Distributed Indexing

CISC-525
Phil Grim

HARRISBURG UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Overview

- Usage
- Implementations
  - Solr
  - ElasticSearch

Usage of Distributed Indexing

- Enables full text and field-based searching of large quantities of unstructured, semi-structured, and structured data
- Complex searches such as faceting, ranges, temporal searching, geospatial searching
- Same concepts apply to indexing as to distributed file systems and NoSQL databases
  - Sharding
  - Replication
  - Parallel processing
  - Fault tolerance

Implementation

- Solr and ElasticSearch are the two most popular implementations
- Both based on Apache Lucene for actual indexing
- Share many features, differ in others

Solr

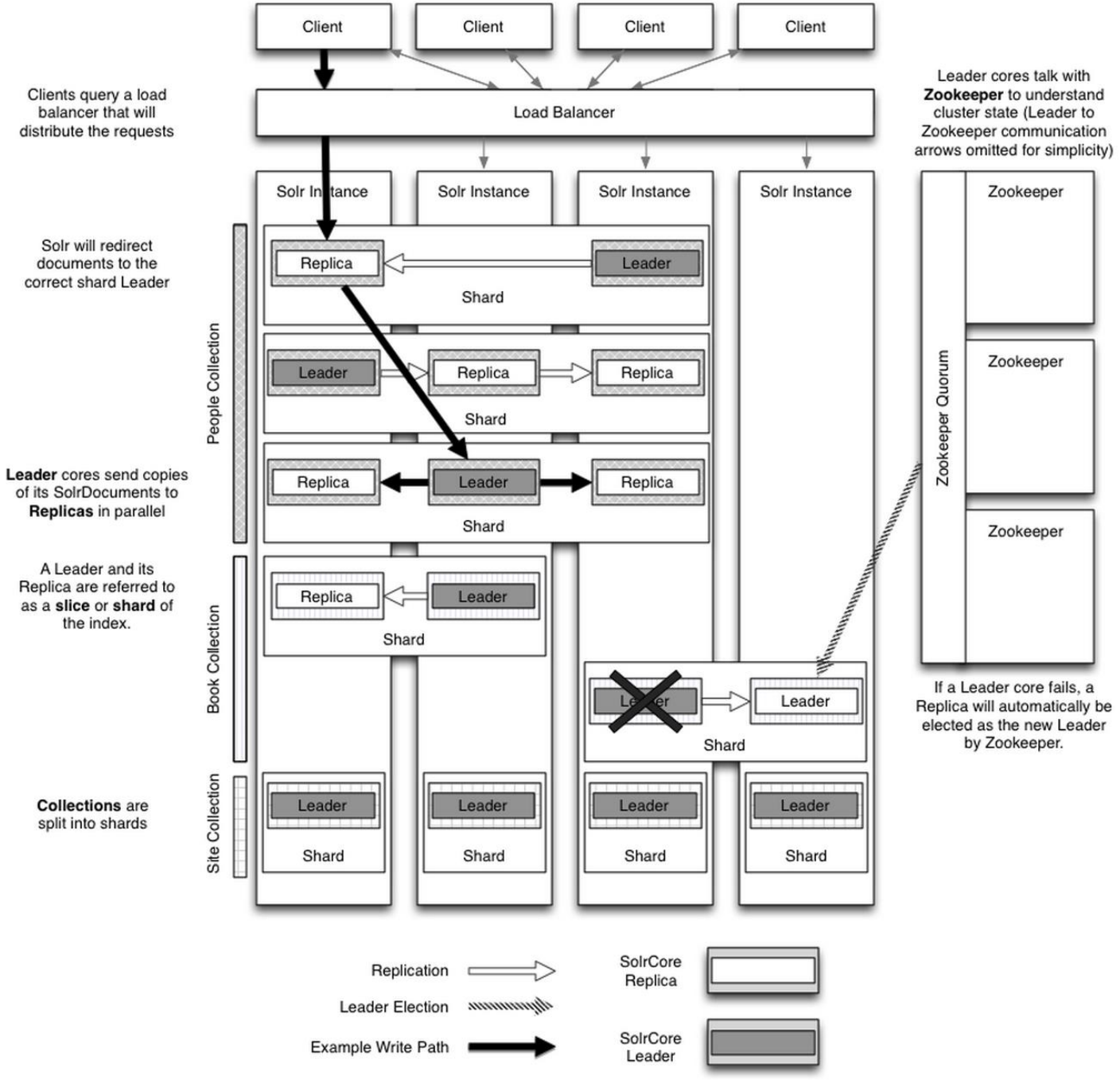
- Apache open-source sub-project under the Lucene project
- Exposes Lucene's Java API as a REST service
- Includes its own Java API
- Adds additional features on top of standard Lucene searching

Solr Features

- Full Text Search
- Faceted Navigation of Search Results
- Recommendation
- Autocomplete/Spelling suggestion
- Custom document ranking
- Snippet Generation
- Highlighting
- More...

Solr Architecture

- ▶Can be standalone, or clustered (SolrCloud)
- ▶Automatic Failover - If a single node goes down, its index is replicated on a different node using a backup
- ▶Consistency - updates to the index must be directed to the correct shard so that one consistent view of the document is maintained
- ▶Automatic shard partitioning - Only needs to know the number of shards, forwards updates to the correct index
- ▶Simple Configuration - Uses ZooKeeper for centralized configuration for the cluster.



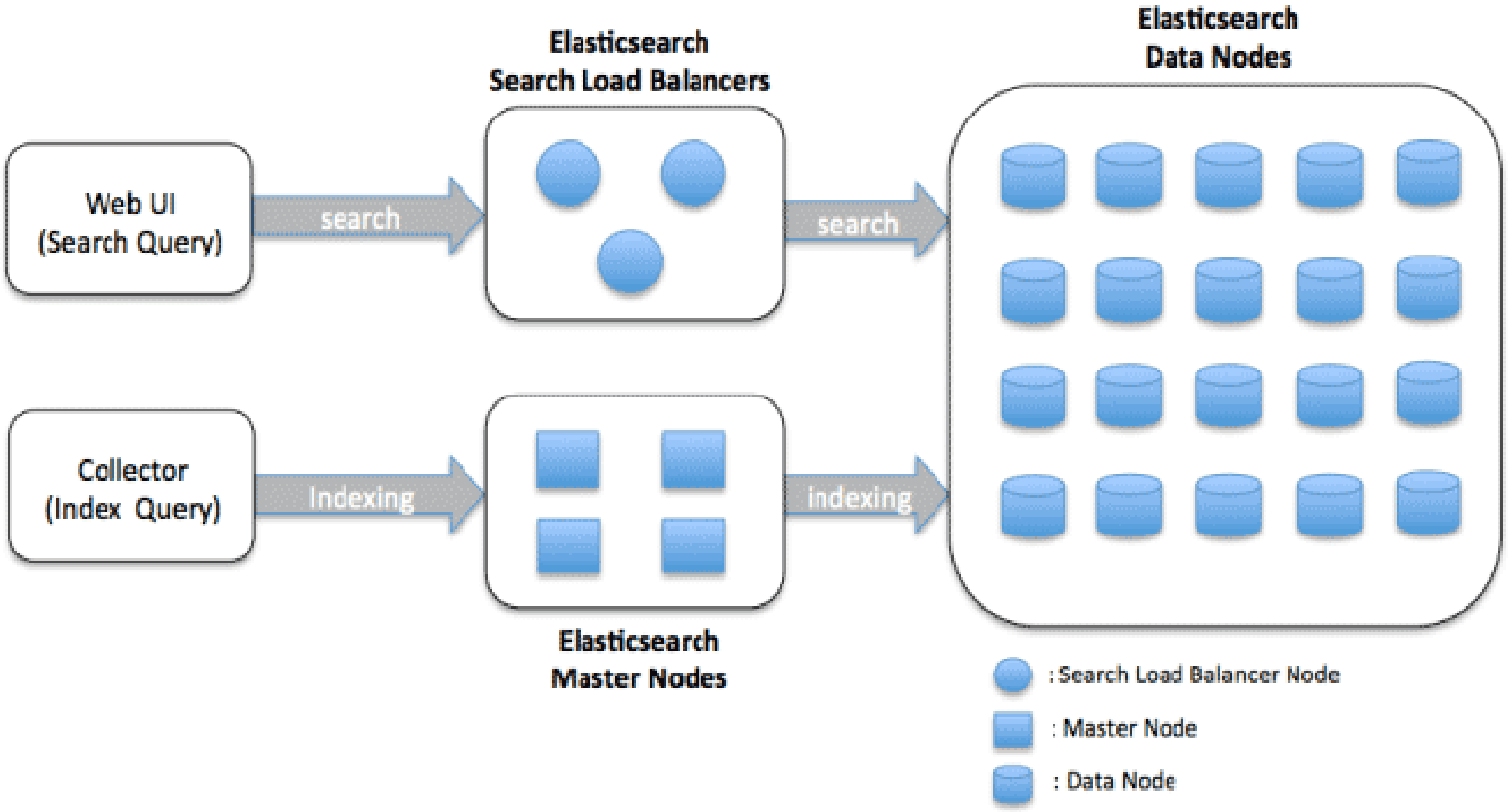
ElasticSearch

- Developed by Shay Banon, originally called Compass, an abstraction layer for Lucene search engine.
- Provides APIs in many languages, including Java, Python, Ruby, REST, and more
- Open Source under the Apache 2 License
- Company known as Elastic formed to provide commercial support for ElasticSearch
- Unlike Solr, was designed from the ground up to be a distributed index
- Is the search engine behind Wikipedia

ElasticSearch Features

- Full Text Search
- Faceted Navigation of Search Results
- Recommendation
- Autocomplete/Spelling suggestion
- Custom document ranking
- Snippet Generation
- Highlighting
- More...

ElasticSearch Architecture



Continued Reading

Solr Website
http://lucene.apache.org/solr

ElasticSearch Website
https://www.elastic.co/products/elasticsearch