The Big Data Battle

Julien Simon

Principal Technical Evangelist

julsimon@amazon.fr

@julsimon

The rules

Same data set

Same SQL queries

Same data format

Same data location

Vanilla setup

Amazon EMR

- Managed service for the Hadoop ecosystem
- Launched in April 2009
  - Apache Hive https://cwiki.apache.org/confluence/display/Hive/Home
- Apache Spark available since June 2015
  - Spark SQL http://spark.apache.org/sql
- Pricing: per instance per hour
- Setup: 10 c4.4xlarge instances
  - 16vCPUs, 30GB RAM each
  - Total cost: $7.96 per hour (on-demand price, us-east-1)

https://aws.amazon.com/emr/

Amazon Athena

- Run read-only SQL queries on S3 data
- Service announced at re:Invent 2016
  - Based on Presto http://prestodb.io
- Fully managed
  - No infrastructure to create, manage or scale
  - No data loading, no indexing, no nothing
- Pricing: $5 per Terabyte scanned

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

Amazon Redshift Spectrum

- Extends the power of Redshift beyond data stored on local disks
- Run read-only SQL queries on S3 data
- New service announced in February 2017
- Pricing:
  - Redshift: per instance per hour
  - Spectrum: $5 per Terabyte scanned
- Setup: 4 dc1.8xlarge nodes (+ managed Spectrum fleet)
  - 32 vCPUs, 244 GB RAM, 2.56TB SSD each
  - Total cost: $19.20 per hour (on-demand price, us-east-1)

Amazon Redshift Spectrum: Now Query Exabytes of Data in S3 https://www.youtube.com/watch?v=gchd2sDhSuY

GDELT Data set

- Global Database of Events, Language and Tone Database
  - 300 categories of political & diplomatic activities around the world
  - Geo-referenced to the city
  - Dating back to 1979 and updated daily
  - http://www.gdeltproject.org/
  - https://aws.amazon.com/public-datasets/gdelt/
  - https://medium.com/@julsimon/exploring-the-gdelt-data-set-with-amazon-athena-a6f7b1d67a6e

- 1612 CSV files in S3 (150 GB)
- 1 table (+ reference tables), 58 columns, 450M lines

Using columnar formats for fun and profit

- Hive makes it easy to convert from CSV to Parquet

  https://docs.aws.amazon.com/athena/latest/ug/convert-to-columnar.html

- Full scan on GDELT data set
  - CSV uncompressed: 136GB scanned, $0.13
  - Parquet compressed: 2.2GB scanned, $0.002

Thank you!

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

amazon web services