rackspace®

the #1 managed cloud company

AWS Athena

amazon web services

AWS DATABASES

Amazon RDS

DynamoDB

Amazon Redshift

Amazon EMR

Athena

Athena - Basics

- *Serverless* interactive query service
- PrestoDB (Facebook) implementation
- Works on data stored in S3 (support for compression and encryption)
- But is still quite fast

Athena - Basics

- Uses Apache Hive Data definition language (DDL)
- Supports ANSI SQL
- Pay as you go model (like Lambda). Charges based on data scanned by the query. 5$ per TB of data scanned
- No charge for creating tables
- S3 costs are separate

Athena – Concepts

- Tables - Metadata that describes your data similar to traditional database tables.
- Tables are like views.
- For e.g. You can delete table definitions without impacting the underlying S3 data

Athena – Concepts

- Databases - Logical grouping of tables. (catalog or a namespace)
- **SerDe** - Serializer/Deserializer - libraries that tell Hive how to interpret data formats.
  - Apache Web Logs, CSV, JSON, ORC (Optimized Row Columnar), Apache Parquet etc
  - Data can also be compressed in GZIP to save costs
  - Data can be partitioned with some formats to improve performance and save costs

Athena - Limitations

- No support for transactions. This includes any transactions found in Hive or Presto. (When you create, update, or delete tables, those operations are guaranteed ACID-compliant.)
- No support for stored procedures

Athena - Limitations

- Limited to 2 AWS regions
  - You can use Athena to query underlying Amazon S3 bucket data that's in a different region
- API/SDK support (Limited to JDBC driver)

Athena – Query Limitations

CREATE VIEW REVENUE0 (SUPPLIER_NO, TOTAL_REVENUE) AS SELECT ...

SELECT S_SUPPKEY, S_NAME FROM SUPPLIER, REVENUE0 WHERE ...

DROP VIEW REVENUE0

Vs

WITH REVENUE0 AS (SELECT ...) SELECT S_SUPPKEY, S_NAME FROM SUPPLIER REVENUE0 WHERE ...

Athena – Service Limits

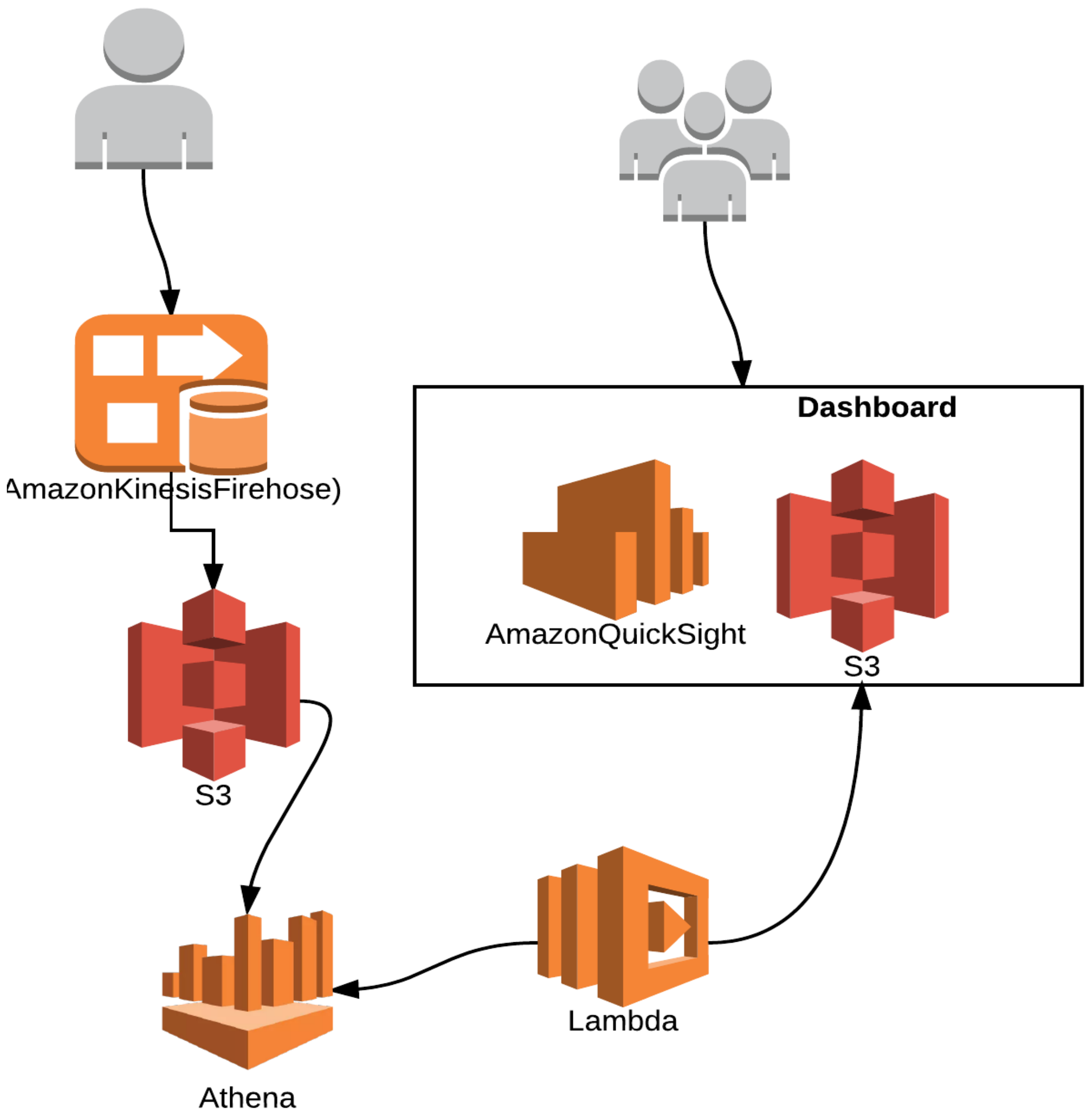
- Query timeout: 30 minutes
- Number of databases: 100
- Table: 100 per database
- Number of partitions: 20k per table
- S3 limits also impact it

Athena – Use Cases

- Querying of any data in S3. If latency is not critical and queries can be run in the background
  - For e.g analyzing log data in S3
  - Integrating with AWS Glue (ETL)

Athena – Use Cases

- Go serverless across the stack. API gateway can be used to accept requests which are handled by Lambda which in turn can leverage Athena for queries. The only persistent service used will be S3.



rackspace
the #1 managed cloud company

Athena – Use Cases

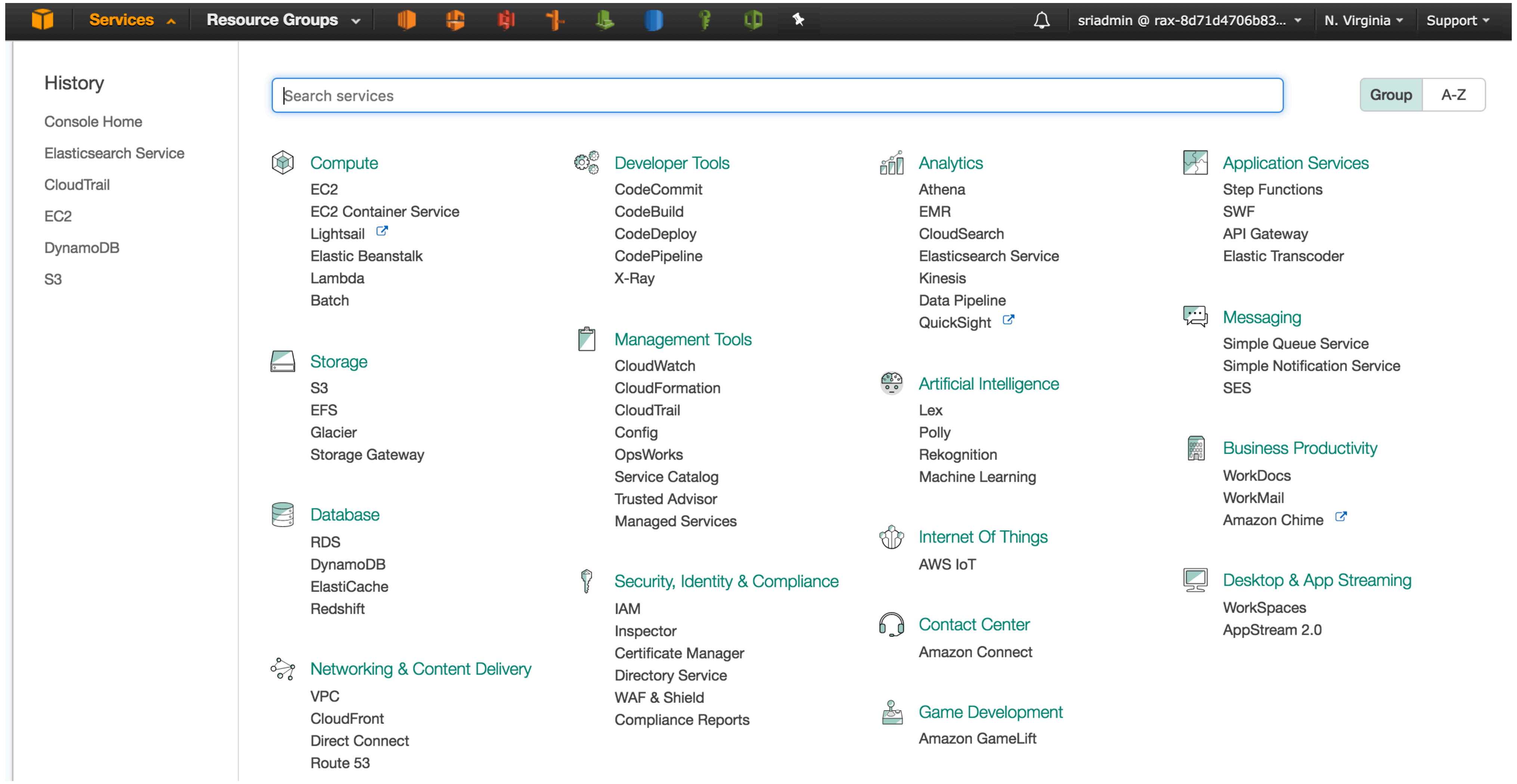
- Mix and match AWS services. Use an on-demand EMR cluster to process data and dump results to S3. Then use Athena to create adhoc tables and run reports.

- Redshift Spectrum uses Athena behind the scenes for a hybrid model (Redshift local + Redshift S3)

Athena – Use Cases

- “Facebook uses Presto for interactive queries against several internal data stores, including their 300PB data warehouse. Over 1,000 Facebook employees use Presto daily to run more than 30,000 queries that in total scan over a petabyte each per day”

Athena – Demo



https://github.com/rackerlabs/athena-playground

rackspace
the #1 managed cloud company

https://rax.io/aws-ks
https://github.com/rackerlabs/athena-playground

rackspace® | YOUR CLOUDS.
OUR EXPERTISE.