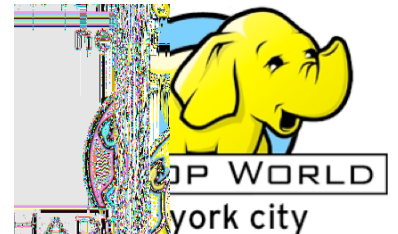


Next Steps for Hadoop

Doug Cutting
Cloudera



Proviso

- Linus Torvalds:
 - “Whatever they contribute.”
 - diverse set of contributors
 - central planning impossible



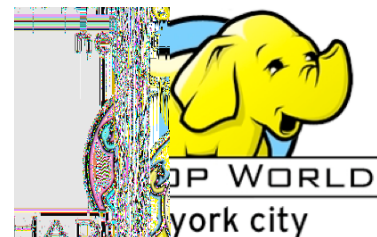
The Dream

- faster, more reliable, available
 - of course
- spreadsheet-like interfaces
 - provide non-programmers
 - with powerful, interactive tools
- easier sharing
 - of data & hardware resources



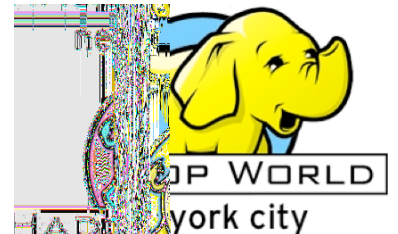
Requirements

- security
 - facilitate sharing of resources
- stable cross-language APIs
 - facilitate diverse tools & apps
- expressive, inter-operable data
 - facilitates sharing of datasets
 - facilitates dynamic analyses



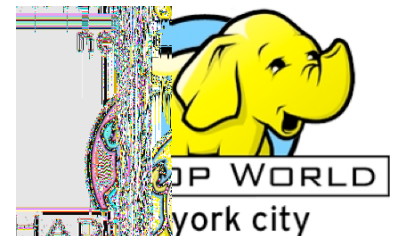
Data Formats

- today in Hadoop:
 - text
 - pro: inter-operable
 - con: not expressive, inefficient
 - - Java Writable
 - pro: expressive, efficient
 - con: platform-specific, fragile



Protocol Buffers & Thrift

- expressive
- efficient (small & fast)
- but not very dynamic
 - cannot browse arbitrary data
 - no DESCRIBE or SHOW
 - viewing a new dataset
 - requires code generation & load
 - writing a new dataset
 - requires generating schema text
 - plus code generation & load



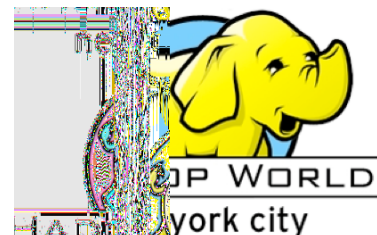
Avro Data

- as expressive
- smaller and faster
- dynamic
 - schema stored with data
 - but factored out of instances
 - API permits reading & creating
 - arbitrary datatypes
 - without generating & loading code



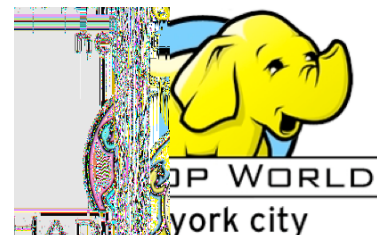
Avro Data

- includes a file format
- includes a textual encoding
- handles versioning
 - if schema changes
 - can still process data
- Hadoop apps can
 - upgrade from text
 - and standardize on Avro for data



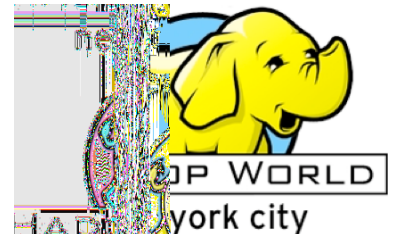
Avro RPC

- leverage versioning support
 - to permit different versions of services to interoperate
- for Hadoop services, will
 - provide cross-language access
 - let apps talk to clusters running different versions



Avro Status

- 1.1 release out
 - added JSON and comparators
- 1.2 soon
 - adds HTTP & UDP-based RPC
- will first appear in Hadoop 0.21
 - as format for job history
 - in sequence files



Avro Near Future

- full mapreduce support
- used for RPC in Hadoop 0.22 (1.0)?



Thanks!

What are your next steps?

