RTB and Big Data

Where Erlang and Hadoop Meet

Aol PLATFORMS.

Agenda

What is RTB in the context of Online Advertising?

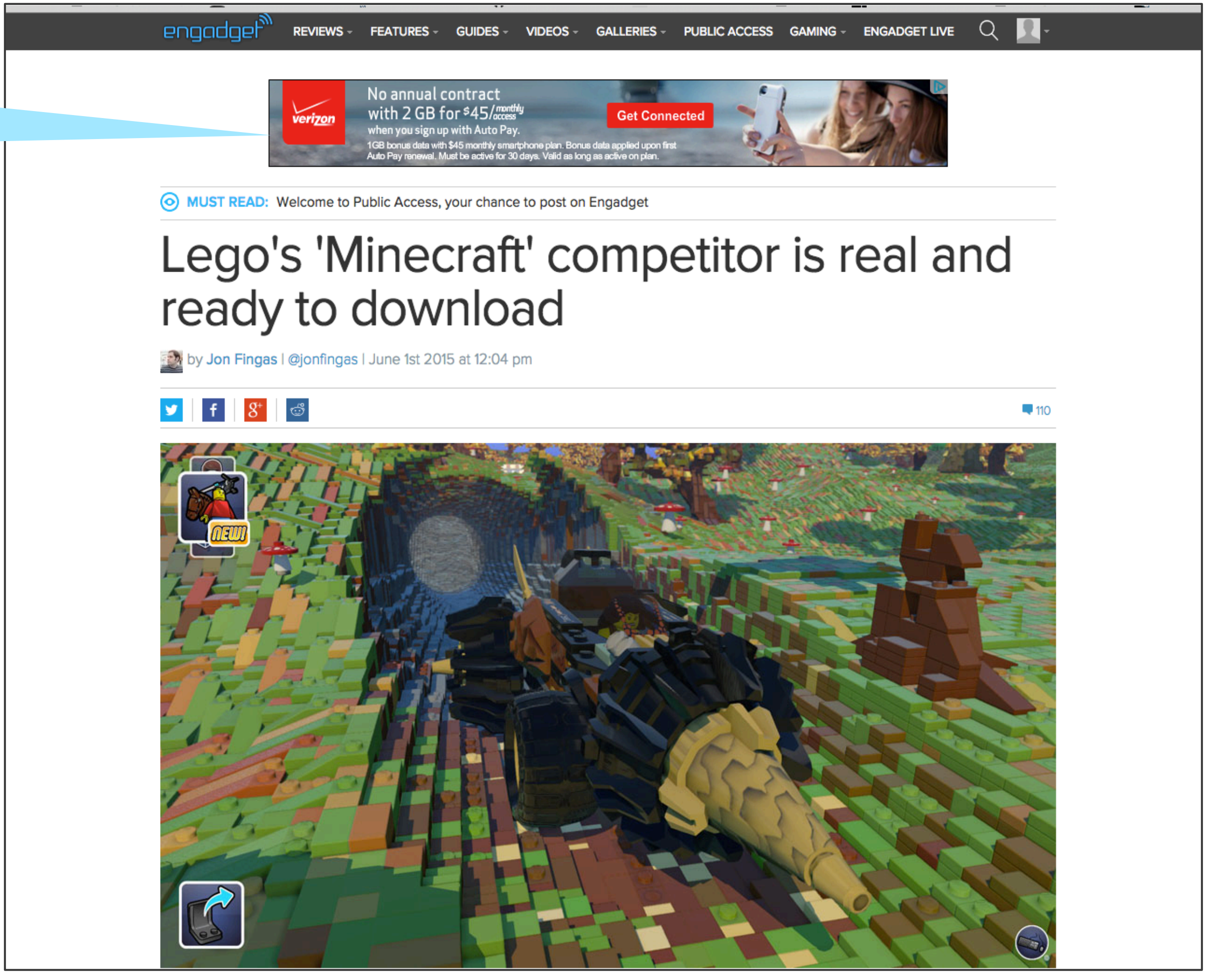
RTB Exchange Architecture

Data Handling with Hadoop

What is RTB ?

Online Advertising

Placement with Advertiser’s Banner



No annual contract
with 2 GB for $45/monthly access
when you sign up with Auto Pay.
1GB bonus data with $45 monthly smartphone plan. Bonus data applied upon first Auto Pay renewal. Must be active for 30 days. Valid as long as active on plan.

MUST READ: Welcome to Public Access, your chance to post on Engadget

Lego's 'Minecraft' competitor is real and ready to download

by Jon Fingas | @jonfingas | June 1st 2015 at 12:04 pm

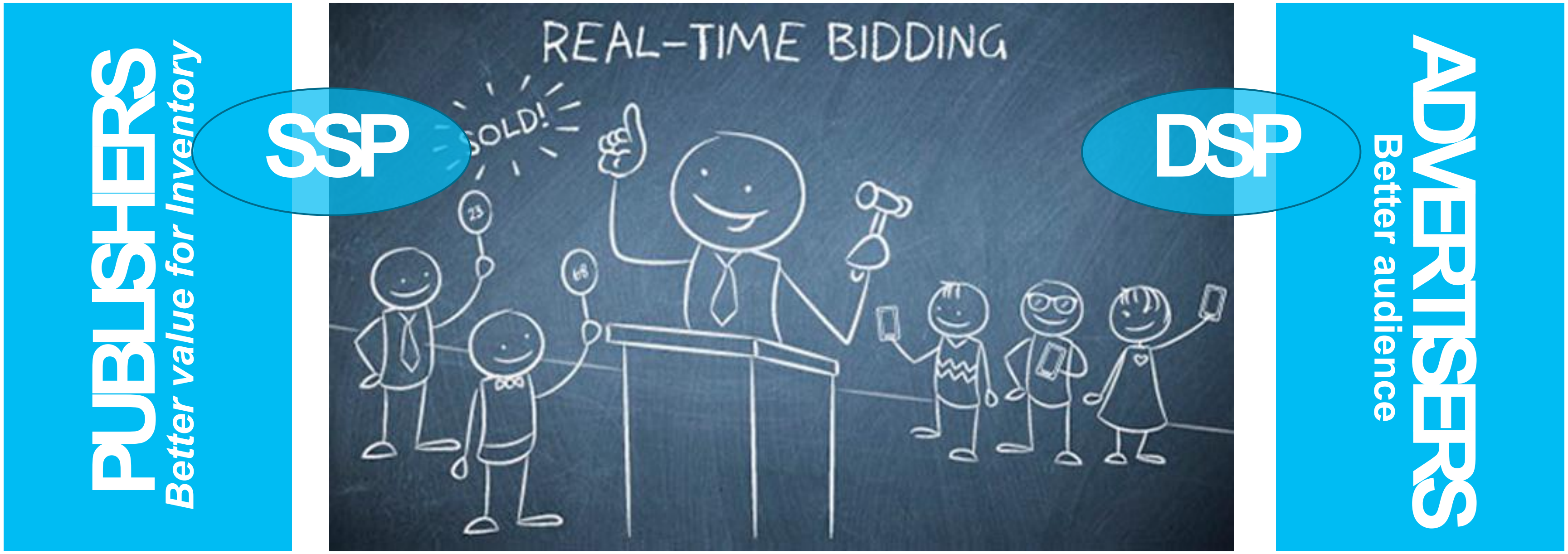
Free Content on WWW

Paid for by Advertising

Upfront agreements between Publishers and Advertisers

What is Real Time Bidding

- The buying a selling of impressions in real time while a page is loading



PUBLISHERS

Better value for Inventory

SSP

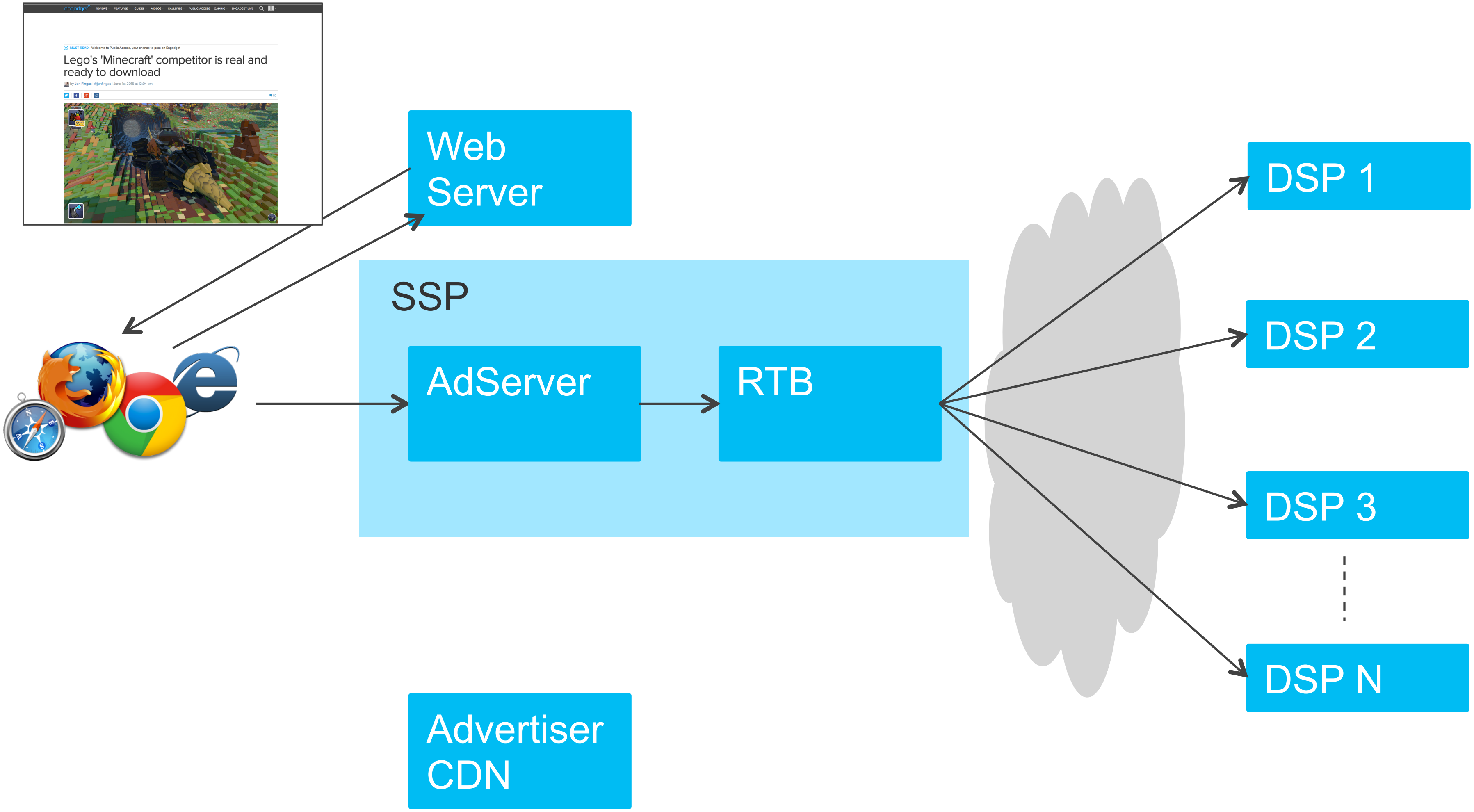
DSP

ADVERTISERS

Better audience

Exoclick, www.exoclick.com (2014)

Adserving Workflow



engadget REVIEWS FEATURES GUIDES VIDEOS GALLERIES PUBLIC ACCESS GAMING ENGADGET LIVE

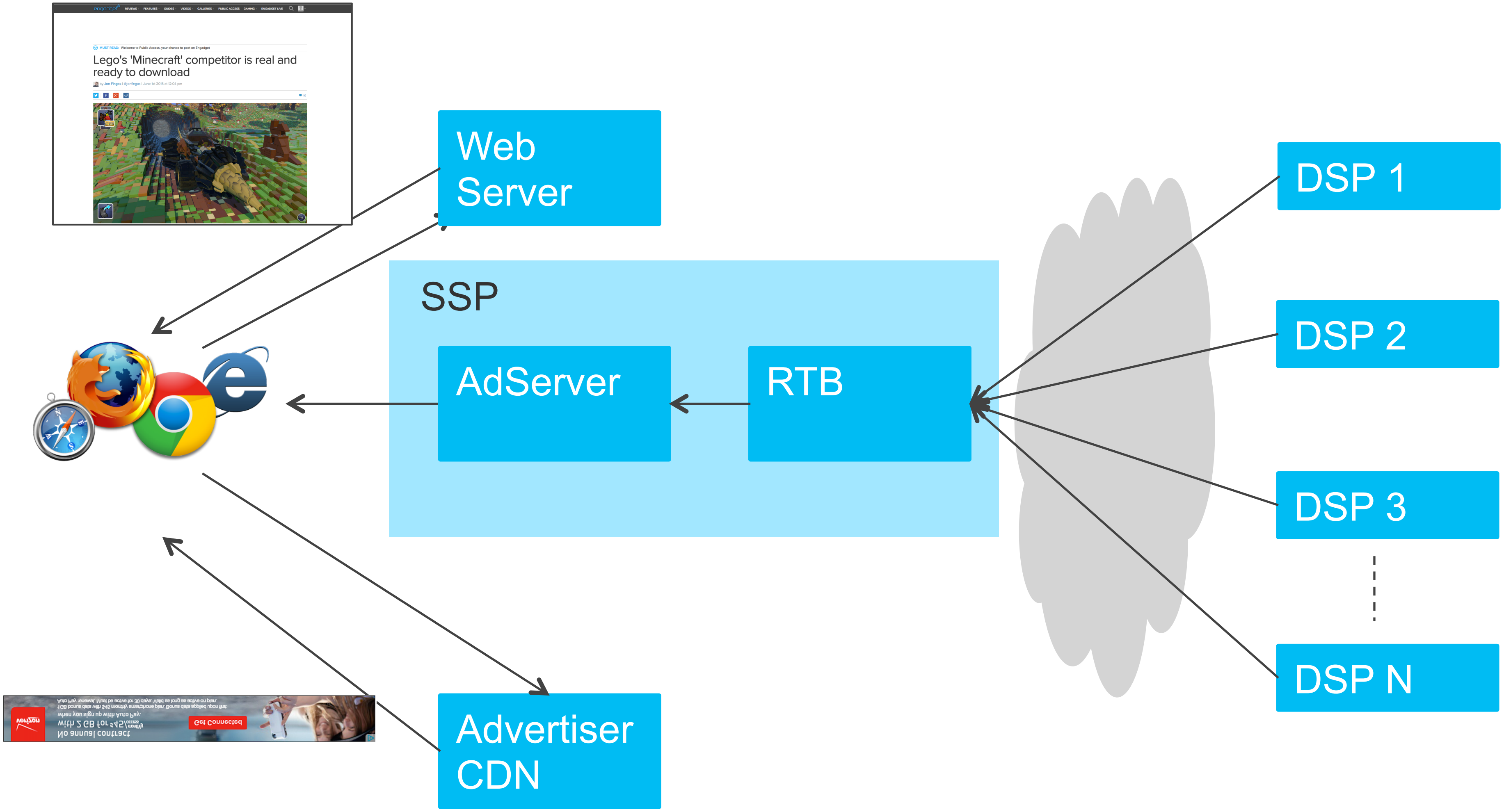
MUST READ: Welcome to Public Access, your chance to post on Engadget

Lego's 'Minecraft' competitor is real and ready to download

by Jon Fingas | @jonfingas | June 1st 2015 at 12:04 pm

110

Adserving Workflow



Web Server

SSP

AdServer

RTB

DSP 1

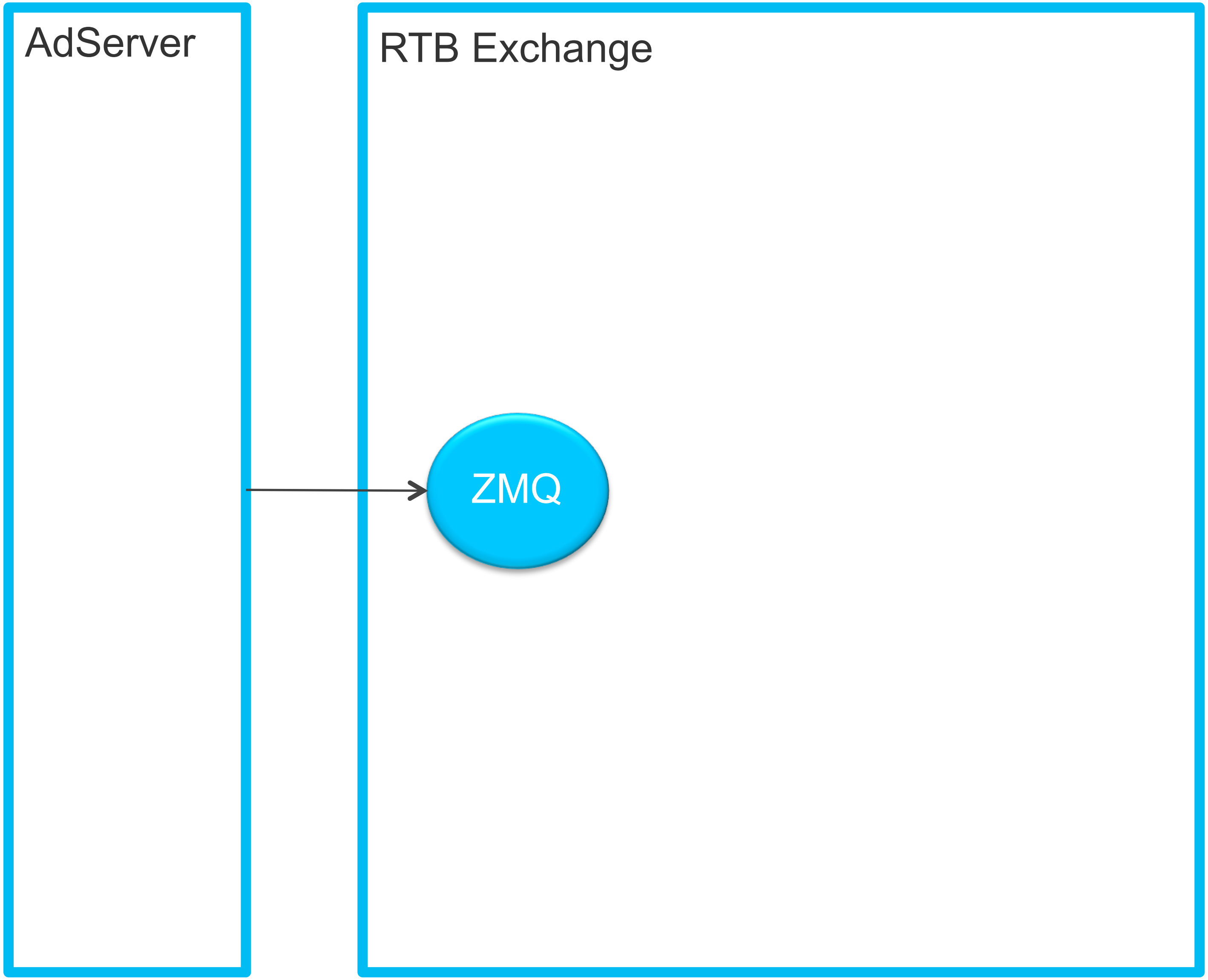
DSP 2

DSP 3

DSP N

Advertiser CDN

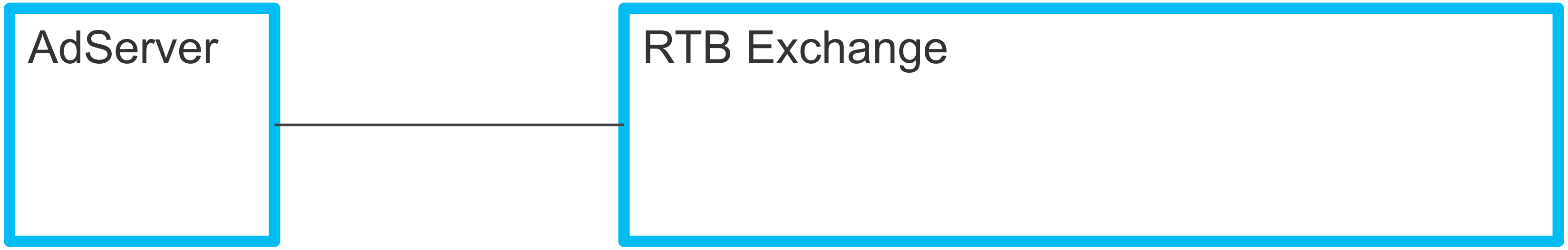
RTB Exchange Architecture



AdServer

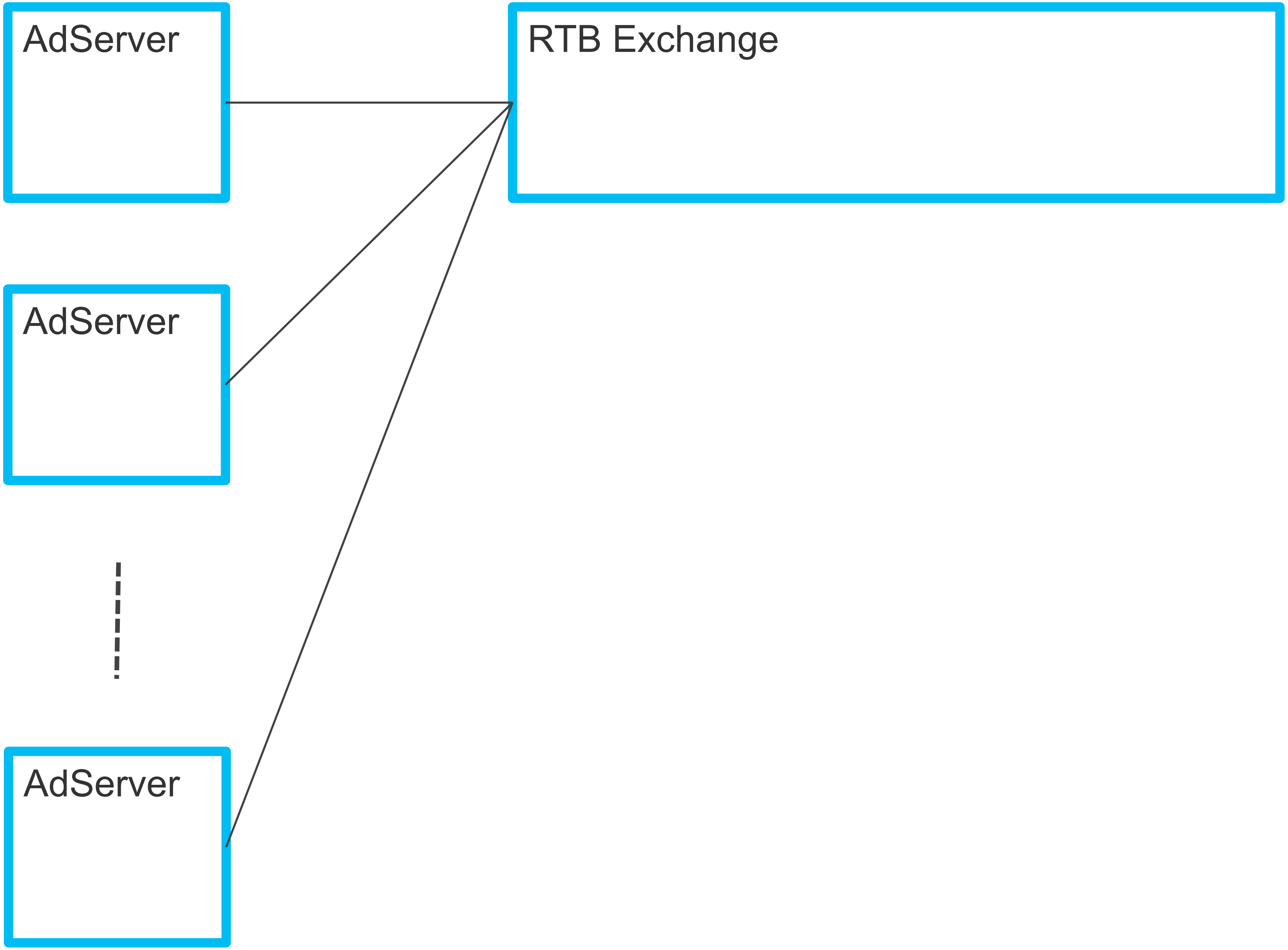
RTB Exchange

ZMQ



AdServer

RTB Exchange

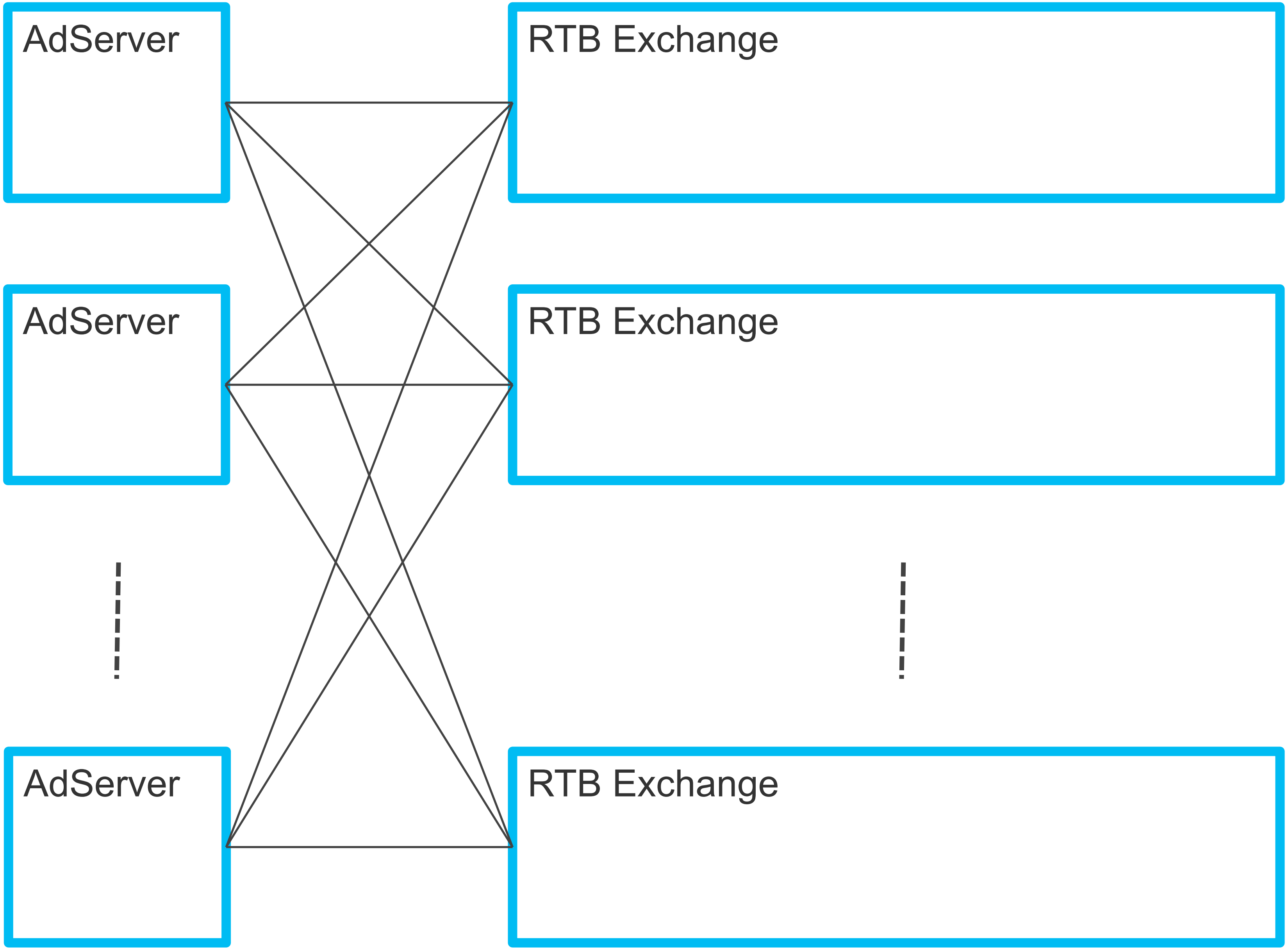


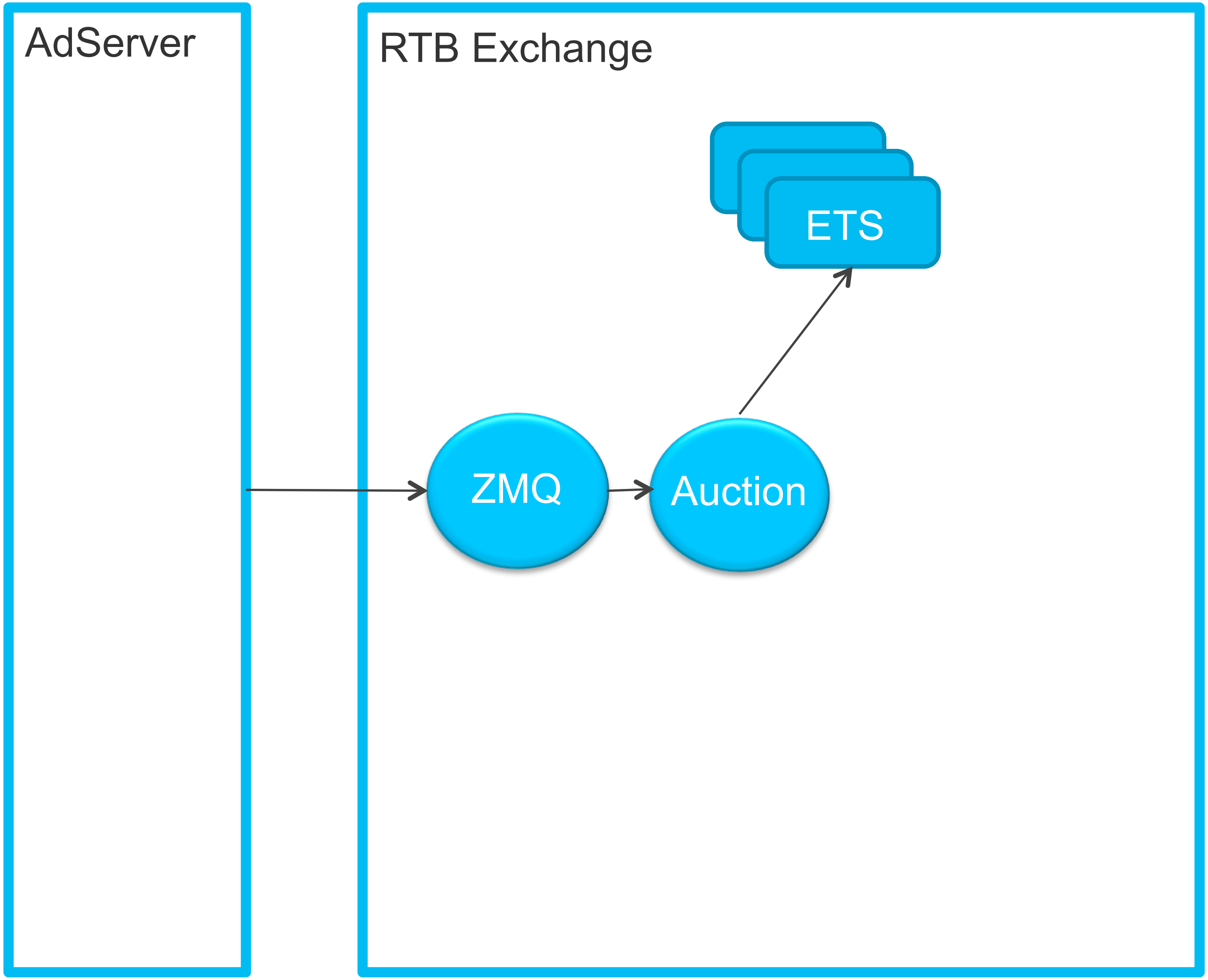
AdServer

AdServer

AdServer

RTB Exchange


AdServer
AdServer
AdServer
RTB Exchange
RTB Exchange
RTB Exchange



AdServer

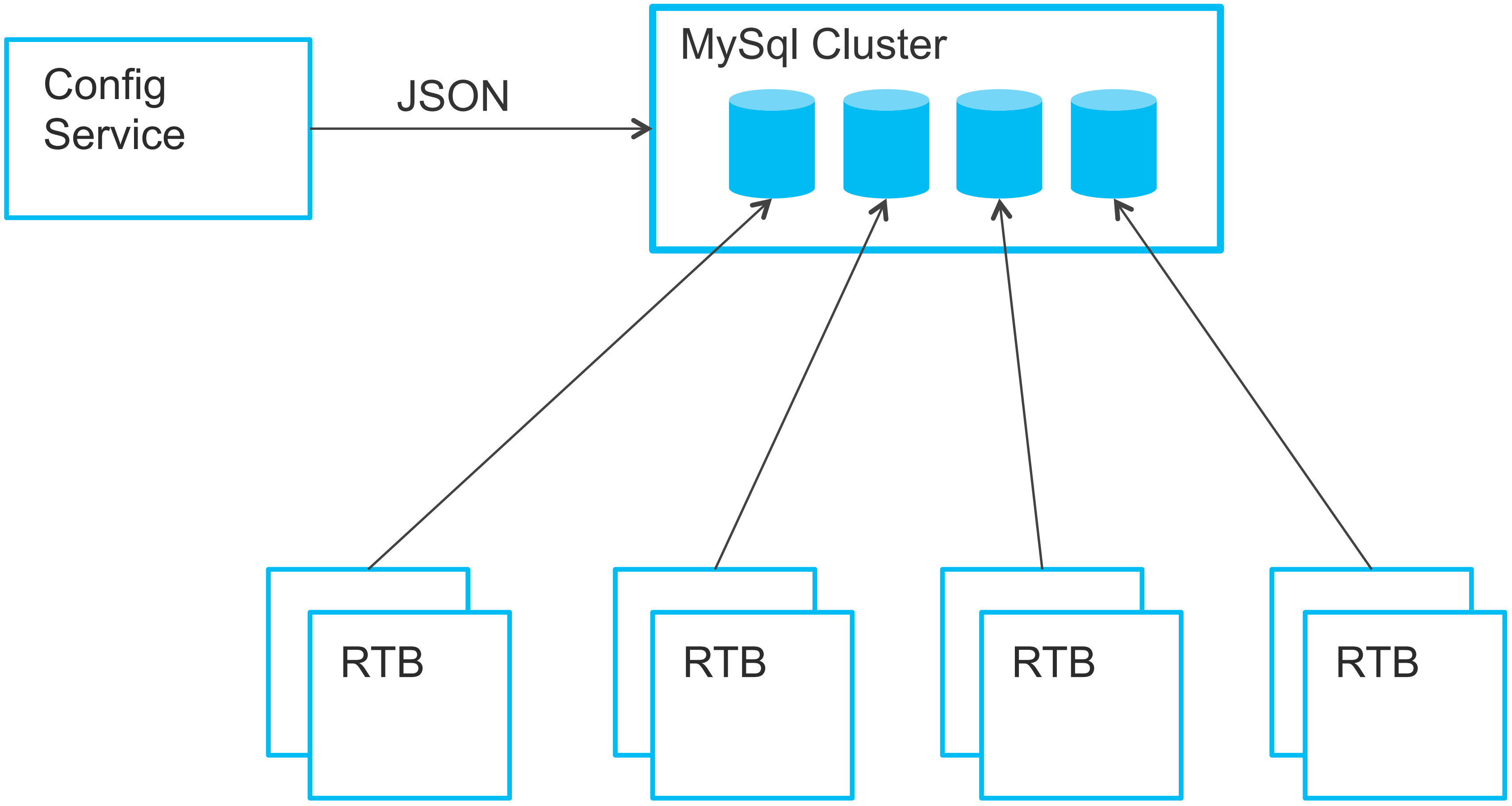
RTB Exchange

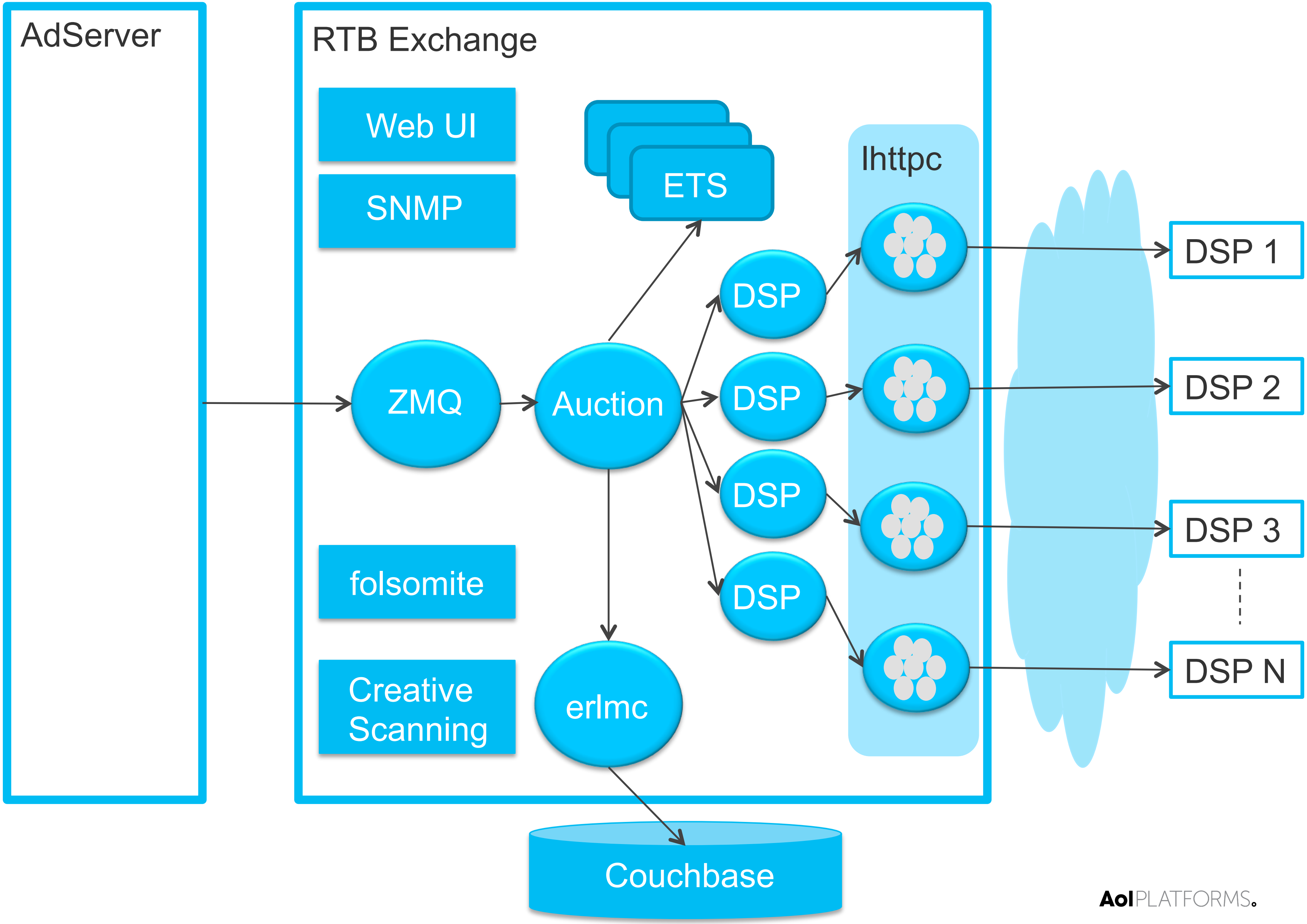
ETS

ZMQ

Auction

Retrieving Campaign Data





AolPLATFORMS.

How Much does it Scale ?

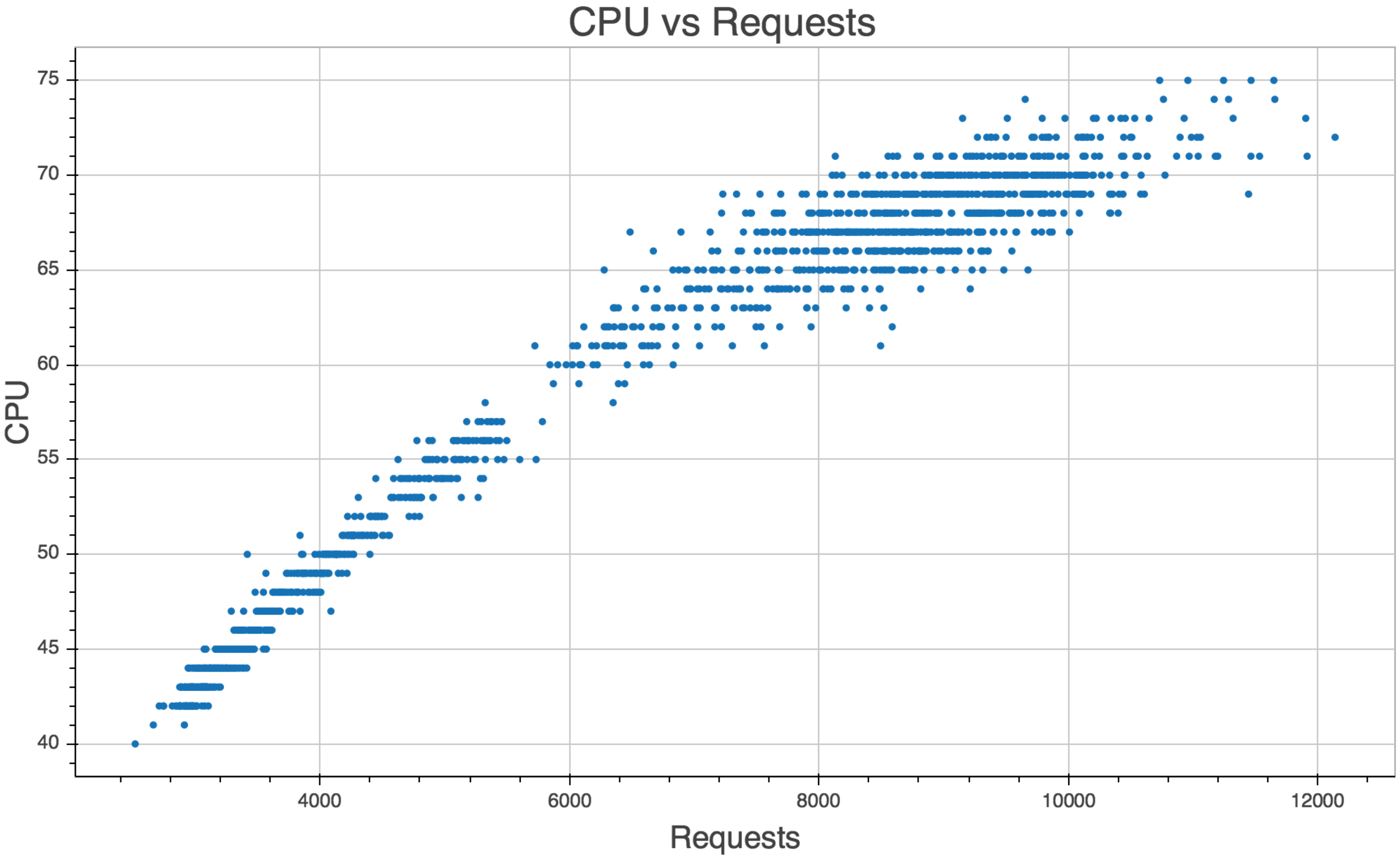
60 Billion

The number of bid requests build and sent to DSPs every day

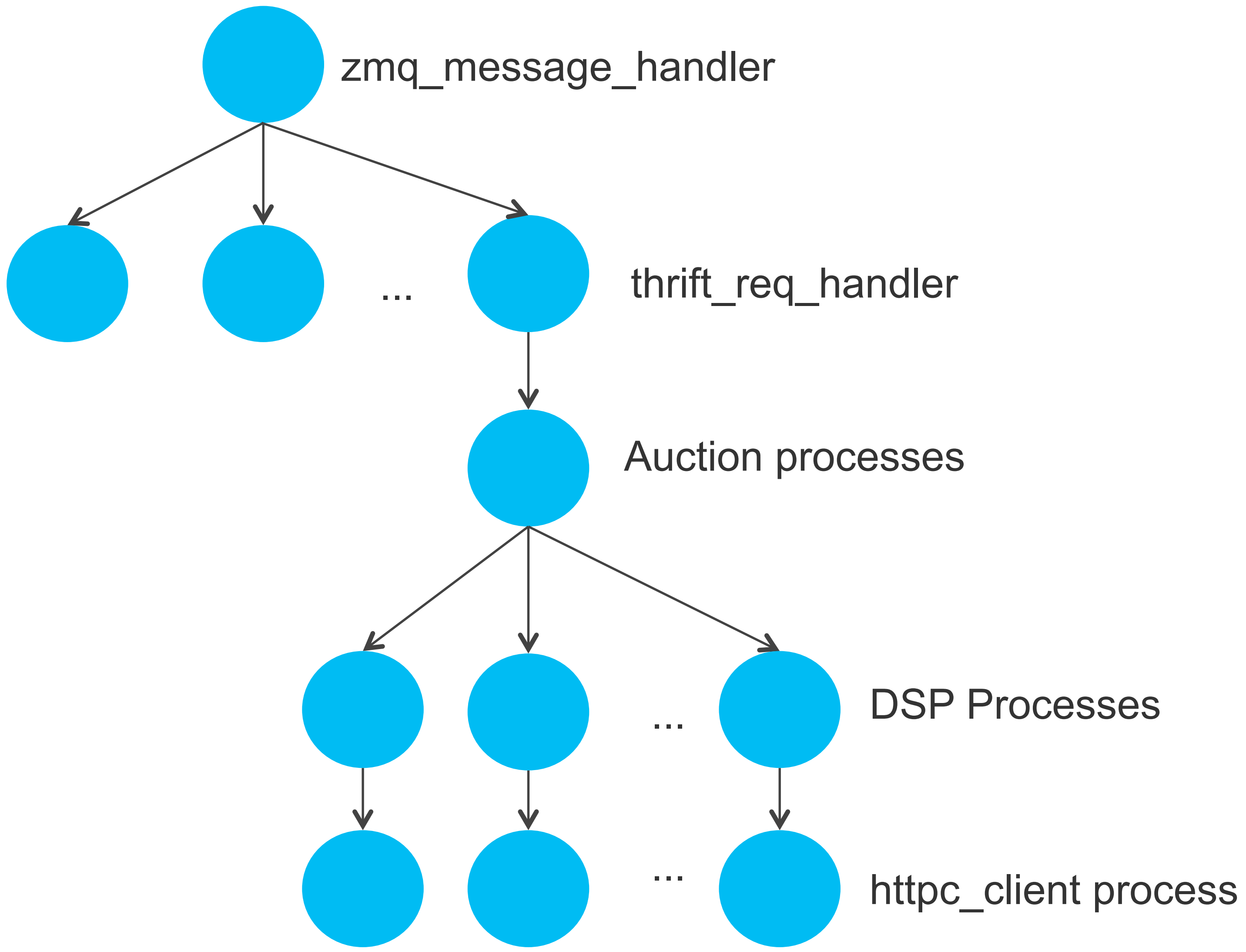
13,000

The number of bid requests build and sent to DSPs every second per host at peak

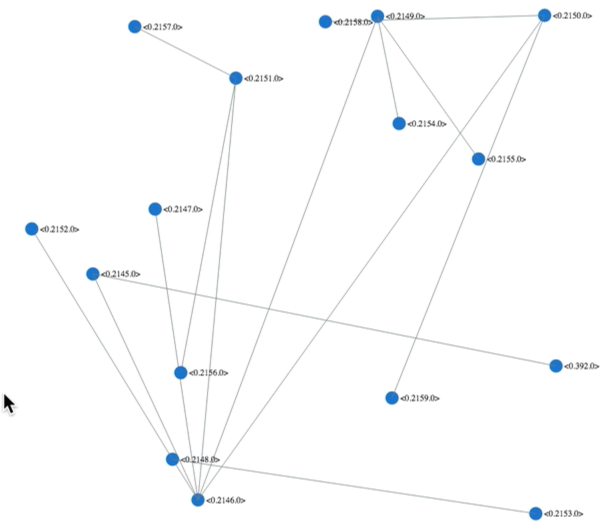
“Linear” Scaling



A Single Request



github.com/aol/erlgraph



Data Handling

Size matters!

What is Big Data?

Columnar Datastore

Linear Scaling

NewSQL

MapReduce

Parallelization

NoSQL

Cluster

Eventual Consitency

K Savety

Some Metrics?

2 x 100

Count of Hadoop nodes

2TB / h

Data Processed

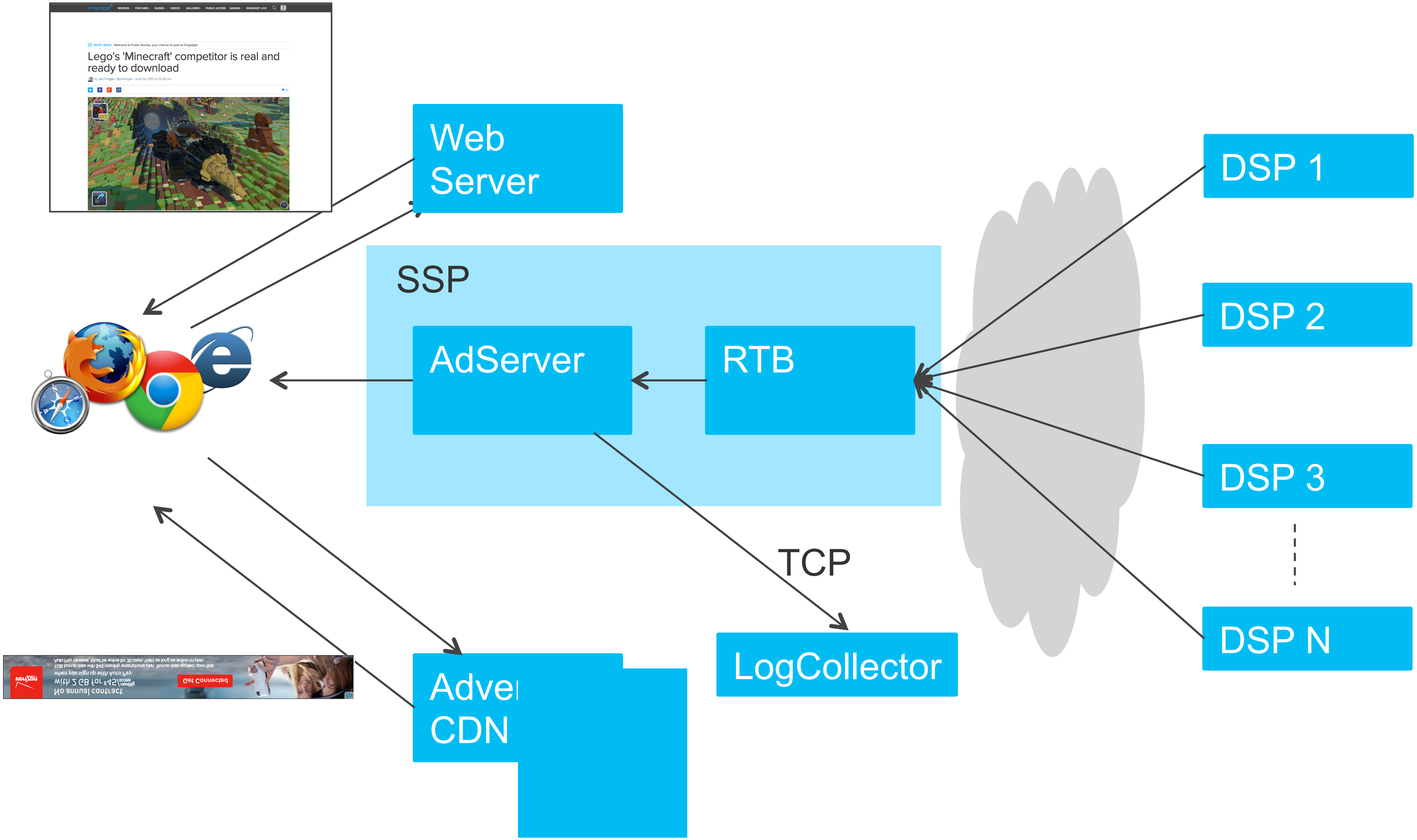
The size in printed paper would equal 500 million pages – a 50km tall pile

20TB

Data stored in Vertica Cluster

Paper pile would reach from Berlin to Stuttgart

Data Handling Logging



LogCollector

- Written in Java
- Based on Netty
- TCP input file output
- Optimized to maximize TCP throughput
  - Other solutions suffered under finer network control
- Circular ringbuffers
- Zero copy
- Gets binary payload + metadata for control flow

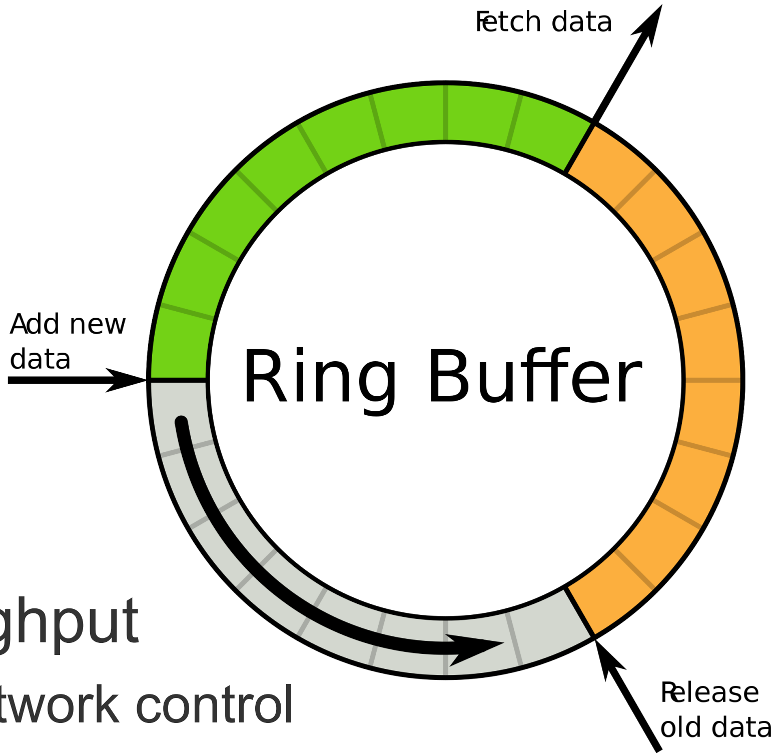


Image source: wikipedia

Avro Log Format

- Binary format
- Structural data support
  - Arrays, Trees etc.
- Compression
- Self descriptive
  - JSON schema header
- Well supported in Hadoop

Schema Example:

Header Part

```
{
  "name": "DataAvroPacket",
  "fields": [
    {
      "name": "SGSHeader",
      "type": {
        "name": "SGSHeader",
        "fields": [
          {
            "name": "VersionID",
            "type": "int"
          }
        ],
        "type": "record"
      }
    },...
```

Partners in Crime

Erlang and MapReduce

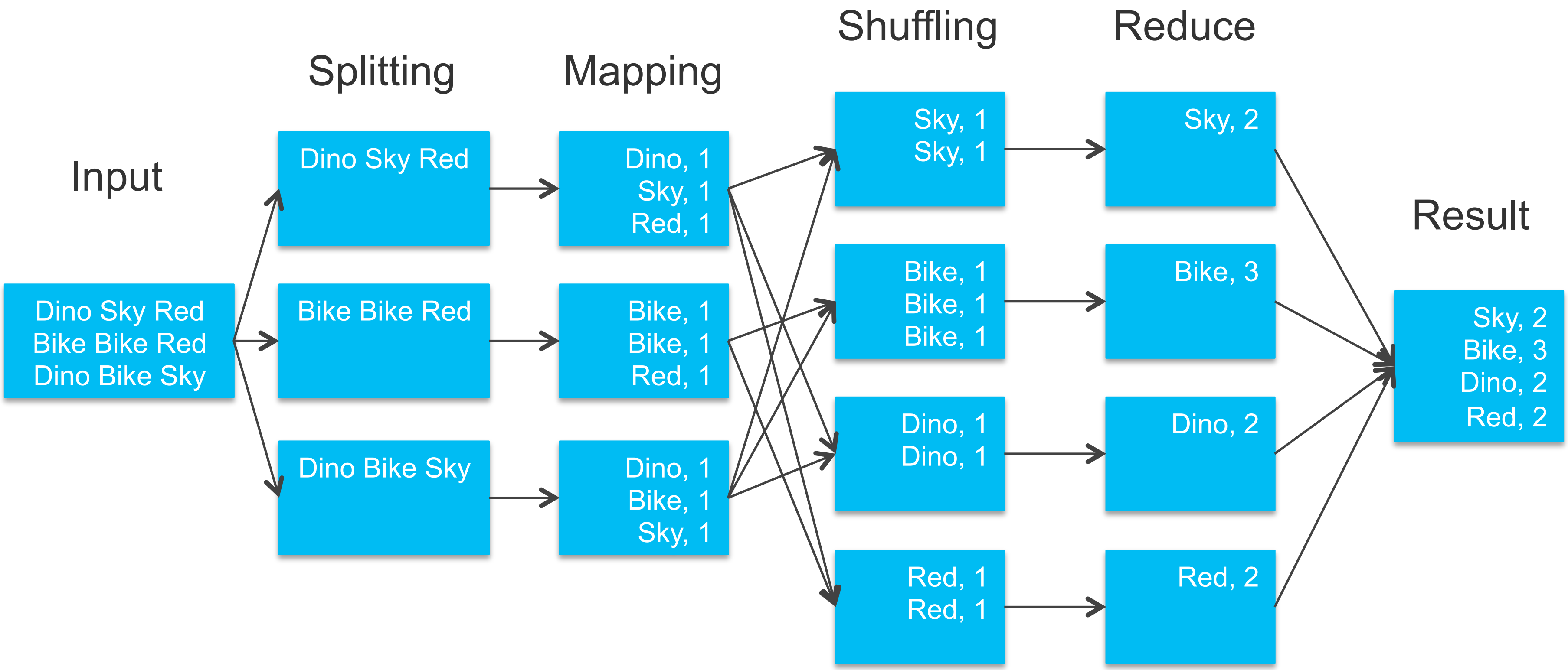
Aol PLATFORMS.

Map Reduce

„MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers (nodes), collectively referred to as a cluster“

(Wikipedia.com)

Map Reduce Example



Input

Dino Sky Red
Bike Bike Red
Dino Bike Sky

Splitting

Dino Sky Red

Bike Bike Red

Dino Bike Sky

Mapping

Dino, 1
Sky, 1
Red, 1

Bike, 1
Bike, 1
Red, 1

Dino, 1
Bike, 1
Sky, 1

Shuffling

Sky, 1
Sky, 1

Bike, 1
Bike, 1
Bike, 1

Dino, 1
Dino, 1

Red, 1
Red, 1

Reduce

Sky, 2

Bike, 3

Dino, 2

Red, 2

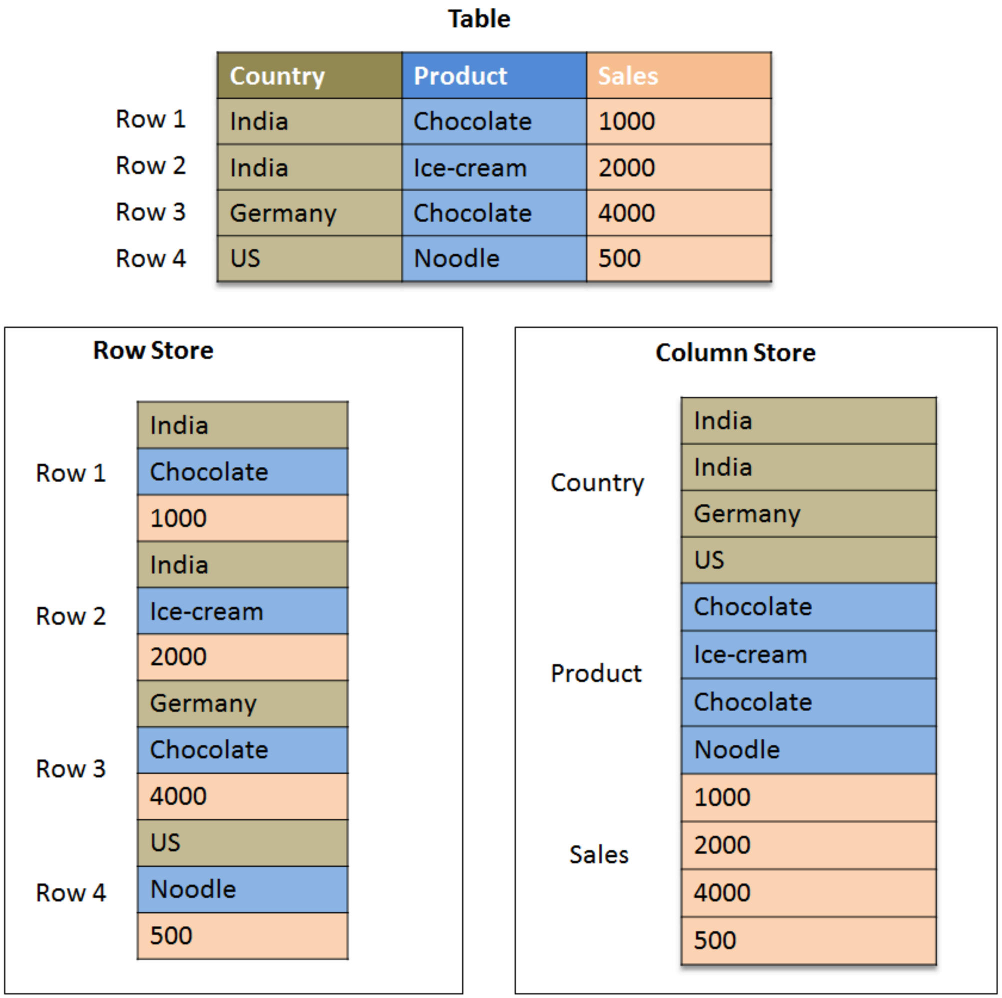
Result

Sky, 2
Bike, 3
Dino, 2
Red, 2

Aol PLATFORMS.

ColumnarDatastore

- We use Vertica
- Significant read performance gain compared to traditional RDBMS
- Each Column end up in own file
- Trick is stream compression combined with smart search (like binary)



Table

	Country	Product	Sales
Row 1	India	Chocolate	1000
Row 2	India	Ice-cream	2000
Row 3	Germany	Chocolate	4000
Row 4	US	Noodle	500

Row Store

Row 1	India
	Chocolate
	1000
Row 2	India
	Ice-cream
	2000
Row 3	Germany
	Chocolate
	4000
Row 4	US
	Noodle
	500

Column Store

Country	India
	India
	Germany
	US
Product	Chocolate
	Ice-cream
	Chocolate
	Noodle
Sales	1000
	2000
	4000
	500

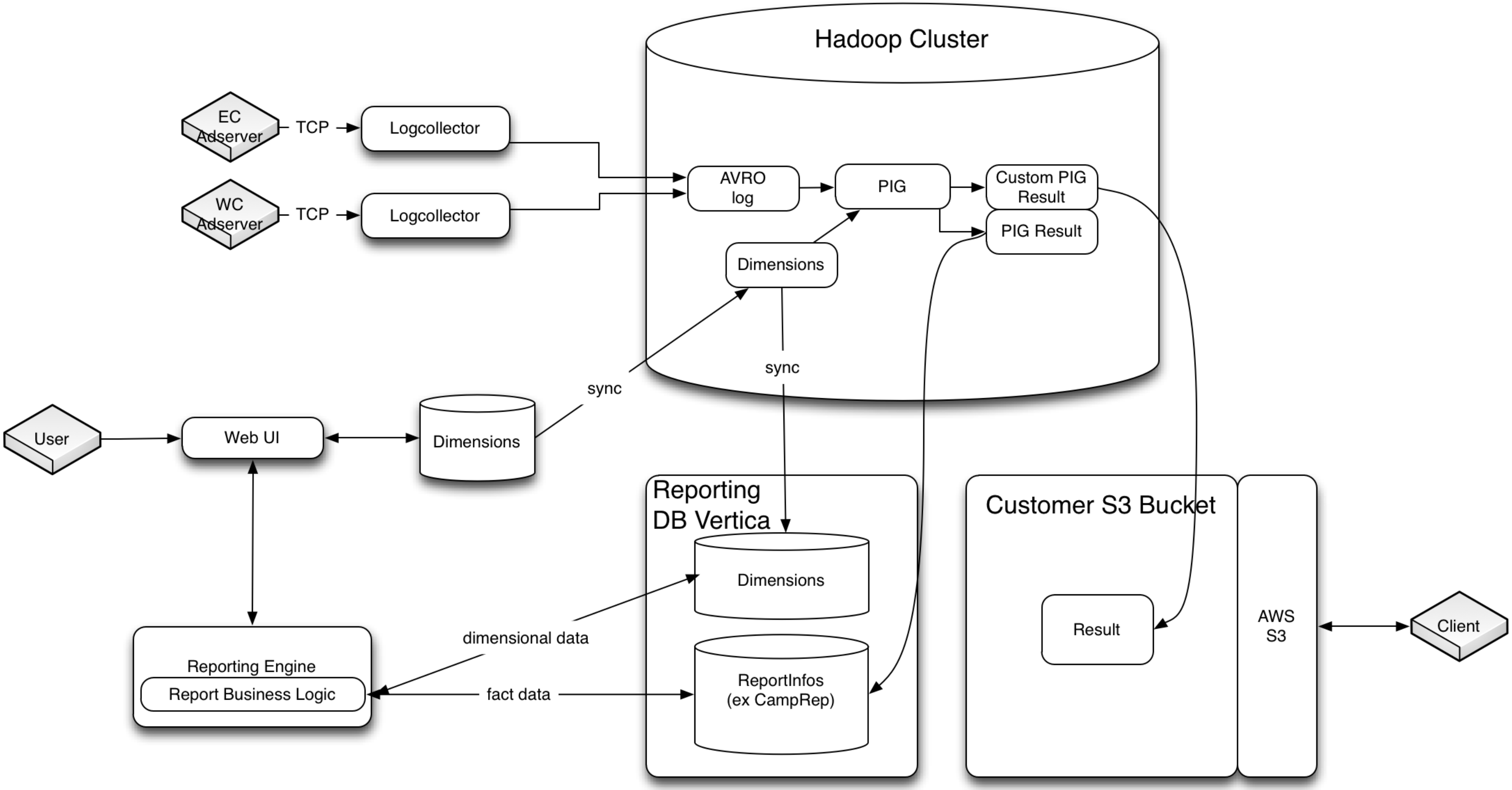
PIG

- Script language
- Creates Map Reduce
- Overcomes the need to write native jobs
- Dataflow oriented

Script Example:

```
REGISTER /usr/lib/pig/contrib/piggybank/java/lib/avro-1.5.4.jar
%default INFILE '/var/tmp/example1.avro’
...
rec1 = LOAD '$INFILE’
USING org.apache.pig.piggybank.storage.avro.AvroStorage ('{}');
rec1Data = FOREACH rec1 GENERATE
SGSMainPacket.PlacementId,SGSMainPacket.CampaignId,
   SGSMainPacket.BannerNumber, $REP_DATE AS DATE, $REP_HOUR AS HOUR;
recGroup = GROUP rec1Data BY ( PlacementId,CampaignId,BannerNumber,DATE,HOUR);
fullCount = FOREACH recGroup GENERATE
   1,    -- VERSION COUNTER
   group.PlacementId,group.CampaignId,group.BannerNumber,group.DATE,group.HOUR,
   COUNT(rec1Data) AS TOTAL;
STORE fullCount INTO '$OUTFILE’
USING org.apache.pig.piggybank.storage.avro.AvroStorage (‘
{
   "schema":
 { "name" : "SummaryHourly”,
  "type" : "record”,
  "fields": [
       { "name": "Version", "type": "int" },
       { "name": "PlacementId", "type": "int" },
       { "name": "CampaignId", "type": "int" },
       { "name": "BannerNumber", "type": "int" },
       { "name": "DateEntered", "type": "int" },
       { "name": "Hour", "type": "int" },
       { "name": "COUNT", "type": "long" }
     ]
}
}');
```

Reporting Architecture



What’s Next?

- Moving into AWS
  - Easier scaling
  - Easy test cluster ramp ups
  - Easy to get additional ressources in error cases to catch up
- Spark
  - Optimized Dataflow
  - Streaming, less intermediate files
  - More functionality
  - Written in Scala

Interested? We’re hiring!