Big Data answers in seconds with Amazon Athena

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

Big Data the way it should be

Questions
(not data!)

We shouldn't have to care about how this really works !

Answers

We shouldn't have to mess with this at all

Data

Want to build it yourself? You need to master this

- Planning capacity for storage and compute
- Handling different data formats, structured and unstructured (CSV, JSON, Parquet, Avro, etc.)
- Learning complex programming models and languages (Map Reduce, Spark, Scala, etc.)
- Keeping costs under control
- Availability, performance, security and a few more

Need help with your own Hadoop?

claranet

cloudera

- Claranet: AWS Premier Consulting Partner
- They can build and run your Cloudera Enterprise platforms on top of AWS
- Claranet has certified AWS and Cloudera experts
- Security & compliance is built-in (ISO 27001, PCI-DSS)
- 24/7 support is available
- Learn more on booth 110. Tell them I sent you ;)

https://www.claranet.fr

THERE IS ANOTHER
quickmeme.com

Amazon Athena

- New service announced at re:Invent 2016
- Run read-only SQL queries on S3 data
- No data load, no indexing, no nothing
- No infrastructure to create, manage or scale
- Availability: us-east-1, us-east-2, us-west-2
- Pricing: $5 per Terabyte scanned

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 Data Definition Language
  - CREATE EXTERNAL_TABLE
- ANSI SQL operators and functions: what Presto supports
- Unsupported operations
  - User-defined functions (UDF or UDAFs)
  - 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
  - delimiter-separated values (CSV, OpenCSV)
  - Tab-separated values (TSV)
  - JSON
- Structured
  - Apache Parquet https://parquet.apache.org/
  - Apache ORC https://orc.apache.org/
  - Apache Avro https://avro.apache.org/
- Compression (Snappy, Zlib, GZIP) & partitioning

Demo

GDELT Data set

- Global Database of Events, Language and Tone Database
  - 300 categories of political & diplomatic activities around the world
  - Georeferenced to the city
  - Dating back to January 1, 1979
  - http://www.gdeltproject.org/

- 1543 CSV files in S3 (146 GB)
- 1 table (+ reference tables), 58 columns, 441M lines
- https://aws.amazon.com/public-datasets/gdelt/

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

- Large request
  - CSV uncompressed : 26 seconds, 136GB scanned, $0.13
  - Parquet compressed : 4 seconds, 2.2GB scanned, $0.002

Athena in a nutshell

- Run SQL queries on S3 data
- No infrastructure
- Multiple input formats supported
- Pretty fast!
- A simple, very cost-efficient option for ad-hoc analysis

AWS User Groups

Meetup

Lille
Paris
Rennes
Nantes
Bordeaux
Lyon
Montpellier
Toulouse
Côte d’Azur (new!)

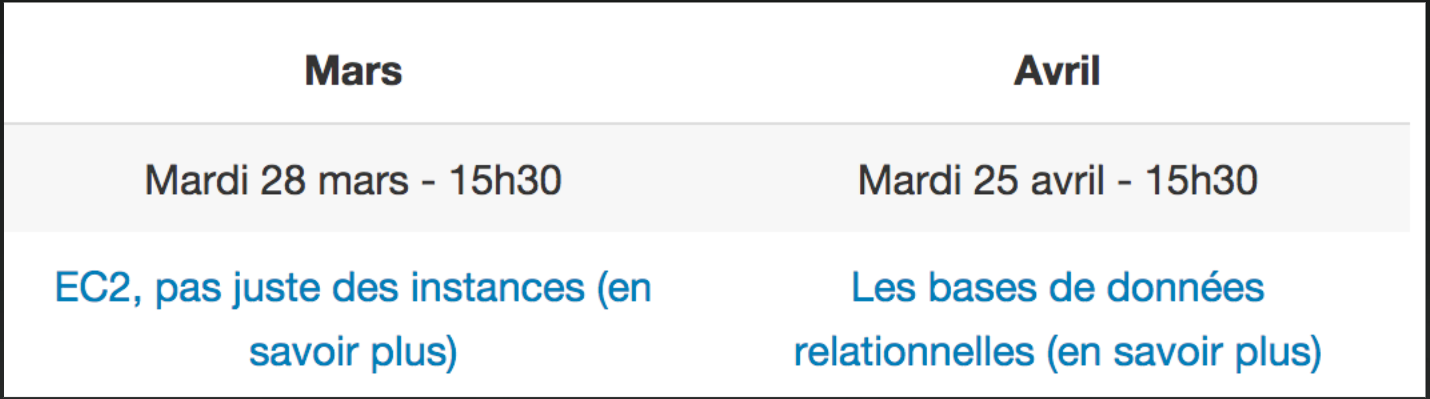
facebook.com/groups/AWSFrance/

@aws_actus

YouTube “Amazon Web Services France”

amazon web services

https://aws.amazon.com/fr/events/webinaires/



Mars	Avril
Mardi 28 mars - 15h30	Mardi 25 avril - 15h30
EC2, pas juste des instances (en savoir plus)	Les bases de données relationnelles (en savoir plus)

Thank you!

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