SOLVING REAL AND BIG (DATA) PROBLEMS USING HADOOP

Eva Andreasson

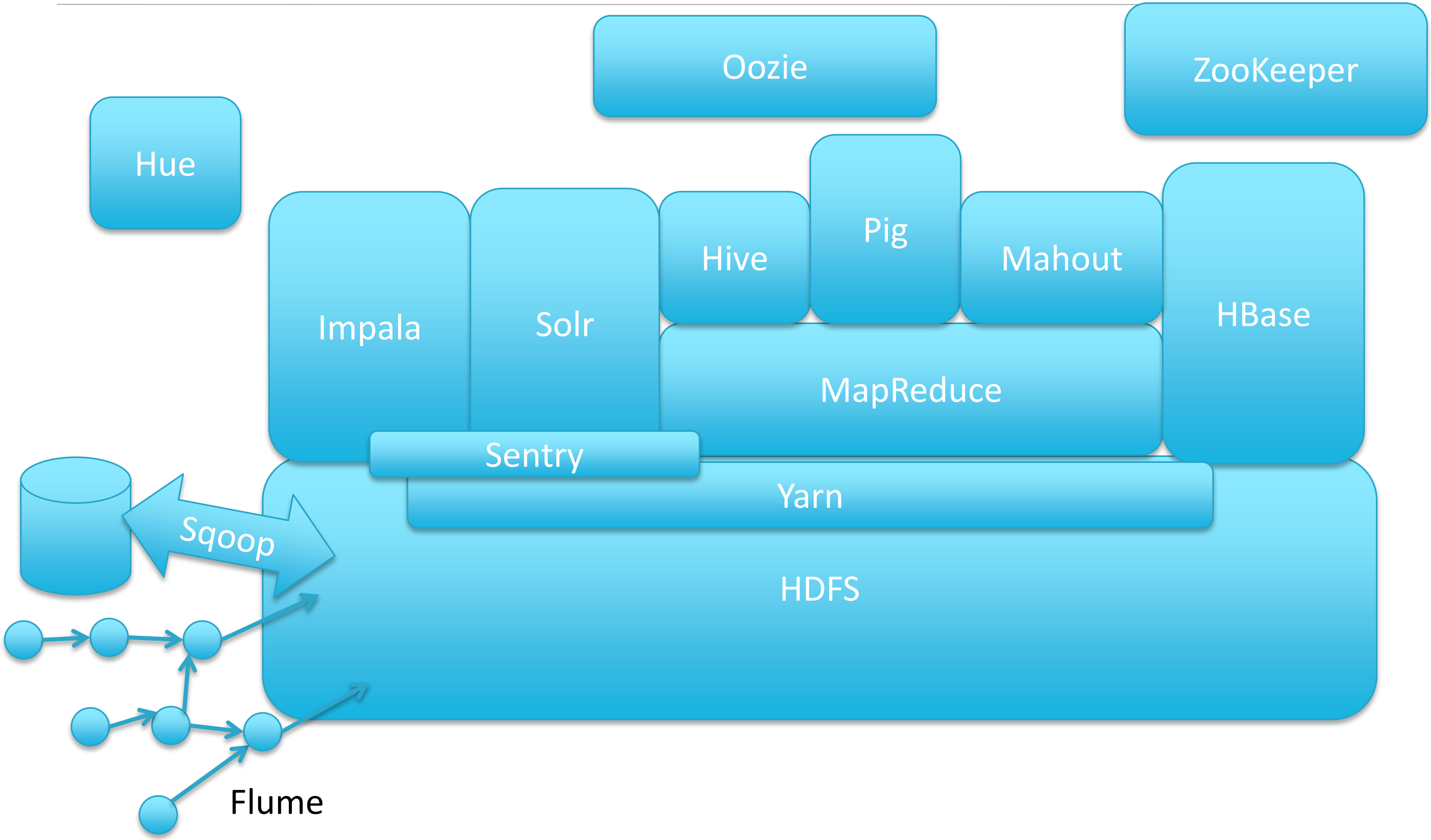
Cloudera

Most FAQ:

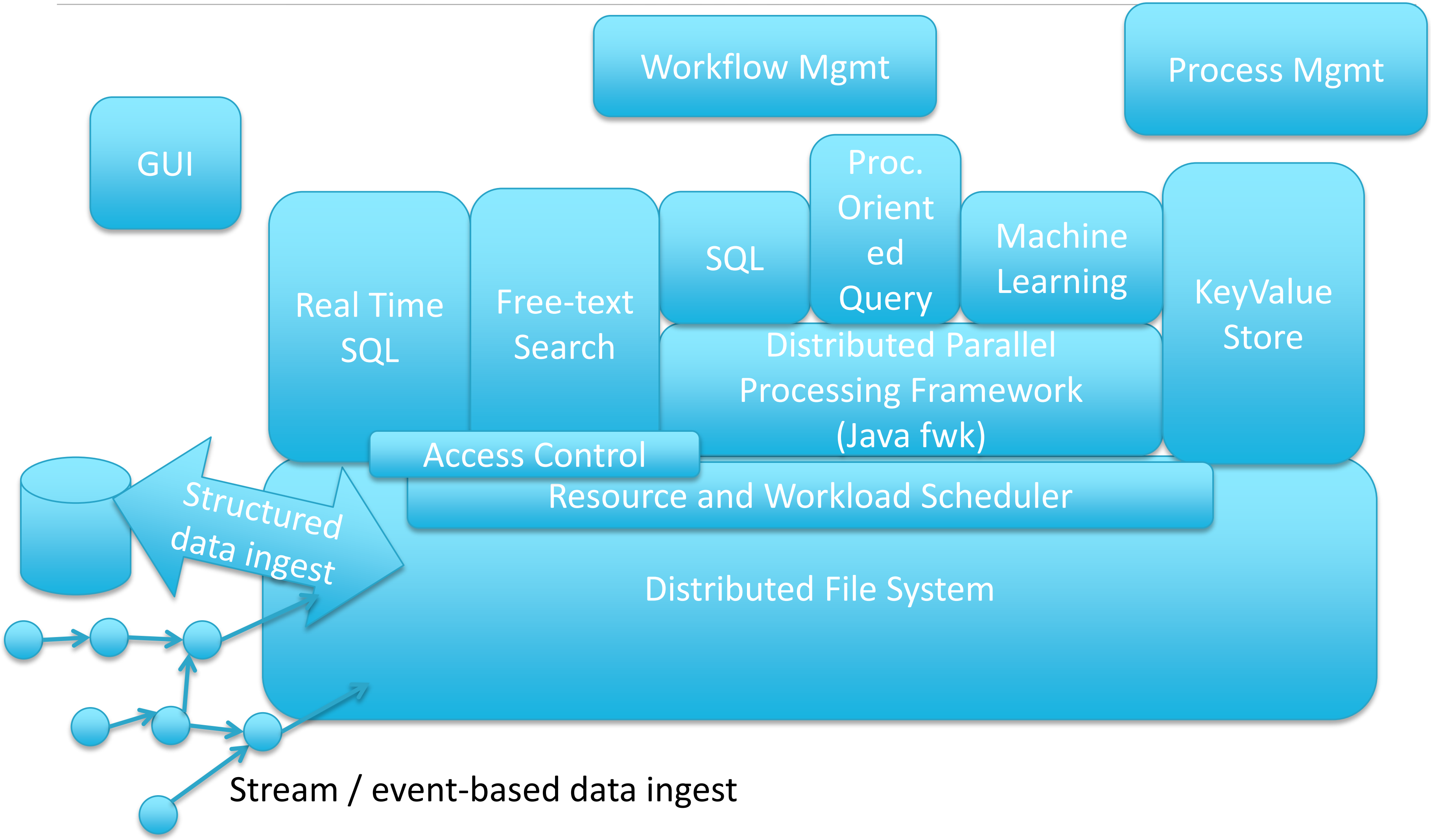
How do organizations use Hadoop?

Super-Quick Overview!

The Apache Hadoop Ecosystem – a Zoo!



The Hadoop Ecosystem – Explained!



Two Views

#1: Scale Data Processing at Low Cost

- Do what I usually do, but on a larger set of data
- Do my complex queries, but within a reasonable time

#2: Break Silos and Ask Bigger Questions

- What *new insights* can we achieve by combining siloed data sets?
- What else can we find by asking questions over new types of data?

There is no box!

Some Typical Use cases

What Organizations Do: Offload ETL

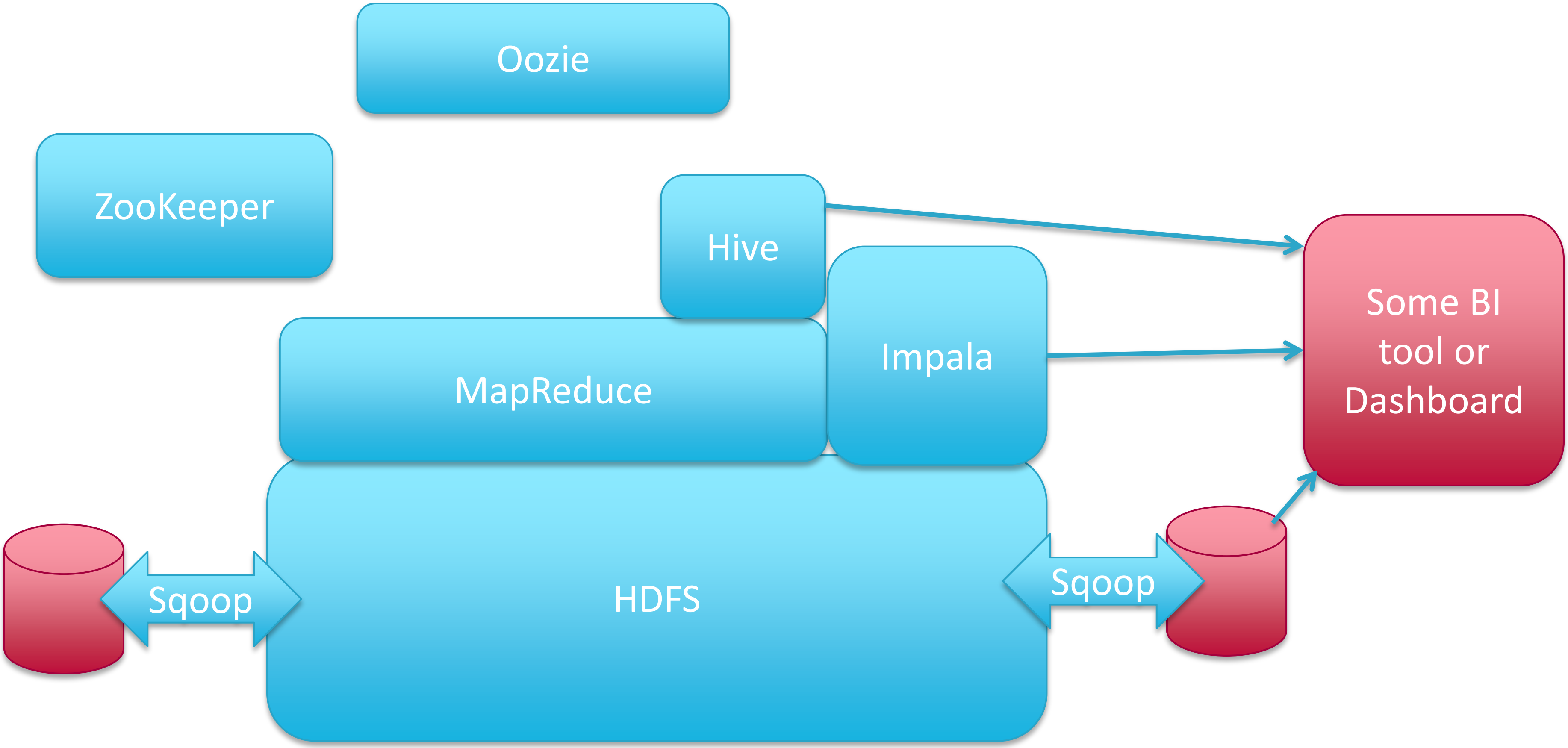
- Common use case:
  - ETL data processing workload
  - Data volume is growing
  - But *fixed time window* for data delivery

- Related side use case
  - Complex queries on the data either take unacceptable time or can’t be deployed at all
    - Cost, volume of records involved, response time, or limited data…

What Organizations Do: Batch ETL

- Example: A Network and Storage solution company – pro-active support
- Challenge
  - 600000 “phone home” machine generated log transmissions needed to be processed every week
  - 40% of the logs need to be transmitted within 18 hours each weekend
  - Expected data growth of ~7TB a month – causing SLA bottlenecks!
  - Complex queries taking weeks or not even possible to run
- Solution
  - Achieved a cost-efficient and linearly scalable storage and data processing solution
  - Can now handle 7TB/month data growth and stay within the 18 hr SLA-bound time window
  - Faster and more flexible analytic capabilities
    - Can now correlate disk latency with manufacturer (a 24 billion records report btw) and achieved a 64x query performance improvement (from weeks to hours)
    - Can now run a pattern matching query that would help detect bugs (a 240 billion record query btw!!)
  - TCO freed up budget for other customer-focused projects

Example Architecture



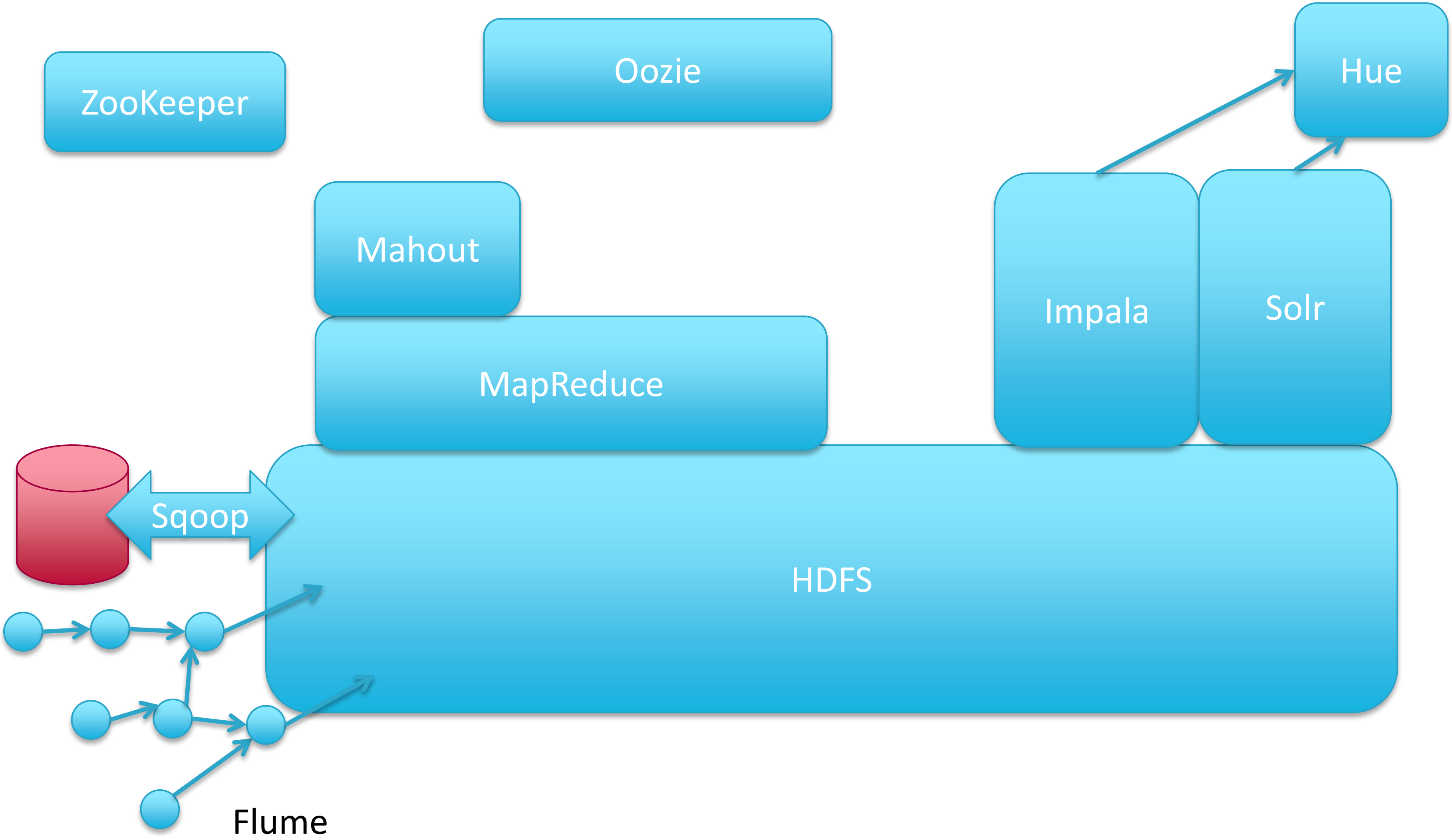
What Organizations Do: Log Processing

- Common Use Case
  - Too many log types, too high volume, and growing…
  - Need for multiple workloads on the same log data
    - Capacity planning
    - Historical load trends in correlation with special activities elsewhere in the org
    - Near real time production issue resolution
    - Anomaly or outlier detection
  - Traditional systems cant easily scale with the load, nor adapt to all the types of data that need to co-exist to answer complex correlation queries

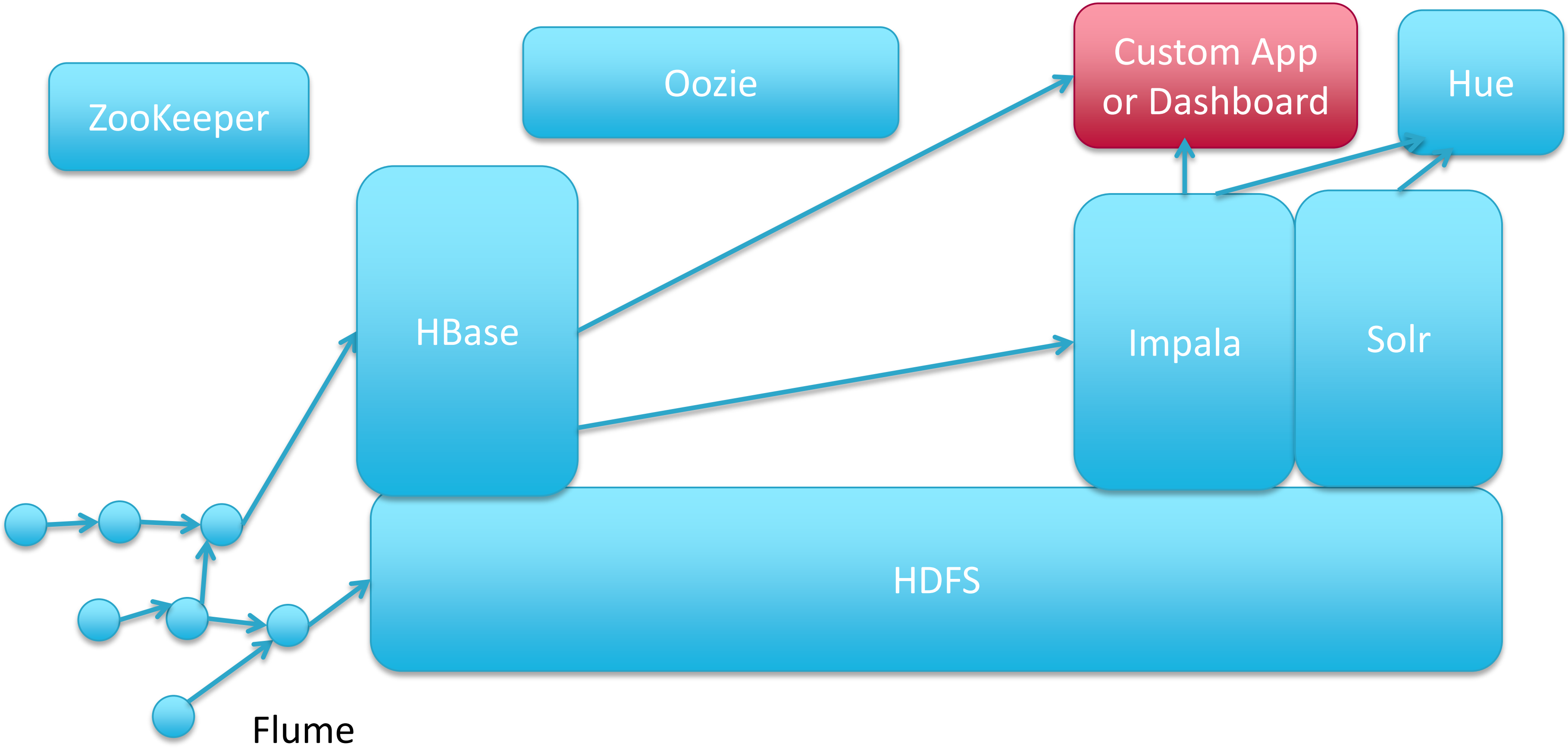
What Organizations Do: Log Processing

- Example: Global Financial Services Firm – anomaly detection
- Challenge
  - Online trading causing data exponential growth
  - Traditional systems could only handle current load, and it took weeks to process current data loads
  - Could not store more than 1 year of data cost-efficiently
- Solution
  - Can now store 200-300TB of data and handle a 2-4TB daily ingestion load
  - Uses Impala for real time queries on that data, e.g. a month data scan happens in 4 seconds vs 4 hours..
  - Monthly reports can now be generated in hours instead of days
  - Saved $30M in IT costs and prevented future growth costs

Example Architecture



Example Architecture



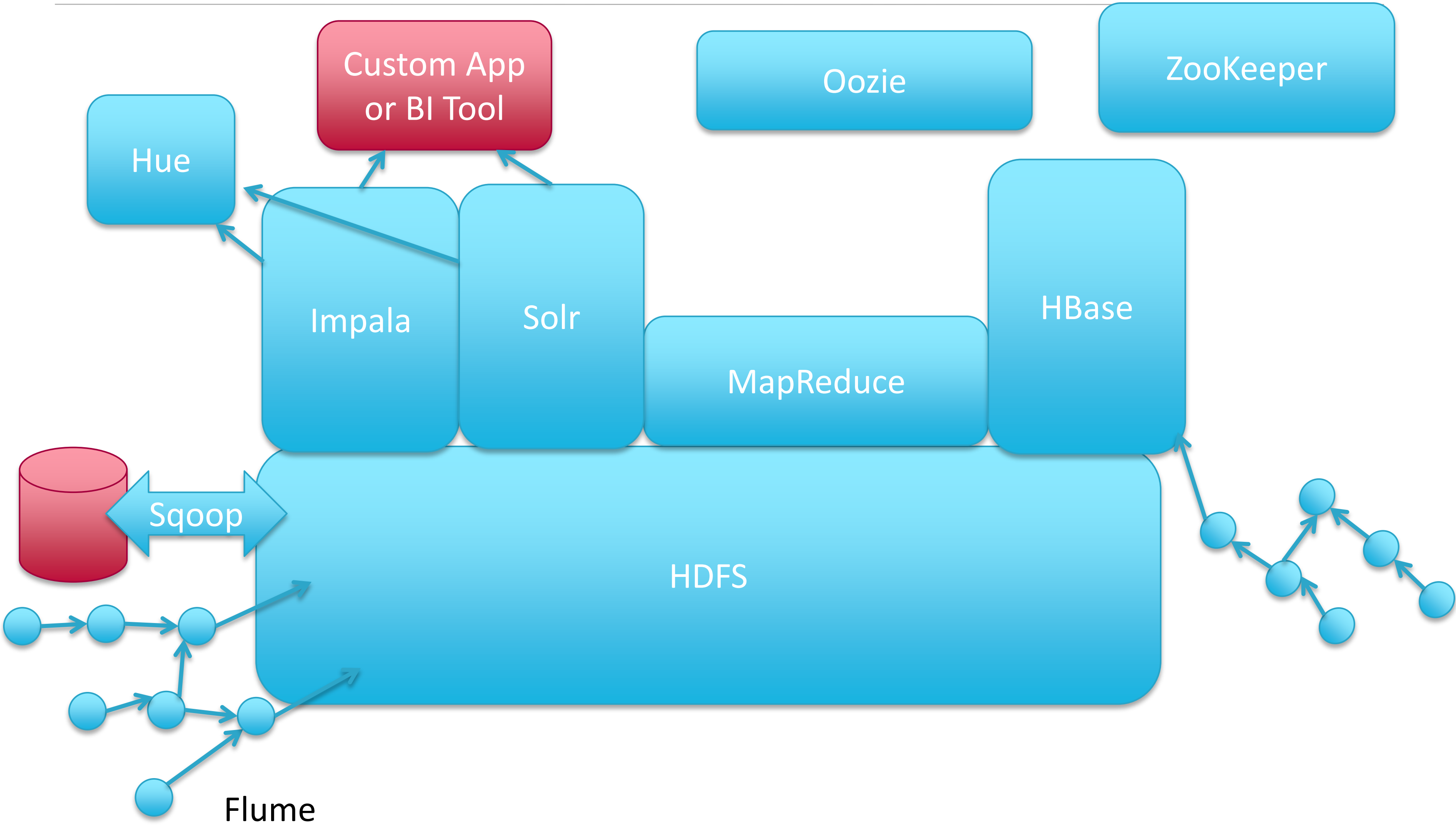
What Organizations Do: Combine Silos

- Common Use Case
  - Customers seek a 360-view of clients, patients, or customers to provide better services, support, competitive offerings, or marketing
  - Data lives in separate (and sometimes old) silos – costly!
    - Maintenance, overlap, access bottlenecks,
  - Some data is “impossible” to access in a timely manner
  - Traditional systems can’t cost-efficiently store all data, handle all data types (and new types added dynamically) and serve the various workloads / clients of the system

What Organizations Do: Combine Silos

- Example: Global On-line retailer
- Challenge
  - Need to correlate online/offline data across disparate, costly legacy DWs
  - Detailed data from every cash register at every store over a 10+ year history across 1,000's product categories (22 subsidiaries?)
  - One data source ~4 weeks to get access to – inhibits productivity
- Solution
  - 250-node cluster of Cloudera + Impala
  - Can now store 1PB over 250 nodes and grow at very low cost
  - Consolidated environment for query and machine learning – no data access bottlenecks anymore
  - Able to correlate all customer, product, and sales data for a 360-degree view of their customer

Example Architecture



And there are many more….

- Image processing
- Suicide prevention / event prediction
- Product and process improvements
- Genome sequence processing
- Hospital – treatment – patient matching
- Travel-logistics-path optimization
- Recommendation engines
- Clickstream analysis and web experience optimizations
- …

What to Consider

- Key benefits of moving a workload to Hadoop
  - Linear scale without the extreme price tag
  - Lots of flexibility – you can always change your ingest pipelines or data models later with low impact and low cost
  - Ability to combine and analyze previously siloed data sets
  - Opens the door to expand business with new questions – cross organizations!
- Questions to investigate:
  - Make sure to have a validated business use case
    - Does your organization have a need to develop a strategy for handling data growth or a need for combining data sets?
    - What workloads can actually move to Hadoop?
      - Is Hive QL compliant with SQL?
      - What about real time workloads and OLTP?
    - What would be gained that the business side would care about?
      - Clear measurable goals makes life easier!
  - Make sure your organization is prepared
    - What training and support is available?
    - What about supportability and production visibility?
    - How does Hadoop integrate with my environment?
  - Make sure you know what would be required for production in your environment
    - What about Security? PCI compliance?
    - What about production visibility?
    - What about HA and DR?

Summary

What you (Hopefully!) Learned Today

How organizations use Hadoop

To Learn More…

1. Read some good stuff
   - Order the Hadoop Operations book (http://shop.oreilly.com/product/0636920025085.do) and/or the Definitive Guide to Hadoop (http://shop.oreilly.com/product/0636920021773.do)
   - Visit Cloudera's blog: blog.cloudera.com/
2. Play on your own
   - Cloudera QuickStart VM: https://ccp.cloudera.com/display/SUPPORT/Cloudera+Manager+Free+Edition+Demo+VM
   - View the videos at gethue.com
3. Get help and training
   - Join or send an email to: cdh-user@cloudera.org
   - Visit the Cloudera dev center: cloudera.com/content/dev-center/en/home.html
   - Get training: university.cloudera.com
4. Contact Cloudera
   - eva@cloudera.com
   - On-line contact form: http://cloudera.com/content/cloudera/en/about/contact-us/contact-form.html

Quizz: What is the *Real* Big Data Challenge

- Technology?
- Knowledge?
- **People?**

Key Take-Away

There is no box!!

Transform the Economics of Data

Traditional Data Warehouse

Add 100 TB =

$2M TO $10M

in incremental spend

With Cloudera

Add 100 TB =

$200K

1/10th the cost of legacy systems

Q&A

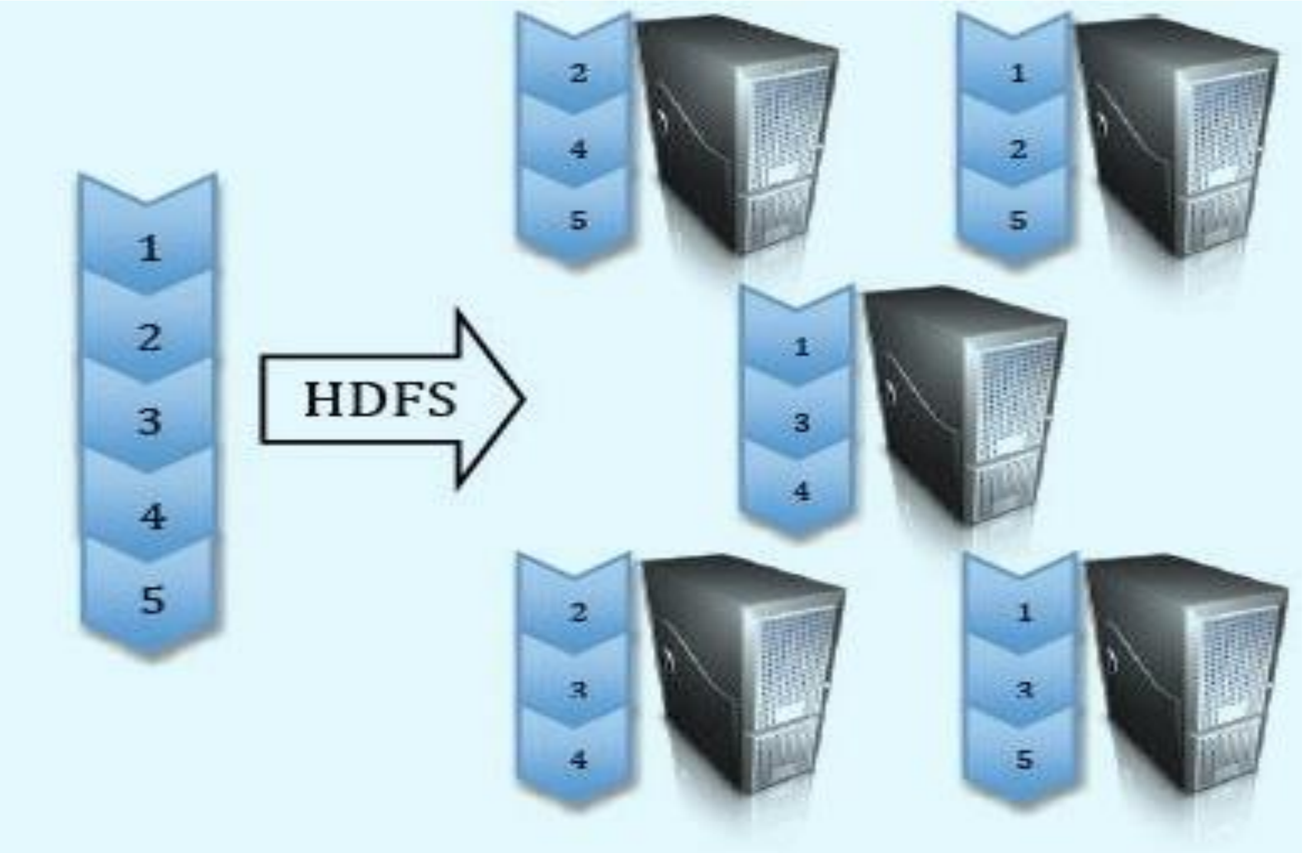
Don't forget to vote!!

cloudera®

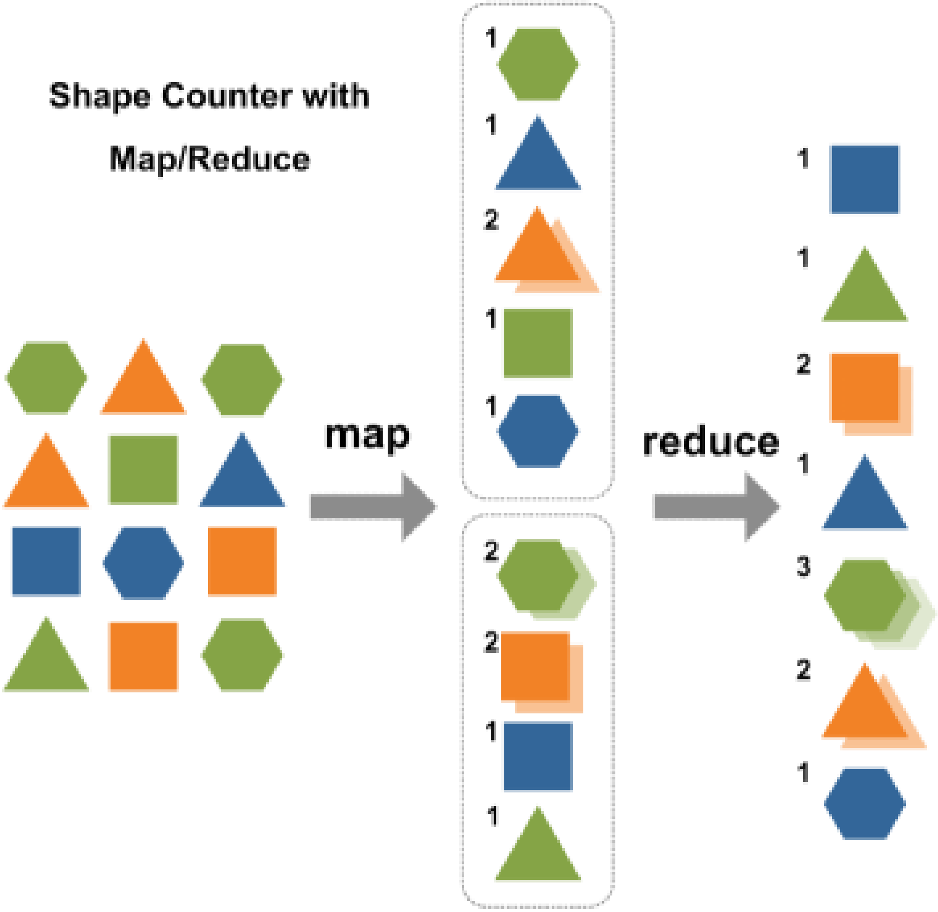
Ask Bigger Questions

©2012 Cloudera, Inc.

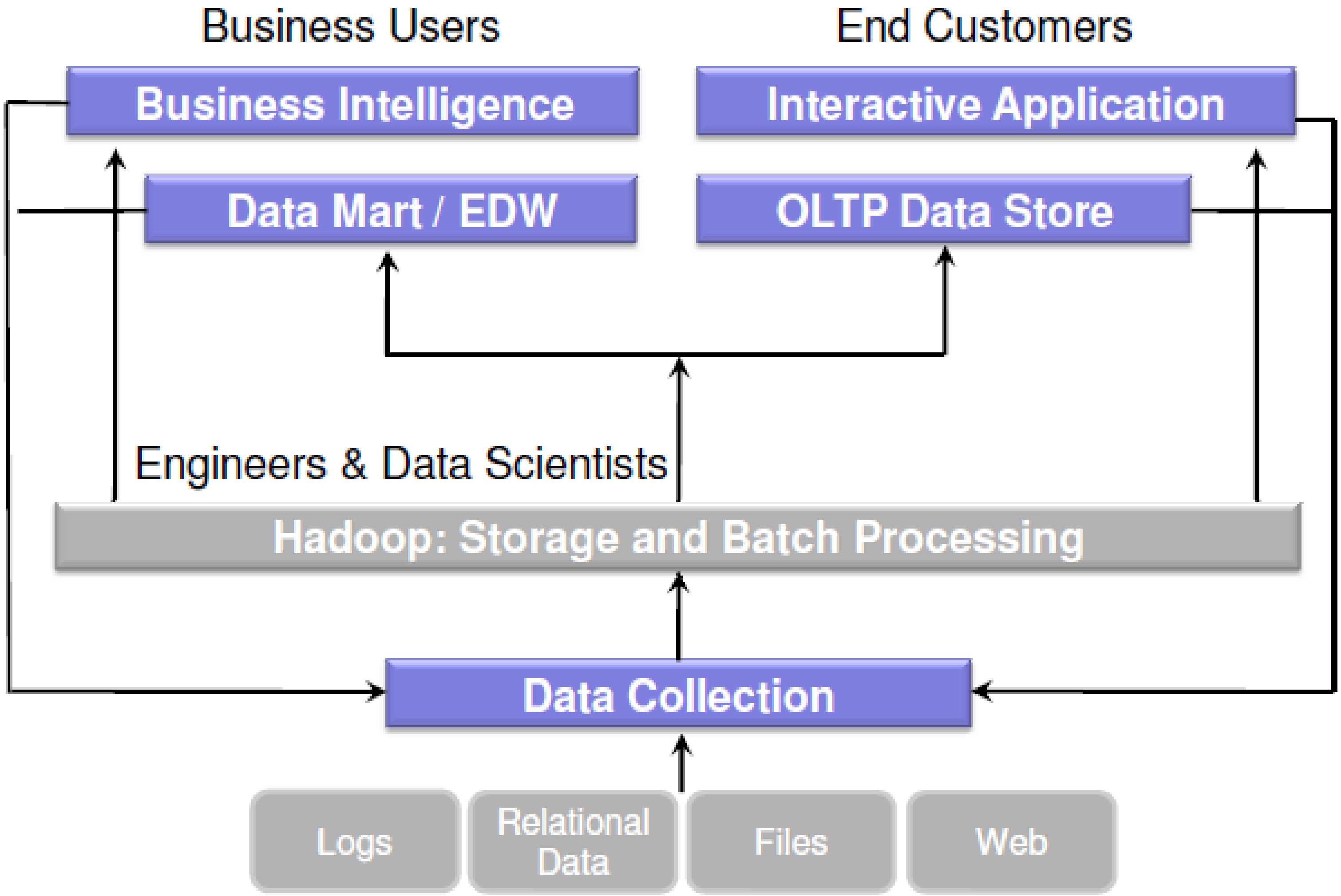
Hadoop Distributed File System (HDFS)



MapReduce: A scalable data processing framework



Architecture for Hadoop in the Enterprise


Business Users
End Customers
Business Intelligence
Interactive Application
Data Mart / EDW
OLTP Data Store
Engineers & Data Scientists
Hadoop: Storage and Batch Processing
Data Collection
Logs
Relational Data
Files
Web