



**AWS
re:Invent**

BDM303

JustGiving: Serverless Data Pipelines, Event Driven ETL, and Stream Processing

Richard T. Freeman, Ph.D., Lead Data Engineer and Architect, JustGiving

November 29, 2016

What to Expect from the Session

- Recap of some AWS services
- Challenges and requirements at JustGiving
- Big data pipelines and extract transform load
- Event-driven data platform at JustGiving
- Five patterns for scalable data pipelines
- Serverless recommendations and best practices

Recap of Some AWS Services

Managed Services



Amazon S3

- Distributed web scale object storage
- Highly scalable, reliable, and secure
- Pay only for what you use
- Supports encryption



AWS Lambda

- Runs code in response to triggers
- Amazon API Gateway requests
- Serverless
- Pay only when code runs
- Automatically scales and high availability



Amazon Redshift

- Managed, massively parallel, petabyte-scale data warehouse
- Load data from S3, DynamoDB, and EMR
- Extensive security features
- Columnar, JDBC/ODBC, ANSI SQL



Amazon EMR

- Batch and real-time processing
- Long-running or transient clusters
- Spot Instance
- Based on Apache Hadoop and Big Top

Amazon Kinesis

Amazon Kinesis Firehose



Easily load massive volumes of streaming data into S3, Amazon Redshift, and Amazon Elasticsearch Service

Amazon Kinesis Analytics



Easily analyze streaming data with standard SQL

Amazon Kinesis Streams



Build your own custom applications that process or analyze streaming data. Scales out using shards

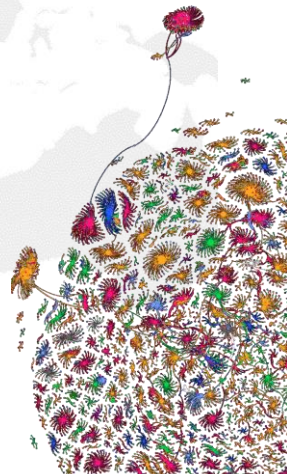
Challenges and Requirements at JustGiving

We are **JustGiving™**

A tech-for-good platform for events based fundraising, charities and crowdfunding


"Ensure no good cause goes unfunded"


- The #1 platform for online social giving in the world
- Peaks in traffic: Ice bucket, natural disasters
- Raised \$4.2bn in donations
- 28.5m users
- 196 countries
- 27,000 good causes
- GiveGraph
 - 91 million nodes
 - 0.53 billion relationships



Fundraising Page

Supporters 41

**Blaine and lizzie**
Good luck
\$50.00
16 days ago

**Brunch**
Our first race! Whey!
£10.00 + £2.50 Gift Aid
16 days ago

102%

£1,125.28

raised of £1,100 target
by 41 supporters

Donate

Share on Facebook

Story


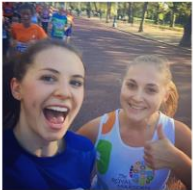
Please text SWJG £5 to 70070 to text donate :)

I'm running the Royal Parks Half Marathon with Lottie in October for Mind.

The sad reality is that more people now suffer from mental health issues than ever before and they should not be made to feel ashamed or isolated because of it. I've had those closest to

Read full story


Updates 2

**Sophie Weatherill**
Run done. Big thank you to everyone who helped me raise £1,125 for Mind. You're all bloody great and I couldn't have done it without you.

Share


Search for a charity, friend or project

Start Fundraising

Home Mike Menu



102%
£1,125.28
raised of £1,100 target
by 41 supporters
Donate
Share on Facebook

**Sophie's Royal Parks Half Marathon page**
Fundraising for Mind - The Mental Health Charity
Event: Royal Parks Half Marathon 2016, 09 Oct 2016

Story

Please text SWJG £5 to 70070 to text donate :)

I'm running the Royal Parks Half Marathon with Lottie in October for Mind.



The sad reality is that more people now suffer from mental health issues than ever before and they should not be made to feel ashamed or isolated because of it. I've had those closest to me affected by mental illness and heard first hand the challenges they face on a daily basis. So I hope that the money

Read full story


Share this story


Facebook Twitter Email


Updates 2


**Sophie Weatherill**
Run done. Big thank you to everyone who helped me raise £1,125 for Mind. You're all bloody great and I couldn't have done it without you.



Supporters 41

**Blaine and lizzie**
Good luck
\$50.00
10 days ago

**Brunch**
Our first race! Whey!
£10.00 + £2.50 Gift Aid
10 days ago

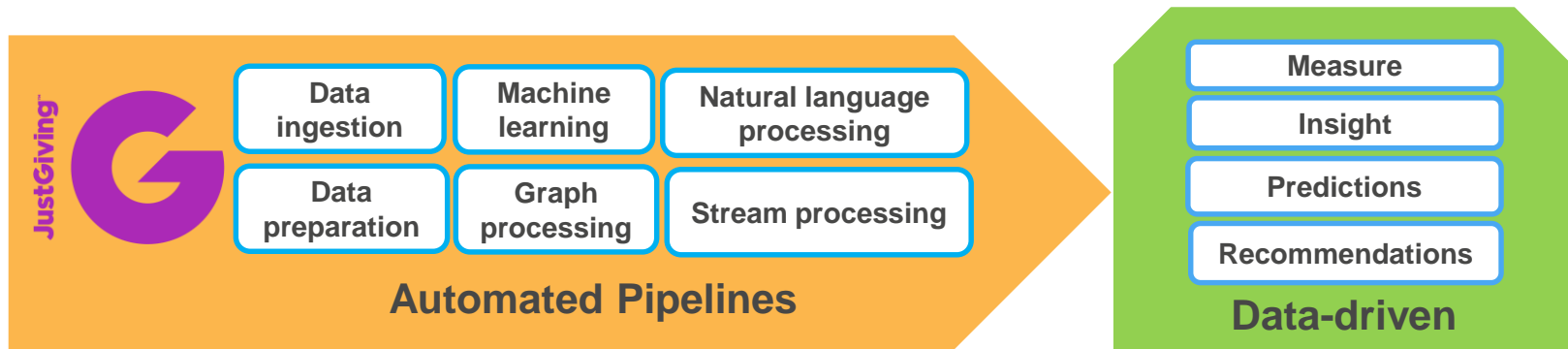
**KSBD**
Good luck on Sunday you beauty!
You're going to smash it! X
£30.00 + £7.50 Gift Aid
11 days ago

**Ben W**
Go for it!
£10.00 + £2.50 Gift Aid
12 days ago

**The JustGiving Team**
Wishing yourself and Lottie all the best in the Royal Parks Half Marathon. What a wonderful cause to be supporting!
£500.00
19 days ago

Our Requirements

- Limitation in existing SQL Server data warehouse
 - **Long running and complex queries** for data scientists
 - **New data sources**: API, clickstream, unstructured, log, behavioural data etc.
- Easy to add **data sources** and **pipelines**
- Reduce time spent on **data preparation** and **experiments**



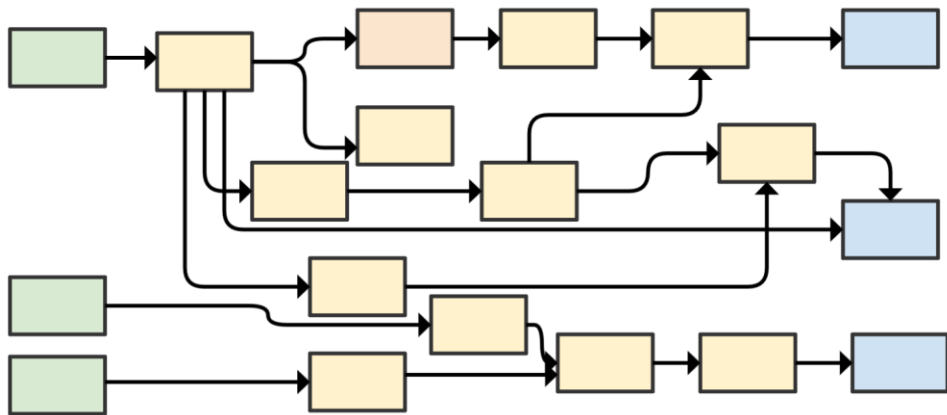
Big Data Pipelines and Extract Transform Load

Big Data Pipelines and Extract Transform Load

- WHY?
 - Automate data preparation
 - Run machine learning algorithms
 - Query and manipulate data efficiently
- Data schemas or models
- Support scheduled or triggered jobs
- Monitoring and failure notifications
- Re-run failed steps
- Traditionally workflow or directed acyclic graphs (DAGs)

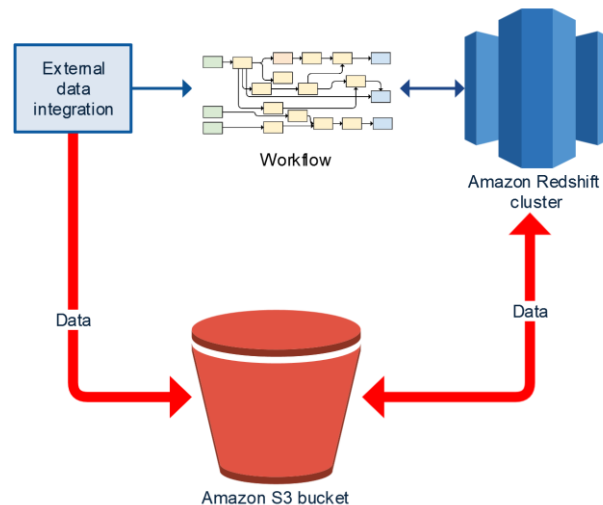
Challenges with Big Data Pipelines and ETL / ELT [1 of 2]

- DAG and ETL workflows can grow complex very quickly, e.g.
 - Pipelines as code or drawn graphically
 - Hand-drawn SQL statements
 - Product or vendor specific
 - Support for schema changes



Challenges with Big Data Pipelines and ETL / ELT [2 of 2]

- Have point-to-point integration
- Don't support several Amazon Redshift or EMR clusters well
- Complex replay or re-load
- Some do not support incremental loads without duplicates
- Different user interface and monitoring
- Not always highly available and resilient



Event-Driven Data Platform at JustGiving

Event Driven Data Platform at JustGiving [1 of 2]

- JustGiving developed in-house analytics and data science platform in AWS called **RAVEN**.
 - Reporting, Analytics, Visualization, Experimental, Networks
- Uses **event-driven** and **serverless pipelines** rather than DAGs
 - Messaging, queues, pub/sub patterns
 - Separate storage from compute
- Supports scalable event driven
 - ETL / ELT
 - Machine learning
 - Natural language processing
 - Graph processing
- Allows users to consume raw tables, data blocks, metrics, KPIs, insight, reports, etc.



Event Driven Data Platform at JustGiving [2 of 2]



RAVEN Platform Overview



Data Sources

Web Analytics

KISSmetrics



Amazon Kinesis Streams

logstash

Qualaroo

Example External API Sources

EVENT STORE

SurveyMonkey

ExactTarget

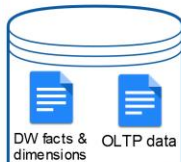
twitter

OPEN

OpenData

API

News & events feeds



DW facts & dimensions

OLTP data dimensions

Data Integration



AWS Lambda



Amazon SNS notification service topics



Raven API integration micro-services



Amazon SQS Redshift queues



Raven content & query manager



Raven Redshift integration



Raven Elastic MapReduce job runner



Raven ingestion service



AWS Identity & Access Management



Amazon Kinesis Firehose



AWS Data Pipeline

Data Storage



Amazon DynamoDB



Amazon Redshift clusters



Amazon S3 bucket



Amazon Elastic MapReduce

Analytics & Data Science



SQL

ExactTarget

tableau

python

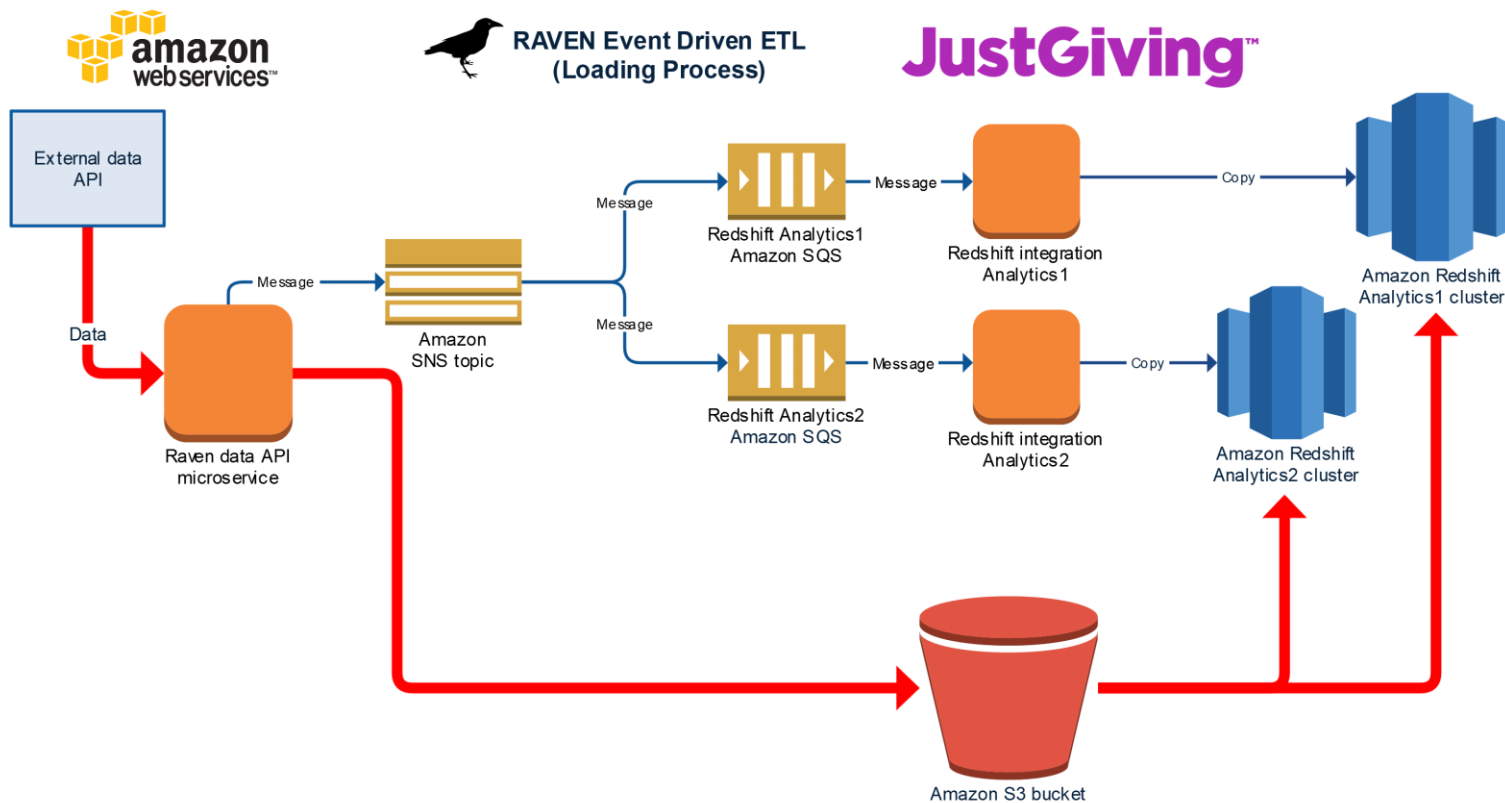
Spark



Amazon Elastic MapReduce

hadoop

Event Driven ETL and Data Pipeline Patterns [1 of 2]



Event Driven ETL and Data Pipeline Patterns [2 of 2]

Pros

- Adds lots of flexibility to loading
- Supports incremental and multi-cluster loads
- Simple reload on failure, e.g., send an Amazon SNS/Amazon SQS message
- Basic sequencing can be done within message

Cons:

- Does not support complex loading workflows and sequencing
- No native FIFO support in SQS (support added Nov 17, 2016)

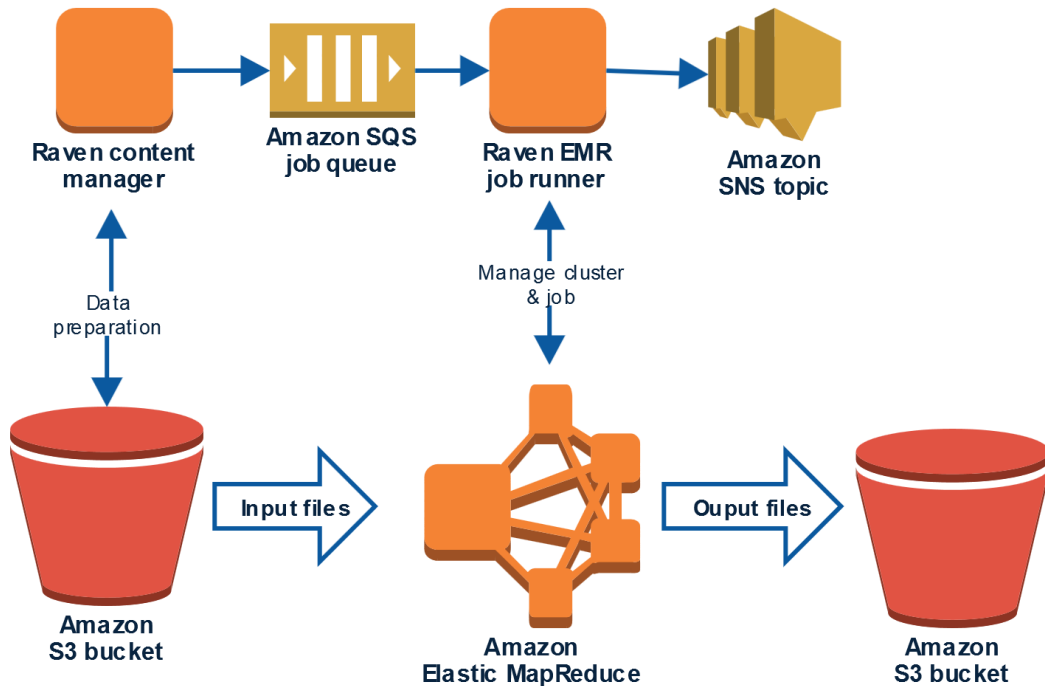
Five Patterns for Scalable Data Pipelines

Pattern 1 - Cluster-based Big Data Pipeline [1 of 2]



RAVEN Platform
Pattern 1 - Cluster-based
big data pipeline

JustGiving™



Pattern 1 - Cluster-based Big Data Pipeline [2 of 2]

Pros

- Spark job e.g. NLP, ML, Graph or ETL at scale
- The Spark job can contain several steps
- Data preparation, separated from Spark job
- EMR steps for sequencing

Cons:

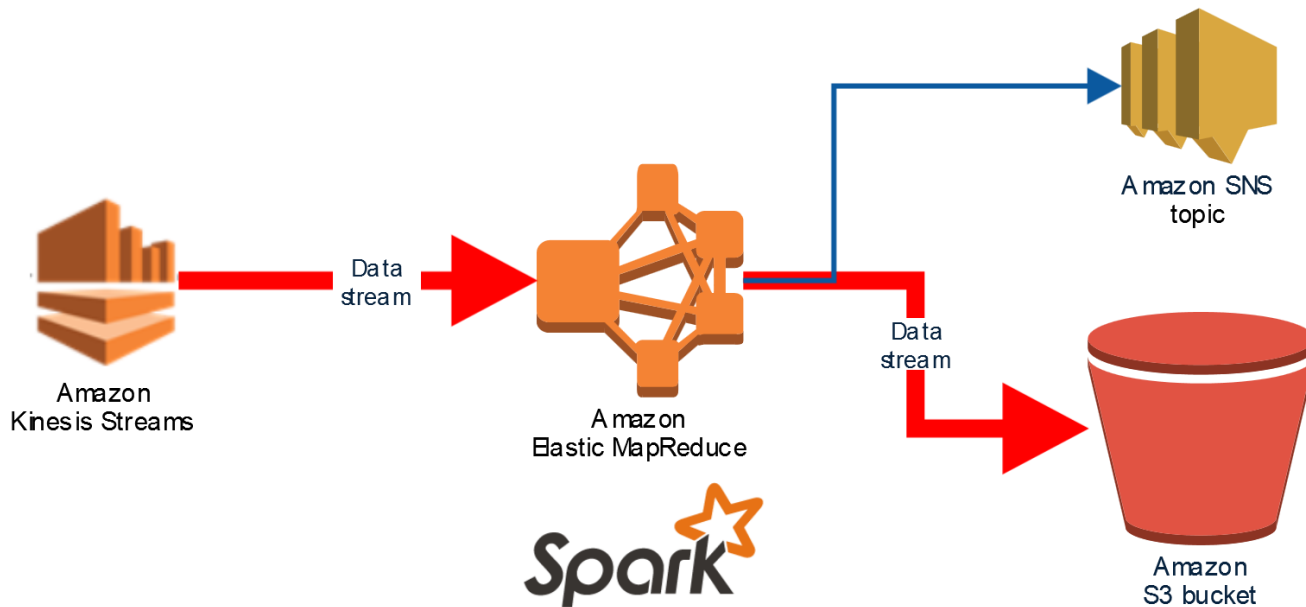
- The cluster launch time ~ 10 minutes
- EMR pay per hour
- Choosing EMR cluster size vs. costs

Pattern 2 - Cluster-based Streaming Events Pipeline [1 of 2]



RAVEN - Platform
Pattern 2 - Cluster-based
streaming events pipeline

JustGiving™



Pattern 2 - Cluster-based Streaming Events Pipeline [2 of 2]

Pros:

- Can process the stream in parallel and fault tolerant
- Rich Spark streaming libraries
- Real-time analytics pipeline

Cons:

- Always on EMR cluster
- Complex to resize
- Needs check-pointing

Serverless

AWS Lambda fully manages stateless compute containers that are event-triggered

- Benefits
 - Zero-maintenance and upgrade
 - Low cost
 - Automatic scaling
 - Secure

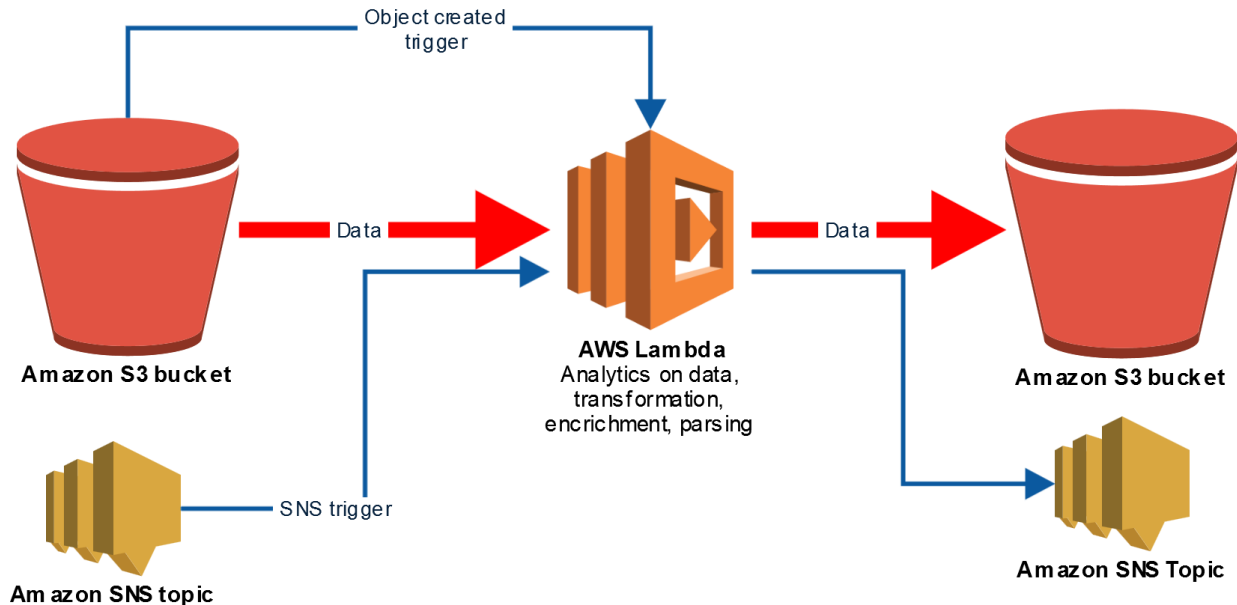


Pattern 3 - Serverless Small Data Pipeline [1 of 2]



RAVEN Platform
Pattern 3 - Serverless small
data pipeline

JustGiving™



Pattern 3 - Serverless Small Data Pipeline [2 of 2]

Pros:

- Batch and incremental possible
- Useful for small and infrequently added files
- Can preserve file name
- Can be used for projection, enrichment, parsing, etc.

Cons:

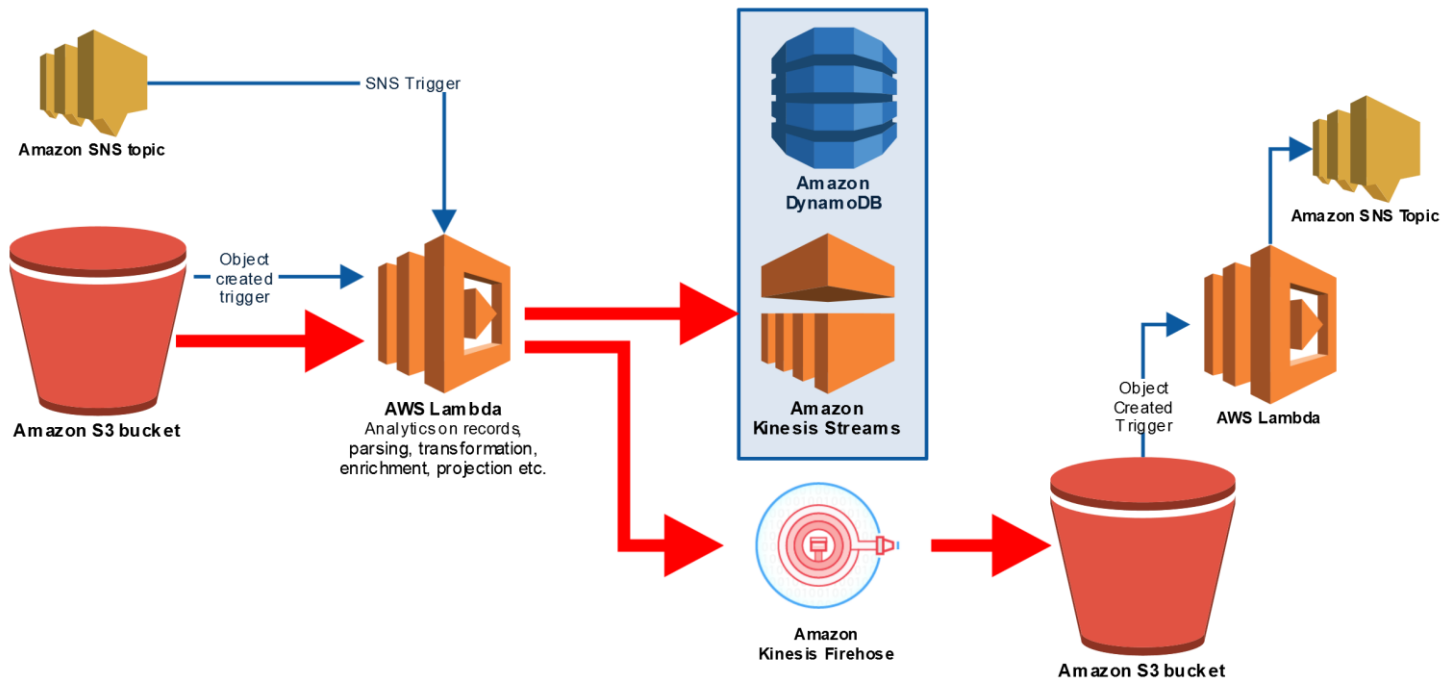
- Lambda limits: 500 MB disk, 1.5GB RAM, complete within 5min
- Complex joins
- One file on input and output

Pattern 4 - Serverless Streamify File and Merge into Larger Files [1 of 2]



RAVEN Platform
Pattern 4 - Serverless streamify
file and merge into larger files

JustGiving™



Pattern 4 - Serverless Streamify File and Merge into Larger Files [2 of 2]

Pros:

- Batch and incremental possible
- Useful for small and frequently added files
- Lambda and Amazon Kinesis Firehose - Serverless way to persist a stream

Cons:

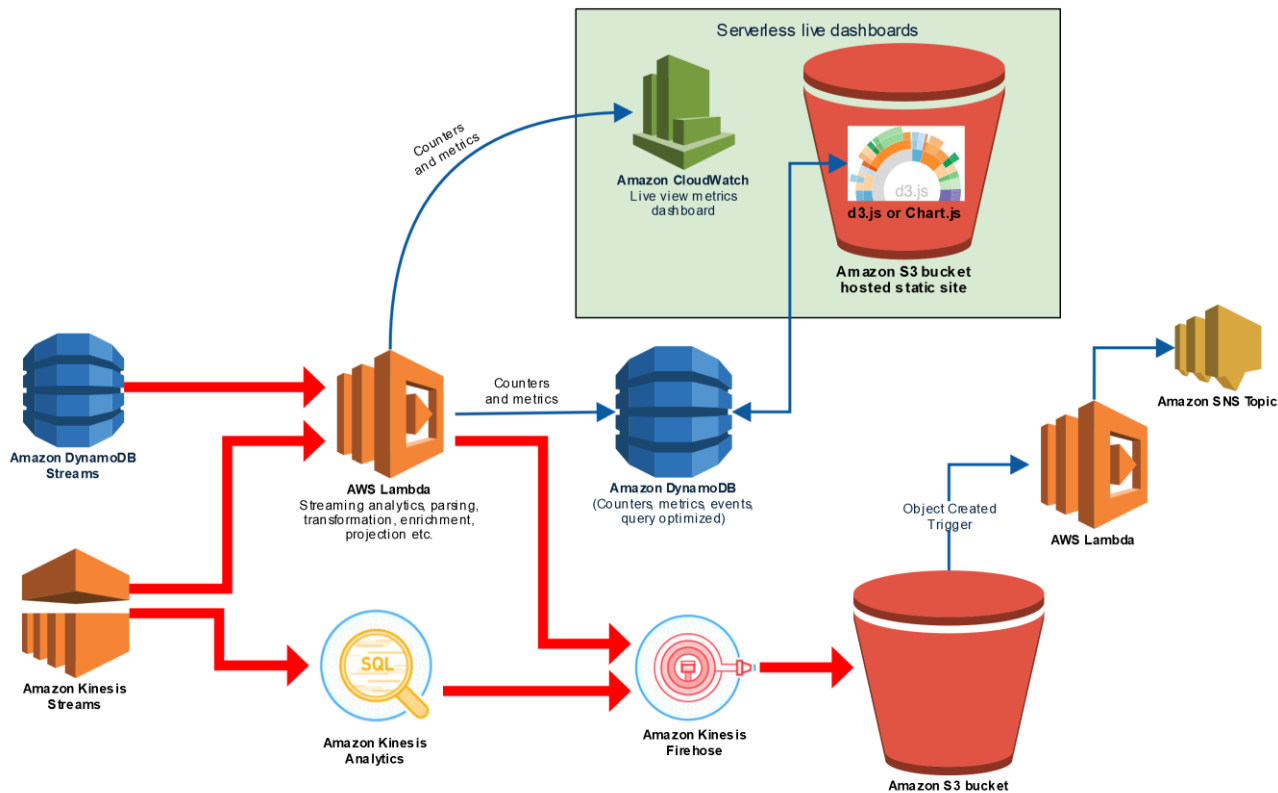
- Lambda limits: 500 MB disk, 1.5GB RAM, complete within 5min
- Firehose limits: 15 min, 128 MB buffer
- Complex joins

Pattern 5 - Serverless Streaming Analytics and Persist Stream [1 of 2]



RAVEN Platform
Pattern 5 - Serverless streaming
analytics and persisting stream

JustGiving™



Pattern 5 - Serverless Streaming Analytics and Persist Stream [2 of 2]

Pros:

- Stream processing without a running cluster
- Lambda and Amazon Kinesis Analytics automatic scaling
- A static website can access DynamoDB metrics
- Choice of language: Python, SQL, Node.js, and Java8
- Code can be changed without interruption

Cons:

- Lambda limits: 500 MB disk, 1.5GB RAM, complete within 5min
- Firehose limits: 15 min, 128 MB buffer
- Complex stream joins

Serverless Recommendations and Best Practices

Serverless Patterns

Pattern	Serverless	Spark on EMR when ...
Pattern 3 - Serverless small data pipeline	<ul style="list-style-type: none">• Small and infrequently added files• Preserves existing file names	<ul style="list-style-type: none">• Big files > 400MB• Joining data sources• Merge small files into one
Pattern 4 - Serverless streamify file and merge into larger files	<ul style="list-style-type: none">• Merge small frequently added files into a stream or larger ones in S3• Streaming analytics and time series	<ul style="list-style-type: none">• Big files > 400MB• Joining streams or data sources
Pattern 5 - Serverless analytics and persist stream	<ul style="list-style-type: none">• Streaming analytics and time series• Real-time dashboard• Persist an Amazon Kinesis stream	<ul style="list-style-type: none">• Joining streams or data sources

Lambda Functions Are Good For...

- Rows & events: parsing, projecting, enriching, filtering etc.
- Working with other AWS offerings: S3, SNS, CloudWatch, IAM, etc.
 - Lambda and Amazon Kinesis Firehose - Serverless way to persist a stream
 - Secure - IAM roles and VPC
- Simple packaging and deployment
- For streams, easily modify code



Understand the Lambda limits

- Memory, local disk, execution time
- Concurrent Lambda functions execution
 - AWS account
 - DynamoDB and Amazon Kinesis shard iterators
- Complex joins harder to implement - better suited for Spark on EMR
- ORC and Parquet - better support in Spark

Lambda Recommendations

- Test with production volumes
- Think deployment and version control
- Reduce execution time
 - Minimize logging
 - Optimize code



Event-Driven Pipeline

- Architect for incremental loads, and re-loads on failures
- Use AWS managed services and Serverless where possible
 - HA and scaling out
- Loosely coupled system
 - SNS/SQS, Lambda event sources
- Stateless systems
 - DynamoDB, S3, RDS
- DAGs can usually be flattened into sequences

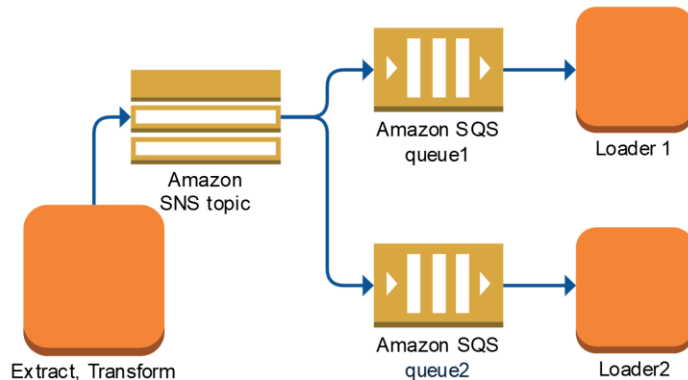
Decouple the L in ETL / ELT

Allows more flexibility

- Loosely coupled via SNS / SQS
- Extraction and transformation once
- Batch and incremental loads
- Load on many clusters

Build for exceptions

- Failed steps, ETL, EMR jobs
- Notification and dashboard
- Support reloads via messaging



Master Data in S3

- Data lake
 - Single source of truth
 - Separate storage and compute
- Make EMR and Amazon Redshift clusters disposable
 - Load / reload and regenerate tables
 - Support many clusters
- Data storage
 - One file type per S3 prefix (folder)
 - Incremental S3 prefix year/month/day/hour
 - If no metadata
 - Store data type in schema file
 - One consistent date time format
- Compress and encrypt the data



Find Out More

- [Serverless Cross Account Stream Replication Using AWS Lambda, Amazon DynamoDB, and Amazon Kinesis Firehose](#) (AWS Compute Blog)
- [Analyze a Time Series in Real Time with AWS Lambda, Amazon Kinesis and Amazon DynamoDB Streams](#) (AWS Big Data Blog)
- [Serverless Dynamic Real-Time Dashboard with AWS DynamoDB, S3 and Cognito](#) (Medium.com)
- <https://github.com/JustGiving>

Summary

- Challenges of existing big data pipelines
- Event-driven ETL and serverless pipelines at JustGiving
 - Alternative to DAGs and workflows
- Five patterns for scalable data pipelines
 1. Cluster-based big data pipeline
 2. Cluster-based Streaming events pipeline
 3. Serverless small data pipeline
 4. Serverless streamify file and merge into larger files
 5. Serverless streaming analytics and persist stream
- Recommendations
 - Serverless and managed services
 - Decouple the L in ETL





**AWS
re:Invent**

Thank you!

JustGiving™

“Ensure no good cause goes unfunded”

Contact:

richard.freeman

@justgiving.com

[https://linkedin.com/in/
drfreeman](https://linkedin.com/in/drfreeman)



**Remember to complete
your evaluations!**

Related Sessions

- BDA201 - Big Data Architectural Patterns and Best Practices on AWS
- BDA301 - Best Practices for Apache Spark on Amazon EMR
- BDA302 - Building a Data Lake on AWS
- BDA402 - Building a Real-Time Streaming Data Platform on AWS
- ARC402 - Serverless Architectural Patterns and Best Practices