cloudera

Hadoop as the Platform for the Smartgrid at TVA

August 26, 2010

Topics

- Introduction
- Retrospective on the openPDC project
- What Is Hadoop?
- Current Smartgrid Obstacles
- Cloudera Enterprise as The New Smartgrid Platform
- Summary

Today's speaker – Josh Patterson

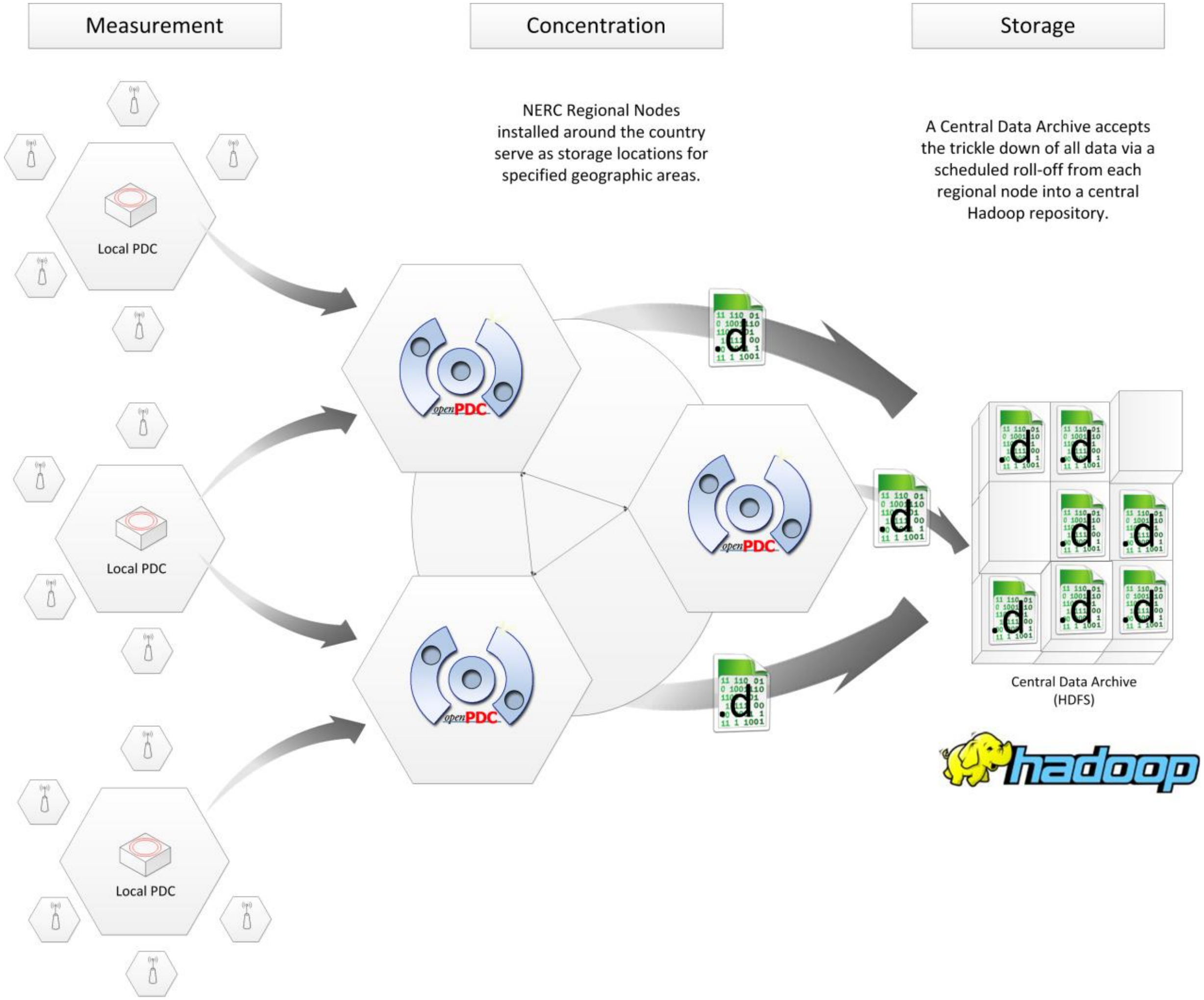
- **josh@cloudera.com**
- Master's Thesis: self-organizing mesh networks
  - Published in IAAI-09: TinyTermite: A Secure Routing Algorithm
- Conceived, built, and led Hadoop integration for the openPDC project at TVA
  - Led small team which designed classification techniques for timeseries and Map Reduce
  - Open source work at http://openpdc.codeplex.com
- Now: Solutions Architect at Cloudera

cloudera

What is the openPDC?

- The openPDC is a complete set of applications for processing streaming time-series data in real-time
  - Measured data is gathered with GPS-time from multiple input sources, time-sorted and provided to user defined actions, dispersed to custom output destinations for archival
- NERC funded
- Started at the Tennessee Valley Authority (TVA)
- Now in use by many government controlled power companies around the world

openPDC Topology



NERC Regional Nodes installed around the country serve as storage locations for specified geographic areas.

A Central Data Archive accepts the trickle down of all data via a scheduled roll-off from each regional node into a central Hadoop repository.

openPDC: Why?

Northeast Blackout of 2003

- Significant failure of US power grid in 2003 due to cascading effects
- SCADA provided a limited at best view of what happened
- NERC mandated that companies collect high resolution data and store for later analysis
- Power grid in US is aging rapidly, cost of needed overhaul is significant

How "Big Data" Challenged the openPDC Project

"We Need More Power, Scotty"

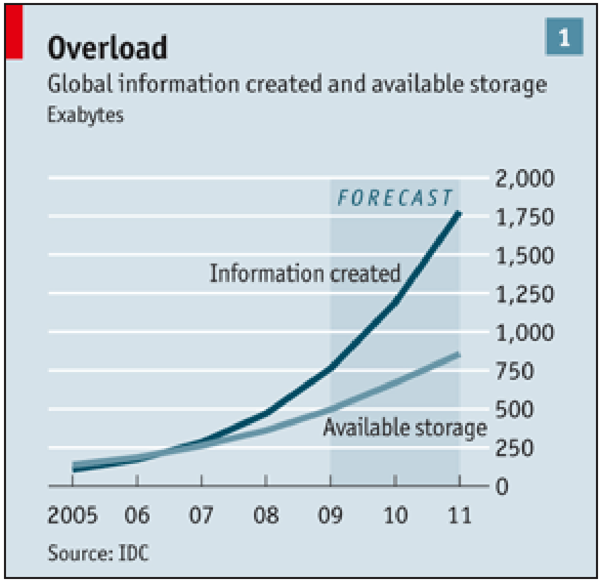
- Data was sampled 30 times a second
- Number of sensors (Phasor Measurement Units / PMU) was increasing rapidly (was 120, heading towards **1000** over next 2 years, currently taking in **4.2 billion** samples per day)
- Cost of SAN storage became excessive
- Little analysis possible on SAN due to poor read rates on large amounts (TBs) of data

Major Themes for Storage and Processing Needs

- Scale Out, not Up
- Linear scalability in cost and processing power
- Robust in the face of hardware failure
- No vendor lock in

Storage Needs: The Data Deluge

- At 1000 PMU sensors we were looking at needing to store 500TB of data
- The Data Deluge
  - “Eighteen months ago, Li & Fung, a firm that manages supply chains for retailers, saw 100 gigabytes of information flow through its network each day. Now the amount has increased tenfold.”
  - http://www.economist.com/opinion/displaystory.cfm?story_id=15579717
- Internet of Things
  - *HP's Peter Hartwell: "one trillion nanoscale sensors and actuators will need the equivalent of 1000 internets: the next huge demand for computing!“*



Source: IDC

cloudera

Processing Needs: Needle in a Haystack

- The “Haystack” in PMU data typically involved in scanning through TBs of info to find the one particular event we were interested in
- RDBMs simply do not work with high resolution timeseries data
- Need for Ad-Hoc processing on data to explore network effects and look at how events cascade across the grid

The Solution: Hadoop

- A scalable fault-tolerant distributed system for data storage and processing (open source under the Apache license)

- Two primary components
  - Hadoop Distributed File System (HDFS): self-healing high-bandwidth clustered storage
  - MapReduce: fault-tolerant distributed processing

- Key value
  - **Flexible ->** store data without a schema and add it later as needed
  - **Affordable ->** cost / TB at a fraction of traditional options
  - **Broadly adopted ->** a large and active ecosystem
  - **Proven at scale ->** dozens of petabyte + implementations in production today

HDFS As Cheap and Scalable Storage

- HDFS is robust in the face of machine failure
- A big thing was cost – we could linearly grow our cluster as needed by just adding new machines
- Ran on commodity hardware – we didn't have to buy expensive (and relatively slow), proprietary SAN setups

MapReduce Provides a Powerful Parallel Processing Framework

- We found Map Reduce to be the perfect framework to quickly process large amounts of PMU (timeseries) data
- Created a machine learning algorithm in Map Reduce which detected “unbounded oscillations” in grid data
- Map Reduce based oscillation scan of a few TBs takes minutes
- A scan of comparable data from a SAN would take days or weeks

What is common across Hadoop-able problems?

Nature of the data

- Complex data
- Multiple data sources
- Lots of it

Nature of the analysis

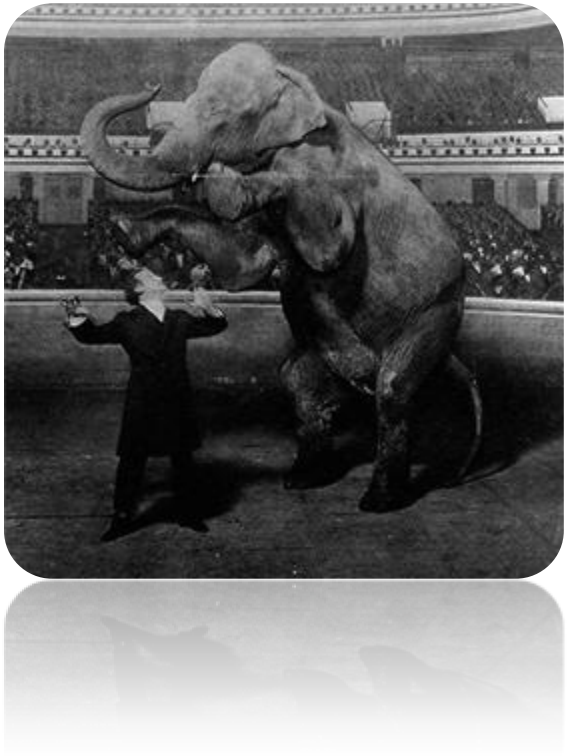
- Batch processing
- Parallel execution
- Spread data over a cluster of servers and take the computation to the data

What Analysis is Possible With Hadoop?

- Text mining
- Index building
- Graph creation and analysis
- Pattern recognition
- Collaborative filtering
- Prediction models
- Sentiment analysis
- Risk assessment

Benefits of Analyzing With Hadoop

- Previously impossible/impractical to do this analysis
- Analysis conducted at lower cost
- Analysis conducted in less time
- Greater flexibility



The Storm of the Data Deluge is Brewing

- Challenges of the openPDC project were just the first wave
- Storage requirements are accelerating
- Disk speeds are relatively constant
- Seeing signs of data deluge, ***GE now using open sourced Hadoop-based timeseries classifiers developed in the openPDC project***

Coming Power Grid Stressors

- Larger fluctuations in power demands
  - Ex: *Millions of new electric cars all charging in the evenings*
- An aging power grid that requires more capital infusion than most companies have allocated for these purposes
  - Grid infrastructure is older than most realize
  - Maintenance policies generally only look at age of equipment

The Power Grid Domain is Slow to Evolve

- Power companies are slow to adopt technology
  - They generally have poor maps of their overall infrastructure
- Coming pressures are going to force power companies to have to analyze TBs and PBs of data
- Ad-Hoc analysis will be needed to explore the complex relationships in this data

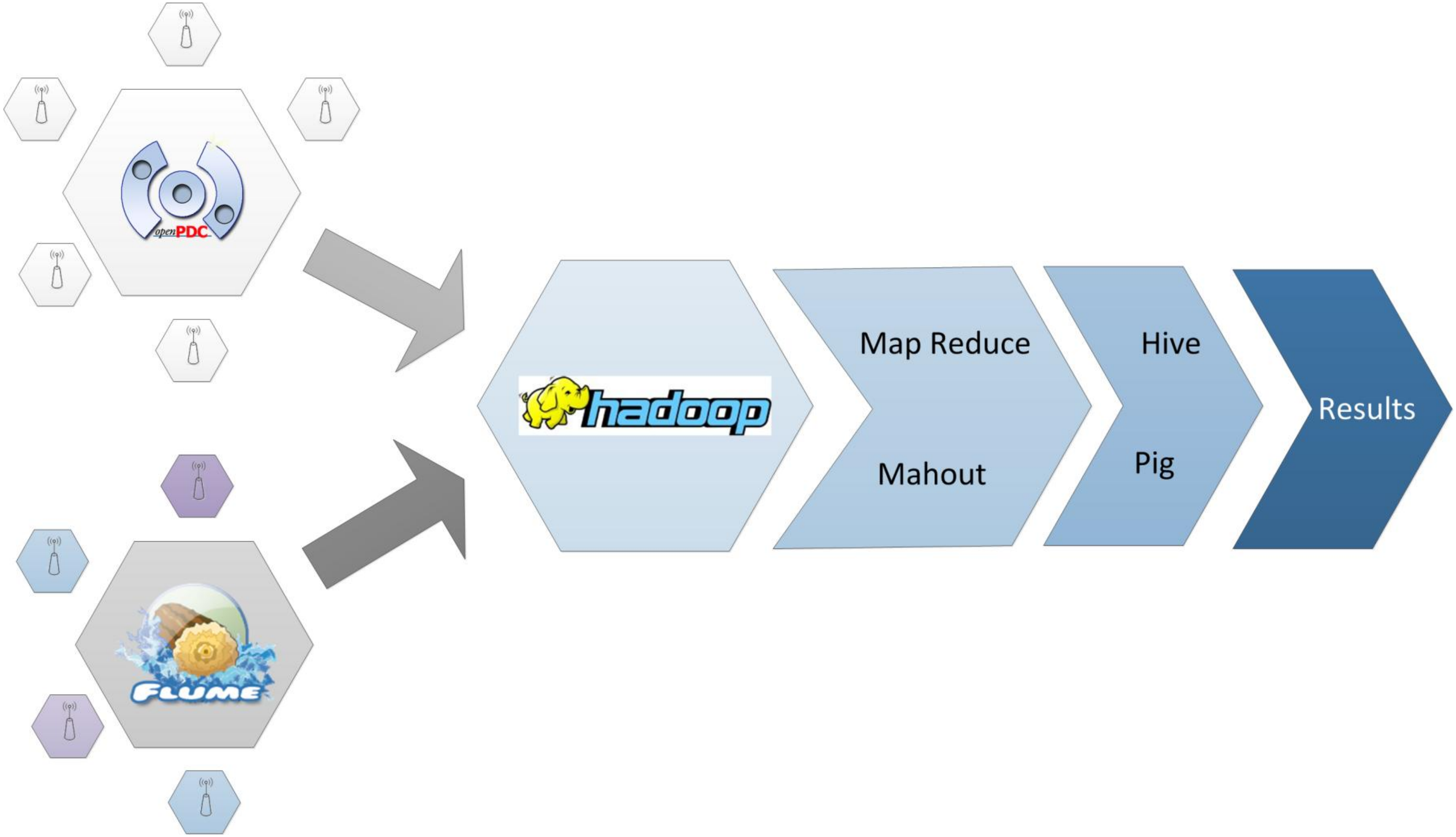
Broader Emerging Smartgrid Themes

- Simply adding lots of sensors is only a very small part of the solution
- Collection, storage, and processing are in themselves all difficult problems
- In order to build a more effective Smartgrid, platforms are needed that handle these things well
- Smartgrid sensor collection is a subset of the larger undercurrent of emerging massive sensor based systems

Even Broader Theme: Internet of Things

- We’re collecting sensor data everywhere, not just the Smartgrid
- Many of the techniques described above can be easily done with Hadoop
  - Open Source generalized collector system is called “Flume”
- Examples:
  - Weather sensors
  - Mesh networks – battlefield UAVs
  - Cell Phones – Google Android as a collector

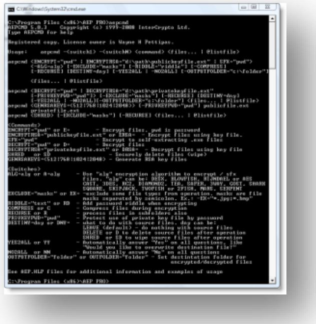
Next Generation Sensor Platform: Hadoop and Related Projects



The Companies That Provide Real Results for Sensor Platforms Will Win

- Much of today's Smartgrid talk is just hype
- Few "solutions" actually fix anything, only put sensors on things
- Analysis is where the true value lies
  - But you need a complete platform to be in position to analyze the data

Harnessing Hadoop Has Its Challenges



```
C:\Program Files (x86)\AEP PRO>aepcmd
AEPCMD 5.0.3    Copyright (c) 1999-2008 InterCrypto Ltd.
Type AEPCMD for help

Registered copy. License owner is Wayne H Pettigas.

Usage:    aepcmd -<switch1> -<switchN> <command> <files... | @listfile>

aepcmd <ENCRYPT="pwd" | ENCRYPTRSA="d:\path\publickeyfile.ext" | SFX="pwd">
       <-ALG=alg> [-EXCLUDE="masks"] [-RIDDLE="riddle"] [-COMPRESS]
       [-RECURSE] [DESTINY=dny] [-YES2ALL | -NO2ALL] [-OUTPUTFOLDER="c:\folder"]

       <files... | @listfile>

aepcmd <DECRYPT="pwd" | DECRYPTRSA="d:\path\privatekeyfile.ext"
       [-PRIVKEYPWD="pwd"]> [-EXCLUDE="masks"] [-RECURSE] [DESTINY=dny]
       [-YES2ALL | -NO2ALL][-OUTPUTFOLDER="c:\folder"] <files... | @listfile>
aepcmd <GENRSAKEYS=<512|768|1024|2048>> [-PRIVKEYPWD="pwd"] publicfile.ext
       privatefile.ext
aepcmd <SHRED> [-EXCLUDE="masks"] [-RECURSE] <files... | @listfile>

<Commands>
ENCRYPT="pwd" or E=         - Encrypt files. pwd is password
ENCRYPTRSA="publickeyfile.ext" or ERSA=  - Encrypt files using key file.
SFX="pwd"                   - Encrypt to self-extracting .exe files
DECRYPT="pwd" or D=         - Decrypt files
DECRYPTRSA="privatekeyfile.ext" or DRSA=  - Decrypt files using key file
SHRED   or SD                 - Securely delete files (wipe)
GENRSAKEYS=<512|768|1024|2048> - Generate RSA key files

<Switches>
ALG=alg or A=alg       - Use "alg" encryption algorithm to encrypt / sfx
                         files. "alg" can be: DESX, BLOWFISH, RIJNDAEL or AES
                         CAST, 3DES, RC2, DIAMOND2, TEA, SAFER, 3WAY, GOST, SHARK
                         SQUARE, SKIPJACK, TWOFISH or 2FISH, MARS, SERPENT
EXCLUDE="masks" or EX= -exclude some file types from operation. masks are file
                         masks separated by semicolon. Ex.: -EX="*.jpg;*.bmp"
RIDDLE="text" or RD    - Add password riddle when encrypting
COMPRESS or C          - Compress files during encryption
RECURSE or R           - process files in subfolders also
PRIVKEYPWD="pwd"       - Protect use of private key file by password
DESTINY=dny or DNY=    - what to do with source files. dny can be:
                         LEAVE (default) - do nothing with source files
                         DELETE or D to delete source files after operation
                         SHRED  or SD to wipe source files after operation
YES2ALL or YY          - Automatically answer "Yes" on all questions, like
                         "Would you like to overwrite destination file?"
NO2ALL  or NN          - Automatically answer "No" on all questions
OUTPUTFOLDER="folder" or OUTFOLDER="folder" - Set destintation folder for
                                              encrypted/decrypted files

See AEP.HLP files for additional information and examples of usage

C:\Program Files (x86)\AEP PRO>
```

Ease of use – command line interface only; data import and access requires development skills

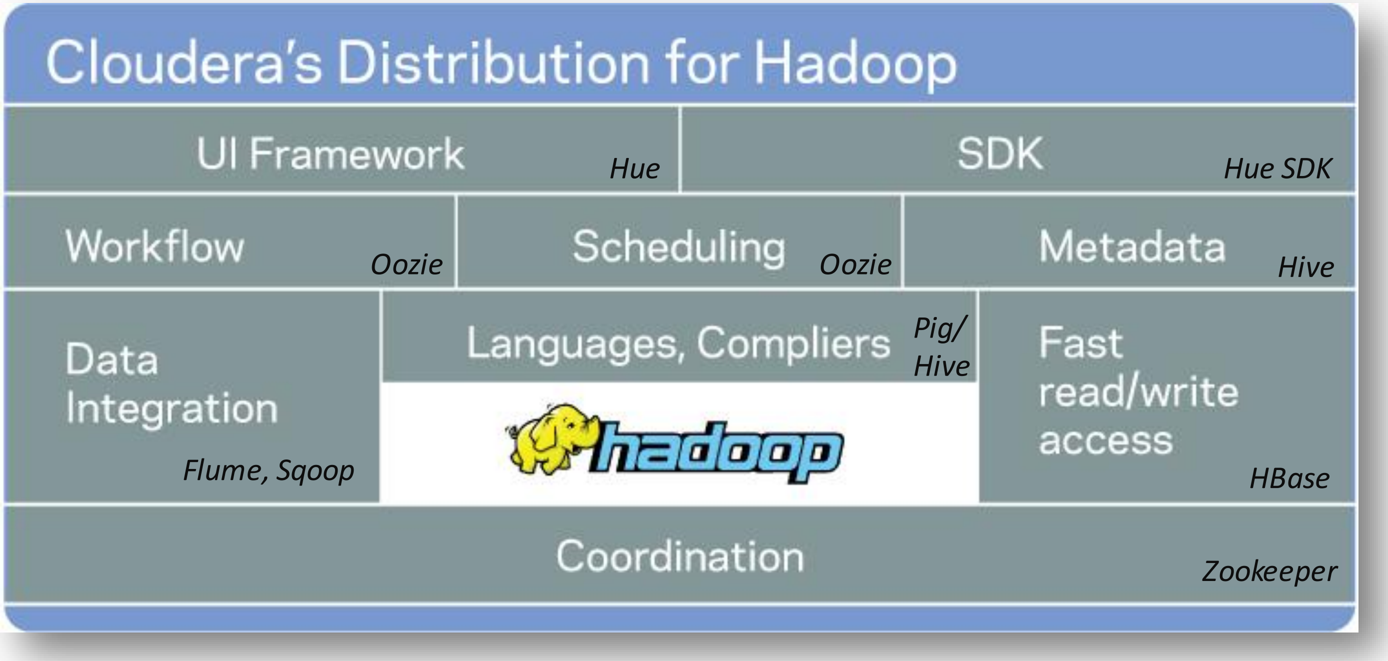
Complexity -- > 12 different components, different versions, dependencies and patch requirements

Manageability – Hadoop is challenging to configure, upgrade, monitor and administer

Interoperability – limited support for popular databases and analytical tools

Cloudera’s Distribution for Hadoop, version 3

The industry’s leading Hadoop distribution



- **Open source –** 100% Apache licensed
- **Simplified –** Component versions & dependencies managed for you
- **Integrated –** All components & functions interoperate through standard API’s
- **Reliable –** Patched with fixes from future releases to improve stability
- **Supported –** Employs project founders and committers for >70% of components

Who is Cloudera?

- Enterprise software & services company providing the industry’s leading Hadoop-based data management platform
  - Founding team came from large Web companies

Google facebook YAHOO!

- Products: Cloudera Enterprise & Cloudera’s Distribution for Hadoop
  - All necessary packages, matched, tested and supported
  - Tools to support production use of Hadoop
  - The leading distribution for the enterprise
- Contributors and committers
  - Fixing, patching and adding features

hadoop

Hear More Examples @ Hadoop World 2010

http://www.cloudera.com/company/press-center/hadoop-world-nyc/

- 2nd annual event focused on practical applications of Hadoop
- Date: October 12th 2010
- Location: Hilton New York
- Keynote from Tim O'Reilly – founder O'Reilly Media
- Pre and post conference training available for Hadoop and related projects
- 36 business and technical focused sessions

HADOOP WORLD
new york city

Confirmed speakers from

ORBITZ
AOL
YAHOO!
ebay
Bank of America
twitter
facebook

cloudera

Questions?