

# Apache Hudi on Amazon EMR Readiness Workshop



# Special OFFERS

**Free Data Lake Assessment**  
for all eligible attendees.

Get started in as little as  
72 hours.



# Apache Hudi Readiness Workshop

## PRESENTERS



**Kireet Kokala**

VP, Big Data & Analytics



**Kalen Zhang**

Partner Solutions Architect  
Data Analytics Segment



**Fernando Gonzalez**

Senior DevOps  
Engineer



**Deepu Mathew**

Senior DevOps  
Engineer





Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 175 fully featured services from data centers globally.

Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.



nClouds is a deeply credentialed, award-winning provider of AWS and DevOps consulting and implementation services. We are an integrated team of skilled engineers, architects, developers, project managers, and sales & marketing professionals who are passionate about client success, software excellence, and innovation. We enable our clients to deliver better products faster and create awesome customer experiences.





## DevOps & Infrastructure Modernization

- ◆ CI/CD pipelines
- ◆ Containers & microservices
- ◆ DevOps-as-a-Service



## Cloud Migration Services

- ◆ Migration Readiness Assessment
- ◆ From simple lift & shift (rehosting) to re-architecting and refactoring
- ◆ CloudChomp C3 Certified Partner
- ◆ VMware/Windows/Linux/Database



## Data & Analytics

- ◆ DataOps: Athena, Aurora, Glue, QuickSight, Data Warehouse, Data Lake, Hadoop, Redshift, ETL / ELT
- ◆ ML & AI: SageMaker, AI, Deep Learning, Alexa



## nOps (SaaS) Cloud Management

- ◆ AWS Well-Architected Reviews
- ◆ Cost optimization
- ◆ Security review

# Trusted by INNOVATIVE BRANDS



# Apache Hudi Readiness Workshop

## AGENDA



### DETAILS *(All times EDT)*

- **1:00 - 1:10 pm** - Intro & Workshop Objectives *by Kireet Kokala, nClouds*
- **1:10 - 1:20 pm** - Architecture: Apache Hudi on Amazon EMR *by Kalen Zhang, AWS*
- **1:20 - 1:30 pm** - Apache Hudi Use Cases *by Fernando Gonzalez, nClouds*
- **1:30 - 1:50 pm** - Demo: Apache Hudi on Amazon EMR *by Deepu Mathew, nClouds*
- **1:50 - 1:55 pm** - Getting Started: Apache Hudi Readiness & Process *by Kireet Kokala, nClouds*
- **1:55 - 2:00 pm** - Q&A

# Apache Hudi Readiness Workshop

## OBJECTIVES



### Apache Hudi on Amazon EMR

Use Cases, Architecture,  
Demo



### Readiness Assessment

Process, Cost Clarity,  
Timing



### nClouds Data & Analytics

Services, Benefits,  
Identifying Next Steps



# Apache Hudi History

## Hadoop Upserts and Incrementals

- Apache Hudi brings stream processing to big data, providing fresh data while being an order of magnitude efficient over traditional batch processing.
- Propped onto the public scene in 2016
- Early adopters saw companies like Uber use Apache Hudi to build large scale, near-real-time pipelines
- In 2020, nClouds uses Apache Hudi to dissect COVID-19 data

# Apache Hudi Value Proposition

- Ease of incremental data processing to handle “time skew.”
- Build more robust and fresh data lakes providing high quality insights by enforcing schematization on data sets.
- Take control of data lakes via seamless ingestion and management of large analytical data sets over distributed file systems.
- Read our recent blog post, *How to accelerate delivery with Apache Hudi on Amazon EMR* [here](#).

# Apache Hudi on Amazon EMR

- **Apache Hudi** is automatically installed in Amazon EMR when you choose Apache Spark, Hive, or Presto when deploying a cluster.
- **You can handle either read or write-heavy use cases**, and Hudi will manage the underlying data stored on S3 (Parquet and Avro).
- **Datasets managed by Hudi** are accessible not only from Spark (and PySpark), but also other engines such as Hive and Presto.
- **Native integration with AWS Database Migration Service** also provides another source for data as it changes.





# Architecture: Apache Hudi on Amazon EMR

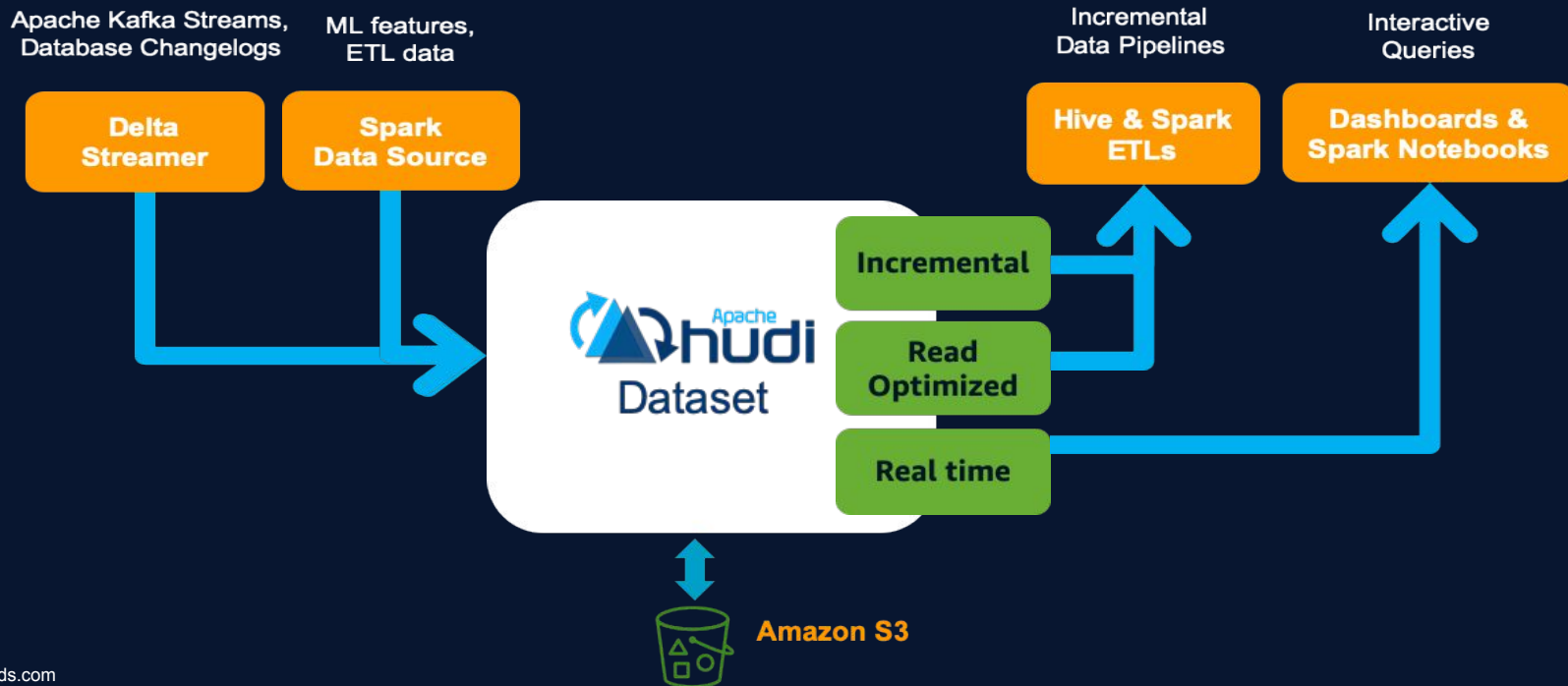


**Kalen Zhang**

Partner Solutions Architect, Data & Analytics Segment



# Apache Hudi Overview



# Apache Hudi Architecture



Hudi Spark  
Data Source



Store & Index  
Data



Index



Data Files



Timeline  
Metadata



Read  
Data



presto



# Apache Hudi Storage Type



**Copy On Write**  
**Read Heavy**



**Merge On Read**  
**Write Heavy**



**Hudi Dataset**

# Apache Hudi **Use Cases**



**Fernando Gonzalez**  
Senior DevOps Engineer





# Apache Hudi Use Cases



## Near-Real-Time Ingestion

- ◆ Compliance with privacy regulations.
- ◆ Upsert late-arriving data into an existing dataset in Amazon S3.
- ◆ Streaming data ingestion, to avoid creating many small files.
- ◆ Hudi provides faster loads via Upserts.



## Near-Real-Time Analytics

- ◆ Interactive SQL solutions on Hadoop such as presto & SparkSQL excel in queries that finish within few seconds.



## Querying of Amazon S3 data sets

- ◆ Directly to provide users with a near-real-time view of your data.



## Analysis of Data

- ◆ As of a specific point in time.

# Use Case POC

- Our use case demo will track COVID-19 cases by date in each U.S. state and show how near-real time the changes to data are updated with Apache Hudi DeltaStreamer.

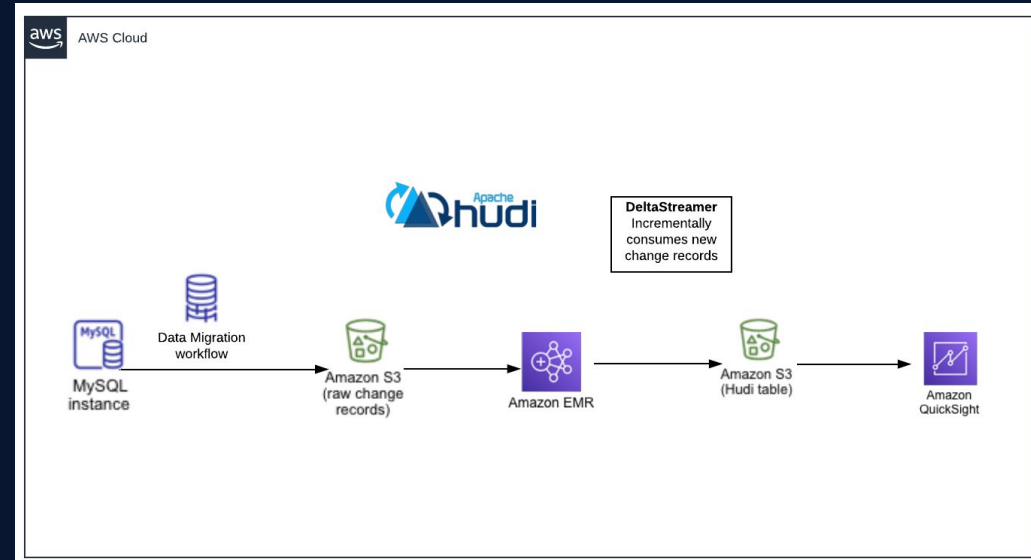
# Apache Hudi

## POC Architecture Review

### 1. Apache Hudi POC Components

Environment needed to run:

- Amazon Relational Database Service (Amazon RDS)
- Amazon Database Migration Service (Amazon DMS) task
- Amazon Elastic MapReduce (Amazon EMR) cluster
- Amazon Simple Storage Service (Amazon S3) buckets



High-level view of the end-to-end architecture

# Apache Hudi

## POC Architecture Review

### 2. Apache Hudi POC Dataset

- ◆ covid\_by\_state datasets from kaggle.com
- ◆ COVID19 data for the U.S. from January 2020 has been ingested for the POC

```
CREATE TABLE covid_by_state(covid_by_state_id INTEGER
NOT NULL AUTO_INCREMENT,date TIMESTAMP DEFAULT
NOW() ON UPDATE NOW(),state VARCHAR(100),fips
INTEGER,cases INTEGER,deaths INTEGER,CONSTRAINT
orders_pk PRIMARY KEY(covid_by_state_id));
```

```
INSERT INTO covid_by_state( date , state, fips, cases,
deaths) VALUES('2020-01-21','Washington',53,1,0);
```

```
INSERT INTO covid_by_state( date , state, fips, cases,
deaths) VALUES('2020-01-21','Illinois',17,1,0);
```

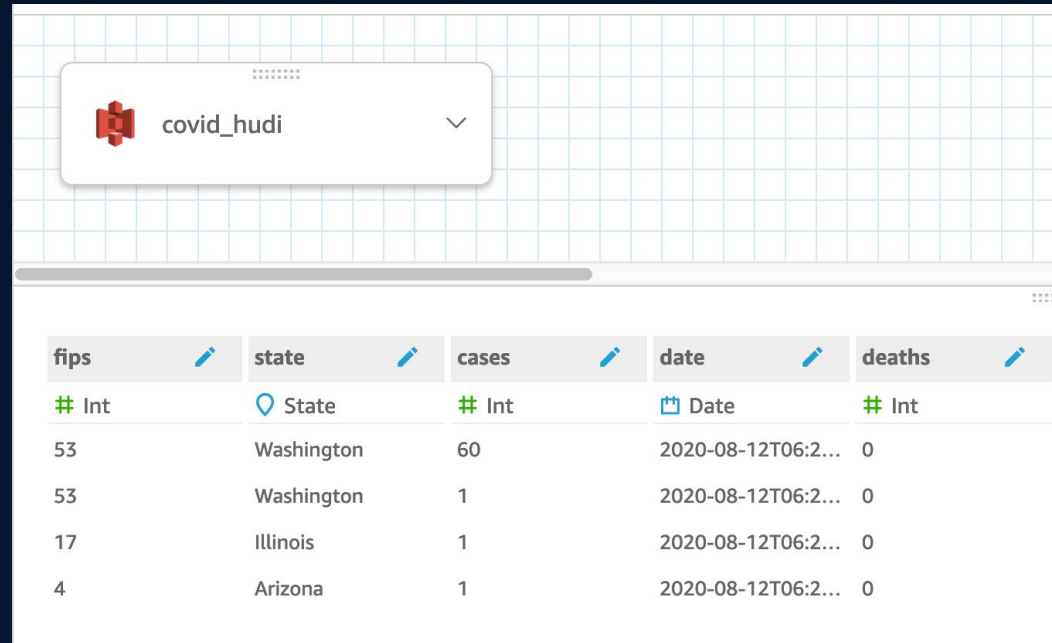
**Schema creation statements**

**Data insertion statements**

# Apache Hudi POC Visualization

## 3. Quicksight Driven Visualization

- ◆ Dashboard view of the data breaks down the number of cases on a certain date by state.



The screenshot shows a Quicksight dashboard interface. At the top, there is a dropdown menu with a red icon and the text 'covid\_hudi'. Below this is a table with the following data:

fips	state	cases	date	deaths
# Int	State	# Int	Date	# Int
53	Washington	60	2020-08-12T06:2...	0
53	Washington	1	2020-08-12T06:2...	0
17	Illinois	1	2020-08-12T06:2...	0
4	Arizona	1	2020-08-12T06:2...	0

# Apache Hudi on **Amazon EMR**



**Deepu Mathew**  
Senior DevOps Engineer



# Apache Hudi Demo



Setup  
Overview



Datasets Used



Results



# Demo



# Getting Started: Apache Hudi Readiness & Process



**Kireet Kokala**

VP, Big Data & Analytics



# Apache Hudi **Readiness & Process**

## Readiness

### 1. Assessing current footprint

- ◆ A combination of architectural, DevOps tools such as nOps, and technical interviews to determine technical landscape.
- ◆ Often via discovery / assessment.

### 2. Data in play

- ◆ Analyze customer data sets around velocity, variety, volume.
- ◆ Determine data security (of data at motion and rest).
- ◆ Hudi managed data will be accessible from Apache Spark and Hive, Presto, etc. We will try to integrate it with the AWS ecosystem via the Data Migration Service. Example: Quicksight for embedded visualization solutions.

# Apache Hudi Readiness & Process



## Process

### 3. Apache Hudi Process, Cost Clarity, Timing

- ◆ Based on our customer's footprint, the nClouds assessment, and the nature of the data of your use case, we provide a Vision document within 48-72 hours.
  - Customer footprint assessment.
  - Recommended high-level solutions.
  - Professional services to design and integrate Hudi into your solution.
  - Examples of deliverables can be found in our latest APN blog [here](#).

# nClouds Data & Analytics



## Assessments

- ◆ Analytics Ecosystem (tooling)
- ◆ Business Intelligence Strategy
- ◆ Solution Architecture Modernization
- ◆ Data Lakes and Data Warehouses



## Data Lakes / Data Warehouse

- ◆ POC
- ◆ Enablement and Implementation
- ◆ BI integration



## Data Movement

- ◆ ETL / ELT
- ◆ Implementation
- ◆ Reporting integration



## Machine Learning

- ◆ POC
- ◆ Sagemaker migration
- ◆ Services integration
- ◆ Solution acceleration



**Q&A**

# Apache Hudi Readiness Workshop

## PRESENTERS



**Kireet Kokala**

VP, Big Data & Analytics



**Kalen Zhang**

Partner Solutions Architect  
Data Analytics Segment



**Fernando Gonzalez**

Senior DevOps  
Engineer



**Deepu Mathew**

Senior DevOps  
Engineer



# Data Lake as Code on AWS Implementation Workshop

Tuesday, November 10 at 10 am PT

[REGISTER HERE](#)

