aws

COMMUNITY DAY

TEL AVIV

Serverless data pipeline and ETL

Michael Haberman | 2018

aws
COMMUNITY DAY

Feeling the buzz?

Everyone talks about serverless
Everyone talks about data

So lets merge them together!

aws
COMMUNITY DAY

Our Goal

Create a clickstream **data pipeline** with **ETL** process which is **serverless**!

aws

COMMUNITY DAY

Save the events

Basic transformation

Some Storage

Data is ready / after some time

Extract

Transform

Load

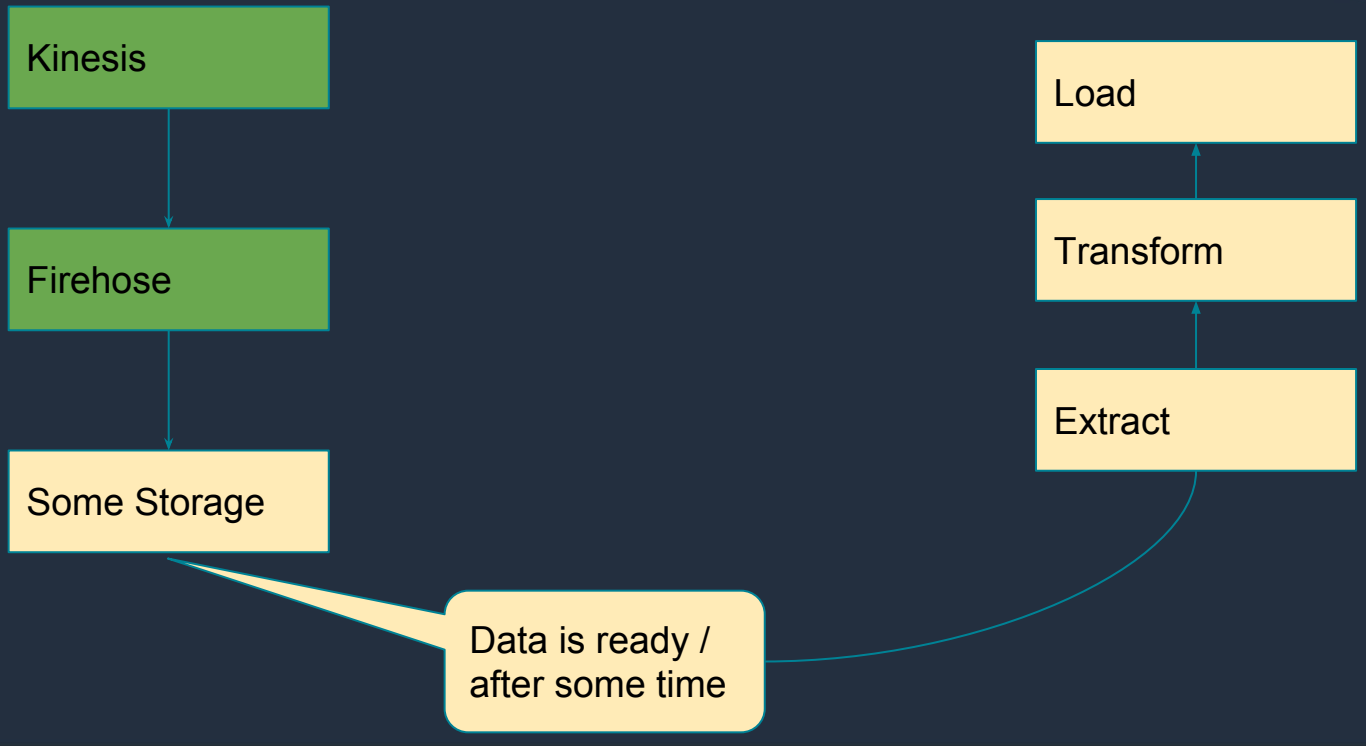
aws
COMMUNITY DAY

aws
COMMUNITY DAY

Kinesis

- Ingest big data - sharded
- Multiple producers and consumers
- Data retention (7 days max)
- Sub products
  - Data analytics - SQL queries over windows
  - Firehose - transformation (Lambda, json to parquet), really fast
- Scale defined by numbers of shard

aws
COMMUNITY DAY



aws
COMMUNITY DAY

Firehose

- Input
  - Kineses
  - Direct
- Transformation
  - Using Lambda
  - Changes schema and type using AWS Glue
- Output
  - S3
  - Redshift
  - Elastic search
- No scaling definition
- Paying per event + Lambda execution

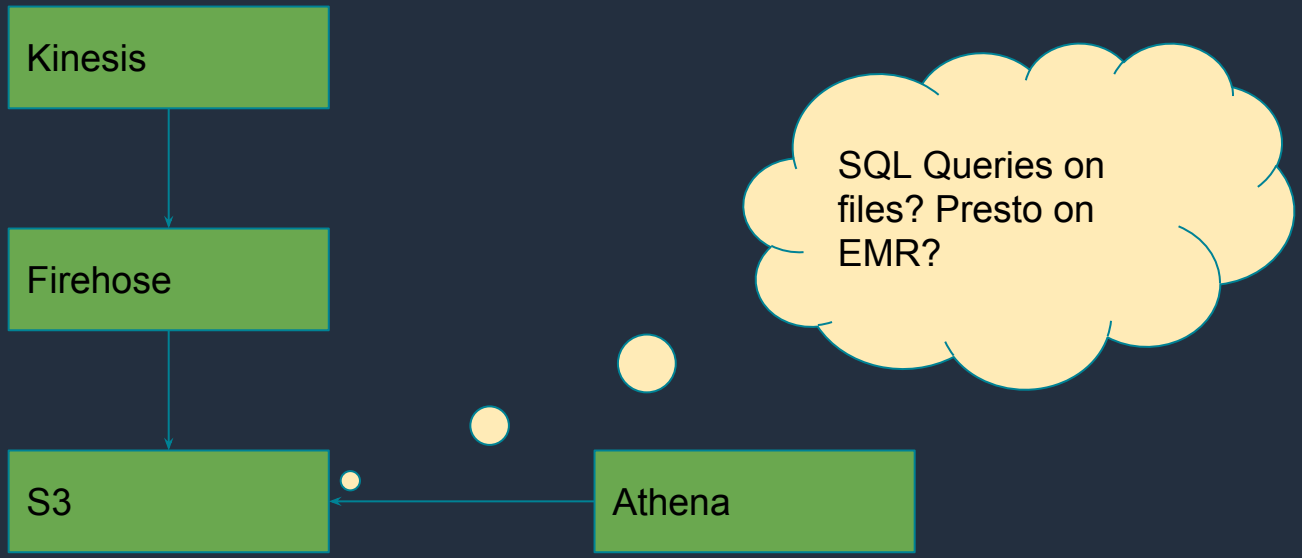
aws
COMMUNITY DAY

aws
COMMUNITY DAY

S3

- I guess you know S3…
- What if I want to query S3?

aws
COMMUNITY DAY

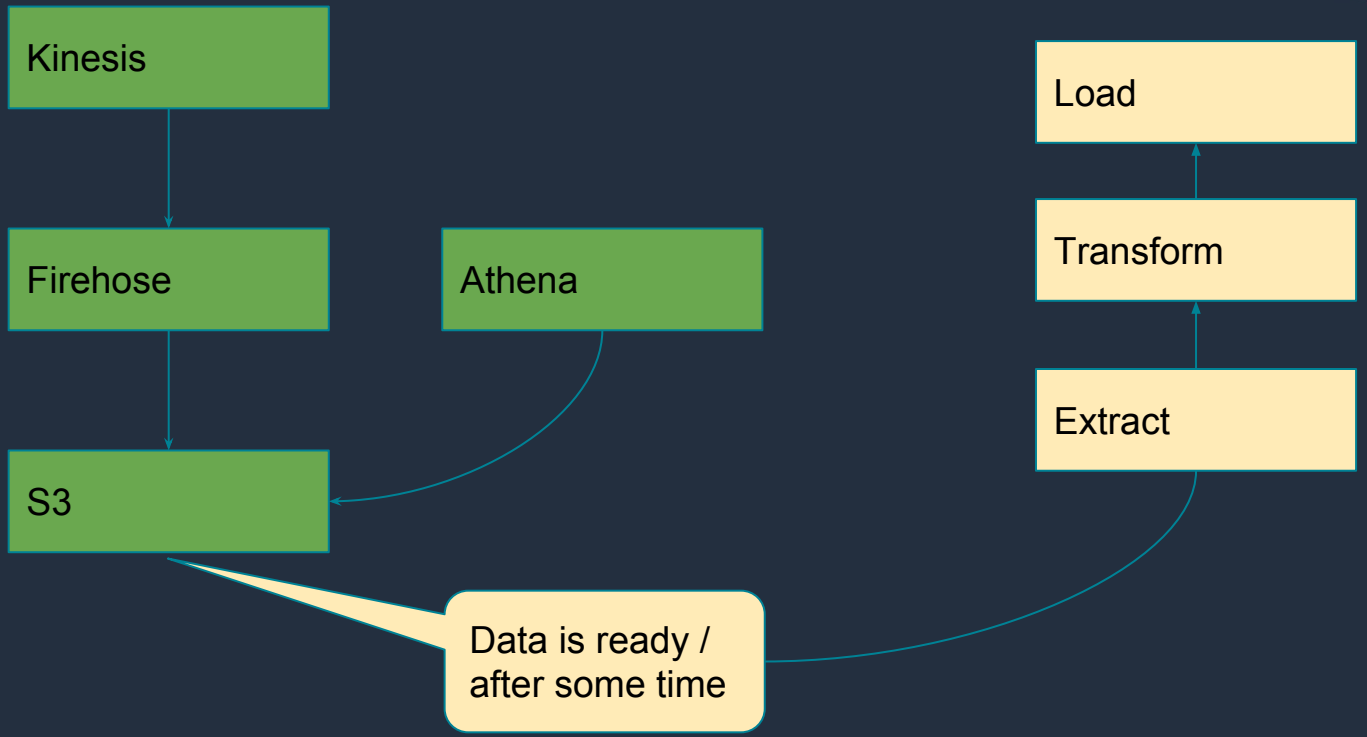


aws
COMMUNITY DAY

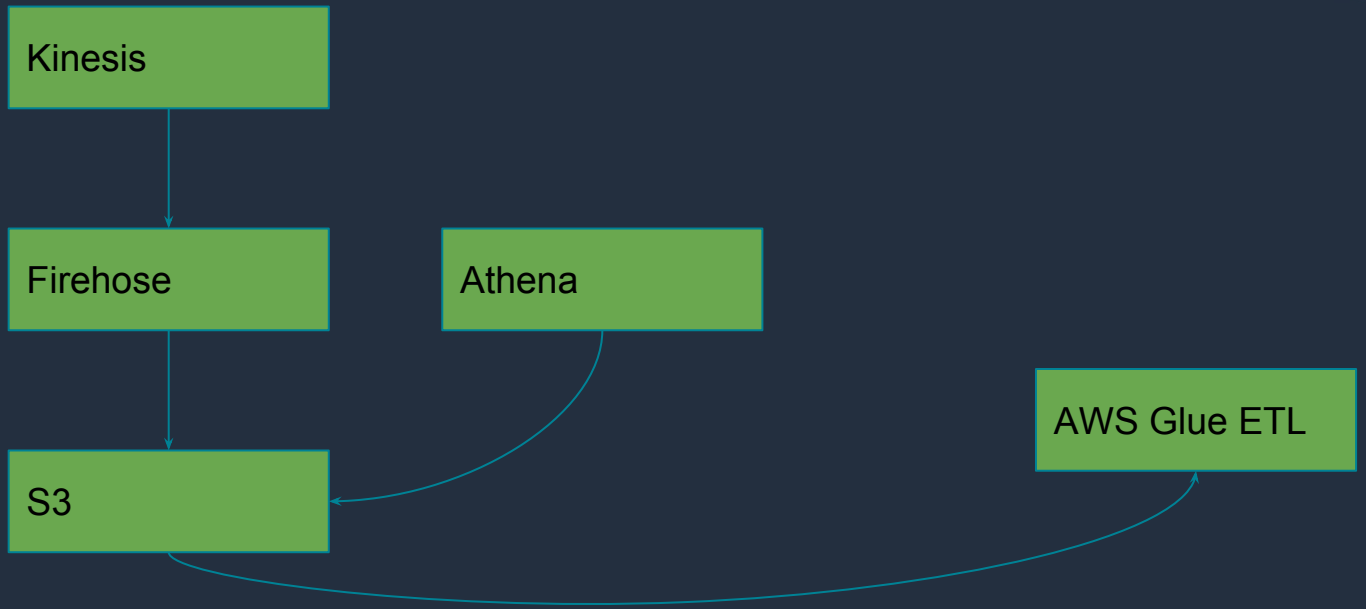
Athena

- Basically Presto
- Paying for scanned data (1TB = 5$)
- Using AWS Glue for metastore
- Can’t insert data

aws
COMMUNITY DAY



aws
COMMUNITY DAY



aws
COMMUNITY DAY

Glue

- Extract - using metastore
- Transform - python code
- Load - S3, Redshift

aws
COMMUNITY DAY

Quicksight

- Analytical dashboards

aws
COMMUNITY DAY

Thanks!

michael@topsight.io