

江蘇





# جستجوی هوشمند با Elasticsearch

ارائه‌دهنده: حدیثه نقوی  
مهسا اشرفی راضیه بهمن یار

نام استاد: دکتر وحیدی پور

# فهرست

درخواست‌های پایه

مراحل نصب

آنالیزورها

مقایسه با Solr

تعریف اولیه

تاریخچه

مفاهیم پایه

اهمیت ایجاد Shard

ویژگی‌ها

کاربرد



# تعریف اولیه

- ◀ موتور جستجوی full-text و آنالیز توزیع شده
- ◀ متن باز و نوشته شده توسط زبان برنامه نویسی جاوا
- ◀ دارای قابلیت multitenancy
- ◀ « ارائه سرویس به تعداد زیادی tenant توسط یک نمونه از برنامه
- ◀ ساخته شده بر پایه کتابخانه های Apache Lucene
- ◀ استفاده شده در وبسایت های مشهور نظیر گیت هاب، موزیلا، استک

اورفلو

# تاریخچه

- ◀ ایده اولیه در زمان انتشار نسخه سوم Compass  
« توسط Shay Banon  
« باز نویسی مجدد برنامه جهت تبدیل آن به برنامه جستجوی توزیع شده  
« قابل استفاده از طریق پروتکل HTTP  
« دریافت و ارسال اطلاعات به فرمت JSON
- ◀ انتشار نخستین نسخه برنامه در فوریه 2010
- ◀ پایه گذاری شرکت Elasticsearch BV در سال ۲۰۱۲  
« جهت ارائه سرویس ها و محصولات تجاری در حیطه کاری  
Elasticsearch

# جایگاه الاستیک سرچ

Rank			DBMS	Database Model	Score		
Dec 2017	Nov 2017	Dec 2016			Dec 2017	Nov 2017	Dec 2016
1.	1.	1.	Oracle +	Relational DBMS	1341.54	-18.51	-62.86
2.	2.	2.	MySQL +	Relational DBMS	1318.07	-3.96	-56.34
3.	3.	3.	Microsoft SQL Server +	Relational DBMS	1172.48	-42.59	-54.17
4.	4.	4.	PostgreSQL +	Relational DBMS	385.43	+5.51	+55.41
5.	5.	5.	MongoDB +	Document store	330.77	+0.29	+2.09
6.	6.	6.	DB2 +	Relational DBMS	189.58	-4.48	+5.24
7.	7.	↑ 8.	Microsoft Access	Relational DBMS	125.88	-7.43	+1.18
8.	↑ 9.	↑ 9.	Redis +	Key-value store	123.24	+2.05	+3.34
9.	↓ 8.	↓ 7.	Cassandra +	Wide column store	123.21	-1.00	-11.07
10.	10.	↑ 11.	Elasticsearch +	Search engine	119.78	+0.37	+16.51

# مفاهیم پایه (1)

Near Realtime یا NRT ◀

« نیاز به صرف زمان خیلی کم از شروع شاخص‌بندی سند تا امکان جستجوی آن

Cluster ◀

« مجموعه‌ای از یک یا چند گره (سرور)

« نگهداری تمام داده‌ها به صورت جمعی

« قابلیت جستجو و شاخص‌بندی جداگانه در تمامی گره‌ها

Document ◀

« واحد پایه اطلاعات قابل شاخص‌گذاری

## مفاهیم پایه (2)



### ◀ Node (گره)

« یک سرور و بخشی از یک کلاستر  
« شرکت در ذخیره‌سازی، جستجو و شاخص‌بندی

### ◀ Index (شاخص)

« شامل مجموعه‌ای از اسناد با ویژگی‌های مشابه

### ◀ Type (نوع)

« دسته‌بندی منطقی شاخص‌ها



## مفاهیم پایه (3)

### Shard ◀

- « راه حلی برای مقابله با محدودیت‌های سخت‌افزاری گره‌ها
- مثال: نیاز به یک ترابایت فضا برای شاخص یک میلیارد سند
- « قرار دادن شاخص روی قسمت‌های مختلف
- « به خودی خود شاخصی مستقل و با کارایی کامل
- « قابل قرارگیری بر روی هر کدام از گره‌های کلاستر

### Replica ◀

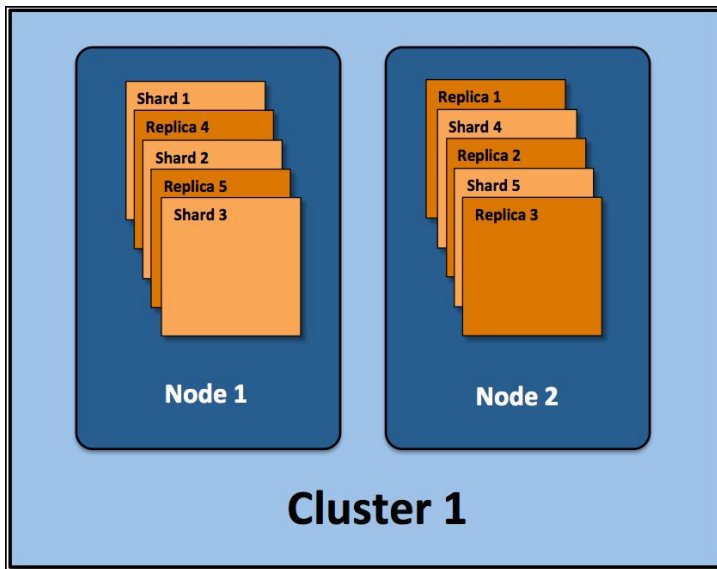
- « کپی کاملی از یک shard روی گره‌های دیگر
- « افزایش ضریب اطمینان برنامه در هنگام بروز مشکل

# اهمیت ایجاد Shard

◀ امکان توزیع محتوا به صورت افقی

◀ موازی سازی عملیات روی چندین گره

◀ افزایش کارایی و خروجی



# ویژگی‌ها (1)

- ◀ سرعت بالا در ساخت شاخص و پاسخگویی به جستجوها
- ◀ مقیاس‌پذیری
- ◀ « قابل استفاده برای حجم بالای داده‌ها
- ◀ سادگی در نصب، مدیریت و بیکربندی
- ◀ عدم نیاز به تعریف اولیه ساختار داده‌ها (Schemaless)
- ◀ کارایی بالا

## ویژگی‌ها (2)

◀ امکان ذخیره، جستجو و آنالیز حجم عظیمی از داده‌ها به صورت آنی

◀ قابلیت استفاده به عنوان تکنولوژی زیرساخت

» نیاز مبرم برنامه‌های کاربردی به جستجو پیشرفته و سریع

◀ قابلیت پشتیبانی از کوئری‌های پیچیده

# کاربرد

- ◀ در فروشگاه‌های آنلاین
- ◀ « جستجو در میان حجم عظیمی از محصولات و دستیابی به مشخصات محصول
- ◀ آنالیز داده‌های مربوط به تراکنش‌ها و لاگ‌ها
- ◀ « پیدا کردن الگوها و نقاط غیرنرمال
- ◀ « تحلیل‌های آماری
- ◀ استفاده از قابلیت Reverse Search
- ◀ « فراهم‌آوری امکان تعیین بازه قیمت محصولات
- ◀ استفاده از Kibana
- ◀ « ساخت داشبوردهای تجاری و به کارگیری هوش تجاری (BI)
- ◀ -ارسال درخواست‌ها به <http>

# مراحل نصب

- ◀ نیاز به نصب جاوا نسخه 8 به بالا
- ◀ دانلود فایل برنامه به صورت زیپ یا tar.gz از وبسایت [elastic.co](http://elastic.co)
- ◀ اجرای فایل `elasticsearch.bat` (ویندوز) / `elasticsearch/` (لینوکس)
- ◀ راه اندازی سرور به صورت پیش فرض روی پورت 9200
- ◀ ارسال تقاضاها به REST API سرور
- ◀ دریافت و ارسال تقاضاها به فرمت JSON

# بررسی آنالیزورها

1. Char Filter
2. Tokenizer
3. Token Filter

- هرکدام از این بخش ها برای حل کردن بخشی از مشکلات تحلیل زبان بکار می روند

## کارکرد بخش char filter

- حل مشکل نیم فاصله در زبان فارسی
- اعداد اعشاری

## کارکرد بخش tokenizer

- تشخیص اعداد فارسی
- تشخیص توکن های تاریخ زمان

## کارکرد بخش token filter

- Stemmer
- Stop word
- Normalizer
- Word deliment
- synonym



# مقایسه با Solr

- ◀ دارای کتابخانه‌های رسمی Javascript، PHP، Groovy، Java و NET. « در مقابل Solr فقط دارای کتابخانه جاوا »
- ◀ قابلیت ورود داده از منابع مختلف نظیر پایگاه داده‌ها و سیستم‌های مختلف « در مقابل پشتیبانی از تعداد محدودی منبع »
- ◀ امکان استفاده از کوئری‌های پیچیده DSL
- ◀ قابلیت جستجو معکوس (ارسال کوئری و سپس ارسال سند برای بررسی تطابق)
- ◀ استفاده آسان‌تر و راحت‌تر نسبت به Solr

# سوال



...با تشکر از توجه شما

