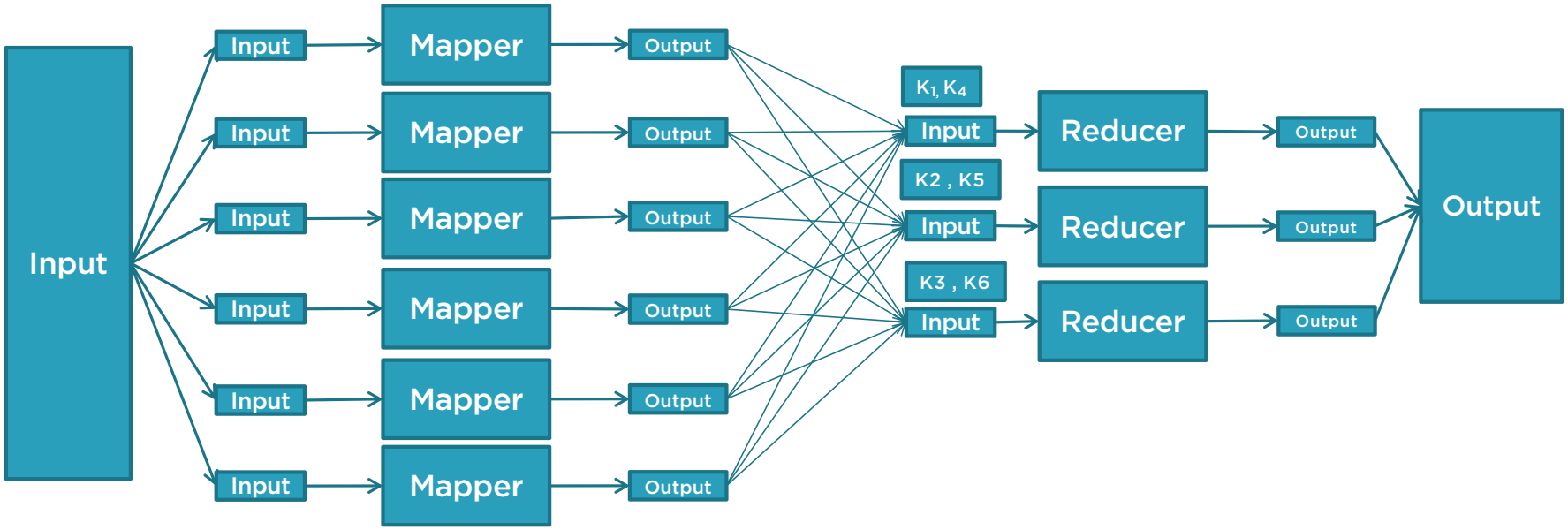
Batch Analytics with Elastic MapReduce (EMR)

Andrew Brust
FOUNDER & CEO, BLUE BADGE INSIGHTS

@andrewbrust www.bluebadgeinsights.com

MapReduce, in a Diagram



Open Source Analytics Technology

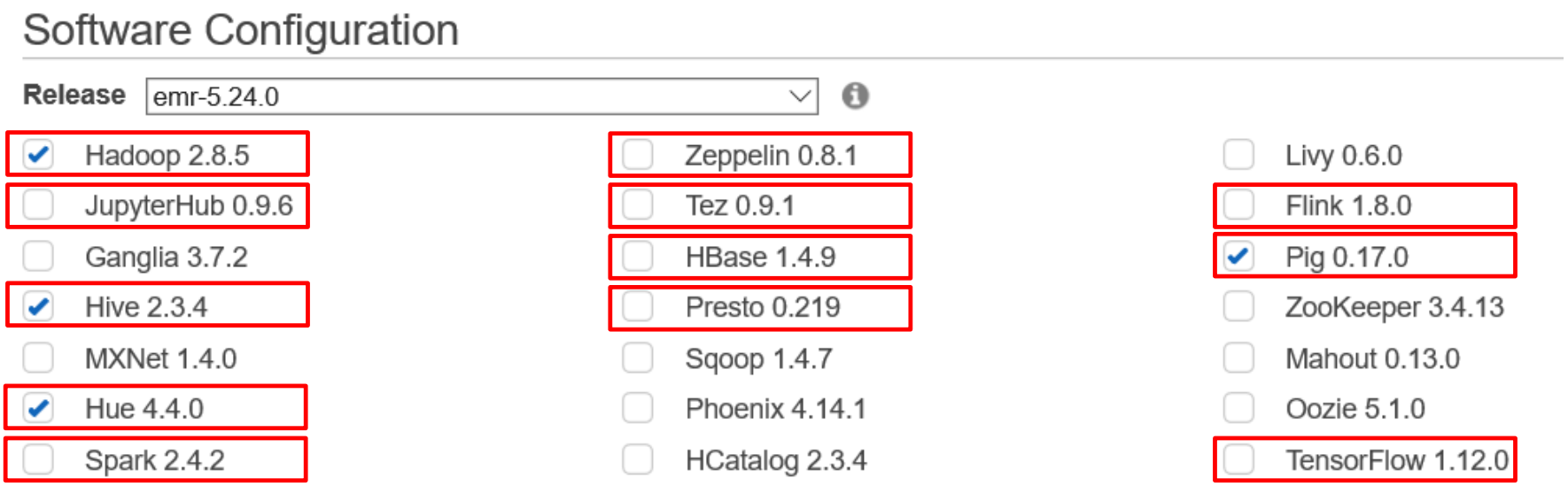
Largely defined by the Hadoop ecosystem

Major components:

- Hadoop and Tez
- Hive (various flavors)
- HBase
- Spark
- Presto

Key feature: many engines; one copy of the data (stored as files in object or distributed storage)

EMR Optional Components



Hue

Editors

- Hive, Syntax completion, Simple charting, Pig

Browsers

- Database Schema, HDFS, S3

File Viewer

Notebooks

Ways to Query with SQL

EMR

- Hive
- Spark SQL
- Presto

Athena

- (Also Presto)

Columnar File Formats

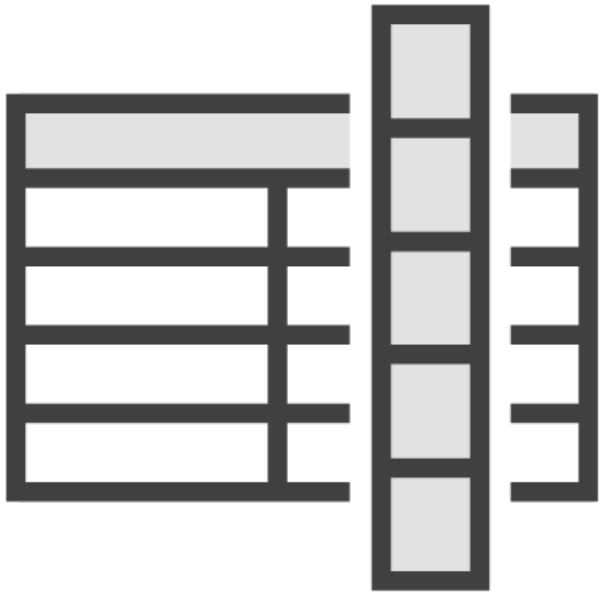
Parquet

ORC

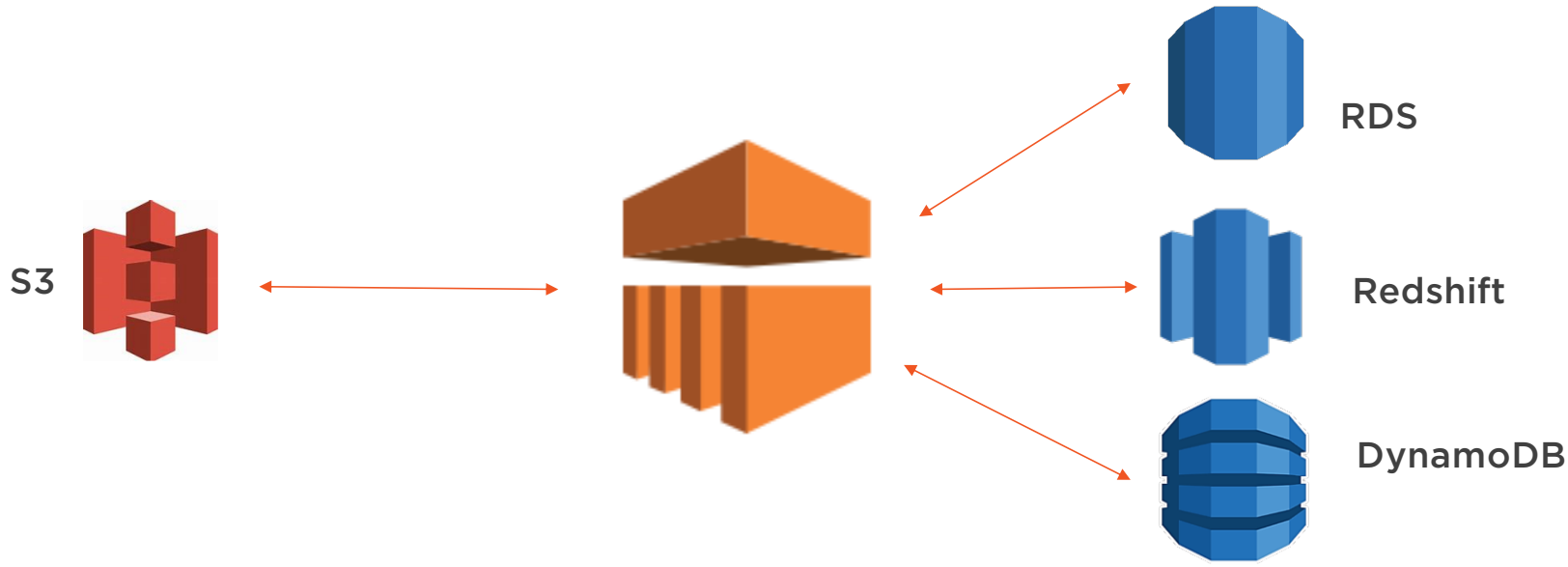
Agnostic formats, yet analytics-optimized

Economical too, due to compression

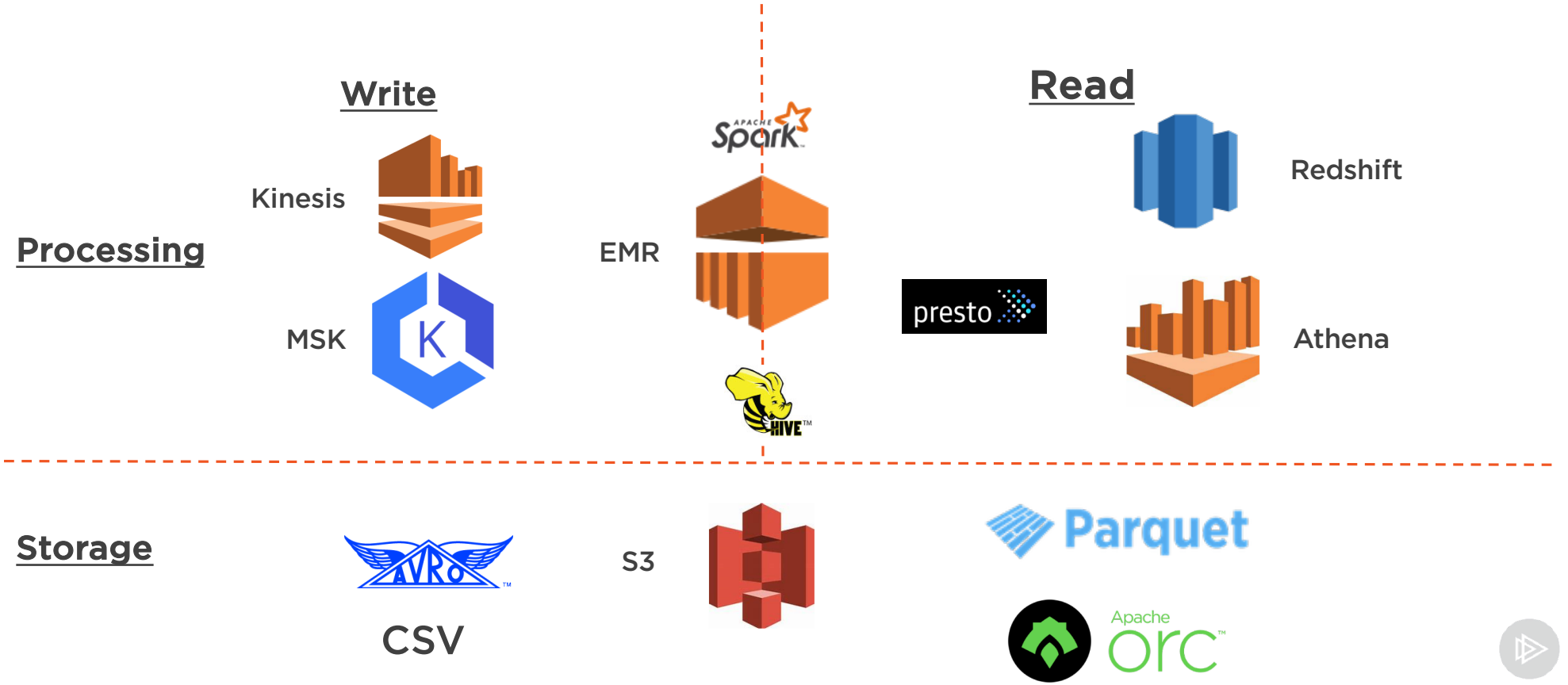
Core to data lakes



EMR: Connections



The Data Lake Stack



Demo

Apache Hue

- MapReduce jobs
- Pig Latin scripts
- Query
  - HBase
  - Hive
  - Spark SQL

Hive query from command line