# What Can Hadoop Do for You?

@EvaAndreasson | Cloudera

2014

# Agenda

- Why Hadoop?

- What is Hadoop?

- Common Use Cases & Real World Examples

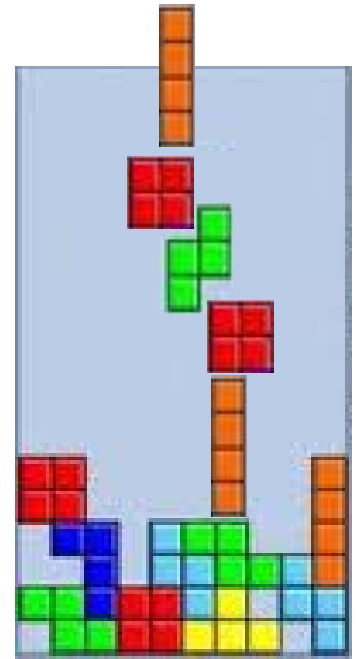- Q&A

# Our Industry is Data Challenged!

# Volume of Data is Growing Exponentially

- Traditional data storage do not scale

  - Pre-mature archiving, limiting timely access

  - Storing minimal data, limiting insight



cloudera
Ask Bigger Questions

# Increased Demands on Processing Efficiency

- Volume increases, SLAs stay the same
- Pace of data generation increasing

# New Data Types Demands Flexibility

- New data types generating valuable information
  - Sensor, Mobile, IoT, Geo, Images, Videos, Voice…
- Traditional systems suffer from inflexible data models
  - Data modeled for questions we know
  - Remodeling = $$$$

# Organizations Need to Rethink Data!

# Practical Driver

- Organizations seek to scale data storage and processing cost-efficiently
  - Do what I usually do, but on larger data sets
  - Do complex queries within SLAs

# Competitive Driver

- Enterprises seek to ask bigger questions
  - What *new insights* can we achieve by combining data silos?
  - What else can we find by asking questions over new types of data?
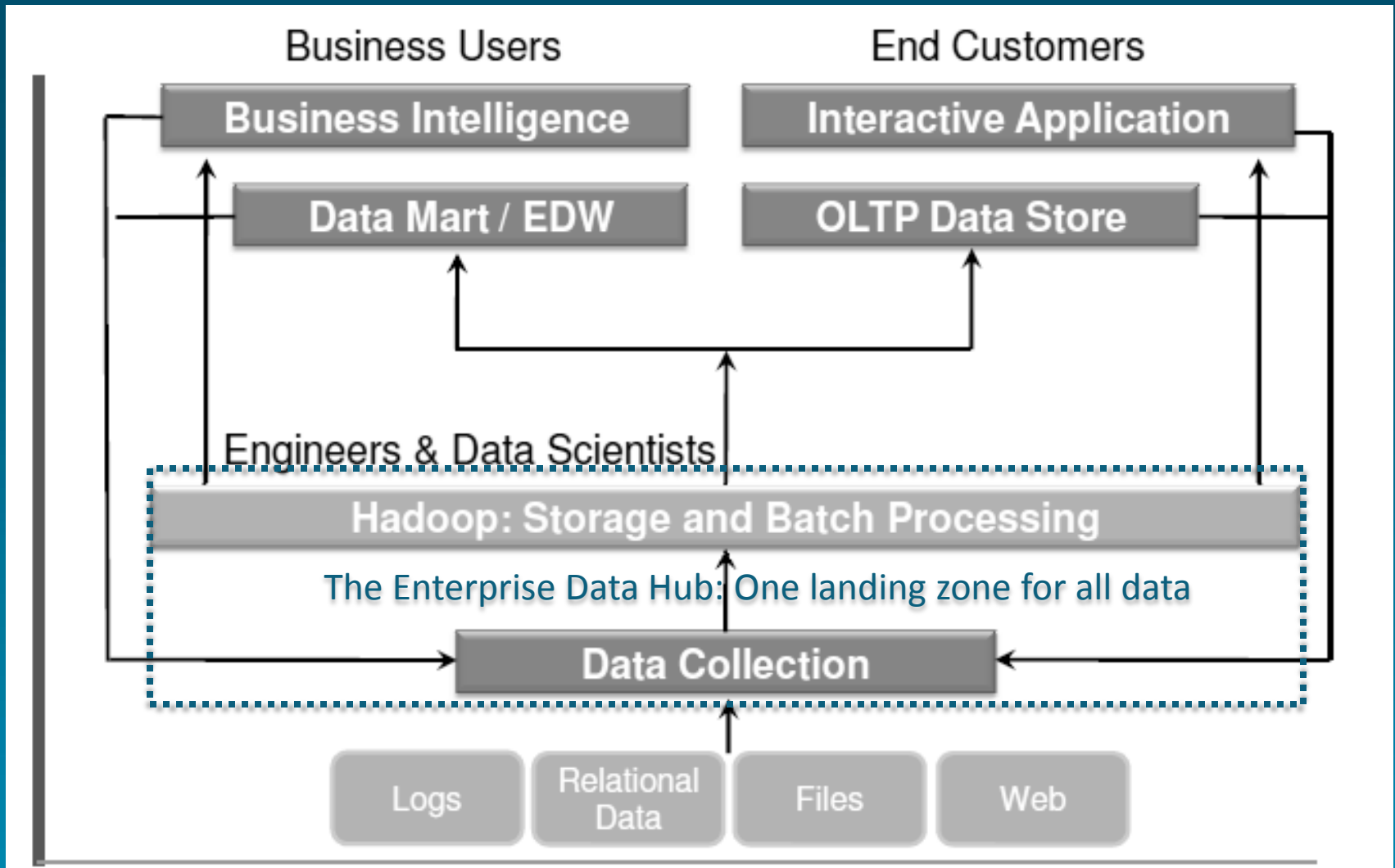
## There is no box!

Hadoop Saves the Day!   ☺
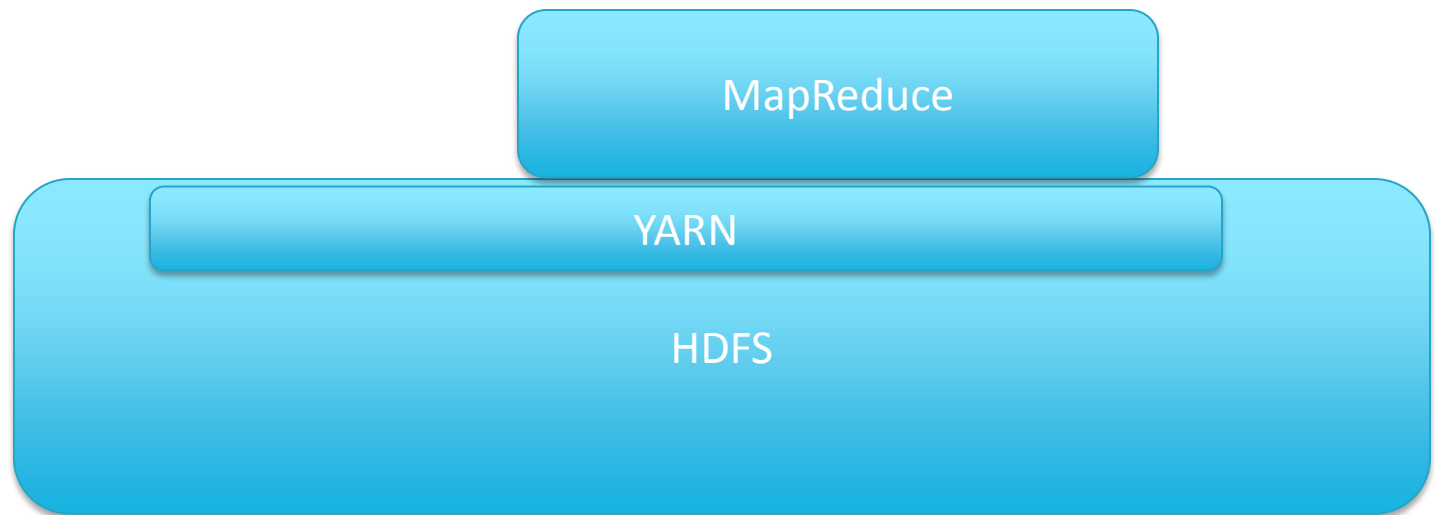
# Hadoop as the Solution

- Distributed, linearly scalable parallel processing AND storage
  - Data stored in equally sized chunks
  - Distributed and replicated across common hardware
  - Parallel processing
- Structure at query time, not at storage time
  - Ask anything at any point

cloudera®
Ask Bigger Questions

# Organizations Use Hadoop as an Enterprise Data Hub



Business Users

End Customers

Business Intelligence

Interactive Application

Data Mart / EDW

OLTP Data Store

Engineers & Data Scientists

Hadoop: Storage and Batch Processing

The Enterprise Data Hub: One landing zone for all data

Data Collection

Logs

Relational Data

Files

Web

cloudera®
Ask Bigger Questions

# Apache Hadoop and its Ecosystem 101

# The Apache Hadoop Ecosystem – a Zoo!
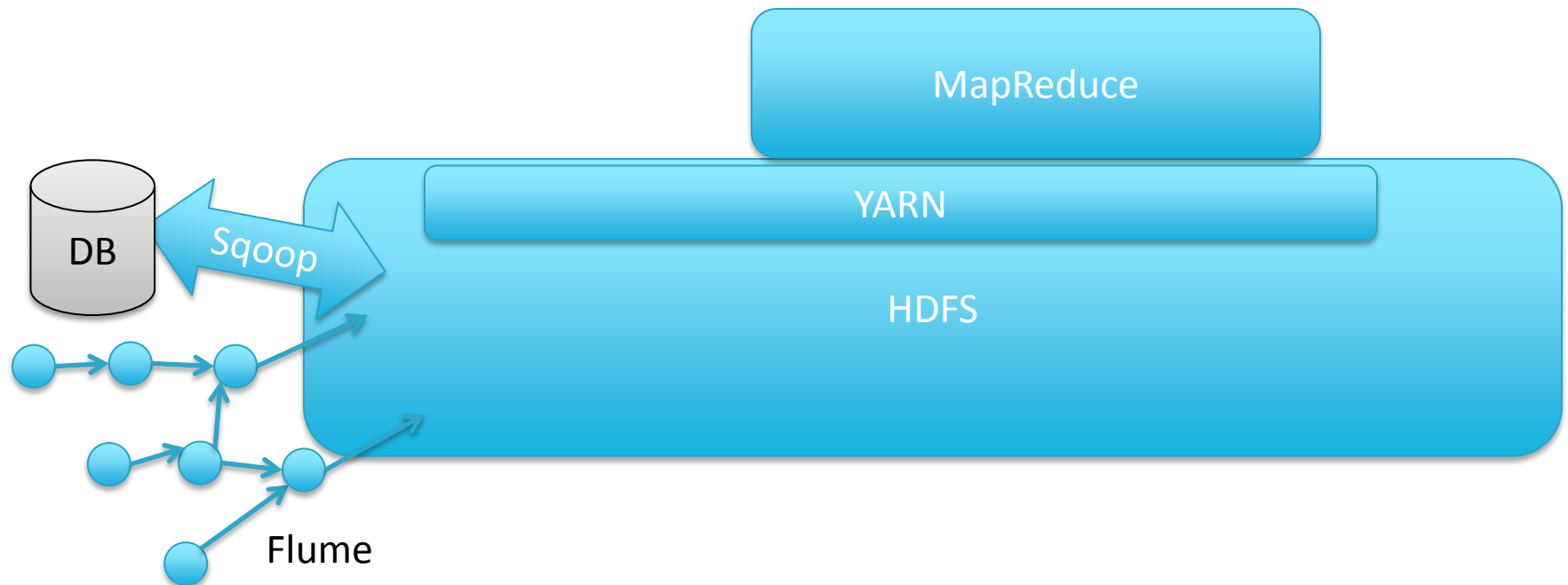
MapReduce

YARN

HDFS

# The Apache Hadoop Ecosystem – a Zoo!

# The Apache Hadoop Ecosystem – a Zoo!

| Hive | Pig | Mahout, Oryx |
| --- | --- | --- |

**MapReduce**

**YARN**

**HDFS**

DB

*Sqoop*

Flume

# The Apache Hadoop Ecosystem – a Zoo!

Impala

Solr

Spark

Hive
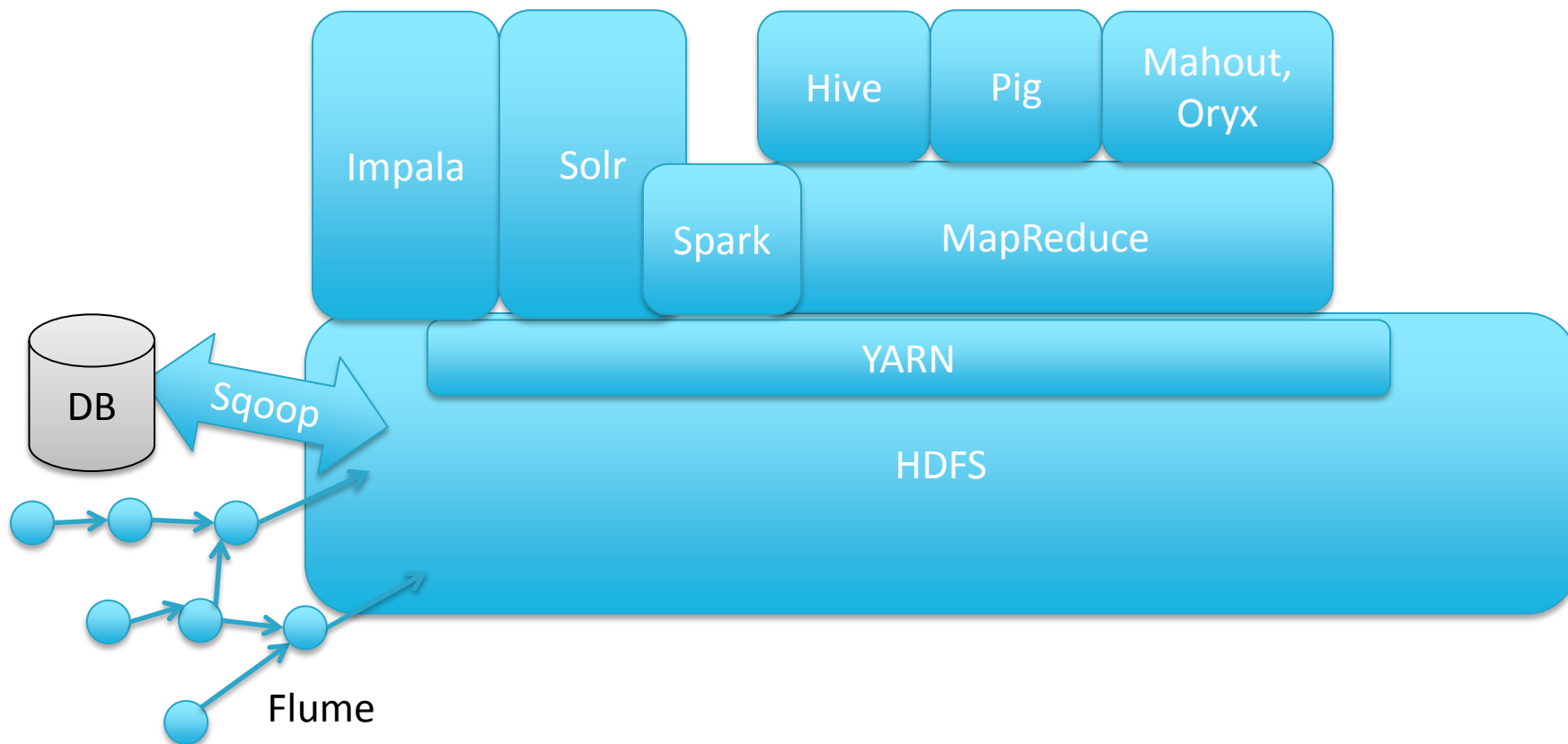
Pig

Mahout, Oryx

MapReduce

YARN

DB

Sqoop
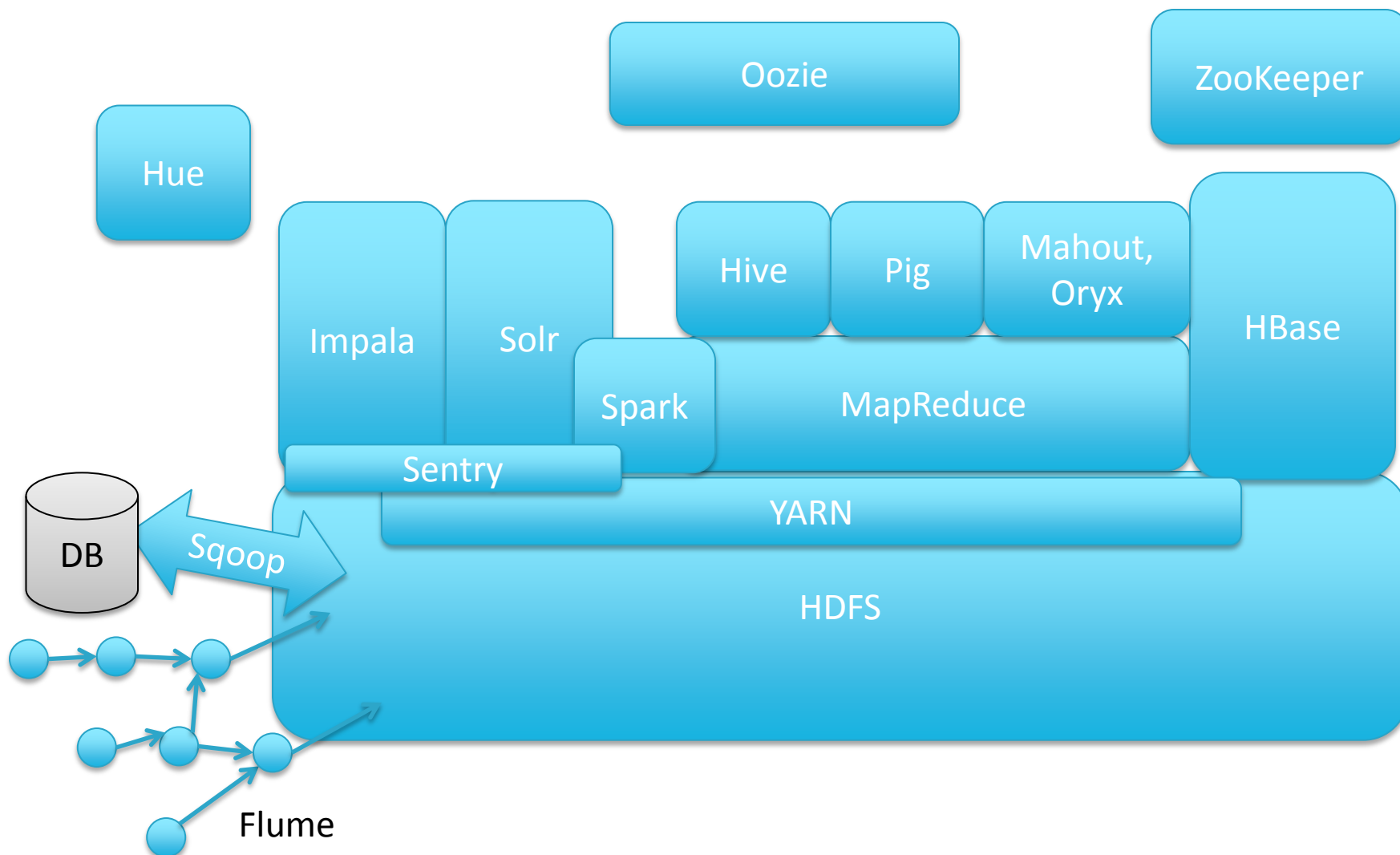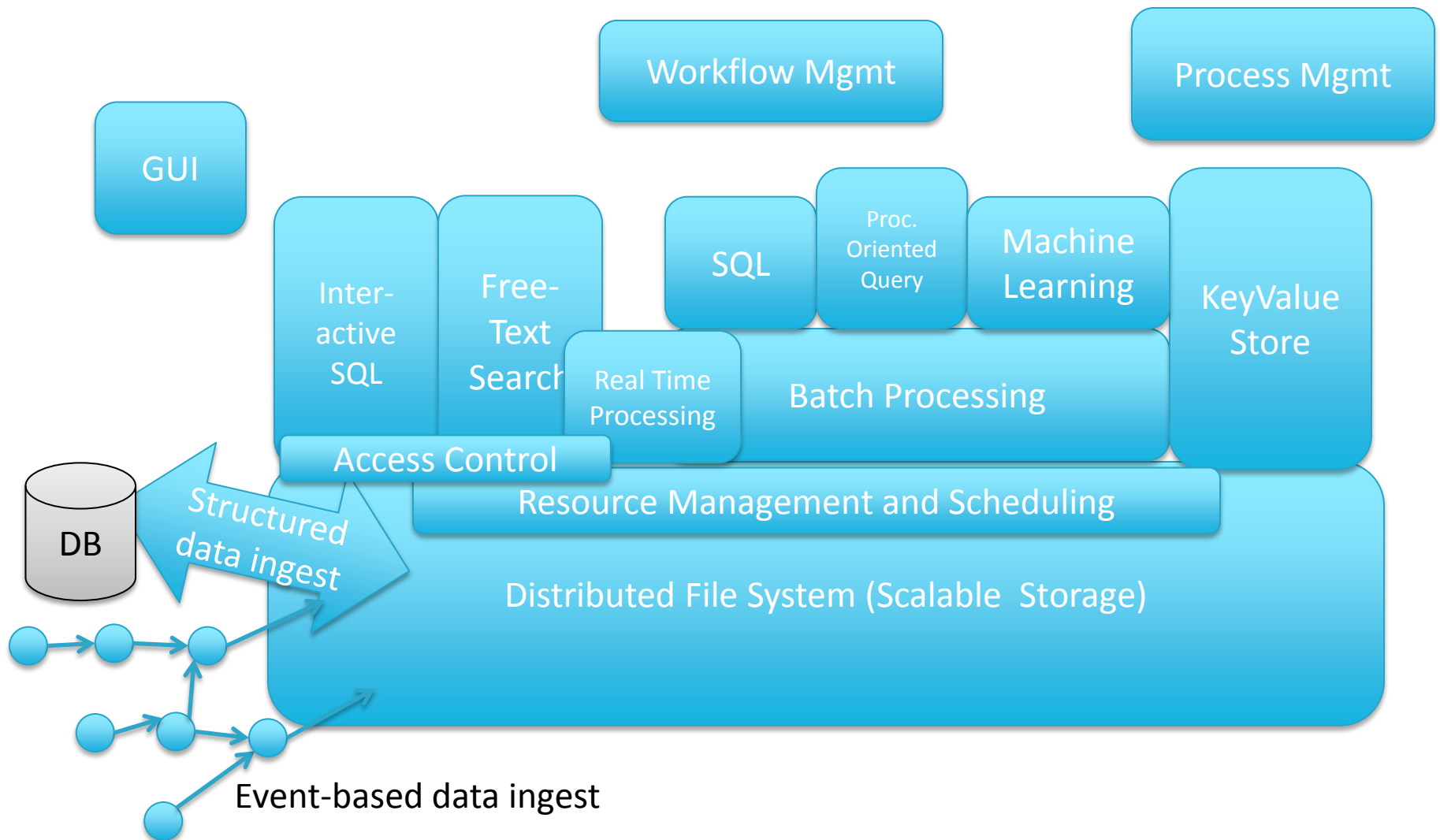
HDFS

Flume
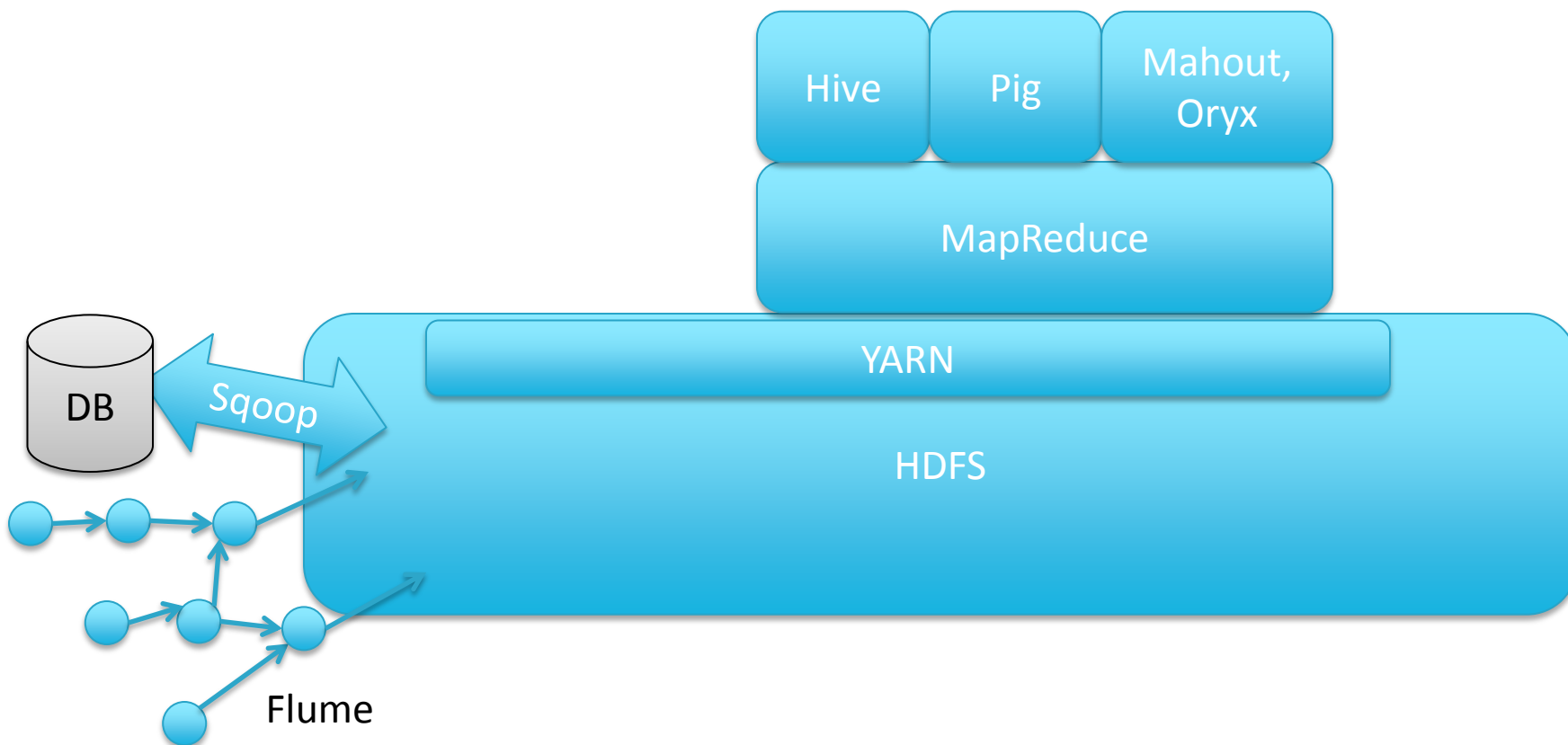
# The Apache Hadoop Ecosystem – a Zoo!
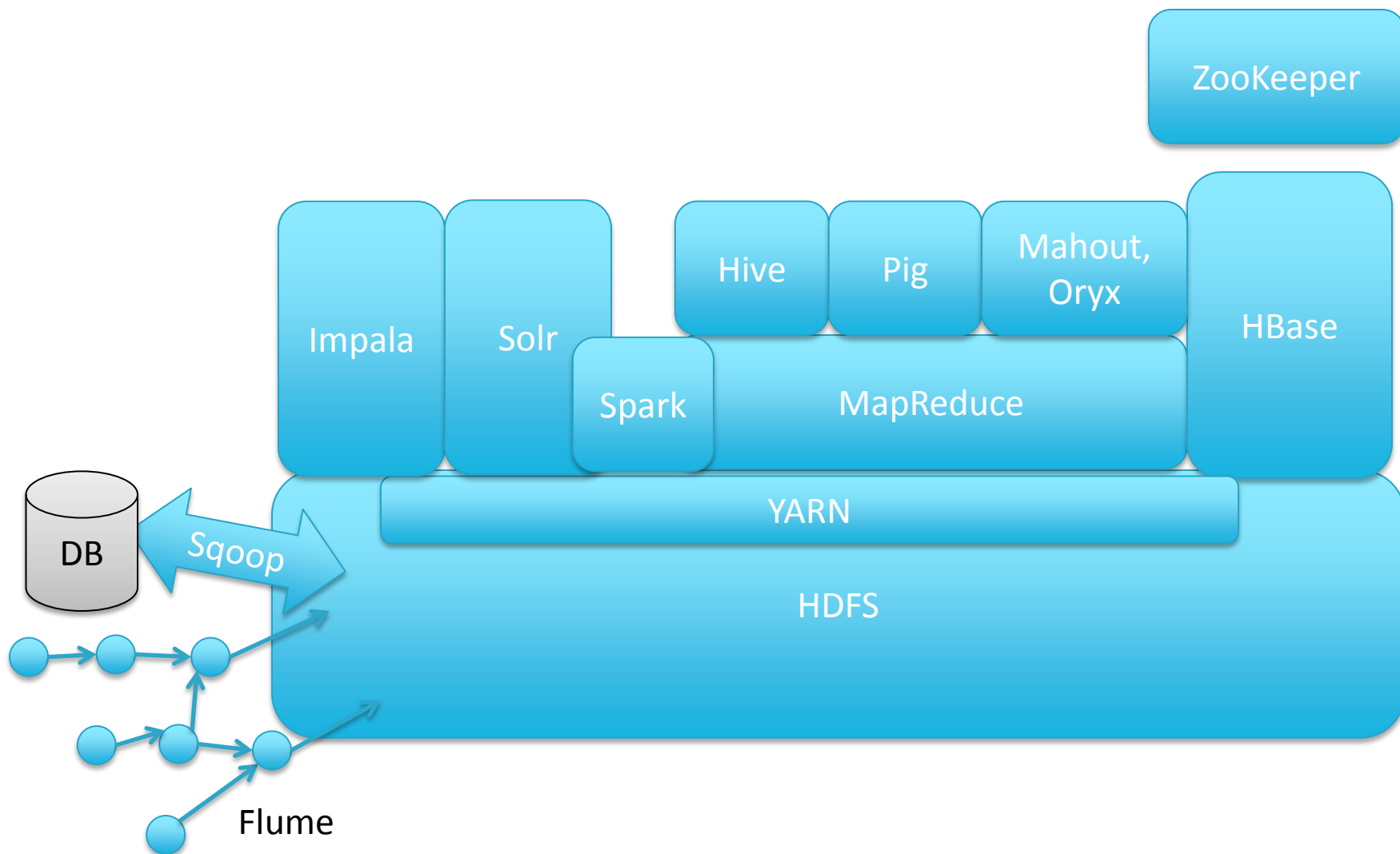
# The Apache Hadoop Ecosystem – a Zoo!

# The Hadoop Ecosystem – Explained!

# Use Cases

# Top 3

- Offload ETL
  - Do more data processing within the same time window
- Enable an active archive
  - All data available at any time, interactively or in batch, to multiple audiences
- Ask bigger questions
  - Combine data silos or new data types for bigger insight (360 view use cases)
  - "Advanced analytics"

# Lots of Interesting Big Questions

- Advanced analytics
  - Event prediction
  - Anomaly detection
  - Customer profiling
  - Recommendation engines
  - Churn analysis
  - Advanced image processing
  - Genome sequencing
  - Clickstream analysis
  - Path & logistic optimization

How do we optimize our replacement logistics?

How do we do damage control cost efficiently?

Which product flavor should we invest in?

Why do we experience a decrease in customers?

How do we increase our online sales?

**cloudera**®
Ask Bigger Questions

# My Personal Favorites

- How do we quickly find paths to provide help to disaster areas?

- How do we prevent suicide?

- How do we reduce world hunger?

- How do we speed up the process of finding cures for cancer and other illnesses?

- How do we mitigate global warming?

- How do we improve education?

# Summary

# Why Hadoop?

- **Scalable**
  - Add hardware, and that's (almost) it
- **Cost efficient**
- **Flexible**
  - Structure at query time
- **It's fun!**
  - Cutting edge has an intrinsic value…

# Plan Ahead

- **Knowledge?**
  - Get certified
  - Find a 3$^{rd}$ party, if recruiting is a problem
- **Real Need?**
  - Is your data too big or costly for current systems?
  - Does your business need insight across multiple data sets or new types of data?
- **Readiness?**
  - Organizational support and strategy?
  - Tools requirements and integration needs with existing environment?

# To Learn More…

1. Read some good stuff
   - Buy any of the Hadoop books from O'Reilly
   - Read great info nuggets at blog.cloudera.com/
2. Play on your own
   - demo.gethue.com
   - Cloudera QuickStart VM
     - https://ccp.cloudera.com/display/SUPPORT/Cloudera+Manager+Free+Edition+Demo+VM
   - View the howto videos at gethue.com
3. Get help and training
   - Join cdh-user@cloudera.org
   - Visit cloudera.com/content/dev-center/en/home.html
   - Get trained and certified by university.cloudera.com
4. Contact us
   - eva@cloudera.com / @EvaAndreasson
   - On-line contact form
     - http://cloudera.com/content/cloudera/en/about/contact-us/contact-form.html

cloudera®
Ask Bigger Questions

**Time to...**

©2012 Cloudera, Inc.