MEMSQL

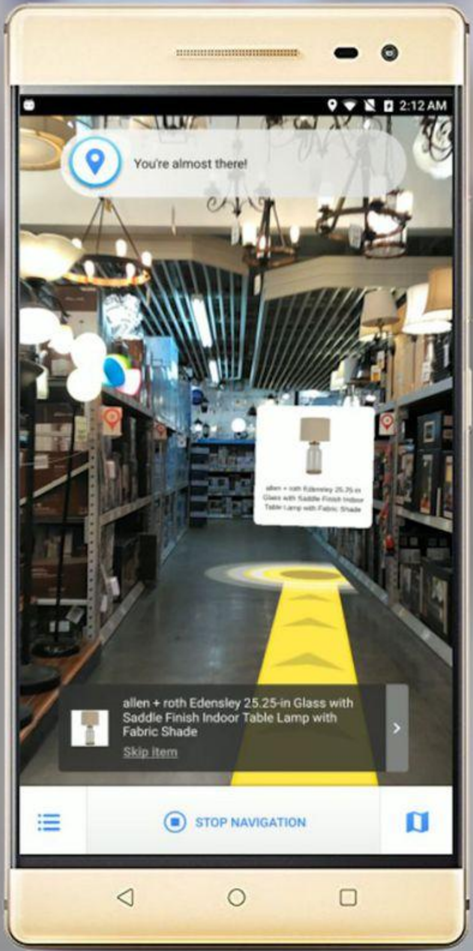
Real-Time Image Recognition with Apache Spark

Nikita Shamgunov, CTO, MemSQL

Spark Summit West 2017

The future of computing is visual

and also
numerical :)

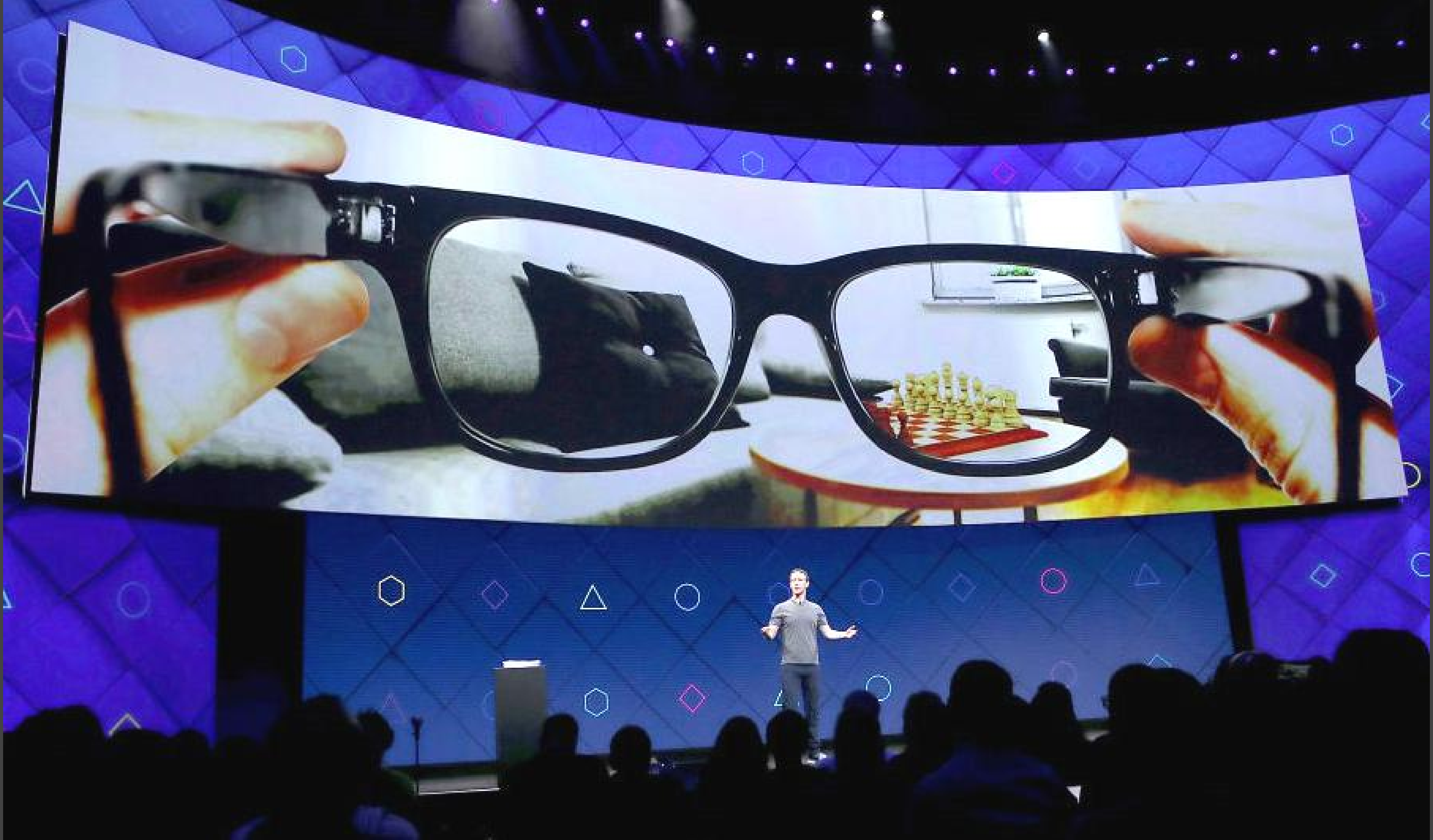


You're almost there!

allen + roth Edensley 25.25-in Glass with Saddle Finish Indoor Table Lamp with Fabric Shade

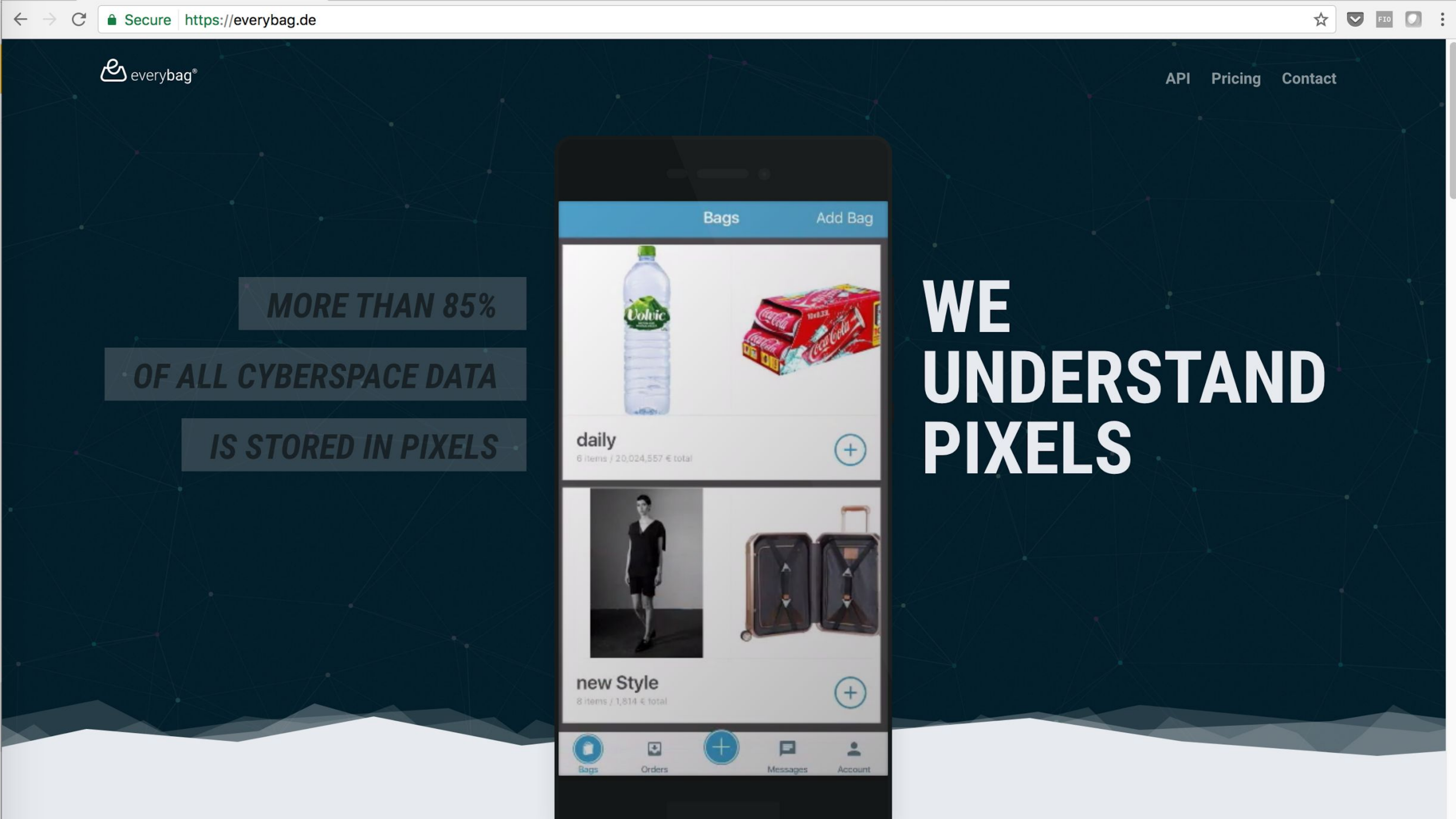
allen + roth Edensley 25.25-in Glass with Saddle Finish Indoor Table Lamp with Fabric Shade

Skip item



add apple image recognition slide

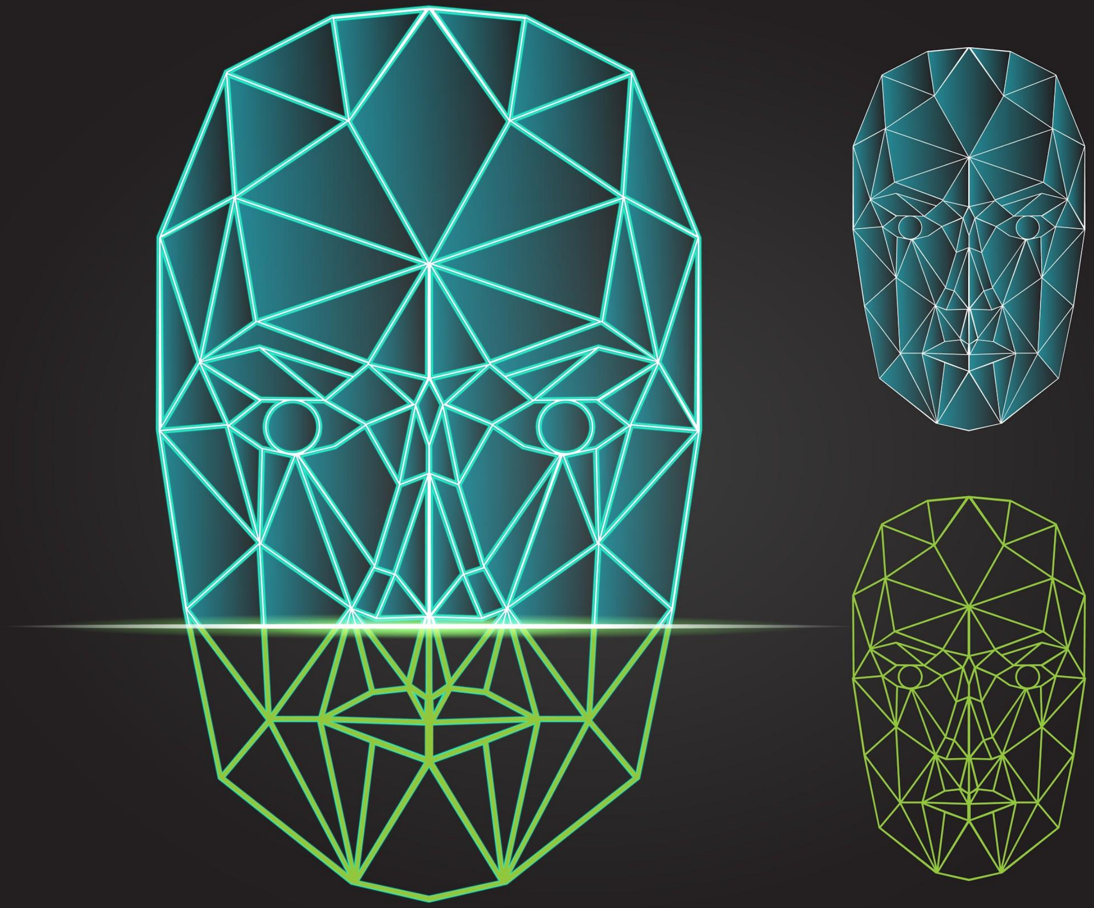
Putting image recognition to work today



MORE THAN 85%
OF ALL CYBERSPACE DATA
IS STORED IN PIXELS

WE UNDERSTAND PIXELS

Video



How It Works

Real-Time Image Recognition Workflow

- Train the model with Spark and TensorFlow
- Use the Model to extract feature vectors from images
  - Model + Image => FV
- You can store every feature vector in a MemSQL table

```
CREATE TABLE features (
  id bigint(11) NOT NULL AUTO_INCREMENT,
  image binary(4096) DEFAULT NULL,
  KEY id (id)USING CLUSTERED COLUMNSTORE
)
```

Working with feature vectors

For every image we store an `ID` and a normalized feature vector in a MemSQL table called `features`.

```
ID  |   Feature Vector
x   |   4KB
```

To find similar images we use this SQL query

```
SELECT
    id
FROM
    features
WHERE
    DOT_PRODUCT(feature * <input>) > 0.9
```

Understanding Dot Product

- Dot Product is an algebraic operation
  - SUM(Xi*Yi) TODO: Put a formula
- With the specific model and normalized feature vectors DOT PRODUCT results in a similarity score.
  - The closer the score is to 1 the more similar are the images

Performance Enhancing Techniques

Achieving best-in-class dot product implementation

- SIMD-powered
- Data compression
- Query parallelism
- Scale out

- Result: Processing at **Memory Bandwidth Speed**

Performance numbers

- Memory Speed: 40GB/sec
- Each vector 4K
- **12.5 Million Images a second** per node
- or
- **1 Billion images a second** on 100 node cluster

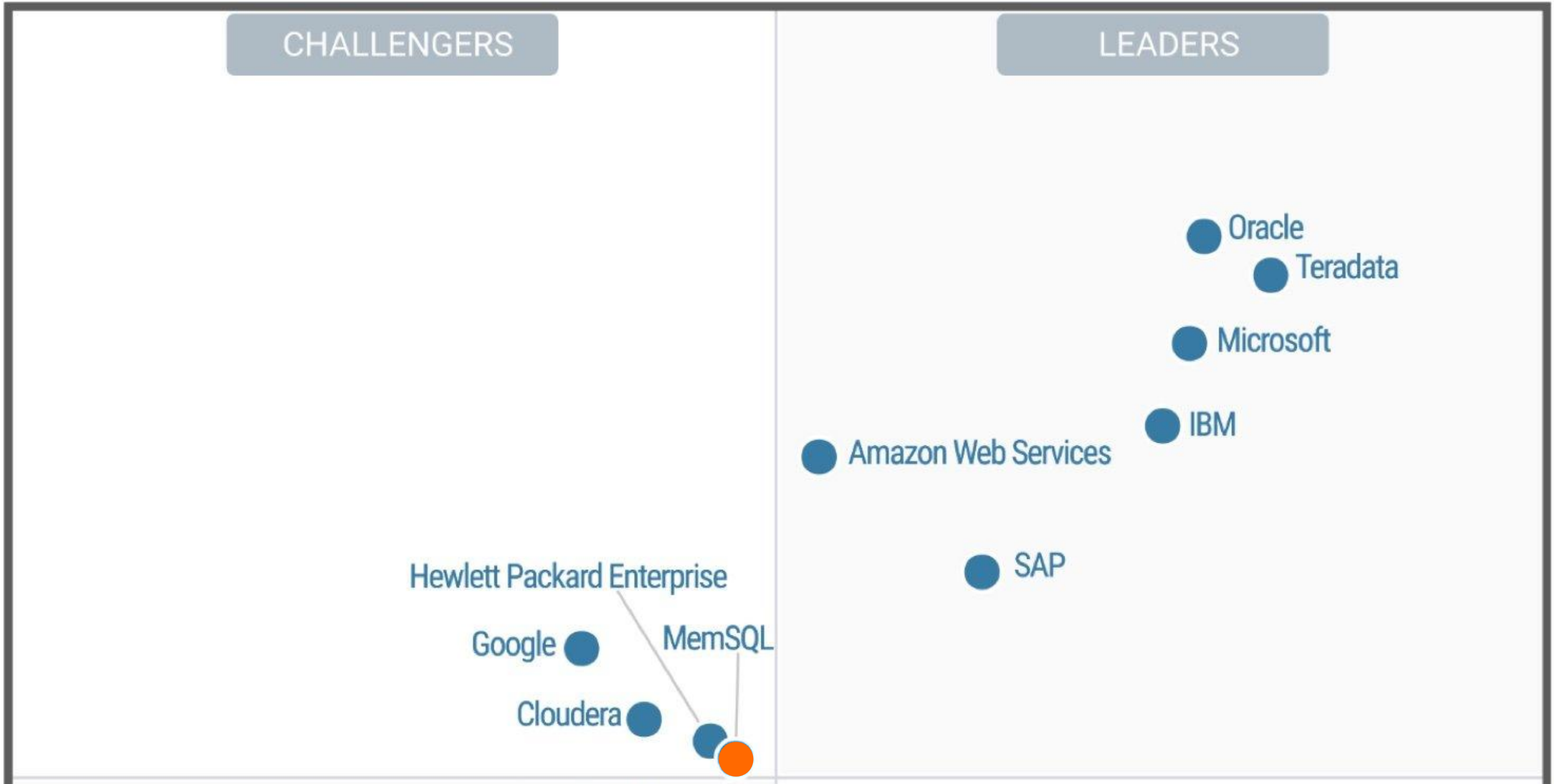
About MemSQL

MemSQL: The Real-Time Data Warehouse

- Scalable
  - Petabyte scale
  - High Concurrency
  - System of record
- Real-time
  - Operational
- Compatible
  - ETL
  - Business Intelligence
  - Kafka
  - Spark

- Deployment
  - MemSQL Cloud Service
  - Any public cloud IaaS
  - On-premises
- Community Edition
  - Unlimited scale
  - Limited high availability and security features

2017 Magic Quadrant for Data Management Solutions for Analytics



About Spark

Apache Spark™ is a fast and general engine for large-scale data processing.

Source: spark.apache.org June 2017

APACHE
Spark™

MEMSQL

Understanding Spark and MemSQL

Spark

Fast, large scale

General processing engine

Great for computation

MemSQL

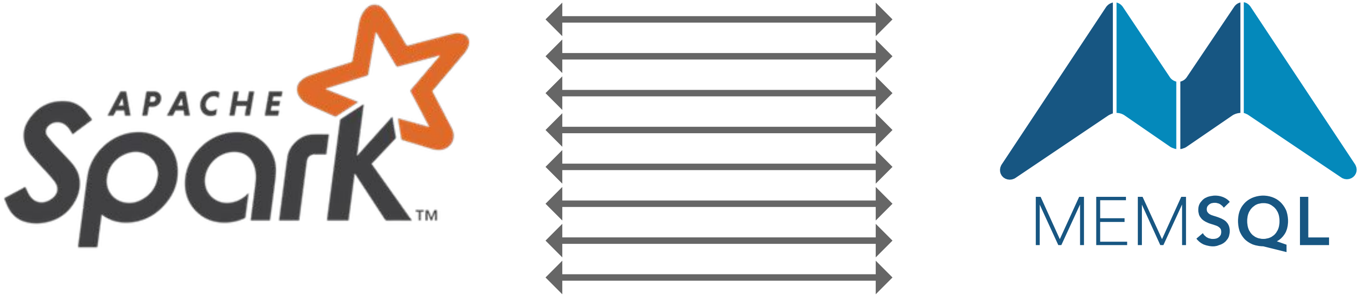
Fast, large scale

Real-time data warehouse

Great for SQL computation, persistence, transactions, application analytics

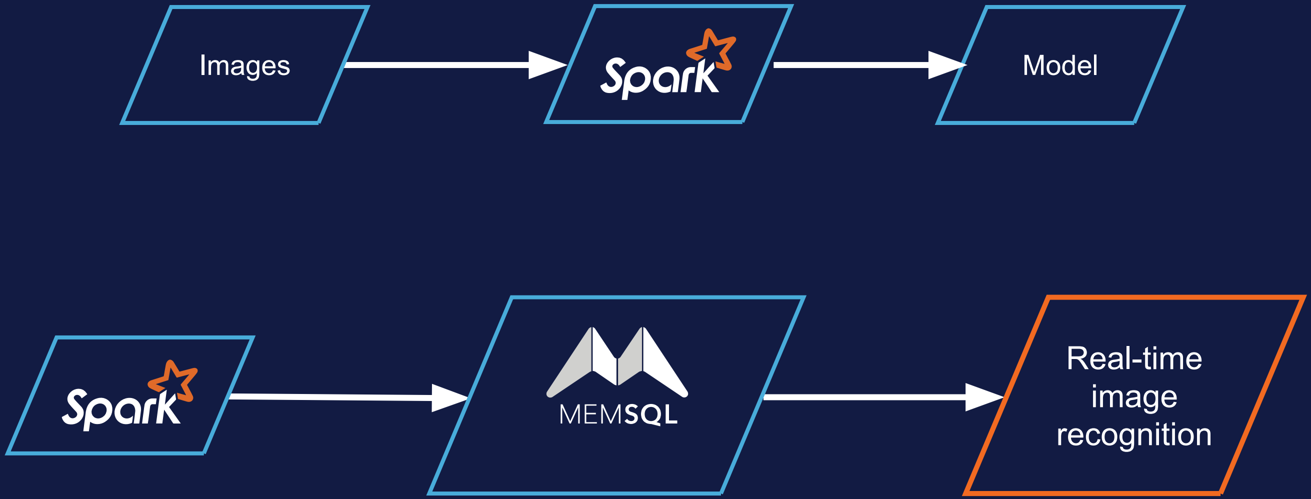
MemSQL Spark Connector 2

Highly parallel, high throughput, bi-directional



Demo

Demo Architecture



Persistent, Queryable Format

```
SELECT
    id
FROM
    features
WHERE
    DOT_PRODUCT(image, 0xa334efa…)
```

MEMSQL

Thank you!

@NikitaShamgunov

www.memsql.com