AWS re:Invent

ANT202-R

Turbocharge your Spark performance with Amazon EMR

Joseph Marques
Principal Engineer
Amazon Web Services

Peter Gvozdjak
Senior Manager Engineering
Amazon Web Services

AWS re:Invent

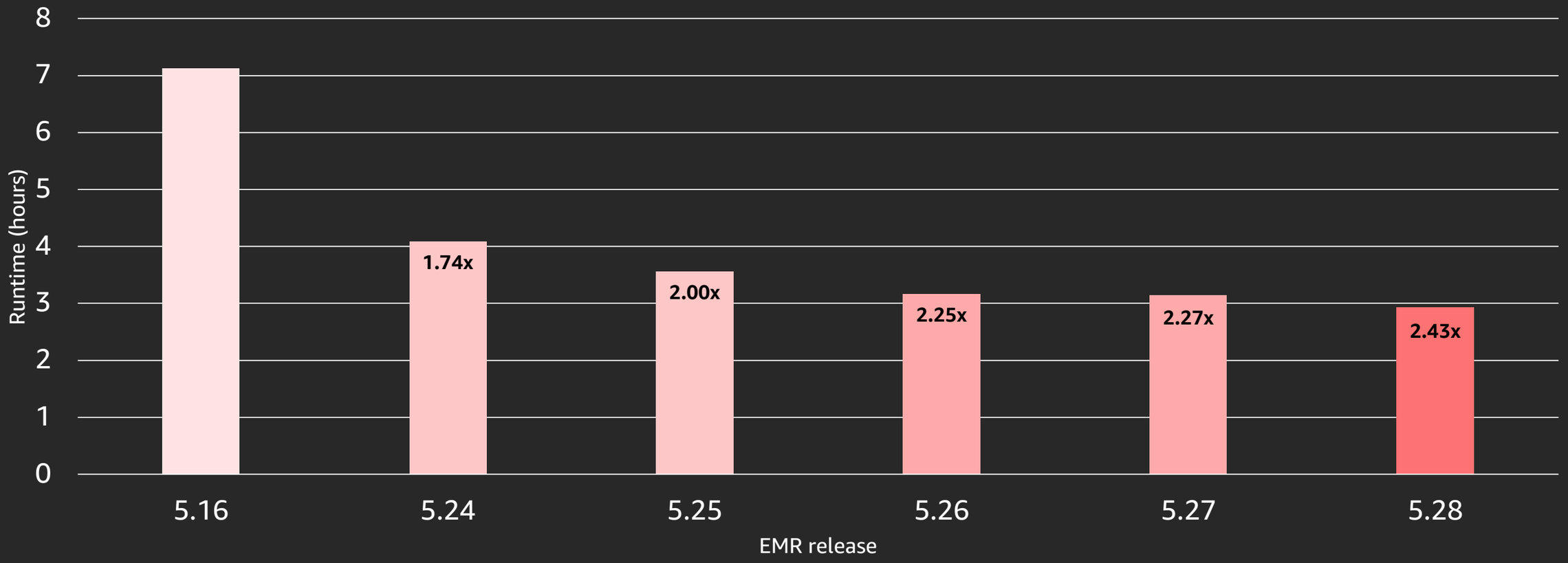
aws

What will you learn

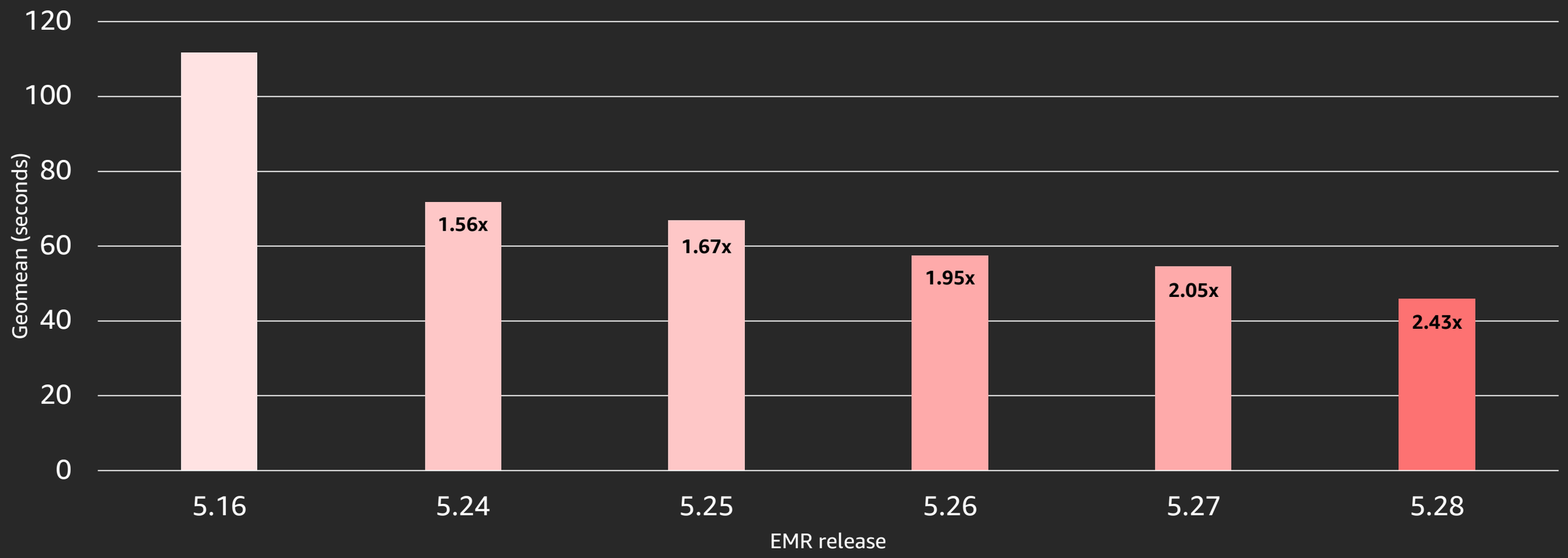
Performance results

Optimization deep dive

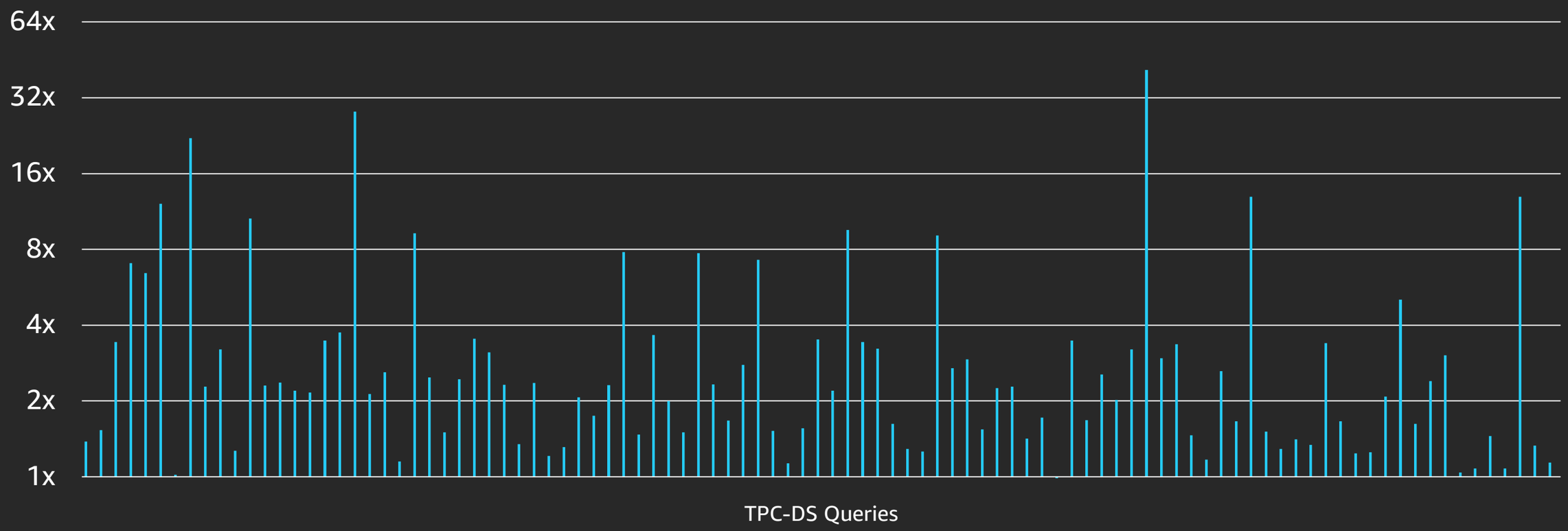
Dollar savings



Time savings



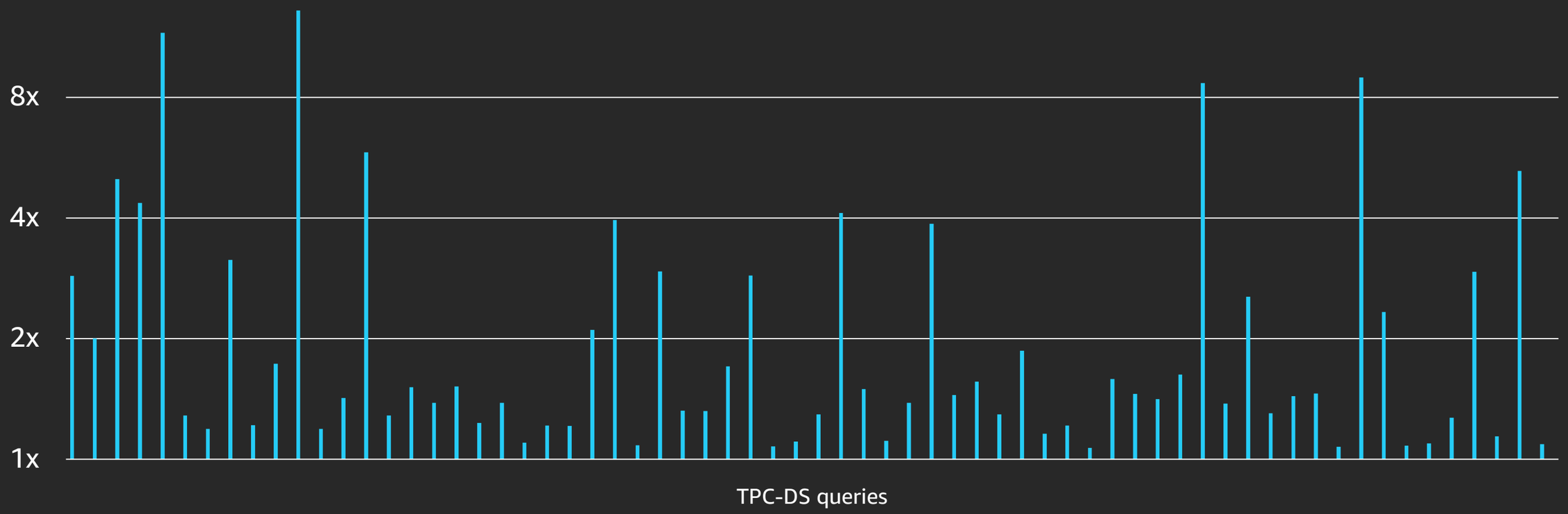
10 TB TPC-DS benchmark



Optimization deep dive

- Configuration
  - CPU/disk ratios, driver/executor conf, native overheads
- Planning/Optimization
  - Dynamic partition pruning, join reordering
- Job startup

Planning/Optimization: Dynamic partition pruning



Related breakouts

ANT301 – Build & optimize Spark data pipelines for incremental data processing

ANT308 – Deep dive into running Apache Spark on Amazon EMR

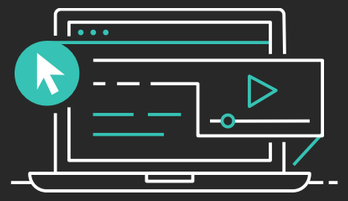
CMP336 – Save on big data workloads like Apache Spark and Hadoop

CMP404 – Running Big Data clusters on Amazon EMR with Spot Instances

CON412 – Performance-tuning tips and tricks for Apache Spark on Kubernetes

Learn big data with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills



New free digital course, Data Analytics Fundamentals, introduces Amazon S3, Amazon Kinesis, Amazon EMR, AWS Glue, and Amazon Redshift

Classroom offerings, including Big Data on AWS, feature AWS expert instructors and hands-on labs

Validate expertise with the **AWS Certified Big Data - Specialty** exam or the new **AWS Certified Data Analytics - Specialty** beta exam

Visit aws.amazon.com/training/paths-specialty/

aws training and certification

Thank you!

Joseph Marques

jmarques@amazon.com

Peter Gvozdjak

petergv@amazon.com

AWS re:Invent

aws

Please complete the session survey in the mobile app.