QCon

Please evaluate my talk via the mobile app!

Design Patterns for Large-Scale Real-Time Learning

QCon London 2014

Sean Owen / Director of Data Science / Cloudera

What We Talk About When We Talk About Data Science

“A data scientist is a statistician who lives in San Francisco.

Data Science is statistics on a Mac.

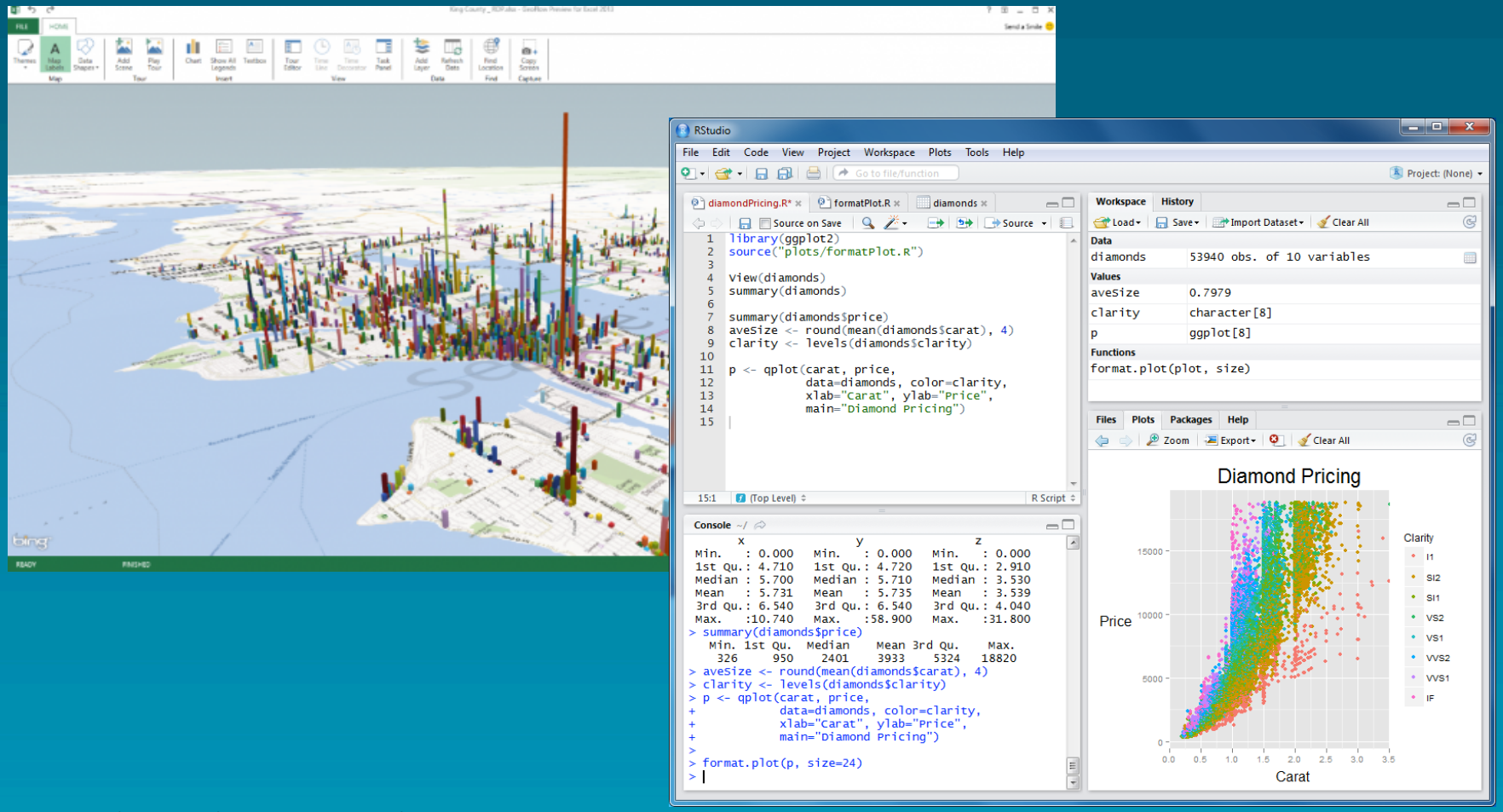
A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.”

www.quora.com/Data-Science/What-is-the-difference-between-a-data-scientist-and-a-statistician

“It’s about finding insights in data
It’s what specialists do to understand data
It’s applied statistics at large scale
It’s predicting the future from data
It’s about analyzing politics,
sports, markets, etc.”

Data Scientist

Data Science Is **Exploratory Analytics?**

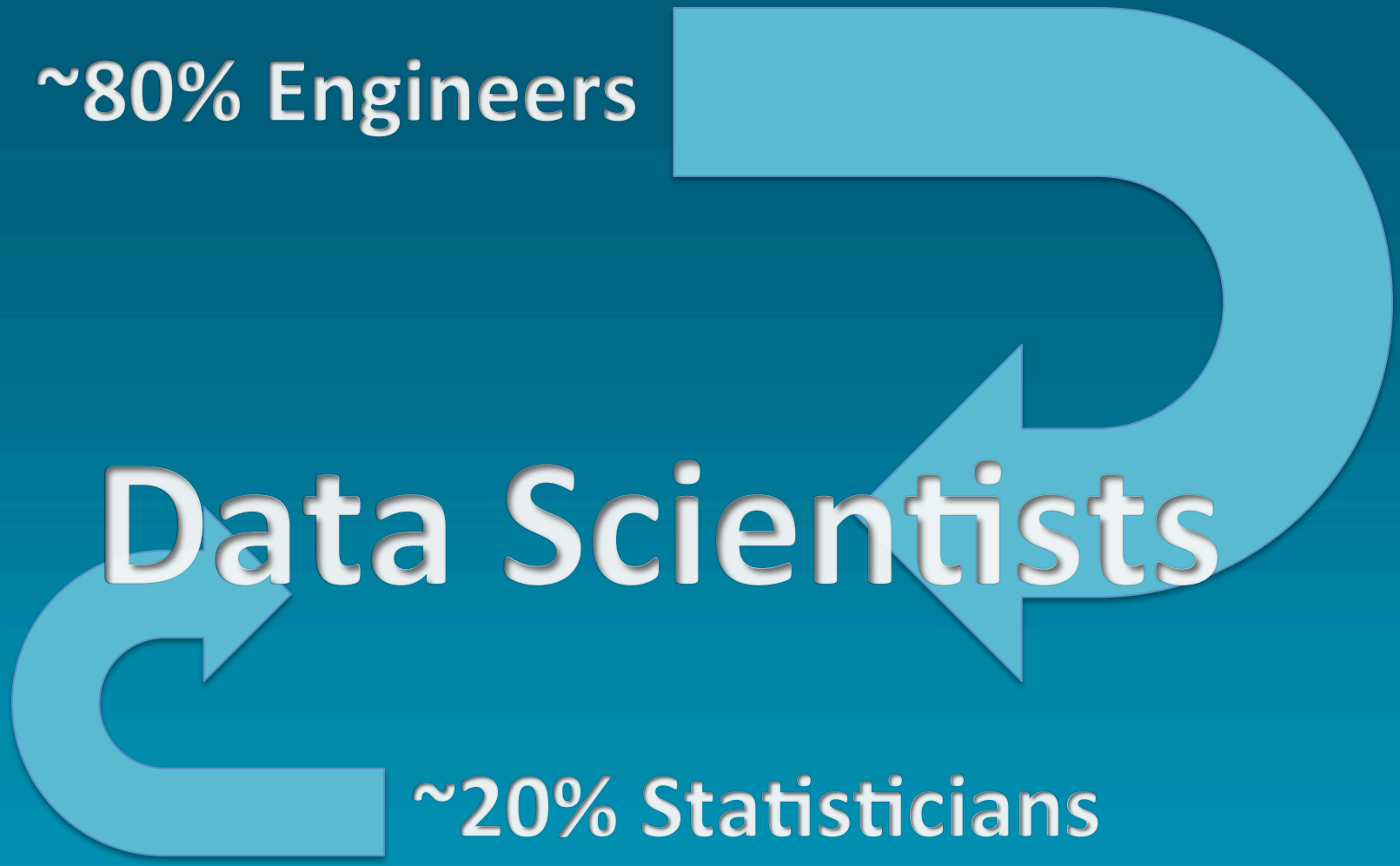


www.tc.umn.edu/~zief0002/Comparing-Groups/blog.html
thenextweb.com/microsoft/2013/07/08/microsoft-brings-the-office-store-to-22-new-markets-adds-power-bi-an-intelligence-tool-to-office-365/

Example: Drug Interactions

Cloudera analysis of FDA drug data: “Our analysis revealed a few drug pairs with surprisingly high correlations with adverse events that did not show up in a search of the academic literature: gabapentin (a seizure medication) taken in conjunction with hydrocodone/paracetamol was correlated with memory impairment, and haloperidol in conjunction with lorazepam was correlated with the patient entering into a coma.”

blog.cloudera.com/blog/2011/11/using-hadoop-to-analyze-adverse-drug-events/


~80% Engineers
Data Scientists
~20% Statisticians

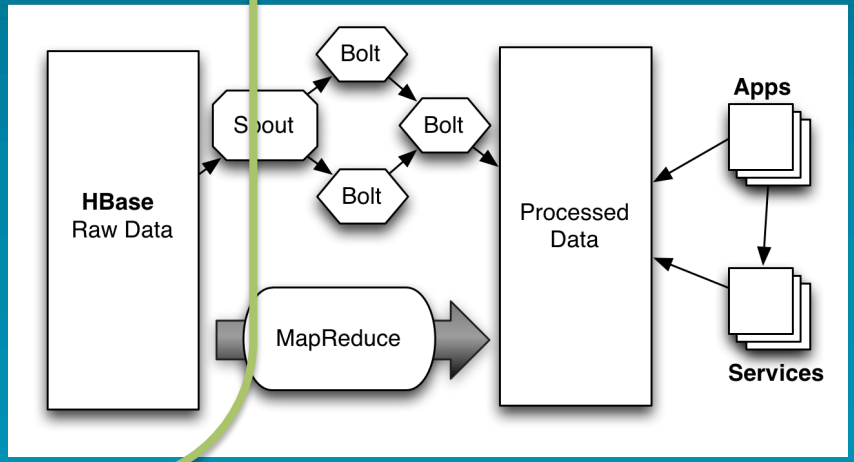
Example: Data Science in the Field

- *[Large European e-commerce site]*
- Wants real-time recommendations for new and returning users
- Data streamed from web server via Flume to HDFS
- Multiple data sources
- 100K+ products, 20M users

Exploratory?

Example: Cerner

- Search, ML over Patient Data
- MapReduce for indexing, learning — Machine Learning
- HBase for storage and fast access
- Also: Storm for incremental update
- And: relational DB for most recent derived data
- API façade for input; API for querying learning



Engineering

engineering.cerner.com/2013/02/near-real-time-processing-over-hadoop-and-hbase/

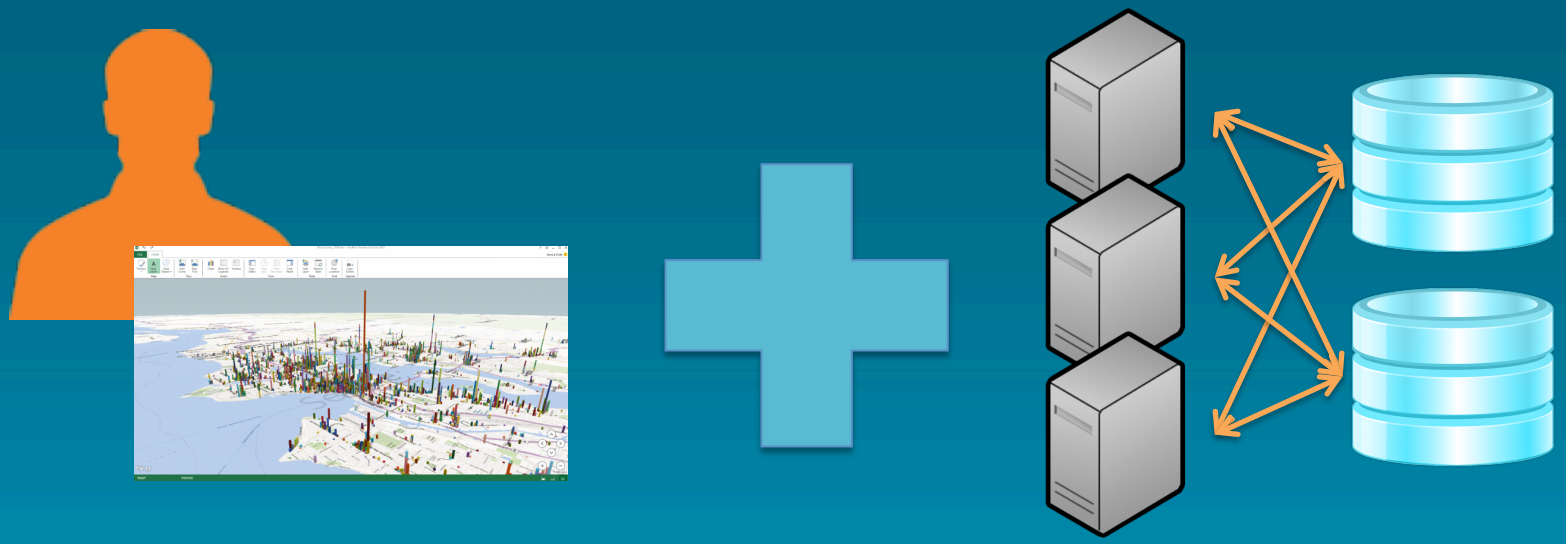
Adding Operational Analytics

2014: Lab to Factory

Data Scientist
Exploratory Analytics

Predictive Data Products
Operational Analytics

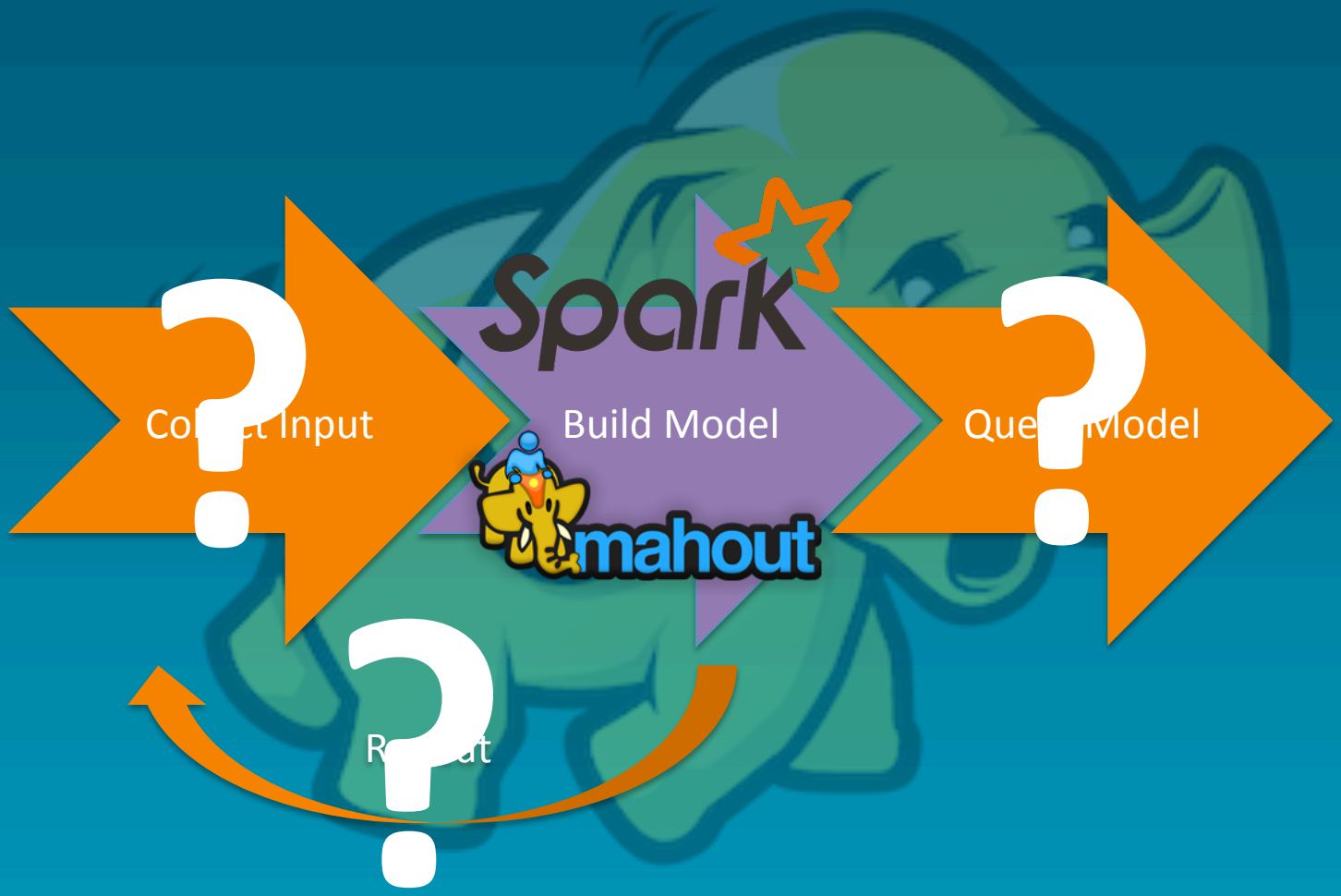
Data Science Will Be **Operational Analytics**



Data Scientist
Exploratory Analytics

Predictive Data Products
Operational Analytics

I Built A Model On Hadoop. Now What?



Example: Oryx

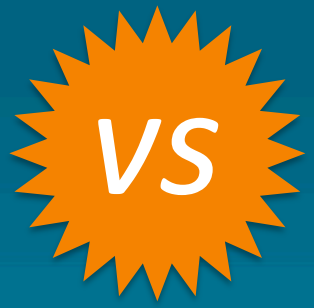
www.mwttl.com/wp-content/uploads/2013/11/IMG_5446_edited-2_mwttl.jpg

Gaps to fill, and Goals

- Model Building
  - Large-scale
  - **Continuous**
  - Apache Hadoop™-based
  - Few, good algorithms
- Model Serving
  - **Real-time query**
  - **Real-time update**
- Algorithms
  - Parallelizable
  - Updateable
  - Works on diverse input
- Interoperable
  - PMML model format
  - **Simple REST API**
  - Open source

Large-Scale or Real-Time?

Large-Scale
Offline
Batch



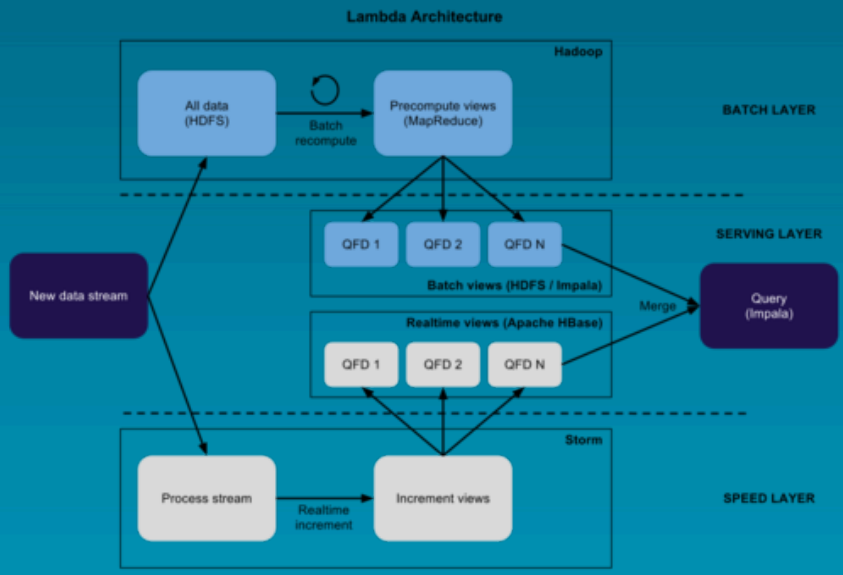
Real-Time
Online
Streaming

Why Don't We Have Both?

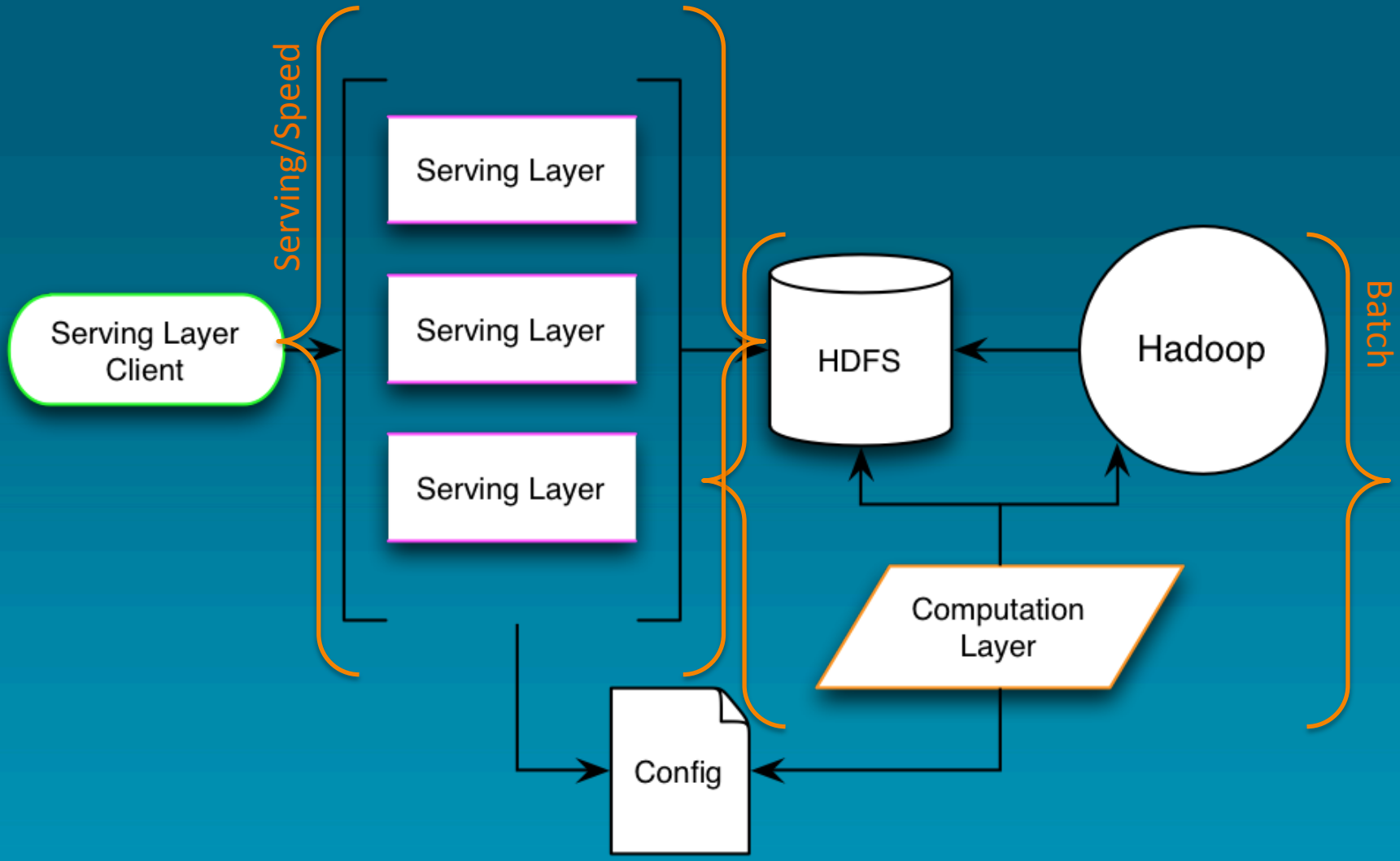
λ!

Lambda Architecture

- Batch, Stream Processing are different
- Tackle separately in 2+ Layers
- **Batch** Layer: offline, asynchronous
- **Serving** / **Speed** Layer: real-time, incremental, approximate



jameskinley.tumblr.com/post/37398560534/the-lambda-architecture-principles-for-architecting


Serving/Speed
Serving Layer
Serving Layer
Serving Layer
Serving Layer Client
HDFS
Hadoop
Batch
Computation Layer
Config

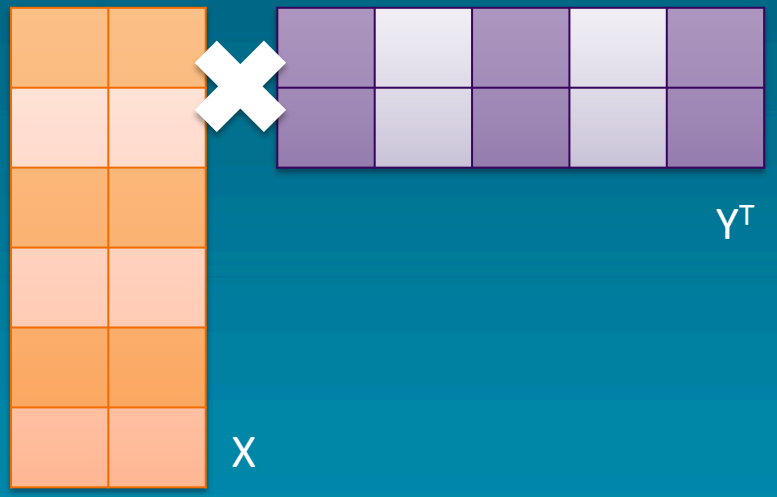
Two Layers

- **Computation Layer**
  - Java-based server process
  - Client of Hadoop 2.x
  - Periodically builds "generation" from recent data and past model
  - Baby-sits MapReduce* jobs (or, locally in-core)
  - Publishes models
- **Serving Layer**
  - Apache Tomcat™-based server process
  - Consumes models from HDFS (or local FS)
  - Serves queries from model in memory
  - Updates from new input
  - Also writes input to HDFS
  - Replicas for scale

* Apache Spark later

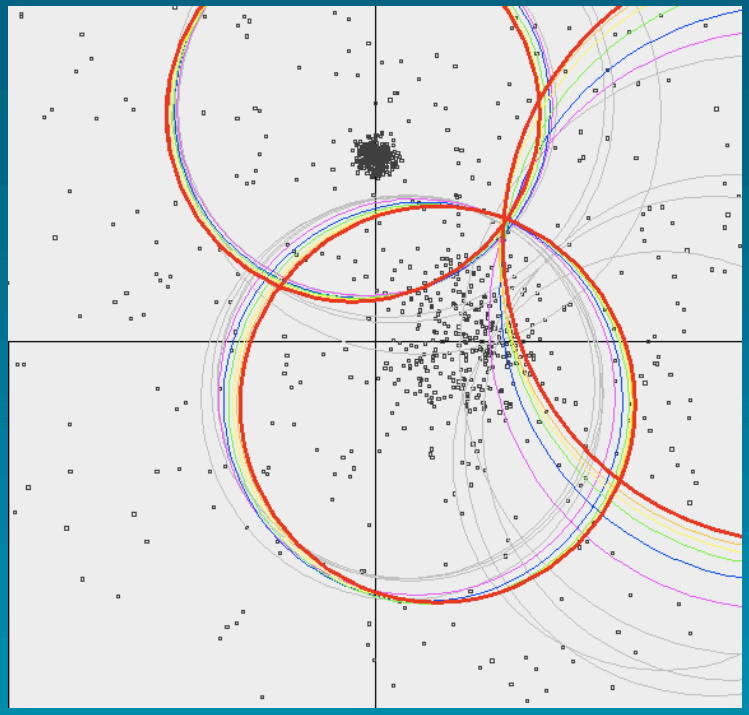
Collaborative Filtering : ALS

- **A**lternating **L**east **S**quares
- Latent-factor model
- Accepts implicit or explicit feedback
- Real-time update via fold-in of input
- No cold-start
- Parallelizable



Clustering : k-means++

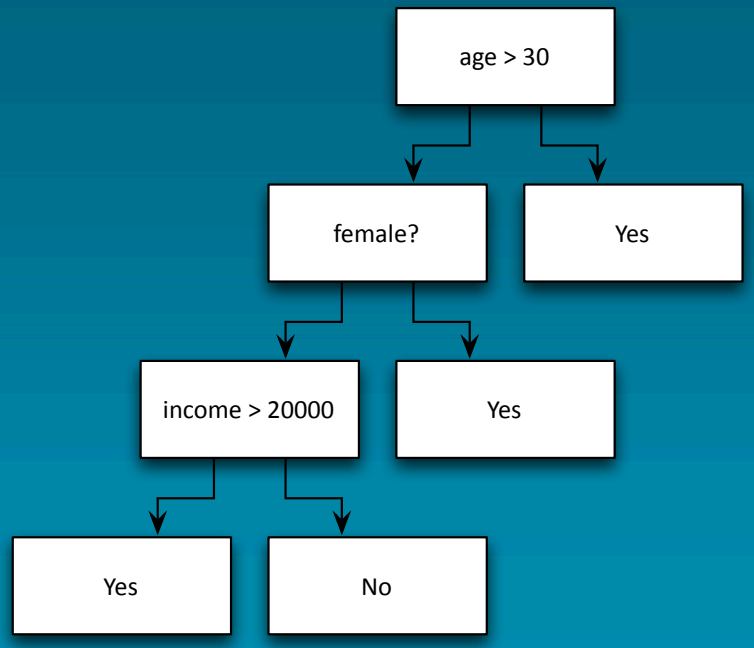
- Well-known and understood
- Parallelizable
- Clusters updateable



cwiki.apache.org/confluence/display/MAHOUT/K-Means+Clustering

Classification / Regression : RDF

- Random Decision Forests
- Ensemble method
- Numeric, categorical features and target
- Very parallel
- Nodes updateable
- Works well on many problems



PMML

- **P**redictive **M**odeling **M**arkup **L**anguage
- XML-based format for predictive models
- Standardized by Data Mining Group (`www.dmg.org`)
- Wide tool support

```
<PMML xmlns="http://www.dmg.org/PMML-4_1"
      version="4.1">
  <Header copyright="www.dmg.org"/>
  <DataDictionary numberOfFields="5">
    <DataField name="temperature"
               optype="continuous"
               dataType="double"/>
   …
  </DataDictionary>
  <TreeModel modelName="golfing"
             functionName="classification">
    <MiningSchema>
      <MiningField name="temperature"/>
      …
    </MiningSchema>
    <Node score="will play">
      <Node score="will play">
        <SimplePredicate field="outlook"
                         operator="equal"
                         value="sunny"/>
        …
      </Node>
    </Node>
  </TreeModel>
</PMML>
```

www.dmg.org/v4-1/TreeModel.html

Extra: Apache Spark as "Crossover Hit"

- Exploratory-friendly
  - REPL
  - Scala closures
  - MLlib
- Operational-friendly
  - Distributed
  - Hadoop integration
  - All Java libraries available

blog.cloudera.com/blog/2014/03/why-apache-spark-is-a-crossover-hit-for-data-scientists/

Thanks!

QCon

Please evaluate my talk via the mobile app!