# An overview of Amazon Athena

Julien Simon, Principal Technical Evangelist, AWS
julsimon@amazon.fr
@julsimon

# Never trust the first image on Google!


"*Amazon Athena is a professional wrestler*" 8-|

On second thought, that's quite relevant!

amazon web services

# Amazon Athena is a professional data wrestler!

- New service announced at re:Invent 2016
- Run interactive SQL queries on S3 data
  - No need to load or aggregate data: ‘schema-on-read’
  - S3 data is never modified
  - Cross-region buckets are supported
- No infrastructure to create, manage or scale
- Availability: us-east-1, us-west-2
- Pricing: $5 per Terabyte scanned
  - Scanned data rounded off to the nearest 10MB
  - Stored data: S3 pricing applies

AWS re:Invent 2016: Introducing Athena (BDA303) https://www.youtube.com/watch?v=DxAuj_Ky5aw

# Athena queries

- Service based on Presto (already available in Amazon EMR)
- Table creation: Apache Hive DDL
  - CREATE EXTERNAL_TABLE only
  - CREATE TABLE AS SELECT is not supported
- ANSI SQL operators and functions: what Presto supports
- Unsupported operations
  - User-defined functions
  - Stored procedures
  - Any transaction found in Hive or Presto

https://prestodb.io
http://docs.aws.amazon.com/athena/latest/ug/known-limitations.html

# Data formats supported by Athena

- Unstructured
  - Apache logs, with customizable regular expression
- Semi-structured
  - Comma-separated values (CSV)
  - Tab-separated values (TSV)
  - Text File with custom delimiters
  - JSON
- Structured
  - Apache Parquet
  - Apache ORC (Optimized Row Columnar)
- Compression formats: Snappy, Zlib, GZIP (no LZO)
  - Less I/O → better performance and cost optimization

https://parquet.apache.org/

https://orc.apache.o

# Using columnar formats for fun and profit

- Apache Parquet and Apache ORC
- Less I/O, better performance (typically 5x-6x faster)

- You can convert your data to columnar formats with EMR and Hive.
  - http://docs.aws.amazon.com/athena/latest/ug/convert-to-columnar.html
- Migrate External Table Definitions from Hive to Athena
  - https://aws.amazon.com/fr/blogs/big-data/migrate-external-table-definitions-from-a-hive-metastore-to-amazon-athena/

# EMR, Redshift or Athena?

- EMR
  - Scale-out data crunching
  - Code / HQL queries running complex transformations on (un)structured data
  - Rich Apache Hadoop ecosystem, at the cost of complexity

- Redshift
  - Petabyte-scale enterprise data warehouse
  - ETL, complex SQL queries and joins on long-lived, structured data
  - Many techniques for performance optimization

- Athena
  - Answering questions in minutes, with zero infrastructure plumbing
  - Ad-hoc SQL queries, with probably a few or no joins
  - Emphasis on simplicity, not on raw performance

# Athena in a nutshell

- Run ad-hoc SQL queries on S3 data in minutes
- No infrastructure
- Multiple input formats supported
- Pretty fast!
- A simpler, very cost-efficient alternative to EMR and Redshift for ad-hoc analysis

# Demo ☺

# Thank you!

Julien Simon, Principal Technical Evangelist, AWS
julsimon@amazon.fr
@julsimon

amazon web services