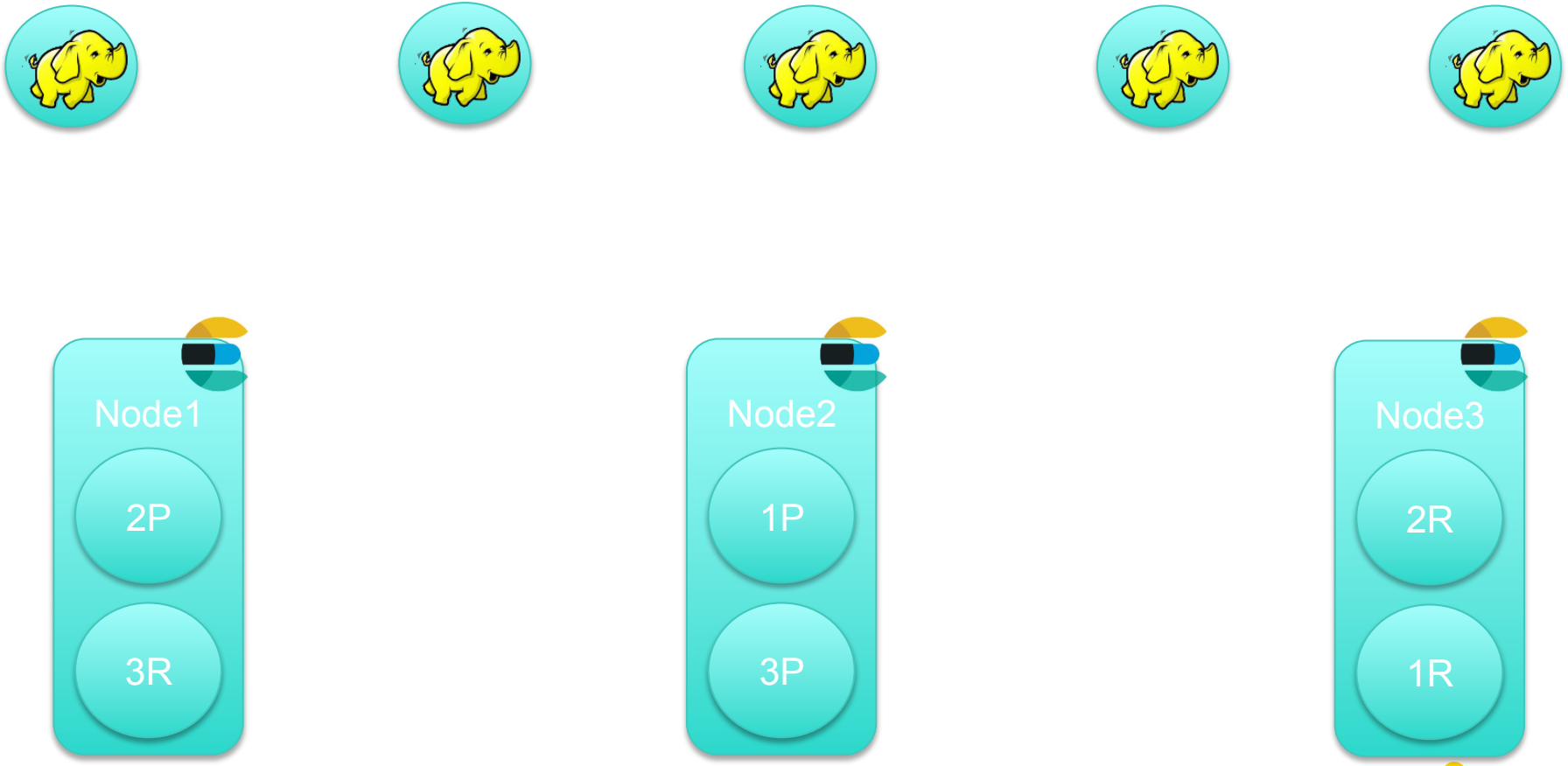
elastic

Elasticsearch Overview



elastic

Partition to partition architecture



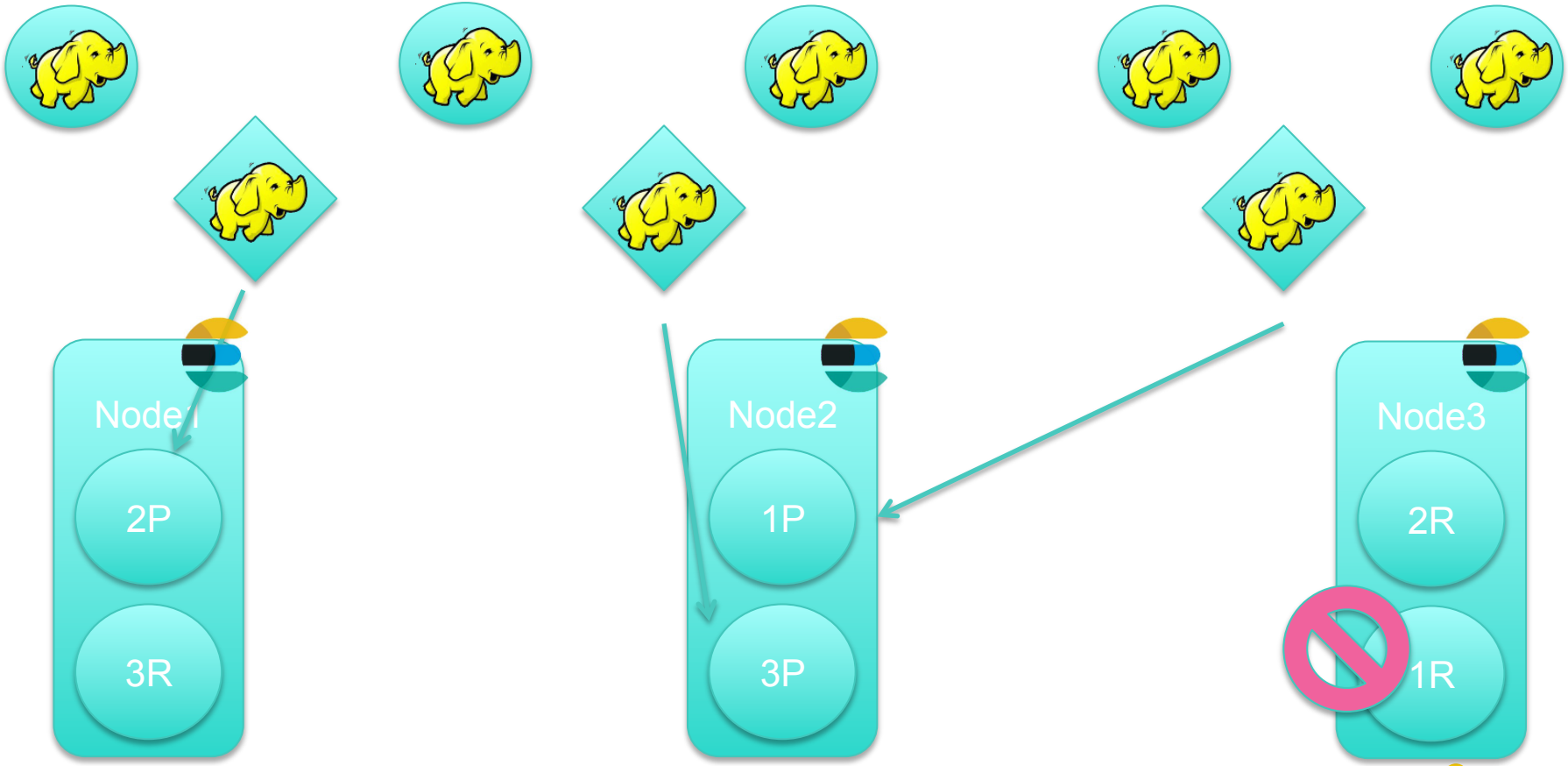
elastic

Dynamic runtime matching



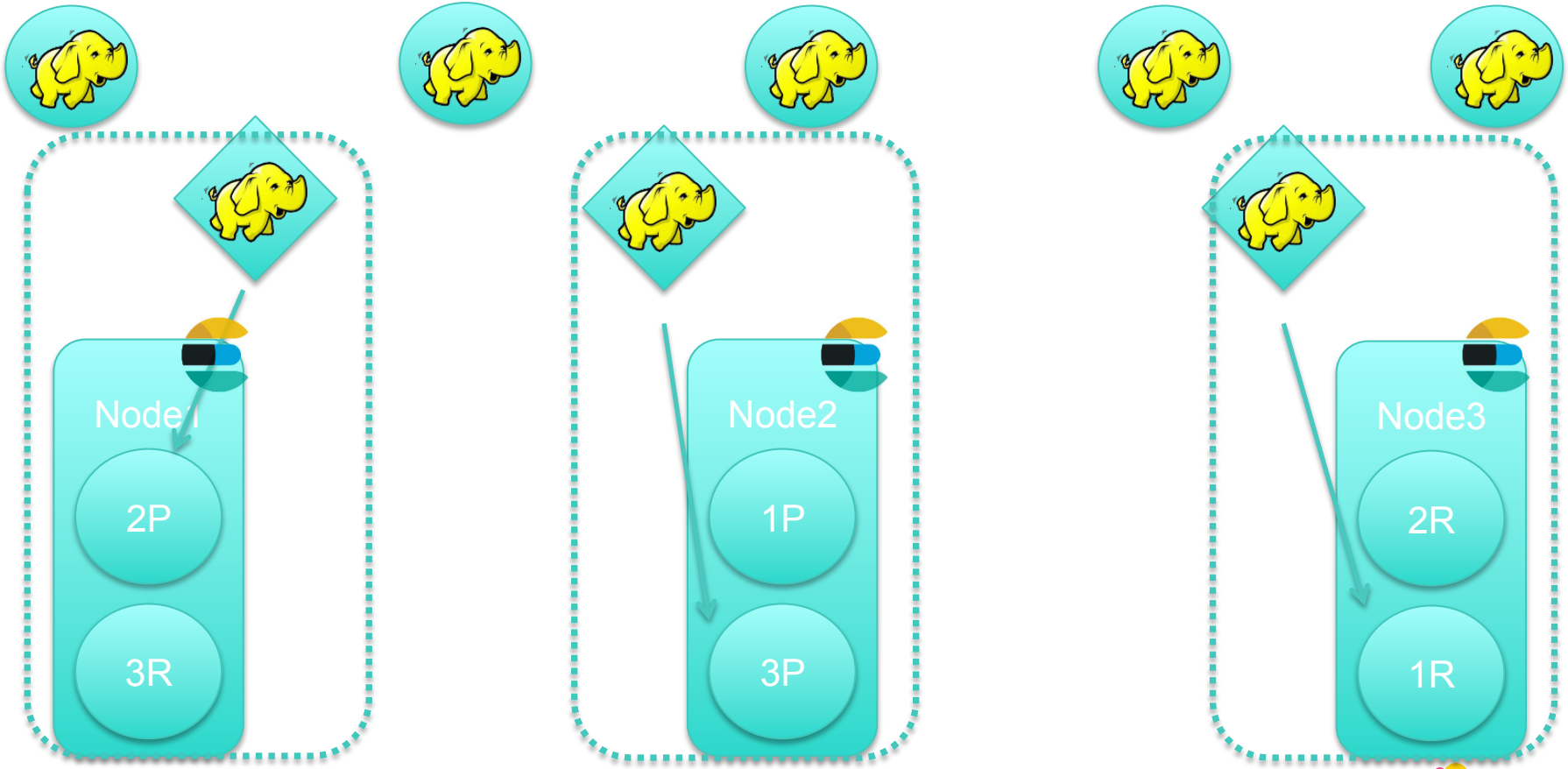
elastic

Failure handling



elastic

Co-location



elastic

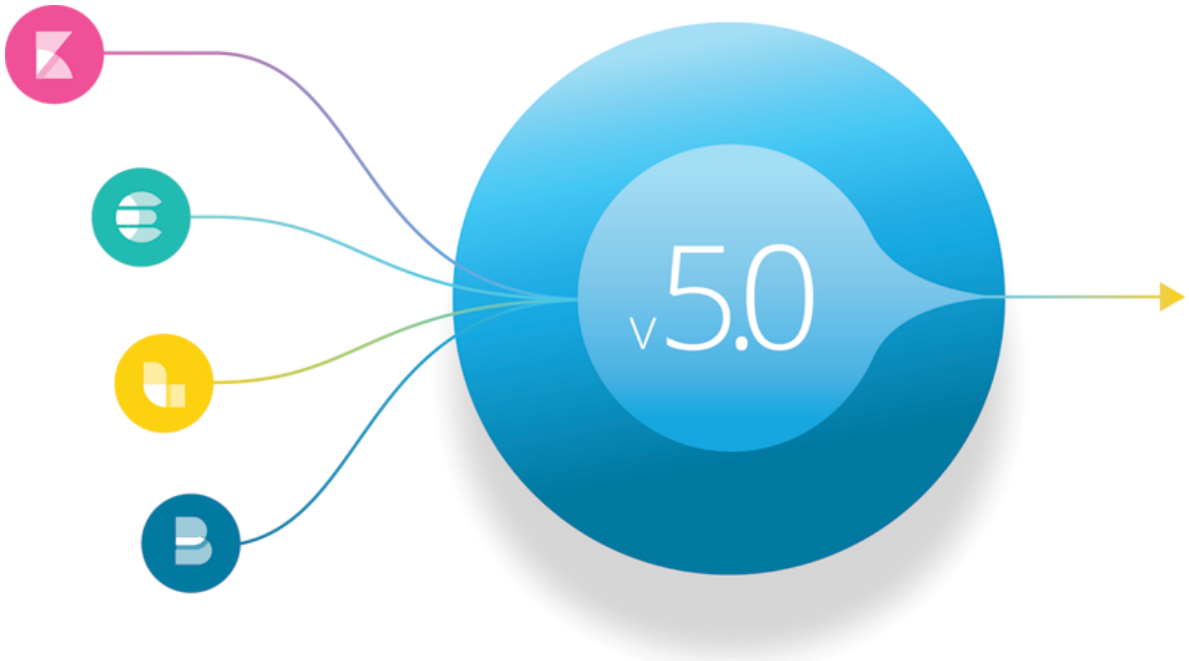
Conclusion

- Bigdata rocks
  - Long-term storage in Hadoop
  - Rich batch process against bigdata(Full-Size)
- Real-time matters
  - Elasticsearch as a transient data store
  - Quick insights and analytics
  - Rich visualization by using Kibana



elastic

Elastic Stack v5 is coming!



https://www.elastic.co/v5

elastic

Community

- 源码 & Issue: http://github.com/elastic/
- 英文社区: http://discuss.elastic.co
- 中文社区: http://elasticsearch.cn
- 官方 QQ 群: 190605846
- 下载: https://www.elastic.co/downloads
- 博客: https://www.elastic.co/blog
- 线下活动: http://elasticsearch.meetup.com/
- IRC: #elasticsearch, #logstash, #kibana, #beats
- 官方 Twitter @elastic



elastic



Thank You