cloudera®
Ask Bigger Questions

Scaling ETL
with Hadoop

Gwen Shapira
@gwenshap
gshapira@cloudera.com

Should DBAs learn Hadoop?

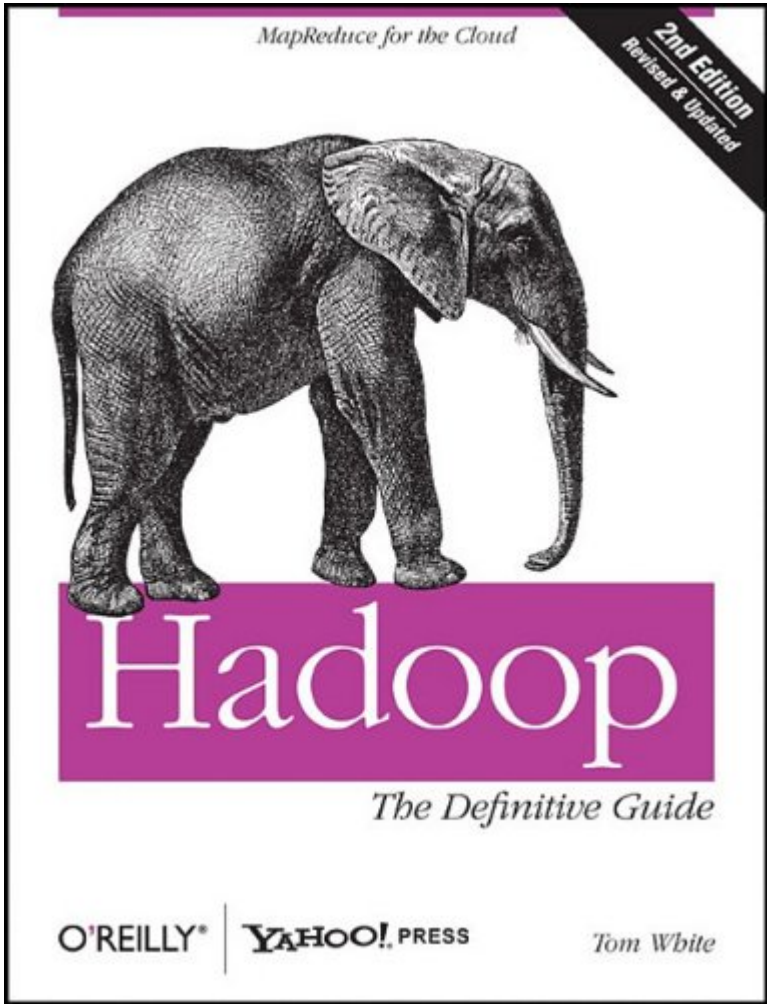
- Hadoop projects are more visible
- 48% of Hadoop clusters are owned by DWH team
- Big Data == Business pays attention to data
- New skills – from coding to cluster administration
- Interesting projects

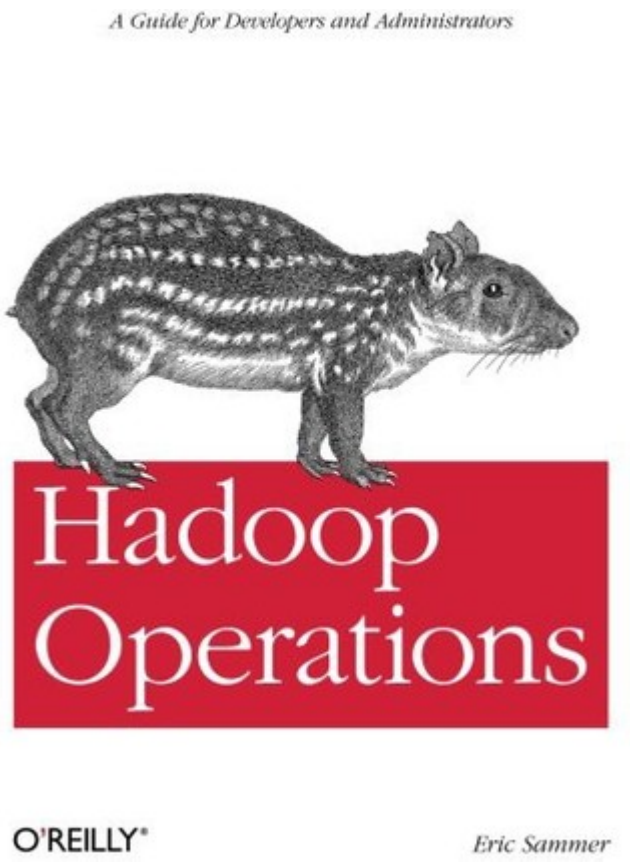
- No, you don't need to learn Java

Beginner Projects

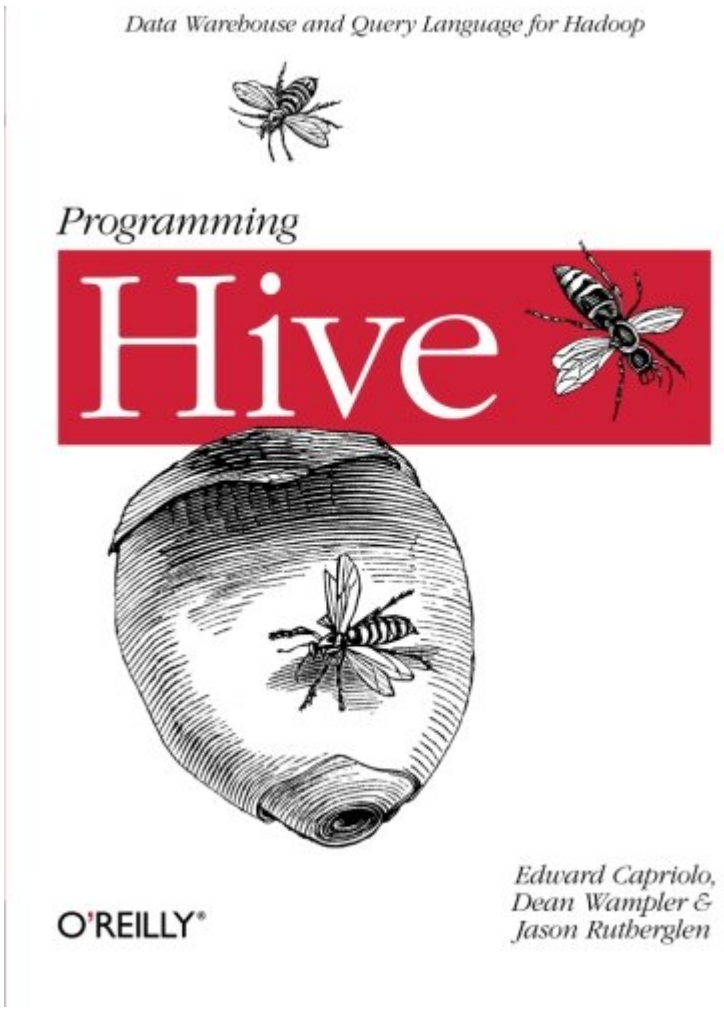
- Take a class
- Download a VM
- Install 5 node Hadoop cluster in AWS
- Load data:
  - Complete works of Shakespeare
  - Movielens database
- Find the 10 most common words in Shakespeare
- Find the 10 most recommended movies
- Run TPC-H
- Cloudera Data Science Challenge
- Actual use-case:
  XML ingestion, ETL process, DWH history

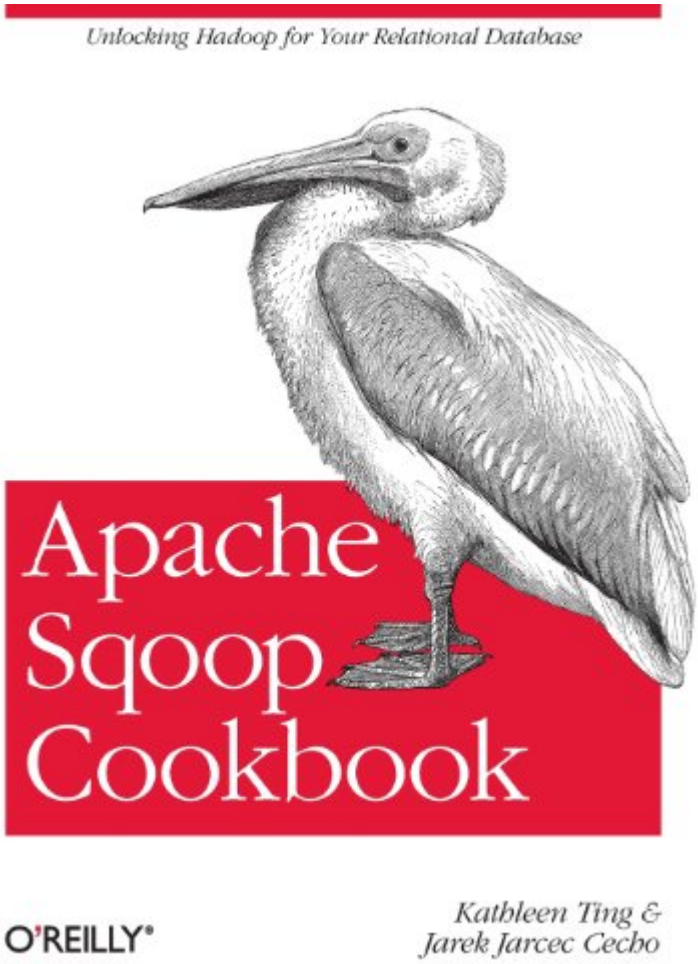
Books





More Books





cloudera
Ask Bigger Questions

Lets talk ETL

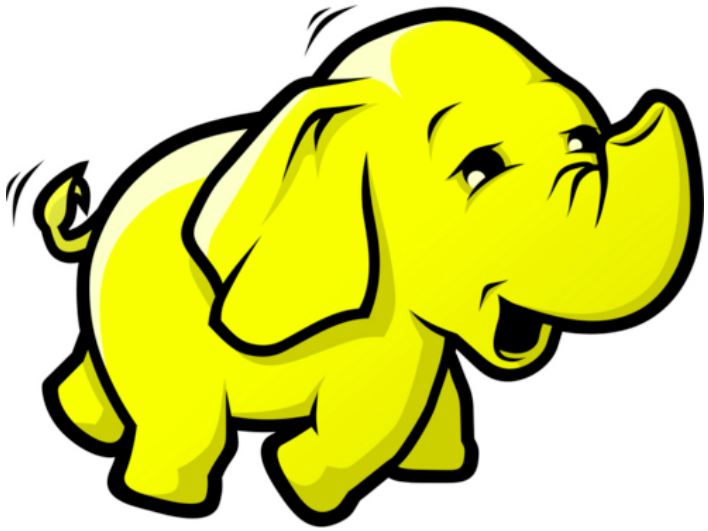
ETL is…

- Extracting data from outside sources
- Transforming it to fit operational needs
- Loading it into the end target

- (Wikipedia: http://en.wikipedia.org/wiki/Extract,_transform,_load)

Hadoop Is…

- HDFS – Massive, redundant data storage
- Map-Reduce – Batch oriented data processing at scale



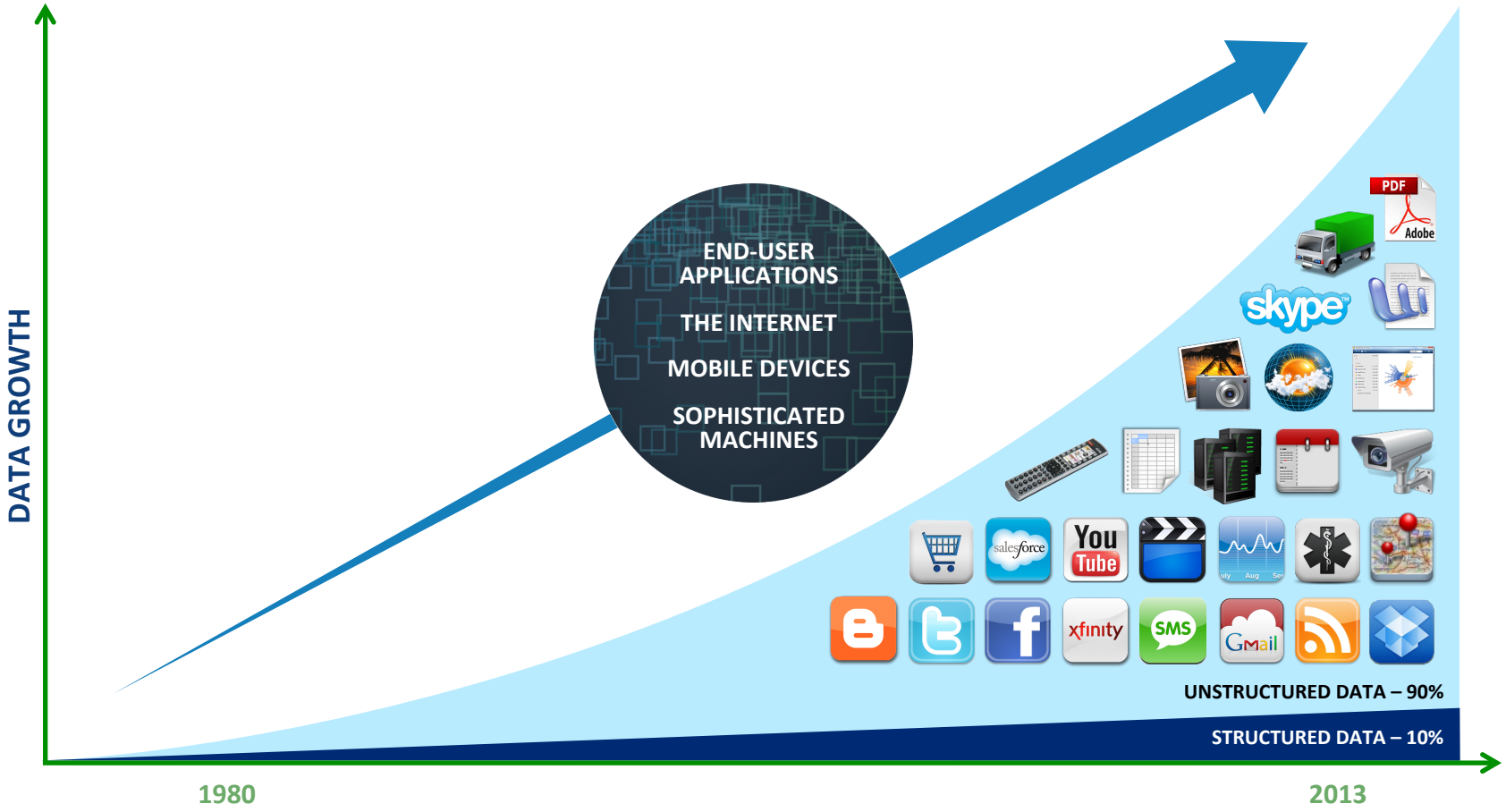
The Ecosystem

- High level languages and abstractions
- File, relational and streaming data integration
- Process Orchestration and Scheduling
- Libraries for data wrangling
- Low latency query language

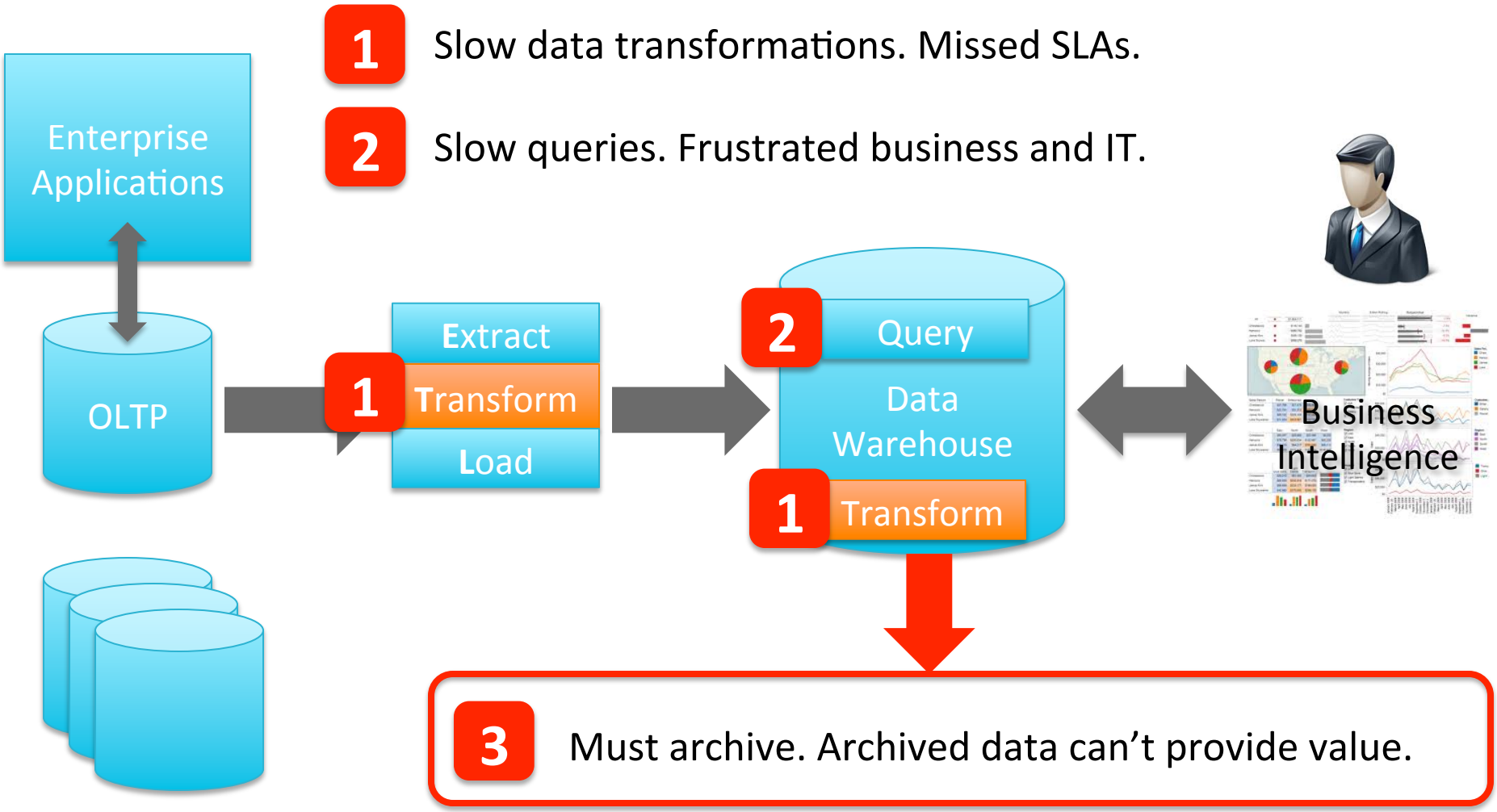
cloudera
Ask Bigger Questions

Why ETL with Hadoop?

Data Has Changed in the Last 30 Years



Volume, Variety, Velocity Cause Problems



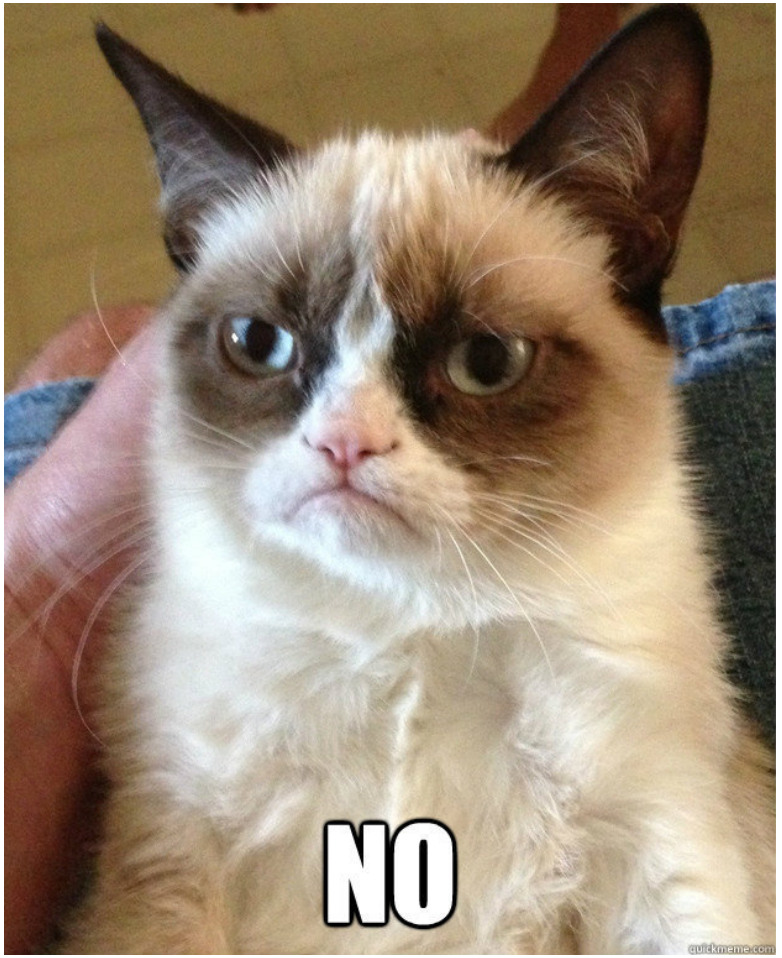
Got unstructured data?

- Traditional ETL:
  - Text
  - CSV
  - XLS
  - XML

- Hadoop:
  - HTML
  - XML, RSS
  - JSON
  - Apache Logs
  - Avro, ProtoBuffs, ORC, Parquet
  - Compression
  - Office, OpenDocument, iWorks
  - PDF, Epup, RTF
  - Midi, MP3
  - JPEG, Tiff
  - Java Classes
  - Mbox, RFC822
  - Autocad
  - TrueType Parser
  - HFD / NetCDF

Replace ETL Clusters

- Cheaper
- MUCH more flexible
- Faster?
- More scalable?
- You can have both

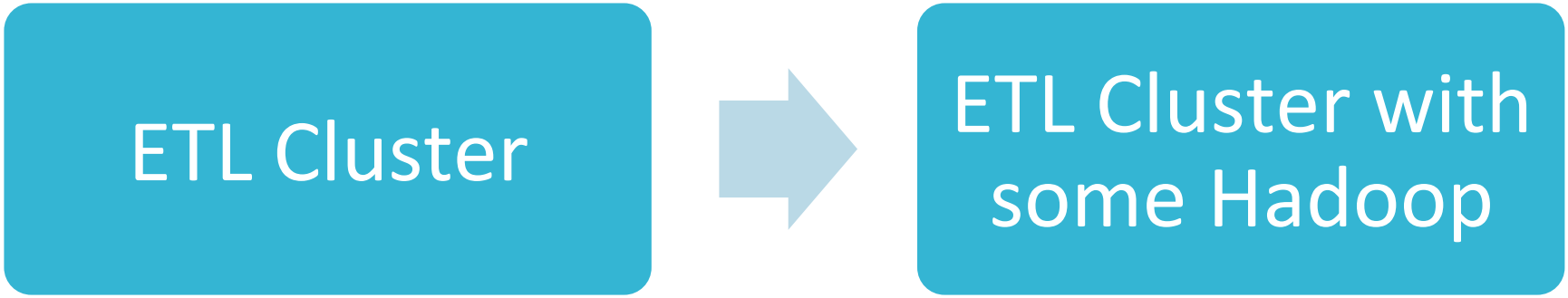


Data Warehouse Offloading

- Reduce storage costs
- Release CPU capacity
- Scale
- on the cheap
- Better tools

What I often see

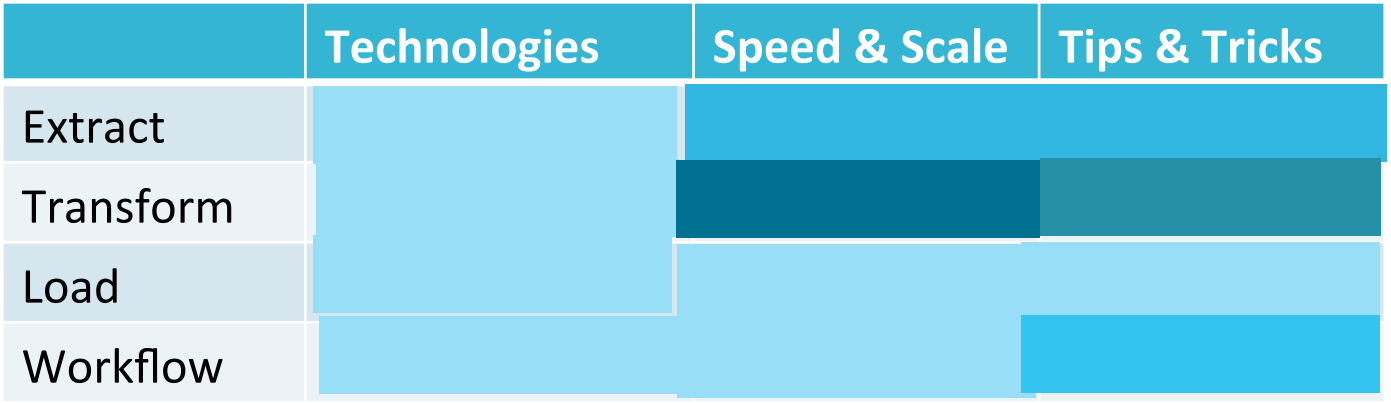
OR



Moving your transformations from the DWH to Hadoop?

Lets do it right.

We'll Discuss:



cloudera®
Ask Bigger Questions

Extract

Let me count the ways

1. From Databases: **Sqoop**
2. Log Data: **Flume + CDK**
3. Copy data to HDFS

Sqoop – The Balancing Act

Scale Sqoop Slowly

- Balance between:
  - Maximizing **network** utilization
  - Minimizing **database** impact
- Start with smallish table (1-10G)
- 1 mapper, 2 mappers, 4 mappers
- Where's the bottleneck?
- vmtat, iostat, mpstat, netstat, iptraf

When Loading Files:

Same principles apply:

- Parallel Copy
- Add Parallelism
- Find Bottlenecks
- Resolve them
- Avoid Self-DDOS

Scaling Sqoop

- Split column - match index or **partitions**
- Compression
- Direct drivers
- Incremental import

OraOOP

- Connection Manager by Quest/Cloudera
- Free! Open Source!
- https://github.com/QuestSoftwareTCD/OracleSQOOPconnector
- Lots of optimizations:
  - Block-wise or Partition-wise
  - Avoids full table scans where possible
  - Disable parallel full-table scans
  - No Logging, parallel direct path writes
- Limitations:
  - Does not support “incremental” jobs

Ingest Tips

- Use file system tricks to ensure consistency
- Directory structure:

  /intent

  /category

  /application (optional)

  /dataset

  /partitions

  /files

- Examples:

  /data/fraud/txs/2011-01-01/20110101-00.avro

  /group/research/model-17/training-tx/part-00000.txt

  /user/gshapira/scratch/surge/

Ingest Tips

- External tables in Hive
- Keep raw data
- Trigger workflows on file arrival

cloudera
Ask Bigger Questions

Transform

Endless Possibilites

- Map Reduce
  (in any language)
- **Hive (i.e. SQL)**
- Pig
- R
- Shell scripts
- Plain old Java

Prototype

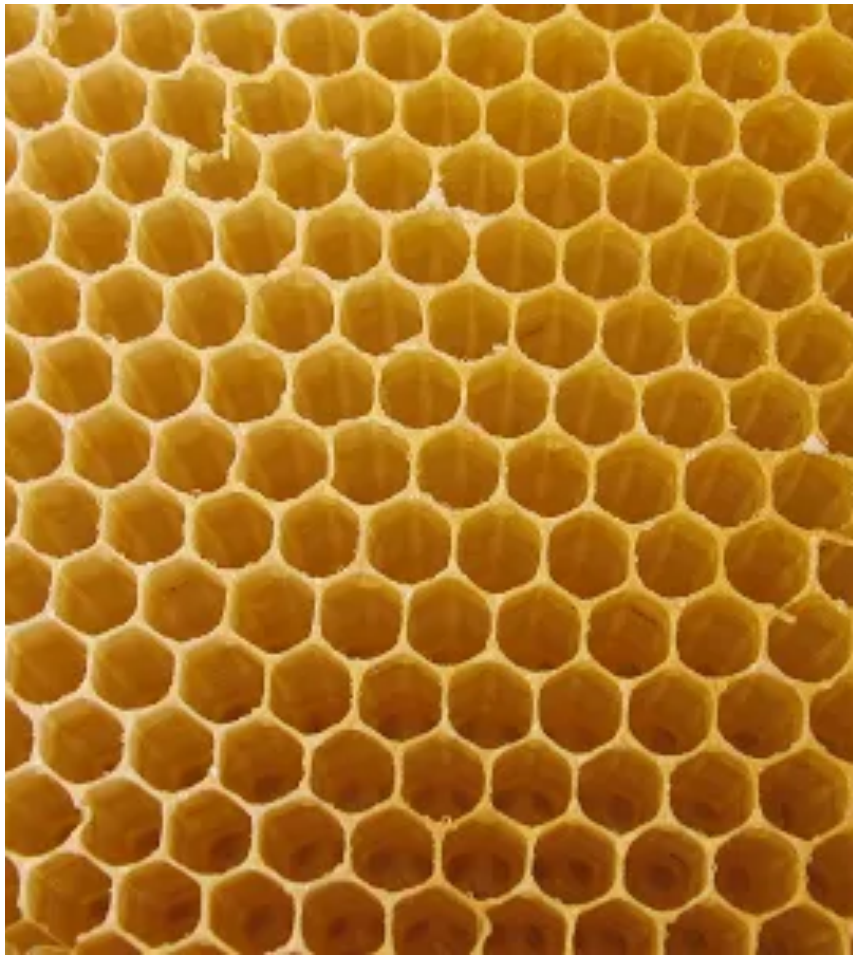
Parallelism –Unit of Work

- Amdahl's Law
- Small Units
- That stay small

- One user?
- One day?
- Ten square meter?

Partitioning

- Hive
- Directory Structure
- Pre-filter
- Adds metadata

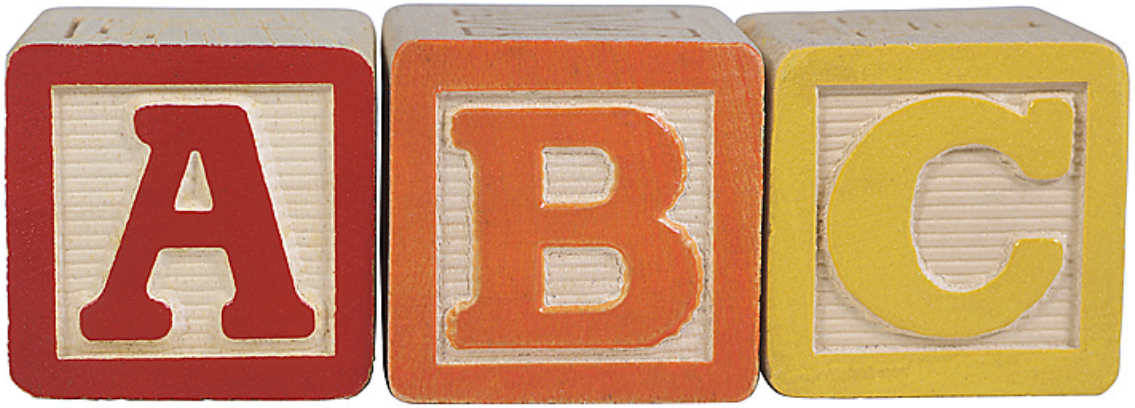


Tune Data Structures

- Joins are expensive
- Disk space is not
- De-normalize
- Store same data in multiple formats

Remember the Basics

- X reduce output is 3X disk IO and 2X network IO
- Less jobs = Less reduces = Less IO = Faster and Scalier
- Know your network and disk throughput
- Have rough idea of ops-per-second



Instrumentation

- Optimize the right things
- Right jobs
- Right hardware
- 90% of the time –
  its not the hardware

Fault and Rebuild

- Tier 0 – raw data
- Tier 1 – cleaned data
- Tier 2 – transformations, lookups and denormalization
- Tier 3 - Aggregations

Few words about Real Time ETL

- What does it even mean?
- Fast reporting?
- No delay from OLTP to DWH?
- Micro-batches make more sense:
  - Aggregation
  - Economy of scale
- Late data happens
- Near-line solutions

Map-Reduce

- Assembly language of data processing
- Simple things are hard, hard things are possible
- Use for:
  - Optimization: Do in one MR job what Hive does in 3
  - Optimization: Partition the data just right
  - GeoSpatial
  - Mahout – Map/Reduce machine learning

cloudera®
Ask Bigger Questions

Load

Technologies

- Sqoop
- Fuse-DFS
- Oracle Connectors
- NoSQLs

Oracle Connectors

- SQL Connector for Hadoop
- Oracle Loader for Hadoop
- ODI with Hadoop
- OBIEE with Hadoop
- R connector for Hadoop

You don’t need BDA

Oracle Loader for Hadoop

- Kinda like SQL Loader
- Data is on HDFS
- Runs as Map-Reduce job
- Partitions, sorts, converts format to Oracle Blocks
- Appended to database tables
- Or written to Data Pump files for later load

Oracle SQL Connector for HDFS

- Data is in HDFS
- Connector creates external table
- That automatically matches Hadoop data
- Control degree of parallelism

- You know External Tables, right?

How not to Load

- Most Hadoop customers don't load data in bulk
- History can stay in Hadoop
- Load only aggregated data
- Or computation results – recommendations, reports.
- Most queries can run in Hadoop
- BI tools often run in Hadoop

cloudera
Ask Bigger Questions

Workflow Management

Tools

- Oozie
- Azkaban

} Native Hadoop

- Pentaho Kettle
- TalenD

} Kinda Open Source

- Informatica

Scaling Challenges

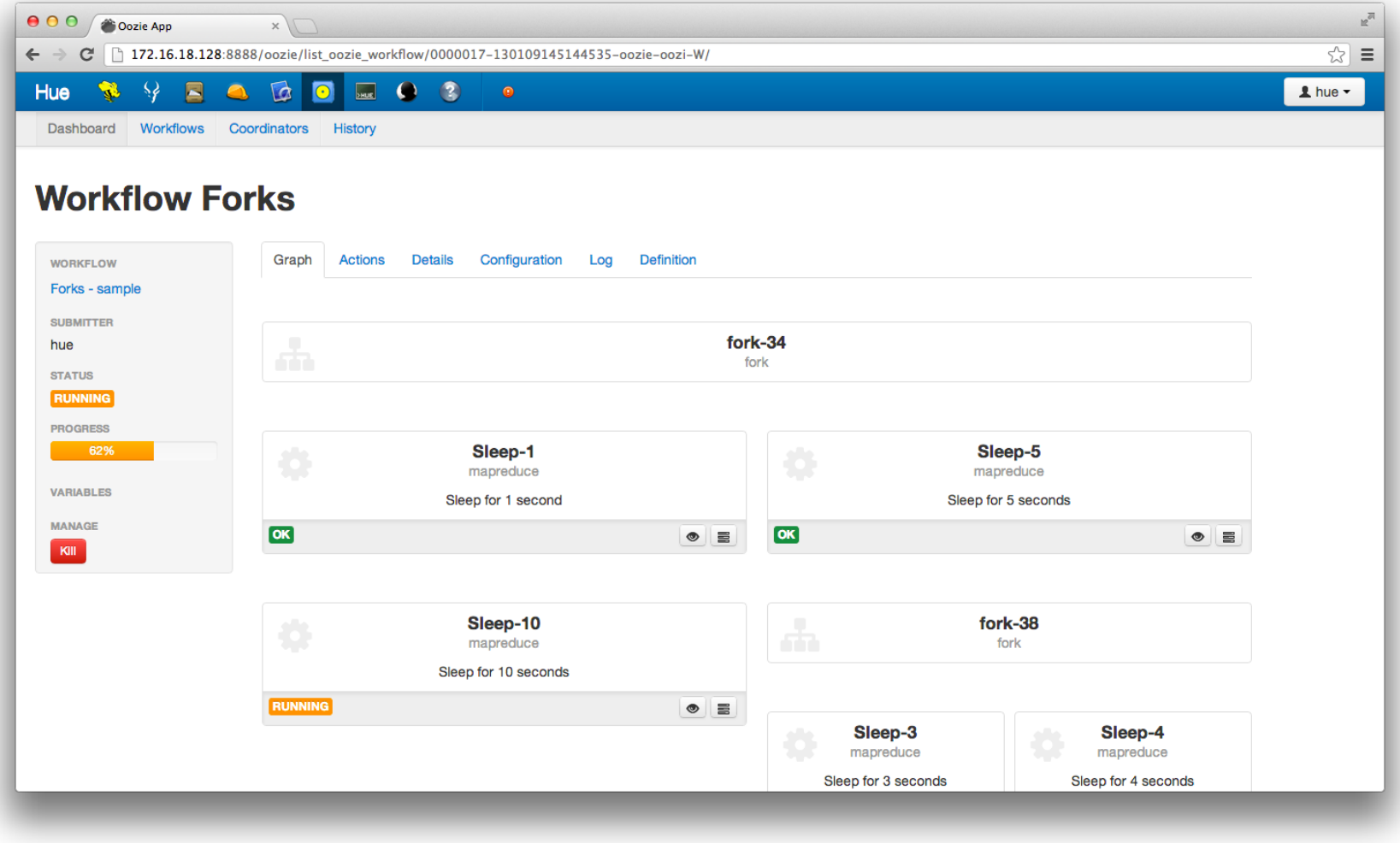
- Keeping track of:
  - Code Components
  - Metadata
  - Integrations and Adapters
  - Reports, results, artifacts
- Scheduling and Orchestration
- Cohesive System View
- Life Cycle
- Instrumentation, Measurement and Monitoring

My Toolbox

- Hue + Oozie:
  - Scheduling + Orchestration
  - Cohesive system view
  - Process repository
  - Some metadata
  - Some instrumentation
- Cloudera Manager for monitoring

- … and way too many home grown scripts

Hue + Oozie



> Writing a workflow engine is the software engineering equivalent of getting involved in a land war in Asia.

— Josh Wills


cloudera®
Ask Bigger Questions