cloudera®

A [Incomplete] Data Tools Landscape [for Hackers] in 2015

Wes McKinney @wesmckinn

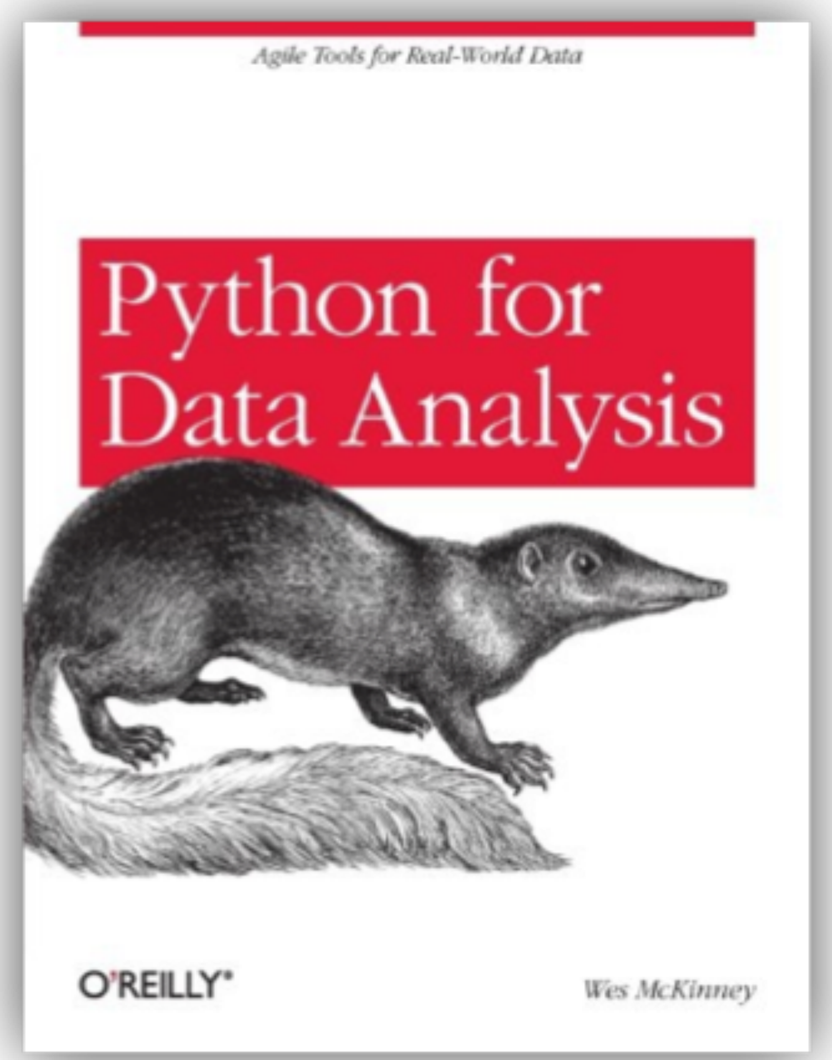
Data^3 Meeting — Minneapolis, MN

This talk

- A partial look at different languages and tools
- Limiting scope to either:
  - Permissively licensed open source software, e.g. Apache-licensed (OSS)
  - Non-dual-licensed copyleft OSS (e.g. GPL)
  - i.e. “do you [the community] have any incentive to create patches?”
- Some trends (that I see, anyway)
- Challenges and opportunities

Who am I?

- Python data firestarter
- Financial analytics in R / Python starting 2007
- pandas project born of frustration in 2008
- 2010-2012
  - Hiatus from gainful employment
  - Make pandas ready for primetime
  - Write "Python for Data Analysis"



Who am I? (cont’d)

- 2013-2014: Co-founder/CEO of DataPad (analytics startup, with early pandas collaborator Chang She)
- Late 2014: DataPad team joins Cloudera
- Now: backend systems and all-things-Python @ Cloudera

SQL: Still a lingua franca

- "SQL: the Fortran of Analytics"
- Often a concise, declarative way to express data transforms, analytics, etc.
- Relatively easy to parse, analyze
- SQL recently has seen resurgence with focus on interactive-speed SQL engines, especially on top of HDFS/Hadoop
- Relevant and impactful features (e.g. JSON support) still arriving in established RDBMS like PostgreSQL

Historical Python Context

- Scientific / HPC computing focus in 1990s, 2000s
  - Python web community developed in parallel, matured faster!
- NumPy became community standard in 2005, born from Numeric + Numarray
- Pyrex, later Cython, easier C / C++ wrapping
- f2py: easy Fortran wrapping
- Anaconda distribution
  - Finally solving Python deployment for all

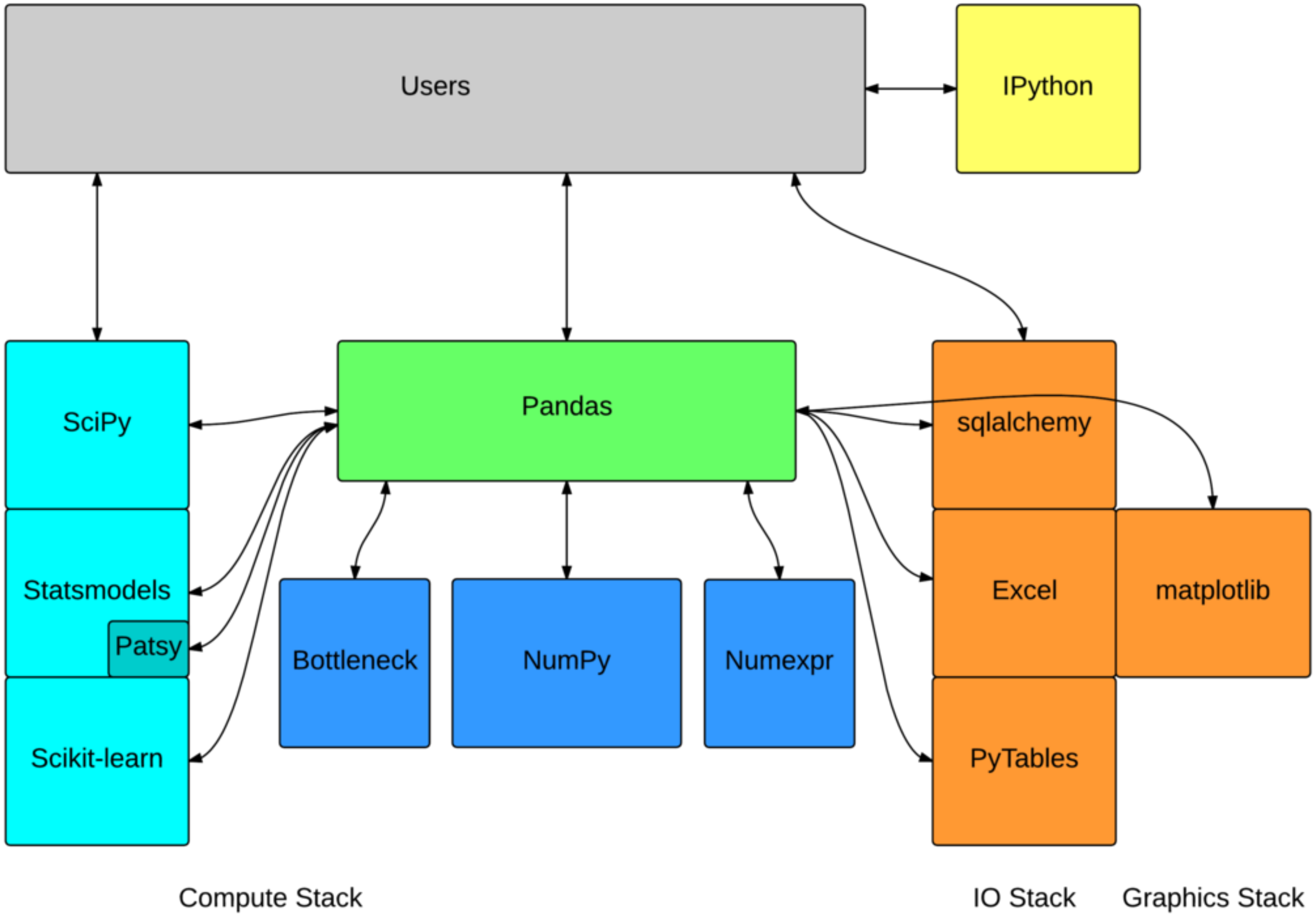
Essential Python stack

- NumPy: low-level array processing
- SciPy: essential computational algos
- pandas: data wrangling
- scikit-learn: machine learning
- matplotlib (+ add-ons, like seaborn): visualization
- numba: numeric hotspot LLVM compiler
- Domain-specific toolkits: nltk, scikit-image, statsmodels, Theano, PyCUDA/ PyOpenCL and many others

pandas

- A Pythonic take on the classic R “data frame” data structure
- Critical piece to make the Python stack useful in everyday work
- Added axis metadata / labeling for representing multidimensional data
- Focus on easy data wrangling, IO, plotting, and basic analytics

Jeff Reback's "pandas as PyData middleware" diagram



cloudera

Newer / Up-and-coming Python projects

- Bokeh: interactive / reactive visualization for the web
- Blaze: uniform data expression API
- Odo: easy data migration

R Project

- Trusted base of statistics libraries
  - Latest and greatest stats research often hits R first
- RStudio
- The "Hadley stack”
  - Visualization: ggplot2 (static) and ggvis (interactive)
  - Data Wrangling: dplyr
  - legacy: plyr / reshape2

dplyr

- Started late 2012 by Hadley Wickham, supported by RStudio
- Composable / chainable analytics and data wrangling expressions
- In-memory and SQL backends
- Has attracted folks back to R from Python in a lot of cases

Some other great R stuff

- shiny: interactive web apps in R
- Rcpp
- data.table
- xts

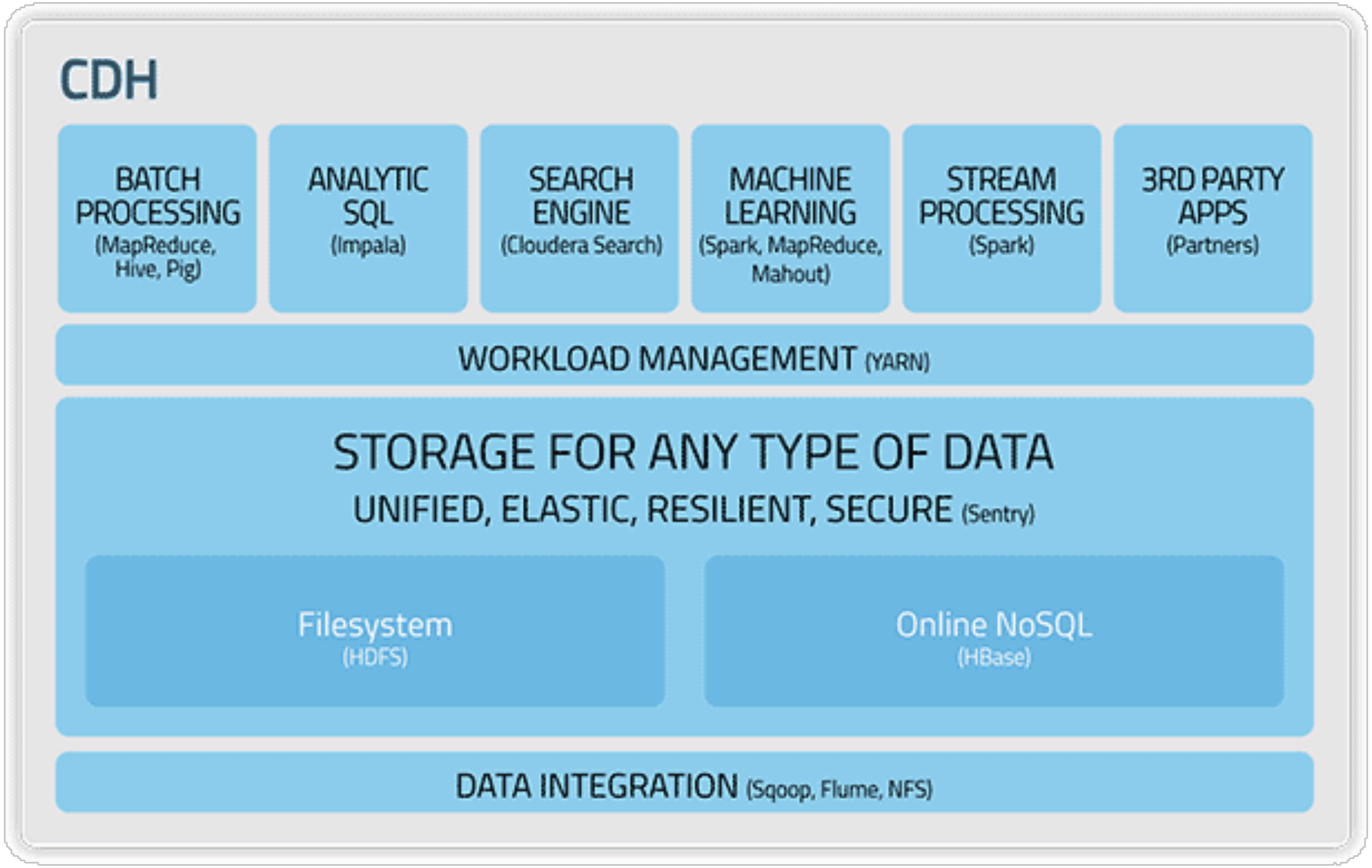
IPython

- IPython started out as a better interactive Python
- Grew to include web-based computational notebook, GUI console, and other components
  - (Google even integrated into Google Drive!)
- IPython Notebook architecture enabled "kernel" processes to be written in nearly any language (even bash!)
- How to build community beyond Python?

Enter Jupyter

- http://jupyter.org
- Breaking out notebook machinery into a standalone non-Python-specific project
- Enable project components to evolve at own pace, without large monolithic releases
- JupyterHub: upcoming multi-user notebook server

A few words about Hadoop + Big Data



CDH

BATCH PROCESSING (MapReduce, Hive, Pig)

ANALYTIC SQL (Impala)

SEARCH ENGINE (Cloudera Search)

MACHINE LEARNING (Spark, MapReduce, Mahout)

STREAM PROCESSING (Spark)

3RD PARTY APPS (Partners)

WORKLOAD MANAGEMENT (YARN)

STORAGE FOR ANY TYPE OF DATA

UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)

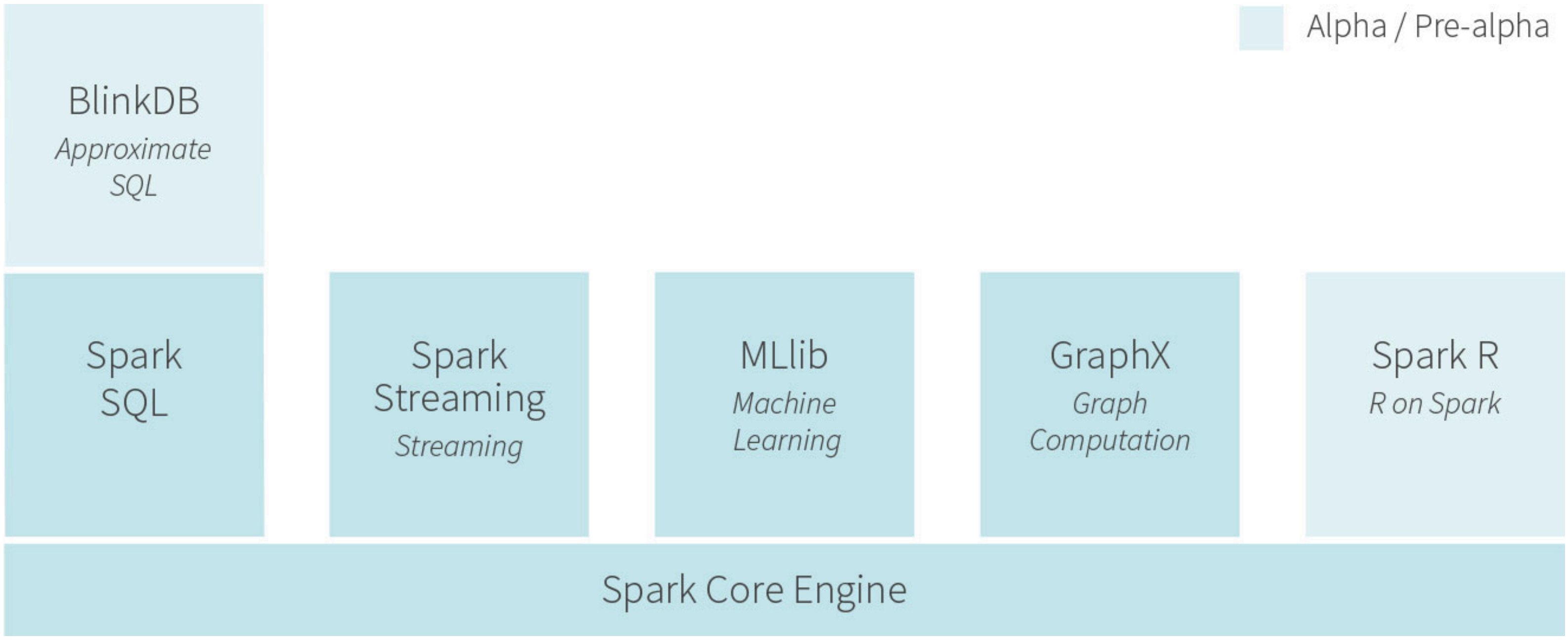
Filesystem (HDFS)

Online NoSQL (HBase)

DATA INTEGRATION (Sqoop, Flume, NFS)

Apache Spark

- Originated from Berkeley AMPLab
- General purpose distributed memory-centric data processing framework
- Official APIs: Scala, Java, Python



□ Alpha / Pre-alpha

BlinkDB
Approximate SQL

Spark SQL

Spark Streaming
Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark R
R on Spark

Spark Core Engine

Source: databricks.com

Spark 1.3: DataFrames!

- R/pandas-inspired API for tabular data manipulation in Scala, Python, etc.
- Logical operation graphs rewritten internally in more efficient form
- Good interop with Spark SQL
- Some interoperability with pandas
- Will help close the semantic gap between Spark and R/Python

Some problems in need of solving

- A Shiny-like quick-and-dirty data app development framework for Python
- IPython/Jupyter notebook collaboration
- A community-standard, Apache-licensed C/C++ data frame library with best-in-class performance
- Ubiquitous support for emerging analytical on-disk storage standards like Parquet

Other interesting stuff to look at

- Torch7 / LuaJIT: high performance ML / deep learning on GPUs
  - Facebook AI group open sourced several ML modules
- Apache Flink
  - Up-and-coming Scala-based data processing framework
  - Some overlap with Spark use cases

Some other interesting industry trends

- Microsoft
  - Acquired Revolution Analytics, leading commercial R vendor
  - Launched Azure ML: R, Python, and more on Azure cloud
- Dato (fka GraphLab)
  - faster, more scalable machine learning, with Python interface (Paid commercial product, free for non-commercial/academic use)
  - Largest-ever VC investment in a data tools company betting big on Python
- Databricks
  - Offering cloud Spark-notebook-as-a-service

cloudera

Thank you

@wesmckinn