



DEEP
LEARNING
INSTITUTE



DLI Accelerated Data Science Teaching Kit

Lecture 12.4 - Learn More About HBase



The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

HBase's “History”

Hadoop & HDFS based on...

- 2003 Google File System (GFS) paper
<https://research.google.com/archive/gfs.html>
- 2004 Google MapReduce paper
<https://research.google.com/archive/mapreduce.html>

← Designed for batch processing

HBase based on ...

- 2006 Google Bigtable paper
<https://research.google.com/archive/bigtable.html>

↑
Designed for random access

RDBMS vs HBase

How are they different?

- Hbase when you don't know the structure/schema
- HBase supports sparse data
 - many columns, values can be absent
- Relational databases good for getting “whole” rows
- HBase: keeps multiple versions of data
- RDBMS support multiple indices, minimize duplications
- Generally a lot cheaper to deploy HBase, for same size of data (petabytes)

More topics to learn about

Other ways to get, put, delete... (e.g., **programmatically** via Java)

- Doing them in **batch**

A lot more to read about cluster administration

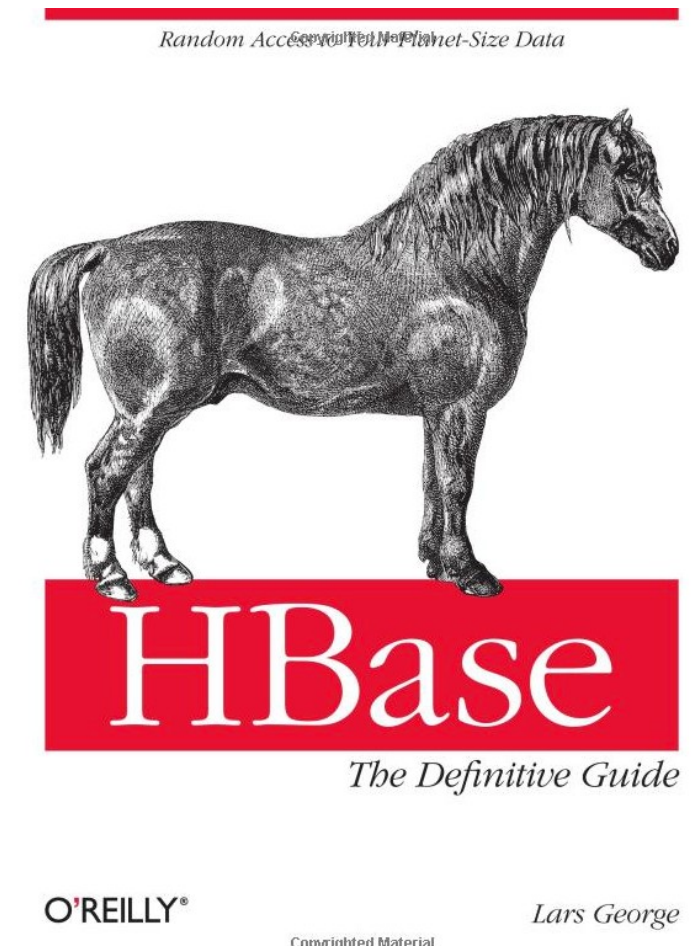
- **Configurations, specs** for master (name node) and workers (region servers)
- Monitoring cluster's health

“Bad key” design (<http://hbase.apache.org/book/rowkey.design.html>)

- **monotonically increasing keys** can decrease performance

Integrating with MapReduce

<https://db-engines.com/en/system/Cassandra%3BHBase>
Cassandra, etc.





DEEP
LEARNING
INSTITUTE



PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You