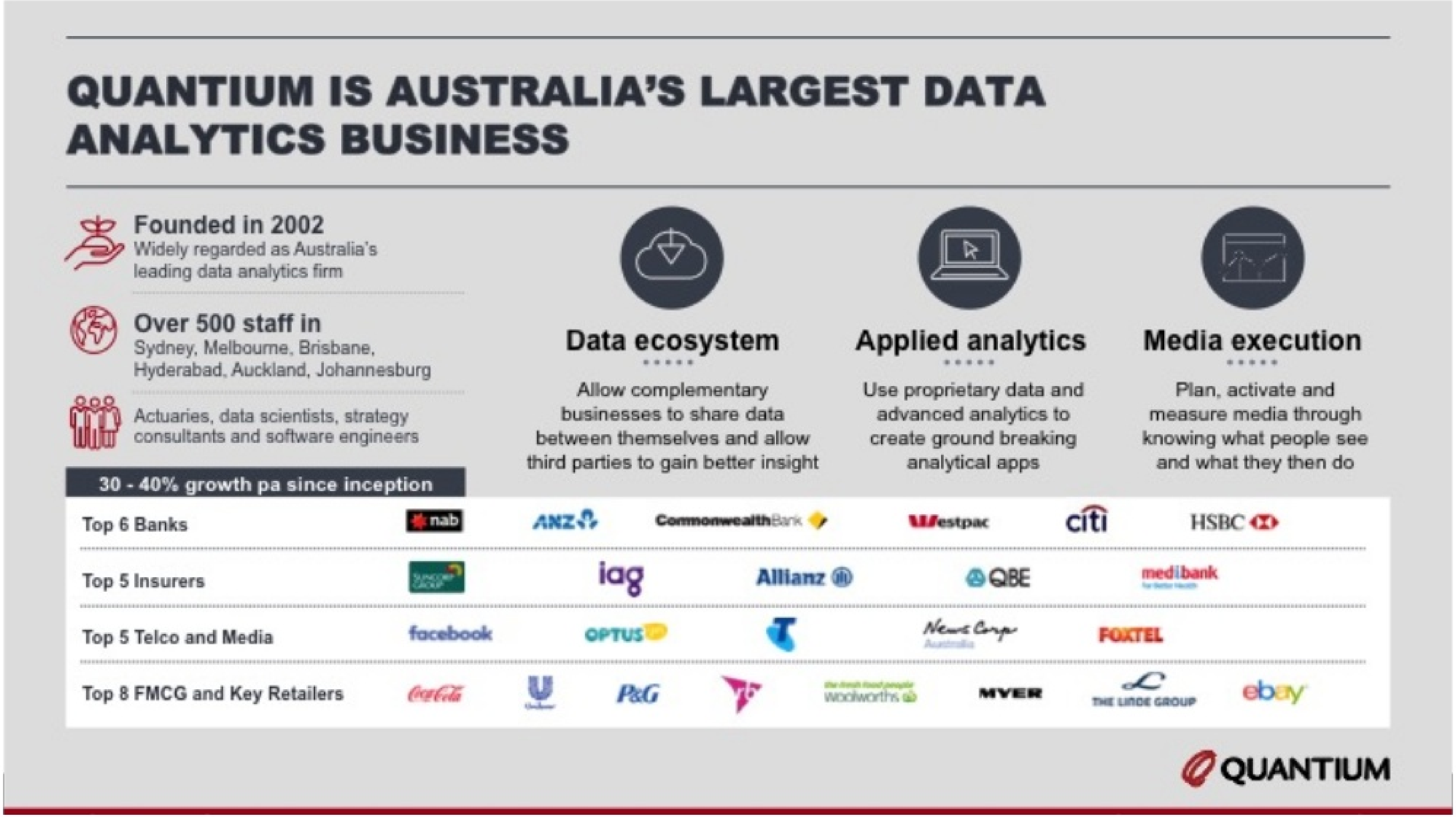
SPARK SUMMIT 2017

Operationalizing Machine Learning at Scale

Sameer Nori, MapR



QUANTIUM IS AUSTRALIA'S LARGEST DATA ANALYTICS BUSINESS

Founded in 2002
Widely regarded as Australia's leading data analytics firm

Over 500 staff in
Sydney, Melbourne, Brisbane, Hyderabad, Auckland, Johannesburg

Actuaries, data scientists, strategy consultants and software engineers

30 - 40% growth pa since inception

Data ecosystem

Allow complementary businesses to share data between themselves and allow third parties to gain better insight

Applied analytics

Use proprietary data and advanced analytics to create ground breaking analytical apps

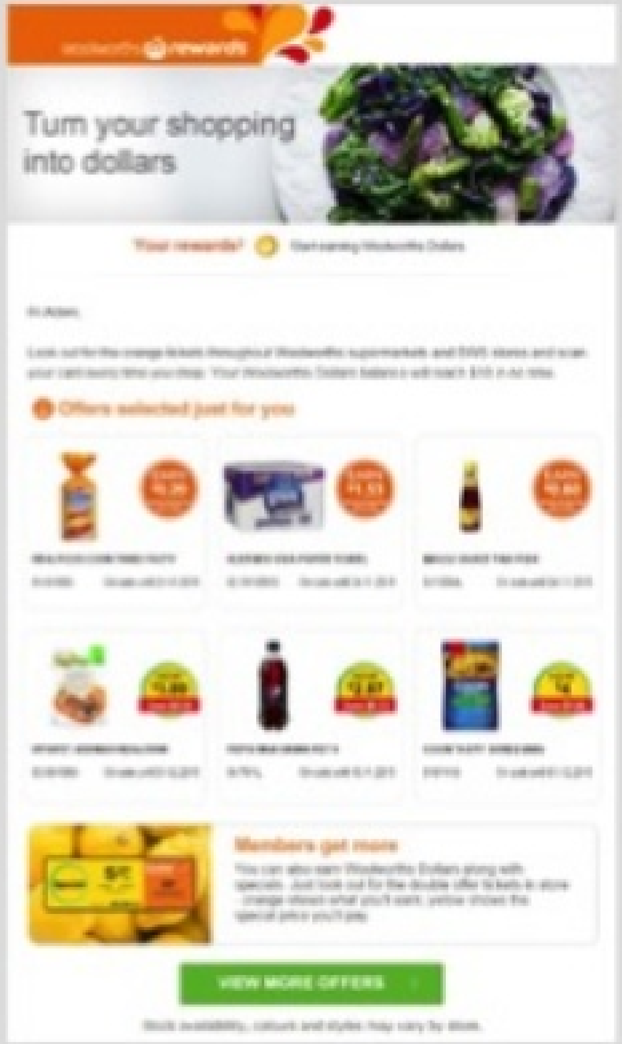
Media execution

Plan, activate and measure media through knowing what people see and what they then do

Top 6 Banks	nab	ANZ	CommonwealthBank	Westpac	citi	HSBC		
Top 5 Insurers	Suncorp Group	iag	Allianz	QBE	medibank For Better Health			
Top 5 Telco and Media	facebook	OPTUS		News Corp Australia	FOXTEL			
Top 8 FMCG and Key Retailers	Coca-Cola	Unilever	P&G		the fresh food people woolworths	MYER	THE LINDE GROUP	ebay

QUANTIUM

PERSONALISATION USES THE COMPLEX RELATIONSHIP BETWEEN EVERY CUSTOMER AND EVERY PRODUCT



woolworths rewards

Turn your shopping into dollars

Your rewards? Start earning Woolworths Dollars

Hi Adam,

[illegible]

Offers selected just for you

[illegible]

Members get more

[illegible]

VIEW MORE OFFERS

Stock availability, colours and styles may vary by store.

Informs wide range of business applications…

✓

Contact customers with the right offers

Across all channels – DM, eDM, online, etc.

• • • • •

Optimising for NPS or revenue growth or other metrics

✓

Provide customers with a store laid out just for them – tailored ranging online

Use full knowledge of customer relationship to inform online interactions

• • • • •

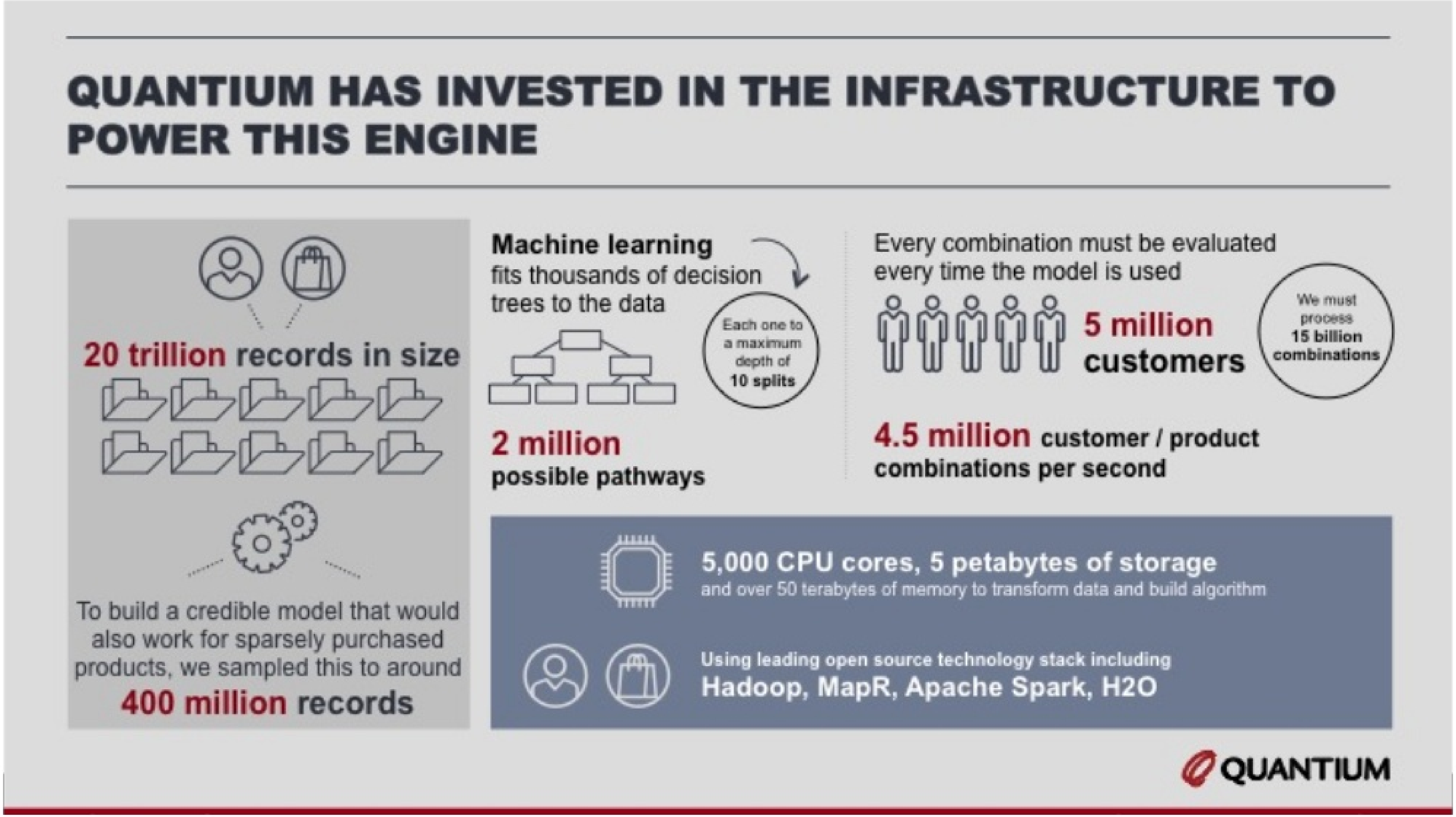
Relevance informs search result, basket autofill or more simply basket suggestions

✓

Bring customer understanding to the centre of business processes

Use insight contained inside the algorithm to inform loyalty strategy, promotional strategy, ranging and marketing for bricks-and-mortar stores

QUANTIUM



QUANTIUM HAS INVESTED IN THE INFRASTRUCTURE TO POWER THIS ENGINE

20 trillion records in size

To build a credible model that would also work for sparsely purchased products, we sampled this to around

400 million records

Machine learning fits thousands of decision trees to the data

Each one to a maximum depth of **10 splits**

2 million possible pathways

Every combination must be evaluated every time the model is used

5 million customers

We must process **15 billion combinations**

4.5 million customer / product combinations per second

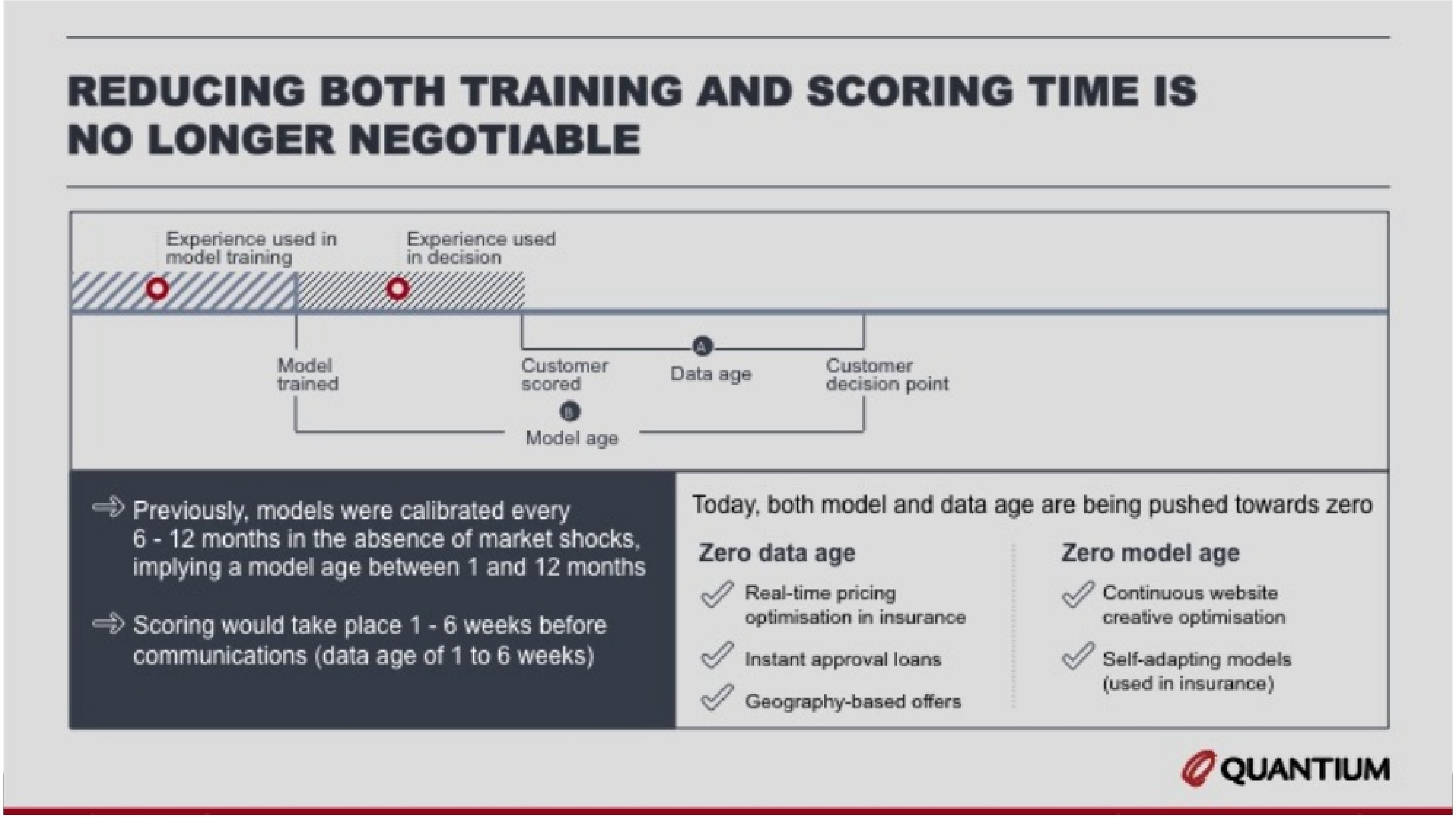
5,000 CPU cores, 5 petabytes of storage

and over 50 terabytes of memory to transform data and build algorithm

Using leading open source technology stack including

Hadoop, MapR, Apache Spark, H2O

QUANTIUM



REDUCING BOTH TRAINING AND SCORING TIME IS NO LONGER NEGOTIABLE

Experience used in model training
Experience used in decision
Model trained
Customer scored
A
Data age
Customer decision point
B
Model age

- Previously, models were calibrated every 6 - 12 months in the absence of market shocks, implying a model age between 1 and 12 months
- Scoring would take place 1 - 6 weeks before communications (data age of 1 to 6 weeks)

Today, both model and data age are being pushed towards zero

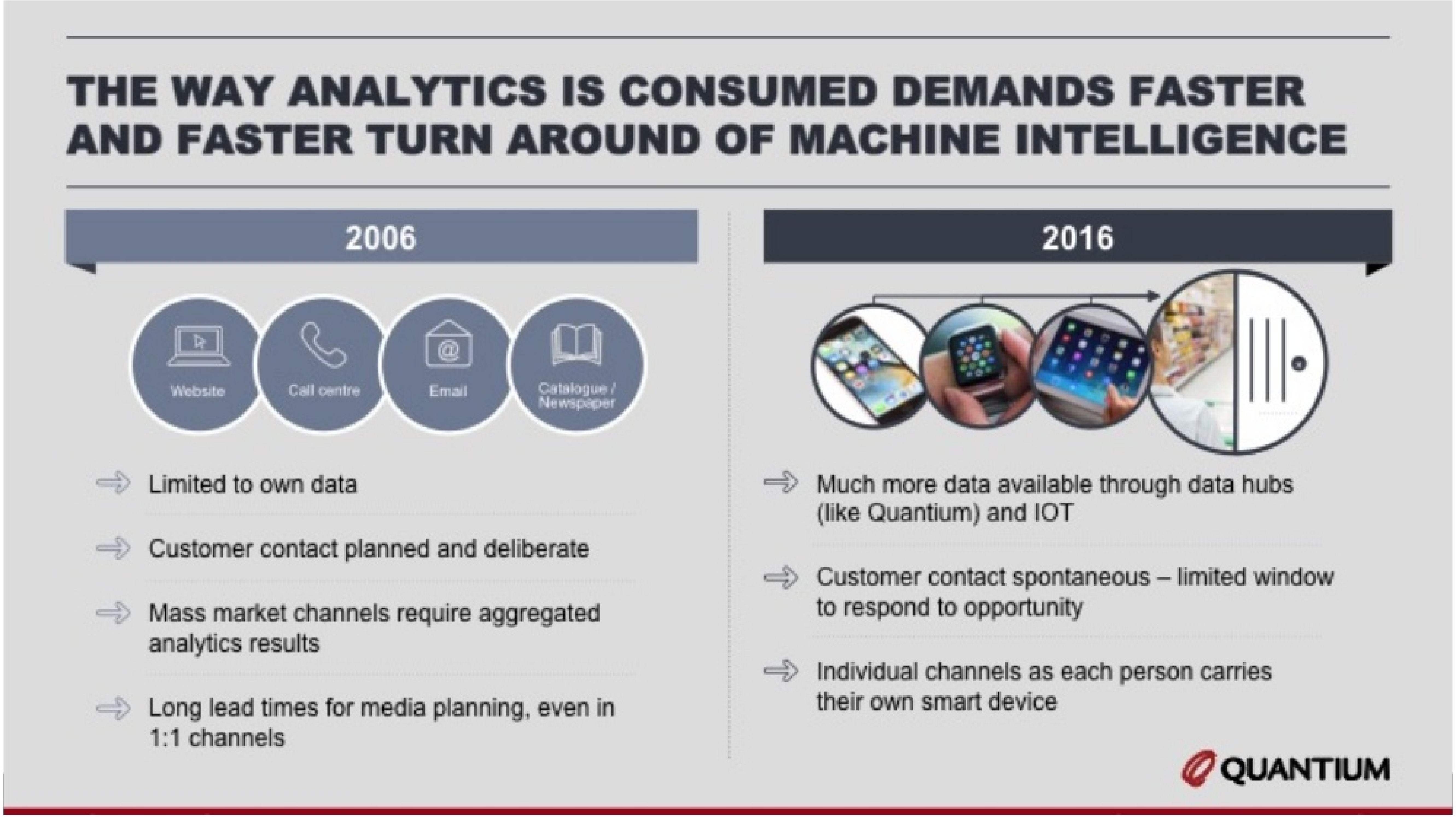
Zero data age

- ✓ Real-time pricing optimisation in insurance
- ✓ Instant approval loans
- ✓ Geography-based offers

Zero model age

- ✓ Continuous website creative optimisation
- ✓ Self-adapting models (used in insurance)

QUANTIUM


Website
Call centre
Email
Catalogue / Newspaper
QUANTIUM

THE WAY ANALYTICS IS CONSUMED DEMANDS FASTER AND FASTER TURN AROUND OF MACHINE INTELLIGENCE

2006

- Limited to own data
- Customer contact planned and deliberate
- Mass market channels require aggregated analytics results
- Long lead times for media planning, even in 1:1 channels

2016

- Much more data available through data hubs (like Quantium) and IOT
- Customer contact spontaneous – limited window to respond to opportunity
- Individual channels as each person carries their own smart device

UNITED HEALTH GROUP (UHG) FRAUD ANALYTICS

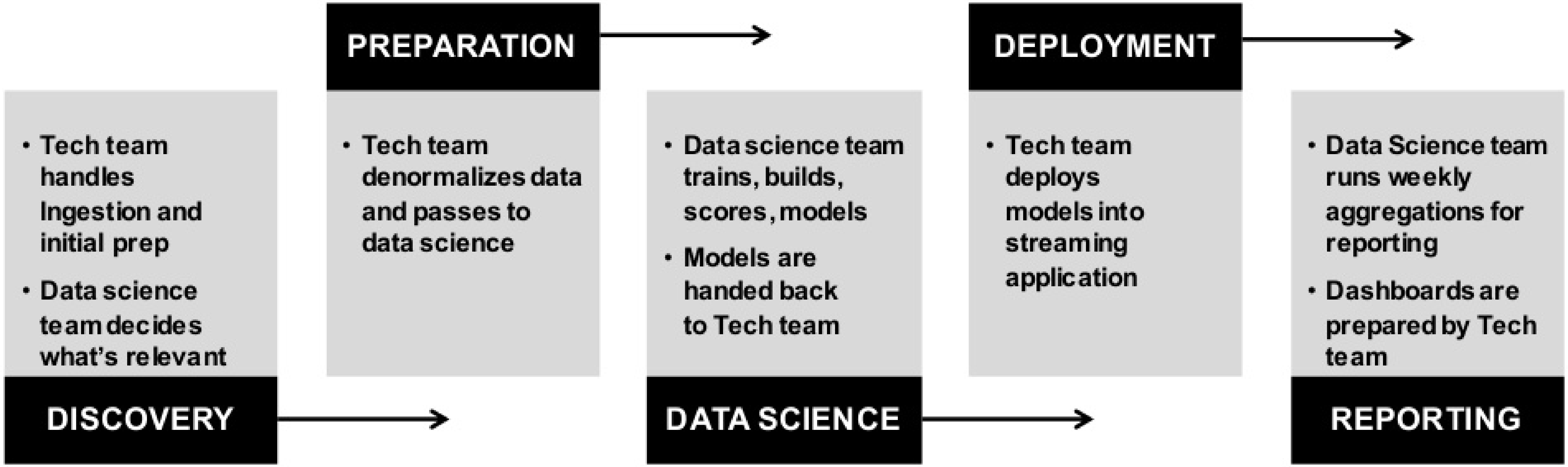
ADVANCED RESEARCH & ANALYTICS GROUP

- **Batch job** for weekly aggregates of claims processes **30 million records**

- The models are deployed into a Streaming Application **(Spark Streaming**) and used to flag suspicious or fraudulent claims for human review

- Models are built in **Python (TorchML) and Spark (MLLib).** Scoring is done in H2O POJO library

PIPELINE PHASES



DISCOVERY

- Tech team handles Ingestion and initial prep
- Data science team decides what's relevant

PREPARATION

- Tech team denormalizes data and passes to data science

DATA SCIENCE

- Data science team trains, builds, scores, models
- Models are handed back to Tech team

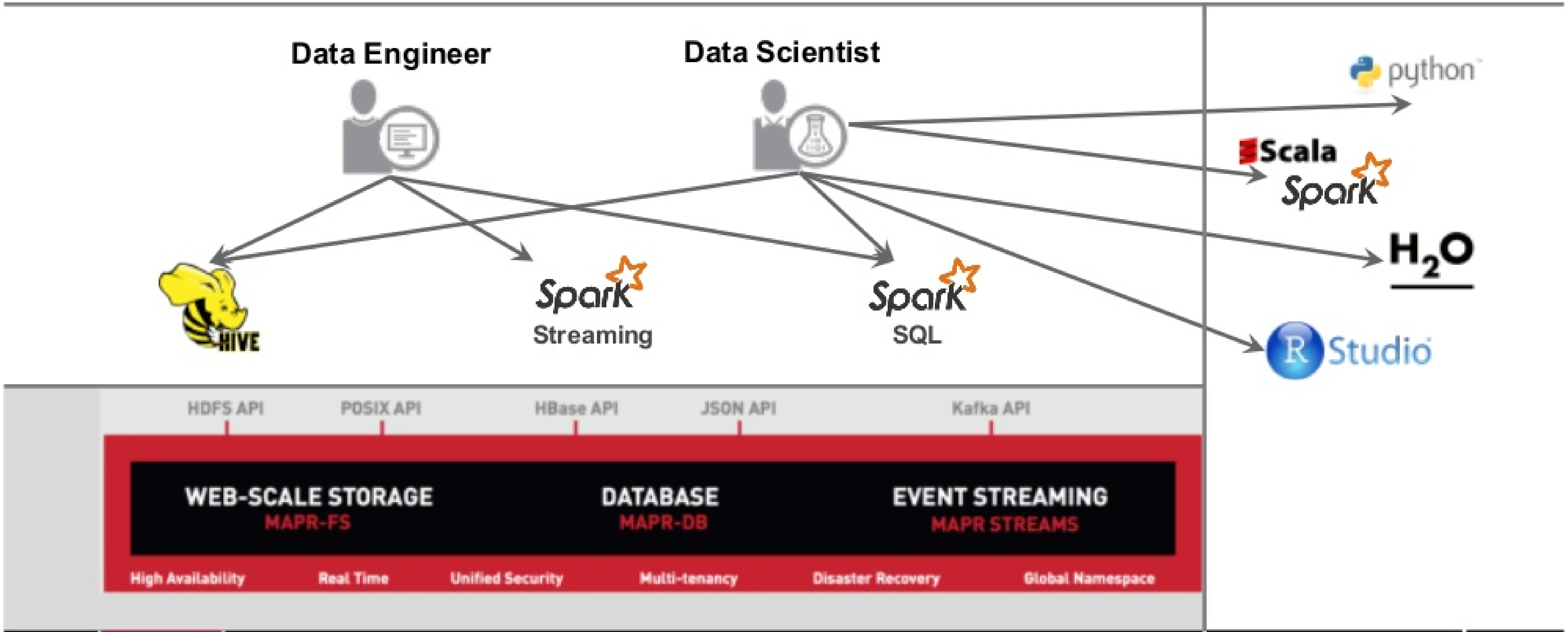
DEPLOYMENT

- Tech team deploys models into streaming application

REPORTING

- Data Science team runs weekly aggregations for reporting
- Dashboards are prepared by Tech team

UHG ARA MACHINE LEARNING ARCHITECTURE



Data Engineer

Data Scientist

HIVE

Spark Streaming

Spark SQL

python

Scala

Spark

H_2O

RStudio

HDFS API

POSIX API

HBase API

JSON API

Kafka API

WEB-SCALE STORAGE
MAPR-FS

DATABASE
MAPR-DB

EVENT STREAMING
MAPR STREAMS

High Availability

Real Time

Unified Security

Multi-tenancy

Disaster Recovery

Global Namespace

UNIQUE MAPR FEATURES AT PLAY

ALSO APPLICABLE FOR PROD ENV.

1. **VOLUMES**

   Data Scientists assigned a specific volume with right level of security policies

   Quotas, Label based scheduling for job prioritization

2. **SNAPSHOTS**

   Versioned Training and Validation Datasets stored efficiently

   New models can be run against these datasets on demand

3. **RANDOM READ/WRITE NFS**

   Modeling language of choice with access to the data in the cluster

   Specialized libraries (C/C++) work out of the box

   Easy to bring new datasets into the mix

CHALLENGES WITH M/L MODELING AND DEPLOYMENT

MODELING PHASE

- Safe sandbox to play in
- Language of choice
- Library of choice
- Reusable ETL functions
- Reusable models

PRODUCTION PHASE

- Variables: libraries, config.
- Hardware/Infra flexibility
- Easy scalability
- Model deployment flexibility
- Model evolution

CHALLENGES WITH M/L MODELING AND DEPLOYMENT

MODELING PHASE

- Safe sandbox to play in
- Language of choice
- Library of choice
- Reusable ETL functions
- Reusable models

PRODUCTION PHASE

- Variables: libraries, config.
- Hardware/Infra flexibility
- Easy scalability
- Model deployment flexibility
- Model evolution

CHALLENGES WITH M/L MODELING AND DEPLOYMENT

MODELING PHASE

- Safe sandbox to play in
- Language of choice
- Library of choice
- Reusable ETL functions
- Reusable models

PRODUCTION PHASE

- Variables: libraries, config.
- Hardware/Infra flexibility
- Easy scalability
- Model deployment flexibility
- Model evolution

SPARK SUMMIT 2017

Thank You.

snori@mapr.com

Distributed Deep Learning

Apache Spark and MapR-DB JSON Integration