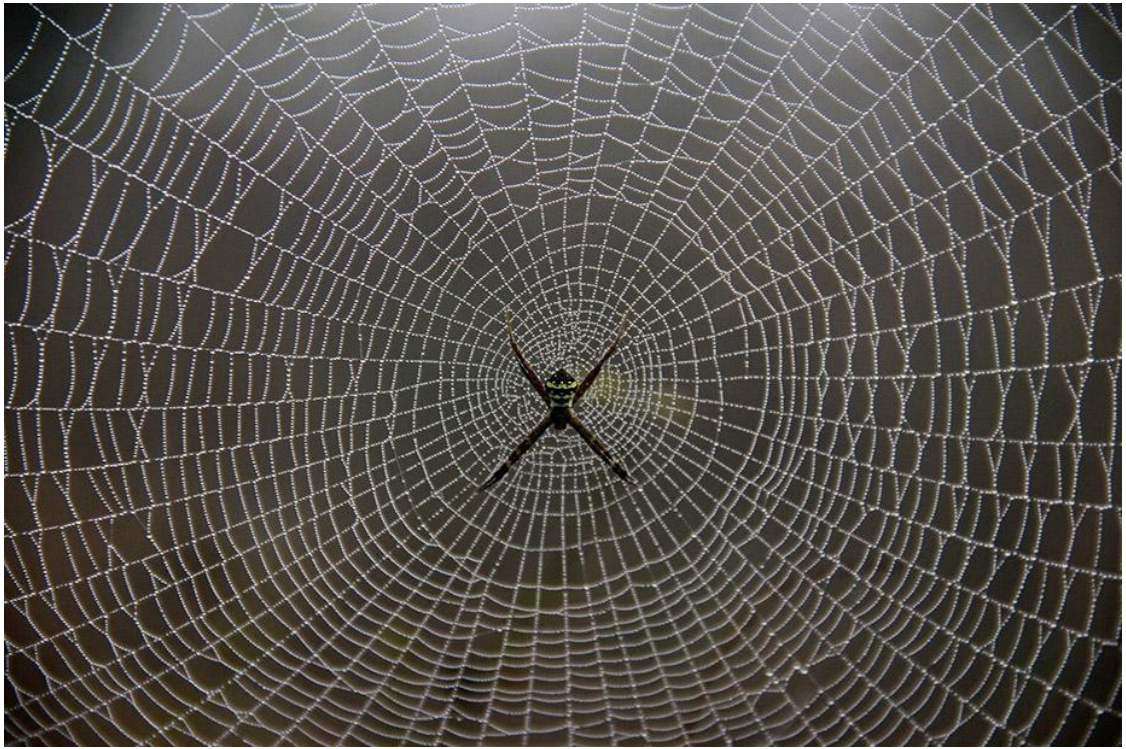
A Tale of Two Technologies

A Story of the IoT “revolution”

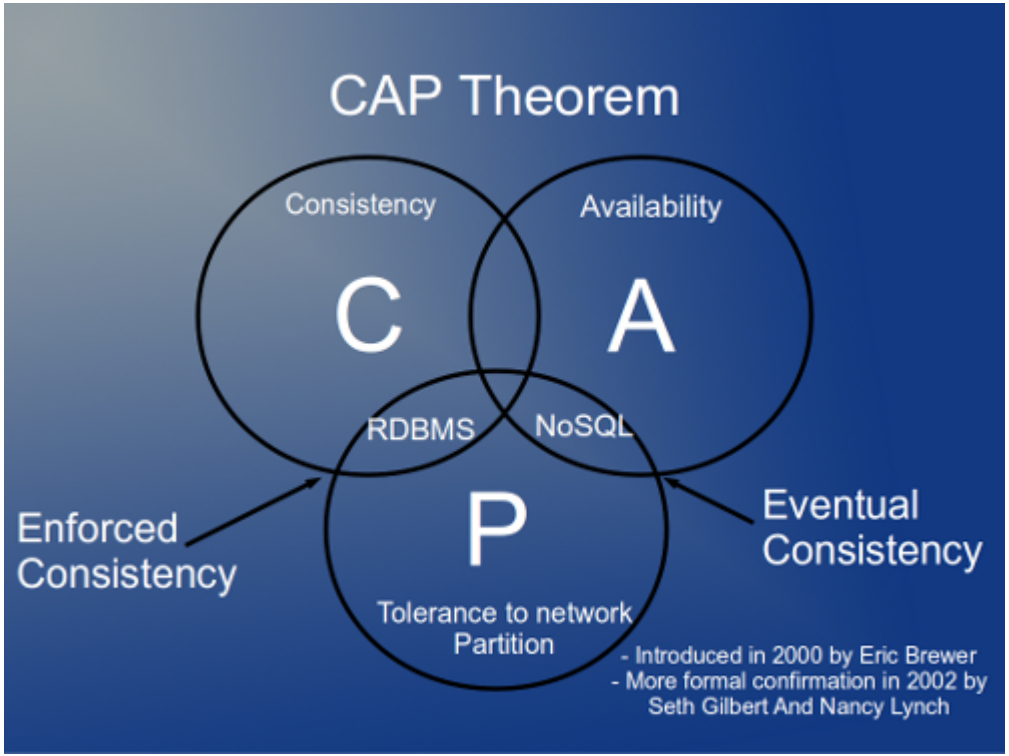
It was the age of connectedness



It was the age of disconnectedness

Failure tolerance is not optiona

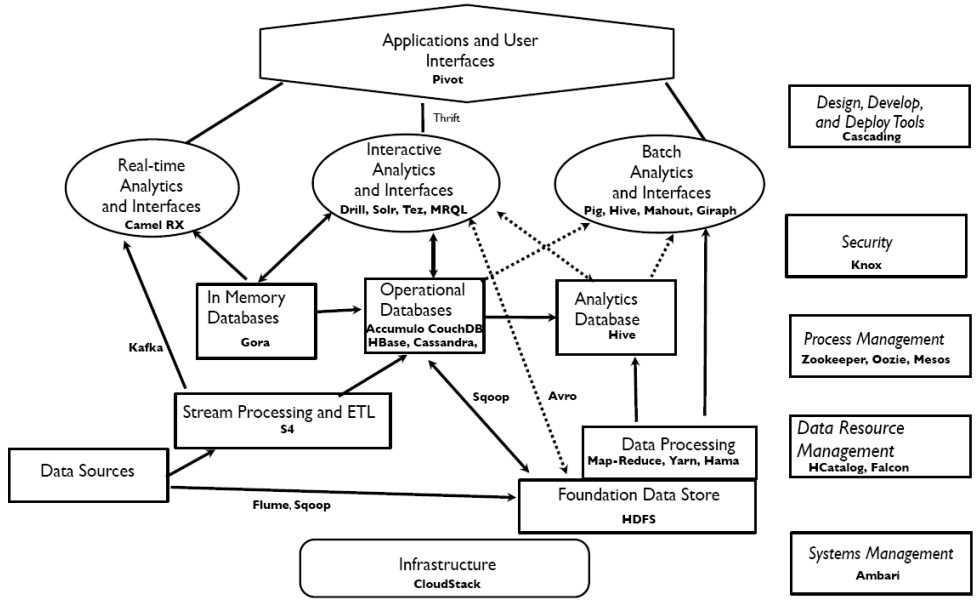
It was the age of wisdom



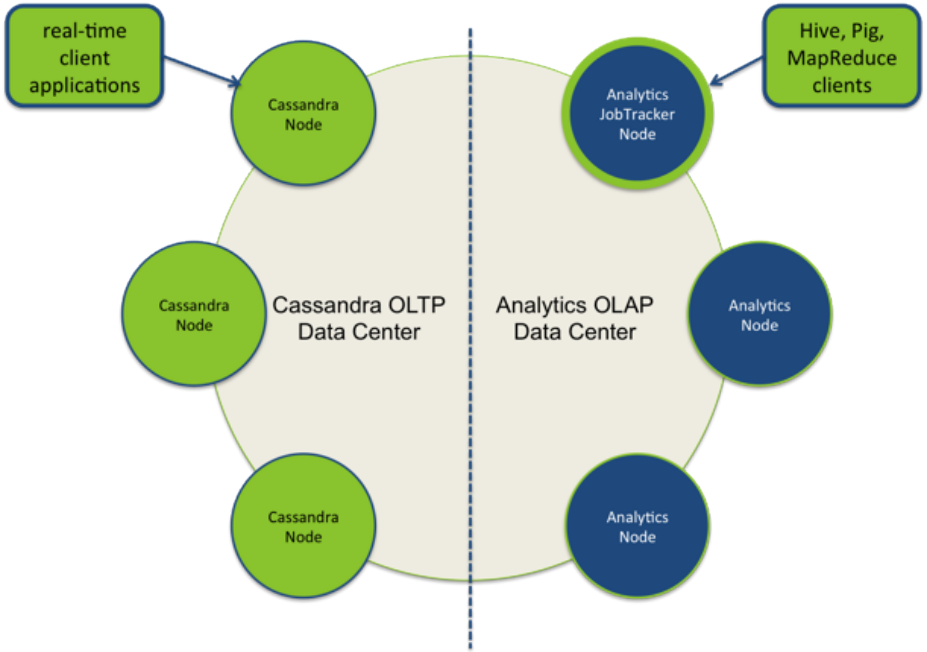
It was the age of foolishness

It was the epoch of complexity

Apache Big Data Framework in Reference Architecture



It was the epoch of simplicity



What is the Internet of Things?

“The Internet of Things (IoT) refers to uniquely identifiable objects and their virtual representations in an Internet-like structure.”
-Wikipedia

No really what is IoT?

- It’s literally the act of connecting “things” to the Internet
- It predates the World Wide Web
- It shouldn’t be surprising to anybody

So IoT is old news?

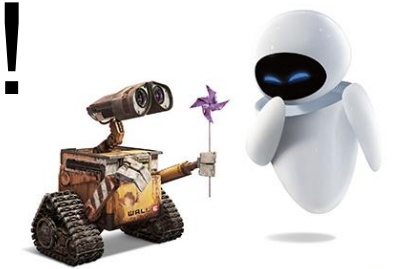
Most definitely

So IoT is just hype?

Fitness trackers

Sensors

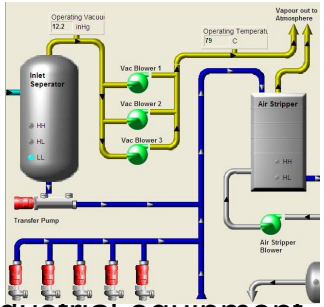
NO!



Medical devices

Home security and automation

Your (driverless?) car



Industrial equipment

Been there done that.

- DataStax and Cassandra have been dealing with event streams (time series) information at scale for years.
- But Cassandra is relatively dumb storage.
- Put data in. Get data out.

What is Cassandra? Chocolate!

- A massively scalable distributed database
- Chooses availability over strong consistency (yes, that really is a fundamental tradeoff)
- With its wide partitions it is able to take advantage of data locality even at “web scale”

What is Spark? Peanut butter!

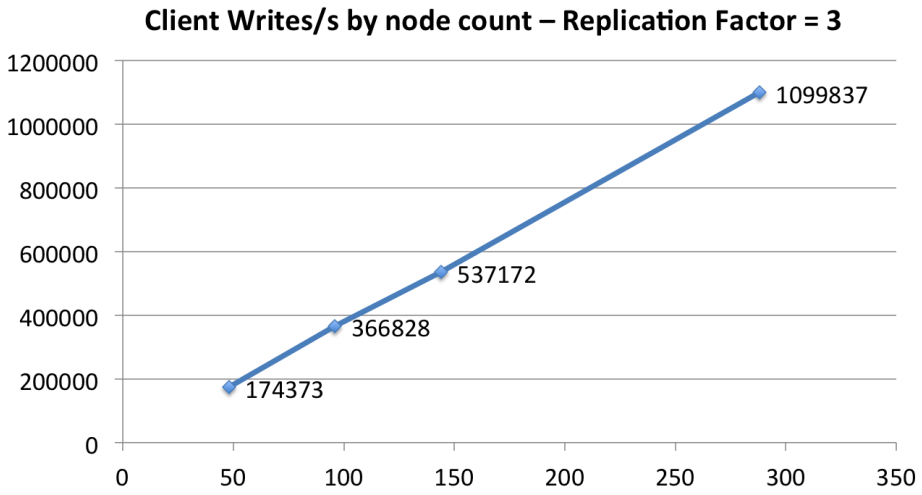
- DAG is a logical superset of M/R
- Adopts much of the Hadoop ecosystem, without being bound by it
- Intelligent use of caching (RDDs) for massive performance gains
- Incorporate Streaming to make ingestion-time processing a first class citizen

What does IoT need from big data?

- Log time-series events -- at scale
- Gather meaning from that data -- at scale
- Report on that data -- at scale
- Take action on that data -- at scale

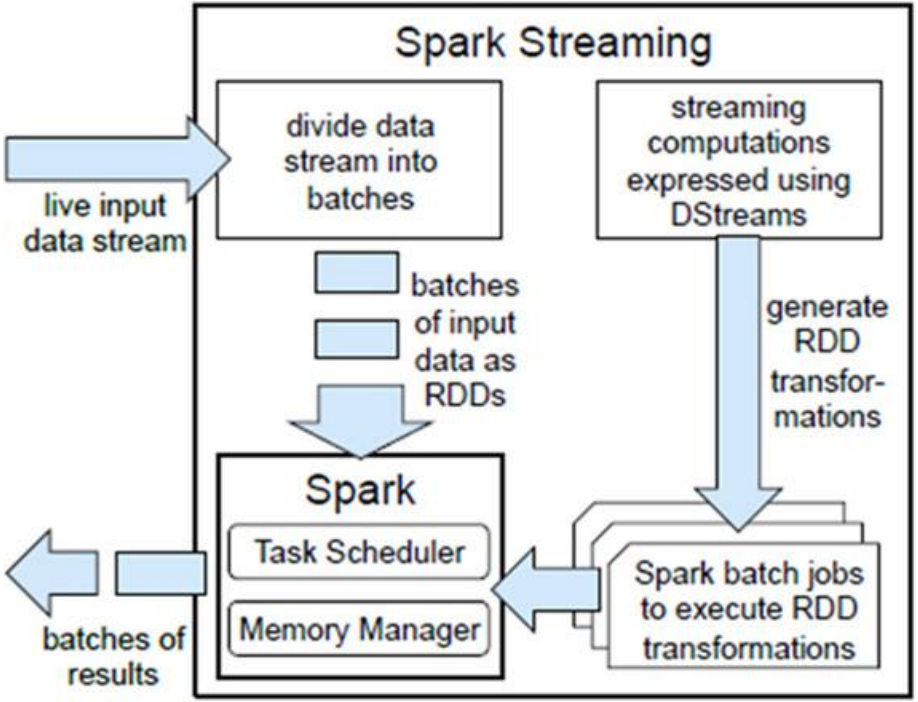
Logging events at scale

Scale-Up Linearity



NETFLIX

Gathering meaning at scale



Reporting at scale

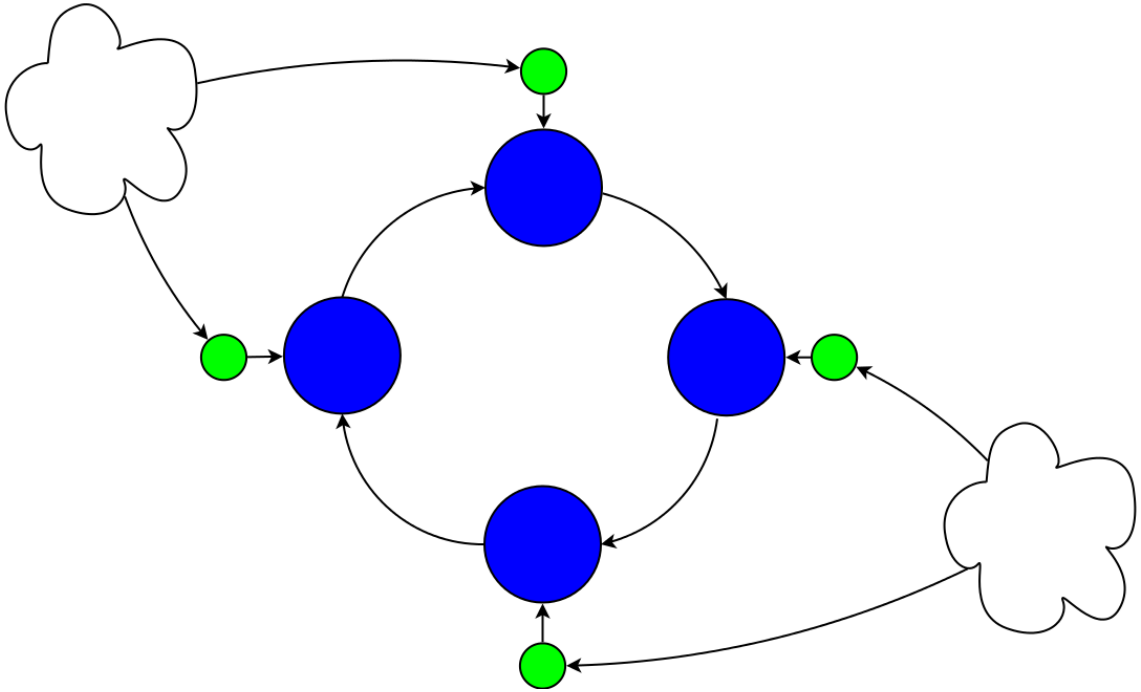
- Canned reports
- Ad-hoc querying and reporting
- Drill down / exploratory
- Alerting
- Aggregation
- Clustering (K-means, et al)
- Generalized machine learning

Take action at scale

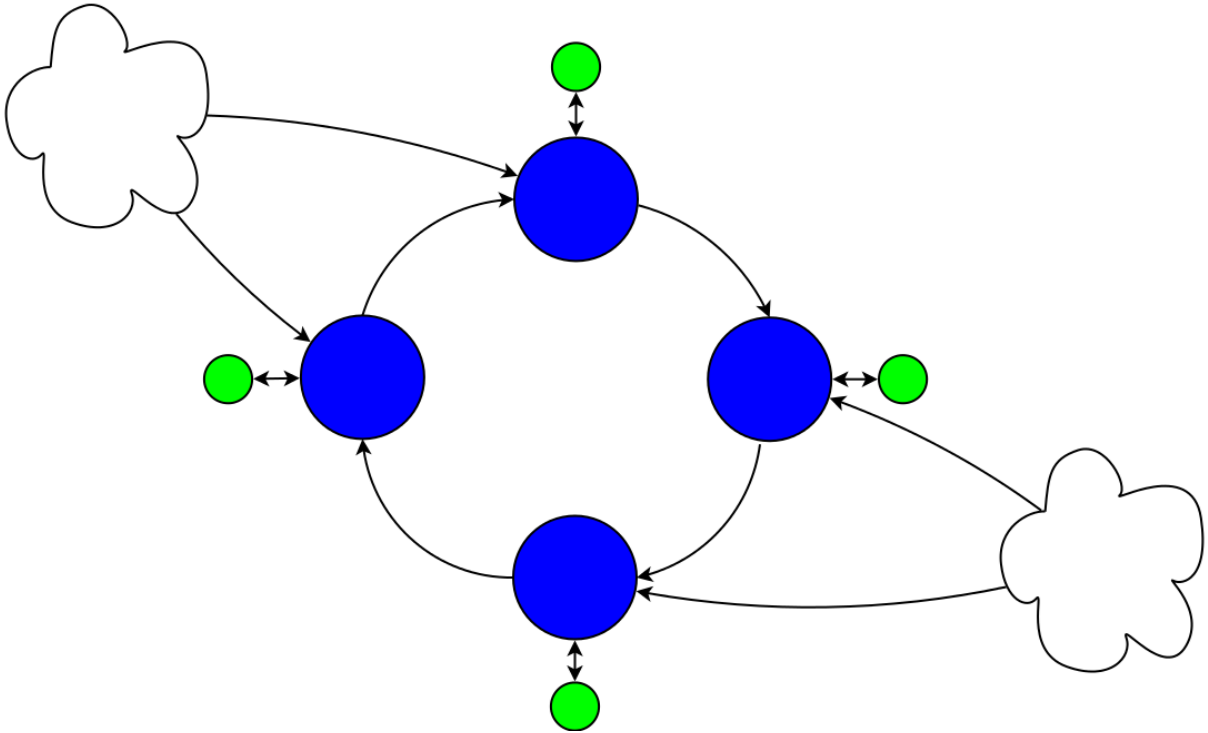
- Stateless application servers
- Horizontally scaled and co-located with Cassandra and Spark in each DC
- Any platform with a CQL driver

The architecture...

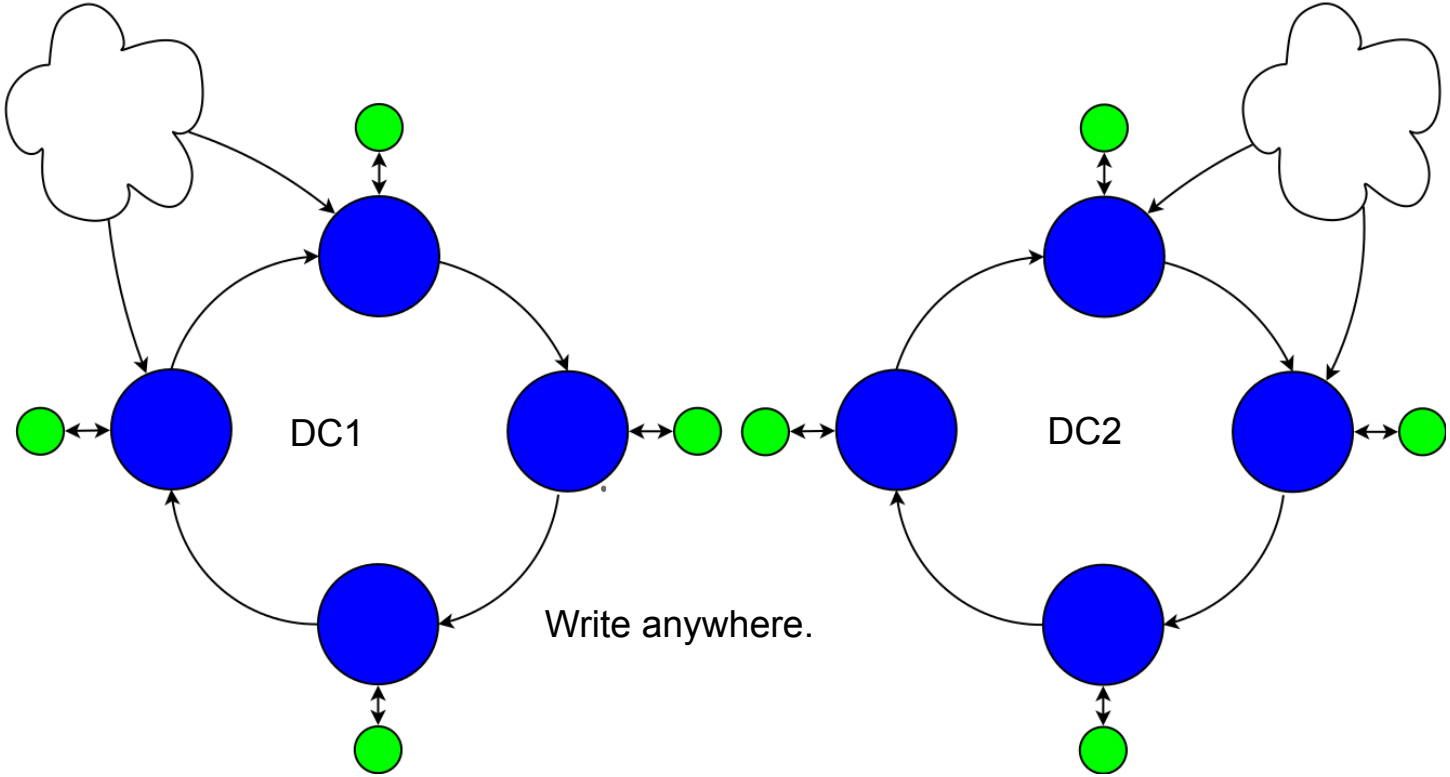
Spark to Cassandra



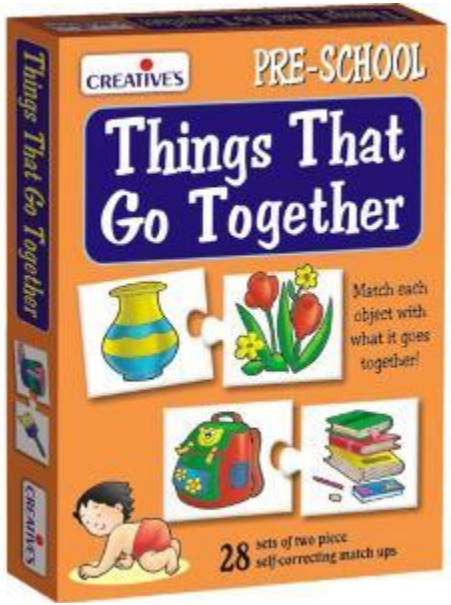
Cassandra to Spark



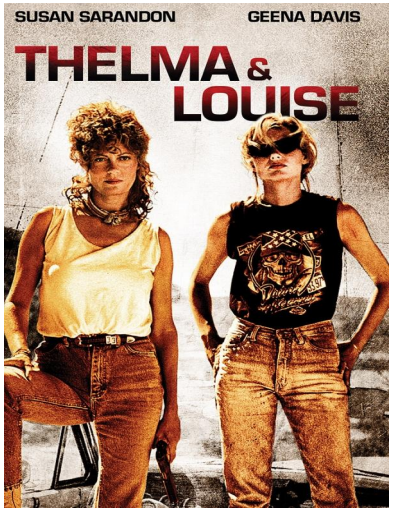
Multi-DC



Things that go together



BEN & JERRY'S



Things that go together

A Tale of Two Summits

Spark Summit 2014