FIELD NOTES

FROM EXPEDITIONS IN THE CLOUD

DR. MATT WOOD
GM, PRODUCT STRATEGY

GOOD MORNING

MATTHEW@AMAZON.COM

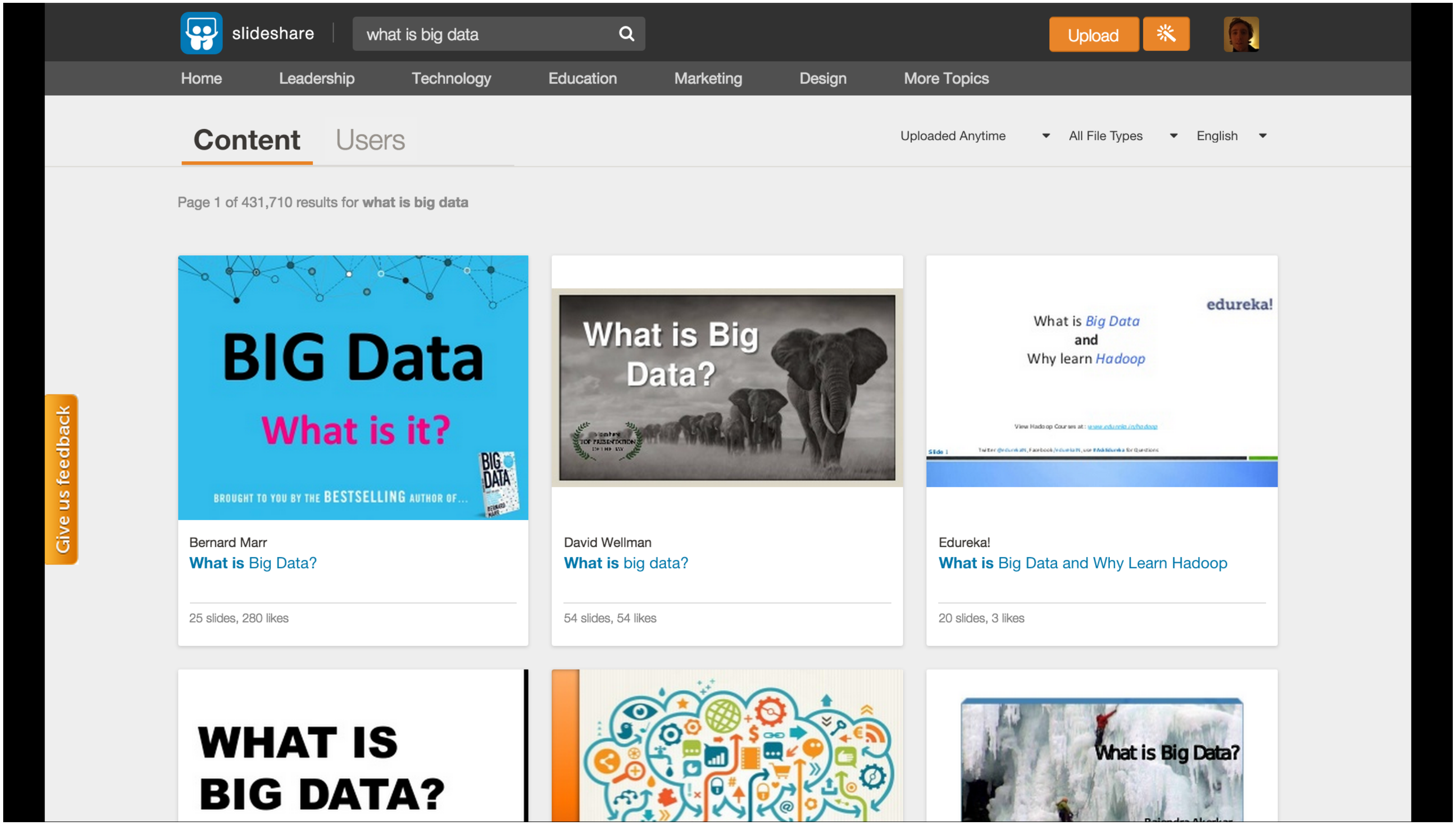
@MZA

THANK YOU

TECHNOLOGY ADOPTION

“WHAT IS CLOUD COMPUTING?”

"WHAT IS BIG DATA?"



Page 1 of 431,710 results for **what is big data**

Bernard Marr

What is Big Data?

25 slides, 280 likes

David Wellman

What is big data?

54 slides, 54 likes

Edureka!

What is Big Data and Why Learn Hadoop

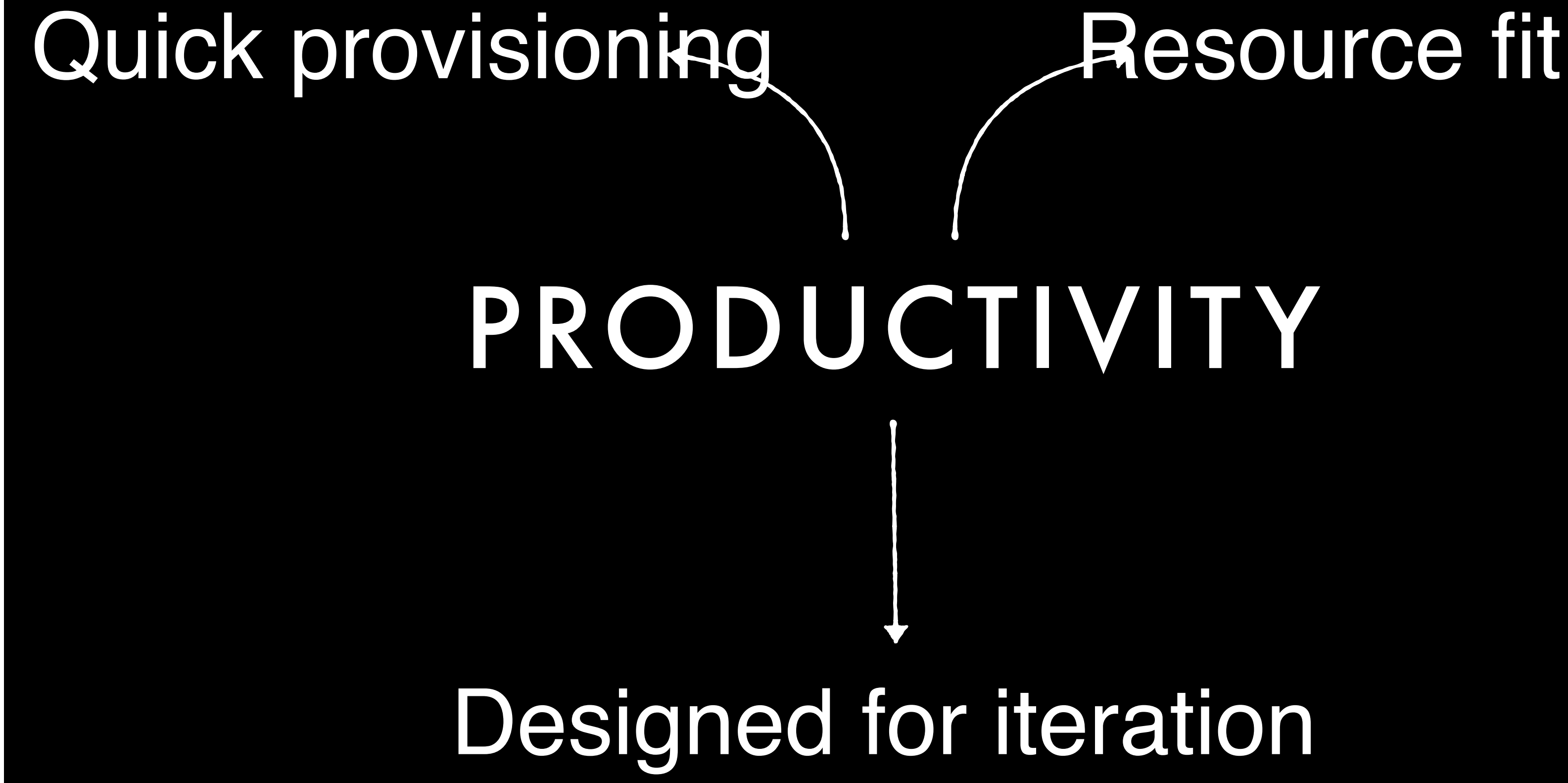
20 slides, 3 likes

VOLUME
VELOCITY
VARIABILITY

CONSTRAINTS

BOXED IN

~~CONSTRAINTS~~



Quick provisioning

Resource fit

PRODUCTIVITY

Designed for iteration

Canonical data stores

Streaming data

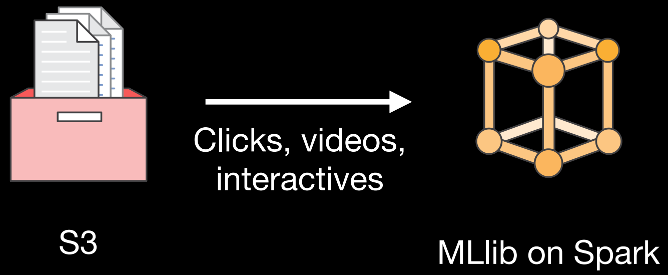
Task clusters

SPARK WITH AMAZON EMR

Production workloads on AWS

yelp	HEARST	The Washington Post	QUIXEY
Machine learning & ad targeting	Web analytics	Ad targeting & recommendations	App search
CROWDSTRIKE	gumgum	RADIUS®	krux
Security event streaming	Revenue forecasting	Predictive marketing	Personalization

The Washington Post

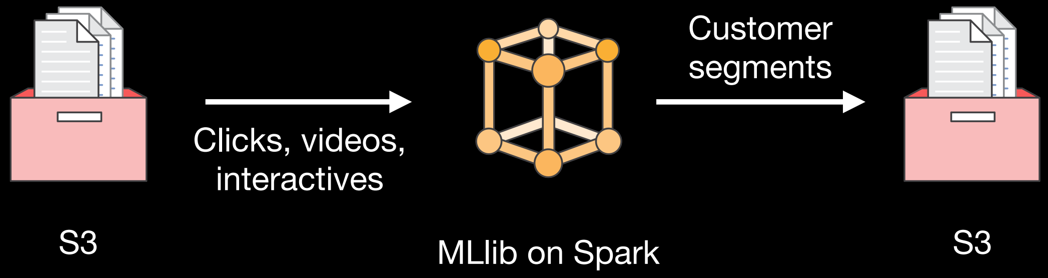


S3

Clicks, videos, interactives

MLlib on Spark

The Washington Post



S3

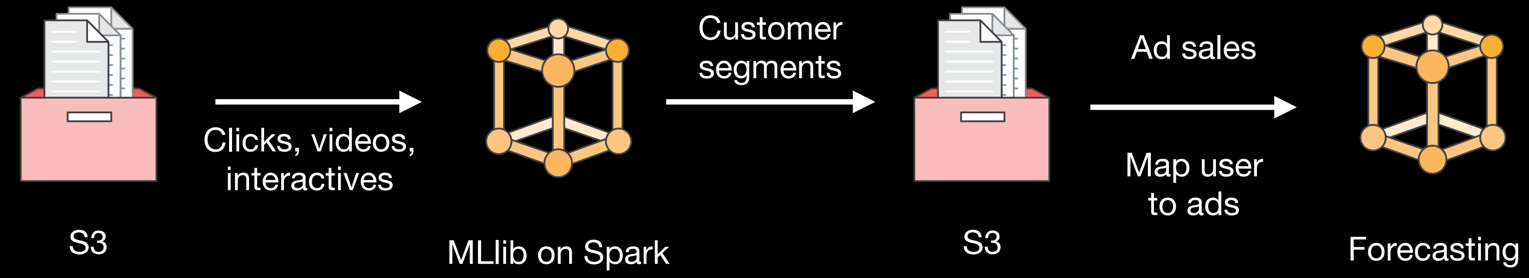
Clicks, videos, interactives

MLlib on Spark

Customer segments

S3

The Washington Post



S3

Clicks, videos, interactives

MLlib on Spark

Customer segments

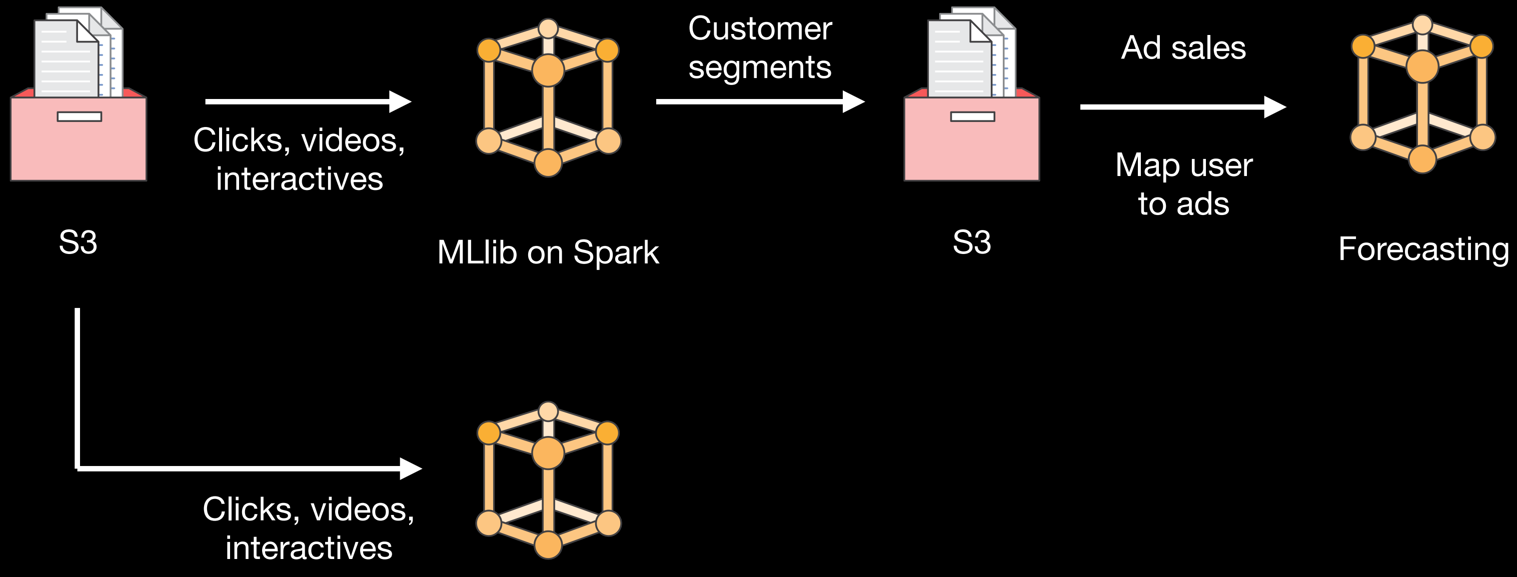
S3

Ad sales

Map user to ads

Forecasting

The Washington Post



S3

Clicks, videos, interactives

MLlib on Spark

Customer segments

S3

Ad sales

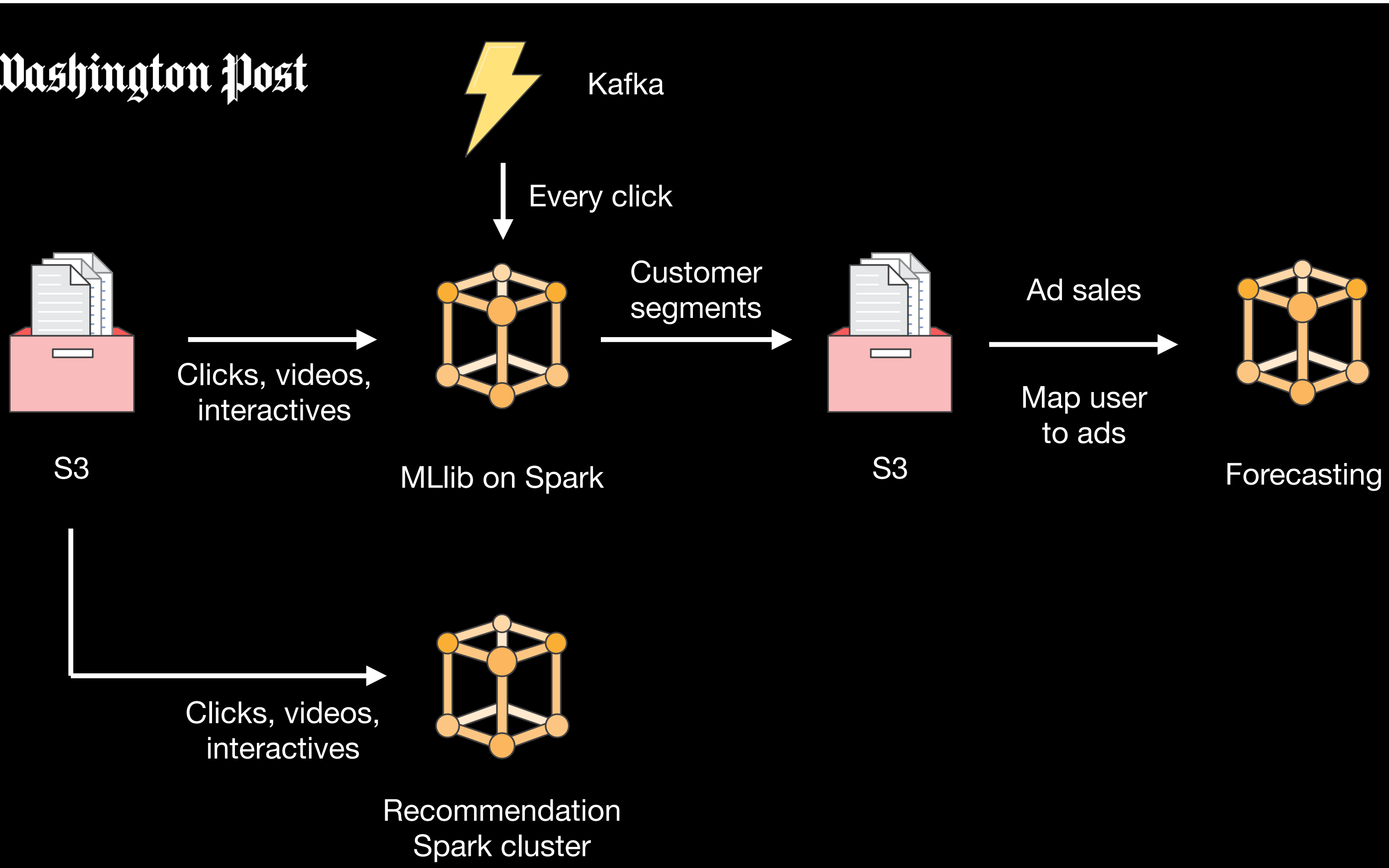
Map user to ads

Forecasting

Clicks, videos, interactives

Recommendation Spark cluster

The Washington Post



Kafka

Every click

S3

Clicks, videos, interactives

MLlib on Spark

Customer segments

S3

Ad sales

Map user to ads

Forecasting

Clicks, videos, interactives

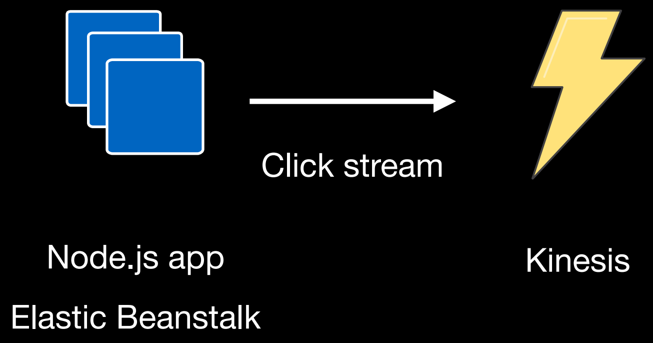
Recommendation Spark cluster

HEARST

Node.js app

Elastic Beanstalk

HEARST



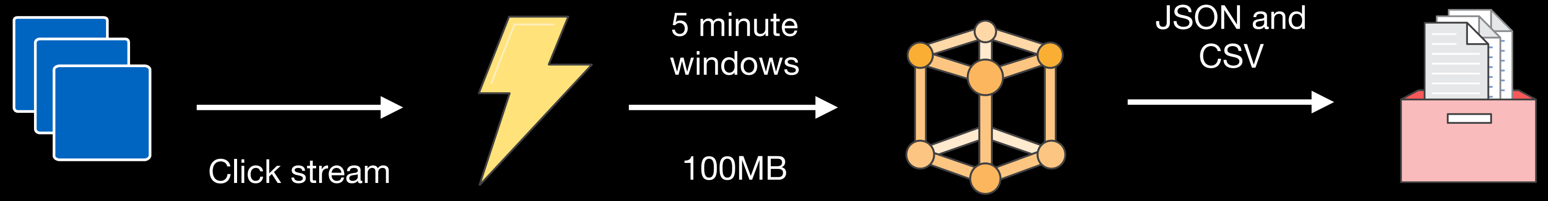
Click stream

Node.js app

Elastic Beanstalk

Kinesis

HEARST



Node.js app

Elastic Beanstalk

Click stream

Kinesis

5 minute windows

100MB

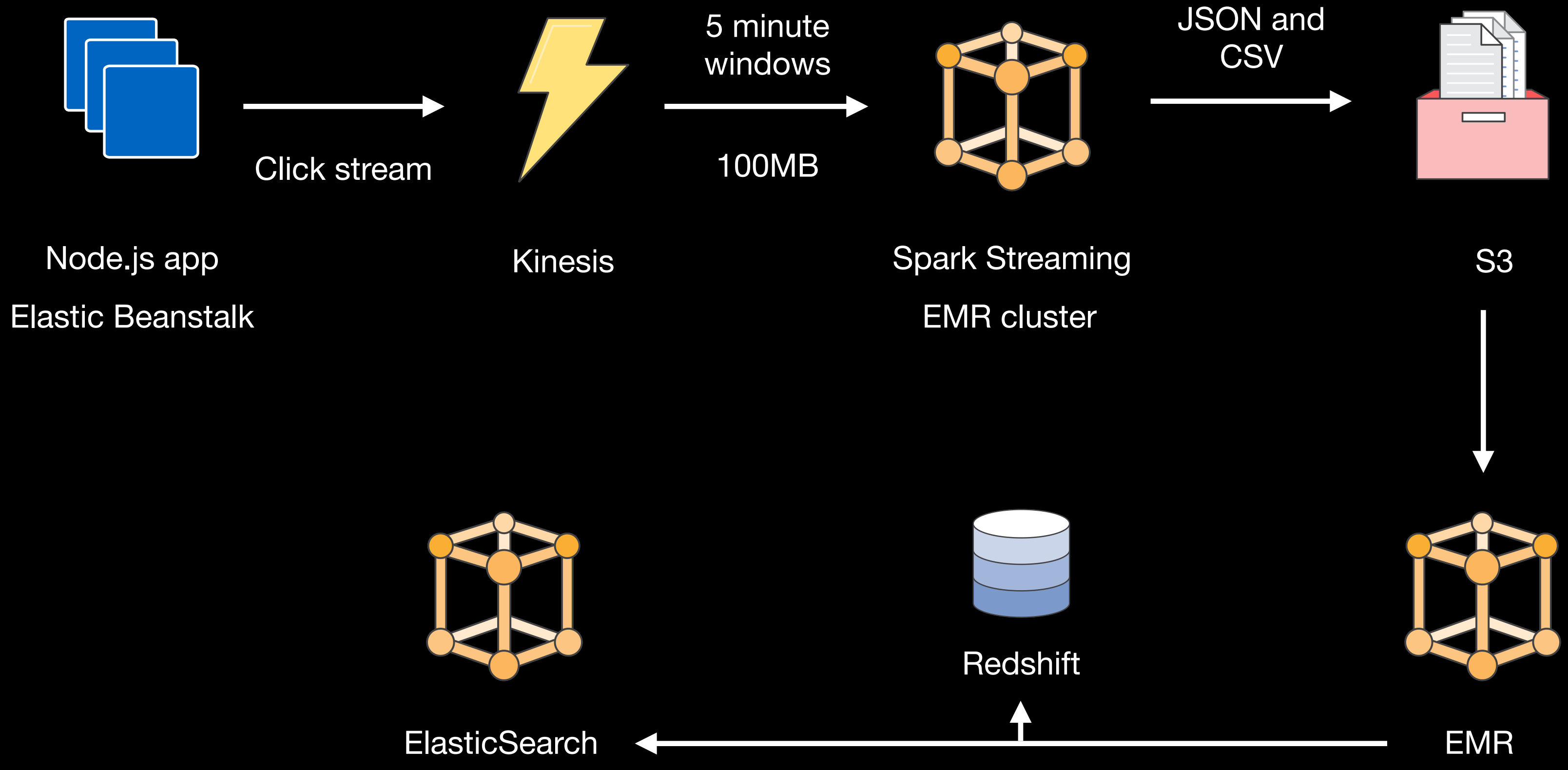
Spark Streaming

EMR cluster

JSON and CSV

S3

HEARST



Node.js app

Elastic Beanstalk

Click stream

Kinesis

5 minute windows

100MB

Spark Streaming

EMR cluster

JSON and CSV

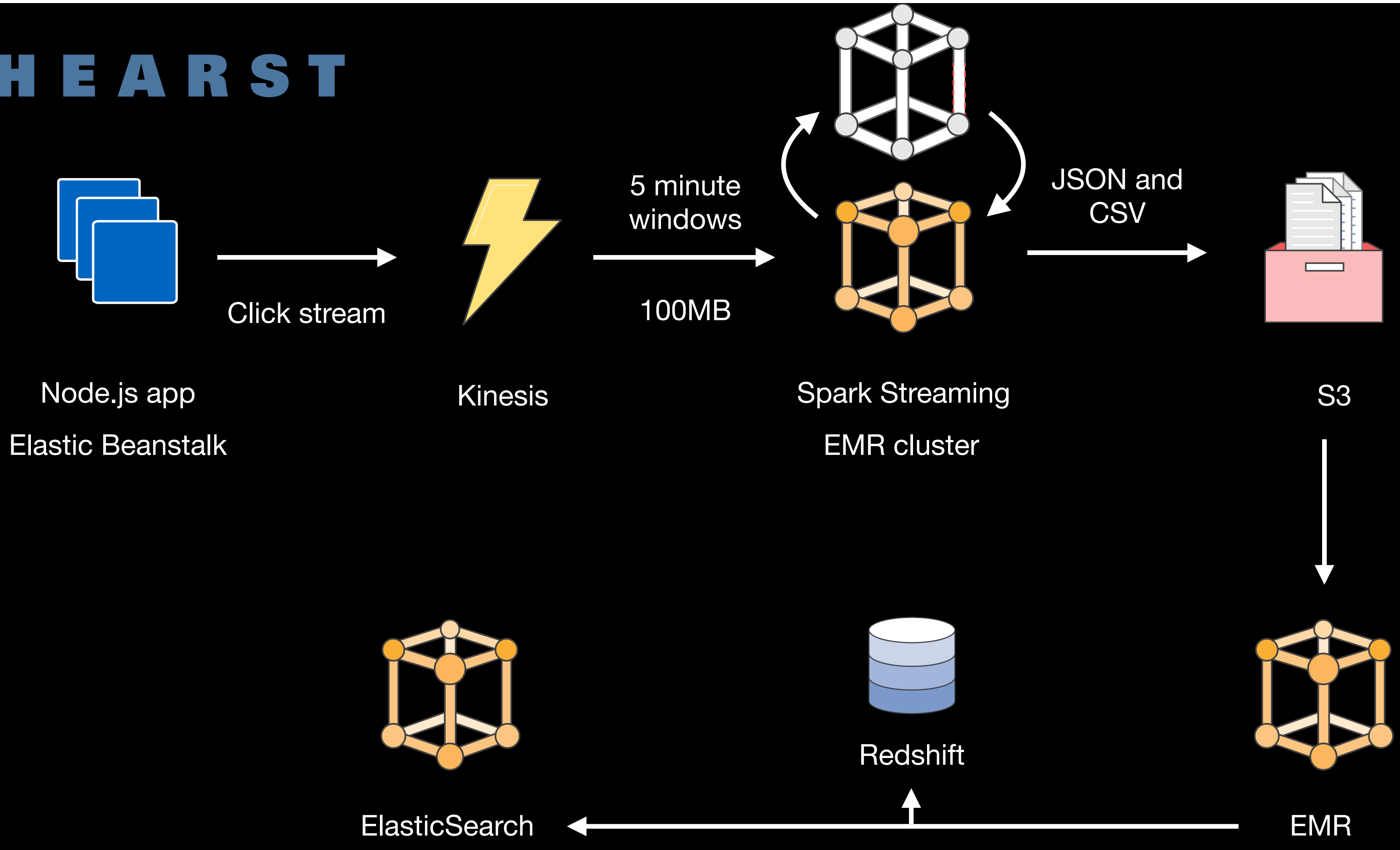
S3

EMR

Redshift

ElasticSearch

HEARST



Node.js app

Elastic Beanstalk

Click stream

Kinesis

5 minute windows

100MB

Spark Streaming

EMR cluster

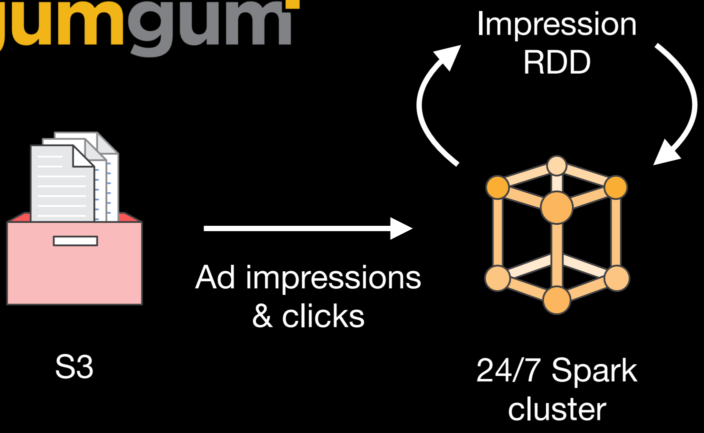
JSON and CSV

S3

EMR

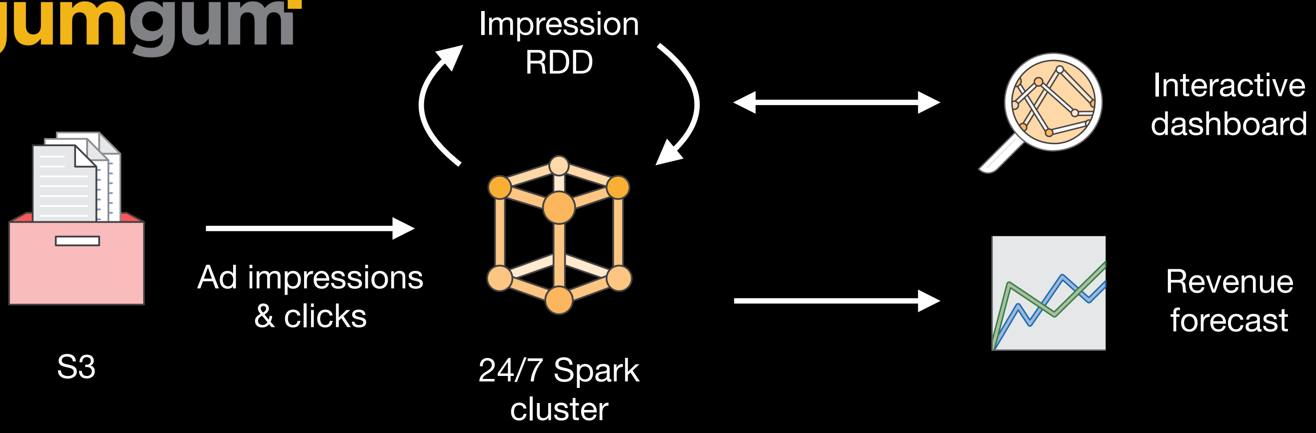
Redshift

ElasticSearch



Impression
RDD
S3
Ad impressions
& clicks
24/7 Spark
cluster

gumgum



Impression RDD

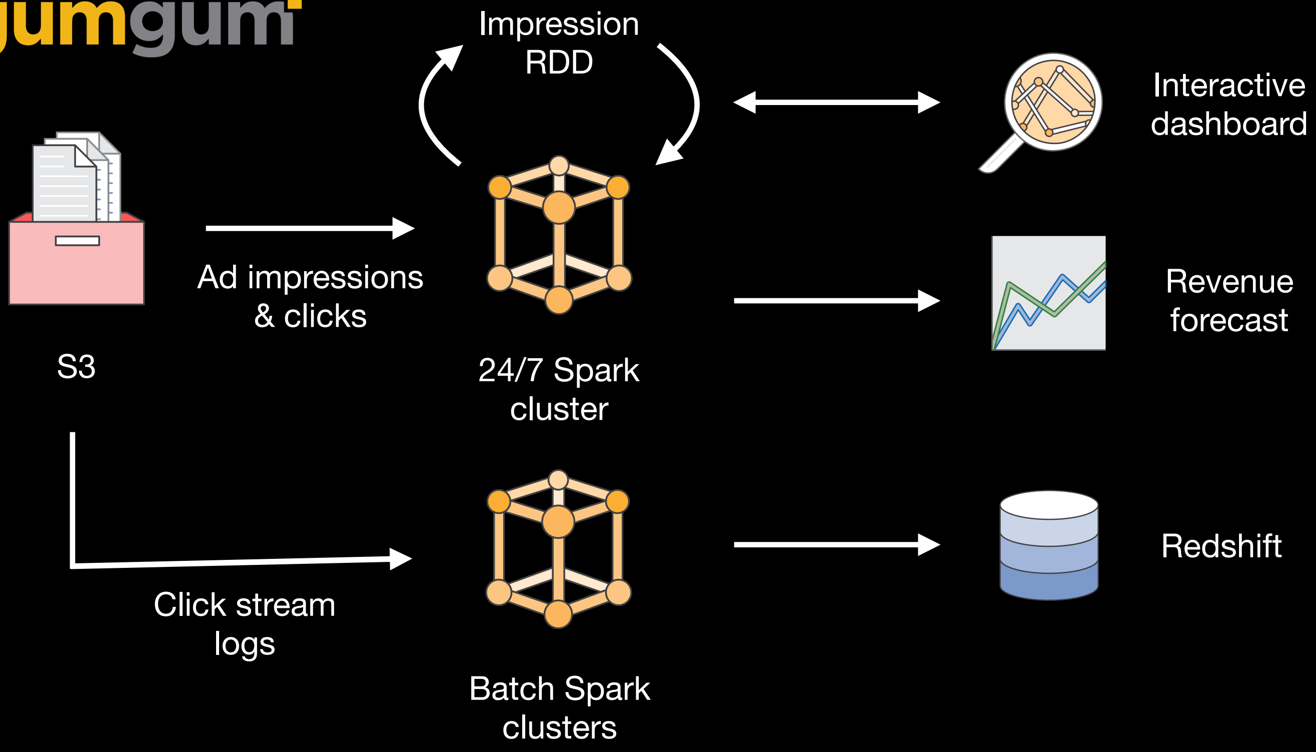
S3

Ad impressions & clicks

24/7 Spark cluster

Interactive dashboard

Revenue forecast



Impression RDD

S3

Ad impressions & clicks

24/7 Spark cluster

Click stream logs

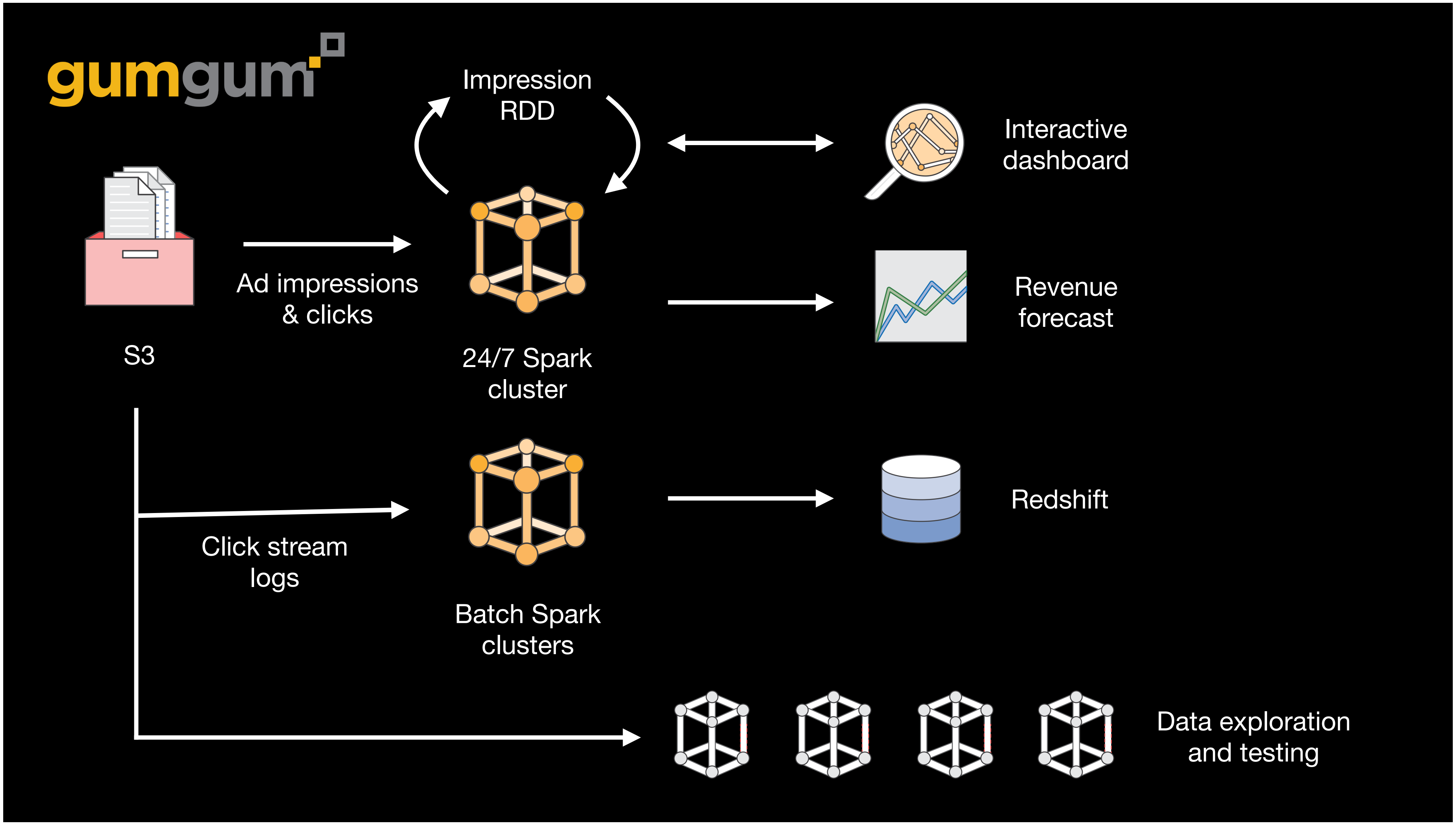
Batch Spark clusters

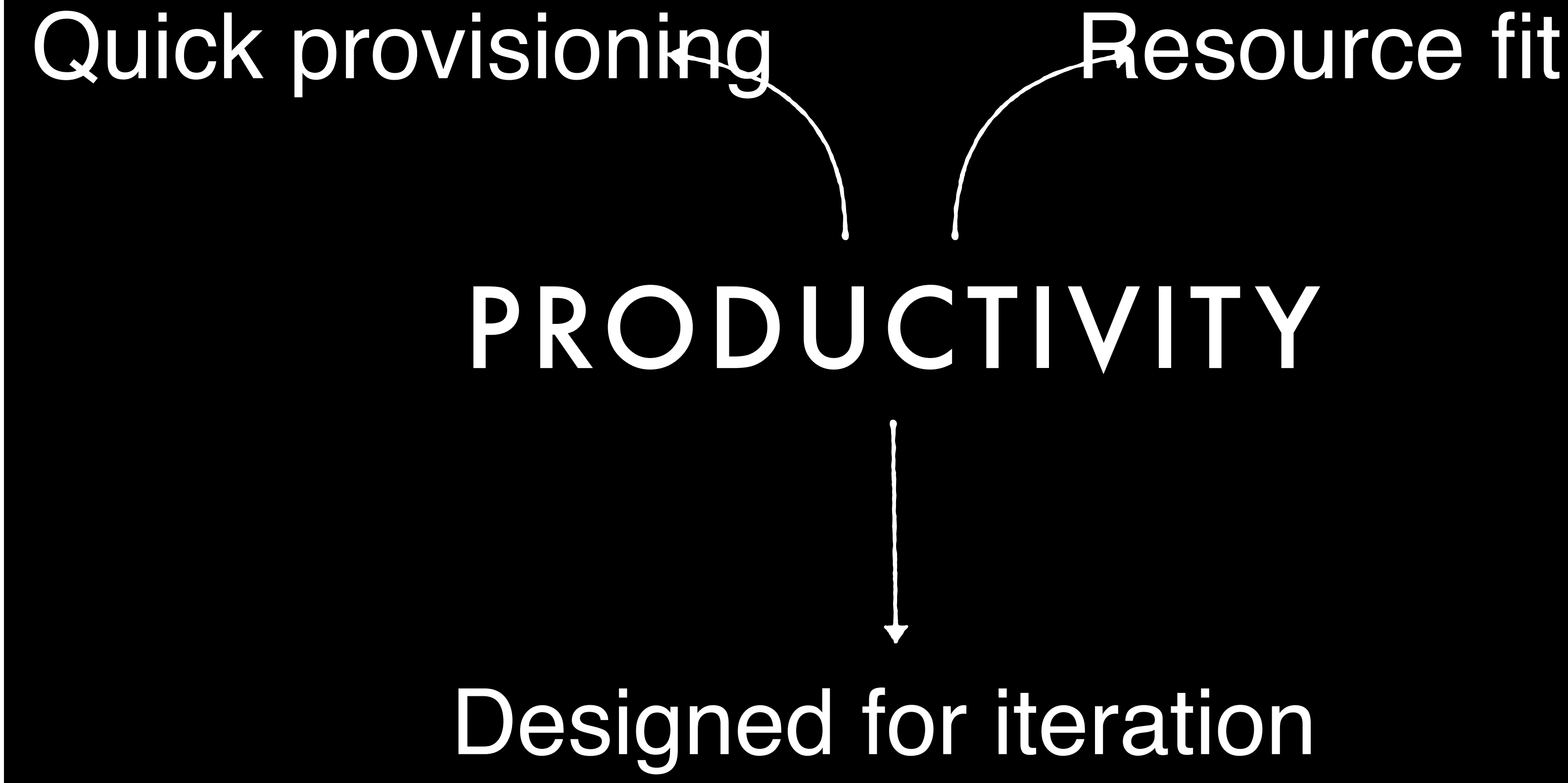
Interactive dashboard

Revenue forecast

Redshift

gumgum


Impression RDD
S3
Ad impressions & clicks
24/7 Spark cluster
Interactive dashboard
Revenue forecast
Click stream logs
Batch Spark clusters
Redshift
Data exploration and testing



Quick provisioning

Resource fit

PRODUCTIVITY

Designed for iteration

SPARK ON EMR

New

Provision and scale managed Spark clusters, as first class citizens

RAPID PROVISIONING OF ELASTIC CLUSTERS

Provision new clusters in minutes

High memory, high CPU, high IO instances

Add or remove capacity on running clusters

Access cluster instances directly

Clusters run within a VPC

DIRECT ACCESS TO DATA ON S3

Access objects directly on Amazon S3

Multiple clusters can access canonical data in S3

No need to copy or manage the data on the cluster

Server-side and client-side encryption with customer controlled keys

Mix S3 and HDFS on a cluster

INTEGRATION WITH THE SPOT MARKET

Bid on under utilized capacity on EC2

“Name your price” clusters

Very low cost at high scale

Lowest cost for time insensitive workloads

Also on-demand and reserve capacity pricing

SPARK ON EMR New

No additional cost. Available today.

aws.amazon.com/emr/spark

THANK YOU

MATTHEW@AMAZON.COM

@MZA