



# How Conviva used Spark to Speed Up Video Analytics by 30x

Dilip Antony Joseph (@DilipAntony)

# Conviva monitors and optimizes online video for premium content providers



# What happens if you don't use Conviva?

PIAZZA AMPCAMP 2012 Q & A Course Page Student

+ New Post Search or add a post...

Note History: Mouse over icons to learn their significance. [Click here](#) to turn tooltips off. Or, use the menu at the top right.

↑ Click the Q&A button above to go back to the home screen


note


Crap, the online stream just totally stopped on me..


edit save to favorites 0 more


followup discussions, for lingering questions and comments


Resolved Unresolved


 **william waddington** (5 hours ago) - ditto

 **lee carraher** (5 hours ago) - ditto

 **David Drum** (5 hours ago) - Same here.

 **Sundeep Reddy** (5 hours ago) - same here

 **Simon Kwoczek** (5 hours ago) - and here

 **Will Harkrider** (5 hours ago) - me too

Just came back for me after retrying the connection.  
[mms://media.citris.berkeley.edu/amp](https://media.citris.berkeley.edu/amp)

Excellent talks thus far. Thank you. 5

★ 1 Spark compared to Pig 6

In Matel's PageRank example is the ... 2

my media.citris.berkeley.edu feed just ... S

Crap, the online stream just totally sto... S

SparkContext("local[TOOMANY]", "Uh... S

Can u Pls enlarge screen fonts!!! i

sbt/sbt compile fails, build is attemptin... S

JVM has some start up delays in gener... S

Scala Question: "class" v. "case class" i

Getting started with spark i

★ 1 Spark determination of 'insufficient' m... i

★ 3 Downsides of Spark S

Integrate Spark+Mesos with CloudBio... S

• 1 Unresolved Followup

★ 2 Common Crawl for Exercises i

If the speakers could repeat the questi... i

★ 2 How Spark work in parallel? i

Streaming quality S

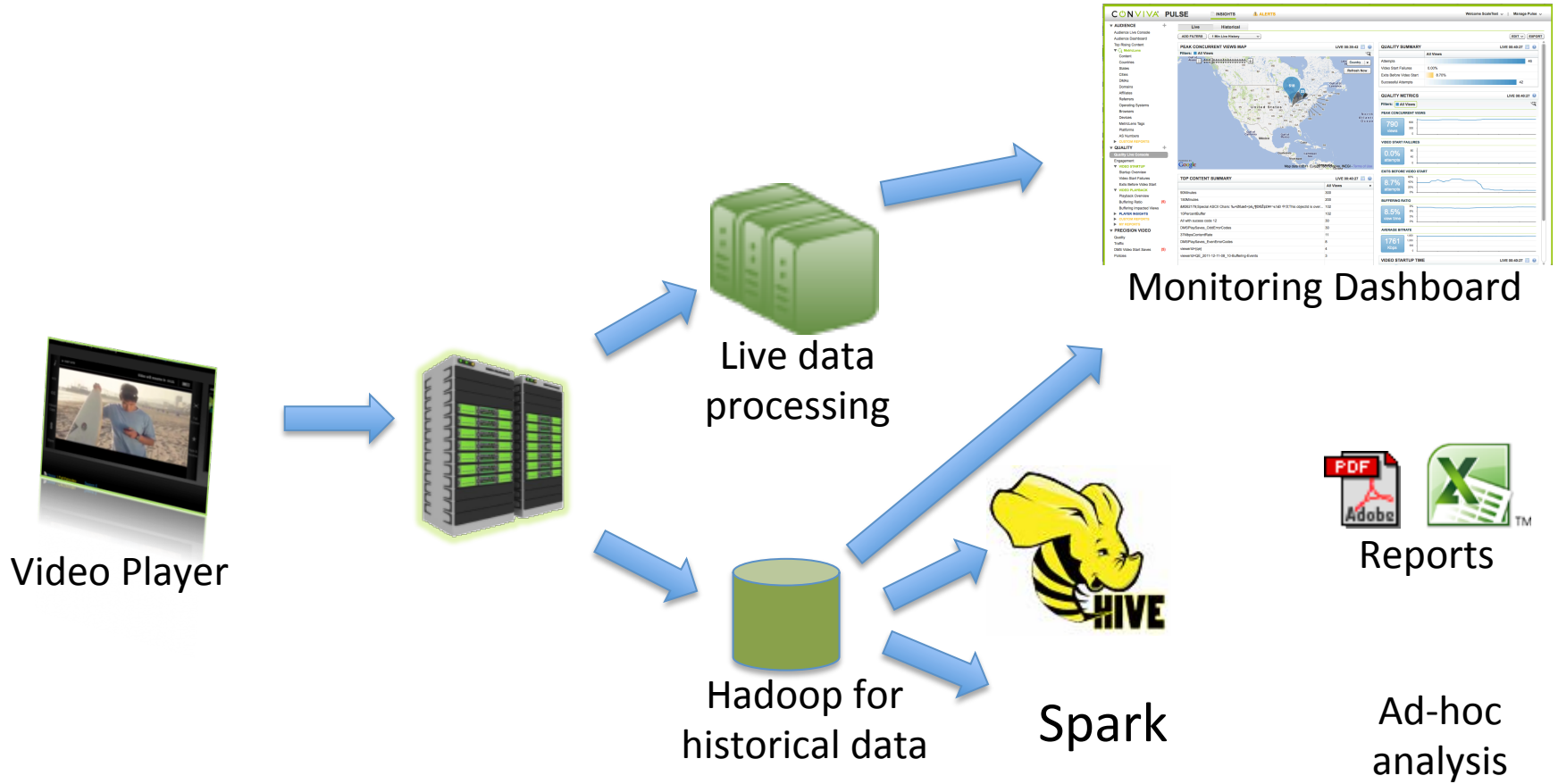
★ 1 What happens when data doesn't fit in... S

★ 1 Will we incur any fees to our Amazon E... S

how many online attendees are there? S

65 million video streams a day

# Conviva data processing architecture



# Group By queries dominate our Reporting workload



```
SELECT videoName, COUNT(1)
FROM summaries
WHERE date='2012_08_22' AND customer='XYZ'
GROUP BY videoName;
```

10s of metrics, 10s of group bys

# Group By query in Spark

```
val sessions = sparkContext.sequenceFile[SessionSummary, NullWritable](
    pathToSessionSummaryOnHdfs,
    classOf[SessionSummary], classOf[NullWritable])
    .flatMap {
        case (key, val) => val.fieldsOfInterest
    }

val cachedSessions = sessions.filter(
    whereConditionToFilterSessionsForTheDesiredDay)
    .cache

val mapFn : SessionSummary => (String, Long) = { s => (s.videoName, 1) }

val reduceFn : (Long, Long) => Long = { (a,b) => a+b }

val results = cachedSessions.map(mapFn).reduceByKey(reduceFn).collectAsMap
```

Spark is 30x faster than Hive

**45 minutes versus 24 hours  
for weekly Conviva Geo Report**

# How much data are we talking about?

- 150 GB / week of compressed summary data
- Compressed ~ 3:1
- Stored as Hadoop Sequence Files
- Custom binary serialization format

Spark is faster because it avoids reading data from disk multiple times



### Group By Country

Read from HDFS  
Decompress  
Deserialize

### Group By State

Read from HDFS  
Decompress  
Deserialize

### Group By Video

Read from HDFS  
Decompress  
Deserialize

10s of Group Bys ...

Cache only  
columns of  
interest

Hive/  
MapReduce  
startup  
overhead

Overhead of  
flushing  
intermediate  
data to disk

## Spark

### Group By Country

Read from HDFS  
Decompress  
Deserialize  
Cache data in memory

### Group By State

Read data from memory

### Group By Video

Read data from memory

10s of Group Bys ...

# Why not <some other big data system>?

- Hive
- Mysql
- Oracle/SAP HANA
- Column oriented dbs

# Spark just worked for us

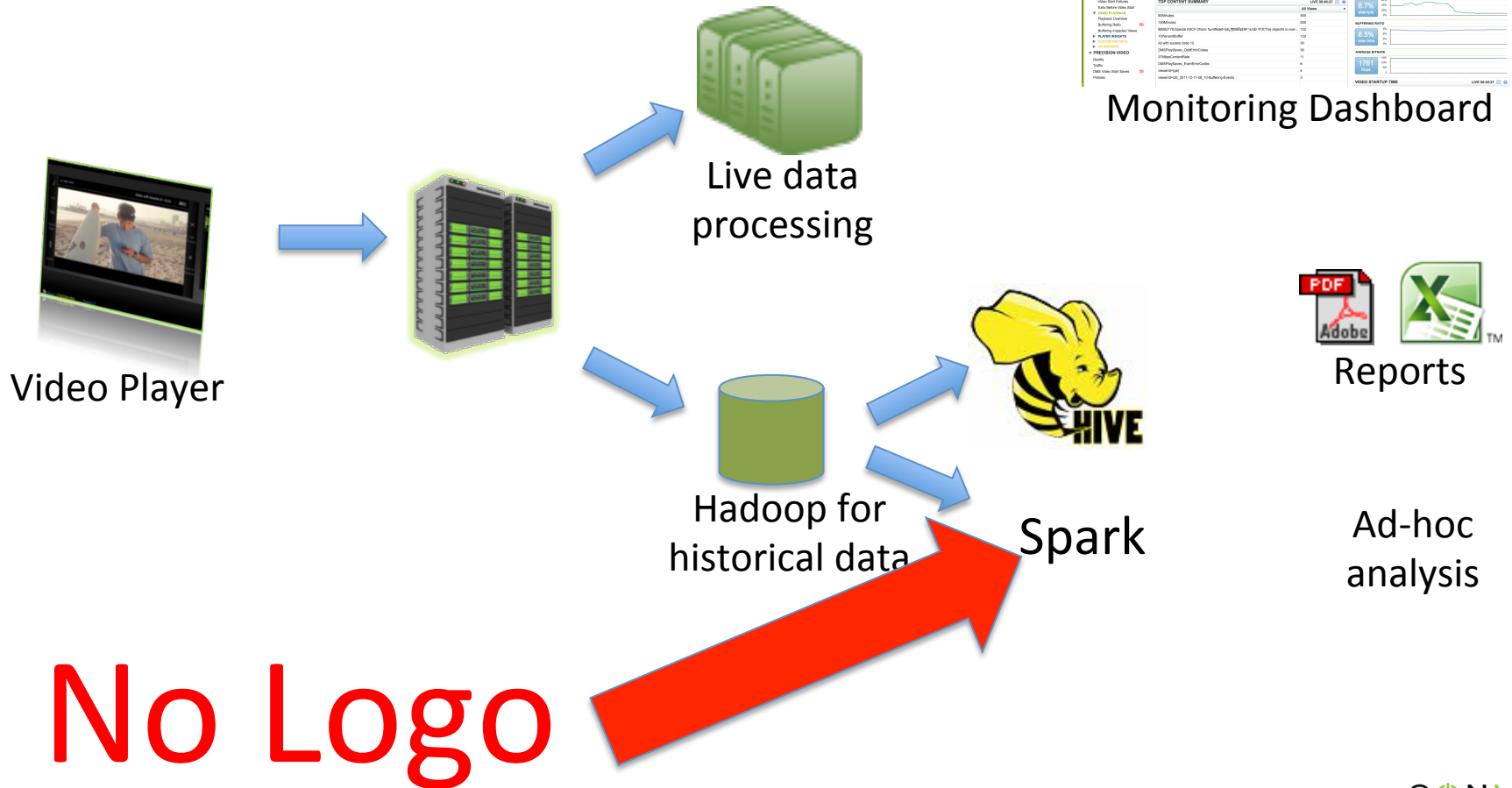
- 30x speed-up
- Great fit for our report computation model
- Open-source
- Flexible
- Scalable

30% of our reports use Spark

# We are working on more projects that use Spark

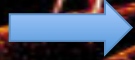
- Streaming Spark for unifying batch and live computation
- **SHadoop** – Run existing Hadoop jobs on Spark

# Problems with Spark





Video Player



Live data processing



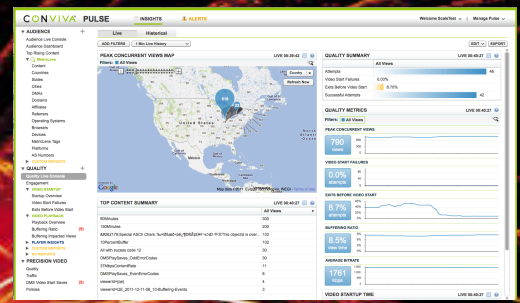
Hadoop



Spark



Reports



Monitoring Dashboard

# Spark queries are not very succinct

```
SELECT videoName, COUNT(1)
FROM summaries
WHERE date='2012_08_22' AND customer='XYZ'
GROUP BY videoName;
```



Spark

```
val sessions = sparkContext.sequenceFile[SessionSummary, NullWritable](
    pathToSessionSummaryOnHdfs,
    classOf[SessionSummary], classOf[NullWritable])
    .flatMap {
        case (key, val) => val.fieldsOfInterest
    }

val cachedSessions = sessions.filter(
    whereConditionToFilterSessionsForTheDesiredDay)
    .cache

val mapFn : SessionSummary => (String, Long) = { s => (s.videoName, 1) }

val reduceFn : (Long, Long) => Long = { (a,b) => a+b }

val results = cachedSessions.map(mapFn).reduceByKey(reduceFn).collectAsMap
```

# There is a learning curve associated with Scala, but ...

- Type-safety offered by Scala is a great boon
  - Code completion via Eclipse Scala plugin
- Complex queries are easier in Scala than in Hive
  - Cascading IF()s in Hive
- No need to write Scala in the future
  - Shark
  - Java, Python APIs

# Additional Hiccups

- Always on the bleeding edge – getting dependencies right
- More maintenance/debugging tools required

Spark has been working great for  
Conviva for over a year

We are Hiring

[jobs@conviva.com](mailto:jobs@conviva.com)

[http://www.conviva.com/blog/engineering/  
using-spark-and-hive-to-process-bigdata-at-  
conviva](http://www.conviva.com/blog/engineering/using-spark-and-hive-to-process-bigdata-at-conviva)