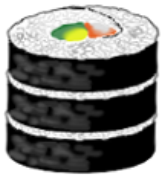

wibidata


UNIVERSITY OF
WASHINGTON



Data Science:
Big Data Beyond MapReduce

Aaron Kimball

Chief Architect – WibiData, Inc.

Models predict the future

- … about business (supply & demand, prices…)
- … about users or customers

(“Bob would like to watch this video next”)

=

Modeling the business

- Often includes classic BI or summary statistics
- Hadoop-based tools include
  - MapReduce
  - Pig
  - Impala

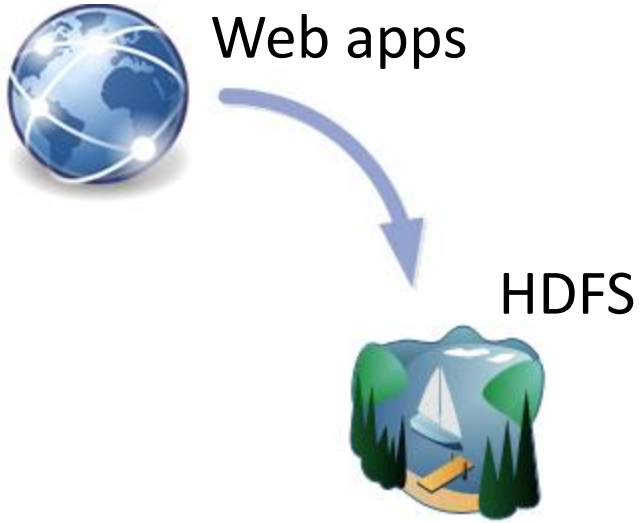
Modeling individual agents

- a.k.a. “Predictive analytics”
- Machine Learning algorithms
  - Classifiers, clustering, collaborative filtering…
- Often involves a heavy time-series component

= Big Data

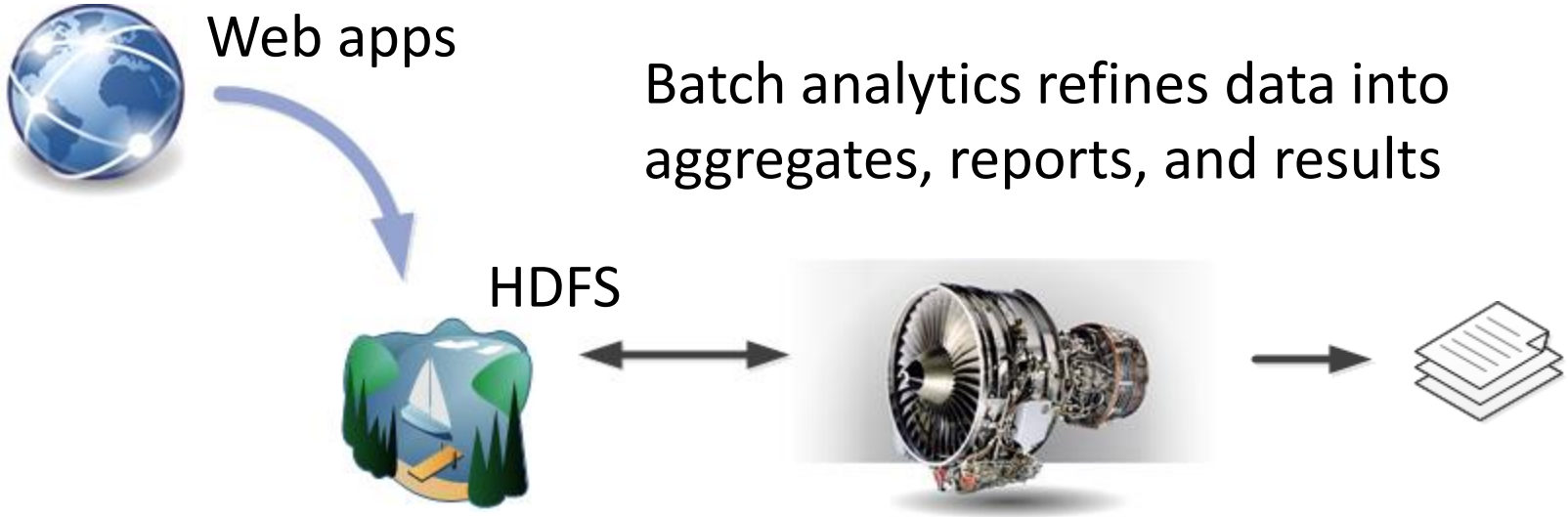
- Storing web-scale data requires HDFS
- HDFS + MapReduce are the foundation of web-scale batch analytics

HDFS: A “data lake”



…Logs, 3rd party or unstructured data

MapReduce: “data refinery”

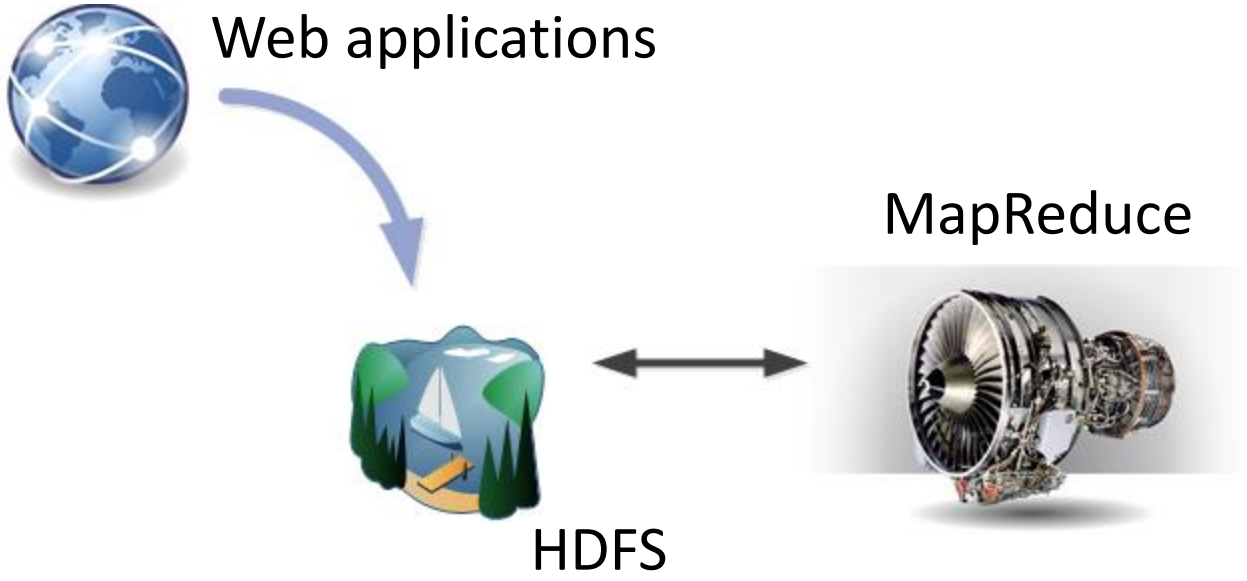


MapReduce: batch analytics

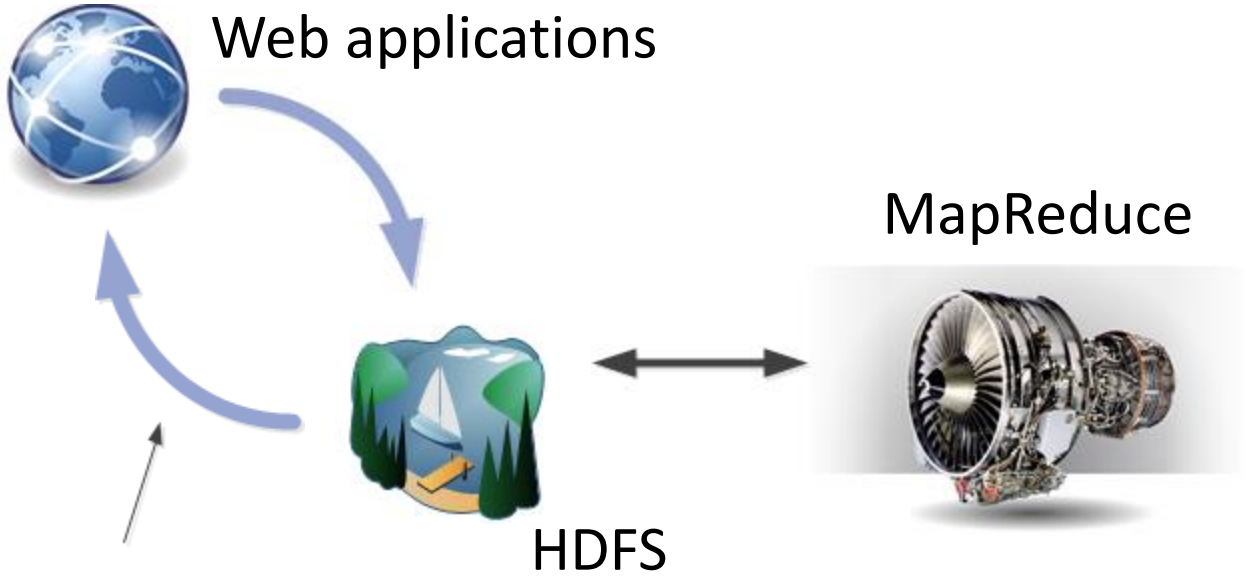
- Jobs process large amounts of data at once, infrequently (at most once per 30 minutes)
- Data set is usually immutable; cannot stream in new data or process incrementally
- Helpful for modeling the *business*

How do you use models of users?

Today's technology



Responding to users in real time



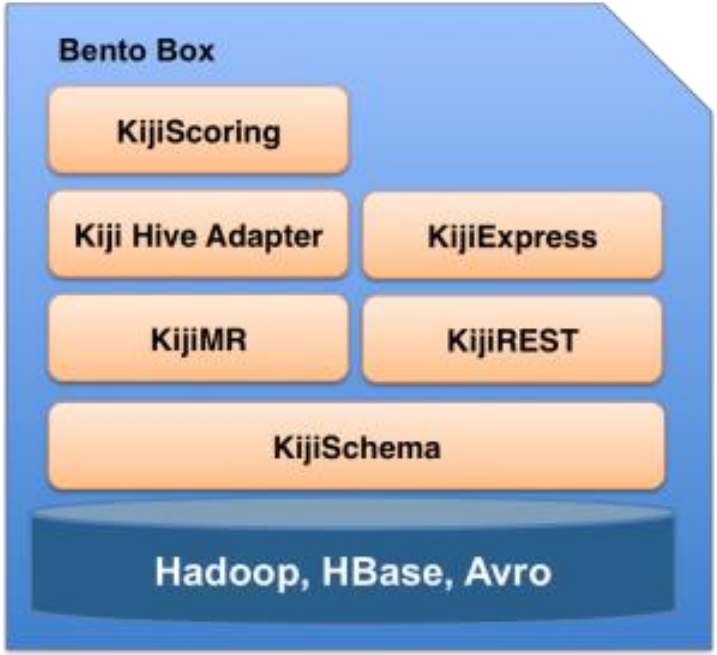
What provides this arrow?

Big Data Applications

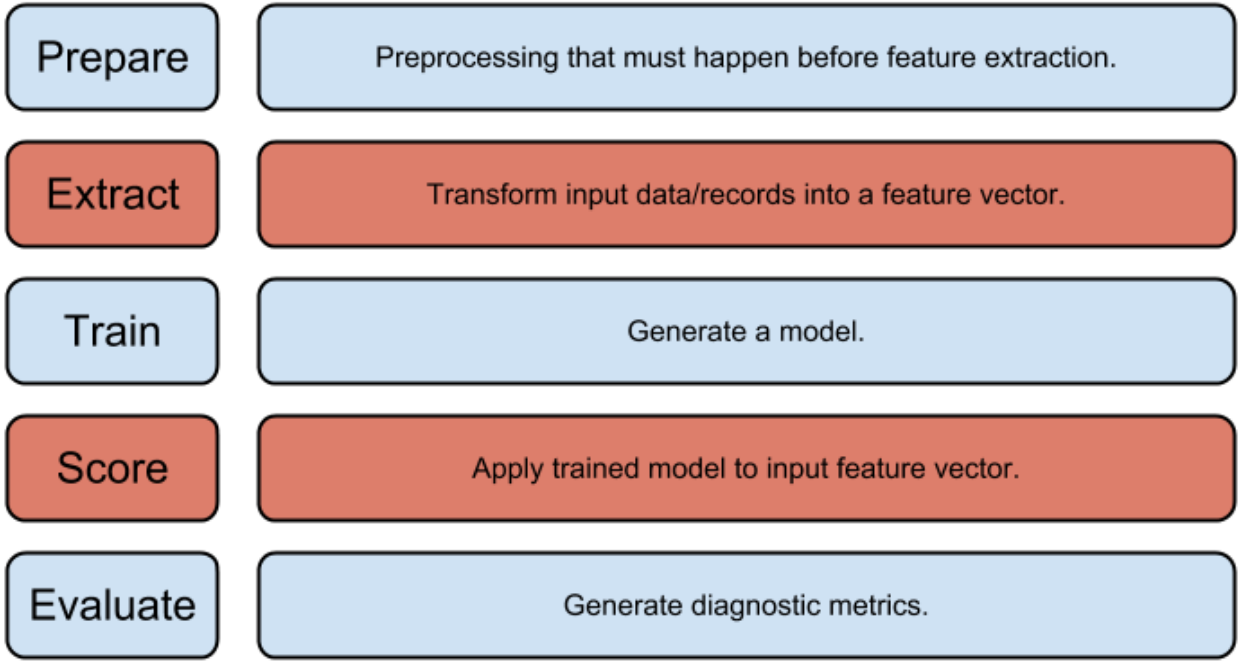
- Track all user behavior in the application
- Use MapReduce to train models
- Respond to users by interactively applying models to their real-time behavior

APACHE HBASE

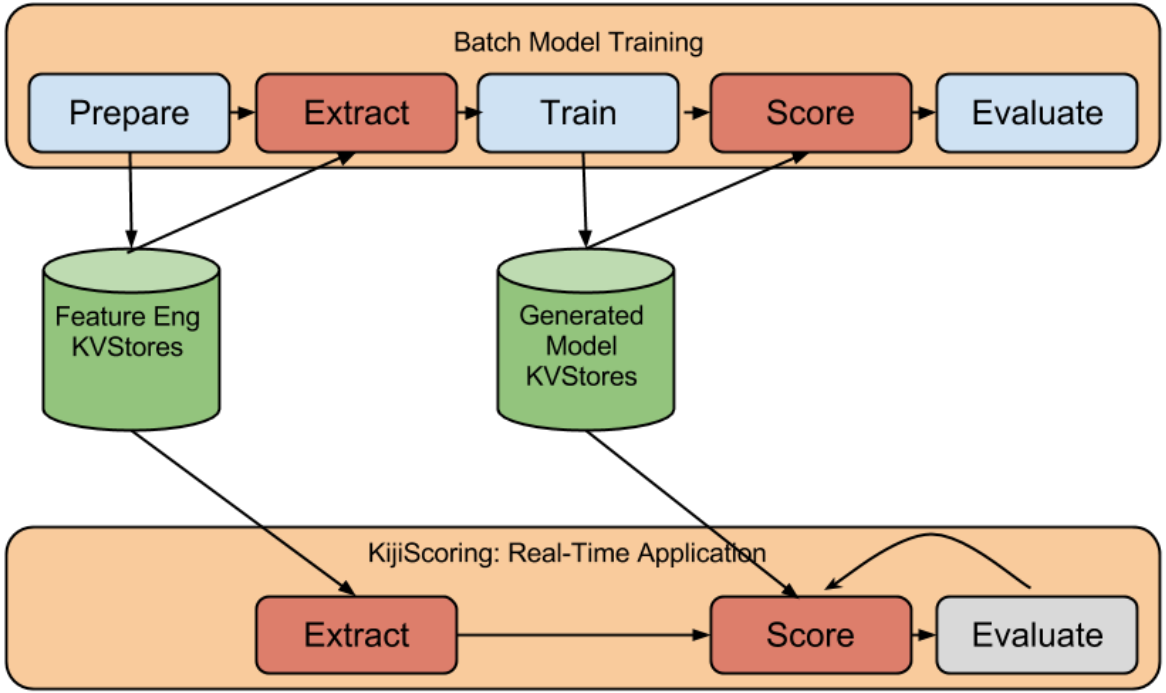
Kiji: A Big Data Application server

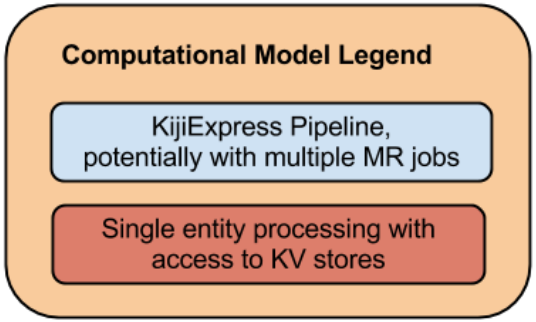


A data scientist workflow



Applying models in real time





Conclusions

- Big Data Applications combine web & mobile frontends, MapReduce batch analytics, and real-time model application
- Big Data Apps are personalized and powerful
- Data Science makes it happen

PS…

- Want to try Kiji? It's open source. See **kiji.org**

- Want to build the next generation of data science tools? See **jobs.wibidata.com**

www.kiji.org

Aaron Kimball – aaron@wibidata.com