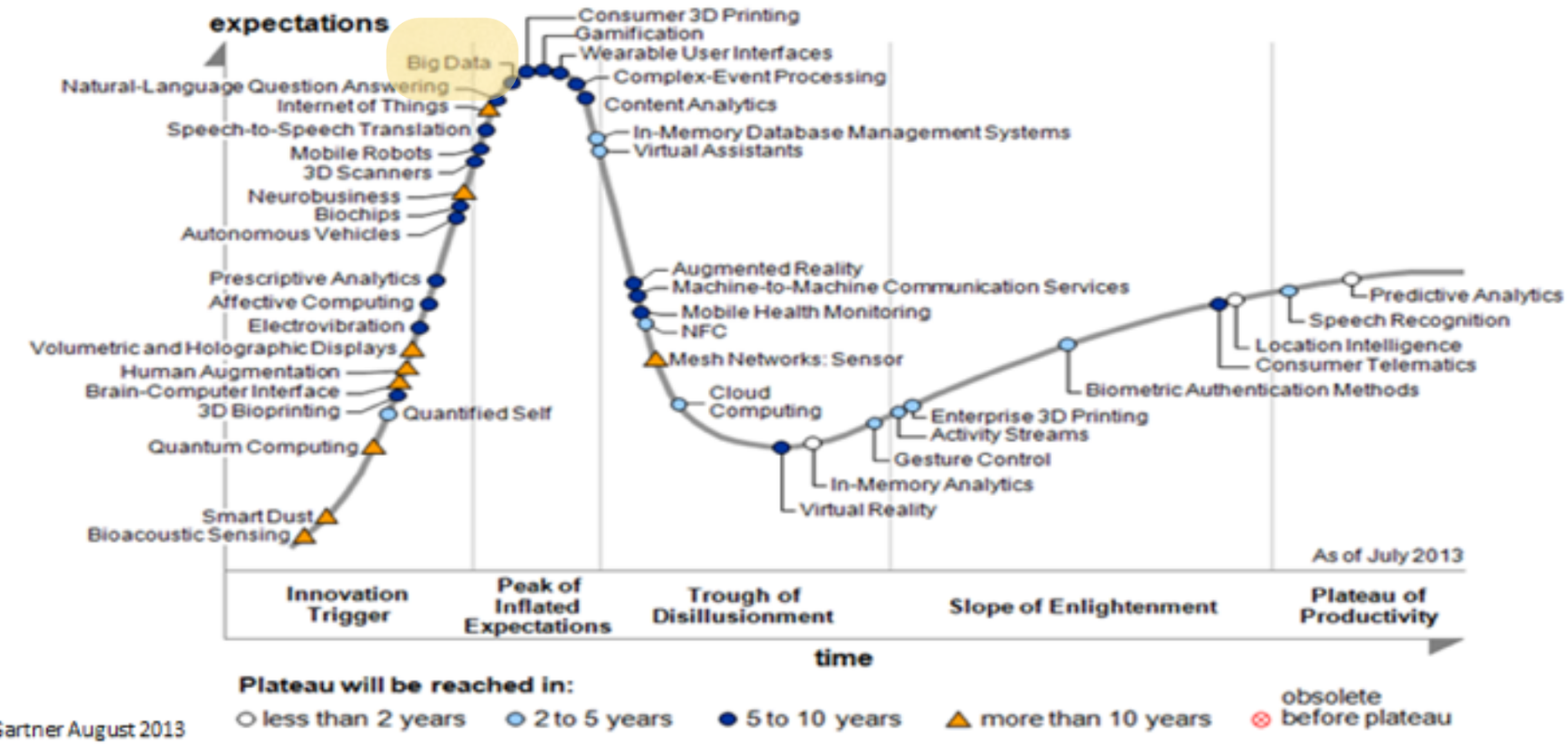
Stream Processing in "Big Data" world

Jags Ramnarayan,
Chief Architect, GemFire
Pivotal

Milind Bhandarkar,
Chief Scientist,
Pivotal

Pivotal

Hype Cycle for Emerging Technologies, 2013



www.facebook.com/EMCacademicalliance

EMC²

Hype Curve

Pivotal

Prediction: Hadoop Will Avoid Hype Curve by Being Flexible....

Pivotal

...Instead, Hype Curve will Apply to Individual Hadoop Components

Pivotal

Hadoop, the Project...

MapReduce

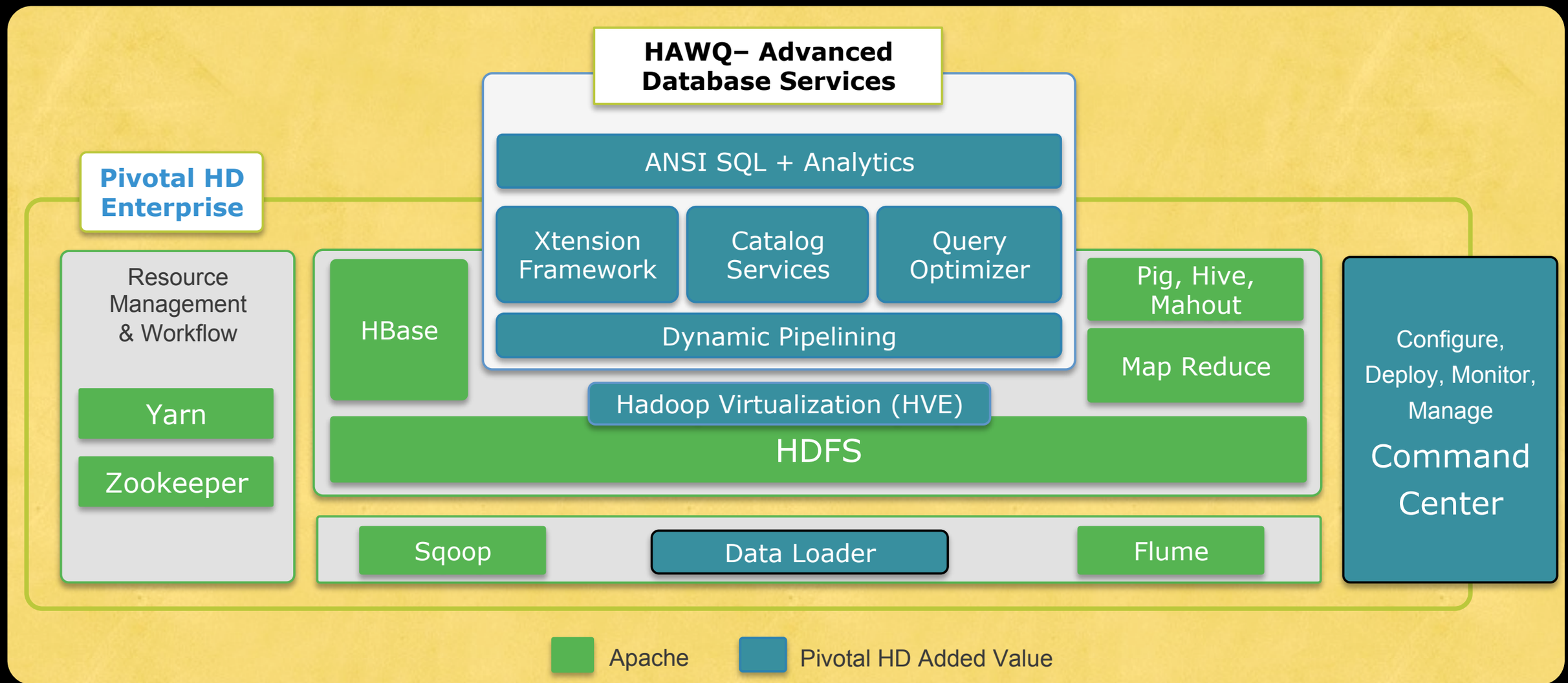
YARN

HDFS

Core

Pivotal

...vs Hadoop, the Product



MapReduce: A major step backwards

on Jan 17 in Database architecture, Database history, Database innovation posted by DeWitt

e.g. MapReduce

Pivotal

MapReduce: Fault Tolerance, Scalability, & Flexibility at the cost of Performance

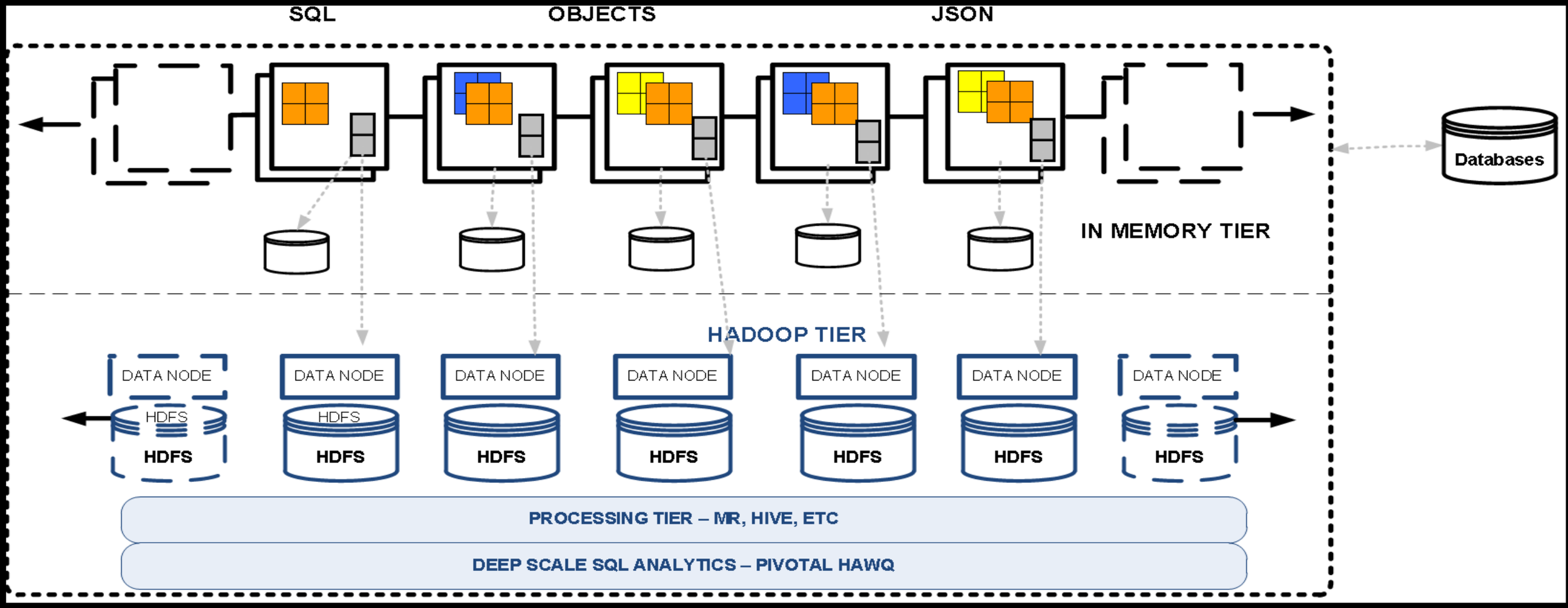
Pivotal

Performance Impact of MapReduce

User intelligence	4.2	198	47X
Sales analysis	8.7	161	19X
Click analysis	2.0	415	208X
Data exploration	2.7	1,285	476X
BI drill down	2.8	1,815	648X

Rise of Fast OLAP-On-Hadoop

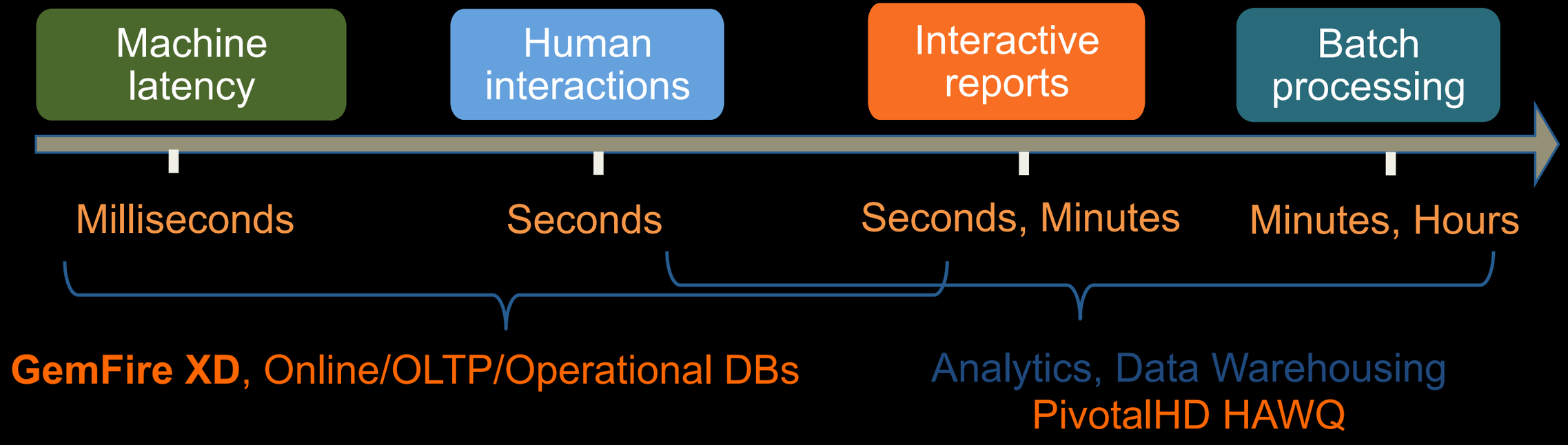
- Pivotal HAWQ (aka Greenplum DB on Hadoop)
- Cloudera Impala
- Hortonworks Stinger (Hive over Tez)
- Drill, BigSQL, PolyBase, Optiq/Lingual
- More to come... (and go, such as Spire...)



Gemfire-XD : Bringing OLTP/Operational DB to Hadoop

Pivotal

Latency Spectrum



Natural Next Step: Streaming in Hadoop

Pivotal

Large-Scale Stream Processing

- Storm (Backtype/Twitter 2010, Apache Incubator 2013)
- S4 (Yahoo 2010, Apache Incubator 2011)
- Spark Streaming (Berkeley AMPLab, 2012)
- Dempsey (Nokia, 2012)
- MUPD8 (@WalmartLabs, 2012)
- MillWheel (Google, 2013)
- Apache Samza (LinkedIn, Apache Incubator 2013)

Properties - 1

- Decoupling Logical Model from Physical Deployment
  - Partitioning, Replication, Colocation with Distributed Reference Datasets
- Event Delivery Model
  - At Least Once, At Most Once, Exactly Once
- Processor State
  - Stateless, Local State, Distributed State

Properties - 2

- Flexible Data Model
- Stream Slice
  - Event-at-a-time, Micro-Batch
- Integration with Hadoop
  - Resource Sharing, Persistence on HDFS, Exactly once writes to persistent store
- Intuitive Programming Model
  - Node, Channel, Flow

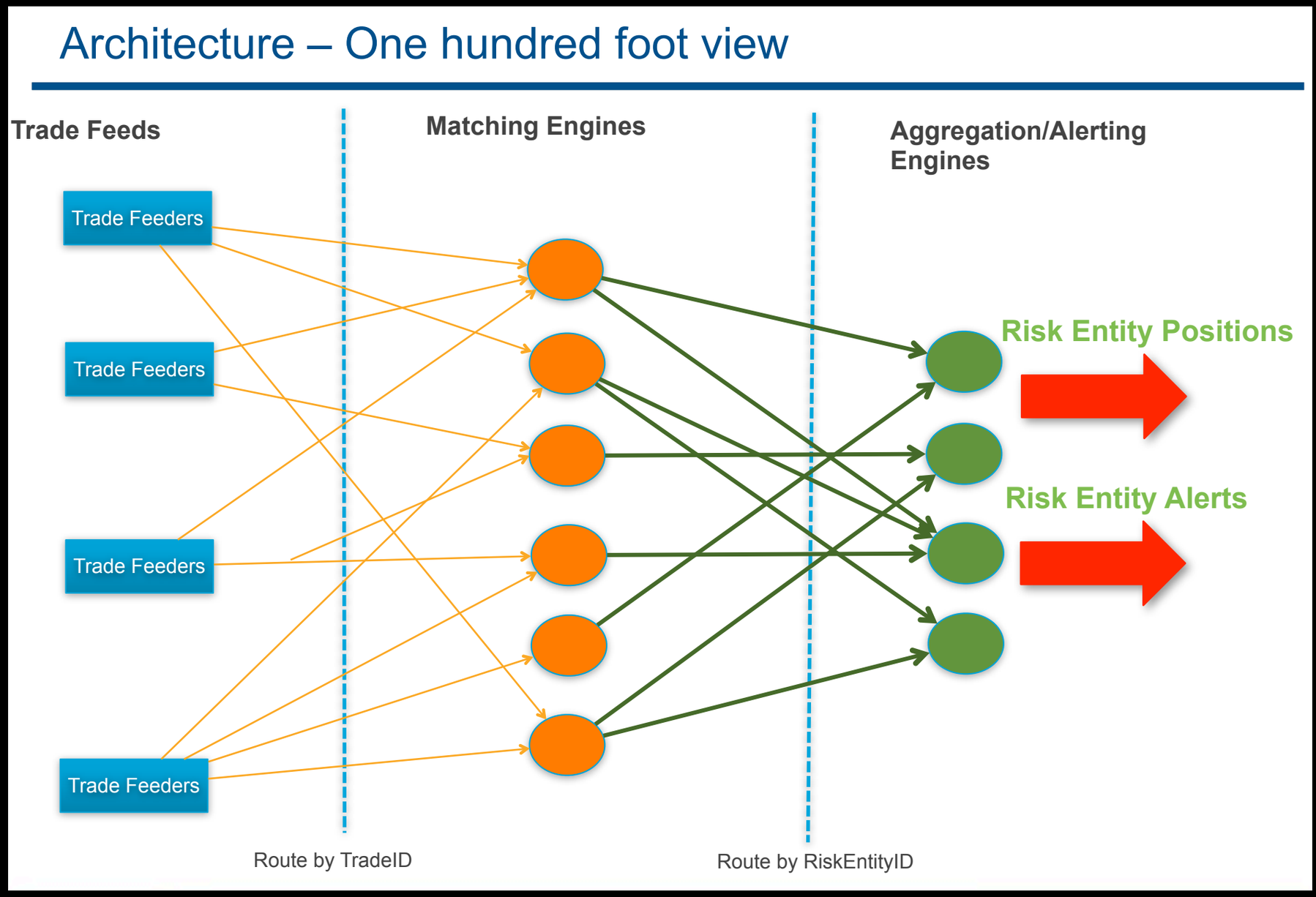
Pivotal

GemStreams

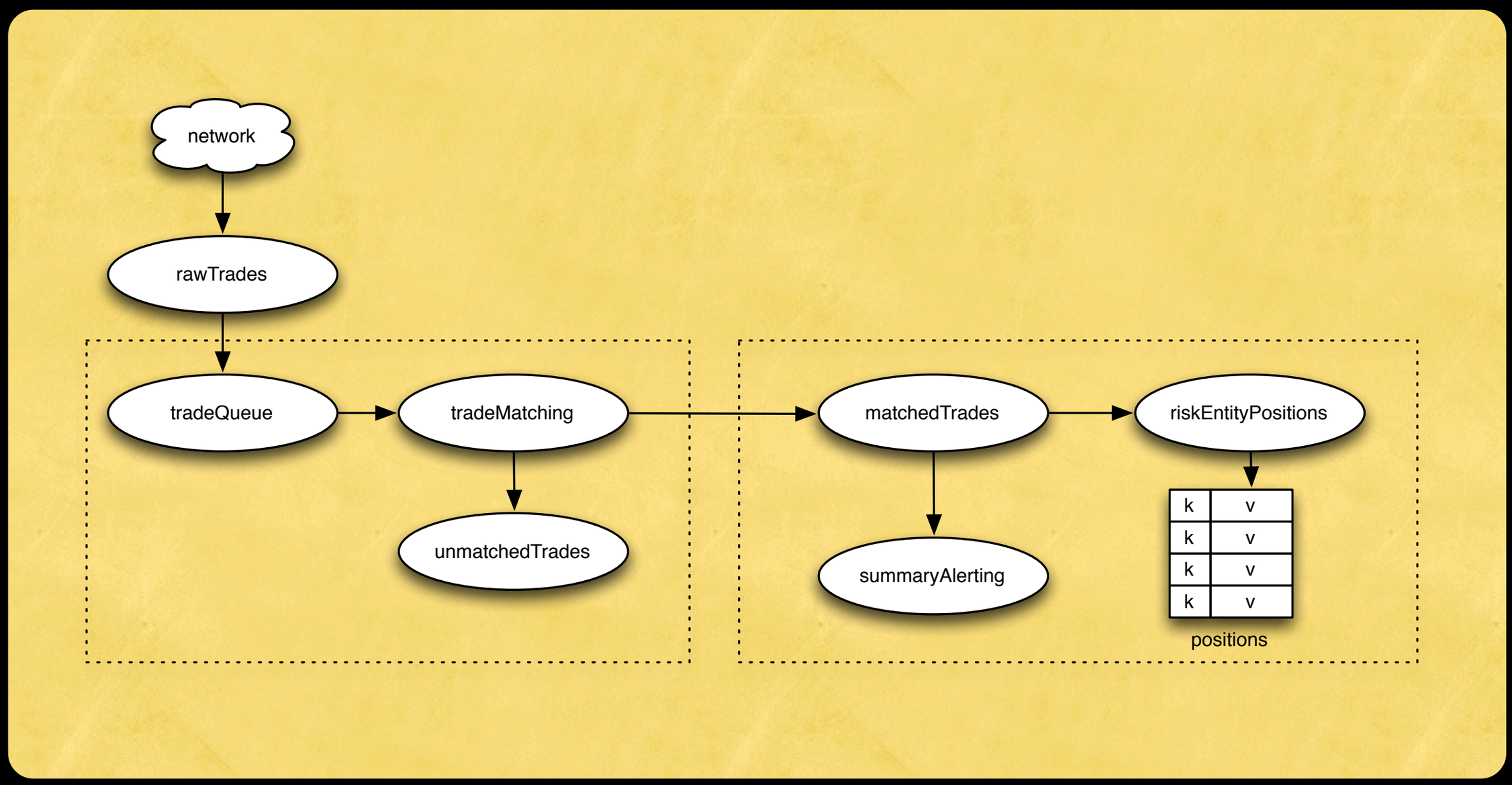
- In-Database Partitioned-Stream Processing
- Integrated Distributed State Management
- Replicated (Slow-Changing) Reference Datasets
- Reliable, Transparent, Exactly-Once persistence to HDFS
- Flexible Event Batching - Single Event or Micro-Batch
- Flexible Data Model - POJOs or Tuples

Pivotal™

Use Case: Trade Matching



Event Flow


network
rawTrades
tradeQueue
tradeMatching
unmatchedTrades
matchedTrades
summaryAlerting
riskEntityPositions
k
v
k
v
k
v
k
v
positions