

# Batch Analytics with Elastic MapReduce (EMR)

---



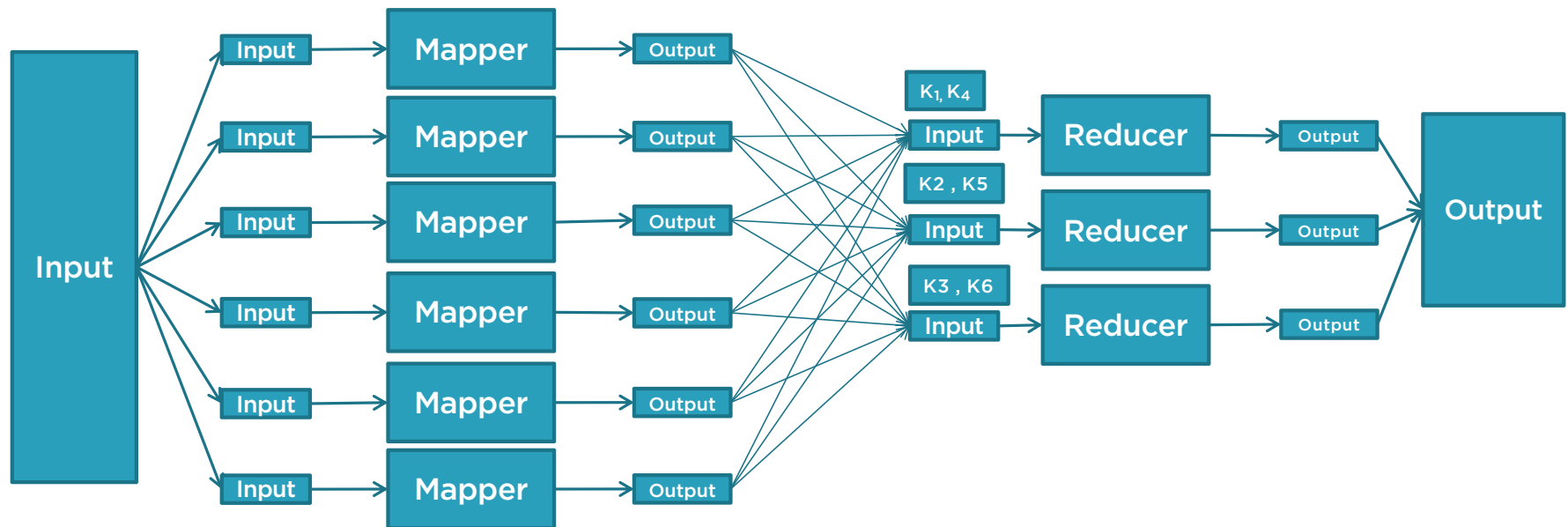
**Andrew Brust**

FOUNDER & CEO, BLUE BADGE INSIGHTS

@andrewbrust [www.bluebadgeinsights.com](http://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

## Software Configuration

Release  

- ☒ Hadoop 2.8.5
- ☐ JupyterHub 0.9.6
- ☐ Ganglia 3.7.2
- ☒ Hive 2.3.4
- ☐ MXNet 1.4.0
- ☒ Hue 4.4.0
- ☐ Spark 2.4.2

- ☐ Zeppelin 0.8.1
- ☐ Tez 0.9.1
- ☐ HBase 1.4.9
- ☐ Presto 0.219
- ☐ Sqoop 1.4.7
- ☐ Phoenix 4.14.1
- ☐ HCatalog 2.3.4

- ☐ Livy 0.6.0
- ☐ Flink 1.8.0
- ☒ Pig 0.17.0
- ☐ ZooKeeper 3.4.13
- ☐ Mahout 0.13.0
- ☐ Oozie 5.1.0
- ☐ TensorFlow 1.12.0



# Hue



## Editors

- Hive, Syntax completion, Simple charting, Pig

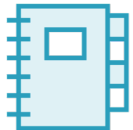


## Browsers

- Database Schema, HDFS, S3



## File Viewer



## Notebooks

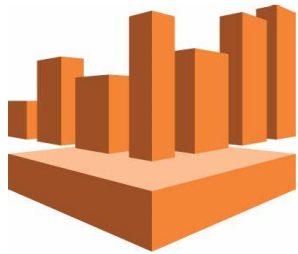


# Ways to Query with SQL



## EMR

- Hive
- Spark SQL
- Presto



## Athena

- (Also Presto)



# Columnar File Formats

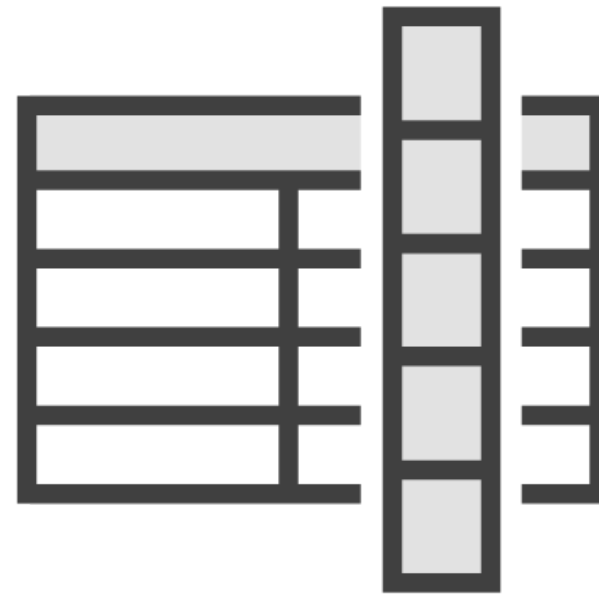
Parquet

ORC

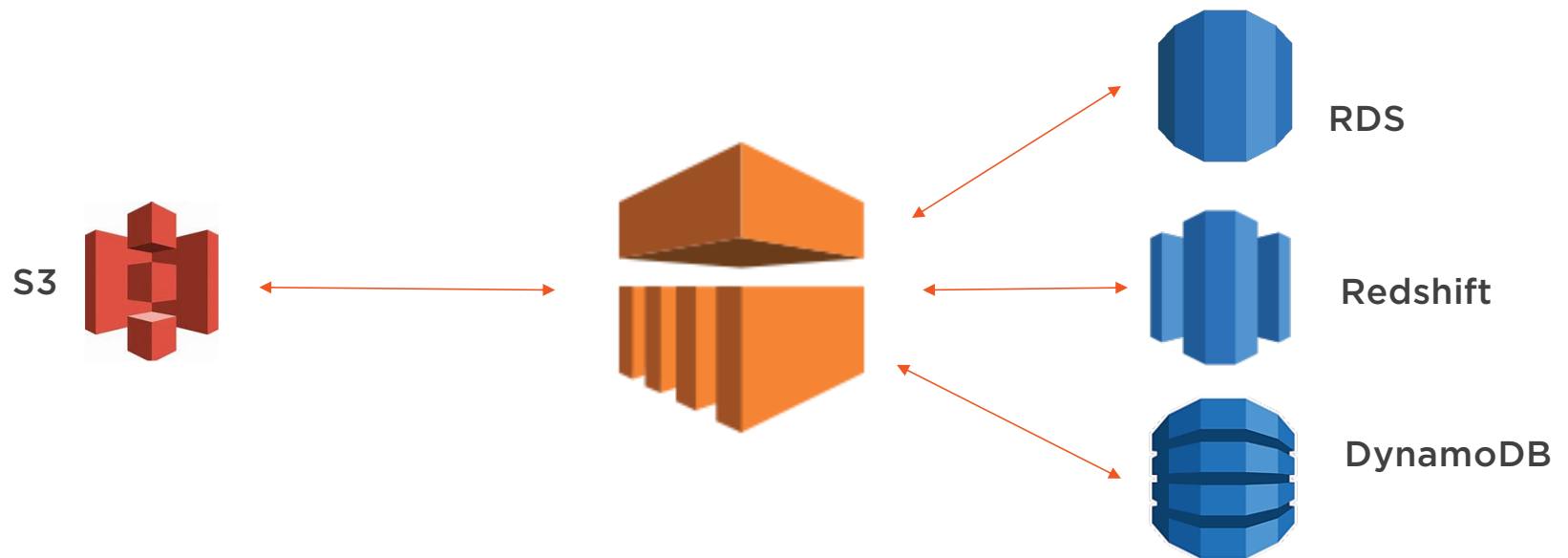
Agnostic formats, yet  
analytics-optimized

Economical too, due  
to compression

Core to data lakes

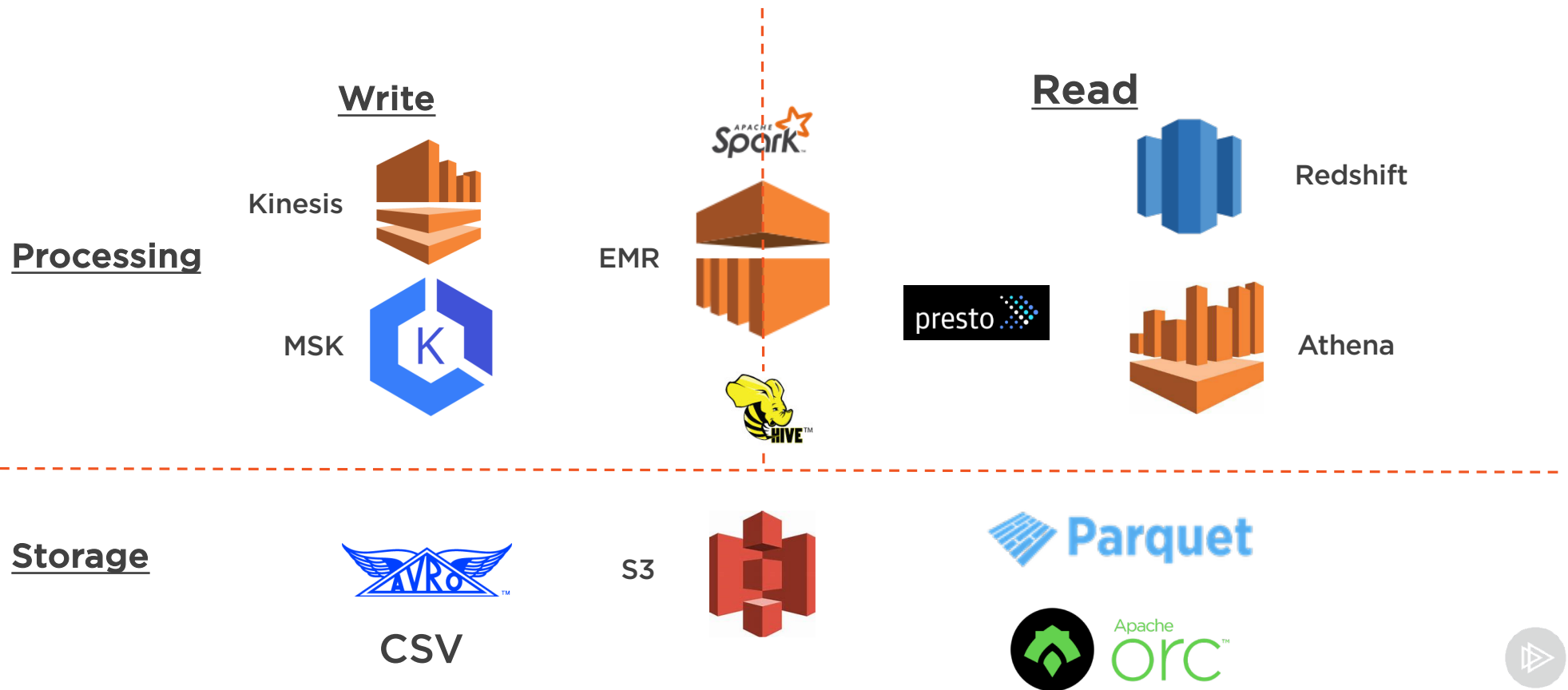


# EMR: Connections

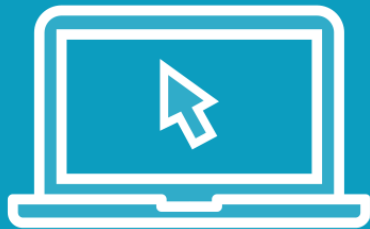




# The Data Lake Stack



# Demo



## Apache Hue

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

Hive query from command line

