Spark and Spark Streaming @ Netflix

Kedar Sadekar & Monal Daxini

Mission

- Enable rapid pace of innovation for Algorithm Engineers

- Business Value – More A/B tests

Experiments

Users with plays

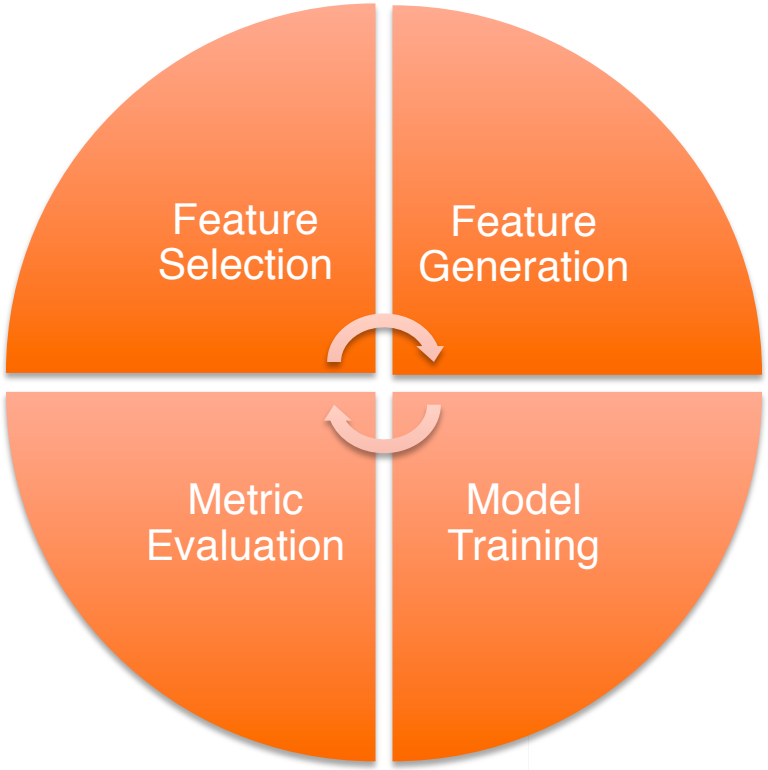
Feature selection

Large sample size

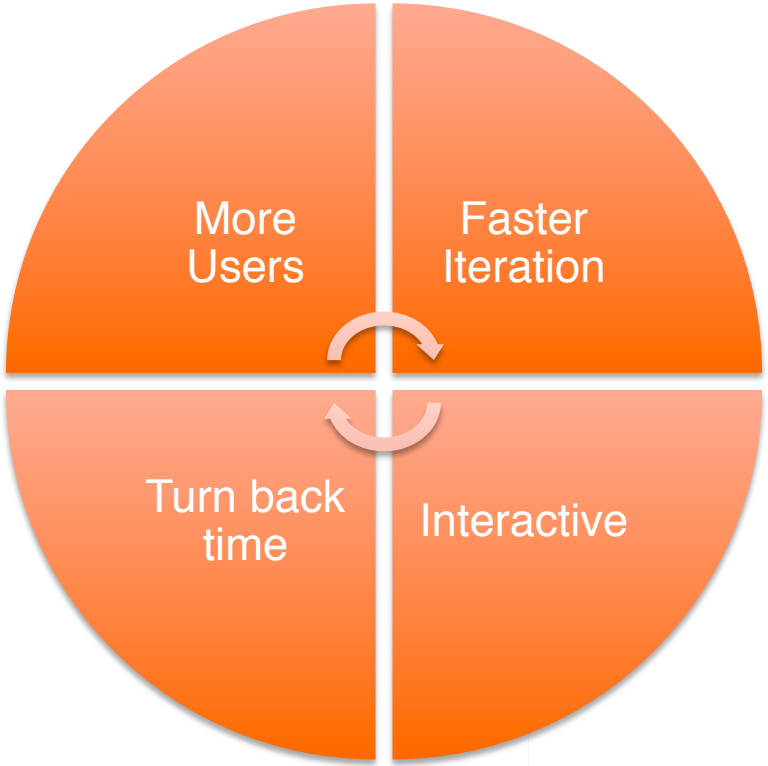
Turn back time

Multiple ideas

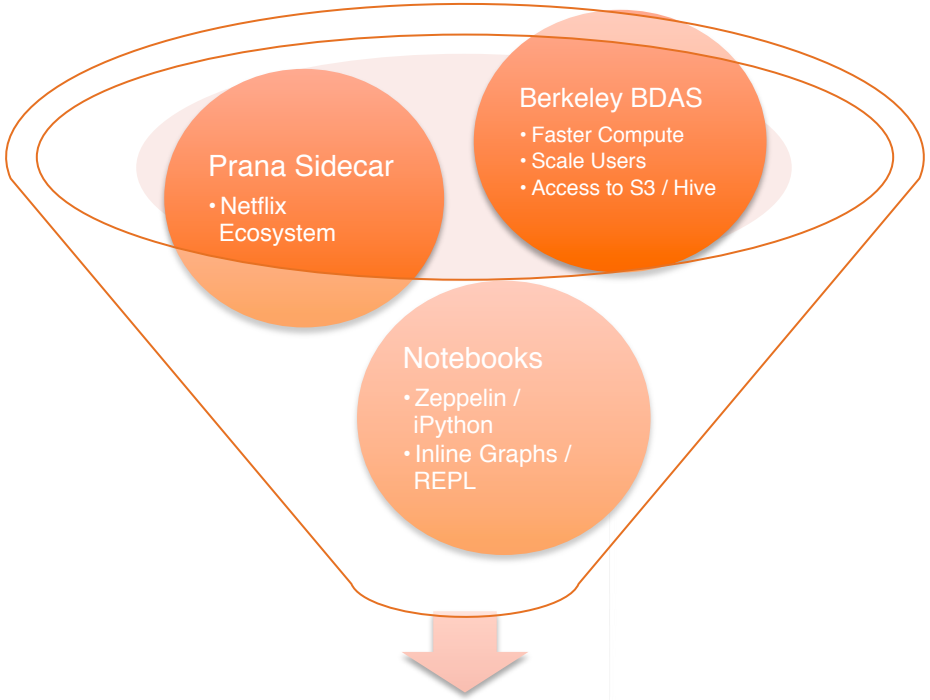
Use Cases


Feature Selection
Feature Generation
Metric Evaluation
Model Training

Use Cases



Solution – Netflix BDAS



Netflix BDAS

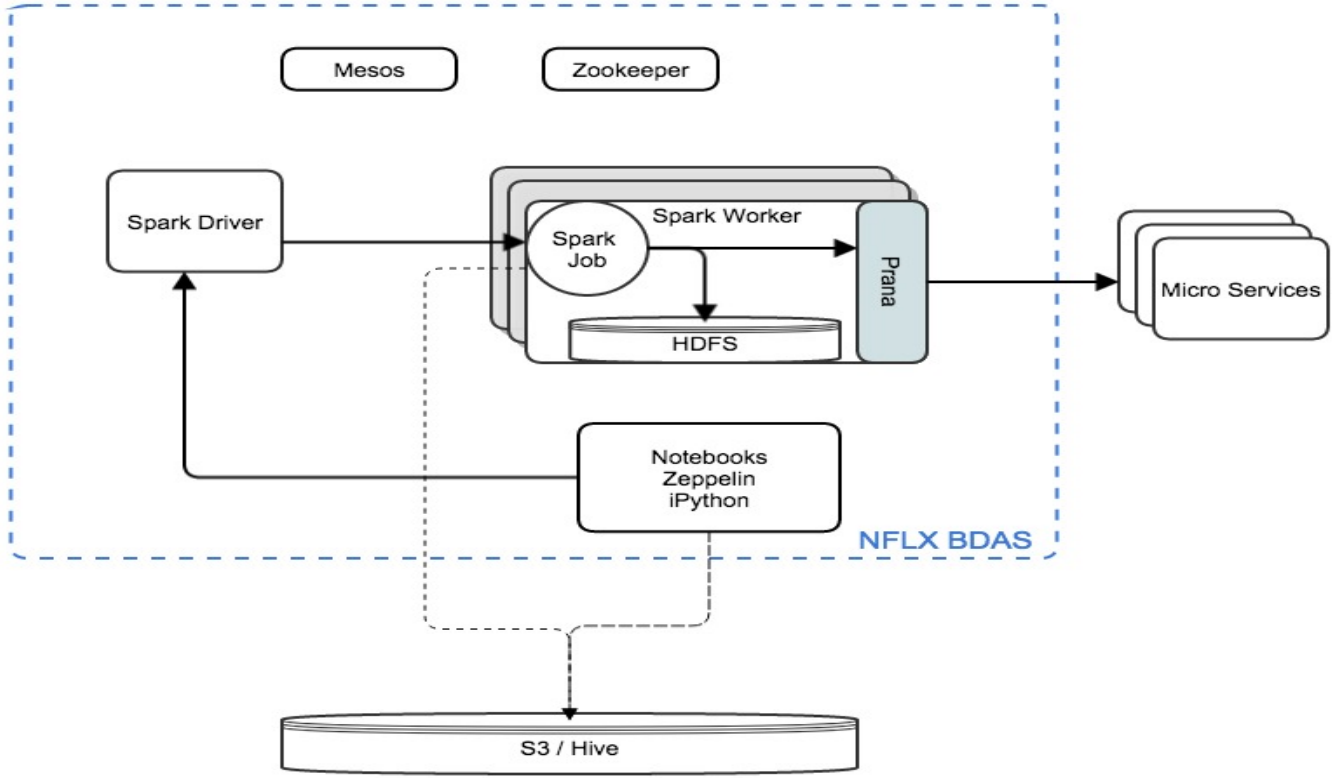
Netflix BDAS - Features

- Simplicity
  - Individual cluster
- Prana - Netflix ecosystem
  - Automatic configuration
  - Classloader isolation
  - Discovery & healthcheck

Netflix BDAS - Features

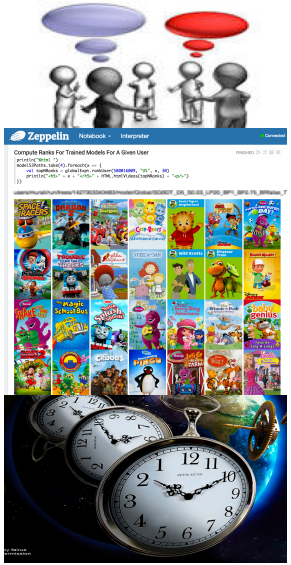
- Ad-hoc experimentation
- Time machine functionality
- Access to Hive data and micro services from single place
  - Access to multiple AWS buckets (S3)

Netflix BDAS – Sample Deployment



Wins

- 8X the number of users
- 5x - 9x faster
- Interactive
- Turn back time



Learnings

- Easy to bring down an online system
- Almost killed 1000's of ETL jobs
  - hive metastore update

- Too many systems and configuration
- Playing catch up with libraries and tools
  - Hadoop, iPython, Zeppelin

- Scala / Spark learning curve
- Debugging
  - files open, no resources etc.

Increased Adoption

Adoption increasing amongst teams

- Multiple Algorithmic Eng. teams
- Personalization Infrastructure
- Marketing
- Security
- A/B Test Engineering

Looking Ahead

- Spark-R / Dataframes support
- Multi-tenancy
  - Job specific configurations
- Debuggability
- Newer notebooks
- Spark Streaming
  - Lambda Architecture
  - Real-time algorithms (trending now)

Netflix Streaming Event Data Pipeline

Monal Daxini

Netflix Event Data Pipeline

Event Streams

Stream processing

Publish

Collect

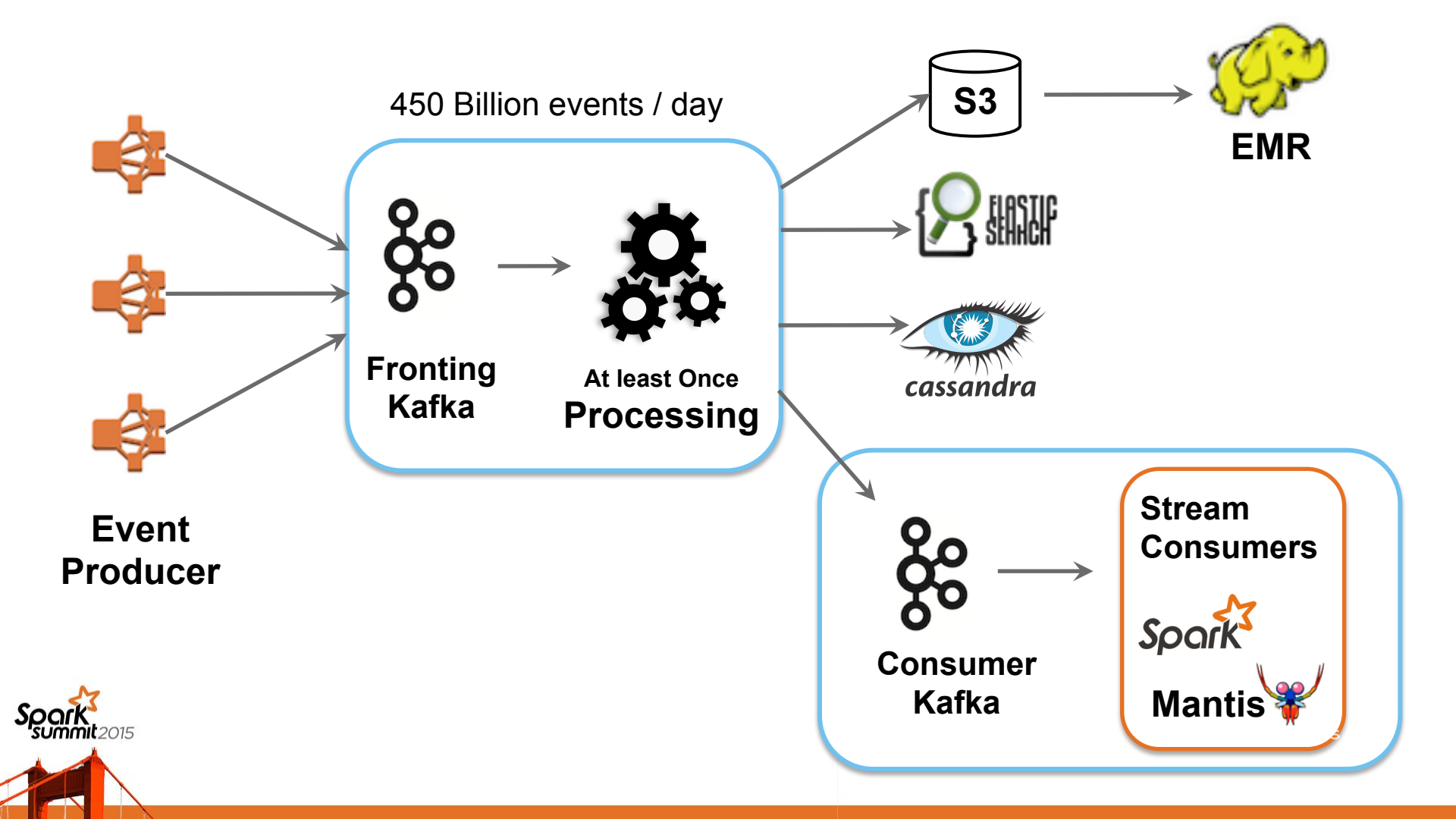
Process

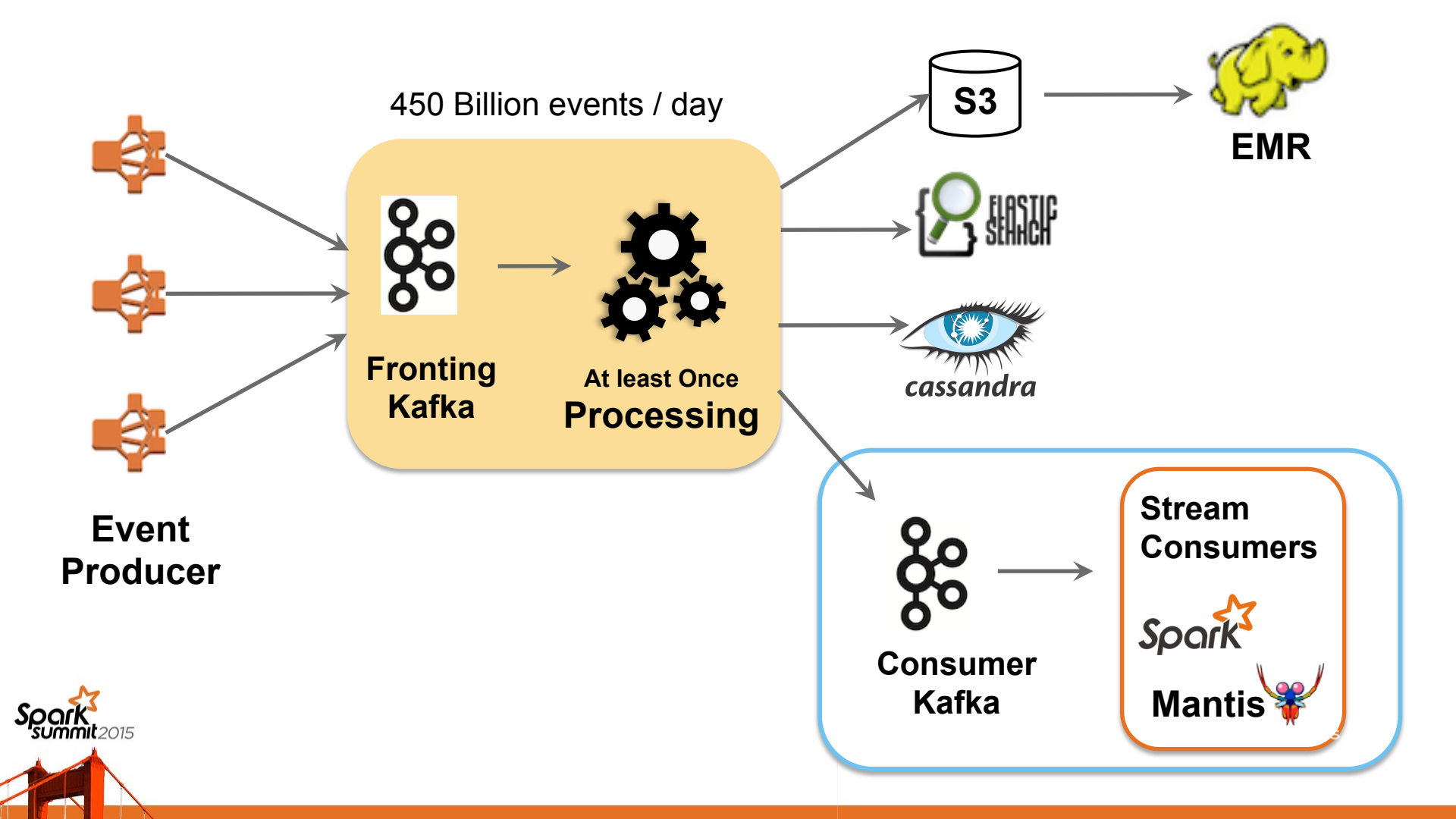
Move

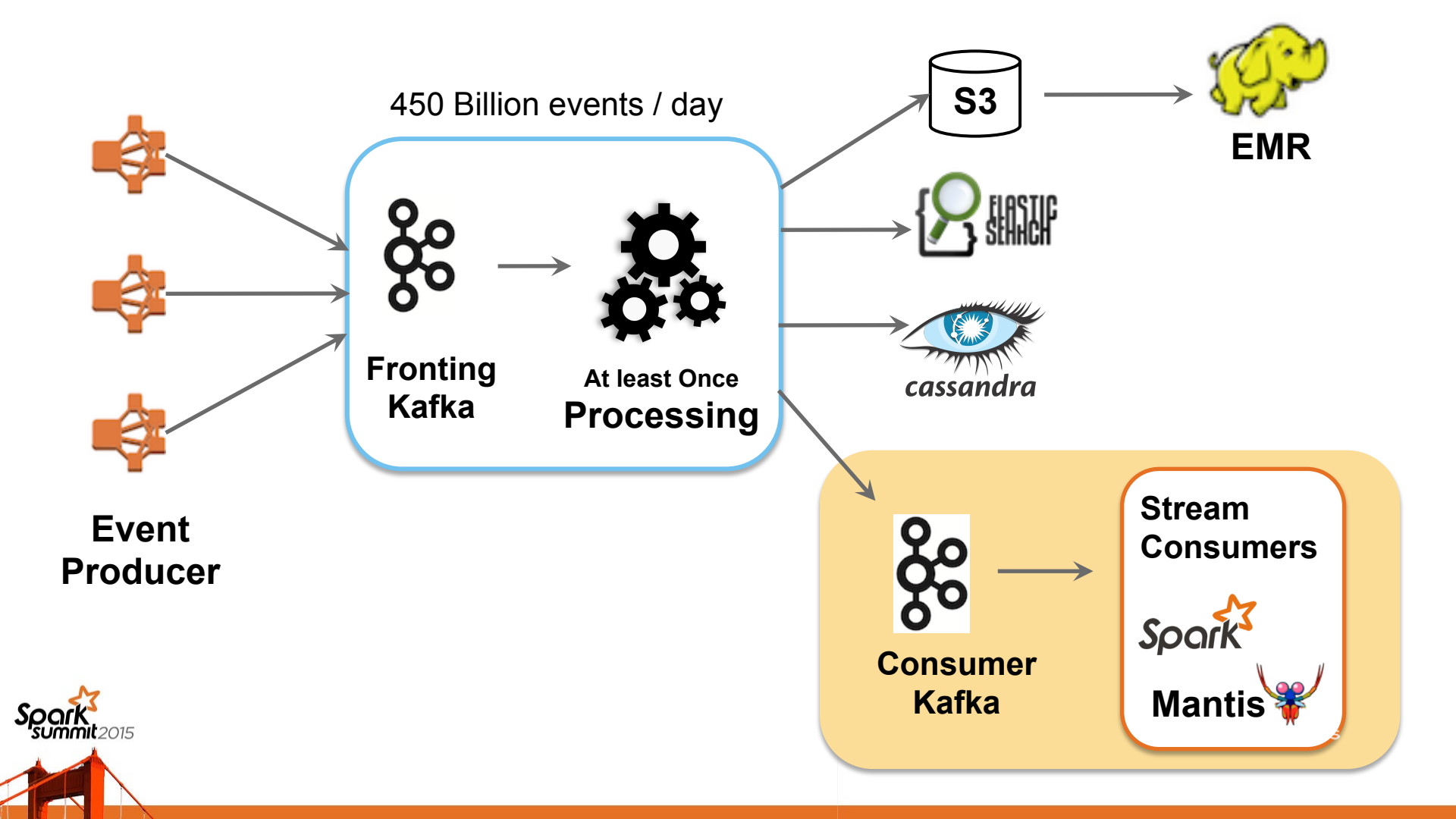
Events @ Cloud Scale

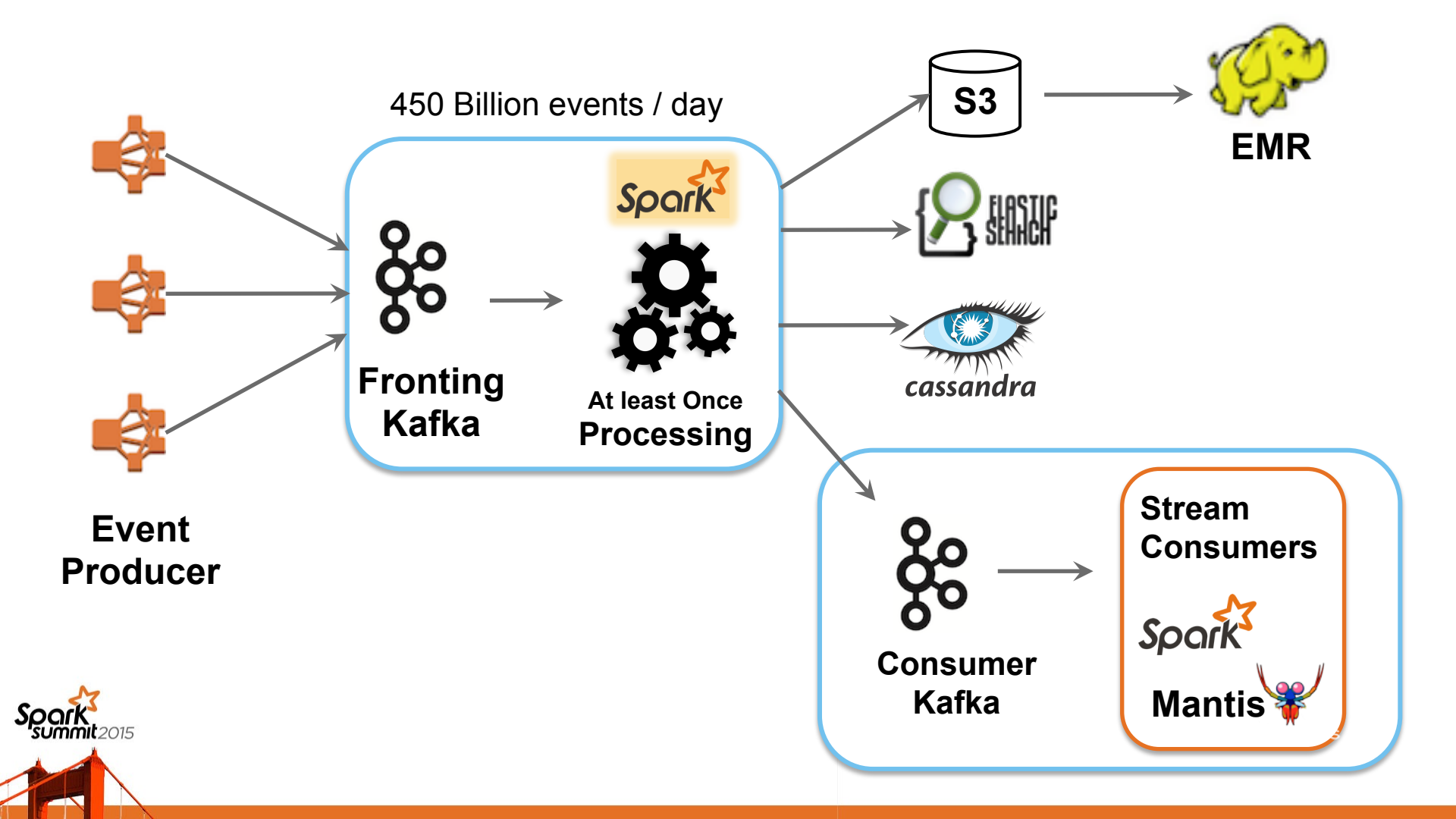
450 Billion Events per Day

8 Million (17 GB) per sec peak

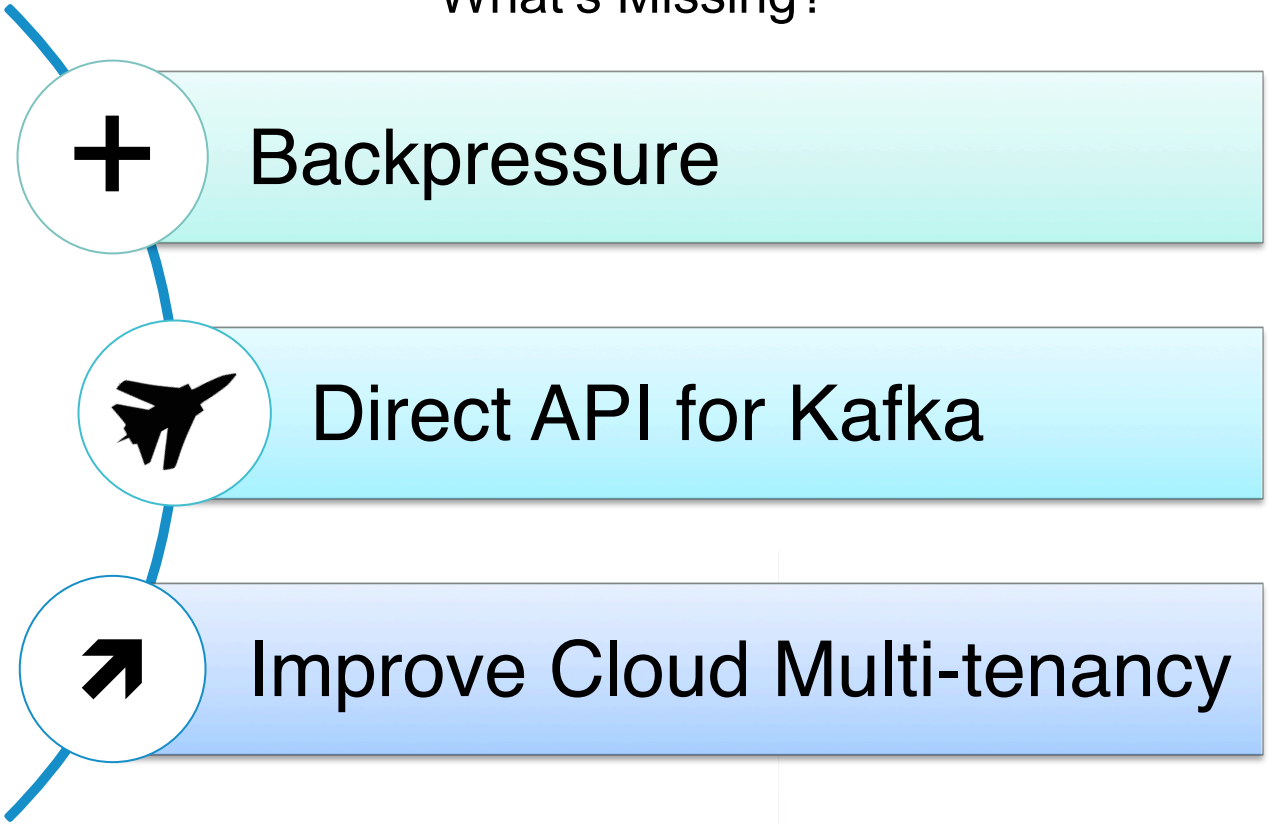

450 Billion events / day
S3
EMR
Fronting
Kafka
At least Once
Processing
cassandra
Event
Producer
Consumer
Kafka
Stream
Consumers
Spark
Mantis


450 Billion events / day
S3
EMR
Fronting Kafka
At least Once Processing
cassandra
Event Producer
Consumer Kafka
Stream Consumers
Spark
Mantis

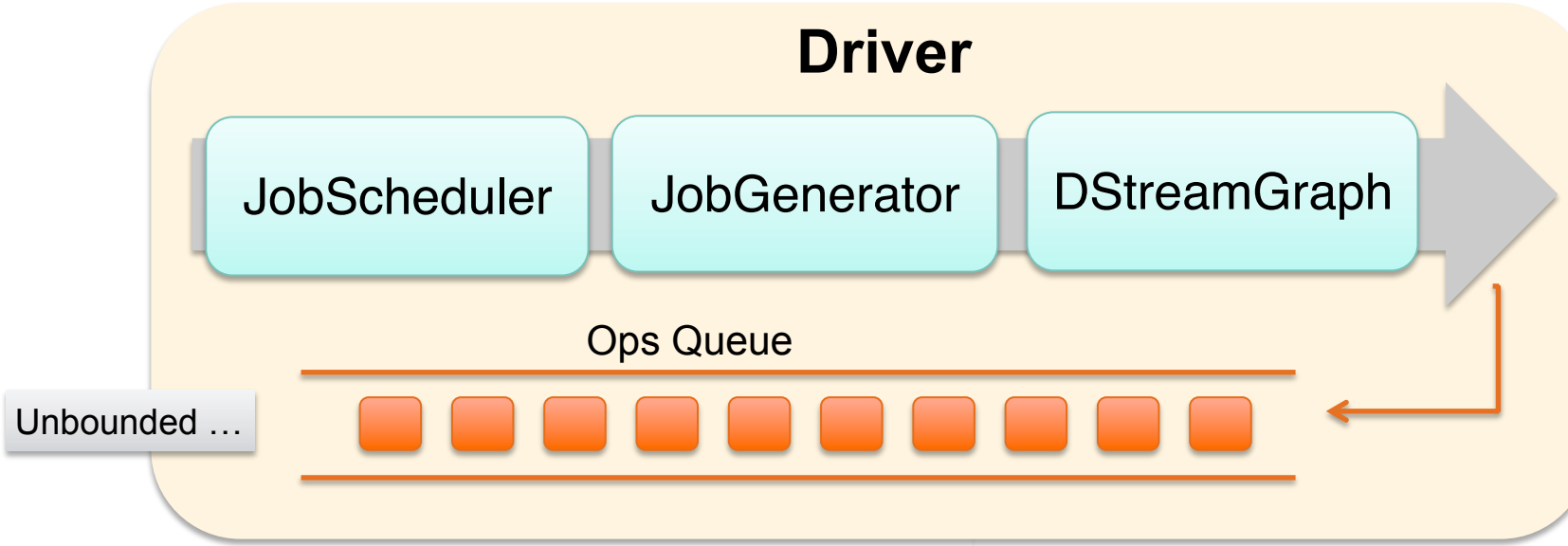

450 Billion events / day
S3
EMR
Fronting
Kafka
At least Once
Processing
cassandra
Event
Producer
Consumer
Kafka
Stream
Consumers
Spark
Mantis


450 Billion events / day
S3
EMR
Spark
ELASTIC SEARCH
cassandra
Fronting Kafka
At least Once Processing
Event Producer
Consumer Kafka
Stream Consumers
Spark
Mantis
Spark summit 2015

What's Missing?



+ Backpressure



00 : 10

Backpressure

SPARK-7398 – Add backpressure to Spark streaming

SPARK-6691 – Add dynamic rate limiter to Spark streaming

+ Backpressure

Backpressure implementation slated for Spark 1.5 release

Direct API for Kafka

Spark 1.3

Kafka Integration

2x
Faster

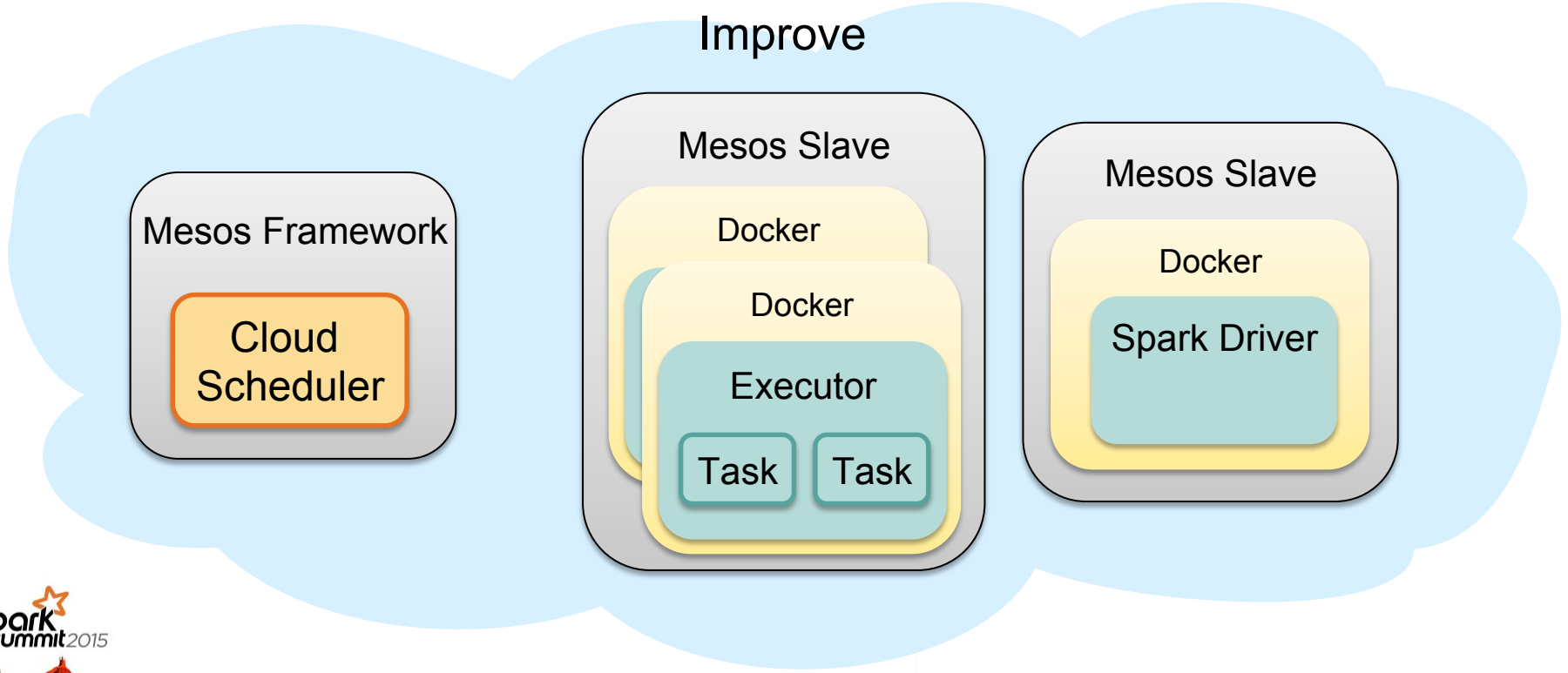
Spark 1.2
Receiver based
Kafka Integration

Direct API for Kafka

Enhance

- Prefetch messages
- Connection reuse (pooling)

Cloud Multi-tenancy



Next?

- Measuring Spark Streaming Latencies
- Spark Streaming Cloud Multi-tenancy