

# SEARCHING BILLIONS OF PRODUCT LOGS IN REAL TIME

Ryan Tabora - Think Big Analytics  
NoSQL Search Roadshow - June 6, 2013

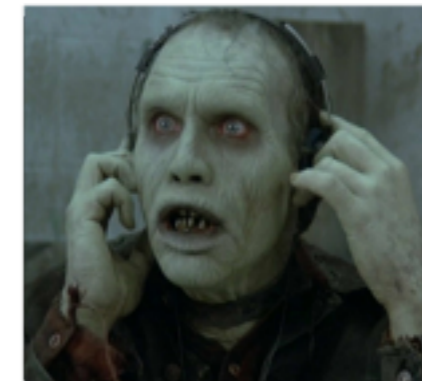
# WHO AM I?



Ryan Tabora

Think Big Analytics - Senior Data Engineer

Lover of dachshunds, bass, and zombies



# OVERVIEW

Primers

What are product logs?

How do they apply to big data?

Real use case

Real issues and designs

Conclusion

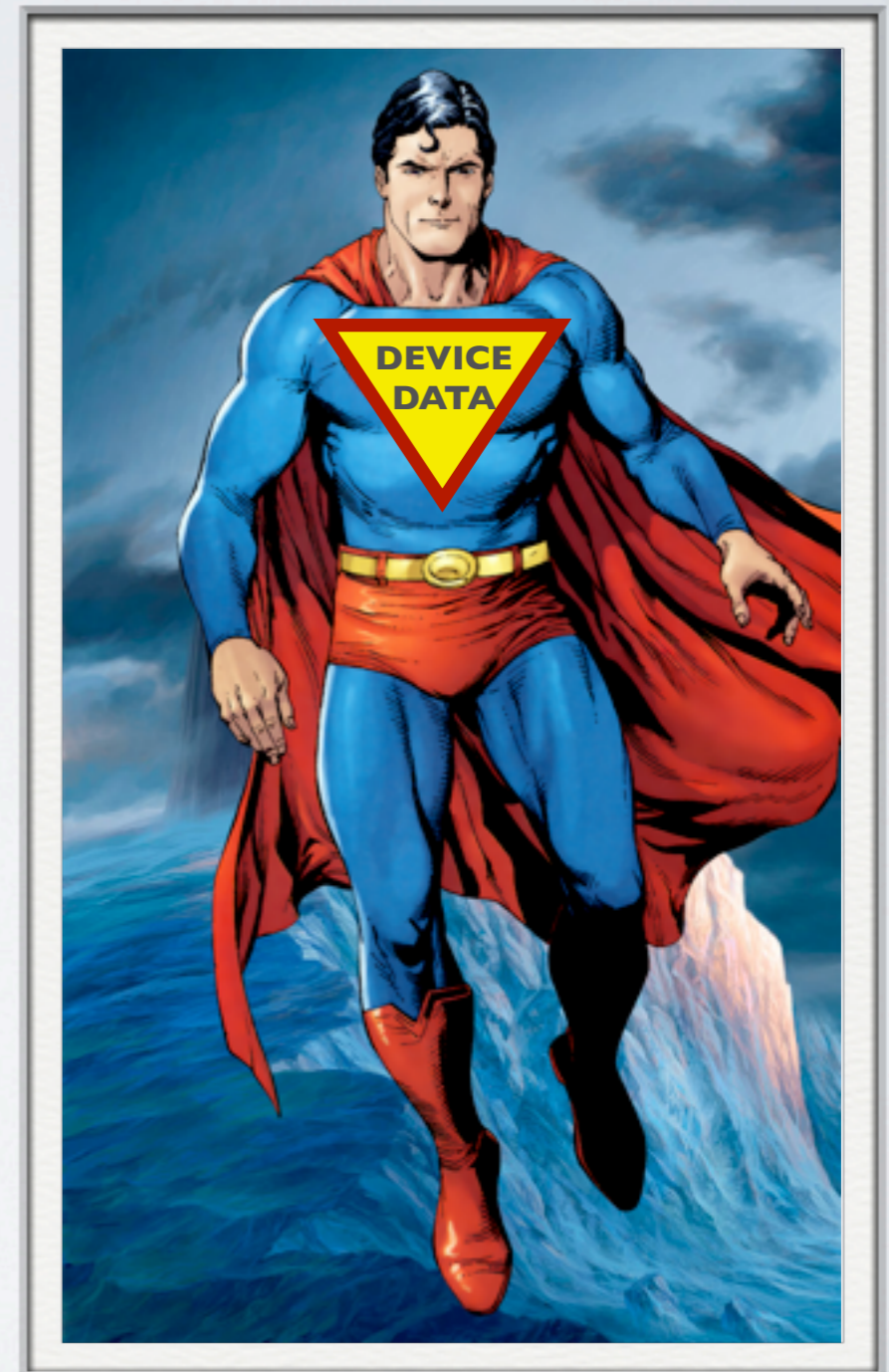
# PRODUCT LOGS?

- Device data
- IT, Energy, Healthcare, Manufacturing, Telecom ...
- These devices are pushing data back home (pull works too!)
- As more devices are sold/installed, more and more data comes back to 'home base'



# POWER OF DEVICE DATA

- Realtime Visualization
- Realtime Response
- Ad Hoc Analysis
- Full Historical Capture
- Blended Data Sets

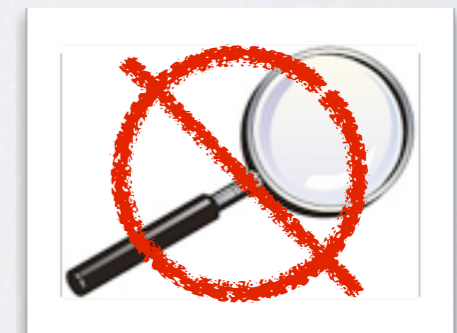
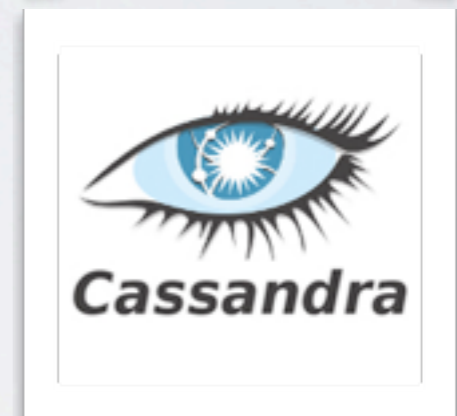
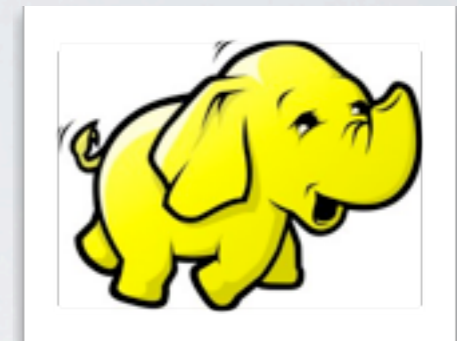


# TRADITIONAL APPROACHES

- SQL: PostGres, MySQL, Oracle, Microsoft
- SQL provides many of the search features required for typical search applications
  - Joins, regex, group by, sorting, etc
- But the these technologies can only scale so far..

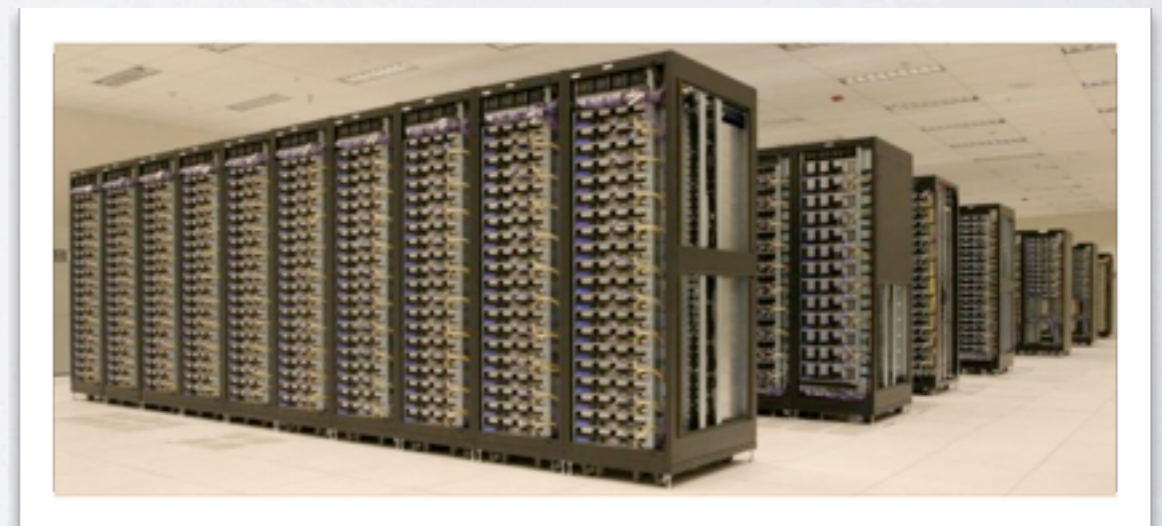
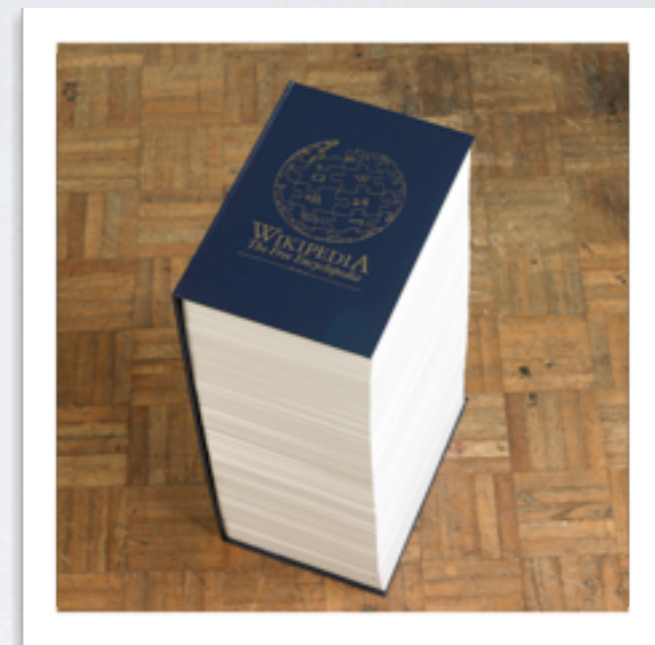
# NEW TECHNIQUES STORING DATA

- Hadoop
- HBase/Cassandra/accumulo
- Search features are very limited
  - HBase row scans, primary key index
  - Cassandra limited secondary indexing



# NEW TECHNIQUES INDEXING DATA

- What is an index?
  - Lucene
- Paralleling Index Creation
  - MapReduce/Flume/Storm
- Real Time Search
  - Searching before it hits disk





# NEW TECHNIQUES SEARCHING DATA

- Solr/ElasticSearch

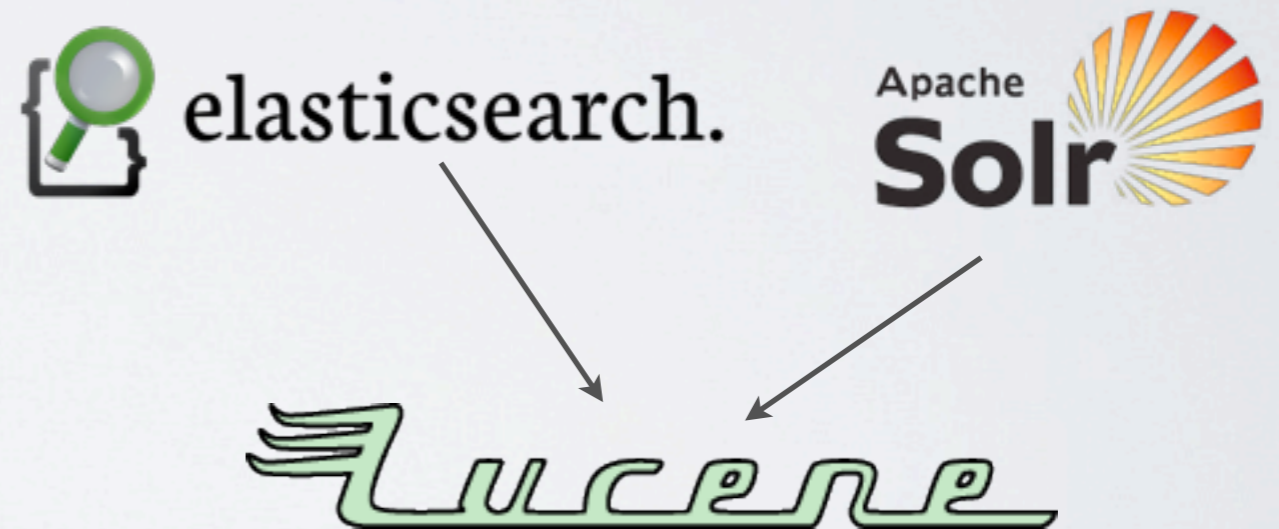
- Both build on top of Lucene

- Search servers

- RESTful HTTP APIs

- Easy to administer

- Add powerful text/numerical search capabilities



# BASIC SEARCH FEATURES

- Boolean logic (AND, OR + -)
- Sorting and Group By
- Range queries
- Phrase/Prefix/Fuzzy queries

# ADVANCED SEARCH FEATURES

- Custom ranking/scoring
- More like this
- Auto suggest
- Faceting/Highlighting
- Geo-spacial search

# SCALING SEARCH

- ElasticSearch and SolrCloud both have distributed features built in
  - Auto-sharding
  - Replication
  - Query routing
  - Transaction log



# USE CASE

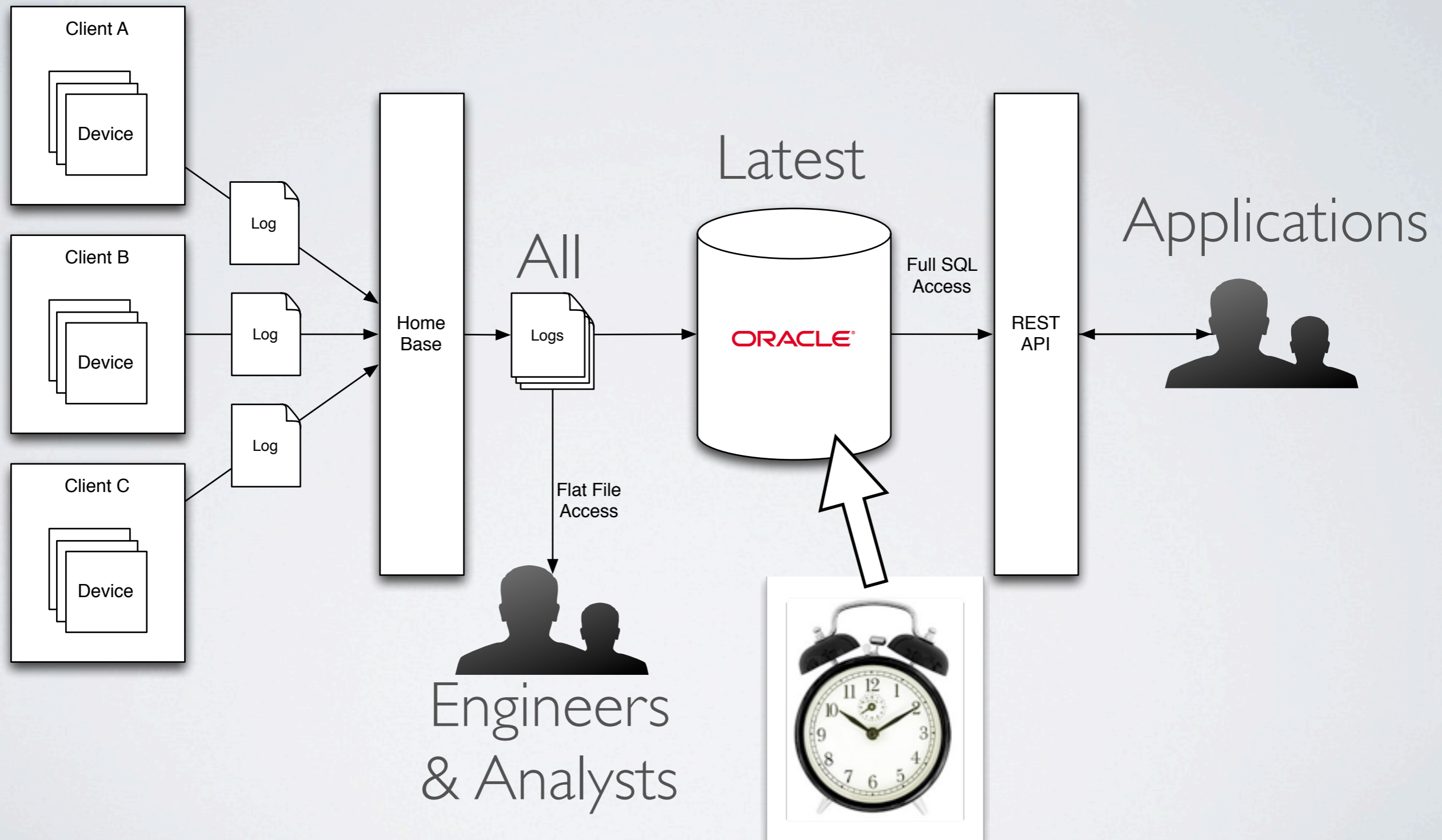
Problem

Sample Solution

Core Design Issues

Other Solutions

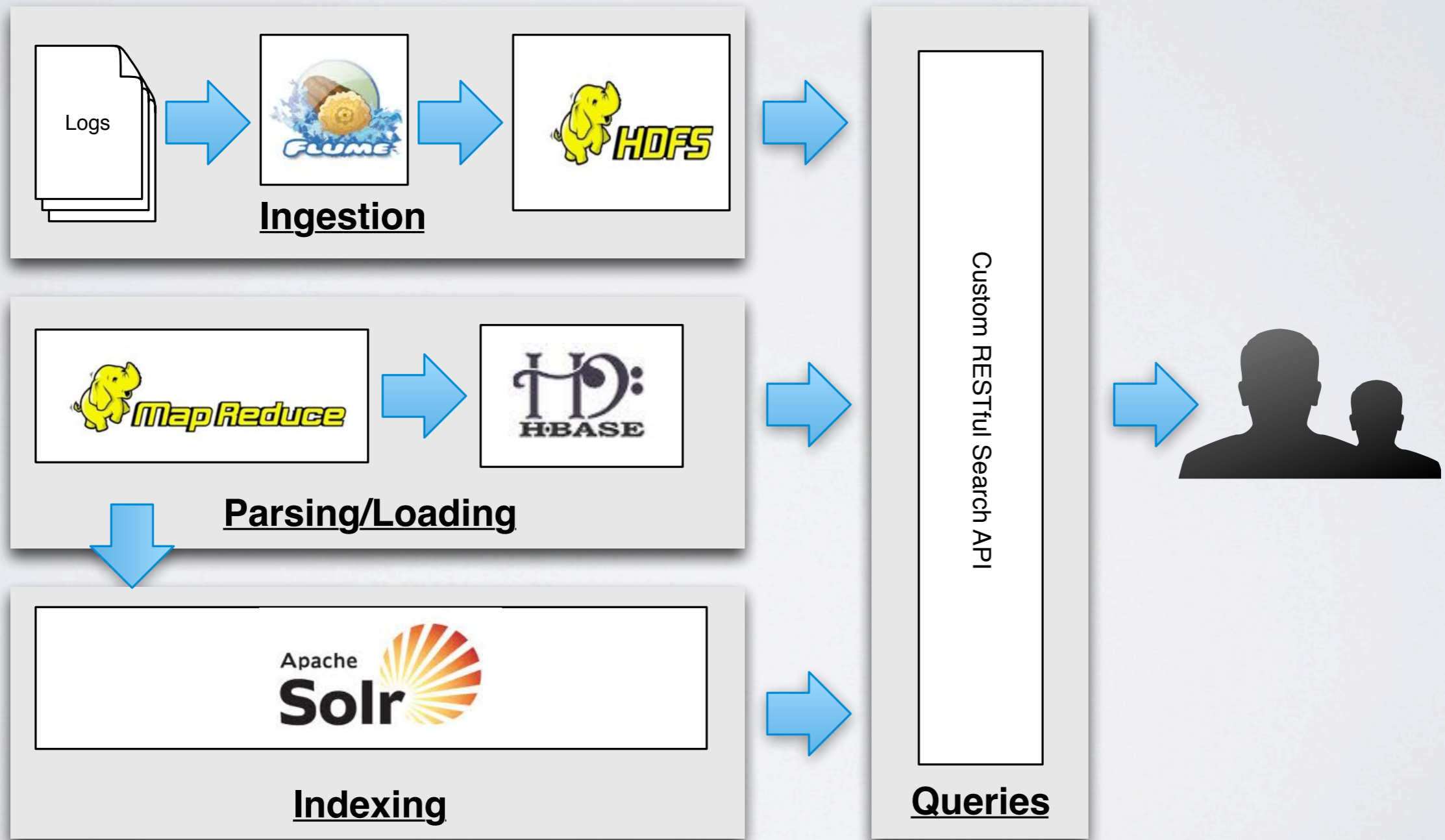
# THE PROBLEM



# SEARCH APPLICATION FEATURES

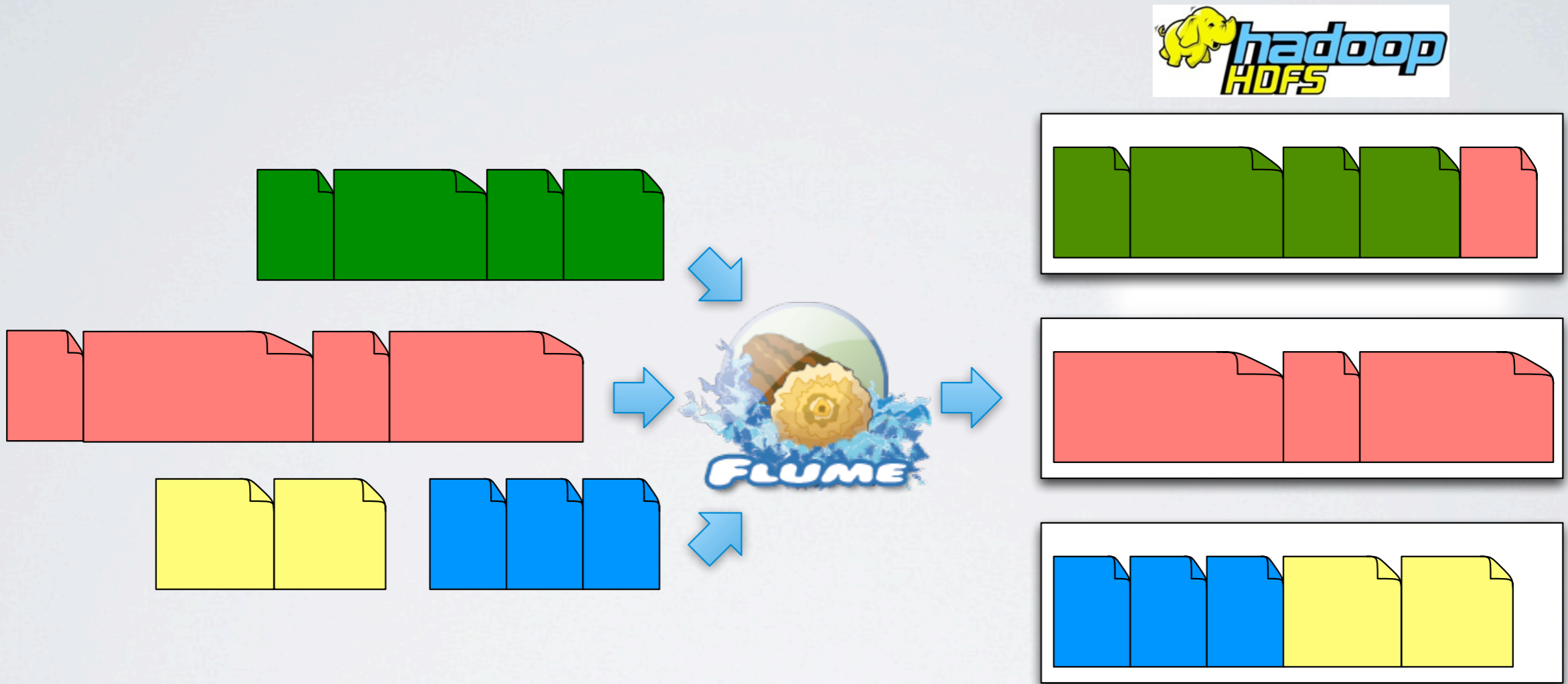
- Find last three days of raw logs from an entire cluster
- Group capacity available grouped by machine serial number and show the largest capacities first
- Search all device header lines for “FAILURE”
- View all hard disk objects that have product number 2341AB
- Find all motherboards with an associated customer ticket

# SAMPLE SOLUTION

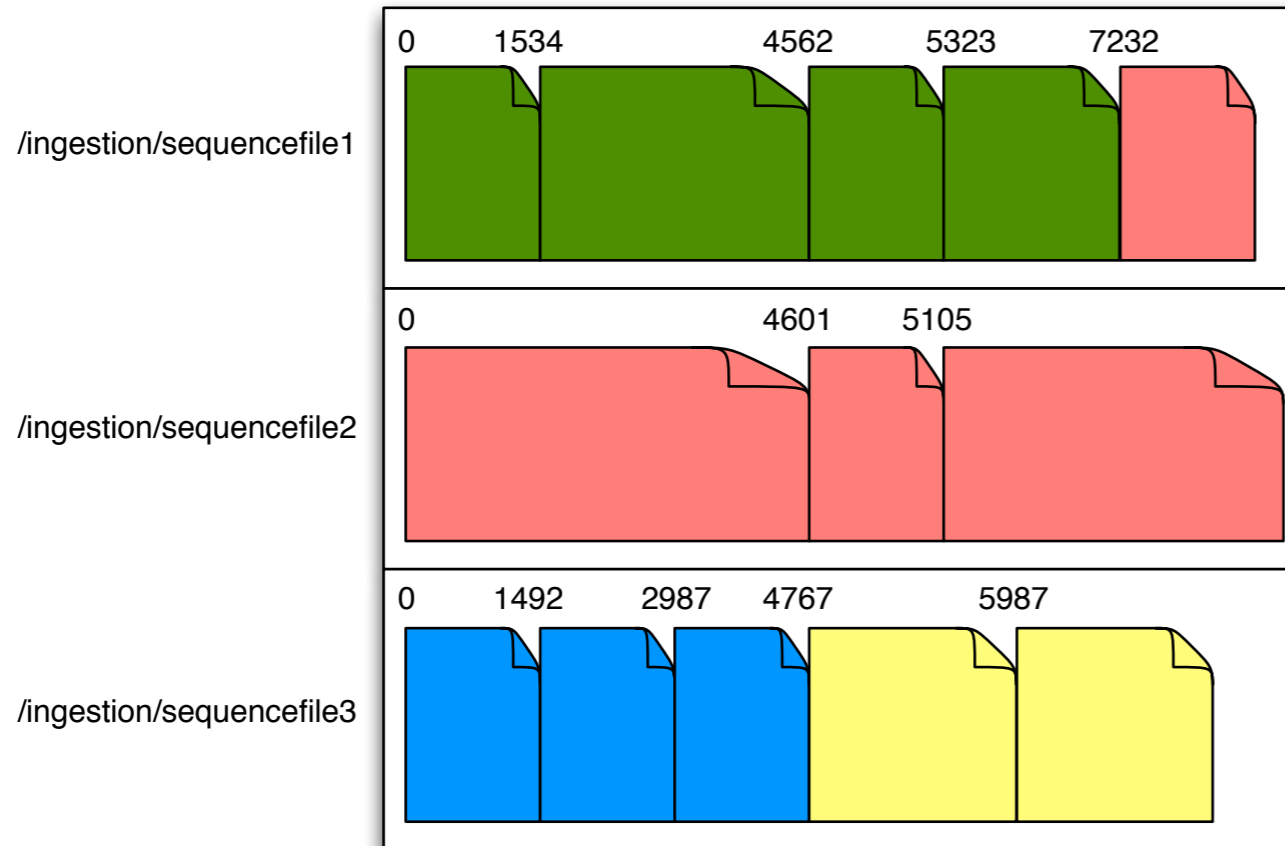




# INGESTION



# PARSING, LOADING, AND INDEXING



Load HBase with parsed objects




Store HBase ROW\_ID  
Store pointer to raw file in HDFS  
Index a number of desired fields

# INSIDE OF HBASE

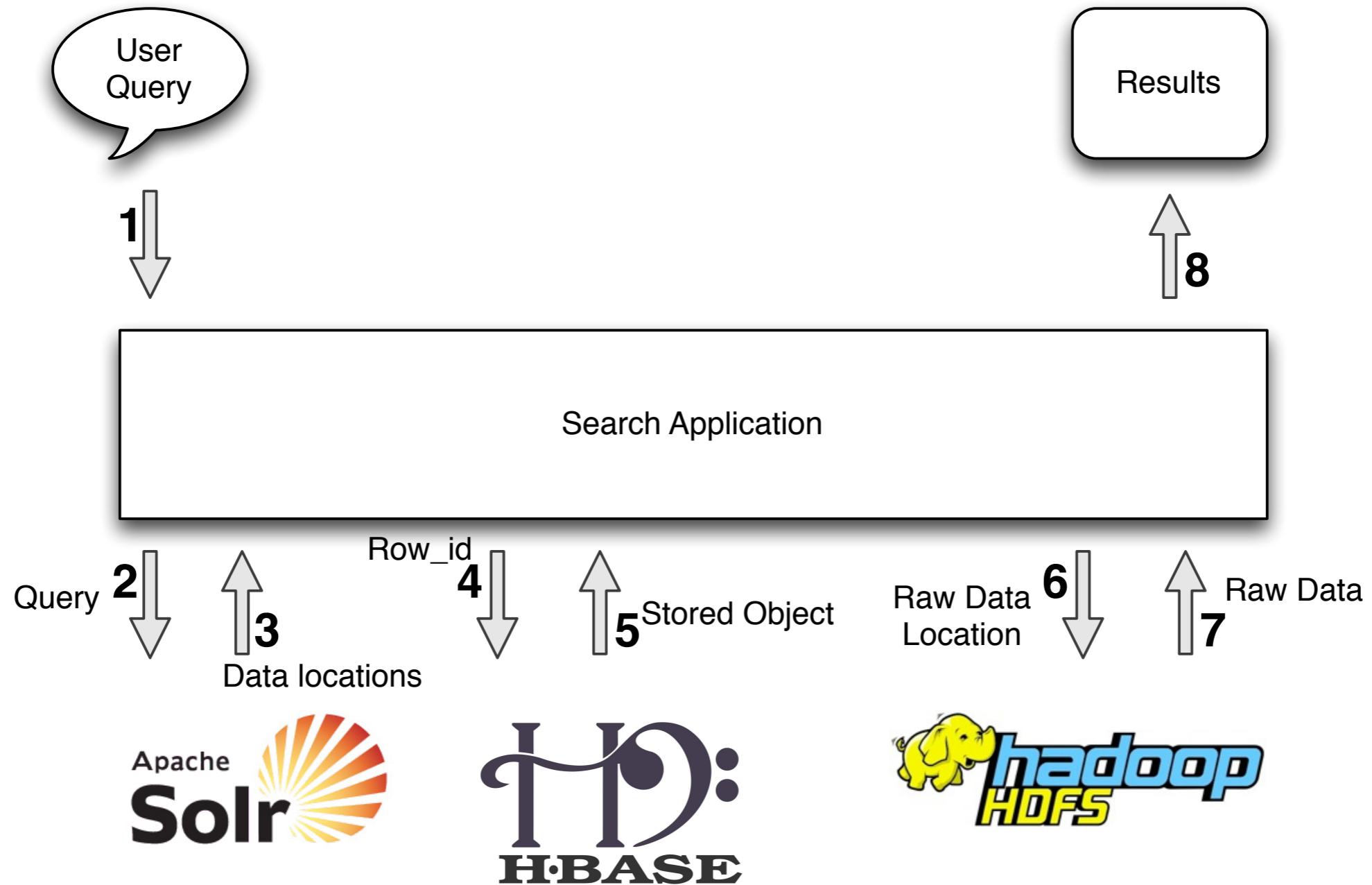
rowkey	object1	object2	object3	object4	object5
	...		...	...	...
	...	...	...	...	...
		...		...	...
...	...		...	...	

# THE SOLR DOCUMENT

## Solr Document

rowkey	
sequenceFile	/ingest/file2
sfOffset	2343
cluster_id	1333-2241-3411
system_id	42ADFF-BZMM
date_sent	2013/05/12
file_name	configs.log
contents	...
header	WARNING: DISK DEAD
...	...

# SEARCH APPLICATION



# CORE DESIGN ISSUES

- Changing the Solr schema (manual reindex)
- Elastic shard scaling (manual reindex)
- No distributed joining (denormalizing the data)
- Replication\*
- Manually managing Solr partitioning/sharding\*
- Write durability\*

# SOLRCLOUD



- Automatic shard creation, routing
- Replication
- Limited to a fixed number of shards defined on initial creation
- ZooKeeper for coordination
- Large community

# ELASTICSEARCH

- Similar feature set to Solr
- Purpose built for easily managing a distributed index
- Rapidly growing community
- Custom built coordination mechanism
- JSON based API





# DATASTAX ENTERPRISE

- Integrates Cassandra and Solr
- Automatic indexing in Solr/storing in Cassandra
- Automatic partitioning
- Automatic reindexing
- Not limited to fixed number of shards
- Proprietary and costs money



# CONCLUSION

- Collecting and analyzing device data/product logs can be a very difficult challenge
- You can use NoSQL and search technologies like Solr or ElasticSearch in unison...
- ...but it is not always easy to integrate search with NoSQL

# QUESTIONS?

- Feel free to reach out if you have any questions or need help with big data/search!
- <http://ryantabora.com>
- <http://thinkbiganalytics.com>
- <http://www.slideshare.net/ratabora>
- [ryan.tabora@thinkbiganalytics.com](mailto:ryan.tabora@thinkbiganalytics.com)
- @ryantabora



# BONUS SLIDES

# HBASE AND SOLR

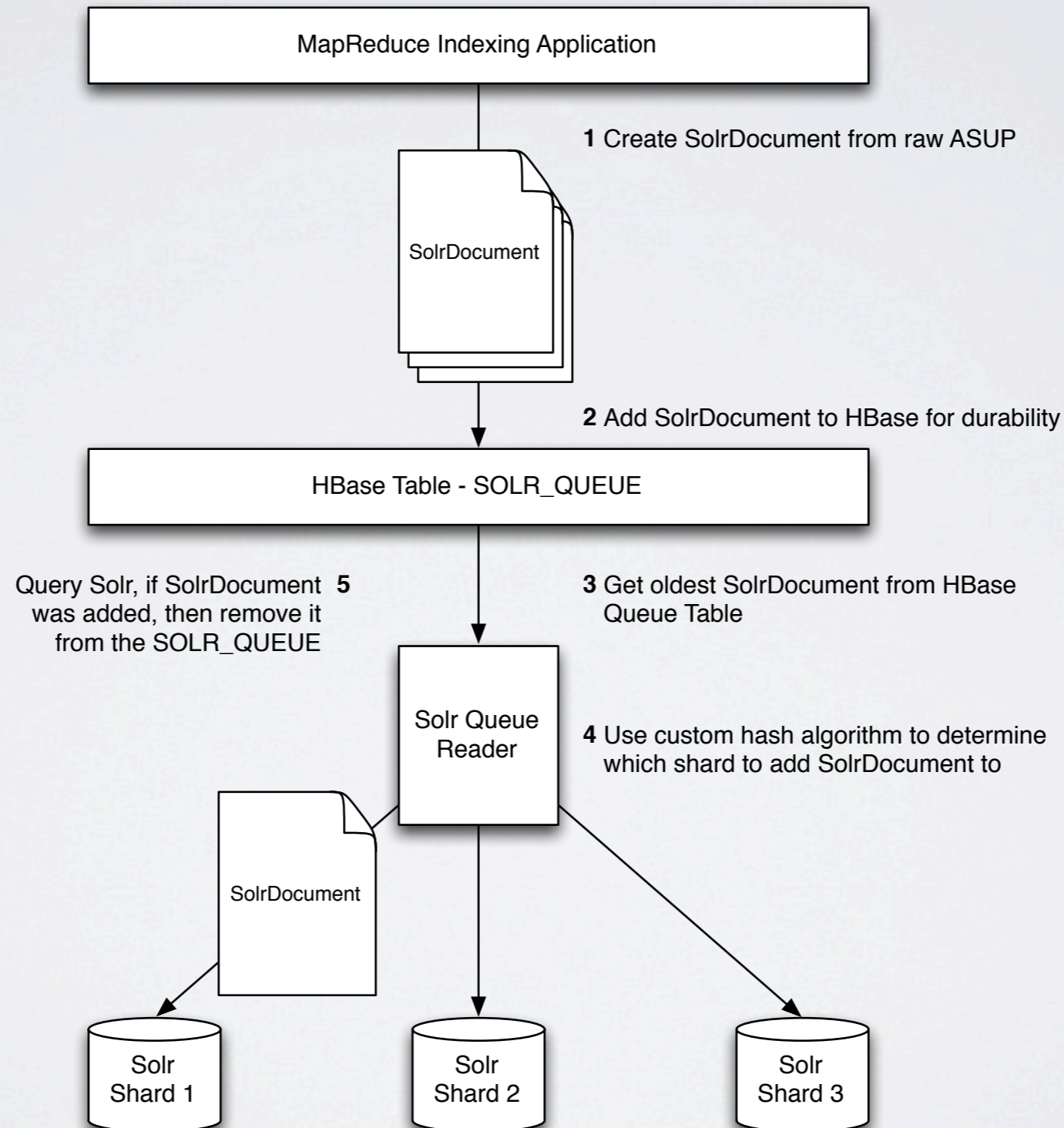
- Automatic partitioning/reindexing
- Automatic index updates on HBase inserts/deletes
- Mapping HBase cells to a Solr schema
- No perfect commercial/open source solution yet
- Many many many more...

# HBASE + SOLR

## AUTOMATIC INDEXING

- HBase coprocessors are like storedprocs/triggers
  - New, powerful, and dangerous
- Triggers on HBase puts/deletes
- Mapping data to a schema?

# HBASE + SOLR WRITE DURABILITY



# HBASE + SOLR ELASTIC SHARDING

- HBase's distributing mechanism uses the concept of regions to split data across many nodes
- Region splitting can be automatic or manual (performance degradation as regions split)
- Piggybacking Solr sharding on HBase Region splitting