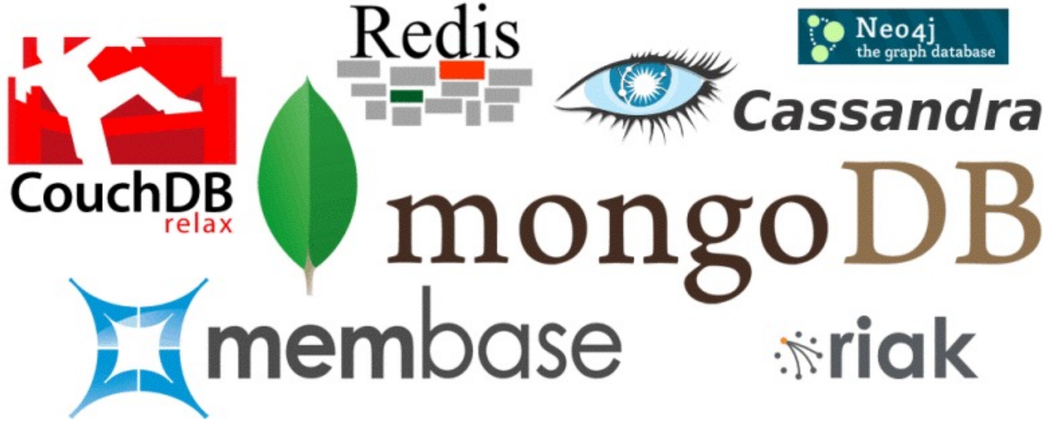
Big Data : Data Agility at Scale

Vivek Pradhan

Sydney 2015

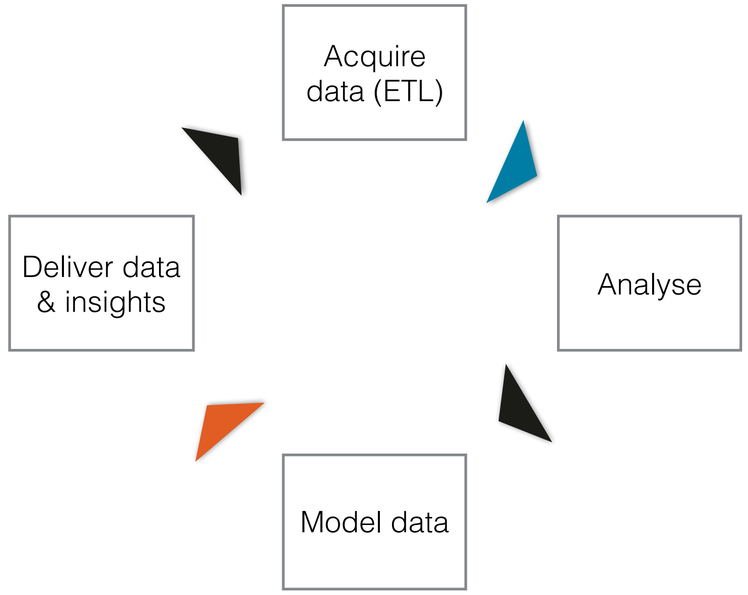
bigdataeverywhere

Agility


Redis
Neo4j
the graph database
Cassandra
CouchDB
relax
mongoDB
membase
riak

Traditional data management

- Schema design upfront
- Integration with polyglot data - hard
- Mostly interactive and batch workloads

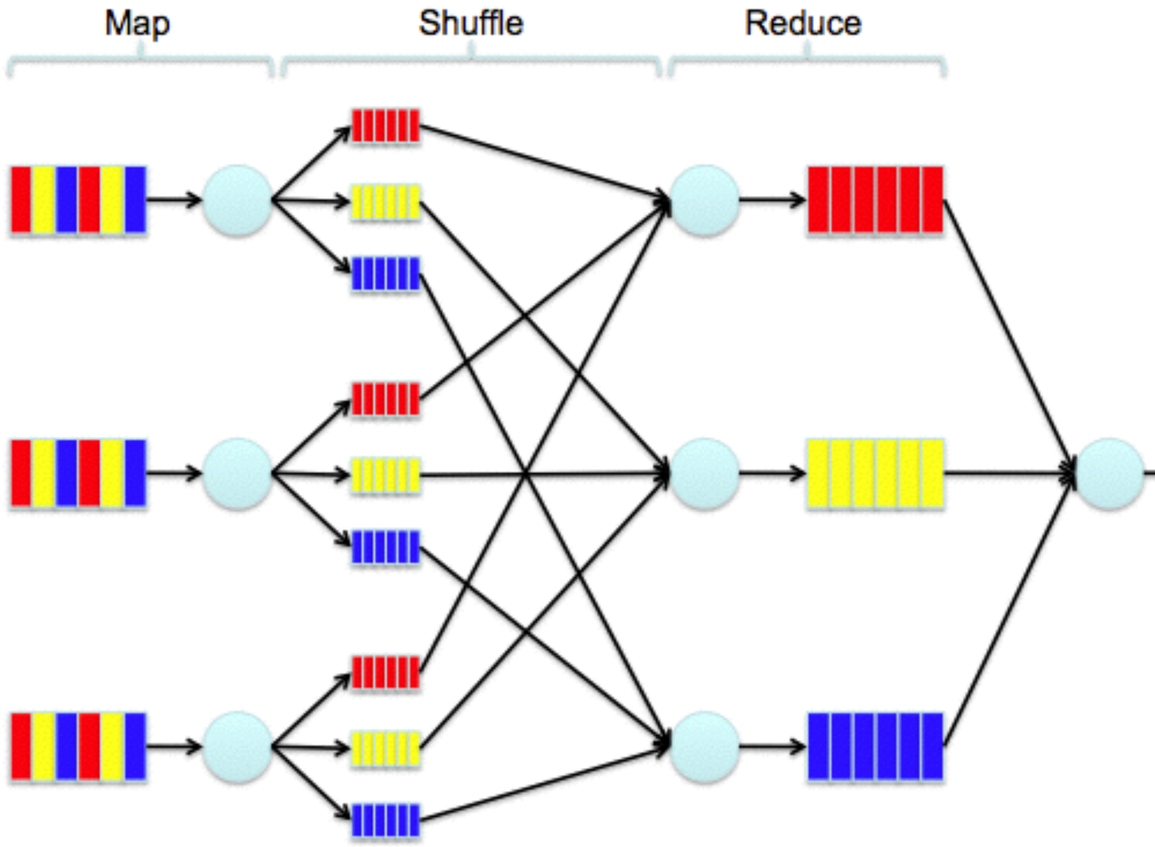


The Hadoop promise

- Very large datasets
- Scale out processing on commodity hardware
- Schema on read

Hadoop: Early capabilities

- Batch oriented
- Solutions required lots of development using Map Reduce pattern



Data Agility at Scale with Hadoop

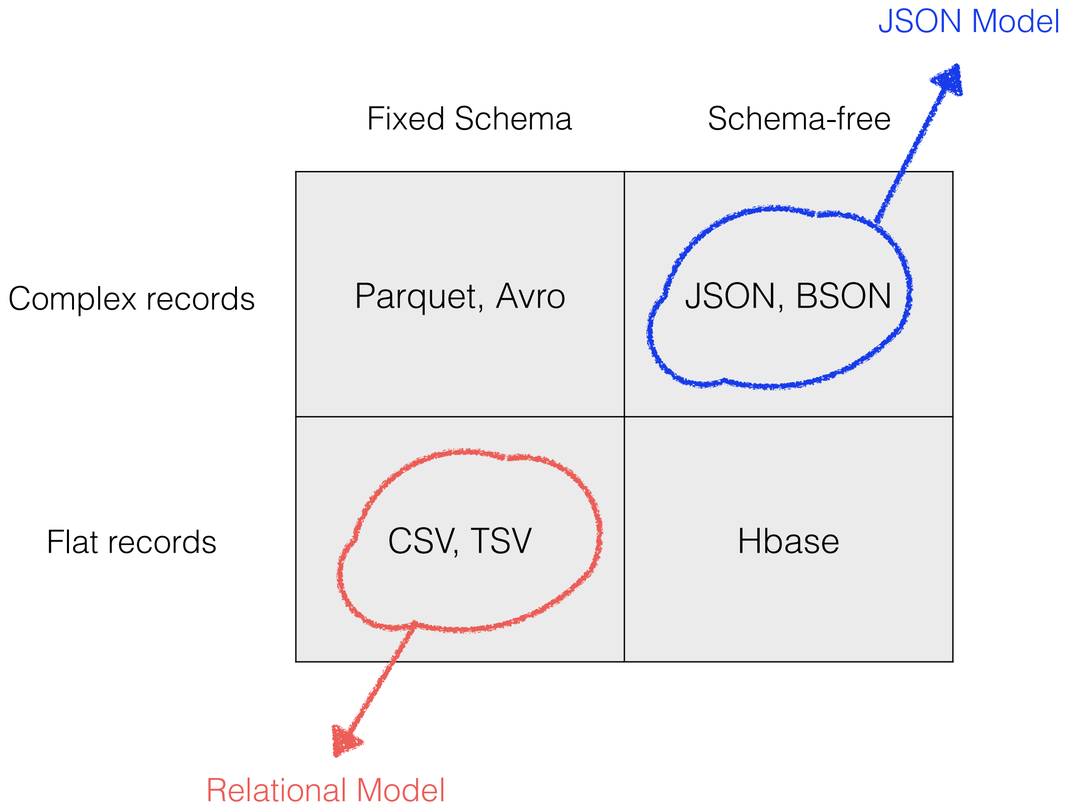
- Apache Spark
- Apache Drill

Apache Spark

- Fast general purpose computing framework - with or without Hadoop
- Support for
  - SQL (Spark SQL)
  - Machine Learning (Spark Mlib)
  - Streaming (Spark Streaming) and
  - GraphX (Graph processing)

Apache Drill

- Low latency ANSI compliant SQL query engine
- Query data in any format - structured CSV, JSON, log files
- Use with existing BI tools



Demonstration

Summary

- Business agility as a competitive differentiator
- Traditional data management not suitable for all data use cases
- Hadoop ecosystem has evolved beyond initial batch oriented patterns
- Projects like Spark and Drill are lowering barriers to entry for Hadoop consumption