## Hello Hydrate! From Stream to Clickhouse with Apache Pulsar and Friends





1



**Tim Spann** Developer Advocate

#### DZone Zone Leader and Big Data MVB Data DJay

- https://www.datainmotion.dev/
- <u>https://github.com/tspannhw/SpeakerProfile</u>
- <u>https://dev.to/tspannhw</u>
- <u>https://sessionize.com/tspann/</u>







Stream Native

- Founded the original developers of Apache Pulsar.
- Passionate and dedicated team.
- StreamNative helps teams to capture, manage, and leverage data using Pulsar's unified messaging and streaming platform.



#### The Need For Real-Time Data



Hybrid and multi-cloud strategies with native geo-replication





**Built for Kubernetes** 

CloudNative migrations with tools



Seamlessly build microservice architectures with support for streaming and messaging workloads

#### 360 degree customer data

multi-tenancy, infinite retention, and extensive connector ecosystem





# **PULSAR**



# **Apache** *PULSAR* is an open source, cloud-native distributed messaging and streaming platform.







#### Apache Pulsar





### A Unified Messaging Platform







### Apache Pulsar Overview

- Pub-Sub
- Geo-Replication
- Pulsar Functions
- Horizontal Scalability
- Multi-tenancy
- Tiered Persistent Storage
- Pulsar Connectors
- REST API
- CLI
- Many clients available
- Four Different Subscription Types
- Multi-Protocol Support
  - MQTT
  - AMQP
  - JMS
  - Kafka
  - o ...





### Pulsar's Publish-Subscribe model

- Producers send messages.
- Topics are an ordered, named channel that producers use to transmit messages to subscribed consumers.
- Messages belong to a topic and contain an arbitrary payload.
- Brokers handle connections and routes messages between producers / consumers.
- Subscriptions are named configuration rules that determine how messages are delivered to consumers.
- Consumers receive messages.





### What is the Pulsar Ecosystem?

- Functions and Connectors
  - Functions: Lightweight stream processing
  - Connectors: Part of "Pulsar IO", includes "Source" and "Sink" APIs
    - Files, Databases, Data tools, Cloud Services, etc
- Protocol Handlers
  - Allows Pulsar to handle additional protocols by an extendable API running in the broker
    - AoP (AMQP), KoP (Kafka), MoP (MQTT)



#### Topics









### MQTT on Pulsar (MoP)







#### Kafka-on-Pulsar (Kop)

<u>کہ</u> چچ





### Moving Data In and Out of Pulsar

IO/Connectors are a simple way to integrate with external systems and move data in and out of Pulsar.

- Built on top of Pulsar Functions
- Built-in connectors <u>hub.streamnative.io</u>





### Moving Data Out of Pulsar to Clickhouse

IO/Connectors are a simple way to integrate with external systems and move data in and out of Pulsar. <u>https://pulsar.apache.org/docs/en/io-jdbc-sink/</u>

- Built on top of Pulsar Functions
- Built-in connectors <u>hub.streamnative.io</u>





### Streaming Events into Altinity Cloud

```
CREATE TABLE iotjetsonjson local
      uuid String, camera String, ipaddress String,
                                                         networktime String,
                                                                                    top1pct String,
      top1 String, cputemp String, gputemp String, gputempf String,
      cputempf String,
                         runtime String,
      host String, filename String,
                                      host name String,
                                                              macaddress String,
      te String, systemtime String, cpu String,
                                                        diskusage String,
      memory String,
                         imageinput String
ENGINE = MergeTree()
  PARTITION BY uuid
 ORDER BY (uuid);
CREATE TABLE iotjetsonjson ON CLUSTER '{cluster}' AS iotjetsonjson local
```

https://docs.altinity.com/altinitycloud/guickstartguide/connectclient/

ENGINE = Distributed('{cluster}', default, iotjetsonjson local, rand());

pulsar-io-jdbc-clickhouse-2.8.0.nar



#### Build Cluster in Altinity Cloud

Cluster Launch Wizard	Resources Configuration	×			
	Name *				
1 Resources Configuration	streamnative				
	Cluster name tag will be used in ClickHouse configuration and it may o	ontain only lowercase letters [a-z], numbers [0-9] and hyphen [-]			
2 High Availability Configuration					
	Node Type				
3 Connection Configuration	m5.large (CPU x2, RAM 7 GB)				
	Node Type will be the same across all ClickHouse hosts				
4 Review & Launch	Node Storage (GB)	Number of Volumes			
	100	1			
	Each ClickHouse host will have specified amount of local volume storage	Network storage can be split to several volumes for a better query performance			
	Volume Type				
	gp2-encrypted (Encrypted)				
	Defines volume claim storage class for each ClickHouse host				
	Number of Shards (nodes per replica)				
	1				
	Each shard will require X number of ClickHouse hosts where X is the number of replicas of this shard (X = 2)				
	ClickHouse Version				
	21.8.10.19 Altinity Stable V				
	ClickHouse Version will be the same across all Cluster nodes				
	ClickHouse User Name				
	admin				
	ClickHouse user will be created with the specified login				
	ClickHouse User Password *	Confirm Password *			
	Enter Password	Confirm Password			
	This password will be assigned to the ClickHouse User The minimum password length is 12 characters. Consider adding digits, capital letters and special symbols to make password more secure	Please confirm the password			

#### Altinity.Cloud streamnative (1/1 nodes online) Access Point 🥑 Health 6/6 checks passed) Shards 1 Replicas 1 Node Storage 100 GB **Node Memory** 7 GB Node CPU 2 Version 21.8.10.19 Latest Backup N/A ≡ CONFIGURE ¥ EXPLORE ACTIONS ~



### Run A Cluster in Altinity Cloud

#### osacon2021 overview

Туре	Kubernetes
ClickHouse Clusters	1
ClickHouse Nodes	1
Zookeeper Clusters	1
Zookeeper Nodes	1
Node Types	6

#### memory





Kubernetes Namespace	osacon2021
Kubernetes Nodes	3
Availability Zones	us-east-1a, us-east-1b, us-east-1c
CH Operator	Version 0.15.0
Monitoring	View in Grafana





UsedFree



#### Run A Cluster in Altinity Cloud

	tinity.Cloud								
∽ васк	CLUSTER	: STREAMNATIVE	· () F	Run DDLs ON CLUSTER	NODE: ANY Y				
Query	Schema P	rocesses							
Query Histo	ry: <	>							
select *	from stocks								
				I					
EXECUTE									
streamna	tive.osacor	2021.altinity.cl	oud:8443 (q	uery time: 0.029s)					
streamna	tive.osacor	2021.altinity.cl	oud:8443 (q	uery time: 0.029s)	open	-close	-high	-volume-	low
streamna	tive.osacor	12021.altinity.cl	oud:8443 (q dt	uery time: 0.029s) 	open			volume- 2198	low
streamna	tive.osacor uuid 6bec81c6 6bec81c6	2021.altinity.cl -1469658966 -1469658966	Dud:8443 (q dtdt	uery time: 0.029s) datetime 2021/01/22 10:06:00 2021/01/22 10:06:00	open- 340.83099 340.83099	-close	-high- 341.38000 341.38000	-volume- 2198 2198	low
streamna ABC IBM	tive.osacor	2021.altinity.cl -1469658966 -1469658966 -1469558966	0ud:8443 (q dt 715224000 715224000 dt	uery time: 0.029s) datetime 2021/01/22 10:06:00 2021/01/22 10:06:00	open- 340.83099 340.83099	close 341.38000 341.38000	-high- 341.38000 341.38000 1	volume 2198 2198 	low



#### Monitor The Cluster in Altinity Cloud

Ô	器 osacon2021 / ClickHouse in Kubernetes Querie	s «ổ					Ģ	<ul> <li>Last 1 hou</li> </ul>	r ∨ ⊖ ℃ 5s ∨
	K8S Namespace osacon2021 ~ K8S Clickhouse Installation	streamnative ~ type	All - top elements 5 - in	nitial user All - query type	all -				
Q	~ Top charts								
88		Top 5	request's rate by type: All; user: All; qu	ery type: all					Types
							max avg ~	1 - successful st	art of query execution
					<ul> <li>INSERT INTO sta</li> </ul>	cks_local(symbol,uuid,ts,dt,da		2 - successful er	d of query execution
	75				- INSERT INTO iot	jetsonjson_local(uuid,camera,i		2 - exception het	ore start of quepu
					<ul> <li>select database</li> <li>SELECT database</li> </ul>	name, engine from system tab		execution	ore start of query
	50				- select timezone	) FORMAT TabSeparatedWithName		4 - exception wh	le query execution
	25								
	0 20:15 20:20 20:25 20:30 20:35	20:40 20:45	20:50 20:55 21:00	21:05 21:10					
	Top slow queries by type: All; user: All; query typ	pe: all	Top memory consumers	s by type: All; user: All; query ty	pe: all	Top f	ailed queries by us	er: All; query type:	all
	/* ddl_entry=query-000000064 */ DROP TABLE default.stocks_local	77.00 2 ms 2	INSERT INTO iotjetsonjson_local(uuid,camera,ipa	address,networktime,top1	2.09 MiB 10	INSERT INTO stocks_local(symbol,uui	d,ts,dt,datetime,op	en,close,high,vo	ExceptionBeforeStart
	/* ddl_entry=query-0000000059 */ CREATE TABLE default.stocks UUID 'e15	3.50 ms 2	INSERT INTO stocks_local(symbol,uuid,ts,dt,date	time,open,close,high,vo	2.04 2 К МіВ 2 К	/* ddl_entry=query-0000 default.iotjetsonjson	000077 */ DROP T	ABLE	ExceptionBeforeStart
	/* ddl_entry=query-0000000057 */ CREATE TABLE IF NOT EXISTS default.st	2.50 ms 2	SELECT database, table, name, type default_type, defa	e, default_kind as	668.00 B 4	/* ddl_entry=query-0000 default.iotjetsonjson_lo	000081 */ DROP T :a	ABLE	ExceptionBeforeStart
	/* ddl_entry=query-000000068 */ CREATE TABLE IF NOT EXISTS default.st	2.50 ms 2	select database, name, engine fron 1 and name	n system.tables where 1 =	348.00 4 B	/* ddl_entry=query-0000 default.iotjetsonjson UU	000074 */ CREATE	TABLE	ExceptionBeforeStart
	/* ddl_entry=query-0000000073 */ CREATE TABLE default.iotjetsonjson UU	2.00 ms 2	select version() FORMAT TabSepar	atedWithNamesAndTypes;	316.00 B 4	/* ddl_entry≖query-0000 default.stocks	000063 */ DROP T	ABLE	ExceptionBeforeStart
	~ Request charts								
æ	Reqs/s by type: All;	user: All; query type: all				Query duration by type:	All; user: All		
~				1.2 ms					3.0 ms
(?) (?)	75			1.0 ms					- 2.5 ms



#### Monitor The Cluster in Altinity Cloud







### StreamNative Cloud





#### How Companies Use StreamNative today





#### StreamNative Cloud

Powered by Apache Pulsar, StreamNative provides a cloud-native, real-time messaging and streaming platform to support multi-cloud and hybrid cloud strategies.

















StreamNative	⊕International ∨ Organizat	tion - sndev \vee 🛛 Instance - g	ke $\sim$						0	tim.spann@streamnative.io $\vee$
gke										
& Tenants	Tenant public	Namespace	default $\vee$							
ANamespaces	OVERVIEW TOPICS P	OLICY		*						
∲.Topics										
	G Search Topic		+ New Topic	-						
⊟Clients	Торіс	Partitions	Domain	Producers	Subscriptions	In Rate	Out Rate	In Throughput	Out Throughput	Storage Size
8 Connector ~	> product	5	persistent	0	0	€ 0	€ o	-Ð 0	€ o	€ o
& Manage	> iotjetsonjson	0	persistent	0	0	<del>년</del> 0	년 0	-Ð 0	<b>e</b> o	- E 0
<b>⊜</b> Service Accounts	> jetsoniotts2	5	persistent	0	0	- C 0	0 B	-D 0	- C	-E o
	> kinesis-output	0	persistent	0	0	- C	0 ট	-D 0	<b>G</b> 0	€ o
≪Pulsar Clusters	> testi	0	persistent	0	0	0 D	€ o	- <b>D</b> o	<b>O</b> O	Ðo
Beinatance Setting	> jetsoniotts	5	persistent	0	0	Ð o	0 ট	- <b>D</b> O	Ð o	€o
	> kinesis-input	0	persistent	0	0	<b>⊖</b> o	€ o	-D 0	<b>e</b> o	Ðo
	> sensors	0	persistent	0	0	<b>⊖</b> 0	€ o	-D o	<b>e</b> o	€o
	> test3	0	persistent	0	0	<b>€</b> 0	Ð 0	-D 0	<b>e</b> o	€ o
	> (TENANT_NAMESPACE_TO PIC)	0	persistent	0	0	Ðo	<del>ପ</del> ୦	Ðo	Ðo	-E o
	> topitems3	5	persistent	0	0	<b>⊖</b> o	€ o	-D 0	<b>e</b> o	€ o
	> data-gen-out	0	persistent	0	0	€ o	€ o	-Ð 0	€ o	<b>⊖</b> 0
	5 TH 112	e -	aut a	2					<b>2</b> -	





StreamNative	⊕International ∨ Organization - sndev ∨ Instance - a	ike 🗸			🚺 tim.spann@streamnative.io ~
gke &Tenants	Tenant public Vamespace defi	ault V <b>Topic</b> iotletson/son V			
▲Namespaces	OVERVIEW SCHEMA MESSAGES STORAGE	POLICIES			
* Topics					
SQL Preview	Storage Size		Entries		Segments
€Clients	$\frown$		$\frown$		
8 Connector ~	2 MB		2К		17
≗ Manage ^					
 @Flink Clusters					
«Pulsar Clusters	Segments				
翻Instance Setting	Ledger ID	Entries	Size	Status	Offload
	1342	5	4.51K	closing	false
	1353	1	882.00	closing	false
	1365	1	940.00	closing	false
	1378	2	1.9K	closing	false
	1386	1	898.00	closing	false
	1975	3	2.79K	closing	false
	2007	1	898.00	closing	false
	2284	1	805.00	closing	false







### Best Practice Architectures



### Example: E-Commerce with Pulsar

- Unified storage with access to underlying data
- Native tiered storage
- Single system to exchange data
- Teams share toolset





### End-to-End Streaming FLiP(N) Apps

Apache Flink - Apache Pulsar - Apache NiFi <-> Events <-> Clickhouse







#### Demo

https://github.com/tspannhw/FLiP-Stream2Clickhouse/





#### IoT Data

**IoT Ingestion:** High-volume streaming sources, sensors, multiple message formats, diverse protocols and multi-vendor devices creates data ingestion challenges.

**Other Sources:** Transit data, news, twitter, status feeds, REST data, stock data and more.





#### **REST JSON "stonks" Events**



{"symbol":"STREAM", "uuid":"10640832-f139-4b82-8780-e3ad37b3d0 ce". "ts":1618529574078. "dt":1612098900000, "datetime":"2021/01/3108:15:00", "open":"12.24500", "close":"12.25500", "high":"12.25500", "volume":"12353", "low":"12.24500"}



stonks

### Wrap-Up



#### Now Available On-Demand Pulsar Training

Academy.StreamNative.io



Platform Engineer [Remote]	San Francisco
Platform Engineer (Flink/Spark) [Remote]	San Francisco
Product Engineer - Cloud [Remote]	San Francisco
Platform Engineer (Flink/Spark) [Remote]	San Francisco
Product Engineer - Cloud [Remote]	San Francisco
Sr. Product Manager [Remote]	San Francisco

#### We're Hiring

streamnative.io/careers/

#### Connect with the Community & Stay Up-To-Date

- Join the Pulsar Slack channel Apache-Pulsar.slack.com
- Follow @streamnativeio and <u>@apache\_pulsar</u> on Twitter
- <u>Subscribe</u> to Monthly Pulsar Newsletter for major news, events, project updates, and resources in the Pulsar community



#### **Interested In Learning More?**





Resources

Flink SQL Cookbook

The Github Source for Flink SQL Demo

The GitHub Source for Demo

Free eBooks

<u>Manning's Apache Pulsar in</u> <u>Action</u>

**O'Reilly Book** 



#### **Upcoming Events**

[11/8] PASS Data Community

[11/18] Developer Week Austin

[11/19] Porto Tech Hub Con

[12/3] Data Science Camp



#### Deeper Content

- <u>https://www.datainmotion.dev/2020/04/building-search-indexes-with-apache.html</u>
- <u>https://github.com/tspannhw/nifi-solr-example</u>
- <u>https://github.com/streamnative/pulsar-flink</u>
- <u>https://www.linkedin.com/pulse/2021-schedule-tim-spann/</u>
- <u>https://github.com/tspannhw/SpeakerProfile/blob/main/2021/talks/20210729 HailHydr</u> <u>ate!FromStreamtoLake TimSpann.pdf</u>
- <u>https://streamnative.io/en/blog/release/2021-04-20-flink-sql-on-streamnative-cloud</u>
- https://docs.streamnative.io/cloud/stable/compute/flink-sql



https://www.pulsardeveloper.com/



# Let's Keep in Touch!



Tim Spann

Developer Advocate



@PassDev



https://www.linkedin.com/in/timothyspann



https://github.com/tspannhw



# Questions



# Thank You!

orthsomet

AMALTICS