What Can Hadoop Do for You?

@EvaAndreasson | Sr. Product Manager, Cloudera

2014

Agenda

- Today’s data challenges
- Apache Hadoop and its ecosystem 101
- Common use cases
- Where to learn more
- Q&A

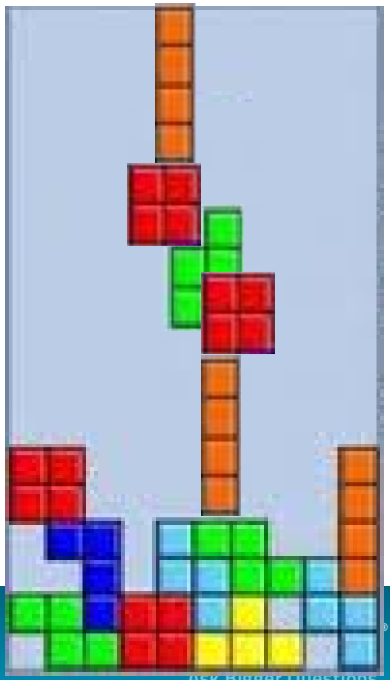
Today’s Data Challenges

Key Challenge #1: Volume

- “Return On Bytes”
  - How to cost efficiently query, manage, and store TBs (or even PBs) of data?
- Pre-mature data death
  - Off-disk and archived data difficult and costly to access
- Forced data silos
  - Costly copies and moves of data
  - Organizational blind-spots

Key Challenge #2: Velocity

- Enough time to process raw data before you need it
  - Data ingest from sensors, cameras, feeds, streaming, logs, user interactions…
  - Raw data structuring for various ETL and DB models



Key Challenge #3: Variety

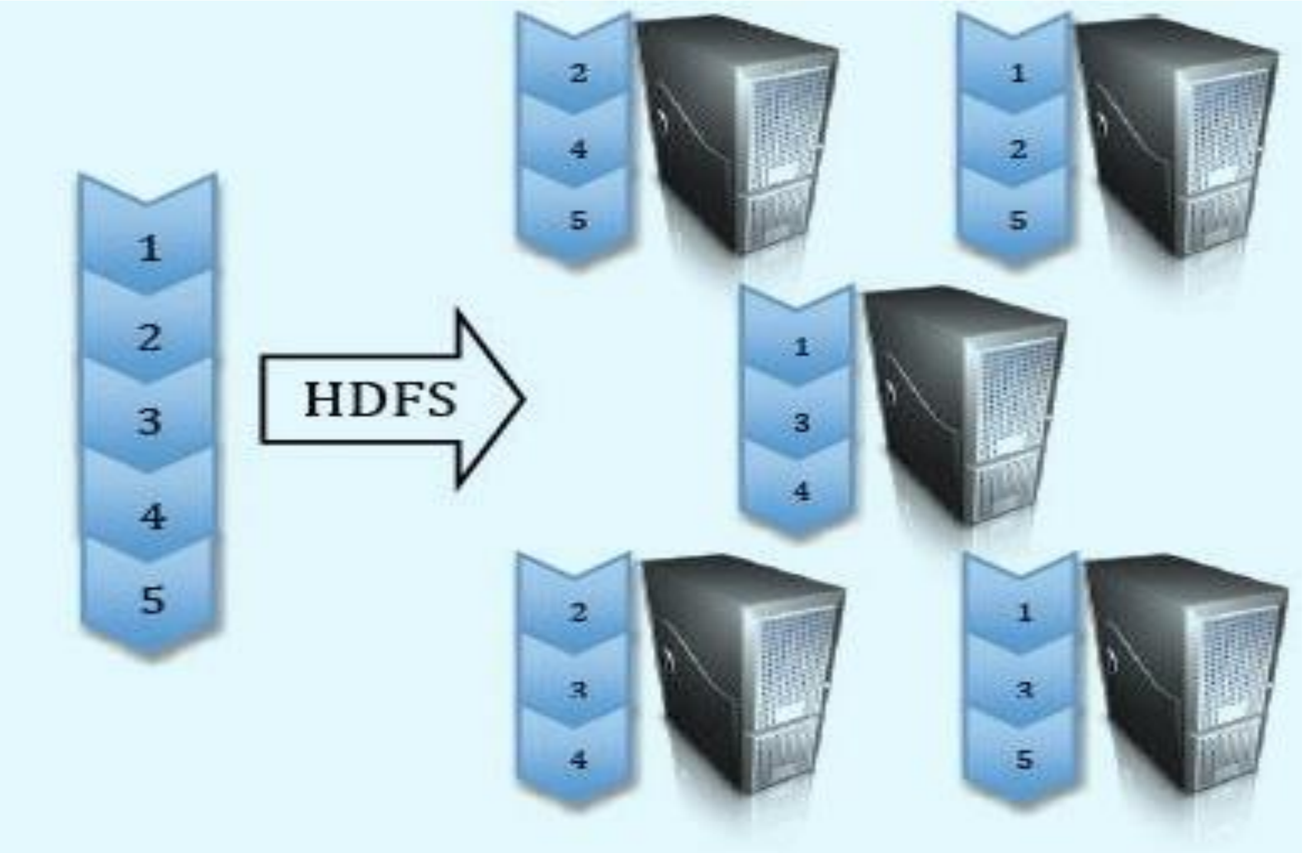
- Costly adaption to new data types
  - Saving account info, images, videos, url clicks, logs, and transactional data – together?
- Inflexible data models
  - Major surgery for future queries
    - Most data is modeled for questions we know will be asked…
    - Raw data value loss

Apache Hadoop and its Ecosystem 101

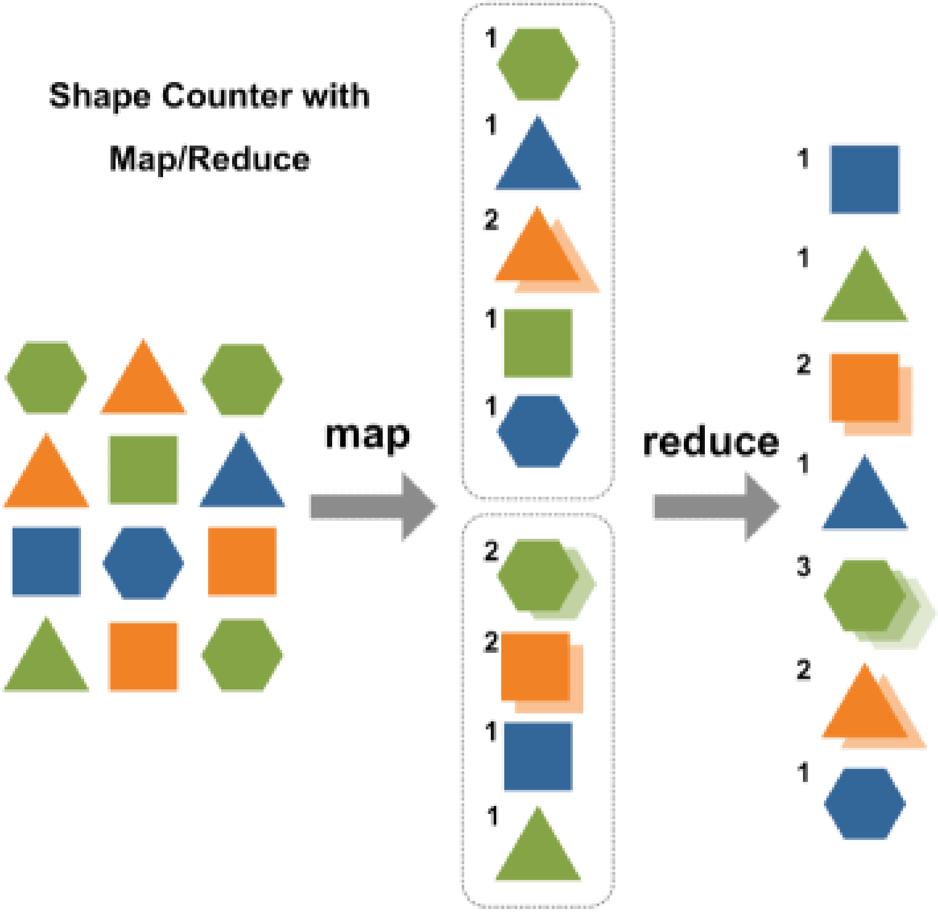
What is Hadoop?

A software framework that lets you find and process data directly *where it is stored* and where you only need to apply structure at query time

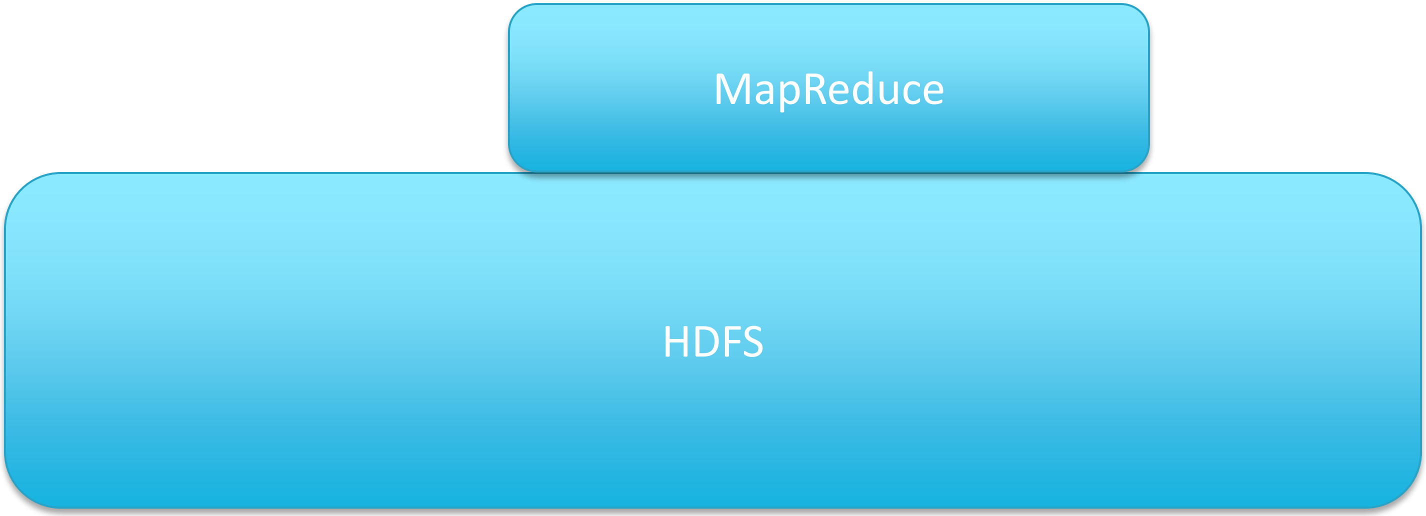
Hadoop Distributed File System (HDFS)



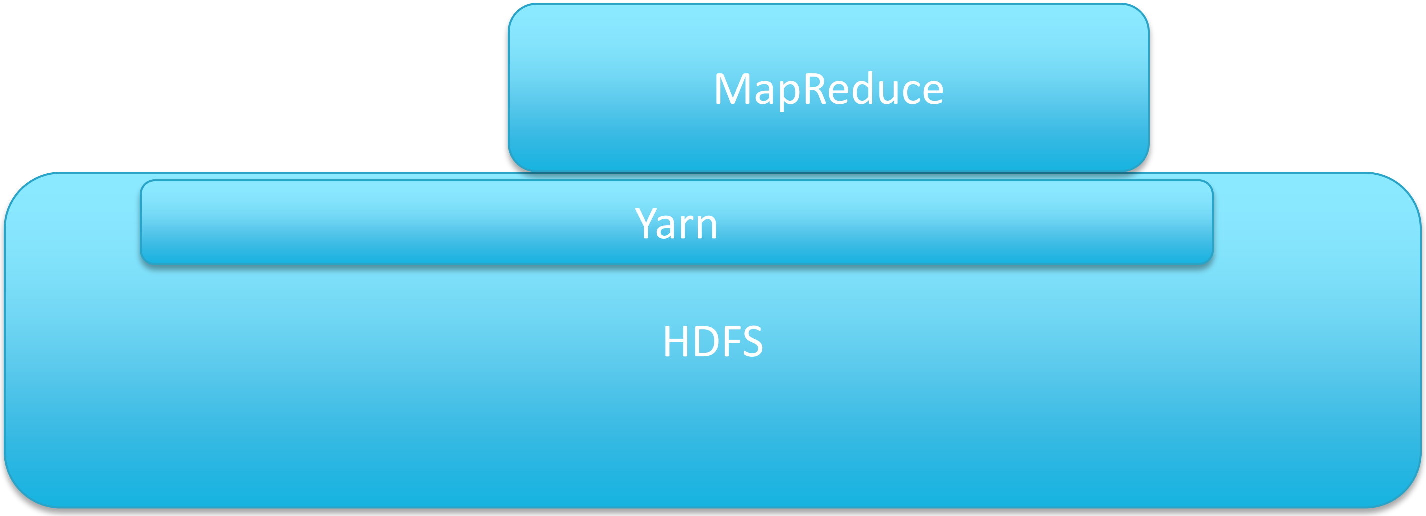
MapReduce: A Parallel, Scalable Processing Framework



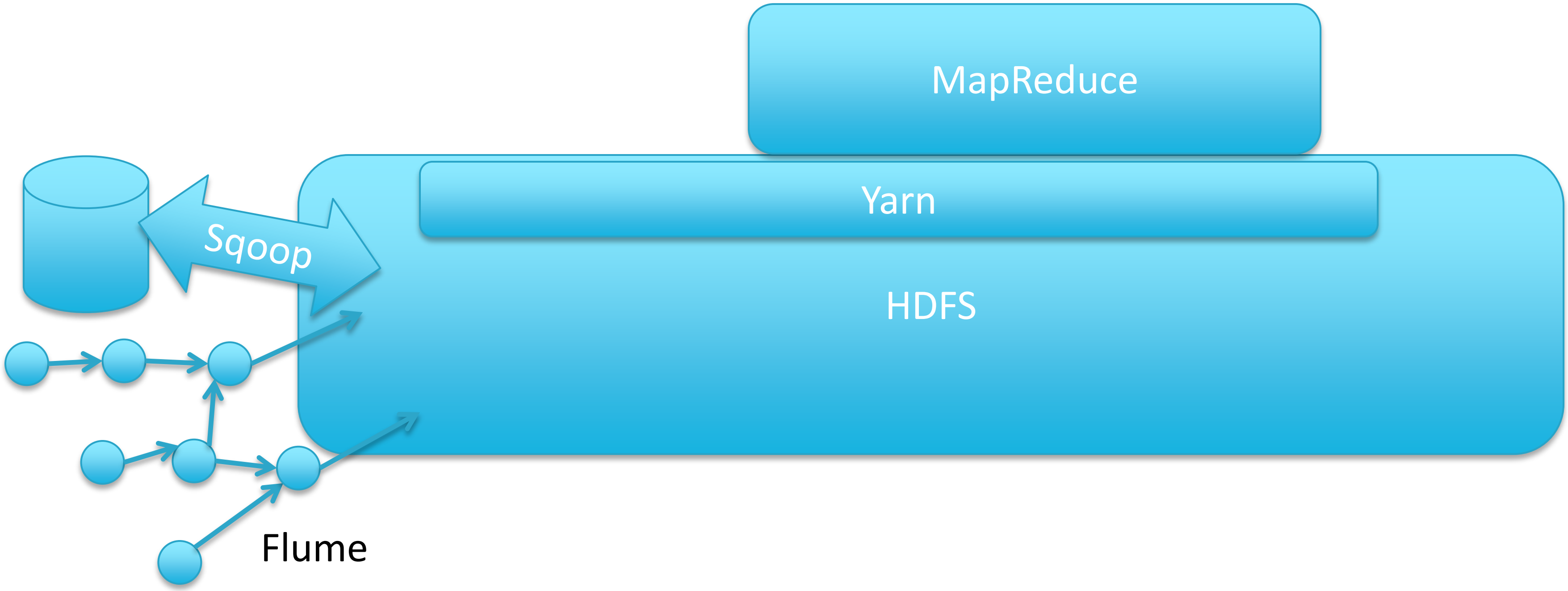
The Apache Hadoop Ecosystem – a Zoo!



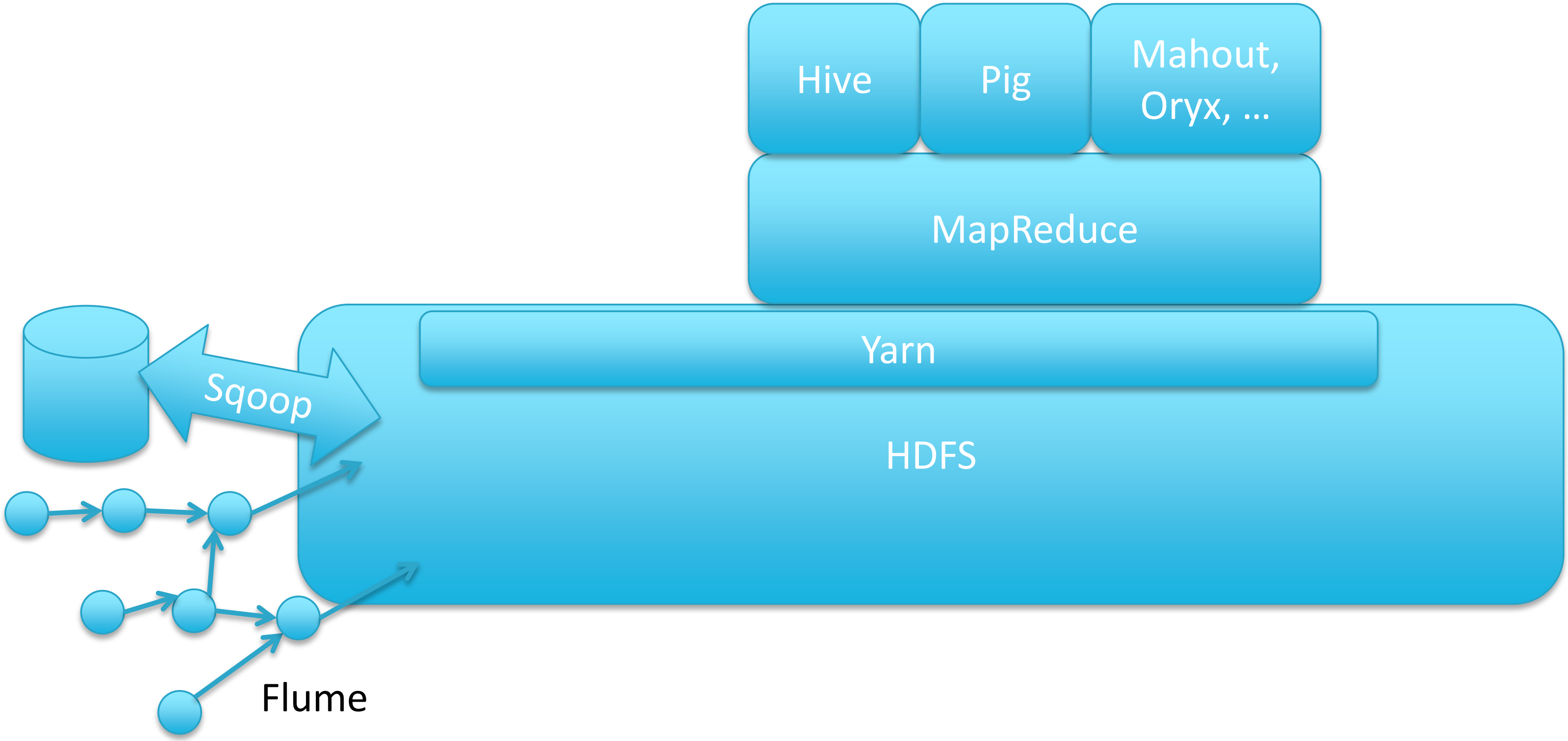
The Apache Hadoop Ecosystem – a Zoo!



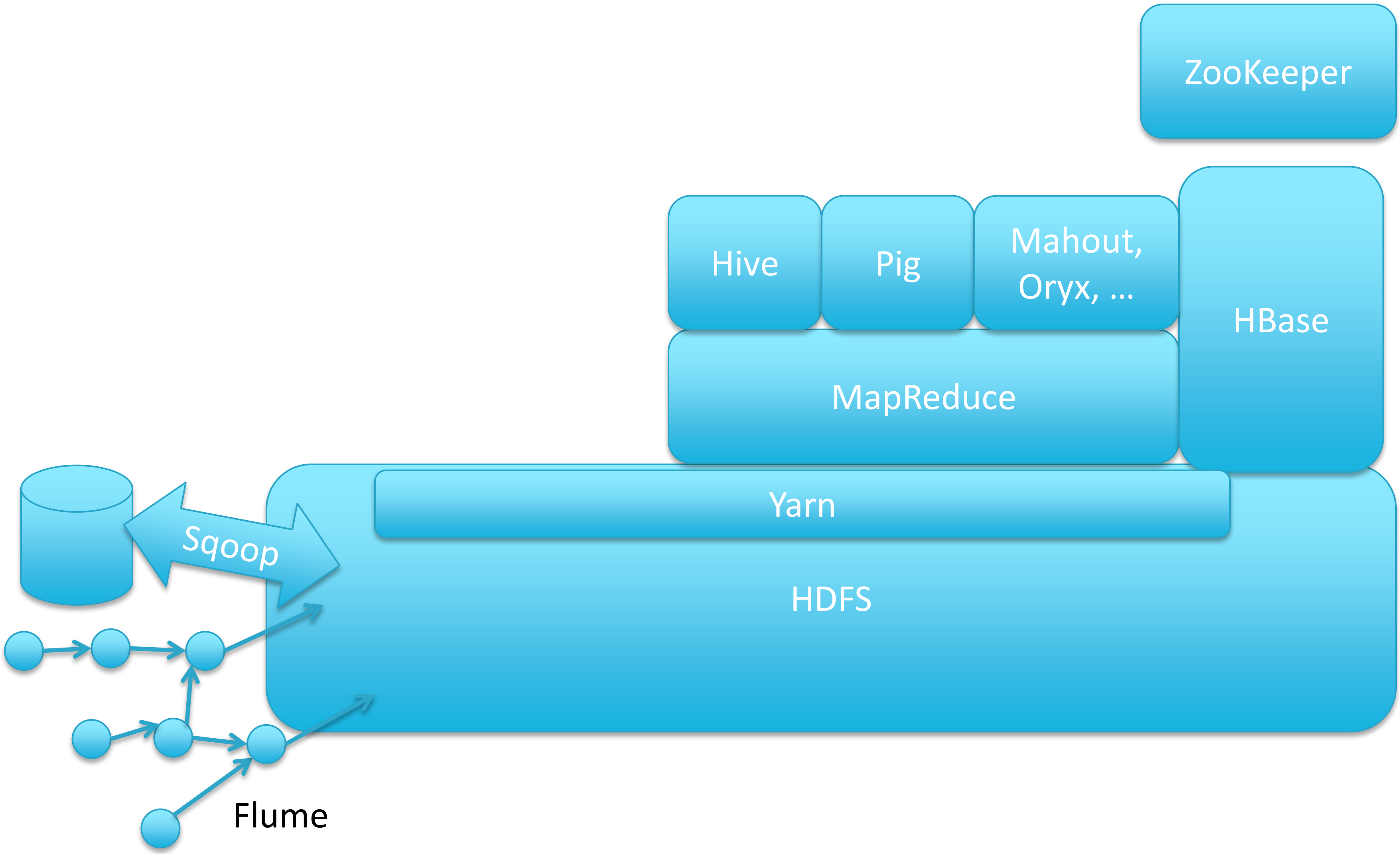
The Apache Hadoop Ecosystem – a Zoo!



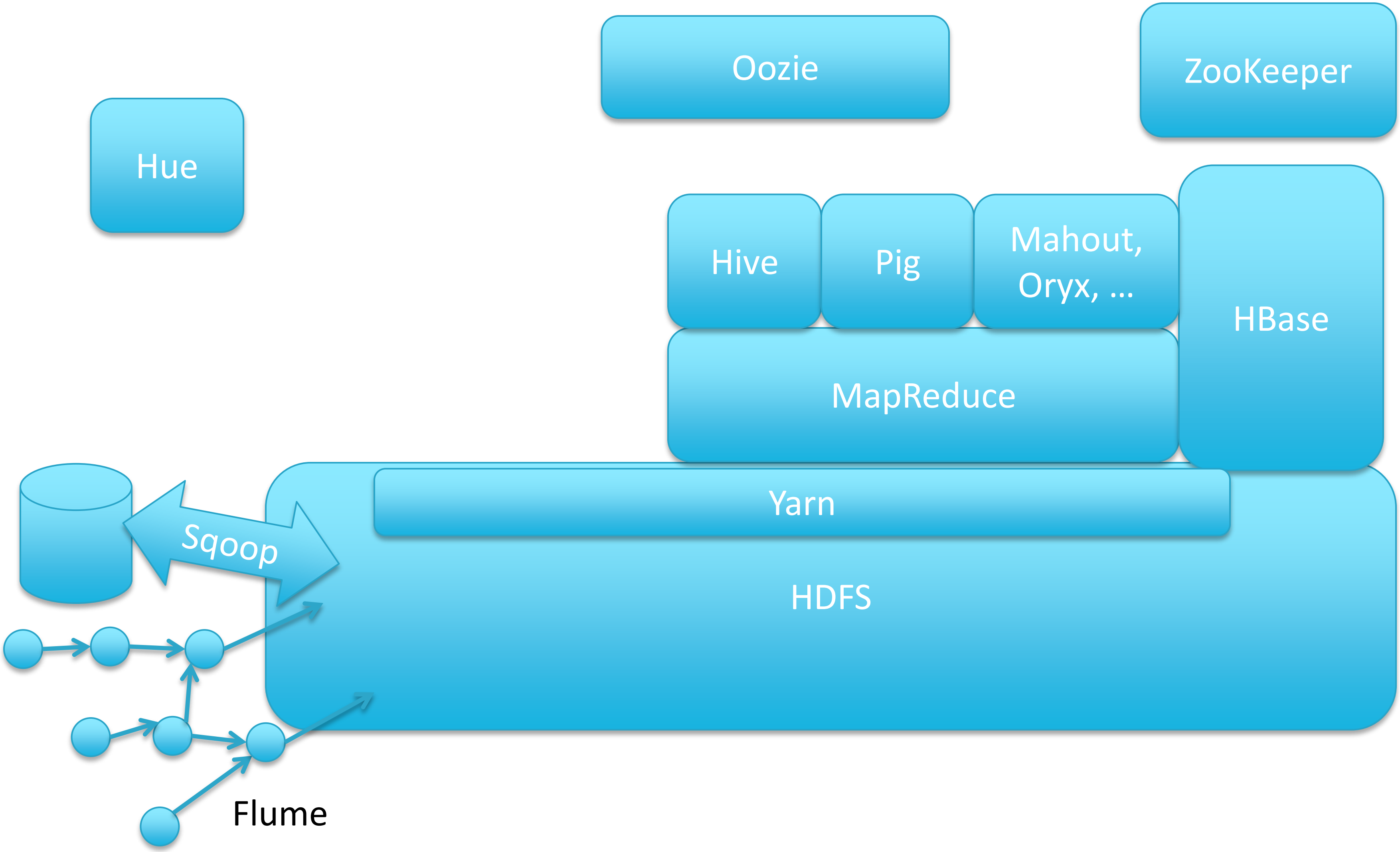
The Apache Hadoop Ecosystem – a Zoo!



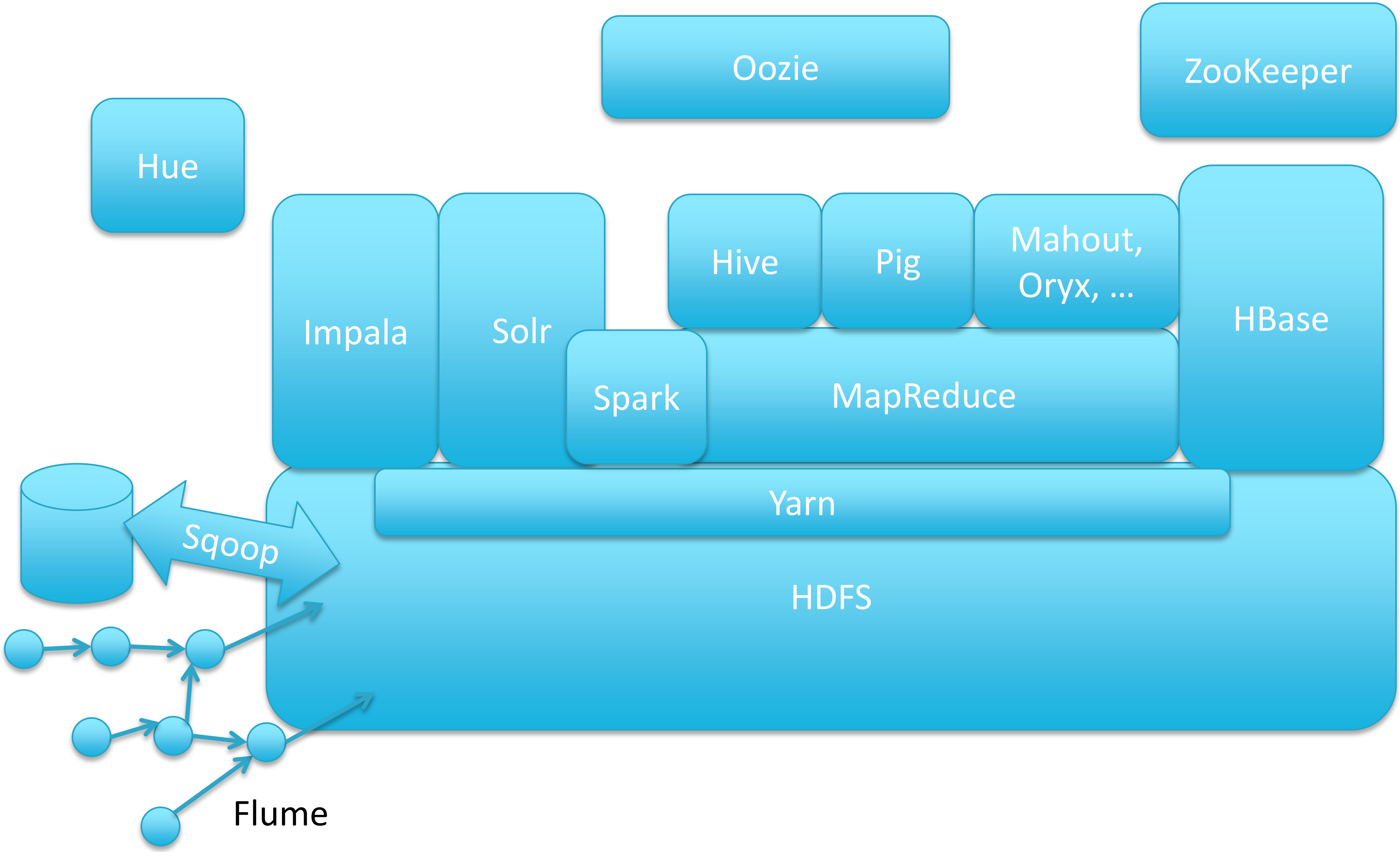
The Apache Hadoop Ecosystem – a Zoo!



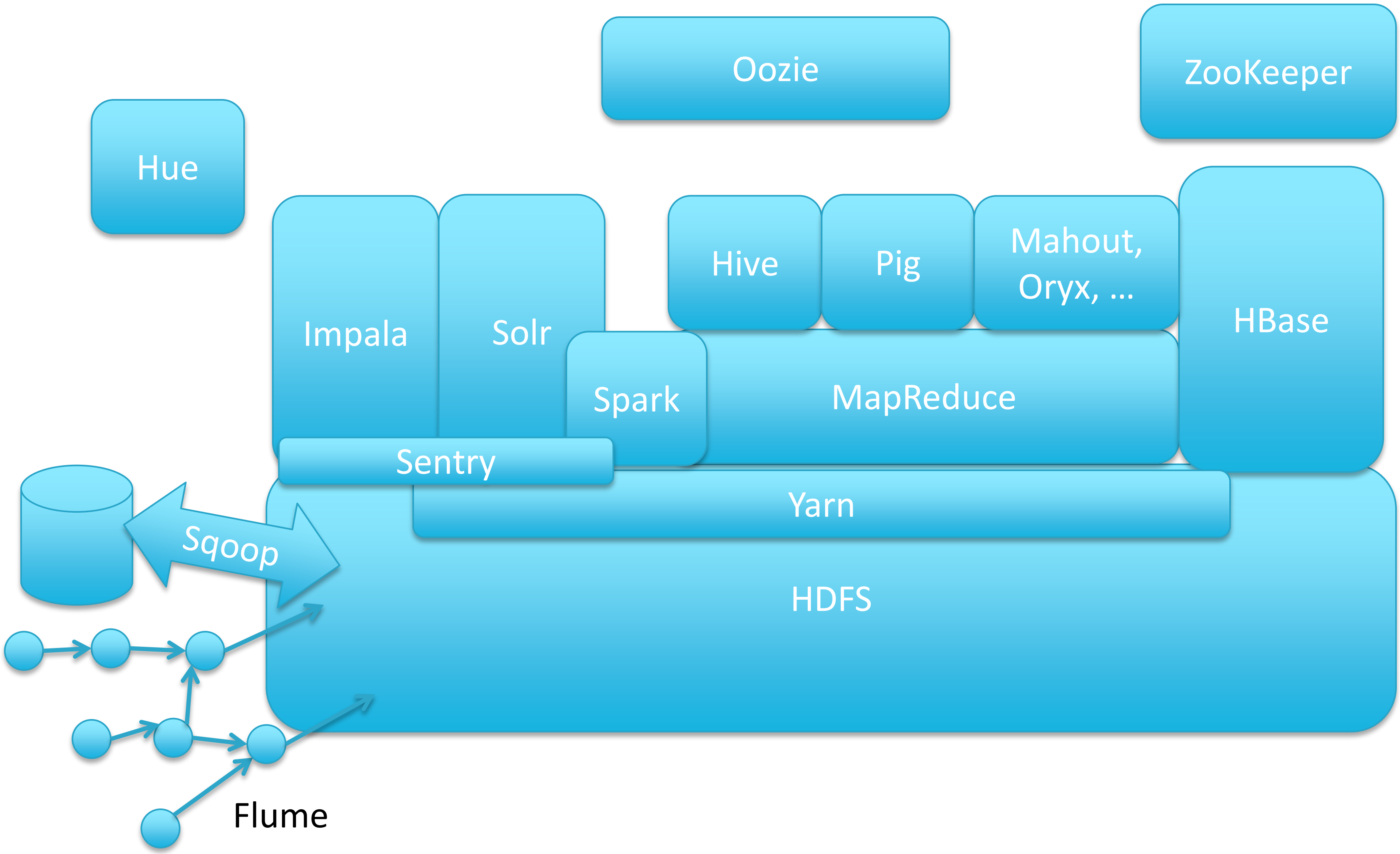
The Apache Hadoop Ecosystem – a Zoo!



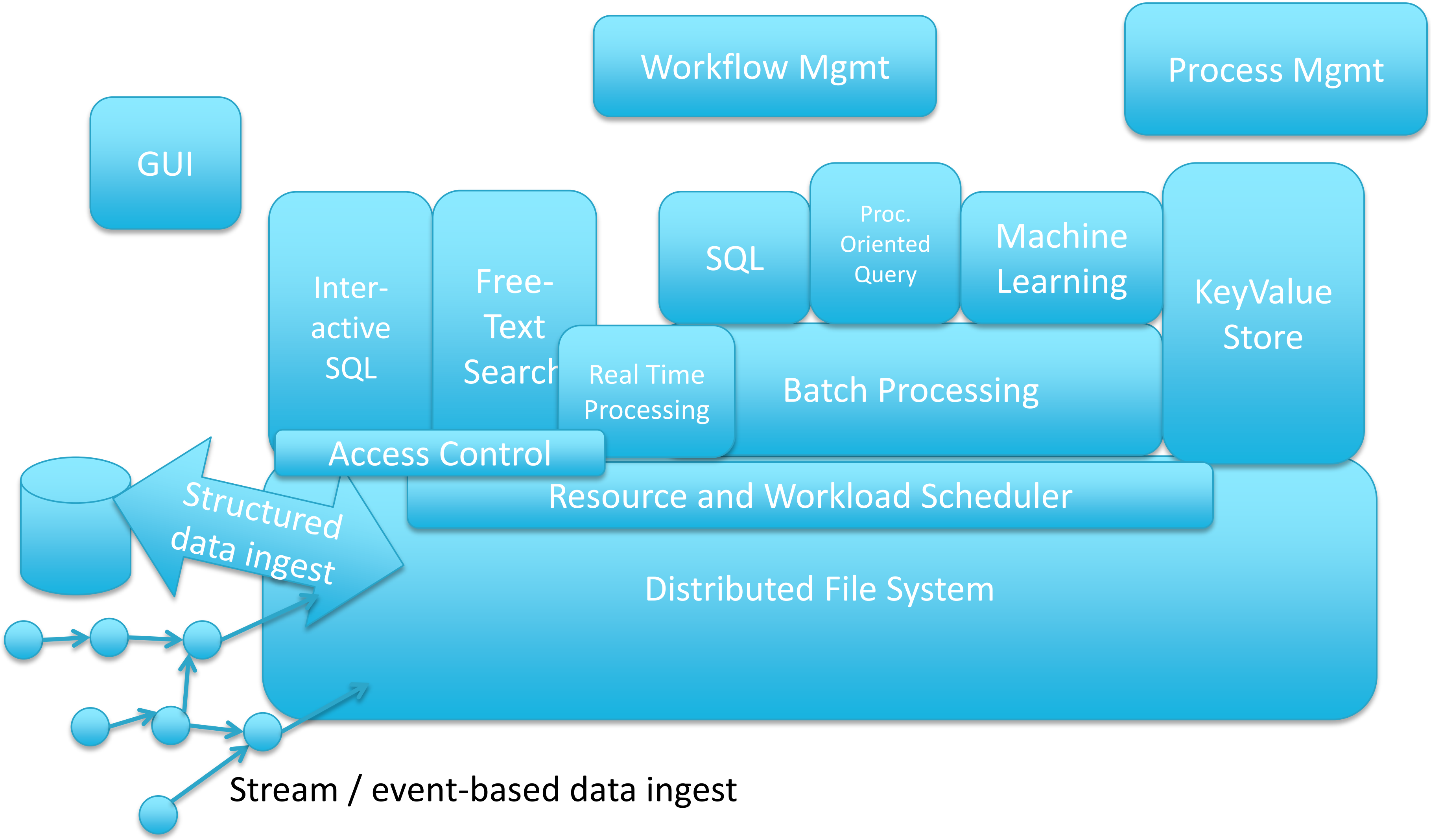
The Apache Hadoop Ecosystem – a Zoo!



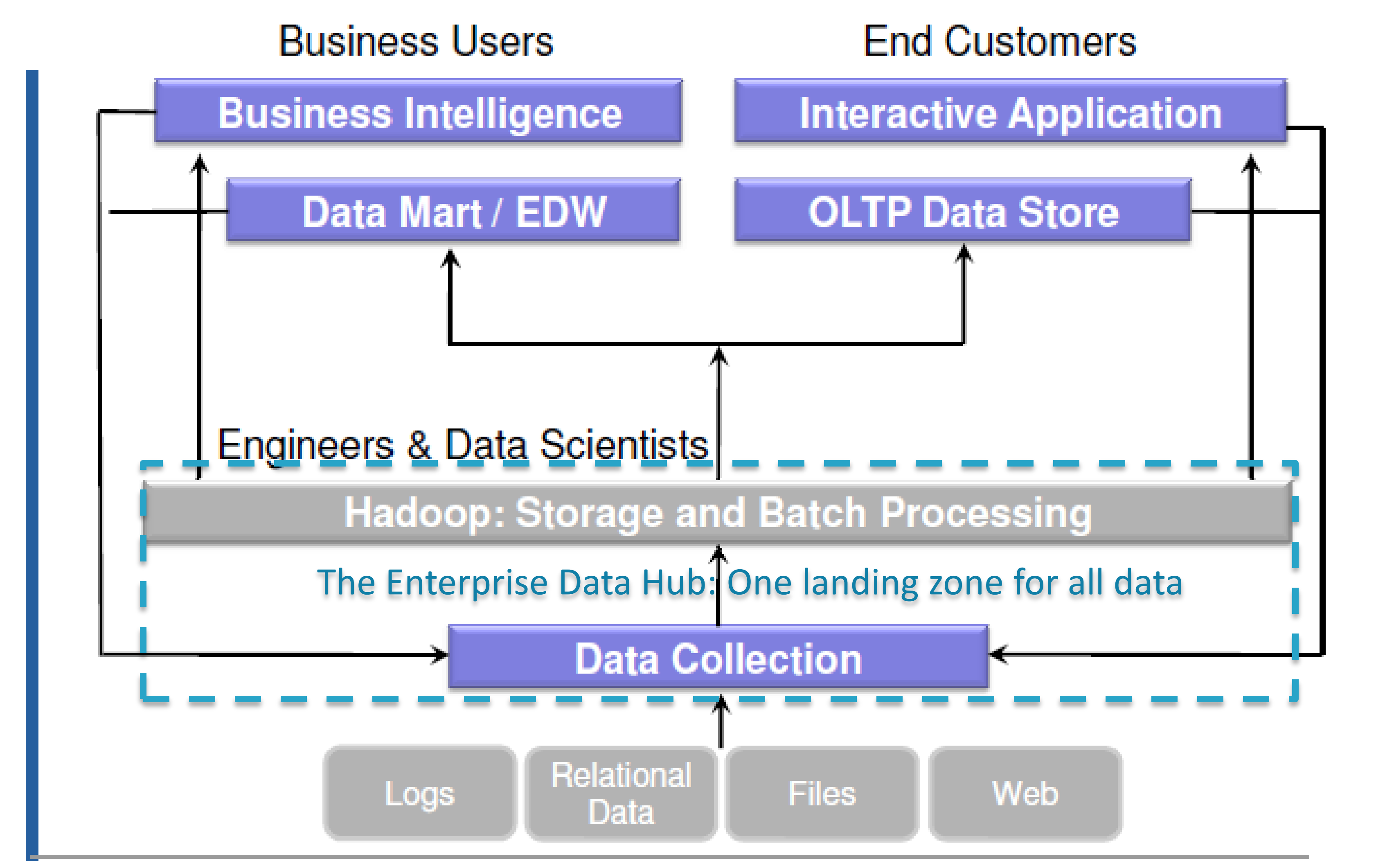
The Apache Hadoop Ecosystem – a Zoo!



The Hadoop Ecosystem – Explained!



Example Enterprise Architecture: Use Hadoop as an EDH



Some Typical Use cases

#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: Create an Active Archive

- Eliminate pre-mature data death
- Data as a (back-chargeable) service across the organization

#3: 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!

There are a Lot of Use Cases…

- Predictive analysis (event prediction)
- Anomaly detection
- Customer profiling
- Recommendation engines
- Clickstream analysis
- Image processing
- Product and process improvements (feedback loops)
- Genome sequence processing
- Object matching
- Path optimization
- …

Do the Same – and More – to a Lower Cost

- Network and storage solution company
- Create pro-active support
  - 600000 “phone home” logs/week
  - SLA: 40% done within 18 hours
  - Complex queries taking weeks or not even possible to run
  - Expected future data growth of ~7TB a month!
- With Hadoop et al and Cloudera’s help
  - Future proof scale
  - Faster and more flexible analytic capabilities
    - Correlation of disk latency with manufacturer (24 billion records)
    - 64x query performance improvement (from weeks to hours)
    - Pattern matching queries in the same infrastructure detecting bugs (240 billion records)
  - TCO freed up budget for other customer-focused projects

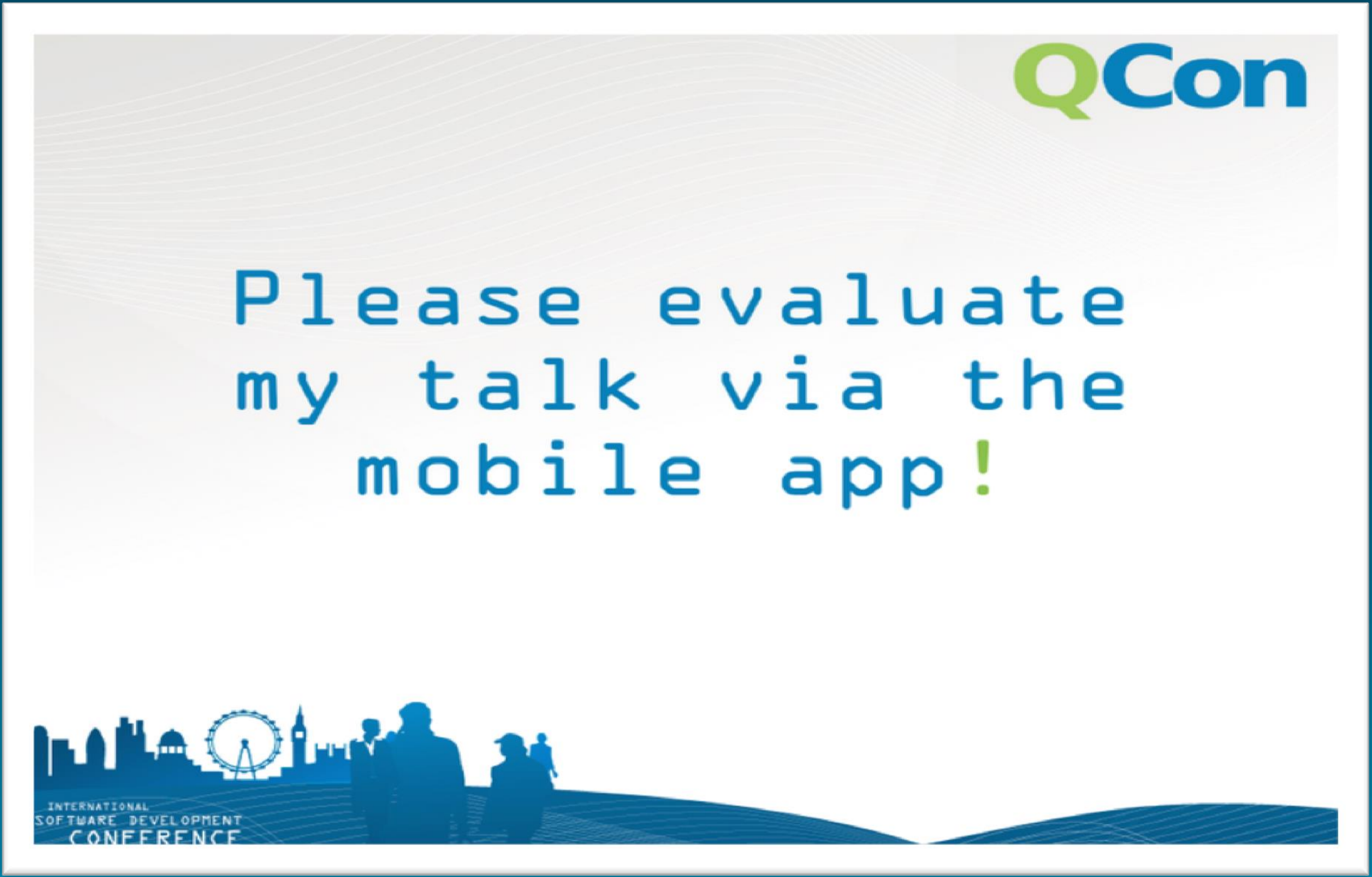
Increasing Revenue Through an Active Archive

- Global on-line retailer
- Need to correlate online/offline data
  - 10+ year history records, 1,000’s product categories
  - Large parts of data archived, complex to access
- With Hadoop et al and Cloudera’s help
  - Unified long-term storage and processing over all data
  - Machine learning and query without data moves
  - Correlate all customer, product, and sales data ad hoc
  - Lead to targeted marketing and increased revenue streams

Where To Learn More?

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



QCon

Please evaluate my talk via the mobile app!

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

Q&A

cloudera®
Ask Bigger Questions