SOLUTION TRACK

Finding the Needle in a Big Data Haystack

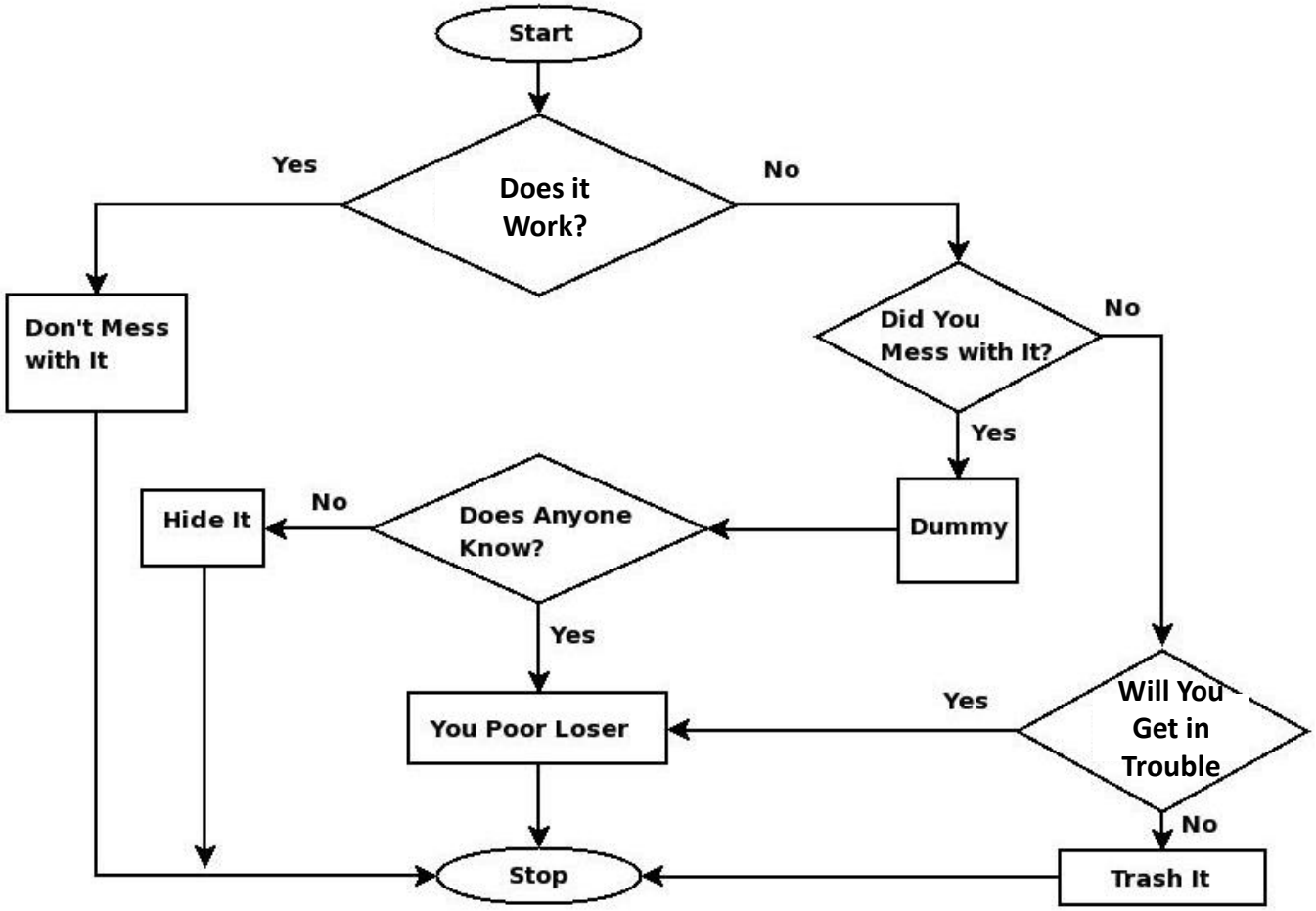
@EvaAndreasson, Innovator & Problem Solver
Cloudera

cloudera
Ask Bigger Questions

Agenda

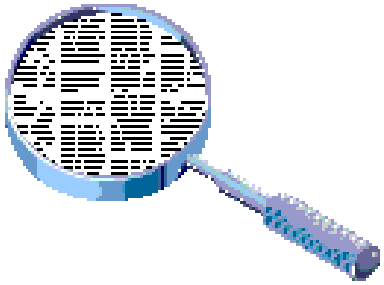
- Problem (Solving)
- Apache Solr + Apache Hadoop et al
- Real-world examples
- Q&A

Problem Solving



Information Driven Problem Solving

- Ask a Question
- Find All Relevant Data to Serve the Question
- Process the Data to Answer the Question

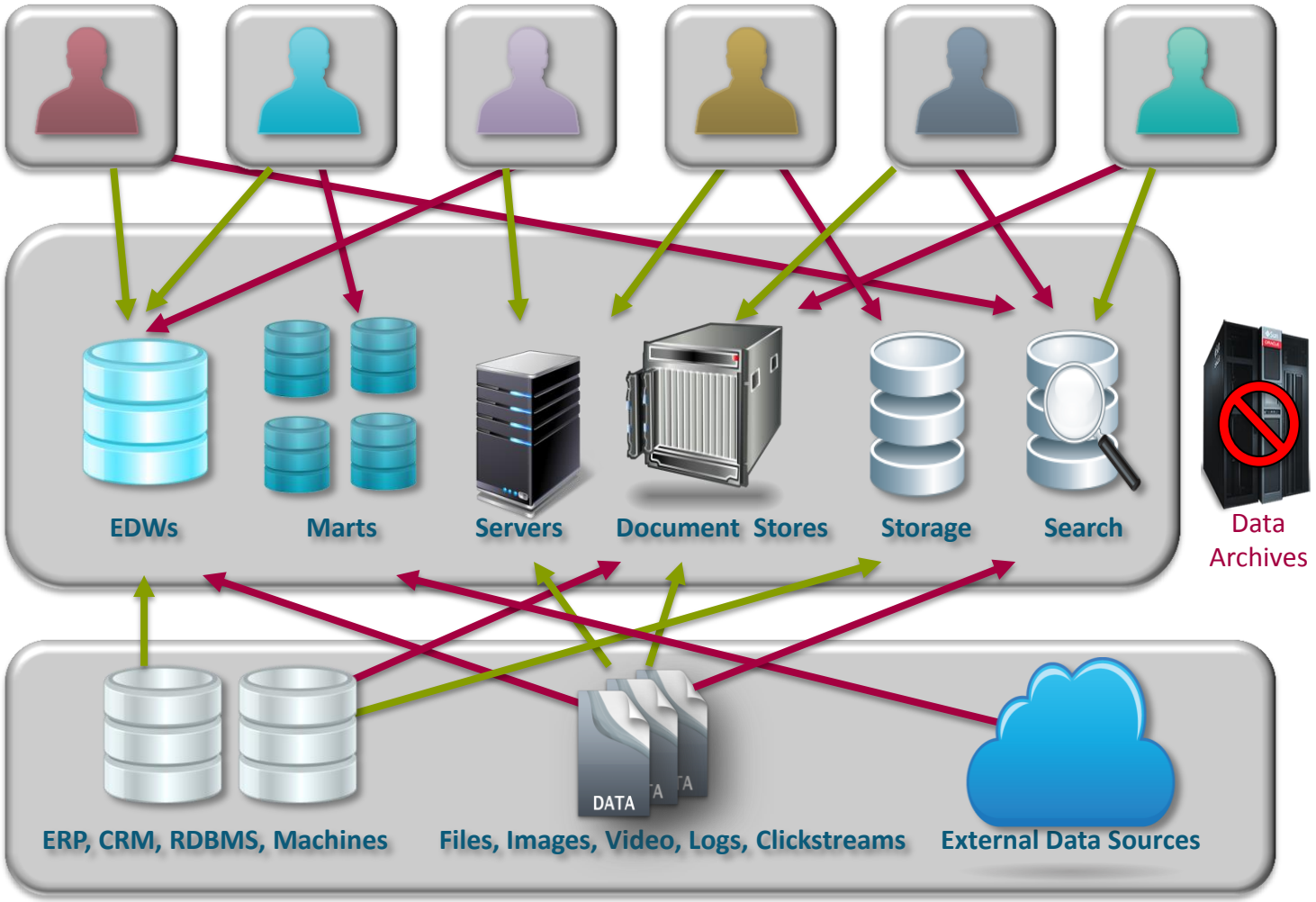

Information Driven Businesses

Problem: Finding the Data (Needle) Across (Hay) Silos

Thousands of employees & lots of data
Difficult to access

Heterogeneous legacy IT Infrastructure
Difficult to manage
Hard to scale

Silos of multi-structured data
Difficult to Integrate
Holds copies of data



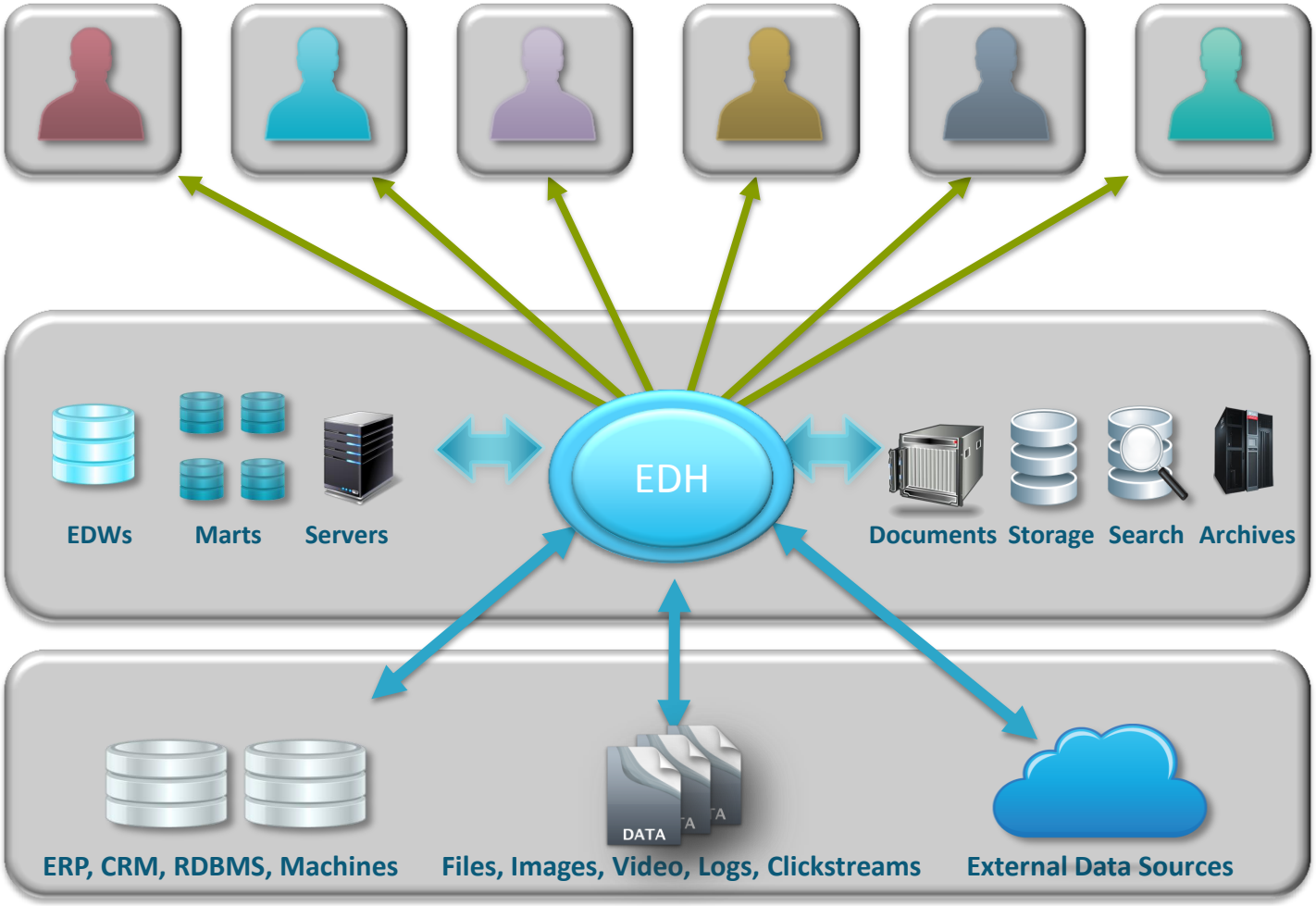
cloudera
Ask Bigger Questions

Solution: The Enterprise Data Hub (EDH)

Independent of audience, topic, or tool, data is accessible

Unified data storage, processing, management and security

Ingest all data any type or any scale Eliminate copies, simplify aggregation and correlation



cloudera
Ask Bigger Questions

Apache Hadoop et al at the Core

- Open source
  - white box, best innovation at all times
- Flexible
- Scalable
  - ingest, storage, processing
- Cost efficient

But an EDH also Needs…

- Security & audit
- Manageability, Visibility and Resource Control
- Open architecture
- Multi-workload support and optimization

CLOUDERA EDH

Problem solved!
We can go home…?

Problem: Finding the Needle in a Big Data Haystack?

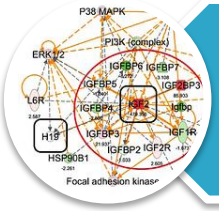
New Audiences, New Challenges

- Non-technical staff needs access to data
  - Same data used in bigger processes
  - Speed up manual introspection
- Technical staff needs to view / explore data
  - View interim results
  - Drill down into mission critical data
  - Explore data to design models
- Cross-workload needs
  - Combine structured and unstructured data

Solution: Everyone Knows Search!

Explore

Navigate



Correlate

Cloudera Search: How We Integrated Solr with Hadoop et al

Cloudera Search

Interactive search for Hadoop

- Full-text and faceted navigation
- Batch, near real-time, and on-demand indexing

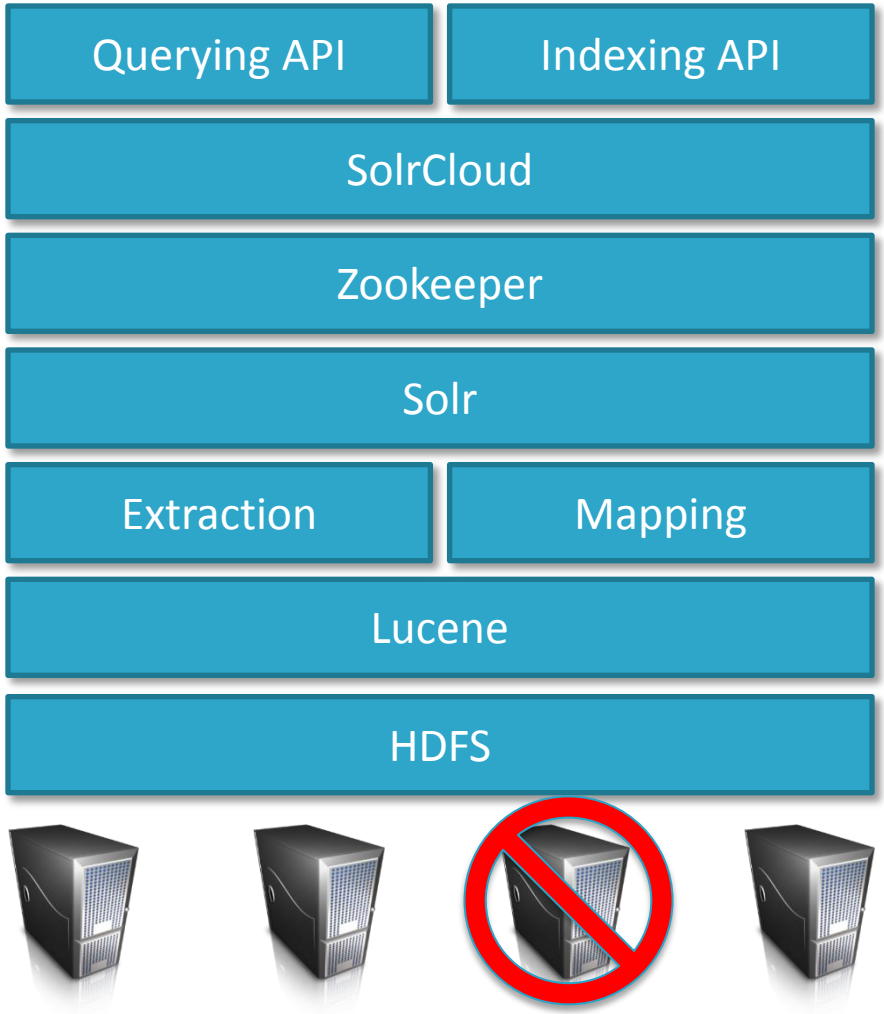
Apache Solr integrated with CDH

- Established, mature search with vibrant community
- Separate runtime like MapReduce, Impala
- Incorporated as part of the Hadoop ecosystem

Open Source

- 100% Apache, 100% Solr
- Standard Solr APIs

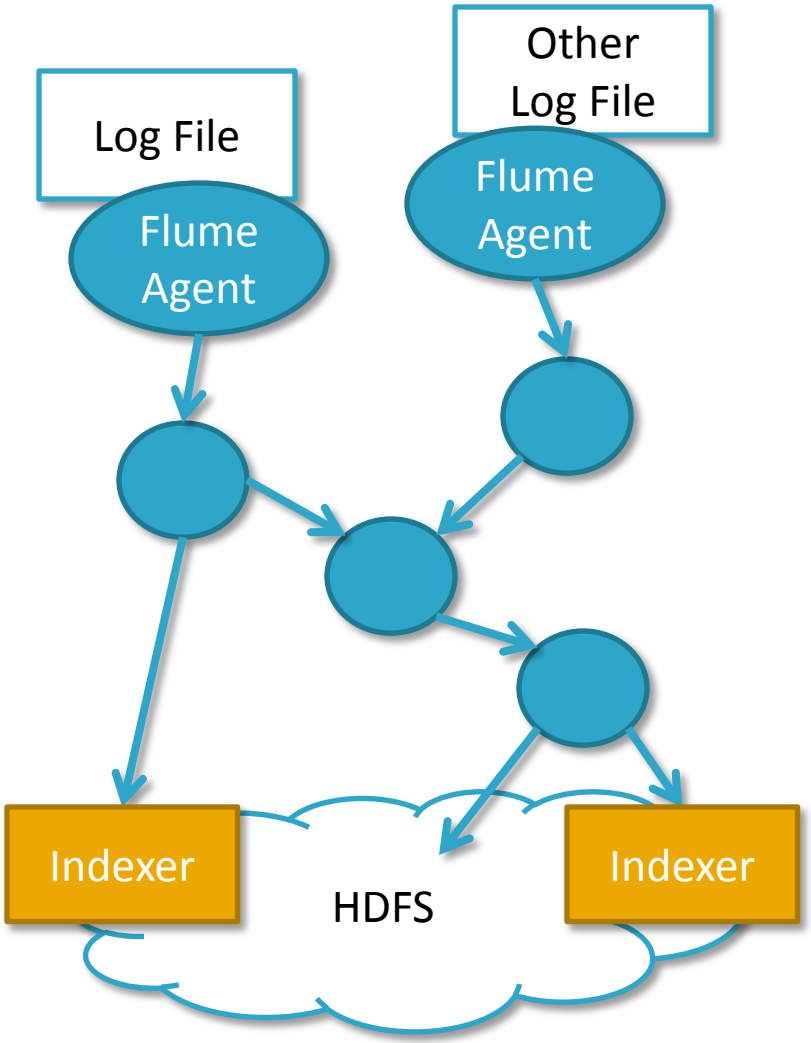
Scalable and Robust Index Storage



Solr and HDFS

- Scalable, cost-efficient index storage
- Higher availability
- Search *and* process data in *one* platform

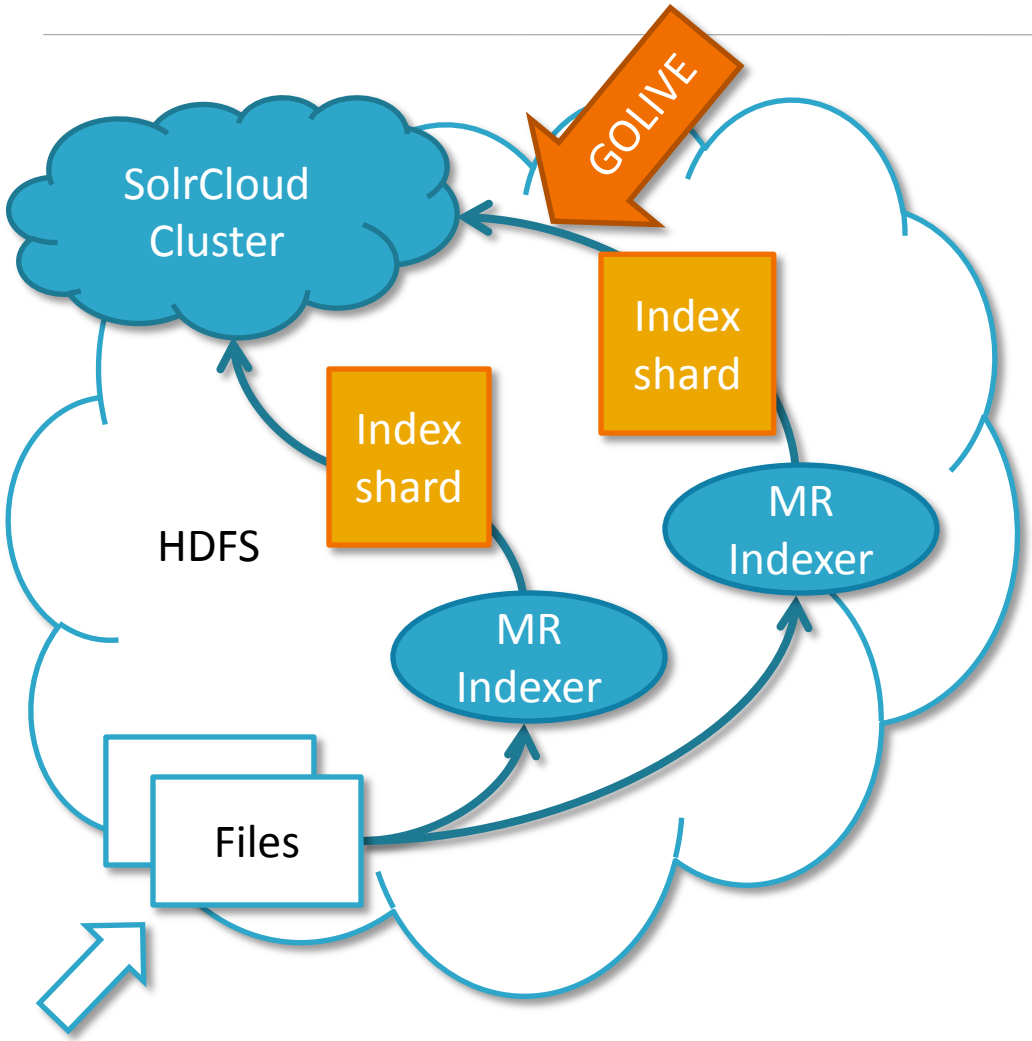
Near Real Time Indexing at Ingest



Solr and Flume

- Data ingest at scale
- Flexible extraction and mapping
- Indexing at data ingest
- Document-level ACL

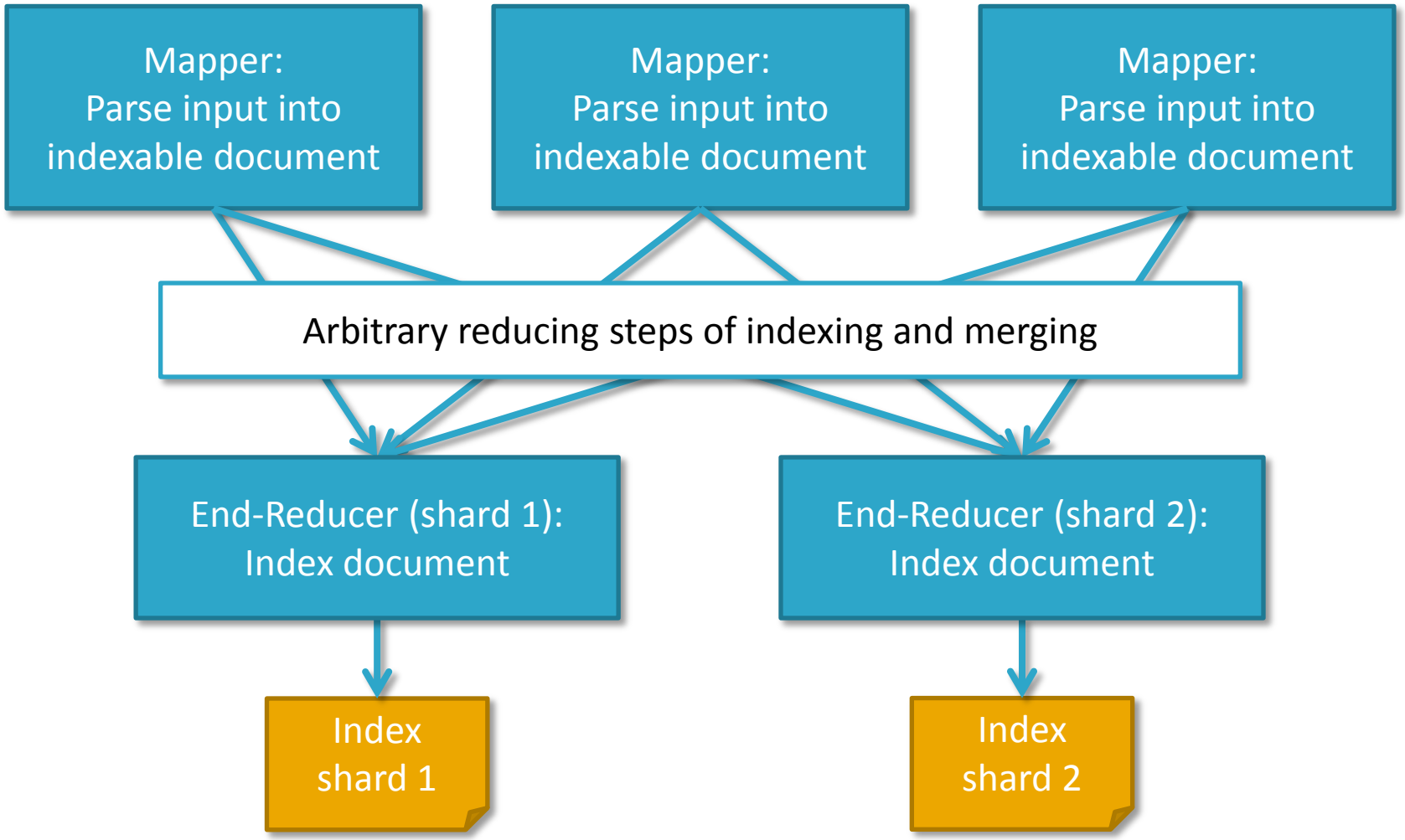
Scalable Batch Indexing



Solr and MapReduce

- Flexible, scalable batch indexing
- GOLIVE: Start serving new indices with no downtime
- On-demand indexing, cost-efficient re-indexing

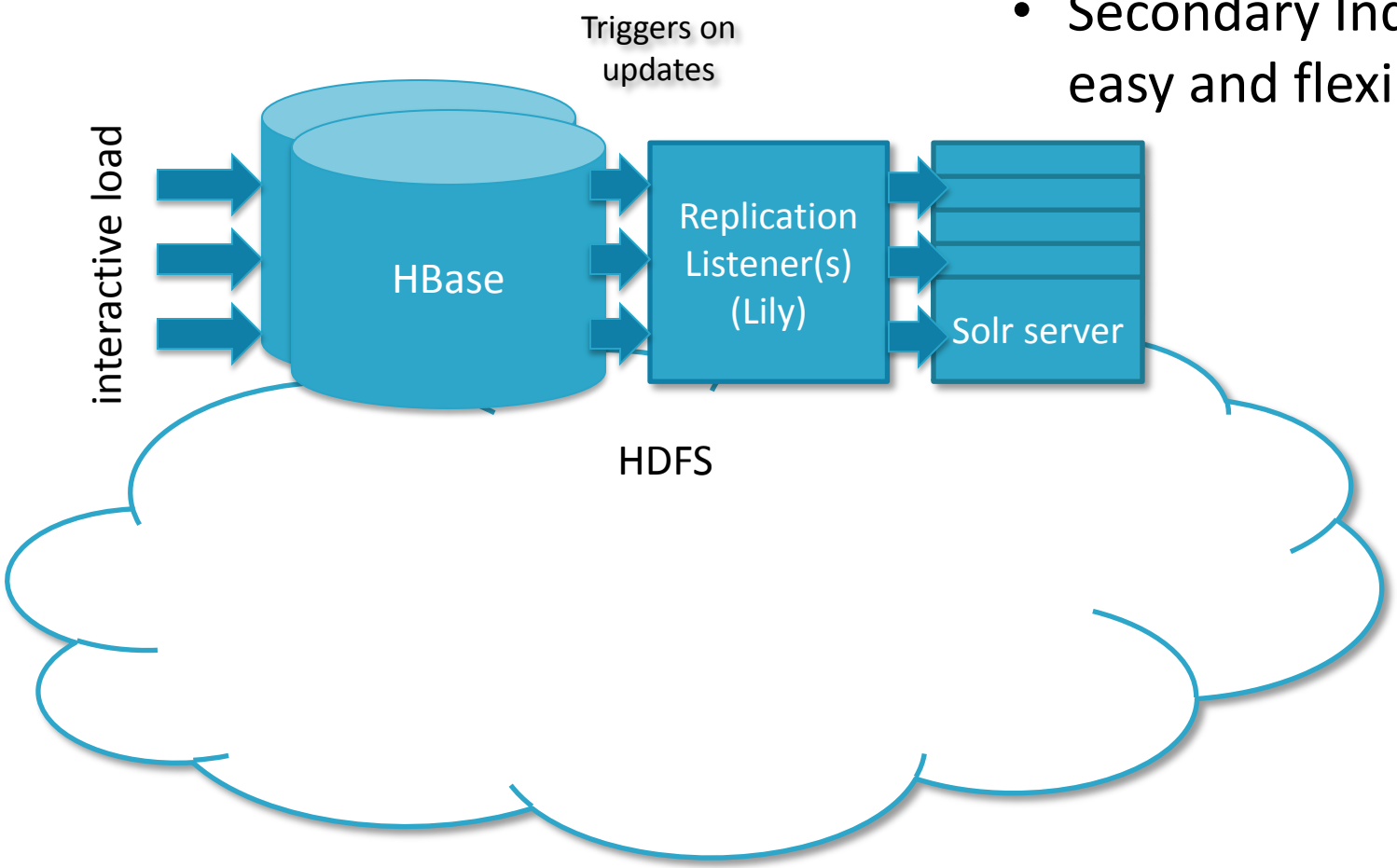
Scalable Batch Indexing



Secondary Indexes, in Real-Time

Solr and HBase

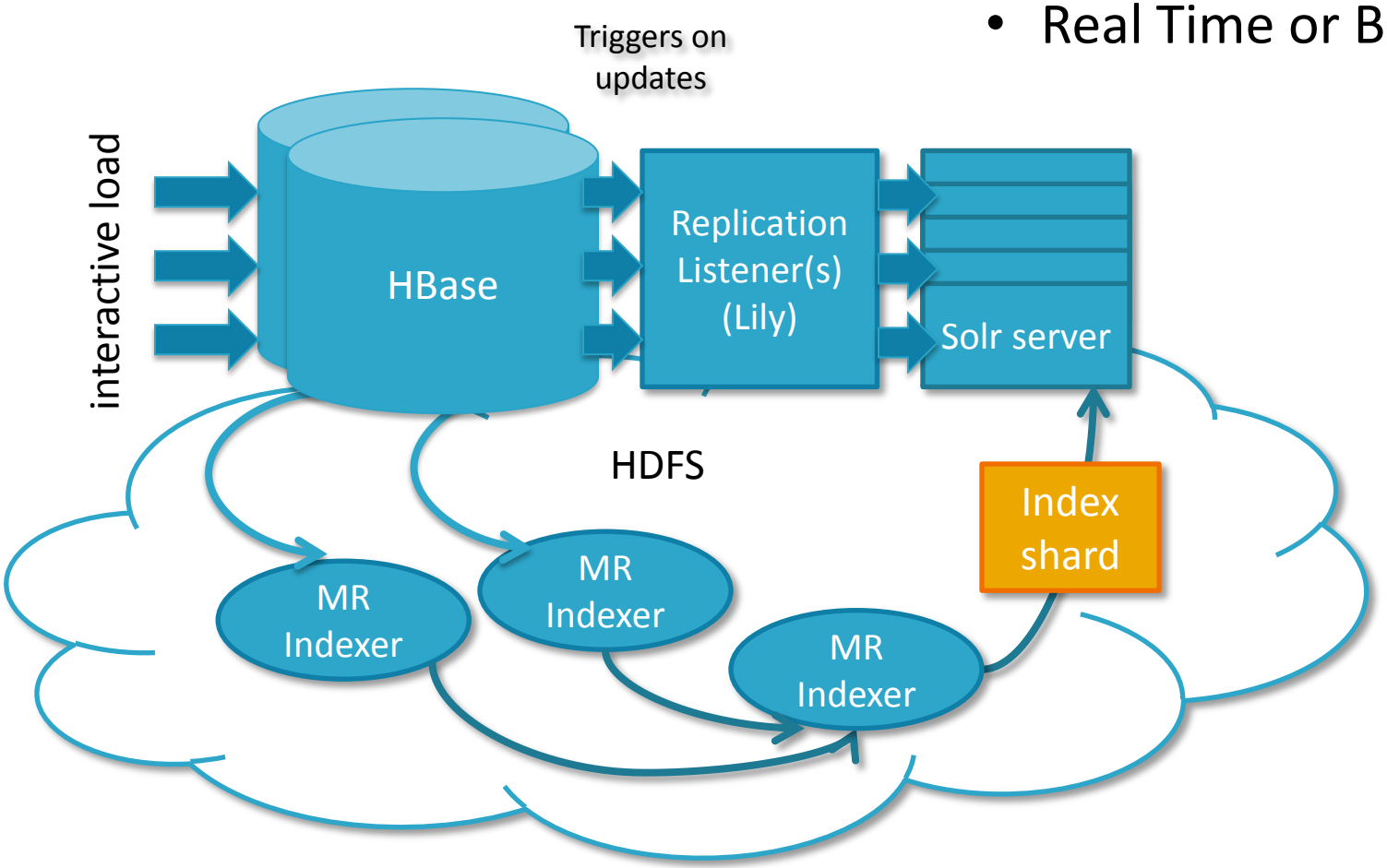
- Secondary Indexes made easy and flexible



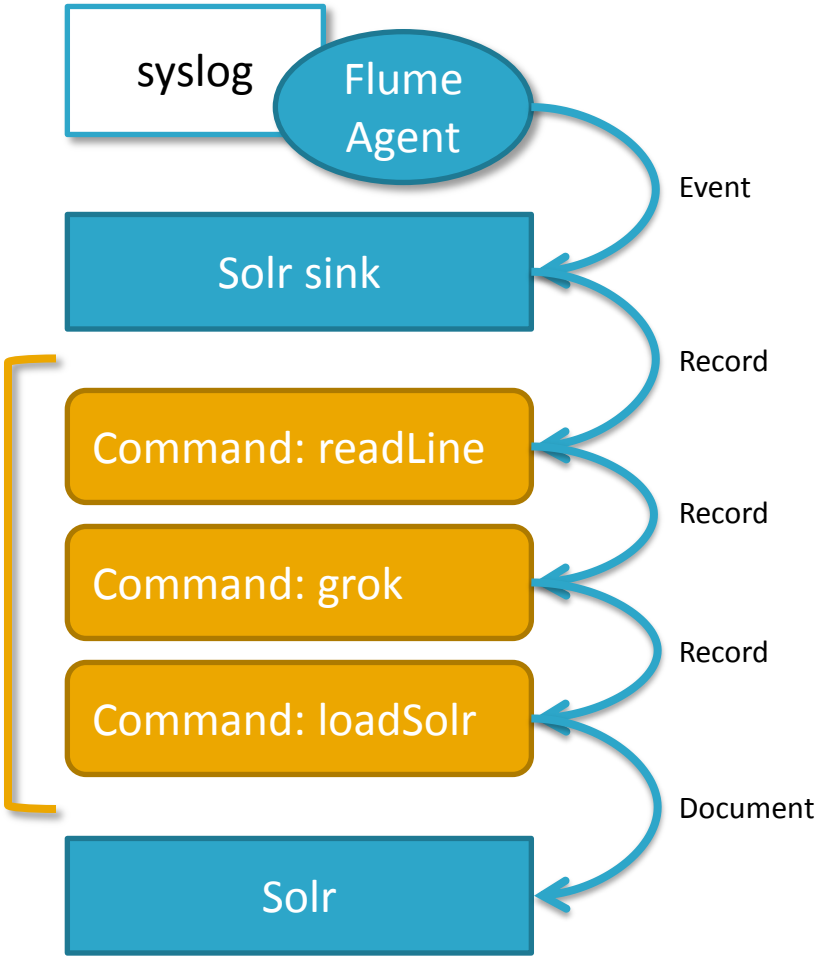
Secondary Indexes, or in Batch

Solr and HBase

- Real Time or Batch



Streamlined Extraction and Mapping



Morphlines

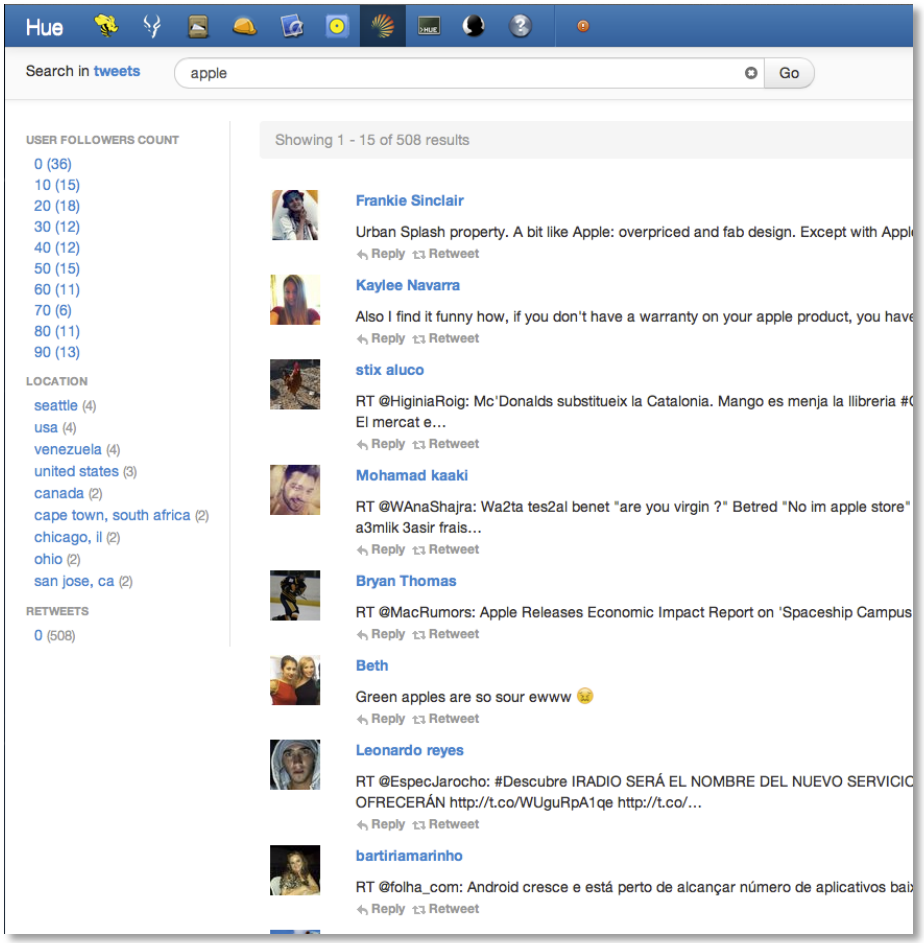
- Simple and flexible data transformation
- Reusable across multiple index (and other) workloads

Security

Cloudera Search + Sentry

- Cluster level access control
- Index level access control

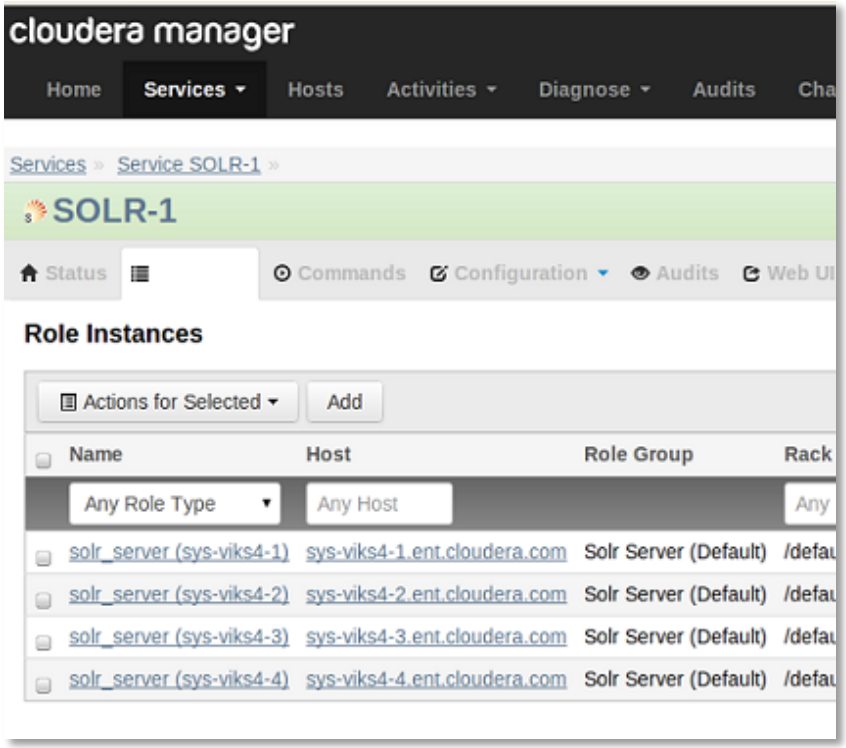
Simple, Customizable Search UI



Hue

- Simple UI
- Navigated, faceted drill down
- Customizable display
- Full text search, standard Solr API and query language

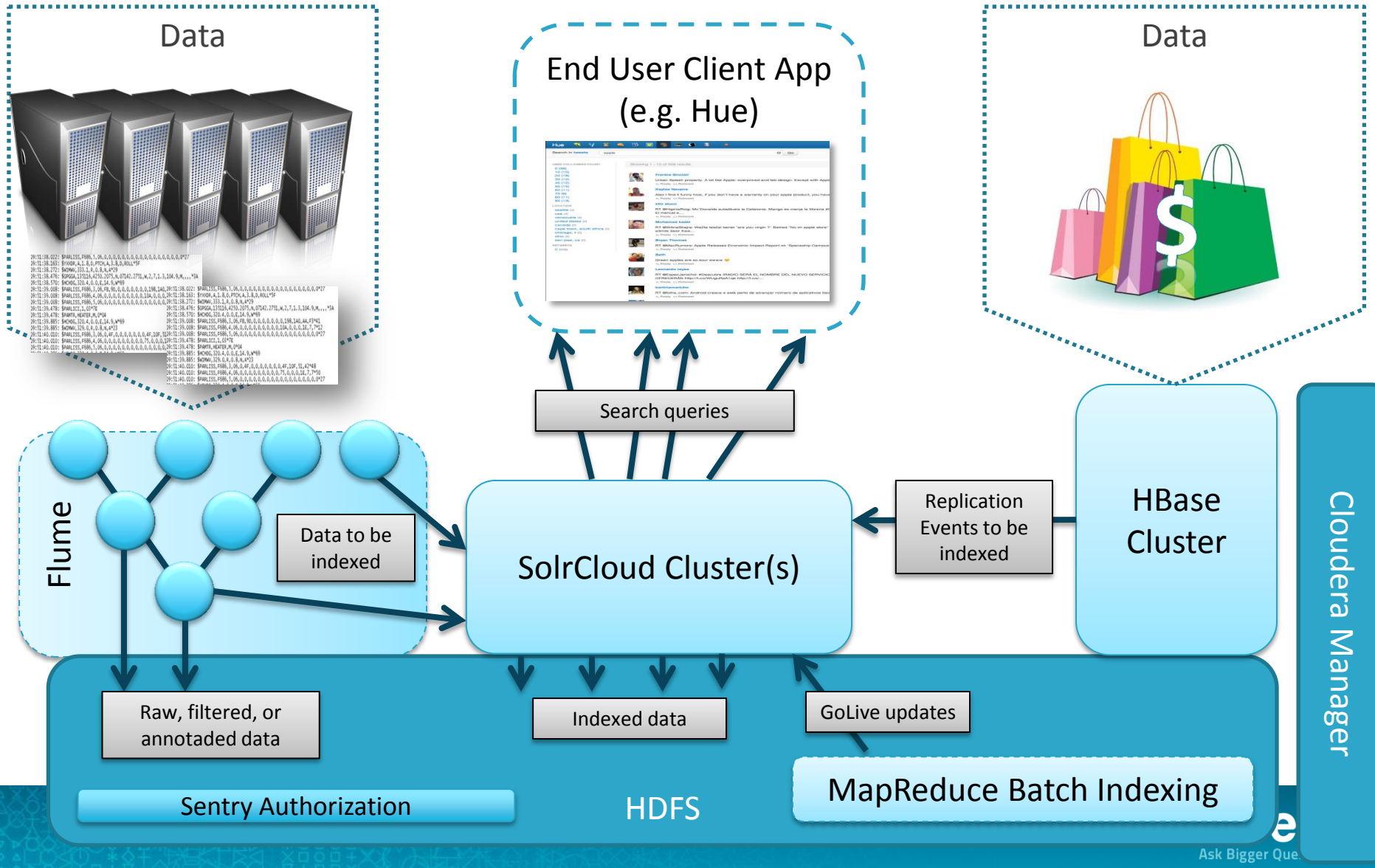
Simplified Management



Cloudera Manager

- Install, configure, deploy SolrCloud on the cluster
- Centralized management and monitoring – cross workloads
- Unified resource management and control

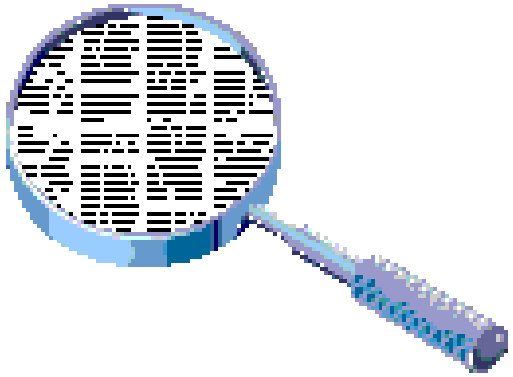
Architecture Overview



Some Real-World Examples

- Image processing and data correlation
  - Quick result checks on new algorithms
  - Correlate image and device logs or external data
  - Image exploration as a service
  - Expedite data modeling processes
- Free-text form matching
  - Claims report correlation, to gather data likely to be similar
  - Serve 360-client or patient records in a speedier way
  - Fraud / pattern extraction
- Log management
  - Real-time drill down
  - Long term trending and capacity planning
  - Anomaly detection over larger sets of data

The EDH - Information-Driven Problem Solving Made Easy!



+

Integration is Key

- Eliminate data moves or copies, break silos
- Create a truly active archive
- Serve non-technical audiences on the same platform as where advanced analytics workloads run
- Analyze and combine structured and non-structured data
- Expedite exploration of various data types
- Find data and do something with it – where it is stored
- Future proof your data management system

Learn More

- Cloudera.com
  - Read our blog
  - Take our online training (or get Cloudera certified)
  - Download whitepapers
  - View webinars
  - Talk to our customers
- Follow/contact me
  - @EvaAndreasson

QCon

Please evaluate my talk via the mobile app!


cloudera®
Ask Bigger Questions