CLOUD COMPUTING APPLICATIONS

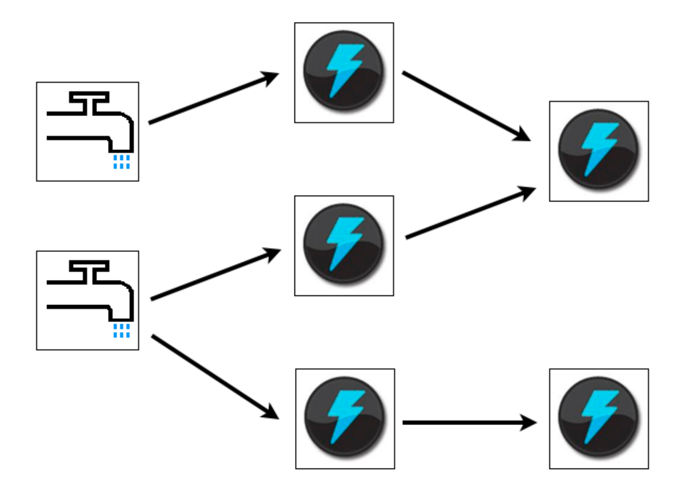
BIG DATA PIPELINES: THE RISE OF REAL-TIME

Matt Ahrens – Yahoo

The Rise of Real-Time

- As Hadoop ramped up to offer batch data availability, a growing need arose to provide data in real-time for analytic and instant feedback use cases

- Storm became stable for production scale in 2012



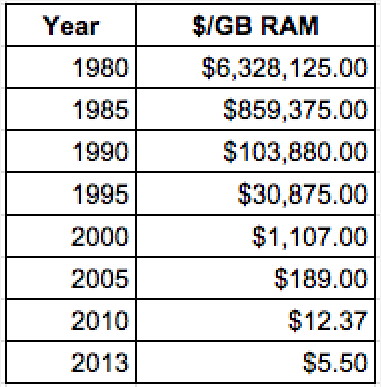
The Storm Fire Hose

- Topologies
  - graph of spouts and bolts that are connected with stream groupings
  - runs indefinitely (no time/batch boundaries)
- Streams
  - unbounded sequence of tuples that is processed and created in parallel in a distributed fashion
- Spouts
  - input source of streams in topology
- Bolts
  - processing container, which can perform transformation, filter, aggregation, join, etc.
  - sinks: special type of bolts that have an output interface

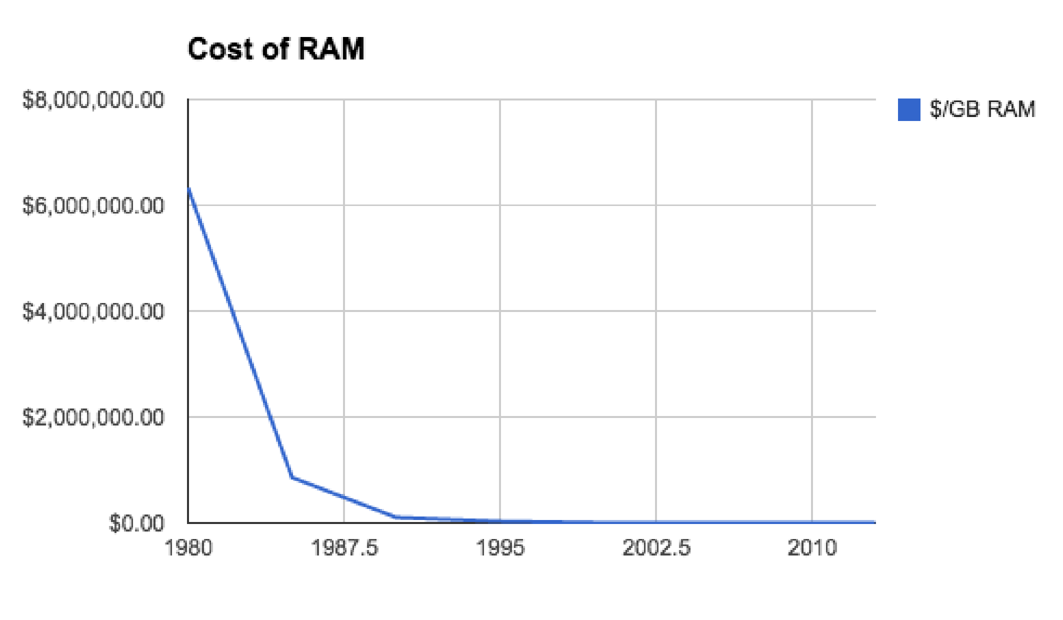
How Did We Get Here?

- People always have wanted data faster

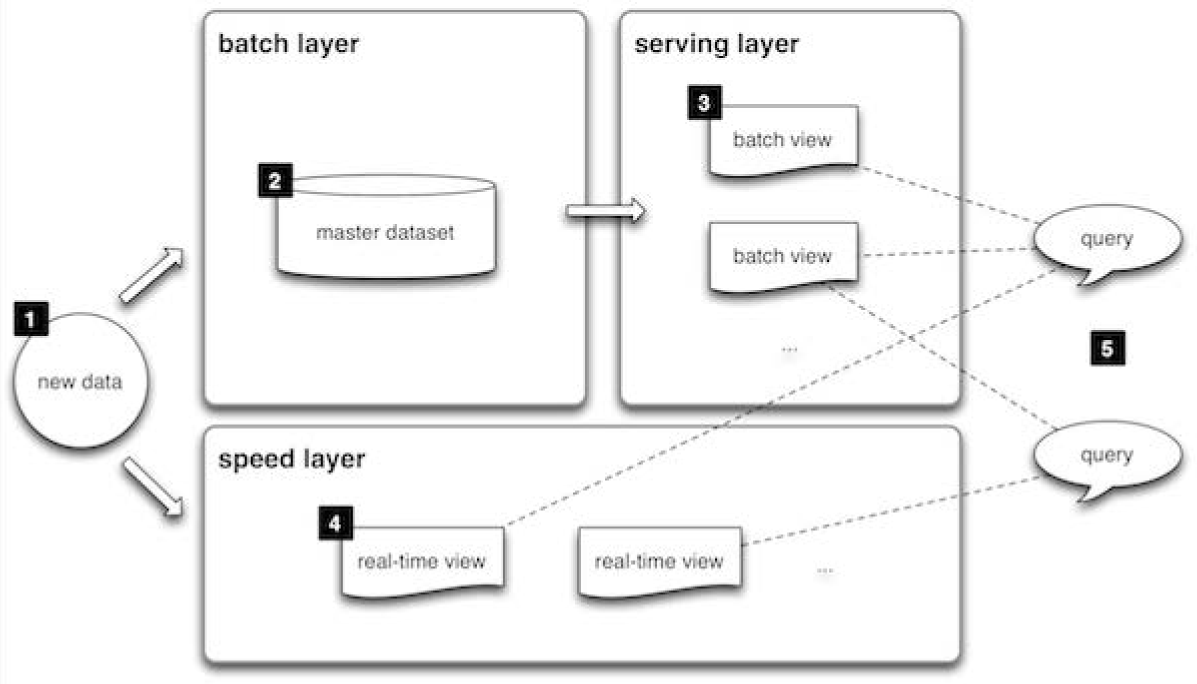
- Finally we had hardware costs that were in line with doing in-memory streaming for billions of events/day



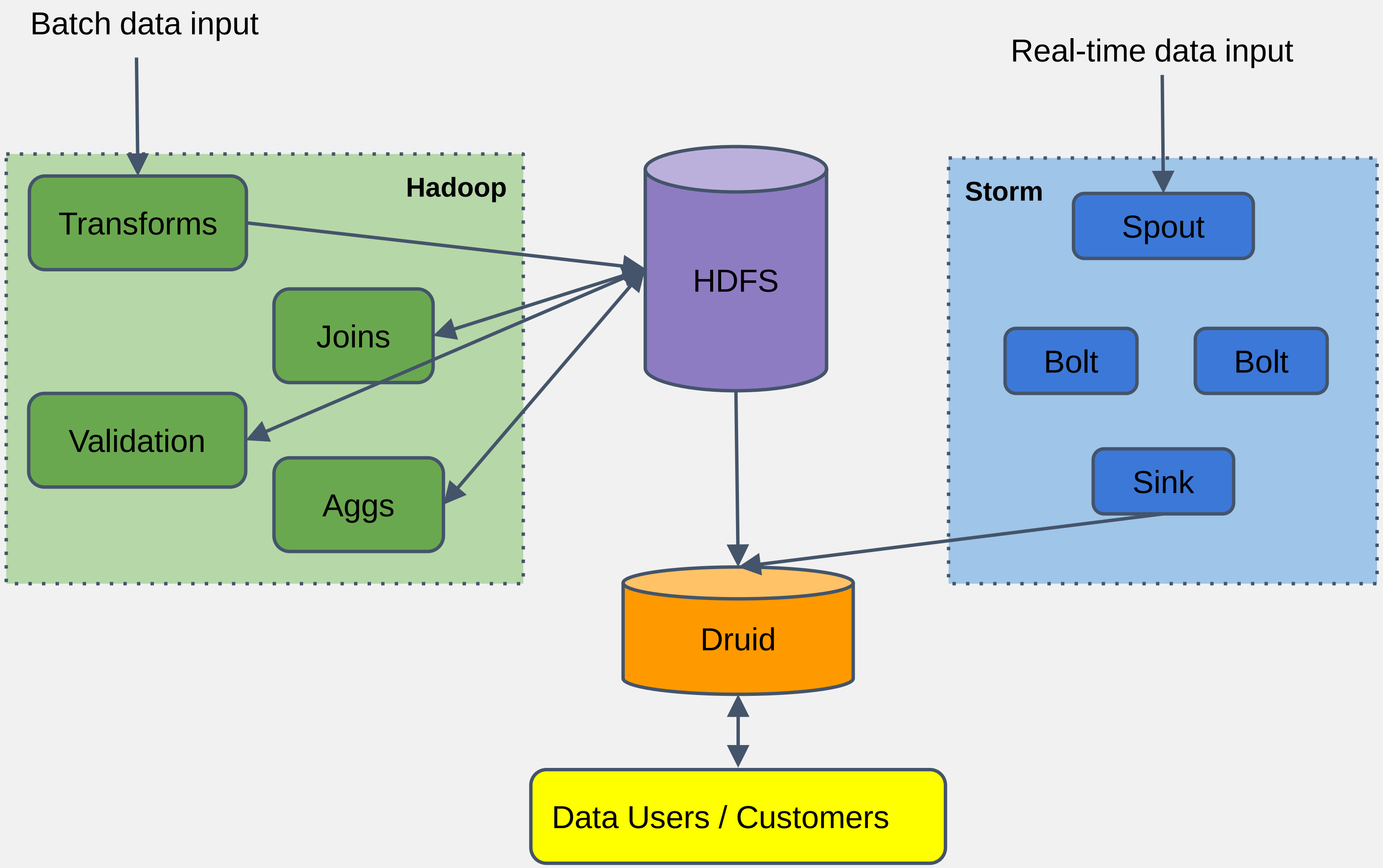
Year	$/GB RAM
1980	$6,328,125.00
1985	$859,375.00
1990	$103,880.00
1995	$30,875.00
2000	$1,107.00
2005	$189.00
2010	$12.37
2013	$5.50



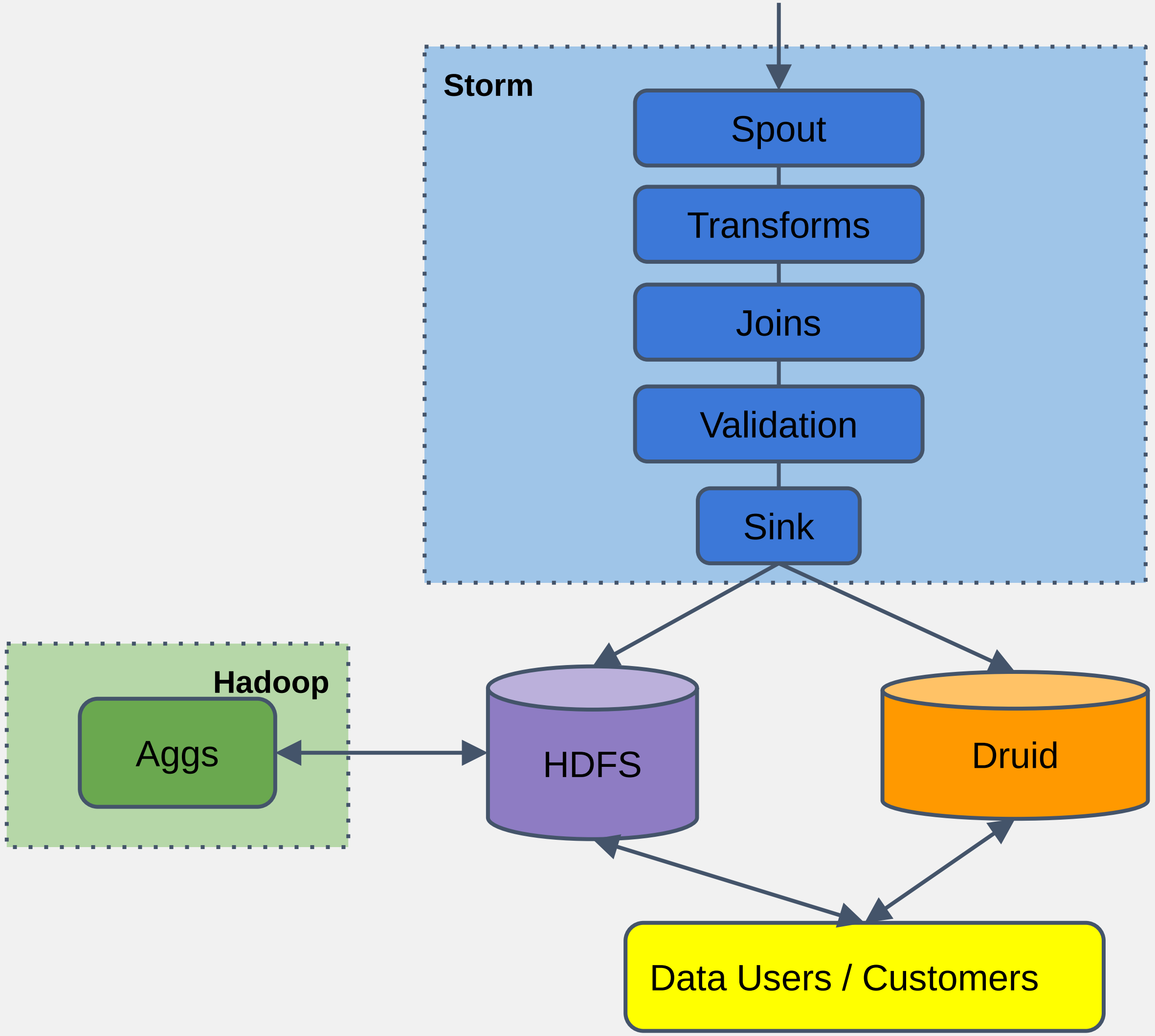
The Lambda Architecture: Real-Time + Batch



The Present Architecture



The Next Frontier: Real-Time as Source of Truth


Storm
Spout
Transforms
Joins
Validation
Sink
Hadoop
Aggs
HDFS
Druid
Data Users / Customers