# 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 - Big Data Consultant

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

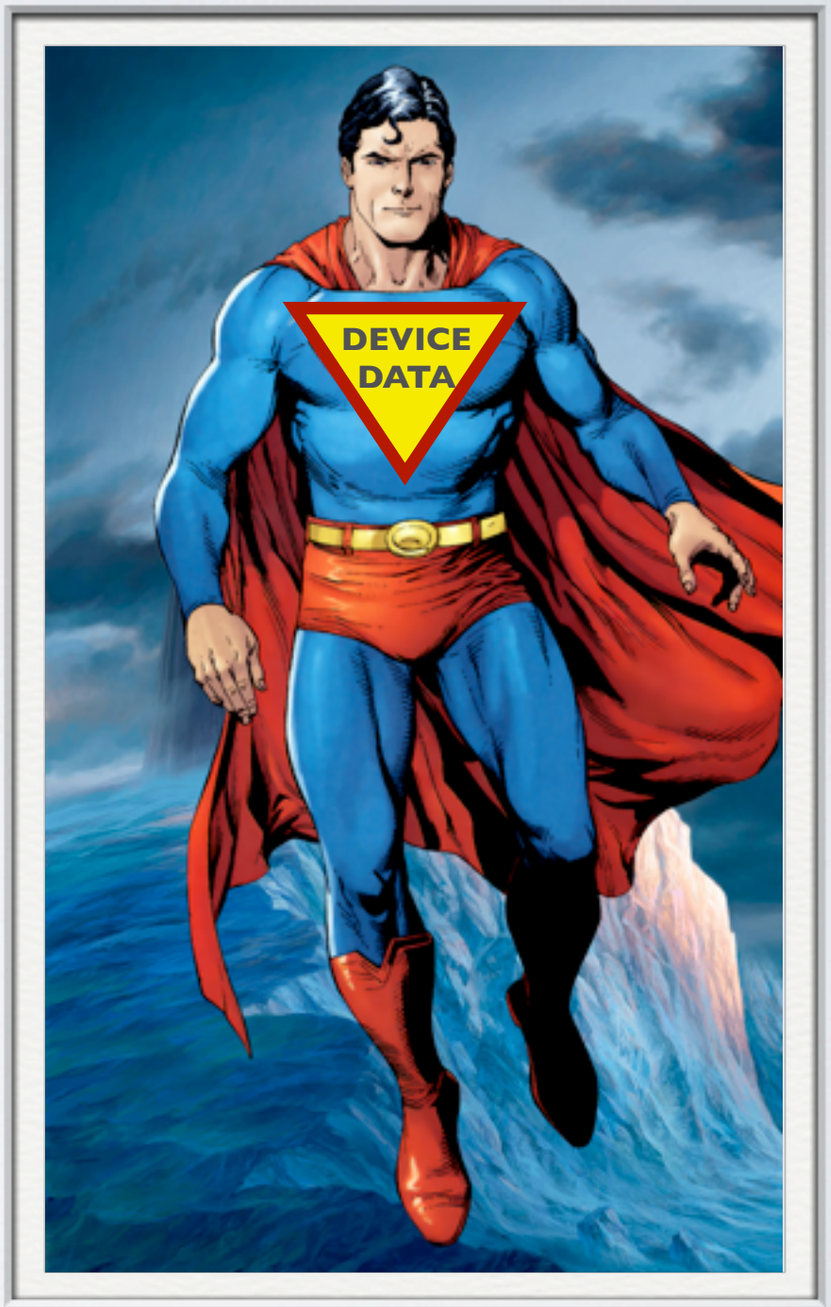
Conclusion

# 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

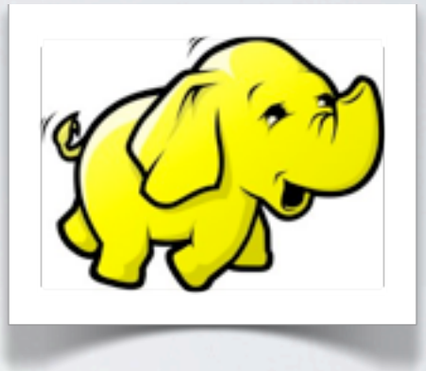
- SQL provides many of the search features required for typical search applications

  - Joins, regex, group by, sorting, etc

- But the these technologies can only scale so far...

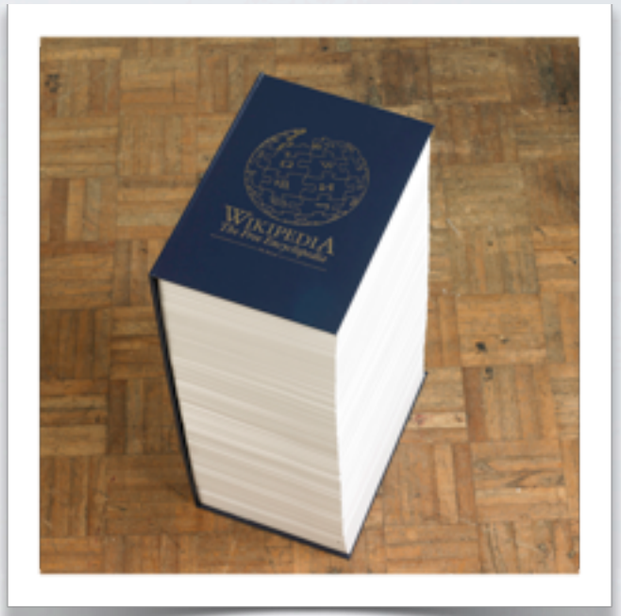
# NEW TECHNIQUES STORING DATA

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



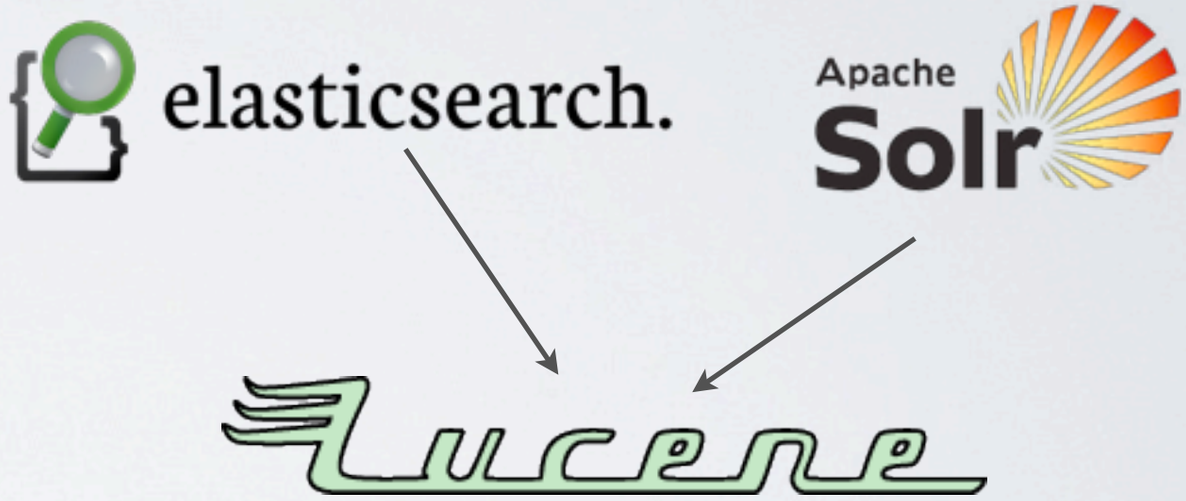

# NEW TECHNIQUES INDEXING DATA

- What is an index?
  - Lucene
- Indexing data across a cluster
  - MapReduce
- Real time search
  - Solr/ElasticSearch




# NEW TECHNIQUES SEARCHING DATA

- Solr/ElasticSearch




- Real time indexing/querying
  - Based on Lucene
- Powerful text/numerical search capabilities

# BASIC SEARCH FEATURES

- Secondary indexing with boolean logic (AND, OR + -)
- Sorting and Group By
- Range queries
- Phrase/Prefix/Fuzzy queries
- Faceting/Highlighting

# ADVANCED SEARCH FEATURES

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

# SCALING SEARCH

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

# USE CASE

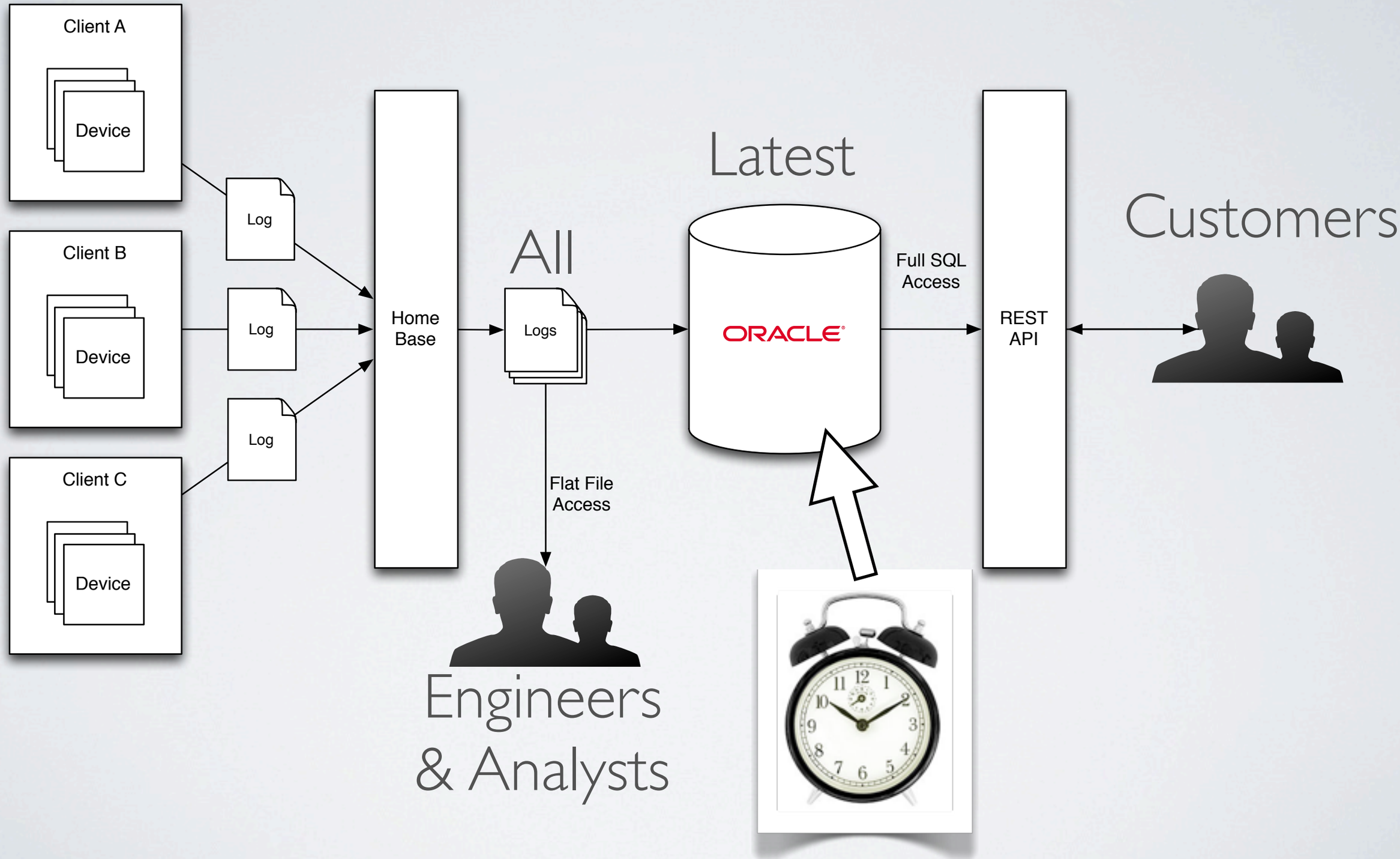
Problem

Sample Solution

Core Design Issues

Other Solutions

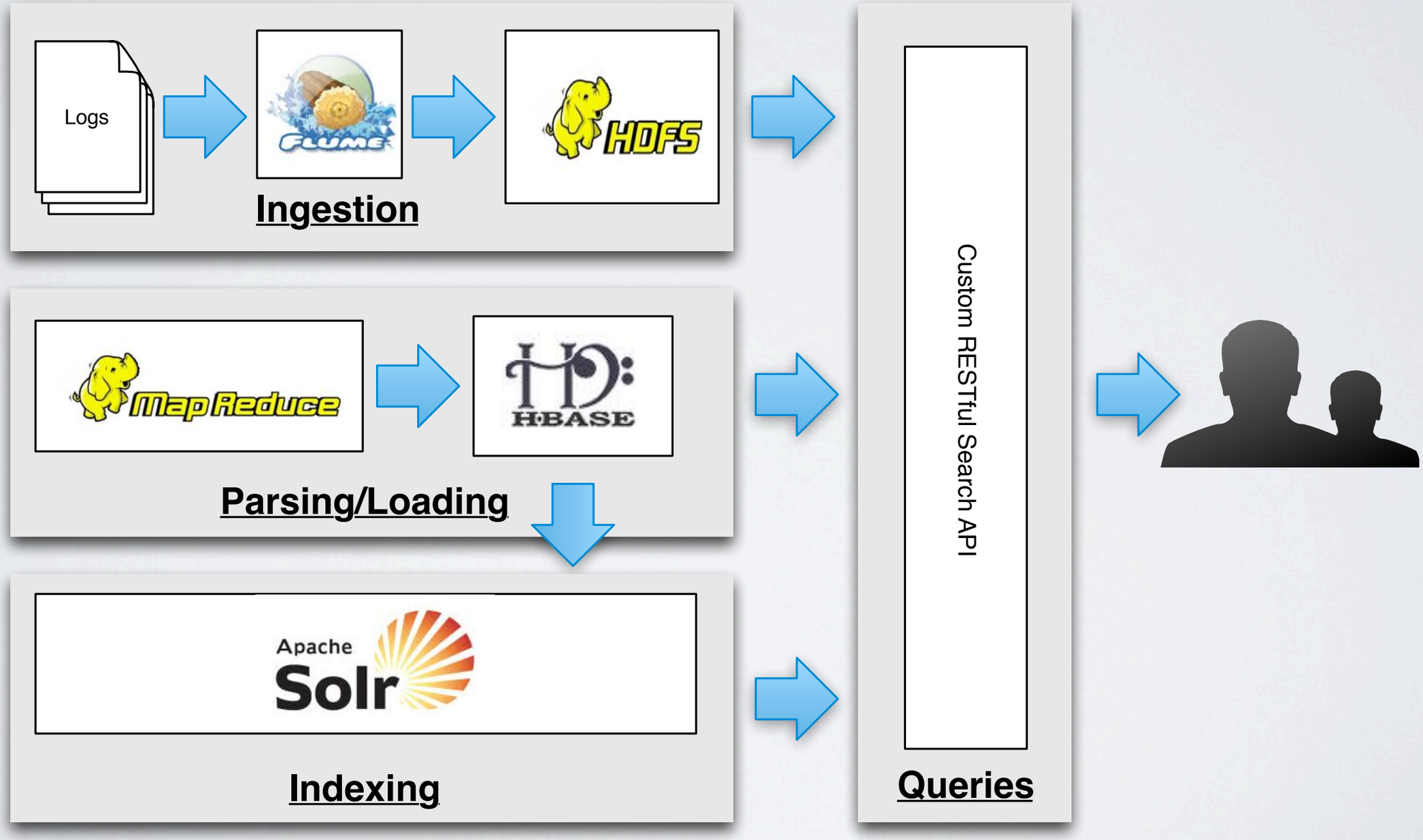
# THE PROBLEM


Client A
Device
Client B
Device
Client C
Device
Log
Log
Log
Home Base
All
Logs
Flat File Access
Engineers & Analysts
Latest
ORACLE
Full SQL Access
REST API
Customers

# SEARCH APPLICATION FEATURES

- Find last three days of raw logs from an entire cluster
- Group all results by machine serial number and show the newest first
- Search all device subject lines for "FAILURE"
- View all hard disk objects that have product number 2341AB
- Find all motherboards with an associated customer ticket

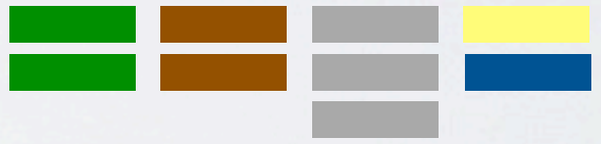
# SAMPLE SOLUTION



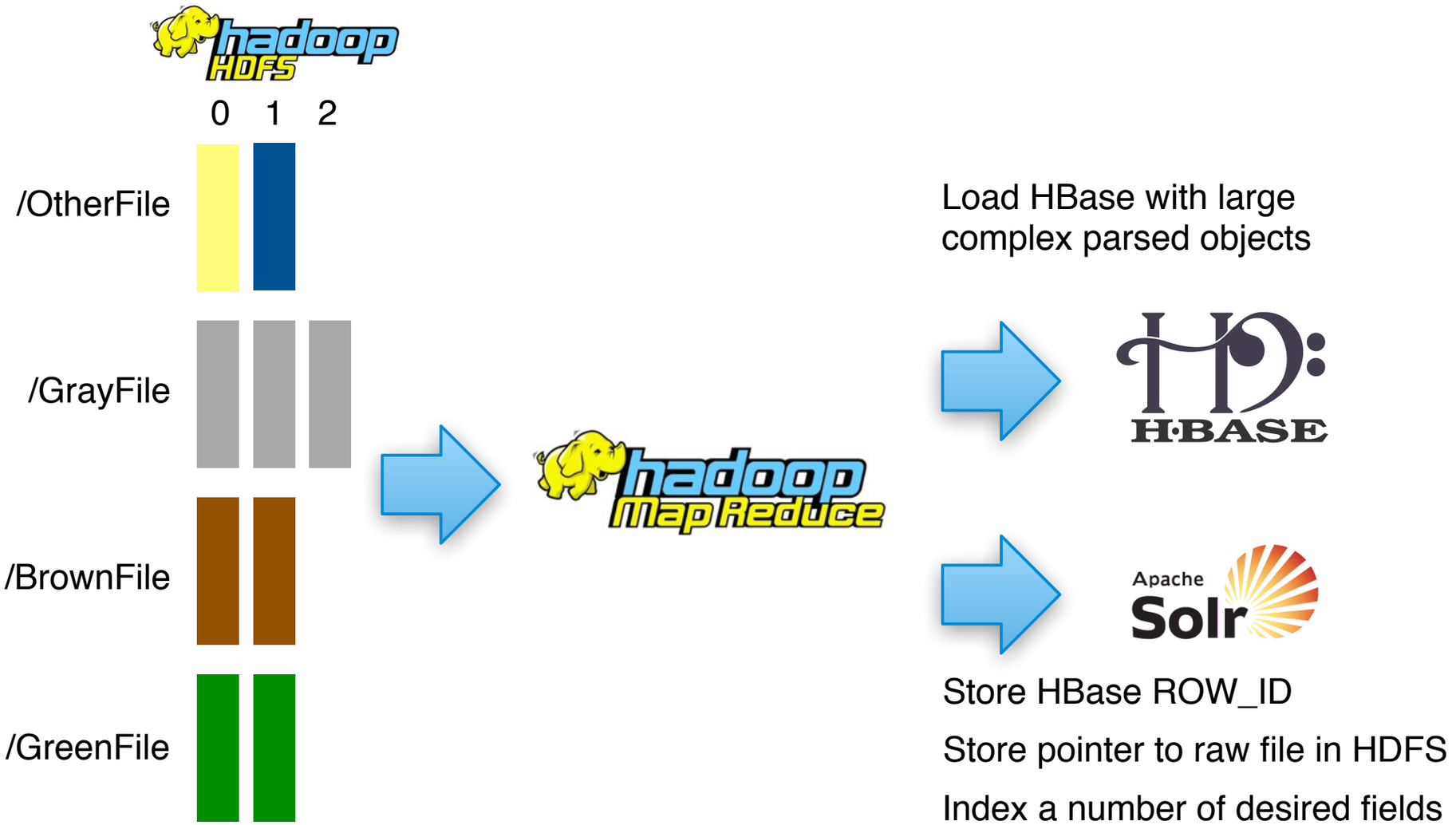
# INGESTION



Green Files
Brown Files
Gray Files
Other Files

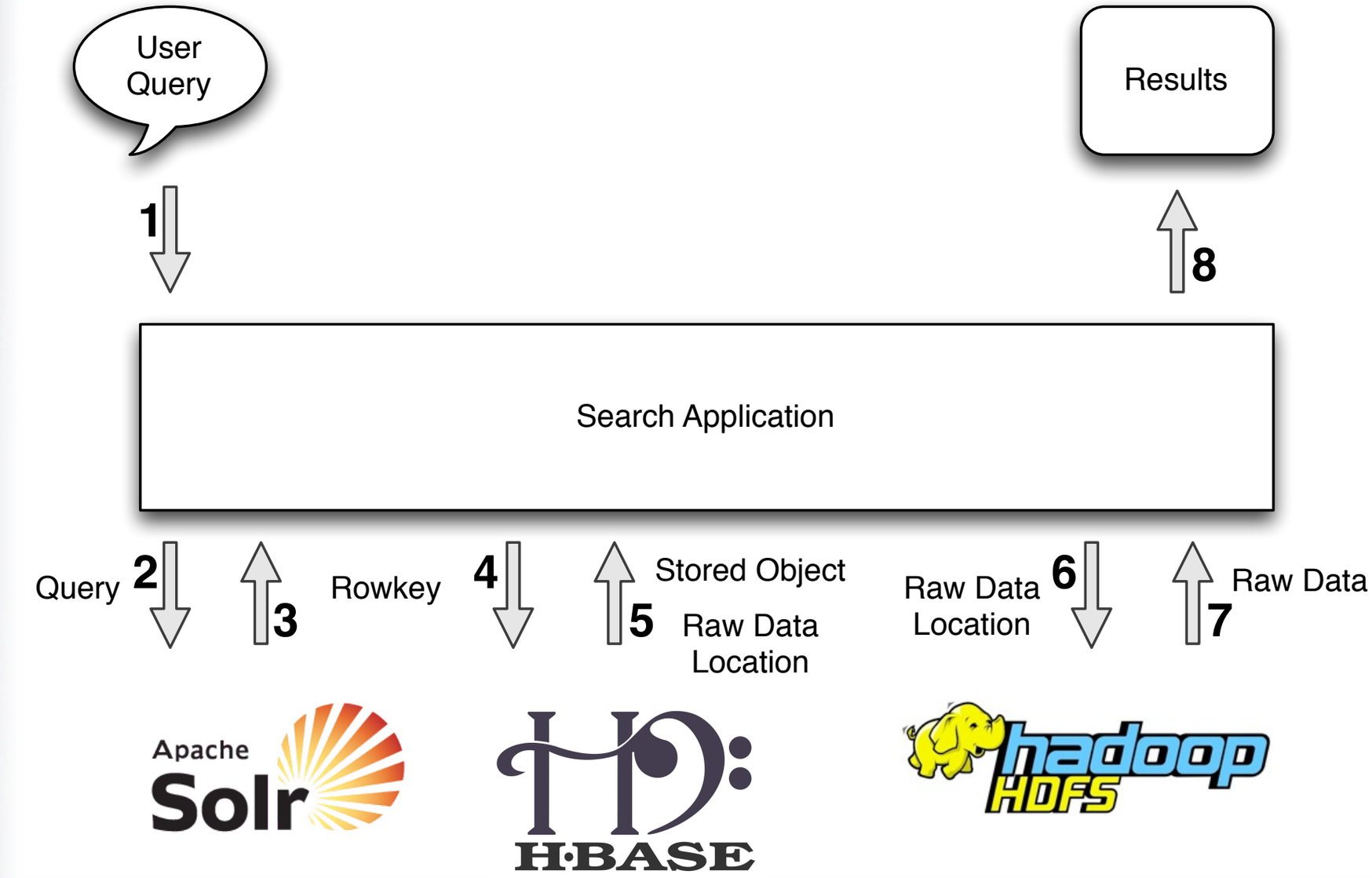


# PARSING, LOADING, AND INDEXING




Storing and Indexing are two very different things!

# SEARCH APPLICATION



# CORE DESIGN ISSUES

- Data de-normalization
- Changing the schema (manual reindex)
- Elastic shard scaling (manual reindex)
- Data replication
- Managing HBase and Solr partitioning/sharding
- Write durability

# 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...

# SOLRCLOUD


- Automatic shard creation, routing
- Replication
- Limited to a fixed number of shards defined on initial creation
- ZooKeeper for coordination
- No support for joins across distributed index
- 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
- @ryantabora
- ryan.tabora@thinkbiganalytics.com